ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ AGRICULTURAL UNIVERSITY OF ATHENS

Βιοπληροφορική

Ενότητα 6:

Στοίχιση ακολουθιών ανά ζεύγη – Σύστημα βαθμολόγησης, 2 ΔΩ

Τμήμα: Βιοτεχνολογίας

Όνομα καθηγητή: **Τ. Θηραίου**

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Μαθησιακοί Στόχοι

- Κατανόηση της σημασίας του συστήματος βαθμολόγησης της στοίχισης.
- Παρουσίαση των πινάκων αντικατάστασης PAM και BLOSUM.
- Επεξήγηση των μοντέλων βαθμολόγησης των ποινών.

Λέξεις Κλειδιά

- Λέξεις κλειδιά: Σύστημα βαθμολόγησης της στοίχισης ακολουθιών ανά ζεύγη, Πίνακας αντικατάστασης, Ποινές για τα κενά.

- Key words: Pairwise sequence alignment scoring scheme, Substitution matrix, PAM / BLOSUM, Gap penalty.

Σύστημα βαθμολόγησης 1/10

- Σ' ένα απλό σύστημα βαθμολόγησης μπορούμε να ορίσουμε σταθερές τιμές για τη στοίχιση:
  - όμοιων καταλοίπων (**match score**).
  - ανόμοιων καταλοίπων (**mismatch score**).
  - καταλοίπου με κενό (**gap penalty**).
- Το τελικό **score ομοιότητας** της στοίχισης ισούται με το **άθροισμα** των επιμέρους scores.

