ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΑΝΟΙΚΤΑ ακαδημαϊκά μαθήματα ΠΠ

Εισαγωγή στη Βιοπληροφορική

Ενότητα 4: Τεχνικές Ανάλυσης και Σύγκρισης Ακολουθιών Βιολογικών Δεδομένων ΙΙ

Μακρής Χρήστος, Τσακαλίδης Αθανάσιος, Περδικούρη Αικατερίνη

Πολυτεχνική Σχολή

Τμήμα Μηχανικών Η/Υ και Πληροφορικής

Σκοποί ενότητας

- Η παρουσίαση των αλγορίθμων προσεγγιστικής εύρεσης προτύπου και στοίχισης συμβολοσειρών
- Η παρουσίαση των αλγορίθμων σύγκρισης ακολουθιών βιολογικών δεδομένων

Περιεχόμενα ενότητας

- Βασικοί ορισμοί
- Στοίχιση ακολουθιών
- Μέθοδος δυναμικού προγραμματισμού
- Προσεγγιστική εύρεση προτύπου
- Εφαρμογές στην ανάλυση ακολουθιών βιολογικών δεδομένων
- Αλγόριθμος BLAST
- Αλγόριθμος FASTA

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 ΙΙ

Τεχνικές Ανάλυσης και Σύγκρισης Ακολουθιών Βιολογικών Δεδομένων

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

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
- edit-distance(S_1->S_2)=5

- Λέγεται και Levenshtein distance, παραμένει το ίδιο είτε αν η ακολουθία πράξεων εφαρμόζεται στο S_1 είτε στο S_2.

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
- edit-distance(S_1->S_2)= RIMDMDMMI

Πηγή Dan Gusfield, Algorithms on Strings, Trees and Sequences, Cambridge University Press, 2010.

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

Πηγή Dan Gusfield, Algorithms on Strings, Trees and Sequences, Cambridge University Press, 2010.

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.**

c t t t a a c - - a - a c
c - - - c a c c c a t - c

Η Μέθοδος του Δυναμικού Προγραμματισμού

- **Δυναμικός Προγραμματισμός:**

  Έστω 2 ακολουθίες S_1 και S_2, θα συμβολίζουμε ως D(i,j) την απόσταση μετασχηματισμού μεταξύ των προθεμάτων $S_1[1..i]$ και $S_2[1..j]$, δηλαδή τον ελάχιστο αριθμό πράξεων μετασχηματισμού που απαιτούνται για να μετασχηματίσουμε τους i πρώτους χαρακτήρες της ακολουθίας S_1 στους j πρώτους χαρακτήρες της ακολουθίας S_2.

- Χρήση 3 βασικών τεχνικών:
  - σχέση αναδρομής- recurrence relation,
  - χρήση πίνακα- tabular computation,
  - σχέση οπισθοχώρησης- traceback.

Παράδειγμα Πίνακα Δυναμικού Πρ/σμου

D(i,j)			***w***	***r***	***i***	***t***	***e***	***r***	***s***
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
	0	0	1	2	3	4	5	6	7
v	**1**	1	1	2	3	4	5	6	7
i	**2**	2	2	2	2	3	4	5	6
n	**3**	3	3	3	3	3	4	5	6
t	**4**	4	4	4	4	*			
e	**5**	5							
r	**6**	6							
s	**7**	7							

Η Σχέση Αναδρομής

- **Σχέση Αναδρομής:**

$$D(i,j)=min[D(i-1,j)+1,D(i,j-1)+1,D(i-1,j-1)+t(i,j)]$$

  - **D(i,j-1)+1:** πρέπει να ενθέσουμε το χαρακτήρα $S_2[j]$
  - **D(i-1,j)+1:** πρέπει να διαγράψουμε το χαρακτήρα $S_1[i]$,
  - **D(i-1,j-1)+1:** για να μετασχηματίσουμε το χαρακτήρα $S_1[i]$ στο χαρακτήρα $S_2[j]$ πρέπει να αντικαταστήσουμε το χαρακτήρα $S_1[i]$, με το χαρακτήρα $S_2[j]$,
  - **D(i-1,j-1):** έχουμε ταίριασμα

