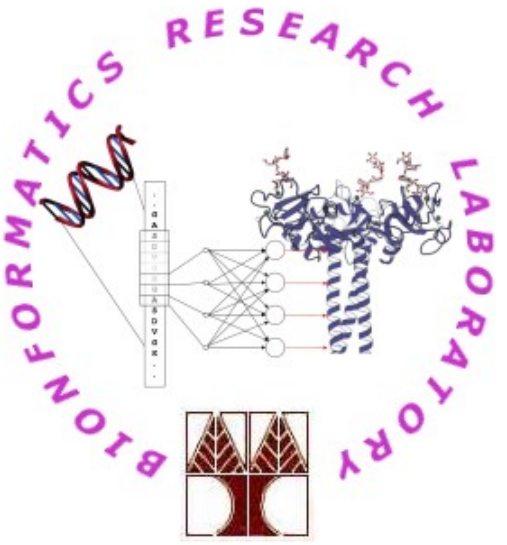
Αλγόριθμοι Εύρεσης Ομοιοτήτων Ακολουθιών
Μέρος ΙΙΙ: Έλεγχος στατιστικής σημαντικότητας

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  GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
                   G+ +VK HGKKV  A ++ +AH+D++    + LS+LH  KL
>P02023|HBB_HUMAN  GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL

β)
>P01922|HBA_HUMAN  GSAQVKGHGKKVADALTNA-----VAHVDDMPNALSALSDLHAHKL
                    + +++ H  KV   +  A     V  V      L  L  +H  K
>P02240|LGB2_LUPLU NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG

γ)
>P01922|HBA_HUMAN  GSAQVKGHGKKVADALT----NAVAHVDDMPNALSALSD----LHAHKL
                         G G  V D+LT        H  D+  A +AL D      AH+
>P91253|GTS7_CAEEL ------GSGYLVGDSLTFVDLLVAQHTADLLAANAALLDEFPQFKAHQE
```

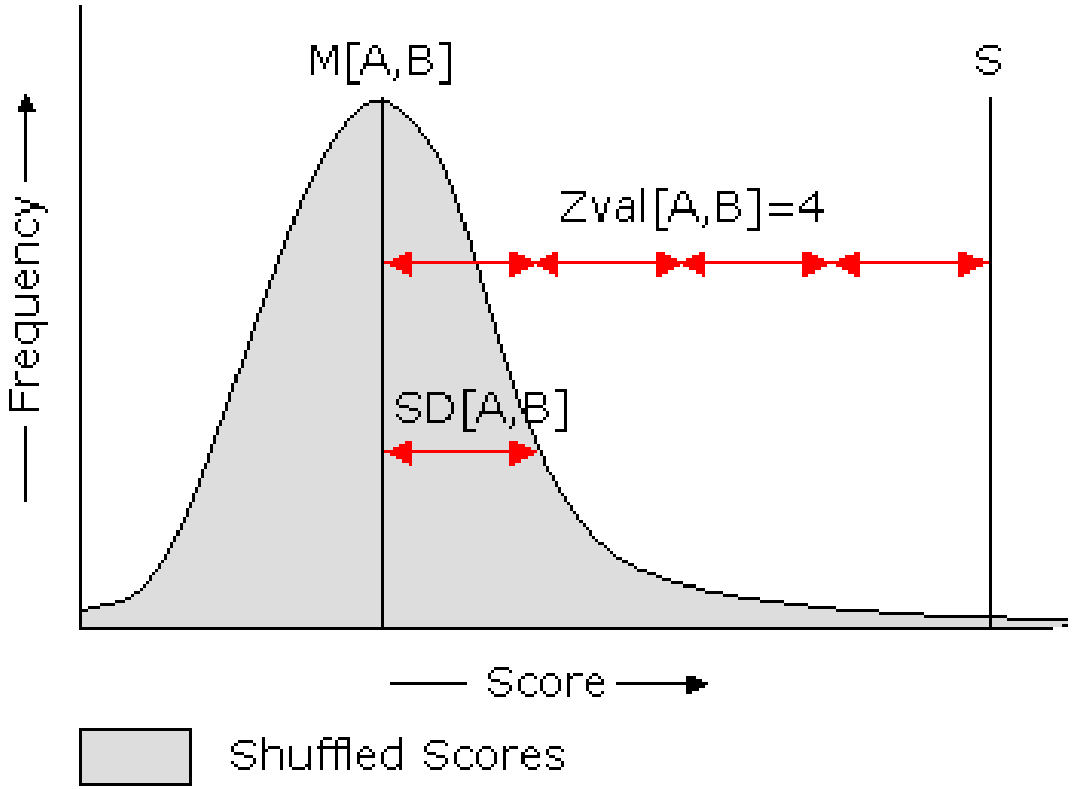
Score
101
17
32

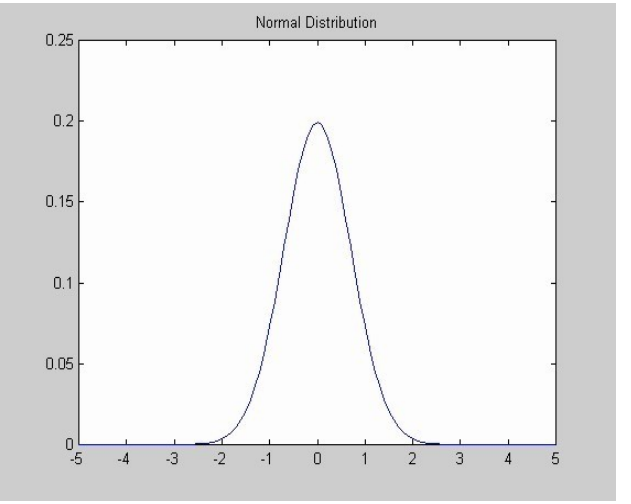

Μετάβαση
Related Structures
Sequences producing significant alignments:
Score (bits)
gi|17647349|ref|NP_523840.1| CG12240-PA, isoform A [Drosoph... 407
gi|40215753|gb|AAL48038.2| LP04971p [Drosophila melanogaste... 380
gi|24762579|ref|NP_611892.1| CG12240-PB, isoform B [Drosoph... 371
gi|58392349|ref|XP_319298.2| ENSANGP00000012354 [Anopheles ... 102
gi|34785543|gb|AAH57849.1| DNAJC4 protein [Homo sapiens] >g... 64
gi|29835195|gb|AAH51008.1| DNAJC4 protein [Homo sapiens] 64
gi|33877996|gb|AAH32137.1| DNAJC4 protein [Homo sapiens] 64
gi|34894012|ref|NP_908331.1| putative DnaJ-like protein [Or... 63
gi|55636307|ref|XP_522047.1| PREDICTED: similar to Vascular... 63
gi|51261928|gb|AAH79955.1| Dnajc4-prov protein [Xenopus tro... 62
gi|18447132|gb|AAL68157.1| AT30646p [Drosophila melanogaster] 62
gi|7299733|gb|AAF54914.1| CG8476-PA [Drosophila melanogaste... 62
gi|28950166|emb|CAD71034.1| hypothetical protein [Neurospor... 62
gi|53765858|ref|ZP_00186718.2| COG0484: DnaJ-class molecula... 62
gi|61859440|ref|XP_587907.1| PREDICTED: similar to DnaJ hom... 61
gi|61843081|ref|XP_613027.1| PREDICTED: similar to DnaJ hom... 61
gi|56755781|gb|AAW26069.1| unknown [Schistosoma japonicum] 61
gi|54032310|ref|ZP_00364442.1| COG2214: DnaJ-class molecula... 60
gi|31541124|gb|AAP56426.1| DnaJ [Mycoplasma gallisepticum R... 59
gi|57099699|ref|XP_533246.1| PREDICTED: similar to hypothet... 59
gi|854466|emb|CAA89929.1| unknown [Saccharomyces cerevisiae... 59
gi|37362683|ref|NP_013941.2| One of several homologs of bac... 59
gi|16800577|ref|NP_470845.1| heat shock protein DnaJ [Liste... 59
