Statistical Properties of Oracles

Jérémie Bourdon[1,2], Irena Rusu[1]
[1] LINA, Université de Nantes, France
[2] IRISA, INRIA Rennes Bretagne Atlantique

Combinatorial Pattern Matching, june 2009

Outline

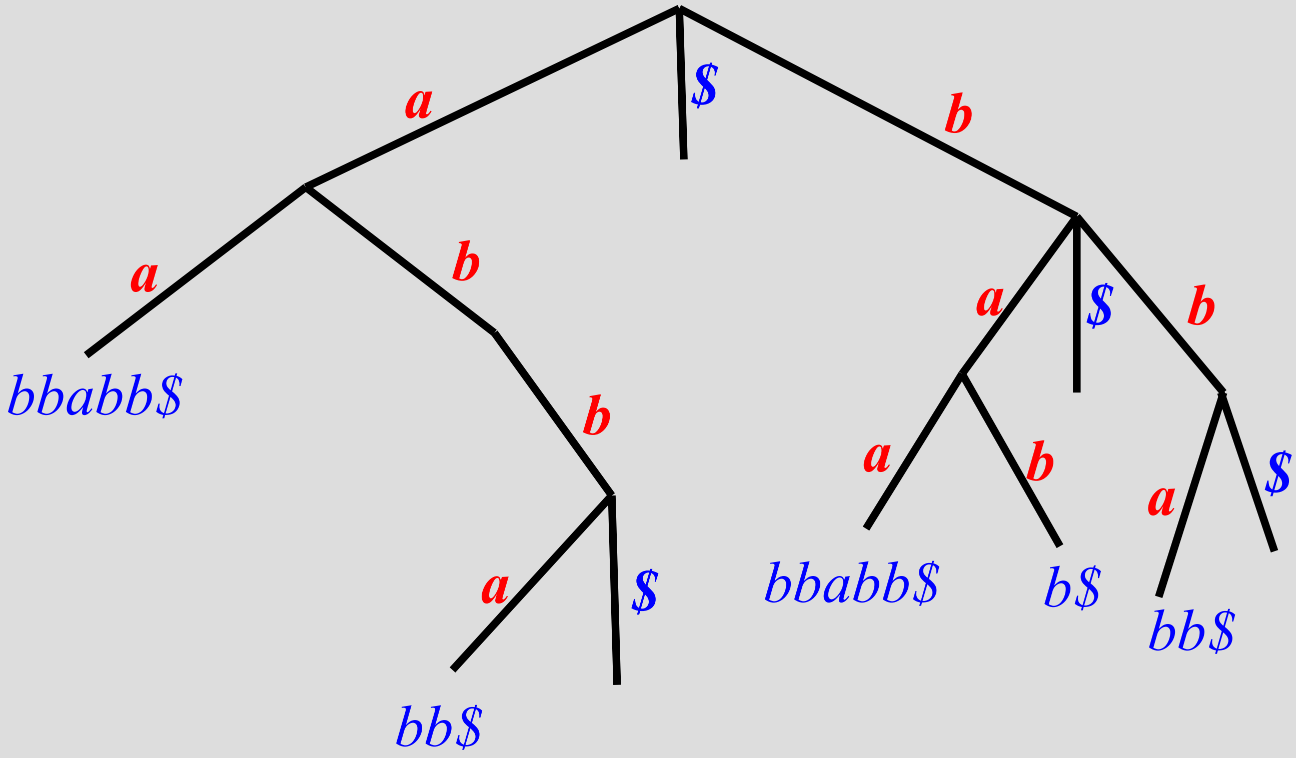
- Suffixes/Factor Oracles : Définition and main properties
- External transition probabilities
- Application : memory size of an oracle
- Conclusion

Index structures

- Problem : design a data structure for storing and retrieving efficiently all the factor/suffixes of a given word
- 3 criteria:
  - Small structure (in a classical use)
  - Store all the factors
  - Store only the factors

