



Συγκριτική Γονιδιωματική

Αλληλούχιση γονιδιωμάτων και σχολιασμός

Ο αριθμός των γονιδιωμάτων που αλληλουχούνται αυξάνει με ταχείς ρυθμούς, **ΌΜΩΣ** ο σχολιασμός τους δεν μπορεί να ακολουθήσει αυτό το ρυθμό.

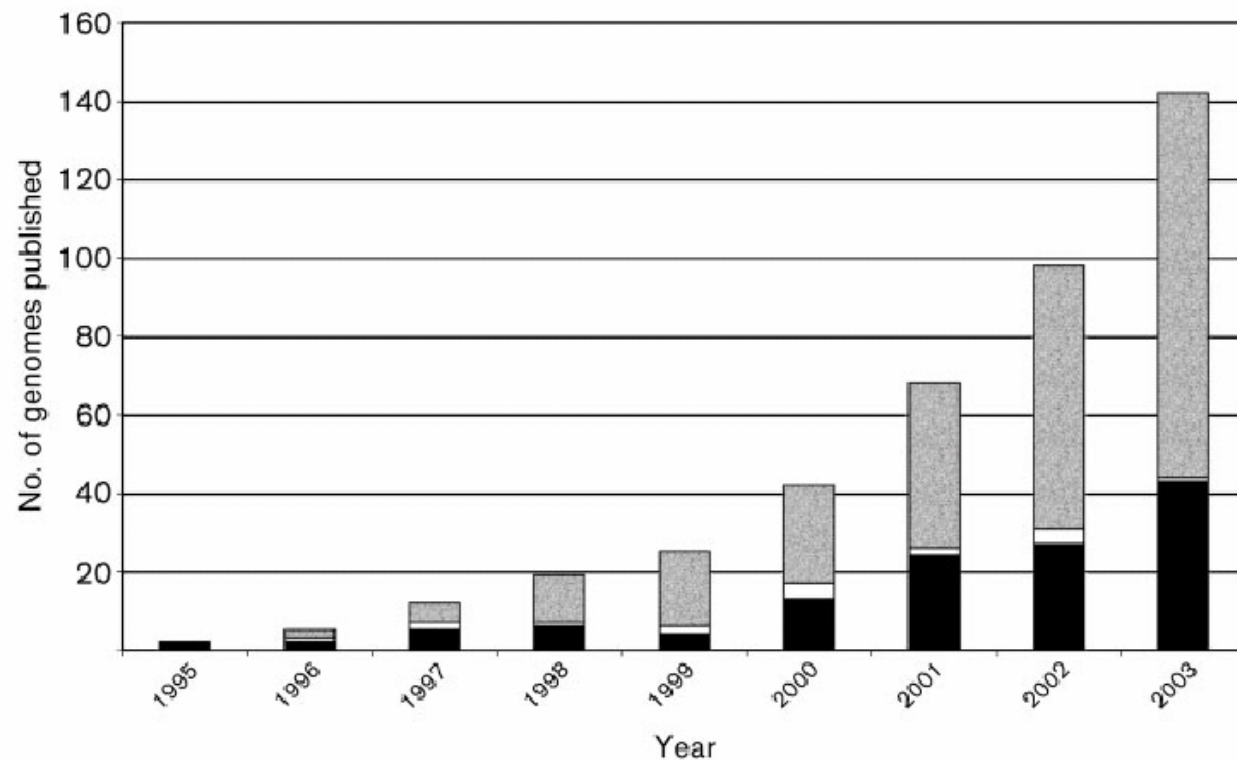


Fig. 1. Number of prokaryotic genomes sequenced each year since 1995. Black, bacterial genomes; white, archaeal genomes; grey, running total.

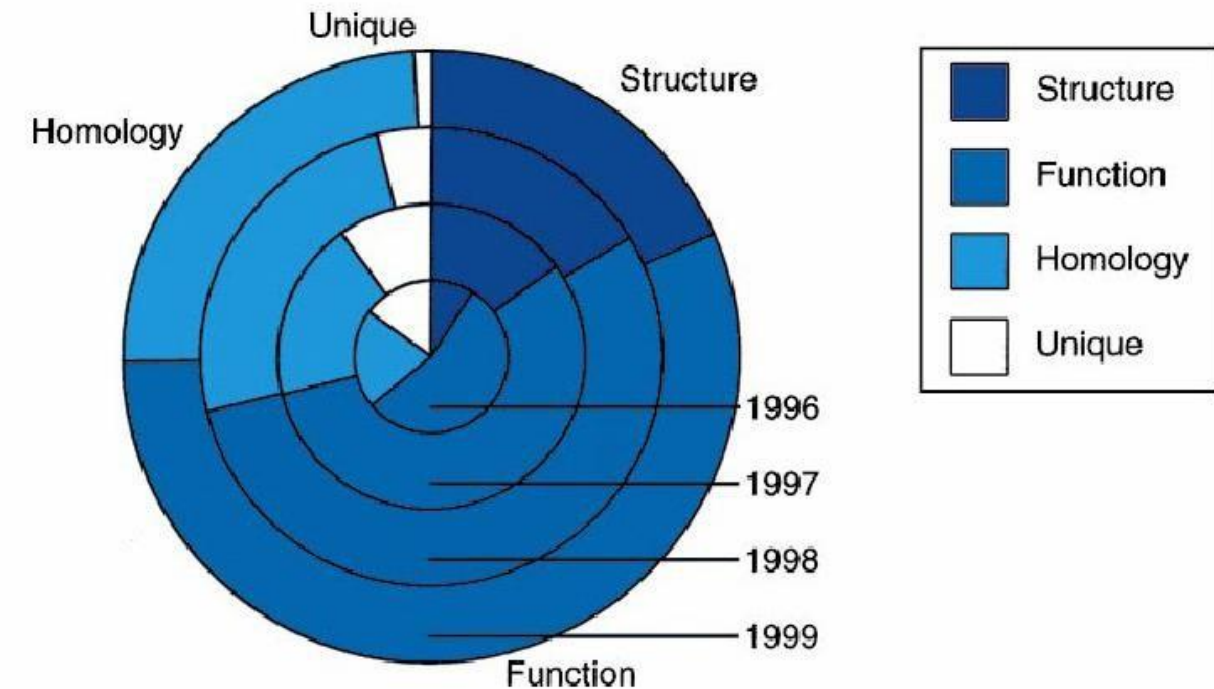


Fig. 2. The ‘function bottleneck’. Information clock illustrating the improvement of annotations identified for a given genome over the years 1996–1999. The four levels of annotation range from homologues of known structure (blue) and homologues of known function (marine) to homologues of unknown function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

Μέθοδοι Ανάλυσης

Για τον σχολιασμό των νεοπροσδιορισθέντων γονιδιωμάτων χρησιμοποιούνται οι ακόλουθες μέθοδοι:

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

Γονιδιωματική

Γονιδιωματική ονομάζουμε τον επιστημονικό κλάδο ο οποίος χρησιμοποιεί διαφορετικές τεχνικές της γενετικής, της μοριακής βιολογίας και της βιοπληροφορικής με **σκοπό**:

1. να βρει την **αλληλουχία**,
2. να κάνει την **συναρμολόγηση** και
3. 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).

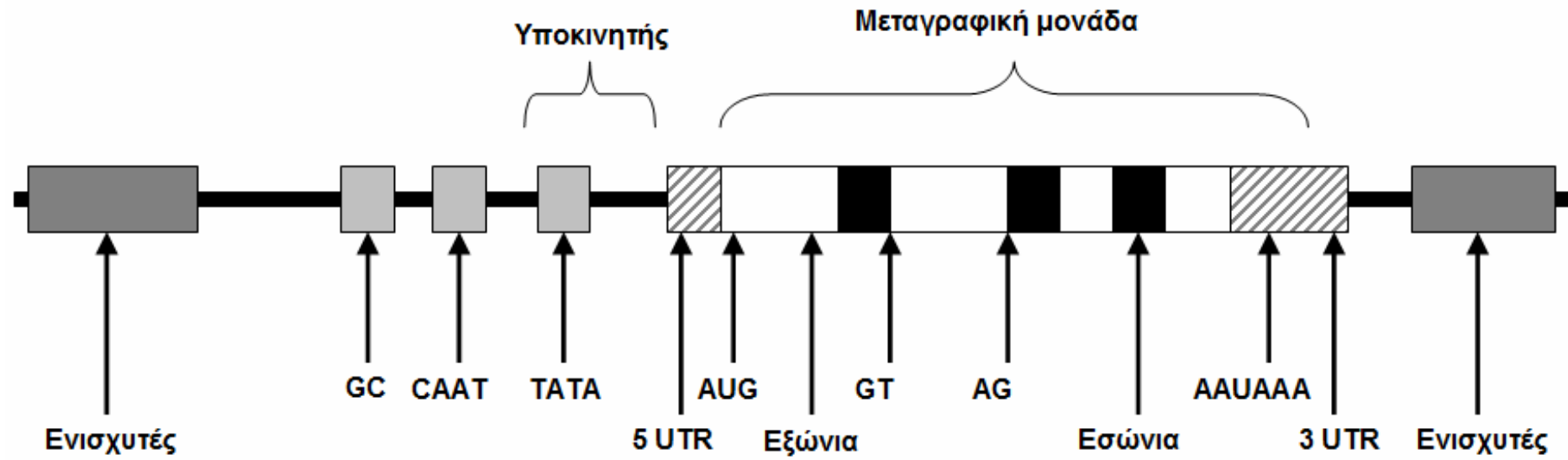
Αλγόριθμοι πρόγνωσης

- 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.

Gene finding



Gene finding

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](#)).

Οι **στατιστικές ιδιότητες** των νουκλεοτιδίων (ακόμα και στο πλαίσιο των αποδεκτών κωδικονίων) διαφέρουν ανάμεσα στις μεγάλες ομάδες οργανισμών. Έτσι, υπάρχουν εξειδικευμένα εργαλεία για ειδικές περιπτώσεις ή εργαλεία που λαμβάνουν υπόψη τους τη φυλογενετική προέλευση του οργανισμού.

Οι μεθοδολογίες που έχουν χρησιμοποιηθεί για τα προβλήματα αυτά, καλύπτουν ένα **μεγάλο εύρος**: από στατιστικές τεχνικές, weight matrices και προφίλ, 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 **δύο ή περισσότερων γονιδιωμάτων** (ή πρωτεωμάτων). Μπορεί να γίνει ακολουθώντας 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.

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

Προσπάθεια εντοπισμού των χαρακτηριστικών των γονιδιωμάτων των θερμόφιλων βακτηρίων που τα διαφοροποιούν σε σχέση με τους υπόλοιπους οργανισμούς.

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

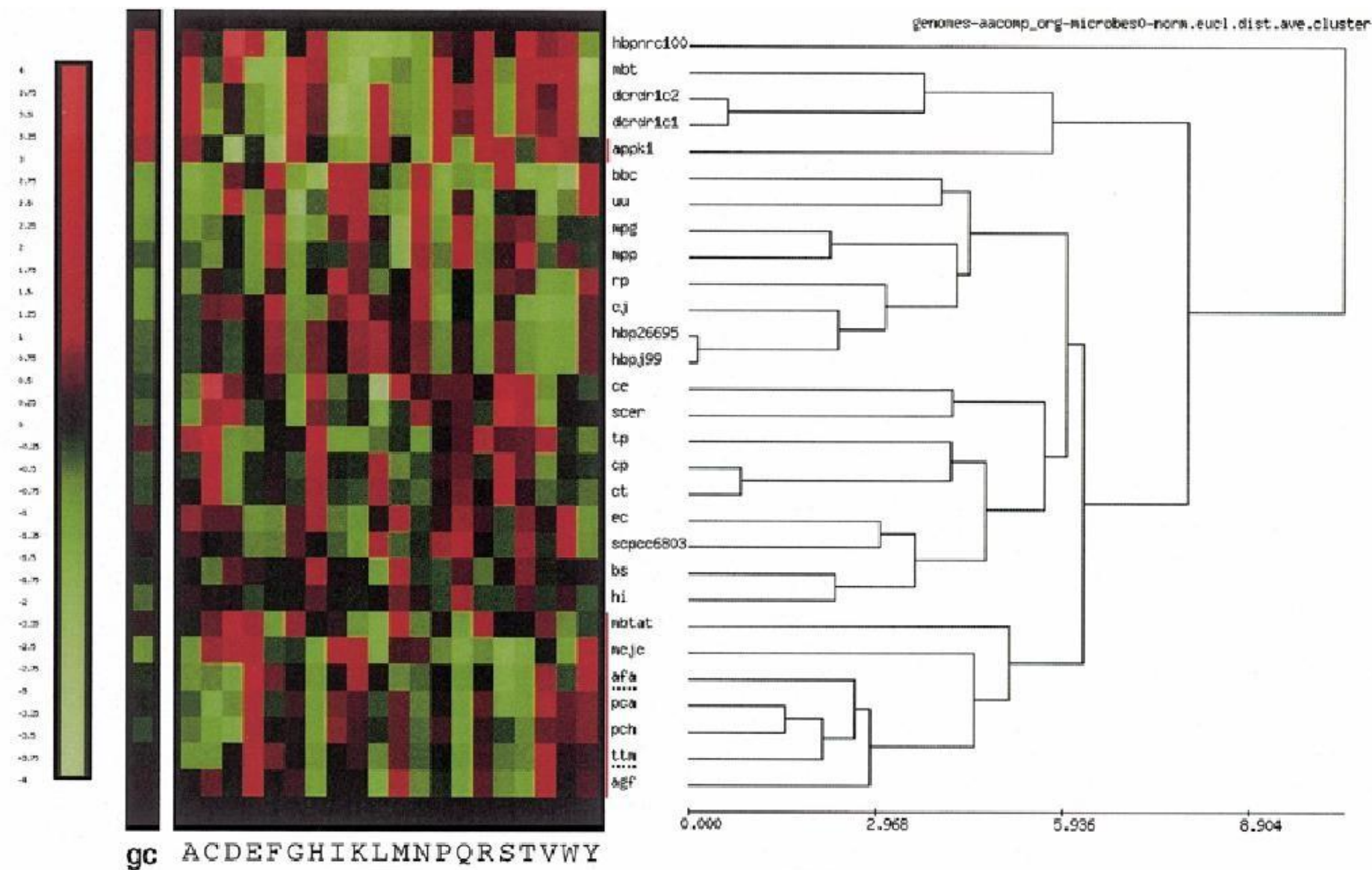


Figure 1. Standardised amino acid composition data of completely sequenced organisms grouped by hierarchical clustering. The GC ratios are shown for reference but were not used for the clustering process. Amino acids are abbreviated by the standard one letter code. The labels indicating the data sets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

