ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΑΝΟΙΚΤΑ ακαδημαϊκά μαθήματα ΠΠ

Εισαγωγή στη Βιοπληροφορική

Ενότητα 3: Τεχνικές Ανάλυσης και Σύγκρισης Ακολουθιών Βιολογικών Δεδομένων Ι

Μακρής Χρήστος, Τσακαλίδης Αθανάσιος, Περδικούρη Αικατερίνη

Πολυτεχνική Σχολή

Τμήμα Μηχανικών Η/Υ και Πληροφορικής

Σκοποί ενότητας

- Σκοπός της ενότητας είναι η παρουσίαση του Δέντρο Επιθεμάτων (Suffix Tree) και οι εφαρμογές του

Περιεχόμενα ενότητας

- Δέντρο επιθεμάτων – Γενικευμένο Δέντρο Επιθεμάτων
- Εφαρμογές στην ανάλυση ακολουθιών βιολογικών δεδομένων

Βασική Βιβλιογραφική Πηγή στην οποία βασίζονται οι διαφάνειες

- Dan Gusfield , Algorithms on Strings, Trees and Sequences,, Cambridge University Press, 10th edition 2007

Τεχνικές Ανάλυσης και Σύγκρισης Ακολουθιών Βιολογικών Δεδομένων Ι

Τεχνικές Ανάλυσης και Σύγκρισης Ακολουθιών Βιολογικών Δεδομένων

- Δέντρο Επιθεμάτων- Suffix Tree
- Γενικευμένο Δέντρο Επιθεμάτων - Generalized Suffix Tree
- Εφαρμογές σε Προβλήματα Μοριακής Βιολογίας

Βασικοί Ορισμοί

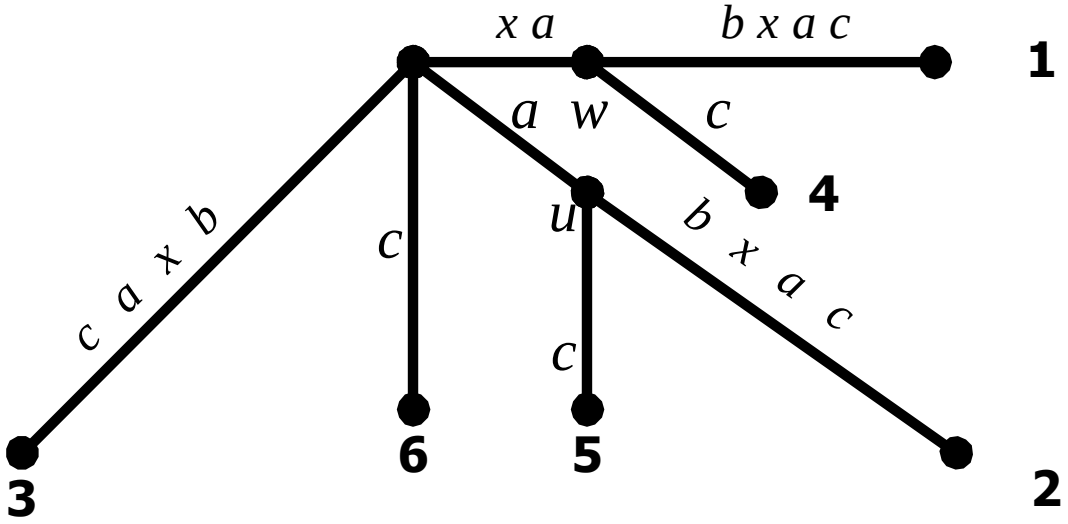
- Συμβολοσειρά-string: x=x[1]x[2]…..x[n], x[i]$\in\Sigma$ & |x|=n
  x= acgttaaaca, |x|=10 & Σ={a,c,g,t}

- Κενή συμβολοσειρά: ε
- Υπο-συμβολοσειρά-substring w: x=uwv
- Πρόθεμα –Prefix w: x=wu
- Επίθεμα-Suffix w: x=uw
- Κάθε συμβολοσειρά S, μήκους |S|=m, έχει m δυνατά μη κενά επιθέματα που είναι τα ακόλουθα: S[1…m], S[2…m], …. S[m-1…m] και S[m].
- Παράδειγμα "sequence" : *sequence, equence, quence, uence, ence, nce, ce, e.*

Το Δέντρο Επιθεμάτων
Suffix Tree

Ορισμός: *«αποθηκεύει όλα τα δυνατά επιθέματα μιας συμβολοσειράς S».*

x= xabxac



Κατασκευή του Δέντρου Επιθεμάτων

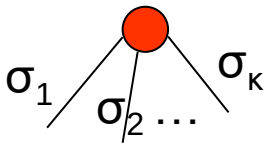
- Μια απλοϊκή θεώρηση:
  - Ένθεση μιας πλευράς στο δέντρο για το επίθεμα S[1…m]$,
  - Διαδοχική ένθεση των επιθεμάτων S[i…m]$, για i=2→m.

Trie

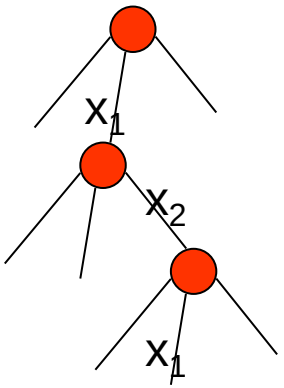
Ορισμός: Έστω σύμπαν $U=\Sigma^l$ για αλφάβητο Σ και l>0.

$$(x \in U : x = d_1 d_2 \cdots d_l)\ S \subseteq U :$$

Trie καλείται το κ-δικό δένδρο (κ=|Σ|) το οποίο περιέχει όλα τα προθέματα των στοιχείων του S. Κάθε επίπεδο του δένδρου αντιστοιχίζεται και σε ένα d_i (d_1→ρίζα). Κάθε στοιχείο $x=x_1x_2...x_l$ τοποθετείται στο υποδένδρο x_1→x_2→… .

