Enumerating the number of RNA structures

Mohammad GANJTABESH[1,2]
Armin MORABBI[1]
Jean-Marc STEYAERT[1]

mohammad.ganjtabesh@polytechnique.edu

1. Laboratoire d'Informatique, Ecole Polytechnique, 91128 Palaiseau Cedex, France.

2. Department of Computer Science, University of Tehran, Tehran 14155-6455, Iran.

Journées ARENA -27 March, 2007 - Lille

Table of Contents

- Primary Structure
  An RNA molecule is a sequence of nucleotides of four possible types, denoted by the letters A, C, G and U , connected by a backbone and is called RNA Primary Structure.

- Base Pairing
  Two nucleotides that are connected via hydrogen bonds are called a base pair. In the Watson-Crick base pairing, A always forms a base pair with U, as does G with C. In the Wobble base pairing, G forms a base pair with U.

- Notation
  An RNA sequence of length n is assumed as a sequence of n points $(1 - 2 - \cdots - n)$, in which each point i, $1 < i < n$, is connected to $i - 1$ and $i + 1$. We write $i.j$ if the nucleotide i is paired with the nucleotide j and $i < j$.

- **RNA Structure**
  An RNA structure is a set S of base pairs $i.j$ with $1 \leq i < j \leq n$, such that $\forall i_1.j_1, i_2.j_2 \in S : i_1 = i_2 \Leftrightarrow j_1 = j_2$. Each base can thus take part in at most one base pairing.

- **Secondary Structure**
  The set S is called secondary structure if $\forall i_1.j_1, i_2.j_2 \in S$ they are nested, i.e. $i_1 < i_2 < j_2 < j_1$, or disjoint, i.e. $i_1 < j_1 < i_2 < j_2$.

- **Pseudoknotted Structure**
  Two base pairs $i_1.j_1, i_2.j_2 \in S$ form a pseudoknot if $i_1 < i_2 < j_1 < j_2$ and S is called a pseudoknotted structure if it contains at least two base pairs which form a pseudoknot.

Enumerating RNA secondary structures

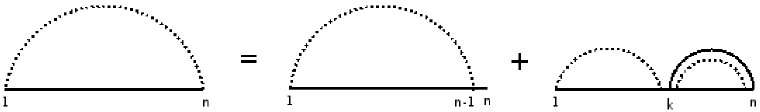
Theorem ([Waterman78])

Let $S(n)$ be the number of secondary structures for n points. Then $S(1) = s(2) = 1$, and for $n > 2$, $S(n)$ satisfies

$$S(n) = S(n-1) + \sum_{k=1}^{n-2} S(k-1)S(n-k-1),$$

where $S(0) \equiv 1$. Also, $S(n) \geq 2^{n-2}$ for $n \geq 2$.

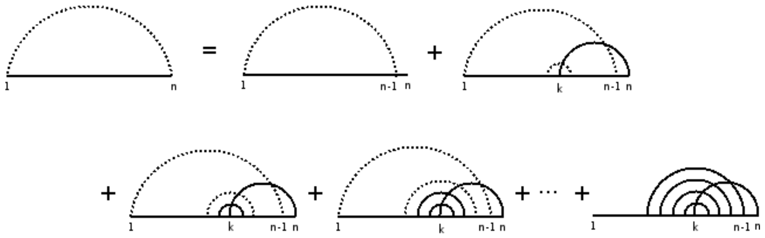
The so-called Catalan numbers...



Enumerating RNA structures

Suppose $P(n)$ denotes the number of RNA structures for a sequence of length n. There are two situations:

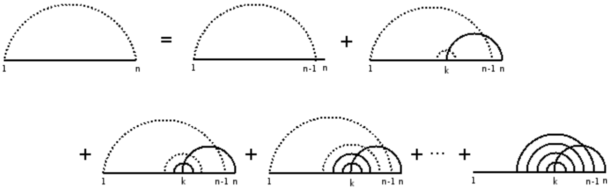
- The last point does not forms a base pairing. In this situation we have $P(n-1)$ structures.
- The last point forms a base pair with another point(say k where $1 \leq k \leq n-2$). In this situation, there are some extra structures! Not the same order of magnitude!



