



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 , Cambridge
University Press**

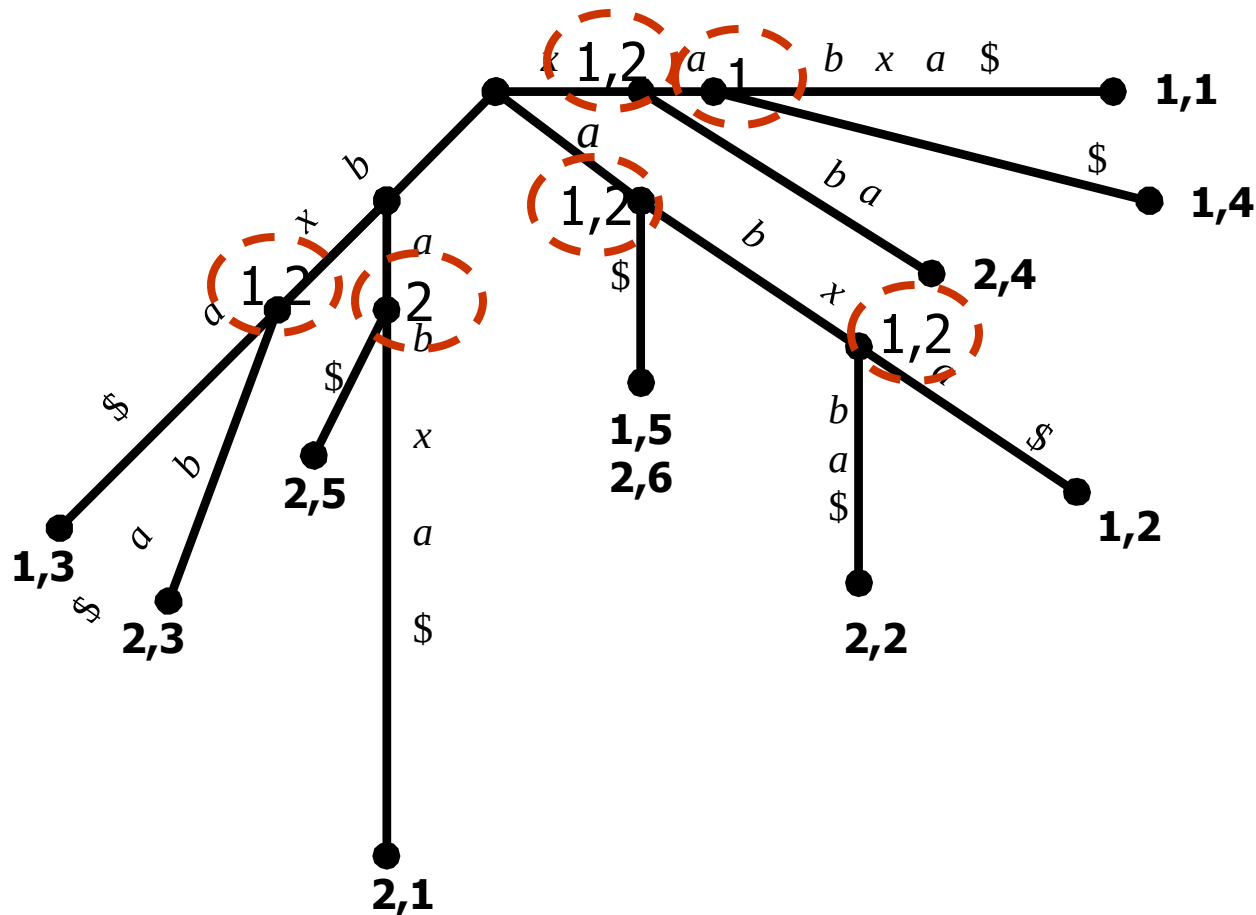
Μέγιστη κοινή υποσυμβολοσειρά 2 ακολουθιών- Longest Common Substring Problem