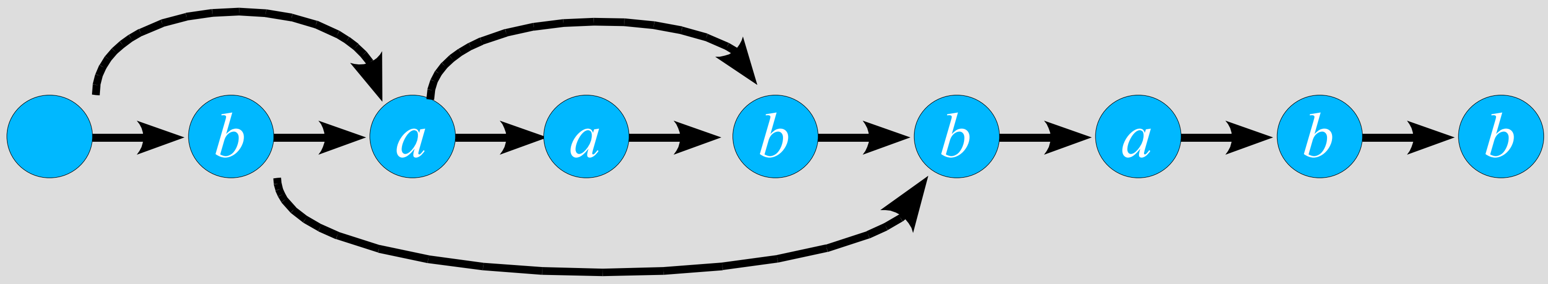
Example 1 : suffix trie

- *baabbabb$*



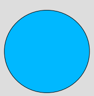
Example 2 : factor oracle [Allauzen & al., 1999]

