ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ AGRICULTURAL UNIVERSITY OF ATHENS

# Βιοπληροφορική

## Ενότητα 12:

Μέθοδοι Πολλαπλής Στοίχισης, 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

- Κατανόηση των μεθόδων πολλαπλής στοίχισης.
- Ανάδειξη των πλεονεκτημάτων και μειονεκτημάτων των τεχνικών αυτών.

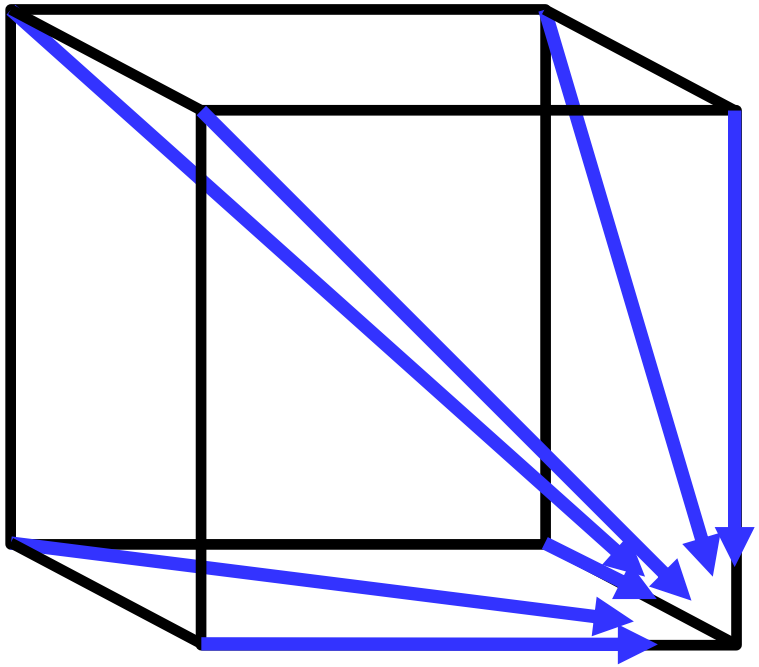
# Λέξεις Κλειδιά

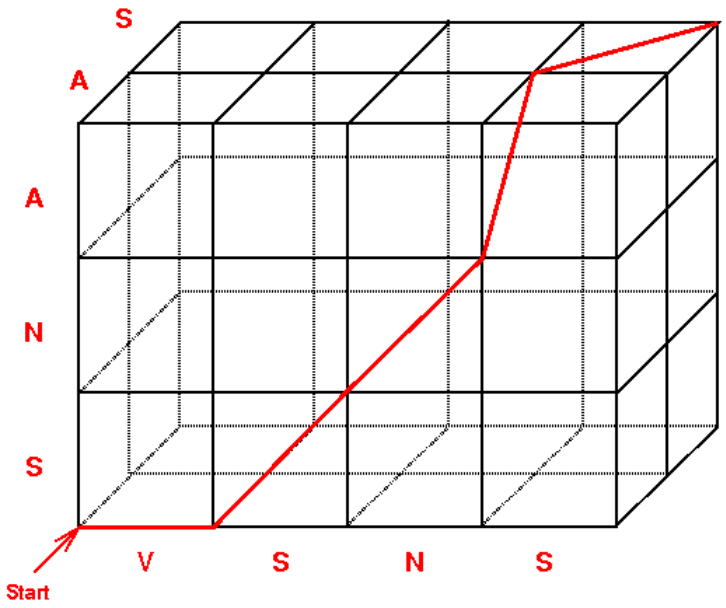
- Λέξεις κλειδιά: Στοίχιση άστρου, Προοδευτική πολλαπλή στοίχιση, Επαναληπτική στοίχιση.

- Key words:Star alignment, Progressive multiple sequence alignment, Iterative alignment.

