A LINEAR ALGORITHM FOR FINDING REPETITIONS
AND ITS APPLICATIONS IN DATA COMPRESSION

by

M. Rodeh*, V. R. Pratt** and S. Even*

Technical Report #72
April 1976

* Dept. of Computer Science, Technion, Haifa, Israel.

** Dept. of Electrical Engineering and Computer Science, Mass. Inst. of Tech.
Cambridge, Mass. 02139.
ABSTRACT

A linear algorithm for finding the longest repetition for each position in a given string is developed. It is related to Weiner's string processing algorithm [1] but simpler. The algorithm is applied to the problem of evaluating the complexity of a finite string as defined by Lempel and Ziv [2]. Using Elias' representation of the integers [4, 5] we produce an asymptotically optimal variable to variable coding scheme for sequential data compression based on the definition of complexity. The scheme requires unbounded memory for strings whose length grows to infinity. To overcome this difficulty Ziv and Lempel [3] invented a variable to block coding scheme that is asymptotically optimal for all ergodic sources with a given finite entropy. We implement their scheme by combining our repetition finder with Weiner's suffix tree construction algorithm. Other related and more practical schemes are described too.
0.1. Introduction.

This paper comprises 3 sections. In Section 1 we throw new light on Welner's string processing algorithm [1] by developing a related algorithm for finding repetitions in a string. In Section 2 we adapt it to constitute a linear implementation of a variable-to-variable data compression scheme based on Lempel and Ziv's [2] definition of the complexity of a finite sequence; this scheme uses unbounded memory. A linear implementation of Ziv and Lempel's [3] variable-to-block coding scheme with bounded memory is described in Section 3; it is a combination of our repetition finder and Welner's algorithm [1] for suffix tree construction.

0.2. Notation.

Let $\Sigma$ be a finite alphabet and $\Sigma^*$ the set of all finite length strings over $\Sigma$. We use the letters a, b, c, ... for elements of $\Sigma$ and $\alpha, \beta, \gamma, ...$ for strings. We let $\varepsilon$ denote the null string, $\ell(\alpha)$ the length of $\alpha$, $\alpha[i]$ the $i$-th letter of $\alpha$, and $\alpha\beta$ the concatenation of $\alpha$ and $\beta$.

Let position in a string denote "between-letter" rather than "at-letter" position. Thus $\alpha$ has $\ell(\alpha)+1$ positions numbered 0, 1, ..., $\ell(\alpha)$. Two positions $i$ and $j$ of $\alpha$ define the subword $\alpha[i+1]...\alpha[j-1]\alpha[j]$ which occurs at $j$ and is denoted by $\alpha[i,j]$;
If \( l \leq j \) then \( \alpha_{l,j} \) is defined to be \( \varepsilon \). \( \beta = \alpha_{0,1} \) for \( 0 \leq l \leq L(\alpha) \) is called a prefix of \( \alpha \); \( \gamma = \alpha_{1,l} \) is called a suffix of \( \alpha \).

A repetition at \( j \) is an occurrence at \( j \) of a subword that also occurs at \( l < j \); thus, every occurrence but the first of a subword is a repetition.

An occurrence of \( \alpha \alpha \) at \( l \) is the left neighbor of the occurrence of \( \alpha \) at \( l \), and the occurrence of \( \alpha \beta \) at \( l+1 \) is the right neighbor.

1. Improvements to Weiner's Algorithm.

1.1. Introduction.

In this section we develop an algorithm for finding repetitions within a string. It is related to Weiner's algorithm but is simpler. We also show that the running time of the algorithm on a RAM can be made independent of the alphabet size. To this end we establish properties of strings which prove the correctness of the algorithm.

1.2. Weiner's Suffix Identifier Finder.

Let \( \alpha \) be a fixed string of length \( n \). The suffix identifier of position \( l \) is the shortest (unique) subword which occurs only at \( l \). To make this definition meaningful we have to add a new alphabet letter \( $ \) to the left of \( \alpha \). Weiner's algorithm
finds all suffix identifiers in $O(n)$ time. (Weiner actually considered prefix identifiers, but the dual case is closer to our algorithm.)

We are interested in repetitions, which is a related concept. Out of the $(n+1)+1$ subwords of $\alpha$, $O(n^2)$ different subwords may be repetitions. Thus in aiming at a linear repetition-finding algorithm we cannot afford to identify every repetition explicitly. The algorithm finds only the longest repetition at each position.

Observe that the longest repetition at position $i$ depends only on $\alpha[i]$ and on nothing to the right of $i$. Thus in the production of longest repetitions in a left-to-right scan, we never have to change the repetitions already found. This is a significant property since in our application we are interested in the scanning process itself and not in the final result. This is in contrast to the situation with the suffix identifiers which depend on the entire $\alpha$. Thus in Weiner's algorithm when scanning $\alpha$ from left-to-right, some subword which occurs at $j>i$ might cause an update operation which includes the suffix identifier for $i$, though this does not compromise its $O(n)$ time bound.

The suffix identifier for $i$ is longer than the longest repetition at $i$. However, scanning the string left-to-right, when we reach position $i$, they differ only by the leftmost letter of the suffix identifier which does not appear in the
longest repetition.

