Text Algorithms

Jaak Vilo
2015 fall

Topics

• Exact matching of one pattern (string)
• Exact matching of multiple patterns
• Suffix trie and tree indexes
  — Applications
• Suffix arrays
• Inverted index
• Approximate matching

Exact pattern matching

• $S = s_1 s_2 \ldots s_n$ (text) $|S| = n$ (length)
• $P = p_1 p_2 \ldots p_m$ (pattern) $|P| = m$
• $\Sigma$ - alphabet $|\Sigma| = c$

• Does $S$ contain $P$?
  — Does $S = S' P S''$ for some strings $S', S''$?
  — Usually $m \ll n$ and $n$ can be (very) large

Find occurrences in text

Algorithms

One-pattern
• Brute force
• Knuth-Morris-Pratt
• Karp-Rabin
• Shift-Or, Shift-AND
• Boyer-Moore
• Factor searches

Multi-pattern
• Aho Corasick
• Commentz-Walter

Indexing
• Trie (and suffix trie)
• Suffix tree

Animations

• EXACT STRING MATCHING ALGORITHMS
  Animation in Java
• Christian Charras - Thierry Lecroq
  Laboratoire d’Informatique de Rouen
  Université de Rouen
  Faculté des Sciences et des Techniques
  76821 Mont-Saint-Aignan Cedex
  FRANCE
• e-mails: [Christian.Charras, Thierry.Lecroq]@laposte.net
Brute force: BAB in text?

A B A C A B A B B A B A B
B A B

Brute Force

\[ P \]
\[ S \]

Identify the first mismatch!

Question:

• Problems of this method? ☺
• Ideas to improve the search? ☺

Brute force or NaiveSearch

1 function NaiveSearch(string s[1..n], string sub[1..m])
2 for i from 1 to n-m+1
3 for j from 1 to m
4 if s[i+j-1] != sub[j]
5 jump to next iteration of outer loop
6 return i
7 return not found

C code

int bf_2( char* pat, char* text , int n ) /* n = textlen */
{
    int m, i, j ;
    int count = 0 ;
    m = strlen(pat);
    for ( i=0 ; i + m <= n ; i++ ) {
        for( j=0; j < m && pat[j] == text[i+j] ; j++) ;
        if( j == m )
            count++ ;
    }
    return(count);
}

int bf_1( char* pat, char* text )
{
    int m ;
    int count = 0 ;
    char *tp;
    m = strlen(pat);
    tp=text ;
    for( ; *tp ; tp++ ) {
        if( strncmp( pat, tp, m ) == 0 ) {
            count++ ;
        }
    }
    return( count );
}
Main problem of Naive

- For the next possible location of P, check again the same positions of S

Goals

- Make sure only a constant nr of comparisons/operations is made for each position in S
  - Move (only) from left to right in S
  - How?
  - After a test of S[i] <> P[j] what do we now?

Knuth-Morris-Pratt

- Make sure that no comparisons “wasted”

- After such a mismatch we already know exactly the values of green area in S!

Knuth-Morris-Pratt

- Make sure that no comparisons “wasted”

- P – longest suffix of any prefix that is also a prefix of a pattern
- Example: ABCABD

Automaton for ABCABD

Automaton for ABCABD

D. Knuth, J. Morris, V. Pratt:
Fast Pattern Matching in strings.

D. Knuth, J. Morris, V. Pratt:
Fast Pattern Matching in strings.

D. Knuth, J. Morris, V. Pratt:
Fast Pattern Matching in strings.
KMP matching

Input: Text $S[1..n]$ and pattern $P[1..m]$
Output: First occurrence of $P$ in $S$ (if exists)

```plaintext
i=1; j=1;
initfail(P) // Prepare fail links
repeat
    if $j=0$ or $S[i] == P[j]$
    then $i++$, $j++$ // advance in text and in pattern
    else $j = fail[j]$ // use fail link
until $j>m$ or $i>n$
if $j>m$ then report match at $i-m$
```

Initialization of fail links

Algorithm: KMP_initfail
Input: Pattern $P[1..m]$
Output: fail[] for pattern $P$

```plaintext
i=1, j=0, fail[1]= 0
repeat
    if $j=0$ or $P[i] == P[j]$
    then $i++$, $j++$ // advance in text and pattern
    else $j = fail[j]$ // use fail link
until $i=m$
```

Time complexity of KMP matching?

Input: Text $S[1..n]$ and pattern $P[1..m]$
Output: First occurrence of $P$ in $S$ (if exists)

```plaintext
i=1; j=1;
initfail(P) // Prepare fail links
repeat
    if $j=0$ or $S[i] == P[j]$
    then $i++$, $j++$ // advance in text and in pattern
    else $j = fail[j]$ // use fail link
until $j>m$ or $i>n$
if $j>m$ then report match at $i-m$
```

Analysis of time complexity

- At every cycle either $i$ and $j$ increase by 1
- Or $j$ decreases ($j=fail[j]$)
- $i$ can increase $n$ (or $m$) times
- Q: How often can $j$ decrease?
  - A: not more than $n$ of increases of $i$
- Amortised analysis: $O(n)$, preprocess $O(m)$

Karp-Rabin


- Compare in $O(1)$ a hash of $P$ and $S[i..i+m-1]$

```plaintext
h($T[i..i+m-1]$)
```

- Goal: $O(n)$.
- $(h(T[i..i+m-1]) -> h(T[i+1..i+m])) = O(1)$
Karp-Rabin


• Compare in $O(1)$ a hash of $P$ and $S[i..i+m-1]$

$$h(T[i..i+m-1])$$

$$h(P)$$

• Goal: $O(n)$.
• $f(h(T[i..i+m-1]) \rightarrow h(T[i+1..i+m])) = O(1)$


Hash

• “Remove” the effect of $T[i]$ and “Introduce” the effect of $T[i+m]$ – in $O(1)$

• Use base $|X|$ arithmetics and treat characters as numbers

• In case of hash match – check all $m$ positions
• Hash collisions => Worst case $O(nm)$

Let’s use numbers

• $T = 57125677$
• $P = 125$ (and for simplicity, $h=125$)

• $H(T[1])=571$
• $H(T[2])= (571-5*100)*10 + 2 = 712$
• $H(T[3]) = (H(T[2]) - \text{ord}(T[1])*10m)*10 + T[3+m-1]$

hash

• $c$ – size of alphabet

• $HS_i = H(S[i..i+m-1])$

• $H(S[i+1..i+m]) = (HS_i - \text{ord}(S[i])*c^{m-1}) * c + \text{ord}(S[i+m])$

• Modulo arithmetic – to fit value in a word!

Karp-Rabin

Input: Text $S[1..n]$ and pattern $P[1..m]$
Output: Occurrences of $P$ in $S$

1. $c=20$; /* Size of the alphabet, say nr. of aminoacids */
2. $q = 35554393$ /* $q$ is a prime */
3. $c^m = c^{m-1} \mod q$
4. $hp = 0$; $hs = 0$
5. for $i = 1 .. m$ do $hp = (hp*c + \text{ord}(p[i])) \mod q$ // $H(P)$
6. for $i = 1 .. m$ do $hs = (hp*c + \text{ord}(s[i])) \mod q$ // $H(S[1..m])$
7. if $hp == hs$ and $P == S[1..m]$ report match at position
8. for $i=2 .. n-m+1$
9. $hs = (hs - \text{ord}(s[i-1])*c^{m-1}) * c + \text{ord}(s[i+m-1]) \mod q$
10. if $hp == hs$ and $P == S[i..i+m-1]$
11. report match at position $i$
More ways to ensure \( O(n) \)?

