Digital trees for DNA sequences

Brigitte CHAUVIN (Versailles)

in collaboration with Peggy CÉNAC (Univ. Bourgogne), Eric FEKETE, Stéphane GINOUILLAC, Nicolas POUYANNE (Versailles)

AofA08

Outline

- Introduction
- Tree representation
- Where randomness is
- What is known
- Results
- Methods

Introduction

- A DNA sequence is an infinite word

$$U = u_1 u_2 \dots u_n \dots \qquad \forall i, u_i \in \{A, C, G, T\}.$$

Introduction

- A DNA sequence is an infinite word

$$U = u_1 u_2 \dots u_n \dots \qquad \forall i, u_i \in \{A, C, G, T\}.$$

- To be seen on a representation:
  - repetition of patterns
  - missing patterns
  - repartition of different possible patterns
  - comparison of different sequences

Introduction

- A DNA sequence is an infinite word

$$U = u_1 u_2 \dots u_n \dots \qquad \forall i, u_i \in \{A, C, G, T\}.$$

- To be seen on a representation:
  - repetition of patterns
  - missing patterns
  - repartition of different possible patterns
  - comparison of different sequences
- Can we identify some characteristics
  - easy to study on the representation
  - different from a species to another species?

Tree representation

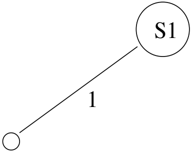
$$U = u_1 u_2 \ldots u_n \ldots$$

Prefixes	Rev.prefixes	Suffixes
u_1	u_1	$u_1 u_2 u_3 u_4 \ldots$
$u_1 u_2$	$u_2 u_1$	$u_2 u_3 u_4 \ldots$
$u_1 u_2 u_3$	$u_3 u_2 u_1$	$u_3 u_4 \ldots$
$\ldots$	$\ldots$	$\ldots$

- suffix trie
- DST of reversed prefixes
- trie of reversed prefixes
- suffix DST

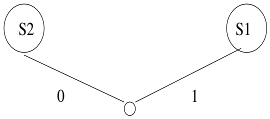
Example. Suffix trie. $U = 1001011001110\ldots$

$S_1 = U = 1001011001110\ldots$



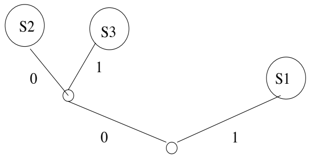
Example. Suffix trie. $U = 1001011001110\ldots$

$S_1 = U = 1001011001110\ldots$
$S_2 = 001011001110\ldots$



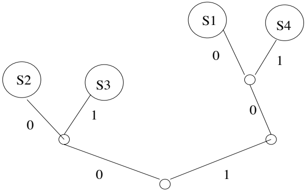
Example. Suffix trie. $U = 1001011001110\ldots$

$$S_1 = U = 1001011001110\ldots$$
$$S_2 = 001011001110\ldots$$
$$S_3 = 01011001110\ldots$$



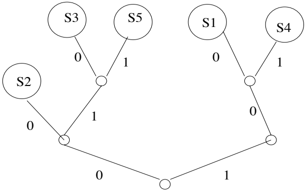
Example. Suffix trie. $U = 1001011001110\ldots$

$S_1 = U = 1001011001110\ldots$
$S_2 = 001011001110\ldots$
$S_3 = 01011001110\ldots$
$S_4 = 1011001110\ldots$



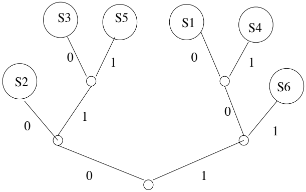
Example. Suffix trie. $U = 1001011001110\ldots$

$S_1 = U = 1001011001110\ldots$
$S_2 = 001011001110\ldots$
$S_3 = 01011001110\ldots$
$S_4 = 1011001110\ldots$
$S_5 = 011001110\ldots$



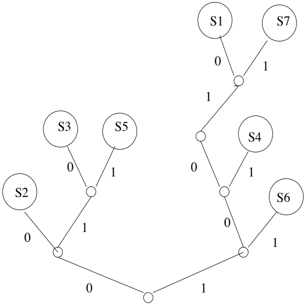
Example. Suffix trie. $U = 1001011001110\ldots$

$S_1 = U = 1001011001110\ldots$
$S_2 = 001011001110\ldots$
$S_3 = 01011001110\ldots$
$S_4 = 1011001110\ldots$
$S_5 = 011001110\ldots$
$S_6 = 11001110\ldots$