- *baabbabb*



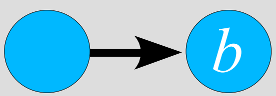
Example 2 : factor oracle

- *baabbabb*



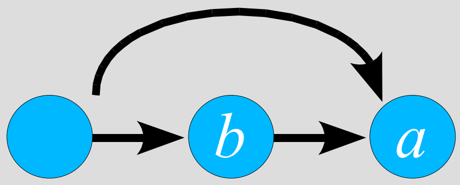
Example 2 : factor oracle

- *b*aabbabb



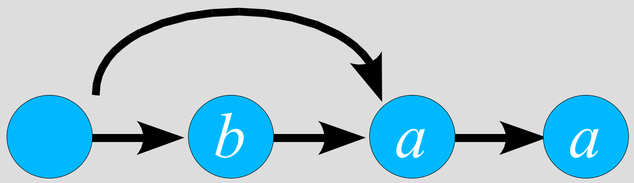
Example 2 : factor oracle

- *ba*abbabb



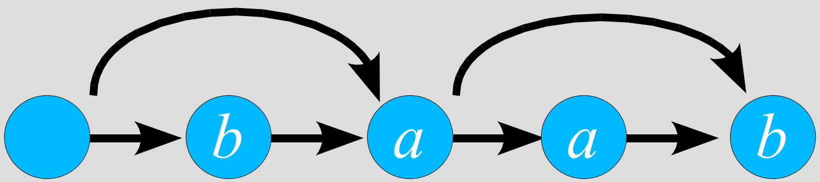
Example 2 : factor oracle

- *baa*bbabb



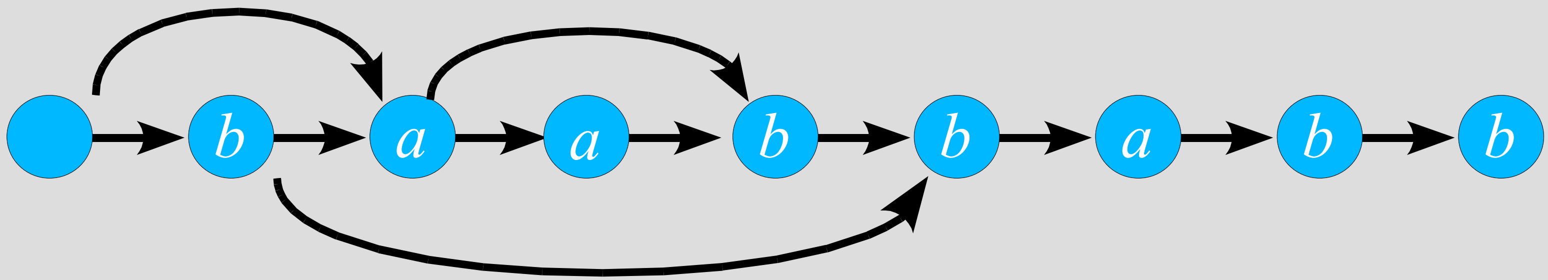
Example 2 : factor oracle

- *baab*babb



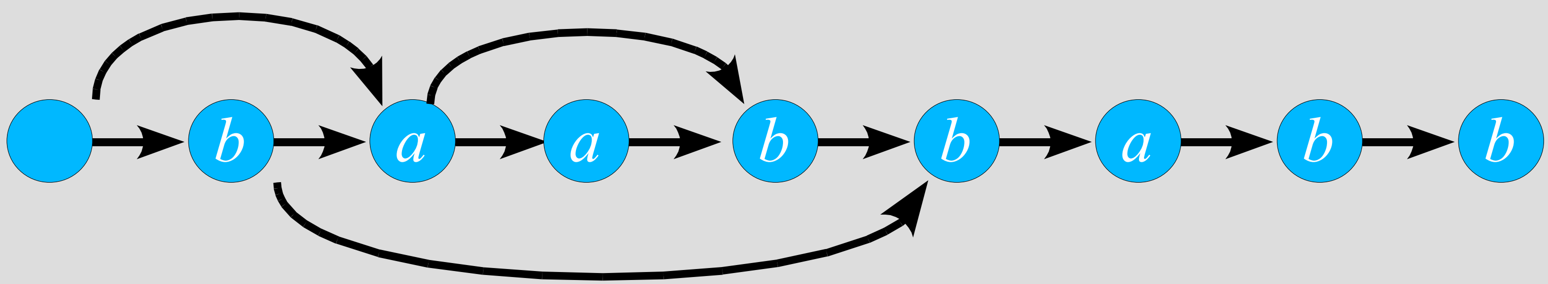
Example 2 : factor oracle

- *baabbabb*



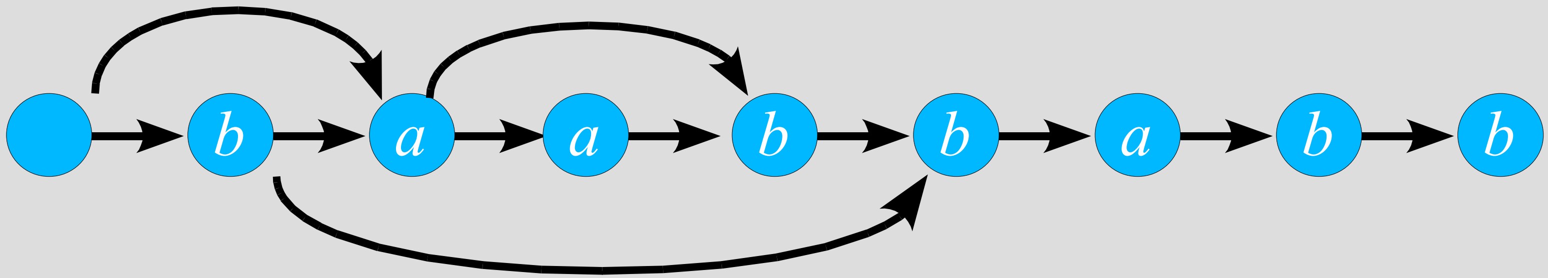
Only the factors ?

- *baabbabb*



- *No ! babba is not a factor !*

An important property



- *min(p)*=*(unique) word recognized in state |p| having minimal length*
- *There exists an external transition between state i=|p| and j=|q.m| iff p is a prefix of q and the first occurrence of min(p).m ends in state j.*

Oracles are useful ?

- as structures indexes ?
- Pattern matching (BOM : if a word is not in the oracle, it is not a factor)
- Musical improvisation (OMAX : by-products are slight variants of factors that can reproduce a musical sequence with errors)
- .....