**Shift-AND / Shift-OR**

- Ricardo Baeza-Yates, Gaston H. Gonnet
- A new approach to text searching
- [ACM Digital Library](http://doi.acm.org/10.1145/135239.135241) [DOI]
- [PDF]

---

**Bit-operations**

- Maintain a set of all prefixes that have so far had a perfect match
- On the next character in text update all previous pointers to a new set
- Bit vector: for every possible character

---

**State: which prefixes match?**

---

**Move to next:**

---

**Track positions of prefix matches**

---

**Shift left <<**

**Bitwise AND**

---
Vectors for every char in $\Sigma$

- $P=\text{aste}$

a s t e b c d .. z
1 0 0 0 0 ...
0 1 0 0 0 ...
0 0 1 0 0 ...
0 0 0 1 0 ...

- $T=\text{lasteaed}$

l a s t e a e d
0 1
0 0
0 0
0 0
0 0

- $T=\text{lasteaed}$

l a s t e a e d
0 1 0 0 0 1
0 0 1 0 0 0
0 0 0 1 0 0
0 0 0 0 1 0
0 0 0 0 1 0

- $T=\text{lasteaed}$

l a s t e a e d
0 1 0 0 0 1
0 0 1 0 0 0
0 0 0 1 0 0
0 0 0 0 1 0

http://www-igm.univ-mlv.fr/~lecroq/string/node6.html

Description:
The Shift Or algorithm uses bitvector techniques. Let $\text{B}$ be a bit array of size $m$. Vector $R_i$ is the value of the array $\text{B}$ after text character $x_i$ has been processed, (see figure 5.2). It contains informations about all matches of prefixes of $\text{B}$ that end at position $i$ in the text for $i = 0, \ldots, m$.

$$R_i[0]=0 \quad \text{if} \quad x_i=1$$
$$R_i[0]=1 \quad \text{otherwise}.$$
Summary

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Worst case</th>
<th>Ave. Case</th>
<th>Preprocess</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brute force</td>
<td>O(mn)</td>
<td>O(n * (1+1/</td>
<td>Σ</td>
</tr>
<tr>
<td>Knuth-Morris-Pratt</td>
<td>O(n)</td>
<td>O(n)</td>
<td>O(m)</td>
</tr>
<tr>
<td>Rabin-Karp</td>
<td>O(mn)</td>
<td>O(n)</td>
<td>O(m)</td>
</tr>
<tr>
<td>Boyer-Moore</td>
<td>O(n/m) ?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BM Horspool</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor search</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shift-Or</td>
<td>O(n)</td>
<td>O(n)</td>
<td>O( m[D] )</td>
</tr>
</tbody>
</table>


Find occurrences in text

- Have we missed anything?
Find occurrences in text

• What have we learned if we test for a potential match from the end?

Our search algorithm may be specified as follows:

```plaintext
void bmInitocc() {
    char a; int j;
    for(a=0; a<alphabetsize; a++)
        occ[a]=-1;
    for (j=0; j<m; j++) {
        a=p[j];
        occ[a]=j;
    }
}
```

Good suffix heuristics

1. S

2. S

delta(j S[[]]) → minimal shift so that matched region is fully covered or that the suffix of match is also a prefix of P

Bad character heuristics

maximal shift on S[i]

delta(j S[[]]) → |m| if pattern does not contain S[j]

First x in pattern (from end)
Boyer-Moore algorithm

Input: Text S[1..n] and pattern P[1..m]
Output: Occurrences of P in S

preprocess_BM() // delta1 and delta2
i=m
while i <= n
    for (j=m; j>0 and P[j]==S[i-m+j]; j--)
        if j==0 report match at position i-m+1
        i = i + max( delta1[ S[i] ], delta2[ j ] )

Simplifications of BM

- There are many variants of Boyer-Moore, and many scientific papers.
- On average the time complexity is sublinear
- Algorithm speed can be improved and yet simplify the code.
- It is useful to use the last character heuristics (Horspool (1980), Baeza-Yates(1989), Hume and Sunday(1991)).

Algorithm BMH (Boyer-Moore-Horspool)

- RN Horspool - Practical Fast Searching in Strings
  Software - Practice and Experience, 10(6):501-506 1980

Input: Text S[1..n] and pattern P[1..m]
Output: occurrences of P in S
1. for a in Σ do delta[a] = m
2. for j=1..m-1 do delta[P[j]] = m-j
3. i=m
4. while i <= n
5.    if S[i] == P[m]
6.        j = m-1
7.        while (j>0 and P[j]==S[i-m+j]) j = j-1;
8.        if j==0 report match at i-m+1
9.        i = i + delta[S[i]]

String Matching: Horspool algorithm

- How the comparison is made?
  Text:  | | | | | | | | | |
  Pattern:  | | | | | | | | | |
  From right to left: suffix search

- Which is the next position of the window?
  Text:  | | | | | | | | | |
  Pattern:  | | | | | | | | | |
  It depends of where appears the last letter of the text, say it ‘a’, in the pattern:  | | | | | | | | | |
  Then it is necessary a preprocess that determines the length of the shift.

Algorithm Boyer-Moore-Horspool-Hume-Sunday (BMHHS)

- Use delta in a tight loop
- If match (delta=0) then check and apply original delta d

Input: Text S[1..n] and pattern P[1..m]
Output: occurrences of P in S
1. for a in Σ do delta[a] = m
2. for j=1..m-1 do delta[P[j]] = m-j
3. d = delta[P[m]]; // memorize d on P[m]
4. delta[P[m]] = 0; // ensure delta on match of last char is 0
5. for (i=m; i<n ; i = i+d)
6.    repeat // skip loop
7.        j=delta[S[i]]; i = i + 1
8.    until i == n
9.    for (j=m-1 ; j>0 and P[j]==S[i-m+j]; j--)
10.       if j==0 report match at i-m+1

BMHHS requires that the text is padded by P: S[n+1]..S[n+m] = P
(in order for the algorithm to finish correctly – at least one occurrence!).
• Daniel M. Sunday: A very fast substring search algorithm [PDF]
  Communications of the ACM August 1990, Volume 33 Issue 8

• Loop unrolling:
  • Avoid too many loops (each loop requires tests) by just repeating code
    within the loop.
  • Line 7 in previous algorithm can be replaced by:

*updated equation*

7. \( i = i + \delta[S[i]] \);  
   \( i = i + \delta[S[i]] \);  
   \( i = \left( t = \delta[S[i]] \right) \);

**Forward-Fast-Search: Another Fast Variant of the Boyer-Moore String Matching Algorithm**

• The Prague Stringology Conference '03
  • Domenico Cantone and Simone Faro

• Abstract: We present a variation of the Fast-Search string matching
  algorithm, a recent member of the large family of Boyer-Moore-like
  algorithms, and we compare it with some of the most effective string
  matching algorithms, such as Horspool, Quick Search, Tuned Boyer-
  Moore, Reverse Factor, Berry-Ravindran, and Fast-Search itself. All
  algorithms are compared in terms of run-time efficiency, number of text
  character inspections, and number of character comparisons. It turns out
  that our new proposed variant, though not linear, achieves very good
  results especially in the case of very short patterns or small alphabets.
  • http://cs.felk.cvut.cz/psc/event/2003/p2.html
  • PS.gz (local copy)

**Factor based approach**

• Optimal average-case algorithms
  -- Assuming independent characters, same probability

• Factor – a substring of a pattern
  -- Any substring
  -- (how many?)

