Στοίχιση Αλληλουχιών Κατά Ζεύγη
(Pairwise sequence alignment)

ΚΑΘ. Γ. ΜΑΤΣΟΠΟΥΛΟΣ

Βασικά Σημεία

Η στοίχιση δύο πρωτεϊνικών ή νουκλεοτιδικών αλληλουχιών έχει ως στόχο τον έλεγχο της ομοιότητάς τους.

Αυτό είναι ιδιαίτερο χρήσιμο για την εξαγωγή συμπερασμάτων για τη δομική, λειτουργική ή και εξελικτική σχέση των συγκεκριμένων αλληλουχιών .

Οι περισσότερες μέθοδοι στοίχισης προσπαθούν να προσομοιώσουν μοριακούς εξελικτικούς μηχανισμούς.

```
HBA_HUMAN       MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GS 53
HBB_HUMAN       MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN 58
                ** *:* :*: *.* ****  :..* *.*** *::: :* *: :*  * ***      *.

HBA_HUMAN       AQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAH 113
HBB_HUMAN       PKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHH 118
                .:**.***** .*::::.:**:*:: .::::**:**..**:*** **:**.: *: .** *

HBA_HUMAN       LPAEFTPAVHASLDKFLASVSTVLTSKYR 142
HBB_HUMAN       FGKEFTPPVQAAYQKVVAGVANALAHKYH 147
                :  ****.*:*: :*.:*.*:..*: **:
```

Παράδειγμα στοίχισης των αλληλουχιών της α και β αλυσίδας της αιμογλοβίνης του ανθρώπου. Ο αστερίσκος “*” υποδηλώνει θέσεις με ταυτόσημα κατάλοιπα στις δύο αλληλουχίες.

Βασικά σημεία

•Μελετώντας το βαθμό ομοιότητας δύο αλληλουχιών μπορεί ένας ερευνητής να κρίνει για παράδειγμα αν είναι δικαιολογημένη η υπόθεση ομολογίας των αλληλουχιών, δηλαδή ύπαρξης κοινού εξελικτικού προγόνου.

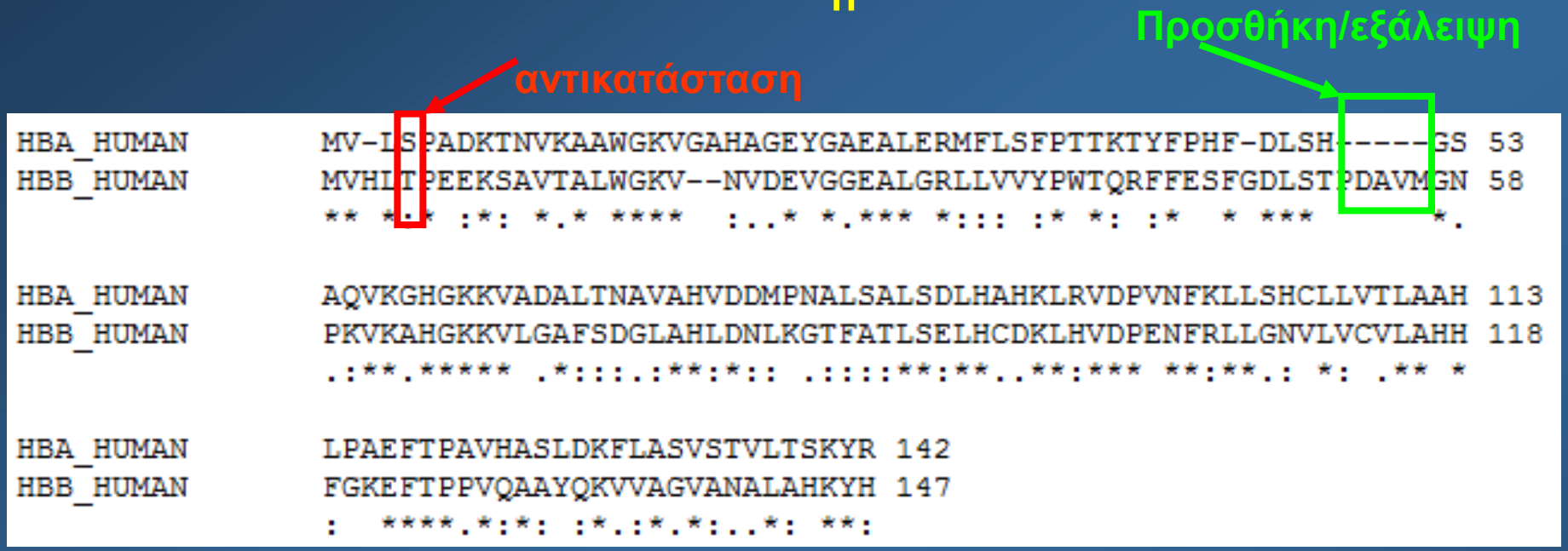
•**Υψηλός βαθμός ομοιότητας είναι ενδεικτικός παρόμοιας λειτουργίας, ενώ χαμηλός βαθμός ομοιότητας υποδηλώνει διαφορετικές λειτουργίες.**

•Μπορούμε να δούμε ποια κομμάτια της αλληλουχίας τείνουν να μένουν αμετάκλητα (είναι συντηρημένα). Αυτά συνήθως είναι και τα σημαντικότερα για τη δράση της πρωτεΐνης.

•Η σύγκριση δύο αλληλουχιών πραγματοποιείται τοποθετώντας τη μία αλληλουχία κάτω από την άλλη και καθορίζοντας ποιος χαρακτήρας της μίας αλληλουχίας θα αντιστοιχηθεί σε κάθε χαρακτήρα της δεύτερης.

•Όμοιοι ή παρόμοιοι χαρακτήρες τοποθετούνται ο ένας κάτω από τον άλλο, στην ίδια στήλη, ενώ ανόμοιοι χαρακτήρες μπορεί είτε να βρίσκονται στην ίδια στήλη είτε να στοιχίζονται με κενά διαστήματα.

Βασικά Σημεία



- Οι στοιχειώδεις αλλαγές που παρατηρούνται κατά τη διάρκεια της απόκλισης από το αρχικό προγονικό μόριο μπορούν να χαρακτηρισθούν ως **αντικαταστάσεις** (**substitutions**), **προσθήκες** (**insertions**) και **εξαλείψεις** (**deletions**) αμινοξικών καταλοίπων ή βάσεων.
- Ανόμοιοι χαρακτήρες που έχουν στοιχηθεί αντιπροσωπεύουν τις αντικαταστάσεις.
- Περιοχές που δεν μπορούν να στοιχηθούν εμφανίζονται ως **κενά διαστήματα** σε μία από τις δύο αλληλουχίες και μπορούν να ερμηνευθούν είτε ως προσθήκες χαρακτήρων στη μία αλληλουχία είτε ως εξαλείψεις στην άλλη.

