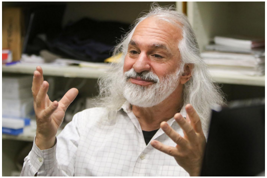
Distribution of endhered patterns in RNA-related secondary structures

Célia Biane[1], Greg Hampikian[2], Sergey Kirgizov[3],
Khaydar Nurligareev[4], Daniel Pinson[3]

[1]LaBRI (Bordeaux), [2]CompGenomics (USA), [3]LIB (Dijon), [4]LIP6 (Paris)

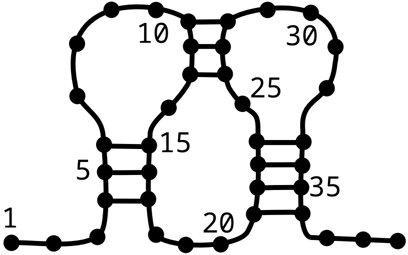
SeqBIM
2024 – 28-29 nov.
Rennes

No occurrences of these patterns in the human genome	CGCTCGACGTA, GTCCGAGCGTA, CGACGAACGGT, CCGATACGTCG

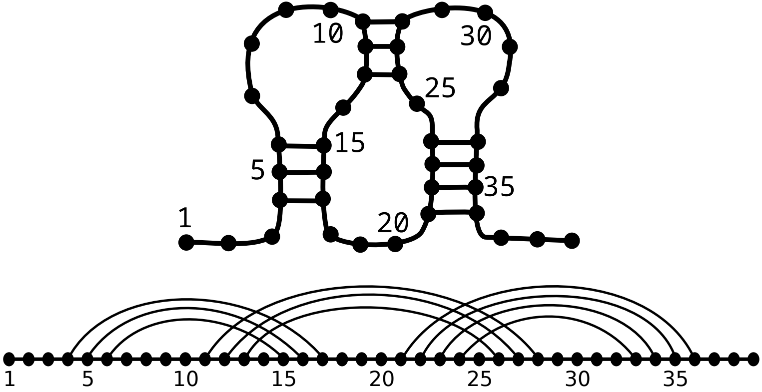
Absent sequences: nullomers and primes, 2007
by Greg Hampikian and Tim Andersen

Explore absense and presence of patterns in RNA secondary structures

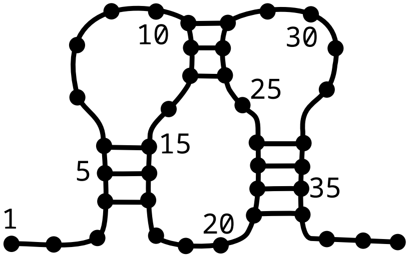
RNA secondary structures and matchings

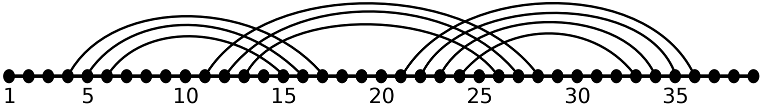


RNA secondary structures and matchings



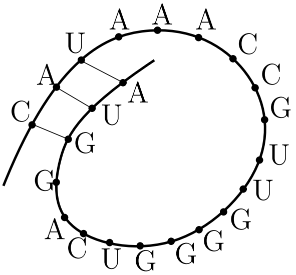
RNA secondary structures and matchings

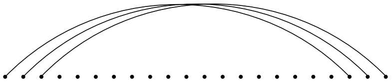




...(((....[[[.)))...((((.]]]....))))...

Snail-like pseudoknotted RNA

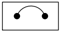




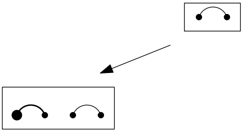
```
([{...............)]}
CAUAAACCGUUGGGUCAGGUA
```

Does it exist in nature ?