Amino acid composition in principal axes

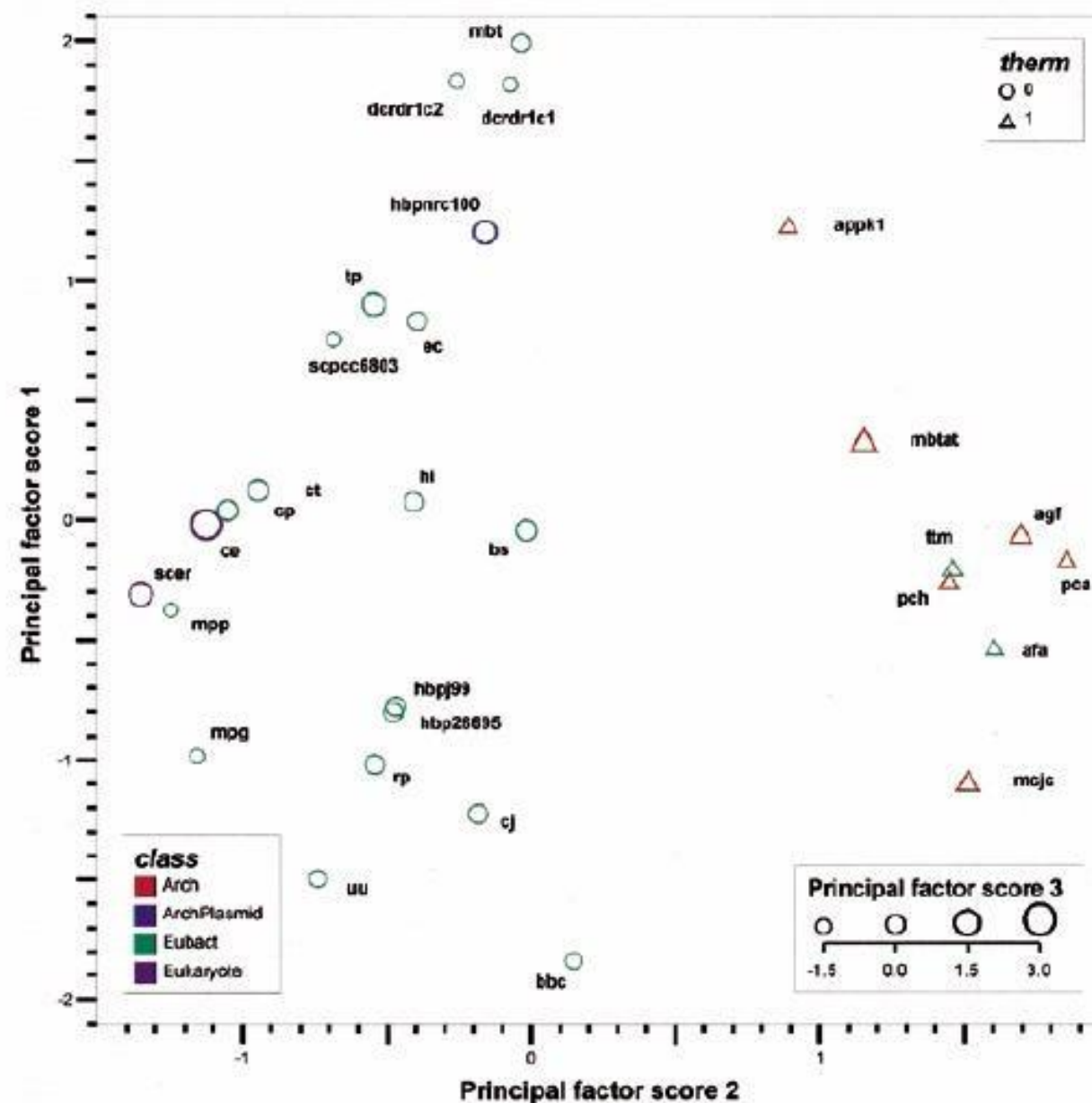


Figure 2. Reduced dimensionality plot showing the main principal components of the global amino acid compositions. The first principal axis (vertical) corresponds to GC ratio (see text). The second principal axis (horizontal) shows a clear separation of thermophiles and mesophiles, denoted by triangles and circles, respectively. The third principal component is depicted by symbol size (see insert for scale). Colour groups the sources into archaea (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)	$\sim 10^{-4}$	<i>t</i> -test
Glu (E)	80%	80%	$\sim 10^{-6}$	2.27 (0.40)	1.73 (0.31)	$\sim 10^{-4}$	<i>t</i> -test
Val (V)	50 to 65%	60%	$\sim 10^{-3}$	1.57 (0.42)	1.40 (0.38)	$\sim 2 \times 10^{-3}$	<i>t</i> -test
Thr (T)	−65%	60%	$\sim 10^{-3}$	−0.84 (0.25)	−1.31 (0.39)	$\sim 5 \times 10^{-3}$	<i>t</i> -test ^b
His (H)	−40 to −60%	−60%	$\sim 10^{-3}$	−0.44 (0.15)	−1.22 (0.42)	1% ^b	<i>t</i> -test ^b
Ser (S)	−30 to −60% ^b	−40%	1%	−1.11 (0.51)	−1.18 (0.54)	5%	<i>t</i> -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.

^aApproximate figures.

^bSee Appendix for discussion.

n/a, not applicable.

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

Διαχωρισμός των θερμοφίλων βακτηρίων από την αμινοξική σύσταση

- 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 θερμοφίλων αρχαιοβακτηρίων** (αρχαίων), **2 θερμοφίλων 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.

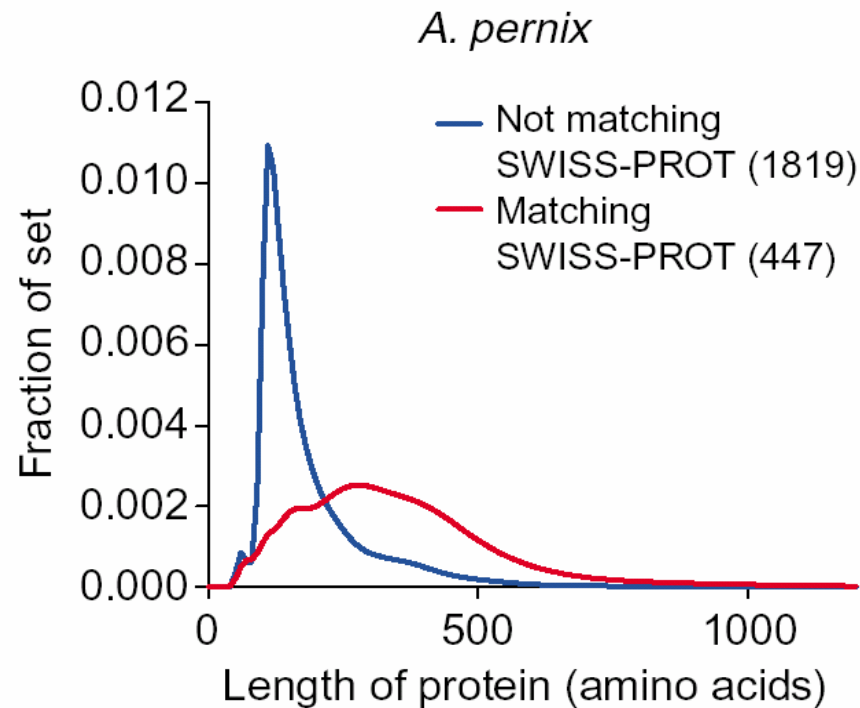
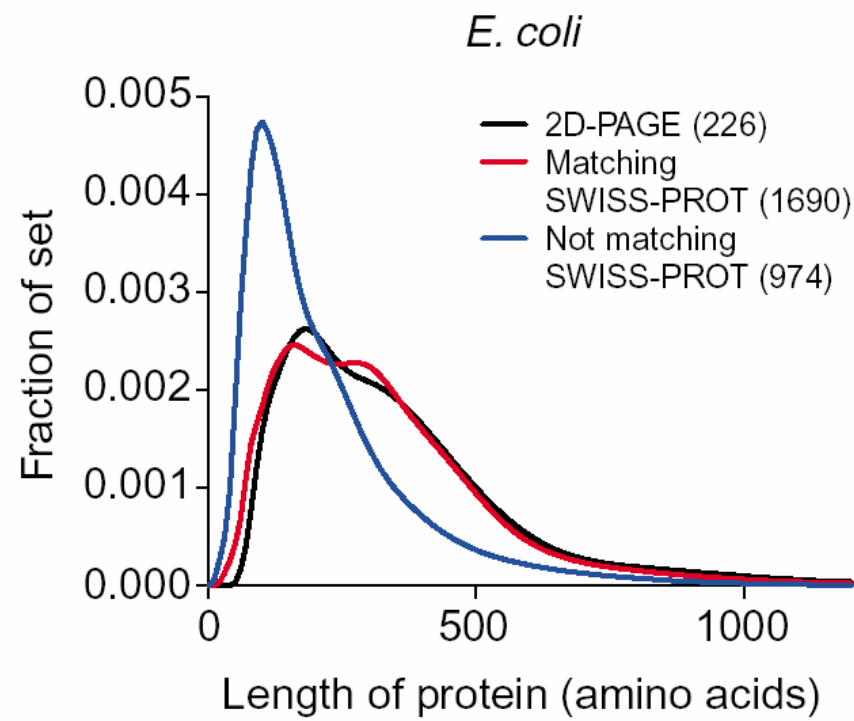
Το μήκος των «πραγματικών πρωτεϊνών» στα πλήρως προσδιορισμένα γονιδιώματα

- ✓ 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 γονιδιώματα

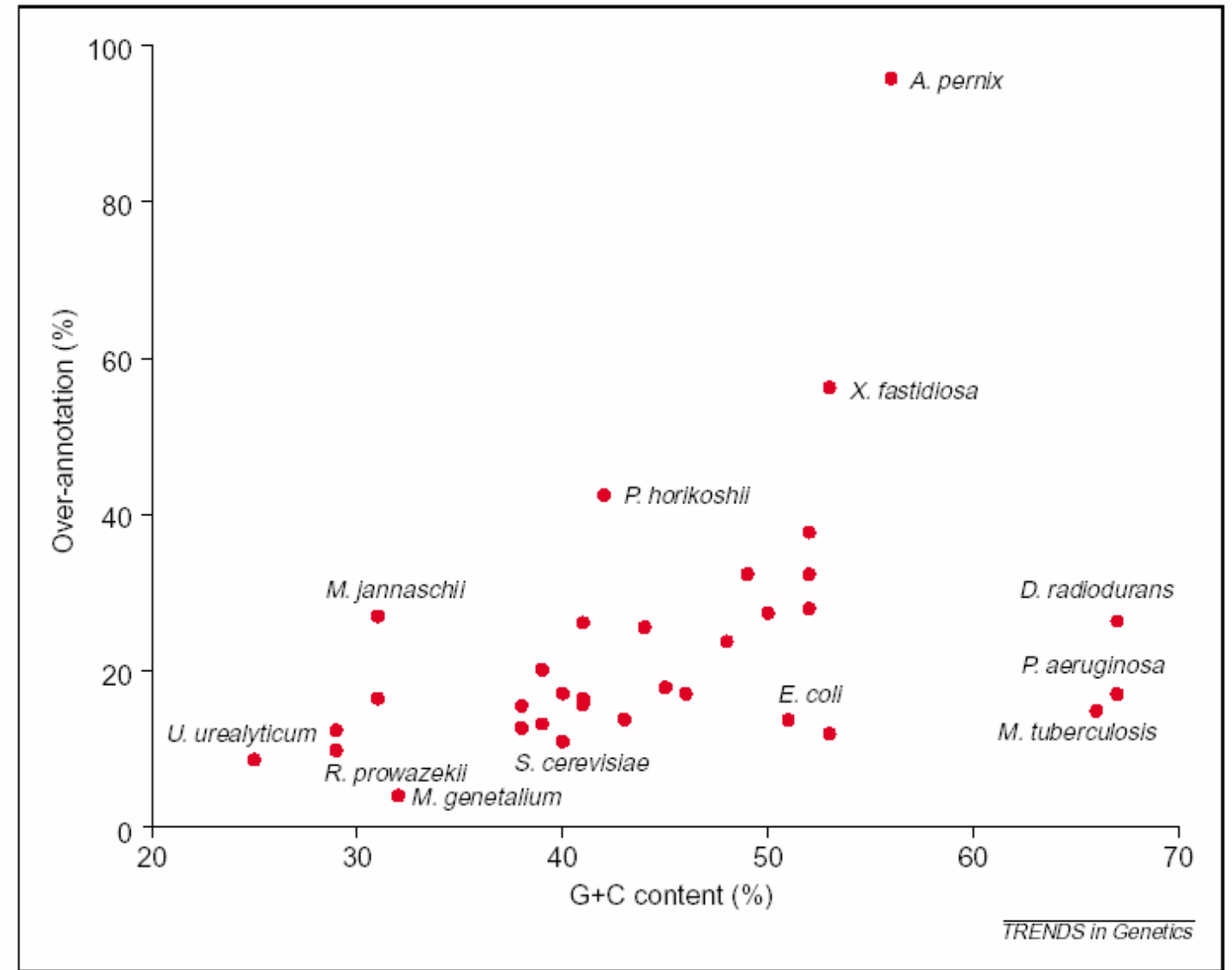
Μεθοδολογία:

- 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



TRENDS in Genetics



TRENDS in Genetics

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

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

GC% content και διαμεμβρανικές πρωτεΐνες

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

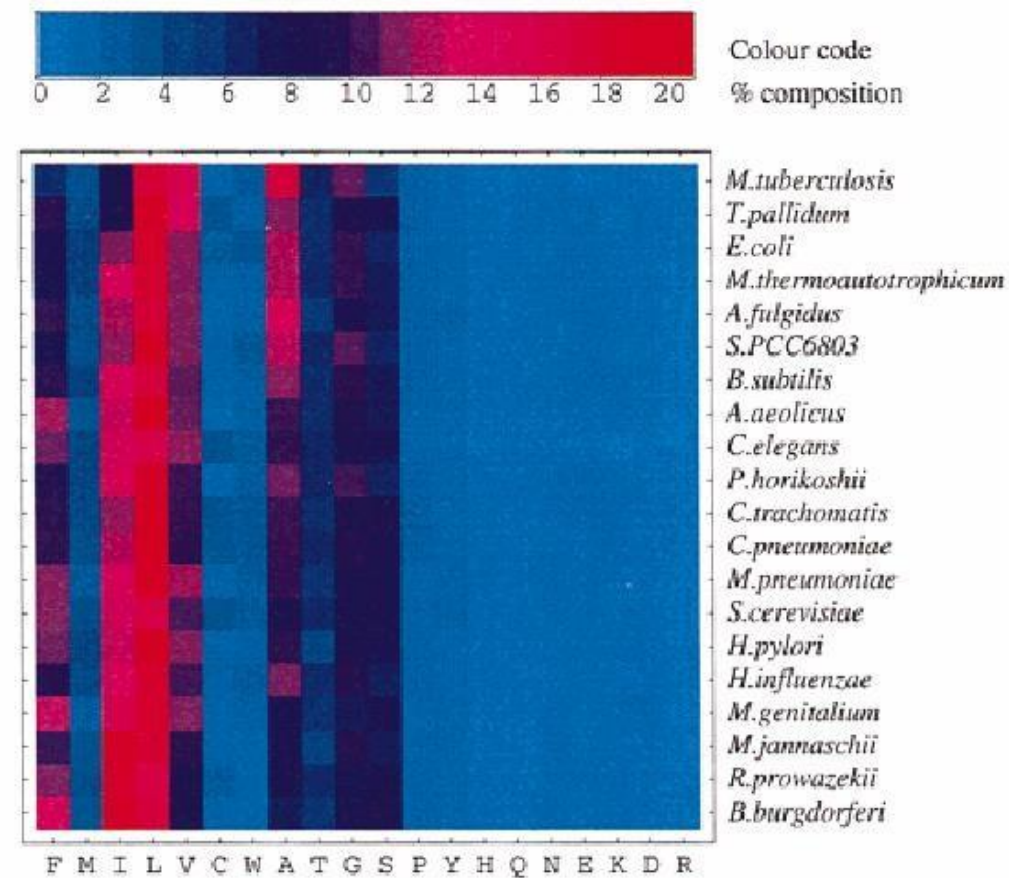


Fig. 1. A plot to show the amino acid composition of the TM domains in each of the proteomes under investigation. Percentage of TM composition exhibited by a residue in each organism increases from blue to red.

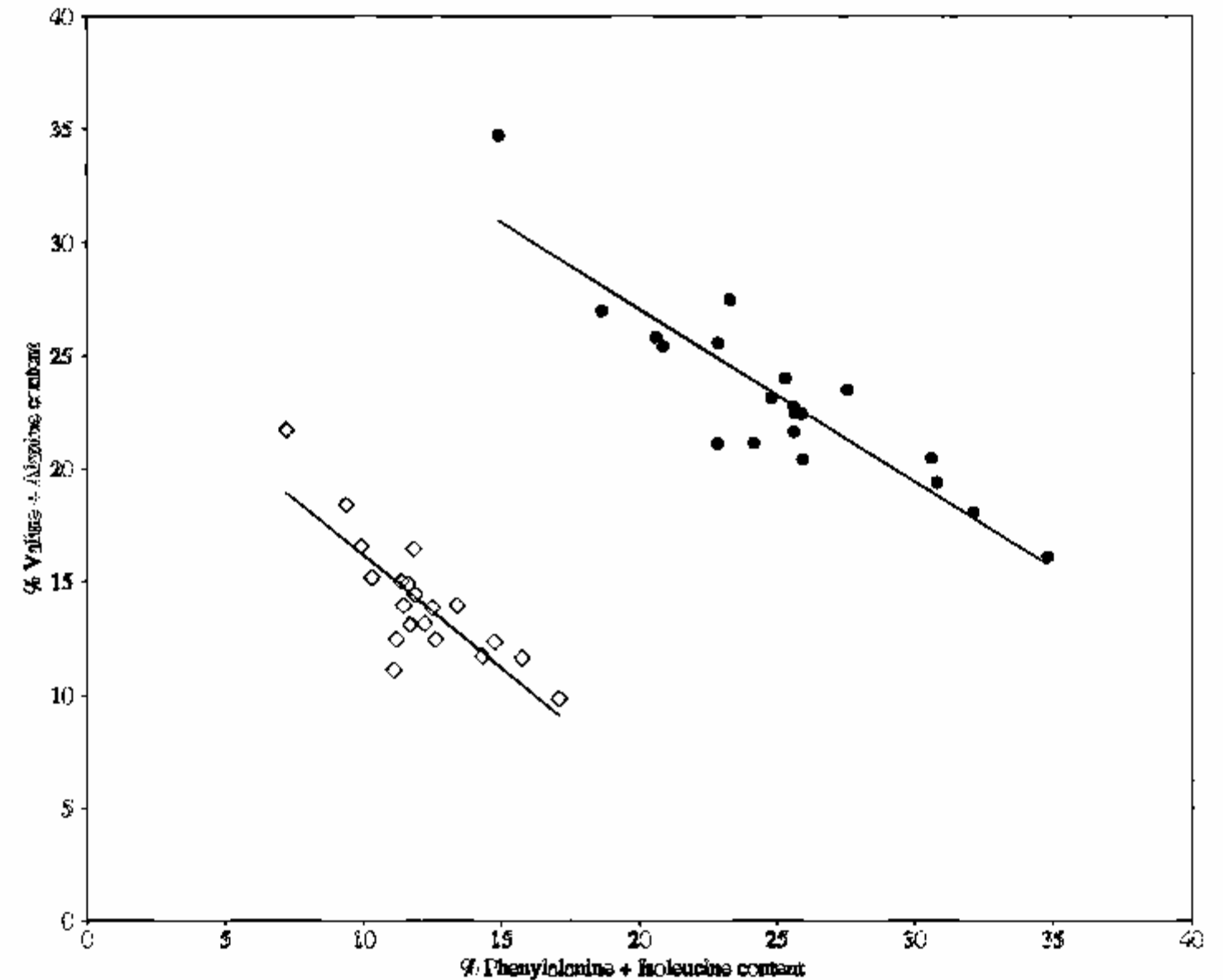
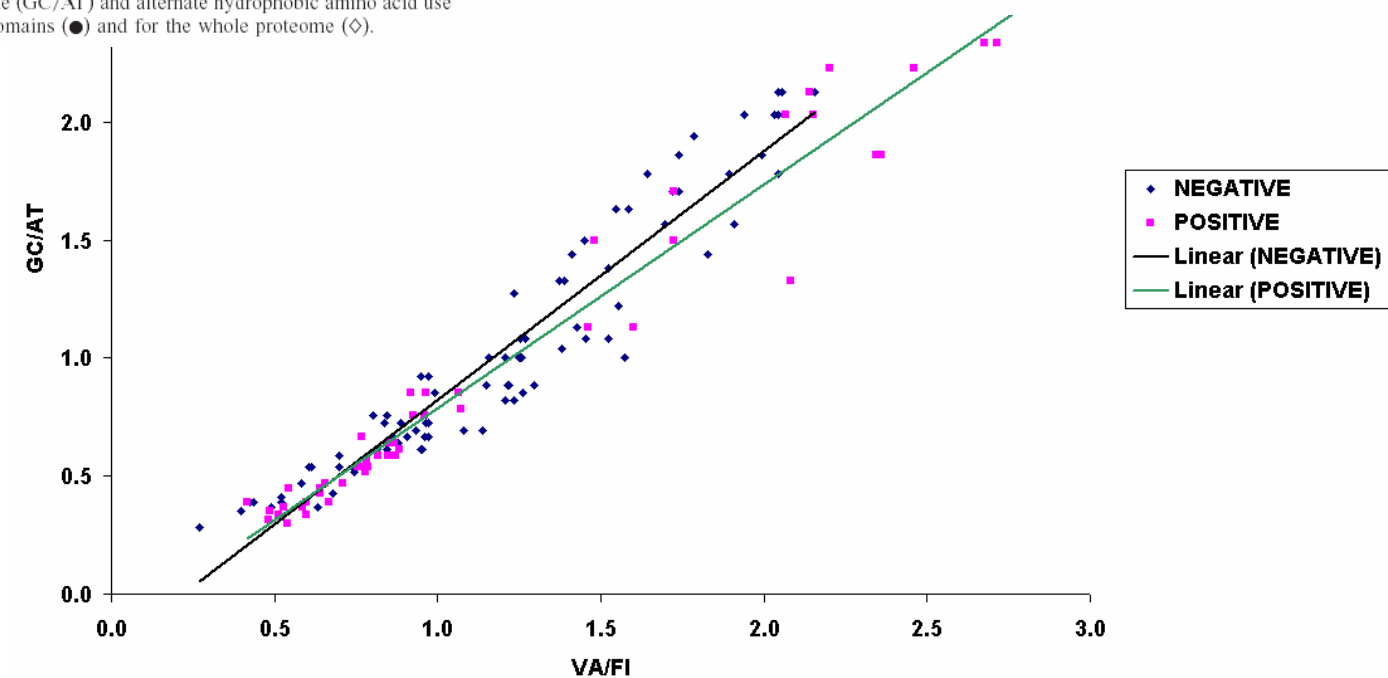
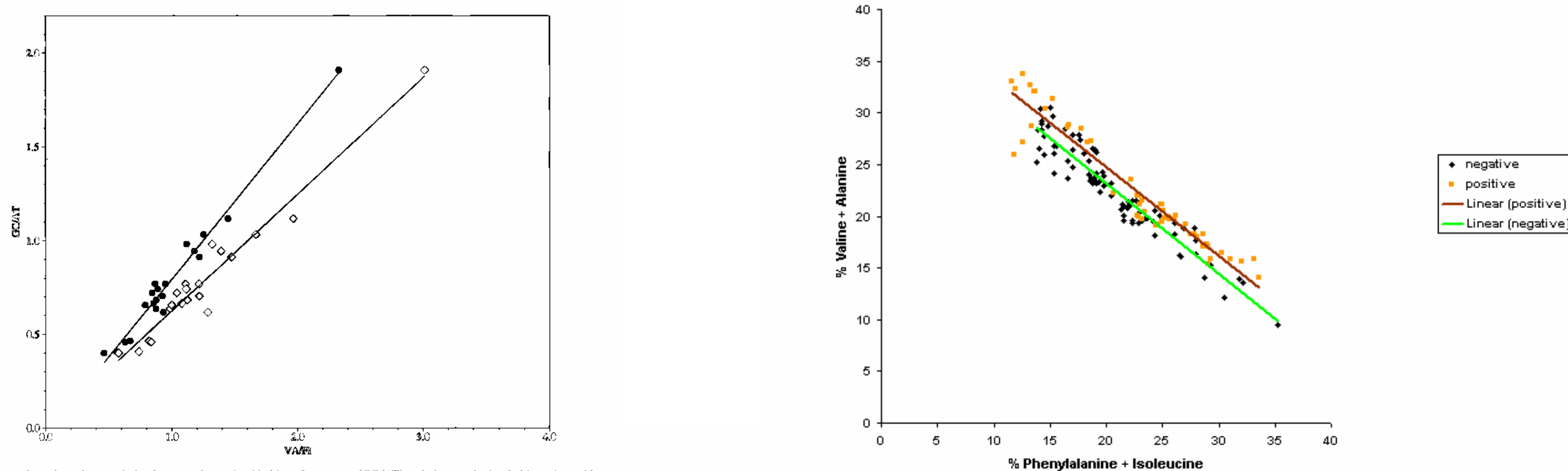


Fig. 2. A graph to show the correlation between the percentage abundance of valine plus alanine residues (VA) and the percentage abundance of phenylalanine plus isoleucine residues (FI). Data are shown for each organism, for both the amino acids within the predicted TM domains (●) and the whole proteome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

