We can further improve string binary search using precomputed information about the lcp's between the strings in $\mathcal{R}$.

Consider again the basic situation during string binary search:

- We want to compare P and S_{mid}.
- We have already compared P against S_{left} and S_{right}, and we know $lcp(S_{left}, P)$ and $lcp(P, S_{right})$.

The values $left$ and $right$ are fully determined by mid independently of P. That is, P only determines whether the search ends up at position mid at all, but if it does, $left$ and $right$ are always the same.

Thus, we can precompute and store the values

$$LLCP[mid] = lcp(S_{left}, S_{mid})$$
$$RLCP[mid] = lcp(S_{mid}, S_{right})$$

Now we know all lcp values between P, S_{left}, S_{mid}, S_{right} except $lcp(P, S_{mid})$. The following lemma shows how to utilize this.

Lemma 1.34: Let A, B, B' and C be strings such that $A \le B \le C$ and $A \le B' \le C$.

(a) If $lcp(A,B) > lcp(A,B')$, then $B < B'$ and $lcp(B,B') = lcp(A,B')$.
(b) If $lcp(A,B) < lcp(A,B')$, then $B > B'$ and $lcp(B,B') = lcp(A,B)$.
(c) If $lcp(B,C) > lcp(B',C)$, then $B > B'$ and $lcp(B,B') = lcp(B',C)$.
(d) If $lcp(B,C) < lcp(B',C)$, then $B < B'$ and $lcp(B,B') = lcp(B,C)$.
(e) If $lcp(A,B) = lcp(A,B')$ and $lcp(B,C) = lcp(B',C)$, then $lcp(B,B') \ge \max\{lcp(A,B), lcp(B,C)\}$.

Proof. Cases (a)–(d) are symmetrical, we show (a). $B < B'$ follows from Lemma 1.26. Then by Lemma 1.25, $lcp(A,B') = \min\{lcp(A,B), lcp(B,B')\}$. Since $lcp(A,B') < lcp(A,B)$, we must have $lcp(A,B') = lcp(B,B')$.

In case (e), we use Lemma 1.25:

$$lcp(B,B') \ge \min\{lcp(A,B), lcp(A,B')\} = lcp(A,B)$$
$$lcp(B,B') \ge \min\{lcp(B,C), lcp(B',C)\} = lcp(B,C)$$

Thus $lcp(B,B') \ge \max\{lcp(A,B), lcp(B,C)\}$. □

Algorithm 1.35: String binary search (with precomputed lcps)
Input: Ordered string set $\mathcal{R} = \{S_1, S_2, \ldots, S_n\}$, arrays LLCP and RLCP, query string P.
Output: The number of strings in $\mathcal{R}$ that are smaller than P.

```
(1)  left ← 0; right ← n + 1
(2)  llcp ← 0; rlcp ← 0
(3)  while right − left > 1 do
(4)      mid ← ⌊(left + right)/2⌋
(5)      if LLCP[mid] > llcp then left ← mid
(6)      else if LLCP[mid] < llcp then right ← mid; rlcp ← LLCP[mid]
(7)      else if RLCP[mid] > rlcp then right ← mid
(8)      else if RLCP[mid] < rlcp then left ← mid; llcp ← RLCP[mid]
(9)      else
(10)         mlcp ← max{llcp, rlcp}
(11)         (x, mlcp) ← LcpCompare(S_mid, P, mlcp)
(12)         if x = "<" then left ← mid; llcp ← mclp
(13)         else right ← mid; rlcp ← mclp
(14) return left
```


Theorem 1.36: An ordered string set $\mathcal{R} = \{S_1, S_2, \ldots, S_n\}$ can be preprocessed in $\mathcal{O}(\Sigma LCP(\mathcal{R}) + n)$ time and $\mathcal{O}(n)$ space so that a binary search with a query string P can be executed in $\mathcal{O}(|P| + \log n)$ time.

Proof. The values $LLCP[mid]$ and $RLCP[mid]$ can be computed in $\mathcal{O}(lcp(S_{mid}, \mathcal{R} \setminus \{S_{mid}\}) + 1)$ time. Thus the arrays $LLCP$ and $RLCP$ can be computed in $\mathcal{O}(\Sigma lcp(\mathcal{R}) + n) = \mathcal{O}(\Sigma LCP(\mathcal{R}) + n)$ time and stored in $\mathcal{O}(n)$ space.

