Στοίχιση ακολουθιών κατά ζεύγη (Pairwise alignment)

Στοίχιση κατά ζεύγη: Τι είναι

- Αντιστοίχιση των νουκλεοτιδίων/αμινοξέων δυο ακολουθιών, ώστε να εντοπιστούν οι ομοιότητες και οι διαφορές τους.
- Χρησιμοποιείται για:
  - Εντοπισμό μεταλλάξεων
  - αναζήτηση ομόλογων γονιδίων/πρωτεϊνών σε βάσεις δεδομένων.
  - Συναρμολόγιση γενωμάτων.
  - Έλεγχος εξειδίκευσης εκκινητών (primers) για PCR.

Στοίχιση κατά ζεύγη: Τι είναι

- Τοποθετούνται οι αντίστοιχοι χαρακτήρες ο ένας κάτω από τον άλλο και μπορεί να γίνει χρήση κενών (gaps)
- Δύο χαρακτήρες μπορεί να είναι:
  - Ίδιοι
  - Παρόμοιοι (κοινές φυσικοχημικές ιδιότητες, π.χ. Ισολευκίνη - βαλίνη)
  - Διαφορετικοί

```
Query  1    MKTPVSAAANLSIQNAGSSGATAIQIIPKTEPVGEEGPMSLDFQSPNLNTSTPNPNKRPG  60
            MKTPVSAAANLS  NAGSSGA AIQI+PKTEPVGEEGPMSLDFQSPNL+TSTPNPNKRPG
Sbjct  1    MKTPVSAAANLSNPNAGSSGAAAIQIVPKTEPVGEEGPMSLDFQSPNLSTSTPNPNKRPG  60

Query  61   SLDLNSKSAKNKRIFAPLVINSPDLSSKTVNTPDLEKILLSNNLMQTPQPGKVFPTKAGP  120
            SLDLNSK AKNKRIFAPLVINSPDL +KTVNTPDLEKILLSNNL+QTPQPGKVFPTKAGP
Sbjct  61   SLDLNSKGAKNKRIFAPLVINSPDLQAKTVNTPDLEKILLSNNLIQTPQPGKVFPTKAGP  120

Query  121  VTVEQLDFGRGFEEALHNLHTNSQAFPSANSAANSAANNTTAAAMTAVNNGISGGTFTYT  180
            VTVEQ DFGRGFEEAL NLHTNSQAFP    A NS ANNTT  AMTAVNNGISGGTFTY
Sbjct  121  VTVEQEDFGRGFEEALKNLHTNSQAFP----AVNSTANNTTGTAMTAVNNGISGGTFTY-  175
```

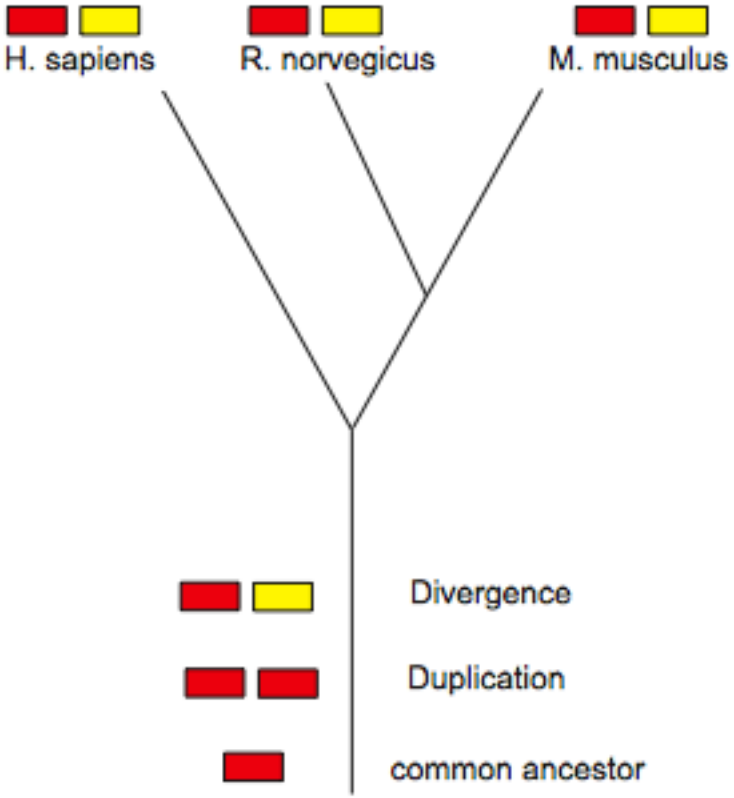
Στοίχιση κατά ζεύγη: Τι είναι

- Για δύο ακολουθίες με 95% ομοιότητα, η στοίχιση μπορεί να γίνει και με το μάτι.
- Τα διαθέσιμα προγράμματα αγγίζουν τα όρια των δυνατοτήτων τους όταν οι ακολουθίες έχουν 18-25% ομοιότητα (ζώνη του λυκόφωτος)

Λίγη εξέλιξη: ομολογία

- Ομόλογα γονίδια: κοινός εξελικτικός πρόγονος. Χιμαιρικές πρωτεΐνες;
- Ορθόλογα γονίδια: προέρχονται από ειδογένεση. Ουσιαστικά, ένα γονίδιο α (μεταλλαγμένο) σε δύο διαφορετικούς οργανισμούς. Συχνά έχουν την ίδια λειτουργία
- Παράλογα γονίδια: προέρχονται από γονιδιακό διπλασιασμό. Ανήκουν στην ίδια οικογένεια
- Ξενόλογα γονίδια: από οριζόντια μεταφορά
- Παράδειγμα με Πυρηνικούς υποδοχείς

Λίγη εξέλιξη: ομολογία (ii)



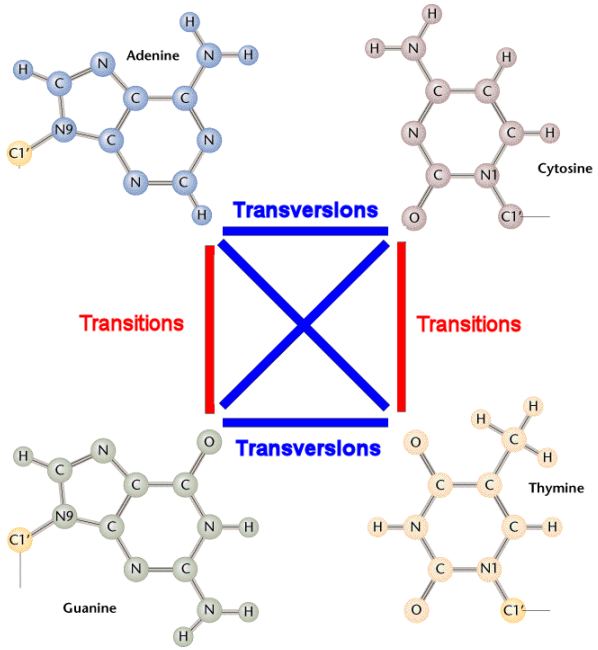
- Γονίδια του ίδιου χρώματος από διαφορετικούς οργανισμούς είναι ορθόλογα.
- Το κόκκινο και το κίτρινο από ένα οργανισμό είναι παράλογα.
- Το κόκκινο από ένα οργανισμό και το κίτρινο από ένα άλλο οργανισμό είναι έξτρα-παράλογα

Βασικότερα είδη μεταλλάξεων

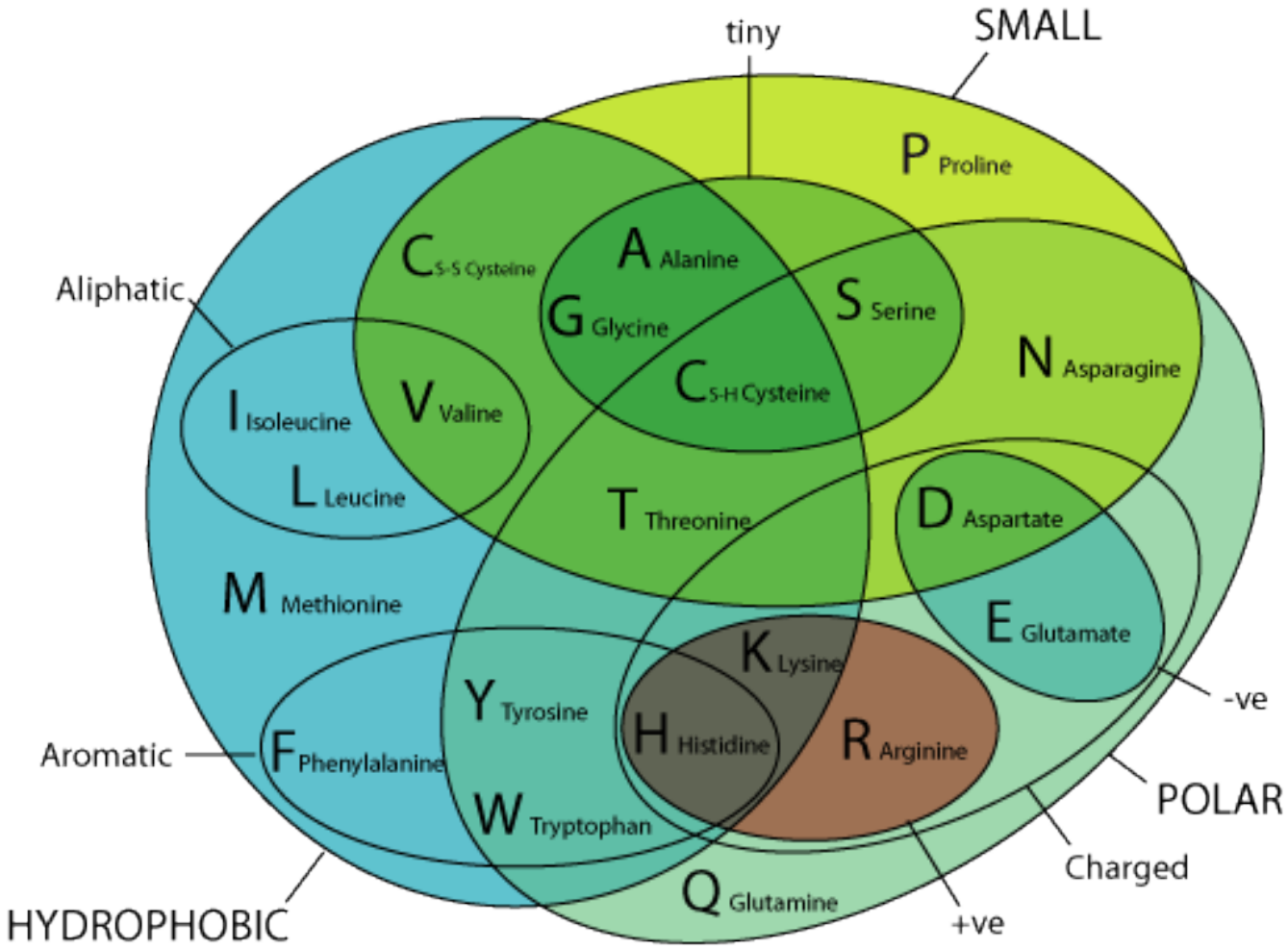
- Μεταλλάξεις σημείου (point mutations)
  - Συνώνυμες (synonymous)
  - Μη-συνώνυμες (non-synonymous)
    - Αμινοξέα με παρόμοιες φυσικοχημικές ιδιότητες
    - Αμινοξέα με διαφορετικές φυσικοχημικές ιδιότητες
  - Κωδικόνια τερματισμού

Μεταπτώσεις-μεταστροφές

- Μεταπτώσεις (Transitions)
  - Δημιουργούνται με μεγαλύτερη συχνότητα
  - Συνήθως οδηγούν σε συνώνυμες μεταλλάξεις
  - Είναι πιο συχνές στα SNPs



Κατηγοριοποίηση αμινοξέων

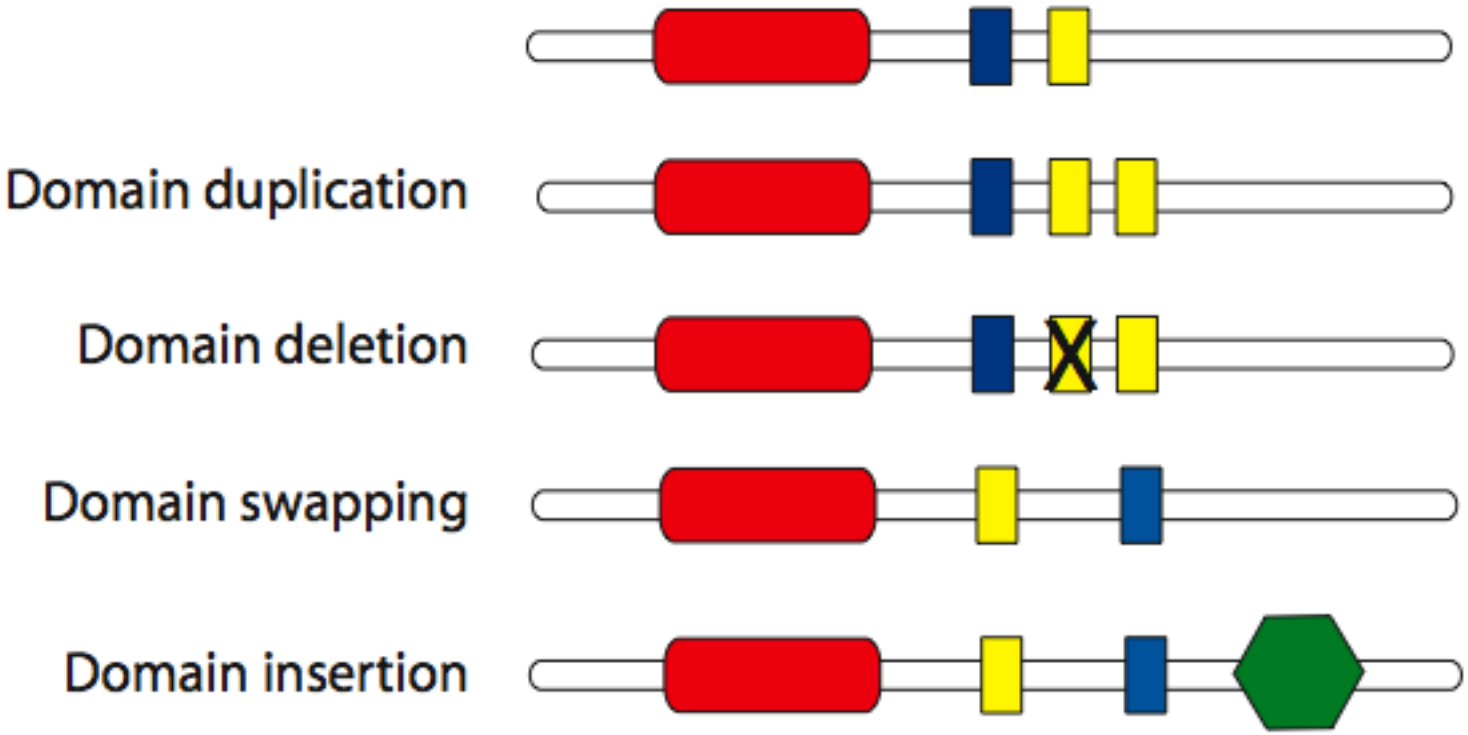


Βασικότερα είδη μεταλλάξεων

- Δομικές Αναδιατάξεις
  - Προσθήκες/απαλείψεις (insertions/deletions)
  - Αναστροφές
  - Διπλασιασμοί

Βασικότερα είδη μεταλλάξεων (ii)

- Αναδιάταξη αυτόνομων λειτουργικών περιοχών μιας πρωτεΐνης (domain rearrangements)



Όλες οι περιοχές μιας πρωτεΐνης δεν μεταλλάσονται με τον ίδιο ρυθμό

- Αυτόνομες λειτουργικές περιοχές (domains): πολύ συντηρημένες

- Περιοχές ενδογενούς δομικής αστάθειας (intrinsically disordered regions). Π.χ, ευέλικτες συνδετικές περιοχές (flexible linkers).
  - Μεταβαλλόμενο μήκος και περιεκτικότητα αμινοξέων, με παρόμοιες όμως φυσικοχημικές ιδιότητες.
  - Μεταλλάσονται γρήγορα. Το εξελικτικό σήμα μπορεί να χαθεί σύντομα
  - Συχνά δεν υπάρχει περιορισμός θέσης (π.χ φωσφορυλίωση)

Γλοβίνες

- πολύ συντηρημένη τριτοταγής δομή, λίγο συντηρημένη πρωτοταγής δομή (~10-20% ομοιότητα)

Είδη στοίχισης κατά ζεύγη (ι)

- Ολική στοίχιση (global alignment)
  - Προσπαθεί να στοιχίσει όσο το δυνατό περισσότερους χαρακτήρες σε ΟΛΟ το μήκος των δύο αλληλουχιών
  - Για ακολουθίες που δεν έχουν αποκλείνει σε μεγάλο βαθμό και επίσης έχουν παρόμοιο μέγεθος
  - Κλασσική μέθοδος: Needleman-Wunsch.
    - Βασίζεται στον δυναμικό προγραμματισμό

Είδη στοίχισης κατά ζεύγη (ii)

- Τοπική στοίχιση (local alignment)
  - Νησίδες στοίχισης.
  - Για ακολουθίες που έχουν αποκλείνει αρκετά και έχουν απομείνει συντηρημένες μόνο κάποιες περιοχές (domains)
  - Για αντιστοίχιση mRNA με γενωμικό DNA
  - Κλασσικές μέθοδοι:
    - Smith-Waterman (δυναμικός προγραμματισμός)
    - Blast (ευρετικές μέθοδοι-heuristics)

