ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ ΙΙ

Κατα ζέυγη στοίχιση και στατιστική σημαντικότητα αυτής

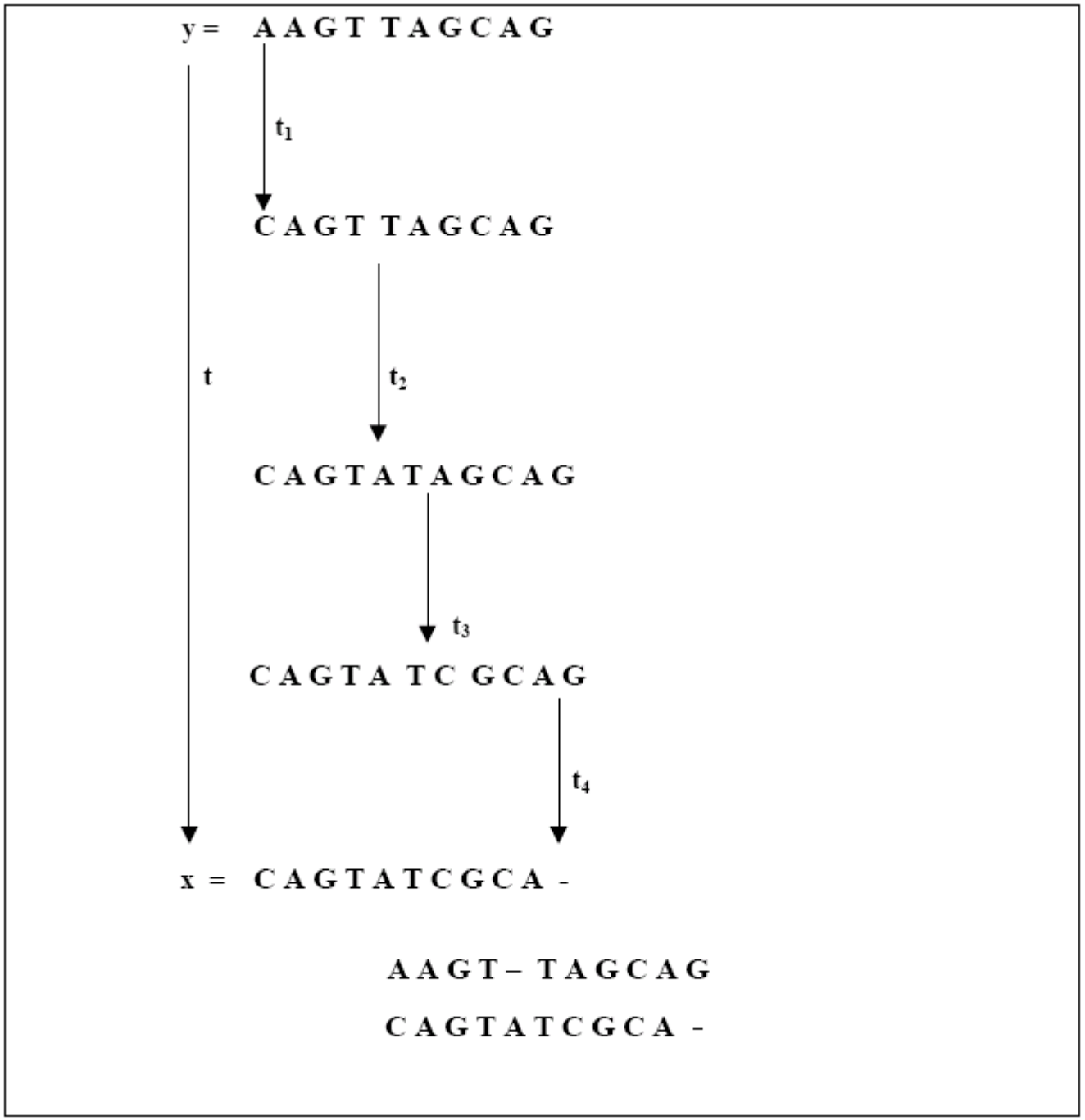
Παντελής Μπάγκος

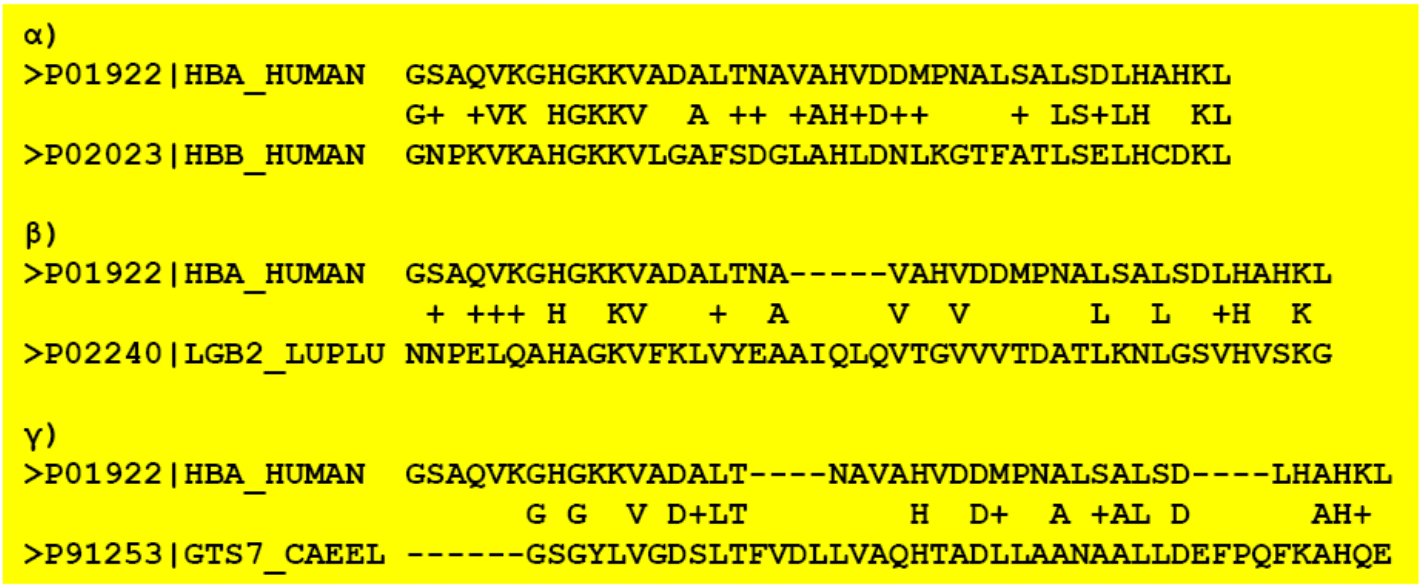
Διάλεξη 2

Αναζήτηση ομοιότητας και κατά ζεύγη στοίχιση ακολουθιών

Κατά ζεύγη στοίχιση ακολουθιών

- Από τα πιο σημαντικά προβλήματα στην Υπολογιστική Βιολογία
- Ιδιαίτερα πλούσια βιβλιογραφία για πάνω από 30 χρόνια
- Η ομοιότητα δυο ακολουθιών αντανακλά κατά βάση την κοινή εξελικτική προέλευση


y = A A G T T A G C A G
t1
C A G T T A G C A G
t
t2
C A G T A T A G C A G
t3
C A G T A T C G C A G
t4
x = C A G T A T C G C A -
A A G T – T A G C A G
C A G T A T C G C A -



Εικόνα 3: Τρεις στοιχίσεις ακολουθιών με τον αλγόριθμο Needleman-Wunsch με ένα τμήμα της άλφα αλυσίδας της ανθρώπινης αιμοσφαιρίνης (SwissProt AC P01922).

