A $2\frac{2}{3}$–Approximation Algorithm for the Shortest Superstring Problem

Chris Armen and Clifford Stein
Dartmouth College Technical Report PCS-TR95-262

November 20, 1995

Abstract

Given a collection of strings $S = \{s_1, \ldots, s_n\}$ over an alphabet Σ, a *superstring* α of S is a string containing each s_i as a substring; that is, for each i, $1 \le i \le n$, α contains a block of $|s_i|$ consecutive characters that match s_i exactly. The *shortest superstring problem* is the problem of finding a superstring α of minimum length.

The shortest superstring problem has applications in both data compression and computational biology. In data compression, the problem is a part of a general model of string compression proposed by Gallant, Maier and Storer (JCSS '80). Much of the recent interest in the problem is due to its application to DNA sequence assembly.

The problem has been shown to be NP-hard; in fact, it was shown by Blum et al.(JACM '94) to be MAX SNP-hard. The first $O(1)$-approximation was also due to Blum et al., who gave an algorithm that always returns a superstring no more than 3 times the length of an optimal solution. Several researchers have published results that improve on the approximation ratio; of these, the best previous result is our algorithm ShortString, which achieves a $2\frac{3}{4}$–approximation (WADS '95).

We present our new algorithm, G-ShortString, which achieves a ratio of $2\frac{2}{3}$. It generalizes the ShortString algorithm, but the analysis differs substantially from that of ShortString. Our previous work identified classes of strings that have a nested periodic structure, and which must be present in the worst case for our algorithms. We introduced machinery to descibe these strings and proved strong structural properties about them. In this paper we extend this study to strings that exhibit a more relaxed form of the same structure, and we use this understanding to obtain our improved result.

1 Introduction

The shortest superstring problem has applications in both computational biology [7, 16, 18] and data compression [10, 20]. We begin by describing the former. DNA sequencing is the task of determining the sequence of nucleotides in a molecule of DNA. These nucleotides are one of adenine, cytosine, guanine, and thymine, and are typically represented by the alphabet $\{a, c, g, t\}$. A molecule of human DNA is about 10^8 nucleotides long. Current laboratory procedures can directly determine the nucleotides of a fragment of DNA up to about 600 nucleotides long. In *shotgun sequencing*, several copies of a DNA molecule are fragmented using various restriction enzymes.

Once the nucleotides of all of the fragments have been determined, the *sequence assembly problem* is the computational task of reconstructing the original molecule from the overlapping fragments. The shortest superstring problem is an abstraction of this problem, in which the shortest reconstruction is assumed to be the most likely on the grounds that it is the most parsimonious. We state the problem as follows.

Given a collection of strings $S = \{s_1, \ldots, s_n\}$ over an alphabet Σ, a *superstring* α of S is a string containing each s_i as a substring, that is, for each i, $1 \le i \le n$, α contains $|s_i|$ consecutive characters that match s_i exactly. The *shortest superstring problem* is the problem of finding a superstring α of minimum length.

The shortest superstring problem is MAX SNP-hard [4]; several heuristics and approximation algorithms have been proposed. One often used algorithm is a greedy algorithm that repeatedly merges the pair of strings with the maximum amount of overlap. Turner [23] and Tarhio and Ukkonen [21] independently proved that the greedy algorithm constructs a superstring that achieves at least half as much overlap as an optimal superstring. However, this does not guarantee a constant approximation with respect to the length of the resulting superstring.

The first bound on the length approximation of the greedy algorithm was provided by Blum et al.[4], who showed that the greedy algorithm returns a string that is no longer than four times optimal; they also give a modified greedy algorithm that returns a string that is within three times optimal. Teng and Yao [22] gave a nongreedy algorithm that finds a string that is within $2\frac{8}{9}$ of optimal. Subsequently, three results appeared that achieved better approximation ratios using very different techniques. Czumaj et al.[6] refined the algorithm of [22] to achieve a $2\frac{5}{6}$ approximation. Kosaraju et al. obtained an improved result for the maximum traveling salesman problem; this more general result can be used by the algorithm of [4] to obtain an approximation slightly better than 2.8 [15]. Our result of $2\frac{3}{4}$ [2, 1] was the best known until recently, and in fact can be combined with the algorithm of [15] to obtain an approximation ratio of about 2.725.

In this report we describe our $2\frac{2}{3}$-approximation algorithm for the shortest superstring problem, which also appears in [3]. Algorithmically, the approach is a generalization of the one taken in [2], but the analysis is very different.

We now give a brief overview of our approach. All of the above mentioned algorithms begin by finding a minimum-weight cycle cover on a graph which has a node for every string and an edge between string u and v of length $|u| - \mathrm{ov}(u, v)$, where $\mathrm{ov}(u, v)$ is the amount of overlap that can be obtained by merging u and v. This cycle cover partitions the strings into cycles; the remaining work is in patching the cycles together to form one cycle covering the whole graph. The key to our new algorithm is to exploit the periodic structure of the cycles of strings that arise in this problem. In particular, the 3-approximation of [4] uses a theorem about infinite periodic functions [8], and the correspondence between periodic functions and strings in cycles. However, the particular instances of cycle patching that appear to be difficult actually involve short periodic strings, that is, strings that are periodic, but whose period may repeat only slightly more than once. We prove several new properties about such strings, allowing us to answer questions of the following form: given a string with some periodic structure, characterize *all* the possible periodic strings that can have a large amount of overlap with the first string. Given this understanding, we will be able to predict the ways in which overlap between certain strings can occur, and thus plan for it algorithmically.

2 Preliminaries

For consistency, we use some notation and definitions of [4] and [22]. We assume, without loss of generality, that the set S of strings is *substring free*, i.e. no s_j is a substring of s_i, $i \ne j$. We use $|s_i|$ to denote the length of string s_i, $|S|$ to denote the sum of the lengths of all the strings, and $\mathrm{opt}(S)$ to denote the length of the shortest superstring of S.

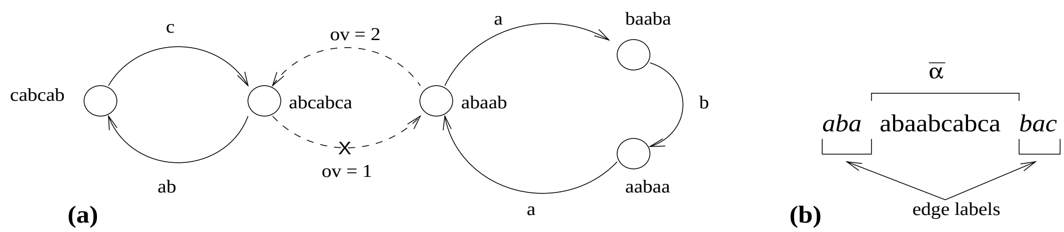

Figure 1: Execution of Generic SuperString Algorithm. Nodes are labeled with strings, edges with prefixes. (a) The graph after Step (4). Solid edges are in C, dashed edges in CC. The edge with an X is the one discarded in Step (4). (b) The final string consisting of $\bar{\alpha}$ (the merge of *abaab* and *abcabca*) along with the labels from the edges of the cycles.

Given two strings s and t, we define $\mathrm{ov}(s,t)$, the *overlap* between s and t, to be the length of the longest string x,such that there exist non-empty u and v with $s = ux$ and $t = xv$. We call u the *prefix* of s with respect to t, $\mathrm{pref}(s,t)$, and refer to $|u|$ as the distance from s to t, $\mathrm{d}(s,t)$. Observe that for any s and t, $\mathrm{ov}(s,t) + \mathrm{d}(s,t) = |s|$. String uxv, the shortest superstring of s and t in which s appears before t is denoted by $\langle s,t\rangle$, and $|\langle s,t\rangle| = |s| + |t| - \mathrm{ov}(s,t)$.

We can map the superstring problem to a graph problem by defining the *distance graph*. We create a graph $G = (V,E)$ with a vertex $v_i \in V$ for each string $s_i \in S$. For every ordered pair of vertices v_i, v_j, we place a directed edge of length $\mathrm{d}(s_i,s_j)$ and label the edge with $\mathrm{pref}(s_i,s_j)$. We can now observe that a minimum length hamiltonian cycle (traveling salesman tour) $v_{\pi_1},\ldots,v_{\pi_n},v_{\pi_1}$, in G, with edge i,j labeled by $\mathrm{pref}(s_{\pi_i},s_{\pi_j})$, almost corresponds to a superstring in S, the only difference being that we must replace $\mathrm{pref}(s_{\pi_n},s_{\pi_1})$ with s_{π_n}. Since $\mathrm{pref}(s_i,s_j) \leq |s|$, we can conclude that $\mathrm{opt}(TSP) \leq \mathrm{opt}(S)$, where $\mathrm{opt}(TSP)$ is the optimal solution to TSP defined above. This TSP is directed (sometimes called *asymmetric*); thus the best known approximation [9] is only within a factor of $O(\log n)$. Therefore, we must exploit more of the structure of the problem in order to achieve better bounds.

Given a directed graph G, with weights on the edges, a *cycle cover* C is a set of cycles such that each vertex is in exactly one cycle. A minimum-cost cycle cover is a cycle cover such that the sum of the weights of the edges in all the cycles is minimized. A minimum-cost cycle cover can be computed in $O(n^3)$ time by a well-known reduction to the assignment problem [17]. Since a tour is a cycle cover, $\mathrm{opt}(C) \leq \mathrm{opt}(TSP)$. When we say that a string s_i is in some cycle c of cycle cover C, we mean that the vertex v_i with which s_i is associated is in cycle c. Throughout the paper, when we refer to a cycle, we will be referring to a cycle that is in a minimum-cost cycle cover in the distance graph.

Because $\mathrm{ov}(s_i,s_j) + \mathrm{d}(s_i,s_j) = |s_i|$, one could also weight the edges by their overlap, find a maximum-cost cycle cover and obtain the same solution. A superstring which has minimum length, or distance, also has maximum overlap. However, this correspondence breaks down for approximations; approximating the largest overlap appears to be an easier problem (cf. [23, 22, 15]) than approximating the shortest superstring.

We now describe a generic superstring algorithm used, in some form, by [4],[22] and [6]. An execution of the algorithm appears as Fig. 1.

<u>Generic SuperString Algorithm</u>

1) Find a minimum cost cycle cover C in the distance graph G.

2) For each cycle $c \in C$, choose one string to be a representative r_c.
Let G' be the subgraph induced by the representative set R.
3) Compute a cycle cover CC on G'.
4) Break each cycle $\gamma \in CC$ by deleting one edge.
5) Concatenate the remaining strings arbitrarily.
6) Extend each representative r_c by the concatenation of the prefixes around c.

The first cycle cover identifies sets of strings that have large amounts of overlap. This allows us to form the second cycle cover, in which approximating overlap and the string length are roughly comparable, so stronger bounds apply. Step (6) correctly extends the superstring for R into a superstring for S, as proved in [22].

We now analyze the generic algorithm in a way that anticipates our improvements. A more detailed analysis appears in [4]. Let $\mathrm{d}(C')$ be the sum of the distances and $\mathrm{ov}(C')$ be the sum of the overlaps of the edges in a cycle cover C'. Consider the second cycle cover CC. Let $\mathrm{opt}(R)$ be the optimal superstring on the strings in $r_c \in R$ and observe that $\mathrm{opt}(R) \leq \mathrm{opt}(S)$. Let $\bar{\alpha}$ be the string produced in Step 5, a superstring of R, and let $\mathrm{opt}(\mathrm{ov}(R)) = |R| - \mathrm{opt}(R)$ be the amount of overlap in the optimal superstring for R. Since the shortest superstring for R is a cycle cover for G', $\mathrm{ov}(CC) \geq \mathrm{opt}(\mathrm{ov}(R))$. However, the superstring $\bar{\alpha}$ does not have as much overlap as CC, since we delete one edge from each cycle.

