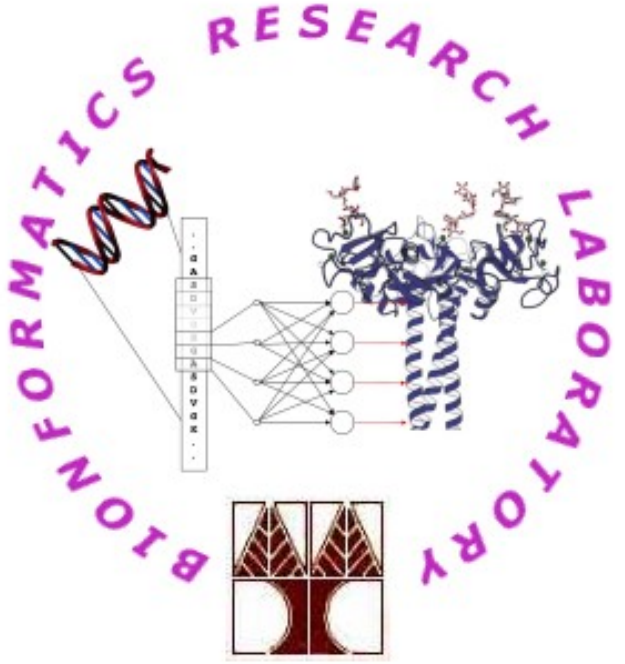
Προγνωστικές μέθοδοι με βάση αλληλουχίες DNA

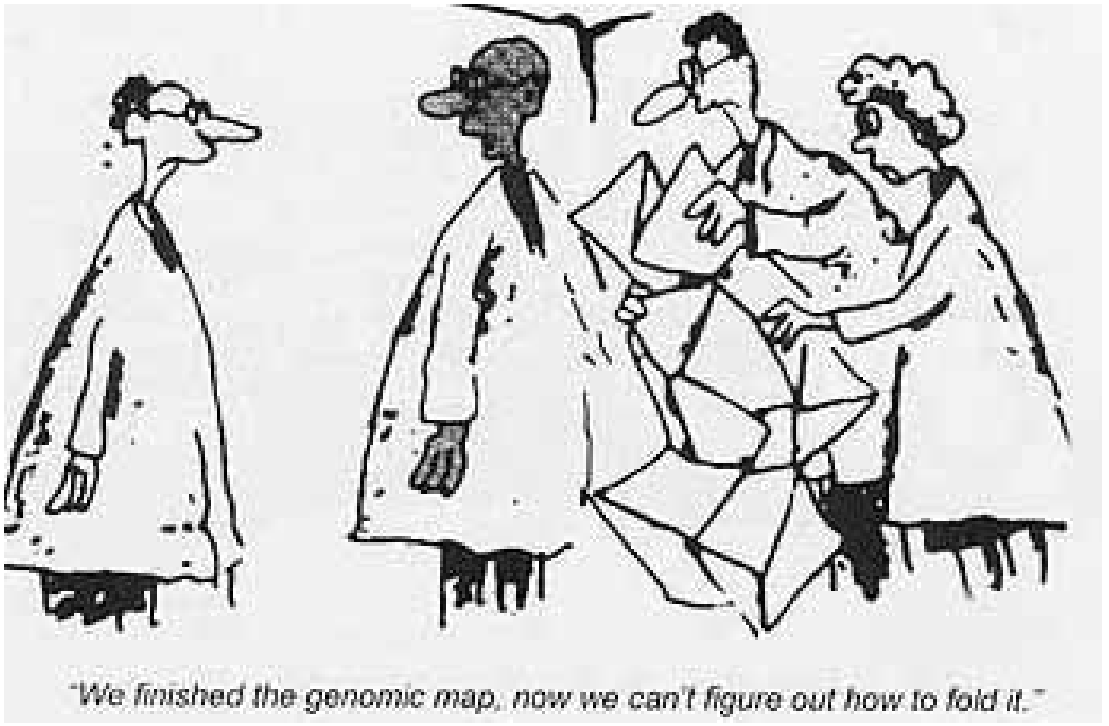


Vasilis Promponas
Bioinformatics Research Laboratory
Department of Biological Sciences
University of Cyprus

ΣΥΝΟΨΗ

- Εισαγωγή
- Αλυσίδες Markov και αλληλουχίες DNA
- Μέτρα κωδικοποίησης αλληλουχιών DNA
- Πρόβλεψη δομής (ευκαρυωτικών) γονιδίων
- Συζήτηση ..