Βασικά Σημεία

- Χρησιμοποιείται κατάλληλο **σύστημα βαθμονόμησης (scoring system)** για τον υπολογισμό ενός μέτρου της ομοιότητας των δύο αλληλουχιών, αποδίδοντας συγκεκριμένη βαθμολογία στη στοίχιση κάθε ζεύγους χαρακτήρων και στην εισαγωγή κενών.

  Π.χ. ταίριασμα όμοιων χαρακτήρων: 2 βαθμοί, ταίριασμα ανόμοιων χαρακτήρων: 0 βαθμοί, εισαγωγή κενού: -1 βαθμός

- Δύο βασικές κατηγορίες **αλγορίθμων στοίχισης** :

  Αλγόριθμοι δυναμικού προγραμματισμού

  Ευριστικοί αλγόριθμοι (π.χ. BLAST, FASTA)

Ολική και Τοπική Στοίχιση

- Ολική στοίχιση (global alignment): έχει στόχο να περιλάβει όσο το δυνατόν περισσότερους χαρακτήρες, σε όλο το μήκος των δύο αλληλουχιών.

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

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

- **Τοπική στοίχιση (local alignment):** κατασκευάζονται «νησίδες» στοίχισης από μεμονωμένες περιοχές που εμφανίζουν ομοιότητα, χωρίς να δίνεται ιδιαίτερο βάρος στην επέκταση της στοίχισης σε όλο το μήκος των αλληλουχιών.

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

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

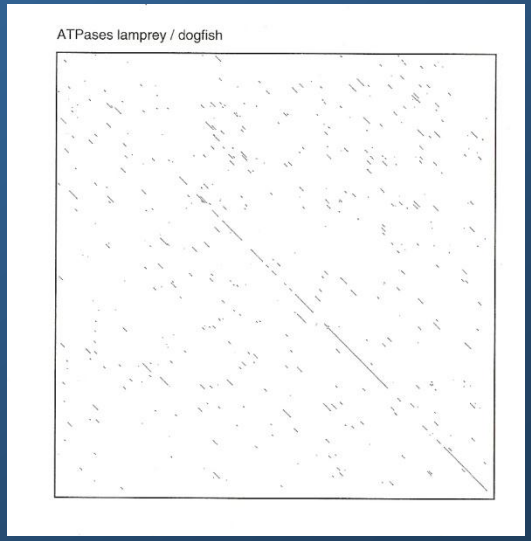
Ολική και Τοπική Στοίχιση

- Οι μέθοδοι τοπικής στοίχισης είναι ιδιαίτερα χρήσιμες όταν εξετάζονται δύο αλληλουχίες διαφορετικού μήκους ή αλληλουχίες που δεν εμφανίζουν ομοιότητα σε όλο το μήκος τους, αλλά τοπική ομοιότητα.
- Στις περιπτώσεις αυτές οι αλγόριθμοι ολικής στοίχισης μπορεί να δώσουν αποτελέσματα που δεν έχει βιολογικό νόημα.
- Το ίδιο ισχύει και για αλληλουχίες που έχουν αποκλίνει σημαντικά (έχουν μεγάλη εξελικτική απόσταση). Στην περίπτωση αυτή συνήθως μόνο ένα τμήμα της αλληλουχίας εμφανίζεται συντηρημένο, ενώ στο υπόλοιπο θα έχουν λάβει χώρα τόσα πολλά γεγονότα μεταλλάξεων, ώστε να μην είναι δυνατή μια στοίχιση που θα εκτείνεται σε όλο το μήκος των αλληλουχιών.

Dot Plots

Μία από τις απλούστερες μεθόδους οπτικοποίησης της ομοιότητας μεταξύ δύο αλληλουχιών.

Σε ένα dot plot συνήθως ο κάθε άξονας αντιστοιχεί σε μία αλληλουχία, και στο πλέγμα που προκύπτει μαρκάρονται με κάποιο τρόπο (πχ. με αστερίσκο) τα σημεία που αντιστοιχούν σε ταυτόσημα κατάλοιπα, ενώ τα υπόλοιπα μένουν κενά. Όταν δύο αλληλουχίες είναι όμοιες κατά μήκος μιας περιοχής, σχηματίζεται μία **διαγώνια γραμμή** στο διάγραμμα (με κατεύθυνση από πάνω αριστερά προς τα δεξιά κάτω)



Παράδειγμα: Σύγκριση του ίδου γονιδίου (ATPase-6) σε δύο διαφορετικά είδη ψαριών (lamprey, dogfish). Η ομοιότητα είναι εμφανέστερη στο δεύτερο μισό των αλληλουχιών.

Dot Plots

- Τα dot plots είναι πολύ εύκολα στην κατασκευή και την ερμηνεία τους, αλλά χαρακτηρίζονται από μεγάλο βαθμό υποκειμενικότητας, γιατί βασίζονται στην ικανότητα διάκρισης προτύπων του ανθρώπινου ματιού.
- Επίσης δεν δίνουν κάποια στοίχιση ως αποτέλεσμα.
- Όταν υπάρχουν πολλά μεμονωμένα ταυτόσημα κατάλοιπα, δημιουργείται «θόρυβος» στην εικόνα και πρέπει να χρησιμοποιηθεί κατάλληλο φίλτρο.
    - Μια μέθοδος φιλτραρίσματος θορύβου αποτελεί η χρήση «παραθύρων ομοιότητας»: διατηρούνται τα τετράγωνα με dots που ο αριθμός τους είναι πάνω από ένα κατώφλι.

Συστήματα Βαθμονόμησης

Κατά τη στοίχιση δύο αλληλουχιών στο τέριασμα δύο «θέσεων» ή «χαρακτήρων» αποδίδεται μια βαθμολογία.

Το ταίριασμα όμοιων αμινοξικών καταλοίπων και κατά δεύτερο λόγο καταλοίπων με παρόμοιες φυσικοχημικές ιδιότητες συνήθως ευνοείται από τα συστήματα βαθμονόμησης.

Σύστημα Βαθμονόμησης: Οι βαθμολογίες για όλα τα δυνατά ζεύγη αμινοξικών καταλοίπων ή νουκλεοτιδικών βάσεων σε συνδυασμό με κάποιους κανόνες για τις ποινές εισαγωγής κενών.

π.χ. ταίριασμα όμοιων χαρακτήρων: 2 βαθμοί, ταίριασμα ανόμοιων χαρακτήρων: 0 βαθμοί, εισαγωγή κενού: -1 βαθμός

Συστήματα Βαθμονόμησης

Σύνθετα συστήματα βαθμονόμησης:

•Για νουκλεοτιδικές αλληλουχίες: Πρότυπο Jukes-Cantor βάσει της υπόθεσης ότι κάθε νουκλεοτίδιο εμφανίζεται με την ίδια συχνότητα ενώ η πιθανότητα αντικατάστασης ενός νουκλεοτιδίου από άλλο είναι η ίδια.

•Για νουκλεοιδικές αλληλουχίες: Πρότυπο Hasegawa-Kishino-Yano (HKY) με το οποίο τέσσερα νουκλεοτίδια εμφανίζονται με διαφορετική συχνότητα και το κόστος μιας μετάπτωσης είναι διαφορετικό από το κόστος της μιας μεταστροφής.