**Factor searches**

Do not compare characters, but find the longest match to any
subregion of the pattern.
Examples

- Backward DAWG Matching (BDM)
  — Crochemore et al 1994
- Backward Nondeterministic DAWG Matching (BNDM)
  — Navarro, Raffinot 2000
- Backward Oracle Matching (BOM)
  — Allauzen, Crochemore, Raffinot 2001

Backward DAWG Matching BDM

Suffix automaton recognises all factors (and suffixes) in $O(n)$

BNDM – simulate using bitparallelism

Bits – show where the factors have occurred so far

BNDM matches an NDA

NDA on the suffixes of ‘announce’

Deterministic version of the same
Backward Factor Oracle
String Matching of one pattern

CTACTACTACGTCTATACTGATCGTAGC
TACTACGGTATGACTAA

1. Prefix search
2. Suffix search
3. Factor search

Multiple patterns

Why?

• Multiple patterns
  • Highlight multiple different search words on the page
  • Virus detection – filter for virus signatures
  • Spam filters
  • Scanner in compiler needs to search for multiple keywords
  • Filter out stop words or disallowed words
  • Intrusion detection software
  • Next-generation sequencing produces huge amounts (many millions) of short reads (20-100 bp) that need to be mapped to genome!

Algorithms

• Aho-Corasick (search for multiple words)
  — Generalization of Knuth-Morris-Pratt
• Commentz-Walter
  — Generalization of Boyer-Moore & AC
• Wu and Manber
  — improvement over C-W
• Additional methods, tricks and techniques

Aho-Corasick (AC)

• Alfred V. Aho and Margaret J. Corasick (Bell Labs, Murray Hill, NJ)
• ACM DOI PDF

ABSTRACT This paper describes a simple, efficient algorithm to locate all occurrences of any of a finite number of keywords in a string of text. The algorithm consists of constructing a finite state pattern matching machine from the keywords and then using the pattern matching machine to process the text string in a single pass. Construction of the pattern matching machine takes time proportional to the sum of the lengths of the keywords. The number of state transitions made by the pattern matching machine in processing the text string is independent of the number of keywords. The algorithm has been used to improve the speed of a library bibliographic search program by a factor of 5 to 10.

References:

Generalization of KMP for many patterns
• Text $S$ like before.
• Set of patterns $P = \{ P_1, \ldots, P_k \}$
• Total length $\mid P \mid = m = \Sigma_{i=1..k} m_i$
• Problem: find all occurrences of any of the $P_i \in P$ from $S$
Idea

1. Create an **automaton** from all patterns

2. Match the automaton

   - Use the PATRICIA trie for creating the main structure of the automaton

**PATRICIA trie**

- **Abstract** PATRICIA is an algorithm which provides a flexible means of storing, indexing, and retrieving information in a large file, which is economical of index space and of reindexing time. It does not require rearrangement of text or index as new material is added. It requires a minimum restriction of format of text and of keys; it is extremely flexible in the variety of keys it will respond to. It retrieves information in response to keys furnished by the user with a quantity of computation which has a bound which depends linearly on the length of keys and the number of their proper occurrences and is otherwise independent of the size of the library. It has been implemented in several variations as FORTRAN programs for the CDC-3600, utilizing disk file storage of text. It has been applied to several large information-retrieval problems and will be applied to others.
- [ACM DOI PDF](http://doi.acm.org/10.1145/321093.321097)

---

**Word trie** - a good data structure to represent a set of words (e.g. a dictionary).

**Trie (data structure)**

**Definition**: A tree for storing strings in which there is one node for every common prefix. The strings are stored in extra leaf nodes.

- See also digital tree, digital search tree, directed acyclic word graph, compact DAWG, Patricia tree, suffix tree.
- **Note**: The name comes from retrieval and is pronounced, "tree."
- To test for a word $p$, only $O(|p|)$ time is used no matter how many words are in the dictionary...

---

[Diagram of Trie for $P=${he, she, his, hers}]

[Diagram of Trie for $P=${he, she, his, hers}]
How to search for words like he, sheila, hi. Do these occur in the trie?

Aho-Corasick

1. Create an automaton \( M_P \) for a set of strings \( P \).
2. Finite state machine: read a character from text, and change the state of the automaton based on the state transitions...
3. Main links: \( \text{goto}[j,c] \) - read a character c from text and go from a state j to state \( \text{goto}[j,c] \).
4. If there are no \( \text{goto}[j,c] \) links on character c from state j, use \( \text{fail}[j] \).
5. Report the output. Report all words that have been found in state j.

AC Automaton (vs KMP)

Output table

state output[ ]
2 ha
5 she, he
7 his
9 hers

AC - matching

Input: Text \([1..n]\) and an AC automaton \( M \) for pattern set \( P \)
Output: Occurrences of patterns from \( P \) in \( S \) (last position)
1. state = 0
2. for \( i = 1..n \) do
3. while \( (\text{goto}[\text{state},S[i]] == \emptyset) \) and \( (\text{fail}[\text{state}] != \text{state}) \) do
4. state = \( \text{fail}[\text{state}] \)
5. \( \text{state} = \text{goto}[\text{state},S[i]] \)
6. if \( (\text{output}[\text{state}] \) not empty \)
7. then report matches \( \text{output}[\text{state}] \) at position \( i \)

Algorithm Aho-Corasick preprocessing I (TRIE)

Input: \( P = \{ P_1, ..., P_k \} \)
Output: \( \text{goto}[\] and \( \text{parDal}[\] output[\]
Assume: \( \text{output}[s] \) is empty when a state \( s \) is created;
\( \text{goto}[s,a] \) is not defined.

procedure enter(\( \alpha_1, ..., \alpha_m \) /* \( P_i = \alpha_1, ..., \alpha_m \) */
begin
1. \( s = 0; j = 1; \)
2. while \( \text{goto}[s,\alpha_j] != \emptyset \) do // follow existing path
3. \( s = \text{goto}[s,\alpha_j]; \)
4. \( j = j + 1; \)
5. for \( p = j \) to \( m \) do // add new path (states)
6. \( \text{news} = \text{new} \);
7. \( \text{goto}[\text{news},\alpha_p] = \text{news}; \)
8. \( s = \text{news}; \)
9. \( \text{output}[s] = \alpha_1, ..., \alpha_m; \)
end

Preprocessing II for AC (FAIL)

begin
10. \( \text{news} = 0; \)
11. for \( i = 1 \) to \( k \) do enter(\( P_i \) )
12. for \( a \in \Sigma \) do
13. if \( \text{goto}[0,a] != \emptyset \) then \( \text{enqueue}(\text{queue}, \text{goto}[0,a]) \)
end

while \( \text{queue} \neq \emptyset \)
begin
1. \( r = \text{take}(\text{queue}) \)
2. for \( a \in \Sigma \) do
3. if \( \text{goto}[r,a] \neq \emptyset \) then \( s = \text{goto}[r,a] \)
4. \( \text{enqueue}(\text{queue}, s) \) // breadth first search
end

state = \( \text{fail}[r] \)
while \( \text{goto}[,state,a] != \emptyset \) do state = \( \text{fail}[\text{state}] \)
\( \text{fail}[s] = \text{goto}[,state,a] \)
\( \text{output}[s] = \text{output}[s] + \text{output}[\text{fail}[s]] \)
Correctness

- Let string t "point" from initial state to state j.
- Must show that fail[j] points to longest suffix that is also a prefix of some word in P.
- Look at the article...

AC matching time complexity

- **Theorem** For matching the $M_p$ on text $S$, $|S|=n$, less than 2n transitions within $M$ are made.
- **Proof** Compare to KMP.
- There is at most n goto steps.
- Cannot be more than n Fail-steps.
- In total -- there can be less than 2n transitions in $M$.