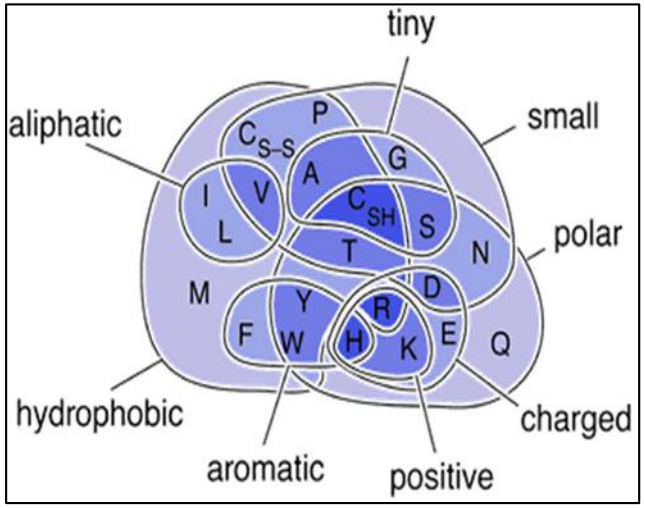
Σύστημα Βαθμολόγησης 2/10

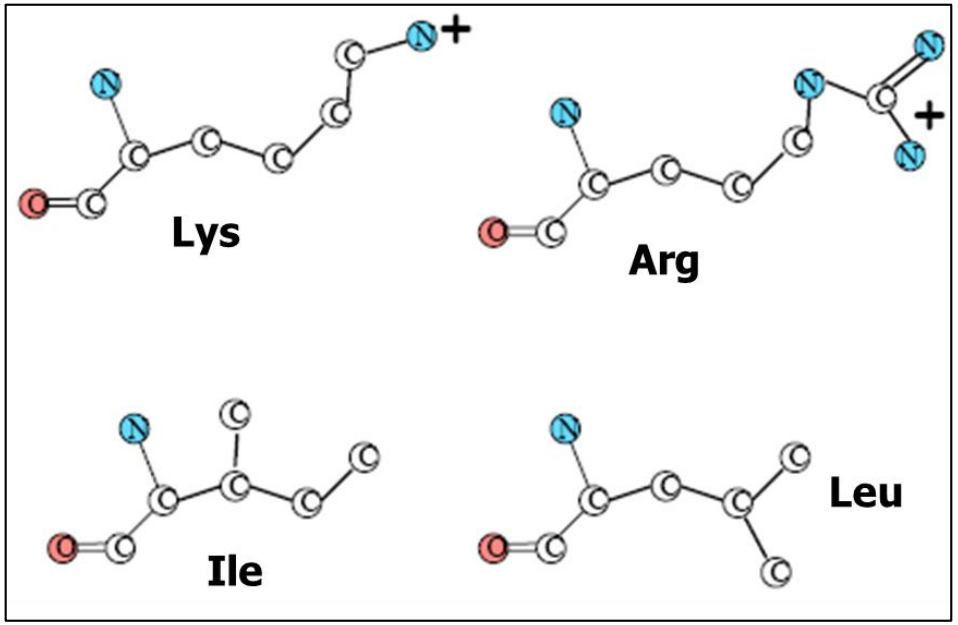
- match score = 1, mismatch score = -1, gap penalty = -2.
  - 18 στοιχίσεις όμοιων καταλοίπων.
  - 32 στοιχίσεις ανόμοιων καταλοίπων.
  - 4 στοιχίσεις καταλοίπων με κενά.
- total score = 18×1 + 32×(-1) + 4×(-2) = -22

```
  X         220         230         240         250                X
  F--SGGNTHIYMNHVEQCKEILRREPKELCELVISGLPYKFRYLSTKE-QLK-Y
  |     |   |||  ||  |   |   | |||       | |       |              |    |
GDFIHTLGDAHIYLNHIEPLKIQLQREPRPFPKLRILRKVEKIDDFKAEDFQIEGYN
  X          260         270         280         290                X
```

Σύστημα Βαθμολόγησης 3/10

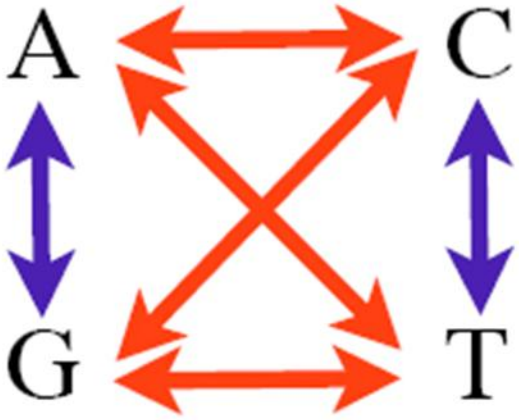
- **συντηρητικές αντικαταστάσεις** (conservative substitutions)
  - Η αντικατάσταση ενός αμινοξέος από κάποιο άλλο με **παρόμοιες** φυσικοχημικές ιδιότητες, είναι **λιγότερο πιθανό** να αλλοιώσει τη δομή και τη λειτουργία μιας πρωτεΐνης.





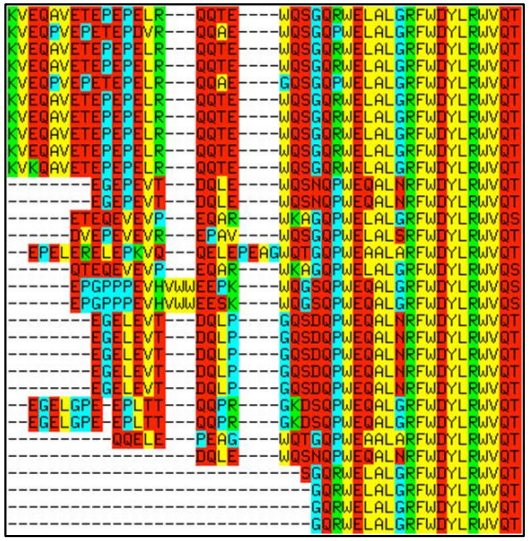
Σύστημα Βαθμολόγησης 4/10

- πουρίνες: αδενίνη (A) και γουανίνη (G).
- πυριμιδίνες: κυτοσίνη (C) και θυμίνη (T).
- **μεταπτώσεις (transitions)**: αλλαγές από πουρίνη σε πουρίνη ή από πυριμιδίνη σε πυριμιδίνη.
- **μεταστροφές (transversions)**: αλλαγές από πουρίνη σε πυριμιδίνη ή αντίστροφα.
- Οι μεταπτώσεις είναι πιθανότερο να συμβούν σε σχέση με τις μεταστροφές.



