The size of the biggest Caterpillar subtree in binary rooted ordered trees

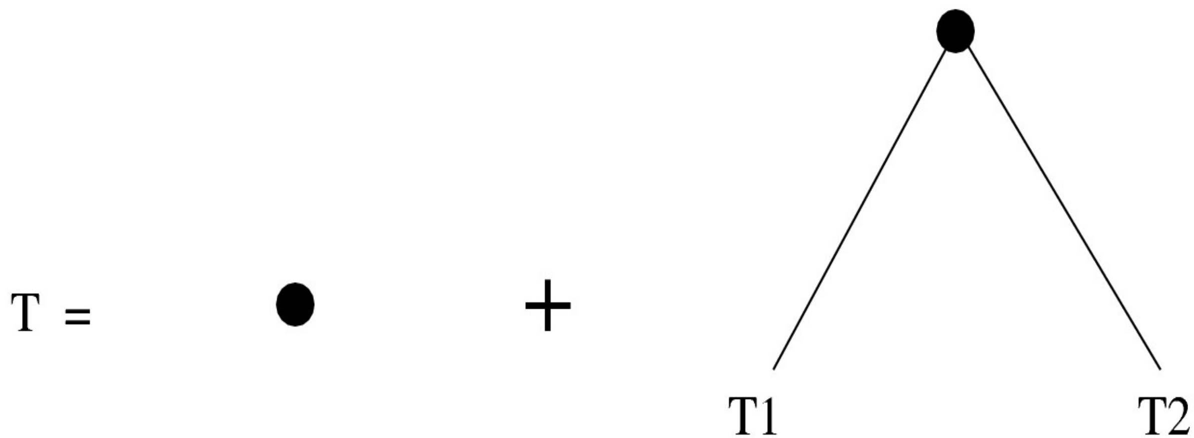
Filippo Disanto

fdisanto@uni-koeln.de

GASCom 2012, Bordeaux

Introduction

Binary rooted ordered trees (#leaves) → Catalan numbers



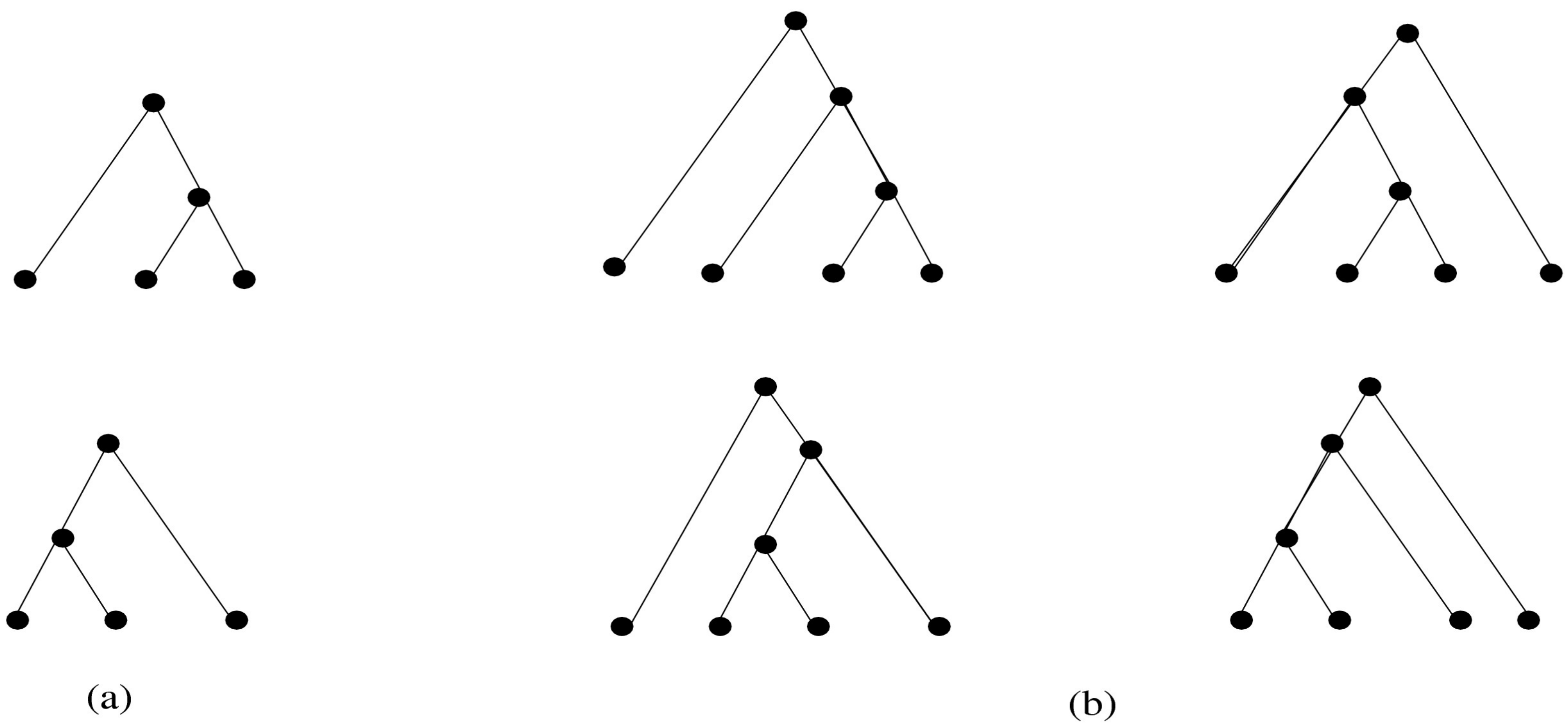
$$C(x) = x + C(x)^2$$

$$C(x) = \frac{1-\sqrt{1-4x}}{2} = x + x^2 + 2x^3 + 5x^4 + 14x^5 + ..$$

Caterpillar trees

A tree is a *caterpillar* if :

each node is either a leaf or has at least one leaf as a child



(a) caterpillars of size 3; (b) caterpillars of size 4

Caterpillar trees & viruses trees

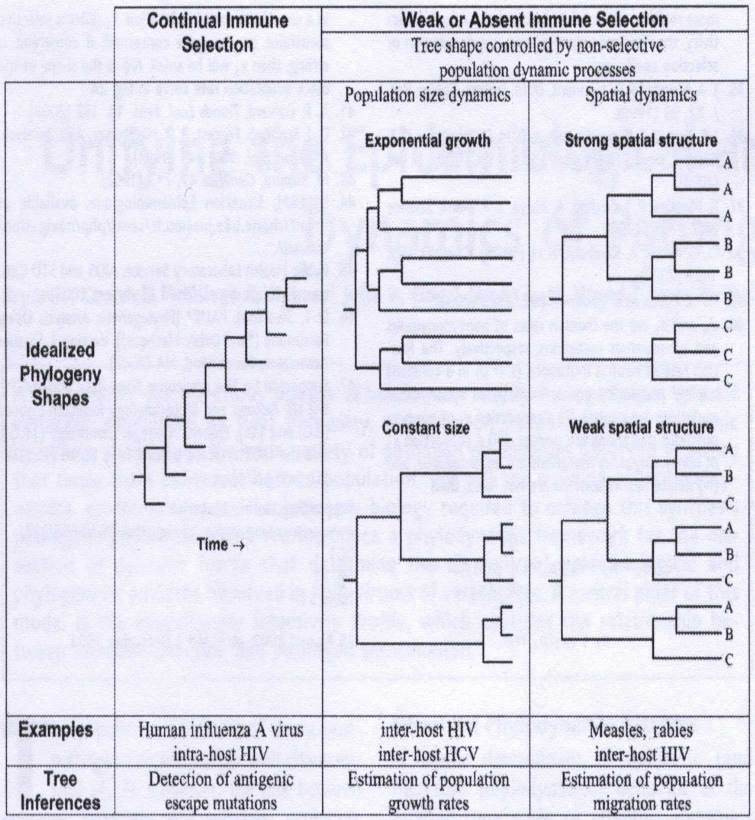


Fig. 3. Idealized tree shapes under different phylodynamic processes. The main division is between those viruses subject to continual immune-driven selection (such as human influenza A virus and intra-host HIV), in which trees have a strong temporal structure, and viruses where immune selection is absent or weak (such as many RNA viruses), in which the trees depict population size and spatial dynamics. The types of evolutionary inference that can be made from the various phylogenies are also indicated. (A, B, and C represent three subpopulations from which viruses have been sampled.)

Bedford *et al. BMC Evolutionary Biology* 2011, **11**:220
http://www.biomedcentral.com/1471-2148/11/220

