
**ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ**

**ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ**

**ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΓΕΝΕΤΙΚΗΣ**

# *Μελέτη Προτύπων Σύστασης, Κατανομές και Συμμετρίες Νουκλεοτιδικών Μοτίβων σε Γονιδιωματική Κλίμακα*

**Διδακτορική Διατριβή του**

**Κωνσταντίνου Αποστόλου Καραμπέλη**

**Βιολόγου**

**Αλεξανδρούπολη 2016**

Η παρούσα διατριβή εκπονήθηκε στο Εργαστήριο *Θεωρητικής Βιολογίας και Υπολογιστικής Γονιδιωματικής* του Ινστιτούτου *Βιοεπιστημών και Εφαρμογών*, του ΕΚΕΦΕ 'Δημόκριτος', σε συνεργασία με το Εργαστήριο *Υπολογιστικής και Δομικής Βιολογίας* του Τμήματος *Μοριακής Βιολογίας και Γενετικής*, του Δημοκρίτειου Πανεπιστημίου Θράκης. Υποστηρίχθηκε δε από υποτροφία που χορηγήθηκε κατόπιν διαγωνισμού, από το ΕΚΕΦΕ 'Δημόκριτος'.

Ήταν η τελευταία χρονιά που προκηρύχθηκαν εσωτερικές υποτροφίες από το 'Δημόκριτο'. Κατόπιν, το σχετικό κονδύλι περικόπηκε από τον προϋπολογισμό του Κέντρου.

## Επταμελής Εξεταστική Επιτροπή

**Γλυκός Νικόλαος,** Επίκουρος Καθηγητής του Τμήματος Μοριακής Βιολογίας και Γενετικής, Δημοκρίτειο Πανεπιστήμιο Θράκης, ως Επιβλέπων και Πρόεδρος της Τριμελούς Συμβουλευτικής Επιτροπής

**Αλμυράντης Ιωάννης,** Ερευνητής Α' βαθμίδος του ΕΚΕΦΕ 'Δημόκριτος', μέλος της Τριμελούς Συμβουλευτικής Επιτροπής

**Βλάση Μεταξία,** Ερευνήτρια Α' βαθμίδος του ΕΚΕΦΕ 'Δημόκριτος', μέλος της Τριμελούς Συμβουλευτικής Επιτροπής

**Καλδούδη Ελένη,** Αναπληρώτρια Καθηγήτρια του Τμήματος Ιατρικής, Δημοκρίτειο Πανεπιστήμιο Θράκης, μέλος

**Μπουλουγούρης Γεώργιος,** Επίκουρος Καθηγητής του Τμήματος Μοριακής Βιολογίας και Γενετικής, Δημοκρίτειο Πανεπιστήμιο Θράκης, μέλος

**Νικολάου Χριστόφορος,** Επίκουρος Καθηγητής του Τμήματος Βιολογίας, Πανεπιστήμιο Κρήτης, μέλος

**Προβατά Αστέρω,** Ερευνήτρια Α' βαθμίδος του ΕΚΕΦΕ 'Δημόκριτος', μέλος

ημερομηνία υποστήριξης,

13η Σεπτεμβρίου του 2016

«Η έγκριση της διδακτορικής διατριβής από Τμήμα Μοριακή Βιολογίας και Γενετικής του Δημοκρίτειου Πανεπιστημίου Θράκης δεν υποδηλώνει αποδοχή των απόψεων του συγγραφέα.»

(Νόμος 5343/32, άρθρο 202 §2 και Νόμος 1268/82, άρθρο 50 §8)

***Ὁ βίος βραχὺς,***
***ἡ δὲ τέχνη μακρὴ,***
***ὁ δὲ καιρὸς ὀξὺς,***
***ἡ δὲ πεῖρα σφαλερὴ,***
***ἡ δὲ κρίσις χαλεπή.***

***ΙΠΠΟΚΡΑΤΗΣ, Ἀφορισμοί***

*ρητό που συχνά ανέφερε ο Γιάννης Αλμυράντης*
*στις πολλές και μακρές συζητήσεις μας*

*Πλησιάζοντας στο τέλος μιας πορείας που προχωρεί, παραπατά, παλινδρομεί, μα πάντα καταφτάνει, θα ήθελα να ευχαριστήσω όλους όσοι, από κοινές επιλογές ή και από συμπτώσεις, υπήρξαν οι συνοδοιπόροι μου. Τον Γιάννη Αλμυράντη, δίχως τον οποίο η διαδρομή που ακολούθησα θα είχε μείνει μονοπάτι αχαρτογράφητο για εμένα. Τον Χριστόφορο Νικολάου, που με βοήθησε να βρω το βηματισμό μου στο δρόμο που διάλεξα. Τον Δημήτρη Πολυχρονόπουλο, συνάδελφο παλαιότερα, τώρα ερευνητή στο Λονδίνο· εύχομαι η επιτυχία να συνοδεύει κάθε του προσπάθεια.*

*Ιδιαίτερα, ευχαριστώ τον Δρ. Νικόλαο Γλυκό και τη Δρ. Μεταξία Βλάση, που δέχθηκαν να παρακολουθήσουν και να επιβλέψουν αυτή την εργασία, καθώς και όλα τα μέλη της εξεταστικής επιτροπής, που με τα σχόλια και τις παρατηρήσεις τους βοήθησαν να περιοριστούν λάθη και παραλείψεις και να αναδειχθεί η πιθανή αξία της παρούσας διατριβής.*

*Δουλεύοντας στο 'Δημόκριτο', στους πρόποδες του Υμηττού, συναντήθηκα με ανθρώπους που όρισαν τα χρόνια αυτά που πέρασαν· ανθρώπους που με περίμεναν, αλλά και ανθρώπους που αναζητούσα. Τα ονόματά τους, ωστόσο, δεν υπάρχουν στο κείμενο αυτό. Γιατί όσο και αν προσπαθώ να ζυγίσω τα πράγματα, δε ωφελεί· ό,τι και αν γράψω, θα αδικεί ίσως μια διατριβή, μα πιο πολύ θα αδικεί αυτούς τους ίδιους...*

Τα βασικά ευρήματα και συμπεράσματα της παρούσας διατριβής έχουν δημοσιευτεί στο ερευνητικό άρθρο:

Apostolou-Karampelis K, Nikolaou C and Almirantis Y. 2016. A novel skew analysis reveals substitution asymmetries linked to genetic code GC-biases and PolIII a-subunit isoforms. DNA Res. 23(4):353-363.