Σύστημα Βαθμολόγησης 5/10

- **Πίνακες αντικατάστασης (Substitution matrices)**:
  - αποδίδουν **διαφορετικά scores ομοιότητας** για την αντιστοίχιση διαφορετικών αμινοξέων.
  - εξάγονται από πολλαπλές στοιχίσεις.
  - score ~ $\log_2$ (observed/expected).
  - observed.
    - παρατηρηθείσα συχνότητα αντικατάστασης.
  - expected
    - αναμενόμενη συχνότητα αντικατάστασης.



Πίνακες αντικατάστασης 1/8

- **PAM (Point Accepted Mutation)** (Dayhoff):
  - 1PAM: μονάδα εξελικτικής απόκλισης που αντιστοιχεί σε αλλαγή 1% αμινοξέων.
  - 250PAM: απόκλιση ~ 80%.
  - Ένα αμινοξύ μπορεί να αλλάξει περισσότερες από μία φορές ή / και να επιστρέψει στον αρχικό του τύπο.

    MA**TC**G

    MA**G**CG

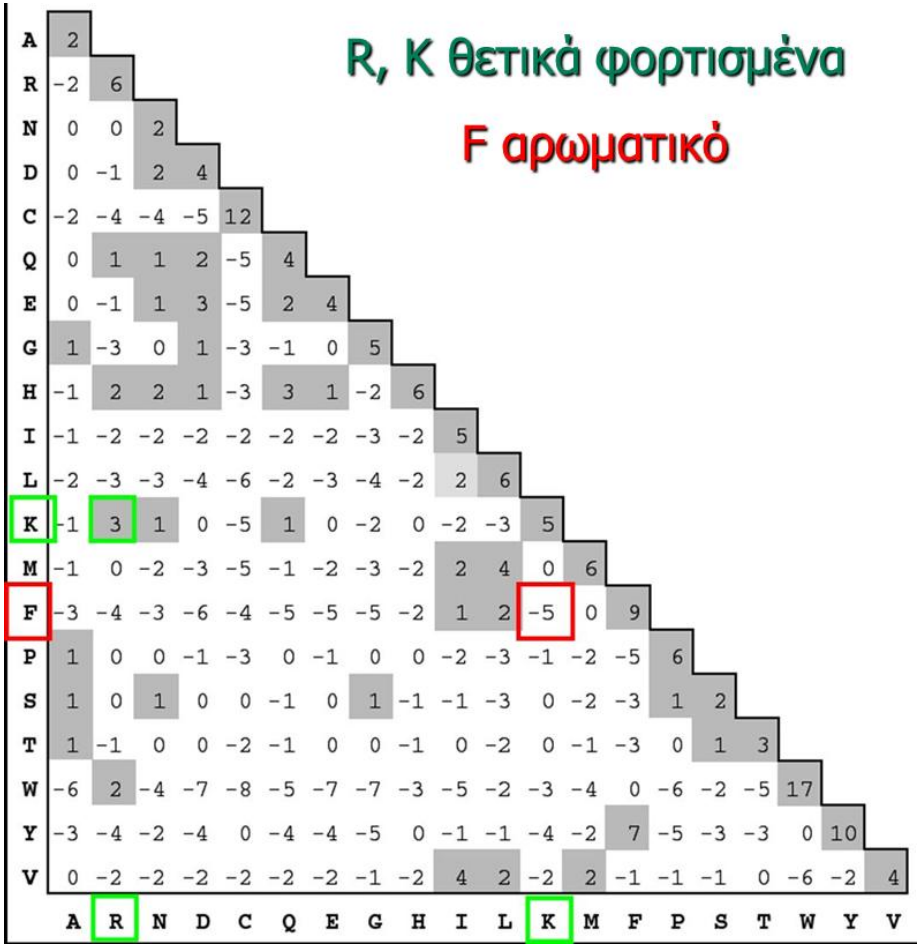
    MA**TG**G
  - Κάθε αμινοξύ αποκλίνει με διαφορετικό ρυθμό.

