To Alberto Apostolico

Repetition length in random sequences

Ph.Chassignet and M. Régnier

INRIA-Team AMIBIO

June, 22nd – 2017

Motivation

Many repetitive structures in genomic sequences:

- microsatellites
- DNA transposons
- long terminal repeats
- long interspersed nuclear elements
- ribosomal DNA
- short interspersed nuclear elements

Treangen&Salzberg2012: half of the genome : repetitive elements.

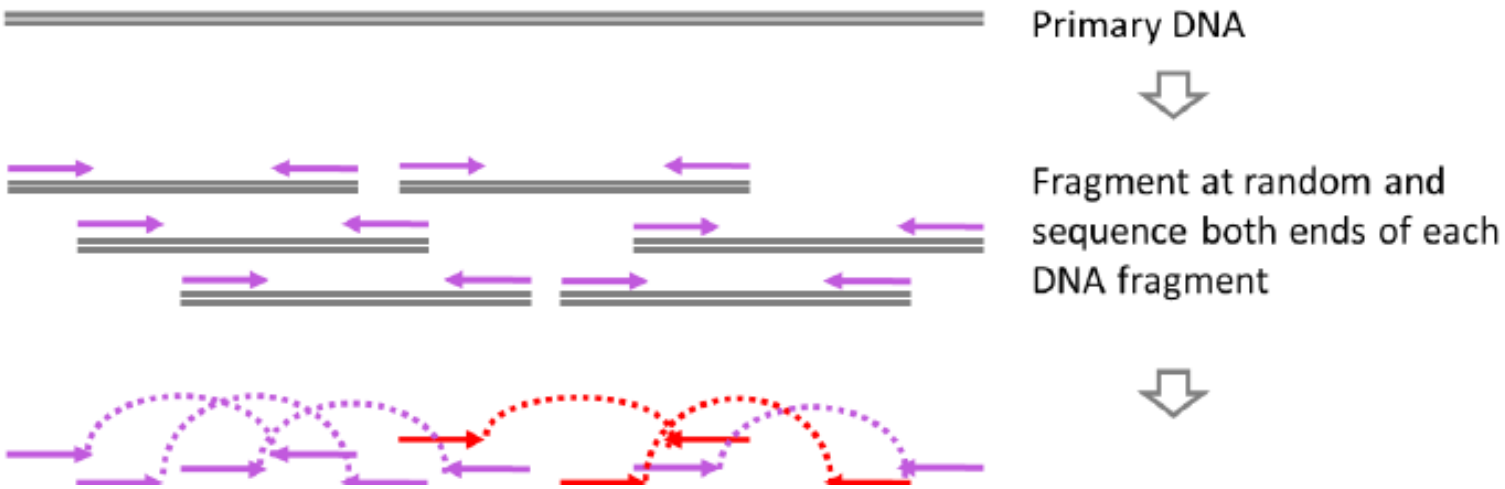
Applications : assembly, de Bruijn graphs, ...

Challenges

1. Direct sequencing: at most 700 bp.
2. Human genome: 10^8 bp. (23 chromosomes)

- Million/milliards pieces
- Without final image (de novo vs resequencing)
- forward, backward ???
- Sequencing errors:
  Pieces do not perfectly match
- Coverage:
  - Multiple copies
  - Some puzzle pieces missing

NGS scheme


Primary DNA
Fragment at random and sequence both ends of each DNA fragment

Assembly strategies (0)