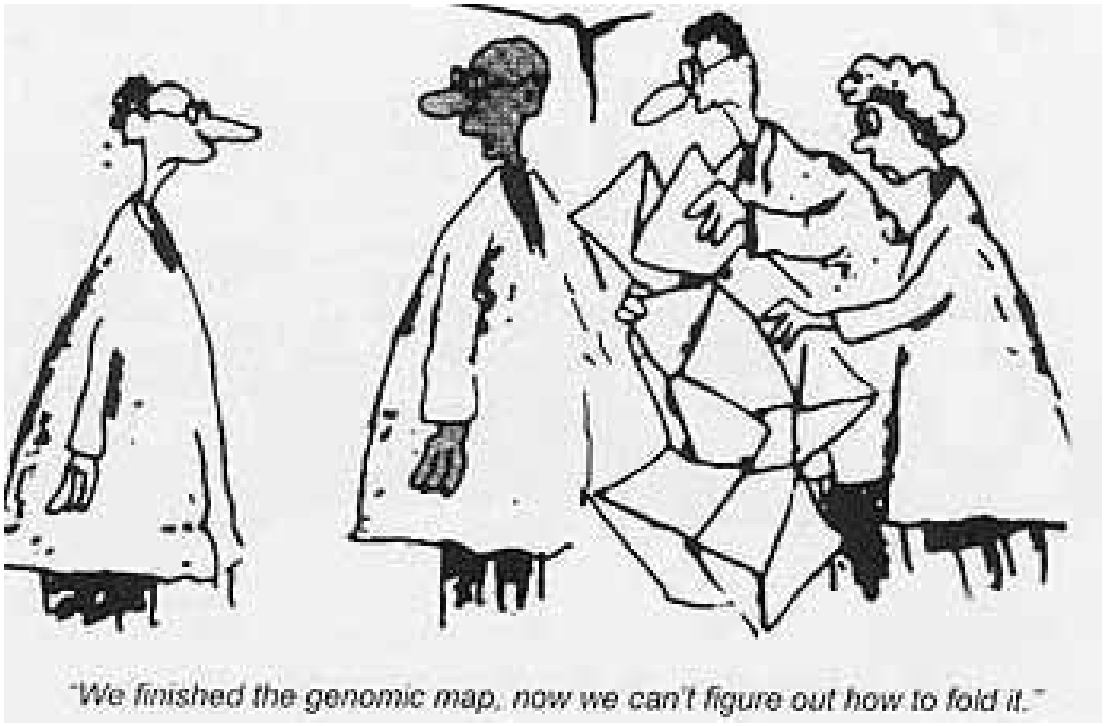
Εισαγωγή
(χωρίς πολλά λόγια)



?

#@!$%@$!#%

Εισαγωγή
(χωρίς πολλά λόγια)




```
#  => E
$  => I
@  => L
%  => S
!  => V
```

ELVISLIVES

..ATCTATGCAG**TCATGCTGACGAGCA**GCTGACTGACGTACGTACGATCGATCG..

?

Πρόβλεψη χαρακτηριστικών από αλληλουχίες DNA

- CpG islands
- Κωδικές περιοχές – Δομή γονιδίων
- Υποκινητές
- …

Συχνά απαιτείται να κατασκευάσουμε ΜΟΝΤΕΛΑ για τις αλληλουχίες που μας ενδιαφέρουν

(Στοχαστικά) Μοντέλα

- Μοντέλο?

  - Ένα σύστημα που προσομοιάζει ένα (πραγματικό) αντικείμενο

- Στοχαστικό?

