AUTOCORRELATION ON WORDS AND ITS APPLICATIONS

Analysis of Suffix Trees by String-Ruler Approach

Third Revision: February 7, 1994

Philippe Jacquet*
INRIA
Rocquencourt
78153 Le Chesnay Cedex
France

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

Abstract

We study in a probabilistic framework some topics concerning the way words can overlap. Our probabilistic models assumes that a word is a sequence of i.i.d. symbols generated from a finite alphabet. This defines the so called Bernoulli model. We investigate the length of a subword that can be recopied, that is, a subword that occurs at least twice in a given word. An occurrence of such repeated substrings is easy to detect in a digital tree called a *suffix tree*. The length of a repeated substring corresponds to the typical depth in the associated suffix tree. Our main finding shows that the typical depth in a suffix tree is asymptotically distributed in the same manner as the typical depth in a digital tree that stores independent keys (i.e., *independent tries*). More precisely, we prove that the typical depth in a suffix tree built from the first n suffixes of a random word is *normally distributed* with the mean asymptotically becoming $1/h_1 \log n$ and the variance $\alpha \cdot \log n$, where h_1 is the entropy of the alphabet, and α is a parameter of the underlying probabilistic model. We prove these results using a novel technique called the *string-ruler* approach. Our results provide new insights into several algorithms on words and data compression schemes. They find direct applications in many facets of science, most notably in molecular biology, coding theory, theory of languages, and design and analysis of algorithms.

`J. Combinatorial Theory, Ser. A`, 1994 (to appear)

*This research was primary 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 INT-8912631, and from Grant AFOSR-90-0107, and in part by Grant R01 LM05118 from the National Library of Medicine.

1. INTRODUCTION

Periodicities, autocorrelations and related phenomena in words are known to play a central role in many facets of science, notably in coding theory, data compression, theory of formal languages, design and analysis of algorithms, and last but not least in molecular sequence comparisons. Several efficient algorithms have been designed to detect the presence of repeated subpatterns and other kinds of avoidable or unavoidable regularities in words [1, 2, 18]. In this paper, we investigate the length of a subword that can be recopied in a random word X, that is, a subword that occurs at least twice in X.

Periodicities, autocorrelations and related phenomena can be also studied on an associated digital tree called a *suffix tree* [1, 2, 3, 30]. A suffix tree is a digital tree that stores suffixes of a given word. In general, a digital tree – that is also called a *trie* – stores a set of words (strings, keys) $\mathcal{W}$ built over a finite alphabet Σ, that is, $\mathcal{W}$ consists of possibly infinite strings of symbols from Σ.

A trie is comprised of branching nodes, called also internal nodes, and external nodes that store the strings from $\mathcal{W}$. We assume that every external node is able to store only one string. The branching policy at any level, say k, is based on the k-th symbol of a string. For example, for a binary alphabet $\Sigma = \{0, 1\}$, if the k-th symbol in a string is "0", then we branch-out left in the trie, otherwise we go to the right. This process terminates the first time we encounter a different symbol between a string that is currently being inserted into the trie and all other strings already in the trie. Then, this new string is stored in a newly generated external node. In other words, the access path from the root to an external node (a leaf of a trie) is the minimal prefix of the information contained in this external node; it is minimal in the sense that this prefix is not a prefix of any other strings. The *depth of a string* is the length of the path from the root to the external node containing this string. The height of a trie is the maximum over all such depths. For more information regarding tries the reader is referred to [2, 17].

A suffix tree is a special trie that is built from suffixes of a *single* word X. We do not compress the trie as in PATRICIA (cf. [17]), that is, in our construction of a suffix tree every edge is labeled by a single character (cf. [1, 2]). Such a suffix tree is also called a noncompact suffix tree.

There is a natural correspondence between lengths of substrings that can be recopied in a word X and depths of suffixes in the associated suffix tree. We found it more convenient to work with suffix trees than the word itself, and most of our main results are presented for such trees. We analyze a random suffix tree in a probabilistic framework called *Bernoulli* *model.* In this model, symbols of a string X are drawn independently from the alphabet Σ, however, it is possible to extend our analysis to some models with dependency between symbols (e.g., Markovian model, see [15]). In passing, we note that a suffix tree has (statistically) *correlated* strings (subwords) which makes the analysis non-trivial.

It is intriguing to compare suffix trees with tries that are built from a set of statistically *independent* strings. We coin a term *independent trie* for the latter digital trees. We prove that that suffix trees do not differ too much (in a probabilistic sense) form independent tries!

The paper is organized as follows. In Section 2, we introduce some measures of correlation among subwords of a word. In particular, we define a self-alignment C_{ij} of any pair of distinct suffixes S_i and S_j of a word X as the length of the longest common prefix of those suffixes. Then, the depth of a *fixed suffix* in the associated suffix tree is the maximum over all self-alignments of the fixed suffix, that is, the depth of the i-th suffix $D_n(i)$ is $D_n(i) = \max\{C_{i1}, C_{i2}, ..., C_{in}\} + 1$. The depth of a *randomly* selected suffix we denote by D_n (cf. definition (2.1) in Section 2). Note that D_n is a random variable even for a *given* word X, and we call it the *typical depth.*

In Section 2 we also present our main results. First of all, our fundamental finding demonstrates that the difference between the distribution functions of the depth in a typical suffix tree and a typical independent trie vanishes to zero as the size of these trees, n, increases to infinity. This allows to apply an abundance of results from independent tries to assess the asymptotic behaviors of suffix trees. In particular, we conclude that the depth of a randomly selected suffix D_n is *normally distributed* for large n with mean $1/h_1 \cdot \log n$ and variance $\alpha \log n$, where h_1 is the entropy of the alphabet and α is a parameter of the underlying probabilistic model. In addition, we show that the average size of a suffix tree (i.e., number of internal nodes) is asymptotically equal to $n/h_1 \cdot (1+P(\log n))$ where $P(\log n)$ is a fluctuating function with a small amplitude.

We delay all proofs till Sections 3 and 4. In Section 3, we prove our main findings concerning the depth D_n. More importantly, this section presents our approach – which seems to be novel – to the analysis of some data structures on strings such as suffix trees, independent tries and so forth. In short, in our new method of attack, we consider an auxiliary string σ called a "ruler", which is used to measure correlations among strings. We call this method the *string-ruler approach.*

The string ruler method is used to show that the depth of a suffix tree *does not* differ significantly from the depth of an independent trie built over the same probabilistic model. We should mention here that such independent tries have been recently extensively analyzed, most notably in [8, 9, 15, 17, 21-24, 26-28]. In particular, Pittel [22], and Jacquet and Régnier [14] derived the limiting distribution for the depth in the independent model, while recently Jacquet and Szpankowski [15] extended this result to the Markovian model.

Finally, in Section 4 we apply the string-ruler approach to prove another result concerning the average size of a suffix tree.

The literature on the analysis of suffix trees is rather scarce, however over last three years a significant progress has been made. Th analysis of the height was initiated by Apostolico and Szpankowski [3], and then continued by Devroye, Szpankowski and Rais [7], and for more general probabilistic models by Szpankowski [29]. In fact, [29] reports asymptotic results for some other suffix tree parameters such as the typical depth, the depth of insertion, and the shortest path. Shields [25] proved almost sure convergence of the external path length in the Markovian model. The size of a suffix tree was investigated by Blumer, Ehrenfeucht and Haussler [4] using a mixture of analytical and simulation tools. In Section 4, we present a rigorous proof of such a result. The limiting distribution of the depth in a suffix tree was left open, and we intend to fill this gap. Preliminary results of this paper were presented in [16].

In passing, we note that using our results from this paper and recent findings of Rais, Jacquet and Szpankowski [23] concerning the limiting distribution of a compact independent tries (known as PATRICIA), we can establish the limiting distribution for the depth in a compact suffix tree (known as a PAT tree).

2. MAIN RESULTS

Let $X = x_1x_2x_3...$ be a string of possibly infinite length built over a finite alphabet Σ of cardinality V, and let $S_i = x_ix_{i+1}...$ be the i-th *suffix* of X. For every off-diagonal pair (i,j) of positions of X, we define the *self-alignment* 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.

Let now n be any fixed integer. We define the *height* H_n of X and the *depth* $D_n(i)$ *of the i-th suffix* of X, as follows

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

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

Furthermore, D_n for a word X is defined as the depth of a randomly selected suffix among the first n suffixes of X. Clearly, we have

$$\Pr\{D_n \le k\} = \frac{1}{n}\sum_{i=1}^{n} \Pr\{D_n(i) \le k\} \tag{2.1c}$$

and we call D_n the typical depth.

Intuitively, $H_n - 1$ is the maximum possible length of 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, $i < j \le n$, such that the occurrence of Z starting at j can be fully recopied from the occurrence starting at i. The depth $D_n(i)$ represents the length of the longest substring started at position i of X that can be recopied, while the depth D_n is the length of a typical recopied substring. The height H_n and the typical depth D_n express structural correlations among the substrings of the word X. Such correlations play a crucial role in many combinatorial and algorithmic constructions, and our above definitions are somewhat reminiscent of notions that have already appeared in the literature, most notably in [10-12, 19, 31].

We illustrate these definitions in the example below.

EXAMPLE 2.1. *Self-alignment matrix*

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 $\mathbf{C} = \{C_{ij}\}_{i,j=1}^{5}$ is as follows.

