Ειδικά Θέματα Βιοπληροφορικής

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας
Λαμία, 2015

ΠΡΟΕΠΙΣΚΟΠΙΣΗ ΜΑΘΗΜΑΤΟΣ

- Εισαγωγή
- Στατιστικές τεχνικές μελέτης των ακολουθιών
- Ομοιότητα ακολουθιών (αλγόριθμοι στόιχισης, στατιστ. σημαντικότητα, ευριστικοί αλγόριθμοι)
- Πολλαπλή στοίχιση/ πρότυπα/ Φυλογενετικές σχέσεις
- Markov Chains – Hidden Markov Models
- Profiles, Profile Hidden Markov Models, Transformational Grammars
- Neural Networks, Support Vector Machines
- Μέθοδοι πρόγνωσης σε ακολουθίες πρωτεϊνών και RNA/DNA
- Δομική Βιοπληφορική
- Ανάλυση δεδομένων γονιδιακής έκφρασης (microarrays)
- Εφαρμογές των τεχνικών σε πραγματικά προβλήματα
- Programming in Perl

Τρόποι μελέτης των ακολουθιών

- Global information

  Η ακολουθία αναπαρίσταται από ένα διάνυσμα σταθερού μήκους (π.χ. τα ποσοστά εμφάνισης των αμινοξέων)

- Local information

  Η ακολουθία αναπαρίσταται διαδοχικά επικαλυπτόμενα «παράθυρα» σταθερού μήκους

Μοντέλο της ανεξαρτησίας

- Έστω μια ακολουθία $\mathbf{x}=x_1,x_2,x_3\ldots,x_n$
- Θεωρούμε τα x_i ανεξάρτητα μεταξύ τους ενδεχόμενα

$$\mathbf{x} = x_1, x_2, ..., x_n \quad \text{με} \quad x_i \in \{A, T, G, C\}$$

$$p_A, p_T, p_G, p_C \ \text{με} \ p_k \geq 0 \ \text{και} \ \sum_{k \in \{A,T,G,C\}} p_k = 1$$

$$p_{o\lambda} = P(\mathbf{x}) = \prod_{i=1}^{n} p_{X_i}$$

Εντροπία

- Μια δεδομένη ακολουθία DNA, όπως την ορίσαμε παραπάνω, λέμε ότι έχει συνάρτηση εντροπίας κατά Shannon ίση με

$$H(\mathbf{x}) = -\sum_i P(x_i) \log P(x_i)$$

- Η εντροπία γίνεται μέγιστη όταν οι βάσεις είναι ισοπίθανες, δηλαδή όταν $pA=pG=pT=PC=1/4$ οπότε θα έχει τιμή ίση με $H(\mathbf{x})=\sum(1/4)\log(1/4)=\log 4$. Συνήθως σε αυτές τις περιπτώσεις παίρνουμε λογάριθμους με βάση το 2, έτσι ώστε η μονάδα μέτρησης να είναι το bit. Η πληροφορία μιας ακολουθίας ορίζεται ως:

$$I(\mathbf{x}) = H_{\max} - H_{obs}$$

Πολυπλοκότητα

- ορίζεται, για ένα παράθυρο μήκους *k* της ακολουθίας, ως εξής:

$$K = \frac{1}{k} \log_{N_\Omega} \left(\frac{k!}{\prod_{\forall s \in \Omega} n_s !} \right)$$

- Στην παραπάνω σχέση, το *ns* είναι ο αριθμός εμφανίσεων του συμβόλου *s* στο παράθυρο και *N*Ω το μέγεθος του αλφάβητου (4 για τα νουκλεοτίδια, 20 για τα αμινοξέα). Διαισθητικά, το μέτρο αυτό δείχνει την ποσότητα της πληροφορίας που απαιτείται σε κάθε θέση της ακολουθίας για να καθορίσει κανείς το σύμβολο (της θέσης), δεδομένης της σύνθεσης όλου του παραθύρου. Για παράδειγμα, ένα παράθυρο 4 νουκλεοτιδίων με σύσταση AAAA, θα έχει πολυπλοκότητα ίση με

$$K = \frac{1}{4} \log_4 \left(\frac{4!}{4!0!0!0!} \right) = \frac{1}{4} \log_4 (1) = 0$$

Σχετική εντροπία

- Μια άλλη σχετική έννοια, είναι αυτή της σχετικής εντροπίας (Relative Entropy). Η σχετική εντροπία δυο καταστάσεων *P*, *Q* (γνωστή και ως μέτρο της απόστασης των Kullback-Leibler) εκφράζει τη σχετική απόσταση, ή διαφορά, μεταξύ των δυο καταστάσεων και δίνεται από τον τύπο

$$H(P,Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)}$$

- Το *P*(*xi*) είναι όπως είδαμε παραπάνω η πιθανότητα εμφάνισης μιας βάσης (A,T,G,C) στην *i* θέση της συγκεκριμένης ακολουθίας, ενώ το *Q*(*xi*) η αντίστοιχη πιθανότητα εμφάνισης μιας βάσης σε μια άλλη ακολουθία. Αυτή η άλλη ακολουθία μπορεί να είναι μια άλλη πραγματική ακολουθία με την οποία θέλουμε να συγκρίνουμε την πρώτη, ή να είναι μια θεωρητική κατανομή, όπως αυτή που υποθέτει ισοπίθανη ή τυχαία εμφάνιση των βάσεων. Προφανώς αν *Q*(*xi*)=1/4 (ισοκατανομή των βάσεων) τότε *H(P,Q)=I(P)*

Αμοιβαία πληροφορία

- Μια άλλη πολύ σημαντική έννοια που θα ξανασυναντήσουμε και στα επόμενα κεφάλαια είναι αυτή της αμοιβαίας πληροφορίας (Mutual Information). Δυο τ.μ *X,Y* έχουν αμοιβαία πληροφορία που δίνεται από τη σχέση:

$$M(X,Y) = \sum_{i,j} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}$$

- Σε αυτή την περίπτωση, έχουμε δυο ακολουθίες, **x** και **y**. Η αμοιβαία πληροφορία μετράει πόση διάφορα έχει η από κοινού κατανομή της σ.π. των *X* και *Y* που συμβολίζουμε με *P*(*xi,yi*), με την υποθετική από κοινού κατανομή που θα είχαν αν ήταν ανεξάρτητες με *P*(*xi,yi*)=*P*(*xi*)*P(yi*). Προφανώς *P*(*xi*) και *P*(*yi*) είναι οι περιθώριες σ.π. των *X,Y* αντίστοιχα. Δηλαδή, η αμοιβαία πληροφορία μετράει το «πόσο ανεξάρτητες» είναι οι δυο κατανομές. Η σχετική εντροπία και η αμοιβαία πληροφορία, βρίσκουν πολλές εφαρμογές όταν μελετάμε ταυτόχρονα πολλές ακολουθίες και σχετικά παραδείγματα θα δούμε στο κεφάλαιο που περιγράφει την πολλαπλή στοίχιση.

Ροές ευνοϊκών αποτελεσμάτων

- Σε μια ακολουθία DNA μήκους n καταλοίπων, ποια η πιθανότητα να εμφανιστούν k συνεχόμενες επαναλήψεις ενός συμβόλου (π.χ. A)?

- Νόμος των Erdos & Renyi (1970)

Σε μια ακολουθία n ανεξάρτητων δοκιμών Bernoulli με πιθανότητα «επιτυχίας» p, με $0 \le p \le 1$ *, το αναμενόμενο μήκος* R_n *μέγιστης δυνατής ροής ευνοϊκών αποτελεσμάτων, είναι ίσο κατά προσέγγιση με* $\log_{1/p}(n)$ ή αλλιώς

$$\frac{R_n}{\log_{1/p}(n)} \to 1 \text{ με πιθανότητα } 1.$$

Αν το ευνοϊκό αποτέλεσμα έχει πιθανότητα p τότε μια ροή x συνεχών ευνοϊκών αποτελεσμάτων έχει πιθανότητα p^x , αν έχουμε n επαναλήψεις $(n\to+\infty)$ τότε έχουμε περίπου n δυνατές ροές και

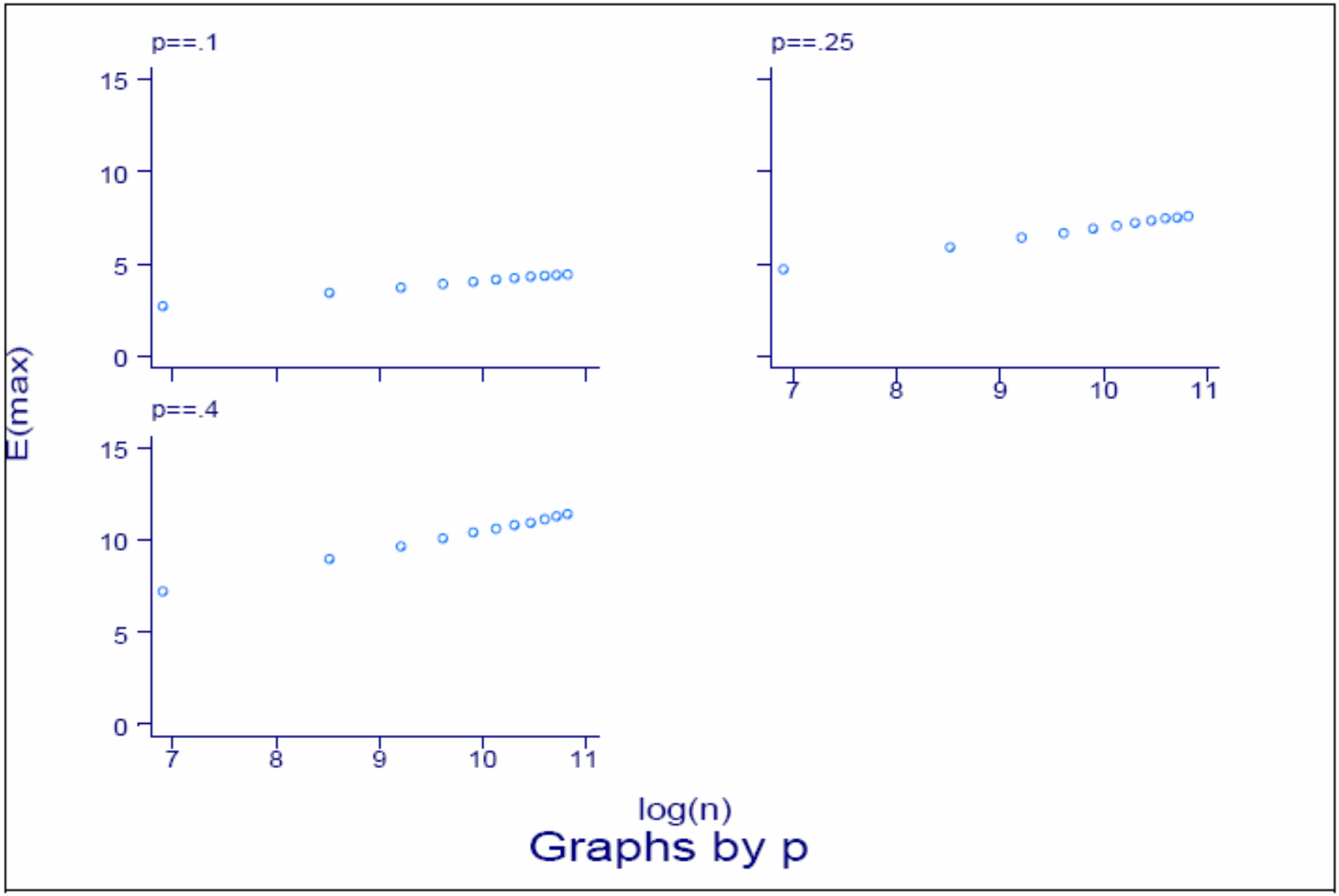
$$E(\#\text{ροων μηκους x}) \cong np^x$$

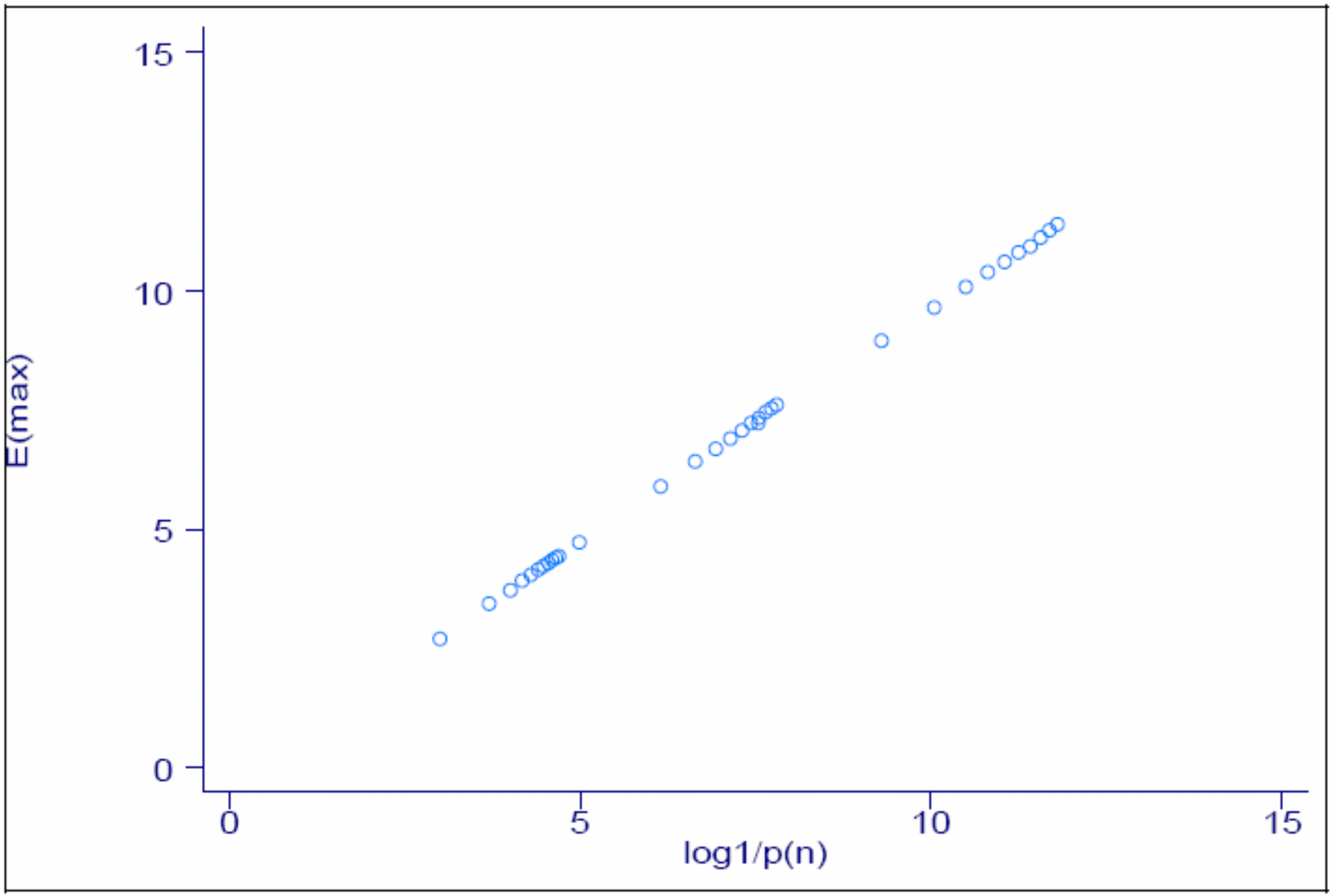
Αν η μέγιστη ροή είναι μοναδική τότε το μήκος της R_n ικανοποιεί τη σχέση $1 = np^{R_n}$ άρα:

$$R_n = \log_{1/p}(n)$$

Παράδειγμα Σε μια ακολουθία $n=10000$ βάσεων του DNA, θεωρώντας αυτές ισοπίθανες (δηλαδή $p_\kappa=1/4$) μας ενδιαφέρει να βρούμε τον αριθμό των μέγιστων επαναλήψεων Α που μπορεί να έχει συμβεί κατά τύχη δηλαδή θεωρώντας ότι η αλληλουχία είναι τυχαία και άρα δεν έχει βιολογική σημασία. Τότε το μέγιστο μήκος ροής από Α θα είναι :

$$R_n = \log_{1/p}(n) \Rightarrow R_n = \log_4(10000) \Rightarrow R_n = \frac{\log_{10} 10000}{\log_{10} 4} = \frac{4}{0.60205} = 6.64$$


p==.1
p==.25
p==.4
E(max)
15
10
5
0
7
8
9
10
11
log(n)
Graphs by p


15
10
5
0
E(max)
0
5
10
15
log1/p(n)

ΘΕΩΡΙΑ ΜΕΓΑΛΩΝ ΑΠΟΚΛΙΣΕΩΝ (LDT)

- Σε μια τυχαία ακολουθία, μας ενδιαφέρει η πιθανότητα εμφάνισης π.χ. μιας περιοχής μήκους L, αποτελούμενης από 100κ% (0<κ<1) επαναλήψεις ενός συμβόλου (π.χ. A)
- Σχετική Εντροπία (Relative Entropy):

$$H(\alpha, p) \equiv \alpha \log\left(\frac{\alpha}{p}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-p}\right) = \log\frac{a^a(1-a)^{1-a}}{p^a(1-p)^{1-a}} = -\log\left(\frac{p}{a}\right)^a\left(\frac{1-p}{1-a}\right)^{1-a}$$

Η συνάρτηση αυτή μετρά τη διαφορά μεταξύ της κατανομής *B(n,p)* από την οποία προέρχονται τα δεδομένα μας (η οποία έχει δώσει γένεση σε μια ακολουθία DNA με πιθανότητα εμφάνισης των βάσεων ίση με p) και μιας άλλης υποθετικής *B(n,α)* για την οποία υποπτευόμαστε ότι έχει δώσει γένεση σε μια υπό- ακολουθία (τοπική) μήκους n στην οποία παρατηρούμε ότι για παράδειγμα η εμφάνιση μιας βάσης, διαφέρει πολύ από την αναμενόμενη καθώς έχει συχνότητα $\alpha=k/n$. Προφανώς $\alpha, p \in (0,1)$.