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

GC% content και διαμεμβρανικές πρωτεΐνες



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

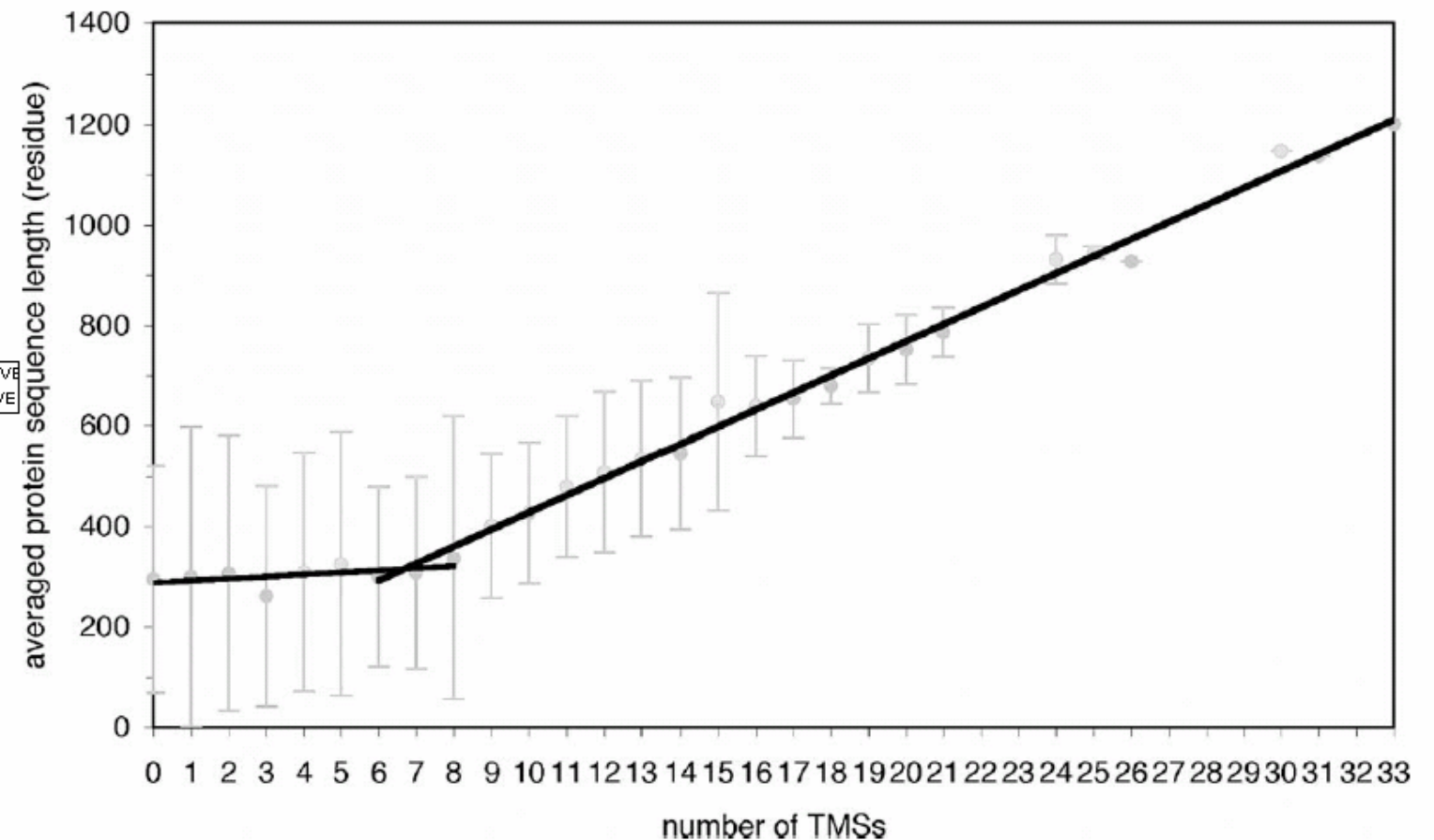
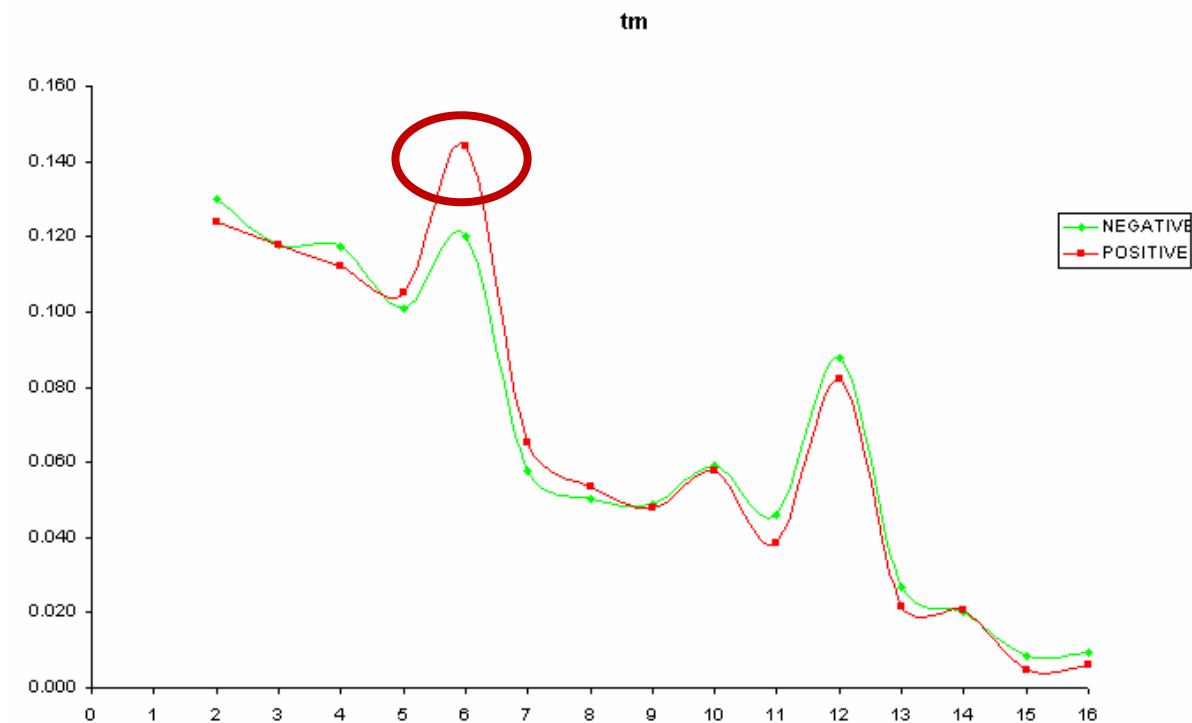


Fig. 4. Relationship between the number of TMSs and the protein sequence length averaged over the 50 genomes. The slope of the line between 7- and 33-tms is 35 residues. The number of TMSs, 0 in the abscissa means soluble proteins.

Μήκος των διαμεμβρανικών πρωτεϊνών και ο εσωτερικός διπλασιασμός

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.:

(A)

```
[1-3] MRKLRILAIIVLIALSIILLIAGGVLLTVAIPGLSSVISSPAGMGACALGCVMLALGIDVL
      *****
[4-6] -----SSVISSPAGMGACALGCVMLALGIDVL

[1-3] LKKREVPFIVLASVTTTPGTGSPRSGISISGADSTIRSLPTYLLDEGHPQSMRKLRI LAIV
      *****
[4-6] LKKREVPFIVLASVTTTPGTGSPRSGISISGADSTIRSLPTYPLDEGHPQSMRKLRI LAIV

[1-3] LIVFSIILLIASGVVLLTVAIPGL-----
      *****
[4-6] LIVFSIILLIASGVVLLTVAIPGLSSIISPAEMGACALGCVMLALGIDVLLKKREVPF
```

(B)

```
[2-3] SSVISSPAGMGACALGCVMLALGIDVLLKKREVPFIVLASVTTTPGTGSPRSGISISGADS
      *****
[4-5] SSVISSPAGMGACALGCVMLALGIDVLLKKREVPFIVLASVTTTPGTGSPRSGISISGADS

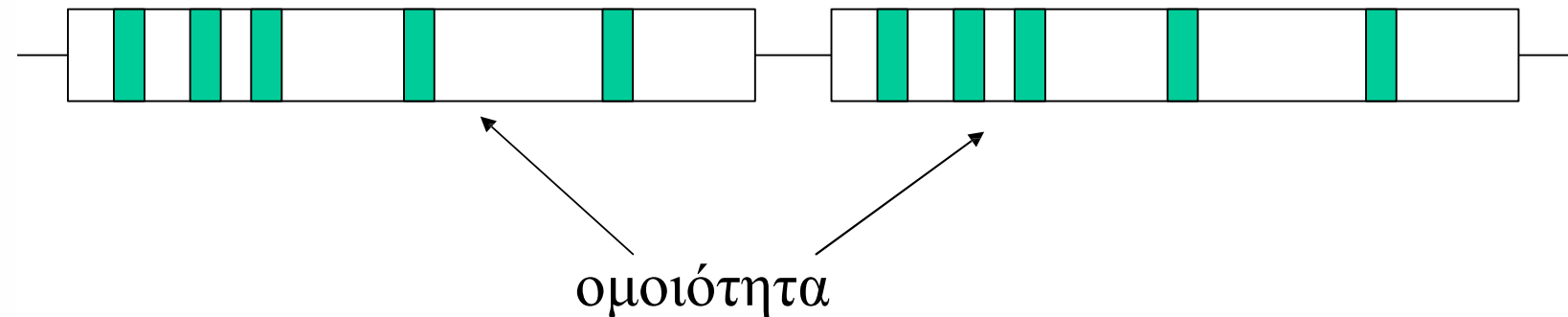
[2-3] TIRSLPTYLLDEGHPQSMRKLRI LAIVLIVFSIILLIASGVVLLTVAIPGL-
      *****
[4-5] TIRSLPTYPLDEGHPQSMRKLRI LAIVLIVFSIILLIASGVVLLTVAIPGLS
```

(C)

```
[1-1] -MRKLRI LAIVLIALSIILLIAGGVLLTVAIPGL
      *****
[3-3] SMRKLRI LAIVLIVFSIILLIASGVVLLTVAIPGL
```

(D)

```
[2-2] SSVISSPAGMGACALGCVMLALGIDVLLKKREVP-
      * _*****
[6-6] S-IISSPAEMGACALGCVMLALGIDVLLKKREVPF
```



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.

Συγκριτική Γονιδιωματική

Η συγκριτική γονιδιωματική πάει ένα βήμα παραπέρα την υπολογιστική ανάλυση των γονιδιωμάτων.

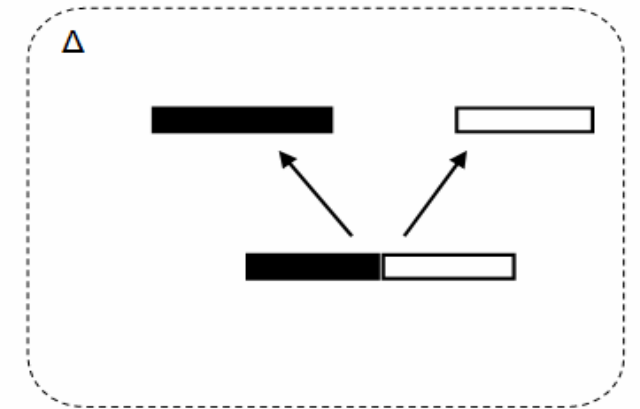
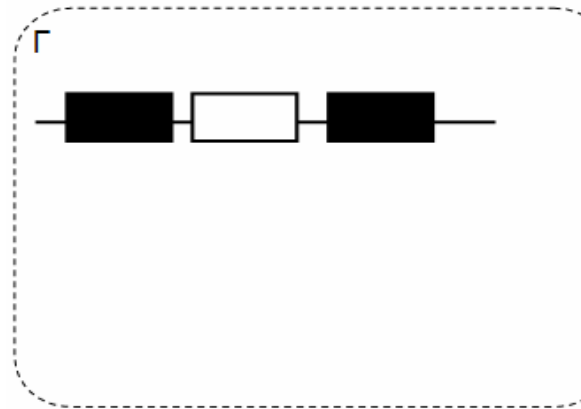
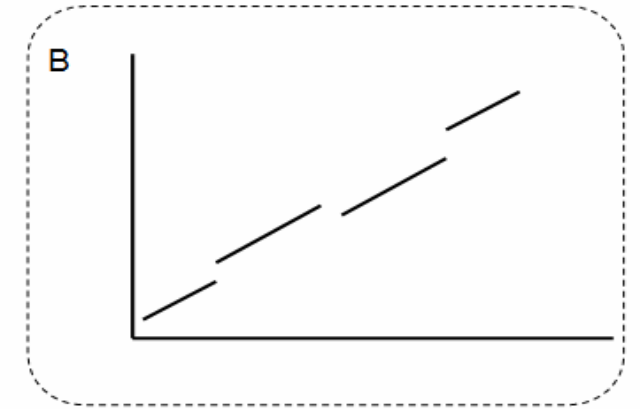
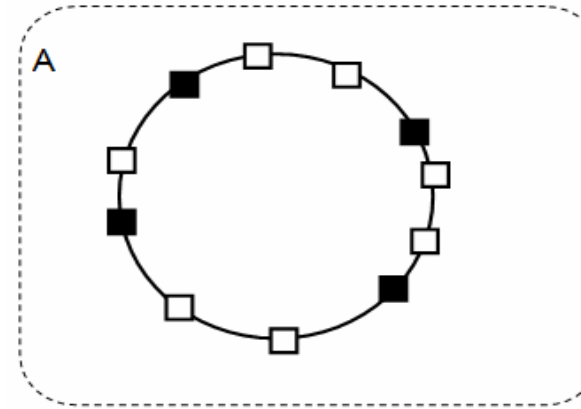
Αντί να εστιάζει μόνο στα συνολικά στατιστικά μέτρα από κάθε γονιδίωμα, όπως π.χ. το ποσοστό GC ή κάποιο άλλο μέτρο, **επιχειρεί να χρησιμοποιήσει τη βασική αρχή της φυλογενετικής ανάλυσης**, ότι δηλαδή τα γονιδιώματα όλων των οργανισμών προέρχονται από προγονικές μορφές και έχουν διαμορφωθεί έτσι όπως είναι σήμερα μετά από αλληπάλληλες 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 διάταξη 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).

Μέθοδοι

- Η μέθοδος «αφαίρεσης» γονιδίων, στην οποία συγκρίνονται σε μια σειρά οργανισμούς τα κοινά γονίδια και εντοπίζονται τα μοναδικά γονίδια.
- 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 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 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).

