Approximate Pattern Matching using the Burrows-Wheeler Transform

Nan Zhang† Amar Mukherjee‡ Don Adjeroh§ Tim Bell¶

November 15, 2002

Abstract. The compressed pattern matching problem is to locate the occurrence(s) of a pattern $P$ in a text string $T$ using a compressed representation of $T$, with minimal (or no) decompression. In this paper, we consider approximate pattern matching directly on BWT compressed text. The BWT provides a lexicographic ordering of the input text as part of its inverse transformation process. Based on this observation, pattern matching is performed by text pre-filtering, using a fast $q$-gram intersection of segments from the pattern $P$ and the text $T$. Algorithms are proposed that solve the $k$-mismatch problem in $O(n \min\{m(m - k) \log \frac{m}{k}, mu \log \frac{m}{k}\})$ time worst case, and the $k$-approximate matching problem in $O(|\Sigma| \log |\Sigma| + \frac{m \log \frac{m}{k}}{\alpha} + O)$ time on average ($\alpha \geq u$), where $u = |T|$ is the size of the text, $m = |P|$ is the size of the pattern, and $\Sigma$ is the symbol alphabet. Each algorithm requires an $O(u)$ auxiliary arrays, which are constructed in $O(u)$ time and space.

1 Introduction

The pattern matching problem is to find the occurrence of a given pattern in a given text string. This is an old problem, which has been approached from different fronts, motivated by both its practical significance and its algorithmic importance. Matches between strings are determined based on the string edit distance. Given two strings $A: a_1 \ldots a_n$, and $B: b_1 \ldots b_m$, over an alphabet $\Sigma$, and a set of allowed edit operations, the edit distance indicates the minimum number of edit operations required to transform one string into the other. Three basic types of edit operations are used: insertion, deletion, and substitution of a symbol. Given two strings $A: a_1 \ldots a_n$, and $B: b_1 \ldots b_m$, over an alphabet $\Sigma$, the exact string matching problem is to check for the existence of a substring of the text that is an exact replica of the pattern string. That is, the edit distance between the substring of $A$ and the pattern should be zero. In $k$-approximate string matching, the task is to find each substring $A_s$ of $A$, such that the edit distance between $A_s$ and $B$ is less than or equal to $k$. Another form of approximate pattern matching is the $k$-mismatch problem. The problem here is to locate all substrings of $A$ that have a maximum of $k$ mismatches with $B$. That is, only the substitution operation is allowed. The parameter $k$ thus acts as a form of threshold to determine the correctness of a match. Various algorithms have been proposed for both exact and approximate pattern matching. See (Gusfield, 1997; Navarro, 2001a) for a survey.

With the sheer volume of data easily available to an ordinary user and the fact that most of this data is increasingly in compressed format, efforts have been made to address the compressed pattern matching

                                                 
\*School of Electrical Engineering & Computer Science, University of Central Florida, Orlando, FL 32816, USA; {xzhang, amaj}@cs.ucf.edu
\†Partially supported by a grant from the National Science Foundation IIS-9977336.
\‡Lane Department of Computer Science & Electrical Engineering, West Virginia University, Morgantown, WV 26506-6109, USA; don@csse.wvu.edu
\§Department of Computer Science, University of Canterbury, New Zealand; {tim}@csc.canterbury.ac.nz
\¶Supported by a grant from the Hitachinaka Techno Center Inc, Japan.
problem. Given $T$ a text string, $P$ a search pattern, and $Z$ the compressed representation of $T$, the problem is to locate the occurrences of $P$ in $T$ with minimal (or no) decompression of $Z$. Initial attempts on compressed pattern matching were directed at compression schemes based on the LZ family (Amir et al., 1996; Farach and Thorup, 1998), where algorithms have been proposed that can search for a pattern in an LZW-compressed text string in $O(n \log^2 (\frac{n}{u}) + m)$ time, where $m = |P|$, $u = |T|$, and $n = |Z|$. Bunke and Csikir (Bunke and Csikir, 1995) proposed methods that can search for patterns in run-length encoded files in $O(um_c)$, when $m_c$ is the length of the pattern when it is compressed. In (Moura et al., 2000), $O(n + m \sqrt{u})$ algorithms were proposed for searching Huffman-encoded files. Special compression schemes that facilitate later pattern matching directly on the compressed file have also been proposed (Manber, 1997; Shibata et al., 1999).