For a cycle γ, let ov^n_γ denote the overlap in the edge deleted in Step 4, and let ov^t_γ denote the remaining overlap in γ. Let $\mathrm{ov}^t = \sum_{\gamma \in CC} \mathrm{ov}^t_\gamma$ and $\mathrm{ov}^n = \sum_{\gamma \in CC} \mathrm{ov}^n_\gamma$. Thus, $|\bar{\alpha}| \leq |R| - \mathrm{ov}^t$. By definition, $|R| \leq \mathrm{opt}(R) + \mathrm{opt}(\mathrm{ov}(R)) \leq \mathrm{opt}(R) + \mathrm{ov}(CC)$. Combining these two inequalities with $\mathrm{ov}(CC) = \mathrm{ov}^n + \mathrm{ov}^t$, gives that $\bar{\alpha} \leq \mathrm{opt}(R) + \mathrm{ov}^n$. We then must extend each cycle, in Step 6. Let $\mathrm{Ext}(\gamma)$ be the cost of extending all cycles $c \in C$ s.t. $r_c \in \gamma$. Then we can express the length of α, the string obtained, as

$$|\alpha| \leq \mathrm{opt}(R) + \sum_{\gamma \in CC} \Big(\mathrm{ov}^n_\gamma + \mathrm{Ext}(\gamma) \Big) \quad . \tag{1}$$

Let $\mathrm{d}(c)$ be the sum of the weights of the edges of a cycle $c \in C$; so $\mathrm{d}(C) = \sum_{c \in C} \mathrm{d}(c)$. To obtain a 3-approximation, observe that the set of edges which contribute to ov^n form a matching M on G'. Now we employ a key lemma from [4]:

Lemma 2.1 ([4]) *Let c, c' be cycles in a minimum cycle cover C with strings $s \in c$ and $s' \in c'$. Then the overlap between s, s' is less than $d(c) + d(c')$.*

Since M is a matching, each cycle c is at an endpoint of a string at most once, and hence $\mathrm{ov}^n \leq \mathrm{d}(C)$. Now, we extend $\bar{\alpha}$ by the edge labels on each cycle, adding a total of $\mathrm{d}(C)$ to the length of the string. Let α be the resulting string. We conclude that

$$|\alpha| \leq \mathrm{opt}(R) + \sum_{\gamma \in CC} \mathrm{ov}^n_\gamma + \mathrm{Ext}(\gamma) \leq \mathrm{opt}(R) + \mathrm{d}(C) + \mathrm{d}(C) \leq 3\mathrm{opt}(S) \ , \tag{2}$$

since both $\mathrm{d}(C)$ and $\mathrm{opt}(R)$ are lower bounds on $\mathrm{opt}(S)$.

The analysis above makes it clear that the cycle cover CC actually partitions the cycles in the cycle cover C, and hence each cycle in CC can be analyzed separately. As was observed by [22] in their $2\frac{8}{9}$ algorithm, if γ has three or more vertices, then $\mathrm{ov}^n_\gamma \leq \frac{2}{3} \sum_{c \in \gamma} \mathrm{d}(c)$.

Thus we can restrict our attention to 2-cycles in CC. We will analyze each 2-cycle in CC separately, and obtain a $2\frac{2}{3}$ bound by proving structural properties of these cycles.

Given a representative $v = r_c$ for some cycle c, we use c_v to denote the cycle c of which v is a representative. We summarize this discussion with the following lemma:

Lemma 2.2 *An algorithm following the framework of the generic algorithm above, that, for each 2-cycle γ in CC consisting of vertices v and t, attains a bound of $ov^n_\gamma + Ext(\gamma) \leq \beta(d(c_v) + d(c_t))$, for some $\beta \geq \frac{5}{3}$, is a $(1+\beta)$-approximation algorithm for the shortest superstring problem.*

We define a few terms describing the structure of cycles. The reader is referred to [4] for a more complete discussion. We call a string s *irreducible* if all cyclic shifts of s yield unique strings, and *reducible* otherwise. We say that s has *periodicity* x if there exists a string t with $|t| = x$ such that s is substring of t^∞. Let $\text{per}(c)$ be the string formed by concatenating all the labels on the edges of a cycle c. Then for each string $s \in c$, s is a substring of $\text{per}(c)^\infty$. Note that $\text{per}(c)$ must be irreducible; otherwise a cycle with less total distance could generate the same strings, contradicting the minimality of the cycle cover. The irreducibility of the periods of cycles in a minimum cycle cover will figure prominently in many of our proofs.

We can now state a corollary to Lemma 2.1 that we will also use frequently in our proofs.

Corollary 2.3 **([4])** *Let w be a substring of both $(\sigma_j)^\infty$ and $(\sigma_k)^\infty$. Then if $|w| \geq |\sigma_j| + |\sigma_k|$, either σ_j or σ_k is reducible.*

3 Repeaters and their Characteristics

In the previous section, we saw that in order to obtain a better approximation for the shortest superstring problem it is sufficient to consider 2-cycles in the second cycle cover of the generic superstring algorithm. In this section we describe the machinery for describing 2-cycles developed in [2].

Suppose we choose v and t as representatives of two cycles of the first cycle cover C, and they form a 2-cycle in CC in which one of $\text{ov}(v,t)$ or $\text{ov}(t,v)$ is large but the other is small. In Step 4 we will break the 2-cycle to form a string, and since we are trying to maximize overlap, the obvious choice is to keep the high-overlap edge and discard the other. But if both edges have high overlap, we must discard one of them. In a 2-cycle this will cost us up to half of the overlap, which is the "worst case" of the generic algorithm. We observe that both edges in such a 2-cycle cannot participate in an optimal solution; in this sense the second cycle cover has achieved "false overlap". We formalize the idea of a "high-overlap 2-cycle" as follows:

Definition 3.1 Let γ be a 2-cycle in the second cycle cover CC of the GENERIC algorithm, consisting of vertices r_j and r_k, the representatives of cycles c_j and c_k in C. Without loss of generality assume that $\text{d}(c_j) \geq \text{d}(c_k)$. Then γ is a (g,h)-*HO2-cycle* if $\min\{\text{ov}(r_j,r_k), \text{ov}(r_k,r_j)\} \geq g\text{d}(c_j) + h\text{d}(c_k)$.

Our strategy is to anticipate, when we choose representatives, the potential of each string to participate in a $(\frac{2}{3},\frac{2}{3})$-HO2-cycle. In particular we evaluate the potential of each string to play the role of the larger-period string in the 2-cycle. Such a string must have a very specific structure; if we find a string without such a structure, we use it as representative. Otherwise we know a great deal about the structure of the entire

$z =$	$\underline{\text{abababrst}}\text{ababab}$	$z =$	$\underline{\text{ababadababadab}}\text{ababadababadabab}$
	) (		( ) ( ) (
$y = y_\ell =$	ababab	$y = y_\ell =$	ababadababadababa
$y_r =$	ababab	$y_r =$	ababadababadabab
$\sigma =$	ab	$\sigma =$	ababad
	(a)		(b)

Figure 2: Positive and Negative Characteristics. Per(c) is underlined. (a) shows a negative characteristic. (b) shows a positive characteristic$y$ and σ are also shown.

cycle and can trade off the amount of two-way overlap against the cost of extending the representative to include the rest of the cycle.

In order to have the potential to be the larger-period string in a high-overlap 2-cycle, a string z must have a significant prefix that has some smaller period. This smaller period might correspond to the period of another cycle in the cover, and hence some other representative w such that $\text{ov}(w, z)$ would be large. The suffix of z must simlarly have the same smaller period, so that $\text{ov}(z, w)$ would be large. We require some notation to describe this potential.

Definition 3.2 Let z be a string in cycle c and let σ be an irreducible string with $|\sigma| < d(c)$. Then σ is a *(g, h)-repeater of z* if there exist witnesses y_ℓ and y_r, such that

1. y_ℓ is a prefix of z and y_r is a suffix of z.
2. y_ℓ and y_r are substrings of $(\sigma)^\infty$.
3. $|y_\ell|, |y_r| > g\text{d}(c) + h|\sigma|$.

We will always choose y_ℓ and y_r to be the maximum length prefix and suffix that satisfy conditions 1–3 above.

Consider the string z in Fig. 2b and let $g = h = \frac{2}{3}$. Here $\text{per}(c) = ababadababadab, \sigma = ababad$, $y_\ell = ababadababadababa$ and $y_r = ababadababadabab$. So $|y_\ell|, |y_r| > \frac{2}{3}d(c) + \frac{2}{3}|\sigma|$, and we say that σ is a $(\frac{2}{3}, \frac{2}{3})$-repeater of z.

Note that in our example y_ℓ and y_r are almost the same; this is not a complete coincidence. All the repeaters we will be considering in this paper will have $g \geq \frac{1}{2}$ and hence y_ℓ and y_r *must* overlap, often significantly (as in this example). For convenience we will define one witness y_σ which contains both y_ℓ and y_r; that is, we define y_σ to be the maximum-length substring of $(\sigma)^\infty$ that is also a substring of $\text{per}(c)^\infty$. In other words, if you took σ and tried to repeat it as many times as possible, in both directions, while being consistent with c, you get y_σ. In the example above $y_\sigma = y_\ell$. When the context is clear, we will drop the σ and just refer to witness y.

Henceforth when discussing and proving properties of cycles, we will refer to the maximal witness y_σ rather than to the underlying pair of witnesses y_ℓ and y_r. This simplification is conservative.

The idea behind (g, h)-repeaters is to identify periodic substrings of the period of a cycle in C. We will also be interested in identifying that portion of a cycle that is *not* consistent with some (g, h)-repeater σ. Note that a copy of y_σ begins every $\text{d}(c)$ in $\text{per}(c)^\infty$, and that $|y| < 2\text{d}(c)$, since by Corollary 2.3, $|y| < \text{d}(c) + |\sigma| \leq 2\text{d}(c)$.

Definition 3.3 Let c be a cycle with (g,h)-repeater σ and maximal witness y. Fix a copy of y in $\mathrm{per}(c)^\infty$. The point just to the left of the first character of y is the *head of* y. Index this point as 0 and continue the indices between each character leftward and rightward to cover the interval $[-\mathrm{d}(c)..\mathrm{d}(c)]$. Now mark the point $|y| - \mathrm{d}(c)$ and call it the *tail* of σ. The *characteristic* of σ, X_σ, is the interval from the head to the tail. If $|y| - \mathrm{d}(c) > 0$ we call $[0..|y| - \mathrm{d}(c)]$ a *positive characteristic* X_σ. If $|y| - \mathrm{d}(c) \leq 0$ we call $[|y| - \mathrm{d}(c)..0]$ a *negative characteristic* X_σ.

We can picture the characteristics of the repeaters of a cycle c in terms of parentheses. Fig. 2b illustrates this idea for positive characteristics. The left and right ends of y_σ are marked with left and right parentheses; these correspond to the head and tail of adjacent copies of X_σ.

A negative characteristic appears in Fig. 2a and can be pictured as a single solid entity (perhaps of size zero) which spans the gap between copies of y. In this example rst is the negative characteristic. Each characteristic appears once every $d(c)$. Intuitively, the characteristic of a repeater borders the portion of $\mathrm{per}(c)$ which must be included as a prefix and suffix of some string z if z is to participate in a high-overlap 2-cycle. Recall that we defined (g,h)-repeaters (Def. 3.2) in terms of some string z in a cycle c which contained witnesses y_ℓ and y_r as a prefix and suffix. In general there might be several such strings in c which could satisfy the definition. We say that σ is *active* in each of these strings. We say that two characteristics X_{σ_i}, X_{σ_j} are *nested* if X_{σ_i} is a positive characteristic and X_{σ_j} falls within X_{σ_i}. We say that two characteristics X_{σ_i}, X_{σ_j} are *disjoint* if their intervals are disjoint. Otherwise we say that X_σ and $X_{\sigma'}$ are *linked.*

We will frequently be interested in the relationship between two substrings of $\mathrm{per}(c)^\infty$, for instance between two witness strings y and y'. As noted above, a copy of any substring of $\mathrm{per}(c)^\infty$ occurs every $\mathrm{d}(c)$ in $\mathrm{per}(c)^\infty$. We overload our notation for $\mathrm{d}(,)$ and $\mathrm{ov}(,)$ in the obvious way to refer to prefix distance $\mathrm{d}(y,y')$ and overlap $\mathrm{ov}(y,y')$. We also define the *suffix distance* $\tilde{\mathrm{d}}(y,y')$ to be the distance from the last character of a copy of y to the last character of the first copy of y' that ends after y.