Individual node (goto)

- Full table
- List
- Binary search tree(?)
- Some other index?

AC thoughts

- Scales for many strings simultaneously.
- For very many patterns -- search time (of grep) improves(??) 
  - See Wu-Manber article
- When k grows, then more fail[] transitions are made (why?)
- But always less than n.
- If all goto[a] are indexed in an array, then the size is $|M_p|^{|E|}$, and the running time of AC is $O(n)$.
- When k and c are big, one can use lists or trees for storing transition functions.
- Then, $O(n \log(\min(k, c)))$.

Advanced AC

- Precalculate the next state transition correctly for every possible character in alphabet
- Can be good for short patterns

Problems of AC?

- Need to rebuild on adding / removing patterns
- Details of branching on each node(?)
Commentz-Walter

• Generalization of Boyer-Moore for multiple sequence search
• Beate Commentz-Walter

A String Matching Algorithm Fast on the Average

You can download here my algorithm StringMatchingFastOnTheAverage (PDF, ~17.2 MB) or here StringMatchingFastOnTheAverage (extended abstract) (PDF, ~3 MB)

Commentz-Walter [CW79]

• Commentz-Walter [CW79] presented an algorithm for the multi-pattern matching problem that combines the Boyer-Moore technique with the Aho-Corasick algorithm. The Commentz-Walter algorithm is substantially faster than the Aho-Corasick algorithm in practice. Hume [Hu91] designed a tool called gre based on this algorithm, and version 2.0 of fgrep by the GNU project [Ha93] is using it.

• Baeza-Yates [Ba89] also gave an algorithm that combines the Boyer-Moore-Horspool algorithm [Ho80] (which is a slight variation of the classical Boyer-Moore algorithm) with the Aho-Corasick algorithm.

C-W description

• Aho and Corasick [AC75] presented a linear-time algorithm for this problem, based on an automata approach. This algorithm serves as the basis for the UNIX tool fgrep. A linear-time algorithm is optimal in the worst case, but as the regular string-searching algorithm by Boyer and Moore [BM77] demonstrated, it is possible to actually skip a large portion of the text while searching, leading to faster than linear algorithms in the average case.

Idea of C-W

• Build a backward trie of all keywords
• Match from the end until mismatch...
• Determine the shift based on the combination of heuristics

Horspool for many patterns

Search for ATGTATG, TATG, ATAAT, ATGTG

1. Build the trie of the inverted patterns

2. imin=4

3. Table of shifts

4. Start the search

Horspool for many patterns

Search for ATGTATG, TATG, ATAAT, ATGTG

ACGTATG

The text ACATGCTATGTGACA…

ACGT

4 (imin) 2

ACGT

imin=4

2

AT

imin=4 (imin)
Horspool for many patterns
Search for ATGTATG, TATG, ATAAT, ATGTG

The text ACATGCTATGTGACA...

Slides courtesy of Xavier Messeguer Peypoch (http://www.lsi.upc.es/~alggen)

Short Shifts!
What are the possible limitations for C-W?

• Many patterns, small alphabet – minimal skips
• What can be done differently?

Wu-Manber

• Citeseer: http://citeseer.ist.psu.edu/wu94fast.html [Postscript]
• We present a different approach that also uses the ideas of Boyer and Moore. Our algorithm is quite simple, and the main engine of it is given later in the paper. An earlier version of this algorithm was part of the second version of agrep [WM92a, WM92b], although the algorithm has not been discussed in [WM92b] and only briefly in [WM92a]. The current version is used in glimpse [MW94]. The design of the algorithm concentrates on typical searches rather than on worst-case behavior. This allows us to make some engineering decisions that we believe are crucial to making the algorithm significantly faster than other algorithms in practice.

Key idea

• Main problem with Boyer-Moore and many patterns is that, the more there are patterns, the shorter become the possible shifts...
• Wu and Manber: check several characters simultaneously, i.e. increase the alphabet.

Horspool to Wu-Manber

How do we can increase the length of the shifts?

With a table shift of l-mers with the patterns ATGTATG, TATG,ATAAT, ATGTG

1 simbol | 2 simbols
---|---
A | 1
C | 4 (min)
G | 2
T | 1

AT | 1
AG | 3
CA | 3
CC | 3
CG | 3
TA | 2
TG | 2

Experimental length: $\log_2 |\Sigma| 2^{\min \star r}$

Wu-Manber algorithm

Search for ATGTATG, TATG, ATAAT, ATGTG

into the text: ACATGCTATGTGACATAATA

...
Backward Oracle

- Set Backwards oracle SBDM, SBOM
- Pages 68-72

3.4.3 Set Backward Oracle Matching algorithm

The Set Backward Oracle Matching algorithm (SBOM) [AR09] uses a factor oracle of the set of strings. The factor oracle of $P$ recognizes at least all the factors of the strings in $P$. The search algorithm is similar to SDBM. We slide a window of size $\min$ along the text, reading backward a suffix of the window in the factor oracle. If we fall on a letter $\sigma$, we can safely shift the window past $\sigma$. If not, we reach the beginning of the window and verify a subset of $P$ against the text.

3.4.3.1 Factor oracle of a set of strings

The factor oracle construction on a set of strings resembles the Aho-Corasick automaton construction. The only difference appears when going down the supply path looking for an outgoing transition labeled by $\sigma$. In the Aho-Corasick automaton construction, if this transition does not exist, we just jump to the next state on the supply path (Section 3.2.2). In the factor oracle construction, we create in addition a transition labeled by $\sigma$ from each state on the supply path to the state where the original transition leads.

String matching of many patterns

Slides courtesy of Xavier Messeguer Peypoch (http://www.lsi.upc.es/~alggen)

5 strings

10 strings
11.12.15

100 strings

1000 strings

**Factor Oracle**

**Factor Oracle: safe shift**

**Factor Oracle:**

**Factor oracle**
Factor oracle

  - http://portal.acm.org/citation.cfm?id=647009.712672&coll=GUIDE&dl=GUIDE&CFID=31549541&CTOKEN=18116417
  - http://www.igm.univ-mlv.fr/~allauzen/work/softsem.ps

Multiple Shift-AND

- P={P1, P2, P3, P4}. Generalize Shift-AND
  - Bits = P4 P3 P2 P1
  - Start = 1 1 1 1
  - Match = 1 1 1 1
Text Algorithms (6EAP)

Full text indexing

Jaak Vilo
2010 fall

E.g. Dictionary problem

- Does P belong to a dictionary $D=\{d_1, \ldots, d_n\}$
  - Build a binary search tree of $D$
  - $B$-Tree of $D$
  - Hashing
  - Sorting + Binary search
- Build a keyword trie: search in $O(|P|)$
  - Assuming alphabet has up to a constant size $c$
  - See Aho-Corasick algorithm, Trie construction

Problem

- Given $P$ and $S$ – find all exact or approximate occurrences of $P$ in $S$
- You are allowed to preprocess $S$ (and $P$, of course)
- Goal: to speed up the searches

Sorted array and binary search

$O(\log n)$

$O(|P|)$
S != set of words

• S of length n

• How to index?

• Index from every position of a text

• Prefix of every possible suffix is important

Suffix tree

• Definition: A compact representation of a trie corresponding to the suffixes of a given string where all nodes with one child are merged with their parents.

• Definition (suffix tree). A suffix tree T for a string S (with n = |S|) is a rooted, labeled tree with a leaf for each non-empty suffix of S.

• Each internal node, other than the root, has at least two children;

• Each edge leaving a particular node is labeled with a non-empty substring of S from which the first symbol is unique among all first symbols of the edge labels of the edges leaving this particular node;

• For any leaf in the tree, the concatenation of the edge labels on the path from the root to this leaf exactly spells out a non-empty suffix of S.