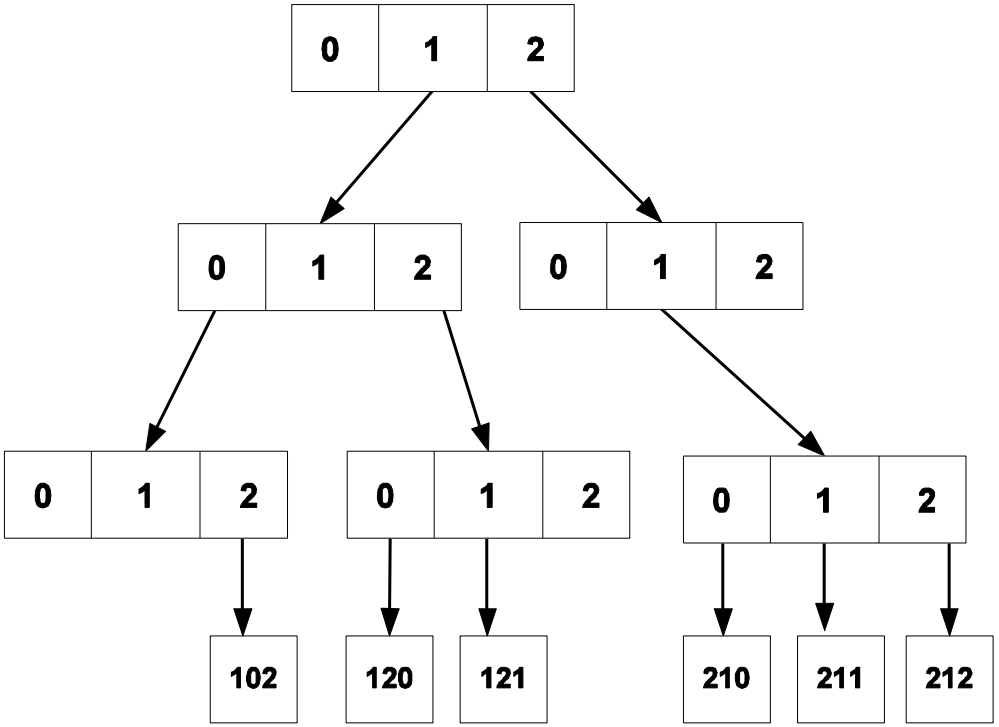


$$(\sigma_i \in \Sigma)$$



Trie - παράδειγμα

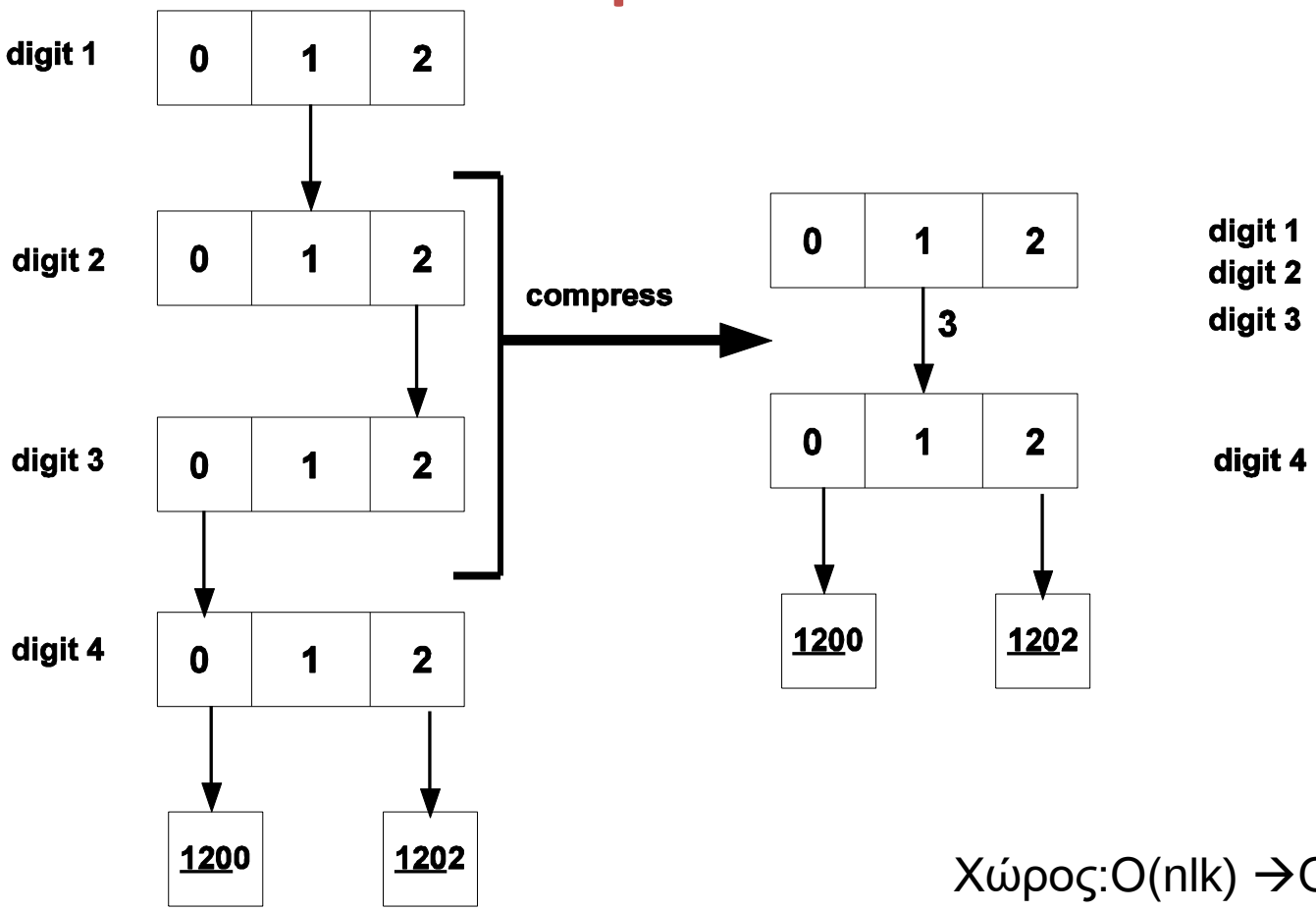
S={ 102, 120, 121, 212, 211, 120}, Σ={0,1,2}



Χρόνος ins/del/search : O(l)