gi|16803512|ref|NP_464997.1| heat shock protein DnaJ [Liste... 59
gi|49651399|emb|CAG78338.1| unnamed protein product [Yarrow... 59
gi|57088065|ref|XP_536990.1| PREDICTED: similar to DnaJ hom... 58
gi|4007007|emb|CAA66720.1| l(2)tid [Drosophila melanogaster... 58
gi|46907700|ref|YP_014089.1| chaperone protein DnaJ [Lister... 58
gi|42527589|ref|NP_972687.1| DnaJ domain protein [Treponema... 58
gi|62897771|dbj|BAD96825.1| DnaJ (Hsp40) homolog, subfamily... 58
gi|55643335|ref|XP_510781.1| PREDICTED: DnaJ (Hsp40) homolo... 58
gi|50912997|ref|XP_467906.1| DNAJ heat shock N-terminal dom... 58
gi|61363502|gb|AAX42402.1| DnaJ-like subfamily A member 3 [... 58
gi|28950130|emb|CAD70988.1| related to SCJ1 protein [Neuros... 58

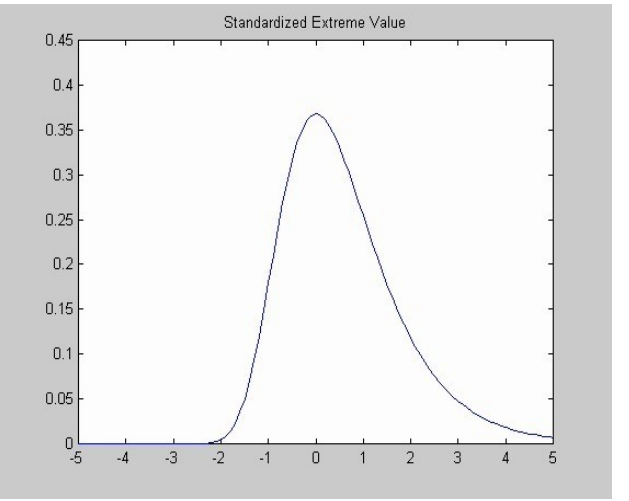
Στατιστική Σημαντικότητα
(Τυπική Προσέγγιση)

- 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: $Z = (s-s_{mean})/sd$
    - $Z < 3$ SD => improbable
    - 3 SD $< Z <$ 5 SD => marginal
    - 5 SD $< Z <$ 10 SD => probable
    - $Z >$ 10 SD => certain (???)

Αλλά είναι «προσέγγιση»

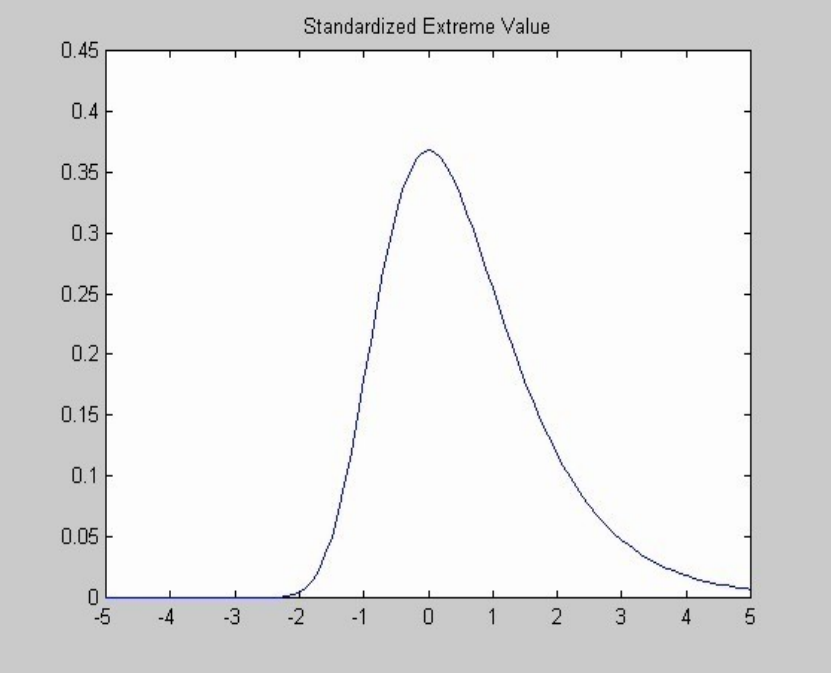






Η κατανομή των Ακραίων Τιμών
(Gumbel, 1958)

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



- 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”: $A=\{\alpha_1,\alpha_2,..,\alpha_r\}$
  - Πρωτεΐνες: r=20
  - DNA/RNA: r=4
- και μια “τυχαία” αλληλουχία
  - $S=s_1s_2...s_m$
- η οποία έχει προκύψει με τυχαία “δειγματοληψία” του A με κατανομή πιθανοτήτων $P=\{p_1,p_2,..,p_r\}$

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

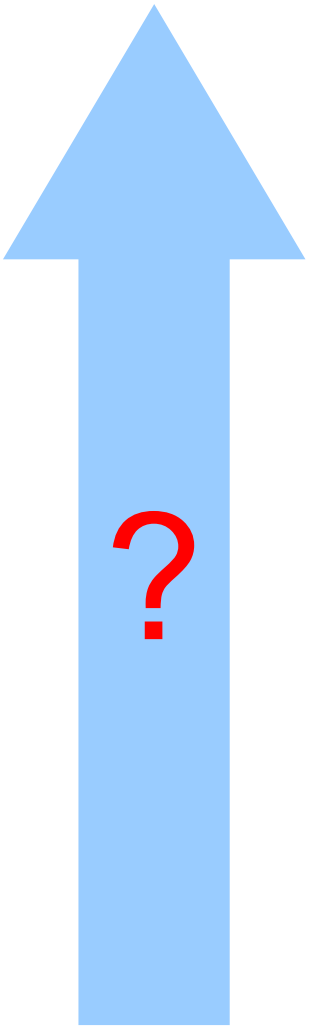
$$E(S \geq s) = Kmne^{-\lambda s}$$

- n: μήκος βάσης δεδομένων
- mn: μέγεθος χώρου αναζήτησης
- λs: κανονικοποιημένο score
- K: σταθερά
- K,λ εξαρτώνται από το σύστημα βαθμονόμησης
    - ΤΙ ΕΚΦΡΑΖΕΙ ΤΟ ***E***?

Μετάβαση

Related Structures

Sequences producing significant alignments:	Score (bits)	E Value	