α) Ξεκάθαρη ομοιότητα με τη βήτα αλυσίδα της ανθρώπινης αιμοσφαιρίνης (AC P02023).

β) Δομικά συμβατή στοίχιση με την leghemoglobin II (AC P02240) του δικοτυλίδονου Lupinus luteus.

γ) 'Παραπλανητική' στοίχιση με ομόλογη της S-τρανφεράσης της γλουταθειόνης (AC P91253) του νηματώδη σκώληκα C. elegans.

Σημαντικά ζητήματα στη στοίχιση ακολουθιών

- Το είδος των στοιχίσεων που μας ενδιαφέρουν
- Το σύστημα βαθμονόμισης (scoring system)
- Ο αλγόριθμος που θα χρησιμοποιήσουμε για την έυρεση της καλής ή και της βέλτιστης στοίχισης
- Ο τρόπος προσδιορισμού της στατιστικής σημαντικότητας μιας στοίχισης

Παράδειγμα

- Έστω 2 ακολουθίες **x,y** (ίδιου ή διαφορετικού μήκους)

  $\mathbf{x}=x_1,x_2,\ldots,x_n$

  $\mathbf{y}=y_1,y_2,\ldots,y_m$
- Μας ενδιαφέρει η εύρεση της μέγιστης κοινής περιοχής τους (πλήρης ταύτιση)
- Η απλή απαρίθμιση όλων των πιθανών κοινών υπό-περιοχών είναι απαγορευτική:

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

- Χρειαζόμαστε έναν πιο αποδοτικό αλγόριθμο (δυναμικός προγραμματισμός)

Score

Θεωρούμε δυο πιθανότητες: την πιθανότητα ανεξάρτητης (τυχαίας) ταύτισης, και αυτή της μη τυχαίας

$$P(\mathbf{x},\mathbf{y} \mid R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

$$P(\mathbf{x},\mathbf{y} \mid M) = \prod_i p_{x_i y_i}$$

Αν πάρουμε το λόγο των δυο πιθανοφανειών (likelihood ratio):

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

Και αν δουλέψουμε σε λογαριθμική κλίμακα:

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

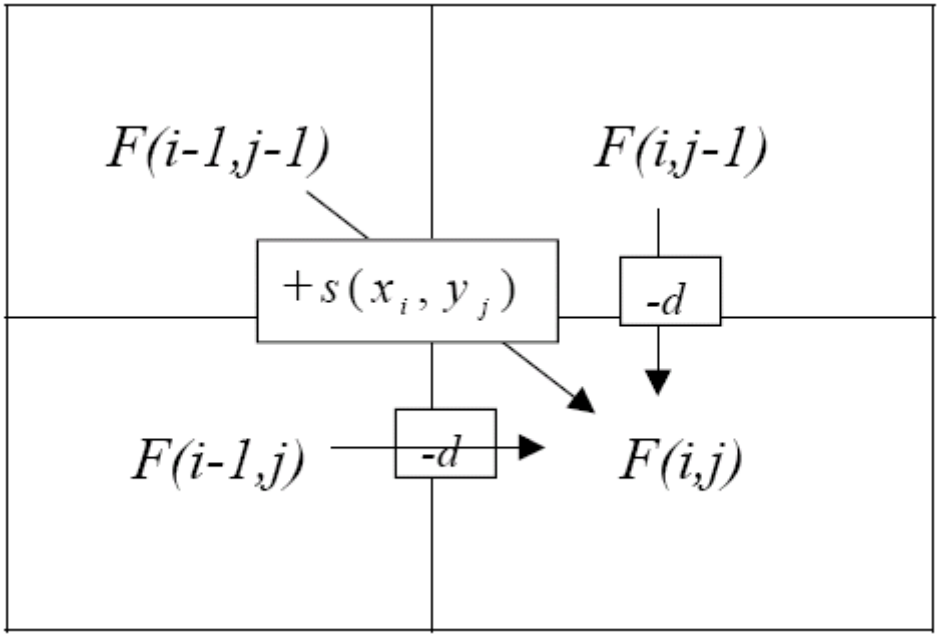
Πίνακες ομοιότητας

Μπορούμε έτσι να ορίσουμε έναν πίνακα ομοιότητας με διαστάσεις όσο το μέγεθος του αλφαβήτου (4x4 για DNA, 20x20 για πρωτεϊνες), π.χ.:

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

Για τη μη-ταύτιση (mismatch), μπορούμε να ορίσουμε μια πολύ μεγάλη ποινή (-∞) έτσι ώστε να απαγορεύουμε πρακτικά την ταύτιση μη ομοίων καταλοίπων

Δυναμικός προγραμματισμός



Ποινές για τα κενά (gap penalties)

Απλή ποινή για τα κενά:

$$\gamma(g) = -gd$$

Σύνθετη ποινή για τα κενά:

$$\gamma(g) = -d - (g-1)e$$

Ολική στοίχιση (Needleman and Wunsch, 1970)

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

$F(i,0) = -id,$
$F(0,j) = -jd$

Παράδειγμα

Έστω δυο ακολουθίες:

$$\mathbf{x} = AAGTTAGCAG$$

$$\mathbf{y} = CAGTATCGCA$$

Αν έχουμε για τα κενά:

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

d=1

Τότε η καλύτερη ολική στοίχιση θα είναι:

A A G T – T A G C A G

C A G T A T C G C A –

συνέχεια…

	-	*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

A A G T – T A G C A G

C A G T A T C G C A –

Τοπική στοίχιση (Smith and Waterman, 1981)

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

$F(i,0)=0,$
$F(0,j)=0$

Η μόνη διαφορά από την ολική στοίχιση είναι το 0 το οποίο εξασφαλίζει διακοπή της στοίχισης όταν το score γίνει αρνητικό

Παράδειγμα

Στα δεδομένα του προηγούμενου παραδείγματος, θα έχουμε:

	-	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

A G T – T A G C A

A G T A T C G C A

Αλγοριθμική πολυπλοκότητα

Πρέπει εδώ να τονίσουμε ότι ο απαιτούμενος χρόνος για να τρέξουν οι παραπάνω αλγόριθμοι δυναμικού προγραμματισμού είναι ανάλογος του γινόμενου των μήκων των ακολουθιών και συμβολίζεται *O(mn)*. Το σύμβολο *O(mn) (big-O notation)* σημαίνει ότι μια συνάρτηση *f(t)=O(nm)* αν καθώς $t \to \infty$ υπάρχει σταθερά c τέτοια ώστε ,

$$|f(t)| \le c.n.m$$

Σύνθετες ποινές για τα κενά

- Απαιτείται μια συνάρτηση $\gamma()$
- Τότε, οι παραπάνω αλγόριθμοι γίνονται:

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

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

Μειονέκτημα

- Η αλγοριθμική πολυπλοκότητα αυξάνει σε $O(n^3)$
- Ο Gotoh (1982), έδειξε ότι για σύνθετες συναρτήσεις του τύπου:

$$\gamma(g) = -d - (g-1)e$$

Μπορούμε να έχουμε πολυπλοκότητα της τάξης του $O(n^2)$ μόνο με αύξηση της μνήμης

Άλλοι αλγόριθμοι

- Υπάρχουν επίσης ειδικές περιπτώσεις στοίχισης (π.χ. προσαρμογή)
- Θέλουμε δηλαδή να εντοπίσουμε, μια μικρή ακολουθία αν συναντάται σε μια μεγαλύτερη

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:

$\mathbf{x} = TCGCGGTATGGCATGATAGCGCCCGGAA$

και η αλληλουχία του υποκινητή είναι

$\mathbf{y} = TATAAT$

συνέχεια…

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

$F(i,0)=-id$

$F(0,j)=0$

	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

Και η ακολουθία του πιθανού υποκινητή είναι:

C A T G A T

Ευριστικοί αλγόριθμοι στοίχισης (Heuristic alignment algorithms)

- Είναι αναγκαίοι για τη μείωση του απαιτούμενου υπολογιστικού χρόνου, ειδικά σε αναζητήσεις σε βάσεις δεδομένων

Απαραίτητα χαρακτηριστικά τους:

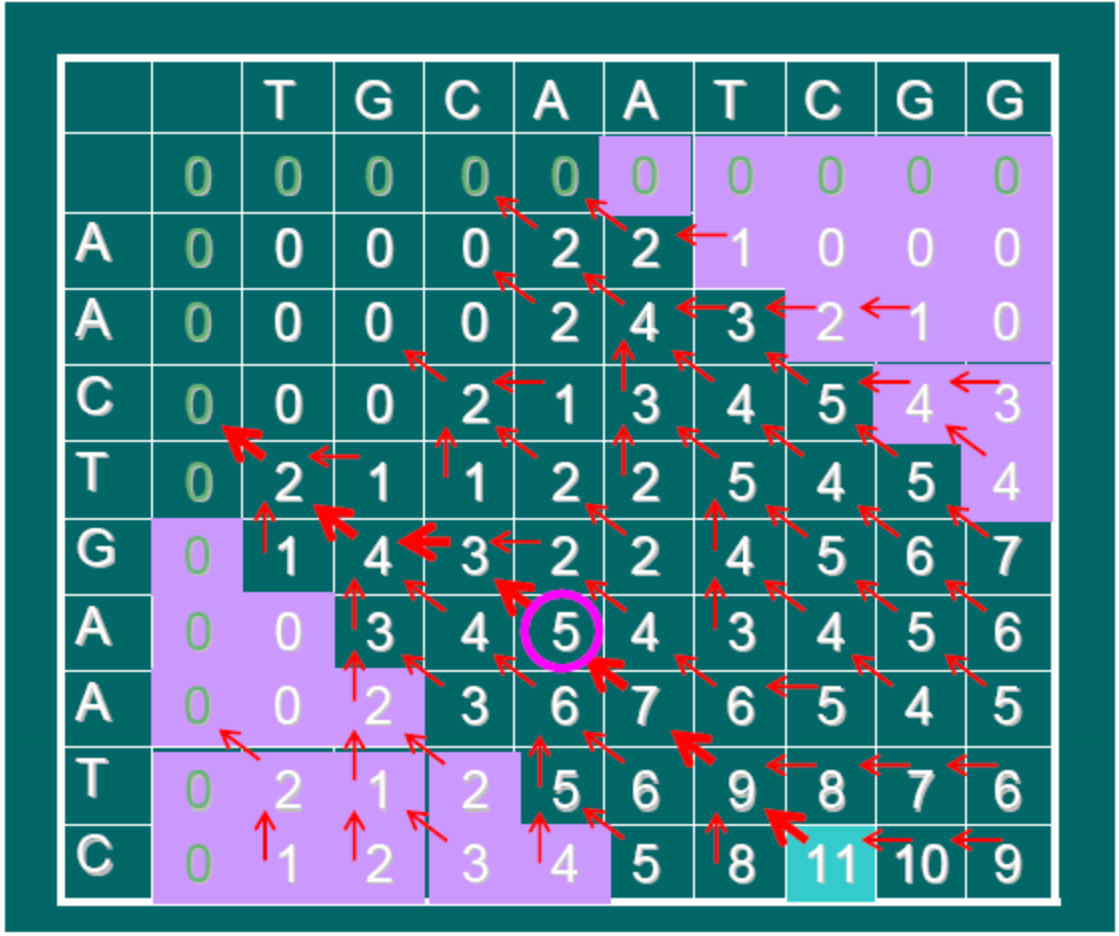
- Να μη διαφέρουν σημαντικά από τις «ακριβείς» (μαθηματικά βέλτιστες) λύσεις των μεθόδων δυναμικού προγραμματισμού.
- Να μην αποκλείουν βιολογικά πιθανές λύσεις.

Βασικές κατηγορίες τέτοιων αλγορίθμων:

- Μέθοδος «κοπής γωνιών» (banded alignment)
- Μέθοδος FASTA
- Μέθοδος BLAST

Μέθοδος «κοπής γωνιών»

- Αυτή είναι ίσως η απλούστερη «βελτίωση» που θα μπορούσε να σκεφτεί κανείς. Η ιδέα είναι πραγματικά πολύ έξυπνη και απλή, περιορίζοντας στην ουσία τους υπολογισμούς των πινάκων Δυναμικού Προγραμματισμού σε μια «ζώνη» γύρω από τη διαγώνιο του πίνακα. Όπως γίνεται εμφανές, η επιλογή του πλάτους της ζώνης στην οποία θα εκτελεστούν οι υπολογισμοί επηρεάζει άμεσα την εξοικονόμηση πόρων κατά τη στοίχιση ακολουθιών.
- Μπορεί να δώσει μια «οικονομία» υπολογιστικών πόρων της τάξης του 30%.
- Σε ακραίες περιπτώσεις