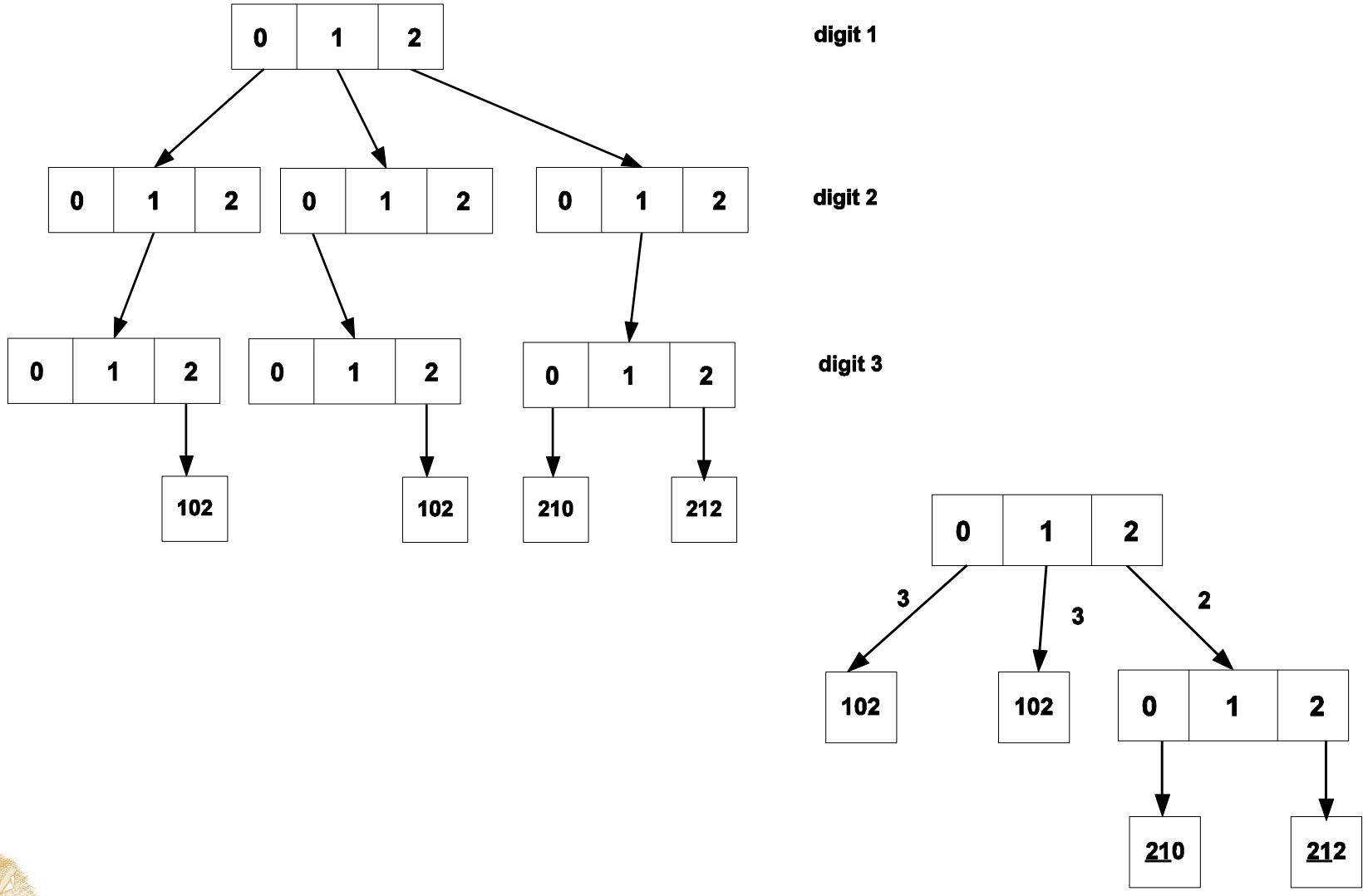
Χώρος: O(nlk)

Compressed Trie



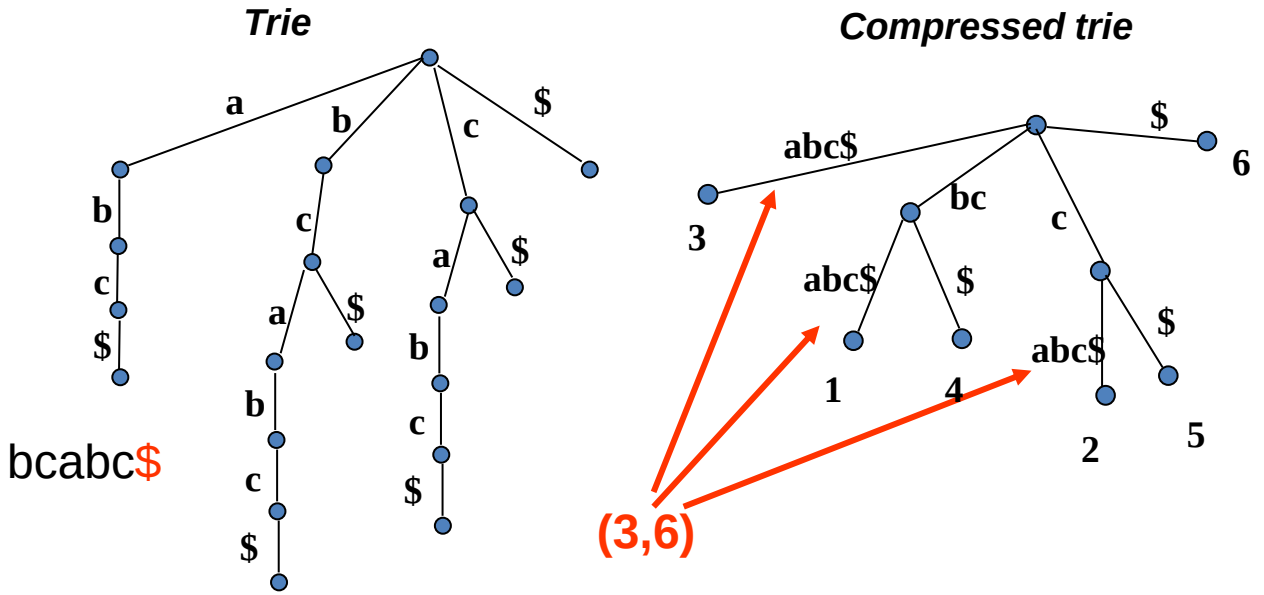
Χώρος:O(nlk) →O(nk)=O(n)

Compressed Trie - example



Suffix Tree

Ορισμός: Το suffix tree μιας συμβολοσειράς S[1…n] είναι ένα *συμπαγές* **trie** που περιέχει ως κλειδιά όλα τα επιθέματα S[i…n], 1≤i≤n.