Το κλειδί στην κατανόηση των μεγάλων αποκλίσεων, είναι το γεγονός ότι έχουμε να κάνουμε με δυο διαφορετικές πιθανότητες *(α,p)* στον ίδιο χώρο πιθανών εκβάσεων.

Θεώρημα (Arratia and Gordon,1989)

Έστω $0 < p < \alpha < 1$, *n=1,2,3…Αν* $Y \sim B(n,p)$ *και* $H = H(\alpha, p)$ *όπως ορίσαμε παραπάνω τότε:*

$$P(Y \geq \alpha n) \leq e^{-nH}$$

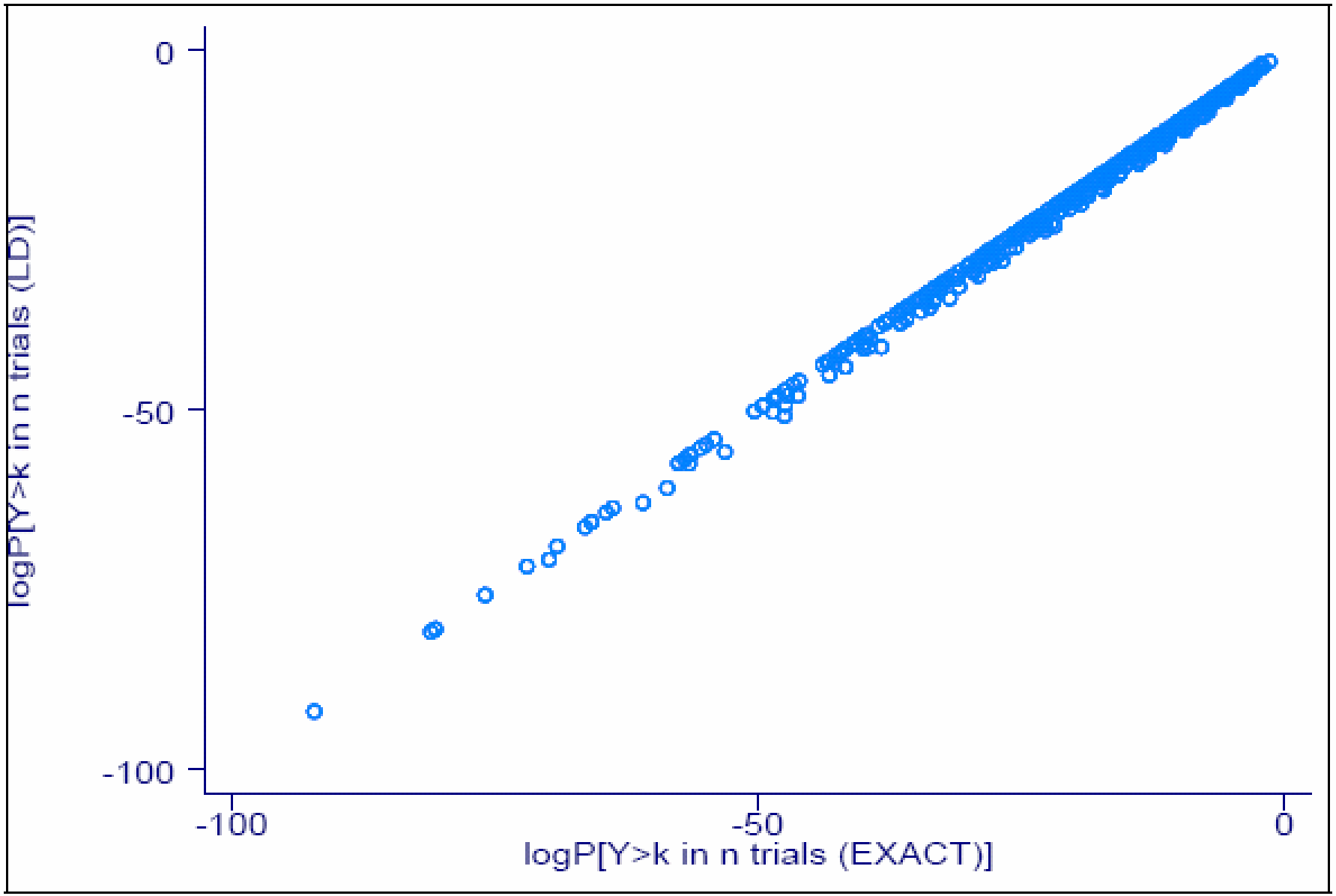
Έστω $0 < p < \alpha < 1$, *n=1,2,3…Αν* $Y \sim B(n,p)$ *και* $H = H(\alpha, p)$, *r όπως τα ορίσαμε παραπάνω τότε*:

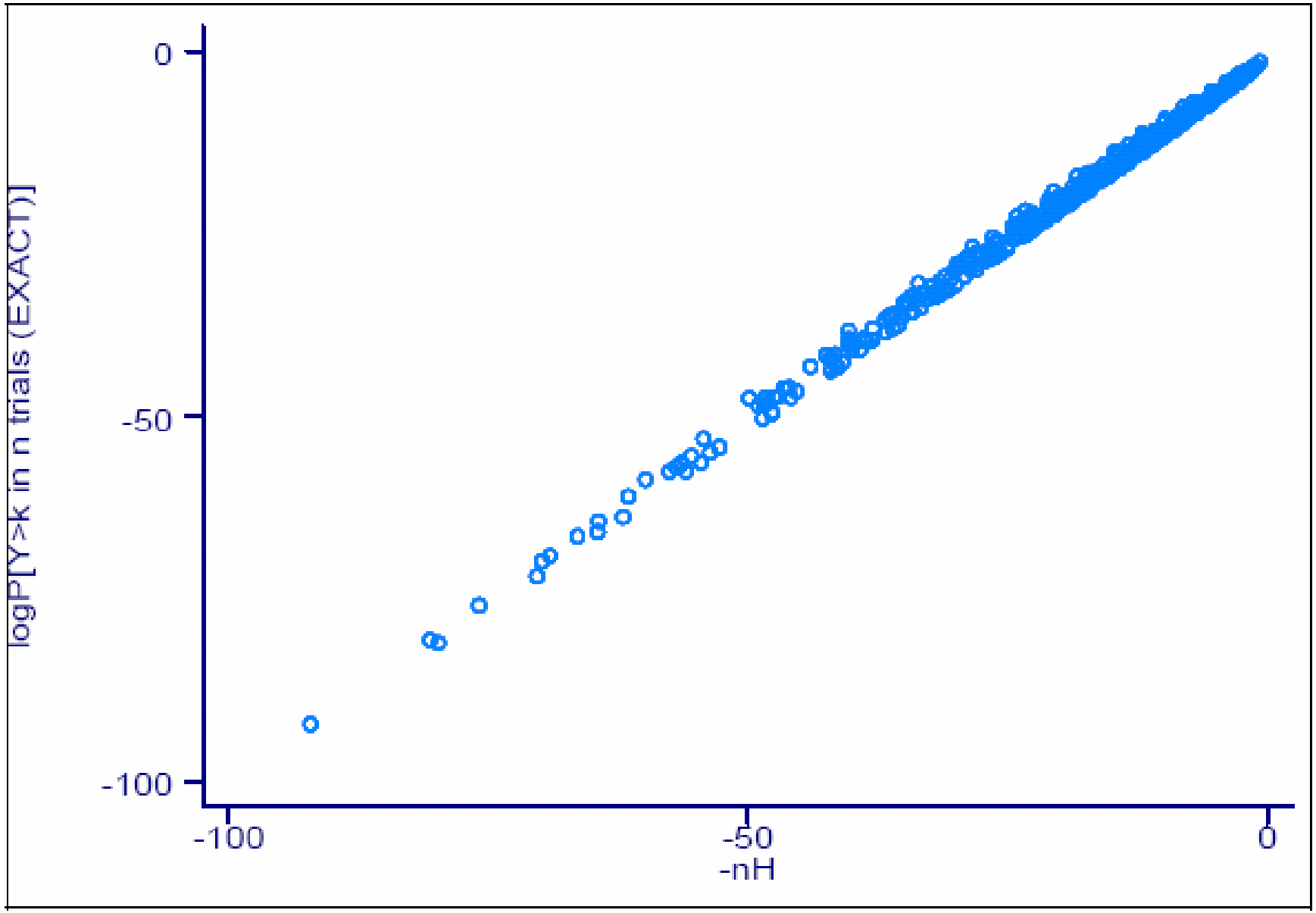
$$P(Y \geq \alpha n) \approx e^{-nH} \Leftrightarrow \log P(Y \geq \alpha n) \sim -nH \text{ και}$$

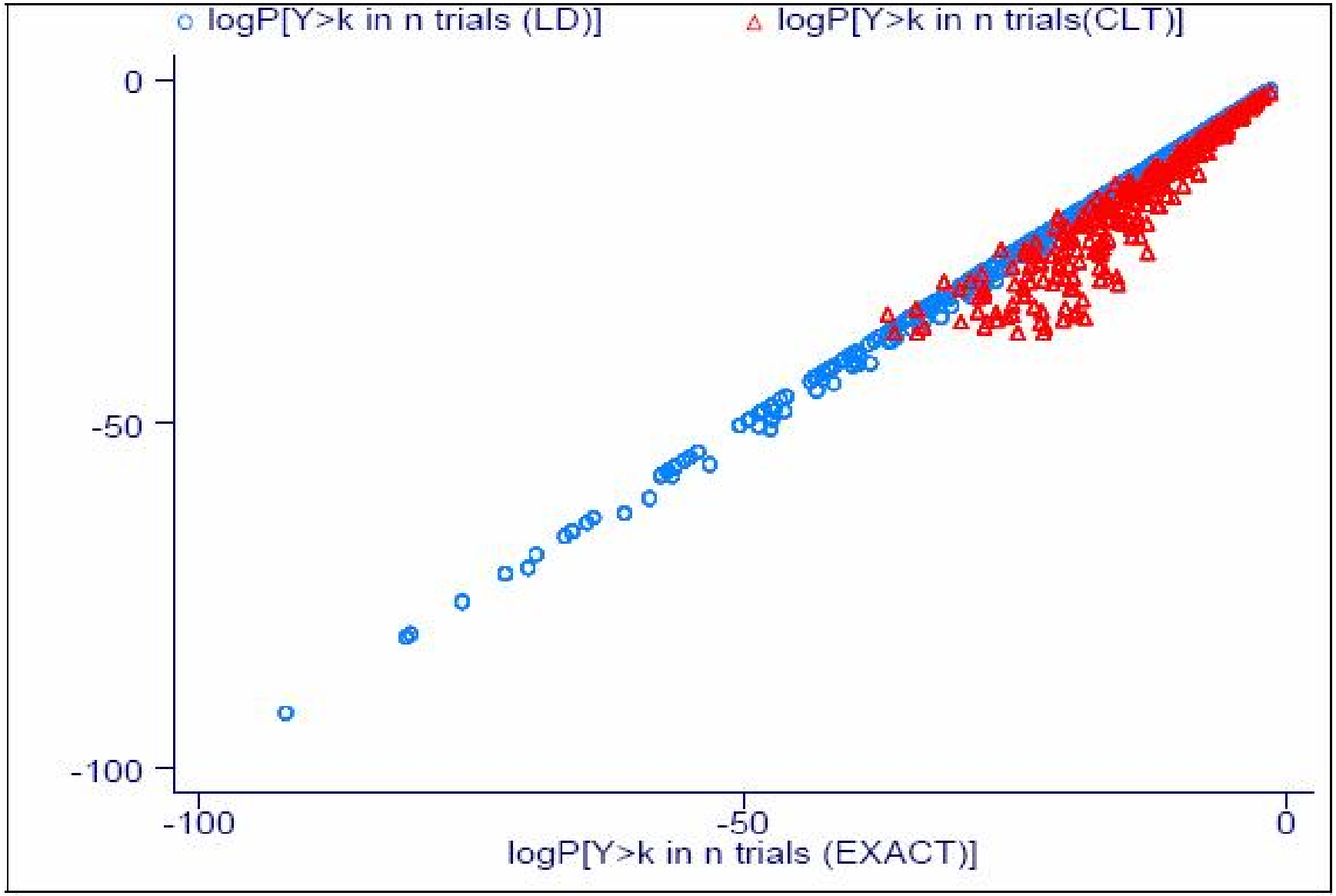
$$P(Y \geq \alpha n) \sim \frac{1}{1-r}\left(\frac{1}{\sqrt{2\pi\alpha(1-\alpha)n}}\right)e^{-nH}$$

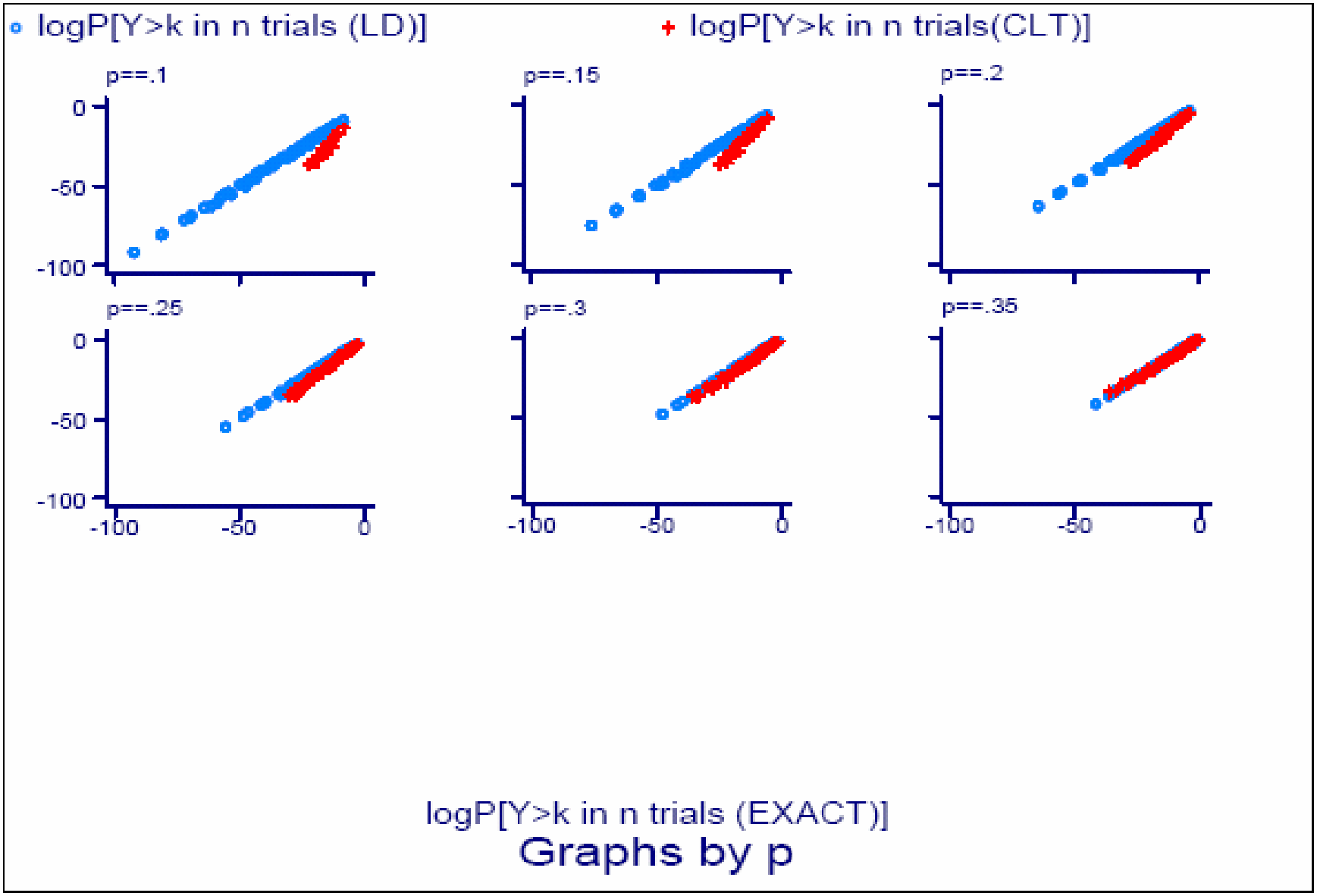
$$P(Y = \alpha n + i) \sim \left(\frac{1}{\sqrt{2\pi\alpha(1-\alpha)n}}\right).r^i e^{-nH}, i = 0,1,2\ldots$$

$$P(Y = \alpha n + i \mid Y \geq \alpha n) \to r^i(1-r), i = 0,1,2\ldots$$


0
-50
-100
logP[Y>k in n trials (LD)]
-100
-50
0
logP[Y>k in n trials (EXACT)]


0
-50
-100
logP[Y>k in n trials (EXACT)]
-100
-50
0
-nH


logP[Y>k in n trials (LD)]
logP[Y>k in n trials(CLT)]
0
-50
-100
-100
-50
0
logP[Y>k in n trials (EXACT)]


logP[Y>k in n trials (LD)]
logP[Y>k in n trials(CLT)]
p==.1
p==.15
p==.2
p==.25
p==.3
p==.35
0
-50
-100
-100
-50
0
logP[Y>k in n trials (EXACT)]
Graphs by p

Παράδειγμα Ας υποθέσουμε ότι κάτω από τις προϋποθέσεις του τυχαίου μοντέλου που είδαμε πιο πάνω, θέλουμε να υπολογίσουμε την πιθανότητα σε μια ακολουθία 20 βάσεων DNA να έχουμε 16 ή περισσότερες εμφανίσεις A

Έχουμε $n=20,\ p=0.25,\ \alpha=0.8$. Με έναν ακριβή (exact) υπολογισμό από τη διωνυμικη κατανομή $(X \sim bin(20,0.25))$ έχουμε $P(X\geq 16)= 0.3865*10^{-6}$ και η μεγαλύτερη συνεισφορά σ' αυτή την πιθανότητα είναι η $P(X= 16)=0.3569*10^{-6}$.

