HAL
open science

Autocorrelation on words and its applications. Analysis of suffix trees by string-ruler approach

Philippe Jacquet, Wojciec Szpankowski

To cite this version:

Philippe Jacquet, Wojciec Szpankowski. Autocorrelation on words and its applications. Analysis of suffix trees by string-ruler approach. RR-1106, INRIA. 1989. inria-00075453

HAL Id: inria-00075453

https://inria.hal.science/inria-00075453v1

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNITE DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 1106

Programme 1
Programmation, Calcul Symbolique
et Intelligence Artificielle

AUTOCORRELATION ON WORDS AND ITS APPLICATIONS

Analysis of suffix trees by string-ruler approach

Philppe JACQUET
Wojciech SZPANKOWSKI

Octobre 1989

RR_1106

AUTOCORRELATION ON WORDS AND ITS APPLICATIONS
Analysis of suffix trees by string-ruler approach

Philippe Jacquet†
INRIA
Domaine de Voluceau Rocquencourt
78153 Le Chesnay
FRANCE

Wojciech Szpankowski‡
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

Abstract

This paper studies in a probabilistic framework some topics concerning the way words (strings) can overlap, and relationship of this to the depth of a digital tree associated with this set of words. A word is defined as a (random) sequence of (possibly infinite) symbols over a finite alphabet. By depth of a word (or the associated suffix tree) we mean the average length of a string that can be recopied. In other words, the depth is a measure of compressions for words, and as such finds many applications in computer sciences and telecommunications, most notably in coding theory, theory of languages, and design and analysis of algorithms. Our main finding shows that the depth of a word (suffix tree) is asymptotically equivalent to the depth of a set of independent words (i.e., independent tries). More precisely, let us consider the first n suffixes of a random word. Then, the depth of a suffix tree built from these suffixes is *normally distributed* with the mean $1/h_1 \cdot \log n$ and the variance $\alpha \cdot \log n$ where h_1 is entropy of the alphabet, and α is a parameter of the probabilistic model. Finally, we present some consequences of our findings for the design and analysis of algorithms on words. In particular, we consider compression schemes that compress repated patterns by sending pointers to them instead of the patterns themselves, and prove that such a compression is not efficient for random strings.

†This research was partially supported by NATO Collaborative Grant 0057/89.

‡This research was primary done while the author was visiting INRIA in Rocquencourt, France. Support was provided in part by NATO Collaborative Grant 0057/89, in part by NSF grants NCR-8702115 and CCR-8900305, and in part by grant R01 LM05118 from the National Library of Medicine.

PAPIER RECUPERE ET RECYCLE

AUTOCORRELATION DANS LES CHAINES DE CARACTERES ET APPLICATIONS

Analyse des arbres suffixes par une méthode de mot coulissant

Résumé

Ce papier analyse dans une perspective probabiliste les chevauchements de mots dans des chaînes de caractères et les met en correspondance avec la profondeur de l'arbre digital associé (arbre suffixe). On définit ainsi la profondeur d'une chaîne de caractères (ou de l'arbre suffixe associé) comme étant la longueur moyenne de toutes les séquences que l'on peut lire à au moins deux endroits différents dans la chaîne. En d'autres termes, la profondeur est une mesure de l'efficacité de la compression susceptible d'être appliquée à une chaîne donnée de caractères, et trouve de nombreuses applications en informatique et en télécommunications, notamment dans la théories des codes, des languages et dans la conception et l'analyse d'algorithme. Notre contribution principale réside dans la démonstration du fait que la profondeur d'une chaîne aléatoire de caractères est asymptotiquement équivalente à celle d'un arbre digital aléatoire construit à partir d'insertions indépendantes. En conséquence, si on considère les n premiers suffixes d'une chaîne aléatoire, alors la profondeur de l'arbre suffixe associé tend à être distribuée selon la loi *normale* de moyenne $1/h_1 \cdot \log n$ et de variance $\alpha \cdot \log n$ quand $n \to \infty$. Le paramètre h_1 est l'entropie de l'alphabet considéré et α un autre paramètre similaire du deuxième ordre. Nous terminons en décrivant quelques conséquences pratiques de ce résultat. Nous établissons que la procèdure de compression de chaîne qui consiste simplement à remplacer les séquences répètées par des pointeurs se révèle inefficace lorsque les chaînes sont purement aléatoires.

1. INTRODUCTION

Periodicities, autocorrelations and related phenomena in words are known to play a central role in many facets of theoretical computer science, notably in coding theory, in the theory of formal languages and in the design and analysis of algorithms. In this latter field, several efficient algorithms have been set up to date both to detect and to exploit the presence of repeated subpatterns and other kinds of avoidable or unavoidable [LO] regularities in words. In this paper, our interest lies in estimating the average depth of a random word X, where average depth is defined as the average length of a substring of X that can be recopied. As such, the depth can be regarded as a measure of compression that can be achieved over the code X.

On the other hand, periodicities, autocorrelations and related phenomena can be equivalently studied through an associated digital tree called a *suffix tree* [WE, AHU, AA, AS]. A suffix tree of a word X is a (noncompact) *trie* built over keys (subwords) that are suffixes of the initial word X. We do not compress the trie as in PATRICIA (cf. [KN]), that is, in our construction of suffix trees no substrings are collapsed into one node. Such a digital tree has correlated keys which makes the analysis non-trivial. Parameters of interest are depth and height of the tree, that are of course in a one-to-one correspondence with the depth and height of the word X (see below for more detailed definitions). We found it more convenient to work with suffix trees, and most of our main results are presented in terms of parameters of these trees. In particular, we evaluate the depth of a suffix tree built from suffixes of a random string X. We shall assume that symbols of the string are drawn independently from some finite alphabet, however, it is possible to extend our analysis to some models with dependency between symbols (e.g., Markovian model, see [JS]).

The paper is organized as follows. In Section 2, we introduce some measures of correlation among subwords of a word. We find it convenient to assume our words unbounded, and as we shall see this does not limit our analysis since finiteness of a word contributes negligibly for long words. In particular, we define a self-alignment C_{ij} of any pair of distinct suffixes S_i and S_j of an unbounded random word X, as the length of the longest common prefix of those suffixes. Then the depth of a fixed key in a tree (or equivalently of a fixed suffix in the word X) is the maximum over all self-alignments of a fixed suffix, that is, the depth of the i-th suffix $D_n(i)$ is $D_n(i) = \max\{C_{i1}, C_{i2}, ..., C_{in}\}$. Finally, the *average* depth D_n is defined as the length of a randomly selected suffix among the n suffixes of X (for more detailed definition see Sections 2 and 3.2).

In Section 2 we also present our main results. We show that the depth D_n of a suffix tree built over the first n suffixes of a random word X, is *normally distributed* for large n. In particular, the mean depth ED_n is asymptotically equal to $1/h_1 \cdot \log n$, where h_1 is the entropy of the alphabet. Moreover, the variance $varD_n$ for large n is equal to $\alpha \cdot \log n$, where α is a parameter of the probabilistic model. In addition, we show that the average size of a suffix tree is asymptotically equal to $n/h_1 \cdot (1 + P(\log n))$ where $P(\log n)$ is a fluctuating function with small amplitude.

We delay all proofs till Section 3, 4 and 5. In Section 3 we prove our main findings concerning the depth. This section presents our novel approach to the analysis of some digital data structures like suffix trees. In short, in our new method of attack we consider

an auxiliary string σ called a "ruler", which is used to measure correlation among strings. We call this method the *string-ruler approach.* We shall show that the depth of a suffix tree *does not* differ significantly (e.g., $O(1/n^{\epsilon})$) from a trie built from *independent words.* Such independent tries have been recently extensively analysed, most notably in [FL, KN, JR, RJ, PI1, PI2, SZ1, SZ3]. In particular, Pittel [PI2], and Jacquet and Régnier [JR] derived a limiting distribution for the depth in the independent model, while recently Jacquet and Szpankowski have obtained a limiting distribution for the Markovian model [JS]. This finding is used in our paper to prove our main results. Finally, in Section 4 we provide a simple proof of our another result concerning the average size of suffix trees, and Section 5 (Appendix) contains some remaining proofs.

The literature on the analysis of suffix trees is very scare. To the best of our knowledge, the analysis of the height of the suffix tree was initialized by Apostolico and Szpankowski [AS], and recently Devroye, Szpankowski and Rais [DSR] have established exact asymptotics for the height. The size of a suffix tree was investigated by Blumer, Ehrenfeucht and Haussler [BEH] using a mixture of analytical and simulation tools (so some more work is needed here), and in in Section 4 we present rigorous analysis for the average size of suffix trees. The depth of the suffix tree (which, as we shall argue below, is the hardest to analyze) was left open, and this paper is intended to fill this gap.

Finally, we point out that our probabilistic results can be applied to the average case analysis of some simple algorithmic solutions of important problems on words (for more details see [AS]). In particular, we find that building the suffix tree for such a word, which takes linear time by clever methods [MC], takes $O(n \log n)$ time by the direct (natural) method (see [AS] for definition of the direct method); detecting all squares in that word, which takes optimal $O(n \log n)$ time by clever methods [AP1, CR, ML], takes $O(n \log n)$ expected time by a simpler method; computing the full statistics without overlap of all substrings of that word, which takes $O(n \log^2 n)$ time by clever methods [AP1, AP2], takes $O(n \log n)$ expected time by a simpler method [AS], etc. And last but not least, we consider a natural compression algorithm that compresses repeated patterns by sending only pointers to them. Using our results we show that a compression coefficient defined as the ratio between the average length of repeated patterns and the length of overhead information is larger than one for uniform and "close" to uniform distributions of symbols. This implies that random codes should not be compressed unless the alphabet is very asymmetric and repeated patterns are likely to be long or a more efficient and more sophisticated algorithm is proposed.