Το ταίριασμα όμοιων αμινοξικών καταλοίπων και κατά δεύτερο λόγο καταλοίπων με παρόμοιες φυσικοχημικές ιδιότητες συνήθως ευνοείται από τα συστήματα βαθμονόμησης.

Αυτές είναι οι λεγόμενες **«συντηρητικές αντικαταστάσεις»** (πχ. ισολευκίνη με βαλίνη, σερίνη με θρεονίνη κα.) που παίρνουν υψηλότερη βαθμολογία από τις **«μη συντηρητικές αντικαταστάσεις»,** γιατί αμινοξέα με παρόμοιες φυσικοχημικές ιδιότητες έχουν περισσότερες πιθανότητες να αντικαταστήσουν το ένα το άλλο κατά τη διάρκεια της εξέλιξης.

Συστήματα Βαθμονόμησης

Μια προσέγγιση του συστήματος βαθμονόμησης θα ήταν να μετρηθούν οι συχνότητες εμφάνισης των διάφορων ζευγών καταλοίπων σε στοιχίσεις που είναι γνωστό ότι είναι σωστές και να χρησιμοποιηθούν για τη ρύθμιση των σχετικών βαθμολογιών.

Το τέριασμα κάθε ζεύγους χαρακτήρων θεωρείται ότι δεν επηρεάζεται από το τέριασμα των άλλων ζευγών χαρακτήρων των αλληλουχιών.

Δηλ. ανεξάρτητα ή διαφορετικά γεγονότα – μαρκοβιανή αλυσίδα: το κατάλοιπο σε μία θέση δεν σχετίζεται με άλλα κατάλοιπα σε άλλες θέσεις και μπορεί να χρησιμοποιηθεί ένα αθροιστικό σύστημα βαθμονόμησης.

Η προσέγγιση είναι καλή για την περίπτωση πρωτεϊνών και αλληλουχιών DNA

Πίνακες αντικατάστασης: πίνακες με τέσσερις σειρές και στήλες στην περίπτωση των νουκλεοτιδίων και είκοσι στην περίπτωση των αμινοξέων, που δίνουν τη βαθμολογία για το ταίριασμα δύο χαρακτήρων.

Πίνακες Αντικατάστασης PAM

Η κατασκευή τους βασίστηκε στη συγκέντρωση δεδομένων για τις αντικαταστάσεις αμινοξικών καταλοίπων από στοιχίσεις πολύ όμοιων αλληλουχιών, μικρής εξελικτικής απόστασης.

Βασική αρχή: οι υπάρχουσες πρωτεΐνες αποτελούν προϊόντα αποδεκτών σημειακών μεταλλαγών (**P**oint **A**ccepted **M**utations), δηλαδή αντικαταστάσεων στις προγονικές αλληλουχίες που έγιναν αποδεκτές από τη φυσική επιλογή.

Για να γίνει αποδεκτή μια αντικατάσταση: το «νέο» αμινοξικό κατάλοιπο θα έχει παρόμοιες ιδιότητες με το «παλιό».

Η αλλαγή από ένα αμονοξικό κατάλοιπο α στο β θεωρείται ισοπίθανη με την αντίστροφη αλλαγή από το β στο α.

Η κατασκευή των Πινάκων PAM έγινε με βάση την κατασκευή των φυλογενετικών δένδρων από στοιχίσεις 71 οικογενειών πρωτεϊνών από τις οποίες κάθε ζεύγος αλληλουχίας παρουσιάζει έως 80% ομοιότητα.

Πίνακες Αντικατάστασης PAM

Στη σειρά πινάκων PAM, οι πιθανότητες αντικατάστασης στον πίνακα PAM-n εκφράζουν τις αλλαγές που παρατηρούνται όταν υπάρχουν **n%** αποδεκτές σημειακές μεταλλαγές (αυτό ονομάζεται εξελικτική απόσταση n PAM).

Η χρήση πινάκων με μικρό n ενδείκνυται σε περιπτώσεις μικρής εξελικτικής απόστασης, όταν δηλαδή οι εξεταζόμενες αλληλουχίες περιμένουμε να είναι πολύ όμοιες μεταξύ τους.

Οι Πίνακες PAM δεν χρησιμοποιούνται καθώς τα κατάλοιπα δεν παρουσιάζουν πρακτικά την ίδια πιθανότητα να μεταλλαχθούν καθώς επίσης η κάθε αμινοξική αλλαγή μπορεί να είναι εντελώς τυχαία και να μη συνδέεται με την εκτίμηση αντικαταστάσεων που προϋποθέτει το μοντέλο.

Ο πίνακας Αντικατάστασης PAM250

The Dayhoff PAM250 matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala (A)	2																			
Arg (R)	−2	6																		
Asn (N)	0	0	2																	
Asp (D)	0	−1	2	4																
Cys (C)	−2	−4	−4	−5	12															
Gln (Q)	0	1	1	2	−5	4														
Glu (E)	0	−1	1	3	−5	2	4													
Gly (G)	1	−3	0	1	−3	−1	0	5												
His (H)	−1	2	2	1	−3	3	1	−2	6											
Ile (I)	−1	−2	−2	−2	−2	−2	−2	−3	−2	−5										
Leu (L)	−2	−3	−3	−4	−6	−2	−3	−4	−2	2	6									
Lys (K)	−1	3	1	0	−5	1	0	−2	0	−2	−3	5								
Met (M)	−1	0	−2	−3	−5	−1	−2	−3	−2	2	4	0	6							
Phe (F)	−3	−4	−3	−6	−4	−5	−5	−5	−2	1	2	−5	0	9						
Pro (P)	1	0	0	−1	−3	0	−1	0	0	−2	−3	−1	−2	−5	6					
Ser (S)	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	2				
Thr (T)	1	−2	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	0	1	3			
Trp (W)	−6	2	−4	−7	−8	−5	−7	−7	−3	−5	−2	−3	−4	0	−6	−2	−5	17		
Tyr (Y)	−3	−4	−2	−4	0	−4	−4	−5	0	−1	−1	−4	−2	7	−5	−3	−3	0	10	
Val (V)	0	−2	−2	−2	−2	−2	−2	−1	−2	4	2	−2	2	−1	−1	−1	0	−6	−2	4

Πίνακες Αντικατάστασης BLOSUM (BLOcks Substitution Matrices)

Εμπειρικοί πίνακες αντικατάστασης.

Η κατασκευή τους βασίζεται στην απαρίθμηση ζευγών αμινοξικών καταλοίπων που εμφανίζονται σε συντηρημένες περιοχές σε πολλές ολικές στοιχίσεις αλληλουχιών μικρής ομοιότητας και απομακρυσμένης εξελικτικής σχέσης.

Οι περιοχές αυτές δεν έχουν κενά και ονομάζονται BLOCKs.

