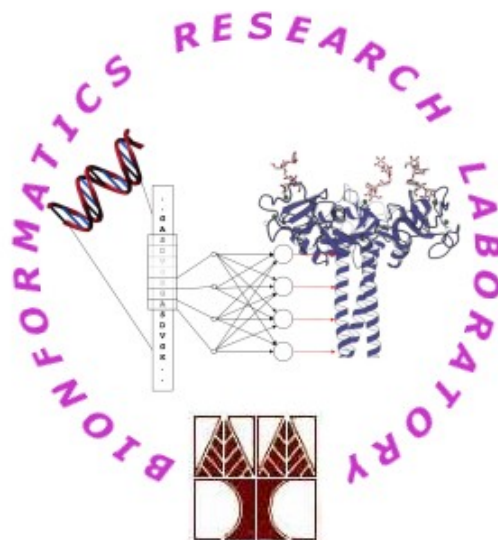


# Αλγόριθμοι Εύρεσης Ομοιοτήτων Ακολουθιών Μέρος III: Έλεγχος στατιστικής σημαντικότητας

## Πίνακες αντικατάστασης για σύγκριση ακολουθιών



Vasilis Promponas  
Bioinformatics Research Laboratory  
Department of Biological Sciences  
University of Cyprus

# ΣΥΝΟΨΗ

- Έλεγχος στατιστικής σημαντικότητας
  - Το πρόβλημα ...
  - Απλοϊκή προσέγγιση
  - Στατιστική προσέγγιση – η εξίσωση Karlin-Altschul
- Συστήματα Βαθμονόμησης Στοιχίσεων και Πίνακες Αντικατάστασης
  - Διαισθητική προσέγγιση
  - Εξελικτικές/Στατιστικές προσεγγίσεις
    - Αποδεκτές Σημειακές Μεταλλαγές και Πίνακες Αντικατάστασης PAM
    - Πίνακες BLOSUM
    - Εναλλακτικές Προσεγγίσεις
- Συζήτηση
  - ...

# Το πρόβλημα ...

α)  
>P01922|HBA\_HUMAN GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDDLHAKL  
G+ +VK HGKKV A ++ +AH+D++ + LS+LH KL  
>P02023|HBB\_HUMAN GNPVKKAHGKKVLGAFSDGLAHLNLDLKGTFATLSELHCDKL

β)  
>P01922|HBA\_HUMAN GSAQVKGHGKKVADALTNA-----VAHVDDMPNALSALSDDLHAKL  
+ +++ H KV + A V V L L +H K  
>P02240|LGB2\_LUPLU NNPELQAHAGKVFVKLVYEAAIQVQVTDATLKNLGSVHVSKE

γ)  
>P01922|HBA\_HUMAN GSAQVKGHGKKVADALT----NAVAHVDDMPNALSALSD----LHAKL  
G G V D+LT H D+ A +AL D AH+  
>P91253|GTS7\_CAEEEL -----GSGYLVGDSLTFVDLLVAQHTADLLAANAALLDEFPQFKAHQE

Score

101

17

32

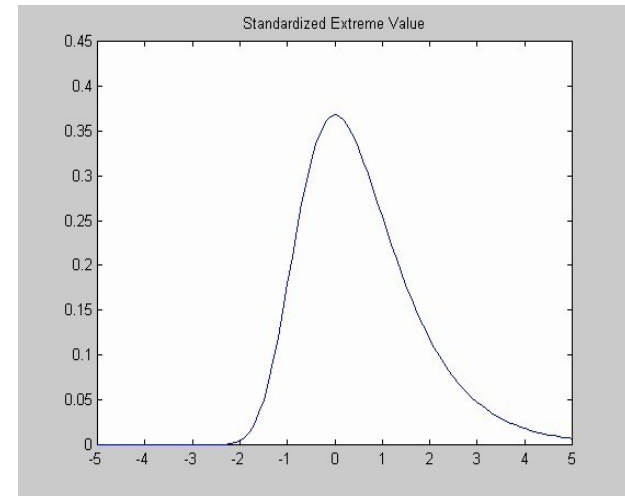
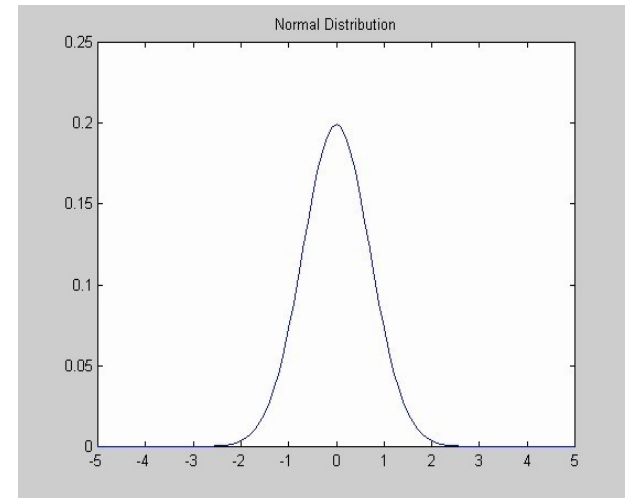
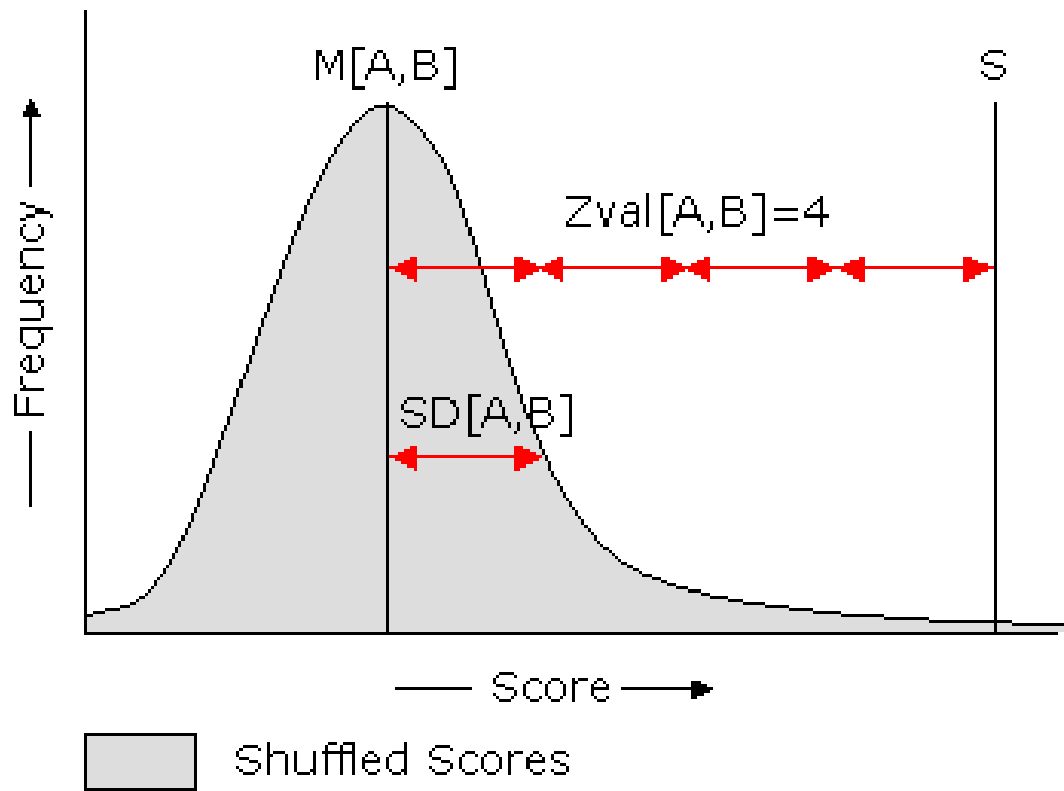
[Related Structures](#)

