On the Closest String and Substring Problems

MING LI

University of Waterloo, Waterloo, Ont., Canada

BIN MA

University of Western Ontario, London, Ont., Canada

AND

LUSHENG WANG

City University of Hong Kong, Kowloon, Hong Kong, China

Abstract. The problem of finding a center string that is "close" to every given string arises in computational molecular biology and coding theory. This problem has two versions: the Closest String problem and the Closest Substring problem. Given a set of strings $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$, each of length m, the Closest String problem is to find the smallest d and a string s of length m which is within Hamming distance d to each $s_i \in \mathcal{S}$. This problem comes from coding theory when we are looking for a code not too far away from a given set of codes. Closest Substring problem, with an additional input integer L, asks for the smallest d and a string s, of length L, which is within Hamming distance d away from a substring, of length L, of each s_i. This problem is much more elusive than the Closest String problem. The Closest Substring problem is formulated from applications in finding conserved regions, identifying genetic drug targets and generating genetic probes in molecular biology. Whether

The results in Section 2 and some preliminary results in Sections 3 and 4 have been presented in *Proceedings of the 31st Annual ACM Symposium on Theory of Computing* (Atlanta, Ga., May). ACM, New York, 1999, pp. 473–482.

The full solution of the Closest Substring problem in Sections 3 and 4 was presented in *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching* (Montreal, Que., Canada, June). 2000, pp. 99–107.

M. Li was supported in part by the NSERC Research Grant OGP0046506, a CGAT grant, a CITO grant, the E. W. R. Steacie Fellowship, and the NSF/ITR Grant ACI00-85801.

B. Ma was supported in part by NSERC Research Grant OGP0046506 and the HK RGC Grant CityU 1130/99E .

L. Wang, was fully supported by the HK RGC Grant CityU 1130/99E.

Authors' present addresses: M. Li, University of California, Santa Barbara, Department of Computer Science, Santa Barbara, CA 93106, e-mail: mli@cs.ucsb.edu; B. Ma, Department of Computer Science, University of Western Ontario, London, Ont. N6A 5B7, Canada, e-mail: bma@csd.uwo.ca; L. Wang, Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, e-mail: lwang@cs.cityu.edu.hk.

there are efficient approximation algorithms for both problems are major open questions in this area. We present two polynomial-time approximation algorithms with approximation ratio $1 + \epsilon$ for any small ϵ to settle both questions.

Categories and Subject Descriptors: F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems; J.3 [**Life and Medical Sciences**]

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Closest string and substring, computer applications, polynomial-time approximation scheme

1. *Introduction*

Many problems in molecular biology involve finding similar regions common to each sequence in a given set of DNA, RNA, or protein sequences. These problems find applications in locating binding sites and finding conserved regions in unaligned sequences [Stormo and Hartzell 1991; Lawrence and Reilly 1990; Hertz and Stormo 1995; Stormo 1990], genetic drug target identification [Lanctot et al. 1999], designing genetic probes [Lanctot et al. 1999], universal PCR primer design [Lucas et al. 1991; Dopazo et al. 1993; Proutski and Holme 1996; Lanctot et al. 1999], and, outside computational biology, in coding theory [Frances and Litman 1997; Gąsieniec et al. 1999]. Such problems may be considered to be various generalizations of the common substring problem, allowing errors. Many measures have been proposed for finding such regions common to every given string. A popular and most fundamental measure is the Hamming distance. Moreover, two popular objective functions are used in these areas. One is the total sum of distances between the center string (common substring) and each of the given strings. The other is the maximum distance between the center string and a given string. In this article, we focus on the second objective function for Hamming distance. The first objective function for Hamming distance and other measures, like the relative entropy measure used by Stormo and his coauthors [Hertz and Stormo 1995], is treated in Li et al. [2001]. It requires completely different techniques.

Throughout the article, we use a fixed finite alphabet Σ. Let s and s' be finite strings over Σ. Let $d(s, s')$ denote the Hamming distance between s and s'. $|s|$ is the length of s. $s[i]$ is the i-th character of s. Thus, $s = s[1]s[2]\cdots s[|s|]$. The following are the problems we study in this article:

CLOSEST STRING. Given a set $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ of strings each of length m, find a center string s of length m minimizing d such that for every string $s_i \in \mathcal{S}$, $d(s, s_i) \leq d$.

CLOSEST SUBSTRING. Given a set $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ of strings each of length m, and an integer L, find a center string s of length L minimizing d such that for each $s_i \in \mathcal{S}$ there is a length L substring t_i of s_i with $d(s, t_i) \leq d$.

Throughout this article, we call the number d in the definitions of CLOSEST STRING and CLOSEST SUBSTRING the radius (or the cost) of the solution.

CLOSEST STRING has been widely and independently studied in different contexts. In the context of coding theory, it was shown to be NP-hard [Frances and Litman 1997]. In DNA-sequence-related topics, Berman et al. [1997] gave an exact algorithm when the distance d is a constant. Ben-Dor et al. [1997] and Gąsieniec et al. [1999] gave near-optimal approximation algorithms only for large d

(superlogarithmic in number of sequences); however, the straightforward LP (linear programming) relaxation technique does not work when d is small because the randomized rounding procedure introduces large errors. This is exactly the reason why Gąsieniec et al. [1999] and Lanctot et al. [1999] analyzed more involved approximation algorithms, and obtained the ratio $4/3$ approximation algorithms. Note that the small d is the key in applications such as genetic drug target search where we look for similar regions to which a complementary drug sequence would bind. It is a major open problem [Frances and Litman 1997; Berman et al. 1997; Ben-Dor et al. 1997; Gąsieniec et al. 1999; Lanctot et al. 1999] to achieve the best approximation ratio for this problem. (Justifications for using Hamming distance can also be found in these references, especially Lanctot et al. [1999]). We present a polynomial time approximation scheme (PTAS), settling the problem.

CLOSEST SUBSTRING is a more general version of the CLOSEST STRING problem. Obviously, it is also NP-hard. In applications such as drug target identification and genetic probe design, the radius d is usually small. Moreover, when the radius d is small, the center strings can also be used as *motifs* in *repeated-motif* methods for multiple sequence alignment problems [Gusfield 1997; Posfai et al. 1989; Schuler et al. 1991; Waterman 1986; Waterman et al. 1984; Waterman and Perlwitz 1984; Pevzner 2000], that repeatedly find motifs and recursively decompose the sequences into shorter sequences. The problem turns out to be much more elusive than CLOSEST STRING. We extend the techniques developed for closest string here to design a PTAS for the CLOSEST SUBSTRING problem when d is small, that is, $d \leq O(\log(nm))$. Using a *random sampling* technique, and combining our methods for CLOSEST STRING, we then design a PTAS for CLOSEST SUBSTRING, for all d.

It is perhaps cautious to point out that the approximation algorithms developed here may not be directly applicable to bioinformatics practice because of their high time complexity. Heuristics would help to improve the running time while losing some theoretical vigor. We have implemented one such system, COPIA [Liang et al. 2001].

Let us introduce more notations. Let s and t be strings of length m. A multiset $P = \{j_1, j_2, \ldots, j_k\}$ such that $1 \leq j_1 \leq j_2 \leq \cdots \leq j_k \leq m$ is called a *position set*. By $s|_P$ we denote the string $s[j_1]s[j_2]\cdots s[j_k]$. We also write $d^P(s,t)$ to mean $d(s|_P, t|_P)$.