# Μέθοδοι Πολλαπλής Στοίχισης

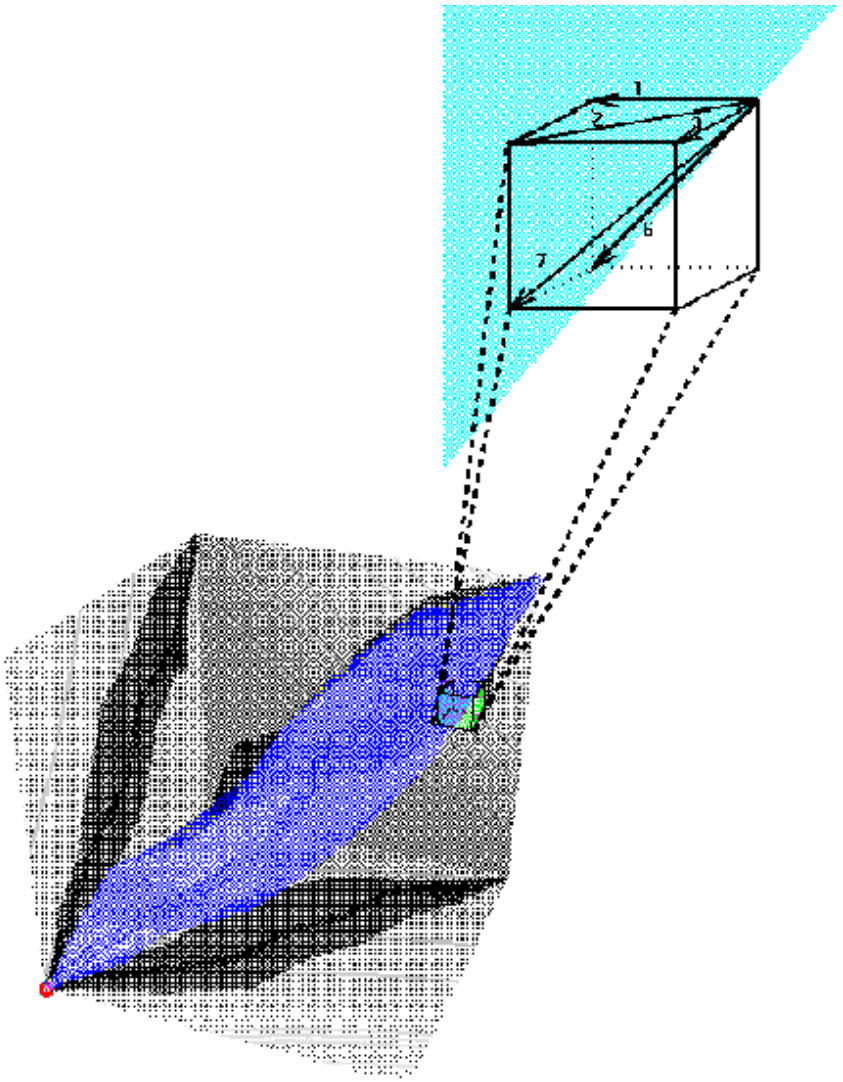
- Δυναμικός προγραμματισμός (dynamic programming)