1. Find overlapping reads

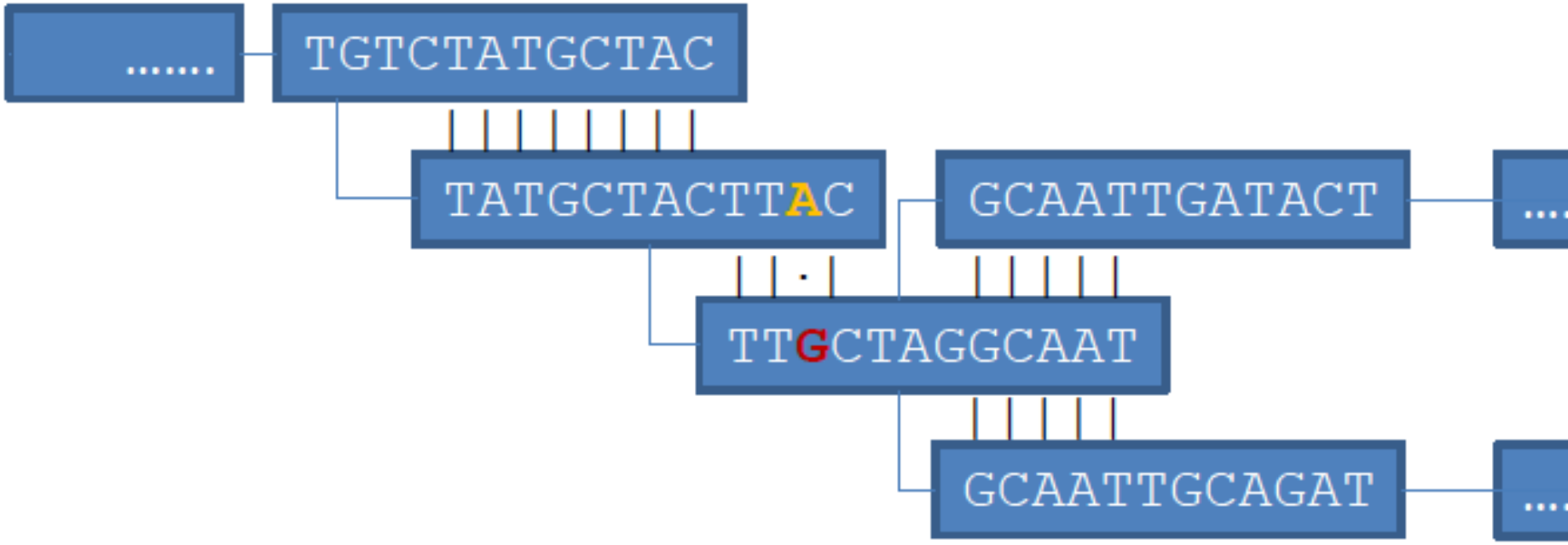
2. Merge good pairs of reads into longer contigs

3. Derive consensus sequence

$$..ACGATTACAATAGGTT..$$

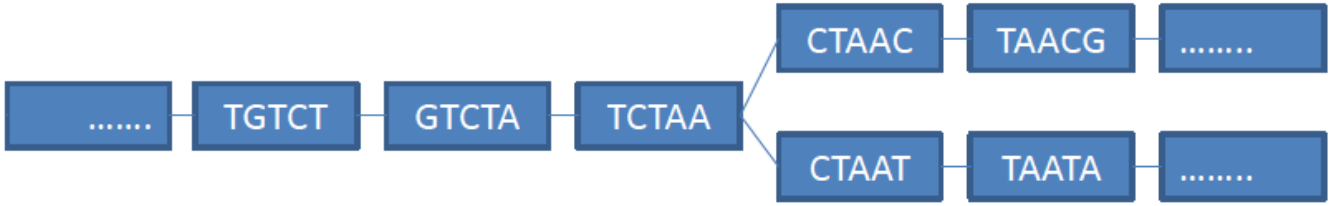
Assembly strategies (1)

Overlap graph, layout graph.



Assembly strategies (2)

de Bruijn graph.



- Reads $\rightarrow$ k-mers
- Node $=$ one k-mers
- Edge $\rightarrow$ 1 $(k-1)$-mer

State of the art

Model: trie versus (word,sequence) repetition

Deviations from uniformity

- **Flajolet&Nigel** : binary alphabet Σ; uniform Bernoulli model:
  - almost all words of length $\leq k$ appear.
  - almost no word of length $> k$ appear.
- **Park&al. 2009**; binary alphabet; biased Bernoulli model:
  transition domain for trie profile:
  "many" words of length k appear.