  - οι διαφορετικές καταστάσεις του μοντέλου προκύπτουν με διαφορετική πιθανότητα

Στοχαστικά Μοντέλα

Ένα παράδειγμα [από Durbin et al, 1998]

$$\text{Η πιθανότητα } p_i \text{ να έρθει } i\,(i=1,2,..,6)$$

$$\text{Προφανώς, } p_i \geq 0 \text{ και } \sum_{i=1}^{6} p_i = 1$$

$$\text{ΤΙΜΙΟ ΖΑΡΙ : } p_i = p_j \;\forall\, i, j$$

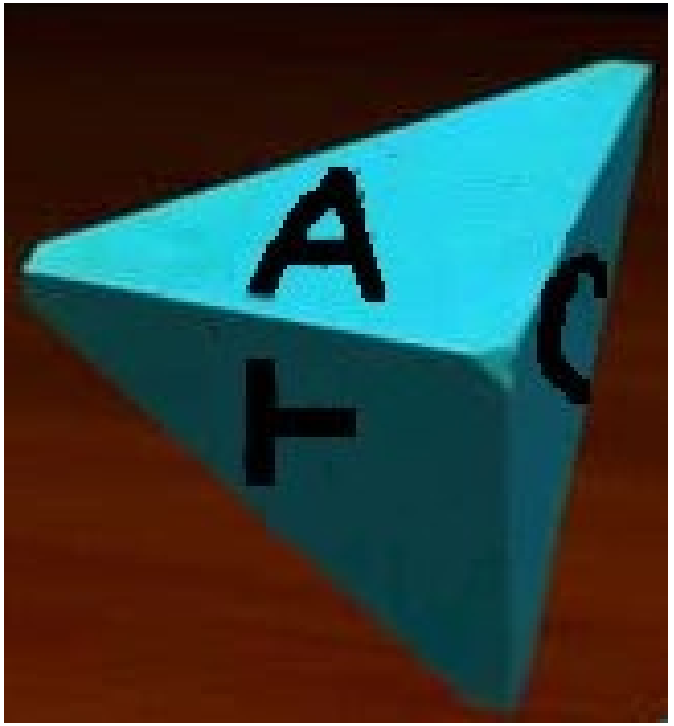
$$\text{ΣΤΗΜΕΝΟ ΖΑΡΙ : } \exists i, j : p_i \neq p_j$$

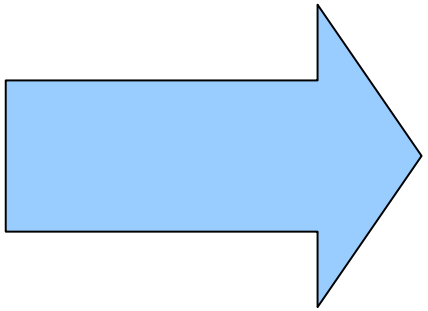
$$\text{Ένα μοντέλο των αποτελεσμάτων τριών ανεξάρτητων διαδοχικών ζαριών :}$$

$$P[a, b, c] = p_a\, p_b\, p_c, a, b, c \in \{1,2,..,6\}$$

Στοχαστικά Μοντέλα

Ένα Βιολογικό παράδειγμα





...ATACGAG...

Αλυσίδες Markov

Markov Chain, models/processes

- Ορισμός:
  - Μια στοχαστική διεργασία κατά την οποία η πιθανότητα παρατήρησης μιας κατάστασης τη χρονική στιγμή t εξαρτάται από πεπερασμένο πλήθος (k) προηγούμενων παρατηρήσεων
- k: τάξη (order) της διεργασίας
- Όταν δεν αναφέρεται η τάξη υπονοούμε ότι k=1

Αλυσίδες Markov

Markov Chains [από Durbin et al, 1998]