*DNA Research*, 2016, 23(4), 353–363
doi: 10.1093/dnares/dsw021
Advance Access Publication Date: 26 June 2016
Full Paper

OXFORD

Full Paper

# A novel skew analysis reveals substitution asymmetries linked to genetic code GC-biases and PolIII a-subunit isoforms

**Konstantinos Apostolou-Karampelis[1,*], Christoforos Nikolaou[2], and Yannis Almirantis[1,*]**

[1]Institute of Biosciences and Applications, National Center for Scientific Research "Demokritos", 15310 Athens, Greece, and [2]Computational Genomics Group, Department of Biology, University of Crete, 71409 Heraklion, Greece

*To whom correspondence should be addressed. Tel: +302106503601. Email: apostolou@bio.demokritos.gr (K.A.K.); Email: yalmir@bio.demokritos.gr (Y.A.)

Edited by Dr Yuji Kohara

Received 11 February 2016; Accepted 9 May 2016

# ΣΥΝΤΜΗΣΕΙΣ ΚΑΙ ΑΡΚΤΙΚΟΛΕΞΑ

| | | |
|---|---|---|
| **BER** | **b**ase-**e**xcision **r**epair | επιδιόρθωσης εκτομής βάσης |
| **BPR** | **b**ase-**p**airing **r**ule | κανόνα του ζευγαρώματος των συμπληρωματικών βάσεων |
| **bps** | **b**ase **p**air**s** | ζεύγη βάσεων |
| **CDS** | **c**o**d**ing **s**equence | κωδικές αλληλουχίες |
| **CPD** | **c**yclobutane **p**yrimidine **d**imer | κυκλοβουτανικά διμερή πυριμιδίνης |
| **GGR** | **g**lobal **g**enome **r**epair | σύστημα καθολικής επιδιόρθωσης του γονιδιώματος |
| **IMP** | **i**ntegral **m**embrane **p**roteins | ενσωματωμένες μεμβρανικές πρωτεΐνες |
| **KL** divergence | **K**ullback–**L**eibler divergence | απόκλιση Kullback–Leibler |
| **MFD** | **m**utation **f**requency **d**ecline | πτώσης των μεταλλακτικών ρυθμών |
| **MMR** | **m**is**m**atch **r**epair system | σύστημα επιδιόρθωσης αταίριαστων ζευγών βάσεων |
| **NER** | **n**ucleotide-**e**xcision **r**epair | επιδιόρθωσης εκτομής νουκλεοτιδίων |
| **ORF** | **o**pen **r**eading **f**rame | ανοιχτό πλαίσιο ανάγνωσης |
| ***ori*** | **ori**gin of replication | σημείο έναρξης της αντιγραφής |
| **PolIII** | DNA **pol**ymerase **III** | DNA πολυμεράση III |
| **PR1** | **p**arity **r**ule **1** | ο πρώτος κανόνας της ισοδυναμίας |
| **PR2** | **p**arity **r**ule **2** | ο δεύτερος κανόνας της ισοδυναμίας |
| **Rep** | **rep**lication | αντιγραφή |

| | | |
|---|---|---|
| **rms** | **r**oot **m**ean **s**quare | τετραγωνικός μέσος |
| **RR** | **r**ecombination **r**epair | επιδιόρθωσης μέσω ανασυνδυασμού |
| **ssDNA** | **s**ingle-**s**tranded **DNA** | δίκλωνο DNA στην μονόκλωνη κατάστασή του |
| **TAM** | **t**ranscription-**a**ssociated **m**utations | μεταλλάξεις σχετιζόμενες με τη μεταγραφή |
| **TCR** | **t**ranscription **c**oupled **r**epair | συζευγμένη με τη μεταγραφή επιδιόρθωση (του DNA) |
| **TEC** | **t**ernary **e**longation **c**omplex | τριμερές σύμπλοκο επιμήκυνσης |
| ***ter*** | **ter**minus of replication | σημείο λήξης της αντιγραφής |
| **TLS** | **t**rans**l**esion **s**ynthesis | αντιγραφή του DNA διαμέσου βλάβης |
| **TRCF** | **t**ranscription-**r**epair **c**oupling **f**actor | ο παράγοντας σύζευξης μεταγραφής-επιδιόρθωσης |
| **Trs** | **tr**an**s**cription | μεταγραφή |
| **VSP** | **v**ery-**s**hort **p**atch repair system | επιδιόρθωση πολύ βραχέος τμήματος |

# ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

# ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

***ΠΕΡΙΕΧΟΜΕΝΑ***

# Ι. ΠΕΡΙΛΗΨΗ

Η σύσταση του DNA εξελίσσεται συμμετρικά όταν δεν υπάρχουν ειδικές ανά κλώνο πολώσεις των μεταλλακτικών ρυθμών και των επιλεκτικών πιέσεων. Η συμμετρία των ρυθμών υποκατάσταση αντιστοιχεί σε μία μηδενική υπόθεση για την εξέλιξη του γενετικού υλικού και οδηγεί σε χαρακτηριστικές κανονικότητες της σύστασης του DNA, στην κλίμακα ολόκληρου του γονιδιώματος. Αποκλίσεις από αυτές τις κανονικότητες υποδηλώνουν την παρουσία ασυμμετριών μεταξύ των αντιστρόφως συμπληρωματικών κλώνων.