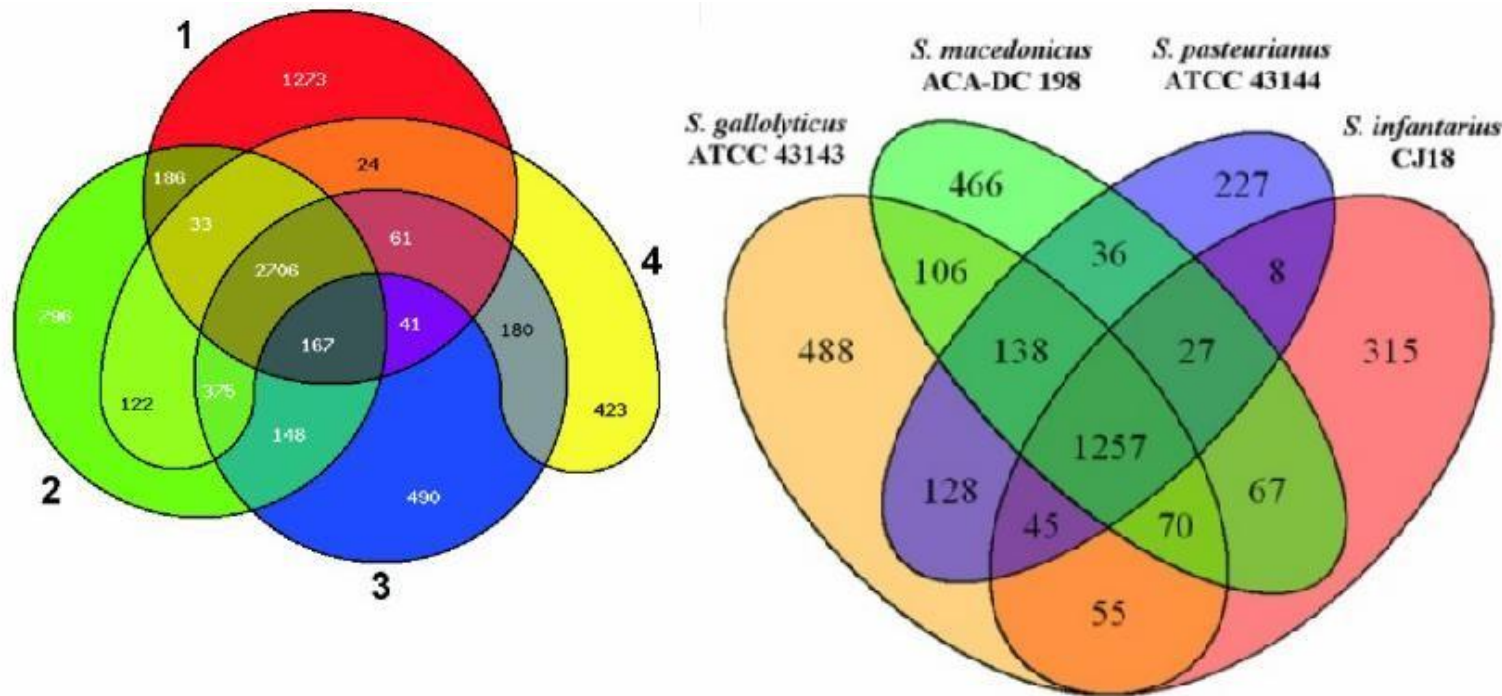


Όλες οι παραπάνω μεθοδολογίες λειτουργούν με χρήση της **ομοιότητας** των γονιδίων και των πρωτεϊνικών προϊόντων τους και κάνουν χρήση της πληροφορίας από τη **σχετική θέση των γονιδίων** (ή και την ίδια την ύπαρξή τους) σε διαφορετικούς οργανισμούς. Παρόλα αυτά, οι μεθοδολογίες αυτές **εντοπίζουν διαφορετικού είδους λειτουργικές συσχετίσεις μεταξύ των γονιδίων**. Προσφέρουν δηλαδή διαφορετικά αποτελέσματα, γι' αυτό και στη μεγάλη τους πλειοψηφία **δρουν συμπληρωματικά**.

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 *Streptococcus* με το R ([Papadimitriou et al., 2014](#))

Η μέθοδος «αφαίρεσης» γονιδίων

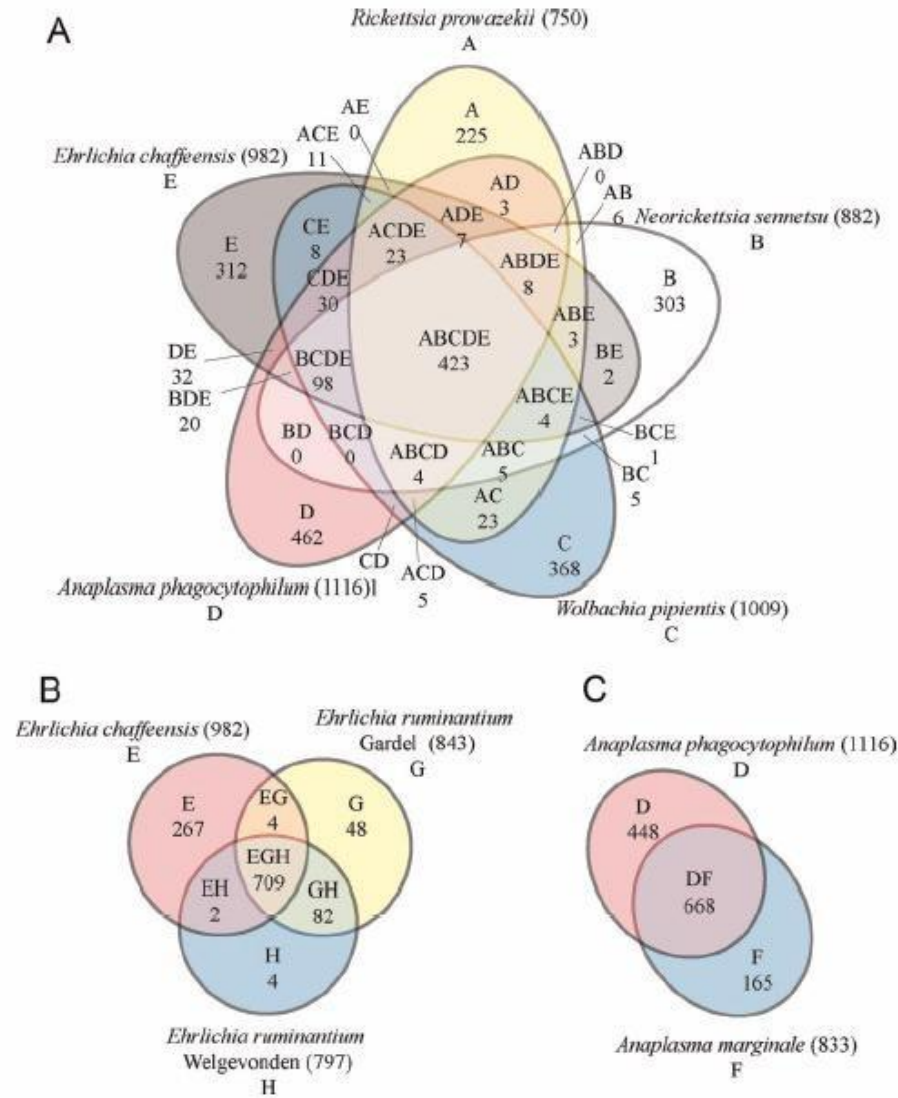
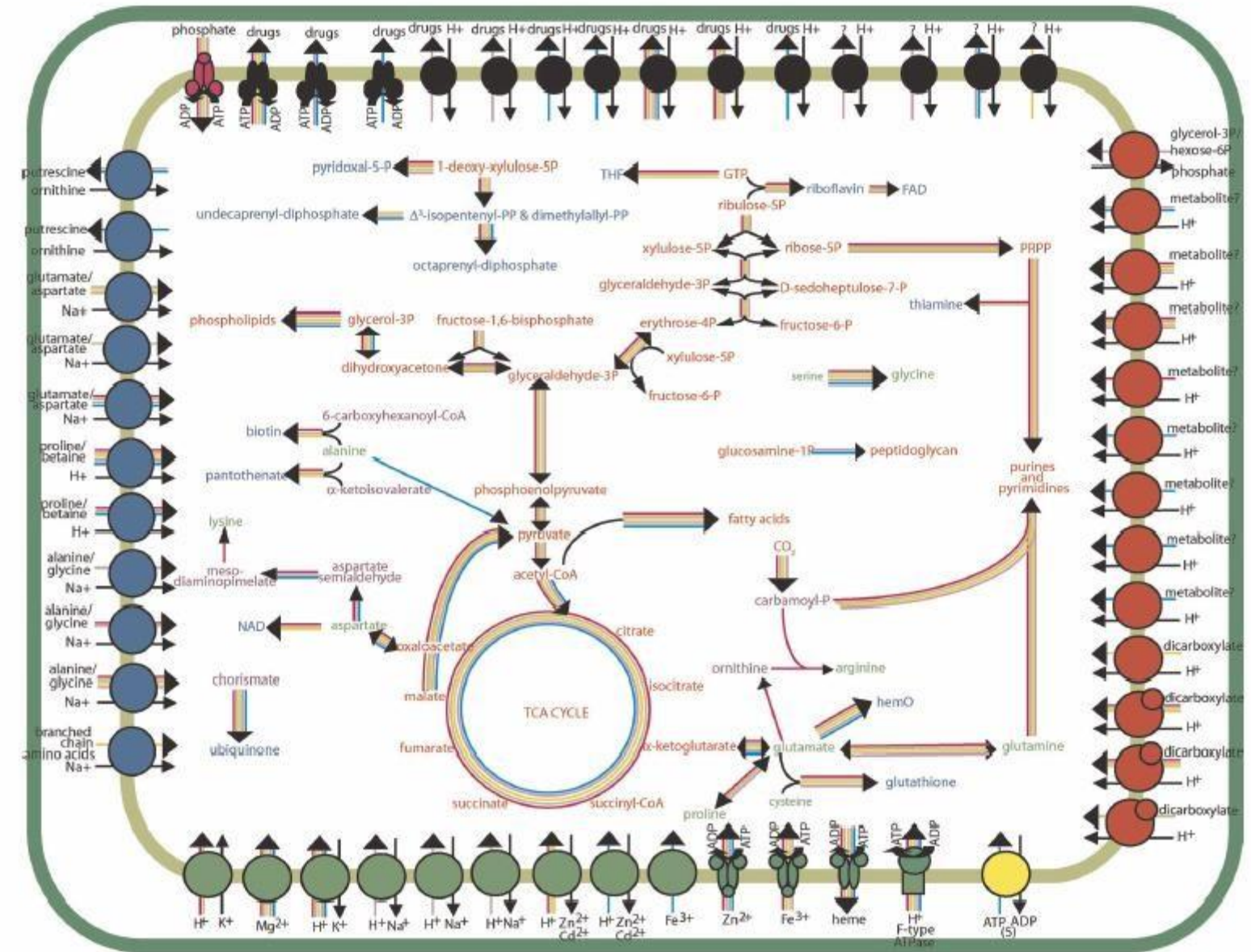


Figure 4. Comparison of the Rickettsiales Gene Sets

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.



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, Kunitz, & Goldovsky, 2006](#)).

Ελάχιστον αριθμός απαραίτητων γονιδίων

- *Mycoplasma genitalium* (468 γονίδια που κωδικοποιούν για πρωτεΐνες)
- *Haemophilus influenzae* (1703 γονίδια)
- 240 *M. genitalium* γονίδια είναι ορθόλογα με γονίδια του *H. influenzae*.
- 22 μη ορθόλογες μετατοπίσεις

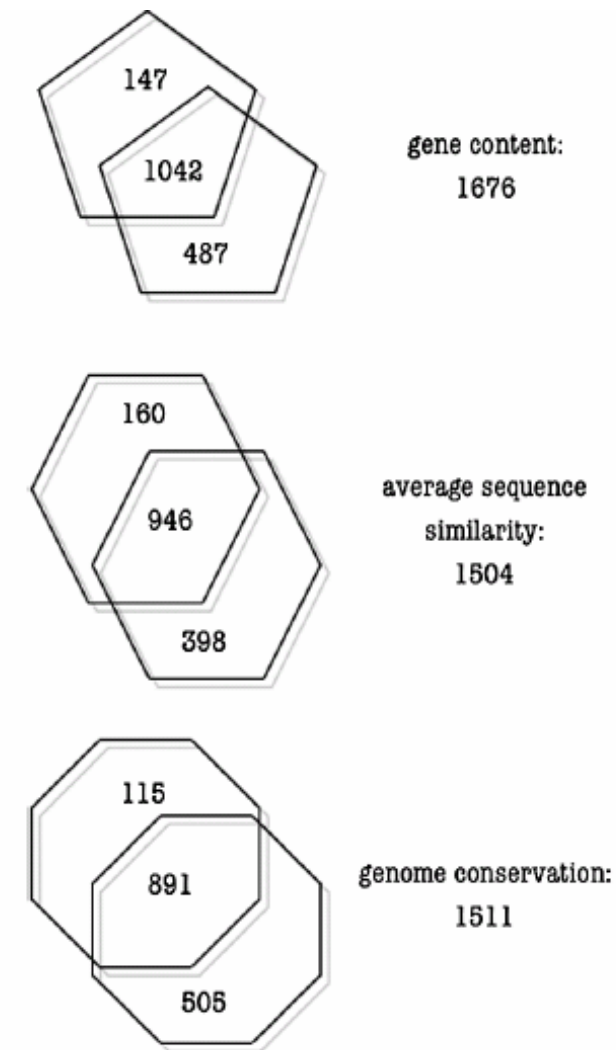


Fig. 1. A representation of the minimal gene content for LUCA. Upper diagrams represent analyses without eukaryotes, lower diagrams represent analyses with eukaryotes; pentagons represent gene content (CT), hexagons represent average sequence similarity (AS), octagons represent genome conservation (GC)—see 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).