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

From $\mathbf{C}$ and the expressions (2.1), we obtain $H_n = 3$ and $D_5(1) = 3$, $D_5(2) = 2$, $D_5(3) = 3$, $D_5(4) = 3$ and $D_5(5) = 3$. Moreover, for *given* X the random variable D_5 is distributed as follows: with probability $1/5$ we have $D_5 = 2$, and with probability $4/5$ we have $D_5 = 3$. In passing we note that if X is random, then $D_n(i)$ becomes a random variable, too. $\square$

For every self-alignment matrix $\mathbf{C}$, we can construct the associated suffix tree built from the first n suffixes of X. As explained above, it consists of branching (internal) nodes and external nodes. At a branching node at level k, we look at the k-th symbol of all suffixes, and – for example, for $\Sigma = \{a, b\}$ – depending whether this symbol is a or b we move *right* or *left* down into the suffix tree. At the first time two suffixes differ (split) we construct two external nodes that contain these suffixes. This is illustrated in the next example.

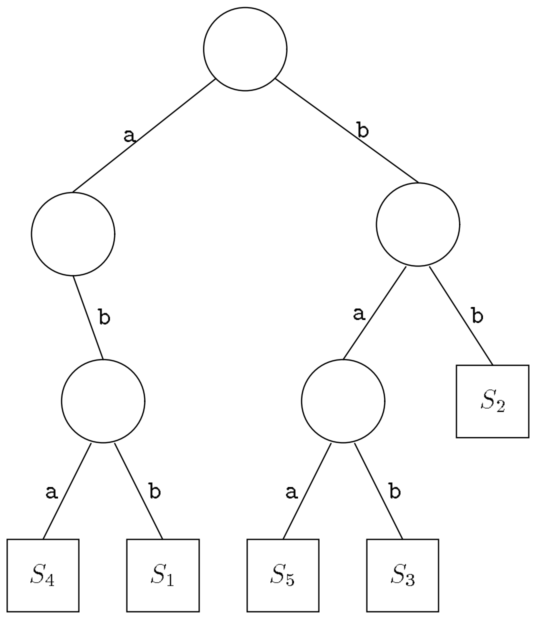

Figure 1: Suffix tree from Example 2.1

EXAMPLE 2.2. *Suffix tree for X from Example 2.1*

Let, as in Example 2.1, $X = abbabaa...$ and $n = 5$. Then, the associated suffix tree is presented in Figure 1, where circles represent branching nodes and squares are external nodes. The depths $D_n(i)$, D_n and the height H_n of the suffix tree are computed in Example 2.1. □

In this paper, we present a probabilistic analysis of the depth D_n in a probabilistic framework known as *Bernoulli model.* We assume: *symbols of X are drawn independently from Σ, and the i-th symbol of Σ occurs in any position of X with probability p_i where $i = 1, 2, ..., V$ and $\sum_{i=1}^{V} p_i = 1$.*

Let us consider the depth of a fixed suffix, say the first one. According to (2.1b) we have $D_n(1) = \max_{2\le j\le n}\{C_{1j}\} + 1$. 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* [5] one immediately proves

$$\Pr\{D_n(1) > k\} = \sum_{r=1}^{n-1} (-1)^{r+1} \sum_{i_1,\ldots,i_r} \Pr\{C_{1,i_1} \ge k, ..., C_{1,i_r} \ge k\}, \qquad (2.2)$$

where the i_j's are distinct and $2 \le i_j \le n$ for every $1 \le j \le r$. Interestingly enough, the distribution of D_n depends on all terms in the right-hand side (RHS) of (2.2). This is partially due to the appearance of the alternating sum in (2.2). We conclude, therefore, that one needs a very precise estimate for the joint distribution $\Pr\{C_{1,i_1} \geq k, ..., C_{1,i_r} \geq k\}$ in order to assess the distribution of D_n. This seems to be a challenging problem.

To illustrate our previous point, we apply (2.2) to independent tries, as we did in [15] for the Markovian model in which the occurrence of the next symbol in X depends on the previous symbol (cf. [28]). For tries, the alignment C_{ij} is defined as the length of a common prefix of the ith and jth independent strings. We have the following result. In the Bernoulli model, for every r-tuple $(i_1, ..., i_r)$ the joint distribution from (2.2) becomes [15, 28]

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

where hereafter, for simplicity of presentation, we assumed a binary alphabet with $p_1 = p$ and $p_2 = q = 1 - p$. From (2.2) and (2.3) we easily obtain the generating function of the depth, the average value, etc. For example, the generating function $Ez^{D_n} = Ez^{D_n(i)}$ for all $i = 1, \ldots, n$, and it becomes 9cf. [15])

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

Asymptotics of (2.4) were extensively studied in the past through the Mellin transform [8, 9, 14, 15, 17, 23, 24, 26, 27] and through probabilistic methods [6, 22]. For instance, the average depth ED_n is equal to $ED_n = 1/h_1 \cdot \log n + 1/h_1 \cdot (\gamma + h_2/(2h_1)) + P(\log n) + O(n^{-1})$, 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. A similar technique works for a Markovian model in which symbols depend in a Markovian fashion but strings are still independent (cf. [15]).

How one can use the above approach to analyze the typical depth in suffix trees? We note that (2.2) holds for any tree since it is based only on the inclusion-exclusion formula. The independence between strings was used to derive the joint distribution of the (self)-alignments in (2.3). In the suffix tree case we must cope with overlapping, and this causes some problems, especially since we need a very precise estimate of the joint distribution of self-alignments.

To illustrate some difficulties arising in the evaluation of this joint distribution, consider the following probability $\Pr\{C_{1,5} > 10, C_{1,8} > 10, C_{1,20} > 10\}$. One can convince himself that this probability is equal to $p^{29} + q^{29}$. This is quite different than (2.3). However, when suffixes are separated by at least k symbols, then, in the Bernoulli model, they are independent on their first k symbols. More precisely, let us define a set of integers $r_1, r_2, ..., r_\ell$ such that for any $i \leq \ell - 1$ the following holds $r_{i+1} - r_i > k$ and $r_1 > k$. Then, (2.3) is true in the following sense

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

It is plausible, however, that the probability of overlapping is rather small and one can expect that formula (2.4) is still approximately true. Then, it is reasonable to expect identical asymptotics for the independent trie and the suffix tree models. Nevertheless, it is rather hard to justify this idea rigorously due to the fact that (2.2) contains an alternating sum. In the next section, we adopt a quite different and novel approach to circumvent this difficulty.

Now we are in a position to summarize our main results. Our major finding deals with a comparison between the independent tries 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 $F_n^T(k) = \Pr\{D_n^T \leq k\}$ and $F_n^S(k)$, respectively. Note that for independent tries $\Pr\{D_n^T \leq k\} = \Pr\{D_n^T(i) \leq k\}$ for any key i, while for suffix tree we have $\Pr\{D_n^S \leq k\} = \frac{1}{n}\sum_{i=1}^n \Pr\{D_n^S(i) \leq k\}$ as in (2.1c). 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 suffix trees and some other related data structures. It basically says that suffix trees do not differ too much from independent tries. But, tries have been analyzed extensively over last few years, 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 [14, 24], Pittel [22] and Szpankowski [26] 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 $h_1 = -\sum_{i=1}^V p_i \log p_i$ *and* $h_2 = \sum_{i=1}^V p_i^2 \log p_i$, *and* $P_1(x), P_2(x)$ *are fluctuating fluctuating functions with small amplitudes, and an explicit formula for the constant* C *can be found in [24]. In the symmetric case, i.e.,* $p_1 = p_2 = ... = p_V = 1/V$, *the variance becomes*

$$varD_n = \frac{\pi^2}{6\log^2 V} + \frac{1}{12} + P_3(\log n) + O\left(\frac{1}{n^\epsilon}\right) \ . \tag{2.7c}$$

(ii) *For the asymmetric model* $(D_n - ED_n)/\sqrt{varD_n}$ *is asymptotically normal with mean zero and variance one, that is, for all* $x \in R$

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

and for all integer m

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

For the symmetric case, one proves that uniformly in $x \ge 0$

$$\lim_{n\to\infty} \sup_x \mid \Pr\{D_n \le x\} - \exp(-nV^{-x}) \mid = 0 \tag{2.7d}$$

uniformly in $x \ge 0$. ■

In some applications, size of a 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 a suffix tree determines space requirements, and therefore the space complexity of any algorithm based on suffix trees, while depth D_n is rather responsible for the time complexity of a string algorithm. The next proposition presents one result in this direction, namely the average size EL_n of a suffix tree 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

$$EL_n^S = \frac{n}{h_1}(1 + P_4(\log n)) + o(n) \ , \tag{2.8b}$$

where $P_4(\log n)$ is a fluctuating function with a small amplitude. ■

Finally, to get some idea about the accuracy of our asymptotics (in particular Proposition 1) we have performed some simulation studies which are discussed in [16]. These results confirmed – as expected – our theoretical findings, and in addition they show good accuracy of the asymptotics even for small values of n.

3. ANALYSIS THROUGH STRING-RULER APPROACH

In this section, we prove our main result (i.e., Proposition 1) using a novel method called the *string-ruler* approach. To the best of our knowledge, this new technique resembles only the work of Guibas and Odlyzko [10, 11, 12] (see also [21]). In fact, the method described in Section 3.1 is used to analyze independent tries (cf. Section 3.2) as well as suffix trees (cf. Section 3.3).

