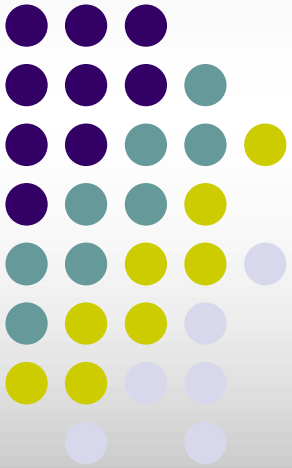
Ζευγαρωτή αντιστοιχία σειρών



Εισαγωγή

- Μεγάλη η σημασία της σύγκρισης των πρωτεϊνικών σειρών (αλληλουχιών), καθώς πρωτεΐνες με όμοιες αλληλουχίες έχουν σχετιζόμενες λειτουργίες ή / και προέρχονται από κοινό πρόγονο.
- Σύγκριση πρωτοταγούς πρωτεϊνικής δομής (σειρές) vs. Σύγκριση δευτεροταγούς τριτοταγούς πρωτεϊνικής δομής.
- Για τον προσδιορισμό της λειτουργίας μίας πρωτεΐνης βρίσκουμε μία ομόλογη αυτής σειρά, για την οποία να είναι γνωστή η λειτουργία της.
- Ως ομόλογες (**homologs**) χαρακτηρίζονται σειρές ή δομές οι οποίες έχουν προέλθει από ένα κοινό πρόγονο μέσα από την εξελικτική διαφοροποίηση.
- Η ομολογία (**homology**) δεν μπορεί να προσδιοριστεί άμεσα, αλλά πρέπει να διαπιστωθεί μέσω της ομοιότητας των εν λόγω σειρών.
- Η διαδικασία αντιστοίχισης δύο σειρών ονομάζεται sequence alignment αυτών.

Ζευγαρωτή αντιστοιχία αλληλουχιών

Αναζήτηση στις βάσεις δεδομένων

- Η αναζήτηση όμοιων αλληλουχιών σε βάσεις δεδομένων μας δίνει τη δυνατότητα
  - ανάκτησης αλληλουχιών, που είναι όμοιες με μια ζητούμενη (query) αλληλουχία, και επίσης τη δυνατότητα
  - ποσοτικοποίησης αυτής της ομοιότητας.

  - Το μέγεθος της ομοιότητας επιτρέπει την αναγνώριση
    - της δομής,
    - της λειτουργίας, ή
    - της οικογενείας της ζητούμενης αλληλουχίας.

- Δύο αλληλουχίες DNA ή αλληλουχίες πρωτεϊνών που είναι πολύ όμοιες πιθανόν να έχουν σχετιζόμενες λειτουργίες και επίσης μπορεί να σχετίζονται επειδή έχουν έναν κοινό πρόγονο.

Αντιστοιχία αλληλουχιών

- Μια από τις πιο χρήσιμες αναπαραστάσεις της ομοιότητας αλληλουχιών είναι η αντιστοιχία. Ας θεωρήσουμε ένα απλό παράδειγμα όπου θέλουμε να συγκρίνουμε τις δύο παρακάτω αλληλουχίες DNA:
  - X= A A T C T G A T A G A A G C C C T A
  - Y= C C A A T C C A G A A C G C C C A
- Μπορούμε να μετασχηματίσουμε την X σε Y (ή αντίστροφα) με μια σειρά απλών αλλαγών βάσεων, μεταλλάξεων ή επεμβατικών λειτουργιών. Οι επιτρεπτές λειτουργίες είναι:
  - Ομοιότητα (match): παραμένει η βάση αμετάβλητη
  - Μη- ομοιότητα (mismatch): αντικατάσταση μιας βάσης από διαφορετική βάση
  - Κενό (gap): εισαγωγή / διαγραφή μιας βάσης

- Μια αντιστοιχία της Χ και Υ είναι η απεικόνιση των επεμβατικών λειτουργιών, οι οποίες είναι απαραίτητες για τον μετασχηματισμό μιας σειράς σε μια άλλη.

- Υπάρχει μεγάλος αριθμός πιθανών αντιστοιχιών της Χ κα Υ, που αντιστοιχούν σε όλους τους δυνατούς συνδυασμούς όπου οι αλληλουχίες θα μπορούσαν να αποκλίνουν από μια κοινή προγονική αλληλουχία. Μια τέτοια αντιστοιχία είναι η παρακάτω:

X	**-**	**-**	**A**	**A**	**T**	**C**	**T**	**G**	**A**	**T**	**A**	**G**	**A**	**A**	**G**	**C**	**C**	**C**	**T**	**A**
			:	**:**	**:**	**:**		**:**	**:**	*****	**:**		**:**	*****	**:**	**:**	**:**	**:**		**:**
Y	**C**	**C**	**A**	**A**	**T**	**C**	**-**	**G**	**A**	**G**	**A**	**-**	**A**	**C**	**G**	**C**	**C**	**C**	**-**	**A**
Θέση	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Όπου,

: σημαίνει ομοιότητα

***** σημαίνει μη-ομοιότητα

- σημαίνει κενό λόγω της εισαγωγής μιας βάσης σε μια αλληλουχία, ή αντίστοιχα η διαγραφή μιας βάσης στην άλλη αλληλουχία.

- Σύμφωνα με την παραπάνω αντιστοιχία, για τον μετασχηματισμό της X στην Y θα πρέπει να γίνει:
  - Αντικατάσταση της G από T στη θέση 10
  - Αντικατάσταση της A από C στη θέση 14
  - Εισαγωγή της C στις θέσεις 1, 2
  - Διαγραφή της T στις θέσεις 7, 19
  - Διαγραφή της G στην θέση 12
- Οπότε η αντιστοιχία περιέχει 13 ομοιότητες, 2 μη-ομοιότητες και 5 κενά.
- Το συνολικό μήκος της αντιστοιχίας είναι 20.

- Για οποιοδήποτε ζεύγος αλληλουχιών θα υπάρχουν πολλαπλές δυνατές αντιστοιχίες. Για παράδειγμα, χρησιμοποιώντας μερικές διαφορετικές επεμβατικές λειτουργίες, μια εναλλακτική αντιστοιχία για τις παραπάνω σειρές X και Y είναι η παρακάτω:

X		-	-	A	A	T	C	T	G	A	T	A	G	A	A	-	G	C	C	C	T	A
				:	:	:	:		:			:	:	:	:		:	:	:	:		:
Y		C	C	A	A	T	C	-	G	-	-	A	G	A	A	C	G	C	C	C	-	A
Θέση	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	

- Η οποία περιέχει 14 ομοιότητες, 0 αντικαταστάσεις και 7 κενά.

- Τώρα το μήκος της αντιστοιχίας είναι 21 και το ποσοστό ομολογίας έχει αυξηθεί σε (14/21)x100=66.7%.

Στατιστικές μετρήσεις για τη σημαντικότητα αντιστοιχίας στην αναζήτηση σε βάσεις δεδομένων

- Η αντιστοιχία αλληλουχιών πραγματοποιείται με τη χρήση προγραμμάτων υπολογιστών.
- Αυτά τα προγράμματα παρέχουν κάποια στατιστική εκτίμηση δηλώνοντας το επίπεδο αξιοπιστίας που θα πρέπει να σχετίζεται σε μια αντιστοιχία.
- Τα συνηθισμένα στατιστικά μεγέθη είναι το p-value και E-value.

- Το p-value σχετίζει το αποτέλεσμα μιας αντιστοιχίας με την πιθανότητα να είναι τυχαίο
  - (όσο πιο πολύ προσεγγίζει το μηδέν, τόσο μεγαλύτερη αξιοπιστία υπάρχει ότι το αποτέλεσμα είναι πραγματικό).

- Το E-value περιγράφει τον αριθμό επιτυχιών (ομοιοτήτων) που αναμένεται να είναι τυχαία στην αναζήτηση μιας βάσης δεδομένων συγκεκριμένου μεγέθους
  - (όταν το E-value πάρει την τιμή 1 για ένα ταίριασμα, αυτό μπορεί να ερμηνευτεί ότι στην τρέχουσα έρευνα, αναμένεται μόνο από τύχη να βρεθεί μια ομοιότητα με ίδιο αποτέλεσμα.
  - Μια τιμή 0 δηλώνει ότι κανένα δεν αναμένεται να είναι τυχαίο, δηλ. είναι απίθανο η αντιστοιχία να είναι από τυχαία ομοιότητα).

  - αντιπροσωπεύουν την πιθανότητα της αντιστοίχισης που συμβαίνει τυχαία. Πρόκειται για στατιστικό υπολογισμό που βασίζεται στην ποιότητα της αντιστοίχισης (βαθμολογία) και στο μέγεθος της βάσης δεδομένων.
  - Μια E-value 0.001 λέει ότι υπάρχει μια πιθανότητα 0.001 ότι αυτή η αντιστοίχιση θα υπάρξει στην βάση δεδομένων τυχαία, δηλαδή, αν η βάση δεδομένων περιέχει 10000 ακολουθίες, τότε ίσως αναμένετε ότι η αντιστοίχιση θα συμβεί ίσως 10 φορές.
  - Μια τιμή E-value 0 είναι στην πραγματικότητα μια στρογγυλευμένη πιθανότητα (ίσως 1e-250 ή κάτι τέτοιο), και απλά λέει ότι υπάρχει (σχεδόν) καμία πιθανότητα ότι η αντιστοίχιση μπορεί να συμβεί τυχαία.

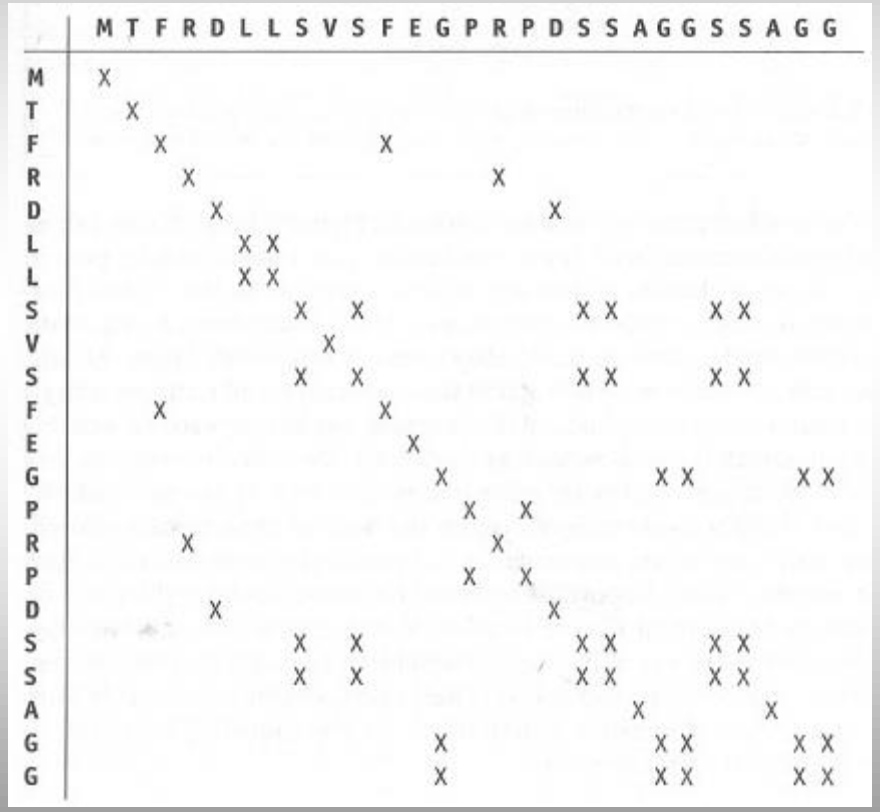
Διάγραμμα ακίδων

Το διάγραμμα ακίδων είναι μια γραφική παράσταση της ομοιότητας δύο αλληλουχιών.

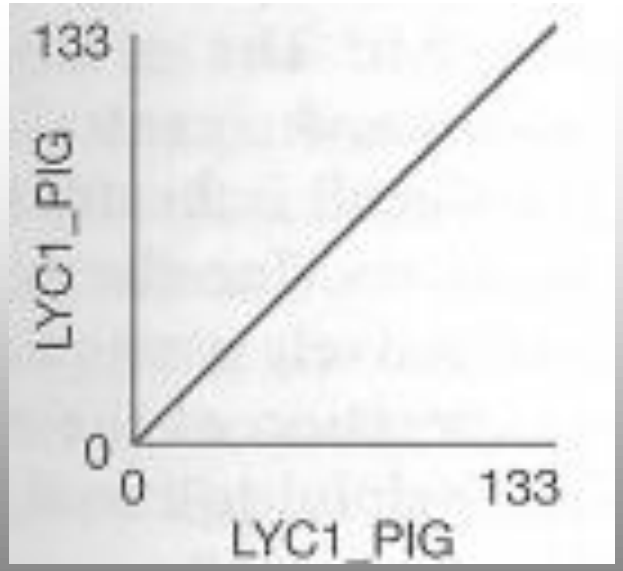
- Ας θεωρήσουμε δύο σειρές, Α και Β, με διαφορετικά μήκη. Σε ένα διάγραμμα ακίδων, δημιουργούμε έναν ορθογώνιο πίνακα όπου π.χ. τα αμινοξέα (residues) της Α τοποθετούνται πάνω στο χ-άξονα και τα αμινοξέα του Β πάνω στο y-άξονα.
  - Τα κελιά για τα οποία ισχύει Ai=Bj παίρνουν την τιμή 1 αλλιώς την τιμή 0.

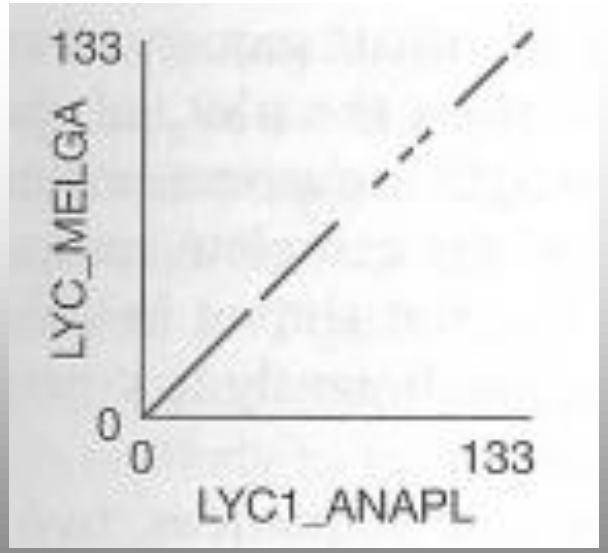
- Στην γραφική απεικόνιση το 1 συμβολίζεται με ακίδα και το 0 με κενό και ο πίνακας παρουσιάζεται με ένα διάγραμμα. Για παράδειγμα, από τη σύγκριση των δύο σειρών δημιουργήθηκε το παρακάτω διάγραμμα.

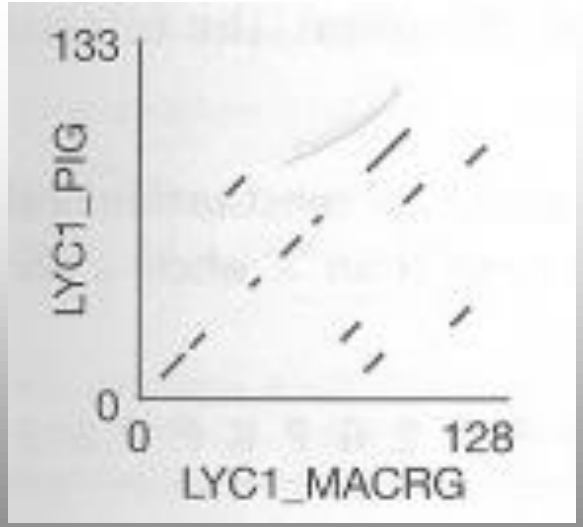
Το διάγραμμα χαρακτηρίζεται από μερικές τυχαίες ακίδες (τυχαίο σφάλμα/θόρυβος) και μια κεντρική διαγώνια γραμμή (μονοπάτι), όπου οι συνεχόμενες ακίδες υψηλής πυκνότητας (σήμα) δηλώνουν τις περιοχές με την μέγιστη ομοιότητα μεταξύ των δύο αλληλουχιών.



- Δύο πανομοιότυπες αλληλουχίες απεικονίζονται με μία απλή συνεχόμενη διαγώνια γραμμή κατά μήκος του διαγράμματος.
- Δύο παρόμοιες αλληλουχίες θα απεικονίζονται με μια διακεκομμένη διαγώνια γραμμή, όπου οι περιοχές με τις διακοπές δηλώνουν μη-ομοιότητα.
- Δύο διαφορετικές αλλά σχετιζόμενες αλληλουχίες θα απεικονίζονται από διαγώνιες ομάδες ακίδων, παράλληλες με την κεντρική διαγώνιο.







Δυναμικός προγραμματισμός – Αλγόριθμος του Needleman και Wunsch

- Στην αντιστοιχία σειρών, η καλύτερη αντιστοιχία, ή το καλύτερο μονοπάτι, μπορεί να βρεθεί χρησιμοποιώντας δυναμικό προγραμματισμό.

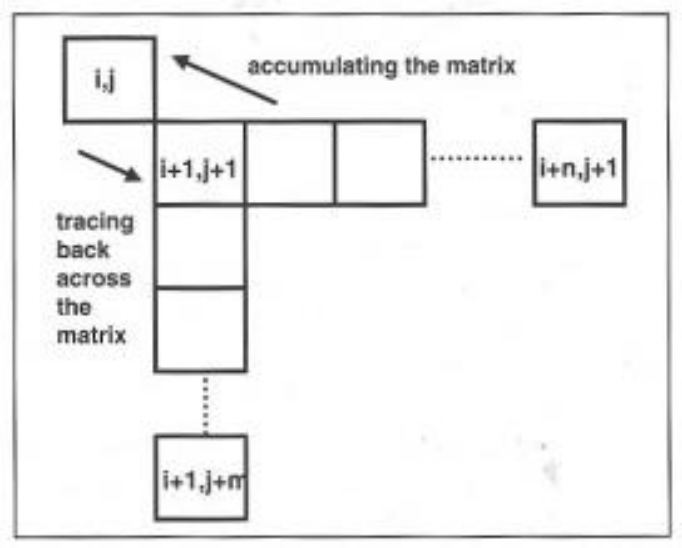
- Ο δυναμικός προγραμματισμός είναι μια τεχνική βελτιστοποίησης για την ανάλυση βαθμολογημένων πινάκων, ο οποίος βρίσκει το υψηλότερα βαθμολογημένο μονοπάτι σε ένα πίνακα, δηλ. βρίσκει την καλύτερη αντιστοιχία μεταξύ δύο σειρών.