Crucial point : what amount of memory is needed to store an oracle ?

Outline

- Suffixes/Factor Oracles : Définition and main properties
- External transition probabilities
- Application : memory size of an oracle
- Conclusion

Average size of an oracle ?

- Average size => which probabilistic model ?
- Each symbols drawn independently with prob. 1/2.

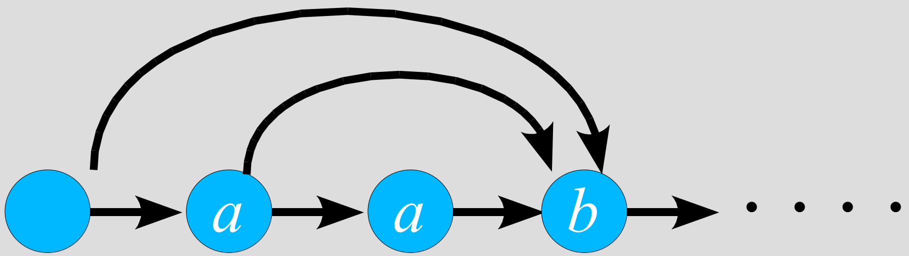
Hence $prob\{W\}=1/2^n$

- if W is a random word of size n on $\{a,b\}$, what is the value of $E[size(oracle(W))]$ *as a function of* n ?
- *Sub problem : What is the probability p(i) that an external transition leaves state i ?*
- $E[size(oracle(W))]=\Sigma\, p(i)$

Simple case n°1 : p(0)

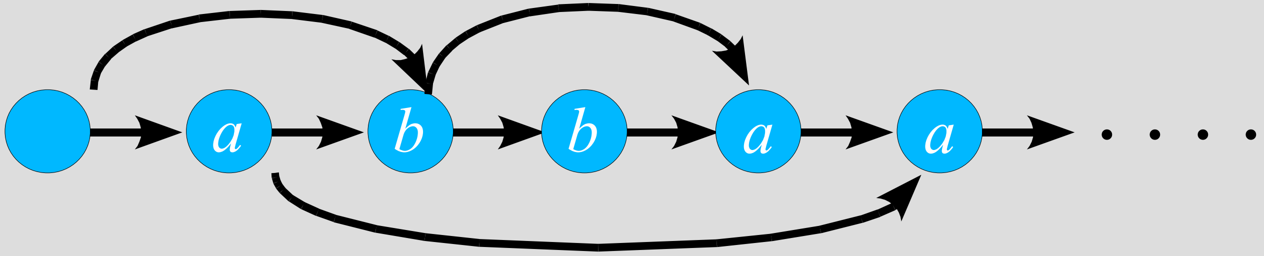
- There exists an external transition leaving state 0 if W is of the form $a...abX$, *where* X is any word.

$$p(0) = 2\sum_{j=2}^{n} \frac{1}{2^j} = 1 - \frac{1}{2^{n-1}}$$



Simple case n°2 : p(1)

- If W begins with *aa,* same case as *p(0)*
- If W begins with *ab*, there is a transition between state 1 and the end of the first occurrence of *aa*. *I.e.*, If W is of the form *ab(b+ab)*aaX*, where *X* is any word.



(b+ab)* and Fibonacci

- $L_0 = \{\varepsilon\}$
- $L_1 = \{b\}$
- $L_2 = \{ab, bb\}$
- $L_3 = \{bab, abb, bbb\}$
- $L_4 = \{abab, bbab, babb, abbb, bbbb\}$
- $L_k = L_{k-2}.ab \cup L_{k-1}.b$

$$p(1) = 1 - \frac{F_{n-1}+1}{2^{n-1}}$$

General case

Let w be a given word.

