1 String Matching

1.1 definitions

- $\sum$ = alphabet
- X, Y, Z = strings over $\sum$
- XY = string X concatenated with string Y
- $X^3 = XXX$
- $|X|$ = length of string X
- $Y(i)$ = ith element of Y
- $n = |Y|$
- $m = |X|$

1.2 string matching problem

- input: pattern X, and string Y
- output: does X occur in Y? $\exists Z, W : ZXW = Y$

The obvious solution is to check for X starting at each $Y(i)$, giving $O(nm)$

Less obvious solutions by KMP(Knuth, Morris, and Pratt) and BM (Boyer, Moore) are $O(m+n)$ proceed in 2 phases:

1. analyze X so that a failure to match at point $Y(j)$ starting at $Y(i)$ will allow you to restart the pattern match later in Y than $Y(i+1)$
2. Do the pattern match.

1.3 more definitions

- X a period of Y if Y is a prefix of X^∞ and $|X| < |Y|$
- Y is periodic if $P \leq \frac{m}{2}$

1.4 periodicity lemma

If Y has periods p, q with $p + q \leq m$ then Y has period $GCD(p, q)$.

1.4.1 proof:

if $p = q$ done.

else, assume $p > q$

claim: y has period $p - q$ or $Y(i) = Y(i + (p - q))$ for $1 \leq i \leq m - (p - q)$

proof:

1. $Y(i) = Y(i+p)$ for $1 \leq i \leq m-p$
2. $Y(i) = Y(i-q)$ for $q+1 \leq i \leq m$
3. 1+2 $\Rightarrow Y(i-q) = Y(i+p-q)$ for $q+1 \leq i \leq m-p+q$
4. 2+3 $\Rightarrow Y(i) = Y(i-q) = Y(i+(p-q))$ for $q+1 \leq i \leq m-(p-q)$
5. You can also get: $Y(i) = Y(i+p) = Y(i+p-q)$ for $1 \leq i \leq m-p$

We just need $m-p \geq q$ to satisfy $p+q \leq m$. Now, apply Euclids theorem to get that Y has period $GCD(p,q)$

1.5 finding periods

To find the period, P, of Y, just offset Y by P and check for matching overlaps.

1.6 Another lemma:

$|Y| = m$ with minimum period P and Z with $|Z| = n \geq m$ then

1. if Y occurs in Z at i and j then $|i-j| \geq p$
2. if Y occurs in Z at i and $i+d$ where $d \leq m-p$ then p divides d.

1.6.1 proof:

if $i-j \geq m$ then we are done.

The rest of the proof is by pictures.

1.7 The witness array

$|Y| = m$ and is of size $min\{P, \frac{m}{2}\}$

- $\phi_{wit}(1) = 0$
- $\phi_{wit}(i) = k \leq m+1-i$ s.t. $Y(k) \neq Y(i+k-1)$

2 The algorithm

$Duel(i,j)$

input: pattern Y plus a witness array and a text, Z with $|i-j| \leq (p, \frac{m}{2})$

output: return i or j

1. compare Y to Z starting at $Z(i)$ and $Z(j)$. Check the witness point.

We have eliminated one element of a pair of patterns in unit time.

3 text analysis in the nonperiodic case

Input: pattern Y with witness vector, text Z

1. Divide Z into chunks, T_i of size $\frac{m}{2}$
2. In parallel, check if Y starts in T_i using $Duel$.
3. Eliminate all but one occurence by using $Duel$.
4. Check the last case by brute force.

3.1 Timing analyses

- steps 2 and 3 are $T(n,m) = log(m)$ and $W(n,m) = m * \frac{n}{m} = O(n)$
- step 4 $T(n) = log(n)$ and $W(n,m) = \frac{n}{m} * m = n$

So the total is $T(n) = log(n)$ and $W(n) = n$