Enumerating RNA structures

Hence, in order to calculate $P(n)$, we can use the following formula:

$$P(n) = \begin{cases} 1 & \text{if } 0 \leq n \leq 2, \\ 2 & \text{if } n = 3, \\ P(n-1) + \sum_{k=1}^{n-2} \left(\sum_{t=0}^{min(k-1,n-k-1)} P(n-2t-2) \right) & \text{otherwise.} \end{cases}$$



Enumerating RNA structures

Hence, in order to calculate $P(n)$, we can use the following formula:

$$P(n) = \begin{cases} 1 & \text{if } 0 \leq n \leq 2, \\ 2 & \text{if } n = 3, \\ P(n-1) + \sum_{k=1}^{n-2} \left(\sum_{t=0}^{min(k-1,n-k-1)} P(n-2t-2) \right) & \text{otherwise.} \end{cases}$$

Recurrence Formula[Stadler97]

$$P(n) = P(n-1) + (n-1)P(n-2) - P(n-3) + P(n-4)$$

$$\forall n \geq 4$$

Enumerating RNA structures

The number of different RNA structures for sequences of length n $(1 \leq n \leq 20)$

n	$P(n)$	n	$P(n)$
1	1	11	16526
2	1	12	64351
3	2	13	259471
4	5	14	1083935
5	13	15	4668704
6	37	16	20732609
7	112	17	94607409
8	363	18	443476993
9	1235	19	2130346450
10	4427	20	10482534517

Some properties and results

Some properties of $P(n)$:

- $P(n) > 0$ for $n \geq 0$

Some properties and results

Some properties of $P(n)$:

- $P(n) > 0$ for $n \geq 0$
- $P(n) > P(n-1)$ for $n \geq 3$

Some properties and results

Some properties of $P(n)$:

- $P(n) > 0$ for $n \geq 0$
- $P(n) > P(n-1)$ for $n \geq 3$
- $P(n) > \sum_{i=0}^{n-1} P(i)$ for $n \geq 4$

One can prove all these properties by induction.

Some properties and results

- Involution
  An involution on a set S is a permutation $\pi : S \mapsto S$, such that $\forall s \in S, \pi^2(s) = s$.

Some properties and results

- **Involution**
  An involution on a set S is a permutation $\pi : S \mapsto S$, such that $\forall s \in S, \pi^2(s) = s$.
- **Number of involutions**
  The number of involutions on a set S of size n is given by the recursion formula $Q(n) = Q(n-1) + (n-1)Q(n-2)$.

Some properties and results

- Involution
  An involution on a set S is a permutation $\pi : S \mapsto S$, such that $\forall s \in S, \pi^2(s) = s$.
- Number of involutions
  The number of involutions on a set S of size n is given by the recursion formula $Q(n) = Q(n-1) + (n-1)Q(n-2)$.
- Asymptotic behaviour
  $Q(n) \backsim \frac{1}{\sqrt{2}} n^{n/2} e^{(-\frac{n}{2}+\sqrt{n}-\frac{1}{4})}$. [Chowla52]

Some properties and results

- Involution
  An involution on a set S is a permutation $\pi : S \mapsto S$, such that $\forall s \in S, \pi^2(s) = s$.
- Number of involutions
  The number of involutions on a set S of size n is given by the recursion formula $Q(n) = Q(n-1) + (n-1)Q(n-2)$.
- Asymptotic behaviour
  $Q(n) \backsim \frac{1}{\sqrt{2}} n^{n/2} e^{(-\frac{n}{2}+\sqrt{n}-\frac{1}{4})}$. [Chowla52]
- Observation
  $\forall n \geq 1$ we have $Q(n) \geq P(n)$, which means $\frac{Q(n)}{P(n)} \geq 1$.

Some properties and results

Lemma

For $n \geq 10$ *we have* $P(n) \geq 2n^{1/4}P(n-1)$.

Some properties and results

Lemma

For $n \geq 10$ *we have* $P(n) \geq 2n^{1/4}P(n-1)$.