V S N _ S
_ S N A _
_ _ _ A S

# Carillo and Lipman



# Στοίχιση Άστρου 1/5

δεδομένα

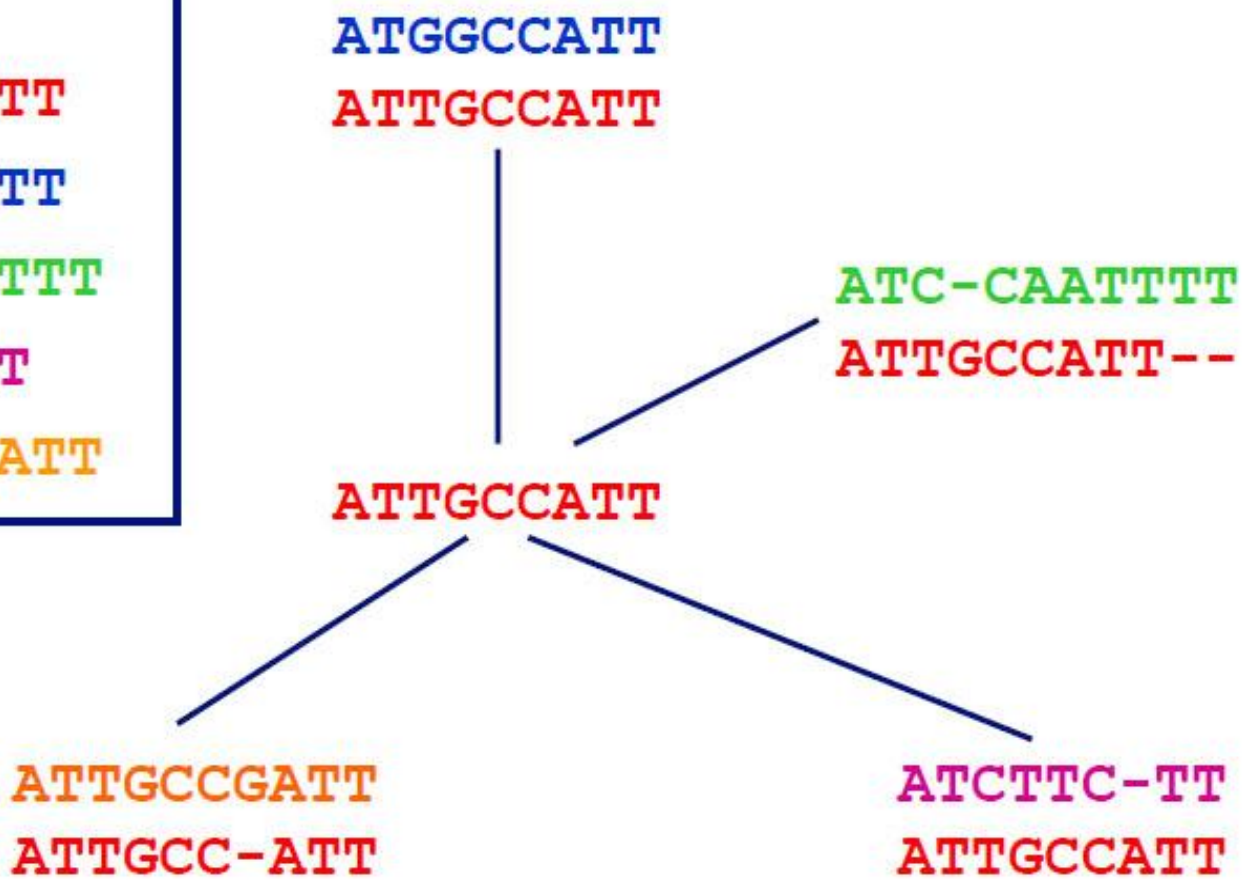
ATTGCCATT

ATGGCCATT

ATCCAATTTT

ATCTTCTT

ATTGCCGATT



# Στοίχιση Άστρου 2/5

| ζεύγος ακολουθιών προς ενσωμάτωση | στοίχιση |
|---|---|
| 1. ATGGCCATT<br>ATTGCCATT | ATTGCCATT<br>ATGGCCATT |
| 2. ATC-CAATTTT<br>ATTGCCATT-- | ATTGCCATT--<br>ATGGCCATT--<br>ATC-CAATTTT |