Πίνακες αντικατάστασης 2/8

- **PAM (Point Accepted Mutation)** (Dayhoff):
  - πίνακας **PAM1**:
    - υπολογίσθηκε βάσει **ολικών στοιχίσεων** από 71 οικογένειες πρωτεϊνών με **>85% ομοιότητα**.
  - πίνακας **PAM250**:
    - υπολογίζεται ως η **250-ιοστή δύναμη** του πίνακα PAM1.
  - Με τον ίδιο τρόπο κατασκευάζονται πίνακες PAM με διαφορετικούς δείκτες.
  - Οι πίνακες PAM με **μικρότερους δείκτες** χρησιμοποιούνται για τη στοίχιση **κοντινότερων εξελικτικά** ακολουθιών.

Πίνακες Αντικατάστασης 3/8

- **PAM (Point Accepted Mutation)** (Dayhoff).



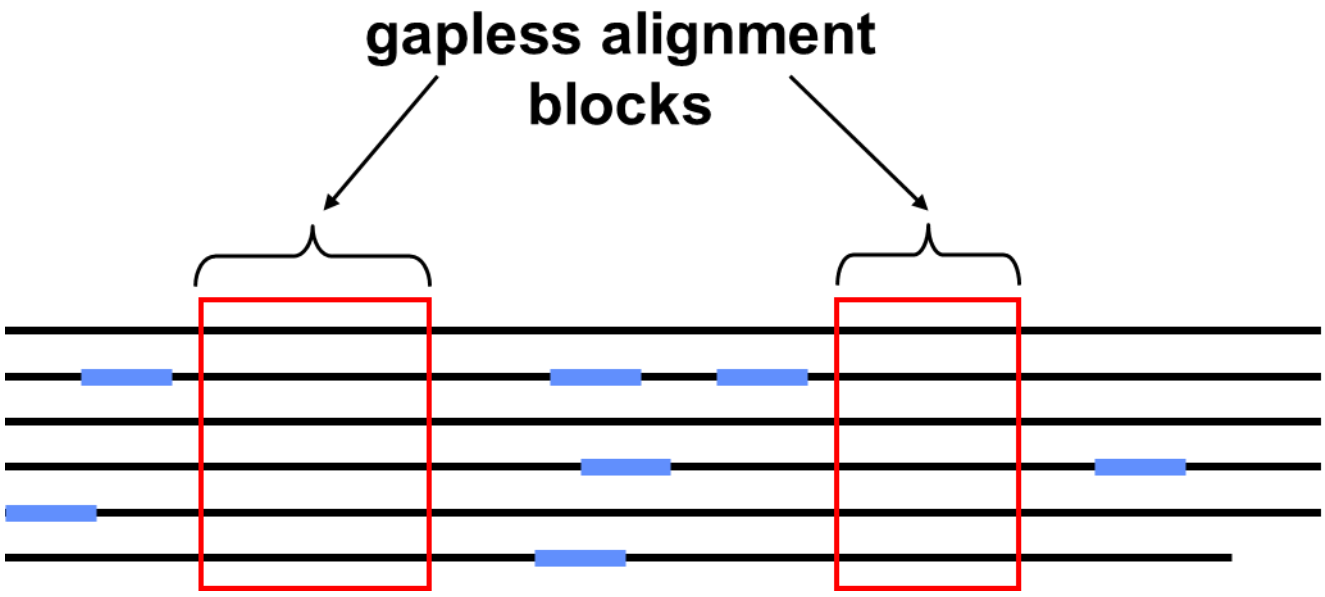
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I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