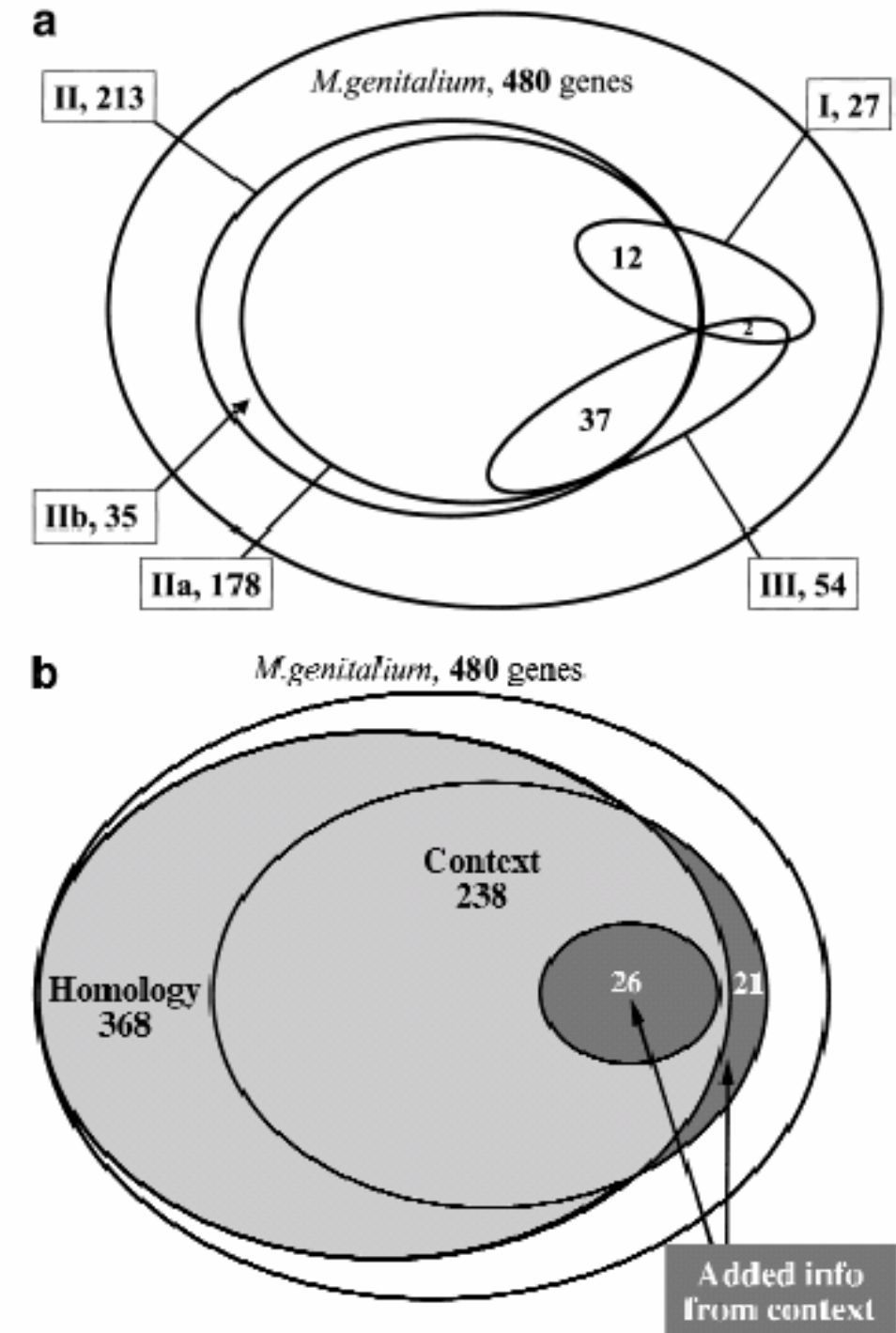
Συνθετική βιολογία και εξωβιολογία

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>.

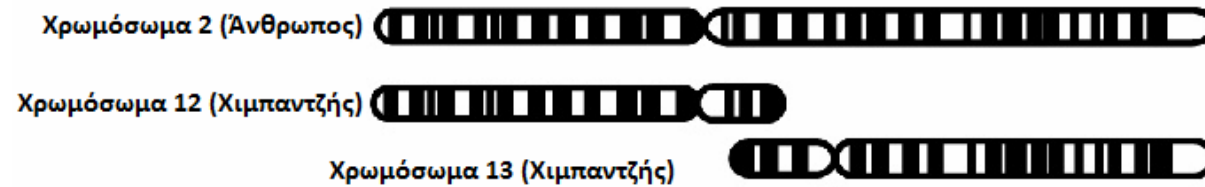


Η μέθοδος στοίχισης ολόκληρων γονιδιωμάτων

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.

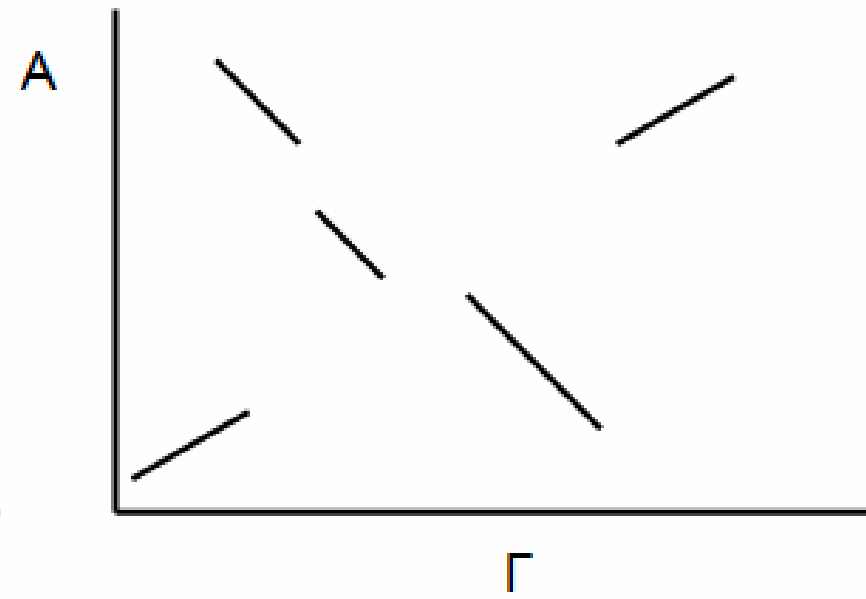
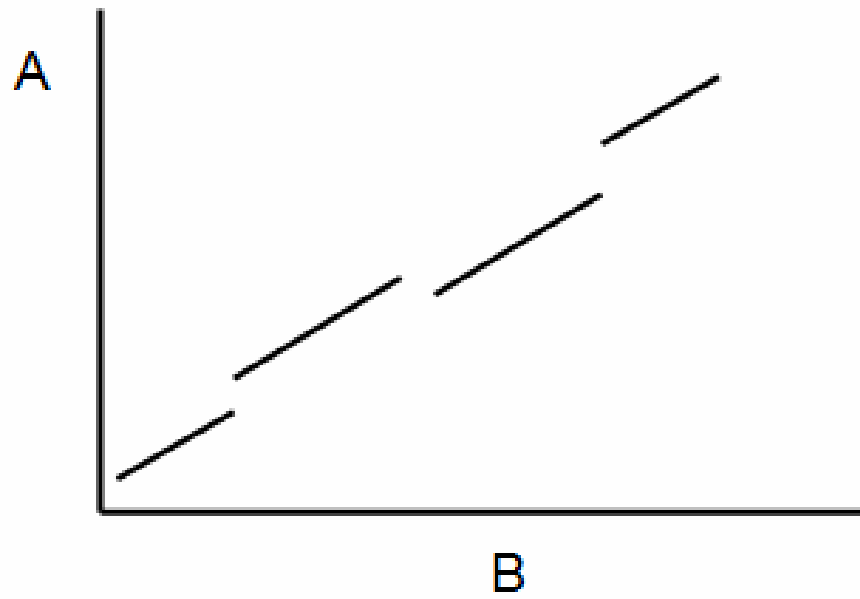


Στοίχιση ολόκληρων γονιδιωμάτων μπορεί να γίνει με διαφορετικούς τρόπους:

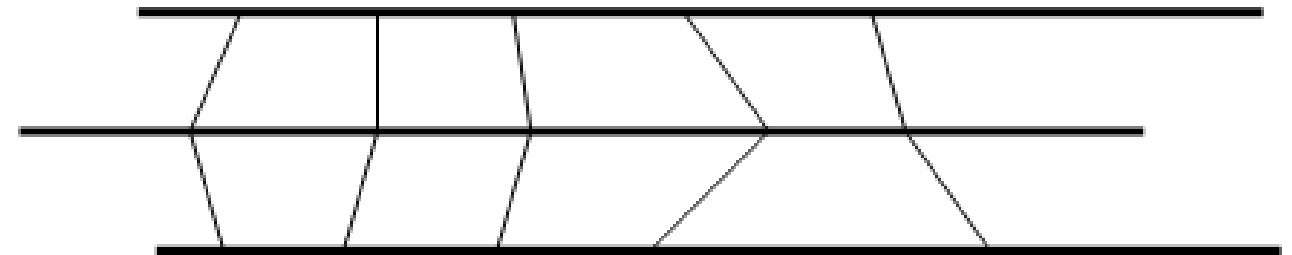
Ο πιο **απλός** είναι με μια παραλλαγή του διαγράμματος σημείων (**dot-plot**) η οποία επεκτείνεται σε όλο το γονιδίωμα, ή με επέκταση κάποιου γνωστού αλγόριθμου στοίχισης (όπως το **BLAST**) η οποία να επιτρέπει χρήση μεγάλων ακολουθιών. Οι πιο σύγχρονες μεθοδολογίες **συνδυάζουν** τόσο τους **αλγορίθμους** 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.

Οι τεχνικές αυτές είναι πολύ πιο εύκολο να εφαρμοστούν σε βακτηριακά ή ιικά γονιδιώματα, τόσο γιατί είναι πιο μικρά όσο και γιατί είναι ενιαία, καθώς τα πολλαπλά χρωμοσώματα των ευκαρυωτικών οργανισμών απαιτούν σύγκριση ένα με ένα.

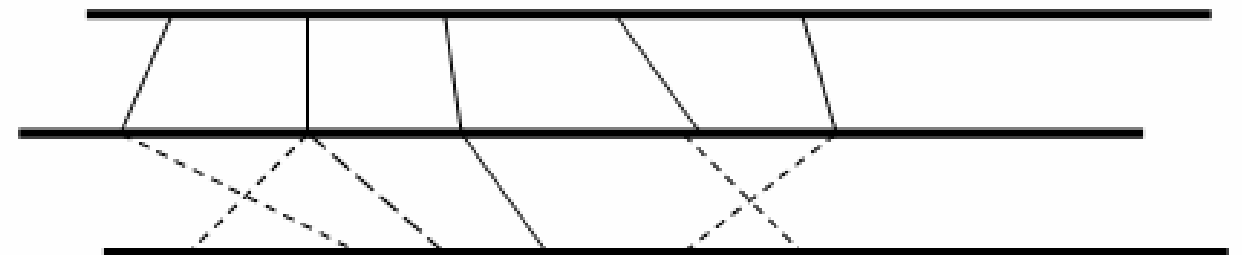
Η μέθοδος στοίχισης ολόκληρων γονιδιωμάτων



A



B



Η μέθοδος στοίχισης ολόκληρων γονιδιωμάτων

- 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).

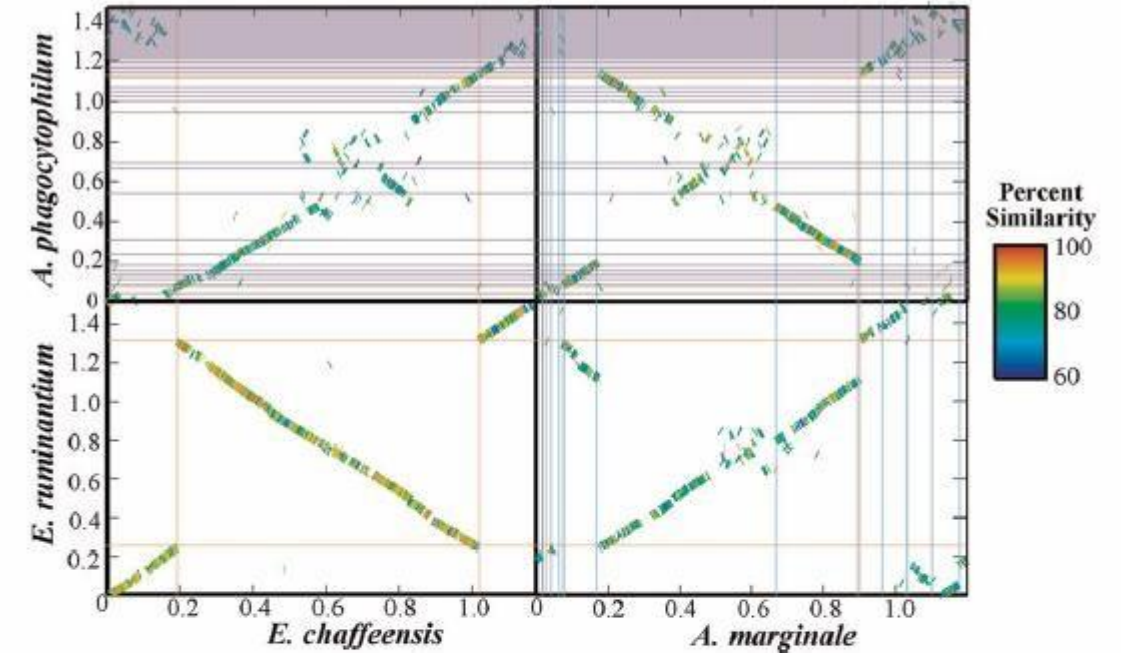


Figure 3. Synteny between *Anaplasma* spp. and *Ehrlichia* spp.

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 *msh2*- and *msh3*-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.

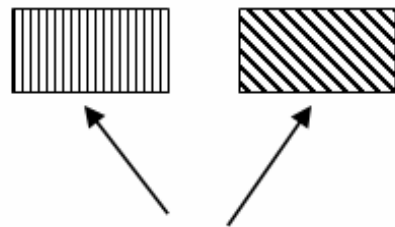
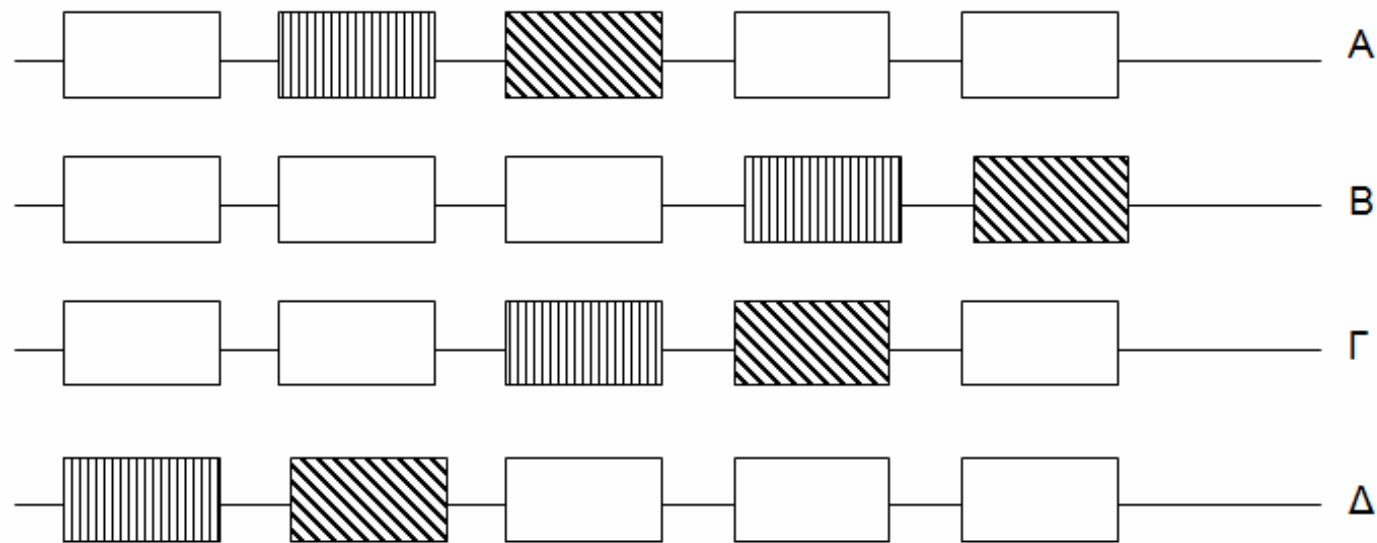
Η βασική αρχή της μεθόδου μοιάζει διαισθητικά με την αρχή της σύνδεσης στη γενετική, μόνο που εδώ χρησιμοποιείται σε μεγαλύτερη κλίμακα χρόνου.

