Βιοπληροφορική

Ενότητα 5: Πίνακες αντικατάστασης BLOSUM και οπτική σύγκριση αλληλουχιών

Αν. καθηγητής Αγγελίδης Παντελής
e-mail: paggelidis@uowm.gr
ΕΕΔΙΠ Μπέλλου Σοφία
e-mail: sbellou@uowm.gr

Τμήμα Μηχανικών Πληροφορικής και Τηλεπικοινωνιών

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

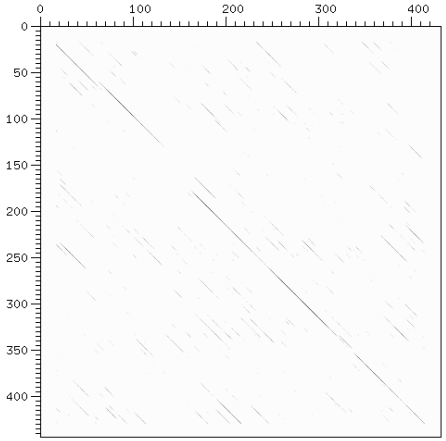
- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ψηφιακά Μαθήματα στο Πανεπιστήμιο Δυτικής Μακεδονίας**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σκοπός του μαθήματος

- Πίνακες αντικατάστασης PAM.
- The log odds form (the mutation data matrix) of PAM250.
- Πίνακες αντικατάστασης BLOSUM.
- Nucleic acid PAM Scoring Matrices.
- Οι ποινές για τα κενά.
- Στατιστική σημαντικότητα της στοίχισης αλληλουχιών.
- P-value, E-value, Z-score.
- Οπτική σύγκριση αλληλουχιών.
- Διαγράμματα πινάκων σημείων – Dot plots.
- Θόρυβος» στα dot plots.

Πίνακες αντικατάστασης BLOSUM και οπτική σύγκριση αλληλουχιών



Σοφία Μπέλλου, sbellou@uowm.gr

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	−2	1	3																		T
P	−3	1	0	6																	P
A	−2	1	1	1	2																A
G	−3	1	0	−1	1	5															G
N	−4	1	0	−1	0	0	2														N
D	−5	0	**0**	−1	0	1	2	4													D
E	−5	**0**	**0**	−1	0	0	1	3	4												E
Q	−5	−1	−1	**0**	0	−1	1	2	2	4											Q
H	−3	−1	−1	0	−1	−2	2	1	1	3	6										H
R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6									R
K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5								K
M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6							M
I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5						I
L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6					L
V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4				V
F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9			F
Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10		Y
W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Γιατί σύγκριση αλληλουχιών;;;

- Συμπεράσματα για λειτουργία ενός γονιδίου (μίας πρωτεΐνης).
  - Όταν δύο γονίδια έχουν μεγάλο ποσοστό ομοιότητας, τότε πιθανά κωδικοποιούν πρωτεΐνες με παρόμοια λειτουργία.
- Συμπεράσματα για σημαντικές περιοχές στην αλληλουχία γονιδίων/πρωτεϊνών.
  - Όταν πρωτεΐνες με συγκεκριμένο χαρακτηριστικό μοιράζονται κοινή περιοχή, τότε αυτή η κοινή περιοχή είναι πιθανά υπεύθυνη για το συγκεκριμένο χαρακτηριστικό.
- Συμπεράσματα για την εξελεγκτική απόσταση μεταξύ 3 ειδών.
  - Όταν η αλληλουχία μίας πρωτεΐνης είναι σχεδόν η ίδια μεταξύ 2 ειδών (ποντίκι και αρουραίος), τότε τα είδη αυτά είναι «κοντά».

Σύγκριση πρωτεϊνών - Εξαγωγή συμπερασμάτων για τη λειτουργία (1/4)

- **Protein A expressed in colon cancer:**

TCGCTGCGAAGGACATTTGGGCTGTGTGTGCGACGCGGGTCGGAGGGGCAGTCGGGGGAACC
GCGAAGAAGCCGAGGAGCCCGGAGCCCCGCGTGACGCTCCTCTCTCAGTCCAAAAGCGGCTTT
TGGTTCGGCGCAGAGAGACCCGGGGGTCTTCAGGACAGCGATTGTCATTGCTGAAGCTTTTCCT
CGAAAAGCGCCGCCCTGCCCTTGGCCCCGAGAA

- **Protein B expressed in colon cancer :**

CAGACAAAGAGCACCGCAGGGCCGATCACGCTGGGGGCGCTGAGGCCGGCCATGGTCATGGA
AGTGGGCACCCTGGACGCTGGAGGCCTGCGGGCGCTGCTGGGGGAGCGAGCGGCGCAATGCC
TGCTGCTGGACTGCCGCTCCTTCTTCGCTTTCAACGCCGGCCACATCGCCGGCTCTGTCAACGT
GCGCTTCAGCACCATCGTGCGGCGCCGGGCCAAGGGCGCCATGGGCCTGGAGCACATCGTGC
CCAACGCCGAGCTCCGCGGTCAGGACAGCGATTGTCATTGCTGA

- **Protein C expressed in colon cancer:**

CCGCCTGCTGGCCGGCGCCTACCACGCCGTGGTGTTGCTGGACGAGCGCAGCGCCGCCCTGG
ACGGCGCCAAGCGCGACGGCACCCTGGCCCTGGCGGCCGGCGCGCTCTGCCGCGAGGCGCG
CGCCGCGCAAGTCTTCTTCCTCAAAGGAGGATACGAAGCGTTTTCGGCTTCCTGCCCGGAGCTC
AGGACAGCGATTGTCATTGCTGATGTGCAGCAAACAGTCGACCCCCATGGGGCTCAGCCTTCCC
CTGAGTACTAGCGTCCCTGACAGCGCGGAATCTGGGTGCAGTT

Σύγκριση πρωτεϊνών - Εξαγωγή συμπερασμάτων για τη λειτουργία (2/4)

- **Protein D expressed in colon cancer:**

TAACTGCCTTGATCAACGTCTCAGCCAATTGTCCCAACCATTTTGAGGGTCACTACCAGTACAAG
AGCATCCCTGTGGAGGACAACCACAAGGCAGACATCAGCTCCTGGTTCAACGAGGCCATTGACT
TCATAGACTCCATCAAGAATGCTGGAGGAAGGGTGTTTGTCCACTGCCAGGCAGGCATTTCCCG
GTCAGCCACCATCTGCCTTGCTTACCTTATGAGGACTAATCGAGTCATCAGGACAGCGATTGTCA
TTGCTGAAGCTGGACGAGGCCTTTGAGTTTGTGAAGCAGAGGC

- **Protein E expressed in colon cancer:**

GAAGCATCATCTCTCCCAACTTCAGCTTCATGGGCCAGCTGCTGCAGTTTGAGTCCCAGGTGCTG
GCTCCGCACTGTTCGGCAGAGGCTGGGAGCCCCGCCATGGCTGTGCTCGACCGAGGCACCTCC
ACCACCACCGTGTTCAACTTATCAGGACAGCGATTGTCATTGCTGACCCCGTCTCCATCCCTGTC
CACTCCACGAACAGTGCGCTGAGCTACCTTCAGAGCCCCATTACGACCTCTCCCAGCTGCTGAA
AGGCCACGGGAGGTGAGGCTCTTCACATCCCATTGGGACTC

- **Protein F with unknown function:**

CATGCTCCTTGAGAGGAGAAATGCAATAACTCTGGGAGGGGCTCAGGACAGCGATTGTCATTGC
TGATCGAGAGGGCTGGTCCTTATTTATTTAACTTCACCCGAGTTCCTCTGGGTTTCTAAGCAGTTA
TGGTGATGACTTAGCGTCAAGACATTTGCTGAACTCAGCACATTCGGGACCAATATATAGTGGGT
ACATCAAGTCCATCTGACAAAATGGGGCAGAAGAGAAAGGACTCAGTGTGTGATCCGGTTTCTTT
TTGCTCGCCCCTGTTTTTTGTAGAATCTCTTCATGCTTGACATACCTACCAGTATTATTCCCGACG
ACACATATACATATGAGAATATACCTTATTTATTTTTGTGTAGGTGTCTGCCTTCACAAATGTCATT
GTCTACTCCTAGAAGAACCAAATACCTCAATTTTTGTTTTTGAGTACTGTACTATCCTGTAAATATA
TCTTAAGCAGGTTTGTTTTCA

Σύγκριση πρωτεϊνών - Εξαγωγή συμπερασμάτων για τη λειτουργία (3/4)

Protein A expressed in colon cancer:

TCGCTGCGAAGGACATTTGGGCTGTGTGTGCGACGCGGGTCGGAGGGGCAGTCGGGGGAACCGCGAAGAAGCC
GAGGAGCCCGGAGCCCCGCGTGACGCTCCTCTCTCAGTCCAAAAGCGGCTTTTGGTTCGGCGCAGAGAGACCCGG
GGGTCT**TCAGGACAGCGATTGTCATTGCTGA**AGCTTTTCCTCGAAAAGCGCCGCCCTGCCCTTGGCCCCGAGAA

Protein B expressed in colon cancer :

CAGACAAAGAGCACCGCAGGGCCGATCACGCTGGGGGCGCTGAGGCCGGCCATGGTCATGGAAGTGGGCACCCT
GGACGCTGGAGGCCTGCGGGCGCTGCTGGGGGAGCGAGCGGCGCAATGCCTGCTGCTGGACTGCCGCTCCTTCT
TCGCTTTCAACGCCGGCCACATCGCCGGCTCTGTCAACGTGCGCTTCAGCACCATCGTGCGGCGCCGGGCCAAGGG
CGCCATGGGCCTGGAGCACATCGTGCCCAACGCCGAGCTCCGCGG**TCAGGACAGCGATTGTCATTGCTGA**

Protein C expressed in colon cancer :

CCGCCTGCTGGCCGGCGCCTACCACGCCGTGGTGTTGCTGGACGAGCGCAGCGCCGCCCTGGACGGCGCCAAGC
GCGACGGCACCCTGGCCCTGGCGGCCGGCGCGCTCTGCCGCGAGGCGCGCGCCGCGCAAGTCTTCTTCCTCAAAG
GAGGATACGAAGCGTTTTCGGCTTCCTGCCCGGAGC**TCAGGACAGCGATTGTCATTGCTGA**TGTGCAGCAAACAG
TCGACCCCCATGGGGCTCAGCCTTCCCCTGAGTACTAGCGTCCCTGACAGCGCGGAATCTGGGTGCAGTT

Σύγκριση πρωτεϊνών - Εξαγωγή συμπερασμάτων για τη λειτουργία (4/4)

Protein D expressed in colon cancer :

TAACTGCCTTGATCAACGTCTCAGCCAATTGTCCCAACCATTTTGAGGGTCACTACCAGTACAAGAGCATCCCTGTGGAGGACAACCAC
AAGGCAGACATCAGCTCCTGGTTCAACGAGGCCATTGACTTCATAGACTCCATCAAGAATGCTGGAGGAAGGGTGTTTGTCCACTGCC
AGGCAGGCATTTCCCGGTCAGCCACCATCTGCCTTGCTTACCTTATGAGGACTAATCGAGTCA**TCAGGACAGCGATTGTCATTGCTGA**A
GCTGGACGAGGCCTTTGAGTTTGTGAAGCAGAGGC

Protein E expressed in colon cancer :

GAAGCATCATCTCTCCCAACTTCAGCTTCATGGGCCAGCTGCTGCAGTTTGAGTCCCAGGTGCTGGCTCCGCACTGTTCGGCAGAGGC
TGGGAGCCCCGCCATGGCTGTGCTCGACCGAGGCACCTCCACCACCACCGTGTTCAACTTA**TCAGGACAGCGATTGTCATTGCTGA**CC
CCGTCTCCATCCCTGTCCACTCCACGAACAGTGCGCTGAGCTACCTTCAGAGCCCCATTACGACCTCTCCCAGCTGCTGAAAGGCCACG
GGAGGTGAGGCTCTTCACATCCCATTGGGACTC

Protein F with unknown function:

CATGCTCCTTGAGAGGAGAAATGCAATAACTCTGGGAGGGGC**TCAGGACAGCGATTGTCATTGCTGA**TCGAGAGGGCTGGTCCTTAT
TTATTTAACTTCACCCGAGTTCCTCTGGGTTTCTAAGCAGTTATGGTGATGACTTAGCGTCAAGACATTTGCTGAACTCAGCACATTCGG
GACCAATATATAGTGGGTACATCAAGTCCATCTGACAAAATGGGGCAGAAGAGAAAGGACTCAGTGTGTGATCCGGTTTCTTTTTGCTC
GCCCCTGTTTTTTGTAGAATCTCTTCATGCTTGACATACCTACCAGTATTATTCCCGACGACACATATACATATGAGAATATACCTTATTTATT
TTTGTGTAGGTGTCTGCCTTCACAAATGTCATTGTCTACTCCTAGAAGAACCAAATACCTCAATTTTTGTTTTTGAGTACTGTACTATCCTG
TAAATATATCTTAAGCAGGTTTGTTTTCA

Συμπέρασμα: Πιθανά και η πρωτεΐνη F παίζει ρόλο στον καρκίνο παχύ εντέρου

Σύγκριση πρωτεϊνών - Εξαγωγή συμπερασμάτων για λειτουργικό τμήμα πρωτεΐνης (1/2)

Protein A with membrane localization:

TCGCTGCGAAGGACATTTGGGCTGTGTGTGCGACGCGGGTCGGAGGGGCAGTCGGGGGAACCGCGAAGAAGCCGAGGAGCCCGGAGCCCCGCGTGACG
CTCCTCTCTCAGTCCAAAAGCGGCTTTTGGTTCGGCGCAGAGAGACCCGGGGGTCTTCAGGACAGCGATTGTCATTGCTGAAGCTTTTCCTCGAAAAGCGCC
GCCCTGCCCTTGGCCCCGAGAACAGACAAAGAGCACCGCAGGGCCGATCACGCTGGGGGCGCTGAGGCCGGCCATGGTCATGGAAGTGGGCACCCTGGAC
GCTGGAGGCCTGCGGGCGCTGCTGGGGGAGCGAGCGGCGCAATGCCTGCTGCTGGACTGCCGCTCCTTCTTCGCTTTCAACGCCGGCCACATCGCCGGCTC
TGTCAACGTGCGCTTCAGCACCATCGTGCGGCGCCGGGCCAAGGGCGCCATGGGCCTGGAGCACATCGTGCCCAACGCCGAGCTCCGCGGTCAGGACAGC
GATTGTCATTGCTGACCGCCTGCTGGCCGGCGCCTACCACGCCGTGGTGTTGCTGGACGAGCGCAGCGCCGCCCTGGACGGCGCCAAGCGCGACGGCACCC
TGGCCCTGGCGGCCGGCGCGCTCTGCCGCGAGGCGCGCGCCGCGCAAGTCTTCTTCCTCAAAGGAGGATACGAAGCGTTTTCGGCTTCCTGCCCGGAGCTC
AGGACAGCGATTGTCATTGCTGATGTGCAGCAAACAGTCGACCCCCATGGGGCTCAGCCTTCCCCTGAGTACTAGCGTCCCTGACAGCGCGGAATCTGGGTG
CAGTTCCTGCAGTACCCCACTCTACGATCAGGGTGGCCCGGTGGAAATCCTGCCCTTTCTGTACCTGGGCAGTGCGTATCACGCTTCCCGCAAGGACATGCTG
GATGCCTTGGGCA

Protein B with membrane localization :

