ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ ΙΙ

Συγκριτική Γονιδιωματική

Παντελής Μπάγκος

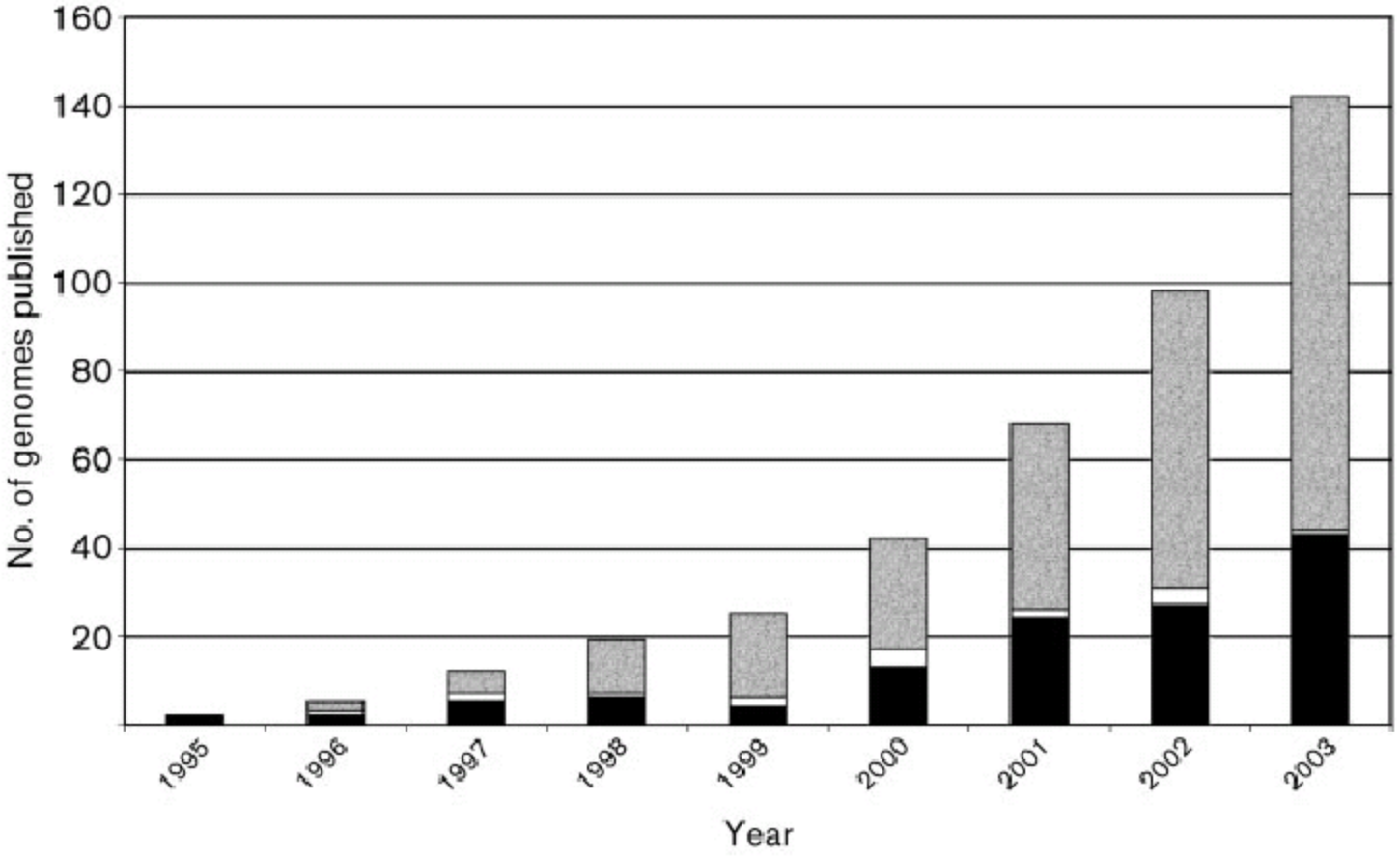


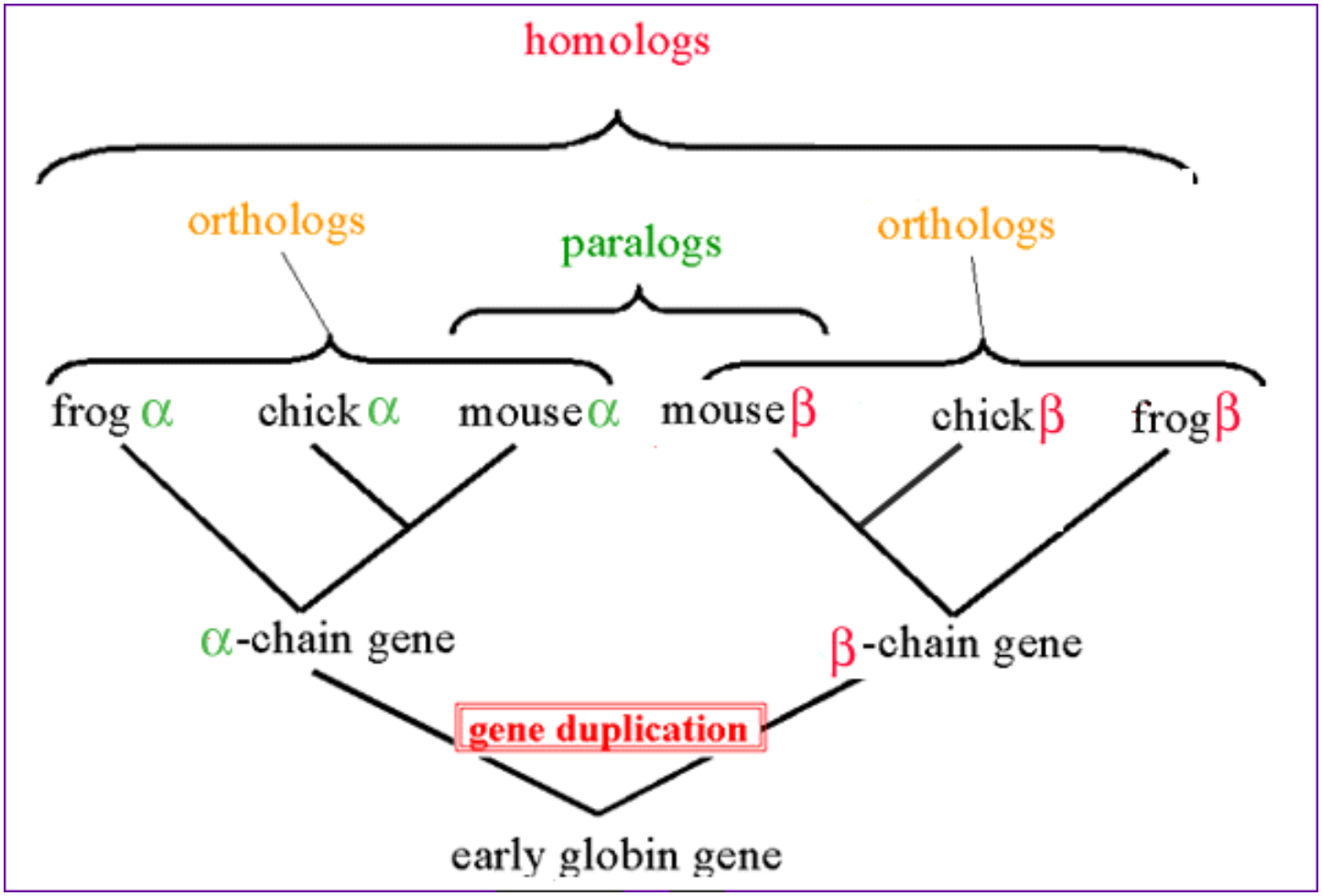
Fig. 1. Number of prokaryotic genomes sequenced each year since 1995. Black, bacterial genomes; white, archaeal genomes; grey, running total.

Μέθοδοι Ανάλυσης

- Μέθοδοι βασισμένες στην ομοιότητα ακολουθιών
  - Τοπική ομοιότητα
  - Ολική ομοιότητα
- Προγνωστικές μέθοδοι
  - Δευτεροταγής δομή
  - Διαμεμβρανικά τμήματα
  - Πεπτίδια οδηγητές
  - Λειτουργικά χαρακτηριστικά, κλπ

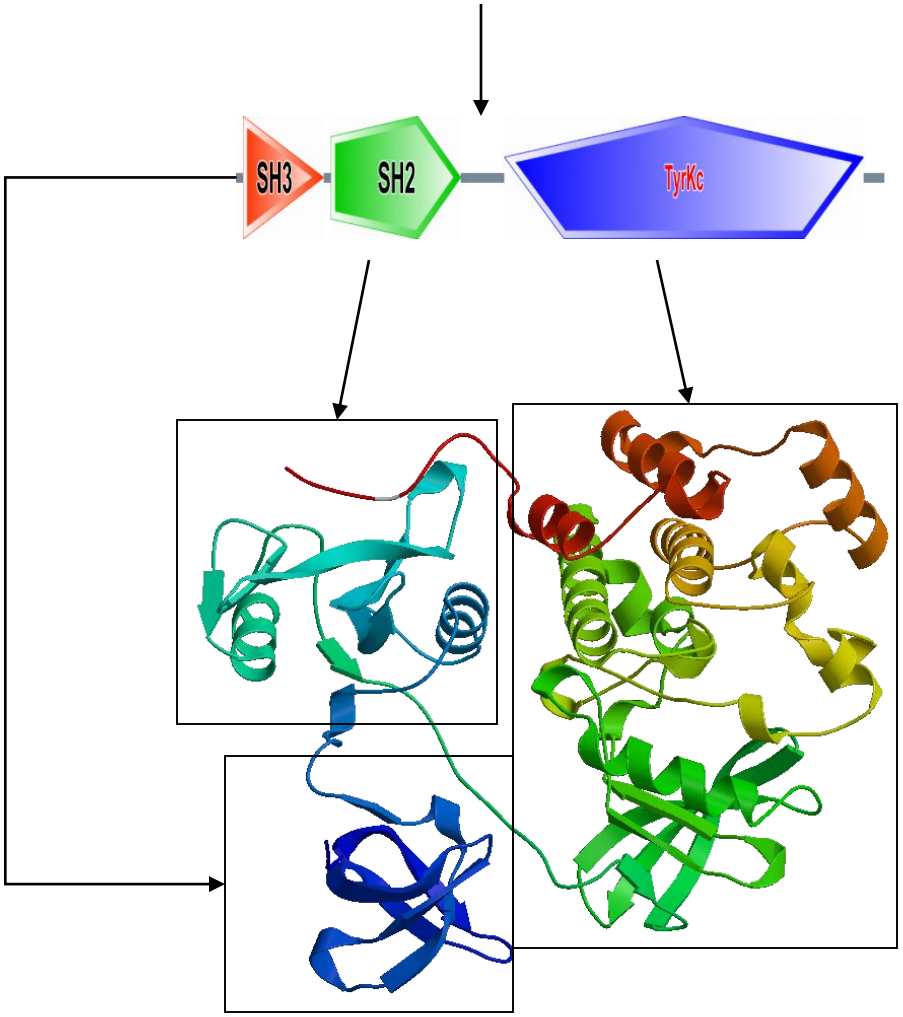
Κατά ζεύγη στοίχιση ακολουθιών

- Από τα πιο σημαντικά προβλήματα στην Υπολογιστική Βιολογία
- Ιδιαίτερα πλούσια βιβλιογραφία για πάνω από 30 χρόνια
- Ένα θέμα κυρίως αλγοριθμικό, αλλά με μεγάλη βιολογική σημασία
- Η ομοιότητα δυο ακολουθιών αντανακλά κατά βάση την κοινή εξελικτική προέλευση

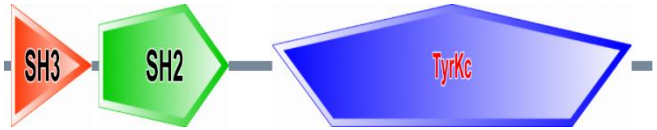

homologs
orthologs
paralogs
orthologs
frog α
chick α
mouse α
mouse β
chick β
frog β
α-chain gene
β-chain gene
gene duplication
early globin gene

Protein Domains

SGIRIIVVALYDYEAIHHEDLSFQKGDQMVVLEESGEWWKARSLATRKEGYIPSNYVARV
DSLETEEWFFKGISRKDAERQLLAPGNMLGSFMIRDSETTKGSYSLSVRDYDPRQGDTVK
HYKIRTLDNGGFYISPRSTFSTLQELVDHYKKGNDGLCQKLSVPCMSSKPQKPWEKDAWE
IPRESLKLEKKLGAGQFGEVWMATYNKHTKVAVKTMKPGSMSVEAFLAEANVMKTLQHDK
LVKLHAVVTKEPIYIITEFMAKGSLLDFLKSDEGSKQPLPKLIDFSAQIAEGMAFIEQRN
YIHRDLRAANILVSASLVCKIADFGLARVIEDNEYTAREGAKFPIKWTAPEAINFGSFTI
KSDVWSFGILLMEIVTYGRIPYPGMSNPEVIRALERGYRMPRPENCPEELYNIMMRCWKN
RPEERPTFEYIQSVLDDFYTATESQEEIP