It is worth noting here that neither Weiner's nor our algorithm proceeds from position $i$ to position $i+1$ at constant time although both operate within a time bound linear in $n$.

1.3. An Algorithm for Finding Repetitions.

As pointed out above, we are interested in longest repetitions only. For example, in `abcabebe`, the longest
repetitions at each position 0 through 8 respectively are
ε,ε,ε,ε,a,ab,abc,b,bc. We shall always use \textit{longest repetition}
to mean "longest repetition at a given position".

Notice that in this example later longest repetitions do
not start in earlier positions. This is always true:

**Theorem 1**: Let \( \beta, \gamma \) be longest repetitions at \( l,j \) respectively.
If \( l < j \) (i.e. If \( \beta \) ends before \( \gamma \)) then \( \gamma \) does not start before \( \beta \).

**Proof**: Any prefix of this occurrence of \( \gamma \) must be a repetition.
Thus if \( \gamma \) started before \( \beta \), \( \beta \) could not be a longest
repetition.

This theorem suggests the basic algorithm, which simply
moves two pointers (i.e. increments two counters) through \( \alpha \). We
represent the pointers as brackets and call the in-between
subword \textit{bracketed}.

**Algorithm 1.1**: \texttt{start with} \([\alpha]\);
\texttt{while the picture is} \( \ldots[\beta]\alpha\ldots \) \texttt{do}
\texttt{if} \( \beta \alpha \) \texttt{is not a repetition} \texttt{then move} [\texttt{else move} ].

"move [" and "move ]" mean move the indicated bracket one
position to the right. If \( \varepsilon \) is bracketed, "move [" also implies
"move ]" to avoid having ":[" overtake "]".

Notice that when "]" first arrives at a position, the
bracketed subword is the longest repetition for that position.

One question remains: How do we tell efficiently whether $\beta a$ is a repetition?

1.3. Testing for repetitions.

A remarkable property of longest repetitions supplies the key to testing whether $\beta a$ is a repetition.

**Lemma 1:** If no occurrence of $\beta$ in a string is a longest repetition then every occurrence of $\beta$ has the same left neighbor.

**Proof:** We use induction on the number of occurrences of $\beta$ in the string. The Lemma is vacuously true for zero or one occurrence. If there are two or more, the rightmost occurrence must have the same left neighbor as some other occurrence of $\beta$, since it is not a longest repetition. By induction all occurrences have the same left neighbor. \(\square\)

Putting the Lemma in its contrapositive form we have:

**Lemma 1'** Let $\beta$ occur at 1 and at $j$, $1 < j$, not with the same left neighbor. Then there exists $k < j$ such that $\beta$ is the longest repetition at $k$.

**Proof:** Apply Lemma 1 to the first $j$ characters of the string. \(\square\)

**Theorem 2:** If $\beta a$ is a longest repetition at $j$, there is some $k < j$ such that $\beta$ is the longest repetition at $k$.
Proof: Let the left neighbor of the occurrence of $\beta a$ at $j$ be $b\beta a$. There must be an earlier occurrence of $\beta a$ (since the one at $j$ is a repetition) whose left neighbor is not $b\beta a$ (since the occurrence of $\beta a$ at $j$ is a longest repetition). Then $\beta$ satisfies the conditions of Lemma 1', yielding the desired $k$. □

In order to test whether $\beta a$ is a repetition (and therefore a longest repetition by the correctness of Algorithm 1.1), we ask whether $\beta$ is a previously recorded longest repetition, and if so whether some earlier occurrence of $\beta$ has right neighbor $\beta a$. To answer the first question we keep a record of longest repetitions seen so far, represented as vertices of a graph $G$. Each vertex corresponds to a distinct element $\beta$ of $\Sigma^*$, and so represents all repetitions of a given subword $\beta$. In our discussion we refer to a vertex by underlining the string which it represents. To answer the second question, we associate with each vertex $\beta$ a record for each letter $a$ of whether $\beta a$ has occurred, and if so where it first occurred. We let $\beta.a$ name a register associated with $\beta$, which we may inspect and update. If $\beta.a$ is undefined, $\beta a$ has not occurred before; if defined, $\beta a$ first occurs at $\beta.a$. (The single dot suggests the first occurrence.) There are $|\Sigma|$ such registers associated with each vertex, all initially undefined.

We postpone for the moment the question of how we keep this information about right neighbors up to date, as it depends on the details of how we navigate in and update the
In order to locate $\beta$ when asking whether $\beta a$ has occurred before, we shall maintain a pointer to the vertex in $G$ representing the currently bracketed subword. Clearly we need enough information in $G$ to allow us to trace a trajectory through $G$ corresponding to the sequence of bracketed words resulting from moving the brackets.

By Theorem 2, Algorithm 1.1 must keep moving "[" until some word is bracketed for which there is a vertex in $G$. Hence for each vertex $\beta \in G$, all we need is an edge (which we label $S$ for suffix) from $\beta$ to the vertex which represents the longest proper suffix of $\beta$ in $G$. We denote by $S(\beta)$ a register associated with $\beta$ containing this edge (i.e. pointer). In this way, we may jump directly from $\beta$ to $S(\beta)$, following possibly several moves of "[".

