Ειδικά Θέματα Βιοπληροφορικής

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας
Λαμία, 2015

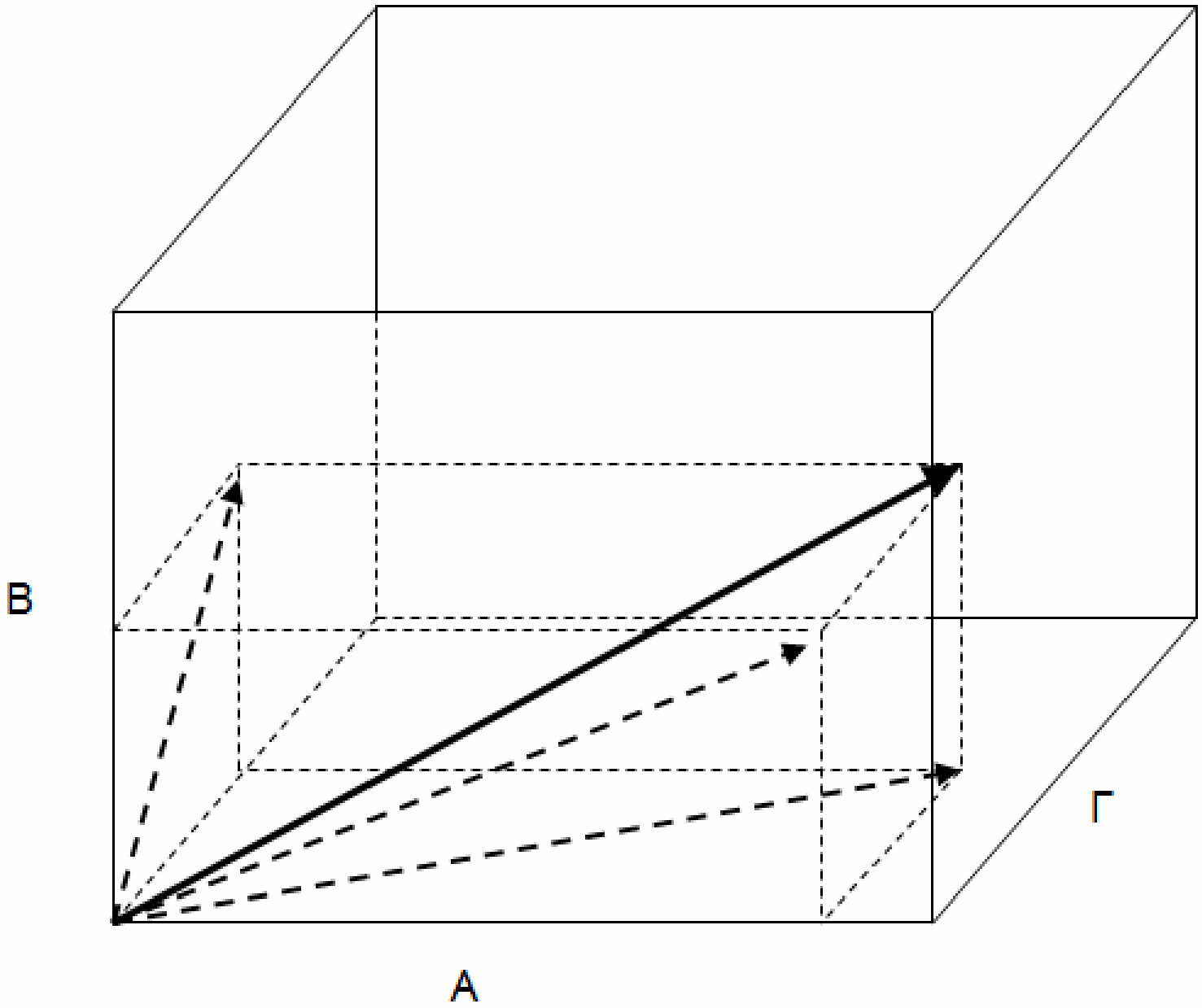
Πολλαπλή στοίχιση ακολουθιών και φυλογενετικά δέντρα

Πολλαπλή στοίχιση

- Αναφέρεται στην ταυτόχρονη στοίχιση περισσοτέρων 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 Ν διαστάσεις

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


B
Г
A

$$\mathbf{X_1} = x_{11}x_{12}\ldots x_{1n}$$
$$\mathbf{X_2} = x_{21}x_{22}\ldots x_{2n}$$
.....
$$\mathbf{X_r} = x_{r1}x_{r2}\ldots x_{rm}$$

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

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

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

$$\Delta_i.x = \begin{cases} (x), \alpha\nu\ \Delta_i = 1 \\ (-), \alpha\nu\ \Delta_i = 0 \end{cases}$$

$$S = \sum_i \log\left(\frac{p_{x1_i x2_i \ldots xr_i}}{q_{x1_i} q_{x2_i} \ldots q_{xr_i}} \right) = \sum_i s(x1_i, x2_i, \ldots, xr_i)$$

Πολυδιάστατο Score

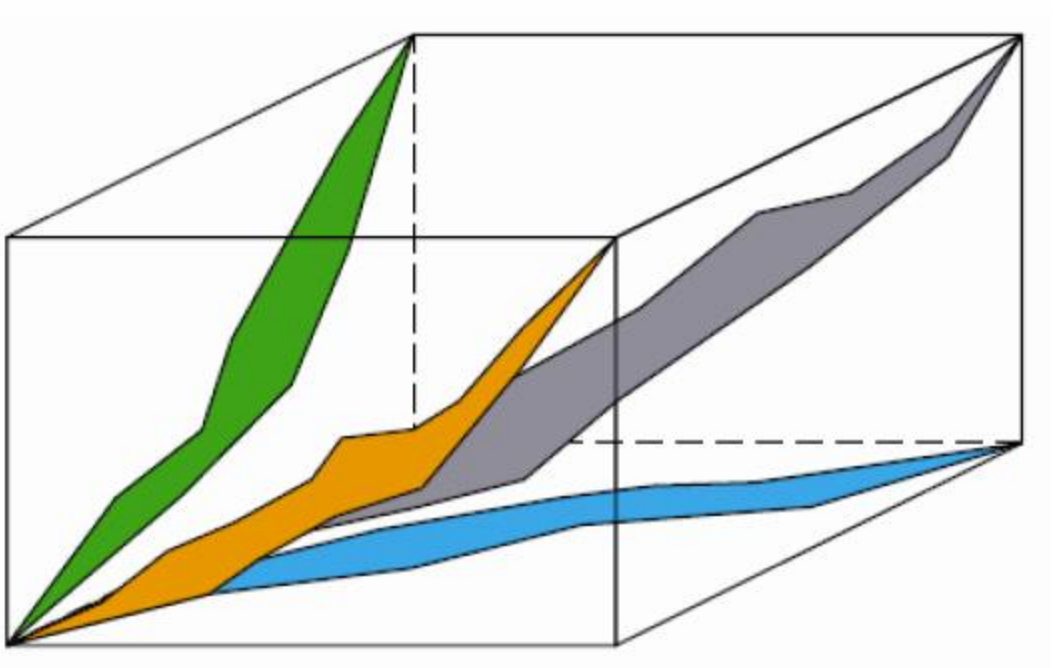
$$SP = \sum_i \left\{ \log\left(\frac{p_{x1_i x2_i}}{q_{x_i} q_{x2_i}} \right) + \ldots. + \log\left(\frac{p_{x(r-1)_i xr_i}}{q_{x(r-1)_i} q_{xr_i}} \right) \right\}$$