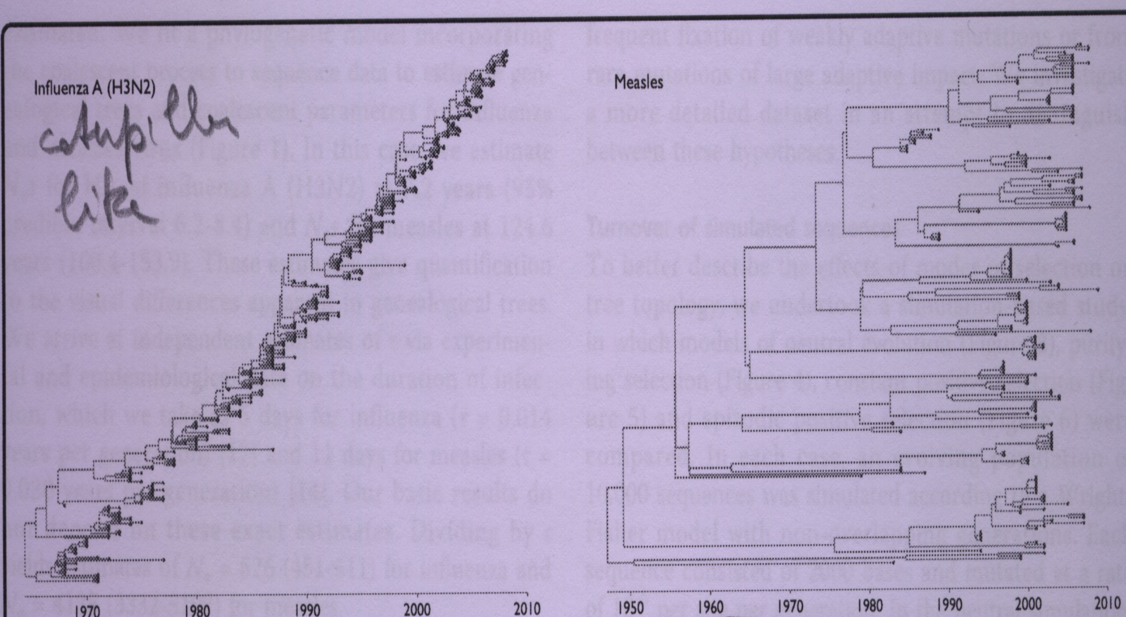
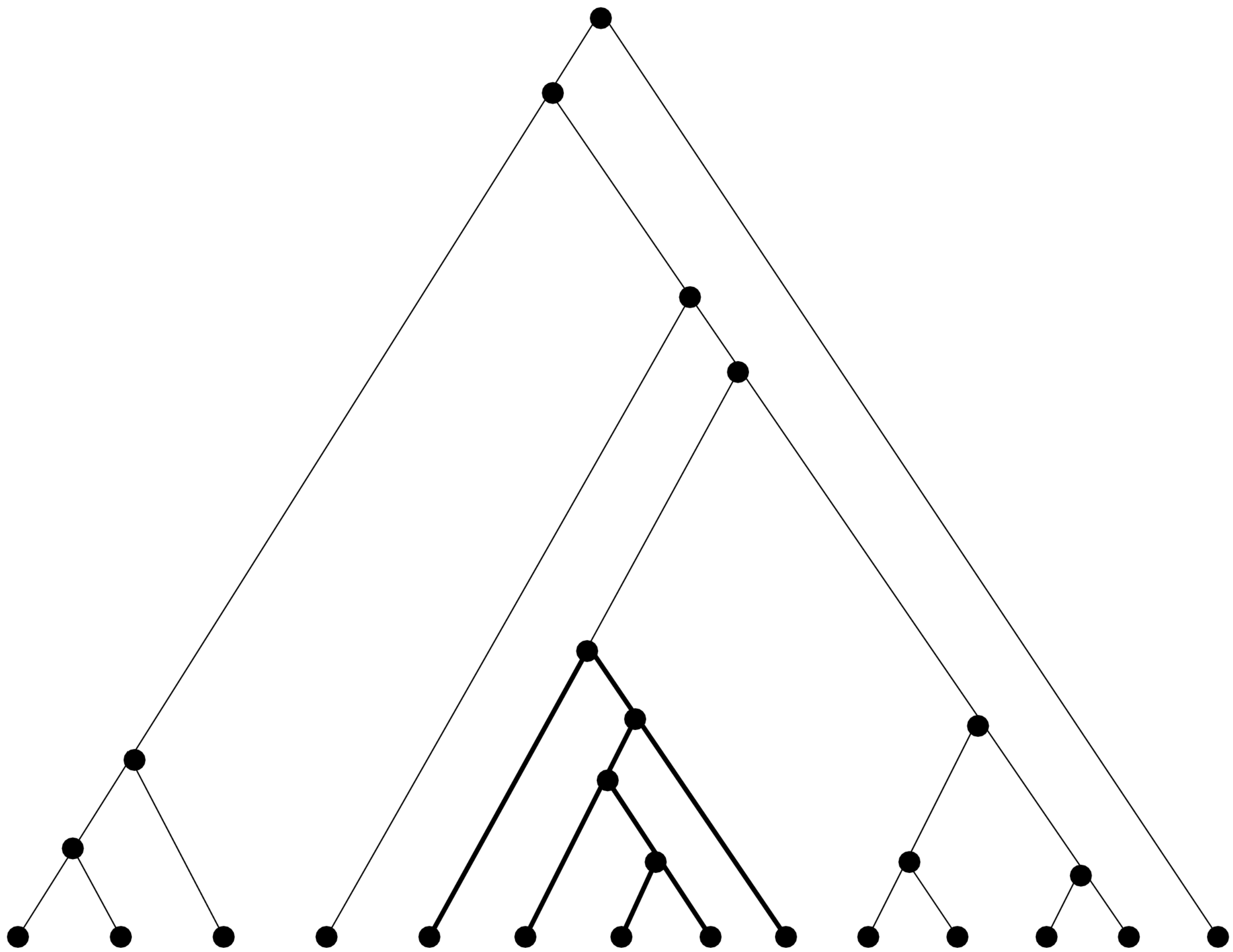


Figure 1 Genealogical trees derived from hemagglutinin sequences of influenza A (H3N2) and measles virus. The influenza tree, sampled between 1968 and 2008, appears long and spindly with a distinct lack of deep branches. It only takes a few years for contemporaneous strains to find a common ancestor. The measles tree, sampled between 1979 and 2009, looks very different, harboring many deep branches. It takes approximately 50 years for contemporaneous measles strains to find a common ancestor.

The size of the biggest caterpillar subtree (γ)



$\gamma = 5$

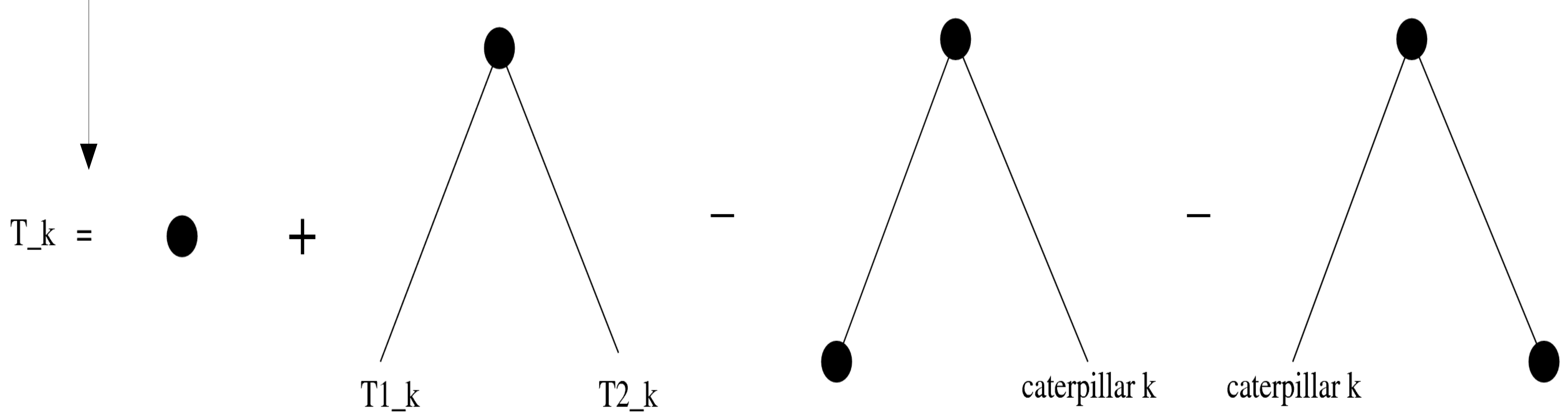
Counting trees wrt size and γ by inclusion-exclusion

$$F_k \text{ is the g.f. for trees with } \gamma \text{ at most } k,$$

$$F_k = x + F_k^2 - 2^{k-1}x^{k+1},$$

$$F_k(x) = \frac{1 - \sqrt{1 - 4x + 2^{k+1}x^{k+1}}}{2},$$

$$F_k - F_{k-1} \text{ counts trees with } \gamma = k.$$



Asymptotic growth

Let ρ_k be the positive dominant singularity of F_k,

$$\rho_k = \frac{1}{4}(1 + 2^{k+1}\rho_k^{k+1}),$$

bootstrapping

$$\rho_k = \frac{1}{4} + \frac{1}{2^{k+3}} + O\left(k\left(\frac{2}{5}\right)^k\right).$$

In particular $\rho_2 = \frac{1}{4}(\sqrt{5} - 1) = 0.3090169...$

Trees with no pitchforks $\to \frac{1}{4}\sqrt{\frac{4\rho_2 - 24\rho_2^3}{\pi n^3}}\left(\frac{1}{\rho_2}\right)^n.$

The average size of the biggest caterpillar subtree E_n(γ)

Let $f_k^{(n)} = [x^n]F_k(x)$ and $C^{(n)} = [x^n]C(x)$

$$
\begin{aligned}
E_n(\gamma) &= \frac{1f_1^{(n)} + \sum_{k\geq 1}(k+1)(f_{k+1}^{(n)} - f_k^{(n)})}{C^{(n)}} \\
&= \frac{-f_1^{(n)} - ... - f_{n-1}^{(n)} + nf_n^{(n)} + \sum_{k\geq n}(k+1)(f_{k+1}^{(n)} - f_k^{(n)})}{C^{(n)}} \\
&= \frac{-f_1^{(n)} - ... - f_{n-1}^{(n)} + nC^{(n)} + \sum_{k\geq n}(C^{(n)} - f_k^{(n)})}{C^{(n)}} \\
&= \frac{\sum_{k=1}^{n-1}(C^{(n)} - f_k^{(n)}) + C^{(n)} + \sum_{k\geq n}(C^{(n)} - f_k^{(n)})}{C^{(n)}} \\
&= \frac{C^{(n)} + \sum_{k\geq 1}(C^{(n)} - f_k^{(n)})}{C^{(n)}} \\
&= 1 + \frac{\sum_{k\geq 1}(C^{(n)} - f_k^{(n)})}{C^{(n)}}
\end{aligned}
$$

The average size of the biggest caterpillar subtree E_n(γ)

$$U(x) = \sum_{k\geq 1}(C(x) - F_k(x)) = \frac{\sqrt{1-4x}}{2}\sum_{k\geq 1}\left(\sqrt{1+\frac{2^{k+1}x^{k+1}}{1-4x}} - 1\right).$$

We want a function $\tilde{U}$ which estimates U near the dominant singularity $1/4$.