gi|17647349|ref|NP_523840.1| CG12240-PA, isoform A [Drosoph...	407	e-112	G
gi|40215753|gb|AAL48038.2| LP04971p [Drosophila melanogaste...	380	e-104	
gi|24762579|ref|NP_611892.1| CG12240-PB, isoform B [Drosoph...	371	e-102	G
gi|58392349|ref|XP_319298.2| ENSANGP00000012354 [Anopheles ...	102	6e-21	G
gi|34785543|gb|AAH57849.1| DNAJC4 protein [Homo sapiens] >g...	64	2e-09	G
gi|29835195|gb|AAH51008.1| DNAJC4 protein [Homo sapiens]	64	2e-09	G
gi|33877996|gb|AAH32137.1| DNAJC4 protein [Homo sapiens]	64	2e-09	G
gi|34894012|ref|NP_908331.1| putative DnaJ-like protein [Or...	63	6e-09	G
gi|55636307|ref|XP_522047.1| PREDICTED: similar to Vascular...	63	6e-09	G
gi|51261928|gb|AAH79955.1| Dnajc4-prov protein [Xenopus tro...	62	8e-09	G
gi|18447132|gb|AAL68157.1| AT30646p [Drosophila melanogaster]	62	8e-09	
gi|7299733|gb|AAF54914.1| CG8476-PA [Drosophila melanogaste...	62	8e-09	G
gi|28950166|emb|CAD71034.1| hypothetical protein [Neurospor...	62	1e-08	G
gi|53765858|ref|ZP_00186718.2| COG0484: DnaJ-class molecula...	62	1e-08	
gi|61859440|ref|XP_587907.1| PREDICTED: similar to DnaJ hom...	61	2e-08	G
gi|61843081|ref|XP_613027.1| PREDICTED: similar to DnaJ hom...	61	2e-08	G
gi|56755781|gb|AAW26069.1| unknown [Schistosoma japonicum]	61	2e-08	
gi|54032310|ref|ZP_00364442.1| COG2214: DnaJ-class molecula...	60	5e-08	
gi|31541124|gb|AAP56426.1| DnaJ [Mycoplasma gallisepticum R...	59	7e-08	G
gi|57099699|ref|XP_533246.1| PREDICTED: similar to hypothet...	59	7e-08	G
gi|854466|emb|CAA89929.1| unknown [Saccharomyces cerevisiae...	59	9e-08	
gi|37362683|ref|NP_013941.2| One of several homologs of bac...	59	9e-08	G
gi|16800577|ref|NP_470845.1| heat shock protein DnaJ [Liste...	59	1e-07	G
gi|16803512|ref|NP_464997.1| heat shock protein DnaJ [Liste...	59	1e-07	G
gi|49651399|emb|CAG78338.1| unnamed protein product [Yarrow...	59	1e-07	G
gi|57088065|ref|XP_536990.1| PREDICTED: similar to DnaJ hom...	58	2e-07	G