Είδη στοίχισης κατά ζεύγη

- Στοίχιση αλληλεπικάλυψης (overlap ή ends-free alignment) για συναρμολόγηση γονιδιώματος από μικρά αλληλεπικαλυπτόμενα κομμάτια DNA

Είδη στοίχισης κατά ζεύγη (III)

```
Global FTFTALILLAVAV
       F--TAL-LLA-AV

Local  FTFTALILL-AVAV
       --FTAL-LLAAV--
```

Στιγμοπίνακες (dotplots)

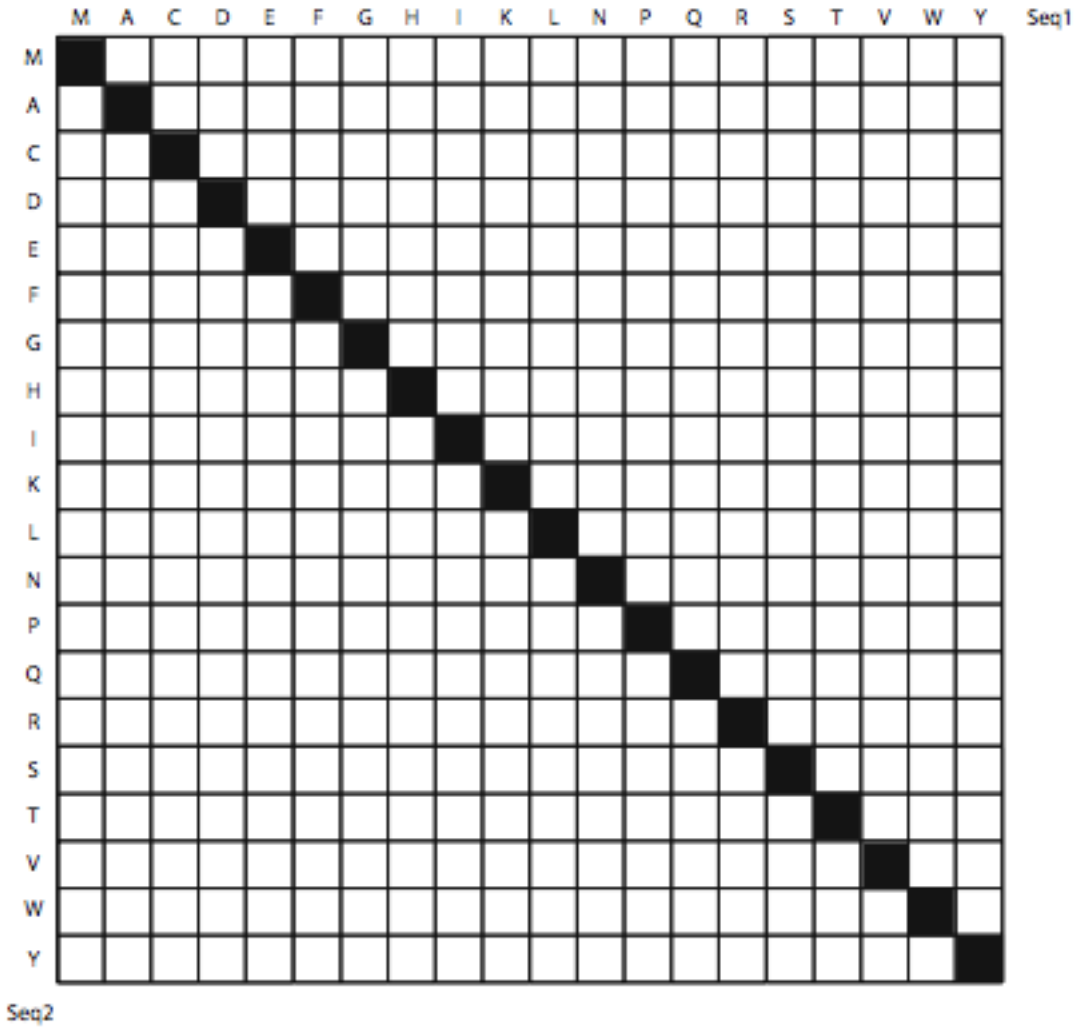
- Εισήχθησαν από τους Gibbs & McIntyre το 1970.
- Χρησιμοποιούνται για σύγκριση 2 ακολουθιών (π.χ. Πρωτεϊνών ή DNA).
- Αποκαλύπτουν
  - Προσθήκες - Εξαλείψεις
  - Ευθείες ή ανεστραμμένες επαναλήψεις (π.χ χρήσιμοι για RNA)
  - Περιοχές χαμηλής πολυπλοκότητας
  - Αναστροφές

- Διάφορα προγράμματα (π.χ Dotlet)

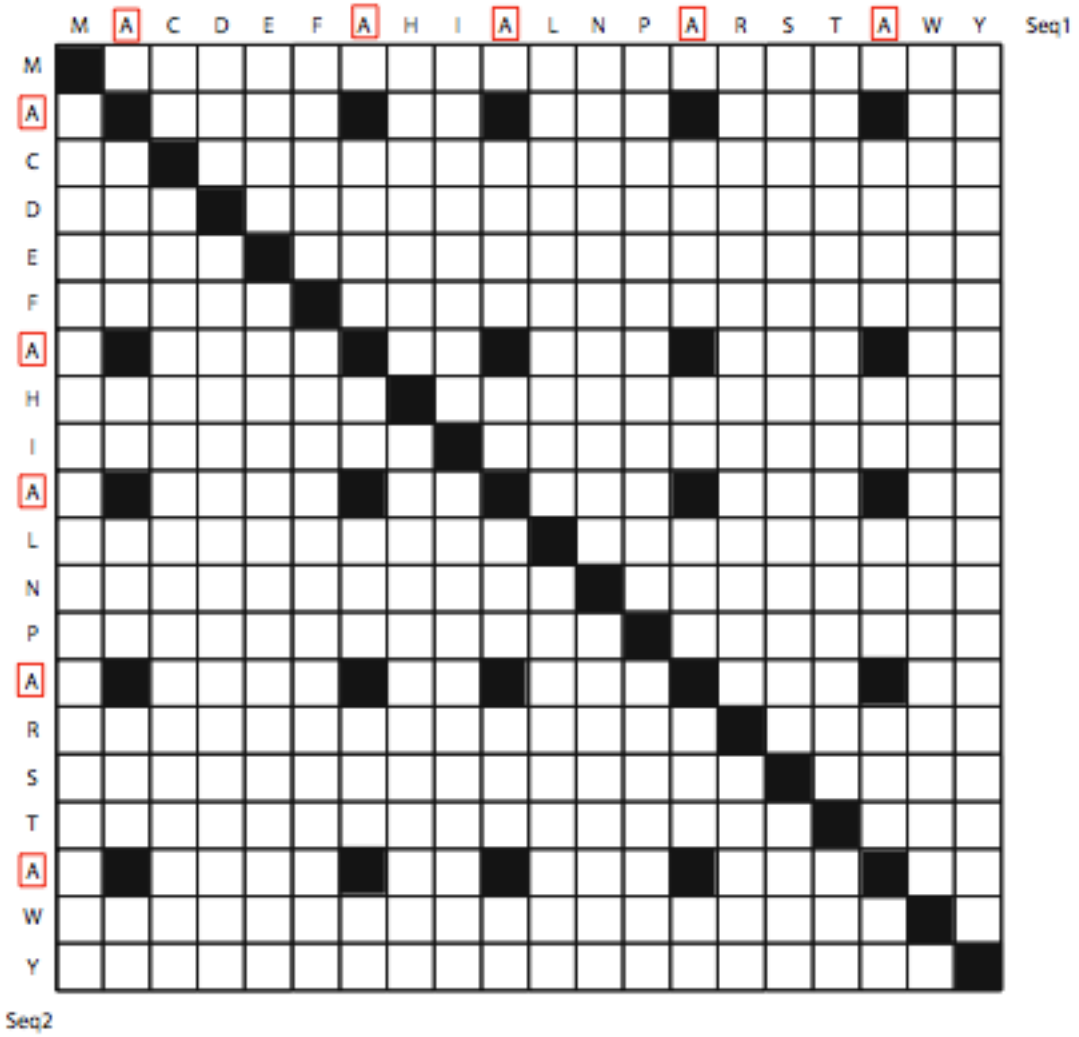
- Σε ένα βαθμό, εισέρχεται η υποκειμενικότητα στην ερμηνεία των αποτελεσμάτων.

Στιγμοπίνακες

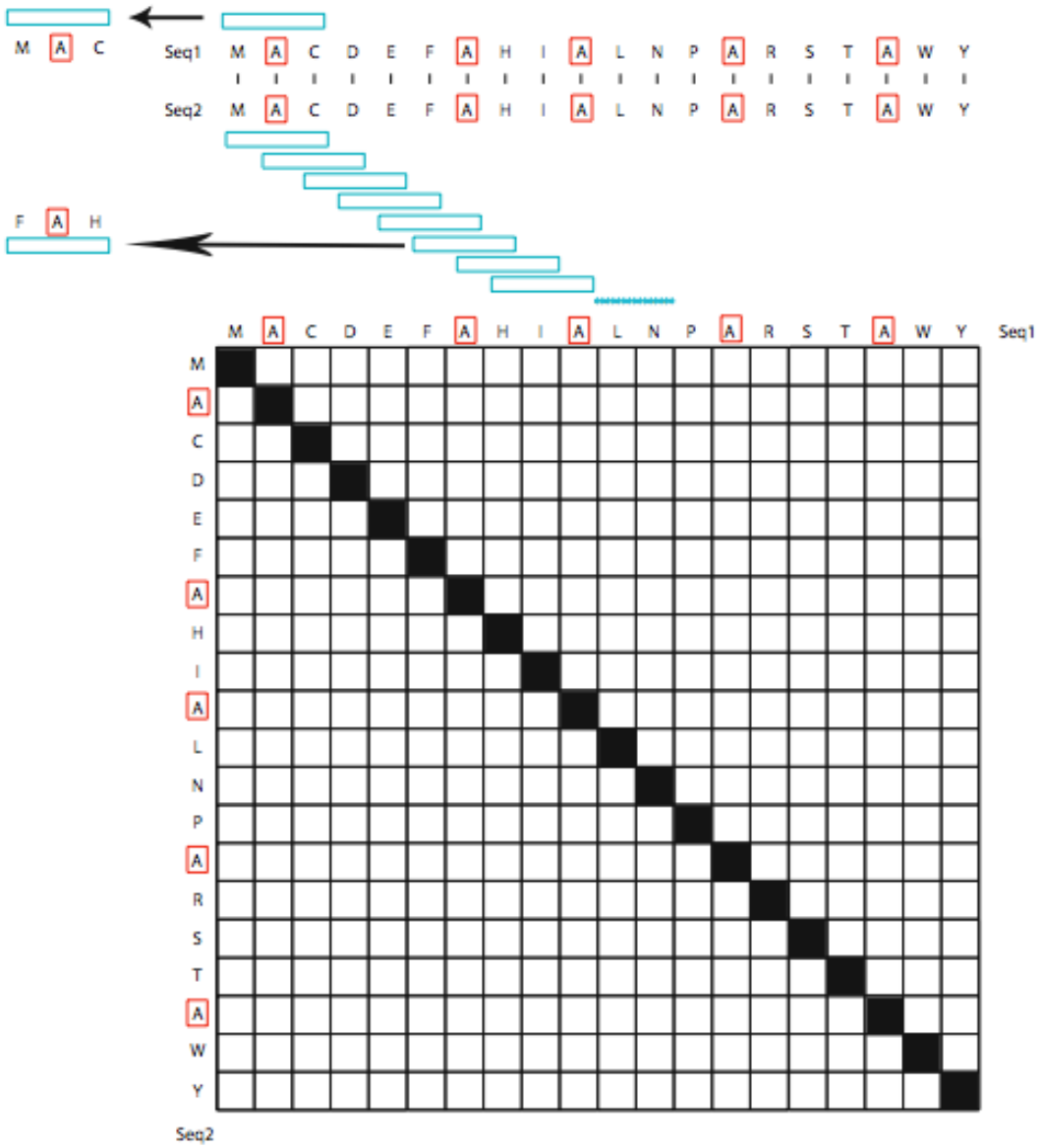
Seq1	M	A	C	D	E	F	G	H	I	K	L	N	P	Q	R	S	T	V	W	Y
	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
Seq2	M	A	C	D	E	F	G	H	I	K	L	N	P	Q	R	S	T	V	W	Y



