



Εισαγωγή στη Βιοπληροφορική

Μάθημα Ελεύθερης Επιλογής Τομέα Λογικού των Υπολογιστών

Εαρινό Εξάμηνο Ακαδ. Έτους 2024-2025

Ορισμένα σημεία διαφανειών είναι από το βιβλίο του
Dan Gusfield, Algorithms on Strings, Trees, and Sequences:
Computer Science and Computational Biology

Τεχνικές Ανάλυσης και Σύγκρισης Ακολουθιών Βιολογικών Δεδομένων

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

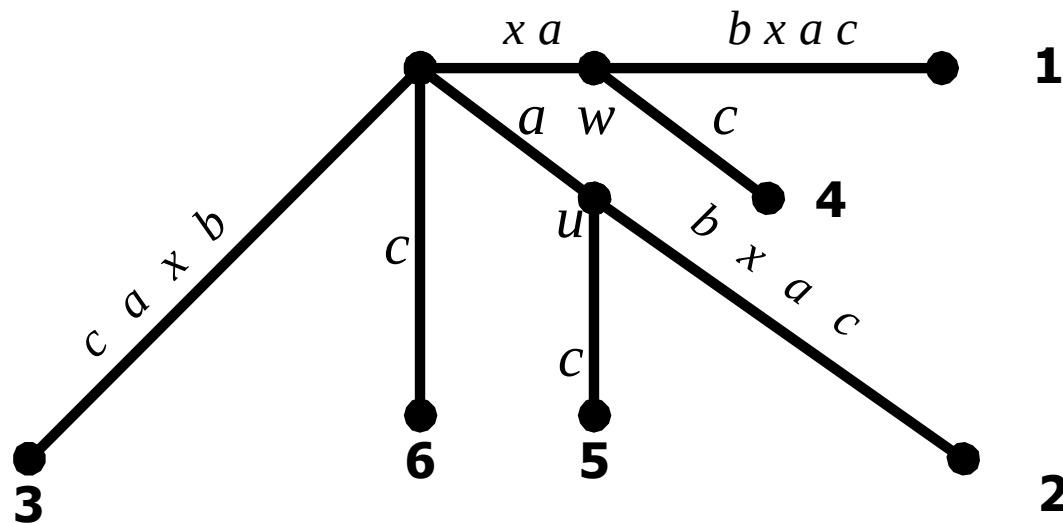
Βασικοί Ορισμοί

- Συμβολοσειρά-string: $x = x[1]x[2] \dots x[n]$, $x[i] \in \Sigma$ & $|x| = n$
 $x = \text{acgttaaca}$, $|x| = 10$ & $\Sigma = \{a, c, g, t\}$
- Κενή συμβολοσειρά: ϵ
- Υπο-συμβολοσειρά-substring w : $x = uwn$
- 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: $S[1 \dots m]$, $S[2 \dots m]$,
.... $S[m-1 \dots m]$ 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 ».

$p = \text{xabxac}$



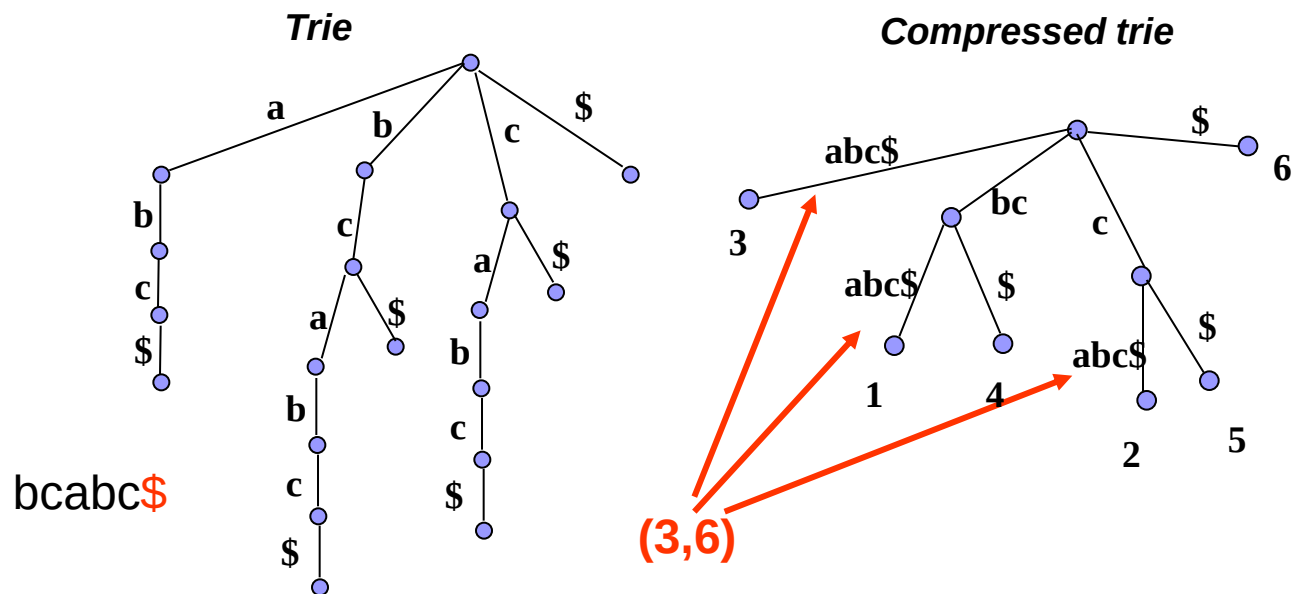
Ορισμός Ένα δέντρο επιθεμάτων T για μια συμβολοσειρά χαρακτήρων n χαρακτήρων είναι ένα rooted κατευθυνόμενο δέντρο με ακριβώς n φύλλα αριθμημένα από το 1 έως το n . Κάθε εσωτερικός κόμβος, εκτός από τη ρίζα, έχει τουλάχιστον δύο παιδιά και κάθε ακμή επισημαίνεται με μια μη κενή υποσυμβολοσειρά. Δεν είναι εφικτό από ένα κόμβο να υπάρχουν δύο ακμές που ξεκινούν από τον ίδιο χαρακτήρα. Το βασικό χαρακτηριστικό του δέντρου επιθεμάτων είναι ότι για οποιοδήποτε φύλλο i , η συνένωση των ακμών-ετικετών στη διαδρομή από τη ρίζα στο φύλλο i είναι ίση με το επίθεμα της συμβολοσειράς που ξεκινά από την θέση i .

Συνήθως βάζουμε στο τέλος της συμβολοσειράς εισόδου ένα σύμβολο εκτός αλφαβήτου (κατά σύμβαση το $\$$) ώστε να μην είναι κάποιο επίθεμα, πρόθεμα κάποιου άλλου.