- Μεταξύ δύο σειρών που φαίνονται διαφορετικές, συχνά υπάρχουν πολλαπλές δυνατές αντιστοιχίες. Ο δυναμικός προγραμματισμός επιτρέπει τον έλεγχο διαφορετικών μονοπατιών που αντιστοιχούν σε διαφορετικές αντιστοιχίες με υψηλή ομολογία (βαθμολόγηση), συνυπολογίζοντας διάφορες παραμέτρους (π.χ. κυρώσεις από κενά).

  Στην ουσία, προσπαθεί να ταιριάξει τον μέγιστο αριθμό από ζεύγη πανομοιότυπων αμινοξέων αλλά και ταυτόχρονα επιτρέποντας το ελάχιστο αριθμό εισαγωγών και διαγραφών σε δύο σειρές,

- Στο τέλος, επιλέγεται το καλύτερο από όλα τα μονοπάτια, ως η τελική αντιστοιχία.

- Το καλύτερο μονοπάτι από κάθε (i,j) κελί θα προκύψει από τη μεγαλύτερη βαθμολογία των κελιών, της γραμμής (j+1) ή της στήλης (i+1) που βρίσκεται από δεξιά και προς τα κάτω του, στο πίνακα.



	A	D	L	G	A	V	F	A	L	C	D	R	Y	F	Q
A	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0
D	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
L	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
G	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
D	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

	A	D	L	G	A	V	F	A	L	C	D	R	Y	F	Q
A	1	0	0	0	1	0	0	1	0	4	3	2	1	1	0
D	0	1	0	0	0	0	0	0	0	4	4	2	1	1	0
L	0	0	1	0	0	0	0	0	1	4	3	2	1	1	0
G	0	0	0	1	0	0	0	0	5	4	3	2	1	1	0
R	0	0	0	0	0	0	0	0	5	4	3	3	1	1	0
T	0	0	0	0	0	0	0	0	5	4	3	2	1	1	0
Q	0	0	0	0	0	0	0	0	5	4	3	2	1	1	1
N	0	0	0	0	0	0	0	0	5	4	3	2	1	1	0
C	0	0	0	0	0	0	0	0	4	5	3	2	1	1	0
D	0	1	0	0	0	0	0	0	3	3	4	2	1	1	0
R	0	0	0	0	0	0	0	0	2	2	2	3	1	1	0
Y	0	0	0	0	0	0	0	0	2	2	2	2	2	1	0
Y	0	0	0	0	0	0	0	0	1	1	1	1	2	1	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Η τελική αντιστοιχία είναι

ADLGAVFALCDRYFQ

: : : : : : : : :

ADLGRTQN-CDR YYQ

	A	D	L	G	A	V	F	A	L	C	D	R	Y	F	Q
A	9	7	6	6	7	6	6	7	5	4	3	2	1	1	0
D	7	8	6	6	6	6	6	6	5	4	4	2	1	1	0
L	6	6	7	5	5	5	5	5	6	4	3	2	1	1	0
G	5	5	5	6	5	5	5	5	5	4	3	2	1	1	0
R	5	5	5	5	5	5	5	5	5	4	3	3	1	1	0
T	5	5	5	5	5	5	5	5	5	4	3	2	1	1	0
Q	5	5	5	5	5	5	5	5	5	4	3	2	1	1	1
N	5	5	5	5	5	5	5	5	5	4	3	2	1	1	0
C	4	4	4	4	4	4	4	4	4	5	3	2	1	1	0
D	3	4	3	3	3	3	3	3	3	3	4	2	1	1	0
R	2	2	2	2	2	2	2	2	2	2	2	3	1	1	0
Y	2	2	2	2	2	2	2	2	2	2	2	2	2	1	0
Y	1	1	1	1	1	1	1	1	1	1	1	1	2	1	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Scoring model for sequence alignment

Κατά την αντιστοίχηση δύο σειρών η όποια σχέση προκύψει μπορεί να οφείλεται είτε στο ότι όντως οι σειρές αυτές σχετίζονται είτε η μεταξύ τους σχέση μπορεί να είναι τυχαία. Στο διαχωρισμό των δύο αυτών περιπτώσεων σημαντικό ρόλο παίζει το scoring model που θα χρησιμοποιηθεί για την αξιολόγηση της συσχέτισης.

Οι αλλαγές οι οποίες παρουσιάζονται ανάμεσα σε δύο ομόλογες σειρές κατά την πορεία της εξέλιξης, οφείλονται σε:

- μεταλλάξεις (**mutations**)
  - αντικαταστάσεις (**substitutions**) βάσεων (DNA ή RNA), ή αμινοξέων (πρωτεΐνες)
  - **gaps/indels**
    - εισαγωγές (**insertions**) βάσης ή αμινοξέως
    - διαγραφές (**deletions**) βάσης ή αμινοξέως
- φυσική επιλογή (**selection**).

Απλό scoring model:

επιμέρους βαθμοί που αποδίδονται για κάθε αντιστοίχηση βάσεων / αμινοξέων μεταξύ των δύο σειρών (θετικός όρος) + επιμέρους βαθμοί που αποδίδονται για κάθε κενό στις σειρές (αρνητικός όρος)

Παράδειγμα

- **1ο βήμα:**

  Εντοπίζουμε τις ακριβείς αντιστοιχίες μεταξύ των δύο σειρών και αποδίδουμε βαθμολογία στην καθεμία.

  ACCGG**TATC**C - - - **GAC**
  : : : : : : : * : : :
  ACC - - **TATC**T**TAGGAC**

- **2ο βήμα:**

  Εντοπίζουμε τις συντηρητικές αντικαταστάσεις και αποδίδουμε σε αυτές τους ανάλογους βαθμούς.

  ACCGGTATC**C** - - - GAC
  : : : : : : **:* : : :**
  ACC - - TATC**T**TAGGAC

- **3ο βήμα:**

  Αποδίδουμε την κατάλληλη βαθμολογία (ή ποινή) σε κάθε κενό ή εισαγωγή στις σειρές. Το μήκος ενός κενού είναι ο αριθμός των indels που το αποτελούν. Στο απλό αυτό παράδειγμα συναντούμε δύο κενά (ένα κενό σε κάθε σειρά), μήκους 2 και 3.

  ACC**GG**TATCC - - - GAC
  : : : : : : :* : : :
  ACC - - TATCT**TAG**GAC

Βαθμολόγηση / Ποινές κενών (Gap penalties)

Πολύ σημαντική διαδικασία η επιλογή του τρόπου βαθμολόγησης των κενών σε μία σειρά.

Τα κενά αυξάνουν την αβεβαιότητα στην αντιστοίχηση. Από βιολογικής απόψεως θεωρείται πιο εύκολο για μία πρωτεΐνη να 'δεχθεί' την αντικατάσταση ενός residue σε μία θέση, αντί για την εισαγωγή ή διαγραφή τμημάτων της αλληλουχίας. Επομένως τα κενά (gaps)/ εισαγωγές (insertions) θα έπρεπε να είναι πιο σπάνια από τις αντικαταστάσεις.

Αυθαίρετη εισαγωγή κενών χωρίς ποινή θα σήμαινε τελικά να προκύπτει αντιστοιχία μεταξύ οποιονδήποτε σειρών, ακόμα και μεταξύ σειρών τελείως άσχετων μεταξύ τους.

Πρέπει να λαμβάνεται υπόψην η εκάστοτε περίπτωση:

- κενά σε introns vs. κενά σε exons
- κενά που συναντώνται σε περιοχές κωδικοποίησης μίας πρωτεΐνης, κ.ο.κ.

Βαθμολόγηση / Ποινές κενών (Gap penalties)

Τρόποι βαθμολόγησης κενών:

- **Σταθερός (constant)** : βαθμολόγηση / αξιολόγηση κενού ανεξάρτητη από το μήκος του
- **Καθορισμένος (affine)** : ποινή ανοιχτού κενού – **gap open penalty**(ποινή για το πρώτο residue κάθε νέου κενό στη σειρά) & ποινή κενού επέκτασης – **gap extension penalty** (ποινή για κάθε επιπλέον residue στο κενό)
- **Κυρτός (convex)** : κάθε επιπλέον κενό συμβάλει σε μικρότερο βαθμό στην ολική / τελική βαθμολογία
- **Αυθαίρετος (arbitrary)** : κάποια αυθαίρετη συνάρτηση βασιζόμενη στο μήκος του εκάστοτε κενού

  Π.χ., συνάρτηση της μορφής: $\gamma(g) = -gd$,

  όπου : g: μήκος του κενού,

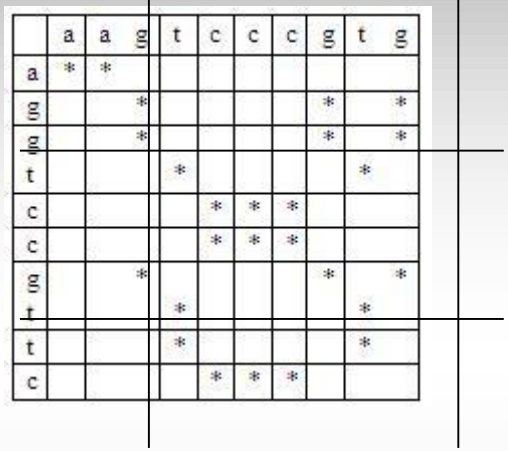
  d: σταθερά και

  γ(g): βαθμολογία του κενού συναρτήσει του μήκους του.

Dot plots

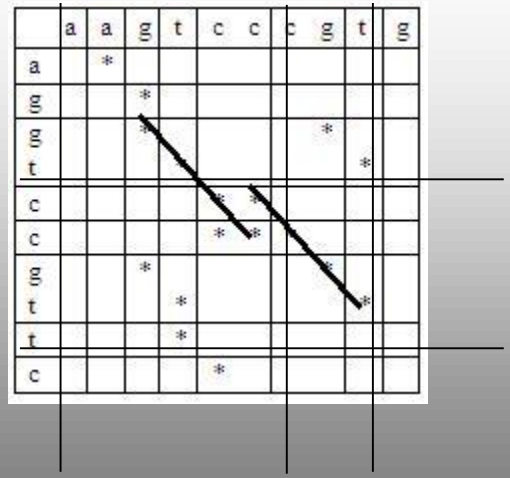
(i) Dot plot (διάγραμμα ακίδων) όπου αναπαρίσταται η αντιστοίχηση δύο σειρών. Τα κελιά του πίνακα τα οποία σχετίζονται με αντιστοιχισμένα στοιχεία των δύο σειρών έχουν σημαδευτεί με ακίδα / αστερίσκο.

- Τα κοινά τμήματα αναπαρίστανται στον πίνακα σαν διαγώνιες.
- Οι εισαγωγές και ο διαγραφές εμφανίζονται σαν διακοπές στη διαγώνιο.
- Περιοχές με τοπική αντιστοίχηση αναπαριστώνται με μικρές διαγωνίους.



(ii) Μόνο τα κελιά τα οποία αντιστοιχούν σε tuples δύο και τεσσάρων βάσεων έχουν σημαδευτεί και έχει επισημανθεί το βέλτιστο μονοπάτι - αντιστοίχηση.

Σημείωση : *k-tuple* είναι μία σειρά από *k residues* μέσα σε μία σειρά. Π.χ., ένα *2-tuple* αντιστοιχεί σε δύο συνεχόμενα residues

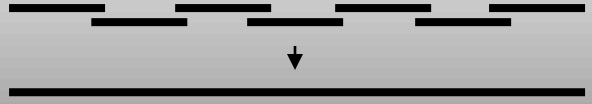


Είδη αντιστοίχησης (Alignment types)

- **Καθολική αντιστοίχηση** δύο σειρών (**global alignment**). Η αντιστοίχηση αυτή χρησιμοποιείται σε περιπτώσεις όπου οι σειρές έχουν ακριβώς το ίδιο ή σχεδόν το ίδιο μήκος.

- **Τοπική αντιστοίχηση** δύο σειρών (**local alignment**). Χρησιμοποιείται για την εύρεση κοινών υποσειρών μέσα στις σειρές.

- **Αντιστοίχηση με ελεύθερα άκρα (Ends free alignment)**. Για την εύρεση ενώσεων (joins) / επικαλύψεων (overlaps).



Δυναμικός Προγραμματισμός

- Μεταξύ δύο σειρών συχνά υπάρχουν πολλαπλές δυνατές αντιστοιχίες, οι οποίες μέσα σε έναν πίνακα βαθμολόγησης (score matrix) εμφανίζονται ως μονοπάτια.
- Ένας πίνακας βαθμολόγησης αποτελεί μία πιο εξελιγμένη μορφή των dot plots. Οι πίνακες βαθμολόγησης χρησιμοποιούνται για να 'κρατούν' τη βαθμολόγηση των αντικαταστάσεων που συναντώνται σε μία αντιστοίχηση, και ειδικά σε περιπτώσεις αντιστοίχησης σειρών πρωτεϊνών. Χρησιμοποιούνται όμως και με σειρές DNA.
- Για την εύρεση της καλύτερης αντιστοιχίας, ή του καλύτερου μονοπατιού, χρησιμοποιείται ο δυναμικός προγραμματισμός.
- Ο δυναμικός προγραμματισμός είναι μία μέθοδος βελτιστοποίησης για την ανάλυση πινάκων βαθμολόγησης (score matrices).
- Ο δυναμικός προγραμματισμός επιτρέπει τον έλεγχο διαφορετικών μονοπατιών τα οποία αντιστοιχούν σε διαφορετικές αντιστοιχίες με υψηλή ομολογία (βαθμολόγηση), συνυπολογίζοντας διάφορες παραμέτρους (π.χ. ποινές κενών).
- Στην ουσία, προσπαθεί να ταιριάξει τον μέγιστο αριθμό από ζεύγη πανομοιότυπων αμινοξέων ανάμεσα στις δυο σειρές, επιτρέποντας ταυτοχρόνως το ελάχιστο αριθμό εισαγωγών και διαγραφών στις δύο αυτές σειρές.
- Ως τελική αντιστοιχία επιλέγεται το βέλτιστο απ' όλα τα μονοπάτια.

Αλγόριθμος Needleman – Wunsch

- Ο αλγόριθμος Needleman – Wunsch είναι αλγόριθμος δυναμικού προγραμματισμού ο οποίος χρησιμοποιείται σε περιπτώσεις καθολικής αντιστοίχησης δύο σειρών.
- Καθολική αντιστοίχηση = Συμπεριλαμβάνει **όλες** τις βάσεις και από τις δύο σειρές στην αντιστοίχηση και στην βαθμολόγηση.
- Τα κενά προστίθονται στο εσωτερικό, ή στα άκρα κάθε σειράς με αποτέλεσμα το μήκος των δύο σειρών (βάσεις + κενά) να είναι ακριβώς το ίδιο.
- Κάθε βάση ή κενό στην κάθε σειρά αντιστοιχίζεται με μία βάση ή κενό στην άλλη σειρά.
- Η βασική ιδέα του αλγορίθμου είναι να χτιστεί η βέλτιστη αντιστοίχηση χρησιμοποιώντας προηγούμενες λύσεις από βέλτιστες αντιστοιχίσεις μικρότερων υποσειρών.

Αλγόριθμος Needleman – Wunsch

Ας υποθέσουμε τώρα πως έχουμε δύο σειρές *S* και *T*.

- Η σειρά *S* αποτελείται από *n* βάσεις, επομένως έχει μήκος *n*, και η σειρά *T* αποτελείται από *m* βάσεις (έχει λοιπόν μήκος *m*).
- Θέτουμε μία αυθαίρετη ποινή για τα κενά της τάξης του -1 για κάθε indel.
- Συμβολίζουμε την αντιστοίχηση μεταξύ της βάσης *i* στη σειρά *S* με ένα κενό στη σειρά *T* ως: (*Si*, -).
- Η βαθμολογία στην περίπτωση αυτή συμβολίζεται ως: σ(*Si*, -) = -1.
- Συμβολίζουμε την αντιστοίχηση μεταξύ της βάσης *i* στη σειρά *S* με τη βάση *j* στη σειρά *T* ως: (*Si*, *Tj*).
- Η βαθμολογία στην περίπτωση αυτή συμβολίζεται ως: σ(*Si*, *Tj*) και για mismatch είναι σ(*Si*, *Tj*) = -1, ενώ για match είναι σ(*Si*, *Tj*) = 2.

Αλγόριθμος Needleman – Wunsch

- Για πίνακα βαθμολόγησης, φτιάχνουμε ένα πίνακα διαστάσεων n+1 επί m+1.

- Η γραμμή 0 και η στήλη 0 του πίνακα αναπαριστούν το κόστος που θα είχαμε προσθέτοντας διαδοχικά κενά και στις δύο σειρές κατά την έναρξη της αντικατάστασης.

- Η βαθμολογία για κάθε κελί του πίνακα υπολογίζεται αναδρομικά, με βάση τις βαθμολογίες των γύρω και προηγούμενων από αυτό κελιών, βρίσκοντας τη βέλτιστη επιλογή ανάμεσα σε αυτά η οποία αναπαριστά είτε κενό, είτε επιτυχία / αποτυχία αντιστοίχησης.

  Ας δούμε ένα παράδειγμα εφαρμογής του αλγορίθμου:

Παράδειγμα Needleman – Wunsch

- Έχουμε τις ακόλουθες δύο σειρές :
  S = ACCGGTAT
  T = ACCTATC

- Μήκος της σειράς S : n = 8
  Μήκος της σειράς T : m = 7
  Τα δύο μήκη είναι σχεδόν τα ίδια, επομένως μπορούμε να χρησιμοποιήσουμε καθολική αντιστοίχηση.

- Φτιάχνουμε τον πίνακα V διαστάσεων (n+1=)9 επί (m+1=)8.

- Ορίζουμε την τιμή του στοιχείου V(0,0) = 0.

- Η πρώτη γραμμή και η πρώτη στήλη συμπληρώνονται σαν να προσθέταμε διαδοχικά κενά και στις δύο σειρές.

Παράδειγμα Needleman – Wunsch

Πίνακας διαστάσεων 9x8 :

		A	C	C	G	G	T	A	T	
	0 →	-1 →	-2 →	-3 →	-4 →	-5 →	-6 →	-7 →	-8	
A	-1 ↓									
C	-2 ↓									
C	-3 ↓									
T	-4 ↓									
A	-5 ↓									
T	-6 ↓									
C	-7									
(T)										

Παράδειγμα Needleman – Wunsch

- Συμπληρώνουμε τον υπόλοιπο πίνακα κινούμενοι γραμμή – γραμμή από πάνω αριστερά προς κάτω δεξιά.

