

# Βιοπληροφορική Ι

Παντελής Μπάγκος  
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας  
Λαμία, 2015

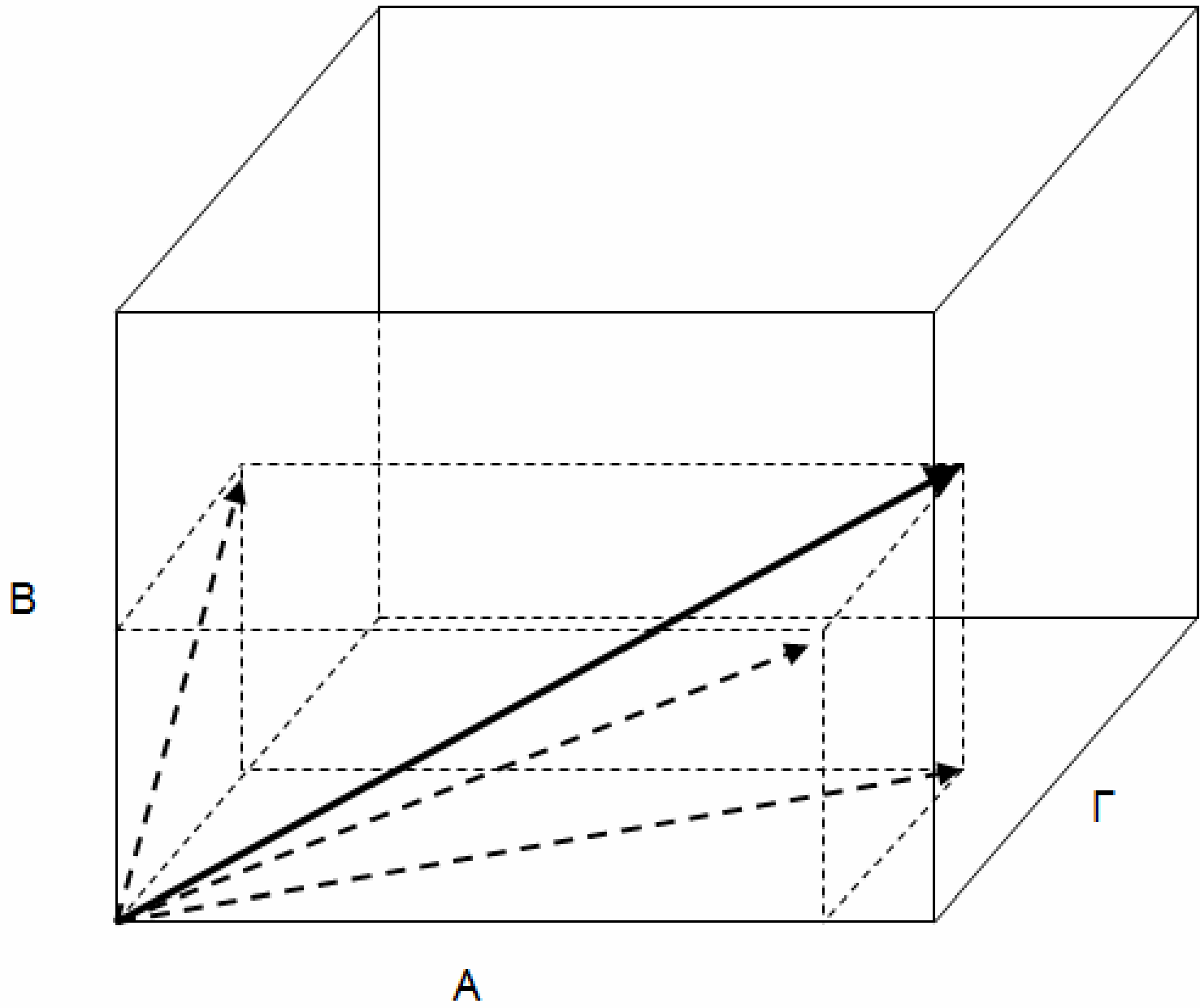
# Πολλαπλή στοίχιση ακολουθιών και φυλογενετικά δέντρα

# Πολλαπλή στοίχιση

- Αναφέρεται στην ταυτόχρονη στοίχιση περισσότερων από 2 ακολουθιών
- Ιδιαίτερα σημαντική καθώς έτσι μπορούμε να εντοπίσουμε οικογένειες σχετιζόμενων ακολουθιών και να μελετήσουμε τα λειτουργικά χαρακτηριστικά τους
- Βασίζεται στις ίδιες αρχές με την κατά ζεύγη στοίχιση, αλλά υπάρχουν πιο πολλές δυσκολίες
- Μπορεί να μας δώσει μια εικόνα της φυλογενετικής προέλευσης των ακολουθιών

# Δυναμικός προγραμματισμός σε N διαστάσεις

- Επέκταση των αλγορίθμων δυναμικού προγραμματισμού (SW και NW) σε περισσότερες διαστάσεις
- Υπολογιστικές δυσκολίες
- Τρόπος υπολογισμού της ομοιότητας (?)



$$\mathbf{X}_1 = x_{11}x_{12}\dots x_{1n}$$

$$\mathbf{X}_2 = x_{21}x_{22}\dots x_{2n}$$

.....

$$\mathbf{X}_r = x_{r1}x_{r2}\dots x_{rn}$$

$$S(m) = G + \sum_i S(m_i)$$

Έτσι τώρα ο αλγόριθμος που υπολογίζει αναδρομικά τα στοιχεία του πίνακα είναι (Sankoff and Kruskal, 1983; Waterman, 1995):

$$a_{i1}, a_{i2}, \dots, a_{in} = \max_{\Delta_1 + \dots + \Delta_n} \left\{ a_{i1-\Delta_1, i2-\Delta_2, \dots, in-\Delta_n} + S(\Delta_1 \cdot x_{i1}^1, \Delta_2 \cdot x_{i2}^2, \dots, \Delta_n \cdot x_{in}^n) \right\}$$

$$\Delta_i \cdot x = \begin{cases} (x), & \text{αν } \Delta_i = 1 \\ (-), & \text{αν } \Delta_i = 0 \end{cases}$$

$$S = \sum_i \log \left( \frac{P_{x1, x2, \dots, xr_i}}{q_{x1} q_{x2} \dots q_{xr_i}} \right) = \sum_i s(x1_i, x2_i, \dots, xr_i)$$

Πολυδιάστατο Score

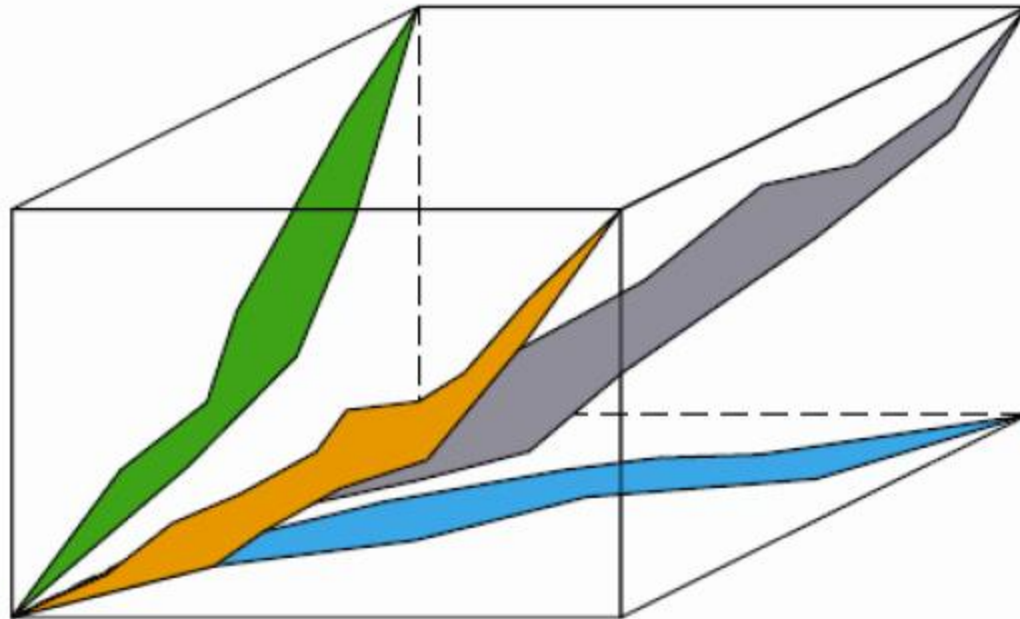
$$SP = \sum_i \left\{ \log \left( \frac{P_{x1, x2_i}}{q_{x1} q_{x2_i}} \right) + \dots + \log \left( \frac{P_{x(r-1)_i, xr_i}}{q_{x(r-1)_i} q_{xr_i}} \right) \right\}$$

Sum of pairs Score

(αυτό χρησιμοποιείται)