When we move "]", we go from $\beta$ to $\beta a$. If $\beta a$ is already in $G$, an edge (labelled :a) from $\beta$ to $\beta a$ suffices; for each letter a we associate with $\beta$ a register named $\beta : a$ to contain $\beta a$. (The two dots suggests that $\beta a$ has occurred at least twice). If $\beta a$ is not yet in $G$, we must create it. This is sufficiently complicated to warrant its own section.

Except for the details of vertex creation, and of keeping $\beta : a$ up to date for all $a$ and for all $\beta$, we may now flesh out Algorithm 1.1 to yield Algorithm 1.2.

In the following, $\beta$ denotes the currently bracketed
subword and its position (thus \( \text{I} \) represents "["]. The function \text{GENERATE} returns a new vertex, with all associated registers undefined. Note that we do not have to represent "[" explicitly as \( \alpha \) keeps track of it implicitly.

**Algorithm 1.2:**

Initialize \( \alpha \rightarrow \text{GENERATE}; \)
\[ \alpha \rightarrow \epsilon; \ \text{I+1}; \]
Iterate \( \text{while I}\neq \text{n do} \)
\[ \begin{align*}
\text{begin} & \ a+a<1>; \\
\text{Repetition?} & \text{ if } \alpha.a \text{ undefined then} \\
\text{Move} & \text{ if } \alpha=\epsilon \text{ then } \alpha+S(\alpha) \\
& \text{ or } [ ] \text{ else I+1+1} \\
\text{Create?} & \text{ else begin if } \alpha.a \text{ undefined then } \text{CREATEVERTEX}; \\
& \alpha+\beta:a; \ I+1+1 \\
\text{end} \\
\text{end}; \\
\end{align*} \]

We turn now to vertex creation, and postpone the details of updating right neighbors for the moment.

1.5. Creating Vertices.

Merely generating a new vertex \( \beta a \) is easy, as is updating \( \beta:a \). It is more difficult to correctly update the \( S \) edges leading to and from \( \beta a \). By definition, only one edge leads from \( \beta a \). Fortunately, when it is created, at most one edge leads to
Theorem 3: Given a and \( \tau \), there exists at most one subword \( \pi \) such that \( \pi \alpha \tau \) is in \( G \) and \( S(\pi \alpha \tau) = \tau \).

Proof: Let \( \pi \alpha \tau \) and \( \pi' \alpha \tau \) be two such distinct subwords, both represented in \( G \). Neither may be a suffix of the other, since \( S(\pi \alpha \tau) = S(\pi' \alpha \tau) = \tau \). Hence \( \pi \alpha \tau \) and \( \pi' \alpha \tau \) must be of the forms \( \eta \rho \pi \alpha \tau \) and \( \eta' \rho' \pi \alpha \tau \) (\( f \neq g \)) respectively where any of \( \eta, \eta', \rho \) may be null. Thus by Lemma 1, \( \pi \alpha \tau \) occurs as a longest repetition no later than the later occurrence of \( \pi \alpha \tau \) and \( \pi' \alpha \tau \), contradicting \( S(\eta \rho \pi \alpha \tau) = \tau \).

Corollary 3.1: For each letter \( c \) and vertex \( \beta \) we may introduce an edge (register) labelled \( *c: \) from \( \beta \) to the (necessarily unique) vertex \( \chi \) of the form \( \psi c \beta \) such that \( S(\chi) = \beta \) (The "*" denotes the \( \psi \) in \( \psi c \beta \)). Thus \( S(*c: \beta) = \beta \), provided \( *c: \beta \) is defined.

Corollary 3.2: When creating a vertex \( \beta \alpha \), at most one vertex \( \chi \) exists such that \( S(\chi) = \beta \alpha \).

Proof: By Lemma 1 all previous occurrences of \( \beta \alpha \) must have the same left neighbor \( b \beta \alpha \), if any. Hence, if \( \beta \alpha \) were a suffix of a bracketed subword, it should have \( b \beta \alpha \) as left neighbor and by Theorem 3, there exists at most one vertex \( \chi \) such that \( S(\chi) = \beta \alpha \).

Hence our problem is reduced to that of finding the \( \chi \) which is \( S(\beta \alpha) \), and the \( \delta \), if any, for which \( S(\delta) = \beta \alpha \), and
updating all the edges between $\delta$, $\beta a$ and $y$.

Either $y = \varepsilon$ or $y = y'a$, for some $y'$. If the latter, then by Theorem 2, since $y$ is in $G$, so is $y'$. To find $y'$ we simply follow along $S$ edges, starting from $\delta$, until we find a vertex $\pi$ for which $\pi : a$ is defined; $\pi : a$ is the required $y$. If there is no such $\pi$ (the search ends with $\pi = \varepsilon$ and $\varepsilon : a$ is undefined) then we are in the first case and $y = \varepsilon$.

To find $\delta$, note that before $\beta a$ is created, $S(\delta) = y$. If $\beta a = n c y$ for some $n$ and $c$, then $\delta = *c : y$. To see that $*c : y$ must yield a $\delta$ of which $\beta a$ is a suffix, note that if it didn't, then $*c : y$ would have to point to both $\beta a$ and $\delta$, contrary to Corollary 3.1.