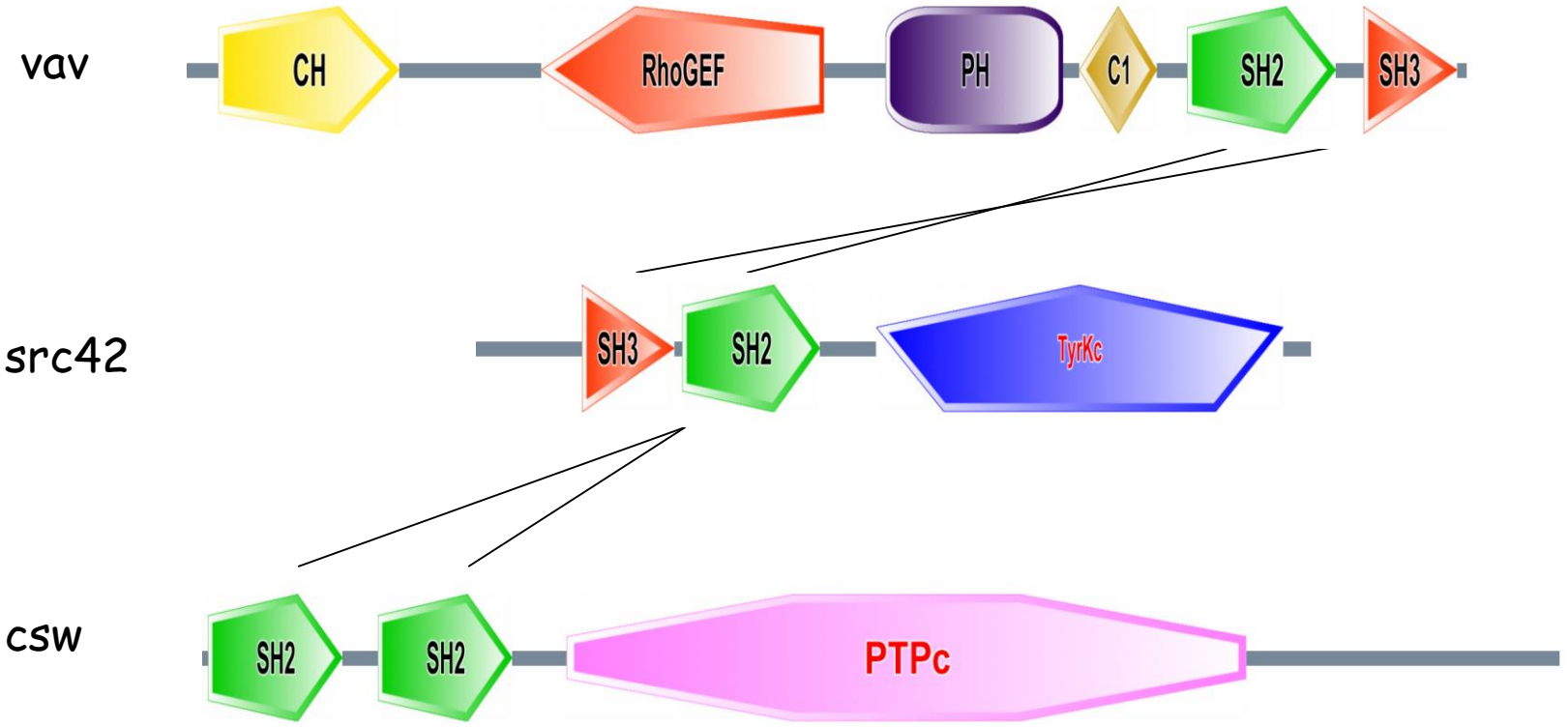
Protein Families

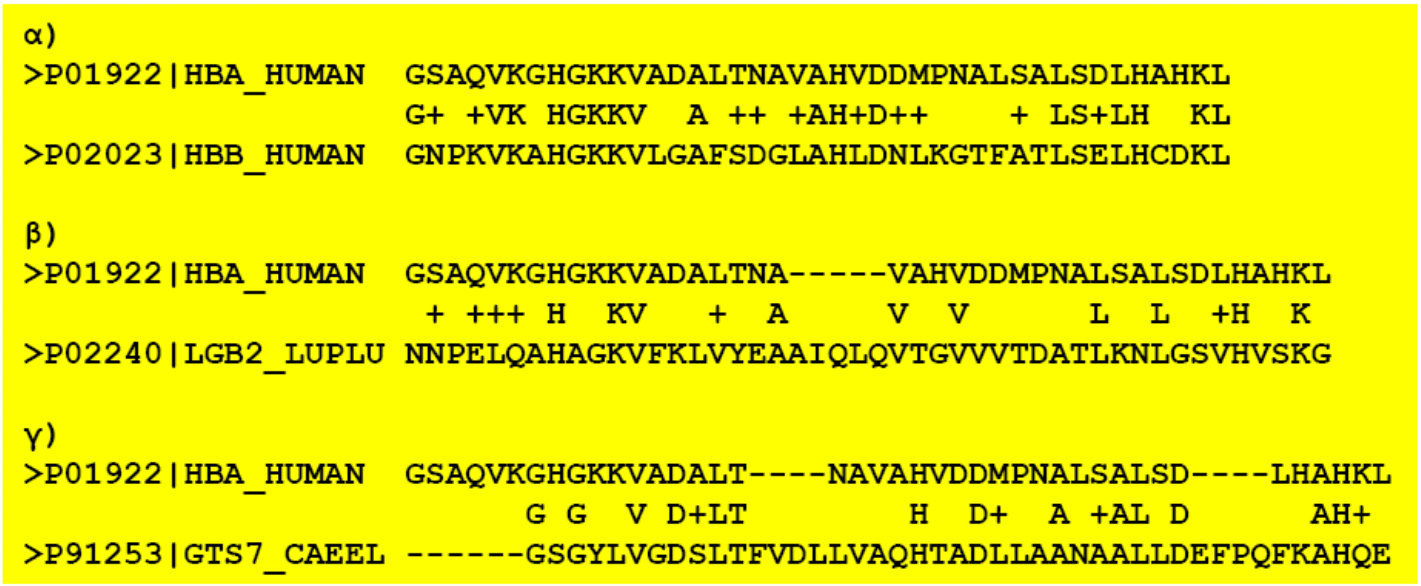


src-like protein tyrosine kinase - 5 in *Drosophila* proteome

38 tyrosine kinases
43 SH2 domain containing
110 SH3 domain containing

Local Similarity





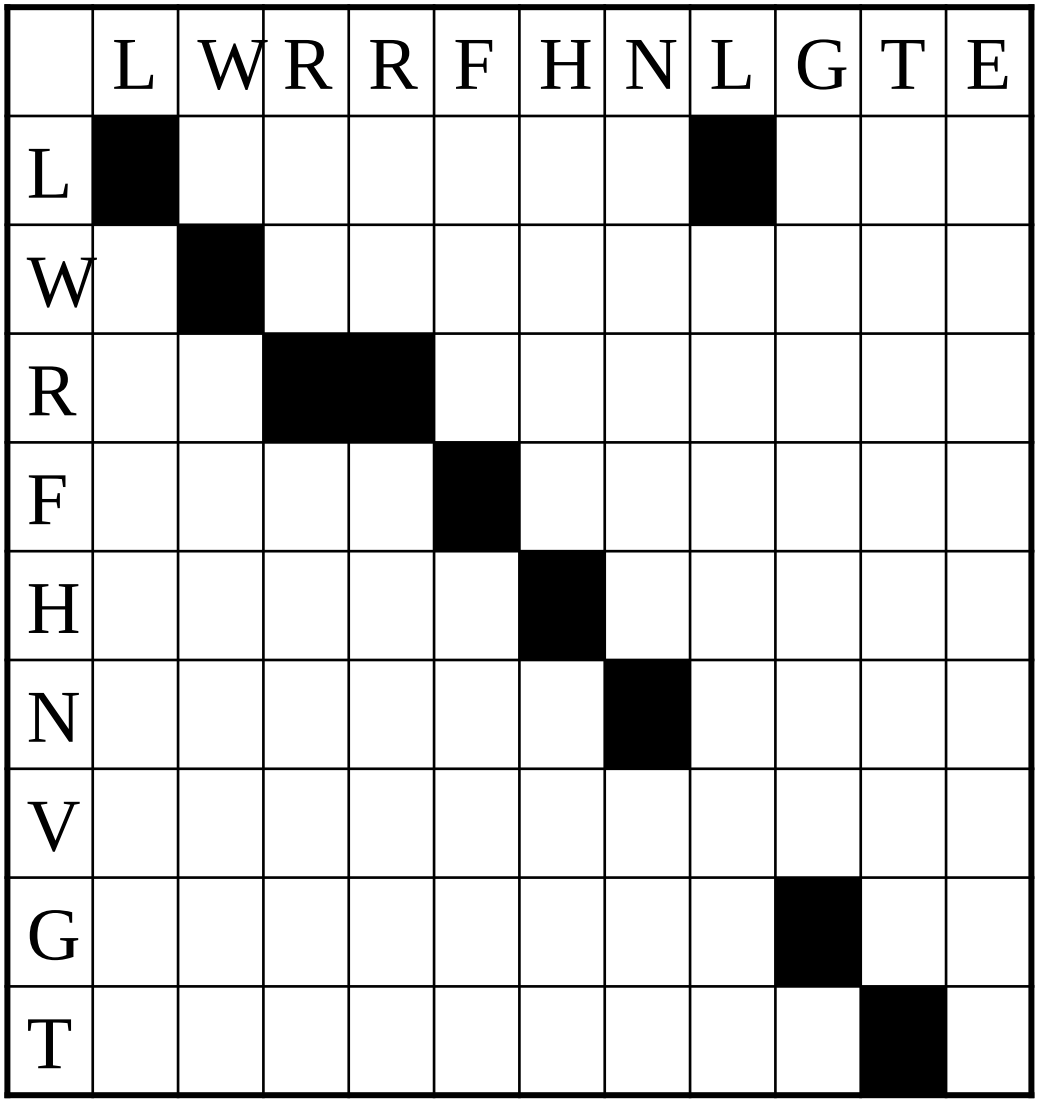
Εικόνα 3: Τρεις στοιχίσεις ακολουθιών με τον αλγόριθμο Needleman-Wunsch με ένα τμήμα της άλφα αλυσίδας της ανθρώπινης αιμοσφαιρίνης (SwissProt AC P01922).

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

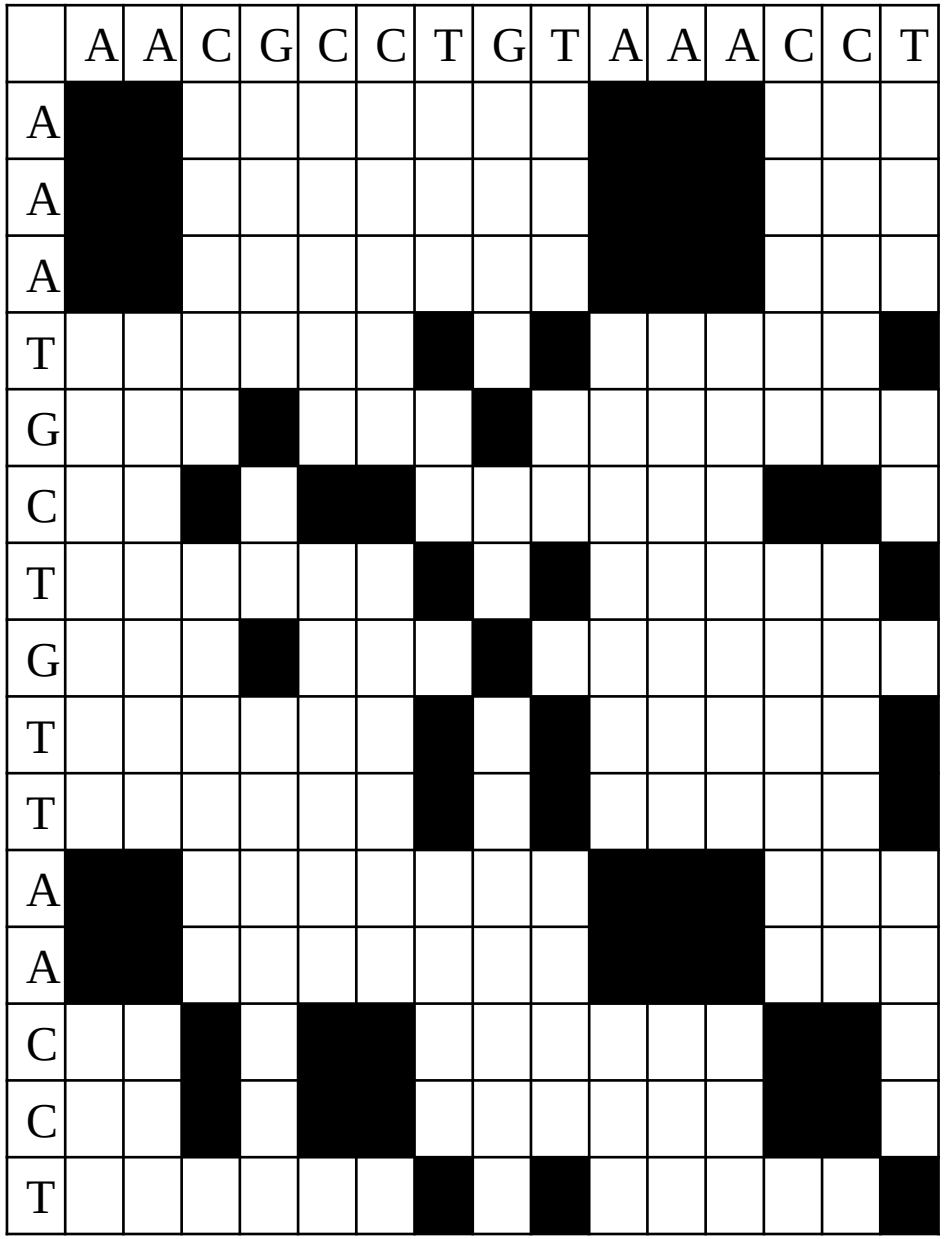
β) Δομικά συμβατή στοίχιση με την leghemoglobin II (AC P02240) του δικοτυλίδονου Lupinus luteus.

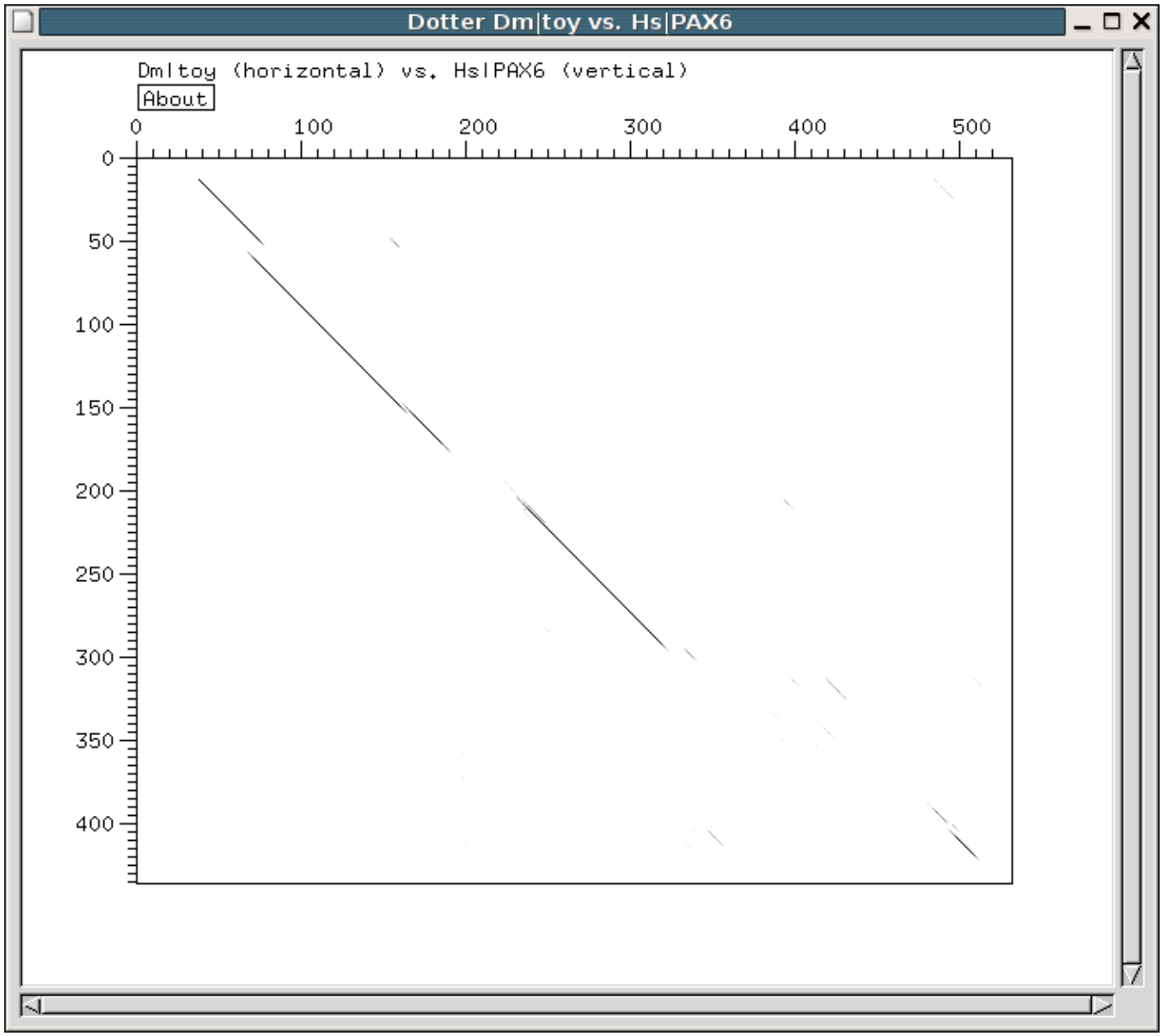
γ) 'Παραπλανητική' στοίχιση με ομόλογη της S-τρανφεράσης της γλουταθειόνης (AC P91253) του νηματώδη σκώληκα C. elegans.