# MSA (Carillo and Lipman, 1988, Lipman et al, 1989)

- Περιορισμός του εύρους αναζήτησης του δυναμικού προγραμματισμού
- Χρήση του Sum of Pairs Score



# Προοδευτική πολλαπλή στοίχιση

- Αρχικές κατά ζεύγη στοιχίσεις όλων των ακολουθιών
- Με βάση αυτές, κατάσκημή πίνακα αποστάσεων και ενός δέντρου οδηγού (guide tree)
- Προοδευτική στοίχιση των πιο όμοιων ακολουθιών μεταξύ τους, μέχρι τέλους

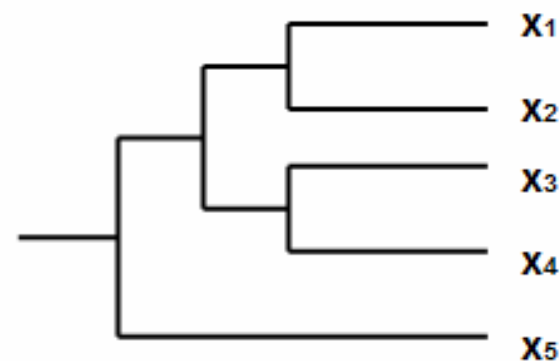


**x1**    **ATCTCGAGA**  
**x2**    **ATCCGAGA**  
**x3**    **ATGTCGACGA**  
**x4**    **ATGTCGACAGA**  
**x5**    **ATTCAACGA**

Όλες οι κατά ζεύγη  
 στοίχισεις (\*)  
 →  
 Δημιουργία του πίνακα  
 αποστάσεων (\*)

x1	-	-	-	-	-
x2	0.11	-	-	-	-
x3	0.20	0.30	-	-	-
x4	0.27	0.36	0.09	-	-
x5	0.30	0.33	0.20	0.27	-
	x1	x2	x3	x4	x5

↓ Κατασκευή του  
 δέντρου-οδηγού (\*)



← Στοίχιση των δύο πιο  
 όμοιων (x1 και x2)

**ATCTCGAGA**  
**ATC-CGAGA**

↓ Στοίχιση του επόμενου  
 ζευγαριού όμοιων (x3  
 και x4)

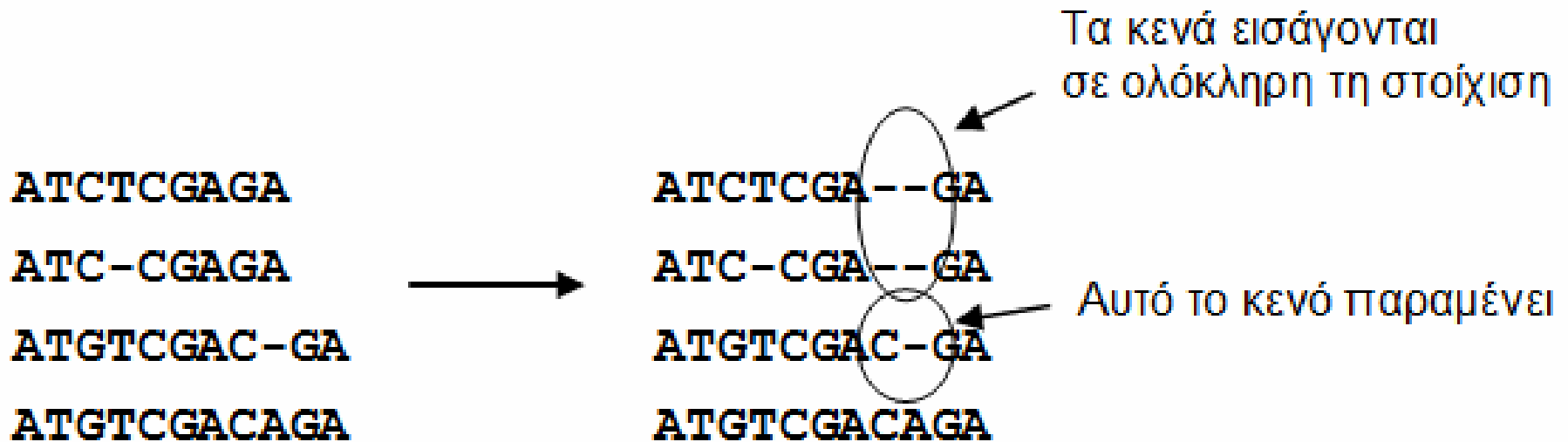
**ATGTCGAC-GA**  
**ATGTCGACAGA**

→ Στοίχιση των  
 ζευγαριών (x1-x2)  
 και (x3-x4) (\*)

**ATCTCGA--GA**  
**ATC-CGA--GA**  
**ATGTCGAC-GA**  
**ATGTCGACAGA**

→ Προσθήκη της  
 x5 στη στοίχιση  
 (\*)

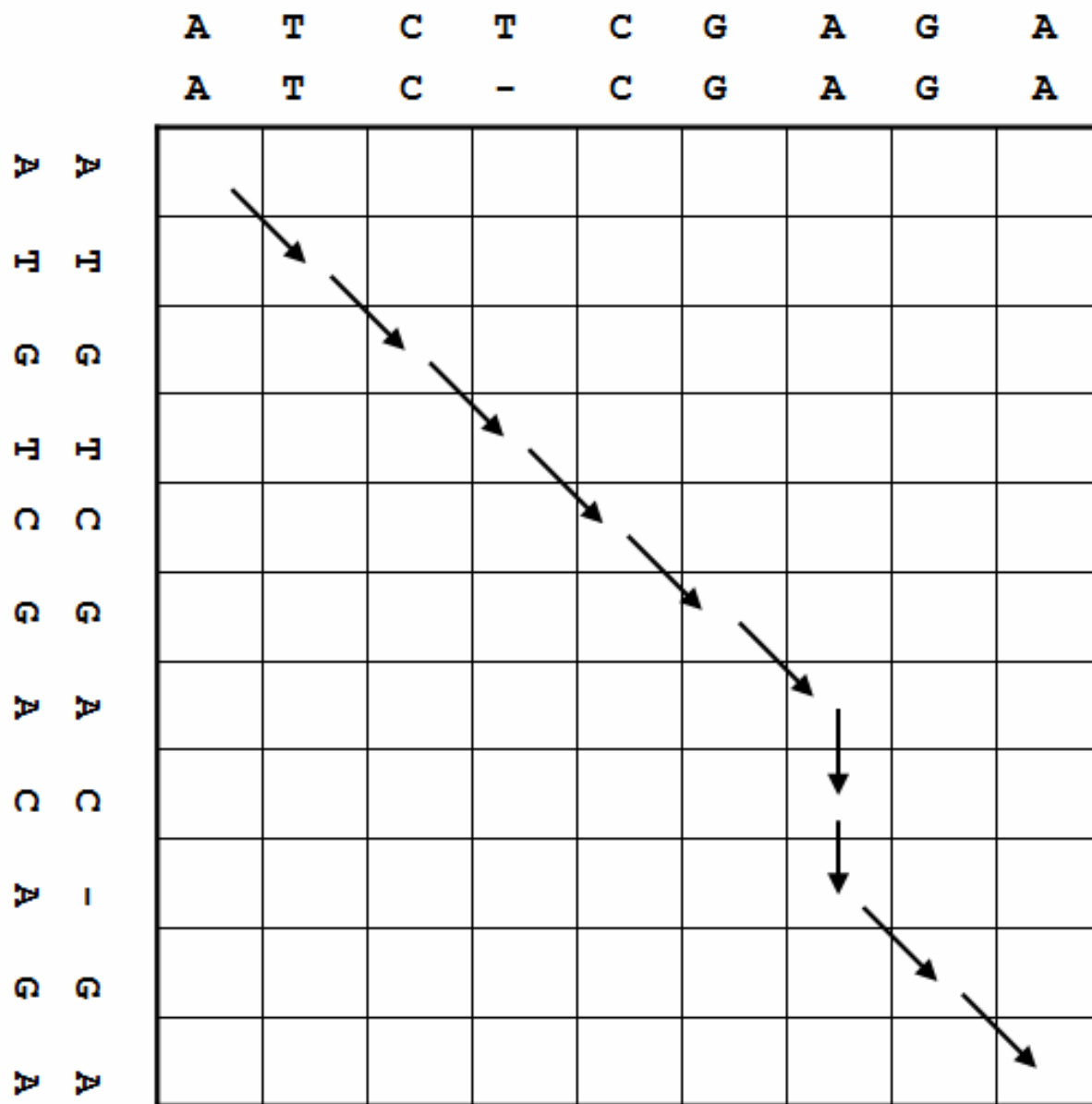
**ATCTCGA--GA**  
**ATC-CGA--GA**  
**ATGTCGAC-GA**  
**ATGTCGACAGA**  
**AT-TCAAC-GA**