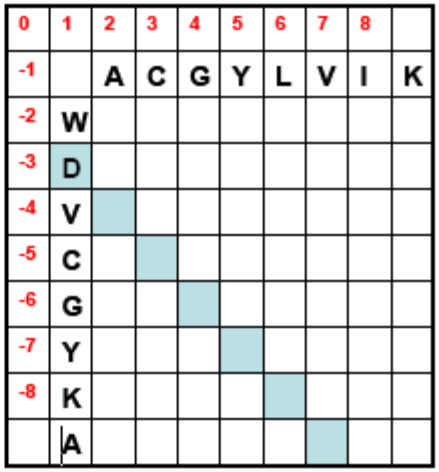


*Εικόνα 3: Πίνακας Δυναμικού Προγραμματισμού για τη μέθοδο «Κοπής Γονιών». Οι τιμές όλων των κελιών έχουν τοποθετηθεί στα κελιά (δείτε σημειώσεις προηγούμενης διάλεξης). Με τη μέθοδο αυτή, υποθέτουμε ότι ένα «καλό μονοπάτι» (δηλ. μια καλή στοίχιση) δεν αναμένουμε να διέρχεται από τις σκιασμένες περιοχές του πίνακα (πάνω δεξιά και κάτω αριστερή γωνία). Με τα παχιά βέλη υποδηλώνεται η βέλτιστη διαδρομή, όπως υπολογίζεται με τον κλασικό Δυναμικό Προγραμματισμό. Παρατηρήστε ότι από τα 100 (=10*10) κελιά του πίνακα απαιτείται το γέμισμα μόνο των 70, κερδίζοντας έτσι 30% σε μνήμη (και χρόνο).*

Μέθοδος FASTA

- Η βασική ιδέα έγκειται στη δημιουργία ενός ευρετηρίου με τις θέσεις όλων των k-tuples (τυπικό μήκος για αμινοξικές ακολουθίες 1 ή 2) που υπάρχουν και στις δύο ακολουθίες (Εικόνα 4, αριστερά).
- Από τη διαφορά των θέσεών τους στις δύο ακολουθίες εντοπίζεται η διαγώνιος στην οποία βρίσκονται (Εικόνα 4, δεξιά), 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 (Εικόνα 5), και
- Τελικά πραγματοποιείται η διαδικασία πλήρους δυναμικού προγραμματισμού (με τον επιλεγμένο πίνακα αντικατάστασης), περιορισμένου σε μια ταινία γύρω από τις συγκεκριμένες διαγωνίους (Εικόνα 5).

K-tuple (K=1)	position in SEQ1	position in SEQ2	offset SEQ1-SEQ2
A	1	8	-7
C	2	4	-2
D	-	2	X
G	3	5	-2
I	7	-	X
K	8	7	1
L	5	-	X
V	3	6	-3
W	-	1	X
Y	4	6	-2

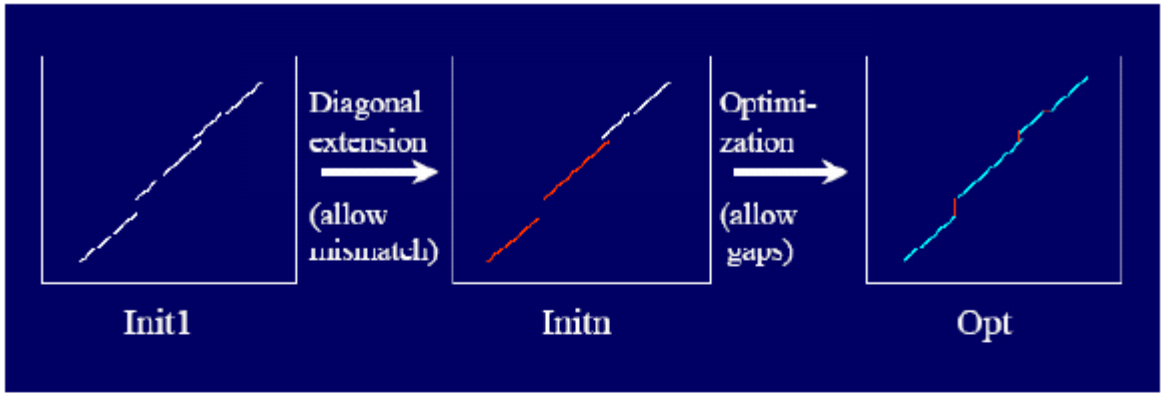


Εικόνα 4: Αριστερά - Κατασκευή ευρετηρίου για k-tuples, με k=1 για τις ακολουθίες SEQ1:ACGYLVIK και SEQ2:WDVCGYKA. Για απλότητα οι ακολουθίες του παραδείγματος περιέχουν από μία φορά κάθε k-tuple. Η διαφορά της θέσης μιας k-tuple στη μία ακολουθία με μιά ταυτόσημή της στην άλλη εκφράζει ένα μέτρο «μετατόπισης» της μιας ακολουθίας ως προς την άλλη για να στοιχιθούν μεταξύ τους τα συγκεκριμένα k-tuples.

Δεξιά - Όλες οι διαγώνιες του πίνακα προσδιορίζονται συμβολικά σε σχέση με την κυρία διαγώνιο με τους κόκκινους αριθμούς οι οποίοι φαίνονται στην 1η γραμμή και πρώτη στήλη αντίστοιχα. Οι αριθμοί αυτοί αντιστοιχούν στην κοινή διαφορά (offset) που έχουν οι δείκτες i, j οι οποίοι ορίζουν τα στοιχεία της συγκεκριμένης διαγωνίου. Προφανώς, η κύρια διαγώνιος (για κάθε κελί της οποίας ισχύει i=j) αντιστοιχεί σε offset 0.

Στην πράξη οι 10 καλύτερες διαγώνιες εντοπίζονται με αυτόν τον τρόπο.

Σημείωση: Ο πίνακας του σχήματος αντιστοιχεί με Dot Matrix Plot, εύκολα μπορεί η ιδέα να αποτυπωθεί σε ένα πίνακα δυναμικού προγραμματισμού.



Εικόνα 5: Σχηματική αναπαράσταση των σταδίων για τη σύκριση ακολουθιών με τη μέθοδο FASTA. Μόνο οι «καλύτερες» διαγώνιοι (όπως προέκυψαν μετά τη δημιουργία ευρετηριών) καθορίζουν την περιοχή στην οποία θα υπολογιστεί τελικά η στοίχιση.