- Γνωρίζοντας τις τιμές των *V(i-1, j-1)*, *V(i-1, j)* και *V(i, j-1)* μπορεί να υπολογιστεί η τιμή του *V(i, j)*, η οποία δίδεται ως εξής:

$$V(i,j) = \max \begin{cases} V(i-1,j-1) + \sigma(Si, Tj) \\ V(i-1,j) + \sigma(Si, -) \\ V(i,j-1) + \sigma(-, Tj) \end{cases}$$

Η τιμή του *V(i, j)* δηλαδή βρίσκεται μέσω του βέλτιστου μονοπατιού...

Παράδειγμα Needleman – Wunsch

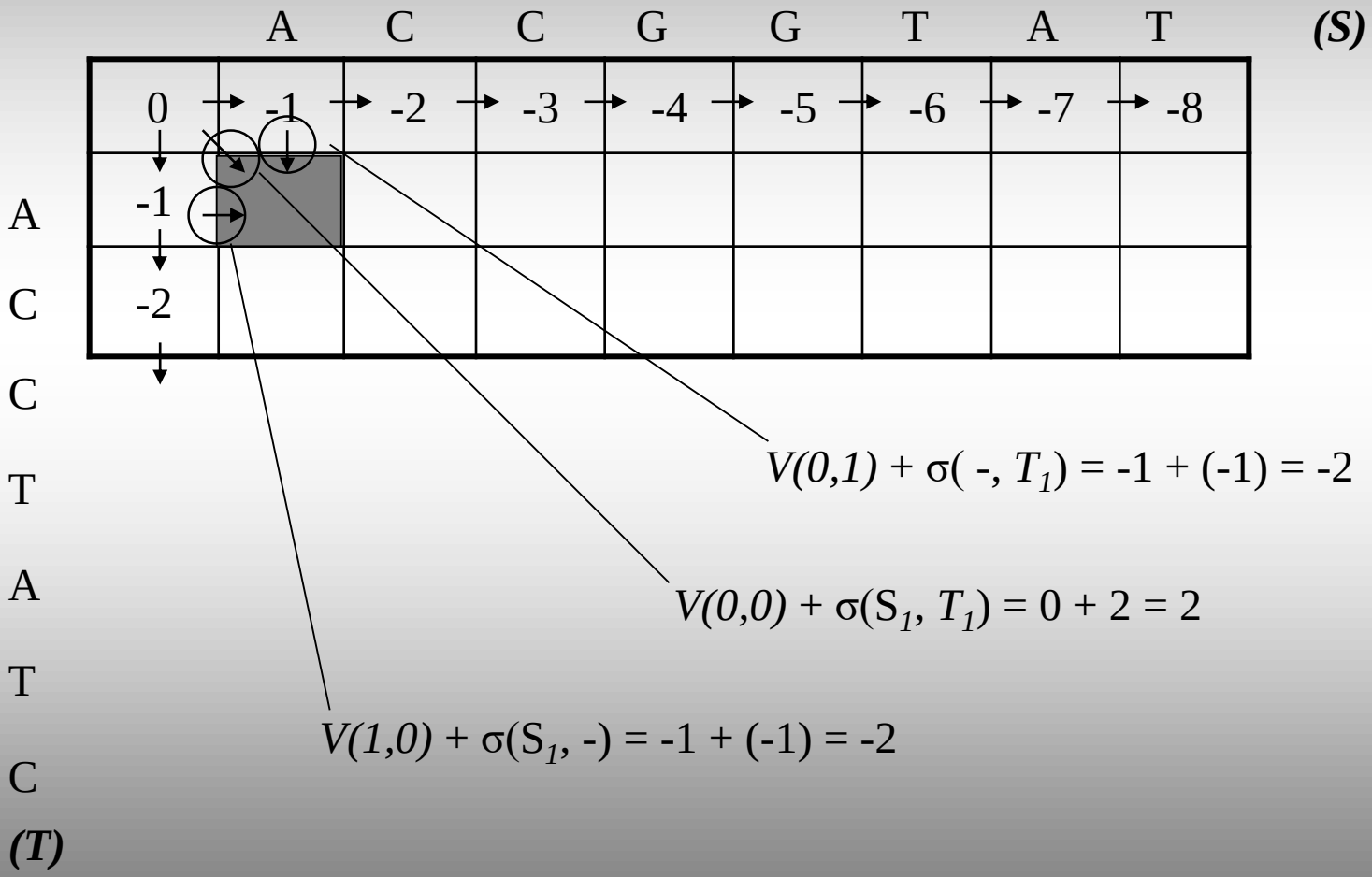
Κανόνες συμπλήρωσης πίνακα βαθμολόγησης:

- Σε περίπτωση μη ταύτισης μπορούμε να κινηθούμε είτε διαγωνίως, είτε οριζοντίως ή καθέτως.
  - Κινούμενοι καθέτως, θεωρούμε πάντα ότι εισάγουμε κενό στη σειρά *S* η οποία αναπαρίσταται οριζόντια στον πίνακα. Επομένως, η βαθμολογία που προστίθεται σε αυτή του στοιχείου *V(i, j-1)* είναι πάντα η ποινή (σ(*Si*, -)) που αντιστοιχεί σε κενό.
  - Το ίδιο ισχύει πάντα και όταν κινούμαστε οριζοντίως, με τη μόνη διαφορά ότι το κενό εισάγεται στην κάθετη σειρά *T* και η ποινή του είναι (σ(-, *Ti*)).
  - Κινούμενοι διαγωνίως, προσθέτουμε την (αρνητική) ποινή που έχει οριστεί για τη μη-ταύτιση.

Παράδειγμα Needleman – Wunsch

- Σε περίπτωση που υπάρχει ακριβής ταύτιση των αντιστοιχούμενων βάσεων μεταξύ των δύο σειρών, μπορούμε να κινηθούμε πάλι είτε διαγωνίως, είτε οριζοντίως ή καθέτως.
  - Κινούμενοι καθέτως ή οριζοντίως θεωρούμε, ακόμα και στην περίπτωση της ταύτισης, ότι εισάγουμε κενό στη σειρά *S* ή στη σειρά *T* αντιστοίχως. Επομένως, η βαθμολογία που προστίθεται σε αυτή του στοιχείου *V(i, j-1)* είναι αντίστοιχα η ποινή (σ(*Si*, -)) ή η ποινή (σ(-, *Ti*)) που αντιστοιχεί σε κενό.
  - Μόνο κινούμενοι διαγωνίως, προσθέτουμε τον προκαθορισμένο βαθμό για την απόλυτη ταύτιση (στην προκειμένη σ(S*i*, *Ti*) = 2) στην τιμή του στοιχείου *V(i-1, j-1)*.

Παράδειγμα Needleman – Wunsch



Παράδειγμα Needleman – Wunsch

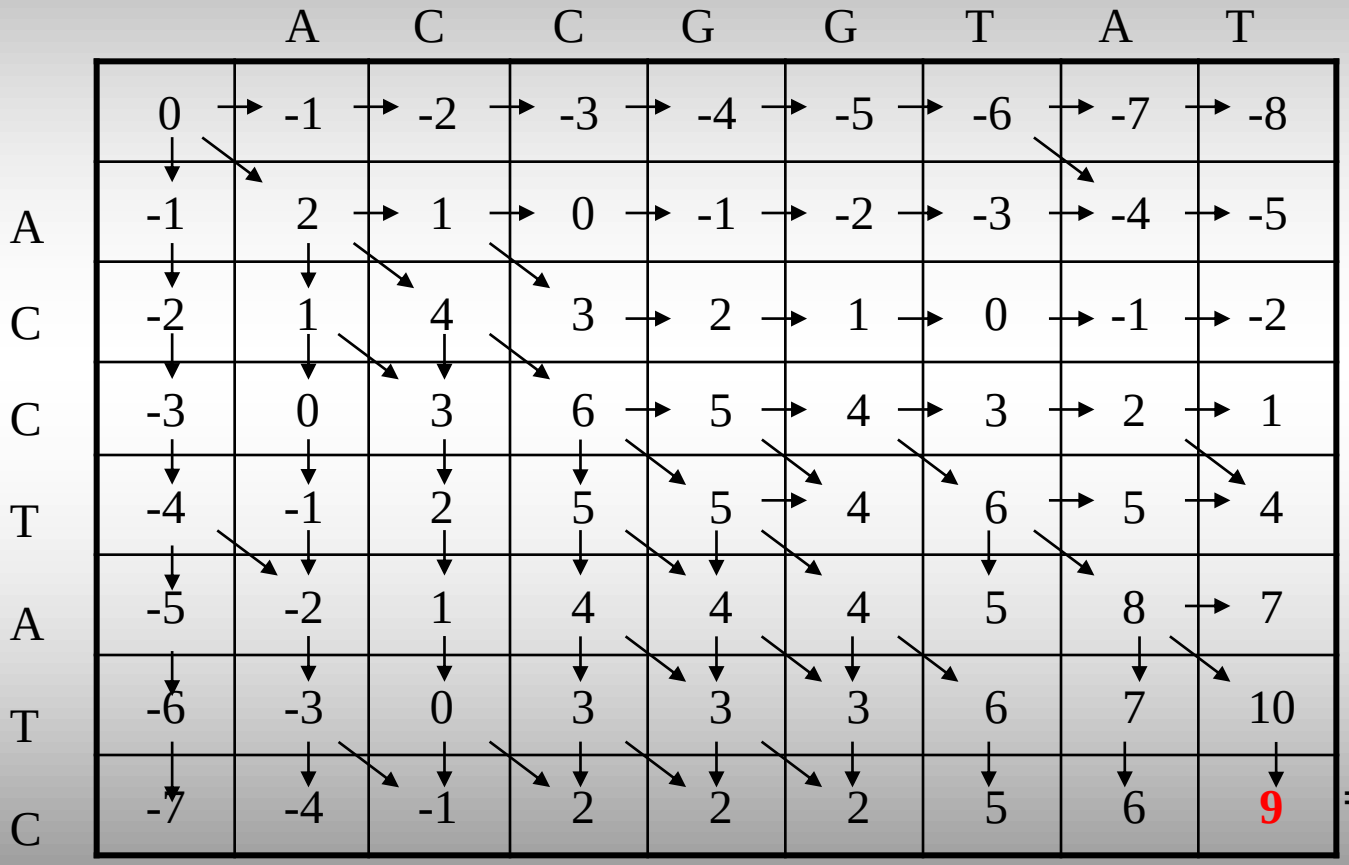
- Επομένως στο παράδειγμά μας θα είναι :

$$V(1,1) = \max \begin{cases} V(0,0) + \sigma(S_1, T_1) = 0 + 2 = 2 \\ V(0,1) + \sigma(S_1, -) = -1 + (-1) = -2 \\ V(1,0) + \sigma(-, T_1) = -1 + (-1) = -2 \end{cases}$$

Άρα: $V(1,1) = 2$

Με τον ίδιο τρόπο συμπληρώνουμε και τα υπόλοιπα στοιχεία του πίνακα....

Παράδειγμα Needleman – Wunsch



= Ολικό score

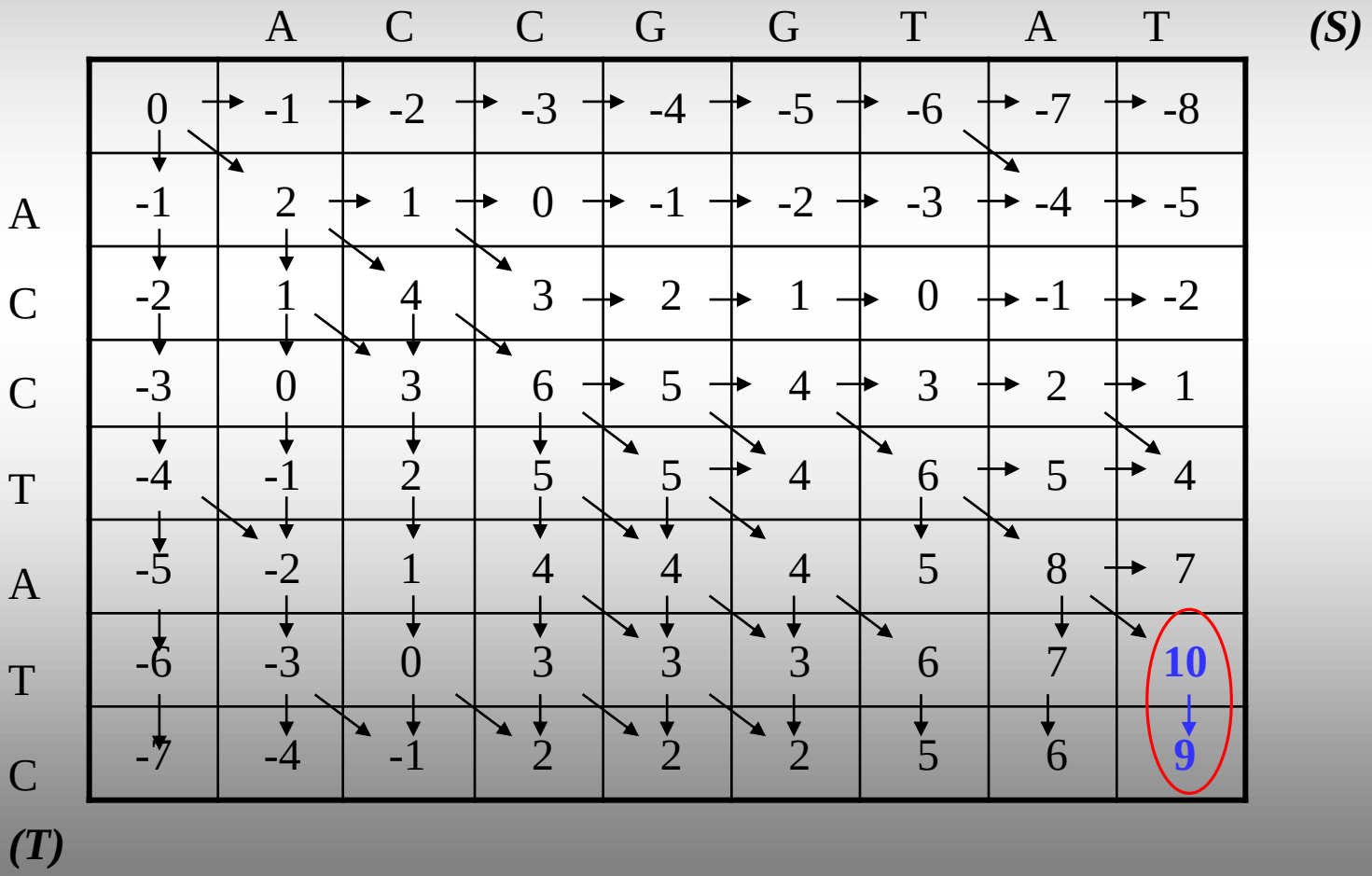
Παράδειγμα Needleman – Wunsch

Για την ανακατασκευή της αντιστοίχησης των σειρών:

- Βρίσκουμε το οριακό στοιχείο του πίνακα (δηλαδή στοιχείο που βρίσκεται στην ακριανή γραμμή ή ακριανή στήλη και απ' το οποίο δεν ξεκινά δείκτης προς άλλο στοιχείο) και το οποίο έχει τη μέγιστη τιμή.
- Με αφετηρία το στοιχείο αυτό κινούμαστε αναδρομικά ακολουθώντας τους δείκτες που δείχνουν από ποιο προηγούμενο στοιχείο οδηγηθήκαμε στο παρόν.
- Εάν μία τιμή ενός στοιχείου έχει προκύψει από τα δύο ή και τα τρία προηγούμενα στοιχεία του πίνακα, οι δείκτες απ' όλα τα στοιχεία κρατούνται και έτσι προκύπτουν αντίστοιχα δύο ή τρία διαφορετικά μονοπάτια.
- Κάθε μονοπάτι απεικονίζει και μία διαφορετική αντιστοίχηση και η επιλογή του βέλτιστου μονοπατιού / αντιστοίχησης, γίνεται πλέον με γνώμονα τη μέγιστη βαθμολόγηση που αντιστοιχεί σε κάποιο από αυτά. Σε περίπτωση που τα μονοπάτια είναι ισότιμα, η επιλογή γίνεται αυθαίρετα.
- Η μέγιστη τιμή του οριακού στοιχείου του πίνακα απ' όπου ξεκινήσαμε αποτελεί την ολική βαθμολογία της αντιστοίχησης.