Πίνακες αντικατάστασης 4/8

- **PAM (Point Accepted Mutation)** (Dayhoff):
  - Θεωρεί ότι η πιθανότητα παρατήρησης μιας μετάλλαξης είναι ανεξάρτητη από:
    - την περιοχή της πρωτεΐνης.
    - προηγούμενες μεταλλάξεις στην ίδια θέση.
  - Στηρίζεται μόνο στις ιδιότητες σφαιρικών υδατοδιαλυτών πρωτεϊνών.

Πίνακες Αντικατάστασης 5/8

- **BLOSUM (Blocks Amino Acid Substitution Matrix)** (Henikoff).
- BLOCKS database:
  - **τοπική** πολλαπλή στοίχιση χωρίς κενά, εξελικτικά απομακρυσμένων πρωτεϊνών.

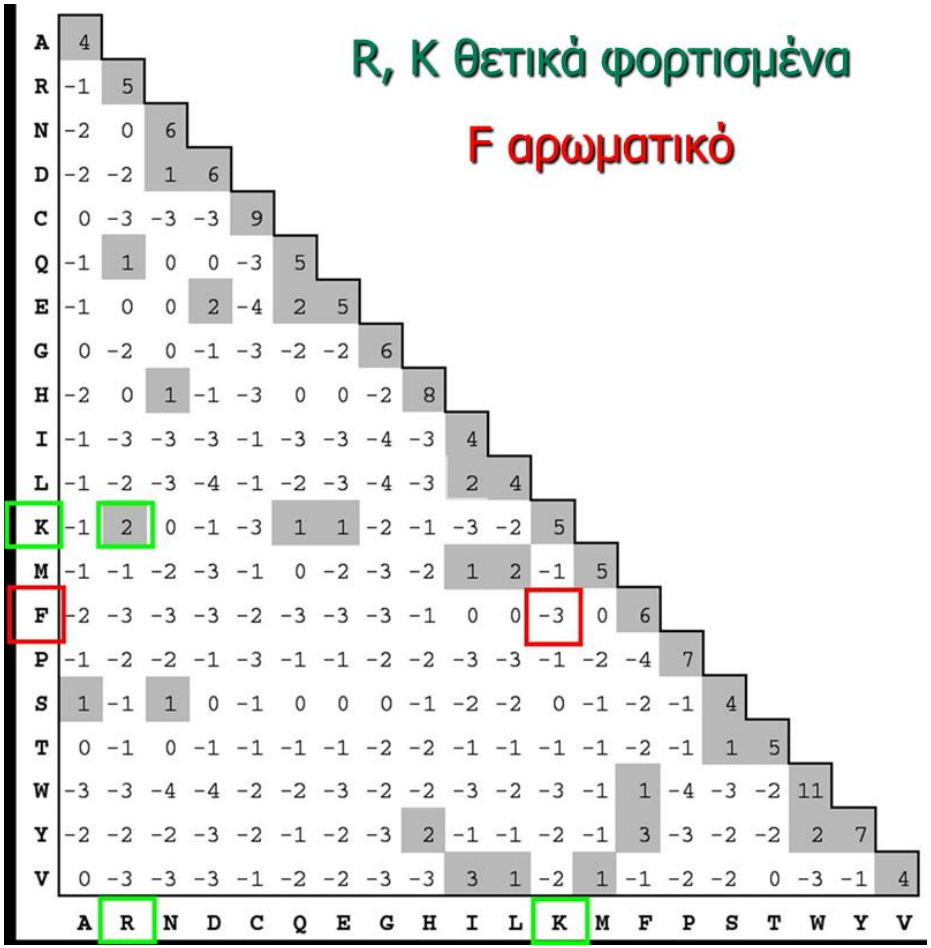


Πίνακες αντικατάστασης 6/8

- δείκτες πινάκων BLOSUM:
  - ποσοστό ομοιότητας των ακολουθιών που χρησιμοποιήθηκαν για την κατασκευή του πίνακα.
  - BLOSUM62: ακολουθίες με ομοιότητα >62% έχουν ομαδοποιηθεί.
- Οι πίνακες BLOSUM με **μικρότερους δείκτες** χρησιμοποιούνται για τη σύγκριση **περισσότερο απομακρυσμένων** εξελικτικά ακολουθιών.
- Στηρίζονται σε πραγματικές στοιχίσεις και δεν προκύπτουν αναγωγικά όπως οι πίνακες PAM.

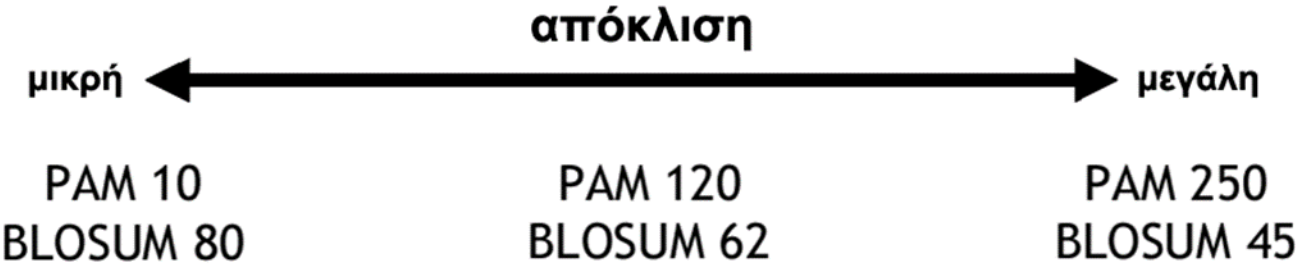
Πίνακες Αντικατάστασης 7/8

- **BLOSUM (Blocks Amino Acid Substitution Matrix)** (Henikoff).



Πίνακες Αντικατάστασης 8/8

- **PAM**:
  - designed to track the evolutionary origins of proteins.