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)**

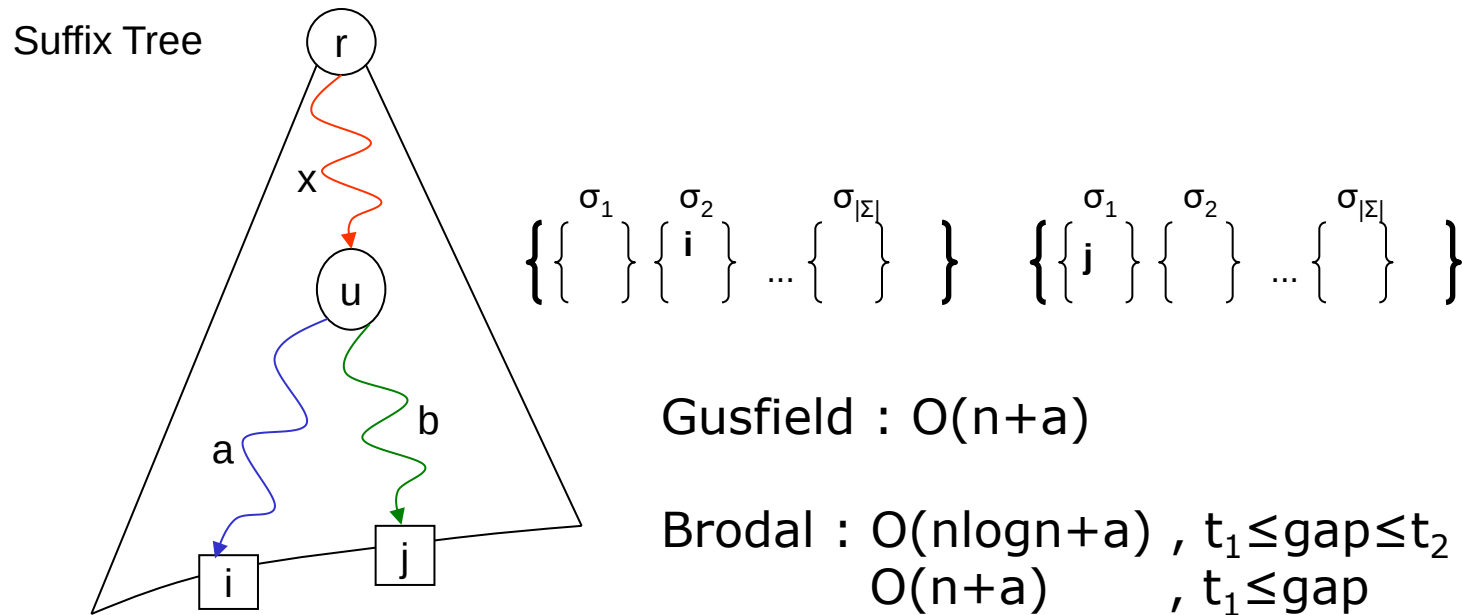
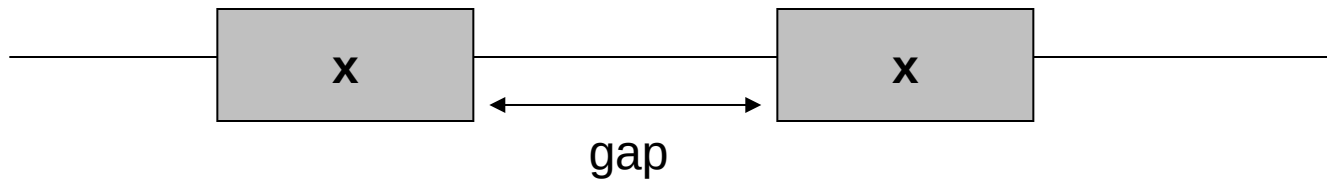
■ 3η κατηγορία:

- **SINE-Short Interspersed Nuclear Sequences (π.χ.: *Alu family*)**
- **LINE-Long Interspersed Nuclear Sequences.**