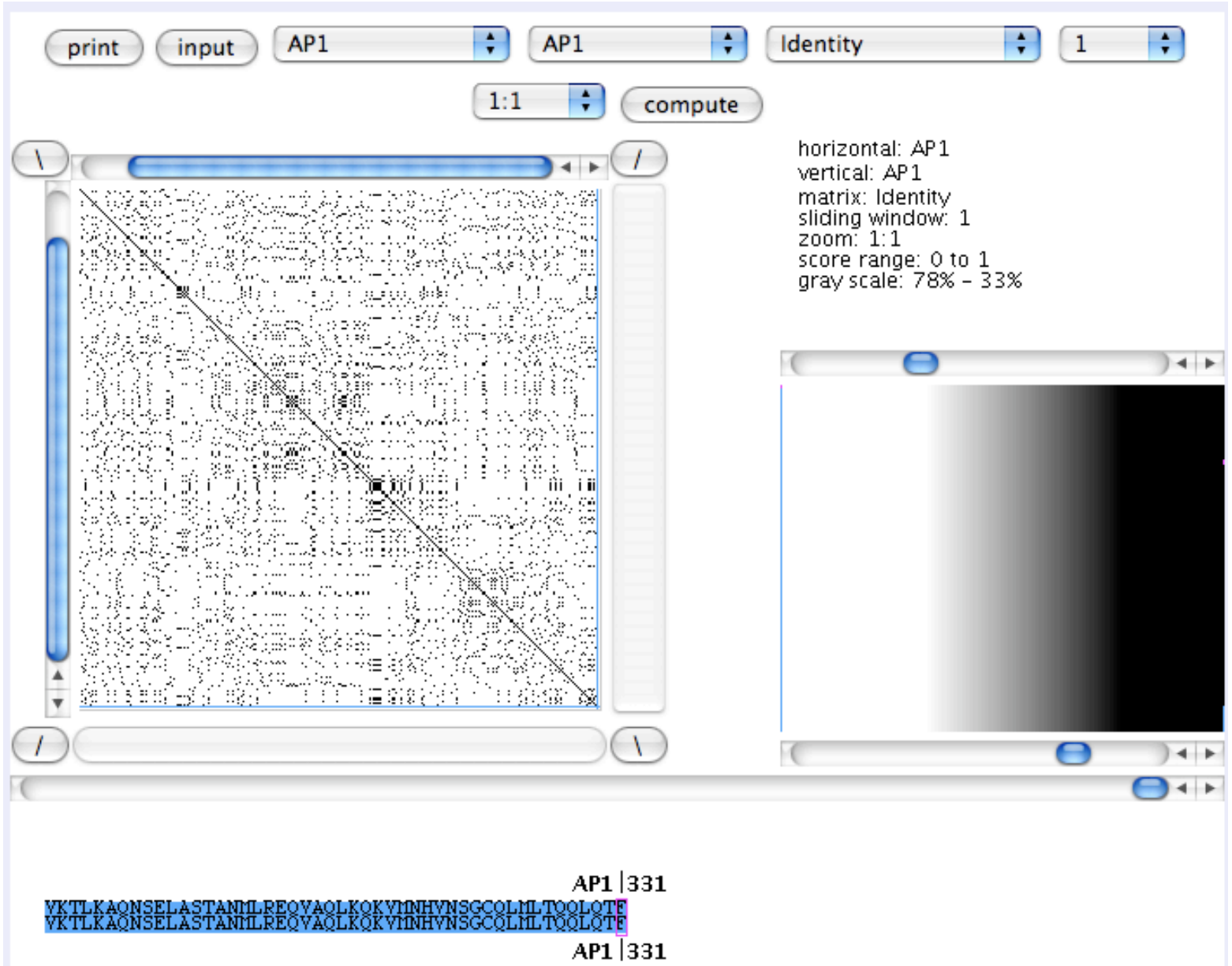
Στιγμοπίνακες - Θόρυβος



Στιγμοπίνακες -
Απαλοιφή Θορύβου με συρόμενα παράθυρα

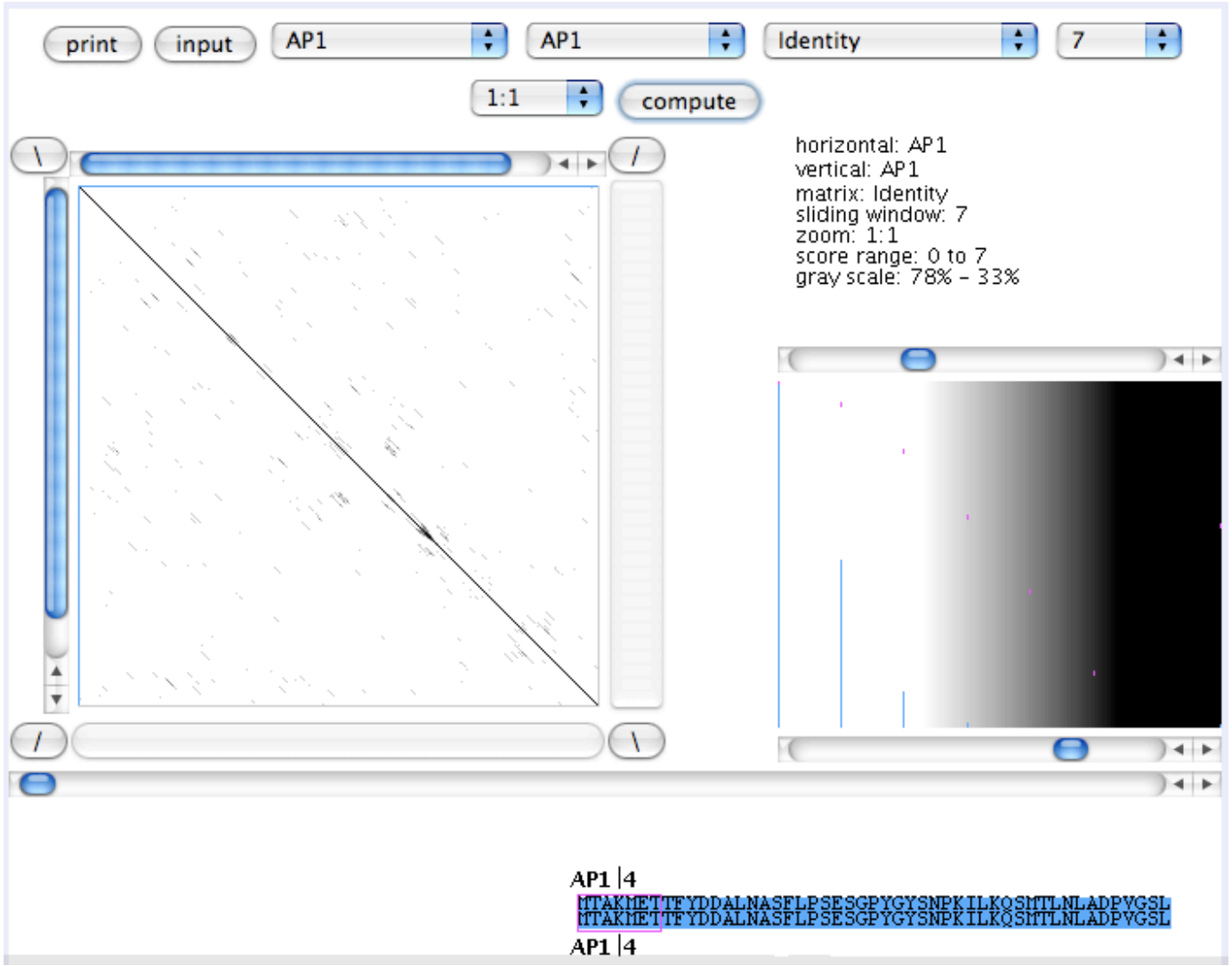


Στιγμοπίνακες



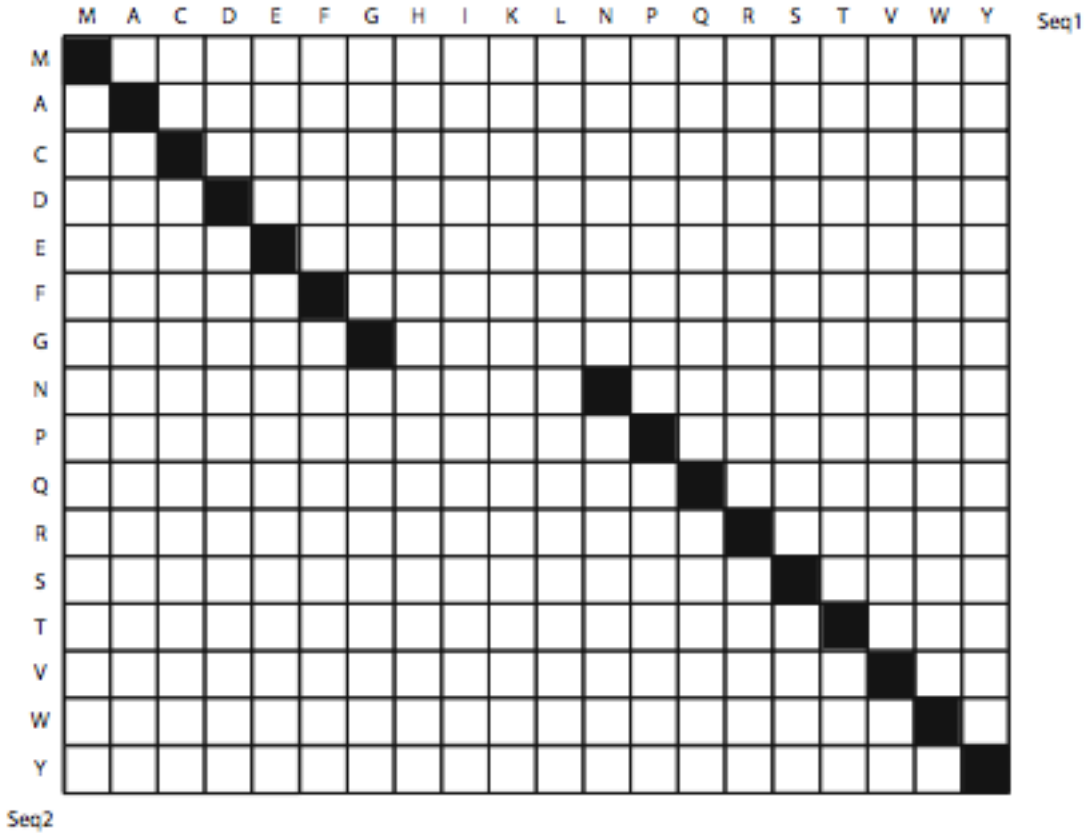
Στιγμοπίνακες

- Απαλοιφή θορύβου με συρόμενα παράθυρα
- Ο Mount προτείνει:
  - Για DNA: παράθυρο 15 χαρακτήρων με τουλάχιστον 10 αντιστοιχίσεις
  - Για πρωτεΐνες: παράθυρο 2-3 χαρακτήρων με τουλάχιστον 2 αντιστοιχίσεις



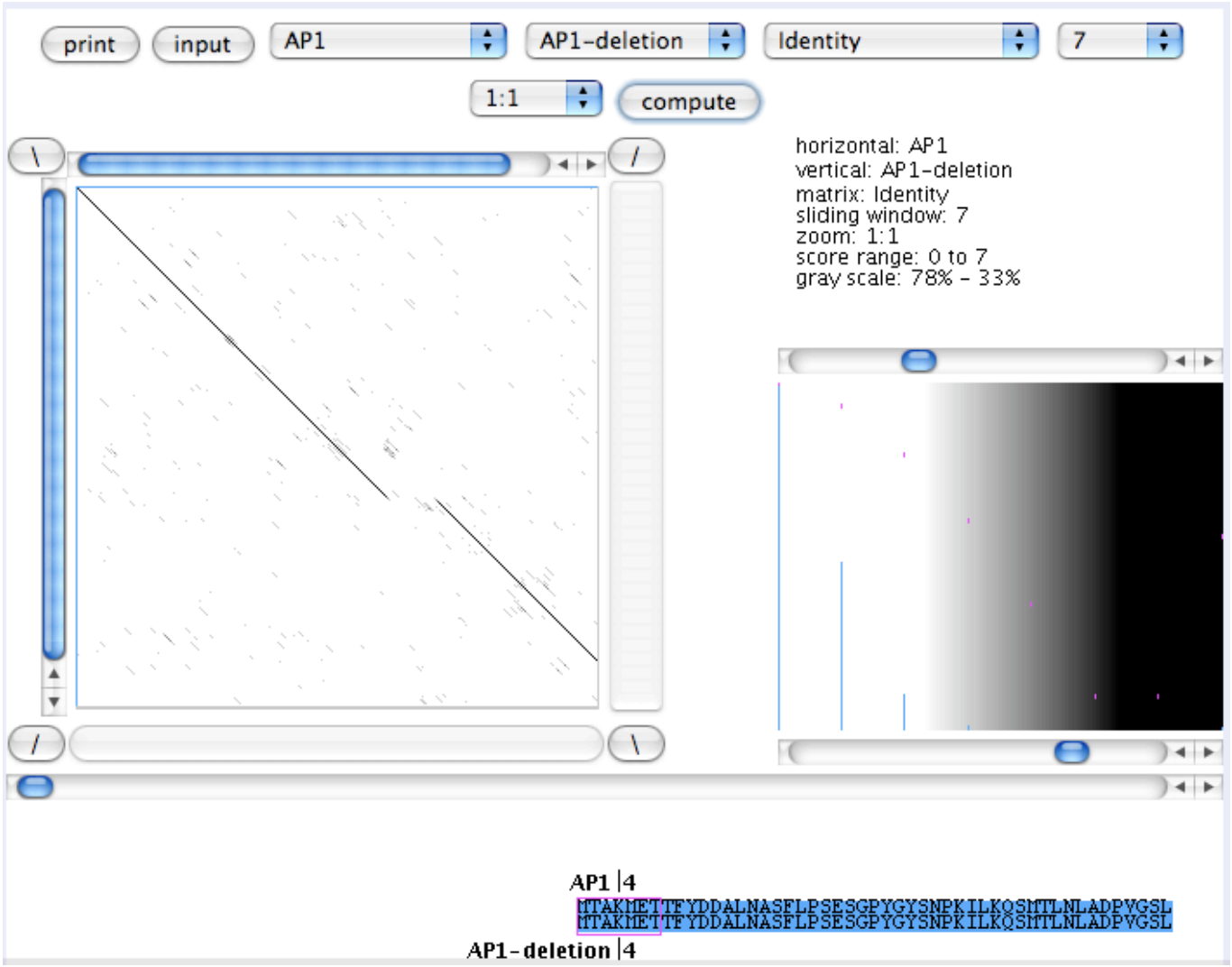
Στιγμοπίνακες insertion - deletions

Seq1	M	A	C	D	E	F	G	H	I	K	L	N	P	Q	R	S	T	V	W	Y
	I	I	I	I	I	I	I					I	I	I	I	I	I	I	I	I
Seq2	M	A	C	D	E	F	G	–	–	–	–	N	P	Q	R	S	T	V	W	Y

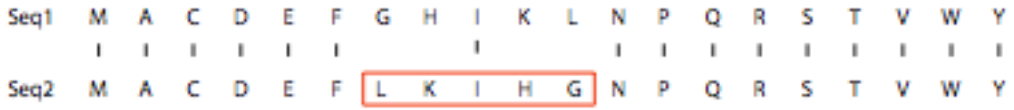


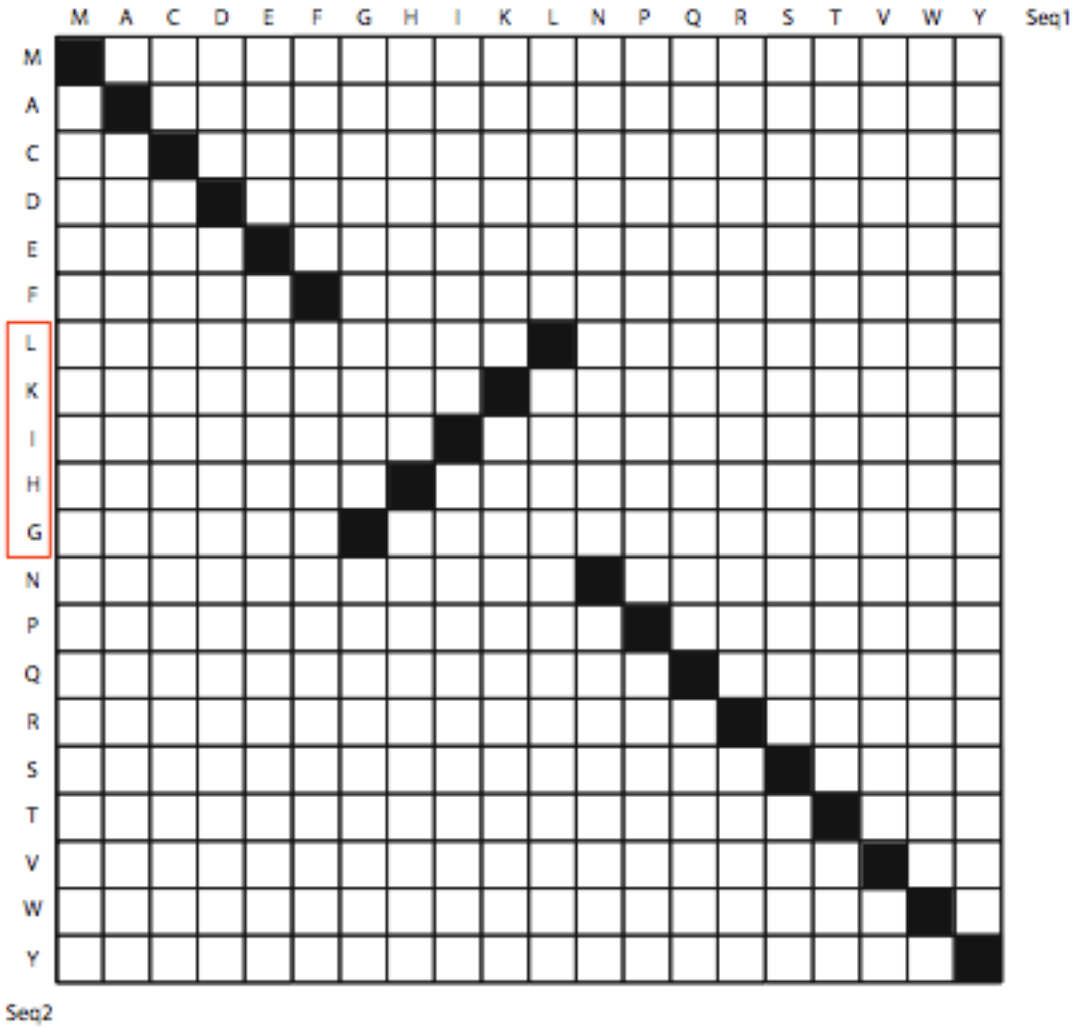
Στιγμοπίνακες

- Insertions/deletions (indels)