Sum of pairs Score

(αυτό χρησιμοποιείται)

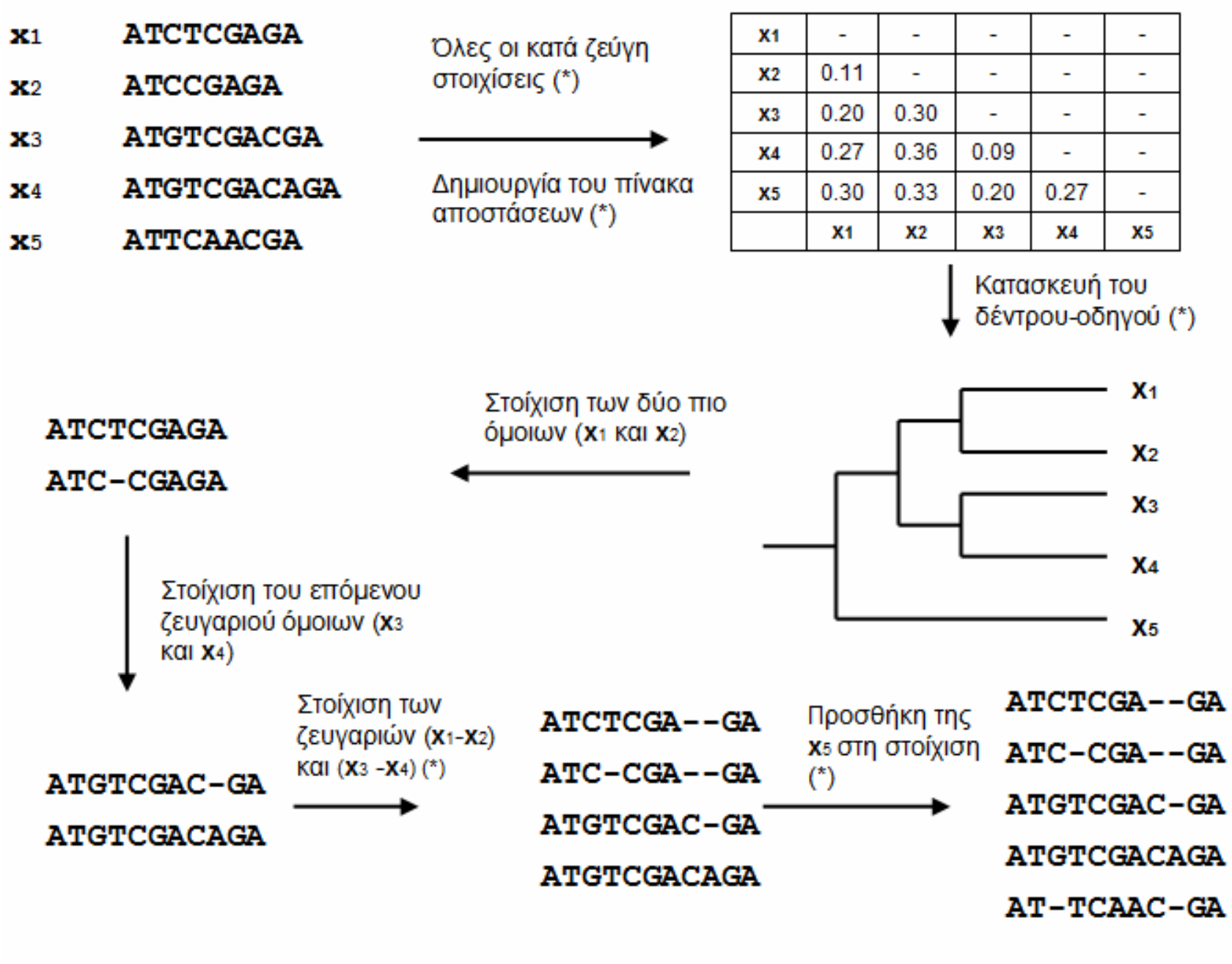
MSA (Carillo and Lipman,1988, Lipman et al,1989)

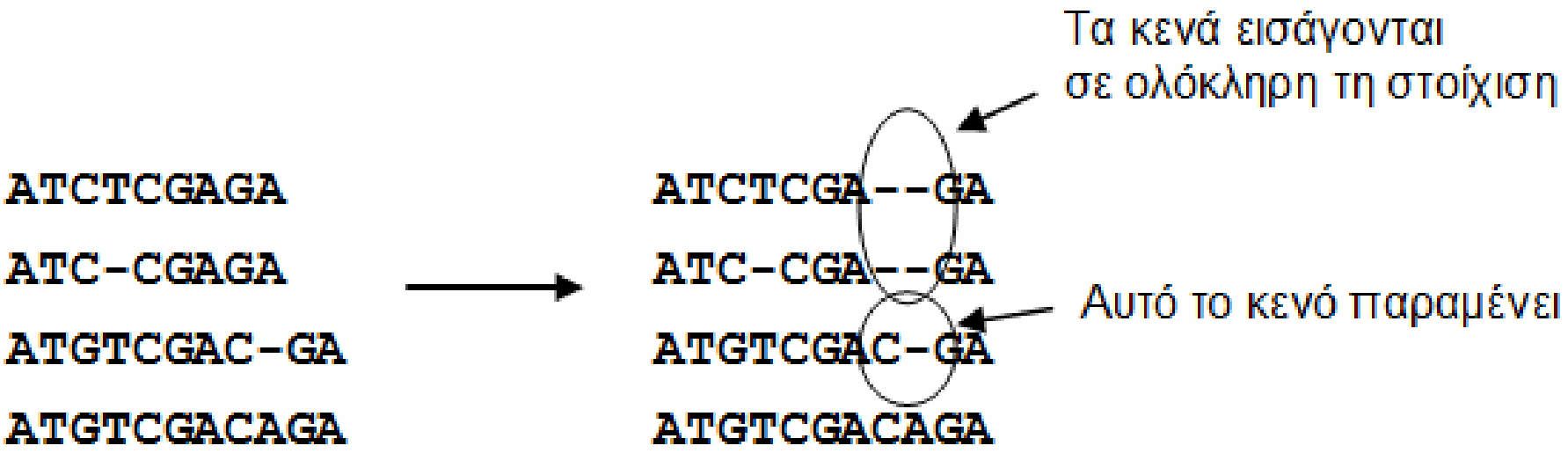
- Περιορισμός του εύρους αναζήτησης του δυναμικού προγραμματισμού
- Χρήση του Sum of Pairs Score



