Κεφάλαιο 3

Αλγόριθμοι Στοίχισης Αλληλουχιών

Σύνοψη
Στο κεφάλαιο αυτό θα παρουσιαστούν αρχικά, τα απαραίτητα μαθηματικά μοντέλα που περιγράφουν τις αλληλουχίες μακρομορίων και κάποια βασικά ασυμπτωτικά αποτελέσματα που αναφέρονται σε αυτές. Στη συνέχεια θα παρουσιαστούν τα βασικά θεωρητικά αποτελέσματα που αφορούν στη στοίχιση βιολογικών αλληλουχιών. Θα παρουσιαστούν οι τρόποι βαθμονόμησης της στοίχισης, οι τρόποι εύρεσης της στοίχισης, καθώς και τα διαφορετικά είδη αλγορίθμων στοίχισης, ενώ ιδιαίτερη έμφαση θα δοθεί στην αξιολόγηση της στατιστικής σημαντικότητας μιας στοίχισης. Τέλος, θα παρουσιαστούν οι βασικοί ευριστικοί αλγόριθμοι τοπικής στοίχισης (FASTA, BLAST), οι οποίοι χρησιμοποιούνται καθημερινά στη βιοπληροφορική.

Προαπαιτούμενη γνώση
Προαπαιτούμενη γνώση για το κεφάλαιο αυτό, είναι η γνώση των βασικών νόμων των πιθανοτήτων και στοιχειώδεις γνώσεις σχετικά με τις βιολογικές αλληλουχίες..

3. Εισαγωγή

Η ομοιότητα αλληλουχιών είναι ένα από τα θεμελιώδη ζητήματα στη Βιοπληροφορική, καθώς πλέον αποτελεί αναπόσπαστο τμήμα των αναλύσεων που πραγματοποιεί καθημερινά οποιοσδήποτε ασχολείται με το γνωστικό αυτό αντικείμενο, αλλά, ακόμα περισσότερο, ο καθένας που ασχολείται ερευνητικά με τη μοριακή βιολογία με οποιονδήποτε τρόπο. Η ομοιότητα των βιολογικών αλληλουχιών τις περισσότερες φορές υποδηλώνει ομολογία (δηλαδή, κοινή εξελικτική προέλευση), και κατά συνέπεια (ειδικά για τις πρωτεΐνες), παρόμοια τρισδιάστατη δομή και παρόμοια λειτουργία.

Τα προβλήματα που καλείται κάποιος να λύσει, όταν μελετάει την ομοιότητα αλληλουχιών, είναι πολλαπλά. Με ποιον αλγόριθμο θα πραγματοποιήσει την «στοίχιση» των δύο αλληλουχιών (δηλαδή την εύρεση της καλύτερης περιοχής ομοιότητας τους); Πώς θα ποσοτικοποιήσει αυτή την ομοιότητα; Τι υποθέσεις θα αναγκαστεί να κάνει; Και τέλος, πως θα αξιολογήσει αν μια στοίχιση είναι σημαντική ή όχι; Το τελευταίο, είναι ίσως και το σπουδαιότερο από τα θέματα αυτά, γιατί όλοι καταλαβαίνουν ότι αν δυο πρωτεϊνικές αλληλουχίες είναι ταυτόσημες λ.χ. σε ποσοστό 99%, τότε υπάρχει πολύ μεγάλη πιθανότητα να είναι και όμοιας δομής και παρόμοιας λειτουργίας (εκτός ίσως από τις περιπτώσεις στις οποίες οι λίγες αλλαγές συμβαίνουν στο ενεργό κέντρο ενός ενζύμου και αναστέλλουν τη δράση του). Με 80% ομοιότητα περιμένουμε ότι πάλι οι πρωτεΐνες θα έχουν μεγάλη ομοιότητα στη δομή. Ποιο είναι όμως το όριο όσο κατεβαίνουμε στο επίπεδο ομοιότητας; Παραδοσιακά οι βιολόγοι χρησιμοποιούν τον εμπειρικό κανόνα του «30% ομοιότητα σε μήκος στοίχισης μεγαλύτερο από 80 αμινοξικά κατάλοιπα», κανόνας που σε γενικές γραμμές λειτουργεί σωστά, αλλά χρειαζόμαστε περισσότερη ακρίβεια σε τέτοια ζητήματα, ειδικά όσο οι βάσεις δεδομένων μεγαλώνουν και οι πιθανότητες εμφάνισης μιας τυχαίας ομοιότητας αυξάνονται.

Στο κεφάλαιο αυτό, θα προσπαθήσουμε να παρουσιάσουμε τα βασικά θεωρητικά εργαλεία που θα μας βοηθήσουν να καταλάβουμε τις απαντήσεις των παραπάνω ερωτημάτων, αλλά επίσης και να αναγνωρίσουμε τους περιορισμούς τους. Για το λόγο αυτό, θα ξεκινήσουμε από τη στατιστική μελέτη των βιολογικών αλληλουχιών και θα παρουσιάσουμε το βασικό μοντέλο της ανεξαρτησίας, το οποίο αποτελεί το «θεωρητικό» ή, σε μια πιο στατιστική ορολογία, αποτελεί το μοντέλο από το οποίο προκύπτει η «μηδενική υπόθεση» έναντι της οποίας θα συγκρίνουμε τα ευρήματα μιας αναζήτησης ομοιότητας, έτσι ώστε να μπορέσουμε να καταλάβουμε αν η δεδομένη στοίχιση είναι «σημαντική» ή όχι. Στη συνέχεια θα παρουσιαστούν οι κύριοι αλγόριθμοι εύρεσης ομοιότητας, και θα συζητηθούν πρακτικά θέματα που προκύπτουν, ειδικά σε αναζητήσεις σε βάσεις δεδομένων.

3.1 Η ακολουθία ως σειρά ανεξάρτητων γεγονότων

Το πιο απλό μοντέλο που περιγράφει μια βιολογική αλληλουχία (DNA, RNA ή πρωτεΐνης) είναι το μοντέλο της ανεξαρτησίας, δηλαδή το μοντέλο που θεωρεί ότι η αλληλουχία των γραμμάτων του αλφάβητου Ω, -στην

περίπτωση του DNA των τεσσάρων νουκλεοτιδίων-, είναι μια σειρά *n* ανεξάρτητων δοκιμών με τέσσερις διακριτές εκβάσεις. Οι πιθανότητες για τα 4 ενδεχόμενα (A, T, G, C) είναι αντίστοιχα:

$$p_A, p_T, p_G, p_C \quad \text{με } p_k \geq 0 \text{ και } \sum_{k\in\{A,T,G,C\}} p_k = 1.$$

Όμοια, ισχύουν και στην περίπτωση των πρωτεϊνών, μόνο που θα έχουμε 20 διαφορετικά σύμβολα και 20 διαφορετικές πιθανότητες. Μια δεδομένη ακολουθία DNA, $\mathbf{x}=x_1, x_2,...,x_n$ με $x_i \in \{A,T,G,C\}$ έχει συνολική πιθανότητα να παρατηρηθεί κάτω από τις προϋποθέσεις του «τυχαίου» αυτού μοντέλου ίση με:

$$p_{o\lambda} = P(\mathbf{x}) = \prod_{i=1}^{n} p_{X_i}$$

Προφανώς το άθροισμα των πιθανοτήτων όλων των πιθανών ακολουθιών (που είναι όσες οι δυνατές διατάξεις των 4 στοιχείων ανά *n* με επανάληψη δηλαδή 4^n) είναι ίσο με 1 δηλαδή:

$$\sum_j P(\mathbf{x}_j) = 1.$$

Έστω **x** μια τέτοια τυχαία ακολουθία *n* βάσεων του DNA. Η συχνότητα της εμφάνισης των 4 βάσεων ακολουθεί την πολυωνυμική κατανομή, δηλαδή:

$$P\left(n_A, n_T, n_G, n_C\right) = \frac{n!}{n_A!\, n_T!\, n_G!\, n_C!}\, p_A^{n_A}\, p_T^{n_T}\, p_G^{n_G}\, p_C^{n_C} \tag{3.1}$$

Αν τώρα θεωρήσουμε τις συχνότητες εμφάνισης κάθε μιας από τις βάσεις ξεχωριστά, τότε αυτές ακολουθούν τη διωνυμική κατανομή, δηλαδή :

$$P\left(X = x\right) = \binom{n}{x} p_A^x \left(1 - p_A\right)^{n-x} \tag{3.2}$$

και όμοια για τις άλλες 3 βάσεις (T, G, C). Έτσι η μια ακολουθία των βάσεων του DNA μπορεί να θεωρείται ως μια ακολουθία ανεξάρτητων δοκιμών Bernoulli με $p=p_A$ και $q=1-p_A$. Όμοια θεώρηση μπορεί να γίνει και για τις άλλες 3 βάσεις. Προφανώς η κατανομή των συχνοτήτων εμφάνισης των βάσεων του DNA (ή των αμινοξέων μιας πρωτεΐνης) δεν είναι επαρκής πληροφορία για να περιγράψει τη βιολογική πληροφορία μιας δεδομένης αλληλουχίας. Η βιολογική σημασία μιας αλληλουχίας βάσεων (ή αμινοξέων) έγκειται στην ακριβή αλληλουχία των 4 βάσεων (ή των 20 αμινοξέων), δηλαδή στον τρόπο που το ένα σύμβολο διαδέχεται το άλλο. Εντούτοις, η παραπάνω θεώρηση της τυχαίας και ανεξάρτητης εμφάνισης των συμβόλων μας είναι ιδιαίτερα χρήσιμη καθώς μας προμηθεύει με μια μηδενική υπόθεση (H_0) έναντι της οποίας θα μπορούμε να συγκρίνουμε μια δεδομένη αλληλουχία για να διαπιστώσουμε αν η -συγκεκριμένη αλληλουχία- είναι δυνατόν να έχει προκύψει τυχαία, ή αν, αντίθετα, έχει κάποια βιολογική σημασία (Durbin, Eddy, Krogh, & Mithison, 1998).

Στο μοντέλο αυτό, μια επιπλέον υπόθεση που μπορούμε να κάνουμε (αν δεν έχουμε λόγους να πιστεύουμε το αντίθετο, όπως για παράδειγμα αν έχουμε μια αλληλουχία από ένα γονιδίωμα με γνωστές τις συχνότητες εμφάνισης των βάσεων) είναι ότι τα ενδεχόμενα εμφάνισης των βάσεων εκτός από ανεξάρτητα είναι και ισοπίθανα, Δηλαδή:

$$p_A = p_T = p_G = p_C = 1/4$$

Πριν προχωρήσουμε παρακάτω θα πρέπει να κάνουμε μια μικρή παρένθεση για να παραθέσουμε κάποιους ορισμούς δανεισμένους από την Θεωρία Πληροφορίας (Information Theory). Συγκεκριμένα θα αποδώσουμε τον ορισμό της έννοιας της εντροπίας και της πληροφορίας. Μια δεδομένη αλληλουχία DNA, όπως την ορίσαμε παραπάνω, λέμε ότι έχει συνάρτηση εντροπίας κατά Shannon ίση με:

$$H(\mathbf{x}) = -\sum_i P(x_i) \log P(x_i) \tag{3.3}$$

Η εντροπία γίνεται μέγιστη όταν οι βάσεις είναι ισοπίθανες, δηλαδή όταν $p_A=p_G=p_T=P_C=1/4$ οπότε θα έχει τιμή ίση με $H(\mathbf{x})=\sum(1/4)\log(1/4)=\log 4$. Συνήθως σε αυτές τις περιπτώσεις παίρνουμε λογάριθμους με βάση το 2, έτσι ώστε η μονάδα μέτρησης να είναι το bit. Η πληροφορία μιας ακολουθίας ορίζεται ως:

$$I(\mathbf{x}) = H_{\max} - H_{obs} \tag{3.4}$$

άρα αν έχουμε μια αλληλουχία με σύσταση βάσεων διαφορετική από την αναμενόμενη με βάση το τυχαίο μοντέλο η εντροπία της θα είναι μικρότερη από τα 2 bits, και η πληροφορία που φέρει αυτή η αλληλουχία θα είναι μεγαλύτερη από το 0.

Ένα διαφορετικό μέτρο για το πληροφοριακό περιεχόμενο μιας βιολογικής αλληλουχίας, έχει δοθεί από τους (Wootton & Federhen, 1993) και βασίζεται επίσης στη Θεωρία Πληροφορίας. Αυτό το μέτρο, που ονομάστηκε "πολυπλοκότητα της σύνθεσης" (compositional complexity), ορίζεται, για ένα παράθυρο μήκους k της ακολουθίας, ως εξής:

$$K = \frac{1}{k}\log_{N_\Omega}\left(\frac{k!}{\prod_{\forall s\in\Omega} n_s!}\right) \tag{3.5}$$

Στην παραπάνω σχέση, το n_s είναι ο αριθμός εμφανίσεων του συμβόλου s στο παράθυρο και N_Ω το μέγεθος του αλφάβητου (4 για τα νουκλεοτίδια, 20 για τα αμινοξέα). Διαισθητικά, το μέτρο αυτό δείχνει την ποσότητα της πληροφορίας που απαιτείται σε κάθε θέση της ακολουθίας για να καθορίσει κανείς το σύμβολο (της θέσης), δεδομένης της σύνθεσης όλου του παραθύρου. Για παράδειγμα, ένα παράθυρο 4 νουκλεοτιδίων με σύσταση AAAA, θα έχει πολυπλοκότητα ίση με

$$K = \frac{1}{4}\log_4\left(\frac{4!}{4!0!0!0!}\right) = \frac{1}{4}\log_4(1) = 0\,.$$

Πράγμα που σημαίνει, ότι αν ξέρουμε την ακολουθία του παραθύρου, δεν χρειαζόμαστε καμιά άλλη πληροφορία για να βρούμε ποιο κατάλοιπο βρίσκεται σε μια δεδομένη θέση. Αντίθετα, ένα παράθυρο με σύσταση ATGC θα έχει

$$K = \frac{1}{4}\log_4\left(\frac{4!}{1!1!1!1!}\right) = \frac{1}{4}\log_4(24) = 0.573$$

Οι Wootton και Federhen χρησιμοποίησαν επίσης και την εντροπία, δίνοντας όμως έναν ισοδύναμο ορισμό:

$$H_k = -\sum_{\forall s\in\Omega}\frac{n_s}{k}\left(\log_2\frac{n_s}{k}\right) \tag{3.6}$$

Στην ίδια εργασία, έδειξαν ότι η εντροπία και η πολυπλοκότητα, είναι ασυμπτωτικά ισοδύναμες ποσότητες (δηλαδή, όταν το παράθυρο είναι πολύ μεγάλο θα δίνουν το ίδιο αποτέλεσμα), ενώ έκαναν την παρατήρηση ότι η πολυπλοκότητα όπως την όρισαν, είναι σύμφωνη με τον ορισμό περί εντροπίας του Boltzman (σε αντίθεση με τον κλασικό ορισμό της έννοιας της εντροπίας κατά Shannon).

Η χρήση της εντροπίας, της πολυπλοκότητας και της πληροφορίας, βρίσκουν πολλές εφαρμογές σε προκαταρκτικές περιγραφικές αναλύσεις γονιδιωμάτων, αλλά και σε άλλες πιο εξειδικευμένες αναλύσεις. Οι Wootton και Federhen για παράδειγμα, χρησιμοποίησαν τα μέτρα αυτά για τον εντοπισμό περιοχών χαμηλής πολυπλοκότητας σε αμινοξικές αλληλουχίες πρωτεϊνών ή σε γονιδιώματα. Η ανεύρεση τέτοιων περιοχών είναι σημαντική, γιατί στην περίπτωση αμινοξικών αλληλουχιών πρωτεϊνών η ύπαρξή τους μπορεί να επηρεάσει τα στατιστικά της στοίχισης και τα αποτελέσματα της αναζήτησης ομοιότητας (βλ παρακάτω), ενώ στην περίπτωση DNA μπορεί να σηματοδοτεί την ύπαρξη ρυθμιστικών περιοχών. Τέλος, όπως θα δούμε στο επόμενο κεφάλαιο, η εντροπία χρησιμεύει στην περιγραφή και στην ποσοτικοποίηση μιας πολλαπλής στοίχισης αλληλουχιών.

Μια άλλη σχετική έννοια, είναι αυτή της σχετικής εντροπίας (Relative Entropy). Η σχετική εντροπία δυο καταστάσεων P, Q (γνωστή και ως μέτρο της απόστασης των Kullback-Leibler) εκφράζει τη σχετική απόσταση, ή διαφορά, μεταξύ των δυο καταστάσεων και δίνεται από τον τύπο:

$$H(P,Q) = \sum_i P(x_i)\log\frac{P(x_i)}{Q(x_i)} \tag{3.7}$$

Το $P(x_i)$ είναι όπως είδαμε παραπάνω η πιθανότητα εμφάνισης μιας βάσης (A,T,G,C) στην i θέση της συγκεκριμένης ακολουθίας, ενώ το $Q(x_i)$ η αντίστοιχη πιθανότητα εμφάνισης μιας βάσης σε μια άλλη ακολουθία. Αυτή η άλλη ακολουθία μπορεί να είναι μια άλλη πραγματική ακολουθία με την οποία θέλουμε να συγκρίνουμε την πρώτη, ή να είναι μια θεωρητική κατανομή, όπως αυτή που υποθέτει ισοπίθανη ή τυχαία εμφάνιση των βάσεων. Προφανώς αν $Q(x_i)=1/4$ (ισοκατανομή των βάσεων) τότε $H(P,Q)=I(P)$

Μια άλλη πολύ σημαντική έννοια που θα ξανασυναντήσουμε και στα επόμενα κεφάλαια είναι αυτή της αμοιβαίας πληροφορίας (Mutual Information). Δυο τ.μ X,Y έχουν αμοιβαία πληροφορία που δίνεται από τη σχέση:

$$M(X,Y)=\sum_{i,j}P(x_i,y_j)\log\frac{P(x_i,y_j)}{P(x_i)P(y_j)} \tag{3.8}$$

Σε αυτή την περίπτωση, έχουμε δυο ακολουθίες, **x** και **y**. Η αμοιβαία πληροφορία μετράει πόση διάφορα έχει η από κοινού κατανομή της σ.π. των X και Y που συμβολίζουμε με $P(x_i,y_i)$, με την υποθετική από κοινού κατανομή που θα είχαν αν ήταν ανεξάρτητες με $P(x_i,y_i)=P(x_i)P(y_i)$. Προφανώς $P(x_i)$ και $P(y_i)$ είναι οι περιθώριες σ.π. των X,Y αντίστοιχα. Δηλαδή, η αμοιβαία πληροφορία μετράει το «πόσο ανεξάρτητες» είναι οι δυο κατανομές. Η σχετική εντροπία και η αμοιβαία πληροφορία, βρίσκουν πολλές εφαρμογές όταν μελετάμε ταυτόχρονα πολλές ακολουθίες και σχετικά παραδείγματα θα δούμε στο κεφάλαιο που περιγράφει την πολλαπλή στοίχιση.

3.2 Ροές - Νόμος Erdos και Renyi

Το επόμενο θέμα που θα μας απασχολήσει είναι γνωστό στη βιβλιογραφία ως το πρόβλημα της μέγιστης ροής όμοιων αποτελεσμάτων (longest run of heads). Η πιο απλή του εφαρμογή είναι η απάντηση στο ερώτημα «ποια είναι η αναμενόμενη τιμή για το μέγιστο αριθμό επαναλήψεων-κορώνα ή γράμματα-σε μια διαδοχική σειρά από n διαδοχικά στριψίματα ενός νομίσματος (δίτιμες δοκιμές Bernoulli)». Το θέμα αυτό είναι πολύ σημαντικό, καθώς στη θεωρία των ροών βασίζονται τα στατιστικά της τοπικής στοίχισης ακολουθιών, τα οποία θα μελετήσουμε παρακάτω.

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

Εικόνα 3.1: Μια ροή από 16 συνεχόμενες Α σε ένα μόριο DNA

Η πρώτη απάντηση που δόθηκε στο ερώτημα αυτό είναι γνωστή ως νόμος του $\log(n)$, ή αλλιώς γνωστός ως νόμος των Erdos και Renyi (Erdos & Renyi, 1970). Το θεώρημα λέει ότι σε μια ακολουθία n ανεξάρτητων δοκιμών Bernoulli με πιθανότητα «επιτυχίας» p, με $0\le p\le 1$, το αναμενόμενο μήκος R_n μέγιστης δυνατής ροής ευνοϊκών αποτελεσμάτων, είναι ίσο κατά προσέγγιση με $\log_{1/p}(n)$ ή αλλιώς:

$$\frac{R_n}{\log_{1/p}(n)}\to 1 \text{ με πιθανότητα 1.} \tag{3.9}$$

Η απόδειξη είναι αρκετά περίπλοκη αλλά μια διαισθητική ερμηνεία του αποτελέσματος μπορεί να γίνει ως εξής (M. S. Waterman, 1995): αν το ευνοϊκό αποτέλεσμα έχει πιθανότητα p τότε μια ροή k συνεχών ευνοϊκών αποτελεσμάτων έχει πιθανότητα p^k. Αν έχουμε n επαναλήψεις ($n\to+\infty$) τότε έχουμε περίπου n δυνατές ροές και

$$E(\text{αριθμός ροών μήκους } x)=np^k$$

Αν τώρα, η μέγιστη ροή είναι μοναδική, το μήκος της, R_n, ικανοποιεί τη σχέση $1=np^k$, άρα:

$$R_n=\log_{1/p}(n)$$

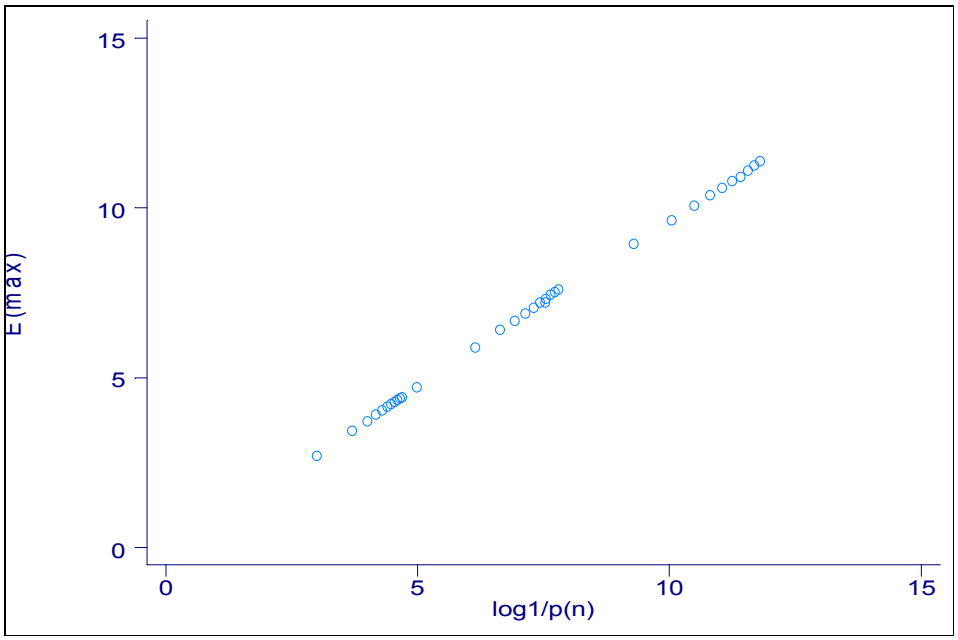
Παρατηρούμε, ότι όσο μεγαλώνει το μήκος της ακολουθίας, τόσο μεγαλώνει και το μήκος της μέγιστης ροής που αναμένουμε να βρούμε λόγω τύχης.