Before ending this section, we present two lemmas that will be used in both Section 2 and Section 4. Lemma 1.1 is commonly known as Chernoff's bounds [Motwani and Raghavan 1995, Theorems 4.2 and 4.3].

LEMMA 1.1 [MOTWANI AND RAGHAVAN 1995]. *Let $X_1, X_2, \ldots, X_n$ be n independent random* 0–1 *variables, where X_i takes* 1 *with probability p_i*, $0 < p_i < 1$. *Let $X = \sum_{i=1}^{n} X_i$, and $\mu = E[X]$. Then, for any $\delta > 0$*,

(1) $\mathbf{Pr}(X > (1+\delta)\mu) < [\frac{e^\delta}{(1+\delta)^{(1+\delta)}}]^\mu$,
(2) $\mathbf{Pr}(X < (1-\delta)\mu) \leq exp(-\frac{1}{2}\mu\delta^2)$.

From Lemma 1.1, we can prove the following lemma:

LEMMA 1.2. *Let X_i, $1 \leq i \leq n$, X and μ be defined as in Lemma* 1.1. *Then for any* $0 < \epsilon \leq 1$,

(1) $\mathbf{Pr}(X > \mu + \epsilon\, n) < exp(-\frac{1}{3}n\epsilon^2)$,
(2) $\mathbf{Pr}(X < \mu - \epsilon\, n) \le exp(-\frac{1}{2}n\epsilon^2)$.

PROOF

(1) Let $\delta = \epsilon n/\mu$. By Lemma 1.1,

$$\begin{aligned}\mathbf{Pr}(X > \mu + \epsilon n) &< \left[\frac{\mathbf{e}^{\epsilon \mathbf{n}/\mu}}{(1+(\epsilon n/\mu))^{(1+(\epsilon n/\mu))}}\right]^{\mu} \\ &= \left[\frac{\mathbf{e}}{(1+(\epsilon n/\mu))^{(1+(\mu/\epsilon n))}}\right]^{\epsilon n} \\ &\le \left[\frac{\mathbf{e}}{(1+\epsilon)^{1+(1/\epsilon)}}\right]^{\epsilon n},\end{aligned}$$

where the last inequality is because $\mu \le n$ and that $(1+x)^{(1+(1/x))}$ is increasing for $x \ge 0$. One can verify that for $0 < \epsilon \le 1$,

$$\frac{\mathbf{e}}{(1+\epsilon)^{1+(1/\epsilon)}} \le \exp\left(-\frac{\epsilon}{3}\right).$$

Therefore, (1) is proved.

(2) Let $\delta = \epsilon n/\mu$. By Lemma 1.1, (2) is proved. □

2. *Approximating* CLOSEST STRING

In this section, we give a PTAS for CLOSEST STRING. Let $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ be a set of n strings, each of length m. We note that a direct application of LP relaxation in Ben-Dor et al. [1997] to all the m positions of the strings does not work when the optimal radius is small. Rather, we extend an idea in Lanctot et al. [1999] to do LP relaxation only to a fraction of the m positions.

We first describe the basic idea behind our approach intuitively. Given a subset of r strings from $\mathcal{S}$. Line them up. Consider the characters where they all agree. Intuitively, there is a high likelihood that the optimal solution should also agree with these characters on corresponding positions. Indeed, this will only slightly worsen the solution. Lemma 2.1 shows that this is true for at least one subset of r strings. Thus, all we need to do is to optimize on the positions (characters) where the r strings do not agree, by LP relaxation and randomized rounding.

Let r be an integer. Roughly, our algorithm will be running in time $n^{O(r)}$ and with approximation ratio $1 + O(1/r)$. So for each fixed r, our algorithm is a polynomial time algorithm. As r grows, our approximation becomes better, but the time complexity also grows.

Let $P = \{j_1, j_2, \ldots, j_k\}$ be a set and $1 \le j_1 \le j_2 \le \cdots \le j_k \le m$. Let s be a string of length m. Recall, we write $s|_P$ to denote the string $s[j_1]s[j_2]\cdots s[j_k]$.

Let s be the optimal solution with radius d_{opt}. For any $s_i \in \mathcal{S}$, if we use s_i as the approximation of s, then $d(s_i, s_j) \le 2d_{opt}$ for any $s_j \in \mathcal{S}$. Thus, this *center-replacing* algorithm has performance ratio 2. Now, we generalize the ratio-2 algorithm by considering k strings $s_{i_1}, s_{i_2}, \ldots, s_{i_k}$ in $\mathcal{S}$ at a time. Let $Q_{i_1,i_2,\ldots,i_k}$ be the set of positions where $s_{i_1}, s_{i_2}, \ldots, s_{i_k}$ agree. We can show that there exist indices $i_1, i_2, \ldots, i_r$ such that the unique characters at positions in $Q_{i_1,i_2,\ldots,i_r}$ form a good approximation of an optimal center string s, that is, for any $s_l \in \mathcal{S}$,

$$d\big(s_l|_{Q_{i_1,i_2,\ldots,i_r}}, s_{i_1}|_{Q_{i_1,i_2,\ldots,i_r}}\big) - d\big(s_l|_{Q_{i_1,i_2,\ldots,i_r}}, s|_{Q_{i_1,i_2,\ldots,i_r}}\big) \le \frac{1}{2r-1} d_{opt}. \qquad (1)$$

Let $P_{i_1,i_2,\ldots,i_k} = \{1, 2, \ldots, m\} - Q_{i_1,i_2,\ldots,i_k}$. Then the size of $P_{i_1,i_2,\ldots,i_k}$ is reduced so that either the LP-relaxization approach or an enumeration approach can work at the positions in $P_{i_1,i_2,\ldots,i_k}$.

To show (1), we do not directly compare s_{i_1} and s at positions in $Q_{i_1,i_2,\ldots,i_r}$. Instead, we consider the relaxed sets $J(l) = \{j \in Q_{i_1,i_2,\ldots,i_r} \mid s_{i_1}[j] \ne s_l[j]$ and $s_{i_1}[j] \ne s[j]\}$ for all $s_l \in \mathcal{S}$. We show that for each $s_l \in \mathcal{S}$, the size of $J(l)$ is small, that is, $|J(l)| \le 1/(2r-1)d_{opt}$.

For any $2 \le k < r$, and $1 \le i_1, i_2, \ldots, i_k \le n$, let $p_{i_1,i_2,\ldots,i_k}$ be the number of mismatches between s_{i_1} and s at the positions in $Q_{i_1,i_2,\ldots,i_k}$. Let

$$\rho_k = \min_{1 \le i_1, i_2, \ldots, i_k \le n} \frac{p_{i_1,i_2,\ldots,i_k}}{d_{opt}}.$$

Increasing k, the number of positions where s_{i_j} all agree decreases, that is, the size of $Q_{i_1,i_2,\ldots,i_k}$ decreases. By definition, ρ_k decreases, too.