State of the art

Model: trie versus (word,sequence) repetition

Deviations from uniformity

- **Flajolet&Nigel** : binary alphabet Σ; uniform Bernoulli model:
  - almost all words of length $\leq k$ appear.
  - almost no word of length $> k$ appear.
- **Park&al. 2009**; binary alphabet; biased Bernoulli model: transition domain for trie profile: "many" words of length k appear.

General alphabets ?

Method

Analytic combinatorics

- functional equation on a generating function, or an induction.
- asymptotics of coefficients of G.F. (Mellin, saddle point; ...)
- Bernoulli-Poisson cycle

Method

Analytic combinatorics

- functional equation on a generating function, or an induction.
- asymptotics of coefficients of G.F. (Mellin, saddle point; ...)
- Bernoulli-Poisson cycle

- probability $\Rightarrow$ coefficients
- Lagrange multipliers

Words and tries

Axiom: repeat $\Leftrightarrow$ internal node

Words and tries

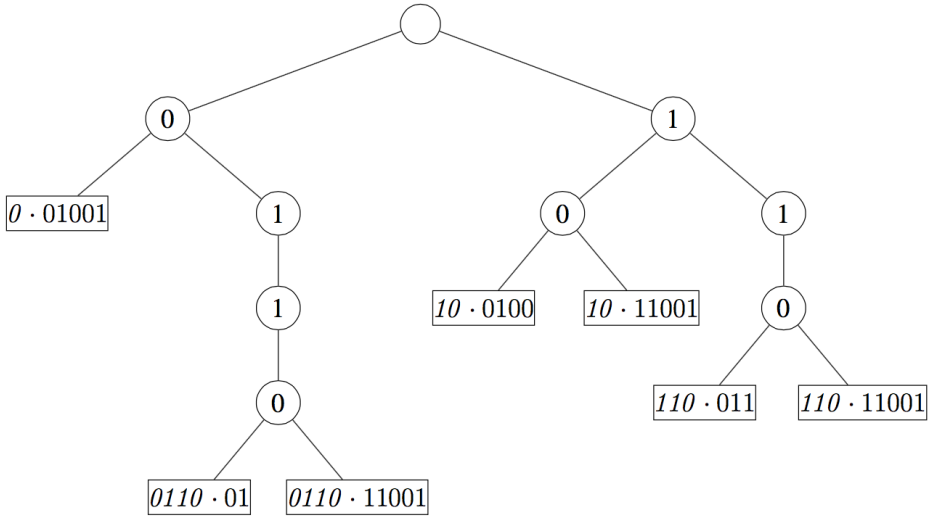
Axiom: repeat $\Leftrightarrow$ internal node

Unique k-mer :

wa : once; w : twice; $|wa| = k$

- In the sequence : $wa \cdots wb$
  w : (right) maximal repeat
- In a trie :
  w : internal node ; w : leaf

Myriad virtues of Tries (and Suffix arrays)



Notations

n words OR sequence of length n

$$
\begin{aligned}
B(n,k) &= \#\text{unique } k\text{-mers} \\
\mu(n,k-1) &= E(B(n,k)) \\
\alpha &= \frac{k}{\log n}
\end{aligned}
$$

Notations

n words OR sequence of length n

$$
\begin{aligned}
B(n,k) &= \#\text{unique } k\text{-mers} \leq n \\
\mu(n,k-1) &= E(B(n,k)) \sim B(n,k)\text{: LLN} \\
\alpha &= \frac{k}{\log n} \quad 0 \cdots \infty
\end{aligned}
$$