Although there has been substantial work in compressed pattern matching, a recent survey (Bell et al., 2001) shows that little has been done on searching directly on text compressed with the Burrows-Wheeler Transform (BWT). Even less has been done on compressed domain approximate pattern matching on BWT text. Yet, in terms of data compression, empirical evidence (Burrows and Wheeler, 1994; Fenwick, 1996; Balkenhol et al., 1999) shows that the BWT is significantly better than the more popular LZ-based methods (such as GZIP and COMPRESS), and is only second to the PPM* algorithm (Cleary and Witten, 1984). In terms of running time, the BWT is much faster than PPM*, but comparable with LZ-based algorithms. So far, the major reported work on searching on BWT-compressed text are those of Sadakane (Sadakane, 2000) and Ferragina and Manzini (Ferragina and Manzini, 2000), who proposed $O(m \log u + \eta_{loc} \log^2 u)$ and $O(m + \eta_{loc} \log^2 u)$ time algorithms respectively, to locate all $\eta_{loc}$ occurrences of $P$ in $T$, where $0 < \epsilon < 1$. In (Adjeroh et al., 2002; Bell et al., 2002), methods were reported that can search for exact matches in BWT text in $O(m \log \frac{u}{m})$ time. The methods, however, considered only exact pattern matching. In general, compressed pattern matching algorithms have focused mainly on exact matching. One exception is (Mäkinen et al., 2001) where they considered RLE-encoded text. Although the one-page abstract of (Adjeroh et al., 2002) discussed $k$-approximate matching very briefly, there was no mention of $k$-mismatches.

In this paper, we provide algorithms for solving both the $k$-mismatch problem and the $k$-approximate matching problem directly on BWT compressed text. Although the BWT has been referenced mainly for its performance in compression, in a sense, this paper also views the BWT as a mechanism that facilitates searching on the text.

2 The Burrows-Wheeler Transform

The BWT performs a permutation of the characters in the text, such that characters in lexically similar contexts will be near to each other. The important procedures in BWT-based compression/decompression are the forward and inverse BWT, and the subsequent encoding of the permuted text.

The forward transform. Given an input text $T = t_1 t_2 \ldots t_u$, the forward BWT is composed of three steps: 1) Form a $u \times u$ matrix $M'$, with each row in $M'$ representing one permutation of $T$; 2) Sort the rows of $M'$ lexicographically to form another matrix $M$. $M$ includes $T$ as one of its rows; 3) Record $L$, the last column of the sorted permutation matrix $M$, and $id$, the row number for the row in $M$ that corresponds to the original text string $T$.

The output of the BWT is the pair, $(L, id)$. Generally, the effect is that the contexts that are similar in $T$ are made to be closer together in $L$. This similarity in nearby contexts can be exploited to achieve compression. As an example, suppose $T = \text{mississippi}$. Let $F$ and $L$ denote the array of first and last characters respectively. Then, $F = \text{iiiiimpsssss}$ and $L = \text{psmipissii}$. The output of the transformation will be the pair: $(\text{psmipissii}, 5)$ (indices are from 1 to $u$).

The inverse transform. The BWT is reversible. It is quite striking that given only the $(L, id)$ pair, the original text can be recovered exactly. The inverse transformation can be performed using the following steps (Burrows and Wheeler, 1994): 1) Sort $L$ to produce $F'$, the array of first characters; 2) Compute $V$, the transformation vector that provides a one-to-one mapping between the elements of $L$ and $F'$, such that $F[V[j]] = L[j]$. That is, for a given symbol $\sigma \in \Sigma$, if $L[j]$ is the $c$-th occurrence of $\sigma$ in $L$, then $V[j] = i$,
where $F[i]$ is the $c$-th occurrence of $\sigma$ in $F$; 3) Generate the original text $T$, since the rows in $M$ are cyclic rotations of each other, the symbol $L[i]$ cyclically precedes the symbol $F[i]$ in $T$. That is, $L[V[j]]$ cyclically precedes $L[j]$ in $T$.

For the example with mississippi, we will have $V = [6\ 8\ 9\ 5\ 1\ 7\ 2\ 10\ 11\ 3\ 4]$. Given $V$ and $L$, we can generate the original text by iterating with $V$. This is captured by a simple algorithm: $T[u + 1 - i] = L[V^{i-1}[i]]$, where $V^0[i] = s$; and $V^{i+1}[i] = V[V^i[s]]$, $1 \leq s \leq u$. In practical implementations, the transformation vector $V$ is computed by use of two arrays of counts $C = c_1, c_2, \ldots, c_{|\Sigma|}$, and $R = r_1, r_2, \ldots, r_u$: $V[i] = R[i] + C[L[i]]$, $0 \leq i < u$ where, for a given index, $c_i$, $C[c_i]$ stores the number of occurrences in $L$ of all the characters preceding $\sigma$, the $c$-th symbol in the $\Sigma$, and $R[j]$ keeps count of the number of occurrences of character $L[j]$ in the prefix $L[1, 2, \ldots, j]$ of $L$ (can be calculated on the fly using an array of $\Sigma$ counts, which accumulates the number of times each character has appeared in $L$ so far.) With $V$, we can use the relation between $L, F$, and $V$ to avoid the sorting required to obtain $F$. Thus, we can compute $F$ in $O(u)$ time.