- (S)

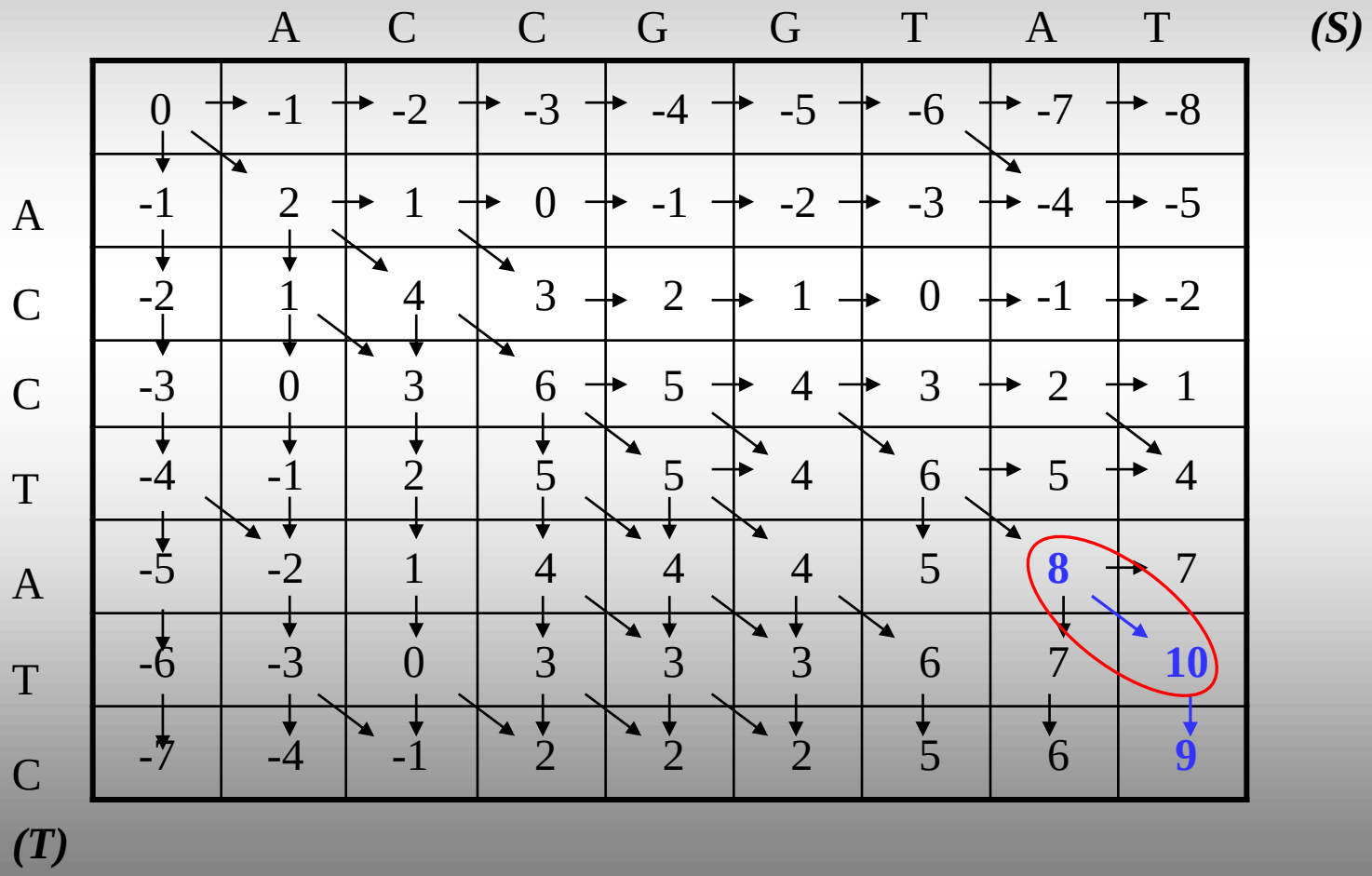
C (T)