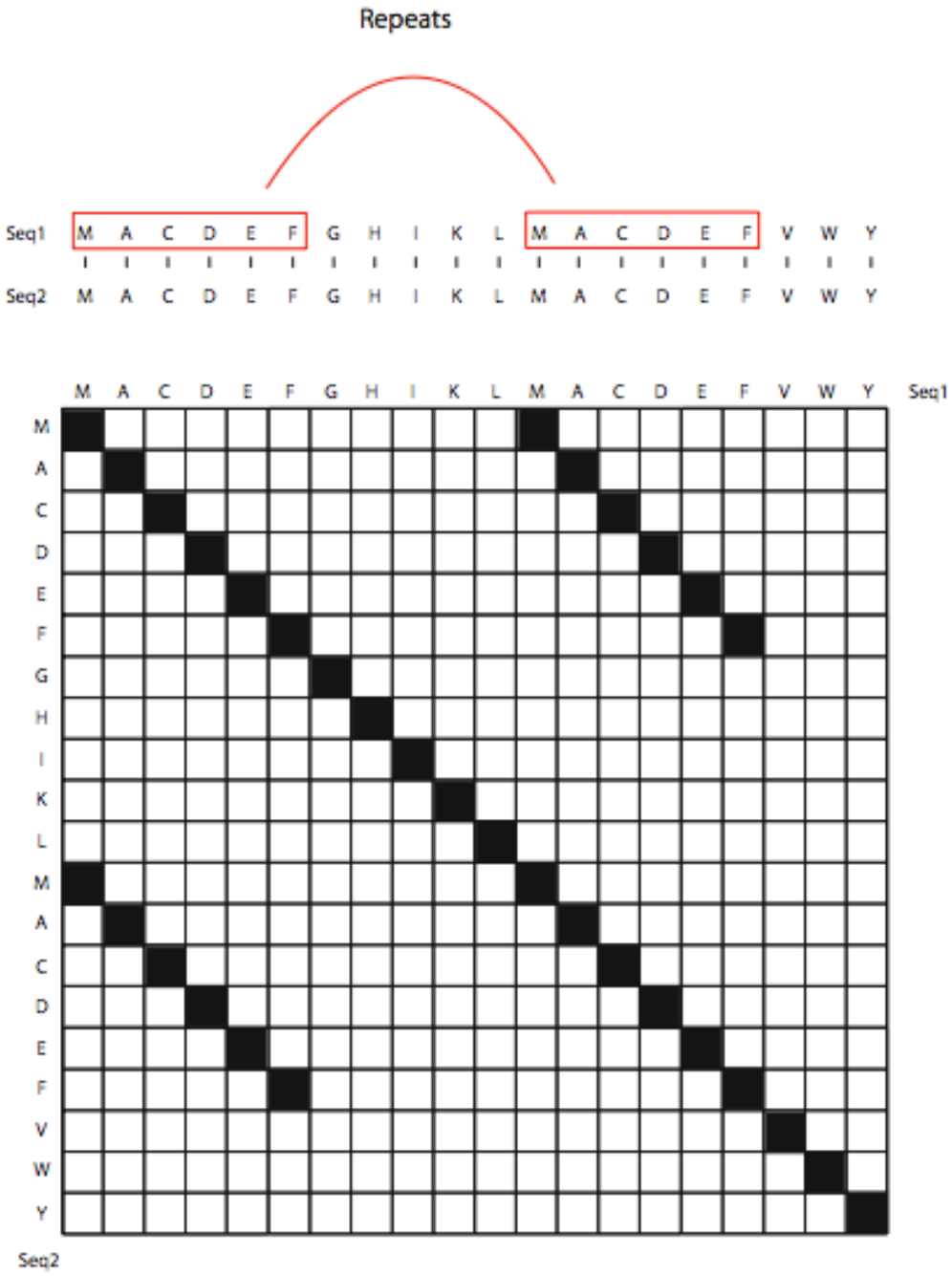
Στιγμοπίνακες - Αναστροφές





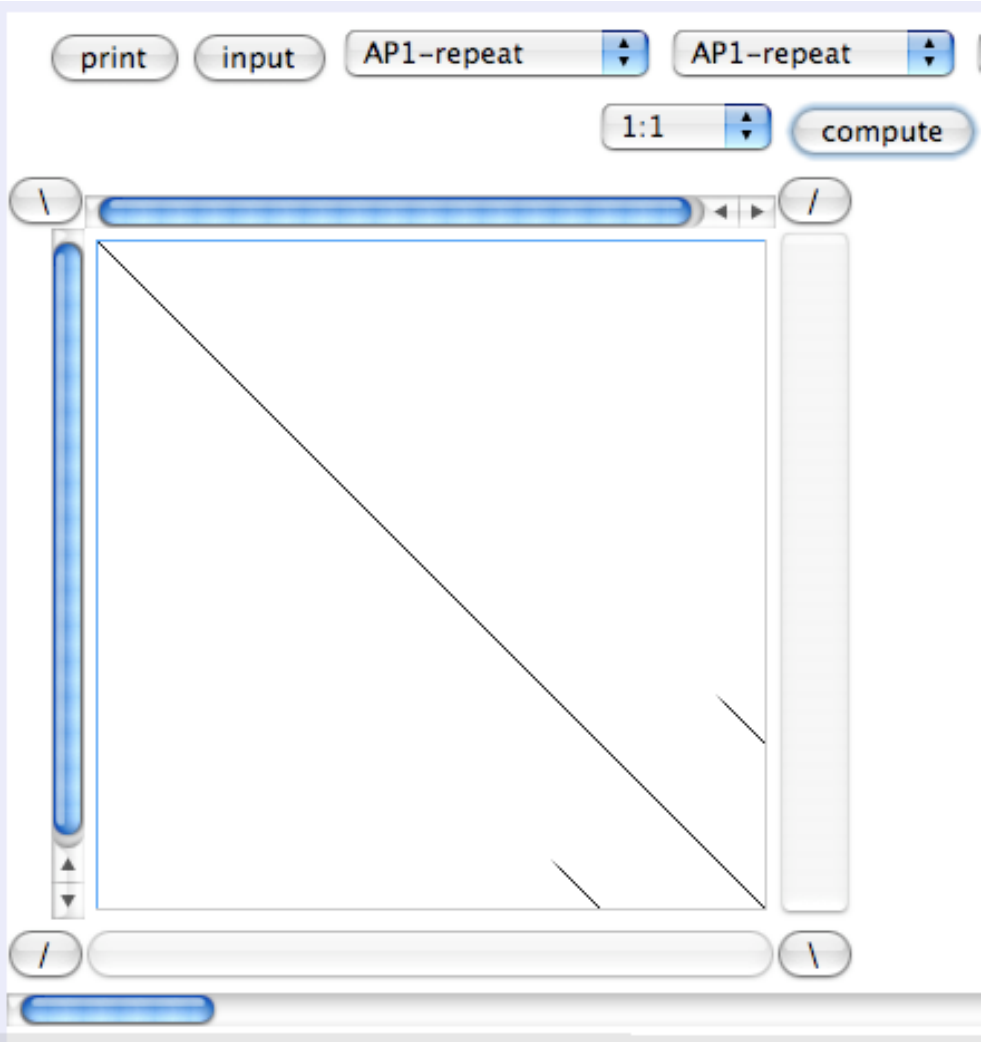
Στιγμοπίνακες

Επαναλήψεις



Στιγμοπίνακες

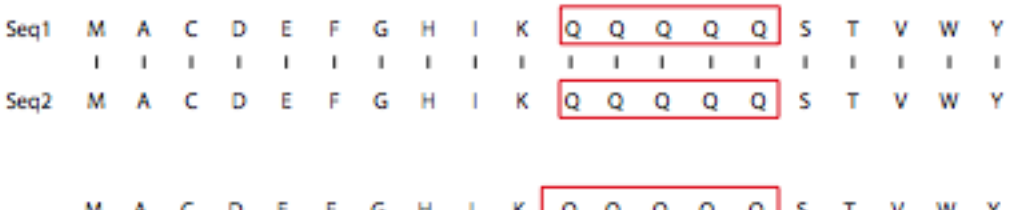
- Επαναλήψεις

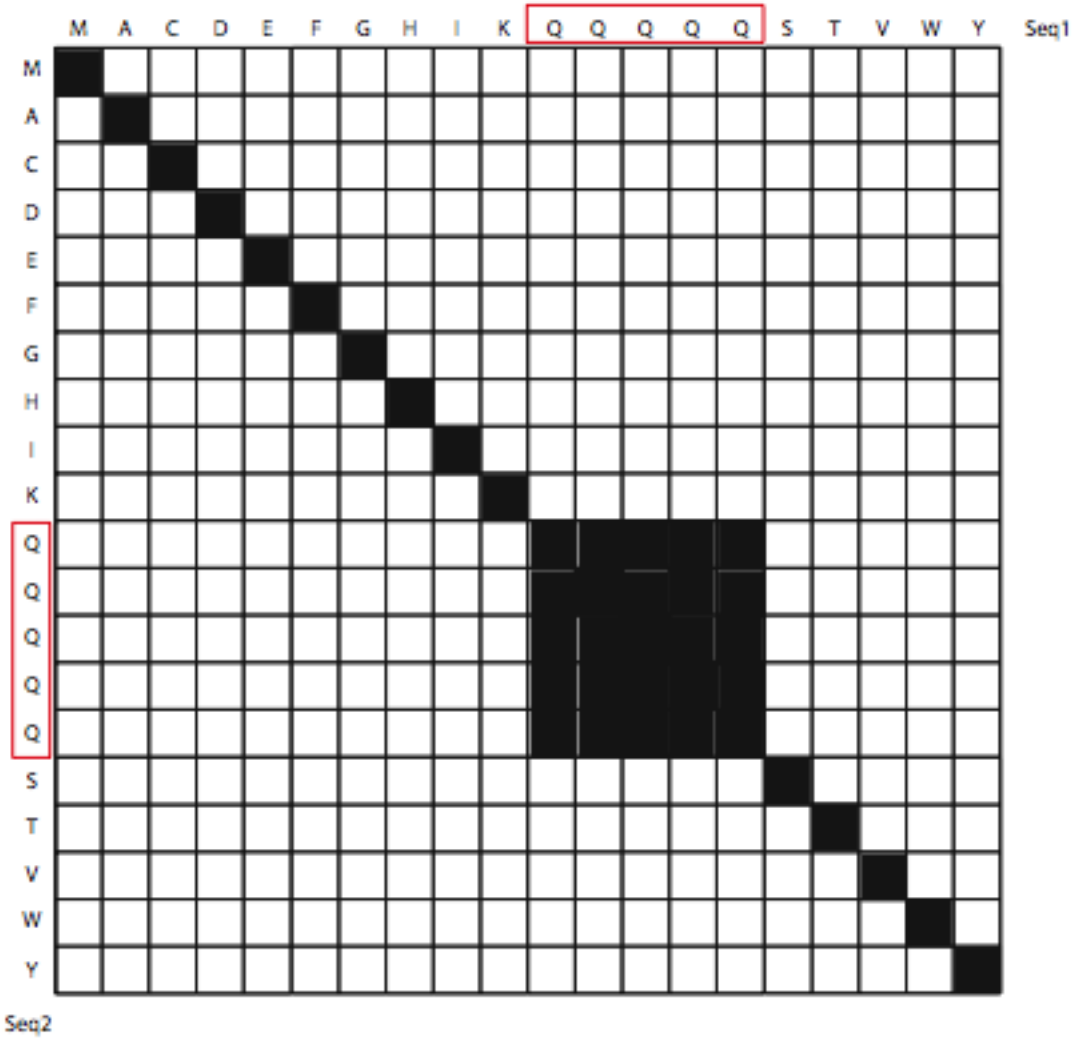


Στιγμοπίνακες

Επαναλήψεις

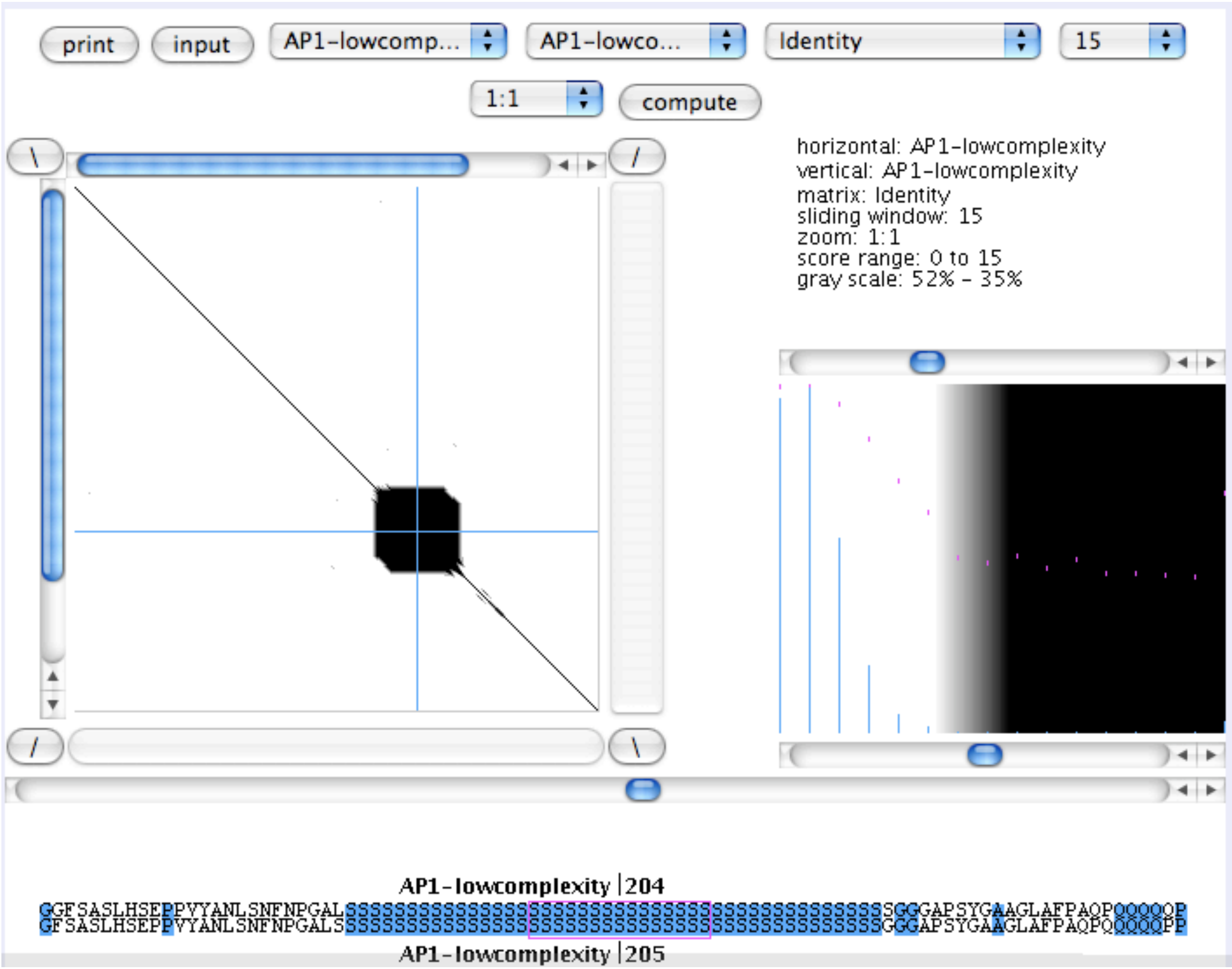
Περιοχές χαμηλής πολυπλοκότητας





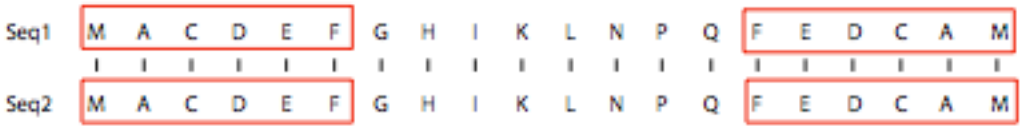
Στιγμοπίνακες

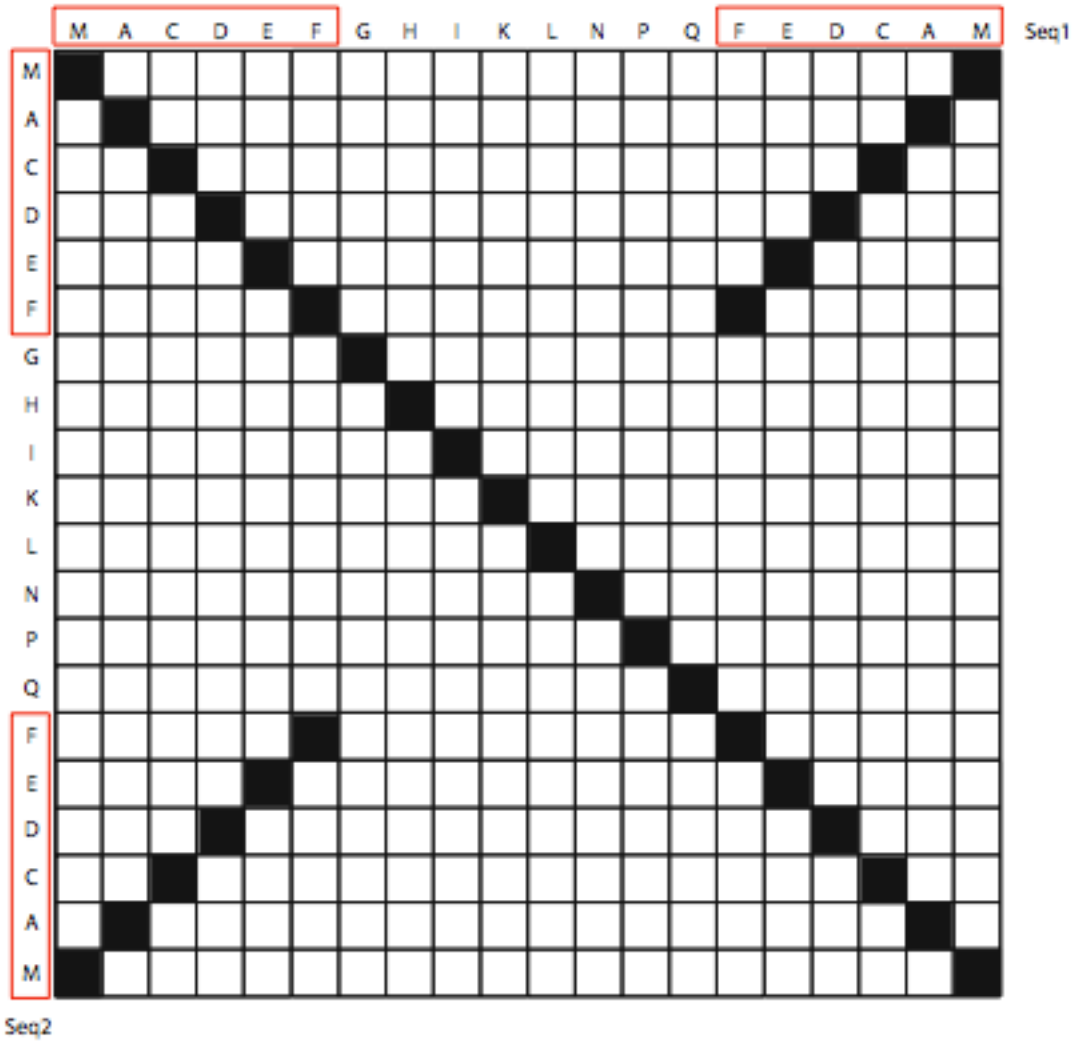
- Επαναλήψεις - Περιοχές χαμηλής πολυπλοκότητας