Sequences producing significant alignments:	Score (bits)
<a href="#">gi 17647349 ref NP_523840.1 </a> CG12240-PA, isoform A [Drosoph...	<a href="#">407</a>
<a href="#">gi 40215753 gb AAL48038.2 </a> LP04971p [Drosophila melanogaste...	<a href="#">380</a>
<a href="#">gi 24762579 ref NP_611892.1 </a> CG12240-PB, isoform B [Drosoph...	<a href="#">371</a>
<a href="#">gi 58392349 ref XP_319298.2 </a> ENSANGP00000012354 [Anopheles ...	<a href="#">102</a>
<a href="#">gi 34785543 gb AAH57849.1 </a> DNAJC4 protein [Homo sapiens] >g...	<a href="#">64</a>
<a href="#">gi 29835195 gb AAH51008.1 </a> DNAJC4 protein [Homo sapiens]	<a href="#">64</a>
<a href="#">gi 33877996 gb AAH32137.1 </a> DNAJC4 protein [Homo sapiens]	<a href="#">64</a>
<a href="#">gi 34894012 ref NP_908331.1 </a> putative DnaJ-like protein [Or...	<a href="#">63</a>
<a href="#">gi 55636307 ref XP_522047.1 </a> PREDICTED: similar to Vascular...	<a href="#">63</a>
<a href="#">gi 51261928 gb AAH79955.1 </a> Dnajc4-prov protein [Xenopus tro...	<a href="#">62</a>
<a href="#">gi 18447132 gb AAL68157.1 </a> AT30646p [Drosophila melanogaster]	<a href="#">62</a>
<a href="#">gi 7299733 gb AAF54914.1 </a> CG8476-PA [Drosophila melanogaste...	<a href="#">62</a>
<a href="#">gi 28950166 emb CAD71034.1 </a> hypothetical protein [Neurospor...	<a href="#">62</a>
<a href="#">gi 53765858 ref ZP_00186718.2 </a> COG0484: DnaJ-class molecula...	<a href="#">62</a>
<a href="#">gi 61859440 ref XP_587907.1 </a> PREDICTED: similar to DnaJ hom...	<a href="#">61</a>
<a href="#">gi 61843081 ref XP_613027.1 </a> PREDICTED: similar to DnaJ hom...	<a href="#">61</a>
<a href="#">gi 56755781 gb AAW26069.1 </a> unknown [Schistosoma japonicum]	<a href="#">61</a>
<a href="#">gi 54032310 ref ZP_00364442.1 </a> COG2214: DnaJ-class molecula...	<a href="#">60</a>
<a href="#">gi 31541124 gb AAP56426.1 </a> DnaJ [Mycoplasma gallisepticum R...	<a href="#">59</a>
<a href="#">gi 57099699 ref XP_533246.1 </a> PREDICTED: similar to hypothet...	<a href="#">59</a>
<a href="#">gi 854466 emb CAA89929.1 </a> unknown [Saccharomyces cerevisiae...	<a href="#">59</a>
<a href="#">gi 37362683 ref NP_013941.2 </a> One of several homologs of bac...	<a href="#">59</a>
<a href="#">gi 16800577 ref NP_470845.1 </a> heat shock protein DnaJ [Liste...	<a href="#">59</a>
<a href="#">gi 16803512 ref NP_464997.1 </a> heat shock protein DnaJ [Liste...	<a href="#">59</a>
<a href="#">gi 49651399 emb CAG78338.1 </a> unnamed protein product [Yarrow...	<a href="#">59</a>
<a href="#">gi 57088065 ref XP_536990.1 </a> PREDICTED: similar to DnaJ hom...	<a href="#">58</a>
<a href="#">gi 4007007 emb CAA66720.1 </a> 1(2)tid [Drosophila melanogaster...	<a href="#">58</a>
<a href="#">gi 46907700 ref YP_014089.1 </a> chaperone protein DnaJ [Lister...	<a href="#">58</a>
<a href="#">gi 42527589 ref NP_972687.1 </a> DnaJ domain protein [Treponema...	<a href="#">58</a>
<a href="#">gi 62897771 dbj BAD96825.1 </a> DnaJ (Hsp40) homolog, subfamily...	<a href="#">58</a>
<a href="#">gi 55643335 ref XP_510781.1 </a> PREDICTED: DnaJ (Hsp40) homolo...	<a href="#">58</a>
<a href="#">gi 50912997 ref XP_467906.1 </a> DNAJ heat shock N-terminal dom...	<a href="#">58</a>
<a href="#">gi 61363502 gb AAW42402.1 </a> DnaJ-like subfamily A member 3 [...	<a href="#">58</a>
<a href="#">gi 28950130 emb CAD70988.1 </a> related to SCJ1 protein [Neuros...	<a href="#">58</a>

# Στατιστική Σημαντικότητα (Τυπική Προσέγγιση)

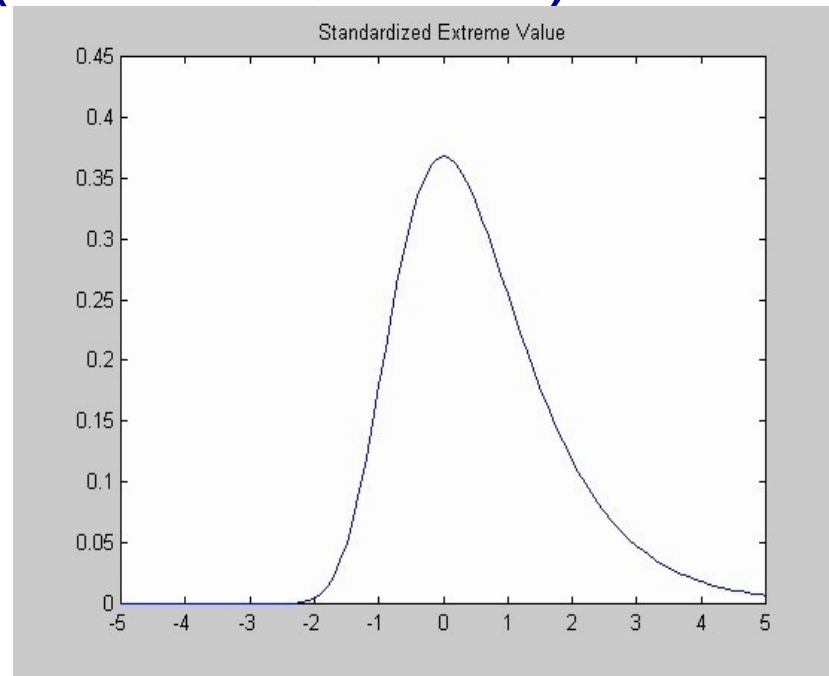
- Σύγκριση του score της πραγματικής στοίχισης με αυτά τυχαίων στοιχίσεων
  - Δημιουργία δείγματος τυχαίων scores
  - Υπολογισμός Μέσης Τιμής και Τυπικής Απόκλισης
  - Υπολογισμός της Απόκλισης του score της στοίχισης από τη Μέση Τιμή:  $Z = (s - s_{\text{mean}}) / sd$ 
    - $Z < 3 \text{ SD} \Rightarrow$  improbable
    - $3 \text{ SD} < Z < 5 \text{ SD} \Rightarrow$  marginal
    - $5 \text{ SD} < Z < 10 \text{ SD} \Rightarrow$  probable
    - $Z > 10 \text{ SD} \Rightarrow$  certain (???)