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

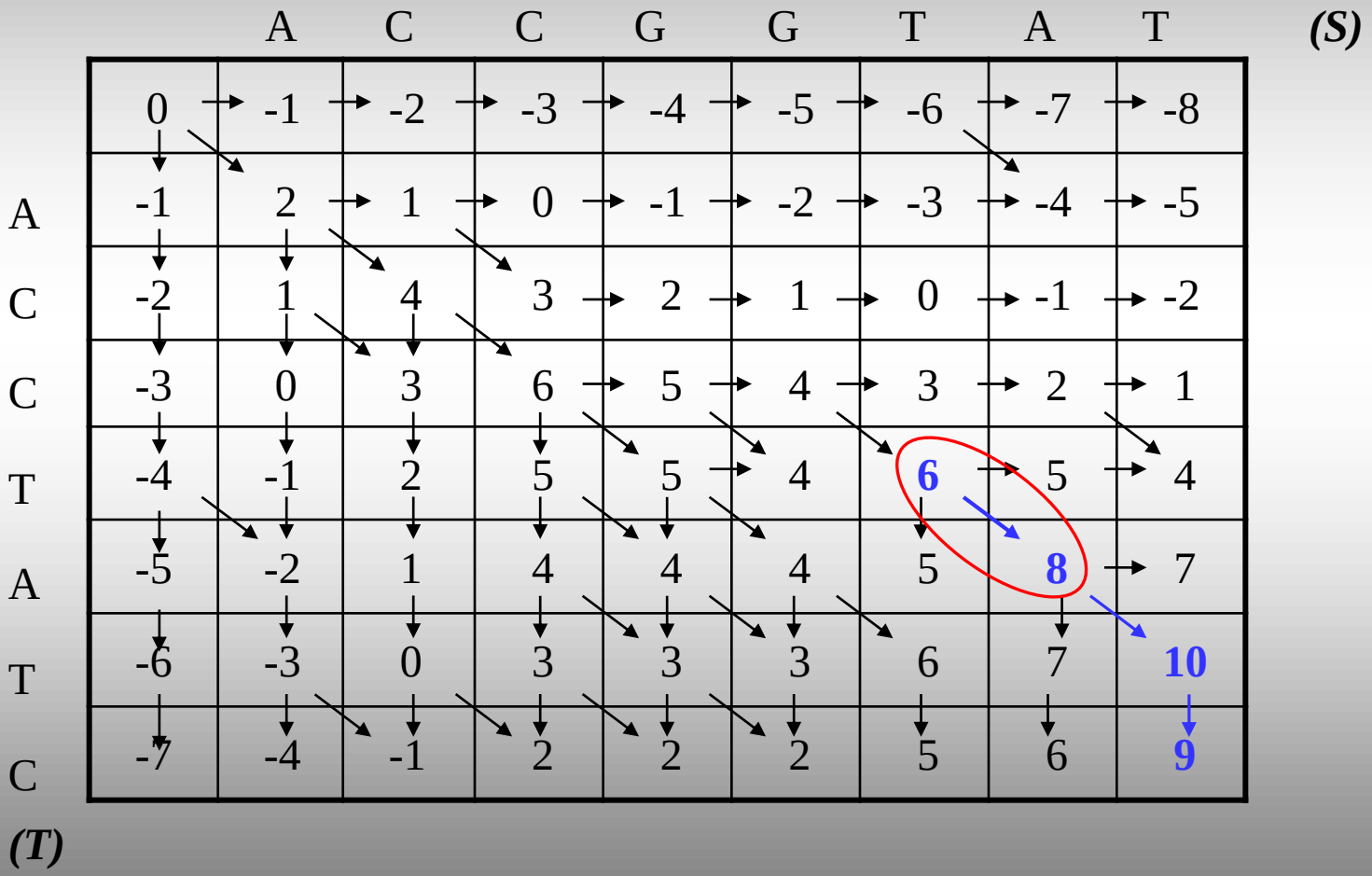
T - ***(S)***

|

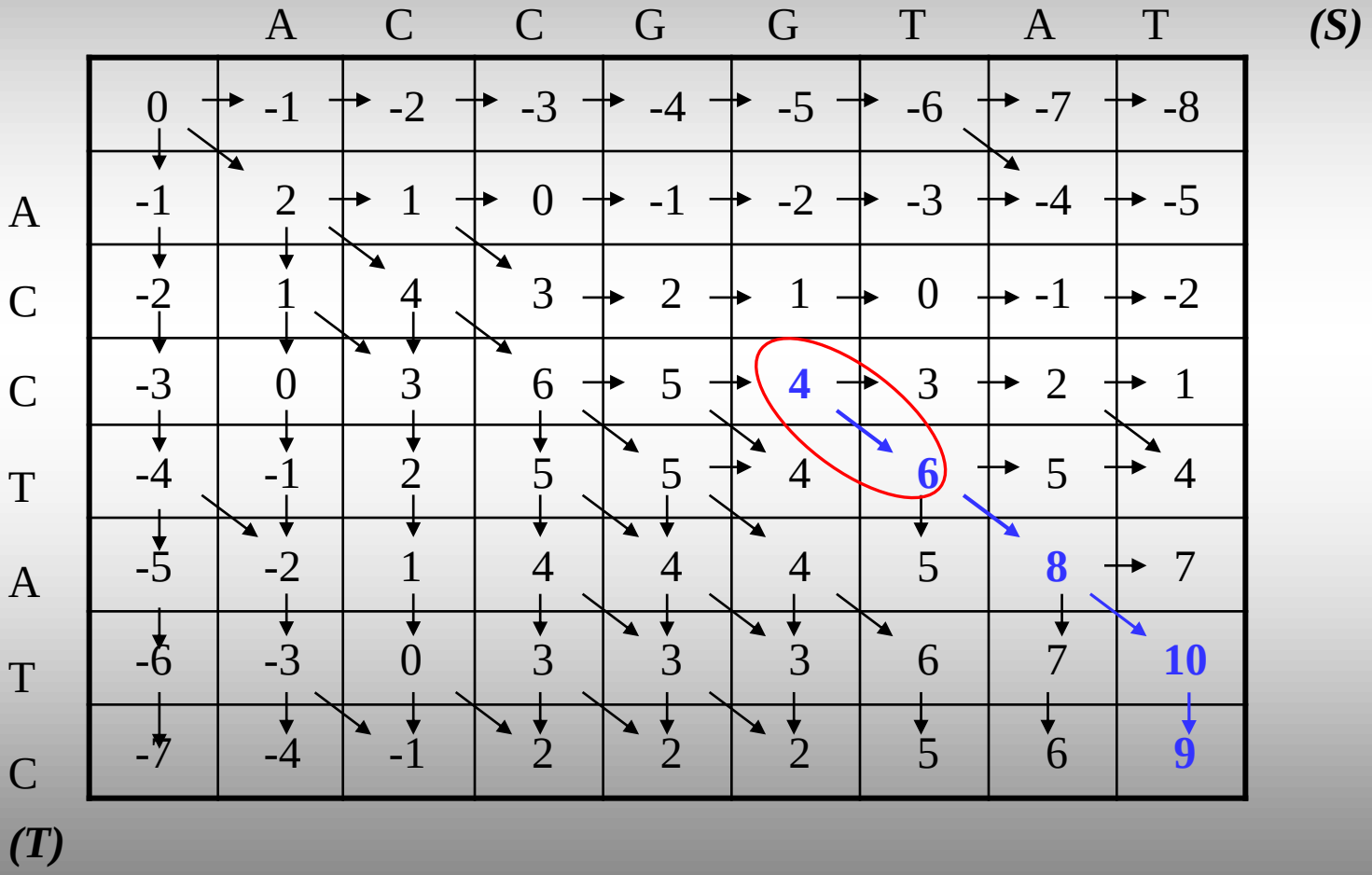
T C ***(T)***


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

A T - *(S)*
| |
A T C *(T)*


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

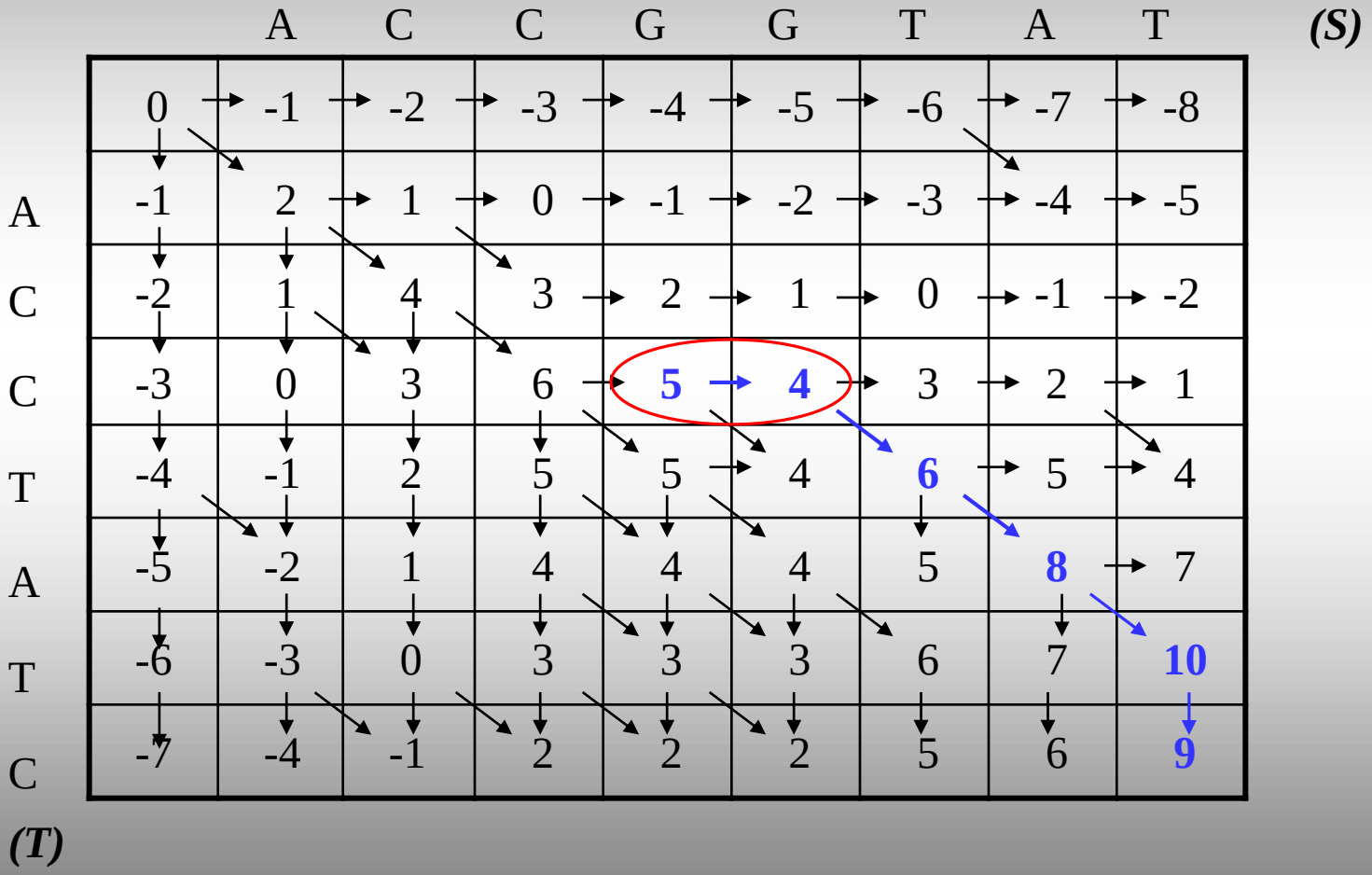
```
TAT-   (S)
|||
TATC   (T)
```


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

G T A T - ***(S)***

| | | |

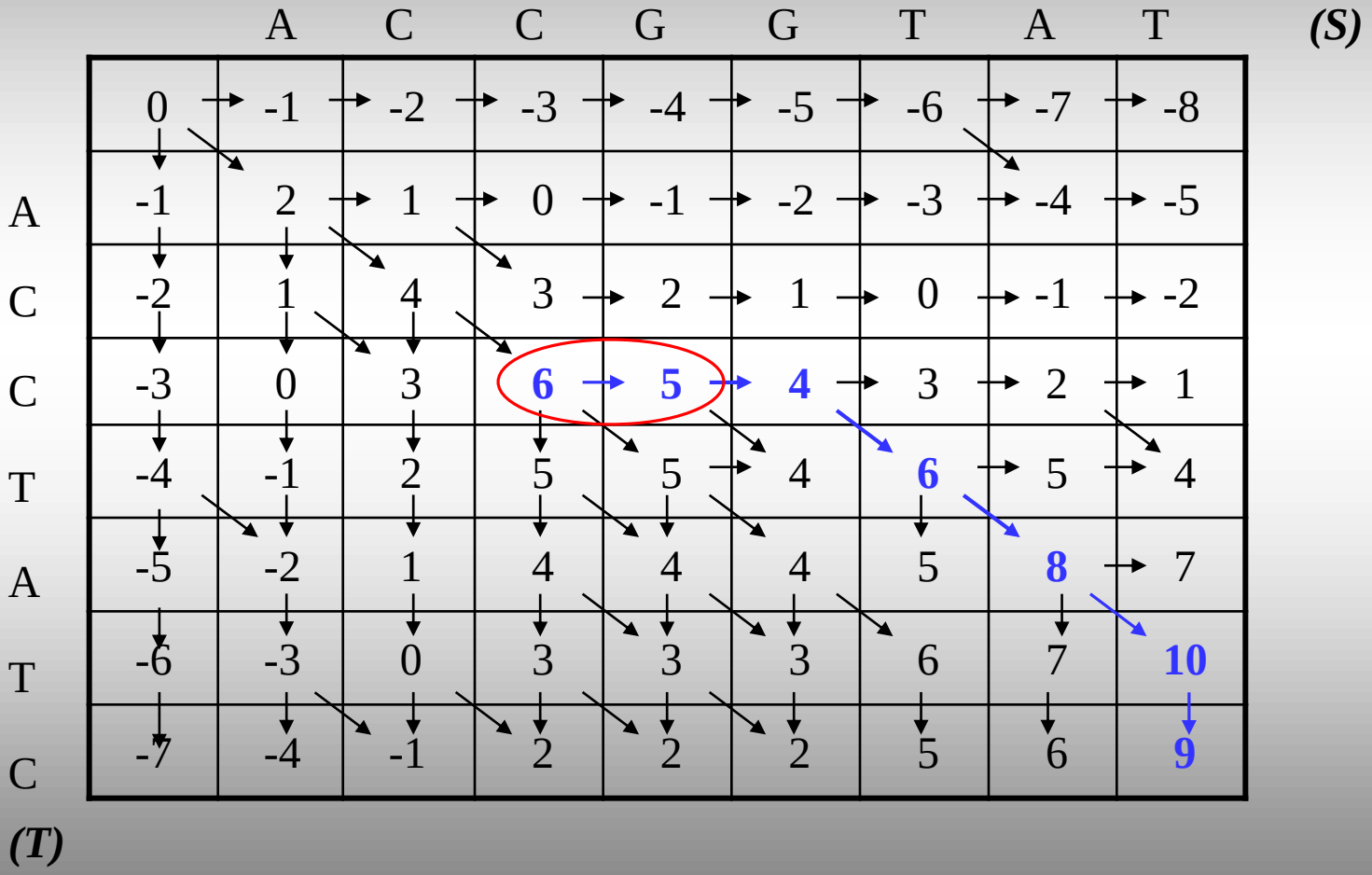
- T A T C ***(T)***


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

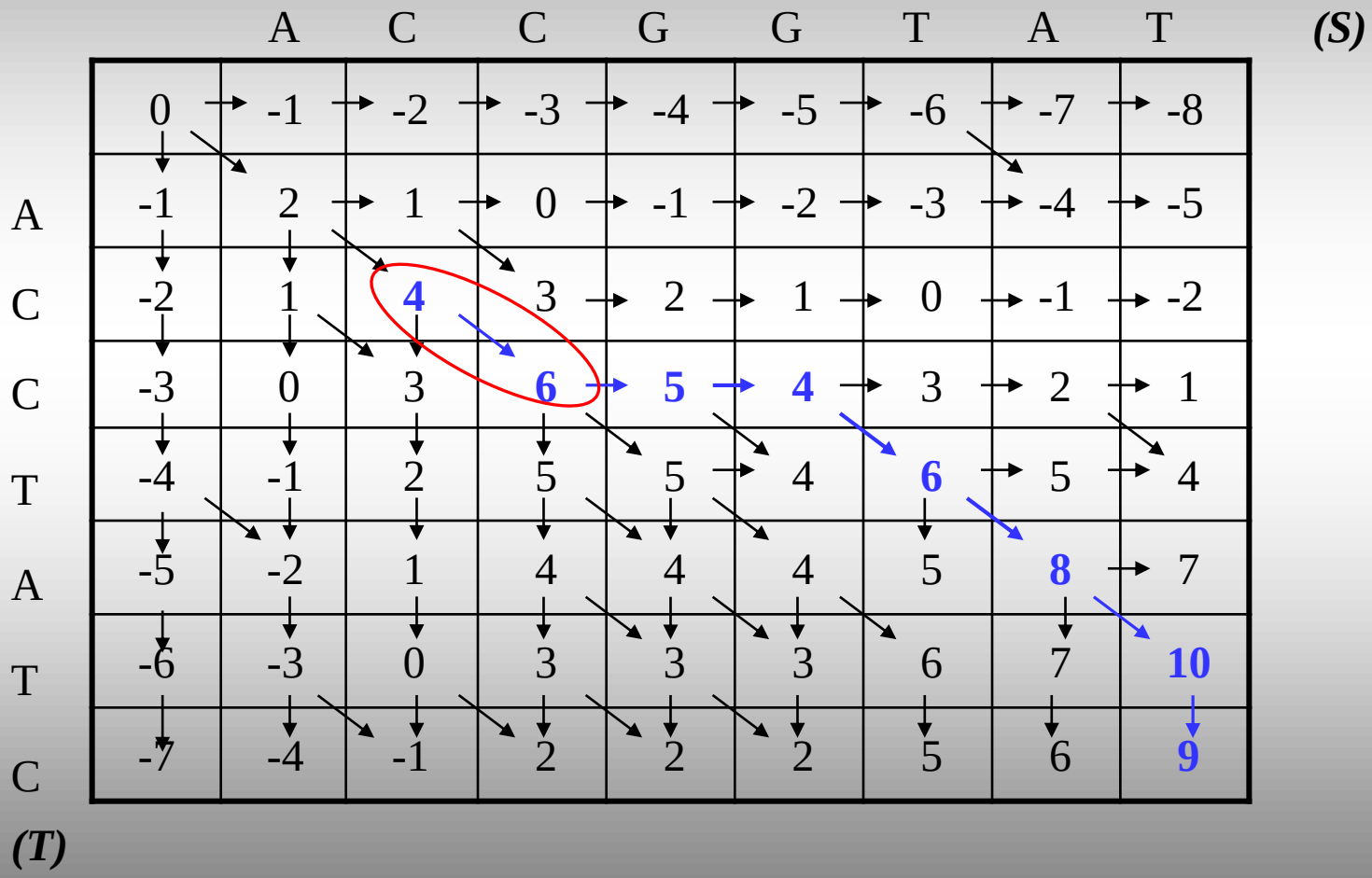
G G T A T - *(S)*

| | | |

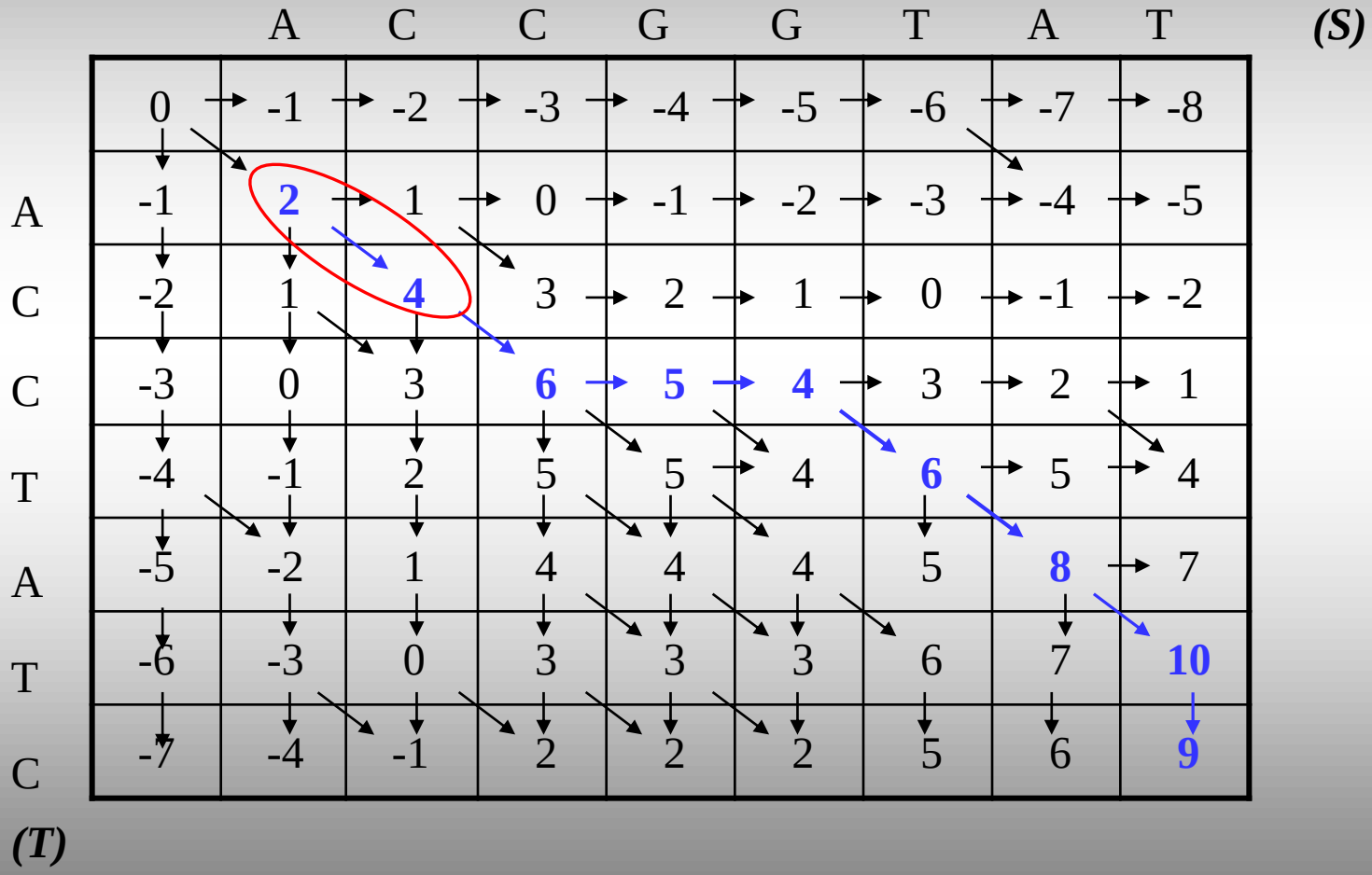
- - T A T C *(T)*



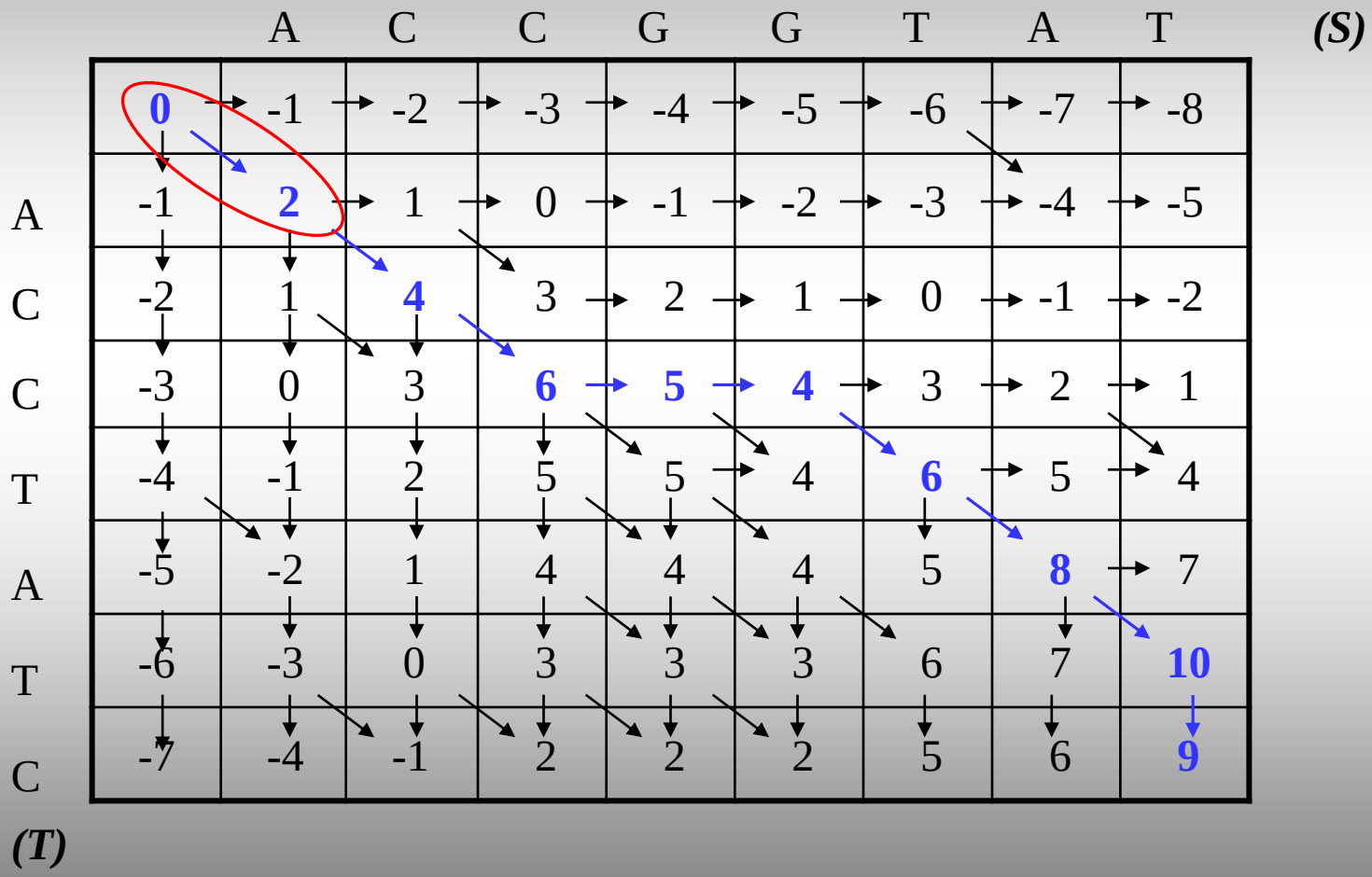
C G G T A T - *(S)*
| | | | |
C - - T A T C *(T)*



```
C C G G T A T -   (S)
| |     | | |
C C - - T A T C   (T)
```



```
A C C G G T A T -   (S)
| | |     | | |
A C C - - T A T C   (T)
```



Παράδειγμα Needleman – Wunsch

Η αντιστοίχηση που προκύπτει για τις σειρές S και T είναι η ακόλουθη:

```
A C C G G T A T -    (S)
| | |       | | |
A C C - - T A T C    (T)
```

- Λάβαμε υπόψη όλες τις βάσεις σε κάθε σειρά.

- Υπολογίζουμε τη βαθμολογία της αντιστοίχησης :
  6 ταυτίσεις (matches) = 6 x 2 = 12 βαθμοί
  3 κενά = 3 x (-1) = -3 βαθμοί ποινής
  0 mismatches
  Άρα συνολικά : 12 + (-3) = 9 βαθμοί
  Η βαθμολογία αυτή συμπίπτει με τη βαθμολογία που προκύπτει από τον πίνακα βαθμολόγησης

Τοπική αντιστοίχηση (Local Alignment)

- Σε ορισμένες περιπτώσεις κρίνεται προτιμότερο να βρούμε τη βέλτιστη αντιστοίχηση ανάμεσα σε *υποσειρές* δύο σειρών (όταν π.χ. υπάρχει περίπτωση δύο πρωτεΐνες να έχουν μία κοινή περιοχή (domain)), αντί για αντιστοίχηση ολόκληρων των σειρών. Σε αυτήν την περίπτωση χρησιμοποιούμε την **τοπική αντιστοίχηση (local alignment)**.

- Η τοπική αντιστοίχηση θεωρείται πιο ευαίσθητη μέθοδος αντιστοίχησης δύο σειρών.

- Είναι χρήσιμη σε περιπτώσεις που δύο σειρές, αν και ομόλογες, έχουν διαφοροποιηθεί πολύ μέσα στην εξελικτική τους πορεία. Τότε μόνο ορισμένες μόνο περιοχές των σειρών θα μπορούν να ταυτιστούν, καθώς οι υπόλοιπες θα έχουν διαφοροποιηθεί τόσο πολύ εξαιτίας του προσαρτώμενου σε αυτές θορύβου.

- Η τοπική αντιστοίχηση παρουσιάζει κοινά με την καθολική:
  - χρησιμοποιεί τον ίδιο τύπο πινάκων βαθμολόγησης και ποινών για τα κενά,
  - κάνει χρήση αλγορίθμων δυναμικού προγραμματισμού.

Αλγόριθμος Smith - Waterman

Εφαρμογή της τοπικής αντιστοίχησης είναι ο αλγόριθμος Smith – Waterman.

- Ο πίνακας βαθμολόγησης για τον αλγόριθμο Smith – Waterman κατασκευάζεται με τον ίδιο τρόπο όπως και για τον αλγόριθμο Needleman – Wunsch.

- Κανόνες κατασκευής του πίνακα βαθμολόγησης για τον αλγόριθμο Smith – Waterman :
  - Εάν η βάση *Si* ταυτίζεται με την *Ti* τότε σ(*Si*, *Tj*) >= 0
  - Σε περίπτωση μη – ταύτισης ή κενού τότε σ(*Si*, *Tj*) <= 0
  - Μόνη διαφορά από τον Needleman – Wunsch ότι για τον Smith – Waterman η κατώτερη τιμή που μπορεί να αποθηκευτεί για κάποιο στοιχείο του πίνακα είναι 0.

    Επιλέγουμε να δώσουμε σε ένα στοιχείο του πίνακα την τιμή 0, αντί μίας αρνητικής τιμής, γιατί έτσι σηματοδοτούμε στην έναρξη μίας νέας αντιστοιχίας. Η λογική είναι ότι σε περίπτωση που προκύψει αρνητικός βαθμός σε μία αντιστοίχηση είναι προτιμότερο να ξεκινήσει μία νέα αντιστοιχία υποσειρών αντί να συνεχιστεί η προηγούμενη.

Αλγόριθμος Smith - Waterman

- Για πίνακα βαθμολόγησης *V*, γνωρίζοντας τις τιμές των *V(i-1, j-1)*, *V(i-1, j)* και *V(i, j-1)*, οι τιμές των στοιχείων του προκύπτουν ως εξής:
  - *V(i, 0)= 0, V(0, j) = 0, για κάθε i, j* και
  - $$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + \sigma(Si, Tj) \\ V(i-1,j) + \sigma(Si, -) \\ V(i,j-1) + \sigma(-, Tj) \end{cases}$$

Ας δούμε όμως ένα παράδειγμα:

- Έστω ότι έχουμε τις σειρές : S = ACCGGTAT μήκους n = 8
  T = TTGTATC μήκους m = 7
- Φτιάχνουμε τον πίνακα V διαστάσεων (n+1=)9 επί (m+1=)8.
- Ορίζουμε V(i, 0) = 0 και V(0, j) = 0, για κάθε i, j.

Παράδειγμα Smith - Waterman

Πίνακας διαστάσεων 9x8 :

		A	C	C	G	G	T	A	T	***(S)***
	0	0	0	0	0	0	0	0	0	
T	0									
T	0									
G	0									
T	0									
A	0									
T	0									
C	0									

(T)

Παράδειγμα Smith - Waterman

Συμπληρώνουμε τα υπόλοιπα στοιχεία του πίνακα (ελάχιστη τιμή στον πίνακα είναι το 0) :

		A	C	C	G	G	T	A	T
	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	2	1	2
T	0	0	0	0	0	0	2	1	3
G	0								
T	0								
A	0								
T	0								
C	0								

(S)

(T)

Παράδειγμα Smith - Waterman

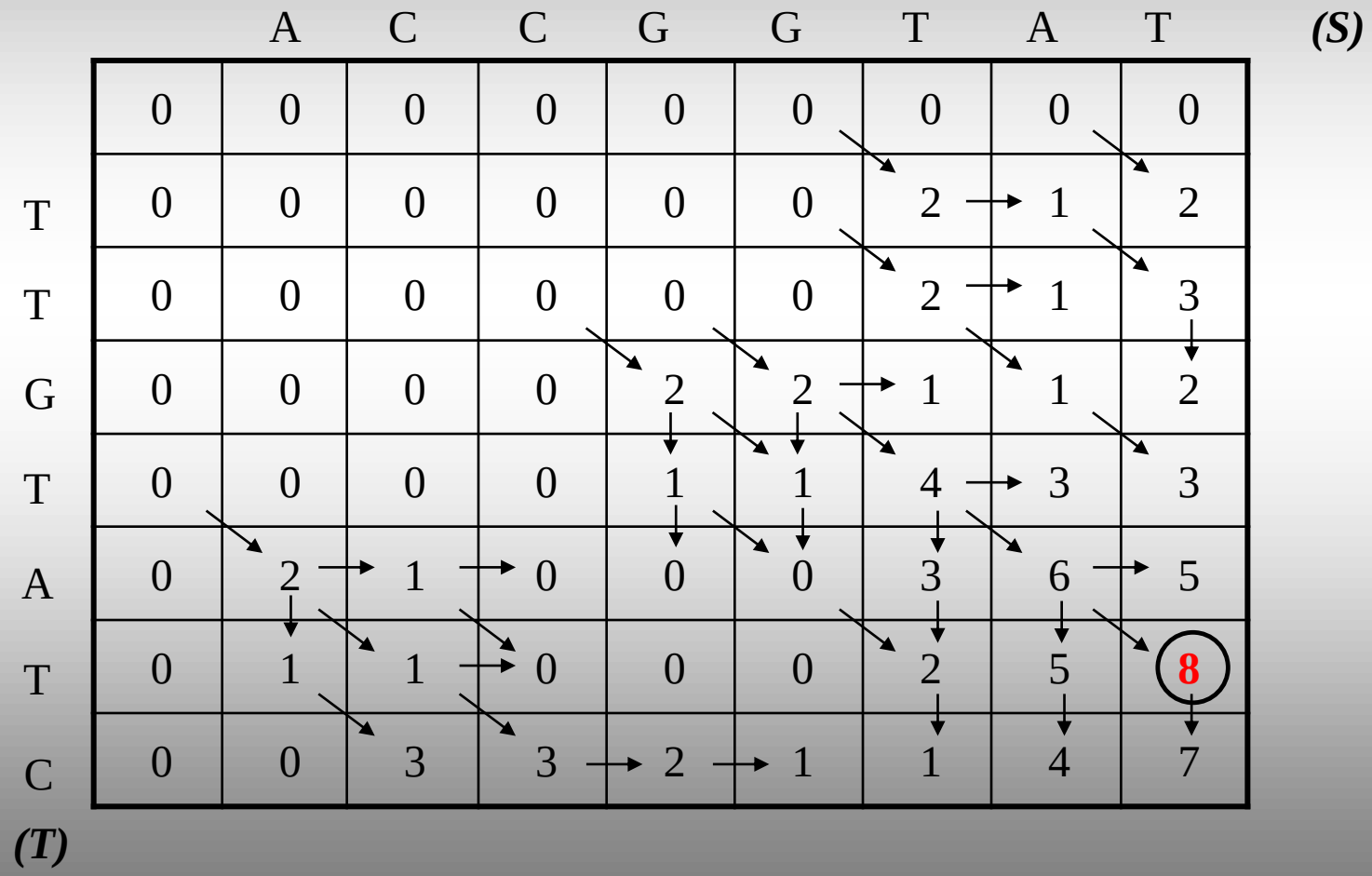
Συμπληρώνουμε τα υπόλοιπα στοιχεία του πίνακα (ελάχιστη τιμή στον πίνακα είναι το 0) :

		A	C	C	G	G	T	A	T
	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	2	1	2
T	0	0	0	0	0	0	2	1	3
G	0	0	0	0	2	2	1	1	2
T	0	0	0	0	1	1	4	3	3
A	0	2	1	0	0	0	3	6	5
T	0	1	1	0	0	0	2	5	8
C	0	0	3	3	2	1	1	4	7

(S)

(T)

- Αφού συμπληρωθεί ο πίνακας, βρίσκουμε, σε οποιαδήποτε θέση του πίνακα, το στοιχείο το οποίο έχει τη μεγαλύτερη τιμή. Στο παράδειγμά μας το στοιχείο αυτό είναι το *V(9, 7) = 8.*



- Με αφετηρία το στοιχείο αυτό και ακολουθώντας τους δείκτες, κινούμαστε αναδρομικά, μέχρις ότου συναντήσουμε στοιχείο με την τιμή 0 και το οποίο μπορεί να βρίσκεται σε οποιαδήποτε θέση του πίνακα.

