DIMACS Technical Report 97-65
October 1997

Longest Common Consecutive Substring in two Random Strings (Preliminary Version)

by

Sarmad Abbasi[1]
Dept. of Computer Science
Rutgers University
Piscataway, New Jersey 08855
email: sabbasi@paul.rutgers.edu

[1]This work was done while the author was supported by DIMACS

ABSTRACT

Let Σ be a finite alphabet with C letters. For any two strings $x = x_0 \cdots x_{n-1}$ and $y = y_0 \cdots y_{n-1}$ of length n, we let $S(x, y)$ denote the size of the longest common consecutive substring between x and y; that is, $S(x, y)$ is the largest k such that,

$$x_i \cdots x_{i+k-1} = y_j \cdots y_{j+k-1}$$

for some $i, j \in \{0, \ldots, n-1\}$ (Here the indices are modulo n.) We show that for x and y chosen uniformly among all possible strings of length n, $S(x, y)$ is highly concentrated around $2 \log_C n$. More precisely, for any $a \geq 1$

$$\Pr[|S(x, y) - 2 \log_C n| > a] \leq \frac{18C^{-a}}{n^2}.$$

This enables us to show that the expected value of $S(x, y)$ is within a constant additive term of $2 \log_C n$.

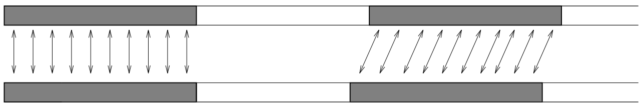

Figure 1: Non overlapping substrings in x and y

1 Introduction

Let Σ be a finite alphabet and $x = x_0, \ldots, x_{n-1}$ and $y = y_0, \ldots, y_{n-1}$ be two strings of length n over Σ. We say that x and y have a k-matching if they have common (consecutive) substring of length k; that is,

$$x_i x_{i+1} \ldots x_{i+k-1} = y_j y_{j+1} \cdots y_{j+k-1}$$

for some $i, j \in \{0, \ldots, n-1\}$. Here all the indices are taken modulo n; hence we are considering x and y as "circular" strings. When $\Sigma = \{A, T, G, C\}$ is the alphabet of the genetic code, x and y are two DNA's. In molecular biology parlance a common substring of length k in x and y is called a k-homology between x and y. Homology is an important measure of similarity between two DNA [7]. We let $S(x, y)$ denote the largest k such that x and y have a k-matching. The purpose of this note is to study $S(x, y)$ when x and y are randomly chosen with uniform distribution over Σ^n.

Chvátal and Sankoff [3] in a seminal paper consider the function $L(x, y)$ the length of the longest common substring in x, y. Two strings x and y have a common substring of length k if there exist $s, t \in \{0, \ldots, n-1\}$ and $0 \leq i_1 < \cdots i_k < n$ such that,

$$x_{s+i_j} = y_{t+i_j} \text{ for all } j = 1, \ldots, k.$$

In case of $L(x, y)$ it is known that $\frac{E[L(x,y)]}{n} \to \rho_C$ for some constant ρ_C. However, even the value of ρ_2 is not known. See [4, 5] for some recent progress on this problem. In another direction Abbasi and Sengupta [1] consider the following algorithmic question: Given x find a y which minimizes $S(x, y)$. They show that one can find such a y in time $O(n \log n)$. The analysis of the algorithm depends on showing that for every $x \in \Sigma^n$ there is a y such that, $S(x, y) \leq \log_C n + \log_C \log_C n + 4$.

When we refer to a random string of length n the underlying probability distribution is assumed to be the uniform distribution over Σ^n. Note a random string of length n can be generated by generating each of its n letters uniformly and independently from Σ. We only refer to consecutive substrings of a string, hence we drop "consecutive" as no confusion arises.

It is known that $S(x, y)$ is around $2 \log_C n$ almost always; more precisely the following result can be found in [8].

Theorem 1.1 *Let Σ be an alphabet with C letters. For every $\epsilon, \delta > 0$ there is a threshold n_0 such that the following statement holds for all $n \geq n_0$: If x and y are two random strings of length n over Σ then*

$$\Pr[|S(x, y) - 2 \log_C n| > \epsilon \log_C n] \leq \delta.$$

In this section we strengthen Theorem 1.1 considerably. The goal of this paper is to show the following result.