# Αλλά είναι «προσέγγιση»



# Η κατανομή των Ακραίων Τιμών (Gumbel, 1958)

$$P(S \leq s) \approx e^{-Kmne^{\lambda s}}$$



Η πιθανότητα να προκύψει στοίχιση με score  $> s$

$$P\text{-value} \equiv P(S \geq s) = 1 - e^{-Kmne^{\lambda s}}$$

Αναμενόμενη τιμή (Expect value)

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

# ... και λίγο “στατιστική” ακολουθιών

Έστω το “αλφάβητο”:  $A = \{a_1, a_2, \dots, a_r\}$

- Πρωτεΐνες:  $r=20$
- DNA/RNA:  $r=4$

και μια “τυχαία” αλληλουχία

$$S = s_1 s_2 \dots s_m$$

η οποία έχει προκύψει με τυχαία “δειγματοληψία” του  $A$  με κατανομή πιθανοτήτων  $P = \{p_1, p_2, \dots, p_r\}$



# Η εξίσωση Karlin-Altschul (Karlin, S. and S. F. Altschul, 1990)

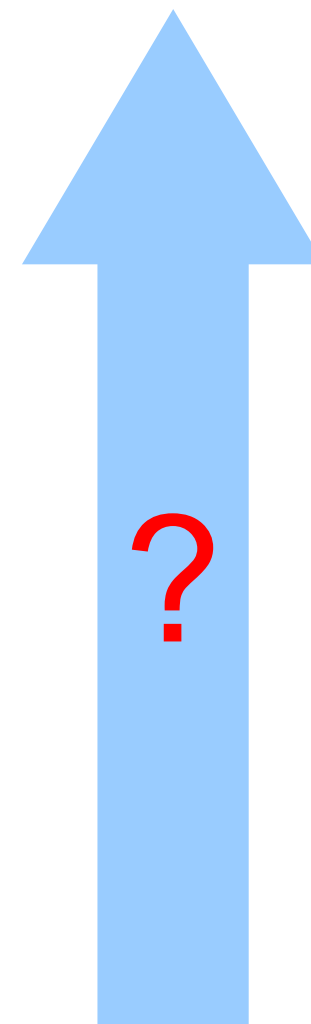
$$E(S \geq s) = Kmne^{-\lambda s}$$

- $n$ : μήκος βάσης δεδομένων
- $mn$ : μέγεθος χώρου αναζήτησης
- $\lambda s$ : κανονικοποιημένο score
- $K$ : σταθερά
- $K, \lambda$  εξαρτώνται από το σύστημα βαθμονόμησης

**ΤΙ ΕΚΦΡΑΖΕΙ ΤΟ  $E$ ?**

[Related Structures](#)

Sequences producing significant alignments:	Score (bits)	E Value	
<a href="#">gi 17647349 ref NP_523840.1 </a> CG12240-PA, isoform A [Drosoph...	<a href="#">407</a>	e-112	<a href="#">G</a>
<a href="#">gi 40215753 gb AAL48038.2 </a> LP04971p [Drosophila melanogaste...	<a href="#">380</a>	e-104	
<a href="#">gi 24762579 ref NP_611892.1 </a> CG12240-PB, isoform B [Drosoph...	<a href="#">371</a>	e-102	<a href="#">G</a>
<a href="#">gi 58392349 ref XP_319298.2 </a> ENSANGP00000012354 [Anopheles ...	<a href="#">102</a>	6e-21	<a href="#">G</a>
<a href="#">gi 34785543 gb AAH57849.1 </a> DNAJC4 protein [Homo sapiens] >g...	<a href="#">64</a>	2e-09	<a href="#">G</a>
<a href="#">gi 29835195 gb AAH51008.1 </a> DNAJC4 protein [Homo sapiens]	<a href="#">64</a>	2e-09	<a href="#">G</a>
<a href="#">gi 33877996 gb AAH32137.1 </a> DNAJC4 protein [Homo sapiens]	<a href="#">64</a>	2e-09	<a href="#">G</a>
<a href="#">gi 34894012 ref NP_908331.1 </a> putative DnaJ-like protein [Or...	<a href="#">63</a>	6e-09	<a href="#">G</a>
<a href="#">gi 55636307 ref XP_522047.1 </a> PREDICTED: similar to Vascular...	<a href="#">63</a>	6e-09	<a href="#">G</a>
<a href="#">gi 51261928 gb AAH79955.1 </a> Dnajc4-prov protein [Xenopus tro...	<a href="#">62</a>	8e-09	<a href="#">G</a>
<a href="#">gi 18447132 gb AAL68157.1 </a> AT30646p [Drosophila melanogaster]	<a href="#">62</a>	8e-09	
<a href="#">gi 7299733 gb AAF54914.1 </a> CG8476-PA [Drosophila melanogaste...	<a href="#">62</a>	8e-09	<a href="#">G</a>
<a href="#">gi 28950166 emb CAD71034.1 </a> hypothetical protein [Neurospor...	<a href="#">62</a>	1e-08	<a href="#">G</a>
<a href="#">gi 53765858 ref ZP_00186718.2 </a> COG0484: DnaJ-class molecula...	<a href="#">62</a>	1e-08	
<a href="#">gi 61859440 ref XP_587907.1 </a> PREDICTED: similar to DnaJ hom...	<a href="#">61</a>	2e-08	<a href="#">G</a>
<a href="#">gi 61843081 ref XP_613027.1 </a> PREDICTED: similar to DnaJ hom...	<a href="#">61</a>	2e-08	<a href="#">G</a>
<a href="#">gi 56755781 gb AAW26069.1 </a> unknown [Schistosoma japonicum]	<a href="#">61</a>	2e-08	
<a href="#">gi 54032310 ref ZP_00364442.1 </a> COG2214: DnaJ-class molecula...	<a href="#">60</a>	5e-08	
<a href="#">gi 31541124 gb AAP56426.1 </a> DnaJ [Mycoplasma gallisepticum R...	<a href="#">59</a>	7e-08	<a href="#">G</a>
<a href="#">gi 57099699 ref XP_533246.1 </a> PREDICTED: similar to hypothet...	<a href="#">59</a>	7e-08	<a href="#">G</a>
<a href="#">gi 854466 emb CAA89929.1 </a> unknown [Saccharomyces cerevisiae...	<a href="#">59</a>	9e-08	
<a href="#">gi 37362683 ref NP_013941.2 </a> One of several homologs of bac...	<a href="#">59</a>	9e-08	<a href="#">G</a>
<a href="#">gi 16800577 ref NP_470845.1 </a> heat shock protein DnaJ [Liste...	<a href="#">59</a>	1e-07	<a href="#">G</a>
<a href="#">gi 16803512 ref NP_464997.1 </a> heat shock protein DnaJ [Liste...	<a href="#">59</a>	1e-07	<a href="#">G</a>
<a href="#">gi 49651399 emb CAG78338.1 </a> unnamed protein product [Yarrow...	<a href="#">59</a>	1e-07	<a href="#">G</a>
<a href="#">gi 57088065 ref XP_536990.1 </a> PREDICTED: similar to DnaJ hom...	<a href="#">58</a>	2e-07	<a href="#">G</a>
<a href="#">gi 4007007 emb CAA66720.1 </a> 1(2)tid [Drosophila melanogaster...	<a href="#">58</a>	2e-07	
<a href="#">gi 46907700 ref YP_014089.1 </a> chaperone protein DnaJ [Lister...	<a href="#">58</a>	2e-07	<a href="#">G</a>
<a href="#">gi 42527589 ref NP_972687.1 </a> DnaJ domain protein [Treponema...	<a href="#">58</a>	2e-07	<a href="#">G</a>
<a href="#">gi 62897771 dbj BAD96825.1 </a> DnaJ (Hsp40) homolog, subfamily...	<a href="#">58</a>	2e-07	<a href="#">G</a>
<a href="#">gi 55643335 ref XP_510781.1 </a> PREDICTED: DnaJ (Hsp40) homolo...	<a href="#">58</a>	2e-07	<a href="#">G</a>
<a href="#">gi 50912997 ref XP_467906.1 </a> DNAJ heat shock N-terminal dom...	<a href="#">58</a>	2e-07	<a href="#">G</a>
<a href="#">gi 61363502 gb AAW42402.1 </a> DnaJ-like subfamily A member 3 [...	<a href="#">58</a>	2e-07	
<a href="#">gi 28950130 emb CAD70988.1 </a> related to SCJ1 protein [Neuros...	<a href="#">58</a>	2e-07	