Παράδειγμα: σχέση αναδρομής

D(i,j)			***w***	***r***	***i***	***t***	***e***	***r***	***s***
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
	0	0	1	2	3	4	5	6	7
v	**1**	1	1	2	3	4	5	6	7
i	**2**	2	2	2	2	3	4	5	6
n	**3**	3	3	3	3	3	4	5	6
t	**4**	4	4	4	4	*			
e	**5**	5							
r	**6**	6							
s	**7**	7							

D(4,4)= D(3,3)=3, αφού $S_1(4)=S_2(4)=t$.

Η Σχέση Οπισθοχώρησης

- **Σχέση Οπισθοχώρησης:**
  - από την (i,j) θέση προς την (i,j-1) αν $D(i,j)=D(i,j-1)+1$ (ένθεση χαρακτήρα)
  - από την (i,j) θέση προς την (i-1,j) αν $D(i,j)=D(i-1,j)+1$ (διαγραφή χαρακτήρα)
  - από την (i,j) θέση προς την (i-1,j-1) αν $D(i,j)=D(i-1,j-1)+t(i,j)$ (αντικατάσταση χαρακτήρα ή ταίριασμα)

Προσθήκη δεικτών οπισθοχώρησης

D(i,j)			w	r	i	t	e	r	s
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
	0	0	? 1	? 2	? 3	? 4	? 5	? 6	? 7
v	**1**	?1	? 1	? ? 2	? ? 3	? ? 4	? ? 5	? ? 6	? ? 7
i	**2**	?2	? ?2	? 2	? 2	? 3	? 4	? 5	? 6
n	**3**	?3	? ?3	? ?3	? ?3	? 3	? ?4	? ?5	? ? 6
t	**4**	?4	? ?4	? ?4	? ?4	? 3	? ?4	? ?5	? ? 6
e	**5**	?5	? ?5	? ?5	? ?5	?4	? 4	? ?5	? ? 6
r	**6**	?6	? ?6	? ?6	? ?6	?5	? 4	? ?5	? ? 6
s	**7**	?7	? ?7	? 6	? ? ?7	?6	?5	? 4	? 5

Ερμηνεία δεικτών οπισθοχώρησης

D(i,j)			w	r	i	t	e	r	s
		0	1	2	3	4	5	6	7
	0	0	? 1	? 2	? 3	? 4	? 5	? 6	? 7
v	1	?1	? 1	? ? 2	? ? 3	? ? 4	? ? 5	? ? 6	? ? 7
i	2	?2	? ?2	? 2	? 2	? 3	? 4	? 5	? 6
n	3	?3	? ?3	? ?3	? ?3	? 3	? ?4	? ?5	? ? 6
t	4	?4	? ?4	? ?4	? ?4	? 3	? ?4	? ?5	? ? 6
e	5	?5	? ?5	? ?5	? ?5	?4	? 4	? ?5	? ? 6
r	6	?6	? ?6	? ?6	? ?6	?5	? 4	? ?5	? ? 6
s	7	?7	? ?7	? 6	? ? ?7	?6	?5	? 4	? 5

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

Πολυπλοκότητα της μεθόδου Δυναμικού Προγραμματισμού

- Αρχικοποίηση: O(n) + O(m)
- Σχέση Αναδρομής: O(n*m)
- Δείκτες Οπισθοχώρησης: O(n+m)
- Πολυπλοκότητα: $O(n^2)$
- Ισοδυναμία με πρόβλημα της θεωρίας γραφημάτων, όπου κάθε κόμβος έχει ετικέτα ένα ζεύγος (i,j)

Βασικοί Ορισμοί (γ)

- **Ζυγισμένη Απόσταση Μετασχηματισμού - Weighted Edit Distance:** το ελάχιστο κόστος των πράξεων μετασχηματισμού που απαιτούνται για να μετασχηματίσουμε την πρώτη συμβολοσειρά στη δεύτερη. Κάθε πράξη μετασχηματισμού έχει συγκεκριμένο κόστος - βάρος. Έστω ότι οι βασικές πράξεις μετασχηματισμού έχουν τα ακόλουθα βάρη:
  - ένθεση ή διαγραφή: d
  - αντικατάσταση: r
  - ταίριασμα: m.
- Παράδειγμα: S_1: vintner και S_2: writers
- weighted edit-distance(S_1->S_2)= **r+4d+4m**.