2. AUTOCORRELATION PARAMETERS IN WORDS

Let $X = x_1 x_2 x_3 \ldots$ be a string of unbounded length formed by symbols from an alphabet Σ of cardinality V, and let $S_i = x_i x_{i+1} \ldots$ be the i-th *suffix* of X. For every off-diagonal pair (i, j) of positions of X, we define C_{ij} as the length of the longest string that is a prefix of both S_i and S_j. We leave C_{ij} undefined when $i = j$. Thus, $C_{ij} = k$ iff S_i and S_j agree exactly on their first k symbols, but differ on their $(k+1)$-st. Clearly, $C_{ij} = C_{ji}$ for all meaningful choices of i and j.

Let now n be any fixed integer. The following two parameters, namely the n-th *height* H_n of X and the n-th *depth* $D_n(i)$ of the i-th suffix of X, which find many applications in

practice, are defined as follows .

$$H_n = \max_{1 \le i < j \le n} \{C_{ij}\}, \tag{2.1a}$$

$$D_n(i) = \max_{1 \le j \le n, j \ne i} \{C_{ij}\} . \tag{2.1b}$$

By the n-th depth D_n, called also the *average* depth, one understands the depth of a randomly selected symbol among the first n symbols of X. Intuitively, H_n is the maximum possible length for a substring Z of X that has at least two occurrences in X, both starting within the first n positions of X. Thus, there are two positions i and j of X, $i < j < n$, such that the occurrence of Z starting at j can be fully recopied from the occurrence starting at i. The average depth is D_n and represents the average length, over the first n positions of X, of the longest substring of X that can be recopied from the past. The height H_n and the depth D_n express structural correlations among the substrings of string X. Such correlations play a crucial role in many combinatorial and algorithmic constructions, and our three definitions above are somewhat reminiscent of notions that have already appeared in the literature, notably, in [LZ, ZL,GO2, GO3].

For a given n, the symmetric table collecting all meaningful values C_{ij} is the n-th *self-alignment matrix* of X. In the following, we refer to a generic off-diagonal entry of this matrix by one of the terms *self-alignment* or *common* , the latter term being mnemonic for length of the longest prefix common to a generic pair of suffixes of X". The following example illustrates the notions introduced so far.

EXAMPLE. *Illustrating definitions*

Let $X = abbabaa...$ and $n = 5$. Then $S_1 = X$, $S_2 = bbabaa...$, $S_3 = babaa...$, $S_4 = abaa...$ and $S_5 = baa...$. The corresponding self-alignment matrix $C = \{C_{ij}\}_{i,j=1}^{5}$ is as follows:

$$\mathbf{C} = \begin{bmatrix} \star & 0 & 0 & 2 & 0 \\ 0 & \star & 1 & 0 & 0 \\ 0 & 1 & \star & 0 & 2 \\ 2 & 0 & 0 & \star & 0 \\ 0 & 1 & 2 & 0 & \star \end{bmatrix}$$

¿From $\mathbf{C}$ and the expressions (2.1), we obtain $H_n = 2$ and the average depth $D_n 9/5$. ■

We deal here with the probabilistic analysis of the above quantities, in particular the depth, under the *Bernoulli model* assumptions, that is: *the symbols of X are drawn independently from Σ, and the i-th symbol of Σ occurs in X with probability p_i, $i = 1, 2, ..., V$,* $\sum_{i=1}^{V} p_i = 1$. Naturally, the height H_n and the depth D_n of a random word X are equal to the height H_n and the depth D_n of the associated suffix tree constructed from the first n suffixes of the word X. Therefore, we shall further reason in terms of these parameters for the suffix tree.

The height H_n of a suffix tree was very recently studied in [AS, DSR]. In [SZ3] Szpankowski (see also [AS]) has initialized investigations of the height through the self-alignment matrix, and in [AS] Apostolico and Szpankowski have obtained an upper bound for the height of suffix trees. In particular, they derived a distribution function for the

self-alignments C_{ij}, which is crucial to obtain more detailed information regarding the height. Using this, Devroye, Szpankowski and Rais very recently have proved that H_n tends *in probability* to $\log_Q n$, where $Q = \left(\sum_{i=1}^{V} p_i^2\right)^{-1}$ as n goes to infinity. This remarkable result can be compared with an *identical* formula (the leading term of asymptotics) for the height of an *independent trie*, that is, a digital tree built from n independent keys $X_1, ..., X_n$. Does a similar identity (in an asymptotic sense) hold for the depth ? We shall prove in this paper that the answer to this question is affirmative, however, we shall first demonstrate below difficulties arising in the analysis of depth D_n.

Let us consider a depth of a fixed suffix, say the first one, and let us denote it as $D_n(1)$. According to (2.1b) $D_n(1) = \max_{1\le j\le n}\{C_{1j}\}$. Note that the self-alignments $C_{1,j}$ are *strongly* dependent. In particular, to compute the distribution function $\Pr\{D_n(1) > k\}$ we need *all* joint distributions of the self-alignments. To be more precise, using *inclusion-exclusion formula* [BO] one immediately proves

$$\Pr\{D_n(1) > k\} = \sum_{r=2}^{n}(-1)^r \sum_{i_1,...,i_r} \Pr\{C_{1,i_1} > k, ..., C_{1,i_r} > k\}, \tag{2.2}$$

where i_j are distinct and $2 \le i_j \le n$ for every $1 \le j \le r$. An interesting fact is that, due to an alternating sum in (2.2), we *have to* take into account all terms of the above sum, and we need an *exact* formula for the joint distribution $\Pr\{C_{1,i_1} > k, ..., C_{1,i_r} > k\}$. For example, for independent tries, one easily proves [SZ3, JS] that for every r-tuple $(i_1, ..., i_r)$ the following holds

$$\Pr\{C_{1,i_1} > k, ..., C_{1,i_r} > k\} = (p^r + q^r)^{k+1} \tag{2.3}$$

where hereafter for simplicity we shall consider only binary model with $p_1 = p$ and $p_2 = q = 1 - p$. Form (2.2) and (2.3) we easily obtain the generating function of the depth, the average value, etc. For example, the average value $ED_n = ED_n(1)$ becomes

$$ED_n = \frac{1}{n}\sum_{r=2}^{n}(-1)^r\binom{n}{r}r\frac{p^r + q^r}{1 - p^r - q^r}. \tag{2.4}$$

The asymptotics of (2.4) were extensively studied in the past through the Mellin transform [KN, FL, JR, RJ, SZ1, SZ2] and probabilistics methods [DE], and one easy proves that $ED_n = 1/h_1 \cdot \log n + 1/h_1 \cdot (\gamma + h_2/(2h_1) + P(\log n) + O(n^{-1})$, where $h_1 = -p\log p - q\log q$ is the entropy of the alphabet, $h_2 = p^2\log p + q^2\log q$, and $P(\log n)$ is a fluctuating function (cf. [KN, FL, FRS, JR, SZ1]).

How one can use the above approach to analyse the depth of suffix trees ? We note that (2.2) holds for any trees since it is based only on the inclusion-exclusion formula. The independency was used to derive the joint distribution of the self-alignments (2.3). In the suffix tree case we must cope with overlapping, and this causes some troubles in the analysis, especially that, as mentioned above, we need an exact formula for the joint distribution. To illustrate some difficulties arising in the evaluation of this joint distribution, consider the following probability $\Pr\{C_{1,5} > 10, C_{1,8} > 10, C_{1,20} > 10\}$, and note that it is equal to $p^{28} + q^{28}$ which is quite different than (2.3). On the other hand,

when suffixes are separated by at least k symbols, then we can take advantage of the independence between symbols. More precisely, let us define a set of integers $r_1, r_2, ..., r_\ell$ such that for any $i < \ell - 1$ the following holds $r_{i+1} - r_i > k$. Then, (2.3) is true in the following sense

$$\Pr\{C_{1r_1} > k, ..., C_{1,r_\ell} > k\} = (p^r + q^r)^{k+1}. \tag{2.5}$$

Noting, in addition, that the probability of overlapping is very small we can expect that formula (2.4) approximately works out. Then, it is reasonable to expect identical asymptotics for the independent and the suffix tree models. The point is, however, that it is very hard to justify rigorously this idea due to the fact that (2.4) contains an alternating sum. In the next section, we adopt quite different and novel approach called "string-ruler" method to prove this fact.

Now we are in a position to summarized our main results. The first result deals with a comparison between the independent trie and suffix trees. Let, for a moment, D_n^T, D_n^S denote the depths in an independent trie and a suffix tree with n keys, respectively. In addition, we define the appropriate distribution functions as $F_n^T(k) = \Pr\{D_n^T < k\}$ and $F_n^S(k)$, respectively. The following proposition is proved in Section 3 (cf. Theorem 14).

PROPOSITION 1.

There exist $\beta > 1$ and $\epsilon > 0$ such that uniformly in k and n the below holds

$$|F_n^T(k) - F_n^S(k)| = O\left(\frac{1}{n^\epsilon \beta^k}\right). \tag{2.6}$$

In addition, all moments of the depth for suffix trees are in the same relationship to the appropriate moments of the depth for independent tries.

Proposition 1 establishes a methodological tool to analyze some *dependent* data structures such as suffix tree. It basically says that suffix tree does not differ too much from independent tries. But, tries have been analysed extensively over last decade, and virtually we know almost everything about them. In particular, the limiting distribution of the depth is known, the average depth and the variance are also well known. Therefore, Proposition 1 and recent results of Jacquet and Régnier [JR, RJ], Pittel [PI2] and Szpankowski [SZ1] imply our next main result.

PROPOSITION 2.