Στιγμοπίνακες

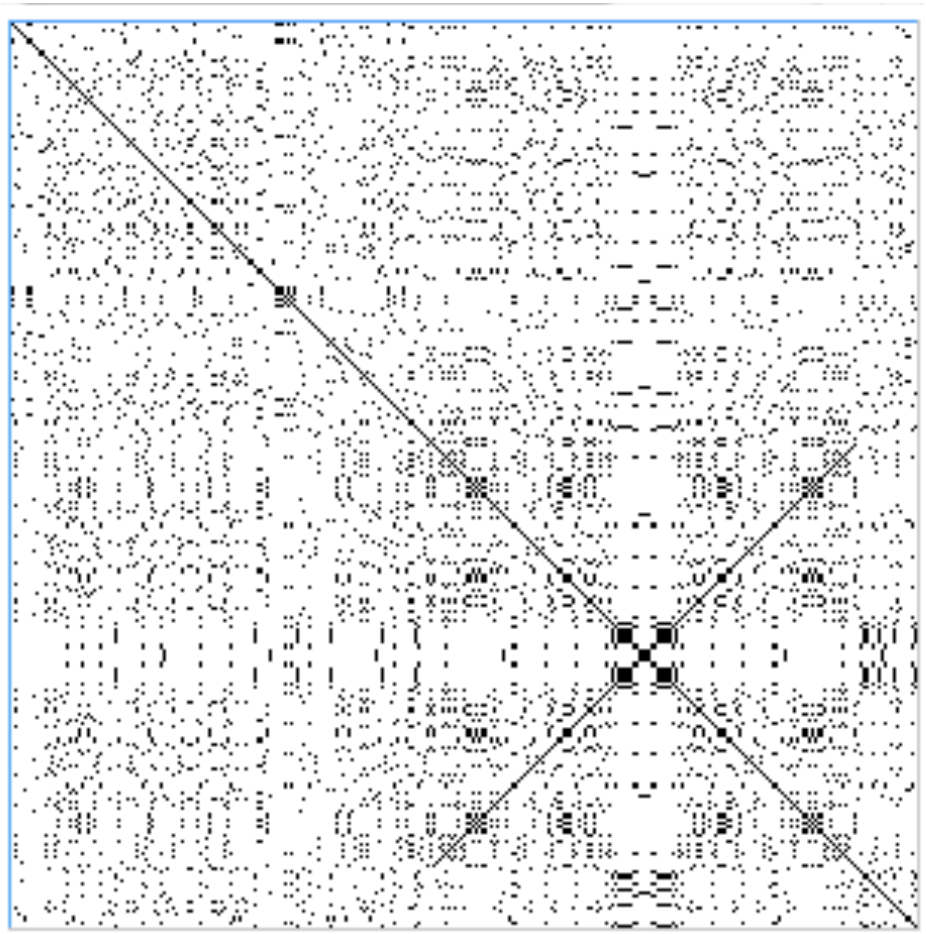
Ανεστραμμένες Επαναλήψεις





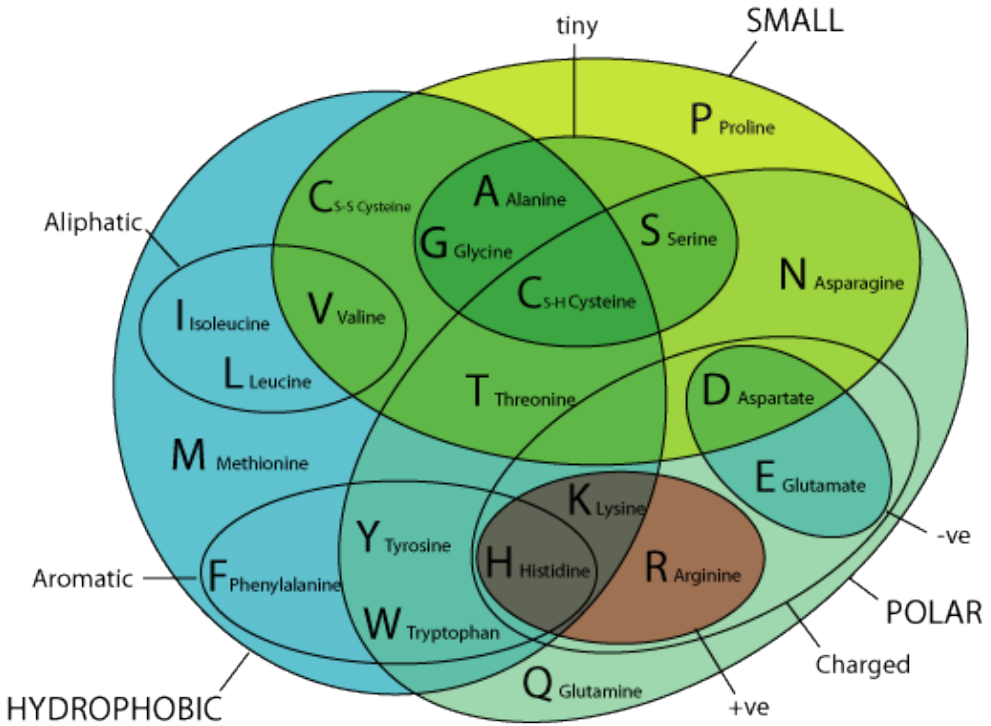
Στιγμοπίνακες

- Ανεστραμμένες επαναλήψεις



Στιγμοπίνακες

- Αν συγκρίνουμε 2 πρωτεΐνες που έχουν αποκλίνει αρκετά, αντί να ελέγξουμε για ακριβές ταίριασμα των αμινοξέων, μπορούμε να ελέγξουμε για ταίριασμα αμινοξέων με παρόμοιες φυσικοχημικές ιδιότητες.
- Χρησιμοποιούμε πίνακες αντικατάστασης (π.χ. PAM, Blosum)
- Για το συρόμενο παράθυρο υπολογίζεται ένα σκορ με βάση τους χρησιμοποιούμενους πίνακες αντικατάστασης.



Δυναμικός προγραμματισμός

- Δίνει την βέλτιστη στοίχιση (Μαθηματικά αποδεδειγμένο).
- Και για ολικές και για τοπικές στοιχίσεις.
- Η στοίχιση εξαρτάται από το βαθμολογικό σύστημα που εφαρμόζεται.

Δυναμικός προγραμματισμός

- Το βαθμολογικό σύστημα πρέπει:
  - Να δίνει βαθμούς για κάθε θέση που οι χαρακτήρες ταιριάζουν απόλυτα
  - Να δίνει βαθμούς (λιγότερους) για κάθε θέση που οι χαρακτήρες έχουν παρόμοιες ιδιότητες
  - Να μην δίνει βαθμούς για μια θέση που οι χαρακτήρες είναι τελείως διαφορετικοί
  - Να βάζει ποινή για κάθε κενό που εισάγεται
  - Να βάζει ποινή (μικρότερη) για κάθε κενό που επεκτείνεται

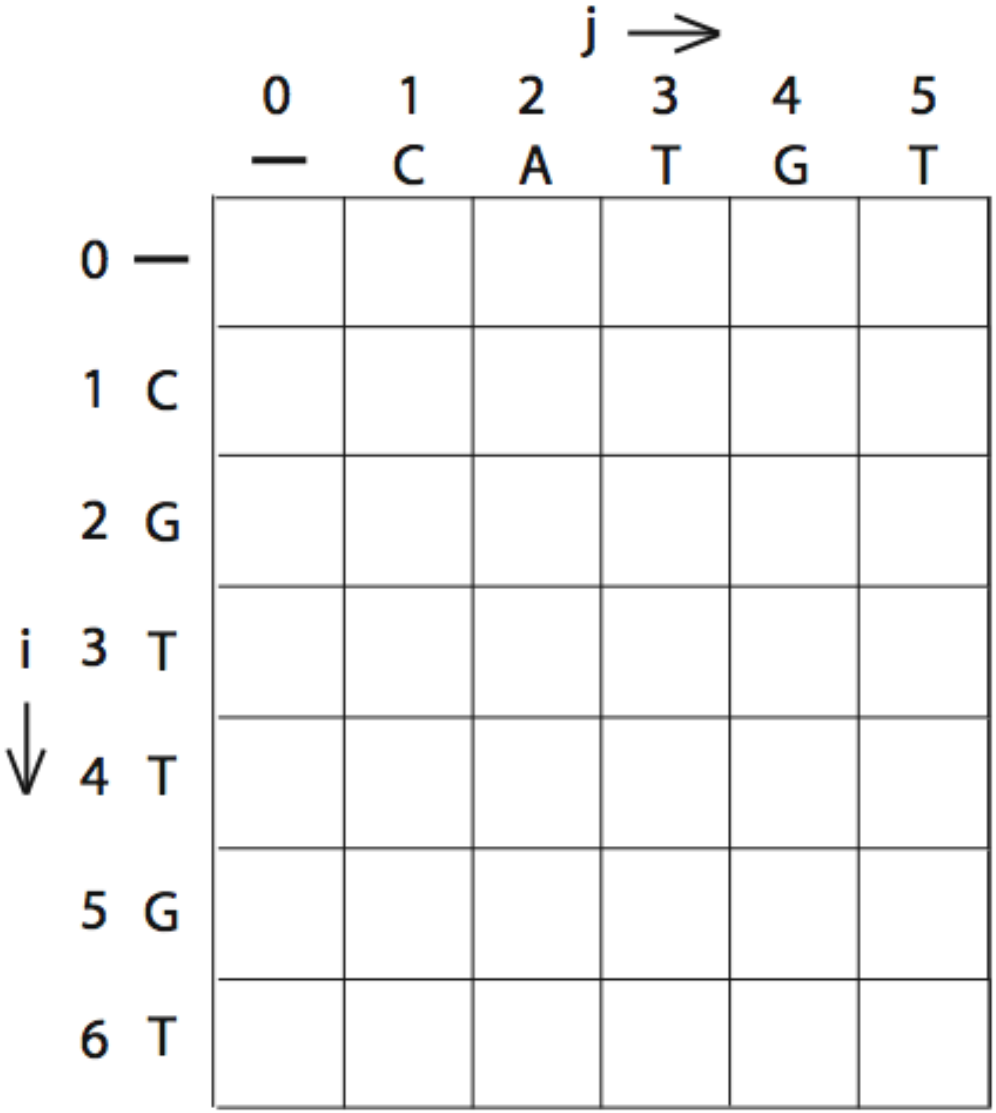
Δυναμικός προγραμματισμός

Το βαθμολογικό σύστημα

```
sequence 1   V   D   S   -   C   Y
sequence 2   V   E   S   L   C   Y
SCORE        4   2   4 -11   9   7   SCORE = SUM OF AMINO ACID PAIR SCORES
(26)                                 MINUS SINGLE GAP PENALTY (11) = 15
```

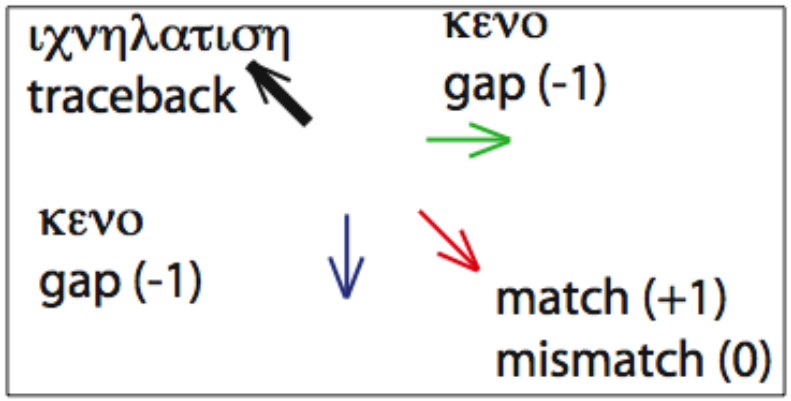
Figure 3.7. Example of scoring a sequence alignment with a gap penalty. The individual alignment scores are taken from an amino acid substitution matrix.

Δ.Π. Ολική στοίχιση παράδειγμα (i)



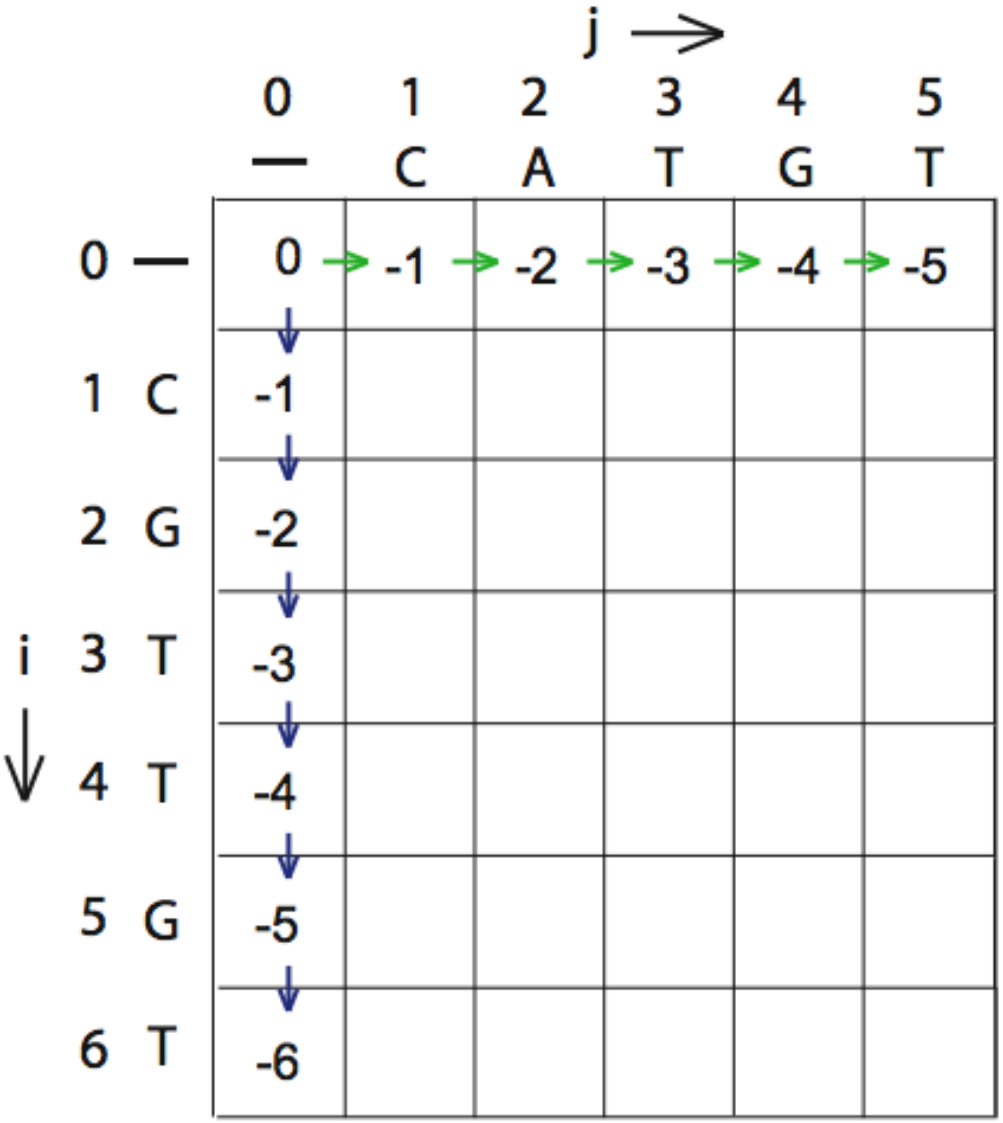
Scoring-System

match=1
mismatch=0
gap=-1



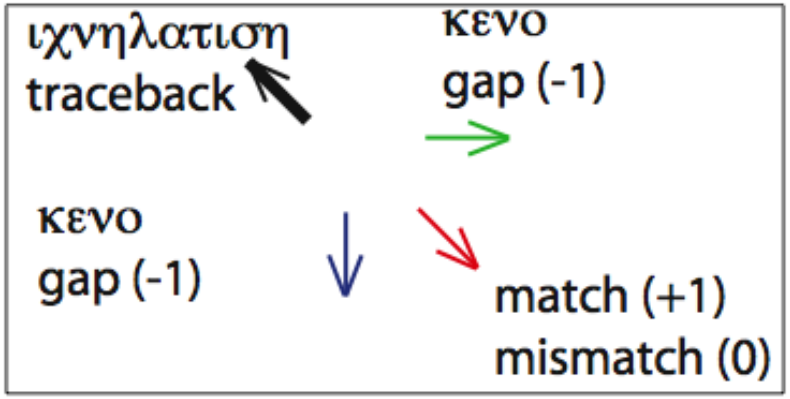
Δ.Π. Ολική στοίχιση παράδειγμα (ii)

Εκκίνηση του πίνακα



Scoring-System

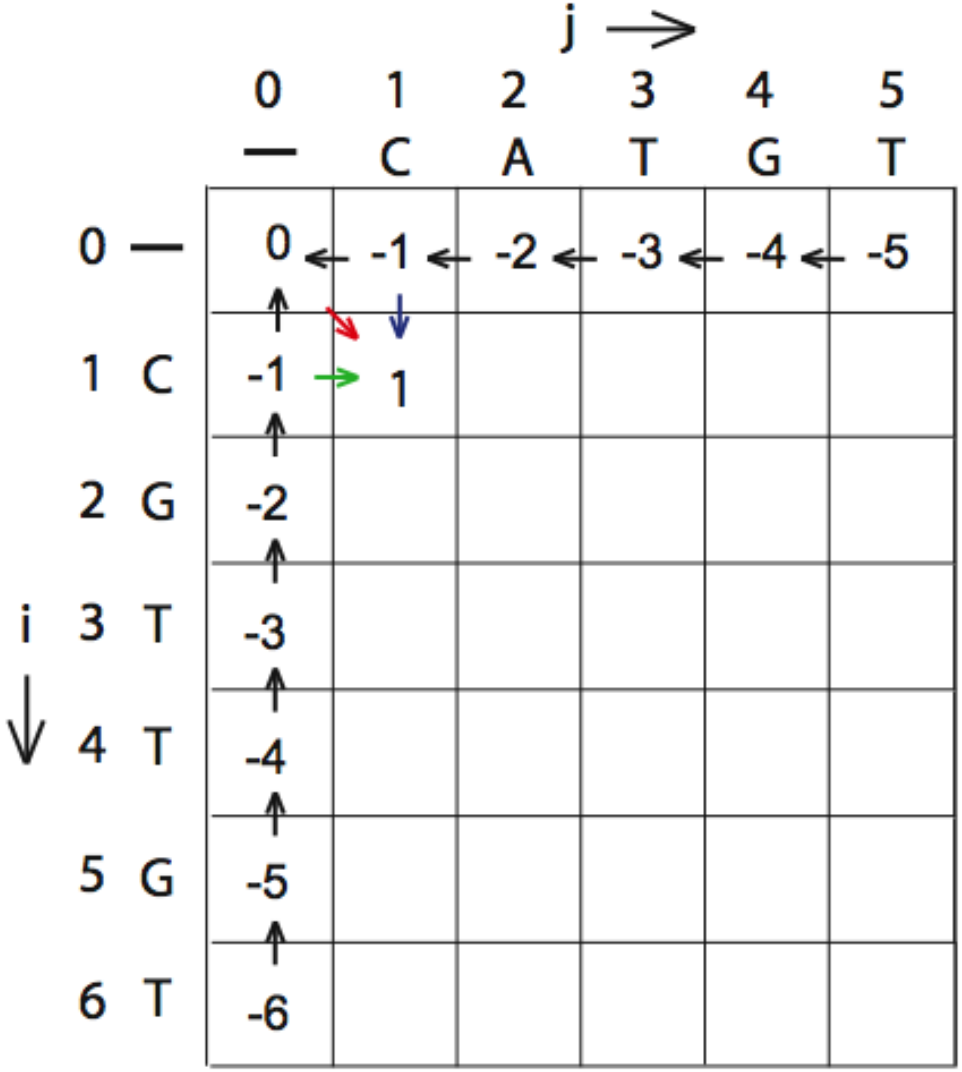
match=1
mismatch=0
gap=-1



Δ.Π. Ολική στοίχιση παράδειγμα (iii)