TAACTGCCTTGATCAACGTCTCAGCCAATTGTCCCAACCATTTTGAGGGTCACTACCAGTACAAGAGCATCCCTGTGGAGGACAACCACAAGGCAGACATCA
GCTCCTGGTTCAACGAGGCCATTGACTTCATAGACTCCATCAAGAATGCTGGAGGAAGGGTGTTTGTCCACTGCCAGGCAGGCATTTCCCGGTCAGCCACCA
TCTGCCTTGCTTACCTTATGAGGACTAATCGAGTCATCAGGACAGCGATTGTCATTGCTGAAGCTGGACGAGGCCTTTGAGTTTGTGAAGCAGAGGCGAAGC
ATCATCTCTCCCAACTTCAGCTTCATGGGCCAGCTGCTGCAGTTTGAGTCCCAGGTGCTGGCTCCGCACTGTTCGGCAGAGGCTGGGAGCCCCGCCATGGCT
GTGCTCGACCGAGGCACCTCCACCACCACCGTGTTCAACTTATCAGGACAGCGATTGTCATTGCTGACCCCGTCTCCATCCCTGTCCACTCCACGAACAGTGC
GCTGAGCTACCTTCAGAGCCCCATTACGACCTCTCCCAGCTGCTGAAAGGCCACGGGAGGTGAGGCTCTTCACATCCCATTGGGACTCCATGCTCCTTGAGA
GGAGAAATGCAATAACTCTGGGAGGGGCTCAGGACAGCGATTGTCATTGCTGATCGAGAGGGCTGGTCCTTATTTATTTAACTTCACCCGAGTTCCTCTGGG
TTTCTAAGCAGTTATGGTGATGACTTAGCGTCAAGACATTTGCTGAACTCAGCACATTCGGGACCAATATATAGTGGGTACATCAAGTCCATCTGACAAAATG
GGGCAGAAGAGAAAGGACTCAGTGTGTGATCCGGTTTCTTTTTGCTCGCCCCTGTTTTTTGTAGAATCTCTTCATGCTTGACATACCTACCAGTATTATTCCCG
ACGACACATATACATATGAGAATATACCTTATTTATTTTTGTGTAGGTGTCTGCCTTCACAAATGTCATTGTCTACTCCTAGAAGAACCAAATACCTCAATTTTT
GTTTTTGAGTACTGTACTATCCTGTAAATATATCTTAAGCAGGTTTGTTTTCA

Συμπέρασμα: Πιθανά το κοινό τμήμα να είναι υπεύθυνο για τη μεμβρανική εντόπιση των πρωτεϊνών

Σύγκριση πρωτεϊνών - Εξαγωγή συμπερασμάτων για λειτουργικό τμήμα πρωτεΐνης (2/2)

Protein A with membrane localization:

TCGCTGCGAAGGACATTTGGGCTGTGTGTGCGACGCGGGTCGGAGGGGCAGTCGGGGGAACCGCGAAGAAGCCGAGGAGCCCGGAGCCCCGCGTGACGCTCCTC
TCTCAGTCCAAAAGCGGCTTTTGGTTCGGCGCAGAGAGACCCGGGGGTCT**TCAGGACAGCGATTGTCATTGCTGA**AGCTTTTCCTCGAAAAGCGCCGCCCTGCCCTT
GGCCCCGAGAACAGACAAAGAGCACCGCAGGGCCGATCACGCTGGGGGCGCTGAGGCCGGCCATGGTCATGGAAGTGGGCACCCTGGACGCTGGAGGCCTGCGG
GCGCTGCTGGGGGAGCGAGCGGCGCAATGCCTGCTGCTGGACTGCCGCTCCTTCTTCGCTTTCAACGCCGGCCACATCGCCGGCTCTGTCAACGTGCGCTTCAGCAC
CATCGTGCGGCGCCGGGCCAAGGGCGCCATGGGCCTGGAGCACATCGTGCCCAACGCCGAGCTCCGCGG**TCAGGACAGCGATTGTCATTGCTGA**CCGCCTGCTGG
CCGGCGCCTACCACGCCGTGGTGTTGCTGGACGAGCGCAGCGCCGCCCTGGACGGCGCCAAGCGCGACGGCACCCTGGCCCTGGCGGCCGGCGCGCTCTGCCGCG
AGGCGCGCGCCGCGCAAGTCTTCTTCCTCAAAGGAGGATACGAAGCGTTTTCGGCTTCCTGCCCGGAGC**TCAGGACAGCGATTGTCATTGCTGA**TGTGCAGCAAAC
AGTCGACCCCCATGGGGCTCAGCCTTCCCCTGAGTACTAGCGTCCCTGACAGCGCGGAATCTGGGTGCAGTTCCTGCAGTACCCCACTCTACGATCAGGGTGGCCCG
GTGGAAATCCTGCCCTTTCTGTACCTGGGCAGTGCGTATCACGCTTCCCGCAAGGACATGCTGGATGCCTTGGGCA

Protein B with membrane localization :

TAACTGCCTTGATCAACGTCTCAGCCAATTGTCCCAACCATTTTGAGGGTCACTACCAGTACAAGAGCATCCCTGTGGAGGACAACCACAAGGCAGACATCAGCTCCT
GGTTCAACGAGGCCATTGACTTCATAGACTCCATCAAGAATGCTGGAGGAAGGGTGTTTGTCCACTGCCAGGCAGGCATTTCCCGGTCAGCCACCATCTGCCTTGCTT
ACCTTATGAGGACTAATCGAGTCA**TCAGGACAGCGATTGTCATTGCTGA**AGCTGGACGAGGCCTTTGAGTTTGTGAAGCAGAGGCGAAGCATCATCTCTCCCAACTT
CAGCTTCATGGGCCAGCTGCTGCAGTTTGAGTCCCAGGTGCTGGCTCCGCACTGTTCGGCAGAGGCTGGGAGCCCCGCCATGGCTGTGCTCGACCGAGGCACCTCC
ACCACCACCGTGTTCAACTTA**TCAGGACAGCGATTGTCATTGCTGA**CCCCGTCTCCATCCCTGTCCACTCCACGAACAGTGCGCTGAGCTACCTTCAGAGCCCCATTA
CGACCTCTCCCAGCTGCTGAAAGGCCACGGGAGGTGAGGCTCTTCACATCCCATTGGGACTCCATGCTCCTTGAGAGGAGAAATGCAATAACTCTGGGAGGGGC**TC**
AGGACAGCGATTGTCATTGCTGATCGAGAGGGCTGGTCCTTATTTATTTAACTTCACCCGAGTTCCTCTGGGTTTCTAAGCAGTTATGGTGATGACTTAGCGTCAAGA
CATTTGCTGAACTCAGCACATTCGGGACCAATATATAGTGGGTACATCAAGTCCATCTGACAAAATGGGGCAGAAGAGAAAGGACTCAGTGTGTGATCCGGTTTCTT
TTTGCTCGCCCCTGTTTTTTGTAGAATCTCTTCATGCTTGACATACCTACCAGTATTATTCCCGACGACACATATACATATGAGAATATACCTTATTTATTTTTGTGTAGG
TGTCTGCCTTCACAAATGTCATTGTCTACTCCTAGAAGAACCAAATACCTCAATTTTTGTTTTTGAGTACTGTACTATCCTGTAAATATATCTTAAGCAGGTTTGTTTTCA

Συμπέρασμα: Πιθανά το κοινό τμήμα να είναι υπεύθυνο για τη μεμβρανική εντόπιση των πρωτεϊνών

Σύστημα βαθμονόμησης στοίχισης - DNA & RNA

Sequence 1: ATCGGATCT

Sequence 2: ACGGACT **ΠΡΟΣΟΧΗ: Συγκρίνουμε νουκλεοτίδια, μόνο 4 πιθανές βάσεις**

A	T	C	G	G	A	T	-	C	T
|		|		|			|	|
A	-	C	-	G	G	-	A	C	T

A	T	C	G	G	A	T	C	T
|		|	|	|	|		|	|
A	-	C	G	G	A	-	C	T

- Πιθανό σύστημα βαθμονόμησης:
  - όμοιο κατάλοιπο: +2
  - διαφορετικό κατάλοιπο: -1
  - κενό: –2
- **Alignment 1:** 5 x 2 – 1(1) – 4(2) = 10 – 1 – 8 = 1
- **Alignment 2:** 7 x 2 – 0(1) – 2 (2) = 14 – 0 – 4 = 10

Στοίχιση πρωτεϊνών;;;

A	S	K	T	M	P	I
I	I	?	?	I	I	I
A	S	R	H	M	P	I

A	S	K	T	M	P	I
I	I	?	?	I	I	|
A	S	Y	H	M	P	I

- **Πώς στοιχίζουμε πρωτεΐνες.**
- **Αμινοξέα με παρόμοιες ιδιότητες.**
- **Αμινοξέα με διαφορετικές ιδιότητες.**

Πίνακες βαθμονόμησης (scoring matrices) ή πίνακες αντικατάστασης

- Για την αντικατάσταση αμινοξέων:
  - PAM.
  - BLOSUM.
- Για την αντικατάσταση νουκλεοτιδίων (DNA):
  - Το DNA είναι λιγότερο συντηρημένο από τις πρωτεΐνες.
  - Δεν είναι αποτελεσματικό να συγκρίνουμε κωδικοποιούσες περιοχές, από τις οποίες προκύπτουν οι πρωτεΐνες, σε επίπεδο νουκελοτιδίων.

The log odds form (the mutation data matrix) of PAM250

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	−2	1	3																		T
P	−3	1	0	6																	P
A	−2	1	1	1	2																A
G	−3	1	0	−1	1	5															G
N	−4	1	0	−1	0	0	2														N
D	−5	0	**0**	−1	0	1	2	4													D
E	−5	**0**	**0**	−1	0	0	1	3	4												E
Q	−5	−1	−1	**0**	0	−1	1	2	2	4											Q
H	−3	−1	−1	0	−1	−2	2	1	1	3	6										H
R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6									R
K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5								K
M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6							M
I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5						I
L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6					L
V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4				V
F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9			F
Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10		Y
W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Tryptophan: W

Cysteine: C

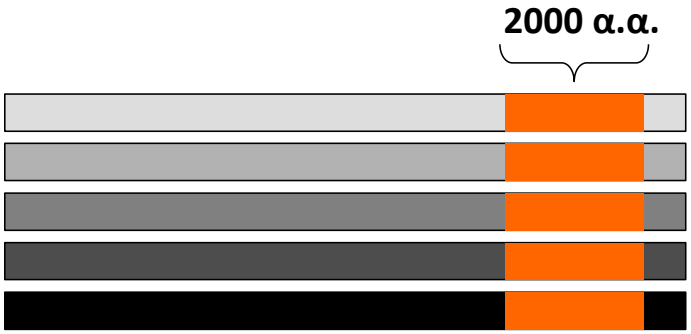
Πίνακες αντικατάστασης PAM (Percent Accepted Mutation - PAM or Dayhoff Matrices)

- Μελετήθηκαν από την Margaret Dayhoff.
- Αξιολογούν - βαθμολογούν την αντικατάσταση ενός αμινοξέος από ένα άλλο.
- Αξιολογήθηκαν:
  - 1572 αμινοξικές αντικαταστάσεις από,
  - 71 πρωτεϊνικά γκρουπ με,
  - 85% ομοιότητα τουλάχιστον ,
- Η κατασκευή τους βασίζεται στις:
  - στοιχίσεις πολλών **όμοιων πρωτεϊνών**, μικρής εξελεγκτικής απόστασης
  - αποδεκτές σημειακές μεταλλάξεις - accepted mutations (αντικατάσταση αμινοξέος από κάποιο άλλο που δεν επηρεάζει τη λειτουργία της πρωτεΐνης).

Πίνακες αντικατάστασης BLOSUM (BLOcks amino acid SUbstitution Matrices)

- Αναπτύχθηκαν από τους Stephen και Georgia Henikoff, 1992.
- Βασίζονται στην απαρίθμηση αντικατάστασης αμινοξέων από άλλα αμινοξέα σε συντηρημένες περιοχές αλληλουχιών.
  - **μικρής ομοιότητας**, και
  - **απομακρυσμένης εξελεγκτικής σχέσης.**
- Εξετάστηκαν πάνω από 2000 συντηρημένες περιοχές (blocks) που ανήκουν σε περισσότερες από 500 οικογένειες συσχετιζόμενων πρωτεϊνών.
- **Οι περιοχές (blocks) δεν περιέχουν κενά.**

BLOSUM MATRICES



: Συντηρημένη περιοχή

- Οι περιοχές βρέθηκαν σε πρωτεϊνικές βάσεις δεδομένων και αντιπροσωπεύουν περισσότερες από 500 οικογένειες πρωτεϊνών.
- Οι πίνακες BLOSUM βασίζονται σε ανάλυση διαφορετικού τύπου αλληλουχιών και μεγαλύτερου αριθμού δεδομένων από τους πίνακες PAM.
- **Βάση δεδομένων: PROSITE.**
- **Prosite:** Database with protein domains. It provides list of proteins that are in the same family because they have a similar biochemical action. For each family, a pattern of amino acids that are characteristic of that function is provided.

Πίνακες αντικατάστασης BLOSUM (BLOcks SUbstitution Matrices)

- **BLOSUMn**: Το n δείχνει το ποσοστό ομοιότητας των αλληλουχιών που χρησιμοποιήθηκαν για να προκύψει ο συγκεκριμένος πίνακας.
- Αλληλουχίες με ομοιότητα τουλάχιστον 62% δίνουν τον πίνακα αντικατάστασης **BLOSUM62.**
- **BLOSUM με μεγάλο n: αλληλουχίες με μεγάλη ομοιότητα.**
- BLOSUM με μεγάλο n -> PAM με μικρό n.
- Γενικά, οι πίνακες BLOSUM είναι καλύτεροι για την εύρεση τοπικών στοιχίσεων.

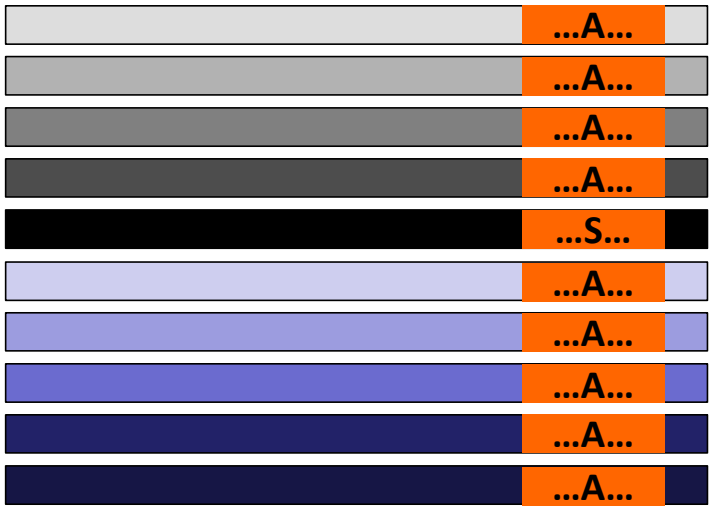
Παράδειγμα BLOSUM62

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

Ερμηνεία τιμής του πίνακα BLOSUM (1/4)

- Η τιμή κάθε ζεύγους αμινοξέος δείχνει την πιθανότητα να εμφανιστεί η συγκεκριμένη αντικατάσταση κατά την εξέλιξη.
- Δίνεται από το λόγο:
  - **λ = συχνότητα εμφάνισης του ζεύγους / αναμενόμενη συχνότητα εμφάνισης του ζεύγους.**
  - **$\lambda = q_{ij}/e_{ij}$,**
    - **q: expected, and**
    - **e: observed frequency of appearance.**

Ερμηνεία τιμής του πίνακα BLOSUM (2/4)



Pairs	Frequency of appearance, 1st	Frequency of appearance, 2nd
AA	9	9
AS	9	1
SS	1	1

- **Pair frequency c_{ij},** for each pair i and j, for each column k using:
  - For "like" comparisons $c_{ii}^{(k)} = \binom{n_i}{2}$
  - For "unlike" comparisons $c_{ij}^{(k)} = n_i n_j$

Ερμηνεία τιμής του πίνακα BLOSUM (3/4)

...A...
...A...
...A...
...A...
...S...
...A...
...A...
...A...
...A...
...A...