Η παρούσα εργασία προσφέρει πλήθος ευρημάτων που αποδεικνύουν πως οι συσχετίσεις μεταξύ των 1$^{ης}$ τάξης γειτονικών βάσεων αποτελούν ένα χαρακτηριστικό της σύστασης του DNA που είναι έντονα πολωμένο μεταξύ των κλώνων του DNA. Το γεγονός αυτό συνεπάγεται συστηματικές ασυμμετρίες των υποκαταστάσεων οι οποίες εξαρτώνται από την ταυτότητα των παρακείμενων βάσεων (neighbor-dependend substitutions). Προκειμένου να εντοπίσουμε τέτοιες ασυμμετρίες, εισάγουμε ένα μέτρο που δεν προϋποθέτει την στοίχιση ομόλογων αλληλουχιών, τις *αποκλίσεις των σταθμισμένων συχνοτήτων των δινουκλεοτιδίων*. Οι κατανομές αυτών των αποκλίσεων κατά μήκος των κωδικών αλληλουχιών επιτρέπουν την ανασυγκρότηση των φυλογενετικών σχέσεων των βακτηριών. Συνεπώς, τα πρότυπα των υποκαταστάσεων που εξαρτώνται από την ταυτότητα των 1$^{ης}$ τάξης γειτονικών βάσεων δεν είναι κοινά μεταξύ εξελικτικά απομακρυσμένων οργανισμών, αλλά αντίθετα είναι ανά είδος καθορισμένα (species-specific).

Αναλύοντας τις αποκλίσεις που εμφανίζονται ανά θέση κωδικονίων, σε συνάρτηση και με την ταυτότητα των 1$^{ης}$ τάξης γειτονικών βάσεων, οδηγούμαστε σε σημαντικά συμπεράσματα σχετικά με την προέλευση των ασυμμετριών που εκδηλώνουν οι ρυθμοί υποκατάστασης. Εισάγοντας ένα απλό μοντέλο που περιγράφει την πιθανότητα εμφάνισης κωδικονίων συναρτήσει ενός ελάχιστου αριθμού παραμέτρων που είναι συμμετρικές ως προς τους κλώνους του DNA, υποστηρίζουμε ότι η δομή του γενετικού κώδικα επιβάλλει ασύμμετρα πρότυπα υποκαταστάσεων ως απόκριση στις μεταλλακτικές πιέσεις που κατευθύνουν την

σύσταση των κωδικών περιοχών προς ένα συγκεκριμένο GC περιεχόμενο. Συγκεκριμένα, η οργάνωση του γενετικού κώδικα σε ομάδες συνώνυμων κωδικονίων μπορεί να οδηγεί σε ασυμμετρίες της κατανομή των νουκλεοτιδίων μεταξύ διαφορετικών κωδικών θέσεων, ακόμα και όταν η επιλογή για συγκεκριμένα κωδικόνια και αμινοξέα δεν λαμβάνονται υπόψιν.

Εν συνεχεία, η μελέτης μας καταδεικνύει ότι εγγενείς ασυμμετρίες του καταλυτικού κέντρου της PolIII, που διαφοροποιούν την ενεργότητα ενσωμάτωσης των νουκλεοτιδίων και την επιδιορθωτική ενεργότητα της α-υπομονάδας κατά μήκος των δύο κλώνων της αντιγραφής, επάγουν συστηματικά ειδικές ανά κλώνο πολώσεις των ρυθμών υποκατάστασης, στην κλίμακα ολόκληρου του γονιδιώματος. Επίσης, εξετάζουμε τον ρόλο ποικίλων μηχανισμών τροποποίησης και επιδιόρθωσης του DNA στην διαμόρφωση των παρατηρούμενων αποκλίσεων από το πρότυπο της συμμετρικής εξέλιξης των κλώνων.

# II. ABSTRACT

The second Parity Rule (PR2) corresponds to a null expectation of DNA composition, assuming strand-symmetric evolution. Motivated by the extensive corpus of studies on deviations from PR2, the present thesis focuses on the context-dependency of strand asymmetries, in terms of nearest-neighbor preferences and codon site-specific composition. Based on the analysis of a large collection of bacterial genomes, this thesis provides new insights into strand biased mutation and fixation processes and points out their far reaching evolutionary implications.

To assess the context-dependency of strand asymmetries, the correlations between neighboring bases are accounted for. Dinucleotides are the primary ordering units of DNA bases. First-neighbor correlations are inferred by means of *relative abundances* of dinucleotides, which filter out the effect of the underlying mononucleotide composition. Contrary to what is hitherto accepted, statistically solid evidence is provided which shows that the *relative abundances* of dinucleotides exhibit significant strand asymmetries related to DNA replication and transcription. These asymmetries are quantified by an alignment free index (*dinucleotide relative abundance skews*), which is introduced for the first time by this thesis. This skew index effectively detects chromosomal regions where context-dependent substitutions are strand-biased.

More interestingly, the present analysis demonstrates that *dinucleotide relative abundance skews* are able to retrace phylogenetic relationships between bacteria. It follows that strand biases of first-neighbor correlations do not reflect a universal trend in sequence evolution, but are rather species-specific. Thus, it can be argued that the strand bias of first-neighbor correlations constitute an idiosyncratic genomic feature that is linked to the evolutionary dynamics of the DNA strands.

The uneven distribution of nucleotides among different codon sites has

been widely attributed to functional constraints reflecting the optimization of transcription or translation efficiency and the selective pressures acting at the level of protein function. Nonetheless, position-dependent biases of DNA composition can be the outcome of processes acting strictly at the nucleotide level, without accounting for codon preferences or amino-acid specific constraints in the primary structure of proteins. Through a simple model the present thesis shows that GC biases of synonymous codons suffice to yield patterns of correlations between codon usage and site-specific skews which are qualitatively identical to the ones observed in bacterial genomes. The importance of this finding lies in the realization that the GC-biased structure of genetic code *per se* can induce compositional asymmetries in a position-dependent manner solely as a response to GC mutational pressure. It thus provides the ground for an *a priori* estimation of skew patterns at each codon site given the GC content of the genome. These baseline skews should be taken into account when estimating the expected number of substitutions per site in comparative genomic analysis.