$S(w)=\{s\in\{a,b\}^*, w \text{ is not a factor of } s\}$

$T(w)=\{s\in\{a,b\}^*, w \text{ occurs once as a suffix of } s\}$

$C(w)=\{s\in\{a,b\}^*, \exists u,v\in\{a,b\}^*, w=s.u=v.s\}$

Related by [Guibas-Odlyzko, 80]

$S(w).\Sigma + \{\varepsilon\} = S(w)+T(w)$

$S(w).w=T(w).C(w)$

Relation with oracles

Main result : *There exists an external transition from state* $|p|$ *to state* $|p|+j+1$ *in the oracle associated to* W *iff* W *is a word of the set*

$$p.m' \,.\, [(D(min(p).m) + T(min(p).m)) \cap \Sigma^j] .\, \Sigma^*,$$

where m' is the opposite letter of m *and*

$D(q.m)=\{s \in \{a,b\}^*,\ \exists u,v \in \{a,b\}^*,\ q.m=s.u \text{ and } q.m'=v.s\}$

Computing p(i)

- Easy to compute the probability that a random word is in *S(w), T(w) or C(w)* for a given word *w, [Regnier et al. 95]*
- Sum over all words p of size i
- Group according to the length of the minimal word.
- neglict the possible correlations

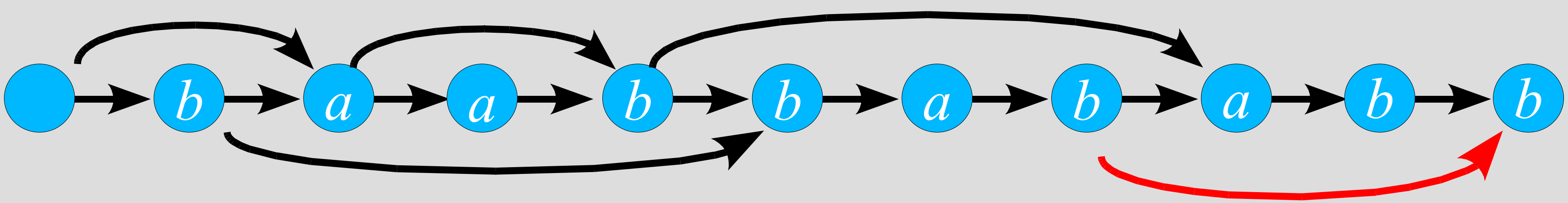
$$p(i) = 1 - \sum_{k=1}^{i} Prob[Min_i = k] \left(1 + \frac{1}{2^{k+1}}\right)^{i-n} + O(2^{i-n+1})$$

Computing Prob[Min_i=k]

- Let short(u) be the shortest non repeated suffix of u.
- Fact 1: min(u) is a non repeated suffix of u, then $|min(u)| \geq |short(u)|$
- Construct the short-oracle of W, built by using the rule « *There exists an external transition between i=|p| and j=|q.m| iff p is a prefix of q and the first occurrence of short(p).m ends in state j.* »

