Ειδικά Θέματα Βιοπληροφορικής

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας
Λαμία, 2015

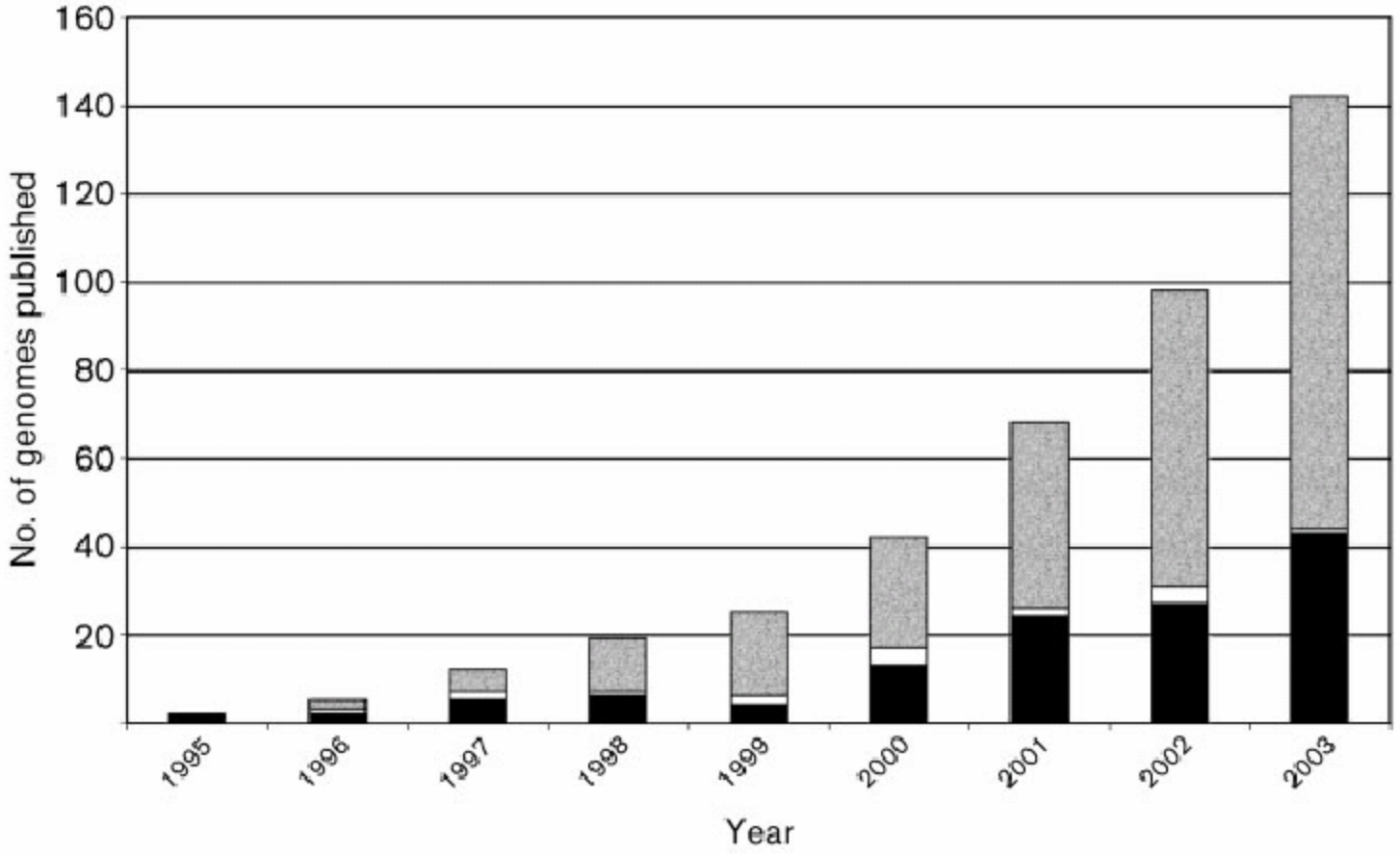


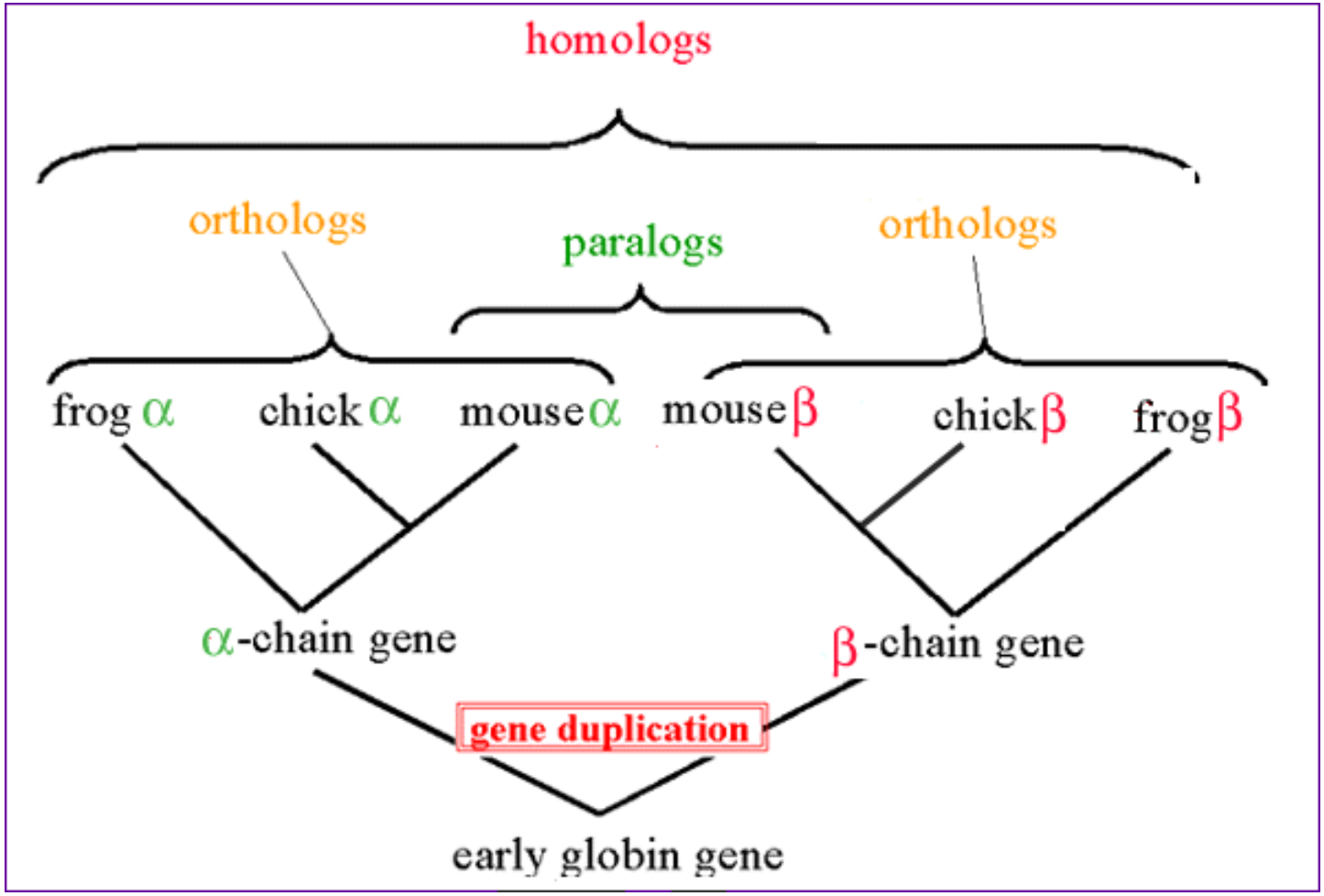
Fig. 1. Number of prokaryotic genomes sequenced each year since 1995. Black, bacterial genomes; white, archaeal genomes; grey, running total.

Μέθοδοι Ανάλυσης

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

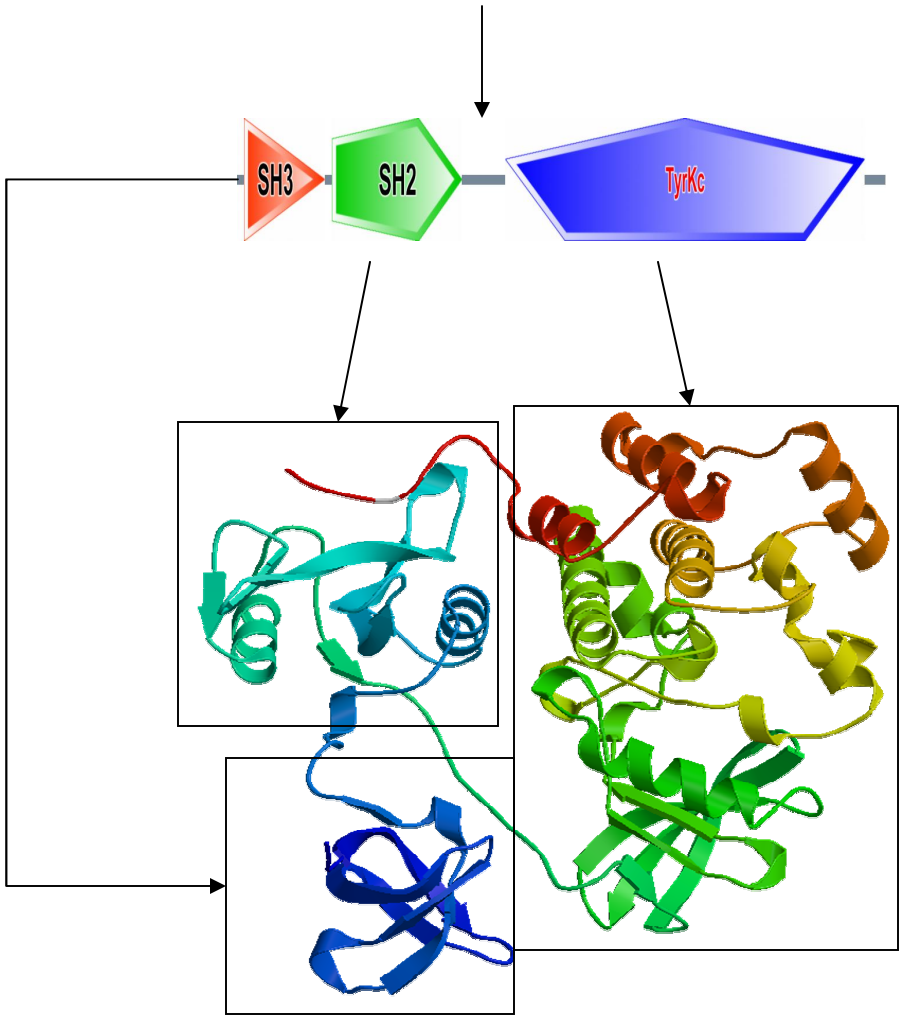
Κατά ζεύγη στοίχιση ακολουθιών

- 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

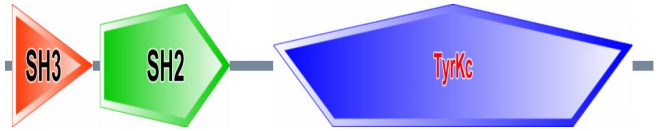

homologs
orthologs
paralogs
orthologs
frog α
chick α
mouse α
mouse β
chick β
frog β
α-chain gene
β-chain gene
gene duplication
early globin gene

Protein Domains

SGIRIIVVALYDYEAIHHEDLSFQKGDQMVVLEESGEWWKARSLATRKEGYIPSNYVARV
DSLETEEWFFKGISRKDAERQLLAPGNMLGSFMIRDSETTKGSYSLSVRDYDPRQGDTVK
HYKIRTLDNGGFYISPRSTFSTLQELVDHYKKGNDGLCQKLSVPCMSSKPQKPWEKDAWE
IPRESLKLEKKLGAGQFGEVWMATYNKHTKVAVKTMKPGSMSVEAFLAEANVMKTLQHDK
LVKLHAVVTKEPIYIITEFMAKGSLLDFLKSDEGSKQPLPKLIDFSAQIAEGMAFIEQRN
YIHRDLRAANILVSASLVCKIADFGLARVIEDNEYTAREGAKFPIKWTAPEAINFGSFTI
KSDVWSFGILLMEIVTYGRIPYPGMSNPEVIRALERGYRMPRPENCPEELYNIMMRCWKN
RPEERPTFEYIQSVLDDFYTATESQEEIP



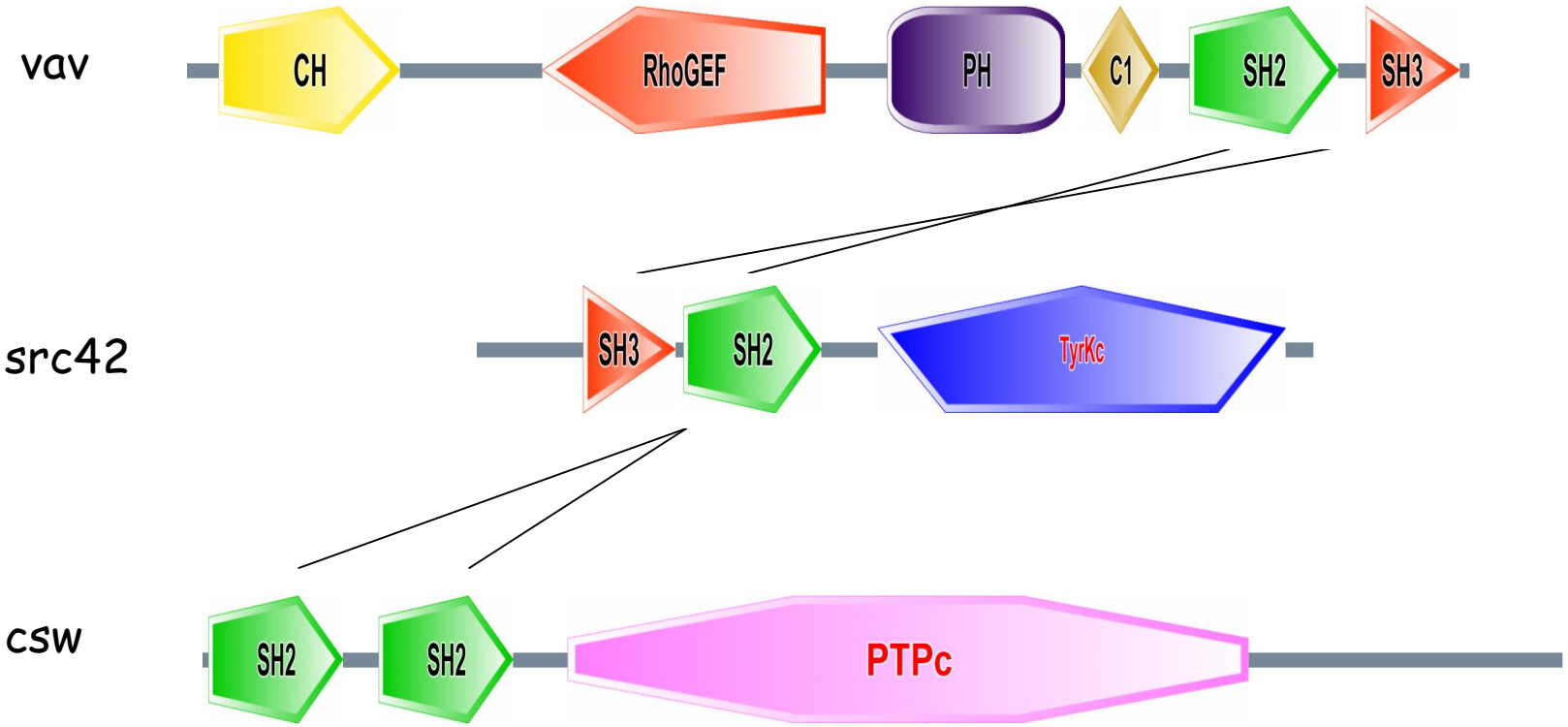
Protein Families



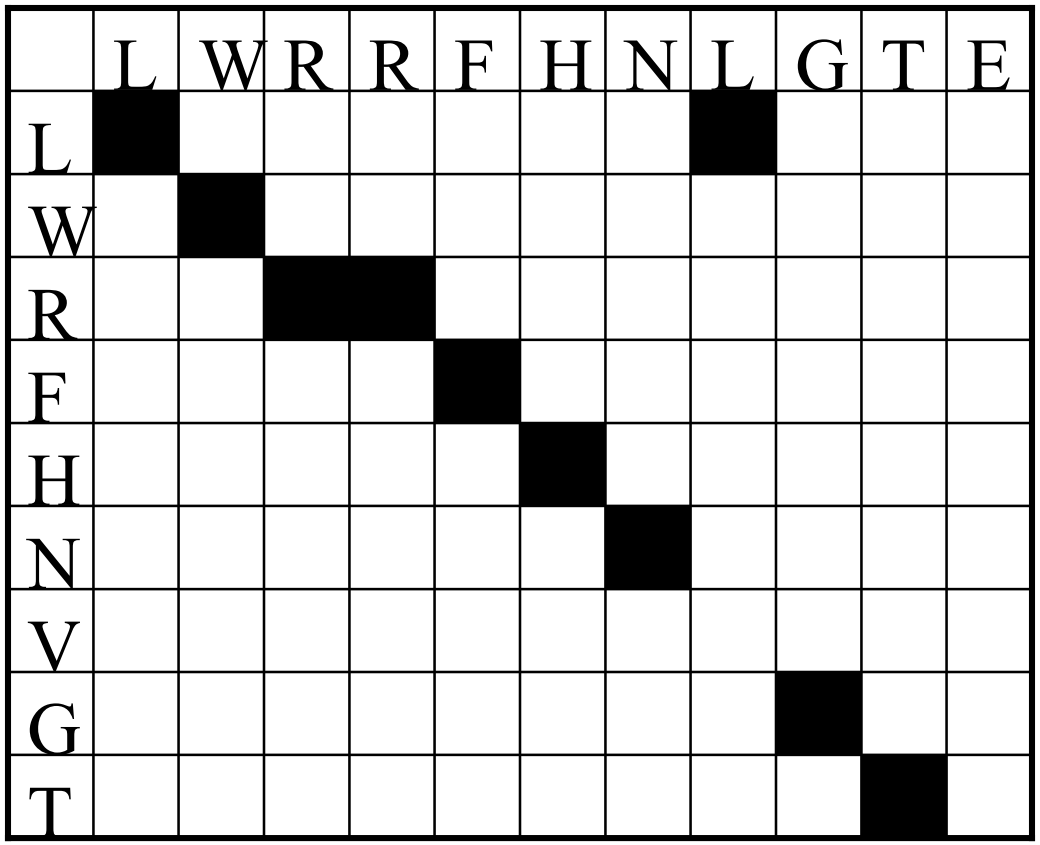
src-like protein tyrosine kinase - 5 in *Drosophila* proteome

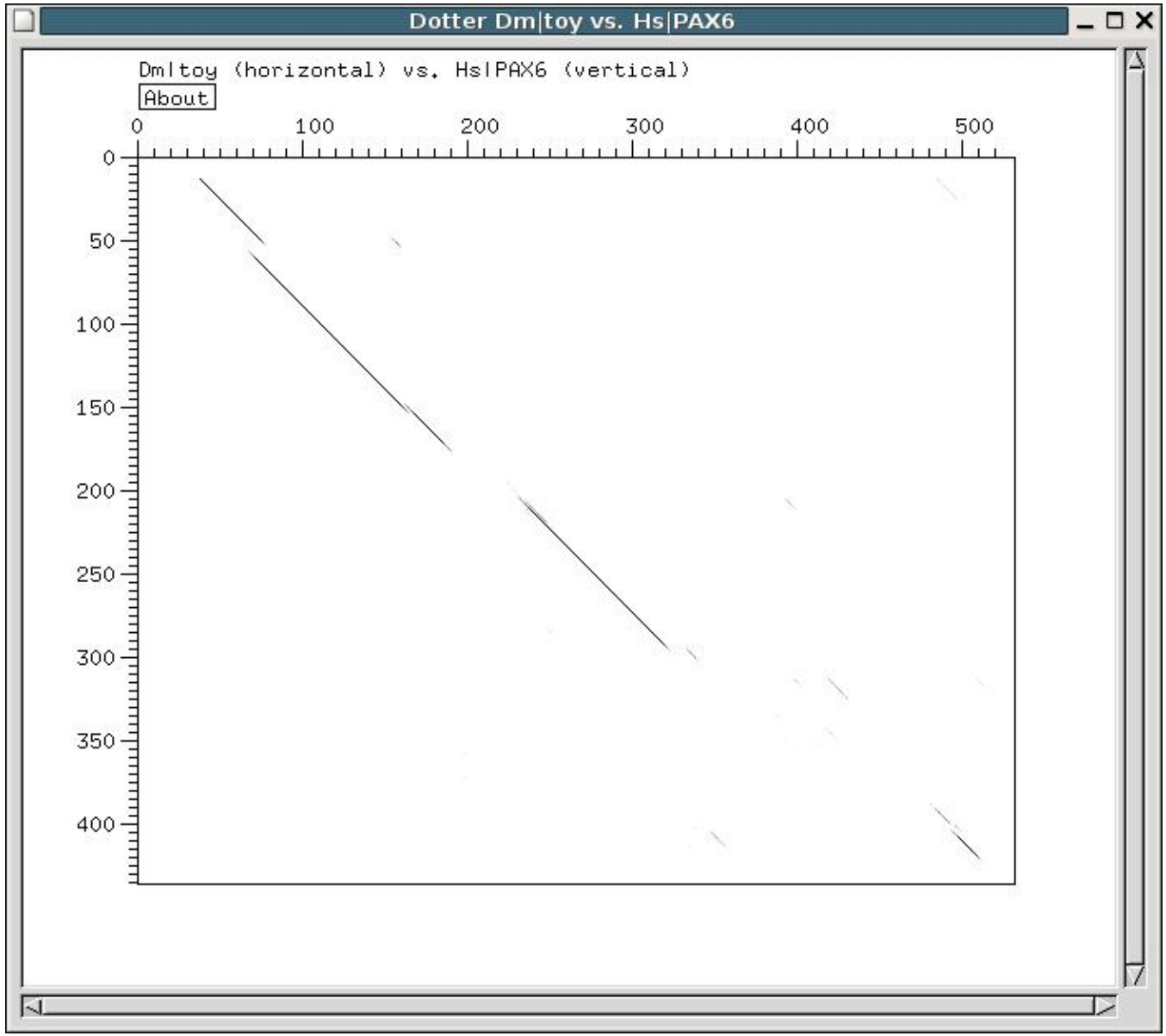
38 tyrosine kinases
43 SH2 domain containing
110 SH3 domain containing

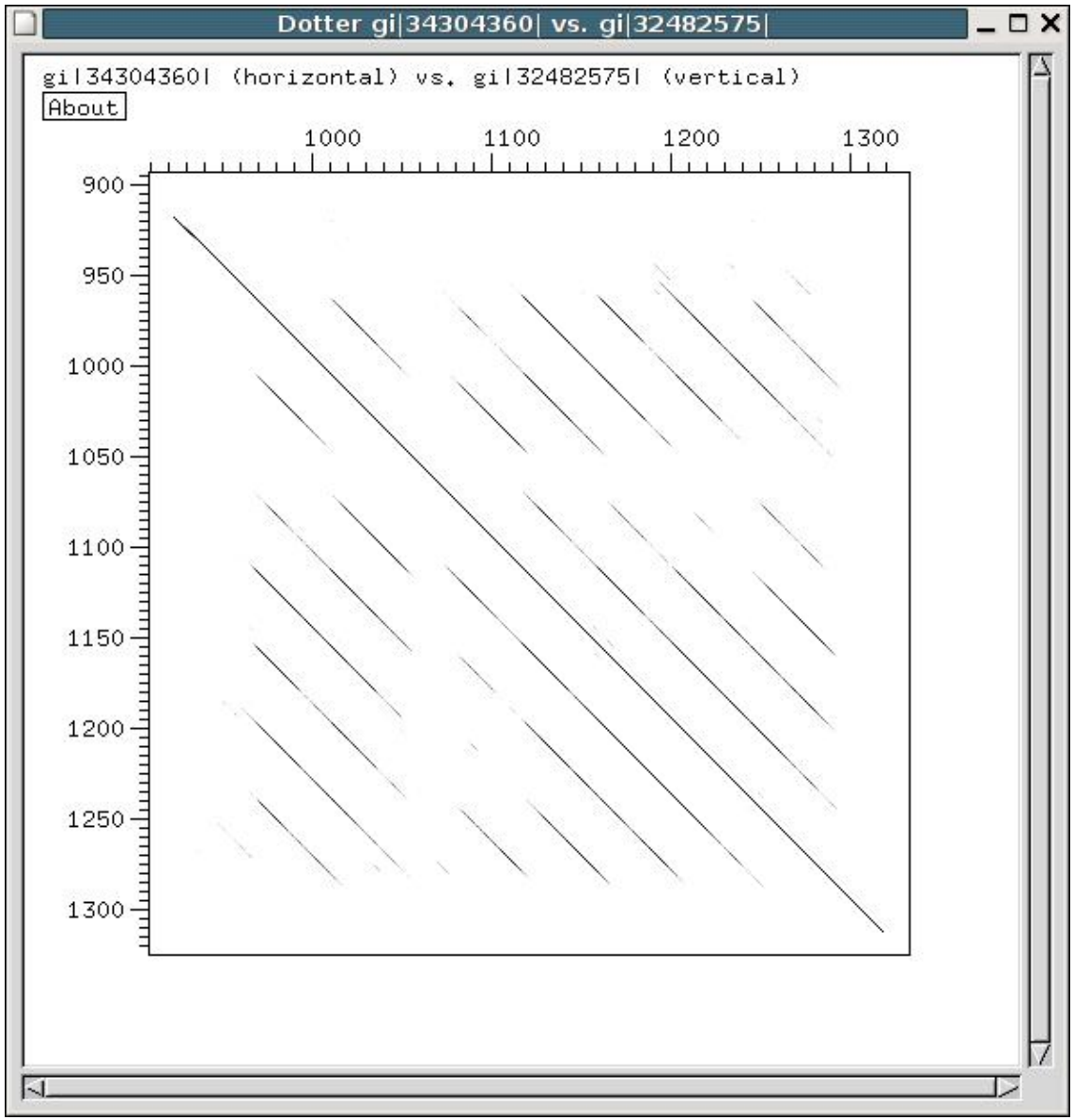
Local Similarity

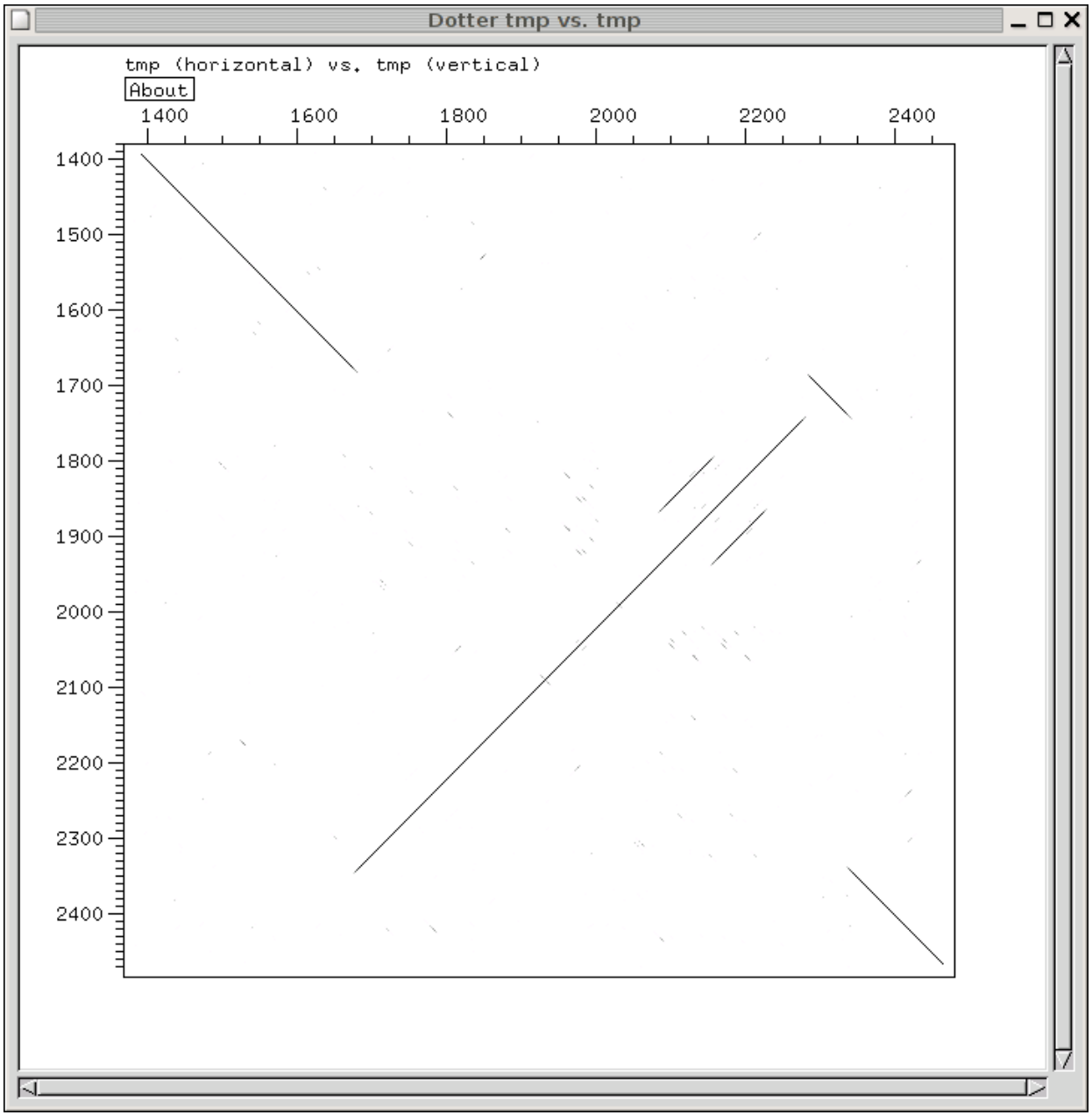


Dotplot




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

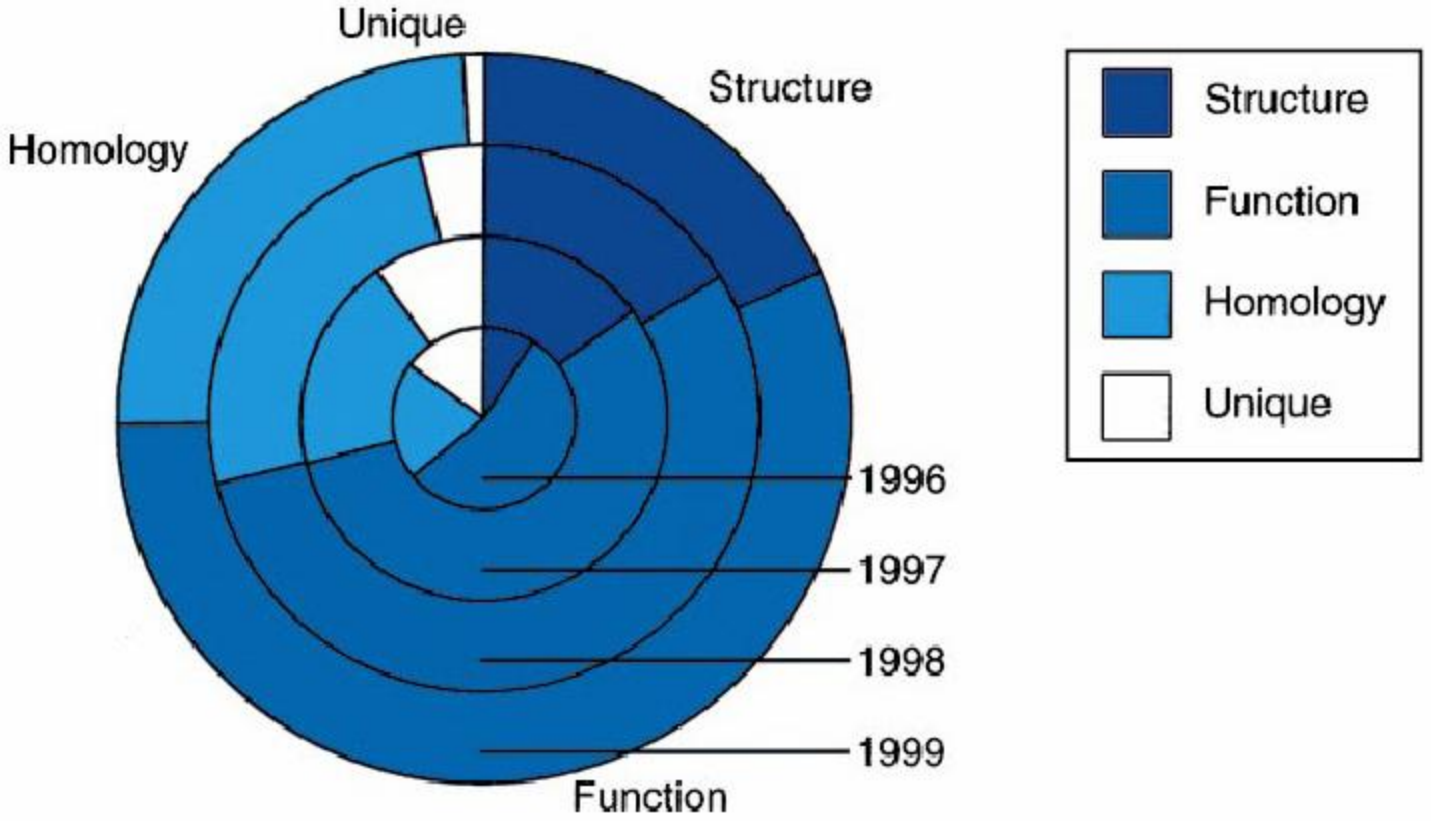


Fig. 2. The 'function bottleneck'. Information clock illustrating the improvement of annotations identified for a given genome over the years 1996–1999. The four levels of annotation range from homologues of known structure (blue) and homologues of known function (marine) to homologues of unknown function (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/.