Η Σχέση Αναδρομής με βάρη

- **Σχέση Αναδρομής:**

$$D(i,j)=min[D(i-1,j)+d,D(i,j-1)+d,D(i-1,j-1)+t(i,j)],$$

- όπου:
  - $t(i,j)= e$, αν $S_1(i)=S_2(j)$,
  - $t(i,j)=r$, αν $S_1(i)\neq S_2(j)$ και
  - $D(i,0)=i*d$ και $D(0,j)=j*d$.

Η σχέση μπορεί να υπολογιστεί κινούμενοι κατά στήλες από αριστερά προς τα δεξιά, αφού έχουμε υπολογίσει τη πρώτη γραμμή και τη πρώτη στήλη.

ΕΠΕΚΤΑΣΕΙΣ

- Ζυγισμένη Απόσταση Μετασχηματισμού βάσει Αλφαβήτου - Weighted Edit Distance.

- Σε εφαρμογές Μοριακής Βιολογίας τα βάρη αντικατάστασης χαρακτήρων αποθηκεύονται σε Πίνακες Αντικατάστασης - Substitution Matrix: PAM και BLOSUM

- Καλύτερη (σημασιολογικά) η αντιμετώπιση σαν alignment και η ενσωμάτωση του score.

- Κατά κανόνα match είναι θετικό, οτιδήποτε άλλο 0 ή αρνητικό

Υπολογισμός score-function στοίχισης ακολουθιών

1	2	3	4	5	6	7
g	*a*	*g*	*-*	*t*	*c*	*t*
g	*a*	*c*	*c*	*t*	*c*	*-*

S	**a**	**c**	**g**	**t**	**_**
a	1	-1	-2	0	-1
c		3	-2	-1	0
g			0	-4	-2
t				3	-1
_					0

- **Score-function=** 0+1-2+0+3+3-1=4

Δυναμικός Προγραμματισμός & Ομοιότητα Ακολουθιών βάσει αλφαβήτου

- **Σχέση Αναδρομής για την ομοιότητα ακολουθιών:**

$$V(i,j)= max[V(i-1,j-1)+s(S_1(i), S_2(j)), V(i-1,j)+ s(S_1(i),\_), V(i,j-1)+ s(\_,S_2(j))],$$

- όπου:
  - s(x,y): η τιμή στοίχισης του χαρακτήρα x με τον y
  - $V(0,j)=\Sigma s(\_,S_2(k)),\ 1<=k<=j$ και
  - $V(i,0)= \Sigma\ s(S_1(k),\_),\ 1<=k<=i.$

Επεκτάσεις

- Longest Common Subsequence (match weight 1, otherwise 0)
- End-space free variant that encourages one string to align in the interior of the other, or the suffix of one to align with the prefix of the other (initial conditions 0, everything is countable in the last row and the last column) – shotgun sequence assembly
- Approximate occurrence of P in T (the optimal alignment of P to a substring of T has distance δ from the optimal alignment) (initial conditions, as previously 0).

  -- locate a cell (n,j) with value greater than δ.

  -- traverse backpointers from (n,j) to (0,k).

  -- occurrence in T[k,j]

Πηγή Dan Gusfield, Algorithms on Strings, Trees and Sequences, Cambridge University Press, 2010.