Our strategy is to show that (1) $|J(l)| \le (\rho_k - \rho_{k+1})d_{opt}$ for any $s_l \in \mathcal{S}$ and any $2 \le k \le r$; and (2) one of $\rho_2 - \rho_3, \rho_3 - \rho_4, \ldots, \rho_r - \rho_{r+1}$ is at most $1/(2r-1)$.

Let $\rho_0 = \max_{1 \le i,j \le n} d(s_i, s_j)/d_{opt}$. By the triangle inequality, $\rho_0 \le 2$ in worst case. The following lemma shows that (1) holds when the center-replacing algorithm does not give a high-quality approximation, that is, $\rho_0 > 1 + 1/(2r-1)$. (Otherwise, we can simply use the center-replacing algorithm.) This lemma is key to our approximation algorithm.

LEMMA 2.1. *For any constant r, $2 \le r < n$, if $\rho_0 > 1 + (1/(2r-1))$, then there are indices $1 \le i_1, i_2, \ldots, i_r \le n$ such that for any $1 \le l \le n$,*

$$d\big(s_l|_{Q_{i_1,i_2,\ldots,i_r}}, s_{i_1}|_{Q_{i_1,i_2,\ldots,i_r}}\big) - d\big(s_l|_{Q_{i_1,i_2,\ldots,i_r}}, s|_{Q_{i_1,i_2,\ldots,i_r}}\big) \le \frac{1}{2r-1} d_{opt}.$$

PROOF. First, we show the following claim.

CLAIM 2.2. *For any k such that $2 \le k \le r$, where r is a constant, there are indices $1 \le i_1, i_2, \ldots, i_r \le n$ such that for any $s_l \in \mathcal{S}$,*

$$|J(l)| \le (\rho_k - \rho_{k+1})\, d_{opt}.$$

PROOF. Consider the indices $1 \le i_1, i_2, \ldots, i_k \le n$ such that $p_{i_1,i_2,\ldots,i_k} = \rho_k d_{opt}$. Then, for any $1 \le i_{k+1}, i_{k+2}, \ldots, i_r \le n$ and $1 \le l \le n$, we have

$$\begin{aligned}
|J(l)| &= \big|\{j \in Q_{i_1,i_2,\ldots,i_r} \big| s_{i_1}[j] \ne s_l[j] \text{ and } s_{i_1}[j] \ne s[j]\}\big| \\
&\le \big|\{j \in Q_{i_1,i_2,\ldots,i_k} \big| s_{i_1}[j] \ne s_l[j] \text{ and } s_{i_1}[j] \ne s[j]\}\big| \qquad (2)\\
&= \big|\{j \in Q_{i_1,i_2,\ldots,i_k} \big| s_{i_1}[j] \ne s[j]\} - \{j \in Q_{i_1,i_2,\ldots,i_k} \big| s_{i_1}[j] = s_l[j] \\
&\qquad \text{and } s_{i_1}[j] \ne s[j]\}\big| \\
&= \big|\{j \in Q_{i_1,i_2,\ldots,i_k} \big| s_{i_1}[j] \ne s[j]\} - \{j \in Q_{i_1,i_2,\ldots,i_k,l} \big| s_{i_1}[j] \ne s[j]\}\big| \\
&= \big|\{j \in Q_{i_1,i_2,\ldots,i_k} \big| s_{i_1}[j] \ne s[j]\}\big| - \big|\{j \in Q_{i_1,i_2,\ldots,i_k,l} \big| s_{i_1}[j] \ne s[j]\}\big| \qquad (3)\\
&= p_{i_1,i_2,\ldots,i_k} - p_{i_1,i_2,\ldots,i_k,l} \\
&\le (\rho_k - \rho_{k+1})d_{opt},
\end{aligned}$$

where Inequality (2) comes from the fact that $Q_{i_1,i_2,\dots,i_r} \subseteq Q_{i_1,i_2,\dots,i_k}$ and Equality (3) is because $\{j \in Q_{i_1,i_2,\dots,i_k,l} \mid s_{i_1}[j] \neq s[j]\} \subseteq \{j \in Q_{i_1,i_2,\dots,i_k} \mid s_{i_1}[j] \neq s[j]\}$. □

We now give an upper bounded for $\min\{\rho_2 - \rho_3, \rho_3 - \rho_4, \dots, \rho_r - \rho_{r+1}\}$. Consider the sum of the $r-1$ terms,

$$(\rho_2 - \rho_3) + (\rho_3 - \rho_4) + \cdots + (\rho_r - \rho_{r+1}) = \rho_2 - \rho_{r+1} \leq \rho_2 \leq 1.$$

Thus, one of $\rho_2 - \rho_3, \rho_3 - \rho_4, \dots, \rho_r - \rho_{r+1}$ is at most $1/(r-1)$. We can give a better bound by considering the weighted average of r terms $\rho_0 - 1$, $\rho_2 - \rho_3$, $\rho_3 - \rho_4, \dots, \rho_r - \rho_{r+1}$.

CLAIM 2.3. *For* $2 \leq r < n$, *min*$\{\rho_0 - 1, \rho_2 - \rho_3, \rho_3 - \rho_4, \dots, \rho_r - \rho_{r+1}\} \leq 1/(2r-1)$.

PROOF. Consider $1 \leq i, j \leq n$ such that $d(s_i, s_j) = \rho_0 d_{opt}$. Then, among the positions where s_i mismatches s_j, for at least one of the two strings, say, s_i, the number of mismatches between s_i and s is at least $(\rho_0/2)d_{opt}$. Thus, among the positions where s_i matches s_j, the number of mismatches between s_i and s is at most $1 - (\rho_0/2)d_{opt}$. Therefore, $\rho_2 \leq 1 - (\rho_0/2)$. So,

$$\frac{(1/2)(\rho_0 - 1) + (\rho_2 - \rho_3) + (\rho_3 - \rho_4) + \cdots + (\rho_r - \rho_{r+1})}{(1/2) + r - 1}$$
$$\leq \frac{(1/2)\rho_0 + \rho_2 - 1/2}{r - 1/2} \leq \frac{1}{2r-1}.$$

Thus, at least one of $\rho_0 - 1$, $\rho_2 - \rho_3$, $\rho_3 - \rho_4, \dots, \rho_r - \rho_{r+1}$ is less than or equal to $1/(2r-1)$. □

If $\rho_0 > 1 + 1/(2r-1)$, then from Claim 2.3, there must be a $2 \leq k \leq r$ such that $\rho_k - \rho_{k+1} \leq 1/(2r-1)$. From Claim 2.2, $|J(l)| \leq 1/(2r-1)\, d_{opt}$. Hence, there are at most $1/(2r-1)\, d_{opt}$ positions in $Q_{i_1,i_2,\dots,i_r}$ where s_l differs from s_{i_1} while agrees with s. The lemma is proved. □

Lemma 2.1 gives a way to find a good approximation of an optimal center string s at positions in $Q_{i_1,i_2,\dots,i_r}$ for some $i_1, i_2, \dots, i_r$. Lemma 2.5 shows how to use this partial approximation solution from Lemma 2.1 to construct a good center string at all L positions.

Let $P_{i_1,i_2,\dots,i_k} = \{1, 2, \dots, m\} - Q_{i_1,i_2,\dots,i_k}$. We have

CLAIM 2.4. $|P_{i_1,i_2,\dots,i_k}| \leq kd_{opt}$ *and* $|Q_{i_1,i_2,\dots,i_k}| \geq m - kd_{opt}$.