**BWT-based compression.** Compression with the BWT is usually accomplished in four phases, viz: but $\rightarrow$ mtf $\rightarrow$ rle $\rightarrow$ ule, where but is the forward BWT transform; mtf is move-to-front encoding (Bentley et al., 1986) used to further transform $L$ for better compression (this usually produces runs of the same symbol); rle is run length encoding of the runs produced by the mtf; and ule is variable length coding of the rle output using entropy encoding methods, such as Huffman or arithmetic coding.

### 3 Overview of Our Approach

The motivation for our approach is the observation that the BWT provides a lexicographic ordering of the input text as part of its inverse transformation process. The decoder only has limited information about the sorted context, but it is possible to exploit this to perform an initial match on two symbols (a character and its context), and then decode only that part of the text to see if the pattern match continues. We give our description in terms of the $F, L$ and $V$ arrays, (i.e. the output of the but transformation — before the mtf and further encoding). The methods can be modified to search directly on the encoded output.

Given $F$ and $L$, we can obtain a set of bi-grams for the original text sequence $T$. Let $Q_f$ and $Q_p$ be the set of bi-grams for the text string $T$ and the pattern $P$ respectively. We can use these bi-grams for at least two purposes:

**Pre-filtering.** To search for potential matches, we consider only the bi-grams that are in the set $Q_f \cap Q_p$.

If the intersection is empty, it means that the pattern does not occur in the text, and we don’t need to do any further decompression.

**Approximate pattern matching.** We can generalize the bi-grams to the more usual $q$-grams, and perform $q$-gram intersection on $Q_f$ and $Q_p$ — the set of $q$-grams from $T$ and $P$. At a second stage we verify if the $q$-grams in the intersection are part of a true $k$-approximate match to the pattern.

**Example.** Suppose $T = abracadabra$, and $P = rac$. We will have $L = carab$, and $F = aaabcr$. Using $F$ and $L$, we can obtain the bi-grams: $Q_f = \{ac, ab, ba, ca, ra\}$ and $Q_p = \{ra, ac\}$. Intersecting the two, we see that only $\{ra, ac\}$ are in the intersection. For exact pattern matching, $ac$ will be eliminated, and thus we will only need to check in the area in $T$ that contains $ra$, since any match must contain $ra$.

### 3.1 Auxiliary Arrays

The inverse BWT transformation is defined as: $\forall i, T[u + 1 - i] = L[V^{i-1}[i]]$, where $V^0[i] = V[V[\ldots V[s]]] (i$ times) and $V^0[s] = s$. Since $V^i[s]$ is just one more iteration on $V^{i-1}[s]$, we can reduce the time required by storing the intermediate results, to be used at the next iteration of the loop. Since $F$ is already sorted, and $F[i] = L[V[s]]$, we can use a mapping between $T$ and $F$, (rather than $L$), so that we can use binary search on $F$. We can use an auxiliary array $H$ (or its reverse $Hr$) to hold the intermediate steps of the indexing using $V$: $\forall i, T[i] = F[H[u + 1 - i]]$; $Hr = \text{reverse}(H)$, and $T[i] = F[Hr[i]]$. 


Example. The mapping vectors are shown below for $T = \text{abracadabra}$, $u = |T| = 6$. $Hrs$ is an index vector to the elements of $Hr$ in sorted order. It is defined as the inverse of $Hr$, that is, $F[i] = T[Hrs[i]]$. It may be observed that $Hrs$ also corresponds to the suffix array of the original text $T$, an important data structure for searching and sequence analysis (Manber and Myers, 1993; Gusfield, 1997). Thus, it provides the index to the lexicographically sorted list of all the suffixes from the text. We denote this matrix of sorted suffixes as the $S$ matrix. The table below shows the auxiliary vectors and the $S$ matrix for the example text $T = \text{abracadabra}$.

We have inserted the special character $\#$, ($\# \in \Sigma, \# < \sigma_i, \forall \sigma_i \in \Sigma$), for consistency in the suffixes, and to delineate the sorted suffixes in the $M$ matrix of the BWT. The generation of $V$ requires $O(u)$ time, and hence the computation of all the auxiliary arrays also requires $O(u)$ time.