# Ιδιότητες της εξίσωσης K-A

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

- **Παραδοχές:**
  - Τοπικές Στοιχίσεις ΧΩΡΙΣ κενά
  - Οι ακολουθίες είναι ανεξάρτητες και προϊόντα της ίδιας κατανομής (iid variables)
  - Έχουν ΑΠΕΙΡΟ μήκος
  - Υπάρχει τουλάχιστον μια θετική τιμή ταιριάσματος δυο καταλοίπων
  - Η αναμενόμενη τιμή ταιριάσματος καταλοίπων είναι αρνητική
- K: “κανονίζει” γειτονικές στοιχίσεις (τυπικά  $K \sim 0.1$ )
- Η τιμή **E** ελαττώνεται ΕΚΘΕΤΙΚΑ με το **S**
- Η τιμή **E** αυξάνεται ΓΡΑΜΜΙΚΑ με το **mn**

# Στοιχίσεις με κενά

- Δεν καλύπτονται από την εξίσωση ***K-A!!***
- Αναλυτικά δεν μπορώ να υπολογίσω  $K$ ,  $\lambda$
- Εξάρτηση από τιμές ποινής κενών
- Εμπειρικός υπολογισμός
  - Στοιχίσεις τυχαίων ακολουθιών
  - Καταγραφή χαρακτηριστικών HSPs (scores, συχνότητες υποβάθρου, μήκη)
  - Υπολογισμός  $K$ ,  $\lambda$  με βάση το “πλησιέστερο” σύστημα βαθμονόμησης (χωρίς κενά)
- Προσομοιώσεις έδειξαν ότι η προσέγγιση δεν είναι και άσχημη ( $\sim$ EVD)

# Στοιχίσεις με κενά (μέρος Β)

Gap open	Gap extend	$\lambda$	k	H (nats)
No gaps allowed	No gaps allowed	0.318	0.134	0.40
11	2	0.297	0.082	0.27
10	2	0.291	0.075	0.23
7	2	0.239	0.027	0.10

From Ian Korf, Mark Yandell & Joseph Bedelle, 2004

# Διορθώσεις σχετικά με τα μήκη

- Χώρος αναζήτησης  $mn$

- Υποθέτει

1. είναι δυνατόν να βρω HSPs σε όλο το μήκος των ακολουθιών με την ίδια πιθανότητα **[ΓΙΑΤΙ ΟΧΙ??]**

2. η βάση δεδομένων είναι ΜΙΑ ακολουθία **[ΓΙΑΤΙ??]**

- ΛΥΣΗ K-A

Υπολογισμός του ~μήκους ενός HSP (expected HSP length)

$$l = \ln(Kmn)/H$$

Επομένως, ο χώρος αναζήτησης είναι μικρότερος

$$m'n' = (m-l)(n - \text{DBsize} * l)$$

# Βαθμολογίες και μήκος HSP

Ορίζουμε:

$$s_{bit} = \frac{\lambda s - \ln K}{\ln 2} \Leftrightarrow s = \frac{s_{bit} \ln 2 + \ln K}{\lambda}$$

Οπότε:

$$\begin{aligned} E\text{-value} &\equiv E(S_{bit} > s_{bit}) \\ &= Kmne^{-\lambda s} = Kmne^{-(s_{bit} \ln 2 + \ln K)} \\ &= Kmne^{-s_{bit} \ln 2} e^{-\ln K} = Kmne^{\ln 2^{-s_{bit}}} \frac{1}{K} \\ &= mn 2^{-s_{bit}} \end{aligned}$$

# Αξιολόγηση με βάση E-values

- Η τιμή κατωφλίου στατιστικής σημαντικότητας εξαρτάται από τον τύπο των συμπερασμάτων στα οποία θέλουμε να καταλήξουμε
- Κατάλληλη επιλογή συστήματος βαθμονόμησης [NEXT LECTURE!!]
- Έλεγχος στοιχίσεων και αξιοποίηση κάθε άλλης διαθέσιμης πληροφορίας
- Δύο Απλοί Κανόνες (???)



# Αξιολόγηση με βάση E-values

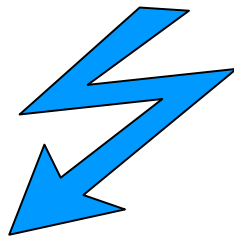
<b>Τύπος Ακολουθίας</b>	<b>E-value</b>	<b>Ταυτότητα Καταλοίπων</b>
Νουκλεοτιδική	$<10^{-6}$	$>70\%$
Αμινοξική	$<10^{-3}$	$>25\%$

# Αξιολόγηση με βάση E-values (II)

- Η εμφάνιση της ομοιότητας σε επαρκές μήκος των ακολουθιών (π.χ. 80 κατάλοιπα, που αντιστοιχούν σε συνήθη μήκη δομικών και λειτουργικών περιοχών)
- Η ύπαρξη σημαντικού ποσοστού ταυτόσημων καταλοίπων (π.χ. ταυτότητα 30% συνεπάγεται, συνήθως, παρόμοιο δίπλωμα) στην περιοχή της στοίχισης
- Η ταυτόχρονη εμφάνιση χαρακτηριστικών δομικών-λειτουργικών μοτίβων
- Η ταύτιση καταλοίπων τα οποία είναι γνωστό πειραματικά ότι είναι σημαντικά για τη δομή-λειτουργία

# Η σημασία των συστημάτων βαθμολόγησης

**Η  
Επιτακτική  
Ανάγκη**



**ΣΥΓΚΡΙΣΗ ΑΚΟΛΟΥΘΙΩΝ  
(DNA/RNA/ΠΡΩΤΕΪΝΩΝ)**

**Απαιτεί**

**ΔΗΜΙΟΥΡΓΙΑ  
(ΒΕΛΤΙΣΤΩΝ)  
ΣΤΟΙΧΙΣΕΩΝ**

**ΧΡΗΣΗ ΚΑΤΑΛΛΗΛΟΥ ΣΥΣΤΗΜΑΤΟΣ  
ΒΑΘΜΟΝΟΜΗΣΗΣ (Scoring System)**

- ΠΙΝΑΚΕΣ ΑΝΤΙΚΑΤΑΣΤΑΣΗΣ
- ΠΟΙΝΕΣ ΕΙΣΑΓΩΓΗΣ ΚΕΝΩΝ

# Χρήση “αθροιστικών” συστημάτων

- Για κάθε θέση σε μία στοίχιση αντιστοιχώ μια “βαθμολογία”
  - Τι μπορεί να εκφράζει??
- Η βαθμολογία της στοίχισης προκύπτει από το άθροισμα των βαθμολογιών
- Τι γίνεται όμως με τον Έλεγχο Υποθέσεων ( $H_0, H_1$ )??
  - Θα ήθελα να έχω πιθανότητες => ολική πιθανότητα της στοίχισης
  - Βολεύει?