από το βιβλίο: Dan Gusfield Algorithms on Strings, Trees and Sequences, Cambridge University Press,

Suffix Tree

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



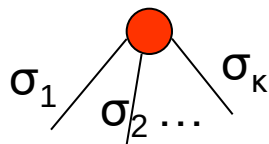
Trie

Ορισμός:

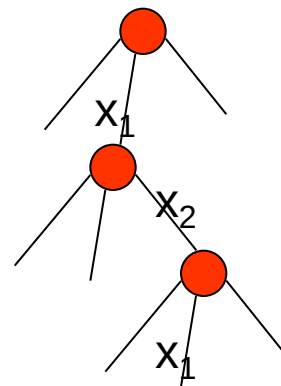
Έστω σύμπαν $U = \Sigma^0 \cup \dots \cup \Sigma^l$ για αλφάβητο Σ και $l > 0$.

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

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

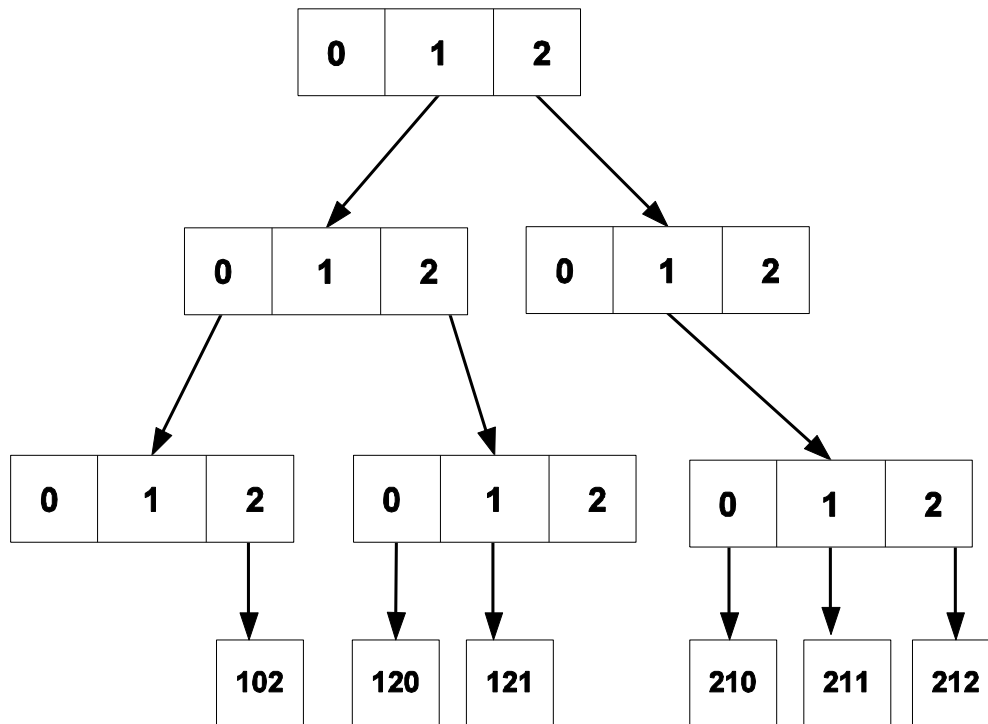


$$(\sigma_i \in \Sigma)$$



Trie - παράδειγμα

$S = \{102, 120, 121, 212, 211, 120\}$, $\Sigma = \{0, 1, 2\}$



digit 1

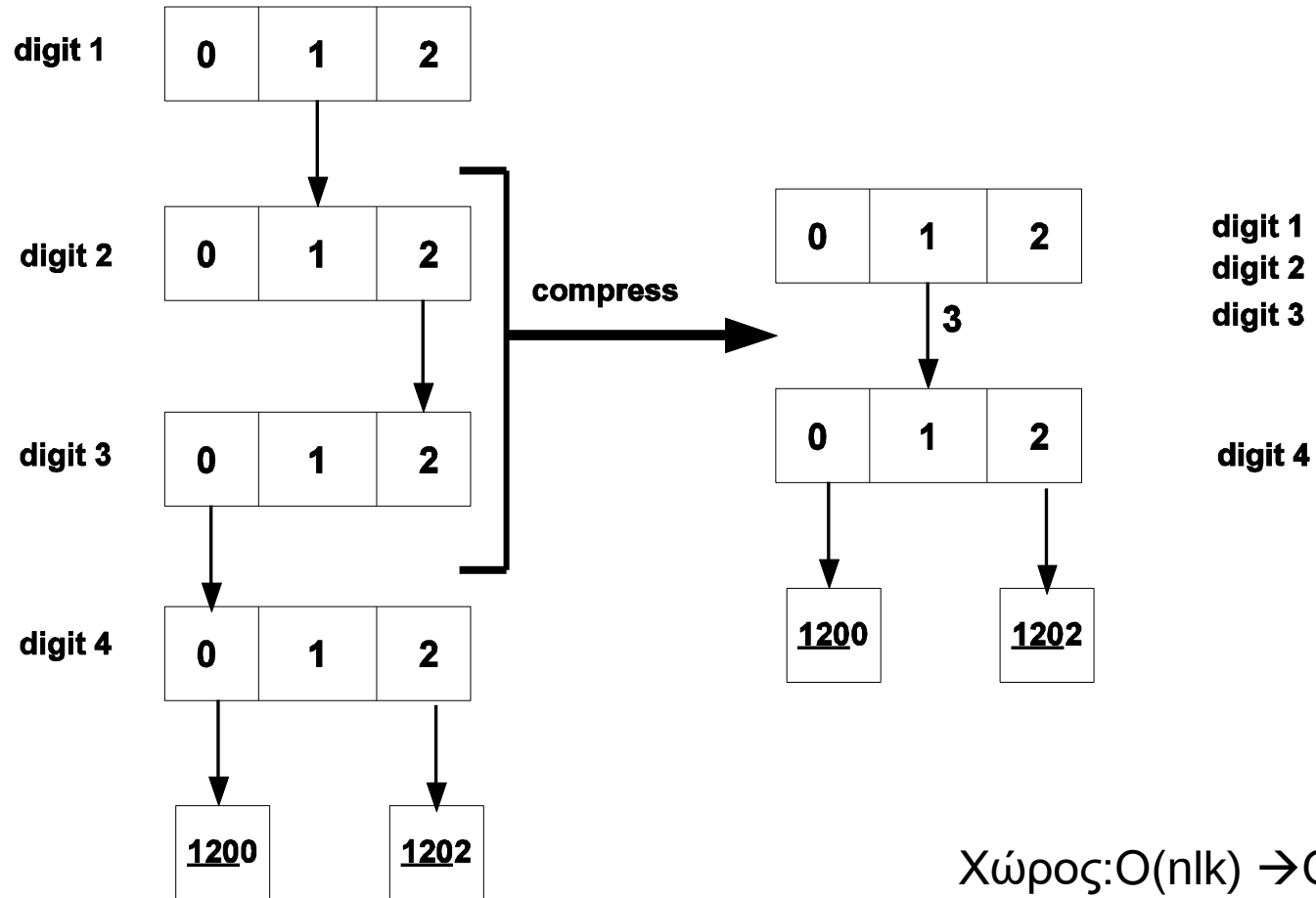
Χρόνος ins/del/search :
 $O(l)$

digit 2

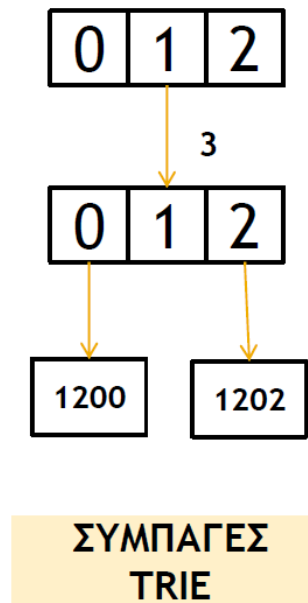
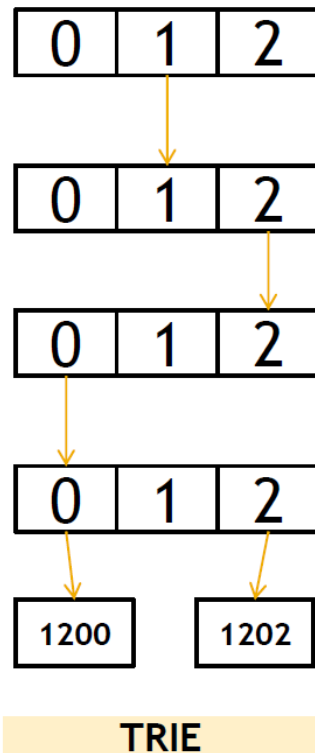
Χώρος:
 $O(nlk)$

digit 3

Compressed Trie



Compressed Trie - example



External construction

(https://en.wikipedia.org/wiki/Suffix_tree)

- Έχουν γίνει πρακτικές εργασίες για την κατασκευή δέντρων επιθεμάτων σε μαγνητικούς δίσκους που κλιμακώνονται σε (λίγα) GB/ώρα. Οι σύγχρονες μέθοδοι είναι: TDD, TRELLIS, DiGeST, and B2ST.
- Οι τεχνικές TDD και το TRELLIS κλιμακώνονται σε ολόκληρο το ανθρώπινο γονιδίωμα με αποτέλεσμα ένα δέντρο επιθεμάτων βασισμένο σε δίσκο μεγέθους δεκάδων gigabytes. Ωστόσο, αυτές οι μέθοδοι δεν μπορούν να χειριστούν αποτελεσματικά συλλογές ακολουθιών που υπερβαίνουν τα 3 GB. Το DiGeST αποδίδει σημαντικά καλύτερα και είναι σε θέση να χειριστεί συλλογές ακολουθιών της τάξης των 6GB σε περίπου 6 ώρες. Η τεχνική B²ST, κλιμακώνεται για να χειριστεί εισόδους που δεν ταιριάζουν στην κύρια μνήμη. Η τεχνική ERA μπορεί να ευρετηριάσει ολόκληρο το ανθρώπινο γονιδίωμα σε 19 λεπτά σε έναν υπολογιστή 8 πυρήνων με μνήμη RAM 16 GB.

Implementation

(https://en.wikipedia.org/wiki/Suffix_tree)

	Lookup	Insertion	Traversal
Sibling lists / unsorted arrays	$O(\sigma)$	$\Theta(1)$	$\Theta(1)$
Bitwise sibling trees	$O(\log \sigma)$	$\Theta(1)$	$\Theta(1)$
Hash maps	$\Theta(1)$	$\Theta(1)$	$O(\sigma)$
Balanced search tree	$O(\log \sigma)$	$O(\log \sigma)$	$O(1)$