Προοδευτική πολλαπλή στοίχιση

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

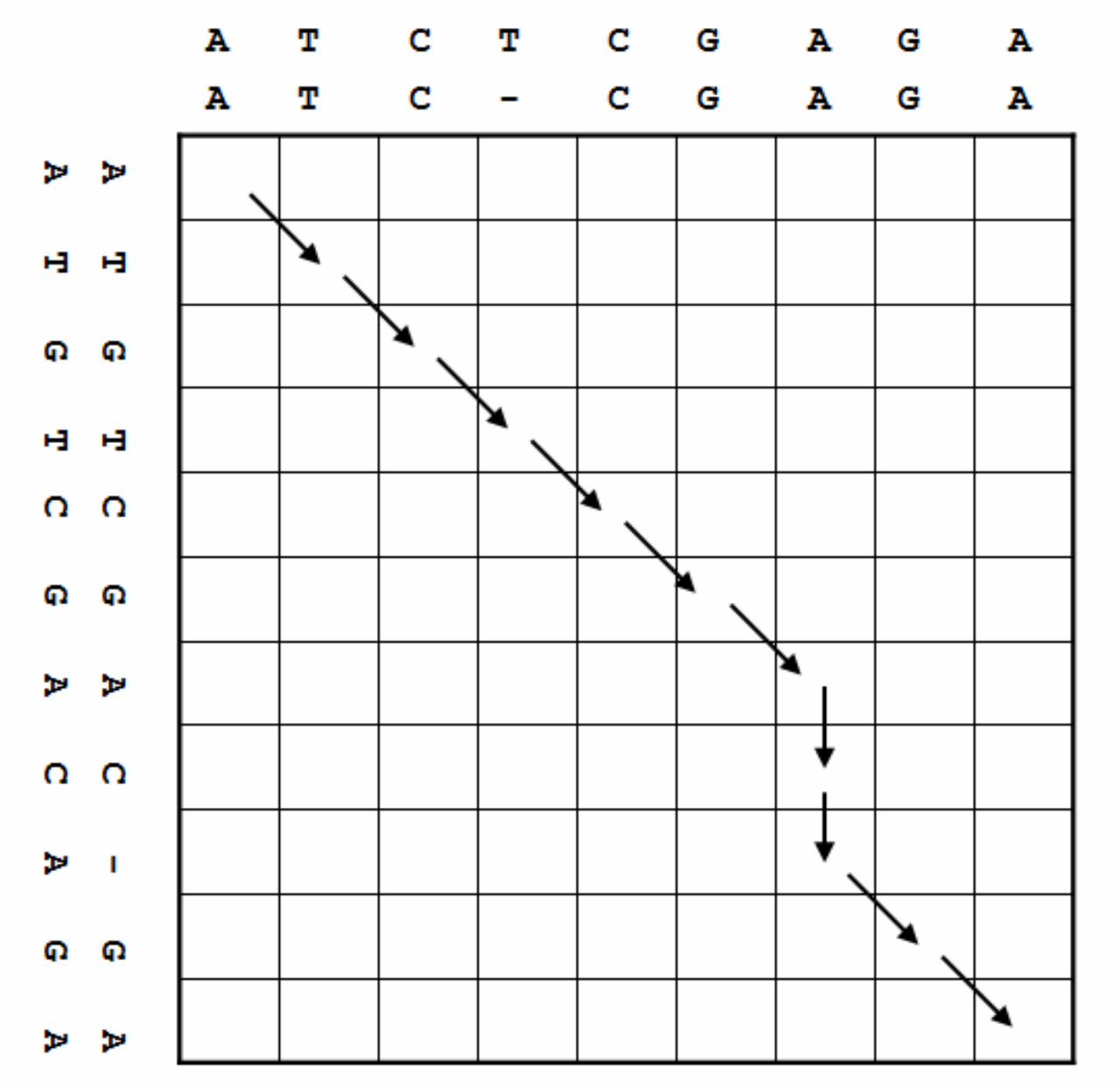

x1 ATCTCGAGA
x2 ATCCGAGA
x3 ATGTCGACGA
x4 ATGTCGACAGA
x5 ATTCAACGA
Όλες οι κατά ζεύγη στοιχίσεις (*)
Δημιουργία του πίνακα αποστάσεων (*)
x1 - - - - -
x2 0.11 - - - -
x3 0.20 0.30 - - -
x4 0.27 0.36 0.09 - -
x5 0.30 0.33 0.20 0.27 -
x1 x2 x3 x4 x5
Κατασκευή του δέντρου-οδηγού (*)
x1
x2
x3
x4
x5
Στοίχιση 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

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

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


A T C T C G A G A
A T C - C G A G A
A T G T C C G A C C A G A
A T G T C C G A C - G A

Feng and Doolitle (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) αλλα και με δυναμικό προγραμματισμό
- Οι αποστάσεις υπολογίζονται απευθείας από την επί τοις εκατό ομοιότητα των ακολουθιών x (D=1-x/100)
- Δέντρο οδηγός με Neibhor-Joining method
- .Η μέθοδος είναι διαθέσιμη στη διεύθυνση 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

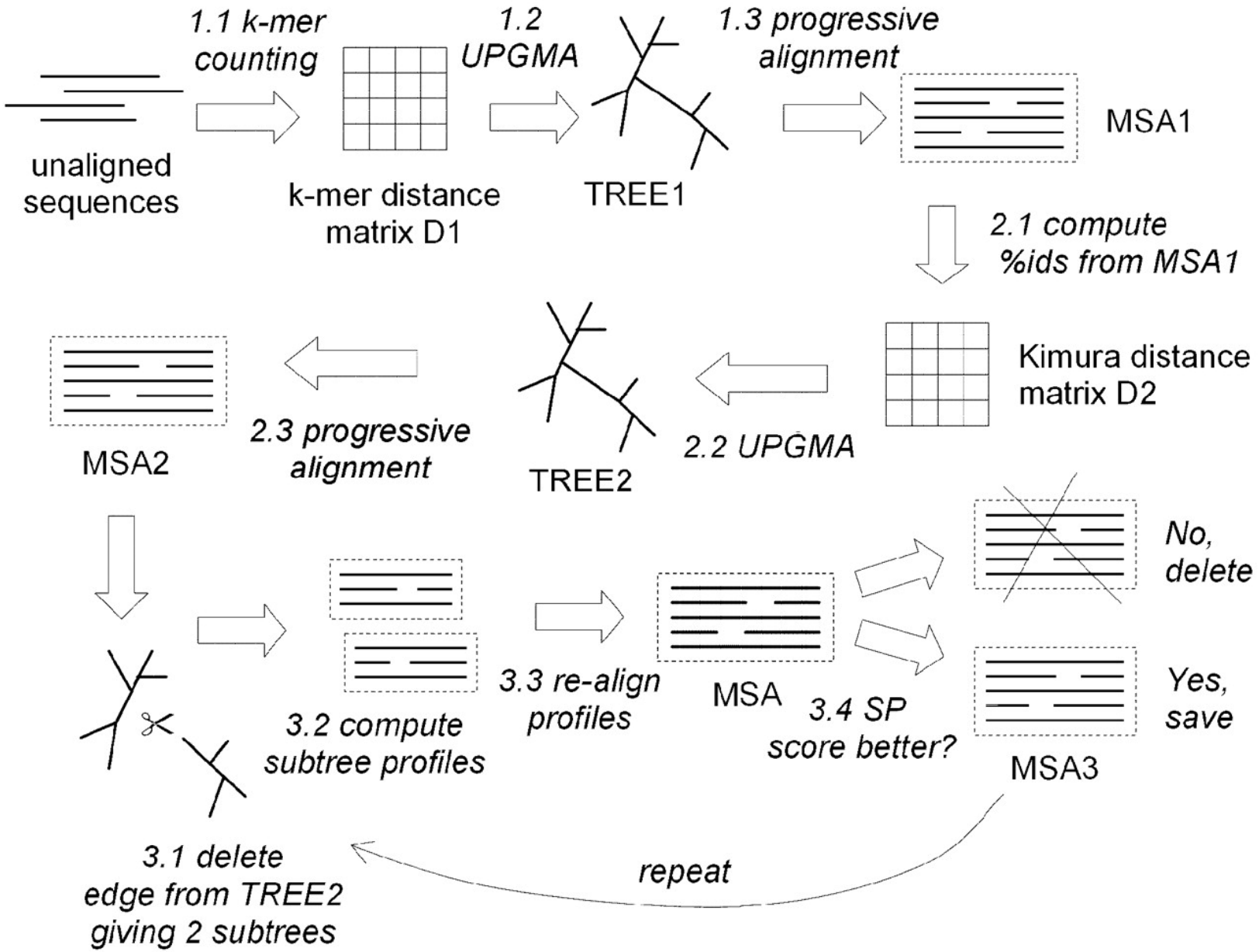
Επαναληπτικές μέθοδοι και μέθοδοι που βασίζονται στη συνέπεια