Το Πρόβλημα Τοπικής Στοίχισης Επιθέματος- Local Suffix Alignment Problem

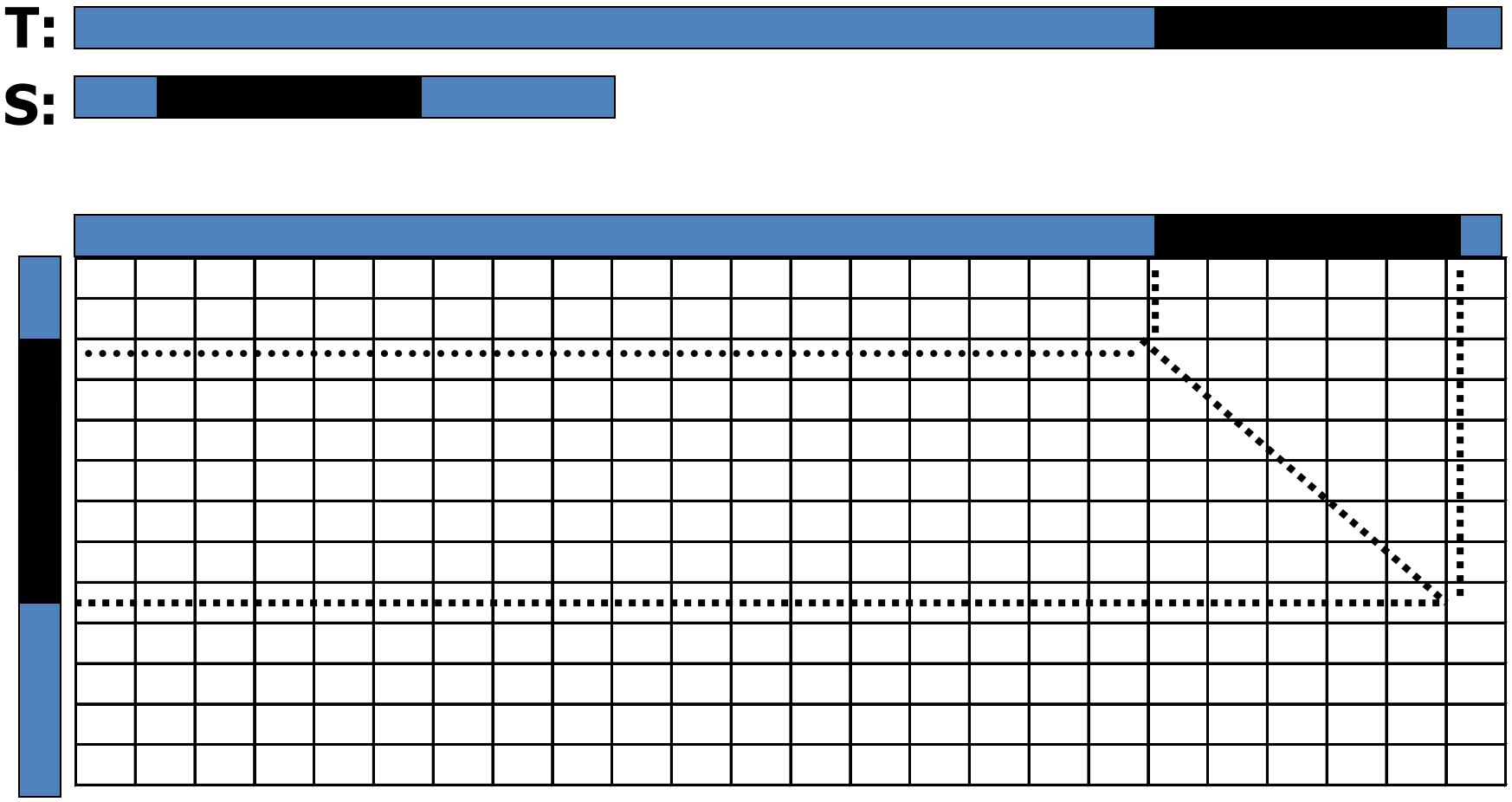
- **Local suffix alignment problem:** για δυο ακολουθίες S_1 και S_2 εντόπισε ένα επίθεμα α του $S_1[1..i]$ (με την πιθανότητα να είναι κενό) και ένα επίθεμα β του $S_2[1..j]$ (πιθανόν κενό) τέτοια ώστε το V(α,β) να έχει τη μέγιστη τιμή από όλα τα άλλα δυνατά ζεύγη επιθεμάτων των $S_1[1..i]$ και $S_2[1..j]$. Συμβολίζουμε ως υ(i,j) τη βέλτιστη τοπική στοίχιση επιθεμάτων για τις τιμές i και ,j (i<n και j<m).