We will require the following bound on the length of a witness string:

Lemma 3.4 ([2]) *Let y_σ be a maximal witness for some (g,h)-repeater σ in a cycle c. Then $|y_\sigma| < d(c) + |\sigma| < 2d(c)$.*

4 The Algorithm

We present our algorithm G-ShortString, which is a $2\frac{2}{3}$-approximation algorithm for the shortest superstring problem. We describe the algorithm in Section 4.1. In order to prove our bound on its approximation ratio in Section 4.3, we present some technical lemmas on the structure of cycles with $(\frac{2}{3},\frac{2}{3})$-repeaters in Section 4.2.

4.1 Algorithm G-ShortString

In order to achieve a bound of $2\frac{2}{3}$ within the framework of Generic, Lemma 2.2 states that we need to concentrate on $(\frac{2}{3},\frac{2}{3})$-HO2-cycles. As in [2], our strategy is to anticipate, when we select a representative r_j, the possible involvement of r_j as the larger-period string in a (g,h)-HO2-cycle. In [2] we used criteria for doing so which were based on our detailed knowledge of the structure of $(\frac{3}{4},\frac{3}{4})$-repeaters. G-ShortString does not depend on such knowledge.

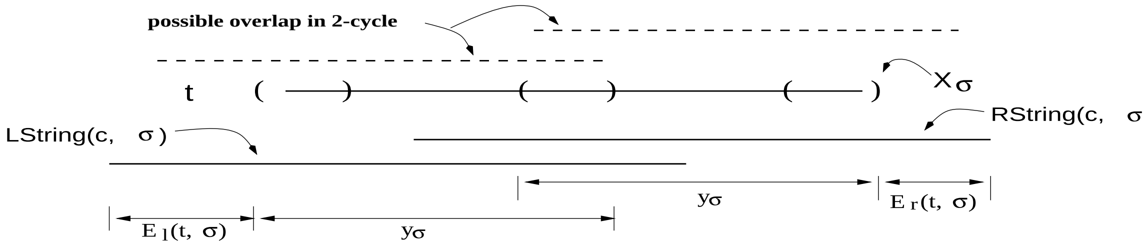

Figure 3: Definitions 4.2 and 4.3.

Our new procedure for selecting representatives is to evaluate a cost function for each string in a cycle, and to select the string with the best *worst-case* cost. We identify a cost function which resembles the desired bounds, and we explicitly attempt to minimize this function in the algorithm. We achieve our improved bound by more careful extension of each representative r_j of a cycle c_j that is also the larger-period string in a $(\frac{2}{3}, \frac{2}{3})$-HO2-cycle. We therefore need some new ideas about extension and some notation for expressing it.

Definition 4.1 Let σ be a (g, h)-repeater with maximal witness y_σ in an m-cycle c. Index the strings s_i such that $\mathrm{d}(y_\sigma, s_i) < \mathrm{d}(y_\sigma, s_{i+1})$, $1 \le i < m$. Then we define the *right string of σ in c*, $\mathrm{RString}(c, \sigma) = s_m$. The *left string of σ in c*, $\mathrm{LString}(c, \sigma)$ is defined symmetrically; reindex the strings s_i such that $\tilde{\mathrm{d}}(s_i, y_\sigma) > \tilde{\mathrm{d}}(s_{i+1}, y_\sigma)$, $1 \le i < m$. Then we define $\mathrm{LString}(c, \sigma) = s_1$.

In other words, if we align a copy of each of the strings in c in such a way that the first one begins as soon after a copy of y_σ as possible, then the rightmost string is $\mathrm{RString}(c, \sigma)$. The idea is that if we choose as representative a string t in which σ is active, and t becomes the larger-period string in a (g, h)-HO2-cycle, then $\mathrm{RString}(c, \sigma)$ is the rightmost string which we will have to include if we extend to the right. Figure 3 illustrates Definitions 4.1 and 4.2.

Definition 4.2 Let σ be a (g, h)-repeater which is active in a string t in cycle c. Then the *right σ-extension with respect to t*, $\mathrm{E}_r(t, \sigma) = \tilde{\mathrm{d}}(y_\sigma, \mathrm{RString}(c, \sigma))$. The *left σ-extension with respect to t*, $\mathrm{E}_\ell(t, \sigma) = \mathrm{d}(\mathrm{LString}(c, \sigma), y_\sigma)$.

Given a $(\frac{2}{3}, \frac{2}{3})$-repeater σ which is active in a string t in cycle c_t, we wish to calculate the cost of choosing t and having t involved in a $(\frac{2}{3}, \frac{2}{3})$-HO2-cycle γ with some string v such that $\mathrm{per}(c_v) = \sigma$. In particular, by Lemma 2.2 we are interested in anticipating $\mathrm{ov}_\gamma^n + \mathrm{Ext}(c_t) + \mathrm{Ext}(c_v)$. Consider without loss of generality the right end of t, and let y_r be the suffix of t which is the witness string for σ. Then we know that $\mathrm{ov}_\gamma^n \le |y_r| \le |y_\sigma|$. If there is slack in either of these inequalities, then we use the slack as part of our upper bound on extension cost. We have to extend to the right well beyond the end of y_σ in any case, so it does not matter whether we charge $|y_\sigma| - |y_\ell|$ to ov_γ^n or to $\mathrm{Ext}(c_t)$. From the end of y_σ, we need to extend to the right only as far as necessary to include $\mathrm{RString}(c, \sigma)$. We also have to extend v to include the remaining strings in c_v; we assume the cost of full extension. This motivates the following definition.

Definition 4.3 Let σ be a (g, h)-repeater that is active in string t in cycle c. Then the *anticipated cost* of choosing t as representative and forming a 2-cycle with a string with period σ is

$$\text{Cost}(t,\sigma) = |y_\sigma| + \min\{\text{E}_\ell(t,\sigma), \text{E}_r(t,\sigma)\} + |\sigma|.$$

What we seek, then, is to *minimize*, in our choice of representative t, the *maximum* over all $(\frac{2}{3},\frac{2}{3})$-repeaters active in t, the anticipated cost $\text{Cost}(t,\sigma)$. Allowing $\sigma \in t$ to mean "σ active in t", we seek .

$$\text{BestRep}(c) = \text{argmin}_{t\in c}\left\{\max_{\sigma\in t}\{\text{Cost}(t,\sigma)\}\right\}$$

Procedure G-FindReps(c), shown below, calculates the anticipated cost for each pair (t,σ) such that t is a string in c and σ is active in t.

Procedure G-FindReps(c_j)
1) Find all $(\frac{2}{3},\frac{2}{3})$-repeaters and associated characteristics in c_j.
2) **If** any string t has no $(\frac{2}{3},\frac{2}{3})$-repeaters
    Then $r_j = t$;
3) **Else**
    $r_j = \text{BestRep}(c_j)$;
4) **Return** r_j.

The main body of G-ShortString is exactly ShortString, except that representatives are selected in Step 2) by a call to procedure G-FindReps(c), and the parameters of the (g,h)-HO2-cycle in Step 4) are different.

Algorithm G-ShortString
(1) Form minimum cycle cover C on distance graph G.
(2) For each cycle $c \in C$
    Call G-FindReps(c) to choose representative r_c.
    Add r_c to R.
  Let G' be the subgraph induced by R.
(3) Form minimum cycle cover CC on G'.
(4) For each cycle γ in CC:
    if γ is a $(\frac{2}{3},\frac{2}{3})$-HO2-cycle (v,t)
  (a)   **then if** $\text{ov}(t,v) + \text{E}_r(t,\text{per}(c_v)) \le \text{ov}(v,t) + \text{E}_\ell(t,\text{per}(c_v))$
        then Extend $\langle v,t\rangle$;
        else Extend $\langle t,v\rangle$;
  (b) **else** discard edge of cycle γ with least overlap; Extend each vertex w by $\text{d}(c_w)$
(5) Concatenate strings from (4) to form superstring α

In Step (4a) above, the instruction "Extend $\langle v,t\rangle$" is shorthand for the following idea. We extend v to the left to include all of the strings in c_v; we assume in the analysis in Section 4.3 that this length is $\text{d}(c_v)$. We extend t to the right, as far as is necessary to include $\text{RString}(c,\text{per}(c_v))$. Extending $\langle t,v\rangle$ is done symmetrically.

The algorithm G-ShortString correctly computes a superstring of the set of strings S. This follows from the correctness of Generic. Our method of choosing representatives for each cycle is a special case of the method of Generic, which chooses an arbitrary string as representative. In step (4b), we do exactly what Generic does. In step (4a), we use a different criterion for breaking a cycle $\gamma \in CC$, and we only extend each representative far enough to "cover" all of the strings in its cycle. Each string is therefore included in the solution α.

G-ShortString runs in polynomial time. The distance graph G can be built in $O(|S|+n^2)$ time [11], and the cycle cover computations take $O(n^3)$ time [17]. These two results determine the running time of Generic. In addition, our algorithm must find all of the $(\frac{2}{3},\frac{2}{3})$-repeaters in each cycle $c \in C$ in G-FindReps(c). This can be done naively in polynomial time by examining a prefix and suffix of each string, and determining whether the prefix and suffix have periodicity $2 \le j < \mathrm{d}(c)$.

In order to analyze the approximation ratio achieved by G-ShortString, we require a few technical lemmas pertaining to $(\frac{2}{3},\frac{2}{3})$-repeaters.

4.2 Properties of Strings with $(\frac{2}{3},\frac{2}{3})$-Repeaters

A *small* (g,h)-repeater in a cycle c is one whose minimum witness length is less than $\mathrm{d}(c)$. A $(\frac{2}{3},\frac{2}{3})$-repeater σ is small if $|\sigma| < \frac{1}{2}\mathrm{d}(c)$. Generally we are interested in avoiding small (g,h)-repeaters. To see why this is so, suppose that we choose a representative r_j for cycle c_j, and r_j is involved in a $(\frac{2}{3},\frac{2}{3})$-HO2-cycle with another representative r_k of cycle c_k, and $\mathrm{per}(c_k)=\sigma$. Then we will want to bound the extension cost we incur, $\mathrm{Ext}(c_j)$, in terms of $\mathrm{d}(c_j)+\mathrm{d}(c_k) = \mathrm{d}(c_j)+|\sigma|$. So if σ is larger, then our extension cost, as a fraction of $\mathrm{d}(c_j)+\mathrm{d}(c_k)$, is smaller.

There may be several small $(\frac{2}{3},\frac{2}{3})$-repeaters in a cycle, but we are able to bound the number of small $(\frac{2}{3},\frac{2}{3})$-repeaters in a string.