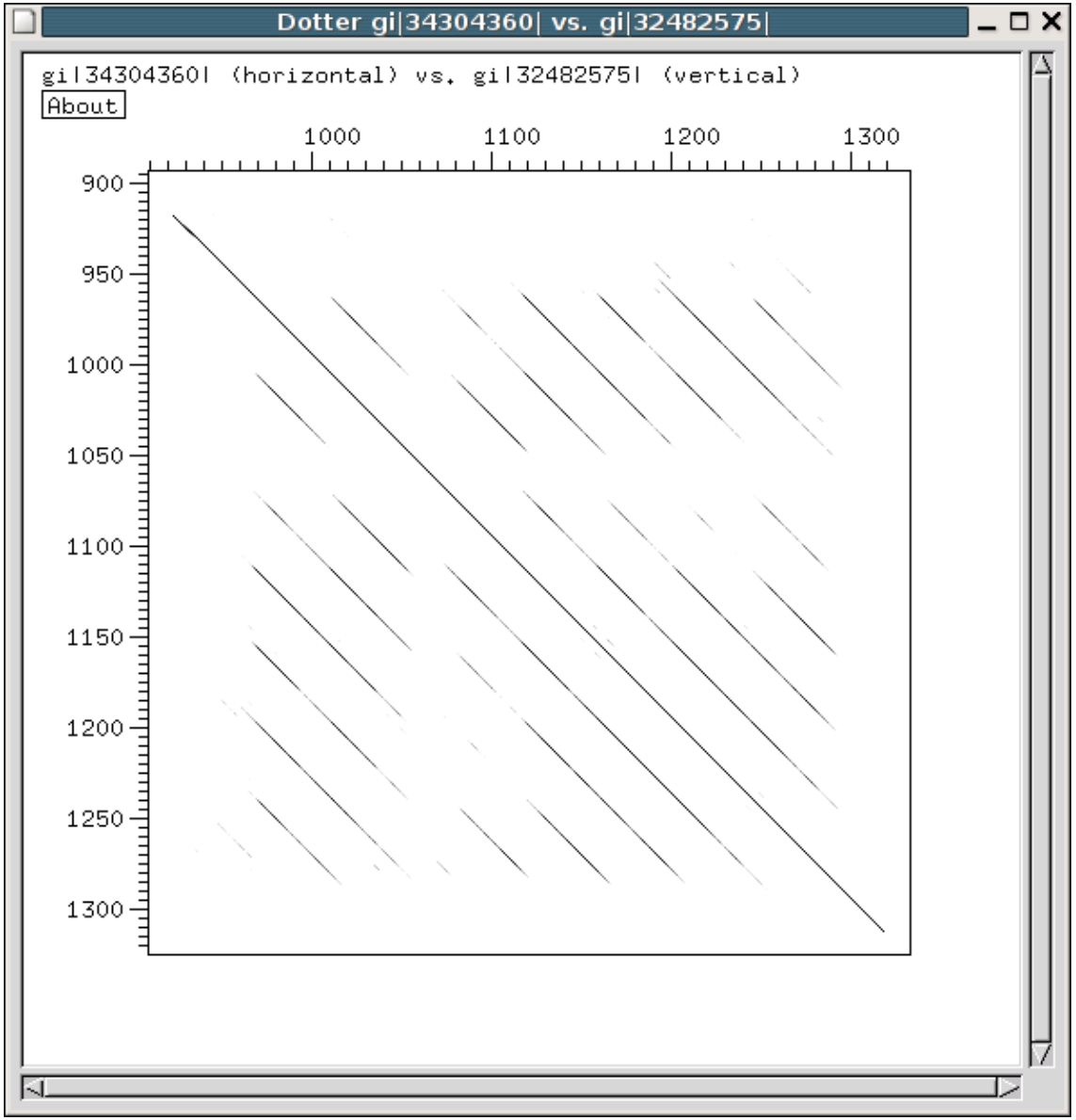
Dotplot

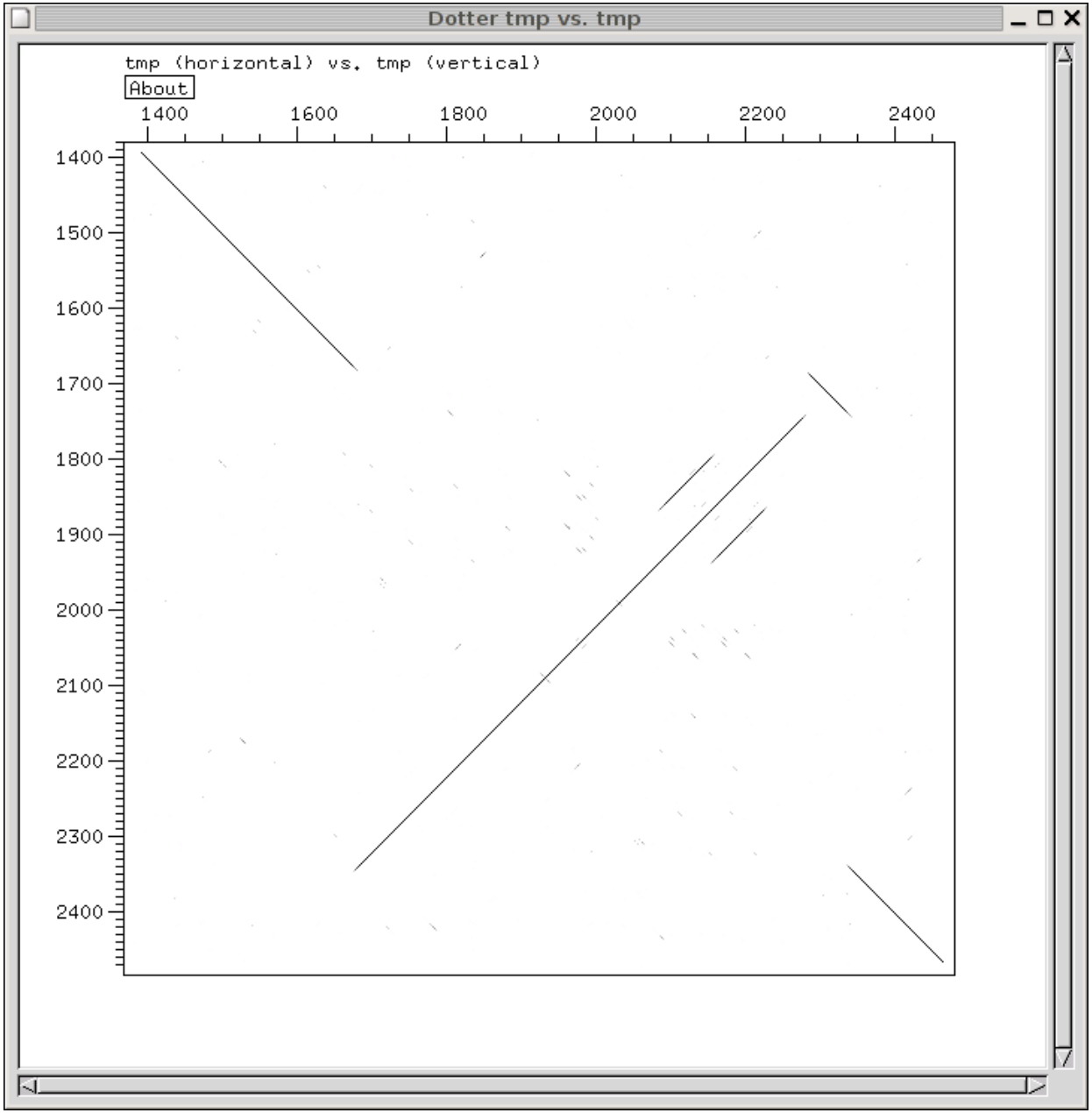


DNA




Dotter Dm|toy vs. Hs|PAX6
Dm|toy (horizontal) vs. Hs|PAX6 (vertical)
About
0
100
200
300
400
500
0
50
100
150
200
250
300
350
400


Dotter gi|34304360| vs. gi|32482575|
gi|34304360| (horizontal) vs. gi|32482575| (vertical)
About
1000
1100
1200
1300
900
950
1000
1050
1100
1150
1200
1250
1300


Dotter tmp vs. tmp
tmp (horizontal) vs. tmp (vertical)
About
1400
1600
1800
2000
2200
2400
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400

Ευριστικοί Αλγόριθμοι (heuristics)

- BLAST (www.ncbi.nlm.nih.gov/BLAST/)
- FASTA (www.ebi.ac.uk/fasta33/)

1. Αποδίδουν «σχεδόν» το ίδιο καλά με τους αλγορίθμους Δυναμικού Προγραμματισμού
2. Απαραίτητοι καθώς αυξάνεται διαρκώς το μέγεθος των βάσεων δεδομένων

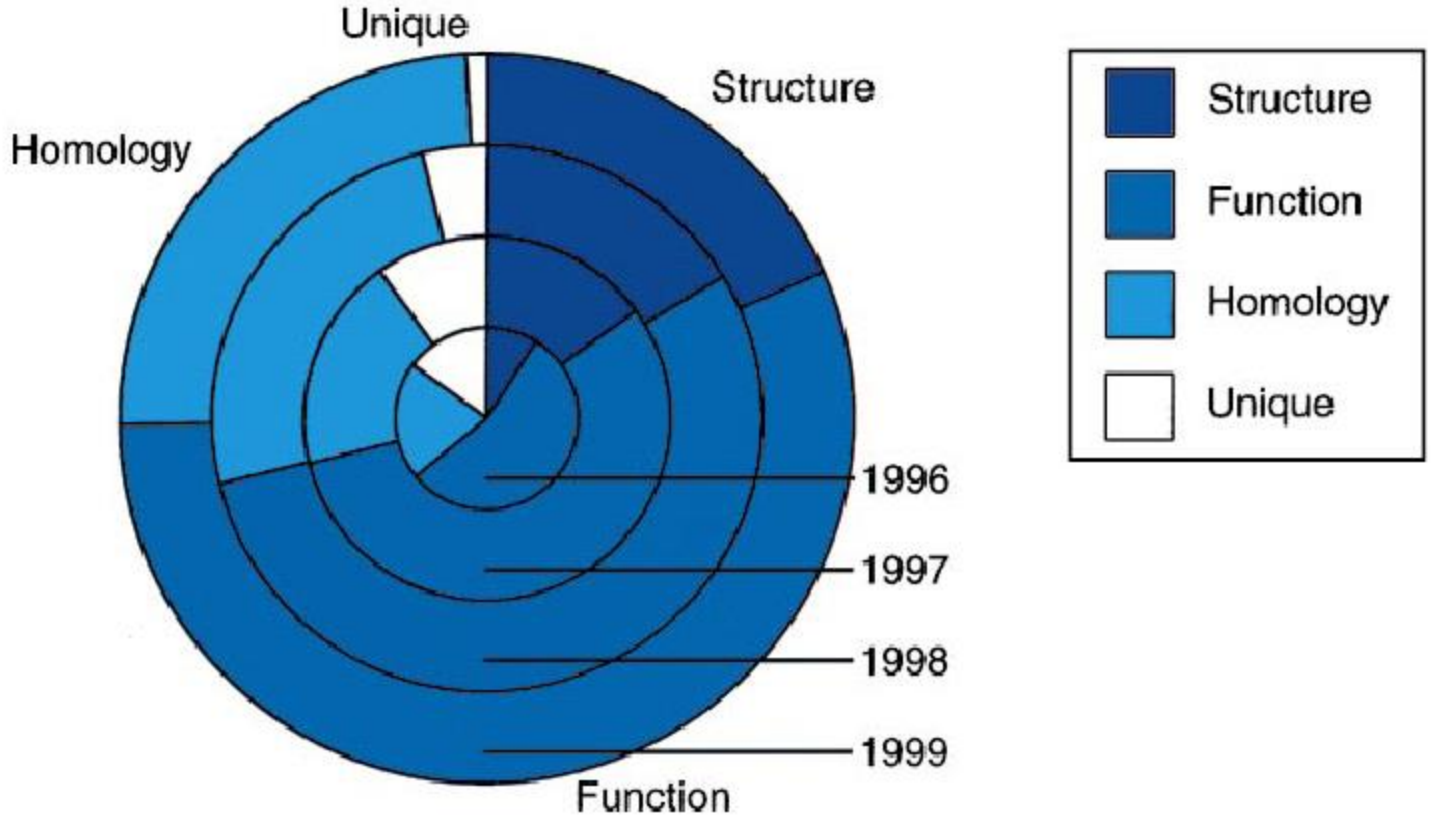


Fig. 2. The 'function bottleneck'. Information clock illustrating the improvement of annotations identified for a given genome over the years 1996–1999. The four levels of annotation range from homologues of known structure (blue) and homologues of known function (marine) to homologues of unknown function (cyan) and unique sequences (white). Note that although structure and homology increased over the years, the function prediction level stalled. Data from the GeneQuiz system, still available at: http://www.ebi.ac.uk/research/cgg/services/.