- Αρχικές συνθήκες v(i,0)=0 και v(0,j)=0 καθώς μπορούμε να επιλέξουμε κάθε άδειο επίθεμα.

- $$v(i,j)=max[0,v(i-1,j-1)+s(S_1(i), S_2(j)), v(i-1,j)+ s(S_1(i),\_), v(i,j-1)+ s(\_,S_2(j))]$$

Πηγή Dan Gusfield, Algorithms on Strings, Trees and Sequences, Cambridge University Press, 2010.

Locally Similar Strings



Πηγή Dan Gusfield, Algorithms on Strings, Trees and Sequences, Cambridge University Press, 2010.

Παρατηρήσεις

- Η ολική στοίχιση καλείται Needleman-Wunsch στοίχιση.
- Η τοπική στοίχιση καλείται Smith-Waterman στοίχιση.
- Η Smith-Waterman στοίχιση μπορεί να εντοπίσει περιοχές υψηλής ομοιότητας πραγματοποιώντας trace-back from από κάθε κελί (i,j) προς τα πίσω έτσι ώστε να εντοπίσει ένα ζευγάρι με ομοιότητα v(i,j).

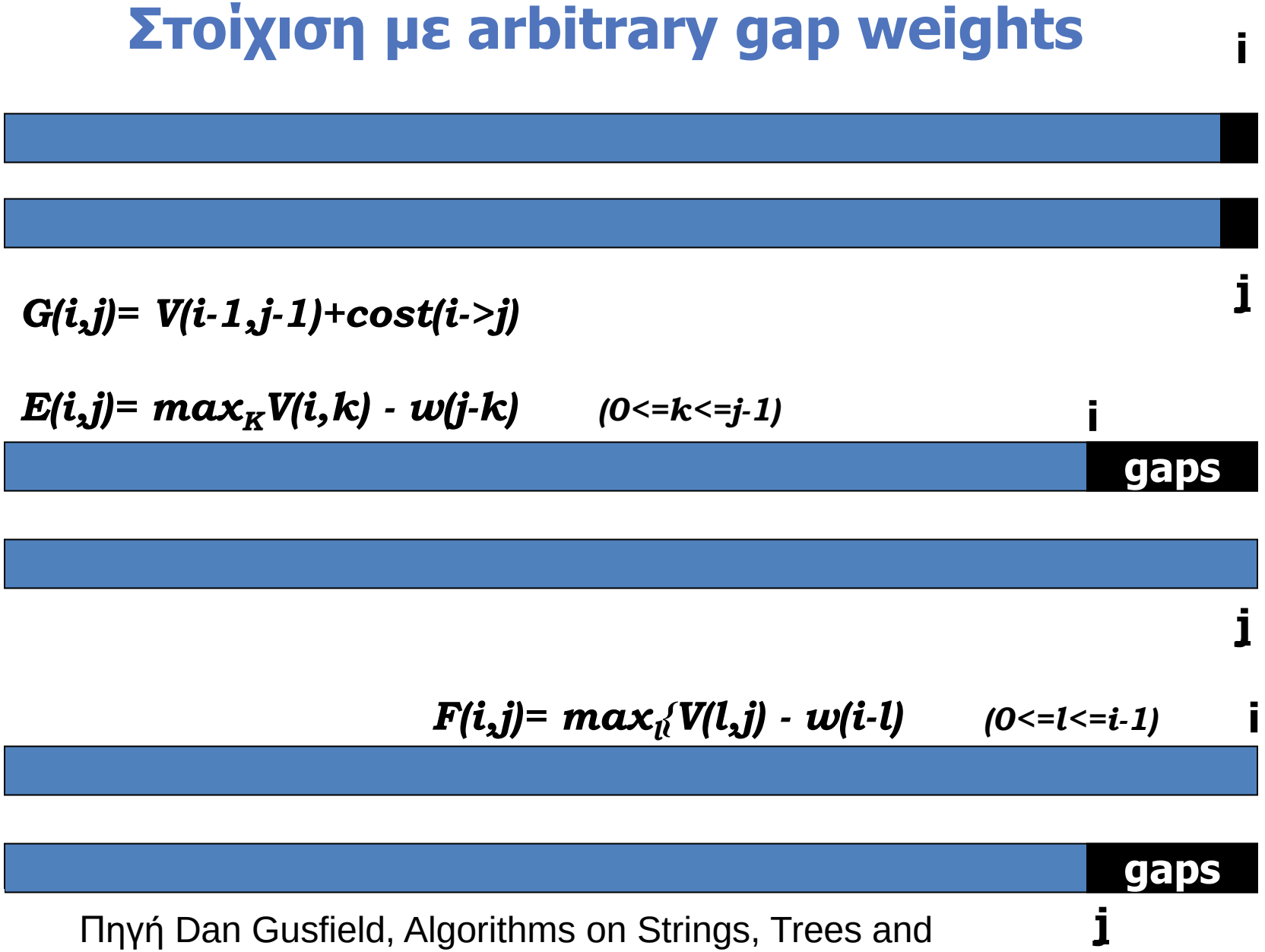
Στοίχιση Ακολουθιών με κενά

- **Έννοια κενού:** συνεχόμενα spaces, θέλουμε να ελέγχουμε την κατανομή των κενών.
- **Εισαγωγή Κενών:** Για να συμπεριλάβουμε το κόστος που η εισαγωγή κενών εισάγει στη στοίχιση 2 ακολουθιών, μπορούμε σε μια απλή προσέγγιση να θεωρήσουμε ότι κάθε κενό συνεισφέρει ένα σταθερό βάρος W_g, ανεξάρτητα από το μήκος του.
- τιμή στοίχισης που περιέχει "k" κενά: $\sum_{i=1}^{l} s(S_1'(j_i), S_2'(j_j)) - kW_g$

- μία καλύτερη προσέγγιση είναι η χρησιμοποίηση μίας συνάρτησης του μήκους του κενού. Τότε μπορούμε να γεμίσουμε ένα πίνακα: V(i,j)=max[E(i,j), F(i,j), G(i,j)]