Μια καλύτερη λύση, είναι το λεγόμενο *profile alignment*, το οποίο μετράει τη σχετική συνεισφορά όλων των ακολουθιών της κάθε στοίχισης και τελικά πραγματοποιεί την στοίχιση λαμβάνοντας υπόψη όλες τις ακολουθίες. Τα μαθηματικά της μεθόδου είναι πολύπλοκα, αλλά μπορούν να απλοποιηθούν αν θεωρήσουμε, όπως και παραπάνω, το κενό σαν ένα πέμπτο σύμβολο (-), οπότε θα έχουμε και γραμμική ποινή για τα κενά

$$\sum_i SP(m_i) = \sum_i \sum_{j < j'} s(m_i^j, m_i^{j'})$$

$$= \sum_i \sum_{j < j' \leq n} s(m_i^j, m_i^{j'}) + \sum_i \sum_{n < j < j' \leq N} s(m_i^j, m_i^{j'}) + \sum_i \sum_{j \leq n, n < j' \leq N} s(m_i^j, m_i^{j'})$$



# Feng and Doolittle (1987)

- Αρχικές κατά ζεύγη στοιχίσεις με NW
- υπολογισμός της απόστασης δυο ακολουθιών
$$D = -\log S = \log \frac{S_{obs} - S_{rand}}{S_{max} - S_{rand}}$$
- Κατασκευή δέντρου οδηγού με ιεραρχικό Clustering
- Η ακολουθία στοιχίζεται με ΟΛΕΣ τις ακολουθίες μιας στοίχισης (με το γνωστό αλγόριθμο δυναμικού προγραμματισμού).
- Η ακολουθία προστίθεται στην πολλαπλή στοίχιση με βάση τη στοίχιση κατά ζεύγη του προηγούμενου βήματος που είχε το μεγαλύτερο score.

# Clustal (Higgins et al, 1992; Thompson et al, 1994)

- Προοδευτική στοίχιση
- Κατά ζεύγη ομοιότητα με FASTA (default) αλλά και με δυναμικό προγραμματισμό
- Οι αποστάσεις υπολογίζονται απευθείας από την επί τοις εκατό ομοιότητα των ακολουθιών  $x$  ( $D=1-x/100$ )
- Δέντρο οδηγός με Neighbor-Joining method
- Η μέθοδος είναι διαθέσιμη στη διεύθυνση [www.ebi.ac.uk/clustalw](http://www.ebi.ac.uk/clustalw)

# συνέχεια

- Τέλος, χρησιμοποιεί μια σειρά από πολύ προσεκτικά επιλεγμένες ευριστικές τεχνικές οι οποίες μεγιστοποιούν το αποτέλεσμα. Για παράδειγμα, οι πολύ όμοιες ακολουθίες λαμβάνουν μικρό σχετικό βάρος (weight) έτσι ώστε να μην επηρεάζουν τόσο πολύ και να μην κατευθύνουν την πολλαπλή στοίχιση.
- Μια άλλη ιδιαιτερότητα είναι ότι ο πίνακας ομοιότητας δεν είναι σταθερός, αλλά επιλέγεται από τον αλγόριθμο ανάλογα με το ποσοστό ομοιότητας που εντοπίζεται στις υπό μελέτη ακολουθίες.
- Επιπλέον, οι ποινές για τα κενά, δεν είναι σταθερές, αλλά ειδικές ανά θέση (υδρόφοβες περιοχές λαμβάνουν μεγαλύτερη ποινή για τα κενά, με συνέπεια να καθίσταται πιο δύσκολη η εισαγωγή κενών σε αυτές τις περιοχές, αντίθετα, η ποινή μειώνεται αν βρεθούν πάνω από 5 συνεχόμενα υδρόφιλα κατάλοιπα).
- Τέλος, οι ποινές για τα κενά αυξάνονται αν στην ίδια στήλη της στοίχισης δεν υπάρχουν κενά, αλλά αντίθετα υπάρχει κάπου δίπλα μια περιοχή με πολλά κενά. Αυτό έχει σαν συνέπεια τα κενά να «συσσωρεύονται» σε συγκεκριμένες θέσεις σε μια στοίχιση.
- Όλες αυτές οι τεχνικές, έχουν βελτιωθεί με τα χρόνια και έχουν κάνει το CLUSTAL να είναι ένα από τα πιο αξιόπιστα εργαλεία πολλαπλής στοίχισης, παρόλο που κατά βάση στηρίζεται σε μια απλή ευριστική μέθοδο

# Kalign

- Στο Kalign, όλες οι επιλογές της προοδευτικής πολλαπλής στοίχισης είναι βελτιστοποιημένες με σκοπό την ταχύτητα
- Χρήση του προσεγγιστικού αλγόριθμου ταύτισης συμβολοσειρών, των Wu και Manber, ο οποίος είναι γραμμικός ως προς το μήκος της ακολουθίας
- UPGMA
- profile alignment
- <http://msa.sbc.su.se/cgi-bin/msa.cgi>

# Επαναληπτικές μέθοδοι και μέθοδοι που βασίζονται στη συνέπεια

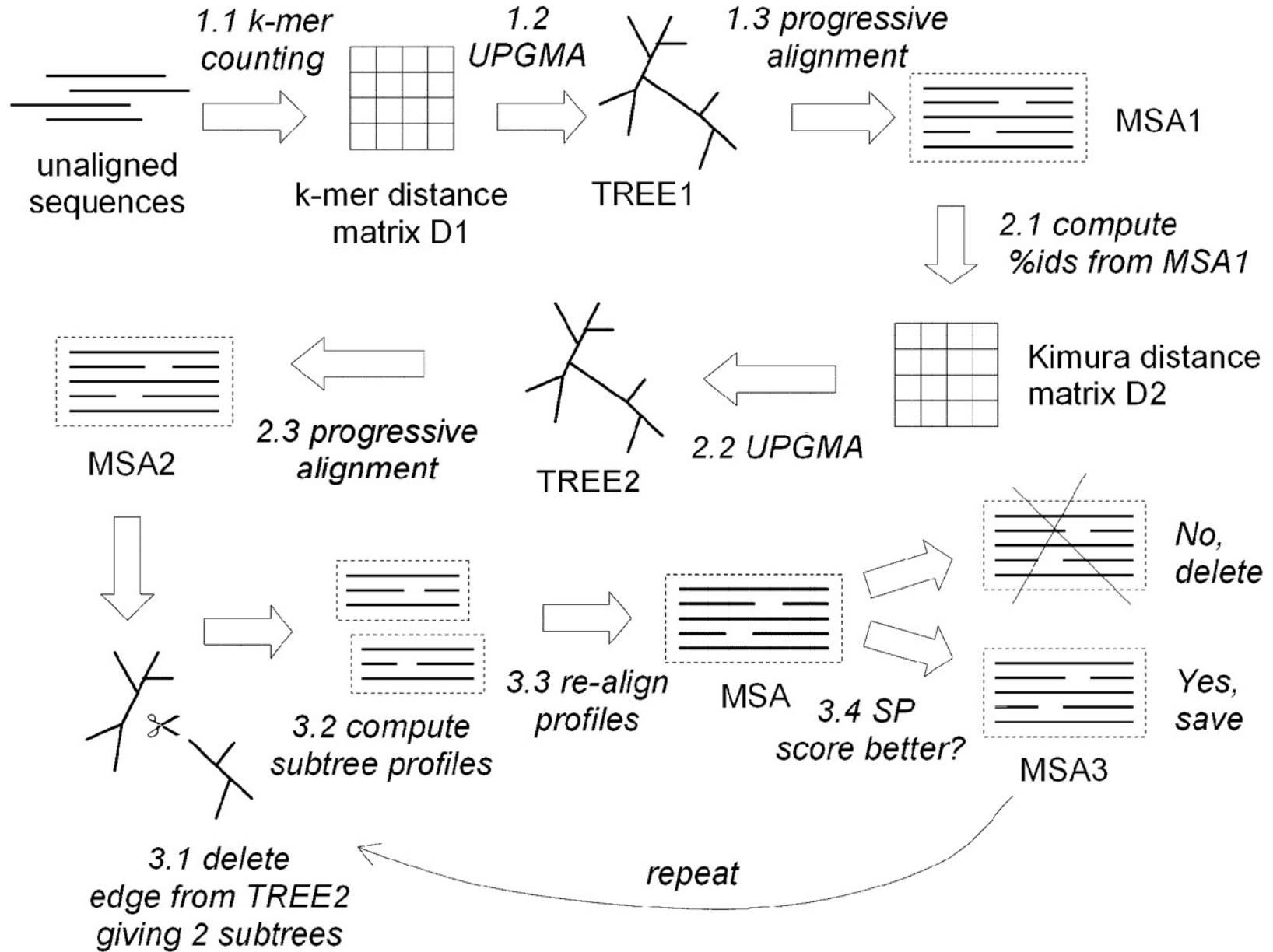
- Η βασική ιδέα των επαναληπτικών μεθόδων, είναι να χρησιμοποιηθεί κάποιου είδους προοδευτική πολλαπλή στοίχιση, αλλά αυτή η διαδικασία να γίνει επαναληπτικά έτσι ώστε λάθη που είναι πιθανό να εισχωρήσουν σε αρχικά στάδια της στοίχισης, να μπορούν να αναιρεθούν σε κάποιο μετέπειτα βήμα. Η επαναληπτική διαδικασία, είναι σε γενικές γραμμές μια εύκολα υλοποιήσιμη ιδέα, και εμπειρικές αναλύσεις έχουν δείξει ότι μπορεί να χρησιμοποιηθεί ακόμα και σε ήδη υπάρχοντες αλγόριθμους, αυξάνοντας σημαντικά την απόδοσή τους. Για παράδειγμα, η ακρίβεια του CLUSTALW αυξάνει κατά 6% με αυτή τη διαδικασία