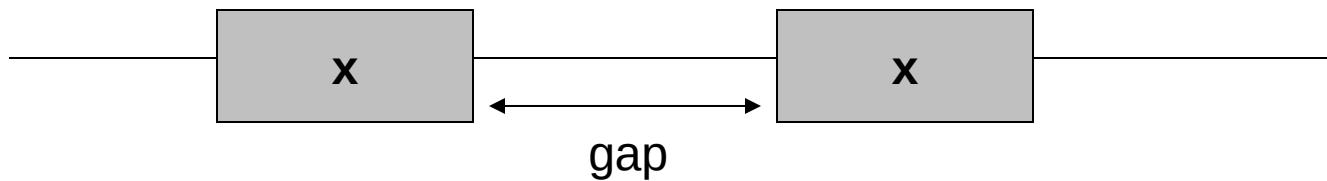
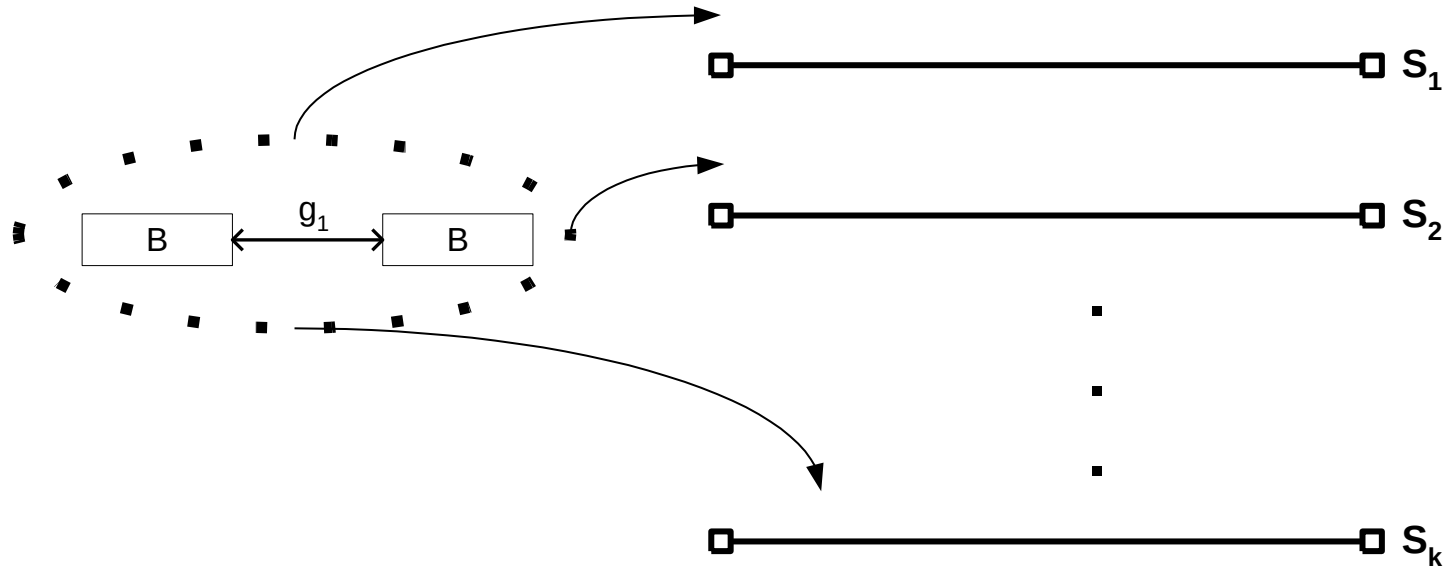
Παλίνδρομα