Πηγή Dan Gusfield, Algorithms on Strings, Trees and Sequences, Cambridge University Press, 2010.

Στοίχιση με arbitrary gap weights



Πηγή Dan Gusfield, Algorithms on Strings, Trees and Sequences, Cambridge University Press, 2010.

- $V(i,j)=\max\{E(i,j), F(i,j), G(i,j)\}$
- $V(i,0)=-w(i)$
- $V(0,j)=-w(j)$
- $E(i,0)=-w(i)$
- $F(0,j)=-w(j)$
- $G(0,0)=0$
- Assuming that $|S_1|=n$ and $|S_2|=m$ the recurrences can be evaluated in $O(nm^2+n^2m)$

Πηγή Dan Gusfield, Algorithms on Strings, Trees and Sequences, Cambridge University Press, 2010.

Εφαρμογές σε Προβλήματα Μοριακής Βιολογίας

- **<u>Το Πρόβλημα της Πολλαπλής Στοίχισης- multiple sequence alignment problem:</u>** Μία πολλαπλή ολική στοίχιση από k>2 συμβολοσειρές S={ S_1, S_2,…., $S_κ$} είναι μία φυσική γενίκευση της στοίχισης για δύο συμβολοσειρές.

Γιατί μας ενδιαφέρει η πολλαπλή στοίχιση ακολουθιών

- Η πολλαπλή στοίχιση ακολουθιών χρησιμοποιείται:
  - στην αναγνώριση και αναπαράσταση πρωτεϊνικών οικογενειών και υπερ-οικογενειών,
  - στην αναπαράσταση των χαρακτηριστικών που μεταφέρονται στις ακολουθίες DNA ή στις πρωτεϊνικές ακολουθίες,
  - στην αναπαράσταση της εξελικτικής ιστορίας (φυλογενετικά δέντρα) από ακολουθίες DNA ή πρωτεϊνών.