gi|4007007|emb|CAA66720.1| l(2)tid [Drosophila melanogaster...	58	2e-07	
gi|46907700|ref|YP_014089.1| chaperone protein DnaJ [Lister...	58	2e-07	G
gi|42527589|ref|NP_972687.1| DnaJ domain protein [Treponema...	58	2e-07	G
gi|62897771|dbj|BAD96825.1| DnaJ (Hsp40) homolog, subfamily...	58	2e-07	G
gi|55643335|ref|XP_510781.1| PREDICTED: DnaJ (Hsp40) homolo...	58	2e-07	G
gi|50912997|ref|XP_467906.1| DNAJ heat shock N-terminal dom...	58	2e-07	G
gi|61363502|gb|AAX42402.1| DnaJ-like subfamily A member 3 [...	58	2e-07	
gi|28950130|emb|CAD70988.1| related to SCJ1 protein [Neuros...	58	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~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, λ
- Εξάρτηση από τιμές ποινής κενών
- Εμπειρικός υπολογισμός
  - Στοιχίσεις τυχαίων ακολουθιών
  - Καταγραφή χαρακτηριστικών HSPs (scores, συχνότητες υποβάθρου, μήκη)
  - Υπολογισμός K, λ με βάση το “πλησιέστερο” σύστημα βαθμονόμησης (χωρίς κενά)
- Προσομοιώσεις έδειξαν ότι η προσέγγιση δεν είναι και άσχημη (~EVD)

Στοιχίσεις με κενά (μέρος Β)

Gap open	Gap extend	λ	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
      - $m'n'=(m-l)(n-DBsize*l)$

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 (???)

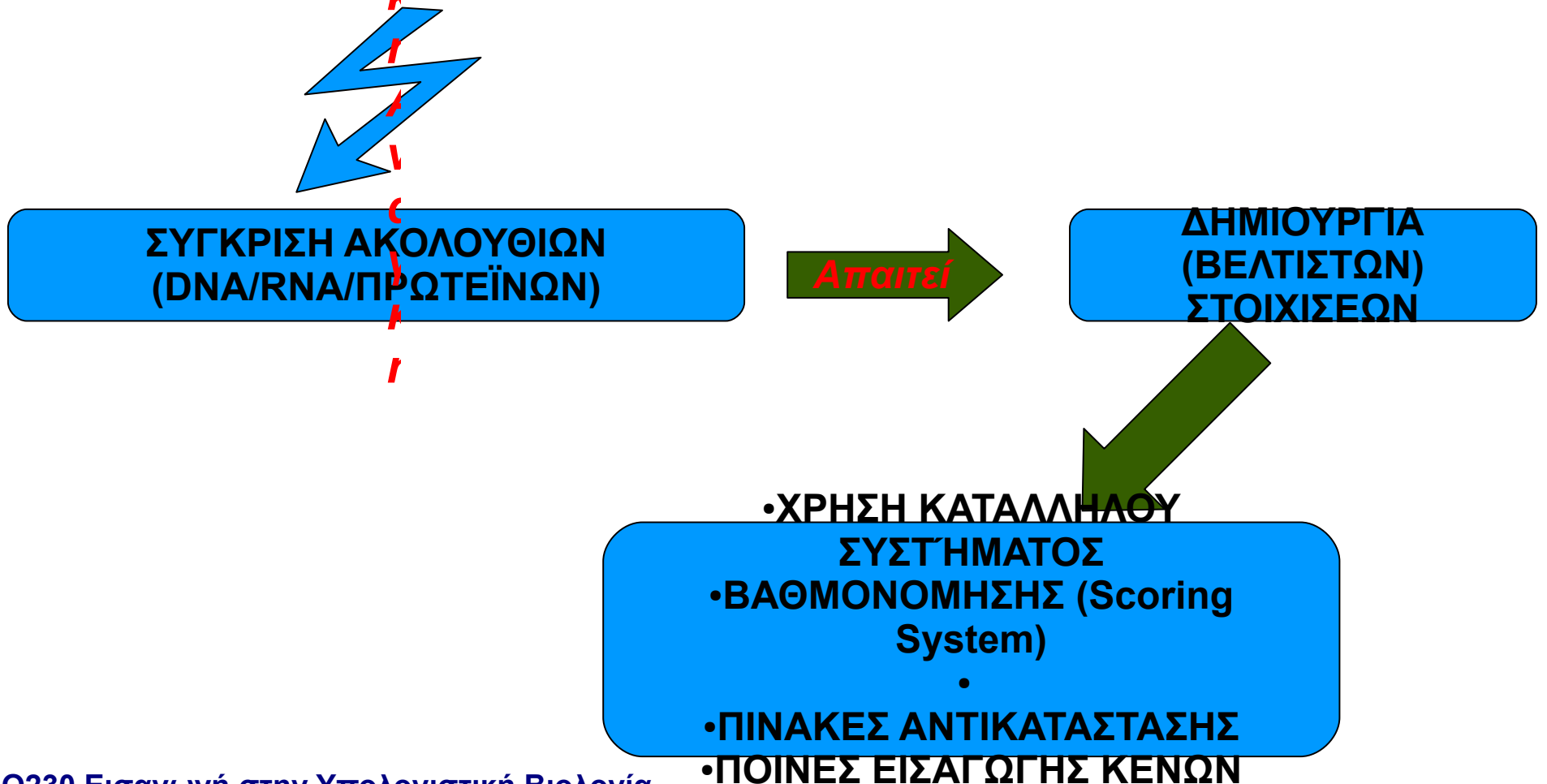
Αξιολόγηση με βάση E-values

Τύπος Ακολουθίας	**E-value**	**Ταυτότητα Καταλοίπων**
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 βαθμονόμησης



Χρήση “αθροιστικών” συστημάτων

- Για κάθε θέση σε μιά 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 συγκεκριμμένο εξελικτικό μοντέλο [??]
- Για Αμινοξικές Ακολουθίες:
  - Ορισμένες μεταλλάξεις αναμένεται να είναι λιγότερο πιθανές [K=>R, K=>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)

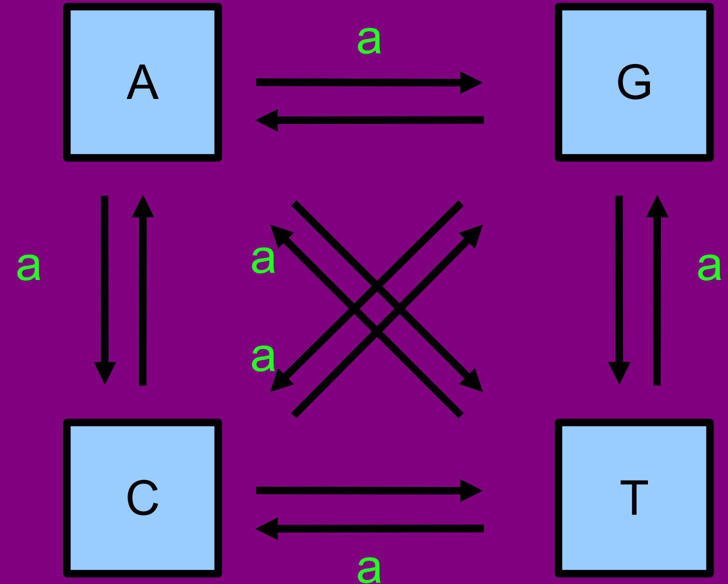
ΚΕΝΤΡΙΚΗ ΙΔΕΑ PAM (Point Accepted Mutations)

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??

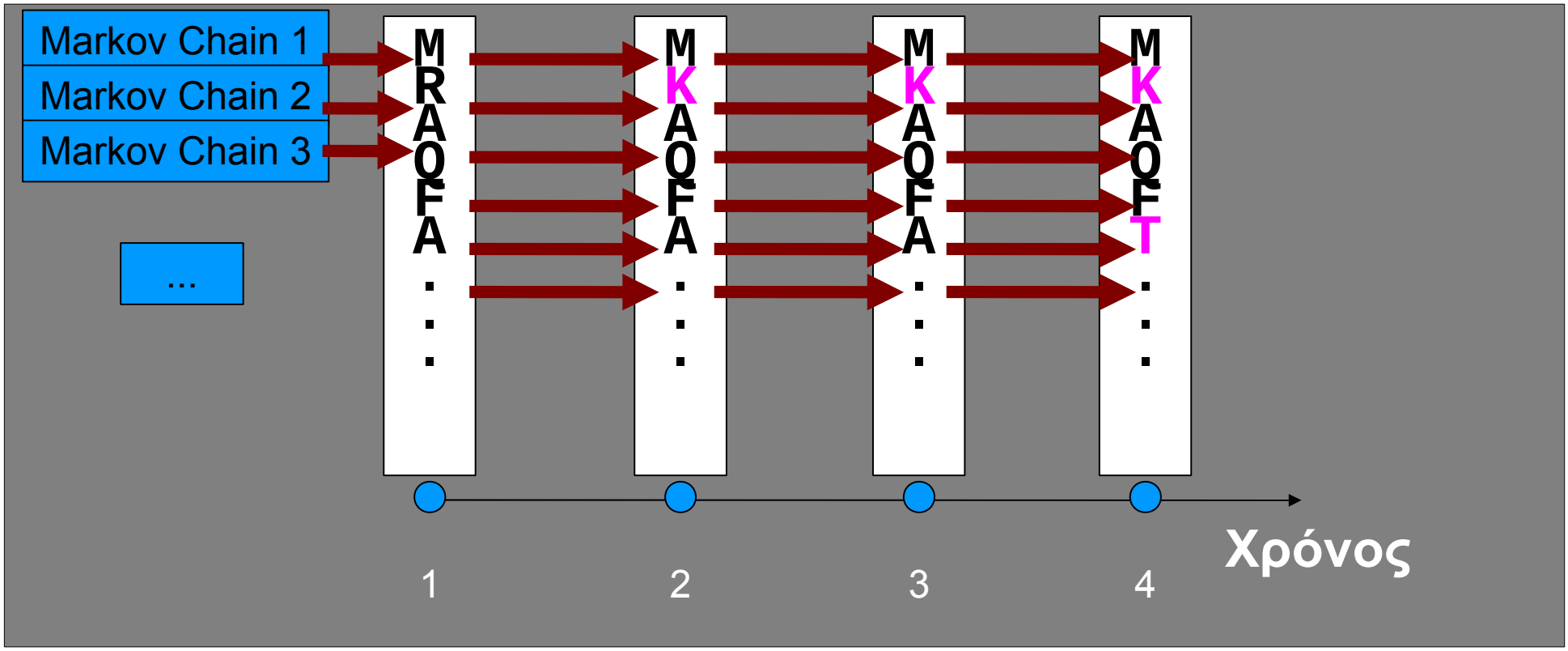