- **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*

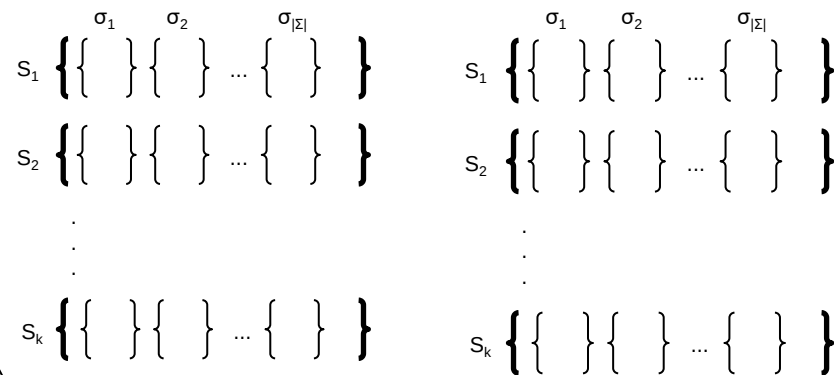
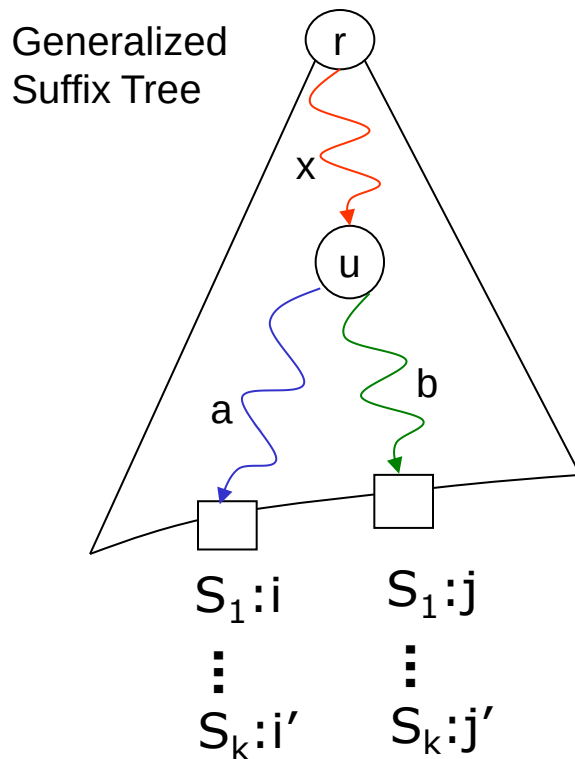
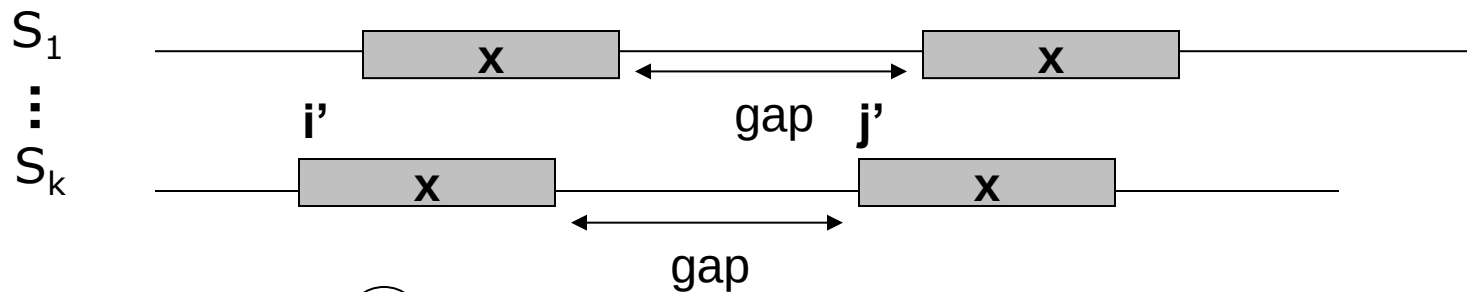
Maximal Pairs