Αν θέλαμε να υπολογίσουμε αυτήν την πιθανότητα μέσω του ΚΟΘ (κανονική προσέγγιση) θα είχαμε $B(n,p) \xrightarrow{d} N(np, np(1-p))$ δηλαδή $\mu=np=5$, $\sigma^2=np(1-p)=3.75$.

Η τιμή του z που αντιστοιχεί σε $k \geq 16$ είναι $\frac{(k-np)}{\sigma} = 5.68$, το οποίο δίνει ένα *p-value*$=0.0069*10^{-6}$.

Αν είχαμε ενσωματώσει και τη διόρθωση συνέχειας πάλι θα είχαμε $p=0.03038*10^{-6}$. Βλέπουμε ότι αυτές οι τιμές είναι πολύ μικρότερες (12-55 φορές) από την ακριβή τιμή της πιθανότητας. Παρατηρούμε βέβαια ότι οι προϋποθέσεις για την εφαρμογή του ΚΟΘ δεν ισχύουν πλήρως εδώ ($np \geq 10, n \geq 30$). Ας προχωρήσουμε για να δούμε πως εφαρμόζονται οι προσεγγιστικές σχέσεις που δώσαμε παραπάνω.

Κατ' αρχήν θα έπρεπε να υπολογίσουμε τη σχετική εντροπία

$$H(\alpha, p) \equiv \alpha \log\left(\frac{\alpha}{p}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-p}\right)$$

$$\Rightarrow H = 0.8\log(3.2) + 0.2\log\left(\frac{4}{15}\right) = 0.666$$

Από την σχέση (2.9) έχουμε $P(Y \geq \alpha n) \leq e^{-nH} = 1.64*10^{-6}$. Επίσης

$r=1/12$, και $\sqrt{2\pi a n(1-a)} = 4.48$ οπότε από την σχέση βρίσκουμε

$$P(Y \geq \alpha n) = 0.398*10^{-6}$$

Παρατηρούμε δηλαδή ότι η προσεγγιστική αυτή μέθοδος δίνει αποτελέσματα πολύ κοντά στην ακριβή τιμή της πιθανότητας, παρ' ότι οι προϋποθέσεις για την εφαρμογή του ΚΟΘ δεν ισχύουν πλήρως.

Αν τώρα θέλουμε να υπολογίσουμε την αναμενόμενη τιμή που έχει ένα τέτοιο «παράθυρο» 20 βάσεων με τουλάχιστον 16 Α, να εμφανιστεί κατά τύχη σε μια ακολουθία DNA μήκους 10^6 βάσεων, θα πολ/ζαμε την παραπάνω προσέγγιση της πιθανότητας με το 10^6 (για την ακρίβεια με το 999.981 επειδή σε μήκος βάσεων $n=10^6$ μπορούν να σχηματιστούν $n-k+1$=999.981 παράθυρα μήκους $k=20$ βάσεων). Παρατηρούμε ότι με την προσεγγιστική αυτή μέθοδο (large deviation) και με τον ακριβή υπολογισμό από την διωνυμικη παίρνουμε *E-value (αναμενόμενη τιμή)* ≈ 0.4 ενώ με την προσέγγιση από το ΚΟΘ παίρνουμε *E-value* ≈ 0.0069 και υπάρχει κίνδυνος να αποδώσουμε στατιστική σημαντικότητα σε ένα γεγονός που οφείλεται στην τύχη.

Επεκτάσεις στο νόμο Erdos-Renyi

Θεώρημα (Erdös και Renyi,1970; Erdös και Revesz, 1975)

Σε μια ακολουθία n ανεξάρτητων δοκιμών Bernoulli με πιθανότητα «επιτυχίας» p , με $0 \le p < a \le 1$, *το πλήθος* R_n^a *διαδοχικών δοκιμών που περιέχουν 100α% ευνοϊκά αποτελέσματα, ικανοποιεί τη σχέση:*

$$\frac{R_n^a}{\log(n)} \to \frac{1}{H(a,p)}$$ *με πιθανότητα 1*

Μια διαισθητική ερμηνεία του αποτελέσματος έχει ως εξής: Από τη θεωρία των μεγάλων αποκλίσεων (Large Deviations) βρίσκουμε ότι μια περιοχή μήκους x η οποία περιέχει 100α.% ευνοϊκά αποτελέσματα, έχει περίπου $e^{-xH(a,p)}$ πιθανότητα.

Επειδή τώρα κάθε ροή έχει περίπου $n\text{-}k+1 \approx n$ δυνατές περιοχές έναρξης έχουμε:

$$1 = ne^{-xH(a,p)} \Rightarrow R_n^a \equiv x = \frac{\log(n)}{H(a,p)}$$

Παρατηρούμε ότι για $a = 1 \Rightarrow H(a,p) = \log(1/p)$

Παράδειγμα Σε μια αλληλουχία 1000000 βάσεων DNA ο μέγιστη περιοχή (ροή) R_n^a που να περιέχει κατ' ελάχιστο 80% βάσεις Αδενίνης (A) είναι :

$$R_n^a = \frac{\log(n)}{H(a,p)} = \frac{\log(1000000)}{0.666} = 20.744$$ (να σημειωθεί εδώ ότι όταν γράφουμε log εννοούμε λογάριθμο με βάση το e)

Η κατανομή της μέγιστης ροής/Extreme Value Distribution (EVD)

- Μας ενδιαφέρει εδώ, η ακριβής στατιστική κατανομή που ακολουθεί η τ.μ. της μέγιστης ροής

Πιο αυστηρά αν έχουμε ένα δείγμα $X_1, X_2, ..., X_n$ από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές (iid) τότε μας ενδιαφέρει ποια είναι η πιθανή οριακή κατανομή του:

$$M_n = a_n[\max(X_1, X_2, ..., X_n) - b_n], n \to \infty$$

Όπου a_n, b_n κατάλληλες σταθερές κανονικοποίησης ώστε να προκύπτει μη τετριμμένη κατανομή.

Extreme Value Distribution (EVD)

1. $F(y) = \exp(-e^{-y}), -\infty \le y \le \infty$ (Gumbel)

2. $F(y) = \begin{cases} 0, y \le 0 \\ \exp(-y^{-a}), y \ge 0, a > 0 \end{cases}$ (Frechet)

3. $F(y) = \begin{cases} \exp(-(-y)^{a}), y < 0, a > 0 \\ 1, y \ge 0 \end{cases}$ (Weibull)

Η κατανομή που αφορά την δική μας περίπτωση είναι η Gumbel, και προκύπτει από την γενικευμένη μορφή της κατανομής των ακραίων τιμών (Generalized Extreme Value Distribution – GEVD):

$$H(y) = \exp\left\{-\left(1 + k\left(\frac{y-a}{b}\right)\right)^{-\frac{1}{k}}\right\} \text{με } -\infty < \alpha, \mathrm{k} < \infty, \mathrm{b} > 0$$

η οποία ορίζεται όταν $1 + k\left(\frac{y-a}{b}\right) > 0$, ως το όριο καθώς $k \to 0$

Αν θέσουμε $z = \left(\frac{y-a}{b} \right)$, και $t = -\frac{1}{k}$ θα έχουμε :

$$H(y) = \exp\left\{ -\left(1 - \frac{z}{t} \right)^t \right\}$$

και αν πάρουμε το όριο καθώς το $k \to 0 \Rightarrow t \to \infty$ επειδή είναι γνωστή η σχέση:

$$\lim_{n\to\infty} \left(1 - \frac{a}{n} \right)^n = e^{-a}$$

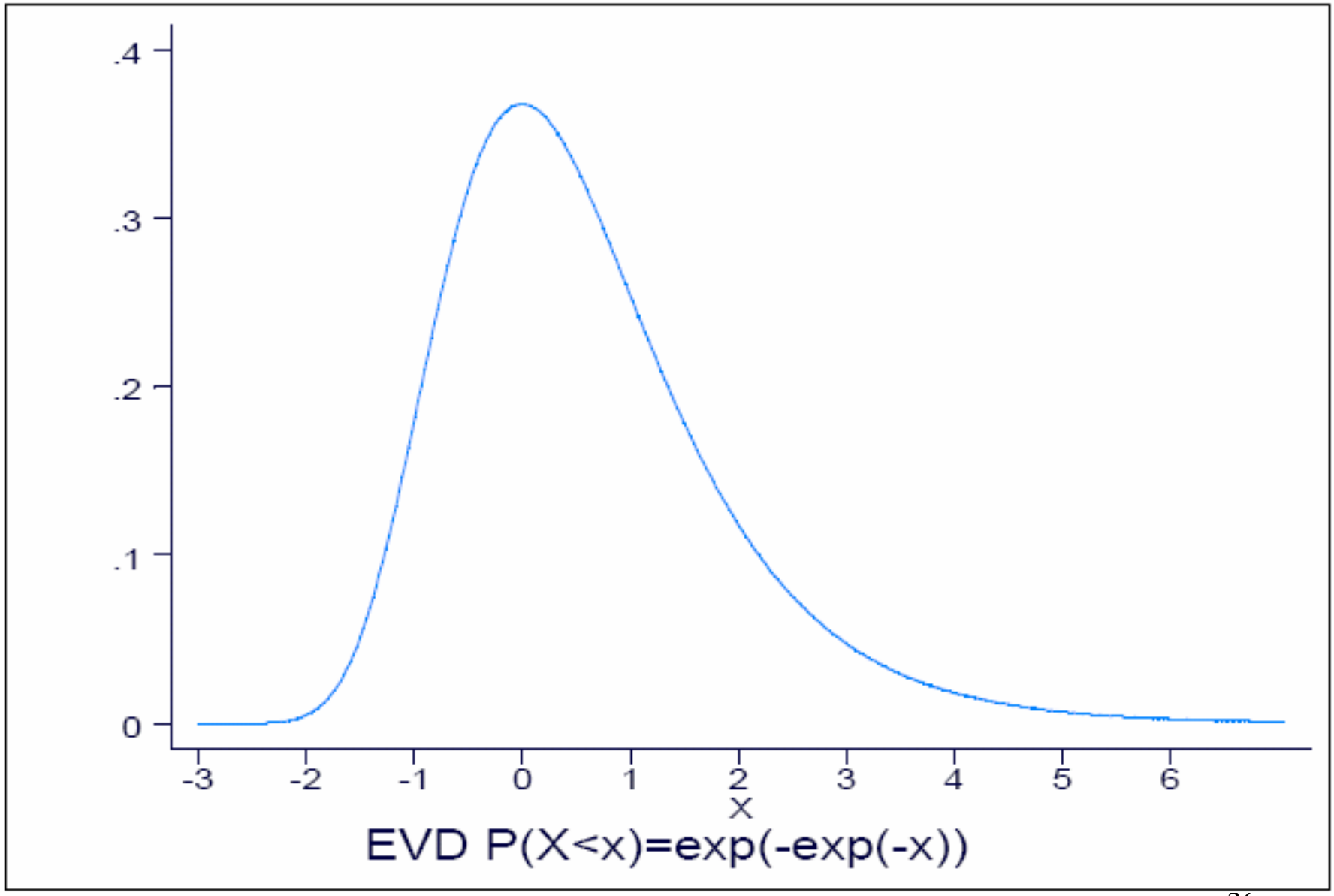
θα έχουμε

$$\lim_{t\to\infty} H(y) = \lim_{t\to\infty} \exp\left\{ -\left(1 - \frac{z}{t} \right)^t \right\} = \exp\left\{ -e^{-z} \right\} = \exp\left\{ -e^{-\left(\frac{y-a}{b} \right)} \right\}$$

Έτσι η κατανομή του $Y_n = \max(X_1, X_2, ..., X_n)$ γίνεται (Gumbel, 1958):

$$F(Y) = \exp(-e^{-\frac{(y-a)}{b}}), -\infty \le y \le \infty$$

με $E = a - b\Gamma'(1)$ και $V = \frac{b^2\pi^2}{6}$


.4
.3
.2
.1
0
-3
-2
-1
0
1
2
3
4
5
6
X
EVD P(X<x)=exp(-exp(-x))

Αποδεικνύεται (Arratia et al, 1986; Arratia et al 1990; Waterman 1995) ότι στην περίπτωση της συνεχούς ροής ενός αποτελέσματος (νόμισμα η βάσεις DNA) για $a_n = \frac{\log(qn)}{\lambda}, b_n = \frac{1}{\lambda}$ όπου $\lambda = \log(1/p)$ ισχύει:

$$\lim_{n\to\infty}\left(R_n < \frac{\log(nq)}{\lambda} + \frac{y}{\lambda} \right) = \exp\left(-e^{-y}\right)$$

Για την ΑΣΚ της τ.μ. R_n θα ισχύει:

$$F(y) = P(R_n \le y) \approx \exp\left(-\exp\left(-\frac{y - \log(nq)/\lambda}{1/\lambda} \right) \right) \qquad (2.19)$$

Η ερμηνεία του αποτελέσματος αυτού έχει ως εξής (Waterman, 1995):

Η κατανομή της εμφάνισης μιας βάσης (A) είναι διωνυμικη, και για συνεχόμενες εμφανίσεις έχουμε γεωμετρική κατανομή $P(Z_i = m) = qp^m$ άρα $R_n \approx \max_{1\le i\le nq} Z_i$

Επιπλέον μπορεί να δειχθεί ότι $Z_i = [W_i]$ όπου $W \sim \exp(\lambda)$ με $\lambda = \log\left(1/\mathrm{p}\right)$, όπου [x] το ακέραιο μέρος του x, άρα

$$R_n \approx \left[\max_{1 \le i \le nq} W_i \right]$$

και αρα (Waterman, 1995):

$$R_n \approx \left[\frac{\log(nq)}{\lambda} + \frac{y}{\lambda} \right]$$

Από τα παραπάνω προκύπτει :

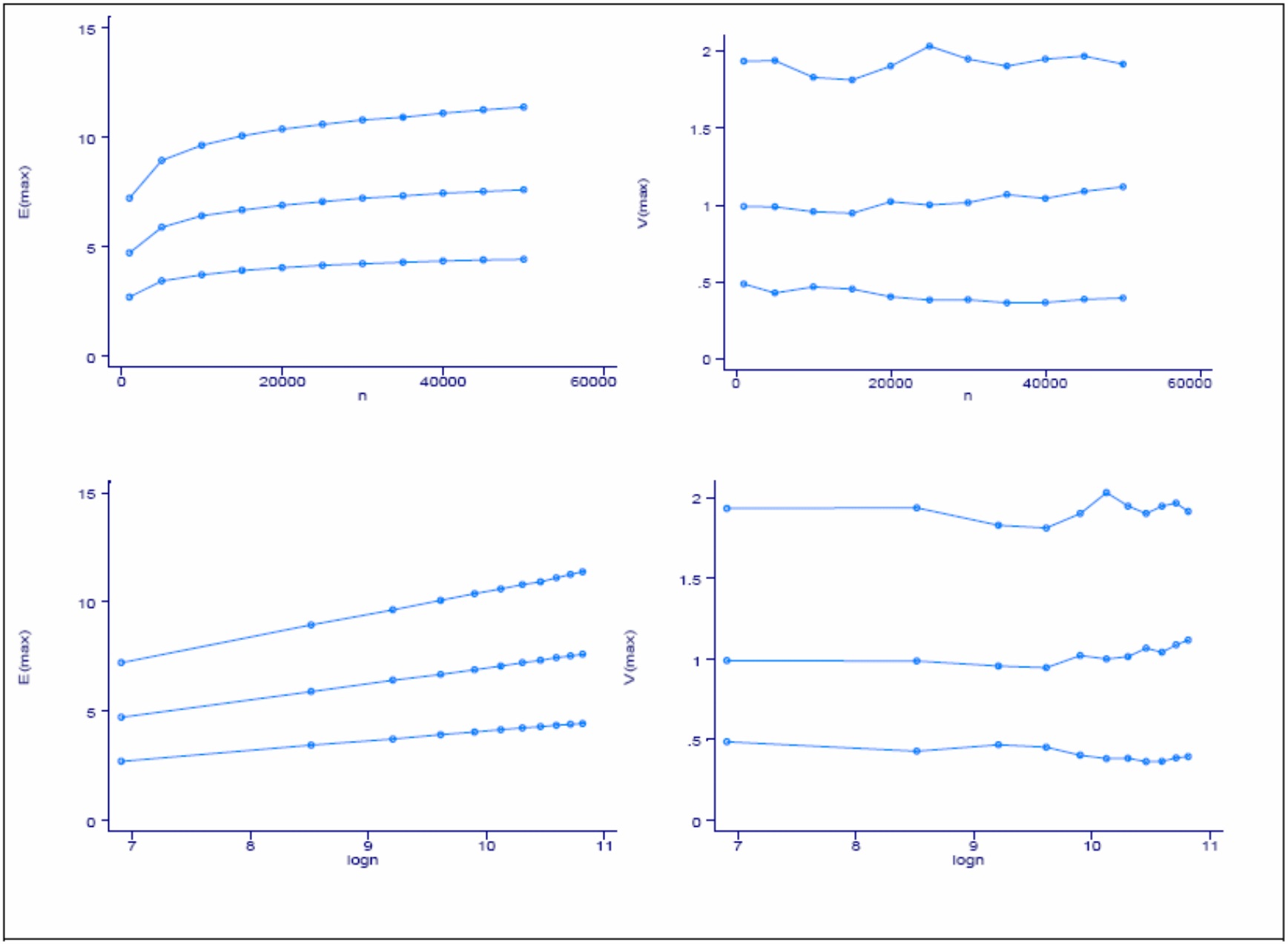
$$E(R_n) \approx \frac{\log(n)}{\lambda} + \frac{\log(q)}{\lambda} + \frac{\gamma}{\lambda} - \frac{1}{2}$$