Lemma 4.4 *Let s be a string in a cycle c. Then at most one small $(\frac{2}{3},\frac{2}{3})$-repeater can be active in s.*

Proof: Suppose for purpose of contradiction that there exist two such $(\frac{2}{3},\frac{2}{3})$-repeaters σ and σ'. Let $y_\ell(\sigma)$ and $y_\ell(\sigma')$ be the prefixes of s which are the left witness strings of σ and σ' respectively. Let $y_\ell = \mathrm{argmin}\{|y_\ell(\sigma)|, |y_\ell(\sigma')|\}$ be the prefix of s which is periodic in both σ and σ'. Applying Corollary 2.3, Definition 3.2, and the fact that $|\sigma'| < \frac{1}{2}\mathrm{d}(c)$, we get

$$\begin{aligned} |\sigma| &> |y_\ell| - |\sigma'| \\ &> \frac{2}{3}\mathrm{d}(c) - \frac{1}{3}|\sigma'| \\ &> \frac{1}{2}\mathrm{d}(c), \end{aligned}$$

a contradiction since σ is a small $(\frac{2}{3},\frac{2}{3})$-repeater. ■

The following lemma gives us a lower bound on the size of a $(\frac{2}{3},\frac{2}{3})$-repeater whose characteristic has the characteristic of another $(\frac{2}{3},\frac{2}{3})$-repeater nested within it.

Lemma 4.5 *Let X_σ be a positive characteristic in cycle c and $X_{\sigma'}$ a characteristic nested within X_σ with $|\sigma| > |\sigma'|$. Then $|\sigma| > \frac{1}{2}d(c)$.*

Proof: In this case the witness y' is completely contained within the witness y. We apply Corollary 2.3 and the definition of $(\frac{2}{3},\frac{2}{3})$-repeater to get

$$\begin{aligned} |\sigma| + |\sigma'| &> \mathrm{ov}(y,y') \\ &> |y'| \\ &> \frac{2}{3}(\mathrm{d}(c)+|\sigma'|) \\ \Rightarrow |\sigma| &> \frac{2}{3}\mathrm{d}(c) - \frac{1}{3}|\sigma'|, \end{aligned}$$

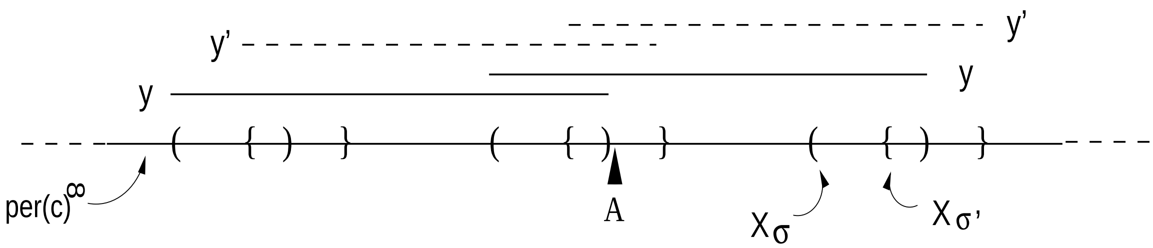

Figure 4: The characteristics X_σ and $X_{\sigma'}$ are linked, as in Lemma 4.7.

which implies $|\sigma| > \frac{1}{2}\mathrm{d}(c)$ because $|\sigma| > |\sigma'|$. ■

Because $(\frac{2}{3}, \frac{2}{3})$-repeaters may not be well parenthesized, we will often be faced in our analysis with situations in which two positive characteristics are linked, as pictured in Figure 4. (Recall that two positive characteristics are linked if they overlap, but neither contains the other.) The following lemma and its corollary gives us strong bounds on the size of the two $(\frac{2}{3}, \frac{2}{3})$-repeaters and on their difference. In order to prove the lemma, we require a proof technique introduced in [2], the *shift argument*. We describe this technique below.

We apply the shift argument to cycles that include two or more repeaters. We are generally interested in proving that some property holds; we assume that it does not, and use the shift argument to derive a contradiction. We begin with the following observation, which can easily be verified by the definition of maximal witness.

Observation 4.6 *Let y be the maximal witness for a (g,h)-repeater σ in a cycle c, and fix a copy y^* of y in $per(c)^\infty$. Index the character positions of $per(c)^\infty$ with the character to the left of y^* as 0, the first character of y^* as 1, and continuing to the right beyond the end of y^*. Let Char(i) be the character in position i. Then*

$$a)\, Char(0) \neq Char(|\sigma|)$$

and

$$b)\, Char(|y^*| - |\sigma| + 1) \neq Char(|y^*| + 1).$$

In each shift argument our goal will be to show that either inequality a) or b) in Observation 4.6 is violated and the terms are indeed equal. We will do so by making a series of *shifts* between characters, which we know to be identical, by the periodic structure of the strings. In particular, within any y_σ, any two characters that are σ apart are identical, and in $\mathrm{per}(c)^\infty$, any two characters that are $\mathrm{d}(c)$ apart are identical. We call such shifts *valid*. We will begin at either the character immediately preceding or following a copy of y or y', and perform a series of shifts which will bring us to the position whose character is supposed to be unequal. If these shifts are valid, then the two characters must be equal, contradicting our initial assumption that the characteristics X_σ and $X_{\sigma'}$ could overlap.

We introduce notation to describe the sequence of shifts. We give a starting position and a position at which we wish to arrive, relative to the starting position. We also give the series of moves and a set of requirements, that is, conditions on the various parameters that must be met in order for the moves to all be valid. Below the box, we show that the conditions for validity are indeed satisfied, which gives us a contradiction for the region in which the shifts are valid.

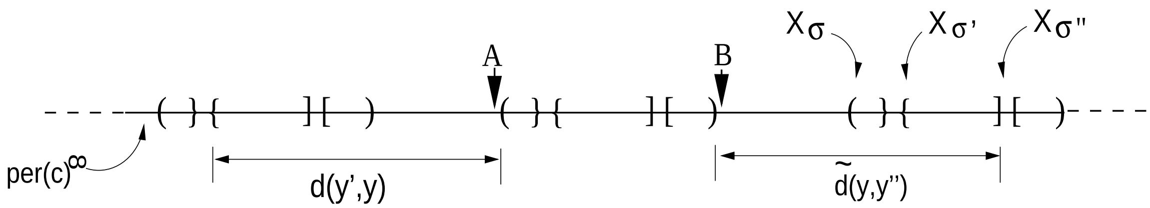

Figure 5: Proof of Lemma 4.9.

Lemma 4.7 *Let σ and σ' be two $(\frac{2}{3},\frac{2}{3})$-repeaters with positive characteristics in a cycle c, with $|\sigma| > |\sigma'|$, and X_σ and $X_{\sigma'}$ linked. Let $k = \left\lfloor \frac{|\sigma|}{|\sigma'|} \right\rfloor$. Then $|\sigma| - k|\sigma'| > |y_\sigma| - d(c)$.*

Proof: We apply the following shift argument, using start position (A) in Figure 4:

Start: (A)			**Goal:** $-\lvert\sigma\rvert$
No.	*Move*	*Requirement*	*Comments*
1.	$+k\lvert\sigma'\rvert$	$k\lvert\sigma'\rvert < \mathrm{d}(c)$	$k\lvert\sigma'\rvert < \lvert\sigma\rvert < \mathrm{d}(c)$
2.	$-\lvert\sigma\rvert$	$\lvert\sigma\rvert - k\lvert\sigma'\rvert \leq \lvert y_\sigma\rvert - \mathrm{d}(c)$	See below.
3	$-k\lvert\sigma'\rvert$	$\lvert\sigma\rvert < \mathrm{d}(c)$	Def. Repeater.

Because only move #2 is the only one whose validity is conditional, we conclude the negation of that condition, i.e. $|\sigma| - k|\sigma'| > |y_\sigma| - \mathrm{d}(c)$. ■

Corollary 4.8 *Let σ and σ' be two $(\frac{2}{3},\frac{2}{3})$-repeaters with positive characteristics in a cycle c, with $|\sigma| > |\sigma'|$, and X_σ and $X_{\sigma'}$ linked. Let $k = \left\lfloor \frac{|\sigma|}{|\sigma'|} \right\rfloor$. Then $|\sigma'| > |y_\sigma| - d(c)$.*

Proof: By the choice of k and Lemma 4.7,

$$|\sigma'| > |\sigma| - k|\sigma'| > |y_\sigma| - \mathrm{d}(c).$$

■

In our analysis, we will be interested in *lower bounds* on the size of potentially small $(\frac{2}{3},\frac{2}{3})$-repeaters in terms of some measure of distance which will correspond to extension cost. The following two lemmas provides such bounds for two important cases in which three characteristics are involved. Our choice of dimensions for identifying the relative positions of the three characteristics will seem unnatural now, but will simplify our task in Section 4.3.

Lemma 4.9 *Let σ,σ' and σ'' be $(\frac{2}{3},\frac{2}{3})$-repeaters in cycle c, with $X_{\sigma'}$ and $X_{\sigma''}$ disjoint, and with $X_{\sigma'}$ nested to the left of $X_{\sigma''}$ within X_σ, Then $|\sigma'| > d(y',y) + |y| - 2d(c)$ and $|\sigma''| > \tilde{d}(y,y'') + |y| - 2d(c)$.*

Proof: Figure 5 illustrates the start positions of our shift arguments.

Start: (A)			**Goal:** $+\lvert\sigma\rvert$
No.	*Move*	*Requirement*	*Comments*
1.	$+\lvert\sigma''\rvert$	$\lvert\sigma''\rvert \leq \tilde{\mathrm{d}}(y,y'') + \lvert y\rvert - 2\mathrm{d}(c)$	See below.
2.	$+\lvert\sigma\rvert$	$\lvert\sigma\rvert + \lvert\sigma''\rvert < \lvert y\rvert - \mathrm{d}(c) + \tilde{\mathrm{d}}(y,y'')$	See below.
3	$-\lvert\sigma''\rvert$	$\lvert\sigma\rvert + \lvert\sigma''\rvert > \lvert y\rvert - \mathrm{d}(c) + \lvert\sigma''\rvert$	See below.

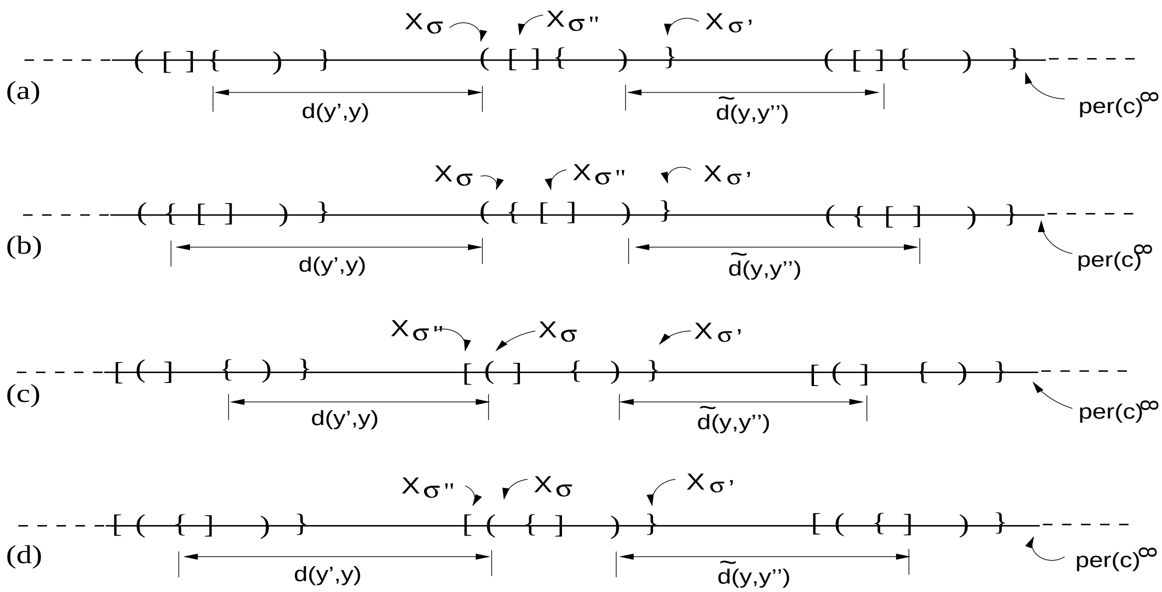

Figure 6: Lemma 4.10: $X_{\sigma'}$ is linked with X_σ. $X_{\sigma''}$ may be nested within $X_{\sigma'}$ (b), or only nested within X_σ(a). If $X_{\sigma''}$ is linked with X_σ, it may or may not also be linked with $X_{\sigma'}$ ((d) and (c) respectively).