Παράδειγμα 3.2.1

Σε μια ακολουθία n=10000 βάσεων του DNA, θεωρώντας αυτές ισοπίθανες (δηλαδή p_κ=1/4), μας ενδιαφέρει να βρούμε τον αριθμό των μέγιστων επαναλήψεων Α που μπορεί να έχει συμβεί κατά τύχη. Δηλαδή, θεωρώντας ότι η αλληλουχία είναι τυχαία, τότε το μέγιστο μήκος ροής από Α θα είναι:

$$R_n=\log_{1/p}(n)\Rightarrow R_n=\log_4(10000)\Rightarrow R_n=\frac{\log_{10}10000}{\log_{10}4}=\frac{4}{0.60205}=6.64$$

Στα παρακάτω διαγράμματα φαίνονται τα αποτελέσματα των προσομοιώσεων για τη μέση τιμή της ροής ευνοϊκών αποτελεσμάτων για n=1000 έως 50000 και για p=0.1, 0.25, 0.4 (1000 επαναλήψεις) οι οποίες επιβεβαιώνουν τη σχέση (3.9).



Εικόνα 3.2: *Αποτελέσματα προσομοίωσης για τη μέγιστη ροή ευνοϊκών αποτελεσμάτων, σε ένα μοντέλο διωνυμικής κατανομής με πιθανότητες p=0.1, 0.25 και 0.4. Το n κυμαίνεται από 1,000 μέχρι 50,000.*

3.3 Επεκτάσεις στον Νόμο Erdos και Renyi

Μέχρι τώρα ασχοληθήκαμε μόνο με τον αριθμό των επαναλήψεων σε μια σειρά ανεξάρτητων δοκιμών. Παρ' όλα αυτά όμως, ξέρουμε ότι υπάρχουν περιπτώσεις στις οποίες μπορεί να μας ενδιαφέρει ο αριθμός των δοκιμών που περιέχουν π.χ. 90% επαναλήψεις από «επιτυχίες». Ένα τέτοιο παράδειγμα, είναι ο εντοπισμός περιοχών με «πολλά» και συνεχόμενα υδρόφοβα κατάλοιπα σε μια πρωτεΐνη. Ξέρουμε ότι οι περιοχές με συνεχόμενα υδρόφοβα κατάλοιπα είναι πιθανό να είναι διαμεμβρανικά τμήματα, αλλά δεν αναμένουμε να συναντήσουμε περιοχές αποκλειστικά με υδρόφοβα αμινοξέα (είναι γνωστό, ότι ακόμα και μέσα σε πραγματικά διαμεμβρανικά τμήματα πρωτεϊνών συναντάμε περιστασιακά 1-2 πολικά κατάλοιπα). Επίσης, ένα άλλο χαρακτηριστικό που μπορεί να μας ενδιαφέρει, είναι το να προσδιορίσουμε τη στατιστική σημαντικότητα μιας τέτοιας παρατήρησης.

Σ' αυτήν την ενότητα θα παρουσιάσουμε κάποια πορίσματα της θεωρίας των μεγάλων αποκλίσεων (Large Deviation Theory) και θα τα χρησιμοποιήσουμε για να επεκτείνουμε τα αποτελέσματα των προηγούμενων παραγράφων. Στην Εικόνα 3.3 φαίνεται γραμμοσκιασμένη μια περιοχή 20 βάσεων η οποία περιέχει 16 Αδενίνες. Σε μια αλληλουχία που θεωρείται τυχαία (και άρα η συχνότητα εμφάνισης των βάσεων δεν έχει λόγο να αποκλίνει από τη συνολική συχνότητα εμφάνισης σε όλη την αλληλουχία), μας ενδιαφέρει το πόσο συχνά μπορεί να εμφανιστεί μια τέτοια περιοχή.

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

Εικόνα 3.3: *Μια ροή 20 νουκλεοτιδίων που περιέχει 80% A.*

Η σχετική εντροπία δυο καταστάσεων α, p εκφράζει τη σχετική απόσταση, δηλαδή τη διαφορά μεταξύ των δυο καταστάσεων και, ειδικά για την περίπτωση της διωνυμικής κατανομής δίνεται από τον τύπο:

$$H(\alpha, p) \equiv \alpha \log\left(\frac{\alpha}{p}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-p}\right) = \log\frac{a^a(1-a)^{1-a}}{p^a(1-p)^{1-a}} = -\log\left(\frac{p}{a}\right)^a\left(\frac{1-p}{1-a}\right)^{1-a} \qquad (3.10)$$

Η συνάρτηση αυτή μετρά τη διαφορά μεταξύ της κατανομής B(k,p) από την οποία προέρχονται τα δεδομένα μας (η οποία έχει δώσει γένεση σε μια ακολουθία DNA με πιθανότητα εμφάνισης των βάσεων ίση με p), και μιας άλλης, υποθετικής, B(k,α) για την οποία υποπτευόμαστε ότι έχει δώσει γένεση σε μια τοπική υπό-ακολουθία μήκους n στην οποία παρατηρούμε ότι για παράδειγμα η εμφάνιση μιας βάσης, διαφέρει πολύ

από την αναμενόμενη καθώς έχει συχνότητα $\alpha=s/k$. Προφανώς $\alpha, p \in (0,1)$. Το κλειδί στην κατανόηση των μεγάλων αποκλίσεων, είναι το γεγονός ότι έχουμε να κάνουμε με δυο διαφορετικές πιθανότητες (α,p) στον ίδιο χώρο πιθανών εκβάσεων. Ένα από τα αποτελέσματα της θεωρίας μεγάλων αποκλίσεων, έχει αρκετό ενδιαφέρον και βρίσκει εφαρμογές στον υπολογισμό διωνυμικών πιθανοτήτων. Συγκεκριμένα, αν έχουμε $0 \le p \le a \le 1$ και $Y \sim B(k,p)$, τότε μια προσεγγιστική σχέση για την διωνυμική πιθανότητα $P(Y \ge ak)$, δίνεται από τον τύπο:

$$P(Y \ge \alpha k) \approx e^{-kH(a,p)} \tag{3.11}$$

Τώρα που έχουμε δώσει τον ορισμό της έννοιας της σχετικής εντροπίας μπορούμε να προχωρήσουμε και να επεκτείνουμε τη σχέση (3.9). Το αποτέλεσμα αυτό, λέει ότι σε μια ακολουθία n ανεξάρτητων δοκιμών Bernoulli με πιθανότητα «επιτυχίας» p, με $0 \le p \le a \le 1$, το πλήθος R_n^a διαδοχικών δοκιμών που περιέχουν $100\alpha\%$ ευνοϊκά αποτελέσματα, ικανοποιεί τη σχέση (Erdos & Renyi, 1970; Erdos & Revesz, 1975):

$$\frac{R_n^a}{\log(n)} \to \frac{1}{H(a,p)} \text{ με πιθανότητα 1} \tag{3.12}$$

Μια διαισθητική ερμηνεία του αποτελέσματος έχει ως εξής: Από τη θεωρία των μεγάλων αποκλίσεων (Large Deviations) της σχέσης (3.11) βρίσκουμε ότι μια περιοχή μήκους k η οποία περιέχει $100\alpha\%$ ευνοϊκά αποτελέσματα, έχει πιθανότητα περίπου ίση με $e^{-kH(a,p)}$. Επειδή τώρα κάθε ροή έχει περίπου $n-k+1 \approx n$ δυνατές περιοχές έναρξης έχουμε:

$$1 = ne^{-kH(a,p)} \Rightarrow R_n^a = \frac{\log(n)}{H(a,p)}$$

Παρατηρούμε, με την χρήση του κανόνα De L' Hospital, ότι για $a = 1 \Rightarrow H(a,p) = \log(1/p)$ και τα αποτελέσματα συμφωνούν με τη σχέση (3.9).

Παράδειγμα 3.3.1

Σε μια αλληλουχία 10.00.000 βάσεων DNA η μέγιστη περιοχή (ροή) R_n^a που να περιέχει κατ' ελάχιστο 80% βάσεις Αδενίνης (A) είναι :

$$R_n^a = \frac{\log(n)}{H(a,p)} = \frac{\log(1000000)}{0.666} = 20.744$$ (να σημειωθεί εδώ ότι όταν γράφουμε log εννοούμε λογάριθμο με βάση το e)

3.4 Η Κατανομή της Μέγιστης Ροής - Η Κατανομή των Ακραίων Τιμών (EVD)

Επεκτείνοντας τα προηγούμενα, πολλές φορές μπορεί να χρειαστεί να βρούμε την προσεγγιστική κατανομή που ακολουθεί η τυχαία μεταβλητή του μήκους της μέγιστης ροής ενός αποτελέσματος. Η πλήρης κατανομή, μας είναι χρήσιμη, και από θεωρητική σκοπιά, αλλά κυρίως από πρακτική, γιατί με τη γνώση της κατανομής θα μπορούμε να πραγματοποιήσουμε έλεγχο υποθέσεων (χρειαζόμαστε εκτός από τη μέση τιμή, και τη διασπορά της τ.μ.). Όταν αναφερόμαστε σε μέγιστα (ή και σε ελάχιστα) μιας ακολουθίας τυχαίων μεταβλητών καταλήγουμε συνήθως στις κατανομές των ακραίων τιμών (Extreme Value Distributions). Πιο αυστηρά, αν έχουμε ένα δείγμα $X_1, X_2, ..., X_n$, από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές (iid) τότε μας ενδιαφέρει η οριακή κατανομή του:

$$M_n = a_n[\max(X_1, X_2, ..., X_n) - b_n], n \to \infty$$

Όπου αn, bn κατάλληλες σταθερές κανονικοποίησης τέτοιες ώστε να προκύπτει μη τετριμμένη κατανομή. Η απάντηση είναι ότι αν υπάρχει μια μη τετριμμένη και ορισμένη ΑΣΚ (cdf) για κάποιες ακολουθίες αn, bn, τότε πρέπει να ανήκει σε μια από τις περιπτώσεις (Davison, 1998):

1. $F(y) = \exp(-e^{-y}), -\infty \le y \le \infty$ (Gumbel)
2. $F(y) = \begin{cases} 0, y \le 0 \\ \exp(-y^{-a}), y \ge 0, a > 0 \end{cases}$ (Frechet)

3. $F(y)=\begin{cases}\exp(-(-y)^{a}), y<0, a>0\\ 1, y\geq 0\end{cases}$ (Weibull)

Η κατανομή που αφορά την δική μας περίπτωση είναι αυτή του Gumbel, και προκύπτει από την γενικευμένη μορφή της κατανομής των ακραίων τιμών (Generalized Extreme Value Distribution – GEVD):

$$H(y)=\exp\left\{-\left(1+k\left(\frac{y-a}{b}\right)\right)^{-\frac{1}{k}}\right\} \text{ με } -\infty<\alpha,k<\infty,b>0 \tag{3.13}$$

η οποία ορίζεται όταν $1+k\left(\frac{y-a}{b}\right)>0$, ως το όριο καθώς $k\to 0$ (οι άλλες δύο μορφές αντιστοιχούν στην περίπτωση που k>0 (Frechet) και k<0 (Weibull) αντίστοιχα). Αν θέσουμε $z=\left(\frac{y-a}{b}\right)$, και $t=-\frac{1}{k}$ θα έχουμε:

$$H(y)=\exp\left\{-\left(1-\frac{z}{t}\right)^{t}\right\}$$

και αν πάρουμε το όριο καθώς το $k\to 0\Rightarrow t\to\infty$ επειδή είναι γνωστή η σχέση:

$$\lim_{n\to\infty}\left(1-\frac{a}{n}\right)^{n}=e^{-a}$$

θα έχουμε

$$\lim_{t\to\infty}H(y)=\lim_{t\to\infty}\exp\left\{-\left(1-\frac{z}{t}\right)^{t}\right\}=\exp\left\{-e^{-z}\right\}=\exp\left\{-e^{-\left(\frac{y-a}{b}\right)}\right\}$$

Έτσι η κατανομή του $Y_n=\max(X_1,X_2,...,X_n)$ γίνεται (Gumbel, 1958):

$$F(Y)=\exp(-e^{-\frac{(y-a)}{b}}), -\infty\leq y\leq\infty \tag{3.14}$$

$$E=a-b\Gamma'(1),\quad V=\frac{b^2\pi^2}{6} \tag{3.15}$$

Αποδεικνύεται (R. Arratia, Gordon, L. and Waterman, 1986; R. Arratia, Gordon, L. and Waterman, M. S., 1990; M. S. Waterman, 1995), ότι στην περίπτωση της συνεχούς ροής ενός αποτελέσματος (νόμισμα η βάσεις DNA) για $a_n=\frac{\log(qn)}{\lambda}, b_n=\frac{1}{\lambda}$ όπου $\lambda=\log\left(1/p\right)$ ισχύει:

$$\lim_{n\to\infty}\left(R_n<\frac{\log(nq)}{\lambda}+\frac{y}{\lambda}\right)=\exp\left(-e^{-y}\right)$$

Για την ΑΣΚ της τ.μ. R_n θα ισχύει:

$$F\left(y\right)=P\left(R_n\leq y\right)\approx\exp\left(-\exp\left(-\frac{y-\log(nq)/\lambda}{1/\lambda}\right)\right) \tag{3.16}$$

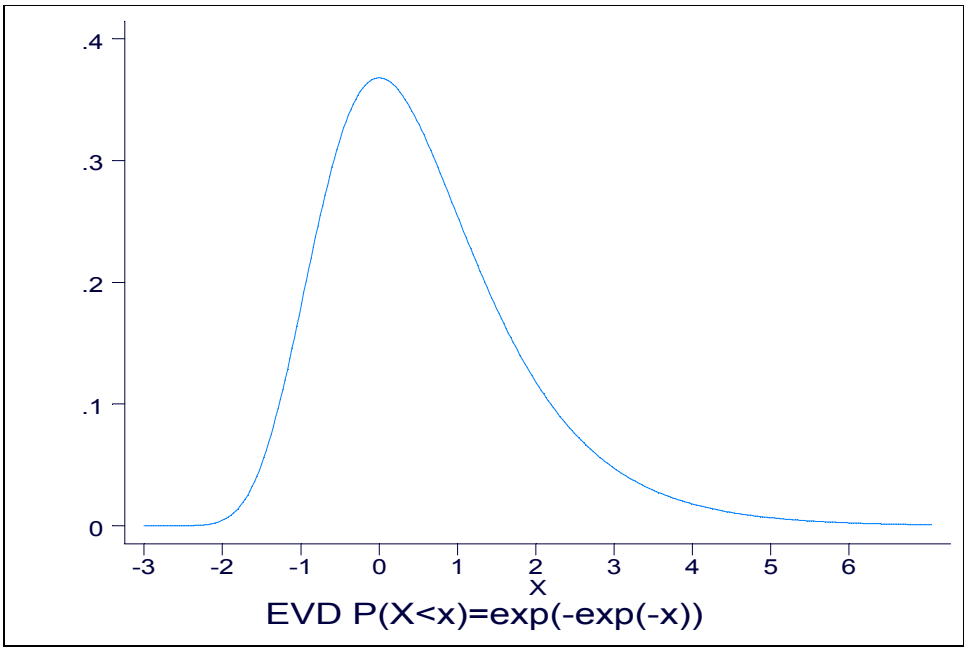
Από τα παραπάνω προκύπτει, ότι η κατανομή της μέγιστης ροής είναι αυτή των ακραίων τιμών του Gumbel. Δηλαδή, οι σχέσεις (3.16) και (3.14) είναι ισοδύναμες. Κατά συνέπεια, θα έχουμε:

$$E\left(R_n\right)\approx\frac{\log(n)}{\lambda}+\frac{\log(q)}{\lambda}+\frac{\gamma}{\lambda}-\frac{1}{2}\Rightarrow E\left(R_n\right)\approx\log_{1/p}(n)+\log_{1/p}(q)+\gamma/\lambda-1/2 \tag{3.17}$$

και

$$\text{var}\left(R_n\right)\approx\frac{\pi^2}{6\lambda^2}+\frac{1}{12} \tag{3.18}$$

όπου $\gamma = -\Gamma'(1) = 0.5772\ldots$ η σταθερά Euler-Mascheroni. Η αφαίρεση από τη μέση τιμή του ½ και η πρόσθεση στη διασπορά 1/12 είναι η διόρθωση συνέχειας του Sheppard, και γίνεται διότι όταν μετατρέπουμε μια συνεχή τ.μ. σε διακριτή αυξάνεται η μέση τιμή της και μειώνεται η διασπορά.



Εικόνα 3.4: *Η γραφική παράσταση της κατανομής του Gumbel*

Παράδειγμα 3.6.1

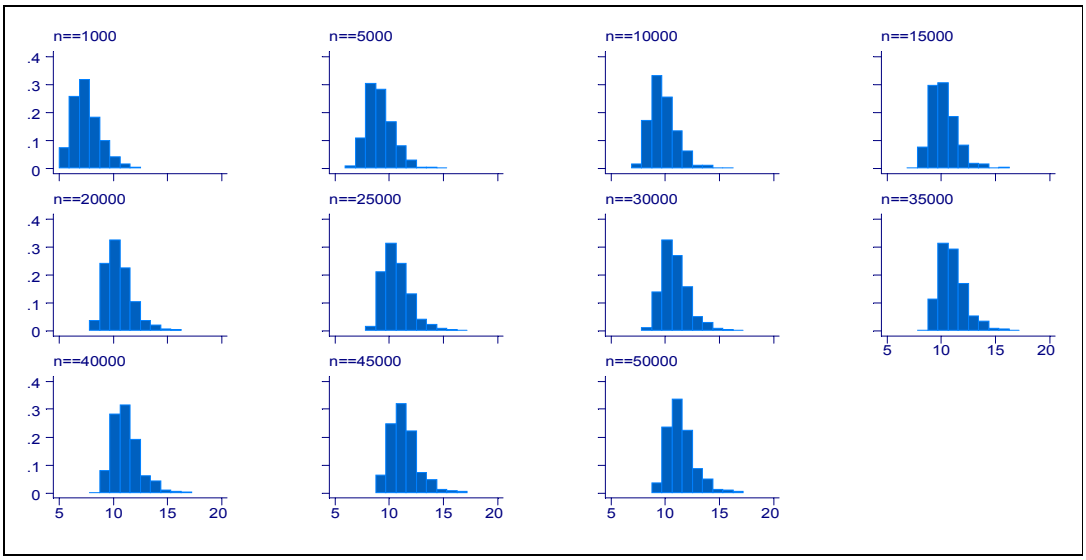
Αν χρησιμοποιήσουμε τις σχέσεις (3.17) και (3.18) στα δεδομένα του παραδείγματος 3.2.1 έχουμε:

$$E(R_n) \approx \log_{1/p}(n) + \log_{1/p}(q) + \gamma/\lambda - 1/2 \Rightarrow$$

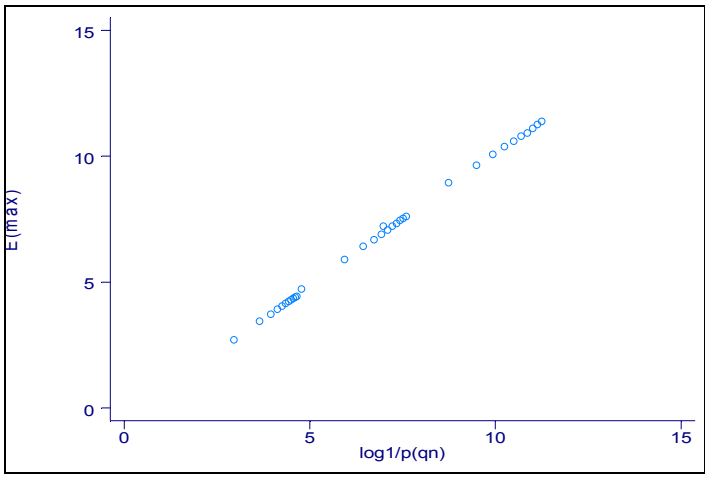
$$E(R_n) \approx \log_4(10000) + \log_4(3/4) + \frac{0.5772}{\log(4)} - 1/2 = 6.3518$$

και $\text{var}(R_n) \approx \dfrac{\pi^2}{6\lambda^2} + \dfrac{1}{12} = 0.939$

Παρατηρούμε, ότι παρόλο που η σχέση (3.17), είναι καλύτερη προσέγγιση του αναμενόμενου μήκους της μέγιστης ροής, η πραγματική διαφορά που βρίσκουμε σε σχέση με την πιο απλή εκδοχή, όπως αυτή αποτυπώθηκε στη σχέση (3.9), είναι αρκετά μικρή. Τούτο συμβαίνει, γιατί ο κύριος παράγοντας που καθορίζει την τελική τιμή εξακολουθεί να είναι η τιμή του $\log(n)$, καθώς το $\log(q)$ και το γ/λ είναι σχετικά μικρές ποσότητες. Προφανώς, όσο το n μεγαλώνει, η διαφορά θα γίνεται ακόμα μικρότερη. Από τις προσομοιώσεις, βλέπουμε ξεκάθαρα ότι το μήκος R_n των ροών ακολουθεί την κατανομή του Gumbel (EVD) με μέση τιμή και διασπορά που δίνονται από τις σχέσεις (3.17) και (3.18).