Για λόγους στάθμισης της επίδρασης των περισσότερο όμοιων αλληλουχιών στο αποτέλεσμα, πραγματοποιήθηκαν ομαδοποιήσεις (clustering) των αλληλουχιών σε κάθε BLOCK με βάση το ποσοστό των ταυτόσημων καταλοίπων. Π.χ. όταν ομαδοποιούνται οι αλληλουχίες που έχουν ομοιότητα τουλάχιστον 62%, προκύπτει ο πίνακας BLOSUM62, που χρησιμοποιείται ευρέως στη βαθμολόγηση στοιχίσεων.

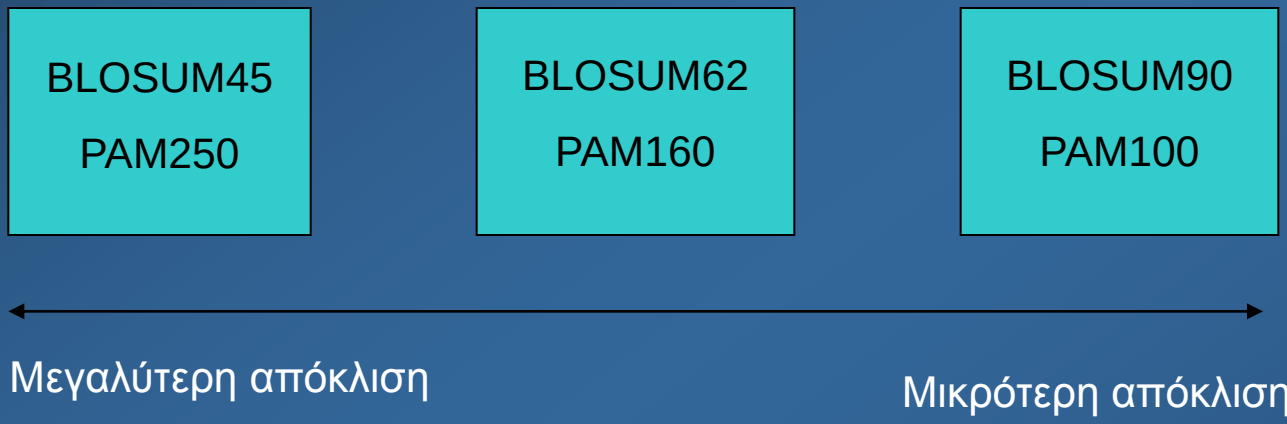
Στη σειρά πινάκων BLOSUM-n, ο αριθμός n δείχνει το μέγιστο επίπεδο ταυτότητας που μπορούν να έχουν οι αλληλουχίες ώστε να συμβάλουν ανεξάρτητα στο μοντέλο.

Ο πίνακας Αντικατάστασης BLOSUM-62

The BLOSUM62 matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala (A)	4																			
Arg (R)	−1	5																		
Asn (N)	−2	0	6																	
Asp (D)	−2	−2	1	6																
Cys (C)	0	−3	−3	−3	9															
Gln (Q)	−1	1	0	0	−3	5														
Glu (E)	−1	0	0	2	−4	2	5													
Gly (G)	0	−2	0	−1	−3	−2	−2	6												
His (H)	−2	0	1	−1	−3	0	0	−2	8											
Ile (I)	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu (L)	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys (K)	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met (M)	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe (F)	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro (P)	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser (S)	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr (T)	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp (W)	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr (Y)	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val (V)	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Αντιστοιχία Πινάκων PAM - BLOSUM



Πίνακες BLOSUM με μεγάλο n αντιστοιχούν σε πίνακες PAM με μικρότερο n και χρησιμοποιούνται για μικρότερες εξελικτικές αποστάσεις.

Οι πίνακες BLOSUM χρησιμοποιούνται ευρέως και είναι καλύτεροι σε γενικές γραμμές για την εύρεση τοπικών στοιχίσεων.

Ποινές για τα Κενά

Η προσθήκη των κενών στις στοιχίσεις γίνεται ώστε να μεγιστοποιείται το ταίριασμα όμοιων ή παρόμοιων χαρακτήρων στα επόμενα ή προηγούμενα τμήματα στοίχισης.

Έχουν προταθεί διάφορες εμπειρικές στρατηγικές. Η πιο κλασική αντιμετώπιση είναι η χρήση «συσχετισμένης» ή «αφινικής» ποινής (affine gap penalty). Προβλέπει την αφαίρεση μιας τιμής για την εισαγωγή ενός κενού και πρόσθετες αφαιρέσεις για επόμενες επεκτάσεις του κενού (εισαγωγή νέων κενών στο ίδιο σημείο):

$$p = - A – B*(n-1)$$

A: «ποινή έναρξης του κενού» (gap opening penalty)
B: η ποινή επέκτασης (gap extension penalty).

Συνήθως χρησιμοποιείται υψηλή τιμή για το A και χαμηλότερη για το B, με τη λογική ότι τα αρχικά γεγονότα εισαγωγής θα είναι σπάνια, αλλά όταν συμβούν είναι πιθανό να περιληφθούν περισσότερα γειτονικά κατάλοιπα.

Ο συσχετισμός των τιμών για τα ταιριάσματα καταλοίπων και την εισαγωγή κενών μπορεί να επιδράσει σημαντικά στο αποτέλεσμα μιας στοίχισης.

Αλγόριθμοι Δυναμικού Προγραμματισμού

Είναι οι βασικότεροι αλγόριθμοι της υπολογιστικής ανάλυσης αλληλουχιών:

Αλγόριθμος ολικής στοίχισης **Needleman-Wunsch**

Αλγόριθμος τοπικής στοίχισης **Smith-Waterman**

Βασίζονται στην αρχή του **«Διαίρει και βασίλευε»,** δηλαδή στην εύρεση της βέλτιστης λύσης ενός προβλήματος συνδυάζοντας τις λύσεις διαδοχικών απλούστερων προβλημάτων. Η βέλτιστη στοίχιση δύο αλληλουχιών μπορεί να βρεθεί από τις βέλτιστες στοιχίσεις διαδοχικών τμημάτων τους.

Το πρόβλημα της βέλτιστης στοίχισης δύο αλληλουχιών μπορεί να αναχθεί στην εύρεση του βέλτιστου μονοπατιού ενός γραφήματος.

Στο γράφημα αυτό όλες οι δυνατές στοιχίσεις των δύο αλληλουχιών αναπαρίστανται με διαδρομές που περνούν από τους κόμβους ενός πλέγματος δύο διαστάσεων που κατασκευάζεται με την ορθογώνια διάταξη των δύο αλληλουχιών.

Αλγόριθμος Ολικής Στοίχισης Needleman-Wunsch

Ο αλγόριθμος Needleman-Wunsch είναι ο κλασικός αλγόριθμος δυναμικού προγραμματισμού για την ολική στοίχιση δύο αλληλουχιών.

Η βασική αρχή του είναι ότι το συνολικό βέλτιστο μονοπάτι σχηματίζεται από μικρότερα βέλτιστα μονοπάτια.

Η διαδικασία εύρεσης λύσης οδηγεί στην κατασκευή ενός πίνακα βαθμολογιών όλων των δυνατών στοιχίσεων δύο αλληλουχιών.