Ορισμός *Αλυσίδα Markov είναι η τριάδα* $(Q,\ p(x_1=s),\ A)$ *όπου:*

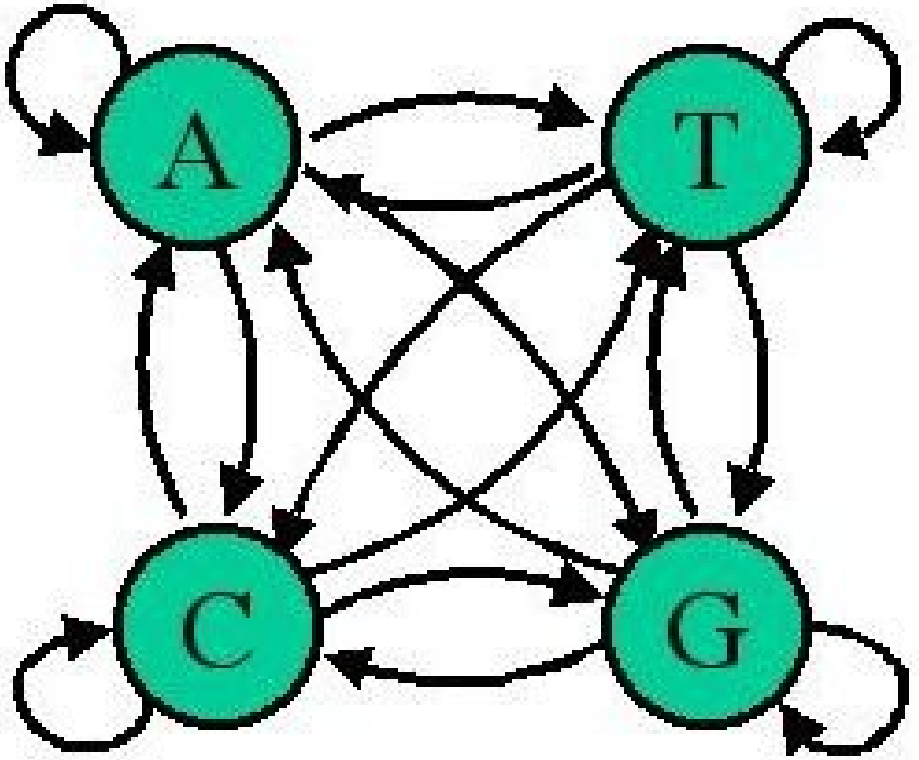
Q *είναι ένα πεπερασμένο σύνολο καταστάσεων*

p *είναι οι αρχικές πιθανότητες καταστάσεων*

A *είναι πιθανότητες μετάβασης* $a_{st},\ \forall s, t \in Q$

$$a_{st} \equiv P(x_i = t | x_{i-1} = s)$$

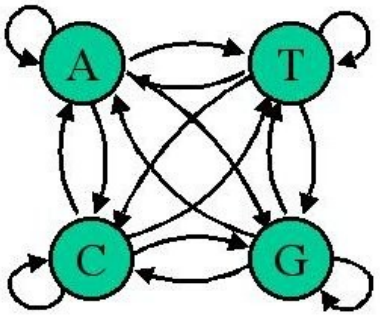
Βιολογικές ακολουθίες και Αλυσίδες Markov [από Durbin et al, 1998]



Ποια η πιθανότητα μιας ακολουθίας δοθέντος του μοντέλου?

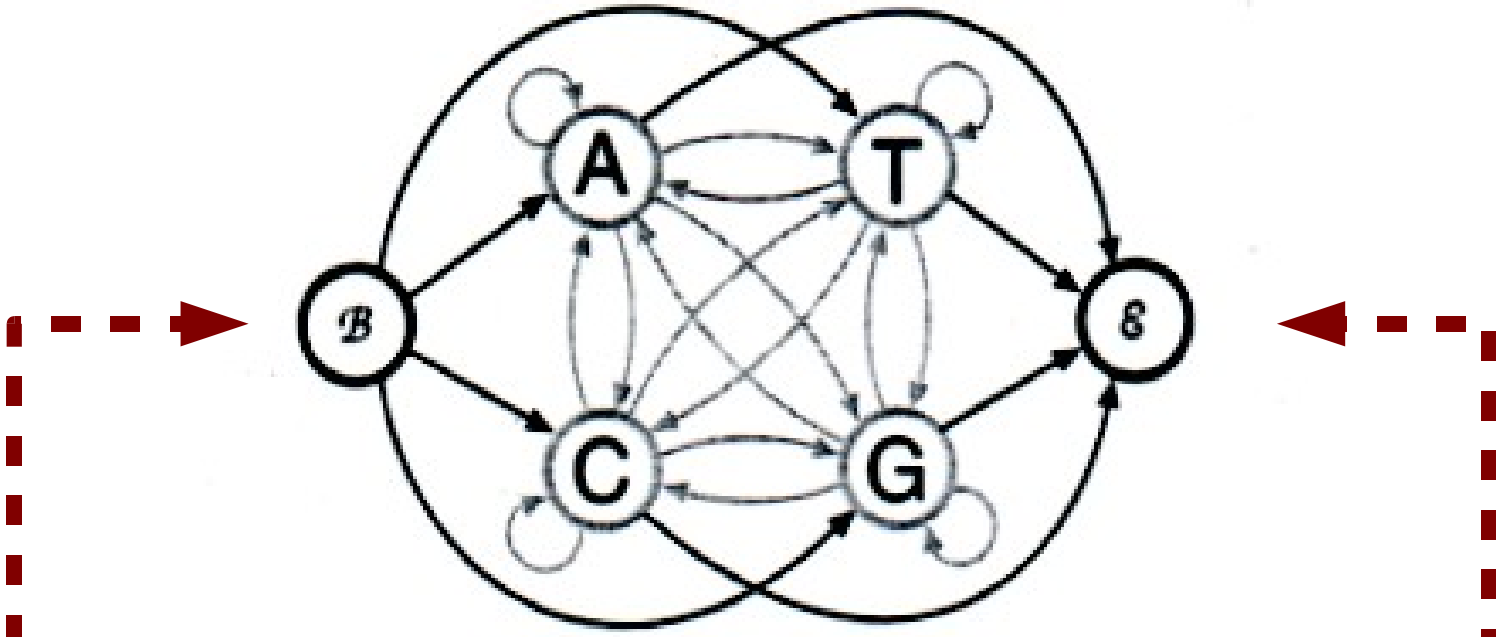
[από Durbin et al, 1998]