Pairs	Frequency of appearance, 1st	Frequency of appearance, 2nd
AA	9	9
AS	9	1
SS	1	1

$$c_{ii}^{(k)} = \binom{n_i}{2}$$

$$c_{ij}^{(k)} = n_i n_j$$

$$c_{AA} = \binom{9}{2} = \frac{9(9-1)}{2} = 36$$

$$c_{SS} = \binom{1}{2} = \frac{1(1-1)}{2} = 0$$

$$c_{AS} = 9 \cdot 1 = 9$$

c_{ij}	A	S
A	36	
S	9	0

Ερμηνεία τιμής του πίνακα BLOSUM - Κανονικοποίηση

...A...
...A...
...A...
...A...
...S...
...A...
...A...
...A...
...A...
...A...

c_{ij}	A	S
A	36	
S	9	0

$$T = \sum c_{ij} = w\frac{n(n-1)}{2}$$

$w : columns$

$n : sequences$

Normalization factor:

$$T = 1\frac{10(10-1)}{2} = 45$$

Observed pair frequency, q_{ij}:

$$q_{AA} = \frac{36}{45} = 0.8$$

$$q_{AS} = \frac{9}{45} = 0.2$$

$$q_{SS} = \frac{0}{45} = 0$$

q_{ij}	A	S
A	0.8	
S	0.2	0

Ερμηνεία τιμής του πίνακα BLOSUM (4/4)

- Η τιμή κάθε ζεύγους αμινοξέος δείχνει την πιθανότητα να εμφανιστεί η συγκεκριμένη αντικατάσταση κατά την εξέλιξη.
- Δίνεται από το λόγο:
  - **λ = συχνότητα εμφάνισης του ζεύγους / αναμενόμενη συχνότητα εμφάνισης του ζεύγους.**
  - **$\lambda = q_{ij}/e_{ij}$,**
    - **q: observed, and**
    - **e: expected frequency of appearance.**

Ερμηνεία τιμής του πίνακα BLOSUM Expected pair frequency, eij (1/3)

$e_{ii} = p_i^2$, for “like” comparisons

$e_{ij} = 2 p_i p_j$, for “unlike” comparisons

where, p_i: expected probability of occurrence of the i*th* residue

$$p_i = q_{ii} + \sum_{j \neq i} \frac{q_{ij}}{2}$$

Ερμηνεία τιμής του πίνακα BLOSUM Expected pair frequency, eij (2/3)

$$e_{ij} = 2 p_i p_j$$

$$p_i = q_{ii} + \sum_{j \neq i} \frac{q_{ij}}{2} \qquad e_{ii} = p_i^2$$

q_{ij}	A	S
A	0.8	
S	0.2	0

Για κάθε κατάλοιπο

$$p_A = q_{AA} + \frac{q_{AS}}{2} = 0.8 + \frac{0.2}{2} = 0.9$$

$$p_S = q_{SS} + \frac{q_{AS}}{2} = 0 + \frac{0.2}{2} = 0.1$$

Για κάθε ζεύγος

$$e_{AA} = p_A^2 = 0.9^2 = 0.81$$

$$e_{SS} = p_S^2 = 0.01$$

$$e_{AS} = 2 p_A p_S = 2 \cdot 0.9 \cdot 0.1 = 0.18$$

Ερμηνεία τιμής του πίνακα BLOSUM Expected pair frequency, eij (3/3)

Observed frequency

q_{ij}	A	S
A	0.8	
S	0.2	0

Expected frequency

$$e_{AA} = p_A^2 = 0.9^2 = 0.81$$

$$e_{SS} = p_S^2 = 0.01$$

$$e_{AS} = 2 p_A p_S = 2 \cdot 0.9 \cdot 0.1 = 0.18$$

Entry: S = $2\log_2$(observed/expected)

$$Entry:$$

$$S_{AA} = 2\log_2 \frac{observed}{\exp ected} = 2 \cdot \log_2 \frac{0.8}{0.81} = -0.03$$

$$S_{AS} = 2\log_2 \frac{observed}{\exp ected} = 2 \cdot \log_2 \frac{0.2}{0.18} = 0.3$$

$$S_{SS} = 2\log_2 \frac{observed}{\exp ected} = 2 \cdot \log_2 \frac{0}{0.01} = ?$$

Δεν παρέχονται πληροφορίες για το ζεύγος SS

Παράδειγμα - Pair frequency cij

Sequence 1	A	A	I
Sequence 2	S	A	L
Sequence 3	T	A	L
Sequence 4	T	A	V
Sequence 5	A	A	L

$$c_{ii}^{(k)} = \binom{n_i}{2} \qquad c_{ij}^{(k)} = n_i n_j$$

Pairs	n_i	n_j
AT	2	2
IL	1	3
LL	3	3
AA	2	2
AA	5	5

Pair frequency c_{ij}, for each pair i and j

c_{ij}	A	I	L	S	T	V
A	10+1					
I	0	0				
L	0	3	3			
S	2	0	0	0		
T	4	0	0	2	1	
V	0	1	3	0	0	0

Παράδειγμα - Normalization

Sequence 1	A	A	I
Sequence 2	S	A	L
Sequence 3	T	A	L
Sequence 4	T	A	V
Sequence 5	A	A	L

$$T = \sum c_{ij} = w\frac{n(n-1)}{2}$$

$w : columns$

$n : sequences$

Normalization factor, T:

$$T = 3\frac{5(5-1)}{2} = 30$$

Pair frequency cij, for each pair i and j

c_{ij}	A	I	L	S	T	V
A	10+1					
I	0	0				
L	0	3	3			
S	2	0	0	0		
T	4	0	0	2	1	
V	0	1	3	0	0	0

cij / 30 = q_{ij}

q_{ij}	A	I	L	S	T	V
A	0.366					
I	0	0				
L	0	0.1	0.1			
S	0.066	0	0	0		
T	0.133	0	0	0.066	0.033	
V	0	0.033	0.1	0	0	0

Παράδειγμα - Observed frequency qij

Sequence 1	A	A	I
Sequence 2	S	A	L
Sequence 3	T	A	L
Sequence 4	T	A	V
Sequence 5	A	A	L

Observed normalized pair frequency c_{ij}, for each pair i and j

c_{ij}	A	I	L	S	T	V
A	0.366					
I	0	0				
L	0	0.1	0.1			
S	0.066	0	0	0		
T	0.133	0	0	0.066	0.033	
V	0	0.033	0.1	0	0	0

Παράδειγμα - Expected frequency of ij pair, eij

$$e_{ii} = p_i^2$$

$$e_{ij} = 2 p_i p_j$$

Αναμενόμενη συχνότητα εμφάνισης ζεύγους

$$p_i = q_{ii} + \sum_{j \neq i} \frac{q_{ij}}{2}$$

Πιθανότητα εμφάνισης ith καταλοίπου

q_{ij}	A	I	L	S	T	V
A	0.366					
I	0	0				
L	0	0.1	0.1			
S	0.066	0	0	0		
T	0.133	0	0	0.066	0.033	
V	0	0.033	0.1	0	0	0

	Πιθανότητα εμφάνισης ith καταλοίπου
p_A	0.366+(0.066+0.133)/2=**0.466**
p_I	0+(0.1+0.033)/2=**0.066**
p_L	0.1+(0.1+0.1)/2=**0.2**
p_S	0+(0.066+0.066)/2=**0.066**
p_T	0.033+(0.133+0.066)/2=**0.133**
p_V	0+(0.033+0.1)/2=**0.066**

Expected frequency of ij pair, eij

$$e_{ii} = p_i^2 \qquad e_{ij} = 2p_i p_j$$

$$p_i = q_{ii} + \sum_{j \neq i} \frac{q_{ij}}{2}$$

$e_{AA}=p_A^2=0.466^2=0.217$

$e_{AI}=2p_A\, p_I=2\cdot0.466\cdot0.066=0.062$

$e_{AI}=2p_A\, p_L=2\cdot0.466\cdot0.2=0.186$

$e_{II}=p_I^2=0.066^2=0.217$

p_i	Πιθανότητα εμφάνισης ith καταλοίπου
p_A	0.366+(0.066+0.133)/2=**0.466**
p_I	0+(0.1+0.033)/2=**0.066**
p_L	0.1+(0.1+0.1)/2=**0.2**
p_S	0+(0.066+0.066)/2=**0.066**
p_T	0.033+(0.133+0.066)/2=**0.133**
p_V	0+(0.033+0.1)/2=**0.066**

Expected frequency of ij pair, e_{ij}

e_{ij}	A	I	L	S	T	V
A	0.217					
I	0.062	0.004				
L	0.186	0.026	0.04			
S	0.062	0.198	0.026	0.004		
T	0.124	0.018	0.053	0.018	0.018	
V	0.062	0.009	0.027	0.009	0.018	0.004

Entry in BLOSUM matrix, Sij

Observed normalized pair frequency cij, for each pair i and j

	A	I	L	S	T	V
A	0.366					
I	0	0				
L	0	0.1	0.1			
S	0.066	0	0	0		
T	0.133	0	0	0.066	0.033	
V	0	0.033	0.1	0	0	0

Expected frequency of ij pair, e_{ij}

e_{ij}	A	I	L	S	T	V
A	0.217					
I	0.062	0.004				
L	0.186	0.026	0.04			
S	0.062	0.198	0.026	0.004		
T	0.124	0.018	0.053	0.018	0.018	
V	0.062	0.009	0.027	0.009	0.018	0.004

$$Entry:$$

$$S_{ij} = 2\log_2 \frac{observed}{\exp ected}$$

$$S_{AA} = 2\log_2 \frac{0.366}{0.217} = 2$$

$$S_{SA} = 2\log_2 \frac{0.066}{0.062} = 0$$

Finally

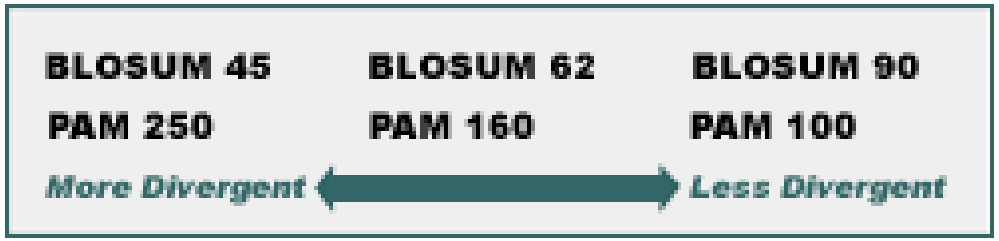
S_{ij}	A	I	L	S	T	V
A	2					
I	?	?				
L	?	4	3			
S	0	?	?	?		
T	0	?	?	4	2	
V	?	4	4	?	?	?

Comparison PAM vs. BLOSUM

- PAM is based on an evolutionary model using phylogenetic trees (85% similarity).
- BLOSUM assumes no evolutionary model, but rather conserved “blocks” of proteins.

Πίνακες αντικατάστασης - Ισοδυναμία

- **PAM100** → **Blosum90**
- **PAM120** → **Blosum80**
- **PAM160** → **Blosum62**
- **PAM200** → **Blosum52**
- **PAM250** → **Blosum45**



Nucleic acid PAM Scoring Matrices (1/2)

- PAM: Point accepted mutations.
- Για τον DNA PAM1 πίνακα θεωρούμε ότι το 99% των αλληλουχιών συντηρείται ενώ το 1% μεταλλάσσεται.
- Υποθέτουμε ότι και οι 4 βάσεις, A-C-G-T, μεταλλάσσονται το ένα στο άλλο με την ίδια πιθανότητα.
- Υποθέτουμε ότι και τα 4 βάσεις έχουν την ίδια συχνότητα εμφάνισης.

A. Model of uniform mutation rates among nucleotides

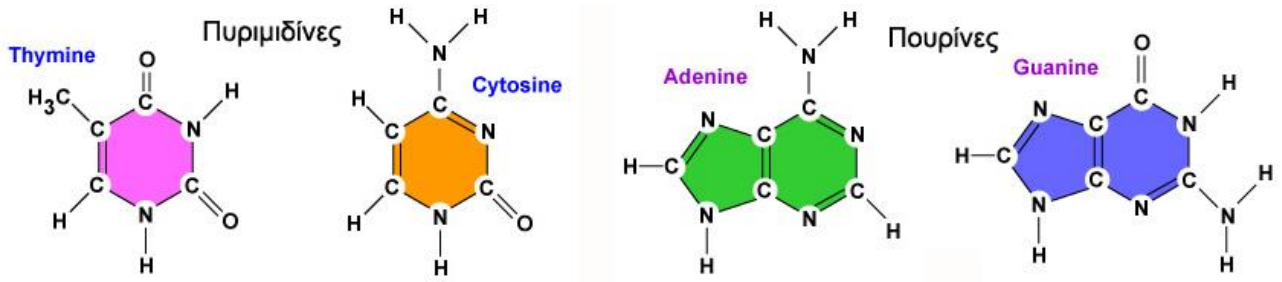
	A	G	T	C
A	0.99			
G	0.00333	0.99		
T	0.00333	0.00333	0.99	
C	0.00333	0.00333	0.00333	0.99

Nucleic acid PAM Scoring Matrices (2/2)

- Θεωρούμε ότι είναι 3 φορές πιο πιθανό οι πυριμιδίνες να μεταλλαχθούν σε πυριμιδίνες (T→C) και οι πουρίνες σε πουρίνες (A →G) σε σχέση με τη μετάλλη των πυριμιδών σε πουρίνες και το αντίστροφο.

B. Model of threefold higher transitions than transversions

	A	G	T	C
A	0.99			
G	0.006	0.99		
T	0.002	0.002	0.99	
C	0.002	0.002	0.006	0.99



4 αζωτούχες βάσεις: Θυμίνη (T), Κυτοσίνη (C), Αδενίνη (A) και Γουανίνη (G)

Οι ποινές για τα κενά (1/2)

- Η εισαγωγή κενών είναι απαραίτητη έτσι ώστε να έχουμε την καλύτερη στοίχιση των αλληλουχιών.
- Συσχετισμένη ή αφινική ποινή (affine gap penalty).
- Πιο πιθανό 1 κενό μήκους K, παρά K κενά μήκους 1.
- Αφαίρεση μιας τιμής για την εισαγωγή ενός κενού και πρόσθετες αφαιρέσεις για επόμενες επεκτάσεις του κενού.
- Το σύστημα βαθμονόμησης θα πρέπει να «τιμωρεί» περισσότερο τα καινούργια κενά.