# Πίνακες Αντικατάστασης (ΠΑ)

(aka Substitution/Mutation/Scoring Matrices)

Στη γενική περίπτωση (π.χ. DNA):

$$S = \begin{pmatrix} S_{a,a} & S_{a,t} & S_{a,c} & S_{a,g} \\ S_{t,a} & S_{t,t} & S_{t,c} & S_{t,g} \\ S_{c,a} & S_{c,t} & S_{c,c} & S_{c,g} \\ S_{g,a} & S_{g,t} & S_{g,c} & S_{g,g} \end{pmatrix}$$

Συνήθως είναι Συμμετρικοί πίνακες ( $s_{i,j} = s_{j,i}$ )

Είναι δυνατόν να εκφράζουν μέτρο

**ΟΜΟΙΟΤΗΤΑΣ** ή **ΑΠΟΣΤΑΣΗΣ**

# Πως μπορώ να δημιουργήσω ένα (χρήσιμο) ΠΑ?

- Για Νουκλεοτιδικές Ακολουθίες:
  - Συνήθως απλοί εμπειρικοί πίνακες [προβλήματα??]
  - Δεν υπονοείται κάποιο συγκεκριμένο εξελικτικό μοντέλο [??]
- Για Αμινοξικές Ακολουθίες:
  - Ορισμένες μεταλλάξεις αναμένεται να είναι λιγότερο πιθανές [ $K \Rightarrow R$ ,  $K \Rightarrow V$  ??]
  - Δομικές/Φυσικοχημικές/Λειτουργικές πιέσεις [εξελικτικές ??]
  - Εξάρτηση από τον τύπο των πρωτεϊνών και την εξελικτική τους συγγένεια
  - Προϊόν ανάλυσης ΠΡΑΓΜΑΤΙΚΩΝ δεδομένων

# Επιθυμητά χαρακτηριστικά ΠΑ

- Τα στοιχεία της κύριας διαγωνίου αναμένουμε να έχουν τη μεγαλύτερη τιμή της αντίστοιχης γραμμής/στήλης [ΓΙΑΤΙ??]
- Θετικά scores για συντηρητικές μεταλλαγές
- Προκειμένου για τοπικές στοιχίσεις αρνητική αναμενόμενη τιμή score (δείτε παρακάτω)
- Δυνατότητα διάκρισης σε διάφορες εξελικτικές αποστάσεις

# Ένας απλός ΠΑ

(για νουκλεοτιδικές ακολουθίες)

- Ο “Μοναδιαίος” Πίνακας Ταυτοτήτων (Identity Matrix)

$$S = \begin{pmatrix} S_{a,a} & S_{a,t} & S_{a,c} & S_{a,g} \\ S_{t,a} & S_{t,t} & S_{t,c} & S_{t,g} \\ S_{c,a} & S_{c,t} & S_{c,c} & S_{c,g} \\ S_{g,a} & S_{g,t} & S_{g,c} & S_{g,g} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$



# .. κι άλλοι δυο απλοί ΠΑ

Ποια η διαφορά??

*Ο Πίνακας Αντικατάστασης για στοιχίσεις Νουκλεοτιδικών ακολουθιών με το λογισμικό NCBI – BLAST*

$$S = \begin{pmatrix} S_{a,a} & S_{a,t} & S_{a,c} & S_{a,g} \\ S_{t,a} & S_{t,t} & S_{t,c} & S_{t,g} \\ S_{c,a} & S_{c,t} & S_{c,c} & S_{c,g} \\ S_{g,a} & S_{g,t} & S_{g,c} & S_{g,g} \end{pmatrix} = \begin{pmatrix} 2 & -1 & -1 & -1 \\ -1 & 2 & -1 & -1 \\ -1 & -1 & 2 & -1 \\ -1 & -1 & -1 & 2 \end{pmatrix}$$

*Ο Πίνακας Αντικατάστασης για στοιχίσεις Νουκλεοτιδικών ακολουθιών με τα λογισμικά FASTA/WU – BLAST*

$$S = \begin{pmatrix} S_{a,a} & S_{a,t} & S_{a,c} & S_{a,g} \\ S_{t,a} & S_{t,t} & S_{t,c} & S_{t,g} \\ S_{c,a} & S_{c,t} & S_{c,c} & S_{c,g} \\ S_{g,a} & S_{g,t} & S_{g,c} & S_{g,g} \end{pmatrix} = \begin{pmatrix} 5 & -4 & -4 & -4 \\ -4 & 5 & -4 & -4 \\ -4 & -4 & 5 & -4 \\ -4 & -4 & -4 & 5 \end{pmatrix}$$

# Πίνακες Αντικατάστασης

(για αμινοξικές ακολουθίες)

Δύο κύριες προσεγγίσεις

- Η οικογένεια των πινάκων PAM
  - Μαρκοβιανό μοντέλο εξέλιξης
  - Φυλογενετικά δέντρα
  - Λογαριθμικοί λόγοι πιθανοφάνειας (log-likelihood ratios)
- Η οικογένεια των πινάκων BLOSUM
  - Λογαριθμικοί λόγοι πιθανοφάνειας (log-likelihood ratios)

# Η Οικογένεια των Πινάκων PAM

(Dayhoff, Schwartz, and Orcutt, 1978)

- Βασίζεται σε στοχαστικές Μαρκοβιανές διαδικασίες και στην κατασκευή φυλογενετικών δέντρων για το “ταίριασμα” με ένα εξελικτικό μοντέλο
- Υπολογισμός Πίνακα μεταβάσεων μετά από απαρίθμηση αντικαταστάσεων
- Υπολογισμός Πίνακα Αντικατάστασης (log-odds ratio)

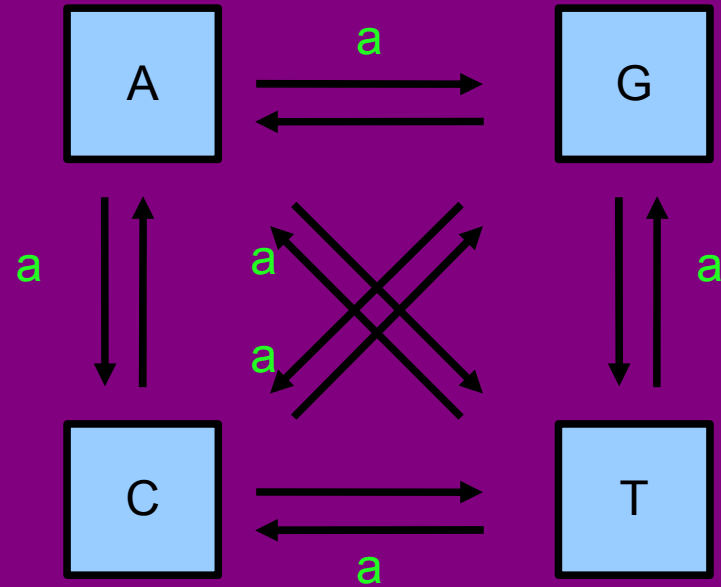
**KENTRIKH IDEIA**  
**PAM**  
**(Point Accepted Mutations)**

Μια Σημειακή Μεταλλαγή λέγεται **αποδεκτή** όταν έχει τη δυνατότητα να εξαπλωθεί και να επικρατήσει σε ένα πληθυσμό

Έχω ξαναδεί πίνακα μεταβάσεων??