# Στοίχιση Άστρου 3/5

ζεύγος ακολουθιών προς ενσωμάτωση στοίχιση

3.

```
ATCTTC-TT
ATTGCCATT
```

```
ATTGCCATT--
ATGGCCATT--
ATC-CAATTTT
ATCTTC-TT--
```

4.

```
ATTGCCGATT
ATTGCC-ATT
```

```
ATTGCC- A TT--
ATGGCC- A TT--
ATC-CA- A TTTT
ATCTTC- - TT--
ATTGCCG A TT--
```

Μετακίνηση ολόκληρης στήλης όταν εισέρχεται κενό

# Στοίχιση Άστρου 4/5

```
ATTGCC- A TT--
ATGGCC- A TT--
ATC-CA- A TTTT
ATCTTC- - TT--
ATTGCCG A TT--
```

ATTGCC-ATT

ATGGCC-ATT

ATCCAATTTT

ATCT-T-CTT

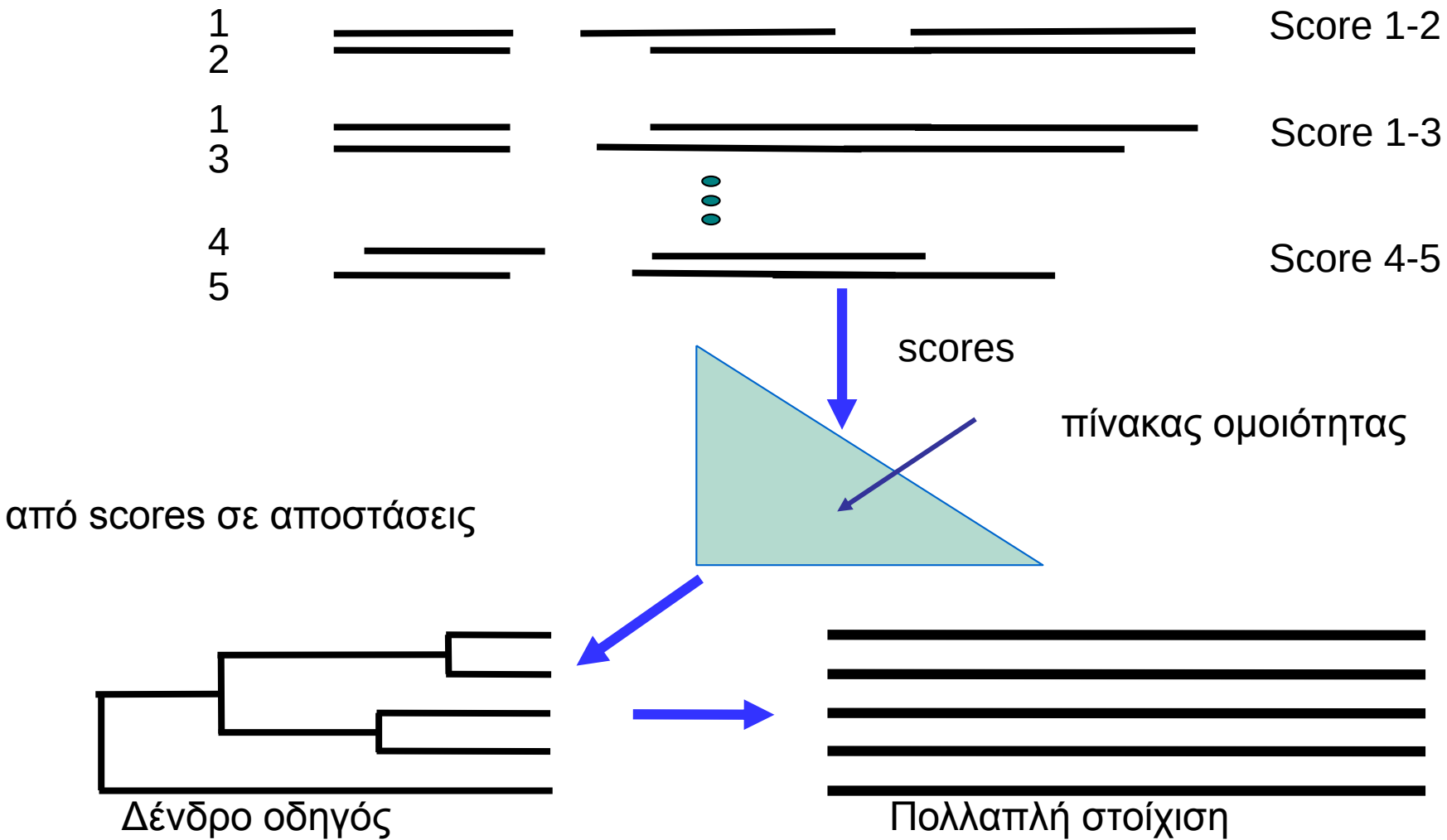
ATTGCCGATT

# Στοίχιση Άστρου 5/5

- Επιλογή μιας ακολουθίας ($x_c$) ως το "**κέντρο**" του άστρου
- Για κάθε μια από τις $x_1, \ldots, x_k$ με $i \neq c$
  - **Στοίχιση ΔΠ**
  - "**Ενσωμάτωση**" στοιχίσεων με την αρχή "**once a gap, always a gap**".
- Για την επιλογή του "κέντρου" του άστρου
  - Δοκιμή όλων και **επιλογή** εκείνης της ακολουθίας $x_c$ που μεγιστοποιεί: $\sum_{i \neq c} \mathrm{sim}(x_i, x_c)$