G T A T *(S)*
| | | |
G T A T *(T)*

		A	C	C	G	G	T	A	T	*(S)*
	0	0	0	0	0	0	0	0	0	
T	0	0	0	0	0	0	2	1	2	
T	0	0	0	0	0	0	2	1	3	
G	0	0	0	0	2	**2**	1	1	2	
T	0	0	0	0	1	1	**4**	3	3	
A	0	2	1	0	0	0	3	**6**	5	
T	0	1	1	0	0	0	2	5	**8**	
C	0	0	3	3	2	1	1	4	7	
(T)										

Παράδειγμα Smith - Waterman

- Η ολική βαθμολογία που αποδίδεται στην αντιστοίχηση υποσειρών είναι η τιμή του στοιχείου απ' όπου ξεκινήσαμε την αναδρομή (η υψηλότερη τιμή στοιχείου στον πίνακα).

  Στο παράδειγμά μας έχουμε 4 ταυτίσεις, επομένως η βαθμολογία της αντιστοίχησης είναι 4x2 = 8, τιμή που συμπίπτει με αυτή που προκύπτει από τον πίνακα.

- Σε ορισμένες περιπτώσεις ενδέχεται το αποτέλεσμα της τοπικής αντιστοίχησης μεταξύ δύο σειρών να είναι υποσύνολο της ολικής αντιστοίχησης των ιδίων σειρών, όμως κάτι τέτοιο δεν πρέπει να θεωρείται πάντα δεδομένο

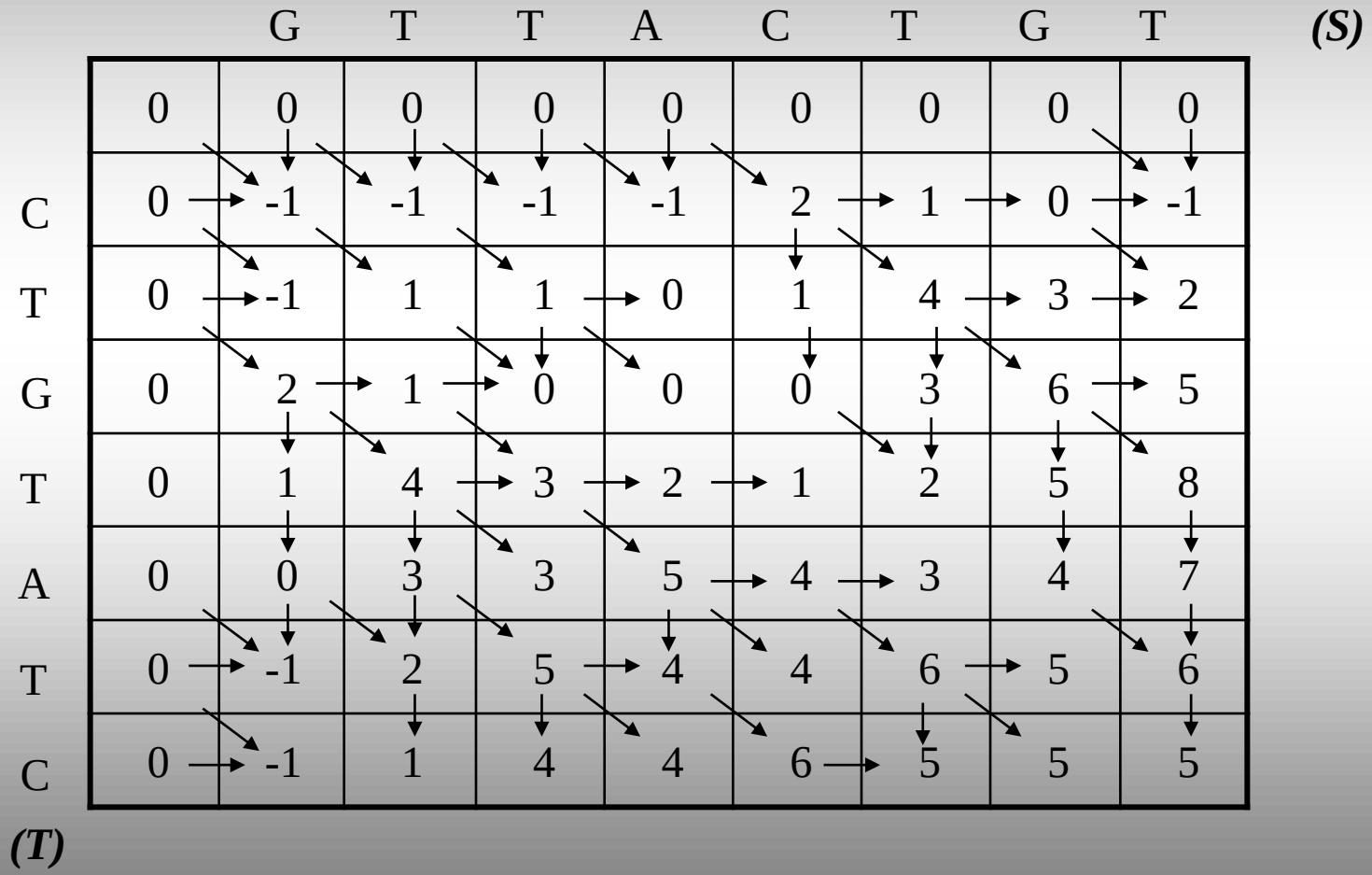
Ends – free alignment

- Σε περιπτώσεις που η μία σειρά από τις δύο που θέλουμε να αντιστοιχίσουμε περιέχει την άλλη, ή ορισμένες περιοχές των δύο σειρών τυγχάνει να επικαλύπτονται, η ολική και η τοπική αντιστοίχηση δεν αποτελούν καλή επιλογή (π.χ. όταν συγκρίνουμε τμήματα σειρών DNA μεταξύ τους ή με μεγαλύτερες σειρές χρωμοσωμάτων).
- Στις περιπτώσεις αυτές χρησιμοποιείται η ends – free αντιστοίχηση.
- Ο πίνακας βαθμολόγησης στην ends – free αντιστοίχηση συμπληρώνεται με τον ίδιο αναδρομικό μοντέλο που εφαρμόστηκε και στην ολική αντιστοίχηση.
- Για πίνακα βαθμολόγησης *V*, γνωρίζοντας τις τιμές των *V(i-1, j-1)*, *V(i-1, j)* και *V(i, j-1)* οι τιμές των στοιχείων του προκύπτουν ως εξής:

$$V(i, 0) = 0,\ V(0, j) = 0, \text{ για κάθε } i, j$$

$$V(i, j) = \max \begin{cases} V(i-1, j-1) + \sigma(Si, Tj) \\ V(i-1, j) + \sigma(Si, -) \\ V(i, j-1) + \sigma(-, Tj) \end{cases}$$

Παράδειγμα ends – free αντιστοίχησης



Παράδειγμα ends – free αντιστοίχησης

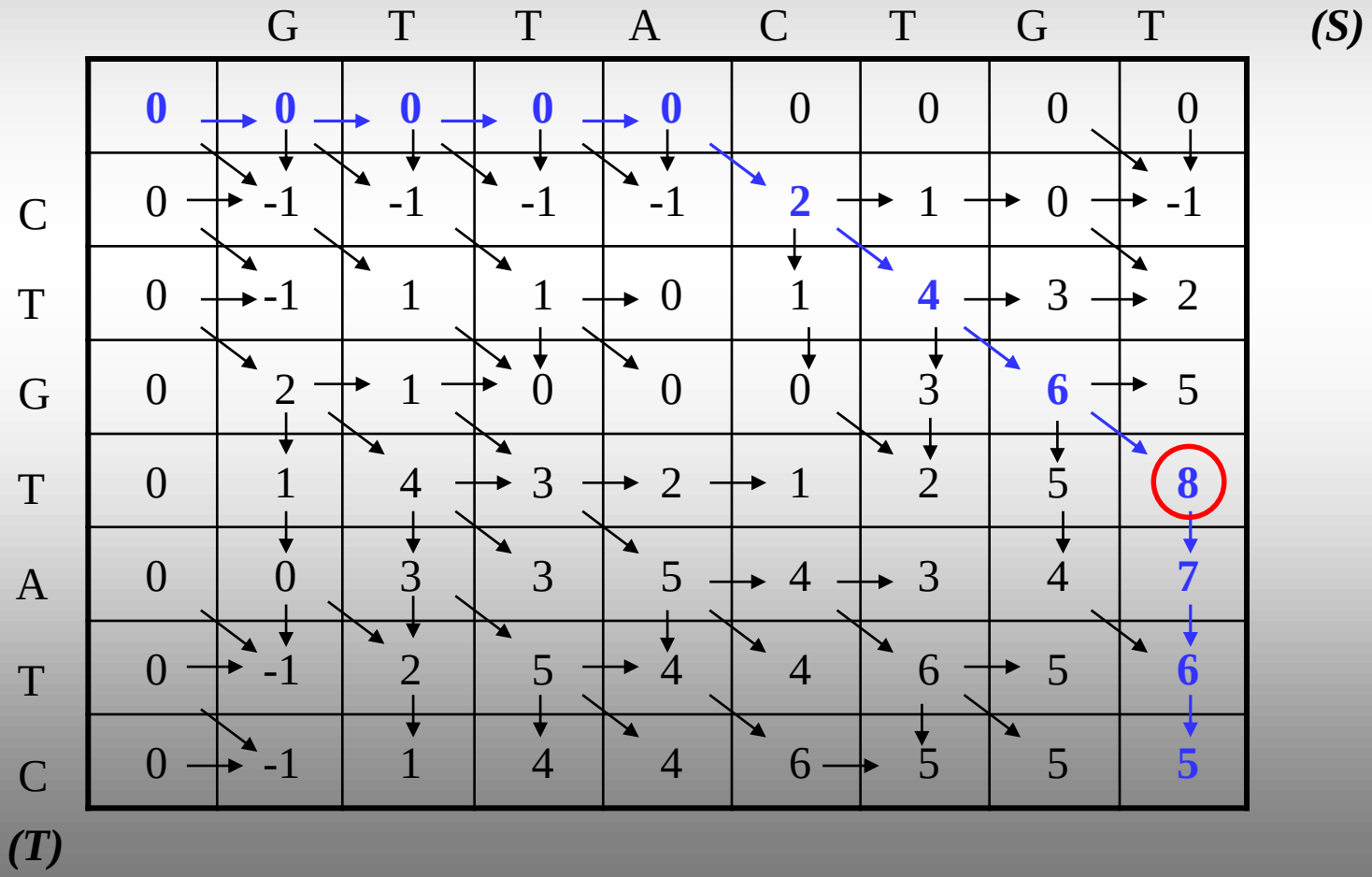
- Για το βέλτιστο μονοπάτι εντοπίζουμε το ακραίο στοιχείο του πίνακα (είτε στην ακραία γραμμή του πίνακα είτε στην ακραία στήλη του) με την καλύτερη τιμή και χρησιμοποιώντας αυτό ως αφετηρία κινούμαστε αναδρομικά και ‘χτίζουμε’ την αντιστοιχία.

- Τα κενά (indels / gaps) επιτρέπονται στην αρχή και στο τέλος των σειρών χωρίς την επιβολή ποινής σε αυτά.

G T T A C T G T - - - ***(S)***

| | | | |

- - - - C T G T A T C ***(T)***



Πίνακες Αντικατάστασης (Substitution Matrices)

- Κάθε πίνακας αντικατάστασης αναπαριστά μία ξεχωριστή εξελικτική θεωρία και αντιστοιχεί σε μία συγκεκριμένη εξελικτική απόσταση.

- Οι τιμές των στοιχείων που αποθηκεύονται σε ένα πίνακα αντικατάστασης αναπαριστούν είτε την **ομοιότητα** – πόσο 'κοντινό' είναι δηλαδή ένα αμινοξύ με αυτό που αντικατέστησε στη σειρά – είτε την **απόσταση** – ποιο είναι το κόστος από την αντικατάσταση ενός αμινοξέως με ένα άλλο. Δηλαδή, βαθμολογούν τις αντικαταστάσεις που συναντόνται σε μία αντιστοίχηση.

- Η λογική πίσω και από τις δύο αυτές προσεγγίσεις είναι οι ίδια, επομένως οι πίνακες αντικατάστασης θα έχουν σχετικά σταθερή μορφή.

Πίνακες PAM (Point Accepted Mutation)

Οι Dayhoff, Schwarz και Orcutt (1978) χρησιμοποιήσαν 71 οικογένειες πρωτεϊνών με ομοιότητα γύρω στο 85% (δηλ., οι σειρές των πρωτεϊνών διαφέρουν το πολύ κατά το 15% των residues τους).

Αντιστοίχησαν τις πρωτεϊνες αυτές και, αγνοώντας την εξελικτική κατεύθυνση, 'έχτισαν' ένα θεωρητικό φυλογενετικό δένδρο (phylogenetic tree – μία γραφική απεικόνιση των εξελικτικών σχέσεων μίας ομάδας οργανισμών).

Βασιζόμενοι σε 1572 αλλαγές residues, κατέγραψαν τη συχνότητα αντικατάστασης ενός residue X από ένα residue Y μέσα σε χρόνο Z. Προέβλεψαν έτσι τα residues τα οποία έχουν τη μεγαλύτερη πιθανότητα να εμφανιστούν στις προγονικές σειρές.

1PAM : ο πρώτος πίνακας PAM και απευθύνεται σε σειρές όπου ο αριθμός των αποδεκτών μεταλλάξεων σε αυτές αποτελεί το 1% του συνολικού μήκους τους (point accepted mutation).

Προκειμένου να αυξηθεί η επιτρεπόμενη απόσταση, ο πίνακας PAM1 μπορεί να πολλαπλασιαστεί και να χρησιμοποιηθούν τα πολλαπλάσιά του. Η πιο διαδεδομένη έκδοση που χρησιμοποιείται είναι ο **PAM250**.

Επιλέγοντας δηλαδή έναν πίνακα PAM με μεγαλύτερη τιμή, επιτρέπουμε αντιστοιχίσεις σειρών με μεγαλύτερη εξελικτική απόσταση.

Πίνακες PAM (Point Accepted Mutation)

Οι πίνακες PAM είναι καλή επιλογή για αντιστοίχηση σειρών με στενή συγγένεια.

Βασίζονται σε δεδομένα όπου οι αντικαταστάσεις είναι πιο πιθανό να συμβούν από αλλαγές μίας μόνο βάσης στα codons.

Κλίνουν προς συντηρητικές μεταλλάξεις στην αλληλουχία του DNA (αντί για αντικαταστάσεις αμινοξέων) οι οποιές επηρεάζουν σε μικρό βαθμό την λειτουργία / δομή.

Μία αντικατάσταση σε κάποιο σημείο της σειράς εξαρτάται αποκλειστικά από το αμινοξύ στο συγκεκριμένο σημείο και από την πιθανότητα που δίνεται από τον πίνακα. Έτσι δεν γίνεται σωστή απεικόνιση των εξελικτικών διαδικασιών, καθώς σειρές με μακρυνή συγγένεια συνήθως έχουν περιοχές υψηλής συντήρησης (blocks).

Νέα έκδοση του PAM δημιουργήθηκε το 1992 από τους Jones, Taylor και Thornton, οι οποίοι χρησιμοποίησαν 59190 αντικαταστάσεις.

Πίνακες BLOSUM (Blocks Substitution Matrix)

Για την κατασκευή των πινάκων BLOSUM ο Henikoff (1991) χρησιμοποίησε σύνολα περιοχών χωρίς κενά. Οι περιοχές αυτές ανήκουν σε οικογένειες πρωτεϊνών που περιέχονται στη βάση BLOCKS (η ΒΔ BLOCKS περιλαμβάνει ομαδοποιημένες (clustered) σύντομες σειρές πρωτεϊνών οι οποίες παρουσιάζουν μεγάλη ομοιότητα. Οι ομάδες αυτές προκύπτουν από τη βάση SWISS-PROT και άλλες βάσεις, εφαρμόζοντας σε αυτές τον αλγόριθμο MOTIF. Στην παρούσα έκδοση περιλαμβάνονται 8656 Blocks πρωτεϊνών.

Στην ίδια ομάδα ανήκουν σειρές οι οποίες έχουν ποσοστό ομοιότητας που υπερβαίνει κάποιο προκαθορισμένο όριο.

Υπολογίζεται και η συχνότητα με την οποία δύο residues τα οποία έχουν αντιστοιχιστεί σε μία ομάδα να τύχει να αντιστοιχιστούν και σε μία άλλη.

Το αποτέλεσμα που προκύπτει είναι ο λόγος των κατεγραμμένων αντικαταστάσεων μεταξύ δύο οποιονδήποτε residues προς όλες τις αντικαταστάσεις που έχουν καταγραφεί, δηλαδή η πιθανότητα ένα συγκεκριμένο residue να αντικατασταθεί από ένα άλλο συγκεκριμένο residue.