Before we plunge into a detailed analysis, let us give a brief overview of our approach. In Section 2 we demonstrated that any analysis of the depth D_n in a digital tree, in particular, in a suffix tree needs a very precise or even exact evaluation of the joint distribution of the self-alignments (e.g., see (2.3) for independent tries). Such an evaluation for suffix trees is very complicated due to strong correlations among overlapping suffixes. Therefore, to circumvent it (in fact, to hide it in a generating function form), we suggest a different, more combinatorial approach. We consider a set of finite strings σ that are used as "rulers" to measure correlation between strings. For example, to evaluate the self-alignment between the i-th suffix S_i and the j-th suffix S_j we first compute the alignment 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}. This – although it looks more complicated than necessary – is the right approach as we shall prove below. We should also stress that this methodology gives a unified approach for analyzing some other digital structures (e.g., independent tries, digital search trees, direct acyclic word graphs (DAWG) [4], etc.).

Using the above idea we shall compute respectively in Sections 3.2 and 3.3 the generating functions of the depths for an independent trie (easy) and a suffix tree (difficult!). These two generating functions are asymptotically compared to show that they do not differ too much for large n (cf. Section 3.4). This will lead to our main result Proposition 1.

It might be worthwhile to point out that along the lines of our proof, we in fact explore autocorrelation properties of strings. This may find many other applications in combinatorics on words, e.g., in squares of strings, in bi-prefix strings, and so forth [1, 3, 18].

3.1 String-Ruler Approach: General Case

As discussed before, a trie is a digital tree built from n possibly infinite strings, say $X_1, \ldots, X_n$. These strings might be statistically dependent or independent; symbols within a string might be statistically correlated or not, etc. In other words, in this section we do *not* assume any specific probabilistic model, and results in this section hold for any probabilistic model.

Let us define for any string σ and a set of $n \geq 1$ strings $\{X_1, \ldots, X_n\}$ a quantity $\langle\sigma\rangle_n$ as follows

$$\langle\sigma\rangle_n = \{i : \ 1 \leq i \leq n \ \text{ and } \sigma \text{ is a prefix of } \ X_i\} \ .$$

In words, $\langle\sigma\rangle_n$ is a set of indices of elements of $\{X_1, \ldots, X_n\}$ for which σ is a prefix. Let C_{ij} be the alignment between the ith and the jth strings, that is, the length of the longest common prefix of X_i and X_j. Then, the depth of the ith string $D_n(i)$ is the depth of the ith string in a trie built over the set $\{X_1, \ldots, X_n\}$, and can be defined using the alignments C_{ij} as in (2.1b). Having in mind the construction of a trie, we immediately establish the following relationships,

$$\{D_n(i) > k\} \qquad \Longleftrightarrow \qquad \exists\sigma \ \exists j \leq n \ : \ |\sigma| = k \text{ and } j \neq i \ : \ \{i, j\} \subset \langle\sigma\rangle_n \tag{3.1a}$$

$$\{D_n(i) \leq k\} \qquad \Longleftrightarrow \qquad \exists\sigma \ : \ |\sigma| = k \text{ and } \langle\sigma\rangle_n = \{i\} \ . \tag{3.1b}$$

Now, consider the set of all strings σ of fixed length $|\sigma| = k$, where $|\sigma|$ denotes the length of σ. Note that the events $\langle\sigma\rangle_n = \{i\}$ and $\langle\sigma'\rangle_n = \{i\}$ are disjoint for distinct strings σ and σ' such that $|\sigma| = |\sigma'| = k$. Hence, we can write

$$\Pr\{D_n(i) \leq k\} = \sum_{|\sigma|=k} \Pr\{\langle\sigma\rangle_n = \{i\}\} \tag{3.1c}$$

where the sum above is over all strings σ of length k. The above provides another characterization of the depth $D_n(i)$. The example below illustrates what we have done so far.

EXAMPLE 3.1 *Depth in a trie as a function of* $\langle\sigma\rangle_n$.

Let us consider a trie built from the following six strings: $X_1 = abaaaba\ldots$, $X_2 = abbabab\ldots$, $X_3 = baaabaa\ldots$, $X_4 = abaabab\ldots$, $X_5 = bbaaaaa\ldots$ and $X_6 = aaaaaba\ldots$. What is $D_6(1)$? Note that:

- $\langle a\rangle_6 = \{1,2,4,6\}$, so $D_6(1) > 1$,
- $\langle ab\rangle_6 = \{1,2,4\}$, so $D_6(1) > 2$,
- $\langle aba\rangle_6 = \{1,4\}$, so $D_6(1) > 3$,
- $\langle abaa\rangle_6 = \{1,4\}$, so $D_6(1) > 4$,
- $\langle abaaa\rangle_6 = \{1\}$, so $D_6(1) \le 5$, and therefore $D_6(1) = 5$.

The reader can check that the depth of the first string in the trie built from the above six strings is really equal to 5. $\Box$

The random variable D_n is the depth of a *randomly* selected string in a trie built from n random strings. If $E[u^{D_n(i)}]$ denotes the ordinary generating function of the depth $D_n(i)$ of the i-th suffix, then the generating function $E[u^{D_n}]$ of D_n becomes

$$E[u^{D_n}] = \frac{1}{n}\sum_{i=1}^{n} E[u^{D_n(i)}] \ , \tag{3.2}$$

and this can be viewed as an alternative definition of D_n (cf. (2.1c)).

It turns out, however, that for our analysis, it is more convenient to work with the bivariate generating function $D(z,u)$ of $E[u^{D_n}]$ defined as

$$D(z,u) = \sum_{n=0}^{\infty} n\, E[u^{D_n}] z^n \ .$$

We express $D(z,u)$ in terms of some other generating functions defined in sequel.

For this, we need another representation of the right-hand side (RHS) of (3.1c). Define an event $A_j = \{j \in \langle\sigma\rangle_n\}$ and the complementary event $\bar{A}_j = \{j \notin \langle\sigma\rangle_n\}$. Then $\Pr\{\langle\sigma\rangle_n = \{i\}\} = \Pr\{\bigcap_{j\neq i} \bar{A}_j \cap A_i\}$. Noting that $\Pr\{\bigcap_{j\neq i} \bar{A}_j \cap A_i\} + \Pr\{\overline{\bigcap_{j\neq i} \bar{A}_j} \cap A_i\} = \Pr\{A_i\}$, and applying the inclusive-exclusive formula [5] to the second probability of the left-hand side (LHS) of the previous expression, we obtain

$$\Pr\{\bigcap_{j\neq i} \bar{A}_j \cap A_i\} = \Pr\{A_i\} - \sum_{j=1}^{n-1}(-1)^{j+1} \sum_{\{i_1,\dots,i_j\}} \Pr\{\bigcap_{k=1}^{j} A_{i_k} \cap A_i\} \ ,$$

where $\{i_1,\dots,i_j\}$ is a j-tuple of *distinct* elements from $\{1,\cdots,n\} - \{i\}$. Note also that $\Pr\{\bigcap_{k=1}^{j} A_{i_k} \cap A_i\} = \Pr\{\{i_1,\dots,i_j,i\} \subset \langle\sigma\rangle_n\}$.

To simplify our last display, let $\mathcal{L}$ be a finite set of integers. We define $P(\mathcal{L},\sigma)$ as the probability of the event "$\mathcal{L} \subset \langle\sigma\rangle_n$", that is, $P(\mathcal{L},\sigma)$ is the probability that σ is a prefix of those strings whose indices belong to the set $\mathcal{L}$. Let $m(\mathcal{L})$ and $|\mathcal{L}|$ denote, respectively, the largest element of $\mathcal{L}$ and the size of $\mathcal{L}$. Then, it is easy to see that for $\mathcal{L} = \{i, i_1, \ldots, i_{j-1}\}$, the above implies the following

$$\Pr\{\langle\sigma\rangle_n = \{i\}\} = \sum_{j=1}^{n} (-1)^{j+1} \sum_{\substack{|\mathcal{L}|=j \\ m(\mathcal{L})\leq n \ , \ i\in\mathcal{L}}} P(\mathcal{L}, \sigma) \ . \tag{3.3}$$

To simplify the above notation, hereafter we consider only such sets $\mathcal{L}$ that $m(\mathcal{L}) \leq n$.

We can generalize (3.3) to include the empty string σ that is further denoted as $*$. We adopt the convention that $\langle * \rangle_n = \{1, 2, \ldots, n\}$, therefore $P(\mathcal{L}, *) = 1$ for every set $\mathcal{L}$. Then (3.3) holds with the first sum starting from $\{j = 0\}$.