  - στοίχιση συγγενών ακολουθιών.
- **BLOSUM**:
  - designed to find the conserved domains of proteins.
  - στοίχιση απομακρυσμένων ακολουθιών.
  - εύρεση τοπικών ομοιοτήτων.



Σύστημα Βαθμολόγησης 7/10

- **Μοντέλα νουκλεοτιδικών υποκαταστάσεων**.

	A	T	C	G	A	T	C	G
	(A) Jukes-Cantor model				(E) HKY model			
A	-	α	α	α	-	βg_T	βg_C	αg_G
T	α	-	α	α	βg_A	-	αg_C	βg_G
C	α	α	-	α	βg_A	αg_T	-	βg_G
G	α	α	α	-	αg_A	βg_T	βg_C	-
	(B) Kimura model				(F) Tamura-Nei model			
A	-	β	β	α	-	βg_T	βg_C	$\alpha_1 g_G$
T	β	-	α	β	βg_A	-	$\alpha_2 g_C$	βg_G
C	β	α	-	β	βg_A	$\alpha_2 g_T$	-	βg_G
G	α	β	β	-	$\alpha_1 g_A$	βg_T	βg_C	-
	(C) Equal-input model				(G) General reversible model			
A	-	αg_T	αg_C	αg_G	-	ag_T	bg_C	cg_G
T	αg_A	-	αg_C	αg_G	ag_A	-	dg_C	eg_G
C	αg_A	αg_T	-	αg_G	bg_A	dg_T	-	fg_G
G	αg_A	αg_T	αg_C	-	cg_A	eg_T	fg_C	-
	(D) Tamura model				(H) Unrestricted model			
A	-	$\beta\theta_2$	$\beta\theta_1$	$\alpha\theta_1$	-	a_{12}	a_{13}	a_{14}
T	$\beta\theta_2$	-	$\alpha\theta_1$	$\beta\theta_1$	a_{21}	-	a_{23}	a_{24}
C	$\beta\theta_2$	$\alpha\theta_2$	-	$\beta\theta_1$	a_{31}	a_{32}	-	a_{34}
G	$\alpha\theta_2$	$\beta\theta_2$	$\beta\theta_1$	-	a_{41}	a_{42}	a_{43}	-