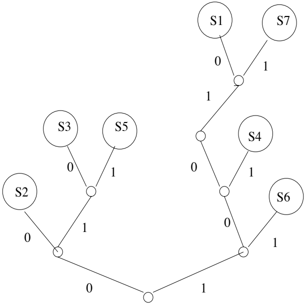
Example. Suffix trie. $U = 1001011001110\ldots$

$S_1 = U = 1001011001110\ldots$
$S_2 = 001011001110\ldots$
$S_3 = 01011001110\ldots$
$S_4 = 1011001110\ldots$
$S_5 = 011001110\ldots$
$S_6 = 11001110\ldots$
$S_7 = 1001110\ldots$



Example. Suffix trie. $U = 1001011001110\ldots$

$S_1 = U = 1001011001110\ldots$
$S_2 = 001011001110\ldots$
$S_3 = 01011001110\ldots$
$S_4 = 1011001110\ldots$
$S_5 = 011001110\ldots$
$S_6 = 11001110\ldots$
$S_7 = 1001110\ldots$



The shape of the tree is closely related to the repetitions of patterns

Where randomness is?

Comes from the production of the letters: $\{0, 1\}$ or $\{A, C, G, T\}$ or from any finite alphabet. For a given word $U = u_1 u_2 \ldots u_n \ldots$,

the tree process $(\mathcal{T}_n)_{n \geq 0}$ is nonrandom.

Where randomness is?

Comes from the production of the letters: $\{0, 1\}$ or $\{A, C, G, T\}$ or an alphabet. For a given word $U = u_1 u_2 \dots u_n \dots$,

the tree process $(\mathcal{T}_n)_{n \geq 0}$ is nonrandom.

Different kinds of sources:

- Memoryless: Bernoulli or asymmetric i.i.d.
- Markov
- Probabilistic dynamical source on an alphabet $\mathcal{A}$:
  - a partition of $[0, 1]$ with open intervals $\mathcal{I}_\alpha, \alpha \in \mathcal{A}$,
  - an encoding mapping $\sigma : [0, 1] \to \mathcal{A}$, s.t. $\sigma_{|\mathcal{I}_\alpha} \equiv \alpha$
  - a transformation T,
  - an initial density f.

Probabilistic dynamical source on an alphabet $\mathcal{A}$:

- a partition of $[0,1]$ with open intervals $\mathcal{I}_\alpha, \alpha \in \mathcal{A}$,
- an encoding mapping $\sigma : [0,1] \to \mathcal{A}$, s.t. $\sigma_{|\mathcal{I}_\alpha} \equiv \alpha$
- a transformation T,
- an initial density f.
    - x_1 is chosen on $[0,1]$ with the density f
    - its orbit is x_1, $T(x_1)$, $T^2(x_1), \ldots$
    - then $U = \sigma(x_1)\sigma(T(x_1))\sigma(T^2(x_1)) \cdots = u_1 u_2 \ldots$

The inserted words (suffixes or reversed prefixes) are NOT independent.

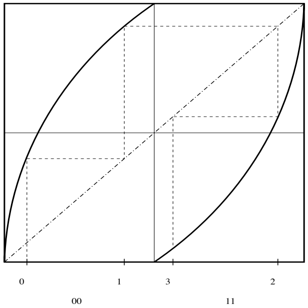


Figure: The shift mapping T

What is known

DST

for independent words

Bernoulli source

- height, insertion depth, profile

  cf. Mahmoud (92)
- $H_n - \log_2 n \xrightarrow{P} 0$

  Aldous-Shields (98)
- Concentration of the height

  Drmota (02)

Suffix tries

Bernoulli source

- size

  Blumer et al. (89)
- height

  Devroye, Szpankowski (92)
- mean, distrib. analysis

  Jacquet, Szpankowski

What is known

DST

for independent words

Bernoulli source

- height, insertion depth, profile

cf. Mahmoud (92)

- $H_n - \log_2 n \xrightarrow{P} 0$

Aldous-Shields (98)

- Concentration of the height

Drmota (02)

iid assymmetric, Markov source

- *Pittel (85)*

insertion depth, height

strong convergences

from an infinite word

- iid or Markov source

Cénac et al. (07)

Suffix tries

Bernoulli source

- size

Blumer et al. (89)

- height

Devroye, Szpankowski (92)

- mean, distrib. analysis

Jacquet, Szpankowski

iid assym., Markov

- average size and

total path length

Fayolle (06)

dynamical source

- *Cénac, Fekete*

(in progress)

Two families of methods:

(1)	(2)
analytic combinatorics generating functions Mellin transform	probability
↓	↓
precise asymptotics on - the average of additive characteristics - distribution of the height	a.s. convergences

Some notations to write the results

- The probability that the source produces a sequence of symbols starting with the pattern m is

$$p_m = \int_{\mathcal{I}_m} f(t)dt.$$

- $s = s_1 s_2 \ldots s_n \ldots$ denotes an infinite deterministic sequence.
- $s^{(n)} = s_1 s_2 \ldots s_n$.