(i) For large n the average ED_n depth of a suffix tree becomes for some $\epsilon > 0$

$$ED_n = \frac{1}{h_1} \cdot \{\log n + \gamma + \frac{h_2}{2h_1}\} + P_1(\log n) + O\left(\frac{1}{n^\epsilon}\right), \tag{2.7a}$$

and the variance $varD_n$ of the depth is

$$varD_n = \frac{h_2 - h_1^2}{h_1^3} \log n + C + P_2(\log n) + O\left(\frac{1}{n^\epsilon}\right), \tag{2.7b}$$

where C is a constant which can be found in [SZ1]. In the symmetric case, i.e., $p_1 = p_2 = ... = p_V = 1/V$, the variance becomes

$$varD_n = \frac{\pi^2}{6\log^2 V} + \frac{1}{12} + O\left(\frac{1}{n^\epsilon}\right), \tag{2.7c}$$

where $h_1 = -\sum_{i=1}^{V} p_i \log p_i$ *and* $h_2 = \sum_{i=1}^{V} p_i^2 \log p_i$, *and* $P_1(x), P_2(x)$ *are fluctuating functions with small amplitudes.*

(ii) For the asymmetric model of suffix trees $(D_n - ED_n)/\sqrt{varD_n}$ *is asymptotically normal with mean zero and variance one. The convergence is in probability:* $\forall x \in R$,

$$\lim_{n\to\infty} \Pr\{D_n \le ED_n + x\sqrt{varD_n}\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \ ,$$

and in moments: for all integer m,

$$\lim_{n\to\infty} E\left[\frac{D_n - ED_n}{\sqrt{varD_n}}\right]^m = 0 \quad \text{when } m \text{ is odd}$$
$$= \frac{m!}{2^{m/2}(\frac{m}{2})!} \quad \text{when } m \text{ is even}$$

(see Jacquet Régnier [JR]). For the symmetric case, one proves that uniformly in $x \ge 0$ *the following convergence:*

$$\lim_{n\to\infty} \Pr\{D_n \le \log_V(xn)\} = e^{-x^{-1}} \tag{2.7d}$$

(see Pittel [PI2]).

In some applications size of suffix tree plays a more dominant role than the depth of the tree. By size of a digital tree we mean the number of (internal) nodes needed to build the tree. Most notably, size of suffix trees determines space requirements, and therefore space-complexity of any algorithm based on the suffix tree. The next proposition presents one result in this direction, namely the average size EL_n of suffix trees built from n suffixes. This result is a consequence of our previous findings, and will be proved in Section 4.

PROPOSITION 3.

There exist such $\epsilon > 0$ *that the average size* EL_n^S *of suffix tree and the average size* EL_n^T *of regular tries satisfy the following relationship*

$$|EL_n^S - EL_n^T| = O(n^{1-\epsilon}) \tag{2.8a}$$

In particular, this implies that

$$EL_n^S = \frac{n}{h_1}(1 + P_3(\log n)) + o(n) \tag{2.8b}$$

where h_1 *is the entropy of the alphabet, and* $P_3(\log n)$ *is a fluctuating function with a small amplitude (cf. [JR]).*

Finally, to get some idea about accuracy of our asymptotics (in particular Proposition 1) we have performed some simulation studies which are presented in the table below. The table compares *theoretical* values of the average depth ED_n^T, the variance $varD_n^T$ and the average size EL_n^T of independent tries with simulation results of ED_n^S, $varD_n^S$ and EL_n^S

respectively for suffix trees in the range of n varying from 700 to 10,000. These results confirm, as expected, our theoretical findings presented above, and in addition they show good accuracy of the asymptotics even for small values of n.

COMPARISON OF TRIES AND SUFFIX TREES

Asymmetric alphabet $p = 0.1$ *and* $q = 0.9$

n	ED_n^T	ED_n^S	$varD_n^T$	$varD_n^S$	EL_n^T	EL_n^S
700	24.483	23.861	112.09	110.85	2153.3	2064.8
800	24.894	24.340	113.78	110.95	2460.9	2385.7
900	25.256	24.766	115.27	113.12	2768.5	2691.1
1000	25.579	25.080	116.61	114.84	3076.1	3014.1
2000	27.720	27.457	125.37	124.81	6152.3	6069.0
3000	28.959	28.802	130.50	130.04	9228.4	9125.3
4000	29.844	29.652	134.14	132.85	12304.6	12155.1
5000	30.531	30.348	136.96	135.87	15380.7	15227.6
10000	32.663	32.574	145.73	144.40	30761.4	30763.2

Some consequences of our findings for design and analysis of algorithms on words are discussed in the introduction. Therefore, here we restrict our discussion to one example, namely *compression* of codes. Consider the following natural compression technique. Instead of sending a whole code one may discover some repetitions in the code that can be recopied by transmitting pointers to the repeated pattern. For example, consider a code ababaaaababb. Noting that the subcode between positions eight and eleven repeats previously transmitted subcode abab (see positions one to four), one may send two pointers *one* and *four* at the position eight instead of repeating the next four symbols. One should note that we trade the lengths of repeated patterns for the length of two pointers. A natural question arises whether it is worth to do it, say in a *random* code. To answer it, we introduce *the compression coefficient* C_V as the ratio between the *average* length of repeated patterns and the length of overhead information, that is, in our case the length of two pointers. But, the average length of repeated patterns is just equal to the depth of the random word X (or the depth of associated suffix tree), while the length of two pointers that can carry information regarding n possible positions is simple $2\log_V n$. Therefore, we just proved that

$$C_V = \frac{2h_1}{\log V}. \tag{2.9}$$

Since the entropy reaches its maximum for uniform distribution (symmetric case) from (2.8) one immediately shows that $C_V \leq 2$. But, for uniform distribution $C_V = 2$ (and $C_V > 1$ for distributions not too far away from the uniform distribution). Therefore, compression, as defined above, is *not* worth to do it. Naturally, for very asymmetric alphabet when $h_1 \leq \log V/2$ compression becomes useful. For example, when a is much more probably to be generated than b, then a pattern aaaaaabaaaaaa is likely to happen, and since longer repeated patterns are more likely to occur, compression becomes efficient. In short, our

analysis confirms what practitioners know very well that a random string is not worth to compress, unless a more sophisticated compression algorithm is used (for more practical compression techniques see [LZ, ZL]).

3. ANALYSIS AND AUTOCORRELATIONS OF STRINGS

In this section we provide a proof of our main result Proposition 1, as well as suggest a new line of investigation for autocorrelations of strings. Our approach to solve the problem seems to be new, and it only resembles some similarities with the work of Guibas and Odlyzko [GO1, GO2, GO3].

Before we plug into detailed analysis let us give an idea of our proof. In Section 2 we have shown that any analysis of the depth D_n of a digital tree, in particular, suffix tree needs an *exact* evaluation of the joint distribution of the self-alignments, as for example shown in (2.3) for independent tries. Such an evaluation for suffix tree is very complicated due to mutual overlapping of suffixes. Therefore, to avoid the alternating sum problem (in fact, to hide it in a generating function form), we suggest a different, more combinatorial approach. Our idea of the analysis of D_n is as follows. We consider a finite string σ that is used as a "ruler" to measure the length of strings, in particular, to estimate overlapping between any two suffixes. For example, to evaluate the self-alignment between the i-th suffix S_i and the j-th suffix S_j we compute first the common digits (i.e., alignments) between S_i and σ, and then the alignments between S_j and σ. These measures can be used to evaluate the self-alignment C_{ij} between S_i and S_j *with respect to* the ruler string σ. Finally, considering all possible ruler-strings σ we evaluate the self-alignments C_{ij}. To evaluate the *unconditional* self-alignment C_{ij} we simple compute the conditional self-alignments over all strings σ. This, although it looks more complicated than necessary, is the right approach as we shall prove below.

Using the above idea we shall compute the generating function of the average depth for suffix trees (difficult !) and independent tries (easy). These two generating functions are asymptotically compared to show that they do not differ too much for large n. This will lead to our main result Proposition 1.

It might be worth to point out that along the lines of our proof, we in fact explore the problem of autocorrelation properties of random strings. This finds much more applications than to suffix tree, e.g., in squares of strings, in biprefix strings, *etc.*

3.1 Some Notations and Preliminary Results

Let us start with a brief introduction to our "string-ruler" approach. As before, X denotes a random string over a finite alphabet, however, for simplicity of analysis we restrict our attention to a binary alphabet $\Sigma = \{a, b\}$ with p (resp. q) denoting the probability of a (resp. b) occurrence in X. Our interest lies in evaluating the correlation between the first n symbols of X. In order to measure it we consider another string, say σ, which is further called a ruler string. The length of σ is denotes $|\sigma|$, and $0 < |\sigma| < \infty$. Let $\langle X, \sigma \rangle$ represent the subset of positions $[1, n)$ of X that σ and X agree, that is, σ overlaps with X starting from any position of $\langle X, \sigma \rangle$. (Note that the sets $\langle X, \sigma \rangle$ and $\langle \sigma, X \rangle$ are not necessary the same.) At this time, we point out that some properties of $\langle X, \sigma \rangle$ were studied by Guibas and Odlyzko [GO1, GO2, GO3]. However, we shall use $\langle X, \sigma \rangle$ to investigate

the depth $D_n(i)$ of the i-th suffix of X. There is a simple relationship between the depth and properties of $\langle X, \sigma \rangle$. Indeed, one immediately notes the following

$$\{D_n(i) \leq k\} \qquad \Longleftrightarrow \qquad \exists \sigma : \ |\sigma| = k \text{ and } \langle X, \sigma \rangle = \{i\} \tag{3.1a}$$

and therefore

$$\Pr\{D_n(i) \leq k\} = \sum_{|\sigma|=k} \Pr\{\langle X, \sigma \rangle = \{i\}\} \ , \tag{3.1b}$$

The example below illustrates what we have done so far.

EXAMPLE 3.1 *Depth of the i-th suffix and* $\langle X, \sigma \rangle$.

