Arbres digitaux et suites d'ADN

Brigitte CHAUVIN (Versailles)

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

INRIA, 26 mai 2008

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?
- objectifs : distance entre les espèces, stat

Tree representation

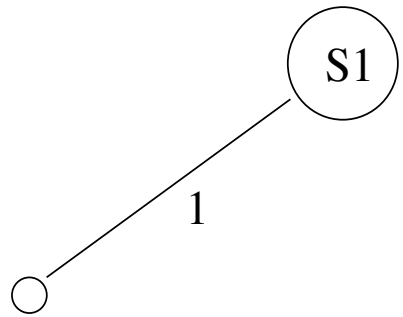
$$U = u_1 u_2 \dots u_n \dots$$

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

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

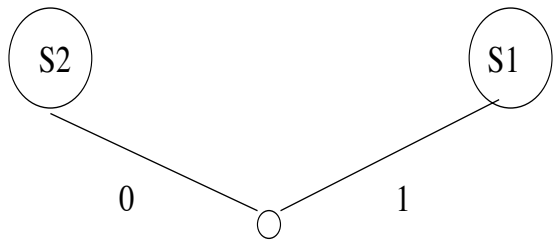
Example. Suffix trie. $U = 10010110011110\ldots$

$S_1 = U = 10010110011110\ldots$



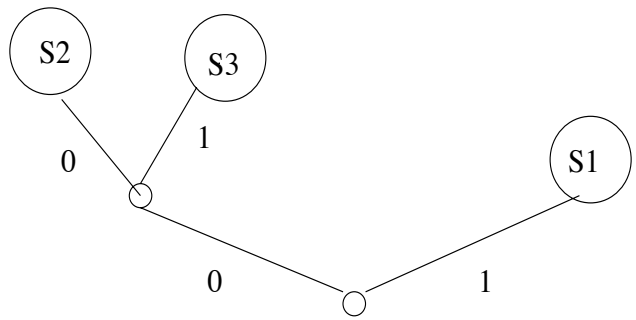
Example. Suffix trie. $U = 10010110011110\ldots$

$S_1 = U = 10010110011110\ldots$
$S_2 = 0010110011110\ldots$



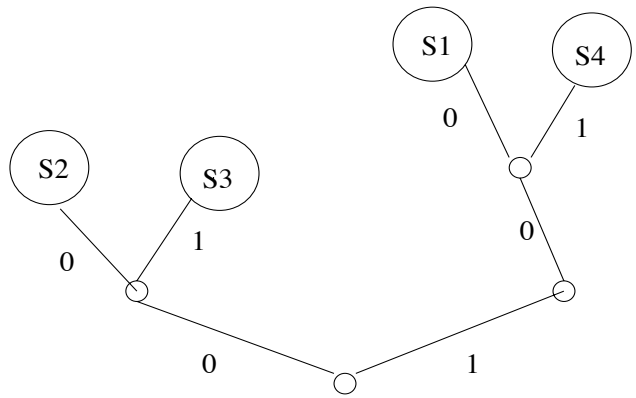
Example. Suffix trie. $U = 10010110011110\ldots$

$S_1 = U = 10010110011110\ldots$
$S_2 = 0010110011110\ldots$
$S_3 = 010110011110\ldots$



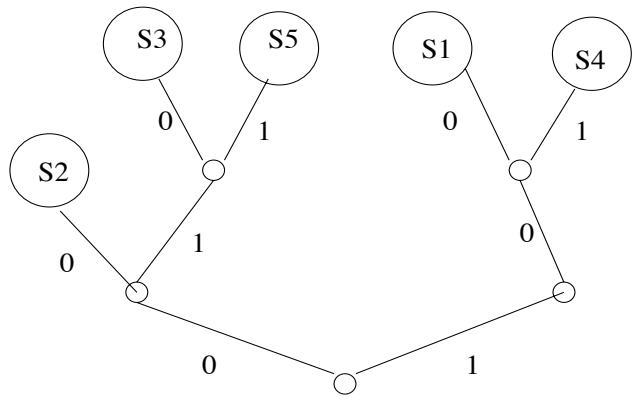
Example. Suffix trie. $U = 10010110011110\ldots$