Let us fix x near $1/4$ and let us consider the threshold function

$$k_0 = \log_2 \frac{1}{|1-4x|}.$$

Then, supposing $k \geq k_0$, we have that

$$\sqrt{1+\frac{2^{k+1}x^{k+1}}{1-4x}} \sim \sqrt{1+\frac{1}{2^{k+1}(1-4x)}} \sim 1 + \frac{1}{2^{k+2}(1-4x)},$$

while if we suppose $k < k_0$ we use the approximation

$$\sqrt{1+\frac{2^{k+1}x^{k+1}}{1-4x}} \sim \sqrt{1+\frac{1}{2^{k+1}(1-4x)}} \sim \sqrt{\frac{1}{2^{k+1}(1-4x)}}.$$

For the fixed x near $1/4$ we estimate $U(x)$ as

$$\tilde{U}(x) = \frac{1}{2} + \frac{1}{2\sqrt{2}} + \sqrt{1-4x}\left(-\frac{1}{\sqrt{2}} + \log_2(\sqrt{1-4x}) + \frac{1}{4}\right),$$

then

$$E_n(\gamma) \sim \frac{[x^n]\tilde{U}(x)}{C^{(n)}},$$

which gives

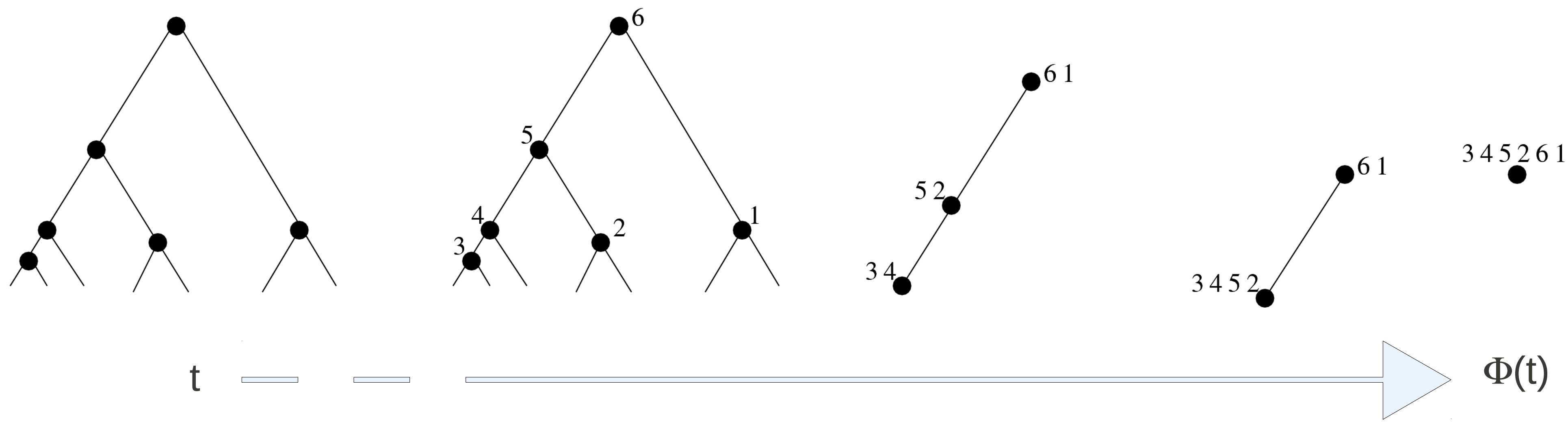
$$\frac{\log_2(n)}{E_n(\gamma)} \sim 1.$$

The average size of the biggest caterpillar subtree E_n(γ)

n	10	20	50	100	200	500	1000
$E_n(\gamma)$	4.535	5.120	6.202	7.107	8.052	9.334	10.318
$\frac{[x^n]\tilde{U}(x)}{C^{(n)}}$	4.032	5.109	6.47	7.490	8.498	9.824	10.825
$\log_2(n)$	3.322	4.322	5.644	6.644	7.644	8.966	9.966

Caterpillars realized in permutations avoiding 132

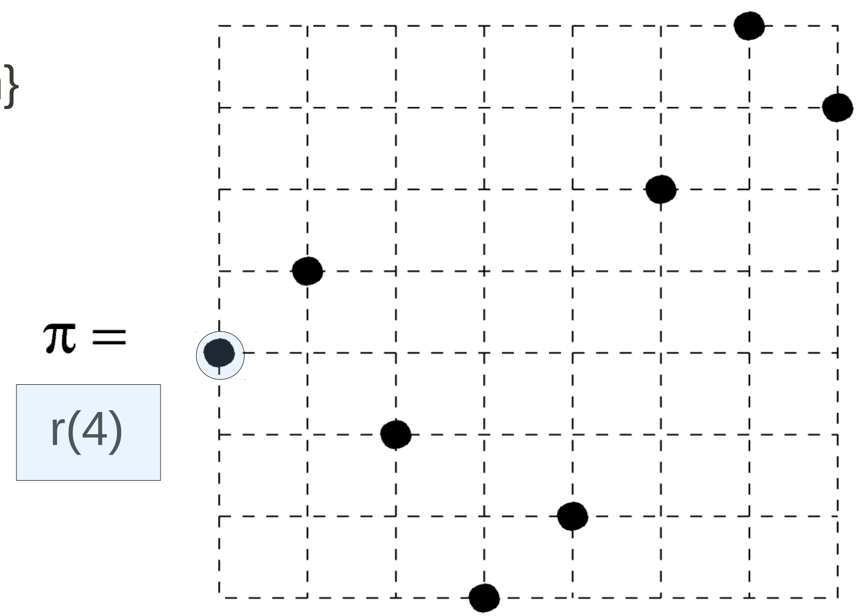
a) Bijection $\Phi:\mathrm{Trees}_{n+1} \rightarrow \mathrm{Permutations}_{n}$



b) $\pi = \pi_1 \pi_2 \ldots \pi_n \rightarrow r(\pi_i) \rightarrow r(\pi) = \{r(\pi_i) : 1 <= i <= n\}$

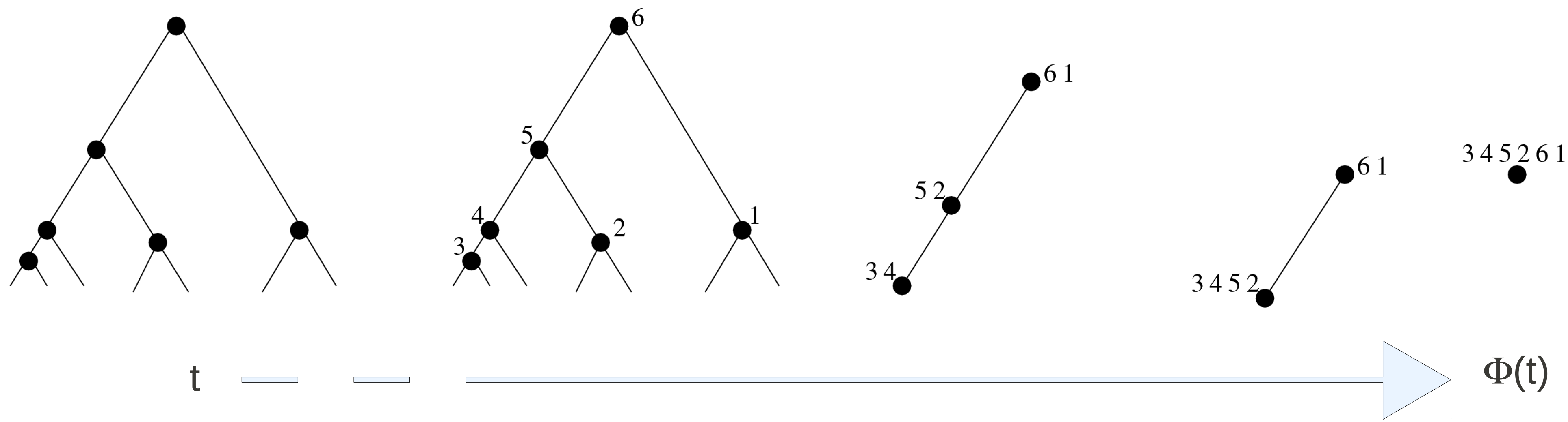
extract those π_j s.t.:

1) $\pi_j <= \pi_i$
2) k in $([j,i] \cup [i,j]) \rightarrow \pi_k <= \pi_i$



Caterpillars realized in permutations avoiding 132

a) Bijection Φ:Trees_{n+1} $\rightarrow$ Permutations_{n}



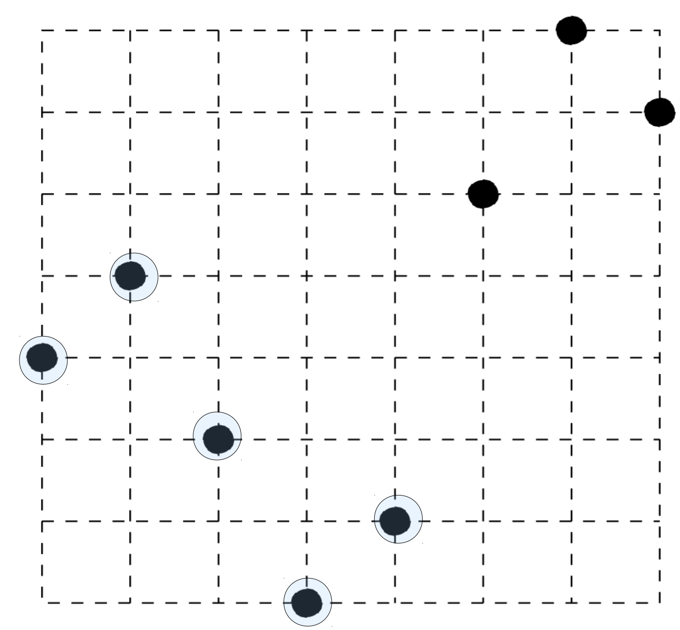
b) $\pi = \pi\_1\ \pi\_2 \ldots \pi\_n$ $\rightarrow$ r(π_i) $\rightarrow$ r(π) = {r(π_i) : 1 <= i <= n}

extract those π_j s.t.:

1) π_j <= π_i
2) k in ([j,i] U [i,j]) $\rightarrow$ π_k <= π_i

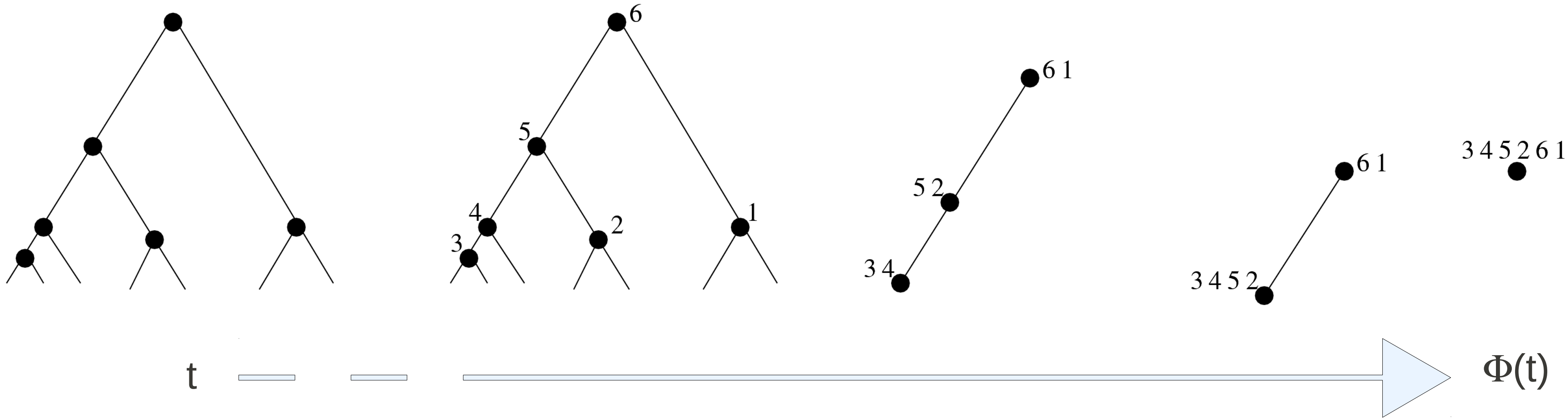
$\pi =$

r(5)



Caterpillars realized in permutations avoiding 132

a) Bijection Φ:Trees_{n+1} → Permutations_{n}

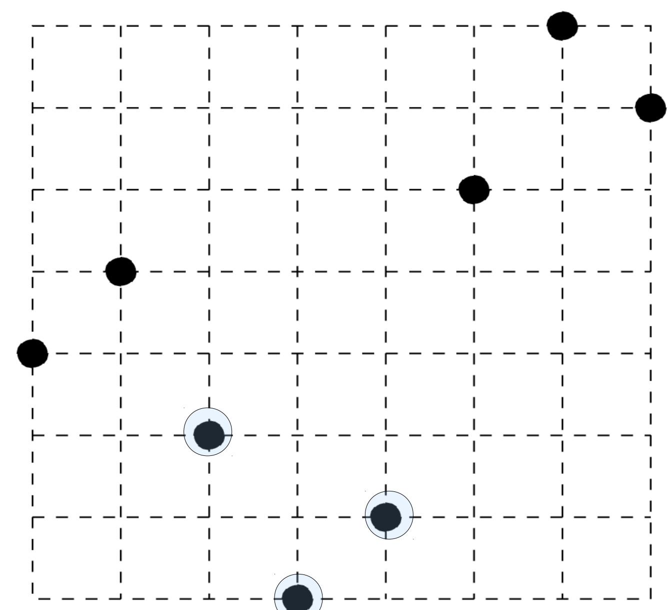


b) $\pi = \pi_1 \pi_2 \ldots \pi_n$ → $r(\pi_i)$ → $r(\pi) = \{r(\pi_i) : 1 <= i <= n\}$

extract those π_j s.t.:
1) $\pi_j <= \pi_i$
2) k in ([j,i] U [i,j]) → $\pi_k <= \pi_i$

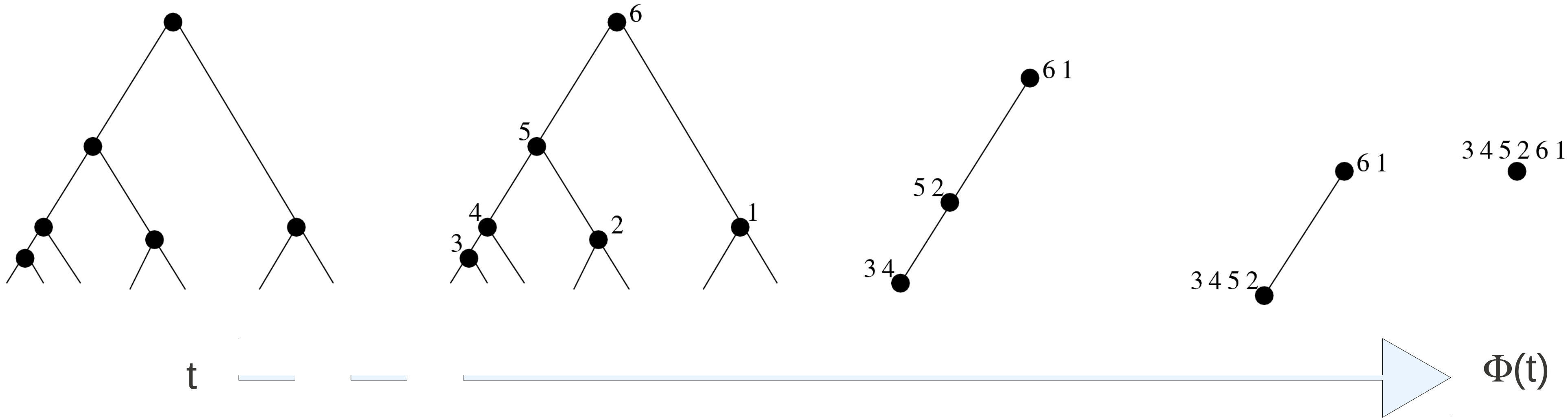
$\pi =$

r(3)



Caterpillars realized in permutations avoiding 132

a) Bijection Φ:Trees_{n+1} $\rightarrow$ Permutations_{n}



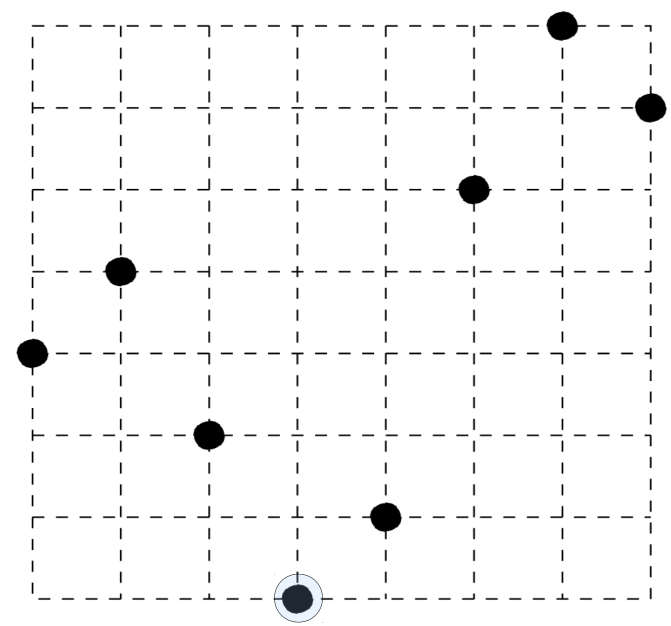
b) $\pi = \pi\_1\ \pi\_2 \ldots \pi\_n$ $\rightarrow$ r(π_i) $\rightarrow$ r(π) = {r(π_i) : 1 <= i <= n}

extract those π_j s.t.:

1) π_j <= π_i
2) k in ([j,i] U [i,j]) → π_k <= π_i

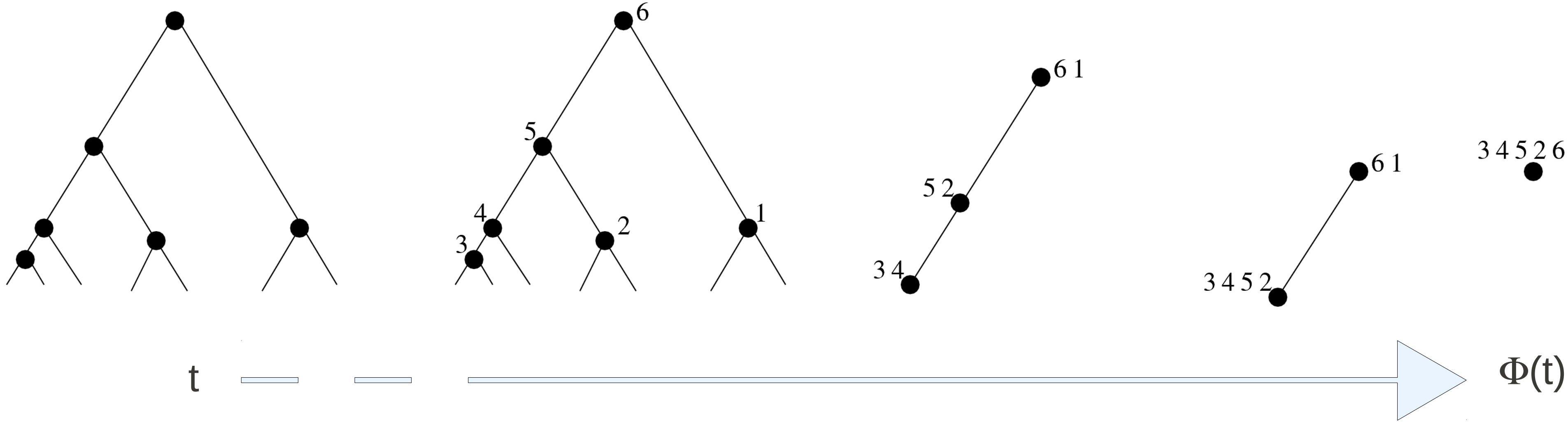
$\pi =$

r(1)