Now, we can compute the bivariate generating function $D(z, u)$. From the definition (2.1c) and (3.1c) we have

$$n \cdot \Pr\{D_n \leq k\} = \sum_{i=1}^{n} \Pr\{D_n(i) \leq k\} = \sum_{i=1}^{n} \sum_{|\sigma|=k} \Pr\{\langle\sigma\rangle_n = \{i\}\} \ .$$

Using (3.3) we can simplify the above by noting that the inner sum of (3.3) after some modifications can be rewritten as

$$\sum_{i=1}^{n} \sum_{\substack{|\mathcal{L}|=j \\ i\in\mathcal{L}}} P(\mathcal{L}, \sigma) = j \cdot \sum_{|\mathcal{L}|=j} P(\mathcal{L}, \sigma) \ .$$

Hence, we obtain

$$n \cdot \Pr\{D_n \leq k\} = \sum_{|\sigma|=k} \sum_{j=0}^{n} (-1)^{j+1} j P_j(\sigma) \tag{3.4}$$

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

Finally, in Theorem 1 below we show that the generating function Eu^{D_n} can be expressed in terms of the following generating functions

$$P_{n,\sigma}(v) = \sum_{j=0}^{n} P_j(\sigma) v^j = \sum_{\{\mathcal{L}:\ m(\mathcal{L})\leq n\}} P(\mathcal{L}, \sigma) v^{|\mathcal{L}|} \qquad , \qquad P_\sigma(z, v) = \sum_{n=1}^{\infty} P_{n,\sigma}(v) z^n \ .$$

Note that for $\sigma = *$ we have $P_j(*) = \binom{n}{j}$, and consequently $P_{n,*}(v) = (1 + v)^n$ as well as $P_*(z, v) = (1 + v)z/(1 - (1 + v)z)$ (see also Lemma 3 below for another representation of $P_\sigma(z, v)$). Hence, after recognizing in the the right-hand side of (3.4) the partial derivative of $P_{n,\sigma}(v)$ with respect to v at $v = -1$, we finally obtain our main result of this subsection.

THEOREM 1.

For $n \geq 1$ we have the identity

$$E[u^{D_n}] = \frac{(1-u)}{n} \sum_{\sigma} u^{|\sigma|} \frac{d}{dv} P_{n,\sigma}(v)|_{(v=-1)}$$

for $|u| < 1$, where the sum above should be interpreted as $\sum_{\sigma} f(\sigma) = \sum_{k=0}^{\infty} \sum_{|\sigma|=k} f(\sigma)$ for any function $f(\cdot)$ defined on strings. ■

As a simple consequence of Theorem 1, we get the following corollary for the bivariate generating function $D(z,u)$.

Corollary 2
For $|u| < 1$, the following holds

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

for $|z| < 1$. ■

Finally, in some analyses (e.g., for suffix trees) we need another form of the generating function $P_{\sigma}(z,v)$ which is presented below.
Lemma 3.
We have the identity for $|z| < 1$

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

with

$$S_{\sigma}(z,v) = \sum_{\mathcal{L}} P(\mathcal{L},\sigma) z^{m(\mathcal{L})} v^{|\mathcal{L}|} \ , \tag{3.5}$$

where $\sum_{\mathcal{L}}$ is a sum over all finite sets $\mathcal{L}$ of positive integers.
Proof : By rearranging the terms in the summation of $P_{\sigma}(z,v)$ we obtain

$$P_{\sigma}(z,v) = \sum_{n=1}^{\infty} z^n \sum_{\{\mathcal{L}:\ n \geq m(\mathcal{L})\}} P(\mathcal{L},\sigma) v^{|\mathcal{L}|} = \sum_{\mathcal{L}} P(\mathcal{L},\sigma) v^{|\mathcal{L}|} \sum_{n=m(\mathcal{L})}^{\infty} z^n \ ,$$

and this is the desired identity. ■

Remark 1. It is worth pointing out that the notation $\sum_{\sigma}$, which is extensively used throughout the entire section, expresses the sum over all finite strings. For example, for a binary alphabet $\Sigma = \{a,b\}$, there are 2^k distinct strings of length k. 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 to the following identities

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

and

$$\sum_{\sigma} x^{|\sigma|_a} \cdot y^{|\sigma|_b} = \sum_{k=0}^{\infty} \sum_{|\sigma|=k} x^{|\sigma|_a} \cdot y^{|\sigma|_b} = \frac{1}{1-x-y} \ ,$$

for suitable values of complex numbers x and y. In passing, we note that the string-ruler σ is *not* a random string but it rather belongs to a finite set of strings. □

3.2 Analysis of Independent Tries

In this section we assume that: (i) the strings X_i $(1 \leq i \leq n)$ are statistically independent; (ii) symbols within a string are generated according to the Bernoulli model. In addition, for simplicity of presentation, we restrict our attention to a binary alphabet $\Sigma = \{a, b\}$ with p (resp. q) denoting the probability of a (resp. b) occurring. We construct an *independent trie* from these n strings within the framework of the Bernoulli model. Let D_n^T denote the depth in such a trie. Using our approach from the previous section, we shall derive the generating functions $E[u^{D_n^T}]$ and $D^T(z, u)$ for independent tries which will be further compared with the generating function of a suffix tree.

To accomplish our task, we need a formula for the generating functions $P_\sigma(z)$ and $P_\sigma(z, v)$. But, in the Bernoulli model we have $P(\mathcal{L}, \sigma) = (p(\sigma))^{|\mathcal{L}|}$ where $\mathcal{L}$ is a finite subset of positive integers. Then, $P_j(\sigma) = \sum_{|\mathcal{L}|=j} P(\mathcal{L}, \sigma) = \binom{n}{j} p^j(\sigma)$ where $p(\sigma)$ is defined as follows. For any finite string σ, the function $p(\sigma)$ is the product $p^{|\sigma|_a} q^{|\sigma|_b}$ where $|\sigma|_a$ is the number of a's in σ and $|\sigma|_b$, the number of b's in σ. [1] Finally, the above leads to $P_{n,\sigma}(v) = (1 + p(\sigma)v)^n$, and finally

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

Then, by Corollary 2 we obtain

$$D^T(z, u) = (1-u) \sum_{\sigma} u^{|\sigma|} \frac{p(\sigma) z}{[1 - z + p(\sigma) z]^2} \ . \tag{3.6}$$

We shall use this formula to compare independent tries with suffix trees which are analyzed in the next section.

[1] We note that although σ is not a random string, $p(\sigma)$ can be viewed as the probability of σ occuring at a given position in a random string X. This is a simple consequence of the fact that $\sum_{|\sigma|=k} p(\sigma) = 1$. Although this notation may cause some confusion, we decided to adopt it because of the latter property.

3.3 Analysis of Suffix Trees

In this section we also adopt the Bernoulli model (with the binary alphabet), however, hereafter we build a suffix tree from the first n suffixes of a random word X. Note that the set of suffixes are *statistically dependent*. This will cause some complications in our analysis.

First of all, we found convenient to introduce the correlation symbol $\langle X, \sigma \rangle$ which represents those (indices of) suffixes of X for which σ is a prefix. For example, if $X = baabbabaaa$ and $\sigma = \{baab\}$, then $\langle X, \sigma \rangle = \{1\}$, but $\langle \sigma, X \rangle = \{1, 4\}$. Note that we can express the correlation $\langle X, \sigma \rangle$ in terms of $\langle \sigma \rangle_n$ as follows

$$\langle \sigma \rangle_n = \langle X, \sigma \rangle \cap \{1, \ldots n\} \ .$$

Now, we apply our approach from Section 3.1 to analyze the depth D_n of a suffix tree. From Corollary 2 we know that $D(z, u)$ depends on the generating function $P_\sigma(z, v)$. This generating function is a function of a string-ruler σ: More formally, it is a function of the probability $P(\mathcal{L}, \sigma) = \Pr\{\mathcal{L} \subset \langle X, \sigma \rangle\}$. This suggests that $P_\sigma(z, v)$ depends on the structure of $\mathcal{L}$ as well as on some autocorrelation properties of σ itself. This is particularly true when there exists a subset of $\mathcal{L}$ consisting of positions separated by less than $|\sigma| = k$ (such a subset will be further called a k-cluster). We illustrate this in the following example.

EXAMPLE 3.2. *Autocorrelation of σ and k-clusters*

Let $X = bbabaabaabaababbbabaaba$ and $\sigma = abaaba$, so $|\sigma| = k = 6$. Note that $\mathcal{L} = \langle X, \sigma \rangle = \{3, 6, 9, 18\}$, and the autocorrelation set of σ denoted as $\langle \sigma, \sigma \rangle$ becomes $\langle \sigma, \sigma \rangle = \{1, 4, 6\}$. There is a relationship between $\langle X, \sigma \rangle$ and $\langle \sigma, \sigma \rangle$, namely those positions of $\langle X, \sigma \rangle$ that are separated by less than $k = 6$ positions – the so called *k-cluster* – are inherently correlated to $\langle \sigma, \sigma \rangle$. Indeed, in our case a k-cluster is $\{1, 3, 6, 9\}$. This cluster is a direct consequence of the fact that the autocorrelation set $\langle \sigma, \sigma \rangle$ includes the position $\{4\}$, so a k-cluster of X with respect to σ can be created if and only if $\langle \sigma, \sigma \rangle - \{1\}$ is nonempty. It should be also clear that all difficulties in evaluating the probability $P(\mathcal{L}, \sigma)$ arise from the necessity of taking into account such k-clusters. $\Box$

In order to investigate k-clusters (for formal definition see below) we need to study some autocorrelation properties of a string-ruler. Let $\mathcal{F}_\sigma = \langle \sigma, \sigma \rangle - \{1\}$ be the autocorrelation set of a string σ, that is, $i \in \mathcal{F}_\sigma$ if and only if σ overlaps with itself from position i. For simplicity we omit the trivial position $i = 1$. For instance, in Example 3.2, $\mathcal{F}_\sigma = \{4, 6\}$. Furthermore, we define the *autocorrelation polynomial* $a_\sigma(z)$ of σ as

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

where σ_{i-1} denotes the prefix of σ of length $i-1$, and $p(\sigma)$ was defined already in the previous section. In particular, the autocorrelation polynomial for $\sigma = abaaba$ becomes $a_\sigma(z) = p^2 q \cdot z^3 + p^3 q^2 \cdot z^5$.