Γονιδιωματική

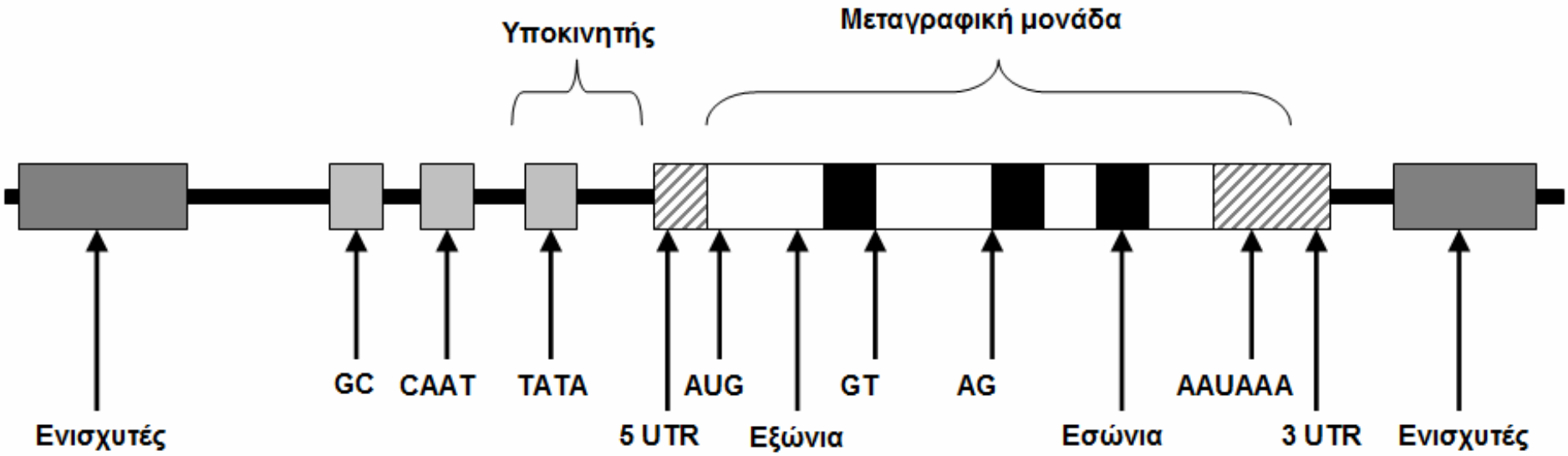
- Γονιδιωματική, ονομάζουμε τον επιστημονικό κλάδο ο οποίος χρησιμοποιεί διαφορετικές τεχνικές της γενετικής, της μοριακής βιολογίας και της βιοπληροφορικής με σκοπό να βρει την αλληλουχία, να κάνει την συναρμολόγηση και να αναλύσει τη δομή και τη λειτουργία των γονιδιωμάτων, δηλαδή, ολόκληρης της γενετικής πληροφορίας που περιέχεται σε ένα κύτταρο ενός οργανισμού. Υπάρχουν πολλές υποδιαιρέσεις της γονιδιωματικής, κυρίως όσον αφορά τις διαφορετικές τεχνικές που είναι δυνατό να χρησιμοποιηθούν κάθε φορά. Για παράδειγμα, η δομική γονιδιωματική ασχολείται με το μαζικό προσδιορισμό τρισδιάστατων δομών πρωτεϊνών από ολόκληρα γονιδιώματα, ενώ η λειτουργική γονιδιωματική ασχολείται κυρίως με τη μελέτη των λειτουργικών περιοχών στα γονιδιώματα (υποκινητές, μικρά RNA κλπ).

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

Gene finding

- Το πιο βασικό πρόβλημα στην περίπτωση αλληλουχιών DNA είναι αυτό της εύρεσης γονιδίων (gene finding), αλλά και αυτό μπορεί να αντιμετωπιστεί με πολλούς τρόπους ενώ μπορεί και να χωριστεί σε μικρότερα «υπο-προβλήματα» (Mathé, Sagot, Schiex, & Rouze, 2002).
- Η εύρεση των πραγματικών γονιδίων που κωδικοποιούνται σε ένα γονιδίωμα, είναι τεράστιας σημασίας πρόβλημα, γιατί όπως έχουμε πει, η αλληλούχιση ενός γονιδιώματος είναι μεν μια δουλειά ρουτίνας, αλλά αυτό δεν σημαίνει ότι και αυτόματα θα έχουμε γνώση των πρωτεϊνών που κωδικοποιεί αυτό το γονιδίωμα. Η εύρεση απλά των ανοιχτών πλαισίων ανάγνωσης, είναι μια σχετικά απλή διαδικασία (ειδικά στους προκαρυωτικούς οργανισμούς), αλλά ακόμα και έτσι υπάρχουν πάρα πολλά ψευδογονίδια ή περιοχές που απλά έτυχε να έχουν το κωδικόνιο έναρξης και λήξης σε διαφορά φάσης (σε απόσταση νουκελοτιδίων που είναι πολλαπλάσιο του 3).
- Έτσι, η εύρεση των κατάλληλων ρυθμιστικών περιοχών (υποκινητές) που καθορίζουν την έκφραση του γονιδίου, είναι μια πολύ σημαντική διαδικασία. Στους δε ευκαρυωτικούς οργανισμούς, στους οποίους τα γονίδια είναι διακοπτόμενα από εσώνια και εξώνια, επιφέρει μια επιπλέον πολυπλοκότητα στους υπολογισμούς καθώς οι ρυθμιστικές αυτές περιοχές πρέπει να αναγνωριστούν πριν καν εντοπιστούν τα ανοιχτά πλαίσια ανάγνωσης. Επιπλέον δε, στους ευκαρυωτικούς οργανισμούς υπάρχουν και άλλες ρυθμιστικές αλληλουχίες πιο μακριά από τον υποκινητή, οι οποίες πρέπει να εντοπιστούν.


Υποκινητής
Μεταγραφική μονάδα
Ενισχυτές
GC
CAAT
TATA
5 UTR
AUG
Εξώνια
GT
AG
Εσώνια
AAUAAA
3 UTR
Ενισχυτές

- Έτσι καταλαβαίνουμε ότι μπορεί να υπάρξουν μια σειρά μικρότερα «προβλήματα» για λύση:
  - μπορεί να υπάρχουν μέθοδοι εύρεσης των σημείων αποκοπής και συρραφής των εξωνίων (exon/intron splice site),
  - μέθοδοι αναγνώρισης του υποκινητή (promoter recognition),
  - μέθοδοι αναγνώρισης του σημείου έναρξης της μεταγραφής (translation initiation site prediction) (Saeys, Abeel, Degroeve, & Van de Peer, 2007),
  - μέθοδοι εύρεσης του σημείου πολυαδενυλίωσης στο mRNA (polyadenylation prediction) (Chang et al., 2011),
  - αλλά και, φυσικά, μέθοδοι που προβλέπουν ολόκληρη τη δομή του γονιδίου.

- Τέλος, οι μέθοδοι έχουν και διαφορετικές στατιστικές ιδιότητες Ανάλογα με την ευαισθησία και την ειδικότητα που μπορεί να έχει η κάθε μία, είναι δυνατόν να αποδίδουν καλύτερα είτε σε απομονωμένες περιοχές DNA ή σε πλήρη γονιδιώματα (Saeys et al., 2007).
- Ένα άλλο σημείο που χρειάζεται προσοχή, είναι η ειδικότητα ανά οργανισμό ή ομάδα οργανισμών, καθώς οι στατιστικές ιδιότητες των νουκλεοτιδίων (ακόμα και στο πλαίσιο των αποδεκτών κωδικωνίων) διαφέρουν ανάμεσα στις μεγάλες ομάδες. Έτσι, υπάρχουν εξειδικευμένα εργαλεία για ειδικές περιπτώσεις ή εργαλεία που λαμβάνουν υπόψη τους τη φυλογενετική προέλευση του οργανισμού.
- Γενικά, υπάρχει μια πληθώρα μεθόδων καθώς η σχετική βιβλιογραφία είχε ξεκινήσει από τη δεκαετία του 1980, ενώ τα πρώτα ολοκληρωμένα προγράμματα εμφανίστηκαν τη δεκαετία του 1990 παράλληλα με τις προσπάθειες αλληλούχισης. Οι μεθοδολογίες που έχουν χρησιμοποιηθεί για τα προβλήματα αυτά, καλύπτουν ένα μεγάλο εύρος: από στατιστικές τεχνικές, weight matrices και προφιλ, νευρωνικά δίκτυα, μαρκοβιανές αλυσίδες μέχρι και Hidden Markov Models.
- Οι μεθοδολογίες που βασίζονται καθαρά σε εκπαίδευση για να κάνουν την πρόγνωση αναφέρονται και ως *ab initio gene finders*, ενώ οι μεθοδολογίες στις οποίες χρησιμοποιείται και πληροφορία από τις ήδη υπάρχουσες γνωστές πρωτεΐνες με σκοπό να «καθοδηγηθεί» η πρόγνωση από τα γνωστά παραδείγματα ονομάζονται *homology-based gene finders*.

Προκαρυωτικοί

- Για τους προκαρυωτικούς οργανισμούς, τα πιο γνωστά και πετυχημένα εργαλεία περιλαμβάνουν τα:
  - **FrameD** (http://tata.toulouse.inra.fr/apps/FrameD/FD)
  - **GeneMark** (http://exon.gatech.edu/GeneMark/gmchoice.html)
  - **Glimmer** (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi)
  - **EasyGene** (http://www.cbs.dtu.dk/services/EasyGene/)
  - **FGENESB** (http://linux1.softberry.com/berry.phtml?topic=fgenesb&group=help&subgroup=gfindb)
  - **Prodigal** (http://prodigal.ornl.gov/)

Ευκαρυωτικοί

- Αντίστοιχα, για τους ευκαρυωτικούς οργανισμούς, τα πιο πετυχημένα αντίστοιχα εργαλεία είναι:
  - **FGENESH** (http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind)
  - **GlimmerHMM** (https://ccb.jhu.edu/software/glimmerhmm/)
  - **HMMgene** (http://www.cbs.dtu.dk/services/HMMgene/)
  - **GeneMark.hmm** (http://exon.gatech.edu/GeneMark/hmmchoice.html)
  - **GeneID** (http://genome.crg.es/software/geneid/geneid.html)
  - **GeneScan** (http://genes.mit.edu/GENSCAN.html)
  - **mGene** (http://raetschlab.org/suppl/mgene)
  - **Grail** (http://compbio.ornl.gov/grailexp/)

Άλλα εργαλεία

- Ειδικά εργαλεία για την έναρξη της μεταγραφής (translation initiation) είναι:
  - **ATGpr** (http://atgpr.dbcls.jp/)
  - **NetStart** (http://www.cbs.dtu.dk/services/NetStart/)
  - **TIS Miner** (http://dnafsminer.bic.nus.edu.sg/Tis.html)
  - **StartScan** (http://bioinformatics.psb.ugent.be/webtools/startscan/)
- Για την πολυαδενυλίωση του mRNA τα διαθέσιμα εργαλεία αυτή τη στιγμή είναι:
  - **Poly(A) Signal Miner** (http://dnafsminer.bic.nus.edu.sg/)
  - **PolyAPred** (http://www.imtech.res.in/raghava/polyapred/help.html)
  - **POLYAH** (http://www.softberry.com/berry.phtml?topic=polyah&group=programs&subgroup=promoter)
  - **PolyApredict** (http://cub.comsats.edu.pk/polyapredict.htm)
- Τέλος, μέθοδοι που εστιάζονται στην εύρεση των σημείων αποκοπής και συρραφής εσωνίων/εξωνίων σε ευκαρυωτικά γονιδιώματα, είναι:
  - **Human Splice Finder** (http://www.umd.be/HSF3/)
  - **NetGene** (http://www.cbs.dtu.dk/services/NetGene2/)
  - **NetPlant** (http://www.cbs.dtu.dk/services/NetPGene/)
  - **GeneSplicer** (https://ccb.jhu.edu/software/genesplicer/)
  - **SpliceView** (http://bioinfo4.itb.cnr.it/~webgene/wwwspliceview_ex.html)
  - **SplicePredictor** (http://bioservices.usd.edu/splicepredictor/

Συγκριτική γονιδιωματική

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

- Σε πρώτο επίπεδο, και με βάση τον γενικότερο ορισμό, υπολογιστική γονιδιωματική είναι και κάθε προσπάθεια ανάλυσης του γονιδιώματος ενός και μόνο οργανισμού, δηλαδή
  - οι τεχνικές αλληλούχισης και συναρμολόγησης του γονιδιώματος (Zerbino & Birney, 2008),
  - η εύρεση γονιδίων (Picardi & Pesole, 2010),
  - η εύρεση ρυθμιστικών περιοχών (Harbison et al., 2004),
  - η εύρεση μικρών RNA (Rigoutsos, 2010; Vlachos & Hatzigeorgiou, 2013)
  - ή η εύρεση περιοχών οριζόντιας γονιδιακής μεταφοράς (Soucy, Huang, & Gogarten, 2015) και η εύρεση του τρόπου γονιδιακής ρύθμισης.

- Σε ένα επόμενο επίπεδο οι τεχνικές που χρησιμοποιούνται είναι απλά εφαρμογές γνωστών μεθόδων και αλγορίθμων που είναι σχεδιασμένοι για αλληλουχίες (π.χ. μέθοδοι πρόγνωσης) σε ολόκληρα γονιδιώματα, και στη συνέχεια στατιστική ανάλυση των αποτελεσμάτων με σκοπό την εξαγωγή γενικότερων κανόνων και συμπερασμάτων. Θα παρουσιάσουμε κάποια τέτοια παραδείγματα με σκοπό να εξοικειωθεί ο αναγνώστης με τη μεθοδολογία.

- Στο επόμενο στάδιο όμως, θα παρουσιαστούν οι πιο ενδιαφέρουσες τεχνικές της συγκριτικής γονιδιωματικής, οι οποίες προσφέρουν κάτι επιπλέον: αξιοποιώντας την πληροφορία για την ύπαρξη, τη θέση και την εσωτερική δομή των γονιδίων στα γονιδιώματα διαφόρων υπό σύγκριση οργανισμών, μπορούν να μας δώσουν επιπλέον πληροφορίες, πληροφορίες που από μια απλή ανάλυση ενός οργανισμού (και του γονιδιώματός του) δεν θα μπορούσαν να εξαχθούν

	X_1	X_2	X_3	X_4	…	X_k
Γονιδίωμα Α						
Γονιδίωμα Β						
Γονιδίωμα Γ						
Γονιδίωμα Δ						
…						

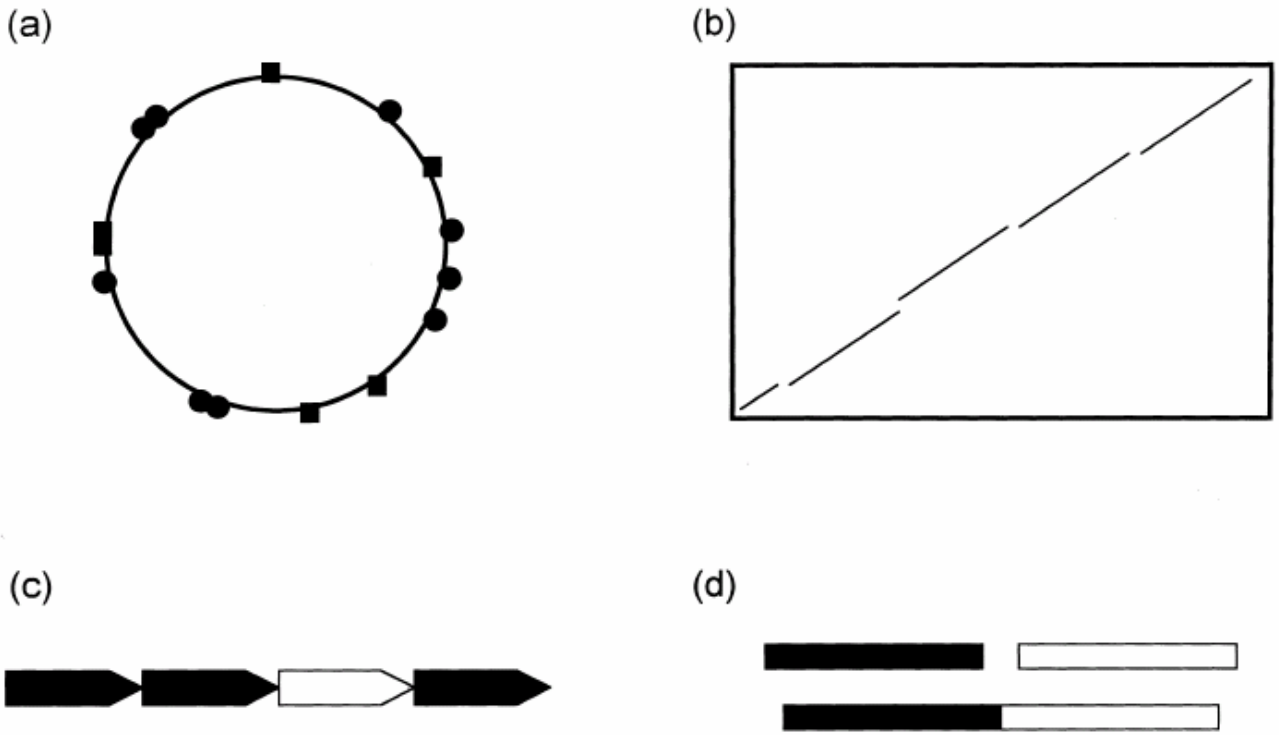


Fig. 1. Pictorial representation of four computational genomics methods. (a) Genome subtraction aims to define species-specific genes. This is achieved by subtracting genes homologous to various elements such as genes orthologous to the species under consideration or phages which are likely to be inserted 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)

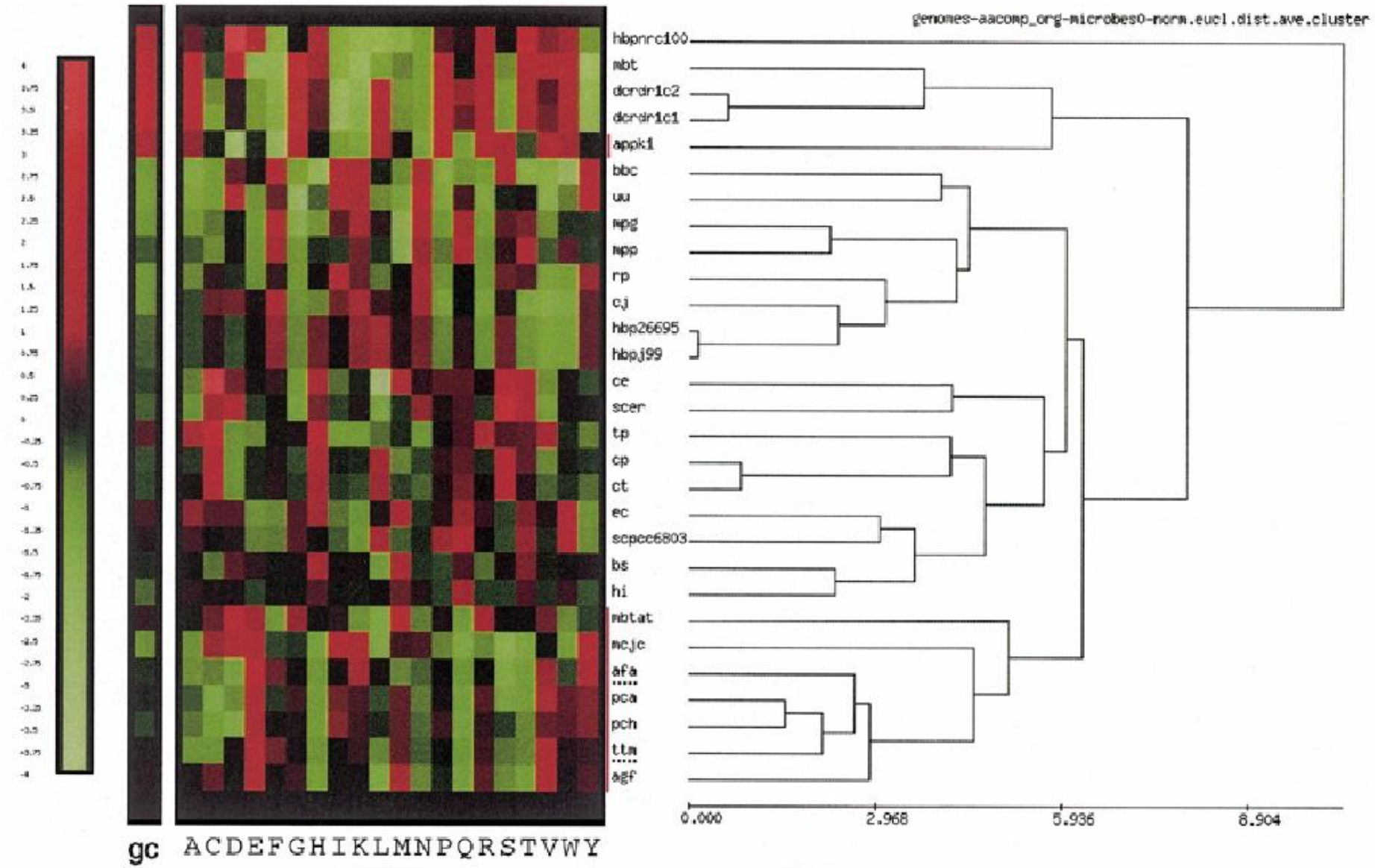


Figure 1. Standardised amino acid composition data of completely sequenced organisms grouped by hierarchical clustering. The GC ratios are shown for reference but were not used for the clustering process. Amino acids are abbreviated by the standard one letter code. The labels indicating the data sets 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.

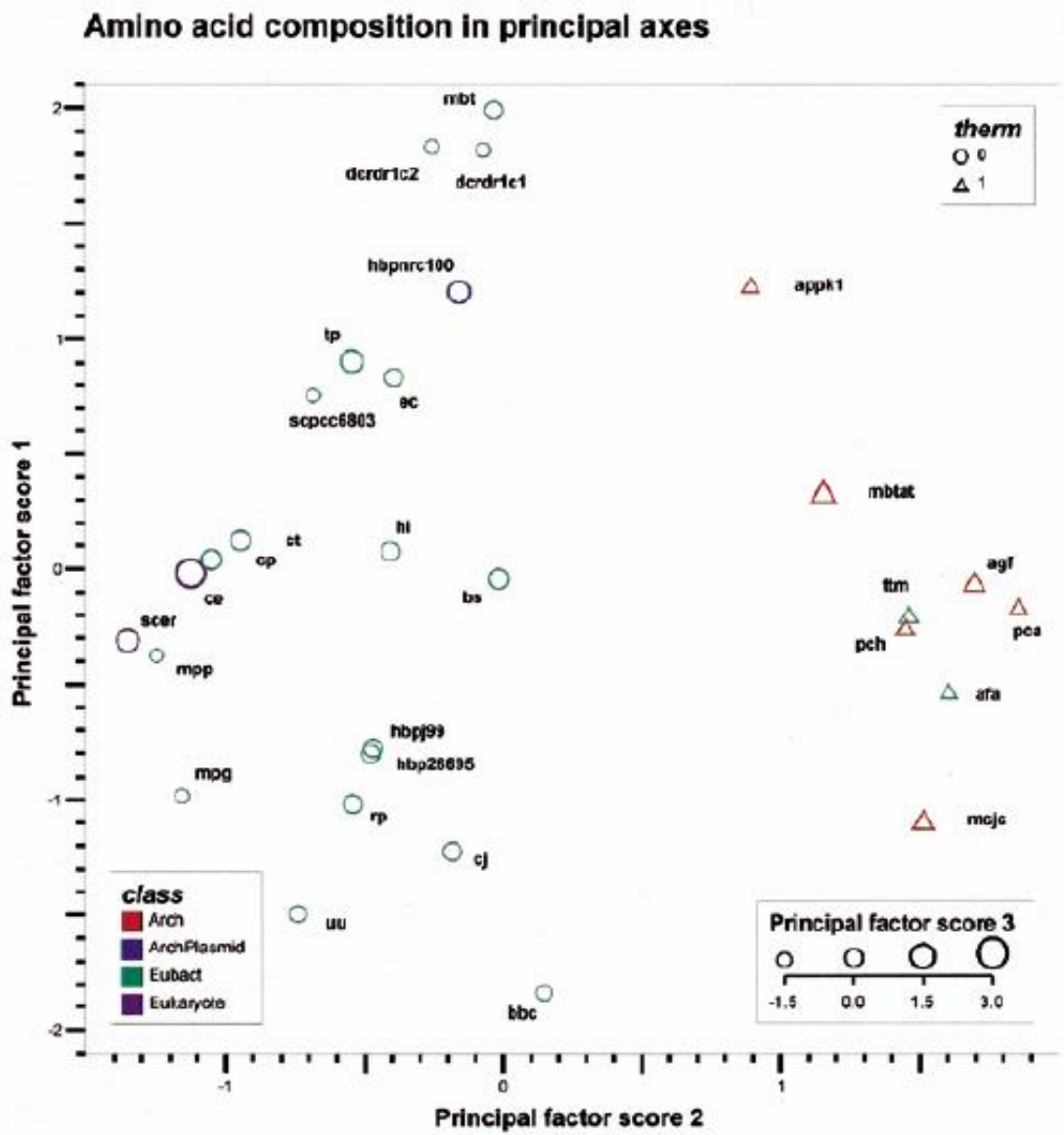


Figure 2. Reduced dimensionality plot showing the main principal components of the global amino acid compositions. The first principal axis (vertical) corresponds to GC ratio (see text). The second principal axis (horizontal) shows a clear separation of thermophiles and mesophiles, denoted by triangles 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%	$\sim 10^{-8}$	–2.18 (0.31)	–1.76 (0.25)	$\sim 10^{-4}$	*t*-test
Glu (E)	80%	80%	$\sim 10^{-6}$	2.27 (0.40)	1.73 (0.31)	$\sim 10^{-4}$	*t*-test
Val (V)	50 to 65%	60%	$\sim 10^{-3}$	1.57 (0.42)	1.40 (0.38)	$\sim 2 \times 10^{-3}$	*t*-test
Thr (T)	–65%	60%	$\sim 10^{-3}$	–0.84 (0.25)	–1.31 (0.39)	$\sim 5 \times 10^{-3}$	*t*-test[b]
His (H)	–40 to –60%	–60%	$\sim 10^{-3}$	–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

- Σε μια από τις πρώτες, αρκετά απλές αλλά ιδιαίτερα πληροφοριακές τέτοιες μελέτες, ο Ouzounis και ο Kreil, ανέλυσαν την αμινοξική σύσταση των πρωτεϊνών που κωδικοποιούν τα γονιδιώματα 6 θερμόφιλων αρχαιοβακτηρίων (αρχαίων), 2 θερμόφιλων βακτηρίων, 17 μεσόφιλων βακτηρίων και 2 ευκαρυωτικών οργανισμών.
- Στην ανάλυση χρησιμοποίησαν την αμινοξική σύσταση και το ποσοστό GC και πραγματοποίησαν ιεραρχική ομαδοποίηση και ανάλυση κύριων συνιστωσών (principal components analysis).
- Παρόλο που το ποσοστό GC είχε μια ξεκάθαρη επιρροή, τα θερμόφιλα είδη μπορούν να αναγνωριστούν με μόνη χρήση της ολικής αμινοξικής σύστασης (Kreil & Ouzounis, 2001).
- Αναλύοντας τα αποτελέσματα, φάνηκε ότι τα θερμόφιλα είδη έχουν λιγότερη Γλουταμίνη (Gln) και περισσότερο Γλουταμικό (Glu) σε σχέση με τα μεσόφιλα. Τα θερμόφιλα, έχουν επίσης περισσότερη Βαλίνη (Val) και λιγότερη Θρεονίνη (Thr) σε σχέση με τα μεσόφιλα. Για τα αμινοξέα Ιστιδίνη (His), Σερίνη (Ser) και Ασπαραγίνη (Asn) υπήρχαν επίσης ενδείξεις αλλά με μικρότερο στατιστικό βάρος.