# Παραδείγματα

- MULTALIN (<http://prodes.toulouse.inra.fr/multalin/multalin.html>)
- MUSCLE (<http://www.drive5.com/muscle>)
- PRPP/PRRN  
([http://www.genome.ist.i.kyoto-u.ac.jp/~aln\\_user/prrn/index.html](http://www.genome.ist.i.kyoto-u.ac.jp/~aln_user/prrn/index.html))
- PRALINE (<http://ibivu.cs.vu.nl/programs/pralinewww/>)
- DIALIGN (<http://bibiserv.techfak.uni-bielefeld.de/dialign/>)
- COBALT (<ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/cobalt>)
- T-COFFEE  
(<http://www.ch.embnet.org/software/TCoffee.html>)

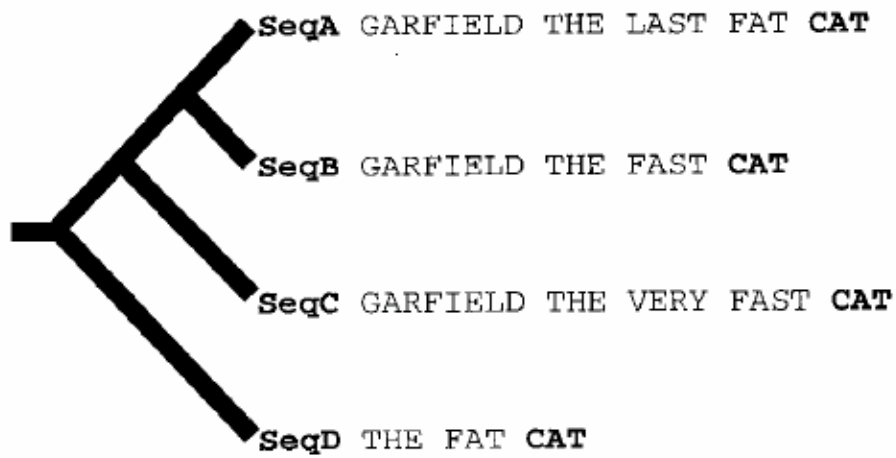
# MUSCLE



# T-Coffee

- Όμοια με το Dialign, το T-Coffee πραγματοποιεί πρώτα όλες τις κατά ζεύγη στοιχίσεις. Παρ' όλα αυτά, το T-Coffee κάνει αυτό το βήμα δυο φορές: μια με χρήση του ClustalW (global) και μια με το Lalign (local-Fasta package).
- Τα αποτελέσματα συνδυάζονται σε μια αρχική βιβλιοθήκη
- Σε ένα βήμα επέκτασης της βιβλιοθήκης, καθορίζεται πως τα ζεύγη καταλοίπων στοιχίζονται σε σχέση με άλλα κατάλοιπα. Τέτοιες τριπλέτες χρησιμοποιούνται για να βρεθεί πόσο καλά οι ακολουθίες στοιχίζονται δοδομένων των υπολοίπων (σε αντίθεση με τον έλεγχο δυο ακολουθιών απομονωμένα).
- Η τελική στοίχιση κατασκευάζεται έπειτα, προοδευτικά με χρήση της πληροφορίας στη βιβλιοθήκη

a) Regular Progressive Alignment Strategy



<b>SeqA</b>	GARFIELD	THE	LAST	FA-T	<b>CAT</b>
<b>SeqB</b>	GARFIELD	THE	FAST	<b>CA-T</b>	---
<b>SeqC</b>	GARFIELD	THE	VERY	FAST	<b>CAT</b>
<b>SeqD</b>	-----	THE	----	FA-T	<b>CAT</b>

b)Primary Library

**SeqA** GARFIELD THE LAST FAT CAT    **Prim. Weight = 88**  
**SeqB** GARFIELD THE FAST CAT ---

**SeqA** GARFIELD THE LAST FA-T CAT    **Prim. Weight = 77**  
**SeqC** GARFIELD THE VERY FAST CAT

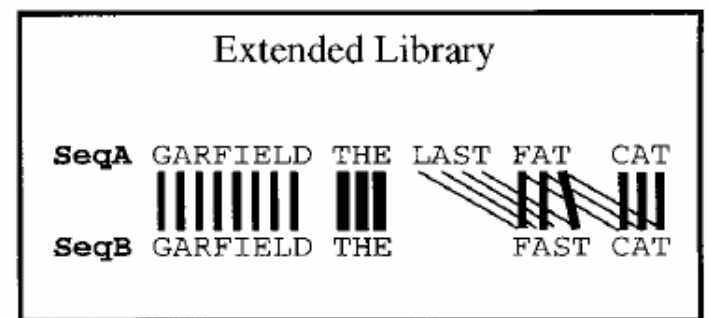
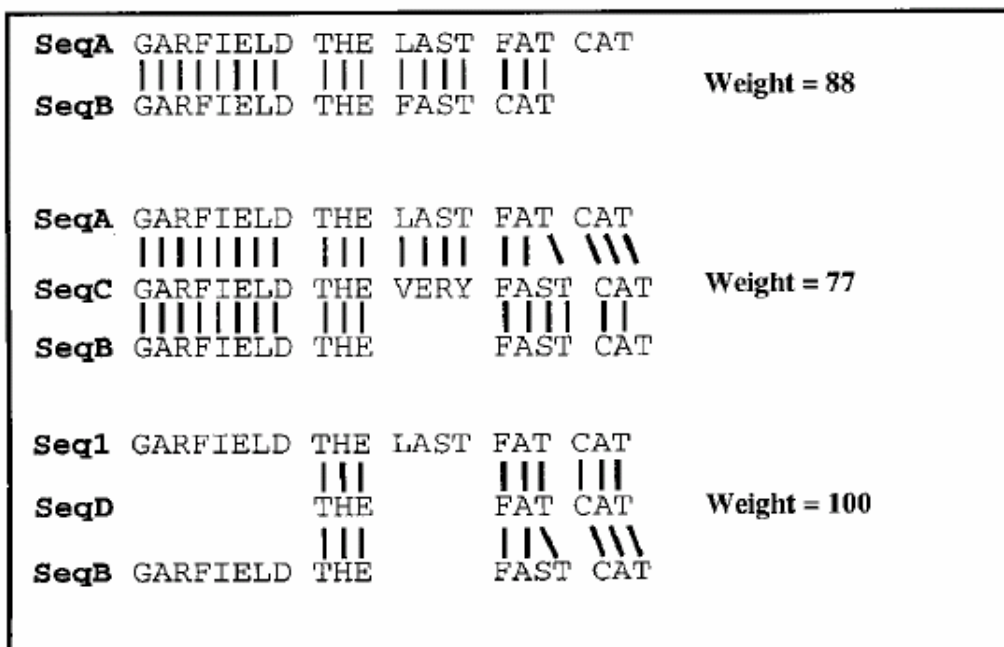
**SeqA** GARFIELD THE LAST FAT CAT    **Prim. Weight =100**  
**SeqD** ----- THE ---- FAT CAT