Εικόνα 3.5: *Αποτελέσματα προσομοίωσης για την κατανομή της μέγιστης ροής ευνοϊκών αποτελεσμάτων, σε ένα μοντέλο διωνυμικής κατανομής με πιθανότητες p=0.1, 0.25 και 0.4. Το n κυμαίνεται από 1,000 μέχρι 50,000.*



Εικόνα 3.6: *Αποτελέσματα προσομοίωσης για τη μέγιστη ροή ευνοϊκών αποτελεσμάτων, σε ένα μοντέλο διωνυμικής κατανομής με πιθανότητες p=0.1, 0.25 και 0.4. Το n κυμαίνεται από 1000 μέχρι 50000.*

3.5 Η Κατανομή του Μέγιστου Τμηματικού Σκορ (Maximal Segment Score)

Στη γενικότερη περίπτωση που ενδιαφερόμαστε για την κατανομή που ακολουθεί η τυχαία μεταβλητή του πλήθους R_n^a διαδοχικών δοκιμών που περιέχουν 100α% ευνοϊκά αποτελέσματα, είναι αναγκαίο να ορίσουμε ένα είδος αθροιστικού σκορ (score), που να το περιγράφει. Η προσέγγιση αυτή, έχει όπως θα δούμε πολλά πλεονεκτήματα, καθώς είναι πολύ γενική αλλά περιλαμβάνει και τη ροή ευνοϊκών αποτελεσμάτων σαν ειδική περίπτωση.

Σύμφωνα με τη μέθοδο αυτή κατά την οποία ενδιαφερόμαστε για την εύρεση μιας περιοχής π.χ. πλούσιας κατά 80% σε A (S. Karlin & Altschul, 1990; Samuel Karlin & Brendel, 1992), πρέπει να ορίσουμε κάποιου είδους σκορ. Τότε, η τυχαία μεταβλητή του πλήθους R_n^a διαδοχικών δοκιμών που περιέχουν 100α% ευνοϊκά αποτελέσματα, μπορεί να περιγράφει με ένα προσθετικό σκορ της μορφής:

$$s_k = \log\left(a_k / p_k\right) \tag{3.19}$$

όπου p_k είναι η πιθανότητα εμφάνισης μιας βάσης σε ολόκληρη την ακολουθία (π.χ. p=1/4) και a_k η πραγματική πιθανότητα εμφάνισης μιας βάσης (target frequency) στο συγκεκριμένο τμήμα της αλληλουχίας

(δηλαδή, σε ένα παράθυρο), το οποίο θέλουμε να ανιχνεύσουμε. Τα *p*,*α* είναι τα ίδια τα οποία συναντήσαμε στην θεωρία μεγάλων αποκλίσεων. Αθροίζοντας τα σκορ για τα *i* κατάλοιπα ενός «παραθύρου» παίρνουμε το τμηματικό σκορ (segment score) και αν το παράθυρο αυτό είναι η μέγιστη περιοχή που περιέχει κατ' ελάχιστο 100α% ευνοϊκά αποτελέσματα, τότε το σκορ ονομάζεται μέγιστο τμηματικό σκορ (maximal segment score), και για μια αλληλουχία μήκους *n* θα συμβολίζεται ως *M*(*n*).

Στον υπολογισμό εργαζόμαστε ως εξής (έστω α=0.8, p=0.25): Από τη σχέση (3.19) έχουμε ότι για κάθε εμφάνιση A, έχουμε συνεισφορά στο σκορ $s_A = \log(0.8/0.25) = 1.163$ και για κάθε εμφάνιση άλλης βάσης θα έχουμε $s_N = \log(0.2/0.25) = -0.223$. Έτσι αν σε ένα τμήμα 20 βάσεων της αλληλουχίας έχουμε 10 A, τότε το σκορ θα είναι s=16*1.163-4*0.223= 17.716.

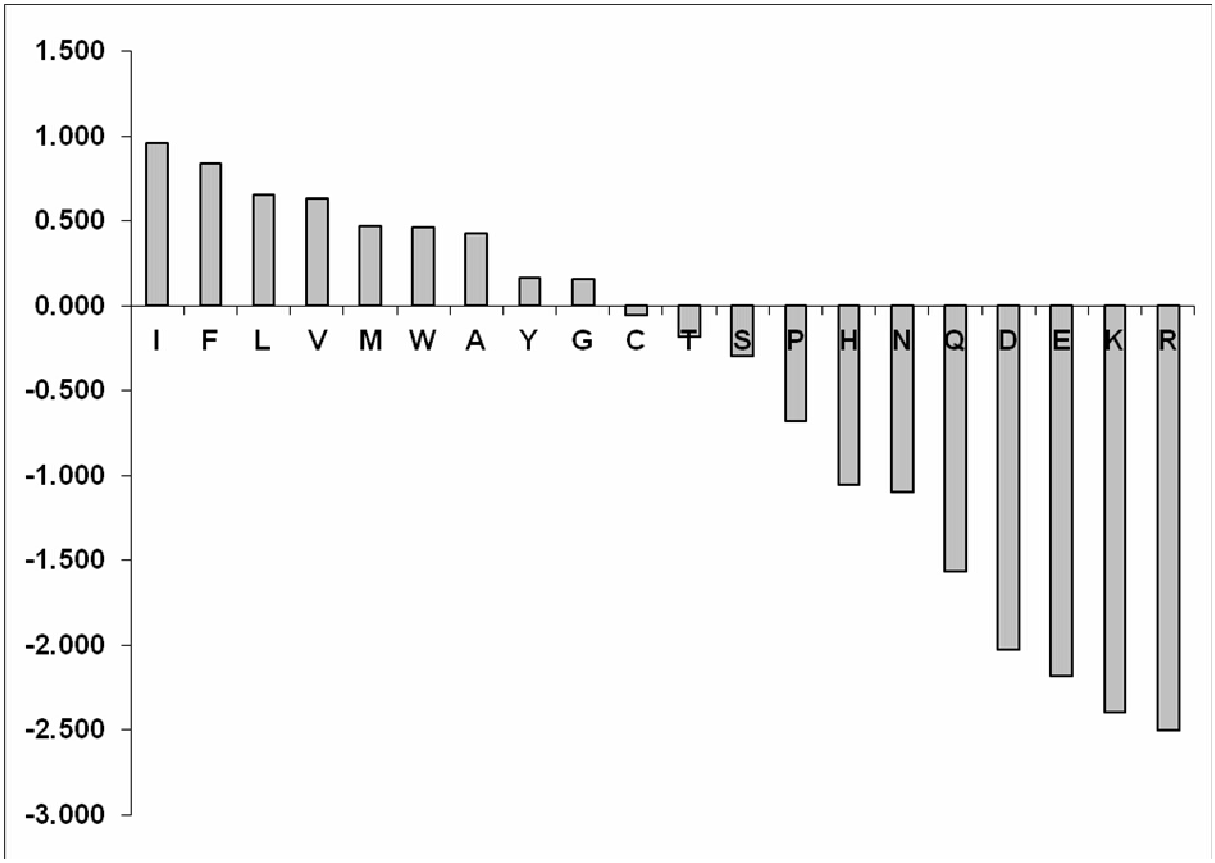
Αμινοξύ (k)	**Πιθανότητα εμφάνισης σε διαμεμβρανικές περιοχές (a_k)**	**Πιθανότητα εμφάνισης σε μη διαμεμβρανικές περιοχές (p_k)**	**Σκορ ($\log(a_k/p_k)$)**
A	0.109	0.071	0.429
C	0.019	0.020	-0.051
D	0.007	0.053	-2.024
E	0.007	0.062	-2.181
F	0.090	0.039	0.836
G	0.082	0.070	0.158
H	0.008	0.023	-1.056
I	0.120	0.046	0.959
K	0.005	0.055	-2.398
L	0.168	0.087	0.658
M	0.040	0.025	0.470
N	0.016	0.048	-1.099
P	0.028	0.055	-0.675
Q	0.009	0.043	-1.564
R	0.005	0.061	-2.501
S	0.053	0.071	-0.292
T	0.050	0.060	-0.182
V	0.115	0.061	0.634
W	0.027	0.017	0.463
Y	0.040	0.034	0.163

Πίνακας 3.1: *Στον πίνακα αυτόν, έχουμε τα στατιστικά στοιχεία από μια ανάλυση 160 διαμεμβρανικών πρωτεϊνών με γνωστή διαμεμβρανική τοπολογία (Krogh, Larsson, von Heijne, & Sonnhammer, 2001). Αν κάποιο αμινοξύ είχε a_k=0 θα έπρεπε να είχαμε κάνει μια μικρή διόρθωση προσθέτοντας μια πολύ μικρή τιμή (πχ 0.0001). Με αυτόν τον τρόπο το σκορ θα έπαιρνε μια πολύ μικρή αρνητική τιμή (πχ -9 ή μικρότερο). Παρατηρήστε ότι η πιθανότητα εμφάνισης των αμινοξέων στις μη-μεμβρανικές περιοχές, είναι πολύ κοντά στη συνολική πιθανότητα εμφάνισης αμινοξέων στη βάση Uniprot (http://web.expasy.org/docs/relnotes/relstat.html).*

Ειδικά στην περίπτωση της ροής R_n^a μιας βάσης, ορίζουμε σκορ =1 (ή κάποιον άλλο θετικό αριθμό) για κάθε εμφάνιση της βάσης αυτής, και -∞ (ή κάποιο άλλο κατ' απόλυτη τιμή πολύ μεγάλο αρνητικό αριθμό) για κάθε άλλη βάση που θα εμφανιστεί. Έτσι μια ροή από π.χ. *k*=10 βάσεις θα δίνει σκορ=10 ενώ κάθε άλλη περίπτωση θα αποκλείεται (σκορ =-∞). Προφανώς, η προσέγγιση αυτή είναι ισοδύναμη με την πρώτη για την ειδική περίπτωση που α=1, καθώς αν θέλουμε να ανιχνεύσουμε την ροή από A, θα πρέπει να δίνουμε σκορ για A, $s_A = \log(1/0.25) = 1.386$, και για κάθε άλλη βάση $s_N = \log(0/0.25) = -\infty$ (πρακτικά, το κάνουμε θέτοντας την αντίστοιχη τιμή ίση με κάποιον πολύ μικρό αρνητικό αριθμό, πχ -10.000). Σε αυτή την περίπτωση το σκορ για μια καθαρή ροή από 16 A, θα είναι 16*1.386=22.176.

Η μέθοδος αυτή, είναι όμως πολύ πιο γενική. Στον Πίνακα 3.1 βλέπουμε τα στατιστικά από μια ανάλυση 160 διαμεμβρανικών πρωτεϊνών με γνωστή διαμεμβρανική τοπολογία (Krogh, et al., 2001). Στις πρωτεΐνες αυτές, έχουμε μελετήσει την αμινοξική σύσταση στα διαμεμβρανικά τμήματα και στις υπόλοιπες (μη μεμβρανικές περιοχές). Με τον τρόπο αυτόν, μπορούμε με τη χρήση της σχέσης (3.19) να κατασκευάσουμε ένα σκορ που θα μπορούμε να το χρησιμοποιήσουμε για τον εντοπισμό περιοχών με μεγάλη πιθανότητα να είναι διαμεμβρανικά τμήματα. Μεγάλες τιμές του σκορ (όπως για παράδειγμα αυτές που έχουν τα αμινοξέα I, F, L, V, M και W), αντιστοιχούν σε αμινοξέα που έχουν μεγαλύτερη πιθανότητα να

εμφανιστούν σε μια διαμεμβρανική περιοχή παρά σε μια μη-μεμβρανική (τα οποία είναι κατά βάση τα υδρόφοβα αμινοξέα). Αντίθετα, τα πολικά αμινοξέα (Q, D, E, K, και R), εμφανίζουν αρνητικές τιμές στο σκορ. Παρόμοια σκορ, είναι δυνατόν να οριστούν με πολλούς άλλους διαφορετικούς τρόπους. Για την ακρίβεια, τέτοιες προσεγγίσεις αποτελούν τη βάση των προγνωστικών αλγορίθμων, τους οποίους θα συναντήσουμε σε επόμενο κεφάλαιο.



Εικόνα 3. 7: *Γραφική παράσταση με τα σκορ των 20 αμινοξέων διαετεταγμένα σε φθίνουσα σειρά μεγέθους. Τα υδρόφοβα αμινοξέα έχουν θετικές τιμές ενώ τα πολικά και τα φορτισμένα, αρνητικές.*

Για τον υπολογισμό του μέγιστου τμηματικού (τοπικού) σκορ πρέπει να θέσουμε και κάποιους περιορισμούς. Συγκεκριμένα:

1. Τουλάχιστον ένα σκορ πρέπει να είναι θετικό
2. Η αναμενόμενη τιμή του σκορ για κάθε βάση να είναι αρνητική, δηλαδή

$$E\left(s_k\right) = \sum p_k s_k = \sum p_k \log\left(\frac{a_k}{p_k}\right) < 0 \tag{3.20}$$

Ο πρώτος περιορισμός είναι απαραίτητος για να είμαστε σίγουροι ότι έχουμε τοπικό σκορ και δεν αναφερόμαστε σε ολόκληρη την ακολουθία, ενώ ο δεύτερος ισχύει σχεδόν πάντα καθώς το $E(s_k)$ είναι ίσο με $-H(\alpha,p)$.

Για την κατανομή που ακολουθεί το μέγιστο τμηματικό score M_n στην γενική περίπτωση, είναι γνωστό το επόμενο θεώρημα (S. Karlin & Altschul, 1990) που λέει ότι η τυχαία μεταβλητή M_n (το μέγιστο τμηματικό score) έχει προσεγγιστική κατανομή την:

$$P\left\{M_n > {\log(n)}/{\lambda} + x\right\} \approx 1 - \exp\left\{-Ke^{-\lambda x}\right\} \tag{3.21}$$

Αυτή είναι η κατανομή των ακραίων τιμών του Gumbel, ενώ K και λ είναι οι σταθερές της και υπολογίζονται με αριθμητικές μεθόδους. Για το λ ειδικά ισχύει το ότι είναι η μοναδική θετική λύση της εξισώσεως:

$$\sum_k p_k \exp\{\lambda s_k\} = 1 \tag{3.22}$$

Οι ίδιοι συγγραφείς, έδειξαν επίσης ότι καθώς το μήκος *n* της τυχαίας ακολουθίας τείνει στο άπειρο, η συχνότητα α_k της εμφάνισης κάποιας βάσης σε ένα τμήμα με αρκετά μεγάλο σκορ προσεγγίζει το $p_k \exp\{\lambda s_k\}$ με πιθανότητα 1. Για την ακρίβεια όταν έχουμε το μέγιστο σκορ, τότε:

$$a_k = p_k \exp\{\lambda s_k\} \tag{3.23}$$

Παρατηρούμε επίσης ότι καθώς η ύπαρξη τέτοιων τμημάτων με μεγάλο σκορ (μεγαλύτερο από *x*) είναι σπάνια γεγονότα (rare events), θα ακολουθούν την κατανομή Poisson με μέση τιμή, E=*Kn*exp(-λ*x*) οπότε η σχέση (3.21) μπορεί να ξαναγραφτεί ως εξής:

$$P(M_n \geq x) \approx 1 - e^{-E} \tag{3.24}$$

Όταν η μέση τιμή-αναμενόμενη τιμή (E-value) είναι πολύ μικρή τότε επειδή ισχύει η προσεγγιστική σχέση:

$$1 - \exp(-\exp(-t)) \approx 1 - (1 - \exp(-t)) = \exp(-t) \tag{3.25}$$

θα έχουμε το P-value περίπου ίσο με το E-value. Επομένως, κάνοντας χρήση της κατανομής Poisson, η πιθανότητα να βρούμε σε μια ακολουθία μήκους *n*, *m* τμήματα με σκορ $S_{(m)}$ μεγαλύτερο ή ίσο από το *x* θα είναι:

$$P(S_{(m)} \geq x) \approx 1 - \exp\left(-Kne^{-\lambda x}\right) \sum_{i=0}^{m-1} \frac{\left(Kne^{-\lambda x}\right)^i}{i!} \tag{3.26}$$

Στην ειδική περίπτωση της ροής R_n, όπως είδαμε παραπάνω, μπορούν να δοθούν κλειστές εκφράσεις για τα *K* και λ και αυτές είναι :

$$K = 1 - p = q \tag{3.27}$$

και

$$\lambda = \log(1/p) \tag{3.28}$$

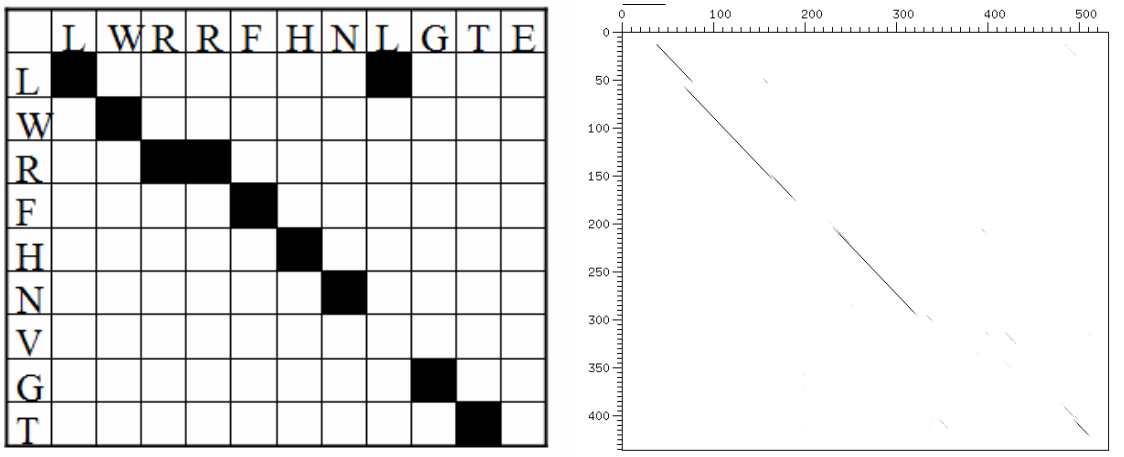
3.6 Στοίχιση αλληλουχιών

Το πρόβλημα της στοίχισης δύο βιολογικών αλληλουχιών, είναι ένα από τα παλιότερα αλλά και πιο σημαντικά θέματα στη βιβλιογραφία της υπολογιστικής βιολογίας. Δύο αλληλουχίες που είναι σε μεγάλο βαθμό «όμοιες», είναι πιθανό να έχουν κοινή εξελικτική προέλευση και, αν μιλάμε για πρωτεΐνες, να έχουν παρόμοια τρισδιάστατη δομή και παρόμοιες λειτουργίες.

Έστω ότι έχουμε δυο βιολογικές αλληλουχίες $\mathbf{x}=x_1,x_2,\ldots,x_n$ και $\mathbf{y}=y_1,y_2,\ldots,y_m$ και θέλουμε να ελέγξουμε κατά πόσο αυτές είναι όμοιες ή όχι. Δημιουργούνται αυτόματα μια σειρά από ερωτήματα:

- Το πρώτο πρόβλημα που προκύπτει είναι με ποιο τρόπο θα μετρήσουμε την ομοιότητα (το πρόβλημα του σκορ)
- Το δεύτερο, αφορά τον τρόπο με τον οποίο θα γίνει η στοίχιση (alignment) των δυο αλληλουχιών (ο αλγόριθμος)
- Το τρίτο, αφορά την επιλογή του είδους της στοίχισης, και τέλος
- Το τελευταίο ερώτημα, αφορά στο πώς θα αποφασίσουμε αν μια δεδομένη στοίχιση είναι σημαντική ή όχι (η στατιστική σημαντικότητα)

Ένας παλιός, αλλά ταυτόχρονα και διαισθητικός τρόπος σύγκρισης δύο αλληλουχιών, είναι το λεγόμενο διάγραμμα σημείων (dot plot). Σύμφωνα με αυτήν την απλοϊκή προσέγγιση, οι δύο αλληλουχίες τοποθετούνται σε δισδιάστατο έναν πίνακα. Σε κάθε κελί του πίνακα, το οποίο αντιστοιχεί σε ένα ζεύγος «γραμμάτων» από τις δύο αλληλουχίες (νουκλεοτίδια ή αμινοξέα), βάζουμε μαύρο χρώμα αν τα δύο σύμβολα είναι όμοια, και λευκό, αν είναι ανόμοια. Διαισθητικά, αναμένουμε ότι αν οι δυο αλληλουχίες είναι 100% όμοιες, το σχήμα που θα παρατηρήσουμε θα είναι μια ευθεία γραμμή στη διαγώνιο. Αν οι οι αλληλουχίες δεν έχουν καμία ομοιότητα, θα περιμένουμε μια τυχαία κατανομή των μαύρων (γραμμοσκιασμένων) κελιών. Προφανώς, σε περιπτώσεις μερικής ομοιότητας, θα περιμένουμε να δούμε «κάτι» που να μοιάζει με γραμμή πάνω ή γύρω από τη διαγώνιο. Αν η ομοιότητα εντοπίζεται μόνο σε ένα ορισμένο σημείο, και δεν εκτείνεται σε όλο το μήκος των αλληλουχιών τότε θα περιμένουμε μια διαγώνιο γραμμή να βρίσκεται κάπου μέσα στον πίνακα (και όχι απαραίτητα στην κύρια διαγώνιο).



