CMSC423: Chapter 9

Introduction to suffix trees

Class so far...

- Deterministic searching (counting, clumps)
- Exact matching (KMP, Z algorithm)
- Randomized searching (Gibbs sampling)
- Branch and bound search (Proteomics)
- Dynamic programming for inexact matching

- This week: exact matching again, for indexing

Stop and think

- Given a text T and pattern P
- Find the longest prefix of P that matches somewhere in T

- Note: KMP solves this for the prefix that is the whole P
- What if the whole of P does not match?

Stop and think...part 2

- Given text T and pattern P
- Find the longest substring of P that matches somewhere in T
- in O(n) time

- Substring – the characters are adjacent (unlike subsequence discussed last week)

- Note: dynamic programming solves the above in O(mn) time (pick the right weights and use local alignment)

Solution...

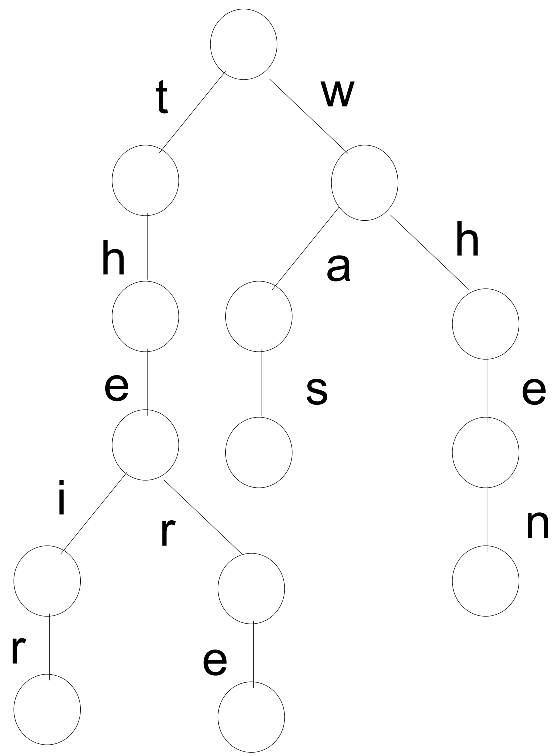
- Note: Donald Knuth did not think O(n) was possible

- Solution:
  - Think of suffixes
  - Each substring is a prefix of a suffix
  - But we know how to solve longest prefix

- How do we organize suffixes?

Many strings: trie

- Basic idea: if many strings share a same sequence only represent it once in the tree



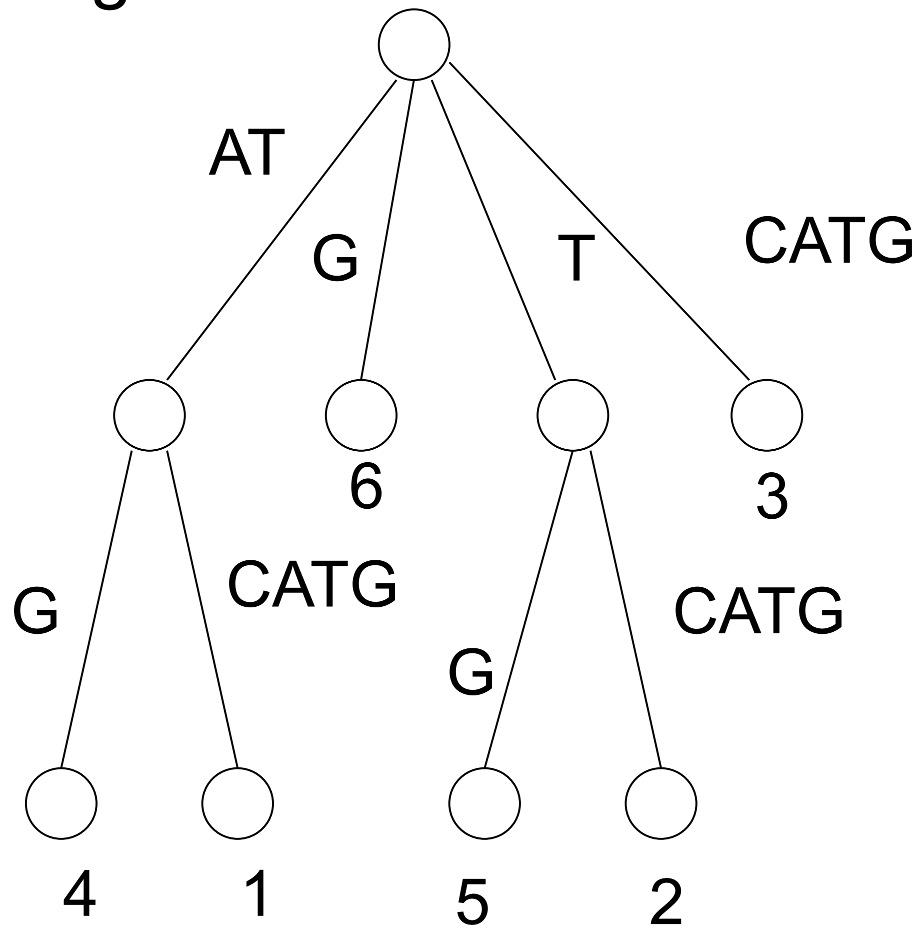
their
there
was
when

Stop and think: How many nodes are in the suffix trie for a string of length N?