**SeqB** GARFIELD THE ---- FAST CAT    **Prim Weight = 100**  
**SeqC** GARFIELD THE VERY FAST CAT

**SeqB** GARFIELD THE FAST CAT    **Prim. Weight = 100**  
**SeqD** ----- THE FA-T CAT

**SeqC** GARFIELD THE VERY FAST CAT    **Prim. Weight = 100**  
**SeqD** ----- THE ---- FA-T CAT

c) Extended Library for seq1 and seq2



Dynamic Programming



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

# Dialign

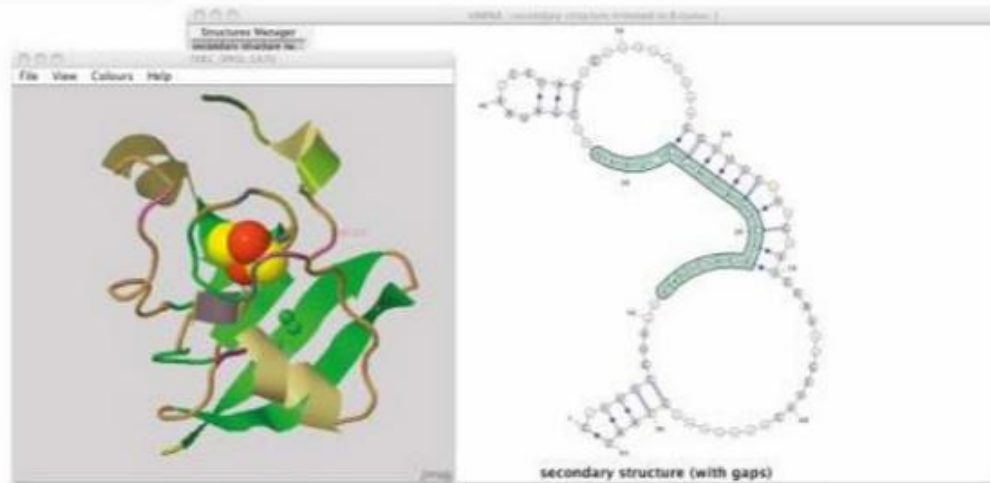
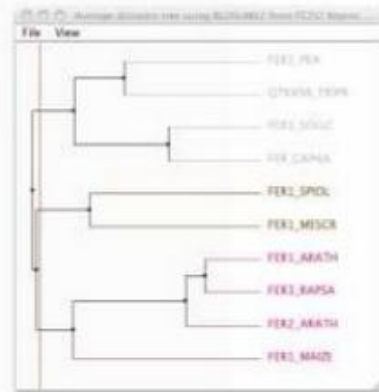
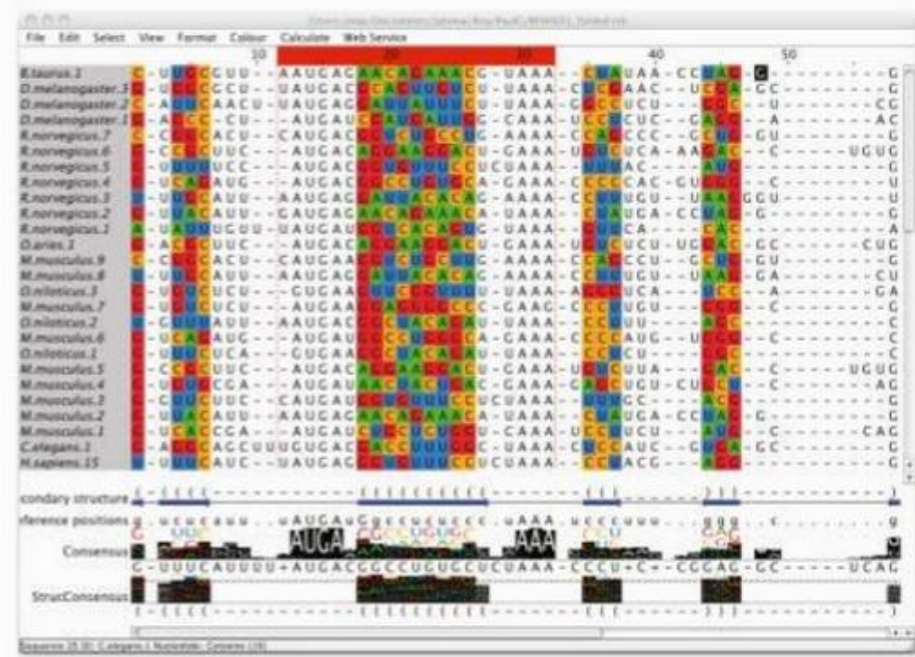
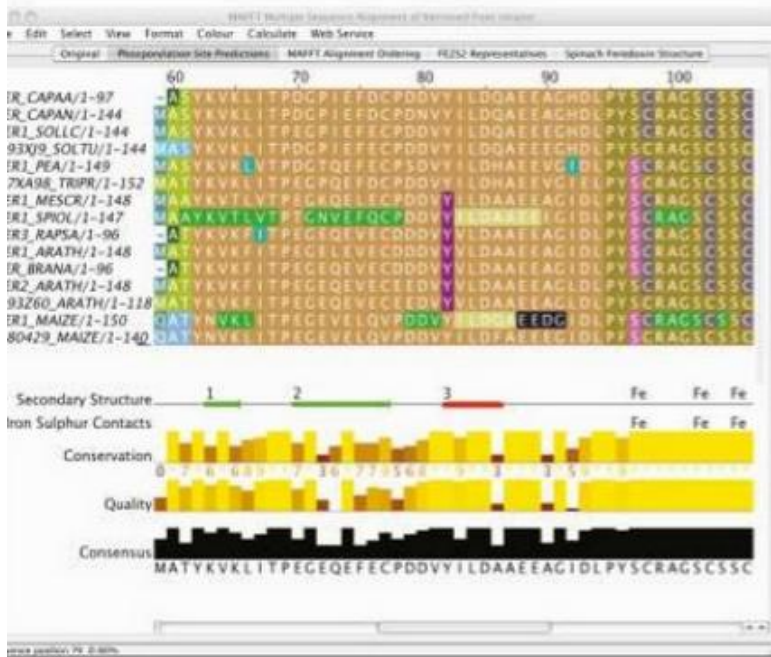
- Το Dialign πραγματοποιεί τοπική στοίχιση και στοιχίζει ολόκληρα τμήματα παρά κατάλοιπα.
- Αρχικά, όλες οι ανά δυο στοιχίσεις πραγματοποιούνται και συλλέγονται οι στοιχισμένες περιοχές στις οποίες δεν υπάρχουν κενά.
- Το όνομα ‘Dialign’ βγαίνει από αυτές τις διαγώνιες περιοχές (diagonal alignments in a dot plot)
- Ένα συνεπές σύνολο από διαγώνιες καθορίζεται έτσι και διαδοχικά προστίθεται στη στοίχιση
- <http://bibiserv.techfak.uni-bielefeld.de/dialign/>

# Οπτικοποίηση και επεξεργασία πολλαπλών στοιχίσεων

- Jalview (<http://www.jalview.org/>)
- Strap (<http://www.bioinformatics.org/strap/>)
- Seqpup  
(<http://iubio.bio.indiana.edu/soft/molbio/seqpup/java/seqpup-doc.html>)
- Seaview (<http://pbil.univ-lyon1.fr/software/seaview.html>)
- Cinema  
(<http://aig.cs.man.ac.uk/research/utopia/cinema/cinema.php>)
- Boxshade  
([http://www.ch.embnet.org/software/BOX\\_form.html](http://www.ch.embnet.org/software/BOX_form.html))
- Bioedit (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>)