Εικόνα 3.8: *Δύο παραδείγματα διαγραμμάτων σημείων (dot plot). Στο αριστερό σχήμα, βλέπουμε το διάγραμμα που αντιστοιχεί στη στοίχιση δύο μικρών πρωτεϊνών, στο οποίο μπορούμε να δούμε τα όμοια και ανόμοια αμινοξέα. Οι δύο αλληλουχίες έχουν μεγάλη ομοιότητα, έστω και αν βλέπουμε 1-2 μικροδιαφορές. Στα δεξιά, βλέπουμε τη σύγκριση δύο πραγματικών πρωτεϊνών μεγάλου μήκους. Στην περίπτωση αυτή δεν μπορούμε να δούμε τα αμινοξέα, αλλά η περιοχή ομοιότητας είναι εμφανής (τουλάχιστον μέχρι το κατάλοιπο 300 των δύο αλληλουχιών)*

Στις επόμενες παραγράφους, θα προσπαθήσουμε να αναλύσουμε περισσότερο τα θέματα αυτά, να τα εξειδικεύσουμε και να τα ποσοτικοποιήσουμε. Θα ξεκινήσουμε λοιπόν, με το πρόβλημα της ποσοτικοποίησης της ομοιότητας. Αν θεωρήσουμε ότι οι δυο αλληλουχίες DNA $\mathbf{x}=x_1,x_2,\ldots,x_n$ και $\mathbf{y}=y_1,y_2,\ldots,y_m$ είναι ασυσχέτιστες (τυχαίες), τότε η πιθανότητα να συμπίπτουν σε κάποιο τμήμα τους είναι:

$$P(\mathbf{x},\mathbf{y} \mid R) = \prod_i q_{x_i} \prod_j q_{y_j} \text{ με } i = 1,2,...,n \text{ και } j = 1,2,...,m \tag{3.29}$$

Εναλλακτικά, αν θεωρήσουμε ότι οι δυο αλληλουχίες είναι συσχετισμένες τότε η πιθανότητα να συμπίπτουν δίνεται από την από κοινού κατανομή της πιθανότητας:

$$P(\mathbf{x},\mathbf{y} \mid M) = \prod_i p_{x_i y_i} \text{ με } i = 1,2,...,n \tag{3.30}$$

όπου για απλότητα (έτσι ώστε να έχει νόημα και η πλήρης ταύτιση-σε όλο το μήκος), μπορούμε να δεχθούμε ότι $n=m$. Ο λόγος των δυο αυτών πιθανοφανειών (likelihood ratio) ονομάζεται και odds ratio και είναι ίσος με:

$$\frac{P(\mathbf{x},\mathbf{y} \mid M)}{P(\mathbf{x},\mathbf{y} \mid R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_j q_{y_j}} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \tag{3.31}$$

Αν πάρουμε τους λογαρίθμους έχουμε:

$$S = \sum_i \log\left(\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}\right) = \sum_i s(x_i, y_i) \tag{3.32}$$

οπότε ορίζουμε με αυτόν τον τρόπο, το score για την ομοιότητα δυο αλληλουχιών το οποίο πλέον έχει προσθετικές ιδιότητες. Για τα 4 νουκλεοτίδια του DNA μπορούν να φτιαχτούν πίνακες 4x4 που να απεικονίζουν τις παραπάνω συνεισφορές στο score για κάθε μια από τις 16 περιπτώσεις ταύτισης βάσεων σε μια στοίχιση αλλά πολλές φορές μπορούμε απλώς να θέσουμε:

$$s(x_i, y_i) = \begin{cases} 1, & \text{αν } x_i = y_i \\ -1, & \text{αν } x_i \neq y_i \end{cases} \tag{3.33}$$

δηλαδή ορίζουμε συνεισφορά στο score 1 για ταύτιση (match) και –1 για διαφορά (mismatch). Ο πίνακας αυτός ονομάζεται πίνακας ταύτισης (match matrix) ενώ αν έχουμε ποινή για την εμφάνιση διαφοράς $\rightarrow$-∞ (π.χ. –10000) αποκλείεται να εμφανιστεί μια στοίχιση που να περιέχει τίποτα άλλο εκτός από απολύτως όμοια νουκλεοτίδια, και ο πίνακας αυτός ονομάζεται ταυτοτικός πίνακας (identity matrix). Οι ονομασίες για τους παραπάνω πίνακες (π.χ. ταυτοτικός) δεν πρέπει να συγχέονται με την αντίστοιχη ονομασία των πινάκων στη

γραμμική άλγεβρα. Η ερμηνεία του ταυτοτικού πίνακα είναι ότι με μια τόσο μεγάλη ποινή για εμφάνιση διαφοράς (→-∞) αποκλείεται να εμφανιστεί μια στοίχιση που να περιέχει τίποτα άλλο εκτός από απολύτως όμοια νουκλεοτίδια.

Αυτό το απλοϊκό σχήμα του περιγράψαμε, θυμίζει αρκετά τη μέθοδο των διαγραμμάτων σημείων. Ειδικά για τις πρωτεΐνες υπάρχουν πάρα πολλοί τέτοιοι πίνακες υποκατάστασης (substitution matrices) οι οποίοι υπολογίζουν τις αντίστοιχες συνεισφορές στο score για μη ταύτιση των διάφορων αμινοξέων (mismatches) στηριζόμενοι σε παρατηρηθείσες αντικαταστάσεις αμινοξέων αλλά με διαφορετικό τρόπο ο καθένας. Οι πίνακες υποκατάστασης όπως έδειξε ο Altschul (Altschul, 1991), έχουν ξεκάθαρη ερμηνεία υπό το πρίσμα της θεωρίας της πληροφορίας. Τέτοιοι πίνακες είναι οι πίνακες των οικογενειών PAM (Dayhoff, Schwartz, & Orcutt, 1978), BLOSUM (Henikoff & Henikoff, 1992), GONNET (Gonnet, Cohen, & Benner, 1992) οι οποίοι επιπλέον, έχουν μια ξεκάθαρη εξελικτική ερμηνεία όπως θα δούμε παρακάτω.

Τέλος, είναι απαραίτητο να προβλέψουμε και την ύπαρξη κενών στις στοιχισμένες αλληλουχίες. Η ύπαρξη αυτή είναι απαραίτητη καθώς θα δούμε παρακάτω ότι ένα από τα βασικά χαρακτηριστικά των μεταλλάξεων μέσω των οποίων προχωράει η εξέλιξη είναι η προσθήκη (insertion) και η απαλοιφή (deletion) νουκλεοτιδίων. Όταν στη στοίχιση δυο αλληλουχιών εμφανίζεται (στη μια από τις δυο), το κενό (gap) δεν είναι δυνατό να ξέρουμε αν αυτό προήλθε (εξελικτικά) από απαλοιφή βάσης σ' αυτή την αλληλουχία, ή από προσθήκη στην άλλη αλληλουχία με την οποία συγκρίνεται. Προφανώς η ύπαρξη του κενού πρέπει να «τιμωρείται» από το score γιατί αλλιώς δυο οποιεσδήποτε αλληλουχίες με την προσθήκη «κατάλληλου» αριθμού κενών θα δίνουν μια άριστη στοίχιση. Η συνεισφορά των κενών (η οποια πρέπει να είναι αρνητική) στο score ορίζεται από μια συνάρτηση $\gamma(g)$, όπου g είναι ο αριθμός των κενών, και μπορεί να είναι είτε γραμμική:

$$\gamma(g) = -gd \tag{3.34}$$

είτε πιο σύνθετη:

$$\gamma(g) = -d - (g-1)e \tag{3.35}$$

όπου d είναι η ποινή για την ύπαρξη κενού (gap open penalty) και e η ποινή για την διεύρυνση του κενού (gap extension penalty). Η σωστή επιλογή της συνάρτησης για τα κενά είναι δύσκολη διαδικασία και υπάρχει πλούσια βιβλιογραφία για αυτό το θέμα (Vingron & Waterman, 1994).

3.7 Πίνακες ομοιότητας

Σε περιπτώσεις στοίχισης αλληλουχιών DNA, χρησιμοποιούνται συνήθως απλοί πίνακες ομοιότητας της μορφής της σχέσης (3.33). Στις περισσότερες περιπτώσεις, αν δεν θέλουμε να επιτρέψουμε πολλές ταυτίσεις ανόμοιων βάσεων χρησιμοποιούμε την τιμή 1 για τη ταύτιση και -3 για τη διαφορά, ενώ στις πιο συνηθισμένες περιπτώσεις, χρησιμοποιούμε μια τιμή 5 για την ταύτιση και -4 για τη διαφορά.

Από την άλλη πλευρά, ειδικά για τις πρωτεΐνες υπάρχουν πάρα πολλοί εξειδικευμένοι πίνακες υποκατάστασης (substitution matrices) οι οποίοι υπολογίζουν τις αντίστοιχες συνεισφορές στο score για μη ταύτιση των διάφορων αμινοξέων (mismatches) στηριζόμενοι σε παρατηρηθείσες αντικαταστάσεις αμινοξέων αλλά με διαφορετικό τρόπο ο καθένας. Οι πίνακες υποκατάστασης όπως έδειξε ο Altschul (Altschul, 1991), έχουν ξεκάθαρη ερμηνεία υπό το πρίσμα της θεωρίας της πληροφορίας. Τέτοιοι πίνακες είναι οι πίνακες των οικογενειών PAM (Dayhoff, et al., 1978), BLOSUM (Henikoff & Henikoff, 1992), GONNET (Gonnet, et al., 1992) αλλά και άλλοι.

Οι πίνακες BLOSUM (Henikoff & Henikoff, 1992), έχουν υπολογισθεί από τμήματα από πολλαπλές στοιχίσεις αλληλουχιών για τις οποίες υπάρχουν ξεκάθαρες ενδείξεις ότι έχουν φυλογενετική σχέση. Τα τμήματα αυτά (blocks), επιλέχθηκαν με προσοχή από ένα μεγάλο εύρος πρωτεϊνικών οικογενειών και διατηρήθηκαν τελικά μόνο τα πιο καλά στοιχισμένα τμήματα (αυτά που δεν περιείχαν κενά). Για τον υπολογισμό χρησιμοποιήθηκε ο τύπος:

$$s_{ij} = \frac{1}{\lambda}\log\left(\frac{q_{ij}}{p_i p_j}\right) \tag{3.36}$$

όπου q_{ij}, είναι η πιθανότητα αντικατάστασης του i από το j σε σχετιζόμενες πρωτεΐνες (target frequencies), p_i, p_j είναι οι πιθανότητες εμφάνισης των αμινοξέων σε οποιαδήποτε θέση (background frequencies) και λ είναι μια σταθερά κανονικοποίησης έτσι ώστε οι τιμές να μετατραπούν σε ακέραιους. Η ομοιότητα με τη σχέση

(3.32) είναι εμφανής. Οι πίνακες αυτοί δεν προϋποθέτουν ένα εξελικτικό μοντέλο αλλά το προσεγγίζουν εμπειρικά.

Η άλλη μεγάλη οικογένεια πινάκων αντικατάστασης, είναι οι πίνακες της οικογένειας Point Accepted Mutations (PAM) (Dayhoff, et al., 1978). Οι συγγραφείς, όρισαν ως «Αποδεκτή Σημειακή Μεταλλαγή» (PAM) σε μια πρωτεΐνη την αντικατάσταση ενός αμινοξικού κατάλοιπού της με ένα κατάλοιπο διαφορετικού τύπου, η οποία έχει γίνει αποδεκτή μέσω της διαδικασίας της Φυσικής Επιλογής. Οι τιμή PAM1 προέκυψε από πολλαπλή στοίχιση αλληλουχιών με γνωστή εξελικτική σχέση και ομοιότητα μεγαλύτερης του 85%, και μέσω αυτής της τιμής με χρήση ενός μαρκοβιανού μοντέλου εξέλιξης (θα παρουσιαστούν στο κεφάλαιο της φυλογενετικής ανάλυσης), προέκυψαν οι πίνακες PAM30, PAM250 κ.ο.κ. καθώς οι πίνακες αυτοί είναι δηλαδή πολλαπλασιαστικοί, καθώς PAMN=$(PAM1)^N$. Η χρήση πινάκων με μικρό N ενδείκνυται όταν οι εξεταζόμενες αλληλουχίες είναι πολύ όμοιες (μικρή εξελικτική απόσταση), ενώ στην περίπτωση περισσότερο απομακρυσμένων ομοιοτήτων χρησιμοποιούμε πίνακες μεγαλύτερου N. Στις περιπτώσεις εκείνες κατά τις οποίες δεν γνωρίζουμε εκ των προτέρων την ομοιότητα των προς σύγκριση αλληλουχιών (π.χ. σε αναζητήσεις έναντι βάσεων δεδομένων) επιλέγουμε έναν ενδιάμεσο πίνακα, όπως τον PAM250, ο οποίος αντιστοιχεί σε συντήρηση της τάξης του 20-25%.

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

Εικόνα 3.9: *Ο Πίνακας BLOSUM62. Παρατηρήστε ότι αμινοξέα τα οποία έχουν παρόμοιες φυσικοχημικές ιδιότητες (π.χ. υδρόφοβα, πολικά, αρωματικά κ.ο.κ.), έχουν γενικά θετικές τιμές για τις μεταξύ τους αντικαταστάσεις*

Παρόλο που οι δύο οικογένειες πινάκων έχουν διαφορές, μπορούμε σε γενικές γραμμές να κάνουμε μια «αντιστοίχιση». Γενικά, μικρές τιμές των πινάκων PAM, και μεγάλες τιμές των πινάκων BLOSUM αντιστοιχούν σε, και κατά συνέπεια πρέπει να χρησιμοποιούνται για, αλληλουχίες με μικρή εξελικτική απόσταση, δηλαδή με μεγάλες ομοιότητες. Αντίθετα, μεγάλες τιμές των πινάκων PAM, και μικρές τιμές των πινάκων BLOSUM αντιστοιχούν σε, και κατά συνέπεια πρέπει να χρησιμοποιούνται για, αλληλουχίες με μεγάλη εξελικτική απόσταση, δηλαδή με μικρότερες ομοιότητες (Πίνακας 3.1).

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

Πίνακας 3.2 *Πίνακας κατά προσέγγιση αντιστοίχιση των πινάκων της οικογένειας PAM με αυτούς της οικογένειας BLOSUM*

Τα παραπάνω προφανώς, ισχύουν σε ειδικές περιπτώσεις, όταν ξέρουμε εκ των προτέρων πόσο αναμένουμε να μοιάζουν δύο υπό σύγκριση αλληλουχίες. Στη γενική περίπτωση όμως που πραγματοποιούμε

αναζήτηση σε μια βάση δεδομένων, τότε η πιο συνετή επιλογή είναι να χρησιμοποιήσουμε έναν πίνακα «ενδιάμεσης» ομοιότητας, όπως τον BLOSUM62. Αν το αποτέλεσμα της αναζήτησης μας δώσει πάρα πολλές πρωτεΐνες τις οποίες δυσκολευόμαστε να διαχωρίσουμε (έχουν όλες μεγάλη ομοιότητα), τότε μπορούμε να επαναλάβουμε την αναζήτηση με έναν πίνακα όπως ο BLOSUM90. Αν, από την άλλη μεριά, η αρχική αναζήτηση μας δώσει λίγα αποτελέσματα, τότε θα πρέπει να ξανακάνουμε αναζήτηση επιλέγοντας για παράδειγμα τον BLOSUM45. Αυτό που πρέπει σε κάθε περίπτωση να θυμάται ο αναγνώστης, είναι ότι αλλαγή στον πίνακα υποκατάστασης, σημαίνει και αλλαγή (μικρή ή μεγάλη) στα αποτελέσματα της αναζήτησης αλλά και της προκύπτουσας στοίχισης.

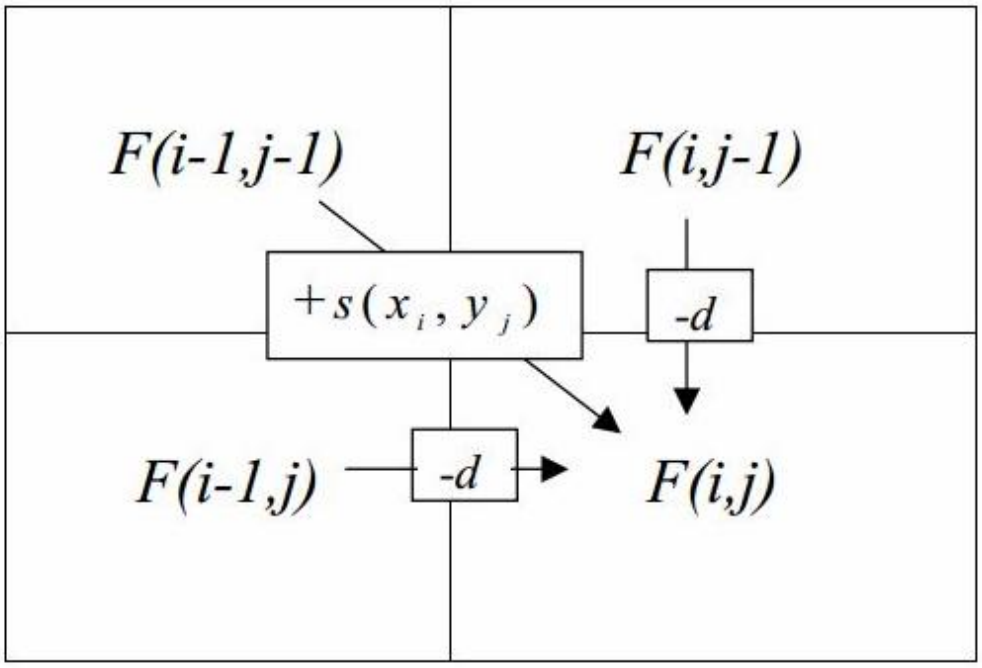
Διαισθητικά, οι πίνακες αυτοί, έχουν την εξής ερμηνεία: αμινοξέα τα οποία έχουν παρόμοιες φυσικοχημικές ιδιότητες (πχ υδρόφοβα, πολικά, αρωματικά κ.ο.κ.), έχουν θετικές τιμές για τις μεταξύ τους αντικαταστάσεις. Αυτό σημαίνει ότι σε γενικές γραμμές, μια αντικατάσταση ενός αμινοξέος με ένα άλλο παρόμοιο, θα είναι «αποδεκτή» διαδικασία για τη δομή και τη λειτουργία της πρωτεΐνης. Αυτό με τη σειρά του, σημαίνει ότι είναι δυνατόν δύο πρωτεΐνες στις οποίες μεγάλο μέρος των αμινοξέων έχουν αντικατασταθεί με «παρόμοια» (και κατά συνέπεια, δεν εμφανίζουν μεγάλη ονομαστική ταύτιση), παρόλα αυτά να θεωρούνται «όμοιες» και να λαμβάνουν μεγάλο score στις στοιχίσεις. Φυσικά, αναμένουμε ότι για κάθε αμινοξύ, τη μεγαλύτερη τιμή για αντικατάσταση θα την έχει ο εαυτός του (οι τιμές στη διαγώνιο) αλλά δεν αναμένουμε όλες οι τιμές της διαγωνίου να είναι ίδιες γιατί οι τιμές αυτές εξαρτώνται και από την πιθανότητα εμφάνισης του κάθε αμινοξέος. Για παράδειγμα στον BLOSUM62, η Κυστεΐνη (C) και η Τρυπτοφάνη (W), οι οποίες είναι τα πιο σπάνια αμινοξέα, έχουν και τις μεγαλύτερες τιμές στη διαγώνιο (9 και 11, αντίστοιχα), ενώ η Αλανίνη (A), η οποία είναι ένα από τα πιο συνηθισμένα, έχει τη μικρότερη τιμή (μόλις 4). Τέλος, πρέπει να τονίσουμε, ότι οι πίνακες που περιγράψαμε είναι φτιαγμένοι για γενική χρήση. Για πιο ειδικά προβλήματα, είναι δυνατόν να κατασκευαστούν ειδικοί πίνακες, όπως για παράδειγμα στην περίπτωση της αναζήτησης για διαμεμβρανικές πρωτεΐνες, ο πίνακας PHAT (Ng, Henikoff, & Henikoff, 2000) και ο SLIM (Muller, Rahmann, & Rehmsmeier, 2001). Ο τελευταίος μάλιστα, είναι και μη-συμμετρικός, ιδιότητα που του επιτρέπει να υπολογίζει καλύτερα την ασυμμετρία που υπάρχει στις κατανομές των αμινοξέων στις μεμβρανικές πρωτεΐνες (σε σχέση με τις γενικές πιθανότητες «υποβάθρου»).

3.8 Αλγόριθμοι δυναμικού προγραμματισμού

Αφού ορίσαμε τον τρόπο ποσοτικοποίησης της ομοιότητας των αλληλουχιών, το επόμενο βήμα είναι να βρούμε τον καλύτερο τρόπο στοίχισης. Για άλλη μια φορά, η αναφορά στο διάγραμμα σημείων είναι χρήσιμη, καθώς είπαμε πριν ότι σε περιπτώσεις μερικής ομοιότητας, θα περιμένουμε να δούμε «κάτι» που να μοιάζει με γραμμή πάνω στη διαγώνιο, ή γύρω από αυτή. Ο σκοπός μιας στοίχισης, είναι να εντοπίσει τη βέλτιστη διαδρομή πάνω σε έναν τέτοιον πίνακα. Όταν η διαδρομή είναι γνωστή, η παράθεση των ζευγών των συμβόλων που αντιστοιχούν στα κελιά του πίνακα με την «καλύτερη» διαδρομή, αντιστοιχεί στην τελική στοίχιση. Προφανώς, «καλύτερη διαδρομή» σημαίνει η διαδρομή η οποία μεγιστοποιεί το σκορ όπως το ορίσαμε λίγο πριν. Οι πιθανές στοιχίσεις όμως, δηλαδή οι πιθανές διαδρομές στον πίνακα, είναι πάρα πολλές. Οι πιθανοί τρόποι παράθεσης των δυο αλληλουχιών η μια κάτω από την άλλη, αν υποθέσουμε ότι μπορεί να υπάρχουν και οσαδήποτε κενά, είναι $\binom{n+m}{n}$ και στην ειδική περίπτωση που $n=m$, έχουμε από τον τύπο του Stirling (Durbin, et al., 1998):

$$\binom{2n}{n} = \frac{(2n!)}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{2\pi n}} \tag{3.37}$$

Προφανώς ο αριθμός αυτός είναι πολύ μεγάλος και δεν υπάρχει τρόπος να υπολογισθούν όλα τα σκορ που αντιστοιχούν σε αυτούς τους συνδυασμούς. Αντίθετα χρησιμοποιούνται αλγόριθμοι δυναμικού προγραμματισμού που με διαδοχικά βήματα βρίσκουν τον καλύτερο τρόπο στοίχισης. Ο δυναμικός προγραμματισμός, είναι μια τεχνική που βρίσκει εφαρμογές σε πολλά δύσκολα προβλήματα στη βιοπληροφορική. Το βασικό χαρακτηριστικό των αλγορίθμων αυτών, είναι ότι «σπάνε» το μεγάλο πρόβλημα (το οποίο απαιτεί πολλούς υπολογισμούς για να λυθεί), σε μικρότερα προβλήματα τα οποία λύνονται πιο εύκολα. Το βασικό σημείο, είναι κάθε φορά, μια επαγωγική απόδειξη η οποία θα δείχνει ότι το άθροισμα των μικρότερων αυτών προβλημάτων, δίνει και τη λύση του μεγάλου προβλήματος.