Theorem 1.2 *Let Σ be an alphabet with C letters. If x and y are two random string of length n over Σ then*

$$\Pr[|S(x,y) - 2\log_C n| > a] \leq \frac{18C^{-a}}{n^2}.$$

for every $a \geq 1$.

This result shows that $S(x,y)$ is concentrated at two values, namely, $\lceil 2\log_C n\rceil$ and $\lfloor 2\log_C n\rfloor$.

Corollary 1.1 *For n large enough*

$$\Pr[S(x,y) \notin \{\lceil 2\log_C n\rceil, \lfloor 2\log_C n\rfloor\}] \leq \frac{9}{n^2}.$$

Using Theorem 1.2 it is also easy to obtain bounds on the expected value of $S(x,y)$.

Corollary 1.2 *For n large enough,*

$$2\log_C n - 2 \leq E[S(x,y)] \leq 2\log_C n + 2.$$

2 Overlapping and Non Overlapping Substrings

In this section we let $k < \frac{n}{2}$. Since we are considering indices modulo n we will consider every index $j > \frac{n}{2}$ to be negative. For any string z of length n we let $z[i,k]$ be the k-substring of z starting at i; that is,

$$z[i,k] = z_i z_{i+1} \ldots z_{i+k-1}.$$

Furthermore, we say $z[i,k]$ and $z[r,k]$ are overlapping if

$$|i - r| < k.$$

Let x and y be chosen uniformly over Σ^n. We let $A_{i,j}$ be the event that k-substring of x starting at i is equal to the k-substring of y starting at j; that is,

$$x[i,k] = y[j,k].$$

Note that

$$\Pr[A_{i,j}] = C^{-k}.$$

We let $X_{i,j}$ be the indicator variable for the event $A_{i,j}$. Hence,

$$E[X_{i,j}] = \Pr[A_{i,j}] = C^{-k}.$$

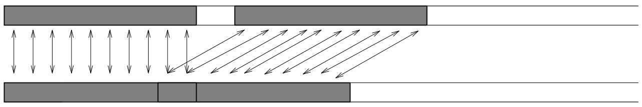

Figure 2: Overlapping substrings in x and non overlapping substring y

Let

$$X = \sum_{i,j} X_{i,j}.$$

X denotes the number of k-matchings between x and y. By linearity of expectation we have,

$$E[X] = \sum_{i,j} E[X_{i,j}] = n^2 C^{-k}.$$

We are interested in the second moment of X; that is,

$$E[X^2] = \sum_{(i,j)(r,s)} E[X_{i,j} X_{r,s}] = \sum_{(i,j),(r,s)} \Pr[A_{i,j} \text{ and } A_{r,s}].$$

By symmetry we observe that,

$$E[X^2] = n^2 \sum_{(i,j)} \Pr[A_{i,j} \text{ and } A_{0,0}].$$

Thus our primary interest is in computing $\Pr[A_{i,j}$ and $A_{0,0}]$.

Remark: Note that our random variables depend on the parameter k. We do not include k in our notation to avoid complicated subscripts.

Now let us take an event $A_{i,j}$. Suppose that $x[0,k]$ and $x[i,k]$ are not overlapping. Furthermore, suppose that $y[0,k]$ and $y[i,k]$ are also not overlapping. In this case, it is easy to see that the events $A_{i,j}$ and $A_{0,0}$ are pairwise independent (See Figure 1). The following observation is also not hard. If say $x[0,k]$ and $x[i,k]$ *are overlapping* and $y[0,k]$ and $y[i,k]$ are *not overlapping* then also $A_{i,j}$ and $A_{0,0}$ are pairwise independent (See Figure 2). To see this consider generating the string x first and let the outcome be $\hat{x}$. Since $y[j,k]$ and $y[0,k]$ are not overlapping, the event $y[j,k] = \hat{x}[i,k]$ is independent of $y[0,k] = \hat{x}[0,k]$. The first event depends on $y_j \cdots y_{j+k-1}$ which are generated independently from $y_0 \cdots y_{s+k-1}$ on which the second event depends. Indeed this observation suffices to prove Theorem 1.1. Perhaps what comes as a little surprise is the following.