Start: (B)			**Goal:** $-\lvert\sigma\rvert$
No.	*Move*	*Requirement*	*Comments*
1.	$-\lvert\sigma'\rvert$	$\lvert\sigma'\rvert \le \mathrm{d}(y', y) + \lvert y\rvert - 2\mathrm{d}(c)$	See below.
2.	$-\lvert\sigma\rvert$	$\lvert\sigma\rvert + \lvert\sigma'\rvert < \lvert y\rvert - \mathrm{d}(c) + \mathrm{d}(y', y)$	See below.
3	$+\lvert\sigma'\rvert$	$\lvert\sigma\rvert + \lvert\sigma'\rvert > \lvert y\rvert - \mathrm{d}(c) + \lvert\sigma'\rvert$	See below.

Requirement #3 for both of the above is always true because $|\sigma| > |y| - \mathrm{d}(c)$ by Lemma 3.4. Because $|\sigma| < \mathrm{d}(c)$, #1 implies #2; therefore #1 must be false:

$$\begin{aligned} |\sigma''| &> \tilde{\mathrm{d}}(y, y'') + |y| - 2\mathrm{d}(c) \\ |\sigma'| &> \mathrm{d}(y', y) + |y| - 2\mathrm{d}(c). \end{aligned}$$

■

Lemma 4.10 *Let σ, σ' and σ'' be $(\frac{2}{3}, \frac{2}{3})$-repeaters in cycle c, with maximal witnesses y, y' and y''. Let $|\sigma| > |\sigma'| > |\sigma''|$, and X_σ and $X_{\sigma'}$ positive. If $X_{\sigma'}$ is linked with X_σ, then $\min\{d(y', y), \tilde{d}(y, y'')\} < \frac{5}{3}d(c) + \frac{2}{3}|\sigma| - |y_\sigma|$.*

Proof: By the condition of the lemma we know that X_σ and $X_{\sigma'}$ are linked, but we do not know the relationship between between $X_{\sigma''}$ and the other two characteristics. The characteristic $X_{\sigma''}$ may be nested within one or both of X_σ and $X_{\sigma'}$ as in Figure 6(a) or (b), or it may be linked with one or both of X_σ and $X_{\sigma'}$ as in Figure 6 (c) or (d). In any of these cases we can apply Corollary 2.3 to the overlap between y' and y'':

$$|\sigma'| + |\sigma''| > \mathrm{d}(y', y) + \tilde{\mathrm{d}}(y, y'') + |y| - 2\mathrm{d}(c),$$

which implies

$$|\sigma'| > \frac{1}{2}(\mathrm{d}(y', y) + \tilde{\mathrm{d}}(y, y'')) + \frac{1}{2}|y| - \mathrm{d}(c). \tag{3}$$

We now use Lemma 4.7 and Equation 3 to obtain

$$\begin{aligned} |\sigma| &> |\sigma'| + |y| - \mathrm{d}(c) \\ &> \frac{1}{2}(\mathrm{d}(y', y) + \tilde{\mathrm{d}}(y, y'')) + \frac{1}{2}|y| - \mathrm{d}(c) + |y| - \mathrm{d}(c) \\ &= \frac{1}{2}(\mathrm{d}(y', y) + \tilde{\mathrm{d}}(y, y'')) + \frac{3}{2}|y| - 2\mathrm{d}(c). \end{aligned}$$

Solving for $\frac{1}{2}(\mathrm{d}(y', y) + \tilde{\mathrm{d}}(y, y''))$ and using Definition 3.2 gives us our result:

$$\begin{aligned} \frac{1}{2}(\mathrm{d}(y', y) + \tilde{\mathrm{d}}(y, y'')) &< |\sigma| - \frac{3}{2}|y| + 2\mathrm{d}(c) \\ &< |\sigma| - |y| + 2\mathrm{d}(c) - \frac{1}{2}(\frac{2}{3}\mathrm{d}(c) + \frac{2}{3}|\sigma|) \\ &= \frac{5}{3}\mathrm{d}(c) + \frac{2}{3}|\sigma| - |y|. \end{aligned}$$

■

4.3 Analysis of the Algorithm

We now analyze our algorithm G-ShortString. The structure of our approach is similar to that of [2], though the analysis we use in each case is completely different than that used for ShortString. We relate the performance of our algorithm to that of Generic; the case of interest is when a cycle in CC is a $(\frac{2}{3}, \frac{2}{3})$-HO2-cycle.

Lemma 4.11 *For each cycle $\gamma \in CC$ which is not a $(\frac{2}{3}, \frac{2}{3})$-HO2-cycle, Algorithm* G-ShortString *produces a superstring no longer than* Generic *would produce on the same cycle γ.*

Proof: We observe that step 4b) of G-Shortstring handles any cycle $\gamma \in CC$ which is not a $(\frac{2}{3}, \frac{2}{3})$-HO2-cycle. It selects an edge e and extends the cycle γ in exactly the same way as Generic. It then fully extends each representative $r_\ell \in \gamma$ to cover the remaining strings in each cycle c_ℓ. The only difference between the two algorithms in their handling of these cycles is that we perform full extension before concatenating with the strings from other cycles in CC. This does not affect the length of the resulting string. ■

We now must show, according to Lemma 2.2, that for each $(\frac{2}{3}, \frac{2}{3})$-HO2-cycle, we attain the bound specified by Lemma 2.2.

Lemma 4.12 *Let γ be a $(\frac{2}{3}, \frac{2}{3})$-HO2-cycle in CC with r_j the representative of cycle c_j and r_k the representative of c_k. Then $ov^n_\gamma + Ext(\gamma) \leq \frac{5}{3}(d(c_j) + d(c_k))$.*

Proof: Assume without loss of generality $\mathrm{d}(c_j) \geq \mathrm{d}(c_k)$. Because r_j has high overlap at both ends with r_k, there must be at least one $(\frac{2}{3}, \frac{2}{3})$-repeater σ' in c_t, with $\sigma' = \mathrm{per}(c_j)$. All strings in c_j must have at least one $(\frac{2}{3}, \frac{2}{3})$-repeater, otherwise we would not have chosen r_j as representative.

We consider two cases:

1. All strings in c_j have a small $(\frac{2}{3}, \frac{2}{3})$-repeater.
2. At least one string in c_j has no small $(\frac{2}{3}, \frac{2}{3})$-repeaters.

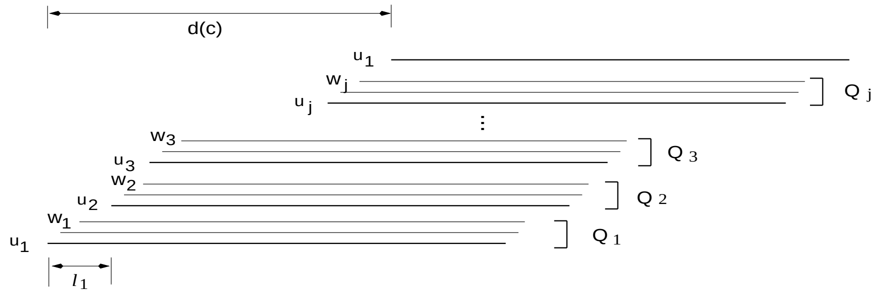

Figure 7: Case 1 of Lemma 4.12.

In each case, we must show that *some* string in c_j must have been able to achieve the bound. Because the representative is selected by comparing the worst case costs of each string, the existence of such a string is sufficient.

Case 1: All strings in c_j have a small $(\frac{2}{3}, \frac{2}{3})$-repeater.

The proof of Lemma 4.4 suggests our strategy: if two strings with different $(\frac{2}{3}, \frac{2}{3})$-repeaters begin near each other, then the sum of their periods must be close to $\mathrm{d}(c)$. If they do not begin near each other, then we can save on extension by the amount of this gap.

Because we're in Case 1, each string has at least one small $(\frac{2}{3}, \frac{2}{3})$-repeater. No string has more than one small $(\frac{2}{3}, \frac{2}{3})$-repeater by Lemma 4.4, and so each string has exactly one small $(\frac{2}{3}, \frac{2}{3})$-repeater. More than one string may have the same small $(\frac{2}{3}, \frac{2}{3})$-repeater active.

Claim 4.13 *Let σ and σ' be small $(\frac{2}{3}, \frac{2}{3})$-repeaters in cycle c. Let Q be the set of strings in which σ is active, and let Q' be the set of strings in which σ' is active. Then there is a rotation of the cyclic ordering of the strings in c such that all of the strings in Q appear before all of the strings in Q'.*

Proof: For purpose of contradiction let t and v be two strings in Q and let t' and v' be two strings in Q' such that they appear in the cyclic order t, t', v, v'. Without loss of generality let $\mathrm{d}(t, v) \leq \frac{1}{2}\mathrm{d}(c)$; otherwise $\mathrm{d}(v, t) \leq \frac{1}{2}\mathrm{d}(c)$ and the same argument follows. Consider the prefixes of t and v which are the left witness for σ; both prefixes must be substrings of the same copy of y_σ. Since t' is between t and v, then it also must have a prefix y'_ℓ which has period σ. The same argument holds for the suffixes of t, v and t', so σ must be active in t'. But then t' has both σ and σ' active, contradicting Lemma 4.4. ■

We resume our analysis of Case 1. Let σ_1 be the largest of the small $(\frac{2}{3}, \frac{2}{3})$-repeaters in c, and let Q_1 be the set of strings in which σ_1 is active. Number the remaining small $(\frac{2}{3}, \frac{2}{3})$-repeaters $s_2, \ldots, s_m$, and let Q_i, $1 \leq i \leq m$, be the set of strings in which σ_i is active. The Q_i partition the strings of the cycle, and by Claim 4.13 the Q_i form a cyclic ordering. Let u_i, $1 \leq i \leq m$ be the leftmost string in each group Q_i, and let w_i, $1 \leq i \leq m$ be the rightmost string in each group Q_i. Let $\ell_i = \mathrm{d}(v_i, v_{i+1})$, $1 \leq i < j$. (See Figure 7.)

First we apply Corollary 2.3 to derive a lower bound on the distance ℓ_1 between u_1 and u_2.

$$
\begin{aligned}
|\sigma_1| + |\sigma_2| &> \mathrm{ov}(|y_{\sigma_1}|, |y_{\sigma_2}|) \\
&\geq |y_{\sigma_1}| - \ell_1 \\
\Rightarrow 2|\sigma_1| &> |y_{\sigma_1}| - \ell_1
\end{aligned}
$$

$$\Rightarrow \ell_1 \quad > \quad |y_{\sigma_1}| - 2|\sigma_1|. \tag{4}$$

Now we bound the anticipated extension cost $\mathrm{Cost}(u_1, \sigma_1)$,

$$\begin{aligned}
\mathrm{Cost}(u_1, \sigma_1) &= |y_{\sigma_1}| + \min\{\mathrm{E}_\ell(u_1, \sigma_1), \mathrm{E}_r(u_1, \sigma_1)\} + |\sigma_1| \\
&\leq |y_{\sigma_1}| + \mathrm{E}_\ell(u_1, \sigma_1) + |\sigma_1|.
\end{aligned}$$

If we extend u_1 to the left, the last string we will have to cover will be u_2, so $\mathrm{E}_\ell(u_1, \sigma_1) = \mathrm{d}(c) - \ell_1$, and then we use Equation 4:

$$\begin{aligned}
&= |y_{\sigma_1}| + \mathrm{d}(c) - \ell_1 + |\sigma_1| \\
&\leq \mathrm{d}(c) + 3|\sigma_1| \\
&\leq \frac{5}{3}(\mathrm{d}(c) + |\sigma_1|).
\end{aligned}$$

The last inequality follows from the fact that σ_1 is a small $(\frac{2}{3}, \frac{2}{3})$-repeater, so $|\sigma_1| < \frac{1}{2}\mathrm{d}(c) < \frac{1}{3}(\mathrm{d}(c) + |\sigma_1|)$.