Αλγόριθμος Ολικής Στοίχισης Needleman-Wunsch

Για τη στοίχιση δύο αλληλουχιών $x_1x_2\ldots x_n$ και $y_1y_2\ldots y_m$ κατασκευάζεται ένας πίνακας **F(i,j),** $0 \le i \le n$, $0 \le j \le m$, όπου το (i,j) στοιχείο είναι η βαθμολογία της βέλτιστης στοίχισης του τμήματος $x_1\ldots x_i$ με το $y_1\ldots y_j$, δηλαδή η **βαθμολογία της βέλτιστης στοίχισης μέχρι τη θέση (i,j).**

Θέτουμε F(0,0)=0 και διατρέχουμε τον πίνακα από πάνω αριστερά προς τα δεξιά και προς τα κάτω και χρησιμοποιούμε επαναλαμβανόμενα τη σχέση:

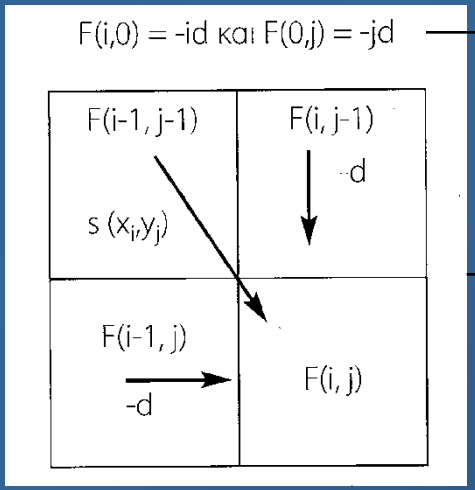
$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

→ Στοίχιση x_i με y_j

→ Στοίχιση x_i με ένα κενό

→ Στοίχιση y_j με ένα κενό

$s(x_i,y_j)$ είναι η βαθμολογία για τη στοίχιση του καταλοίπου x_i με το y_j και d η ποινή για τα κενά.



→ Η πρώτη γραμμή και στήλη

→ Οι τρεις δυνατοί τρόποι επέκτασης της στοίχισης στη θέση (i,j)

Χρονική πολυπλοκότητα, απαιτήσεις σε μνήμη: O(nm) – 3 αθροίσεις και 1 μέγιστο

Παράδειγμα: Ολική Στοίχιση των Αλληλουχιών ASIRVVFALF και ASRFALFF

		A	S	R	F	A	L	F	F
	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1	**2**	1	0	-1	-2	-3	-4	-5
S	-2	1	**4**	3	2	1	0	-1	-2
I	-3	0	**3**	4	3	2	1	0	-1
R	-4	-1	2	**5**	4	3	2	1	0
V	-5	-2	1	**4**	5	4	3	2	1
V	-6	-3	0	**3**	4	5	4	3	2
F	-7	-4	-1	2	**5**	4	5	6	5
A	-8	-5	-2	1	4	**7**	6	5	6
L	-9	-6	3	0	3	6	**9**	8	7
F	-10	-7	-4	-1	2	5	8	**11**	**10**

Βέλτιστη στοίχιση:

A S - R - - F A L F F

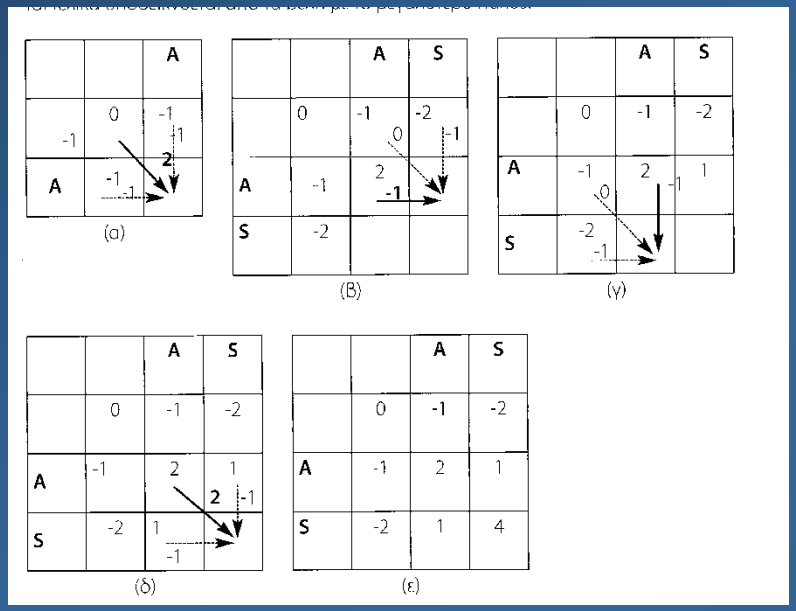
A S I R V V F A L F -

Score=10

Σύστημα βαθμονόμησης: Ταίριασμα ταυτόσημων καταλοίπων +2 βαθμοί, ταίριασμα ανόμοιων καταλοίπων 0 βαθμοί, εισαγωγή κενού -1 βαθμός.

Καθώς διατρέχουμε τον πίνακα και γεμίζουμε τα κελιά, κρατάμε κάθε φορά και ένα δείκτη (βελάκι στην εικόνα) που μας πληροφορεί από ποιο κελί προήλθε η τιμή στην τρέχουσα θέση. Στο τέλος της διαδικασίας η τιμή στο τελευταίο κελί του πίνακα, F(n,m), δίνει την βαθμολογία της βέλτιστης στοίχισης, ενώ για να βρούμε την ίδια τη στοίχιση ακολουθούμε με την αντίστροφη φορά τα βελάκια.

Παράδειγμα (συνέχεια)



Ενδεικτικά στιγμιότυπα από τη συμπλήρωση του πίνακα δυναμικού προγραμματισμού.

Το κελί προς συμπλήρωση είναι σκιασμένο, τα βελάκια δείχνουν τις πιθανές κατευθύνσεις και φέρουν την αντίστοιχη βαθμολογία (που προστίθεται στη βαθμολογία του κελιού από το οποίο ξεκινάνε), ενώ η κατεύθυνση που επιλέγεται τελικά υποδεικνύεται από τα βέλη με το μεγαλύτερο πάχος.

Αλγόριθμος Ολικής Στοίχισης Needleman-Wunsch

Για τον συγκεκριμένο αλγόριθμο βλέπουμε ότι χρειάζεται να αποθηκευθούν (n+1)×(m+1) αριθμοί, καθένας από τους οποίους μας «κοστίζει» έναν σταθερό αριθμό υπολογισμών (τρείς αθροίσεις και μία εύρεση μεγίστου).

Η χρονική δηλαδή πολυπλοκότητα και οι απαιτήσεις σε μνήμη του αλγορίθμου είναι O(nm) (“της τάξης nm”). Για συγκρίσιμα n και m, λέμε ότι η πολυπλοκότητα είναι $O(n^2)$.