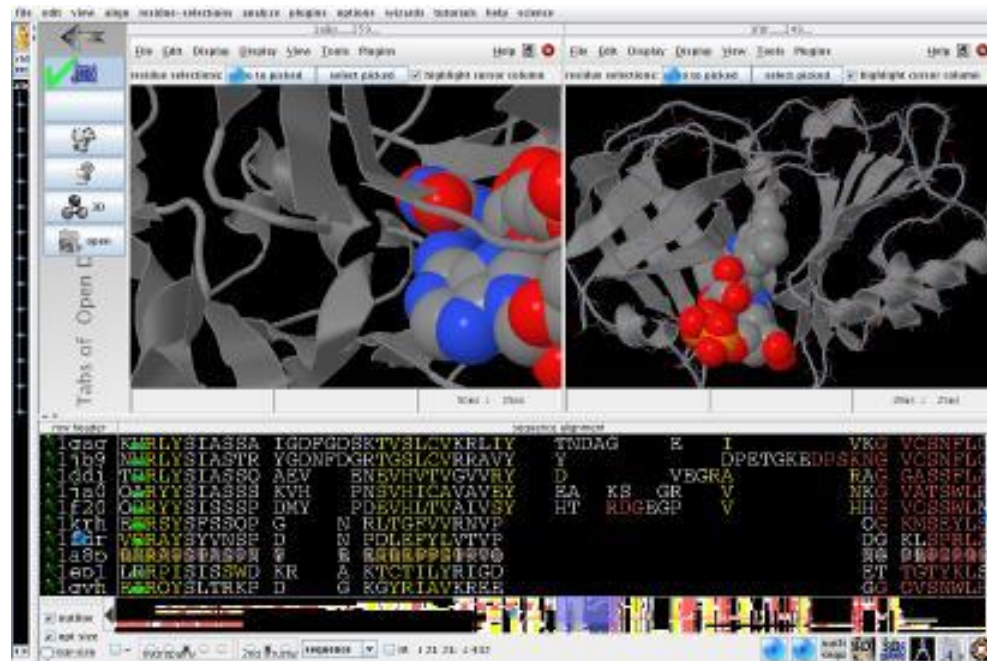
# Jalview



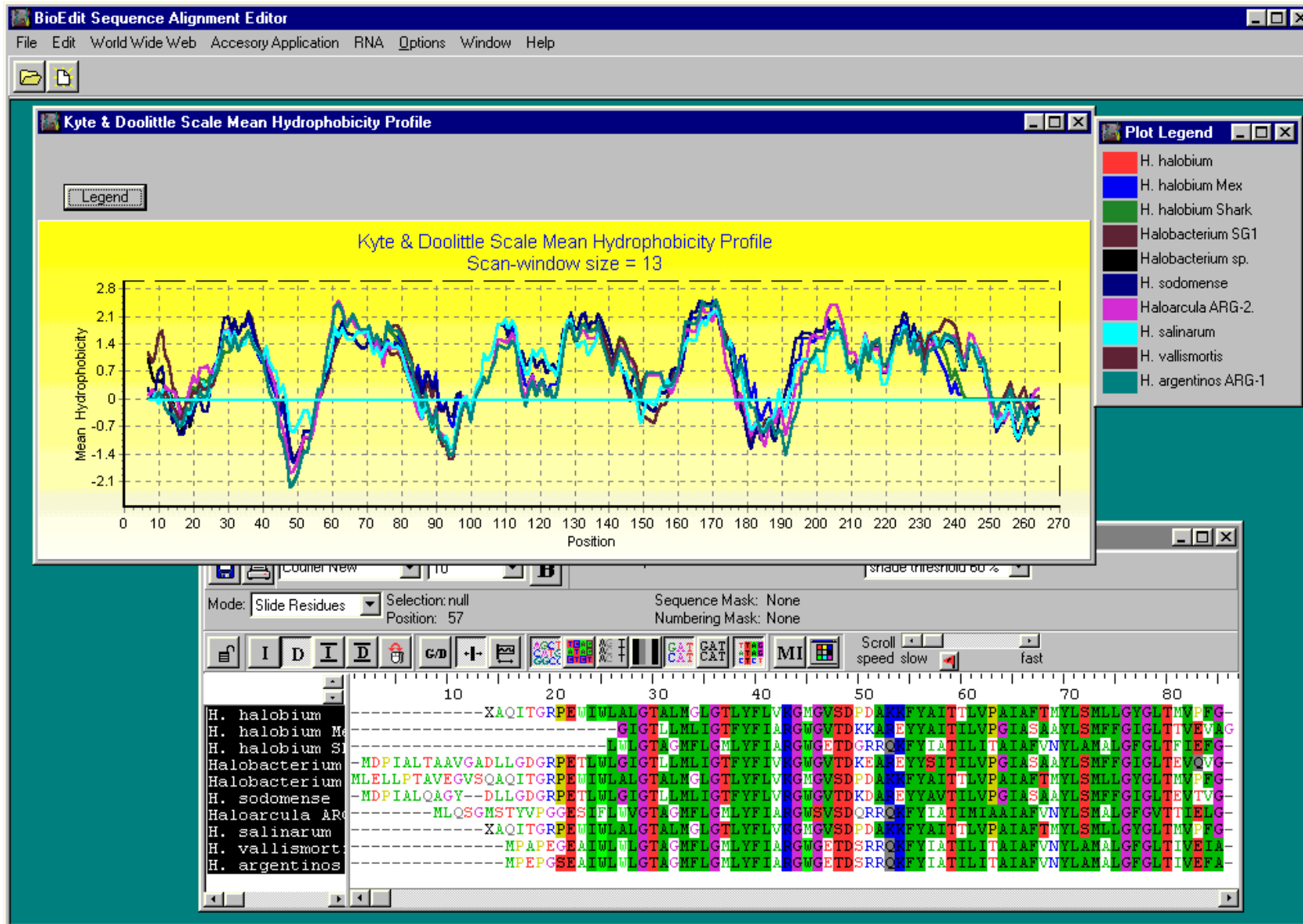
	position 12 ↓	hex HU <u>00000000</u>	sheet
a <sub>1</sub> S. cerevisiae	G <b>Y</b> D <b>R</b> H <b>I</b> T <b>I</b> F <b>S</b> P.	<b>E</b> G <b>R</b> L <b>Y</b> Q <b>V</b> E <b>Y</b> A <b>F</b> K <b>A</b> T <b>N</b> Q <b>T</b> M <b>I</b> N <b>S</b> L <b>A</b> V <b>R</b> G	
a <sub>2</sub> S. cerevisiae	R <b>Y</b> S <b>F</b> S <b>L</b> T <b>T</b> F <b>S</b> P.	<b>S</b> G <b>K</b> L <b>G</b> Q <b>I</b> D <b>Y</b> A <b>L</b> T <b>A</b> V <b>K</b> Q <b>G</b> .V <b>T</b> S <b>L</b> G <b>I</b> K <b>A</b>	
a <sub>3</sub> S. cerevisiae	R <b>Y</b> D <b>S</b> R <b>T</b> T <b>I</b> F <b>S</b> P.	<b>E</b> G <b>R</b> L <b>Y</b> Q <b>V</b> E <b>Y</b> A <b>L</b> E <b>S</b> I <b>S</b> H <b>A</b> .G <b>T</b> A <b>I</b> G <b>I</b> M <b>A</b>	
a <sub>4</sub> S. cerevisiae	G <b>Y</b> D <b>R</b> A <b>L</b> S <b>I</b> F <b>S</b> P.	<b>D</b> G <b>H</b> I <b>F</b> Q <b>V</b> E <b>Y</b> A <b>L</b> E <b>A</b> V <b>K</b> R <b>G</b> .T <b>C</b> A <b>V</b> G <b>V</b> K <b>G</b>	
a <sub>5</sub> S. cerevisiae	E <b>Y</b> D <b>R</b> G <b>V</b> S <b>T</b> F <b>S</b> P.	<b>E</b> G <b>R</b> L <b>F</b> Q <b>V</b> E <b>Y</b> S <b>L</b> E <b>A</b> I <b>K</b> L <b>G</b> .S <b>T</b> A <b>I</b> G <b>I</b> A <b>T</b>	
a <sub>6</sub> S. cerevisiae	N <b>Y</b> D <b>G</b> D <b>T</b> V <b>T</b> F <b>S</b> P.	<b>T</b> G <b>R</b> L <b>F</b> Q <b>V</b> E <b>Y</b> A <b>L</b> E <b>A</b> I <b>K</b> Q <b>G</b> .S <b>V</b> T <b>V</b> G <b>L</b> R <b>S</b>	
a <sub>7</sub> S. cerevisiae	G <b>Y</b> D <b>L</b> S <b>N</b> S <b>V</b> F <b>S</b> P.	<b>D</b> G <b>R</b> N <b>F</b> Q <b>V</b> E <b>Y</b> A <b>V</b> K <b>A</b> V <b>E</b> N <b>G</b> .T <b>T</b> S <b>I</b> G <b>I</b> K <b>C</b>	
a <sub>7</sub> E. cuniculi	G <b>R</b> E <b>K</b> P <b>V</b> T <b>L</b> F <b>S</b> S.	<b>E</b> G <b>K</b> L <b>G</b> Q <b>C</b> D <b>N</b> A <b>L</b> R <b>A</b> A <b>V</b> N <b>G</b> .S <b>L</b> S <b>I</b> G <b>A</b> S <b>A</b>	
a <sub>7</sub> E. cuniculi	G <b>F</b> E <b>Q</b> L <b>A</b> V <b>F</b> S <b>P</b> .	<b>D</b> G <b>R</b> L <b>I</b> Q <b>V</b> E <b>Y</b> A <b>Q</b> Q <b>A</b> S <b>E</b> Q <b>G</b> S <b>L</b> V <b>V</b> F <b>G</b> W <b>D</b> S	
a <sub>7</sub> E. cuniculi	D <b>N</b> K <b>P</b> K <b>S</b> N <b>T</b> F <b>T</b> D.	<b>E</b> G <b>R</b> L <b>P</b> Q <b>V</b> E <b>F</b> A <b>I</b> K <b>N</b> V <b>S</b> R <b>A</b> .G <b>T</b> I <b>I</b> G <b>Y</b> V <b>C</b>	
a <sub>7</sub> E. cuniculi	T <b>Q</b> E <b>V</b> S <b>N</b> I <b>F</b> N <b>S</b> .	<b>D</b> G <b>K</b> L <b>L</b> Q <b>I</b> E <b>Y</b> G <b>L</b> E <b>A</b> V <b>N</b> N <b>G</b> .L <b>P</b> V <b>V</b> T <b>A</b> K <b>S</b>	
a <sub>7</sub> E. cuniculi	D <b>F</b> K <b>Q</b> S <b>V</b> N <b>T</b> Y <b>S</b> S.	<b>E</b> G <b>R</b> I <b>H</b> Q <b>I</b> E <b>Y</b> A <b>M</b> K <b>A</b> M <b>N</b> L <b>G</b> .T <b>T</b> T <b>I</b> G <b>V</b> R <b>T</b>	
a <sub>7</sub> E. cuniculi	A <b>R</b> Y <b>N</b> M <b>F</b> K <b>I</b> F <b>N</b> P.	<b>E</b> G <b>C</b> V <b>K</b> Q <b>L</b> D <b>F</b> I <b>R</b> Q <b>T</b> T <b>E</b> L <b>G</b> .G <b>T</b> A <b>V</b> G <b>L</b> K <b>N</b>	
a <sub>7</sub> E. cuniculi	S <b>N</b> L <b>D</b> F <b>C</b> T <b>I</b> Y <b>T</b> T.	<b>T</b> G <b>Q</b> I <b>D</b> Q <b>L</b> S <b>Y</b> A <b>Q</b> K <b>A</b> A <b>D</b> S <b>G</b> .D <b>T</b> C <b>I</b> G <b>M</b> K <b>S</b>	
a <sub>3</sub> A. thaliana	R <b>Y</b> D <b>S</b> R <b>T</b> T <b>I</b> F <b>S</b> P.	<b>E</b> G <b>R</b> L <b>Y</b> Q <b>V</b> E <b>Y</b> A <b>M</b> E <b>A</b> I <b>G</b> N <b>A</b> .G <b>S</b> A <b>I</b> G <b>I</b> L <b>S</b>	
a <sub>3</sub> C. elegans	R <b>Y</b> D <b>S</b> R <b>T</b> T <b>I</b> F <b>S</b> P <b>L</b> R	<b>E</b> G <b>R</b> L <b>Y</b> Q <b>V</b> E <b>Y</b> A <b>M</b> E <b>A</b> I <b>S</b> H <b>A</b> .G <b>T</b> C <b>L</b> G <b>I</b> L <b>S</b>	
a <sub>3</sub> C. albicans	R <b>Y</b> D <b>S</b> R <b>T</b> T <b>I</b> F <b>S</b> P.	<b>E</b> G <b>R</b> L <b>Y</b> Q <b>V</b> E <b>Y</b> A <b>Q</b> E <b>A</b> I <b>S</b> N <b>A</b> .G <b>T</b> A <b>I</b> G <b>I</b> L <b>S</b>	
a <sub>3</sub> D. melanogaster	R <b>Y</b> D <b>S</b> R <b>T</b> T <b>I</b> F <b>S</b> P.	<b>E</b> G <b>R</b> L <b>Y</b> Q <b>V</b> E <b>Y</b> A <b>M</b> E <b>A</b> I <b>S</b> H <b>A</b> .G <b>T</b> C <b>L</b> G <b>I</b> L <b>A</b>	
a <sub>3</sub> H. sapiens	R <b>Y</b> D <b>S</b> R <b>T</b> T <b>I</b> F <b>S</b> P.	<b>E</b> G <b>R</b> L <b>Y</b> Q <b>V</b> E <b>Y</b> A <b>M</b> E <b>A</b> I <b>G</b> H <b>A</b> .G <b>T</b> C <b>L</b> G <b>I</b> L <b>A</b>	
a <sub>3</sub> O. sativa	R <b>Y</b> D <b>S</b> R <b>T</b> T <b>I</b> F <b>S</b> P.	<b>E</b> G <b>R</b> L <b>Y</b> Q <b>V</b> E <b>Y</b> A <b>M</b> E <b>A</b> I <b>G</b> N <b>A</b> .G <b>S</b> A <b>L</b> G <b>V</b> L <b>A</b>	
a <sub>3</sub> P. falciparum	R <b>Y</b> D <b>S</b> R <b>T</b> T <b>I</b> F <b>S</b> P.	<b>E</b> G <b>R</b> L <b>Y</b> Q <b>V</b> E <b>Y</b> A <b>L</b> E <b>A</b> I <b>N</b> N <b>A</b> .S <b>I</b> T <b>I</b> G <b>L</b> I <b>T</b>	
a <sub>3</sub> S. pombe	S <b>Y</b> D <b>S</b> R <b>T</b> T <b>I</b> F <b>S</b> P.	<b>E</b> G <b>R</b> L <b>Y</b> Q <b>V</b> E <b>Y</b> A <b>L</b> E <b>A</b> I <b>N</b> H <b>A</b> .G <b>V</b> A <b>L</b> G <b>I</b> V <b>A</b>	