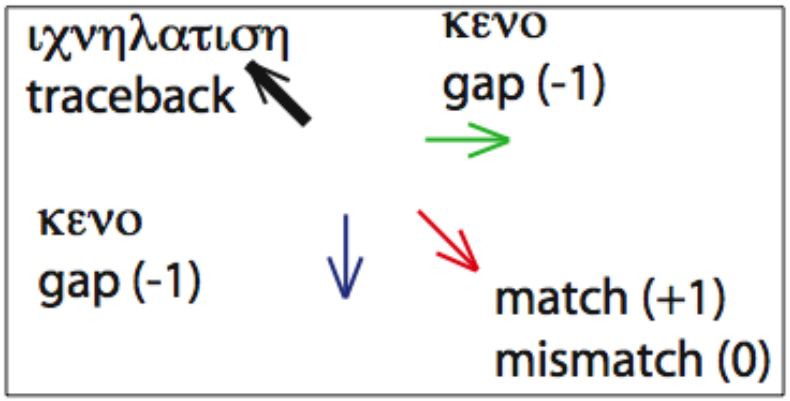
Συμπλήρωση πίνακα

$S_{1,1}$ = MAX [0+1, -1-1, -1-1] = 1



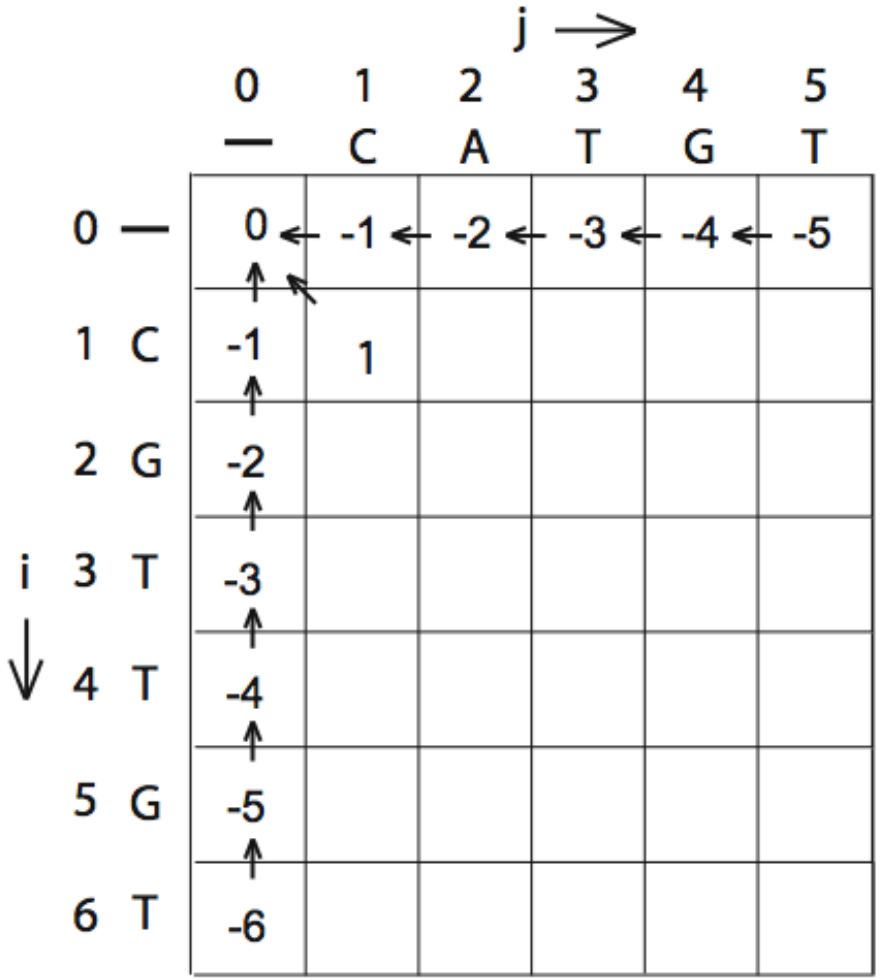
Scoring-System

match=1
mismatch=0
gap=-1



Δ.Π. Ολική στοίχιση παράδειγμα (iv)

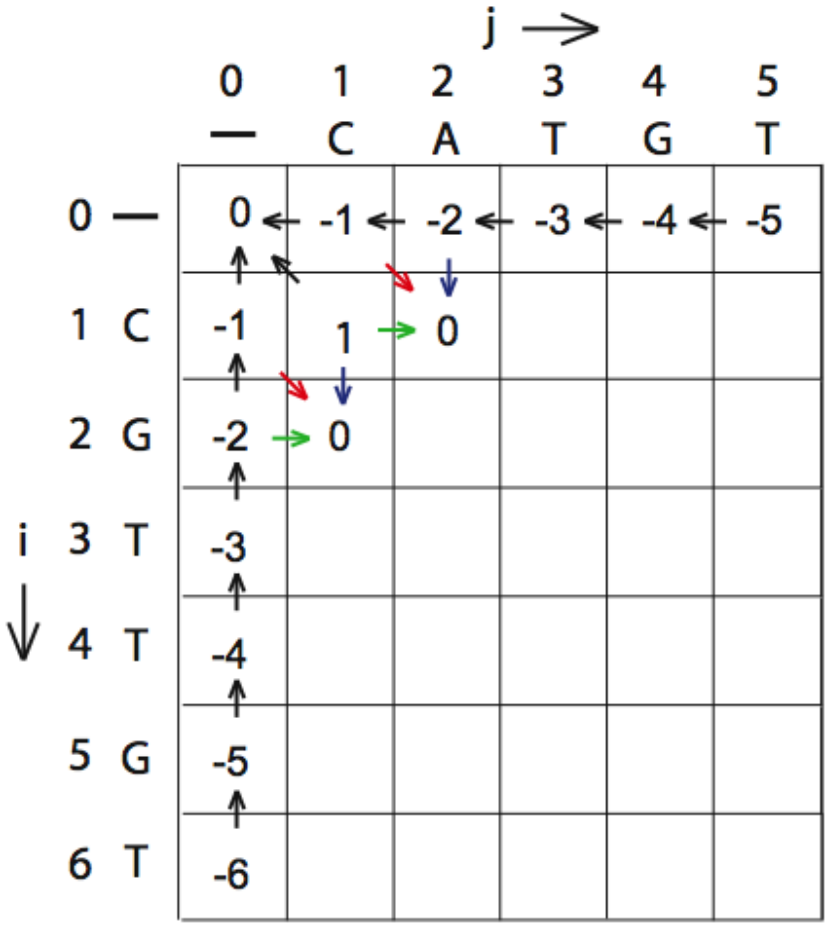
ιχνηλάτηση



Δ.Π. Ολική στοίχιση παράδειγμα (v)

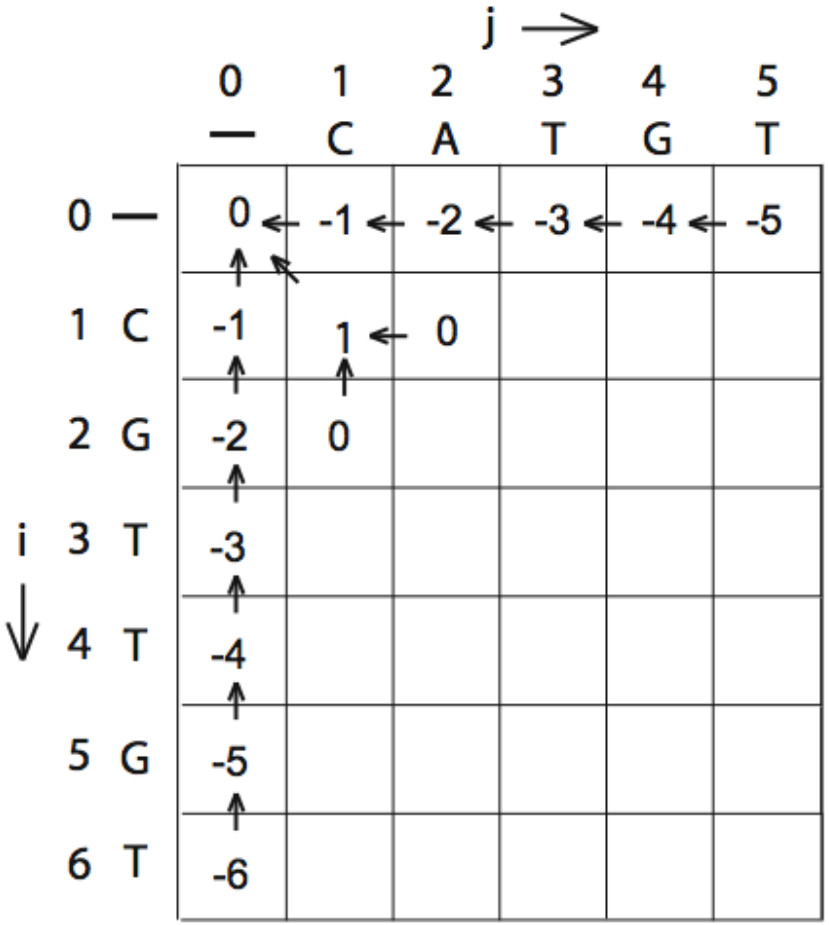
συμπλήρωση

$S_{1,2}$ = MAX [-1+0, 1-1, -2-1] = 0



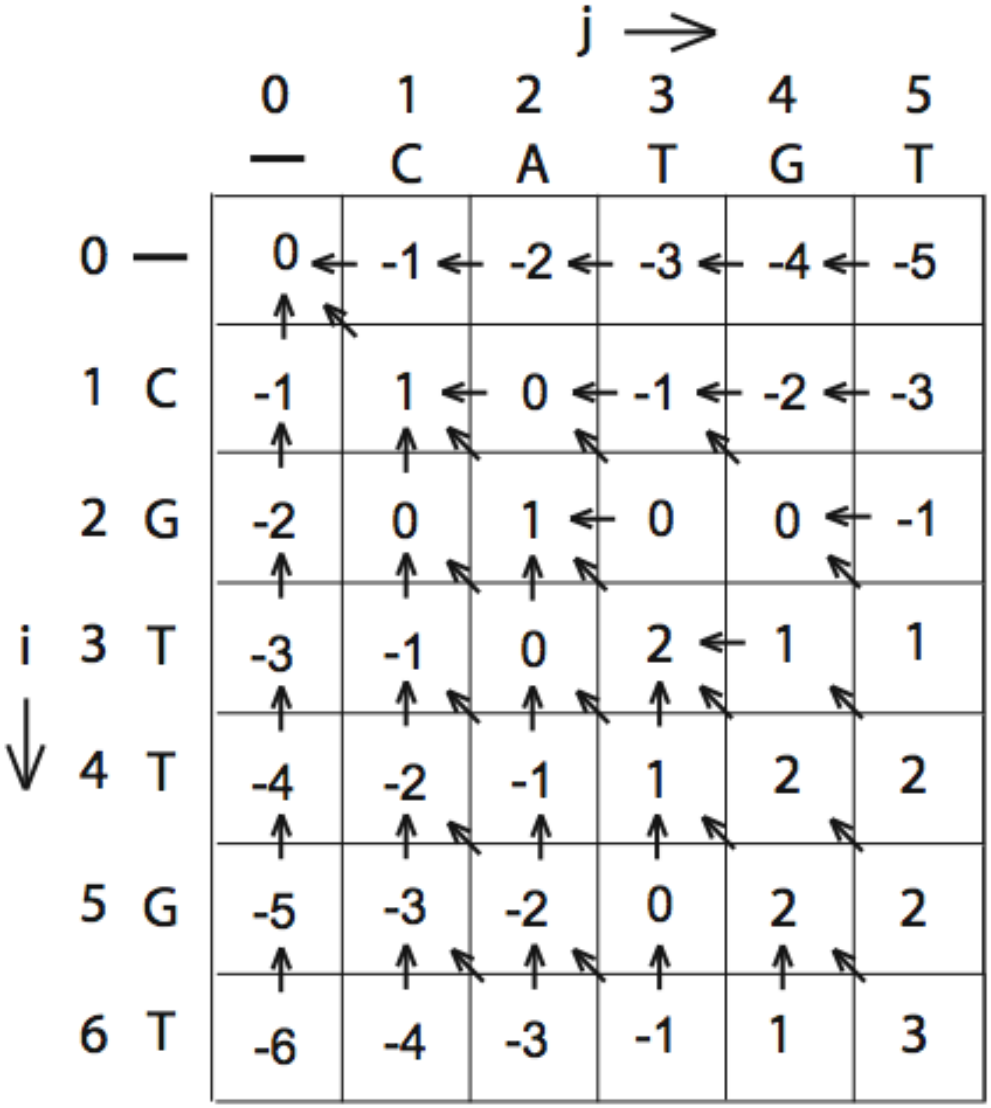
$S_{2,1}$ = MAX [-1+0, -2-1, 1-1] = 0

$S_{1,2}$ = MAX [-1+0, 1-1, -2-1] = 0

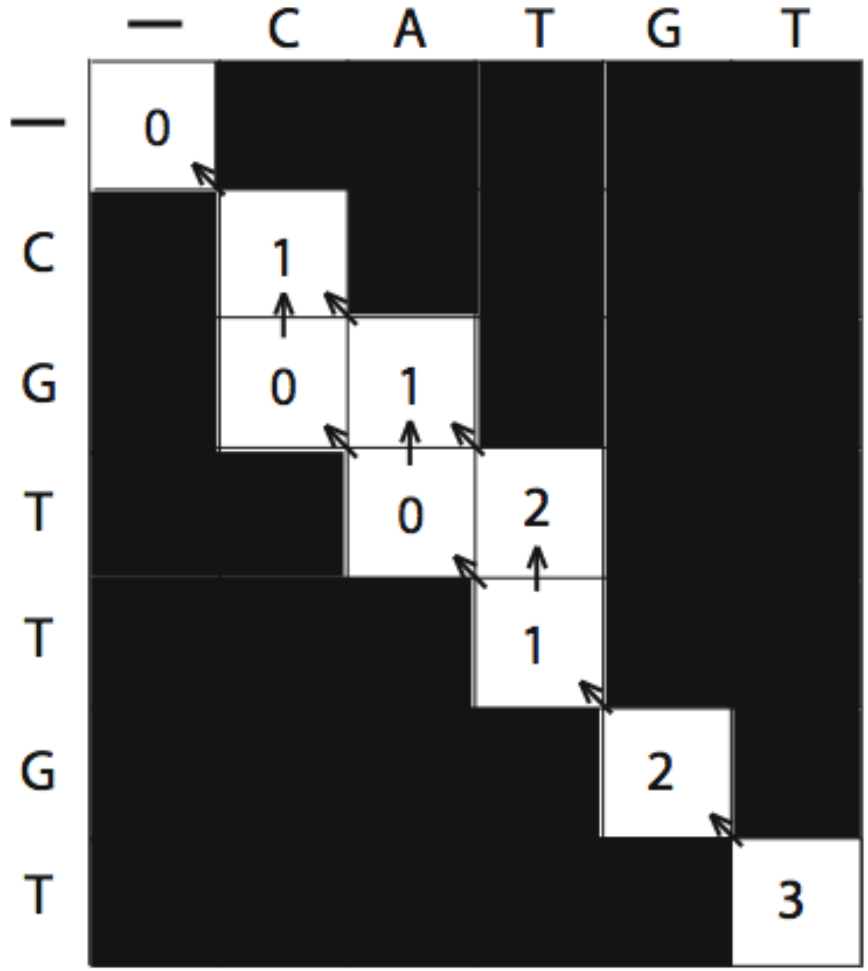


$S_{2,1}$ = MAX [-1+0, -2-1, 1-1] = 0

Δ.Π. Ολική στοίχιση παράδειγμα (vi)



Ολική στοίχιση: ιχνηλάτιση



0←1←1←2←1←2←3

C	A	T	—	G	T	Seq1
I		I		I	I	
C	G	T	T	G	T	Seq2

0←1←1←0←1←2←3

C	A	—	T	G	T	Seq1
I			I	I	I	
C	G	T	T	G	T	Seq2

0←1←0←0←1←2←3

C	—	A	T	G	T	Seq1
I			I	I	I	
C	G	T	T	G	T	Seq2

Πρέπει να βρούμε όλες τις δυνατές πορείες από κάτω-δεξιά -> πάνω-αριστερά.