So we wish to compute $c$. In the absence of a representation of $c$ in $G$, we look in $a$. For this purpose we introduce two registers, for each vertex $\delta$: loc($\delta$) gives the location of the first occurrence of $\beta$ and len($\beta$) holds the length of $\beta$.

Now we can easily find $c$. Let $j = loc(\beta a) - len(y)$; then $c = a < j >$. When setting $*d : \beta a$ to $\delta$ we find the $d$ in the same way except that $j = loc(\delta) - len(\beta a)$, $d = a < j >$.

Let us review all the registers associated with each vertex $\delta$ used in $G$.

$\delta : a = \text{the position of the earliest occurrence of } \beta a$.

$\delta : a = \text{a pointer to the vertex } \beta a \text{ in } G$.

$S(\delta) = \text{a pointer to the longest proper suffix of } \beta$. 
represented in G.

*a:* a pointer to the \( \delta \) for which \( \delta = \pi \alpha \beta \) and \( S(\delta) = \beta \).

\( \text{loc}(\beta) \) = the position of the first occurrence of \( \beta \).

\( \text{len}(\beta) \) = the length of \( \beta \).

The remaining question is that of updating right neighbors of subwords represented in G.

### 1.6. Updating Right Neighbors.

For the algorithm's correctness, it is enough to show a way to update \( \beta \)'s right neighbors no later than the first occurrence of \([\beta]a\). We keep \( \beta \)'s right neighbors up to date from the moment of creating \( \beta \). This is useful in certain applications.

First assume that the earliest occurrence of \( \beta a \) is later than the earliest bracketed occurrence of \( \beta \). Then when \( \) first arrives between \( \beta \) and \( a \), \( \) must be to the left of this occurrence of \( \beta \). Now \( \) may not move while \( \) is to the left of \( \beta \) since this is the first occurrence of \( \beta a \). But in the process of moving to the right, \( \) must come to bracket \( \beta \) since \( \beta \) is already in G. At this time we are in a position to define \( \beta.a \), which can be done with \( \beta.a + 1 \). By only performing this assignment when \( \beta.a \) is undefined, we avoid erroneously redefining \( \beta.a \) at a non-initial occurrence of \( \beta a \).

Now assume that the first occurrence of \( \beta a \) is not later
than the first bracketed occurrence of $\beta$. We have to show that when $\beta$ is created we can immediately record the position of the first $\beta$.

Two cases arise:

(I) The second occurrence of $\beta$ is bracketed. In this case, only one right neighbor is involved, namely that of the first occurrence of $\beta$. Thus when $\beta$ is first bracketed, $\text{loc}(\beta)$ may be used to determine this right neighbor. In this case there exists no $\delta$ such that $S(\delta) = \beta$, since the presence of such a $\delta$ implies at least two earlier occurrences of $\beta$.

(II) The third or later occurrence of $\beta$ is the first to be bracketed. In this case, there must be some $\delta$ such that $S(\delta) = \beta$, because the second occurrence of $\beta$ can avoid being bracketed only by not being the longest repetition at that position. In this case, there may be several right neighbors to be recorded; however, every previous occurrence of $\beta$ must be at the same position as some occurrence of $\delta$, and since all of $\delta$'s right neighbors are up to date (by Induction), we can simply copy their positions, which will then serve as the positions of the corresponding right neighbors of $\beta$.

Now we collect all the pieces to yield the following final algorithm.
1.7. The Algorithm in Full Detail.

Algorithm 1.3

Initialize $\varepsilon$+GENERATE;
len($\varepsilon$)+0;
loc($\varepsilon$)+0;
S($\varepsilon$)+$\varepsilon$;
$\beta$+$\varepsilon$;
l+1;

Iterate while l\leq n do
begin
a+a<l>;

Repetition? If $\beta$.a undefined then

New rt. nbr begin $\beta$.a+1;

Move [ if $\beta$.e then $\beta$.+$S(e)$
or [] else l+1+1

end

Create? else begin if $\beta$.a undefined then CREATEVERTEX;

Move ] $\beta$.+$\beta$.a;
l+1+1

end;

end;
CREATEVERTEX: \texttt{Ba+GENERATE;}

\texttt{len}(\texttt{Ba}) + len(\texttt{Ba}) + 1;
\texttt{loc}("Ba") + "Ba;"

\texttt{Link } \texttt{Ba, Ba} \quad \texttt{Ba} + \texttt{Ba;}

\texttt{Find y} \quad \texttt{y} = \texttt{y} + \texttt{S}(\texttt{y}); \quad \texttt{comment: we need } \texttt{S(e)} = \texttt{e} \texttt{ here;}
\quad \texttt{while } \texttt{y} = \texttt{a} \texttt{ undefined and } \texttt{y} = \texttt{e} \texttt{ do } \texttt{y} = \texttt{y} + \texttt{S}(\texttt{y});
\quad \texttt{y} = \texttt{if } \texttt{y} = \texttt{a} \texttt{ defined and } \texttt{y} = \texttt{e} \texttt{ then } \texttt{y} := \texttt{e} \texttt{ else } \texttt{e;}