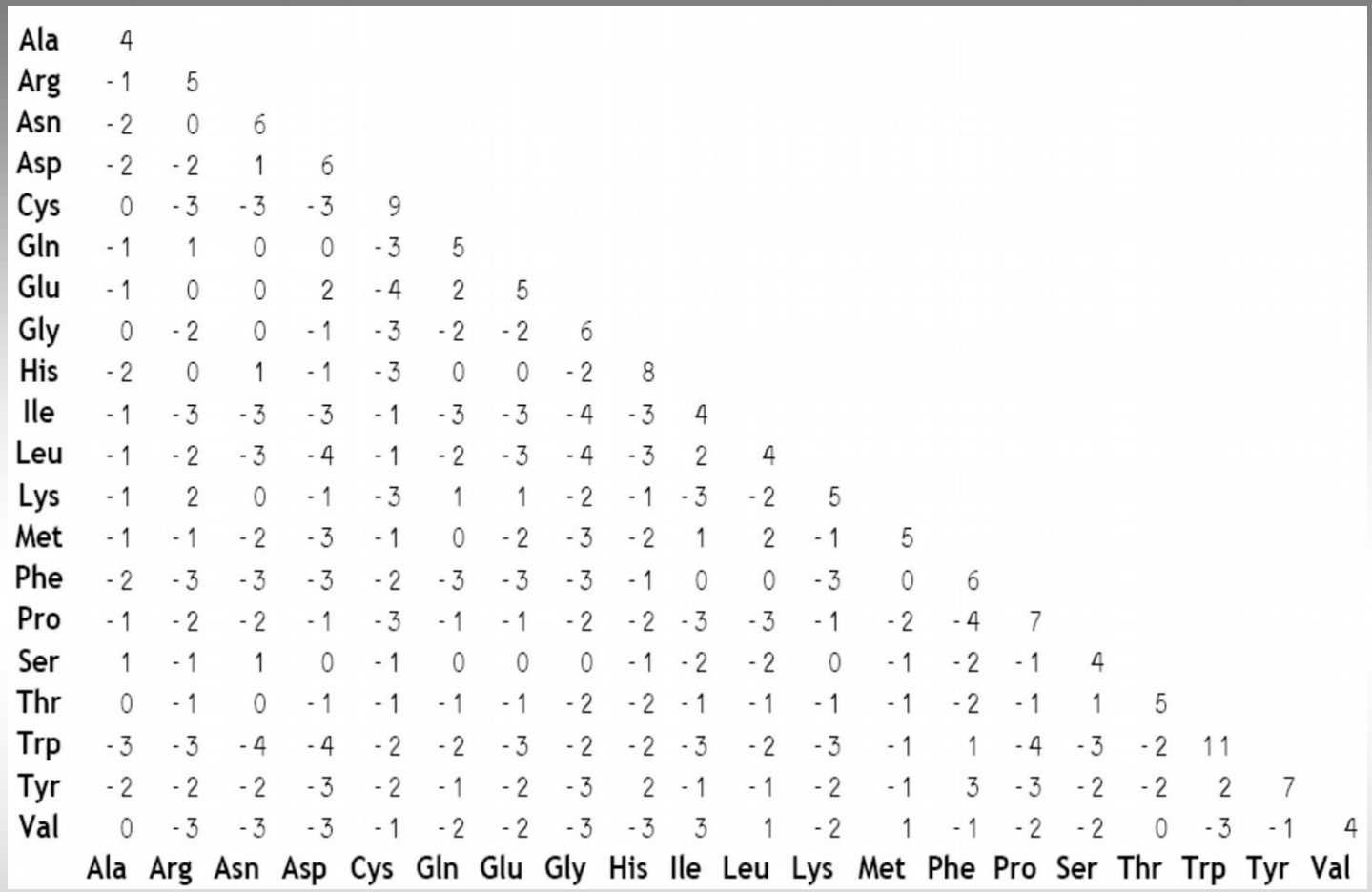
Πίνακες BLOSUM (Blocks Substitution Matrix)

Οι πίνακες BLOSUM είναι η καλύτερη επιλογή για τον εντοπισμό ασθενών πρωτεϊνικών πιθανοτήτων.

Όπως και με τους πίνακες PAM, υπάρχουν πολλές εκδόσεις των πινάκων BLOSUM, με τη διαφορά ότι η αρίθμησή τους είναι αντίστροφη από αυτή των πινάκων PAM.

Έτσι, π.χ. ο πίνακας BLOSUM50 περιλαμβάνει ομάδες σειρών με τουλάχιστον 50% ομοιότητα, ενώ ο BLOSUM62 περιλαμβάνει ομάδες σειρών με τουλάχιστον 62% ομοιότητα.

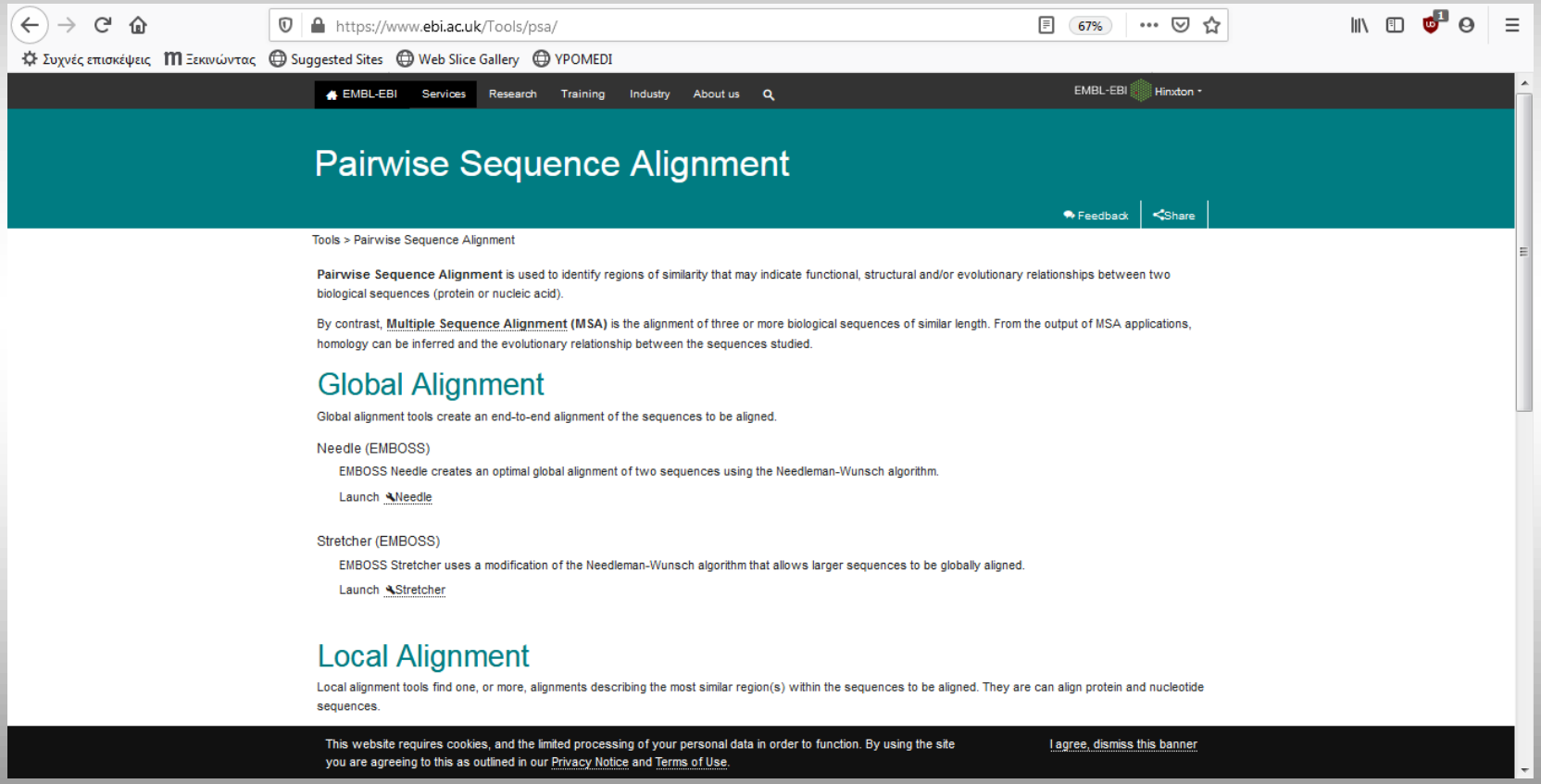
Επιλέγοντας δηλαδή έναν πίνακα BLOSUM με μεγαλύτερη τιμή, επιτρέπουμε αντιστοιχίσεις σειρών με μεγαλύτερο ποσοστό ομοιότητας.

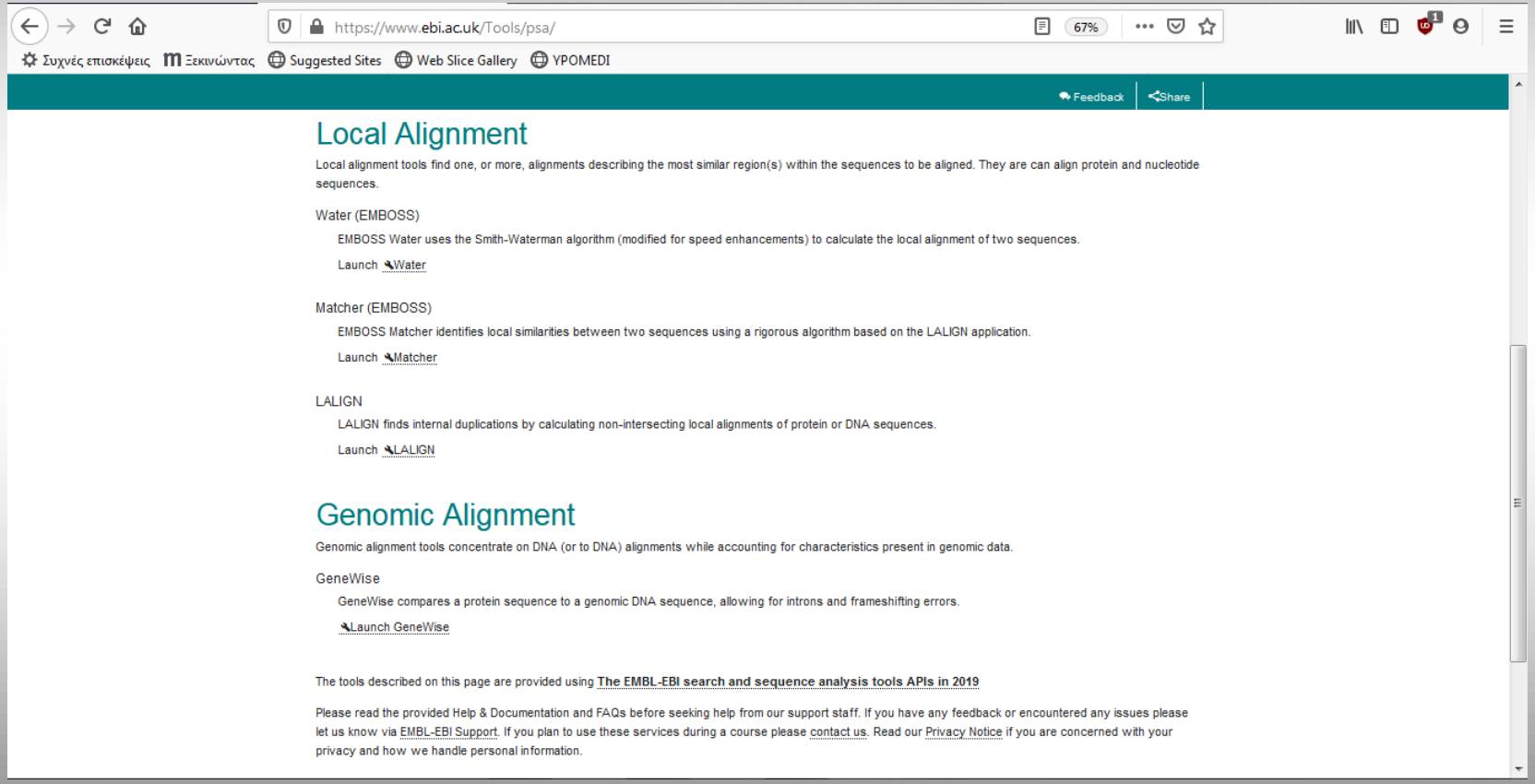


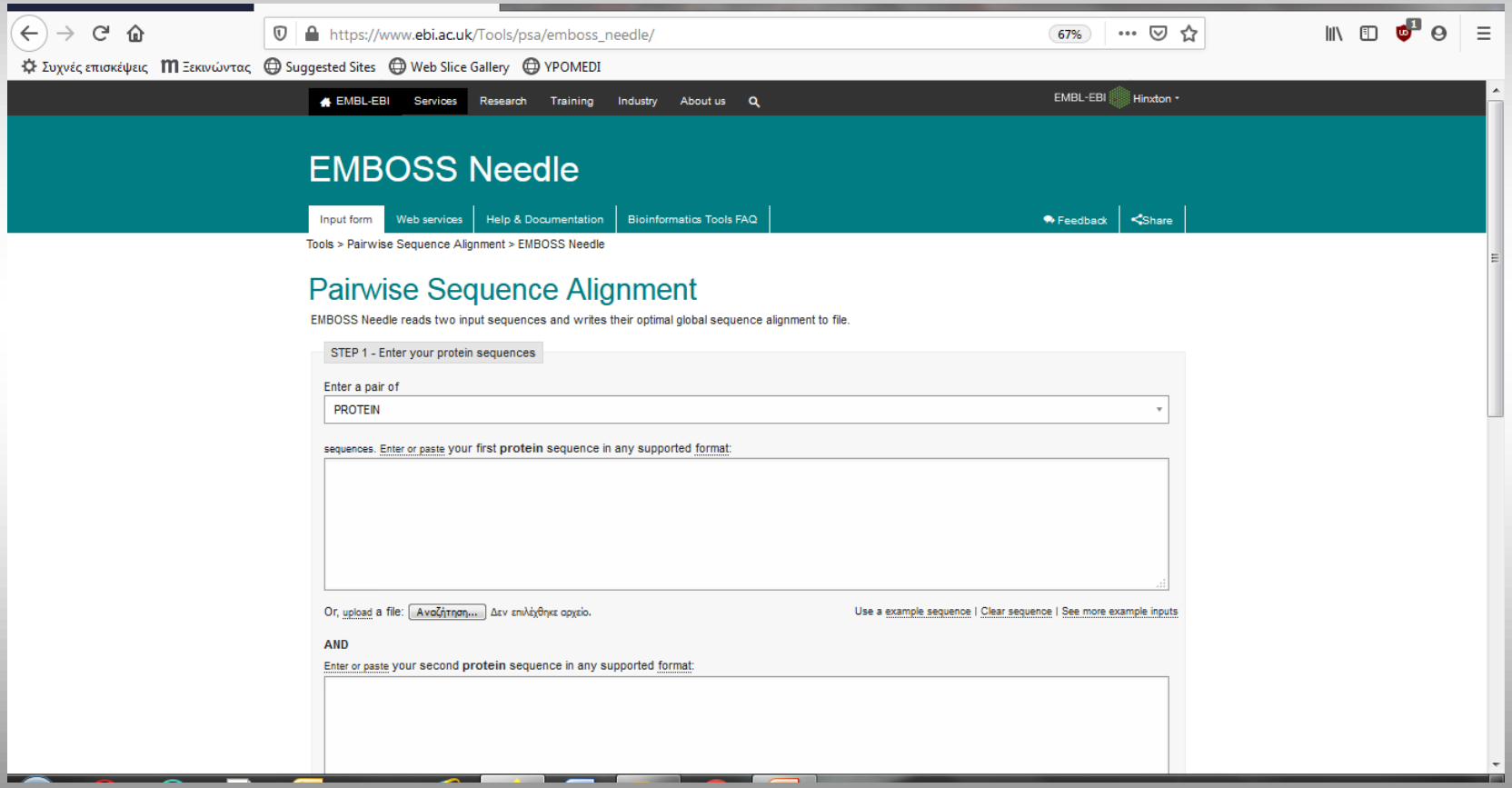
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

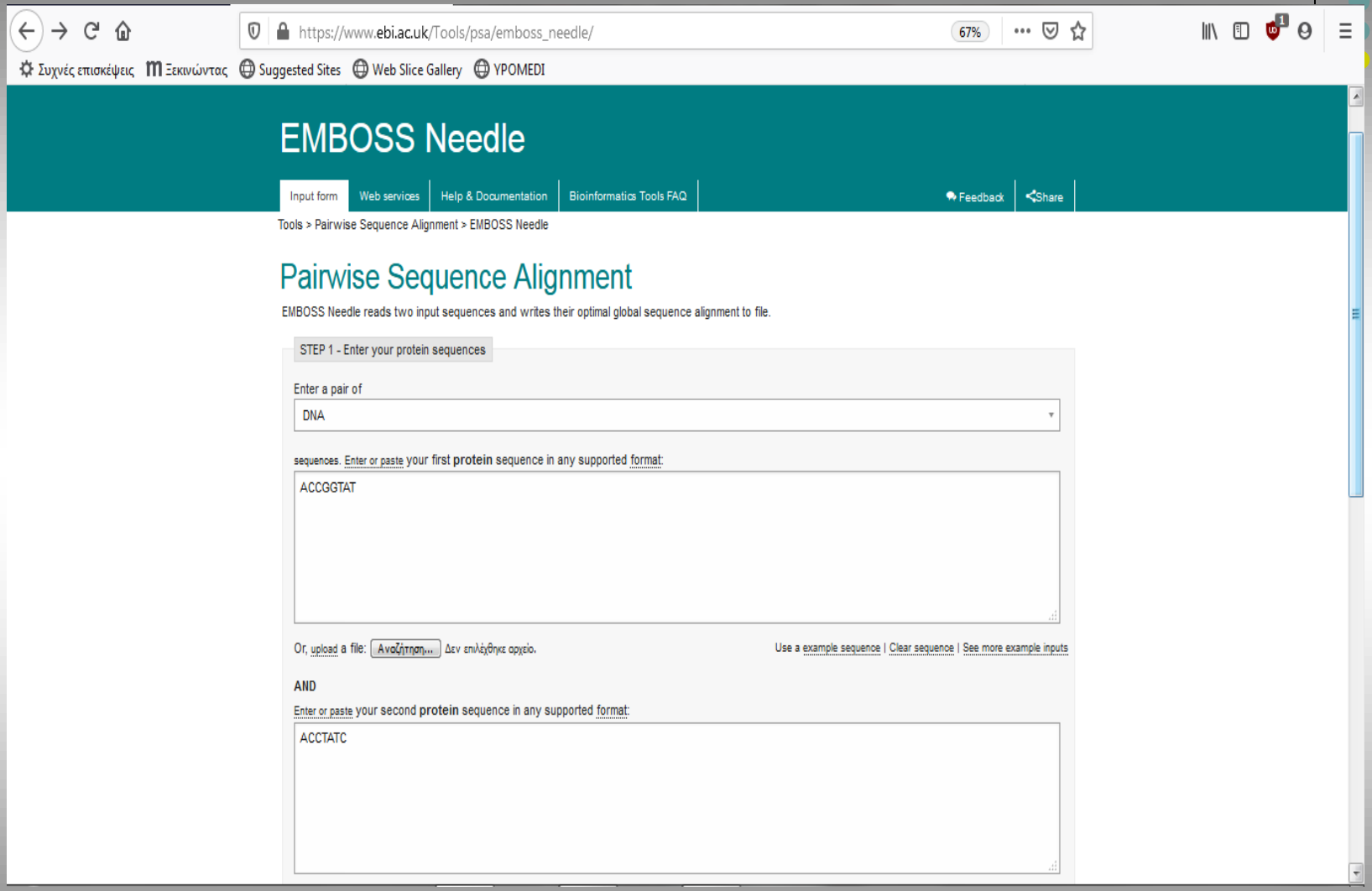
- Ο Πίνακας Αντικατάστασης BLOSUM62. (Εικόνα από Hannes Röst, CC BY SA 2.5, από Wikimedia Commons).