The main while loop in Algorithm 1.35 is executed $\mathcal{O}(\log n)$ times and everything except LcpCompare on line (11) needs constant time.

If a given LcpCompare call performs $t + 1$ symbol comparisons, $mclp$ increases by t on line (11). Then on lines (12)–(13), either $llcp$ or $rlcp$ increases by at least t, since $mlcp$ was $\max\{llcp, rlcp\}$ before LcpCompare. Since $llcp$ and $rlcp$ never decrease and never grow larger than $|P|$, the total number of extra symbol comparisons in LcpCompare during the binary search is $\mathcal{O}(|P|)$. □

String Binary Search Trees

Binary search can be seen as a search on an implicit binary search tree, where the middle element is the root, the middle elements of the first and second half are the children of the root, etc.. The string binary search technique can be extended for arbitrary binary search trees.

- Let S_v be the string stored at a node v in a binary search tree. Let $S_<$ and $S_>$ be the closest lexicographically smaller and larger strings stored at ancestors of v.
- The comparison of a query string P and the string S_v is done the same way as the comparison of P and S_{mid} in string binary search. The roles of S_{left} and S_{right} are taken by $S_<$ and $S_>$.
- If each node v stores the values $lcp(S_<, S_v)$ and $lcp(S_v, S_>)$, then a search in a balanced search tree can be executed in $\mathcal{O}(|P| + \log n)$ time. Other operations including insertions and deletions take $\mathcal{O}(|P| + \log n)$ time too.

Hashing and Fingerprints

Hashing is a powerful technique for dealing with strings based on mapping each string to an integer using a hash function:

$$H : \Sigma^* \to [0..q) \subset \mathbb{N}$$

The most common use of hashing is with hash tables. Hash tables come in many flavors that can be used with strings as well as with any other type of object with an appropriate hash function. A drawback of using a hash table to store a set of strings is that they do not support lcp and prefix queries.

Hashing is also used in other situations, where one needs to check whether two strings S and T are the same or not:

- If $H(S) \neq H(T)$, then we must have $S \neq T$.
- If $H(S) = H(T)$, then $S = T$ and $S \neq T$ are both possible. If $S \neq T$, this is called a collision.

When used this way, the hash value is often called a fingerprint, and its range $[0..q)$ is typically large as it is not restricted by a hash table size.

Any good hash function must depend on all characters. Thus computing $H(S)$ needs $\Omega(|S|)$ time, which can defeat the advantages of hashing:

- A plain comparison of two strings is faster than computing the hashes.
- The main strength of hash tables is the support for constant time insertions and deletions, but inserting a string S into a hash table needs $\Omega(|S|)$ time when the hash computation time is included. Compare this to the $\mathcal{O}(|S|)$ time for a trie under a constant alphabet and the $\mathcal{O}(|S| + \log n)$ time for a ternary trie.

However, a hash table can still be competitive in practice. Furthermore, there are situations, where a full computation of the hash function can be avoided:

- A hash value can be computed once, stored, and used many times.
- Some hash functions can be computed more efficiently for a related set of strings. An example is the Karp–Rabin hash function.

Definition 1.37: The Karp–Rabin hash function for a string $S = s_0 s_1 \ldots s_{m-1}$ over an integer alphabet is

$$H(S) = (s_0 r^{m-1} + s_1 r^{m-2} + \cdots + s_{m-2} r + s_{m-1}) \bmod q$$

for some fixed positive integers q and r.

Lemma 1.38: For any two strings A and B,

$$H(AB) = (H(A) \cdot r^{|B|} + H(B)) \bmod q$$
$$H(B) = (H(AB) - H(A) \cdot r^{|B|}) \bmod q$$

Proof. Without the modulo operation, the result would be obvious. The modulo does not interfere because of the rules of modular arithmetic:

$$(x + y) \bmod q = ((x \bmod q) + (y \bmod q)) \bmod q$$
$$(xy) \bmod q = ((x \bmod q)(y \bmod q)) \bmod q$$

□

Thus we can quickly compute $H(AB)$ from $H(A)$ and $H(B)$, and $H(B)$ from $H(AB)$ and $H(A)$. We will see applications of this later.

If q and r are *coprime*, then r has a multiplicative inverse r^{-1} modulo q, and we can also compute $H(A) = ((H(AB) - H(B)) \cdot (r^{-1})^{|B|}) \bmod q$.

The parameters q and r have to be chosen with some care to ensure that collisions are rare for any reasonable set of strings.