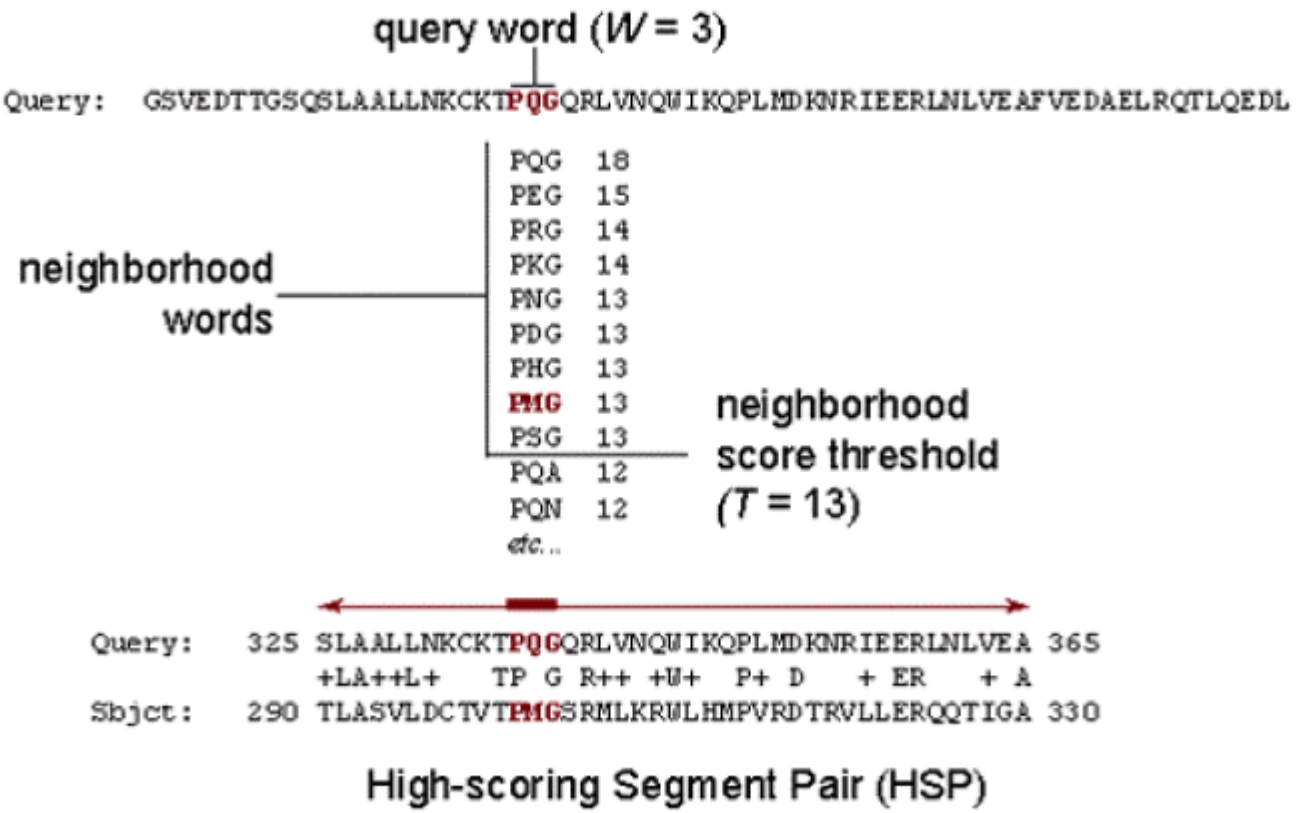
Στο πρώτο στάδιο ενοποιούνται περιοχές της ίδιας διαγωνίου επιτρέποντας την στοίχιση και ανόμοιων καταλοίπων (mismatch) αλλά ΟΧΙ την εισαγωγή κενών. Οι περιοχές που φαίνονται στο αριστερό διάγραμμα ονομάζονται Βέλτιστες Αρχικές Περιοχές (Best Initial Regions) και επιλέγονται με κριτήριο να έχουν βαθμολογία για τη στοίχισή τους (με τη χρήση μόνο ενός πίνακα αντικατάστασης) μεγαλύτερη από μια αρχική τιμή κατωφλίου Init1. Σε παραλλαγές της μεθόδου μπορεί να επιλεχθεί να διαλέγουμε συγκεκριμένο πλήθος από τις τοπικές στοιχίσεις με τα μεγαλύτερα scores ανεξάρτητα από το εάν αυτά ξεπερνούν την τιμή Init1.

Στο επόμενο στάδιο ενοποιούνται περιοχές οι οποίες δεν ανήκουν υποχρεωτικά στην ίδια διαγώνιο. Αυτό προφανώς επιβάλει την εισαγωγή κενών σε κάποια από τις ακολουθίες (πιθανότατα και στις 2). Για την εισαγωγή κενών αφαιρείται μια ποινή για την εισαγωγή κάθε κενού. Τώρα πλέον οι στοιχίσεις επεκτείνονται όσο η τιμή του score παραμένει μεγαλύτερη από μια δεύτερη τιμή κατωφλίου InitN.

Μέθοδος BLAST

- Η διαδικασία της σύγκρισης ξεκινά με την κατασκευή ενός καταλόγου όλων των λέξεων που θα ταίριαζαν με κάποια λέξη της άγνωστης ακολουθίας ξεπερνώντας την τιμή κατωφλίου (προκαθορισμένη τιμή για πρωτεϊνικές ακολουθίες T=13).
- Στη συνέχεια, ο αλγόριθμος αναζητά αυτές τις λέξεις στις ακολουθίες της βάσης δεδομένων και κάθε φορά που εντοπίζει κάποια 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 περιγράφεται στην επόμενη ενότητα.

The BLAST Search Algorithm



The BLAST algorithm. The BLAST algorithm is a heuristic search method that seeks words of length W (default = 3 in blastp) that score at least T when aligned with the query and scored with a substitution matrix. Words in the database that score T or greater are extended in both directions in an attempt to fina a locally optimal ungapped alignment or HSP (high scoring pair) with a score of at least S or an E value lower than the specified threshold. HSPs that meet these criteria will be reported by BLAST, provided they do not exceed the cutoff value specified for number of

Πίνακες αντικατάστασης (substitution matrices)

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

- 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)
- λ είναι μια σταθερά κανονικοποίησης

Εντροπία των πινάκων αντικατάστασης

$$H = \sum_{i=1}^{20} \sum_{j=1}^{20} q_{ij} S_{ij}$$

- 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.

Διάφοροι πίνακες αντικατάστασης

- PAM
- BLOSUM

PAM

- **P**oint **A**ccepted **M**utations (Dayhoff et al)
- Ως Αποδεκτή Σημειακή Μεταλλαγή σε μια πρωτεΐνη θεωρείται η αντικατάσταση ενός αμινοξικού καταλοίπου της με ένα κατάλοιπο διαφορετικού τύπου, η οποία έχει γίνει αποδεκτή μέσω της διαδικασίας της Φυσικής Επιλογής.
- Προέκυψε από πολλαπλή στοιχιση ακολουθιών με γνωστή εξελικτική σχέση και επίπεδο ομοιότητας >85%
- PAM1, PAM30, PAM250 κλπ
- Προυποθέτει ένα Μαρκοβιανό μοντέλο εξέλιξης
- Η χρήση πινάκων με μικρό N ενδείκνυται όταν οι εξεταζόμενες ακολουθίες είναι πολύ όμοιες (μικρή εξελικτική απόσταση), ενώ στην περίπτωση περισσότερο απομακρυσμένων ομοιοτήτων χρησιμοποιούμε πίνακες μεγαλύτερου N. Στις περιπτώσεις εκείνες κατά τις οποίες δε γνωρίζουμε εκ των προτέρων την ομοιότητα των προς σύγκριση ακολουθιών (π.χ. σε αναζητήσεις έναντι βάσεων δεδομένων) επιλέγουμε ένα ενδιάμεσο πίνακα, όπως τον PAM-250, ο οποίος αντιστοιχεί σε συντήρηση της τάξης του 20-25%.