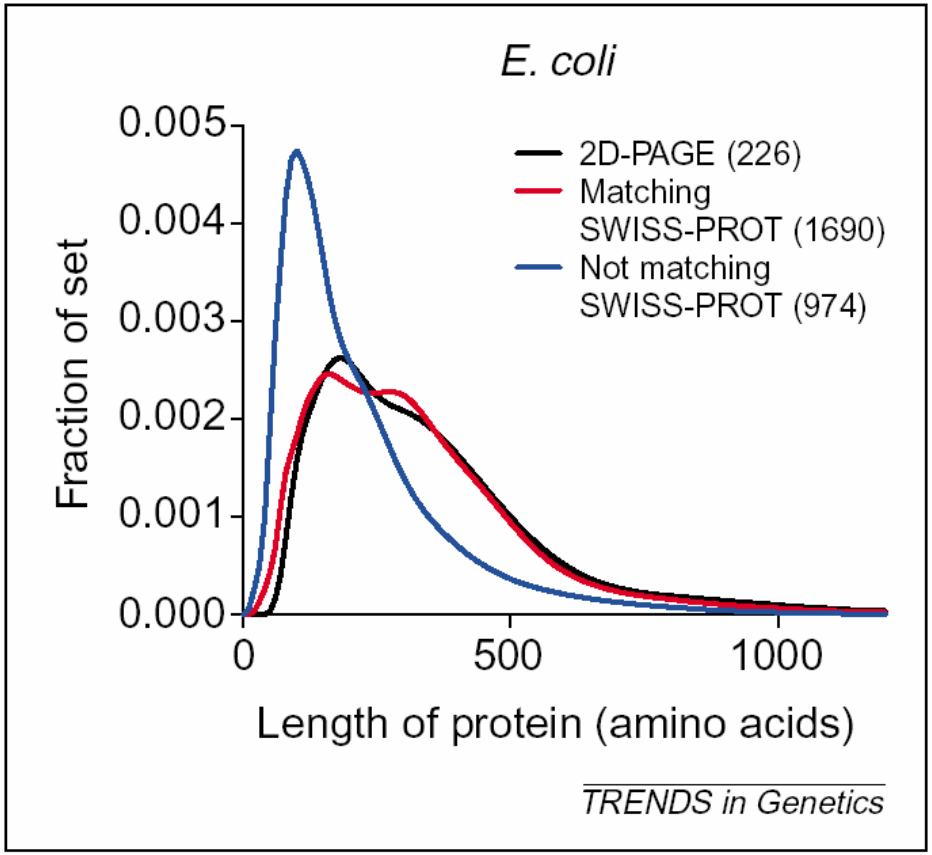
Το μήκος των «πραγματικών πρωτεϊνών» στα πλήρως προσδιορισμένα γονιδιώματα

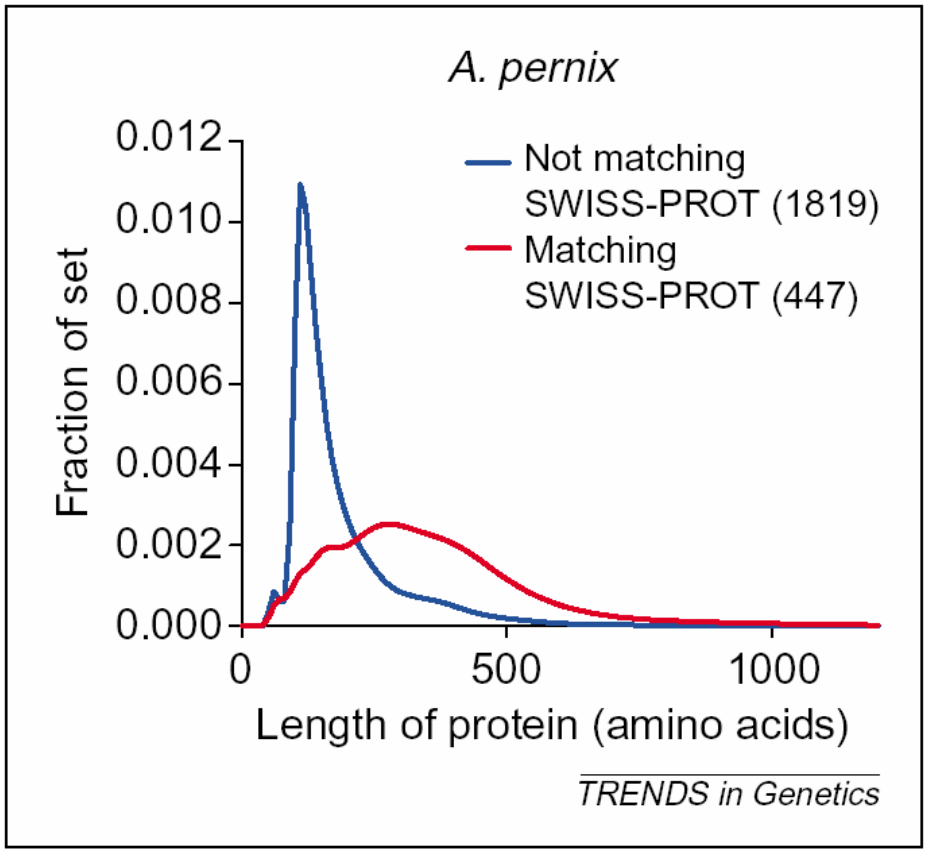
- 64 τριπλέτες (61 για αμινοξέα-3 για λήξη)
- Αν τα νουκλεοτίδια θεωρηθούν ισοπίθανα, έχουμε μια τριπλέτα λήξης περίπου κάθε 21 αμινοξέα
- Οι τριπλέτες λήξης είναι πλούσιες σε AT (TAA, TGA, TAG)
- Κατά συνέπεια το μήκος των «τυχαίων» ORF θα αυξάνει στα πλούσια σε GC γωνιδιώματα

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

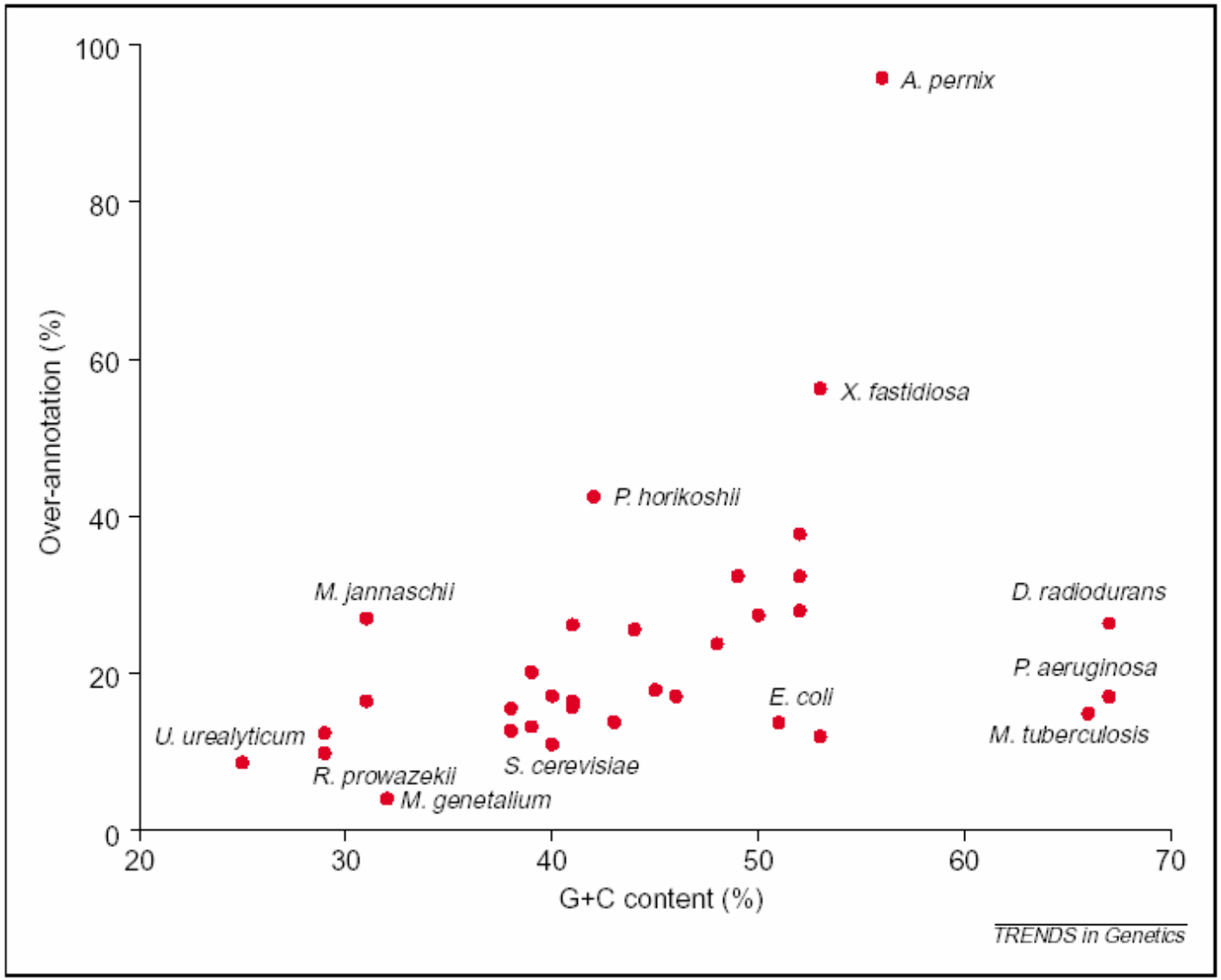
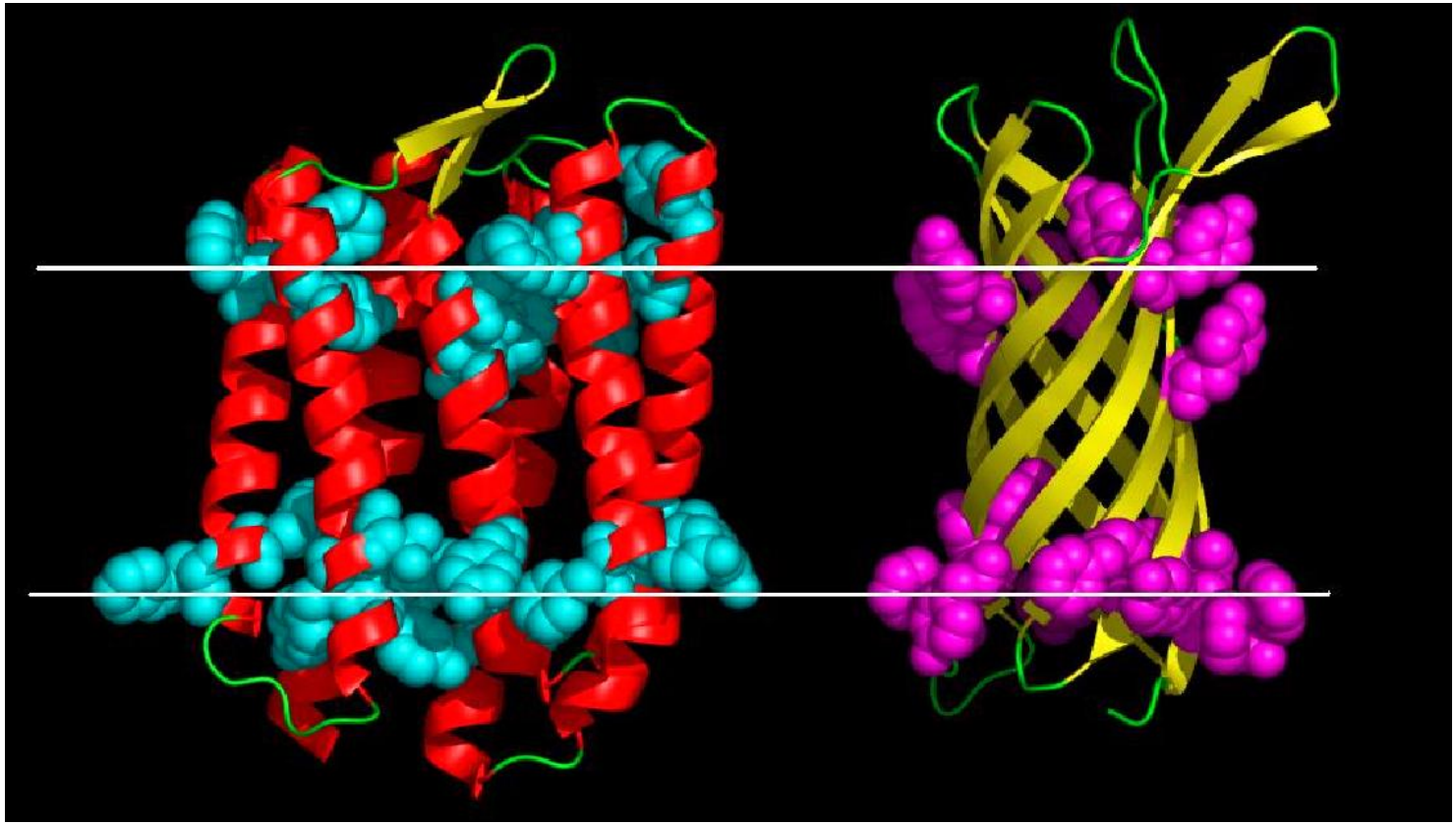


Fig. 1. Estimated over-annotation of genes in sequenced genomes. For each organism the SWISS-PROT-based estimate is calculated and the difference to the number of annotated genes shown in percent of the estimated number of genes.

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

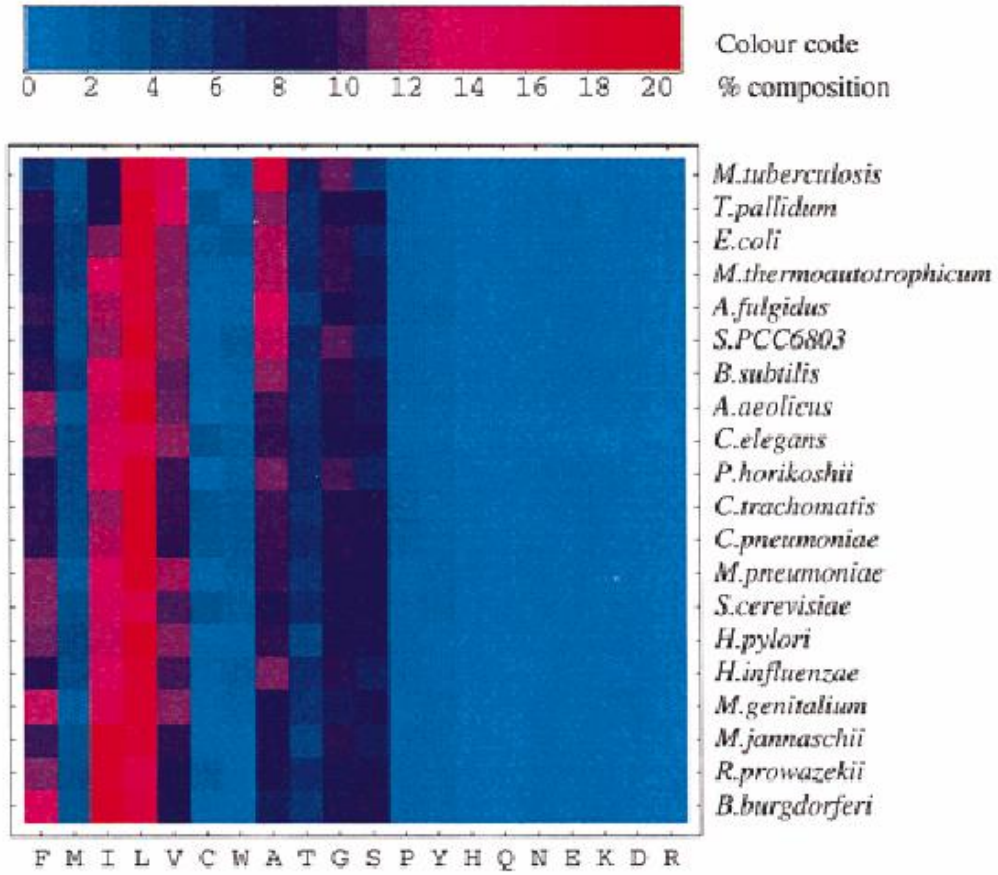


Fig. 1. A plot to show the amino acid composition of the TM domains in each of the proteomes under investigation. Percentage of TM composition exhibited by a residue in each organism increases from blue to red.

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

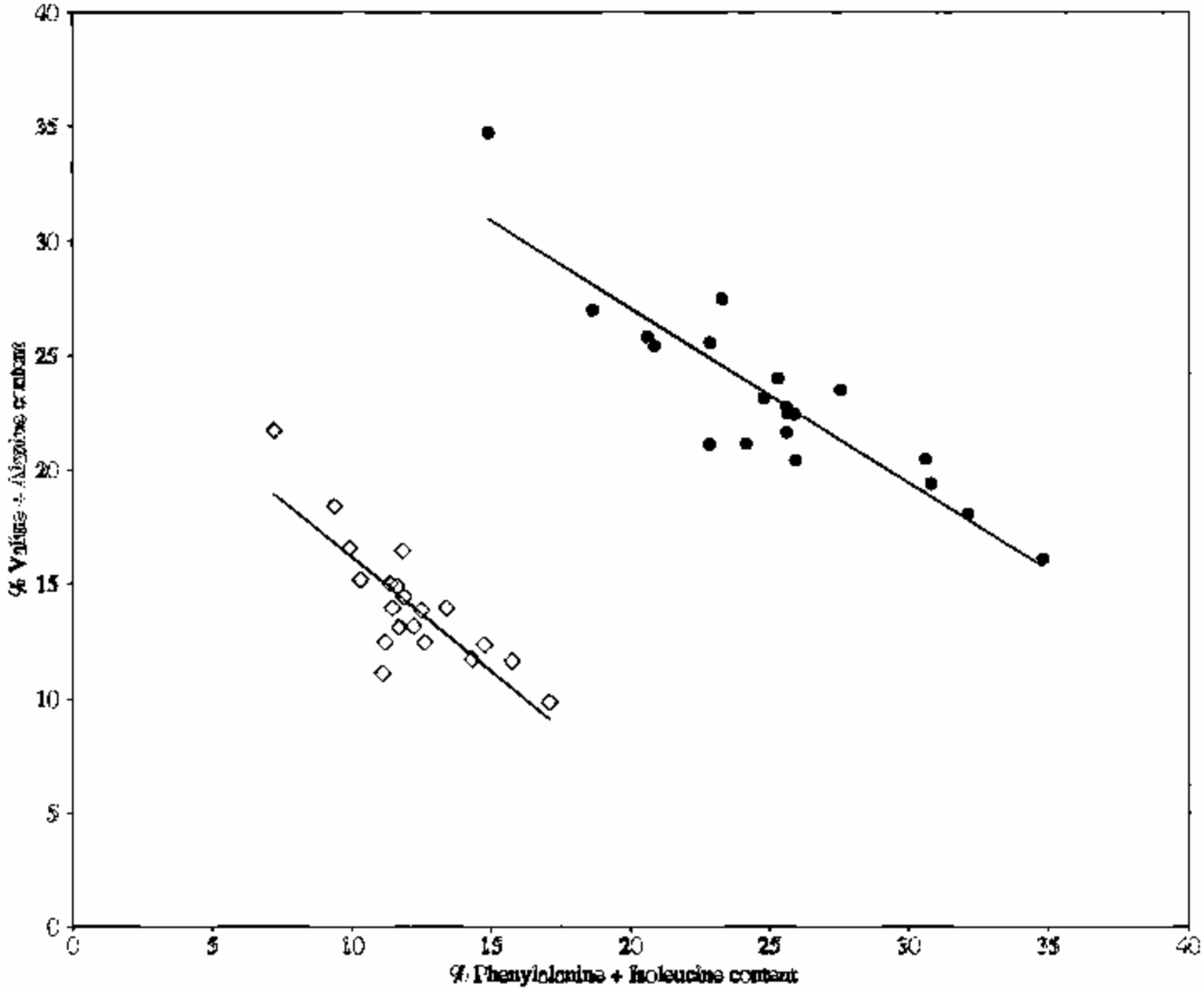


Fig. 2. A graph to show the correlation between the percentage abundance of valine plus alanine residues (VA) and the percentage abundance of phenylalanine plus isoleucine residues (FI). Data are shown for each organism, for both the amino acids within the predicted TM domains (●) and the whole proteome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

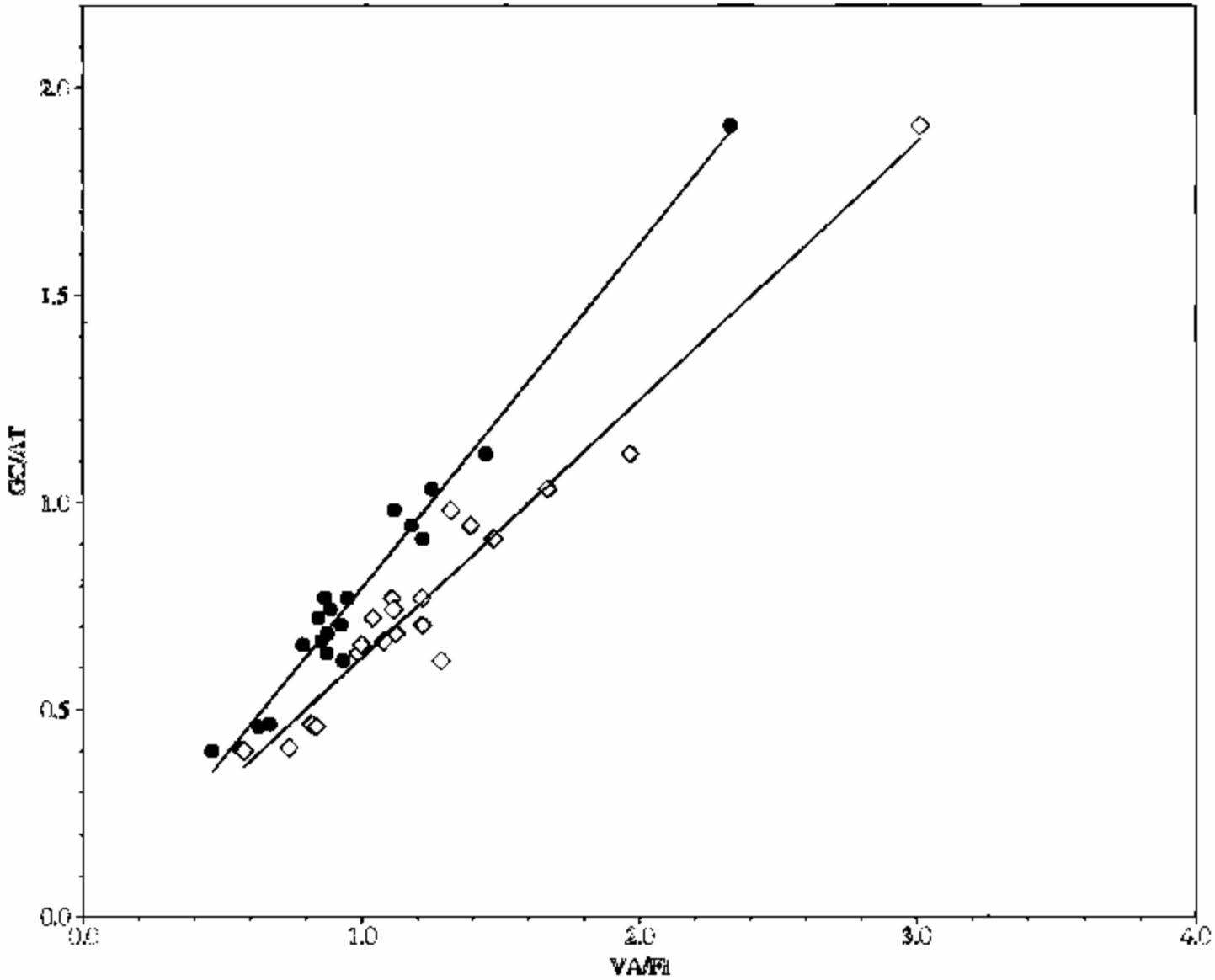
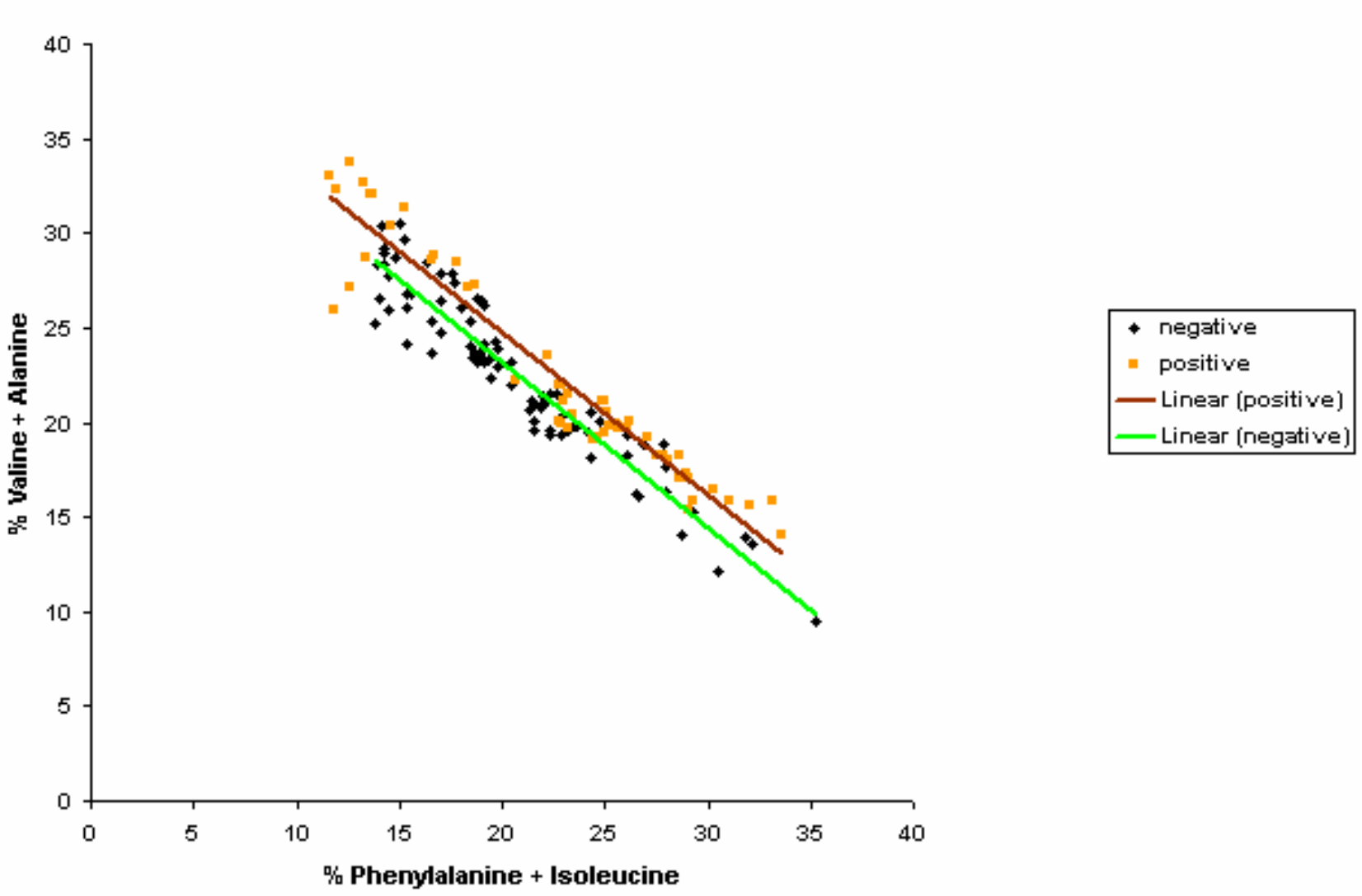
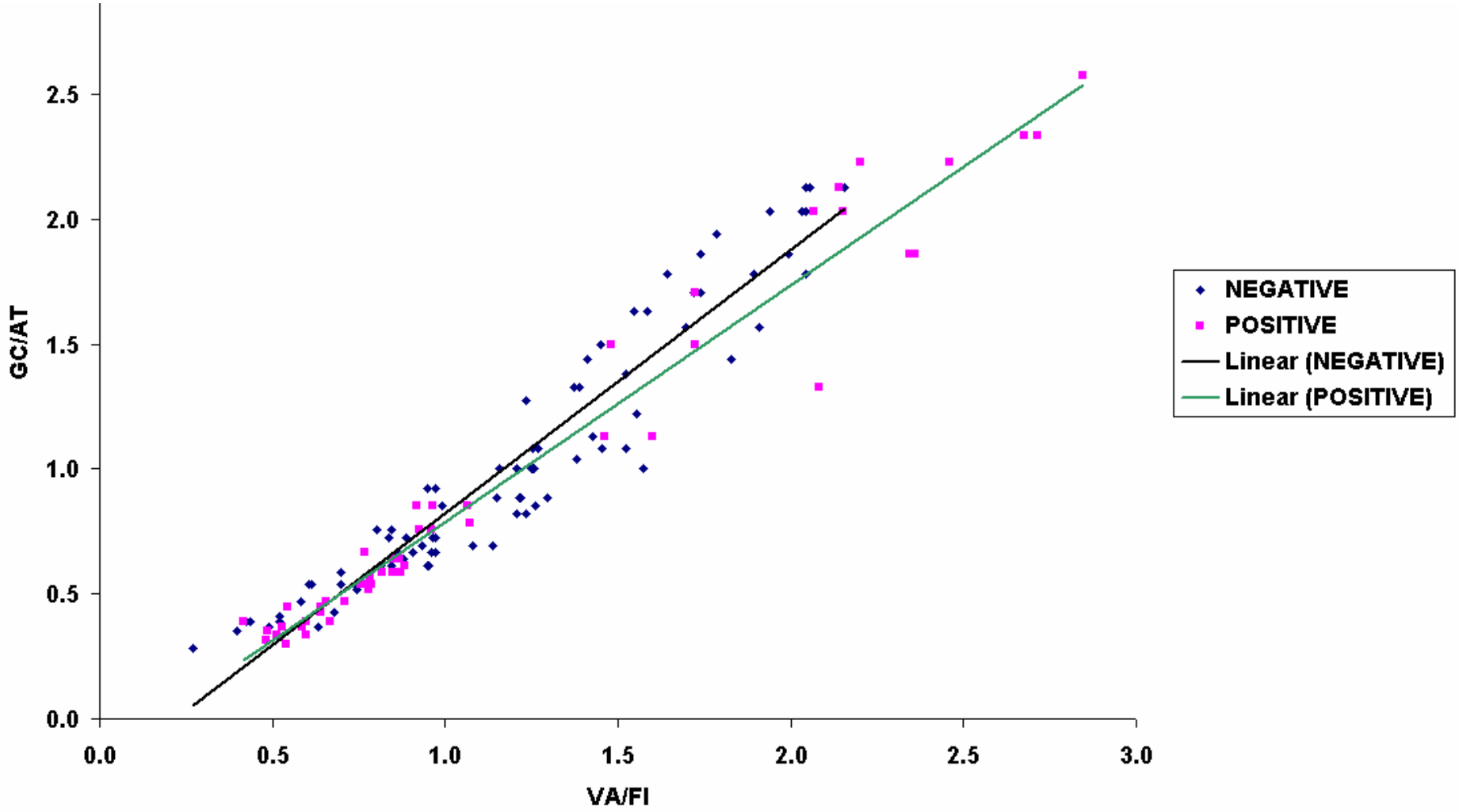


Fig. 3. A graph to show the correlation between the nucleotide bias of a genome (GC/AT) and alternate hydrophobic amino acid use (VA/FI). Data are shown for both the amino acids within the predicted TM domains (●) and for the whole proteome (◇).