- The original choice is $r = \sigma$ and q is a large prime.
- Another possibility is that q is a power of two and r is a small prime ($r = 37$ has been suggested). This is faster in practice, because the slow modulo operations can be replaced by bitwise shift operations. If $q = 2^w$, where w is the machine word size, the modulo operations can be omitted completely.
- If q and r were both powers of two, then only the last $\lceil (\log q)/\log r \rceil$ characters of the string would affect the hash value. More generally, q and r should be coprime, i.e, have no common divisors other than 1.
- The hash function can be randomized by choosing q or r randomly. For example, if q is a prime and r is chosen uniformly at random from $[0..q)$, the probability that two strings of length m collide is at most m/q.
- A random choice over a set of possibilities has the additional advantage that we can change the choice if the first choice leads to too many collisions.

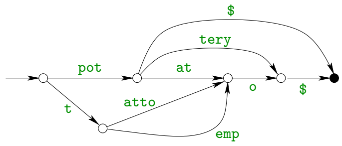
Automata

Finite automata are a well known way of representing sets of strings. In this case, the set is often called a **language**.

A trie is a special type of an automaton.

- Trie is generally not a *minimal* automaton.
- Trie techniques including path compaction and ternary branching can be applied to automata.

Example 1.39: Compacted minimal automaton for $\mathcal{R} = \{\texttt{pot\$}, \texttt{potato\$}, \texttt{pottery\$}, \texttt{tattoo\$}, \texttt{tempo\$}\}$.



Automata are much more powerful than tries in representing languages:

- Infinite languages
- Nondeterministic automata
- Even an acyclic, deterministic automaton can represent a language of exponential size.

Automata do not support all operations of tries:

- Insertions and deletions
- Satellite data, i.e., data associated to each string.

Sets of Strings: Summary

Efficient algorithms and data structures for sets of strings:

- Storing and searching: trie and ternary trie and their compact versions, string binary search and string binary search tree, Karp–Rabin hashing.
- Sorting: string quicksort and mergesort, LSD and MSD radix sort.

Lower bounds:

- Many of the algorithms are optimal.
- General purpose algorithms are asymptotically slower.

The central role of longest common prefixes:

- LCP array $LCP_{\mathcal{R}}$ and its sum $\Sigma LCP(\mathcal{R})$.
- Lcp-comparison technique.

2. Exact String Matching

Let $T = T[0..n)$ be the text and $P = P[0..m)$ the pattern. We say that P occurs in T at position j if $T[j..j + m) = P$.

Example: $P =$ `aine` occurs at position 6 in $T =$ `karjalainen`.

In this part, we will describe algorithms that solve the following problem.

Problem 2.1: Given text $T[0..n)$ and pattern $P[0..m)$, report the first position in T where P occurs, or n if P does not occur in T.

The algorithms can be easily modified to solve the following problems too.

- Existence: Is P a factor of T?
- Counting: Count the number of occurrences of P in T.
- Listing: Report all occurrences of P in T.

The naive, brute force algorithm compares P against $T[0..m)$, then against $T[1..1 + m)$, then against $T[2..2 + m)$ etc. until an occurrence is found or the end of the text is reached.

Algorithm 2.2: Brute force
Input: text $T = T[0 \dots n)$, pattern $P = P[0 \dots m)$
Output: position of the first occurrence of P in T

(1) $i \leftarrow 0; j \leftarrow 0$
(2) while $i < m$ and $j < n$ do
(3)   if $P[i] = T[j]$ then $i \leftarrow i + 1; j \leftarrow j + 1$
(4)   else $j \leftarrow j - i + 1; i \leftarrow 0$
(5) if $i = m$ then return $j - m$ else return n

The worst case time complexity is $\mathcal{O}(mn)$. This happens, for example, when $P = \texttt{a}^{m-1}\texttt{b} = \texttt{aaa..ab}$ and $T = \texttt{a}^n = \texttt{aaaaaa..aa}$.

Knuth–Morris–Pratt

The Brute force algorithm forgets everything when it moves to the next text position.

The Morris–Pratt (MP) algorithm remembers matches. It never goes back to a text character that already matched.

The Knuth–Morris–Pratt (KMP) algorithm remembers mismatches too.

