Ukkonen's Cut-off Heuristic

We can speed up the algorithm using the diagonal monotonicity of the matrix \((g_{ij})\):

A diagonal \(-m \leq d \leq n\), consists of the cells \(g_{ij}\) with \(j - i = d\).

Every diagonal in \((g_{ij})\) is monotonically non-decreasing.

**Example 3.11:** Diagonals -3 and 2.

<table>
<thead>
<tr>
<th>g</th>
<th>r</th>
<th>e</th>
<th>m</th>
<th>a</th>
<th>t</th>
<th>c</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

We can reduce computation using diagonal monotonicity:

- Whenever the value on a diagonal \(d\) grows larger than \(k\), we can discard \(d\) from consideration, because we are only interested in values at most \(k\) on the row \(m\).
- We keep track of the smallest undiscarded diagonal \(d\). Each column is computed only up to diagonal \(d + 1\).

**Example 3.13:** \(P = \text{strict}\), \(T = \text{datastructure}\), \(k = 1\)

<table>
<thead>
<tr>
<th>g</th>
<th>s</th>
<th>d</th>
<th>a</th>
<th>t</th>
<th>r</th>
<th>i</th>
<th>c</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

The row on the current column corresponding to the smallest undiscarded diagonal is kept in a variable \(top\).

**Algorithm 3.14:** Ukkonen's cut-off algorithm

**Input:** text \(T[1..n]\), pattern \(P[1..m]\), and integer \(k\)

**Output:** end positions of all approximate occurrences of \(P\)

1. \(i \leftarrow \min(k + 1, m)\)
2. \(j \leftarrow 0\)
3. \(top \leftarrow 0\)
4. \(g_{top,j} \leftarrow 0\)
5. \(g_{i,j} \leftarrow 0\) for \(i = 1, \ldots, m\)

The algorithm does not compute the \(\Delta d\) values but they are useful in the proofs.

**Myers’ Bitparallel Algorithm**

Another way to speed up the computation is bitparallelism.

Instead of the matrix \((g_{ij})\), we store differences between adjacent cells:

- **Vertical delta:** \(\Delta v_{ij} = g_{ij} - g_{i-1,j}\)
- **Horizontal delta:** \(\Delta h_{ij} = g_{ij} - g_{i,j-1}\)
- **Diagonal delta:** \(\Delta d_{ij} = g_{ij} - g_{i-1,j-1}\)

Because \(g_{ij} = i\) if \(g_{ij} = 0\), \(g_{ij} = \Delta v_{ij} + \Delta h_{ij} + \cdots + \Delta v_{ij} + \cdots + \Delta h_{ij}\)

\(= i + \Delta h_{ij} + \Delta v_{ij} + 1\), and it can be stored in one bit.

The following result, \(\Delta h_{ij}, \Delta v_{ij} \in \{0, 1\}\) for every \(i, j\) that they are defined for.

The proof is left as an exercise.

In the standard computation of a cell:

- **Input** is \(g_{i-1,j}, g_{i,j-1}, g_{i-1,j-1}\) and \(\delta(P[i], T[j])\).
- **Output** is \(g_{ij}\).

In the corresponding bitparallel computation:

- **Input** is \(\Delta v^m = \Delta v_{i-1,j},\ \Delta h^m = \Delta h_{i,j} - 1\) and \(E_{eq} = E_{eq} = 1 - \delta(P[i], T[j])\).
- **Output** is \(\Delta v_{out} = \Delta v_{i,j}\) and \(\Delta h_{out} = \Delta h_{i,j}\).

The algorithm does not compute the \(\Delta d\) values but they are useful in the proofs.
The computation rule is defined by the following result.

**Lemma 3.17:** If $E_0 = 1$ or $\Delta v^0 = -1$ or $\Delta h^0 = -1$, then $\Delta d = 0$, $\Delta d^0 = 1 - \Delta h^0$ and $\Delta h^0 = 1 - \Delta d^0$.