40
35
30
25
20
15
10
5
0
% Valine + Alanine
0
5
10
15
20
25
30
35
40
% Phenylalanine + Isoleucine
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

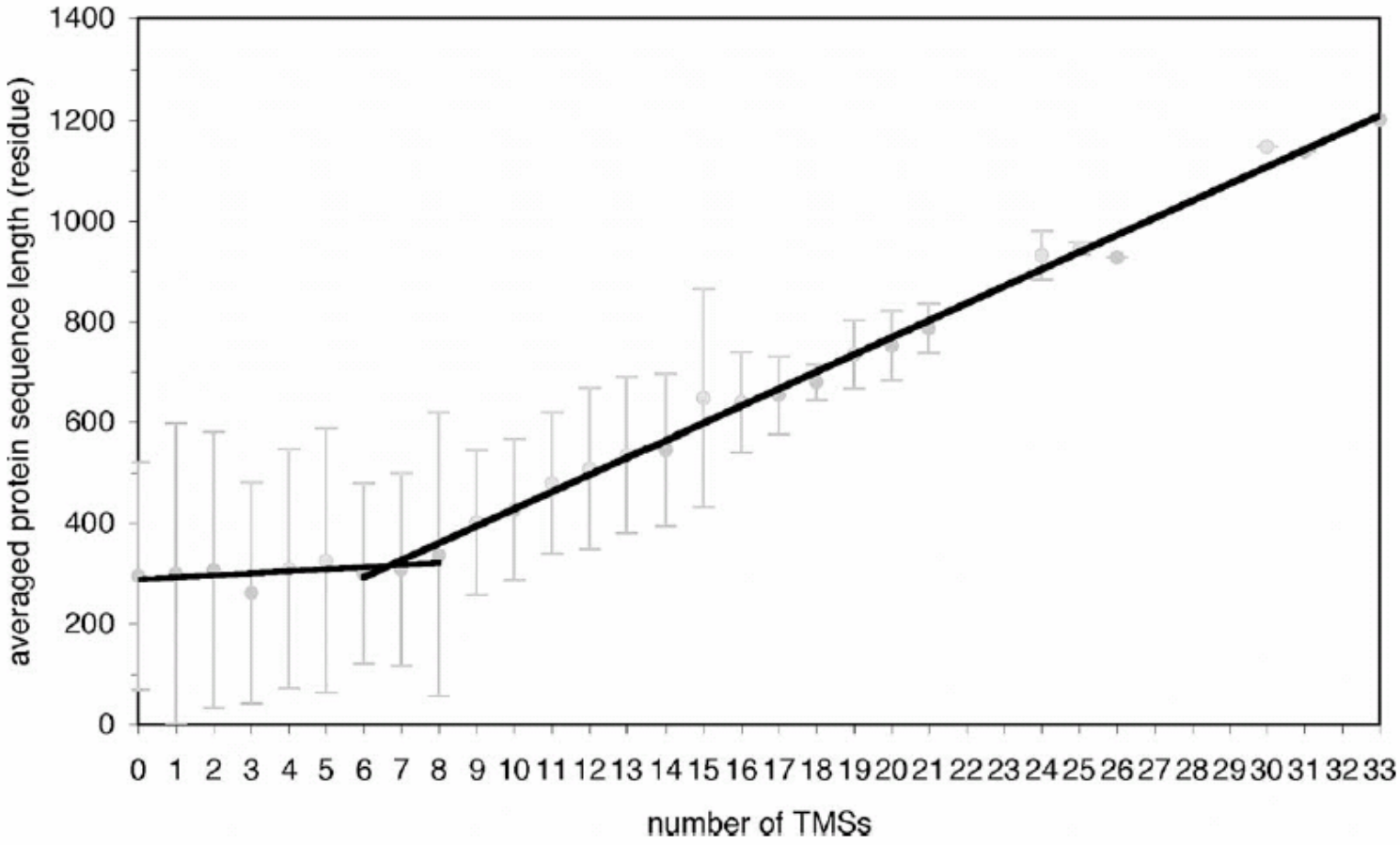
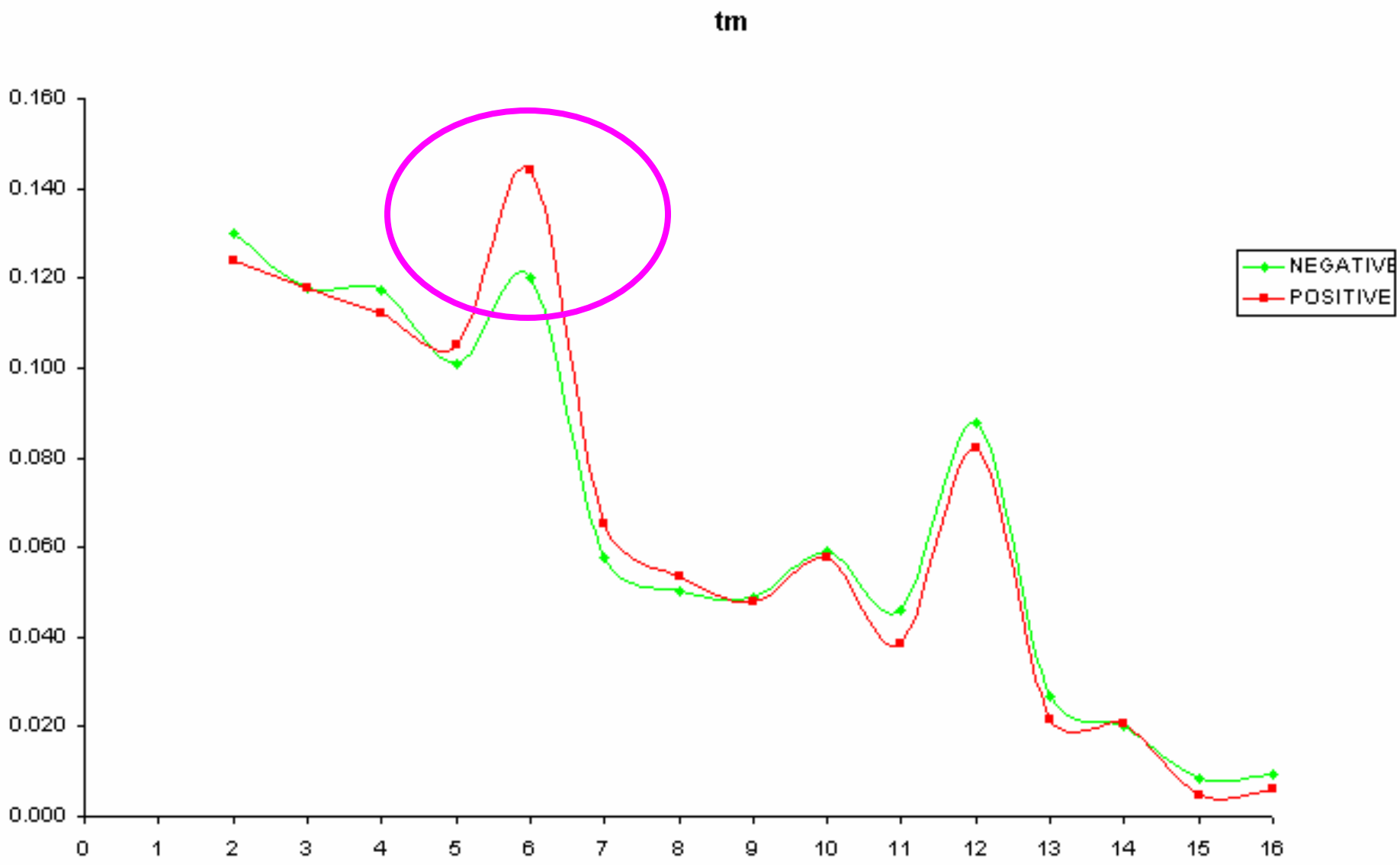
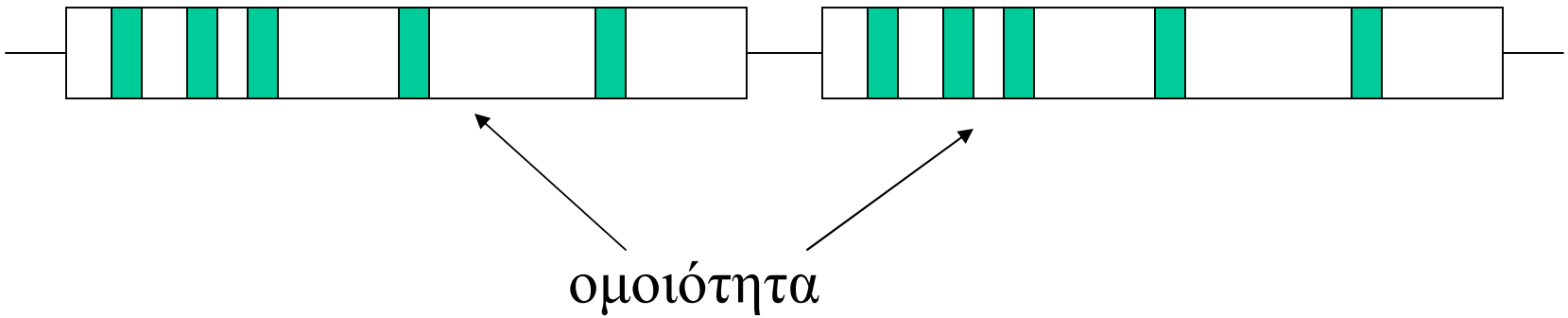


Fig. 4. Relationship between the number of TMSs and the protein sequence length averaged over the 50 genomes. The slope of the line between 7- and 33-tms is 35 residues. The number of TMSs, 0 in the abscissa means soluble proteins.

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

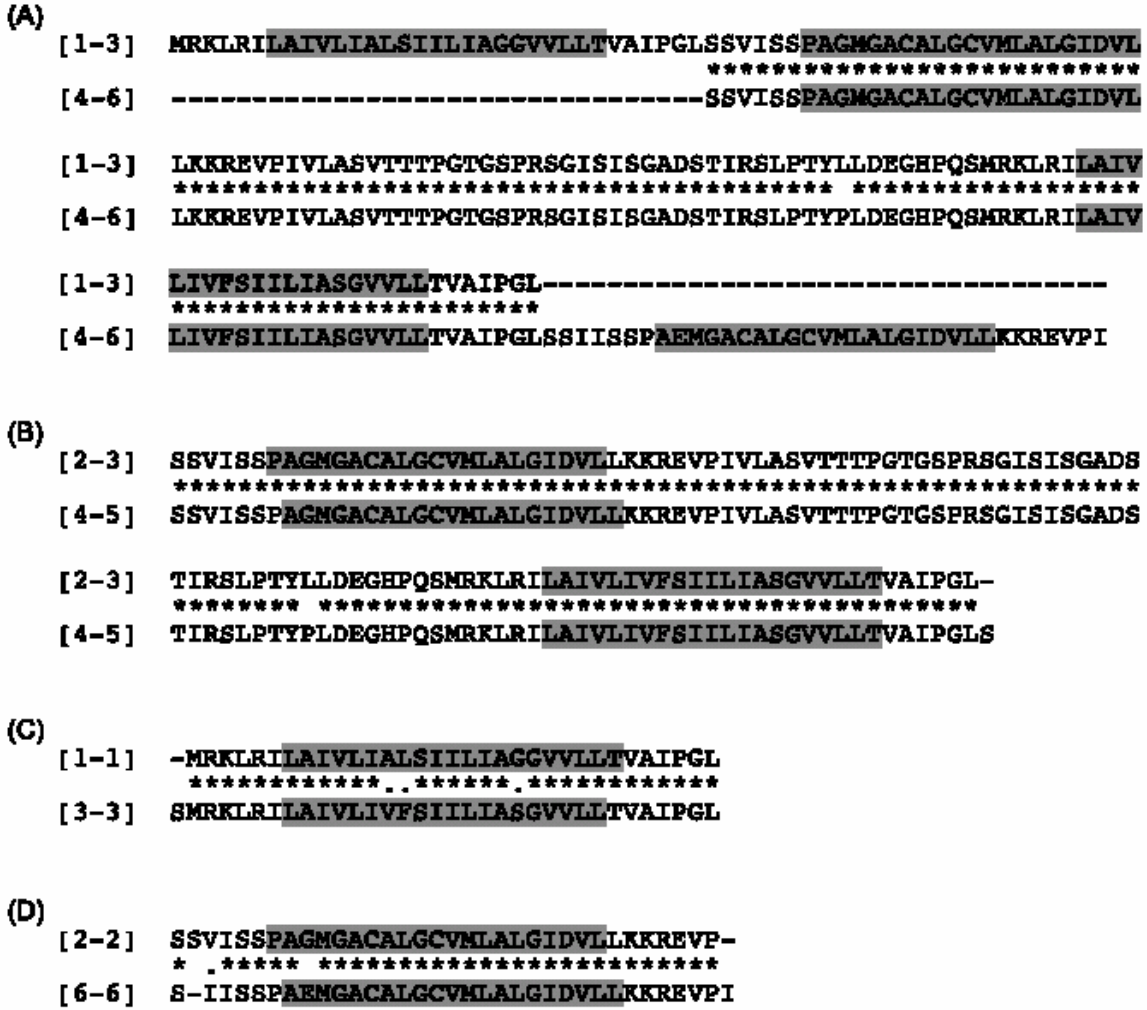


Figure 3. Pairwise alignments between partial sequences of a 6-tms TM protein, CPn0007 (Golgi autoantigen, golgin subfamily A4) from *Chlamydophila pneumoniae* by using the ALIGN program with the setting parameters, i.e. opening gap penalty −12, extension gap penalty −2, and substitution matrix BLOSUM62): (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

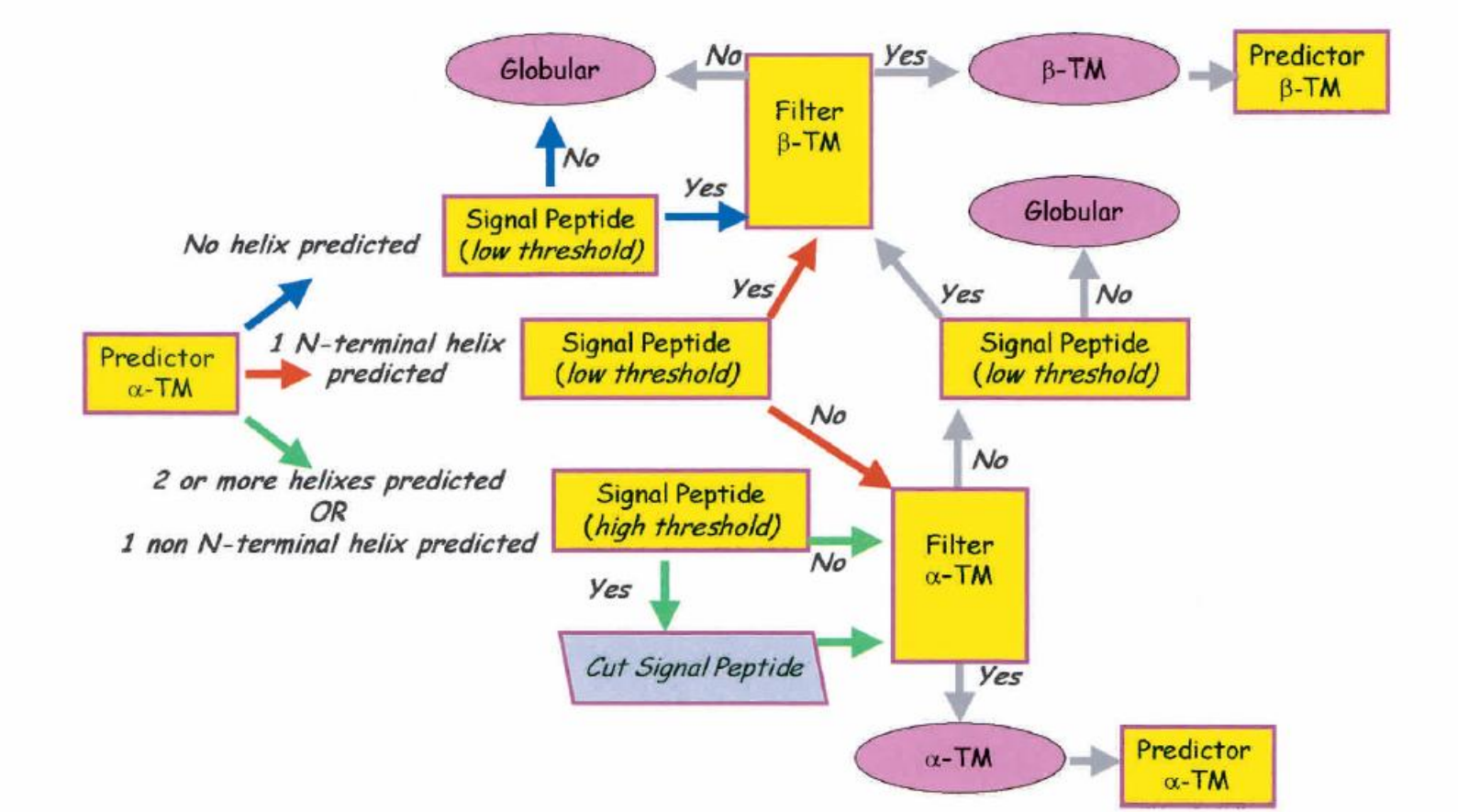


Figure 1. Hunter: The suite of predictors. The flow chart indicates the possible alternatives after the first prediction done with a neural network-based method. Chain flow limiting steps are: a signal peptide predictor (acting with two different threshold values), trained and tested on signal peptides 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.

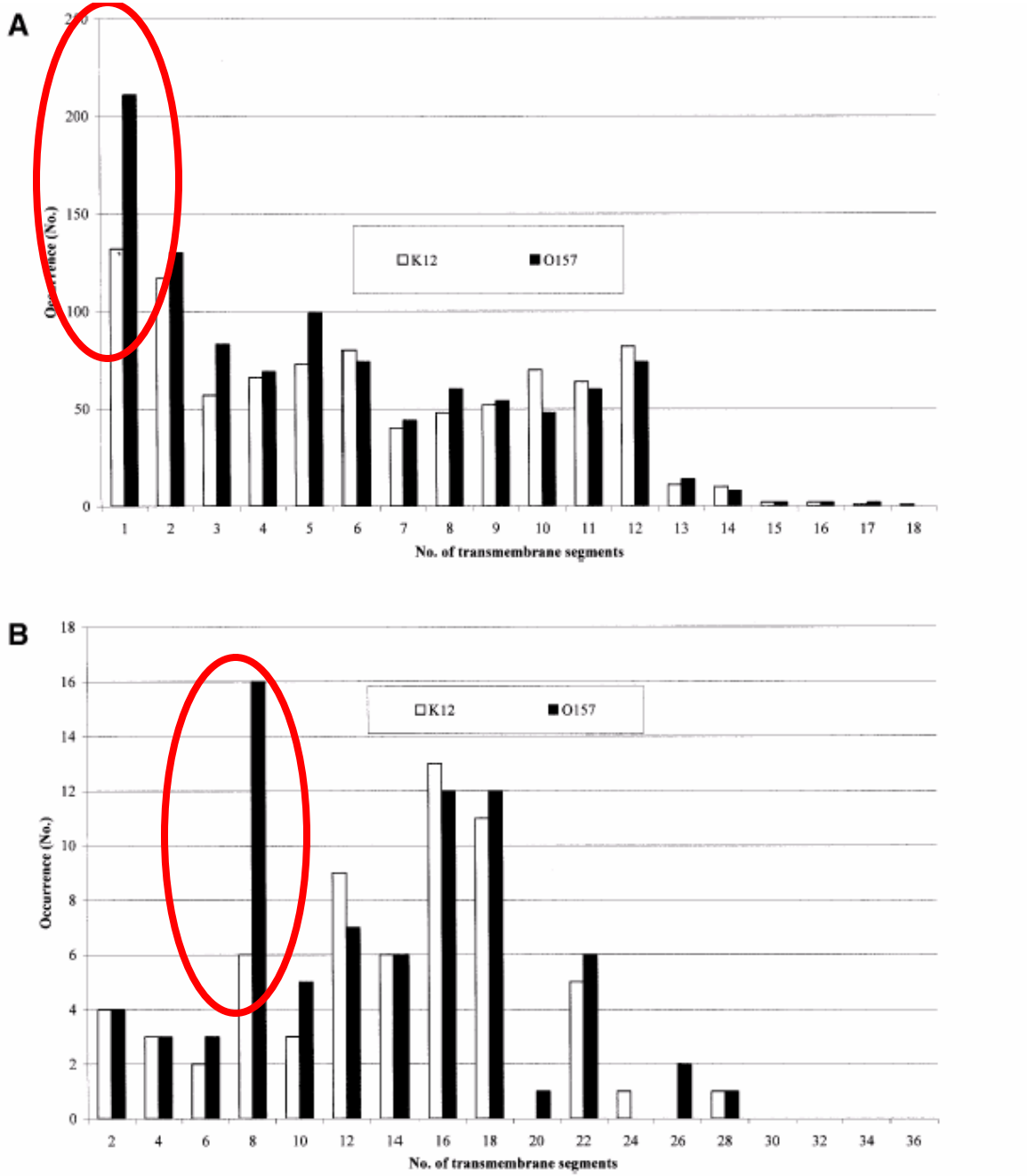
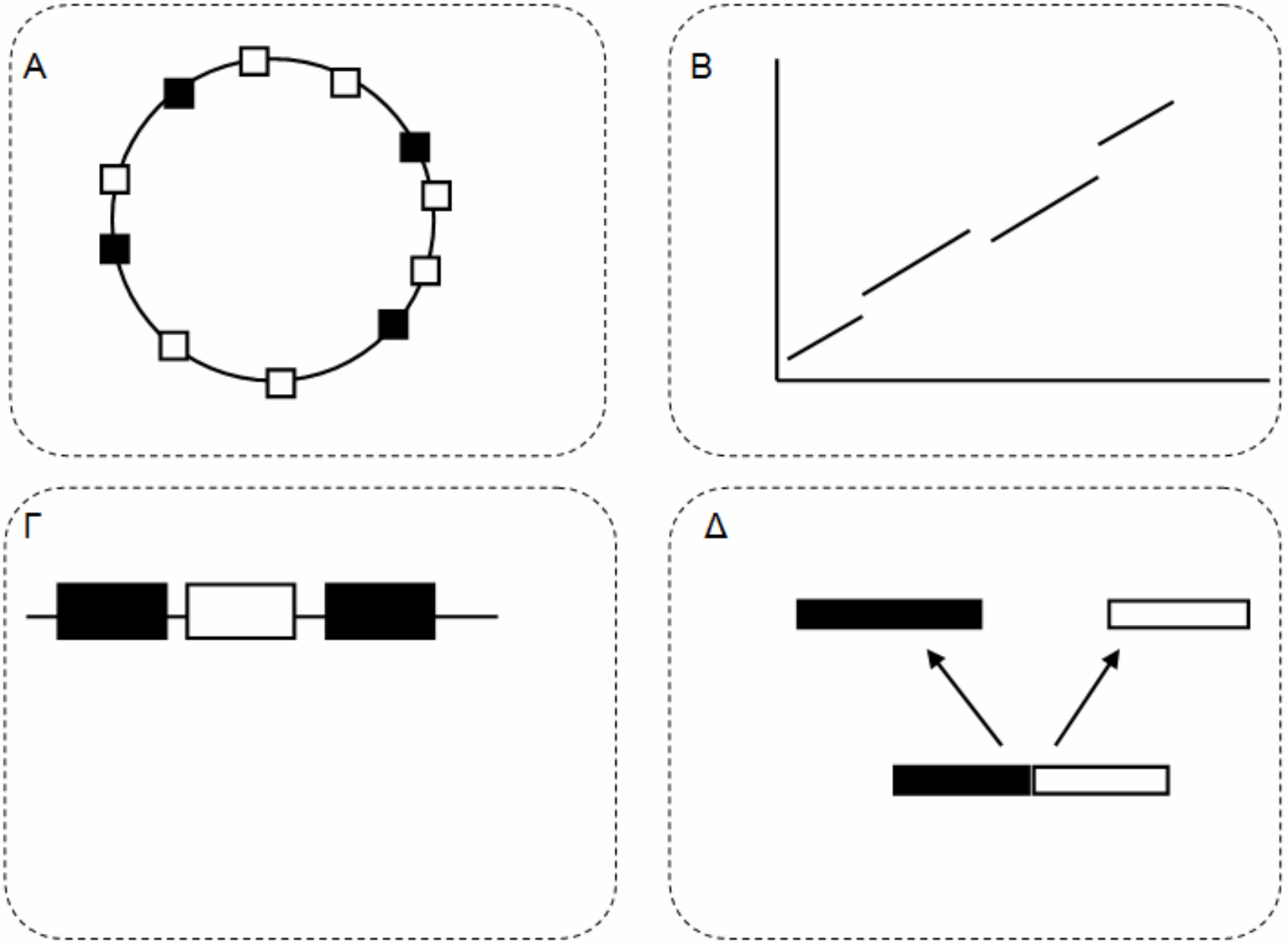


Figure 3. Topography of transmembrane proteins in *E. coli* K12 and O157 as predicted with Hunter. (*A*) Bar plot of inner membrane proteins as a function of the number of transmembrane predicted segments in both strains. (*B*) Bar plot of outer membrane proteins as a function of transmembrane predicted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

συγκριτική γονιδιωματική

- Η συγκριτική γονιδιωματική, πάει ένα βήμα παραπέρα την υπολογιστική ανάλυση των γονιδιωμάτων. Αντί να εστιάζει μόνο στα συνολικά στατιστικά μέτρα από κάθε γονιδίωμα, όπως π.χ. το ποσοστό GC ή κάποιο άλλο μέτρο, επιχειρεί να χρησιμοποιήσει τη βασική αρχή της φυλογενετικής ανάλυσης, ότι δηλαδή τα γονιδιώματα όλων των οργανισμών προέρχονται από προγονικές μορφές και έχουν διαμορφωθεί έτσι όπως είναι σήμερα μετά από αλλεπάλληλες αλλαγές που έγιναν μέσα σε εκατομμύρια χρόνια. Οι αλλαγές αυτές, αφορούν τόσο τα αντίστοιχα ορθόλογα γονίδια και τις αλληλουχίες τους, όσο και το ίδιο το γονιδίωμα, τη δομή του, και τη διάταξή των γονιδίων πάνω σε αυτό.
- Βασικά, η συγκριτική γονιδιωματική κάνει χρήση των κλασικών αλγορίθμων στοίχισης και εύρεσης ομοιότητας μεταξύ γονιδίων ή/και πρωτεϊνών, αλλά συνδυάζοντας αυτή την πληροφορία με τη δομή του γονιδιώματος και τη διάταξη των γονιδίων πάνω σε αυτό, καταφέρνει να εξάγει πολύ σημαντικά συμπεράσματα, που δεν θα μπορούσαν να έχουν εξαχθεί με άλλον τρόπο (ούτε καν με πρόγνωση)