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.

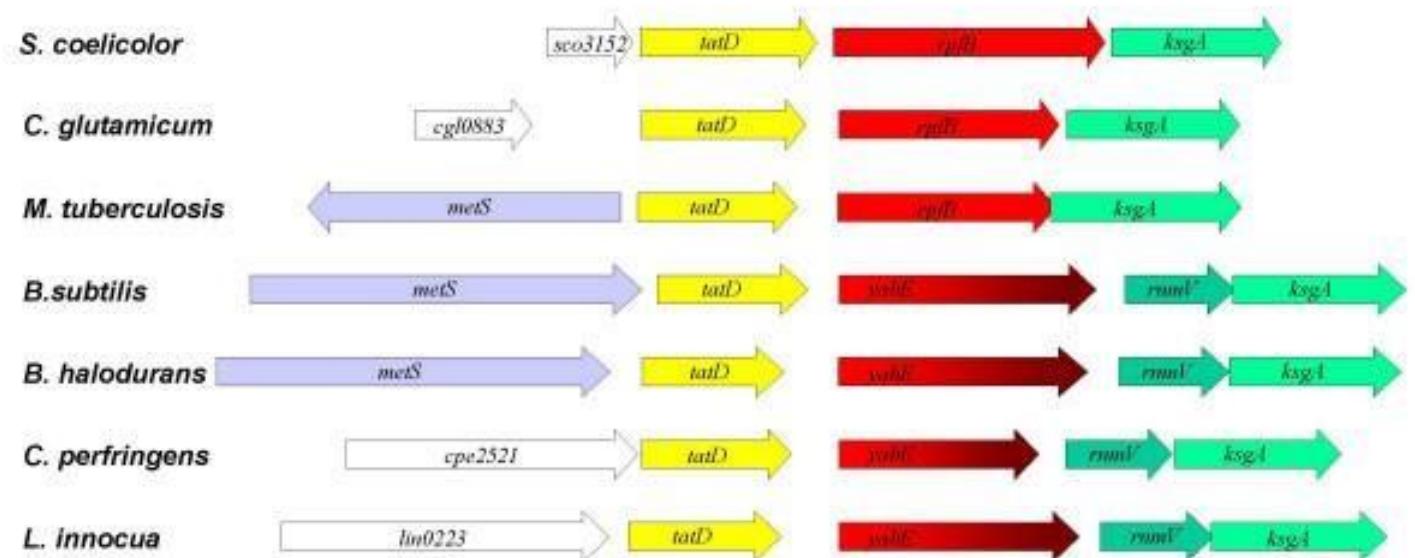
Με τη μέθοδο αυτή είναι δυνατό να εντοπιστούν **συσχετίσεις** μεταξύ γονιδίων που κωδικοποιούν **τελείως διαφορετικές πρωτεΐνες**.

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.

Η μέθοδος σύγκρισης της σειράς των γονιδίων



Τα γονίδια αυτά συνδέονται
με κάποιον τρόπο



Η μέθοδος σύγκρισης της σειράς των γονιδίων

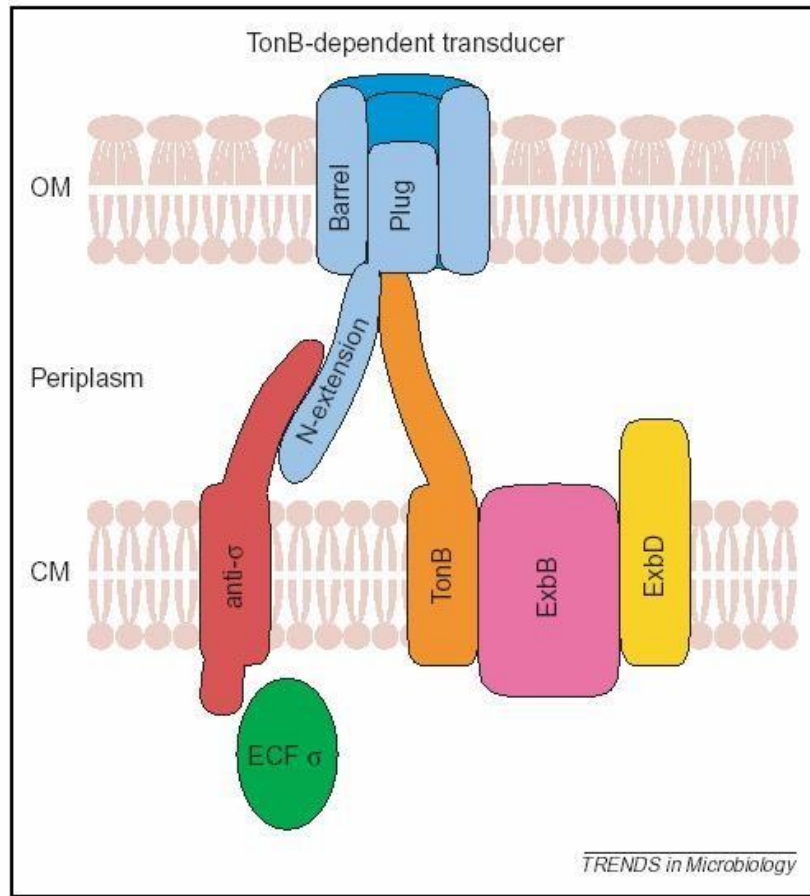


Figure 1. Structural organization of TonB-dependent regulatory systems. TonB-dependent regulatory systems consist of six components, an outer membrane TonB-dependent transducer (blue) in interplay with its energizing TonB-ExbB-ExbD protein complex (orange, pink and yellow), a cytoplasmic membrane-localized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

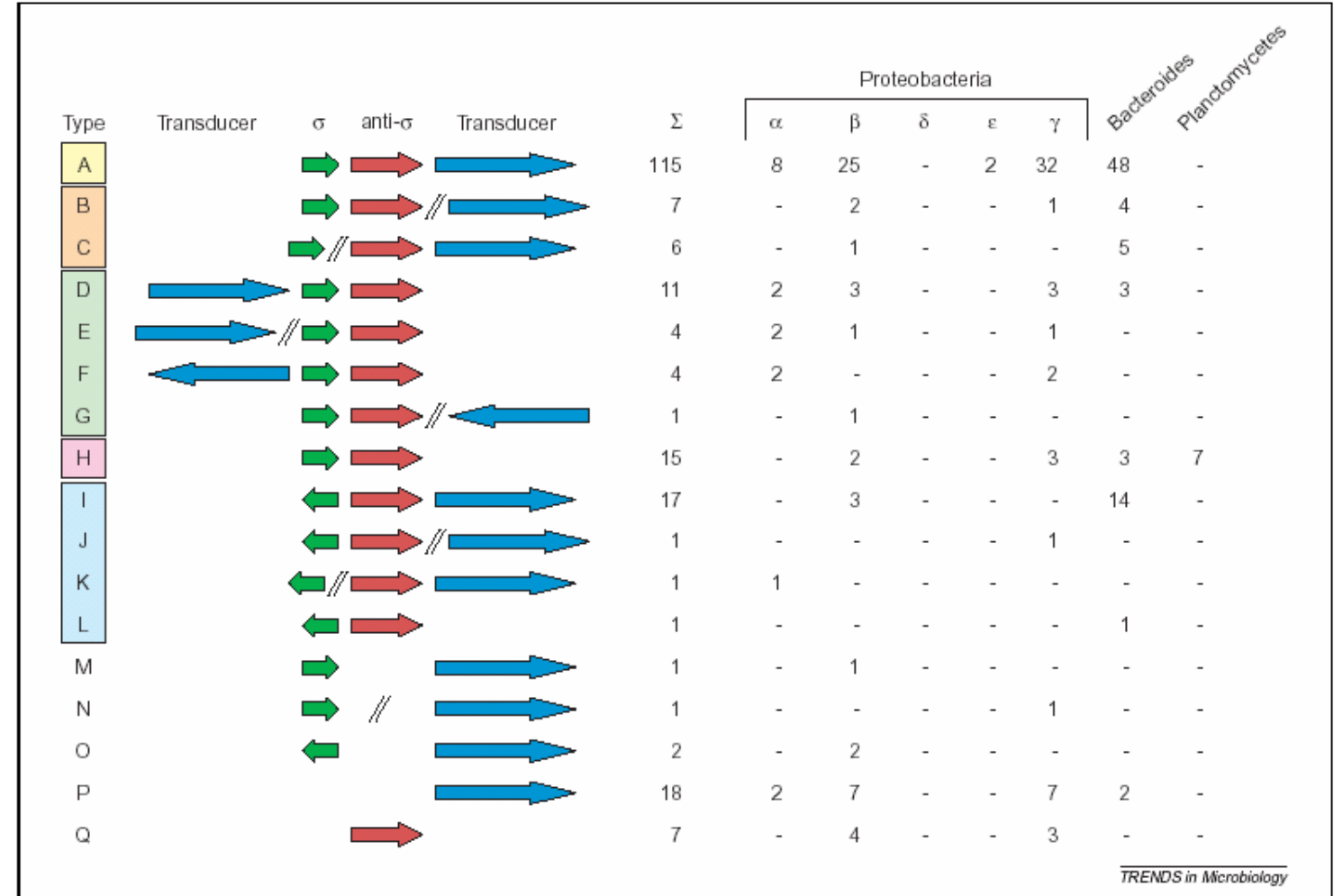


Figure 2. Genetic organization of regulatory systems consisting of a TonB-dependent transducer (blue), an anti-sigma factor (red) and an ECF-subfamily sigma factor (green). The occurrence of the different genetic organizations in proteobacteria (α , β , δ , ϵ and γ), in *Bacteroides*, in planctomycetes 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

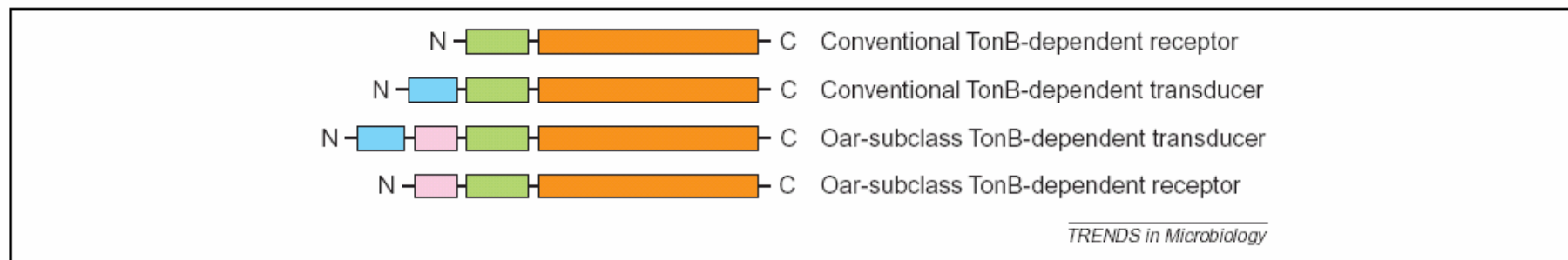


Figure 3. Domain structure of TonB-dependent receptors. All TonB-dependent receptors consist of a C-terminal β -barrel (orange) and a plug domain (green), which seals the barrel (see also Figure 1). TonB-dependent transducers have an N-terminal extension (blue) of ~ 70 amino acids. Receptors from *Bacteroides* 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*

Η μέθοδος εντοπισμού προϊόντων γονιδιακής σύντηξης

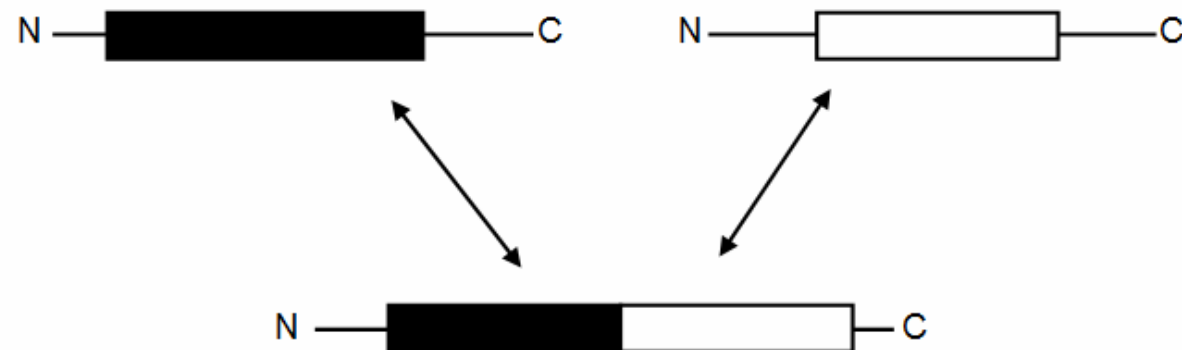
Η βασική αρχή αυτής της μεθόδου βασίζεται στη **σπονδυλωτή φύση των πρωτεϊνών**, δηλαδή, στην ύπαρξη ανεξάρτητων δομικών και λειτουργικών περιοχών (domains).