$S_1 = U = 10010110011110\ldots$
$S_2 = 0010110011110\ldots$
$S_3 = 010110011110\ldots$
$S_4 = 10110011110\ldots$



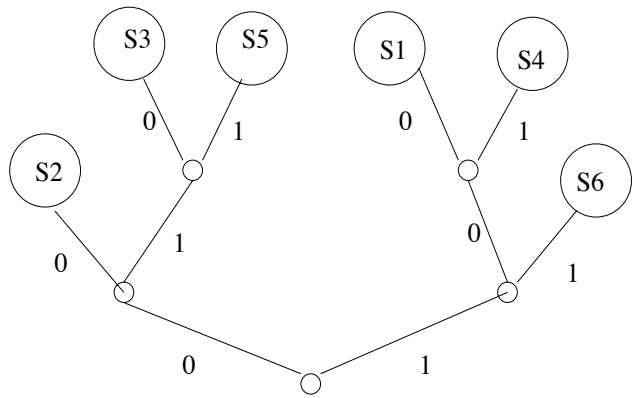
Example. Suffix trie. $U = 10010110011110\ldots$

$S_1 = U = 10010110011110\ldots$
$S_2 = 0010110011110\ldots$
$S_3 = 0101100111 0\ldots$
$S_4 = 1011001110\ldots$
$S_5 = 0110011 10\ldots$



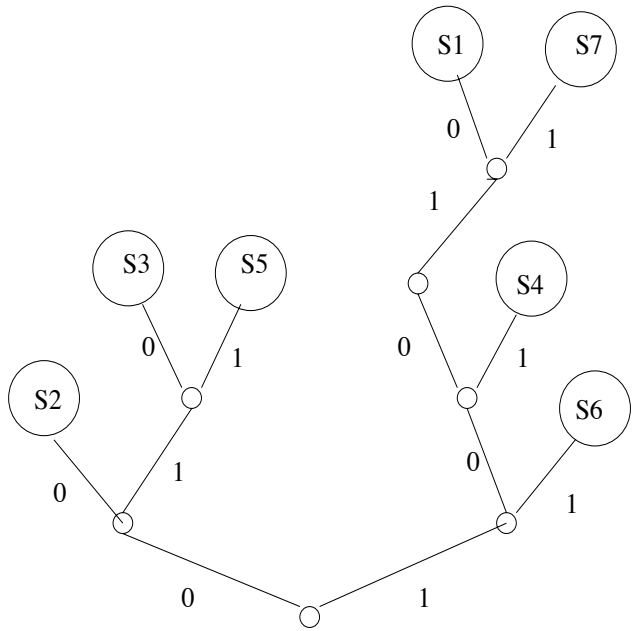
Example. Suffix trie. $U = 10010110011110\ldots$

$S_1 = U = 10010110011110\ldots$
$S_2 = 0010110011110\ldots$
$S_3 = 0101100111 10\ldots$
$S_4 = 1011001110\ldots$
$S_5 = 0110011 10\ldots$
$S_6 = 11001110\ldots$



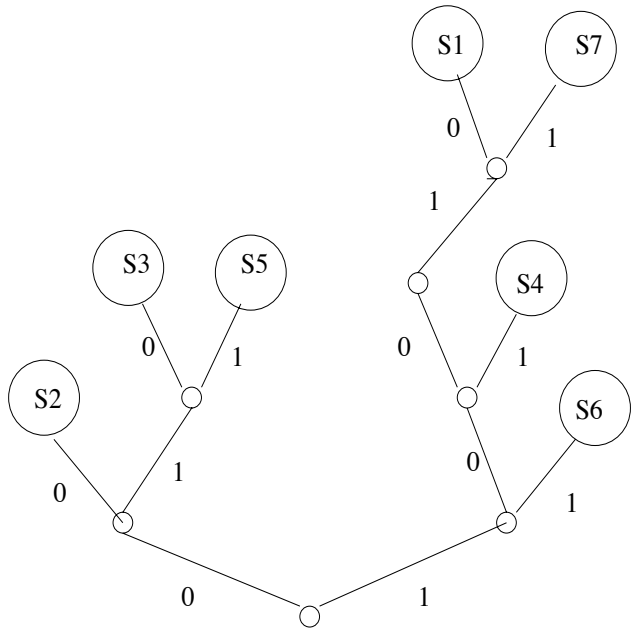
Example. Suffix trie. $U = 10010110011110\ldots$