Παράδειγμα - PAM250

```
C - K H V F C R V C I

C K K C - F C - K C V
```

- **PAM250**
- **Εισαγωγή κενού: -10**

```
Score = 12+5+9+12+12-3-2+4-3x10

Score = 19
```

Οι ποινές για τα κενά (2/2)

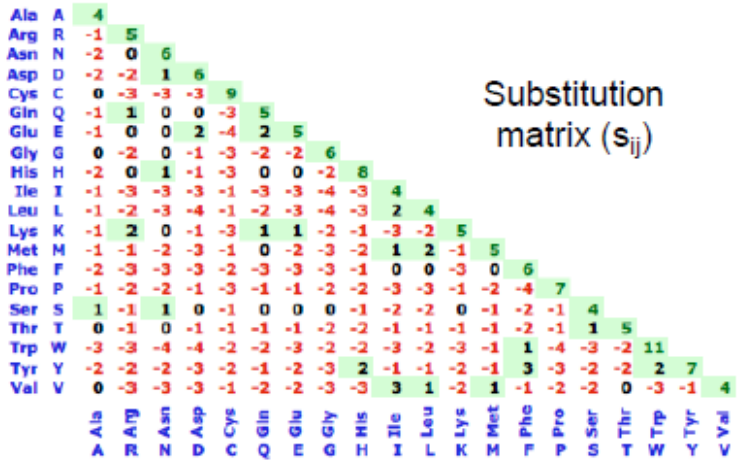
- **p = - A - B·(n-1):**
  - όπου A: ποινή έναρξης του κενού (gap opening penalty).
  - B: ποινή επέκτασης (gap extension penalty).
  - n: μήκος του κενού.
- Συνήθως: A=12 και B=4, οπότε **p = -12 – 4·(n-1)**
- **Πολύ μικρή ποινή: Εισαγωγή πολλών κενών.**
- **Πολύ μεγάλη ποινή: Μη εισαγωγή κενών.**
- **Γενικά:** Η ποινή για ένα κενό n καταλοίπων θα πρέπει να είναι μικρότερη από τη συνολική ποινή για n κενά του 1 καταλοίπου.

Typical score matrix

- DNA:
  - Match = +1.
  - Mismatch = -3.
  - Gap penalty = -5.
  - Gap extension penalty = -2.
- Protein sequences:
  - Blossum62 matrix.
  - Gap open penalty = -11.
  - Gap extension = -1.

Ποινές κενών - Παράδειγμα

```
R  L  A  S  V  -  E  T  D  M  W  T  P  L  T  L  R  Q  H
.  |  .  |  :     :  |  .  :        .  |  .  .  |
T  L  T  S  L  A  Q  T  T  L  -  -  K  A  H  L  G  T  H
-1 +4 +0 +4 +1 -4 +2 +5 -1 +2 -4 -1 -1 -1 -2 +4 -2 -1 +8 = 12
```



Substitution matrix (s_{ij})

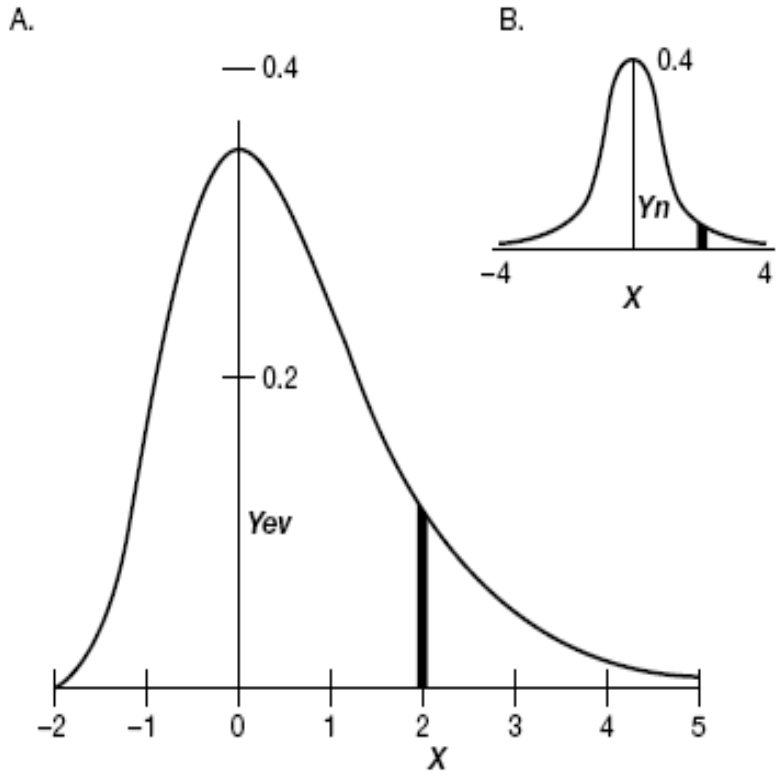
gap penalty (s_{i-})

gap opening	-4
gap extension	-1
end gap	0

Εξετάζοντας τη σημαντικότητα της στοίχισης αλληλουχιών

- Για αλληλουχίες που ανήκουν στην ίδια οικογένεια και είναι προφανές, κάτι τέτοιο δεν είναι απαραίτητο.
- Για αλληλουχίες που δεν είναι ξεκάθαρο ότι είναι παρόμοιες, αυτό είναι αναγκαίο.
- **Συμπέρασμα:** Είναι πράγματι όμοιες ή το αποτέλεσμα της σύγκρισης είναι τυχαίο, δηλαδή το ίδιο με την περίπτωση που οι δύο αλληλουχίες δεν έχουν καμία σχέση.
- Ο έλεγχος θα γίνει σε όλες τις στοιχίσεις έτσι ώστε να επιλεχτεί εκείνη που είναι στατιστικά η σημαντικότερη.

Gumbel extreme value distribution



- **Σκοπός:** Να υπολογιστεί η πιθανότητα το σκορ της στοίχισης μεταξύ δύο τυχαίων αλληλουχιών να φτάσει το σκορ της στοίχισης δύο σχετικών αλληλουχιών. Όσο μικρότερη είναι αυτή η πιθανότητα, τόσο στατιστικά σημαντικό είναι το σκορ στοίχισης των 2 πραγματικών αλληλουχιών.

The distribution of alignment scores between random sequences follows the extreme value distribution (A), not the normal distribution (B)

Βαθμολόγηση στοίχισης αλληλουχιών

- **Score: A number used to assess the biological relevance of a finding.**
- In the context of sequence alignment, a score is a numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity.
- **The score scale depends on the scoring system used (matrix, gap penalty).**

Gap penalty	Alignment	Identity / Similarity	Gaps	Score
0	1 GTC-ATGCTA-GTCGT---GG---GTAGCATTTA-GCT-ATG-TGGG-GT 38 1 -TCGATGCT-GGTCG-CAAGGCAAGTAG---TTATG-TCATGCT---AG- 39	27/50 (54.0%)	23/50	S=135
5	1 GTC-ATGCTAGTCG--TGGGTAGCATTTA-GCT-ATG-TGGGGT 38 1 -TCGATGCTGGTCGCAAGGCAAGTAGTTATG-TCATGCTAG--- 39	26/44 (59.1%)	11/44	S=67
10	1 ----------GTCATGCTAGTCGTGGGTAGC 21 1 TCGATGCTGGTCGCAAGGCAAGTAGTTATGTCATGCTAG---------- 39 22 ATTTAGCTATGTGGGGT 38 39 ---------- 39	10/67 (14.9%)	57/67	S=50

1: 27 matches x 5 = 135. No gap penalty
2: 26 matches x 5 = 130, 7 mismatch x (-4) = -28, 7 gaps x (-5) =-35 TOTAL=130-28-35=67
3: 10 matches x 5 = 50

Converting to Bit-scores

- The bit-score S' is a **normalized score expressed in bits**. Lets you estimate **the magnitude of the search space you would have to look through** before you would expect to find a score as good as or better that this one by chance.

$$S' = \frac{\lambda S - \ln(K)}{\ln 2}$$

where S: raw score. Parameters λ and K depend on the substitution matrix and on the gap penalty

The bit-score is a rescaled version of the raw alignment score that is <u>independent</u> of the size of the search space

Δείκτες στατιστικής σημαντικότητας (1/4)

P-value: Probability that an event occurs by chance

- The P-value associated to a score S is **the <u>probability</u> to obtain by chance a score x at least equal to S**

Low P-value → significant alingment

$$P(S' \geq x) = 1 - \exp\left[-e^{-x}\right] \approx e^{-x}$$

Δείκτες στατιστικής σημαντικότητας (2/4)

- E-value: Correlation of the p-value for multiple testing.
- Expected number of sequences that will produce same or better score by chance.
- The lower the E value, the more significant the score is
- Low E-value → significant alingment.

Αναμενόμενος αριθμός στοιχίσεων με σκορ τουλάχιστον S:

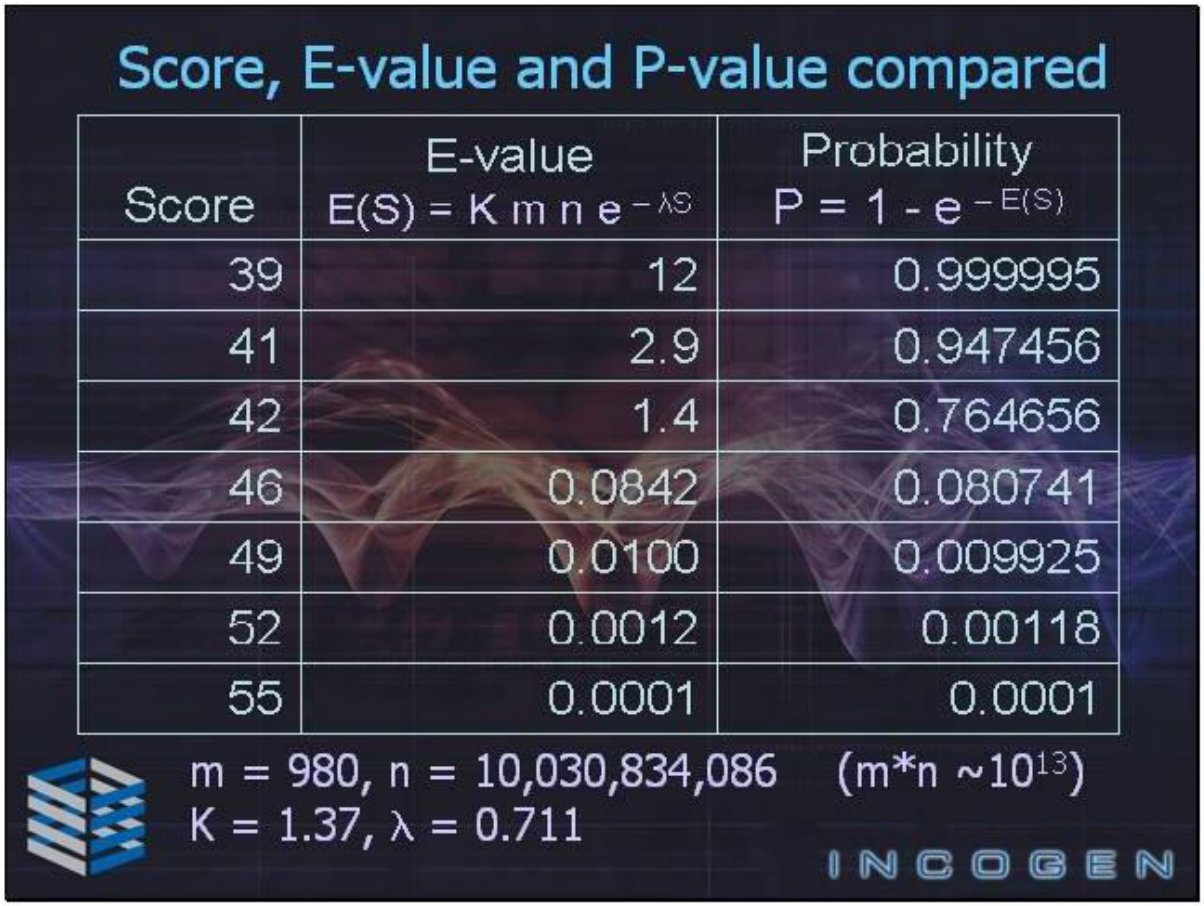
$$E = K \cdot m \cdot n \cdot e^{-\lambda S}$$

- m,n: Μήκος αλληλουχιών
- K, λ: Στατιστικές σταθερές που εξαρτώνται από το σύστημα βαθμονόμησης και τη συχνότητα εμφάνισης των καταλοίπων. π.χ. για τον πίνακα PAM250, K=0.1 και λ=0.229

Δείκτες στατιστικής σημαντικότητας (3/4)

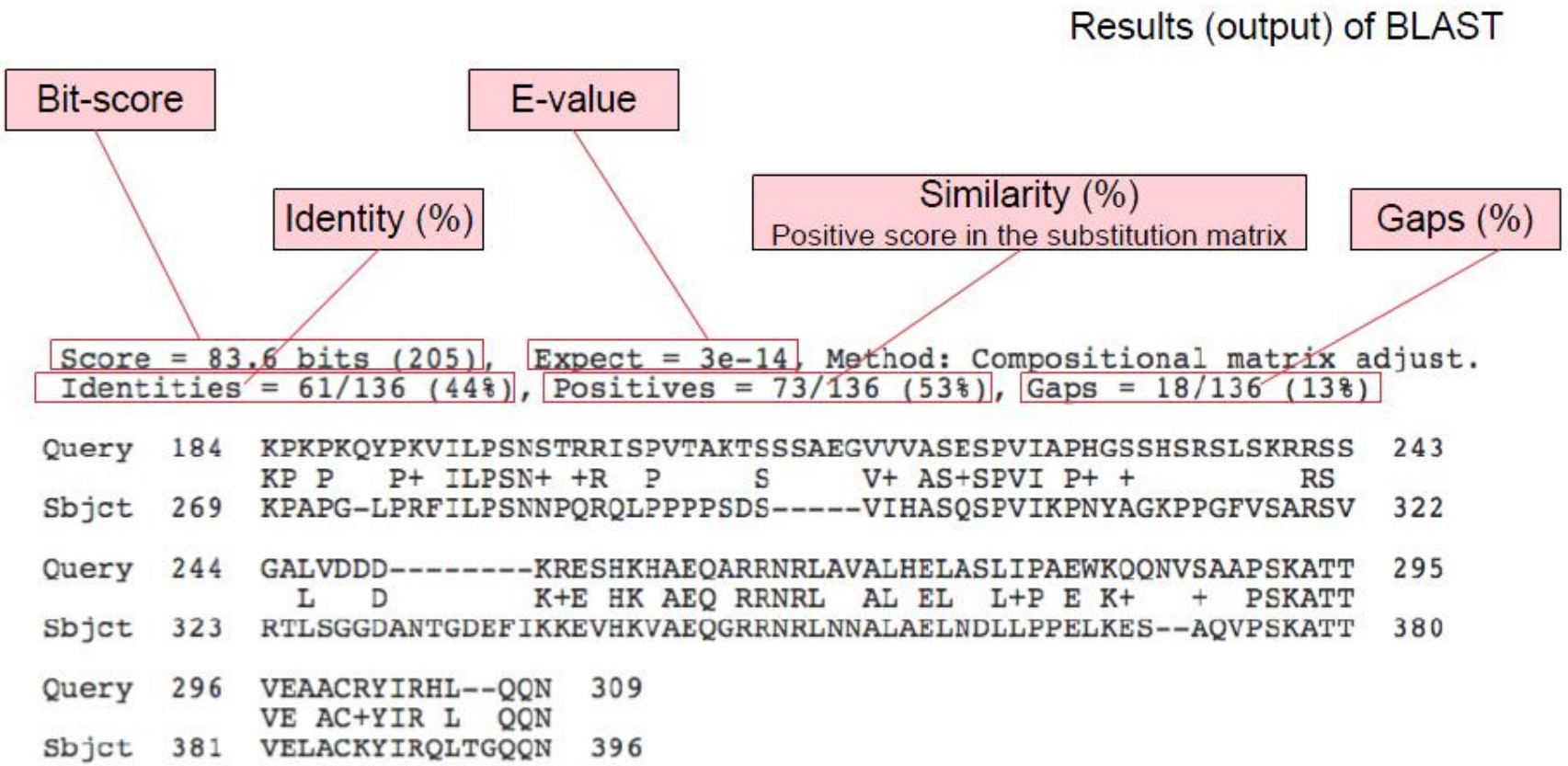
- **Z-score:**
  - Measures how much standard deviations above the mean of the score distribution.