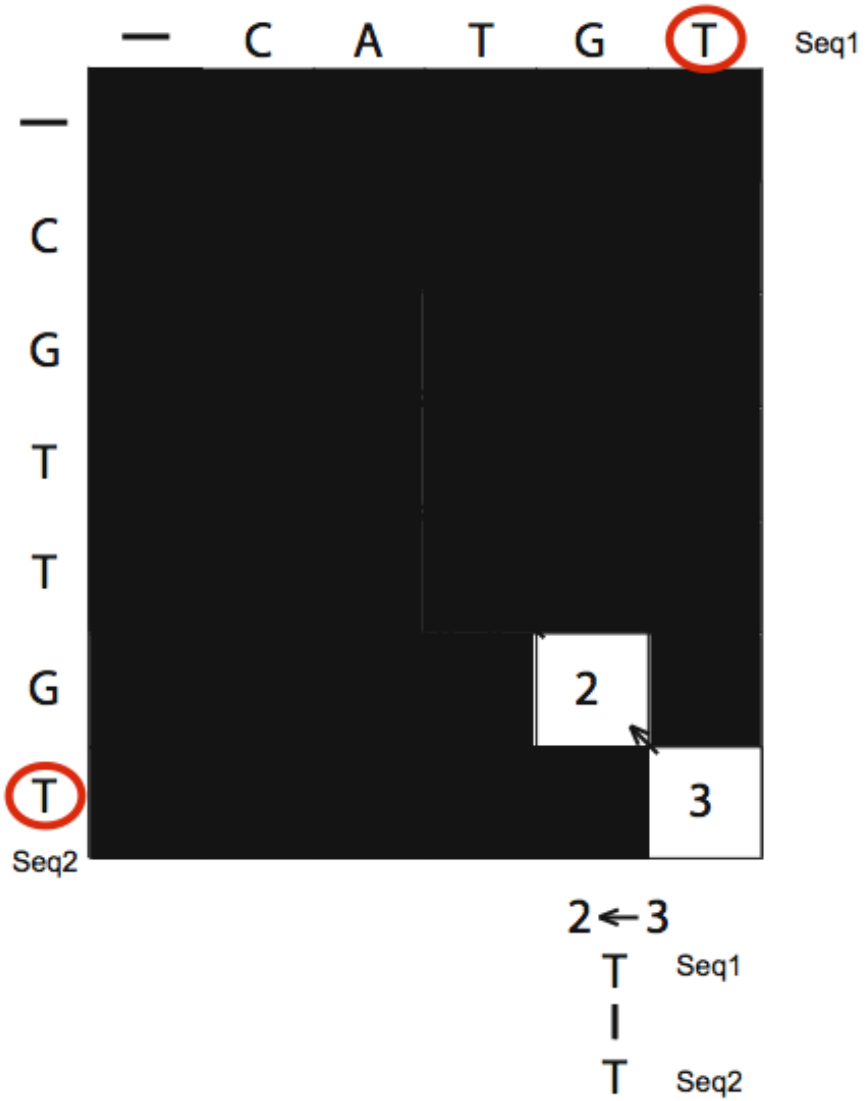
Εδώ: 3 πιθανές πορείες = 3 εξίσου καλές λύσεις

Πώς στοιχίζουμε

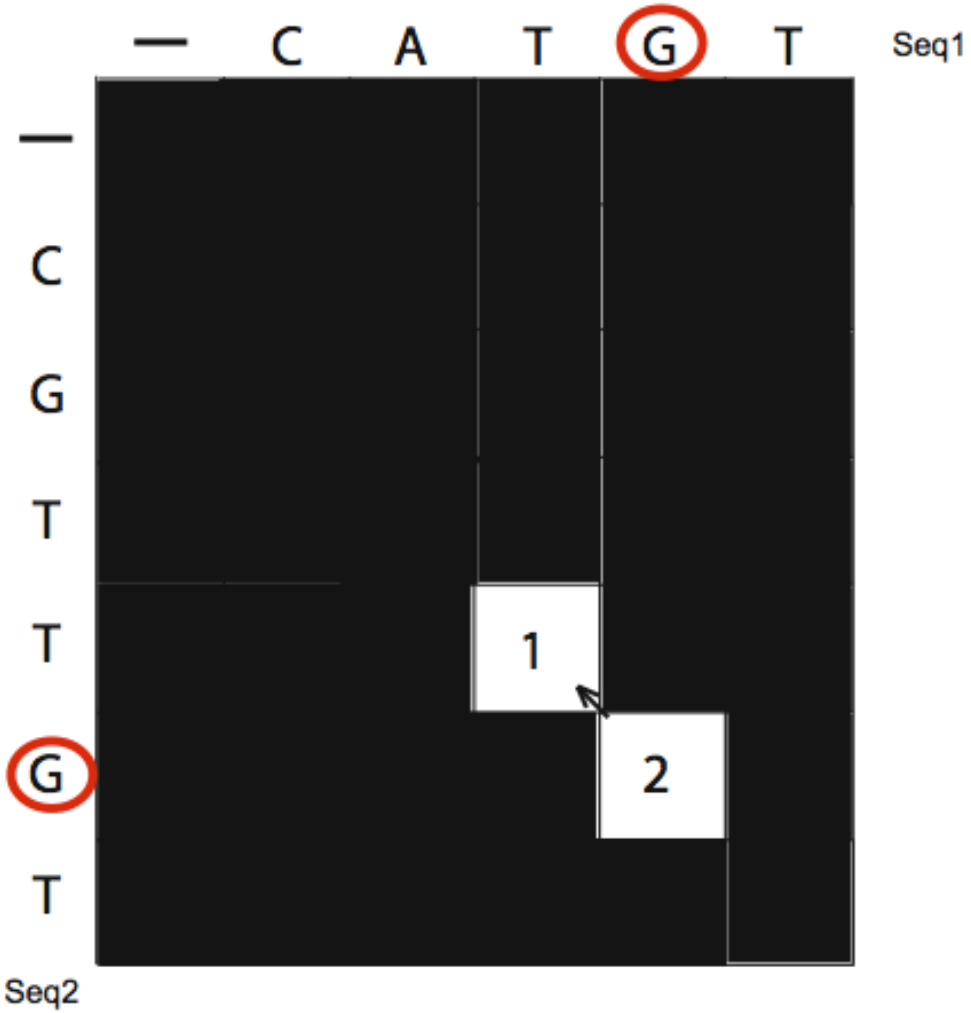
Για κάθε θέση:

- Αν κινηθούμε διαγώνια, τότε στοιχίζουμε τα 2 νουκλεοτίδια/ αμινοξέα που αντιστοιχούν για εκείνη την θέση (είτε ταιριάζουν είτε όχι).
- Αν κινηθούμε οριζόντια ή κάθετα βάζουμε κενό στην ακολουθία που δείχνει το βέλος