# Προοδευτική στοίχιση (progressive alignment)

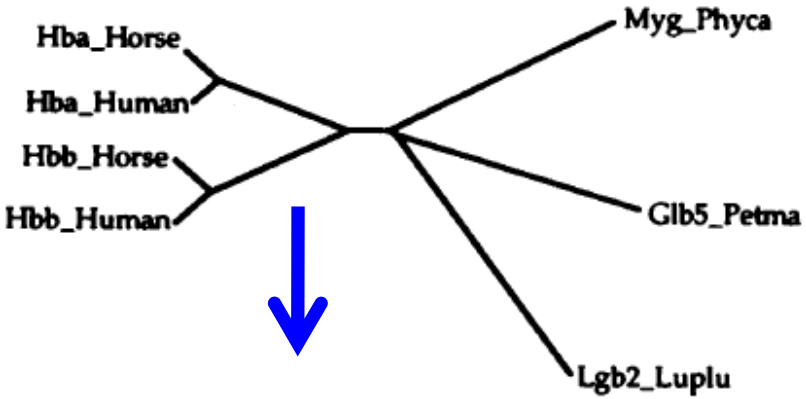
- στοίχιση βάσει ενός δέντρου οδηγού

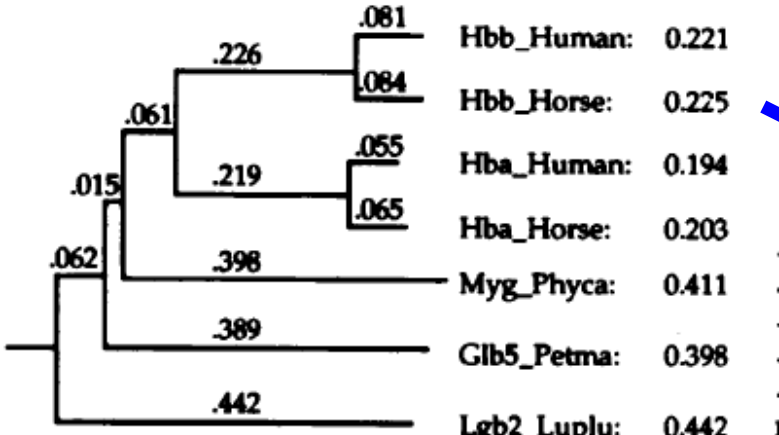


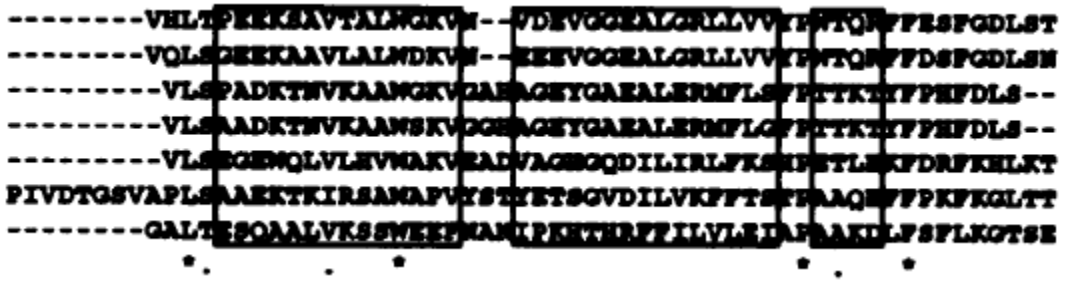
# ClustalW 1/9

- Προοδευτική στοίχιση

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Hbb_Human | 1 | - | | | | | |
| Hbb_Horse | 2 | .17 | - | | | | |
| Hba_Human | 3 | .59 | .60 | - | | | |
| Hba_Horse | 4 | .59 | .59 | .13 | - | | |
| Myg_Phyca | 5 | .77 | .77 | .75 | .75 | - | |
| Glb5_Petma | 6 | .81 | .82 | .73 | .74 | .80 | - |
| Lgb2_Luplu | 7 | .87 | .86 | .86 | .88 | .93 | .90 |
| | | 1 | 2 | 3 | 4 | 5 | 6 |