# Έχω ξαναδεί πίνακα μεταβάσεων??

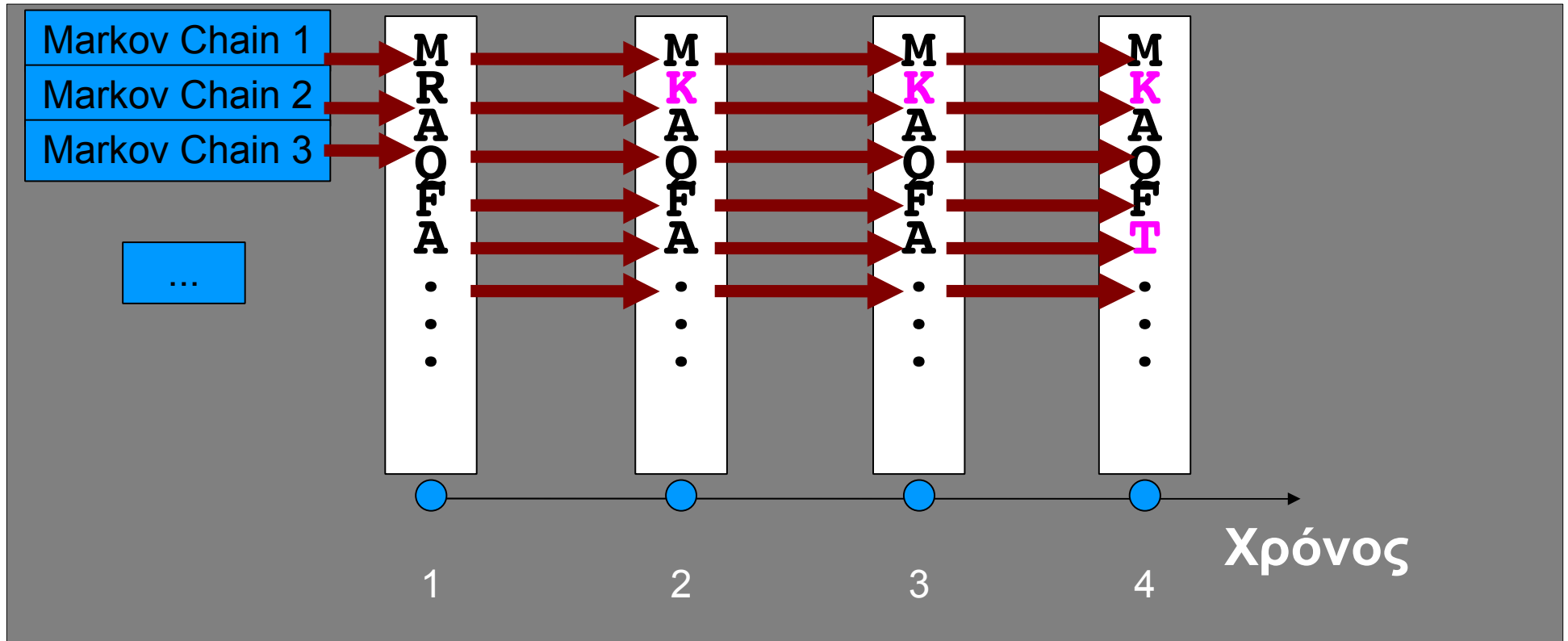
- Να!!!!!!!!!!!!!!!!!!!!!!



To:		A	G	C	T
From:	A	$1-3a$	$a$	$a$	$a$
	G	$a$	$1-3a$	$a$	$a$
	C	$a$	$a$	$1-3a$	$a$
	T	$a$	$a$	$a$	$1-3a$

# Το υποκείμενο εξελικτικό μοντέλο:

Κάθε θέση σε μια ακολουθία εξελίσσεται με βάση μια ανεξάρτητη (από τη θέση) Μαρκοβιανή διαδικασία



- Όλες αυτές οι διαδικασίες έχουν τον ίδιο πίνακα μεταβάσεων  $P$  (20x20)
- Προσδιορισμός από δεδομένα πρωτεϊνικών ακολουθιών

# Ο Πίνακας μεταβάσεων PAM1

Είναι ο πίνακας μεταβάσεων μιας Μαρκοβιανής διαδικασίας για μια χρονική περίοδο στην οποία αναμένουμε **1% των αμινοξικών καταλοίπων να υποστούν αποδεκτή σημειακή μεταλλαγή**

Στοίχιση “όμοιων” πρωτεϊνικών ακολουθιών (>85% id)

???? ΓΙΑΤΙ?

???? ΠΙΝΑΚΑΣ?

Ανακατασκευή φυλογενετικού δέντρου και ακολουθιών στους εσωτερικούς κόμβους

Απαρίθμηση αντικαταστάσεων

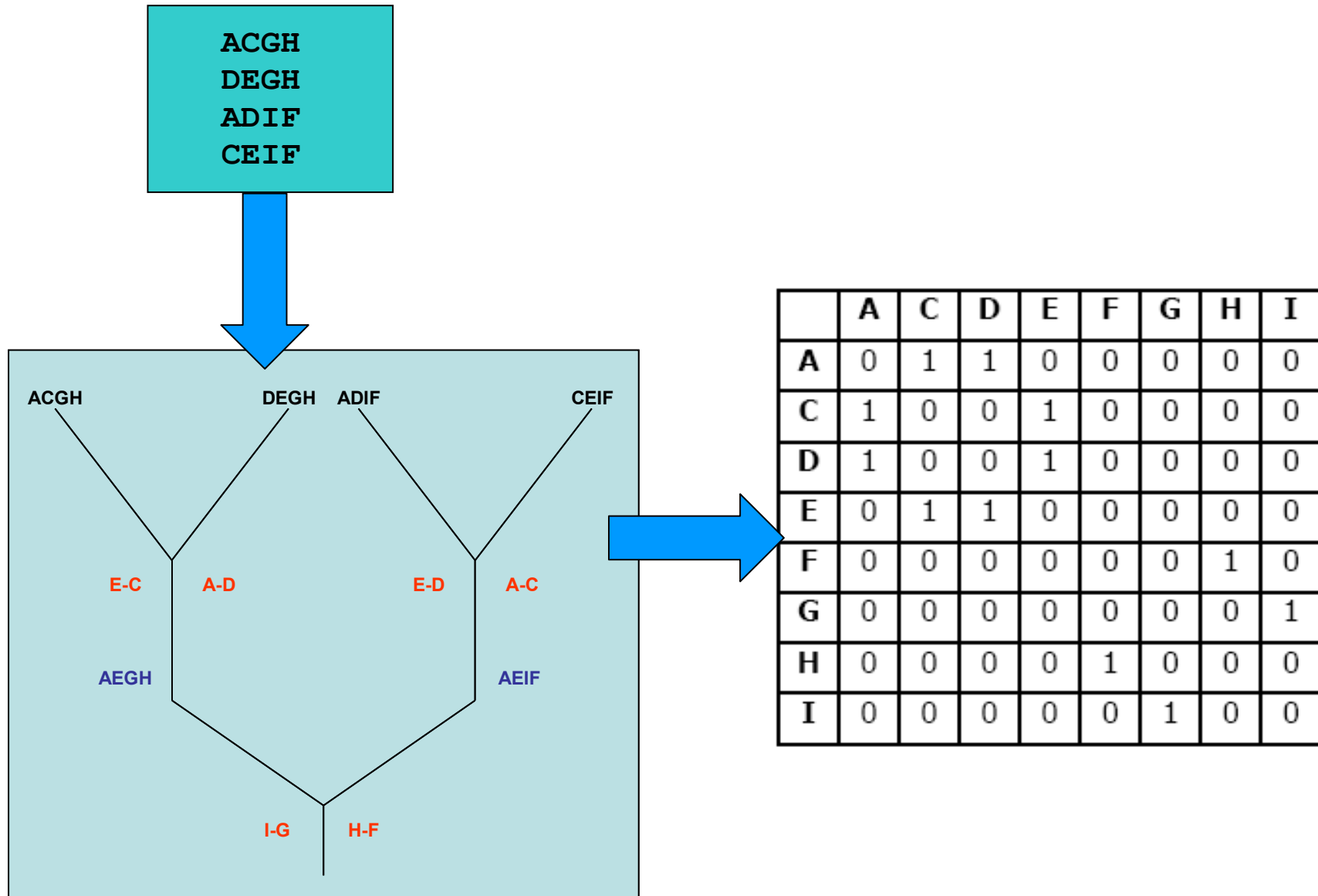
Εκτίμηση πιθανοτήτων αντικαταστάσεων

# PAM1 (pt. II)

- Πολλαπλές στοιχίσεις χωρίς κενά
- 71 πρωτεϊνικές οικογένειες
- Επιλογή καλά διατηρημένων περιοχών
- (Μόνο) 1572 αντικαταστάσεις
- Δημιουργία φυλογενετικών δέντρων => Maximum Parsimony
- Ανακατασκευή “αρχαίων” ακολουθιών και απαρίθμηση