Sorted arrays	$O(\log \sigma)$	$O(\sigma)$	$O(1)$
Hash maps + sibling lists	$O(1)$	$O(1)$	$O(1)$

Δέντρα Επιθεμάτων και gene/protein ενσωματώσεις

- - Τα δέντρα επιθεμάτων παρέχουν γρήγορη αντιστοίχιση υποσυμβολοσειρών για ανάλυση γονιδιώματος.
- - Οι ενσωματώσεις γονιδίων μετατρέπουν τις αλληλουχίες σε συνεχείς αριθμητικές αναπαραστάσεις.
- - Ο συνδυασμός αυτών των τεχνικών επιτρέπει την αποτελεσματική ανάκτηση και ταξινόμηση αλληλουχιών.

Δέντρα Επιθεμάτων - Βασικά Χαρακτηριστικά

- - Πολυπλοκότητα χώρου: $O(n)$ όπου n είναι το μήκος ακολουθίας.
- - Πολυπλοκότητα αναζήτησης: $O(m)$ όπου m είναι το μήκος του ερωτήματος.
- - Εφαρμογές: στοίχιση γονιδιώματος, επαναλαμβανόμενη αναζήτηση μοτίβων.
- - Χρησιμοποιείται σε εργαλεία όπως MUMmer, REPuter και STS.

Ενσωματώσεις Γονιδίων - Βασικές Έννοιες

- - Οι ενσωματώσεις γονιδίων αξιοποιούν μοντέλα εμπνευσμένα από NLP για ανάλυση αλληλουχίας.
- Οι προσεγγίσεις περιλαμβάνουν ενσωματώσεις k-mer, μοντέλα sequence-to-vector και μεθόδους βασισμένες σε μετασχηματιστές.-
- Συλλαμβάνουν σημασιολογικές σχέσεις σε αλληλουχίες DNA, RNA και πρωτεϊνών.
- - Χρησιμοποιούνται σε βιολογικές βάσεις δεδομένων για ταξινόμηση και αναζήτηση.

Γιατί να Συνδυάσουμε Δέντρα Επιθεμάτων και Ενσωματώσεις Γονιδίων?

- Τα επιθέματα είναι αποτελεσματικά για ακριβείς αντιστοιχίες, αλλά στερούνται σημασιολογικής κατανόησης.-
- Οι ενσωματώσεις γονιδίων χειρίζονται κατά προσέγγιση αντιστοιχίσεις, αλλά απαιτούν vector-based αναζήτηση ομοιότητας.
- Ο συνδυασμός και των δύο παρέχει γρήγορες αναζητήσεις και ισχυρή αντιστοίχιση ακολουθίας.-
- Εφαρμογές: Variant detection, genome assembly, functional annotation.

Υβριδική Ευρετηρίαση for Sequence Search

- - Βήμα 1: Κατασκευάστε ένα δέντρο επιθεμάτων για γενετικές αλληλουχίες.
- - Βήμα 2: Δημιουργήστε ενσωματώσεις για κάθε ακολουθία.
- - Βήμα 3: Συσχετίστε κόμβους δέντρων επιθεμάτων με διανύσματα ενσωμάτωσης.
- - Βήμα 4: Χρησιμοποιήστε την αναζήτηση κατά προσέγγιση πλησιέστερου γείτονα (ANN) για κατάταξη ομοιότητας.
- - Όφελος: Γρήγορη αναζήτηση με σημασιολογική κατανόηση των ακολουθιών.