Εικόνα 3.10: *Οι αλγόριθμοι δυναμικού προγραμματισμού, υπολογίζουν κάθε φορά το στοιχείο F(i,j) από τα 3 γειτονικά κελιά του F(i-1,j) F(i,j-1) F(i-1,j-1).*

Οι αλγόριθμοι δυναμικού προγραμματισμού στη στοίχιση αλληλουχιών (Gonnet, et al., 1992) εργάζονται σε γενικές γραμμές ως εξής: τοποθετούν τις δυο αλληλουχίες $\mathbf{x}=x_1,x_2,\ldots,x_n$ και $\mathbf{y}=y_1,y_2,\ldots,y_m$ σε ένα *nm* πίνακα με στοιχεία $F(i,j)$ όπου κάθε στοιχείο αυτού του πίνακα είναι η τιμή του σκορ για την καλύτερη στοίχιση μέχρι το στοιχείο x_i, και το y_j. Στην ουσία, δουλεύουν πάνω στον πίνακα του διαγράμματος σημείων που είδαμε πριν, τοποθετώντας αριθμητικές τιμές στα κελιά του.

Προφανώς, αν γνωρίζουμε την τιμή της συνεισφοράς στο σκορ $s(x_i,y_i)$ για κάθε δυνατό συνδυασμό βάσεων και τη συνάρτηση της ποινής για το κενό τότε με γνωστά τα στοιχεία $F(i\text{-}1,j)$, $F(i,j\text{-}1)$ και $F(i\text{-}1,j\text{-}1)$ μπορούμε να υπολογίσουμε αναδρομικά το $F(i,j)$ και όπως φαίνεται στην Εικόνα 3.10 όπου απεικονίζονται οι τρεις οι πιθανοί τρόποι μετάβασης από το $F(i\text{-}1,j\text{-}1)$ στο $F(i,j)$. Προφανώς κάθε μη διαγώνια μετάβαση σημαίνει την εισαγωγή του κενού σε μια από τις δυο αλληλουχίες. Έχοντας υπολογίσει όλα τα στοιχεία αυτού του πίνακα μπορούμε κινούμενοι προς τα πίσω να βρούμε την καλύτερη δυνατή στοίχιση των δυο αλληλουχιών.

3.9 Ολική στοίχιση - Ο αλγόριθμος των Needleman και Wunsch

Η πρώτη περίπτωση η οποία θα εξετάσουμε είναι η λεγόμενη ολική στοίχιση δυο αλληλουχιών (*global alignment*). Στην περίπτωση αυτή έχουμε δυο αλληλουχίες περίπου ίδιου μήκους και θέλουμε να δούμε ποιος είναι ο καλύτερος δυνατός τρόπος να στοιχηθούν παράλληλα η μια κάτω από την άλλη σε όλο το μήκος τους (π.χ. μπορεί να είναι δυο γονίδια για την ίδια πρωτεΐνη από διαφορετικούς οργανισμούς) ώστε να μπορέσουμε να εξετάσουμε την πιθανή εξελικτική ή λειτουργική σχέση τους.

Ο αλγόριθμος που επιτυγχάνει τα παραπάνω είναι ο αλγόριθμος των *Needleman-Wunsch* (Needleman & Wunsch, 1970). Σύμφωνα με τον αλγόριθμο αυτό το σκορ κάθε κελιού υπολογίζεται με τον αναδρομικό τύπο:

$$F(i,j)=\max\left\{F(i-1,j-1)+s(x_i,y_j),F(i-1,j)-d,F(i,j-1)-d\right\} \tag{3.38}$$

Η τιμή για το κάτω δεξιά στοιχείο του πίνακα είναι εξ' ορισμού το σκορ για την καλύτερη δυνατή στοίχιση, ενώ για την αρχικοποίηση της πρώτης στήλης και της πρώτης γραμμής, έχουμε $F(i,0)=-id$ και $F(0,j)=-jd$. Από το κάτω δεξιά στοιχείο, θα πρέπει να ξεκινήσει μια αναδρομή (recursion) στον πίνακα, η οποία ακολουθώντας κάθε φορά τα μέγιστα θα αποκαλύψει τη βέλτιστη διαδρομή, δηλαδή τη βέλτιστη στοίχιση.

Παράδειγμα 3.10.1 (M. S. Waterman, 1995)
Έστω ότι έχουμε τις εξής δυο αλληλουχίες DNA **y**=CAGTATCGCA και **x**=AAGTTAGCAG. Θέλουμε να δούμε ποια είναι η καλύτερη ολική στοίχιση που μπορούν να έχουν με:

$$s(x_i, y_i) = \begin{cases} 1, & \text{αν } x_i = y_i \\ -1, & \text{αν } x_i \neq y_i \end{cases} \tag{3.39}$$

και d=1, τότε συμπληρώνοντας τον πίνακα έχουμε:

	-	*A*	*A*	*G*	*T*	*T*	*A*	*G*	*C*	*A*	*G*
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
C	-1	-1	-2	-3	-4	-5	-6	-7	-6	-7	-8
A	-2	0	0	-1	-2	-3	-4	-5	-6	-5	-6
G	-3	-1	-1	1	0	-1	-2	-3	-4	-5	-4
T	-4	-2	-2	0	2	1	0	-1	-2	-3	-4
A	-5	-3	-1	-1	1	1	2	1	0	-1	-2
T	-6	-4	-2	-2	0	2	1	1	0	-1	-2
C	-7	-5	-3	-3	-1	1	1	0	2	1	0
G	-8	-6	-4	-2	-2	0	0	2	1	1	2
C	-9	-7	-5	-3	-3	-1	-1	1	3	2	1
A	-10	-8	-6	-4	-4	-2	0	0	2	4	3

οπότε η ολική στοίχιση είναι:

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

η οποία έχει σκορ ίσο με 3 (η τιμή του κελιού κάτω δεξιά στον πίνακα).

3.10 Προσαρμογή αλληλουχιών

Μια άλλη περίπτωση έχουμε όταν θέλουμε να δούμε την προσαρμογή (fit) μιας μικρής αλληλουχίας σε μια μεγαλύτερη, δηλαδή όταν θέλουμε να ανιχνεύσουμε αν μια μικρή αλληλουχία με βιολογική σημασία υπάρχει σε μια μεγαλύτερη. Ο αλγόριθμος αυτός χρησιμοποιεί τη σχέση (3.38) με κάποιες διαφοροποιήσεις όμως. Πιο συγκεκριμένα (Galas, Eggert, & Waterman, 1985):

$$\begin{aligned} &F(i,j) = \max\left\{F(i-1,j-1) + s(x_i, y_j), F(i-1,j) - d, F(i,j-1) - d\right\} \\ &\text{με } F(i,0) = -id \text{ και } F(0,j) = 0 \end{aligned} \tag{3.40}$$

Παράδειγμα 3.11.1 (M. S. Waterman, 1995)
Έστω ότι θέλουμε να ανιχνεύσουμε αν στην αλληλουχία του γονιδίου lacI της E.coli υπάρχει η γνωστή αλληλουχία του υποκινητή (promoter). Έστω ακόμα ότι το τμήμα του γονιδίου έχει αλληλουχία:

x=TCGCGGTATGGCATGATAGCGCCCGGAA,

και η αλληλουχία του υποκινητή είναι:

y=TATAAT

Αν θέσουμε επίσης $s(x_i, y_i) = \begin{cases} 1, & \text{αν } x_i = y_i \\ -1, & \text{αν } x_i \neq y_i \end{cases}$ και d=2, τότε ο πίνακας F παίρνει τη μορφή:

	T	C	G	C	G	G	T	A	T	G	G	C	A	T	G	A	T	A	G	C	G	C	C	C	G	G	A	A
T	1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
A	0	0	-2	-2	-2	-2	-1	2	0	0	-2	-2	0	-1	0	0	-1	2	0	-2	-2	-2	-2	-2	-2	-2	0	0
T	1	-1	-1	-3	-3	-3	-1	0	3	1	-1	-3	-2	1	-1	-1	1	0	1	-1	-3	-3	-3	-3	-3	-3	-2	-1
A	-1	0	-2	-2	-4	-4	-3	0	1	2	0	-2	-2	-1	0	0	-1	2	0	0	-2	-4	-4	-4	-4	-4	-2	-1
A	-3	-2	-1	-3	-3	-5	-5	-2	-1	0	1	-1	-1	-3	-2	1	-1	0	1	-1	-1	-3	-5	-5	-5	-5	-3	-1
T	-3	-4	-3	-2	-4	-4	-4	-4	-1	-2	-1	0	-2	0	-2	-1	2	0	-1	0	-2	-2	-4	-6	-6	-6	-5	-3

Παρατηρούμε ότι ο αλγόριθμος εντόπισε μια αλληλουχία πιθανού υποκινητή

C A T G A T

η οποία έχει σκορ ίσο με 2 (επειδή το 2 είναι το μέγιστο στοιχείο στην τελευταία σειρά του πίνακα, και με αναδρομή στη γραμμοσκιασμένη περιοχή βρίσκουμε την παραπάνω αλληλουχία).

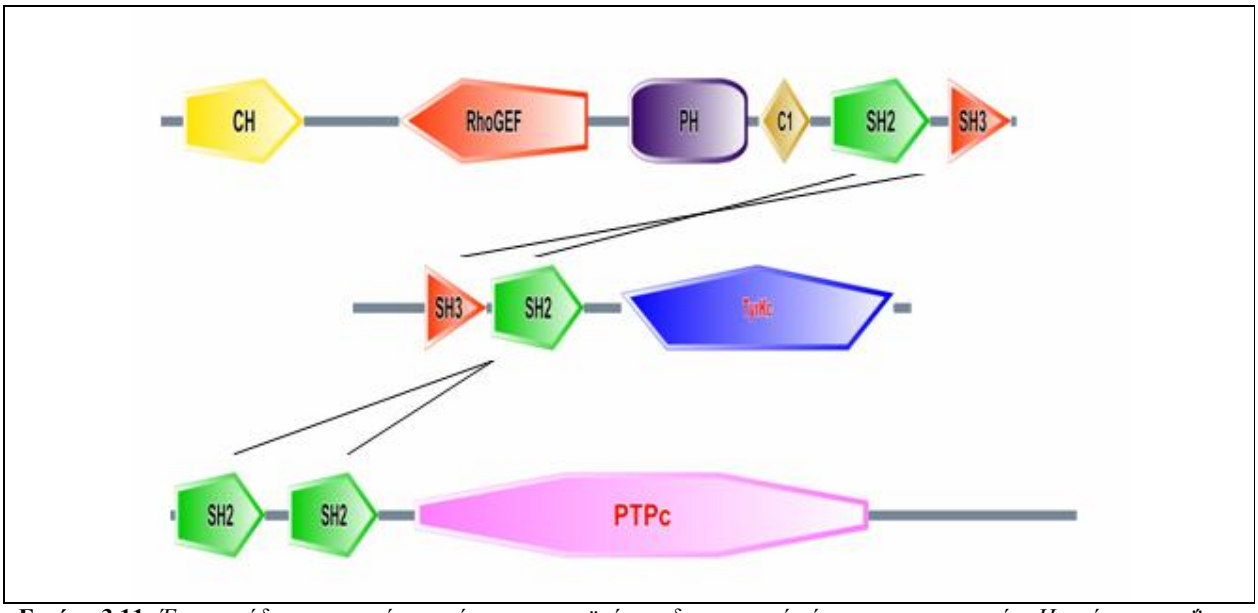
3.11 Τοπική στοίχιση – ο αλγόριθμος Smith και Waterman

Τέλος, μια τρίτη περίπτωση, η οποία όμως παρουσιάζει ιδιαίτερο ενδιαφέρον είναι αυτή που χρησιμοποιείται για τη σύγκριση δυο αλληλουχιών στην περίπτωση που θέλουμε να βρούμε την καλύτερη δυνατή στοίχιση δυο υπό-ακολουθιών τους. Η μέθοδος αυτή ονομάζεται τοπική στοίχιση (local alignment) και δίνει πολλές φορές συνταρακτικά αποτελέσματα ακόμα και σε αλληλουχίες που δεν έχουν καθόλου εμφανή ολική ομοιότητα (ομολογία). Η μέθοδος αυτή είναι η ευρύτερα χρησιμοποιούμενη καθώς μας επιτρέπει και από εξελικτική σκοπιά να διαχωρίζουμε τις αλληλουχίες σε περιοχές που βρίσκονται κάτω από ισχυρή εξελικτική πίεση (και άρα μεταλλάσσονται πολύ αργά) και σε άλλες που μπορεί να διαφέρουν πάρα πολύ (W. R. Pearson & Wood, 2001). Όπως επίσης έχουμε αναφέρει, η μέθοδος έχει μεγάλη σημασία στη σύγκριση πρωτεϊνικών αλληλουχιών, καθώς οι πρωτεΐνες αποτελούνται από διαφορετικούς συνδυασμούς περιοχών (domains), και κατά συνέπεια μας ενδιαφέρει πολλές φορές να μπορούμε να εντοπίσουμε τέτοιου είδους ομοιότητες.

Ο αλγόριθμος που επιτυγχάνει τα παραπάνω είναι ο αλγόριθμος των Smith – Waterman (Smith & Waterman, 1981) και χρησιμοποιεί τον εξής αναδρομικό τύπο:

$$F(i,j) = \max\left\{F(i-1,j-1) + s(x_i, y_j), F(i-1,j) - d, F(i,j-1) - d, 0\right\}$$
$$\text{με } F(i,0) = 0 \text{ και } F(0,j) = 0 \tag{3.41}$$

Παρατηρούμε ότι ο αλγόριθμος είναι ίδιος με αυτόν για την ολική στοίχιση με τη διαφορά ότι όποτε μια στοίχιση δίνει αρνητικό *σκορ* αυτή τερματίζεται και αρχίζει μια νέα. Επίσης, και αυτό είναι πολύ σημαντικό, η αρχικοποίηση του πίνακα είναι διαφορετική για να μπορεί να εντοπίσει ομοιότητες σε οποιοδήποτε σημείο εκτός της κύριας διαγωνίου.



Εικόνα 3.11: *Ένα παράδειγμα τοπικής ομοιότητας πρωτεϊνών με διαφορετική σύσταση των περιοχών. Η πρώτη πρωτεΐνη έχει δύο περιοχές που μοιάζουν με περιοχές της δεύτερης πρωτεΐνης (αλλά δεν βρίσκονται στην ίδια θέση στην αλληλουχία). Αντίθετα, η Τρίτη πρωτεΐνη διαθέτει μόνο μία από τις περιοχές αυτές, αλλά σε δύο αντίγραφα.*

Παράδειγμα 3.12.1 (M. S. Waterman, 1995)

Αν εφαρμόσουμε τη μέθοδο αυτή στις αλληλουχίες του παραδείγματος 3.11.1 έχουμε $s(x_i, y_i) = \begin{cases} 1, & \text{αν } x_i = y_i \\ -1, & \text{αν } x_i \neq y_i \end{cases}$ και d=1 και ο πίνακας F παίρνει τη μορφή:

	-	A	A	G	T	T	A	G	C	A	G
-	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	1	0	0
A	0	1	1	0	0	0	1	0	0	2	1
G	0	0	0	2	1	0	0	2	1	1	3
T	0	0	0	1	3	2	1	1	1	0	2
A	0	1	1	0	2	2	3	2	1	2	1
T	0	0	0	0	1	3	2	2	1	1	1
C	0	0	0	0	0	2	2	1	3	2	1
G	0	0	0	1	0	1	1	3	2	2	3
C	0	0	0	0	0	0	0	2	4	3	2
A	0	1	1	0	0	0	1	1	3	5	4

Επομένως, η καλύτερη τοπική στοίχιση είναι:

A G T A T C G C A
A G T – T A G C A

με σκορ ίσο με 5 (το κελί με τη μεγαλύτερη τιμή στον πίνακα). Πρέπει εδώ να τονίσουμε ότι ο απαιτούμενος χρόνος για να εκτελεστούν οι παραπάνω αλγόριθμοι δυναμικού προγραμματισμού είναι ανάλογος του γινόμενου των μηκών των ακολουθιών και συμβολίζεται $O(mn)$. Το σύμβολο O(mn) (*big-O notation*) δηλώνει ότι ο αριθμός των υπολογισμών που απαιτούνται για να ολοκληρωθεί ο αλγόριθμος, είναι ανάλογος του nm, δηλαδή του πλήθους των κελιών του πίνακα. Κατά συνέπεια, λέμε ότι ο αλγόριθμος είναι γραμμικός ως προς το μήκος των αλληλουχιών. Σε αντιδιαστολή, ο απλοϊκός αλγόριθμος απαρίθμησης όλων των πιθανών στοιχίσεων, είναι εκθετικός ως προς το μήκος των αλληλουχιών (είναι ανάλογος του 2^n).

Πρέπει να τονιστεί τέλος, ότι οι αλγόριθμοι που περιγράφηκαν παραπάνω, αφορούν μόνο την περίπτωση που η ποινή για τα κενά είναι απλή. Σε πραγματικά προβλήματα, απαιτούνται αλγόριθμοι που να υλοποιούν το ρεαλιστικότερο μοντέλο της σύνθετης ποινής για τα κενά. Σε αυτή την περίπτωση, ο αντίστοιχος αλγόριθμος έχει μεγαλύτερη πολυπλοκότητα, της τάξης του $O(nm^2+mn^2)$ γιατί σε κάθε βήμα θα πρέπει να «θυμάται» αν το κενό που έβαλε είναι το πρώτο (open) ή κάποιο από τα επόμενα (extension). Ο αλγόριθμος στην περίπτωση της ολικής στοίχισης γίνεται:

$$F(i,j)=\max\begin{Bmatrix}F(i-1,j-1)+s(x_i,y_j),\\F(k,j)-\gamma(i-k),k=0,..,i-1\\F(i,k)-\gamma(j-k),k=0,..,j-1\end{Bmatrix} \tag{3.42}$$

Όμοια τροποποίηση μπορεί να γίνει και στον αλγόριθμο της τοπικής στοίχισης. Έχουν προταθεί, παρόλα αυτά, τροποποιήσεις οι οποίες πραγματοποιούν τον ίδιο υπολογισμό σε χρόνο της τάξης $O(mn)$, με το αντιστάθμισμα, ότι απαιτείται μεγαλύτερη χρήση της μνήμης. Η βασική απαίτηση, είναι ότι η σύνθετη ποινή για τα κενά θα πρέπει να είναι της μορφής της σχέσης (3.35). Ο αλγόριθμος σε αυτή την περίπτωση απαιτεί 3 διαφορετικούς πίνακες, και θα είναι:

$$F(i,j)=\max\begin{Bmatrix}F\left(i-1,j-1\right)+s\left(x_i,y_j\right)\\I_x\left(i-1,j\right)+s\left(x_i,y_j\right)\\I_y\left(i,j-1\right)+s\left(x_i,y_j\right)\end{Bmatrix} \tag{3.43}$$

με

$$I_x\left(i,j\right)=\max\left\{F\left(i-1,j\right)-d,I_x\left(i-1,j\right)-e\right\} \tag{3.44}$$

και

$$I_y\left(i,j\right)=\max\left\{F\left(i,j-1\right)-d,I_y\left(i,j-1\right)-e\right\} \tag{3.45}$$

3.12 Ο νόμος των Erdos και Renyi για τη σύγκριση αλληλουχιών

Είδαμε στις προηγούμενες παραγράφους, με ποιους τρόπους μπορούμε να βρούμε την καλύτερη στοίχιση δύο αλληλουχιών. Το τελευταίο πρόβλημα που μένει, είναι αυτό της εκτίμησης της στατιστικής σημαντικότητας. Συγκεκριμένα μας ενδιαφέρει το πώς μπορούμε να διαχωρίσουμε «τυχαία» ευρήματα από «σημαντικά». Το P-value ενός στατιστικού ελέγχου (γιατί περί τέτοιου πρόκειται) είναι η πιθανότητα, ένα αποτέλεσμα τόσο ακραίο ή και περισσότερο, να έχει προκύψει κατά τύχη, δεδομένου ότι ισχύει η μηδενική υπόθεση (στην περίπτωση μας, ότι οι δύο αλληλουχίες που συγκρίναμε δεν έχουν καμία σχέση μεταξύ τους). Είναι προφανές ότι αναφερόμαστε σε παραμετρικό έλεγχο και χρειάζεται να ξέρουμε την κατανομή που ακολουθεί η τυχαία μεταβλητή που μας ενδιαφέρει, το σκορ δηλαδή. Για να απαντήσουμε σε αυτό το ερώτημα, θα πρέπει να δούμε πάλι το θέμα των ροών των επιτυχιών και τα δεδομένα της κατανομής του Gumbel.

