Αλγόριθμοι Εύρεσης Ομοιοτήτων Ακολουθιών
Μέρος ΙΙΙ: Έλεγχος στατιστικής σημαντικότητας

Βασίλης Προμπονάς, PhD
Ερευνητικό Εργαστήριο Βιοπληροφορικής

Τμήμα Βιολογικών Επιστημών
Νέα Παν/πολη, Γραφείο Β161
Πανεπιστήμιο Κύπρου
Ταχ.Κιβ. 20537
1678, Λευκωσία
ΚΥΠΡΟΣ

τηλ: 00357-22892879 (εσωτ. 2879)
email: vprobon@ucy.ac.cy, vprobon@biol.uoa.gr

Σύνοψη

- Το πρόβλημα ...
- Απλοϊκή προσέγγιση
- Στατιστική προσέγγιση – η εξίσωση Karlin-Altschul
- Συζήτηση
  - ...

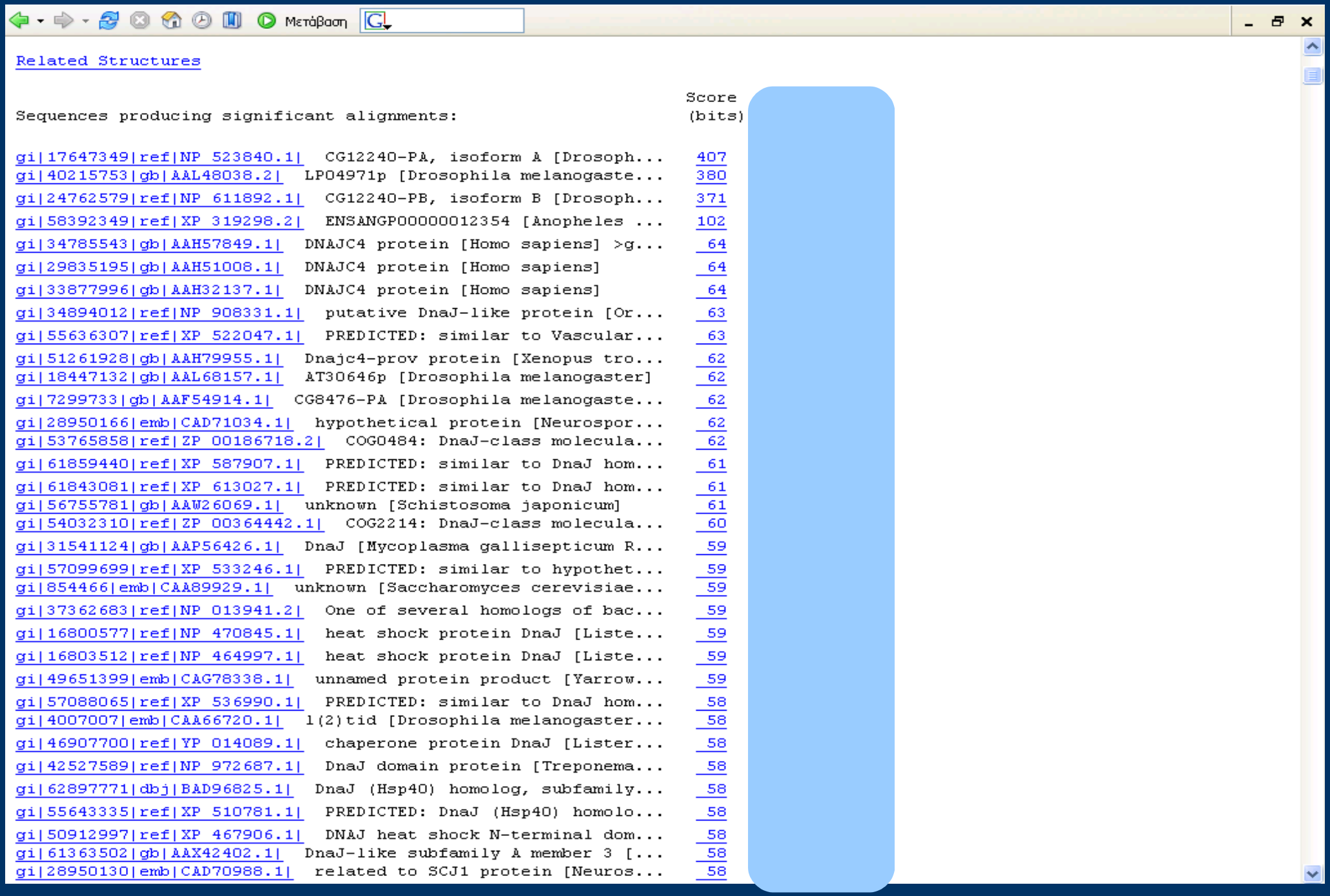
Το πρόβλημα ...

```
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
```