# Ο πίνακας απαρίθμησης





# Σχετική μεταλλαξιμότητα και πίνακας μεταβάσεων

**Σχετική Μεταλλαξιμότητα:** Για κάθε τύπο αμινοξικού καταλοίπου  $X$ , εκτιμώ την πιθανότητα αντικατάστασής του από

$$m_x = \frac{\sum_Y F_{x,y}}{N_x}$$

## Πίνακας Μεταβάσεων

$$\mathbf{M}_{x,y} = \begin{cases} \frac{\lambda m_y F_{x,y}}{\sum_x F_{x,y}}, & \text{αν } X \neq Y \\ 1 - \lambda m_y, & \text{αν } X = Y \end{cases}$$

## ... και το Λάμδα?

- Παράγοντας Στάθμισης
- Τι σταθμίζει?
- Επιλογή τιμής  $\lambda$  ώστε 1% των καταλοίπων να παθαίνουν PAM => **PAM1**
- Η αντίστοιχη εξελικτική απόσταση: **1-PAM**

# Υπολογισμός $\lambda$ για 1-PAM

- Έστω ότι μελετούμε μια θέση μιας ακολουθίας και  $A_n$  είναι το κατάλοιπο τη χρονική στιγμή  $n$
- Η πιθανότητα αλλαγής του μετά από χρόνο 1PAM είναι

$$\begin{aligned} P &= P(A_1 \neq A_0) = \sum_{j=1}^{20} P(A_0 = j, A_1 \neq j) \\ &= \sum_{j=1}^{20} P(A_1 \neq j | A_0 = j) * P(A_0 = j) \\ &\approx \sum_{j=1}^{20} P(A_1 \neq j | A_0 = j) * F_j \end{aligned}$$

όπου  $F_j$  η παρατηρούμενη συχνότητα του καταλοίπου τύπου  $j$

# Υπολογισμός $\lambda$ για 1-PAM (pt II)

$$\begin{aligned} 0.01 &= \sum_{j=1}^{20} P(A_1 \neq j | A_0 = j) * F_j \\ &= \sum_{j=1}^{20} \left( \sum_{k \neq j} P(A_1 = k | A_0 = j) \right) * F_j \\ &\approx \sum_{j=1}^{20} \left( \sum_{k \neq j} p_{j,k} \right) * F_j \\ &\approx \sum_{j=1}^{20} \left( \sum_{k \neq j} \left( \frac{\lambda * m_j * F_{k,j}}{\sum_j F_{k,j}} \right) \right) * F_j \\ &= \lambda * \sum_{j=1}^{20} \left( \sum_{k \neq j} \left( \frac{m_j * F_{k,j}}{\sum_j F_{k,j}} \right) \right) * F_j \end{aligned}$$

$$\text{οπότε } \lambda = \frac{0.01}{\sum_{j=1}^{20} \left( \left( \sum_{k \neq j} \left( \frac{m_j * F_{k,j}}{\sum_j F_{k,j}} \right) \right) * F_j \right)}$$

ORIGINAL AMINO ACID

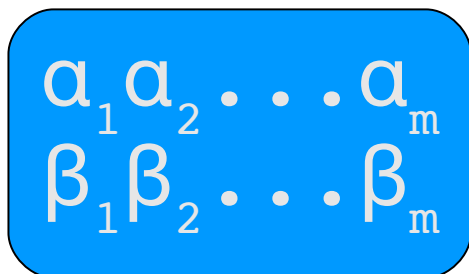
	ORIGINAL AMINO ACID																			
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1 9913		1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1 9822		36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42 9859		0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0 9973		0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0 9876		27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35 9865		4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7 9935		1	0	1	2	1	1	3	21	3	0	0	5
H His	1	2	18	3	1	20	1	0 9912		0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0 9872		9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22 9947		2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1 9926		20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4 9874		1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4 9946		0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1 9926		12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17 9840		38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32 9871		0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0 9976		1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2 9945		1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2 9901	

Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix,  $M_{ij}$ , gives the probability that the amino acid in column  $j$  will be replaced by the amino acid in row  $i$  after a given evolutionary interval, in this case

1 accepted point mutation per 100 amino acids. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

# Κατασκευή ΠΑ ΡΑΜ1

- Έστω οι ακολουθίες  $\alpha = \alpha_1 \alpha_2 \dots \alpha_m$  και  $\beta = \beta_1 \beta_2 \dots \beta_m$
- Η βαθμονόμηση της στοίχισής τους αντιστοιχεί με τον έλεγχο δύο αλληλοαποκλειόμενων υποθέσεων
  - $H_0$ : οι  $\alpha, \beta$  ΔΕΝ ΕΧΟΥΝ εξελικτική σχέση [τυχαία στοίχιση]
  - $H_1$ : οι  $\alpha, \beta$  ΕΧΟΥΝ κοινό πρόγονο [οι  $\alpha, \beta$  εξαρτώνται μέσω της Μαρκοβιανής διαδικασίας]
- Για τη στοίχισή τους με βάση την υπόθεση  $H_0$ :



$$P_{H_0}(\text{στοίχισης}) = \left( \prod_{i=1}^m F_{\alpha_i} \right) * \left( \prod_{i=1}^m F_{\beta_i} \right)$$
$$= \prod_{i=1}^m (F_{\alpha_i} * F_{\beta_i}) \quad , \text{λόγω ανεξαρτησίας}$$



# Κατασκευή ΠΑ PAM1 (pt II)

Με βάση την υπόθεση  $H_1$ :  $P_{H_1}(\text{στοίχισης}) = \prod_{i=1}^m p_{\alpha_i, \beta_i}$

Επιθυμούμε το score να αντικατοπτρίζει την πιθανότητα εξελικτικής σχέσης. Λογική επιλογή του λόγου:

$$\frac{P_{H_1}(\text{στοίχισης})}{P_{H_0}(\text{στοίχισης})} = \frac{\prod_{i=1}^m p_{\alpha_i, \beta_i}}{\prod_{i=1}^m F_{\alpha_i} * F_{\beta_i}} = \prod_{i=1}^m \frac{p_{\alpha_i, \beta_i}}{F_{\alpha_i} * F_{\beta_i}}$$

ή ισοδύναμα

$$\log \frac{P_{H_1}(\text{στοίχισης})}{P_{H_0}(\text{στοίχισης})} = \log \prod_{i=1}^m \frac{p_{\alpha_i, \beta_i}}{F_{\alpha_i} * F_{\beta_i}} = \sum_{i=1}^m \log \frac{p_{\alpha_i, \beta_i}}{F_{\alpha_i} * F_{\beta_i}}$$