↑↑↑  
(F, Y or W)<sub>15</sub>S<sub>16</sub>P<sub>17</sub>

# Strap



# BioEdit



```

9rnt0      ACDYTGNCYSSDVAQAAGYLHEKGVVNSYHKYRYEGDFVSYEYEWILSGDYG
rnt1 aspor ACDYTGNCYSSDVAQAAGYLHEKGVVNSYHKYRYEGDFVSYEYEWILSGDYG
rnpc pench ACAATGVCYSSAIAQAAGYLYSRIIVSI--YHEYRYEGDFVGYEYEWILSGAYG
rnc2 aspcl -CDYTGHCYAAVAVQAQAGYLEKGVVRSRYHQYRYEGDFVGYEYEWILSGSYG
rnpb penbr  ACAATGVCYSSAIAQAAGYLYSRIIVSI--YHEYRYEGDFVGYEYEWILSGKYG
rnms aspsa SCEYTGSTCYWSSDVAKAKKGYLYEKGITIM--YHGYRYEGDFVGYEYEWIMDYDYG
rnn1 neucr  ACMYICGVYSSAIAALNKGYYEKAASSSYHRYRYEGDFVTKWYEFILSGRYG
rnf1 fusla ----TCGSTPYAAQVAAANAACYQSDIAGSTTYHTYRYEGDFVGYQEYIRGG-YG
rnf1_fusmo ----TCGSTNYAAQVAAANAACYQSDIAGSTTYHTYRYEGDFVGYQEYIRGG-YG
rnt1 triha  ----TCGKVFYAAVAAANAACYVRAGIAGSTYHVVRYEGDFRKLKFEYILSGKYG
rnf2 fusla ----TSSKPYAAQVAAANAACYQSDIAGSTTYHTYRYEGDFVGYQEYIRSG-YG
rnul1 ustsp -----CGSTYYSIQVRA-----INNAKGGQYRTGYHTYRYEGDFDYGKREYLKSSSYG
rnu2 ustsp -----CGNVYNDIATAIQGALDVANCRPNI--YHQYRYEASELILGWSEFLVYNGPYS
aga2_pedpe -----YVIQILAHRYPVHPYFIMRGGSHFANRTREFPYVSLPLEYTIQS-----G
Consensus/80% . . . . .sts..asts.lpst..hthphh.spsshs..as+.apsa-tacasltsab-aslbpt.sast
PHD           L.eEe.LL.e..hhhHHHHHHhheee.LLLLLLLLLLLL.e.L...ee.LLL.eEEEE.LLLeEEE.

```

```

9rnt0      GSMGADRVVFEHMLAGVIHAGASGDFVET
rnt1 aspor GSMGADRVVFEHMLAGVIHAGASGDFVET
rnpc pench HSMGADRVVFEHMLAGVIHAGASGDFVET
rnc2 aspcl  LGMGADRVVFEHMLGLIHHAGASGDFVET
rnpb penbr  GSMGADRVVFEHMLAGVIHAGASGDFVET
rnms aspsa GSMGADRVVFEHMLAGVIHAGASGDFVET
rnn1 neucr  GSMGADRVVFEHMLMLIHHAGASGDFVET
rnf1 fusla  GSMGADRVVFEHMLAGVIHAGASGDFVET
rnf1_fusmo GSMGADRVVFEHMLAGVIHAGASGDFVET
rnt1 triha  GSMGADRVVFEHMLAGVIHAGASGDFVET
rnf2 fusla  GSMGADRVVFEHMLAGVIHAGASGDFVET
rnul1 ustsp GSMGADRVVFEHMLAGVIHAGASGDFVET
rnu2 ustsp  RSMGADRVVFEHMLAGVIHAGASGDFVET
aga2_pedpe MYRQPAYVIKDAHMLLPLEYGFSVET
Consensus/80% ststt-+llbsspspbsthls+sttttssal.s.
PHD           LLLLLL.EEEEEeLLLLeEEEEEEeLLLLL.eEEEE