Remark 2. Our definition of the autocorrelation polynomial resembles the autocorrelation function introduced by Guibas and Odlyzko [10, 11, 12], however, in our case the autocorrelation polynomial is additionally weighted by the probability $p(\sigma_{i-1})$. □

Now we are ready to deal with a k-cluster with respect to a string-ruler σ of length k, and find a relationship between the generating function $P_\sigma(z,v)$ and the generating function of $|\sigma| = k$-clusters. A k-cluster can be viewed as a collection of suffixes that are separated by less than k symbols. More formally, we define a k-cluster $\mathcal{C}$ as a finite set of integers which satisfies the following properties: $\mathcal{C}$ contains the integer 1 and either it contains no other element or $\mathcal{C}$ can be considered as an increasing sequence of integers such that the difference between any two consecutive elements is strictly smaller than k.

EXAMPLE 3.3 *Continuation of Example 3.2*

Assume X and σ as in Example 3.2, and define a k-cluster as $\mathcal{C} = \{1,4,7\}$. Then, $\langle X, \sigma\rangle - \{18\} = \mathcal{C} + 2$, where by $\mathcal{C} + i$ we mean that every element of $\mathcal{C}$ is increased by i. Another k-cluster is $\mathcal{C}' = \{1,4\}$. Note that these two k-clusters can be represented as $\mathcal{C} = \{1\} \cup \{\mathcal{C}' + 3\}$. We use this property to derive the generating function $C_\sigma(z,v)$ that enumerates k-clusters with respect to σ. □

Let $C_\sigma(z,v)$ be a bivariate generating function defined as

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

where $P(\mathcal{C},\sigma) = \Pr\{\mathcal{C} \subset \langle X,\sigma\rangle\}$ and $\sum_{\mathcal{C}}$ denotes the summation over all possible k-clusters such that $m(\mathcal{C}) \le n$. Note that for any i we also have $\Pr\{\mathcal{C}+i \subset \langle X,\sigma\rangle\} = \Pr\{\mathcal{C} \subset \langle X,\sigma\rangle\}$. This observation is used in the proof of our next result. □

THEOREM 4.
The generating function $C_\sigma(z,v)$ becomes

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

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

PROOF : A nonempty k-cluster $\mathcal{C}$ such that $P(\mathcal{C},\sigma) > 0$ is either $\{1\}$ or of the form $\{1\} \cup \{\mathcal{C}' + i - 1\}$, where $\mathcal{C}'$ is another cluster and $i \in \mathcal{F}_\sigma$ (see example above). Note that

$P(\mathcal{C}'+i-1,\sigma) = P(\mathcal{C}',\sigma)$. In addition, for every prefix σ_{i-1} of σ we have $P(\mathcal{C},\sigma) = p(\sigma_{i-1})P(\mathcal{C}',\sigma)$, and also $m(\mathcal{C}'+i-1) = m(\mathcal{C}')+i-1$. Hence, for given $i \in \mathcal{F}_\sigma$ we obtain

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

Furthermore, we trivially have $p(\{1\},\sigma)z^{m(\{1\})}v^{|\{1\}|} = p(\sigma)z\,v$, thus enumerating all k-clusters and all positions of $\mathcal{F}_\sigma$, we finally obtain $C_\sigma(z,v) = p(\sigma)z\,v + a_\sigma(z)vC_\sigma(z,v)$, which completes the proof. ■

Now, we are ready to prove a relationship between the generating functions $P_\sigma(z,v)$ and $C_\sigma(z,v)$. The following example illustrates the main idea of our next theorem.

EXAMPLE 3.4 *Factorization of $\mathcal{L}$ into k-clusters*

Let $X = bbababababbbaaabababababbbababb...$ and $\sigma = abab$. Note that $\mathcal{L} = \langle X,\sigma\rangle = \{3,5,13,15,17,23\}$. We have three $k=4$-clusters: $\mathcal{C}_1 = \{1,3\}$, $\mathcal{C}_2 = \{1,3,5\}$ and $\mathcal{C}_3 = \{1\}$ such that the following factorization of $\mathcal{L}$ holds

$$\mathcal{L} = \{\mathcal{C}_1+2\} \cup \{\mathcal{C}_2+12\} \cup \{\mathcal{C}_3+22\} \tag{3.7}$$

where, as before, $\{\mathcal{C}+i\}$ is a translation of $\mathcal{C}$ modulo i. □

THEOREM 5

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

$$P_\sigma(z,v) = \frac{1}{(1-z)} \cdot \frac{C_\sigma(z,v)}{1-z-z^{k-1}C_\sigma(z,v)} \tag{3.8}$$

where $k = |\sigma|$.

PROOF : As in (3.7), a k-factorization of $\mathcal{L}$ is defined as a partition of $\mathcal{L}$ into a certain number of k-clusters that are in their minimal form. That is,

$$\mathcal{L} = \{\mathcal{C}_1+i_1\} \cup \{\mathcal{C}_2+i_2\} \cup \cdots \cup \{\mathcal{C}_d+i_d\} \ ,$$

where $\mathcal{C}_j$ is a k-cluster and $i_1,\cdots,i_d$ are suitable integers. Note that for the Bernoulli model, the above implies

$$P(\mathcal{L},\sigma) = P(\mathcal{C}_1,\sigma)P(\mathcal{C}_2,\sigma)\cdots P(\mathcal{C}_d,\sigma) \ .$$

Furthermore, it is easy to see that $\mathcal{L}$ can be viewed on the integer axis as an increasing sequence of k-clusters such that two consecutive ones are separated by more than $k-1$ units. Let $i-1 = \min\{i_1,\ldots,i_d\}$ and let $\mathcal{C}$ be the corresponding k-cluster associated with i in the factorization of $\mathcal{L}$. It is obvious that $i = \min(\mathcal{L})$, and

(i) if $d = 1$, then $\mathcal{L}$ is a k-cluster itself, modulo translation of $i-1$ unit, i.e. $\mathcal{L} = \mathcal{C} + i - 1$;

(ii) if $d > 1$, then the remaining elements of $\mathcal{L}$ that are not in $\mathcal{C} + i - 1$ are necessarily greater than or equal to $m(\mathcal{C} + i - 1) + k = m(\mathcal{C}) + i - 1 + k$; therefore the following partition holds [2]: $\mathcal{L} = \{\mathcal{C} + i - 1\} \cup \{\mathcal{L}' + i - 1\} + m(\mathcal{C}) - 1 + k$ where $\mathcal{L}'$ is another finite set of integers (which will lead to a k-factorization with $d - 1$ clusters).

Part (i) gives

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

while (ii) leads to

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

Finally, summing over all sets $\mathcal{L}$ and using Lemma 3, we obtain for $S_\sigma(z, v) = (1-z) P_\sigma(z, v)$ (cf. (3.5)) the following expression

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

as needed. ■

Finally, using Corollary 2 and Theorem 4 we obtain our main result of this subsection, namely the generating function $D(z, u)$ expressed in terms of the autocorrelation polynomial $a_\sigma(z)$.

COROLLARY 6.
The bivariate generating function $D(z, u)$ for the depth D_n of suffix trees becomes

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

for every $|u| < 1$ and $|z| < 1$. ■

Remark 3. In several computer science applications the string X is finite, and it is terminated by a special character which does *not* belong to the alphabet Σ (e.g., $X = x_1 x_2 \cdots x_n \$$ with $\$ \notin \Sigma$). Fortunately, it is easy to accommodate this case in our model. Indeed, note that in this case only an entire match between X and σ can take place due to the fact that the last special character \$ cannot match any character of σ. This implies that $m(\mathcal{L}) \leq n - k + 1$

[2] In example 3.4 this partitioning can be represented as follows. The set $\mathcal{L}$ is $\mathcal{L} = \{3, 5, 13, 15, 17, 23\}$ and for $\sigma = abab$ we have $i = 3$, $\mathcal{C}_1 = \{1, 3\}$. Therefore $\mathcal{L} = \{\mathcal{C}_1 + 2\} \cup \{\mathcal{L}' + 8\}$ with $\mathcal{L}' = \{5, 7, 9, 15\}$ and $m(\mathcal{C}_1) = 3$ (thus $i - 1 + m(\mathcal{C}) - 1 + k = 8$).

for $|\sigma| = k$. Consequently, the generating function $P_\sigma(z,v)$ becomes (see proof of Lemma 3)

$$P_\sigma(z,v) = \sum_{n=1}^{\infty} \sum_{\{\mathcal{L}:\ m(\mathcal{L}) \leq n-k+1\}} P(\mathcal{L},\sigma) v^{|\mathcal{L}|} z^n = \sum_{\mathcal{L}} P(\mathcal{L},\sigma) v^{|\mathcal{L}|} z^{m(\mathcal{L})} \sum_{l=k-1}^{\infty} z^l = \frac{z^{k-1}}{1-z} S_\sigma(z,v)\ ,$$

where $S_\sigma(z,v)$ is defined in (3.5). In particular, this implies

$$D(z,u) = (1-u) \sum_{\sigma} (uz)^{|\sigma|} \frac{p(\sigma)}{[(1-z)(1+a_\sigma(z)) + p(\sigma) z^{|\sigma|}]^2} \tag{3.10}$$

for all $|u| < 1$ and $|z| < 1$. Comparing (3.10) and (3.9) one should conclude that finiteness of the string X does not have any significant impact on the asymptotic behavior of suffix trees. This is confirmed by our analysis in the next section. □

3.4 Asymptotics

In this section, we present an asymptotic analysis of the depth D_n through a careful evaluation 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 moments* can be established (since every analytical function is differentiable). In the *second step*, we prove that the expanded generating function has only a single pole that determines the asymptotics (cf. Theorem 11). Finally, the *third step* consists of applying the celebrated Cauchy's theorem [13] to prove asymptotics. However, to simplify our investigation we do not determine directly the asymptotics of the suffix tree, but rather compare the asymptotics of suffix trees with independent tries (cf. Theorem 14) to take advantage of many well established results for tries (cf. [15, 17, 21-24]).