$$\begin{aligned}
&\text{Έστω η ακολουθία } x = x_1 x_2 x_3 \ldots x_L \\
&\text{Εάν έχουμε μια Αλυσίδα Markov } (Q, p, A) \\
&\text{Τότε } P(x) = P(x_L, x_{L-1}, \ldots, x_1) \\
&\qquad = P(x_L | x_{L-1}, \ldots, x_1) P(x_{L-1} | x_{L-2}, \ldots, x_1) \ldots P(x_2 | x_1) P(x_1) \\
&\qquad = P(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i}
\end{aligned}$$



Καταστάσεις Έναρξης/Λήξης

[από Durbin et al, 1998]



"Σιωπηλές" καταστάσεις

Αλυσίδες Markov για διάκριση

Εφαρμογή CpG-island prediction [από Durbin et al, 1998]

- Δημιουργία μοντέλων που περιγράφουν 2 αλληλοαποκλειόμενες κατηγορίες ακολουθιών

- Άγνωστη ακολουθία
  - Εύρεση της πιθανότητας σύμφωνα με κάθε μοντέλο
  - Εύρεση του log-odds ratio

Δημιουργία μοντέλων
[από Durbin et al, 1998]

- Positive examples (CpG islands)

- Negative examples (non-CpG islands)

$$a_{st}^{pos} = \frac{c_{st}^{pos}}{\sum_{t'} c_{st'}^{pos}}$$

$$a_{st}^{neg} = \frac{c_{st}^{neg}}{\sum_{t'} c_{st'}^{neg}}$$

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

Διάκριση ...

[από Durbin et al, 1998]

$$S(x) = \log \frac{P(x|model\ pos)}{P(x|model\ \text{neg})}$$

$$= \sum_{i=1}^{L} \log \frac{a^{pos}_{x_{i-1}x_i}}{a^{\text{neg}}_{x_{i-1}x_i}}$$

$$= \sum_{i=1}^{L} \beta_{x_{i-1}x_i}$$

“Μέτρα κωδικοποίησης”

Coding statistics

- *Μέτρο κωδικοποίησης* ορίζεται μια συνάρτηση η οποία με δεδομένη μια αλληλουχία DNA υπολογίζει έναν πραγματικό αριθμό, ο οποίος σχετίζεται με την πιθανοφάνεια αυτή η αλληλουχία να κωδικοποιεί μια πρωτεΐνη
- Οι συναρτήσεις αυτές είναι εν γένει αυθαίρετες, π.χ.:
  - Codon usage bias
  - Positional (within codons) base composition bias
  - Periodicities