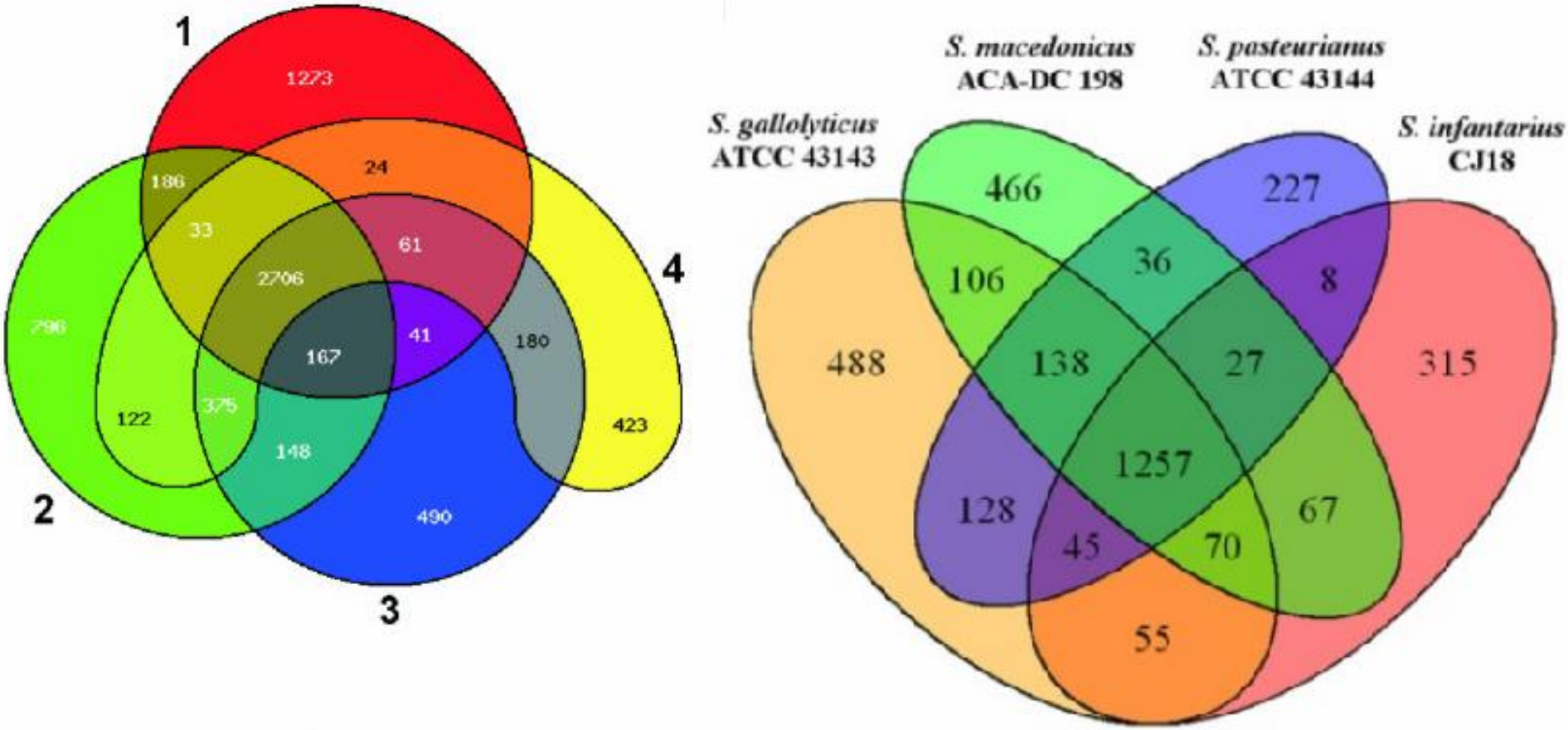

A
B
Γ
Δ

Μέθοδοι

- **Η μέθοδος «αφαίρεσης» γονιδίων**, στην οποία συγκρίνονται σε μια σειρά οργανισμούς τα κοινά γονίδια και εντοπίζονται τα μοναδικά γονίδια.
- **Η μέθοδος σύγκρισης της σειράς των γονιδίων**, σύμφωνα με την οποία εντοπίζονται γονίδια που έχουν την τάση να βρίσκονται κοντά σε όλα τα υπό μελέτη γονιδιώματα,
- **Η μέθοδος στοίχισης ολόκληρων γονιδιωμάτων**, σύμφωνα με την οποία στοιχίζονται ολόκληρα γονιδιώματα και εντοπίζονται οι περιοχές στις οποίες έχουν μεγάλη ομοιότητα, και τέλος
- **Η μέθοδος εντοπισμού προϊόντων γονιδιακής σύντηξης**, στην οποία εντοπίζονται με υπολογιστικό τρόπο γονίδια τα οποία σε κάποιον άλλον οργανισμό βρίσκονται ενωμένα (σύντηξη), λειτουργούν δηλαδή σαν ανεξάρτητες πρωτεϊνικές περιοχές (domains).
- Όλες οι παραπάνω μεθοδολογίες λειτουργούν με χρήση της ομοιότητας των γονιδίων και των πρωτεϊνικών προϊόντων τους και κάνουν χρήση της πληροφορίας από τη σχετική θέση των γονιδίων (ή και την ίδια την ύπαρξή τους) σε διαφορετικούς οργανισμούς. Παρόλα αυτά, οι μεθοδολογίες αυτές εντοπίζουν διαφορετικού είδους λειτουργικές συσχετίσεις μεταξύ των γονιδίων. Προσφέρουν δηλαδή διαφορετικά αποτελέσματα, γι' αυτό και στη μεγάλη τους πλειοψηφία δρουν συμπληρωματικά, όπως θα δούμε παρακάτω

Η μέθοδος «αφαίρεσης» γονιδίων

- Η μέθοδος αυτή, βασίζεται στην εύρεση κοινών, ομόλογων δηλαδή, γονιδίων σε μια σειρά υπό σύγκριση οργανισμών.
- Η βασική αρχή, είναι η γνωστή από παλιά αρχή στη φυλογενετική, ότι οι πιο συγγενικοί οργανισμοί θα έχουν και περισσότερα κοινά χαρακτηριστικά (δηλαδή, γονίδια στην περίπτωση μας). Με την ενσωμάτωση της γνώσης για τη μοριακή λειτουργία αυτών των γονιδίων, μπορούμε να εντοπίσουμε ποια γονίδια είναι χαρακτηριστικά για μια ομάδα οργανισμών και να εξάγουμε χρήσιμα συμπεράσματα για τη φυλογένεση (λειτουργούν δηλαδή ως απομορφικοί χαρακτήρες).
- Εξετάζοντας τα γονίδια που είναι μοναδικά σε κάποιον οργανισμό (ή σε κάποιους οργανισμούς) μπορούμε επίσης να εντοπίσουμε ειδικές λειτουργίες που επιτελεί αυτός ο οργανισμός για να επιβιώσει (π.χ. τα μεθανότροφα βακτήρια έχουν ειδικά μεταβολικά μονοπάτια για να αποικοδομούν το μεθάνιο που βρίσκεται σε περίσσεια στο περιβάλλον τους).



Παραδείγματα διαγραμμάτων Venn. Αριστερά: Σύγκριση στελεχών του Xanthomonas Oryzae με το EDGAR (Blom et al., 2009) Δεξιά: Σύγκριση διαφορετικών ειδών Streptococcous με το R (Papadimitriou et al., 2014)

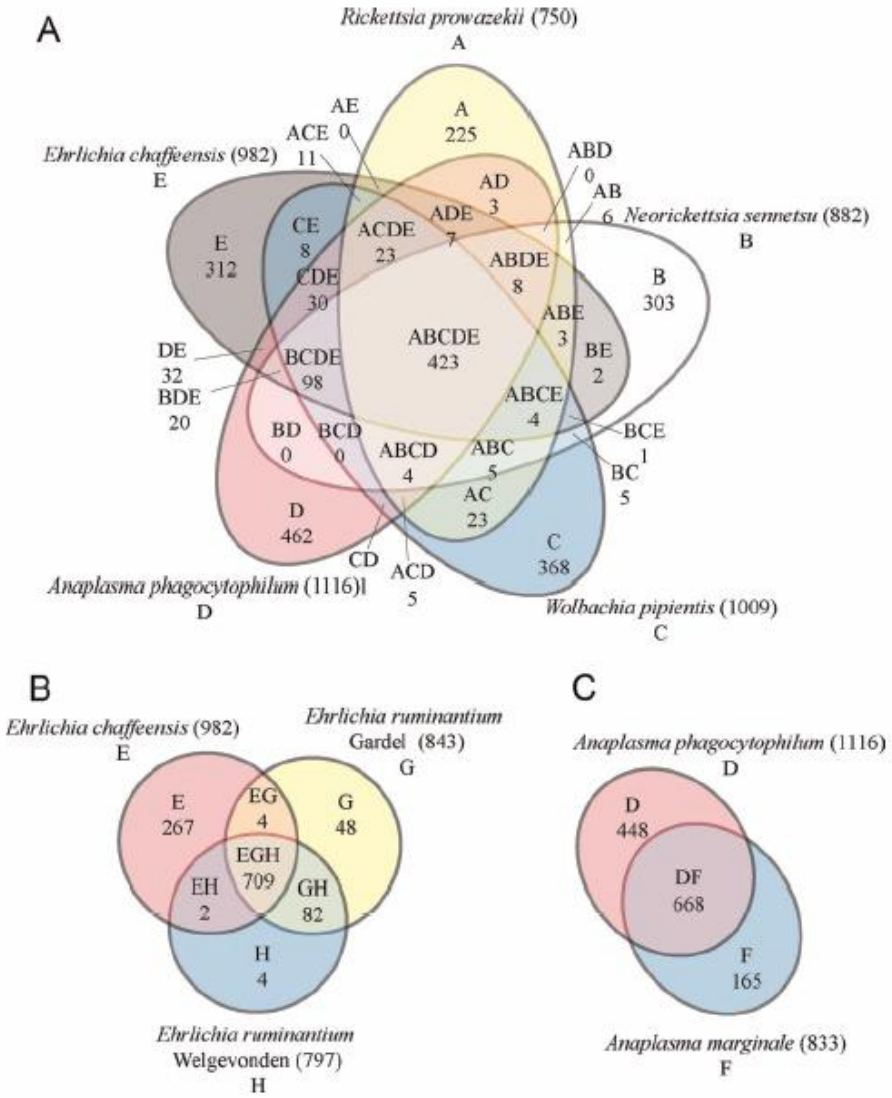


Figure 4. Comparison of the Rickettsiales Gene Sets

The composition of ortholog clusters (see Materials and Methods) of representative Rickettsiales (A), *Ehrlichia* spp. (B), and *Anaplasma* spp. (C) were compared. Numbers within the intersections of different ovals indicate ortholog clusters shared by 2, 3, 4, or 5 organisms. Species compared are indicated in diagram intersections as follows. A, *R. prowazekii;* B, *N. sennetsu;* C, *W. pipientis;* D, *A. phagocytophilum;* E, *E. chaffeensis;* F, *A. marginale;* G, *E. ruminantium* Gardel; and H, *E. ruminantium* Welgevonden.

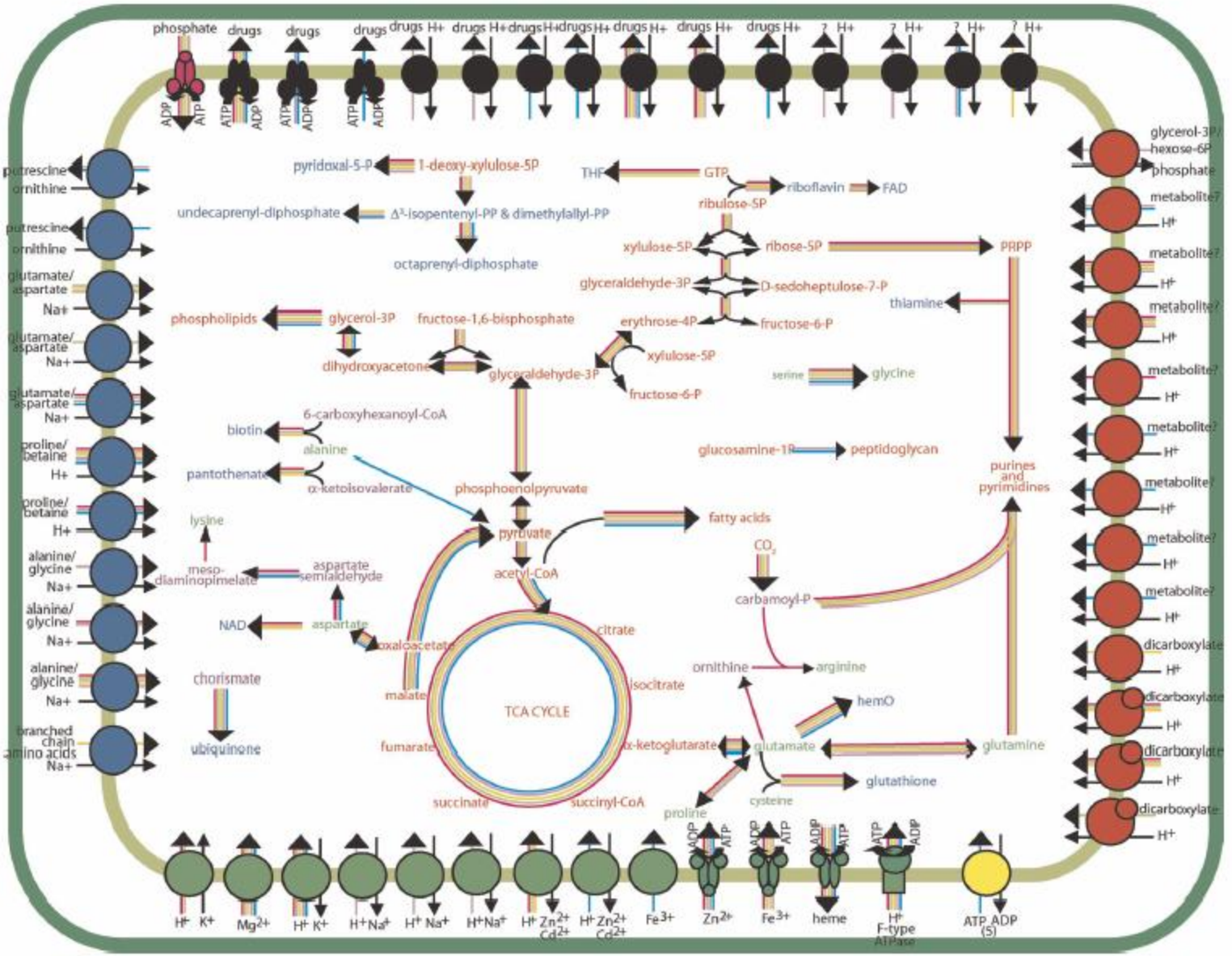


Figure 6. Comparative Metabolic Potential of Select Rickettsiales

Metabolic pathways of *E. chaffeensis* (magenta arrows), *A. phagocytophilum* (green arrows), *N. sennetsu* (gold arrows), *W. pipientis* (lavender arrows), and *R. prowazekii* (cyan arrows) were reconstructed and compared. The networks of some of the more important pathways are shown with metabolites color coded: red and purple, central and intermediary metabolites; blue, cofactors; green, amino acids; and black, cell structures. Transporters are shown in the membrane and are grouped by predicted substrate specificity: green, inorganic cations; magenta, inorganic anions; red, carbohydrates and carboxylates; blue, amino acids/peptides/amines; yellow, nucleotides/nucleosides; and black, drug/polysaccharide efflux or unknown.

DOI: 10.1371/journal.pgen.0020021.g006

Last universal common ancestor (LUCA)

- Τέτοιου είδους αναλύσεις, έχουν χρησιμοποιηθεί για να διαλευκανθεί το ερώτημα που αφορά τον τελευταίο κοινό πρόγονο όλων των σύγχρονων οργανισμών (Last Universal Common Ancestor-LUCA). Οι αναλύσεις ξεκίνησαν με τη μελέτη του οργανισμού με το μικρότερο γονιδίωμα, του βακτηρίου *Mycoplasma genitalium* το οποίο είναι υποχρεωτικό ενδοκυτταρικό παράσιτο και κωδικοποιεί μόλις 468 γονίδια που παράγουν πρωτεΐνες. Ακόμα και σε σύγκριση με κάποιο άλλο βακτήριο, π.χ. με το *Haemophilus influenzae* (1703 γονίδια) γίνεται εμφανές ότι μόνο 240 γονίδια του *M. genitalium* έχουν ορθόλογα γονίδια στον *H. influenzae*.
- Το ερώτημα λοιπόν ήταν αν ο LUCA ήταν ένας οργανισμός με λίγα γονίδια (όπως π.χ. το *Mycoplasma*) ή αν, αντίθετα, ήταν οργανισμός με περισσότερα γονίδια (όπως τα περισσότερα βακτήρια) και τελικά η εξέλιξη οδήγησε κάποιους οργανισμούς να χάσουν τα γονίδια αυτά και άλλους να αποκτήσουν κάποια νέα. Συγκριτικές αναλύσεις, με κάποιες παραδοχές (όπως π.χ. ότι δεν αναμένουμε σε όλους τους οργανισμούς να είναι συντηρημένα όλα τα γονίδια), έδειξε ότι μάλλον η δεύτερη εκδοχή είναι η σωστή.
- Για παράδειγμα, όταν στην ανάλυση συμπεριλήφθηκαν μόνο προκαρυώτες, βρέθηκε ότι ο κοινός πρόγονος όλων των οργανισμών πρέπει να είχε γονίδια μεταξύ 1006 και 1189, ενώ όταν συμπεριλήφθηκαν και οι ευκαρυώτες, ο αριθμός ανέβηκε στο 1344 με 1529, -αριθμοί που είναι πιο κοντά στο μέσο όρο των σημερινών βακτηρίων παρά στο ελάχιστο (δηλαδή στο *Mycoplasma*) (Ouzounis Kunin, Darzentas, & Goldovsky, 2006)

Minimal gene set

- *Mycoplasma genitalium* (468 identified protein-coding genes)
- *Haemophilus influenzae* (1703 genes)
- 240 *M. genitalium* genes have orthologs among the genes of *H. influenzae*.
- 22 nonorthologous displacements

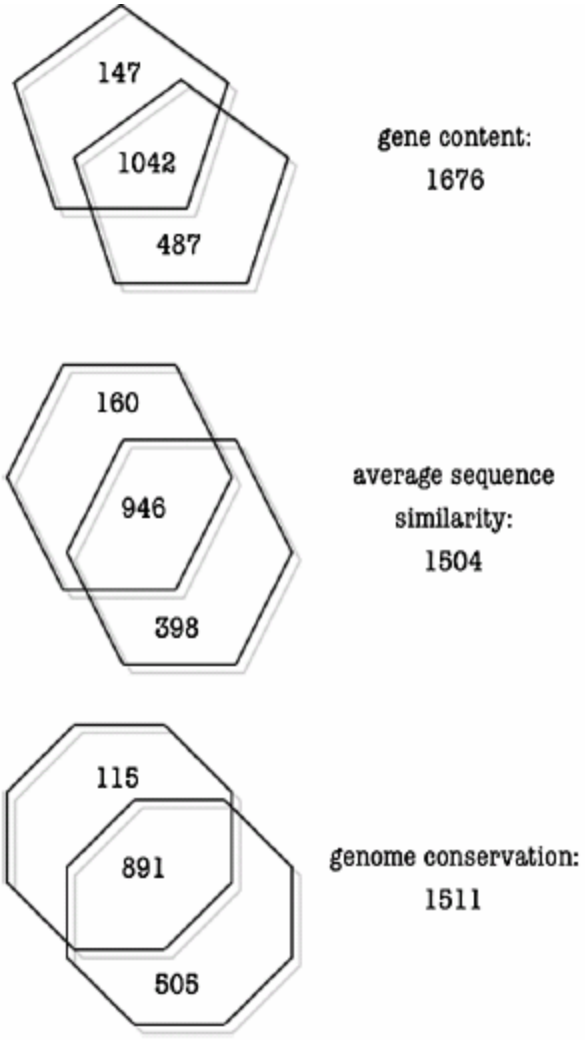


Fig. 1. A representation of the minimal gene content for LUCA. Upper diagrams represent analyses without eukaryotes, lower diagrams represent analyses with eukaryotes; pentagons represent gene content (CT), hexagons represent average sequence similarity (AS), octagons represent genome conservation (GC)—see Section 2. The number of unique (outside the intersection) and common (inside the intersection) gene families per category are given in the diagrams; the number of total unique families is also provided (listed below the corresponding method).

Συνθετική βιολογία και εξωβιολογία

- Παρόμοιες αναλύσεις, έχουν μεγάλο ενδιαφέρον και στη λεγόμενη «εξωβιολογία», τον κλάδο δηλαδή που μελετάει θεωρητικά το πώς αναμένουμε να είναι οι οργανισμοί που ενδεχομένως βρεθούν σε άλλους πλανήτες, αλλά και στη συνθετική βιολογία και τη γενετική μηχανική.
- Για παράδειγμα, τέτοιου είδους αναλύσεις, έκαναν δυνατό τον υπολογισμό των απαραίτητων γονιδίων που απαιτούνται για να συντηρήσουν τη ζωή σε ένα βακτήριο και εφαρμόστηκαν πρόσφατα όταν επιστήμονες συνέθεσαν εξ' ολοκλήρου ένα βακτηριακό γονιδίωμα 1Mbp και το ενσωμάτωσαν σε ένα βακτηριακό κύτταρο από το οποίο είχαν αφαιρέσει το γονιδίωμα.
- Το «νέο» βακτήριο, το οποίο χρησιμοποιεί αποκλειστικά το συνθετικό DNA (*Mycoplasma mycoides* JCVI-syn1.0), είχε τις αναμενόμενες φαινοτυπικές λειτουργίες και ήταν ικανό να αναπαράγεται (Gibson et al., 2010).

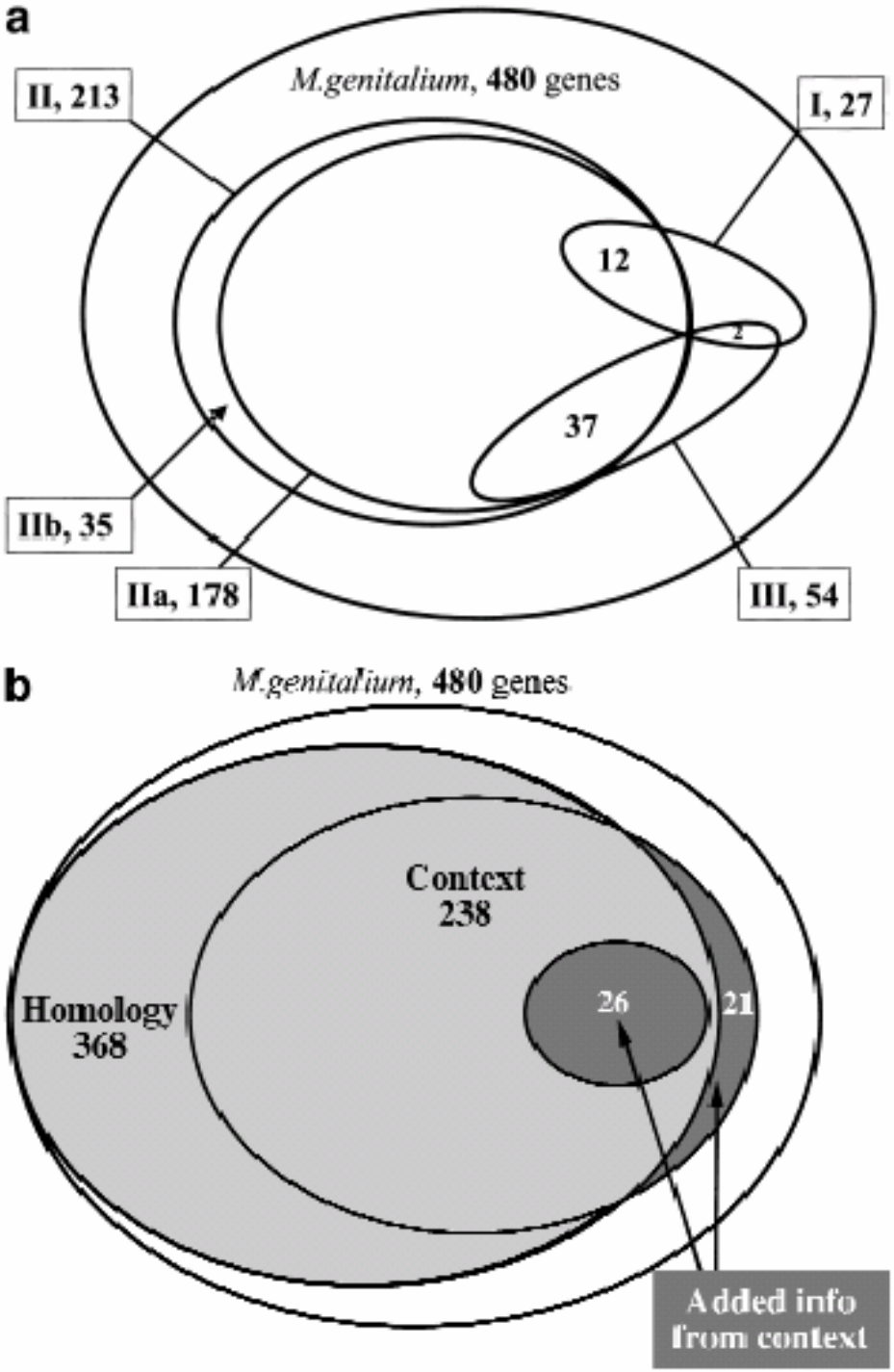
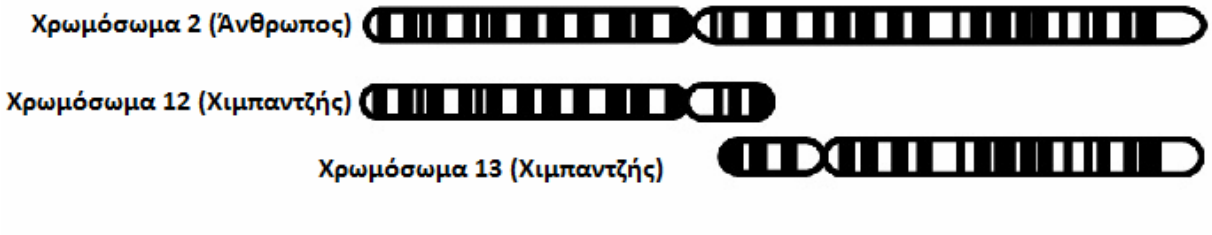


Figure 1 (a) Coverage of and overlap between various types of genomic context for *M. genitalium* genes. Type I is gene-fusion. Type II is the conservation local gene neighborhood, which is separated in type IIa (the conservation of gene order) and type IIb (the co-occurrence of genes within potential operons in absence of the conservation of gene order). Type III is the co-occurrence of genes in genomes. (b) Overlap between genes for which significant genomic context is available and genes for which functional features can be predicted by homology searches. For the latter, only genes that are homologous to genes with known molecular functions were included, which were determined by manual inspection. The dark gray areas in the figure are genes for which new functional features can be predicted by genomic context. They can be homologous to proteins with a known molecular function, in which case the context can indicate in which process this function plays a role (see text for specific examples). A complete list of genes for which new functional features could be predicted by genomic context and, if available, homology to proteins with known function, is available from http://dove.embl-heidelberg.de/MG/Context.

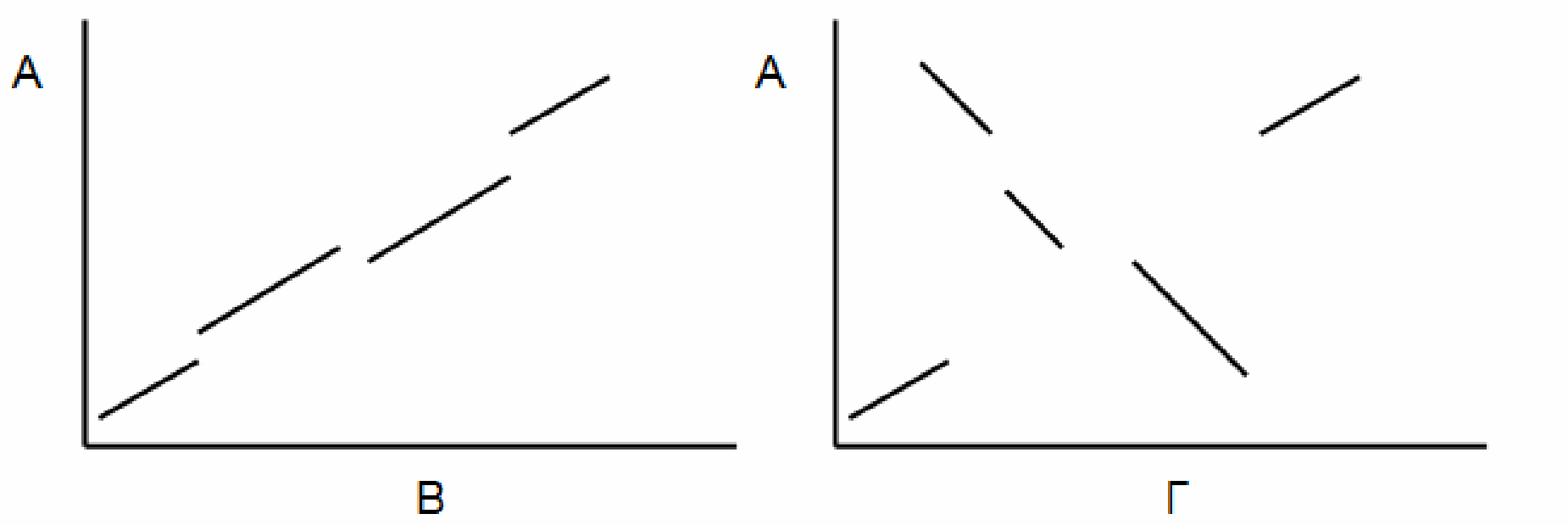
Η μέθοδος στοίχισης ολόκληρων γονιδιωμάτων

- Η μέθοδος αυτή βασίζεται στην ίδια αρχή με τις στοιχίσεις αλληλουχιών (οι συγγενικοί οργανισμοί είναι πιο πιθανό να έχουν μεγάλες ομοιότητες στο γονιδίωμα τους). Με τη μέθοδο αυτή στοιχίζονται ολόκληρα γονιδιώματα και εντοπίζονται οι περιοχές στις οποίες έχουν μεγάλη ομοιότητα. Τέτοιες τεχνικές σε πιο πρώιμη μορφή ήταν γνωστές από παλιά, π.χ. από παρατηρήσεις ότι το ανθρώπινο DNA υβριδοποιείται με το αντίστοιχο του χιμπατζή, είχε γίνει γνωστό ότι τα γονιδιώματα του ανθρώπου και των άλλων μεγάλων πιθήκων έχουν μεγάλη ομοιότητα. Παρόμοιες ανακαλύψεις είχαν γίνει και με τη χρήση καρυότυπου, όταν για παράδειγμα έγινε γνωστό ότι το χρωμόσωμα 2 του ανθρώπου εμφανίζει μερική ομοιότητα με το χρωμόσωμα 12 και 13 του χιμπατζή και έγινε κατανοητό ότι στο απώτατο παρελθόν είχε προκύψει από σύντηξη τελομερών.