We start with a technical – but important – lemma concerning the autocorrelation polynomial $a_\sigma(z)$. Hereafter, it is assume that $p \geq q$ and $p < 1$. We consider all finite strings σ of length k. For any function $f(\sigma)$ of σ such that $|\sigma| = k$, we define $P_k(f(\sigma) \leq y) = \sum_{\{\sigma:\ |\sigma|=k\ ,\ f(\sigma)\leq y\}} p(\sigma)$ for any real y. The next lemma estimates a "typical" form of the autocorrelation polynomial.

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

$$P_k(a_\sigma(\rho) \leq \theta(\rho\delta)^k) \geq 1 - \theta\delta^k\ . \tag{3.11}$$

PROOF : Consider all finite strings σ of length k, and note that $a_\sigma(\rho)$ is a function of σ. It is more convenient for the purpose of this proof to give a probabilistic interpretation of $P_k(\cdot)$. Let us introduce a Bernoulli model restricted to finite strings of length k (we refer to it as the *finite* Bernoulli model). We assume that σ is a prefix of length k of an infinite random string defined in the infinite Bernoulli model. It is clear that the following identity $\Pr\{f(\sigma) < y\} = P_k(f(\sigma) < y)$ holds since the probability weight of σ in this model is exactly $p(\sigma)$ (see also our footnote in Section 3.2).

Our goal is to prove that within our finite Bernoulli model we have $\Pr\{a_\sigma(\rho) \le \theta(\rho\delta)^k\} \ge 1 - \theta\delta^k$. Recall that $\mathcal{F}_\sigma$ is the set of positions that σ overlaps with itself (except the trivial position 1). We shall prove that for some θ and δ the following holds: (i) $\min\{\mathcal{F}_\sigma\} \ge k/2$ implies that $a_\sigma(\rho) \le \theta(\rho\delta)^k$; and (ii) $\Pr\{\min\{\mathcal{F}_\sigma\} \le k/2\} \le \theta\delta^k$. This will imply our lemma since $\Pr\{a_\sigma(\rho) \le \theta(\rho\delta)^k\} \ge 1 - \Pr\{\min\{\mathcal{F}_\sigma\} \le k/2\}$. To proceed along these lines, we first consider the event $\{i+1 \in \mathcal{F}_\sigma\}$ for $i \le k$ which is equivalent to $\{C^\sigma_{1,i+1} \ge k-i\}$, where C^σ_{ij} is the self-alignment (cf. Section 2) between the i-th and the j-th suffixes of σ. Note that the probability of such an event does not change if we consider σ as the prefix of length k of an infinite string generated according to our infinite Bernoulli model. Therefore, we can refer to [3] for a closed formula for this probability, namely

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

where $r = k \bmod i$. In fact, the above can be proved by noting that $\{i+1 \in \mathcal{F}_\sigma\} = \{C^\sigma_{1,i+1} \ge k-i\}$ if and only if $\sigma = (\sigma_i)^{\lfloor \frac{k}{i} \rfloor}\xi$, where ξ is a prefix of σ_i with $|\xi| = r < i$ and σ_i is the prefix of σ of length i (cf. [18]). Using the above, with $p \ge q$ (and $p+q=1$), we obtain

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

Thus, $\Pr\{\min \mathcal{F}_\sigma \le \frac{k}{2}\} \le \sum_{i=1}^{k/2} \Pr\{i+1 \in \mathcal{F}_\sigma\} \le \frac{p^{k/2}}{1-p}$.

Now, consider those strings σ for which $\min\{\mathcal{F}_\sigma\} \ge \frac{k}{2}$. A simple algebra reveals that

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

hence (3.11) follows with $\delta = \sqrt{p}$ and $\theta^{-1} = p(1-p)$. Since there is no restrictions on ρ, we can always guarantee that $\rho\delta < 1$, as needed. ∎

Now, we are ready to prove our next result concerning an analytical continuation of $D(z,u)$, which by Corollary 6 is $D(z,u) = (1-u)\sum_\sigma u^{|\sigma|} p(\sigma) z / [R_\sigma(z)]^2$ where $R_\sigma(z) = (1-z)(1+a_\sigma(z)) + p(\sigma)z^{|\sigma|}$.

Theorem 8

The generating function $D(z,u)$ can be analytically continued to all $|u| < \delta^{-1}$ for some $p \le \delta < 1$.

Proof : Let $|u| < 1$ and $|z| < 1$. Consider the following identity (cf. Remark 1)

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

Therefore, for $|z| < 1$,

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

We have $R_\sigma(z) - (1-z) = (1-z)a_\sigma(z) + p(\sigma)z^k$. By Lemma 7, we note that for all σ such that $|\sigma| = k$,

$$P_k(|R_\sigma(z) - (1-z)| \le (|1-z|+1)\delta^k) \ge 1 - O(\delta^k) \ .$$

Moreover, for any bounded function $f(\sigma)$ such that $f(\sigma) \le f_{max}$ for all σ with $|\sigma| = k$, we also have the following estimate

$$\sum_{|\sigma|=k} p(\sigma) f(\sigma) \le y P_k(f(\sigma) < y) + f_{max}(1 - P_k(f(\sigma) < y)) \ . \tag{3.12}$$

In particular, using the above we obtain

$$\begin{aligned}
D(z,u) - \frac{z}{(1-z)^2} &= (u-1)\sum_{k=0}^{\infty} u^k \Big(P_k(|R_\sigma(z) - (1-z)| \le (|1-z|+1)\delta^k) O(\delta^k) + \\
&+ \ (1 - O(1) P_k(|R_\sigma(z) - (1-z)| \le (|1-z|+1)\delta^k)) \Big) \ ,
\end{aligned}$$

since for all σ we have $|a_\sigma(z)| < (1-p)^{-1}$. The above implies that $D(z,u) - z/(1-z)^2 = O((u-1)/(1-\delta|u|))$, as desired. ∎

The next step is to find singularities of the generating function $D(z,u)$ that contribute to asymptotics. We shall show that $D(z,u)$ does not have any singularities in the disc $|z| < 1$ (cf. Lemma 9), and the only pole of $D(z,u)$ is for $|z| > 1$ (cf. Theorem 11). In addition, in Lemma 10 we provide one technical result required in further proofs; in particular, to apply Rouche's theorem [13] needed in Theorem 11.

Lemma 9

The polynomial $R_\sigma(z)$ has no root in the disc $|z| < (1-p(\sigma))^{-1/k}$ where $|\sigma| = k$. ∎

PROOF: From the proof of Theorem 1 and the final form of $D(z,u)$ (cf. (3.9)) we conclude that

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

Note that

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

hence there is no root of $R_\sigma(z)$ in the disc $|z|<1$. In order to extend this claim to a larger disc, say $|z|<\rho$ where $\rho>1$, we need a better estimate for the probability $\Pr\{\langle\sigma\rangle_n=\{i\}\}$. This probability is definitely smaller than the probability that $i\in\langle X,\sigma\rangle$ and $j\notin\langle X,\sigma\rangle$ for $j\le n$ such that $i-j$ is a multiple of $|\sigma|=k$. Note that suffixes started at positions j defined above do not overlap on the first k symbols, and the number of such nonoverlapping suffixes is greater than $\lfloor\frac{n}{k}\rfloor$. Hence, under our Bernoulli model,

$$\Pr\{\langle\sigma\rangle_n=\{i\}\}\le p(\sigma)(1-p(\sigma))^{\lfloor\frac{n}{k}\rfloor-1}$$

and this immediately leads to

$$\left|\frac{z}{R_\sigma(z)^2}\right|\le\sum_{n=1}^{\infty}n|z|^n(1-p(\sigma))^{\lfloor\frac{n}{k}\rfloor-1}$$

which converges in $|z|<(1-p(\sigma))^{-1/k}$, as stated in Lemma 9. ∎

LEMMA 10

There exist an integer K, a constant $\rho>1$, and a real number $\alpha>0$ such that the following holds

$$|\sigma|\ge K\qquad\Rightarrow\qquad|1+a_\sigma(z)|\ge\alpha\ .$$

for all $|z|\le\rho$ where $\delta\rho<1$. ∎

PROOF: Define ℓ as an integer such that $p+p^\ell<1$, and let $\rho>1$ be such that $p\rho+(p\rho)^\ell<1$. Consider a finite string σ with $|\sigma|=k>\ell$, and let $i+1=\min\mathcal{F}_\sigma$. The integer $i+1$ is the first position from which σ overlaps with itself. We consider two cases: (i) $i\ge\ell$, and (ii) $i<\ell$. First, suppose that $i\ge\ell$. Hence, for every complex number z such that $|z|\le\rho$, we have

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

provided $p\ge q$. Then, the lemma is proved by referring to our definition of ℓ.