Let $X = baabbabaaa...$ and we need to compute the depth $D_{10}(1)$ of the first suffix, that is, we would like to find such k that $\{D_{10}(1) \leq k\}$ holds. Let us first consider all σ such that $|\sigma| = 1$. Naturally, in this case $|\langle X, \sigma \rangle| > 1$, so then all σ with $|\sigma| = 2$ are investigated. But, for $\sigma = ba$ we have $\langle X, \sigma \rangle = \{1, 5, 7\}$. Finally, string-rulers with $|\sigma| = 3$ are studied, and for $\sigma = \{baa\}$ we find that $\langle X, \sigma \rangle = \{1\}$, so (3.1a) holds, and $\{D_{10}(1) \leq 3\}$. ■

The correlation $\langle X, \sigma \rangle$, and in particular (3.1), is used to derive the generating function for the depth D_n. We start with some notations. Let $ED_n(u)$ be the probability generating function of the depth, namely

$$ED_n(u) = \sum_{k=1}^{\infty} \Pr\{D_n = k\} u^k = E[u^{D_n}] \ .$$

and by $D(z, u)$ we denote the bivariate generating function of the $ED_n(u)$'s, that is,

$$D(z, u) = \sum_{n=0}^{\infty} n \, ED_n(u) z^n \ .$$

Note that the following relationship holds between the generating function of D_n and the generating function of $D_n(i)$ defined in (2.1b)

$$E[u^{D_n}] = \frac{1}{n} \sum_{i=1}^{n} E[u^{D_n(i)}] \ .$$

Now, we express $D(z, u)$ in terms of some probabilities defined on $\langle X, \sigma \rangle$. Therefore, let L be a nonempty subset of $[1, n)$, and by $P(L, \sigma)$ we denote the probability that L is included in $\langle X, \sigma \rangle$, that is, σ and X agree on at least the positions of L. For any finite subset L of $[1, \infty)$ we define $m(L)$ as the greatest element of L. Note that $P(L, \sigma)$ in fact does not depend on n, given that $n \geq m(L)$. Thus we can define $P(L, \sigma)$ without any reference to n (with the implicit condition $n \geq m(L)$). Note that our definition of $\langle X, \sigma \rangle$ together with the inclusive-exclusive formula [BO] implies

$$\Pr\{\langle X, \sigma \rangle = \{i\}\} = \sum_{j=1}^{j=n} (-1)^{j+1} \sum_{\substack{|L|=j \\ i \in L}} P(L, \sigma) \ . \tag{3.2}$$

Therefore, (3.1) and (3.2) lead to a formula for the generating function of the depth. Indeed, let us additionally define

$$P_j(\sigma) = \sum_{|L|=j} P(L,\sigma) \ ,$$

with $|L|$ being the number of elements of L, and

$$P_{n,\sigma}(v) = \sum_{j=1}^{j=n} P_j(\sigma)v^j \quad , \quad P_\sigma(z,v) = \sum_{n=1}^{\infty} P_{n,\sigma}(v)z^n \ .$$

Then, the following theorem can be easily proved.

THEOREM 1.
Let $n > 1$, when $|u| < 1$ we have the identity

$$ED_n(u) = \frac{(1-u)}{n} \sum_{\sigma} u^{|\sigma|} \frac{d}{dv} P_{n,\sigma}(v)|_{(v=-1)} \ , \tag{3.3}$$

where $|\sigma|$ is the length of the string σ and $\sum_\sigma$ means the summation on all possible finite string σ (of length greater than zero), that is, $\sum_\sigma f(\sigma) = \sum_{k=1}^{\infty} \sum_{|\sigma|=k} f(\sigma)$ for any function $f(\cdot)$ defined on strings.
Proof: According to (3.1) and (3.2) the depth D_n defined in introduction becomes

$$n \cdot \Pr\{D_n \le k\} = \sum_{i=1}^{i=n} \Pr\{D_n(i) \le k\} = \sum_{i=1}^{i=n} \Pr\{\langle X, \sigma\rangle = \{i\}\} = \sum_{|\sigma|=k} \sum_{j=1}^{j=n} (-1)^{j+1} j P_j(\sigma) \ .$$

After detecting the partial derivative of $P_{n,\sigma}(v)$ in the RHS of the above, the proof is completed. ■

Finally, since $D_1(u) = 1$, we easily get the following corollary.

COROLLARY 2
When $|u| < 1$, we also have the identity

$$D(z,u) = (1-u)z + (1-u) \sum_{\sigma} u^{|\sigma|} \frac{\partial}{\partial v} P_\sigma(z,v)|_{(v=-1)} \ ,$$

for $|z| < 1$. ■

Remark 1. It is worth to point out that the notation $\sum_\sigma$, which is extensively used throughout the entire section, express the sum over all finite string. There are 2^k distinct strings of length k (V^k in a V-ary alphabet). Let $|\sigma|_a$ and $|\sigma|_b$ be respectively the number of symbols a and b in σ. Then, the number of strings of length k such that $|\sigma|_a = i$ and $|\sigma|_b = k - i$ is equal exactly to $\binom{k}{i}$. This leads, for example, to the following identities

$$\sum_{|\sigma|=k} x^{|\sigma|_a} \cdot y^{|\sigma|_b} = (x+y)^k \ ,$$

and

$$\sum_{\sigma} x^{|\sigma|_a} \cdot y^{|\sigma|_b} = \frac{1}{1-x-y} ; \tag{3.4}$$

for suitable values for complex numbers x and y. ■

3.2 Computation of $P_\sigma(z,v)$

¿From Theorem 1 we know that the generating function $D_n(u)$ can be evaluated by computing the generating function $P_\sigma(z,v)$, so in this section we establish formula on it. For this, however, we need to estimate the autocorrelation function of σ, and the overlapping probability of two or more suffixes that are separated by less than, say k, symbols (such a set of overlapping is further called a cluster) in order to compute exactly the joint probability distribution of the alignments, as suggested in Section 2 (cf. formula (2.2) for independent tries).

Let $|\sigma| = k$, and for any finite string ξ, let $p(\xi)$ be the unconditional probability of the occurrence of ξ, namely the product $p^{|\xi|_a} q^{|\xi|_b}$ where $|\xi|_a$ is the number of a in ξ and $|\xi|_b$, the number of b in ξ. Let $F(\sigma) = \langle \sigma, \sigma \rangle - \{1\}$ be the autocorrelation set of the string σ,that is, $i \in F(\sigma)$ means that σ overlaps itself from position i, of course we omit the trivial position $i = 1$. Let $a_\sigma(z)$ be the *autocorrelation* polynomial of σ defined by

$$a_\sigma(z) = \sum_{i \in F(\sigma)} p(\sigma_{i-1}) z^{i-1} ,$$

where σ_{i-1} denotes the prefix of σ of length $i-1$.

Remark 2. Our definition of autocorrelation polynomial resembles an appropriate definition of autocorrelation function introduced by Guibas and Odlyzko [OD1, OD2, OD3], however in our case the autocorrelation polynomial is additionally weighted by the probability $p(\sigma_i)$. It can be also proven that very often $a_\sigma(z) = p(\sigma) z^k O(1)$. But sometimes, for example with $\sigma = aaa \cdots aa$, we have exception, namely $b(\sigma, z) = pz + (pz)^2 + \cdots + (pz)^{k-1}$. See Lemma 7 for a rigorous statement. ■

Now we are ready to deal with a k-cluster, that is, a collection of overlapping suffixes that are separated by less than k symbols. Let ℓ be a finite subset of $[1, \infty)$. The set ℓ is a k-cluster if $1 \in \ell$ and either ℓ contains no other elements or ℓ can be considered as an increasing sequence of integers such that the difference between any two consecutive elements is always strictly less than k. Let $C_\sigma(z,v)$ be the bivariate generating function defined by

$$C_\sigma(z,v) = \sum_{\ell} P(\ell, \sigma) z^{m(\ell)} v^{|\ell|} ,$$

where $P(\ell, \sigma)$ is the probability that ℓ is a k-cluster, and $\sum_\ell$ means the summation on all possible k-cluster. Then, one proves

THEOREM 3.
We have the expression

$$C_\sigma(z,v) = \frac{p(\sigma) z\, v}{1 - a_\sigma(z) v} . \tag{3.5}$$

for all $|u| < 1$ *and* $|z| < 1$.
Proof: A cluster ℓ is either $\{1\}$ itself or of the form $\{1\} \cup (\ell' + i - 1)$, where ℓ' is another cluster, $i \in F(\sigma)$ ($\ell' + i - 1$ means that we add $i - 1$ to each element of ℓ'). Since $P(\ell' + i - 1, \sigma) = P(\ell', \sigma)$ and $m(\ell' + i - 1) = m(\ell') + i - 1$, we have the identity

$$P(\ell, \sigma) z^{m(\ell)} v^{|\ell|} = p(\sigma_{i-1}) z^{i-1} v \, P(\ell', \sigma) z^{m(\ell')} v^{|\ell'|} \, .$$

Furthermore, we trivialy have $p(\{1\}, \sigma) z^{m(\{1\})} z^{|\{1\}|} = p(\sigma) z \, v$, thus the following equation holds $C_\sigma(z, v) = p(\sigma) z \, v + a_\sigma(z) v C_\sigma(z, v)$, which ends the proof. ∎

In order to obtain a final form of the generating function $P_\sigma(z, v)$ we need a little lemma.

LEMMA 4.
We have the identity

$$P_\sigma(z, v) = \frac{1}{1 - z} S_\sigma(z, v) \, ,$$

with

$$S_\sigma(z, v) = \sum_L P(L, \sigma) z^{m(L)} v^{|L|} \, ,$$

where $\sum_L$ *means the summation over all possible finite nonempty subsets* L *of* $[1, \infty)$.
Proof: By rearranging the terms in the summation it is easy to see that

$$P_\sigma(z, v) = \sum_L P(L, \sigma) v^{|L|} \sum_{n=m(L)}^{\infty} z^n \, .$$

as needed. ∎

Finally, combining (3.5) and the above we prove the following.