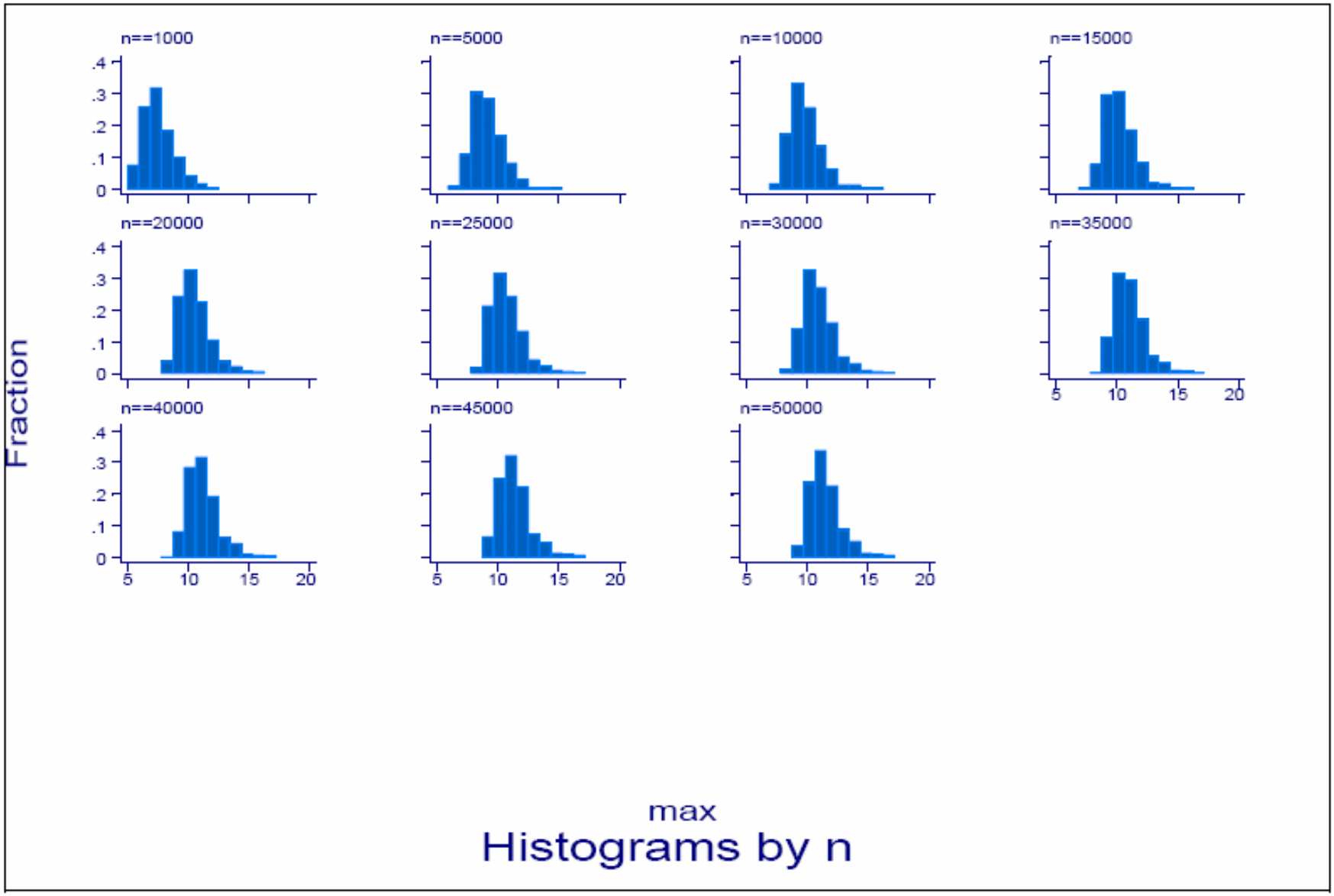
$$\Rightarrow E(R_n) \approx \log_{1/p}(n) + \log_{1/p}(q) + \gamma/\lambda - 1/2$$

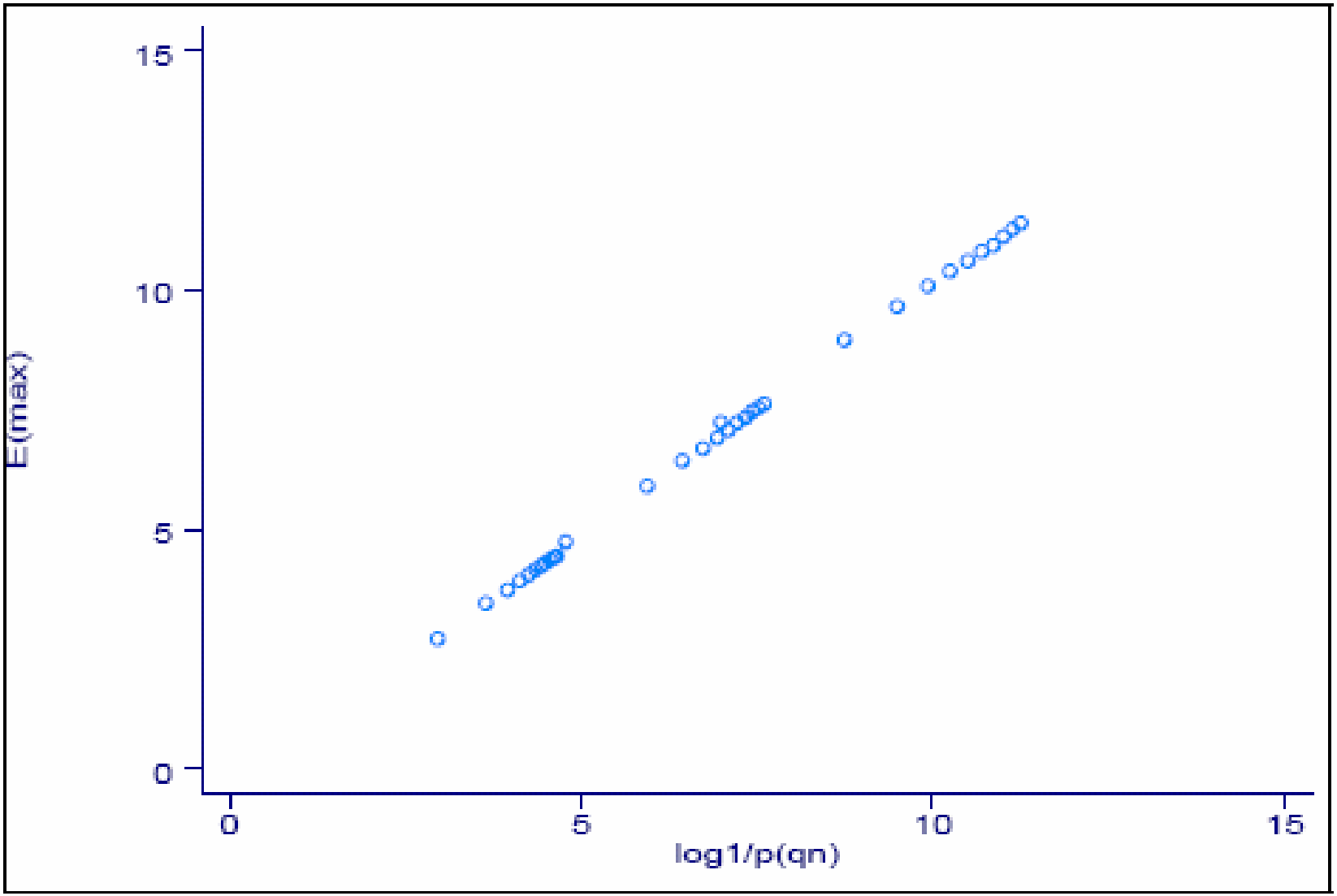
$$\text{Και } Var(R_n) \approx \frac{\pi^2}{6\lambda^2} + \frac{1}{12}$$

όπου $\gamma = -\Gamma'(1) = 0.5772\ldots$ η σταθερά *Euler-Mascheroni*.

Η αφαίρεση από τη μέση τιμή του ½ και η πρόσθεση στη διασπορά 1/12 είναι η διόρθωση συνέχειας του *Sheppard* , και γίνεται διότι όταν μετατρέπουμε μια συνεχή τ.μ. σε διακριτή αυξάνεται η μέση τιμή της και μειώνεται η διασπορά.


E(max)
15
10
5
0
0
20000
40000
60000
n
V(max)
2
1.5
1
.5
0
0
20000
40000
60000
n
E(max)
15
10
5
0
7
8
9
10
11
logn
V(max)
2
1.5
1
.5
0
7
8
9
10
11
logn


n==1000
n==5000
n==10000
n==15000
n==20000
n==25000
n==30000
n==35000
n==40000
n==45000
n==50000
Fraction
0
.1
.2
.3
.4
5
10
15
20
max
Histograms by n


15
10
5
0
E(max)
0
5
10
15
log1/p(qn)

Η κατανομή του μέγιστου τμηματικού score (Maximal Segment Score)

- Αν μας ενδιαφέρει η κατανομή της τ.μ.της πιθανότητας εμφάνισης π.χ. μιας περιοχής μήκους L, αποτελούμενης από 100κ% (0<κ<1) επαναλήψεις ενός συμβόλου (π.χ. A)
- Ορίζουμε το Score:

$$s_k = \log\left(a_k / p_k\right)$$

Για τον υπολογισμό του μέγιστου τμηματικού (τοπικού) score θέτουμε κάποιους περιορισμούς

1. Τουλάχιστον ένα score πρέπει να είναι θετικό
2. Η αναμενόμενη τιμή του score για κάθε βάση να είναι αρνητική, δηλαδή

$$E(s_k) = \sum p_k s_k = \sum p_k \log\left(\frac{a_k}{p_k}\right) < 0$$

Θεώρημα Karlin-Altschul

Θεώρημα (Karlin and Altschul, 1990)

Η τυχαία μεταβλητή M(n) (το μέγιστο τμηματικό score) έχει προσεγγιστική κατανομή την :

$$P\left\{M(n) > \log(n)/\lambda + x\right\} \approx 1 - \exp\left\{-Ke^{-\lambda x}\right\}$$

Αυτή είναι η κατανομή των ακραίων τιμών του Gumbel, ενώ K και λ είναι οι σταθερές της και υπολογίζονται με αριθμητικές (προσεγγιστικές) μεθόδους οι οποίες δεν θα παρουσιαστούν εδώ. Για το λ ειδικά ισχύει το ότι είναι η μοναδική θετική λύση της εξισώσεως:

$$\sum_k p_k \exp\{\lambda s_k\} = 1$$

Ισχύει επίσης το επόμενο:

Θεώρημα (Karlin and Altschul, 1990)

Καθώς το μήκος n της τυχαίας ακολουθίας τείνει στο άπειρο, η συχνότητα a_k *της εμφάνισης κάποιας βάσης σε ένα τμήμα με αρκετά μεγάλο score προσεγγίζει το* $p_k \exp\{\lambda s_k\}$ *με πιθανότητα 1. Για την ακρίβεια όταν έχουμε το μέγιστο score, τότε :*

$a_k = p_k \exp\{\lambda s_k\}.$

Παρατηρούμε επίσης ότι καθώς η ύπαρξη τέτοιων τμημάτων με μεγάλο score (μεγαλύτερο από x) είναι σπάνια γεγονότα (rare events), θα ακολουθούν την κατανομή Poisson με μέση τιμή (παράμετρο) :

$E=Kn\exp\{-\lambda x\}$

$$P(M(n) \geq x) \approx 1 - e^{-E}$$

Όταν το E-value (μέση τιμή-αναμενόμενη τιμή) είναι πολύ μικρό τότε επειδή ισχύει η προσεγγιστική σχέση:

$$1-\exp(-\exp(-t))\approx 1-(1-\exp(-t))=\exp(-t)$$

θα έχουμε το P-value περίπου ίσο με το E-value.

Επομένωςκάνοντας χρήση της κατανομής Poisson, η πιθανότητα να βρούμε σε μια ακολουθία μήκους *n, m* τμήματα με score $S_{(m)}$ μεγαλύτερο η ίσο από το x θα είναι :

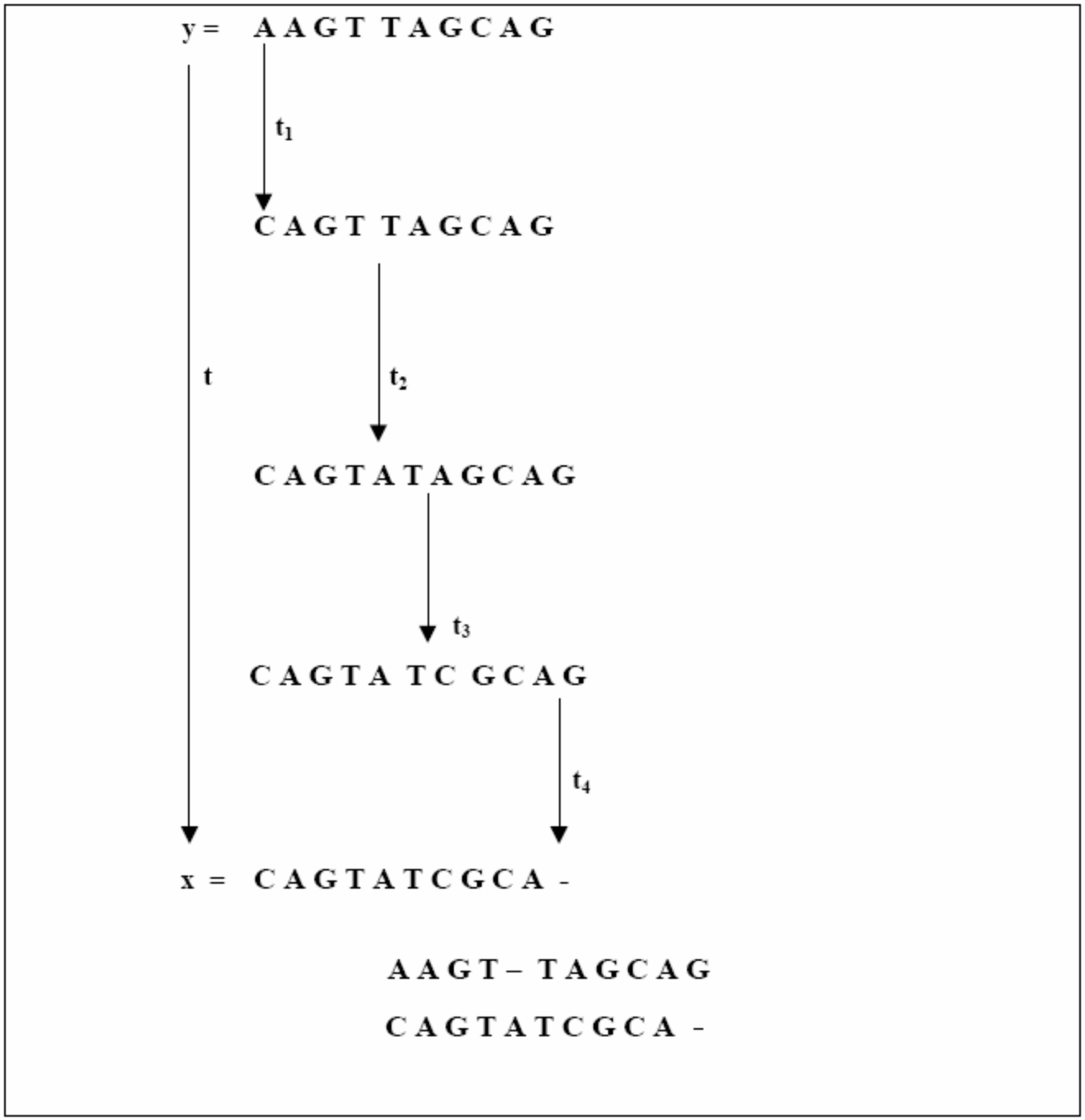
$$P(S_{(m)}\geq x)\approx 1-\exp\left(-Kne^{-\lambda x}\right)\sum_{i=0}^{m-1}\frac{\left(Kne^{-\lambda x}\right)^{i}}{i!}$$