- Ναιιιιιιιιιιιιιιιιιιιιιι!



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

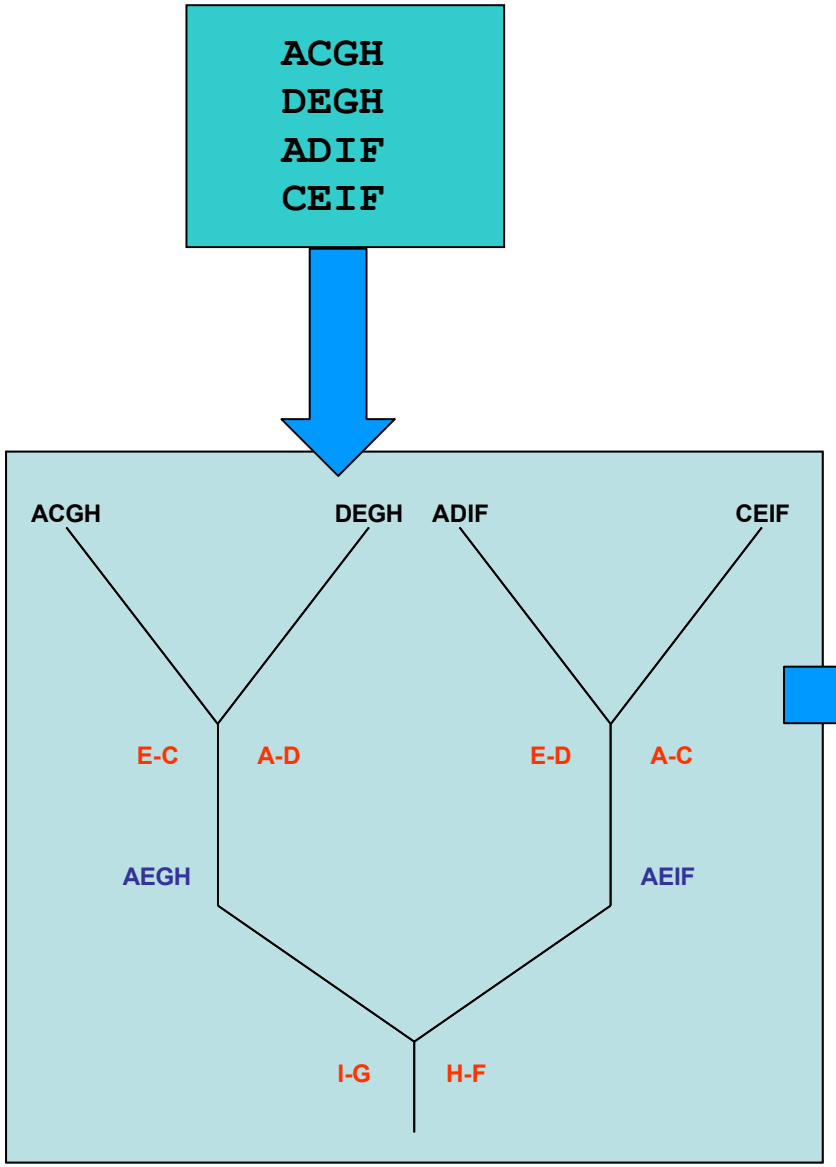
Ο Πίνακας μεταβάσεων PAM1

- 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)

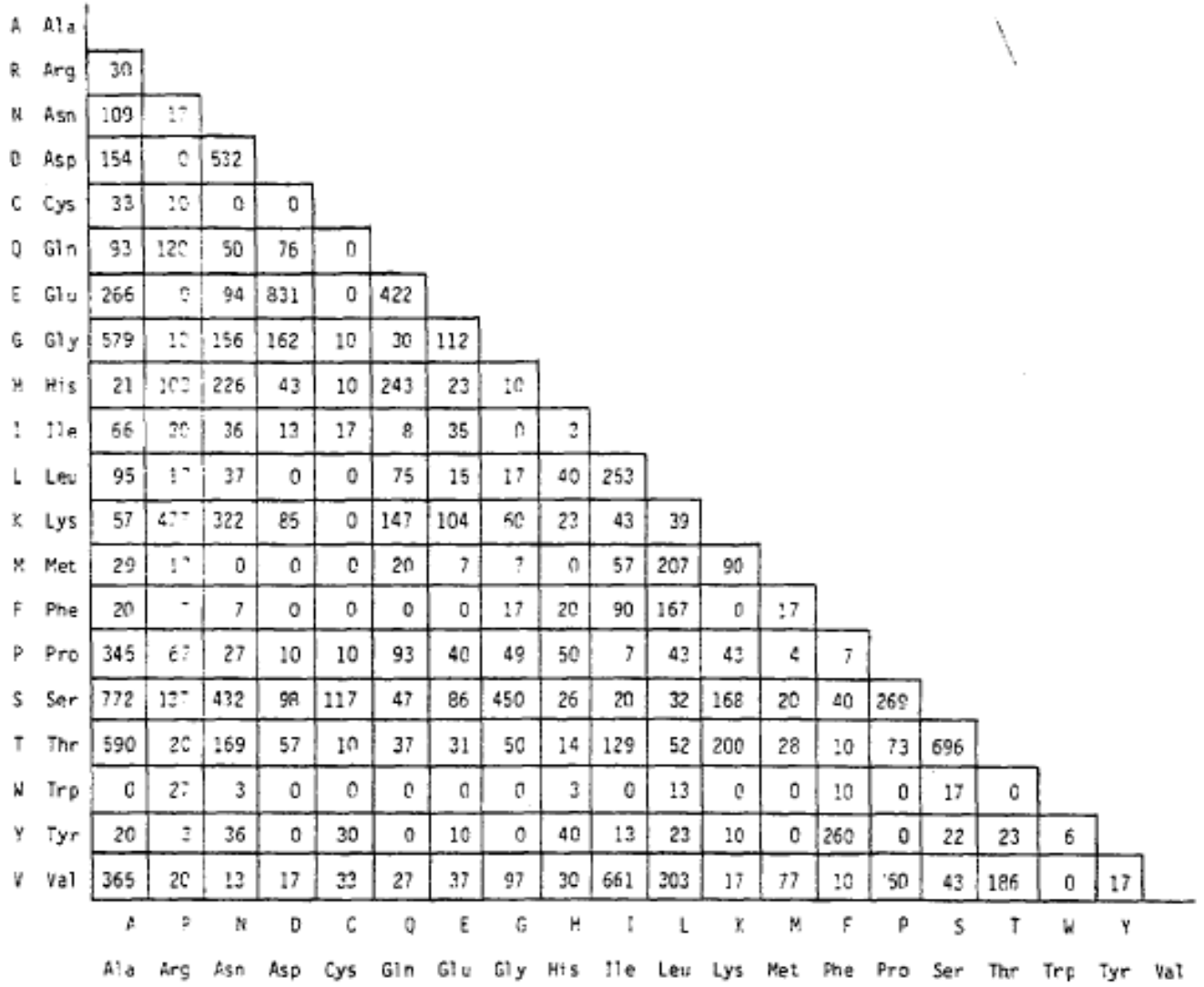
- Πολλαπλές στοιχίσεις χωρίς κενά
- 71 πρωτεϊνικές οικογένειες
- Επιλογή καλά διατηρημένων περιοχών
- (Μόνο) 1572 αντικαταστάσεις
- Δημιουργία φυλογενετικών δέντρων => Maximum Parsimony
- Ανακατασκευή “αρχαίων” ακολουθιών και απαρίθμηση

Ο πίνακας απαρίθμησης



	A	C	D	E	F	G	H	I
A	0	1	1	0	0	0	0	0
C	1	0	0	1	0	0	0	0
D	1	0	0	1	0	0	0	0
E	0	1	1	0	0	0	0	0
F	0	0	0	0	0	0	1	0
G	0	0	0	0	0	0	0	1
H	0	0	0	0	1	0	0	0
I	0	0	0	0	0	1	0	0

... και το Original ...



		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	Ala																				
R	Arg	30																			
N	Asn	109	17																		
D	Asp	154	0	532																	
C	Cys	33	10	0	0																
Q	Gln	93	120	50	76	0															
E	Glu	266	0	94	831	0	422														
G	Gly	579	10	156	162	10	30	112													
H	His	21	100	226	43	10	243	23	10												
I	Ile	66	30	36	13	17	8	35	0	3											
L	Leu	95	17	37	0	0	75	15	17	40	253										
K	Lys	57	427	322	85	0	147	104	60	23	43	39									
M	Met	29	17	0	0	0	20	7	7	0	57	207	90								
F	Phe	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	Pro	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	Ser	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	Thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	Trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	Tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	Val	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	

Figure 80. Numbers of accepted point mutations (X 10) accumulated from closely related sequences. Fifteen hundred and seventy-two exchanges are shown. Fractional exchanges result when ancestral sequences are ambiguous.

Σχετική μεταλλαξιμότητα και πίνακας μεταβάσεων