Note: An element (e_{ij}) of the above substitution matrices stands for the substitution rate from the nucleotide in the i-th row to the nucleotide in the j-th column. g_A, g_T, g_C, and g_G are the nucleotide frequencies. $\theta_1 = g_G + g_C$. $\theta_2 = g_A + g_T$. $g_R = g_A + g_G$. $g_Y = g_T + g_C$.

Σύστημα βαθμολόγησης 8/10

- **Μοντέλα βαθμολόγησης των κενών**:
  - κενό μήκους k καταλοίπων.
  - **Linear gap penalty**:
    - gap penalty a.
    - gp(k) = ak.
  - **Affine gap penalty**:
    - gap-opening penalty b (ποινή για το άνοιγμα κενού).
    - gap-extension penalty a (ποινή για την επέκταση κενού).
    - gp(k) = b + ak , (|b| > |a|).
    - ευνοεί την εισαγωγή λίγων μεγάλων κενών σε σχέση με πολλά μικρότερα κενά.

Σύστημα Βαθμολόγησης 9/10

- Το **ολικό score** της στοίχισης προκύπτει από το άθροισμα των scores των ζευγών καταλοίπων και των ποινών για τα κενά.
- **Εμπειρική επιλογή** τιμών ποινής για την εισαγωγή και την επέκταση κενών.
  - http://www.ebi.ac.uk/help/matrix.html

Protein Query Length	Matrix	Open Gap	Extend Gap
>300	BLOSUM50	-10	-2
85-300	BLOSUM62	-7	-1
50-85	BLOSUM80	-16	-4
>300	PAM250	-10	-2
85-300	PAM120	-16	-4
35-85	MDM40	-12	-2
<=35	MDM20	-22	-4
<=10	MDM10	-23	-4

Σύστημα Βαθμολόγησης 10/10

- Blosum62.
- gap-opening penalty = b = -10.
- gap-extension penalty = a = -1.

- score = S(K,K)+S(A,K)+S(H,H)+S(G,G)+S(K,V)+S(K,T)-(|b|+|a|•1) = 5 + (-1) + 8 + 6 + (-2) + (-1) - (10+1•2) = 3.

K	L	V	A	H	G	K	K
K	-	-	K	H	G	V	T

Βιβλιογραφία

- David Mount, "Bioinformatics: Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press; 2nd edition (March 12, 2013).

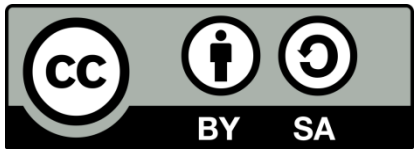
- Jonathan Pevsner, "Bioinformatics and Functional Genomics", Wiley-Blackwell; 2nd edition (May 4, 2009).

- Andreas D. Baxevanis, B. F. Francis Ouellette, "Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins", Wiley-Interscience; 3rd edition (October 29, 2004).

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.

- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδεια χρήσης, η άδεια χρήσης αναφέρεται ρητώς.



Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στο πλαίσιο του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα Γεωπονικού Πανεπιστημίου Αθηνών**» έχει χρηματοδοτήσει μόνο την 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

Σημείωμα Αναφοράς

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.