Κατηγοριοποίηση Μέτρων Κωδικοποίησης

- Ανεξάρτητα μοντέλου
  - Αποτυπώνουν "γενικά" χαρακτηριστικά του κωδικού DNA
  - Δεν απαιτούν παραμετροποίηση - εκπαίδευση
- Βασισμένα σε μοντέλο του "κωδικού" DNA
  - Στοχαστικό μοντέλο
    - Υπολογισμός πιθανότητας κωδικοποίησης από μια αλληλουχία δεδομένου του μοντέλου κωδικών αλληλουχιών
    - Σύγκριση με την πιθανότητα ενός "τυχαίου" μοντέλου
  - Συνήθης δείκτης ο λογάριθμος του λόγου πιθανοφανειών
  - Απαιτούν την εκπαίδευση (εκτίμηση παραμέτρων)

Μέτρα βασισμένα σε μοντελοποίηση κωδικού DNA

- Πλεονεκτήματα
  - Αποτυπώνουν εξειδικευμένα χαρακτηριστικά
  - Εξαρτώνται από πλήθος παραμέτρων
  - Αυξημένες δυνατότητες
- Μειονεκτήματα
  - Απαιτούν αντιπροσωπευτικό δείγμα κωδικών αλληλουχιών από κάθε οργανισμό
  - Το μέγεθος - σύσταση του δείγματος ανάλογο του πλήθους των παραμέτρων

Παράδειγμα: codon usage

0.030759	-- [0] AAA	0.034321	-- [32] GAA
0.020932	-- [1] AAC	0.017247	-- [33] GAC
0.032765	-- [2] AAG	0.032298	-- [34] GAG
0.022274	-- [3] AAU	0.036727	-- [35] GAU
0.015655	-- [4] ACA	0.017533	-- [36] GCA
0.010364	-- [5] ACC	0.010396	-- [37] GCC
0.007708	-- [6] ACG	0.008973	-- [38] GCG
0.017626	-- [7] ACU	0.028530	-- [39] GCU
0.018855	-- [8] AGA	0.024278	-- [40] GGA
0.011281	-- [9] AGC	0.009177	-- [41] GGC
0.010937	-- [10] AGG	0.010223	-- [42] GGG
0.013962	-- [11] AGU	0.022375	-- [43] GGU
0.012527	-- [12] AUA	0.009928	-- [44] GUA
0.018588	-- [13] AUC	0.012805	-- [45] GUC
0.024427	-- [14] AUG	0.017400	-- [46] GUG
0.021581	-- [15] AUU	0.027350	-- [47] GUU
0.019338	-- [16] CAA	0.000915	-- [48] UAA
0.008693	-- [17] CAC	0.013830	-- [49] UAC
0.015213	-- [18] CAG	0.000509	-- [50] UAG
0.013720	-- [19] CAU	0.014682	-- [51] UAU
0.016158	-- [20] CCA	0.018124	-- [52] UCA
0.005316	-- [21] CCC	0.011123	-- [53] UCC
0.008576	-- [22] CCG	0.009221	-- [54] UCG
0.018707	-- [23] CCU	0.025087	-- [55] UCU
0.006262	-- [24] CGA	0.001092	-- [56] UGA
0.003765	-- [25] CGC	0.007089	-- [57] UGC
0.004827	-- [26] CGG	0.012469	-- [58] UGG
0.009007	-- [27] CGU	0.010462	-- [59] UGU
0.009861	-- [28] CUA	0.012630	-- [60] UUA
0.016086	-- [29] CUC	0.020723	-- [61] UUC
0.009840	-- [30] CUG	0.020865	-- [62] UUG
0.024165	-- [31] CUU	0.021847	-- [63] UUU

arab.codon.use
Codon usage for *Arabidopsis thaliana*: 66749
from GB142/gbpln.spsum: 26545376 codons

http://bioinformatics.weizmann.ac.il/blocks/help/CODEHOP/codons/arab.codon.use

Παράδειγμα: codon usage

- Συγκρίνουμε τη συχνότητα εμφάνισης τριπλετών σε μια περιοχή του γονιδιώματος σε κάθε ένα από τα πλαίσια ανάγνωσης με την “τυπική” συχνότητα ...
  - Περιοχές για τις οποίες οι τριπλέτες χρησιμοποιούνται με παρόμοιες με τις τυπικές συχνότητες είναι πιθανότερο να αντιστοιχούν σε κωδικές περιοχές
  - Ας το δούμε με ένα παράδειγμα ...