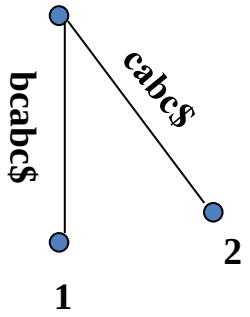


Naïve Κατασκευή

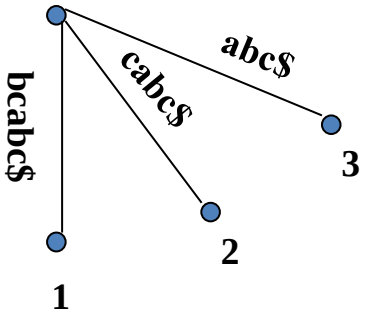
S=bcabc$

1: bcabc$

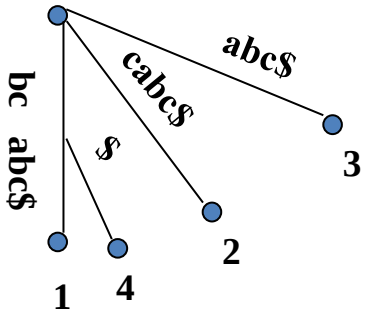
2: cabc$



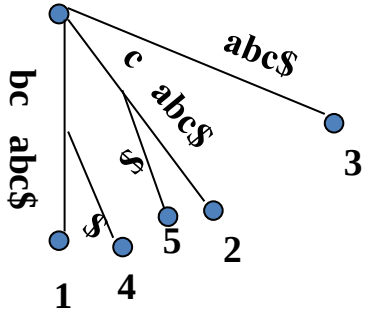
3: abc$



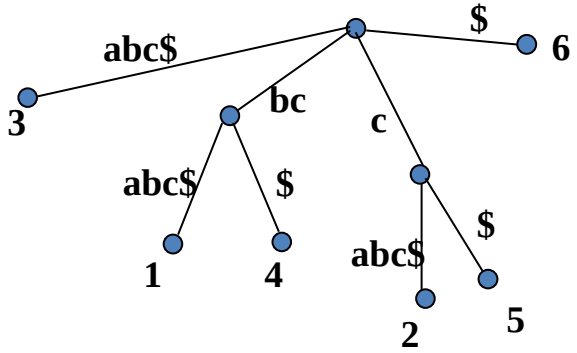
4: bc$



5: c$



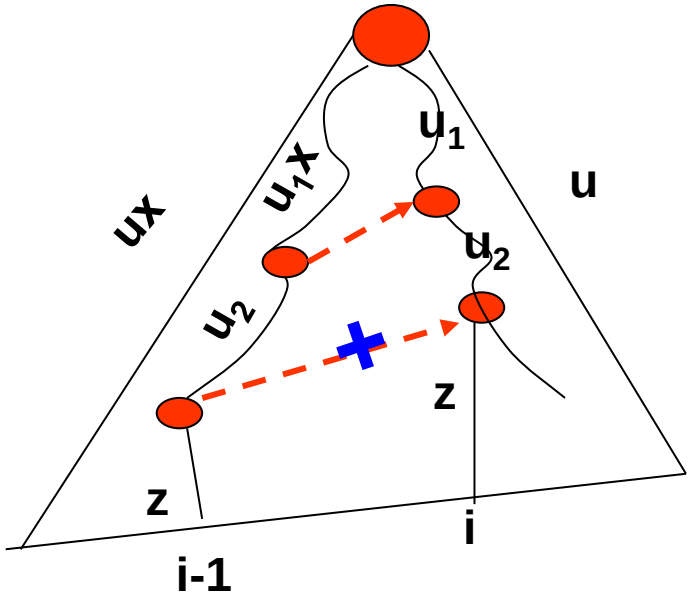
6: $



Χρόνος: $O(n^2)$

S=aaaaa...$

Suffix Links – Speed Up



i-1	i		
x	u		z
x	u_1	u_2	z

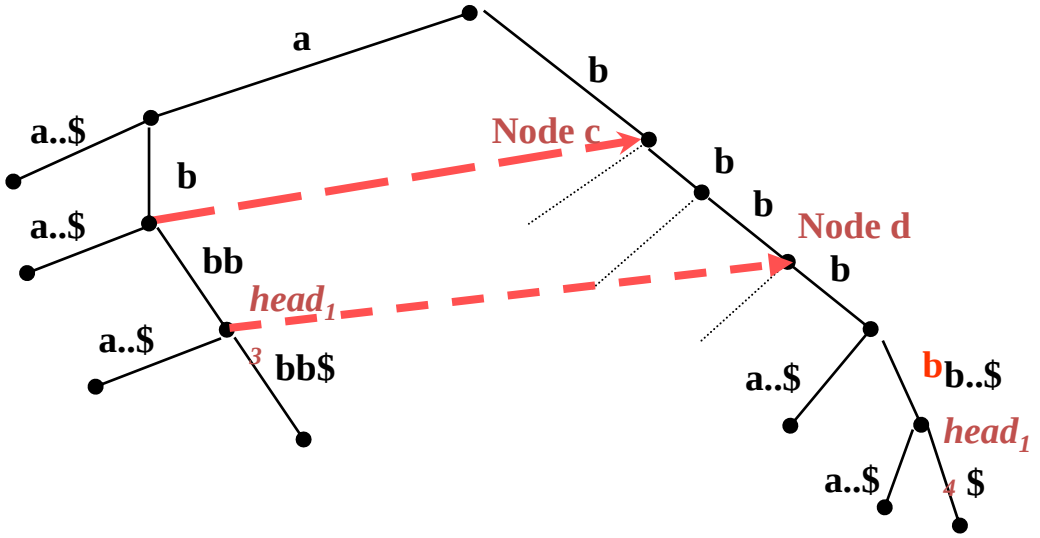
res(i)=|u2|+|z|

res(i+1)|<=res(i)-int_i

Scan cost=length($head_i$)-length($head_{i-1}$)

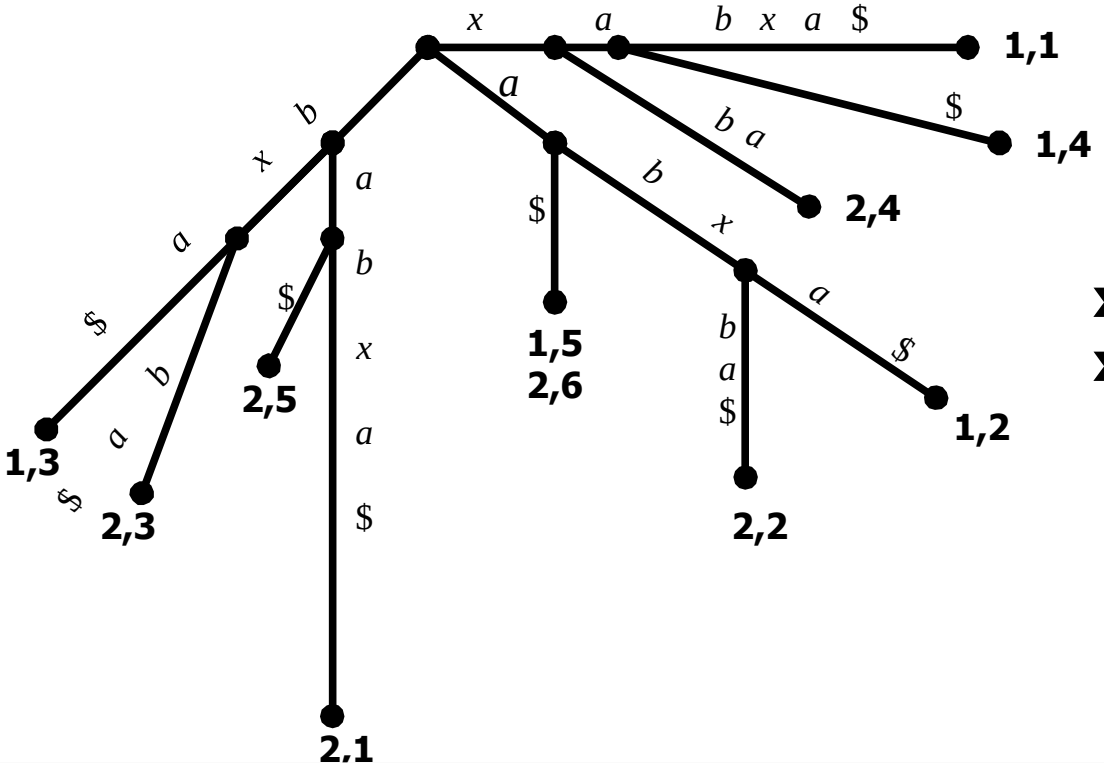
Suffix Links – Speed Up

S=bbbbbababbbaabbbbb$
i-1:**Inserted** S[13..19]=abbbbb$
x=a, u=bbb, z=bb$
u_1=b, u_2=bb
i: **Insert** S[14..19]=bbbbb$