• CPM articles archive: http://www.cs.ucr.edu/~stelo/cpm/


• STXXL: Standard Template Library for Extra Large Data Sets.

http://stxxl.sourceforge.net/

• ANSI C implementation of a Suffix Tree

http://yeda.cs.technion.ac.il/~yona/suffix_tree/

These are old links – check for newer...

Literature on suffix trees

• http://en.wikipedia.org/wiki/Suffix_tree


• CPM articles archive: http://www.cs.ucr.edu/~stelo/cpm/


The suffix tree Tree(T) of T

- data structure **suffix tree**, Tree(T), is **compacted trie** that represents all the suffixes of string T
- linear size: |Tree(T)| = O(|T|)
- can be constructed in linear time O(|T|)
- has **myriad virtues** (A. Apostolico)
- is well-known: 366 000 Google hits

**Suffix trie and suffix tree**

*abaab*

*baab*

*aabb*

*b*

**Trie(abaab)**

**Tree(abaab)**

---

**Suffix tree and suffix array techniques for pattern analysis in strings**

*Esko Ukkonen*

*Univ Helsinki*

*Erice School 30 Oct 2005*

---

**Algorithms for combinatorial string matching?**

- deep beauty?  +-  
- shallow beauty?  +  
- applications?  ++  
- intensive algorithmic miniatures  
- sources of new problems:  
  *text processing, DNA, music,…*

---

**High-throughput genome-scale sequence analysis and mapping using compressed data structures**

*Veli Mäkinen*

*Department of Computer Science*

*University of Helsinki*

---

*ttttttttttttagacggagtctcgctctg*

*tcgccccaggttgagttcgagtggccggg*

*atctcggtctacgtcaagcttgcgtctcctcc*

*cggttcagccattctctgccctagcc*

*tcrcagtagctggactacccgctg*

*cgccactacgcggctaatatttttttttt*

*ttagtagacggaggtttctaccggtttagc*

*cgggatggtctctgttcacctcactcgtg*

*atccgcccctcgccctcacaagttgacg*

*tgggattacaggcgt*

---

Analysis of a string of symbols

• **T** = h a t t i v a t t i  ‘text’
• **P** = a t t  ‘pattern’

• Find the occurrences of P in T:
  hat t i v a t t i

• Pattern synthesis:
  
  - #(t) = 4
  - #(atti) = 2
  - #(tti**t**t) = 2


Pattern finding & synthesis problems

• **T** = t₁t₂ ... tₙ, **P** = p₁p₂ ... pₙ, strings of symbols in finite alphabet

• Indexing problem: Preprocess T (build an index structure) such that the occurrences of different patterns P can be found fast
  - static text, any given pattern P

• Pattern synthesis problem: Learn from T new patterns that occur surprisingly often

• What is a pattern? Exact substring, approximate substring, with generalized symbols, with gaps, ...

The suffix tree Tree(T) of T

• data structure suffix tree, Tree(T), is a compacted trie that represents all the suffixes of string T

• linear size: |Tree(T)| = O(|T|)

• can be constructed in linear time O(|T|)

• has myriad virtues (A. Apostolico)

• is well-known: 366 000 Google hits

Trie(T) can be large

- $|\text{Trie}(T)| = O(|T|^2)$
- bad example: $T = a^n b^n$
- Trie(T) can be seen as a DFA: language accepted = the suffixes of T
- minimize the DFA $\Rightarrow$ directed cyclic word graph (‘DAWG’)

Tree(T) is of linear size

- only the internal branching nodes and the leaves represented explicitly
- edges labeled by substrings of T
- $v = \text{node}(\alpha)$ if the path from root to v spells $\alpha$
- one-to-one correspondence of leaves and suffixes
- $|T|$ leaves, hence $< |T|$ internal nodes
- $|\text{Tree}(T)| = O(|T| + \text{size(edge labels)})$

Tree(hattivatti)

- substring labels of edges represented as pairs of pointers

Tree(T) is full text index

- P occurs in T at locations 8, 31, ...
- P occurs in T $\Rightarrow$ P is a prefix of some suffix of T
- $\Rightarrow$ Path for P exists in Tree(T)
- All occurrences of P in time $O(|P| + \text{occ})$
### Linear time construction of Tree(T)

- Weiner (1973), ‘algorithm of the year’
- McCreight (1976)
- ‘on-line’ algorithm (Ukkonen 1992)

### On-line construction of Trie(T)

- \( T = t_1 t_2 \ldots t_n \$ \)
- \( P_i = t_1 t_2 \ldots t_i \) : \( i \)th prefix of \( T \)
- on-line idea: update \( \text{Trie}(P_i) \) to \( \text{Trie}(P_{i+1}) \)
- \( \Rightarrow \) very simple construction

### Trie(abaab)

- Add next symbol = b
What happens in $Trie(P_i) => Trie(P_{i+1})$?

On-line procedure for suffix trie

1. Create $Trie(P_1)$: nodes root and $v$, an arc $son(root, t_1) = v$, and suffix links $slink(v) = root$ and $slink(root) = root$
2. for $i = 2$ to $n$ do begin
3. \quad $v_{i-1} :=$ leaf of $Trie(t_1 \ldots t_{i-1})$ for string $t_1 \ldots t_{i-1}$ (i.e., the deepest leaf)
4. \quad $v' := v_{i-1}; v := 0$
5. \quad while node $v$ has no outgoing arc for $t_i$ do begin
6. \quad \quad Create a new node $v'$ and an arc $son(v,t_i) = v'$
7. \quad if $v' \neq 0$ then $slink(v) := v'$
8. \quad $v := slink(v); v' := v'$ end
9. \quad for the node $v''$ such that $v'' = son(v,t_i)$ do
10. \quad \quad if $v'' = v'$ then $slink(v') := root$ else $slink(v') := v''$

Suffix trees on-line

• ‘compacted version’ of the on-line trie construction: simulate the construction on the linear size tree instead of the trie => time $O(|T|)$
• all trie nodes are conceptually still needed => implicit and real nodes
Implicit and real nodes

- Pair \((v, \alpha)\) is an *implicit node* in Tree\((T)\) if \(v\) is a node of Tree and \(\alpha\) is a (proper) prefix of the label of some arc from \(v\). If \(\alpha\) is the empty string then \((v, \alpha)\) is a 'real' node (= \(v\)).
- Let \(v = \text{node}(\alpha')\) in Tree\((T)\). Then implicit node \((v, \alpha)\) represents node\((\alpha' \alpha)\) of Trie\((T)\).