Lemma 2.1 *If*

$$i \neq j$$

then the events $A_{i,j}$ and $A_{0,0}$ are pairwise independent.

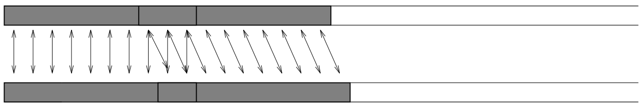

Figure 3: Overlapping substrings in x and y

Proof: From the above discussion we may assume that $x[i,k]$ and $x[0,k]$ are overlapping and also that, $y[j,k]$ and $y[0,k]$ are overlapping (See Figure 3.) Since indices are taken modulo n and $k < \frac{n}{2}$. Hence, $|i| < k$ and $|j| < k$. There are four cases to consider depending on whether i and j are positive or negative. However, because of symmetry it suffice to consider only two cases.

Case 1: Both i and j are positive.

Let us assume that $0 \leq i < j < k$. Let $\mathcal{A}$ be the set of pairs of string (a,b) such that the length of a is $k+i$ and the length of b is $k+j$. Let $\mathcal{S}$ be the pairs of strings $(a,b) \in \mathcal{A}$ which satisfy

$$a[0,k] = b[0,k] \tag{1}$$

and

$$a[i,k] = b[j,k]. \tag{2}$$

In what follows we want to calculate the size of $\mathcal{S}$. Let $t = j - i$ and $l = k - i$. We write

$$l = qt + d$$

where $0 \leq d < t$. Let α be any strings of length t. Furthermore, let us take γ and δ to be any strings of length i. Consider the following two strings (See Figure 4)

$$\hat{a} = \gamma\alpha^q\alpha[0,d]\delta$$

and

$$\hat{b} = \gamma\alpha^{q+1}\alpha[0,d]\delta.$$

The $\hat{a}$ has length $i + qt + d + i = k + i$ and $\hat{b}$ has length $i + qt + t + d + i = k + i + t = k + j$. We have,

$$\hat{a}[0,k] = \gamma\alpha^q\alpha[0,d] = \hat{b}[0,k]$$

and

$$\hat{a}[i,k] = \alpha^q\alpha[0,d]\delta = \hat{b}[j,k];$$

that is, $\hat{a}$ and $\hat{b}$ satisfy Equations 1 and 2. Our next task is to show that *any two* strings which satisfy Equations 1 and 2 must be of this form. Indeed assume that, $a[0,k] = b[0,k]$ and $a[i,k] = b[j,k]$ and we let,

$$\gamma = a[0,i], \alpha = a[i,t], \text{ and } \delta = a[k,i].$$

For $s = 0, \ldots, q$ since $i + st \leq i + qt \leq k$ from Equation 1 we have,

$$a[i+st,t] = b[i+st,t] \tag{3}$$

Also, we have,

$$a[i+st,t] = b[j+st,t].$$

Using the fact that $j = i+t$ and then Equation 3 we get, for $s = 0, \ldots, q-1$

$$a[i+st] = b[i+(s+1)t,t] = a[i+(s+1)t].$$

This yields for $s = 0, \ldots, q-1$,

$$a[i+st,t] = a[i,t] = \alpha.$$

Lastly, since $k = i+qt+d$ we have,

$$a[i+(q-1)t,d] = b[j+(q-1)t,d] = b[i+qt,d] = a[i+qt,d].$$

Hence,

$$a[i+qt,d] = a[i+(q-1)t,d] = \alpha[0,d].$$

This establishes the fact that a is of the required form; that is, $a = \gamma\alpha^q\alpha[0,d]\delta$. Now it is straight forward to check that b must have the required from.

Hence, the number of pairs (a,b) where a and b are strings of length $k+i$ and $k+j$ satisfying Equations 1 and 2 is exactly equal to the number of ways to choose γ, α and δ of sizes i, t, and i respectively. This in turn is equal to

$$C^{i+t+i} = C^{i+j}.$$

Whereas the number of ways of choosing pairs $(a,b) \in \mathcal{A}$ is exactly

$$C^{2k+i+j}$$

Now we are ready to show that $A_{0,0}$ and $A_{i,j}$ are pairwise independent. Note that these are events that,

$$x[0,k] = y[0,k] \text{ and } x[i,k] = y[j,k].$$

Hence, from the above discussion this is equivalent to $(x[0,k+i], y[0,k+j]) \in \mathcal{S}$. Since, x and y are generated uniformly and independently, we have,

$$\Pr[A_{0,0} \text{ and } A_{i,j}] = \Pr[(x[0,k+i], y[0,k+j]) \in \mathcal{S}] = \frac{|\mathcal{S}|}{|\mathcal{A}|} = C^{-2k}.$$

Since,

$$\Pr[A_{0,0} \text{ and } A_{i,j}] = \Pr[A_{0,0}]\Pr[A_{i,j}]$$

they are independent.