The second case $i < \ell$ is more intricate. Let $q = \lfloor \frac{k}{i} \rfloor$. Then, as in the proof of Lemma 7, we have $\sigma = (\sigma_i)^q \xi$, where σ_i is prefix of σ of length i and ξ is a prefix string of σ_i such that $|\xi| < i = |\sigma_i|$. Then,

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

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 $p\rho + 3(p\rho)^{j\ell} < 1$ and choose K such 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. ■

With the above apparatus, we are able to prove the next theorem which is crucial for our further analysis.

THEOREM 11

There exists K' such that for $|\sigma| \geq K'$, there is only one root of the equation $R_\sigma(z) = 0$ in the region $1 < |z| \leq \rho$ for $\delta\rho < 1$.

PROOF : Let K_1 be such that $(p\rho)^{K_1} < \alpha(\rho - 1)$ holds for some α and ρ as in Lemma 10. Denote $K' = \max\{K, K_1\}$, where K is defined in Lemma 10. Note also that for $p \geq q$ the above condition implies that for all σ such that $|\sigma| = k > K'$ we have $p(\sigma)\rho^k < \alpha(\rho - 1)$. Hence, by Lemma 10 for $|\sigma| > K'$ we have $|p(\sigma)z^k| < |(z-1)(1 + a_\sigma(z))|$ on the circle $|z| = \rho > 1$. Therefore, by Rouché's theorem [13] the polynomial $R_\sigma(z)$ has the same number of roots as $(1-z)(1+a_\sigma(z))$ in the disc $|z| \leq \rho$. But, the polynomial $(1-z)(1+a_\sigma(z))$ has only a single root in this disc since by Lemma 10 we have $(1 + a_\sigma(z)) > 0$ in $|z| \leq \rho$. ■

As a consequence of Theorem 11, we conclude that there exists the smallest root of $R_\sigma(z) = 0$ which we denote as A_σ. Let also C_σ and D_σ be the first and the second derivatives of $R_\sigma(z)$ at $z = A_\sigma$, respectively. Using bootstrapping, one easily obtains the following expansions

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

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

Finally, in our last step we compare asymptotics of suffix trees with corresponding 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 the difference between the *probability distribution functions* of the depth in suffix trees and in independent tries, that is

$$Q_n(u) = \frac{1}{1-u}\left(E[u^{D_n}] - E[u^{D_n^T}]\right)$$

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

Then, by (3.6) and by (3.9) of Corollary 6 we obtain

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

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 $|u| \le \beta$ such that $\beta > 1$

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

PROOF : By Cauchy,

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

where the integration is done along a loop contained in the unit disc that encircles the origin. Let σ be such that $|\sigma| \ge K'$, where K' is defined in Theorem 11. From the proof of Theorem 11 we conclude that $R_\sigma(z)$ and $(1-z+p(\sigma)z)$ have only one root in $|z| \le \rho$. Applying the residue formula [13] we obtain

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

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

where

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

To establish a bound for $I_\sigma(\rho)$ we argue exactly in the same manner as in the proof of Theorem 8. This leads for $|\sigma| > K'$ to the following

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

since for all σ we also have $a_\sigma(\rho) \leq 1/(1-p\rho)$ and $R_\sigma(z) = O(\rho^k)$ in the circle $|z| \leq \rho$. Set now $\beta = (\delta\rho)^{-1} > 1$. Then, for $|u| < \beta$ we have the following estimate $\sum_{\{\sigma:\ |\sigma|>K'\}} I_\sigma(u) = O(\rho^{-n})$, and this establish our bound since the other terms ($|\sigma| < K'$) contribute only B^{-n} for some $B > 1$ due to the fact that all roots of $R_\sigma(z)$ have magnitudes greater than 1. ■

Finally, we can formulate our main result of this section. 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 [7]).

THEOREM 14
For all $1 < \beta < \delta^{-1}$, *there exists* $\varepsilon > 0$ *such that uniformly for* $|u| \leq \beta$: $E[u^{D_n^S}] - E[u^{D_n^T}] = (1-u)O(n^{-\varepsilon})$.
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, by Lemma 12 we have

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

Let now $f_\sigma(x)$ be a function defined for x real by

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

By the same arguments as used in proving (3.12) in Theorem 8, we note that $\sum_\sigma u^{|\sigma|} p(\sigma) f_\sigma(x)$ is absolutely convergent for all x and u such that $|u| \leq \beta$. The function $\bar{f}_\sigma(x) = f_\sigma(x) - f_\sigma(0)e^{-x}$ is exponentially decreasing when $x \to +\infty$ and is $O(x)$ when $x \to 0$; therefore its Mellin transform $\bar{f}_\sigma^*(s) = \int_0^\infty \bar{f}_\sigma(x) x^{s-1} dx$ (cf. [9]), is well defined for $\Re(s) > -1$. In this region we obtain

$$\bar{f}_\sigma^*(s) = \Gamma(s) \left(\frac{(\log A_\sigma)^{-s} - 1}{A_\sigma C_\sigma^2} - \frac{(-\log(1-p(\sigma))^{-s} - 1}{1-p(\sigma)} \right) \ ,$$

where $\Gamma(s)$ is the gamma function [13]. Let $g^*(s,u)$ be the Mellin transform of the series $\sum_\sigma u^{|\sigma|} p(\sigma) \bar{f}_\sigma(x)$ which exists at least in the strip $(-1, 0)$. Formally, we have

$$g^*(s,u) = \sum_\sigma u^{|\sigma|} p(\sigma) \bar{f}_\sigma^*(s) \ .$$

We can reverse the Mellin transform $g^*(s,u)$ [13] provided that the following holds.

LEMMA 15
The function $g^*(s,u)$ is analytical in $\Re(s) \in (-1,\varepsilon)$ for some $\varepsilon > 0$. ∎

Assuming Lemma 15 is granted, we have

$$Q_n(u) = \frac{1}{2i\pi}\int_{c-i\infty}^{c+i\infty} g^*(s,u)n^{-s}ds + O(1/n) \; + \sum_\sigma u^{|\sigma|}p(\sigma)f_\sigma(0)e^{-n},$$

with $c \in (0,\varepsilon)$. Note that the last term of the above contributes $O(e^{-n})$, and can be safely ignored. Furthermore, a simple majorization under the integral gives the evaluation $Q_n(u) = O(n^{-c})$ which completes the proof of Theorem 14. ∎

PROOF OF LEMMA 15: We establish the absolute convergence of $g^*(s,u)$ for all s such that $\Re(s) \in (-1,\varepsilon)$ and $|u| \le \beta$. Let us define $h^*(s,u) = \frac{g^*(s,u)}{\Gamma(s)}$. Note that for any fixed s we have the following

$$\begin{aligned}(\log A_\sigma)^{-s} &= \left(\frac{p(\sigma)}{1+a_\sigma(1)}\right)^{-s}(1+O(p(\sigma)))\;,\\ (-\log(1-p(\sigma)))^{-s} &= p(\sigma)^{-s}(1+O(p(\sigma)))\;.\end{aligned}$$

Thus,

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

By Lemma 7, $P_k(1+a_\sigma(1) \le 1+\theta\delta^k) \ge 1-O(\delta^k)$, and hence

$$h^*(s,u) = \sum_{k=0}^{\infty}\left(\sup\{p^{-\Re(s)}, q^{-\Re(s)}\}|u|\delta\right)^k O(1)$$

which absolutely converges for all values of s such that $\Re(s) < \varepsilon$ where $\sup\{p^{-\varepsilon}, q^{-\varepsilon}\} < (\delta\beta)^{-1}$. Since $h^*(0,u) = 0$ by definition, the pole of $\Gamma(s)$ at $s=0$ is canceled in $g^*(s,u)$, and therefore $h^*(s,u)$ does not show any singularities in the strip $\Re(s) \in (-1,\varepsilon)$. ∎

Finally, to prove Proposition 2 we consider asymptotics for the independent tries. A copious literature has been devoted to this topic (cf. [8, 15, 17, 21-28]). Nevertheless, it might be interesting and illuminating to obtain the asymptotics for the depth D_n^T of independent tries directly from the generating function (3.6). This can also be regarded as an additional verification of our approach. First of all, we note that the Cauchy's formula applied to (3.6) implies

$$E[u^{D_n^T}] = (1-u)\sum_\sigma u^\sigma p(\sigma)(1-p(\sigma))^{n-1}$$

and therefore, the Mellin transform $D^*(s,u)$ of $E[u^{D_n^T}]$ becomes for $-1 < \Re(s) < 0$

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

After simple algebra – that uses the formulas 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. ANOTHER APPLICATION: SIZE OF SUFFIX TREES

In this section we apply the string-ruler approach to obtain another characteristic of suffix trees, namely the *average* number of internal nodes in a suffix tree. Such a characteristic is useful in many applications of suffix trees, most notably to assess the space complexity of algorithms that are based on suffix trees.

Let EL_n denote the average size of a suffix tree built over n suffixes. Then, as easy to see, for $n \geq 2$

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

where $|\langle\sigma\rangle_n|$ denotes the cardinality of the set $\langle\sigma\rangle_n$ already defined in Section 3. The above formula is a simple consequence of our discussion in Section 3.1 since $|\langle\sigma\rangle_n| \geq 2$ implies the existence of an internal node in a suffix tree at depth $|\sigma| = k$. Having this is mind, we prove the following theorem.