PROOF. Let q be a position where some of $s_{i_1}, s_{i_2}, \dots, s_{i_k}$ do not agree. Then there exists some s_{i_j} such that $s_{i_j}[q] \neq s[q]$. Since $d(s_{i_j}, s) \leq d_{opt}$, each s_{i_j} contributes at most d_{opt} positions in $P_{i_1,i_2,\dots,i_k}$. Thus, $|P_{i_1,i_2,\dots,i_k}| \leq kd_{opt}$. By definition, $|Q_{i_1,i_2,\dots,i_k}| \geq m - kd_{opt}$. □

For the positions in $P_{i_1,i_2,\dots,i_r} = \{1, 2, \dots, m\} - Q_{i_1,i_2,\dots,i_r}$, we use ideas in Lanctot et al. [1999] and employ the following two strategies: (1) if $|P_{i_1,i_2,\dots,i_r}|$ is large, that is, $d_{opt} > \Omega(\log m)$, we use the LP relaxation to approximate s, since $|P_{i_1,i_2,\dots,i_k}| \leq kd_{opt}$ and thus the conditions for applying the method are satisfied; (2) if $|P_{i_1,i_2,\dots,i_r}|$ is small, that is, $d_{opt} \leq O(\log m)$, we can enumerate $|\Sigma|^{|P_{i_1,i_2,\dots,i_r}|}$ possibilities to approximate s.

LEMMA 2.5. *Let $\mathcal{S} = \{s_1, s_2, \ldots s_n\}$, where $|s_i| = m$ for $i = 1, \ldots, n$. Assume that s is the optimal solution of* CLOSEST STRING *and $\max_{1\le i\le n} d(s_i, s) = d_{opt}$. Given a string s' and a position set Q of size $m - O(d_{opt})$ such that for any $i = 1, \ldots, n$,*

$$d(s_i|_Q, s'|_Q) - d(s_i|_Q, s|_Q) \le \rho\, d_{opt}, \tag{4}$$

where $0 \le \rho \le 1$, one can obtain a solution with radius at most $(1 + \rho + \epsilon)d_{opt}$ in polynomial time for any fixed $\epsilon \ge 0$. The $\epsilon\, d_{opt}$ is considered as the error of the solution.

PROOF. Let $P = \{1, 2, \ldots, m\} - Q$. Then, for any two strings t and t' of length m, we have $d(t|_P, t'|_P) + d(t|_Q, t'|_Q) = d(t, t')$. Thus, for any $i = 1, 2, \ldots, n$,

$$\begin{aligned} d(s_i|_P, s|_P) &= d(s_i, s) - d(s_i|_Q, s|_Q) \\ &\le d(s_i, s) - (d(s_i|_Q, s'|_Q) - \rho\, d_{opt}) \\ &\le (1+\rho)\, d_{opt} - d(s_i|_Q, s'|_Q). \end{aligned}$$

Therefore, the following optimization problem

$$\begin{cases} \min \;\; d; \\ d(s_i|_P, y) \le d - d(s_i|_Q, s'|_Q), \quad i = 1, \ldots, n; |y| = |P|, \end{cases} \tag{5}$$

has a solution $y = s|_P$ with cost $d \le (1+\rho)d_{opt}$. Note that y is a variable that represents a string of length $|P|$. Suppose the optimal solution for the optimization problem (5) is such that the cost is d_0. Then

$$d_0 \le (1+\rho)d_{opt}. \tag{6}$$

Now we solve (5) approximately. Let π be a permutation on $1, 2, \ldots, m$. For any length m string t, we use t^π to denote the string $t[\pi(1)]t[\pi(2)]\ldots t[\pi(m)]$. Clearly, s is an optimal center string for $\{s_1, s_2, \ldots, s_n\}$ if and only if s^π is an optimal center string for $\{s_1^\pi, s_2^\pi, \ldots, s_n^\pi\}$ with the same cost d. Therefore, without loss of generality, we assume that $P = \{1, 2, \ldots, |P|\}$. Similar to Ben-Dor et al. [1997] and Lanctot et al. [1999], we use a 0-1 variable $y_{j,a}$ to indicate whether $y[j] = a$, where $1 \le j \le |P|$ and $a \in \Sigma$. Denote $\chi(s_i[j], a) = 0$ if $s_i[j] = a$ and 1 if $s_i[j] \ne a$. Then, (5) can be formulated to an equivalent 0–1 optimization problem as follows:

$$\begin{cases} \min \;\; d; \\ \sum_{a\in\Sigma} y_{j,a} = 1, \quad j = 1, 2, \ldots, |P|, \\ \sum_{1\le j\le |P|} \sum_{a\in\Sigma} \chi(s_i[j], a)\, y_{j,a} \le d - d(s_i|_Q, s'|_Q), \quad i = 1, 2, \ldots, n. \end{cases} \tag{7}$$

Solve (7) by linear programming to get a fractional solution $y_{j,a} = \overline{y}_{j,a}$ $(1 \le j \le |P|, a \in \Sigma)$ with cost $\overline{d}$. Clearly $\overline{d} \le d_0$. For each $1 \le j \le |P|$, independently, with probability $\overline{y}_{j,a}$, set $y'_{j,a} = 1$ and $y'_{j,a'} = 0$ for any $a' \in \Sigma - \{a\}$. Then, we get a solution $y_{j,a} = y'_{j,a}$ $(1 \le j \le m, a \in \Sigma)$ for the 0–1 optimization problem (7); hence a solution $y = y'$ for (5).

For each $1 \le j \le |P|$, the above random rounding ensures that only one $a \in \Sigma$ is such that $y'_{j,a} = 1$. Therefore, $\sum_{a\in\Sigma} \chi(s_i[j], a)\, y'_{j,a}$ takes 1 or 0 randomly. Since the rounding is independent for different j's, $\sum_{a\in\Sigma}\chi(s_i[j],a)\,y'_{j,a}$ are independent 0–1 random variables for $1\le j\le|P|$. Thus, for any $1\le i\le n$, $d(s_i|_P,y')=\sum_{1\le j\le|P|}\sum_{a\in\Sigma}\chi(s_i[j],a)\,y'_{j,a}$ is a sum of $|P|$ independent 0–1 random variables, and

$$\begin{aligned}E[d(s_i|_P,y')] &= \sum_{1\le j\le|P|}\sum_{a\in\Sigma}\chi(s_i[j],a)\,E[y'_{j,a}]\\ &= \sum_{1\le j\le|P|}\sum_{a\in\Sigma}\chi(s_i[j],a)\,\overline{y}_{j,a}\\ &\le \overline{d}-d(s_i|_Q,s'|_Q)\le d_0-d(s_i|_Q,s'|_Q).\end{aligned}\tag{8}$$

Therefore, for any fixed $\epsilon'>0$, by Lemma 1.2,

$$\mathbf{Pr}\,(d(s_i|_P,y')\ge d_0+\epsilon'|P|-d(s_i|_Q,s'|_Q))\le\exp\left(-\frac{1}{3}\epsilon'^2|P|\right).$$

Considering all of the n strings, we have

$$\begin{aligned}\mathbf{Pr}\,(d(s_i|_P,y') &\ge d_0+\epsilon'|P|-d(s_i|_Q,s'|_Q)\text{ for some }1\le i\le n)\\ &\le n\times\exp\left(-\frac{1}{3}\epsilon'^2|P|\right).\end{aligned}\tag{9}$$

If $|P|\ge(4\ln n)/\epsilon'^2$, then, $n\times\exp(-\frac{1}{3}\epsilon'^2|P|)\le n^{-1/3}$. Thus, we obtain a randomized algorithm to find a solution for (5) with cost at most $d_0+\epsilon'|P|$ with probability at least $1-n^{-1/3}$.