# PAM250 ??

```
#
# This matrix was produced by "pam" Version 1.0.6 [28-Jul-93]
#
# PAM 250 substitution matrix, scale = ln(2)/3 = 0.231049
#
# Expected score = -0.844, Entropy = 0.354 bits
#
# Lowest score = -8, Highest score = 17
#
  A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  2 -2  0  0 -2  0  0  1 -1 -1 -2 -1 -1 -3  1  1  1 -6 -3  0  0  0  0 -8
R -2  6  0 -1 -4  1 -1 -3  2 -2 -3  3  0 -4  0  0 -1  2 -4 -2 -1  0 -1 -8
N  0  0  2  2 -4  1  1  0  2 -2 -3  1 -2 -3  0  1  0 -4 -2 -2  2  1  0 -8
D  0 -1  2  4 -5  2  3  1  1 -2 -4  0 -3 -6 -1  0  0 -7 -4 -2  3  3 -1 -8
C -2 -4 -4 -5 12 -5 -5 -3 -3 -2 -6 -5 -5 -4 -3  0 -2 -8  0 -2 -4 -5 -3 -8
Q  0  1  1  2 -5  4  2 -1  3 -2 -2  1 -1 -5  0 -1 -1 -5 -4 -2  1  3 -1 -8
E  0 -1  1  3 -5  2  4  0  1 -2 -3  0 -2 -5 -1  0  0 -7 -4 -2  3  3 -1 -8
G  1 -3  0  1 -3 -1  0  5 -2 -3 -4 -2 -3 -5  0  1  0 -7 -5 -1  0  0 -1 -8
H -1  2  2  1 -3  3  1 -2  6 -2 -2  0 -2 -2  0 -1 -1 -3  0 -2  1  2 -1 -8
I -1 -2 -2 -2 -2 -2 -2 -3 -2  5  2 -2  2  1 -2 -1  0 -5 -1  4 -2 -2 -1 -8
L -2 -3 -3 -4 -6 -2 -3 -4 -2  2  6 -3  4  2 -3 -3 -2 -2 -1  2 -3 -3 -1 -8
K -1  3  1  0 -5  1  0 -2  0 -2 -3  5  0 -5 -1  0  0 -3 -4 -2  1  0 -1 -8
M -1  0 -2 -3 -5 -1 -2 -3 -2  2  4  0  6  0 -2 -2 -1 -4 -2  2 -2 -2 -1 -8
F -3 -4 -3 -6 -4 -5 -5 -5 -2  1  2 -5  0  9 -5 -3 -3  0  7 -1 -4 -5 -2 -8
P  1  0  0 -1 -3  0 -1  0  0 -2 -3 -1 -2 -5  6  1  0 -6 -5 -1 -1  0 -1 -8
S  1  0  1  0  0 -1  0  1 -1 -1 -3  0 -2 -3  1  2  1 -2 -3 -1  0  0  0 -8
T  1 -1  0  0 -2 -1  0  0 -1  0 -2  0 -1 -3  0  1  3 -5 -3  0  0 -1  0 -8
W -6  2 -4 -7 -8 -5 -7 -7 -3 -5 -2 -3 -4  0 -6 -2 -5 17  0 -6 -5 -6 -4 -8
Y -3 -4 -2 -4  0 -4 -4 -5  0 -1 -1 -4 -2  7 -5 -3 -3  0 10 -2 -3 -4 -2 -8
V  0 -2 -2 -2 -2 -2 -2 -1 -2  4  2 -2  2 -1 -1 -1  0 -6 -2  4 -2 -2 -1 -8
B  0 -1  2  3 -4  1  3  0  1 -2 -3  1 -2 -4 -1  0  0 -5 -3 -2  3  2 -1 -8
Z  0  0  1  3 -5  3  3  0  2 -2 -3  0 -2 -5  0  0 -1 -6 -4 -2  2  3 -1 -8
X  0 -1  0 -1 -3 -1 -1 -1 -1 -1 -1 -1 -1 -2 -1  0  0 -4 -2 -1 -1 -1 -1 -8
* -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8  1
```

# Πίνακες BLOSUM

BLOcks SUBstitution Matrices (Henikoff and Henikoff, 1992)

- Δε γίνεται καμιά αναφορά σε εξελικτικό μοντέλο
- Στατιστική ανάλυση συντηρημένων Blocks χωρίς κενά από σχετιζόμενες πρωτεϊνικές οικογένειες
- Υπολογισμός συχνοτήτων στόχων/υποβάθρου
- log-likelihood ratios
- Πατέντα!!!

# BLOSUM62

```
# Matrix made by matblas from blosum62.11j
# * column uses minimum score
# BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
# Blocks Database = /data/blocks_5.0/blocks.dat
# Cluster Percentage: >= 62
# Entropy = 0.6979, Expected = -0.5209
  A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```

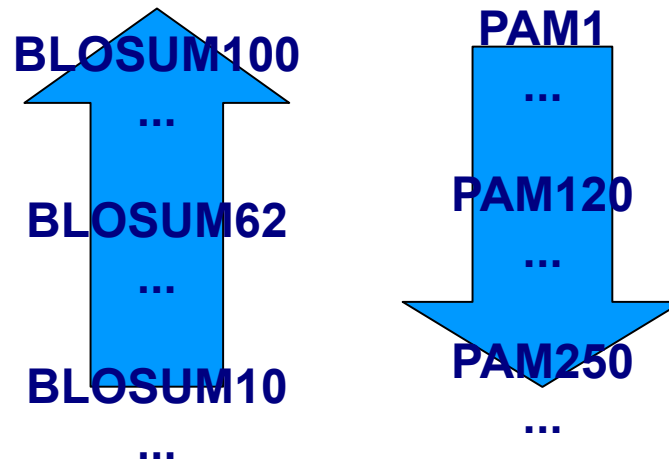
# Τα “νούμερα” των BLOSUM

WYIIR	CASILRKIYIYGPV	GVSRLRTAYGGRK	NRG
WFYVR	CASILRHLYHRSPA	GVGSIITKIYGGRK	RNG
WYYVR	AAAVARHIYLRKTV	GVGRLRKVHGSTK	NRG
WYFIR	AASICRHLYIRSPA	GIGSFEDIYGGRR	RRG
WYYTR	AASIARKIYLRQGI	GVGGFQKIYGGRQ	RNG
WFYKR	AASVARHIYMRKQV	GVGKLNKLYGGAK	SRG
WFYKR	AASVARHIYMRKQV	GVGKLNKLYGGAK	SRG
WYYVR	TASIARRLYVRSPT	GVDALRLVYGGSK	RRG
WYYVR	TASVARRLYIRSPT	GVGALRRVYGGNK	RRG
WFYTR	AASTARHLYLRGGA	GVGSMTKIYGGRQ	RNG
WFYTR	AASTARHLYLRGGA	GVGSMTKIYGGRQ	RNG
WYIVR	AAALLRRVYIDGPV	GVNSLRTHYGGKK	DRG

- Ομαδοποίηση “όμοιων” ακολουθιών
- Κατώφλι ομοιότητας  $x\% \Rightarrow \text{BLOSUM}_x$
- Σταθμισμένη συνεισφορά

# PAM vs BLOSUM

- BLOSUM: σημαντικά μεγαλύτερος όγκος δεδομένων [min F = 2369]
- PAM1 [ακριβής] ~ PAMn [προσεγγιστικά]



# Οι ΠΑ από μια άλλη σκοπιά

$$S_{i,j} = \frac{1}{\lambda} \log \frac{p_{i,j}}{F_i F_j}$$

$p_{i,j}$ : target frequencies

$F_i, F_j$ : background frequencies

Προφανώς, εάν  $p_{i,j} > F_i F_j$  θα ισχύει  $S_{i,j} > 0$

$$\text{Επίσης: } \lambda S_{i,j} = \log \frac{p_{i,j}}{F_i F_j} \text{ ή}$$

$$\frac{p_{i,j}}{F_i F_j} = e^{\lambda S_{i,j}} \text{ οπότε}$$

$$p_{i,j} = F_i F_j e^{\lambda S_{i,j}}$$

# Πληροφοριακό περιεχόμενο ΠΑ

Σχετική εντροπία

$$H = \sum_{i=1}^{20} \sum_{j=1}^{20} p_{i,j} S_{i,j} = \sum_{i=1}^{20} \sum_{j=1}^{20} p_{i,j} \log \frac{p_{i,j}}{F_i F_j}$$



# Συζήτηση ...

**Επιπλέον διδακτικό υλικό:**

<http://troodos.biol.ucy.ac.cy/BRL/courses/BIO331/index.html>