Besides GC content diversity, other causes also contribute to the observed variation of compositional biases along coding sequences, such as transcription- and replication-induced substitution asymmetries. In line with previous experimental data, this study reaffirms that strand asymmetries of single base substitutions is associated to transcription-coupled repair (TCR). Furthermore, it shows that TCR may induce asymmetries of neighbor-dependent substitution patterns. The presence or absence of the transcription-repair coupling factor (TRCF) correlates with a reversion of the observed biases.

In bacteria, both DNA strands are synthesized by the a-subunit dimer of DNA PolIII. Though it has been previously reported that there is no correlation between the α-subunit isoforms and the compositional skews, the herein presented analysis clearly demonstrates that such correlation does exist. Different a-subunit isoforms induce specific strand-biased substitutions on a genome-wide scale, which may be context-dependent, as suggested by the findings of the present study.

This thesis also inquires the role of other DNA repair mechanisms, which are not yet determined to act in a strand-specific manner, in shaping CDS compositional asymmetries. In this context, the present thesis suggests

a framework of *in silico* analysis that may provide valuable insight on the possible strand-biased activity of certain enzymes.

The findings presented in the following study broaden our understanding of strand asymmetries inherent to the dynamics of genome evolution. The corresponding analysis suggests that strand asymmetries can be used to facilitate phylogenetic reconstruction methods. It also shows that strand asymmetries can have a major effect on the long-term molecular evolution of proteins. Moreover, skew analysis provides initial indications for the strand-biased activity of specific molecular systems that interact with the genome.

# III. ΠΡΟΛΟΓΟΣ

Σκοπός της παρούσας διατριβής είναι η μελέτη χαρακτηριστικών του γονιδιώματος, όπως αυτά αποτυπώνονται στη σύσταση της αλληλουχίας του, και η εξαγωγή χρήσιμων συμπερασμάτων σχετικά με τους μηχανισμούς που τα διαμορφώνουν, στη χρονική κλίμακα της εξέλιξης. Εντός κάθε ενός από τους κλώνους του DNA εμφανίζονται χαρακτηριστικά πρότυπα κανονικοτήτων της σύστασής του, τα οποία παρέχουν το έδαφος για την μελέτη της εξελικτικής δυναμικής του γενετικού υλικού. Τα πρότυπα αυτά προϋποθέτουν την συμμετρική εξέλιξη των κλώνων του DNA. Σε αυτό το πλαίσιο, η ύπαρξη συμμετρίας δηλώνει ότι οι ρυθμοί υποκατάστασης των αντιστρόφως συμπληρωματικών συστατικών του DNA είναι ίσοι κατά μήκος καθενός από τους δύο κλώνους ξεχωριστά. Αποκλίσεις από τη συμμετρία μπορούν εύκολα να υπολογιστούν και να χρησιμοποιηθούν ως εργαλεία που μας επιτρέπουν να εστιάσουμε και να αναλύσουμε τη φύση και τις επιμέρους πτυχές των διαδικασιών που διαμορφώνουν το γονιδίωμα. Παραφράζοντας τη γνωστή διατύπωση του Pierre Curie 'C'est la dissymétrie qui crée le phénomène', μπορούμε να πούμε ότι *είναι η ασυμμετρία που δημιουργεί τα φαινόμενα*, της οποίας το έντονο αποτύπωμα στη σύσταση του DNA ανιχνεύουμε στην παρούσα μελέτη.

Στη σχετική βιβλιογραφία ανευρίσκεται πλήθος εργασιών που έχουν ως αντικείμενό τους τις ασυμμετρίες μεταξύ των κλώνων του DNA. Οι μελέτες αυτές επικεντρώνονται σε αποκλίσεις μεταξύ των συχνοτήτων των συμπληρωματικών βάσεων. Επίσης, έχει εξεταστεί η ασύμμετρη κατανομή στους κλώνους του DNA συγκεκριμένων ολιγονουκλεοτιδίων, με γνωστή ή πιθανολογούμενη λειτουργικότητα. Οι ερμηνείες που έχουν προταθεί καλύπτουν ένα ευρύτατο φάσμα μηχανισμών, από τις ειδικές ανά κλώνο υποκαταστάσεις που επάγουν οι εγγενείς ασυμμετρίες της διχάλας της αντιγραφής έως την επιλογή στην χρήση κωδικονίων λόγω συσχέτισή τους με την ενδοκυτταρική συγκέντρωση των αντίστοιχων tRNAs. Παρότι είναι σαφές πως στην εμφάνιση των ενδοκλωνικών αποκλίσεων εμπλέκονται ποικίλοι μηχανισμοί, με επικαλυπτόμενη ή αντικρουόμενη δράση ως προς την ένταση και τη φορά των ασυμμετριών, η

αναζήτηση απλών και συνεκτικών ερμηνευτικών σχημάτων αποτέλεσε πρόκληση για πολλούς ερευνητές. Την παράδοση αυτή επιχειρούμε να εμπλουτίσουμε, αναδεχόμενοι τους κινδύνους που μια τέτοια προσπάθεια εμπεριέχει.