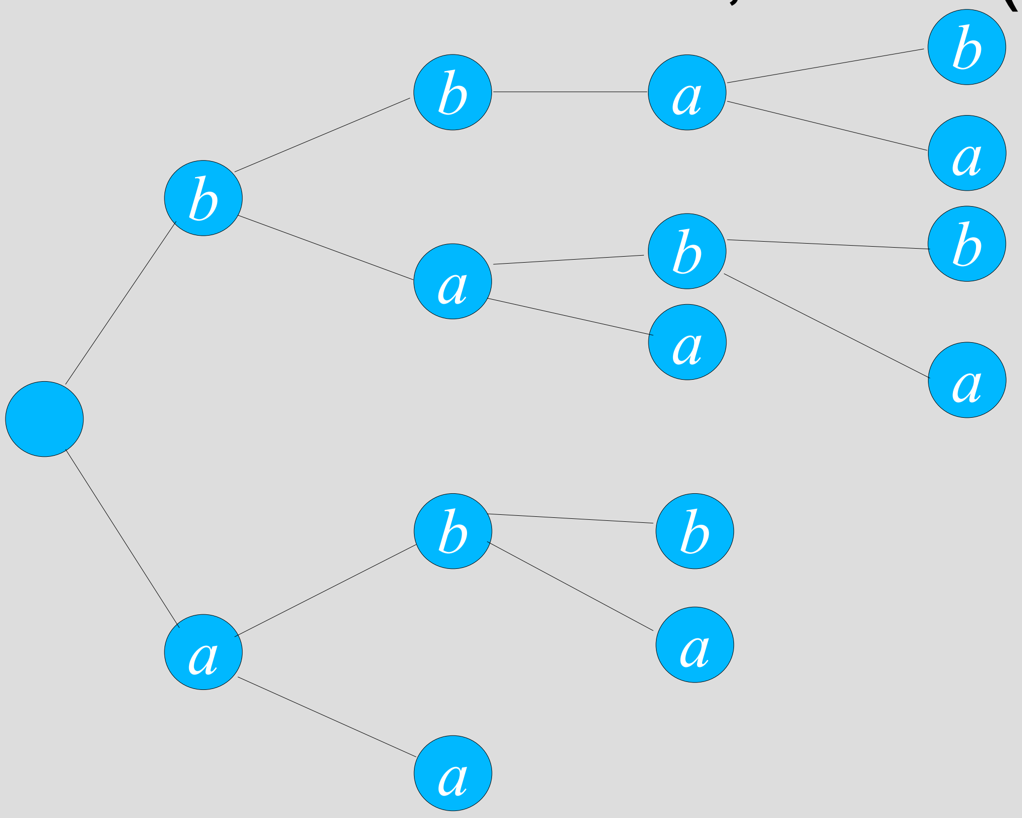
Comparison of oracles

- « short » version of oracles contains more external transitions and recognize more word than the classical version.



Computing Prob[Short$_i$=k]

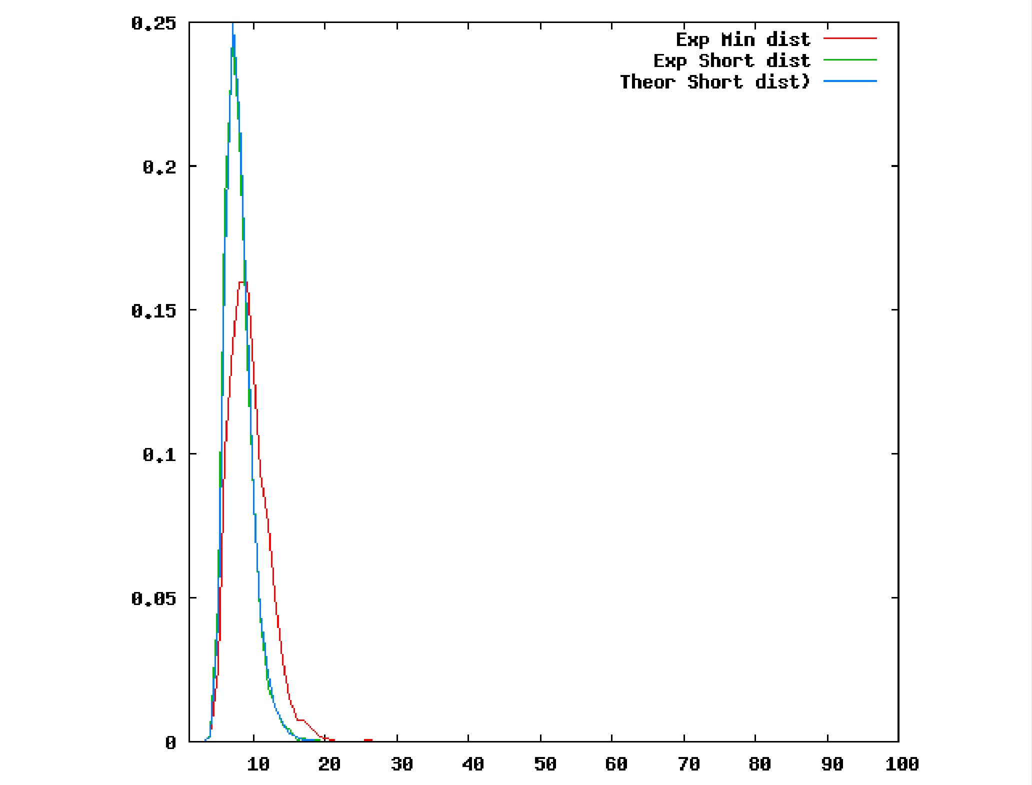
- Short(u) is related to the insertion depth in a suffix trie.
- u=*baabbababb*, mirror(u)=*bbababbaab*