Notations

n words OR sequence of length n
Σ alphabet $\chi_1, \cdots, \chi_V$
Probabilities: $p_1, \cdots, p_V$

$$\beta_i = \log \frac{1}{p_i} \quad .$$

$$p_{min} = \min\{p_i; 1 \leq i \leq V\} \quad \text{and} \quad \alpha_{min} = \frac{1}{\log \frac{1}{p_{min}}} = \frac{1}{\max(\beta_i)}$$

$$p_{max} = \max\{p_i; 1 \leq i \leq V\} \quad \text{and} \quad \alpha_{max} = \frac{1}{\log \frac{1}{p_{max}}} = \frac{1}{\min(\beta_i)}$$

k-mers classification

Barycentric coordinates & objective function

$$\rho(k_1, \cdots, k_V) = \sum_{i=1}^{V} \frac{k_i}{k} \beta_i - \frac{1}{\alpha} \quad . \tag{1}$$

$\sum_{i=1}^{V} \frac{k_i}{k} \beta_i \in [\min(\beta_i), \max(\beta_i)]$

k-mers classification

Barycentric coordinates & objective function

$$\rho(k_1, \cdots, k_V) = \sum_{i=1}^{V} \frac{k_i}{k} \beta_i - \frac{1}{\alpha} \quad . \tag{1}$$

A k-mer $w\chi_i$ is said

- a *common k-mer* if $\rho(k_1, \cdots, k_V) < 0$;
- a *transition k-mer* if $\rho(k_1, \cdots, k_V) \geq 0$ and its ancestor is a common k-mer;
- a *rare k-mer* , otherwise.

k-mers classification

Barycentric coordinates & objective function

$$\rho(k_1, \cdots, k_V) = \sum_{i=1}^{V} \frac{k_i}{k} \beta_i - \frac{1}{\alpha} \quad . \tag{1}$$

A k-mer $w\chi_i$ is said

- a *common k-mer* if $\rho(k_1, \cdots, k_V) < 0$; $E(w\chi_i) > 1$
- a *transition k-mer* if $\rho(k_1, \cdots, k_V) \geq 0$ and its ancestor is a common k-mer; $E(w\chi_i) \leq 1, E(w) > 1$
- a *rare k-mer* ; $E(w) \leq 1$

k-mers classification

Barycentric coordinates & objective function

$$\rho(k_1, \cdots, k_V) = \sum_{i=1}^{V} \frac{k_i}{k} \beta_i - \frac{1}{\alpha} \quad . \tag{1}$$

A k-mer $w\chi_i$ is said

- a *common k-mer* if $\rho(k_1, \cdots, k_V) < 0$; $E(w\chi_i) > 1$
- a *transition k-mer* if $\rho(k_1, \cdots, k_V) \geq 0$ and its ancestor is a common k-mer; $E(w\chi_i) \leq 1, E(w) > 1$
- a *rare k-mer* ; $E(w) \leq 1$

Main contribution for each given level k :transition nodes.

Combinatorial sums

$$\mu(n,k) = n \sum_{k_1+\cdots k_V=k} \binom{k}{k_1,\cdots,k_V} \phi(k_1,\cdots,k_V)\psi_n(k_1,\cdots,k_V) \tag{2}$$

$\phi(k_1,\cdots,k_V) = p_1^{k_1}\cdots p_V^{k_V}$

$\psi \ : \sum_{i=1}^{V} p_i[(1-\phi(k_1,\cdots,k_V)p_i)^{n-1} - (1-\phi(k_1,\cdots,k_V))^{n-1}]$

Combinatorial sums

$$\mu(n,k) = n \sum_{k_1+\cdots k_V=k} \binom{k}{k_1,\cdots,k_V} \phi(k_1,\cdots,k_V)\psi_n(k_1,\cdots,k_V)$$

$\phi(k_1,\cdots,k_V)p_i = p_1^{k_1}\cdots p_V^{k_V} p_i$: $P(w\chi_i)$

ψ : $\sum_{i=1}^{V} p_i[(1-\phi(k_1,\cdots,k_V)p_i)^{n-1} - (1-\phi(k_1,\cdots,k_V))^{n-1}]$