Αλγόριθμοι Πολλαπλής Στοίχισης Ακολουθιών

- Είδη στοίχισης:
  - Extention of DP approach (too costly)
  - Use of pairwise alignment (center star algorithm)
- Αλγόριθμοι πολλαπλής στοίχισης ακολουθιών:
  - FASTA
  - BLAST.

Βιολογικές Βάσεις Δεδομένων (πηγή Wikipedia)

- Γενικευμένες (Generalised) ή Αρχειακές (Archival) βιολογικές βάσεις δεδομένων). Διακρίνονται σε:

  - Πρωτογενείς βάσεις δεδομένων ακολουθιών (Primary Sequence Databases). Περιέχουν νουκλεοτιδικές και αμινοξικές ακολουθίες από γονιδιώματα οργανισμών που είτε έχουν αποκρυπτογραφηθεί πλήρως είτε όχι

  - βάσεις δεδομένων που περιέχουν τρισδιάστατες δομές νουκλεϊνικών οξέων και πρωτεϊνών

- Δευτερευουσες (Secondary) βιολογικές βάσεις δεδομένων που προκύπτουν από ανάλυση των δεδομένων που είναι αποθηκευμένα στις αρχειακές βιολογικές βάσεις δεδομένων και διακρίνονται σε:

  ✓ Δευτερεύουσες ΒΔ ακολουθιών DNA και πρωτεϊνών που προκύπτουν από τις βασικές ΒΔ ακολουθιών και περιλαμβάνουν

  (α) ΒΔ ακολουθιών στις οποίες έχουν απομακρυνθεί οι ακολουθίες που έχουν αποθηκευτεί περισσότερες από μία φορές

  (β) ΒΔ που καταγράφουν μεταλλαγές ή παραλλαγές στις ακολουθίες DNA και πρωτεινών

  (γ) Γονιδιωματικές ΒΔ που είτε ομαδοποιούν συγγενή ή όχι πλήρως αποκρυπτογραφημένα γονιδιώματα είτε ασχολούνται με γονιδιώματα οργανισμών μοντέλων

- ✓ ΒΔ που ασχολούνται με τις ιεραρχήσεις ή/και συσχετίσεις μεταξύ βιομορίων όπως οικογένειες πρωτεϊνών, κοινές δομές πρωτεϊνών κοινά μοτίβα ακολουθιών DNA και πρωτεϊνών.

- ❑ Εξειδικευμένες Β.Δ., κατηγορία στην οποία ανήκουν:
  - ✓ Β.Δ. μικροσυστοιχιών που περιλαμβάνουν πληροφορίες για την έκφραση γονιδίων και πρωτεϊνών
  - ✓ Β.Δ. Μεταβολικών μονοπατιών που περιέχουν πληροφορίες για τις χημικές αντιδράσεις που πραγματοποιούνται στο κύτταρο
- ❑ Βιβλιογραφικές βιολογικές βάσεις δεδομένων
- ❑ Βιολογικές βάσεις δεδομένων ιστοσελίδων που περιλαμβάνουν:
  - ✓ Β.Δ. που περιλαμβάνουν ως εγγραφές βιολογικές βάσεις
  - ✓ Συνδέσμους μεταξύ βιολογικών βάσεων δεδομένων.

Βάσεις Βιολογικών Δεδομένων

- GenBank: NCBI (http://www.ncbi.nlm.nih.gov)
- PIR: Protein Information Resource (http://pir.georgetown.edu)
- Swiss-Prot + TrEMBL: Swiss-Prot.htm (http://tw.expasy.org/sprot/)
- PRINTS: (http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/)
- PROSITE: Prosite (http://tw.expasy.org/prosite/)
- PDB-Protein Data Bank: PDB (http://www.rcsb.org/pdb/)
- SCOP: (http://scop.berkeley.edu/)

  Structural Classification of Proteins
- SCOP: (http://scop.berkeley.edu/)