\texttt{Find } \texttt{a} \quad \texttt{j} = \texttt{loc}(\texttt{Ba}) - \texttt{len}(\texttt{y}) - 1;
\quad \texttt{c} = \texttt{a}(\texttt{j});
\quad \texttt{a} = \texttt{c} + \texttt{y};

\texttt{Link } \texttt{Ba, y} \quad \texttt{S(Ba)} \rightarrow \texttt{y;}
\quad \texttt{*c} = \texttt{y} \rightarrow \texttt{Ba;}
\quad \texttt{if } \texttt{a} \texttt{ defined then begin}
\quad \texttt{Link } \texttt{a, Ba} \quad \texttt{j} = \texttt{loc}(\texttt{a}) - \texttt{len}(\texttt{Ba}) + 1;
\quad \texttt{d} = \texttt{a}(\texttt{j});
\quad \texttt{S(\texttt{d}) + Ba;}
\quad \texttt{*d} = \texttt{Ba} + \texttt{d}
\quad \texttt{end;}

\texttt{Rt. nbrs} \quad \texttt{if } \texttt{a} \texttt{ undefined then } \texttt{Ba, a} \rightarrow \texttt{Ba, a} + 1
\quad \texttt{else for each letter } \texttt{b} \texttt{ do}
\quad \texttt{Ba, b + \texttt{a, b}}
\quad \texttt{end;}
We wish to show that the running time of the algorithm is linear in $n=\ell(a)$, independently of the alphabet size. The argument proceeds in two stages: At first we assume that the alphabet is fixed and finite. Then we consider the influence of the alphabet size, $|\Sigma|$.

The main routine moves one or both brackets to the right at each iteration of the single loop. Hence the body of that loop may not be executed more than $2n$ times.

CREATEVERTEX contains two loops. The second, which transcribes right neighbors is executed $|\Sigma|$ times, thus constant. The first loop searches for $\gamma'$. The search is done by moving from $A$ along the $S$ arcs. We introduce a third pointer, in addition to the two brackets; it points to the starting position of the occurrence of $\gamma'$ at $I$, during this search. We claim that this pointer never moves left through $\alpha$, even between successive vertex creations, and moves right at every iteration of the search loop. The latter claim is non-controversial. To see the former, observe that when the search that started from $A$ terminates, a search that starts from $Ba$ must begin with the just-found $\gamma$. A similar observation applies even when the next search does not start from $Ba$. This establishes the linear time bound for a finite alphabet.

Now consider an unbounded alphabet. The right neighbor transcription loop then becomes a potential time-waster. We
shall show that the cost of copying all defined right neighbors but one can be charged to the account of moving "[" through $\alpha$. Let us assume that the algorithm is at position 1 and transcribes the right neighbors of $B \alpha$ from $\delta$. For every $c$ such that $\delta.c$ is defined, let $k(c)+1$ be the first occurrence of $\beta ac$. We divide the cost of copying $\delta$'s right neighbors into two cases:

(i) $c$ is the letter for which $k(c)$ is minimum, i.e. $k(c) = \text{loc}(B \alpha)$. The cost of transcribing $\delta.c$ is charged to $B \alpha$, and since the number of vertices is bounded by $n$, so is the total cost of these transcriptions.

(ii) For any other letter $d$, consider the moment when "[" is first in position $k(d)$. Since $\beta \alpha$, at that time, is already a repetition (but not a longest one) "[" must be to the left of $\beta \alpha$ (but not adjacent to it). However "[" cannot move to the right before "[" enters $\beta \alpha$ since $\beta \alpha$ is not a repetition at $k(d)+1$. Also, "[" cannot be located at $k(d)-\ell(\beta \alpha)$, since $B \alpha$ is not defined yet. Thus, "[" makes a jump over location $k(d)-\ell(\beta \alpha)$. We charge the transcription of $\delta.d$ (to vertex $B \alpha$) to the jump over this location.

To take advantage of this observation we must devise a method of ignoring undefined right neighbors. This is easily done by maintaining for each vertex $\alpha$ a list of letters such that for each letter $b$ in the list $\alpha.b$ is defined. It is easy
to see that a duplication can never exist in such a list. To access each letter in constant time we need an array of length \(|\Sigma|\) for each vertex \(A\) and each of the tables which contain \(A:a\), \(A:a\) and \(*a:B\). Thus, if we have a storage preset to zero, we achieve an independence of the alphabet size. Instead we can simulate the storage preset to zero in real time ([7], p.71, Ex.2.12). Unfortunately this can be very wasteful of space; in fact the factor of \(|\Sigma|\) has simply crossed over to space bound. In order to keep the space requirement independent of \(|\Sigma|\) we have to pay a time overhead of at most \(\log|\Sigma|\).
2.1. Introduction.

In this section we present a linear scheme for sequential data compression in two steps:

**Step 1:** We design an algorithm for translating strings to words over a certain infinite alphabet \( \Theta \). The translation method is based on Lempel and Ziv's definition of the complexity of finite strings [2]. The implementation uses the repetition finding algorithm. The handling of the data structures is irrelevant here. Therefore we shall refer to Algorithm 1.1 and not to the full detailed version. The linearity will come out as an immediate consequence of the timing considerations of the repetition finding algorithm.