Suffix links and open arcs

On-line procedure for suffix tree

Input: string \(T = t_1 t_2 \ldots t_n\)
Output: Tree\((T)\)

Notation: \(\text{son}(v, \alpha) = w\) if there is an arc from \(v\) to \(w\) with label \(\alpha\).
\(\text{son}(v, \varepsilon) = v\)

Function Canonize\((v, \alpha)\):

\[
\text{while } \text{son}(v, \alpha') \neq 0 \text{ where } \alpha = \alpha' \alpha'' \text{ and } |\alpha'| > 0 \text{ do}
\]

\[
v := \text{son}(v, \alpha') \quad \text{\(\alpha := \alpha''\)}
\]

\[
\text{return } (v, \alpha)
\]

Suffix-tree on-line: main procedure

Create Tree\((T)\); \(\text{slink}(\text{root}) := \text{root}\)
\((v, \alpha) := (\text{root}, \varepsilon)\) \(\Rightarrow (v, \alpha)\) is the start node \(\gamma\)

\[
\text{for } i = 2 \text{ to } n+1 \text{ do}
\]

\[
v' := 0 \quad \text{while there is no arc from } v \text{ with label prefix } \alpha \text{ do}
\]

\[
\text{if } \alpha \neq \varepsilon \text{ then divide the arc } w = \text{son}(v, \alpha) \text{ into two } \gamma^\prime\text{ and } \gamma^\prime'\text{; } \text{son}(v, \alpha) := v^\prime\gamma' \text{ and } \text{son}(v, \alpha') := w
\]

\[
\text{else } \text{son}(v, \alpha) := v^\gamma' \quad \text{v'} := v^\gamma''
\]

\[
\text{if } v' \neq 0 \text{ then } \text{slink}(v') := v'
\]

\[
\text{v'} := v^\gamma' \quad \text{v} := \text{slink}(v); \quad (v, \alpha) := \text{Canonize}(v, \alpha)
\]

\[
\text{if } v' \neq 0 \text{ then } \text{slink}(v') := v
\]

\[
(v, \alpha) := \text{Canonize}(v, \alpha) \quad \Rightarrow (v, \alpha) = \text{start node of the next round } \gamma
\]
The actual time and space

- $|\text{Tree}(T)|$ is about $20|T|$ in practice
- brute-force construction is $O(|T|\log|T|)$ for random strings as the average depth of internal nodes is $O(\log|T|)$
- difference between linear and brute-force constructions not necessarily large (Giegerich & Kurtz)
- truncated suffix trees: $k$ symbols long prefix of each suffix represented (Na et al. 2003)
- alphabet independent linear time (Farach 1997)

Applications of Suffix Trees


- **APL1**: Exact String Matching Search for P from text S. Solution 1: build STree(S) - one achieves the same O(n+m) as Knuth-Morris-Pratt, for example!
- **APL2**: Exact set matching Search for a set of patterns P

- **APL3**: substring problem for a database of patterns
  - Given a set of strings S=S1, ... , Sn --- a database Find all Si that have P as a substring
  - Generalized suffix tree contains all suffixes of all Si
  - Query in time O(|P|), and can identify the LONGEST common prefix of P in all Si

- **APL4**: Longest common substring of two strings
  - Find the longest common substring of S and T.
  - Overall there are potentially O(n^2) such substrings, if n is the length of a shorter of S and T
  - Donald Knuth once (1970) **conjectured** that linear-time algorithm is **impossible**.
  - Solution: construct the STree(S+T) and find the node deepest in the tree that has suffixes from both S and T in subtree leaves.
  - Ex: S= *superiorcalifornialives* T= *sealiver* have both a substring xxxxx.
Simple analysis task: LCSS

- Let $\text{LCSS}(A,B)$ denote the longest common substring two sequences $A$ and $B$. E.g.:
  - $\text{LCSS(AGATCTATCT,CGCTCTATG)}=TCTAT$.
- A good solution is to build suffix tree for the shorter sequence and make a **descending suffix walk** with the other sequence.

**Descending suffix walk**

Read B left-to-right, always going down in the tree when possible. If the next symbol of B does not match any edge label on current position, take suffix link, and try again. (Suffix link in the root to itself emits a symbol). The node $v$ encountered with largest string depth is the solution.

**Generalized suffix tree application**

node info:
- subtree size 4398
- blue sequences 12/15
- red sequences 2/62

**Case study continued**

- 5 blue
- 1 red
- regions with ChIP-seq matches
Properties of suffix tree

- Suffix tree has \( n \) leaves and at most \( n-1 \) internal nodes, where \( n \) is the total length of all sequences indexed.
- Each node requires constant number of integers (pointers to first child, sibling, parent, text range of incoming edge, statistics counters, etc.).
- Can be constructed in linear time.

Applications of Suffix Trees

- **APL4**: Longest common substring of two strings
- Find the longest common substring of \( S \) and \( T \).
- Overall there are potentially \( O(n^2) \) such substrings, if \( n \) is the length of a shorter of \( S \) and \( T \).
- Donald Knuth once (1970) conjectured that linear-time algorithm is impossible.
- Solution: construct the STree\((S+T)\) and find the node deepest in the tree that has suffixes from both \( S \) and \( T \) in subtree leaves.
- Ex: \( S= \) superiorcalifornialives \( T= \) sealiver have both a substring alive.

Properties of suffix tree... in practice

- Huge overhead due to pointer structure:
  - Standard implementation of suffix tree for human genome requires over 200 GB memory!
  - A careful implementation (using \( \log n \)-bit fields for each value and array layout for the tree) still requires over 40 GB.
  - Human genome itself takes less than 1 GB using 2-bits per bp.

Applications of Suffix Trees

- **APL5**: Recognizing DNA contamination Related to DNA sequencing, search for longest strings (longer than threshold) that are present in the \( DB \) of sequences of other genomes.
- **APL6**: Common substrings of more than two strings Generalization of APL4, can be done in linear (in total length of all strings) time

Applications of Suffix Trees

- **APL7**: Building a directed graph for exact matching: **Suffix graph** - directed acyclic word graph (DAWG), a smallest finite state automaton recognizing all suffixes of a string \( S \). This automaton can recognize membership, but not tell which suffix was matched.
- Construction: merge isomorphic subtrees.
- Isomorphic in Suffix Tree when exists suffix link path, and subtrees have equal nr. of leaves.

Applications of Suffix Trees

- **APL8**: A reverse role for suffix trees, and major space reduction Index the pattern, not tree...
- Matching statistics.
- **APL10**: All-pairs suffix-prefix matching For all pairs \( S_i, S_j \) find the longest matching suffix-prefix pair. Used in shortest common superstring generation (e.g. DNA sequence assembly), EST alignment etc.
Applications of Suffix Trees

• APL11: Finding all maximal repetitive structures in linear time
• APL12: Circular string linearization e.g. circular chemical molecules in the database, one wants to linearize them in a canonical way...
• APL13: Suffix arrays - more space reduction will touch that separately

Applications of Suffix Trees

• APL14: Suffix trees in genome-scale projects
• APL15: A Boyer-Moore approach to exact set matching
• APL16: Ziv-Lempel data compression
• APL17: Minimum length encoding of DNA

Applications of Suffix Trees

• Additional applications Mostly exercises...
• Extra feature: CONSTANT time lowest common ancestor retrieval (LCA)
  Andmestrutuur mis võimaldab leida konstantse ajaga alumist (ühist vanemat (see vastab plikimale ihisele prefixile)) on võimalik koostada lineaarse ajaga.
• APL: Longest common extension: a bridge to inexact matching
• APL: Finding all maximal palindromes in linear time
  Palindrome reads from central position the same to left and right. E.g.: kirik, saippuakivikauppias.
  - Build the suffix tree of S and inverted S (aabcbad => aabcbad#dbacbaa ) and using the LCA one can ask for any position pair (i, 2i-1), the longest common prefix in constant time.
  - The whole problem can be solved in O(n).

Applications of Suffix Trees

• APL: Exact matching with wild cards
• APL: The k-mismatch problem
• Approximate palindromes and repeats
• Faster methods for tandem repeats
• A linear-time solution to the multiple common substring problem
• And many-many more ...
Suffix array: example

- suffix array = lexicographic order of the suffixes

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>hattivatti</td>
<td>1</td>
</tr>
<tr>
<td>attivatti</td>
<td>2</td>
</tr>
<tr>
<td>tivatti</td>
<td>3</td>
</tr>
<tr>
<td>ivatti</td>
<td>4</td>
</tr>
<tr>
<td>vatti</td>
<td>5</td>
</tr>
<tr>
<td>atti</td>
<td>6</td>
</tr>
<tr>
<td>tti</td>
<td>7</td>
</tr>
<tr>
<td>i</td>
<td>8</td>
</tr>
<tr>
<td>ti</td>
<td>9</td>
</tr>
<tr>
<td>i</td>
<td>10</td>
</tr>
<tr>
<td>ε</td>
<td>11</td>
</tr>
</tbody>
</table>

Suffix array construction: sort!

