Positions Within Prefixes

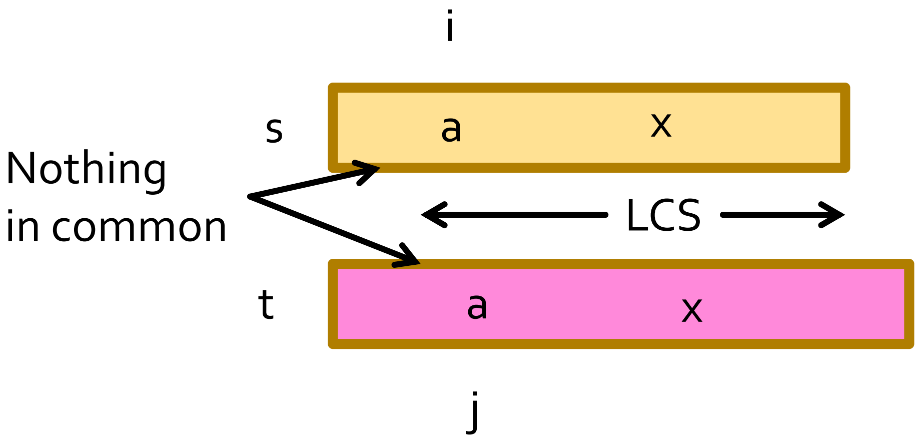
Positions in the Probe and Target Strings

Bounding the Edit Distance

Two-Dimensional Indexes

Exploiting the Position

- If position *i* of probe string *s* is the first position to match a prefix position of string *t*, and it matches position *j*, then the edit distance between *s* and *t* is at least $i + j - 2$.



- The LCS of *s* and *t* is no longer than L-*i* +1, where L is the length of *s*.

Positions/Prefixes – (2)

- If J is the limit on Jaccard distance, then remember E/(E+C) $\leq$ J.
  - E $\geq$ i + j - 2.
  - C $\leq$ L – i + 1.
- Thus, (i + j – 2)/(L + j – 1) $\leq$ J.
- Or, j $\leq$ (JL – J – i +2)/(1 – J).

Positions/Prefixes – Indexing

- Create a 2-attribute index on (symbol, position).
- If string *s* has symbol *a* as the *i* th position of its prefix, add *s* to the bucket (*a*, *i*).
- A B-tree index with keys ordered first by symbol, then position is excellent.

Positions/Prefixes – (3)

- Given probe string *s*, we only need to find a candidate once, so we may as well:

1. Visit positions *i* of *s* in numerical order, and
2. Assume that there have been no matches for earlier positions.
    - That lets us use the upper bound on j when deciding what index buckets we need to look in.

Lookup

- If we want to find matches for probe string *s* of length L, do:

```
for (i=1; i<=J*L+1; i++) {
  let s have a in position i;
  for (j=1;
     j<=(J*L-J-i+2)/(1-J); j++)
     compare s with strings in
     bucket (a, j);
}
```

Example: Lookup

- Suppose J = 0.2.
- Given probe string adegjkmprz, L=10, and the prefix is ade.
- For the i th position of the prefix, we must look at buckets where

$$j \leq (JL - J - i + 2)/(1 - J) = (3.8 - i)/0.8.$$

- For i = 1: $j \leq 3$; for i = 2: $j \leq 2$, and for i = 3: $j \leq 1$.

Example: Lookup – (2)

- Thus, for probe s = adegjkmprz we look in the following buckets: (*a*, 1), (*a*, 2), (*a*, 3), (*d*, 1), (*d*, 2), (*e*, 1).
- Suppose string *t* is in none of these buckets.
- Then the edit distance E is at least 3.
  - Why? Consider where the first common symbol between s and t could be within t.
- The LCS C cannot be longer than s, i.e., 10.
- Thus, J $\geq$ 3/13 > 0.2.