Example 2.3:

Brute force	Morris–Pratt	Knuth–Morris–Pratt
`ainaisesti-ainainen`	`ainaisesti-ainainen`	`ainaisesti-ainainen`
`ainai~~n~~en` (6 comp.)	`ainai~~n~~en` (6)	`ainai~~n~~en` (6)
`~~a~~inainen` (1)	`ai~~n~~ainen` (1)	`~~a~~inainen` (1)
`~~a~~inainen` (1)	`~~a~~inainen` (1)	
`ai~~n~~ainen` (3)		
`~~a~~inainen` (1)		
`~~a~~inainen` (1)		

MP and KMP algorithms never go backwards in the text. When they encounter a mismatch, they find another pattern position to compare against the same text position. If the mismatch occurs at pattern position i, then $fail[i]$ is the next pattern position to compare.

The only difference between MP and KMP is how they compute the failure function $fail$.

Algorithm 2.4: Knuth–Morris–Pratt / Morris–Pratt
Input: text $T = T[0 \dots n)$, pattern $P = P[0 \dots m)$
Output: position of the first occurrence of P in T

(1) compute $fail[0..m]$
(2) $i \leftarrow 0; j \leftarrow 0$
(3) while $i < m$ and $j < n$ do
(4)   if $i = -1$ or $P[i] = T[j]$ then $i \leftarrow i + 1; j \leftarrow j + 1$
(5)   else $i \leftarrow fail[i]$
(6) if $i = m$ then return $j - m$ else return n

- $fail[i] = -1$ means that there is no more pattern positions to compare against this text positions and we should move to the next text position.
- $fail[m]$ is never needed here, but if we wanted to find all occurrences, it would tell how to continue after a full match.

We will describe the MP failure function here. The KMP failure function is left for the exercises.

- When the algorithm finds a mismatch between $P[i]$ and $T[j]$, we know that $P[0..i) = T[j-i..j)$.
- Now we want to find a new $i' < i$ such that $P[0..i') = T[j-i'..j)$. Specifically, we want the largest such i'.
- This means that $P[0..i') = T[j-i'..j) = P[i-i'..i)$. In other words, $P[0..i')$ is the longest proper border of $P[0..i)$.

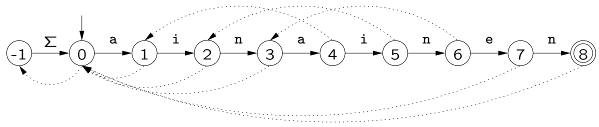
Example: `ai` is the longest proper border of `ainai`.

- Thus $fail[i]$ is the length of the longest proper border of $P[0..i)$.
- $P[0..0) = \varepsilon$ has no proper border. We set $fail[0] = -1$.

Example 2.5: Let $P =$ `ainainen`.

i	$P[0..i)$	border	$fail[i]$
0	ε	–	-1
1	a	ε	0
2	ai	ε	0
3	ain	ε	0
4	aina	a	1
5	ainai	ai	2
6	ainain	ain	3
7	ainaine	ε	0
8	ainainen	ε	0

The (K)MP algorithm operates like an automaton, since it never moves backwards in the text. Indeed, it can be described by an automaton that has a special failure transition, which is an ε-transition that can be taken only when there is no other transition to take.



An efficient algorithm for computing the failure function is very similar to the search algorithm itself!

- In the MP algorithm, when we find a match $P[i] = T[j]$, we know that $P[0..i] = T[j-i..j]$. More specifically, $P[0..i]$ is the longest prefix of P that matches a suffix of $T[0..j]$.
- Suppose $T = \#P[1..m)$, where $\#$ is a symbol that does not occur in P. Finding a match $P[i] = T[j]$, we know that $P[0..i]$ is the longest prefix of P that is a proper suffix of $P[0..j]$. Thus $fail[j+1] = i+1$.

Algorithm 2.6: Morris–Pratt failure function computation
Input: pattern $P = P[0 \dots m)$
Output: array $fail[0..m]$ for P

(1) $i \leftarrow -1; j \leftarrow 0;\ fail[j] \leftarrow i$
(2) while $j < m$ do
(3)   if $i = -1$ or $P[i] = P[j]$ then $i \leftarrow i+1; j \leftarrow j+1;\ fail[j] \leftarrow i$
(4)   else $i \leftarrow fail[i]$
(5) return $fail$

- When the algorithm reads $fail[i]$ on line 4, $fail[i]$ has already been computed.

Theorem 2.7: Algorithms MP and KMP preprocess a pattern in time $\mathcal{O}(m)$ and then search the text in time $\mathcal{O}(n)$.