Στην παρούσα διατριβή εξετάζουμε εάν και κατά πόσο οι ασυμμετρίες των κλώνων του DNA επεκτείνονται από το επίπεδο της νουκλεοτιδικής σύστασης σε εκείνο της διάταξης των βάσεων. Επίσης, μελετάμε τη συσχέτιση των ενδοκλωνικών αποκλίσεων με τη φυλογένεση των βακτηρίων, προκειμένου να διαπιστώσουμε εάν οι ασυμμετρίες του DNA συγκροτούνται σε πρότυπα καθολικά απαντώμενα μεταξύ των οργανισμών ή αντιθέτως είναι ανά είδος καθορισμένες. Αναζητώντας τις πιθανές αιτίες αυτών των ασυμμετριών, ερευνούμε τις συσχετίσεις των αποκλίσεων στις κωδικές περιοχές με τα πρότυπα χρήσης κωδικονίων. Τα ευρήματά μας συνηγορούν υπέρ της άποψης σύμφωνα με την οποία διαδικασίες που δρουν στο επίπεδο της νουκλεοτιδικής σύστασης είναι ικανές να παράγουν τις παρατηρούμενες ασυμμετρίες, δίχως να λαμβάνεται υπόψιν η σύσταση των κωδικοποιούμενων πρωτεϊνών. Στο πλαίσιο αυτό, συνάγονται ενδιαφέροντα συμπεράσματα για την ίδια την οργάνωση του γενετικού κώδικα. Τα αποτελέσματα της ανάλυσής μας υποδηλώνουν ότι η ασύμμετρη εξέλιξη του γονιδιώματος *per se* δύναται να διαμορφώνει σε σημαντικό βαθμό την εξελικτική δυναμική του πρωτεώματος. Παράλληλα, τα αποτελέσματα αυτά ανοίγουν ένα νέο μονοπάτι για διερεύνηση της πιο πάνω υπόθεσης. Τέλος, μέσω των αποκλίσεων μελετάμε πτυχές των μηχανισμών αντιγραφής, τροποποίησης και επιδιόρθωσης του DNA. Όπως φανερώνει η ανάλυσή μας που αφορά την α-καταλυτική υπομονάδα της πολυμεράσης Pol III, οι ενδοκλωνικές αποκλίσεις αποτελούν χρήσιμο εργαλείο που μας προσφέρει σημαντικές ενδείξεις σχετικά με την ασύμμετρη δράση διαφόρων μορίων κατά μήκος των κλώνων του DNA.

# IV. ΕΙΣΑΓΩΓΗ

Η εξέλιξη της σύστασης του γονιδιώματος καθορίζεται από την συνδυασμένη δράση μεταλλακτικών και επιλεκτικών διαδικασιών (Lobry & Lobry 1999). Όταν οι ρυθμοί μεταλλάξεων και οι επιλεκτικές πιέσεις δεν εμφανίζουν πολώσεις μεταξύ των κλώνων του DNA, τότε από τον κανόνα του ζευγαρώματος των συμπληρωματικών βάσεων (base-pairing rule, BPR) συνάγεται ότι οι ρυθμοί υποκατάστασης των συμπληρωματικών βάσεων είναι ίσοι κατά μήκος του κάθε κλώνου ξεχωριστά, όπως δηλώνει ο 1$^{ος}$ *κανόνας της ισοδυναμίας* (parity rule 1, PR1) (Sueoka 1995). Δυνάμει του PR1, η ενδοκλωνική σύσταση ενός χρωμοσώματος αναμένεται να φτάσει σε μια κατάσταση ισορροπίας, στην οποία ισχύουν οι σχέσεις [A] = [T] και [G] = [C]. Οι σχέσεις αυτές είναι γνωστές ως ο 2$^{ος}$ *κανόνας της ισοδυναμίας* (parity rule 2, PR2) (Sueoka 1995, Lobry 1995), ενώ αναφέρονται επίσης και ως ο 2$^{ος}$ *κανόνας του Chargaff*.

Τα περισσότερα βακτηριακά γονιδιώματα συμμορφώνονται προς τον PR2, στην κλίμακα του χρωμοσώματος (Prabhu 1993, Bell & Forsdyke 1999, Mitchell & Bridge 2006). Ωστόσο, εμφανίζουν χαρακτηριστικές αποκλίσεις από τον PR2 στην τοπική κλίμακα, όπως καταδεικνύουν προγενέστερες μελέτες (Lobry 1996, Mrázek & Karlin 1998, Frank & Lobry 1999, Rocha et al. 1999, Tillier & Collins 2000, Lobry & Sueoka 2002, Nikolaou & Almirantis 2005, Rocha et al. 2006, Morton & Morton 2007, Charneski et al. 2011). Οι αποκλίσεις από τον PR2 προσφέρουν την δυνατότητα μελέτης θεμελιωδών διαδικασιών, όπως η αντιγραφή, η μεταγραφή και η επιδιόρθωση του DNA, που δρουν με διαφορετικό τρόπο κατά μήκος καθενός από τους δύο κλώνους και ως εκ τούτου αναμένεται να επάγουν συστηματικές ασυμμετρίες στους ρυθμούς υποκατάστασης. Οι ασυμμετρίες αυτές οδηγούν σε αποκλίσεις από την μηδενική υπόθεση της συμμετρικής εξέλιξης των κλώνων του DNA, οπότε οι σχέσεις [A] = [T] και [G] = [C] παύουν να ισχύουν (Rocha & Danchin 2001, Lobry & Sueoka 2002, Danchin 2003, Klasson & Andersson 2006, Nikolaou & Almirantis 2006, Necşulea & Lobry 2007, Rocha 2008, Charneski et al. 2011).

Ο Lobry (1996) πρότεινε ότι οι παρατηρούμενες ασυμμετρίες της σύστασης

του DNA οφείλονται σε διαφορές της *σχετιζόμενης με την αντιγραφή* μεταλλαξιγένεσης μεταξύ οδηγού και συνοδού κλώνου, και κατοπινές μελέτες ενισχύουν την σχετική επιχειρηματολογία (Rocha et al. 1999, Worning et al. 2006). Οι εν λόγω ασυμμετρίες αποδίδονται επίσης στην άνιση κατανομή των κωδικών κλώνων μεταξύ των δύο κλώνους της αντιγραφής (Bell & Forsdyke 1999, Lopez & Philippe 2001, Nikolaou & Almirantis 2005, Necşulea & Lobry 2007, Charneski et al. 2011). Ο εμπλουτισμό του οδηγού κλώνου σε κωδικούς κλώνους, σε συνδυασμό με την *συζευγμένη με τη μεταγραφή* μεταλλαξιγένεση και επιδιόρθωση (Francino et al. 1996) καθώς επίσης και τις πολώσεις στην χρήση κωδικονίων (Ikemura 1981, Gouy & Gautier 1982, Bulmer 1991, Xia 1998), μπορεί να οδηγεί σε αποκλίσεις από τον PR2 στην κλίμακα ολόκληρου του γονιδιώματος. Στο πλαίσιο αυτό, η επιλογή στην χρήση κωδικονίων και αμινοξέων καθορίζει σε σημαντικό βαθμό τις παρατηρούμενες ασυμμετρίες της σύστασης των κωδικών περιοχών.

