ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ AGRICULTURAL UNIVERSITY OF ATHENS

Βιοπληροφορική

Ενότητα 7:

Στοίχιση ακολουθιών ανά ζεύγη – Τεχνικές Στοίχισης Ακολουθιών,(2/2) 2 ΔΩ

Τμήμα: Βιοτεχνολογίας

Όνομα καθηγητή: **Τ. Θηραίου**

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Μαθησιακοί Στόχοι

- Παρουσίαση της μεθόδου κατασκευής και των εφαρμογών των διαγραμμάτων πινάκων σημείων.
- Επεξήγηση των αλγορίθμων δυναμικού προγραμματισμού για τη στοίχιση ακολουθιών ανά ζεύγη.
- Κατανόηση της διαφοράς μεταξύ της βέλτιστης και της βιολογικά σωστής στοίχισης.

Λέξεις Κλειδιά

- Λέξεις κλειδιά: Διαγράμματα Πινάκων Σημείων, Αλγόριθμοι Δυναμικού Προγραμματισμού.

- Key words: Dot Matrix Plots, Dynamic programming algorithms for pairwise sequence alignment, Needleman-Wunsch algorithm, Smith-Waterman algorithm.

Μέθοδοι Δυναμικού Προγραμματισμού 1/2

- Δημιουργία βέλτιστης στοίχισης χρησιμοποιώντας τις βέλτιστες στοιχίσεις μικρότερων ακολουθιών.
  - **Ολική Στοίχιση (Needleman-Wunsch)**.
  - **Τοπική Στοίχιση (Smith-Waterman)**.
- 2 ακολουθίες x και y.
- πίνακας **F(i,j)**: score της βέλτιστης στοίχισης μεταξύ του αρχικού τμήματος $x_{1...i}$ της x μέχρι το κατάλοιπο x_i και του αρχικού τμήματος $y_{1...j}$ της y μέχρι το κατάλοιπο y_j.
- "γέμισμα" του F(i,j): επαναληπτική διαδικασία με F(0,0) = 0.

Ολική Στοίχιση 1/9

Η πρώτη γραμμή και η πρώτη στήλη του πίνακα συμπληρώνεται βάσει της; σχέσης:

$F_{i,0} = -i \times gap$, $F_{0j} = -j \times gap$

		A	G	T	A
	0	-1×d	-2×d	-3×d	-4×d
A	-1×d				
T	-2×d				
A	-3×d				

για απλότητα d = |gap|

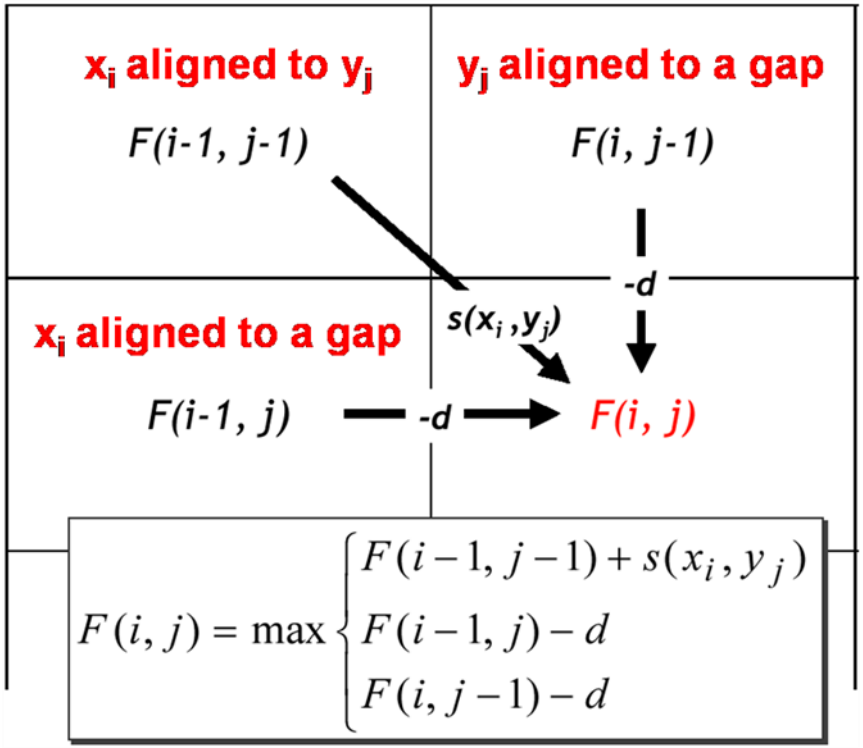
Ολική Στοίχιση 2/9

- Οι τιμές των άλλων κελιών υπολογίζονται **κατά γραμμή**, ξεκινώντας από την επάνω αριστερή γωνία και προχωρώντας συνεχώς **προς τα δεξιά**.
- Όταν φτάνουμε στο τέλος μιας γραμμής ξεκινάμε από το πρώτο κελί της επόμενης.

		A	G	T	A
	0	-1×d	-2×d	-3×d	-4×d
A	-1×d				
T	-2×d				
A	-3×d				

Ολική Στοίχιση 3/9

- Παράλληλα με τον υπολογισμό της τιμής ενός κελιού, φυλάσσεται σ' ένα **πίνακα "ιχνηθέτη"** (trace-back) η διεύθυνση του κελιού βάσει του οποίου υπολογίστηκε η τιμή (από το διαγώνιο κελί, το αριστερό ή το κατακόρυφο).



Ολική Στοίχιση 4/9

- match = 1, mismatch = -1, gap = -1.

F(i,j)	i = 0	1	2	3	4
		A	G	T	A
j = 0	0	-1	-2	-3	-4
1 A	-1	1	0	-1	-2
2 T	-2	0	0	1	0
3 A	-3	-1	-1	0	2

Ολική Στοίχιση 5/9

- Η τιμή στο **κάτω δεξιά κελί** αποτελεί τη **βαθμολογία** της καλύτερης δυνατής στοίχισης των δύο ακολουθιών με βάση το συγκεκριμένο σύστημα βαθμολόγησης.
- Ακολουθώντας τη **διαδρομή στον πίνακα ιχνηθέτη** από το κάτω δεξιά κελί παράγουμε τη **συγκεκριμένη στοίχιση**.
- Για κάθε θέση:
  - Αν κινηθούμε **διαγώνια**, τότε **στοιχίζουμε τα δύο κατάλοιπα** που αντιστοιχούν σ' εκείνη την θέση.
  - Αν κινηθούμε **οριζόντια ή κάθετα**, βάζουμε **κενό** στην ακολουθία που δείχνει το βέλος.

Ολική Στοίχιση 6/9

F(i,j)	i = 0		1	2	3	4
			A	G	T	A
j = 0		0	-1	-2	-3	-4
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3	-1	-1	0	2

στοίχιση

A G T A

A – T A

score της στοίχισης

score **= 2**

Ολική Στοίχιση 7/9

- από διαγώνιο κελί = 0 + 2 = 2.
- από αριστερό κελί = 1 – 1 = 0.
- από κατακόρυφο κελί = 1 - 1 = 0.

		T	G	C	A	A	T	C	G	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	0	-1	-2	-1	-2	-3	-4	-5	-6
A	-2	-1	0	-1	0	1	0	-1	-2	-3
C	-3	-2	-1	2	1	0	1	2	1	0
T	-4	-1	-2	1	2	1	2	1	2	1
G	-5	-2	1	0	1	2	1	2	3	4
A	-6	-3	0	1	2	3	2	1	2	3
A	-7	-4	-1	0	3	4	3	2	1	2
T	-8	-5	-2	-1	2	3	6	5	4	3
C	-9	-6	-3	0	1	2	5	8	7	6

- max(2, 0, 0) = 2

- match = 2
- mismatch = 0
- gap = -1

Ολική Στοίχιση 8/9

```
---TGCAATCGG
AACTG-AATC--
```

← Βέλτιστη Στοίχιση

		T	G	C	A	A	T	C	G	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	0	-1	-2	-1	-2	-3	-4	-5	-6
A	-2	-1	0	-1	0	1	0	-1	-2	-3
C	-3	-2	-1	2	1	0	1	2	1	0
T	-4	-1	-2	1	2	1	2	1	2	1
G	-5	-2	1	0	1	2	1	2	3	4
A	-6	-3	0	1	2	3	2	1	2	3
A	-7	-4	-1	0	3	4	3	2	1	2
T	-8	-5	-2	-1	2	3	6	5	4	3
C	-9	-6	-3	0	1	2	5	8	7	6