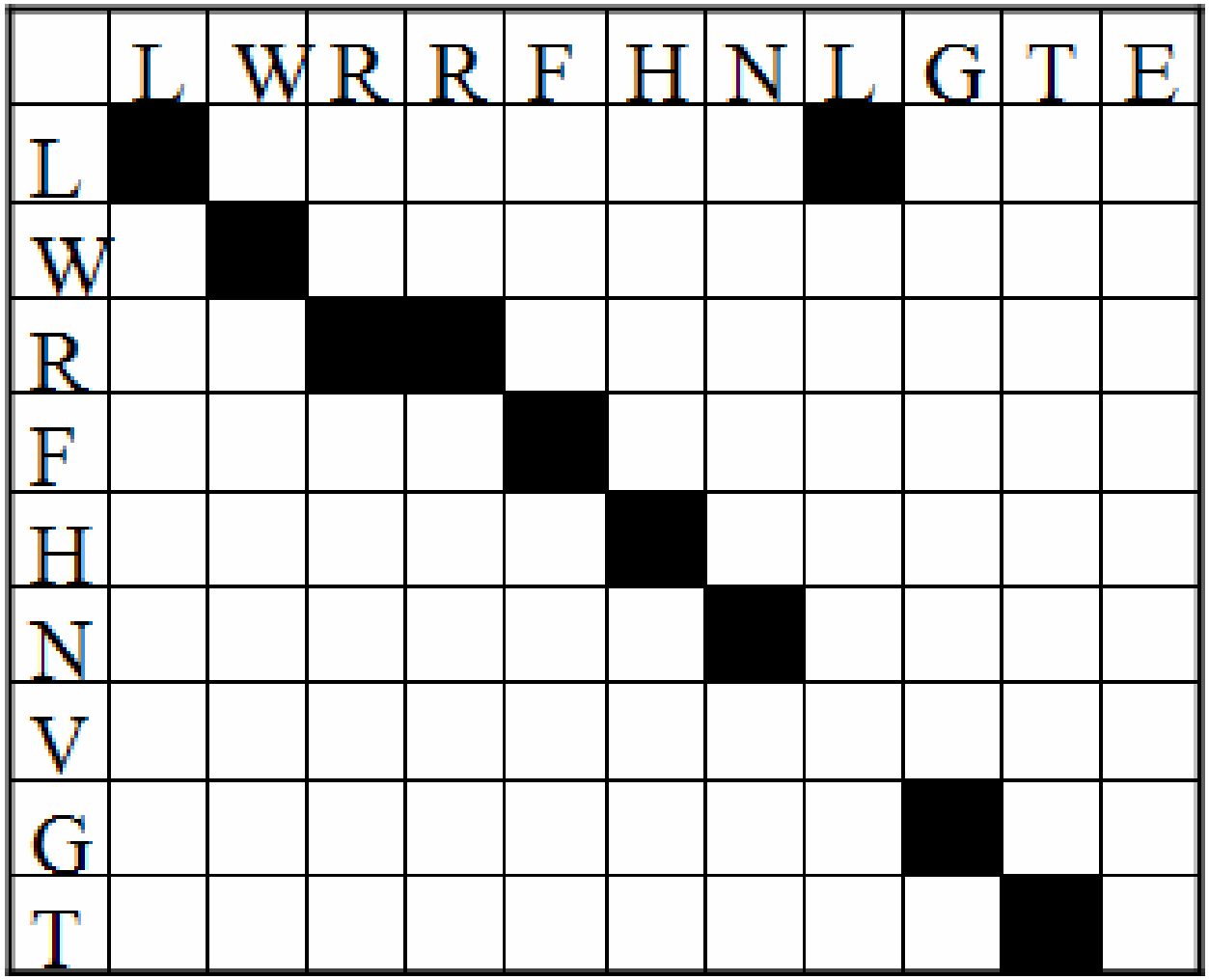
Στην ειδική περίπτωση της ροής R_n, που είδαμε παραπάνω μπορούν να δοθούν κλειστές εκφράσεις για τα K και λ και αυτές είναι :

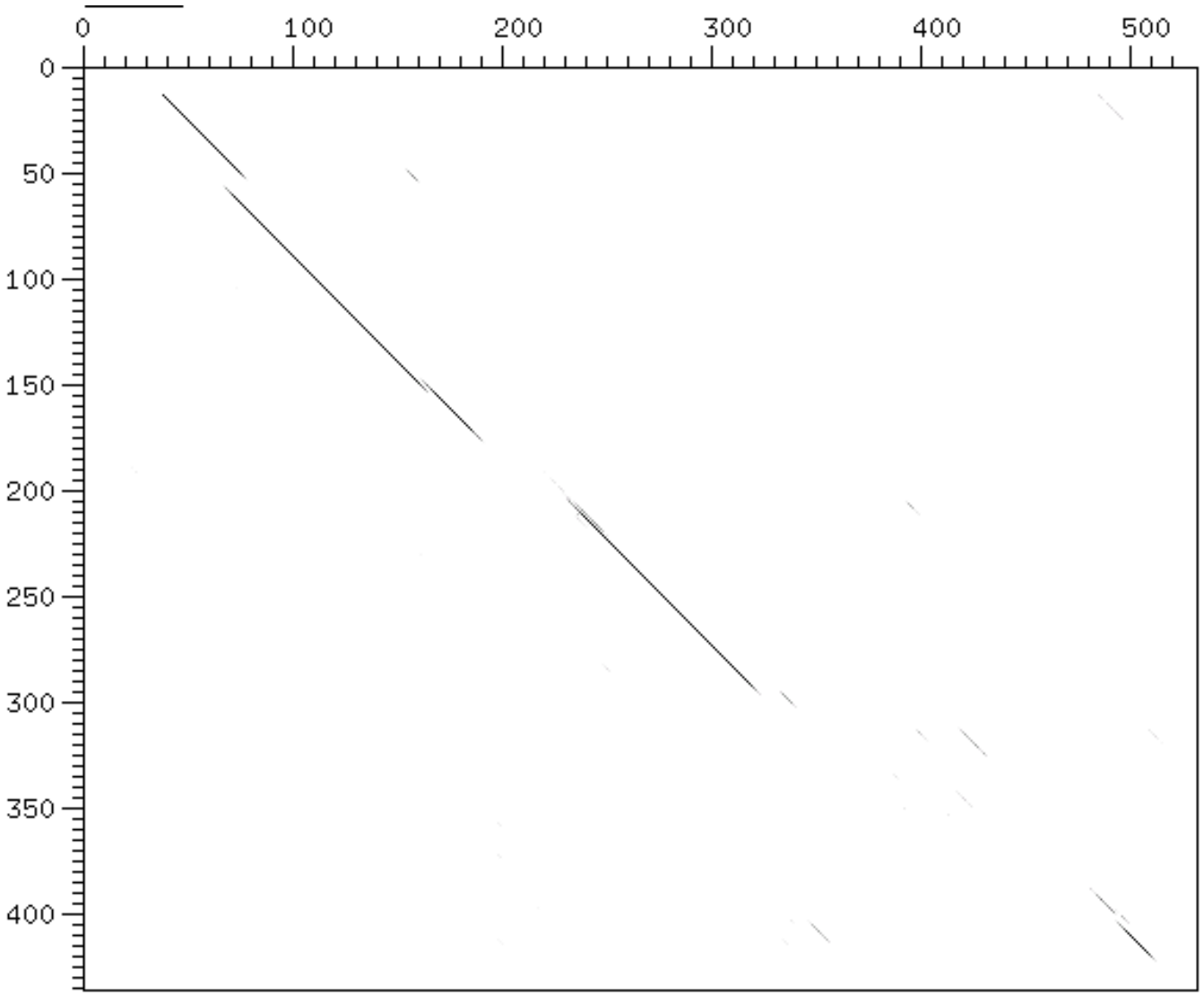
$K=1-p=q$, και $\lambda=\log(1/p)$

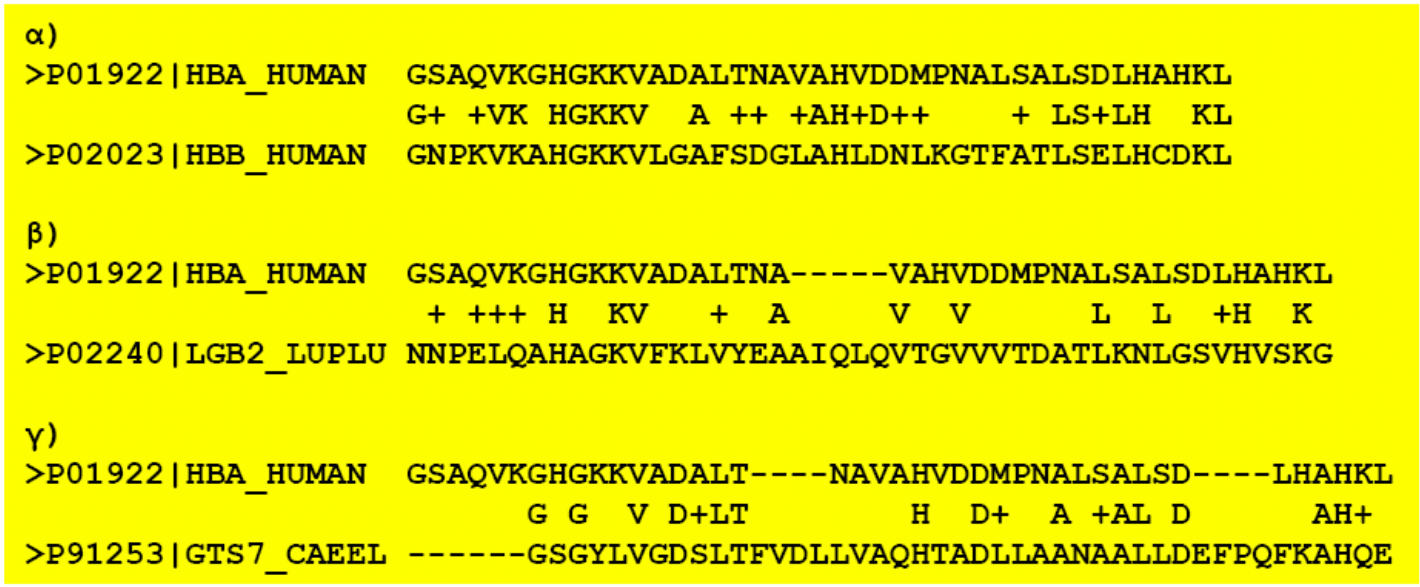
Κατά ζεύγη στοίχιση ακολουθιών

- Από τα πιο σημαντικά προβλήματα στην Υπολογιστική Βιολογία
- Ιδιαίτερα πλούσια βιβλιογραφία για πάνω από 30 χρόνια
- Η ομοιότητα δυο ακολουθιών αντανακλά κατά βάση την κοινή εξελικτική προέλευση


y = A A G T T A G C A G
t1
C A G T T A G C A G
t
t2
C A G T A T A G C A G
t3
C A G T A T C G C A G
t4
x = C A G T A T C G C A -
A A G T – T A G C A G
C A G T A T C G C A -


L W R R F H N L G T E
L
W
R
F
H
N
V
G
T


0
100
200
300
400
500
0
50
100
150
200
250
300
350
400



Εικόνα 3: Τρεις στοιχίσεις ακολουθιών με τον αλγόριθμο Needleman-Wunsch με ένα τμήμα της άλφα αλυσίδας της ανθρώπινης αιμοσφαιρίνης (SwissProt AC P01922).

α) Ξεκάθαρη ομοιότητα με τη βήτα αλυσίδα της ανθρώπινης αιμοσφαιρίνης (AC P02023).

β) Δομικά συμβατή στοίχιση με την leghemoglobin II (AC P02240) του δικοτυλίδονου Lupinus luteus.

γ) 'Παραπλανητική' στοίχιση με ομόλογη της S-τρανφεράσης της γλουταθειόνης (AC P91253) του νηματώδη σκώληκα C. elegans.

Σημαντικά ζητήματα στη στοίχιση ακολουθιών

- Το είδος των στοιχίσεων που μας ενδιαφέρουν
- Το σύστημα βαθμονόμισης (scoring system)
- Ο αλγόριθμος που θα χρησιμοποιήσουμε για την έυρεση της καλής ή και της βέλτιστης στοίχισης
- Ο τρόπος προσδιορισμού της στατιστικής σημαντικότητας μιας στοίχισης

Παράδειγμα

- Έστω 2 ακολουθίες **x,y** (ίδιου ή διαφορετικού μήκους)

  $\mathbf{x}=x_1,x_2,\ldots,x_n$

  $\mathbf{y}=y_1,y_2,\ldots,y_m$
- Μας ενδιαφέρει η εύρεση της μέγιστης κοινής περιοχής τους (πλήρης ταύτιση)
- Η απλή απαρίθμιση όλων των πιθανών κοινών υπό-περιοχών είναι απαγορευτική:

$$\binom{2n}{n} = \frac{(2n!)}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{2\pi n}}$$

- Χρειαζόμαστε έναν πιο αποδοτικό αλγόριθμο (δυναμικός προγραμματισμός)

Score

Θεωρούμε δυο πιθανότητες: την πιθανότητα ανεξάρτητης (τυχαίας) ταύτισης, και αυτή της μη τυχαίας

$$P(\mathbf{x},\mathbf{y} \mid R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

$$P(\mathbf{x},\mathbf{y} \mid M) = \prod_i p_{x_i y_i}$$

Αν πάρουμε το λόγο των δυο πιθανοφανειών (likelihood ratio):

$$\frac{P(\mathbf{x},\mathbf{y} \mid M)}{P(\mathbf{x},\mathbf{y} \mid R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_j q_{y_j}} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$

Και αν δουλέψουμε σε λογαριθμική κλίμακα:

$$S = \sum_i \log\left(\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}\right) = \sum_i s(x_i, y_i)$$

Πίνακες ομοιότητας

Μπορούμε έτσι να ορίσουμε έναν πίνακα ομοιότητας με διαστάσεις όσο το μέγεθος του αλφαβήτου (4x4 για DNA, 20x20 για πρωτεϊνες), π.χ.:

$$s(x_i, y_i) = \begin{cases} 1, & \alpha\nu \quad x_i = y_i \\ -1, & \alpha\nu \quad x_i \neq y_i \end{cases}$$

Για τη μη-ταύτιση (mismatch), μπορόυμε να ορίσουμε μια πολύ μεγάλη ποινή ($-\infty$) έτσι ώστε να απαγορεύουμε πρακτικά την ταύτιση μη ομοίων καταλοίπων

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

Πίνακες αντικατάστασης (substitution matrices)

$$s_{ij} = \frac{1}{\lambda} \log\left(\frac{q_{ij}}{p_i p_j} \right)$$

- q_{ij}, είναι η πιθανότητα αντικατάστασης του i από το j σε σχετιζόμενες πρωτεΐνες (target frequencies)
- p_i, p_j είναι οι πιθανότητες εμφάνισης των αμινοξέων σε οποιαδήποτε θέση (background frequencies)
- λ είναι μια σταθερά κανονικοποίησης

Εντροπία των πινάκων αντικατάστασης

$$H = \sum_{i=1}^{20} \sum_{j=1}^{20} q_{ij} S_{ij}$$

- Η σχετική εντροπία εκφράζει το μέσο ποσό πληροφορίας που είναι διαθέσιμο για κάθε ζεύγος καταλοίπων που στοιχίζεται, και διαχωρίζει την προκύπτουσα στοίχιση από μια τυχαία στοίχιση που οφείλεται απλά στις συχνότητες υποβάθρου. Υψηλότερη τιμή της σχετικής εντροπίας συνεπάγεται εύκολο διαχωρισμό μεταξύ των συχνοτήτων στόχων και υποβάθρου.

Διάφοροι πίνακες αντικατάστασης

- PAM
- BLOSUM

PAM

- **P**oint **A**ccepted **M**utations (Dayhoff et al)
- Ως Αποδεκτή Σημειακή Μεταλλαγή σε μια πρωτεΐνη θεωρείται η αντικατάσταση ενός αμινοξικού καταλοίπου της με ένα κατάλοιπο διαφορετικού τύπου, η οποία έχει γίνει αποδεκτή μέσω της διαδικασίας της Φυσικής Επιλογής.
- Προέκυψε από πολλαπλή στοιχιση ακολουθιών με γνωστή εξελικτική σχέση και επίπεδο ομοιότητας >85%
- PAM1, PAM30, PAM250 κλπ
- Προυποθέτει ένα Μαρκοβιανό μοντέλο εξέλιξης
- Η χρήση πινάκων με μικρό N ενδείκνυται όταν οι εξεταζόμενες ακολουθίες είναι πολύ όμοιες (μικρή εξελικτική απόσταση), ενώ στην περίπτωση περισσότερο απομακρυσμένων ομοιοτήτων χρησιμοποιούμε πίνακες μεγαλύτερου N. Στις περιπτώσεις εκείνες κατά τις οποίες δε γνωρίζουμε εκ των προτέρων την ομοιότητα των προς σύγκριση ακολουθιών (π.χ. σε αναζητήσεις έναντι βάσεων δεδομένων) επιλέγουμε ένα ενδιάμεσο πίνακα, όπως τον PAM-250, ο οποίος αντιστοιχεί σε συντήρηση της τάξης του 20-25%.