Αλγόριθμοι πρόγνωσης

- Στηρίζονται στην εκπαίδευση μιας μεθόδου με κάποια γνωστά παραδείγματα και την ελπίδα ότι τα αποτελέσματα θα γενικεύονται σε άγνωστες πρωτεΐνες
- Διάφορες αλγοριθμικές τεχνικές (στατιστικές μέθοδοι, μέθοδοι μηχανικής μάθησης κλπ)
- Ποσοστά επιτυχίας που ποικίλλουν ανάλογα με το πρόβλημα και τη μεθοδολογία
  - Gene finding
  - Δευτεροταγής δομή
  - Διαμεμβρανικά τμήματα
  - Πεπτίδια οδηγητές
  - Λειτουργικά χαρακτηριστικά, κλπ

Συγκριτική γονιδιωματική

- Η συγκριτική μελέτη δυο η περισσότερων γονιδιωμάτων (ή πρωτεωμάτων)
- Διάφορες προσεγγίσεις

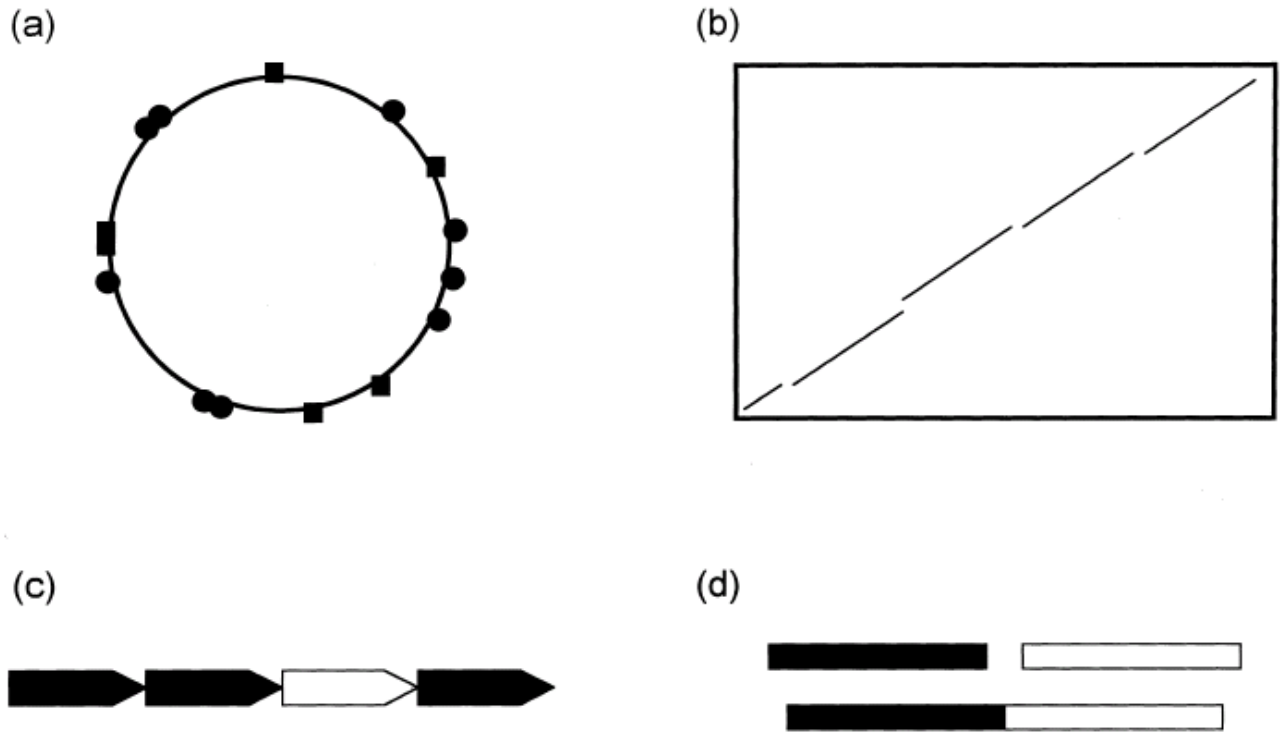


Fig. 1. Pictorial representation of four computational genomics methods. (a) Genome subtraction aims to define species-specific genes. This is achieved by subtracting genes homologous to various elements such as genes orthologous to the species under consideration or phages which are likely to be inserted by horizontal transfer. The method thus detects species-specific genes that can be linked to phenotypic features (represented by squares or circles). (b) Whole-genome alignment for two hypothetical species. Axes indicate genome positions and each point indicates a match between genome sequences. Such genome alignment plots reveal organisational features such as homologous regions or duplications. (c) Functional coupling of gene clusters detects orthologous genes between species which are then used to predict functional networks. The detection of a conserved battery of genes of known function (black arrows) implies that a gene of unknown function (white arrow) may have a related role, on the basis of its presence in the same 'operon'. (d) Schematic representation of fusion analysis. The approach resembles an in silico two-hybrid system and is based on the detection of groups of non-homologous genes in one organism found fused in the corresponding gene in another organism. In the case of genes of unknown function being involved, such associations may be used to infer functional associations.

Codon Bias – GC% content

- Στα γονιδιώματα διαφόρων οργανισμών παρατηρούνται διαφορετικά ποσοστά εμφάνισης GC
- Τα διαφορετικά κωδικόνια για τα ίδια αμινοξέα εμφανίζονται με διαφορετικές συχνότητες
- Στα Gene Finders, χρησιμοποιούνται αυτές οι «προτιμήσεις»

Διαχωρισμός των θερμόφιλων βακτηρίων από την αμινοξική σύσταση

- Συσχέτιση GC με αμινοξική σύσταση
- Η αμινοξική σύσταση καθορίζει την προσαρμογή στο περιβάλλον (π.χ. θερμοφιλία)

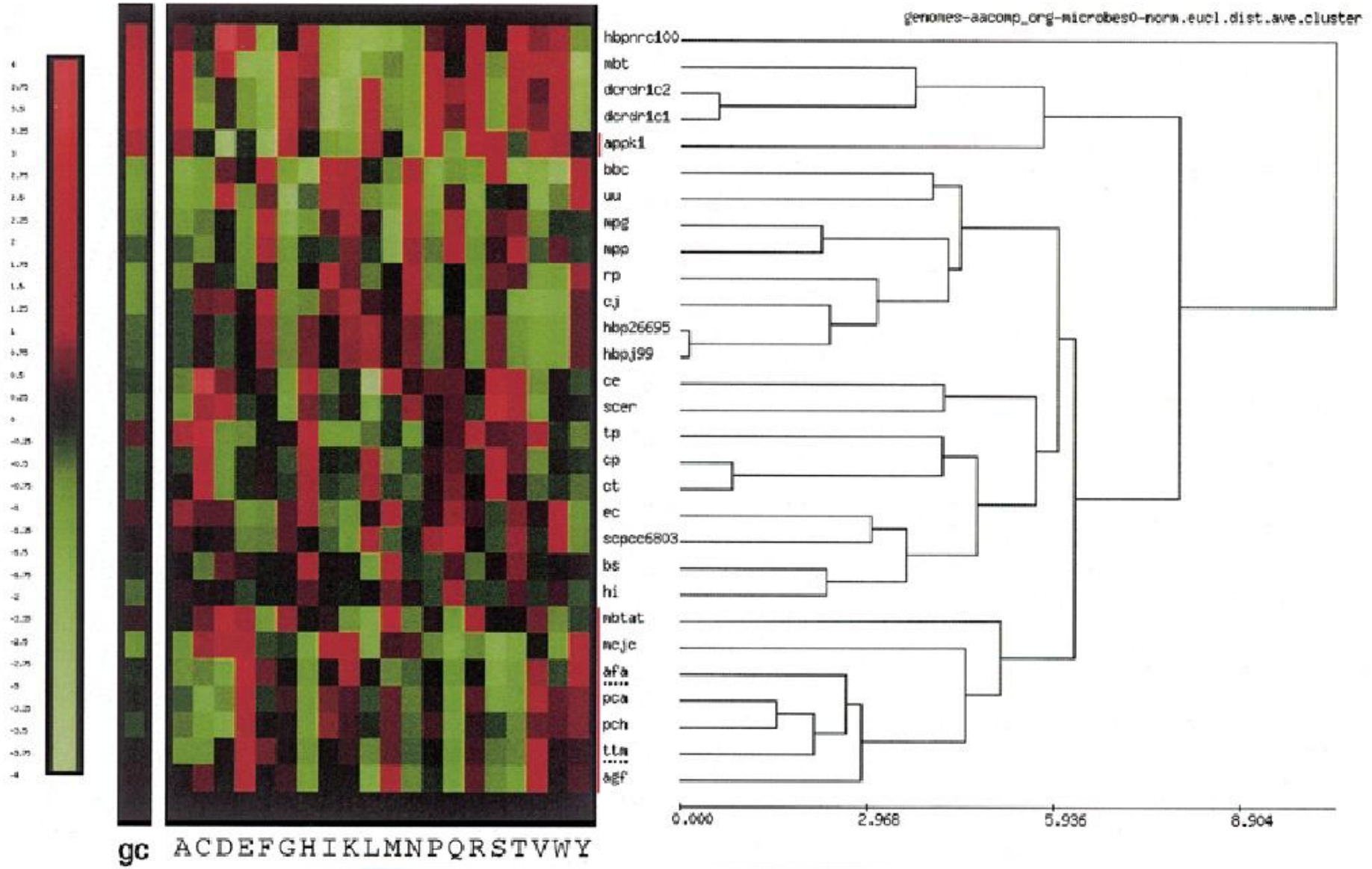


Figure 1. Standardised amino acid composition data of completely sequenced organisms grouped by hierarchical clustering. The GC ratios are shown for reference but were not used for the clustering process. Amino acids are abbreviated by the standard one letter code. The labels indicating the data sets for each row are explained in Table 1. In this figure, labels for thermophiles are marked with a red vertical bar, the thermophilic bacteria are highlighted by a dotted underline. The coloured blocks show normalised values as seen from the colour bar at the left. Red and green mean more and less than average, respectively. The scale for the dendrogram represents Eucledian distance. See Materials and Methods for details.

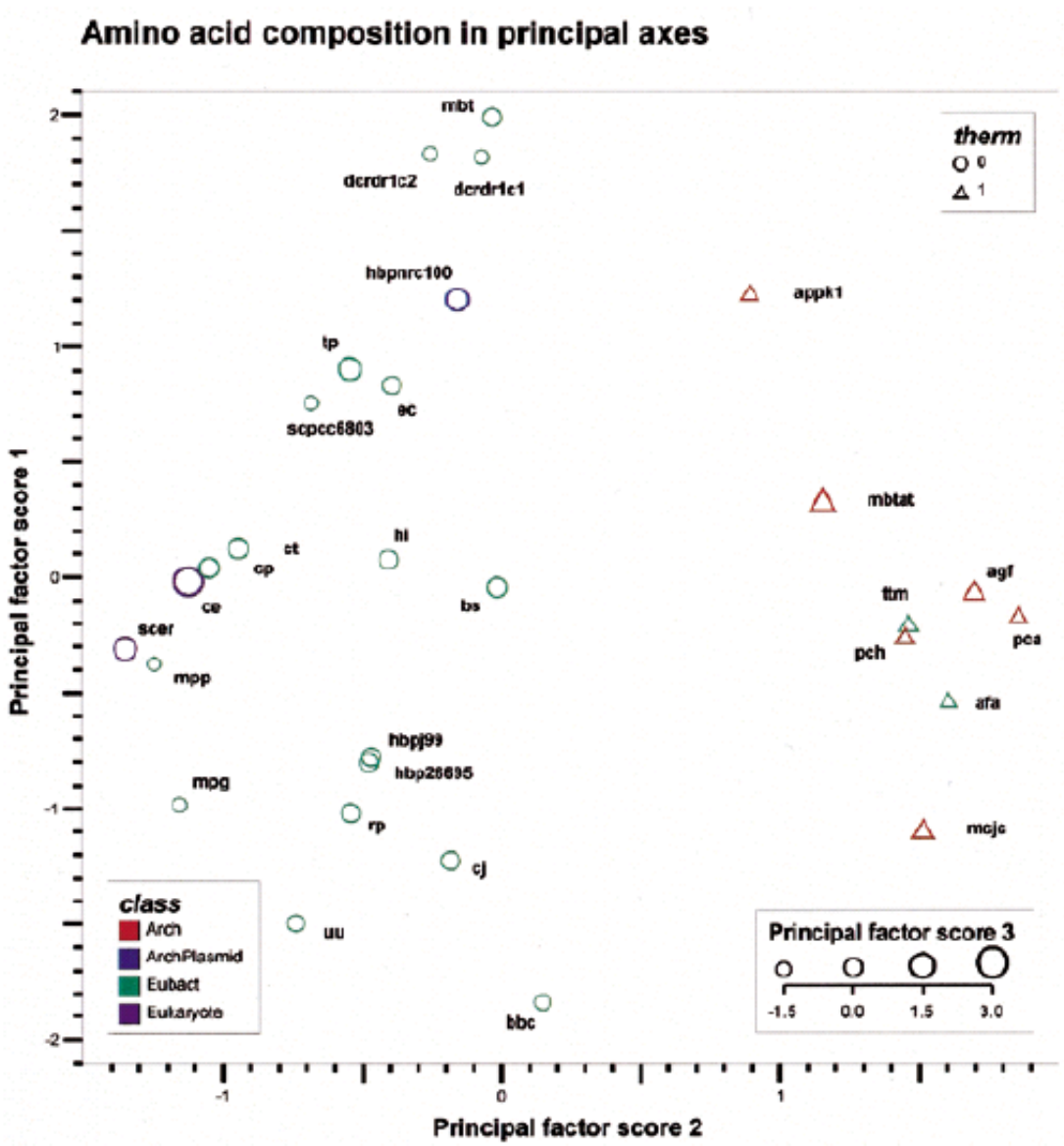


Figure 2. Reduced dimensionality plot showing the main principal components of the global amino acid compositions. The first principal axis (vertical) corresponds to GC ratio (see text). The second principal axis (horizontal) shows a clear separation of thermophiles and mesophiles, denoted by triangles and circles, respectively. The third principal component is depicted by symbol size (see insert for scale). Colour groups the sources into archea (red), bacteria (green) and eukaryotes (purple). The plasmid (the outgroup for hierarchical clustering, Fig. 1) is shown in blue. The graph is a projection, and distances are therefore not directly comparable to the distances observed in Figure 1. See text for discussion. For an explanation of data set labels see Table 1.

Table 2. Statistical evidence sorted by strength

Amino acid	PCA factor loading[a]	Raw correlation[a]	Significance	Raw Δ (S.D.)	Δ (S.D.)	Significance	Statistic
Gln (Q)	–90%	–80%	~10^{-8}	–2.18 (0.31)	–1.76 (0.25)	~10^{-4}	*t*-test
Glu (E)	80%	80%	~10^{-6}	2.27 (0.40)	1.73 (0.31)	~10^{-4}	*t*-test
Val (V)	50 to 65%	60%	~10^{-3}	1.57 (0.42)	1.40 (0.38)	~2×10^{-3}	*t*-test
Thr (T)	–65%	60%	~10^{-3}	–0.84 (0.25)	–1.31 (0.39)	~5×10^{-3}	*t*-test[b]
His (H)	–40 to –60%	–60%	~10^{-3}	–0.44 (0.15)	–1.22 (0.42)	1%[b]	*t*-test[b]
Ser (S)	–30 to –60%[b]	–40%	1%	–1.11 (0.51)	–1.18 (0.54)	5%	*t*-test
Asn (N)	–30 to –40%	–35%	3%	–1.94 (n/a)[b]	–1.05 (n/a)[b]	<2%	Median/Mann–Whitney[b]
Arg (R)	20 to 30%[b]	25%	>5%	>0 (n/a)	>0 (n/a)	1%	Mann–Whitney

For each amino acid, the range of PCA factor loadings for component 2, the raw correlation to the binary variable *therm* and its significance are displayed. The Raw Δ column shows the average difference between thermophiles and mesophiles in raw percentage points; Δ gives the equivalent in standardised scores. The last columns report the significance of the observed difference and the test statistics that have been used.

[a]Approximate figures.

[b]See Appendix for discussion.

n/a, not applicable.