Αλγόριθμος Τοπικής Στοίχισης Smith-Waterman

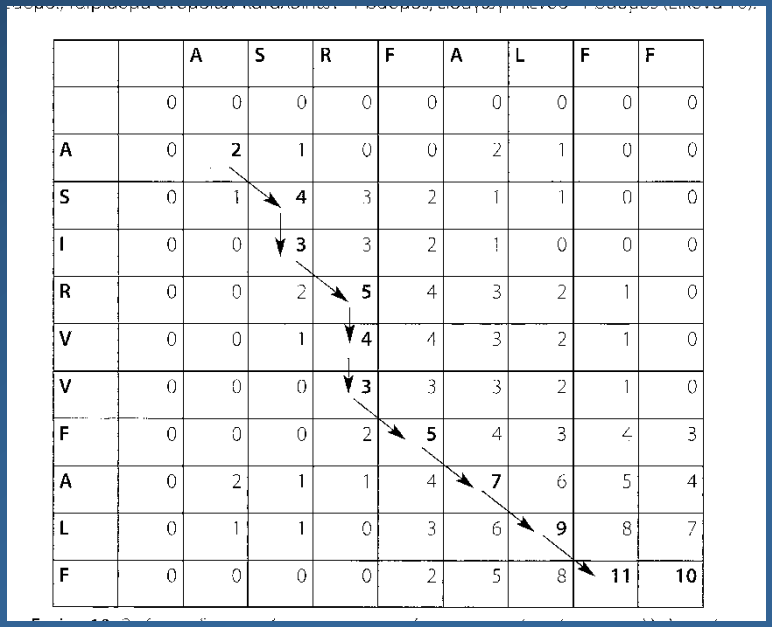
Τροποποίηση του αλγορίθμου Needleman-Wunsch: το βέλτιστο μονοπάτι μπορεί να μη φτάνει στις άκρες του γραφήματος αναζήτησης, αλλά να ξεκινάει και να τελειώνει στο εσωτερικό του. Η σχέση για τη συμπλήρωση των κελιών του πίνακα γίνεται:

$$F(i,j) = \max \begin{Bmatrix} 0 \\ F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{Bmatrix}$$

Έχει προστεθεί μία ακόμη επιλογή: το F(i,j) μπορεί να πάρει την τιμή 0 αν οι υπόλοιπες επιλογές δίνουν αρνητική βαθμολογία, που σημαίνει ότι ξεκινάει μία καινούρια τοπική στοίχιση.

Για την κατασκευή της βέλτιστης στοίχισης ξεκινάμε από τη μέγιστη τιμή F(i,j) οπουδήποτε στον πίνακα.

Παράδειγμα: Τοπική Στοίχιση των αλληλουχιών ASIRVVFALF και ASRFALFF



		A	S	R	F	A	L	F	F
	0	0	0	0	0	0	0	0	0
A	0	**2**	1	0	0	2	1	0	0
S	0	1	**4**	3	2	1	1	0	0
I	0	0	**3**	3	2	1	0	0	0
R	0	0	2	**5**	4	3	2	1	0
V	0	0	1	**4**	4	3	2	1	0
V	0	0	0	**3**	3	3	2	1	0
F	0	0	0	2	**5**	4	3	4	3
A	0	2	1	1	4	**7**	6	5	4
L	0	1	1	0	3	6	**9**	8	7
F	0	0	0	0	2	5	8	**11**	**10**

Βέλτιστη στοίχιση:

A S - R - - F A L F

A S I R V V F A L F

Score=11

Ταίριασμα ταυτόσημων καταλοίπων +2 βαθμοί, ταίριασμα ανόμοιων καταλοίπων και εισαγωγή κενού -1 βαθμός.

Αλγόριθμος Τοπικής Στοίχισης Smith-Waterman

Επειδή χρησιμοποιήθηκαν πολύ απλές αλληλουχίες και συστήματα βαθμονόμησης, το αποτέλεσμα της τοπικής στοίχισης είναι το ίδιο με την περίπτωση της ολικής στοίχισης.

Ωστόσο, φαίνεται η διαφορετική λογική που ακολουθείται κατά τη συμπλήρωση του πίνακα του δυναμικού προγραμματισμού.

Θα πρέπει να σημειωθεί ότι για να λειτουργήσει σωστά ο αλγόριθμος τοπικής στοίχισης είναι απαραίτητη και η χρήση κατάλληλου συστήματος βαθμονόμησης.

Ευριστικοί Αλγόριθμοι

Οι αλγόριθμοι δυναμικού προγραμματισμού είναι σίγουρο ότι θα εντοπίσουν τη μαθηματικά βέλτιστη λύση του προβλήματος της στοίχισης δύο αλληλουχιών για συγκεκριμένο σύστημα βαθμονόμησης.

Αλλά: υψηλή χρονική πολυπλοκότητα, υψηλές αποθηκευτικές απαιτήσεις

$\Rightarrow$Μη πρακτικοί για τη στοίχιση αλληλουχιών μεγάλου μήκους και για αναζητήσεις έναντι μεγάλων βάσεων δεδομένων.

Για βιολογικές αλληλουχίες και κλασσικούς υπολογιστές οι $O(n^2)$ αλγόριθμοι είναι εφικτοί αλλά αργοί, ενώ οι $O(n^3)$ αλγόριθμοι είναι εφικτοί μόνο για πολύ μικρές αλληλουχίες.

Για προβλήματα υψηλών υπολογιστικών απαιτήσεων, χρησιμοποιούνται οι Ευριστικοί Αλγόριθμοι.

Ευριστικοί αλγόριθμοι

Οι ευριστικοί αλγόριθμοι **θυσιάζουν ένα μέρος της ευαισθησίας τους, με στόχο την επιτάχυνση της εξεύρεσης της λύσης**, προσπαθώντας να δουλέψουν μόνο σε κάποιες περιοχές του πίνακα δυναμικού προγραμματισμού όπου αναμένονται τα υψηλότερα σκορ.

Οι αλγόριθμοι **BLAST** και **FASTA** αποτελούν τους πιο συχνά χρησιμοποιούμενους ευριστικούς αλγορίθμους και πραγματοποιούν τη στοίχιση των αλληλουχιών πολύ γρήγορα, βασιζόμενοι στην εύρεση μικρών περιοχών ομοιότητας («λέξεων») και στην επέκταση τους με χρήση δυναμικού προγραμματισμού.

Χρησιμοποιούνται για αναζητήσεις ομοιοτήτων από Βάσεις Δεδομένων.

FASTA

1. Εντοπίζονται όλες οι "**λέξεις**" μήκους ktup: μικρές περιοχές χωρίς κενά, στις οποίες οι δύο αλληλουχίες ταυτίζονται (τυπικό μήκος για αμινοξικές αλληλουχίες 1 ή 2, για νουκλεοτιδικές 4 ή 6).

2. Εντοπίζονται οι **διαγώνιες** με τα περισσότερα k-tuples, από τη διαφορά των θέσεων των k-tuples στις δύο αλληλουχίες.