Αποτελεσματική Ανίχνευση Μεταλλάξεων και Παραλλαγών

- - Τα παραδοσιακά δέντρα επιθεμάτων ανιχνεύουν μόνο ακριβείς αντιστοιχίες.
- - Οι ενσωματώσεις γονιδίων επιτρέπουν την ανίχνευση παρόμοιων αλληλουχιών με μεταλλάξεις.
- - Μέθοδος:
 - 1. Το δέντρο επιθεμάτων προσδιορίζει τις υποψήφιες περιοχές.
 - 2. Οι ενσωματώσεις βελτιώνουν την αναζήτηση με βάση την ομοιότητα.
- - Εφαρμογή: Προσδιορισμός γενετικών παραλλαγών σε μελέτες ασθενειών.

Κατηγοριοποίηση Ακολουθιών και Λειτουργικός Σχολιασμός

- - Τα δέντρα επιθεμάτων αποθηκεύουν αποτελεσματικά τις ακολουθίες για γρήγορη ανάκτηση.
- - Οι ενσωματώσεις ομαδοποιούν παρόμοιες ακολουθίες με βάση τη λειτουργία.
- - Υβριδική προσέγγιση:
 - 1. Χρησιμοποιήστε δέντρα επιθεμάτων για γρήγορη ανάκτηση υποψηφίων.
 - 2. Εφαρμόστε ενσωματώσεις για να ταξινομήσετε τις ακολουθίες σε λειτουργικές ομάδες.
- - Χρησιμοποιείται στην πρόβλεψη της πρωτεϊνικής δομής και στη φυλογενετική.

Βελτιστοποίηση Συναρμολόγησης Γονιδιώματος

- - Τα δέντρα επιθεμάτων βοηθούν στην ανίχνευση επικαλύψεων αλληλουχίας.
- - Οι ενσωματώσεις γονιδίων βελτιώνουν την ακρίβεια ευθυγράμμισης.
- - Συνδυασμένη προσέγγιση:
- 1. Προσδιορίστε τις επικαλύψεις χρησιμοποιώντας δέντρα επιθεμάτων.
- 2. Επικυρώστε τις επικαλύψεις χρησιμοποιώντας ομοιότητα ενσωμάτωσης.
- - Μειώνει τα λάθη και ενισχύει την αναδόμηση του γονιδιώματος.

4. Advantages of Combining Suffix Trees and Gene Embeddings

Feature	Suffix Trees	Gene Embeddings	Combined Approach
Search Speed	$O(m)$ lookup	$O(\log N)$ (ANN)	Fast + Semantic Matching
Memory Usage	High	Lower	Optimized Indexing
Mutation Handling	Weak	Strong	Robust against Variations
Sequence Classification	No	Yes	Yes (Hybrid)
Genomic Assembly	Good	Enhanced	Best Performance

Βήματα Υλοποίησης

- 1. Ακολουθίες προεπεξεργασίας: Εξαγωγή k-mers, δημιουργία ενσωματώσεων.
- 2. Κατασκευάστε ένα δέντρο επιθεμάτων από γονιδιωματικά δεδομένα.
- 3. Βελτιώστε τους κόμβους δέντρων επιθημάτων με διανύσματα ενσωμάτωσης.
- 4. Ενσωματώστε έναν δείκτη ANN για αποτελεσματική κατά προσέγγιση αντιστοίχιση.
- 5. Χρησιμοποιήστε υβριδική αναζήτηση για ανάκτηση ακολουθίας και ταξινόμηση.

- - Τα δέντρα επιθεμάτων επιτρέπουν την αποτελεσματική ακριβή αντιστοίχιση αλληλουχίας.
- - Οι ενσωματώσεις γονιδίων παρέχουν έναν ισχυρό τρόπο χειρισμού των παραλλαγών αλληλουχίας.
- - Μια υβριδική προσέγγιση συνδυάζει ταχύτητα και δυνατότητες σημασιολογικής αναζήτησης.
- - Οι εφαρμογές περιλαμβάνουν ανίχνευση παραλλαγών, συναρμολόγηση γονιδιώματος και ταξινόμηση.
- - Μελλοντική Εργασία: Εφαρμογή μοντέλων βαθιάς μάθησης για περαιτέρω βελτίωση της ενσωμάτωσης επιθημάτων

Κατασκευή του Δέντρου Επιθεμάτων

- Μια απλοϊκή θεώρηση:

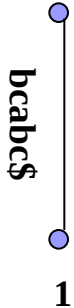
- Ένθεση μιας πλευράς στο δέντρο για το επίθεμα $S[1...m]$,
- Διαδοχική ένθεση των επιθεμάτων $S[i...m]$, για $i=2 \rightarrow m$.