Ωστόσο, σύμφωνα με εργασίες στις οποίες μελετήθηκαν τα πρότυπα συσχέτισης της σύστασης του DNA και των πρωτεϊνών (Sueoka 1961) καθώς και τα προφίλ υδροφοβικότητας του πρωτεώματος συναρτήσει της χρήσης κωδικονίων (D'Onofrio et al. 1999), η νουκλεοτιδική σύσταση των γονιδίων, όπως αυτή εκφράζεται στο GC περιεχόμενό τους, ενδέχεται να είναι η κινητήρια δύναμη που διαμορφώνει καθοριστικά το αμινοξικό περιεχόμενο των κωδικοποιούμενων πρωτεϊνών. Συνεπώς, οι μεταλλακτικές πιέσεις, όπως αντανακλώνται στο GC περιεχόμενο και τις ειδικές ανά κλώνο αποκλίσεις του DNA, δύνανται να επηρεάζουν σημαντικά την σύσταση των πρωτεϊνών στην χρονική κλίμακα της εξέλιξης.

Παρότι στην σχετική βιβλιογραφία ανευρίσκεται πλήθος εργασιών που πραγματεύονται τις αποκλίσεις από τον PR2 στο επίπεδο των μονονουκλεοτιδίων (*μονονουκλεοτιδικές αποκλίσεις*), η κατανομή των ολιγονουκλεοτιδίων μεταξύ των κλώνων του DNA αποτέλεσε το αντικείμενο ενός περιορισμένου αριθμού μελετών (Salzberg et al. 1998, Mascher 2013). Αξίζει μάλιστα να σημειωθεί το περιορισμένο ενδιαφέρον για την αναζήτηση τυχόν ασυμμετριών στο επίπεδο των δινουκλεοτιδίων, παρότι αυτά αποτελούν την πρωταρχική μονάδα διάταξης των βάσεων (primary ordering unit) του DNA (Karlin 1998).

Τα δινουκλεοτίδια εμφανίζουν συστηματικές προτιμήσεις υπο- ή υπερ-εκπροσώπησης, όπως έχει ήδη καταγραφεί από πρωτοπόρες μελέτες (Josse et al. 1961, Nussinov 1981). Η υπο- ή υπερ-εκπροσώπηση των δινουκλεοτιδίων

ποσοτικοποιείται από τις αντίστοιχες *σταθμισμένες συχνότητες* και αντανακλά τις προτιμήσεις που εμφανίζει κάθε νουκλεοτίδιο για τα γειτονικά του. Οι σταθμισμένες δινουκλεοτιδικές συχνότητες εκφράζουν την σχετική αφθονία των δινουκλεοτιδίων, σε αντίθεση με τις παρατηρούμενες συχνότητες εμφάνισής τους. Σε κάθε έναν από τους κλώνους του DNA η σταθμισμένη συχνότητα ενός δεδομένου δινουκλεοτιδίου είναι κατά προσέγγιση ίση με εκείνη του αντιστρόφως συμπληρωματικού του (Nussinov 1984, Shioiri & Takahata 2001, Baisnee 2002), γεγονός που υποδηλώνει την ύπαρξη συμμετρίας των κλώνων του DNA στο επίπεδο των συσχετίσεων μεταξύ των 1$^{ης}$ τάξεως γειτονικών βάσεων. Οι Mrázek και Karlin (1998) αναζήτησαν τυχόν αποκλίσεις από την συμμετρία των σταθμισμένων δινουκλεοτιδικών συχνοτήτων. Ωστόσο, βασιζόμενοι στον περιορισμένο αριθμό των έως τότε διαθέσιμων γονιδιωμάτων που είχαν πλήρως αλληλουχηθεί, συμπέραναν ότι "η σχετική αφθονία των δινουκλεοτιδίων τείνει να είναι συμμετρική και ιδιαίτερα σταθερή όσον αφορά τον οδηγό και το συνοδό κλώνο, παρά την ειδική ανά κλώνο ασυμμετρία της σύστασης" του DNA.

Η παρούσα εργασία προσφέρει ισχυρές αποδείξεις για την ύπαρξη συστηματικών αποκλίσεων από την ισοδυναμία των σταθμισμένων συχνοτήτων των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων, στην τοπική κλίμακα. Οι αποκλίσεις αυτές είναι στατιστικά σημαντικές και αποτελούν ένα εμμενές χαρακτηριστικό της σύστασης του DNA. Οι αποκλίσεις των σταθμισμένων συχνοτήτων που εμφανίζονται κατά μήκος των κωδικών περιοχών συσχετίζονται με την φυλογένεση των βακτηρίων. Μάλιστα, η απόδοσή τους στην φυλογενετική ανασυγκρότηση είναι αντίστοιχη, και σε ορισμένες περιπτώσεις μεγαλύτερη, από εκείνη άλλων ποσοτήτων που περιγράφουν την σύσταση του DNA. Η ανάλυσή μας έχει ως σκοπό την αναζήτηση των αιτιών που βρίσκονται στην ρίζα των παρατηρούμενων αποκλίσεων. Τα ευρήματά μας υποστηρίζουν ότι η οργάνωση του γενετικού κώδικα σε ομάδες συνωνύμων και συγκεκριμένα οι πολώσεις του GC περιεχομένου εντός αυτών των ομάδων δύνανται να διαμορφώνουν τις ασυμμετρίες της σύστασης των κωδικών περιοχών, ακόμα και όταν δεν λαμβάνεται υπόψιν η επιλογή στο επίπεδο των κωδικοποιούμενων αμινοξέων. Επίσης, εξετάζουμε τις ειδικές ανά κλώνο πολώσεις των υποκαταστάσεων που επάγουν οι μηχανισμοί της αντιγραφής, της μεταγραφής και της επιδιόρθωσης του DNA στην κλίμακα ολόκληρου του γονιδιώματος. Η ανάλυση των αποκλίσεων παρέχει αρχικές ενδείξεις που μπορούν να κατευθύνουν την πειραματική διερεύνηση της ειδικής ανά κλώνο δράσης συγκεκριμένων μοριακών μηχανισμών.