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

$$m_X = \frac{\sum_Y F_{X,Y}}{N_X}$$

Πίνακας Μεταβάσεων

$$\mathbf{M}_{X,Y} = \begin{cases} \dfrac{\lambda m_Y F_{X,Y}}{\sum_X F_{X,Y}}, & \text{αν } X \neq Y \\ 1 - \lambda m_Y, & \text{αν } X = Y \end{cases}$$

... και το Λάμδα?

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

Υπολογισμός λ 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

$$P = P(A_1 \neq A_0) = \sum_{j=1}^{20} P(A_0 = j, A_1 \neq j)$$

$$= \sum_{j=1}^{20} P(A_1 \neq j | A_0 = j) * P(A_0 = j)$$

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

όπου F_j η παρατηρούμενη συχνότητα του καταλοίπου τύπου j

Υπολογισμός λ για 1-PAM (pt II)

$$0.01 = \sum_{j=1}^{20} P(A_1 \neq j | A_0 = j) * F_j$$

$$= \sum_{j=1}^{20} \left(\sum_{k \neq j} P(A_1 = k | A_0 = j) \right) * F_j$$

$$\approx \sum_{j=1}^{20} \left(\sum_{k \neq j} p_{j,k} \right) * F_j$$

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

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

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

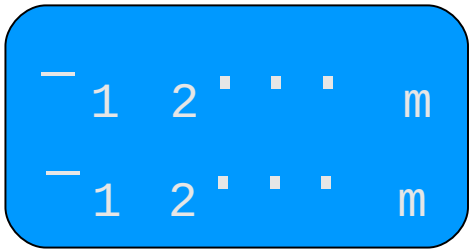
ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V 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	8	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 PAM1

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



$$P_{H_0}(\sigma\tau o\acute{\iota}\chi\iota\sigma\eta\varsigma)=(\prod_{i=1}^{m} F_{\alpha_i})*(\prod_{i=1}^{m} F_{\beta_i})$$

$$=\prod_{i=1}^{m}(F_{\alpha_i}*F_{\beta_i}) \quad ,\lambda\acute{o}\gamma\omega\ \alpha\nu\varepsilon\xi\alpha\rho\tau\eta\sigma\acute{\iota}\alpha\varsigma$$

Κατασκευή ΠΑ PAM1 (pt II)

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

- 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 -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.iij
#  * 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
```