Caterpillars realized in permutations avoiding 132

a) Bijection Φ:Trees_{n+1} → Permutations_{n}

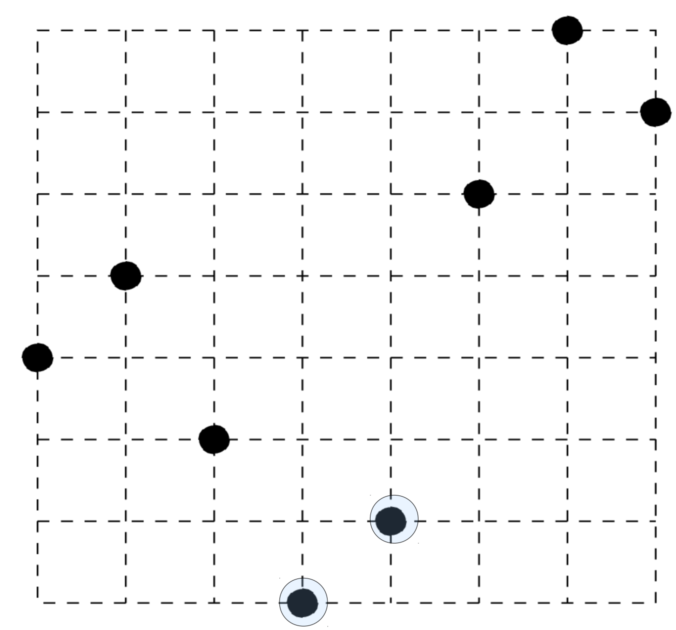


b) π = π_1 π_2 … π_n → r(π_i) → r(π) = {r(π_i) : 1 <= i <= n}

extract those π_j s.t.:
1) π_j <= π_i
2) k in ([j,i] U [i,j]) → π_k <= π_i

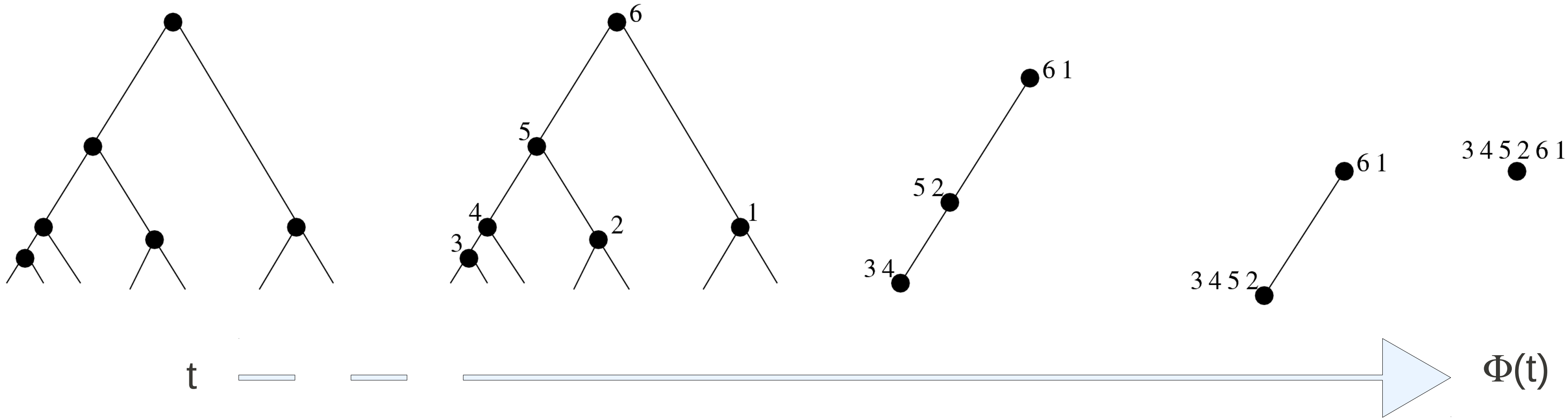
π =

r(2)



Caterpillars realized in permutations avoiding 132

a) Bijection Φ:Trees_{n+1} → Permutations_{n}

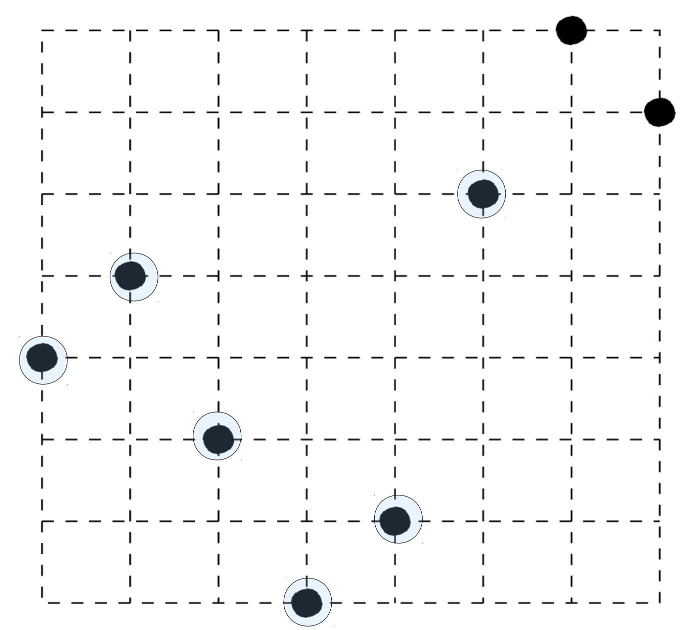


b) π = π_1 π_2 … π_n → r(π_i) → r(π) = {r(π_i) : 1 <= i <= n}

extract those π_j s.t.:
1) π_j <= π_i
2) k in ([j,i] U [i,j]) → π_k <= π_i

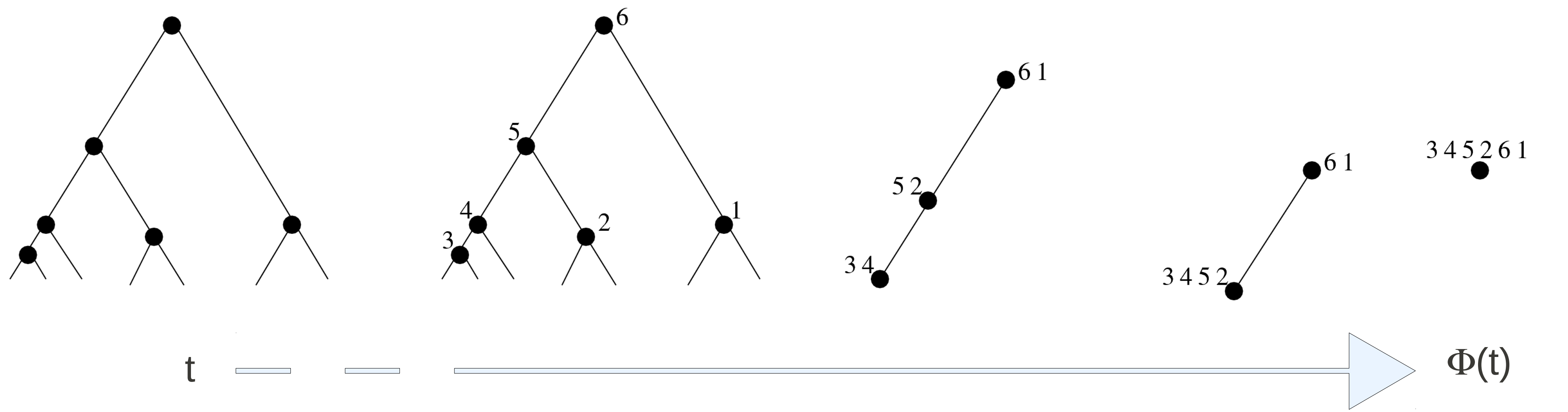
π =

r(6)



Caterpillars realized in permutations avoiding 132

a) Bijection $\Phi:\mathrm{Trees}_{n+1} \to \mathrm{Permutations}_{n}$

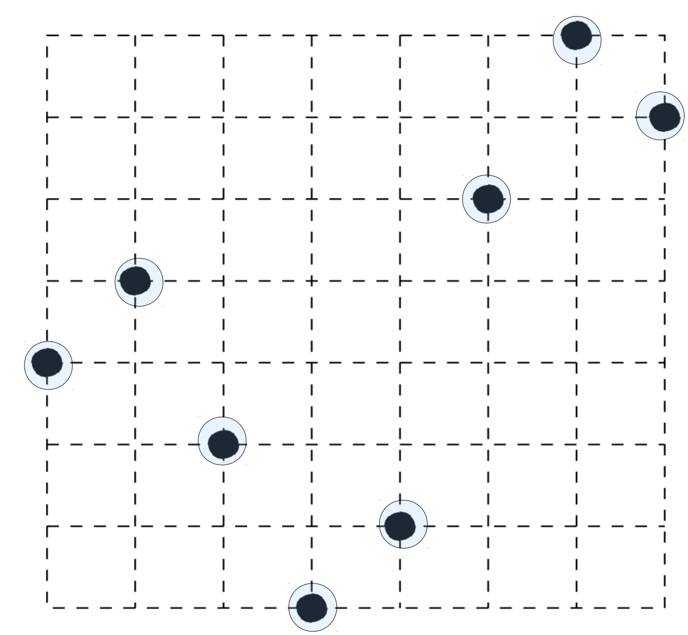


b) $\pi = \pi_1\, \pi_2 \ldots \pi_n \to r(\pi_i) \to r(\pi) = \{r(\pi_i) : 1 <= i <= n\}$

extract those π_j s.t.:
1) $\pi_j <= \pi_i$
2) k in ([j,i] U [i,j]) $\to \pi_k <= \pi_i$

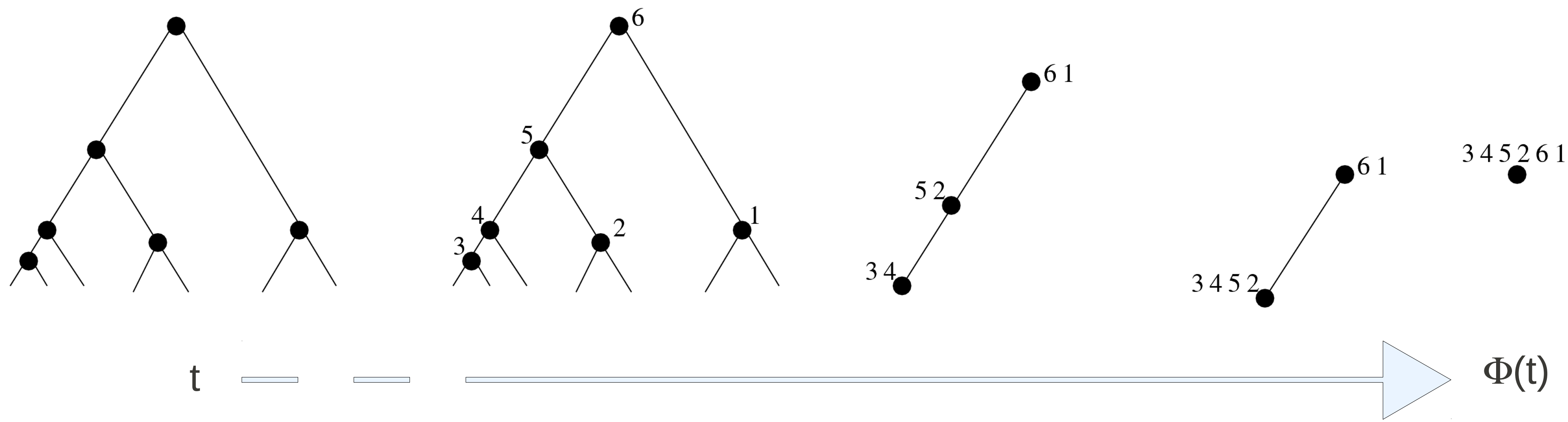
$\pi =$

r(8)



