Algorithms Seminar: November 12, 2004

Lecturer: Steve Skiena

Scribe: Janet Braunstein

Introduction

The hip-hop DJ problem is related to Adrian Fisher's string-compliant maze problem introduced by George Hart on April 16, 2004. Previous discussion on the maze problem included the case where a path with labeled edges, as well as start and finish vertices, is given and we wish to find the shortest word w such that there is a walk from start to finish whose edge labels correspond to $w \dots w$. (*See notes by Andrew Mehler.*) The variation on the problem presented this week is the following:

Problem: Given a string, $S = s_1 s_2 \dots s_n$, find the shortest associated string, $R = r_1 r_2 \dots r_m$, from which S can be reconstucted in the following manner: Consider the letters in R to be the edge labels of a path. Start with r_1 (walk forward across edge r_1). Add either r_1 or r_2 (walk backward across edge r_1 or forward across edge r_2). Continue in this fashion, choosing at each step to either move backward or forward. This can be thought of as a DJ looking to find the shortest record that can be used to play a particular song if he can let the record play forward or backward at any time. In the discussion, only binary strings were considered, and we required that we start at the beginning of the record to play the song.

Examples:

- For song $S = 0001111$, the associated optimal record is $R = 01$.
- Given $S = 1111000$, $R = 110$.
- Given $S = 01101101100$, $R = 01101$.

Claim. *The same letter can never appear more than twice in a row in the optimal record.*

Proof. Any song that can be played by the record containing the string $r_i 111 r_{i+4}$ can also be played by the record obtained from the original by substituting $r_i 1 r_{i+4}$ in place of $r_i 111 r_{i+4}$. Choose any point in the song which has stopped us between r_i and 1 in $r_i 111 r_{i+4}$. Let n equal the number of subsequent 1's to be played from the record before playing either r_i or

r_{i+4}. If n is even, r_i must be played after the set of 1's, and we will be between r_i and 1 in our record. We can play the same set of 1's in the song and finish in the same position by moving back and forth over the 1 in $r_i 1 r_{i+4}$. Similarly, if n is odd, r_{i+4} must be played after the set of 1's, and we will be between 1 and r_{i+4} in our record. We can play the same set of 1's in the song and finish in the same position by moving back and forth over the 1 in $r_i 1 r_{i+4}$. (A similar argument can be used for any point in the song that has stopped us between 1 and r_{i+4} in $r_i 111 r_{i+4}$ in our record.) By applying this idea iteratively, any even number of consecutive letters in a record may be compressed to two of that letter, and any odd number of consecutive letters may be compressed to one of that letter. □

More generally, we have the following claim:

Claim. *Let w be any word in the optimal record. The sequence ww^Tw cannot be contained in this record.*

Proof. This proof follows the same thought process as the proof above, with the substitution ww^Tw for the 111 in $r_i 111 r_{i+4}$ and w for the 1 in $r_i 1 r_{i+4}$. An alternate way of thinking about this is the following: Make creases in the record between w and w^T and between w^T and w, then fold the record back at the first crease and forward at the second crease. This results in three copies of w being stacked on top of each other. Since we have not physically broken the record, we can still play the song in the same fashion. If the record is viewed from above, however, it appears that there is only one copy of w, and we are simply moving back and forth across it. Therefore, any song played with the original record can also be played with the compressed record. □

Clearly, as seen in the third example above, it is possible that a compression can be made even if there are no instances of ww^Tw in R. However, this is no longer true if we require that we finish at the end of the record (s_n corresponds to r_m). If this requirement is made, then we have the following property: *A record R is compressible $\Leftrightarrow$ R contains ww^Tw.*

Proposed Algorithm

Based on the above observations, an algorithm for finding the optimal record, R, for any given song, S, was proposed: Search for any instance of ww^Tw in S. Compress S in the manner described above and repeat. When no more compressions are possible, the optimal record has been found. If the order of compressions doesn't matter, then this algorithm will

work. However, it is not immediately clear that doing two different sets of compressions will always result in the same final record. No counterexamples could be found during the session, but the conjecture could not be proven. (See notes by Xiaotian Yin for further ideas.)