THEOREM 5
The generating function $P_\sigma(z, v)$ *can be expressed as*

$$P_\sigma(z, v) = \frac{1}{(1 - z)} \cdot \frac{C_\sigma(z, v)}{1 - z - z^k C_\sigma(z, v)} \, . \tag{3.6}$$

where $k = |\sigma|$.
Proof: Let L be a finite subset of $[1, \infty)$ (not empty). When ordering L, L resolves itself in a succession of distinct k-clusters, $\ell_1, \ell_2, \cdots, \ell_m$, *modulo* some translation of course, so

$$L = \ell_1 + i_1 \cup \ell_2 + i_2 \cup \cdots \cup \ell_m + i_m \, ,$$

where $i_1, \cdots, i_m$ are suitable integers. Therefore

$$P(L, \sigma) = P(\ell_1, \sigma) P(\ell_2, \sigma) \cdots P(\ell_m, \sigma) \, . \tag{3.7}$$

Let $i = \inf L$, then we alternatively have the following cases
(i) L is a k-cluster itself, *modulo* some translation, i.e., $L = \ell + i - 1$, where ℓ is a k-cluster.

(ii) $L = (\ell + i - 1) \cup (L' + m(\ell) + k - 1 + i - 1)$, where ℓ is a k-cluster and L' another finite subset of $[1, \infty)$.

Point (i) gives

$$P(L, \sigma) z^{m(L)} v^{|L|} = z^{i-1} P(\ell, \sigma) z^{m(\ell)} v^{|\ell|} .$$

Point (ii) gives

$$P(L, \sigma) z^{m(L)} v^{|L|} = z^{i-1} P(\ell, \sigma) z^{m(\ell)} v^{|\ell|} z^{k-1} P(L', \sigma) z^{m(L')} v^{|L'|} .$$

These two points together produce

$$S_\sigma(z, v) = \frac{C_\sigma(z, v)}{1 - z} + \frac{z^k C_\sigma(z, v)}{1 - z} S_\sigma(z, v) .$$

as desired ∎

Therefore, by Theorem 1 and Theorem 5 we finally obtain our main result of this subsection, that is, an evaluation of the generating function $D(z, u)$ as a function of the autocorrelation polynomial $a_\sigma(z)$. Indeed,

COROLLARY 6.
The generating function for the depth D_n of suffix trees becomes

$$D(z, u) = z(1 - u) + (1 - u) \sum_\sigma u^{|\sigma|} \frac{p(\sigma) z}{[(1 - z)(1 + a_\sigma(z)) + p(\sigma) z^{|\sigma|+1}]^2} \tag{3.8}$$

for every $|u| < 1$ and $|z| < 1$. ∎

Remark 3. If, instead of an infinite string X, we take a finite string X, say of length n, we get in Lemma 4

$$P_\sigma(z, v) = \frac{z^{k-1}}{1 - z} S_\sigma(z, v) ,$$

because for any finite subset L of $[1, \infty)$ $P(L, \sigma) \neq 0$ if and only if $m(L) + |k| - 1 \leq n$. Therefore, in this case

$$D(z, u) = z(1 - u) + (1 - u) \sum_\sigma (uz)^{|\sigma|} \frac{p(\sigma)}{[(1 - z)(1 + a_\sigma(z)) + p(\sigma) z^{|\sigma|+1}]^2}$$

for all $|u| < 1$ and $|z| < 1$.

Remark 4. We can define generating functions $ED_n(u)$ and $D(z, u)$ for independent tries built over n independent strings $X_1, ..., X_n$, in the same manner as above. This case is much simpler since strings are not suffixes of any initial word, and do not suffer mutual correlations. We will use the superscript S for suffix trees and superscript T for regular trie when it will be necessary to point out distinction. The generating function $D_n^T(z, u)$ for independent tries can be derived in the same fashion as before. However this time, we define $\langle \mathbf{X}, \sigma \rangle$, where $\mathbf{X} = (X_1, ..., X_n)$, as those strings (i.e., string numbers) that agree with σ.

Then, the depth is evaluated as in (3.1) and Theorem 1, provided $P_\sigma(z)$ and $P_\sigma(z,v)$ can be computed. But independence assumption leads immediately to $P_j(\sigma) = \binom{n}{j} p^j(\sigma)$ and therefore

$$P_\sigma(z,v) = \frac{1}{1 - z[1 + p(\sigma)v]} - \frac{1}{1-z},$$

and finally by Corollary 2 we obtain

$$D^T(z,u) = (1-u)z + (1-u)\sum_\sigma u^{|\sigma|} \frac{p(\sigma)z}{[1 - z + p(\sigma)z]^2} . \tag{3.9}$$

We shall use this formula to compare regular tries with suffix trees.

3.3 Asymptotics

In this section we present an asymptotic analysis of the depth D_n of suffix trees through a careful estimation of the generating function $D(z,u)$ around its singularities. The asymptotics of $D(z,u)$ is carried out in three steps. *At first*, we prove that the generating function $D(z,u)$ can be analytically continued to $|u| < 1 + \epsilon$ (cf. Theorem 8). This strengthens our results in the sense that not only convergence *in distribution* but also convergence *in mean* can be established (we use the well known fact that every analytical function has its derivatives). In the *second step*, we prove that the extended generating function has only a single pole that completely determines the asymptotics (cf. Theorem 11). Finally, the *third step* consists of applying the celebrated Cauchy's theorem [HE] to prove asymptotics. However, to simplify our analysis we directly compare the asymptotics of suffix trees with independent tries (cf. Theorem 14) to take advantage of many well established results for tries (cf. [KN, JS, RS, PI2, SZ1]).

We start with a technical, but important, lemma which says that the autocorrelation polynomial most likely is of the form $O(1)p(\sigma)z^k$ for any string σ of length k, that is, overlapping is rather unlikely for random strings (cf. Remark 2). Let us suppose that $p \geq q$, and we assume that $p < 1$. We consider all finite string σ of length k. Let $f(\sigma)$ be a function of σ and a real number x. In addition, we define $\Pr_k(f_\sigma(x) \leq y)$ as $\sum_{\substack{|\sigma|=k \\ f_\sigma(x)\leq y}} p(\sigma)$. The following lemma estimates a "typical" form of the autocorrelation polynomial.

LEMMA 7
There exists $\delta < 1$ and $\theta > 0$, such that for all $\rho \geq 1$,

$$\Pr_k\{a_\sigma(\rho) \leq \theta(\rho\delta)^k\} \geq 1 - \theta\delta^k .$$

Proof: See Appendix in Section 5. ■

Now we are ready to prove our first main finding as discussed above. Let $R_\sigma(z)$ be $(1-z)(1+a_\sigma(z)) + p(\sigma)z^{|\sigma|+1}$, and $1 > \delta \geq p$. Then, the analytical continuation of $D(z,u)$ is established below.

THEOREM 8

The expression of $D(z,u)$ (and therefore this of $ED_n(u)$), that is,

$$D(z,u) = z(1-u) + (1-u)\sum_{\sigma} u^{|\sigma|}\frac{p(\sigma)z}{(R_\sigma(z))^2} ,$$

can be analytically continued for all $|u| < \delta^{-1}$ for some $\delta < 1$.
Proof: Let $|u| < 1$ and consider the following identity

$$\sum_{\sigma} u^{|\sigma|}\frac{p(\sigma)z}{(1-z)^2} = \frac{z}{(1-u)(1-z)^2} .$$

Therefore, for $|z| < 1$,

$$\begin{aligned} D(z,u) - \frac{z}{(1-z)^2} - (1-u)z &= (1-u)\sum_{\sigma} u^{|\sigma|}p(\sigma)z\left(\frac{1}{(R_\sigma(z))^2} - \frac{1}{(1-z)^2}\right) \\ &= \sum_{\sigma} u^{|\sigma|}p(\sigma)\frac{z}{(R_\sigma(z)(1-z)^2}(R_\sigma(z)-(1-z))(R_\sigma(z)+(1-z)) . \end{aligned}$$

We have $R_\sigma(z) - (1-z) = (1-z)a_\sigma(z) + p(\sigma)z^{k+1}$. Applying Lemma 7, we know that

$$\Pr_k\{|R_\sigma(z) - (1-z)| \le (|1-z|+1)\delta^k\} \ge 1 - O(p\delta)^k) .$$

Since in every case $|a_\sigma(z)| < (1-p)^{-1}$, we get the following evaluation

$$\begin{aligned} D(z,u) - (1-u)z - \frac{z}{(1-z)^2} &= \\ &= \sum_k u^k \left[\Pr_k\{|R_\sigma(z)-(1-z)| \le (|1-z|+1)\delta^k\}O(\delta^k) + \right. \\ &\left. +(1-\Pr_k\{|R_\sigma(z)-(1-z)| \le (|1-z|+1)\delta^k\})O(1)\right] , \end{aligned}$$

and we easily get $D(z,u) - \frac{z}{(1-z)^2} - (1-u)z = O(\frac{1}{1-\delta|u|})$. ∎

The next step in the asymptotic analysis is to find singularities of the generating function $D(z,u)$ that contributes to the asymptotics. We shall show that $D(z,u)$ does not have any singularities in the disk $|z| < 1$ (cf. Lemma 9), and the only pole of the generating function is for $|z| > 1$ (cf. Theorem 11). In addition, in Lemma 10 we provide one technical lemma required in further proofs, in particular to apply Rouche's theorem [HE] needed in Theorem 11. The proofs of Lemmas 9 and 10 are moved to Appendix, and we only show here how to establish Theorem 11.

LEMMA 9
The polynomial $R_\sigma(z)$ has no root in the disk $|z| < (1-p(\sigma))^{-1/k}$.

LEMMA 10

Let ℓ be an integer such that $p + p^\ell < 1$, and let $\rho > 1$ be such that $p\rho + (p\rho)^\ell < 1$. Then, there exists an integer K, and a real number $\alpha > 0$ that the following

$$|\sigma| \geq K \text{ and } |z| \leq \rho \Rightarrow |1 + a_\sigma(z)| \geq \alpha \ .$$

holds.

Then, our second main finding is the following theorem.