Score Βέλτιστης Στοίχισης = 6

Ολική Στοίχιση 9/9

- Στοίχιση των ακολουθιών AGTA και ATA με τη χρήση δυναμικού προγραμματισμού, του **πίνακα αντικατάστασης** Blosum62 και linear gap penalty = -1

- Αντί για match και mismatch score, χρησιμοποιούνται τα scores που είναι καταγεγραμμένα στον **πίνακα αντικατάστασης**.

Τοπική Στοίχιση 1/3

- **F(i, 0) = F(0, j) = 0**
- Στοιχίσεις με αρνητική βαθμολογία δεν παρουσιάζουν ενδιαφέρον.
- Μια βέλτιστη τοπική στοίχιση κατασκευάζεται **ξεκινώντας** από το **κελί με τη μεγαλύτερη τιμή** (όπου κι αν βρίσκεται) και **τερματίζεται** μόλις συναντήσουμε για πρώτη φορά **μηδενική τιμή**.
- Το **score** της τοπικής στοίχισης ισούται με τη **μεγαλύτερη τιμή** του πίνακα.

Ολική-Τοπική Στοίχιση 1/5

Ολική Στοίχιση

		X_1	X_2	X_3	X_4
	0	-1×d	-2×d	-3×d	-4×d
Y_1	-1×d				
Y_2	-2×d				
Y_3	-3×d				

Τοπική Στοίχιση

		X_1	X_2	X_3	X_4
	0	0	0	0	0
Y_1	0				
Y_2	0				
Y_3	0				

d = |gap|

Ολική-Τοπική Στοίχιση 2/5

- Ολική Στοίχιση
  - **F(i, j) = max**
    - **F(i-1, j-1) + s(x_i, y_j)**
    - **F(i-1, j) – |gap|**
    - **F(i, j-1) – |gap|**

- Τοπική Στοίχιση
  - **F(i, j) = max**
    - **F(i-1, j-1) + s(xi, yj)**
    - **F(i-1, j) – |gap|**
    - **F(i, j-1) – |gap|**
    - **0**

Ολική-Τοπική Στοίχιση 3/5

Ολική Στοίχιση

		X_1	X_2	X_3	X_4
	0	-1	-2	-3	-4
Y_1	-1	3	0	-1	-5
Y_2	-2	0	2	6	0
Y_3	-3	-5	-2	0	4

Τοπική Στοίχιση

		X_1	X_2	X_3	X_4
	0	0	0	0	0
Y_1	0	1	3	0	1
Y_2	0	0	0	7	0
Y_3	0	1	0	0	5

Τοπική Στοίχιση 2/3

- από διαγώνιο κελί = 3 + 2 = 5.
- από αριστερό κελί = 4 – 1 = 3.
- από κατακόρυφο κελί = 2 - 1 = 1.
- 0.

- max(5, 3, 1, 0) = 5

- match = 2
- mismatch = 0
- gap = -1

		T	G	C	A	A	T	C	G	G
	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	2	2	1	0	0	0
A	0	0	0	0	2	4	3	2	1	0
C	0	0	0	2	1	3	4	5	4	3
T	0	2	1	1	2	2	5	4	5	4
G	0	1	4	3	2	2	4	5	6	7
A	0	0	3	4	5	4	3	4	5	6
A	0	0	2	3	6	7	6	5	4	5
T	0	2	1	2	5	6	9	8	7	6
C	0	1	2	3	4	5	8	11	10	9

Τοπική Στοίχιση 3/3

TGCAATC
TG-AATC

← Βέλτιστη Στοίχιση

		T	G	C	A	A	T	C	G	G
	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	2	2	1	0	0	0
A	0	0	0	0	2	4	3	2	1	0
C	0	0	0	2	1	3	4	5	4	3
T	0	2	1	1	2	2	5	4	5	4
G	0	1	4	3	2	2	4	5	6	7
A	0	0	3	4	5	4	3	4	5	6
A	0	0	2	3	6	7	6	5	4	5
T	0	2	1	2	5	6	9	8	7	6
C	0	1	2	3	4	5	8	11	10	9