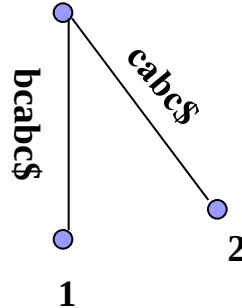
Naïve Κατασκευή

$S = \text{bcabc}\$$

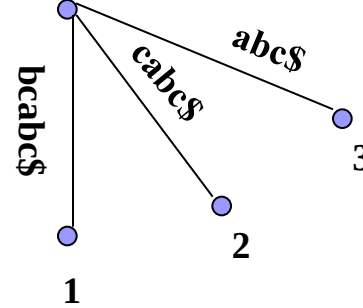
1: $\text{bcabc}\$$



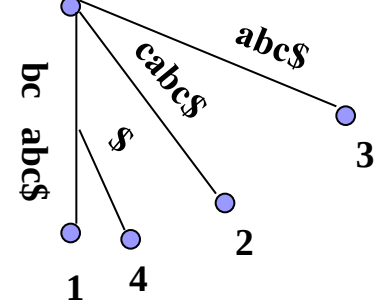
2: $\text{cab}\$$



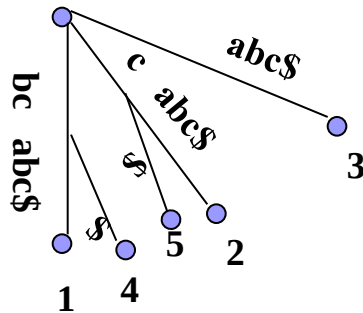
3: $\text{ab}\$$



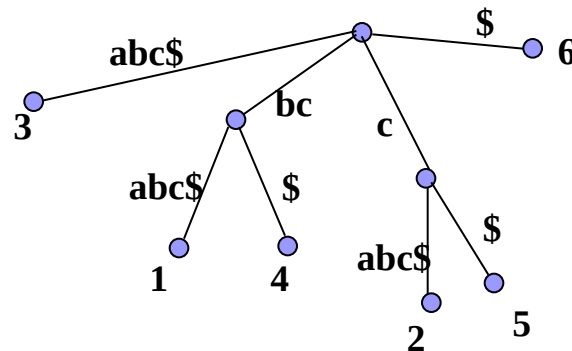
4: $\text{bc}\$$



5: $\text{c}\$$



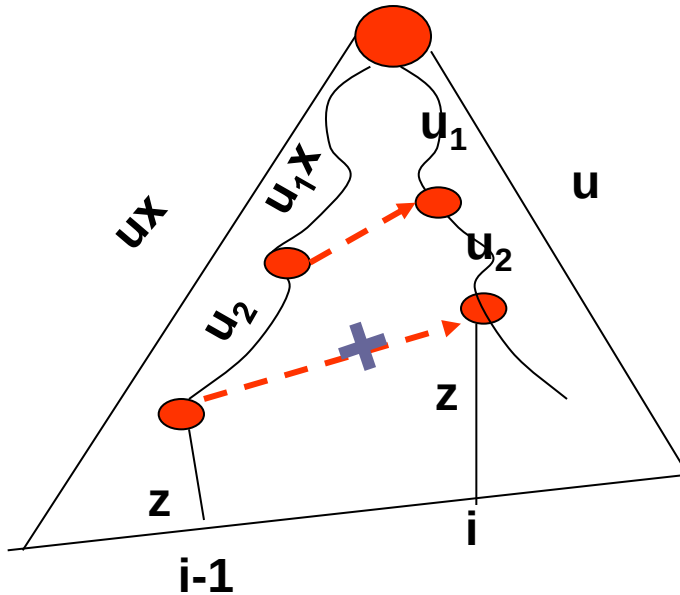
6: $\$$



Χρόνος: $O(n^2)$

$S = \text{aaaaa...}\$$

Suffix Links – Speed Up



$i-1$	i
x	u
x	u_1
u_2	z

$$\text{res}(i) = |u_2| + |z|$$

$$\text{res}(i+1) \leq \text{res}(i) - \text{int}_i$$

$$\text{Scan cost} = \text{length}(\text{head}_i) - \text{length}(\text{head}_{i-1})$$

Suffix Links – Speed Up

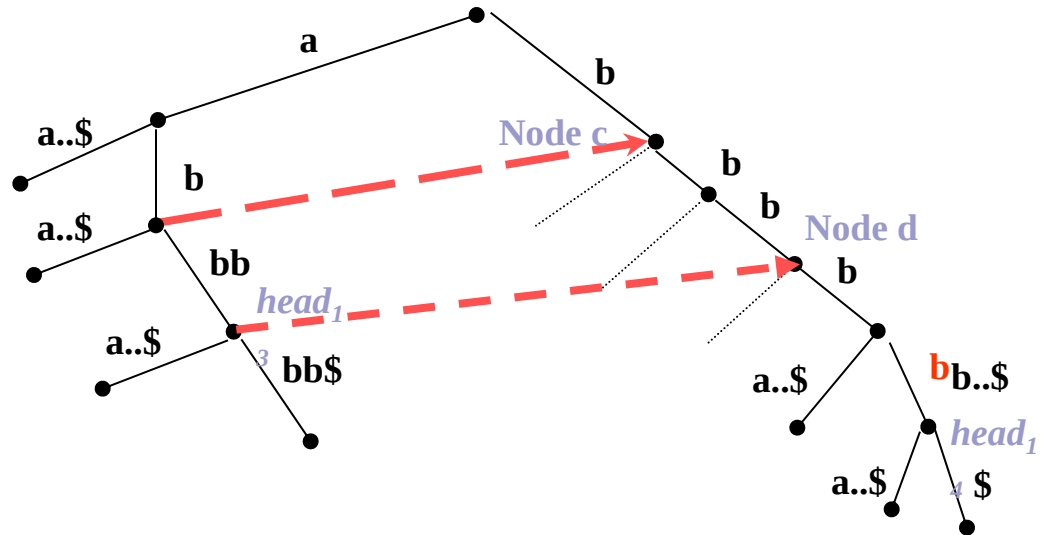
$S = \text{bbbbbababbbbaabbbbbb}\$$

$i-1$: **Inserted** $S[13..19] = \text{abbbbbb}\$$

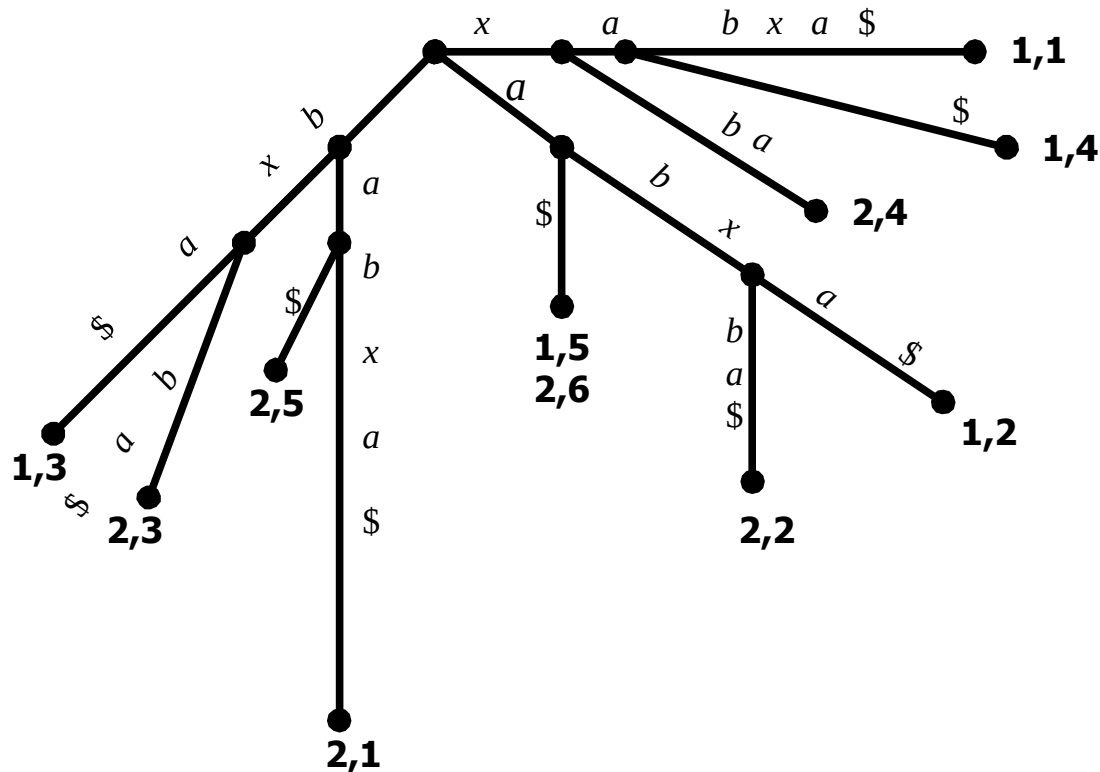
$x = a, u = \text{bbb}, z = \text{bb}\$$

$u_1 = b, u_2 = \text{bb}$

i : **Insert** $S[14..19] = \text{abbbbbb}\$$



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



$x_1 = \text{xabxa\$}$
 $x_2 = \text{babxba\$}$

Κατασκευή του Γενικευμένου Δέντρου Επιθεμάτων

- Μια απλοϊκή θεώρηση:

- Διαδοχική ένθεση όλων των επιθεμάτων των συμβολοσειρών εισόδου.