Suffix tree

- Extends trie of all suffixes of a string
- Collapses non-branching nodes

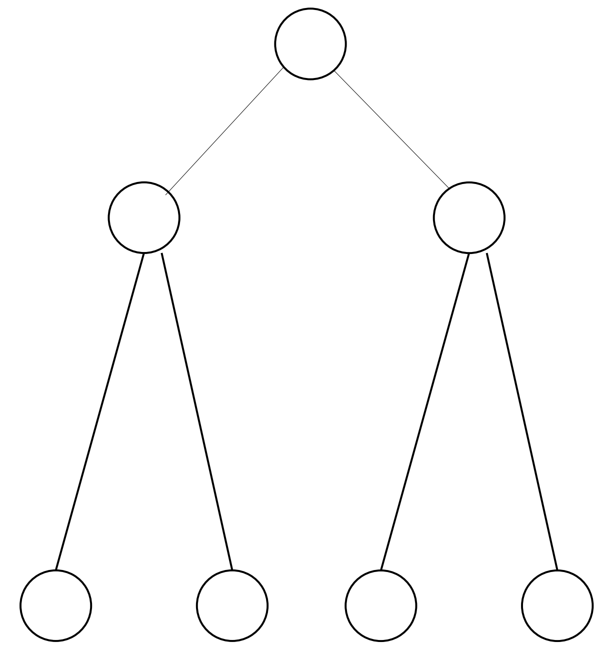
1	ATCATG
2	TCATG
3	CATG
4	ATG
5	TG
6	G



Stop and think:
How many nodes are in the suffix tree for a string of length N?
How much memory do you need to store the suffix tree?

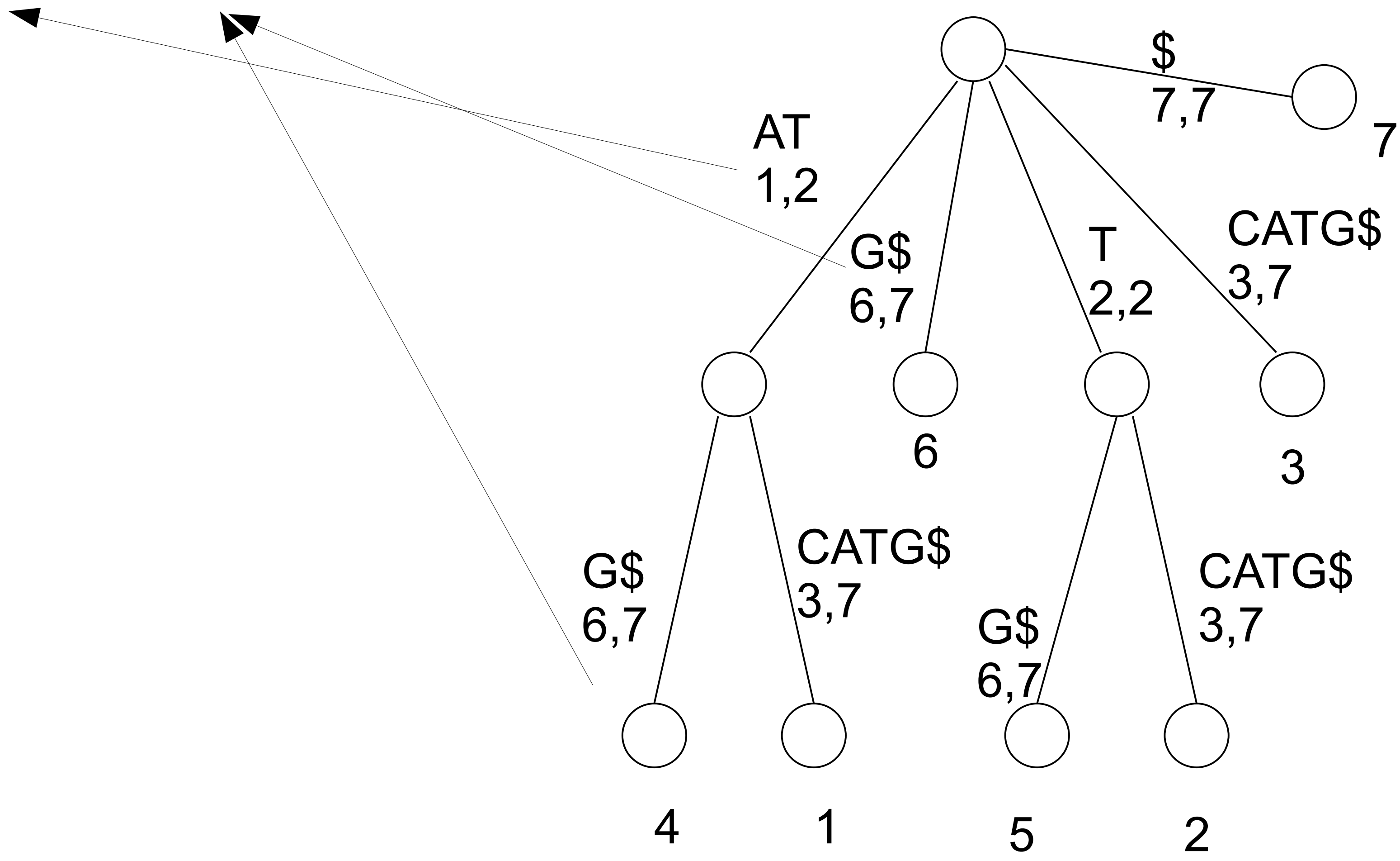
Some answers...

- Number of internal nodes <= number of leaves
- Worst case scenario – complete binary tree: number of internal nodes = number of leaves – 1
- Tree size = # nodes + info in tree
- # nodes = O(N)
  (as many leaves as suffixes)
- info in tree = all suffixes - $O(N^2)$



Suffix tree ...cont

- To store in linear space – just store range in sequence instead of string (constant space per edge/node)
- To ensure suffixes end at leaves, add $ char at end of string
- ATCATG$



Next: using suffix trees to perform matching