Maximal Pairs in Multiple Strings

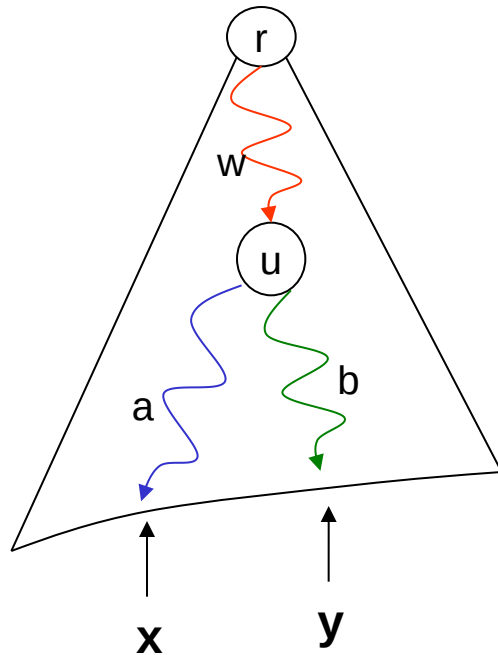


Maximal Pairs in Multiple Strings

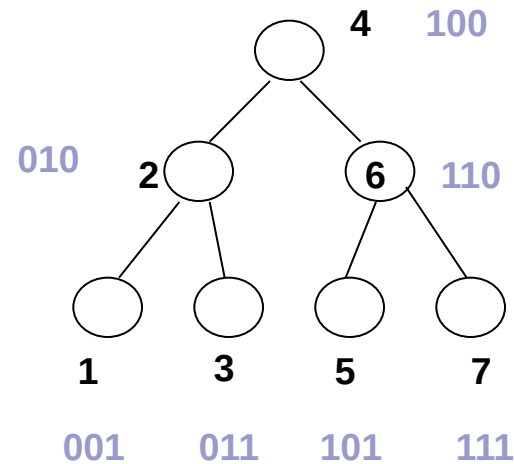


$O(n \log n + ak)$, gap $\leq t_1$
 $O(n + a)$, gap unbounded

Nearest Common Ancestor & Suffix Tree



$nca(x,y)=u$ σε χρόνο $O(1)$



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

Πρόβλημα Μεγαλύτερης κοινής επέκτασης (Longest Common Extension - LCE) :

Δίνονται δύο συμβολοσειρές $S1$ και $S2$ συνολικού μήκους n (σε μία αρχική φάση προεπεξεργασίας) και, μια **μακρά** ακολουθία ερωτημάτων για τα ζεύγη θέσεων (i,j)

Για κάθε καθορισμένο ζεύγος δεικτών (i, j) , πρέπει να βρούμε το μήκος της μεγαλύτερης υποσυμβολοσειράς του $S1$ που ξεκινά από τη θέση i που ταιριάζει με μια υποσυμβολοσειρά του $S2$ που ξεκινά από τη θέση j . Δηλαδή, πρέπει να βρούμε το μήκος του μεγαλύτερου προθέματος του επιθέματος i του $S1$ που ταιριάζει το πρόθεμα του επιθέματος j του $S2$.

Υπόδειξη: Απλή χρήση γενικευμένου δέντρου επιθεμάτων και NCA επεξεργασίας

From Dan Gusfield, **Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology** Cambridge University Press

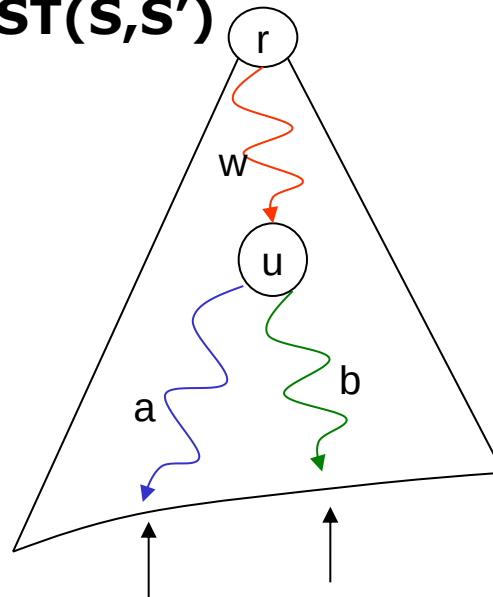
Maximal Palindromes

avva

aabactgaaccaat

taaccaagtcabaa

GST(S,S')