THEOREM 11
There exists K' such that, for each finite string σ which satisfies $|\sigma| \geq K'$, there is only one root, A_σ, of $R_\sigma(z) = 0$, with $1 < |z| \leq \rho$.
Proof: Let k be large enough such that $(p\rho)^k < \alpha(\rho - 1)$. Thus, according to Lemma 9, for all z such that $|z| = \rho$ we have $|p(\sigma)z^k| < |(z-1)(1 + a_\sigma(z)|$. Therefore, by Rouché's theorem [HE] (cf. Lemma 10), expression $R_\sigma(z)$ has the same number of roots as $(1-z)(1+a_\sigma(z))$ in the disk $|z| \leq \rho$. In addition, the polynomial $(1-z)(1+a_\sigma(z)$ has 1 has only single root in this disk (again by Lemma 9). Theorem 8, allows us to conclude that $|A_\sigma| > 1$. ■

As a consequence of Theorem 11 we conclude that there exists the smallest root of $R_\sigma(z)$ which we denote as A_σ. Let also C_σ and D_σ be the first and second derivatives of $R_\sigma(z)$ at $z = A_\sigma$ respectively. We easily get the following expansions

$$\begin{cases} A_\sigma = 1 + \dfrac{1}{1 + a_\sigma(1)} p(\sigma) + O(p(\sigma)^2) \\ C_\sigma = -1 - a_\sigma(1) + (k - \dfrac{2a'(\sigma,1)}{1 + a_\sigma(1)})p(\sigma) + O(p(\sigma)^2) \\ D_\sigma = -2a'(\sigma,1) + (k(k-1) - \dfrac{3a''(\sigma,1)}{1 + a_\sigma(1)})p(\sigma) + O(p(\sigma)^2) \ , \end{cases}$$

where quantities $a'(\sigma,1)$ and $a''(\sigma,1)$ respectively denote the first and second derivatives of $a_\sigma(z)$ at $z = 1$.

Finally, we are ready to the last step in our asymptotic analysis, namely to compare asymptotics of suffix trees with asymptotics of independent tries, to conclude that they do not differ too much (cf. Theorem 14). Let us define two new generating functions $Q_n(u)$ and $Q(z,u)$ that represent difference between the probability distributions of the depth in the suffix tree and in the regular trie built over independent strings. In other words, we introduce the following

$$Q_n(u) = \frac{1}{1-u}\left(ED_n^S(u) - ED_n^T(u)\right)$$

$$Q(z,u) = \sum_{n=0}^{\infty} n\, Q_n(u) z^n = \frac{1}{1-u}\left(D^S(z,u) - D^T(z,u)\right) \ .$$

Then, by Corollary 6 and Remark 4 we find

$$Q(z,u) = \sum_\sigma u^{|\sigma|} p(\sigma) z \left(\frac{1}{R_\sigma(z)^2} - \frac{1}{(1 - z + p(\sigma)z)^2}\right) \ .$$

Then, it is not difficult to establish asymptotics of $Q_n(u)$ by appealing to the Cauchy theorem. This is done in the following lemma.

LEMMA 12
There exists $B > 1$, such that the following evaluation holds for all $|u| \leq \beta$:

$$Q_n(u) = \sum_{\sigma} u^{|\sigma|} p(\sigma) \left(A_\sigma^{-n} \left(\frac{n}{C_\sigma^2 A_\sigma} + \frac{D_\sigma}{C_\sigma^3} \right) - n(1-p(\sigma))^{n-1} \right) + O(B^{-n}) .$$

Proof: By Cauchy

$$Q_n(u) = \frac{1}{2i\pi} \oint Q(z,u) \frac{dz}{z^{n+1}} ,$$

where the integration is done along a loop included in the unit disk and encircling the origin. Let σ such that $|\sigma| \geq K'$, we know, by Rouché, that $R_\sigma(z)$ and $(1 - z + p(\sigma)z)$ has only one root in $|z| \leq \rho$. Applying residues formula [HE]:

$$\frac{1}{2i\pi} \oint u^{|\sigma|} p(\sigma) \frac{dz}{z^n} \left(\frac{1}{R_\sigma(z)^2} - (1 - z + p(\sigma)z)^{-2} \right) =$$
$$= u^{|\sigma|} p(\sigma) \left(A_\sigma^{-n} \left(\frac{n}{C_\sigma^2 A_\sigma} + \frac{D_\sigma}{C_\sigma^3} \right) - n(1-p(\sigma))^{n-1} \right) + + I_\sigma(\rho, u) ,$$

where

$$I_\sigma(\rho, u) = \frac{1}{2i\pi} \int_{|z|=\rho} u^{|\sigma|} p(\sigma) \frac{dz}{z^n} \left(\frac{1}{R_\sigma(z)^2} - (1 - z + p(\sigma)z)^{-2} \right) .$$

In addition, by factorization we obtain

$$\frac{1}{R_\sigma(z)^2} - (1 - z + p(\sigma)z)^{-2} = \frac{[(z-1)a_\sigma(z) + p(\sigma)(z - z^k)](R_\sigma(z) + 1 - z + p(\sigma)z)}{R_\sigma(z)^2 (1 - z + p(\sigma)z)^2} ,$$

and using Lemma 7, as in the proof of Theorem 8 (and the fact that $a_\sigma(\rho) \leq 1/(1 - p\rho)$), we easily get (with $\delta \geq p$)

$$\sum_{|\sigma|=k} p(\sigma) I_\sigma(\rho, u) = O((\delta \rho u)^k \rho^{-n}) .$$

Hence, the evaluation follows

$$\sum_{|\sigma|>K'} p(\sigma) I_\sigma(u) = O(\rho^{-n}) .$$

The terms which correspond to the σ such that $|\sigma| \leq K'$ are in a finite number and are in B^{-n}, since the roots of each $R_\sigma(z)$ are all with modulus strictly greater than 1. ■

Finally, the main theorem of this section (and the paper) follows. The theorem below is our Proposition 1 from Section 2 rephrased in terms of generating functions rather than in probability distribution functions. It says that independent tries very closely approximate suffix trees (in fact, not only from the depth view point; see Proposition 3 and [DSR]).

THEOREM 14
For all $1 < \beta < \delta^{-1}$, there exists $\varepsilon > 0$ such that uniformly for $|u| \leq \beta$: $D_n^S(u) - D_n^T(u) = (1-u)O(n^{-\varepsilon})$.
Proof : The expansion of D_σ, with respect to $p(\sigma)$, and Lemma 7 show that as $n \to \infty$ the following holds $\sum_\sigma u^{|\sigma|}p(\sigma)A_\sigma^{-n}D_\sigma/C_\sigma^3 = O(1)$. Therefore

$$Q_n(u) = n\sum_\sigma u^{|\sigma|}p(\sigma)\frac{A_\sigma^{-n-1}}{C_\sigma^2} - (1-p(\sigma))^{n-1} + O(1) .$$

We omit the proof of the absolute convergence of the summation which can be done by using the expansions of C_σ and A_σ and Lemma 7. Let f_σ be the function defined for x real by

$$f_\sigma(x) = \frac{A_\sigma^{-x-1}}{C_\sigma^2} - (1-p(\sigma))^{x-1} .$$

The summation $\sum_\sigma u^{|\sigma|}p(\sigma)f_\sigma(x)$ is absolutely convergence for all x and u such that $|u| \leq \beta$. The function $f_\sigma(x) - f_\sigma(0)e^{-x}$ is exponentially decreasing when $x \to +\infty$ and is $O(x)$ when $x \to 0$; therefore its Mellin transform $f_\sigma^*(s)$ (for properties of the Mellin transform see [FRS]), is defined for $\Re(s) > -1$. We have

$$f_\sigma^*(s) = \Gamma(s)\left(\frac{(\log A_\sigma)^{-s} - 1}{A_\sigma C_\sigma^2} - \frac{(-\log(1-p(\sigma))^{-s} - 1}{1-p(\sigma)}\right) .$$

Let $g^*(s,u)$ be the Mellin transform of $\sum_\sigma u^{|\sigma|}p(\sigma)(f_\sigma(x) - f_\sigma(0)e^{-x})$. Formally we have

$$g^*(s,u) = \Gamma(s)\sum_\sigma u^{|\sigma|}p(\sigma)f_\sigma^*(s) .$$

where $\Gamma(s)$ is the gamma function [HE].

To be rigorous we need to show that the above summation absolutely converges (see below). For a moment we suppose the above formal identity holds. We will prove that $g^*(s,u)$ is defined (and analytical) for $\Re(s) \in]-1,-\varepsilon[$ with no singularities. In this case the reverse Mellin formula gives [SZ2, HE]

$$\frac{Q_n(u)}{n} = \frac{1}{2i\pi}\int_{c-i\infty}^{c+i\infty} g^*(s,u)n^{-s}ds + O(1) ,$$

with c arbitrary chosen in $]0,\varepsilon[$. A trivial majorization under the sign $\int$ gives the evaluation $Q_n(u)/n = O(n^{-c})$ which shall terminate the proof.