BLOSUM

- **BLO**cks **SU**bstistution **M**atrcices (Henikoff and Henikoff)
- Προέκυψαν από πολλαπλές στοιχίσεις ακολουθιών με γνωστή κάθε φορά εξελικτική σχέση και διαφορετικό επίπεδο ομοιότητας
- Δεν προυποθέτουν ένα εξελικτικό μοντέλο αλλά το προσεγγίζουν εμπειρικά
- BLOSUM50, BLOSUM62, κλπ

PAM-1

PAM-250

BLOSUM100

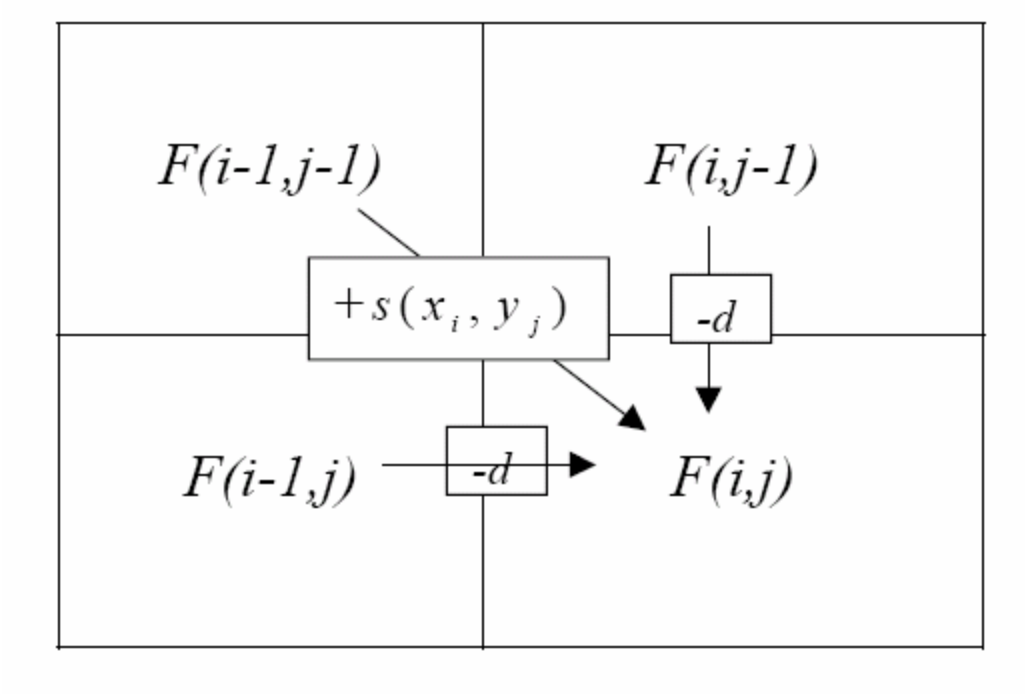
BLOSUM30

Small evolutionary distance
Strong similarity for short sequence

Large evolutionary distance
Weak similarity over stretched length

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

Δυναμικός προγραμματισμός



Ποινές για τα κενά (gap penalties)

Απλή ποινή για τα κενά:

$$\gamma(g) = -gd$$

Σύνθετη ποινή για τα κενά:

$$\gamma(g) = -d - (g-1)e$$

Ολική στοίχιση (Needleman and Wunsch, 1970)

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j), \\ F(i-1, j) - d, \\ F(i, j-1) - d \end{cases}$$

$F(i,0)=-id,$
$F(0,j)=-jd$

Παράδειγμα

Έστω δυο ακολουθίες:

$$\mathbf{x} = AAGTTAGCAG$$

$$\mathbf{y} = CAGTATCGCA$$

Αν έχουμε για τα κενά:

$$s(x_i, y_i) = \begin{cases} 1, & \text{αν} \quad x_i = y_i \\ -1, & \text{αν} \quad x_i \neq y_i \end{cases}$$

d=1

Τότε η καλύτερη ολική στοίχιση θα είναι:

```
A A G T - T A G C A G
C A G T A T C G C A -
```

συνέχεια…

	-	*A*	*A*	*G*	*T*	*T*	*A*	*G*	*C*	*A*	*G*
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
C	-1	-1	-2	-3	-4	-5	-6	-7	-6	-7	-8
A	-2	0	0	-1	-2	-3	-4	-5	-6	-5	-6
G	-3	-1	-1	1	0	-1	-2	-3	-4	-5	-4
T	-4	-2	-2	0	2	1	0	-1	-2	-3	-4
A	-5	-3	-1	-1	1	1	2	1	0	-1	-2
T	-6	-4	-2	-2	0	2	1	1	0	-1	-2
C	-7	-5	-3	-3	-1	1	1	0	2	1	0
G	-8	-6	-4	-2	-2	0	0	2	1	1	2
C	-9	-7	-5	-3	-3	-1	-1	1	3	2	1
A	-10	-8	-6	-4	-4	-2	0	0	2	4	3

A A G T – T A G C A G

C A G T A T C G C A –

Τοπική στοίχιση (Smith and Waterman, 1981)

$$F(i,j) = \max \begin{Bmatrix} F(i-1,j-1) + s(x_i, y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d, \\ 0 \end{Bmatrix}$$

$F(i,0)=0,$
$F(0,j)=0$

Η μόνη διαφορά από την ολική στοίχιση είναι το 0 το οποίο εξασφαλίζει διακοπή της στοίχισης όταν το score γίνει αρνητικό

Παράδειγμα

Στα δεδομένα του προηγούμενου παραδείγματος, θα έχουμε:

	-	A	A	G	T	T	A	G	C	A	G
-	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	1	0	0
A	0	1	1	0	0	0	1	0	0	2	1
G	0	0	0	2	1	0	0	2	1	1	3
T	0	0	0	1	3	2	1	1	1	0	2
A	0	1	1	0	2	2	3	2	1	2	1
T	0	0	0	0	1	3	2	2	1	1	1
C	0	0	0	0	0	2	2	1	3	2	1
G	0	0	0	1	0	1	1	3	2	2	3
C	0	0	0	0	0	0	0	2	4	3	2
A	0	1	1	0	0	0	1	1	3	5	4

A G T – T A G C A

A G T A T C G C A

Αλγοριθμική πολυπλοκότητα

Πρέπει εδώ να τονίσουμε ότι ο απαιτούμενος χρόνος για να τρέξουν οι παραπάνω αλγόριθμοι δυναμικού προγραμματισμού είναι ανάλογος του γινόμενου των μήκων των ακολουθιών και συμβολίζεται *O(*mn*)*. Το σύμβολο *O(*mn*) (big-O notation)* σημαίνει ότι μια συνάρτηση *f(t)=O(nm)* αν καθώς $t \to \infty$ υπάρχει σταθερά c τέτοια ώστε ,

$$|f(t)| \leq c.n.m$$

Σύνθετες ποινές για τα κενά

- Απαιτείται μια συνάρτηση $\gamma()$
- Τότε, οι παραπάνω αλγόριθμοι γίνονται:

$$F(i,j) = \max \begin{Bmatrix} F(i-1,j-1) + s(x_i, y_j), \\ F(k,j) - \gamma(i-k), k = 0,..,i-1 \\ F(i,k) - \gamma(j-k), k = 0,..,j-1 \end{Bmatrix}$$

$$F(i,j) = \max \begin{Bmatrix} F(i-1,j-1) + s(x_i, y_j), \\ F(k,j) - \gamma(i-k), k = 0,..,i-1 \\ F(i,k) - \gamma(j-k), k = 0,..,j-1 \\ 0 \end{Bmatrix}$$

Μειονέκτημα

•Η αλγοριθμική πολυπλοκότητα αυξάνει σε $O(n^3)$

•Ο Gotoh (1982), έδειξε ότι για σύνθετες συναρτήσεις του τύπου:

$$\gamma(g) = -d - (g-1)e$$

Μπορούμε να έχουμε πολυπλοκότητα της τάξης του $O(n^2)$ μόνο με αύξηση της μνήμης

Άλλοι αλγόριθμοι

- Υπάρχουν επίσης ειδικές περιπτώσεις στοίχισης (π.χ. προσαρμογή)
- Θέλουμε δηλαδή να εντοπίσουμε, μια μικρή ακολουθία αν συναντάται σε μια μεγαλύτερη

Έστω ότι θέλουμε να ανιχνεύσουμε αν στην αλληλουχία του γονιδίου *lacI* της *E.coli* υπάρχει η γνωστή αλληλουχία του υποκινητή (promoter). Έστω ακόμα ότι το τμήμα του γονιδίου έχει αλληλουχία:

$\mathbf{x} = TCGCGGTATGGCATGATAGCGCCCGGAA$

και η αλληλουχία του υποκινητή είναι

$\mathbf{y} = TATAAT$

συνέχεια…

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j), \\ F(i-1, j) - d, \\ F(i, j-1) - d \end{cases}$$

$F(i,0) = -id$

$F(0,j) = 0$

	T	C	G	C	G	G	T	A	T	G	G	C	A	T	G	A	T	A	G	C	G	C	C	C	G	G	A	A
T	1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
A	0	0	-2	-2	-2	-2	-1	2	0	0	-2	-2	0	-1	0	0	-1	2	0	-2	-2	-2	-2	-2	-2	-2	0	0
T	1	-1	-1	-3	-3	-3	-1	0	3	1	-1	-3	-2	1	-1	-1	1	0	1	-1	-3	-3	-3	-3	-3	-3	-2	-1
A	-1	0	-2	-2	-4	-4	-3	0	1	2	0	-2	-2	-1	0	0	-1	2	0	0	-2	-4	-4	-4	-4	-4	-2	-1
A	-3	-2	-1	-3	-3	-5	-5	-2	-1	0	1	-1	-1	-3	-2	1	-1	0	1	-1	-1	-3	-5	-5	-5	-5	-3	-1
T	-3	-4	-3	-2	-4	-4	-4	-4	-1	-2	-1	0	-2	0	-2	-1	2	0	-1	0	-2	-2	-4	-6	-6	-6	-5	-3

Και η ακολουθία του πιθανού υποκινητή είναι:

C A T G A T

Ευριστικοί αλγόριθμοι στοίχισης (Heuristic alignment algorithms)

- Είναι αναγκαίοι για τη μείωση του απαιτούμενου υπολογιστικού χρόνου, ειδικά σε αναζητήσεις σε βάσεις δεδομένων

Απαραίτητα χαρακτηριστικά τους:

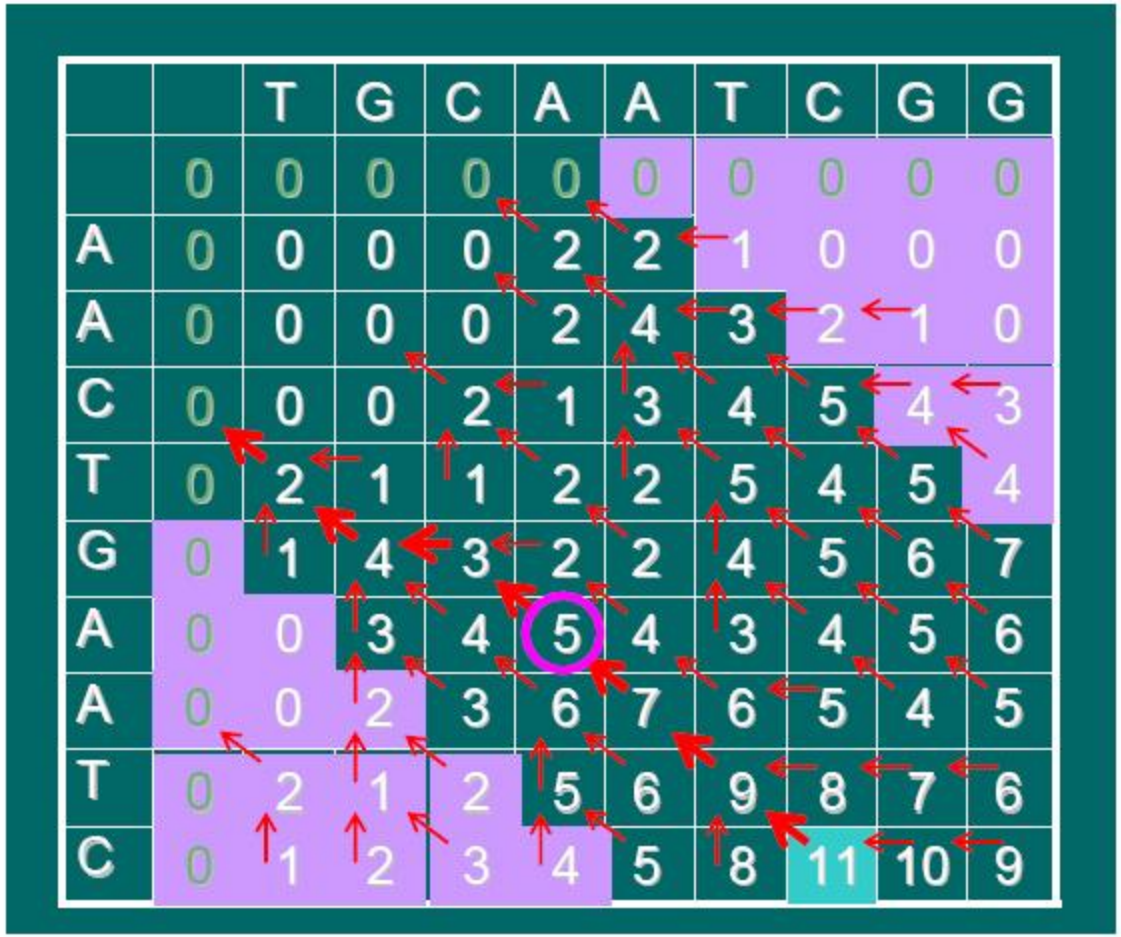
- Να μη διαφέρουν σημαντικά από τις «ακριβείς» (μαθηματικά βέλτιστες) λύσεις των μεθόδων δυναμικού προγραμματισμού.
- Να μην αποκλείουν βιολογικά πιθανές λύσεις.

Βασικές κατηγορίες τέτοιων αλγορίθμων:

- Μέθοδος «κοπής γωνιών» (banded alignment)
- Μέθοδος FASTA
- Μέθοδος BLAST

Μέθοδος «κοπής γωνιών»

- Αυτή είναι ίσως η απλούστερη «βελτίωση» που θα μπορούσε να σκεφτεί κανείς. Η ιδέα είναι πραγματικά πολύ έξυπνη και απλή, περιορίζοντας στην ουσία τους υπολογισμούς των πινάκων Δυναμικού Προγραμματισμού σε μια «ζώνη» γύρω από τη διαγώνιο του πίνακα. Όπως γίνεται εμφανές, η επιλογή του πλάτους της ζώνης στην οποία θα εκτελεστούν οι υπολογισμοί επηρεάζει άμεσα την εξοικονόμηση πόρων κατά τη στοίχιση ακολουθιών.
- Μπορεί να δώσει μια «οικονομία» υπολογιστικών πόρων της τάξης του 30%.
- Σε ακραίες περιπτώσεις



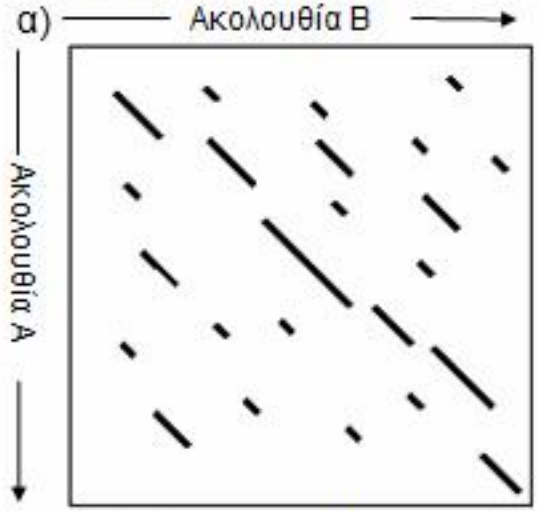
*Εικόνα 3: Πίνακας Δυναμικού Προγραμματισμού για τη μέθοδο «Κοπής Γονιών». Οι τιμές όλων των κελιών έχουν τοποθετηθεί στα κελιά (δείτε σημειώσεις προηγούμενης διάλεξης). Με τη μέθοδο αυτή, υποθέτουμε ότι ένα «καλό μονοπάτι» (δηλ. μια καλή στοίχιση) δεν αναμένουμε να διέρχεται από τις σκιασμένες περιοχές του πίνακα (πάνω δεξιά και κάτω αριστερή γωνία). Με τα παχιά βέλη υποδηλώνεται η βέλτιστη διαδρομή, όπως υπολογίζεται με τον κλασικό Δυναμικό Προγραμματισμό. Παρατηρήστε ότι από τα 100 (=10*10) κελιά του πίνακα απαιτείται το γέμισμα μόνο των 70, κερδίζοντας έτσι 30% σε μνήμη (και χρόνο).*

Μέθοδος FASTA

- Η βασική ιδέα έγκειται στη δημιουργία ενός ευρετηρίου με τις θέσεις όλων των k-tuples (τυπικό μήκος για αμινοξικές ακολουθίες 1 ή 2) που υπάρχουν και στις δύο ακολουθίες (Εικόνα 4, αριστερά).
- Από τη διαφορά των θέσεών τους στις δύο ακολουθίες εντοπίζεται η διαγώνιος στην οποία βρίσκονται (Εικόνα 4, δεξιά), οπότε στο επόμενο βήμα εντοπίζονται οι διαγώνιες με τα περισσότερα k-tuples.
- Ακολούθως, αυτές οι περιοχές ταύτισης συνενώνονται επιτρέποντας την εισαγωγή κενών με τον υπολογισμό της αντίστοιχης ποινής (Εικόνα 5), και
- Τελικά πραγματοποιείται η διαδικασία πλήρους δυναμικού προγραμματισμού (με τον επιλεγμένο πίνακα αντικατάστασης), περιορισμένου σε μια ταινία γύρω από τις συγκεκριμένες διαγωνίους (Εικόνα 5).