Όπως είδαμε ήδη, η σύγκριση ακολουθιών, μοιάζει με τη μελέτη των ροών σε αλληλουχίες. Η διαφορά είναι ότι το πρόβλημα είναι τώρα δισδιάστατο. Κατά συνέπεια, οι ασυμπτωτικοί νόμοι των Erdos και Renyi που ισχύουν για τις ροές, βρίσκουν εφαρμογή και εδώ (M. S. Waterman, 1995). Έστω ότι έχουμε δυο αλληλουχίες $x=x_1,x_2,\ldots,x_n$ και $y=y_1,y_2,\ldots,y_m$. Τότε η μέγιστη περιοχή ταύτισης μεταξύ τους έχει μήκος $M_n \cong \log_{1/p}\left(mn\right)$ ή αλλιώς:

$$\frac{M_n}{\log_{1/p}\left(mn\right)}\to 1 \text{ με πιθανότητα } 1 \tag{3.46}$$

Προφανώς, η πιθανότητα ταύτισης p είναι ίση με $p=P(x_i=y_j)\Leftrightarrow p=p_A^2+p_T^2+p_G^2+p_C^2$ αν η κατανομή των συμβόλων στις δυο αλληλουχίες είναι ίδια. Η διαισθητική ερμηνεία εδώ είναι εντελώς ανάλογη με αυτή που δώσαμε στην περίπτωση της μελέτης μίας ακολουθίας, με τη μόνη διαφορά ότι τώρα υπάρχουν περίπου mn δυνατά σημεία εμφάνισης της περιοχής ταύτισης. Αντίστοιχο αποτέλεσμα θα έχουμε και για το μήκος της περιοχής μη απόλυτης ταύτισης. Αν για παράδειγμα έχουμε δυο αλληλουχίες $\mathbf{x}=x_1,x_2,\ldots,x_n$ και $\mathbf{y}=y_1,y_2,\ldots,y_m$ με $0\le p<\alpha\le 1$. Τότε για τη μέγιστη περιοχή που περιέχει $100\alpha\%$ όμοια νουκλεοτίδια μεταξύ τους ισχύει:

$$\frac{M_n}{\log_{1/p}(mn)} \to \frac{1}{H(a,p)} \text{ με πιθανότητα 1} \tag{3.47}$$

Η ποσότητα $H(a,p)$ είναι η σχετική εντροπία όπως την ορίσαμε παραπάνω. Όπως θα εξηγήσουμε, οι παραπάνω σχέσεις ισχύουν για τοπικές (local) συγκρίσεις ακολουθιών. Γενικά από εδώ και πέρα θα ασχοληθούμε με την κατανομή του Local Similarity Score αφ' ενός μεν γιατί είναι πιο σημαντικό από πρακτική άποψη αφ' ετέρου δε γιατί τα πιο σημαντικά θεωρητικά αποτελέσματα που έχουν βρεθεί, αφορούν αυτό. Επιπλέον, τα παραπάνω αποτελέσματα έχουν επεκταθεί (R. Arratia, Gordon, L. and Waterman, 1986; R. Arratia, Gordon, L. and Waterman, M. S., 1990) και έχουν δοθεί ακόμα πιο ακριβείς τύποι για τη μέση τιμή του μήκους της μέγιστης περιοχής ταύτισης. Για την ακρίβεια, ο Arratia (1990) έδωσε μια καλύτερη προσέγγιση για τη μέση τιμή του μήκους της μέγιστης περιοχής ταύτισης μεταξύ δύο ακολουθιών, η οποία είναι:

$$E(M_n) \approx \frac{\log(mn)}{\lambda} + \frac{\log(q)}{\lambda} + \frac{\gamma}{\lambda} - \frac{1}{2} \tag{3.48}$$

όπου q=1-p, και γ=-Γ'(1) = 0.5772… η σταθερά Euler-Mascheroni, και λ=log(1/p). Παράλληλα, έδωσαν και προσεγγιστικό τύπο για την αντίστοιχη διασπορά:

$$Var(M_n) \approx \frac{\pi^2}{6\lambda^2} + \frac{1}{12} \tag{3.49}$$

Οι σχέσεις αυτές, είναι εντελώς ανάλογες με τις αντίστοιχες σχέσεις (3.17) και (3.18) που είδαμε για τη μελέτη μίας αλληλουχίας. Εντελώς ανάλογη είναι και η ερμηνεία για τις διαφορές των σχέσεων (3.48) και (3.47), καθώς ο κύριος παράγοντας που καθορίζει την τελική τιμή εξακολουθεί να είναι η τιμή του log(mn), μιας το log(q) και το γ/λ είναι σχετικά μικρές ποσότητες. Προφανώς, όσο το m και n μεγαλώνουν, η διαφορά θα γίνεται ακόμα μικρότερη. Οι Arratia και Waterman (R. Arratia & Waterman, 1989), έδωσαν και ακριβή τύπο για τον υπολογισμό της μέγιστης περιοχής ταύτισης δυο αλληλουχιών, με δεδομένο τον αριθμό k των μη όμοιων νουκλεοτιδίων (mismatches). Σε αυτή την περίπτωση, έχουμε δυο αλληλουχίες $\mathbf{x}=x_1,x_2,\ldots,x_n$ και $\mathbf{y}=y_1,y_2,\ldots,y_m$ και μας ενδιαφέρει η μέση τιμή για το μήκος της μέγιστης περιοχής ταύτισης μεταξύ τους, όταν υπάρχουν k μη κοινά νουκλεοτίδια (k mismatches):

$$E(M_n) \approx \log_{1/p}(qn^2) + k\log_{1/p}\log_{1/p}(qn^2) + k\log_{1/p}(q) - \log_{1/p}(k!) + k + \frac{\gamma}{\lambda} - \frac{1}{2} \tag{3.50}$$

Για την αντίστοιχη διασπορά ισχύει όπως και προηγουμένως:

$$\text{var}(M_n) \approx \frac{\pi^2}{6\lambda^2} + \frac{1}{12} \tag{3.51}$$

Όπως και παραπάνω, q=1-p, και γ=-Γ'(1)=0.5772…η σταθερά Euler-Mascheroni, και λ=log(1/p).

3.13 Η ασυμπτωτική κατανομή του local similarity score

Το επόμενο λογικό βήμα είναι να προσδιορίσουμε την ακριβή κατανομή του Local Similarity Score και κατ' επέκταση του μήκους της μέγιστης κοινής υπό-ακολουθίας, έτσι ώστε να μπορούμε να υπολογίσουμε τη στατιστική σημαντικότητα μιας δεδομένης σύγκρισης δυο αλληλουχιών. Δηλαδή, να μπορέσουμε να προσδιορίσουμε αν το αποτέλεσμα μιας τέτοιας σύγκρισης οφείλεται στην τύχη και μόνο, ή αν υπάρχει μια βιολογική σημαντικότητα στην ομοιότητα αυτή των δυο αλληλουχιών. Παραδοσιακά, οι βιολόγοι είχαν αναπτύξει διάφορους εμπειρικούς κανόνες. Ο πιο ακριβής από αυτούς, λέει ότι δυο πρωτεΐνες είναι «όμοιες» αν έχουν τουλάχιστον 30% ομοιότητα (similarity) σε μήκος στοίχισης τουλάχιστον 80 αμινοξικά κατάλοιπα. Θα δούμε, ότι σε γενικές γραμμές ο κανόνας αυτός αποδίδει, αλλά ο ακριβής υπολογισμός της στατιστικής σημαντικότητας μπορεί να προσδώσει πολλά πλεονεκτήματα, ειδικά στις οριακές καταστάσεις, και ακόμα περισσότερο στις αναζητήσεις σε βάσεις δεδομένων, όπου το πρόβλημα των πολλαπλών ελέγχων είναι υπαρκτό.

Πρέπει να τονίσουμε εδώ ότι ακριβή θεωρητικά αποτελέσματα έχουν δοθεί μόνο για την κατανομή του «Local Similarity Score without gaps», δηλαδή για την περίπτωση κατά την οποία στην στοίχιση δεν υπάρχουν κενά, αν και υπάρχουν ενδείξεις ότι τα ίδια αποτελέσματα γενικεύονται και για την περίπτωση ύπαρξης κενών. Κατ' αρχήν πρέπει να δούμε τις δυο ακραίες περιπτώσεις. Πρώτον, όταν ορίσουμε

$$s\left(x_i, y_i\right) = \begin{cases} 1, \text{ αν } x_i = y_i \\ 0, \text{ αν } x_i \neq y_i \end{cases} \text{ και d=0}$$

δηλαδή, όταν δεν υπάρχουν ποινές για κενά και για διαφορές, τότε έχουμε $s(x_i,y_j)\sim cn$ και βρισκόμαστε στη λεγόμενη γραμμική περιοχή για την οποία δεν υπάρχουν προς το παρόν ενδείξεις για την κατανομή του σκορ, αλλά δεν υπάρχει και πρακτικό ενδιαφέρον καθώς με τις παραπάνω ποινές οποιεσδήποτε αλληλουχίες θα μπορούν να στοιχηθούν «καλά». Αντίθετα αν χρησιμοποιήσουμε

$$s\left(x_i, y_i\right) = \begin{cases} 1, \text{ αν } x_i = y_i \\ -\infty, \text{ αν } x_i \neq y_i \end{cases} \text{ και d=}\infty$$

δηλαδή, αν δεν επιτρέπονται καθόλου τα κενά και οι διαφορές, τότε έχουμε $s(x_i,y_j)\sim k\log n$ και βρισκόμαστε στη λογαριθμική περιοχή όπου ισχύει η σχέση (3.46). Προφανώς σ' αυτήν την περιοχή ανήκει και η σχέση (3.47) γιατί και σ' αυτή βλέπουμε ότι το μήκος της περιοχής ταύτισης αυξάνεται με τον λογάριθμο του n. Μειώνοντας σταδιακά τις ποινές για τις διαφορές και τα κενά, μεταπίπτουμε από τη λογαριθμική περιοχή του σκορ, στη γραμμική. Αυτή η μετάπτωση φάσεως (phase transition) έχει περιγραφεί θεωρητικά από τους Arratia, Gordon και Waterman (R. Arratia & Waterman, 1994; M. S. Waterman, 1995; M. S. Waterman, Gordon, & Arratia, 1987), αλλά παρ' όλα αυτά δεν υπάρχει αναλυτική έκφραση για τις τιμές των παραμέτρων m (mismatch) και d (gap), στις οποίες συμβαίνει αυτή η μετάπτωση.

Όπως είναι φανερό εμείς ενδιαφερόμαστε για την κατανομή του σκορ στη λογαριθμική περιοχή. Στην περιοχή αυτή η εμφάνιση θετικών σκορ, δηλαδή η ύπαρξη κοινών υπό-ακολουθιών, είναι σπάνια γεγονότα. Επομένως, ασυμπτωτικά θα περιγράφονται από μια κατάλληλη κατανομή Poisson, με μέση τιμή:

$$\mathrm{E}\left(S \geq x\right) = Kmne^{-\lambda x} = Kmnp^{x} \tag{3.52}$$

όπου K είναι μια σταθερά <1 η οποία διορθώνει τον παράγοντα mn, και λ η μοναδική θετική ρίζα της εξίσωσης $\sum q_i q_j e^{\lambda S} =1$. Στην ιδανική περίπτωση για την οποία δεν επιτρέπονται τα κενά αλλά ούτε και διαφορές, μπορούν να δοθούν κλειστές εκφράσεις για τα K και λ και αυτές είναι :

$$K = 1 - p = q \tag{3.53}$$

και

$$\lambda = \log\left(1/p\right) \tag{3.54}$$

Όταν από την άλλη πλευρά επιτρέπονται διαφορές, τότε τα K, λ υπολογίζονται με αριθμητικές μεθόδους. Πιο συγκεκριμένα, κομβικό ρόλο στη μελέτη των στατιστικών της τοπικής στοίχισης έχει το θεώρημα των Karlin και Altschul (S. Karlin & Altschul, 1990) το οποίο λέει ότι στη σύγκριση δύο αλληλουχιών $\mathbf{x}=x_1,x_2,\ldots,x_n$ και $\mathbf{y}=y_1,y_2,\ldots,y_m$ με σκορ S όπως το ορίσαμε στη σχέση (3.32), η πιθανότητα να προκύψει ένα σκορ μεγαλύτερο από το x (δηλαδή, το p-value), δίνεται από τη σχέση:

$$P\left(S > x\right) \approx 1 - \exp\left(-Kmne^{-\lambda x}\right) \tag{3.55}$$

Για τον υπολογισμό του μέγιστου τοπικού σκορ (local similarity score) πρέπει να θέσουμε κάποιους περιορισμούς, όμοιους με την περίπτωση του maximal segment score. Πιο συγκεκριμένα:

1. Τουλάχιστον ένα σκορ πρέπει να είναι θετικό
2. Η αναμενόμενη τιμή του σκορ για μια τυχαία θέση στη στοίχιση να είναι αρνητική, δηλαδή.

$$E\left(s_{ij}\right) = \sum q_i q_j s_{ij} = \sum q_i q_j \log\left(\frac{q_i q_j}{p_{ij}}\right) < 0 \tag{3.56}$$

Το λ είναι όπως είπαμε ήδη, η μοναδική θετική ρίζα της εξίσωσης $\sum q_i q_j e^{\lambda S} =1$. Προφανώς, οι παραπάνω δυο περιορισμοί είναι απαραίτητοι για να είμαστε σίγουροι ότι το σκορ θα παίρνει τιμές στη λογαριθμική περιοχή, και κατά συνέπεια θα είναι όντως τοπικό. Η σχέση αυτή, γράφεται ισοδύναμα ως:

$$P\left(S \leq x\right) = \exp\left(-Kmne^{-\lambda x}\right) \tag{3.57}$$

Η τελευταία σχέση είναι, με άλλη παραμετροποίηση, η α.σ.κ. της κατανομής των ακραίων τιμών του Gumbel (EVD). Αν κάνουμε μετασχηματισμό, βλέπουμε ότι ισχύει:

$$P\left(S \leq x\right) = \exp\left(-e^{-\frac{(x-a)}{b}}\right), -\infty \leq x \leq \infty \tag{3.58}$$

με

$$E(x) = a - b\Gamma'(1),\ V(x) = \frac{b^2\pi^2}{6} \tag{3.59}$$

Οι παράμετροι α,b είναι προφανώς $a = \frac{\log(kmn)}{\lambda}, b = \frac{1}{\lambda}$ με $\lambda = \log\left(\frac{1}{p}\right)$ και $K = 1 - p = q$, όταν δεν επιτρέπονται διαφορές. Από τις παραπάνω σχέσεις είναι δυνατόν να υπολογιστεί το p-value για ένα δεδομένο σκορ που προέκυψε από τη σύγκριση δυο αλληλουχιών. Αφού τυποποιήσουμε τη μεταβλητή, έχουμε (W. R. Pearson, 1998; W. R. Pearson & Wood, 2001):

$$P(Z \geq z) = 1 - \exp\left(-\exp-\left(\left(\frac{\pi}{\sqrt{6}}\right)z - \Gamma'(1)\right)\right) \tag{3.60}$$

Η σχέση αυτή, είναι ισοδύναμη με την (3.55) αλλά δεν περιέχει σταθερές ενώ το σκορ είναι εκφρασμένο σε τυποποιημένες μονάδες (z). Από την παραπάνω σχέση μπορούμε να υπολογίσουμε την πιθανότητα να εμφανισθεί ένα σκορ που να ξεπερνά κάποιες (z) φορές την τυπική απόκλιση της θεωρητικής κατανομής δεδομένου ότι οι αλληλουχίες είναι τυχαίες (ασυσχέτιστες).

Πρέπει σ' αυτό το σημείο να τονίσουμε ότι διαφορετική ερμηνεία έχει ένα p-value που προκύπτει από τη σύγκριση δυο διαφορετικών αλληλουχιών, και άλλη έχει ένα p-value που θα προκύψει από σύγκριση μεταξύ της ίδιας αλληλουχίας με μια μεγάλη βάση δεδομένων με πολλές χιλιάδες αλληλουχίες. Αυτό ισχύει ακόμα και αν η στοίχιση που προέκυψε στις δύο περιπτώσεις είναι πανομοιότυπη. Έτσι ένα p-value που προκύπτει από τη σύγκριση δυο διαφορετικών αλληλουχιών με τιμή 10^{-4}, μπορεί να φαίνεται στατιστικά σημαντικό αλλά αν πρόκειται για σύγκριση μεταξύ μιας ακολουθίας με μια μεγάλη βάση δεδομένων με 100.000 αλληλουχίες τότε λόγω της τύχης και μόνο αναμένεται να εμφανιστεί τουλάχιστον 10 φορές.

Όταν συγκρίνουμε μια αλληλουχία με μια ολόκληρη βάση δεδομένων, η οποία περιέχει D αλληλουχίες, τότε η παρατήρηση ακολουθιών οι οποίες εμφανίζουν μικρό p-value (μεγάλη ομοιότητα- p-match) είναι σπάνιο ενδεχόμενο, και θα περιγράφεται από την κατανομή Poisson. Άρα (W. R. Pearson & Wood, 2001): $P = \Pr(\text{τουλάχιστον 1 score } S \geq x) = 1 - e^{-Dp}$ και αν το Dp είναι πολύ μικρό (<0.01) θα έχουμε: $P \approx Dp$. Στο ίδιο αποτέλεσμα θα καταλήγαμε αν υπολογίζαμε την αναμενόμενη τιμή για τις εμφανίσεις περιοχών με σκορ S≥x, έπειτα από D συγκρίσεις με τις αλληλουχίες της βάσης δεδομένων. Αυτό το E-value (expectation value) είναι ίσο με $E(S \geq x) = DP(S \geq x)$ όπου D είναι ο αριθμός των ανεξάρτητων αλληλουχιών που περιέχει η υπό έλεγχο βάση δεδομένων.

Για να έχουμε περισσότερο ακριβή αποτελέσματα, μια πιο σωστή προσέγγιση θα προέκυπτε αν λαμβάναμε υπόψη το γεγονός ότι όλες οι αλληλουχίες στη βάση δεδομένων δεν έχουν τον ίδιο αριθμό βάσεων. Πρακτικά αυτό σημαίνει ότι θεωρούμε ολόκληρη τη βάση δεδομένων ως μια τεράστια αλληλουχία από N νουκλεοτίδια (ή αμινοξέα) και συγκρίνουμε με αυτήν τη συγκεκριμένη αλληλουχία μας η οποία έχει μήκος n βάσεις (ή αμινοξέα). Κατά μέσο όρο κάθε μια από τις αλληλουχίες της βάσης περιέχει $m=N/D$ βάσεις, οπότε η πιθανότητα να υπάρχει μια περιοχή με σκορ μεγαλύτερο από x, όπως είπαμε παραπάνω είναι $P(S > x) = 1 - e^{-E(S)} = 1 - \exp\left(-KNne^{-\lambda x}\right)$ ενώ η αναμενόμενη τιμή (E-value) θα είναι $E(S \geq x) = KNne^{-\lambda x} = DKmne^{-\lambda x}$.

Πολλά προγράμματα όπως το BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990), αντί του p-value, αναφέρουν ως αποτέλεσμα (output) αυτήν την τιμή, επειδή είναι πιο εύκολη η ερμηνεία της από κάποιο μη ειδικό, αλλά όπως είδαμε όταν το E-value είναι πολύ μικρό τότε, επειδή ισχύει η προσεγγιστική σχέση (M. S. Waterman, 1995): $1 - \exp\left(-\exp(-t)\right) \approx 1 - \left(1 - \exp(-t)\right) = \exp(-t)$, το p-value θα είναι περίπου ίσο με το E-value. Είναι φανερό ότι σήμερα που οι βάσεις δεδομένων αυξάνονται συνεχώς σε μέγεθος είναι καλύτερο κάθε φορά που γίνονται τέτοιες συγκρίσεις να αναφέρονται τουλάχιστον μαζί το p-value και το E-value.

3.14 Η κατανομή του σκορ όταν υπάρχουν κενά

Όταν στη στοίχιση δυο αλληλουχιών υπάρχουν κενά δεν υπάρχει μαθηματική απόδειξη που να περιγράφει την κατανομή που ακολουθεί το σκορ. Παρ' όλα αυτά πολλοί ερευνητές έχουν προτείνει (Altschul et al.,

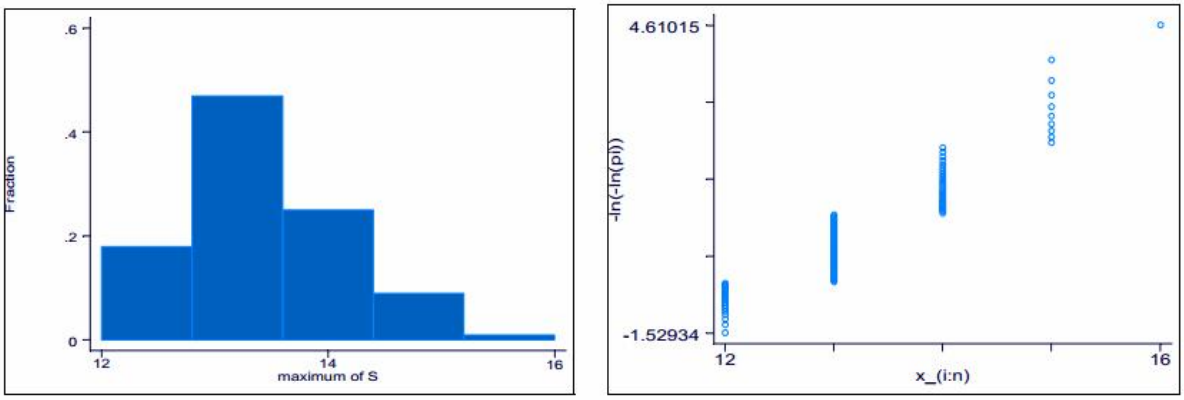
1997; Clote & Backofen, 2000; Mott, 2000), ότι και σ' αυτήν την περίπτωση η κατανομή του σκορ είναι η κατανομή των ακραίων τιμών του Gumbel:

$$P\left(S \le x\right) = \exp\left(-Kmne^{-\lambda x}\right) \tag{3.61}$$

με τη διαφορά ότι οι παράμετροι *K,λ* είναι διαφορετικές από την περίπτωση της στοίχισης χωρίς κενά.