Kreil and Ouzounis, Nucleic Acids Res 29(7); 2001

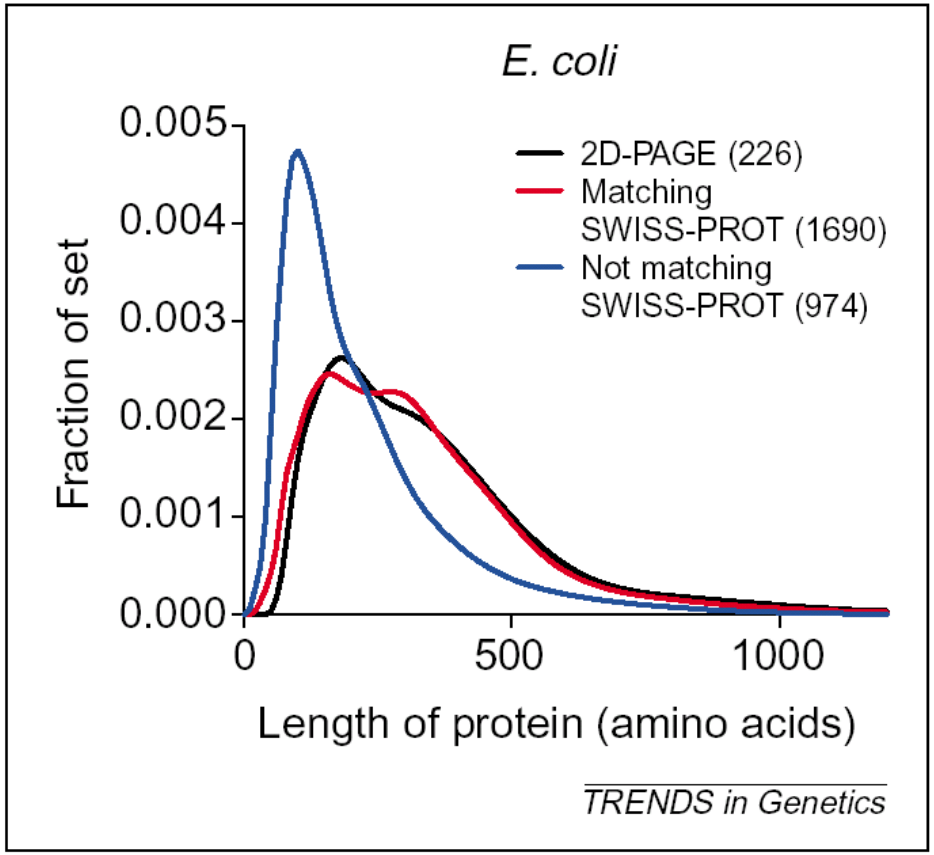
Το μήκος των «πραγματικών πρωτεϊνών» στα πλήρως προσδιορισμένα γονιδιώματα

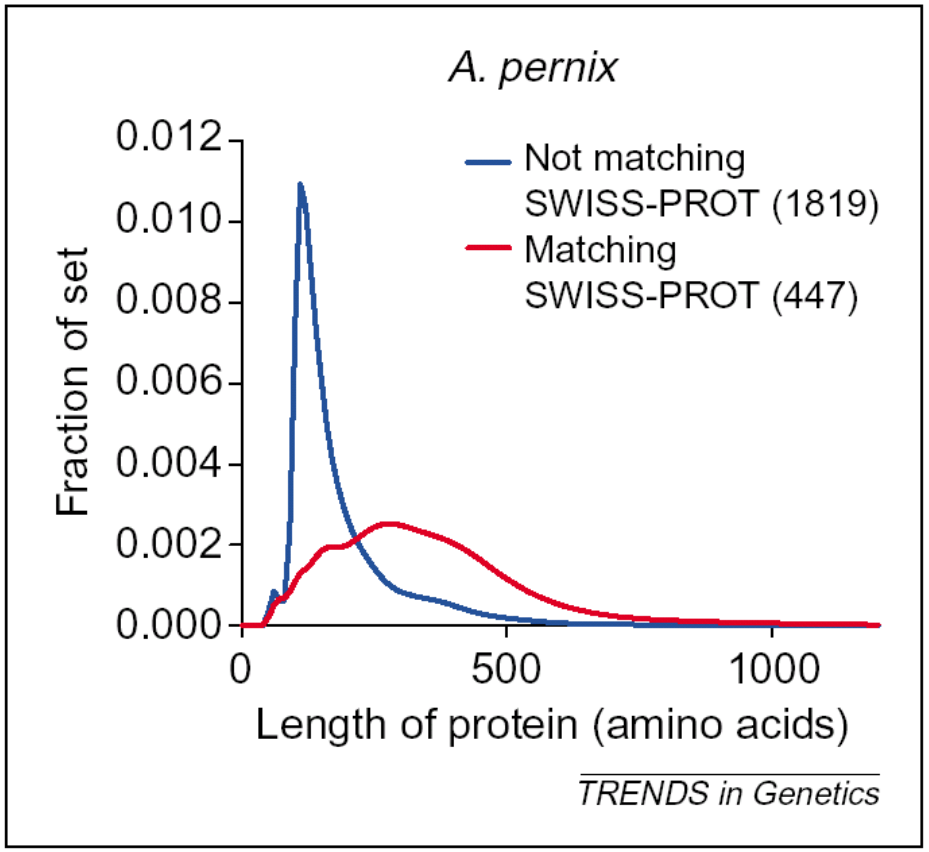
- 64 τριπλέτες (61 για αμινοξέα-3 για λήξη)
- Αν τα νουκλεοτίδια θεωρηθούν ισοπίθανα, έχουμε μια τριπλέτα λήξης περίπου κάθε 21 αμινοξέα
- Οι τριπλέτες λήξης είναι πλούσιες σε AT (TAA, TGA, TAG)
- Κατά συνέπεια το μήκος των «τυχαίων» ORF θα αυξάνει στα πλούσια σε GC γωνιδιώματα

Μεθοδολογία

- Εύρεση όλων ORF από τα βακτηριακά γονιδιώματα (34 εκείνη την εποχή)
- Redundancy Reduction
- Σύγκριση με πραγματικές (non-hypothetical) πρωτεΐνες της SwissProt (E-value<10^{-6})
- Γραφική παράσταση και στατιστική ανάλυση των αποτελεσμάτων

Κατανομή του μήκους





Skovgaard et al, TRENDS in Genetics 17(8); 2001

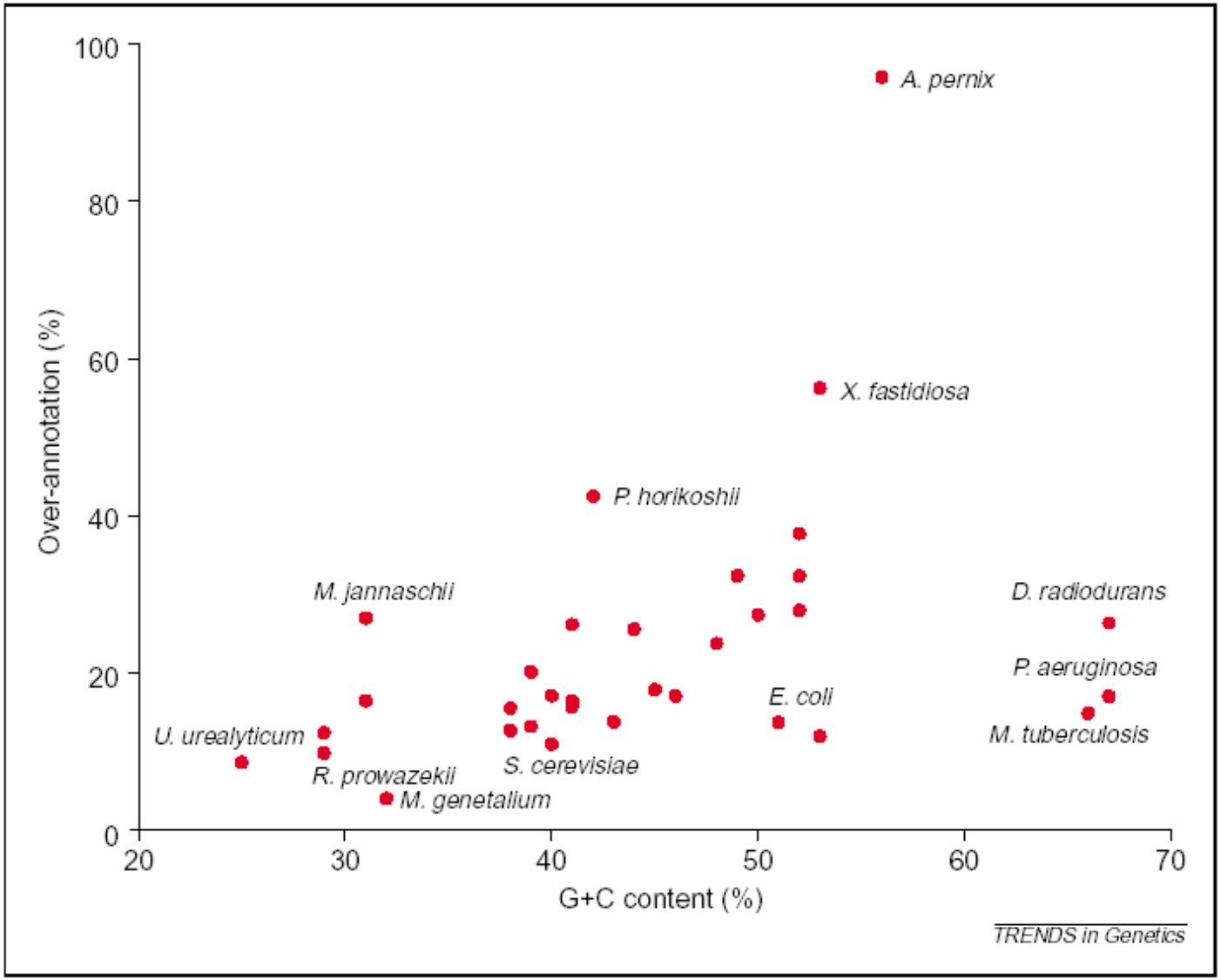
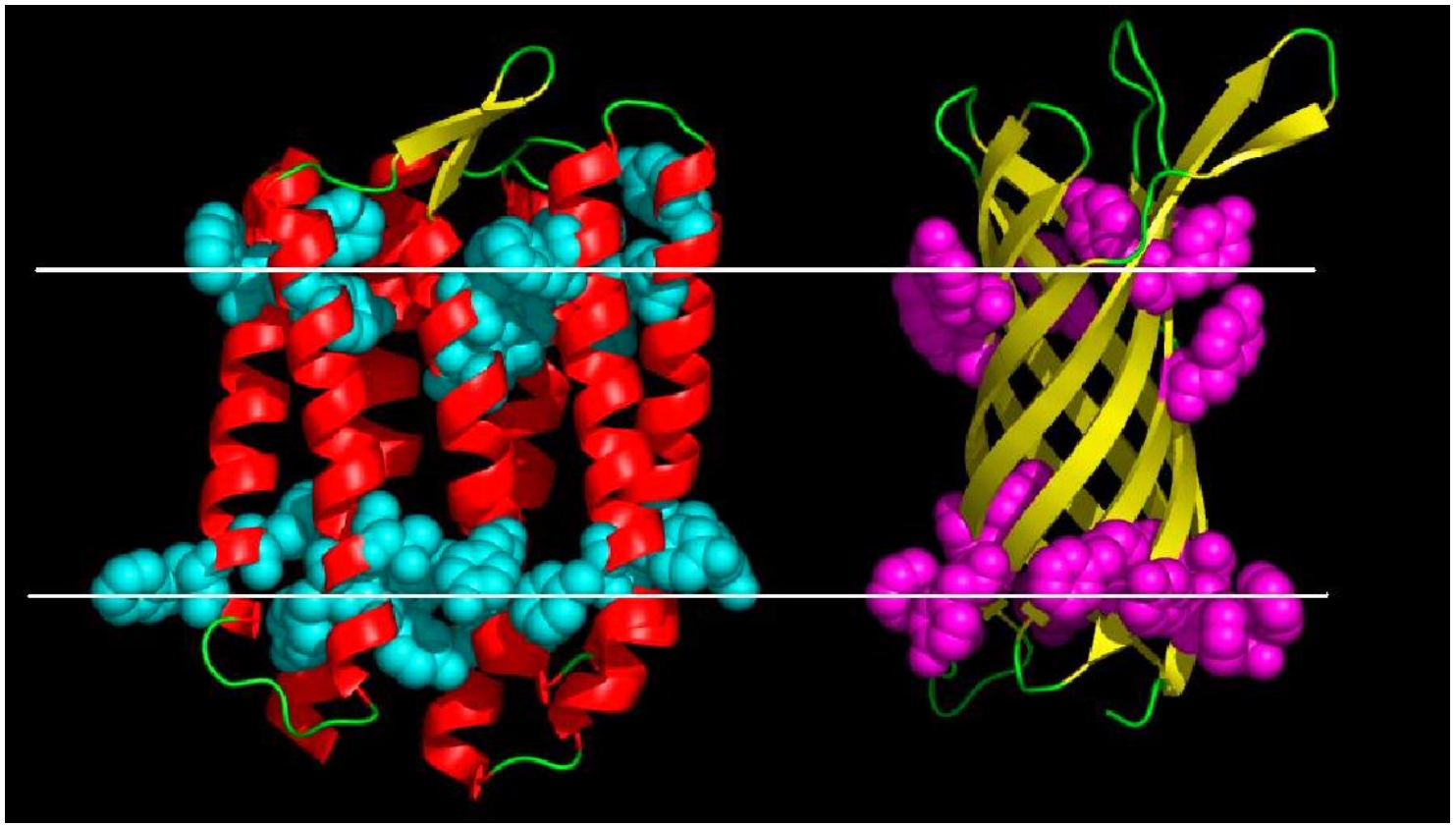


Fig. 1. Estimated over-annotation of genes in sequenced genomes. For each organism the SWISS-PROT-based estimate is calculated and the difference to the number of annotated genes shown in percent of the estimated number of genes.