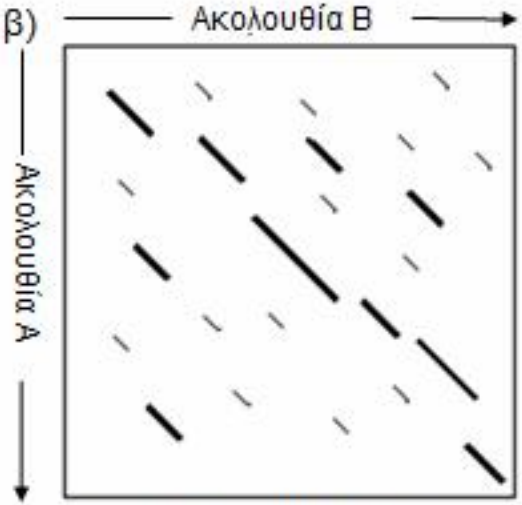
Αλγόριθμος FASTA

α)



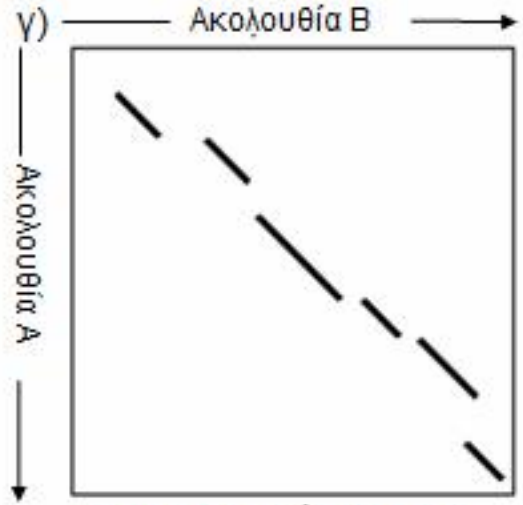
Εύρεση επαναλήψεων ταυτόσημων λέξεων

β)



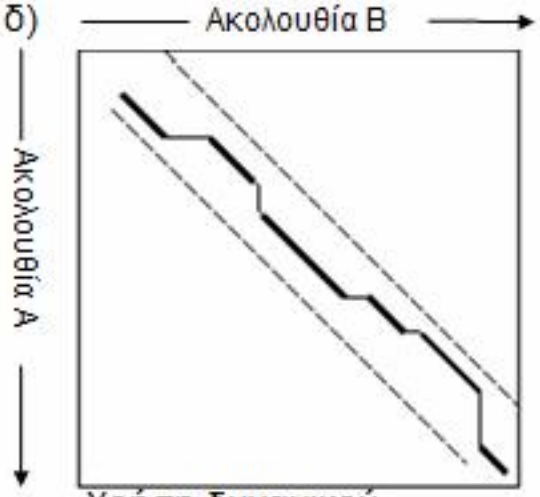
Επαναβαθμολόγηση με τη χρήση του πίνακα PAM. Διατήρηση των τμημάτων με την υψηλότερη βαθμολόγηση.

γ)



Ένωση τμημάτων με τη χρήση κενών, εξάλειψη άλλων τμημάτων

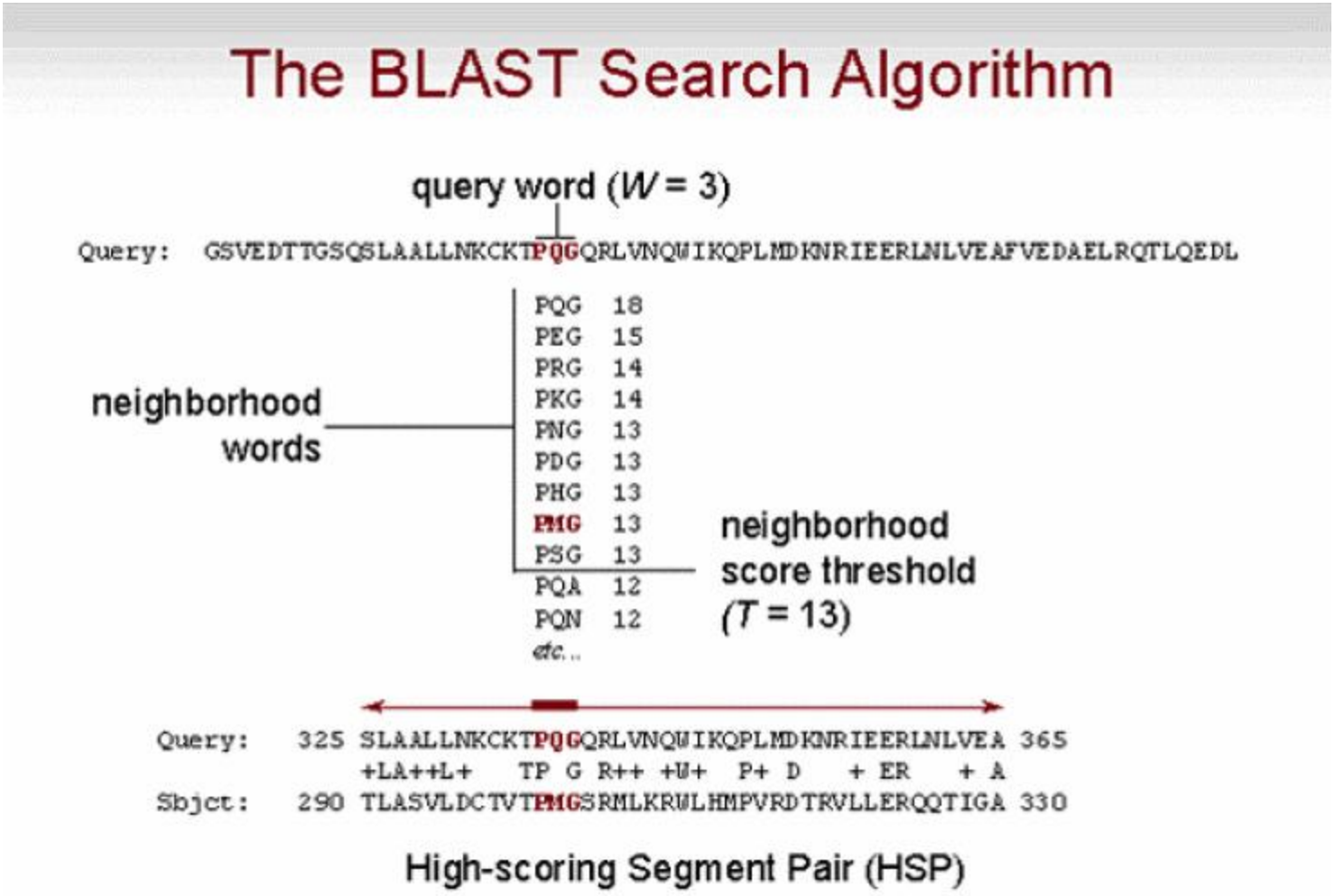
δ)



Χρήση δυναμικού προγραμματισμού για τη διαμόρφωση της βέλτιστης στοίχισης

Μέθοδος BLAST

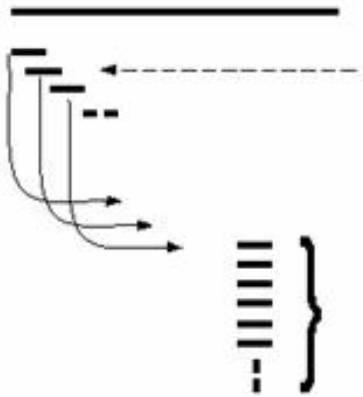
- Η διαδικασία της σύγκρισης ξεκινά με την κατασκευή ενός καταλόγου όλων των λέξεων που θα ταίριαζαν με κάποια λέξη της άγνωστης ακολουθίας ξεπερνώντας την τιμή κατωφλίου (προκαθορισμένη τιμή για πρωτεϊνικές ακολουθίες T=13).
- Στη συνέχεια, ο αλγόριθμος αναζητά αυτές τις λέξεις στις ακολουθίες της βάσης δεδομένων και κάθε φορά που εντοπίζει κάποια ξεκινάει μια διαδικασία επέκτασης του 'ευρήματος' προς τις δύο κατευθύνσεις, όσο η βαθμολογία συνεχίζει και αυξάνει.
- Οι περιοχές μέγιστης βαθμολογίας που εντοπίζονται σε αυτό το στάδιο είναι οι υποψήφιες περιοχές ομοιότητας (HSPs, high scoring pairs).
- Από όλα τα HSPs αναφέρονται στα αποτελέσματα εκείνες οι περιοχές στις οποίες η βαθμολογία υπερβαίνει μια δεύτερη τιμή κατωφλίου S
- Τελικά, επιλέγονται να αναφερθούν εκείνες μόνο οι τοπικές ομοιότητες οι οποίες εμφανίζουν υψηλή στατιστική σημαντικότητα, ο προσδιορισμός της οποίας περιγράφεται στην επόμενη ενότητα.



The BLAST algorithm. The BLAST algorithm is a heuristic search method that seeks words of length W (default = 3 in blastp) that score at least T when aligned with the query and scored with a substitution matrix. Words in the database that score T or greater are extended in both directions in an attempt to fina a locally optimal ungapped alignment or HSP (high scoring pair) with a score of at least S or an E value lower than the specified threshold. HSPs that meet these criteria will be reported by BLAST, provided they do not exceed the cutoff value specified for number of descriptions and/or alignments to report.

Αλγόριθμος BLAST

1) Εύρεση των λέξεων μήκους w με την υψηλότερη βαθμολόγηση για την αναζήτηση



Αναζήτηση αλληλουχίας μήκους L

Μέγιστο των L-w+1 λέξεων
(συνήθως w=3 για πρωτεΐνες)

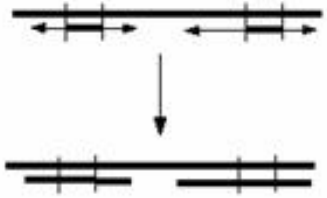
Εύρεση, για κάθε λέξη από την αναζήτηση αλληλουχιών, του καταλόγου των λέξεων με βαθμολόγηση τουλάχιστον T με τη χρήση πίνακα αντικατάστασης (για παράδειγμα PAM 250). Για τις συνήθεις παραμέτρους υπάρχουν περίπου 50 λέξεις ανά κατάλοιπο της αναζήτησης.

2) Σύγκριση του καταλόγου των λέξεων με τη βάση δεδομένων για την εύρεση ταύτισης

Αλληλουχίες της βάσης δεδομένων

Ταύτιση λέξεων από τον κατάλογο λέξεων

3) Για κάθε ταύτιση μιας λέξης, γίνεται επέκταση της στοίχισης και προς τις δυο κατευθύνσεις για την εύρεση στοιχίσεων με βαθμολόγηση μεγαλύτερη από το κατώφλι S



Maximal Segment Pairs (MSPs)

Στατιστική σημαντικότητα των στοιχίσεων

- Αν λαβουμε με οποιοδήποτε τρόπο μια στοίχιση δυο ακολουθιών, θέλουμε να έχουμε έναν τρόπο να την αξιολογήσουμε (να ξέρουμε δηλαδή αν είναι στατιστικά σημαντική)
- Ιδιαίτερο νόημα έχει αυτό σε μια αναζήτηση σε μεγάλες βάσεις δεοδομένων όπου αναμένουμε να δούμε έως και εκατοντάδες «ομόλογες» ακολουθίες
- Χρειαζόμαστε έναν έλεγχο υποθέσεων.

  Ho: οι δυο ακολουθίες είναι ασυσχέτιστες,

  Ha: οι δυο ακολουθίες σχετίζονται με κάποιο τρόπο (είναι ομόλογες)
- Ακόμα και αν βρεθεί στατιστικά σημαντική ομοιότητα, δεν σημαίνει ότι υπάρχει και βιολογική συσχέτιση των ακολουθιών, και το αντίστροφο (εξαρτάται από τις παραμέτρους, gap penalty, substitution matrix, αλγοριθμο στοίχισης κλπ)
- Τα πιο πολλά αποτελέσματα αναφέρονται στην **τοπική στοίχιση**

Ασυμπτωτικά αποτελέσματα

Θεώρημα (Waterman, 1995)

Έστω ότι έχουμε δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, ..., x_n$ *και* $\mathbf{y} = y_1, y_2, ..., y_m$. *Τότε η μέγιστη περιοχή σύμπτωσης (match) μεταξύ τους είναι* $M_n \cong \log_{1/p}(mn)$ *ή αλλιώς:*

$$\frac{M_n}{\log_{1/p}(mn)} \to 1 \text{ με πιθανότητα } 1.$$

Προφανώς η πιθανότητα σύμπτωσης p είναι ίση με $p=P(x_i=y_j) \Leftrightarrow$

$p = p_A^2 + p_T^2 + p_G^2 + p_C^2$ αν η κατανομή των βάσεων στις δυο αλληλουχίες είναι ίδια.

Θεώρημα (Waterman, 1995)

Έστω δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, ..., x_n$ *και* $\mathbf{y} = y_1, y_2, ..., y_m$ *με* $0 \le p < a \le 1$. *Τότε για τη μέγιστη περιοχή που περιέχει 100α% όμοια νουκλεοτίδια μεταξύ τους ισχύει*

$$\frac{M_n}{\log_{1/p}(mn)} \to \frac{1}{H(a,p)} \text{ με πιθανότητα 1.}$$

Θεώρημα (Arratia et al, 1990)

Έστω δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, ..., x_n$ *και* $\mathbf{y} = y_1, y_2, ..., y_m$. *Τότε η μέση τιμή για το μήκος της μέγιστης περιοχής σύμπτωσης (match) μεταξύ τους είναι:*

$$E(M_n) \approx \frac{\log(mn)}{\lambda} + \frac{\log(q)}{\lambda} + \frac{\gamma}{\lambda} - \frac{1}{2}$$

όπου q=1-p, και γ = -Γ'(1) = 0.5772... η σταθερά Euler-Mascheroni, και λ=log(1/p)

Για την αντίστοιχη διασπορά ισχύει :

$$Var(M_n) \approx \frac{\pi^2}{6\lambda^2} + \frac{1}{12}$$

Θεώρημα (Arratia and Waterman, 1989; Waterman, 1995)

Έστω ότι έχουμε δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, ..., x_n$ *και* $\mathbf{y} = y_1, y_2, ..., y_m$. *Τότε η μέση τιμή για το μήκος της μέγιστης περιοχής σύμπτωσης (match) μεταξύ τους, όταν υπάρχουν k μη κοινά νουκλεοτίδια (k mismatches) είναι:*

$$E(M_n) \approx \log_{1/p}(qn^2) + k\log_{1/p}\log_{1/p}(qn^2) + k\log_{1/p}(q) - \log_{1/p}(k!) + k + \frac{\gamma}{\lambda} - \frac{1}{2}$$

Για την αντίστοιχη διασπορά ισχύει :

$$Var(M_n) \approx \frac{\pi^2}{6\lambda^2} + \frac{1}{12}$$

Όπως και παραπάνω, *q=1-p*, και *γ=-Γ'(1)=0.5772...*η σταθερά *Euler-Mascheroni,* και *λ=log(1/p)*

Η κατανομή του Local Similarity Score

- Σε όλες τις τοπικές στοιχίσεις χωρίς κενά, η κατανομή του score είναι η κατανομή των ακραίων τιμών του Gumbel
- Αν υπάρχουν κενά, η κατανομή φαίνεται να συγκλίνει (υπο προυποθέσεις) σε αυτή του Gumbel χωρίς όμως αυτό να μπορεί να αποδειχθεί
- Σε ολικές στοιχίσεις δεν ισχύει τίποτα από τα παραπάνω

Η κατανομή του Local Similarity Score

Δυο ακραίες περιπτώσεις:

$$s(x_i, y_i) = \begin{cases} 1, \text{αν} & x_i = y_i \\ 0, \text{αν} & x_i \neq y_i \end{cases} \text{ και d=0} \qquad s(x_i, y_j) \sim c.n$$

Γραμμική περιοχή

$$s(x_i, y_i) = \begin{cases} 1, & \text{αν} \quad x_i = y_i \\ -\infty, \text{αν} & x_i \neq y_i \end{cases} \text{ και d=}\infty \qquad s(x_i, y_j) \sim k.\log n$$

Λογαριθμική περιοχή

Στη δεύτερη περίπτωση η κατανομή είναι αποδεδειγμένα αυτή του Gumbel, αλλά όταν μπαίνουν κενά δεν υπάρχει τέτοια απόδειξη

Μειώνοντας σταδιακά τις ποινές για διαφορές και κενά, μεταπίπτουμε από τη λογαριθμική περιοχή του score στη γραμμική. Αυτή η μετάπτωση φάσεως (phase transition) έχει περιγραφεί αναλυτικά από τους Arratia, Gordon και Waterman (Waterman et al, 1987; Arratia and Waterman, 1994; Waterman, 1995) αλλά παρ' όλα αυτά δεν υπάρχει αναλυτική έκφραση για τις τιμές των παραμέτρων m (mismatch) και d (gap) στις οποίες συμβαίνει αυτή η μετάπτωση (μπορούν να προσεγγισθούν μόνο με αριθμητικές μεθόδους)

Η κατανομή του Local Similarity Score

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

Θεώρημα **(Karlin and Altschul, 1990)**

Έστω ότι έχουμε δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, \ldots, x_n$ *και* $\mathbf{y} = y_1, y_2, \ldots, y_m$ *και το score S*

Τότε: $P\{S > x\} \approx 1 - \exp\{-Kmne^{-\lambda x}\}$

Τουλάχιστον ένα score πρέπει να είναι θετικό

Η αναμενόμενη τιμή του score για κάθε βάση να είναι αρνητική, δηλαδή

$$E(s_{ij}) = \sum q_i q_j s_{ij} = \sum q_i q_j \log\left(\frac{q_i q_j}{p_{ij}}\right) < 0$$

Το λ είναι όπως είπαμε ήδη, η μοναδική θετική ρίζα της εξίσωσης:

$$\sum q_i q_j e^{\lambda S} = 1$$

Προφανώς οι παραπάνω δυο περιορισμοί είναι απαραίτητοι για να είμαστε σίγουροι ότι το score θα παίρνει τιμές στη λογαριθμική περιοχή, και κατά συνέπεια θα είναι όντως τοπικό. Αν δεν ισχύουν οι παραπάνω προϋποθέσεις, τότε το score θα παίρνει τιμές στη γραμμική περιοχή και κατά συνέπεια θα μιλάμε για ολική στοίχιση.