Case 2: i is positive and j is negative. Let us replace j with $-j$ and assume that j is positive. Let $\mathcal{A}$ be the set of pairs of strings (a,b) where $|a| = k+i$ and $|b| = k+j$. We

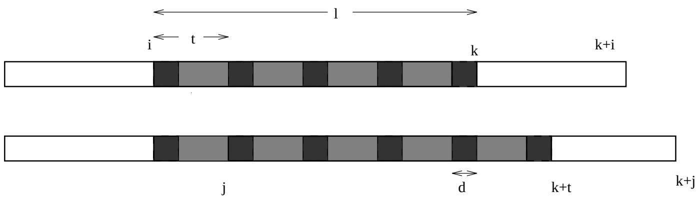

Figure 4: A typical pair of strings in $\mathcal{S}$

index b by starting from $-j$; that is, $b = b_{-j} \cdots b_0 \cdots b_{k-1}$ and $a = a_0 \cdots a_{k+i-1}$. Again we let $\mathcal{S}$ be the set of pairs $(a, b) \in \mathcal{A}$ satisfying

$$a[0, k] = b[0, k] \tag{4}$$

$$a[i, k] = b[-j, k] \tag{5}$$

By following a similar reasoning as in case 1, it suffices to check that, $|\mathcal{S}| = C^{i+j}$. Set $t = i+j$ and $k + i = q_1 t + s_1$ and $k = q_2 t + s_2$, where $s_1, s_2 < t$. Let α and β be any two strings of length i and j and let $\delta = \alpha\beta$. Define,

$$\hat{a} = \delta^{q_1} \delta[0, s_1]$$

and

$$\hat{b} = \beta \delta^{q_2} \delta[0, s_2]$$

It is easy to see that $\hat{a}$ and $\hat{b}$ satisfy Equations 4 and 5. Furthermore, one can show using similar methods as used in case 1 that any pair of strings in $\mathcal{S}$ is of this form (See Figure 5.) Hence, the number of strings in $\mathcal{S}$ is exactly the ways of choosing strings α and β of length i and j respectively. Therefore, $|\mathcal{S}| = C^{i+j}$ and we are done. ∎

Lemma 2.2 *For any i with $|i| < k$ we have,*

$$\Pr[A_{0,0} \textit{ and } A_{i,i}] = C^{-k-|i|}$$

Proof: If i is positive, event in consideration is equivalent to

$$x[0, k+i] = y[0, k+i].$$

Since x and y are chosen at random the probability of the above event is C^{-k-i}. If i is negative the event in consideration is equivalent to

$$x[i, k] = y[i, k]$$

and again the probability of this event is C^{-k+i}. ∎

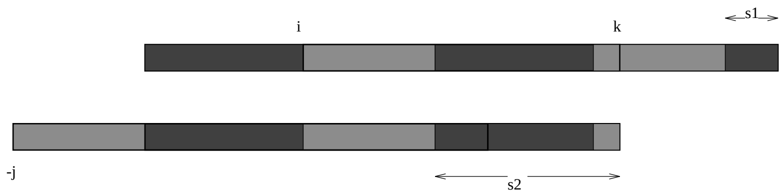

Figure 5: A typical pair of strings in $\mathcal{S}$

3 The Concentration Result

The second moment method is widely use to prove concentration results for various problems in combinatorics [2]. Here we use the same method to obtain our concentration result. For any $k < \frac{n}{2}$ we let

$$\mu := E[X] = \sum_{(i,j)} E[X_{i,j}] = n^2 C^{-k}.$$

As observed before we have,

$$E[X^2] = n^2 \sum_{(i,j)} \Pr[A_{i,j} \text{ and } A_{0,0}] \tag{6}$$

Let $R = \{(i,i) : 1-k \leq i \leq k-1\}$ then

$$\begin{aligned} \sum_{(i,j)\in R} \Pr[A_{i,j} \text{ and } A_{0,0}] &= \sum_{1-k\leq i\leq k-1} \Pr[A_{i,i} \text{ and } A_{0,0}] \\ &= \sum_{1-k\leq i\leq k-1} C^{-k-|i|} \\ &\leq 3C^{-k} \end{aligned}$$