Skovgaard et al, TRENDS in Genetics 17(8); 2001

Διαμεμβρανικές πρωτεΐνες



GC% content και διαμεμβρανικές πρωτεΐνες

- Η εύρεση των α-ελικοειδών διαμεμβρανικών πρωτεϊνών στηρίζεται στην με διάφορους τρόπους αναζήτηση περιοχών πλούσιων σε υδρόφοβα κατάλοιπα
- Τα γονιδιώματα όμως διαφέρουν στο ποσοστό GC%
- Επιπλέον, τα κωδικόνια των υδρόφοβων αμινοξέων περιέχουν GC σε διαφορετικό βαθμό
- Άρα, ένας «γενικής χρήσης» αλγόριθμος πρόγνωσης μπορεί να υπερ- ή υπό-εκτιμά την πρόγνωση διαμεμβρανικών τμημάτων

Table 1. *A table to show the nucleotide bias in the genomes of the organisms studied*

Organism	%GC/%AT
B. burgdorferi	0.400
R. prowazekii	0.408
M. jannaschii	0.458
M. genitalium	0.464
H. influenzae	0.617
H. pylori	0.636
S. cerevisiae	0.656
M. pneumoniae	0.664
C. pneumoniae	0.683
C. trachomatis	0.704
P. horikoshii	0.721
C. elegans	0.742
A. aeolicus	0.769
B. subtilis	0.770
S. PCC6803	0.913
A. fulgidus	0.945
M. thermoautotrophicum	0.982
E. coli	1.032
T. pallidum	1.118
M. tuberculosis	1.908

Table 2. *A table to show the codons for the abundant transmembrane amino acids and the minimum number of AT and GC bases required to code for the amino acid*

Amino acid	Codons	Min. A + T	Min. G + C
Ala	GCX	0	2
Gly	GGX	0	2
Val	GTX	1	1
Leu	CTX TTA/G	1	1
Ile	ATA/C/T	2	0
Phe	TTC/T	2	0

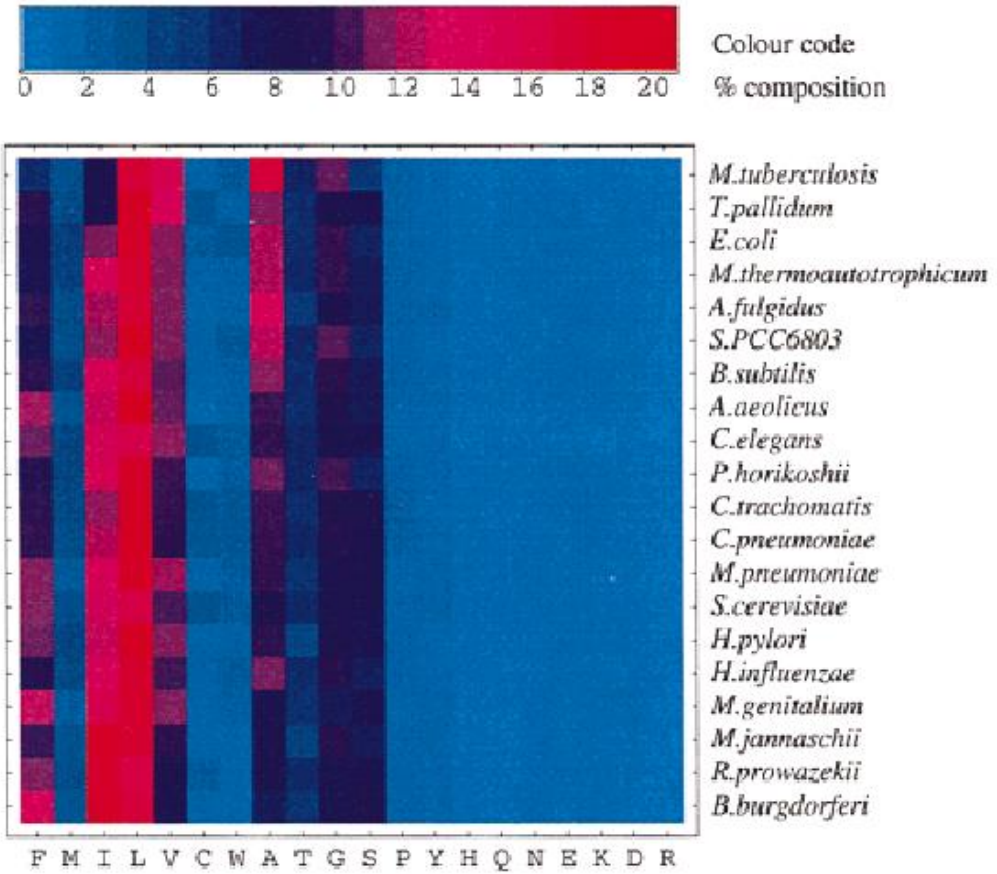


Fig. 1. A plot to show the amino acid composition of the TM domains in each of the proteomes under investigation. Percentage of TM composition exhibited by a residue in each organism increases from blue to red.

Stevens and Arkin, *Protein Science* (2000), *9*:505–511.

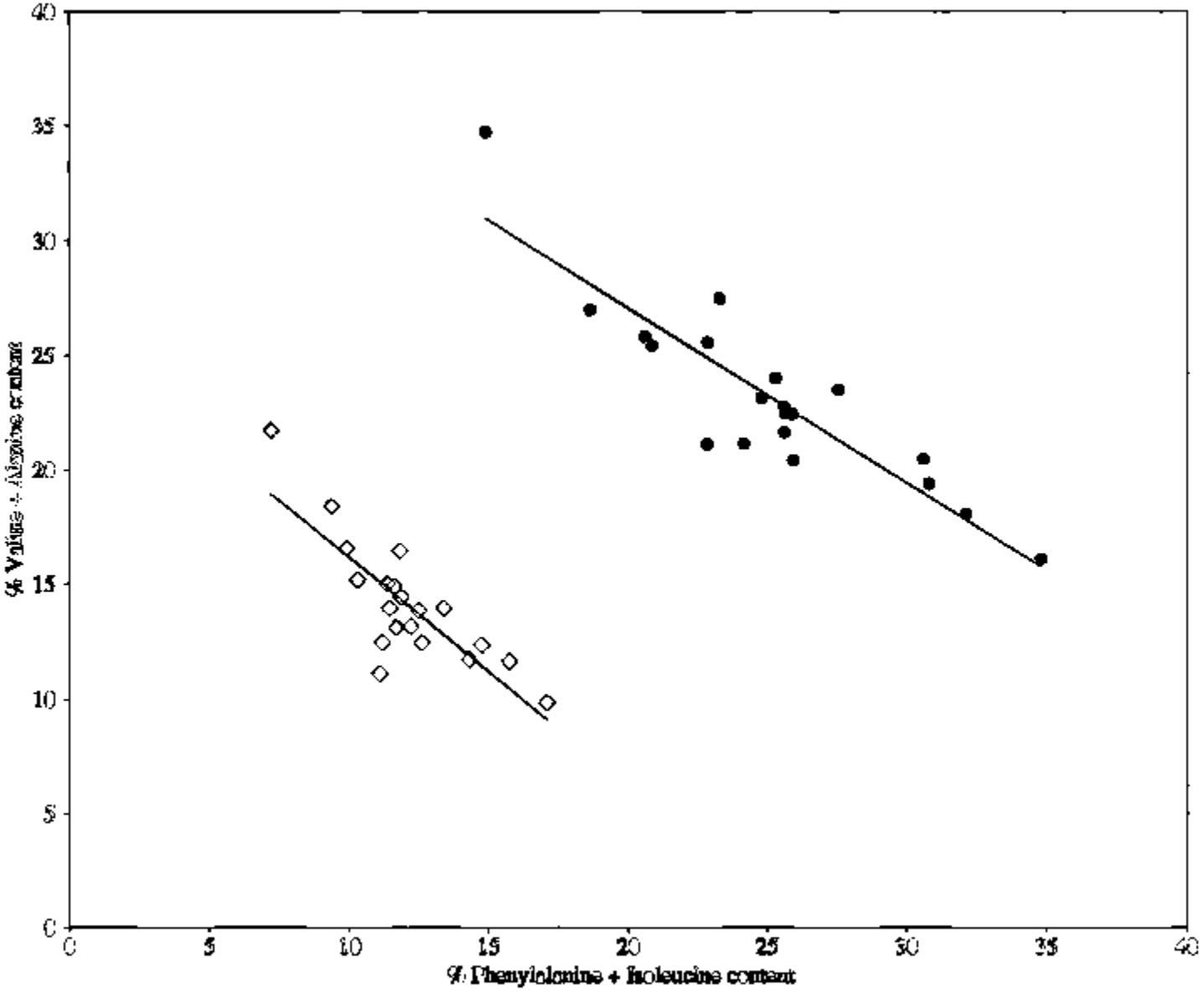


Fig. 2. A graph to show the correlation between the percentage abundance of valine plus alanine residues (VA) and the percentage abundance of phenylalanine plus isoleucine residues (FI). Data are shown for each organism, for both the amino acids within the predicted TM domains (●) and the whole proteome (◇).

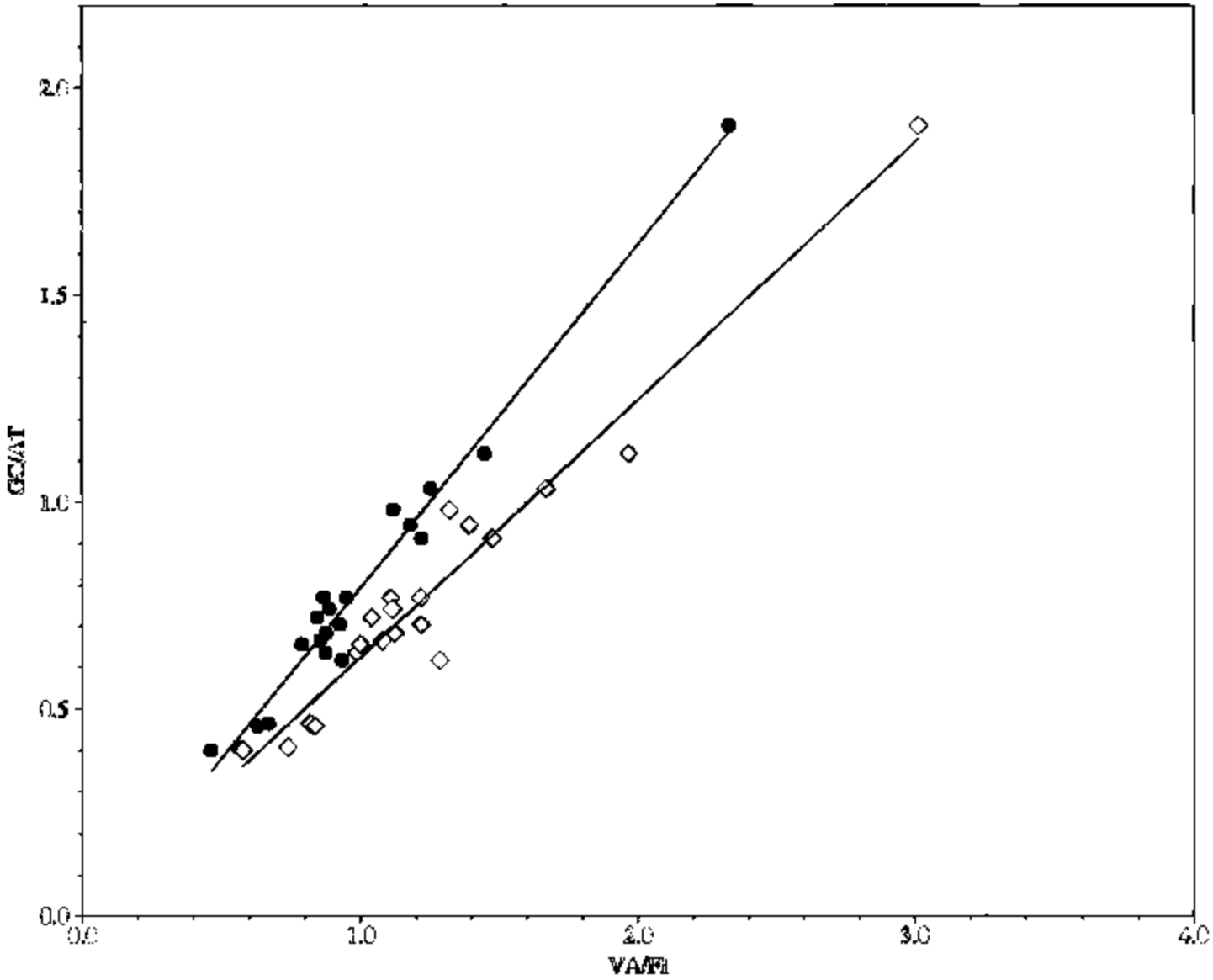
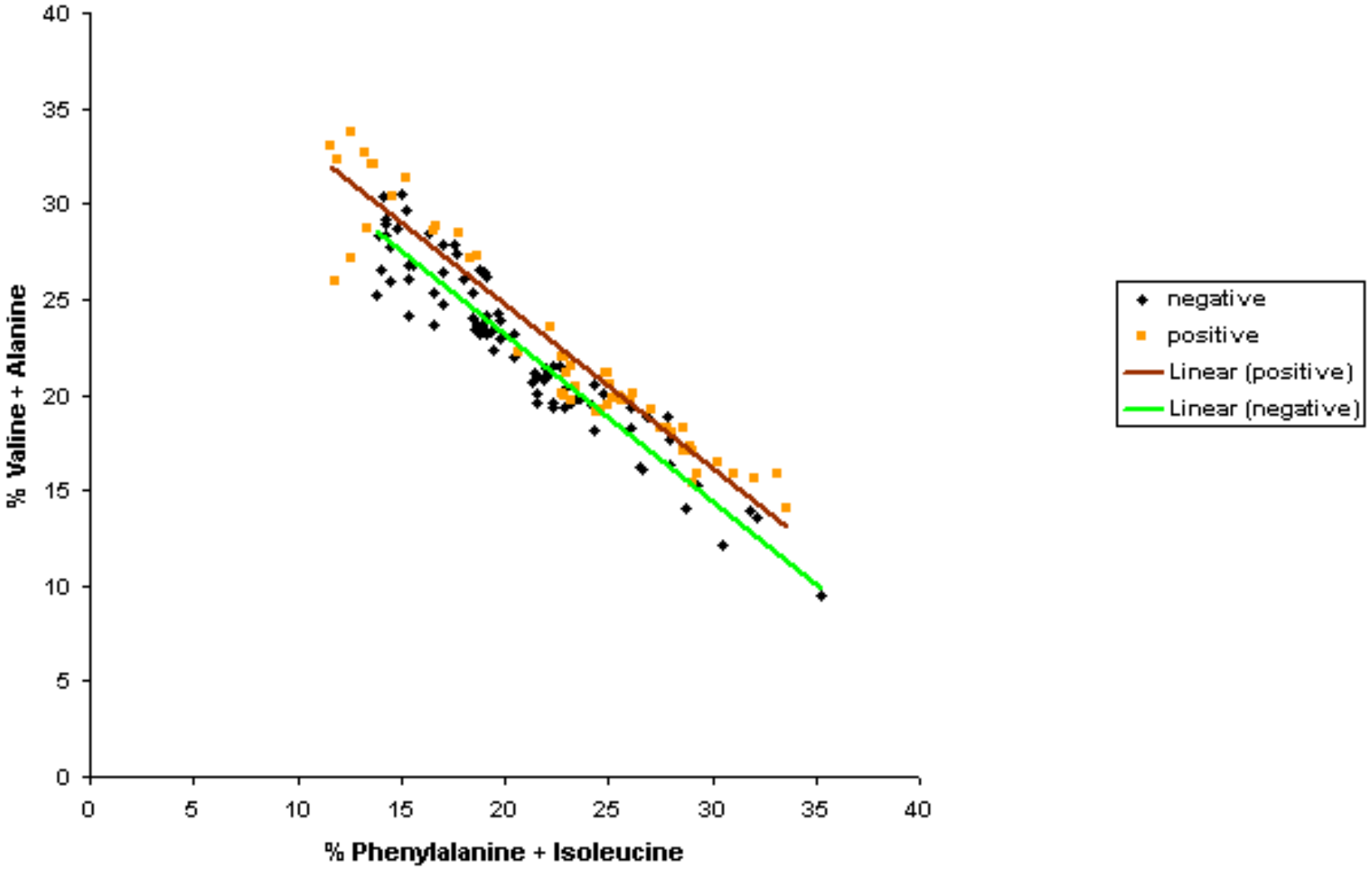
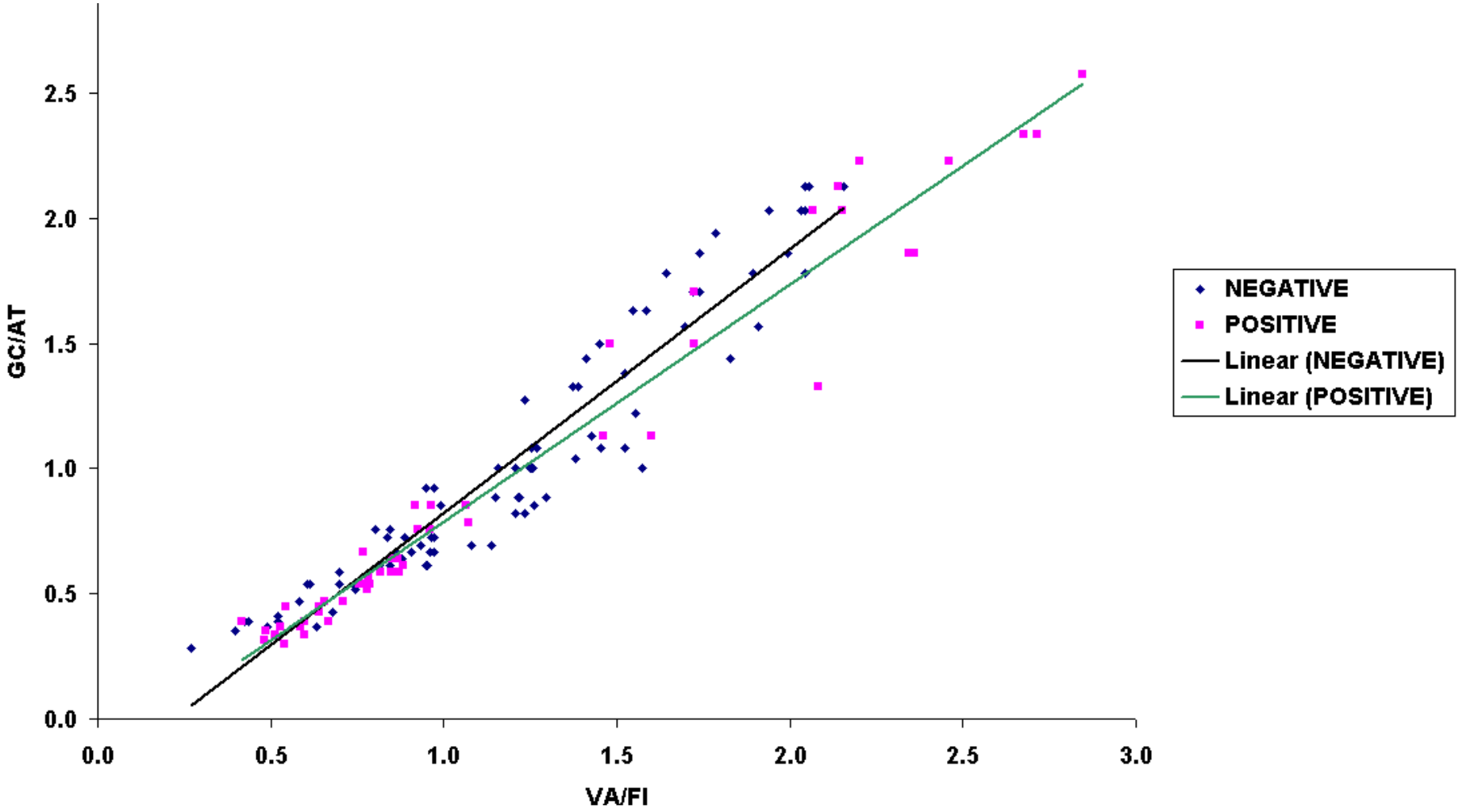