Έτσι, με τη μέθοδο αυτή **εντοπίζονται** με υπολογιστικό τρόπο **γονίδια ενός οργανισμού A** τα οποία **σε κάποιον άλλον οργανισμό B 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 B, **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 A ([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 αναγωγή (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 θυμιδικής συνθετάσης (TS) η οποία συμμετέχει στο ίδιο μονοπάτι (σύνθεση νουκλεοτιδίων) και η οποία στους ευκαρυωτικούς οργανισμούς βρίσκεται σε διαφορετικό γονίδιο.



Χάρτες πρωτεϊνικών αλληλεπιδράσεων ολόκληρων γονιδιωμάτων βασιζόμενοι σε γεγονότα σύντηξης γονιδίων

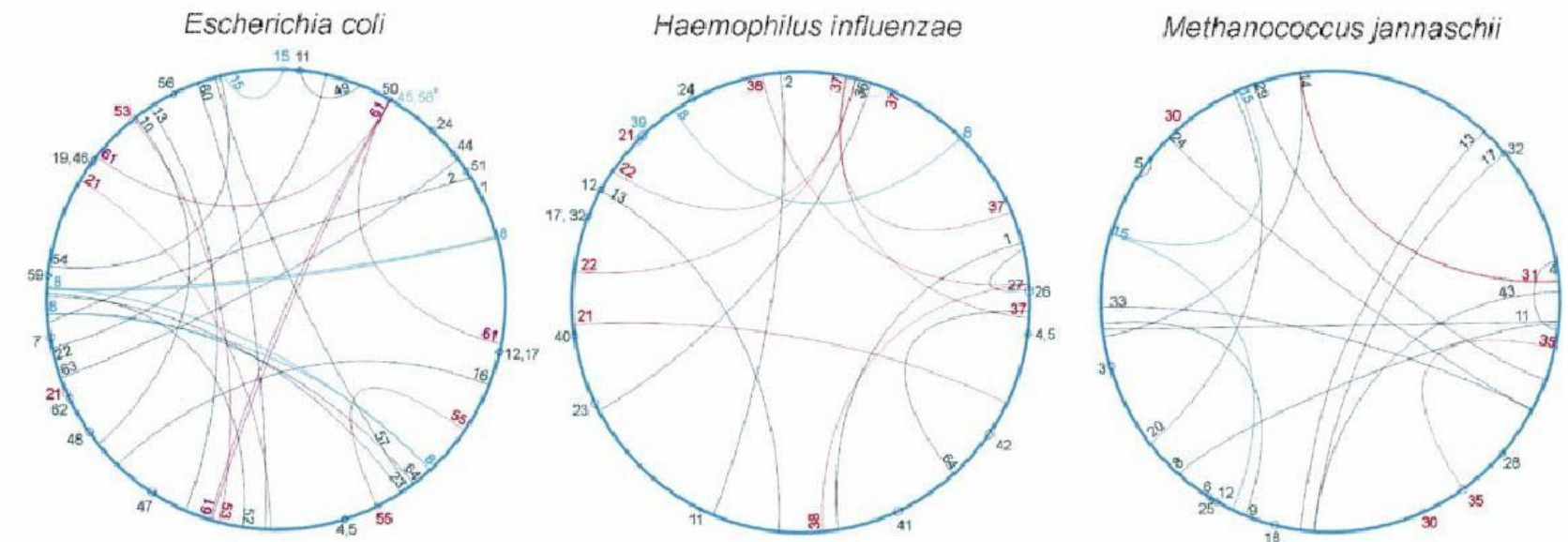
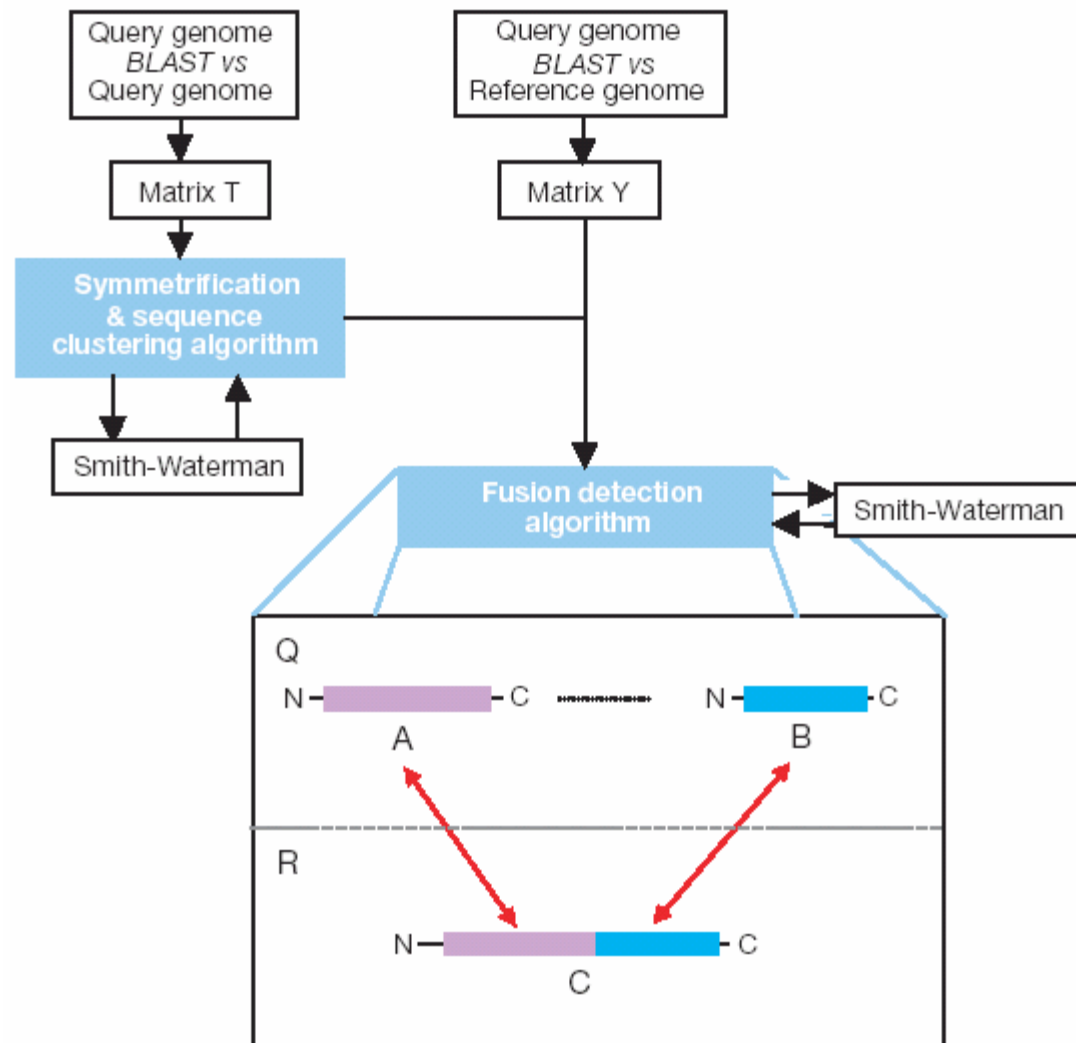


Figure 2 Representation of protein interaction maps for the most likely interactions predicted for *E. coli*, *H. influenzae* and *M. jannaschii*. In the large blue circles, which represent the three genomes, 0° corresponds to the first base pair, and 360° the last base pair of the genome. Predicted 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 εντοπισμού προϊόντων γονιδιακής σύντηξης

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 υποψήφιος 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).

Η μέθοδος εντοπισμού προϊόντων γονιδιακής σύντηξης

- 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

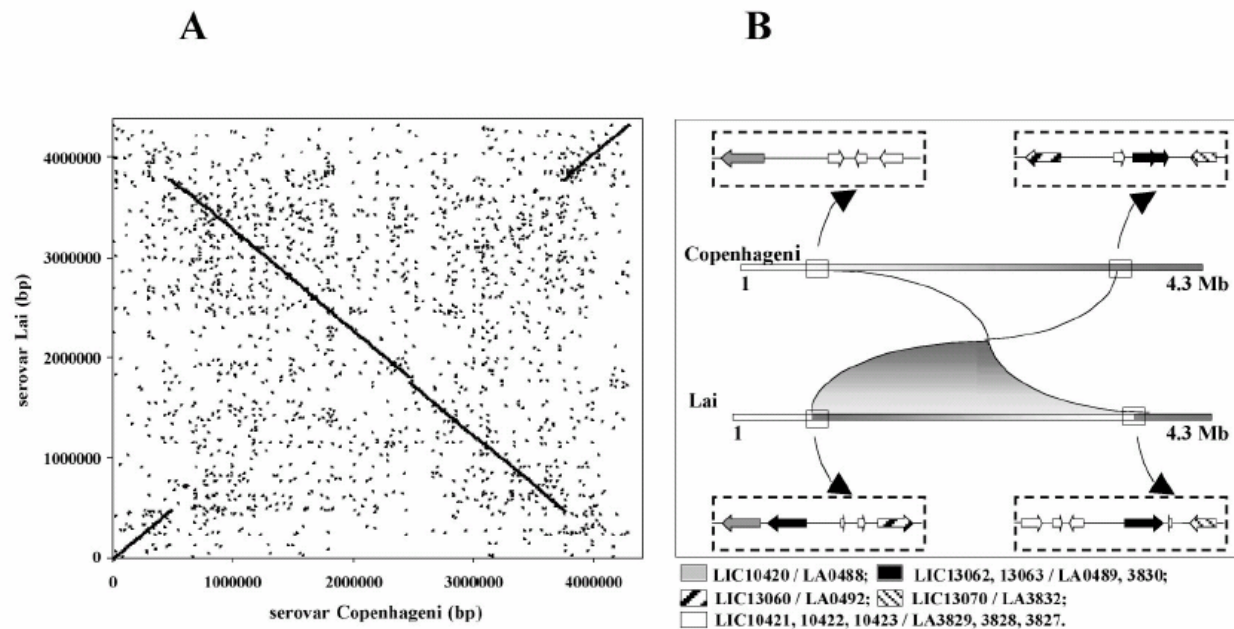


FIG. 1. Inversion of *L. interrogans* serovars Copenhageni and Lai CI chromosomes. (A) Nucleotide alignment obtained by using MUMmer, which relies on exact matches of at least 20 bp. Each dot in the figure is one such match. The dark lines on the two main diagonals result from the high density of points 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.

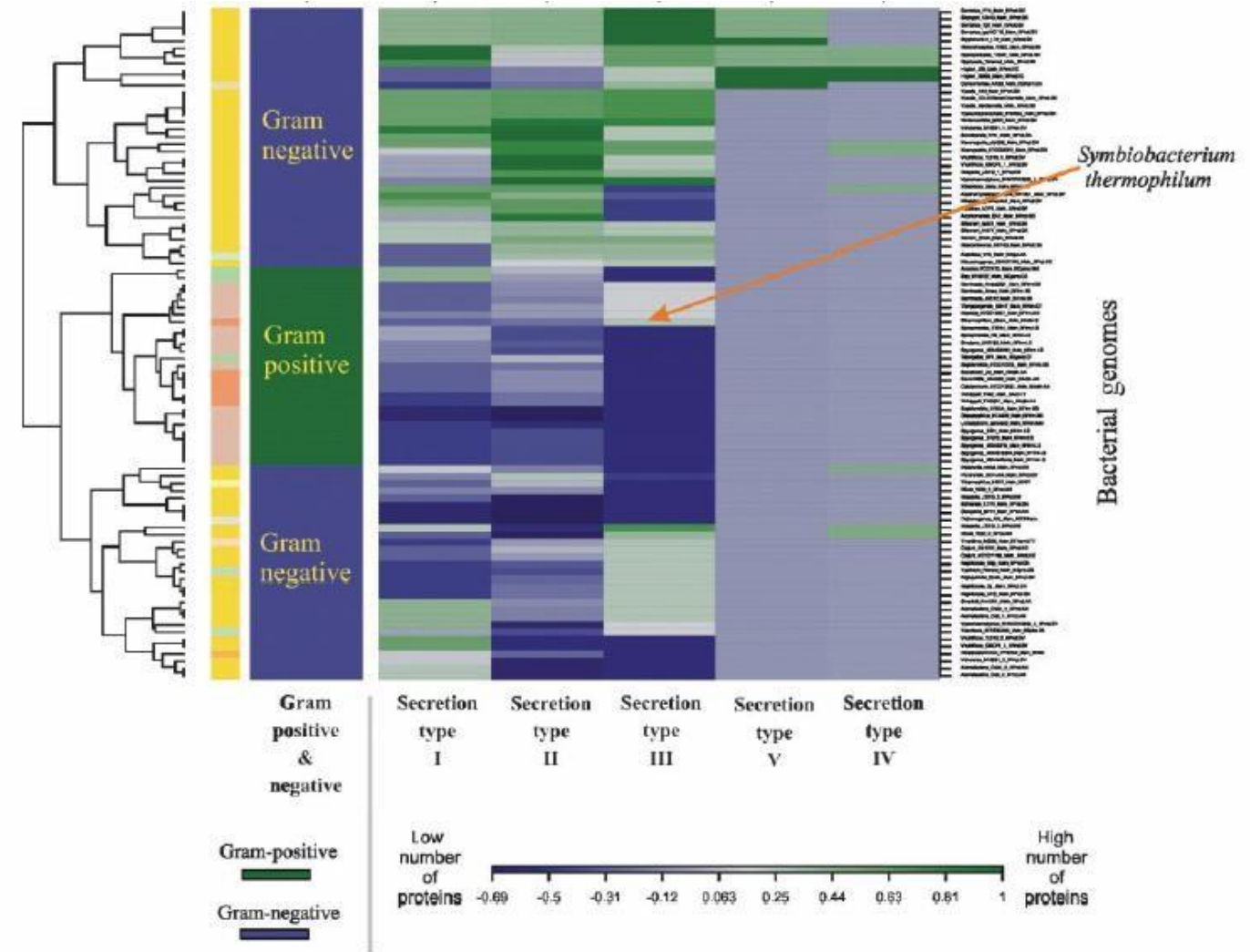


Fig. 1. Two-dimensional clustering (Willenbrock *et al.*, 2005) of bacterial genome sequences versus secretion systems type I-V. Dark blue indicates that a low number of the selected proteins is present for the specific secretion type; dark green represents cases where we find that most of the proteins 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.

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>)

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/>).

Λογισμικό

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, & Baptiste, 2013](#)).