For binary strings, Raghavan [1988] gave a method to derandomize the above randomized algorithm. Also see Motwani and Raghavan [1995]. Here we give a method to derandomize the algorithm for any constant-size alphabet.

Let $y_{j,a}=\overline{y}_{j,a}$ be a fractional solution for (7) for $j\in P$. We can arbitrarily decompose P into disjoint sets $P_1,P_2,\dots,P_k$ such that $|P_i|\ge(4\ln n)/\epsilon'^2$ and $|P_i|<(8\ln n)/\epsilon'^2$ for $i=1,2,\dots,k$. Let $\mu_{l,i}=\sum_{1\le j\le|P_i|}\sum_{a\in\Sigma}\chi(s_l[j],a)\,y_{j,a}$. For each P_i, replacing P with P_i in (9), we know that there exists a string x_i of length $|P_i|$ such that for each $s_l\in\mathcal{S}$,

$$d(s_l|_{P_i},x_i)\le\mu_{l,i}+\epsilon'|P_i|.$$

Thus, we can simply enumerate all string of length $|P_i|$ to find out x_i in polynomial time for any constant size alphabet. Concatenating all x_i's, we have a string x of length $|P|$ such that for each string $s_l\in\mathcal{S}$,

$$d(s_l|_P,x)\le\sum_{i=1}^{k}(\mu_{l,i}+\epsilon'|P_i|)=d_0-d(s_l|_Q,s'|_Q)+\epsilon'|P|.$$

Thus, we obtain a desired string x in polynomial time.

If $|P|<(4\ln n)/\epsilon'^2$, $|\Sigma|^{|P|}<n^{(4\ln|\Sigma|)/\epsilon'^2}$ is polynomial in n. So, we can enumerate all strings in $\Sigma^{|P|}$ to find an optimal solution for (5) in polynomial time. Thus, in both cases, we can obtain a solution $y=y_0$ for the optimization problem (5) with cost at most $d_0+\epsilon'|P|$ in polynomial time. Since $|P|=O(d_{opt})$, $|P|\le c\times d_{opt}$ for a constant c. Let $\epsilon'=\epsilon/c$. Let s^* be the string such that $s^*|_Q=s'|_Q$ and $s^*|_P=y_0$.

```
                              Algorithm closestString
Input      $s_1, s_2, \ldots, s_n \in \Sigma^m$, an integer $r \geq 2$ and a small number $\epsilon > 0$.
Output     a center string $s \in \Sigma^m$.
1.  for each $r$-element subset $\{s_{i_1}, s_{i_2}, \ldots, s_{i_r}\}$ of the $n$ input strings do
      (a) $Q = \{1 \leq j \leq m \mid s_{i_1}[j] = s_{i_2}[j] = \ldots = s_{i_r}[j]\}$, $P = \{1, 2, \ldots, m\} - Q$.
      (b) Let $s' = s_{i_1}$. Solve the optimization problem defined by (5) as described in the proof
          of Lemma 2.5 to get an approximate solution $y = y_0$ within error $\epsilon\, d_{opt}$.
      (c) Let $u$ be a string such that $u|_Q = s_{i_1}|_Q$ and $u|_P = y_0$. Calculate the radius of the
          solution with $u$ as the center string.
2.  for $i = 1, 2, \ldots, n$ do
        calculate the radius of the solution with $s_i$ as the center string.
3.  Output the best solution of the above two steps.
```

FIG. 1. Algorithm for CLOSEST STRING.

From (5),

$$\begin{aligned} d(s_i, s^*) &= d(s_i|_P, s^*|_P) + d(s_i|_Q, s^*|_Q) \\ &= d(s_i|_P, y_0) + d(s_i|_Q, s'|_Q) \\ &\leq d_0 + \epsilon'|P| \leq (1+\rho)d_{opt} + \epsilon d_{opt}, \end{aligned}$$

where the last inequality is from (6). This proves the lemma. □

Now we describe the complete algorithm in Figure 1.

THEOREM 2.6. *The algorithm closestString is a PTAS for* CLOSEST STRING.

PROOF. Given an instance of CLOSEST STRING, suppose s is an optimal solution and the optimal radius is d_{opt}, that is, $\max_{i=1,\ldots,n} d(s, s_i) = d_{opt}$. Let P be defined as in Step 1(a) of Algorithm closestString. Since for every position in P, at least one of the r strings $s_{i_1}, s_{i_2}, \ldots, s_{i_r}$ conflicts with the optimal center string s, we have $|P| \leq r \times d_{opt}$. As far as r is a constant, Step 1(b) can be done in polynomial time by Lemma 2.5. Obviously, the other steps of Algorithm closestString runs in polynomial time, with r a constant. In fact, it can be easily verified that the algorithm runs in time $O((nm)^r\, n^{O(\log|\Sigma| \times r^2/\epsilon^2)})$.

If $\rho_0 - 1 \leq 1/(2r-1)$, then by the definition of ρ_0, the algorithm finds a solution with radius at most $\rho_0 d_{opt} \leq (1 + (1/(2r-1)))d_{opt}$ in Step 2.

If $\rho_0 > 1 + (1/(2r-1))$, then from Lemma 2.1 and Lemma 2.5, the algorithm finds a solution with radius at most $(1 + (1/(2r-1) + \epsilon))d_{opt}$ in Step 1.

Therefore, the performance ratio of Algorithm closestString is $(1 + (1/(2r-1)) + \epsilon)d_{opt}$. This proves the theorem. □

3. *Approximating* CLOSEST SUBSTRING *when D is Small*

In some applications, such as drug target identification, genetic probe design, the radius d is often small. As a direct application of Lemma 2.1, we now present a PTAS for CLOSEST STRING when the radius d is small, that is, $d \leq O(\log(nm))$. Again, we focus on the construction of the center string.

Suppose $\langle \mathcal{S}, L \rangle$ is the given instance, where $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$. Let s be an optimal center string and t_i be the substring from s_i which is the closest to s $(1 \leq i \leq n)$.

```
Algorithm smallSubstring
Input    $s_1, s_2, \ldots, s_n \in \Sigma^m$, an integer $L$ and an integer $r > 0$.
Output   a center string $s \in \Sigma^L$.
1. for each $r$-element subset $\{t_{i_1}, t_{i_2}, \ldots, t_{i_r}\}$, where $t_{i_j}$ is a substring of length $L$ from $s_{i_j}$ do
     (a) $Q = \{1 \le j \le L \mid t_{i_1}[j] = t_{i_2}[j] = \ldots = t_{i_r}[j]\}$, $P = \{1, 2, \ldots, L\} - Q$.
     (b) for every $x \in \Sigma^{|P|}$ do
           let $t$ be the string such that $t|_Q = t_{i_1}|_Q$ and $t|_P = x|_P$; compute the radius of the
           solution with $t$ as the center string.
2. for every length $L$ substring $u$ from any given sequence do
       compute the radius of the solution with $u$ as the center string.
3. select a center string that achieves the best result in Step 1 and Step 2; output the best
   solution of the above two steps.
```

FIG. 2. Algorithm for CLOSEST SUBSTRING when d is small.