**Proof.** We can write the recurrence for $g_{ij}$ as

$$g_{ij} = \min[g_{i-1,j-1} + \delta(P[i,j]), 1, g_{i-1,j} + 1, g_{i,j-1} + 1].$$

Then $\Delta d = g_{ij} - g_{i-1,j-1} = \min[1 - Eq(i,j) \Delta d^0 + 1, \Delta h^0 + 1]$ which is 0 if $Eq = 1$ or $\Delta v^0 = -1$ or $\Delta h^0 = -1$ and 1 otherwise.

Clearly $\Delta d = \Delta v^0 + \Delta h^0 = \Delta h^0 + \Delta v^0$, Thus $\Delta d^0 = \Delta v^0 - \Delta h^0$ and $\Delta h^0 = \Delta v^0 - \Delta h^0$. □

To enable bitparallel operation, we need two changes:

- The $\Delta v$ and $\Delta h$ values are “trits” not bits. We encode each of them with two bits as follows:
  $$P_v = \begin{cases} 1 & \text{if } \Delta v = 1 \\ 0 & \text{otherwise} \end{cases} \quad M_v = \begin{cases} 1 & \text{if } \Delta v = -1 \\ 0 & \text{otherwise} \end{cases}$$
  $$P_h = \begin{cases} 1 & \text{if } \Delta h = 1 \\ 0 & \text{otherwise} \end{cases} \quad M_h = \begin{cases} 1 & \text{if } \Delta h = -1 \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\Delta v = P_v - M_v \quad \Delta h = P_h - M_h$$

- We replace arithmetic operations $(+, -)$ with Boolean (logical) operations $(\land, \lor, \neg)$.

According to Lemma 3.18, the bit representation of the matrix can be computed as follows.

- For $i \leftarrow 1$ to $m$ do
  $$P_{vi0} = 1; \quad M_{vi0} = 0$$
  $$P_{vij} \leftarrow M_{hij-1} \lor \neg(X_{vij} \lor P_{vij-1})$$
  $$M_{vij} \leftarrow P_{vij-1} \land X_{vij}$$

This is not yet bitparallel though.

To obtain a bitparallel algorithm, the columns $P_{vij}, M_{vij}, X_{vij}, Ph_{ij}, Mh_{ij}, Xh_{ij}$ and $Eq_{ij}$ are stored in bitvectors.

Now the second inner loop can be replaced with the code

$$Xv_{ij} \leftarrow Eq_{ij} \lor Mv_{ij-1}$$
$$Pv_{ij} \leftarrow (Xh_{ij} < 1) \lor (Xv_{ij} \lor (Ph_{ij} < 1))$$
$$Mv_{ij} \leftarrow (Ph_{ij} < 1) \land Xv_{ij}$$

A similar attempt with the for the first inner loop leads to a problem:

$$Xh_{ij} \leftarrow Eq_{ij} \lor Mh_{ij-1}$$
$$Ph_{ij} \leftarrow Mv_{ij-1} \land \neg(Xh_{ij} \lor P_{vij-1})$$
$$Mh_{ij} \leftarrow P_{vij} \land Xh_{ij}$$

Now the vector $Mh_{ij}$ is used in computing $Xh_{ij}$ before $Mh_{ij}$ itself is computed! Changing the order does not help, because $Xh_{ij}$ is needed to compute $Mh_{ij}$.

To get out of this dependency loop, we compute $Xh_{ij}$ without $Mh_{ij}$ using only $Eq_{ij}$ and $P_{vij-1}$ which are already available when we compute $Xh_{ij}$.

At first sight, we cannot use Lemma 3.19 to compute even a single bit in constant time, let alone a whole vector $Xh_{ij}$. However, it can be done, but we need more bit operations:

- Let $\land$ denote the xor-operation: $0 \land 1 = 1 \land 0 = 1$ and $0 \land 0 = 1 \land 1 = 0$.
- A bitvector is interpreted as an integer and we use addition as a bit operation. The carry mechanism in addition plays a key role. For example $0001 + 0111 = 1000$.