Η πιθανότητα να υπάρχει μια κοινή υπο-ακολουθία με μήκος μεγαλύτερο από x , όπως είπαμε παραπάνω είναι (ασυμπτωτικά):

$$P\left(S > x = \log_{1/p}(mn) + T\right) = 1 - e^{-E(S)} = 1 - \exp\left(-Kmne^{-\lambda x}\right)$$

οπότε $P(S \le x) = \exp(-Kmne^{-\lambda x})$

Η τελευταία σχέση είναι η α.σ.κ. της κατανομής των ακραίων τιμών του Gumbel (EVD).

$$P(S \le x) = \exp(-e^{-\frac{(x-a)}{b}}), -\infty \le x \le \infty$$

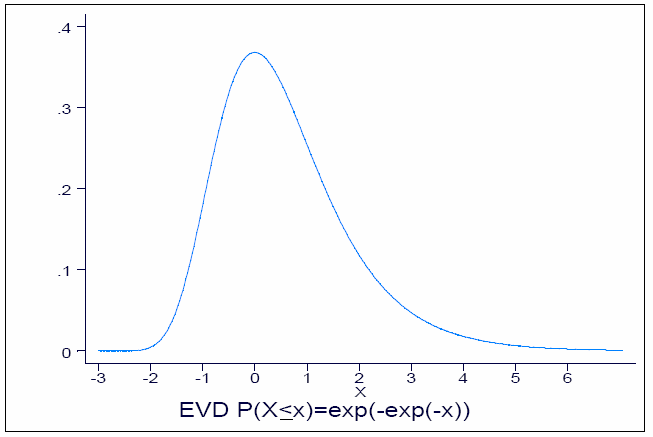
με

$$E(x) = a - b\Gamma'(1) \text{ και } V(x) = \frac{b^2\pi^2}{6}.$$

Οι παράμετροι a, b είναι προφανώς $a = \frac{\log(kmn)}{\lambda}, b = \frac{1}{\lambda}$ με $\lambda = \log\left(1/p\right)$ και K=1-p=q,

όταν δεν επιτρέπονται διαφορές. Από τις παραπάνω σχέσεις είναι δυνατόν να υπολογιστεί το p-value για ένα δεδομένο score που προέκυψε από την σύγκριση δυο ακολουθιών. Αφού τυποποιήσουμε τη μεταβλητή μας έχουμε (Pearson, 1998; Pearson and Wood, 2001):

$$P(Z \ge z) = 1 - \exp\left(-\exp-\left(\left(\pi/\sqrt{6}\right)z - \Gamma'(1)\right)\right)$$



Όταν συγκρίνουμε μια ακολουθία με μια ολόκληρη βάση δεδομένων, η οποία περιέχει D ακολουθίες, τότε η παρατήρηση ακολουθιών οι οποίες εμφανίζουν μικρό p-value (μεγάλη ομοιότητα- p-match) είναι σπάνιο ενδεχόμενο, και θα περιγράφεται από την κατανομή Poisson. Άρα (Pearson and Wood, 2001):

$$P = \Pr(\text{τουλαχιστον 1 score } S \geq x) = 1 - e^{-Dp}$$

και αν το Dp είναι πολύ μικρό (<0.01) θα έχουμε :

$$P \approx Dp\,.$$

Στο ίδιο αποτέλεσμα θα καταλήγαμε αν υπολογίζαμε την αναμενόμενη τιμή για τις εμφανίσεις περιοχών με score $S \geq x$, έπειτα από D συγκρίσεις με τις ακολουθίες της βάσης δεδομένων. Αυτό το E-value (expectation value) είναι ίσο με $E(S \geq x) = D.P(S \geq x)$ όπου D είναι ο αριθμός των ανεξάρτητων ακολουθιών που περιέχει η υπό έλεγχο βάση δεδομένων.

Για να έχουμε περισσότερο ακριβή αποτελέσματα, μια πιο σωστή προσέγγιση θα προέκυπτε αν λαμβάναμε υπόψη το γεγονός ότι όλες οι ακολουθίες στη βάση δεδομένων δεν έχουν τον ίδιο αριθμό βάσεων. Πρακτικά αυτό σημαίνει ότι θεωρούμε ολόκληρη τη βάση δεδομένων ως μια τεράστια ακολουθία από N νουκλεοτίδια (βάσεις) και συγκρίνουμε με αυτήν τη συγκεκριμένη ακολουθία μας η οποία έχει μήκος n βάσεις. Κατά μέσο όρο κάθε μια από τις ακολουθίες της βάσης περιέχει m=N/D βάσεις, οπότε η πιθανότητα να υπάρχει μια περιοχή με score μεγαλύτερο από x , όπως είπαμε παραπάνω είναι :

$$P(S > x) = 1 - e^{-E(S)} = 1 - \exp\left(-KNne^{-\lambda x}\right)$$

ενώ η αναμενόμενη τιμή (E-value) θα είναι:

$$E(S \geq x) = KNne^{-\lambda x} = DKmne^{-\lambda x}.$$

Πολλά προγράμματα όπως το BLAST (Altschul et al, 1990), αντί του p-value, αναφέρουν ως αποτέλεσμα (output) αυτή την τιμή, επειδή είναι πιο εύκολη η ερμηνεία της από κάποιο μη ειδικό, αλλά όπως είδαμε όταν το E-value είναι πολύ μικρό τότε, επειδή ισχύει η προσεγγιστική σχέση (Waterman, 1995):

$$1 - \exp(-\exp(-t)) \approx 1 - (1 - \exp(-t)) = \exp(-t),$$

το p-value θα είναι περίπου ίσο με το E-value. Είναι φανερό ότι σήμερα που οι βάσεις δεδομένων αυξάνονται σε μέγεθος συνεχώς είναι καλύτερο κάθε φορά που γίνονται τέτοιες συγκρίσεις να αναφέρονται τουλάχιστον μαζί το p-value και το e-value, και τέτοια παραδείγματα θα δούμε σε παρακάτω κεφάλαια.

Κάτι άλλο που πρέπει να τονιστεί είναι ότι, λόγω του γεγονότος ότι πολλές φορές χρησιμοποιούνται διαφορετικά σχήματα για το score (gap penalties, mismatches), είναι αναγκαίο να αναφέρεται και μια αντικειμενική τιμή για το score. Αυτό μπορεί να επιτευχθεί κανονικοποιώντας το score όπως είδαμε και σε προηγούμενα κεφάλαια με βάση το bit (Altschul et al, 1990; Altschul et al, 1997)):

$$S_{bit} = \frac{\lambda S_{raw} - \log K}{\log 2}$$

όπου S_{raw}, είναι το score που υπολογίστηκε με κάποιες συγκεκριμένες τιμές για κενά και διαφορές. Αντικαθιστώντας τώρα στην σχέση (4.15) θα έχουμε

$$E(S_{bit}) = m.n.2^{-S_{bit}} .$$

$$m' = m - \frac{\log(kmn)}{H} \text{ και } n' = n - \frac{\log(kmn)}{H}$$

δηλαδή το λειτουργικό μήκος της ακολουθίας και της βάσης δεδομένων προσαρμόζεται (μειώνεται), για να λάβει υπόψη το γεγονός ότι με αυτά τα μήκη και τον δεδομένο πίνακα (substitution matrix) δεν επιτρέπονται όλες οι στοιχίσεις. Το H είναι η σχετική εντροπία του πίνακα για τη δεδομένη σύσταση και το μήκος των ακολουθιών που συγκρίνονται.

Η κατανομή όταν υπάρχουν κενά

- Η μέθοδος του Mott (1992)
- Η μέθοδος Direct Estimation (Waterman, 1995)
- Η μέθοδος Poisson declumping (Waterman and Vingron, 1994)
- Η μέθοδος weighted regression του Pearson (1995)

Η μέθοδος του Mott (1992)

- Παραλλαγή της εκτίμησης στην κατανομή του Gumbel

$$P(S \le x) = \exp(-e^{\frac{(x-A)}{B}})$$

όπου :

$$A = a_0 + \frac{a_1}{\lambda} + \frac{a_2 \log(mn)}{\lambda}, \quad B = \frac{b_1}{\lambda}.$$

Το λ είναι και πάλι η μοναδική θετική ρίζα της εξίσωσης:

$$\sum q_i q_j e^{\lambda S} = 1$$

Η μέθοδος Direct Estimation (Waterman, 1995)

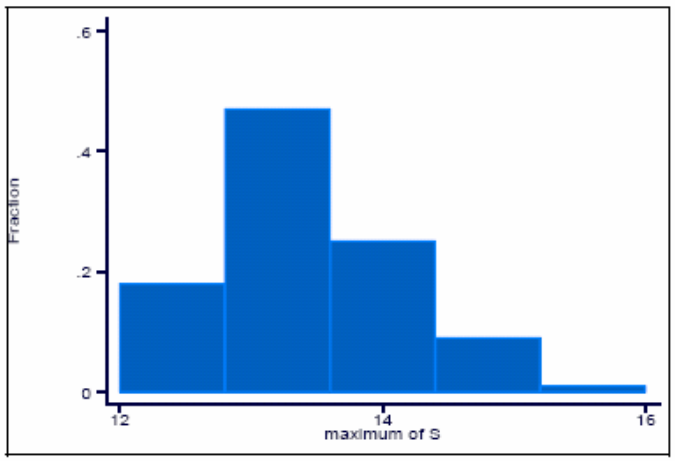
- Πραγματοποιεί Maximum Likelihood fit, σε εμπειρικά δεδομένα
- Απαιτεί αποτελέσματα από πολλές αναζητήσεις
- Απλή στην εκτέλεση (linear regression)

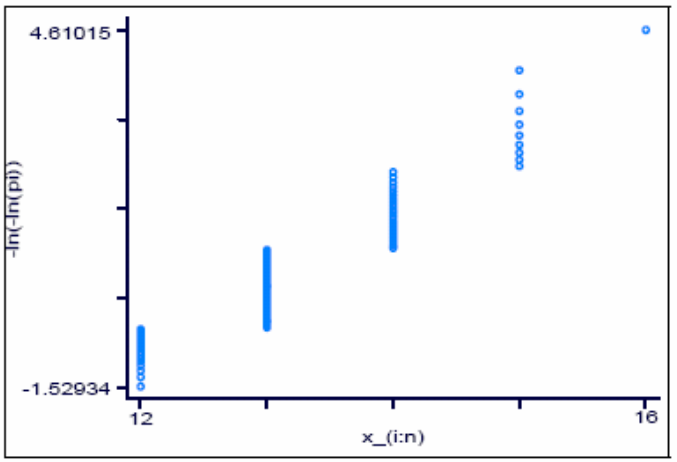
$$P(S \le x) = \exp(-Kmne^{-\lambda x})$$

$$\log P(S \le x) = -Kmne^{-\lambda x} \Leftrightarrow$$

$$\log\left(-\log P(S \le x)\right) = \log\left(Kmne^{-\lambda x}\right) \Leftrightarrow$$

$$\log\left(-\log P(S \le x)\right) = -\lambda x + \log\left(Kmn\right)$$





Παραλλαγές

- Η αναζήτηση μπορεί να γίνει σε τυχαίες ακολουθίες με προκαθορισμένη σύνθεση
- Η αναζήτηση μπορεί να γίνει σε shuffled ακολουθίες με σύνθεση όμοια με αυτή της ακολουθίας εισόδου
- Αν πρόκειται για αναζήτηση σε βάση δεδομένων μπορεί να χρησιμοποιηθούν τα αποτελέσματα της αναζήτησης (αφου απομακρυνθούν οι πολύ όμοιες και οι πολύ ανόμοιες ακολουθίες)
- Χρειάζονται το λιγότερο 100-1000 ακολουθίες, άρα είναι χρονοβόρα διαδικασία

Η μέθοδος Poisson declumping (Waterman and Vingron, 1994)

- Παραλλαγή της προηγούμενης μεθόδου
- Πολύ πιο αποδοτική και γρήγορη
- Στηρίζεται στην προσέγγιση Poisson declumping
- Για κάθε ακολουθία χρησιμοποιεί το διατεταγμένο δείγμα:

$$S_{(1)} \geq S_{(2)} \geq \ldots \geq S_{(k)}$$

και όχι μονο το μεγιστο

- Τα score από κάθε ακολουθία ακολουθούν κατανομή Poisson:

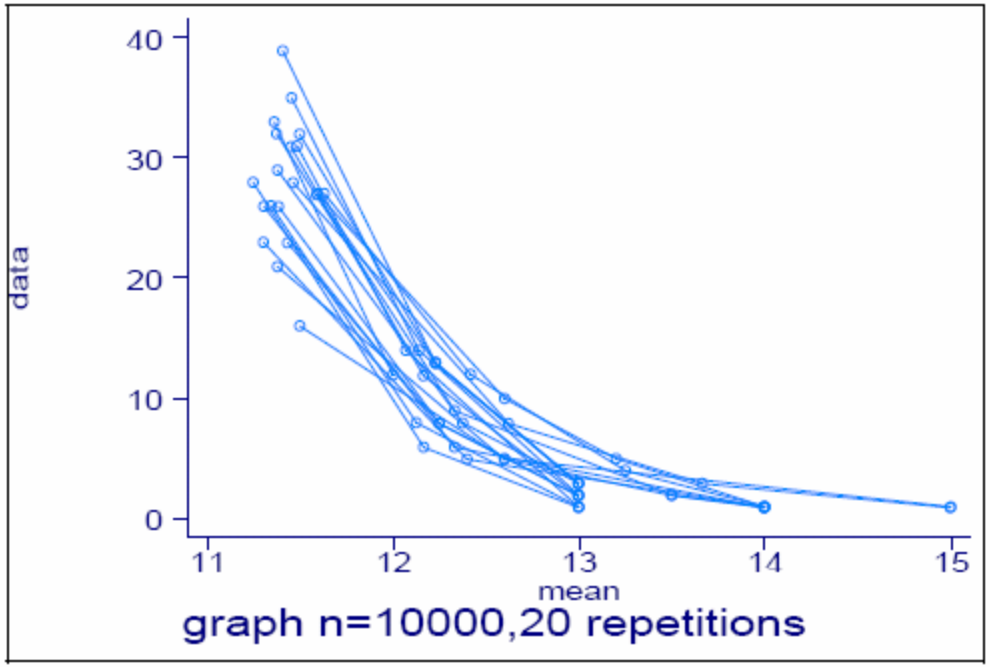
$$E(S \geq x) = Kmne^{-\lambda x} .$$

- Άρα η πιθανότητα να υπάρχουν k περιοχές με score>x θα είναι:

$$P(S_{(k)} > x) \approx 1 - \exp\left(-Kmne^{-\lambda x}\right) \sum_{i=0}^{k-1} \frac{\left(Kmne^{-\lambda x}\right)^i}{i!}$$

συνέχεια…

- Επομένως, παριστάνοντας γραφικά το λογάριθμο του αριθμού τοπικών περιοχών με score πάνω από κάποιο όριο σε σχέση με τη μέση τιμή του score για τις περιοχές πάνω από το όριο αυτό παίρνουμε ευθεία γραμμή και μια απλή γραμμική παλινδρόμηση δίνει αμέσως εκτιμήτριες για τα K,λ.
- Απαιτεί πολύ λιγότερες ακολουθίες (~10-20), άρα είναι πολύ πιο γρήγορη μέθοδος



Η μέθοδος weighted regression του Pearson (1995)

- Χρησιμοποιείται σε αναζητήσεις σε βάσεις δεδομένων
- Η βάση δεδομένων χωρίζεται σε *k* υποσύνολα σύμφωνα με το μήκος των ακολουθιών $n_1, n_2,\ldots, n_k$
- Υπολογίζονται όλα τα score S, για την τοπική ομοιότητα των ακολουθιών και στη συνέχεια μια ευθεία σταθμισμένης γραμμικής παλινδρόμησης (weighted linear regression) για τη σχέση:

$$S = a + b\log(n_i).$$

- Όπου n_i, είναι το μήκος των ακολουθιών του i υποσυνόλου της βάσης δεδομένων, ενώ το $\log(n_i)$ είναι σταθμισμένο με την αντίστροφη διασπορά (1/var) των scores σε αυτό το υποσύνολο, καθώς τμήματα με πολύ μεγάλο score θα έχουν και μεγάλη διασπορά. Υπολογίζεται τέλος η εκτιμήτρια της διασποράς , των υπολοίπων της παλινδρόμησης (residual variance) η οποία καθορίζει το z-score.

$$z-score = \frac{S-(a+b\log(n_i)}{\mathrm{var}}$$

$$P(Z \geq z) = 1 - \exp\left(-\exp-\left(\left(\pi/\sqrt{6}\right)z - \Gamma'(1)\right)\right)$$

Διαθέσιμο Software

- SW (http://www-hto.usc.edu/software/seqaln/seqaln-query.html)
- BLAST (www.ncbi.nlm.nih.gov/BLAST/)
- WU-BLAST (http://blast.wustl.edu/)
- FASTA (www.ebi.ac.uk/fasta33/)