$(1-\phi(k_1,\cdots,k_V)p_i)^{n-1}$: no other $w\chi_i$

$(1-\phi(k_1,\cdots,k_V))^{n-1}$: at least an other w

Combinatorial sums

$$S(k) = n \sum_{D_k(n)} \binom{k}{k_1 \cdots k_V} \phi(k_1, \cdots, k_V) \psi_n(k_1, \cdots, k_V) \ ;$$

$$T(k) = n \sum_{E_k(n)} \binom{k}{k_1 \cdots k_V} \phi(k_1, \cdots, k_V) \psi_n(k_1, \cdots, k_V) \ .$$

Tech: two diff. approx. when

- w : rare or transition
- w : common

Computable for moderate k.

Lagrange multipliers

Large Deviation Principle

$$np_1^{k_1} \cdots p_V^{k_V} \quad = \quad e^{-k\rho(k_1,\cdots,k_V)}$$

$$\binom{k}{k_1,\cdots,k_V} \phi(k_1,\cdots,k_V) \quad \to e^{-k\sum_i \frac{k_i}{k}\log\frac{k_i}{k}}$$

Dominating contribution $S(k), T(k)$: $\rho(k_1,\cdots,k_V) = 0$.

Large Deviation principle

Main contribution

For each given level k :transition nodes.

Maximization problem

$\sim \max\{-\sum_{i=1}^{V} \frac{k_i}{k} \log \frac{k_i}{k}; \rho(k_1, \cdots, k_V) = 0\}$

Rewrite :

$\max\{\sum_{i=1}^{V} \theta_i \log \frac{1}{\theta_i}; \sum_{i=1}^{V} \theta_i = 1; \sum_{i=1}^{V} \beta_i \theta_i = \frac{1}{\alpha}; 0 \leq \theta_i \leq 1\}$

Lagrange multipliers and Large Deviation Principle

Lagrange multipliers

$\max\{\sum_{i=1}^{V} \theta_i \log \frac{1}{\theta_i}; \sum_{i=1}^{V} \theta_i = 1; \sum_{i=1}^{V} \beta_i \theta_i = \frac{1}{\alpha}; 0 \leq \theta_i \leq 1\}$

Implicit equation solution

Let τ_α be the unique real root of the equation

$$\frac{1}{\alpha} = \frac{\sum_{i=1}^{V} \beta_i e^{-\beta_i \tau}}{\sum_{i=1}^{V} e^{-\beta_i \tau}} \tag{2}$$

Let ψ be the function defined in $[\alpha_{min}, \alpha_{ext}]$ as

$$\alpha_{min} \leq \alpha \leq \bar{\alpha} \quad : \quad \psi(\alpha) = \tau_\alpha + \alpha \log(\sum_{i=1}^{V} e^{-\beta_i \tau_\alpha}) \ ;$$

$$\bar{\alpha} \leq \alpha \quad : \quad \psi(\alpha) = 2 - \alpha \log \frac{1}{\sigma_2} \ .$$

Results and interpretation

0 —— α_{min} —— $\tilde{\alpha}$ —— $\bar{\alpha}$ —— α_{max} —— α_{ext} ——

- $\alpha \leq \alpha_{min}$: **all nodes are common** : $\frac{\log \mu(n,k)}{\log n} \leq 0.$
- **common, transition and rare** :
- **all nodes are rare**
  - $\alpha_{max} \leq \alpha \leq \alpha_{ext}$: LLN

$$\frac{\log \mu(n,k)}{\log n} = \psi_2(\alpha) = 2 - \alpha \log \frac{1}{\sigma_2}$$