Case 2: At least one string in c_j has no small $(\frac{2}{3}, \frac{2}{3})$-repeaters.

Throughout the proof of this case, we fix s to be a particular string; in some cases, but not all, s will prove to be a good choice of representative. When it does not, we will show that there is another string whose anticipated cost is small enough.

Let A be the set of m' strings which do not have a small repeater; there is at least one such string because we are in Case 2. For the purpose of identifying s, rename the strings in A, $a_1, \ldots, a_{m'}$. Let σ_i be the smallest repeater which is active in each of the strings a_i. Then let $s = a_k$, with k chosen such that $|\sigma_k| \geq |\sigma_i|$, $1 \leq i \leq m'$. In other words, s is the string whose smallest $(\frac{2}{3}, \frac{2}{3})$-repeater is the largest, over all the strings in c.

By our choice of s, we know that for any other string t in c_j, t has at least one $(\frac{2}{3}, \frac{2}{3})$-repeater σ' such that $|\sigma'| \leq |\sigma|$.

Our strategy will be to show that either s can be extended to include any other strings in c within our bound, or that there is some particular string t whose position and length causes the extension of s to be too costly. In the latter case we show that t can be extended within our bounds.

We will consider four cases, which depend on the the composition of the cycle c.

Case 2A: $\min\{\mathrm{E}_\ell(s, \sigma), \mathrm{E}_r(s, \sigma)\} \leq \frac{5}{3}\mathrm{d}(c) + \frac{2}{3}|\sigma| - |y_\sigma|$.

In this case we can extend s either to the left if $\mathrm{E}_\ell(s, \sigma) \leq \mathrm{E}_r(s, \sigma)$, or to the right otherwise to cover the remaining strings in c. We bound $\mathrm{Cost}(s, \sigma)$:

$$\begin{aligned}
\mathrm{Cost}(s, \sigma) &\leq |y_\sigma| + \min\{\mathrm{E}_\ell(s, \sigma), \mathrm{E}_r(s, \sigma)\} + |\sigma| \\
&\leq |y_\sigma| + \frac{5}{3}\mathrm{d}(c) + \frac{2}{3}|\sigma| - |y_\sigma| + |\sigma| \\
&= \frac{5}{3}(\mathrm{d}(c) + |\sigma|).
\end{aligned}$$

This concludes the analysis of Case 2A. If Case 2A does not apply, then as in Figure 8 there must be a string $t = \mathrm{LString}(c, \sigma)$ and a string $u = \mathrm{RString}(c, \sigma)$, not necessarily distinct, which extend to the left and right, respectively, too far for s to be extended within the bounds of Case 2A. In particular, let X_σ^ℓ and X_σ^r be the copies of X_σ in which s begins and ends. Then t must extend into X_σ^ℓ, because otherwise $\mathrm{E}_\ell(s, \sigma) \leq$

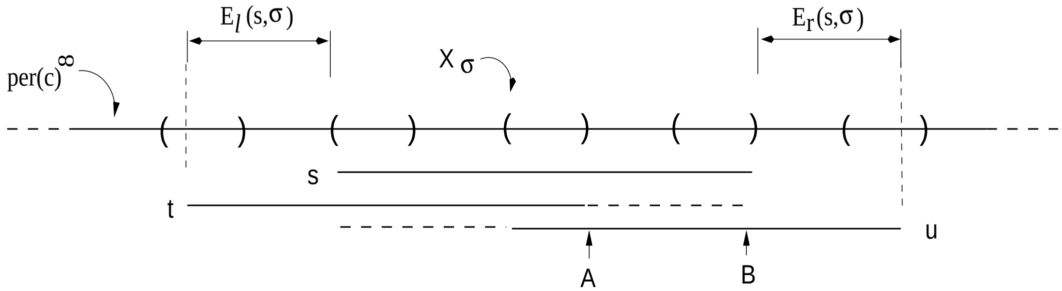

Figure 8: Case 2 of Lemma 4.12. Determining the range of possible t and u.

$2\mathrm{d}(c) - |y_\sigma| \leq \frac{5}{3}\mathrm{d}(c) + \frac{2}{3}|\sigma| - |y_\sigma|$, since $|\sigma| > \frac{1}{2}\mathrm{d}(c)$. We also note that t cannot extend to the left beyond X^ℓ_σ, or we could simply shift it over $\mathrm{d}(c)$ to the right. Therefore the left end of t is in X^ℓ_σ. The right end of t must also be within $\mathrm{d}(c)$ of the right end of s, or between points A and B marked in Figure 8. Similarly, the right end of u is in X^r_σ, and the left end may be anywhere within $\mathrm{d}(c)$ to the right of the left end of s.

Because each string in c must have at least one $(\frac{2}{3}, \frac{2}{3})$-repeater active, let σ' be the smallest $(\frac{2}{3}, \frac{2}{3})$-repeater active in t, and σ'' the smallest $(\frac{2}{3}, \frac{2}{3})$-repeater active in u. The position of the right end of t (left end of u) will determine whether $X_{\sigma'}$ ($X_{\sigma''}$) is nested within X_σ or linked with it. The remaining cases which we consider all have $\min\{\mathrm{E}_\ell(s,\sigma), \mathrm{E}_r(s,\sigma)\} > \frac{5}{3}\mathrm{d}(c) + \frac{2}{3}|\sigma| - |y_\sigma|$ and are determined by whether $t = u$ and whether $X_{\sigma'}$ and $X_{\sigma''}$ are linked with or nested within X_σ.

In order to simplify our analysis, we will often assume that a string with an active repeater σ extends from the left end of one copy of y_σ to the right end of another copy of y_σ. This assumption is pessimistic in two ways; first, we may be over-charging for extension, if a string does not go as far as the right end of y_σ and we assume it does. Second, witnesses longer than the minimum for $(\frac{2}{3}, \frac{2}{3})$-repeaters give us stronger results when we apply Corollary 2.3.

Case 2B: $\min\{\mathrm{E}_\ell(s,\sigma), \mathrm{E}_r(s,\sigma)\} > \frac{5}{3}\mathrm{d}(c) + \frac{2}{3}|\sigma| - |y_\sigma|$, $t = u$.
We will show that t can be extended within the desired bounds. Recall that σ' is the smallest $(\frac{2}{3}, \frac{2}{3})$-repeater active in t. Observe that $\mathrm{E}_\ell(s,\sigma)$ and $\mathrm{E}_r(s,\sigma)$ span the length of a single copy of $y_{\sigma'}$ with some overlap between two copies of X_σ. This observation gives rise to the following identity:

$$\mathrm{E}_\ell(s,\sigma) + \mathrm{E}_r(s,\sigma) = |y_{\sigma'}| \ + 2\mathrm{d}(c) - |y_\sigma|. \tag{5}$$

Now consider extending t to the right. Any string t' which begins within $\mathrm{d}(t,s)$ of the beginning of t must end before s due to the no-substring assumption, and we will only need to extend t by $\tilde{\mathrm{d}}(y_{\sigma'}, y_\sigma)$, to the end of X_σ. (See Figure 9(a).) We will also have to consider the case where a string v begins to the right of s and extends beyond the right end of s. We call v an *interloper*. We first consider the case where there are no interlopers, then when there is an interloper on one side, and finally when there is an interloper on each side.

If there are no interlopers, then by the definition of interloper, we only have to extend t left or right to the end of string s. Therefore $\mathrm{E}_\ell(t,\sigma') \leq \mathrm{d}(c) - \mathrm{E}_\ell(s,\sigma)$ and $\mathrm{E}_r(t,\sigma') \leq$

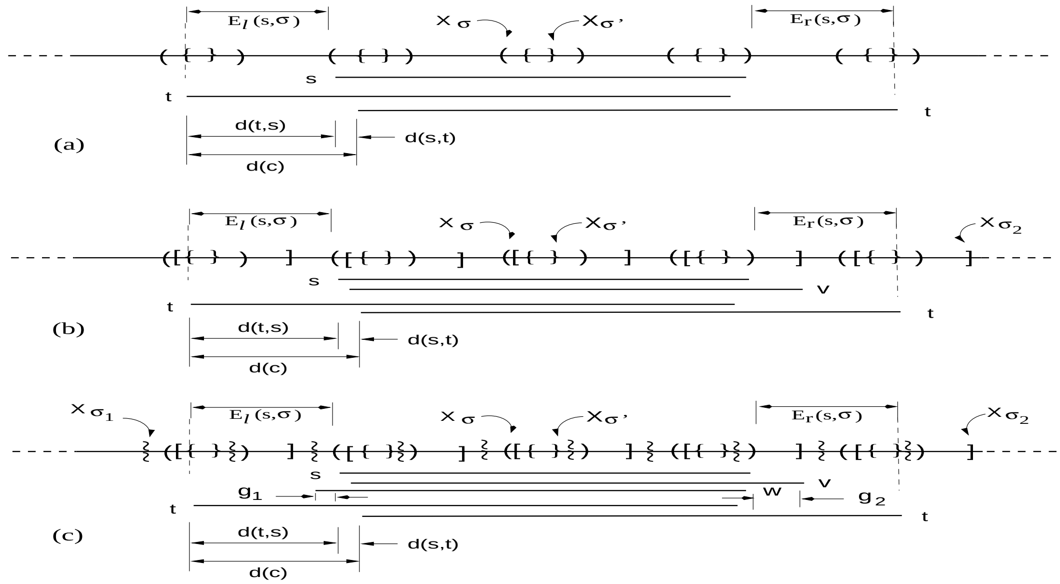

Figure 9: Case 2B of Lemma 4.12. (a) Without an interloper. (b) With one interloper v. (c) With two interlopers v and w.

$\mathrm{d}(c) - \mathrm{E}_r(s, \sigma)$:

$$
\begin{aligned}
\mathrm{Cost}(t, \sigma') &= |y_{\sigma'}| + \min\{\mathrm{d}(c) - \mathrm{E}_\ell(s, \sigma), \mathrm{d}(c) - \mathrm{E}_r(s, \sigma)\} + |\sigma'| \\
&\le |y_{\sigma'}| + \mathrm{d}(c) - \tfrac{1}{2}(\mathrm{E}_\ell(s, \sigma) + \mathrm{E}_r(s, \sigma)) + |\sigma'| \\
&= \tfrac{1}{2}|y_{\sigma'}| + \tfrac{1}{2}|y_\sigma| + |\sigma'| && \text{(Eq. 5.)} \\
&< \mathrm{d}(c) + \tfrac{1}{2}|\sigma| + \tfrac{3}{2}|\sigma'| && \text{(Lemma 3.4.)} \\
&< \tfrac{3}{2}(\mathrm{d}(c) + |\sigma'|).
\end{aligned}
$$

Suppose there is an interloper on one side. Let v be the interloper which extends the furthest to the right as in Figure 9(b). Because all strings must have an active $(\frac{2}{3}, \frac{2}{3})$-repeater, let σ_2 be the smallest $(\frac{2}{3}, \frac{2}{3})$-repeater active in v. By our conditions on where v starts and ends, X_{σ_2} must be linked with X_σ and contain $X_{\sigma'}$ as shown. We know by our choice of s and σ that $|\sigma| > |\sigma_2|$. By Lemma 4.5, $|\sigma_2| > \frac{1}{2}\mathrm{d}(c)$, so we apply Lemma 4.7 to conclude that

$$|y_\sigma| < \frac{3}{2}\mathrm{d}(c). \tag{6}$$

If v goes beyond X_σ to the right as in the Figure, we will extend t to the left. As above when there were no interlopers, we use $\mathrm{E}_\ell(t, \sigma') = \mathrm{d}(c) - \mathrm{E}_\ell(s, \sigma)$:

$$
\begin{aligned}
\mathrm{Cost}(t, \sigma') &= |y_{\sigma'}| + \mathrm{d}(c) - \mathrm{E}_\ell(s, \sigma) + |\sigma'| \\
&= |y_{\sigma'}| + \mathrm{d}(c) - (|y_{\sigma'}| + 2\mathrm{d}(c) - |y_\sigma| - \mathrm{E}_r(s, \sigma)) + |\sigma'| && \text{(Eq. 5.)} \\
&= |y_\sigma| + \mathrm{E}_r(s, \sigma) - \mathrm{d}(c) + |\sigma'| \\
&< |y_\sigma| + |\sigma'| && (\mathrm{E}_\ell(s, \sigma) < \mathrm{d}(c).) \\
&< \tfrac{3}{2}(\mathrm{d}(c) + |\sigma'|). && \text{(Eq. 6.)}
\end{aligned}
$$

Finally, suppose that there is an interloper in each direction, say w and v with $(\frac{2}{3}, \frac{2}{3})$-repeaters σ_1 and σ_2 respectively, as in Figure 9(c). Although this seems to present some difficulties, the situation also gives us stronger bounds because multiple characteristics are linked and we can employ Lemma 4.7.