Proof. The proof is by induction.

$$P(n) \geq P(n-1) + (n-1)P(n-2) - P(n-3) \text{ and}$$
$$2n^{1/4}P(n-2) + 2n^{1/4}(n-2)P(n-3) \geq 2n^{1/4}P(n-1)$$

Some properties and results

Lemma

For $n \geq 10$ *we have* $P(n) \geq 2n^{1/4}P(n-1)$*.*

Proof. The proof is by induction.

$$P(n) \geq P(n-1) + (n-1)P(n-2) - P(n-3) \text{ and}$$
$$2n^{1/4}P(n-2) + 2n^{1/4}(n-2)P(n-3) \geq 2n^{1/4}P(n-1)$$

$$P(n-1) + (n-1)P(n-2) - P(n-3) \geq 2n^{1/4}P(n-2) + 2n^{1/4}(n-2)P(n-3)$$

Some properties and results

Lemma

For $n \geq 10$ *we have* $P(n) \geq 2n^{1/4}P(n-1)$.

Proof. The proof is by induction.

$$P(n) \geq P(n-1) + (n-1)P(n-2) - P(n-3) \text{ and}$$
$$2n^{1/4}P(n-2) + 2n^{1/4}(n-2)P(n-3) \geq 2n^{1/4}P(n-1)$$

$$P(n-1) + (n-1)P(n-2) - P(n-3) \geq 2n^{1/4}P(n-2) + 2n^{1/4}(n-2)P(n-3)$$

$$P(n-2) \geq 2\left[n^{1/4} - (n-1)^{1/4}\right]P(n-2) + 2\left[(n-2)n^{1/4} - (n-2)(n-2)^{1/4}\right]P(n-3) + P(n-3)$$

Some properties and results

Lemma

For $n \geq 10$ we have $P(n) \geq 2n^{1/4}P(n-1)$.

Proof. The proof is by induction.

$$P(n) \geq P(n-1) + (n-1)P(n-2) - P(n-3) \text{ and}$$
$$2n^{1/4}P(n-2) + 2n^{1/4}(n-2)P(n-3) \geq 2n^{1/4}P(n-1)$$

$$P(n-1) + (n-1)P(n-2) - P(n-3) \geq 2n^{1/4}P(n-2) + 2n^{1/4}(n-2)P(n-3)$$

$$P(n-2) \geq 2\left[n^{1/4} - (n-1)^{1/4}\right]P(n-2) + 2\left[(n-2)n^{1/4} - (n-2)(n-2)^{1/4}\right]P(n-3) + P(n-3)$$

$$\left[1 - \frac{1}{2(n-1)^{3/4}}\right]P(n-2) \geq (n-2)^{1/4}P(n-3) + P(n-3)$$

Some properties and results

Lemma

For $n \geq 10$ we have $P(n) \geq 2n^{1/4}P(n-1)$.

Proof. The proof is by induction.

$$P(n) \geq P(n-1) + (n-1)P(n-2) - P(n-3) \text{ and}$$
$$2n^{1/4}P(n-2) + 2n^{1/4}(n-2)P(n-3) \geq 2n^{1/4}P(n-1)$$

$$P(n-1) + (n-1)P(n-2) - P(n-3) \geq 2n^{1/4}P(n-2) + 2n^{1/4}(n-2)P(n-3)$$

$$P(n-2) \geq 2\left[n^{1/4} - (n-1)^{1/4}\right]P(n-2) + 2\left[(n-2)n^{1/4} - (n-2)(n-2)^{1/4}\right]P(n-3) + P(n-3)$$

$$\left[1 - \frac{1}{2(n-1)^{3/4}}\right]P(n-2) \geq (n-2)^{1/4}P(n-3) + P(n-3)$$

$$\left[1 - \frac{1}{(n-1)^{3/4}}\right](n-2)^{1/4} \geq 1$$

□

Some properties and results

Lemma

There exists a constant number $M > 1$ *and an integer* N*, such that for* $n \geq N$, $\frac{Q(n)}{P(n)} \leq M$.

Some properties and results

Lemma

There exists a constant number $M > 1$ and an integer N, such that for $n \geq N$, $\frac{Q(n)}{P(n)} \leq M$.

Proof.

Supp. $R(n) = \frac{Q(n)}{P(n)} \Rightarrow R(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} + Q(n)\left(\frac{P(n-3)-P(n-4)}{P(n)(P(n-1)+(n-1)P(n-2))}\right)$

Some properties and results

Lemma

There exists a constant number $M > 1$ and an integer N, such that for $n \geq N$, $\frac{Q(n)}{P(n)} \leq M$.