BLOSUM

- **BLO**cks **SU**bstistution **M**atrcices (Henikoff and Henikoff)
- Προέκυψαν από πολλαπλές στοιχίσεις ακολουθιών με γνωστή κάθε φορά εξελικτική σχέση και διαφορετικό επίπεδο ομοιότητας
- Δεν προυποθέτουν ένα εξελικτικό μοντέλο αλλά το προσεγγίζουν εμπειρικά
- BLOSUM50, BLOSUM62, κλπ

PAM-1

PAM-250

BLOSUM100

BLOSUM30

Small evolutionary distance
Strong similarity for short sequence

Large evolutionary distance
Weak similarity over stretched length

Στατιστική σημαντικότητα των στοιχίσεων

- Αν λαβουμε με οποιοδήποτε τρόπο μια στοίχιση δυο ακολουθιών, θέλουμε να έχουμε έναν τρόπο να την αξιολογήσουμε (να ξέρουμε δηλαδή αν είναι στατιστικά σημαντική)
- Ιδιαίτερο νόημα έχει αυτό σε μια αναζήτηση σε μεγάλες βάσεις δεοδομένων όπου αναμένουμε να δούμε έως και εκατοντάδες «ομόλογες» ακολουθίες
- Χρειαζόμαστε έναν έλεγχο υποθέσεων.

  Ho: οι δυο ακολουθίες είναι ασυσχέτιστες,

  Ha: οι δυο ακολουθίες σχετίζονται με κάποιο τρόπο (είναι ομόλογες)
- Ακόμα και αν βρεθεί στατιστικά σημαντική ομοιότητα, δεν σημαίνει ότι υπάρχει και βιολογική συσχέτιση των ακολουθιών, και το αντίστροφο (εξαρτάται από τις παραμέτρους, gap penalty, substitution matrix, αλγοριθμο στοίχισης κλπ)
- Τα πιο πολλά αποτελέσματα αναφέρονται στην **τοπική στοίχιση**

Ασυμπτωτικά αποτελέσματα

Θεώρημα (Waterman, 1995)

Έστω ότι έχουμε δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, ..., x_n$ *και* $\mathbf{y} = y_1, y_2, ..., y_m$. *Τότε η μέγιστη περιοχή σύμπτωσης (match) μεταξύ τους είναι* $M_n \cong \log_{1/p}(mn)$ *ή αλλιώς:*

$$\frac{M_n}{\log_{1/p}(mn)} \to 1 \text{ με πιθανότητα } 1.$$

Προφανώς η πιθανότητα σύμπτωσης 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 βάσεων στις δυο αλληλουχίες είναι ίδια.

Θεώρημα (Waterman, 1995)

Έστω δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, ..., x_n$ *και* $\mathbf{y} = y_1, y_2, ..., y_m$ *με* $0 \le p < a \le 1$. *Τότε για τη μέγιστη περιοχή που περιέχει 100α% όμοια νουκλεοτίδια μεταξύ τους ισχύει*

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

Θεώρημα (Arratia et al, 1990)

Έστω δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, ..., x_n$ *και* $\mathbf{y} = y_1, y_2, ..., y_m$. *Τότε η μέση τιμή για το μήκος της μέγιστης περιοχής σύμπτωσης (match) μεταξύ τους είναι:*

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

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

Για την αντίστοιχη διασπορά ισχύει :

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

Θεώρημα (Arratia and Waterman, 1989; Waterman, 1995)

Έστω ότι έχουμε δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, ..., x_n$ *και* $\mathbf{y} = y_1, y_2, ..., y_m$. *Τότε η μέση τιμή για το μήκος της μέγιστης περιοχής σύμπτωσης (match) μεταξύ τους, όταν υπάρχουν k μη κοινά νουκλεοτίδια (k mismatches) είναι:*

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

Για την αντίστοιχη διασπορά ισχύει :

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

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

Η κατανομή του Local Similarity Score

- Σε όλες τις τοπικές στοιχίσεις χωρίς κενά, η κατανομή του score είναι η κατανομή των ακραίων τιμών του Gumbel
- Αν υπάρχουν κενά, η κατανομή φαίνεται να συγκλίνει (υπο προυποθέσεις) σε αυτή του Gumbel χωρίς όμως αυτό να μπορεί να αποδειχθεί
- Σε ολικές στοιχίσεις δεν ισχύει τίποτα από τα παραπάνω

Η κατανομή του Local Similarity Score

Δυο ακραίες περιπτώσεις:

$$s(x_i, y_i) = \begin{cases} 1, \text{αν} & x_i = y_i \\ 0, \text{αν} & x_i \neq y_i \end{cases} \text{ και d=0} \qquad s(x_i, y_j) \sim c.n$$

Γραμμική περιοχή

$$s(x_i, y_i) = \begin{cases} 1, & \text{αν} \quad x_i = y_i \\ -\infty, \text{αν} & x_i \neq y_i \end{cases} \text{ και d=}\infty \qquad s(x_i, y_j) \sim k.\log n$$

Λογαριθμική περιοχή

Στη δεύτερη περίπτωση η κατανομή είναι αποδεδειγμένα αυτή του Gumbel, αλλά όταν μπαίνουν κενά δεν υπάρχει τέτοια απόδειξη

Μειώνοντας σταδιακά τις ποινές για διαφορές και κενά, μεταπίπτουμε από τη λογαριθμική περιοχή του score στη γραμμική. Αυτή η μετάπτωση φάσεως (phase transition) έχει περιγραφεί αναλυτικά από τους Arratia, Gordon και Waterman (Waterman et al, 1987; Arratia and Waterman, 1994; Waterman, 1995) αλλά παρ' όλα αυτά δεν υπάρχει αναλυτική έκφραση για τις τιμές των παραμέτρων m (mismatch) και d (gap) στις οποίες συμβαίνει αυτή η μετάπτωση (μπορούν να προσεγγισθούν μόνο με αριθμητικές μεθόδους)

Η κατανομή του Local Similarity Score

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

Θεώρημα **(Karlin and Altschul, 1990)**

Έστω ότι έχουμε δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, ..., x_n$ *και* $\mathbf{y} = y_1, y_2, ..., y_m$ *και το score S*

Τότε: $P\{S > x\} \approx 1 - \exp\{-Kmne^{-\lambda x}\}$

Τουλάχιστον ένα score πρέπει να είναι θετικό

Η αναμενόμενη τιμή του score για κάθε βάση να είναι αρνητική, δηλαδή

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

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 score θα παίρνει τιμές στη λογαριθμική περιοχή, και κατά συνέπεια θα είναι όντως τοπικό. Αν δεν ισχύουν οι παραπάνω προϋποθέσεις, τότε το score θα παίρνει τιμές στη γραμμική περιοχή και κατά συνέπεια θα μιλάμε για ολική στοίχιση.