Note that X_{σ_1} and X_{σ_2} are linked, as are X_{σ_1} and X_{σ}. Let $g_1 = \mathrm{d}(y_{\sigma_1}, y_\sigma)$ be the amount that w extends to the left beyond X_σ, and let $g_2 = \tilde{\mathrm{d}}(y_\sigma, y_{\sigma_2})$ be the amount that v extends to the right beyond X_σ. We derive a lower bound on $|\sigma|$:

$$\begin{array}{rrll}
 & |\sigma| > & |\sigma_1| + |y_\sigma| - \mathrm{d}(c) & \text{(Lemma 4.7.)} \\
 & > & |\sigma_2| + |y_{\sigma_1}| - \mathrm{d}(c) + |y_\sigma| - \mathrm{d}(c) & \text{(Lemma 4.7.)} \\
 & > & |y_{\sigma_1}| + |y_\sigma| - \frac{3}{2}\mathrm{d}(c) & \text{(Lemma 4.5.)} \\
\Rightarrow & \frac{1}{3}|\sigma| > & \frac{2}{3}|\sigma_1| - \frac{1}{6}\mathrm{d}(c) & \text{(Def. 3.2.)} \\
\Rightarrow & |\sigma_1| < & \frac{1}{4}\mathrm{d}(c) + \frac{1}{2}|\sigma|. &
\end{array}$$

Without loss of generality let $|\sigma_1| > |\sigma_2|$. We will choose to extend in the direction of the larger of σ_1 and σ_2, so in this case we will extend t to the left. Since $g_1 = \mathrm{d}(y_{\sigma_1}, y_\sigma)$ and X_{σ_1} and X_σ are linked, we conclude that

$$g_1 < |y_{\sigma_1}| - \mathrm{d}(c). \tag{7}$$

We use Equation 7, Lemma 4.7, and Equation 4.3 to bound g_1:

$$g_1 < |y_{\sigma_1}| - \mathrm{d}(c) < |\sigma_1| - |\sigma_2| < \frac{1}{4}\mathrm{d}(c) + \frac{1}{2}|\sigma| - |\sigma_2| < \frac{1}{2}|\sigma| - \frac{1}{4}\mathrm{d}(c). \tag{8}$$

We now calculate the anticipated cost of extending t to the left (in the direction of σ_1, the larger of σ_1 and σ_2):

$$\begin{array}{rcll}
\mathrm{Cost}(t, \sigma') & \leq & |y_{\sigma'}| + \mathrm{E}_\ell(t, \sigma') + |\sigma'| & \\
 & \leq & |y_{\sigma'}| + \mathrm{d}(c) - \mathrm{E}_\ell(s, \sigma) + g_1 + |\sigma'| & \\
 & < & |y_{\sigma'}| + \mathrm{d}(c) - (\frac{5}{3}\mathrm{d}(c) + \frac{2}{3}|\sigma| - |y_\sigma|) + g_1 + |\sigma'| & \text{(Case bound.)} \\
 & = & |y_{\sigma'}| - \frac{11}{12}\mathrm{d}(c) - \frac{1}{6}|\sigma| + |y_\sigma| + |\sigma'| & \text{(Eq. 8.)} \\
 & < & |y_{\sigma'}| - \frac{11}{12}\mathrm{d}(c) - \frac{1}{6}|\sigma| + (|\sigma| - |\sigma_1| + \mathrm{d}(c)) + |\sigma'| & \text{(Lemma 4.7.)} \\
 & = & |y_{\sigma'}| + \frac{1}{12}\mathrm{d}(c) + \frac{5}{6}|\sigma| - |\sigma_1| + |\sigma'|. &
\end{array}$$

In the last inequality above we were able to apply Lemma 4.7 because X_σ and X_{σ_1} are linked; now we can apply it again, because X_{σ_1} and X_{σ_2} are also linked.

$$\begin{array}{cll}
< & |y_{\sigma'}| + \frac{1}{12}\mathrm{d}(c) + \frac{5}{6}|\sigma| - (|\sigma_2| + |y_{\sigma_1}| - \mathrm{d}(c)) + |\sigma'| & \text{(Lemma 4.7.)} \\
= & |y_{\sigma'}| + \frac{13}{12}\mathrm{d}(c) + \frac{5}{6}|\sigma| - |\sigma_2| - |y_{\sigma_1}| + |\sigma'| & \\
< & |y_{\sigma'}| + \frac{13}{12}\mathrm{d}(c) + \frac{5}{6}|\sigma| - |\sigma_2| - (\frac{2}{3}\mathrm{d}(c) + \frac{2}{3}|\sigma_1|) + |\sigma'| & \text{(Def. 3.2.)} \\
< & |y_{\sigma'}| + \frac{5}{12}\mathrm{d}(c) + \frac{5}{6}|\sigma| - \frac{5}{3}|\sigma_2| + |\sigma'| & (|\sigma_1| > |\sigma_2|.) \\
< & \frac{9}{4}\mathrm{d}(c) + 2|\sigma'| - \frac{5}{3}|\sigma_2| & \text{(Lemma 3.4.)} \\
< & \frac{17}{12}\mathrm{d}(c) + 2|\sigma'| & \\
< & \frac{29}{18}(\mathrm{d}(c) + |\sigma'|). & (|\sigma'| < \frac{1}{2}\mathrm{d}(c).)
\end{array}$$

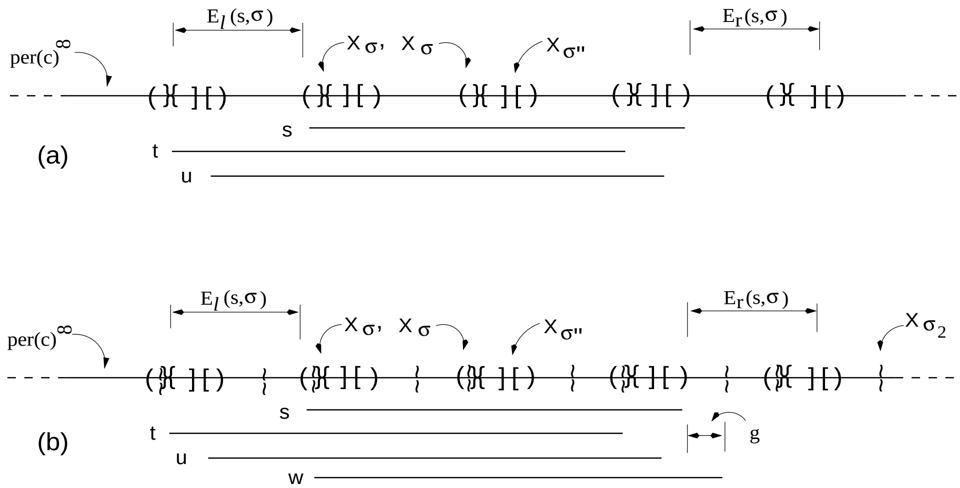

Figure 10: Case 2C of Lemma 4.12. (a) Without an interloper. (b) With an interloper w.

This concludes the analysis of Case 2B. In the remaining two cases, $t \neq u$; that is, $\mathrm{LString}(c,\sigma) \neq \mathrm{RString}(c,\sigma)$. Let σ' be the smallest $(\frac{2}{3},\frac{2}{3})$-repeater active in t and σ'' be the smallest $(\frac{2}{3},\frac{2}{3})$-repeater active in u, and without loss of generality let $|\sigma'| > |\sigma''|$. By our choice of s we know that $|\sigma| > |\sigma'| > |\sigma''|$.

If $X_{\sigma'}$ is linked with X_σ, we observe that $\mathrm{E}_\ell(s,\sigma) = \mathrm{d}(y',y)$ and $\mathrm{E}_r(s,\sigma) = \tilde{\mathrm{d}}(y'',y)$, so we can apply Lemma 4.10 and conclude that $\min\{\mathrm{E}_\ell(s,\sigma), \mathrm{E}_r(s,\sigma)\} < \frac{5}{3}\mathrm{d}(c) + \frac{2}{3}|\sigma| - |y_\sigma|$. This satisfies the bound for Case 2A. We therefore only need to consider two remaining cases: when neither $X_{\sigma'}$ nor $X_{\sigma''}$ is linked with X_σ (Case 2C), and when only $X_{\sigma''}$ is linked with X_σ (Case 2D).

Case 2C: $\min\{\mathrm{E}_\ell(s,\sigma), \mathrm{E}_r(s,\sigma)\} > \frac{5}{3}\mathrm{d}(c) + \frac{2}{3}|\sigma| - |y_\sigma|$, $X_{\sigma'}$ and $X_{\sigma''}$ both nested.
We show that t can be extended to the right within our bounds. (See Figure 10a.) Here again interlopers are possible, so we will first consider the case without an interloper, and then the case with an interloper on at least one side.