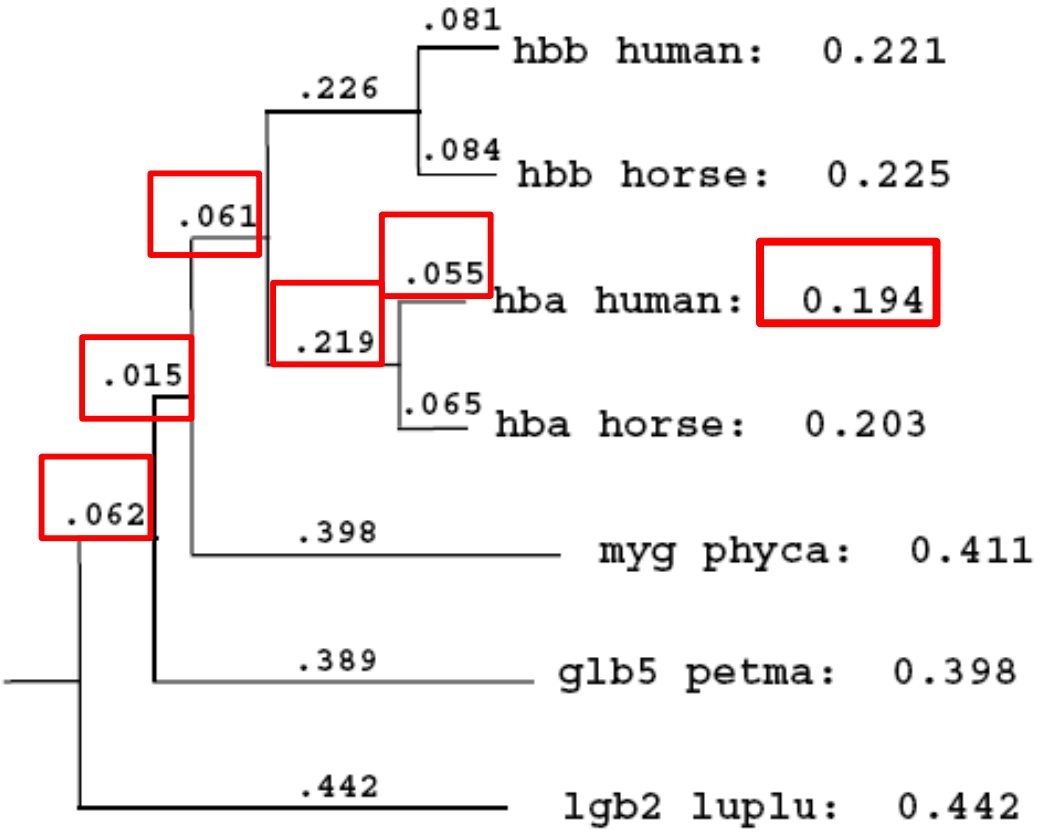



# ClustalW 2/9

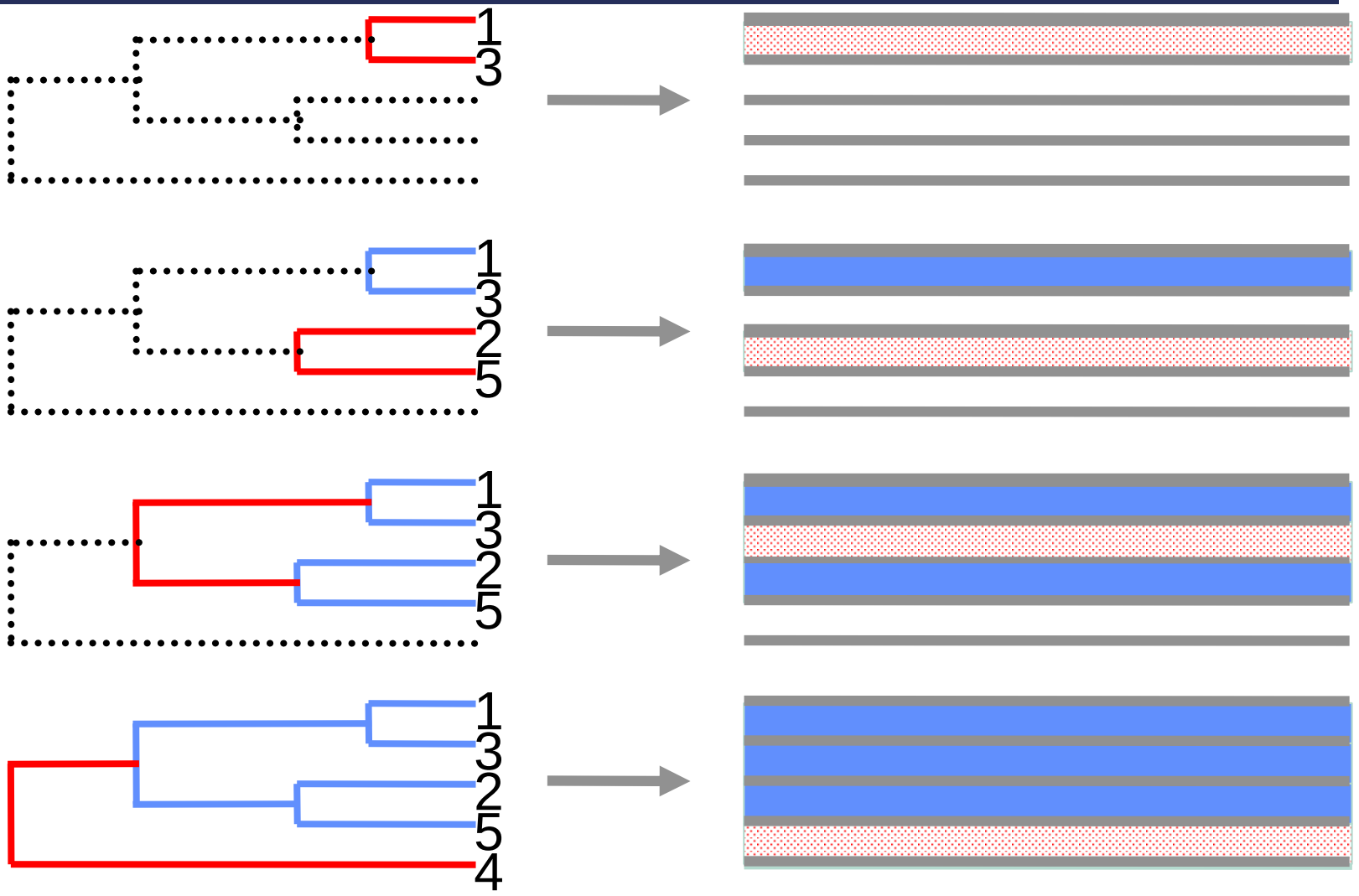
- στατιστικό βάρος βάσει των αποστάσεων των ακολουθιών στο δέντρο-οδηγό

W =

0.055/1 +
0.219/2 +
0.061/4 +
0.015/5 +
0.062/6 =
0.194



# ClustalW 3/9



# ClustalW 4/9

$S_{3,4}$
S3: ATACCCTG
S4: ATACCCGG

$S_{1,2}$
S2: AAACCGG
S1: AAATCGG

| | | AA | TT | AA | CC | CC | CC | TG | GG |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |
| AA | | | | | | | | | |
| AA | | | | | | | | | |
| AA | | | | | | | A | B | |
| CT | | | | | | | C | D | |
| CC | | | | | | | | | |
| GG | | | | | | | | | |
| GG | | | | | | | | | |

$$D = \max \begin{cases} A + SP(CT, TG) \\ B - g \\ C - g \end{cases}$$

# ClustalW 5/9

Ακολουθία A (βάρος **a**) .....**K**..... (στοίχιση 1)

Ακολουθία B (βάρος **b**) .....**I**..... (στοίχιση 1)

Ακολουθία C (βάρος **c**) .....**L**..... (στοίχιση 2)

Ακολουθία D (βάρος **d**) .....**V**.....(στοίχιση 2)

- χωρίς στάθμιση με τη χρήση στατιστικών βαρών
  - μέσος όρος των scores των ανά δύο στοιχίσεων των ακολουθιών της πρώτης ομάδας με εκείνες της δεύτερης