$S_1 = U = 10010110011110\ldots$
$S_2 = 0010110011110\ldots$
$S_3 = 010110011110\ldots$
$S_4 = 10110011110\ldots$
$S_5 = 0110011110\ldots$
$S_6 = 110011110\ldots$
$S_7 = 10011110\ldots$



Example. Suffix trie. $U = 10010110011110\ldots$

$S_1 = U = 10010110011110\ldots$
$S_2 = 0010110011110\ldots$
$S_3 = 010110011110\ldots$
$S_4 = 10110011110\ldots$
$S_5 = 0110011110\ldots$
$S_6 = 110011110\ldots$
$S_7 = 10011110\ldots$



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 $U = u_1 u_2 \dots u_n \dots$,

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
- General probabilistic source
  - choose an infinite word $U = u_1 u_2 \dots u_n \dots$ with distribution μ
  - call T the shift,
  - add mixing assumptions (later).

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

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)

What is known

Suffix tries

- height

Devroye, Szpankowski (92) (i.i.d. source)

- depth, fill-up level, height

Jacquet, Szpankowski (93) (general source + mixing)

- average size and total path length

Fayolle (06) (iid assym., Markov)

- fill-up level

Cénac, Fekete (general source + not too strong mixing)
(in progress)

Two families of methods:

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

common: correlations, overlapping of words

Some notations to write the results

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

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

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

Some notations to write the results

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

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

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

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

Results

ℓ_n = 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

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

Results

$\ell_n =$ 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

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

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

In progress

$\ell_n =$ fill-up level

$\mathcal{L}_n =$ length of the longest branch of the tree.

$D_n =$ insertion depth

Theorem

For the suffix trie for a general source with mixing conditions

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

Methods - 1 - Runs well

(works for the DST and for the suffix trie)

- $s = s_1 s_2 \dots s_n \dots$ 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

(works for the DST and for the suffix trie)

- $s = s_1 s_2 \dots s_n \dots$ denotes an infinite deterministic sequence.
- $s^{(n)} = s_1 s_2 \dots s_n$
- $T_k(s) \stackrel{\text{def}}{=}$ 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 $\{X_n(s) \geq k\} = \{T_k(s) \leq n\}$.

$$P(\ell_n \leq k-1) \leq \dots$$

Methods - 1 - Runs well

(works for the DST and for the suffix trie)

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

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

$$P(\ell_n \leq k-1) \leq \sum_{s^{(k)}} P(T_k(s) > n)$$

Methods - 1 - Runs well

(works for the DST)

$$P(\ell_n \leq k-1) \leq \sum_{s^{(k)}} P(T_k(s) > n)$$

- $s = s_1 s_2 \dots s_n \dots$ denotes an infinite deterministic sequence.
- $s^{(n)} = s_1 s_2 \dots s_n$
- $T_k(s) \stackrel{\text{def}}{=}$ size of the first tree where is inserted $s^{(k)}$,

$$T_k(s) = \sum_{r=1}^{k} T_r(s) - T_{r-1}(s) = \sum_{r=1}^{k} Z_r(s)$$

$Z_r(s) =$ waiting time of the first occurrence of $s^{(r)}$ in U after T_{r-1}

Hyp Markov $\Rightarrow$ *the r. v.* $Z_r(s)$ *are independent*

Un peu de proba

(works for the DST)