Caterpillars realized in permutations avoiding 132

a) Bijection Φ:Trees_{n+1} → Permutations_{n}

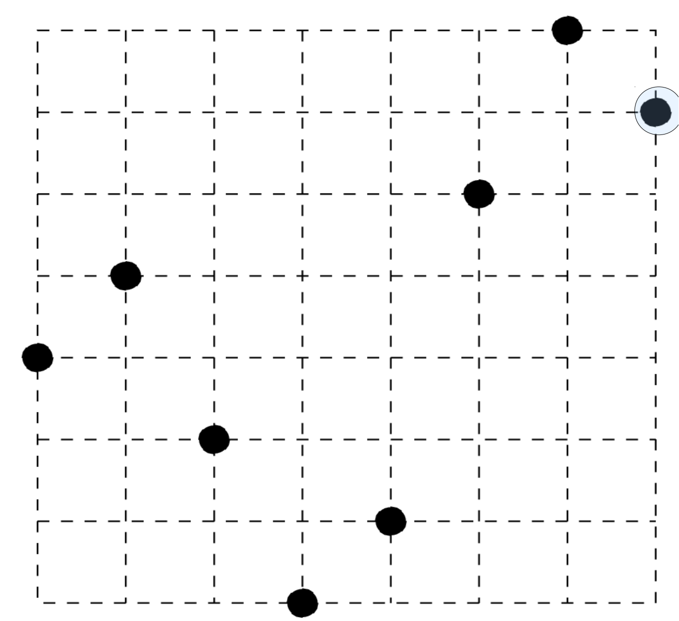


b) π = π_1 π_2 … π_n → r(π_i) → r(π) = {r(π_i) : 1 <= i <= n}

extract those π_j s.t.:
  1) π_j <= π_i
  2) k in ([j,i] U [i,j]) → π_k <= π_i

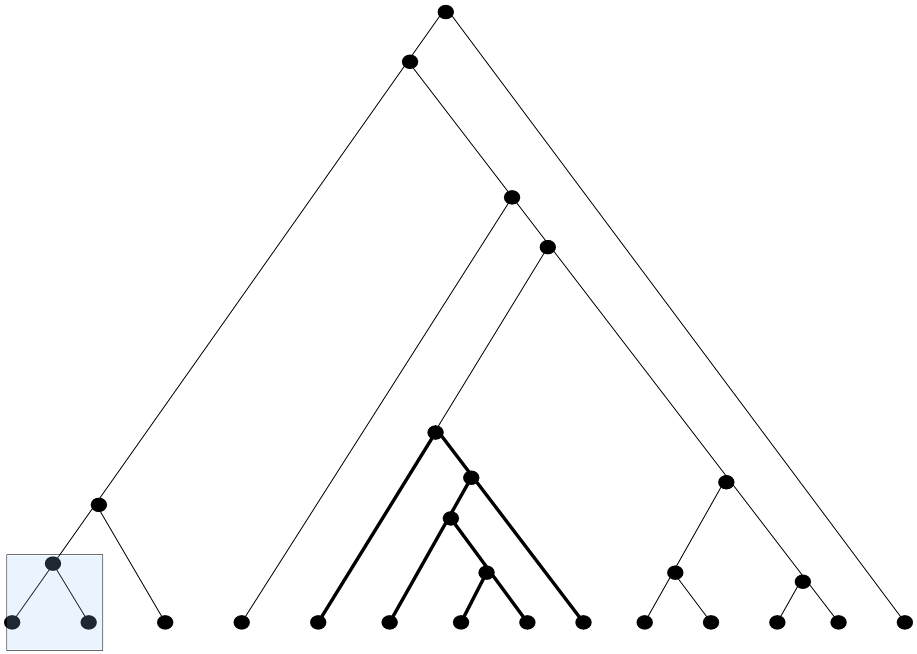
π =

r(7)



Caterpillars realized in permutations avoiding 132

Fact 1): Caterpillar subtrees of *t* correspond to those permutations in *r(ϕ(t))* avoiding 231, which means having no peaks (132 is forbidden, then 231 ↔ peak).

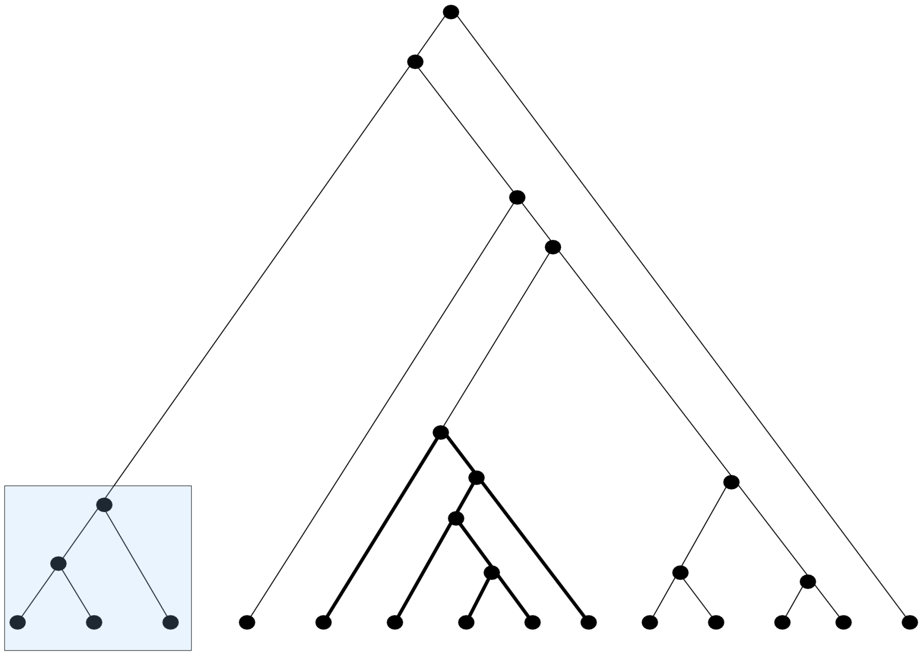


10 11 12 9 7 5 4 6 8 2 3 1 13

r(10) = 1

Caterpillars realized in permutations avoiding 132

Fact 1): Caterpillar subtrees of *t* correspond to those permutations in *r(ϕ(t))* avoiding 231, which means having no peaks (132 is forbidden, then 231 ↔ peak).

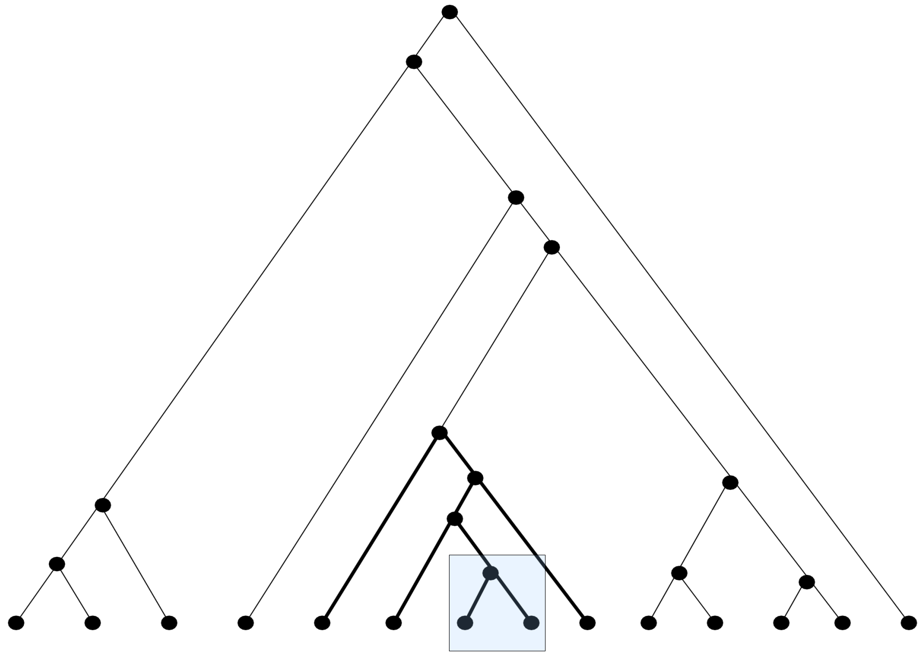


10 11 12 9 7 5 4 6 8 2 3 1 13

r(11) = 1 2

Caterpillars realized in permutations avoiding 132

Fact 1): Caterpillar subtrees of t correspond to those permutations in $r(\phi(t))$ avoiding 231, which means having no peaks (132 is forbidden, then 231 $\leftrightarrow$ peak).