Η πιθανότητα να υπάρχει μια κοινή υπο-ακολουθία με μήκος μεγαλύτερο από x , όπως είπαμε παραπάνω είναι (ασυμπτωτικά):

$$P\left(S > x = \log_{1/p}(mn) + T\right) = 1 - e^{-E(S)} = 1 - \exp\left(-Kmne^{-\lambda x}\right)$$

οπότε $P(S \le x) = \exp(-Kmne^{-\lambda x})$

Η τελευταία σχέση είναι η α.σ.κ. της κατανομής των ακραίων τιμών του Gumbel (EVD).

$$P(S \le x) = \exp(-e^{-\frac{(x-a)}{b}}), -\infty \le x \le \infty$$

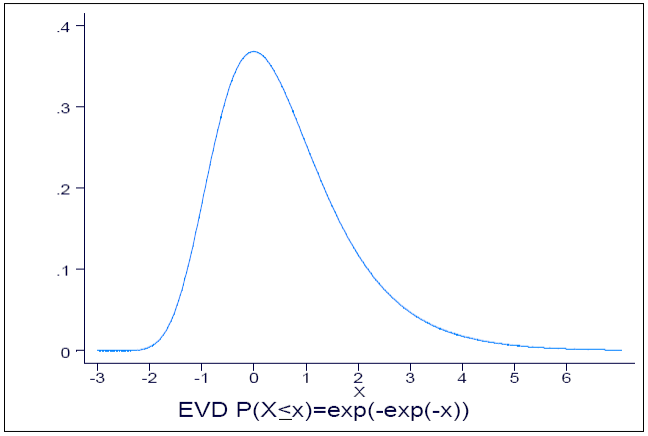
με

$$E(x) = a - b\Gamma'(1) \text{ και } V(x) = \frac{b^2\pi^2}{6}.$$

Οι παράμετροι a, b είναι προφανώς $a = \frac{\log(kmn)}{\lambda}, b = \frac{1}{\lambda}$ με $\lambda = \log\left(1/p\right)$ και K=1-p=q,

όταν δεν επιτρέπονται διαφορές. Από τις παραπάνω σχέσεις είναι δυνατόν να υπολογιστεί το p-value για ένα δεδομένο score που προέκυψε από την σύγκριση δυο ακολουθιών. Αφού τυποποιήσουμε τη μεταβλητή μας έχουμε (Pearson, 1998; Pearson and Wood, 2001):

$$P(Z \ge z) = 1 - \exp\left(-\exp-\left(\left(\pi/\sqrt{6}\right)z - \Gamma'(1)\right)\right)$$



Όταν συγκρίνουμε μια ακολουθία με μια ολόκληρη βάση δεδομένων, η οποία περιέχει D ακολουθίες, 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 (Pearson and Wood, 2001):

$$P = \Pr(\text{τουλαχιστον 1 score } S \geq x) = 1 - e^{-Dp}$$

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 score $S \geq x$, έπειτα από D συγκρίσεις με τις ακολουθίες της βάσης δεδομένων. Αυτό το E-value (expectation value) είναι ίσο με $E(S \geq x) = D.P(S \geq x)$ όπου D είναι ο αριθμός των ανεξάρτητων ακολουθιών 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 ακολουθίες στη βάση δεδομένων δεν έχουν τον ίδιο αριθμό βάσεων. Πρακτικά αυτό σημαίνει ότι θεωρούμε ολόκληρη τη βάση δεδομένων ως μια τεράστια ακολουθία από N νουκλεοτίδια (βάσεις) και συγκρίνουμε με αυτήν τη συγκεκριμένη ακολουθία μας η οποία έχει μήκος n βάσεις. Κατά μέσο όρο κάθε μια από τις ακολουθίες της βάσης περιέχει m=N/D βάσεις, οπότε η πιθανότητα να υπάρχει μια περιοχή με score 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 et al, 1990), αντί του p-value, αναφέρουν ως αποτέλεσμα (output) αυτή 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 (Waterman, 1995):

$$1 - \exp(-\exp(-t)) \approx 1 - (1 - \exp(-t)) = \exp(-t),$$

το p-value θα είναι περίπου ίσο με το E-value. Είναι φανερό ότι σήμερα που οι βάσεις δεδομένων αυξάνονται σε μέγεθος συνεχώς είναι καλύτερο κάθε φορά που γίνονται τέτοιες συγκρίσεις να αναφέρονται τουλάχιστον μαζί το p-value και το e-value, και τέτοια παραδείγματα θα δούμε σε παρακάτω κεφάλαια.

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

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

όπου S_{raw}, είναι το score που υπολογίστηκε με κάποιες συγκεκριμένες τιμές για κενά και διαφορές. Αντικαθιστώντας τώρα στην σχέση (4.15) θα έχουμε

$$E(S_{bit}) = m.n.2^{-S_{bit}} .$$

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

δηλαδή το λειτουργικό μήκος της ακολουθίας και της βάσης δεδομένων προσαρμόζεται (μειώνεται), για να λάβει υπόψη το γεγονός ότι με αυτά τα μήκη και τον δεδομένο πίνακα (substitution matrix) δεν επιτρέπονται όλες οι στοιχίσεις. Το H είναι η σχετική εντροπία του πίνακα για τη δεδομένη σύσταση και το μήκος των ακολουθιών που συγκρίνονται.

Η κατανομή όταν υπάρχουν κενά

- Η μέθοδος του Mott (1992)
- Η μέθοδος Direct Estimation (Waterman, 1995)
- Η μέθοδος Poisson declumping (Waterman and Vingron, 1994)
- Η μέθοδος weighted regression του Pearson (1995)

Η μέθοδος του Mott (1992)

- Παραλλαγή της εκτίμησης στην κατανομή του Gumbel

$$P(S \le x) = \exp(-e^{\frac{(x-A)}{B}})$$

όπου :

$$A = a_0 + \frac{a_1}{\lambda} + \frac{a_2 \log(mn)}{\lambda}, \quad B = \frac{b_1}{\lambda}.$$

Το λ είναι και πάλι η μοναδική θετική ρίζα της εξίσωσης:

$$\sum q_i q_j e^{\lambda S} = 1$$

Η μέθοδος Direct Estimation (Waterman, 1995)

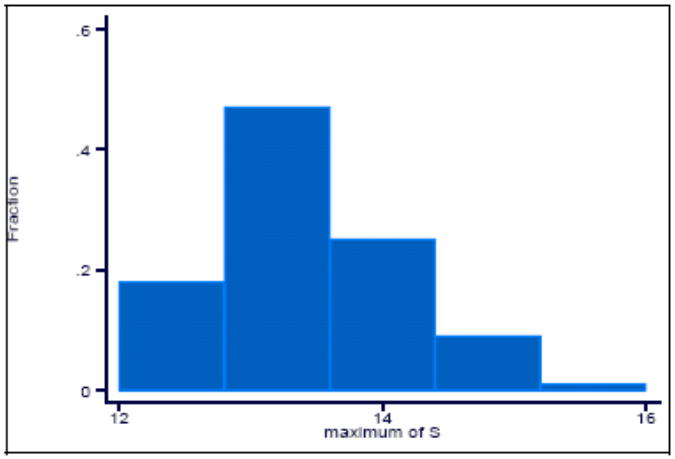
- Πραγματοποιεί Maximum Likelihood fit, σε εμπειρικά δεδομένα
- Απαιτεί αποτελέσματα από πολλές αναζητήσεις
- Απλή στην εκτέλεση (linear regression)