```

## Chroma

eubact.mase

File ▾ Props ▾ Sites ▾ Species ▾ Footers ▾ Search:  Goto:  Edit ▾ Help

sel=6 411 Seq:6 Pos:438|399 [Flavobact] 469

ThermusTh	AATGGGCGCAAGCCTGACGGAGCGACGCCGCTTGGAGGAA GAA-GCCCTTCGGGCTGTA
Chlamydia	AATGGACGAAAAGTCTGACGAAGCGACGCCGCTGTGTGATGAA-GCCTCTAGGCTTCTA
Flavobact	AATGGAGGGAACTCTGAACCAGCCATGCGCGTGCAGGAAACACAGCCCTCTGGCTCCTA
Bacteroid	AATGGGCGCTAGCCTGAACCAGCCAAGTAGCGTGAAGGATGAAAGGCTCTATGGGTCTA
Anacystis	AATGGGCGCAAGC--GACGGAGCAACGCCGCTGGGGGAGGAA-GGTTTTTGGACTGTA
Streptomy	AATGGGCGAAAAGCCTGATGCAGCGACGCCGCTGAGGATGAC-GCCCTTCGGCTTCTA
Mycobacte	AATGGGCGCAAGCCTGATGCAGCGACGCCGCTGGGGGATGAC-GCCCTTCGGCTTCTA
Mycoplhyo	AATAAGCGAAAAGCTTGATGGAGCGACACAGCGTGCAGGATGAA-GTCTTTCGGGATGTA
Mycoplcap	AATGGACGAAAAGTCTGATGAAGCAATGCCGCGTGAGTATGAC-GCCCTTCGGCTTCTA
Heliobact	AATGGGCGAAAAGCCTGACGGAGCAATGCCGCGTGGGGGATGAA-GCTCTTCGGATTCTA
Bacillus	AATGGACGAAAAGTCTGACGGAGCAACGCCGCGTGAGTATGAA-GCTTTTTCGGATCCTA
Proteus	AATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAA GAA-GCCCTTAGGCTTCTA
Escherich	AATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAA GAA-GCCCTTCGGCTTCTA
Pseudoaer	AATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGTGTGAA GAA-GCTCTTCGGATTCTA
Pseudotes	AATGGGCGAAAAGCCTGATCCAGCAATGCCGCGTGCAGGATGAA-GCCCTTCGGCTTCTA
Neisseria	AATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGTCTGAA GAA-GCCCTTCGGCTTCTA
all seqs	-----XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX--
Comments	-----[---helix 45---]-----

[><-+\_] [◀▶] [▶▶]

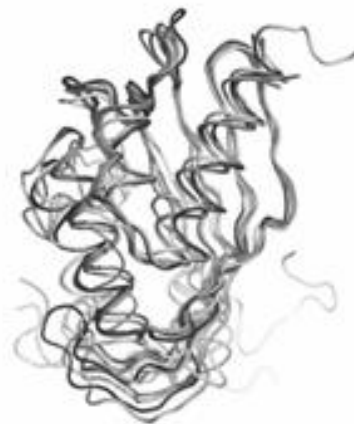
Seaview

# Αξιολόγηση των πολλαπλών στοιχίσεων

- Βασίζεται κυρίως σε δομικές στοιχίσεις ελεγμένες από ειδικούς. Παρ' όλα αυτά δεν υπάρχει ακόμα «απολύτως αντικειμενικός» και αποδεκτός τρόπος για αξιολόγηση
- BaliBase (<http://www-igbmc.u-strasbg.fr/BioInfo/BAlIbASE/index.html> )
- OxBench (<http://www.compbio.dundee.ac.uk/>)
- SABmark(<http://bioinformatics.vub.ac.be/databases/databases.html> )
- PREFAB (<http://drive5.com/muscle/prefab.htm>. )

Δομές με όλες τις πρωτεΐνες της οικογένειας

Δομική στοίχιση



Εξαγωγή των ακολουθιών

Ακολουθία 1	...	W	G	K	V	G	A	-	-	H	A	G	E	...
Ακολουθία 2	...	W	G	K	V	-	-	-	-	N	V	D	E	...
Ακολουθία 3	...	W	G	K	V	E	A	-	-	D	V	A	G	...
Ακολουθία 4	...	W	K	D	F	N	A	-	-	N	I	P	K	...
Ακολουθία 5	...	W	E	E	I	A	G	A	D	N	G	A	G	...

Πολλαπλή στοίχιση (στοίχιση αναφοράς)

# Συμπεράσματα

- Παρόλο που οι επιμέρους μελέτες διαφέρουν πολλές φορές ως προς τη μεθοδολογία, μπορούμε να εξάγουμε κάποια γενικά συμπεράσματα. Για παράδειγμα, τα περισσότερα από τα σύγχρονα εργαλεία που αναφέραμε παραπάνω, σε ένα ευρύ φάσμα συνθηκών αποδίδουν πολύ καλά, πετυχαίνοντας πάνω από 50% επιτυχία στην ανακατασκευή των στοιχίσεων αναφοράς, ακόμα και σε οικογένειες με μέσο ποσοστό ομοιότητας γύρω στο 20%. Το T-Coffee, το ProbCons και το ProbAlign είναι σε γενικές γραμμές οι πιο αποδοτικοί αλγόριθμοι, αλλά είναι και πιο χρονοβόροι και με μεγάλες απαιτήσεις σε μνήμη (ιδιαίτερα τα δύο τελευταία). Το ClustalW και το MUSCLE, ακολουθούν με μικρή διαφορά στην απόδοση, αλλά υπερτερούν σε ταχύτητα εκτέλεσης και σε απαιτήσεις σε μνήμη. Το Prnp/Prnp είναι επίσης καλό, αλλά πιο αργό. Το Kalign, είναι σε γενικές γραμμές ελαφρώς χειρότερο, αλλά είναι έως και 10 φορές γρηγορότερο από το CLUSTALW (πολύ δε περισσότερο από τα υπόλοιπα), και κατά συνέπεια καλύτερο για αναλύσεις μεγάλου όγκου δεδομένων σε καθημερινή βάση. Τέλος, οι αλγόριθμοι που κάνουν ολική στοίχιση, αποδίδουν σε γενικές γραμμές καλύτερα, εκτός αν στις πολλαπλές στοιχίσεις υπάρχουν μεγάλες περιοχές στο αμινοτελικό ή στο καρβοξυτελικό άκρο, οι οποίες δεν ταυτίζονται σε όλα τα μέλη της οικογένειας (δηλαδή, αν υπάρχουν οικογένειες με μέλη τα οποία εμφανίζουν τοπική ομοιότητα). Το T-Coffee γενικά, είναι ένας καλός συμβιβασμός, καθώς τα καταφέρνει σχετικά καλά σε όλες τις περιπτώσεις, ενώ το Dialign αποδεικνύεται καλύτερο μόνο σε κάποια από τα σετ με τέτοιες ακολουθίες (στις πιο ακραίες περιπτώσεις).



- Τα δύο τελευταία χαρακτηριστικά, δηλαδή η ταχύτητα και η ικανότητα σωστής στοίχισης σε περιπτώσεις τοπικής ομοιότητας πρέπει να ελέγχονται προσεκτικά και να λαμβάνονται σοβαρά υπόψη στην επιλογή προγράμματος. Η ταχύτητα για παράδειγμα, δεν είναι σημαντική όταν κάνουμε μια μελέτη μιας συγκεκριμένης οικογένειας (θέλουμε να πάρουμε την καλύτερη δυνατή στοίχιση και δεν μας πειράζει να περιμένουμε λίγο). Από την άλλη όμως, είναι ένας σημαντικός παράγοντας αν πρόκειται τις πολλαπλές στοιχίσεις να τις χρησιμοποιούμε λ.χ. για την υποβοήθηση μιας μεθόδου πρόγνωσης, γιατί σε αυτή την περίπτωση θα χρειάζεται να επαναλαμβάνουμε τις στοιχίσεις καθημερινά (για παράδειγμα, αν φτιάχνουμε μια διαδικτυακή εφαρμογή). Κάτι αντίστοιχο ισχύει και για τις τοπικές ομοιότητες των πρωτεϊνών. Αν μελετάμε μια συγκεκριμένη οικογένεια πρωτεϊνών, κατά πάσα πιθανότητα θα ξέρουμε τι είδους στοίχιση να περιμένουμε. Αν όμως πρόκειται η πολλαπλή στοίχιση να χρησιμοποιείται σε μια αυτοματοποιημένη διαδικασία, τότε δεν έχουμε αυτή την πολυτέλεια. Τέλος, ένας άλλος παράγοντας που πρέπει να λαμβάνεται υπόψη είναι και η ευκολία προς τον απλό χρήστη. Τα περισσότερα από τα προγράμματα που αναφέραμε (CLUSTALW, T-Coffee, Dialign, Kalign, MUSCLE, ProbAlign, Prrp/Prrn), προσφέρονται σαν διαδικτυακές εφαρμογές αλλά και σαν τοπικές εφαρμογές τις οποίες ο χρήστης μπορεί να εγκαταστήσει στον υπολογιστή του. Τα περισσότερα από αυτά, είναι ιδιαίτερα εύκολα στην εγκατάσταση σε όλα τα συστήματα (Windows, Linux, Mac), αλλά το COBALT και το PRALINE, τα οποία απαιτούν χρήση και άλλων προγραμμάτων (PSI-BLAST κλπ), είναι πιο δύσκολα στη ρύθμιση (και για την ακρίβεια, για το PRALINE δεν είμαστε σίγουροι αν υπάρχει και διαθέσιμη εφαρμογή πέραν της διαδικτυακής). Όλα τα παραπάνω είναι παράγοντες που πρέπει να λαμβάνονται σοβαρά υπόψη από τον χρήστη πριν επιλέξει με ποιο πρόγραμμα θα πραγματοποιήσει την ανάλυση του, και σε κάθε περίπτωση, είναι χρήσιμο πάντα κάποιος να δοκιμάζει αρκετές εναλλακτικές προτάσεις.