Proof. We show that the text search requires $\mathcal{O}(n)$ time. Exactly the same argument shows that pattern preprocessing needs $\mathcal{O}(m)$ time.

It is sufficient to count the number of comparisons that the algorithms make. After each comparison $P[i] = T[j]$, one of the two conditional branches is executed:

then Here j is incremented. Since j never decreases, this branch can be taken at most $n+1$ times.

else Here i decreases since $fail[i] < i$. Since i only increases in the then-branch, this branch cannot be taken more often than the then-branch.

□

Shift-And (Shift-Or)

When the MP algorithm is at position j in the text T, it computes the longest prefix of the pattern $P[0..m)$ that is a suffix of $T[0..j]$. The Shift-And algorithm computes all prefixes of P that are suffixes of $T[0..j]$.

- The information is stored in a bitvector D of length m, where $D.i = 1$ if $P[0..i] = T[j-i..j]$ and $D.i = 0$ otherwise. ($D.0$ is the least significant bit.)
- When $D.(m-1) = 1$, we have found an occurrence.

The bitvector D is updated at each text position j:

- There are precomputed bitvectors $B[c]$, for all $c \in \Sigma$, where $B[c].i = 1$ if $P[i] = c$ and $B[c].i = 0$ otherwise.
- D is updated in two steps:
  1. $D \leftarrow (D << 1) + 1$ (the bitwise shift). Now D tells, which prefixes would match if $T[j]$ would match every character.
  2. $D \leftarrow D\ \&\ B[T[j]]$ (the bitwise and). Remove the prefixes where $T[j]$ does not match.

Let w be the wordsize of the computer, typically 64. Assume first that $m \le w$. Then each bitvector can be stored in a single integer.

Algorithm 2.8: Shift-And
Input: text $T = T[0 \dots n)$, pattern $P = P[0 \dots m)$
Output: position of the first occurrence of P in T
Preprocess:
  (1) for $c \in \Sigma$ do $B[c] \leftarrow 0$
  (2) for $i \leftarrow 0$ to $m-1$ do $B[P[i]] \leftarrow B[P[i]] + 2^i$ // $B[P[i]].i \leftarrow 1$
Search:
  (3) $D \leftarrow 0$
  (4) for $j \leftarrow 0$ to $n-1$ do
  (5) $D \leftarrow ((D << 1) + 1)\ \&\ B[T[j]]$
  (6)   if $D\ \&\ 2^{m-1} \neq 0$ then return $j - m + 1$ // $D.(m-1) = 1$
  (7) return n

Shift-Or is a minor optimization of Shift-And. It is the same algorithm except the roles of 0's and 1's in the bitvectors have been swapped. Then & on line 5 is replaced by | (bitwise or). The advantage is that we don't need that "+1" on line 5.

Example 2.9: $P =$ `assi`, $T =$ `apassi`, bitvectors are columns.

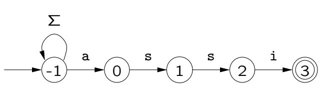
$B[c]$, $c \in \{$`a`,`i`,`p`,`s`$\}$

	a	i	p	s
a	1	0	0	0
s	0	0	0	1
s	0	0	0	1
i	0	1	0	0

D at each step

		a	p	a	s	s	i
a	0	1	0	1	0	0	0
s	0	0	0	0	1	0	0
s	0	0	0	0	0	1	0
i	0	0	0	0	0	0	1

The Shift-And algorithm can also be seen as a bitparallel simulation of the nondeterministic automaton that accepts a string ending with P.



After processing $T[j]$, $D.i = 1$ if and only if there is a path from the initial state (state -1) to state i with the string $T[0..j]$.

On an integer alphabet when $m \le w$:

- Preprocessing time is $\mathcal{O}(\sigma + m)$.
- Search time is $\mathcal{O}(n)$.

If $m > w$, we can store the bitvectors in $\lceil m/w \rceil$ machine words and perform each bitvector operation in $\mathcal{O}(\lceil m/w \rceil)$ time.

- Preprocessing time is $\mathcal{O}(\sigma \lceil m/w \rceil + m)$.
- Search time is $\mathcal{O}(n \lceil m/w \rceil)$.

If no pattern prefix longer than w matches a current text suffix, then only the least significant machine word contains 1's. There is no need to update the other words; they will stay 0.

- Then the search time is $\mathcal{O}(n)$ on average.

Algorithms like Shift-And that take advantage of the implicit parallelism in bitvector operations are called bitparallel.