  - $\alpha \geq \alpha_{ext}$:

$$\frac{\log \mu(n,k)}{\log n} \leq 0$$

Results and interpretation

0 —— α_{min} —— $\tilde{\alpha}$ —— $\bar{\alpha}$ —— α_{max} —— α_{ext} ——

common, transition and rare

- $\alpha_{min} \leq \alpha \leq \tilde{\alpha}$: transition k-mers increase

$$\frac{\log \mu(n,k)}{\log n} = \psi_1(\alpha)$$

- $\tilde{\alpha} \leq \alpha \leq \bar{\alpha}$: transition k-mers decrease

$$\frac{\log \mu(n,k)}{\log n} = \psi_1(\alpha)$$

- $\bar{\alpha} \leq \alpha_{max}$: transition k-mers decrease

$$\frac{\log \mu(n,k)}{\log n} = \psi_2(\alpha) = 2 - \alpha \log \frac{1}{\sigma_2}$$

Simulations

k	observed $B(k+1)$	predicted $S(k)$	$T(k)$	$\mu(N,k)$	observed $\frac{\log B(k+1)}{\log N}$	asymptotic $\psi(\alpha)$	$\psi(\alpha)+\xi(\alpha)$	
11	0.29	0.0	0.3	0.3	-0.0803			
12	7.91	0.0	8.3	8.3	0.1341			k_{min}
13	87.87	0.1	86.9	87.1	0.2902	**0.0843**	*0.0012*	
14	552.88	1.2	550.3	551.5	0.4094	**0.3340**	*0.2485*	
15	2456.77	86.6	2366.4	2453.0	0.5061	**0.4962**	*0.4085*	
16	8269.20	209.4	8069.1	8278.5	0.5848	**0.6181**	*0.5282*	
17	22516.20	406.1	22097.7	22503.8	0.6497	**0.7136**	*0.6218*	
18	51085.15	4823.8	46267.2	51091.0	0.7028	**0.7897**	*0.6960*	
19	99387.01	6636.1	92717.6	99353.7	0.7460	**0.8504**	*0.7549*	
20	169303.03	37415.5	131882.6	169298.1	0.7805	**0.8984**	*0.8013*	
21	256358.10	42003.9	214454.4	256458.3	0.8074	**0.9357**	*0.8370*	
22	349801.23	137615.9	212264.2	349880.1	0.8276	**0.9635**	*0.8634*	
23	434625.83	134807.6	299824.7	434632.4	0.8416	**0.9830**	*0.8814*	
24	495572.93	122283.1	373279.8	495562.8	0.8501	**0.9949**	*0.8919*	
25	522788.19	255284.4	267476.3	522760.7	0.8536	**0.9998**	*0.8955*	$\tilde{k}$
26	513374.76	211204.2	302252.5	513456.7	0.8524	**0.9982**	*0.8926*	
27	472126.51	315154.7	157087.0	472241.6	0.8470	**0.9906**	*0.8838*	
28	408946.76	242583.4	166360.3	408943.7	0.8377	**0.9772**	*0.8692*	
29	335080.05	273441.0	61579.7	335020.7	0.8248	**0.9582**	*0.8491*	
30	260999.29	198163.4	62712.5	260875.9	0.8086	**0.9339**	*0.8236*	
31	194100.36	137502.0	56463.1	193965.1	0.7894	**0.9043**	*0.7930*	$\bar{k}$
32	138437.13	122218.3	16090.9	138309.2	0.7675	**0.8699**	*0.8136*	
33	95017.33	80937.1	14067.8	95004.9	0.7431	**0.8346**	*0.7783*	