Computing Prob[Short$_i$=k]

- [Park et al. 2008]

$$Prob[Short_i = k] = (1-2^{-k})^{n-1} - (1-2^{-k+1})^{n-1}$$



Outline

- Suffixes/Factor Oracles : Définition and main properties
- External transition probabilities
- Application : memory size of an oracle
- Conclusion and perspectives

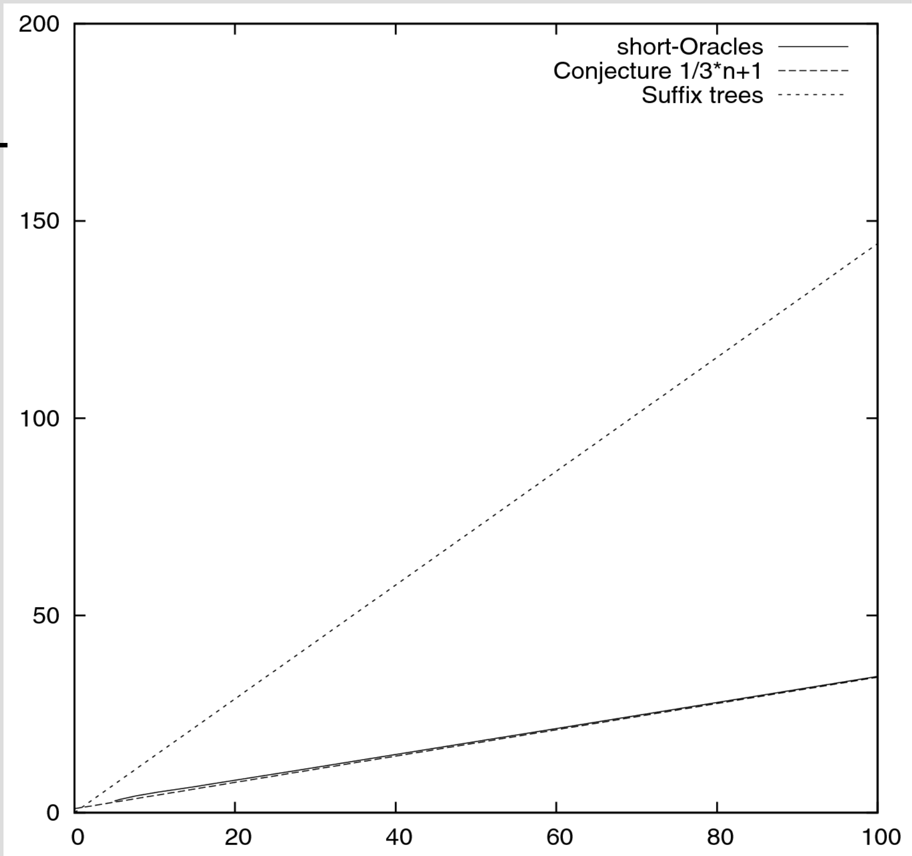
Average space occupancy

$$E[S_n] \leqslant n - \sum_{k=0}^{n} f_n(k,k) - f_n(k-1,k),$$

$$f_n(m,k) = \frac{\gamma_m^{k-1}\lambda_{k+1}^{k-n} - \gamma_m^{n-2}\lambda_{k+1}^{-1}}{1-\gamma_m\lambda_{k+1}}$$

$$\gamma_m = 1 - \frac{1}{2^m}, \lambda_m = 1 + \frac{1}{2^m}$$

$$E[T_n] \approx n/\log 2 \ \ \textit{internal nodes}$$



Conclusion

- Conjecture of n/3 for the number of external transitions of short-oracles (tractable)
- Extension to non binary alphabet (easy)
- Extension to biased probabilistic model of word creation (tractable+boring)
- Study the number of by-products if the short oracle (tractable if better approximations)
- Compute Prob[Min_i=k] (hard)
- Study of compacted versions (~easy)

Thank you !