From Dan Gusfield, **Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology** Cambridge University Press

Maximal Palindromes

Υπάρχει μία απλή μέθοδος γραμμικού χρόνου για την εύρεση όλων των μέγιστων παλίνδρομων άρτιου μήκους στο S *(για περαιτέρω να εφαρμοστεί αντίστοιχος αλγόριθμος)*:

1. Σε γραμμικό χρόνο, δημιουργήστε την αντίστροφη συμβολοσειρά S^r από το S και προεπεξεργαστείτε τις δύο συμβολοσειρές, έτσι ώστε οποιοδήποτε Longest Common Extension ερώτημα να μπορεί να επιλυθεί σε σταθερό χρόνο.
2. Για κάθε q από 1 έως $n - 1$, λύστε το Longest Common Extension ερώτημα για το ζεύγος $(q+1, n-q+1)$ σε S και S^r , αντίστοιχα. Εάν η επέκταση έχει μη μηδενικό μήκος k , τότε υπάρχει ένα μέγιστο παλίνδρομο ακτίνας k κεντραρισμένο στο q .

Η μέθοδος απαιτεί **χρόνο $O(n)$** αφού το δέντρο επιθεμάτων μπορεί να δημιουργηθεί και να προεπεξεργαστεί για Longest Common Extension ερωτήματα σε αυτό το χρονικό διάστημα και κάθε ένα από τα **ερωτήματα επέκτασης** επιλύεται σε σταθερό χρόνο.

From Dan Gusfield, **Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology** Cambridge University Press