Δείκτες στατιστικής σημαντικότητας (4/4)



http://www.incogen.com

Alignment - Output



Μέθοδοι σύγκρισης αλληλουχιών

1. Οπτικοί.
2. Με αλγόριθμους δυναμικού προγραμματισμού.
3. Με ευρετικούς αλγόριθμους, που βασίζονται στην έννοια των «λέξεων».

- Πρέπει να λαμβάνεται υπόψη ο **λόγος για τον οποίο γίνεται η στοίχιση αλληλουχιών.**
- Ανάλογα επιλέγεται και η μέθοδος σύγκρισης.
- **Ομοιότητα της τάξης το 90%:** Εύκολη κατασκευή της στοίχισης με όλους τους αλγόριθμους.
- **Ομοιότητα της τάξης του 25% (ζώνη του λυκόφωτος, «twilight zone of sequence alignment»):** Οριακή για ασφαλή εξαγωγή συμπερασμάτων.

Διαγράμματα πινάκων σημείων - Dot plots

- Αποτελούν μία από τις απλούστερες μεθόδους οπτικοποίησης της ομοιότητας μεταξύ δύο αλληλουχιών.
- Μέθοδος:
  - Σε ένα δισδιάστατο ορθογώνιο διάγραμμα ο κάθε άξονας αντιστοιχεί και σε μία αλληλουχία.
    - **Αλληλουχία 1: A C A G C G C G A G C**
    - **Αλληλουχία 2: A C A C G A G G**
  - Στο πλέγμα που προκύπτει:

Dot plots - Μέθοδος (1/3)

Σε ένα δισδιάστατο ορθογώνιο διάγραμμα ο κάθε άξονας αντιστοιχεί και σε μία αλληλουχία

	A	C	A	G	C	G	C	G	A	G	C
A											
C											
A											
C											
G											
A											
G											
G											

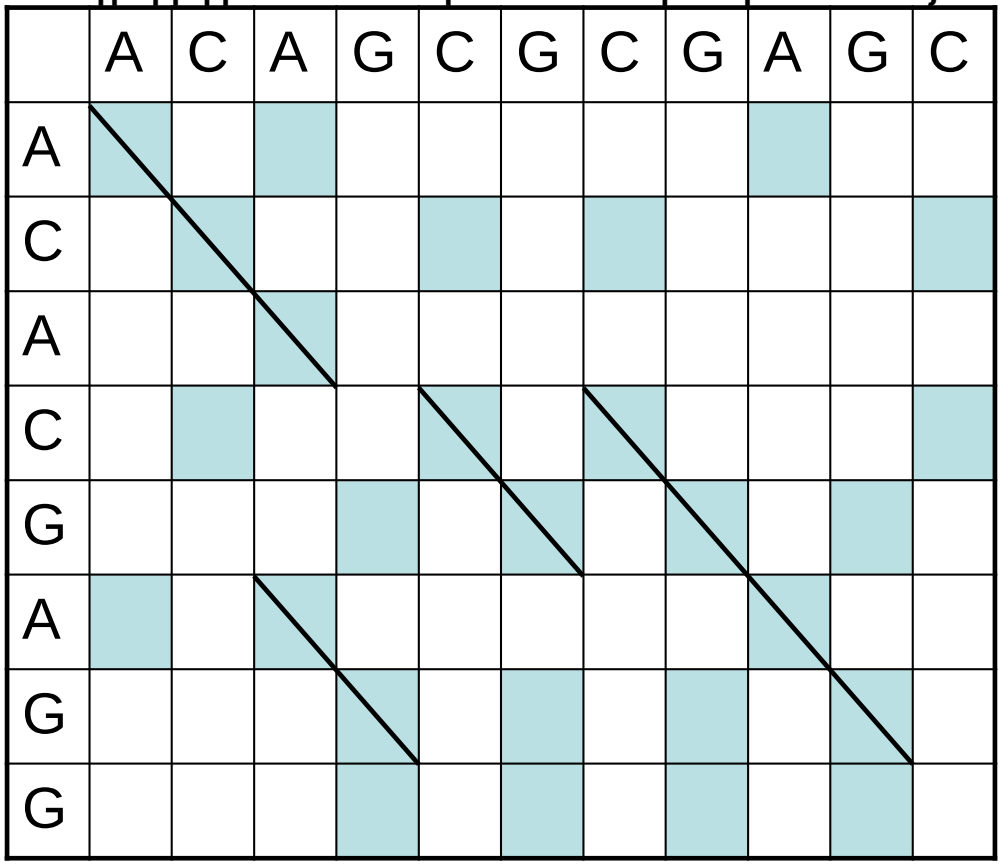
Dot plots - Μέθοδος (2/3)

Τα τετράγωνα που αντιστοιχούν σε ταυτόσημα κατάλοιπα σημειώνονται (χρώμα)

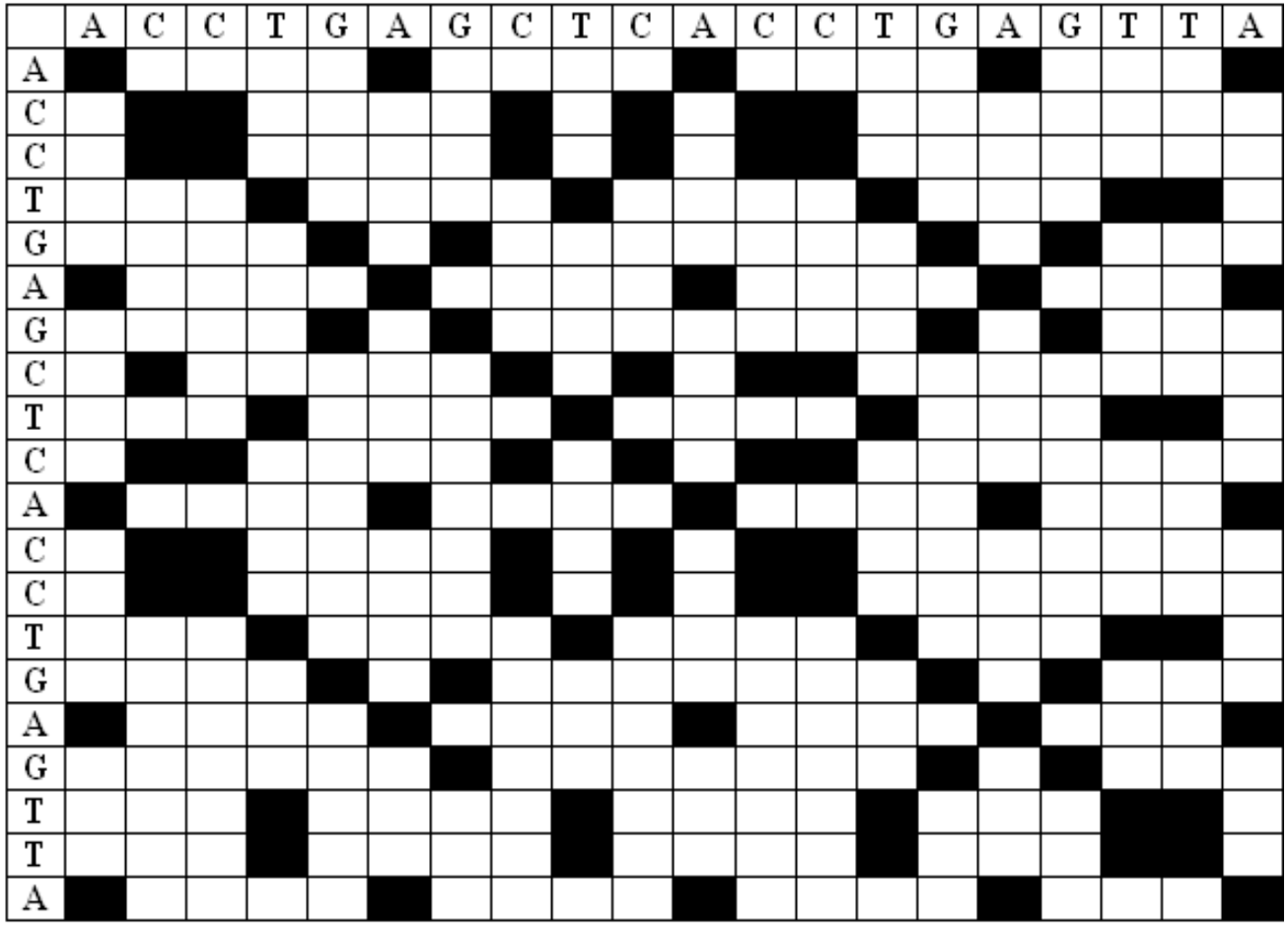
	A	C	A	G	C	G	C	G	A	G	C
A	■		■						■		
C		■			■		■				■
A			■								
C		■			■		■				■
G				■		■		■		■	
A	■		■						■		
G				■		■		■		■	
G				■		■		■		■	

Dot plots - Μέθοδος (3/3)

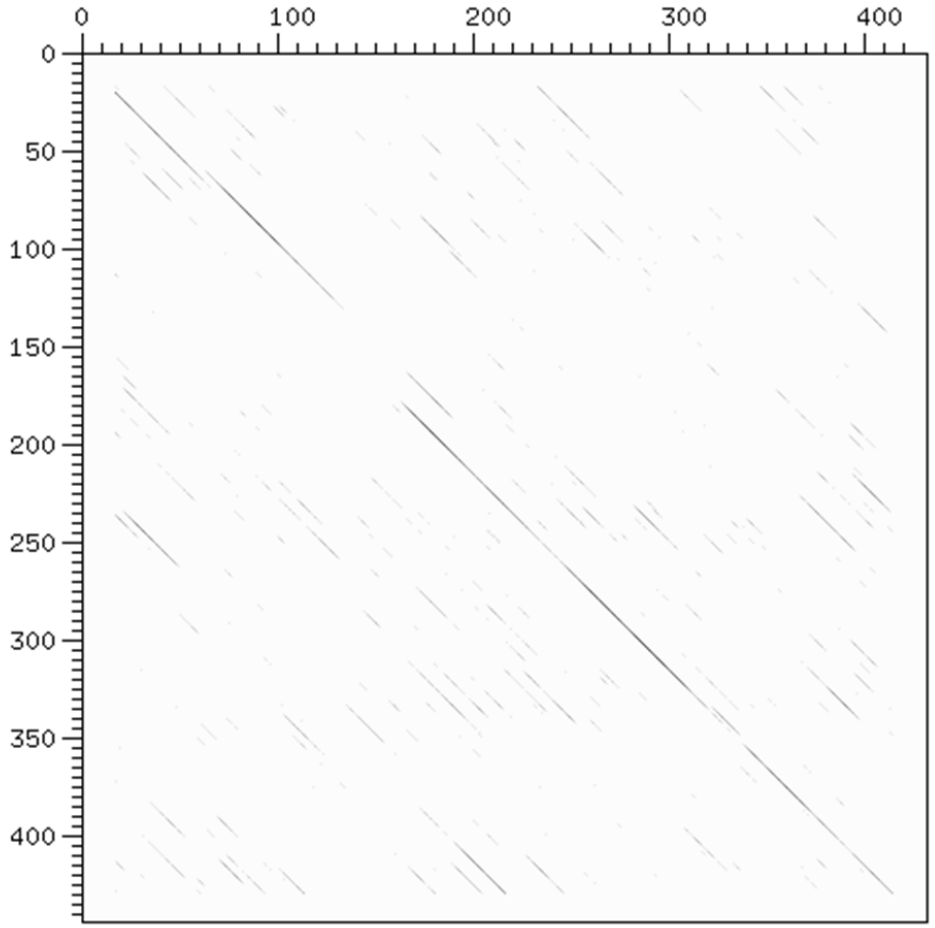
Όταν οι αλληλουχίες είναι όμοιες κατά μήκος μιας περιοχής, σχηματίζεται μία διαγώνιο γραμμή με κατεύθυνση από πάνω αριστερά κάτω δεξιά



Dot plots - Παράδειγμα (1/2)



Dot plots - Παράδειγμα (2/2)

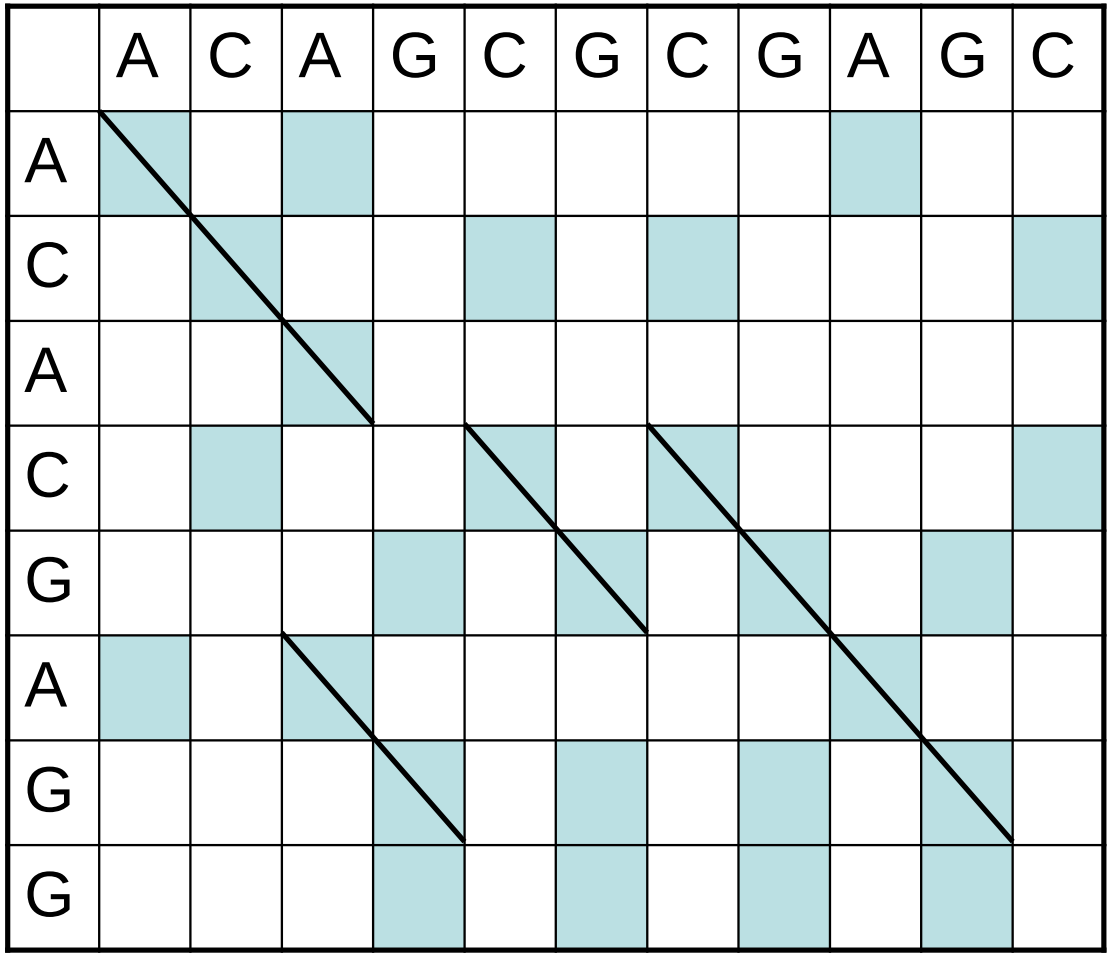


http://myhits.isb-sib.ch/cgi-bin/dotlet
http://www.vivo.colostate.edu/molkit/dnadot/