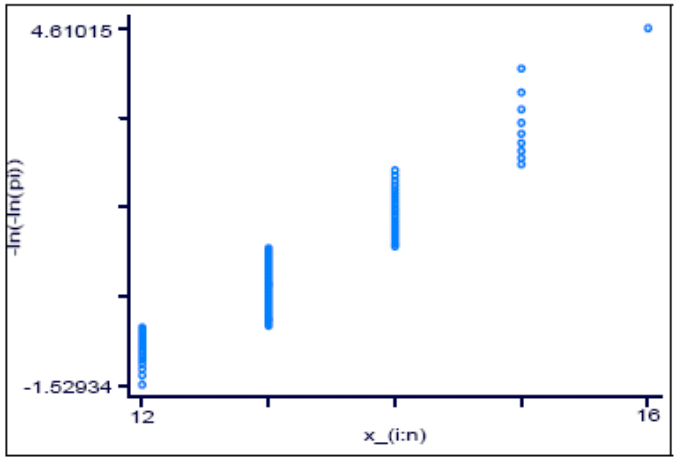
$$P(S \le x) = \exp(-Kmne^{-\lambda x})$$

$$\log P(S \le x) = -Kmne^{-\lambda x} \Leftrightarrow$$

$$\log\left(-\log P(S \le x)\right) = \log\left(Kmne^{-\lambda x}\right) \Leftrightarrow$$

$$\log\left(-\log P(S \le x)\right) = -\lambda x + \log\left(Kmn\right)$$





Παραλλαγές

- Η αναζήτηση μπορεί να γίνει σε τυχαίες ακολουθίες με προκαθορισμένη σύνθεση
- Η αναζήτηση μπορεί να γίνει σε shuffled ακολουθίες με σύνθεση όμοια με αυτή της ακολουθίας εισόδου
- Αν πρόκειται για αναζήτηση σε βάση δεδομένων μπορεί να χρησιμοποιηθούν τα αποτελέσματα της αναζήτησης (αφου απομακρυνθούν οι πολύ όμοιες και οι πολύ ανόμοιες ακολουθίες)
- Χρειάζονται το λιγότερο 100-1000 ακολουθίες, άρα είναι χρονοβόρα διαδικασία

Η μέθοδος Poisson declumping (Waterman and Vingron, 1994)

- Παραλλαγή της προηγούμενης μεθόδου
- Πολύ πιο αποδοτική και γρήγορη
- Στηρίζεται στην προσέγγιση Poisson declumping
- Για κάθε ακολουθία χρησιμοποιεί το διατεταγμένο δείγμα:

$$S_{(1)} \geq S_{(2)} \geq \ldots \geq S_{(k)}$$

και όχι μόνο το μέγιστο

- Τα score από κάθε ακολουθία ακολουθούν κατανομή Poisson:

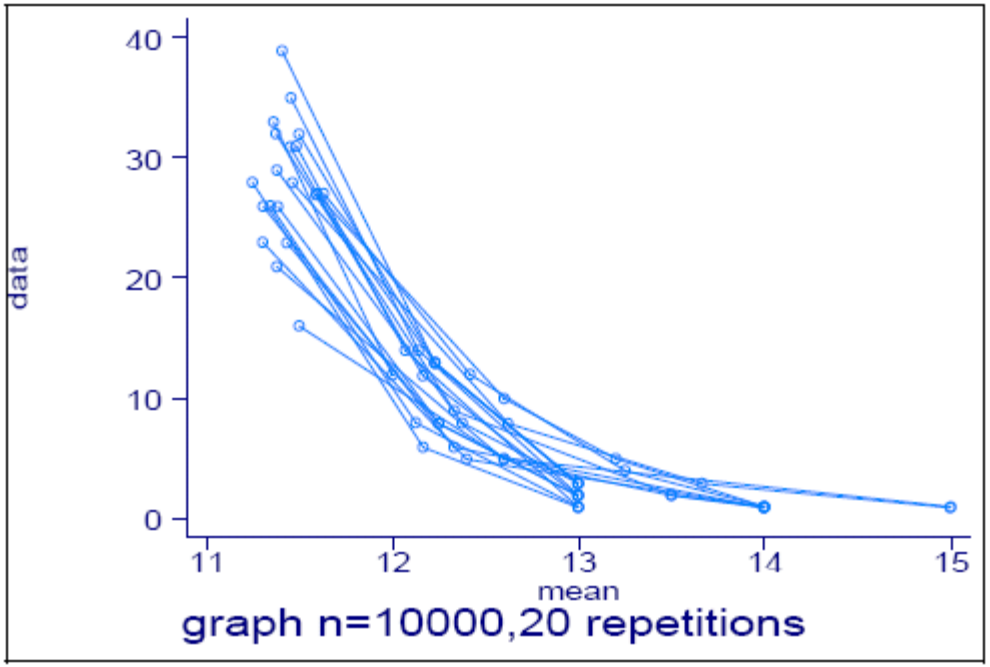
$$E(S \geq x) = Kmne^{-\lambda x} .$$

- Άρα η πιθανότητα να υπάρχουν k περιοχές με score>x θα είναι:

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

συνέχεια…

- Επομένως, παριστάνοντας γραφικά το λογάριθμο του αριθμού τοπικών περιοχών με score πάνω από κάποιο όριο σε σχέση με τη μέση τιμή του score για τις περιοχές πάνω από το όριο αυτό παίρνουμε ευθεία γραμμή και μια απλή γραμμική παλινδρόμηση δίνει αμέσως εκτιμήτριες για τα K,λ.
- Απαιτεί πολύ λιγότερες ακολουθίες (~10-20), άρα είναι πολύ πιο γρήγορη μέθοδος



Η μέθοδος weighted regression του Pearson (1995)

- Χρησιμοποιείται σε αναζητήσεις σε βάσεις δεδομένων
- Η βάση δεδομένων χωρίζεται σε *k* υποσύνολα σύμφωνα με το μήκος των ακολουθιών n_1, n_2,…, n_k
- Υπολογίζονται όλα τα score S, για την τοπική ομοιότητα των ακολουθιών και στη συνέχεια μια ευθεία σταθμισμένης γραμμικής παλινδρόμησης (weighted linear regression) για τη σχέση:

$$S = a + b\log(n_i).$$

- Όπου n_i, είναι το μήκος των ακολουθιών του i υποσυνόλου της βάσης δεδομένων, ενώ το $\log(n_i)$ είναι σταθμισμένο με την αντίστροφη διασπορά (1/var) των scores σε αυτό το υποσύνολο, καθώς τμήματα με πολύ μεγάλο score θα έχουν και μεγάλη διασπορά. Υπολογίζεται τέλος η εκτιμήτρια της διασποράς , των υπολοίπων της παλινδρόμησης (residual variance) η οποία καθορίζει το z-score.

$$z-score = \frac{S-(a+b\log(n_i)}{\text{var}}$$

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

Διαθέσιμο Software

- SW (http://www-hto.usc.edu/software/seqaln/seqaln-query.html)
- BLAST (www.ncbi.nlm.nih.gov/BLAST/)
- WU-BLAST (http://blast.wustl.edu/)
- FASTA (www.ebi.ac.uk/fasta33/)