Το Γενικευμένο Δέντρο Επιθεμάτων Generalized Suffix Tree

Ορισμός: *«Το **Γενικευμένο Δέντρο Επιθεμάτων**, αποτελεί ένα Δέντρο Επιθεμάτων το οποίο αποθηκεύει όλα τα δυνατά επιθέματα ενός συνόλου συμβολοσειρών* $S=\{S_1,S_2,...S_n\}$*».*



x_1 = xabxa
x_2 = babxba

Κατασκευή του Γενικευμένου Δέντρου Επιθεμάτων

- Μια απλοϊκή θεώρηση:
  - Διαδοχική ένθεση όλων των επιθεμάτων των συμβολοσειρών εισόδου.

Εφαρμογές στην ανάλυση ακολουθιών βιολογικών δεδομένων

- Ακριβής Εύρεση Προτύπου - Exact pattern matching problem
- Ακριβής Εύρεση Πολλαπλών Προτύπων- Σύγκριση με το Αυτόματο Aho- Corasick
- Μέγιστη κοινή υποσυμβολοσειρά 2 ακολουθιών- Longest Common Substring Problem
- DNA Contamination Problem
- Εύρεση Κοινών Μοτίβων σε 2 ή περισσότερες Βιολογικές Ακολουθίες
- Εύρεση Επαναλήψεων σε Βιολογικές Ακολουθίες

Ακριβής Εύρεση Προτύπου - Exact pattern matching problem

- Κατασκευή του Δ.Ε. για την ακολουθία εισόδου T σε O(|T|) χρόνο,
- Ξεκινώντας από τη ρίζα, σύγκρινε έναν προς έναν τους χαρακτήρες του P, ακολουθώντας το κατάλληλο μονοπάτι.
  - Εάν εμφανιστεί κάποιο μη-ταίριασμα, τότε το πρότυπο δεν εμφανίζεται στην ακολουθία,
  - διαφορετικά ανέφερε ως απάντηση όλα τα φύλλα που βρίσκονται κάτω από τον κόμβο του τελευταίου χαρακτήρα του P.
- Η αναζήτηση στοιχίζει O(n+k) χρόνο, |P|=n & k: το πλήθος εμφανίσεων του P στο T.

Ακριβής Εύρεση Πολλαπλών Προτύπων

- Κατασκευή του Δ.Ε. για την ακολουθία εισόδου Τ σε Ο(|Τ|) χρόνο,
- Ξεκινώντας από τη ρίζα, αναζητούμε διαδοχικά όπως και στην προηγούμενη εφαρμογή το σύνολο των προτύπων $P=\{P_1, P_2, ..\}$
- Η αναζήτηση στοιχίζει $O(n+|P|+ k_p)$ χρόνο, $|P|= |P_1| + |P_2| + ...$ & k_p: το πλήθος εμφανίσεων των προτύπων στο Τ.

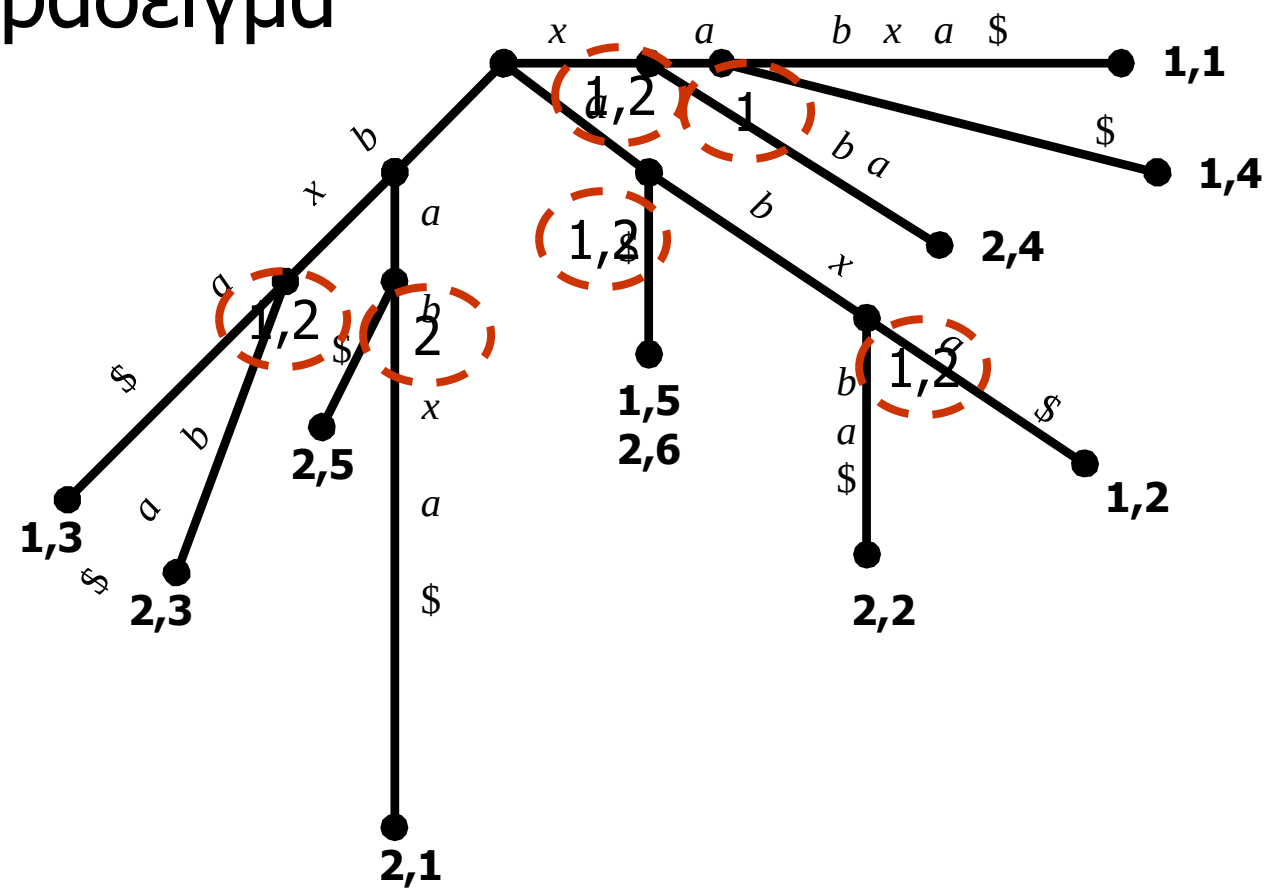
- Κατασκευή του Αυτομάτου σε χρόνο Ο(|P|)
- Αναζήτηση: Ο(m+k)