Proof.

$$\text{Supp. } R(n) = \frac{Q(n)}{P(n)} \Rightarrow R(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} + Q(n)\left(\frac{P(n-3)-P(n-4)}{P(n)(P(n-1)+(n-1)P(n-2))}\right)$$

$$\text{Supp. } S(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} \Rightarrow S(n) \leq max\{R(n-1), R(n-2)\}.$$

Some properties and results

Lemma

There exists a constant number $M > 1$ and an integer N, such that for $n \geq N$, $\frac{Q(n)}{P(n)} \leq M$.

Proof.

$$\text{Supp. } R(n) = \frac{Q(n)}{P(n)} \Rightarrow R(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} + Q(n)\left(\frac{P(n-3)-P(n-4)}{P(n)(P(n-1)+(n-1)P(n-2))}\right)$$

$$\text{Supp. } S(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} \Rightarrow S(n) \leq max\{R(n-1), R(n-2)\}.$$

$$R(n) \leq max\{R(n-1), R(n-2)\}\left(1 - \frac{1}{(n-1)(n-2)^{1/4}}\right)^{-1}$$

Some properties and results

Lemma

There exists a constant number $M > 1$ and an integer N, such that for $n \geq N$, $\frac{Q(n)}{P(n)} \leq M$.

Proof.

$$\text{Supp. } R(n) = \frac{Q(n)}{P(n)} \Rightarrow R(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} + Q(n)\left(\frac{P(n-3)-P(n-4)}{P(n)(P(n-1)+(n-1)P(n-2))}\right)$$

$$\text{Supp. } S(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} \Rightarrow S(n) \leq max\{R(n-1), R(n-2)\}.$$

$$R(n) \leq max\{R(n-1), R(n-2)\}\left(1 - \frac{1}{(n-1)(n-2)^{1/4}}\right)^{-1}$$

$$\left(1 - \frac{1}{(i-1)(i-2)^{1/4}}\right)^{-1} \leq \left(1 + \frac{2}{(i-1)(i-2)^{1/4}}\right)$$

Some properties and results

Lemma

There exists a constant number $M > 1$ and an integer N, such that for $n \geq N$, $\frac{Q(n)}{P(n)} \leq M$.

Proof.

$$\text{Supp. } R(n) = \frac{Q(n)}{P(n)} \Rightarrow R(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} + Q(n)\left(\frac{P(n-3)-P(n-4)}{P(n)(P(n-1)+(n-1)P(n-2))}\right)$$

$$\text{Supp. } S(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} \Rightarrow S(n) \leq max\{R(n-1), R(n-2)\}.$$

$$R(n) \leq max\{R(n-1), R(n-2)\}\left(1 - \frac{1}{(n-1)(n-2)^{1/4}}\right)^{-1}$$

$$\left(1 - \frac{1}{(i-1)(i-2)^{1/4}}\right)^{-1} \leq \left(1 + \frac{2}{(i-1)(i-2)^{1/4}}\right)$$

$$\prod_{i=N}^{\infty}\left(1 - \frac{1}{(i-1)(i-2)^{1/4}}\right)^{-1} \leq \prod_{i=N}^{\infty}\left(1 + \frac{2}{(i-1)(i-2)^{1/4}}\right) \rightsquigarrow M_0$$

Some properties and results

Lemma

There exists a constant number $M > 1$ and an integer N, such that for $n \geq N$, $\frac{Q(n)}{P(n)} \leq M$.