- 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)
- 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)

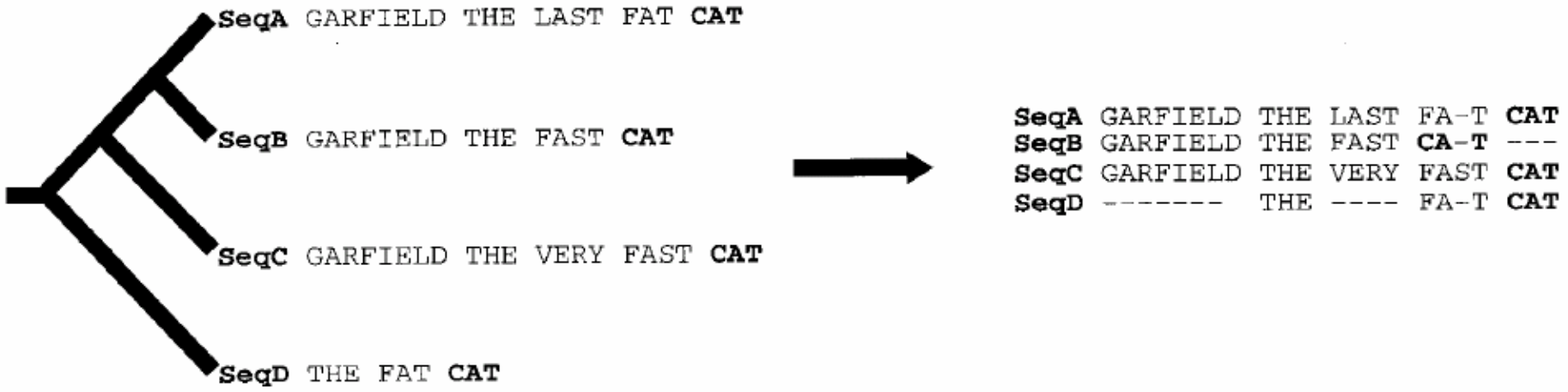
MUSCLE


1.1 k-mer counting
1.2 UPGMA
1.3 progressive alignment
unaligned sequences
k-mer distance matrix D1
TREE1
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
3.1 delete edge from TREE2 giving 2 subtrees
3.2 compute subtree profiles
3.3 re-align profiles
MSA
3.4 SP score better?
No, delete
Yes, save
MSA3
repeat

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)Primary Library

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

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

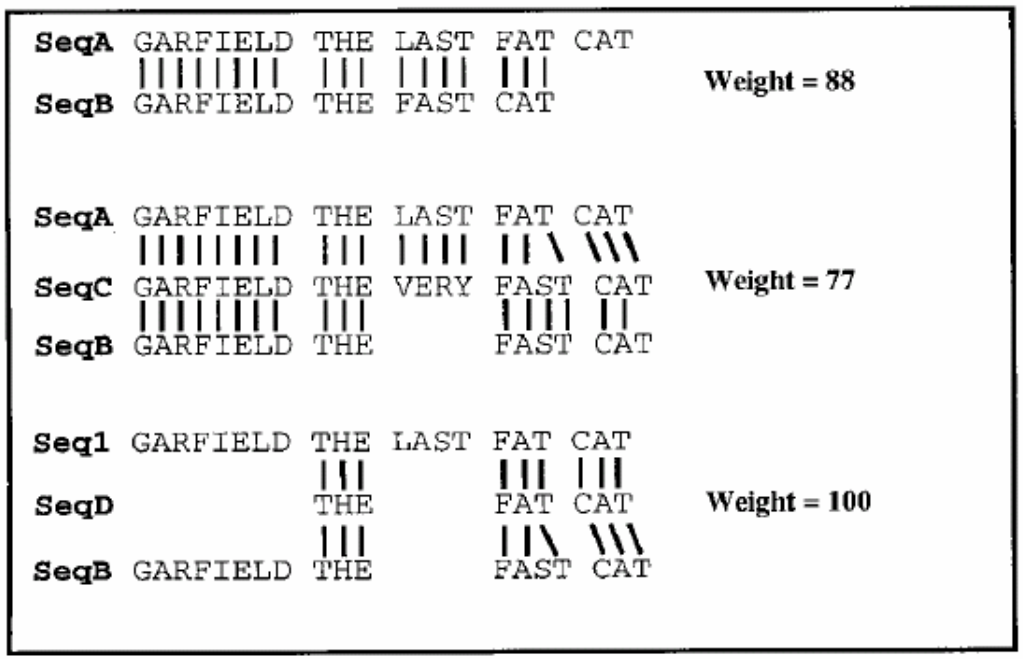
```
SeqA GARFIELD THE LAST FAT CAT
SeqD -------  THE ---- FAT CAT
```
Prim. Weight =100

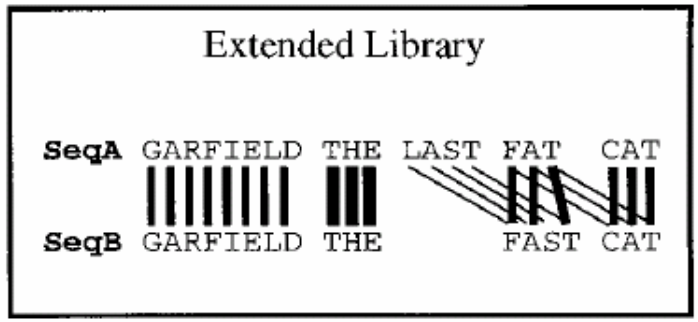
```
SeqB GARFIELD THE ---- FAST CAT
SeqC GARFIELD THE VERY FAST CAT
```
Prim Weight = 100

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

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

c)Extended Library for seq1 and seq2





Dynamic Programming

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

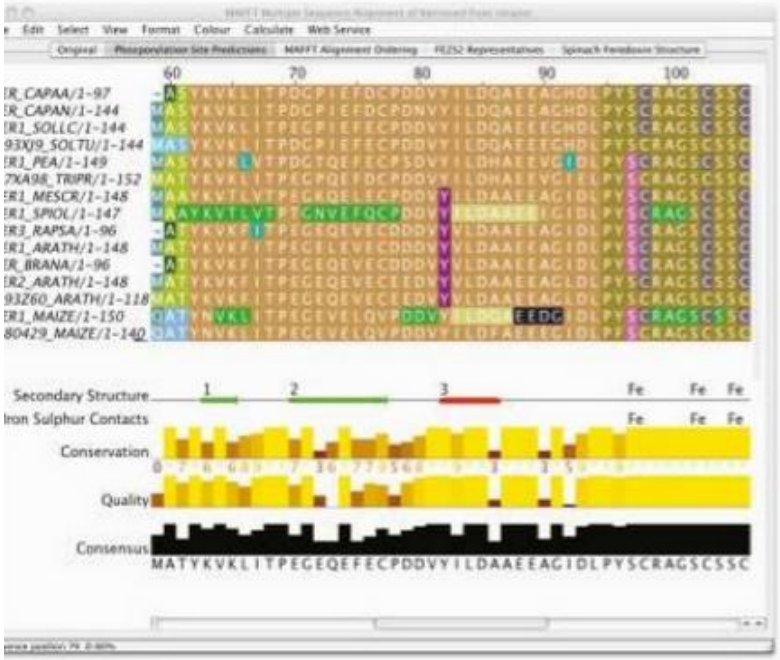
Dialign

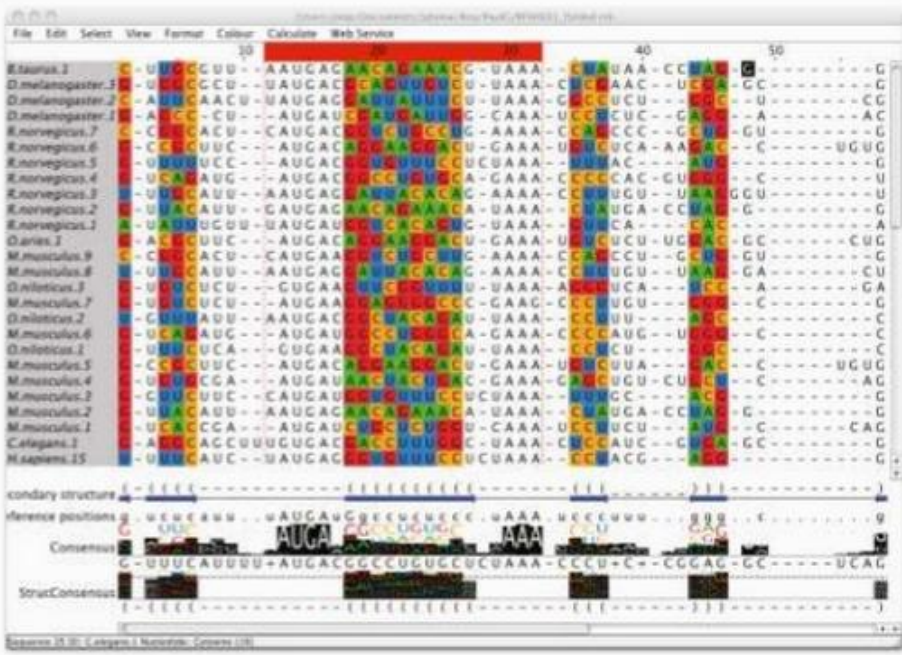
- Το Dialign πραγματοποιεί τοπική στοίχιση και στοιχίζει ολόκληρα τμήματα παρά κατάλοιπα.
- Αρχικά, όλες οι ανά δυο στοιχίσεις πραγματοποιούνται και συλλέγονται οι στοιχισμένες περιοχές στις οποίες δεν υπάρχουν κενά.
- Το όνομα 'Dialign' βγαίνει από αυτές τις διαγώνιες περιοχές (diagonal alignments in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

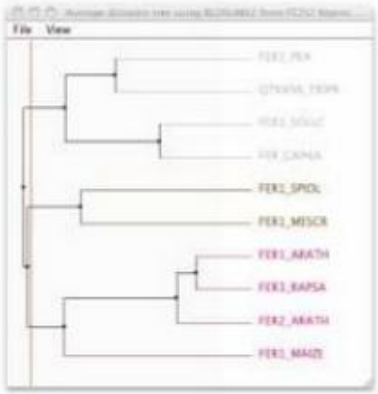
Οπτικοποίηση και επεξεργασιά πολλαπλών στοιχίσεων

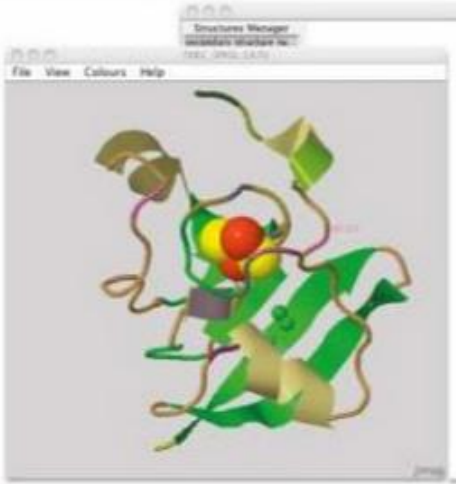
- Jalview (http://www.jalview.org/)
- Strap (http://www.bioinformatics.org/strap/)
- Seqpup (http://iubio.bio.indiana.edu/soft/molbio/seqpup/java/seqpup-doc.html)
- Seaview (http://pbil.univ-lyon1.fr/software/seaview.html)
- Cinema (http://aig.cs.man.ac.uk/research/utopia/cinema/cinema.php)
- Boxshade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

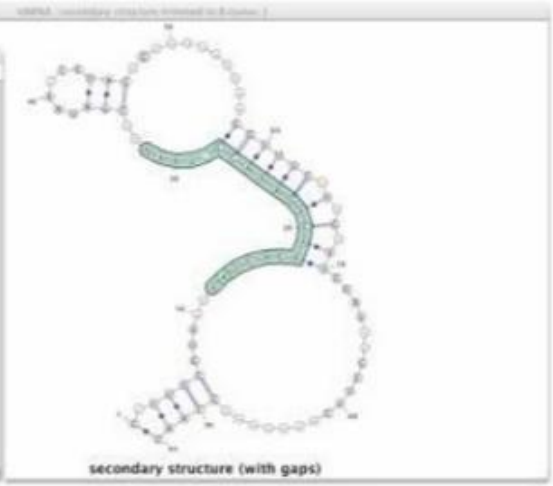
Jalview

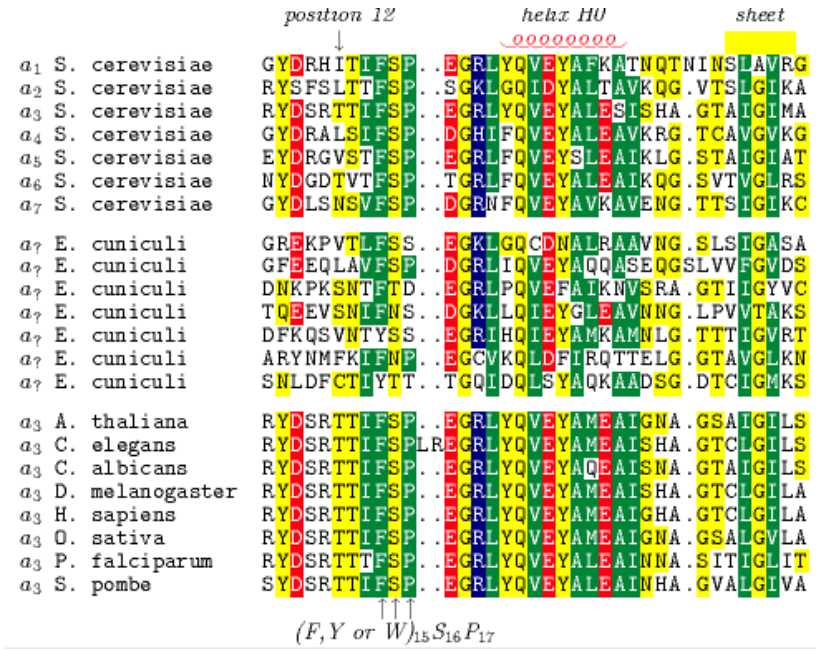

Secondary Structure
Iron Sulphur Contacts
Conservation
Quality
Consensus
MATYKVKLITPEGEQEFECPDDVYILDAAEEAGIDLPYSCRAGSCSSC


Consensus
StrucConsensus

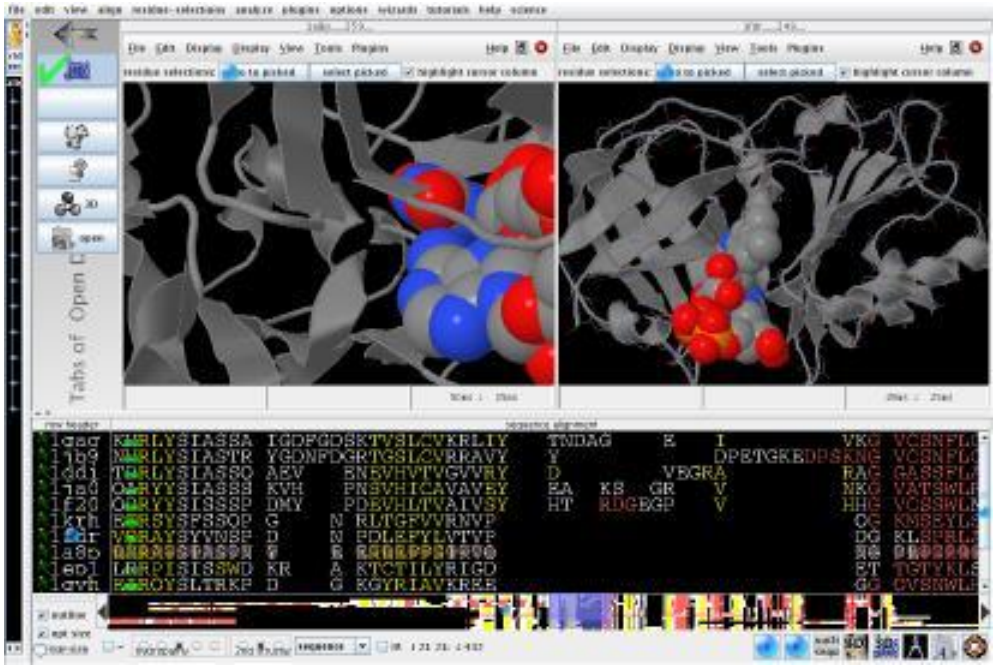

FER1_SPIOL
FER1_MESCR
FER1_ARATH
FER1_BRAPA
FER2_ARATH
FER1_MAIZE



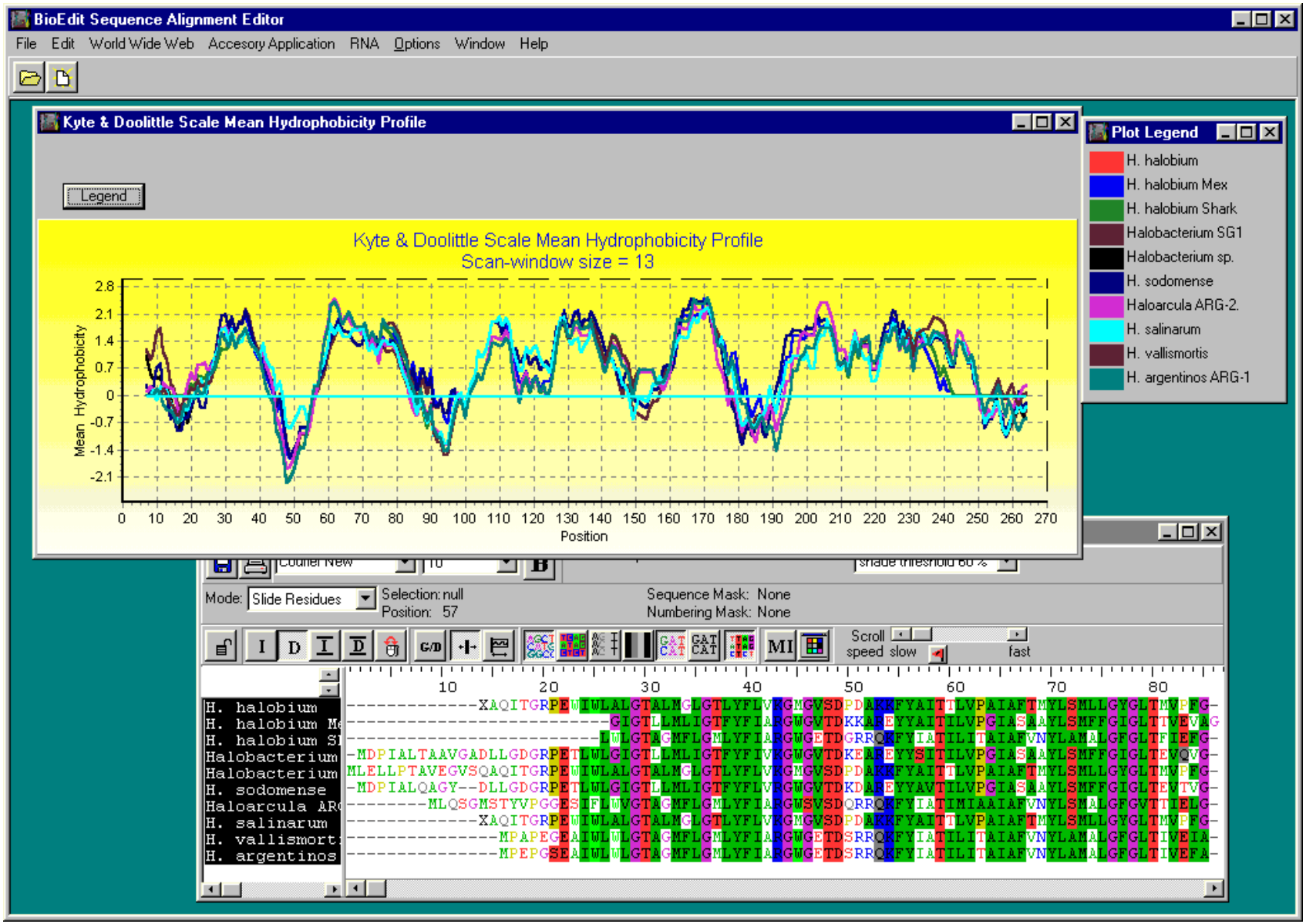

secondary structure (with gaps)

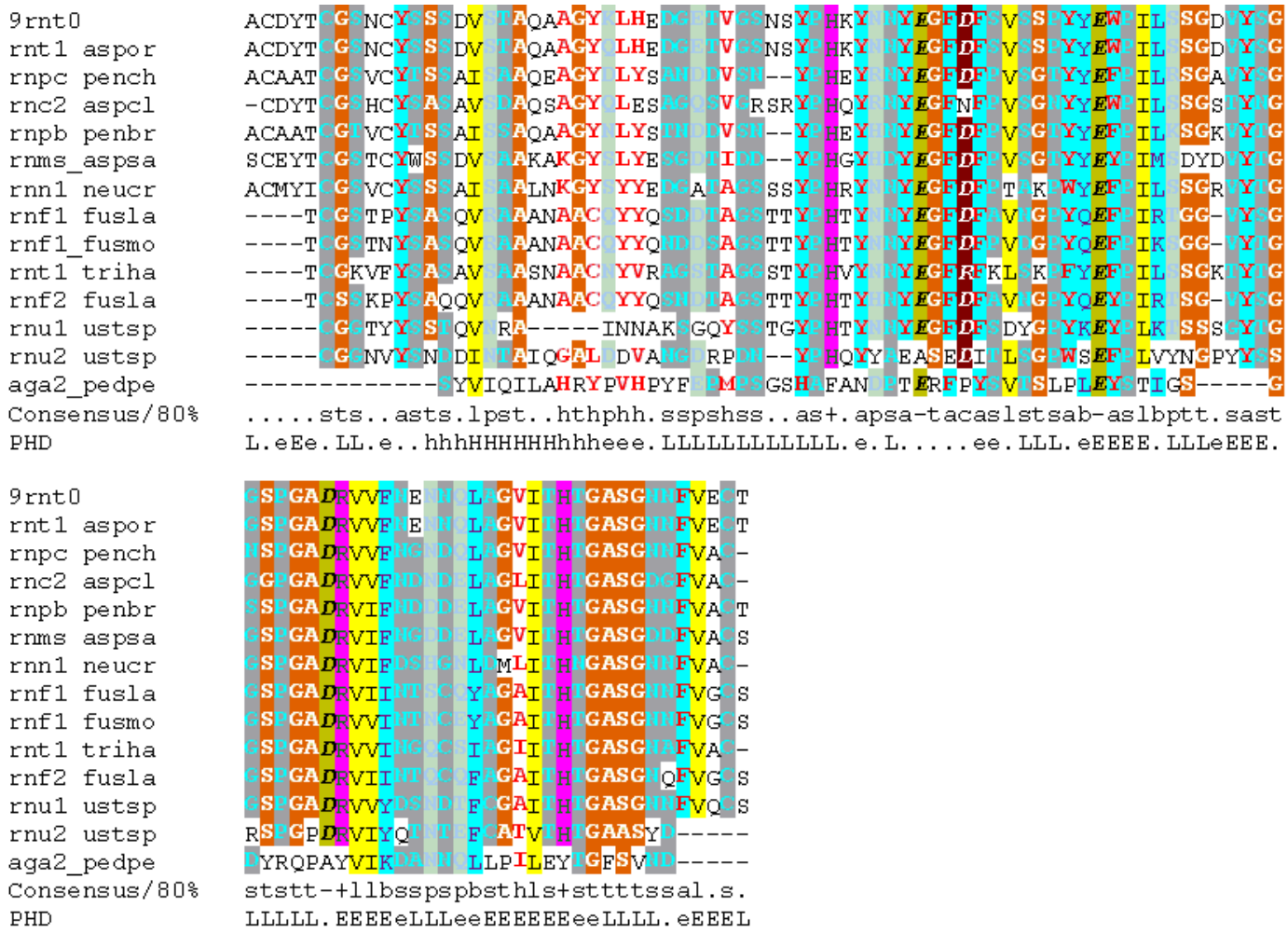

position 12
helix HU
sheet
a_1 S. cerevisiae GYDRHITIFSP..EGRLYQVEYAFKATNQTNINSLAVRG
a_2 S. cerevisiae RYSFSLTTFSP..SGKLGQIDYALTAVKQG.VTSLGIKA
a_3 S. cerevisiae RYDSRTTIFSP..EGRLYQVEYALESISHA.GTAIGIMA
a_4 S. cerevisiae GYDRALSIFSP..DGHIFQVEYALEAVKRG.TCAVGVKG
a_5 S. cerevisiae EYDRGVSTFSP..EGRLFQVEYSLEAIKLG.STAIGIAT
a_6 S. cerevisiae NYDGDTVTFSP..TGRLFQVEYALEAIKQG.SVTVGLRS
a_7 S. cerevisiae GYDLSNSVFSP..DGRNFQVEYAVKAVENG.TTSIGIKC
a_7 E. cuniculi GREKPVTLFSS..EGKLGQCDNALRAAVNG.SLSIGASA
a_7 E. cuniculi GFEEQLAVFSP..DGRLIQVEYAQQASEQGSLVVFGVDS
a_7 E. cuniculi DNKPKSNTFTD..EGRLPQVEFAIKNVSRA.GTIIGYVC
a_7 E. cuniculi TQEEVSNIFNS..DGKLLQIEYGLEAVNNG.LPVVTAKS
a_7 E. cuniculi DFKQSVNTYSS..EGRIHQIEYAMKAMNLG.TTTIGVRT
a_7 E. cuniculi ARYNMFKIFNP..EGCVKQLDFIRQTTELG.GTAVGLKN
a_7 E. cuniculi SNLDFCTIYTT..TGQIDQLSYAQKAADSG.DTCIGMKS
a_3 A. thaliana RYDSRTTIFSP..EGRLYQVEYAMEAIGNA.GSAIGILS
a_3 C. elegans RYDSRTTIFSPLREGRLYQVEYAMEAISHA.GTCLGILS
a_3 C. albicans RYDSRTTIFSP..EGRLYQVEYAQEAISNA.GTAIGILS
a_3 D. melanogaster RYDSRTTIFSP..EGRLYQVEYAMEAISHA.GTCLGILA
a_3 H. sapiens RYDSRTTIFSP..EGRLYQVEYAMEAIGHA.GTCLGILA
a_3 O. sativa RYDSRTTIFSP..EGRLYQVEYAMEAIGNA.GSALGVLA
a_3 P. falciparum RYDSRTTTFSP..EGRLYQVEYALEAINNA.SITIGLIT
a_3 S. pombe SYDSRTTIFSP..EGRLYQVEYALEAINHA.GVALGIVA
$(F, Y \text{ or } W)_{15} S_{16} P_{17}$

Strap

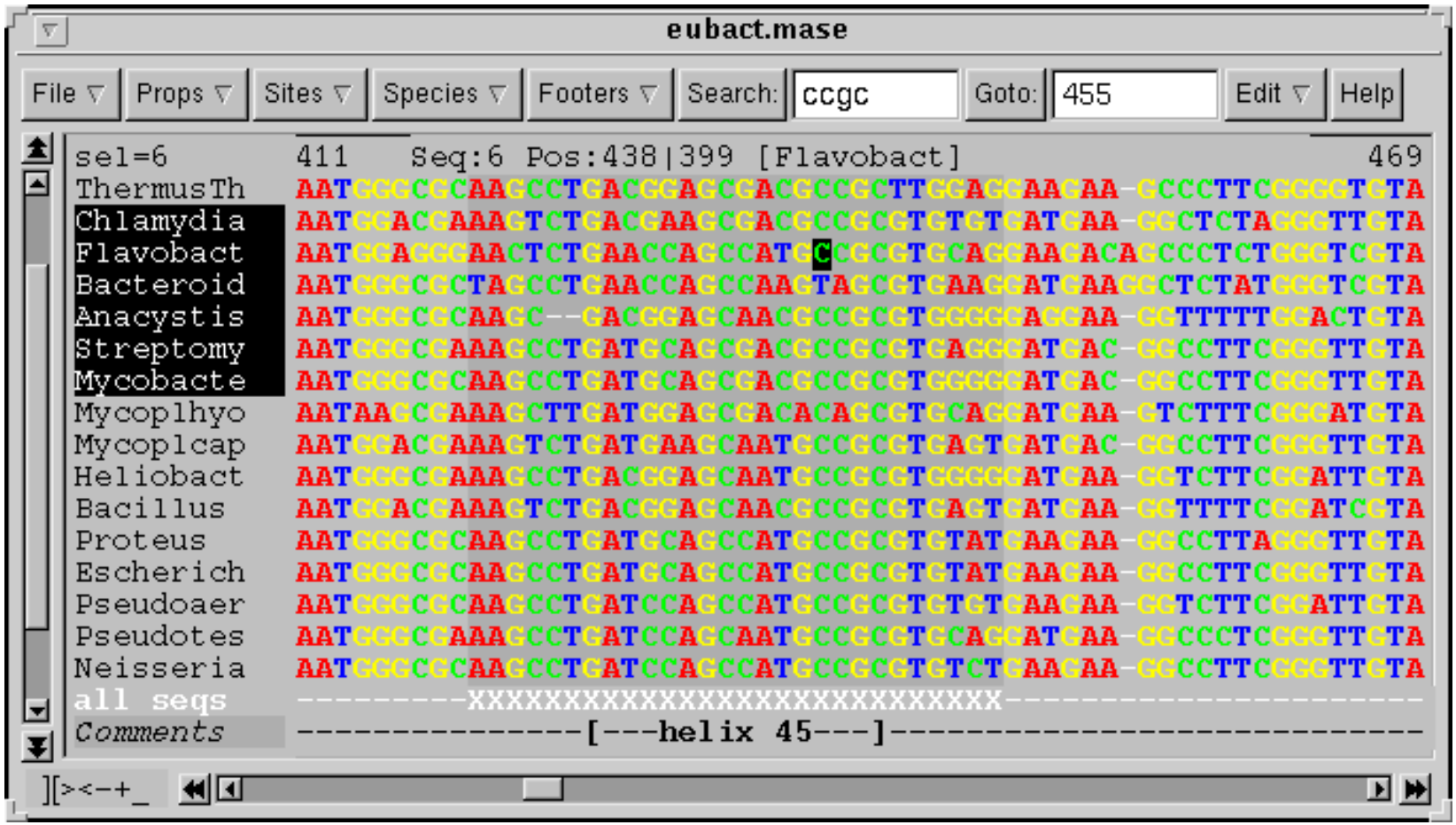

Tabs or Open

BioEdit


BioEdit Sequence Alignment Editor
File Edit World Wide Web Accesory Application RNA Options Window Help
Kyte & Doolittle Scale Mean Hydrophobicity Profile
Legend
Kyte & Doolittle Scale Mean Hydrophobicity Profile
Scan-window size = 13
Mean Hydrophobicity
Position
Plot Legend
H. halobium
H. halobium Mex
H. halobium Shark
Halobacterium SG1
Halobacterium sp.
H. sodomense
Haloarcula ARG-2.
H. salinarum
H. vallismortis
H. argentinos ARG-1
Mode: Slide Residues
Selection: null
Position: 57
Sequence Mask: None
Numbering Mask: None
Scroll speed slow fast
H. halobium
H. halobium M
H. halobium S
Halobacterium
Halobacterium
H. sodomense
Haloarcula AR
H. salinarum
H. vallismort
H. argentinos


9rnt0
rnt1 aspor
rnpc pench
rnc2 aspcl
rnpb penbr
rnms_aspsa
rnn1 neucr
rnf1 fusla
rnf1_fusmo
rnt1 triha
rnf2 fusla
rnu1 ustsp
rnu2 ustsp
aga2_pedpe
Consensus/80%
PHD

Chroma


eubact.mase
File
Props
Sites
Species
Footers
Search:
ccgc
Goto:
455
Edit
Help
sel=6
411 Seq:6 Pos:438|399 [Flavobact] 469
ThermusTh
Chlamydia
Flavobact
Bacteroid
Anacystis
Streptomy
Mycobacte
Mycoplhyo
Mycoplcap
Heliobact
Bacillus
Proteus
Escherich
Pseudoaer
Pseudotes
Neisseria
all seqs
Comments
[---helix 45---]

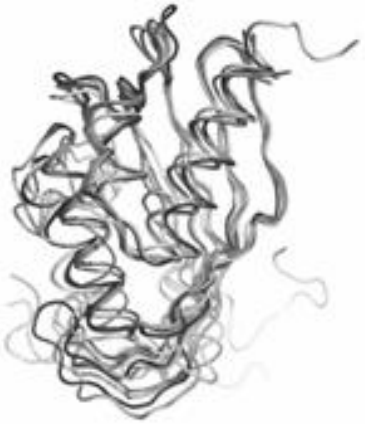
Seaview

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 Prrp/Prrn 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-Cofffee, 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.