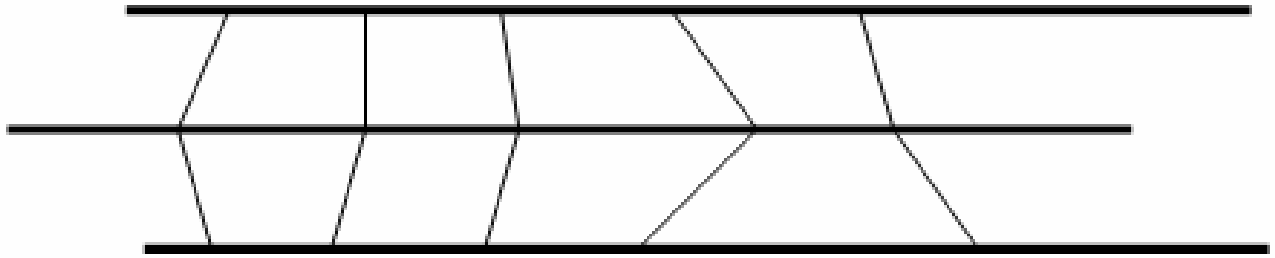


Μεθοδολογίες

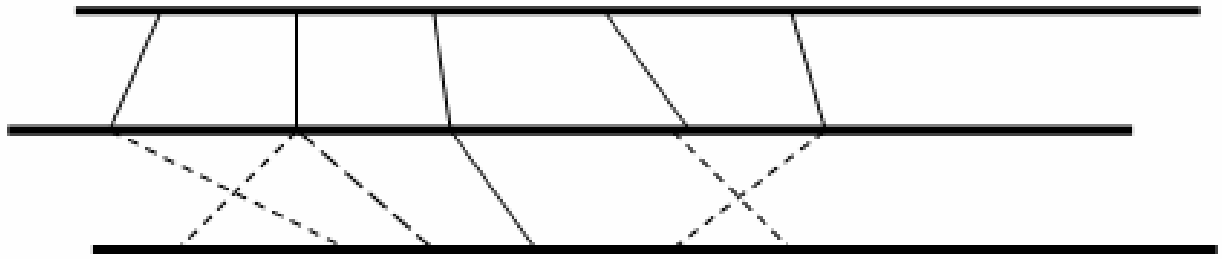
- Στοίχιση ολόκληρων γονιδιωμάτων, μπορεί να γίνει με διαφορετικούς τρόπους. Ο πιο απλός είναι με μια παραλλαγή του γνωστού διαγράμματος σημείων (dot-plot) η οποία επεκτείνεται σε όλο το γονιδίωμα ή με κάποια επέκταση κάποιου γνωστού αλγόριθμου στοίχισης (όπως το BLAST) η οποία να επιτρέπει χρήση μεγάλων ακολουθιών. Οι πιο σύγχρονες μεθοδολογίες, συνδυάζουν τόσο τους αλγορίθμους τοπικής ή ολικής στοίχισης (για κάθε ζευγάρι ομόλογων γονιδίων) με τη θέση των γονιδίων αυτών στο γονιδίωμα, δείχνοντας π.χ. με διαφορετικό χρωματισμό τα ζευγάρια, ενώ κάποιες από τις τεχνικές αυτές επιτρέπουν και πολλαπλή στοίχιση. Όπως γίνεται εύκολα αντιληπτό, οι τεχνικές αυτές είναι πολύ πιο εύκολο να εφαρμοστούν σε βακτηριακά ή ϊικά γονδιώματα, τόσο γιατί είναι πιο μικρά όσο και γιατί είναι ενιαία, καθώς τα πολλαπλά χρωμοσώματα των ευκαρυωτικών οργανισμών απαιτούν σύγκριση ένα με ένα.


A
B
A
Г

A



B



- Οι μεθοδολογίες ολικής στοίχισης γονιδιωμάτων είναι δυνατό να δώσουν πολλές πληροφορίες για τις αλλαγές που έχουν συμβεί στα γονιδιώματα στο πέρασμα του εξελικτικού χρόνου.
- Για παράδειγμα, μια στοίχιση και ένα διάγραμμα σημείων περίπου στο ύψος της διαγωνίου δείχνει την κοινή προέλευση και τη στενή σχέση των δύο οργανισμών.
- Επιπλέον, αλλαγές μεγάλης κλίμακας όπως αναστροφές και διπλασιασμοί είναι ιδιαίτερα εύκολο να εντοπιστούν.
- Τέλος, περιοχές μη ομοιότητας ανάμεσα σε 2 κατά κανόνα «όμοια» γονιδιώματα είναι δυνατό να δείξουν πρόσφατη απόκτηση γενετικού υλικού (είτε με οριζόντια μεταφορά είτε με κάποιον άλλο τρόπο ενσωμάτωσης DNA).

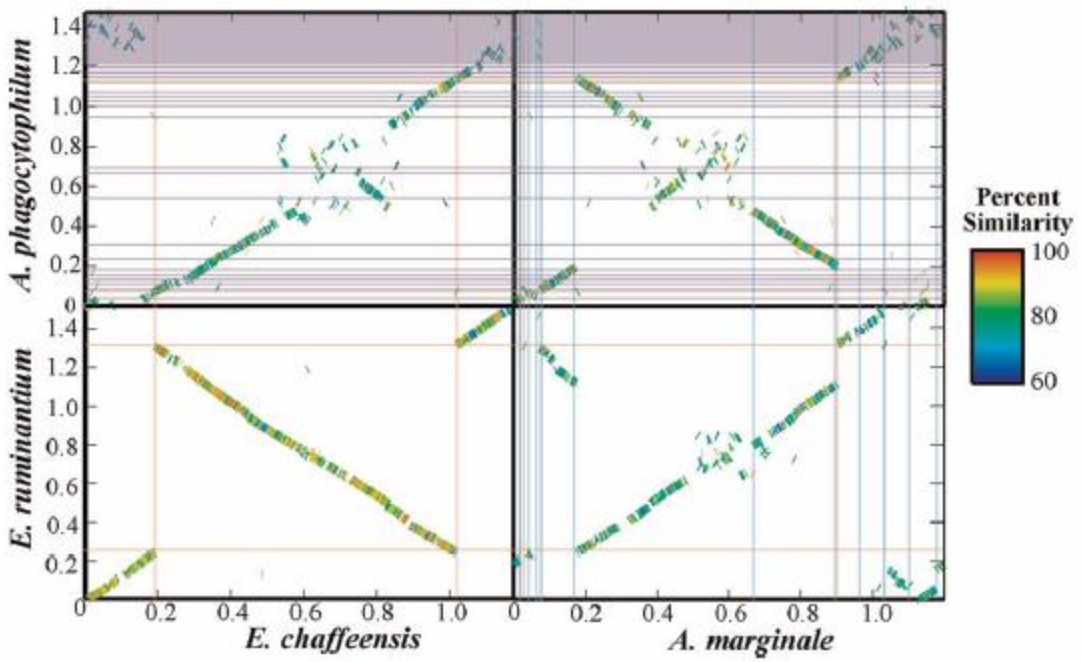


Figure 3. Synteny between *Anaplasma* spp. and *Ehrlichia* spp.

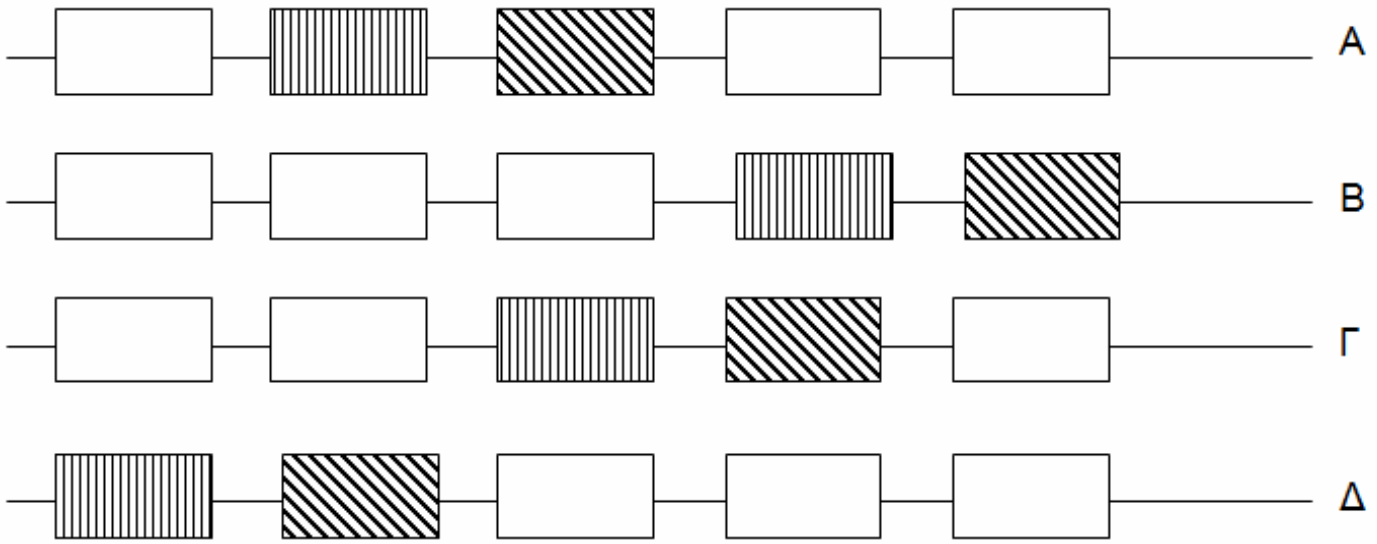
Anaplasma spp. and *Ehrlichia* spp. share conserved gene order (synteny) across their chromosomes. *E. ruminantium* and *E. chaffeensis* have a single symmetrical inversion near two duplicate Rho termination factors (approximate positions shown in pink). Genomic rearrangements between these Rho termination factors are also apparent in *A. marginale* (pink). In addition to the synteny breaks near the Rho termination factors, *A. marginale* has rearrangements located near the *msp2*- and *msp3*-expression locus and pseudogenes (approximate positions shown in light blue). Likewise, in *A. phagocytophilum,* numerous changes in genome arrangement are located near the homologous *p44* expression locus and silent genes (approximate positions shown in lavender).

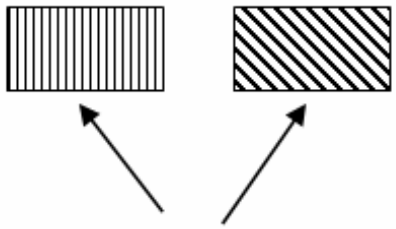
DOI: 10.1371/journal.pgen.0020021.g003

Η μέθοδος σύγκρισης της σειράς των γονιδίων

- Σύμφωνα με μέθοδο αυτή εντοπίζονται γονίδια που έχουν την τάση να βρίσκονται κοντά σε όλα ή στα περισσότερα τα υπό μελέτη γονιδιώματα. Η βασική αρχή της μεθόδου μοιάζει διαισθητικά με την αρχή της σύνδεσης στη γενετική, μόνο που εδώ χρησιμοποιείται σε μεγαλύτερη κλίμακα χρόνου. Η ιδέα είναι ότι γονίδια που βρίσκονται σε πολλούς οργανισμούς δίπλα-δίπλα, το κάνουν για κάποιο λόγο (π.χ. εκφράζονται μαζί ή συμμετέχουν σε κάποιο κοινό μεταβολικό μονοπάτι). Ειδικά στα βακτήρια, είναι γνωστό ότι ομάδες γονιδίων που συμμετέχουν στο ίδιο μονοπάτι, βρίσκονται οργανωμένα σε ομάδες που ονομάζονται οπερόνια, ομάδες οι οποίες εκφράζονται και ελέγχονται ταυτόχρονα.

- Με τη μέθοδο αυτή είναι δυνατό να εντοπιστούν συσχετίσεις μεταξύ γονιδίων που κωδικοποιούν τελείως διαφορετικές πρωτεΐνες. Για παράδειγμα, αν υποθέσουμε ότι στο οπερόνιο της λακτόζης, ξέραμε τη λειτουργία της γαλακτοσιδάσης (lacZ) αλλά όχι αυτή της περμεάσης (lacY), με την παρατήρηση ότι σε μια σειρά από οργανισμούς τα δύο γονίδια βρίσκονται πάντα μαζί, θα μπορούσαμε να συμπεράνουμε ότι αποτελούν και τα δύο τμήμα κάποιου οπερονίου. Δεν θα ξέραμε φυσικά ακριβώς τη λειτουργία του νέου γονιδίου, αλλά συνδυάζοντας κάποιες απλές μεθόδους πρόγνωσης, όπως για παράδειγμα την πρόγνωση διαμεμβρανικών τμημάτων, θα βλέπαμε ότι πρόκειται για διαμεμβρανική πρωτεΐνη με 12 πιθανά διαμεμβρανικά τμήματα και αμέσως θα υποθέταμε ότι πρόκειται για κάποιον διαμεμβρανικό υποδοχέα που πιθανότατα εμπλέκεται με το μεταβολισμό της λακτόζης. Μια τόσο λεπτομερής πρόβλεψη για τη λειτουργία μιας πρωτεΐνης δεν θα μπορούσε με κανέναν τρόπο να γίνει δυνατή με χρήση μόνο της ακολουθίας της, αλλά βλέπουμε ότι αυτό συμβαίνει όταν χρησιμοποιήσουμε την πληροφορία από τη σειρά των γονιδίων και τη συντήρησή της στα γονδιώματα.





Τα γονίδια αυτά συνδέονται
με κάποιον τρόπο

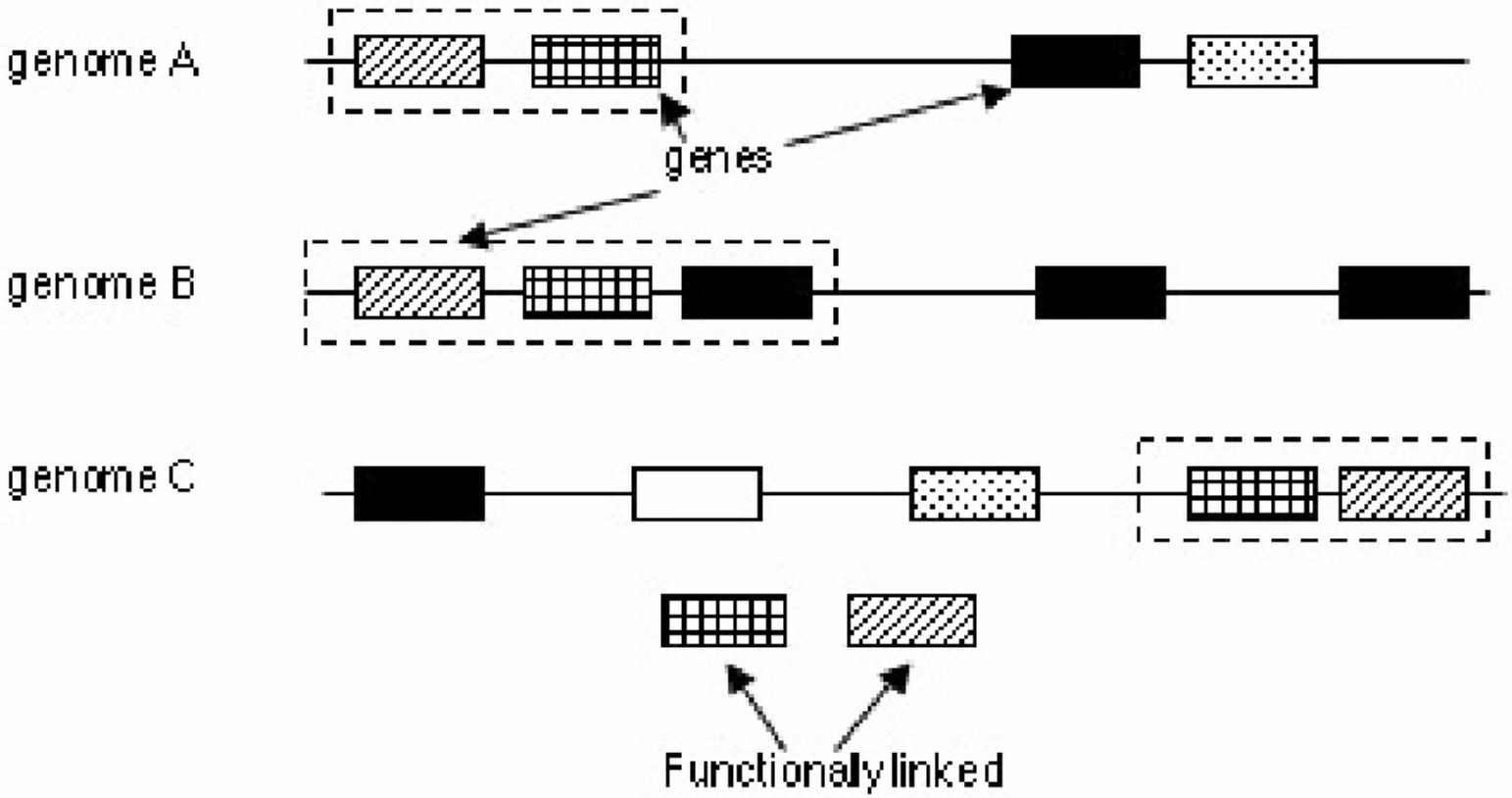
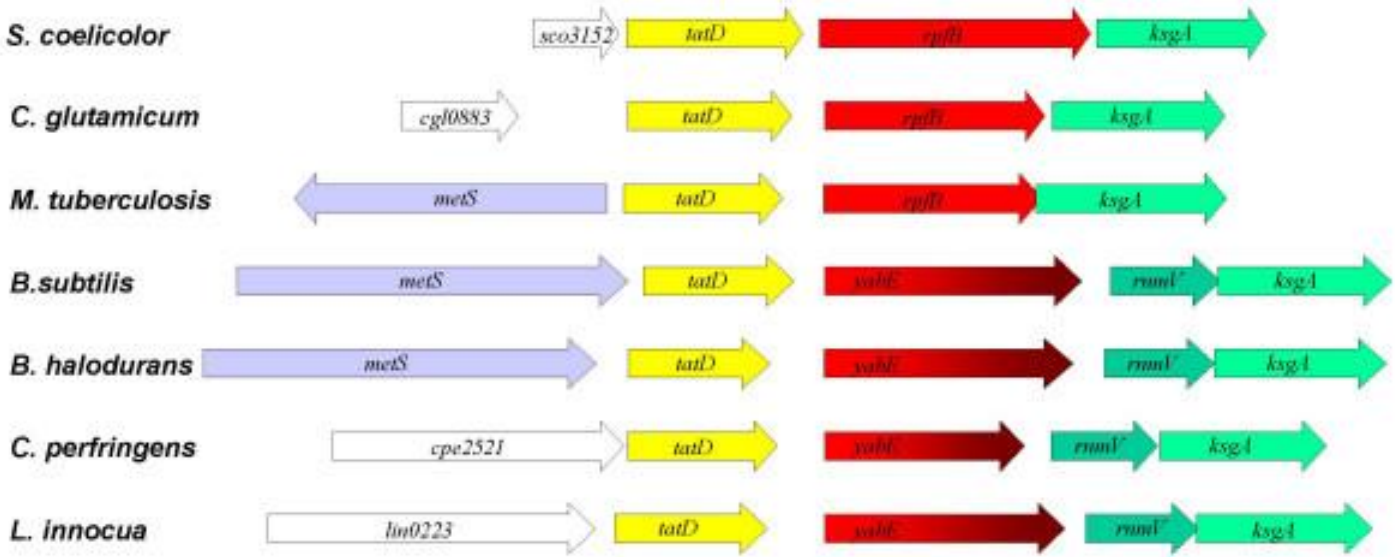
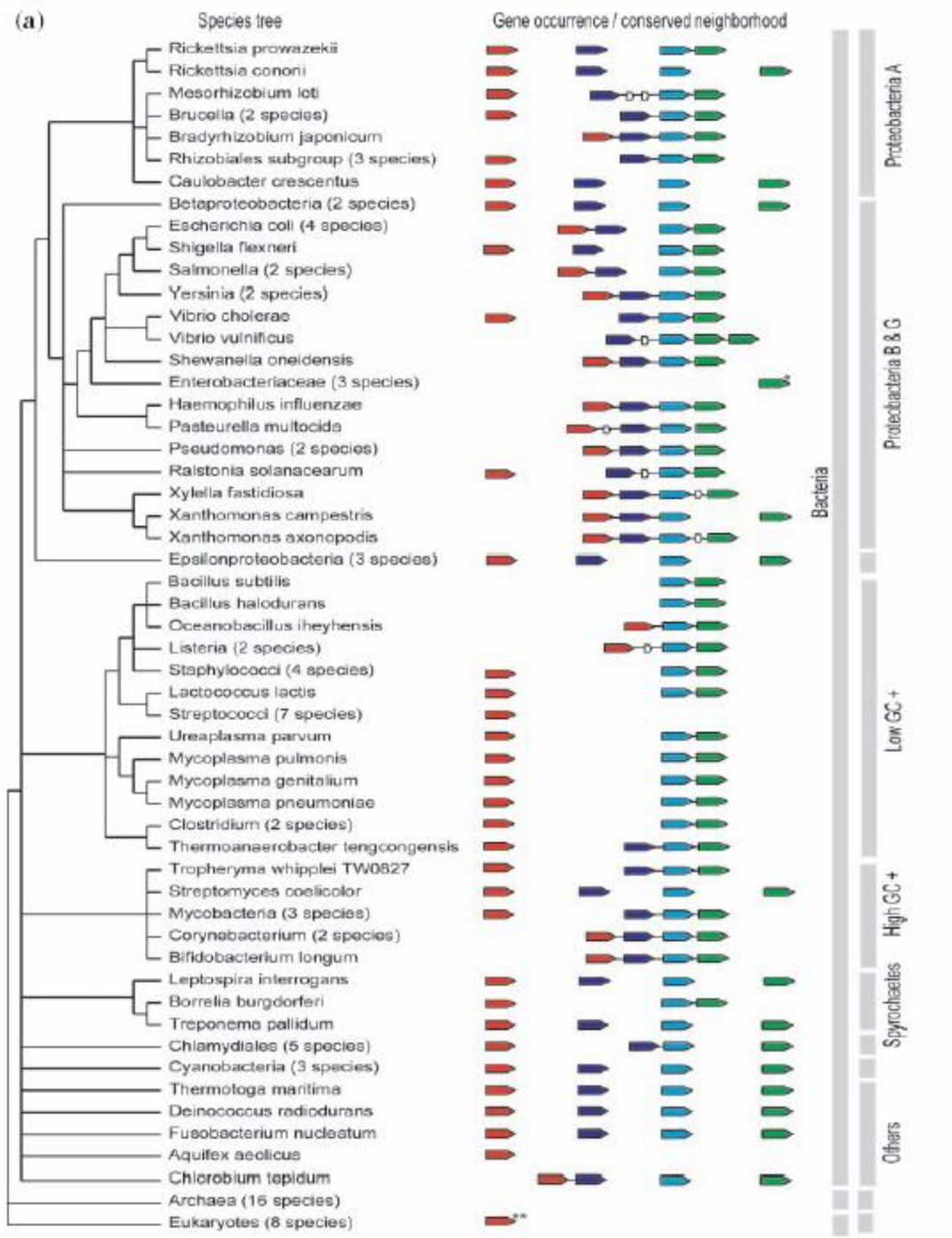


Fig. 2. Conservation of gene order/neighborhood. Genes that are consistent neighbors across multiple genomes may be function-


S. coelicolor
sco3152
tatD
rpfB
ksgA
C. glutamicum
cgl0883
tatD
rpfB
ksgA
M. tuberculosis
metS
tatD
rpfB
ksgA
B.subtilis
metS
tatD
yabE
rnmV
ksgA
B. halodurans
metS
tatD
yabE
rnmV
ksgA
C. perfringens
cpe2521
tatD
yabE
rnmV
ksgA
L. innocua
lin0223
tatD
yabE
rnmV
ksgA


(a)
Species tree
Gene occurrence / conserved neighborhood
Rickettsia prowazekii
Rickettsia conorii
Mesorhizobium loti
Brucella (2 species)
Bradyrhizobium japonicum
Rhizobiales subgroup (3 species)
Caulobacter crescentus
Betaproteobacteria (2 species)
Escherichia coli (4 species)
Shigella flexneri
Salmonella (2 species)
Yersinia (2 species)
Vibrio cholerae
Vibrio vulnificus
Shewanella oneidensis
Enterobacteriaceae (3 species)
Haemophilus influenzae
Pasteurella multocida
Pseudomonas (2 species)
Ralstonia solanacearum
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas axonopodis
Epsilonproteobacteria (3 species)
Bacillus subtilis
Bacillus halodurans
Oceanobacillus iheyhensis
Listeria (2 species)
Staphylococci (4 species)
Lactococcus lactis
Streptococci (7 species)
Ureaplasma parvum
Mycoplasma pulmonis
Mycoplasma genitalium
Mycoplasma pneumoniae
Clostridium (2 species)
Thermoanaerobacter tengcongensis
Tropheryma whipplei TW0827
Streptomyces coelicolor
Mycobacteria (3 species)
Corynebacterium (2 species)
Bifidobacterium longum
Leptospira interrogans
Borrelia burgdorferi
Treponema pallidum
Chlamydiales (5 species)
Cyanobacteria (3 species)
Thermotoga maritima
Deinococcus radiodurans
Fusobacterium nucleatum
Aquifex aeolicus
Chlorobium tepidum
Archaea (16 species)
Eukaryotes (8 species)
**
Bacteria
Proteobacteria A
Proteobacteria B & G
Low GC +
High GC +
Spyrochaetes
Others

(b)

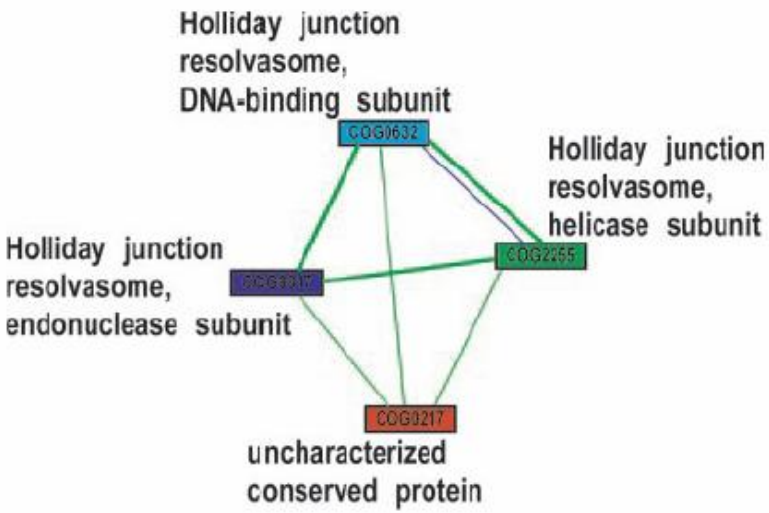


Figure 2. (**a**) Case study: evidence linking COG0217 to well-annotated proteins. Species tree showing conserved operon architecture and co-occurrence of genes coding for subunits of Holliday junction resolvasome (COG0217: uncharacterized conserved protein (red), COG0632: Holliday junction resolvasome, DNA-binding subunit (light blue), COG2255: Holliday junction resolvasome, helicase subunit (dark blue), COG0817: Holliday junction resolvasome, endonuclease subnit (green). Single asterisk (*), not present in Buchnera species and double asterisks(**), not present in Encephalitozoon cuniculi. (**b**) Network representation of evidence related to COG0217 (red). The network edges represent the predicted functional associations. An edge may be drawn with up to three different colour lines—these lines represent the existence of the three types of evidence used in predicting the associations. A red line indicates the presence of fusion evidence; a green line represents the neighbourhood evidence; and a blue line the co-occurrence evidence. Line thickness correlates linearly with STRING scores.

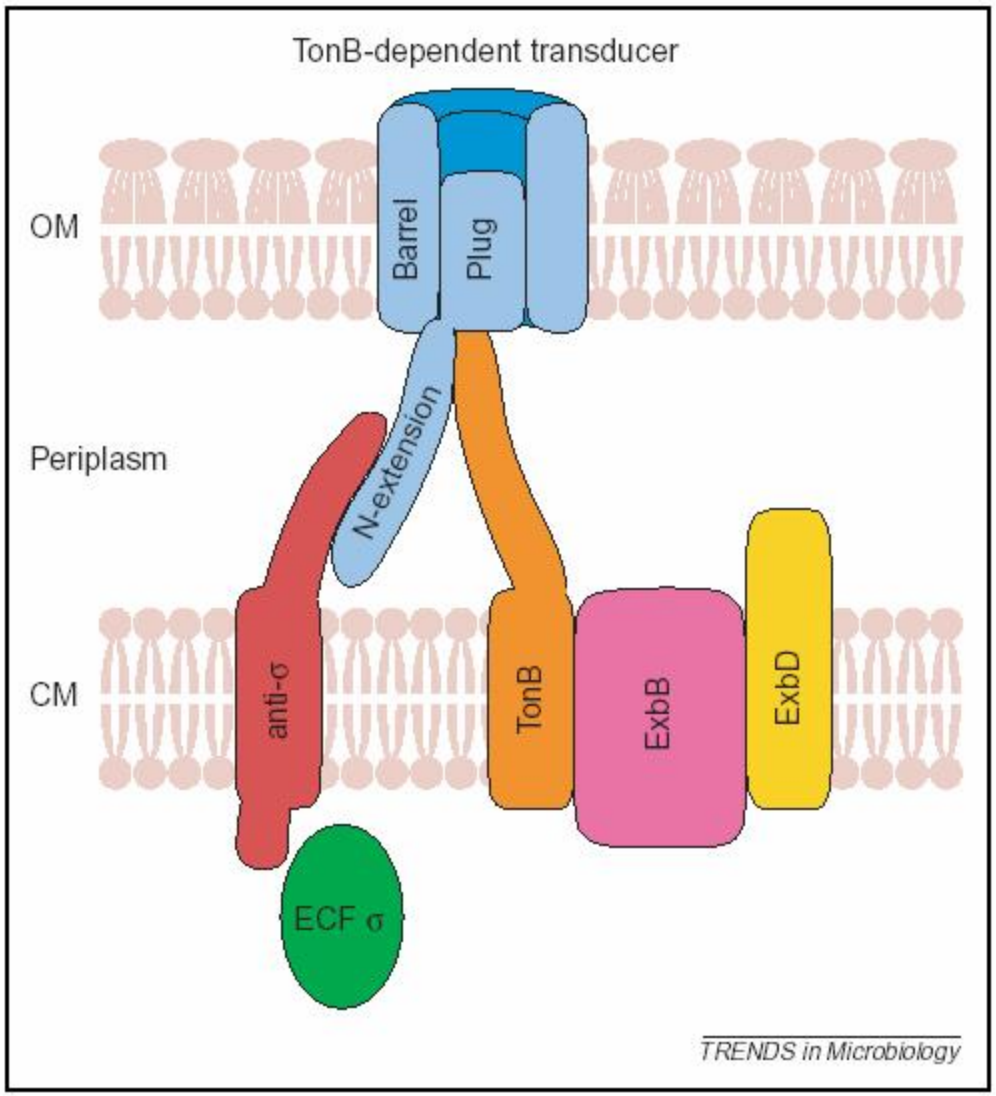


Figure 1. Structural organization of TonB-dependent regulatory systems. TonB-dependent regulatory systems consist of six components, an outer membrane TonB-dependent transducer (blue) in interplay with its energizing TonB–ExbBD protein complex (orange, pink and yellow), a cytoplasmic membrane-localized anti-sigma factor (red) and an ECF-subfamily sigma factor (green). Abbreviations: CM, cytoplasmic membrane; OM, outer membrane.