Proof.

$$\text{Supp. } R(n) = \frac{Q(n)}{P(n)} \Rightarrow R(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} + Q(n)\left(\frac{P(n-3)-P(n-4)}{P(n)(P(n-1)+(n-1)P(n-2))}\right)$$

$$\text{Supp. } S(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} \Rightarrow S(n) \leq max\{R(n-1), R(n-2)\}.$$

$$R(n) \leq max\{R(n-1), R(n-2)\}\left(1 - \frac{1}{(n-1)(n-2)^{1/4}}\right)^{-1}$$

$$\left(1 - \frac{1}{(i-1)(i-2)^{1/4}}\right)^{-1} \leq \left(1 + \frac{2}{(i-1)(i-2)^{1/4}}\right)$$

$$\prod_{i=N}^{\infty}\left(1 - \frac{1}{(i-1)(i-2)^{1/4}}\right)^{-1} \leq \prod_{i=N}^{\infty}\left(1 + \frac{2}{(i-1)(i-2)^{1/4}}\right) \rightsquigarrow M_0$$

$$R(n) \leq M_0 \times max\{R(N-1), R(N-2)\}$$

Some properties and results

Lemma

There exists a constant number $M > 1$ and an integer N, such that for $n \geq N$, $\frac{Q(n)}{P(n)} \leq M$.

Proof.

$$\text{Supp. } R(n) = \frac{Q(n)}{P(n)} \Rightarrow R(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} + Q(n)\left(\frac{P(n-3)-P(n-4)}{P(n)(P(n-1)+(n-1)P(n-2))}\right)$$

$$\text{Supp. } S(n) = \frac{Q(n-1)+(n-1)Q(n-2)}{P(n-1)+(n-1)P(n-2)} \Rightarrow S(n) \leq max\{R(n-1), R(n-2)\}.$$

$$R(n) \leq max\{R(n-1), R(n-2)\}\left(1 - \frac{1}{(n-1)(n-2)^{1/4}}\right)^{-1}$$

$$\left(1 - \frac{1}{(i-1)(i-2)^{1/4}}\right)^{-1} \leq \left(1 + \frac{2}{(i-1)(i-2)^{1/4}}\right)$$

$$\prod_{i=N}^{\infty}\left(1 - \frac{1}{(i-1)(i-2)^{1/4}}\right)^{-1} \leq \prod_{i=N}^{\infty}\left(1 + \frac{2}{(i-1)(i-2)^{1/4}}\right) \rightsquigarrow M_0$$

$$R(n) \leq M_0 \times max\{R(N-1), R(N-2)\}$$

Let $M = M_0 \times max\{R(N-1), R(N-2)\}$.

□

Some properties and results

Theorem

$P(n) = \Theta(Q(n))$

Proof.

Using the fact $\frac{1}{M} \leq \frac{P(n)}{Q(n)} \leq 1$, and the definition of Θ.

$\square$

Some properties and results

Theorem

$P(n) = \Theta(Q(n))$

Proof.

Using the fact $\frac{1}{M} \leq \frac{P(n)}{Q(n)} \leq 1$, and the definition of Θ.

□

Asymptotic behaviour

$$P(n) \sim \frac{1}{\sqrt{2}} n^{n/2} e^{\left(-\frac{n}{2}+\sqrt{n}-\frac{1}{4}\right)}$$

Conclusions and Prespectives

- Some basic definitions
- Enumerate the number of RNA structures
- Some properties for the recurrence $P(n)$
- Asymtotic behaviour for $P(n)$

Conclusions and Prespectives

- Some basic definitions
- Enumerate the number of RNA structures
- Some properties for the recurrence $P(n)$
- Asymtotic behaviour for $P(n)$

Interesting problems:

- Generalize the method: Differential Equation for the exponential generating function, holonomic function; solve via Laplace transform; asymptotics from Cauchy formula with saddlepoint integration
- Enumerating the number of Bi-Secondary structures
- Applying some restrictions to the structure such as minimum length for hairpin loops and stems

References

- M. Waterman, *Secondary structure of single - stranded nucleic acids*, Academic Press N.Y., **1**, 167 – 212, 1978.
- I. L. Hofacker, P. Schuster, and P. F. Stadler, *Combinatorics of RNA Secondary Structures*, Discr. Appl. Math., **88**, 207–237, 1998.
- P. Stadler and C. Haslinger, *RNA structures with pseudo-knots: Graph theoretical and combinatorial properties*, Bull. Math. Biol., Preprint 97-03-030, 1997.
- M. Nebel, *Combinatorial Properties of RNA secondary Structures*,2001.
- M. Régnier, *Generating Functions in Computational Biology*, Inria, March 3, 1997.
- *http://www.research.att.com/∼njas/sequences/A000085*.

End.