3. Επεκτείνονται οι στοιχίσεις που ξεκινούν από τα k-tuples πάνω σε κάθε διαγώνιο επιτρέποντας τη στοίχιση ανόμοιων καταλοίπων. Έτσι προκύπτουν οι λεγόμενες **Βέλτιστες Αρχικές Περιοχές (Best Initial Regions).** Κατά τη διαδικασία αυτή μπορεί να ενωθούν κάποια k-tuples που βρίσκονται στην ίδια διαγώνιο.

4. Η **εισαγωγή κενών** γίνεται επιτρεπτή με κάποια ποινή και ενοποιούνται περιοχές που δεν ανήκουν απαραίτητα στην ίδια διαγώνιο, όσο η βαθμολογία παραμένει υψηλότερη από ένα κατώφλι.

5. Πραγματοποιείται **πλήρης στοίχιση με δυναμικό προγραμματισμό** σε μια περιοχή γύρω από τις συγκεκριμένες διαγωνίους με το υψηλότερο σκορ , όπως έχουν προσδιορισθεί στα προηγούμενα βήματα.

Υπολογιστικό Πακέτο BLAST

Βασική αρχή: Οι βέλτιστες στοιχίσεις περιέχουν μικρές περιοχές (λέξεις) όπου η βαθμολογία της στοίχισης είναι μεγαλύτερη από μια τιμή κατωφλίου. Αυτές οι λέξεις θεωρούνται πιθανά σημεία έναρξης μιας καλής τοπικής στοίχισης. Το τυπικό τους μήκος είναι 3 για πρωτεϊνικές αλληλουχίες και 28 για νουκλεοτιδικές.

Ο αλγόριθμος τις εντοπίζει και προσπαθεί να επεκτείνει τη στοίχιση και προς τις δύο κατευθύνσεις, όσο η βαθμολογία αυξάνεται. Οι περιοχές που προκύπτουν έτσι είναι τα λεγόμενα ζεύγη υψηλής βαθμολογίας **(High Scoring Pairs, HSPs),** από τα οποία επιλέγονται εκείνα που έχουν βαθμολογία μεγαλύτερη από κάποιο κατώφλι και εμφανίζουν υψηλή στατιστική σημαντικότητα (δηλαδή, όπως θα εξηγηθεί στη συνέχεια, θεωρείται ότι δεν έχουν προκύψει τυχαία).

Η βασική υλοποίηση του αλγορίθμου BLAST εντοπίζει τοπικές στοιχίσεις χωρίς κενά, ωστόσο έχουν αναπτυχθεί διάφορες εξειδικευμένες παραλλαγές που δίνουν και τοπικές στοιχίσεις με κενά.

Το **υπολογιστικό πακέτο BLAST** περιλαμβάνει προγράμματα για την εύρεση τοπικών ομοιοτήτων μεταξύ μιας αλληλουχίας επερώτησης (query sequence) και μιας βάσης δεδομένων, τόσο για πρωτεϊνικές όσο και για νουκλεοτιδικές αλληλουχίες.

Υπολογιστικό πακέτο BLAST

Μία οικογένεια προγραμμάτων. Παραδείγματα επιμέρους προγραμμάτων:

BLASTp: στοίχιση πρωτεϊνικής αλληλουχίας έναντι βάσης δεδομένων πρωτεϊνικών αλληλουχιών

BLASTn: στοίχιση νουκλεοτιδικής αλληλουχίας έναντι βάσης δεδομένων νουκλεοτιδικών αλληλουχιών

Blastx: μετάφραση αλληλουχίας DNA στα έξι πιθανά πλαίσια ανάγνωσης και αναζήτηση έναντι βάσης δεδομένων πρωτεϊνών.

tblastn: στοίχιση πρωτεϊνικής αλληλουχίας έναντι βάσης δεδομένων μεταφρασμένων αλληλουχιών DNA.

Υπολογιστικό πακέτο BLAST

- Αποδοτικός στην αναζήτηση ομοιοτήτων έναντι (τεράστιων) βάσεων δεδομένων (ταχύτερος, υπολογίζει τη στατιστική σημαντικότητα του αποτελέσματος)
- Ουσιαστικά ψάχνει (μικρές) περιοχές που δίνουν πολύ καλή στοίχιση και προσπαθεί να τις επεκτείνει.
- Διατίθονται ελεύθερα υλοποιήσεις είτε για χρήση μέσω διαδικτύου ή και για download.

  π.χ. οι ευρέως χρησιμοποιούμενες υλοποιήσεις του αλγορίοθμου blast από το **NCBI** (National Center for Biotechnology Information) http://blast.ncbi.nlm.nih.gov/Blast.cgi και το **EBI** (European Bioinformatics Institute) http://www.ebi.ac.uk/Tools/blast/

  Ο χρήστης μπορεί να βρει εκεί λεπτομερείς οδηγίες για αποδοτικό τρόπο χρήσης των διάφορων προγραμμάτων.

Στατιστική Σημαντικότητα

Οι αλγόριθμοι στοίχισης δίνουν ως αποτέλεσμα κάποια στοίχιση ακόμη και για αλληλουχίες που δεν έχουν πραγματικά βιολογική σχέση.

Για να διακρίνουμε τις στοιχίσεις με βιολογικό νόημα από εκείνες που θα μπορούσαν να προκύψουν κατά τύχη, υπολογίζουμε τη στατιστική σημαντικότητα της βαθμολογίας που αντιστοιχεί σε μία στοίχιση.

Μπορούμε να συγκρίνουμε τη βαθμολογία μίας στοίχισης με τις βαθμολογίες που προκύπτουν κατά τη στοίχιση τυχαίων αλληλουχιών παρόμοιας νουκλεοτιδικής/αμινοξικής σύστασης.

Π.χ. ανακατεύουμε πολλές φορές τυχαία τα κατάλοιπα της μίας αλληλουχίας και έτσι προκύπτουν πολλές τυχαίες αλληλουχίες, οι οποίες θα στοιχηθούν με τη δεύτερη αλληλουχία. Αν οι τυχαίες αλληλουχίες δίνουν στοιχίσεις με εξίσου καλή βαθμολογία, τότε είναι πιθανό η στοίχιση των δύο αλληλουχιών να μην είναι στατιστικά σημαντική.

Αναμενόμενη Κατανομή Τυχαίων Βαθμολογιών

Κατανομή ακραίων τιμών (Extreme Value Distribution, EVD)

$$P(S \leq s) \cong e^{-Kmne^{\lambda s}}$$

S: η τυχαία μεταβλητή της βαθμολογίας των στοιχίσεων, K, λ: οι παράμετροι της κατανομής που προσαρμόζονται στο σύστημα βαθμονόμησης, m,n: τα μήκη των αλληλουχιών.