Some notations to write the results

- $$p_m = \int_{\mathcal{I}_m} f(t)dt$$
- $s = s_1 s_2 \ldots s_n \ldots$ denotes an infinite deterministic sequence.
- $s^{(n)} = s_1 s_2 \ldots s_n$.
- Entropies

$$h_+ = \lim_{n \to +\infty} \frac{1}{n} \max_{s^{(n)}} \left\{ \ln \left(\frac{1}{p_{s^{(n)}}} \right) \right\},$$

$$h_- = \lim_{n \to +\infty} \frac{1}{n} \min_{s^{(n)}} \left\{ \ln \left(\frac{1}{p_{s^{(n)}}} \right) \right\},$$

$$h = \lim_{n \to +\infty} \frac{1}{n} E \left[\ln \left(\frac{1}{p\left(U^{(n)}\right)} \right) \right].$$

Some notations to write the results

- $s = s_1 s_2 \dots s_n \dots$ denotes an infinite deterministic sequence.
  $s^{(n)} = s_1 s_2 \dots s_n$.
-
$$h_+ = \lim_{n \to +\infty} \frac{1}{n} \max_{s^{(n)}} \Big\{ \ln \Big(\frac{1}{p_{s^{(n)}}} \Big) \Big\}, h_- = \lim_{n \to +\infty} \frac{1}{n} \min_{s^{(n)}} \Big\{ \ln \Big(\frac{1}{p_{s^{(n)}}} \Big) \Big\},$$
$$h = \lim_{n \to +\infty} \frac{1}{n} E \Big[\ln \Big(\frac{1}{p\big(U^{(n)}\big)} \Big) \Big].$$
- ℓ_n = length shortest branch of the tree = fill-up level
  $\mathcal{L}_n$ = length of the longest branch of the tree.
  D_n = insertion depth

Results

ℓ_n = length shortest branch of the tree = fill-up level
$\mathcal{L}_n$ = length of the longest branch of the tree.
D_n = insertion depth

Theorem

(Cénac et al. (07))
For the DST for a memoryless source or a Markovian source

$$\frac{\ell_n}{\ln n} \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{h_+}, \quad \textit{and} \quad \frac{\mathcal{L}_n}{\ln n} \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{h_-}.$$

Results

ℓ_n = length shortest branch of the tree = fill-up level
$\mathcal{L}_n$ = length of the longest branch of the tree.
D_n = insertion depth

Theorem

For the DST for a memoryless source or a Markovian source

$$\frac{\ell_n}{\ln n} \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{h_+}, \quad \text{and} \quad \frac{\mathcal{L}_n}{\ln n} \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{h_-}.$$

$$\frac{D_n}{\ln n} \xrightarrow[n\to\infty]{\text{P}} \frac{1}{h}$$

Results

ℓ_n = length shortest branch of the tree = fill-up level
$\mathcal{L}_n$ = length of the longest branch of the tree.
D_n = insertion depth

Theorem

For the DST for a memoryless source or a Markovian source

$$\frac{\ell_n}{\ln n} \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{h_+}, \quad \textit{and} \quad \frac{\mathcal{L}_n}{\ln n} \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{h_-}.$$

$$\frac{D_n}{\ln n} \xrightarrow[n\to\infty]{\text{P}} \frac{1}{h}$$

For the suffix trie for a dynamical source with a ϕ−mixing condition

$$\frac{\ell_n}{\ln n} \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{h_+}.$$

Methods - 1 - Runs well

- $s = s_1 s_2 \dots s_n \dots$ denotes an infinite deterministic sequence.
- $s^{(n)} = s_1 s_2 \dots s_n$

$X_n(s) \stackrel{\text{def}}{=}$ length of the branch corresponding to s in the tree $\mathcal{T}_n$

$$\ell_n = \min_s X_n(s) \quad \text{and} \quad \mathcal{L}_n = \max_s X_n(s).$$

Methods - 1 - Runs well

- $s = s_1 s_2 \ldots s_n \ldots$ denotes an infinite deterministic sequence.
- $s^{(n)} = s_1 s_2 \ldots s_n$
- $T_k(s) \stackrel{\text{def}}{=}$ size of the first tree where is inserted $s^{(k)}$,
  $X_n(s) \stackrel{\text{def}}{=}$ length of the branch corresponding to s in $\mathcal{T}_n$.