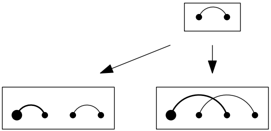
Perfect matchings



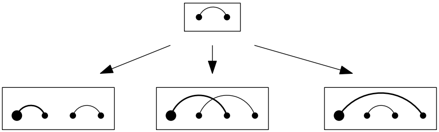
Perfect matchings



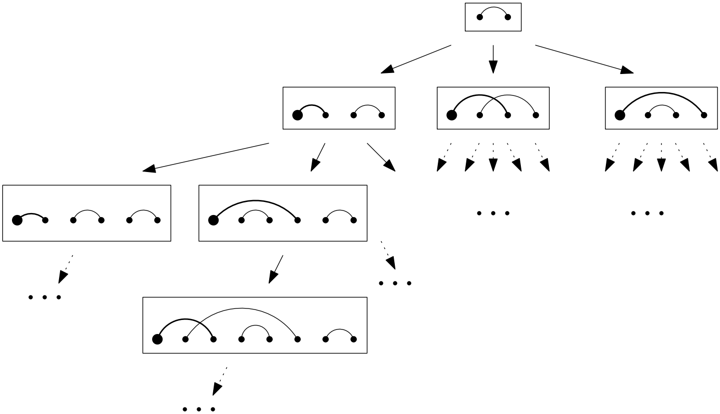
Perfect matchings



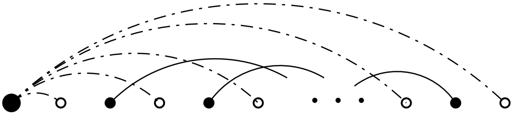
Perfect matchings



Perfect matchings



Perfect matchings



$2(n-1)$ old points

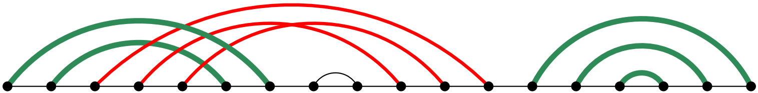
$2(n-1)+1$ possible new arcs endings

$a_n = (2n-1)\cdot(2n-3)\cdots 5\cdot 3\cdot 1 = (2n-1)!!$

A1147 in Sloane's Encyclopedia : 1, 3, 15, 105, 945, 10395,...

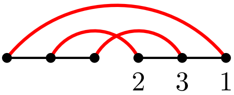
Endhered patterns in perfect matchings

(endhered = **end**-ad**hered**)

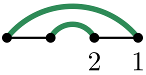


This matching contains

1 occurrence of pattern



3 occurrences of pattern



We write patterns in condensed form, indicating sequentially the order of starting points corresponding to arc ends.

(En français.... les motifs collex = **coll**és par leurs **ex**trémités)

What's the distribution of endhered patterns in perfect matchings?

Let's start with  and 

Distribution of  in perfect matchings

k \ n	1	2	3	4	5	6	7	8	9	OEIS
0	1	2	10	68	604	6584	85048	1269680	21505552	A165968
1	0	1	4	30	272	3020	39504	595336	10157440	A179540
2	0	0	1	6	60	680	9060	138264	2381344	
3	0	0	0	1	8	100	1360	21140	368704	
4	0	0	0	0	1	10	150	2380	42280	
5	0	0	0	0	0	1	12	210	3808	
6	0	0	0	0	0	0	1	14	280	
7	0	0	0	0	0	0	0	1	16	
8	0	0	0	0	0	0	0	0	1	

Let $a_{n,k}$ be the number of matchings with n arcs and k occurrences of pattern . Exponential generating function is

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n} a_{n+1,k}\frac{z^n}{n!}u^k = \frac{e^{z(u-1)}}{\sqrt{(1-2z)^3}}$$

Distribution of  in perfect matchings

k \ n	1	2	3	4	5	6	7	8	9	OEIS
0	1	2	10	68	604	6584	85048	1269680	21505552	A165968
1	0	1	4	30	272	3020	39504	595336	10157440	A179540
2	0	0	1	6	60	680	9060	138264	2381344	
3	0	0	0	1	8	100	1360	21140	368704	
4	0	0	0	0	1	10	150	2380	42280	
5	0	0	0	0	0	1	12	210	3808	
6	0	0	0	0	0	0	1	14	280	
7	0	0	0	0	0	0	0	1	16	
8	0	0	0	0	0	0	0	0	1	

Let $a_{n,k}$ be the number of matchings with n arcs and k occurrences of pattern . Asymptotics

$$a_{n,k} \sim \frac{1}{2^k k!} \left(\frac{2}{e}\right)^{n+1/2} n^n \qquad \frac{a_{n,k}}{a_{n,k+1}} \sim 2(k+1).$$

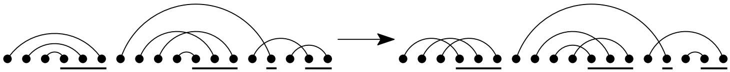
Does the distribution of  differ from the distribution of ?