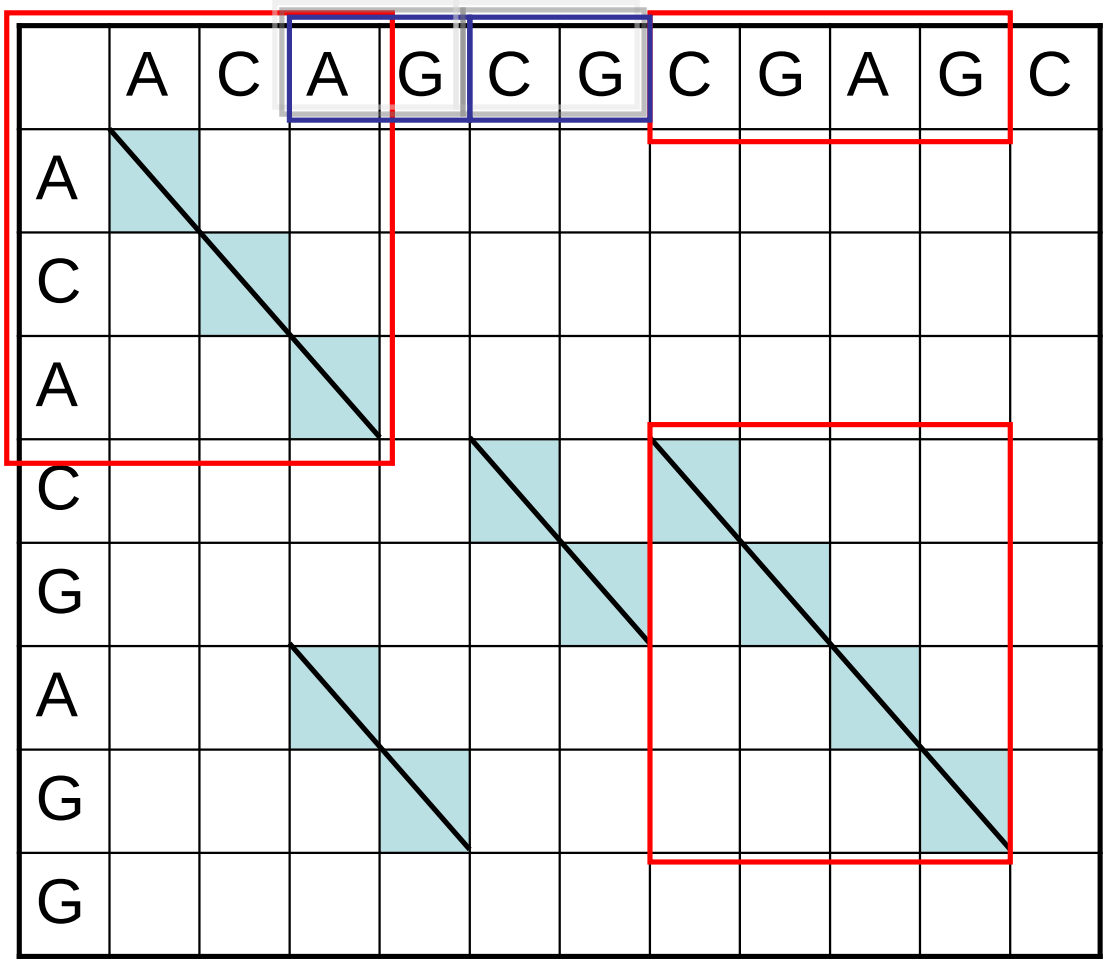
«Θόρυβος» στα dot plots

- 1 από τις 4 βάσεις έχει την πιθανότητα να ταιριάζει τυχαία.
- **Μέθοδος φιλτραρίσματος:** μία περιοχή σκιάζεται αν υπερβαίνει κάποιον συγκεκριμένο αριθμό (φίλτρο) όμοιων καταλοίπων.

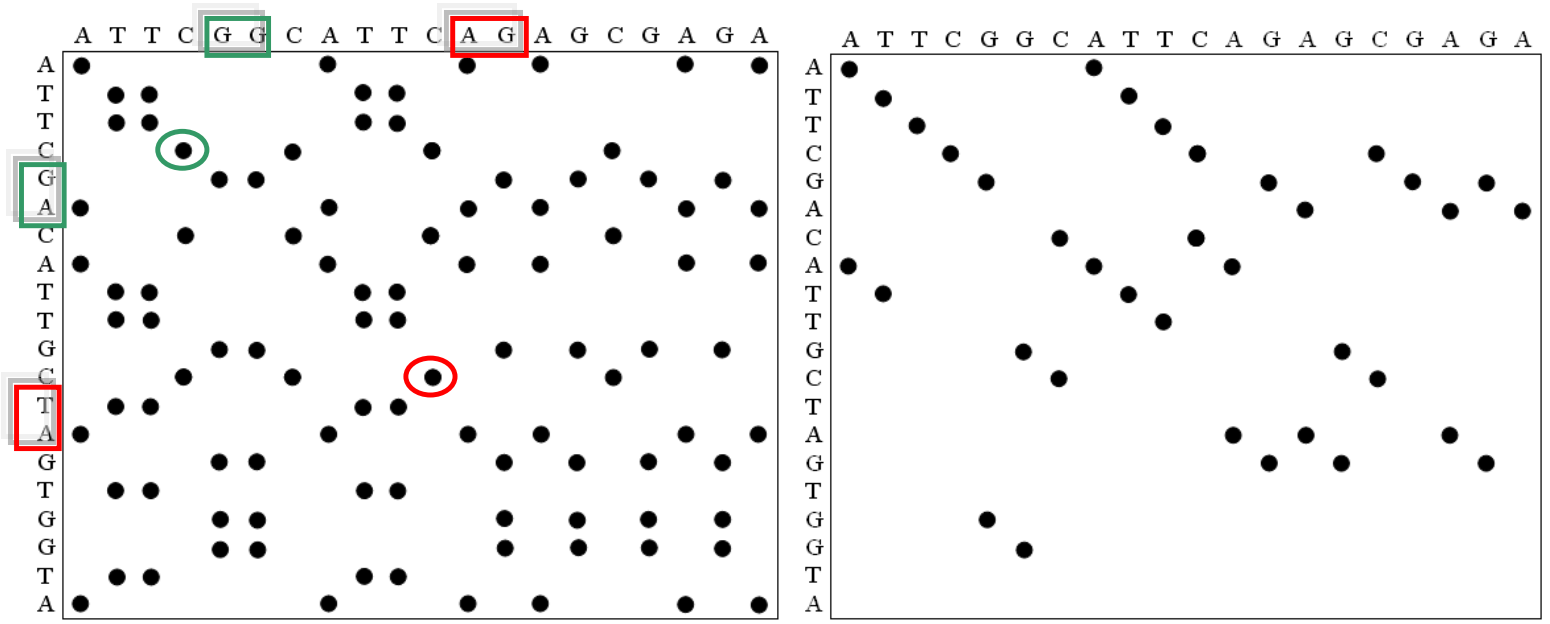
Μείωση θορύβου στα dot plots (1/3)



Μείωση θορύβου στα dot plots (2/3)



Μείωση θορύβου στα dot plots (3/3)

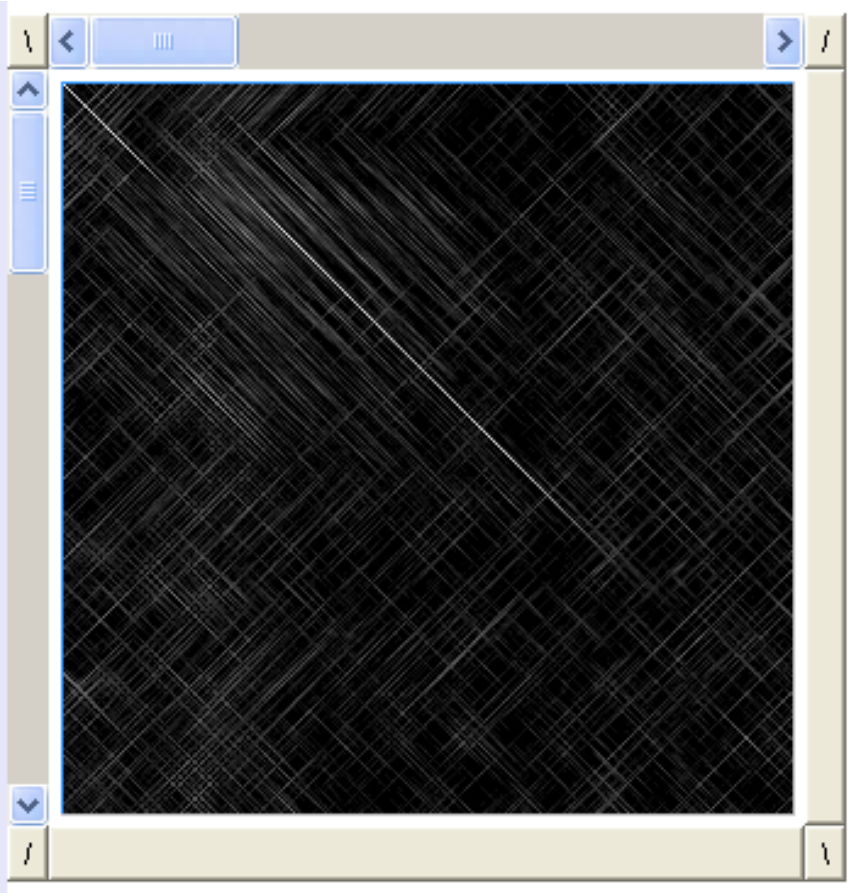


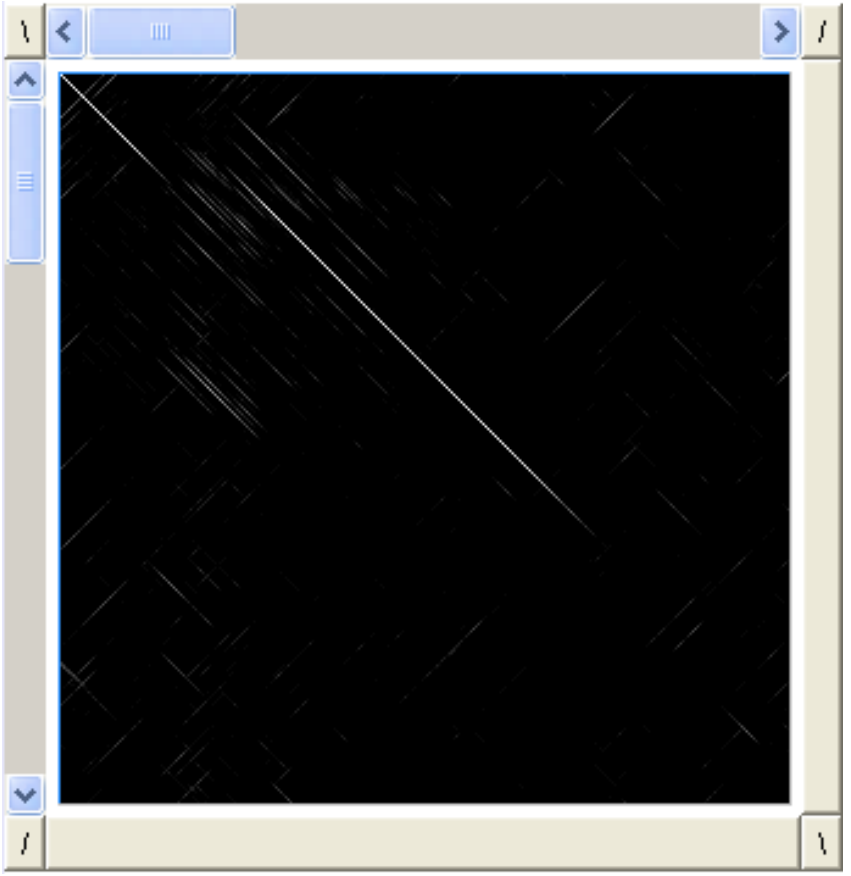
Window: 2, Stringency: 1

Window size: Number of nucleotides compared each time

Stringency: The minimum number of nucleotides in the window must be match, so the dot can be placed

Μείωση θορύβου στα dot plots (Dotlet)

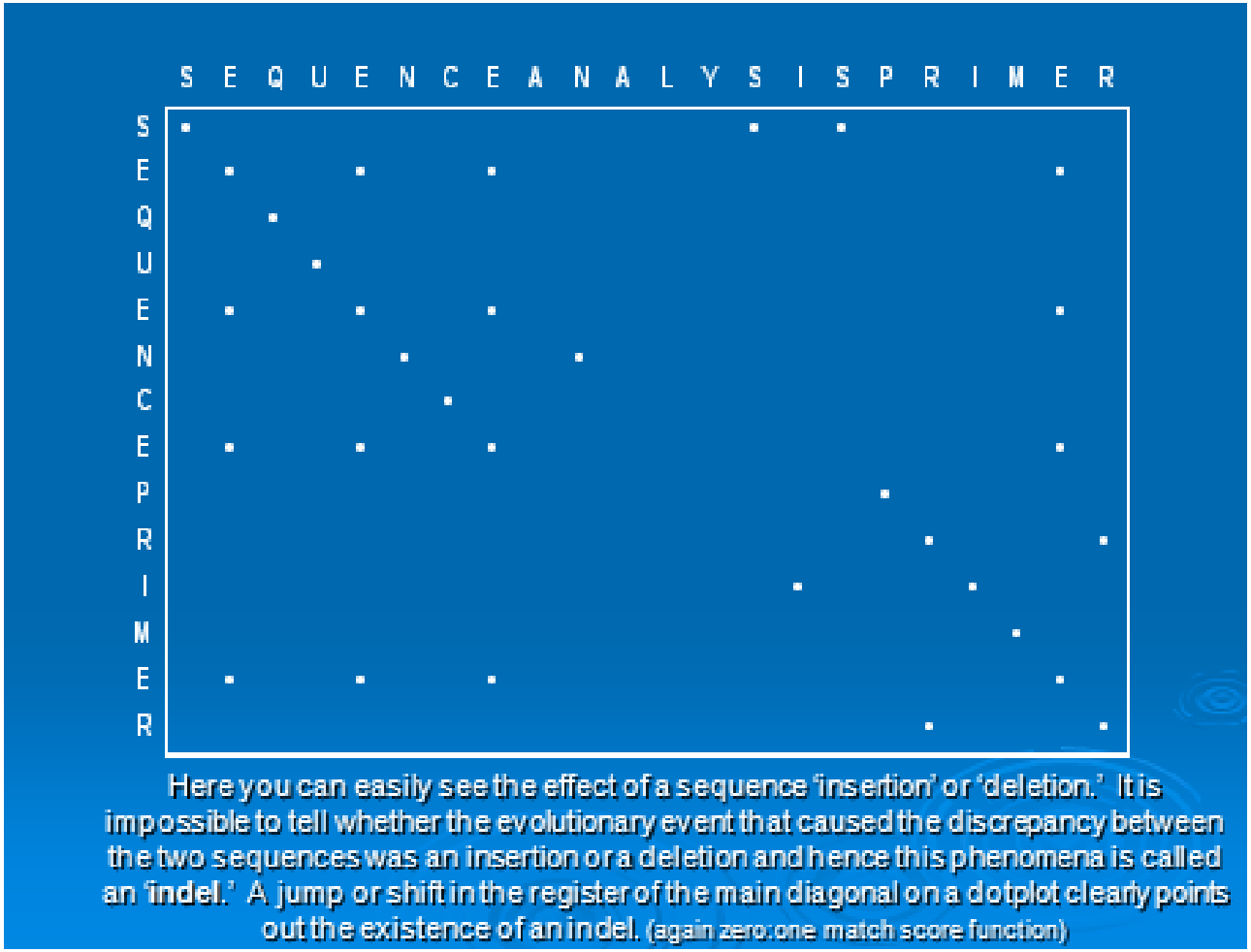




Πληροφορίες από τα dot plots

1. **Περιοχές ομοιότητας:**
   - Διαγώνιος.
2. **Προσθήκες (insertions) / εξαλείψεις (deletions).**
3. **Επαναλαμβανόμενες περιοχές:**
   - Ορθές.
   - Αντιστρέψιμες (αντίστροφη φορά).