Simulations

k	observed $B(k+1)$	predicted $S(k)$	predicted $T(k)$	predicted $\mu(N,k)$	observed $\frac{\log B(k+1)}{\log N}$	asymptotic $\psi(\alpha)$	asymptotic $\psi(\alpha)+\xi(\alpha)$	
12	7.91	0.0	8.3	8.3	0.1341			k_{min}
13	87.87	0.1	86.9	87.1	0.2902	**0.0843**	*0.0012*	
19	99387.01	6636.1	92717.6	99353.7	0.7460	**0.8504**	*0.7549*	
24	495572.93	122283.1	373279.8	495562.8	0.8501	**0.9949**	*0.8919*	
25	522788.19	255284.4	267476.3	522760.7	0.8536	**0.9998**	*0.8955*	$\tilde{k}$
26	513374.76	211204.2	302252.5	513456.7	0.8524	**0.9982**	*0.8926*	
27	472126.51	315154.7	157087.0	472241.6	0.8470	**0.9906**	*0.8838*	
29	335080.05	273441.0	61579.7	335020.7	0.8248	**0.9582**	*0.8491*	
31	194100.36	137502.0	56463.1	193965.1	0.7894	**0.9043**	*0.7930*	$\bar{k}$
32	138437.13	122218.3	16090.9	138309.2	0.7675	**0.8699**	*0.8136*	
34	63082.67	60397.1	2744.6	63141.7	0.7165	**0.7993**	*0.7430*	
36	25679.21	23888.2	1817.4	25705.6	0.6582	**0.7286**	*0.6724*	
38	9645.84	9455.0	194.2	9649.2	0.5948	**0.6580**	*0.6018*	
40	3433.87	3426.4	12.1	3438.5	0.5278	**0.5874**	*0.5311*	
42	1188.84	1189.0	0.3	1189.3	0.4590	**0.5167**	*0.4605*	
43	692.28	694.8	0.2	695.0	0.4240	**0.4814**	*0.4252*	k_{max}
44	402.75	405.1	0.0	405.1	0.3889	**0.4461**	*0.3899*	
46	135.42	137.0	0.0	137.0	0.3182	**0.3755**	*0.3192*	
48	44.69	46.2	0.0	46.2	0.2463	**0.3048**	*0.2486*	
50	14.57	15.6	0.0	15.6	0.1737	**0.2342**	*0.1780*	
52	4.76	5.2	0.0	5.2	0.1012	**0.1636**	*0.1073*	
54	1.74	1.8	0.0	1.8	0.0359	**0.0929**	*0.0367*	
56	0.64	0.6	0.0	0.6	-0.0289	**0.0223**	*-0.0339*	k_{ext}
57	0.32	0.3	0.0	0.3	-0.0739	**-0.0130**		
59	0.16	0.1	0.0	0.1	-0.1188	**-0.0836**		
61	0.08	0.0	0.0	0.0	-0.1637	**-0.1543**		

Extensions

- Right to left maximality
- Maximal repeats
- Markov model
- Errors

Thank you !


QUELQUES IRREDUCTIBLES ...
Vuillemimus (Cosmique honoraire)
Gandefrejus (esclave)
Camélix (druide)
Erlix
Vaxumix
Algorithmix (chef)
Roulembrix
Mirabelix
Sylvix
Agecanonix
Croarix (barde)
Superfrancix
Paralytix
légionnaire quipeuleplus

A basic scheme

$$
\begin{array}{ccc}
\mathcal{R} : f(n) = \sum \cdots & \rightarrow & f(n) \sim \cdots : \mathcal{R} \\
\downarrow & 1983 & \downarrow \\
\mathcal{C} : F(z) = \sum_n f(n) z^n & \rightarrow & f(n) \sim \cdots : (\textit{singularities!})\mathcal{C}
\end{array}
$$

Generating functions

Combinatorial object and a size : trees, words, ...
Generating functions :

$$\begin{aligned} F(z) &= \sum_n f(n) z^n \quad \textit{ordinary} \\ &= \sum_n f(n) \frac{z^n}{n!} \quad \textit{exponential} \end{aligned}$$

algebraic or probability

monovariate or multivariate

A systematic approach

$$\{f_n\}_{n\geq 1} \leftrightarrow F(z) = \sum_n f_n z^n$$

Induction: Recursive combinatorial properties

$$\Downarrow$$

Functional equation on $F(z)$

$$\Downarrow$$

Asymptotics :
n large: $f_n \sim \beta_n \rho^{-n}$
where ρ is the root of some algebraic equation.