THEOREM 16

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

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

PROOF : To compute $\Pr\{|\langle\sigma\rangle_n| \geq 2\}$ we need to evaluate the following two probabilities: $\Pr\{|\langle\sigma\rangle_n| = 0\}$ and $\Pr\{|\langle\sigma\rangle_n| = 1\}$. For the former probability, let A_i denote an event that σ does not match X starting at position i, that is, $i \notin \langle\sigma\rangle_n$. Then, as in Section 3, we obtain

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

On the other hand, as in the derivation of Theorem 1 (see (3.1)-(3.3)), we obtain

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

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

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

Using Theorem 16 and extending slightly our analysis from Sections 3.2 and 3.3, we obtain the following results concerning the average size of suffix trees and independent tries.

COROLLARY 17

(i) `Suffix Tree`. *The generating function $L(z)$ for the size of a suffix tree becomes*

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

for $|z| < 1$, where $R_\sigma(z) = (1-z)(1+a_\sigma(z)) + p(\sigma)z^{|\sigma|}$.

(ii) `Independent Trie`. *The generating function $L_n^T(z)$ for the size of independent tries is*

$$L^T(z) = z - \sum_{\sigma} \left(\frac{zp(\sigma)}{[1-z+p(\sigma)z]^2} - \frac{zp(\sigma)}{(1-z)(1-z+p(\sigma)z)} \right) \tag{4.5}$$

for $|z| < 1$. ■

The next step is to obtain asymptotics for the average size of a suffix tree. 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.

LEMMA 18

We have the following estimate when $n \to \infty$

$$EL_n^S - EL_n^T = O(n^{1-\varepsilon})$$

for some $0 < \varepsilon < 1$.

PROOF : Arguing as in Section 3.3, we apply the Cauchy residue formula to (4.4) and (4.5) to obtain

$$EL_n^T - EL_n^S = d_{n,1} + d_{n,2} + O(B^{-n}) \ ,$$

for some $B > 1$ and $d_{n,1} = \sum_\sigma \{p(\sigma)[A_\sigma^{-n}(n/(C_\sigma^2 A_\sigma) + D_\sigma/C_\sigma^3) - n(1-p(\sigma))^{n-1}]\}$, and the second term is $d_{n,2} = \sum_\sigma \{p(\sigma)A_\sigma^{-n}/[(1-A_\sigma)C_\sigma] - (1-p(\sigma))^n\}$. The term $d_{n,1}$ is the same as the one analyzed in Section 3 except for the factor n that shows up in $d_{n,1}$. Hence $d_{n,1} = O(n^{1-\varepsilon})$. The term $d_{n,2}$ is more intricate since we do not know even whether the series in $d_{n,2}$ is convergent.

To estimate $d_{n,2}$ we need an extended expansion of the root A_σ of $R_\sigma(z)$. As in Section 3, using one more step in the bootstrapping method we arrive at

$$A_\sigma = 1 + \frac{1}{1+a_\sigma(1)} p(\sigma) + \left(\frac{k}{(1+a_\sigma(1))^2} - \frac{a'_\sigma(1)}{(1+a_\sigma(1))^3} \right) p(\sigma)^2 + O(p(\sigma)^3) \ .$$

Therefore, $(1-A_\sigma)C_\sigma = p(\sigma) + O(p(\sigma)^2 a'_\sigma(1) + p(\sigma)^3)$, and $\frac{p(\sigma)A_\sigma^{-n}}{(1-A_\sigma)C_\sigma} - (1-p(\sigma))^n = (1-p(\sigma))^n O(np(\sigma){a'}_\sigma(1) + np(\sigma)^2)$. Since $a'_\sigma(1) \le k a_\sigma(1)$, by Lemma 7 we know that the series $\sum_\sigma p(\sigma) a'_\sigma(1)$ converges like $\sum_k k\delta^k$, hence we conclude that $d_{n,2}$ converges in the same manner.

Now, we are ready to apply the Mellin transform to prove our asymptotic approximation for $d_{n,2}$. Let us introduce a new function $f_\sigma(x)$ defined as

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

We have $d_{n,2} = \sum_\sigma f_\sigma(n)$. The function $f_\sigma(x)$ decreases for $x \to \infty$ and $f_\sigma(0) \neq 0$, hence the Mellin transform of $f_\sigma(x)$ exists in the strip $(0, \infty)$. The Mellin transform of $\sum_\sigma f_\sigma(x)$, however, does not exist. Indeed, the function $\sum_\sigma f_\sigma(x)$ is non-zero when $x = 0$, hence the transform would exist only in the strip $\Re(s) > 0$. But, the function becomes $O(x)$ when $x \to \infty$, and this would require $\Re(s) < -1$ for the existence of the transform, which contradicts our previous estimate.

In order to circumvent this problem, we use of the same trick as in Theorem 14, namely we introduce a new function $\bar{f}_\sigma(x) = f_\sigma(x) - f_\sigma(0)e^{-x} - (f'_\sigma(0) + f_\sigma(0))xe^{-x}$. Note that $d_{n,2} = \sum_\sigma \bar{f}_\sigma(n) + O(ne^{-n})$. We also have $\bar{f}_\sigma(x) = O(x^2)$ when $x \to 0$, hence its Mellin transform is defined on the larger strip $(-2, \infty)$ and the Mellin transform $g^*(s)$ of $\sum_\sigma \bar{f}_\sigma(x)$ is now well defined in the strip $(-2, -1)$. It is easy to factorize $\bar{f}^*_\sigma(s) = \Gamma(s) h^*_\sigma(s)$, and then elementary algebra (using the extended expansion of A_σ) shows that the series $\sum_\sigma h^*_\sigma(s)$ converges like $\sum_\sigma |p(\sigma)^{-s} k\delta^k|$ for $\Re(s) < -1 + \varepsilon$ and some $\varepsilon > 0$. Therefore, the Mellin transform $g^*(s)$ exists in the larger strip $(-2, -1+\varepsilon)$. This leads to our estimate $n^{1-\varepsilon}$, after applying the reverse Mellin transform in the same manner as in Theorem 14. ■

Finally, estimating the size of independent tries (cf. [24]) we prove the following result.

COROLLARY 19

For large n the following holds

$$EL_n^S = EL_n^T + O(n^{1-\epsilon}) = \frac{n}{h_1}(1 + P_4(\log n)) + O(n^{1-\epsilon}) \tag{4.8}$$

where $\varepsilon > 0$ and $P_4(\log n)$ is a fluctuating function with a small amplitude. ■

In the above corollary we have used the fact that for independent tries $EL_n^T = n/h_1 \cdot (1 + P_4(\log n)) + O(1)$ [14, 24]. Of course, we can obtain this result directly from (4.5). Indeed, by the Cauchy's formula and applying Remark 1 we obtain the following

$$\begin{aligned} EL_n &= -\sum_\sigma \{np(\sigma)[1-p(\sigma)]^{n-1} + [1-p(\sigma)]^n - 1\} \\ &= -\sum_\sigma \{n\sum_{\ell=2}^{n}(-1)^{\ell+1}\binom{n-1}{\ell-1}p^\ell(\sigma) + \sum_{\ell=2}^{n}(-1)^\ell\binom{n}{\ell}p^\ell(\sigma)\} \\ &= \sum_{\ell=2}^{n}(-1)^\ell\binom{n}{\ell}\frac{\ell-1}{1-p^\ell-q^\ell} = \frac{n}{h_1}\cdot(1+P_4(\log n)) + O(1)\ . \end{aligned}$$

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 [25].

ACKNOWLEDGEMENT

We sincerely thank an anonymous referee for a very careful reading of our manuscripts, and several suggestions that led to elimination of some minor slips in our analysis, and improvements of our presentation.

REFERENCES

[1] A. Apostolico, The Myriad Virtues of Suffix Trees, *Combinatorial Algorithms on Words,* pp. 85-96, Springer-Verlag, 3I F12 (1985).

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

[3] A. Apostolico, W. Szpankowski, Self-alignments in Words and Their Applications, *J. Algorithms*, (1991), to appear.

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

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

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

[7] L. Devroye, W. Szpankowski and B. Rais, A Note of the Height of Suffix Trees, *SIAM J. Computing*, 20, 5 (1991).

[8] P. Flajolet, On the Performance Evaluation of Extendible Hashing and Trie Searching, *Acta Informatica,* 20, 345-369 (1983).

[9] 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).

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

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

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

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

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

[15] P. Jacquet and W. Szpankowski, Analysis of Tries With Markovian Dependency, *IEEE Trans. Information Theory*, 37, 1470-1475 (1991).

[16] P. Jacquet and W. Szpankowski, What can we learn about suffix trees from independent tries? *1991 Workshop on Algorithms and Data Structures*, Lecture Notes in Computer Science, 519 (eds. F. Dehne, J. Sack and N. Santoro), Springer-Verlag, pp. 228–229 (1991).

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

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

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

[20] E.M. McCreight, A Space Economical Suffix Tree Construction Algorithm, *JACM*, 23, 262-272 (1976).

[21] B. Pittel, Asymptotic Growth of a Class of Random Trees, *Annals of Probability,* 18, 414 - 427 (1985).

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

[23] B. Rais, P. Jacquet and W. Szpankowski, Limiting Distribution for the Depth in Patricia Tries, *SIAM Discrete Mathematics*, in press.

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

[25] P. Shields, Entropy and Prefixes, *Annals of Probability*, 20, 403-409 (1992).

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

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

[28] W. Szpankowski, On the Height of Digital Trees and Related Problems, *Algorithmica*, 6, 256-277 (1991).

[29] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, to appear.

[30] P. Weiner, Linear Pattern Matching Algorithms, *Proc. of the 14-th Annual Symposium on Switching and Automata Theory,* 1-11 (1973).

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