We first describe the intuition behind the proof. The basic idea is that for any fixed constant $r > 0$, by trying all the choices of r substrings from $\mathcal{S}$, we can assume that we know the $t_{i_1}, t_{i_2}, \ldots, t_{i_r}$ that satisfy Lemma 2.1 by replacing s_i by t_i and s_{i_j} by t_{i_j}. Since $d \le O(\log(nm))$, the r substrings $t_{i_1}, t_{i_2}, \ldots, t_{i_r}$ disagree at at most $O(\log(nm))$ positions. Therefore, we can keep the characters at the positions where $t_{i_1}, t_{i_2}, \ldots, t_{i_r}$ all agree, and try all possibilities for the rest of the positions. From Lemma 2.1, we will find a good approximation to s. The complete algorithm is described in Figure 2.

THEOREM 3.1. *Algorithm smallSubstring is a PTAS for* CLOSEST SUBSTRING *when the radius d is small, that is, $d \le O(log(nm))$.*

PROOF. Obviously, the size of P in Step 1 is at most $O(r \times \log(nm))$. Step 1 takes $O((mn)^r \times \Sigma^{O(r \times \log(nm))} \times mnL) = O((nm)^{r+1} \times (nm)^{O(r \times \log|\Sigma|)}) = O((nm)^{O(r \times \log|\Sigma|)})$ time. Other steps take less than that time. Thus, the total time required is $O((nm)^{O(r \times \log|\Sigma|)})$, which is polynomial in terms of input size for any constant r.

From Lemma 2.1, the performance ratio of the algorithm is $1 + (1/(2r - 1))$. □

4. *A PTAS for* CLOSEST SUBSTRING

In this section, we further extend the algorithm for CLOSEST STRING to a PTAS for CLOSEST SUBSTRING, making use of a *random sampling* strategy. Note that Algorithm smallSubstring runs in exponential time for general radius d and the algorithm closestString does not work for CLOSEST SUBSTRING since we do not know how to construct an optimization problem similar to (5)—The construction of (5) requires us to know all the n strings (substrings) in an optimal solution of CLOSEST STRING (CLOSEST SUBSTRING). Thus, the choice of a "good" substring from every string s_i is the only obstacle on the way to the solution. For each $s_i \in \mathcal{S}$, there are $O(m)$ ways to choose t_i. Thus, we have $O(m^n)$ possible combinations, which is exponential in terms of n. We use random sampling to handle this.

Now let us outline the main ideas. Let $\langle \mathcal{S} = \{s_1, s_2, \ldots, s_n\}, L\rangle$ be an instance of CLOSEST SUBSTRING, where each s_i is of length m, for $1 \le i \le n$. Suppose that s, of length L, is its optimal solution and t_i is a length L substring of s_i which is the closest to s ($1 \le i \le n$). Let $d_{opt} = \max_{i=1}^n d(s, t_i)$. The key step of our

```
Algorithm closestSubstring
Input     $s_1, s_2, \ldots, s_n \in \Sigma^m$, an integer $1 \le L \le m$, an integer $r \ge 2$ and a small number $\epsilon > 0$.
Output    the center string $s$.
1.  for each $r$-element subset $\{t_{i_1}, t_{i_2}, \ldots, t_{i_r}\}$, where $t_{i_j}$ is a substring of length $L$ from $s_{i_j}$ do
        (a) $Q = \{1 \le j \le L \mid t_{i_1}[j] = t_{i_2}[j] = \ldots = t_{i_r}[j]\}$, $P = \{1, 2, \ldots, L\} - Q$.
        (b) Let $R$ be a multiset containing $\lceil \frac{4}{\epsilon^2} \log(nm) \rceil$ uniformly random positions from $P$.
        (c) for every string $x$ of length $|R|$ do
            (i) for $i$ from 1 to $n$ do
                Let $t'_i$ be a length $L$ substring of $s_i$ minimizing $f(t'_i) = d(x, t'_i|_R) \times \frac{|P|}{|R|} + d(t_{i_1}|_Q, t'_i|_Q)$.
            (ii) Using the method provided in the proof of Lemma 2.5, solve the optimization problem
                 defined by (10) approximately to get a solution $y = y_0$ within error $\epsilon\,|P|$.
            (iii) Let $s'$ be the string such that $s'|_P = y_0$ and $s'|_Q = t_{i_1}|_Q$. Let $c = \max_{i=1}^n$
                  $\min_{\{t_i \text{ is a substring of } s_i\}} d(s', t_i)$.
2.  for every length-$L$ substring $s'$ of $s_1$ do
        Let $c = \max_{i=1}^n \min_{\{t_i \text{ is a substring of } s_i\}} d(s', t_i)$.
3.  Output the $s'$ with minimum $c$ in Step 1(c)(iii) and Step 2.
```

FIG. 3. The PTAS for the closest substring problem.

algorithm is to construct a function $f(u)$ for u being a length L substring from any $s_i \in \mathcal{S}$, such that $f(u)$ approximates $d(s, u)$ very well. Therefore, we can find an "approximation" to each t_i by minimizing $f(u)$ for u being a substring from s_i $(1 \le i \le n)$.

More specifically, for any fixed $r > 0$, by trying all the choices of r substrings from $\mathcal{S}$, we can assume that $t_{i_1}, t_{i_2}, \ldots, t_{i_r}$ are the r substrings that satisfy Lemma 2.1 by replacing s_i by t_i and s_{i_j} by t_{i_j}. Let Q be the set of positions where $t_{i_1}, t_{i_2}, \ldots, t_{i_r}$ agree and $P = \{1, 2, \ldots, L\} - Q$. By Lemma 2.1, $t_{i_1}|_Q$ is a good approximation to $s|_Q$. We want to approximate $s|_P$ by the solution of y of the following optimization problem (10), where t'_i is a substring of s_i and is up to us to choose.

$$\begin{cases} \min \quad d; \\ d(t'_i|_P, y) \le d - d(t'_i|_Q, t_{i_1}|_Q), \quad i = 1, \ldots, n; |y| = |P|. \end{cases} \tag{10}$$

Note that t'_i's are not variables of problem (10), but rather they define the constraints. We need to choose them properly to construct the right problem (10) to solve. The ideal choice is $t'_i = t_i$, that is, t'_i is the closest to s among all substrings of s_i. However, we only approximately know the corresponding characters of s in Q and know nothing about those positions of s in P so far. So, we randomly pick $O(\log(mn))$ positions from P. Suppose the multiset of these random positions is R. (Here we use multiset, since each time we have to independently choose a position from P without changing the distribution. Thus, we can estimate the expected value of the $|R|$ independent random variables and apply the Chernoff bound and Lemma 1.2 in our proof.) By trying all length $|R|$ strings, we can assume that we know $s|_R$. Then, for each $1 \le i \le n$, we find the substring t'_i from s_i such that $f(t'_i) = d(s|_R, t'_i|_R) \times |P|/|R| + d(t_{i_1}|_Q, t'_i|_Q)$ is minimized. Then, t'_i potentially belongs to the substrings which are the closest to s.