Η πρώτη προσπάθεια να υπολογιστούν οι παράμετροι της κατανομής του Gumbel όταν υπάρχουν κενά έγινε από τον Mott, το 1992 (Mott, 1992). Συγκεκριμένα θεώρησε μια παραλλαγή της σχέσης (3.58) και έθεσε $A = a_0 + \frac{a_1}{\lambda} + \frac{\alpha_2 \log\left(mn\right)}{\lambda}, B = \frac{b_1}{\lambda}$. Τις σταθερές α_0, α_1, α_2 και b_1 τις εκτίμησε με μέγιστη πιθανοφάνεια. Το λ είναι και πάλι η μοναδική θετική ρίζα της εξίσωσης $\sum q_i q_j e^{\lambda S} = 1$.

Ένας πιο απλός, αλλά και χρονοβόρος τρόπος υπολογισμού των παραμέτρων αυτών είναι η απευθείας εκτίμηση (direct estimation) (M. S. Waterman, 1995; M. S. Waterman & M. Vingron, 1994). Στην περίπτωση αυτή απαιτείται ένας μεγάλος αριθμός προσομοιώσεων (συγκρίσεις με τυχαίες αλληλουχίες). Πιο συγκεκριμένα, αφού πραγματοποιήσουμε την τοπική στοίχιση των δυο αλληλουχιών, πραγματοποιούμε πολλές (στη βιβλιογραφία αναφέρεται ότι πρέπει να είναι τουλάχιστον 1000) συγκρίσεις μεταξύ τυχαίων αλληλουχιών με παρόμοια σύσταση βάσεων με τις αρχικές (αυτό ονομάζεται shuffling, ανακάτεμα αλληλουχιών). Κατόπιν, αφού υπολογίσουμε την εμπειρική α.σ.κ. (e.c.d.f.) και εφαρμόσουμε κατάλληλο μετασχηματισμό (log[-log[cdf]]) κάνουμε μια απλή γραμμική παλινδρόμηση του log[-log[cdf]] με το *S*. Η κλίση (slope) της ευθείας ελαχίστων τετραγώνων θα είναι ίση με –λ και η σταθερά της (constant) θα είναι ίση με log(*Kmn*).



Εικόνα 3.12: *Κατανομή της μέγιστης κοινής υπο-ακολουθίας από τις συγκρίσεις αλληλουχιών DNA μήκους 10000 βάσεων. Αριστερά βλέπουμε την κατανομή του σκορ ενώ δεξιά, τον μετασχηματισμό log[-log[cdf]]από τον οποίο θα εκτιμήσουμε τις σταθερές K και λ της κατανομής.*

Εναλλακτικά, αν πραγματοποιείται σύγκριση μιας αλληλουχίας με μια βάση δεδομένων, είναι δυνατόν για την εύρεση της κατανομής, να χρησιμοποιηθούν τα αποτελέσματα από τις συγκρίσεις που έγιναν κατά τη διάρκεια της αναζήτησης. Ιδιαίτερη προσοχή εδώ θέλει το γεγονός, ότι για να πετύχει η μέθοδος πρέπει να έχουν απομακρυνθεί όλες οι αλληλουχίες της βάσης με πολύ μεγάλα (z>7) και πολύ μικρά (z<-3) σκορ, έτσι ώστε να αποφευχθεί μεροληψία στα αποτελέσματα (systematic error – bias) (W. R. Pearson, 1998).

Μια άλλη μέθοδος που προτάθηκε από τους Waterman και Vingron (M. S. Waterman & M. Vingron, 1994; M. S. Waterman & M. Vingron, 1994), αναδεικνύει τη δύναμη της προσέγγισης Poisson (Poisson Approximation - (R. Arratia, Goldstein, & Gordon, 1989; Chen, 1975)) και ονομάζεται "de-clumping estimation". Η μέθοδος αυτή χρησιμοποιεί το διατεταγμένο δείγμα $S_{(1)} \ge S_{(2)} \ge \ldots \ge S_{(k)}$ και όχι μόνο το μέγιστο σκορ. Βασική προϋπόθεση εδώ είναι ότι οι τοπικές περιοχές που θα δώσουν αυτά τα σκορ πρέπει να είναι ανεξάρτητες μεταξύ τους, δηλαδή να μην ταυτίζονται σε κάποιο μικρότερο τμήμα τους. Οι διατεταγμένες παρατηρήσεις $S_{(i)}$ ακολουθούν ασυμπτωτικά την κατανομή Poisson με μέση τιμή, όπως είδαμε :

$$E\left(S \ge x\right) = Kmne^{-\lambda x} \tag{3.62}$$

Άρα, η πιθανότητα να υπάρχουν *k* ανεξάρτητες μεταξύ τους περιοχές, με σκορ μεγαλύτερο από *x* θα είναι:

$$P\left(S_{(k)} > x\right) \approx 1 - \exp\left(-Kmne^{-\lambda x}\right) \sum_{i=0}^{k-1} \frac{\left(Kmne^{-\lambda x}\right)^i}{i!} \tag{3.63}$$

και η μέση τιμή αυτής της κατανομής θα δίνεται από τη σχέση (3.52). Επομένως, παριστάνοντας γραφικά το log[data] (το λογάριθμο του αριθμού τοπικών περιοχών με σκορ πάνω από κάποιο όριο) σε σχέση με το $Kmne^{-\lambda x}$ (τη μέση τιμή του σκορ για τις περιοχές πάνω από το όριο αυτό) παίρνουμε ευθεία γραμμή και μια απλή γραμμική παλινδρόμηση δίνει αμέσως εκτιμήτριες για τα K,λ. Ισοδύναμα μια παλινδρόμηση Poisson μεταξύ αριθμού παρατηρήσεων που ξεπερνούν κάποιο σημείο, και της μέσης τιμής του σκορ για παρατηρήσεις πάνω από το κάθε σημείο, δίνει τις εκτιμήτριες για τα K,λ.

Οι δυο προηγούμενες μέθοδοι δίνουν ταυτόσημα αποτελέσματα αλλά η δεύτερη είναι πολλές φορές γρηγορότερη καθώς απαιτεί λιγότερες προσομοιώσεις. Τούτο συμβαίνει διότι για κάθε προσομοίωση η μέθοδος του Poisson υπολογίζει μόνο μια φορά τον πίνακα nm από τον αλγόριθμο Smith-Waterman, και από αυτόν υπολογίζει τα k υποβέλτιστα σκορ (sub-optimal alignments). Οι Waterman και Vingron, αναφέρουν ότι χρειάζονται περίπου 10 προσομοιώσεις, με κατάλληλο αριθμό sub-optimal scores για να πάρουμε καλούς εκτιμητές για τα K,λ.

Μια άλλη προσέγγιση στην εύρεση της στατιστικής σημαντικότητας όταν επιτρέπεται η ύπαρξη κενών, είναι αυτή που προτάθηκε από τον Pearson (W. R. Pearson, 1995, 1998; W. R. Pearson & Wood, 2001), και αφορά τη σύγκριση μιας αλληλουχίας με μια βάση δεδομένων. Είναι δηλαδή παραλλαγή της μεθόδου απευθείας εκτίμησης. Κατά τη διαδικασία αυτή, η βάση δεδομένων χωρίζεται σε k υποσύνολα σύμφωνα με το μήκος των αλληλουχιών n_1, n_2,…, n_k που περιέχουν, έτσι ώστε τα υποσύνολα αυτά να διαφέρουν το καθένα από το επόμενο στο μέσο μήκος αλληλουχιών που περιέχουν, κατά περίπου 10%. Υπολογίζονται κατόπιν, όλα τα σκορ S, για την τοπική ομοιότητα των αλληλουχιών, και στη συνέχεια μια ευθεία σταθμισμένης γραμμικής παλινδρόμησης (weighted linear regression) για τη σχέση:

$$S = a + b\log\left(n_i\right) \tag{3.64}$$

Εδώ, το n_i, είναι το μήκος των αλληλουχιών του i υποσυνόλου της βάσης δεδομένων, ενώ το $\log(n_i)$ είναι σταθμισμένο με την αντίστροφη διασπορά (1/var) των σκορ σε αυτό το υποσύνολο, καθώς τμήματα με πολύ μεγάλο σκορ θα έχουν και μεγάλη διασπορά. Υπολογίζεται τέλος η εκτιμήτρια της διασποράς $\hat{\sigma}^2$, των καταλοίπων της παλινδρόμησης (residual variance) η οποία καθορίζει το z-score:

$$z = \frac{S - \left(a + b\log\left(n_i\right)\right)}{\text{var}} \tag{3.65}$$

Οι αλληλουχίες με πολύ μεγάλη διασπορά του σκορ εξαιρούνται, επειδή θεωρούνται ότι είναι αυτές με μεγάλη ομοιότητα, και άρα θα προσδώσουν συστηματικό σφάλμα στις εκτιμήσεις των παραμέτρων. Η όλη διαδικασία επαναλαμβάνεται έως και 5 φορές, για να απομακρυνθούν όλες οι συσχετισμένες (με μεγάλο σκορ) αλληλουχίες. Τελικά υπολογίζονται όλα τα z-scores, για τις αλληλουχίες της βάσης, και από τη σχέση :

$$P\left(Z \geq z\right) = 1 - \exp\left(-\exp-\left(\left(\frac{\pi}{\sqrt{6}}\right)z - \Gamma'(1)\right)\right) \tag{3.66}$$

υπολογίζεται η στατιστική σημαντικότητα (p-value, E-value), για κάθε μια από τις συγκρίσεις που έχουν πραγματοποιηθεί. Η μέθοδος αυτή είναι πολύ χρήσιμη καθώς επιτρέπει «εσωτερική» ρύθμιση για την ακρίβεια των παραμέτρων που εκτιμούμε, και έχει αποδειχθεί ότι δίνει πολύ καλά αποτελέσματα στην πράξη.

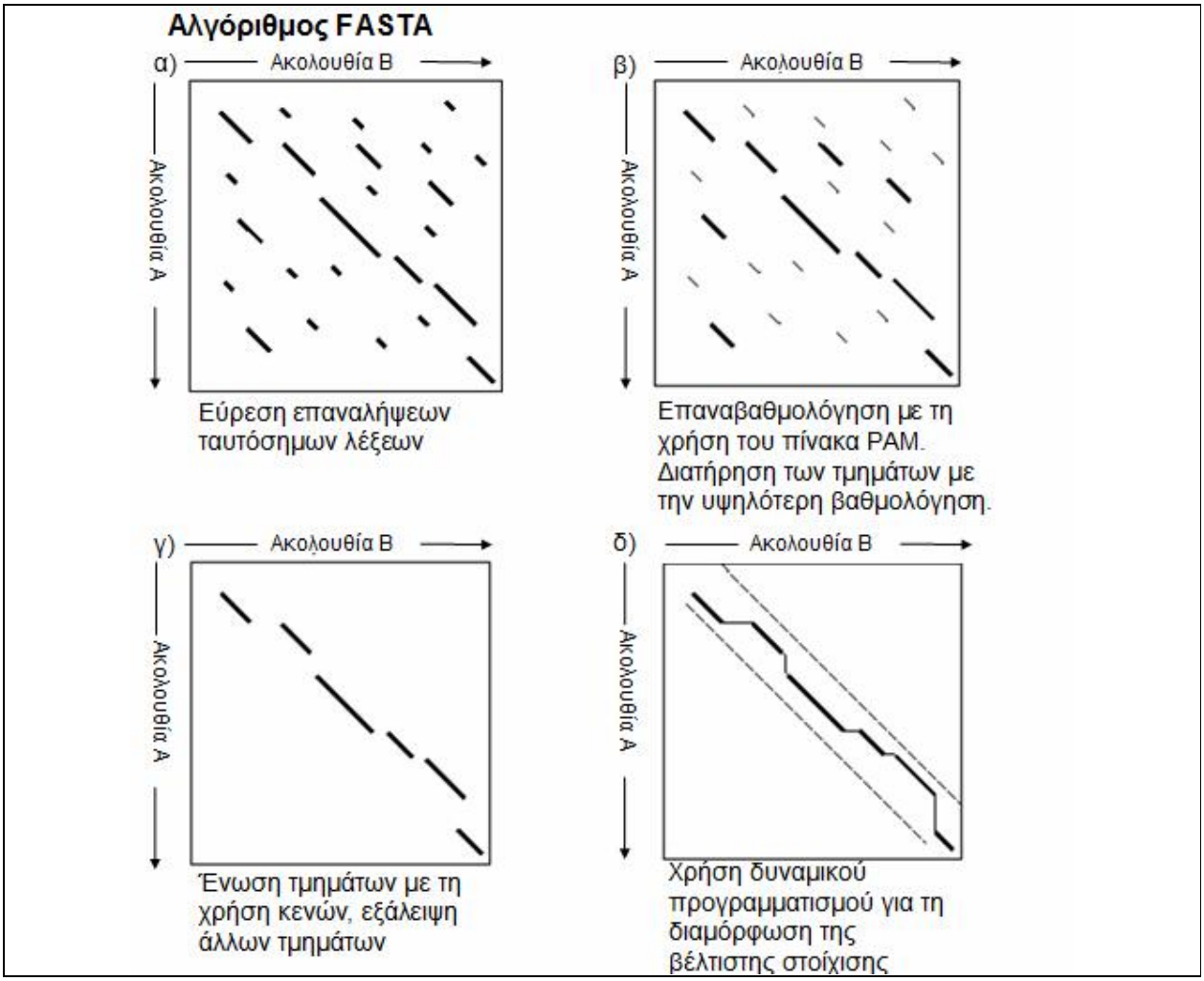
Ο Mott (Mott, 2000), πρότεινε τέλος, μια άλλη μέθοδο για τον υπολογισμό των παραμέτρων K, λ της κατανομής του Gumbel όταν υπάρχουν κενά. Συγκεκριμένα χρησιμοποιώντας μια μικρή τροποποίηση του αλγόριθμου των Smith-Waterman, και ένα συνδυασμό αριθμητικών και στατιστικών μεθόδων, κατόρθωσε να δώσει ακριβείς τύπους για τον υπολογισμό των K, λ ως συνάρτηση της ποινής για τα κενά. Οι τύποι αυτοί, στην περίπτωση που δεν επιτρέπονται κενά, ανάγονται στις αντίστοιχες παραμέτρους όπως τις προβλέπει η γενική θεωρία. Η παραπάνω μέθοδος δίνει επίσης πολύ καλά αποτελέσματα και επιπλέον λαμβάνει υπόψη τη διαφορετική σύνθεση κάθε αλληλουχίας της βάσης δεδομένων.

Όπως είναι φανερό, αν και δεν έχουμε το πλήρες θεωρητικό πλαίσιο για να εκτιμήσουμε τη στατιστική σημαντικότητα από συγκρίσεις αλληλουχιών όταν επιτρέπονται τα κενά, έχουμε στη διάθεση μας πληθώρα μεθόδων, που δίνουν πολύ καλά προσεγγιστικά αποτελέσματα. Το ποια μέθοδος θα χρησιμοποιηθεί, πέραν από τη διαθεσιμότητα των προγραμμάτων και τους πρακτικούς περιορισμούς, εξαρτάται κυρίως από το είδος των αλληλουχιών που συγκρίνουμε και το είδος της ομοιότητας που ελπίζουμε να ανακαλύψουμε.

3.15 Ευριστικοί αλγόριθμοι - BLAST και FASTA

Όπως είπαμε ήδη, ο αλγόριθμος Smith και Waterman δίνει σε κάθε περίπτωση την καλύτερη δυνατή στοίχιση μεταξύ δύο αλληλουχιών, λαμβάνοντας υπόψιν τον πίνακα ομοιότητας και τις ποινές για τα κενά. Παρόλα αυτά, σε πρακτικές εφαρμογές, είναι δύσχρηστος. Τούτο συμβαίνει, όχι τόσο για την περίπτωση που ενδιαφερόμαστε για τη σύγκριση δύο αλληλουχιών, αλλά περισσότερο για την αναζήτηση ομοιότητας σε μια βάση δεδομένων.

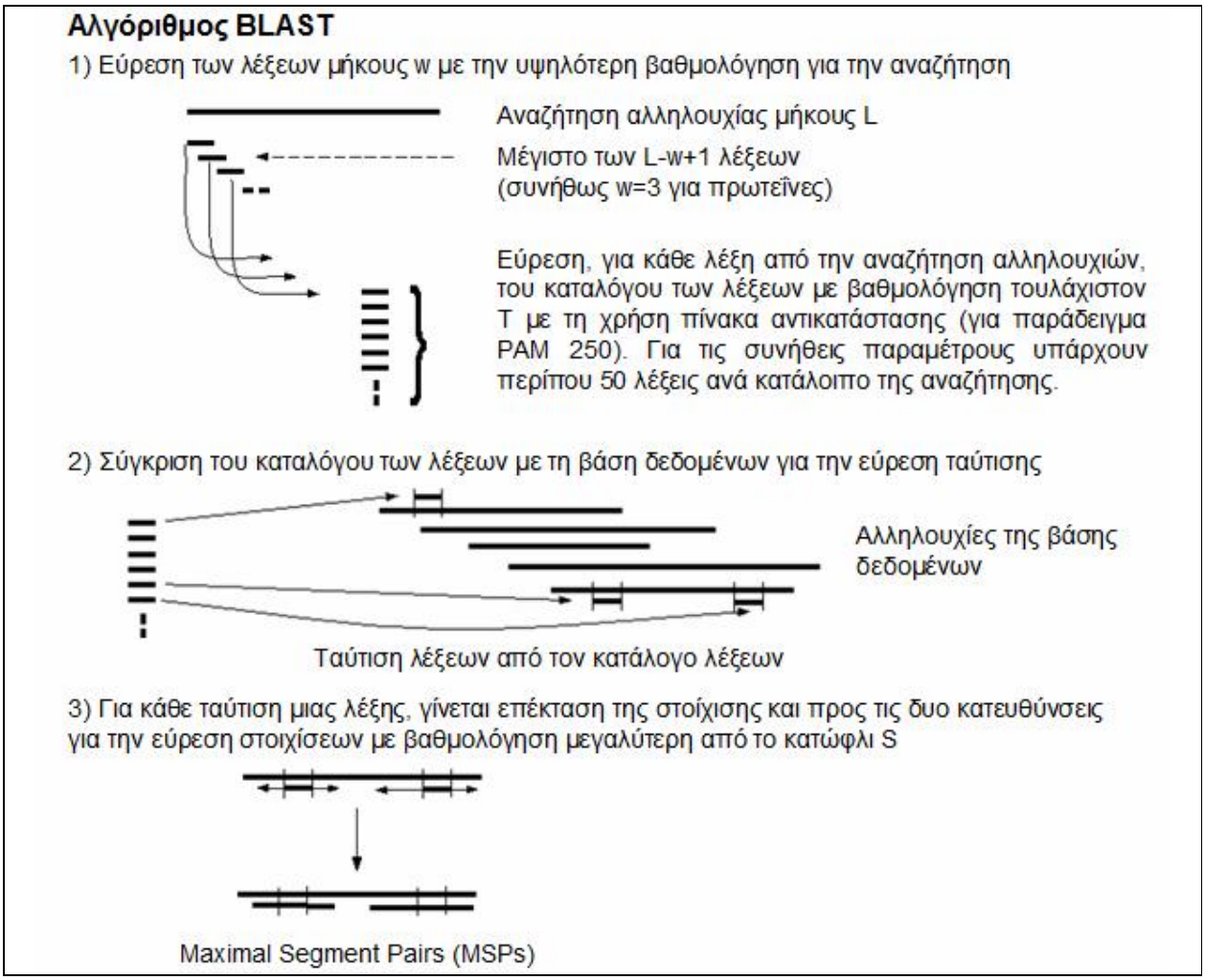
Οι αναζητήσεις στις βάσεις δεδομένων, είναι ένα βασικό εργαλείο στην υπολογιστική ανάλυση αλληλουχιών και είναι στην πραγματικότητα, μέρος της καθημερινής ρουτίνας ακόμα και των εργαστηριακών μοριακών βιολόγων. Το μεγάλο πρόβλημα προκύπτει, όπως έχουμε δει στο κεφάλαιο 2, από τη συνεχή αύξηση του όγκου των δεδομένων που βρίσκονται κατατεθειμένα στις δημόσιες βάσεις. Είδαμε, ότι ο αριθμός των καταχωρήσεων διπλασιάζεται σε λιγότερο από δύο χρόνια, και ο ρυθμός αυτός είναι ίσως και πιο γρήγορος από την αύξηση της υπολογιστικής ισχύος. Αυτό, το είχαν αντιληφθεί ήδη από τη δεκαετία του 1980, οπότε και ξεκίνησε η έρευνα για τη δημιουργία γρήγορων και αποδοτικών ευριστικών αλγορίθμων, οι οποίοι θα κάνουν την ίδια δουλειά αλλά σε μικρότερο χρόνο. Η βασική απαίτηση από έναν ευριστικό (heuristic) αλγόριθμο, είναι να αποδίδει «σχεδόν» πάντα το ίδιο καλά με τον αυστηρά μαθηματικό αλγόριθμο, αλλά να πραγματοποιεί τις αναλύσεις πολλές φορές πιο γρήγορα. Το «σχεδόν πάντα», δεν μπορεί να αποδειχθεί θεωρητικά αλλά μπορεί να τεκμηριωθεί με εμπειρικές αναλύσεις. Οι δύο πιο σημαντικοί αλγόριθμοι αυτής της κατηγορίας, είναι το Το BLAST (Altschul, et al., 1990; Altschul, et al., 1997) και το FASTA (Lipman & Pearson, 1985; Wilbur & Lipman, 1983).