<table>
<thead>
<tr>
<th>idx</th>
<th>T</th>
<th>L</th>
<th>F</th>
<th>V</th>
<th>Hr</th>
<th>Hrs</th>
<th>Matrix of suffixes S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>c</td>
<td>a</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>a#</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>a</td>
<td>a</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>abracadabra#</td>
</tr>
<tr>
<td>3</td>
<td>r</td>
<td>r</td>
<td>a</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>aca#</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>a</td>
<td>b</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>braca#</td>
</tr>
<tr>
<td>5</td>
<td>c</td>
<td>a</td>
<td>c</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>ca#</td>
</tr>
<tr>
<td>6</td>
<td>a</td>
<td>b</td>
<td>r</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>raca#</td>
</tr>
</tbody>
</table>

3.2 Fast $q$-gram generation

Permissible $q$-grams.

The key to a faster approach is to generate only the $q$-grams that are necessary, using $F$ and the auxiliary arrays. We call these $q$-grams that are necessary the permissible $q$-grams — they are the only $q$-grams that are permissible given $u,m$, and the fact that matching cannot proceed beyond the last characters in $T$ and $P$.

Further, if we wish to perform exact-pattern matching for a pattern $P$, where $|P| = m$, all we need will be the $m$-length $q$-grams (i.e. the $m$-grams) in the text $T$. The $m$-length $q$-grams (and excluding the $q$-gram from the rotations of the text) are the permissible $q$-grams. In general, we have a total of $u-q+1$ permissible $q$-grams for a $u$-length text. The major problems are to find cheap ways to generate the $q$-grams from $T$, and then how to perform the intersection quickly.

Fast $q$-gram generation. $Hr$ (also $H$) represents a one-to-one mapping between $F$ and $T$. By simply using $F$ and $Hr$, we can access any character in the text string, without using $T$ itself — which is not available without complete decompression. Therefore, there is a simple algorithm to generate the $q$-grams, for any given $q$: $\forall x=1,2,...,u-q+1, Q^T_q[x] = F[Hr[x]] \cdots F[Hr[x+q-1]]$;

These $q$-grams are not sorted. However, we can obtain the sorted $q$-grams directly by picking out the $x$’s according to their order in $Hr$, and then use $F$ to locate them in $T$. The index vector $Hrs$ provides information about the ordering, and is defined such that: $\forall i, F[i] = T[Hrs[i]] \in F[Hr[Hrs[i]]]$. This means that we have effectively generated the $q$-grams in constant time. Since the availability of $F$, $Hr$ and $Hrs$ implies constant-time access to any area in the text string $T$. The $x$ used in the previous description is simply an index on the elements of $T$.

3.3 Fast $q$-gram intersection

Based on the nature of the different arrays used by the BWT, and the new auxiliary arrays previously described, algorithms have been proposed to perform fast $q$-gram intersection (Adjeroh et al., 2002). Let $MQ_q = Q^T_q \cap Q^F_q$. We call $MQ_q$, the set of matching $q$-grams. For each $q$-gram, we use indexing on $F$, $Hr$ and $Hrs$ to pick up the required areas in $T$, and then match the patterns. To compute $MQ_q$, we need to search for the occurrence of each member of $Q^T_q$ in $Q^F_q$. 

4
For the special case when $m = q$ (i.e. exact pattern matching), (Adjeroh et al., 2002; Bell et al., 2002) proposed $O(m \log \frac{B}{m})$ to perform the required set intersection and thus locate the exact matches of $P$ in $T$, after an $O(u)$ preprocessing of $Z$, the BWT transformed output $T$.

4 Locating $k$-mismatches

We propose a $k$-mismatch algorithm based on the fast $q$-gram generation algorithm. The pattern matching operation is performed with all possible alignment of the pattern with the text indirectly via the matrix of sorted suffixes specified by the vector $Hrs$. The characters of the pattern $P$ are compared with characters in successive columns of the permissible $q$-grams which, in fact, is the suffix matrix $S$ if they exist. If there is a mismatch between the characters at corresponding locations of $P$ and $T$ for a given row in $S$, the number of the errors or mismatch is incremented by 1 without considering the possibility of inserting or deleting a character in the text to make a match. Since the $S$ matrix is lexicographically sorted, the match or mismatch takes place with the entire group of consecutively located rows in $S$. We record the number of mismatches ($count$) for the group as well as the start(st) and end (ed) positions for the group in the form of a triplet ($st, ed, count$). We place the triplet in an output list called Candidates. If $count$ is still less than $k$, we will continue to search in the group. If in a given row, the suffix length becomes less than the pattern length, it means that for this alignment of the pattern, the text has run out of characters. So, for each additional operation a mismatch count in its row triplet has to be added by one as long as $count \leq k$.

The operation proceeds until the last character of $P$ is processed yielding a final partition of the suffixes in $S$ having maximum of $k$ mismatch with $P$. This will correspond to the set of triplets that survive with less than or equal to $k$ mismatches. The triplets remaining in Candidates at this point correspond to positions with a maximum of $k$-mismatches to the pattern. The algorithm is given below.