Score= {score(K,L ) + score(I,L) + score(K,V) + score(I,V)}/4

# ClustalW 6/9

Ακολουθία A (βάρος **a**) .....**K**..... (στοίχιση 1)

Ακολουθία B (βάρος **b**) .....**I**..... (στοίχιση 1)

Ακολουθία C (βάρος **c**) .....**L**..... (στοίχιση 2)

Ακολουθία D (βάρος **d**) .....**V**.....(στοίχιση 2)

- στάθμιση με τη χρήση στατιστικών βαρών
  - μέσος όρος των σταθμισμένων scores των ανά δύο στοιχίσεων των ακολουθιών της πρώτης ομάδας με εκείνες της δεύτερης

Score= {a*c*score(K,L ) + b*c*score(I,L) + a*d*score(K,V) + b*d*score(I,V)}/4

# ClustalW 7/9

- Ολική στοίχιση όλων των ακολουθιών ανά δύο με δυναμικό προγραμματισμό και μετατροπή των scores ομοιότητας σε εξελικτικές αποστάσεις.
- Δημιουργία του δέντρου-οδηγού βάσει του πίνακα αποστάσεων με τη μέθοδο Neighbor-joining.
  - χαμηλότερης ακρίβειας από ένα φυλογενετικό δέντρο
  - απόδοση στατιστικών βαρών (weights) στις ακολουθίες

# ClustalW 8/9

- Προοδευτική στοίχιση βάσει του δέντρου-οδηγού με δυναμικό προγραμματισμό
  - στοίχιση
    - ακολουθίας με ακολουθία
    - ακολουθίας με στοίχιση
    - στοίχισης με στοίχιση
  - στάθμιση με τη χρήση των στατιστικών βαρών από το δέντρο-οδηγό

# ClustalW 9/9

+ Χρήση διαφορετικού πίνακα αντικατάστασης ανάλογα με την απόσταση των ακολουθιών στο δέντρο-οδηγό
+ Προσαρμογή των ποινών για τα κενά
  - ομοιότητα και μήκος ακολουθιών
  - προϋπάρχοντα κενά
  - γειτνίαση με κενά
  - υδρόφιλα τμήματα
  - αμινοξέα

– Εξαρτάται από τη σειρά των στοιχίσεων.
– "Once a gap, Always a gap"

# T-Coffee

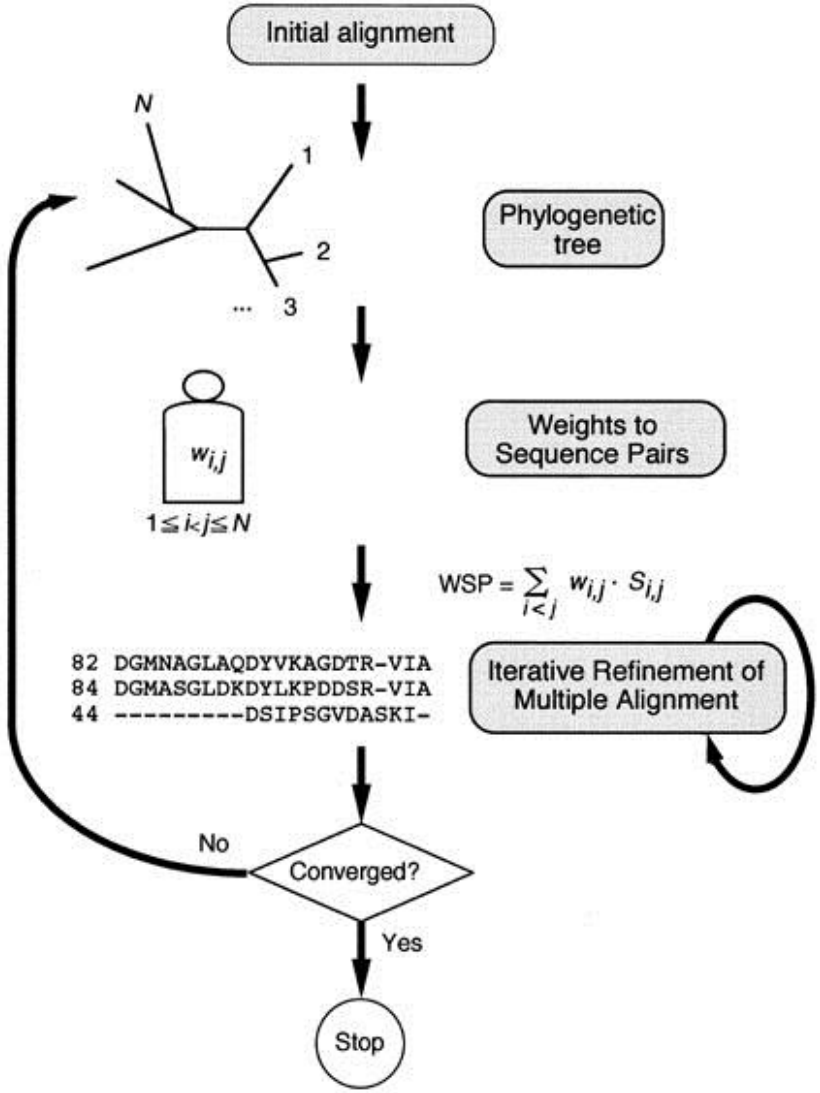
- Δημιουργία βιβλιοθήκης ανά δύο στοιχίσεων
  - Ολική στοίχιση
  - Τοπική στοίχιση
  - Δομική στοίχιση
- Προοδευτική στοίχιση βάσει της πληροφορίας όλων των επιμέρους στοιχίσεων

```
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST CA-T ---
SeqC GARFIELD THE VERY FAST CAT
SeqD ------- THE ---- FA-T CAT
```