Sequence Database Searching

Βήματα καθορισμού πρωτεϊνικής ακολουθίας

1. Σύγκριση της νέας ακολουθίας με PROSITE και BLOCKS για εύρεση well-characterized sequence motifs.
2. Ψάξιμο στις DNA και protein sequence databases (Genbank, Swiss-Prot, etc.) για εντοπισμό ακολουθιών τοπικά παρόμοιων (με χρήση ενός κριτηρίου τοπικής ομοιότητας) – χρήση FASTA και BLAST
3. Εάν τα παραπάνω ψαξίματα δίνουν ενδιαφέρον αποτέλεσμα, τότε καταφεύγουμε στη χρήση της τεχνικής του δυναμικού προγραμματισμού.
4. Όταν χρειαστεί να εμπλακούν amino acid substitution matrices, συνήθως χρησιμοποιείται μία παραλλαγή του Dayhoff PAM matrix και του BLOSUM matrix.

Ο Αλγόριθμος BLAST (Συντηρείται από το NCBI)

- BLAST: Basic Local Alignment Search Tool, Altschul et. Al. 1990
- Βασική ιδέα: εντοπισμός κοινών υπο-ακολουθιών ίδιου μήκους (segment pairs) που εμφανίζονται και στη δοσμένη ακολουθία μικρού μήκους (input query sequence) και στο σύνολο των ακολουθιών μίας βάσης δεδομένων. Στη συνέχεια επέκταση για εύρεση maximal segment pairs

Αλγόριθμος	Είδος ερώτησης	Είδος ακολουθίας
BLASTP	Πρωτείνη	Πρωτεϊνη
BLASTN	Νουκλεοτίδιο	Νουκλεοτίδιο
BLASTX	Νουκλεοτίδιο	Πρωτεϊνη
TBLASTN	Πρωτεϊνη	Νουκλεοτίδιο
TBLASTX	Νουκλεοτίδιο	Νουκλεοτίδιο

1° βήμα: τμηματοποίηση της δοσμένης ακολουθίας σε διαδοχικές υπο-λέξεις μεγέθους w=3																								
Query sequence:																								
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y
Words																								
a	b	c																						
	b	c	d																					
		c	d	e																				
			d	e	f																			
2° βήμα: Εντοπισμός των υπο-λέξεων με μέγιστη τιμή στοίχισης για το όλες τις ακολουθίες																								
High-scoring matching words:																								
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y
a	b	c																						
d	s	k	o	w	j	j	d	f	k	s	l	m	n	k	d	k	j	d	f	k	k	j	d	f
											l	m	n											
m	s	l	z	m	s	o	w	u	r	n	f	k	s	a	d	e	f	a	q	m	a	z	m	s
															d	e	f							
3° βήμα: επέκταση των high-scoring words																								
a	b	c	d	w	f	h	h	f	j	s	l	m	n	k	d	k	j	d	e	h	k	k	j	f
a	b	c	=							⇐	l	m	n	=										

Ο Αλγόριθμος FASTA

(φιλοξενείται στο EBI)

- FASTA: Fast – All, Lipman et al. 1985
- Κεντρική ιδέα: εύρεση μικρών λέξεων (words ή k-tuples) που εμφανίζονται και στις δύο ακολουθίες. Στην περίπτωση πρωτεϊνικών ακολουθιών το μήκος των λέξεων είναι 1-2 βάσεις ενώ για ακολουθίες DNA το μήκος μίας λέξης μπορεί να φτάνει στις 6 βάσεις

Τα βήματα του αλγορίθμου FASTA

- 1ο βήμα: αναζητούμε λέξεις μήκους ktup στον πίνακα δυναμικού προγραμματισμού: 'hot spots' (pairs *(i,j)*)
- 2ο βήμα: εντοπίζουμε τις δέκα καλύτερες διαγώνιες τροχιές – diagonal runs από 'hot-spots' στον πίνακα

  (a hot spot define the (i-j)-diagonal, the score is the sum of scores of hot-spots plus weighted decreasing as the distance increases)
- 3ο βήμα: συνδυάζουμε «καλές υπο-στοιχίσεις»
- 4ο βήμα: παράγουμε το βέλτιστο μονοπάτι

Indexing Approaches

- k-gram indexing
- direct indexing
- vector space indexing

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.