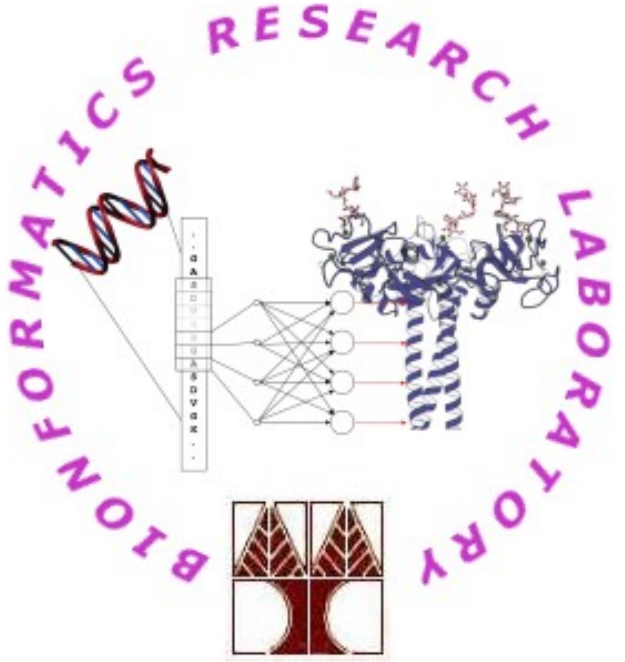
Πολλαπλές στοιχίσεις ακολουθιών (Προοδευτικές μέθοδοι)



Vasilis Promponas
Bioinformatics Research Laboratory
Department of Biological Sciences
University of Cyprus

Σύνοψη

- Εισαγωγή
- Πολλαπλή στοίχιση και ΔΠ
- Βαθμονόμηση Πολλαπλών Στοιχίσεων
- Μέθοδοι Προοδευτικής Πολλαπλής Στοίχισης
- Συζήτηση
  - ...

ΠΟΛΛΑΠΛΗ ΣΤΟΙΧΙΣΗ ΑΚΟΛΟΥΘΙΩΝ

- Από Υπολογιστική Πλευρά:

  "… two strings good, four strings better …"

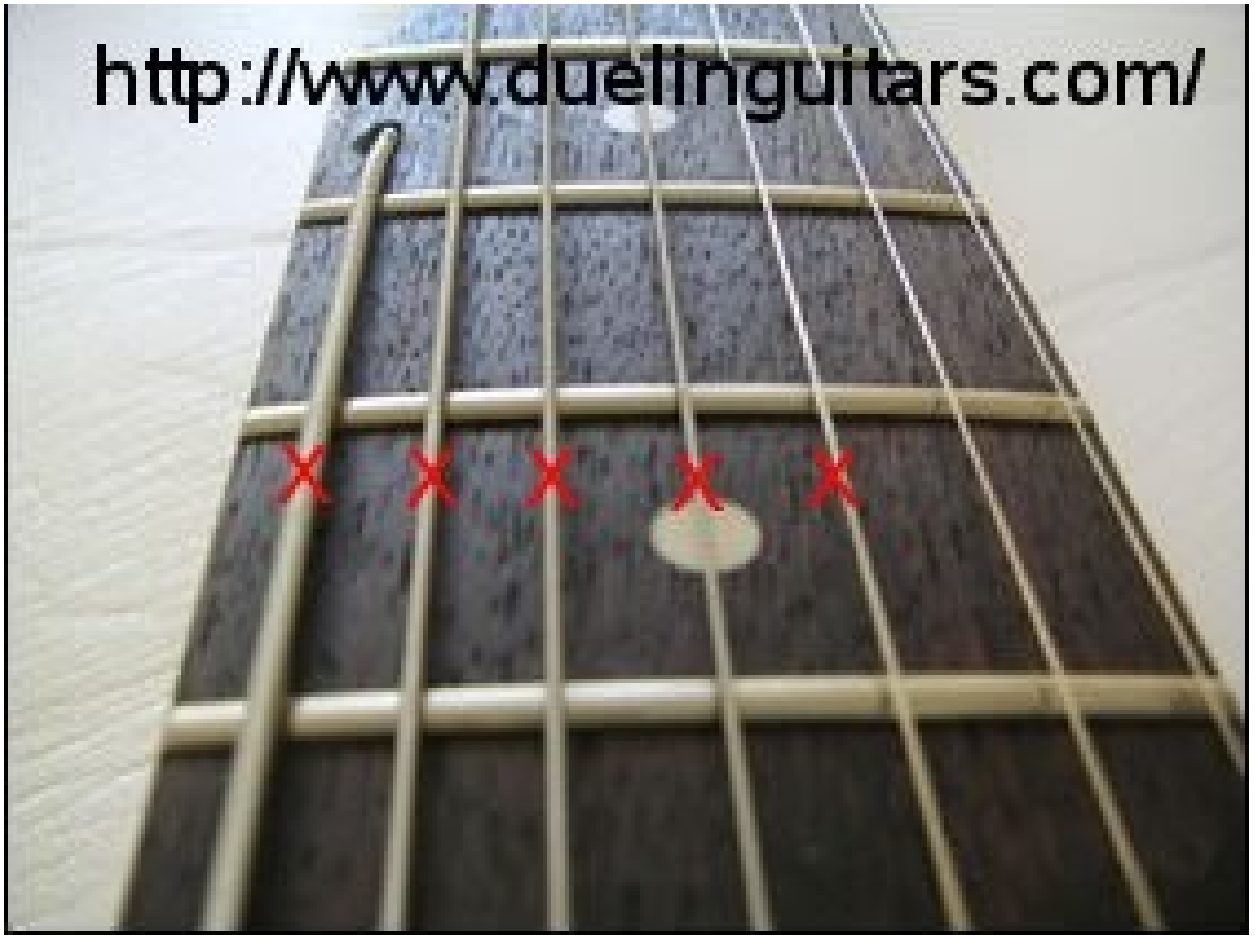
  (Gusfield, 1997)

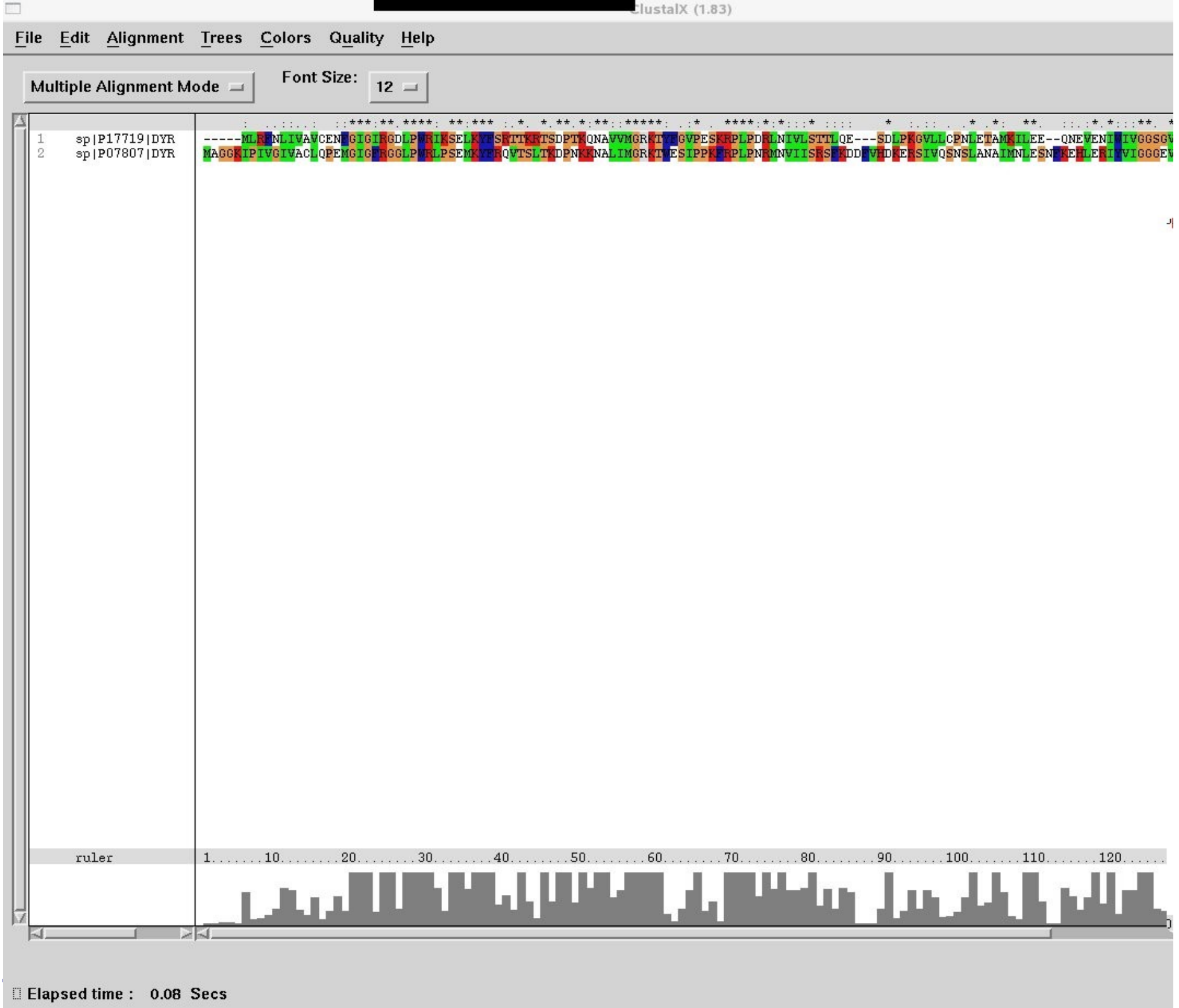
- Από Βιολογική Πλευρά

  "One or two homologous sequences whisper … a full multiple alignment shouts out loud"

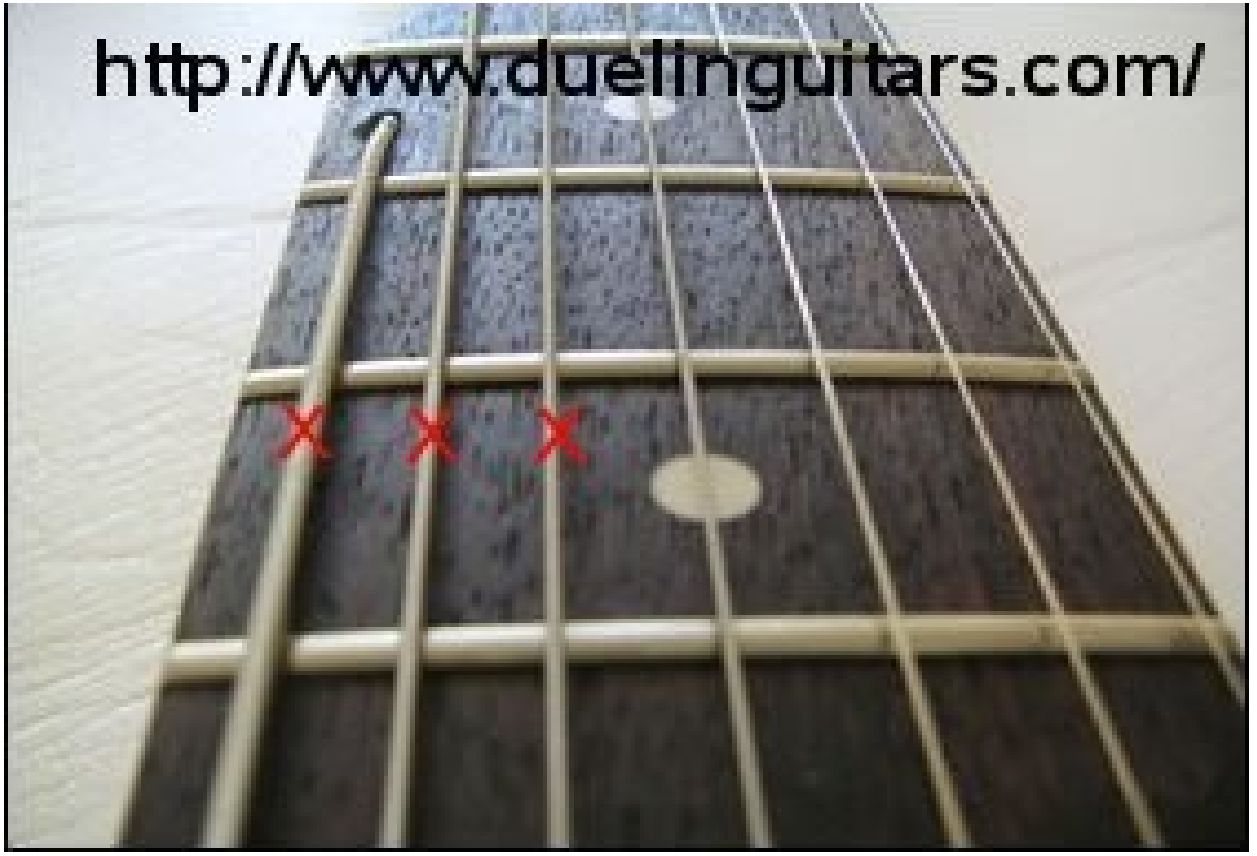
  (Hubbard, Lesk and Tramontano, 1996)

Two strings good ...




ClustalX (1.83)
File Edit Alignment Trees Colors Quality Help
Multiple Alignment Mode
Font Size:
12
1 sp|P17719|DYR
2 sp|P07807|DYR
ruler
1.....10.....20.....30.....40.....50.....60.....70.....80.....90.....100.....110.....120.....
Elapsed time : 0.08 Secs

... four strings better!!




ClustalX (1.83)
File Edit Alignment Trees Colors Quality Help
Multiple Alignment Mode
Font Size: 14
1 DRTS1_AR
2 sp|P1771
3 sp|P0780
4 sp|Q0451
ruler
1.....10.....20.....30.....40.....50.....60.....70.....
50.....60.....70
Elapsed time : 0.12 Secs

42 !?

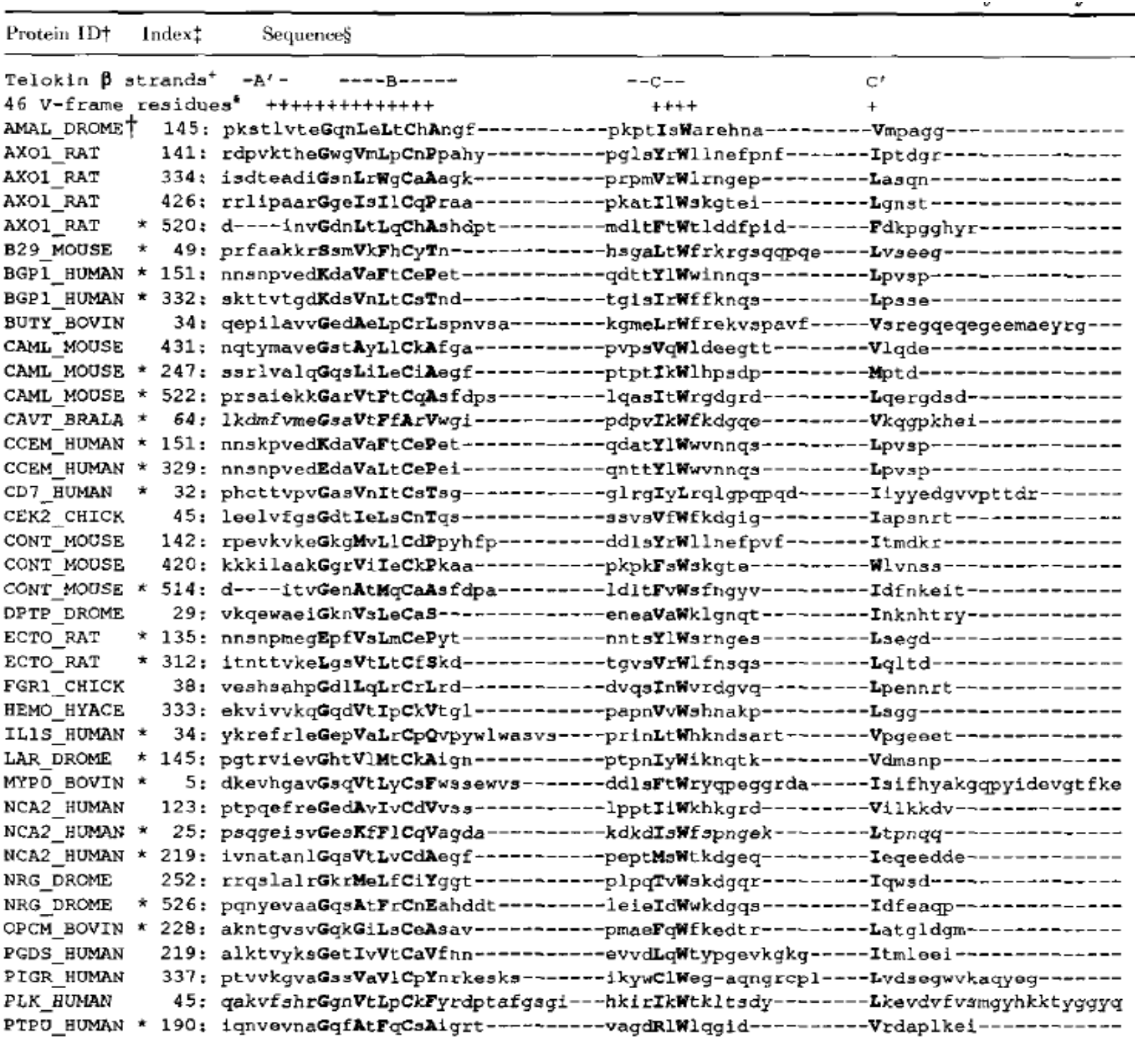
45++++

J. Mol. Biol. (1994) **238**, 528–539

Many of the Immunoglobulin Superfamily Domains in Cell Adhesion Molecules and Surface Receptors Belong to a New Structural Set Which is Close to That Containing Variable Domains

Yahouda Harpaz[1] and Cyrus Chothia[1,2]

[1]*Cambridge Centre for Protein Engineering and*
[2]*MRC Laboratory of Molecular Biology*
Hills Road, Cambridge CB2 2QH, U.K.


Protein ID† Index‡ Sequence§
Telokin β strands* -A'- ----B----- --C-- C'
46 V-frame residues* ++++++++++++++ ++++ +
AMAL_DROME† 145: pkstlvteGqnLeLtChAngf----------pkptIsWarehna---------Vmpagg----------
AXO1_RAT 141: rdpvktheGwgVmLpCnPpahy---------pglsYrWllnefpnf-------Iptdgr----------
AXO1_RAT 334: isdteadiGsnLrWgCaAagk----------prpmVrWlrngep---------Lasqn----------
AXO1_RAT 426: rrlipaarGgeIsIlCqPraa----------pkatIlWskgtei---------Lgnst----------
AXO1_RAT * 520: d----invGdnLtLqChAshdpt---------mdltFtWtlddfpid-------Fdkpgghyr----------
B29_MOUSE * 49: prfaakkrSsmVkFhCyTn----------hsgaLtWfrkrgsqqpqe----Lvseeq----------
BGP1_HUMAN * 151: nnsnpvedKdaVaFtCePet----------qdttYlWwinnqs---------Lpvsp----------
BGP1_HUMAN * 332: skttvtgdKdsVnLtCsTnd----------tgisIrWffknqs---------Lpsse----------
BUTY_BOVIN 34: qepilavvGedAeLpCrLspnvsa--------kgmeLrWfrekvspavf-----Vsregqeqegeemaeyrg---
CAML_MOUSE 431: nqtymaveGstAyLlCkAfga----------pvpsVqWldeegtt--------Vlqde----------
CAML_MOUSE * 247: ssrlvalqGqsLiLeCiAegf----------ptptIkWlhpsdp---------Mptd----------
CAML_MOUSE * 522: prsaiekkGarVtFtCqAsfdps---------lqasItWrgdgrd---------Lqergdsd----------
CAVT_BRALA * 64: lkdmfvmeGsaVtFfArVwgi----------pdpvIkWfkdgqe---------Vkqgpkhei----------
CCEM_HUMAN * 151: nnskpvedKdaVaFtCePet----------qdatYlWwvnnqs---------Lpvsp----------
CCEM_HUMAN * 329: nnsnpvedEdaVaLtCePei----------qnttYlWwvnnqs---------Lpvsp----------
CD7_HUMAN * 32: phcttvpvGasVnItCsTsg----------glrgIyLrqlgpqpqd------Iiyyedgvvpttdr-------
CEK2_CHICK 45: leelvfgsGdtIeLsCnTqs----------ssvsVfWfkdgig---------Iapsnrt----------
CONT_MOUSE 142: rpevkvkeGkgMvLlCdPpyhfp---------ddlsYrWllnefpvf-------Itmdkr----------
CONT_MOUSE 420: kkkilaakGgrViIeCkPkaa----------pkpkFsWskgte---------Wlvnss----------
CONT_MOUSE * 514: d----itvGenAtMqCaAsfdpa---------ldltFvWsfngyv---------Idfnkeit----------
DPTP_DROME 29: vkqewaeiGknVsLeCaS----------eneaVaWklgnqt---------Inknhtry----------
ECTO_RAT * 135: nnsnpmegEpfVsLmCePyt----------nntsYlWsrnges---------Lsegd----------
ECTO_RAT * 312: itnttvkeLgsVtLtCfSkd----------tgvsVrWlfnsqs---------Lqltd----------
FGR1_CHICK 38: veshsahpGdlLqLrCrLrd----------dvqsInWvrdgvq---------Lpennrt----------
HEMO_HYACE 333: ekvivvkqGqdVtIpCkVtgl----------papnVvWshnakp---------Lsgg----------
IL1S_HUMAN * 34: ykrefrleGepVaLrCpQvpywlwasvs----prinLtWhkndsart-------Vpgeeet----------
LAR_DROME * 145: pgtrvievGhtVlMtCkAign----------ptpnIyWiknqtk---------Vdmsnp----------
MYPO_BOVIN * 5: dkevhgavGsqVtLyCsFwssewvs-------ddlsFtWryqpeggrda-----Isifhyakgqpyidevgtfke
NCA2_HUMAN 123: ptpqefreGedAvIvCdVvss----------lpptIiWkhkgrd---------Vilkkdv----------
NCA2_HUMAN * 25: psqgeisvGesKfFlCqVagda----------kdkdIsWfspngek--------Ltpnqq----------
NCA2_HUMAN * 219: ivnatanlGqsVtLvCdAegf----------peptMsWtkdgeq---------Ieqeedde----------
NRG_DROME 252: rrqslalrGkrMeLfCiYggt----------plpqTvWskdgqr---------Iqwsd----------
NRG_DROME * 526: pqnyevaaGqsAtFrCnEahddt---------leieIdWwkdgqs---------Idfeaqp----------
OPCM_BOVIN * 228: akntgvsvGqkGiLsCeAsav----------pmaeFqWfkedtr---------Latgldgm----------
PGDS_HUMAN 219: alktvyksGetIvVtCaVfnn----------evvdLqWtypgevkgkg-----Itmleei----------
PIGR_HUMAN 337: ptvvkgvaGssVaVlCpYnrkesks-------ikywClWeg-aqngrcpl----Lvdsegwvkaqyeg-------
PLK_HUMAN 45: qakvfshrGgnVtLpCkFyrdptafgsgi---hkirIkWtkltsdy--------Lkevdvfvsmgyhkktyggyq
PTPU_HUMAN * 190: iqnvevnaGqfAtFqCsAigrt----------vagdRlWlqgid---------Vrdaplkei----------

ΠΟΙΟ ΕΙΝΑΙ ΤΟ ΖΗΤΟΥΜΕΝΟ;

- ΑΝΤΙΣΤΟΙΧΙΣΗ ΚΑΤΑΛΟΙΠΩΝ (ΣΤΗΝ ΙΔΙΑ ΣΤΗΛΗ) ΠΟΥ ΕΧΟΥΝ ΠΡΟΕΛΘΕΙ ΑΠΟ ΕΝΑ «ΑΡΧΕΓΟΝΟ» ΚΑΤΑΛΟΙΠΟ
  - Αντίστοιχη ΛΕΙΤΟΥΡΓΙΚΗ/ΔΟΜΙΚΗ σημασία
  - Εξελικτική σχέση
- ΕΙΣΑΓΩΓΗ ΚΕΝΩΝ
- ΥΠΟΘΕΤΙΚΟ ΜΟΝΤΕΛΟ ΜΟΡΙΑΚΗΣ ΕΞΕΛΙΞΗΣ (?)

ΑΝΤΙΚΕΙΜΕΝΙΚΟ ΚΡΙΤΗΡΙΟ ΠΟΙΟΤΗΤΑΣ?

ΒΗΜΑΤΑ ΠΟΥ ΑΠΑΙΤΟΥΝΤΑΙ

- ***ΣΥΛΛΟΓΗ*** ΠΙΘΑΝΩΝ ΟΜΟΛΟΓΩΝ ΑΚΟΛΟΥΘΙΩΝ
- ***ΥΠΟΛΟΓΙΣΜΟΣ*** ΠΟΛΛΑΠΛΗΣ ΣΤΟΙΧΙΣΗΣ
- ***ΕΛΕΓΧΟΣ*** ΚΑΙ ΠΙΘΑΝΗ ΤΡΟΠΟΠΟΙΗΣΗ (EDITING)
- ***ΠΑΡΟΥΣΙΑΣΗ*** ΑΠΟΤΕΛΕΣΜΑΤΩΝ

ΒΗΜΑΤΑ ΠΟΥ ΑΠΑΙΤΟΥΝΤΑΙ

- ***ΣΥΛΛΟΓΗ*** ΠΙΘΑΝΩΝ ΟΜΟΛΟΓΩΝ ΑΚΟΛΟΥΘΙΩΝ
  - ΠΕΙΡΑΜΑΤΑ
  - ΑΝΑΖΗΤΗΣΗ ΣΕ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ
  - ΒΙΒΛΙΟΓΑΦΙΑ
- ***ΥΠΟΛΟΓΙΣΜΟΣ*** ΠΟΛΛΑΠΛΗΣ ΣΤΟΙΧΙΣΗΣ
- ***ΕΛΕΓΧΟΣ*** ΚΑΙ ΠΙΘΑΝΗ ΤΡΟΠΟΠΟΙΗΣΗ (EDITING)
- ***ΠΑΡΟΥΣΙΑΣΗ*** ΑΠΟΤΕΛΕΣΜΑΤΩΝ

ΒΗΜΑΤΑ ΠΟΥ ΑΠΑΙΤΟΥΝΤΑΙ

- ***ΣΥΛΛΟΓΗ*** ΠΙΘΑΝΩΝ ΟΜΟΛΟΓΩΝ ΑΚΟΛΟΥΘΙΩΝ
- ***ΥΠΟΛΟΓΙΣΜΟΣ*** ΠΟΛΛΑΠΛΗΣ ΣΤΟΙΧΙΣΗΣ
  - ΕΠΙΛΟΓΗ ΚΑΤΑΛΛΗΛΗΣ ΜΕΘΟΔΟΛΟΓΙΑΣ ΚΑΙ ΠΑΡΑΜΕΤΡΩΝ
  - ΛΟΓΙΣΜΙΚΟ (WWW, ΤΟΠΙΚΟ)
- ***ΕΛΕΓΧΟΣ*** ΚΑΙ ΠΙΘΑΝΗ ΤΡΟΠΟΠΟΙΗΣΗ (EDITING)
- ***ΠΑΡΟΥΣΙΑΣΗ*** ΑΠΟΤΕΛΕΣΜΑΤΩΝ

ΒΗΜΑΤΑ ΠΟΥ ΑΠΑΙΤΟΥΝΤΑΙ

- ***ΣΥΛΛΟΓΗ*** ΠΙΘΑΝΩΝ ΟΜΟΛΟΓΩΝ ΑΚΟΛΟΥΘΙΩΝ
- ***ΥΠΟΛΟΓΙΣΜΟΣ*** ΠΟΛΛΑΠΛΗΣ ΣΤΟΙΧΙΣΗΣ
- ***ΕΛΕΓΧΟΣ*** ΚΑΙ ΠΙΘΑΝΗ ΤΡΟΠΟΠΟΙΗΣΗ (EDITING)
  - ΑΝΤΙΠΑΡΑΒΟΛΗ ΜΕ ΠΕΙΡΑΜΑΤΙΚΑ (ή ΑΛΛΑ ΥΠΟΛΟΓΙΣΤΙΚΑ ΔΕΔΟΜΕΝΑ)
    - ΜΕ ΤΟ ... ΜΑΤΙ !!!
    - ΑΥΤΟΜΑΤΟΠΟΙΗΜΕΝΑ
- ***ΠΑΡΟΥΣΙΑΣΗ*** ΑΠΟΤΕΛΕΣΜΑΤΩΝ

ΒΗΜΑΤΑ ΠΟΥ ΑΠΑΙΤΟΥΝΤΑΙ

- ***ΣΥΛΛΟΓΗ*** ΠΙΘΑΝΩΝ ΟΜΟΛΟΓΩΝ ΑΚΟΛΟΥΘΙΩΝ
- ***ΥΠΟΛΟΓΙΣΜΟΣ*** ΠΟΛΛΑΠΛΗΣ ΣΤΟΙΧΙΣΗΣ
- ***ΕΛΕΓΧΟΣ*** ΚΑΙ ΠΙΘΑΝΗ ΤΡΟΠΟΠΟΙΗΣΗ (EDITING)
- ***ΠΑΡΟΥΣΙΑΣΗ*** ΑΠΟΤΕΛΕΣΜΑΤΩΝ
  - ΕΞΟΔΟΣ ΛΟΓΙΣΜΙΚΟΥ
  - ΧΡΗΣΗ ΕΞΕΙΔΙΚΕΥΜΕΝΟΥ ΛΟΓΙΣΜΙΚΟΥ

Υπολογισμός (βέλτιστης?) MSA

- Τύποι Σύγκρισης
- Μέτρο Σύγκρισης
- Αντικειμενικότητα
- Σημαντικότητα

ΕΠΕΚΤΕΙΝΟΝΤΑΙ ΟΙ ΑΛΓΟΡΙΘΜΟΙ ΔΠ;

- ΝΑΙ
  - "Πολυδιάστατος" ΔΠ
- ΟΧΙ ΑΠΟΔΟΤΙΚΑ!!!
  - ΠΡΑΚΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ ΓΙΑ ΜΙΚΡΟ ΜΟΝΟ ΠΛΗΘΟΣ ΑΚΟΛΟΥΘΙΩΝ
  - ΓΙΑ n ΑΚΟΛΟΥΘΙΕΣ ΜΗΚΟΥΣ L1,L2,…,Ln:
    - ΜΝΗΜΗ O(L1*L2*…*Ln)
    - ΧΡΟΝΟΣ O(2n * L1*L2*…*Ln)

ΑΝΤΙΚΕΙΜΕΝΙΚΟ ΚΡΙΤΗΡΙΟ ΠΟΙΟΤΗΤΑΣ ??

SUM OF PAIRS (SP)

- ΧΡΗΣΗ ΠΙΝΑΚΑ ΑΝΤΙΚΑΤΑΣΤΑΣΗΣ (s)

- Για κάθε στήλη m_i:

$$SP(m_i) = \sum_{j=0}^{l-1} \sum_{j<k\le r} s(m_i^k, m_i^j)$$

Blosum50

s(L-L) = 5

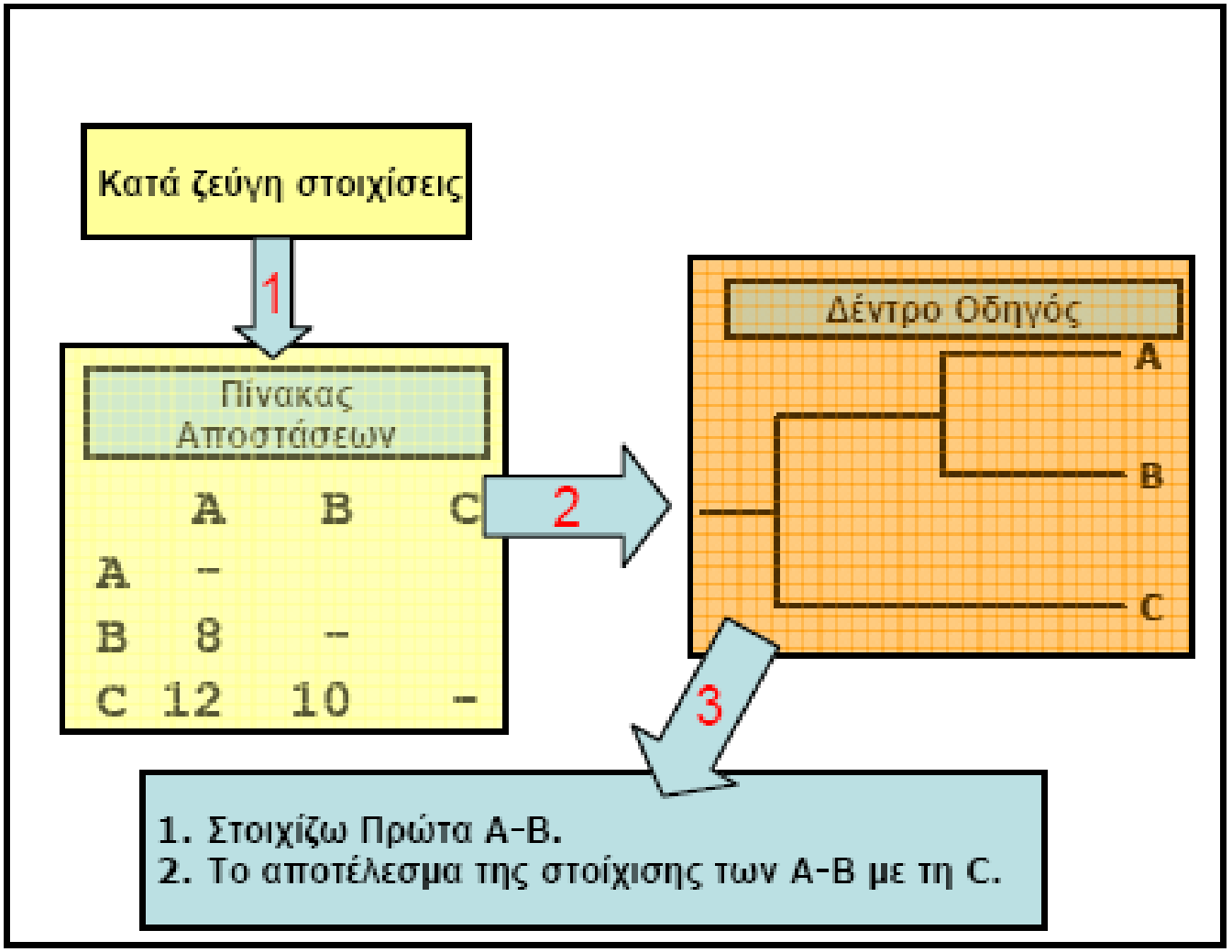
s(L-G) = -4

```
Seq1 :  ...ALLE...
Seq2 :  ...GLLD...
Seq3 :  ...WLGD...
```

SP(2)=15

SP(3)=-3

ΕΥΡΙΣΤΙΚΗ ΠΡΟΟΔΕΥΤΙΚΗ ΠΟΛΛΑΠΛΗ ΣΤΟΙΧΙΣΗ (Ι)



ΕΥΡΙΣΤΙΚΗ ΠΡΟΟΔΕΥΤΙΚΗ ΠΟΛΛΑΠΛΗ ΣΤΟΙΧΙΣΗ (II)

- Ο ΑΛΓΟΡΙΘΜΟΣ Feng-Doolittle (Feng and Doolittle, 1987)
  - ΔΙΑΓΩΝΙΟΣ ΠΙΝΑΚΑΣ ΑΠΟΣΤΑΣΕΩΝ (D= - log((Sobs-Srand)/(Smax-Srand))
  - ΔΕΝΤΡΟ-ΟΔΗΓΟΣ (ΔΟ) (αλγόριθμος Fitch-Margoliash, 1967)
  - ΠΡΟΟΔΕΥΤΙΚΗ ΣΤΟΙΧΙΣΗ ΒΑΣΕΙ ΔΟ
    - ΑΚΟΛΟΥΘΙΑ-ΑΚΟΛΟΥΘΙΑ (κλασσικός ΔΠ)
    - ΑΚΟΛΟΥΘΙΑ-ΣΤΟΙΧΙΣΗ (winner takes all)
    - ΣΤΟΙΧΙΣΗ-ΣΤΟΙΧΙΣΗ (winner takes all)

Φενγκ! Κάνε λίγο...

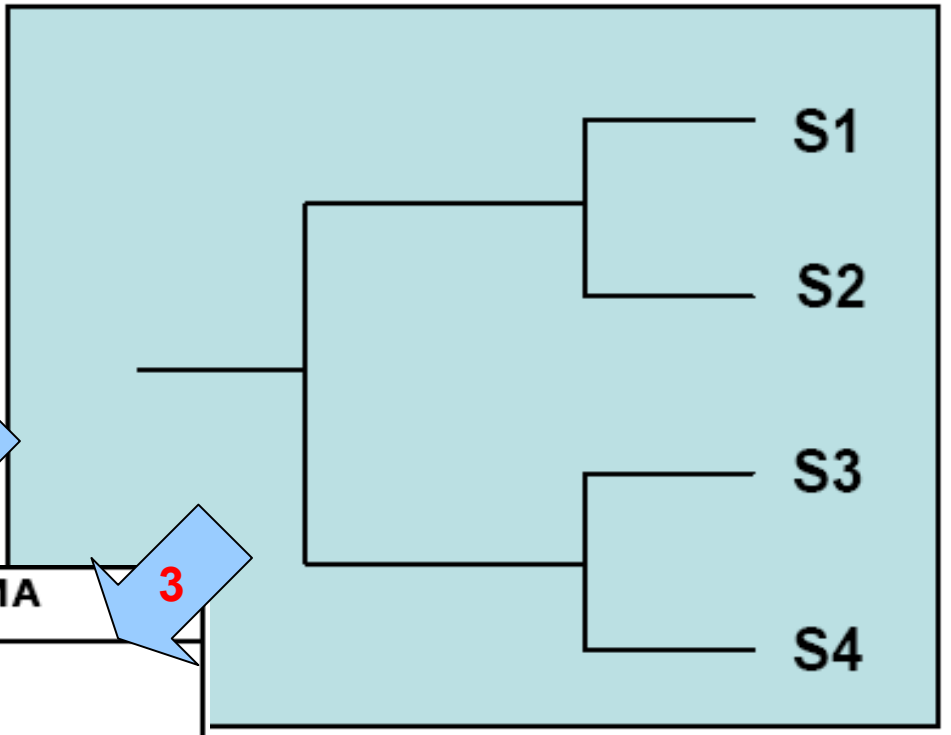
Έστω, οι ακολουθίες:

S1: AAATCGG, S2: AAACCGG, S3: ATACCCTG, S4: ATACCCGG

1

	S1	S2	S3	S4
S1	-			
S2	1	-		
S3	3	2	-	
S4	2	2	1	-

2

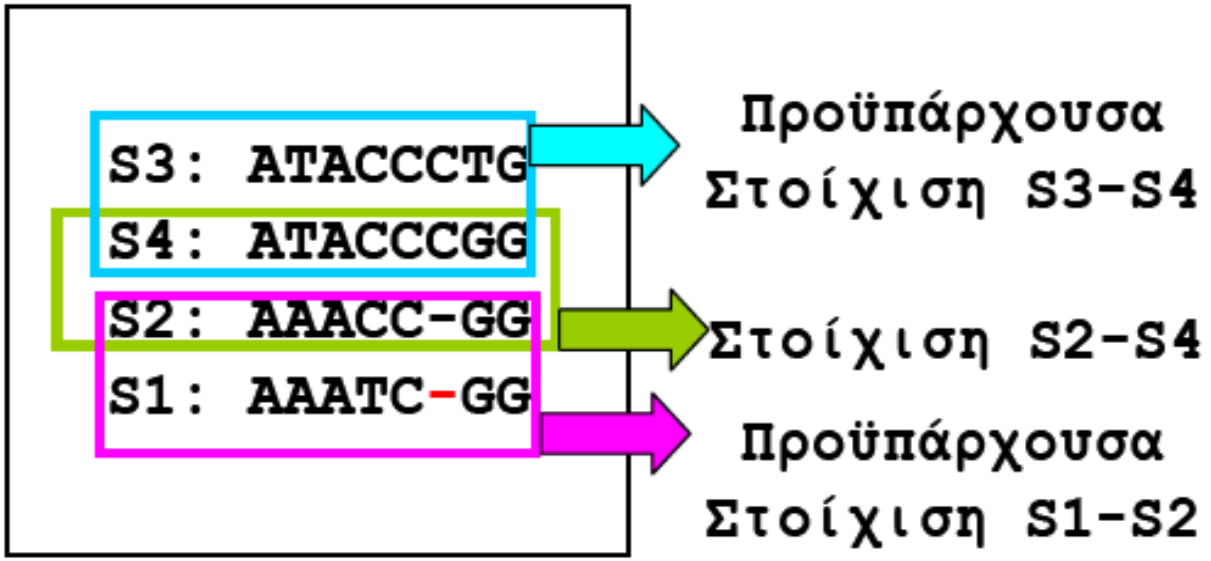


3

	ΟΔΗΓΙΑ	ΑΠΟΤΕΛΕΣΜΑ
1	Στοίχισε S1-S2 => $S_{1,2}$	AAATCGG ||| ||| AAACCGG
2	Στοίχισε S3-S4 => $S_{3,4}$	ATACCCTG |||||| | ATACCCGG
3	Στοίχισε $S_{1,2}$ – $S_{3,4}$	Δείτε παρακάτω ...

... και λίγο ακόμα ...

	Στοίχιση	Score
S1-S3	AAATC-GG | | | | | ATACCCTG	3
S1-S4	AAATCGG | | | | ATACCCG	4
S2-S3	AAACCG-G | ||| | ATACCCTG	4
S2-S4	AAACC-GG | ||| || ATACCCGG	5



Η Μέθοδος CLUSTAL

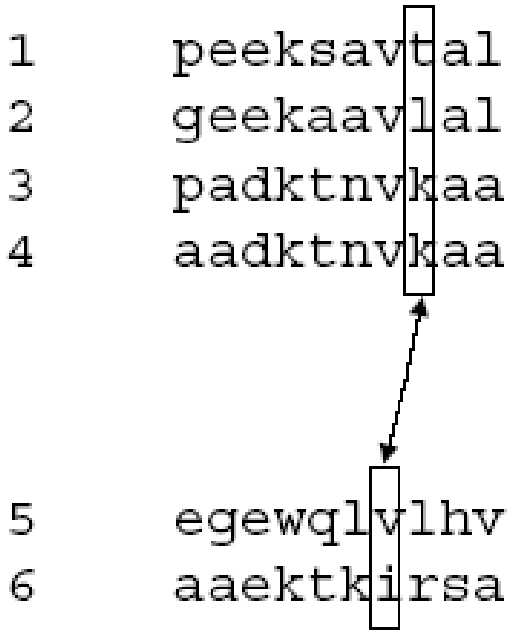
(Higgins and Sharp, 1988)

- ΔΙΑΓΩΝΙΟΣ ΠΙΝΑΚΑΣ ΑΠΟΣΤΑΣΕΩΝ
  (μετατροπή scores-απόσταση: αριθμός διαφορών ανά θέση)
- ΔΕΝΤΡΟ-ΟΔΗΓΟΣ (ΔΟ)
  (μέθοδος NJ, Saitou and Nei, 1987) => sequence weighting
- ΠΡΟΟΔΕΥΤΙΚΗ ΣΤΟΙΧΙΣΗ ΒΑΣΕΙ ΔΟ
  - ΑΚΟΛΟΥΘΙΑ-ΑΚΟΛΟΥΘΙΑ
  - ΑΚΟΛΟΥΘΙΑ-ΣΤΟΙΧΙΣΗ
  - ΣΤΟΙΧΙΣΗ-ΣΤΟΙΧΙΣΗ

CLUSTALW
ΒΑΣΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

- Profile Alignment
- Ποινές εισαγωγής κενών
  - “Once a gap, Always a gap”
  - GOP, GEP
  - Εξαρτώμενες από τη θέση
- Sequence weighting
  - Δείτε σε λίγο ...

PROFILE ALIGNMENT



Sequence weights
$w_1,\ldots,w_6$

Score: $\frac{1}{8}[M(t,v)w_1w_5 + M(t,i)w_1w_6 + \ldots + M(k,i)w_4w_6]$

ΔΠ και Profile Alignment

$S_{3,4}$

S3: ATACCCTG

S4: ATACCCGG

$S_{1,2}$

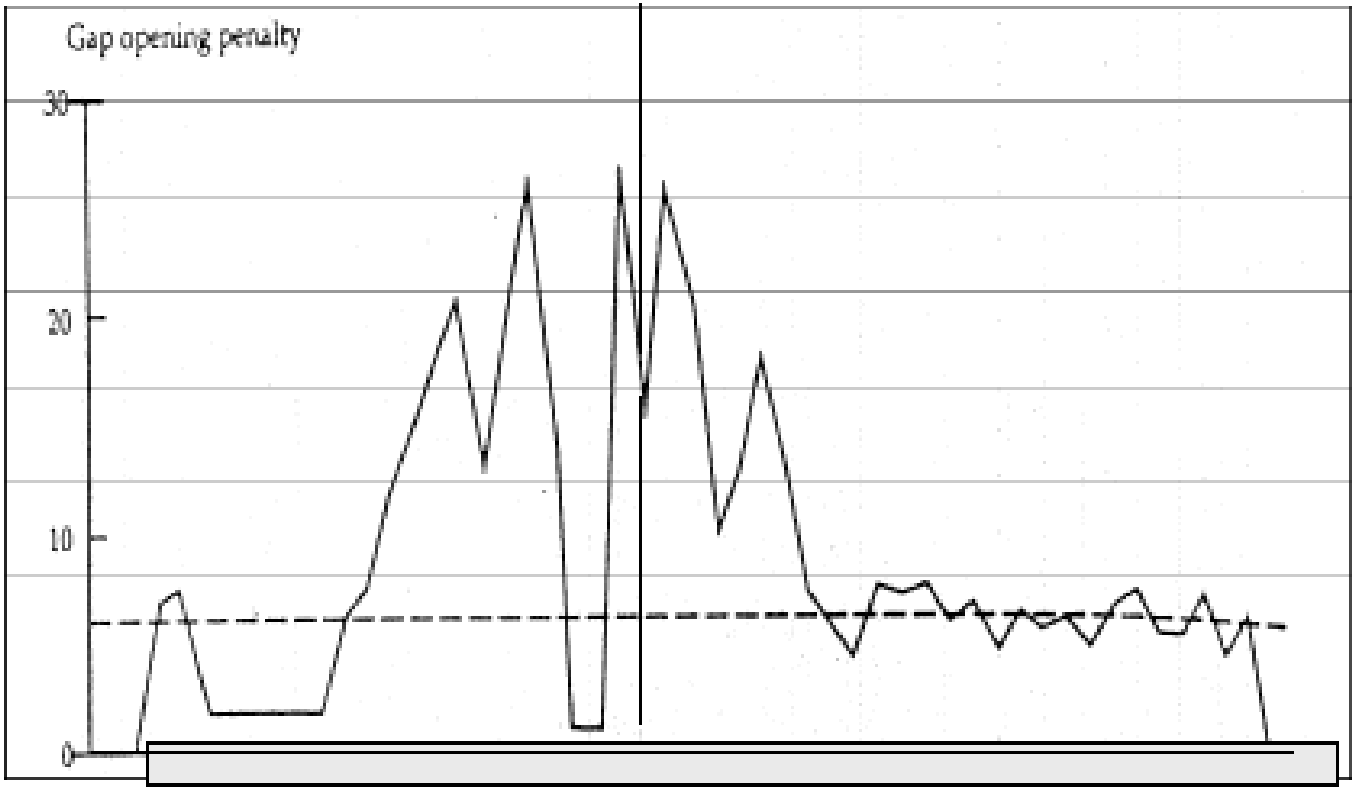
S2: AAACCGG

S1: AAATCGG

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

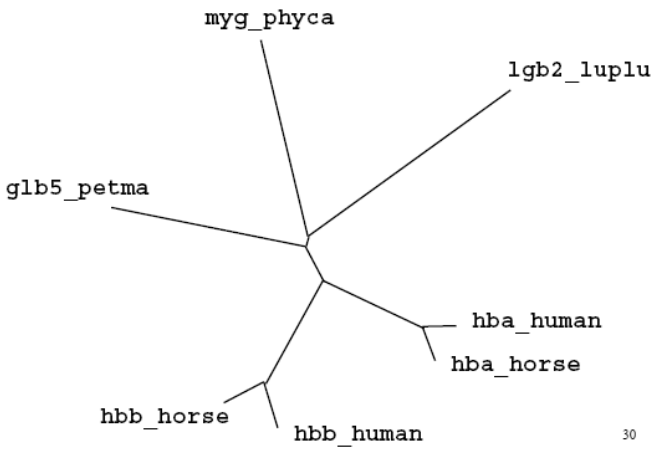
$$D = \max \begin{cases} A + SP(CT, TG) \\ B - g \\ C - g \end{cases}$$

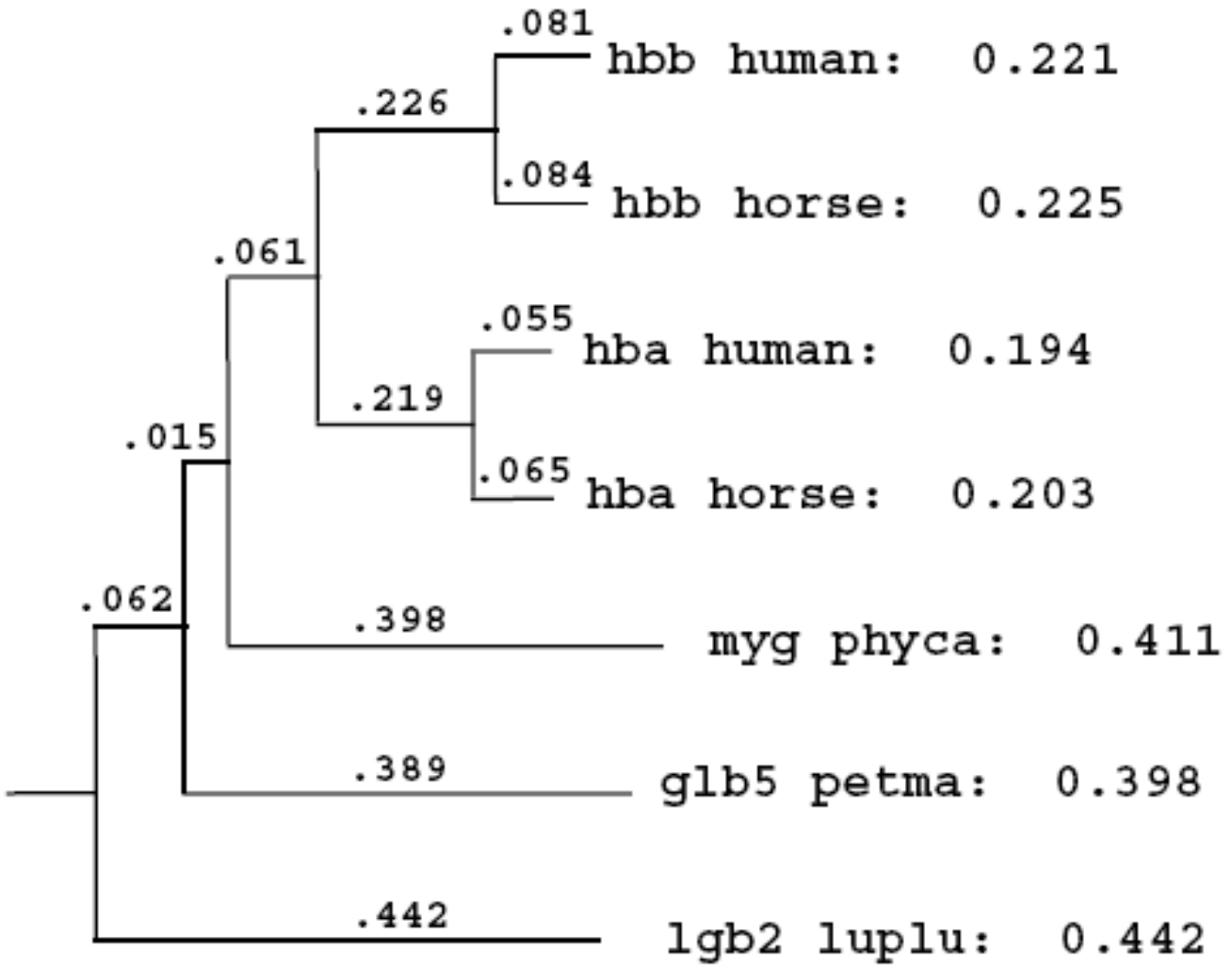
Ποινές Εισαγωγής Κενών εξαρτώμενες από τη θέση




```
HLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDL
QLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDL
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLS
```

Sequence Weighting





ΑΞΙΟΛΟΓΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΠΡΟΟΔΕΥΤΙΚΗΣ ΣΤΟΙΧΙΣΗΣ

+ ΑΠΟΔΟΤΙΚΟΙ

+ ΣΧΕΤΙΚΑ ΑΞΙΟΠΙΣΤΟΙ

- ΑΝΟΜΟΙΕΣ ΑΚΟΛΟΥΘΙΕΣ ;;;;

- GREEDY => "ΠΑΓΩΜΕΝΕΣ" ΕΠΙΜΕΡΟΥΣ ΣΤΟΙΧΙΣΕΙΣ

ΛΥΣΗ 1: ΒΕΛΤΙΩΜΕΝΟΙ ΑΛΓΟΡΙΘΜΟΙ ΠΡΟΟΔΕΥΤΙΚΗΣ ΣΤΟΙΧΙΣΗΣ (T-COFFEE, Notredame, Higgins and Herringa, 2000)

ΛΥΣΗ 2: ΕΠΑΝΑΛΗΠΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ (Barton and Sternberg, 1987; Berger and Manson, 1991; Gotoh, 1993)

ΛΥΣΗ 3: Πολλαπλή Στοίχιση με Profile HMMs

Συζήτηση ...

- Υλικό, κατά τα γνωστά, στην ιστοσελίδα του μαθήματος ...