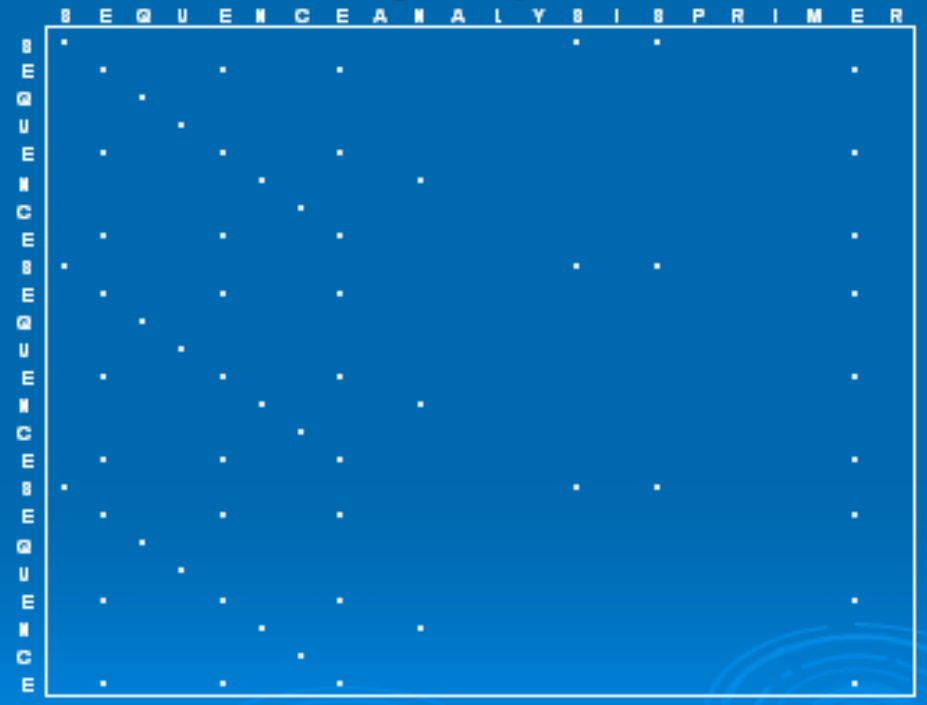
Check out the «mutated» inter-sequence comparison below:



Here you can easily see the effect of a sequence 'insertion' or 'deletion.' It is impossible to tell whether the evolutionary event that caused the discrepancy between the two sequences was an insertion or a deletion and hence this phenomena is called an '**indel**.' A jump or shift in the register of the main diagonal on a dotplot clearly points out the existence of an indel. (again zero:one match score function)

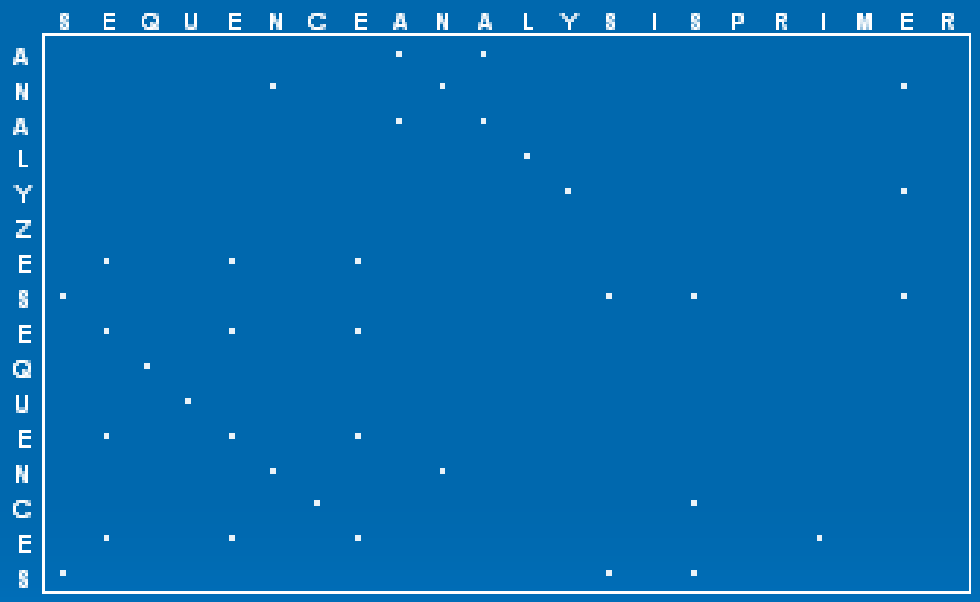
Easy to visualize with dot matrix

Another phenomenon that is very easy to visualize with dot matrix analysis are duplications or direct repeats. These are shown in the following example:



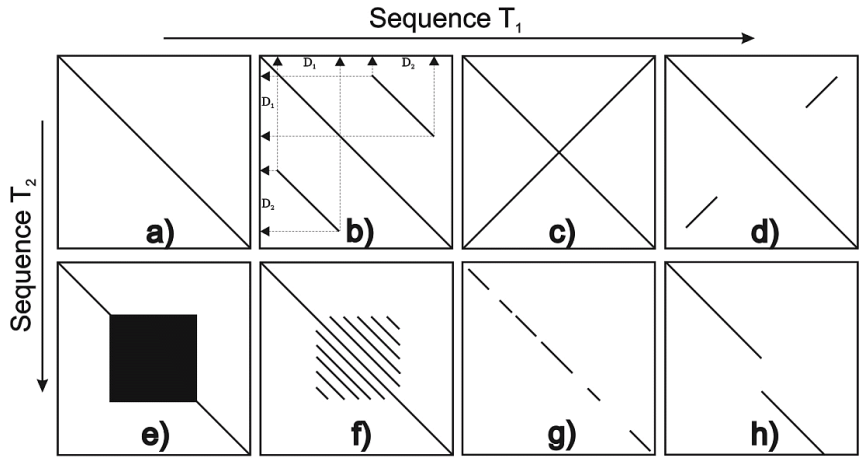
The 'duplication' here is seen as a distinct column of diagonals; whenever you see either a row or column of diagonals in a dotplot, you are looking at direct repeats.

Now consider the more complicated «mutation» in the following comparison:



Again, notice the diagonals. However, they have now been displaced off of the center diagonal of the plot and, in fact, in this example, show the occurrence of a 'transposition.' Dot matrix analysis is one of the only sensible ways to locate such transpositions in sequences. Inverted repeats still show up as perpendicular lines to the diagonals, they are just now not on the center of the plot. The 'deletion' of 'PRIMER' is shown by the lack of a corresponding diagonal.

Dot plots - Παράδειγμα



a) **Απόλυτη ομοιότητα.**
b) **Επαναλαμβανόμενες περιοχές.**
c) **Ολική παλινδρόμηση.**
d) **Μερική παλινδρόμηση.**
e) **Επαναλαμβανόμενο σύμβολο (αμινοξύ ή νουκλεοτίδιο) και στις δύο αλληλουχίες.**
f) **Επαναλαμβανόμενες περιοχές και στις δύο αλληλουχίες.**
g) **Διακοπές - Κοινός πρόγονος.**
h) **Προσθήκη στην αλληλουχία 1 ή εξάλειψη στην αλληλουχία 2.**

Size of the search space (1/2)

- Size of the search space ~ query sequence length (n) · the sum of the lengths of the sequences in the database (m), **N=n · m.**
- **Size of the search space = N · K**, where K: Altschul coefficient.
- **Example:** Calculate the search space for a protein of 235a.a. length against a protein database of size m=12496420a.a. The value of K is given equal to 0.13.
- Size of the search space = 0.13*235*12,496,420=0.38 billion.
- In this case, a bit score of 30 (which correspond to a space of 2^{30} = 1 billion) may have occurred by chance alone.

Converting to Bit - scores

$$S' = \frac{\lambda \cdot S - \ln K}{\ln 2}$$

$$\Rightarrow S' \cdot \ln 2 = \lambda \cdot S - \ln K$$

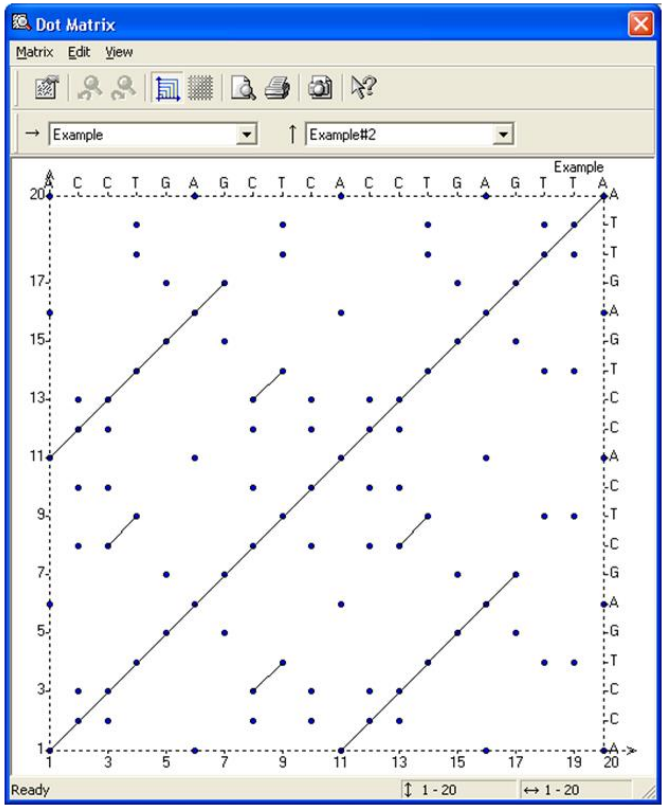
$$\Rightarrow \ln K = \lambda \cdot S - S' \ln 2$$

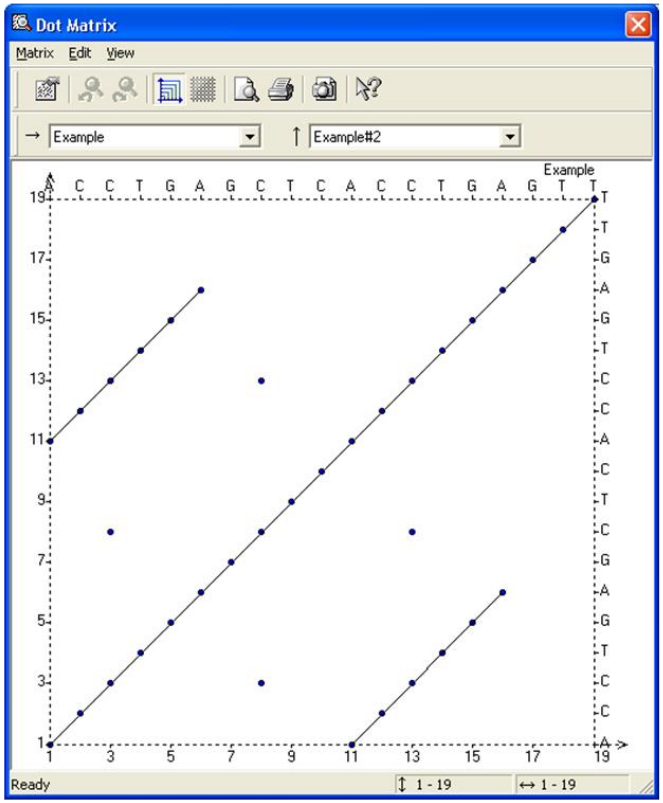
$$\Rightarrow K = e^{\lambda S - S' \ln 2}$$

$$E = K \cdot m \cdot n \cdot e^{-\lambda S}$$

$$E = \frac{m \cdot n \cdot e^{-\lambda S} \cdot e^{\lambda S}}{e^{S' \cdot \ln 2}} = m \cdot n \cdot 2^{-S'}$$

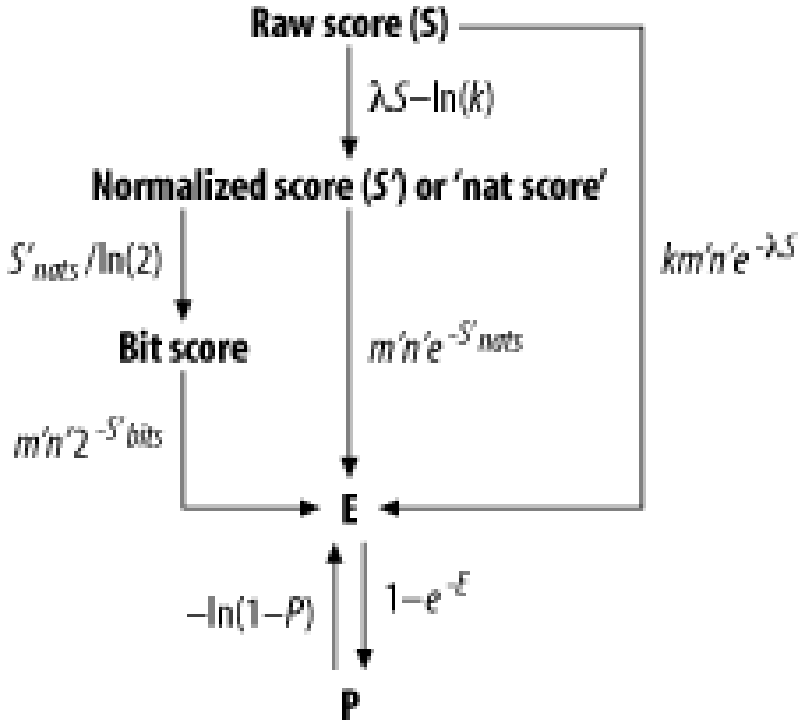
Μείωση θορύβου στα Dot Plots (1/2)