For any $(i,j) \notin R$, $A_{i,j}$ and $A_{0,0}$ are independent. Hence,

$$\begin{aligned} E[X^2] &= n^2 \sum_{(i,j)} \Pr[A_{i,j} \text{ and } A_{0,0}] \\ &\leq n^4 C^{-2k} + 3C^{-k} \end{aligned}$$

This yields the variance σ^2 of X is bounded as follows:

$$\sigma^2 \leq 3C^{-k}$$

Proof(Of Theorem 1.2) Now we are ready to give our concentration result. We use the following two consequences of the Chebyschev's inequality [6]. If X is a nonnegative variable then

$$\Pr[X = 0] \leq \frac{\sigma^2}{\mu^2}. \tag{7}$$

If X is a nonnegative variable with $\mu \le \frac{1}{2}$ then

$$\Pr[X \ne 0] \le 4\sigma^2. \tag{8}$$

Let $a \ge 1$ and Let $k = 2\log_C n - a$ then the probability that $S(x,y) < k$ is the same as the probability that $X = 0$. By 7 this is bounded by

$$\frac{\sigma^2}{\mu^2} \le \frac{3C^{-k}}{n^4 C^{-2k}} = \frac{3C^{-a}}{n^2}.$$

Similarly, if we let $k = 2\log_C n + a$ then the probability that $S(x,y) \ge k$ is the same as the probability $X \ge 1$. By 8 this is at most

$$4\sigma^2 \le 12C^{-k} = \frac{12C^{-a}}{n^2}.$$

Here we use the fact that if $k = 2\log_C n + a$ and $a \ge 1$ then $\mu = C^{-a} \le \frac{1}{2}$. ∎

4 Conclusion and Open Problems

Our initial study of $E[S(x,y)]$ was motivated the algorithmic question considered in [1]: Given x find a y which minimized $S(x,y)$. The first approach was to solve the problem approximately. If we pick y at random we hope that $S(x,y)$ would not be too large. However, our result shows that most of the time this strategy is at least $\log_C n - \log_C \log_C n$ away from the optimal. As it has been shown that, for every x there is a y such that $S(x,y) \le \log_C n + \log_C \log_C n + 12$. Let us define $D(x) = \min_y S(x,y)$, hence $D(x) \le \log_C n + \log_C \log_C n + 12$. What can one say about the average value of $D(x)$ if x is chosen uniformly at random? Can one prove a concentration result for $D(x)$?

Acknowledgments I would like to thank Endre Szemerédi for his constant encouragement and guidance.

References

[1] S. Abbasi and A. Sengupta, *An $O(n \log n)$ algorithm for finding dissimilar strings*, Information Processing Letters, **62**, (1997) 135-139

[2] N. Alon and J. Spencer, *The Probabilistic Method*, (1992) John Wiley, New York.

[3] Chvátal, V. and Sankoff, D. *Longest Common subsequences of two random sequences*, J. Applied Probability, **12**, 306-315

[4] V. Dancik, *Common subsequences and supersequences and their expected length*, Proceedings of 6th Annual Symposium Combinatorial Pattern Matching, Lecture Notes in Computer Science **937** (1995), 55-63.

[5] V. Dancik and M. S. Paterson *Upper bounds for the expected length of a longest common subsequence of two binary sequences*, Random Structures and Algorithms **6** (1995) 449-458.

[6] W. Feller, *An introduction to Probability Theory and its Applications.* Volume I, 3rd ed. John Wiley and Sons, Inc, Canada.

[7] D. M. Freifelder, *Molecular Biology*, (1987) Jones and Bartlett, Boston.

[8] M. S. Waterman, *Introduction to Computational Biology*, (1995) Chapman and Hall, London.