Hyp Markov $\Rightarrow$ *the r. v.* $Z_r(s)$ *are independent*

$$
\begin{aligned}
T_k(s) &= \sum_{r=1}^{k} T_r(s) - T_{r-1}(s) = \sum_{r=1}^{k} Z_r(s) \\
&= \sum_{r=1}^{k} \left[Z_r(s) - \mathbb{E} Z_r(s) \right] + \sum_{r=1}^{k} \mathbb{E} Z_r(s) \\
&= \sum_{r=1}^{k} \epsilon_r(s) + \mathbb{E}\, T_k(s) \\
&= \text{martingale } M_k(s) + \mathbb{E}\, T_k(s)
\end{aligned}
$$

$$
\begin{aligned}
T_k(s) = \sum_{r=1}^{k} Z_r(s) &= \sum_{r=1}^{k} \left[Z_r(s) - \mathbb{E} Z_r(s) \right] + \sum_{r=1}^{k} \mathbb{E} Z_r(s) \\
&= \sum_{r=1}^{k} \epsilon_r(s) + \mathbb{E}\, T_k(s) \\
&= \text{martingale } M_k(s) + \mathbb{E}\, T_k(s)
\end{aligned}
$$

$$
\begin{aligned}
\log T_k(s) &= \log \mathbb{E}\, T_k(s) + \log \left(1 + \frac{M_k(s)}{\mathbb{E}\, T_k(s)}\right) \\
&\sim kh(s) \quad + \quad \downarrow
\end{aligned}
$$

$$
\forall \alpha > 0, \frac{M_k(s)}{\mathbb{E}\, T_k(s)} = o(k^{1+\alpha/2})
$$

$$
\frac{\log T_k(s)}{k} \xrightarrow[k \to \infty]{\text{a.s.}} h(s)
$$

Un peu de proba

(works for the DST)

Hyp Markov $\Rightarrow$ *the r. v.* $Z_r(s)$ *are independent*

$$\begin{aligned} P(\ell_n \leq k-1) &\leq \sum_{s^{(k)}} P(T_k(s) > n) \\ &\leq \sum_{s^{(k)}} t^{-n} \mathbb{E}(t^{T_k(s)}), \qquad t > 1 \end{aligned}$$

$T_k(s) = \sum_{r=1}^{k} Z_r(s)$

$$\mathbb{E}(t^{T_k(s)}) = \prod_{r=1}^{k} \mathbb{E}(t^{Z_r(s)}) \qquad \textit{Daudin-Robin (99)}$$

(for the suffix trie)

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

(works for the suffix trie)

- $T_k(s) \stackrel{\text{def}}{=}$ 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

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

- sufficient:

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

$\sum_{s^{(k)}} P(t^1_{s^{(k)}} > n/2)$ 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.

It is sufficient to prove

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

$\sum_{s^{(k)}} P(t^1_{s^{(k)}} > n/2)$ 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:

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

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

$\uparrow$ for a pattern m

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

$\sim$ *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

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

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

μ stationary, ergodic measure is the distribution of the words.
T is the shift (or the transformation in a dynamical system)
A is a word depending on the first m letters
B is a word depending on the suffix after $m + n$.

mixing

$$\lim_{n\to\infty} \mu(A \cap T^{-n}B) - \mu(A)\mu(B) = 0.$$

$\uparrow$

$\phi-$mixing (*Paccaut (99)*):
$\exists \phi \to 0$, s.t.

$$|\mu(A \cap T^{-n}B) - \mu(A)\mu(B)| \leq \phi(n)\mu(B).$$

$\uparrow$

$\psi-$mixing (*Szpankowski (93), Galves-Schmidt (97)*):
$\exists \psi$ decreasing, positive, tending to 0 s.t.

$$\mu(A \cap T^{-n}B) - \mu(A)\mu(B) \leq \psi(n)\mu(A)\mu(B)$$

Questions

Le cas particulier des systèmes dynamiques apporte-t-il quelque chose dans l'utilisation des conditions de mixing ?

Que se passe-t-il si on ne met pas ou peu d'hypothèse de mixing ?