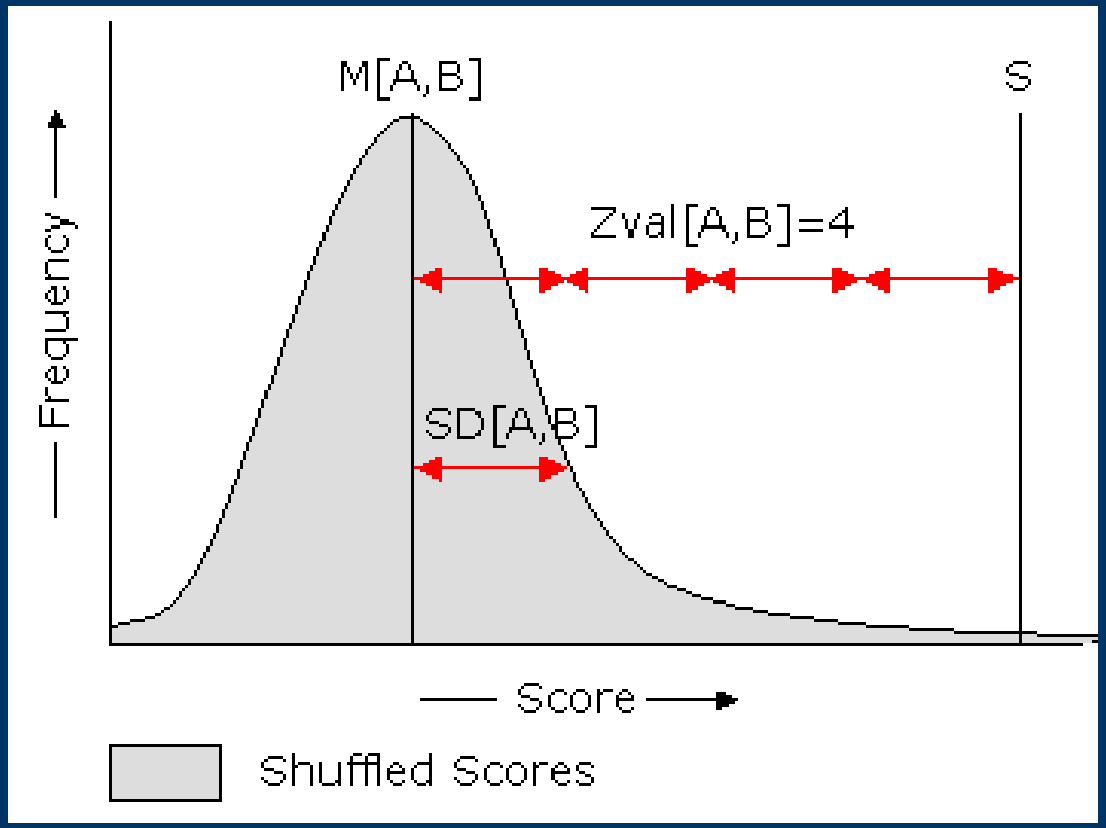
Score
101
17
32

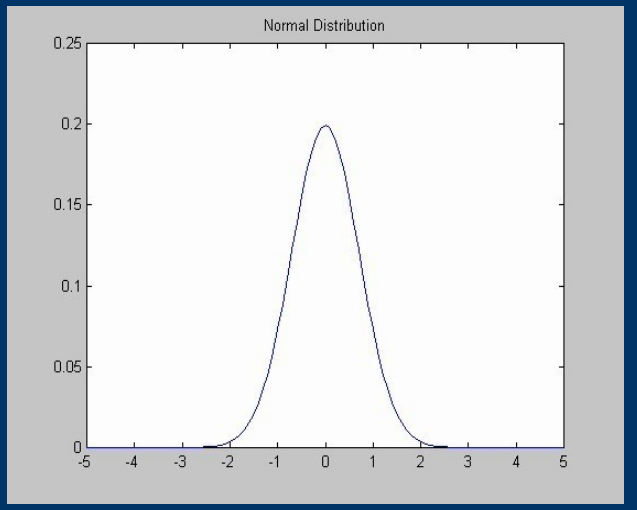

Μετάβαση
Related Structures
Sequences producing significant alignments:
Score (bits)
gi|17647349|ref|NP_523840.1| CG12240-PA, isoform A [Drosoph... 407
gi|40215753|gb|AAL48038.2| LP04971p [Drosophila melanogaste... 380
gi|24762579|ref|NP_611892.1| CG12240-PB, isoform B [Drosoph... 371
gi|58392349|ref|XP_319298.2| 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

