

Πίνακες αντικατάστασης PAM και BLOSUM και εναλλακτικές προσεγγίσεις

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

- Συστήματα Βαθμονόμησης Στοιχίσεων και Πίνακες Αντικατάστασης
- Διαισθητική προσέγγιση
- Εξελικτικές/Στατιστικές προσεγγίσεις
 - Αποδεκτές Σημειακές Μεταλλαγές και Πίνακες Αντικατάστασης PAM
 - Πίνακες BLOSUM
- Εναλλακτικές Προσεγγίσεις
- Συζήτηση
 - ...

Η σημασία των συστημάτων βαθμονόμησης

Η
Επιτακτική
Ανάγκη



ΣΥΓΚΡΙΣΗ ΑΚΟΛΟΥΘΙΩΝ
(DNA/RNA/ΠΡΩΤΕΙΝΩΝ)

Απαιτεί



ΔΗΜΙΟΥΡΓΙΑ
(ΒΕΛΤΙΣΤΩΝ)
ΣΤΟΙΧΙΣΕΩΝ

ΧΡΗΣΗ ΚΑΤΑΛΛΗΛΟΥ ΣΥΣΤΗΜΑΤΟΣ
ΒΑΘΜΟΝΟΜΗΣΗΣ (Scoring System)

- ΠΙΝΑΚΕΣ ΑΝΤΙΚΑΤΑΣΤΑΣΗΣ
 - ΠΟΙΝΕΣ ΕΙΣΑΓΩΓΗΣ ΚΕΝΩΝ
- 

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 \Rightarrow ολική πιθανότητα της στοίχισης
 - Βολεύει?

Πίνακες Αντικατάστασης (ΠΑ) (aka Substitution/Mutation/Scoring Matrices)

Στη γενική περίπτωση (π.χ. DNA):

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

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

Είναι δυνατόν να εκφράζουν μέτρο **ΟΜΟΙΟΤΗΤΑΣ**
ή **ΑΠΟΣΤΑΣΗΣ**

Πως μπορώ να δημιουργήσω ένα (χρήσιμο) ΠΑ?

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

Επιθυμητά χαρακτηριστικά ΠΑ

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

Ένας απλός ΠΑ

(για νουκλεοτιδικές ακολουθίες)

- Ο “Μοναδιαίος” Πίνακας Ταυτοτήτων (Identity Matrix)

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

.. κι άλλοι δυο απλοί ΠΑ

Ποια η διαφορά??

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

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

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

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

Πίνακες Αντικατάστασης

(για αμινοξικές ακολουθίες)

Δύο κύριες προσεγγίσεις

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

Η Οικογένεια των Πινάκων PAM

(Dayhoff, Schwartz, and Orcutt, 1978)

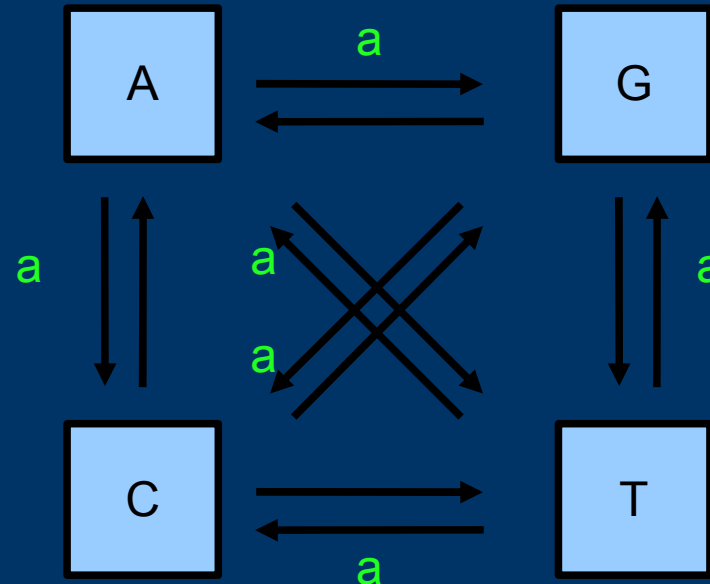
- Βασίζεται σε στοχαστικές Μαρκοβιανές διαδικασίες και στην κατασκευή φυλογενετικών δέντρων για το “ταίριασμα” με ένα εξελικτικό μοντέλο
- Υπολογισμός Πίνακα μεταβάσεων μετά από απαρίθμηση αντικαταστάσεων
- Υπολογισμός Πίνακα Αντικατάστασης (log-odds ratio)

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??