Πώς στοιχίζουμε



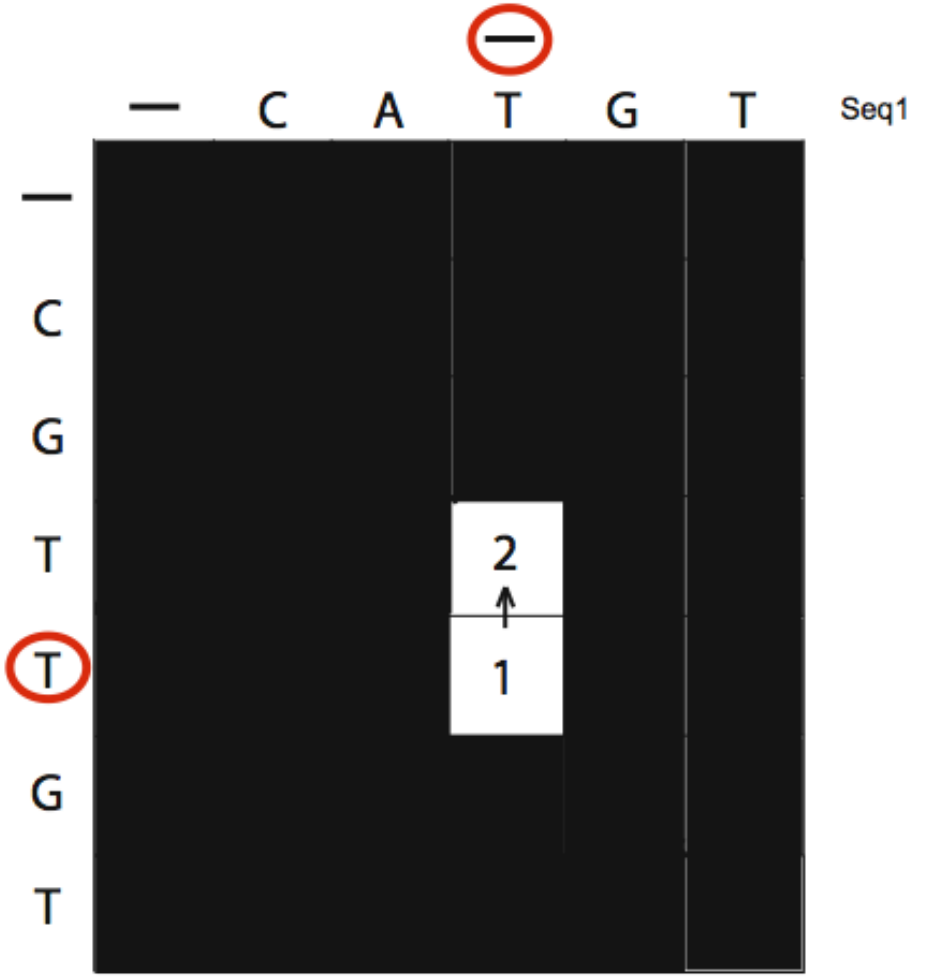
Πώς στοιχίζουμε



1←2

G	T	Seq1
I	I	
G	T	Seq2

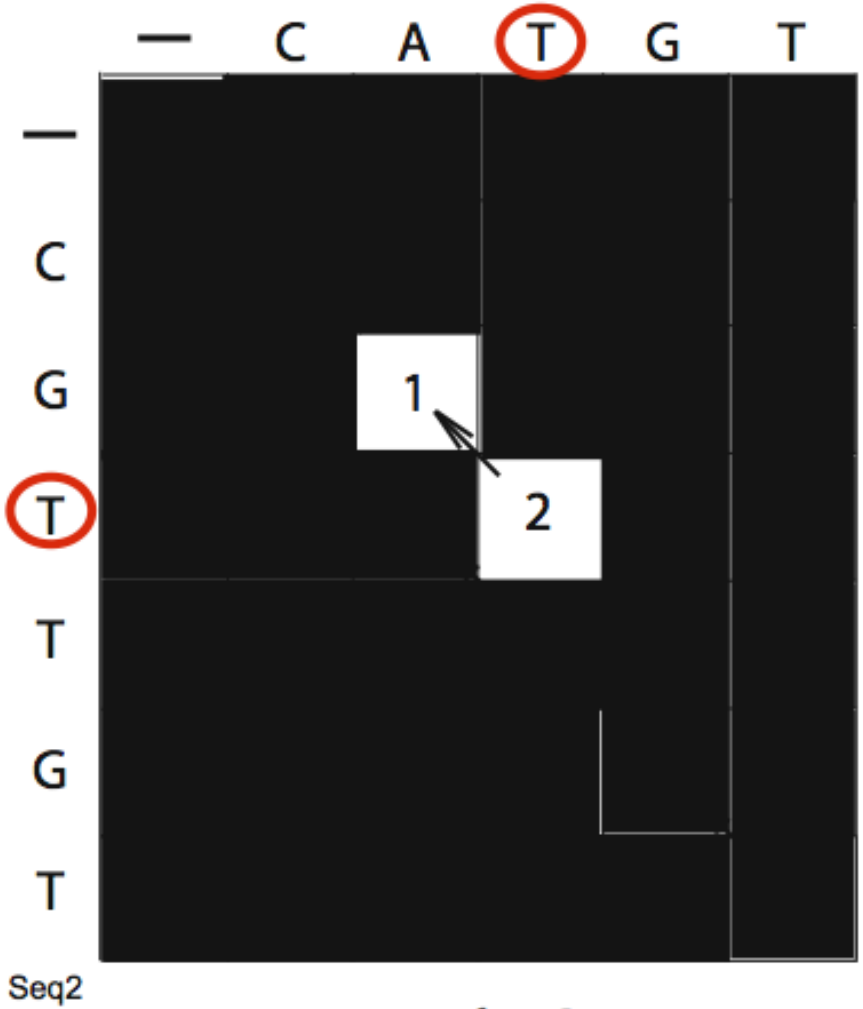
Πώς στοιχίζουμε



2←1

—	G	T	Seq1
I	I	I	
T	G	T	Seq2

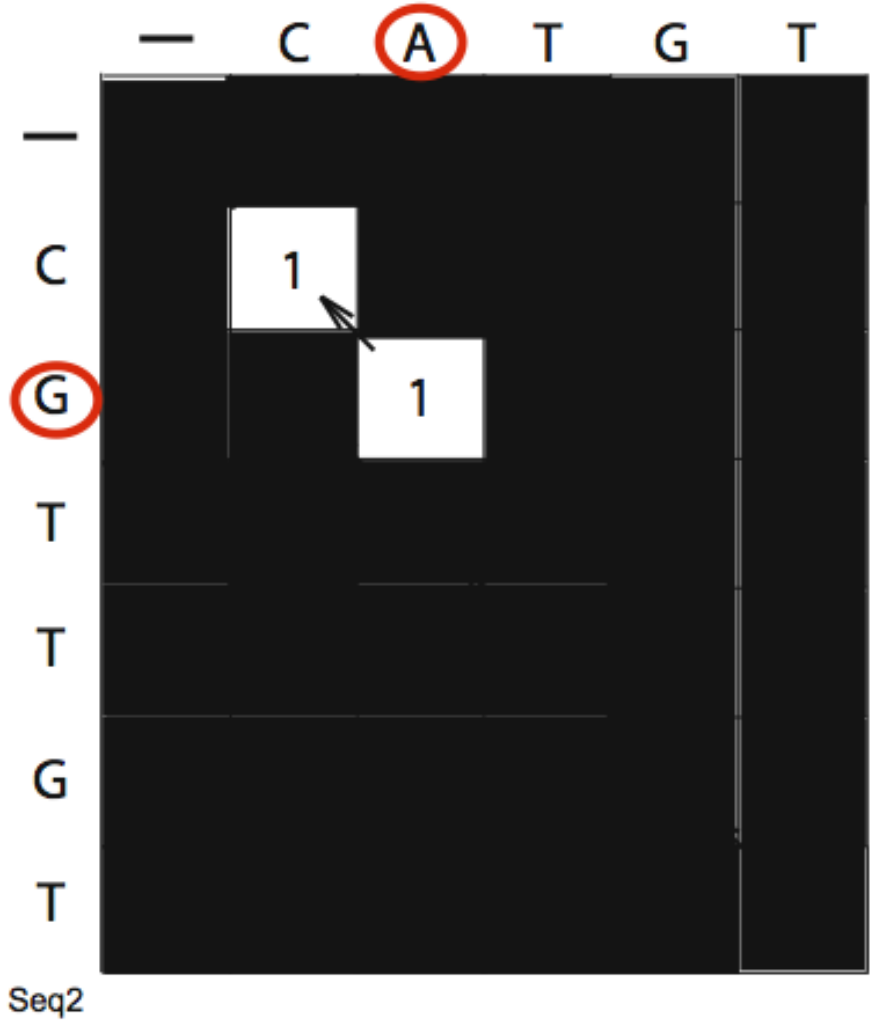
Πώς στοιχίζουμε



1←2

T	—	G	T	Seq1
I	I	I	I	
T	T	G	T	Seq2

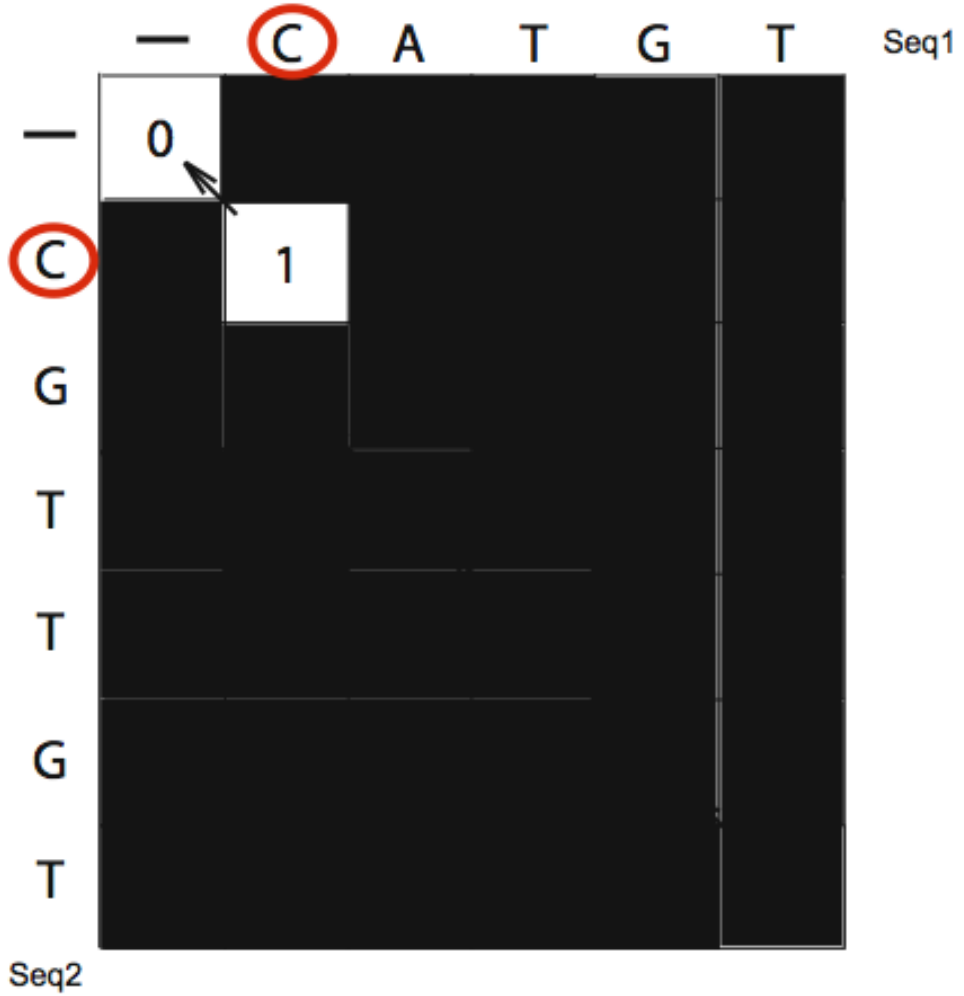
Πώς στοιχίζουμε



1←1

A	T	—	G	T	Seq1
	I		I	I	
G	T	T	G	T	Seq2

Πώς στοιχίζουμε



0←1

C	A	T	—	G	T	Seq1
I		I		I	I	
C	G	T	T	G	T	Seq2

Δυναμικός προγραμματισμός τοπική στοίχιση

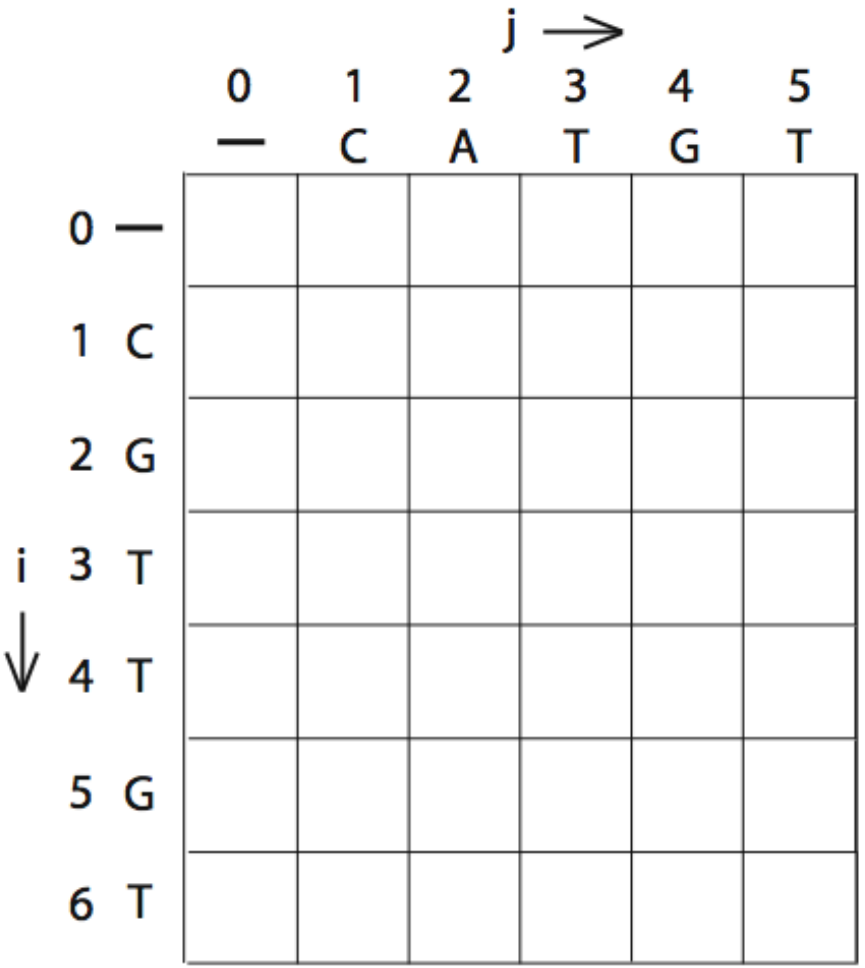
- Ενδείκνυται για
  - μακρομόρια διαφορετικού μεγέθους
  - Συντηρημένη μόνο μια μικρή περιοχή
  - Στοίχιση ώριμου mRNA με το γονίδιό του
  - 2 γονίδια με συντηρημένα εξόνια αλλά αποκλείνοντα ιντρόνια

- Αλγόριθμος Smith-Waterman (1981)

Δυναμικός προγραμματισμός τοπική στοίχιση

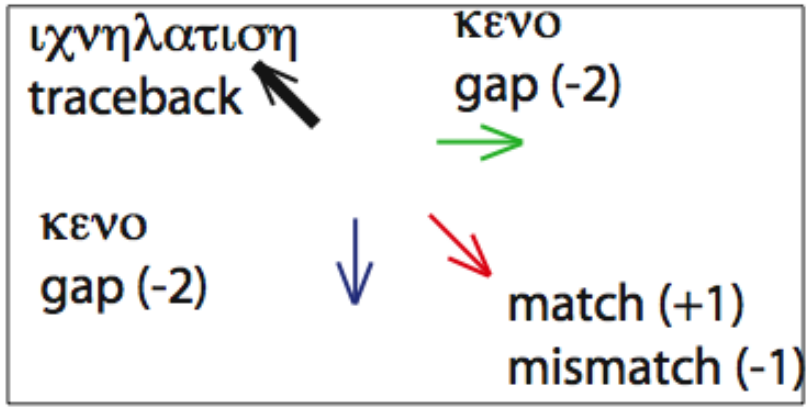
- Αλγόριθμος παρόμοιος με ολική στοίχιση
- Διαφορές:
  - Οι ασυμφωνίες δίνουν αρνητική βαθμολογία.
  - Όταν μια τιμή του πίνακα βγαίνει αρνητική, μηδενίζεται.

Δ.Π τοπική στοίχιση παράδειγμα (i)

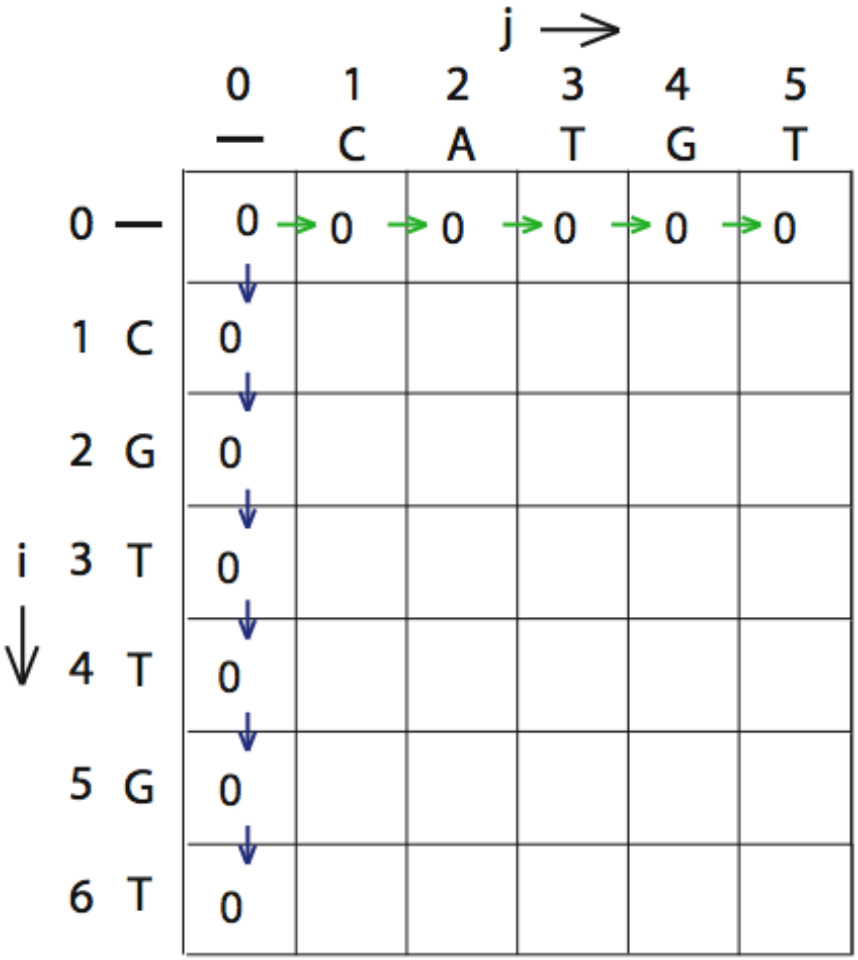


Scoring-System

match=1
mismatch=-1
gap=-2

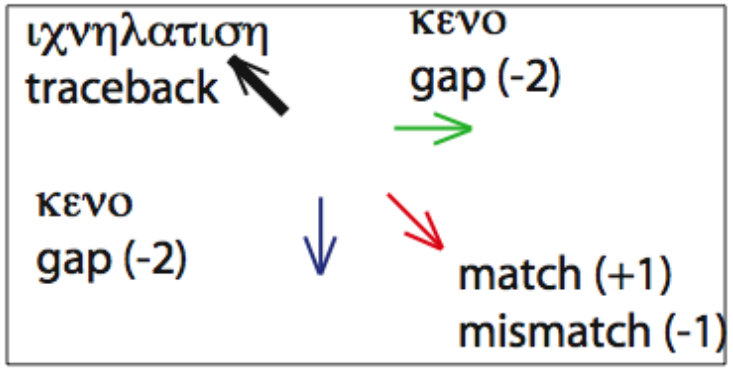


Δ.Π τοπική στοίχιση παράδειγμα (ii)



Scoring-System

match=1
mismatch=-1
gap=-2



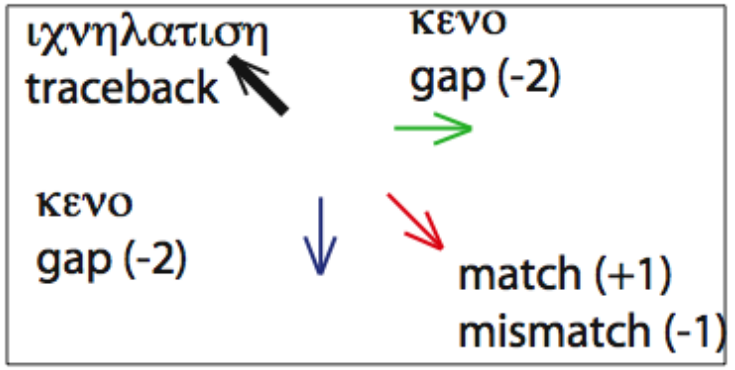
Δ.Π τοπική στοίχιση παράδειγμα (iii)

$S_{1,1}$ = MAX [0+1, 0-2, 0-2, 0] = 1

		j → 0 —	1 C	2 A	3 T	4 G	5 T
0	—	0	0	0	0	0	0
1	C	0	1				
2	G	0					
i ↓ 3	T	0					
4	T	0					
5	G	0					
6	T	0					

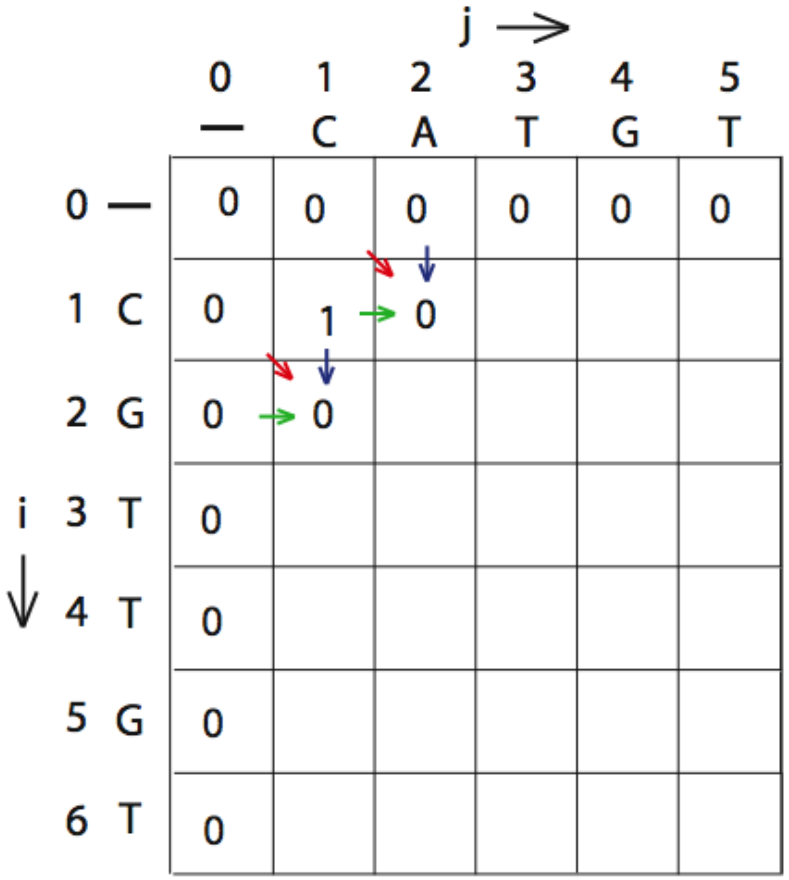
Scoring-System

match=1
mismatch=-1
gap=-2



Δ.Π τοπική στοίχιση παράδειγμα (iv)

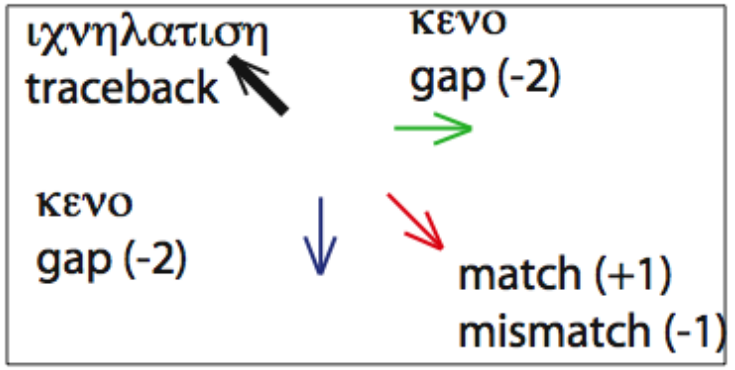
$S_{1,2}$ = MAX [0-1, 1-2, 0-2, 0] = 0



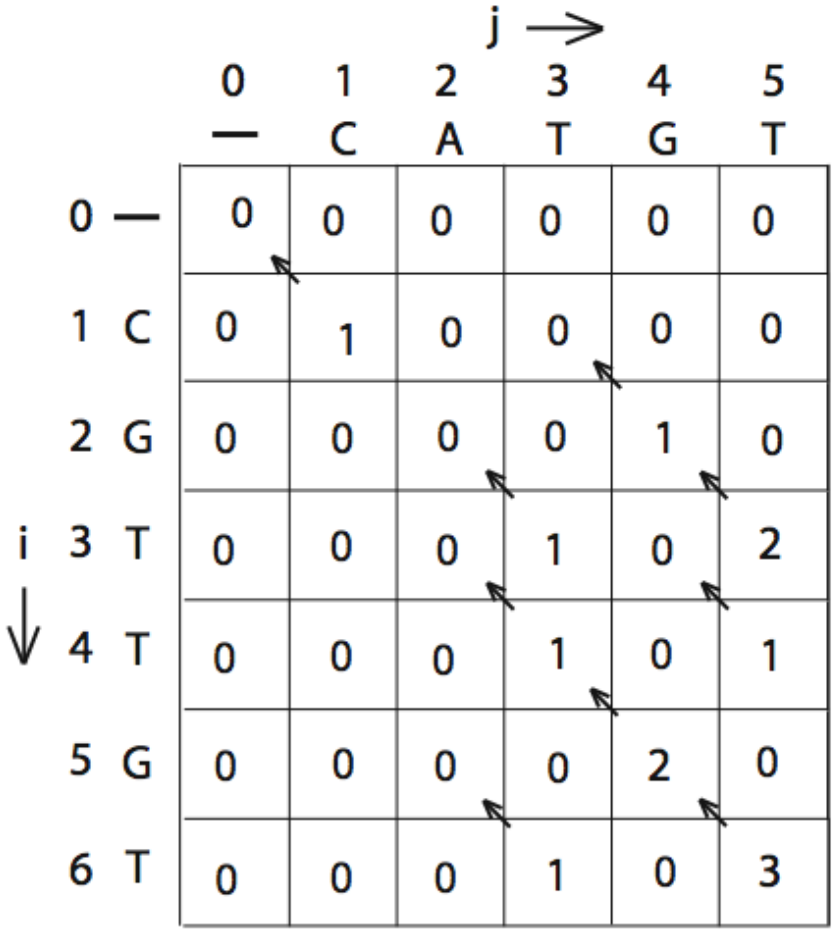
$S_{2,1}$ = MAX [0-1, 0-2, 1-2, 0] = 0

Scoring-System

match=1
mismatch=-1
gap=-2



Δ.Π τοπική στοίχιση παράδειγμα (v)



0←1←2←3

T	G	T
I	I	I
T	G	T