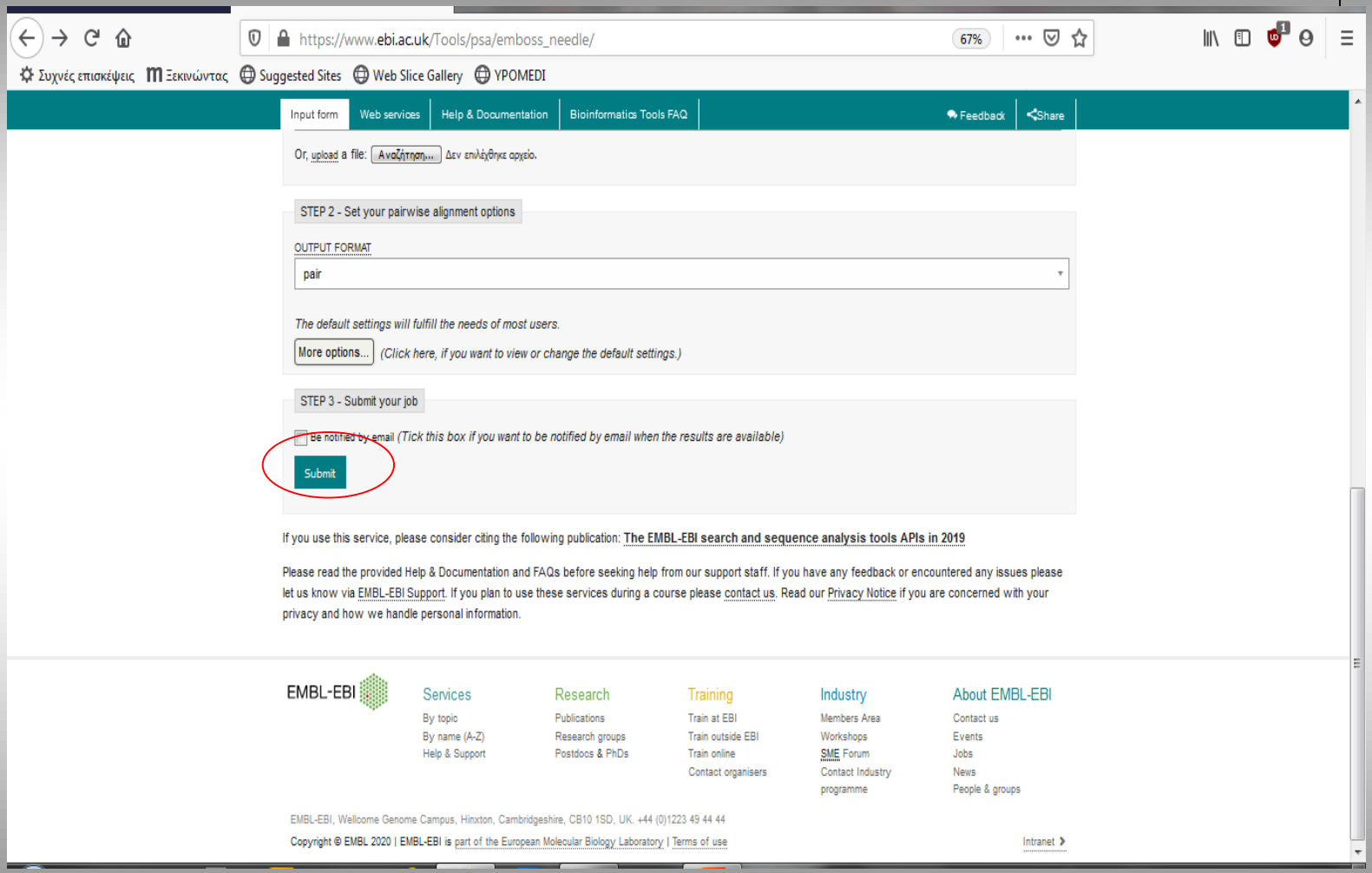
Μπορουμε να χρησιμοποιησουμε την http://www.ebi.ac.uk/Tools/psa/

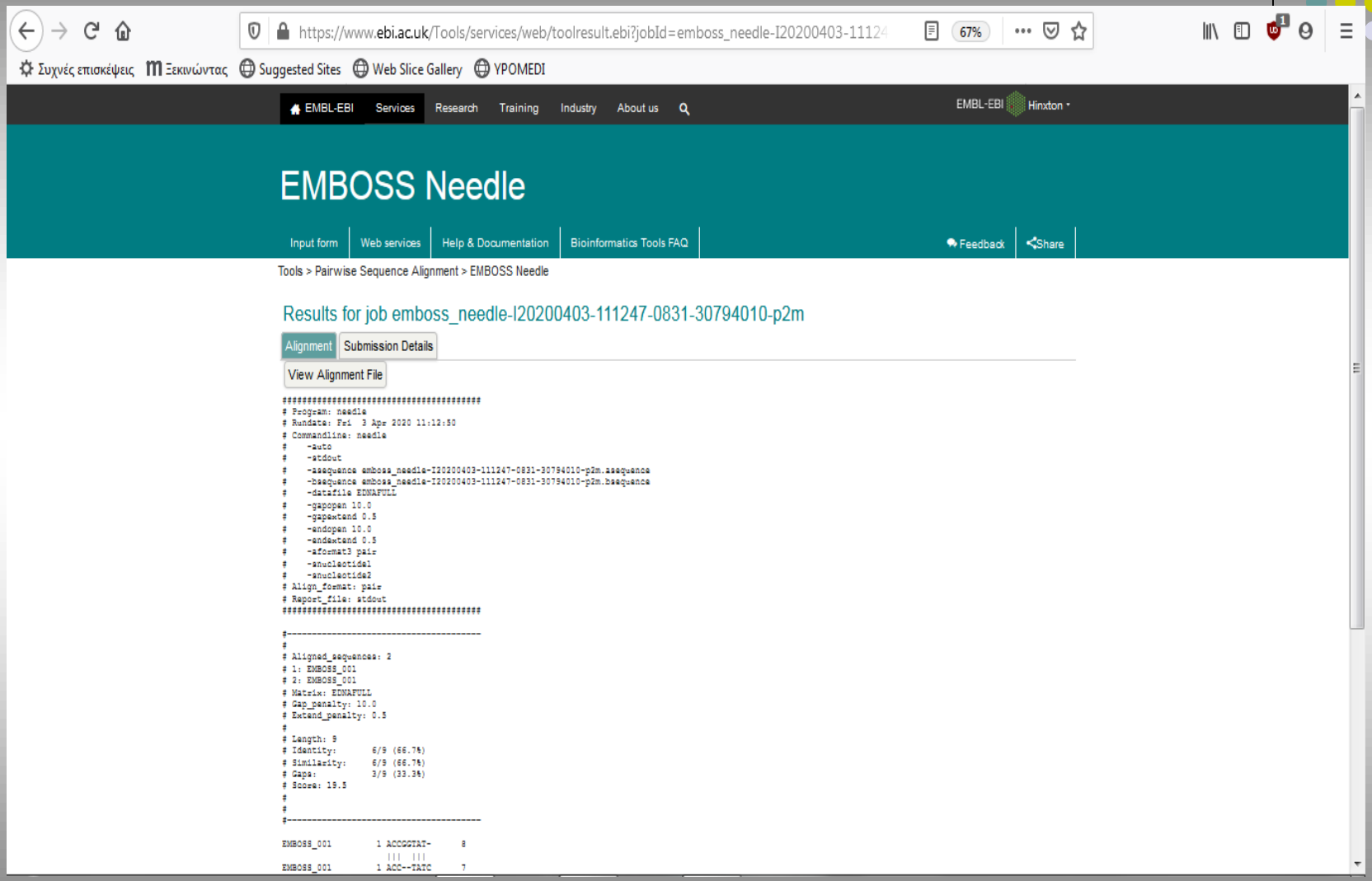


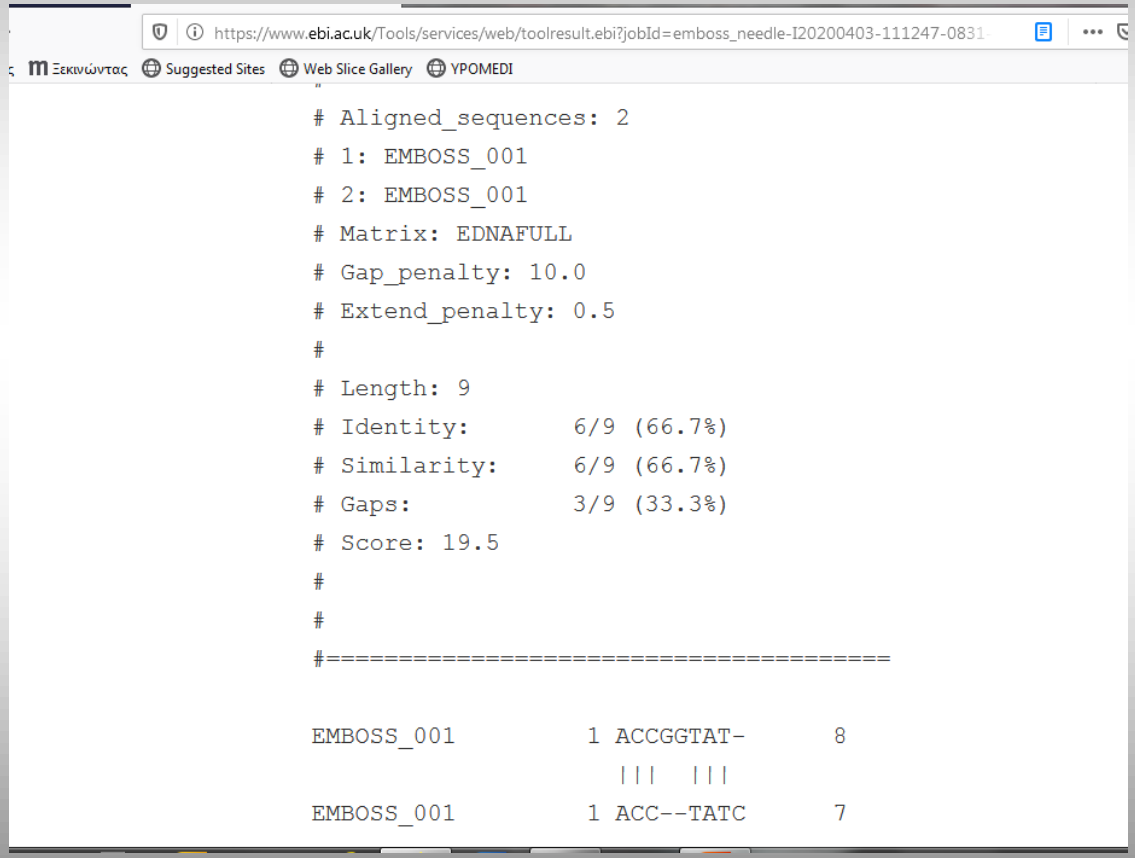

https://www.ebi.ac.uk/Tools/psa/
67%
Συχνές επισκέψεις
Ξεκινώντας
Suggested Sites
Web Slice Gallery
YPOMEDI
Feedback
Share
Local Alignment
Local alignment tools find one, or more, alignments describing the most similar region(s) within the sequences to be aligned. They are can align protein and nucleotide sequences.
Water (EMBOSS)
EMBOSS Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.
Launch Water
Matcher (EMBOSS)
EMBOSS Matcher identifies local similarities between two sequences using a rigorous algorithm based on the LALIGN application.
Launch Matcher
LALIGN
LALIGN finds internal duplications by calculating non-intersecting local alignments of protein or DNA sequences.
Launch LALIGN
Genomic Alignment
Genomic alignment tools concentrate on DNA (or to DNA) alignments while accounting for characteristics present in genomic data.
GeneWise
GeneWise compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.
Launch GeneWise
The tools described on this page are provided using The EMBL-EBI search and sequence analysis tools APIs in 2019
Please read the provided Help & Documentation and FAQs before seeking help from our support staff. If you have any feedback or encountered any issues please let us know via EMBL-EBI Support. If you plan to use these services during a course please contact us. Read our Privacy Notice if you are concerned with your privacy and how we handle personal information.


https://www.ebi.ac.uk/Tools/psa/emboss_needle/
67%
Συχνές επισκέψεις
Ξεκινώντας
Suggested Sites
Web Slice Gallery
YPOMEDI
EMBL-EBI
Services
Research
Training
Industry
About us
EMBL-EBI Hinxton
EMBOSS Needle
Input form
Web services
Help & Documentation
Bioinformatics Tools FAQ
Feedback
Share
Tools > Pairwise Sequence Alignment > EMBOSS Needle
Pairwise Sequence Alignment
EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.
STEP 1 - Enter your protein sequences
Enter a pair of
PROTEIN
sequences. Enter or paste your first protein sequence in any supported format:
Or, upload a file: Αναζήτηση... Δεν επιλέχθηκε αρχείο.
Use a example sequence | Clear sequence | See more example inputs
AND
Enter or paste your second protein sequence in any supported format:


https://www.ebi.ac.uk/Tools/psa/emboss_needle/
67%
Συχνές επισκέψεις
Ξεκινώντας
Suggested Sites
Web Slice Gallery
YPOMEDI
EMBOSS Needle
Input form
Web services
Help & Documentation
Bioinformatics Tools FAQ
Feedback
Share
Tools > Pairwise Sequence Alignment > EMBOSS Needle
Pairwise Sequence Alignment
EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.
STEP 1 - Enter your protein sequences
Enter a pair of
DNA
sequences. Enter or paste your first protein sequence in any supported format:
ACCGGTAT
Or, upload a file: Αναζήτηση... Δεν επιλέχθηκε αρχείο.
Use a example sequence | Clear sequence | See more example inputs
AND
Enter or paste your second protein sequence in any supported format:
ACCTATC


https://www.ebi.ac.uk/Tools/psa/emboss_needle/
67%
Συχνές επισκέψεις
Ξεκινώντας
Suggested Sites
Web Slice Gallery
YPOMEDI
Input form
Web services
Help & Documentation
Bioinformatics Tools FAQ
Feedback
Share
Or, upload a file:
Αναζήτηση...
Δεν επιλέχθηκε αρχείο.
STEP 2 - Set your pairwise alignment options
OUTPUT FORMAT
pair
The default settings will fulfill the needs of most users.
More options...
(Click here, if you want to view or change the default settings.)
STEP 3 - Submit your job
Be notified by email (Tick this box if you want to be notified by email when the results are available)
Submit
If you use this service, please consider citing the following publication: The EMBL-EBI search and sequence analysis tools APIs in 2019
Please read the provided Help & Documentation and FAQs before seeking help from our support staff. If you have any feedback or encountered any issues please let us know via EMBL-EBI Support. If you plan to use these services during a course please contact us. Read our Privacy Notice if you are concerned with your privacy and how we handle personal information.
EMBL-EBI
Services
By topic
By name (A-Z)
Help & Support
Research
Publications
Research groups
Postdocs & PhDs
Training
Train at EBI
Train outside EBI
Train online
Contact organisers
Industry
Members Area
Workshops
SME Forum
Contact Industry programme
About EMBL-EBI
Contact us
Events
Jobs
News
People & groups
EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. +44 (0)1223 49 44 44
Copyright © EMBL 2020 | EMBL-EBI is part of the European Molecular Biology Laboratory | Terms of use
Intranet


https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=emboss_needle-I20200403-111247
67%
Συχνές επισκέψεις
Ξεκινώντας
Suggested Sites
Web Slice Gallery
YPOMEDI
EMBL-EBI
Services
Research
Training
Industry
About us
EMBL-EBI Hinxton
EMBOSS Needle
Input form
Web services
Help & Documentation
Bioinformatics Tools FAQ
Feedback
Share
Tools > Pairwise Sequence Alignment > EMBOSS Needle
Results for job emboss_needle-I20200403-111247-0831-30794010-p2m
Alignment
Submission Details
View Alignment File
##
Program: needle
Rundate: Fri 3 Apr 2020 11:12:50
Commandline: needle
-auto
-stdout
-asequence emboss_needle-I20200403-111247-0831-30794010-p2m.asequence
-bsequence emboss_needle-I20200403-111247-0831-30794010-p2m.bsequence
-datafile EDNAFULL
-gapopen 10.0
-gapextend 0.5
-endopen 10.0
-endextend 0.5
-aformat3 pair
-snucleotide1
-snucleotide2
Align_format: pair
Report_file: stdout
##
#---------------------------------------
#
Aligned_sequences: 2
1: EMBOSS_001
2: EMBOSS_001
Matrix: EDNAFULL
Gap_penalty: 10.0
Extend_penalty: 0.5
#
Length: 9
Identity: 6/9 (66.7%)
Similarity: 6/9 (66.7%)
Gaps: 3/9 (33.3%)
Score: 19.5
#
#
#---------------------------------------
EMBOSS_001 1 ACCGGTAT- 8
||| |||
EMBOSS_001 1 ACC--TATC 7


https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=emboss_needle-I20200403-111247-0831
Ξεκινώντας
Suggested Sites
Web Slice Gallery
YPOMEDI
Aligned_sequences: 2
1: EMBOSS_001
2: EMBOSS_001
Matrix: EDNAFULL
Gap_penalty: 10.0
Extend_penalty: 0.5
#
Length: 9
Identity: 6/9 (66.7%)
Similarity: 6/9 (66.7%)
Gaps: 3/9 (33.3%)
Score: 19.5
#
#
#=======================================
EMBOSS_001 1 ACCGGTAT- 8
||| |||
EMBOSS_001 1 ACC--TATC 7