Thus it remains establish absolute convergence of $g^*(s,u)$ for all s such that $\Re(s) \in]-1,\varepsilon[$ and $|u| \leq \beta$. Let us define $h^*(s,u) = \frac{g^*(s,u)}{\Gamma(s)}$. In this perspective, using Taylor

expansion in $p(\sigma)$ of the coefficients A_σ, C_σ and D_σ, we get for *any fixed* s, the uniform evaluations

$$(\log A_\sigma)^{-s} = \left(\frac{p(\sigma)}{1+a_\sigma(1)}\right)^{-s}(1+O(p(\sigma)))$$
$$(-\log(1-p(\sigma))^{-s} = p(\sigma)^{-s}(1+O(p(\sigma)))\ .$$

Thus

$$\frac{(\log A_\sigma)^{-s}-1}{A_\sigma C_\sigma^2} - \frac{(-\log(1-p(\sigma))^{-s}-1}{1-p(\sigma)} =$$
$$= p(\sigma)^{-s}\left[(1+a_\sigma(1))^s(1+O(p(\sigma))-(1+O(p(\sigma))\right]+O(p(\sigma))\ .$$

Using Lemma 7 about $\Pr_k\{1+a_\sigma(1)\le 1+\theta\delta^k\}$, we get also

$$h^*(s,u) = \sum_k \left(\sup\{p^{-\Re(s)}, q^{-\Re(s)}\}|u|\delta\right)^k O(1)$$

which absolutely converges for all values of s such that $\Re(s) < \frac{\log(|u|\delta)}{\log q}$ which is strictly positive. The function $h^*(s,u)$ is now analytically defined for $\Re(s) < \varepsilon$. Since $h^*(0,u)=0$ is trivial, the pole of $\Gamma(s)$ at $s=0$ is canceled in $g^*(s,u)$ which therefore is analytical and shows no singularities for $\Re(s) \in]-1,\varepsilon[$. ■

Finally, to prove Proposition 2 one needs to investigate asymptotics for the independent tries. A copious literature has been devoted to this topic (cf. [FL, JS, RS, SZ1]). Nevertheless, it might be interesting and illuminating to obtain the asymptotics for the depth D_n directly from the generating function (3.9). This can be also regarded as additional verification of our approach. First of all, we note that the Cauchy's formula applied to (3.9) implies

$$ED_n(u) = (1-u)\sum_\sigma u^\sigma p(\sigma)(1-p(\sigma))^{n-1}$$

and therefore, the Mellin transform $D^*(s,u)$ of $ED_n(u)$ becomes

$$D^*(s,u) = (1-u)\sum_\sigma u^\sigma \frac{p(\sigma)}{1-p(\sigma)} p(\sigma)\log(1-p(\sigma))^{-s}\Gamma(s).$$

After simple algebra that takes into account (3.4) from Remark 1 we obtain

$$D^*(s,u) = \frac{(1-u)\Gamma(s)}{1-u(p^{1-s}+q^{1-s})} + O(1).$$

The first term of the above was extensively analyzed for independent tries, and easily leads to our Proposition 2.

4. APPLICATION: SIZE OF SUFFIX TREES

In this section we show how the string-ruler approach can be applied to obtain other characteristics of suffix trees. In particular, our interest lies in computing the *average* number of internal nodes in a suffix tree (in general, any digital trees). Such a characteristic is useful in many applications of suffix trees, most notably to evaluate space complexity of any algorithm that is based on suffix trees.

Let EL_n be the mean size of the suffix tree built over n suffixes. Then, it is not difficult to notice that the average EL_n can be evaluated from $\langle X, \sigma \rangle$ as follows

$$EL_n = \sum_{\sigma} \Pr\{|\langle X, \sigma \rangle| \geq 2\}, \tag{4.1}$$

where, as before, $|\langle X, \sigma \rangle|$ denotes the cardinality of the set $\langle X, \sigma \rangle$ defined in Section 3. The last formula follows from the fact that if σ agrees with at least two suffixes, then these suffixes must agree on σ, and hence there is a node in the associated suffix tree at depth $|\sigma|$. Having this is mind, it is not difficult to prove the following theorem.

THEOREM 15
When $n \geq 2$, we have the identity

$$EL_n = -\sum_{\sigma} \{P_{n,\sigma}(-1) + \frac{dP_{n,\sigma}(v)}{dv}|_{(v=-1)}\} \ . \tag{4.2}$$

Proof: To compute $\Pr\{|\langle X, \sigma \rangle| \geq 2\}$ required for EL_n we need to evaluate the following two probabilities: $\Pr\{|\langle X, \sigma \rangle| = 0\}$ and $\Pr\{|\langle X, \sigma \rangle| = 1\}$. For the former probability, let A_i denote an event that there is mismatch at position i between X and σ, that is, $i \notin \langle X, \sigma \rangle$. Then,

$$\begin{aligned} \Pr\{|\langle X, \sigma \rangle| = 0\} &= \Pr\{\bigcap_{i=1}^{n} A_i\} = 1 - \Pr\{\bigcup_{i=1}^{n} \bar{A}_i\} \\ &= 1 - \sum_{i=1}^{n} (-1)^{i+1} \sum_{|L|=i} \Pr\{\bar{A}_{k_1} \cap \ldots \cap \bar{A}_{k_{|L|}}\} = 1 + P_{n,\sigma}(-1) \ . \end{aligned}$$

On the other hand, the latter probability is equal to

$$\Pr\{|\langle X, \sigma \rangle| = 1\} = \sum_{i=1}^{n} \Pr\{\langle X, \sigma \rangle = \{i\}\} = \frac{dP_{n,\sigma}(v)}{dv}|_{(v=-1)} \ .$$

The theorem is proved by taking into account (4.1) and the above. ■

Now we are ready to evaluate the generating function $L(z)$ of EL_n defined as below

$$L(z) = \sum_{n=1}^{\infty} EL_n z^n \ . \tag{4.3}$$

Using Theorem 15 and (3.6) together with (3.5) we obtain the following corollary.

COROLLARY 16
The generating function $L(z)$ becomes

$$L(z) = z - \sum_{\sigma}\{\frac{p(\sigma)z}{[(1-z)(1+a_\sigma(z))+p(\sigma)z^{\sigma+1}]^2} - \frac{p(\sigma)z}{(1-z)[(1-z)(1+a_\sigma(z)+p(\sigma)z^{\sigma+1}]}\}$$

for $|z| < 1$.■

Remark 5. It is easy to obtain equivalent expression to the above for the size of independent tries. Using the estimates from Remark 4, after some simple algebra, one proves that

$$L^T(z) = z - \sum_{\sigma}\left(\frac{zp(\sigma)}{[1-z+p(\sigma)z]^2} - \frac{zp(\sigma)}{(1-z)(1-z+p(\sigma)z)}\right) \tag{4.4}$$

as expected. ■

The next step is to obtain asymptotics for the suffix tree. To derive them we adopt the same approach as before, namely we prove that the asymptotics for suffix trees are not far away from the asymptotics for independent tries. Therefore, define a new generating function $D(z) = L^T(z) - L^S(z)$. Formulas (4.3) and (4.4) suggest to split this generating function into two natural terms denoted as $D_1(z)$ and $D_2(z)$. But, $D_1(z)$ resembles $Q(z,u)$ for $u = 1$ from Section 3, and repeating the same analysis as before we immediately conclude that the coefficient $d_{n,1}$ at z^n of $D_1(z)$ becomes $d_{n,1} = O(n^{1-\epsilon})$ for some positive ϵ. So, it remains only to consider the second term $D_2(z)$ which becomes

$$D_2(z) = \sum_{\sigma}\left(\frac{zp(\sigma)}{(1-z)[(1-z)(1+a_\sigma(z))+p(\sigma)z^{\sigma+1}]} - \frac{zp(\sigma)}{(1-z)(1-z+p(\sigma)z)}\right) . \tag{4.5}$$

Applying Cauchy's formula to (4.5) one immediately proves that the coefficient $d_{n,2}$ at z^n of $D_2(z)$ can be expressed as below

$$d_{n,2} = \sum_{\sigma}\left(\frac{A_\sigma^{-n}p(\sigma)}{(1-A_\sigma)C_\sigma} - [1-p(\sigma)]^n\right) , \tag{4.6}$$

where, as before A_σ and C_σ denote the root and the first derivatives of the denominator of (4.4) at A_σ.

To obtain asymptotics of (4.6) we proceed as in Section 3. Let $f_\sigma(x)$ be the expression under the sum in (4.6). Note that the Mellin transform of $f_\sigma(x) - f_\sigma(0)e^{-x}$ exists in the strip $\Re(s) > -1$, and

$$f_\sigma^*(s) = \Gamma(s)\left(\frac{(\log A_\sigma)^{-s}-1}{(1-A_\sigma)C_\sigma} - \log(1-p(\sigma))^{-s}\right) . \tag{4.7}$$

Therefore, the Mellin transform $D_2^*(s)$ of $D_2(z)$ becomes $D_2^*(s) = \Gamma(s)\sum_\sigma f_\sigma^*(s)$, and repeating the same analysis as in the proof of Theorem 14 we show that $D_2^*(s)$ is analytical in $]-1,-\epsilon[$, which implies that $d_{n,2} = O(n^{-\epsilon})$. Finally we conclude the following.

COROLLARY 17
For large n the following holds

$$EL_n^S = EL_n^T + O(n^{1-\epsilon}) = \frac{n}{h_1}(1 + P(\log n)) + O(n^{1-\epsilon}) \ . \tag{4.8}$$

where ϵ is some small positive number and $P(\log n)$ is a fluctuating function with a small amplitude [RJ]. ■

In the corollary above we have used the fact that for independent tries $EL_n^T = n/h_1 \cdot (1 + P(\log n)) + O(1)$ [JR, RJ]. Of course, we can obtain this result from (4.4). Indeed, inverting it by the Cauchy's formula and applying Remark 1 we obtain the following

$$\begin{aligned} EL_n &= -\sum_\sigma \{np(\sigma)[1-p(\sigma)]^{n-1} + [1-p(\sigma)]^n - 1\} \\ &= -\sum_\sigma \{n \sum_{\ell=2}^{n} (-1)^{\ell+1} \binom{n-2}{\ell-1} p^\ell(\sigma) + \sum_{\ell=1}^{n} (-1)^\ell \binom{n}{\ell} p^\ell(\sigma)\} \\ &= \sum_{\ell=2}^{n} (-1)^\ell \binom{n}{\ell} \frac{\ell-1}{1-p^\ell-q^\ell} = \frac{n}{h_1} \cdot (1 + P(\log n)) + O(1) \ . \end{aligned}$$

The last asymptotics is a simple consequence of a general asymptotic formula for an alternating sum of the form $\sum_{k=2}^{n} (-1)^k \binom{n}{k} \binom{k}{r} f_k$ for any well-behaved sequence f_k. For details see [SZ2].