Μέγιστη κοινή υποσυμβολοσειρά 2 ακολουθιών- Longest Common Substring Problem

- Κατασκευή του Γ.Δ.Ε. για τις ακολουθίες εισόδου S_1, S_2, ...
- Σημειώνουμε κάθε εσωτερικό κόμβο του δέντρου *u,* με "1" ή "2", αν εμπεριέχει στο υπόδεντρο του *u*, κάποιο φύλλο που αναπαριστά κάποιο επίθεμα της ακολουθίας S_1 ή S_2.
- Η ετικέτα μονοπατιού - path label, κάθε εσωτερικού κόμβου που σημειώνεται ταυτόχρονα με "1" και "2", αποτελεί μια κοινή υπο-συμβολοσειρά των δυο ακολουθιών S_1 και S_2,.

Μέγιστη κοινή υποσυμβολοσειρά 2 ακολουθιών- Longest Common Substring Problem

- Παράδειγμα



DNA Contamination Problem

- ***DNA Contamination Problem:*** *Για μια δοσμένη ακολουθία DNA S_1, που έχει πρόσφατα απομονωθεί και ταυτοποιηθεί και μια ήδη γνωστή ακολουθία S_2, (επιμέρους τμήματα που πιθανά έχουν μολυνθεί), αναζητούμε όλες τις υπο-συμβολοσειρές της S_2 που εμφανίζονται στην S_1, με μήκος μεγαλύτερο από λ.*
- Κατασκευάζουμε το **Γενικευμένο Δέντρο Επιθεμάτων** για τις ακολουθίες S_1 και S_2.
- Ακολουθούμε τη μεθοδολογία και αναφέρουμε όλους τους κόμβους με βάθος *string-depth(u)* $\geq$ λ.

Εύρεση Κοινών Μοτίβων σε 2 ή περισσότερες Βιολογικές Ακολουθίες

- ***<u>Το Πρόβλημα της Εύρεσης κοινών μοτιβων:</u>*** *Για ένα σύνολο K ακολουθιών με συνολικό μήκος Σ(|K|)= n, και έναν ακέραιο k, (2<k<K), ορίζουμε ως λ(k), το μήκος του μέγιστου μοτίβου που εμφανίζεται σε τουλάχιστον k υπο-συμβολοσειρές. Το πρόβλημα ανάγεται στον υπολογισμό όλων των δυνατών τιμών του λ(k) και λύνεται σε γραμμικό χρόνο O(n), ως προς το μήκος των ακολουθιών εισόδου.*

Εύρεση Κοινών Μοτίβων σε 2 ή περισσότερες Βιολογικές Ακολουθίες

- Παράδειγμα
- Έστω *K={sandollar, sandlot, handler, grand, pantry}*.

k	*l (k)*	*μοτίβο*
2	*4*	*sand*
3	*3*	*and*
4	*3*	*and*
5	*2*	*an*

Εύρεση Επαναλήψεων σε Βιολογικές Ακολουθίες

- Οι επαναλήψεις σε βιολογικές ακολουθίες κατηγοριοποιούνται στις εξής 3 βασικές κατηγορίες:
  - επαναλήψεις περιορισμένου μήκους που εμφανίζονται σε τοπικό επίπεδο, και των οποίων η λειτουργία είναι γνωστή,
  - επαναλήψεις περιορισμένου μήκους που εμφανίζονται σε όλο το μήκος της ακολουθίας, και των οποίων η λειτουργία δεν είναι απόλυτα γνωστή,
  - δομημένες επαναλήψεις μεγάλου μήκους των οποίων η λειτουργία δεν έχει προσδιοριστεί.

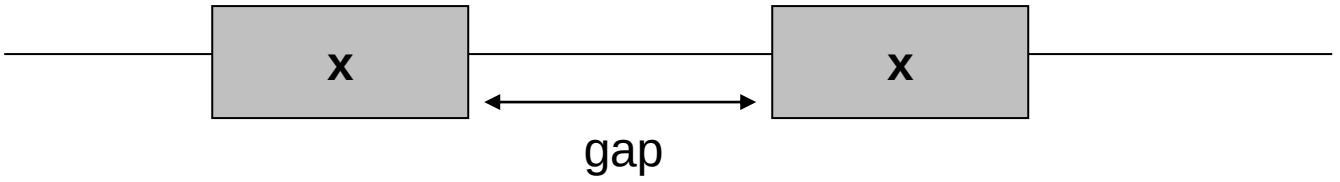
Παραδείγματα Επαναλήψεων

- 1η κατηγορία:
  - τα **συμπληρωματικά παλίνδρομα** σε ακολουθίες DNA & RNA, που ρυθμίζουν τη μετεγγραφή του DNA,
  - τα **εμφωλευμένα συμπληρωματικά παλίνδρομα** σε ακολουθίες tRNA
- 2η κατηγορία:
  - **συνεχόμενες επαναλήψεις- tandem repeats**,
  - **δορυφορικά τμήματα DNA**- **satellite DNA**, (**micro** & **mini satellite DNA**)
- 3η κατηγορία:
  - **SINE-Short Interspersed Nuclear Sequences (π.χ.:** *Alu family)*
  - **LINE-Long Interspersed Nuclear Sequences**.