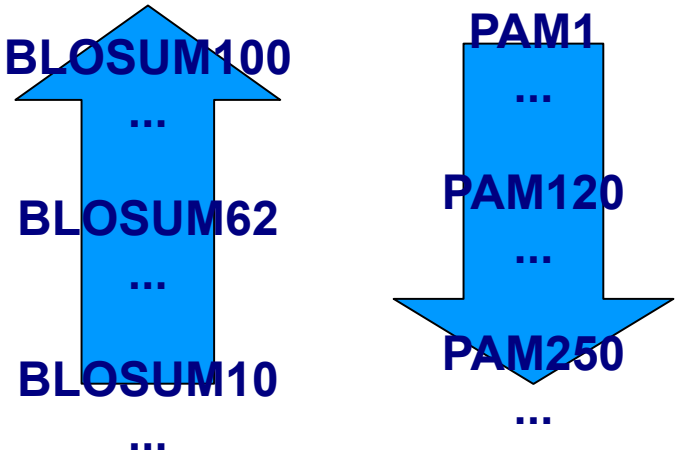
Τα “νούμερα” των BLOSUM

```
WWYIR  CASILRKIYIYGPV  GVSRLRTAYGGRK  NRG
WFYVR  CASILRHLYHRSPA  GVGSITKIYGGRK  RNG
WYYVR  AAAVARHIYLRKTV  GVGRLRKVHGSTK  NRG
WYFIR  AASICRHLYIRSPA  GIGSFEKIYGGRR  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
WWYVR  AAALLRRVYIDGPV  GVNSLRTHYGGKK  DRG
```

- Ομαδοποίηση “όμοιων” ακολουθιών
- Κατώφλι ομοιότητας x % => BLOSUMx
- Σταθμισμένη συνεισφορά

PAM vs BLOSUM

- BLOSUM: σημαντικά μεγαλύτερος όγκος δεδομένων [min F = 2369]
- PAM1 [ακριβής] ~ PAMn [προσεγγιστικά]
- BLOSUM1-BLOSUM100 [ακριβής]



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* $S_{i,j} > 0$

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

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

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

Πληροφοριακό περιεχόμενο ΠΑ

- Σχετική εντροπία

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

Συζήτηση ...

Επιπλέον διδακτικό υλικό:

http://troodos.biol.ucy.ac.cy/