**Step 2:** We use Elias' universal binary representation of the integers [4,5] to produce a binary encoding of strings over \( \Theta \). For a given integer \( m \) this representation can be built within a time bound linear in the length of the binary representation of \( m \).

The composition of the two steps yields a universal linear

*The generalization to an arbitrary finite alphabet is straightforward.
variable to variable encoding scheme for strings, whose compression ratio tends to be optimal, for ergodic sources, as the length of the input string grows to infinity. We shall prove this by referring to results of Lempel and Ziv [3]. The disadvantages of the scheme are that the memory requirements grow with the length of the input string and the rate of convergence to the optimal compression ratio is slow. We shall treat the memory problem in Section 3.

2.2. Translating strings to $\Theta^*$.

Let $\Theta = \{<i,j,a> | i,j$ are integers, $a \in \Sigma \}$. The letters of $\Theta$ are triples whose first two components are integers, and the third is a letter of $\Sigma$. A word over $\Theta$ is a sequence of such triples. Define the translation of $\alpha$ to $\Theta^*$, denoted by $T(\alpha)$, by the following algorithm:

Algorithm 2.1.

$p \leftarrow 0; \ k \leftarrow 0$

while $p < l(\alpha)$ do

begin $m$ = maximum integer for which there exists a $j < p$

satisfying $\alpha_{j,j+m} = \alpha_{p,p+m}$ and $p+m < l(\alpha)$;

$1$ + some $j$ satisfying the above;
$k + k+1; \ p + p+m+1;$
\text{t}_k + <i,m,a(p)>$

end;


It can easily be seen that \( T(\alpha) \) exists and is well defined. Lempel and Ziv [2] define \( \ell(\ell(\alpha)) \), the number of triples in the translation of \( \alpha \), which is the same as the final value of \( k \) in Algorithm 2.1, to be the complexity of \( \alpha \); denote it by \( C(\alpha) \).

Assume that we have translated \( \alpha(0,p) \) into \( t_1 \ldots t_k \), and computed \( t_{k+1} = (l,m,\alpha(p+m+1)) \). Thus \( \alpha(p,p+m) \) is a repetition at \( p+m \). If \( p+m < \ell(\alpha) \) then \( \alpha(p,p+m+1) \) is not a repetition. In terms of the brackets movements of Algorithm 1.1 we may say that when "\( ] \)" first arrives at \( p+m \) "\( [ \)" is not to the right of \( p \) while when "\( ] \)" is at \( p+m+1 \) "\( [ \)" is to the right of \( p \). The following algorithm implements the ideas above:

We let \( p[\cdot,p] \) denote the current position of \( [ \) and \( ] \) respectively; \( \text{loc}(\beta) \) the location of the first occurrence of \( \beta \), and \( p \) the position immediately after the current translated prefix of \( \alpha \).
Algorithm 2.2.

\[ k=0; \ p=0; \]

**start with []a;**

**while the picture is ...[β]a... do**

**begin if βa is a repetition then move ] else**

**if β≠ε then begin**

\[ j+p_j-p; \]

\[ l+\text{loc}(β)-j; \]

**move [;**

**if p_j>p then begin**

\[ k+k+1; \]

\[ t_k=⟨l,j,a⟩; \]

\[ p+p_j+1 \]

**end**

**else begin if p_j=p then begin**

\[ k+k+1; \]

\[ t_k=⟨0,0,a⟩; \]

\[ p+p+1 \]

**end;**

**move [ end**

**end;**

**if p<l(α) then begin**

\[ j+\ell(α)-p-1; \]

\[ l+\text{loc}(β)-j; \]

\[ k+k+1; \]

\[ t_k=⟨l,j,a\langle\ell(α)⟩⟩ \]

**end;**
It is easy to see that the movements of "[" and "]" on \( x \) are identical to those in Algorithm 1.1. The production of \( T(a) \) is added to the identification of longest repetitions and does not change the flow of the repetition finding algorithm. To each iteration of the single loop, we add a constant number of operations. Therefore, Algorithm 2.2 is as linear as Algorithm 1.1. We can say a little more: The production cost of each triple of \( T(a) \) is linear in the length of the substring which it represents. We refer to the movements of the three pointers introduced in Section 1. Denote two consecutive substrings of \( x \) each of which is translated into a single letter of \( \Theta \), by \( \gamma \) and \( \gamma' \). While "[" moves along \( \gamma \), "]" moves along \( \gamma' \). The third pointer is bounded between these two, and thus can move no more than \( l(\gamma'\gamma) \) steps. Thus, the total cost is bounded by \( O(l(\gamma')+l(\gamma)) \). The contribution of \( l(\gamma') \) may be charged to the translation of \( \gamma' \). Thus the cost of translating \( \gamma \) into a single triple is \( O(l(\gamma)) \).