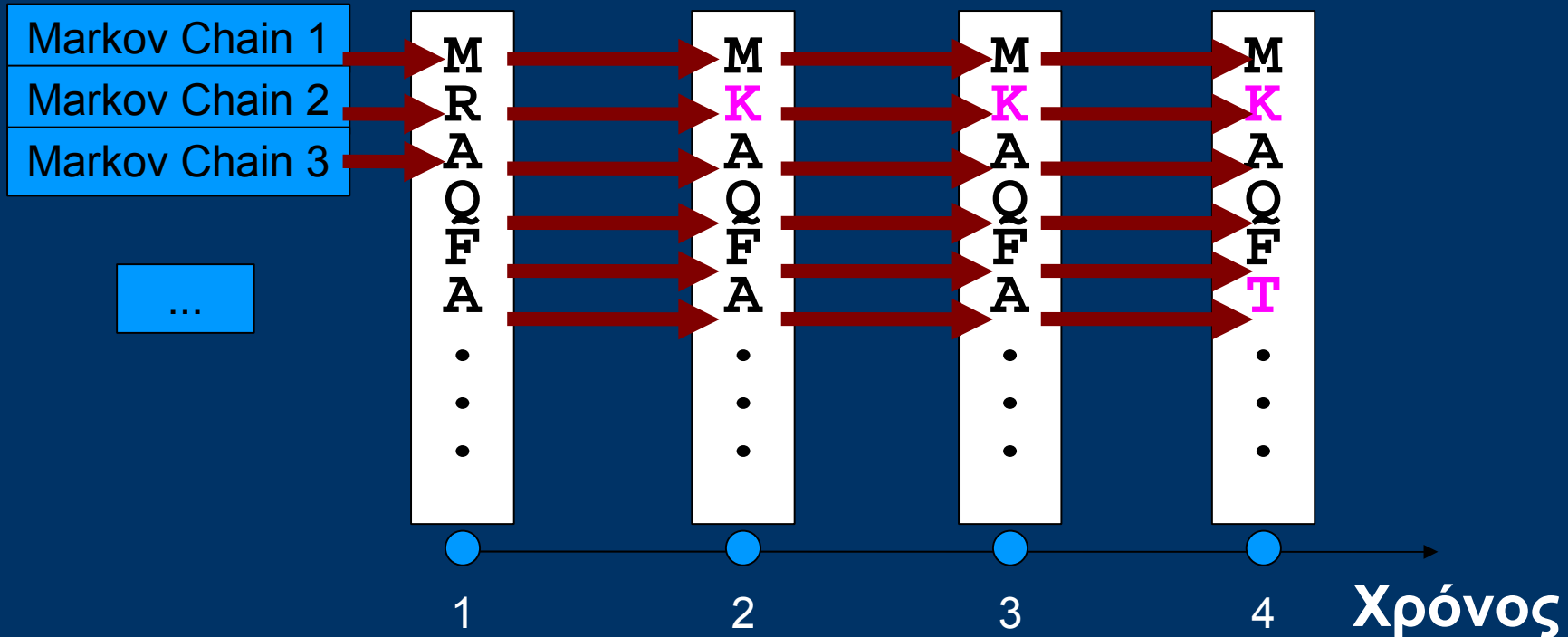
- Ναιiiiiiiiiiiii!



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 (20×20)
- 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

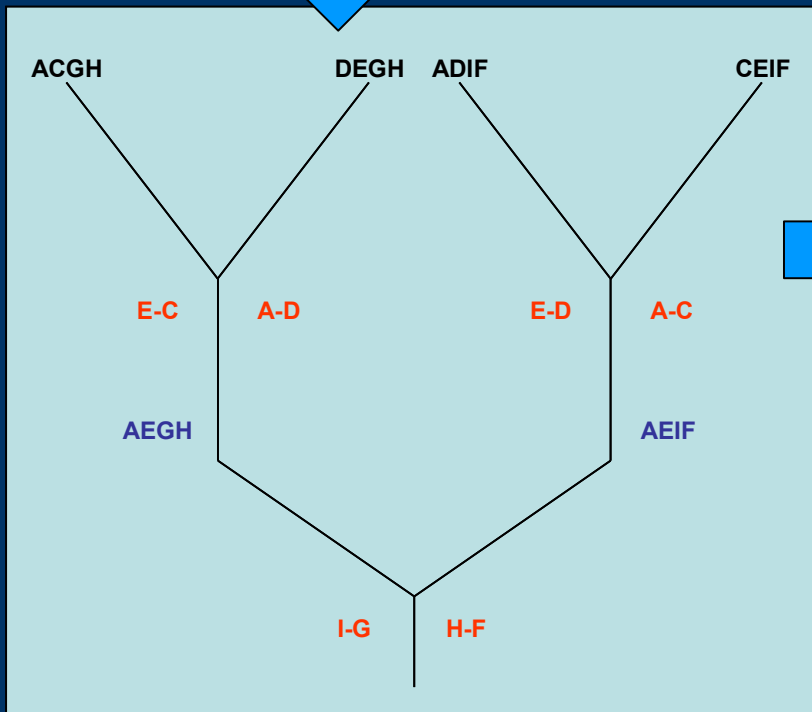
- Στοιχείση “όμοιων” πρωτεϊνικών ακολουθιών (>85% id)
???? ΓΙΑΤΙ?
???? ΠΙΝΑΚΑΣ?
- Ανακατασκευή φυλογενετικού δέντρου και ακολουθιών στους εσωτερικούς κόμβους
- Απαρίθμηση αντικαταστάσεων
- Εκτίμηση πιθανοτήτων αντικαταστάσεων