Endhered twist

Left endhered twist
All runs of consecutive **starting points** are reversed.

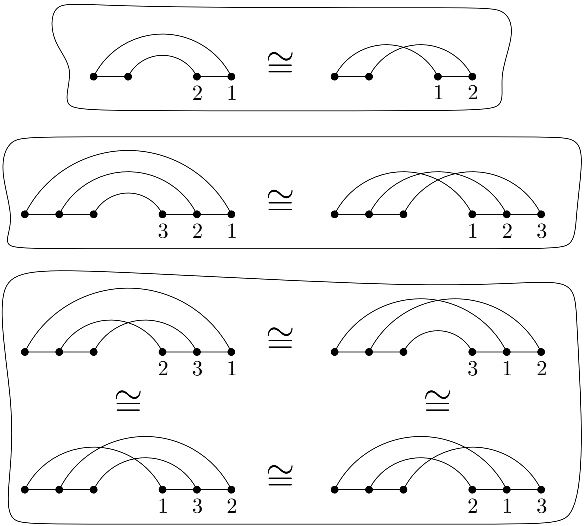
Right endhered twist
All runs of consecutive **ending points** are reversed.

Example of right twist:



Thanks to an endhered twist
 **and**  **have the same distribution!**

Endhered pattern equivalence



equivalence = same distribution

Goulden-Jackson cluster method and endhered patterns

Endhered pattern enumeration in matchings

Imagine we have g.f. for a distribution of a given pattern μ:

$$D_\mu(z,u) = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} d_{n,k} z^n u^k.$$

There are $d_{n,k}$ matchings of size n with k occurrences of μ.

Endhered pattern enumeration in matchings

Imagine we have g.f. for a distribution of a given pattern μ:

$$D_\mu(z,u) = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} d_{n,k} z^n u^k.$$

There are $d_{n,k}$ matchings of size n with k occurrences of μ.

We label some occurrences by variable v,
i.e. u is replaced either by 1 or by v.

$$H_\mu(z,v) = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} h_{n,k} z^n v^k = D_\mu(z, 1+v).$$

It is simplier to construct H_μ than D_μ !

Endhered pattern enumeration in matchings

Imagine we have g.f. for a distribution of a given pattern μ:

$$D_\mu(z,u) = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} d_{n,k} z^n u^k.$$

There are $d_{n,k}$ matchings of size n with k occurrences of μ.

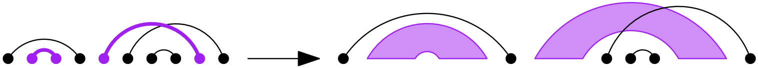
We label some occurrences by variable v,
i.e. u is replaced either by 1 or by v.

$$H_\mu(z,v) = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} h_{n,k} z^n v^k = D_\mu(z,1+v).$$

It is simplier to construct H_μ than D_μ !
Then we recover by the symbolic inclusion-exclusion:
$D_\mu(z,u) = H_\mu(z,u-1)$

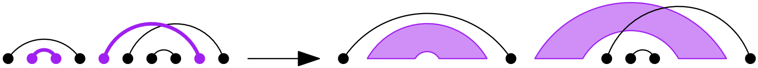
Idea: replace some arcs by patterns

(simple case without self-overlappings)



Idea: replace some arcs by patterns

(simple case without self-overlappings)



We label certain arcs by v, these arcs will be replaced by occurrences of pattern μ.