2.3. Encoding of Words Over \( \Theta \).

For simplicity we assume that \( \Sigma=\{0,1\} \). Let \( \tau=t_1t_2\ldots t_k \) be a word over \( \Theta \). Each \( t_q \) has the form \( <i,j,a> \). We use Elias' representation of the integers to encode \( i \) and \( j \). The code of \( t_q \) is formed by the concatenation of these codes and \( a \). We get a uniquely decipherable binary encoding of \( \tau \).

Let \( B(n) \) denote that binary representation of \( n \), for which
the most significant bit is 1. \( \log x \) is an abbreviation for 
\([\log(x+1)]\) where for the binary case, the base of the logarithm
is 2, and \([x]\) is the least integer \( \geq x; \log^2 x = \log(\log x), \)
\( \log^{1+1} x = \log^{1} (\log x). \)

Define \( R'(n) \) as follows:

\[
R'(n) = \begin{cases} 
  b_1 b_2 b_3 & \text{if } n \leq 7 \\
  R'(\log n) B(n) & \text{if } n > 7 
\end{cases}
\]

where \( b_1 b_2 b_3 \) is a binary representation of \( n. \)

Next we define \( R(n) \) by
\[
R(n) = R'(n) 0.
\]

\( R(n) \) is the representation of \( n \) which we use, and it has the
prefix property (there exists no \( m < n \) for which \( R(m) \) is a prefix
of \( R(n) \)). It can be proven [5] that for large \( n, \)
\[
(R(n)) = 1 + \sum_{i=1}^{\log n=3} \log^i n
\]

and \( (R(n)) \leq \log n + 2 \log^2 n. \)

To produce \( R(n) \) for \( n \geq 4 \) perform the following algorithm:

**Algorithm 2.3.**

\[
\begin{align*}
\text{m} & \leftarrow n; \\
\text{R(n)} & \leftarrow 0; \\
\text{while } m > 3 \text{ do } & \xi + \text{B(m)}; \\
\text{R(n)} & \leftarrow \xi \text{R(n)}; \\
\text{m} & \leftarrow \text{the number of bits in } \xi \\
\text{end}
\end{align*}
\]
The cost of producing $R(n)$ is $O(2^{|R(n)|})$. But $2^{|R(n)|} \leq 2 \lg n$. Therefore, the production time of $R(n)$ is linear in the length of the binary representation of $n$.

2.4. Encoding of Strings.

Let us show that the compression ratio of the composed encoding scheme tends to be optimal as the length of $\alpha$ grows to infinity.

Denote the complexity of $\alpha$ by $C(\alpha)$. Lempel and Ziv [2] show that for almost all strings emitted by an ergodic source with entropy $h$

$$C(\alpha) \leq \frac{h_n}{\lg n}.$$  

Shperling [6] shows that $\frac{h_n}{\lg n}$ is the limit in probability of $C(\alpha)$, under more relaxed conditions.

Let $T(\alpha)=t_1 t_2 \ldots t_k$ where $t_q=\langle l_q, j_q, a_q \rangle$. Then the length of the encoding of $\alpha$ is bounded by

$$k \sum_{q=1}^{k} \lg l_q + 2 \lg^2 l_q + \lg j_q + 2 \lg^2 j_q + 1.$$  

$\lg$ is a convex function. Thus, the length is bounded by

$$C(\alpha) \cdot (\lg n + 2 \lg^2 n + \lg \frac{n}{\lg n} + 2 \lg^2 \frac{n}{\lg n} + 1).$$  

For large $n$ this tends to

$$\frac{h_n}{\lg n} \cdot (\lg n + 2 \lg^2 n + \lg \frac{\lg n}{h} + 2 \lg^2 \frac{\lg n}{h}).$$

Therefore the compression ratio tends to $h$ for almost all strings, which is the best possible result.

More about the asymptotic complexity of sequences appears in [6].
Clearly, the main handicap of this data compression scheme is that the memory requirement is unbounded. In the next section we shall describe a method more suitable in cases of bounded memory.

3. Algorithm for Data Compression with Bounded Memory.

3.1. Introduction.

Ziv and Lempel [3] invented a compression method which first parses an input string into bounded length substrings and then uses an adaptive coding scheme which maps these substrings sequentially into uniquely decipherable code-words of fixed length over the desired alphabet. For a given $\epsilon$ and for all finite memory sources with entropy $h$, the scheme uses a fixed amount of memory and its compression ratio is proven to be close to optimal up to $\epsilon$.

We implement the proposed compression method by combining Weiner's algorithm with the repetition finding algorithm described in Section 1. The two parameters of the algorithm are $N$, the memory size, and $F$, a bound on the length of source substring to be represented by a single code word.

3.2. The Compression Method.

Let $\alpha$ be an input string. We shift a window of size $N$ along $\alpha$. It is set initially to hold zeros and a prefix of
length $F$ of $\alpha$. We encode $\alpha$ step by step. At each step we assume that the $N-F$ leftmost letters within the window have been encoded already and we encode as much as we can out of the rightmost $F$ letters. Let $\alpha_1$ denote the longest prefix of these $F$ letters such that there exists an earlier occurrence of $\alpha <0, \ell(\alpha)-1>$ within the window. Denote by $p_1$ the start position of such an occurrence (the window's left boundary is considered to be position 0). Then we encode $\alpha_1$ into 

$$C_1 = \langle p_1, \ell(\alpha_1)-1, \alpha_1 \ell(\alpha_1) \rangle,$$

and shift the window $\ell(\alpha_1)$ positions to the right.