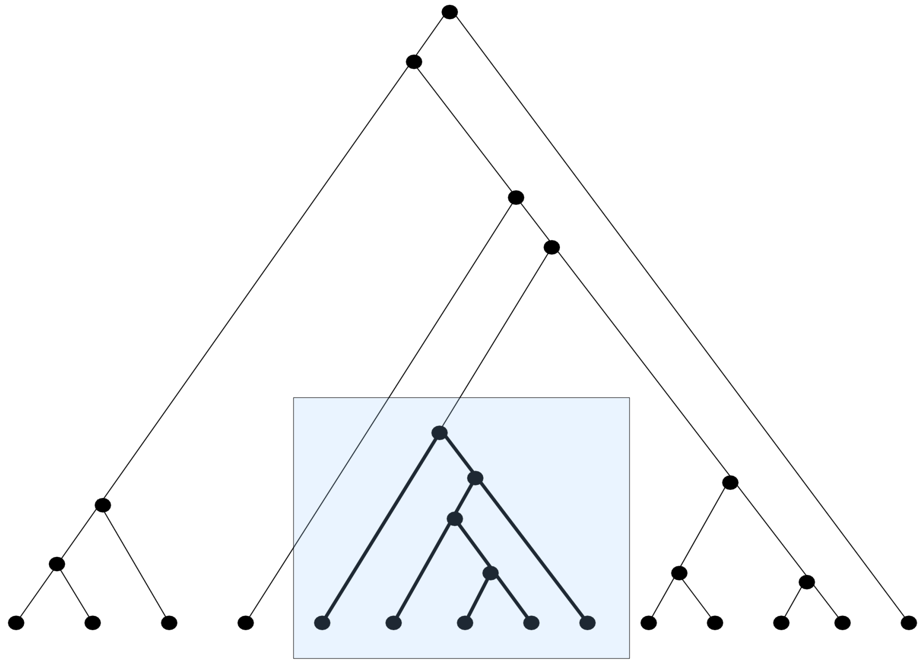


10 11 12 9 7 5 **4** 6 8 2 3 1 13

r(4) = 1

Caterpillars realized in permutations avoiding 132

Fact 1): Caterpillar subtrees of *t* correspond to those permutations in *r(ϕ(t))* avoiding 231, which means having no peaks (132 is forbidden, then 231 ↔ peak).

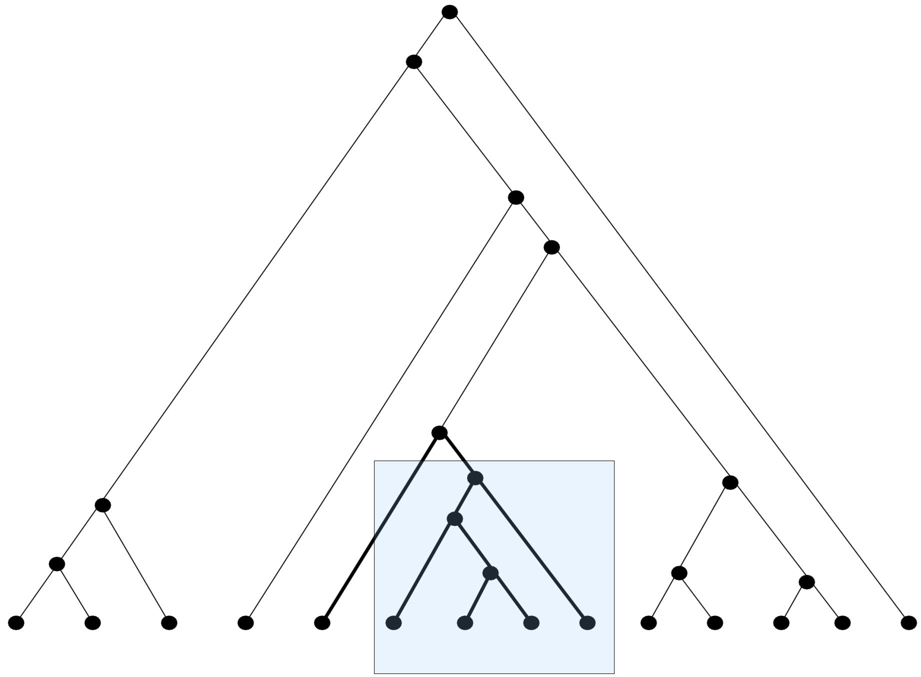


10 11 12 9 **7 5 4 6** 8 2 3 1 13

r(7) = 4 2 1 3

Caterpillars realized in permutations avoiding 132

Fact 1): Caterpillar subtrees of *t* correspond to those permutations in *r(ϕ(t))* avoiding 231, which means having no peaks (132 is forbidden, then 231 ↔ peak).

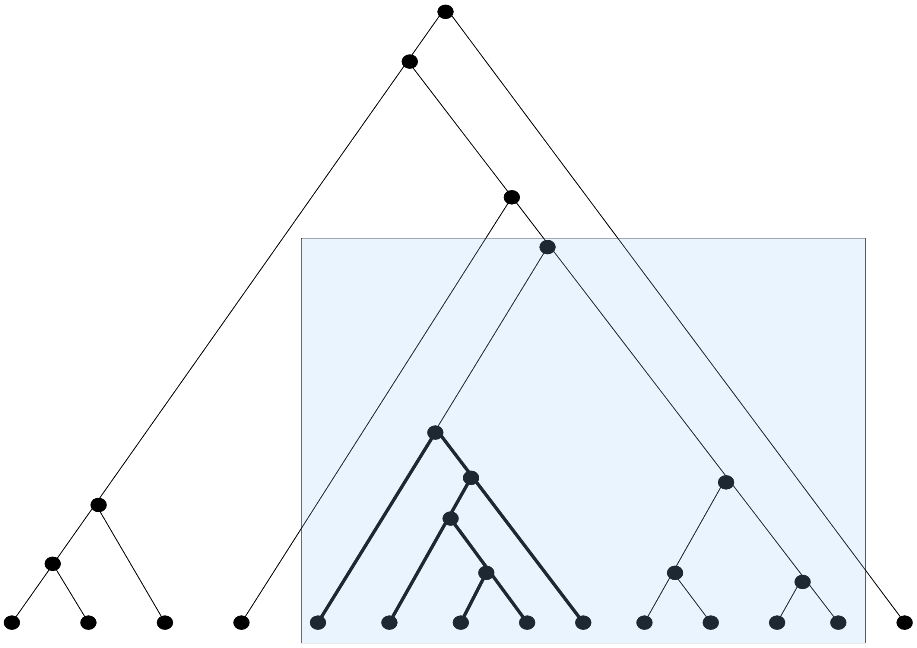


10 11 12 9 7 **5 4 6** 8 2 3 1 13

r(6) = 2 1 3

Caterpillars realized in permutations avoiding 132

Fact 1): Caterpillar subtrees of *t* correspond to those permutations in *r(ϕ(t))* avoiding 231, which means having no peaks (132 is forbidden, then 231 ↔ peak).

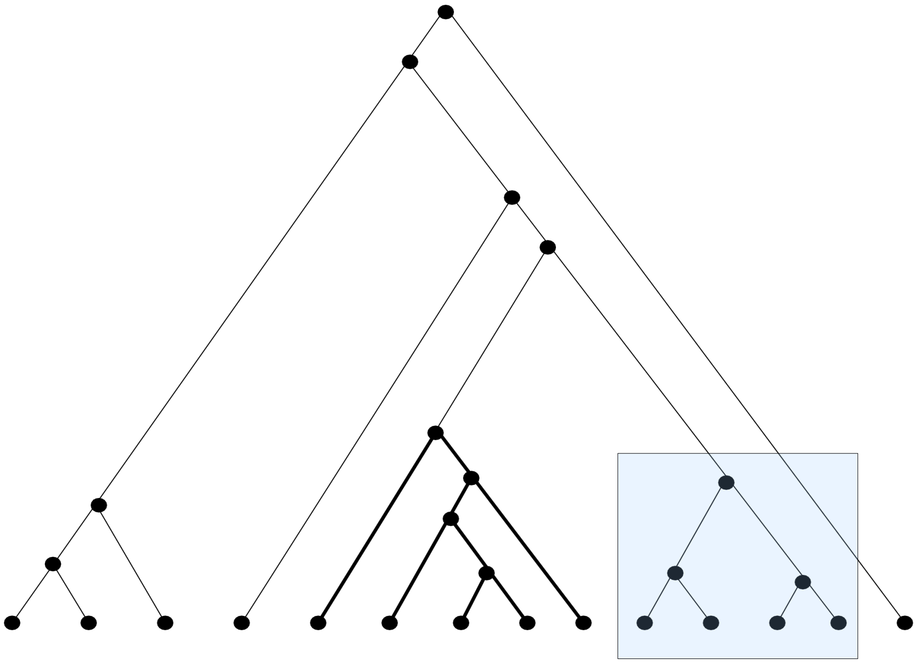


10 11 12 9 **7 5 4 6 8 2 3 1** 13

R(8) = 7 5 4 6 8 2 3 1 <u>8 and 3 are peaks</u>

Caterpillars realized in permutations avoiding 132

Fact 1): Caterpillar subtrees of *t* correspond to those permutations in *r(ϕ(t))* avoiding 231, which means having no peaks (132 is forbidden, then 231 ↔ peak).

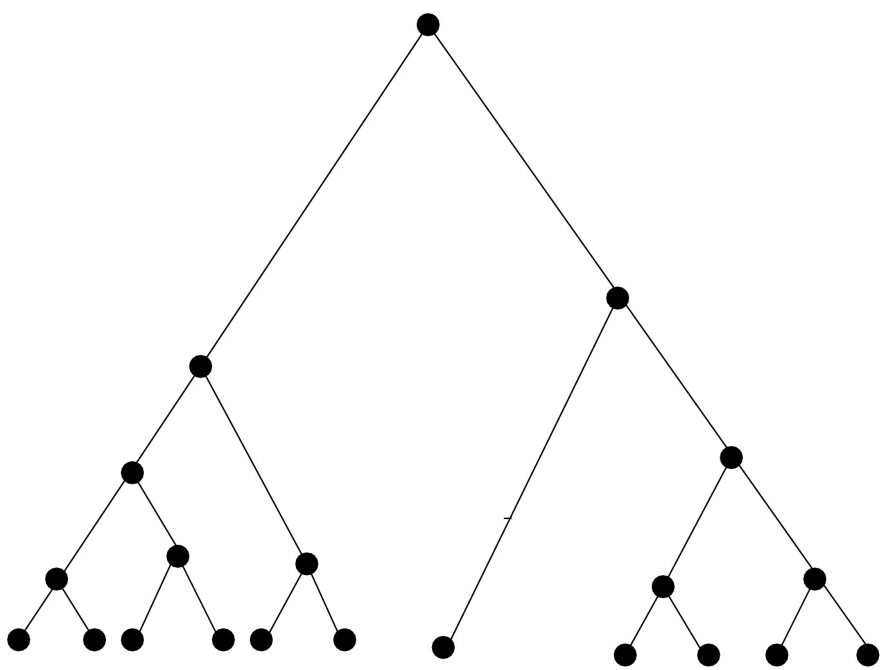


10 11 12 9 7 5 4 6 8 **2 3 1** 13

R(3) = 2 3 1 <u>3 is a peak</u>

Fact 2): Trees with <u>no pitchforks</u> correspond to those <u>permutations π avoiding 132 such that each element of *r*(π) whose size is greater than one contains the pattern 231, which means containing at least one peak</u> (132 is forbidden, then 231 ↔ peak).