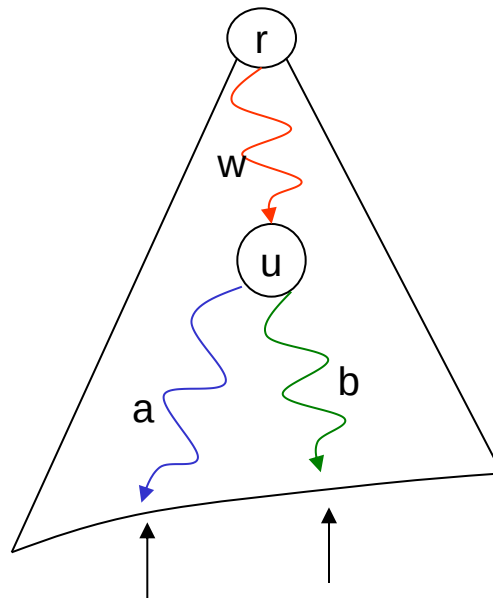
Exact Matching with wild cards

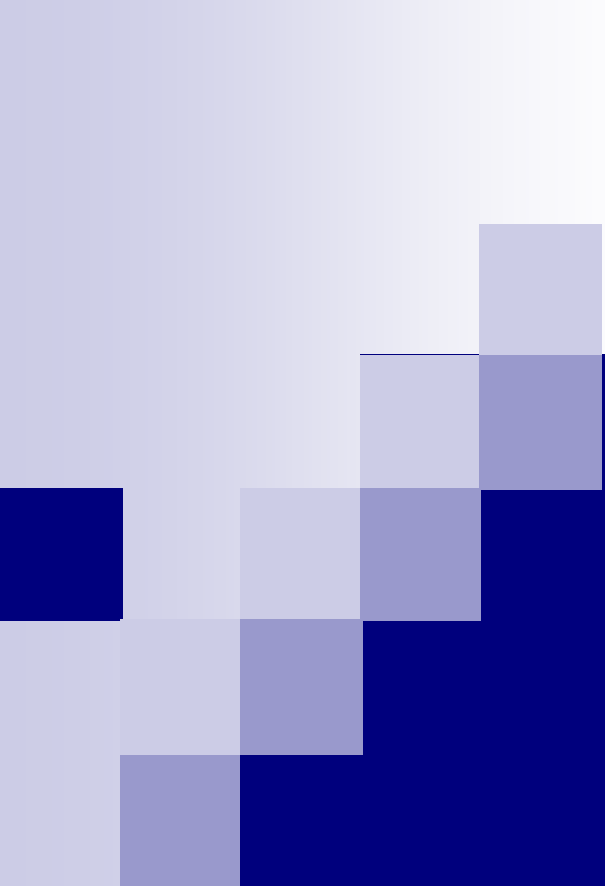
text _____

acgtttaacctttgagttgggcv

pattern * * * * *

a**t





Εισαγωγή στη Βιοπληροφορική Προσεγγιστικό ταΐριασμα

Τεχνικές Ανάλυσης και Σύγκρισης Ακολουθιών Βιολογικών Δεδομένων

- Προσεγγιστική Εύρεση Προτύπου - Approximate Pattern Matching
- Στοίχιση Ακολουθιών - Multiple Sequence Alignment
- Εφαρμογές σε Προβλήματα Μοριακής Βιολογίας

Παρατηρήσεις

- Η ολική στοίχιση λέγεται Needleman-Wunsch alignment (1970)
- Η τοπική στοίχιση λέγεται Smith-Waterman alignment (1981)
- FASTA (Pearson and Lipman, 1988), BLAST (Atschul al. 1990)
- Ο Smith-Waterman μπορεί να βρει περιοχές με μεγάλη ομοιότητα απλά εκτελώντας ένα ίχνος από οποιοδήποτε κελί (i,j) προς τα πίσω και να εντοπίσει ένα ζευγάρι με ομοιότητα (i,j) .

Βασικοί Ορισμοί (α)

- **Απόσταση Μετασχηματισμού - Edit Distance:** για 2 συμβολοσειρές ορίζουμε το ελάχιστο πλήθος των πράξεων μετασχηματισμού που απαιτούνται για να μετασχηματίσουμε την πρώτη συμβολοσειρά στη δεύτερη. Οι βασικές πράξεις μετασχηματισμού είναι η **ένθεση**, **διαγραφή** και **αντικατάσταση** συμβόλων.
- Παράδειγμα: S_1 : vintner και S_2 : writers
- $\text{edit-distance}(S_1 \rightarrow S_2) = 5$

Βασικοί Ορισμοί (β)

- **Ακολουθία Μετασχηματισμού - Edit Transcript:** για το μετασχηματισμό μιας συμβολοσειράς ορίζεται ως η ακολουθία των πράξεων μετασχηματισμού που απαιτούνται για να μετασχηματίσουμε την πρώτη συμβολοσειρά στη δεύτερη. Οι βασικές πράξεις μετασχηματισμού αναπαρίστανται ως εξής:
 - **ένθεση: I**
 - **διαγραφή: D**
 - **αντικατάσταση: R**
 - **ταίριασμα: M**
- Παράδειγμα: S_1 : vintner και S_2 : writers
- $\text{edit-distance}(S_1 \rightarrow S_2) = \text{RIMDMDMMI}$

Στοίχιση Ακολουθιών

- **Στοίχιση Ακολουθιών- Sequence Alignment:** τοποθετούμε τη μια ακολουθία κάτω από την άλλη έτσι ώστε οι κοινοί χαρακτήρες να τοποθετούνται στις ίδιες θέσεις.

-	V	I	N	T	N	E	R	-
		↓		↓		↓	↓	
W	R	I	-	T	-	E	R	S

Στοίχιση Ακολουθιών επιτρέποντας κενά

- Στοίχιση δυο ακολουθιών με την εισαγωγή 7 κενών χαρακτήρων σε 4 θέσεις, που μεταφράζεται ως μετάλλαξη της ακολουθίας του DNA στις αντίστοιχες θέσεις.

<i>c</i>	<i>t</i>	<i>t</i>	<i>t</i>	<i>a</i>	<i>a</i>	<i>c</i>	-	-	<i>a</i>	-	<i>a</i>	<i>c</i>
<i>c</i>	-	-	-	<i>c</i>	<i>a</i>	<i>c</i>	<i>c</i>	<i>c</i>	<i>a</i>	<i>t</i>	-	<i>c</i>