Εφαρμογές στην ανάλυση ακολουθιών βιολογικών δεδομένων

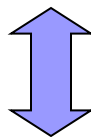
- Ακριβής Εύρεση Προτύπου - Exact pattern matching problem
- Ακριβής Εύρεση Πολλαπλών Προτύπων- Σύγκριση με το Αυτόματο Aho- Corasick
- Μέγιστη κοινή υποσυμβολοσειρά 2 ακολουθιών- Longest Common Substring Problem
- DNA Contamination Problem
- Εύρεση Κοινών Μοτίβων σε 2 ή περισσότερες Βιολογικές Ακολουθίες
- 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 $O(m+k)$ χρόνο, $|P|=m$ & k : το πλήθος εμφανίσεων του P στο T .

Ακριβής Εύρεση Πολλαπλών Προτύπων

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



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

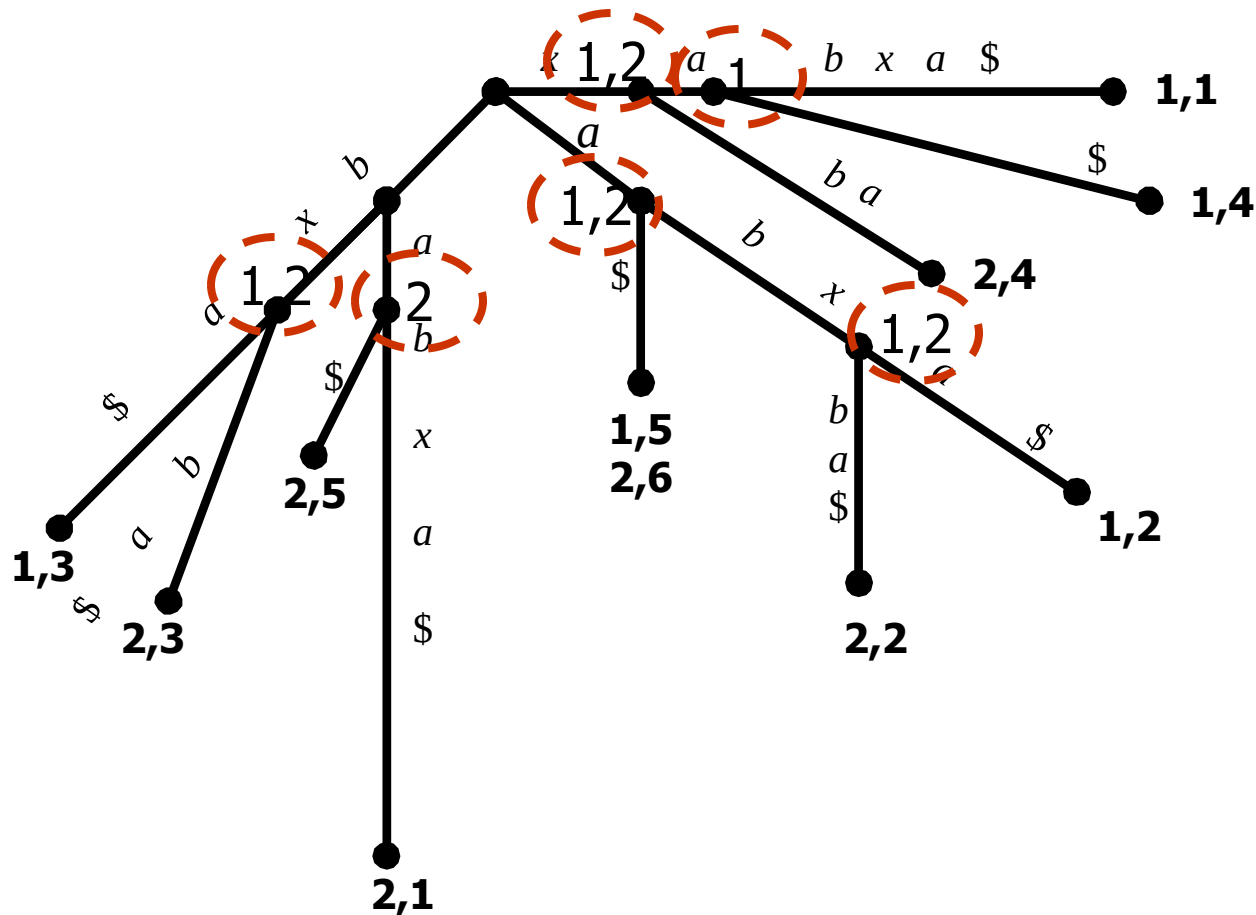
Μέγιστη κοινή υποσυμβολοσειρά 2 ακολουθιών- Longest Common Substring Problem

s t r i n g m a t c h e r s a n i a c h e r s
t e a c h e r A

- Κατασκευή του Γ.Δ.Ε. για τις ακολουθίες εισόδου $S_1, S_2, \dots$
- 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 υποσυμβολοσειρά 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 $string-depth(u) \geq \lambda$.

Εύρεση Κοινών Μοτίβων σε 2 ή περισσότερες Βιολογικές Ακολουθίες

- **Το Πρόβλημα της Εύρεσης κοινών μοτίβων:** Για ένα σύνολο K ακολουθιών με συνολικό μήκος $\Sigma(|K|) = n$, και έναν ακέραιο k , ($2 \leq k \leq K$), ορίζουμε ως $\lambda(k)$, το μήκος του μέγιστου μοτίβου που εμφανίζεται σε τουλάχιστον k υποσυμβολοσειρές. Το πρόβλημα ανάγεται στον υπολογισμό όλων των δυνατών τιμών του $\lambda(k)$ και λύνεται σε γραμμικό χρόνο $O(n)$, ως προς το μήκος των ακολουθιών εισόδου.

