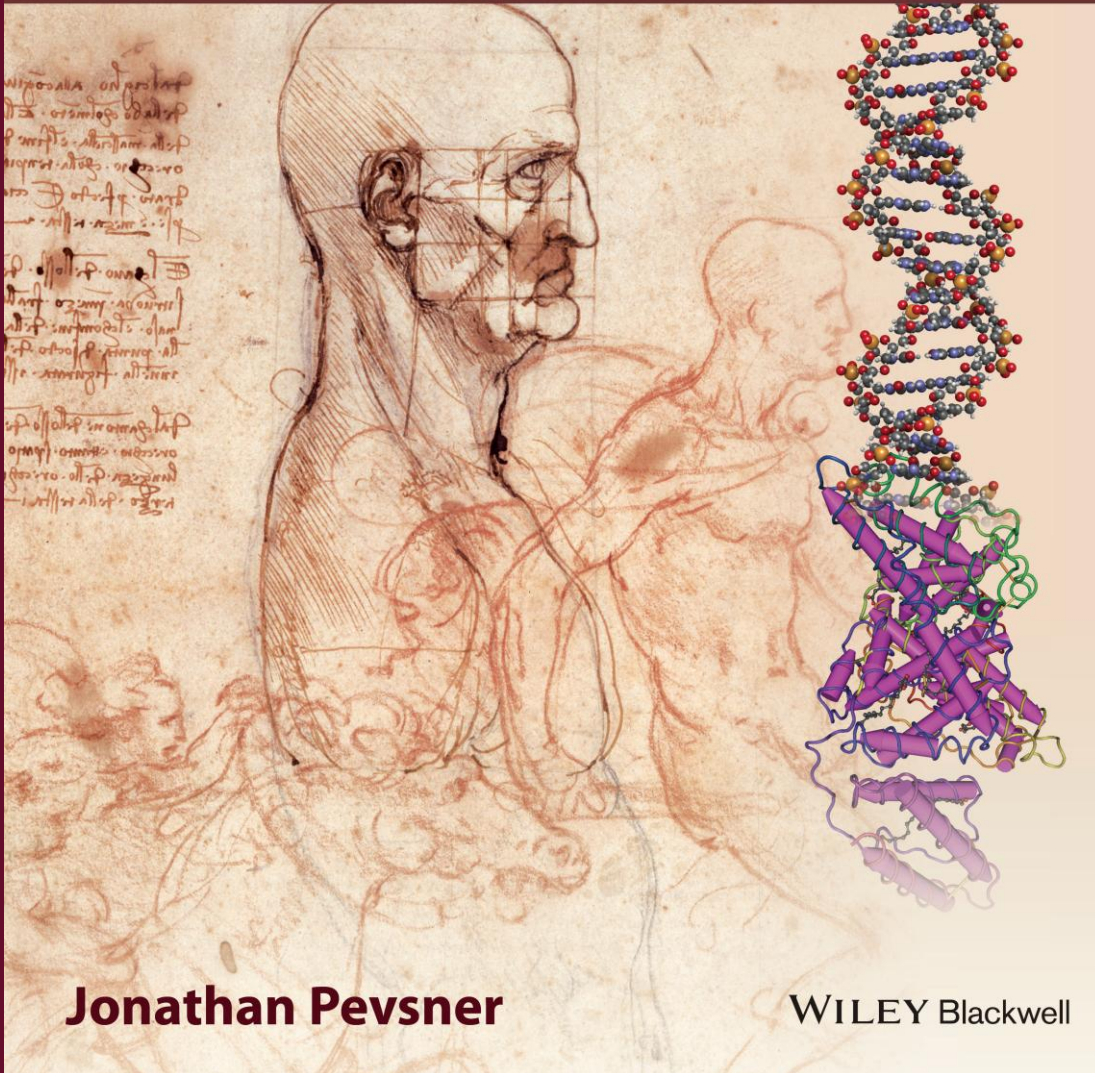


BIOINFORMATICS AND FUNCTIONAL GENOMICS

third edition



Jonathan Pevsner

WILEY Blackwell

Κεφάλαιο 6

Πολλαπλή στοίχιση αλληλουχιών

Ακαδημαϊκές
Εκδόσεις



STEP 1 - Enter your input sequences

Enter or paste a set of Protein sequences in any supported format:

```
>beta_globin 2hhbB NP_000509.1 [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDVAVMGNPVKVKAHGGKKVLGAFSDGLAHLDNLKGTFATLSLHCD
KLHVDPENFRLLGNVLCVLAHHEGKEFTPPVQAAAYQKVAVANALAHKYH
>myoglobin 2MM1 NP_005359.1 [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLRFLKGGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAOSHAT
KHKIPVKYLEFISECIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
>neuroglobin 1QJ6A NP_067080.1 [Homo sapiens]
MERPEPELURQSWRAVSRSPLEHGTVLFARLFALPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHR
```

Or, upload a file:

STEP 2 - Set your Pairwise Alignment Options

Alignment Type: ☒ Slow ☐ Fast

Slow Pairwise Alignment Options

Protein Weight Matrix	GAP OPEN	GAP EXTENSION
Gonnet	10	0.1

STEP 3 - Set your Multiple Sequence Alignment Options

Protein Weight Matrix	GAP OPEN	GAP EXTENSION	GAP DISTANCES	NO END GAPS
BLOSUM	10	0.20	5	no

ITERATION	NUMITER	CLUSTERING
none	1	NJ

Output Options

FORMAT	ORDER
Clustal w/ numbers	input

STEP 4 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

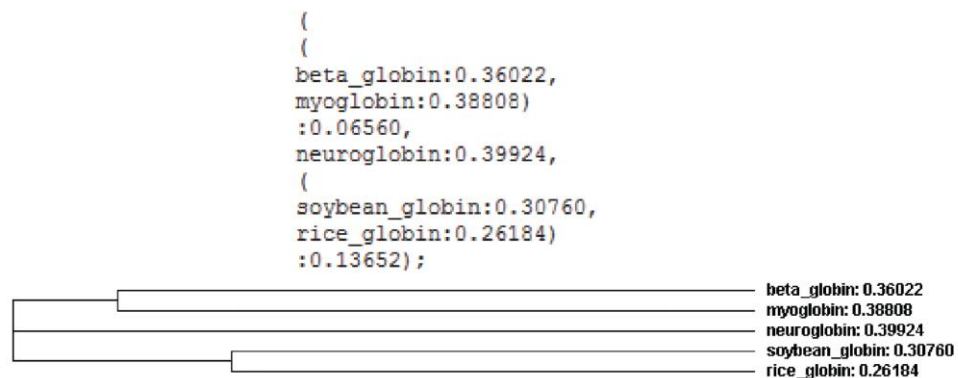
Εικόνα 6.1 Στοιχισή πέντε εξελικτικά απομακρυσμένων σφαιρινών με το ClustalW. Οι αλληλουχίες είναι στη μορφή FASTA.

(α) Βήμα 1: Κατά ζεύγη στοιχίσεις αλληλουχιών

SeqA	Name	Length	SeqB	Name	Length	Score
1	beta_globin	147	2	myoglobin	154	25.17
1	beta_globin	147	3	neuroglobin	151	15.65
1	beta_globin	147	4	soybean_globin	144	13.19
1	beta_globin	147	5	rice_globin	166	21.09
2	myoglobin	154	3	neuroglobin	151	16.56
2	myoglobin	154	4	soybean_globin	144	8.33
2	myoglobin	154	5	rice_globin	166	12.99
3	neuroglobin	151	4	soybean_globin	144	17.36
3	neuroglobin	151	5	rice_globin	166	18.54
4	soybean_globin	144	5	rice_globin	166	43.06

1

(β) Βήμα 2: Δημιουργία ενός δενδρογράμματος-οδηγού (υπολογισμένου από τον πίνακα αποστάσεων)



Εικόνα 6.2 Η μέθοδος προοδευτικής στοίχισης των Feng και Doolittle (1987) που χρησιμοποιείται από πολλά προγράμματα στοιχίσεων, π.χ. από το ClustalW. Στο πρώτο στάδιο γίνονται όλες οι πιθανές κατά ζεύγη στοιχίσεις των πέντε σφαιρινών που εξετάζουμε σε αυτό το παράδειγμα (Εικόνα 6.1). Σημειώστε ότι η καλύτερη βαθμολογία αντιστοιχεί στη στοίχιση δύο φυτικών σφαιρινών (βαθμολογία = 43, βέλος 1). Στο δεύτερο στάδιο, και με βάση τις βαθμολογίες από το προηγούμενο βήμα, υπολογίζεται ένα δενδρογράμμα-οδηγός το οποίο περιγράφει τις σχέσεις μεταξύ των πέντε αλληλουχιών. Μπορείτε να δείτε μια γραφική παράσταση του δενδρογράμματος μέσω του προγράμματος JalView από την ιστοσελίδα του ClustalW. Τα μήκη των κλάδων αντανakλούν τις αποστάσεις μεταξύ των αλληλουχιών και υποδεικνύονται στο δέντρο. Συγκρίνετε αυτή τη γραφική παράσταση με την Εικόνα 6.4.

Πολλαπλή στοίχιση αλληλουχιών με τον αλγόριθμο CLUSTAL 2.1

```

beta_globin      -----MVHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin       -----MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGHPETLEKFDKFK- 48
neuroglobin     -----MERPEPELIRQSWRAVSRSPLEHGTVLFLARLFALEPDLLPLFQYNCR 47
soybean_globin  -----MVAFTKQDALVSSSFEEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice_globin     MALVEDNNAVAVSFS EEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSLR- 59
                  :   :   :   :   .   .   .   :   *   *
beta_globin      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAT-----LSELHCDKLHVDP 101
myoglobin       HLKSEDEMKASEDLKKGATVLTALGGTLKKGHHEAEIKP-----LAQSHATKHKIPV 102
neuroglobin     QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEY---LASLGRKHRAVGVKLS 104
soybean_globin  --NGVDPT--NPKLTGHA EKLFALVRDSAGQLKASGTVVAD---AALGSVHAQKAVTDP 101
rice_globin     --NSDVPLEKNPKLKT HAMSVFVMTCEAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
                  .   .   .   *   .   :   :   :   :   :   *   *
beta_globin      ENFRLLGNVLVCVLAHHFGKEFTPPVQAAVQKVVAGVANALAHKYH----- 147
myoglobin       KYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin     SFSTVGESLLYMLEKCLG-PAFTPATRAAWSQLYGAVVQAMSRGWDGE--- 151
soybean_globin  QFVVVKEALLKTIKAAVG--DKWSELSRAWEVAYDELA AAIKKA----- 144
rice_globin     HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166

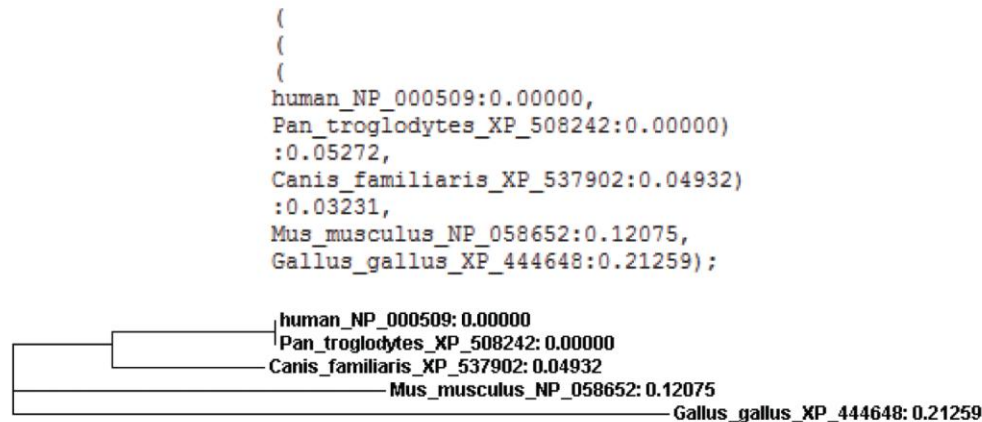
```

Εικόνα 6.3 Στοίχιση πέντε εξελικτικά απομακρυσμένων σφαιρινών με το πρόγραμμα ClustalW χρησιμοποιώντας τον αλγόριθμο προοδευτικής στοίχισης των Feng και Doolittle. Στο τρίτο στάδιο (βλ. Εικόνα 6.2 για τα πρώτα δύο στάδια) δημιουργείται αυτή καθαυτή η στοίχιση όλων των αλληλουχιών. Οι δύο πιο συγγενικές αλληλουχίες στοιχίζονται πρώτες (σε αυτό το παράδειγμα οι σφαιρίνες από τη σόγια και το ρύζι) και στη συνέχεια προστίθενται σταδιακά και οι υπόλοιπες. Η σειρά με την οποία προστίθενται βασίζεται στη θέση τους στο δενδρόγραμμα-οδηγό. Στη στοίχιση που βλέπετε εδώ, τα απόλυτα συντηρημένα κατάλοιπα υποδεικνύονται με αστερίσκο, οι συντηρητικές αντικαταστάσεις με άνω και κάτω τελεία και οι λιγότερο συντηρητικές αντικαταστάσεις με τελεία. Οι πρωτεΐνες είναι η ανθρώπινη β-σφαιρίνη (καταχώριση NP_000509, αναγνωριστικό 2hhb στην Protein Data Bank), η ανθρώπινη μυοσφαιρίνη (NP_005359, 3RGK), η ανθρώπινη νευροσφαιρίνη (NP_067080, 1OJ6A), η λεγκαιμοσφαιρίνη της σόγιας (1FSL) και η μη συμβιωτική αιμοσφαιρίνη του ρυζιού (1D8U). Με κόκκινα γράμματα επισημαίνονται οι περιοχές των α-ελίκων (που ορίζονται στο Κεφάλαιο 13) όπως ταυτοποιήθηκαν με κρυσταλλογραφία ακτίνων Χ. Τρία εξαιρετικά συντηρημένα κατάλοιπα υποδεικνύονται με κεφαλές βέλους και έντονα μπλε γράμματα (είναι τα αμινοξέα Phe44, His65 και His94 σύμφωνα με την αρίθμηση της μυοσφαιρίνης). Αυτές οι δύο ιστιδίνες είναι σημαντικές για την πρόσδεση της ομάδας της αίμης. Το τμήμα της στοίχισης που περιβάλλεται από το πράσινο πλαίσιο θα το ξαναδούμε όταν θα συγκρίνουμε τη στοίχιση από το ClustalW με τις στοιχίσεις από άλλα προγράμματα στοίχισης (Εικόνα 6.7)

(α) Βήμα 1: Κατά ζεύγη στοίχισεις αλληλουχιών (στενά συγγενικών σφαιρινών)

SeqA	Name	Length	SeqB	Name	Length	Score
1	human_NP_000509	147	2	Pan_troglodytes_XP_508242	147	100.0
1	human_NP_000509	147	3	Canis_familiaris_XP_537902	147	89.8
1	human_NP_000509	147	4	Mus_musculus_NP_058652	147	80.27
1	human_NP_000509	147	5	Gallus_gallus_XP_444648	147	69.39
2	Pan_troglodytes_XP_508242	147	3	Canis_familiaris_XP_537902	147	89.8
2	Pan_troglodytes_XP_508242	147	4	Mus_musculus_NP_058652	147	80.27
2	Pan_troglodytes_XP_508242	147	5	Gallus_gallus_XP_444648	147	69.39
3	Canis_familiaris_XP_537902	147	4	Mus_musculus_NP_058652	147	78.91
3	Canis_familiaris_XP_537902	147	5	Gallus_gallus_XP_444648	147	71.43
4	Mus_musculus_NP_058652	147	5	Gallus_gallus_XP_444648	147	66.67

(β) Βήμα 2: Δημιουργία δενδρογράμματος-οδηγού (υπολογισμένου από τον πίνακα αποστάσεων)



Εικόνα 6.4. Παράδειγμα στοίχισης πέντε σφαιρινών με υψηλή εξελικτική συγγένεια χρησιμοποιώντας τη μέθοδο προοδευτικής στοίχισης των Feng και Doolittle όπως εφαρμόζεται από το πρόγραμμα ClustalW. Συγκρίνετε αυτές τις βαθμολογίες με εκείνες των εξελικτικά απομακρυσμένων σφαιρινών (Εικόνα 6.2). Παρατηρήστε ότι οι βαθμολογίες των κατά ζεύγη στοίχισεων είναι σταθερά υψηλότερες και οι μεταξύ τους αποστάσεις (που αντανακλώνται στα μήκη των κλάδων του δέντρου) είναι πολύ μικρότερες.

Πολλαπλή στοίχιση αλληλουχιών με τον αλγόριθμο CLUSTAL 2.1

```

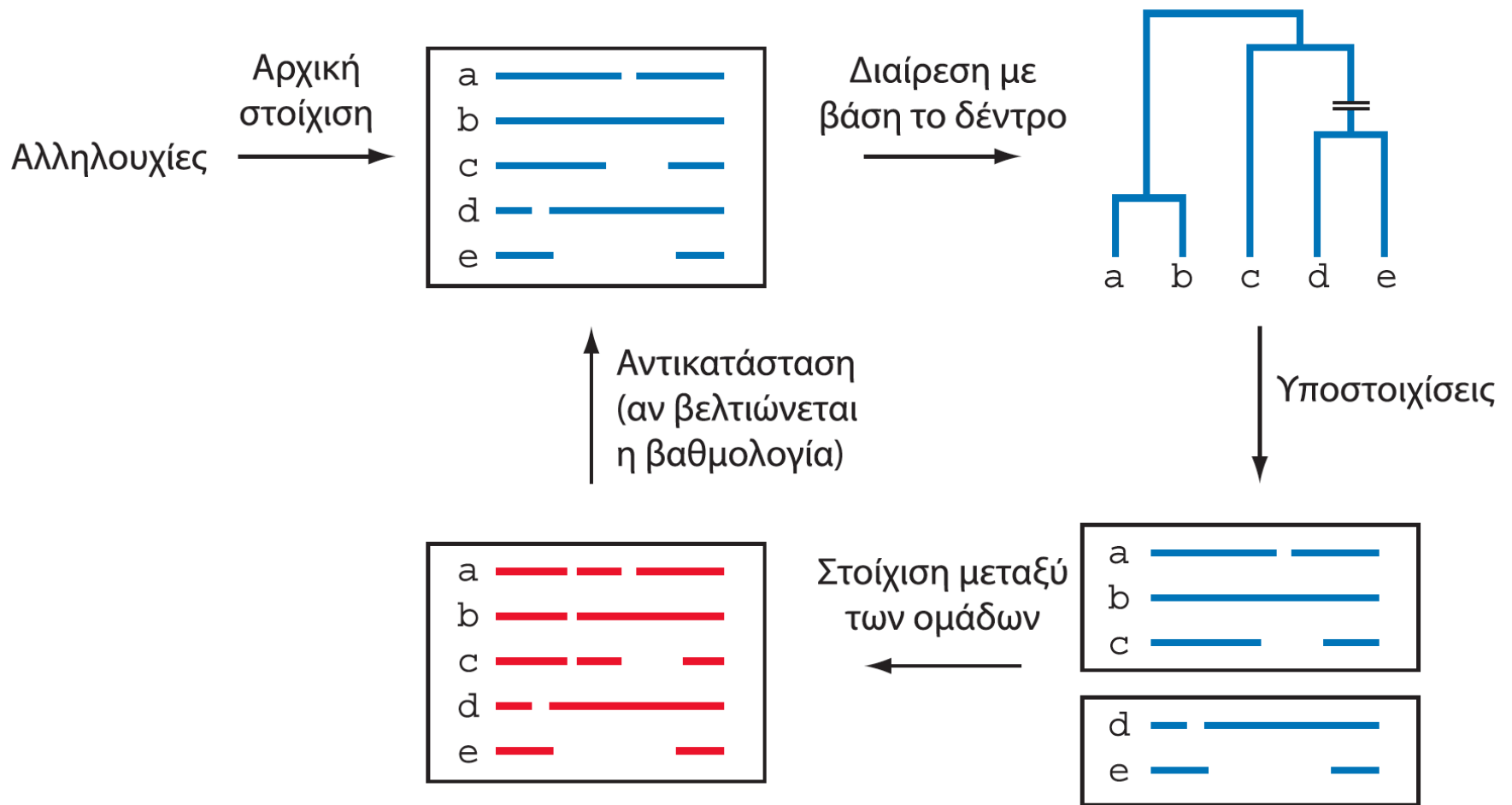
human_NP_000509          MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS 50
Pan_troglodytes_XP_508242 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS 50
Canis_familiaris_XP_537902 MVHLTAE EKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLS 50
Mus_musculus_NP_058652    MVHLTDAEKSAVSCLWAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLS 50
Gallus_gallus_XP_444648   MVHWTAE EKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLS 50
*** *  **  :: **.* **  * *.***.***:*****:* ***: **

human_NP_000509          TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD 100
Pan_troglodytes_XP_508242 TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD 100
Canis_familiaris_XP_537902 TPDAVMSNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVD 100
Mus_musculus_NP_058652    SASAIMGNPKVKAHGKKVITAFNEGLKNLDNLKGTFASLSELHCDKLHVD 100
Gallus_gallus_XP_444648   SPTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVD 100
.: *::*. *::*****: :*::: :****:*.***: *****

human_NP_000509          PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
Pan_troglodytes_XP_508242 PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
Canis_familiaris_XP_537902 PENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALAHKYH 147
Mus_musculus_NP_058652    PENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVAGVATALAHKYH 147
Gallus_gallus_XP_444648   PENFRLLGDILIIVLAAHFSKDFTECQAAWQKLVRVVAHALARKYH 147
****:****: :: **  *:.* **  * *.***.***:*****:* ***: **

```

Εικόνα 6.5. Στοίχιση των πέντε στενά συγγενικών β-σφαιρινών (βλ. Εικόνα 6.4). Η εικόνα που βλέπετε προέρχεται απευθείας από το πρόγραμμα ClustalW. Τα βέλη αντιστοιχούν στα κατάλοιπα Phe44, His72 και His104 (με την αρίθμηση της ανθρώπινης β-σφαιρίνης) και το πράσινο πλαίσιο δείχνει την ίδια περιοχή με την Εικόνα 6.3. Ο χρωματικός κώδικας είναι επιλογή του προγράμματος και δείχνει ιδιότητες καταλοίπων, όπως όξινα αμινοξέα (μπλε), βασικά αμινοξέα (μοβ) και υδρόφοβα κατάλοιπα (κόκκινο). Οι αστερίσκοι δείχνουν τις θέσεις της στοίχισης που είναι απόλυτα συντηρημένες.



Εικόνα 6.6 Η μέθοδος επαναληπτικής βελτιστοποίησης που χρησιμοποιείται από το πρόγραμμα MAFFT. Η μέθοδος βασίζεται στην επαναληπτική επαναστοίχιση ομάδων αλληλουχιών, με τις ομάδες να προκύπτουν από τη διαίρεση του δενδρογράμματος-οδηγού. Αν μετά από έναν κύκλο επαναστοίχισης η βαθμολογία της νέας στοίχισης είναι καλύτερη από την αρχική, τότε αυτή η νέα στοίχιση αντικαθιστά την αρχική και η διαδικασία επαναλαμβάνεται μέχρι συγκλίσεως.

(α) Στοίχιση εννέα σφαιρινών με το πρόγραμμα MAFFT FFT-NS-2 (v7.058b) (DSSP χρωματικός κώδικας: **στροφή**, **α-έλικα**, **βρόχοι**, 3₁₀-έλικα)

hbb_human	-----MVHLT PEEK SAVTAL WG VNVD-- EV GGEALGRLL VV YPWTQRFFE-SFG	▼1
hbb_chimp	-----MVHLT PEEK SAVTAL WG VNVD-- EV GGEALGRLL VV YPWTQRFFE-SFG	
hbb_dog	-----MVHLT AEEK SLV SGL WG VNVD-- EV GGEALGRLL IV YPWTQRFFD-SFG	
hbb_mouse	-----MVHLT DAEK SAV SCL WAK VN PD -- EV GGEALGRLL VV YPWTQRYFD-SFG	
hbb_chicken	-----MVHWT AEEK QLIT GL WG VN VA -- EC GA EAL ARLL IV YPWTQRFFA-SFG	
myoglobin	-----MGL S DGE W QLVLNV WG KVEADIPGHGQEV LIR L FK GH PET LEKFD-K FK	
neuroglobin	-----MER PE ELIRQ SW RAV SR S PLEHGTVL FAR L FA LEPDLLPL FQ YN CR	
soybean	-----MVA FT E KQ DALV SS S FE AFKAN I PQYSVV FY TS IL E K AP AA KDL S -FL A	
rice	MAL V ED NN AVAV SF SE EQ EALVLK S WAIL KK DSANIALR FF L KI FEV AP SAS Q M S -FL R	
	: : : .. . :: * *	

hbb_human	DL ST PD AV MGN PK VKAHGKKVLGAFSDGLAH---LDNL---KGTFATLSELHCDKLHVDP	▼2	▼3
hbb_chimp	DLSTPD AV MGN PK VKAHGKKVLGAFSDGLAH---LDNL---KGTFATLSELHCDKLHVDP		
hbb_dog	DL ST PD AV MSNA KV KAHGKKVLNSFS DGL KN---LDNL---KGTF AK LSELHCDKLHVDP		
hbb_mouse	DL SS ASAIMGN PK VKAHGKKVITAFNEGL KN ---LDNL---KGTFASLSELHCDKLHVDP		
hbb_chicken	NL SS PTAILGNPMVRAHGKKVLTSFGDAV KN ---LDNI---KNTFSQ L SELHCDKLHVDP		
myoglobin	HL K SEDEM KAS EDLKKHGATVLTALGGIL KK ---KGHH---EAEIKPLAQSHATKHIPV		
neuroglobin	Q FSS PE DC L SS PEFLDHIRKVMLVIDAAVT N VEDLSSL---EEYLASLGRKH-RAVG VKL		
soybean	NGVDP---TN PK L TG HA EKL FALVRDSAGQLKAS GT V-VADAA---LGS VH -AQKAV TD		
rice	NSD VP --LEKN PK L KT HAMS V FVMTCEAA AQL RKA GV TVRDTTLKRLGATH-LKYGV GD		
	. . . * .:: : . . * . *		

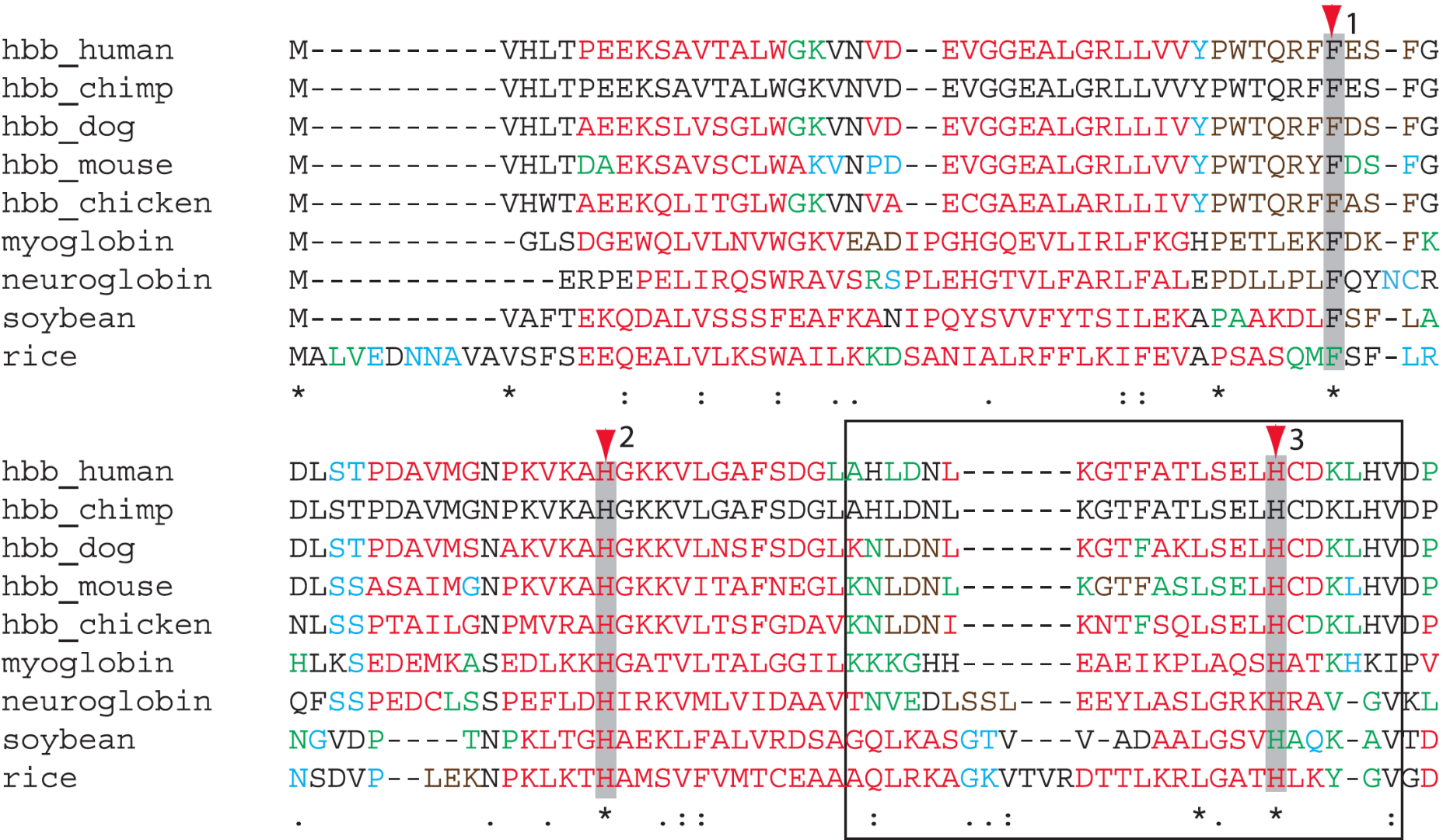
Εικόνα 6.7 Στοίχιση των αλληλουχιών εννέα σφαιρινών χρησιμοποιώντας τα προγράμματα: (α) MAFFT, (β) MUSCLE, (γ) ProbCons και (δ) T-COFFEE. Ο χρωματικός κώδικας αντιστοιχεί στα στοιχεία δευτεροταγούς δομής των αντίστοιχων δομών σύμφωνα με το πρόγραμμα DSSP (Kabsch and Sander, 1983) και τις καταχωρίσεις της Protein Data Bank (Κεφάλαιο 13) ως εξής: α-έλικες (κόκκινο), 3₁₀-έλικα (καφέ), στροφές και βρόχοι (κυανό και πράσινο αντίστοιχα), χωρίς απόδοση δευτεροταγούς δομής (μαύρο). Η τρισδιάστατη δομή που αντιστοιχεί στη δεύτερη αλληλουχία (hbb_chimp) δεν είναι γνωστή, γεγονός που ερμηνεύει την απουσία χρωμάτων. Τα βέλη αντιστοιχούν στα κατάλοιπα Phe44, His72 και His104 (με την αρίθμηση της ανθρώπινης β-σφαιρίνης όπως και στην Εικόνα 6.5).

(β) Στοιχισή εννέα σφαιρινών με το πρόγραμμα MUSCLE (3.8)

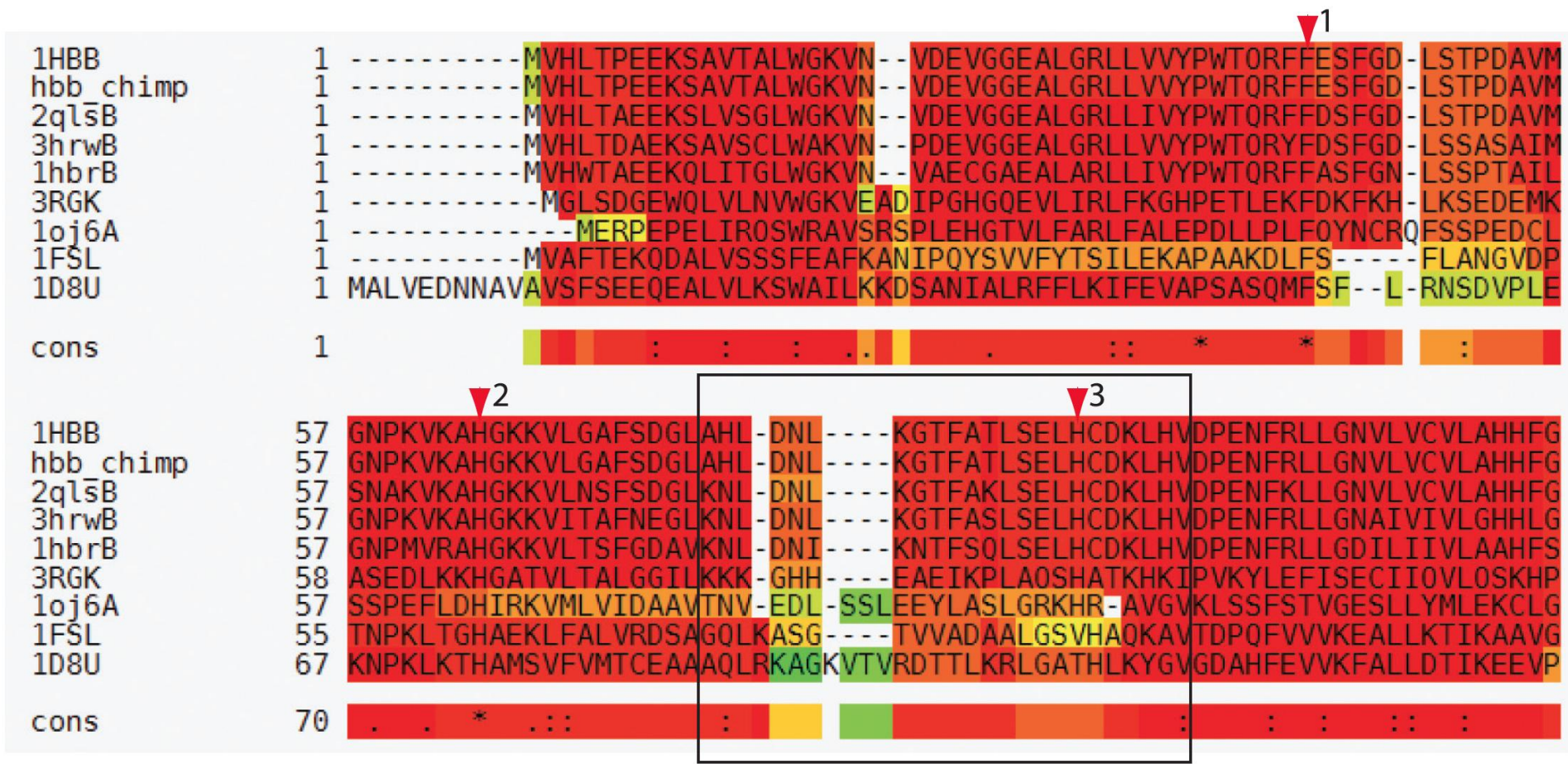
hbb_human	-----MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLLVVYPWTQRRFFE-SFG	▼1
hbb_chimp	-----MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLLVVYPWTQRRFFE-SFG	
hbb_dog	-----MVHLTAEKSLVSGLWGKVNVD--EVGGEALGRLLIVYPWTQRRFFD-SFG	
hbb_mouse	-----MVHLTDAEKSAVSCSLWAKVNPDP--EVGGEALGRLLLVVYPWTQRYFD-SFG	
hbb_chicken	-----MVHWTAEKQLITGLWGKVNVA--ECGAEALARLLIVYPWTQRRFFA-SFG	
myoglobin	-----MGLSDGEWQLVLNVWGKVEADIPGHGQEV LIRLFKGH PETLEKFD-KFK	
neuroglobin	-----MERPEPELIRQSWRAVSRSPLEHGT VLFARLFALEPDLLPLFQYNCR	
soybean	-----MVAFTTEKQDALVSSSF EAFKANIPQYSVV FYTSILEKAPAAKDLFS-FLA	
rice	MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKI FEVAPSASQMFS-FLR	
	: : : .. . :: *	
hbb_human	DLSTPDAVMGNPKVKAHGKKVLGAFSDG LAHL---DNLKGT FATLSELHCDK--LHVDPE	▼2
hbb_chimp	DLSTPDAVMGNPKVKAHGKKVLGAFSDG LAHL---DNLKGT FATLSELHCDK--LHVDPE	
hbb_dog	DLSTPDAVMSNAKVKAHGKKVLNSFS DGLKNL---DNLKGT FAKLSELHCDK--LHVDPE	
hbb_mouse	DLSSASAIMGNPKVKAHGKKVITAFNEGLKNL---DNLKGT FASLSELHCDK--LHVDPE	
hbb_chicken	NLSSPTAILGNPMVRAHGKKVLTSFGDAVKNL---DNIKNTFSQLSELHCDK--LHVDPE	
myoglobin	HLKSEDEMKASEDLKKHGATVLTALGGILKKK---GHHEAEIKPLAQSHATK--HKIPVK	▼3
neuroglobin	QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNV---EDLSSLEEYLASLGRKHRAVGVKLS	
soybean	NGVDPT---NPKLTGHA EKLFALVRDSAGQL---KASGT VVADAALG SVHAQKAVTDP	
rice	NSDVP--LEKNPKLKT HAMS VFVMTCEAA AQLRKAGKVTVRDTTLKRLGATHLKYGVGDA	
	. . * .:: :	

Εικόνα 6.7. Στοιχισμός των αλληλουχιών των εννέα σφαιρινών με το πρόγραμμα MUSCLE (3.8). Τα στοιχεία που είναι σκιασμένα (σκιασμένες στήλες) και στον αριθμό και στη θέση των κενών (βλ. περιοχές με πλαίσιο). Οι πρωτεΐνες που χρησιμοποιήθηκαν σε αυτές τις στοιχίσεις (διαθέσιμες μέσω του Web Document 6.3) μαζί με τους κωδικούς τους RefSeq και PDB είναι οι ακόλουθες: (1) hbb_human (human NP_000509.1, 1HBB), (2) hbb_chimp (Pan_troglodytes XP_508242.1, χωρίς δομή), (3) hbb_dog (Canis lupus familiaris NP_001257813,1, 2QLS), (4) hbb_mouse (Mus_musculus NP_058652.1, 3HRW), (5) hbb_chicken (Gallus_gallus NP_990820.1, 1HBR), (6) μυοσφαιρίνη (myoglobin, human NP_005359.1, 3RGK), (7) νευροσφαιρίνη (neuroglobin, human NP_067080.1, 10J6), (8) σφαιρίνη της σόγιας (soybean, Glycine max leghemoglobin A, NP_001235928,1, 1FSL) και (9) σφαιρίνη του ρυζιού (rice, Oryza sativa NonSymbiotic Plant Hemoglobin NP_001049476.1, NP_001049476.1, 1D8U). Στην εικόνα εμφανίζονται μόνο τα πρώτα δύο τρίτα κάθε στοιχίσης.

(γ) Στοίχιση εννέα σφαιρινών με το πρόγραμμα ProbCons (έκδοση 1.12)



(δ) Στοιχισή εννέα σφαιρινών με το πρόγραμμα T-COFFEE (Expresso έκδοση 10.00)



Family: *Globin* (PF00042)


 34
architectures


 6000
sequences


 5 interactions


 2886 species


 1971 structures

Summary

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & model

Species

Interactions

Structures

Jump to...

Alignments

We store a range of different sequence alignments for families. As well as the seed alignment from which the family is built, we provide the full alignment, generated by searching the sequence database using the family HMM. We also generate alignments using four [representative proteomes](#)¹ (RP) sets, the NCBI sequence database, and our metagenomics sequence database. [More...](#)

View options

We make a range of alignments for each Pfam-A family. You can see a description of each [above](#). You can view these alignments in various ways but please note that some types of alignment are never generated while others may not be available for all families, most commonly because the alignments are too large to handle.

	Seed (73)	Full (6000)	Representative proteomes				NCBI (5331)	Meta (34)
			RP15 (348)	RP35 (594)	RP55 (949)	RP75 (1261)		
Jalview	✓	✓	✓	✓	✓	✓	✓	✓
HTML	✓	—	✓	✓	✓	✓	×	×
PP/heatmap	X ₁	—	✓	✓	✓	✓	×	×
Pfam viewer	✓	✓	×	×	×	×	×	×

¹Cannot generate PP/Heatmap alignments for seeds; no PP data available

Key: ✓ available, × not generated, — not available.

Εικόνα 6.8 Η βάση δεδομένων Pfam είναι μια πλήρης και έγκυρη πηγή για τη μελέτη των οικογενειών πρωτεϊνών. (α) Μια τυπική καταχώριση από την Pfam, στο συγκεκριμένο παράδειγμα για τις σφαιρίνες. Στην κορυφή της σελίδας υπάρχουν σύνδεσμοι για την αρχιτεκτονική της οικογένειας (δηλαδή την οργάνωση επικρατειών), τις αλληλουχίες, τις αλληλεπιδράσεις, τα είδη και τις δομές. Αριστερά υπάρχουν σύνδεσμοι για τις στοιχίσεις και άλλες πληροφορίες. Αυτές οι στοιχίσεις μπορεί να είναι οι στοιχίσεις εκκίνησης (seed alignments, αποτελούμενες σε αυτή την περίπτωση από 73 αντιπροσωπευτικές σφαιρίνες), οι πλήρεις στοιχίσεις ή αντιπροσωπευτικά σύνολα πρωτεϊνών. Κάνοντας κλικ στα διάφορα κελιά αυτού του πίνακα, μπορούμε να εξετάσουμε τις αντίστοιχες στοιχίσεις.

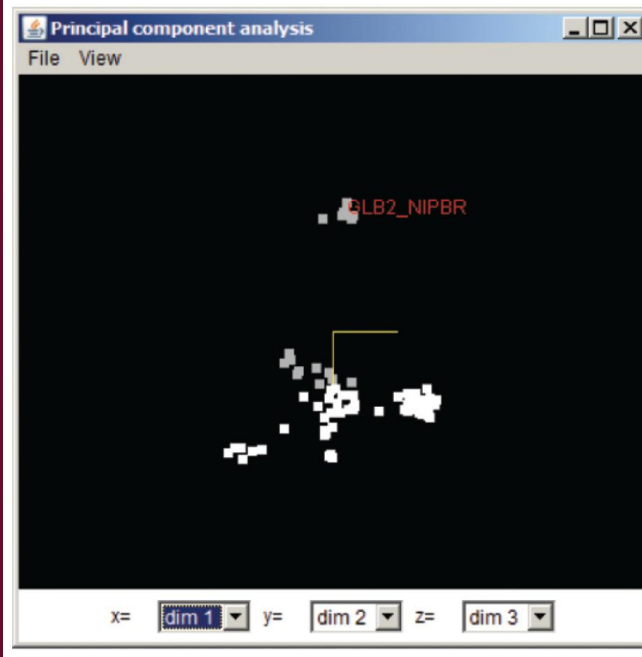
(β) Τμήμα μιας στοίχισης εκκίνησης από την Pfam

Seed sequence alignment for PF00042

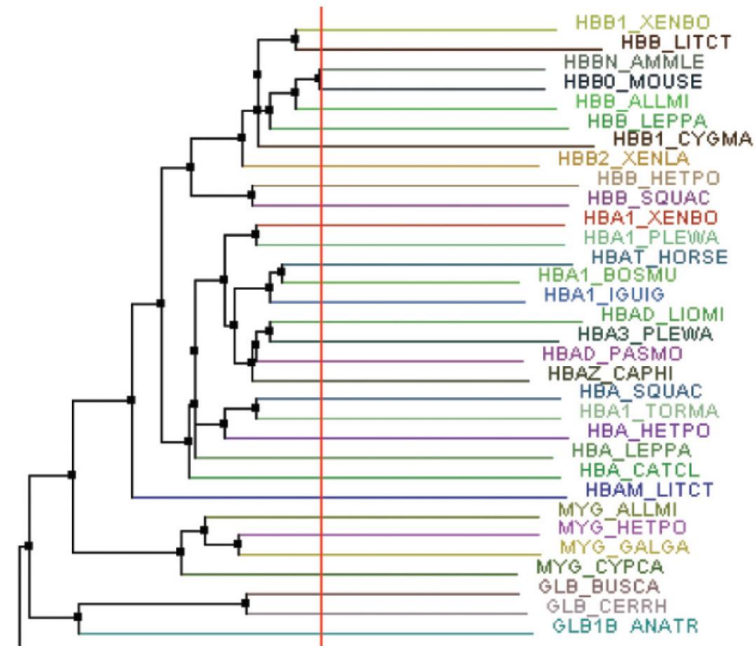
```
Q20638_CAEEL/74-184      EKELLRRRTWSD.EFD.....NLYELGSAIYCYIFDHNPNCKQLFP.F.ISKYQGDEWKESKEFRSQALKFVQTLAQVVK
Q19601_CAEEL/105-215     ERILLEQSWRK.TRK.....TGADHIGSKIFFMVLTAQPDIIKAIFG.L..EKIPTGRLKYDPRFRQHALVYTKTILDFVIR
Q18311_CAEEL/32-140      TKKLVIQEWPR.VLA.....QCPELFTEIWHKSATRSTSIKLAFG.I.AE.N..ESPMQNAAFGLSSTIQAFFYKLI
GLB4_LUMTE/11-120        DRREIRHIWDD.VWSSS.FTDRRVAIVRAVFDDLFKHYPTSKALFERVKIDEF.....ESGEFKSHLVRVANGDLLIN
GLB4_LUMTE/11-120 (SS)   HHHHHHHHHH.HS--S.SCHHHHHHHHHHHHHHHHHHSGGGGGGGGCCCTTTST.....TSSHHHHHHHHHHHHHHHHHT
GLB3_TYLHE/8-117         DRHEVLDNWKG.IWSAE.FTGRRVAIGQAI FQELFALDPNAKGVFGRVNV.D.K.....PSEADWKAHVIRVINGIDLAVN
GLB4_TYLHE/8-117         DRREVQALWRS.IWSAE.DTGRRTLIGRLLFEELEIDGATKGLFKRVNVDDT.....HSPFEFAHVLRVNVGLDTLIG
GLB1_TYLHE/7-110         QRIKVKQQWAQ.VYSV...GESRIDFAIDVFNNFFRTNPD.RS.LFNRVNGDNV.....YSPFEKAHMVRVFEAGFDILIS
GLB2_TYLHE/9-115         QRLKVKQQWAK.AYGV...GHERVELGIALWKSMAQDNDARDLFKRVHGEDV.....HSPAFAEHMARVFNGLDRVIS
GLB2_LUMTE/8-114         EGLKVKSEWGR.AYGS...GHDREAFSQAIWRATFAQVPESRSLFKRVHGGDDT.....SHPAFIAHAERVVGLGLDIAIS
GLB2_LUMTE/8-114 (SS)   HHHHHHHHHH.HH-S...HHHHHHHHHHHHHHHHHHH-GGGGGGGGGGTTT-T.....TSHHHHHHHHHHHHHHHHHHC
GLB_TUBTU/6-112         QRFKVKHQWAE.AFGT...SHHRLDFGLKLWNSIFRDAPEIRGLFKRVVDGD.N.....AYSAEFEAHAERVVGLGLDMTIS
GLB3_LAMSP/7-113        QRLKVKRQWAE.AYGS...GNDREEFGHFIWTHVFKDAPSARDLFKRVRGDNI.....HTPAFRAHATRVVGLGLDMCIA
```

Εικόνα 6.8 Η βάση δεδομένων Pfam είναι μια πλήρης και έγκυρη πηγή για τη μελέτη των οικογενειών πρωτεϊνών. (β) Τμήμα της στοίχισης εκκίνησης με τη μορφοποίηση HTML. Προσέξτε ότι για τις πρωτεΐνες γνωστής δομής υπάρχουν διπλές καταχωρίσεις: μία γραμμή περιέχει την αλληλουχία αυτή καθαυτή και μία γραμμή [κάτω από την αλληλουχία, με το διακριτικό «SS» (Secondary Structure)] δείχνει την αντίστοιχη δευτεροταγή δομή. Οι κωδικοί για τη δευτεροταγή δομή είναι: έλικα (H), στροφή (T), βρόχοι (S), 310-έλικα (G), β-κλώνοι (E).

(α) Ανάλυση κύριων συνιστωσών (PCA)



(β) Δέντρο που υπολογίστηκε με τη μέθοδο Neighbor-Joining



Εικόνα 6.9 Μπορούμε να εξετάσουμε τις στοίχισης της Pfam και μέσω του προγράμματος JalView. Το JalView μπορεί όχι μόνο να προβάλει τη στοίχιση, αλλά και να πραγματοποιήσει μια σειρά πολύπλοκων αναλύσεων. (α) Τα αποτελέσματα από την ανάλυση μιας στοίχισης σε κύριες συνιστώσες (PCA, Principal Component Analysis). Κάθε αλληλουχία αντιστοιχεί σε ένα σημείο και οι σχέσεις μεταξύ των αλληλουχιών (π.χ. ομοιότητα) αντιστοιχούν στις αποστάσεις των σημείων. Η γραφική παράσταση που βλέπετε στην εικόνα αυτή είναι στην πραγματικότητα τρισδιάστατη και μπορούμε ελεύθερα να την περιστρέψουμε γύρω από τους τρεις άξονες (που αντιστοιχούν στις τρεις πρώτες κύριες συνιστώσες). Το συγκεκριμένο παράδειγμα είναι από μια στοίχιση σφαιρινών και η PCA δείχνει ότι οι πέντε αλληλουχίες μιας ομάδας (είναι εκείνα τα σημεία στο επάνω τμήμα του διαγράμματος που περιλαμβάνουν και την αλληλουχία GLB2_NIPBR) είναι παρόμοιες μεταξύ τους, αλλά διαφέρουν σημαντικά από τις υπόλοιπες σφαιρίνες. Θα συζητήσουμε αναλυτικά την PCA στο Κεφάλαιο 11. (β) Τμήμα ενός φυλογενετικού δέντρου που υπολογίστηκε από το JalView με τη μέθοδο Neighbor-Joining (βλ. Κεφάλαιο 7). Η κόκκινη γραμμή που φαίνεται στο διάγραμμα επιτρέπει στον χρήστη του προγράμματος να περιορίσει την ανάλυση σε ένα υποσύνολο των αλληλουχιών της στοίχισης. Το JalView περιγράφηκε από τους Waterhouse et al. (2009).

Πίνακας 6.1 Βάσεις δεδομένων στις οποίες βασίζεται το Interpro (έκδοση 51.0). Οι καταχωρίσεις έχουν στρογγυλοποιηθεί στην πλησιέστερη εκατοντάδα.

Βάση δεδομένων	Περιεχόμενα (καταχωρίσεις)
PANTHER 9.0	60.000
Pfam 27.0	14.800
PIRSF 3.01	3.300
PRINTS 42.0	2.000
ProDom 2006.1	1.900
PROSITE 20.105 patterns	1.300
PROSITE 20.105 profiles	1.100
SMART 6.2	1.000
TIGRFAMs 15.0	4.500
CATH-Gene3D 3.5.0	2.600
SUPERFAMILY 1.75	2.000
UniProtKB 2015_04	47.300.000
UniProtKB/Swiss-Prot 2015_04	531.000
UniProtKB/TrEMBL 2015_04	46.715.000
GO Classification	27.000

Πηγή: http://www.ebi.ac.uk/interpro/έκδοση_notes.html. Πρόσβαση Απρίλιος 2015.

(α) Το γονίδιο *HBB* (σμίκρυνση 1,5x στα 2.409 ζεύγη βάσεων)

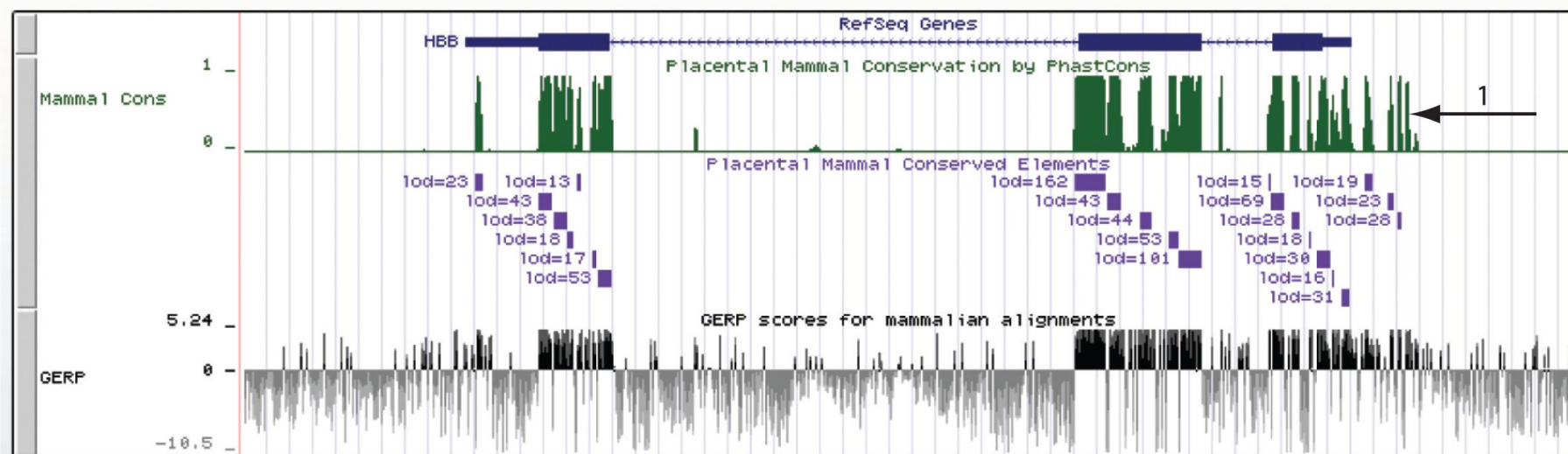
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

chr11:5,246,295-5,248,703 2,409 bp. chr11:5,246,295-5,248,703

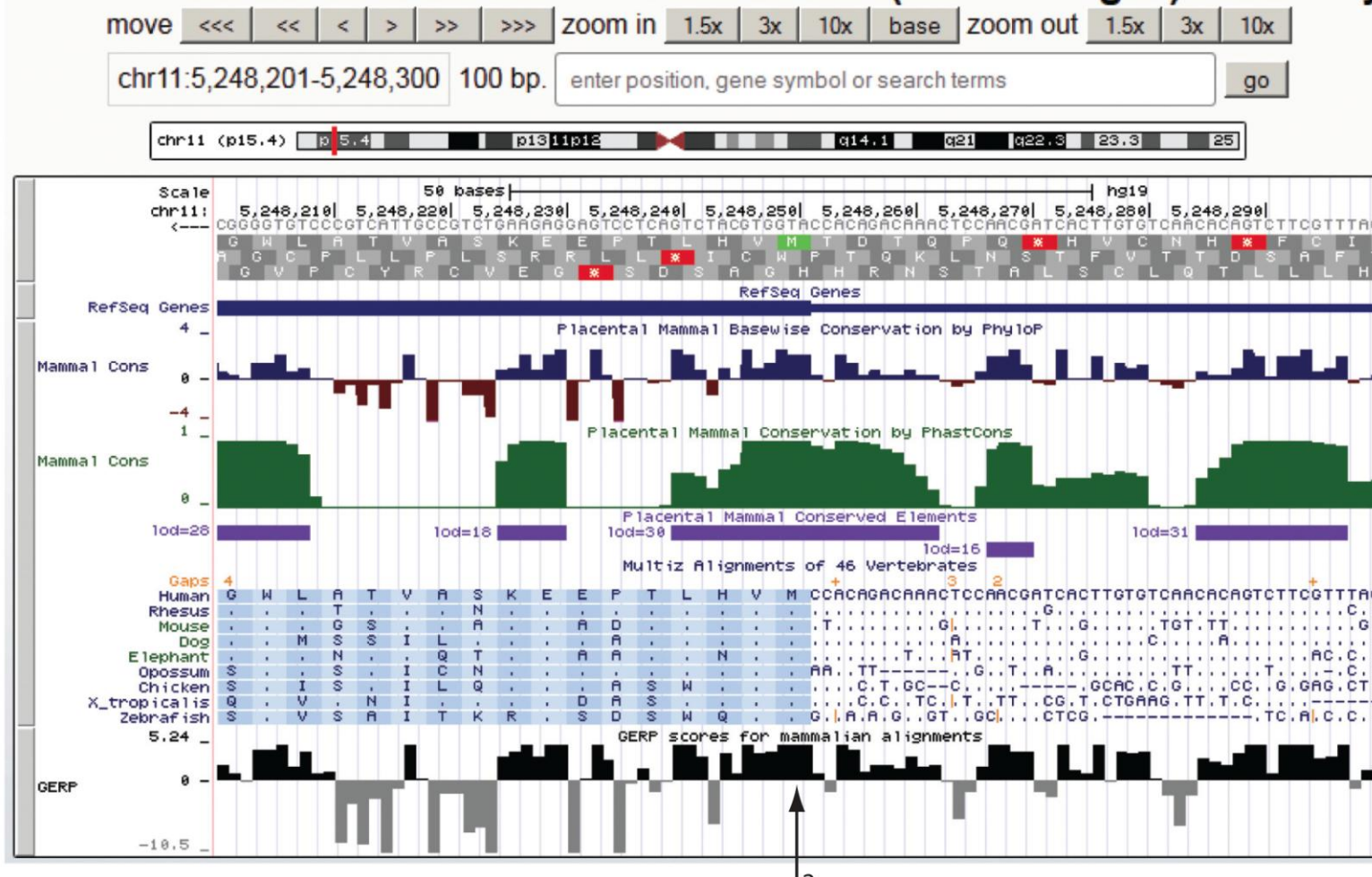
go

chr11 (p15.4) p5.4 p13 p11 p12 q14.1 q21 q22.3 23.3 25



Εικόνα 6.10 Στοίχιση του γονιδίου της ανθρώπινης β-σφαιρίνης (*HBB*) και ορθόλογων γονιδίων άλλων σπονδυλωτών. (α) Προβολή στο πρόγραμμα περιήγησης γονιδιωμάτων του UCSC της β-σφαιρίνης (σμίκρυνση 1,5 X). Τα εξόνια παρουσιάζονται ως συμπαγή παραλληλόγραμμα στο τμήμα της γραφικής παράστασης με τίτλο «RefSeq Genes» και τείνουν να είναι εξαιρετικά συντηρημένα μεταξύ μιας ομάδας γονιδιωμάτων σπονδυλωτών. Φαίνονται επίσης τρεις γραφικές παραστάσεις που δείχνουν το πόσο συντηρημένες είναι οι αλληλουχίες (Placental Mammal Conservation από το πρόγραμμα PhastCons, Placental Mammal Conserved Elements και βαθμολογίες GERP). Είναι προφανές ότι οι περιοχές των εξονίων είναι πολύ καλά συντηρημένες. Οι συντηρημένες περιοχές που δεν αντιστοιχούν σε εξόνια (όπως αυτές που δείχνει το βέλος 1) μπορεί να αντιπροσωπεύουν ρυθμιστικά στοιχεία.

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly



Εικόνα 6.10 (β) Εστιάζοντας σε 100 ζεύγη βάσεων, εμφανίζονται οι ίδιες γραφικές παραστάσεις, καθώς και η στοίχιση Multiz των 46 σπονδυλωτών (στην εικόνα φαίνονται εννέα). Τα στοιχισμένα νουκλεοτίδια εμφανίζονται στο δεξί μισό, ενώ τα αμινοξέα ξεκινούν με τη μεθειονίνη έναρξης (βέλος 2) και εκτείνονται προς τα αριστερά, ταιριάζοντας με την αρχή της πρωτεΐνης NP_000509.1, αλληλουχία MNHLTPREEKS. Κάνοντας κλικ στα γραφήματα (π.χ. σε μια κορυφή από το τμήμα PhastCons), μπορείτε να κατεβάσετε δεδομένα από τις αντίστοιχες στοιχίσεις.

(α) Η καταχώριση του γονιδίου *HBB* στο Ensembl

Gene: HBB ENSG00000244734

Description: hemoglobin, beta [Source:HGNC Symbol;Acc:4827]

Location: [Chromosome 11: 5,246,694-5,250,625](#) reverse strand.

INSDC coordinates: chromosome:GRCh37:CM000673.1:5246694:5250625:1

Transcripts: This gene has 4 transcripts (splice variants) [Hide transcript table](#)

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CDS incomplete	CCDS
HBB-001	ENST00000335295	754	ENSP00000333994	147	Protein coding	-	CCDS7753
HBB-004	ENST00000380315	502	ENSP00000369671	90	Protein coding	3'	-
HBB-002	ENST00000485743	680	No protein product	-	Retained intron	-	-
HBB-003	ENST00000475226	319	No protein product	-	Retained intron	-	-

Genomic alignments

Alignment: -- Select an alignment -- [Go](#)

Go to a track: 6 primates EPO, 13 eutherian mammals EPO, 20 amniota vertebrates Pecan, 36 eutherian mammals EPO LOW COVERAGE

Key: Pairwise alignments

Features: Alpaca (Vicugna pacos) - blastz, Anole lizard (Anolis carolinensis) - translated blat, Amadillo (Dasypus novemcinctus) - blastz, Bushbaby (Oryzomys flavescens) - lastz, Cat (Felis catus) - lastz, Chicken (Gallus gallus) - lastz, Chicken (Gallus gallus) - translated blat, Chimpanzee (Pan troglodytes) - lastz, Chinese softshell turtle (Pelodiscus sinensis) - lastz, Ciona intestinalis - translated blat, Ciona savignyi - translated blat, Cod (Gadus morhua) - translated blat, Coelacanth (Latimeria chalumnae) - translated blat

Human > chr11

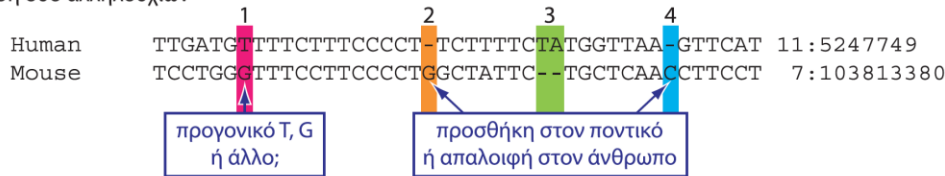
Human AGA TTTGAACAACATGATAAACCACATCCCATAGATGAGTGTCAATG
Human CCT ACTTAAGAAAGATTATAGACTGGAGTAAGGAAATGGAATCTGTCT
Human TTG GGGCTGGAATAAAGTAGAATAGACCTGCACCTGCTGTGACATCCAT
Human AGT CTGATTTAGATTGAAACTGAGGCTCTGACCATAACCAATTTGCAC
Human GCC TGTCCCTGCAGGTTATTATGGGTAATAGAAAGAAAGTCTGCTTACG
Human TTG CAGTTGCCAACAAGAGAGAGGATCCATAGTTTCAATTTAAAAAAG
Human TTT TTCTGCCAATCAGGATTTCAAAGCTCTTGTCTTGACAAATTTGGTC
Human TAT TGCATAAGACATATTCAACTTCGCGAGAACACTTTATTTACATAT
Human AAC TTAAATTTAATAAATAAATCCAAATCTAACAGCCAAGTCAAT
Human TTA TGATACAGCTGTGCTAGTACCTCATTGCTTTAGTTTTCACAGAGG

(β) Πολλαπλή στοίχιση αλληλουχιών στο Ensembl (προγράμματα Enredo/Pecan/Ortheus)

```
Homo sapiens      11: 5246983 TTCATACCTCTT-ATCTTCCCTCCCACAGCTCCTGGGCAACGTGCTGG
Gorilla gorilla gorilla 11: 5181973 TTCATACCTCTT-GTCTTCCCTCCCACAGCTCCTGGGCAATGTGCTGG
Pongo abelii      11: 65239065 TTCATACCTCTT-GTCTCCCTCCCACAGCTCCTGGGCAATGTGCTGG
Oryctolagus cuniculus 1: 146237264 TTCATGCTCTTCT--TCTCTTCCCTACAGCTCCTGGGCAACGTGCTGG
Mus musculus      7: 103812810 TTGATGGTTCTT--CCATCTTCCCACAGCTCCTGGGCAATATGATCG
Bos taurus        15: 49339417 CCCTTGCTTAATG-TCTTTTCCACACAGCTCCTGGGCAACGTGCTAG
Bos taurus        15: 49074455 CCCTTGCTTAATG-TCTTTTCCACACAGCTCCTGGGCAACGTGCTGG
Sus scrofa        9: 5633260 CCCTCTCTTTTTA-TCTCTCTCCCACAGCTCCTGGGCAACGTGATAG
Equus caballus    7: 73936736 CCCCTCTTTT-TT-TCTCTTCCCACAGCTCCTGGGCAACGTGCTGG
Canis lupus familiaris 21: 28179266 CACATGCCCTTTG-TCT--TCCCACAGCTGCTGGGCAACGTGTTGG
```

Εικόνα 6.11 Ανάλυση πολλαπλών στοιχίσεων αλληλουχιών μέσω του Ensembl. (α) Μετά από μια αναζήτηση για την ανθρώπινη β-σφαιρίνη (HBB), επιλέγουμε πρώτα το «Genomic alignments» (βέλος 1) και στη συνέχεια μια ομάδα που θέλουμε, π.χ. 6 πρωτευόντων ή 36 θηλαστικών (βέλος 2). Οι στοιχισμένες αλληλουχίες εμφανίζονται με τα εξόνια σημειωμένα με κόκκινο χρώμα (βέλος 3). (β) Παρουσιάζεται ένα τμήμα των αποτελεσμάτων από τη διαδικασία EPO (Enredo/Pecan/Ortheus).

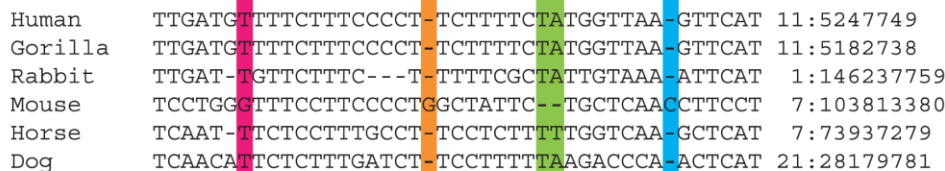
(α) Στοιχίση δύο αλληλουχιών



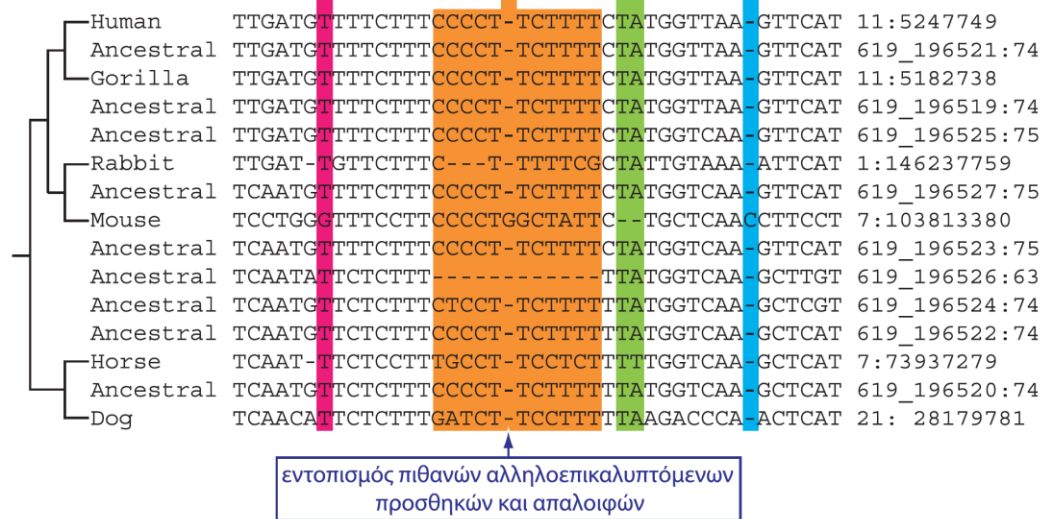
(β) Στοιχίση που περιλαμβάνει την προγονική αλληλουχία



(γ) Πολλαπλή στοιχίση αλληλουχιών



(δ) Πολλαπλή στοιχίση αλληλουχιών που περιλαμβάνει τις προγονικές τους αλληλουχίες



Εικόνα 6.12 Το πρόγραμμα Ortheus, που είναι τμήμα της ανάλυσης EPO του Ensembl, παρέχει στοιχίσεις αλληλουχιών προσανατολισμένες στη φυλογένεση και περιλαμβάνει τη δυνατότητα της ανασύνθεσης προγονικών αλληλουχιών. (α) Στη στοιχίση μεταξύ DNA ανθρώπου και ποντικού από την περιοχή της β-σφαιρίνης, δεν είναι σαφές αν οι στοιχισμένες βάσεις T και G στη στήλη 1 προέρχονται από ένα προγονικό T, G ή κάποιο άλλο νουκλεοτίδιο. Είναι επίσης αδύνατο να αποφανθούμε για το αν τα κενά αντιστοιχούν σε γεγονότα προσθήκης στο ένα είδος ή απάλειψης στο άλλο (βλ. στήλες 2, 3 και 4). (β) Συνάγοντας μια προγονική αλληλουχία ανθρώπου/ποντικού μπορούμε να συμπεράνουμε τα προγονικά αλληλόμορφα και να προσδιορίσουμε αν οι θέσεις με κενά αντιστοιχούν σε προσθήκες ή απάλειψές και σε ποιο είδος. (γ) Μια πολλαπλή στοιχίση αλληλουχιών προσφέρει περαιτέρω στοιχεία και ενισχύει τα συμπεράσματα από την προηγούμενη ανάλυση. (δ) Εισαγωγή των προγονικών αλληλουχιών στην πολλαπλή στοιχίση αλληλουχιών μάς δίνει πληροφορίες για κάθε κόμβο του φυλογενετικού δέντρου που συνδέει αυτές τις αλληλουχίες και επιλύει ερωτήματα σχετικά με την προέλευση προσθηκών, απάλειψών και σύνθετων συμβάντων, όπως αυτά που περιλαμβάνονται στο πορτοκαλί πλαίσιο. Δεξιά από κάθε αλληλουχία είναι σημειωμένα το χρωμόσωμα και η θέση για τις σημερινές αλληλουχίες ή ένα αναγνωριστικό για τις προγονικές. Μπορείτε να δείτε τις αλληλουχίες που εμφανίζονται σε αυτή την εικόνα ακολουθώντας τη διαδικασία που περιγράφει η Εικόνα 6.11 και επιλέγοντας «13 eutherian mammals EPO». Στη συνέχεια, κάντε κλικ στο «Configure this page» για να προσθέσετε ή να αφαιρέσετε είδη καθώς και προγονικές αλληλουχίες.