If there is no interloper, then we only have to extend t to the right as far as the end of X_σ. We use Lemma 4.9, the Case bound on $\mathrm{E}_\ell(s,\sigma)$ and $\mathrm{E}_r(s,\sigma)$, and the fact that $\mathrm{E}_\ell(s,\sigma) + \mathrm{E}_r(s,\sigma) = |y_{\sigma'}| + 2\mathrm{d}(c) - |y_\sigma|$:

$$
\begin{aligned}
\mathrm{Cost}(t,\sigma') &\leq |y_{\sigma'}| + \mathrm{E}_\ell(t,\sigma') + |\sigma'| \\
&\leq |y_{\sigma'}| + \tilde{\mathrm{d}}(y',y) + |\sigma'| \\
&= |y_{\sigma'}| + (|y_\sigma| - |y_{\sigma'}| - \mathrm{d}(y',y) + |\sigma'| \\
&= |y_\sigma| + \mathrm{E}_\ell(s,\sigma) - \mathrm{d}(c) + |\sigma'| \\
&= \tfrac{5}{3}|\sigma'| + |y_\sigma| + \mathrm{E}_\ell(s,\sigma) - \mathrm{d}(c) - \tfrac{2}{3}|\sigma'|.
\end{aligned}
$$

We apply Lemma 4.9 and the fact that $\mathrm{E}_\ell(s,\sigma) = \mathrm{d}(y', y)$ to bound the last term above,

$$\begin{array}{ll} \leq & \frac{5}{3}|\sigma'| + |y_\sigma| + \mathrm{E}_\ell(s,\sigma) - \mathrm{d}(c) - \frac{2}{3}(\mathrm{E}_\ell(s,\sigma) + |y_\sigma| - 2\mathrm{d}(c)) \\ = & \frac{5}{3}|\sigma'| + \frac{1}{3}|y_\sigma| + \frac{1}{3}\mathrm{E}_\ell(s,\sigma) + \frac{1}{3}\mathrm{d}(c) \\ < & \frac{5}{3}|\sigma'| + \frac{4}{3}\mathrm{d}(c) \qquad \text{(Lemma 3.4.)} \\ < & \frac{5}{3}(\mathrm{d}(c) + |\sigma'|). \end{array}$$

Because $|\sigma'| > |\sigma''|$ and σ' is active in t, t was our choice of representative and we elected to extend to the right. Therefore the only interloper which concerns us is one like w in Figure 10b. Let σ_2 be the smallest $(\frac{2}{3}, \frac{2}{3})$-repeater active in w. Due to our choice of s, $|\sigma| > |\sigma_2|$, we can apply Lemma 4.7 to obtain:

$$|\sigma_2| < |\sigma| - |y_\sigma| + \mathrm{d}(c). \tag{9}$$

Let $g = \tilde{\mathrm{d}}(y, y_{\sigma_2})$ be the distance that the interloper w extends beyond X_σ. We observe that

$$g < |y_{\sigma_2}| - \mathrm{d}(c) - (\mathrm{E}_\ell(s,\sigma) - 2\mathrm{d}(c) - |y_\sigma|) = |y_{\sigma_2}| + \mathrm{d}(c) - \mathrm{E}_\ell(s,\sigma) - |y_\sigma|. \tag{10}$$

We now calculate the cost of extending t to the right:

$$\begin{array}{rcll} \mathrm{Cost}(t,\sigma') & \leq & |y_{\sigma'}| + \mathrm{E}_r(t,\sigma') + |\sigma'| & \\ & \leq & |y_{\sigma'}| + \tilde{\mathrm{d}}(y_{\sigma'}, y_{\sigma_2}) + g + |\sigma'| & \\ & = & |y_{\sigma'}| + (|y_\sigma| - (\mathrm{d}(c) - \mathrm{E}_\ell(s,\sigma)) - |y_{\sigma'}|) + g + |\sigma'| & \\ & = & |y_\sigma| - \mathrm{d}(c) + \mathrm{E}_\ell(s,\sigma) + g + |\sigma'| & \\ & < & |y_\sigma| - \mathrm{d}(c) + \mathrm{E}_\ell(s,\sigma) & \\ & & +(|y_{\sigma_2}| + \mathrm{d}(c) - \mathrm{E}_\ell(s,\sigma) - |y_\sigma|) + |\sigma'| & \text{(Eq. 10.)} \\ & = & |y_{\sigma_2}| + |\sigma'| & \\ & < & \mathrm{d}(c) + |\sigma_2| + |\sigma'| & \text{(Lemma 3.4.)} \\ & < & \mathrm{d}(c) + |\sigma'| + (|\sigma| - |y_\sigma| + \mathrm{d}(c)) & \text{(Eq. 9.)} \\ & < & \mathrm{d}(c) + |\sigma'| + \frac{1}{3}|\sigma| + \frac{1}{3}\mathrm{d}(c) & \text{(Def. 3.2.)} \\ & < & \frac{5}{3}(\mathrm{d}(c) + |\sigma'|). & \end{array}$$

Case 2D: $\min\{\mathrm{E}_\ell(s,\sigma), \mathrm{E}_r(s,\sigma)\} > \frac{5}{3}\mathrm{d}(c) + \frac{2}{3}|\sigma| - |y_\sigma|$, $X_{\sigma''}$ (but not $X_{\sigma'}$) linked with X_σ.

In this case $X_{\sigma'}$ might be nested within $X_{\sigma''}$ (Figure 11 a), or not (Figure 11b. It is an unlikely case to give us trouble, because here the smaller $(\frac{2}{3}, \frac{2}{3})$-repeater has the larger characteristic, and it turns out that we achieve a stronger bound than in other cases.

Subcase (i). Because $X_{\sigma''}$ contains $X_{\sigma'}$, Lemma 4.5 applies, so $|\sigma''| > \frac{1}{2}\mathrm{d}(c)$. Since Lemma 4.7 also applies we have

$$|y_\sigma| < |\sigma| - |\sigma''| + \mathrm{d}(c) < \frac{3}{2}\mathrm{d}(c). \tag{11}$$

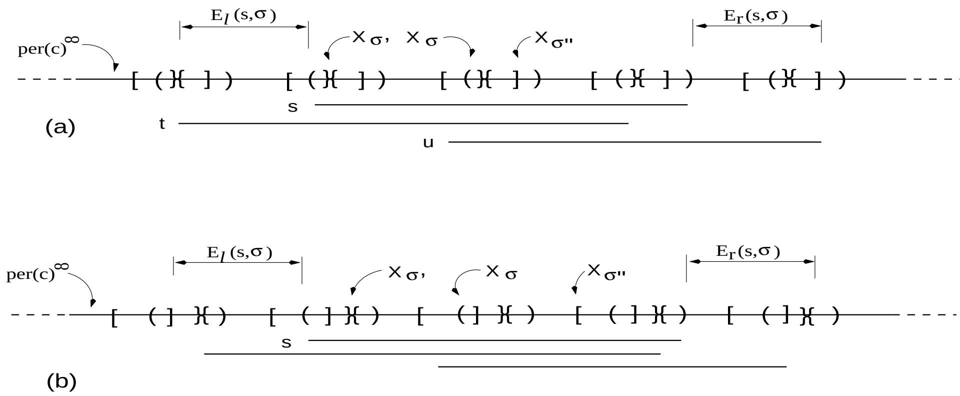

Figure 11: Case 2D of Lemma 4.12. (a) $X_{\sigma'}$ nested within $X_{\sigma''}$. (b) $X_{\sigma'}$ not nested within $X_{\sigma''}$.

If there are no interlopers, we can now bound the anticipated cost of extending t to the right as follows:

$$\begin{aligned}
\mathrm{Cost}(t,\sigma') &\leq |y_{\sigma'}| + \mathrm{E}_\ell(t,\sigma') + |\sigma'| \\
&\leq |y_{\sigma'}| + \tilde{\mathrm{d}}(y_{\sigma'}, y_\sigma) + |\sigma'| \\
&= |y_{\sigma'}| + (|y_\sigma| - |y_{\sigma'}| - (\mathrm{d}(c) - \mathrm{E}_\ell(s,\sigma))) + |\sigma'| \\
&< |y_\sigma| + |\sigma'| && (\mathrm{E}_\ell(s,\sigma) < \mathrm{d}(c)) \\
&< \tfrac{3}{2}(\mathrm{d}(c) + |\sigma'|). && \text{(Eq. 11)}
\end{aligned}$$

Now suppose there was an interloper v with smallest $(\frac{2}{3},\frac{2}{3})$-repeater σ_2. Then X_{σ_2} would be linked with $X_{\sigma''}$ and X_σ, and Lemma 4.10 would apply as in Figure 6(d), and we would once again be in Case 2A.

Subcase (ii). Now $X_{\sigma'}$ is not nested within $X_{\sigma''}$, as in Figure 11b. If there are no interlopers, then we only have to extend t to the right to the end of X_σ:

$$\begin{aligned}
\mathrm{Cost}(t,\sigma') &\leq |y_{\sigma'}| + \mathrm{E}_\ell(t,\sigma') + |\sigma'| \\
&\leq |y_{\sigma'}| + \mathrm{E}_\ell(s,\sigma) + \mathrm{d}(c) - |y_{\sigma'}| - (2\mathrm{d}(c) - |y_\sigma|) + |\sigma'| \\
&= |y_\sigma| + \mathrm{E}_\ell(s,\sigma) - \mathrm{d}(c) + |\sigma'|.
\end{aligned}$$

We apply Lemma 4.9 to complete the analysis:

$$\begin{aligned}
\mathrm{Cost}(t,\sigma') &< |y_\sigma| + (|\sigma'| - |y_\sigma| + 2\mathrm{d}(c)) - \mathrm{d}(c) + |\sigma'| \\
&= 2|\sigma'| + \mathrm{d}(c) \\
&< \frac{3}{2}(\mathrm{d}(c) + |\sigma'|.
\end{aligned}$$

As in Case (i), if there is an interloper then Lemma 4.10 will apply (Figures 6c or d), and we have Case 2A.

This completes the proof of Case IId, which completes the proof of the lemma. ■

We now combine Lemmas 2.2, 4.11, and 4.12 to obtain:

Theorem 4.14 *Algorithm* G-SHORTSTRING(S) *is a* $2\frac{2}{3}$*-approximation for the shortest superstring problem.*

References

[1] C. Armen and C.Stein. Improved length bounds for the shortest superstring problem. In *Proceedings of Workshop on Algorithms and Data Structures*, pages 494–505, 1995.

[2] C. Armen and C. Stein. Short supertrings and the structure of overlapping strings. To appear in J. of Computational Biology, 1995.

[3] Chris Armen. *Approximation Algorithms for the Shortest Superstring Problem.* PhD thesis, Dartmouth College, July 1995.

[4] A. Blum, T. Jiang, M. Li, J. Tromp, and M. Yannakakis. Linear approximation of shortest superstrings. In *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing*, pages 328–336, 1991.

[5] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms.* MIT Press/McGraw-Hill, 1990.

[6] A. Czumaj, L. Gasieniec, M. Piotrow, and W. Rytter. Parallel and sequential approximations of shortest superstrings. In *Proceedings of Fourth Scandinavian Workshop on Algorithm Theory*, pages 95–106, 1994.

[7] A. Lesk (edited). *Computational Molecular Biology, Sources and Methods for Sequence Analysis.* Oxford University Press, 1988.

[8] N. Fine and H. Wilf. Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society*, 16:109–114, 1965.

[9] A.M. Frieze, G. Galbiati, and F. Maffoli. On the worst case performance of some algorithms for the asymmetric travelling salesman problem. *Networks*, 12:23–39, 1982.

[10] J. Gallant, D. Maier, and J. Storer. On finding minimal length superstrings. *Journal of Computer and System Sciences*, 20:50–58, 1980.

[11] D. Gusfield, G. Landau, and B. Schieber. An efficient algorithm for the all pairs suffix-prefix problem. *Information Processing Letters*, (41):181–185, March 1992.

[12] Tao Jiang and Ming Li. Approximating shortest superstrings with constraints. *Therotical Computer Science*, (134):473–491, 1994.

[13] J.D. Kececioglu and E.W. Myers. Combinatorial algorithms for dna sequence assembly. *Algorithmica*, 13(1/2):7–51, 1995.

[14] John D. Kececioglu. *Exact and approximation algorithms for DNA sequence reconstruction.* PhD thesis, University of Arizona, 1991.

[15] R. Kosaraju, J. Park, and C. Stein. Long tours and short superstrings. In *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, November 1994.

[16] M. Li. Towards a DNA sequencing theory (learning a string). In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, pages 125–134, 1990.

[17] Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization, Algorithms and Complexity.* Prentice-Hall, Englewood Cliffs, NJ, 1982.

[18] H. Peltola, H. Soderlund, J. Tarjio, and E. Ukkonen. Algorithms for some string matching problems arising in molecular genetics. In *Proceedings of the IFIP Congress*, pages 53–64, 1983.

[19] Graham A. Stephen. *String searching algorithms.* World Scientific, 1994.

[20] J. Storer. *Data compression: methods and theory.* Computer Science Press, 1988.

[21] J. Tarhio and E. Ukkonen. A greedy approximation algorithm for constructing shortest common superstrings. *Theoretical Computer Science*, 57:131–145, 1988.

[22] Shang-Hua Teng and Frances Yao. Approximating shortest superstrings. In *Proceedings of the 34th Annual Symposium on Foundations of Computer Science*, pages 158–165, November 1993.

[23] J. Turner. Approximation algorithms for the shortest common superstring problem. *Information and Computation*, 83:1–20, 1989.