$$\ell_n = \min_s X_n(s) \quad \text{and} \quad \mathcal{L}_n = \max_s X_n(s).$$

- X_n and T_k are in duality

$$\{X_n(s) \geq k\} = \{T_k(s) \leq n\}.$$

$$P(\ell_n \leq k-1) \leq \sum_{s^{(k)}} P(T_k(s) > n) = \sum_{s^{(k)}} P(t^0_{s^{(k)}} + t^1_{s^{(k)}} > n)$$

Methods - 1 - Runs well

- $T_k(s) \stackrel{\text{def}}{=}$ size of the first tree where is inserted $s^{(k)}$,

$$\ell_n = \min_s X_n(s)$$

$$P(\ell_n \leq k-1) \leq \sum_{s^{(k)}} P(T_k(s) > n) = \sum_{s^{(k)}} P(t^0_{s^{(k)}} + t^1_{s^{(k)}} > n)$$

where (for the suffix trie)

$$t^0_m = \text{hitting time of pattern } m$$
$$t^1_m = \text{return time of pattern } m.$$

- sufficient:

$$\sum_{s^{(k)}} P(t^0_{s^{(k)}} > n/2) \text{ is the g.t. of a conv. series}$$

$$\sum_{s^{(k)}} P(t^1_{s^{(k)}} > n/2) \text{ is the g.t. of a conv. series}$$

Methods - 1 - Runs well

$$t_m^0 = \text{hitting time of pattern } m$$
$$t_m^1 = \text{return time of pattern } m.$$

It is sufficient to prove

$$\sum_{s^{(k)}} P(t_{s^{(k)}}^0 > n/2) \text{ is the g.t. of a conv. series}$$

$$\sum_{s^{(k)}} P(t_{s^{(k)}}^1 > n/2) \text{ is the g.t. of a conv. series}$$

Methods - 1 - Runs well

$t_m^0 =$ hitting time of pattern m
$t_m^1 =$ return time of pattern m.

To prove:

$$\sum_{s^{(k)}} P(t^0_{s^{(k)}} > n/2) \text{ is the g.t. of a conv. series}$$

$$\boxed{\sum_{s^{(k)}} P(t^1_{s^{(k)}} > n/2) \text{ is the g.t. of a conv. series}}$$

$\uparrow$ for a pattern m

$$\boxed{|P(t_m^1 > t) - Ce^{-\xi_m t}| \leq C' t^{\beta}}$$

$\sim$ *Galves-Schmidt (97)*

Methods - 2 - Less easy

The more auto-correlated a word is, the more easily it may reappear and the smaller its return time is.

Methods - 2 - Less easy

The more auto-correlated a word is, the more easily it may reappear and the smaller its return time is.

To achieve this

(1)
work on the assumptions
add independence
↓
Bernoulli
Markov
dynamical source + mixing assumptions .

(2)
tools
auto-correlation polynomials

Meaning of such mixing conditions:
When two parts of a word

$$w = \dots w_0 | w_1 w_2 \dots w_n | w_{n+1} \dots$$

are far (more than n letters) from each other, then, these two parts are "almost" independent.

The mixing assumptions

Assumptions on the geometry of the branches of the dynamical system (T, f):
- branches of class C^2
- bounded distorsion of the branches

$\downarrow$

weak $\phi-$mixing condition (*Paccaut (99)*):
μ stationary measure, $\exists C, \exists \xi \in]0,1[$ s.t. $\forall P, Q$ borelians in $[0,1]$,

$$|\mu(P \cap T^{-n}Q) - \mu(P)\mu(Q)| \leq C\xi^n \mu(Q)$$

$\uparrow$

$\phi-$mixing condition (*Galves-Schmidt (97)*):
$\exists \phi$ decreasing, positive, tending to 0 s.t.

$$\sup_{P \in \mathcal{F}_n, Q} \frac{\mu(P \cap T^{-(n+l)}Q) - \mu(P)\mu(Q)}{\mu(P)\mu(Q)} \leq \phi(l)$$

$\phi-$mixing condition (*Galves-Schmidt (97)*):
$\exists\phi$ decreasing, positive, tending to 0 s.t.

$$\sup_{P\in\mathcal{F}_n,Q}\frac{\mu(P\cap T^{-(n+l)}Q)-\mu(P)\mu(Q)}{\mu(P)\mu(Q)}\leq\phi(l)$$

$$\downarrow$$

$$\downarrow$$

$$|P(t_m^1>t)-Ce^{-\xi_m t}|\leq C't^{\beta}$$

Still to do

Still to do

- convergence rates
- central limit theorem

Still to do

- convergence rates
- central limit theorem
- mixing conditions

Still to do

- convergence rates
- central limit theorem
- mixing conditions

- statistical point of view

to be continued...