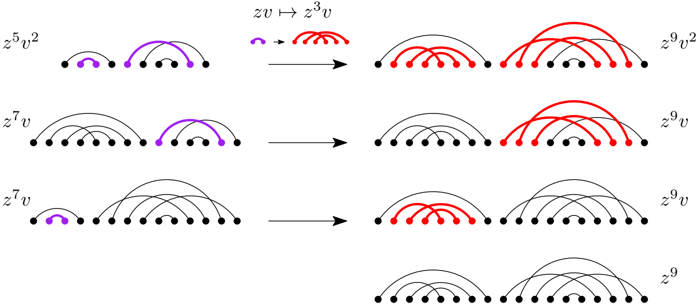
$$F(z + zv),$$

where $F(z)$ is the ordinary g.f. for all matchings.

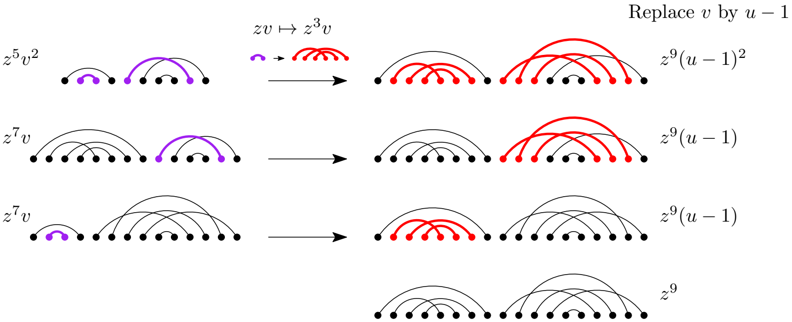
$$F(z) = \sum_{n=0}^{\infty} (2n-1)!!\, z^n = 1 + z + 3z^2 + 15z^3 + 105z^4 + \dots$$

Consider, for instance, the endhered pattern

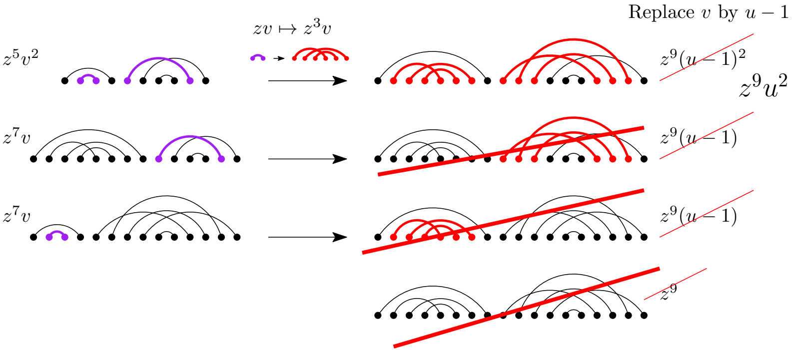
Zero self-overlappings



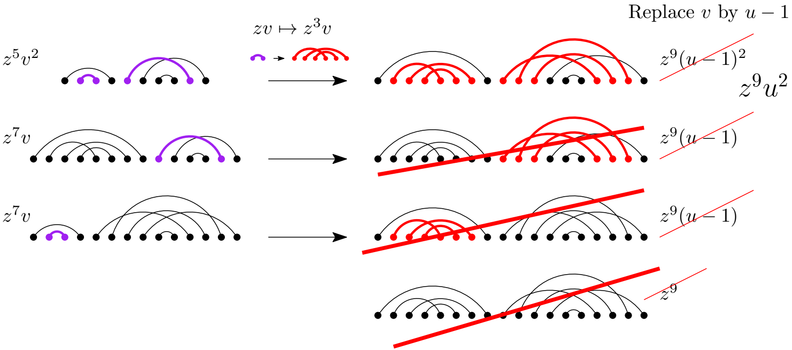
Zero self-overlappings



Zero self-overlappings



Zero self-overlappings



$$H_\mu(z, v) = F(z + z^\ell v),$$

$$D_\mu(z, u) = H_\mu(z, u-1),$$

where ℓ is the size of pattern μ.

WITH SELF-OVERLAPPINGS ?

Autocorrelation encodes self-overlappings

An *autocorrelation polynomial* $A(\pi; z)$ for an endhered pattern π of size n is

$$A(\pi; z) = 1 + \sum_{k \in S} z^{n-k},$$

where S is the set of lengths of possible overlappings of two different occurrences of the pattern π in some matching.
In other words, z^{n-k} means that two occurrences have k edges in common.