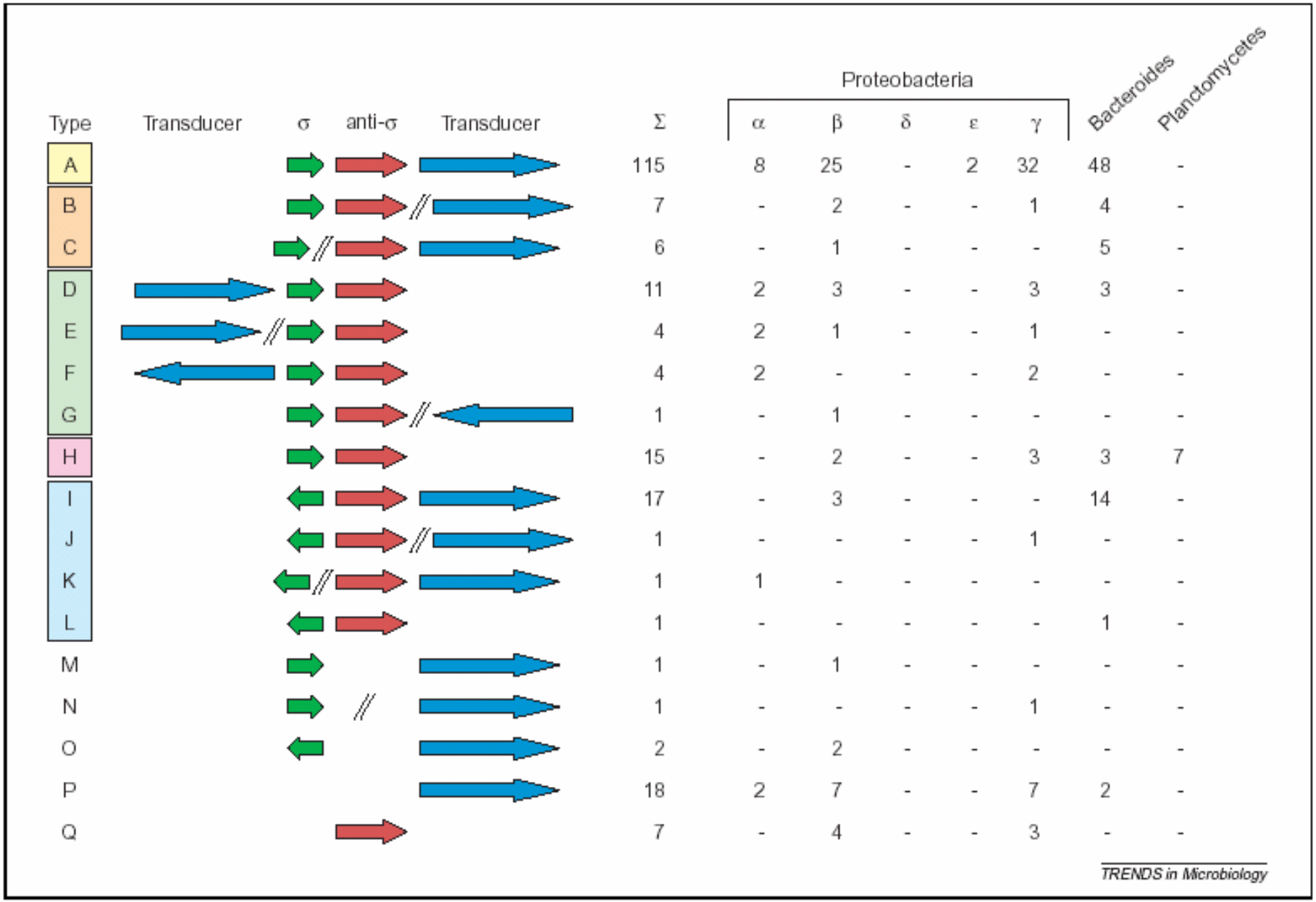


Figure 2. Genetic organization of regulatory systems consisting of a TonB-dependent transducer (blue), an anti-sigma factor (red) and an ECF-subfamily sigma factor (green). The occurrence of the different genetic organizations in proteobacteria (α, β, δ, ε and γ), in *Bacteroides*, in planctomycetes and in total (Σ) is shown. A diagonal double line indicates that one or a few additional genes are present between the TonB-dependent regulatory genes. Pseudogenes have been included in this representation. The color code of the type (left) corresponds to that of Supplementary Table 2.

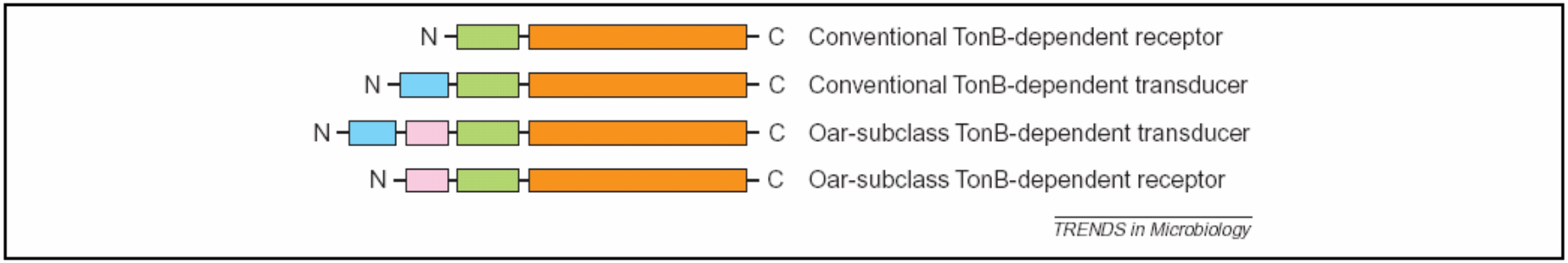


Figure 3. Domain structure of TonB-dependent receptors. All TonB-dependent receptors consist of a C-terminal β-barrel (orange) and a plug domain (green), which seals the barrel (see also Figure 1). TonB-dependent transducers have an N-terminal extension (blue) of ~70 amino acids. Receptors from *Bacteroides* often have another additional domain in the N-terminal region (pink). A related protein domain is also found in the Oar protein from *M. xanthus* and in a few receptors from *Xanthomonas* and *Xylella* species.

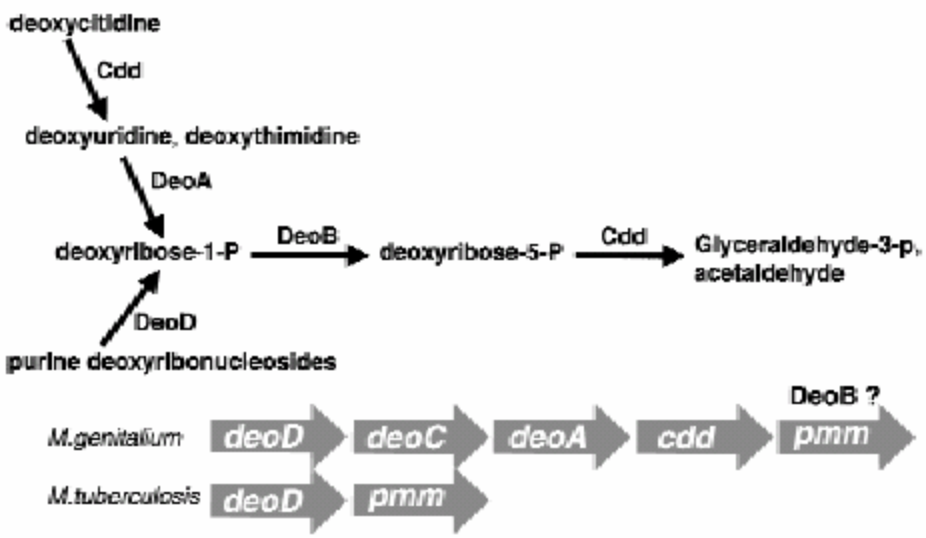


Figure 3 Genomic context predicts substrate specificity of proteins involved in a nucleoside salvage pathway in *M. genitalium*. A cluster of five genes in *M. genitalium* encodes four genes of a nucleoside salvage pathway. The "standard" gene for this fifth reaction in the pathway, phosphoribomutase (*deoB*), is absent. The fifth gene in the operon is homologous to phosphomannomutases and phosphoglucomutases. *M. genitalium* does not contain any other candidate for a phosphoribomutase. The most likely candidate for the phosphoribomutase is thus MG053. The significance of the location of a homolog of MG053 in a run with *deoD* is supported by the location of a homolog of the *M. genitalium* gene MG053 beside *deoD* in *Mycobacterium tuberculosis*.

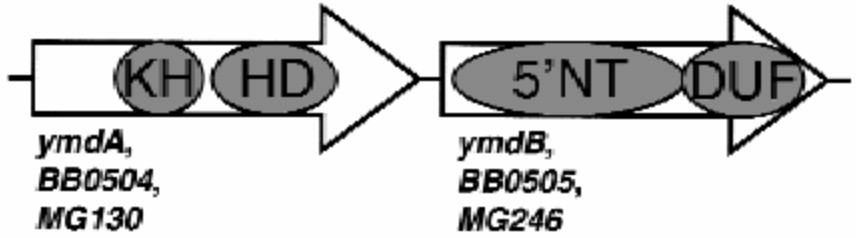
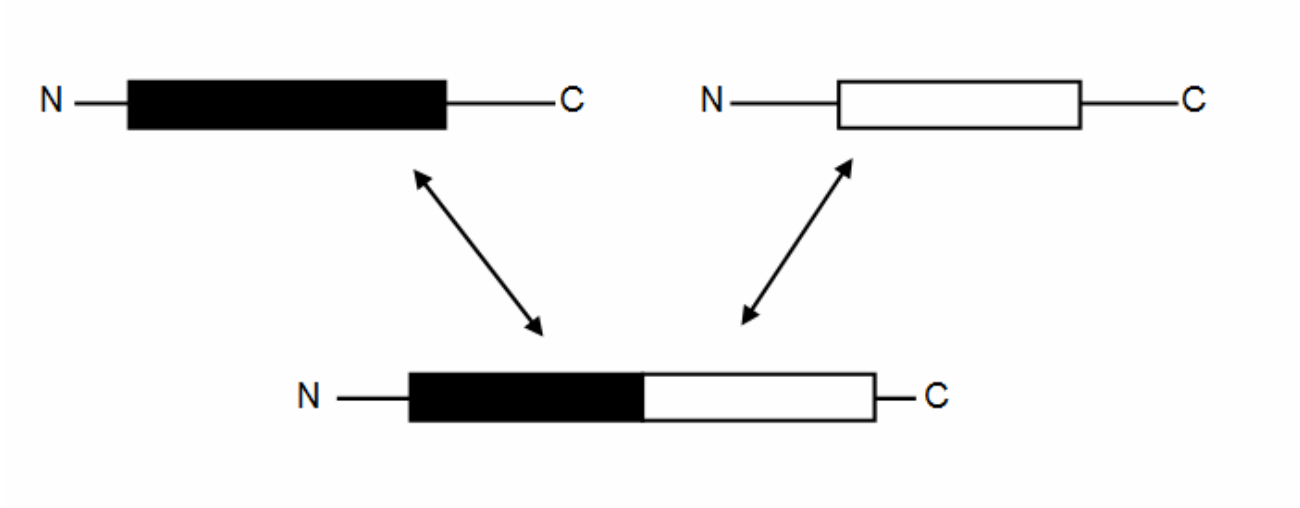


Figure 4 Domain organization of two proteins that are encoded by neighboring genes on *B. subtilis* (*ymdA* and *ymdB*) and *B. burgdorferi* (BB0504 and BB0505), and that are both present in *M. genitalium* (MG130 and MG246). The three domains that have functionally been characterized, KH, HD, and 5′NT, can all be related to ribonucleotide metabolism. KH binds (single-stranded) RNA; HD hydrolyzes phosphates from nucleotides; and 5′NT hydrolyzes NMP to nucleosides. A fourth, uncharacterized sequence domain (DUF) is present at C-terminus of MG246 and its orthologs.

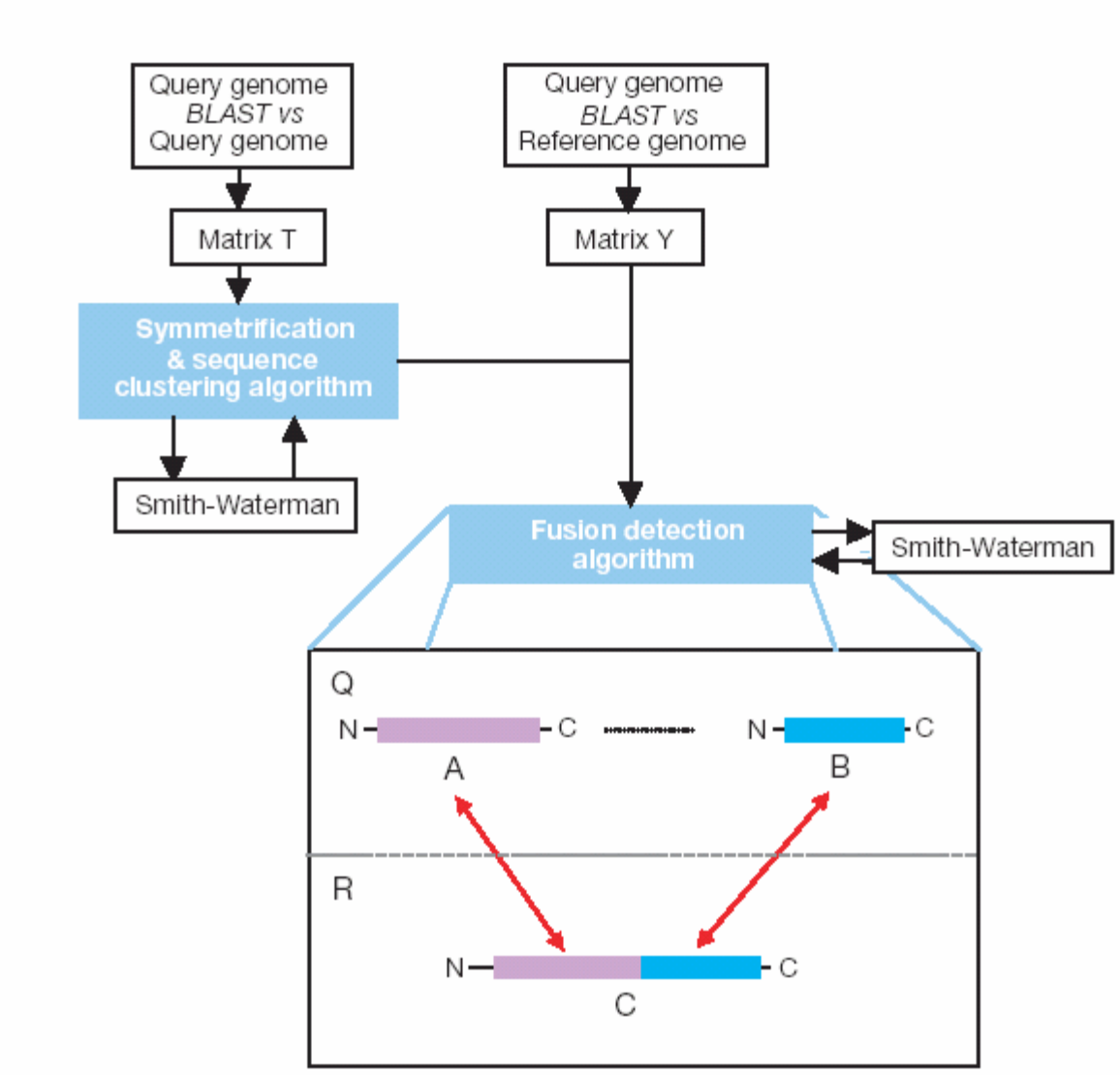
Η μέθοδος εντοπισμού προϊόντων γονιδιακής σύντηξης

- Η βασική αρχή αυτής της μεθόδου βασίζεται στη σπονδυλωτή φύση των πρωτεϊνών, δηλαδή, στην ύπαρξη ανεξάρτητων δομικών και λειτουργικών περιοχών (domains). Έτσι, με τη μέθοδο αυτή εντοπίζονται με υπολογιστικό τρόπο γονίδια ενός οργανισμού Α τα οποία σε κάποιον άλλον οργανισμό Β βρίσκονται ενωμένα, λειτουργούν δηλαδή σαν ανεξάρτητες περιοχές της ίδιας πρωτεΐνης. Η εξήγηση είναι ότι σε κάποια προγονική μορφή, είτε τα γονίδια βρίσκονταν ανεξάρτητα και συνενώθηκαν (σύντηξη γονιδίων)με το πέρασμα του χρόνου στον οργανισμό Β, είτε ότι σε κάποια προγονική μορφή τα γονίδια βρίσκονταν ενωμένα, ήταν δηλαδή πρωτεϊνικές περιοχές και κατόπιν στην πορεία της εξέλιξης αυτή η σχέση διακόπηκε στον οργανισμό Α (Enright, Iliopoulos, Kyrpides, & Ouzounis, 1999). Με τη μέθοδο αυτή, δεν μπορούμε να διακρίνουμε ποια από τις δύο εναλλακτικές όντως συνέβη, αλλά αυτό δεν αποτελεί πρόβλημα σε αυτές τις αναλύσεις, γιατί μπορούμε να εξάγουμε ούτως ή άλλως σημαντικά συμπεράσματα για πρωτεΐνες που ούτε ομοιότητα έχουν, αλλά και ούτε βρίσκονται κοντά στο γονιδίωμα.



- Συνήθως τέτοιες περιπτώσεις γονιδίων αφορούν ένζυμα τα οποία εμπλέκονται στον ίδιο μεταβολικό δρόμο, πιθανότατα το προϊόν του ενός να είναι αντιδρόν στο άλλο και με αυτόν τον τρόπο διευκολύνονται οι μεταβολικές οδοί. Ένα κλασικό παράδειγμα, είναι η διυδροφολική αναγωγάση (DHFR) η οποία στους ευκαρυωτικούς οργανισμούς αποτελεί μια πρωτεΐνη με μια μοναδική πρωτεϊνική περιοχή, αλλά στα βακτήρια στο ίδιο μόριο συνυπάρχει και η λειτουργική περιοχή της θυμιδιλικής συνθετάσης (TS) η οποία συμμετέχει στο ίδιο μονοπάτι (σύνθεση νουκλεοτιδίων) και η οποία στους ευκαρυωτικούς οργανισμούς βρίσκεται σε διαφορετικό γονίδιο.

Protein interaction maps for complete genomes based on gene fusion events



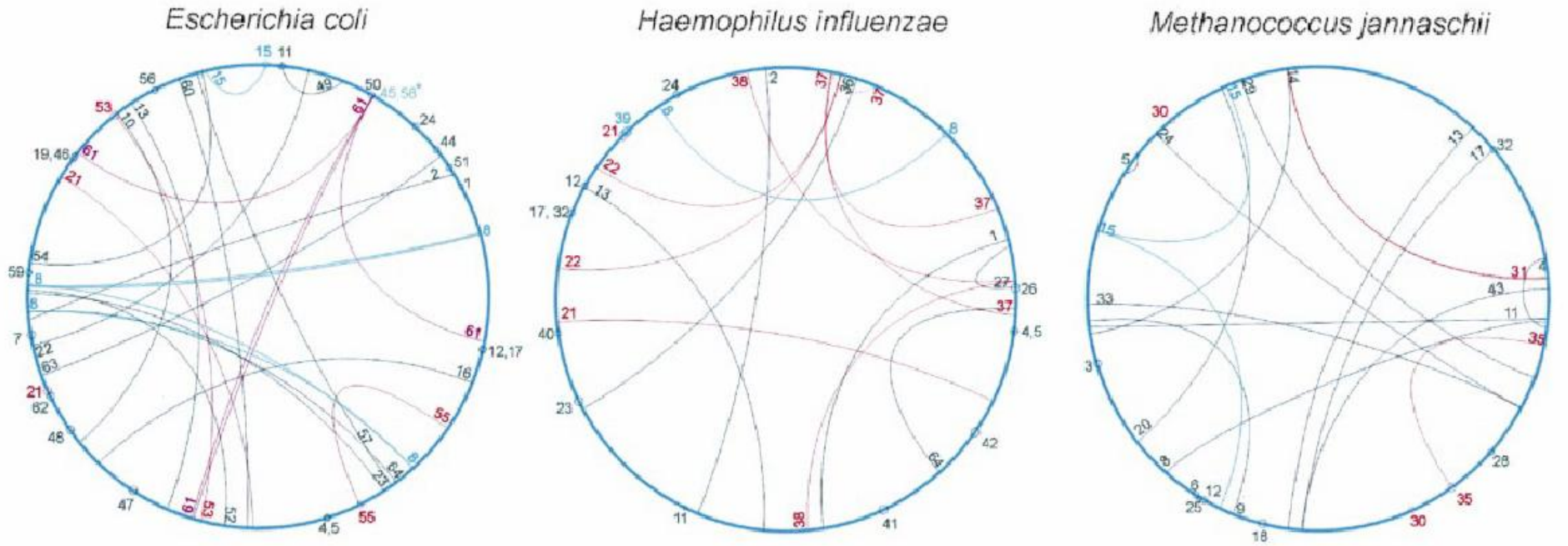


Figure 2 Representation of protein interaction maps for the most likely interactions predicted for *E. coli*, *H. influenzae* and *M. jannaschii*. In the large blue circles, which represent the three genomes, 0° corresponds to the first base pair, and 360° the last base pair of the genome. Predicted interactions are indicated by linking the circular map positions of the genes involved. In cases of neighbouring genes (<5°), a small circle indicates the predicted interaction between two genes at that region; otherwise, an arc links the two genes in question. Multiple interactions are not cross-labelled. Some paralogous cases are resolved and only the most likely case is indicated by an arc. All cases are numbered according to Table 1. Predictions are colour coded: black, pairwise interactions; blue, multiple interactions; red/purple, cases where, due to paralogy, more than one pairwise interaction is possible (red, two possibilities; purple, more than two possibilities); green (marked by asterisk), because of a large number of paralogues, no interaction can be easily resolved. The source of the prediction (composite protein from a given species) is not indicated.

Enright et al, NATURE,402, 1999

- Η μέθοδος αυτή, είναι υπολογιστικά απαιτητική καθώς απαιτεί μία προς μία στοιχίσεις όλων των πρωτεϊνών του ενός οργανισμού, με όλες τις πρωτεΐνες του άλλου οργανισμού, ενώ απαιτείται και επιπλέον επεξεργασία για να διασφαλιστεί ότι οι δύο υποψήφιες πρωτεΐνες μοιάζουν μεν με μια άλλη πρωτεΐνη του άλλου οργανισμού αλλά σε διαφορετική περιοχή (δηλαδή, ότι δεν μοιάζουν μεταξύ τους). Από την άλλη μεριά, ένα σημαντικό πλεονέκτημα της μεθόδου σε σχέση με τις υπόλοιπες μεθόδους που αναλύθηκαν παραπάνω, είναι το ότι καθώς δεν χρησιμοποιεί τη σειρά των γονιδίων, μπορεί να εφαρμοσθεί με ακριβώς τον ίδιο τρόπο σε κάθε είδους ζευγάρια ή ομάδες οργανισμών ανεξάρτητα τόσο της εξελικτικής τους απόστασης όσο και του αριθμού χρωμοσωμάτων τους. Μπορεί με άλλα λόγια, να χρησιμοποιηθεί για τη σύγκριση του ανθρώπου με ένα βακτήριο και να δώσει χρήσιμα συμπεράσματα σε αντίθεση με τις προηγούμενες μεθόδους οι οποίες αποδίδουν καλύτερα και πρέπει να χρησιμοποιούνται κυρίως σε συγγενικούς οργανισμούς (και κατά βάση, σε βακτήρια).

Σύνοψη

- Χρήση μόνο ομοιότητας ακολουθιών
- 88 Fusion events σε 3 γονιδιώματα
- Αναγνωρίζονται πολλές μακρινές (στο γονιδίωμα) αλληλεπιδράσεις
- Ορισμένες περιοχές στο γονιδίωμα έχουν μεγάλη τάση για fusions

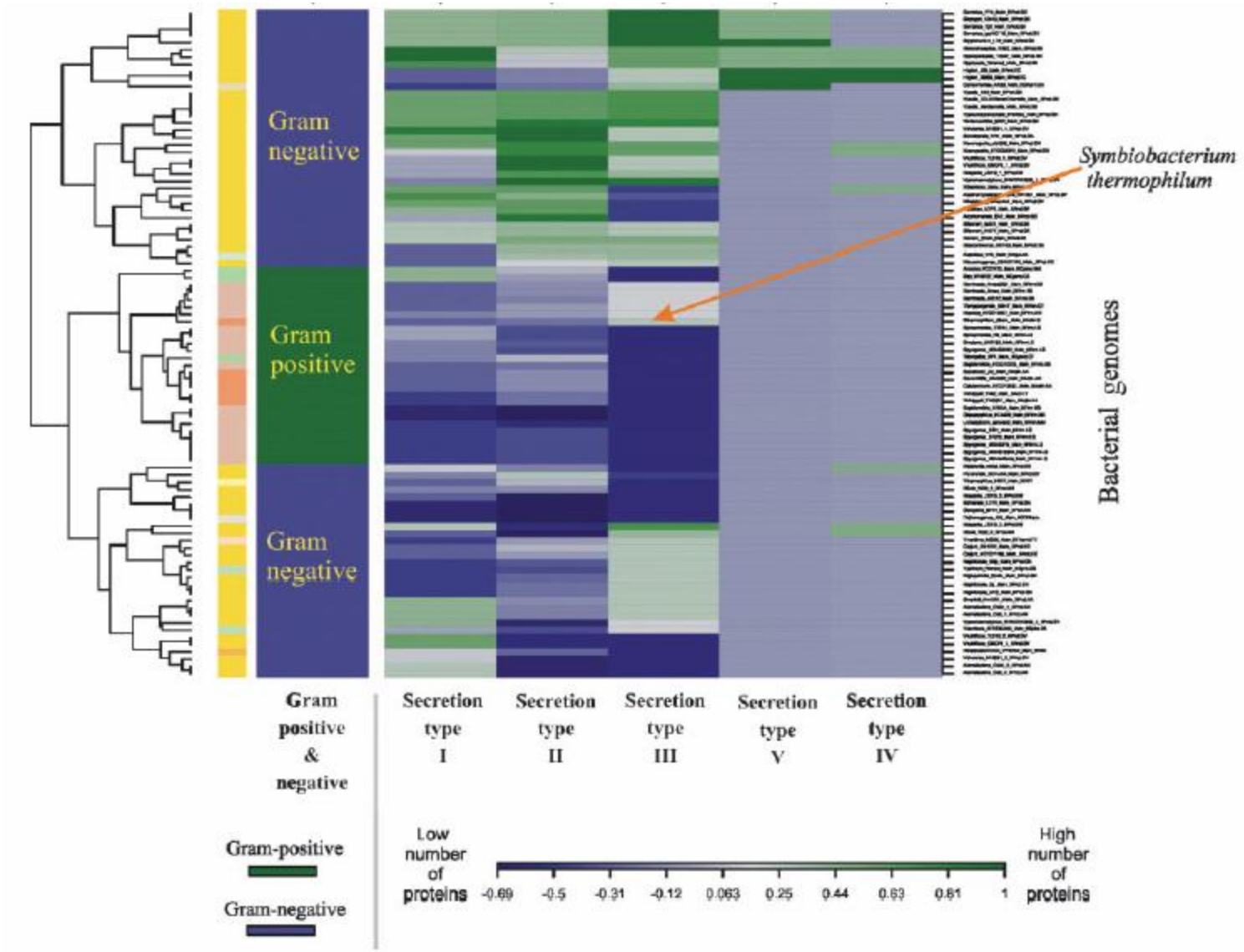


Fig. 1. Two-dimensional clustering (Willenbrock *et al.*, 2005) of bacterial genome sequences versus secretion systems type I–V. Dark blue indicates that a low number of the selected proteins is present for the specific secretion type; dark green represents cases where we find that most of the proteins for a given secretion system are present. It should be noted that data within each column are normalized around the centre using minimum and maximum values.