Η Πιθανότητα τυχαίας στοίχισης με βαθμολογία μεγαλύτερη ή ίση με το s:

$$p\_value \equiv P(S \geq s) = 1 - e^{-Kmne^{\lambda s}}$$

Αναμενόμενο πλήθος τυχαίων στοιχίσεων με βαθμολογία μεγαλύτερη ή ίση με το s:

$$E(S \geq s) = Kmne^{-\lambda s}$$

Η βαθμολογία εξαρτάται από τον πίνακα αντικατάστασης και ποινές κενών. Για σύγκριση αποτελεσμάτων από διαφορετικά συστήματα βαθμονόμησης: Κανονικοποιημένα score:

$$s_{bit} = \frac{\lambda s - \ln K}{\ln 2}$$

$$E\text{-}value = E(S_{bit} > s_{bit}) = mn\, 2^{-s_{bit}}$$

Όσο πλησιάζει η τιμή του E-value στο μηδέν, τόσο μεγαλύτερη είναι η πιθανότητα η εξεταζόμενη στοίχιση να μην έχει προκύψει κατά τύχη.

Επιλογή Μεθόδου Στοίχισης

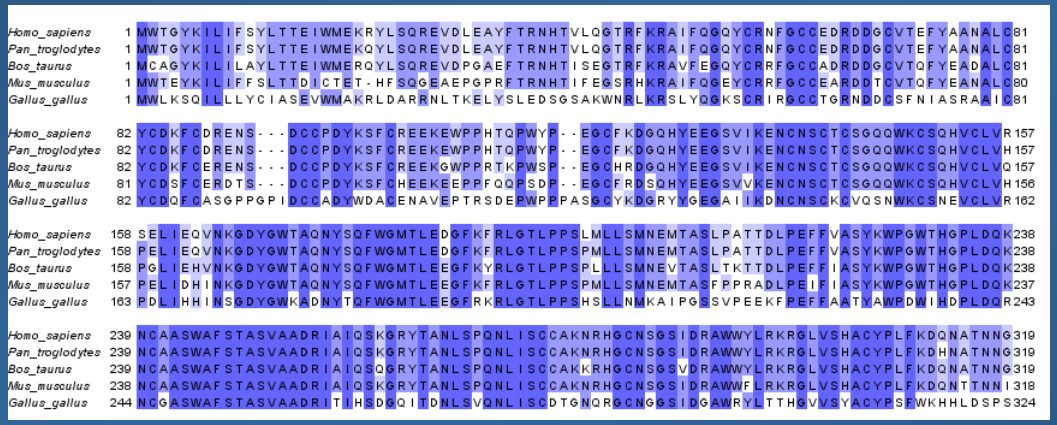
Κατά την επιλογή της μεθόδου που θα χρησιμοποιηθεί για την πραγματοποίηση μιας στοίχισης θα πρέπει πάντοτε να λαμβάνεται υπόψη ο λόγος για τον οποίο γίνεται η στοίχιση των αλληλουχιών. Ανάλογα με το στόχο κάποια προγράμματα στοίχισης μπορεί να είναι πλεονεκτικότερα συγκριτικά με άλλα.

Άλλη σημαντική απόφαση είναι η επιλογή τοπικής ή ολικής στοίχισης και η ρύθμιση σημαντικών παραμέτρων όπως ο πίνακας αντικατάστασης και οι ποινές για τα κενά που θα χρησιμοποιηθούν.

Για αλληλουχίες που είναι πολύ όμοιες (πχ. ομοιότητα της τάξης του 90%) η κατασκευή της στοίχισης είναι συνήθως πολύ εύκολη με όλους τους αλγορίθμους. Για πρωτεϊνικές αλληλουχίες, ομοιότητα της τάξης του 25% θεωρείται οριακή για την ασφαλή εξαγωγή συμπερισμάτων. Η στοίχιση αλληλουχιών σε αυτό το επίπεδο ομοιότητας χαρακτηρίζεται ως «ζώνη του λυκόφωτος»

Πολλαπλή Στοίχιση Αλληλουχιών (multiple sequence alignment)

Ένας σημαντικός τρόπος αναζήτησης λειτουργικών, δομικών και εξελικτικών πληροφοριών για ένα σύνολο (σχετιζόμενων) πρωτεϊνικών ή νουκλεοτιδικών αλληλουχιών.



Παράδειγμα πολλαπλής στοίχισης τμήματος μιας πρωτεΐνης σε 5 διαφορετικά είδη. Σκούρες αποχρώσεις του μπλε υποδηλώνουν μεγαλύτερη ομοιότητα.

Εκτός από τις περιπτώσεις πολύ όμοιων αλληλουχιών, η αυτόματη κατασκευή μίας πολλαπλής στοίχισης αποτελεί σύνθετο υπολογιστικό πρόβλημα. Συχνά η βιολόγοι επεξεργάζονται με το χέρι το αποτέλεσμα μιας πολλαπλής στοίχισης για να τη βελτιώσουν/διορθώσουν με βάση επιπλέον στοιχεία που ενδεχομένως διαθέτουν.

Βασικοί Αλγόριθμοι Πολλαπλής Στοίχισης (1)

- **Αλγόριθμοι δυναμικού προγραμματισμού**. Μη πρακτικοί.
- **Προοδευτικοί ή ιεραρχικοί αλγόριθμοι (progressive/hierarchical/tree methods).**

  Η πιο γνωστή και σχετικά αποδοτική μέθοδος πολλαπλής στοίχισης.

  Φτιάχνουν μία πολλαπλή στοίχιση βασιζόμενοι στις στοιχίσεις κατά ζεύγη των επιμέρους αλληλουχιών, ξεκινώντας από το ζευγάρι με τη μεγαλύτερη βαθμολογία και προχωρώντας σταδιακά στα επόμενα.

  Η βαθμολογία της ολικής στοίχισης υπολογίζεται συχνά ως το άθροισμα των βαθμολογιών των επιμέρους στοιχίσεων (Sum of Pairs, SPs).

  Αλληλουχίες με μεγάλη ομοιότητα συχνά σταθμίζονται για λόγους «δικαιοσύνης» στη συμβολή τους στο αποτέλεσμα.

  π.χ. **CLUSTALW** (http://www.ebi.ac.uk/Tools/clustalw2/index.html) (ένα από τα πιο δημοφιλή και εύχρηστα λογισμικά)

  T-Coffee (http://www.ebi.ac.uk/Tools/t-coffee/index.html)

Βασικοί Αλγόριθμοι Πολλαπλής Στοίχισης (2)

- **Επαναληπτικοί αλγόριθμοι (iterative algorithms).**

  Παρόμοιοι με τους προοδευτικούς αλγορίθμους, με τη διαφορά ότι πραγματοποιούν επαναλαμβανόμενες στοιχίσεις των αλληλουχιών κάθε φορά που προσθέτουν μία καινούρια.

  π.χ. **DIALIGN** (http://bibiserv.techfak.uni-bielefeld.de/dialign/)
  MUSCLE (http://www.drive5.com/muscle/)

- **Hidden Markov Models (HMMs)**

  Στατιστική μέθοδος που αποδίδει πιθανότητες σε όλους τους δυνατούς συνδυασμούς ταιριασμάτων, μη ταιριασμάτων και κενών για την κατασκευή της πιθανότερης στοίχισης.

  π.χ. **HMMER** (http://hmmer.janelia.org/)

- Και πολλοί άλλοι...