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

ACGH
DEGH
ADIF
CEIF



	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

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

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

... και το Λάμδα?

- Παράγοντας Στάθμισης
- Τι σταθμίζει?
- Επιλογή τιμής λ ώστε 1% των καταλοίπων να παθαίνουν PAM \Rightarrow *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

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

όπου F_j η παρατηρούμενη συχνότητα του καταλοίπου τύπου j

Υπολογισμός λ για 1-PAM (pt II)

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

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

ORIGINAL AMINO ACID

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

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

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

Κατασκευή ΠΑ ΡΑΜ1

- Έστω οι ακολουθίες $\alpha = \alpha_1 \alpha_2 \dots \alpha_m$ και $\beta = \beta_1 \beta_2 \dots \beta_m$
- Η βαθμονόμηση της στοίχισής τους αντιστοιχεί με τον έλεγχο δύο αλληλοαποκλειόμενων υποθέσεων
 - H_0 : οι 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 :

$\alpha_1 \alpha_2 \dots \alpha_m$
 $\beta_1 \beta_2 \dots \beta_m$

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

Κατασκευή ΠΑ PAM1 (pt II)

Με βάση την υπόθεση H_1 :

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

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

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

ή ισοδύναμα

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

PAM250 ??

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

Πίνακες *BLOSUM*

BLOcks SUBstitution Matrices

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

BLOSUM62

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

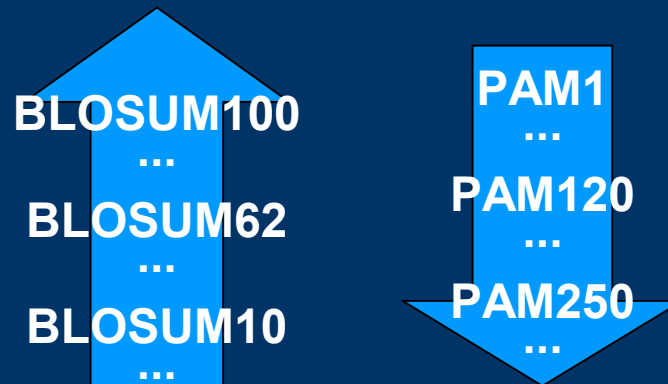
Τα νούμερα των BLOSUM

WWYIR	CASILRKIYIYGPV	GVSRLRTAYGGRK	NRG
WFYVR	CASILRHLYHRSPA	GVGSI TKIYGGRK	RNG
WYYVR	AAAVARHIYLRKTV	GVGRLRKVHGSTK	NRG
WYFIR	AASICRHLYIRSPA	GIGSF EKIYGGR	RRG
WYYTR	AASIARKIYLRQGI	GVGGFQKIYGGRQ	RNG
WFYKR	AASVARHIYMRKQV	GVGKLNKLYGGAK	SRG
WFYKR	AASVARHIYMRKQV	GVGKLNKLYGGAK	SRG
WYYVR	TASIARRLYVRSPT	GVDALRLVYGGSK	RRG
WYYVR	TASVARRLYIRSPT	GVGALRRVYGGNK	RRG
WFYTR	AASTARHLYLRGGA	GVGSM TKIYGGRQ	RNG
WFYTR	AASTARHLYLRGGA	GVGSM TKIYGGRQ	RNG
WYVVR	AAALLRRVYIDGPV	GVNSLRTHYGGKK	DRG

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

PAM vs BLOSUM

- BLOSUM: σημαντικά μεγαλύτερος όγκος δεδομένων [min F = 2369]
- PAM1 [ακριβής] ~ PAMn [προσεγγιστικά]
- 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 $S_{i,j} > 0$

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

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

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

Πληροφοριακό περιεχόμενο ΠΑ

Σχετική εντροπία

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

Συζήτηση

• ...