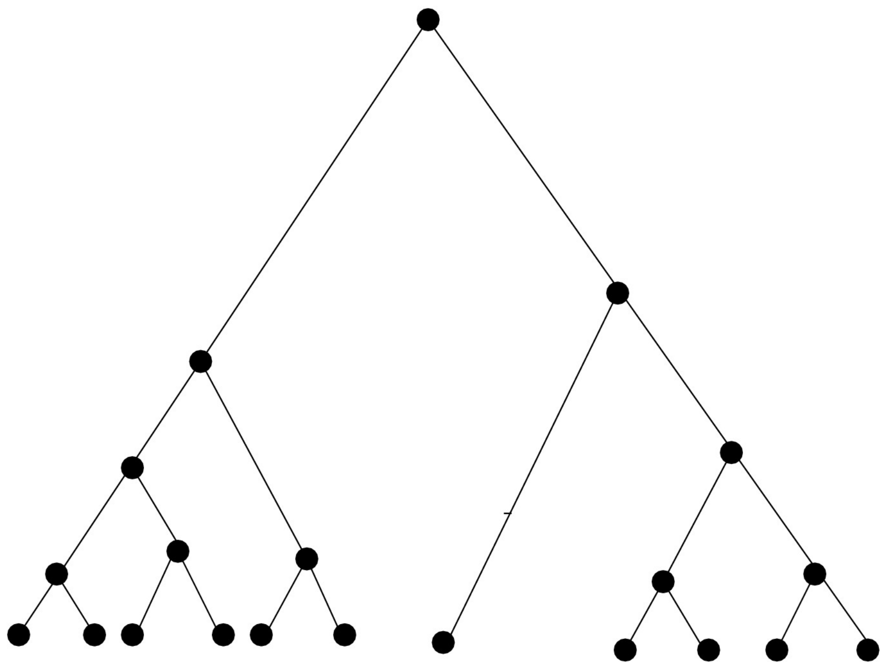
$$\frac{F_2(x)}{x} - 1 = \frac{1 - 2x - \sqrt{1 - 4x + 8x^3}}{2x}$$



7 8 6 9 5 10 4 **2 3 1**

Fact 2): Trees with no pitchforks correspond to those permutations π avoiding 132 such that each element of $r(\pi)$ whose size is greater than one contains the pattern 231, which means containing at least one peak (132 is forbidden, then 231 $\leftrightarrow$ peak).

$$\frac{F_2(x)}{x} - 1 = \frac{1 - 2x - \sqrt{1 - 4x + 8x^3}}{2x}$$

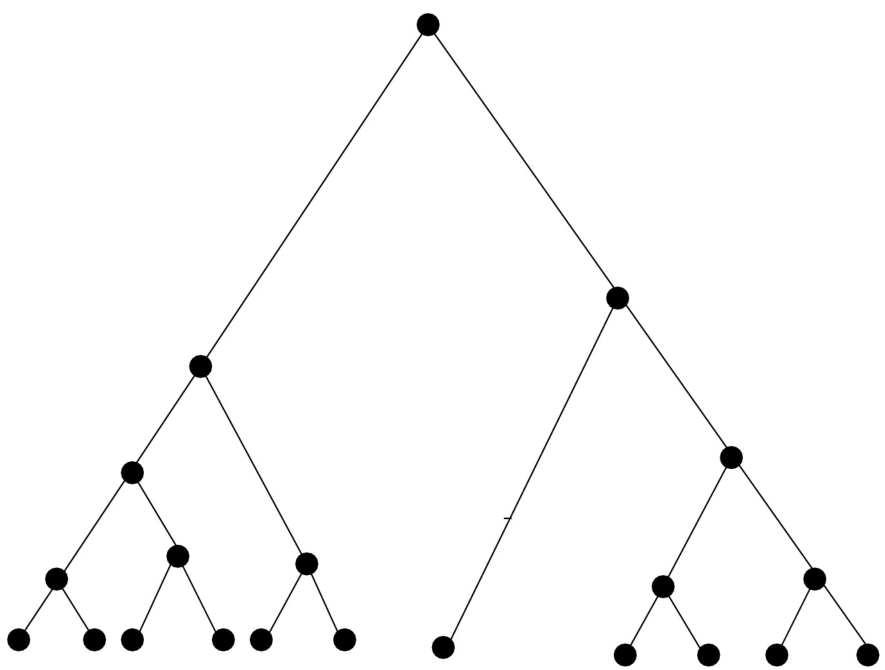


7 8 6 9 5 10 **4 2 3 1**

Caterpillars realized in permutations avoiding 132

Fact 2): Trees with no pitchforks correspond to those permutations π avoiding 132 such that each element of *r*(π) whose size is greater than one contains the pattern 231, which means containing at least one peak (132 is forbidden, then 231 ↔ peak).

$$\frac{F_2(x)}{x} - 1 = \frac{1 - 2x - \sqrt{1 - 4x + 8x^3}}{2x}$$

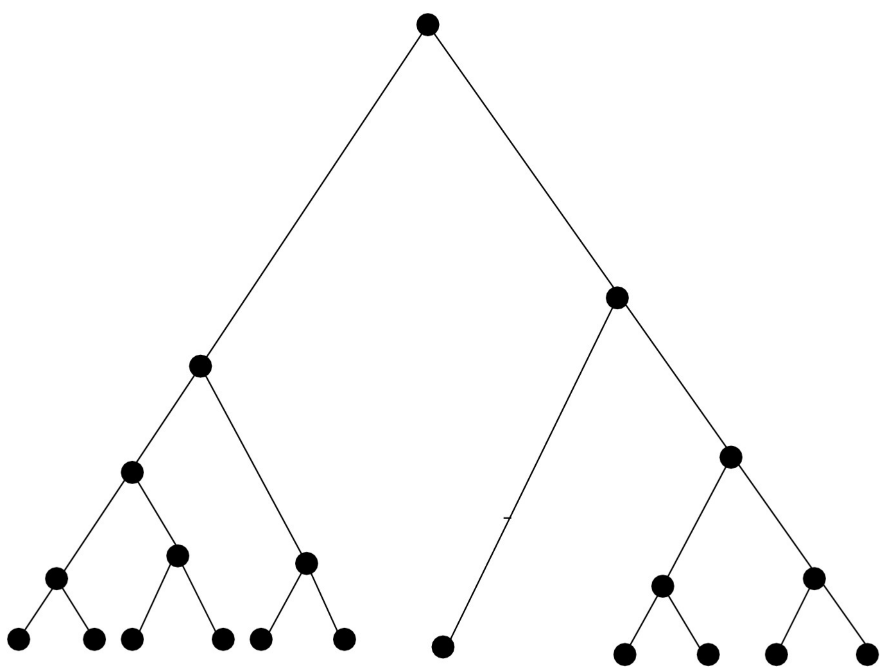


7 8 6 9 5 10 4 2 3 1

Caterpillars realized in permutations avoiding 132

Fact 2): Trees with no pitchforks correspond to those permutations π avoiding 132 such that each element of $r(\pi)$ whose size is greater than one contains the pattern 231, which means containing at least one peak (132 is forbidden, then 231 $\leftrightarrow$ peak).

$$\frac{F_2(x)}{x} - 1 = \frac{1 - 2x - \sqrt{1 - 4x + 8x^3}}{2x}$$

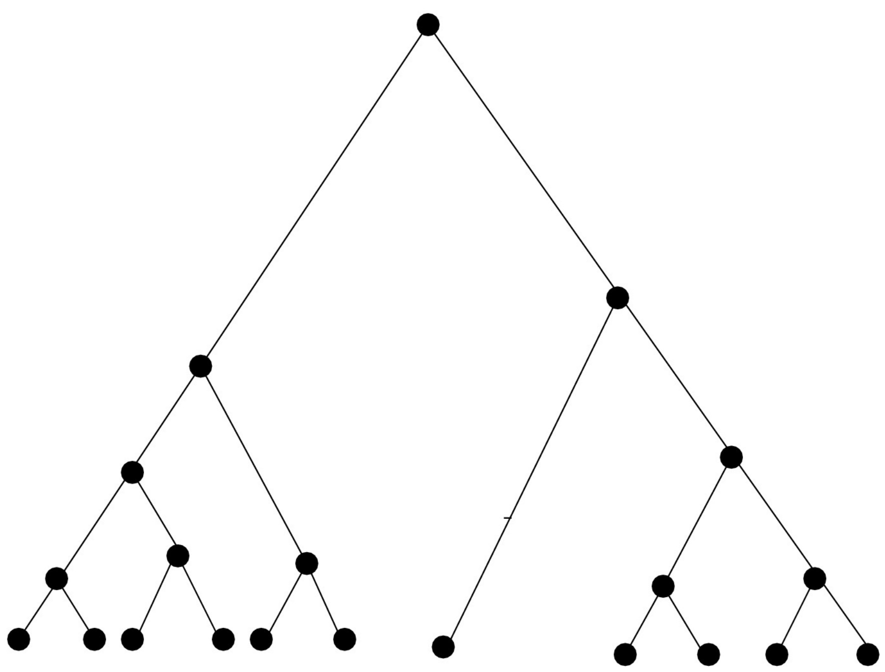


7 8 6 9 5 10 4 2 3 1

Caterpillars realized in permutations avoiding 132

Fact 2): Trees with no pitchforks correspond to those permutations π avoiding 132 such that each element of *r*(π) whose size is greater than one contains the pattern 231, which means containing at least one peak (132 is forbidden, then 231 ↔ peak).

$$\frac{F_2(x)}{x} - 1 = \frac{1 - 2x - \sqrt{1 - 4x + 8x^3}}{2x}$$



7 8 6 9 5 10 4 2 3 1

Further works, thanks and questions

a) caterpillars and binary rooted un-ordered trees

b) further studies on $r(\pi)$ and permutations

Thank you for your attention,

any question?