$k$-mismatch Algorithm

Input: pattern $P$, the arrays $F, C$, auxiliary arrays $Hr, Hrs$, parameter $k$

Output: Candidates, the set of surviving triplets

1. Initialize Candidates:
   for each symbol $\sigma_c \in \Sigma$, (the $c$-the symbol in $\Sigma$), and $\sigma_c \in F$ do
     create a triplet with
     
     \[
     st = C[c] + 1;
     \]
     if $c < |\Sigma|$ then $ed = C[c] + 1$ else $ed = u$;
     if $\sigma_c = P[1]$ then $count = 0$ else $count = 1$;
     if $count \leq k$ then append triplet to Candidates.
   end

2. for $j = 2$ to $m$ do
   for each element in Candidates that survive the $(j-1)$-th iteration do
     \begin{itemize}
     \item Remove the triplet $(st, ed, count)$ from Candidates
     \item for each distinct symbol $\sigma_c \in F$ do
     \end{itemize}

     - locate the start and end position
       $st'$ and $ed'$ in $F$ between $st$ and $ed$ using binary search on the $j$-th column of the suffix matrix $S$

     Note that we do not actually create the $j$-gram for each row in the $S$ matrix during binary search. Instead, given the row index $pos$ of a $(j-1)$-gram in $S$, the $j$-th symbol $s$ can be accessed in constant time as: $s = F[Hr[Hrs[pos] + j]]$.

     - if $\sigma_c = P[j]$, then add triplet $(st', ed', count)$ to Candidates. (Since it is a match to $P[j]$, there is no change to $count$).
     - else, if $count + 1 \leq k$, then add triplet $(st', ed', count + 1)$ to Candidate. Increment $count$ by 1.

5
3. Report the $m$-length patterns between $st$ and $ed$ for each element of $Candidates$ as the $k$-mismatches. The row positions in $F$ can be converted to the corresponding positions in $T$ using $Hrs$, as explained earlier.

We demonstrate the algorithm with an example: $P=\text{"ssis"}, T=\text{"mississippi"}, k=2$.

$$
\begin{array}{cccc}
F & FS & FST & FSTT \\
1 & i & i & X \\
2 & (1,4,1) & ip (2,3,2) & ipp & X \\
3 & is (3,4,1) & iss (3,4,2) & issi & X \\
4 & i & is & issi \\
5 & m (5,5,1) & mi (5,5,2) & mis & X \\
6 & p (6,7,1) & pi(6,6,2) & pi & X \\
7 & pp (7,8,2) & ppi (7,8,2) & ppi & X \\
8 & s & si (8,9,1) & sip (8,8,2) & sipp & X \\
9 & s(8,11,0) & si & sis (9,9,2) & siss (9,9,2) \\
10 & s & ss (10,11,0) & ssi (10,11,0) & ssip (10,11,1) \\
11 & s & ss ssi & sssi (10,11,0) & \\
\end{array}
$$

When the first character of the pattern ‘$s$’ is checked, the first column of q-gram is partitioned into four groups. These are represented by the four triplets recording the start and end position in the sorted matrix as well as the number of errors that occurred. For example, row 8 to row 11 has the character ‘$s$’ matching $P[1]$. This has triplet $(8,11,0)$. Row 1 to row 4 has the character ‘$i$’ that is a mismatch to $P[1]$, hence the triplet $(1,4,1)$. Then continue matching with the next character in the pattern by checking the second character of the bigrams in the text. The first row corresponds to the suffix starting from the last character of the text, thus the bigram beginning at this position is not a valid or permissible bigram. The match operation, therefore, has to be terminated. Terminated matches are indicated with ‘X’ in the above example. Then suffixes that belong to rows two to four are split into 2 groups, one with a mismatch giving the triplet $(2,3,2)$ and the other a match giving triplet $(3,4,1)$. Continuing with the third character of the pattern, the suffix in row two has a mismatch so that the error count is greater than $k$. The corresponding triplet is removed from the candidate set, denoted by ‘X’. This also happens for the suffix in row 5. Finally, suffixes in rows 9, 10 and 11 lead to pattern match with the error count 2, 1 and 0, respectively.

**Complexity Analysis.** The preprocessing cost for preparing the auxiliary arrays is $O(u)$. For each iteration of the innermost loop, binary search is used to locate all the groups with the same $j$-gram. At most $u$ groups will be generated. Thus each loop takes $O(u \log \frac{u}{m})$ time in the worst case. The maximum number of triplets that can be generated will be in $O(|\Sigma|^k)$. But this cannot be more than $u$, the text size. In practice, many groups or triplets are dropped because the error count becomes greater than $k$. The worst case time to search the whole pattern will be $O(\min\{m(m-k)|\Sigma|^k \log \frac{u}{m}, u \log \frac{u}{m^k}\})$ Figure 1 shows experimental results of the maximum and average number of triplets generated over the test corpus for different word lengths.