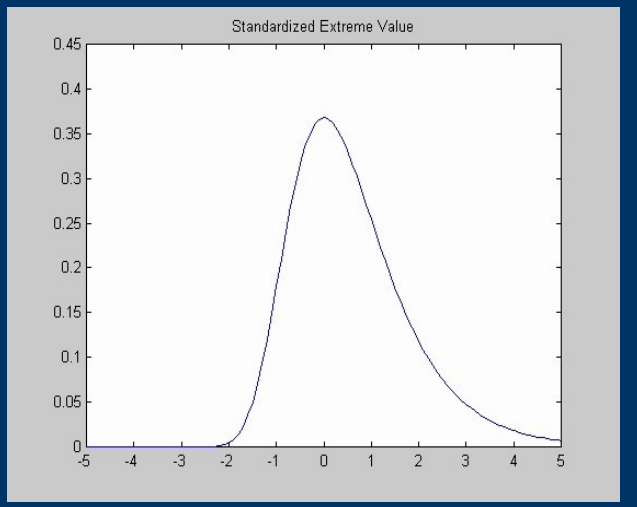
Στατιστική Σημαντικότητα (Τυπική Προσέγγιση)

- 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»

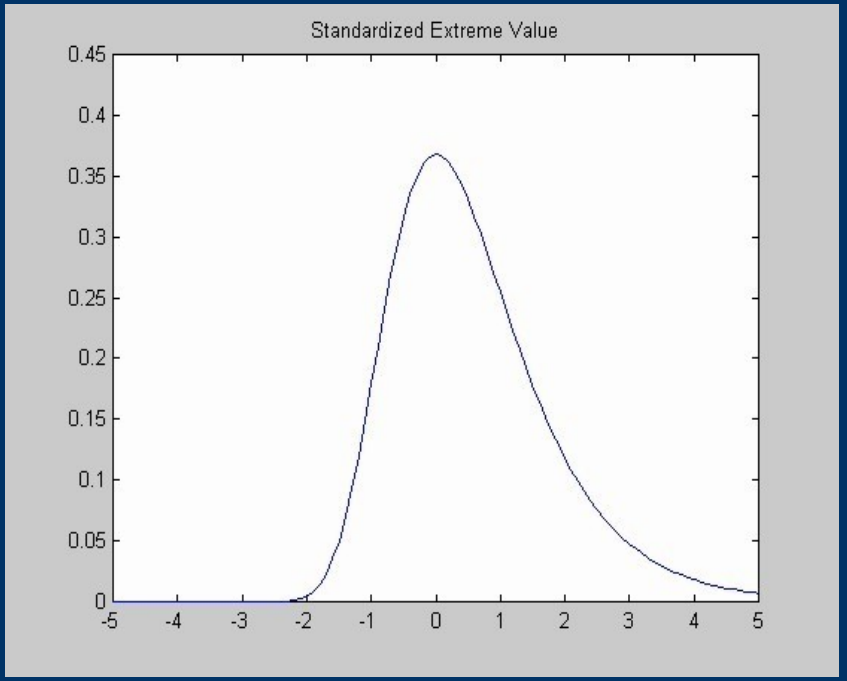






Η κατανομή των Ακραίων Τιμών
(Gumbel, 1958)

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



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

$S=s_1 s_2 \ldots s_m$

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)

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

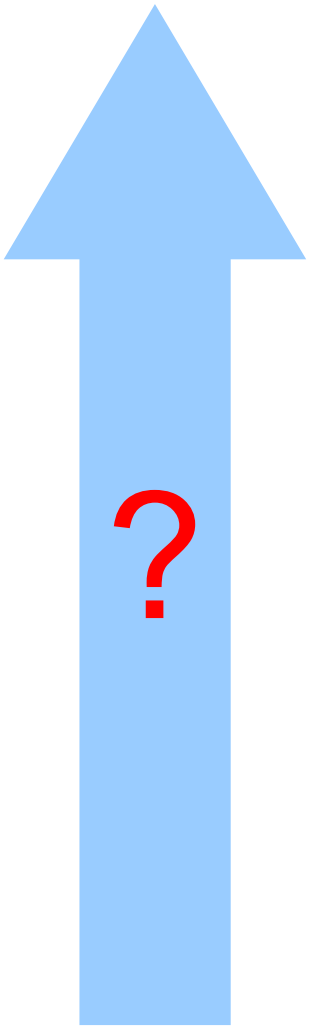
- n: μήκος βάσης δεδομένων
- mn: μέγεθος χώρου αναζήτησης
- λs: κανονικοποιημένο score
- K: σταθερά
- K,λ εξαρτώνται από το σύστημα βαθμονόμησης

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 (???)

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

Συζήτηση

- ...