Score Βέλτιστης Στοίχισης = 11

Μέθοδοι Δυναμικού Προγραμματισμού 2/2

- Οι αλγόριθμοι δυναμικού προγραμματισμού εγγυώνται τη **βέλτιστη στοίχιση** για τις **παραμέτρους** που χρησιμοποιούνται.
- Διαφορετικές παράμετροι ⇨ διαφορετική στοίχιση.
- Η βέλτιστη στοίχιση δεν είναι απαραίτητα και η βιολογικά σωστή.

BLOSUM62

gap opening penalty = **-3** gap extension penalty = -0.1 score = **6.3**	1 ...VLSPADKFLTNV 12 |||| 1 VFTELSPAKTV.... 11
gap opening penalty = **0** gap extension penalty = -0.1 score = **11.3**	1 V...LSPADKFLTNV 12 | |||| | | | 1 VFTELSPA.K..T.V 11

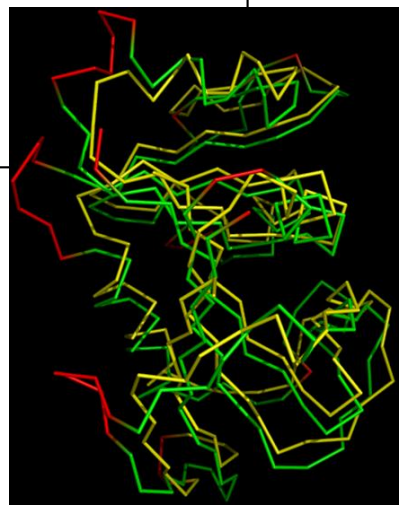
Στοίχιση Βάσει Της Δομικής Υπέρθεσης

```
D_HUM vgSLNCIVAVSQNMGIGKNGDLpWPPLRNEFRYFQRMTTtssvegkQNLVIMGKKTWFSI
ident       | |      ||          |      | | |           ||||  || ||
D_ECO --MISLIAALAVDRVIGMENAM-PFNLPADLAWFKRNTL-------DKPVIMGRHTWESI

D_HUM PeknRPLKGRINLVLSRELkEPPQGAhFLSRSLDDALKLTEqpelanKVDMVWIVGGSSV
ident     ||| || |  ||                | | |            |      ||  |
D_ECO G---RPLPGRKNIILSSQP-GTDDRV-TWVKSVDEAIAACG------DVPEIMVIGGGRV

D_HUM YKEAMNHpghLKLFVTRIMQDFESDTFFPEIDLEKYKLLPeypgvlSDVQEE---KGIKY
ident |           ||  | |    | || ||                  |           |
D_ECO YEQFLPK--aQKLYLTHIDAEVEGDTHFPDYEPDDWESVF------SEFHDAdaqNSHSY

D_HUM KFEVYEKNd
ident  ||  |
D_ECO CFEILERR-
```



Προγράμματα Στοίχισης Ακολουθιών

- Δυναμικός Προγραμματισμός:
  - EMBOSS Needle. http://www.ebi.ac.uk/Tools/psa/emboss_needle/
  - EMBOSS Water. http://www.ebi.ac.uk/Tools/psa/emboss_water/

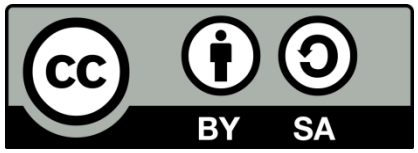
Βιβλιογραφία

- David Mount, "Bioinformatics: Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press; 2nd edition (March 12, 2013).
- Jonathan Pevsner, "Bioinformatics and Functional Genomics", Wiley-Blackwell; 2nd edition (May 4, 2009).
- Andreas D. Baxevanis, B. F. Francis Ouellette, "Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins", Wiley-Interscience; 3rd edition (October 29, 2004).

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.

- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδεια χρήσης, η άδεια χρήσης αναφέρεται ρητώς.



Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στο πλαίσιο του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα Γεωπονικού Πανεπιστημίου Αθηνών**» 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.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

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.