# 1. ΘΕΩΡΗΤΙΚΟ ΜΕΡΟΣ

Η αναζήτηση μοτίβων που επαναλαμβάνονται συστηματικά στη Φύση ανοίγει ένα παράθυρο για την από μέρους μας κατανόηση της ουσίας και των λειτουργιών της. Έτσι, η στοιχειομετρική ισότητα των συμπληρωματικών βάσεων, ήτοι [A] = [T] και [G] = [C], που προσδιορίστηκε βιοχημικά κατά τη μελέτη του γενετικού υλικού, συνέβαλλε καθοριστικά στον προσδιορισμό της δομής του δίκλωνου DNA. Η ισότητα αυτή, που αποτελεί απαράβατο κανόνα στο επίπεδο του δίκλωνου μορίου, σε πολλές περιπτώσεις ισχύει και σε κάθε έναν από τους κλώνους του ξεχωριστά. Προϋπόθεση είναι σε κάθε κλώνο του DNA οι ρυθμοί υποκατάστασης των συμπληρωματικών βάσεων να ισούνται μεταξύ τους. Συνεπώς, απλοί στοιχειομετρικοί υπολογισμοί κατά μήκος καθενός εκ των δύο κλώνων μας επιτρέπουν να αντλήσουμε σημαντικές πληροφορίες για τους ρυθμούς υποκατάστασης και συνεπώς για την εξελικτική δυναμική του γενετικού υλικού.

Μεταλλακτικές διαδικασίες και επιλεκτικές πιέσεις που κατά μήκος κάθε κλώνου επάγουν ίσους ρυθμούς αντιστρόφως συμπληρωματικών υποκαταστάσεων, θεωρούμε ότι δρουν συμμετρικά στο δίκλωνο μόριο του DNA (βλ. Εικόνα 1). Αποκλίσεις από την ενδοκλωνική ισότητα των συμπληρωματικών βάσεων υποδηλώνουν την παρουσία μηχανισμών με ασύμμετρη δράση, που οδηγούν σε ρυθμούς υποκατάστασης πολωμένους προς έναν από τους δύο κλώνους των χρωμοσωμάτων. Τέτοιοι μηχανισμοί μπορεί να είναι επιλεκτικοί ή μεταλλακτικοί (βλ. ενότητες 1.4.1 και 1.4.2, αντίστοιχα). Στην κατηγορία των επιλεκτικών μηχανισμών συγκαταλέγονται:

(*α*) η *επιλογή των συχνοτήτων εμφάνισης αμινοξέων* και *κωδικονίων*, που επάγει ασυμμετρίες μεταξύ κωδικού και μεταγραφόμενου κλώνου

(*β*) η *επιλογή του προσανατολισμού των γονιδίων*, που οδηγεί στον εμπλουτισμό του οδηγού κλώνου σε κωδικούς κλώνους, με αποτέλεσμα οι ασυμμετρίες των κωδικών περιοχών να εκδηλώνονται και ως ασυμμετρίες μεταξύ οδηγού και συνοδού κλώνου, και

(*γ*) η *επιλογή της θέσης και του προσανατολισμού ολιγονουκλεοτιδίων* που λειτουργούν ως αλληλουχίες σηματοδότησης.

Στην κατηγορία των μεταλλακτικών μηχανισμών περιλαμβάνονται:

(*α*) πρωτίστως *οι μηχανισμοί της αντιγραφής και της μεταγραφής*, οι οποίοι είναι εγγενώς ασύμμετροι και συνεπώς μπορούν (*i*) είτε αφεαυτές να επάγουν ασύμμετρους ρυθμούς υποκαταστάσεων (*ii*) είτε να πολώνουν τους ρυθμούς αυθόρμητων μεταλλάξεων προς τον έναν από τους δύο κλώνους (οδηγό ή συνοδό και κωδικό ή μεταγραφόμενο, αντίστοιχα), και

(*β*) μηχανισμοί τροποποίησης και επιδιόρθωσης του DNA.

Η μελέτη των αποκλίσεων από τις ενδοκλωνικές ισότητες που προαναφέραμε μας προσφέρει την δυνατότητα να αξιολογήσουμε τη συμβολή καθενός από τους παραπάνω μηχανισμούς στην εξέλιξη συγκεκριμένων πτυχών/χαρακτηριστικών της σύστασης του γονιδιώματος. Αλλά και αντίστροφα, η μελέτη της σύστασης του γονιδιώματος σε όρους ενδοκλωνικών αποκλίσεων μας επιτρέπει να συνάγουμε χρήσιμα συμπεράσματα σχετικά με πιθανές ασυμμετρίες της δράσης θεμελιωδών μοριακών και εξελικτικών μηχανισμών (βλ. ενότητα 3.12).