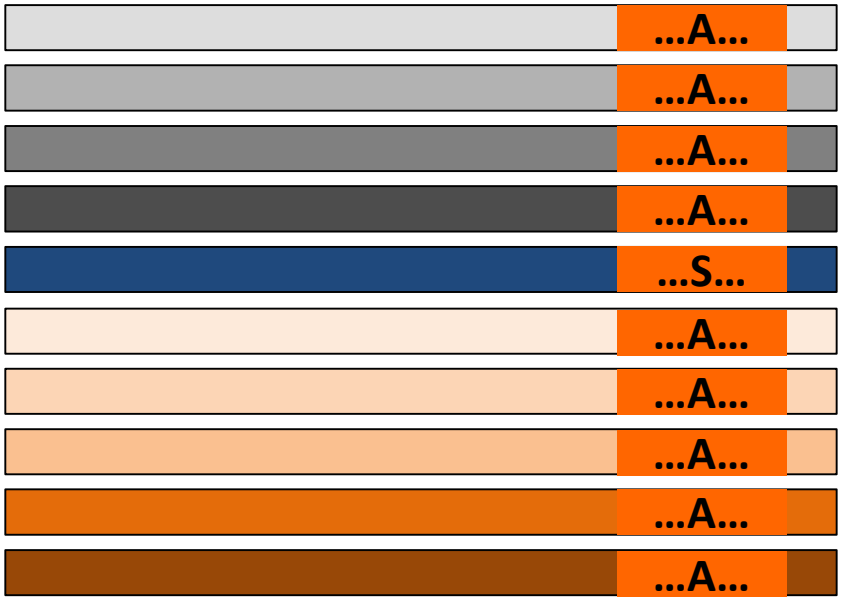


Self alignment of **ACCTGAG**CTC**ACCTGAG**TTA

Μείωση θορύβου στα Dot Plots (2/2)



Ερμηνεία τιμής του πίνακα BLOSUM - Παράδειγμα



- AA pairs: 8+7+…+2+1=36 possible AA pairs, $F_{AA.}$
- AS pairs: 9 possible AS pairs, $f_{AS.}$
- Η συχνότητα εμφάνισης του ζεύγους AA: **$q_{AA}=f_{AA} / f_{AA}+f_{AS} = 36/(36+9)=0.8.$**
- Η συχνότητα εμφάνισης του ζεύγους AS: **$q_{AS}=f_{AS} / f_{AA}+f_{AS} = 9/(36+9)=0.2.$**
- Η αναμενόμενη συχνότητα το A να βρίσκεται σε ζεύγος: **$p_A=q_{AA}+q_{AS}/2=0.8+0.2/2=0.9.$**
- Η αναμενόμενη συχνότητα το S να βρίσκεται σε ζεύγος: **$p_S=q_{AS}/2=0.2/2=0.1.$**
- Η αναμενόμενη συχνότητα εμφάνισης του ζεύγους AA: $e_{AA}=$ **$p_A x\ p_A = 0.9x0.9=0.81.$**
- Η αναμενόμενη συχνότητα εμφάνισης του ζεύγους AS: $e_{AS}=$ **$2xp_A x\ p_S=0.18.$**
- **Η τιμή στον πίνακα για το ζεύγος AA: $e_{AA}/q_{AA}=0.8/0.81=0.988.$**
- **Η τιμή στον πίνακα για το ζεύγος AS: $e_{AS}/q_{AS}=0.18/0.2=1.111.$**
- Οι τιμές μετατρέπονται σε λογαρίθμους με βάση το 2 και πολλαπλασιάζονται με το 2.

Size of the search space (2/2)

- **Bit-score S':** A log-scaled version of a score, **$\log_2$(score)=S'**
- Sequence: m

Database of protein sequences: Total length n.

Size of the search space ~ m*n.

Size of the search space =K·m*n, K:

- **Example:**

Protein database with sequences of total length: 12496420 a.a.

Sequence: 235 a.a. length

For protein database, K=0.13

Size of the search space:

Significance of a local alignment

- Altschul and Gish (1996) have provided estimations of **K=0.09 and λ=0,229** for PAM250 matrix, for a typical amino acid distribution and for an alignment score based on using a very high gap penalty.

$$S' = \lambda S - \ln(Kmn) = 0.229 \cdot 73 - \ln(0.09 \cdot 250 \cdot 250)$$

$$S' = 16.72 - 8.63 = 8.09 bits$$

$$P(S' \geq 8.55) = e^{-8.09} = 3.2 \cdot 10^{-4}$$

- ***Note that the calculated S' of 8.09 bits in the previous step is approximately the same as the 9 bits calculated by the simpler method***

Relative Entropy

$$H = \sum_{i \geq j} q_{ij} s_{ij}$$

- Indicates power of scoring scheme to distinguish from "background noise" (i.e., randomness).
- **Can use *H* to compare different scoring matrices.**
- Relative entropy of a random alignment should be negative.
- **Scores can be related to biology:**
  - **negative=dissimilarity,**
  - **zero=indifference, positive=similar.**

Στατιστική σημαντικότητα στοίχισης αλληλουχιών χωρίς κενά

- The raw score of an alignment is determined as the sum of the log-odds scores and the gap penalties and is given in the units of log to the base 2 ($\log_2$) or bits.
- **Bit-score S':** A log-scaled version of a score, **$\log_2$(score)=S'** .
- Αν μετατρέψουμε τις τιμές σε $\log_2 x$ έχουμε bits πληροφορίας.
- Όταν χρησιμοποιούμε bit score system, το **κατώφλι της στατιστικής σημαντικότητας** δίνεται από τη σχέση: **$\log_2(m \cdot n)$.**

Στατιστική σημαντικότητα στοίχισης αλληλουχιών χωρίς κενά - Παράδειγμα (1/2)

- 2 sequences, each 250 a.a. long
- Significance cut off: $\log_2(250*250)=16$ bit
- F W L E V E G N S M T A P T G
- F W L D V Q G D S M T A P A G

Using **PAM250** the score is calculated:

- S=9+17+6+3+4+2+5+2+2+6+3+2+6+1+5=73
- **S=73**

Στατιστική σημαντικότητα στοίχισης αλληλουχιών χωρίς κενά - Παράδειγμα (2/2)

- S is in $10*\log_{10}x$.
- Convert S to a bit score, i.e. calculate $\log_2 x$.

$$S = 10\log_{10} x \Rightarrow \log_{10} x = \frac{S}{10}$$

$$\Rightarrow \frac{\log_2 x}{\log_2 10} = \frac{S}{10}$$

$$\Rightarrow \log_2 x = \frac{S}{10}\log_2 10$$

$$\Rightarrow \log_2 x \approx \frac{1}{3}S$$

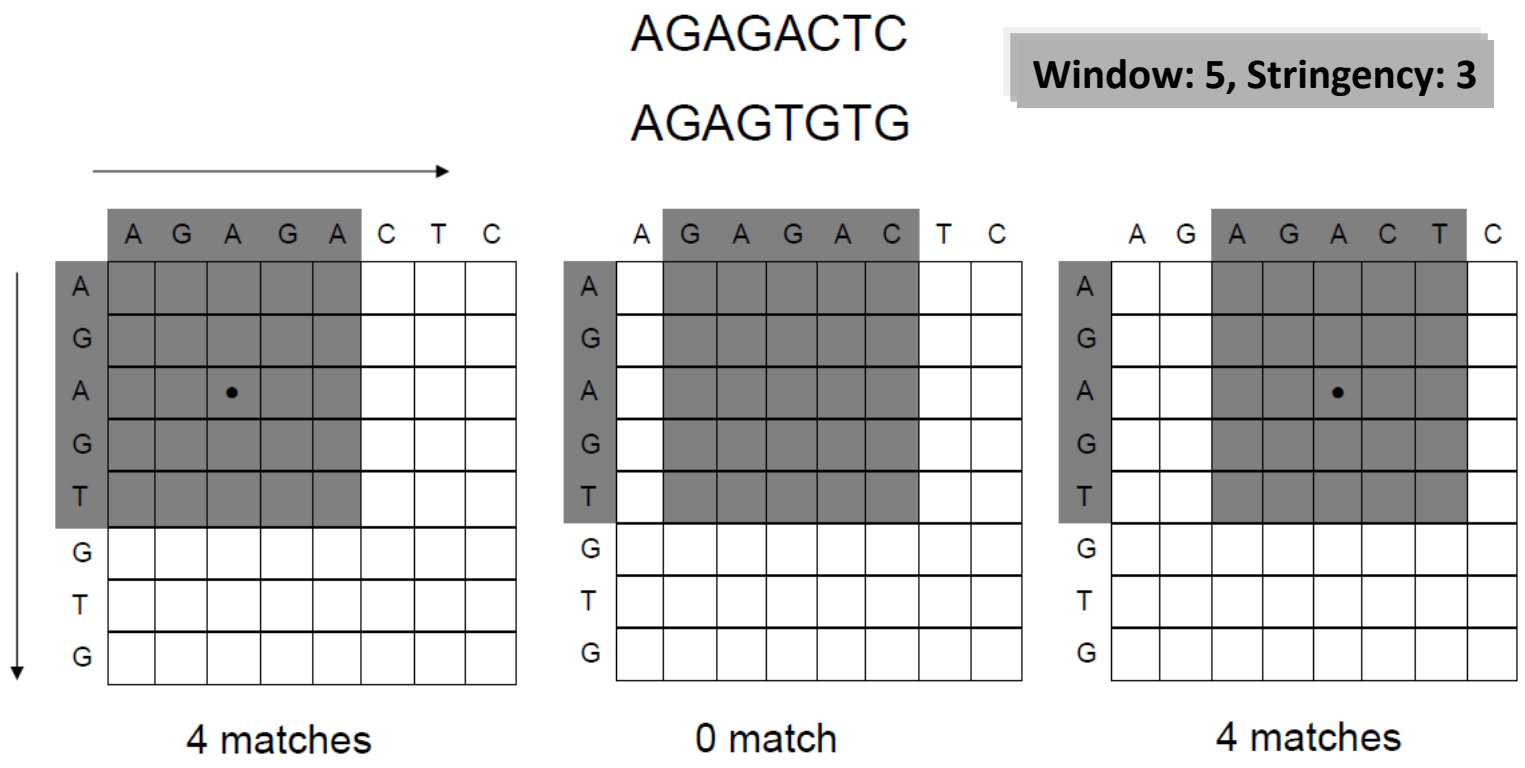
$$\Rightarrow S' = \frac{1}{3}S$$

Στατιστική σημαντικότητα στοίχισης αλληλουχιών χωρίς κενά - Παράδειγμα

$$S' = \frac{1}{3} S = \frac{1}{3} 73 = 24.333 bits$$

- Significance cut-offs ($\log_2(m \cdot n)$) = 16 bits.
- **Το S' είναι 9 bits μεγαλύτερο** από το κατώφλι της στατιστικής σημαντικότητας, επομένως η στοίχιση στο συγκεκριμένο σημείο των αλληλουχιών είναι **στατιστικά σημαντική.**

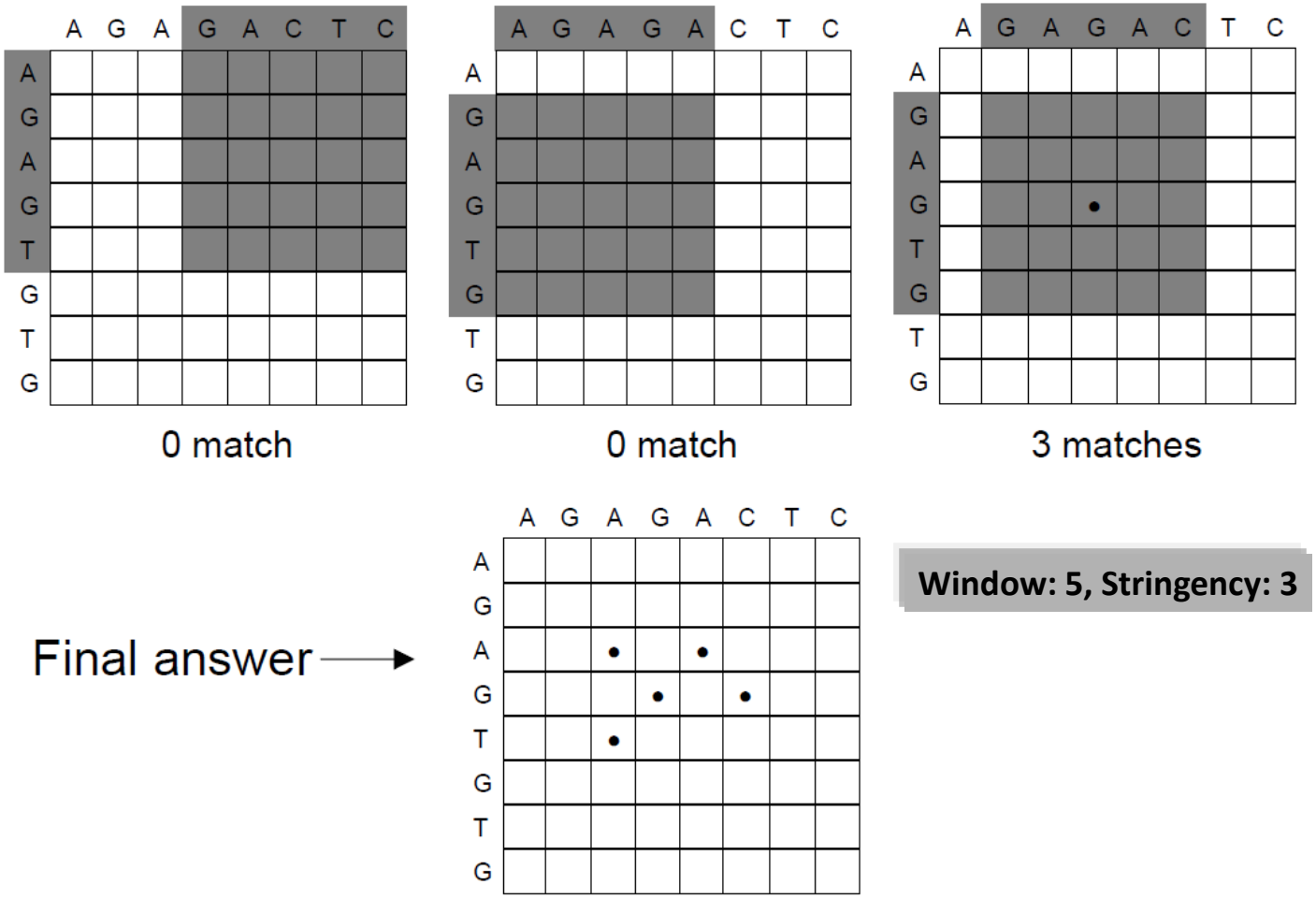
Μείωση θορύβου στα dot plots (1/2)



Window size: Number of nucleotides compared each time

Stringency: The minimum number of nucleotides in the window must be match, so the dot can be placed

Μείωση θορύβου στα dot plots (2/2)



Τέλος Ενότητας

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημείωμα Αναφοράς

- Copyright Πανεπιστήμιο Δυτικής Μακεδονίας, Τμήμα Μηχανικών Πληροφορικής και Τηλεπικοινωνιών, Αγγελίδης Παντελής. «**Βιοπληροφορική**». Έκδοση: 1.0. Κοζάνη 2015. Διαθέσιμο από τη δικτυακή διεύθυνση:

  https: //eclass.uowm.gr/courses/ICTE102/

Σημείωμα Αδειοδότησης

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.

Σημείωμα Χρήσης Έργων Τρίτων

Το Έργο αυτό κάνει χρήση των ακόλουθων έργων:

Εικόνες:

- http: //blog.com.mk/send/121903
- http: //foter.com/Cmyk/
- http: //www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/what-are-protein-signatures/signature-types/what-ar-0
- http: //contentinacottage.blogspot.ca/2012_01_29_archive.html
- https: //www.cartoonstock.com/