In the following, for a bitvector $B$, we will write

$$B = B[1..m] = B[m]B[m-1] \ldots B[1]$$

The reverse order of the bits reflects the interpretation as an integer.
The following calculation shows that $Y[i] = 1$ in this case:

\[ E[\ldots i] = 00\ldots 01 \]

$P[\ldots i] = b1\ldots 1t$

\[ (E \land P)[\ldots i] = 00\ldots 0t \]

\[ ((E \land P) + P)[\ldots i] = 01\ldots 1t \]

\[ Y = (((E \land P) + P) \lor P) \lor P)[\ldots i] = 11\ldots 1t \]

where $b$ is the unknown bit $P[i]$, $c$ is the possible carry bit coming from the summation of bits $1 \ldots \ell - 1$, and $b$ and $c$ are their negations.

c) Because for all bitvectors $B$, $0 \lor B = 0$ and $B = B$, we get $Y = (((0 \land P) + P) \lor P) \lor P = 0$.

d) Consider the calculation in case b). A key point there is that the carry bit in the summation travels from position $\ell$ to $i$ and produces $b$ to position $i$. The difference in this case is that at least one bit $P[\ell]$, $\ell < k < i$, is zero, which stops the carry at position $k$. Thus

\[ ((E \land P) + P)[i] = b \] and $Y[i] = (b \oplus b) \lor 0 = 0$. □

In the integer alphabet model, when $m \leq w$:

- Pattern preprocessing time is $O(m + n)$.
- Search time is $O(n)$.

When $m > w$, we can store each bit vector in $[m/w]$ machine words:

- The worst case search time is $O(n/m[w])$.
- Using Ukkonen’s cut-off heuristic, it is possible reduce the average case search time to $O(n/k[w])$.

There are also algorithms for bitparallel simulation of a nondeterministic automaton that recognizes the approximate occurrences of the pattern.

**Example 3.22:**

\[ P = \text{patrono}, \; k = 3 \]

\[ \begin{array}{cccccccc}
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array} \]

**Baeza-Yates–Perleberg Filtering Algorithm**

A filtering algorithm for approximate string matching searches the text for factors having some property that satisfies the following conditions:

1. Every approximate occurrence of the pattern has this property.
2. Strings having this property are reasonably rare.
3. Text factors having this property can be found quickly.

Each text factor with the property is a potential occurrence, which is then verified for whether it is an actual approximate occurrence.

The Karp–Rabin algorithm is a filtering algorithm for exact string matching. The property we are looking for in that case is having the same fingerprint as the pattern.

Filtering algorithms can achieve linear or even sublinear average case time complexity.

The algorithm has two phases:

**Filtration:** Search the text $T$ for exact occurrences of the pattern factors $P$.

Using the Aho–Corasick algorithm this takes $O(n)$ time in the constant alphabet model.

**Verification:** An area of length $O(m)$ surrounding each potential occurrence found in the filtration phase is searched using the standard dynamic programming algorithm in $O(m^2)$ time.

The worst case time complexity is $O(m^2n)$, which can be reduced to $O(mn)$ by combining any overlapping areas to be searched.

As a final detail, we compute the bottom row values $g_{mn}$ using the equalities $g_{mn} = m \; j = g_{m+1} + \Delta h_{m+1}$.