Then we solve (10) approximately by the method provided in the proof of Lemma 2.5 to get a solution $y = y_0$ and combine y_0 at the positions in P and t_{i_1} at the positions in Q. The resulting string should be a good approximation to s. The detailed algorithm (Algorithm closestSubstring) is given in Figure 3. We prove Theorem 4.1 in the rest of the section.

THEOREM 4.1. *Algorithm closestSubstring is a PTAS for the closest substring problem.*

PROOF. Let s be an optimal center string and t_i be the length-L substring of s_i that is the closest to s ($1 \le i \le n$). Let $d_{opt} = \max_{1\le i\le n} d(s, t_i)$. Let ϵ be any small positive number and $r \ge 2$ be any fixed integer. Let $\rho_0 = \max_{1\le i,j\le n} d(t_i, t_j)/d_{opt}$. If $\rho_0 \le 1 + (1/(2r-1))$, then clearly we can find a solution s' within ratio ρ_0 in Step 2. So, we assume that $\rho_0 \ge 1 + (1/(2r-1))$ from now on.

FACT 1. *There exists a group of substrings $t_{i_1}, t_{i_2}, \ldots, t_{i_r}$ chosen in Step 1 of Algorithm closestSubstring, such that for any $1 \le l \le n$, $d(t_l|_Q, t_{i_1}|_Q) - d(s_l|_Q, s|_Q) \le 1/(2r-1)d_{opt}$, where Q is defined in Step 1(a) of the algorithm.*

PROOF. The fact follows from Lemma 2.1 directly. □

Obviously, the algorithm takes x as $s|_R$ at some point in Step 1(c). Let $x = s|_R$ and $t_{i_1}, t_{i_2}, \ldots, t_{i_r}$ satisfy Fact 1. Let P and Q be defined as in Step 1(a), and t_i' be defined as in Step 1(c)(i). Let s^* be a string such that $s^*|_P = s|_P$ and $s^*|_Q = t_{i_1}|_Q$. Then we claim:

FACT 2. *With high probability, $d(s^*, t_i') \le d(s^*, t_i) + 2\epsilon|P|$ for all $1 \le i \le n$.*

PROOF. Recall, for any position multiset T, we denote $d^T(t_1, t_2) = d(t_1|_T, t_2|_T)$ for any two strings t_1 and t_2. Let $\rho = |P|/|R|$. Consider any length L substring t' of s_i satisfying

$$d(s^*, t') \ge d(s^*, t_i) + 2\epsilon|P|. \tag{11}$$

Let $f(u) = \rho d^R(s^*, u) + d^Q(t_{i_1}, u)$ for any length L string u. The fact $f(t') \le f(t_i)$ implies either $f(t') \le d(s^*, t') - \epsilon|P|$ or $f(t_i) \ge d(s^*, t_i) + \epsilon|P|$. Thus, we have the following inequality:

$$\begin{aligned}&\mathbf{Pr}(f(t') \le f(t_i))\\ &\quad \le \mathbf{Pr}(f(t') \le d(s^*, t') - \epsilon|P|) + \mathbf{Pr}(f(t_i) \ge d(s^*, t_i) + \epsilon|P|).\end{aligned} \tag{12}$$

It is easy to see that $d^R(s^*, t')$ is the sum of $|R|$ independent random 0-1 variables $\sum_{i=1}^{|R|} X_i$, where $X_i = 1$ indicates a mismatch between s^* and t' at the ith position in R. Let $\mu = E[d^R(s^*, t')]$. Obviously, $\mu = d^P(s^*, t')/\rho$. Therefore, by Lemma 1.2 (2),

$$\begin{aligned}&\mathbf{Pr}(f(t') \le d(s^*, t') - \epsilon|P|)\\ &\quad = \mathbf{Pr}(\rho d^R(s^*, t') + d^Q(s^*, t') \le d(s^*, t') - \epsilon|P|)\\ &\quad = \mathbf{Pr}(\rho d^R(s^*, t') \le d^P(s^*, t') - \epsilon|P|)\\ &\quad = \mathbf{Pr}(d^R(s^*, t') \le \mu - \epsilon|R|)\\ &\quad \le \exp\left(-\frac{1}{2}\epsilon^2|R|\right) \le (nm)^{-2},\end{aligned} \tag{13}$$

where the last inequality is due to the setting $|R| = \lceil 4/\epsilon^2 \log(nm)\rceil$ in Step 1(b) of the algorithm. Similarly, using Lemma 1.2 (1), we have

$$\mathbf{Pr}(f(t_i) \ge d(s^*, t_i) + \epsilon|P|) \le (nm)^{-4/3}. \tag{14}$$

Combining (12), (13), and (14), we know that for any t' that satisfies (11),

$$\mathbf{Pr}(f(t') \le f(t_i)) \le 2\,(nm)^{-4/3}. \tag{15}$$

For any fixed $1 \le i \le n$, there are less than m substrings t' that satisfy (11). Thus, from (15) and the definition of t_i',

$$\mathbf{Pr}(d(s^*, t_i') \ge d(s^*, t_i) + 2\epsilon|P|) \le 2\,n^{-4/3}m^{-1/3}. \tag{16}$$

Summing up all $i \in \{1, 2, \dots, n\}$, we know that with probability at least $1 - 2\,(nm)^{-1/3}$, $d(s^*, t_i') \le d(s^*, t_i) + 2\epsilon|P|$ for all $i = 1, 2, \dots, n$. $\square$

From Fact 1, $d(s^*, t_i) = d^P(s, t_i) + d^Q(t_{i_1}, t_i) \le d(s, t_i) + 1/(2r-1)\,d_{opt}$. Combining with Fact 2 and $|P| \le r\,d_{opt}$, we get

$$d(s^*, t_i') \le (1 + (1/(2r-1)) + 2\epsilon\,r)d_{opt}. \tag{17}$$

By the definition of s^*, $s|_P$ is a solution for the optimization problem defined by (10) such that $d \le (1 + (1/(2r-1)) + 2\epsilon\,r)d_{opt}$. We can solve the optimization problem (10) within error $\epsilon\,|P| \le \epsilon\,rd_{opt}$ by the method in the proof of Lemma 2.5. Let $y = y_0$ be the approximate solution of the optimization problem in (10). Then, by (10), for any $1 \le i \le n$,

$$d(t_i'|_P, y_0) \le (1 + (1/(2r-1)) + 2\epsilon\,r)d_{opt} - d(t_i'|_Q, t_{i_1}|_Q) + \epsilon\,rd_{opt}. \tag{18}$$

Let s' be defined in Step 1(c)(iii), then, by (18),

$$\begin{aligned} d(s', t_i') &= d(y_0, t_i'|_P) + d(t_{i_1}|_Q, t_i'|_Q) \\ &\le (1 + (1/(2r-1)) + 2\epsilon r)d_{opt} + \epsilon\,rd_{opt} \\ &\le (1 + (1/(2r-1)) + 3\epsilon r)d_{opt}. \end{aligned}$$

For any small constant $\delta > 0$, by setting $r = 1/\delta + 1/2$ and $\epsilon = \delta^2/6$, we assure that $1/(2r-1) + 3\epsilon r \le \delta$. Therefore, with high probability, Algorithm closest-Substring outputs in polynomial time a center string s' such that $d(t_i', s')$ is no more than $(1+\delta)d_{opt}$ for every $1 \le i \le n$, where t_i' is a substring of s_i. It can be easily verified that the running time of the algorithm is $O((n^2m)^{O((\log|\Sigma|)/\delta^4)})$.