Εικόνα 3.13: *Διαγραμματική απεικόνιση του αλγόριθμου FASTA*

Η βασική ιδέα του FASTA (www.ebi.ac.uk/fasta33/), είναι να εντοπίσει κατά προσέγγιση τη διαγώνιο γύρω από την οποία βρίσκεται η στοίχιση, για να περιορίσει έτσι κατά πολύ το εύρος της αναζήτησης. Για το σκοπό αυτό χρησιμοποιεί τα εξής βήματα:

- Στην αρχή δημιουργείται ένα ευρετήριο με τις θέσεις όλων των *k*-tuples (τυπικό μήκος για αμινοξικές αλληλουχίες 1 ή 2) που υπάρχουν ταυτόχρονα και στις δύο αλληλουχίες.
- Από τη διαφορά των θέσεων τους στις δύο αλληλουχίες εντοπίζεται η διαγώνιος στην οποία βρίσκονται, οπότε στο επόμενο βήμα εντοπίζονται οι διαγώνιες με τα περισσότερα *k*-tuples.
- Ακολούθως, αυτές οι περιοχές ταύτισης συνενώνονται επιτρέποντας την εισαγωγή κενών με τον υπολογισμό της αντίστοιχης ποινής, και
- Τελικά πραγματοποιείται η διαδικασία πλήρους δυναμικού προγραμματισμού (με τον επιλεγμένο πίνακα αντικατάστασης), περιορισμένου όμως μόνο σε μια ζώνη γύρω από τις συγκεκριμένες διαγωνίους.



Εικόνα 3.14: *Διαγραμματική απεικόνιση του αλγόριθμου BLAST*

Η διαδικασία του BLAST (www.ncbi.nlm.nih.gov/BLAST/), μοιάζει στα αρχικά στάδια, αλλά είναι ακόμα πιο γρήγορη καθώς πολλές παραμέτρους τις έχει προϋπολογισμένες και αποφεύγει τον δυναμικό προγραμματισμό:

- Η διαδικασία της σύγκρισης ξεκινά με την κατασκευή ενός καταλόγου όλων των λέξεων που θα ταίριαζαν με κάποια λέξη της άγνωστης αλληλουχίας και ξεπερνούν την τιμή κατωφλίου (προκαθορισμένη τιμή για πρωτεϊνικές αλληλουχίες T=13).

- Στη συνέχεια, ο αλγόριθμος αναζητά αυτές τις λέξεις στις αλληλουχίες της βάσης δεδομένων και κάθε φορά που εντοπίζει κάποια τέτοια ξεκινάει μια διαδικασία επέκτασης του 'ευρήματος' προς τις δύο κατευθύνσεις, όσο η βαθμολογία συνεχίζει και αυξάνει.
- Οι περιοχές μέγιστης βαθμολογίας που εντοπίζονται σε αυτό το στάδιο είναι οι υποψήφιες περιοχές ομοιότητας (HSPs, high scoring pairs).
- Από όλα τα HSPs αναφέρονται στα αποτελέσματα εκείνες οι περιοχές στις οποίες η βαθμολογία υπερβαίνει μια δεύτερη τιμή κατωφλίου S
- Τελικά, επιλέγονται να αναφερθούν εκείνες μόνο οι τοπικές ομοιότητες οι οποίες εμφανίζουν υψηλή στατιστική σημαντικότητα, ο προσδιορισμός της οποίας βασίζεται στο θεώρημα Karlin και Altschul.

Οι αρχικές εκδόσεις του BLAST, δεν επέτρεπαν την εισαγωγή κενών και έτσι ο αλγόριθμος ήταν ένα απλά εύχρηστο και γρήγορο εργαλείο για τον εντοπισμό όμοιων αλληλουχιών. Από τη 2η έκδοση όμως του προγράμματος και μετά, προστέθηκε και η δυνατότητα εισαγωγής κενών με συνέπεια το BLAST να μπορεί να χρησιμοποιηθεί και σαν γενικό πρόγραμμα στοίχισης. Το BLAST γενικά, έχει κερδίσει την αποδοχή της κοινότητας, τόσο γιατί είναι ελεύθερα διαθέσιμο και συνδεδεμένο με τις βάσεις του NCBI, όσο και γιατί είναι ο πιο γρήγορος από τους αλγόριθμους στοίχισης.

Διαφορετικός είναι ο τρόπος υπολογισμού της στατιστικής σημαντικότητας των ευρημάτων. Ενώ το BLAST υπολογίζει τις παραμέτρους της κατανομής (K, λ) από προσομοιώσεις, που έχει πραγματοποιήσει από πριν και έχει αποθηκευμένες τις παραμέτρους, το FASTA τις υπολογίζει από όλες τις άλλες αλληλουχίες της βάσης δεδομένων και για αυτόν τον λόγο είναι και πιο αργό.

Οι νεότερες εκδόσεις του BLAST αυτών περιέχουν πολλές τροποποιήσεις που επιτρέπουν πιο ακριβείς υπολογισμούς με χρήση profile και ειδικών ανά θέση πινάκων ομοιότητας (PSI-BLAST) (Altschul, et al., 1997), τεχνική που θα περιγράψουμε σε επόμενο κεφάλαιο.

Το BLAST, χρησιμοποιεί επίσης και μια σειρά βελτιστοποιήσεων για τον ακριβή υπολογισμό της στατιστικής σημαντικότητας, πέραν της κλασικής θεωρίας των Karlin και Altschul. Έτσι, χρησιμοποιεί επιπλέον και μια διόρθωση στο μήκος των αλληλουχιών για να λάβει υπόψη το «αποτελεσματικό μήκος» (effective length) της αλληλουχίας και της βάσης δεδομένων. Συγκεκριμένα θέτει

$$m' = m - \frac{\log(Kmn)}{H} \text{ και } n' = n - \frac{\log(Kmn)}{H}$$

δηλαδή αναπροσαρμόζει το αποτελεσματικό μήκος της αλληλουχίας και της βάσης δεδομένων για να λάβει υπόψη το γεγονός ότι με αυτά τα μήκη και τον δεδομένο πίνακα (substitution matrix) δεν επιτρέπονται όλες οι στοιχίσεις. Πρακτικά, αυτό δίνει διαφορά όταν οι αλληλουχίες της βάσης είναι μικρές. Θεωρητικά, αν η βάση ήταν ολόκληρη μια τεράστια αλληλουχία, το αποτελεσματικό μήκος της αλληλουχίας και το πραγματικό, θα ήταν τα ίδια. Το H είναι η σχετική εντροπία του πίνακα για τη δεδομένη σύσταση και το μήκος των αλληλουχιών που συγκρίνονται.

$$H = \sum_{i=1}^{20} \sum_{j=1}^{20} q_{ij} s_{ij} \tag{3.67}$$

Η σχετική εντροπία εκφράζει το μέσο ποσό πληροφορίας που είναι διαθέσιμο για κάθε ζεύγος καταλοίπων που στοιχίζεται, και διαχωρίζει την προκύπτουσα στοίχιση από μια τυχαία στοίχιση που οφείλεται απλά στις συχνότητες υποβάθρου. Υψηλότερη τιμή της σχετικής εντροπίας συνεπάγεται εύκολο διαχωρισμό μεταξύ των συχνοτήτων στόχων και υποβάθρου. Η ποσότητα αυτή, εκφράζει την αναμενόμενη τιμή του σκορ της αντικατάστασης (expected substitution score per position), και σε αντίθεση με την παρόμοια ποσότητα στη σχέση (3.56), η οποία εκφράζει την αναμενόμενη τιμή του σκορ για κάθε θέση της στοίχισης (expected per position alignment score), πρέπει να είναι θετική.

Κάτι άλλο που πρέπει να τονιστεί είναι ότι, λόγω του γεγονότος ότι πολλές φορές χρησιμοποιούνται διαφορετικά σχήματα για το σκορ (gap penalties, mismatches), είναι αναγκαίο να αναφέρεται και μια αντικειμενική τιμή για το σκορ. Αυτό μπορεί να επιτευχθεί κανονικοποιώντας το σκορ με βάση το bit (Altschul, et al., 1990; Altschul, et al., 1997):

$$S_{bit} = \frac{\lambda S_{raw} - \log K}{\log 2} \tag{3.68}$$

όπου S_{raw}, είναι το σκορ που υπολογίστηκε με κάποιες συγκεκριμένες τιμές για κενά και διαφορές. Αντικαθιστώντας τώρα στην σχέση (3.47) θα έχουμε:

$$E\left(S_{bit}\right) = mn2^{-S_{bit}} \tag{3.69}$$

Η τελευταία σχέση, δίνει ακριβώς ίδιες τιμές με την (3.61) αλλά είναι πιο εύκολη στον υπολογισμό, όταν έχουμε σαν δεδομένο το bit Score.

Το FASTA ενσωματώνει επίσης ακριβέστερους τρόπους υπολογισμού της στατιστικής σημαντικότητας ενός ευρήματος όταν υπάρχουν κενά (W. R. Pearson, 1998). Πρέπει όμως να τονιστεί, ότι το BLAST σε αντίθεση με το FASTA δεν μπορεί να χρησιμοποιήσει κάθε νέα μέθοδο, ειδικά αυτές που για τον υπολογισμό της σημαντικότητας χρησιμοποιούν τα αποτελέσματα της αναζήτησης στη βάση. Τούτο συμβαίνει γιατί το BLAST για τις αλληλουχίες για τις οποίες δεν βρήκε κάποια ομοιότητα, δεν θα έχει υπολογισμένο κάποιο σκορ της στοίχισης.

Τέλος, πρέπει να σημειώσουμε, ότι τα πακέτα αυτά περιέχουν πολλές εκδόσεις που επιτρέπουν τη σύγκριση αλληλουχιών DNA με DNA, Πρωτεΐνες με Πρωτεΐνες, αλλά και εναλλακτικούς συνδυασμούς δηλαδή τη σύγκριση ενός γονίδιου (DNA) με μια βάση δεδομένων πρωτεϊνών (μετάφραση του γονιδίου), αλλά και τη σύγκριση μιας πρωτεΐνης με μια βάση αλληλουχιών DNA, και τέλος τη σύγκριση DNA με DNA αφού πρώτα αυτά μεταφραστούν (δηλαδή σύγκριση DNA-DNA στο πρωτεϊνικό επίπεδο). Σε γενικές γραμμές και το BLAST και το FASTA παρέχουν αποτελέσματα σχεδόν παραπλήσια με τους κλασικούς αλγόριθμους δυναμικού προγραμματισμού και το ποιο πακέτο θα χρησιμοποιηθεί από κάποιον είναι θέμα που εξαρτάται κυρίως από το πού αποσκοπεί η έρευνά του (ακρίβεια), από την ταχύτητα και από τις ανάγκες παραμετροποίησης που έχει (είδος ακολουθίας που συγκρίνεται, πλήθος των πινάκων του σκορ, ποινές για κενά κλπ).

Βιβλιογραφία

Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol, 219*(3), 555-565.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25*(17), 3389-3402.

Arratia, R., Goldstein, L., & Gordon, L. (1989). Two moments suffice for Poisson approximation: The Chen-Stein method. *Ann. Probab., 17*, 9-25.

Arratia, R., Gordon, L. and Waterman. (1986). An extreme value theory for sequence matching. *Ann. Statist., 14*, 971-993.

Arratia, R., Gordon, L. and Waterman, M. S. (1990). The Erdos-Renyi law in distribution, for coin tossing and sequence matching. *Ann. Statist., 18*, 539-570.

Arratia, R., & Waterman, M. S. (1989). The Erdos-Renyi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab., 17*, 1152-1169.

Arratia, R., & Waterman, M. S. (1994). A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab., 4*, 200-225.

Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Probab., 3*, 534-545.

Clote, P., & Backofen, R. (2000). *Computational Molecular Biology, an Introduction.*: John Wiley and Sons, Ltd. USA.

Davison, A. C. (1998). Extreme Values *Encyclopedia of Biostatistics*: John Wiley & Sons, Ltd.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). A model of evolutionary change in Proteins. In M. Dayhoff (Ed.), *In Atlas of protein sequence and structure* (Vol. 5, Suppl. 3, pp. 345-352): National biomedical research foundation, Silver Spring, MD.

Durbin, R., Eddy, S., Krogh, A., & Mithison, G. (1998). *Biological sequence analysis, probabilistic models of proteins and nucleic acids.*: Cambridge University Press.

Erdos, P., & Renyi, A. (1970). On a new law of large numbers. *J. Anal. Math., 22*, 103-111.

Erdos, P., & Revesz, P. (1975). On the length of the longest head-run. *Topics in Inforrmation Theory. Colloquia Math. Soc. J. Bolyai, 16*, 219-228.

Galas, D. J., Eggert, M., & Waterman, M. S. (1985). Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from Escherichia coli. *J Mol Biol, 186*(1), 117-128.

Gonnet, G. H., Cohen, M. A., & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science, 256*(5062), 1443-1445.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitutions matrices from protein blocks. *Proceedings of the National Academy of Sciences (USA), 89*, 10915-10919.

Karlin, S., & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the USA., 87*, 2264-2268.

Karlin, S., & Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis. *Science, 257*(5066), 39-49.

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol, 305*(3), 567-580.

Lipman, D. J., & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science, 227*(4693), 1435-1441.

Mott, R. (1992). Maximum likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bulletin of Mathematical Biology, 54*, 59-75.

Mott, R. (2000). Accurate formula for P-values of gapped local sequence and profile alignments. *J Mol Biol, 300*(3), 649-659.

Muller, T., Rahmann, S., & Rehmsmeier, M. (2001). Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics, 17 Suppl 1*, S182-189.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol, 48*(3), 443-453.

Ng, P. C., Henikoff, J. G., & Henikoff, S. (2000). PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics, 16*(9), 760-766.

Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Science, 4*, 1145-1160.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J Mol Biol, 276*(1), 71-84.

Pearson, W. R., & Wood, T. C. (2001). Statistical significance in biological sequence comparison. In D. J. Balding, M. Bishop & C. Cannings (Eds.), *In handbook of statistical genetics.* (pp. 39-65): John Wiley and Sons, Ltd. England.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol, 147*(1), 195-197.

Vingron, M., & Waterman, M. S. (1994). Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J Mol Biol, 235*(1), 1-12.

Waterman, M. S. (1995). *Introduction to Computational Biology*: Chapman and Hall, London.

Waterman, M. S., Gordon, L., & Arratia, R. (1987). Phase transitions in sequence matches and nucleic acid structure. *Proceedings of the National Academy of Sciences of the USA., 84*, 1239-1243.

Waterman, M. S., & Vingron, M. (1994). Rapid and accurate estimates of statistical significance for sequence database searches. *Proceedings of the National Academy of Sciences of the USA., 91*, 4625-4628.

Waterman, M. S., & Vingron, M. (1994). Sequence comparison significance and Poisson approximation. *Statistical Science, 2*, 367-381.

Wilbur, W. J., & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences of the USA., 80*, 726-730.

Wootton, J. C., & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers & chemistry, 17*(2), 149-163.

Ερωτήσεις

1) Στον ορισμό της σχετικής εντροπίας

$$H(\alpha, p) \equiv \alpha \log\left(\frac{\alpha}{p}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-p}\right)$$

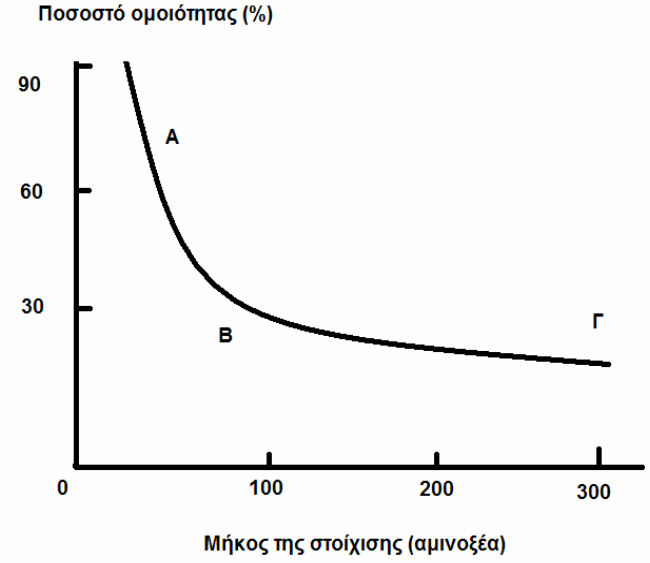
διερευνήστε τι θα συμβεί στην οριακή περίπτωση που το α=1. Τι επιπτώσεις θα έχει αυτή η λύση για τις σχέσεις (3.11) και (3.12);

2) Για τα δεδομένα του πίνακα 3.1, δείξτε ότι ισχύει η ισότητα:

$$E(s_k) = \sum p_k s_k = \sum p_k \log\left(\frac{a_k}{p_k}\right) < 0$$

Τι επιπτώσεις μπορεί να έχει αυτό για την πιθανή χρήση των δεδομένων του πίνακα;

3) Δίνεται η παρακάτω γραφική παράσταση στην οποία αντιστοιχίζονται οι τιμές του ποσοστού ομοιότητας (%) σε συνάρτηση του μήκους μιας κατά ζεύγη στοίχισης δύο πρωτεϊνικών αλληλουχιών.



Α) Τι αναπαριστά η καμπύλη; Ποια είναι η σημασία των δύο περιοχών στις οποίες διαχωρίζει το επίπεδο;
Β) Τι μπορείτε να πείτε για τα σημεία Α, Β και Γ;

4) Δύο αλληλουχίες μήκους 250 αμινοξέων στοιχίζονται με αλγόριθμο τοπικής στοίχισης και τον πίνακα PAM250, και προκύπτει η στοίχιση:

```
F W L E V E G N S M T A P T G
F W L D V Q G D S M T A P A G
```

Υπολογίστε το σκορ της στοίχισης και τη στατιστική σημαντικότητα, αν το K=0.09, και το λ=0.229. Ποιο είναι το bit-score αυτής της στοίχισης;

5) Εκτελούμε τοπική στοίχιση μιας αλληλουχίας Α μήκους 300 αμινοξέων με μια αλληλουχία Β μήκους 550 αμινοξέων. Η στοίχιση που προκύπτει δίνει μια ομοιότητα (similar residues) σε 61 από τα 166 στοιχισμένα κατάλοιπα, ενώ το Bit Score της στοίχισης είναι ίσο με 39.
Α) Ποιο είναι το E-value που προκύπτει από την παραπάνω στοίχιση και πως προκύπτει;
Β) Τι θα συνέβαινε αν το μήκος της στοίχισης ήταν το μισό με αντίστοιχη μείωση του Σκορ; Τι θα συνέβαινε αν το μήκος της στοίχισης ήταν το ίδιο και το Σκορ μειωνόταν στο μισό; Τι θα συνέβαινε αν το μήκος της στοίχισης ήταν το διπλάσιο και το Σκορ παρέμενε ίδιο;

Γ) Ποιο θα ήταν το E-value αν η ίδια στοίχιση είχε προκύψει πραγματοποιώντας αναζήτηση της αλληλουχίας Β έναντι μιας βάσης δεδομένων που περιέχει 500.000 αλληλουχίες με ίδιο μήκος με την Α;

6) Δίνεται τμήμα του αποτελέσματος από την αναζήτηση ομοιότητας με το BLAST μιας πρωτεΐνης έναντι της βάσης δεδομένων NR του NCBI.

Αλληλουχία Α

```
 Score = 34.3 bits (77),  Expect = XXXXX
 Identities = 28/85 (32%), Positives = 44/85 (51%), Gaps = 11/85 (12%)

Query  96   INDWASIYGVVGVGYGKFQTTEYPTY---KHDTSDYGFSYGAGLQ--FNPMENVALDFSY  150
            I++   I+G +G  YG+ +T+  P +     D S +G SYGAG++  FNP     L+  +
Sbjct  118  ISEQFDIFGKLGTTYGRTKTSGNPGFGVATGDDSGFGLSYGAGVRWAFNPQWAAVLE--W  175

Query  151  EQSRIR----SVDVGTWIAGVGYRF  171
            E+ R+       DV     GV YR+
Sbjct  176  ERHRLHFADGKSDVDMTTIGVQYRY  200
```

Αλληλουχία Β

```
 Score = 77.4 bits (189),  Expect = XXXXX
 Identities = 62/201 (30%), Positives = 101/201 (50%), Gaps = 32/201 (15%)

Query  1    MKKIACLSALAAVLAFTAGTSVAAT---STVTGGY--AQSDAQGQMNKMGGFNLKYRYEE  55
            M+K+      AA+    +G    A+   ST++ GY    ++  G  +++ G N+KYRYE
Sbjct  1    MRKLYAAILSAAICLAVSGAPAWASEHQSTLSAGYLHVSTNVPGS-DELNGINVKYRYEF  59

Query  56   DNSPLGVIGSFTY--------TEKSRTASSGDYNKNQYYGITAGPAYRINDWASIYGVVG  107
             ++ LG++ SF+Y        T  S T    D  +N+++ + AGP+ R+N+W S Y + G
Sbjct  60   TDT-LGMVTSFSYAGDKNRQLTHYSDTRWHEDSVRNRWFSVMAGPSVRVNEWFSAYAMAG  118

Query  108  VGYGKFQT--------TEYPTYKHDT---------SDYGFSYGAGLQFNPMENVALDFSY  150
            + Y +  T         T+     HD          S+   ++GAG+Q NP E+VA+D +Y
Sbjct  119  MAYSRVSTFSGDYLRVTDNKGKTHDVLTGSDDGRHSNTSLAWGAGVQVNPTESVAIDIAY  178

Query  151  EQSRIRSVDVGTWIAGVGYRF  171
            E S         +I GVGY+F
Sbjct  179  ECSGSGDWRTDGFIVGVGYKF  199
```

```
Gapped
Lambda     K      H
   0.267   0.0410    0.140
Number of Sequences: 4496249
Length of query: 171
Length of database: 1544746084
Length adjustment: 122
Effective length of query: 49
Effective length of database: 996203706
Effective search space: 48813981594
Effective search space used: 48813981594
```

Α) Υπολογίστε το E-value (Expectation) για τις δυο παραπάνω στοιχίσεις. Ποια συμπεράσματα βγάζετε για τη στατιστική σημαντικότητα των στοιχίσεων αυτών?
Β) Είναι τα συμπεράσματα αυτά σύμφωνα με τους εμπειρικούς κανόνες για την ομοιότητα δυο αλληλουχιών?
Γ) Τι θα συνέβαινε αν οι δύο παραπάνω στοιχίσεις είχαν προκύψει σε μία κατά ζεύγη στοίχιση και όχι σε μια αναζήτηση στη βάση δεδομένων;