- suffix array = lexicographic order of the suffixes

Suffix array

- suffix array \( SA(T) \) = an array giving the lexicographic order of the suffixes of \( T \)
- space requirement: \( 5|T| \)
- practitioners like suffix arrays (simplicity, space efficiency)
- theoreticians like suffix trees (explicit structure)

Reducing space: suffix array

Pattern search from suffix array

- Many algorithms on suffix tree can be simulated using suffix array...
  - ... and couple of additional arrays...
  - ... forming so-called enhanced suffix array...
  - ... leading to the similar space requirement as careful implementation of suffix tree
- Not a satisfactory solution to the space issue.
What we learn today?

- We learn that it is possible to replace suffix trees with *compressed suffix trees* that take 8.8 GB for the human genome.
- We learn that *backtracking* can be done using *compressed suffix arrays* requiring only 2.1 GB for the human genome.
- We learn that *discovering* interesting motif seeds from the human genome takes 40 hours and requires 9.3 GB space.

Recent suffix array constructions

- Manber&Myers (1990): $O(|T| \log |T|)$
- linear time via suffix tree
- January / June 2003: direct linear time construction of suffix array
  - Kim, Sim, Park, Park (CPM03)
  - Kärkkäinen & Sanders (ICALP03)
  - Ko & Aluru (CPM03)

Kärkkäinen-Sanders algorithm

1. Construct the suffix array of the suffixes starting at positions $i \mod 3 \neq 0$. This is done by reduction to the suffix array construction of a string of two thirds the length, which is solved recursively.
2. Construct the suffix array of the remaining suffixes using the result of the first step.
3. Merge the two suffix arrays into one.

Running example

```
T[0,n) = y a b b a d a b b a d o 0 0 ...
SA = (12,1,6,4,9,3,8,2,7,5,10,11,0)
```

Notation

- string $T = T[0,n) = t_0 t_1 ... t_{n-1}$
- suffix $S_i = T[i,0) = t_i t_{i+1} ... t_{n-1}$
- for $C \subset [0,n]$: $S_C = \{S_i \mid i \in C\}$

$\text{suffix array } SA[0,n] \text{ of } T \text{ is a permutation of } [0,n] \text{ satisfying } S_{SA[0]} < S_{SA[1]} < ... < S_{SA[n]}$

Step 0: Construct a sample

- for $k = 0,1,2$
  
  $B_k = \{i \in [0,n] \mid i \mod 3 = k\}$
- $C = B_1 \cup B_2 \text{ sample positions}$
- $S_C \text{ sample suffixes}$

- Example: $B_1 = \{1,4,7,10\}$, $B_2 = \{2,5,8,11\}$, $C = \{1,4,7,10,2,5,8,11\}$
**Step 1: Sort sample suffixes**

- for \( k = 1, 2 \), construct 
  \[ R_k = [t_k t_{k+1} t_{k+2}] [t_{k+3} t_{k+4} t_{k+5}] \ldots [t_{\text{max}B_k} t_{\text{max}B_k+1} t_{\text{max}B_k+2}] \]
  \[ R = R_1 \land R_2 \] concatenation of \( R_1 \) and \( R_2 \)

Suffixes of \( R \) correspond to \( S_\mathcal{C} \); suffix \([t_i t_{i+1} \ldots] \) corresponds to \( S_i \); correspondence is order preserving.

Sort the suffixes of \( R \): radix sort the characters and rename with ranks to obtain \( R' \). If all characters different, their order directly gives the order of suffixes. Otherwise, sort the suffixes of \( R' \) using Kärkkäinen-Sanders. Note: \(| R' | = 2n/3 \).

**Step 1 (cont.)**

- once the sample suffixes are sorted, assign a rank to each: \( \text{rank}(S_i) = \text{the rank of } S_i \text{ in } S_\mathcal{C}; + \text{rank}(S_{n+1}) = \text{rank}(S_{n+2}) = 0 \)
- Example:
  \[ R = [\text{abb}][\text{ada}][\text{bba}][\text{do0}][\text{bba}][\text{dab}][\text{bad}][\text{a00}] \]
  \[ R' = (1, 2, 4, 6, 4, 5, 3, 7) \]
  \[ \text{SA}_{R'} = (8, 0, 1, 6, 4, 2, 5, 3, 7) \]
  \[ \text{rank}(S_i) \begin{array}{c} -1 \end{array} 4 \begin{array}{c} -2 \end{array} 6 \begin{array}{c} -5 \end{array} 3 \begin{array}{c} -7 \end{array} 8 \begin{array}{c} -0 \end{array} \]

**Step 2: Sort nonsample suffixes**

- for each non-sample \( S_i \in S_{\mathcal{B}0} \) (note that \( \text{rank}(S_{i+1}) \) is always defined for \( i \in \mathcal{B}0 \)):
  \[ S_i \leq S_j \iff (t_i, \text{rank}(S_{i+1})) \leq (t_j, \text{rank}(S_{j+1})) \]
  - radix sort the pairs \((t_i, \text{rank}(S_{i+1}))\).

  - Example: \( S_{12} < S_8 < S_3 < S_0 \) because \((0, 0) < (a, 5) < (a, 7) < (b, 2) < (y, 1)\)

**Step 3: Merge**

- merge the two sorted sets of suffixes using a standard comparison-based merging:
  - to compare \( S_i \in S_\mathcal{C} \) with \( S_j \in S_{\mathcal{B}0} \), distinguish two cases:
    - \( i \in \mathcal{B}1 \): \( S_i \leq S_j \iff (t_i, \text{rank}(S_{i+1})) \leq (t_j, \text{rank}(S_{j+1})) \)
    - \( i \in \mathcal{B}2 \): \( S_i \leq S_j \iff (t_i, t_{i+1}, \text{rank}(S_{i+2})) \leq (t_j, t_{j+1}, \text{rank}(S_{j+2})) \)
  - note that the ranks are defined in all cases!
    - \( S_1 < S_0 \) as \((a, 4) < (a, 5)\) and \( S_3 < S_4 \) as \((b, a, 6) < (b, a, 7)\)

**Running time \( O(n) \)**

- excluding the recursive call, everything can be done in linear time
- the recursion is on a string of length \( 2n/3 \)
- thus the time is given by recurrence 
  \[ T(n) = T(2n/3) + O(n) \]
- hence \( T(n) = O(n) \)

**Implementation**

- about 50 lines of C++
- code available e.g. via Juha Kärkkäinen’s home page
LCP table

- Longest Common Prefix of successive elements of suffix array:
- \( \text{LCP}[i] = \) length of the longest common prefix of suffixes \( S_{SA[i]} \) and \( S_{SA[i+1]} \)
- build inverse array \( SA^{-1} \) from \( SA \) in linear time
- then LCP table from \( SA^{-1} \) in linear time (Kasai et al, CPM2001)

Suffix tree vs suffix array

- suffix tree \( \leftrightarrow \) suffix array + LCP table

Substring motifs of string \( T \)

- string \( T = t_1 \ldots t_n \) in alphabet \( A \).
- Problem: what are the frequently occurring (ungapped) substrings of \( T \)? Longest substring that occurs at least \( q \) times?
- Thm: Suffix tree \( Tree(T) \) gives complete occurrence counts of all substring motifs of \( T \) in \( O(n) \) time (although \( T \) may have \( O(n^2) \) substrings!)