Fig. 3. A graph to show the correlation between the nucleotide bias of a genome (GC/AT) and alternate hydrophobic amino acid use (VA/FI). Data are shown for both the amino acids within the predicted TM domains (●) and for the whole proteome (◇).


% Valine + Alanine
% Phenylalanine + Isoleucine
0
5
10
15
20
25
30
35
40
negative
positive
Linear (positive)
Linear (negative)


GC/AT
VA/FI
0.0
0.5
1.0
1.5
2.0
2.5
3.0
NEGATIVE
POSITIVE
Linear (NEGATIVE)
Linear (POSITIVE)

Μήκος των διαμεμβρανικών πρωτεϊνών και ο εσωτερικός διπλασιασμός

- Ανάλυση όλων ORF από τα βακτηριακά γονιδιώματα (50 στο σύνολο)
- Εύρεση διαμεμβρανικών τμημάτων
- Αφαίρεση πεπτιδίων οδηγητών
- Στατιστική ανάλυση

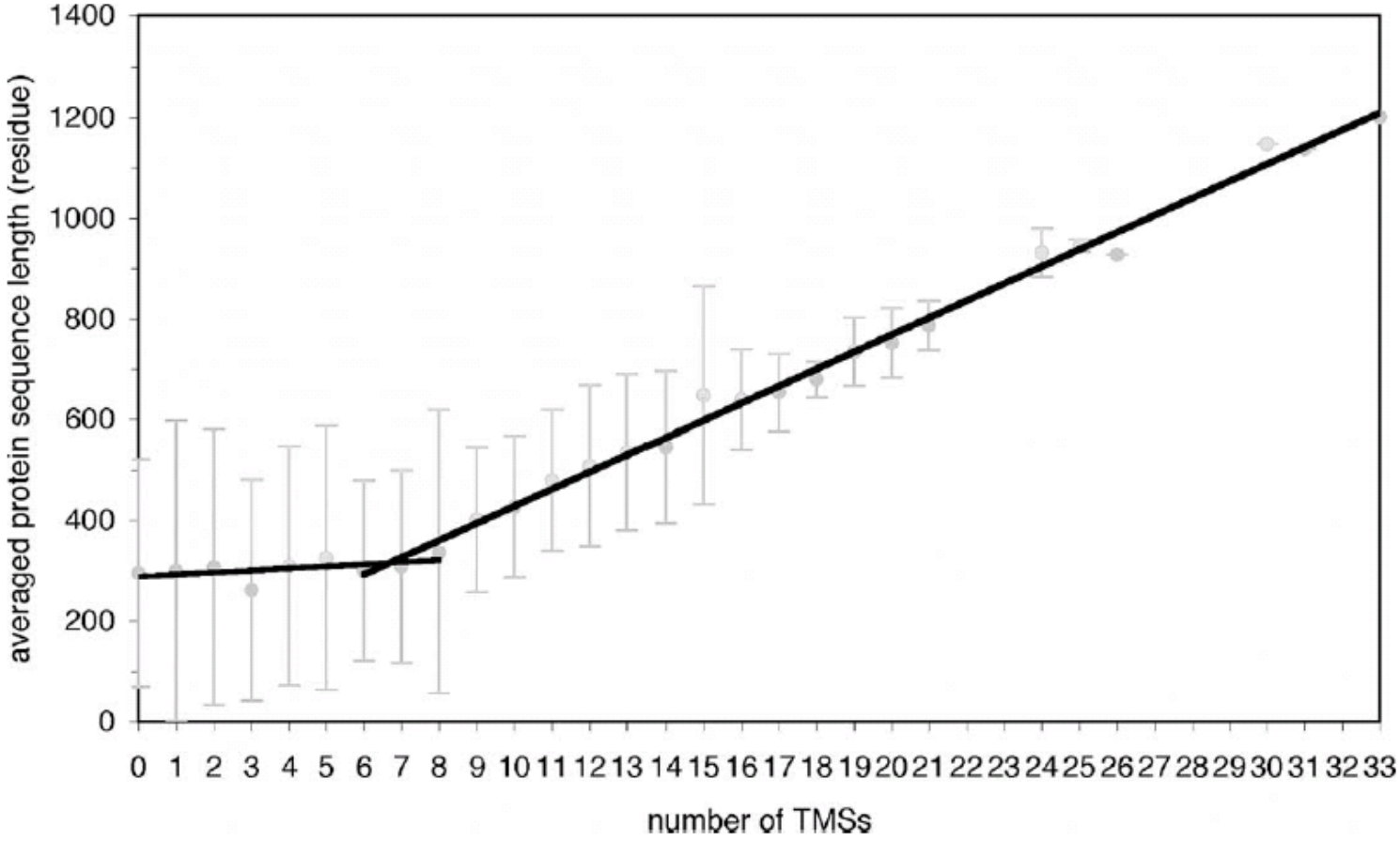
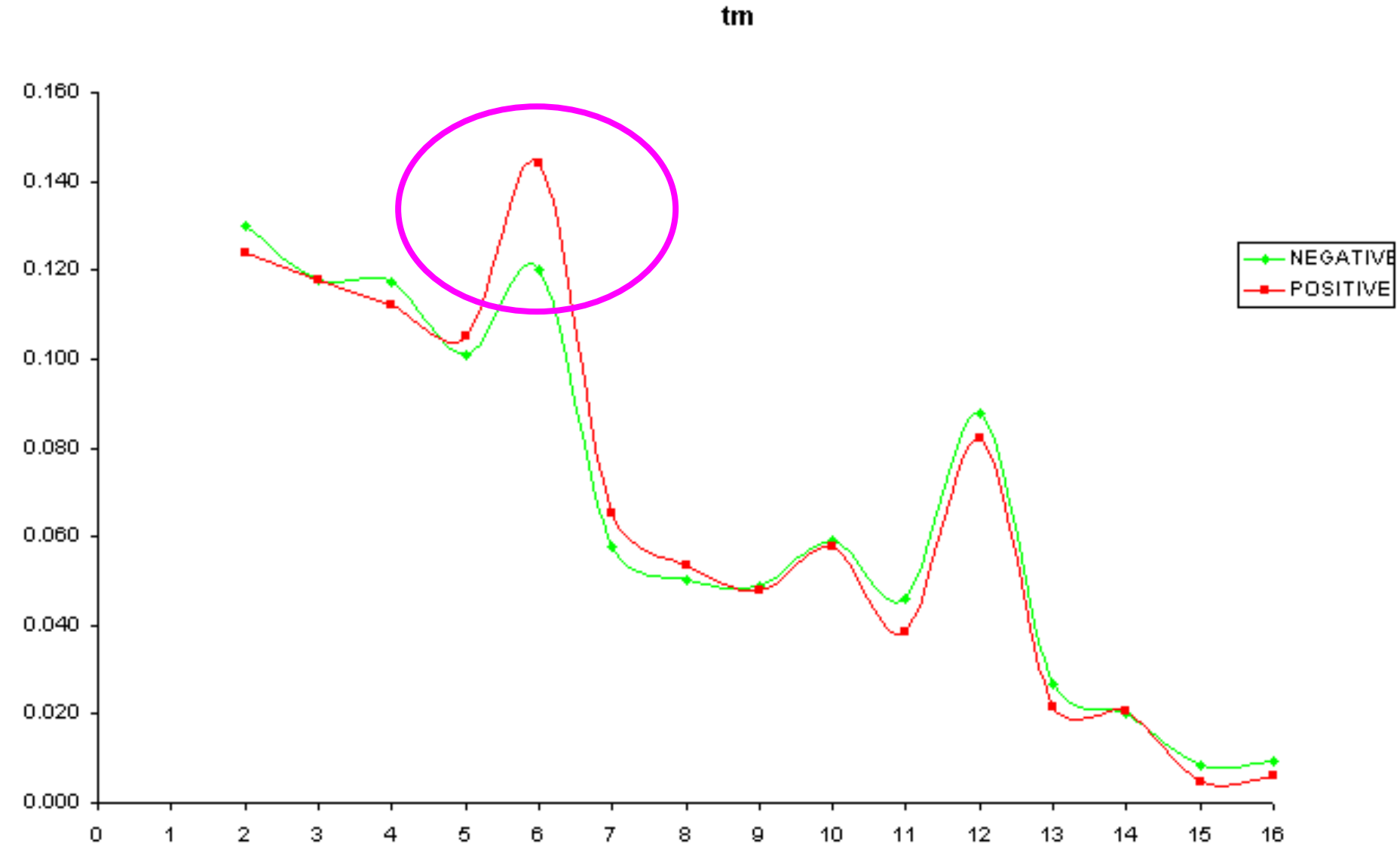
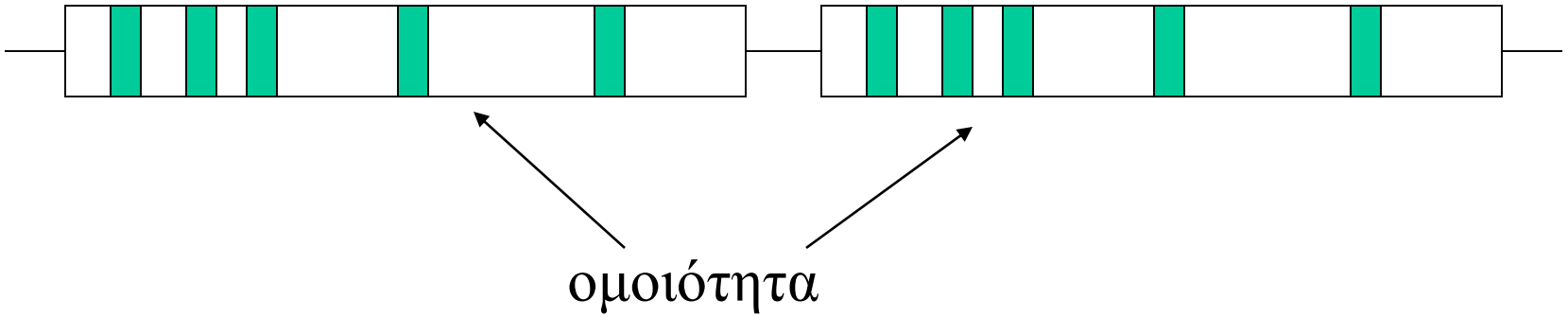


Fig. 4. Relationship between the number of TMSs and the protein sequence length averaged over the 50 genomes. The slope of the line between 7- and 33-tms is 35 residues. The number of TMSs, 0 in the abscissa means soluble proteins.