For simplicity assume that the source and code alphabets are binary. To write each $C_1$ we use

$$L = \lceil \log_2(N-F) \rceil + \lceil \log_2(F) \rceil + 1$$

bits (padding with zeros if necessary). The code of $\alpha$ is the concatenation of its $C_1$'s.

### 3.3. Implementation of the Compression Method.

The compression method just described is similar to that in Section 2. There are two main differences:

1. The length of each $\alpha_1$ is bounded by $F$. To handle this constraint all we have to do is to follow the "$\rangle$" movements and when it goes "too far", produce a new $C_1$.

2. The size of the window is fixed. Thus when we shift its right end, we have to shift its left end too.

Algorithm 1.3 is designed to handle strings given in a
left to right mode. It currently holds information about a portion of the input string which starts at a fixed leftmost position, \( P_1 \), and ends in a position that advances to the right. If the fixed start position lies inside the window, it can simulate the window's right side movements. In fact the suffix of length \( F-1(a_1) \) of the window's content is ignored in such a simulation.

Weiner's algorithm handles strings in a right to left mode. Assume that Weiner's algorithm is applied from some fixed rightmost position, \( P_2 \), to the left and processed \( K \) letters. While the algorithm is proceeding, a stack is built which holds all actions on the data structure. Using this stack we can "reverse the algorithm" and get a new algorithm which acts upon the string from left to right in linear time while its right end is fixed. (Actually only a substructure of the data is necessary in our application and the algorithm can be improved accordingly.) If this fixed position lies inside the window the reversed algorithm can simulate the window's left side movements.

To implement the above compression method we have to be able to check for the existence of an occurrence of a string whose length is bounded by \( F-1 \) within the window. Therefore we take \( P_2 = P_1 + F-2 \). In this way if there exists such an occurrence, it occurs either on the left side of \( P_2 \) or on the right side of \( P_1 \); thus it is treated by at least one of the two algorithms.
To ensure that there always exist two positions $P_1$ and $P_2$ inside the window we design a sequence of such couples: Note that the $N-F$ leftmost letters are initially zero. To make the discussion uniform define $\xi$ to be $0^{N-F}$ and put the window on the first $N$ letters of $\xi$. In the sequel "position $j$" abbreviates "position $j$ in $\xi$". Define $P_1^1$ and $P_2^1$ by:

$$
P_1^1 = \lfloor \frac{(N-F)}{2} \rfloor - (F-2)$$

$$
P_2^1 = \lfloor \frac{(N-F)}{2} \rfloor ;
$$

For $l \geq 1$:

$$
P_2^{l+1} = P_1^l + \lfloor \frac{(N-F)}{2} \rfloor 
$$

$$
P_1^{l+1} = P_2^l - F + 2 .
$$

Usually $F << N$. More explicitly we assume that $N > 9F-14$. In this case the number of couples inside the window is at most three; the leftmost is being used and the others are being developed. When the couple in use is shifted out, the one next to it is ready, and we start using it. At this moment, $P_2^{l+1}$ is located at position $\lfloor (N-F)/2 \rfloor$. Thus Weiner's algorithm has had enough time to produce the desired data structure. Therefore the composed algorithm is linear.

The assumption that $N > 9F-14$ is sound since for a given entropy $h$, $F$ should satisfy

$$
F \approx \log N/h ,
$$

and, as $N$ increases the ratio between $F$ and $N$ decreases logarithmically. Thus, a reasonable choice of $N$ will satisfy the inequality.
3.4. Remarks.

Ziv and Lempel [3] do not use the shift facility of the window to prove that the compression ratio of their scheme is close to the entropy; the input is partitioned into sections, each of which is of length N, which is the buffer size. Thus we can develop an algorithm which is very similar to Algorithm 2.2 but has a bound on the length of substrings that are represented by a single code word. When the current section is encoded the memory is cleared and the encoding of a new section is started. The intuitive disadvantage of such a scheme is that when we restart, we cannot use known substrings and therefore for a while the compression is poor. However we gain simplicity and may hold a larger portion of the input string in core.

A compromise between the above scheme and the one described in 3.2 is the following: Design a sequence of positions $P_I$ defined by:

$$P_1 = 0,$$

For $I \geq 1$,

$$P_{I+1} = P_I + \lceil (N-F)/2 \rceil.$$ 

There are at least two such positions inside a window of size N. When $P_I$ first enters into the window, we initiate the repetition finder and stop it when it represents $\varepsilon \langle P_I, P_I + I + F - 2 \rangle$. Denote the corresponding data structure by $D_I$. When $P_I$ is shifted out, drop $D_I$ out of memory.
Clearly, we use $D_i$ and $D_{i+1}$ for matching with a new portion of $\xi$. In this way we may always refer to at least half of the window and get a more effective scheme.

The ideas above illustrate two variations of the two basic algorithms; others are possible too.
REFERENCES