We observe that the maximum number of triplets remains relatively constant for different pattern lengths. The average number of triplets decreases with increasing pattern length, $m$. Both are, however, relatively small compared to the file size (1.13MB on average over the test corpus). Figure 2 shows a typical evolution of the number of triplets during a search for $k$-mismatches, (with $m = |P| = 7, u = 125KB$ and $k = 2, 3, 4$). The number of triplets approaches a maximum at the $k$-th iteration, then and drops quickly as more characters from $P$ are checked.
5 Locating \( k \)-approximate matches

There are two phases in our approach. In the first phase, we locate areas in the text that contain potential matches by performing some filtering operations using *appropriately sized q-grams*. In the second phase, we verify the results that are hypothesized by the filtering operations. The verification stage is generally slow, but usually, it will be performed on only a small proportion of the text. Thus, the overall performance depends critically on the number of hypothesis generated.

**Locating potential matches.** The first phase is based on a known fact in approximate pattern matching:

**Lemma 2: \( k \)-approximate match (Baeza-Yates and Perleberg, 1992)** 
*Given a text \( T \), a pattern \( P \), \( (m = |P|) \), and \( k \), for a \( k \)-approximate match of \( P \) to occur in \( T \), there must exist at least one \( r \)-length block of symbols in \( P \) that form an exact match to some \( r \)-length substring in \( T \), where \( r = \frac{m}{k+1} \).*

This is trivially the case for exact pattern matching, in which \( k = 0 \), and hence \( r = m \). With the lemma, we can perform the filtering phase in three steps: 1) Compute \( r \), the minimum block size for the \( q \)-grams. 2) Generate \( Q_T^r \) and \( Q_P^r \), the permissible \( r \)-grams from the text \( T \) and the pattern \( P \), respectively. 3) Perform \( q \)-gram intersection of \( Q_T^r \) and \( Q_P^r \).
Let $M_k = \Omega^F \cap \Omega^T$, and $\eta_k = |M_k|$. Let $M_k[j]$ be the $i$-th matching $q$-gram, Let $M_k[j]$ be the $j$-th character in $M_k[j]$, $j = 1,2,\ldots, r$. Further, let $\bar{M}_k[i]$ be the index of the first character of $M_k[i]$ in the array of first characters, $F$. That is, $\bar{M}_k[i] = x$, if $F[x] = M_k[i][1]$. We call $M_k$ the matching $q$-grams at $k$. Its size is an important parameter for the next phase.

Based on the discussion in 3.2, the indices $j$ and $\bar{M}_k[i]$ can be generated in $O(1)$ time. Similarly, step 2 above can be done in constant time and space. The cost of step 3, will grow slower than $\frac{m^2}{k^2} \log u$. The time for hypothesis generation is simply the time needed for $q$-gram intersection, where $q$ is given by Lemma 2. Plugging these into the analysis for $q$-gram algorithm (Adjeroh et al., 2002), we have the following:

**Lemma 3: Locating potential $k$-approximate matches.** Given $T = t_1 t_2 \ldots t_u$, transformed with the BWT, $P = p_1 p_2 \ldots p_m$, an alphabet $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_{|P|}\}$, and the arrays $F$, $H$ and $I$. Let $k$ be given. The hypothesis phase can be performed in time proportional to: $Z_d \log |\Sigma| + \frac{m}{k+1} \sum_{i=1}^{|I|} Z_i^2 \log Z_i^2$.

**Verifying the matches.** Here, we need to verify if the $r$-blocks that were hypothesized in the first phase are true matches. The time required will depend critically on $\eta_k$. We perform the verification in two steps: 1) Using $Hr$ and $F$ determine the matching neighborhood in $T$ for each $r$-block in $M_k$. The maximum size of the neighborhood will be $m + 2k$; 2) Verify if there is a $k$-approximate match within the neighborhood.

Let $N_k$ be the neighborhood in $T$ for $\Omega_k[i]$, the $i$-th matching $q$-gram. Let $t$ be the position in $T$ where $\Omega_k[i]$ starts. That is, $t = Hr[F[\bar{M}_k[i]]]$. The neighborhood is defined by the left and right limits: $t_{left}$ and $t_{right}$ viz: $t_{left} = t - k$ if $t - k \geq 1$; $t_{left} = 1$ otherwise; $t_{right} = t + m + k$ if $t + m + k \leq u$; $t_{right} = u$ otherwise. Hence, the $i$-th matching neighborhood in $T$ is given by: $N_i = [t_{left} \ldots t_{right}]$. Thus, $|N_i| \leq m + 2k$, $\forall i = 1, 2, \ldots, \eta_k$. We then obtain a set of matching neighborhoods $\mathcal{S}_M \Omega_k = \{N_1, N_2, \ldots, N_{\eta_k}\}$. Verifying a match within any given $N_i$ can be done with any fast algorithm for $k$-approximate matching, for instance, Ukkonen’s $O(ku)$ algorithm (Ukkonen, 1985). The cost of the first step in the verification will be in $O(\eta_k)$. The cost of the second step will thus be in $O(\eta_k k(m + 2k)) \leq O(\eta_k k(3m)) \approx O(\eta_k km)$.