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)**

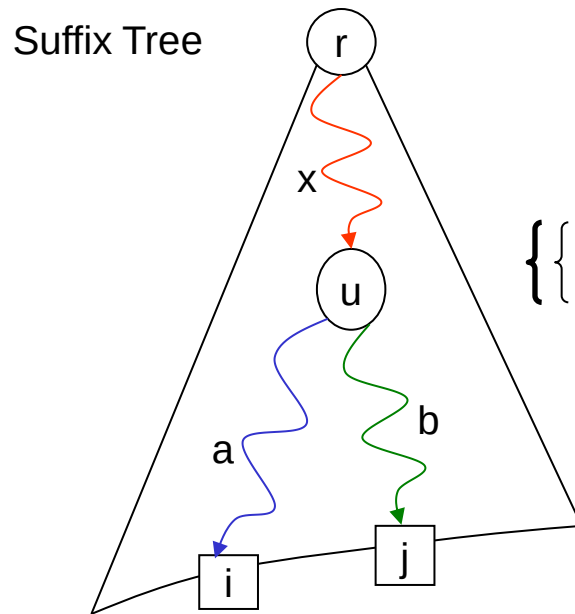
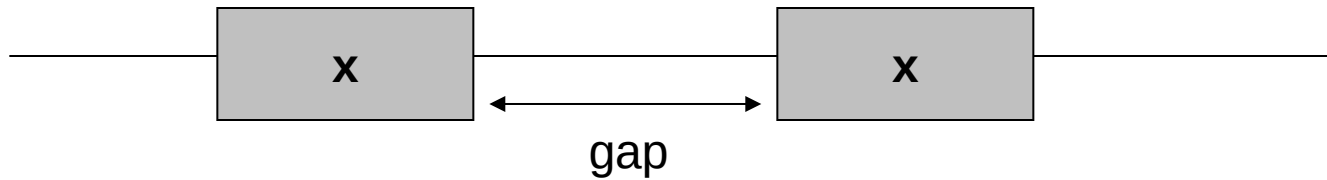
■ 3η κατηγορία:

- **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

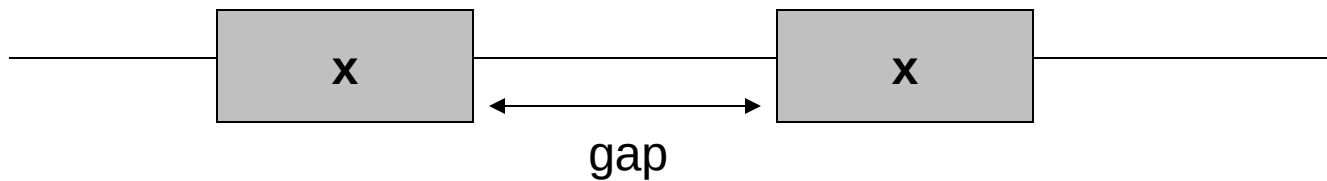
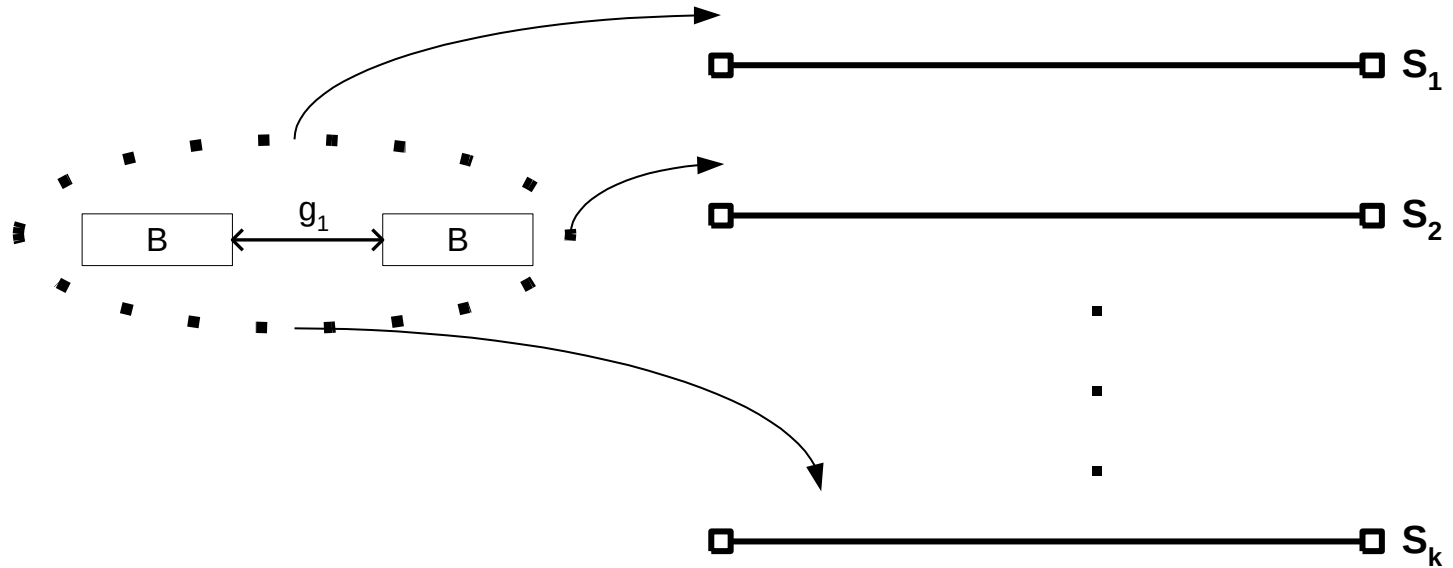


$$\left\{ \left\{ \sigma_1 \right\} \left\{ i \right\} \dots \left\{ \sigma_{|\Sigma|} \right\} \right\} \quad \left\{ \left\{ j \right\} \left\{ \sigma_1 \right\} \left\{ \sigma_2 \right\} \dots \left\{ \sigma_{|\Sigma|} \right\} \right\} \right\}$$