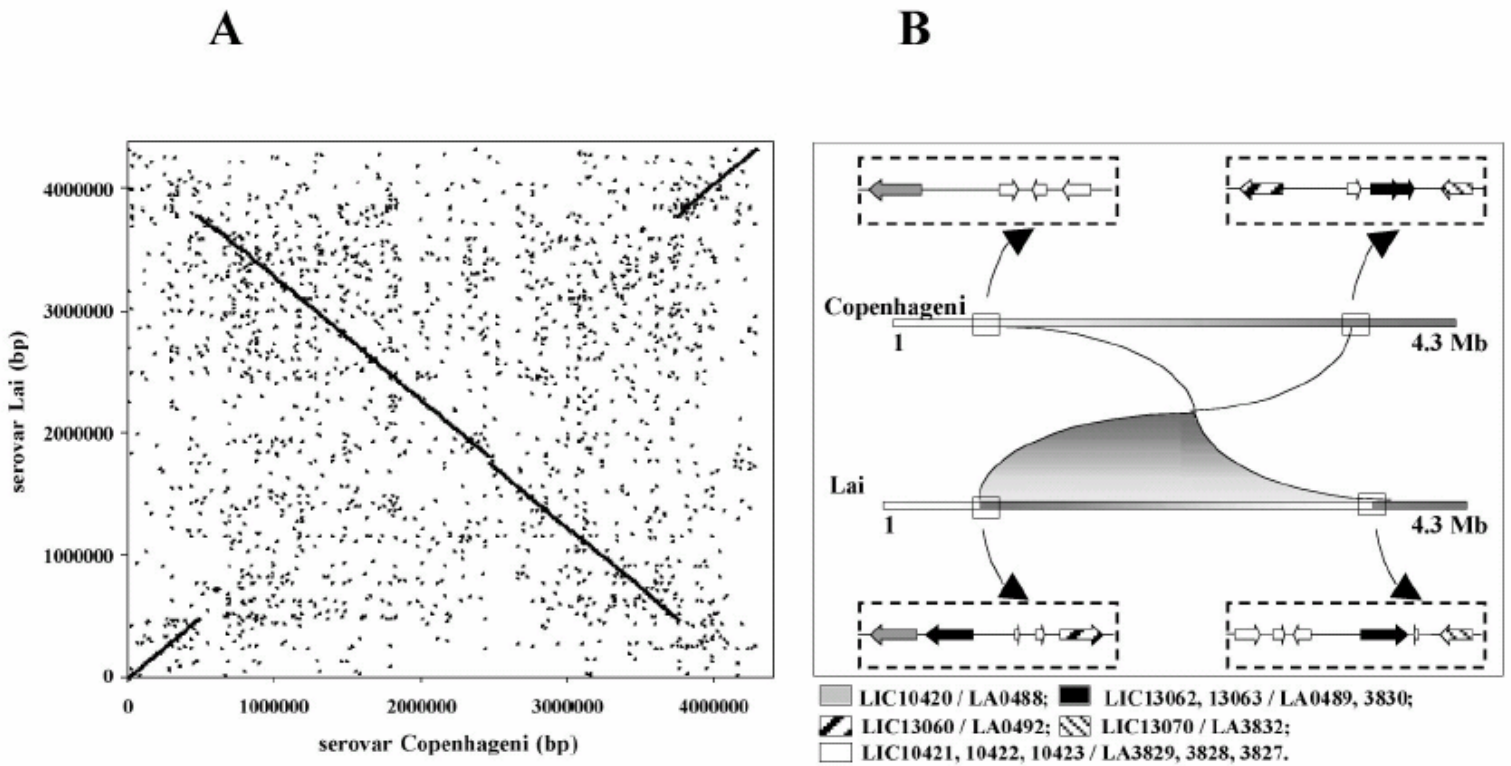


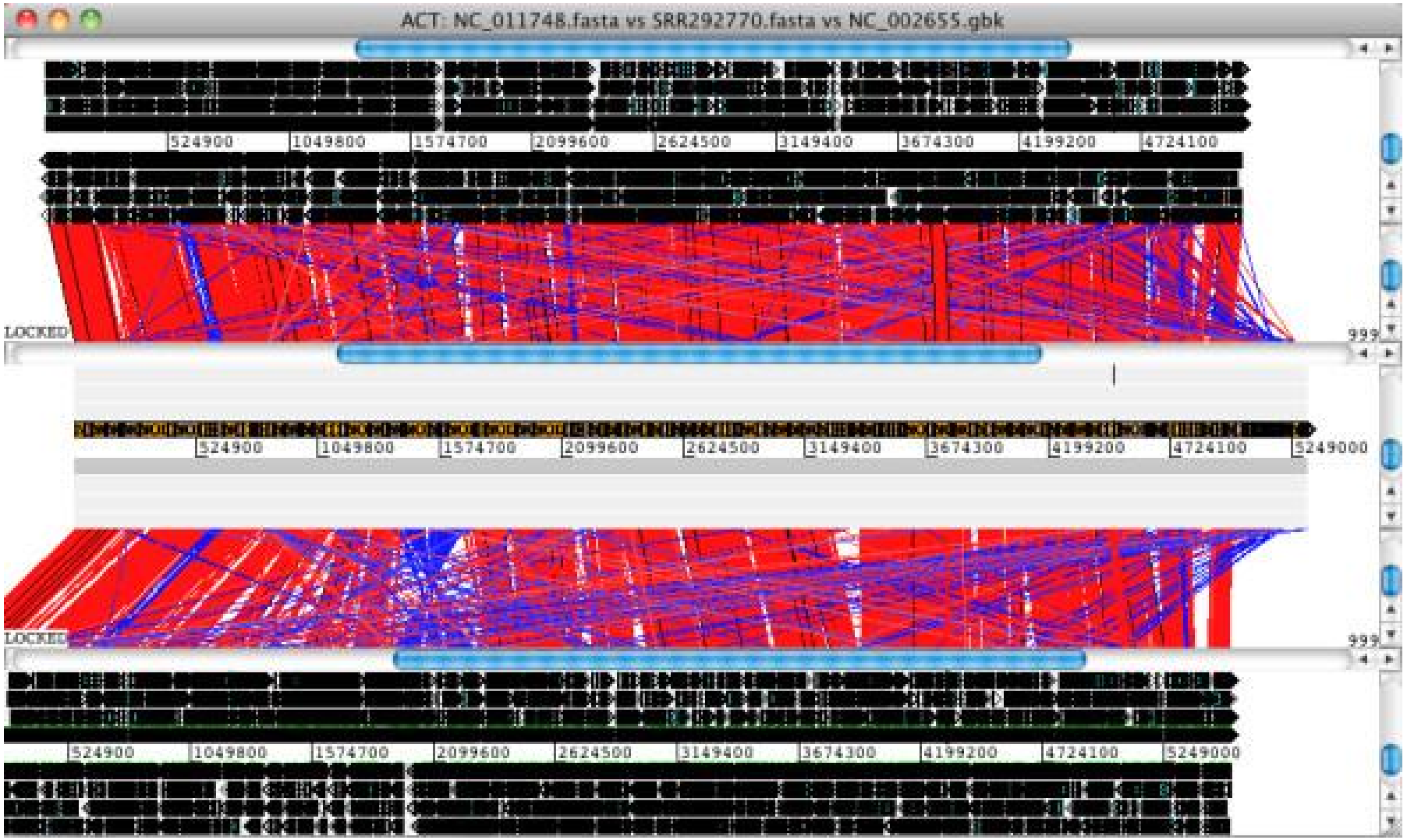
FIG. 1. Inversion of *L. interrogans* serovars Copenhageni and Lai CI chromosomes. (A) Nucleotide alignment obtained by using MUMmer, which relies on exact matches of at least 20 bp. Each dot in the figure is one such match. The dark lines on the two main diagonals result from the high density of points with sequence identity along chromosome I of the two serovars. The scattered points outside the main diagonals represent other short regions of sequence identity. (B) Scheme showing predicted genes flanking the inversion breakpoints. Pairs of ortholog genes have the same pattern code. The black arrows represent IS elements.

Λογισμικό

- **ACT** (https://www.sanger.ac.uk/resources/software/act/)
- **MAUVE** (http://darlinglab.org/mauve/mauve.html)
- **EDGAR** (http://edgar.cebitec.uni-bielefeld.de)
- **CGAT** (http://mbgd.genome.ad.jp/CGAT/)
- **BRIG** (BLAST Ring Image Generator, http://sourceforge.net/projects/brig/)
- **VISTA** (http://genome.lbl.gov/vista/index.shtml)

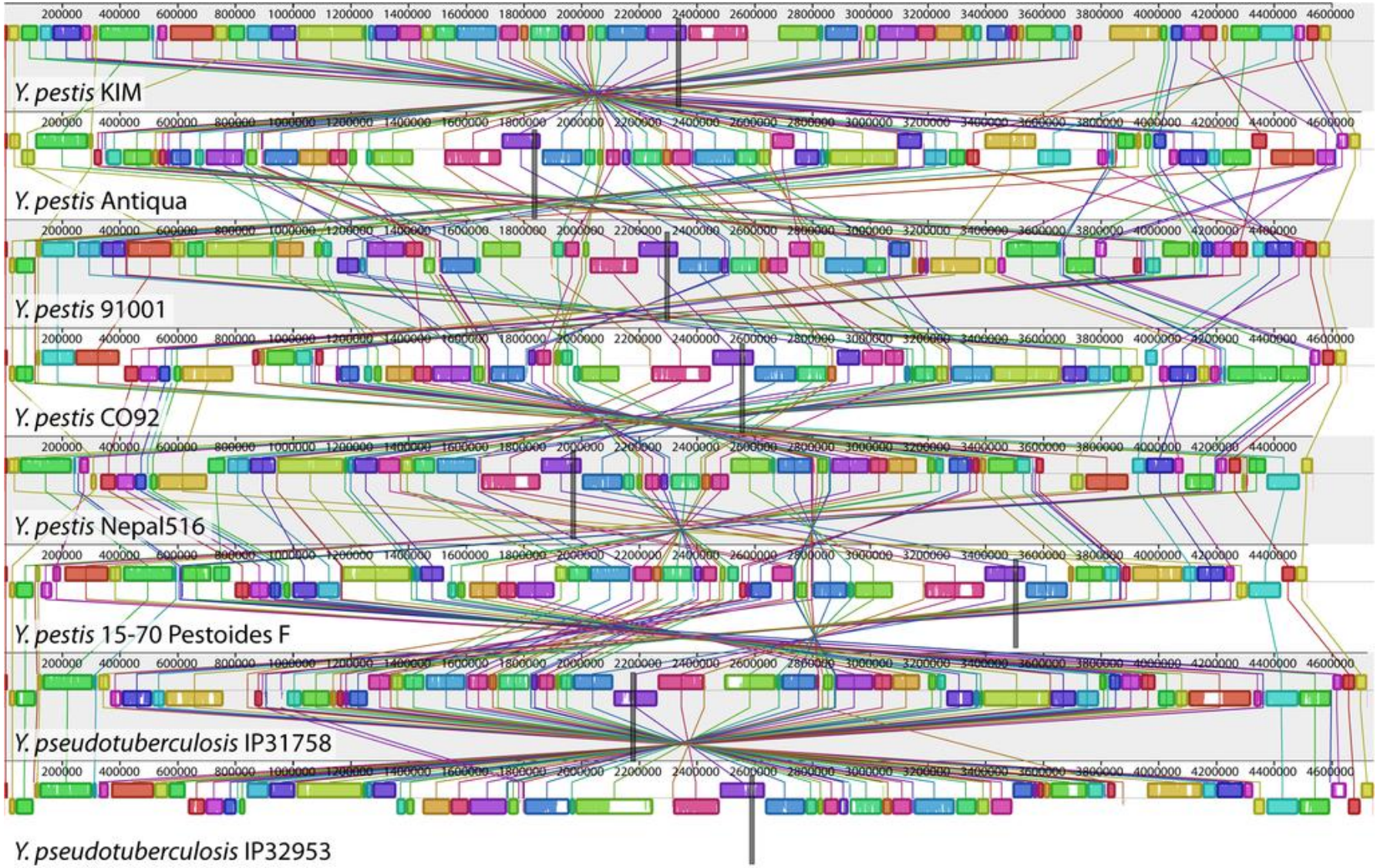
ACT

- Το **ACT** (https://www.sanger.ac.uk/resources/software/act/) είναι ένα εργαλείο βασισμένο στη Java το οποίο επιτρέπει την οπτικοποίηση γονιδιωμάτων και τη σύγκρισή τους. Για τη στοίχιση των αλληλουχιών χρησιμοποιεί το BLAST. Κατόπιν τα δύο γονιδιώματα και το αποτέλεσμα από την αναζήτηση του BLAST εισάγονται στο ACT για οπιτικοποίηση της σύγκρισης. Επιπλέον, το εργαλείο μπορεί να οπτικοποιήσει ταυτόχρονα περισσότερες από μία συγκρίσεις γονδιωμάτων. Οι ομόλογες περιοχές οι οποίες βρίσκονται στην ίδια κατεύθυνση στο γονιδίωμα χρωματίζονται με κόκκινο ενώ αυτές που βρίσκονται σε αντίθετες κατευθύνσεις, με μπλε. Η ένταση του χρωματισμού αντικατοπτρίζει το επίπεδο ομοιότητας. Τα πλεονεκτήματα του ACT περιλαμβάνουν τη δυνατότητα να απεικονίζει τη στοίχιση σε διαφορετικές μεγεθύνσεις (zoom in – zoom out) έτσι ώστε να μπορεί να απεικονίσει είτε τη στοίχιση ολόκληρου του γονιδιώματος, είτε να εστιάσει σε συγκεκριμένα γονίδια ενδιαφέροντος, αλλά και τη δυνατότητα που προσφέρει στο χρήστη να προσθέσει δικό του σχολιασμό για τα γονιδιώματα που αναλύονται (Carver et al., 2005).


ACT: NC_011748.fasta vs SRR292770.fasta vs NC_002655.gbk
524900
1049800
1574700
2099600
2624500
3149400
3674300
4199200
4724100
LOCKED
999
524900
1049800
1574700
2099600
2624500
3149400
3674300
4199200
4724100
5249000
LOCKED
999
524900
1049800
1574700
2099600
2624500
3149400
3674300
4199200
4724100
5249000

MAUVE

- Το **MAUVE** (http://darlinglab.org/mauve/mauve.html) είναι επίσης ένα εργαλείο βασισμένο στη Java κατάλληλο για συγκρίσεις γονιδιωμάτων. Διαθέτει ενσωματωμένο σύστημα απεικόνισης αλλά και τη δυνατότητα να εξάγει την πληροφορία από τη σύγκριση των γονιδιωμάτων σε διάφορες μορφές. Το MAUVE μπορεί να εργαστεί με δεδομένα αλληλούχισης νέας γενιάς, και έτσι παρέχει τη δυνατότητα να τοποθετήσει και να διατάξει μια σειρά από contigs απέναντι σε ένα ολόκληρο γονιδίωμα. Το εργαλείο δέχεται σαν είσοδο τις τελικές μορφές των γονιδιωμάτων και δημιουργεί μια στοίχιση αυτών. Αναγνωρίζει περιοχές με μεγάλη ομολογία και αναθέτει ένα ξεχωριστό χρώμα σε κάθε μία. Κατόπιν, κάθε γονιδίωμα απεικονίζεται σαν μια ακολουθία τέτοιων χρωματιστών περιοχών. Με τον τρόπο αυτό, γίνεται εύκολος ο εντοπισμός περιοχών με μοναδικά γονίδια. Επίσης, το MAUVE μπορεί χρησιμοποιηθεί (καθώς δουλεύει όπως αναφέραμε και με δεδομένα αλληλούχισης νέας γενιάς) και για τον εντοπισμό νουκλεοτιδικών πολυμορφισμών (SNPs) οι οποίοι μπορούν να χρησιμοποιηθούν παρακάτω για φυλογενετικές, εξελικτικές ή ιατρικές αναλύσεις (Darling, Mau, & Perna, 2010).


Y. pestis KIM
Y. pestis Antiqua
Y. pestis 91001
Y. pestis CO92
Y. pestis Nepal516
Y. pestis 15-70 Pestoides F
Y. pseudotuberculosis IP31758
Y. pseudotuberculosis IP32953

EDGAR

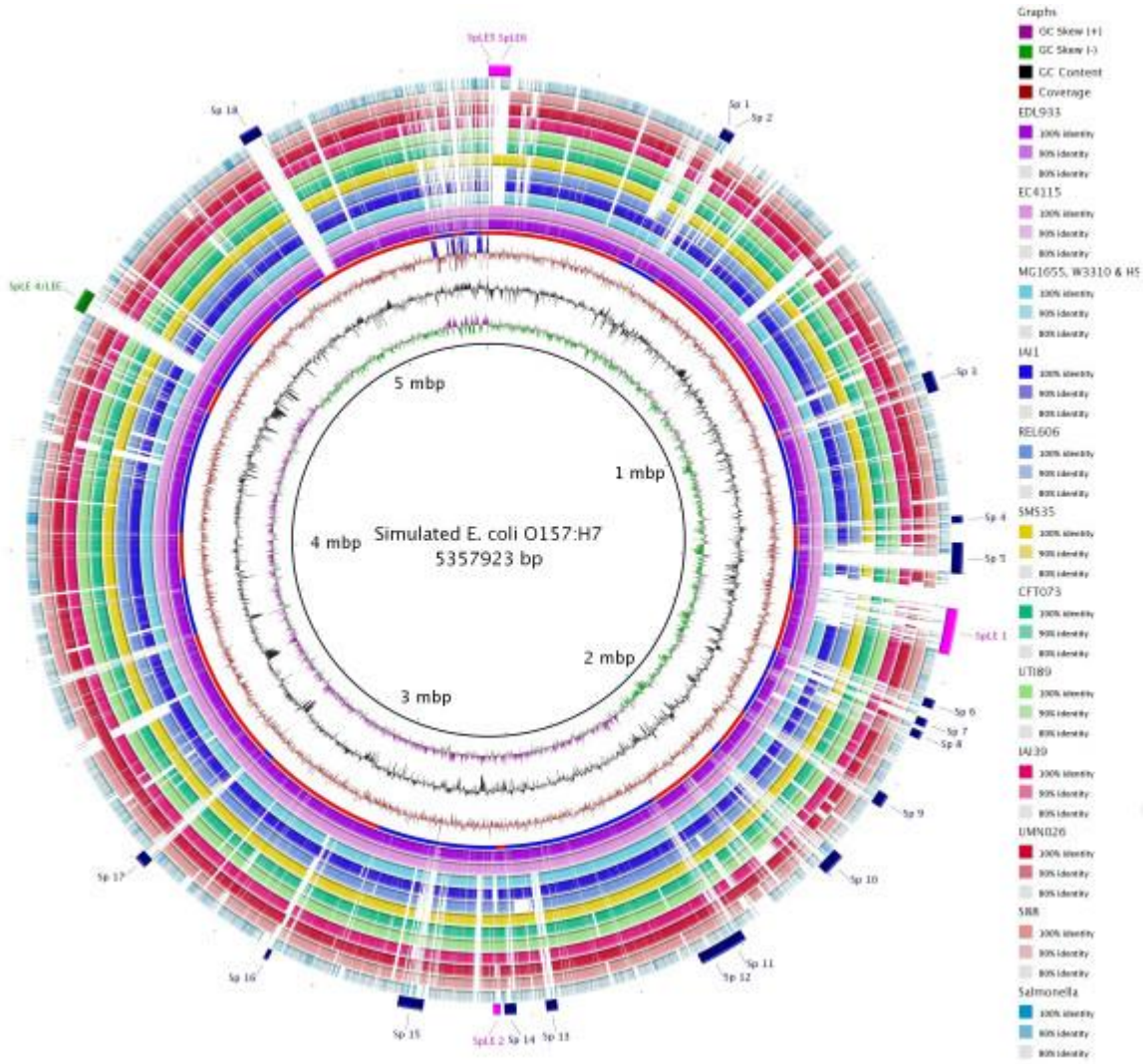
- Το **EDGAR** (http://edgar.cebitec.uni-bielefeld.de) είναι ένα ακόμα σύγχρονο διαδικτυακό εργαλείο συγκριτικής γονιδωματικής το οποίο μπορεί να δεχτεί και δεδομένα αλληλούχισης. Το EDGAR είναι σχεδιασμένο έτσι ώστε να διευκολύνει το χρήστη και να απλοποιεί τις διαδικασίες. Ενσωματώνει τις βάσεις δεδομένων του NCBI και έχει στη βάση δεδομένων του όλα τα αποτελέσματα γνωστών γονιδιωμάτων προ-υπολογισμένα, ενώ έχει και τη δυνατότητα να απεικονίσει εξελικτικές και φυλογενετικές σχέσεις οι οποίες πολλές φορές διαλευκάνουν υποθέσεις σύγκρισης γονιδιωμάτων. Επίσης, υποστηρίζει μια σειρά από τρόπους απεικόνισης των αποτελεσμάτων όπως τα διαγράμματα στοίχισης γονιδιωμάτων (synteny plots) και διαγράμματα Venn για τα κοινά γονίδια (Blom, et al., 2009).

CGAT

- Το **CGAT** (http://mbgd.genome.ad.jp/CGAT/) είναι ένα ακόμα παρόμοιο εργαλείο που δημιουργήθηκε για να διευκολύνει τις συγκρίσεις συγγενικών βακτηριακών γονιδιωμάτων. Το CGAT λειτουργεί με αρχιτεκτονική client-server, στην οποία ο client AlignmentViewer (μια εφαρμογή Java) συνεργάζεται με τον DataServer (προγράμματα Perl). Το εργαλείο οπτικοποιεί στοιχίσεις γονιδιωμάτων τόσο στη μορφή των διαγραμμάτων σημείων όσο και στη μορφή των στοιχίσεων. Ο χρήστης μπορεί να προσθέσει πληροφορία στη σύγκριση, όπως για παράδειγμα την ύπαρξη επαναληπτικών αλληλουχιών και αλλαγές στη συχνότητα κωδικονίων έτσι ώστε να διευκολυνθεί στην εξαγωγή συμπερασμάτων. Εκτός από την οπτικοποίηση, ένα πλεονέκτημα του CGAT είναι η ευελιξία του καθώς επιτρέπει τη χρήση πολλών διαφορετικών αλγόριθμων στοίχισης γονιδιωμάτων (Uchiyama, Higuchi, & Kobayashi, 2006).

BRIG

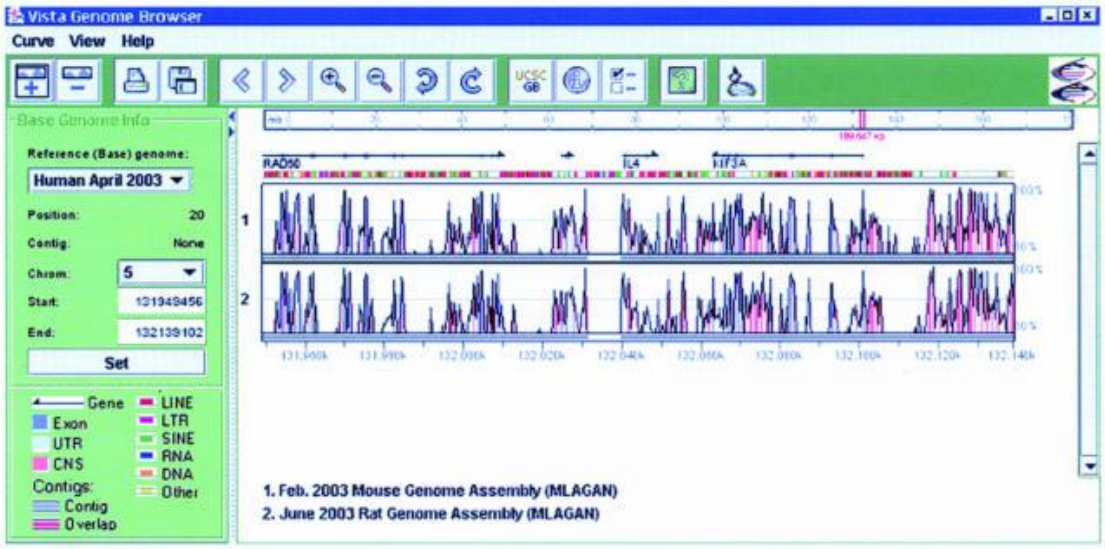
- Το **BRIG** (BLAST Ring Image Generator, http://sourceforge.net/projects/brig/) είναι ένα άλλο εργαλείο βασισμένο στη Java, το οποίο οπτικοποιεί τη σύγκριση ενός γονιδιώματος αναφοράς με μία ή περισσότερες άλλες αλληλουχίες. Χρησιμοποιεί έναν ιδιαίτερο τρόπο οπτικοποίησης, σύμφωνα με τον οποίο τα γονιδιώματα αναπαρίστανται ως σειρές από επάλληλους κύκλους (δαχτυλίδια), με ειδικό χρωματισμό, για να δηλώνει την παρουσία μιας περιοχής ή ενός γονιδίου στο γονιδίωμα αναφοράς. Το BRIG είναι αρκετά ευέλικτο και μπορεί να χρησιμοποιηθεί για να απαντήσει πλήθος ερωτημάτων, ανάλογα με την επιλογή των γονιδιωμάτων υπό σύγκριση. Αυτό που πρέπει να τονιστεί είναι το γεγονός ότι η αναπαράσταση είναι εξαρτώμενη από το γονιδίωμα αναφοράς. Με άλλα λόγια, ενώ το εργαλείο απεικονίζει ποιες περιοχές είναι παρούσες ή απούσες από τα γονιδιώματα σύγκρισης, δεν μπορεί να δείξει περιοχές των γονιδιωμάτων αυτών που λείπουν από το γονιδίωμα αναφοράς. Γι' αυτό το λόγο η επιλογή του γονιδιώματος αναφοράς είναι ιδιαίτερα σημαντική (Alikhan, Petty, Ben Zakour, & Beatson, 2011).


Simulated E. coli O157:H7
5357923 bp
1 mbp
2 mbp
3 mbp
4 mbp
5 mbp
Graphs
GC Skew (+)
GC Skew (-)
GC Content
Coverage
EDL933
EC4115
MG1655, W3310 & H5
IAI1
REL606
SMS35
CFT073
UTI89
IAI39
UMN026
SBR
Salmonella

VISTA

- Το **VISTA** (http://genome.lbl.gov/vista/index.shtml) ήταν ένα από τα πρώτα εργαλεία οπτικοποίησης στοιχίσεων γονιδιωμάτων και είχε παρουσιαστεί το 2000. Σήμερα, έχει εξελιχθεί σε μια ολοκληρωμένη σουίτα προγραμμάτων τα οποία καλύπτουν κάθε ανάγκη συγκριτικής ανάλυσης γονιδιωμάτων. Διαθέτει ειδικά εργαλεία για διάφορες συγκριτικές αναλύσεις γονιδιωμάτων, διασύνδεση με τις βάσεις δεδομένων γονιδιωμάτων, ενώ διαθέτει και αποθηκευμένα προ-υπολογισμένα αποτελέσματα για τα γνωστά γονιδιώματα (ακόμα και των σπονδυλοτών). Διαθέτει ειδικό σύστημα οπτικοποίησης (VISTA Browser) το οποίο επιτρέπει στο χρήστη να υποβάλει και το δικό του γονιδίωμα για ανάλυση στους διάφορους εξυπηρετητές (VISTA servers, rVista, mVISTA, phyloVISTA, gVISTA κ.ο.κ.) στους οποίους ο χρήστης μπορεί να επιτελέσει στοιχίσεις με διαφορετικούς αλγόριθμους, οπτικοποίηση με διαφορετικούς τρόπους, αλλά και ενσωμάτωση διαφορετικών ειδών πληροφορίας όπως φυλογενετικές σχέσεις, ρυθμιστικές περιοχές κ.ο.κ. (Frazer, Pachter, Poliakov, Rubin, & Dubchak, 2004). Μια επιπλέον δυνατότητα του VISTA είναι το ότι διαθέτει και μια standalone εφαρμογή με σχεδόν τις ίδιες δυνατότητες, το GenomeVISTA, το οποίο μπορεί να εγκατασταθεί ελεύθερα στον υπολογιστή του χρήστη και να εκτελέσει εκεί τις ίδιες λειτουργίες με τη διαδικτυακή εκδοχή thus provides, προσφέροντας μεγαλύτερη ασφάλεια των δεδομένων και ίσως και ταχύτητα (Poliakov, Foong, Brudno, & Dubchak, 2014).

(a)



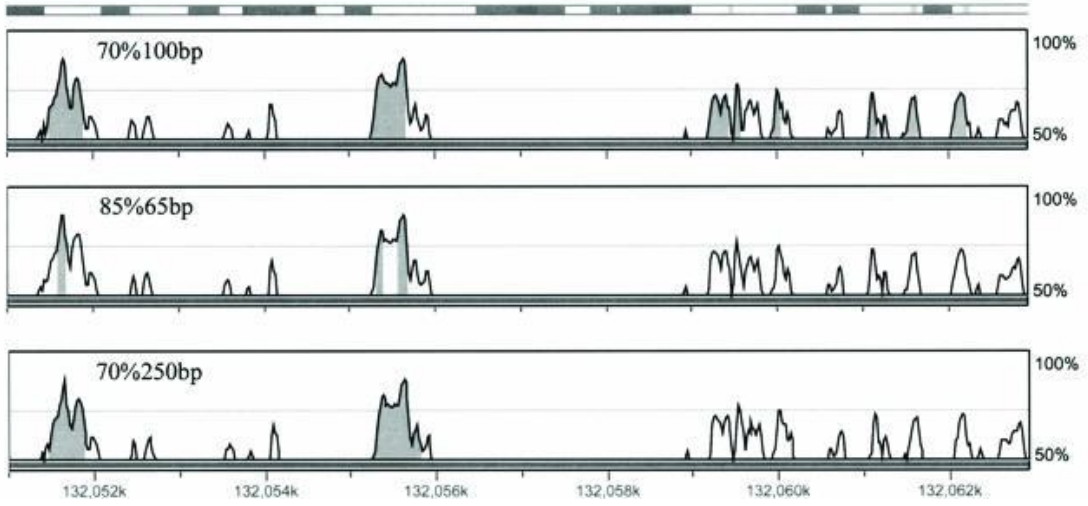
(b)

```
Criteria: 70% identity over 100 bp

*************** Conserved Regions - Human (Mouse) ***************

131952851  (54292441)  to 131953108  (54292210)  =     258bp  at  69.4%  noncoding
131954117  (54291314)  to 131954245  (54291186)  =     129bp  at  89.9%  exon
131954246  (54291185)  to 131954339  (54291091)  =      98bp  at  71.4%  noncoding
131954479  (54290969)  to 131954644  (54290804)  =     166bp  at  87.3%  exon
131954759  (54289473)  to 131954891  (54289341)  =     135bp  at  71.1%  noncoding
131955242  (54288804)  to 131955435  (54288611)  =     194bp  at  89.7%  exon
131956186  (54288222)  to 131956392  (54288016)  =     207bp  at  73.4%  exon
131957525  (54284506)  to 131957654  (54284379)  =     130bp  at  70.0%  noncoding
131957779  (54284180)  to 131957961  (54283998)  =     183bp  at  85.2%  exon
```

(c)



- Όπως είδαμε, τα περισσότερα από τα προαναφερθέντα πακέτα λογισμικού παρέχουν τη δυνατότητα χρήσης διαφορετικών αλγορίθμων στοίχισης γονιδιωμάτων. Κάποια, διαθέτουν και δικούς τους αλγόριθμους στοίχισης αλλά τα περισσότερα δίνουν τα δυνατότητα ενσωμάτωσης και άλλων εξειδικευμένων αλγορίθμων.
- Οι πιο γνωστοί από αυτούς είναι
  - το **MUMMER** (http://mummer.sourceforge.net/),
  - το **MEGA-BLAST** (http://www.ncbi.nlm.nih.gov/BLAST/),
  - το **LAGAN** (http://bioperl.org/wiki/LAGAN) και
  - το **MGA** (http://bibiserv.techfak.uni-bielefeld.de/mga/).

- Όσον αφορά τους αλγόριθμους εύρεσης σύντηξης γονιδίων, μέθοδος η οποία είναι η πιο «απόμακρη» (ή ξεχωριστή) από τις υπόλοιπες, υπάρχουν επίσης μια σειρά από επιλογές οι οποίες έχουν πολλαπλασιαστεί ιδιαίτερα τα τελευταία χρόνια με την έλευση της αλληλούχισης νέας γενιάς με τη χρήση τέτοιων τεχνικών σε διάφορες άλλες εφαρμογές, ακόμα και ιατρικές (Carrara et al., 2013). Ενδεικτικά, αναφέρουμε
  - τον αρχικό αλγόριθμο των Ouzounis και συνεργατών, το GeneRAGE (Enright & Ouzounis, 2000), αλλά και μερικές νεότερες εφαρμογές όπως
  - το FusionMap (http://www.omicsoft.com/fusionmap) (Ge et al., 2011)
  - και το MosaicFinder (http://sourceforge.net/projects/mosaicfinder) (Jachiet, Pogorelcnik, Berry, Lopez, & Bapteste, 2013).

Βιβλιογραφία

- Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA.Nature. **Protein interaction maps for complete genomes based on gene fusion events.**1999 Nov 4;402(6757):86-90.
- Alfonso Valencia and Florencio Pazos. **Computational methods for the prediction of protein interactions.** Current Opinion in Structural Biology 2002, 12:368–373
- Tsoka S, Ouzounis CA. **Recent developments and future directions in computational genomics**.FEBS Lett. 2000 Aug 25;480(1):42-8.