Έστω το τμήμα αλληλουχίας
S=AAGAAA
του *Arabidopsis thaliana*

```
0.030759        -- [ 0] AAA
0.020932        -- [ 1] AAC
0.032765        -- [ 2] AAG
0.022274        -- [ 3] AAU
0.015655        -- [ 4] ACA
0.010364        -- [ 5] ACC
0.007708        -- [ 6] ACG
0.017626        -- [ 7] ACU
0.018855        -- [ 8] AGA
0.011281        -- [ 9] AGC
0.010937        -- [10] AGG
0.013962        -- [11] AGU
0.012527        -- [12] AUA
0.018588        -- [13] AUC
...
...
```

P1 = P(S κωδική| μοντέλο codon usage) = F(AAG) x F(AAA)
≈ 0.033 x 0.031
≈ 0.001

Τυχαίο μοντέλο: όλα τα κωδικόνια ισοπίθανα (1/64=0.0156)

P0 = P(S μη κωδική|τυχαίο μοντέλο) = 0.0156 x 0.0156
= 0.000244

log(*P1 / P2*) > *0*

Η ίδια διαδικασία πρέπει να ακολουθηθεί για

ΟΛΑ τα πλαίσια ανάγνωσης

Μέτρα ανεξάρτητα μοντέλου

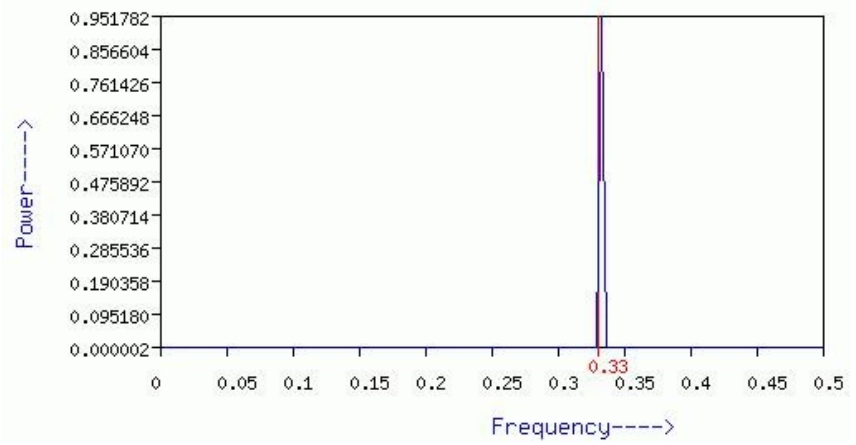
- Δε βασιζόμαστε σε κάποιο *a priori* στοχαστικό μοντέλο των κωδικών περιοχών
- Χρήσιμα για οργανισμούς για τα γονίδια των οποίων δεν έχουμε αρκετά δεδομένα
- Μια γενική παραδοχή είναι ότι οι κωδικές περιοχές χαρακτηρίζονται από *μικρότερη τυχαιότητα*
  - Η παρέκκλιση από την τυχαιότητα είναι υποδεικνύει κωδικές περιοχές
  - Εμπειρικά μέτρα

Παράδειγμα: Περιοδικές συσχετίσεις

- Κωδικές Περιοχές
  - Συσχετίσεις τοπικής εμβέλειας
  - Η τρίτη βάση επηρεάζει λιγότερο το αμινοξικό κατάλοιπο
- Μη κωδικές περιοχές
  - Δεν παρουσιάζεται κάποιο γενικό μοτίβο (Tsonis AA, Elsner JB, Tsonis PA., 1991)
  - Διαδεδομένες Μη κωδικές επαναλήψεις (SINES, LINES)
- FFT
  - Αποδοτικός εντοπισμός περιοδικοτήτων για σχετικά μεγάλα μήκη ακολουθιών DNA

Περιοδικότητες σε κωδικές περιοχές

Η ΙΔΑΝΙΚΗ ΠΕΡΙΠΤΩΣΗ
ΚΑΘΑΡΗ ΠΕΡΙΟΔΙΚΟΤΗΤΑ
T=0.33, f=3



ΣΤΗΝ ΠΡΑΞΗ ΤΟ ΦΑΣΜΑ ΠΕΡΙΕΧΕΙ ΟΛΕΣ ΤΙΣ ΣΥΧΝΟΤΗΤΕΣ

Συζήτηση ...

Διδακτικό υλικό:

http://troodos.biol.ucy.ac.cy/BRL/courses/BIO003/index.html