**Algorithm 3.21:** Myers’ bitparallel algorithm

Input: text $T[1..n]$, pattern $P[1..m]$, and integer $k$

Output: end positions of all approximate occurrences of $P$

(1) for $c \in \Sigma$ do $B[c] \leftarrow 0^n$

(2) for $i \leftarrow 1$ to $n$ do $B[P[i]] \leftarrow 1$

(3) $Pv \leftarrow 1^n$, $Mv \leftarrow 0$, $g \leftarrow m$

(4) for $j \leftarrow 1$ to $n$ do

(5) $Eq \leftarrow B[\ell][j]$

(6) $Xh \leftarrow (((Eq \land Pv) + Pv) \lor Pv) \lor Eq$

(7) $Ph \leftarrow Mh \lor (Xh \lor Pv)$

(8) $Mh \leftarrow Ph \times Xh$

(9) $Xv \leftarrow Eq \lor Mv$

(10) $Pv \leftarrow (Mh < 1) \lor (Xh \lor (Ph < 1))$

(11) $Mv \leftarrow (Ph < 1) \lor Xv$

(12) $g \leftarrow g + Ph[m] - Mh[m]$

(13) if $g \leq k$ then output $j$

Another way to utilize Lemma 3.15 $(\Delta h_{ij}, \Delta v_{ij} \in \{-1,0,1\})$ is to use precomputed tables to process multiple matrix cells at a time.

- There are at most 3$^r$ different columns. Thus there exists a deterministic automaton with $3^r$ states and $3^r$ transitions that can find all approximate occurrences in $O(n)$ time. However, the space and constructions time of the automaton can be too big to be practical.

- There is an algorithm for the packed string model that processes $O(\log n)$ characters at a time and $O(\log^2 n)$ matrix cells at a time using lookup tables of size $O(n)$. This gives time complexity $O((mn)/\log^2 n)$.

- A practical variant uses smaller lookup tables to compute multiple entries of a column at a time.

The following lemma shows the property used by the Baeza-Yates–Perleberg algorithm and proves that it satisfies the first condition.

**Lemma 3.23:** Let $P[1]P_2 \ldots P_{k+1} = P$ be a partitioning of the pattern $P$ into $k+1$ nonempty factors. Any string $S$ with $n d(P,S) \leq k$ contains $P$ as a factor for some $i \in [1..k+1]$.

**Proof.** Each single symbol edit operation can change at most one of the pattern factors $P_i$. Thus any set of at most $k$ edit operations leaves at least one of the factors untouched. □

Let us analyze the average case time complexity of the verification phase.

- The best pattern partitioning is as even as possible. Then each pattern factor has length at least $r = \lceil m/(k+1) \rceil$.
- The expected number of exact occurrences of a random string of length $r$ in a random text of length $n$ is at most $n/r^e$.
- The expected total verification time is at most

\[ O \left( (k+1) \frac{n}{r^e} \cdot m^2 \right) \leq O \left( \frac{n^2 m^2}{\sigma^2} \right). \]

This is $O(n)$ if $r \geq 3 \log_m m$.

- The condition $r \geq 3 \log_m m$ is satisfied when $(k+1) \leq m/ (3 \log_m m + 1)$.

**Theorem 3.24:** The average case time complexity of the Baeza-Yates–Perleberg algorithm is $O(n)$ when $k \leq m/(3 \log_m m + 1) - 1$. 

Many variations of the algorithm have been suggested:

- The filtration can be done with a different multiple exact string matching algorithm.
- The verification time can be reduced using a technique called hierarchical verification.
- The pattern can be partitioned into fewer than \( k + 1 \) pieces, which are searched allowing a small number of errors.

A lower bound on the average case time complexity is \( \Omega\left( n(k + \log \sigma m) / m \right) \), and there exists a filtering algorithm matching this bound.

**Summary: Approximate String Matching**

We have seen two main types of algorithms for approximate string matching:

- Basic dynamic programming time complexity is \( O(mn) \). The time complexity can be improved to \( O(kn) \) using diagonal monotonicity, and to \( O(n(m/w)) \) using bitparallelism.
- Filtering algorithms can improve average case time complexity and are the fastest in practice when \( k \) is not too large. The partitioning into \( k + 1 \) factors is a simple but effective filtering technique.

More advanced techniques have been developed but are not covered here (except in study groups).

Similar techniques can be useful for other variants of edit distance but not always straightforwardly.

**Selected Literature**

- Survey
  

- Edit distance
  

- Dynamic programming
  

  and many other independent discoveries.

- Approximate string matching
  

  Ukkonen’s cut-off algorithm
  

  Myers’ bitparallel algorithm
  

  Baeza-Yates–Perleberg filtering algorithm
  