- Καλύτερης ποιότητας στοιχίσεις
- Πιο αργός υπολογισμός

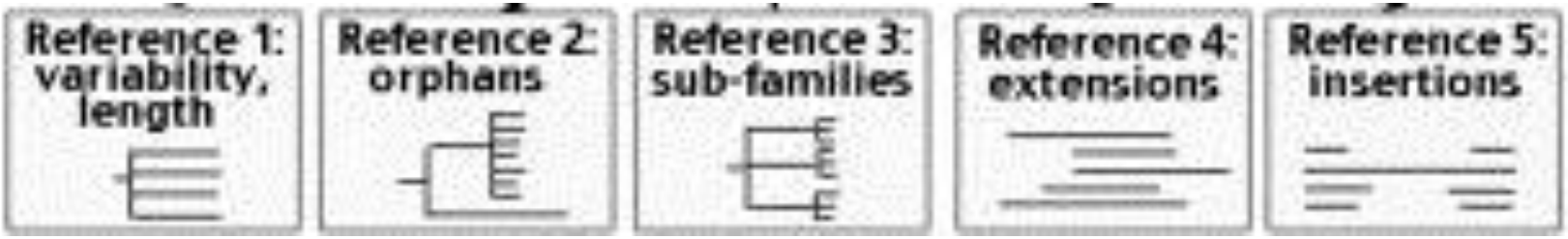
```
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE ---- FAST CAT
```

# Επαναληπτική Στοίχιση (iterative alignment)

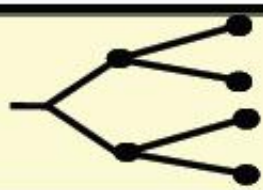
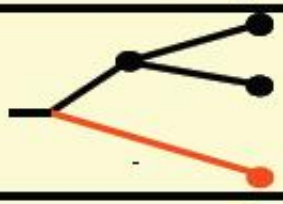
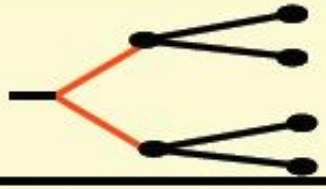
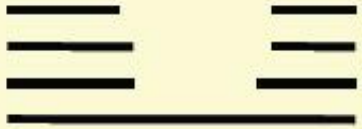
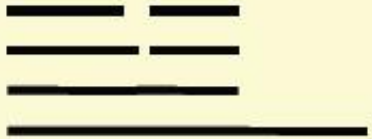


# BAliBASE

- http://www-bio3d-igbmc.u-strasbg.fr/balibase/
- multiple alignment benchmark
  - υψηλής ποιότητας στοιχίσεις βασισμένες στην υπέρθεση τρισδιάστατων δομών, που έχουν ελεγχθεί από εξειδικευμένους επιστήμονες



# Επιλογή Μεθόδου Στοίχισης

| PROBLEM | Program |
|---|---|
|  | ClustalW, T-coffee, MUSCLE, ProbCons |
|  | T-Coffee, MUSCLE, ProbCons |
|  | ProbCons, MUSCLE, MAFFT |
|  | Dialign II, ProbCons, T-Coffee |
|  | Dialign II, ProbCons, MAFFT |

# Προγράμματα Πολλαπλής Στοίχισης

- Clustal Omega
  - http://www.ebi.ac.uk/Tools/msa/clustalo/
- T-Coffee
  - http://tcf_dev.vital-it.ch/apps/tcoffee/index.html
- MAFFT
  - http://mafft.cbrc.jp/alignment/server/
- MUSCLE
  - http://www.ebi.ac.uk/Tools/msa/muscle/

# Βιβλιογραφία


- David Mount, "Bioinformatics: Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press; 2nd edition (March 12, 2013).

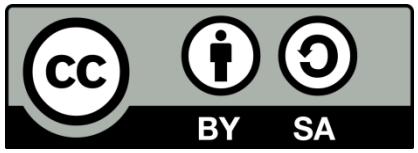
- Jonathan Pevsner, "Bioinformatics and Functional Genomics", Wiley-Blackwell; 2nd edition (May 4, 2009).

- 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.