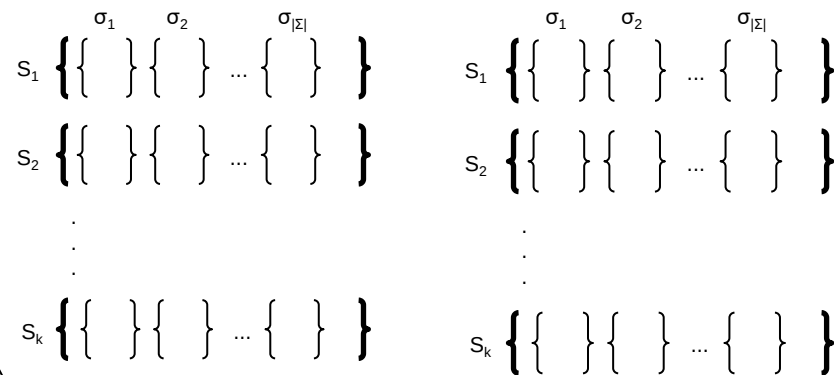
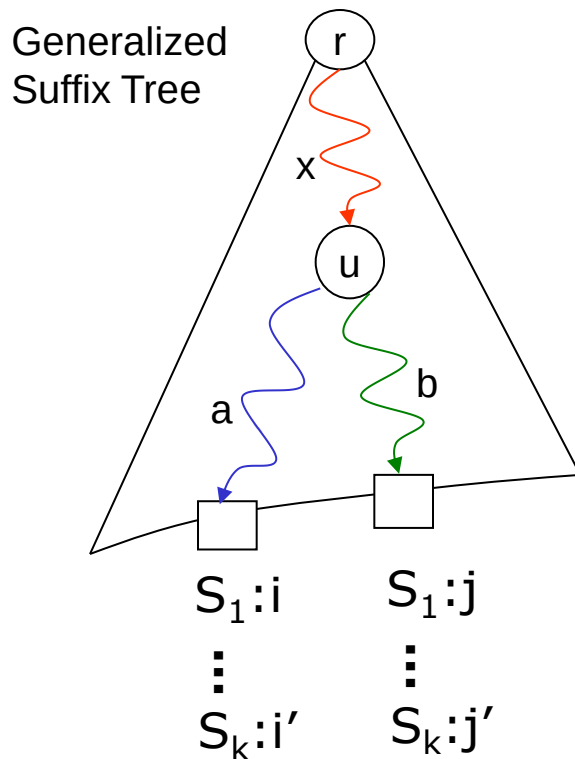
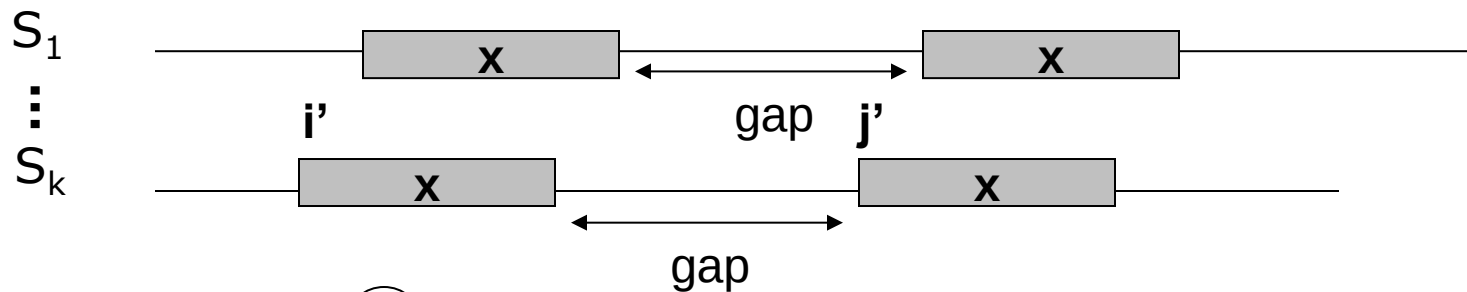
Gusfield : $O(n+a)$

Brodal : $O(n \log n + a)$, $t_1 \leq \text{gap} \leq t_2$
 $O(n+a)$, $t_1 \leq \text{gap}$

Maximal Pairs in Multiple Strings

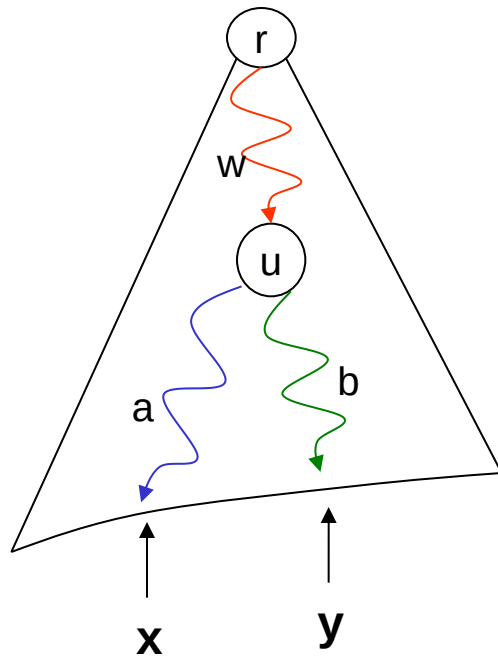


Maximal Pairs in Multiple Strings

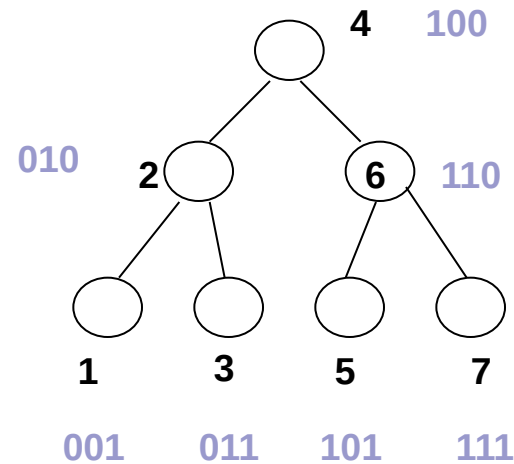


$O(n \log n + ak)$, gap $\leq t_1$
 $O(n + a)$, gap unbounded

Nearest Common Ancestor & Suffix Tree



$nca(x,y)=u$ σε χρόνο $O(1)$



$nca(\mathbf{001}, \mathbf{101}) = \text{leftmost}_1(\text{XOR}(\mathbf{001}, \mathbf{101})) = \mathbf{100} = 100$
 $nca(\mathbf{001}, \mathbf{111}) = \text{leftmost}_1(\text{XOR}(\mathbf{001}, \mathbf{111})) = \mathbf{110} = 100$
 $nca(\mathbf{011}, \mathbf{010}) = \text{leftmost}_1(\text{XOR}(\mathbf{011}, \mathbf{010})) = \mathbf{010}$

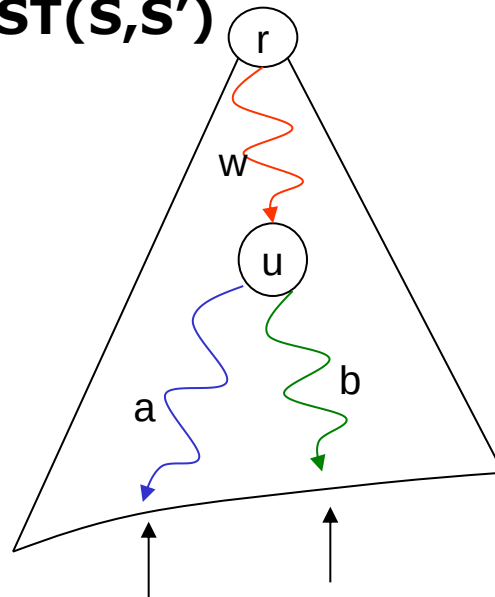
Maximal Palindromes

ävva

aabactgaaccaat

taaccaagtcabaa

$\text{GST}(S, S')$



Exact Matching with wild cards

text _____

acgtttaacctttgagttgggcv

pattern * * * * *

a**t