$A(21; z) = A($  $; z) = 1 + z,$

$A(12; z) = A($  $; z) = 1 + z,$

$A(132; z) = A($  $; z) = 1,$

$A(321; z) = A($  $; z) = 1 + z + z^2,$

$A(3412; z) = A($  $; z) = 1 + z^2,$

$$A(7564231; z) = 1 + z^3 + z^6.$$

Enumeration and asymptotics

Let π be an endhered pattern of size ℓ, with autocorrelation $A(\pi; z) = 1 + z^m + \dots$ (m is the smallest positive power)
If $A(\pi; z) = 1$, then we let $m = 0$.
Generating function:

$$\sum_{n,k \geq 0} a_{n,k}\, z^n u^k = F\left(z + \frac{(u-1)z^\ell}{1 - (u-1)(A(\pi; z) - 1)}\right)$$

Enumeration and asymptotics

Let π be an endhered pattern of size ℓ, with autocorrelation $A(\pi; z) = 1 + z^m + \ldots$ (m is the smallest positive power)
If $A(\pi; z) = 1$, then we let $m = 0$.
Generating function:

$$\sum_{n,k \geq 0} a_{n,k}\, z^n u^k = F\left(z + \frac{(u-1)z^\ell}{1 - (u-1)(A(\pi; z) - 1)}\right)$$

Asymptotics by Borinsky's approach: as $n \to \infty$,

$$\frac{a_{n,k}}{(2n-1)!!} \sim \begin{cases} \dfrac{1}{k!\, 2^{k(\ell-1)}} \cdot \dfrac{1}{n^{k(\ell-2)}} & \text{if} \quad m = \ell - 1 \text{ or } m = 0 \\ \dfrac{1}{(2n)^{k(\ell-2)}} \displaystyle\sum_{s=1}^{k} \frac{1}{s!\, 2^s} \binom{k-1}{s-1} & \text{if} \quad m = \ell - 2 \\ \dfrac{1}{2(2n)^{km+(\ell-2-m)}} & \text{if} \quad 0 < m < \ell - 2 \end{cases}$$

Well... what about real-world data ?

Real-world data

Data comes from PDB, we have used X3DNA-DSSR to obtain dot-bracket notations from 3D coordinates of atoms.
FR3D Python can also be used.

Our database looks like this:

```
2GQ5 ((((((((&))))))))&((((((((&))))))))
2IRN (((..(((.&)))..))).
2IZ8 (((.((....))))).&((.((....))))
3VAL ......&......&......&.....
437D ..(((((..[[[.)))))).......]]]
4M40 ((((((.(((((.((.......((((.(.[.(..])..).)))))).))))).))))))
6IV6 ....(((((.....))))).......
7K16 {{...((((((((....))))))((((...[[[...))))}}.))...]]]
```

Interactive web application by Daniel Pinson

https://rna.kirgizov.link

Possible research directions

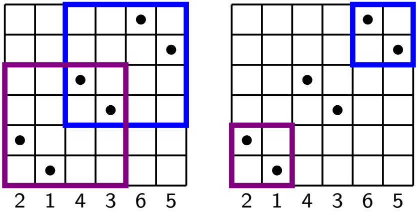
- Explain possible autocorrelations or endhered patterns, (in other words, period sets)

  1, 2, 4, 4, 7, 7, 11, 12, 18, ... ? (shift of A304178 ?)
  Bijection with sets of palindrome prefix lengths, over all binary palindromes of length n ???

- Characterise real-world RNA secondary structures by pattern distributions (avoidance-presence)

Endhered patterns in matchings and RNA
Célia Biane, Greg Hampikian, Sk, Khaydar Nurligareev
https://arxiv.org/abs/2404.18802
To appear in *Journal of Computational Biology*

Asymptotics of self-overlapping permutations
Sk and Khaydar Nurligareev
https://arxiv.org/abs/2311.11677
To appear in *Discrete Mathematics*



Interactive web application by Daniel Pinson *et al.*
https://rna.kirgizov.link

Clusters of endhered patterns in permutations and matchings. In preparation.

Merci !