**Example.** Let $T = abraca$ and $P = brace$, with $k = 1$. Then, $r = 2$, and permissible $q$-grams will be $\Omega^P = \{ac, br, ce, ra\}$, $\Omega^T = \{ab, ac, br, ca, ra\}$, yielding $M_1 = \{ac, br, ra\}$, and $N_1 = [3 \ldots 6]; N_2 = [1, \ldots, 6]; N_3 = [2, \ldots, 6]$, Matches will be found in $N_1$ and $N_2$ at positions 1 and 2 in $T$, respectively.

We conclude this section with the following theorem:

**Theorem 2: $k$-approximate matching in BWT-compressed text.** Given a text string $T = t_1 t_2 \ldots t_u$, a pattern $P = p_1 p_2 \ldots p_m$, and symbol alphabet $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_{|P|}\}$. Let $Z$ be the BWT output when $T$ is the input. There is an algorithm that can locate the $k$-approximate matches of $P$ in $T$, using only $Z$, (i.e. without full decomposition nor with off-line index structures) in $O(|\Sigma| \log |\Sigma| + \frac{n^2}{k} \log \frac{n^2}{k} + k|\psi_{\Omega_k}|)$ time on average, ($|\psi_{\Omega_k}| \leq u$), and in $O(ku + |\Sigma| \log |\Sigma| + \frac{n^2}{k^2} \log \frac{n^2}{k})$ worst case, after an $O(u)$ preprocessing on $Z$.

The proof is based on using the $O(m \log \frac{n^2}{k})$ BWT-based search algorithm (Adjeroh et al., 2002; Bell et al., 2002) to match all the $m - r + 1$ $r$-grams that will be generated, where $r = \lfloor \frac{u}{k+1} \rfloor$. We omit the detailed proof for brevity. With the $q$-gram approach, we can treat exact pattern matching as no different from $k$-approximate pattern matching. We just have $k=0$, and hence no verification stage.

6 Results

To test the proposed methods for compressed approximate pattern matching, we selected 133 files from three different corpora. The files included the text files in the Canterbury corpus, html and C program files from the Calgary corpus, and the AP, DOE and FR files from Disk 1 in the TIPSTER corpus. The file sizes ranged from 11,150 characters (fields.c) to 4,161,115 characters (FR89011 in TIPSTER). The average file size was 1.13Mbytes.
Experiments were performed using 10 sets of sample patterns (words). Each set has 100 words with the same length, \( m = |P| = 2, 3, \ldots, 11 \). The words are the most frequently used from the English dictionary. The tests were run on Sun Ultra-5 workstation (200 MHz) and Solaris 2.5 Operating System. The overhead of inverse BWT and auxiliary arrays was 1.03 seconds per file, on average. Time is measured in seconds in all reported results.

We compare the proposed compressed domain \( k \)-mismatch algorithm with an algorithm described in Gusfield’s book (Gusfield, 1997) which is based on suffix trees. Here, the \( k \)-mismatch check at any position \( i \) in \( T \) is performed using at most \( k \) longest common extension computation. Each computation can be performed in constant time, after the longest common ancestor table has been constructed. The suffix tree, however, usually requires a large storage, (about 21\( u \) bytes), although the construction is in \( O(u) \) time. The search time used by the two algorithms are shown in Figure 3 (left column).

For \( k \)-approximate matches, we tested the proposed method with two popular approximate pattern matching algorithms: AGREP (Wu and Manber, 1992), and NRGREP (Navarro, 2001b). Both algorithms are based on bit-wise operations using the patterns and text. The two algorithms operate on the raw (uncompressed) text. The results for the search time are shown in Figure 3, right column. Results for the proposed BWT-based approach is labeled BWT-\( \text{DFA} \). Here, the hypothesis phase is performed by finding the exact matches for the \( r \)-grams using the QGRAM algorithm (Adjeroh et al., 2002), while the verification phase is performed using Ukkonen’s DFA (Ukkonen, 1985). We have also included two other results: BWT-AGREP and BWT-NRGREP. These correspond to when we used the proposed \( q \)-gram filtering approach, but with the verification phase performed with AGREP and NRGREP respectively.