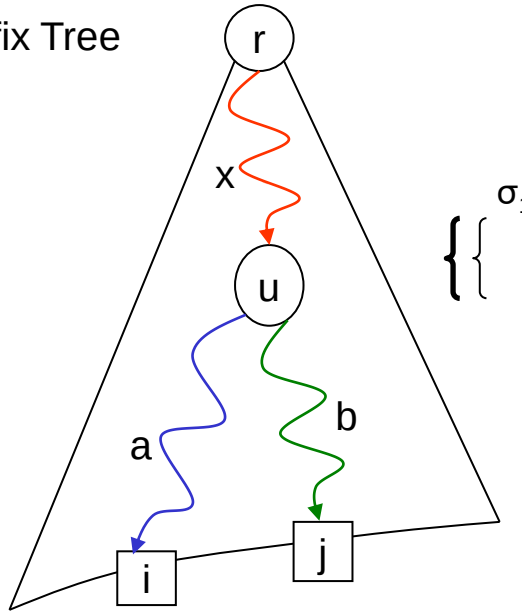
Παλίνδρομα

- ***Ορισμός:*** *Ένα* ***παλίνδρομο- palindrome*** *αποτελεί την επαναλαμβανόμενη εμφάνιση της υπο-συμβολοσειράς που διαβάζεται ως ίδια και προς τις 2 κατευθύνσεις (από αριστερά προς τα δεξιά και από δεξιά προς τα αριστερά): xyaayx*
- ***Ορισμός:*** *Ένα παλίνδρομο σε μια ακολουθία DNA ή RNA, ονομάζεται* ***συμπληρωματικό παλίνδρομο- complemented palindrome,*** *αν προκύπτει από την αντικατάσταση όλων των χαρακτήρων από την αρχή έως τη μέση με τις αντίστοιχες συμπληρωματικές βάσεις: agctcgcgagct*

Maximal Pairs



Suffix Tree

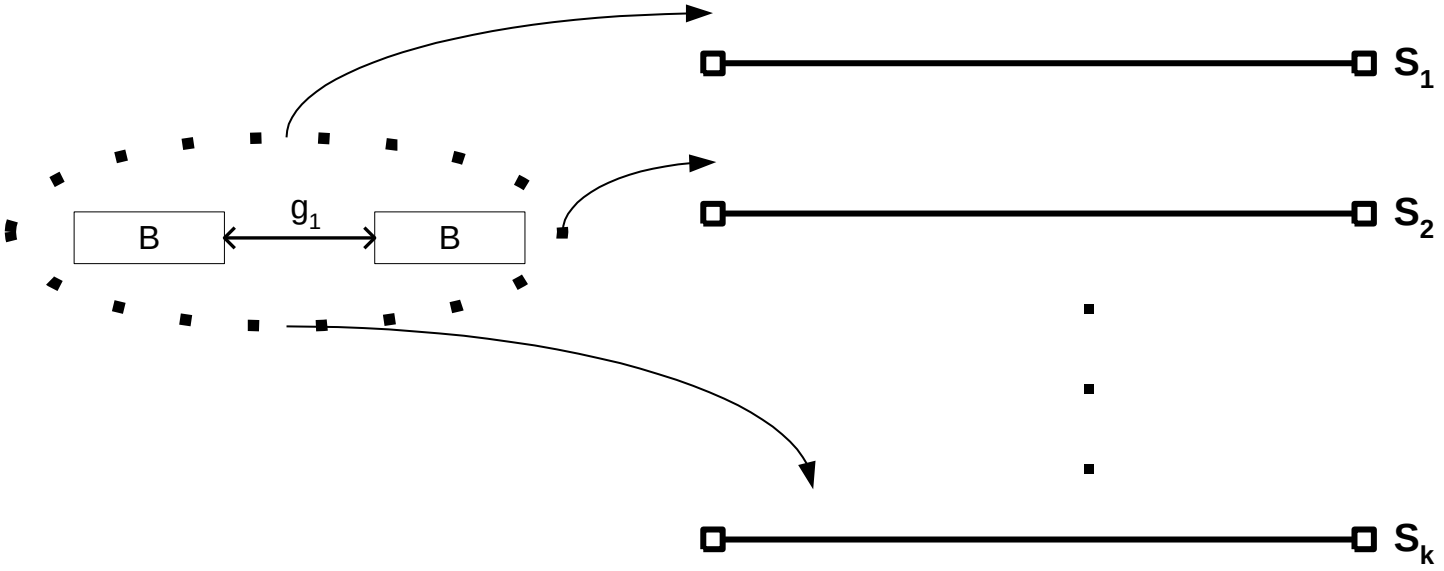


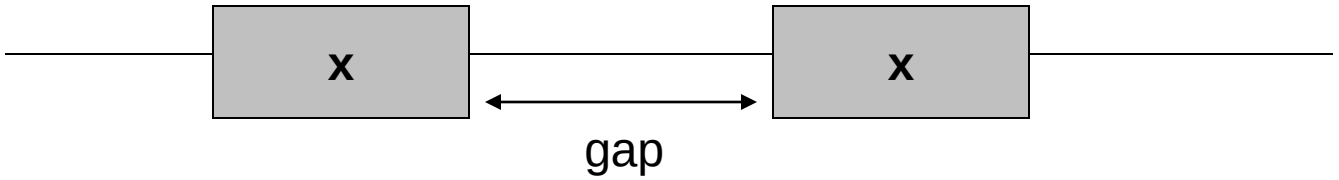
$\{\{\ \}\ \{i\}\ \ldots\ \{\ \}\}\quad \{\{j\}\ \{\ \}\ \ldots\ \{\ \}\}$ (over $\sigma_1\ \sigma_2\ \ldots\ \sigma_{|\Sigma|}$)

Gusfield : O(n+a)

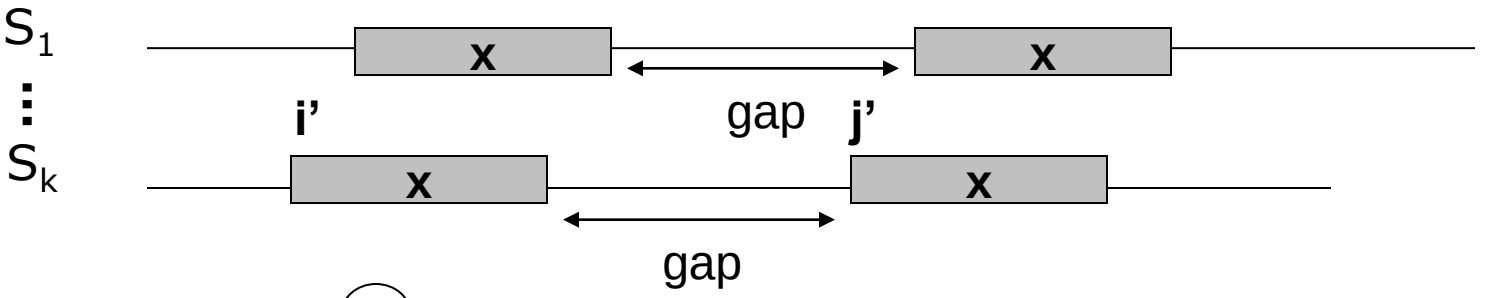
Brodal : O(nlogn+a) , $t_1 \le gap \le t_2$
O(n+a) , $t_1 \le gap$