Counting the substring motifs

- internal nodes of \( Tree(T) \) \( \leftrightarrow \) repeating substrings of \( T \)
- number of leaves of the subtree of a node for string \( P = \) number of occurrences of \( P \) in \( T \)


Example - Word of the Day, Fourth

PAT index - by Gaston Gonnet (ta on samuti Maple tarkvara üks loojatest ning hiljem molekulaabioologia tarkvarapaketti väljanõudja)
- PAT index is essentially a suffix array. To save space, indexed only from first character of every word
- XML-tagging (or SGML, at that time!) also indexed
- To mark certain fields of XML, the bit vectors were used.
- Main concern - improve the speed of search on the CD - minimize random accesses.
- For slow medium even 15-20 accesses is too slow...
Counts for the $O(n)$ maximal motifs shown

Finding repeats in DNA

- human chromosome 3
- the first 48,999,930 bases
- 31 min CPU time (8 processors, 4 GB)

- Human genome: $3 \times 10^9$ bases
- Tree(Human Genome) feasible

Longest repeat?

Occurrences at: 28395980, 28401554
Length: 2559

```
ttagggtacatgtgcacaacgtgcaggtttgttacatatgtatacacgtgccatgatggtgtgctgcacccattaactcgtcatttagcgttaggtatatctccgaat
gctatccctcccccctccccccaccccacaacagtccccggtgtgtgatgttccccttcctgtgtccatgtgttctcattgttcaattcccacctatgagtgagaa
catgcggtgtttggttttttgtccttgcgaaagtttgctgagaatgatggtttccagcttcatccatatccctacaaaggacatgaactcatcatttttttatggctgcat
agtattccatggtgtatatgtgccacattttcttaacccagtctacccttgttggacatctgggttggttccaagtctttgctattgtgaatagtgccgcaataaacat
acgtgtgcatgtgtctttatagcagcatgatttataatcctttgggtatatacccagtaatgggatggctgggtcaaatggtatttctagttctagatccctgaggaat
caccacactgacttccacaatggttgaactagtttacagtcccagcaacagttcctatttctccacatcctctccagcacctgttgtttcctgactttttaatgatcgc
cattctaactggtgtgagatggtatctcattgtggttttgatttgcatttctctgatggccagtgatgatgagcattttttcatgtgttttttggctgcataaatgtcttctttt
agagaagtgtctgttcatatccttcgcccacttttgatggggttgtttgtttttttcttgtaaatttgttggagttcattgtagattctgggtattagccctttgtcagatgagt
aggttgcaaaaattttctcccattctgtaggttgcctgttcactctgatggtggtttcttctgctgtgcagaagctctttagtttaattagatcccatttgtcaattttggct
ttttgttgccatagcttttggtgttttagacatgaagtccttgcccatgcctatgtcctgaatggtattgcctaggttttcttctagggtttttatggttttaggtctaacatg
taagtctttaatccatcttgaattaattataaggtgtatattataaggtgtaattataaggtgtataattatatattaattataaggtgtatattaattataaggtgtaagga
gagggatccagtttcagctttctacatatggctagccagttttccctgcaccatttattaaatagggaatcctttccccattgcttgtttttgtcaggtttgtcaaagatc
agatagttgtagatatgcggcattatttctgagggctctgttctgttccattggtctatatctctgttttggtaccagtaccatgctgttttggttactgtagccttgtagt
atagtttgaagtcaggtagcgtgatggttccagctttgttcttttggcttaggattgacttggcaatgtgggctcttttttggttccatatgaactttaaagtagttttttc
caattctgtgaagaaattcattggtagcttgatggggatggcattgaatctataaattaccctgggcagtatggccattttcacaatattgaatcttcctacccatga
gcgtgtactgttcttccatttgtttgtatcctcttttatttcattgagcagtggtttgtagttctccttgaagaggtccttcacatcccttgtaagttggattcctaggtatttt
ttctctttgaagcaattgtgaatgggagttcactcatgatttgactctctgtttgtctgttattggtgtataagaatgcttgtgatttttgcacattgattttgtatcctgag
actttgctgaagttgcttatcagcttaaggagattttgggctgagacgatggggttttctagatatacaatcatgtcatctgcaaacagggacaatttgacttcctcttttcctaattgaatacccgttatttccctctcctgcctgattgccctggccagaacttccaacactatgttgaataggagtggtgagagagggcatccctgtcttgt
gccagttttcaaagggaatgcttccagtttttgtccattcagtatgatattggctgtgggtttgtcatagatagctcttattattttgagatacatcccatcaatacctaa
tttattgagagtttttagcatgaagagttcttgaattttgtcaaaggccttttctgcatcttttgagataatcatgtggtttctgtctttggttctgtttatatgctggagtac
gtttattgattttcgtatgttgaaccagccttgcatcccagggatgaagcccacttgatcatggtggataagctttttgatgtgctgctggattcggtttgccagtattt
tattgaggatttctgcatcgatgttcatcaaggatattggtctaaaattctctttttttgttgtgtctctgtcaggctttggtatcaggatgatgctggcctcataaaatga
gttagg```

Ten occurrences?

Occurrences at: 10130003, 11421803, 18695837, 26652515, 42971130, 47398125
In the reversed complement at: 17858493, 41463059, 42431718, 42580925

Length: 277

```	tttttttttgagacgggttcctgctctgctgcccaggctggagtgcagtggcgggatctcggctctcctgtgcttcagctcgggttcaccgcctctcctgcctcagcctcc
cagagtagctgggactacaggcgcccgccactacgcccggctaattttttgtatttttagttagagacggggtttcaccgttttagccgggatggtctcgatctcctgacctcgtgatccgcccgcctcggcctcccaaagtgctgggattacaggcgt```

Using suffix trees: plagiarism

- find longest common substring of strings $X$ and $Y$
- build Tree($X$Y) and find the deepest node which has a leaf pointing to $X$ and another pointing to $Y$

Using suffix trees: approximate matching

- edit distance: insertions, deletions, changes
- STOCKHOLM vs TUKHOLMA
### String distance/similarity functions

STOCKHOLM vs TUKHOLMA

STOCKHOLM_ _TU_ KHOLMA

=> 2 deletions, 1 insertion, 1 change

### Approximate string matching

- A: STOCKHOLM
- B: TUKHOLMA

- Minimum number of ‘mutation’ steps:
  - $a \rightarrow b$
  - $a \rightarrow e$
  - $e \rightarrow b$...
- $d_{ID}(A, B) = 5$
- $d_{Levenshtein}(A, B) = 4$
- Mutation costs $\Rightarrow$ probabilistic modeling
- Evaluation by dynamic programming $\Rightarrow$ alignment

### Dynamic programming

$$d_{ij} = \min \begin{cases} 
  0 & \text{if } a_i = b_j \\
  d_{i-1,j-1} & \text{else if } d_{ij} \leq \infty \\
  d_{i-1,j} + 1 & \text{else if } d_{ij} > \infty \\
  d_{i,j-1} + 1 & \text{else if } d_{ij} > \infty 
\end{cases}$$

= distance between i-prefix of A and j-prefix of B (substitution excluded)

- $m \times n$ table

### Search problem

- Find approximate occurrences of pattern P in text T: substrings $P'$ of T such that $d(P, P')$ small
- Dynamic programming with small modification: $O(mn)$
- Lots of (practical) improvement tricks

### Index for approximate searching

- Dynamic programming: $P \times \text{Tree}(T)$ with backtracking