The construction time for Ukkonen’s DFA is shown in Table 1.

AGREP performed constantly better than the NRGREP, albeit marginally. The original AGREP and NRGREP report only one pattern occurrence for each line in the text. The QGRAM algorithm however reports all the occurrences of the pattern in a text. For a fair comparison, the QGRAM algorithm was modified to report only one occurrence of the pattern. Thus, the actual time to find all the occurrences of the patterns would be slightly larger than the results reported.

We observe that, the search time for AGREP and NRGREP increased quite significantly as \( k \) increased. But the proposed methods only show little or no changes. The search time for BWT-\( \text{DFA} \) is almost constant once the DFA is built. Also, using BWT-AGREP and BWT-NRGREP produced shorter search times than the traditional AGREP and NRGREP algorithms, which perform matches on the de-compressed text. Their reported search time does not include the time required for decompression.

### Table 1: Construction time for Ukkonen’s DFA

<table>
<thead>
<tr>
<th>( k )</th>
<th>( L=1 )</th>
<th>( L=2 )</th>
<th>( L=3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0006</td>
<td>0.0008</td>
<td>0.0009</td>
</tr>
<tr>
<td>3</td>
<td>0.0008</td>
<td>0.0009</td>
<td>0.0011</td>
</tr>
<tr>
<td>4</td>
<td>0.0009</td>
<td>0.0011</td>
<td>0.0014</td>
</tr>
<tr>
<td>5</td>
<td>0.0011</td>
<td>0.0014</td>
<td>0.0018</td>
</tr>
<tr>
<td>6</td>
<td>0.0013</td>
<td>0.0016</td>
<td>0.0021</td>
</tr>
<tr>
<td>7</td>
<td>0.0015</td>
<td>0.0018</td>
<td>0.0024</td>
</tr>
<tr>
<td>8</td>
<td>0.0017</td>
<td>0.0020</td>
<td>0.0027</td>
</tr>
<tr>
<td>9</td>
<td>0.0019</td>
<td>0.0022</td>
<td>0.0029</td>
</tr>
<tr>
<td>10</td>
<td>0.0021</td>
<td>0.0025</td>
<td>0.0033</td>
</tr>
<tr>
<td>11</td>
<td>0.0023</td>
<td>0.0027</td>
<td>0.0035</td>
</tr>
<tr>
<td>12</td>
<td>0.0025</td>
<td>0.0030</td>
<td>0.0039</td>
</tr>
<tr>
<td>13</td>
<td>0.0027</td>
<td>0.0032</td>
<td>0.0042</td>
</tr>
<tr>
<td>14</td>
<td>0.0029</td>
<td>0.0034</td>
<td>0.0045</td>
</tr>
<tr>
<td>15</td>
<td>0.0031</td>
<td>0.0037</td>
<td>0.0049</td>
</tr>
<tr>
<td>16</td>
<td>0.0033</td>
<td>0.0040</td>
<td>0.0054</td>
</tr>
<tr>
<td>17</td>
<td>0.0035</td>
<td>0.0042</td>
<td>0.0060</td>
</tr>
<tr>
<td>18</td>
<td>0.0037</td>
<td>0.0045</td>
<td>0.0071</td>
</tr>
</tbody>
</table>

### 7 Conclusion

Although the performance of the BWT has made it an important addition to the long list of text compression algorithms, very little has been reported in searching directly on text compressed with the BWT. The BWT with its sorted contexts however provides an ideal platform for compressed domain pattern matching. This paper has described algorithms for approximate pattern matching directly on BWT-transformed text.

The proposed algorithms could be further improved. For instance, the space requirement could be...
reduced by considering the compression blocks typically used in BWT-based compression schemes, while the time requirement could be further reduced by using faster pattern matching techniques for the q-gram intersection. One way to reduce the relatively high overheads will be to consider pattern matching on blocked BWT-compressed files, since these will typically involve smaller text sizes.

The overhead of DFA construction is rather high compared to the search time. But once a DFA is constructed for a pattern, it can be used to find any other pattern with edit distance less than or equal to $k$ without changing the DFA. The proposed approach will thus be very effective in dictionary matching, where one may wish to search for a pattern in multiple text sequences. One possible improvement on the hypothesis verification phase could be to use dynamic construction of the DFA. The basic concept in the method does not depend on any particular verification algorithm. This could also be seen from the results with BWT-AGREP and BWT-NRGREP. Thus, one could abandon Ukkonen’s DFA altogether, and look for alternative faster verification algorithms.

We note that the methods as described basically operate on the output of the BWT transformation stage. One challenge is to extend the approach to operate beyond the BWT output, i.e. after the later encoding stages in the BWT compression process.

References


Figure 3: Results for k-mismatches (left column) and k-approximate match (right column)