The algorithm can be derandomized by standard methods. For instance, instead of randomly and independently choosing $O(\log(nm))$ positions from P, we can pick the vertices encountered on a random walk of length $O(\log(nm))$ on a constant degree expander [Gillman 1993]. Obviously, the number of such random walks on a constant degree expander is polynomial in terms of nm. Thus, by enumerating all random walks of length $O(\log(nm))$, we have a polynomial time deterministic algorithm. (Also see Arora et al. [1995]). $\square$

5. *Discussions*

We have designed polynomial time approximation schemes for both the closest string and closest substring problems. A more general problem called *distinguishing string selection* problem is to search for a string of length L which is "close" (as an approximate substring) to some sequences (of harmful germs) and "far" from some other sequences (of humans). See Lanctot et al. [1999] for a formal definition. The problem is originated from genetic drug target identification, and universal

PCR primer design [Lucas et al. 1991; Dopazo et al. 1993; Proutski and Holme 1996]. A trivial ratio-2 approximation was given in Lanctot et al. [1999] and the first nontrivial algorithm with approximation ratio $2 - (2/(2|\Sigma| + 1))$ was given in Li et al. [1999]. An open problem is whether a polynomial-time approximation scheme exists for this problem.

ACKNOWLEDGMENTS. We would like to thank Tao Jiang, Kevin Lanctot, Joe Wang, and Louxin Zhang for discussions and suggestions on related topics.

REFERENCES

ARORA, S., KARGER, D., AND KARPINSKI, M. 1995. Polynomial time approximation schemes for dense instances of NP-hard problems. In *Proceedings of the 27th Annual ACM Symposium on Theory of Computing*. ACM, New York, pp. 284–293.

BEN-DOR, A., LANCIA, G., PERONE, J., AND RAVI, R. 1997. Banishing bias from consensus sequences. In *Proceedings of the 8th Annual Combinatorial Pattern Matching Conference*. pp. 247–261.

BERMAN, P., GUMUCIO, D., HARDISON, R., MILLER, W., AND STOJANOVIC, N. 1997. A linear-time algorithm for the 1-mismatch problem. In *Proceedings of the Workshops on Algorithms and Data Structures*. pp. 126–135.

DOPAZO, J., RODRÍGUEZ, A., SÁIZ, J. C., AND SOBRINO, F. 1993. Design of primers for PCR amplification of highly variable genomes. *CABIOS 9*, 123–125.

FRANCES, M., AND LITMAN, A. 1997. On covering problems of codes. *Theoret. Comput. Syst. 30*, 113–119.

GĄSIENIEC, L., JANSSON, J., AND LINGAS, A. 1999. Efficient approximation algorithms for the Hamming center problem. In *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms*. ACM, New York, pp. S905–S906.

GILLMAN, D. 1993. A Chernoff bound for random walks on expanders. In *Proceedings of the 34th Annual Symposium on Foundations of Computer Science.* IEEE Computer Society Press, Los Alamitos, Calif., pp. 680–691.

GUSFIELD, D. 1993. Efficient methods for multiple sequence alignment with guaranteed error bounds. *Bull. Math. Biol. 30*, 141–154.

GUSFIELD, D. 1997. *Algorithms on Strings, Trees, and Sequences*. Cambridge Univ. Press.

HERTZ, G., AND STORMO, G. 1995. Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps. In *Proceedings of the 3rd International Conference on Bioinformatics and Genome Research*. pp. 201–216.

LAWRENCE, C., AND REILLY, A. 1990. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins 7*, pp. 41–51.

LUCAS, K., BUSCH, M., MÖSSINGER, S., AND THOMPSON, J. A. 1991. An improved microcomputer program for finding gene- or gene family-specific oligonucleotides suitable as primers for polymerase chain reactions or as probes. *CABIOS*, 7, 525–529.

LANCTOT, K., LI, M., MA, B., WANG, S., AND ZHANG, L. 1999. Distinguishing string selection problems. In *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms*. ACM, New York, pp. 633–642. Also to appear in *Inf. Comput.*

LI, M., MA, B., AND WANG, L. 1999. Finding similar regions in many strings. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing* (Atlanta, Ga.). ACM, New York, pp. 473–482.

LI, M., MA, B., AND WANG, L. 2001. Finding similar regions in many sequences. *J. Comput. Syst. Sci.* (special issue for *31st Annual ACM Symposium on Theory of Computing*) to appear.

LIANG, C., LI, M., AND MA, B. 2001. COPIA—A consensus pattern identification and analysis software system. Manuscript. Software available at: http://dna.cs.ucsb.edu/copia/copia_submit.html.

MA, B. 2000. A polynomial time approximation scheme for the closest substring problem. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching* (Montreal, Ont., Canada). pp. 99–107.

MOTWANI, R., AND RAGHAVAN, P. 1995. *Randomized Algorithms*, Cambridge Univ. Press.

PEVZNER, P. A. 2000. *Computational molecular biology—An algorithmic approach.* The MIT Press, Cambridge, Mass.

POSFAI, J., BHAGWAT, A., POSFAI, G., AND ROBERTS, R. 1989. Predictive motifs derived from cytosine methyltransferases. *Nucl. Acids Res. 17*, 2421–2435.

PROUTSKI, V., AND HOLME, E. C. 1996. Primer master: A new program for the design and analysis of PCR primers. *CABIOS 12*, 253–255.

RAGHAVAN, P. 1988. Probabilistic construction of deterministic algorithms: Approximate packing integer programs. *J. Comput. Syst. Sci. 37*, 2, 130–143.

SCHULER, G. D., ALTSCHUL, S. F., AND LIPMAN, D. J. 1991. A workbench for multiple alignment construction and analysis. *Proteins: Struct. Funct. Genet. 9*, 180–190.

STORMO, G. 1990. Consensus patterns in DNA. In *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, R. F. Doolittle, Ed. *Methods in Enzymology*, vol. 183, pp. 211–221.

STORMO, G., AND HARTZELL III, G. W. 1991. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA 88*, 5699–5703.

WATERMAN, M. 1986. Multiple sequence alignment by consensus. *Nucl. Acids Res. 14*, 9095–9102.

WATERMAN, M., ARRATIA, R., AND GALAS, D. 1984. Pattern recognition in several sequences: Consensus and alignment. *Bull. Math. Biol. 46*, 515–527.

WATERMAN, M., AND PERLWITZ, M. 1984. Line geometries for sequence comparisons. *Bull. Math. Biol. 46*, 567–577.

RECEIVED FEBRUARY 2000; REVISED JANUARY 2002; ACCEPTED JANUARY 2002