5. APPENDIX

We start with a technical lemma that is used in the next proofs.
LEMMA 17
For $|z| < 1$ we have

$$\left| \frac{z}{(R_\sigma(z))^2} \right| \le \frac{|z|}{(1-|z|)^2} \ .$$

Proof: We note that (3.1) and (3.2) imply

$$\frac{z}{R_\sigma(z)^2} = \sum_{n=1}^{\infty} \sum_{i=1}^{i=n} \Pr\{D_n(i) \le k \mid \text{the ith suffix agrees on } \sigma\} z^n \ .$$

The proof follows. ■

Proof of Lemma 7 : We consider all finite strings, σ, of length k. We recall that $F(\sigma)$ is the set of positions that σ overlaps with itself (except trivial position 1). Let i be an integer not greater than k. Then, $i+1 \in F(\sigma)$, if and only if $\sigma = (\sigma_i)^{\lfloor \frac{k}{r} \rfloor} \xi$, where ξ is a prefix of σ_i with $|\xi| = r < i$ and σ_i is the prefix of σ of length i. Thus (for more details see [AS])

$$\Pr_k\{i \in F(\sigma)\} = (p^{\lfloor \frac{k}{r} \rfloor + 1} + q^{\lfloor \frac{k}{r} \rfloor + 1})^r (p^{\lfloor \frac{k}{r} \rfloor} + q^{\lfloor \frac{k}{r} \rfloor})^{i-r} \ .$$

The following evaluation is easy, given that $p \geq q$ (and $p+q=1$),

$$\begin{aligned}\Pr_k\{i \in F(\sigma)\} &\leq (p\,p^{\lfloor \frac{k}{i} \rfloor} + q\,p^{\lfloor \frac{k}{i} \rfloor})^r (p\,p^{\lfloor \frac{k}{i} \rfloor - 1} + q\,p^{\lfloor \frac{k}{i} \rfloor - 1})^{i-r} \\ &= p^{i\lfloor \frac{k}{i} \rfloor + r - i} = p^{k-i} \ .\end{aligned}$$

Thus $\Pr_k\{\inf F(\sigma) \leq \frac{k}{2}\} \leq \sum_{i=1}^{i=k/2} \Pr_k\{i \in F(\sigma)\} \leq \frac{p^{k/2}}{1-p}$. Let σ be such that $\inf F(\sigma) \geq \frac{k}{2}$, then

$$a_\sigma(\rho) \leq \rho^k \sum_{i=k/2}^{i=k} p^i \leq \rho^k \frac{p^{k/2}}{1-p} \ .$$

The proof is completed if one takes $\delta = \sqrt{p}$ and $\theta = (1-p)^{-1}$. ■

Proof of Lemma 9 : We have

$$\frac{p(\sigma)z}{(R_\sigma(z))^2} = \sum_{n=1}^{\infty} \sum_{i=1}^{i=n} \Pr\{\langle X, \sigma \rangle = \{i\}\} z^n \ .$$

Lemma 17 tells us that there is no root for $|z| < 1$. Let us now consider the i-th suffix The probability that i is the only suffix that agrees on σ is less than the probability that i agree on σ and other suffixes j such that $i-j$ is a multiple of k. Since these suffixes do not overlap on k first symbols and the number is greater than $\lfloor \frac{n}{k} \rfloor$, we find

$$\Pr\{\langle X, \sigma \rangle = \{i\}\} \leq p(\sigma)(1-p(\sigma))^{\lfloor \frac{n}{k} \rfloor - 1}$$

and

$$\left| \frac{z}{R_\sigma(z)} \right| \leq \sum_n n|z|^n (1-p(\sigma))^{\lfloor \frac{n}{k} \rfloor - 1}$$

which converges on the appropriate disk. ■

Proof of Lemma 10 : Let σ be a finite string such that $|\sigma| = k > \ell$, and let $i+1 = \inf F(\sigma)$. The integer i is, therefore ,the first position from which σ overlaps with itself. Let us suppose that $i \geq \ell$. Hence, for every complex number, z, such that $|z| \leq \rho$, we have

$$\begin{aligned}|1 + a_\sigma(z)| &\geq 1 - \frac{(p\rho)^\ell - (p\rho)^k}{1-p\rho} \\ &\geq \frac{1 - p\rho - (p\rho)^\ell}{1-p\rho} \ .\end{aligned}$$

Thus the proposition is established in the case where $i \geq \ell$. The alternative case $i < \ell$ is somewhat more delicate to deal with. Let $q = \lfloor \frac{k}{i} \rfloor$. Let σ_i be prefix of σ before position i. Of course we have $\sigma = (\sigma_i)^q \xi$, where ξ is a prefix string of σ_i such that $|\xi| < i = |\sigma_i|$. Then,

$$1 + a_\sigma(z) = \frac{1 - (p(\sigma_i)z^i)^{q+1}}{1 - p(\sigma_i)z} + (p(\sigma_i)z^i)^q a_\xi(z) \ .$$

and

$$|1 + a_\sigma(z)| \geq \frac{1-(p\rho)^{i(q+1)}}{1+(p\rho)^i} - (p\rho)^{qi}\frac{1-(p\rho)^{|\xi|}}{1-p\rho} .,$$

Let j be an integer such that $1 - p\rho - 3(p\rho)^{j\ell} < 1$ and choose such K that $K = (j+1)\ell$ (and $|\sigma| \geq K$). Thus

$$|1 + a_\sigma(z)| \geq \frac{1 - p\rho - 3(p\rho)^{qi}}{1 + p\rho} .$$

Since $qi > k - \ell$, the proof is completed. ■

REFERENCES

[AA] A. Apostolico, The Myriad Virtues of Suffix Trees, *Combinatorial Algorithms on Words,* pp. 8596, Springer-Verlag, ASI F12 (1985).

[AHU] A.V. Aho, J.E. Hopcroft and J.D. Ullman, *The Design and Analysis of Computer Algorithms,* Addison-Wesley (1974).

[AP1] A. Apostolico and F.P. Preparata, Optimal Off-line Detection of Repetitions in String, *Theoretical Computer Science,* 22, 287315 (1983).

[AP2] A. Apostolico, P.F., Preparata, Structural Properties of The String Statistics Problem, *Journal of Computer and Systems Science*, 31, 2, 394-411 (1985).

[AS] A. Apostolico, W. Szpankowski, Self-alignments in Words and Their Applications, Purdue CSD-TR-732 (1987).

[BEH] A. Blumer, A. Ehrenfeucht and D. Haussler, Average Size of Suffix Trees and DAWGS, *Discrete Applied Mathematics*, 24, 37-45 (1989).

[BO] B. Bollobás *Random Graphs*, Academic Press, London (1985).

[CR] M., Crochemore, An Optimal Algorithm for Computing the Repetitions in a Word, *Inf. Proc. Letters* 12, 5, 244-250 (1981).

[DE] L. Devroye, A Note on the Average Depth of Tries, *Computing*, 28, 367-371 (1982).

[DSR] L., Devroye, W. Szpankowski and B. Rais, A note of the height of suffix trees, preprint (1989).

[FL] P. Flajolet, On the Performance Evaluation of Extendible Hashing and Trie Searching, *Acta Informatica,* 20, 345369 (1983).

[FRS] P. Flajolet, M. Regnier and R. Sedgewick, Some Uses of the Mellin Transform Techniques in the Analysis of Algorithms, in *Combinatorial Algorithms on Words*, Springer NATO ASI Ser. F12, 241-254 (1985).

[GO1] L. Guibas and A. Odlyzko Maximal Prefix-Synchronized Codes, *SIAM J. Appl. Math*, 35, 401-418 (1978).

[GO2] L. Giubas and A. Odlyzko, Periods in Strings *Journal of Combinatorial Theory*, Series A, 30, 19-43 (1981).

[GO3] L. Guibas and A. W. Odlyzko, String Overlaps, Pattern Matching, and Nontransitive Games, *Journal of Combinatorial Theory,* Series A, 30, 183-208 (1981).

[HE] P. Henrici, *Applied and Computational Complex Analysis*, John Wiley & Sons (1977).

[JR] P. Jacquet and M. Regnier, Trie Partitioning Process: Limiting Distribution, *Proc. CAAP'86*, Lecture Notes in Computer Science **214**, 194-210 (1986).

[JS] P. Jacquet and W. Szpankowski, Analysis of Tries With Markovian Dependency, in preparation (1989).

[KN] D. Knuth, *The Art of Computer Programming. Sorting and Searching,* Addison-Wesley (1973).

[LO] M. Lothaire, *Combinatorics on Words,* Addison-Wesley (1982).

[LZ] A. Lempel and J. Ziv, On the Complexity of Finite Sequences, *IEEE Information Theory* 22, 1, 75-81 (1976).

[MC] E.M. McCreight, A Space Economical Suffix Tree Construction Algorithm, *JACM*, 23, 262272 (1976).

[ML] M. G. Main and R. J. Lolentz, An All Repetitions in a String, *Journal of Algorithms*, 422-432 (1984).

[PI1] B. Pittel, Asymptotic growth of a class of random trees, *The Annalas of Probability* ,18, 414 - 427 (1985).

[PI2] B. Pittel, Paths in a Random Digital Tree: Limiting Distributions, *Adv. Appl. Prob.*, 18, 139-155 (1986).

[RJ] M. Regnier and P. Jacquet, New Results on the Size of Tries, *IEEE Trans. Information Theory*, 35, 203-205 (1989).

[SZ1] W. Szpankowski, Some Results on V-ary Asymmetric Tries, *Journal of Algorithms,* 9, 224-244 (1988).

[SZ2] W. Szpankowski, The Evaluation of an Alternating Sum with Applications to the Analysis of Some Data Structures, *Information Processing Letters*, 28, 13-19 (1988).

[SZ3] W. Szpankowski, Digital Data Structures and Order Statistics, *Proc. WADS'89*, Lectures Notes in Computer Science **382**, 206-217 (1989).

[WE] P. Weiner, Linear Pattern Matching Algorithms, *Proc. of the 14-th Annual Symposium on Switching and Automata Theory,* 111 (1973).

[ZL] J. Ziv and A. Lempel, A Universal Algorithm for Sequential Data Compression, *IEEE Information Theory*, 23, 3, 337-343 (1977).

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique