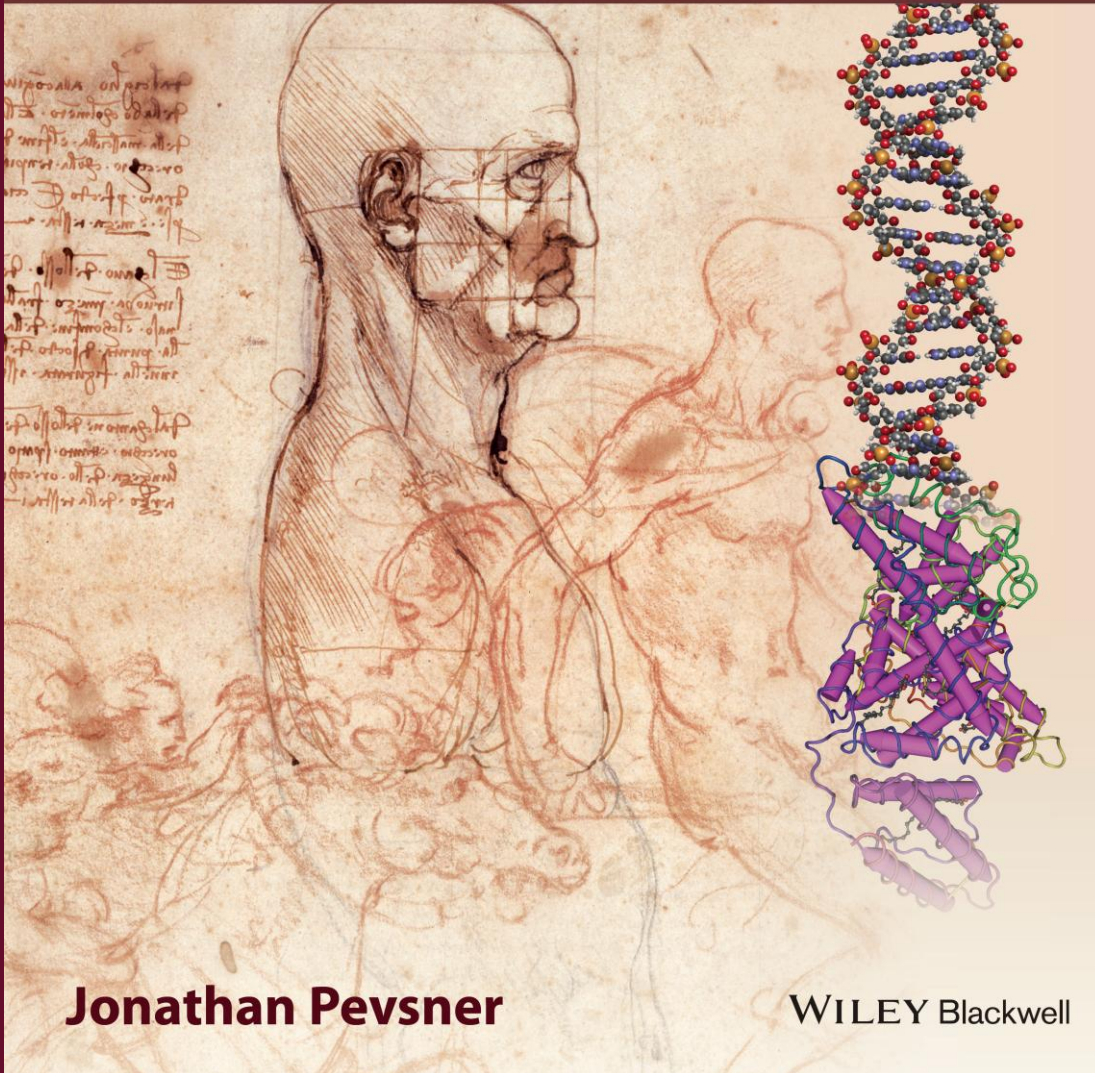


BIOINFORMATICS AND FUNCTIONAL GENOMICS

third edition



Jonathan Pevsner

WILEY Blackwell

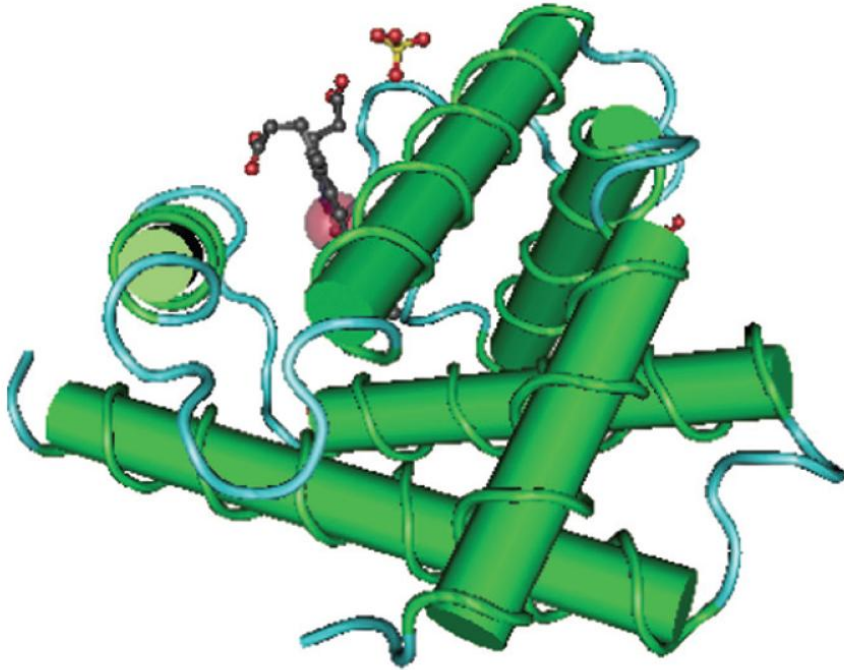
Κεφάλαιο 3

Στοίχιση αλληλουχιών κατά ζεύγη

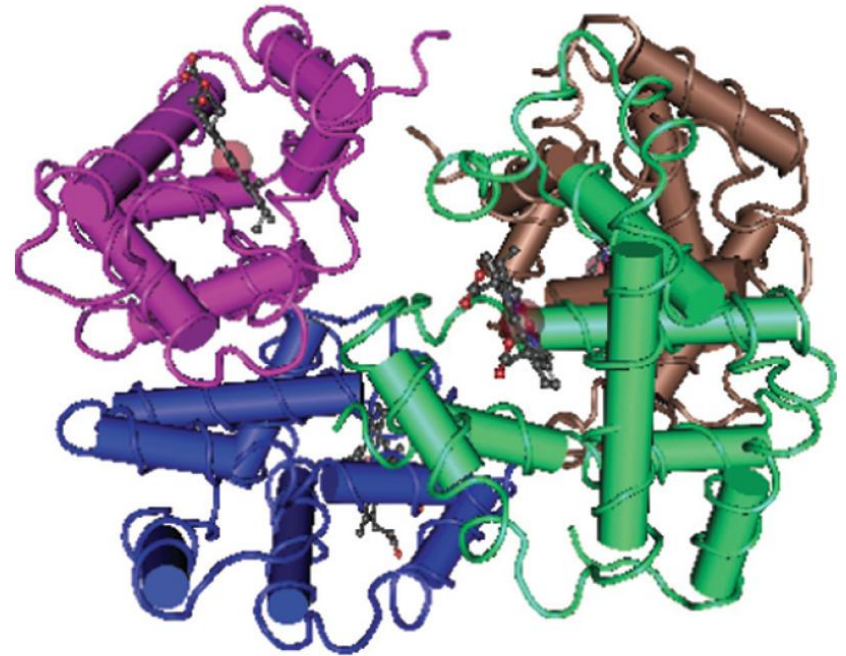
Ακαδημαϊκές
Εκδόσεις



(α) Ανθρώπινη μυοσφαιρίνη (3RGK)

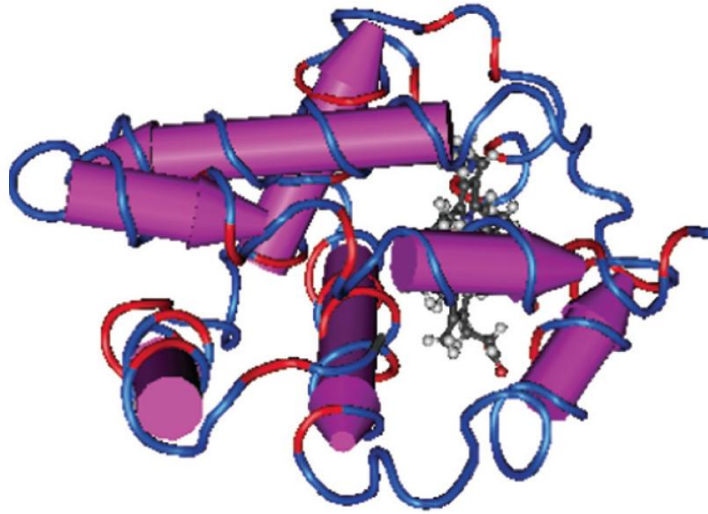


(β) Τετραμερές ανθρώπινης αιμοσφαιρίνης (2H35)

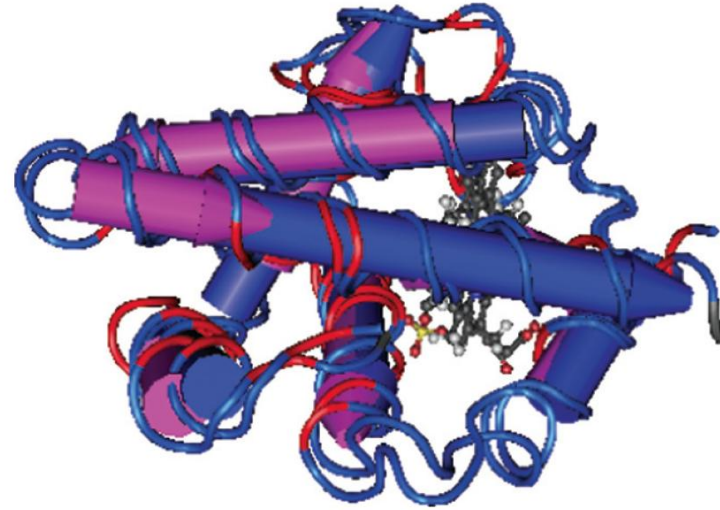


Εικόνα 3.1 Τρισδιάστατες δομές: (α) της μυοσφαιρίνης (καταχώριση 3RGK), (β) του τετραμερούς αιμοσφαιρίνης (2H35).

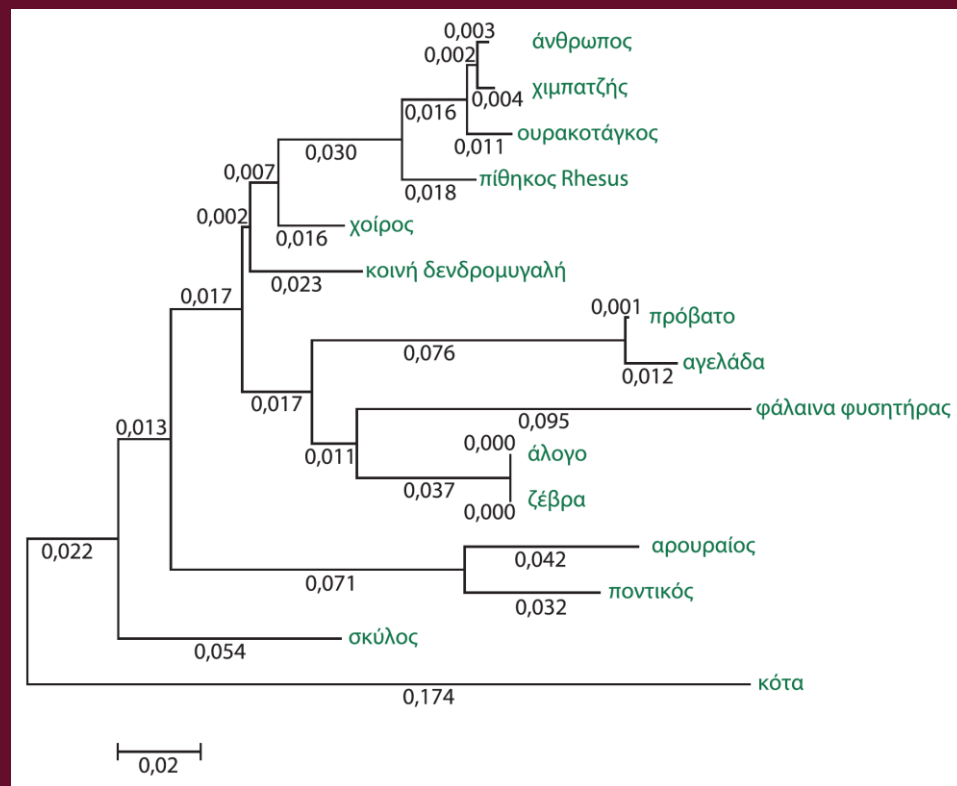
(γ) Ανθρώπινη β-σφαιρίνη (υπομονάδα 2H35)



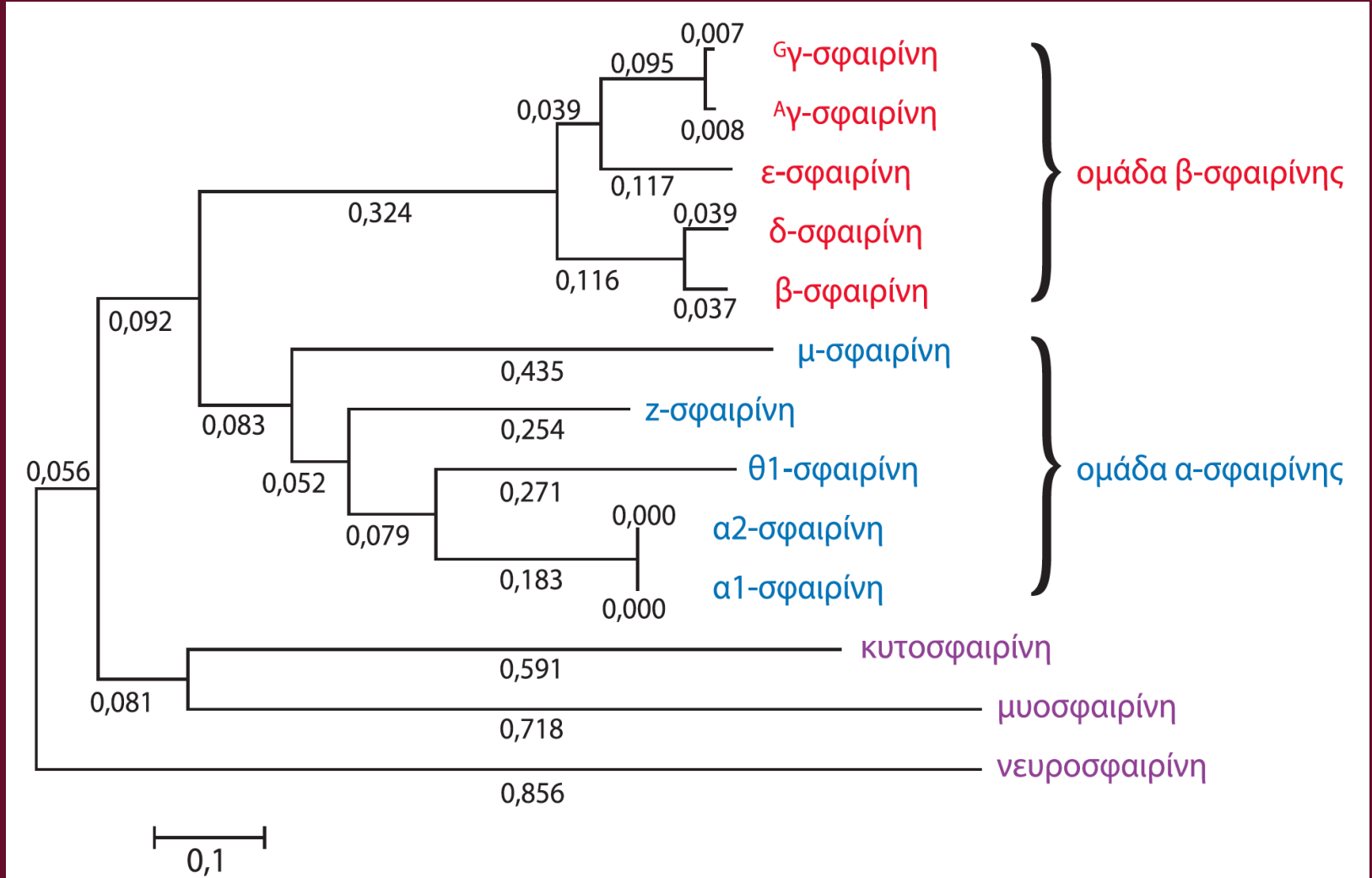
(δ) Υπέρθεση των δομών της β-σφαιρίνης και της μυοσφαιρίνης



Εικόνα 3.1 Τρισδιάστατες δομές: (γ) της υπομονάδας β-σφαιρίνης της αιμοσφαιρίνης και (δ) της υπέρθεσης των δομών της μυοσφαιρίνης και της β-σφαιρίνης. Οι εικόνες δημιουργήθηκαν με το πρόγραμμα Cn3D (βλ. Κεφάλαιο 13). Αυτές οι πρωτεΐνες είναι ομόλογες (προέρχονται από έναν κοινό πρόγονο) και έχουν παρόμοιες τρισδιάστατες δομές. Ωστόσο, η στοίχιση κατά ζεύγη των αλληλουχιών τους αποκαλύπτει ότι εμφανίζουν πολύ περιορισμένη ταύτιση αμινοξέων.



Εικόνα 3.2 Φυλογενετικό δέντρο ορθόλογων αλληλουχιών μυοσφαιρίνης. Για την κατασκευή του δέντρου αρχικά πραγματοποιήθηκε πολλαπλή στοίχιση (Κεφάλαιο 6) των αλληλουχιών και στη συνέχεια χρησιμοποιήθηκε μία μέθοδος γνωστή ως Neighbor-Joining (μέθοδος ένωσης γειτόνων, Κεφάλαιο 7). Οι αριθμοί καταχώρισης των αλληλουχιών και τα ονόματα των ειδών έχουν ως εξής: άνθρωπος NP_005359 (*Homo sapiens*), χιμπατζής XP_001156591 (*Pan troglodytes*), ουρακοτάγκος P02148 (*Pongo pygmaeus*), πίθηκος Rhesus XP_001082347 (*Macaca mulatta*), χοίρος NP_999401 (*Sus scrofa*), κοινή δενδρομυγαλή P02165 (*Tupaia glis*), αλόγο P68082 (*Equus caballus*), ζέβρα P68083 (*Equus burchellii*), σκύλος XP_850735 (*Canis familiaris*), φάλαινα φυσητήρας P02185 (*Physeter catodon*), πρόβατο P02190 (*Ovis aries*), αρουραίος NP_067599 (*Rattus norvegicus*), ποντικός NP_038621 (*Mus musculus*), αγελάδα NP_776306 (*Bos taurus*), κότα XP_416292 (*Gallus gallus*). Μπορείτε να βρείτε τις αλληλουχίες στο Web Document 3.1 (<http://www.bioinfbook.org/chapter3>). Στα φυλογενετικά δέντρα, όσο πιο στενά συνδεδεμένες μεταξύ τους είναι δύο αλληλουχίες, τόσο πιο κοντά τοποθετούνται. Σημειώστε ότι, καθώς συνεχίζεται η αλληλούχιση ολόκληρων γονιδιωμάτων, στις περισσότερες οικογένειες ορθόλογων πρωτεϊνών ο αριθμός των μελών αυξάνεται ταχέως.



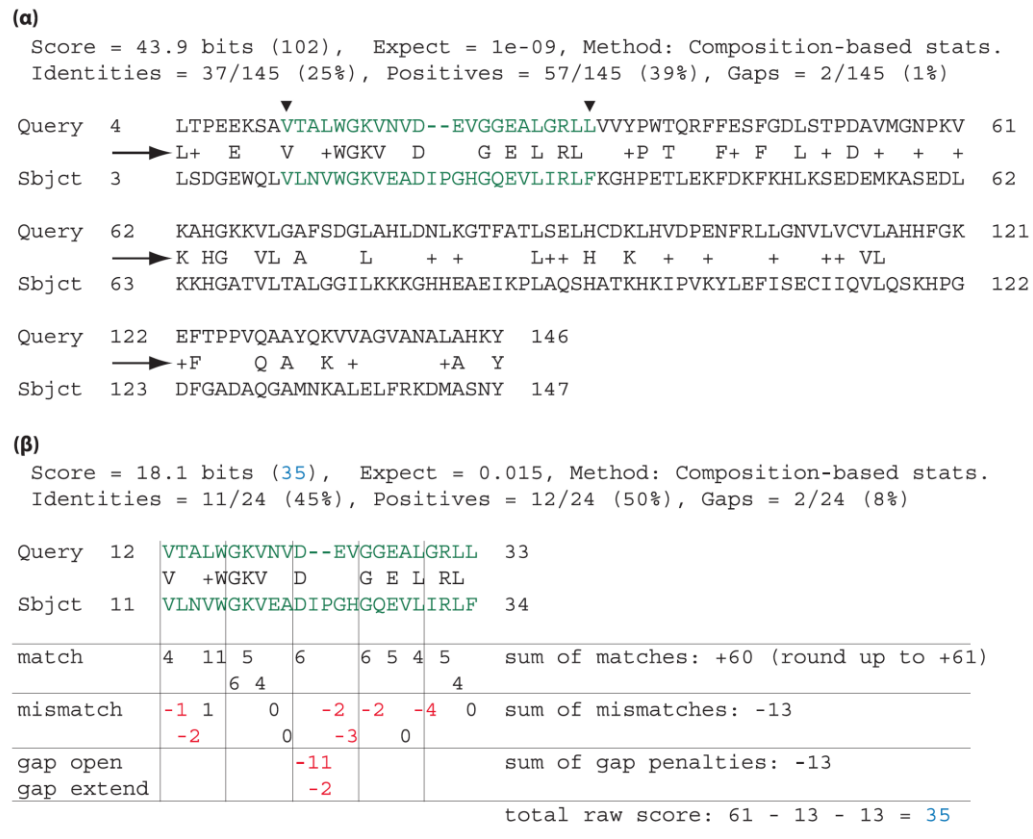
Εικόνα 3.3 Παράλογες ανθρώπινες σφαιρίνες. Όλες αυτές οι πρωτεΐνες είναι ανθρώπινες και ανήκουν στην οικογένεια των σφαιρινών. Αυτό το δέντρο χωρίς ρίζα (unrooted tree) κατασκευάστηκε με τον αλγόριθμο Neighbor-Joining στο MEGA (βλ. Κεφάλαιο 7). Οι πρωτεΐνες και οι RefSeq αριθμοί καταχώρισής τους (Web Document 3.2) είναι οι εξής: δ-σφαιρίνη (NP_000510), γ2-σφαιρίνη (NP_000175), β-σφαιρίνη (NP_000509), γ1-σφαιρίνη (NP_000550), ε-σφαιρίνη (NP_005321), ζ-σφαιρίνη (NP_005323), α1-σφαιρίνη (NP_000549), α2-σφαιρίνη (NP_000508), θ1-σφαιρίνη (NP_005322), αλυσίδα mu αιμοσφαιρίνης (NP_001003938), κυτοσφαιρίνη (NP_599030), μυοσφαιρίνη (NP_005359) και νευροσφαιρίνη (NP_067080). Χρησιμοποιήθηκε το μοντέλο διόρθωσης Poisson (βλ. Κεφάλαιο 7).

The screenshot shows the NCBI BLAST web interface for protein alignment. The page title is 'Align Sequences Protein BLAST'. The interface is divided into several sections:

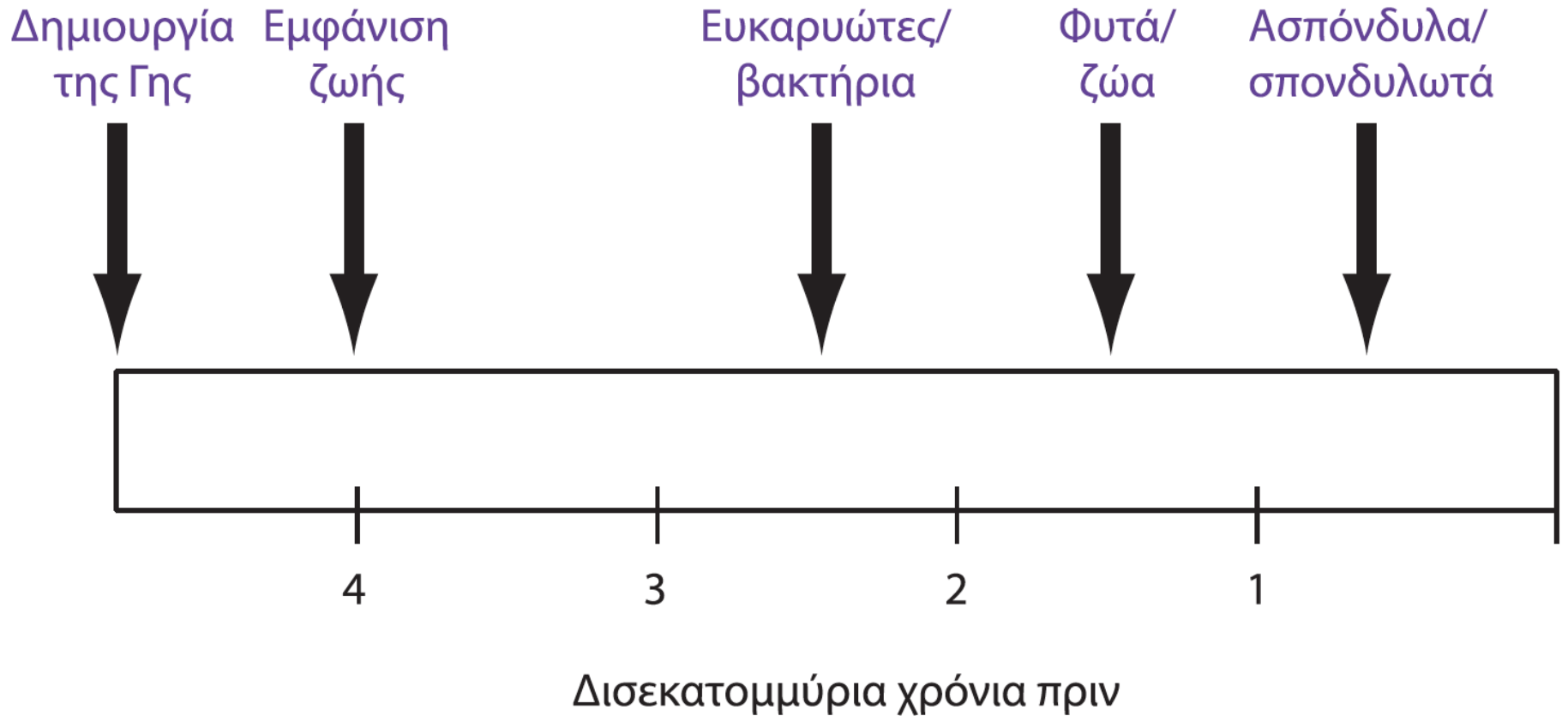
- Enter Query Sequence:** This section contains a text input field for the query sequence, a 'Clear' button, and a 'Query subrange' section with 'From' and 'To' input fields. A red arrow labeled '1' points to the query sequence input field, which contains a FASTA-formatted sequence for hemoglobin subunit beta.
- Enter Subject Sequence:** This section contains a text input field for the subject sequence, a 'Clear' button, and a 'Subject subrange' section with 'From' and 'To' input fields. A red arrow labeled '3' points to the subject sequence input field, which contains the accession number 'NP_005359'.
- Program Selection:** This section contains a dropdown menu for the algorithm, currently set to 'blastp (protein-protein BLAST)'. A red arrow labeled '2' points to the 'Align two or more sequences' checkbox, which is checked.
- BLAST Button:** A large blue button labeled 'BLAST' is located at the bottom of the form. A red arrow labeled '4' points to this button.

Below the BLAST button, there is a checkbox for 'Show results in a new window' and a link for '+ Algorithm parameters'.

Εικόνα 3.4 Τα προγράμματα της οικογένειας BLAST στην ιστοσελίδα του NCBI επιτρέπουν τη σύγκριση δύο αλληλουχιών DNA ή πρωτεϊνών. Εδώ χρησιμοποιείται το πρόγραμμα BLASTP για τη σύγκριση δύο πρωτεϊνών (βέλος 2). Η ανθρώπινη β-σφαιρίνη (NP_000509) εισάγεται με τη μορφή αλληλουχίας FASTA (βέλος 1), ενώ για την ανθρώπινη μυοσφαιρίνη (NP_005359) εισάγεται η αναφορά του κωδικού καταχώρισής της (βέλος 3). Πατώντας το κουμπί BLAST (βέλος 4), πραγματοποιείται η στοίχιση. Προσέξτε τον σύνδεσμο κάτω αριστερά που μας επιτρέπει να δούμε και να τροποποιήσουμε τις παραμέτρους του αλγορίθμου.



Εικόνα 3.5 Στοίχιση μεταξύ της ανθρώπινης β-σφαιρίνης (ως «αλληλουχία αναζήτησης») και της μυοσφαιρίνης (ως «αντικείμενο»). (α) Το αποτέλεσμα της στοίχισης από τον υπολογισμό που παρουσιάστηκε στην Εικόνα 3.4. Σημειώστε ότι αυτή η στοίχιση είναι τοπική (δηλαδή δε στοιχίζονται σε όλη τους την έκταση οι δύο πρωτεΐνες). Πολλά αμινοξέα είναι ταυτόσημα στις δύο αλληλουχίες [επισημαίνονται στις ενδιάμεσες γραμμές μεταξύ των αλληλουχιών (δείτε τα βέλη)]. Η στοίχιση περιέχει ένα εσωτερικό κενό που υποδεικνύεται με δύο παύλες στην αλληλουχία αναζήτησης. (β) Εικονογράφηση του πώς υπολογίζεται η πρωτογενής βαθμολογία (raw score) μιας στοίχισης. Στο παράδειγμα αυτό ως αλληλουχία αναζήτησης έχουν χρησιμοποιηθεί μόνο τα αμινοξέα 12-33 της ανθρώπινης β-σφαιρίνης, τα οποία στο (α) βρίσκονται ανάμεσα σε δύο κεφαλές βέλους και υποδεικνύονται με πράσινα γράμματα. Η πρωτογενής βαθμολογία είναι 35 και αντιπροσωπεύει το άθροισμα τριών όρων: των βαθμολογιών (από τον πίνακα BLOSUM62) για τα στοιχισμένα και τα μη στοιχισμένα αμινοξέα, την ποινή/κόστος δημιουργίας κενού (που εδώ έχει τιμή −11) και την ποινή επέκτασης κενού (που εδώ έχει τιμή −1). Οι πρωτογενείς βαθμολογίες μετατρέπονται στη συνέχεια σε bit σκορ (βλ. Κεφάλαιο 4).



Εικόνα 3.6 Επισκόπηση της ιστορίας της ζωής στη Γη. Οι αλληλουχίες των πρωτεϊνών και του DNA αναλύονται υπό το πρίσμα της εξέλιξης. Ποιοι οργανισμοί έχουν ορθόλογα γονίδια; Πότε στην πορεία της εξέλιξης προέκυψαν οι οργανισμοί αυτοί; Πόσο σχετίζονται μεταξύ τους οι ανθρώπινες και οι βακτηριακές σφαιρίνες;



Εικόνα 3.7 Η προσέγγιση της Dayhoff για τον προσδιορισμό των αμινοξικών αντικαταστάσεων. (α) Τμήμα πολλαπλής στοίχισης ανάμεσα στις αλληλουχίες της α1-σφαιρίνης, της β-σφαιρίνης, της δ-σφαιρίνης και της μυοσφαιρίνης του ανθρώπου. Με κόκκινο χρώμα υποδεικνύονται τέσσερις στήλες στις οποίες η α1-σφαιρίνη και η μυοσφαιρίνη έχουν διαφορετικά αμινοξέα. Για παράδειγμα, το A στοιχίζεται με το G (βέλος). (β) Φυλογενετικό δέντρο που παρουσιάζει τις τέσσερις υπάρχουσες αλληλουχίες (1-4), καθώς και δύο εσωτερικούς κόμβους που αντιπροσωπεύουν τις προγονικές αλληλουχίες (5, 6). Οι συναγόμενες προγονικές αλληλουχίες [σημειώνονται ως αλληλουχίες 5 και 6 στο (α)] προσδιορίστηκαν μέσω της αρχής της μεγίστης φειδωλότητας χρησιμοποιώντας το λογισμικό RAUP (Κεφάλαιο 7). Από την ανάλυση αυτή είναι φανερό ότι σε καθεμία από τις στήλες που επισημαίνονται με κόκκινο χρώμα δεν υπήρξαν άμεσες αντικαταστάσεις αμινοξέων μεταξύ α1-σφαιρίνης και μυοσφαιρίνης. Οι αλληλουχίες των δύο αυτών πρωτεϊνών προέκυψαν μέσω αντικαταστάσεων στην αλληλουχία του κοινού τους προγόνου. Για παράδειγμα, στη θέση που επισημαίνεται με βέλος στο (α) δε συνέβη μια άμεση αντικατάσταση αλανίνης από γλυκίνη, αλλά ένα (προγονικό) κατάλοιπο glu άλλαξε σε ala και gly στην α1-σφαιρίνη και στη μυοσφαιρίνη αντίστοιχα.

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	γ	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

Εικόνα 3.8 Πλήθος αποδεκτών σημειακών μεταλλαγών (πολλαπλασιασμένο επί 10) σε 1.572 περιπτώσεις αντικαταστάσεων αμινοξέων μεταξύ στενά συγγενικών πρωτεϊνών. Τα αμινοξέα παρουσιάζονται αλφαβητικά σύμφωνα με τον κώδικα των τριών γραμμάτων. Παρατηρήστε ότι ορισμένες αντικαταστάσεις (πράσινα κελιά) είναι πολύ συχνά αποδεκτές (όπως μεταξύ V και I ή S και T). Άλλα αμινοξέα, όπως το C και το W, σπάνια αντικαθίστανται από οποιοδήποτε άλλο αμινοξύ (πορτοκαλί κελιά).

Πίνακας 3.1 Κανονικοποιημένες συχνότητες αμινοξέων. Οι τιμές αυτές αθροίζονται στο 1. Αν τα 20 αμινοξέα αντιπροσωπεύονταν εξίσου στις πρωτεΐνες, αυτές οι τιμές θα ήταν όλες 0,05 (δηλαδή 5%). Ωστόσο, η συχνότητα εμφάνισής τους διαφέρει.

Gly	0,089	Arg	0,041
Ala	0,087	Asn	0,040
Leu	0,085	Phe	0,040
Lys	0,081	Gln	0,038
Ser	0,070	Ile	0,037
Val	0,065	His	0,034
Thr	0,058	Cys	0,033
Pro	0,051	Tyr	0,030
Glu	0,050	Met	0,015
Asp	0,047	Trp	0,010

Πηγή: Dayhoff (1972). Αναδημοσιεύεται κατόπιν αδείας του National Biomedical Research Foundation.

Πίνακας 3.2 Σχετικές μεταλλαξιμότητες αμινοξέων. Η σχετική μεταλλαξιμότητα της αλανίνης ορίζεται αυθαίρετα ως 100.

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

Πηγή: Dayhoff (1972). Αναδημοσιεύεται κατόπιν αδείας του National Biomedical Research Foundation.

		Αρχικό αμινοξύ																			
		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
Αμινοξύ που αντικαθιστά το αρχικό	A	98,7	0,0	0,1	0,1	0,0	0,1	0,2	0,2	0,0	0,1	0,0	0,0	0,1	0,0	0,2	0,4	0,3	0,0	0,0	0,2
	R	0,0	99,1	0,0	0,0	0,0	0,1	0,0	0,0	0,1	0,0	0,0	0,2	0,0	0,0	0,0	0,1	0,0	0,1	0,0	0,0
	N	0,0	0,0	98,2	0,4	0,0	0,0	0,1	0,1	0,2	0,0	0,0	0,1	0,0	0,0	0,0	0,2	0,1	0,0	0,0	0,0
	D	0,1	0,0	0,4	98,6	0,0	0,1	0,5	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0
	C	0,0	0,0	0,0	0,0	99,7	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0
	Q	0,0	0,1	0,0	0,1	0,0	98,8	0,3	0,0	0,2	0,0	0,0	0,1	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0
	E	0,1	0,0	0,1	0,6	0,0	0,4	98,7	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	G	0,2	0,0	0,1	0,1	0,0	0,0	0,1	99,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,1
	H	0,0	0,1	0,2	0,0	0,0	0,2	0,0	0,0	99,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	I	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	98,7	0,1	0,0	0,2	0,1	0,0	0,0	0,1	0,0	0,0	0,3
	L	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,2	99,5	0,0	0,5	0,1	0,0	0,0	0,0	0,0	0,0	0,2
	K	0,0	0,4	0,3	0,1	0,0	0,1	0,1	0,0	0,0	0,0	0,0	99,3	0,2	0,0	0,0	0,1	0,1	0,0	0,0	0,0
	M	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,1	0,0	98,7	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	F	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,1	0,0	0,0	99,5	0,0	0,0	0,0	0,0	0,3	0,0
	P	0,1	0,1	0,0	0,0	0,0	0,1	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	99,3	0,1	0,0	0,0	0,0	0,0
	S	0,3	0,1	0,3	0,1	0,1	0,0	0,1	0,2	0,0	0,0	0,0	0,1	0,0	0,0	0,2	98,4	0,4	0,1	0,0	0,0
	T	0,2	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,1	0,1	0,0	0,1	0,3	98,7	0,0	0,0	0,1
	W	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	99,8	0,0	0,0
	Y	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,0	99,5	0,0
	V	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,6	0,1	0,0	0,2	0,0	0,0	0,0	0,1	0,0	0,0	99,0

Εικόνα 3.9 Ο πίνακας πιθανότητας μεταλλαγής PAM1. Οι στήλες αντιστοιχούν στα αρχικά αμινοξέα (*j*) και οι σειρές στα αμινοξέα που τα αντικαθιστούν (*i*). Η Dayhoff και οι συνεργάτες της πολλαπλασίασαν τις τιμές επί 10.000 (για να αυξήσουν την ακρίβεια), ενώ εδώ έχουμε πολλαπλασιάσει με το 100, έτσι ώστε, για παράδειγμα, η τιμή του πρώτου κελιού 98,7 να αντιστοιχεί σε πιθανότητα ίση με 98,7% (που στο συγκεκριμένο παράδειγμα αντιστοιχεί στην πιθανότητα μια αλανίνη να μη μεταλλαχθεί σε αυτό το εξελικτικό διάστημα).

NP_002037.2	164	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSKLWRDGRGALQNII	207
XP_001162057.1	164	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSKLWRDGRGALQNII	207
NP_001003142.1	162	IHDHFGIVEGLMTTVHAIITATQKTVDGPGSKMWRDGRGAAQNII	205
XP_893121.1	168	IHDNFGIMEGLMTTVHAIITATQKTVDGPGSKLWRDGRGAAQNII	211
XP_576394.1	162	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSKLWRDGRGAAQNII	205
NP_058704.1	162	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSKLWRDGRGAAQNII	205
XP_001070653.1	162	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSKLWRDGRGAAQNII	205
XP_001062726.1	162	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSKLWRDGRGAAQNII	205
NP_989636.1	162	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSKLWRDGRGAAQNII	205
NP_525091.1	161	INDNFEIVEGLMTTVHATTATQKTVDGPGSKLWRDGRGAAQNII	204
XP_318655.2	161	INDNFGILEGLMTTVHATTATQKTVDGPGSKLWRDGRGAAQNII	204
NP_508535.1	170	INDNFGIIEGLMTTVHAVTATQKTVDGPGSKLWRDGRGAGQNII	213
NP_595236.1	164	INDTFGIEEGLMTTVHATTATQKTVDGPSSKDWRGGRGASANII	207
NP_011708.1	162	INDAFGIEEGLMTTVHSLTATQKTVDGPSSHKDWRGGRTASGNII	205
XP_456022.1	161	INDEFGIDEALMTTVHSITATQKTVDGPSSHKDWRGGRTASGNII	204
NP_001060897.1	166	IHDNFGIIEGLMTTVHAIITATQKTVDGPSSKDWRGGRAASFNII	209

Εικόνα 3.10 Πολλαπλή στοίχιση ενός τμήματος της GAPDH από 13 οργανισμούς: *Homo sapiens* (άνθρωπος), *Pan troglodytes* (χιμπατζής), *Canis lupus* (σκύλος), *Mus musculus* (ποντικός), *Rattus norvegicus* (αρουραίος, τρεις μορφές), *Gallus gallus* (κοτόπουλο), *Drosophila melanogaster* (μύγα των φρούτων), *Anopheles gambiae* (κουνούπι), *Caenorhabditis elegans* (σκώληκας), *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae* (ζυμομύκητας), *Kluyveromyces lactis* (ένας ακόμα μύκητας) και *Oryza sativa* (ρύζι). Οι στήλες στις οποίες υπάρχει έστω και μία αμινοξική αλλαγή υποδεικνύονται με βέλη. Αριστερά δίνονται οι αριθμοί πρόσβασης. Η στοίχιση δημιουργήθηκε πραγματοποιώντας στο HomoloGene του NCBI μια αναζήτηση με τον όρο «gapdh».

		Αρχικό αμινοξύ							
Αμινοξύ που αντικαθιστά το αρχικό	PAM0	A	R	N	D	C	Q	E	G
	A	100	0	0	0	0	0	0	0
	R	0	100	0	0	0	0	0	0
	N	0	0	100	0	0	0	0	0
	D	0	0	0	100	0	0	0	0
	C	0	0	0	0	100	0	0	0
	Q	0	0	0	0	0	100	0	0
	E	0	0	0	0	0	0	100	0
	G	0	0	0	0	0	0	0	100
		Αρχικό αμινοξύ							
Αμινοξύ που αντικαθιστά το αρχικό	PAM ∞	A	R	N	D	C	Q	E	G
	A	8,7	8,7	8,7	8,7	8,7	8,7	8,7	8,7
	R	4,1	4,1	4,1	4,1	4,1	4,1	4,1	4,1
	N	4,0	4,0	4,0	4,0	4,0	4,0	4,0	4,0
	D	4,7	4,7	4,7	4,7	4,7	4,7	4,7	4,7
	C	3,3	3,3	3,3	3,3	3,3	3,3	3,3	3,3
	Q	3,8	3,8	3,8	3,8	3,8	3,8	3,8	3,8
	E	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0
	G	8,9	8,9	8,9	8,9	8,9	8,9	8,9	8,9

Εικόνα 3.12 Τμήμα των πινάκων PAM που αντιστοιχούν σε μηδενική τιμή PAM (PAM0, άνω) και άπειρη τιμή PAM (PAM ∞ , κάτω). Στον PAM ∞ (που προκύπτει αν ο PAM1 πολλαπλασιαστεί άπειρες φορές με τον εαυτό του) οι αριθμητικές τιμές προσεγγίζουν την κανονικοποιημένη συχνότητα εμφάνισης του αμινοξέος αντικατάστασης (βλ. Πίνακα 3.1). Ο πίνακας PAM2000 έχει παρόμοιες τιμές με τον PAM ∞ , οι οποίες επίσης τείνουν να προσεγγίζουν την κανονικοποιημένη συχνότητα του αμινοξέος αντικατάστασης. Στον PAM2000 οι πρωτεΐνες που συγκρίνονται είναι σχεδόν πλήρως μη σχετιζόμενες. Αντίθετα, στον PAM0 δεν υπάρχουν μεταλλαγές και τα αμινοξέα των πρωτεϊνών είναι απολύτως συντηρημένα.

		Αρχικό αμινοξύ																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Αμινοξύ που αντικαθιστά το αρχικό	A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
	R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
	N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
	D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
	C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
	Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
	E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
	G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
	H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
	I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
	L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
	K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
	M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
	F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
	P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
	S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
	T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
	W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
	Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
	V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

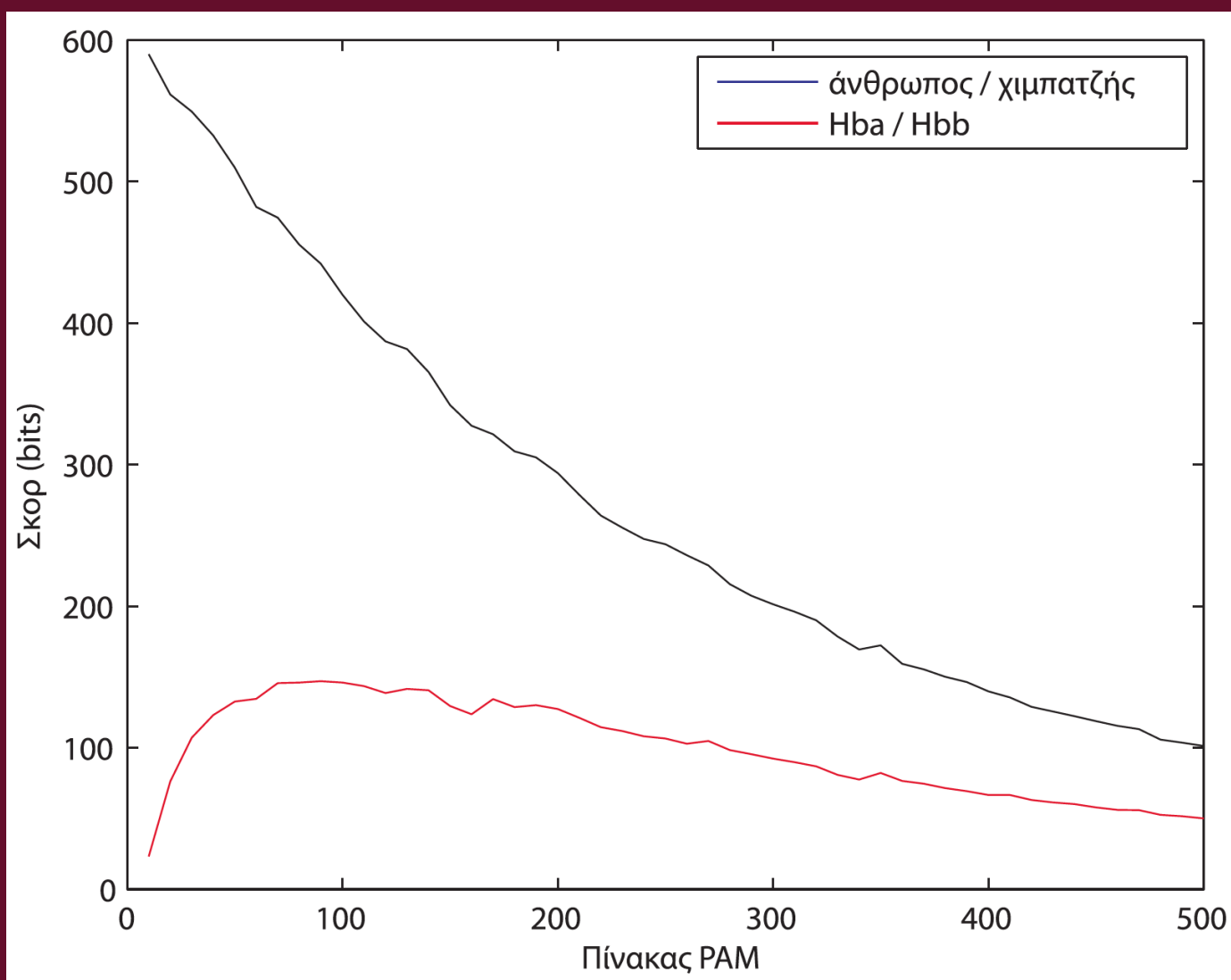
Εικόνα 3.13 Ο πίνακας πιθανότητας μεταλλαγής PAM250. Σε αυτή την εξελικτική απόσταση, μόνο ένα στα πέντε αμινοξέα παραμένει αμετάβλητο. Οι στήλες του πίνακα αντιστοιχούν στο αρχικό αμινοξύ, οι γραμμές στο αμινοξύ αντικατάστασης και η κλίμακα είναι τέτοια ώστε οι αριθμητικές τιμές κάθε στήλης να έχουν άθροισμα 100.

A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	-2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Εικόνα 3.14 Ο πίνακας λογαριθμικού λόγου (συμπληρωματικών) πιθανοτήτων PAM250. Οι υψηλές τιμές PAM (π.χ. PAM250) είναι χρήσιμες για τη στοίχιση αλληλουχιών που αποκλίνουν σημαντικά. Μια ποικιλία αλγορίθμων για στοίχιση κατά ζεύγη, για πολλαπλή στοίχιση και για την αναζήτηση σε βάσεις δεδομένων (π.χ. BLAST) επιτρέπει την επιλογή από μια ποικιλία πινάκων PAM, όπως οι PAM250, PAM70 και PAM30.

A	7																			
R	-10	9																		
N	-7	-9	9																	
D	-6	-17	-1	8																
C	-10	-11	-17	-21	10															
Q	-7	-4	-7	-6	-20	9														
E	-5	-15	-5	0	-20	-1	8													
G	-4	-13	-6	-6	-13	-10	-7	7												
H	-11	-4	-2	-7	-10	-2	-9	-13	10											
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9										
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7									
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7								
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12							
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9						
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8					
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7				
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8			
W	-2	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-19	13		
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-8	10	
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Εικόνα 3.15 Ο πίνακας λογαριθμικού λόγου (συμπληρωματικών) πιθανοτήτων PAM10. Οι πίνακες με χαμηλές τιμές PAM, όπως αυτός, είναι χρήσιμοι για τη στοίχιση στενά συγγενικών αλληλουχιών. Συγκρίνετε τον PAM10 με τον PAM250 (Εικόνα 3.14) και σημειώστε ότι στον PAM10 υπάρχουν μεγαλύτερες θετικές βαθμολογίες για τις ταυτοσημίες και μεγαλύτερες ποινές για τις αναντιστοιχίες.



Εικόνα 3.16 Βαθμολογία ολικής στοίχισης κατά ζεύγη χρησιμοποιώντας μια σειρά πινάκων PAM. Δύο στενά συγγενικές σφαιρίνες (η β-σφαιρίνη του ανθρώπου και του χιμπατζή, μαύρη γραμμή) στοιχίστηκαν χρησιμοποιώντας μία σειρά από πίνακες PAM (άξονας x) και προσδιορίστηκαν τα bit σκορ (άξονας y). Όταν στοιχίζονται δύο σφαιρίνες με απομακρυσμένη συγγένεια (η ανθρώπινη α-σφαιρίνη με την ανθρώπινη β-σφαιρίνη, κόκκινη γραμμή), τα bit σκορ είναι μικρότερα στους χαμηλούς PAM (όπως οι PAM1 έως PAM20), διότι οι αναντιστοιχίες επισείουν υψηλές ποινές.

A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-1	1	1	-2	-1	-3	-2	5								
M	-1	-2	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Εικόνα 3.17 Ο πίνακας βαθμολόγησης BLOSUM62 των Henikoff και Henikoff (1992). Για την κατασκευή αυτού του πίνακα έχουν συγχωνευτεί σε μία αλληλουχία όλες οι πρωτεΐνες που έχουν 62% ταύτιση αμινοξέων ή μεγαλύτερη. Ο BLOSUM62 είναι καλύτερος σε σχέση με άλλους πίνακες BLOSUM ή πίνακες PAM για την ανίχνευση της ομολογίας μεταξύ πρωτεϊνών που έχουν αποκλίνει σημαντικά και έτσι εμφανίζουν σχετικά περιορισμένη ομοιότητα. Για τον λόγο αυτό είναι ο προεπιλεγμένος πίνακας βαθμολόγησης στα περισσότερα προγράμματα αναζήτησης σε βάσεις δεδομένων, για παράδειγμα στο BLAST (Κεφάλαιο 4).

BLOSUM90

PAM30

Λίγο
αποκλίνουσες
αλληλουχίες

Σύγκριση της β-σφαιρίνης του
ανθρώπου με αυτή του χιμπατζή

BLOSUM62

PAM120

BLOSUM45

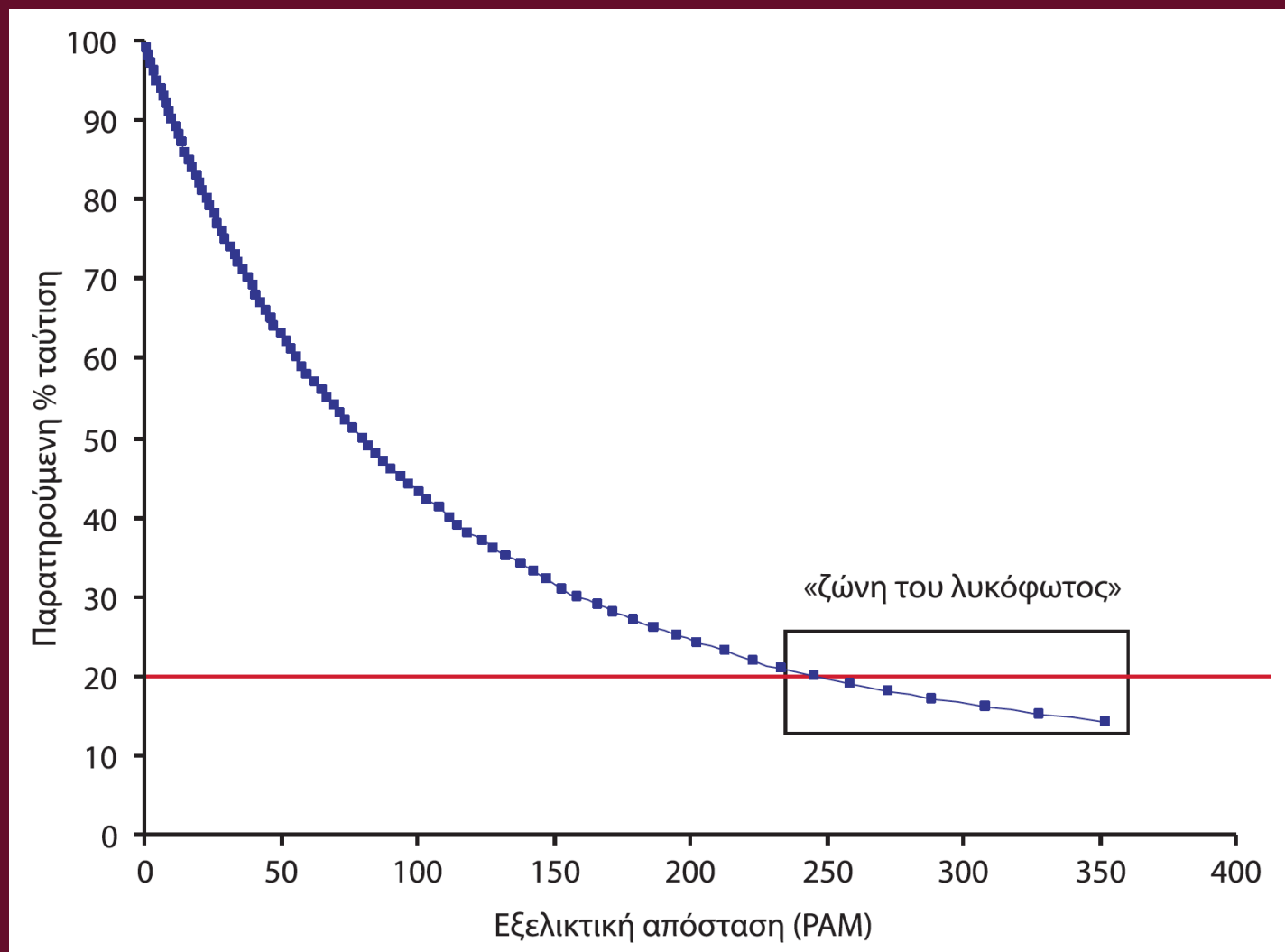
PAM250

Πολύ
αποκλίνουσες
αλληλουχίες

Σύγκριση των σφαιρινών του
ανθρώπου με αυτές των βακτηρίων



Εικόνα 3.18 Περίληψη των πινάκων PAM και BLOSUM. Οι υψηλής τιμής πίνακες BLOSUM και οι χαμηλής τιμής πίνακες PAM είναι οι πλέον κατάλληλοι για τη σύγκριση καλά συντηρημένων πρωτεϊνών όπως η β-σφαιρίνη του ποντικού και του αρουραίου. Οι χαμηλής τιμής πίνακες BLOSUM (π.χ. BLOSUM45) και οι υψηλής τιμής πίνακες PAM είναι οι πλέον κατάλληλοι για την ανίχνευση της ομολογίας μεταξύ πρωτεϊνών που έχουν αποκλίνει σημαντικά. Θυμηθείτε ότι σε έναν πίνακα BLOSUM45 όλα τα μέλη μιας οικογένειας πρωτεϊνών με ταύτιση αμινοξέων μεγαλύτερη από 45% συγχωνεύονται, κάνοντας τον πίνακα κατάλληλο για μελέτη πρωτεϊνών με ταύτιση μικρότερη από 45%.

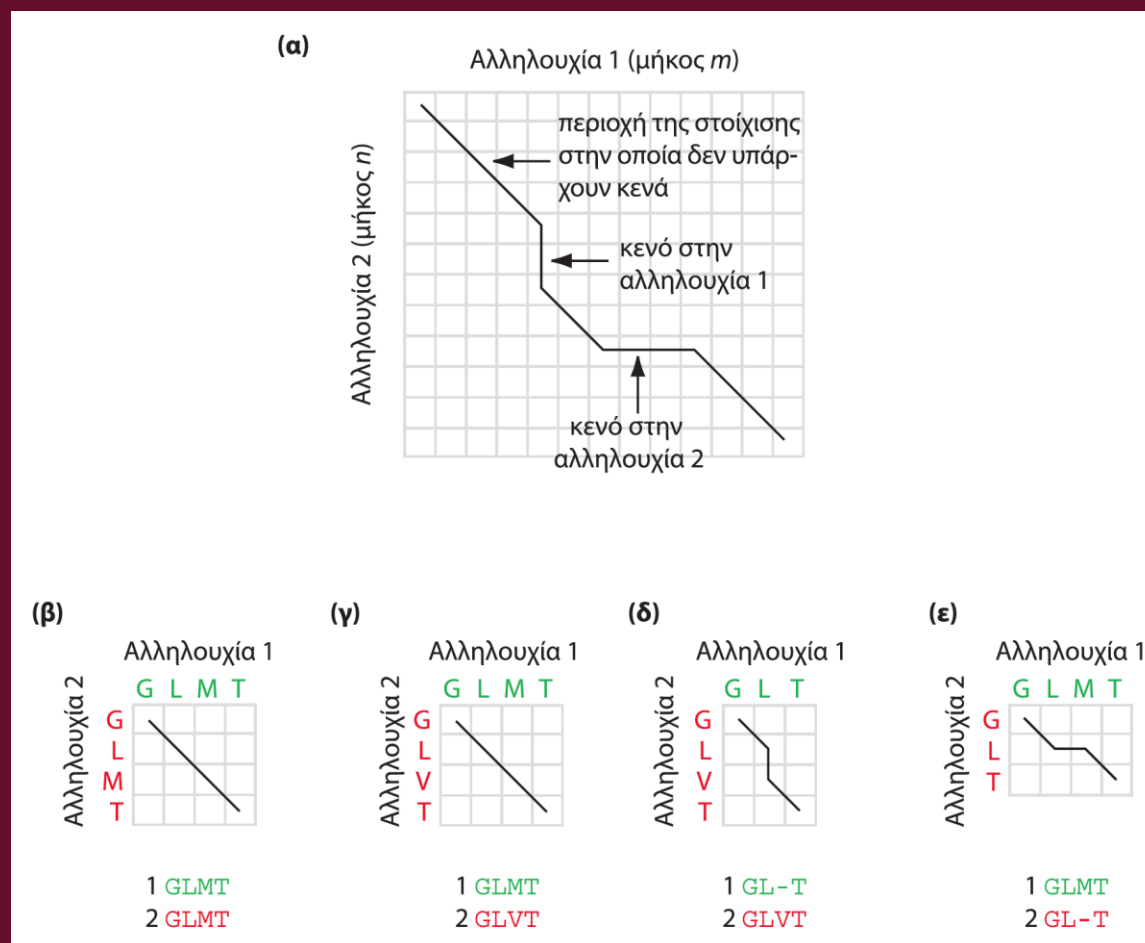


Εικόνα 3.19 Δύο τυχαία αποκλίνουσες αλληλουχίες πρωτεϊνών αλλάζουν με αρνητικά εκθετικό τρόπο. Αυτή η γραφική παράσταση δείχνει τον παρατηρούμενο αριθμό ταυτώσεων αμινοξέων ανά 100 αμινοξέα των δύο αλληλουχιών (άξονας y) σε σχέση με τον αριθμό των αλλαγών που αναμένεται να έχουν συμβεί (δηλαδή την εξελικτική τους απόσταση σε μονάδες PAM). Ως «ζώνη του λευκόφωτος» (Doolittle, 1987) αναφέρεται η εξελικτική απόσταση που αντιστοιχεί σε περίπου 20% ταύτιση μεταξύ δύο πρωτεϊνών. Οι πρωτεΐνες με αυτό το ποσοστό ταύτισης μπορεί να είναι ομόλογες, αλλά η ομολογία τους είναι δύσκολο να ανιχνευθεί.

Πίνακας 3.3 Σχέση μεταξύ της εξελικτικής απόστασης και του παρατηρούμενου αριθμού αμινοξικών διαφορών ανά 100 κατάλοιπα σε δύο στοιχισμένες αλληλουχίες πρωτεϊνών. Δίνεται σε μονάδες PAM ο αριθμός των αλλαγών που θα πρέπει να έχουν συμβεί.

Παρατηρούμενος αριθμός αμινοξικών διαφορών ανά 100 κατάλοιπα	Εξελικτική απόσταση σε PAM
1	1
5	5,1
10	10,7
15	16,6
20	23,1
25	30,2
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246

Πηγή: Dayhoff (1972). Αναδημοσιεύεται κατόπιν αδείας του National Biomedical Research Foundation.



Εικόνα 3.20 Ολική στοίχιση δύο πρωτεϊνικών αλληλουχιών χρησιμοποιώντας τον αλγόριθμο δυναμικού προγραμματισμού των Needleman και Wunsch (1970). (α) Η στοίχιση ανάμεσα στις δύο αλληλουχίες αντιστοιχεί σε μια διαγώνια διαδρομή στον πίνακα, η οποία, όταν είναι απαραίτητο να εισαχθούν κενά, αποκλίνει οριζόντια ή κάθετα. (β) Η στοίχιση ανάμεσα σε δύο πανομοιότυπες αλληλουχίες αντιστοιχεί σε μία χωρίς αποκλίσεις διαγώνια διαδρομή στον πίνακα. (γ) Αν υπάρχει αναντιστοιχία (ή αναντιστοιχίες), η διαδρομή εξακολουθεί να μην αποκλίνει από τη διαγώνιο, ωστόσο ένα σύστημα βαθμολόγησης επιβάλλει ποινές στις αναντιστοιχίες. Αν η στοίχιση περιλαμβάνει ένα κενό στην πρώτη αλληλουχία (δ) ή στη δεύτερη αλληλουχία (ε), τότε στη διαδρομή εισέρχεται μια κάθετη ή μια οριζόντια γραμμή.

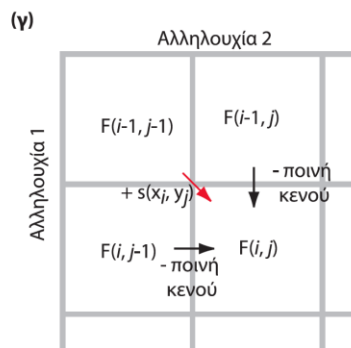
(α) Αλληλουχία 2

	F	M	D	T	P	L	N	E
0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2							
K	-4							
H	-6							
M	-8							
E	-10							
D	-12							
P	-14							
L	-16							
E	-18							

(β)

$$\text{Σκορ} = \text{Max} \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - \text{ποινή εισαγωγής κενού} \\ F(i, j-1) - \text{ποινή εισαγωγής κενού} \end{cases}$$

Σκορ (στο παράδειγμα αυτό) = +1 (ταυτοσημίες)
 -2 (αναντιστοιχίες)
 -2 (ποινή κενού)



(δ) Αλληλουχία 2

	F	M
0	0	-2
F	-2	-4
K	-4	-4

(ε) Αλληλουχία 2

	F	M
0	0	-2
F	-2	-4
K	-4	-4

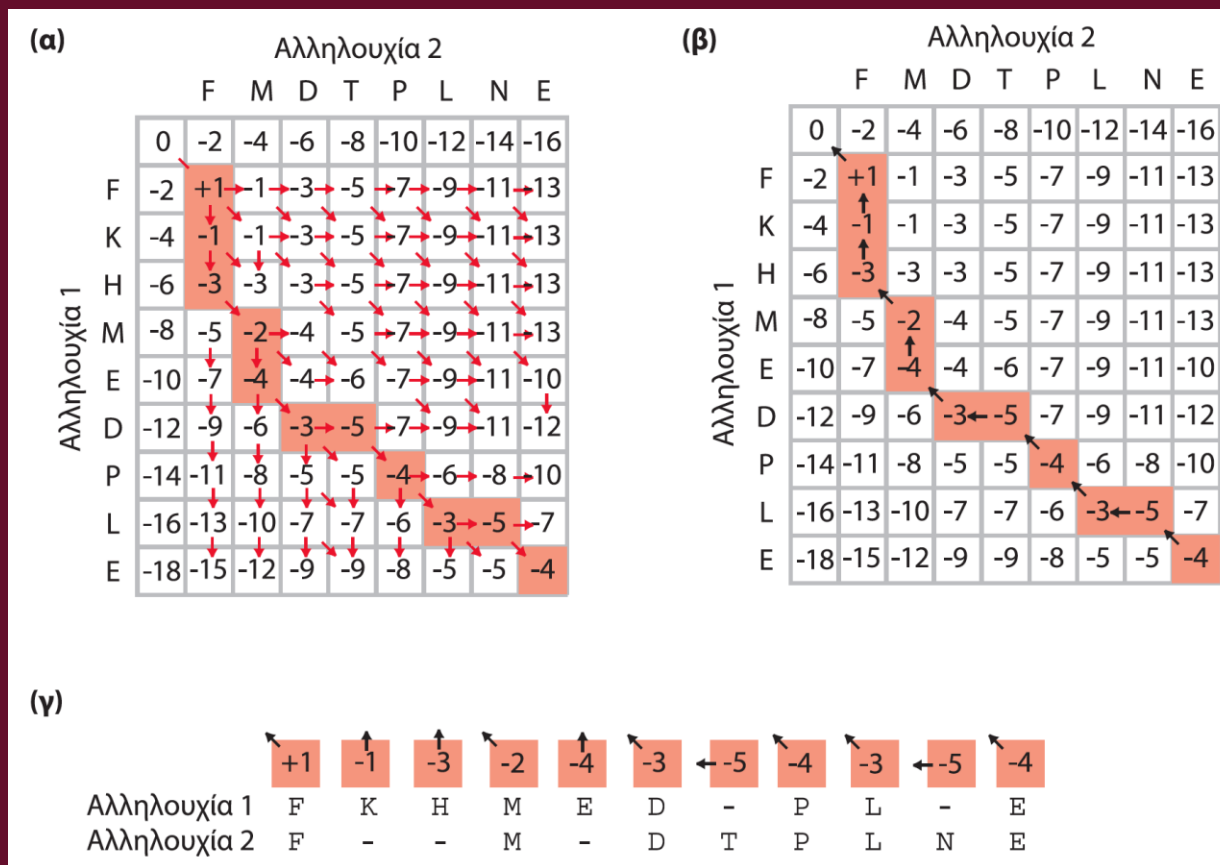
(στ) Αλληλουχία 2

	F	M	D	T	P	L	N	E
0	0	-2	-4	-6	-8	-10	-12	-14
F	-2	+1	-1	-3	-5	-7	-9	-11
K	-4	-1	-1	-3	-5	-7	-9	-11
H	-6							
M	-8							
E	-10							
D	-12							
P	-14							
L	-16							
E	-18							

(ζ) Αλληλουχία 2

	F	M	D	T	P	L	N	E
0	0	-2	-4	-6	-8	-10	-12	-14
F	-2	+1	-1	-3	-5	-7	-9	-11
K	-4	-1	-1	-3	-5	-7	-9	-11
H	-6	-3	-3	-3	-5	-7	-9	-11
M	-8	-5	-2	-4	-5	-7	-9	-11
E	-10	-7	-4	-4	-6	-7	-9	-11
D	-12	-9	-6	-3	-5	-7	-9	-11
P	-14	-11	-8	-5	-5	-4	-6	-8
L	-16	-13	-10	-7	-7	-6	-3	-5
E	-18	-15	-12	-9	-9	-8	-5	-4

Εικόνα 3.21 Ολική στοίχιση δύο αλληλουχιών χρησιμοποιώντας τον αλγόριθμο δυναμικού προγραμματισμού των Needleman και Wunsch (1970). (α) Για δύο αλληλουχίες μήκους m και n σχηματίζουμε έναν πίνακα διαστάσεων $m + 1$ με $n + 1$ και εισάγουμε τις ποινές εισαγωγής κενών στην πρώτη σειρά και στην πρώτη στήλη του πίνακα, όπως φαίνεται στην εικόνα. Κάθε κενό έχει βαθμολογικό κόστος -2 . Τα κελιά που αντιστοιχούν σε θέσεις όπου οι δύο αλληλουχίες έχουν το ίδιο αμινοξύ είναι σκιασμένα με γκρι χρώμα. (β) Το σύστημα βαθμολόγησης για αυτό το παράδειγμα είναι $+1$ για τις ταυτοσημίες, -2 για τις αναντιστοιχίες και -2 για τα κενά. Για κάθε θέση του πίνακα η βαθμολογία υπολογίζεται με την ίδια μέθοδο, η οποία βασίζεται στο να βρούμε την υψηλότερη βαθμολογία μεταξύ τριών πιθανών σκορ. (γ) Για κάθε θέση $F(i, j)$ υπολογίζουμε τις βαθμολογίες που προκύπτουν ακολουθώντας τη διαδρομή από το άνω αριστερά κελί [και προσθέτουμε το σκορ αυτού του κελιού στο σκορ του $F(i, j)$], από το αριστερό κελί (και αφαιρούμε το κόστος εισαγωγής κενού) και από το κελί ακριβώς από πάνω (αφαιρώντας πάλι το κόστος εισαγωγής κενού). (δ) Για να υπολογίσουμε το σκορ στο κελί της δεύτερης σειράς και στήλης, λαμβάνουμε το μέγιστο των τριών βαθμών $+1$, -4 , -4 . Η καλύτερη βαθμολογία ($+1$) αντιστοιχεί στη διαδρομή του κόκκινου βέλους. Διατηρούμε την πληροφορία της καλύτερης διαδρομής για το σκορ κάθε κελιού προκειμένου να προσ-διορίσουμε αργότερα τη στοίχιση. (ε) Για να υπολογίσουμε τη βαθμολογία στη δεύτερη σειρά, τρίτη στήλη, παίρνουμε πάλι το μέγιστο των τριών βαθμών -4 , -1 , -4 . Η καλύτερη βαθμολογία προκύπτει από το αριστερό κελί (κόκκινο βέλος). (στ) Συνεχίζουμε να συμπληρώνουμε τις βαθμολογίες στην πρώτη σειρά του πίνακα. (ζ) Ο συμπληρωμένος πίνακας περιλαμβάνει τη βαθμολογία της βέλτιστης στοίχισης (-4 , βλ. κελί κάτω δεξιά, που αντιστοιχεί στο καρβοξυτελικό άκρο των πρωτεϊνών). Τα κόκκινα βέλη υποδεικνύουν τη διαδρομή με την οποία είχαμε βρει την υψηλότερη βαθμολογία για κάθε κελί.



Εικόνα 3.22 Ολική στοίχιση δύο αλληλουχιών με αλγορίθμους δυναμικού προγραμματισμού: υπολογισμός της υψηλότερης βαθμολογίας και προσδιορισμός της βέλτιστης στοίχισης μέσω της διαδικασίας αντιστροφής της διαδρομής. (α) Ο πίνακας της Εικόνα 3.21ζ. Τα πορτοκαλί κελιά αντιστοιχούν στα κελιά εκείνα μέσω των οποίων φτάνουμε στη βέλτιστη βαθμολογία (κάτω δεξιά). (β) Μια ισοδύναμη αναπαράσταση στην οποία τα βέλη δείχνουν την (προς τα πίσω) πηγή της βέλτιστης βαθμολογίας κάθε κελιού. (γ) Αυτή η διαδικασία αντιστροφής της διαδρομής μάς επιτρέπει να προσδιορίσουμε τη βέλτιστη στοίχιση. Τα κάθετα και οριζόντια βέλη αντιστοιχούν στις θέσεις στις οποίες πρέπει να μπουν κενά στη στοίχιση, ενώ οι διαγώνιες γραμμές αντιστοιχούν σε ταυτοσημίες (ή σε αναντιστοιχίες). Η τελική βαθμολογία (-4) ισούται με το άθροισμα των τιμών που προέρχονται από τις ταυτοσημίες ($6 \times 1 = 6$), τις αναντιστοιχίες (καμία σε αυτό το παράδειγμα) και τα κενά ($5 \times -2 = -10$).

(α)

NP_824492.1	1	MCGDMTVHTVEYIRYRIPEQQSAEFLAAYTRAAQQLAAAPQCVDYELARC	50
NP_337032.1	1		0
NP_824492.1	51	EEDFEHFVLRITWTSTEDHIEGFRKSELFDFLAEIRPYISSIEEMRHYK	100
NP_337032.1	1		0
NP_824492.1	101	PTTVRGTGA [▼] AVPTLYAWAGGAEAFARLTEVFYEKVLKDDVLAPVFEGMAP	150
		:.: ... : .:.:. . .: .: .. :	
NP_337032.1	1	MEGMDQMPKSFYDAVGGAKTFDAIVSRFYAQVAEDEVLRVY----P	43
NP_824492.1	151	EH-----AAHVALWLGEVFGGPAAYSETQGGHGHMVAKHLGKNITEVQRR	195
	: .: ::.: :..: . .	
NP_337032.1	44	EDDLAGAEERLRMFLEQYWGGPRTYSE-QRGHPRLRMRHAPFRISLIERD	92
NP_824492.1	196	RWVNLLQDAADDAGLPT-DAEFRSAFLAYA [▼] EWGTRLAVYFSGPDAVPPAE [▼]	244
		. :..:.. :. .	
NP_337032.1	93	AWLRCMHTAVASIDSETLDDEHRRELLDYLEMAAHSLV--NSPF	134
NP_824492.1	245	QPVPQWSWGAMPPYQP	260
NP_337032.1	135		134

Εικόνα 3.23 (α) Ολική στοίχιση πρωτεϊνών που περιέχουν επικράτειες σφαιρίνης και προέρχονται από τα βακτήρια *Streptomyces avermitilis* MA-4680 (NP_824492) και *Mycobacterium tuberculosis* CDC1551 (NP_337032). Ως πίνακας βαθμολόγησης χρησιμοποιήθηκε ο BLOSUM62. Οι στοιχισμένες πρωτεΐνες παρουσιάζουν 14,7% ταύτιση (39/266 στοιχισμένα αμινοξέα), 22,6% ομοιότητα (60/266) και 51,9% κενά (138/266).

(β)

NP_824492.1	113	<div> <div>TLYAWAGGAEAFARL</div> <div>TEVFY</div> <div>EKVLKDDV</div> <div>LAPVFEGMAPEH</div> <div>----</div> <div>AAHVA</div> </div> <div> <div>.. ... </div> <div> :.. </div> <div>:... </div> <div> :.. </div> <div>: .. </div> <div>:</div> <div> .</div> <div>....:</div> </div>	157
NP_337032.1	10	<div> <div>SFYDAVGGAKTFDA</div> <div>IVSRFYA</div> <div>QVAEDEVL</div> <div>RRVY----</div> <div>PEDDLAGAEERLR</div> </div> <div> <div>.. ... </div> <div> :.. </div> <div>:... </div> <div> :.. </div> <div>:</div> <div> .</div> <div>....:</div> </div>	55
NP_824492.1	158	<div> <div>LWLGEVFGGPAAY</div> <div>SETQGGHGHM</div> <div>VAKHLGKN</div> <div>ITEVQRRRWVN</div> <div>LLQDAADD</div> </div> <div> <div>.. ... </div> <div> :.. </div> <div>:... </div> <div> :.. </div> <div>:</div> <div> .</div> <div>....:</div> </div>	207
NP_337032.1	56	<div> <div>MFLEQYWGGPRTY</div> <div>SE-QRGHPRL</div> <div>MRHAPFRIS</div> <div>LIERDAWLRC</div> <div>MHTAVAS</div> </div> <div> <div>.. ... </div> <div> :.. </div> <div>:... </div> <div> :.. </div> <div>:</div> <div> .</div> <div>....:</div> </div>	104
NP_824492.1	208	<div> <div>AGLPT-DAEFRSA</div> <div>FLAYAE</div> </div> <div> <div>.... </div> <div> .. </div> <div>... </div> <div>.. </div> </div> <div>225</div>	
NP_337032.1	105	<div> <div>IDSETLDDEHR</div> <div>RELLDYLE</div> </div> <div> <div>.... </div> <div> .. </div> <div>... </div> <div>.. </div> </div> <div>123</div>	

Εικόνα 3.23 (β) Οι ίδιες αλληλουχίες αλλά μετά από τοπική στοίχιση. Προσέξτε ότι η τοπική στοίχιση δεν περιλαμβάνει τα αμινοτελικά και τα καρβοξυτελικά άκρα των πρωτεϊνών. Η ταύτιση ανέρχεται σε 30% (35/115 στοιχισμένα αμινοξέα). Η στοίχιση στο (β) αντιστοιχεί στη σκιασμένη περιοχή του (α). Τα τρία βέλη στο (α) δείχνουν τη θέση ταυτόσημων αμινοξέων που παρ’ όλα αυτά δεν περιέχονται στην τοπική στοίχιση. Κατά την πραγματοποίηση τοπικών στοιχίσεων (όπως γίνεται στο BLAST, Κεφάλαιο 4), ορισμένες περιοχές μεταξύ των οποίων υπάρχουν ομοιότητες μπορεί να μη συμπεριληφθούν στην τελική στοίχιση.

(α)

Αλληλουχία 1

C A G C C U C G C U U A G

Αλληλουχία 2

	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
A	0,0	0,0	1,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,0	0,0
A	0,0	0,0	1,0	0,7	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,0	0,7
U	0,0	0,0	0,0	0,7	0,3	0,0	1,0	0,0	0,0	0,0	1,0	1,0	0,0	0,7
G	0,0	0,0	0,0	1,0	0,3	0,0	0,0	0,7	1,0	0,0	0,0	0,7	0,7	1,0
C	0,0	1,0	0,0	0,0	2,0	1,3	0,3	1,0	0,3	2,0	0,7	0,3	0,3	0,3
C	0,0	1,0	0,7	0,0	1,0	3,0	1,7	1,3	1,0	1,3	1,7	0,3	0,0	0,0
A	0,0	0,0	2,0	0,7	0,3	1,7	2,7	1,3	1,0	0,7	1,0	1,3	1,3	0,0
U	0,0	0,0	0,7	1,7	0,3	1,3	2,7	2,3	1,0	0,7	1,7	2,0	1,0	1,0
U	0,0	0,0	0,3	0,3	1,3	1,0	2,3	2,3	2,0	0,7	1,7	2,7	1,7	1,0
G	0,0	0,0	0,0	1,3	0,0	1,0	1,0	2,0	3,3	2,0	1,7	1,3	2,3	2,7
A	0,0	0,0	1,0	0,0	1,0	0,3	0,7	0,7	2,0	3,0	1,7	1,3	2,3	2,0
C	0,0	1,0	0,0	0,7	1,0	2,0	0,7	1,7	1,7	3,0	2,7	1,3	1,0	2,0
G	0,0	0,0	0,7	1,0	0,3	0,7	1,7	0,3	2,7	1,7	2,7	2,3	1,0	2,0
G	0,0	0,0	0,0	1,7	0,7	0,3	0,3	1,3	1,3	2,3	1,3	2,3	2,0	2,0

(β)

Αλληλουχία 1

GCC-UCG

Αλληλουχία 2

GCCAUUG

(γ)

Αλληλουχία 1

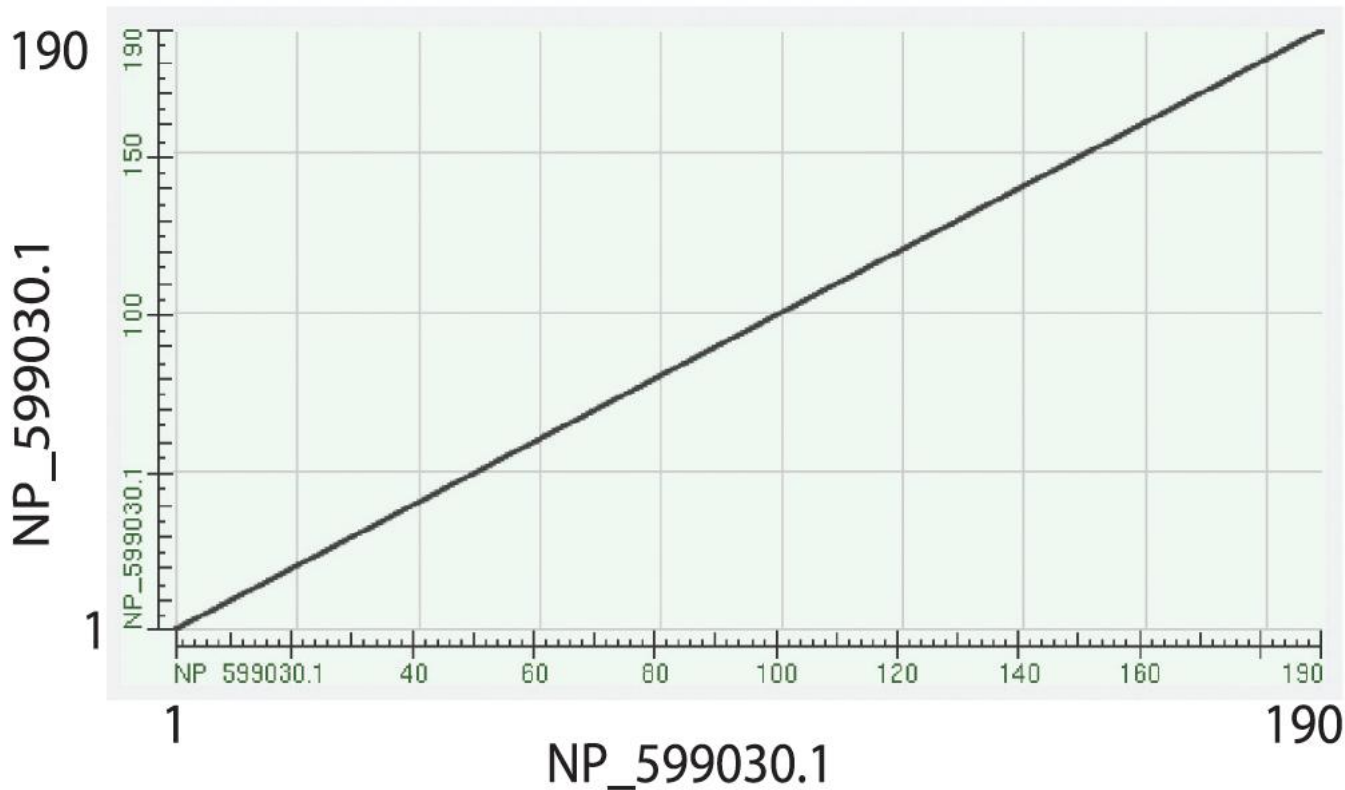
CA-GCC-UCGCUUAG

Αλληλουχία 2

AAUGCCAUUGACG-G

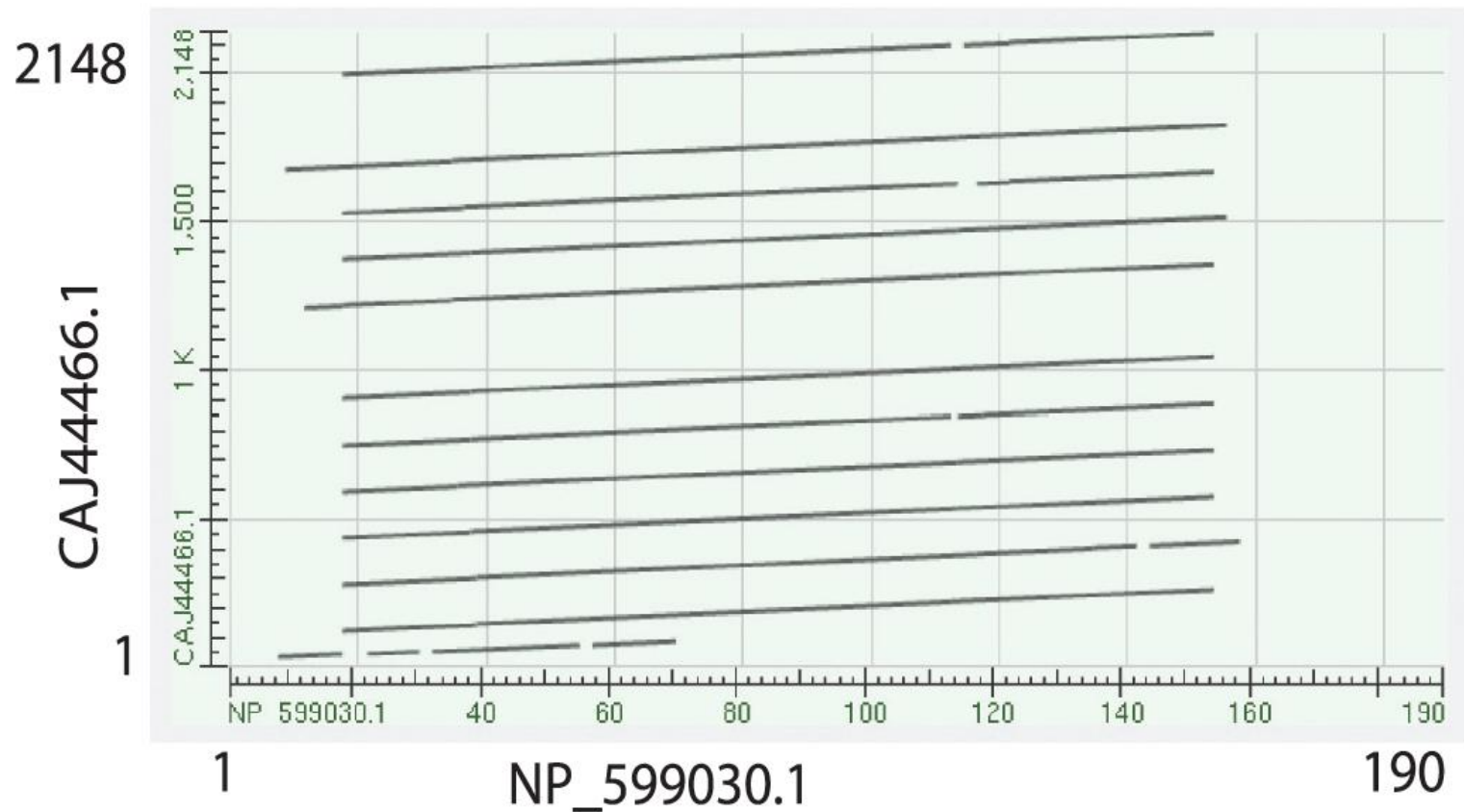
Εικόνα 3.24 Η μέθοδος τοπικής στοίχισης των Smith και Waterman (1981). (α) Σε αυτό το παράδειγμα, ο πίνακας κατασκευάζεται με δύο αλληλουχίες RNA (CAGCCUCGCUUAG και AAUGCCAUUGACGG). Αν και αυτός δεν είναι ένας πίνακας ταύτισης (όπως της Εικόνας 3.21α), έχουμε σημειώσει τις θέσεις ταυτόσημων νουκλεοτιδίων (με ροζ στην περιοχή της τοπικής στοίχισης, με γκρι εκτός αυτής). Το σύστημα βαθμολόγησης εδώ περιλαμβάνει τρεις τιμές: +1 για τις ταυτοσημίες, $-1/3$ για τις αναντιστοιχίες, ενώ το κόστος εισαγωγής κενού είναι η διαφορά μεταξύ μίας ταυτοσημίας και μίας αναντιστοιχίας ($-1,3$ για ένα κενό μήκους ένα). Η βέλτιστη βαθμολογία είναι το μέγιστο μεταξύ τριών πιθανών τιμών και της τιμής 0 (ώστε να μην μπορεί να υπάρξουν αρνητικές βαθμολογίες). Η υψηλότερη τιμή του πίνακα (εδώ 3,3) υποδεικνύει τη θέση στην οποία τελειώνει η βέλτιστη τοπική στοίχιση και τα στοιχισμένα αμινοξέα (πράσινο χρώμα) εκτείνονται προς τα πάνω και προς τα αριστερά μέχρι να φτάσουμε σε ένα κελί με τιμή μηδέν. (β) Η τοπική στοίχιση που προκύπτει από αυτόν τον πίνακα. Σημειώστε ότι περιλαμβάνει ταυτοσημίες, αναντιστοιχίες και κενά. (γ) Μια ολική στοίχιση των δύο αλληλουχιών η οποία παρουσιάζεται για λόγους σύγκρισης με την τοπική στοίχιση.

(α) Σύγκριση της ανθρώπινης κυτοσφαιρίνης με τον εαυτό της



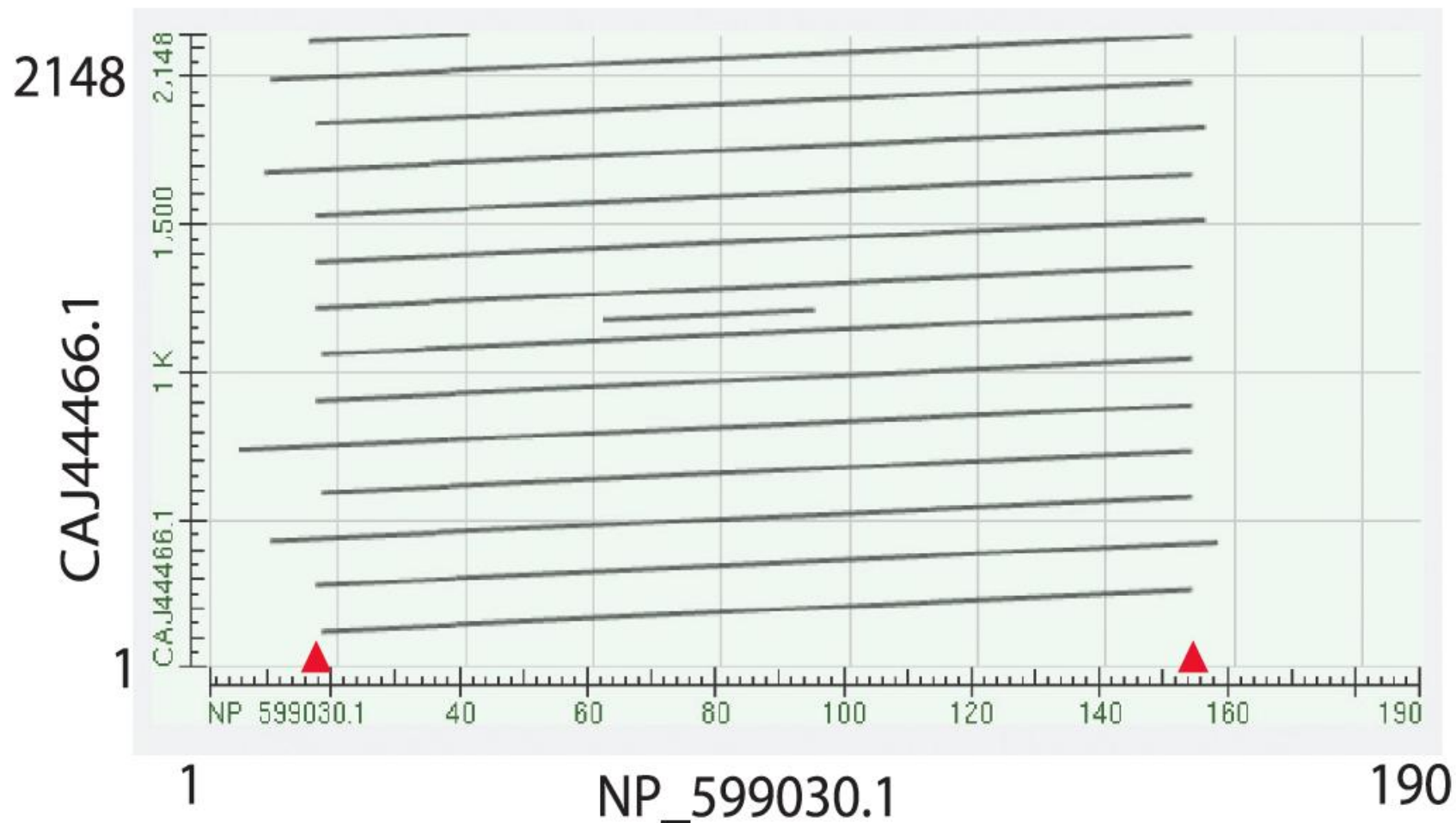
Εικόνα 3.25 Τα διαγράμματα κουκκίδων που το BLAST περιλαμβάνει στην παρουσίαση των αποτελεσμάτων του για τις κατά ζεύγη στοιχίσεις αποτελούν μια απεικόνιση στην οποία είναι εύκολο να εντοπίσει κανείς τις περιοχές των δύο πρωτεϊνών που εμφανίζουν τη μεγαλύτερη ομοιότητα. Το BLAST χρησιμοποιείται όπως περιγράφεται στην Εικόνα 3.4. (α) Στη σύγκριση της ανθρώπινης κυτοσφαιρίνης (NP_599030.1, μήκος 190 αμινοξέα) με τον εαυτό της, το διάγραμμα κουκκίδων έχει την αλληλουχία της κυτοσφαιρίνης και στους δύο άξονές του και τα σημεία δεδομένων (data points) που υποδεικνύουν τις θέσεις ταυτοσημίας σχηματίζουν μια διαγώνια γραμμή.

(β) Σύγκριση της ανθρώπινης κυτοσφαιρίνης με μία αιμοσφαιρίνη του σαλιγκαριού *B. glabrata* (BLOSUM62)



(β) Στη σύγκριση της ανθρώπινης κυτοσφαιρίνης με μια σφαιρίνη από το σαλιγκάρι *Biomphalaria glabrata* (καταχώριση CAJ44466.1, μήκος 2.148 αμινοξέα) αποκαλύπτεται ότι η αλληλουχία της κυτοσφαιρίνης (άξονας x) έχει 12 περιοχές ομοιότητας με ισάριθμες εσωτερικές επαναλήψεις της πρωτεΐνης του σαλιγκαριού. Σε αυτή τη σύγκριση έχει χρησιμοποιεί ο προεπιλεγμένος πίνακας βαθμολόγησης BLOSUM62.

(γ) Σύγκριση της ανθρώπινης κυτοσφαιρίνης με μία αιμοσφαιρίνη του σαλιγκαριού *B. glabrata* (PAM250)



Εικόνα 3.25 (γ) Αν χρησιμοποιηθεί ως πίνακας βαθμολόγησης ο PAM250, εμφανίζονται 13 επαναλήψεις της σφαιρίνης του σαλιγκαριού οι οποίες ταιριάζουν με την αλληλουχία της κυτοσφαιρίνης.

(δ) Στοίχιση κατά ζεύγη της ανθρώπινης κυτοσφαιρίνης με μία επανάληψη της αιμοσφαιρίνης του σαλιγκαριού

haemoglobin type 1 [Biomphalaria glabrata]

Sequence ID: [emb|CAJ44466.1|](#) Length: 2148 Number of Matches: 15

Range 1: 1529 to 1669 [GenPept](#) [Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps
55.0 bits(189)	4e-13	Composition-based stats.	36/141(26%)	83/141(58%)	4/141(2%)
Query 18	ELSEAERKAVQAMWARLYANCEDV---GVAILVRFFVNFPSAKQYFSQFKHMEDPLEMER				74
	LSE++R+A+++ W RL A ++V GV ++++FF N+P+ ++ F++F + +				
Sbjct 1529	GLSETDRRALDSSWKRLTAGENGVOKAGVNLVWFFNNIPNMRERFTKFDANQADDALRA				1588
Query 75	SPQLRKHACRVMGALNTVVENLHDPDKVSSVLALVGKAH-ALKHKVEPVYFKILSGVILE				133
	P+++K+ ++G+L++ +++++DP + + + V+ AH ++ V YF LS I				
Sbjct 1589	DPEFQKQVNVIVGGLKSFLDSVNDPIALQANMDRVAEHLSDMPVVGVPYFSALSQNIHR				1648
Query 134	VVAEEFASDFPPETQRAWAKL				154
	+ ++ ++ +AW+ L				
Sbjct 1649	FIEISLGVITADSDESQAWTDL				1669

(δ) Μια στοίχιση κατά ζεύγη των δύο αλληλουχιών δείχνει ότι οι επαναλήψεις σφαιρίνης του σαλιγκαριού στοιχίζονται με τα αμινοξέα 18-154 της κυτοσφαιρίνης. Αυτό αντικατοπτρίζεται και στα διαγράμματα κουκκίδων, όπου φαίνεται πως το τμήμα στον άξονα x που αντιστοιχεί στα αμινοξέα της κυτοσφαιρίνης 1-17 και 155-190 (κόκκινες κεφαλές βέλους στο γ) δε στοιχίζεται με την αλληλουχία του σαλιγκαριού. Το BLASTP παράγει ένα ολόκληρο σύνολο εναλλακτικών στοιχίσεων, από τις οποίες μόνο η πρώτη παρουσιάζεται εδώ.

Πληροφορία σύμφωνα με την πλέον αξιόπιστη μέθοδο ανάλυσης
(π.χ. την ομοιότητα σε επίπεδο τρισδιάστατης δομής)

Οι αλληλουχίες είναι
ομόλογες

Οι αλληλουχίες δεν
είναι ομόλογες

Αποτέλεσμα της
στοίχισης: οι αλλη-
λουχίες είναι συγγε-
νικές

Αληθώς
θετικές

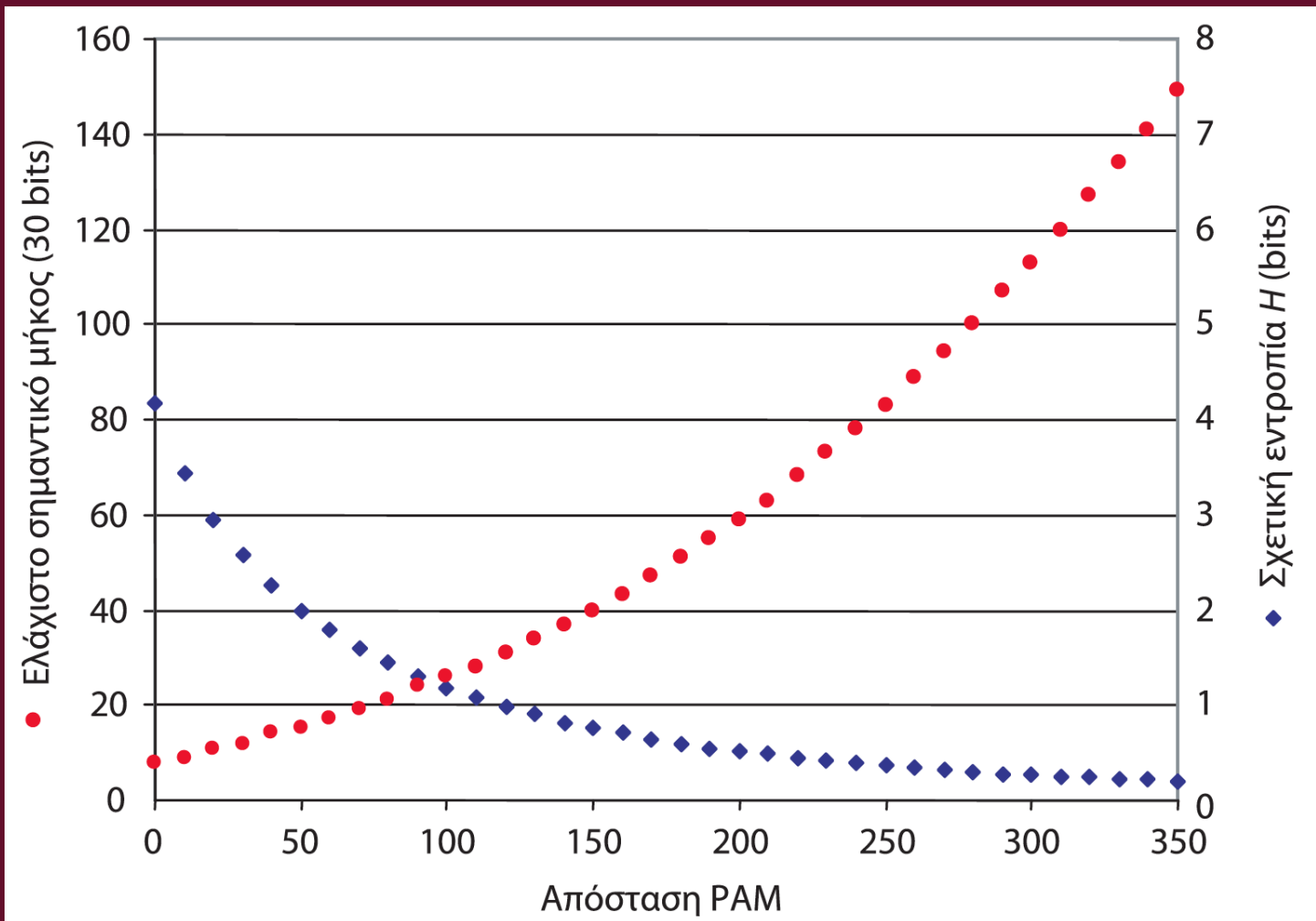
Ψευδώς
θετικές

Αποτέλεσμα της
στοίχισης: οι αλλη-
λουχίες δεν είναι
συγγενικές (ή δεν
εντοπίζεται συγγε-
νική αλληλουχία)

Ψευδώς
αρνητικές

Αληθώς
αρνητικές

Εικόνα 3.26 Οι στοιχίσεις των αλληλουχιών μπορούν να ταξινομηθούν ως αληθείς ή ψευδείς και ως θετικές ή αρνητικές. Οι στατιστικές αναλύσεις αποτελούν την κύρια μέθοδο αξιολόγησης του κατά πόσο μια στοίχιση είναι, για παράδειγμα, αληθώς θετική (δηλαδή οι δύο στοιχισμένες πρωτεΐνες είναι πραγματικά ομόλογες). Στην ιδανική περίπτωση, ένας αλγόριθμος στοίχισης μεγιστοποιεί τόσο την ευαισθησία όσο και την ειδικότητά του.



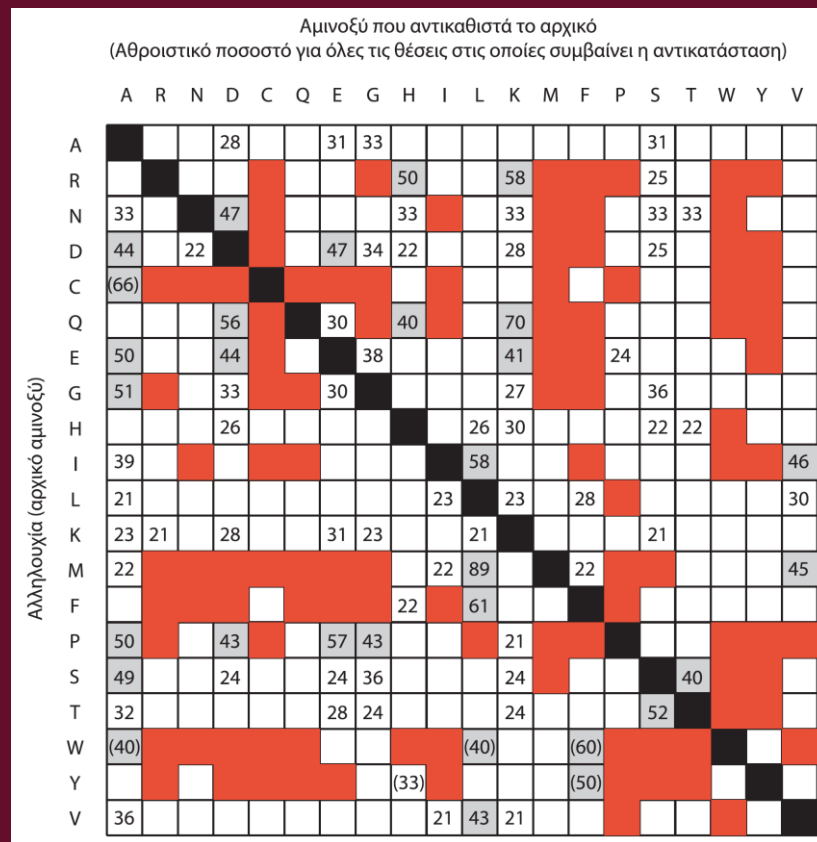
Εικόνα 3.27 Η σχετική εντροπία (H) ως συνάρτηση της απόστασης PAM. Για τους πίνακες PAM με χαμηλή τιμή (π.χ. PAM10), η σχετική εντροπία σε bits είναι υψηλή και το ελάχιστο μήκος που απαιτείται για την ανίχνευση μιας σημαντικής ομοιότητας μεταξύ δύο αλληλουχιών είναι μικρό (π.χ. περίπου 10 αμινοξέα). Χρησιμοποιώντας έναν πίνακα PAM10, η ομολογία μεταξύ δύο στενά συγγενικών πρωτεϊνών είναι ανιχνεύσιμη ακόμα και αν συγκριθεί μόνο μία σχετικά μικρή περιοχή τους. Για τους PAM250 και άλλους πίνακες PAM με υψηλές τιμές, η σχετική εντροπία (ή το πληροφοριακό περιεχόμενο της αλληλουχίας) είναι χαμηλή και είναι απαραίτητο να συγκρίνουμε σχετικά μεγάλης έκτασης περιοχές τους (π.χ. 80 αμινοξέα) προκειμένου να μπορέσουμε να ανιχνεύσουμε την ενδεχόμενη μεταξύ τους ομολογία.

Πίνακας 3.4 Αλγόριθμοι ολικής στοίχισης κατά ζεύγη.

Πρόγραμμα	Ιστότοπος	URL
BLAST	NCBI	http://www.ncbi.nlm.nih.gov/BLAST/
Needle, πακέτο EMBOSS (ολική στοίχιση κατά ζεύγη)	EBI	http://www.ebi.ac.uk/Tools/emboss/
Water πακέτο EMBOSS (τοπική στοίχιση κατά ζεύγη)	EBI	http://www.ebi.ac.uk/emboss/align/
Pairwise	Informagen (δυνατότητα επιλογής ολικής ή τοπικής στοίχισης κατά ζεύγη)	http://informagen.com/Applets/Pairwise/
Stretcher	Ινστιτούτο Παστέρ (ολική στοίχιση κατά ζεύγη)	http://bioweb2.pasteur.fr/docs/EMBOSS/stretcher.html

Πίνακας 3.5 Αλγόριθμοι τοπικής στοίχισης κατά ζεύγη.

Πρόγραμμα	Περιγραφή	URL
BLAST	NCBI	http://www.ncbi.nlm.nih.gov/BLAST/
est2genome	Πακέτο EMBOSS από το Ινστιτούτο Παστέρ. Στοιχίζει ετικέτες εκφραζόμενης αλληλουχίας (EST, βλ. Κεφάλαιο 10) με γονιδιωματικό DNA	http://bioweb.pasteur.fr/docs/EMBOSS/est2genome.html
LALIGN	Εντοπίζει πολλαπλές υποπεριοχές ομοιότητας σε δύο αλληλουχίες	http://www.ch.embnet.org/software/LALIGN_form.html
Pairwise	Εργαλείο στοίχισης δύο αλληλουχιών (δυνατότητα επιλογής ολικής ή τοπικής στοίχισης κατά ζεύγη)	http://informagen.com/Applets/Pairwise/
PRSS	Εργαλείο στοίχισης δύο αλληλουχιών από το Πανεπιστήμιο της Virginia (Bill Pearson)	http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=shuffle
SIM	Εργαλείο στοίχισης πρωτεϊνικών αλληλουχιών από το ExPASy	http://web.expasy.org/sim/
SSEARCH	Εργαλείο στοίχισης πρωτεϊνικών αλληλουχιών από το Protein Information Resource	http://pir.georgetown.edu/pirwww/search/pairwise.shtml



Εικόνα 3.28 Συχνότητες αμινοξικών αλλαγών στις σφαιρίνες (προσαρμοσμένες από τη δημοσίευση Zuckerkandl and Pauling, 1965, σελ. 118). Η σειρά των αμινοξέων είναι αλφαβητική (σύμφωνα με το πλήρες όνομά τους στα αγγλικά). Τα αποτελέσματα προέρχονται από τη στοίχιση αλληλουχιών αιμοσφαιρίνης και μυοσφαιρίνης από τον άνθρωπο και άλλα πρωτεύοντα, από το άλογο, την αγελάδα, τον χοίρο, τη μύρινα και τον κυπρίνο. Οι σειρές του πίνακα αντιστοιχούν στο αρχικό αμινοξύ και οι στήλες αντιστοιχούν στο αμινοξύ υποκατάστασης. Οι τιμές του πίνακα είναι το ποσοστό των αλλαγών μεταξύ των αντίστοιχων αμινοξέων. Αυτό το ποσοστό υπολογίζεται επί του συνόλου των θέσεων στις αλληλουχίες που είχαν το αρχικό αμινοξύ. Για παράδειγμα, η τιμή 33% στην πρώτη γραμμή του πίνακα (που συνδέει την αλανίνη με τη γλυκίνη) υποδηλώνει ότι από όλες τις θέσεις των αλληλουχιών που είχαν αλανίνη παρατηρήθηκε μεταλλαγή προς γλυκίνη στο 33%. Τα λευκά κελιά αντιστοιχούν στις περιπτώσεις εκείνες που το ποσοστό αντικατάστασης είναι μικρότερο από 20%. Τα κόκκινα κελιά αφορούν περιπτώσεις στις οποίες η σχετική αντικατάσταση δεν παρατηρήθηκε ποτέ. Τα γκρίζα κελιά αφορούν περιπτώσεις ιδιαίτερα συντηρητικών αντικαταστάσεων (ποσοστά $\geq 40\%$). Για παράδειγμα, στο 89% των θέσεων της στοίχισης που υπήρχαν μεθειονίνες παρατηρήθηκε και λευκίνη (ή το ίδιο πράγμα ανάποδα: μόνο στο 11% των θέσεων με μεθειονίνες δεν παρατηρήθηκαν καθόλου λευκίνες). Τα ταυτόσημα αμινοξέα υποδεικνύονται με μαύρα συμπαγή τετράγωνα. Οι τιμές σε παρένθεση υποδεικνύουν μικρό μέγεθος δείγματος, γεγονός που υποδηλώνει ότι η εξαγωγή συμπερασμάτων από τα δεδομένα αυτά πρέπει να γίνεται με ιδιαίτερη προσοχή.