Maximal Pairs in Multiple Strings

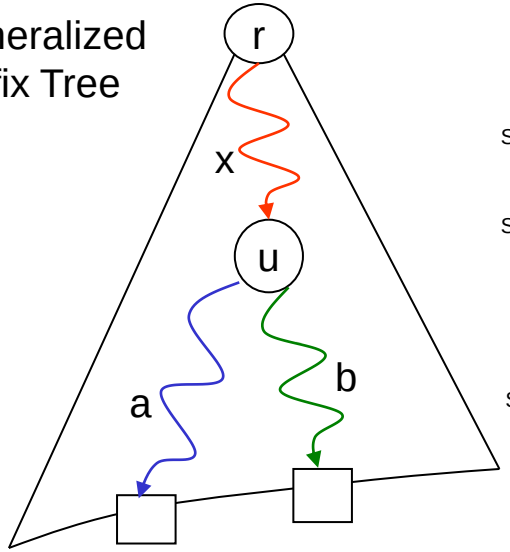




Maximal Pairs in Multiple Strings

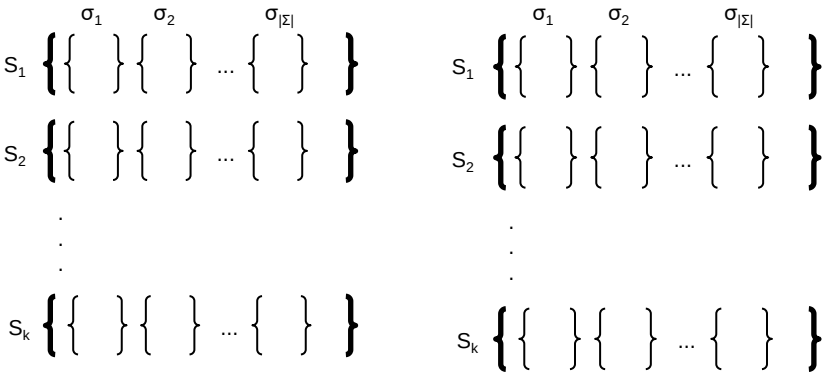


Generalized Suffix Tree



S_1:i S_1:j

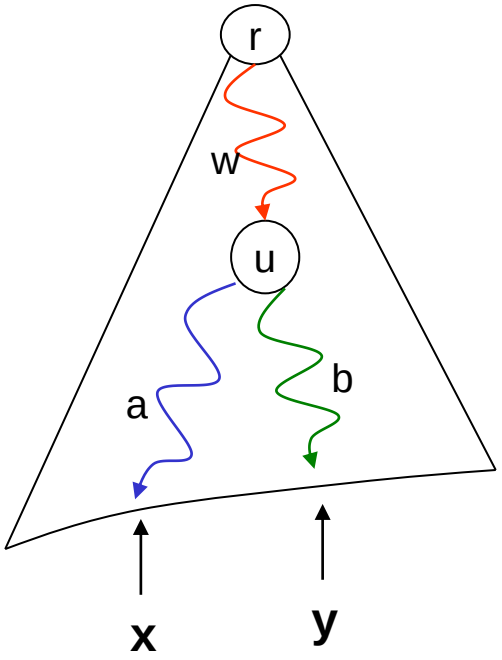
S_k:i' S_k:j'



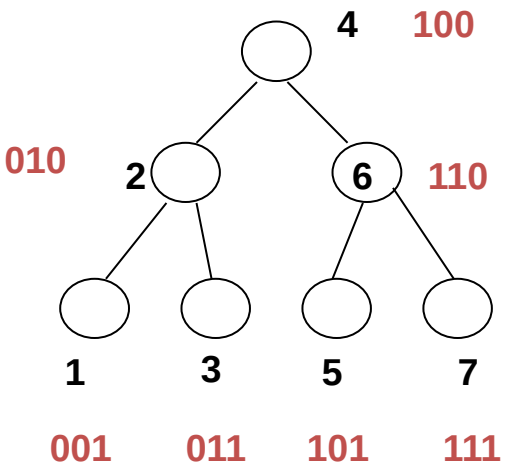
$O(n\log n + ak)$, gap $\leq t_1$

$O(n+a)$, gap unbounded

Nearest Common Ancestor & Suffix Tree



nca(x,y)=u σε χρόνο O(1)



nca(**001**, **101**)=leftmost$_1$(XOR(**001**, **101**)) = 100 =100

nca(**001**, **111**)= leftmost$_1$(XOR(**001**, **111**)) = 110 =100

nca(**011**, **010**)=leftmost$_1$(XOR(**011**, **010**) = 010

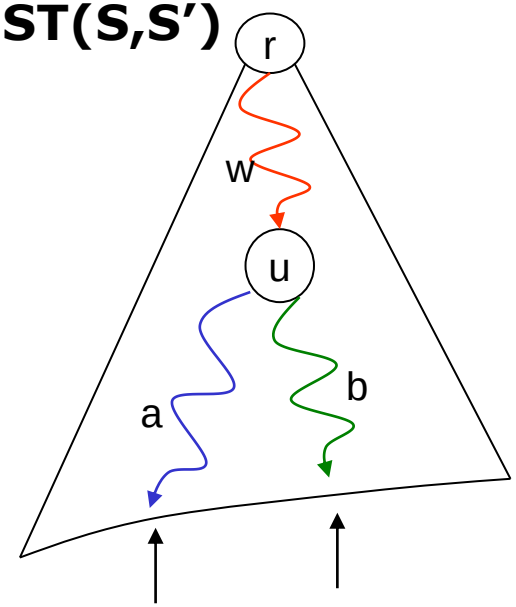
Maximal Palindromes

ȧvvɑ

aabactgaaccaat

taaccaagtcabaa

GST(S,S')



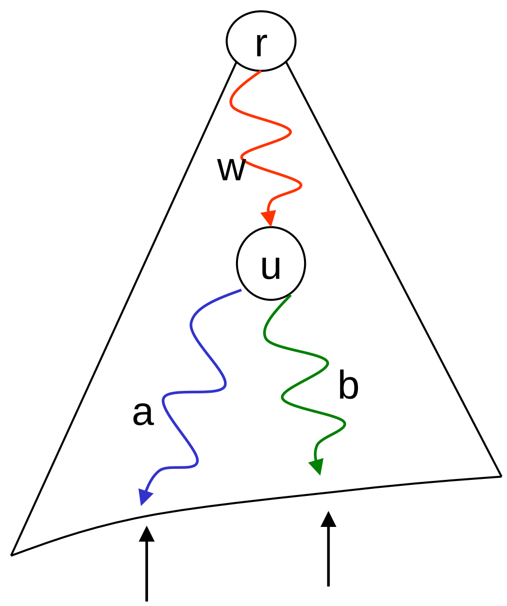
Exact Matching with wild cards

text ————————————————

pattern —*—*—*—*——*———*——

acgtttaaccttttgagttttgggcv

at**



Τέλος Ενότητας

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στο πλαίσιο του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Πατρών**» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Σημειώματα

Σημείωμα Ιστορικού Εκδόσεων Έργου

Το παρόν έργο αποτελεί την έκδοση 1.0.

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.