Arai et al. Gene; 304 (2003) 77–86


tm
0.160
0.140
0.120
0.100
0.080
0.060
0.040
0.020
0.000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
NEGATIVE
POSITIVE

- Ανάμεσα σε 38,174 διαμεμβρανικές πρωτεΐνες από 87 γονιδιώματα, 377 βρέθηκαν να έχουν παραχθεί από ένα μηχανισμό εσωτερικού διπλασιασμού
- Κυρίως με 8, 10 και 12 διαμεμβρανικά τμήματα
- Διάφοροι μηχανισμοί εσωτερικού διπλασιασμού προτάθηκαν, π.χ.:



Shimizu et al, J. Mol. Biol. (2004) 339, 1–15

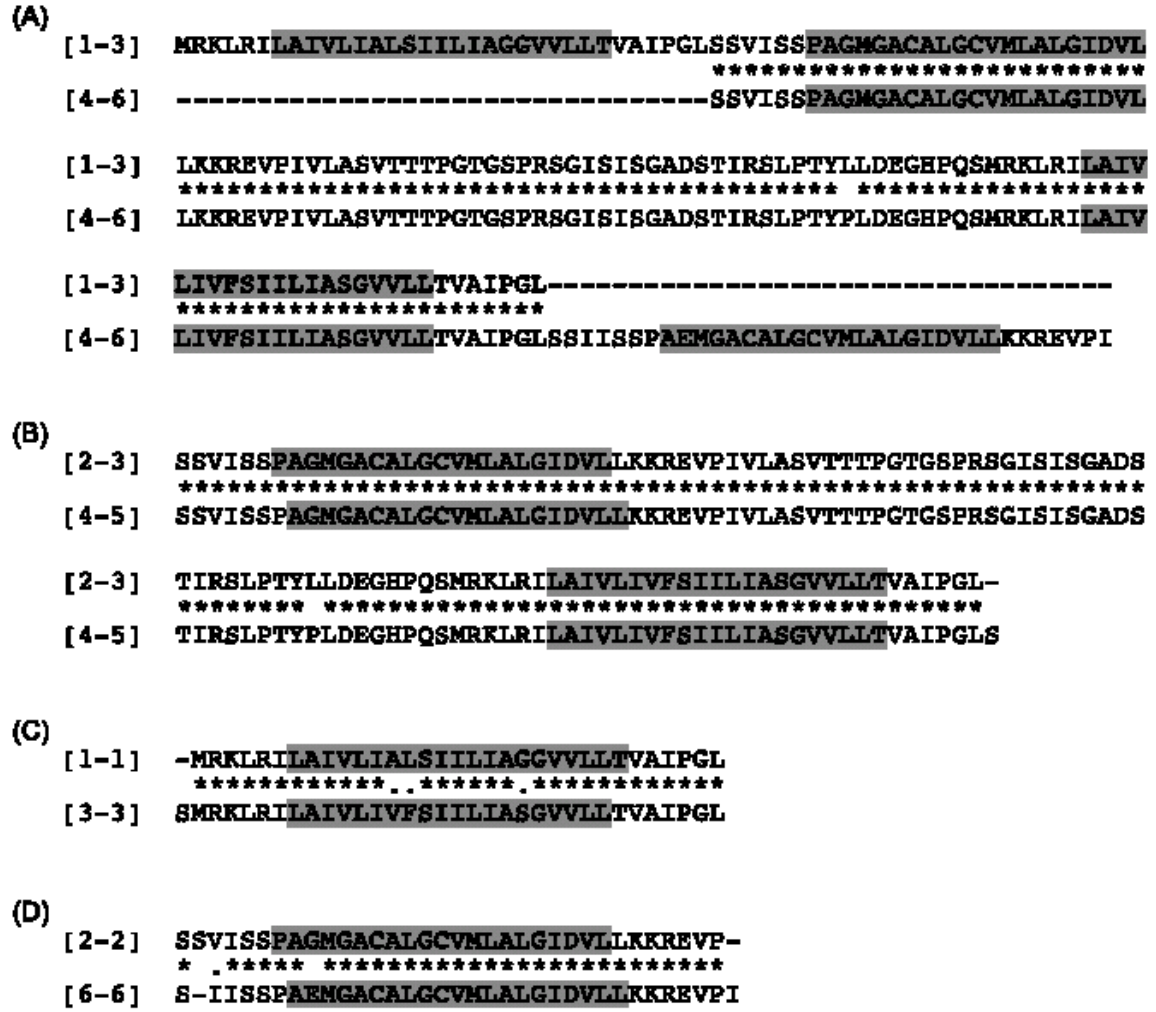


Figure 3. Pairwise alignments between partial sequences of a 6-tms TM protein, CPn0007 (Golgi autoantigen, golgin subfamily A4) from *Chlamydophila pneumoniae* by using the ALIGN program with the setting parameters, i.e. opening gap penalty −12, extension gap penalty −2, and substitution matrix BLOSUM62): (A) {1-2-3} *versus* {4-5-6} (identity, 61.2%). (B) {2-3} *versus* {4-5} (identity, 98.2%). (C) {1} *versus* {3} (identity, 88.2%). (D) {2} *versus* {6} (identity, 88.6%). The shaded boxes indicate the TMSs.

Shimizu et al, J. Mol. Biol. (2004) 339, 1–15

Εύρεση διαμεμβρανικών β-βαρελιών

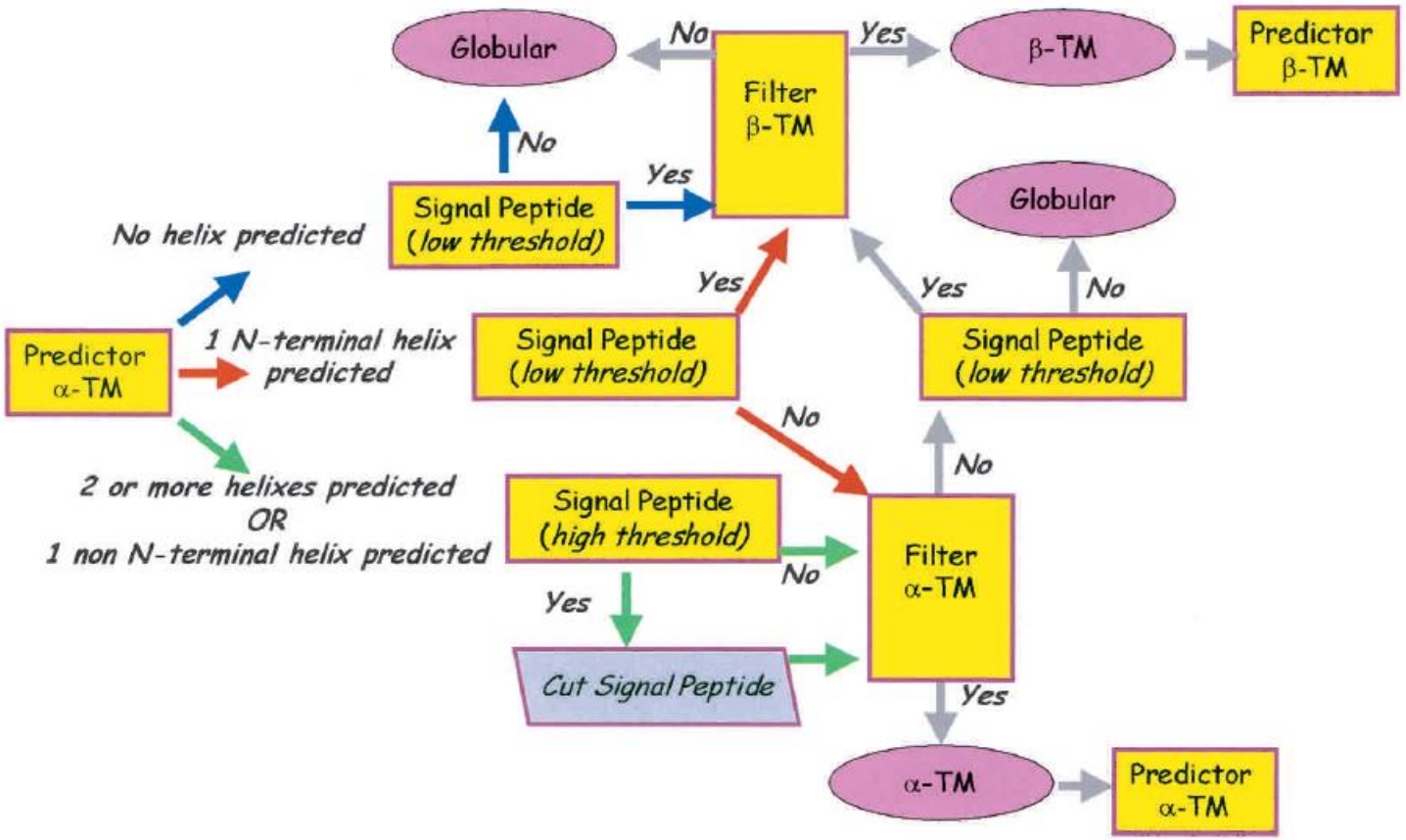


Figure 1. Hunter: The suite of predictors. The flow chart indicates the possible alternatives after the first prediction done with a neural network-based method. Chain flow limiting steps are: a signal peptide predictor (acting with two different threshold values), trained and tested on signal peptides of Gram-negatives; a hidden Markov model-based filter for outer membrane proteins; a neural network-based filter for all α transmembrane proteins. All the predictors are described in the Materials and Methods section. See text for details.

Casadio et al, *Protein Science* (2003), 12:1158–1168.

Table 1. *Predicting well and partially annotated proteins of* Escherichia coli *K12*[a] *with Hunter*

	Prediction			
	α-TM	β-TM	Globular	Total
Well annotated proteins				
Annotation				
α-TM	389	0	33	422
β-TM	0	28	6	34
Globular	50	3	1651	1704
Total	439	31	1690	2160
Partially annotated proteins				
Annotation				
α-TM	317	0	35	352
β-TM	0	14	4	18
Globular	15	2	373	390
Total	332	16	412	760

[a] Annotation of *Escherichia coli* K12 is according to EcoGene (Rudd 2000).

Table 4. *Fishing new globular, inner, and outer membrane proteins in the* E. coli *0157 genome with Hunter*

New globular proteins	1564
New inner membrane proteins	327
New outer membrane proteins	10

NCBI code	Homolog[a] in *E. coli* K12	Length	No. of predicted TM strands	No. of other homologous in Swiss-Prot	Annotation of homologs (first homolog, % identity of local and global alignments)
13359635	UP05_ECOLI	810	18	5	Surface antigen (D152_HAEIN: 45%; 45%)
13359780	YAGZ_ECOLI	195	2	0	
13360600	YMCA_ECOLI	698	20	1	Probable lipoprotein (YJBH_ECOLI: 65%; 64%)
13361464	OMPN_ECOLI	123	4	24	Outer membrane porin (OMS2_SALTI: 85%; 26%)
13361566	YDDB_ECOLI	790	24	1	Hypothetical protein (YDDB_HAEIN: 26%; 23%)
13361895	YDIY_ECOLI	252	12	0	
13362260	CIRA_ECOLI	715	14	22	Colicin receptor; TonB dependent transport (Y262_HAEIN: 24%; 23%)
13362608	YFAZ_ECOLI	187	8	0	
13364489	YJBH_ECOLI	698	22	1	Hypothetical protein (YMCA_ECOLI: 65%; 64%)
13364675	YTFM_ECOLI	577	12	1	Hypothetical protein (YTFM_HAEIN: 44%; 42%)

[a] Homolog = with an E-value $\leq 10^{-7}$.

Table 5. *Predicting globular, inner, and outer membrane proteins in genomes of Gram-negative bacteria with Hunter*

Organism	Outer membrane	Inner membrane	Globular	Total
Escherichia coli K12	65 (1.6%)	907 (21.7%)	3201 (76.7%)	4173
New[a]	18	136	1099	1253
Escherichia coli O157:H7	78 (1.5%)	1034 (19.3%)	4249 (79.2%)	5361
New	10	327	1564	1901
Chlamidia pneumoniae CWL029	12 (1.1%)	290 (27.6%)	750 (71.3%)	1052
New	2	181	236	419
Salmonella typhimurium LT2	70 (1.6%)	1002 (22.5%)	3379 (75.9%)	4451
New	0	2	21	23
Neisseria meningitidis MC58	34 (1.7%)	372 (18.4%)	1619 (80.0%)	2025
New	6	176	662	844
Helicobacter pylori 26695	36 (2.3%)	352 (22.5%)	1178 (75.2%)	1566
New	10	141	445	596
Haemophylus influentiae Rd	23 (1.3%)	348 (20.4%)	1338 (78.3%)	1709
New	5	121	430	556
Thermotoga maritima	18 (1.0%)	370 (20.0%)	1458 (79.0%)	1846
New	11	203	559	773
Pseudomonas aeruginosa	131 (2.4%)	1292 (23.2%)	4142 (74.4%)	5565
New	62	616	1867	2545

[a] The number of new proteins predicted in the class with Hunter out of the nonannotated region.

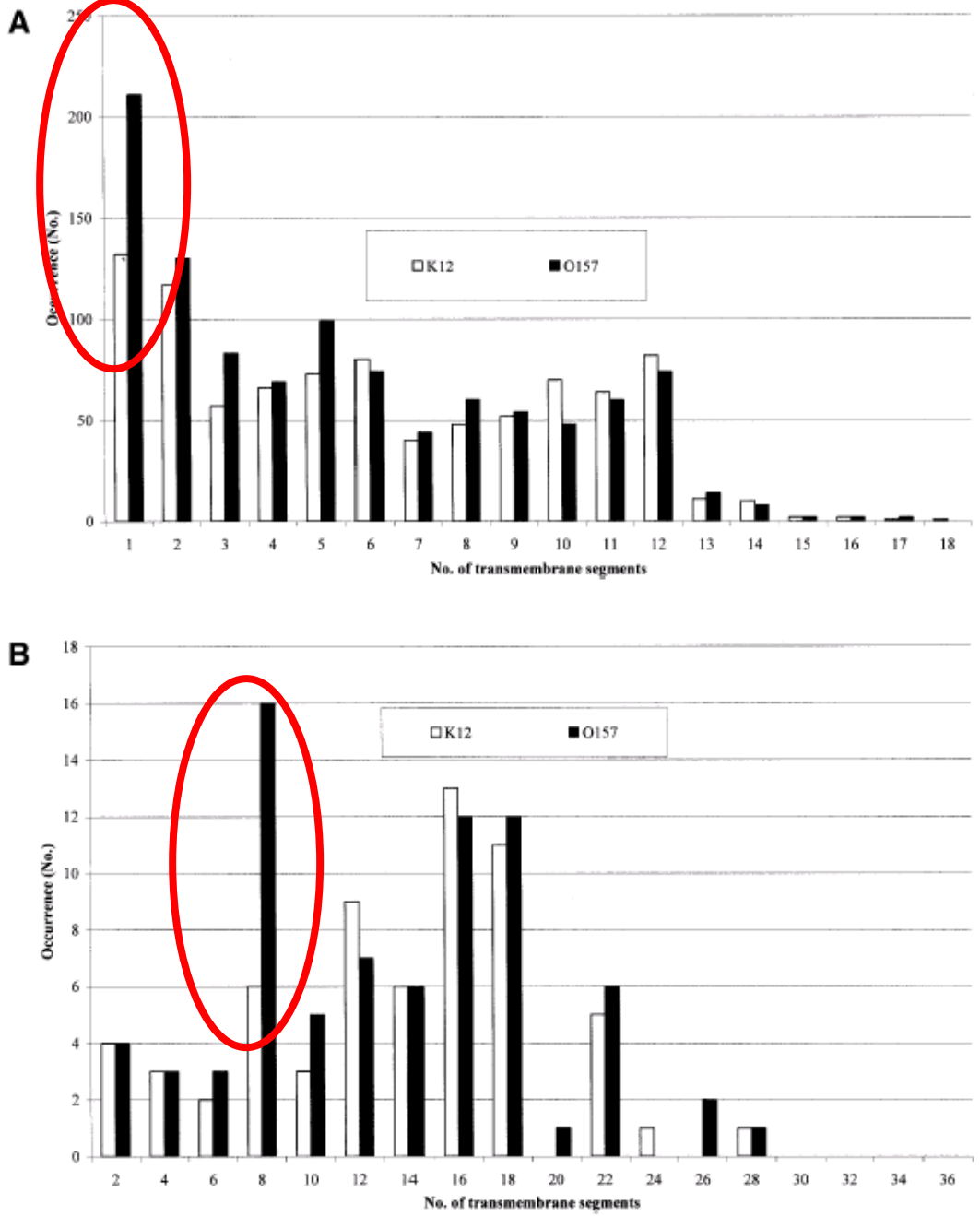


Figure 3. Topography of transmembrane proteins in *E. coli* K12 and O157 as predicted with Hunter. (*A*) Bar plot of inner membrane proteins as a function of the number of transmembrane predicted segments in both strains. (*B*) Bar plot of outer membrane proteins as a function of transmembrane predicted β strands in the barrel in both strains.