Εκτός από τα πρότυπα συμμετρίας και τις αντίστοιχες αποκλίσεις που εκδηλώνονται στο επίπεδο των συχνότητες των συμπληρωματικών βάσεων, χαρακτηριστικά μοτίβα εμφανίζονται και στο επίπεδο της διάταξης των βάσεων κατά μήκος της αλληλουχίας του DNA. Η πρωταρχική μονάδα διάταξης των βάσεων είναι τα δινουκλεοτίδια. Η συχνότητα εμφάνισης κάθε δινουκλεοτιδίου καθορίζεται τόσο από τη συχνότητα των μονονουκλεοτιδίων που το απαρτίζουν όσο και από την ύπαρξη συσχετίσεων μεταξύ τους. Σταθμίζοντας τη συχνότητα εμφάνισης των δινουκλεοτιδίων κατά τρόπο ώστε να απαλείφεται η επίδραση της συχνότητας των αντίστοιχων μονονουκλεοτιδίων (βλ. ενότητες 1.8, 2.4), μπορούμε να εστιάσουμε στις συσχετίσεις που εκδηλώνονται μεταξύ των κοντινότερων γειτονικών βάσεων του DNA. Οι συσχετίσεις αυτές αντανακλούν:

(*α*) τις φυσικοχημικές ιδιότητες των βάσεων, οι οποίες, όταν γειτνιάζουν, αλληλεπιδρούν και καθορίζουν την στερεοχημική διαμόρφωση (conformation) των δινουκλεοτιδίων και την ενέργεια πακεταρίσματός τους (stacking energy) (βλ. ενότητα 1.8.1), και

(*β*) τον τύπο και τον ρυθμό των υποκαταστάσεων που εξαρτώνται από τις εκάστοτε γειτονικές τους βάσεις (neighbor-dependent substitution) και αποτελούν τη συνισταμένη τόσο επιλεκτικών όσο και μεταλλακτικών πιέσεων (βλ. ενότητα 1.8.2)

Οι σταθμισμένες δινουκλεοτιδικές συχνότητες αποκαλύπτουν συστηματικά πρότυπα υπο- και υπερ-εκπροσώπησης των δινουκλεοτιδίων. Τα πρότυπα αυτά είναι στενά συνδεδεμένα με την εξελικτική πορεία του κάθε οργανισμού, σε

βαθμό που να θεωρείται ότι αποτελούν χαρακτηριστικές υπογραφές του κάθε γονιδιώματος (βλ. ενότητα 1.8.3). Σύμφωνα με την έως τώρα γνώση μας, οι σταθμισμένες δινουκλεοτιδικές συχνότητες θεωρείται ότι εξελίσσονται συμμετρικά ως προς τους κλώνους του DNA, παρά τις έντονες ασυμμετρίες της σύστασής τους σε όρους συχνοτήτων εμφάνισης των νουκλεοτιδικών βάσεων (βλ. ενότητα 1.8.3.3). Στην παρούσα διατριβή ανασκευάζουμε αυτόν τον ισχυρισμό και αποδεικνύουμε ότι οι σταθμισμένες συχνότητες των δινουκλεοτιδίων αποτελούν το αντικείμενο ασύμμετρων εξελικτικών διαδικασιών (βλ. ενότητες 3.3-3.5, 3.3, 3.8) που είναι ανά είδος καθορισμένες (species-specific, βλ. ενότητα 3.9).

Στις αναλύσεις της σύστασης του DNA σημαντική θέση κατέχει η μελέτη του περιεχομένου του σε κατάλοιπα γουανίνης και κυτοσίνης (GC περιεχόμενο). Το ζευγάρωμα των συμπληρωματικών βάσεων στο δίκλωνο DNA δηλώνει ότι οι συχνότητες των καταλοίπων G και C στον έναν κλώνο ισούνται με τις συχνότητες των καταλοίπων C και G, αντίστοιχα, στον αντιστρόφως συμπληρωματικό κλώνο. Συνεπώς το GC περιεχόμενο, που ορίζεται ως το άθροισμα των συχνοτήτων των G και C, είναι το ίδιο και για τους δύο κλώνους του DNA. Με άλλα λόγια, το GC περιεχόμενο είναι από τη φύση του ένα συμμετρικό χαρακτηριστικό της σύστασης του γονιδιώματος. Πλήθος ερμηνειών, συχνά αντιφατικών, έχουν προταθεί προκειμένου να εξηγήσουν την ποικιλότητα που εμφανίζει το GC περιεχόμενο μεταξύ διαφορετικών γονιδιωμάτων (βλ. ενότητα 1.9). Εδώ, εστιάζουμε στις μεταλλάξεις που κατευθύνουν το GC περιεχόμενο των χρωμοσωμάτων προς χαμηλότερες ή υψηλότερες τιμές (GC κατευθύνουσες μεταλλακτικές πιέσεις). Οι μεταλλάξεις αυτές διαδραματίζουν καθοριστικό ρόλο στη διαφοροποίηση (diversification) των βακτηριακών γονιδιωμάτων (βλ. ενότητες 1.9.1, 1.9.2).

Μεταλλακτικές πιέσεις που καθοδηγούν την εξελικτική δυναμική του γονιδιώματος, διαμορφώνουν σε μεγάλο βαθμό τα πρότυπα χρήσης κωδικονίων, αλλά και το ίδιο το αμινοξικό περιεχόμενο των πρωτεϊνών. Στις πιέσεις αυτές συγκαταλέγονται:

(*α*) μεταλλάξεις που είναι πολωμένες προς έναν από τους δύο κλώνους του DNA, και συνεπώς επάγουν ασυμμετρίες στην εξέλιξη του DNA (βλ. ενότητα 1.9.4), και

(*β*) GC κατευθύνουσες μεταλλακτικές πιέσεις που καθορίζουν το GC περιεχόμενο του γονιδιώματος, το οποίο είναι καθαυτό συμμετρικό ως προς τους κλώνους του DNA (βλ. ενότητες 1.9.3, 1.9.5)

Τα ευρήματα της παρούσας διατριβής συγκεράζουν τους δύο αυτούς παράγοντες, καταδεικνύοντας ότι το GC περιεχόμενο των χρωμοσωμάτων, αν και συμμετρικό, συμβάλλει καθοριστικά στην ασύμμετρη εξέλιξη των κωδικών περιοχών του DNA (βλ. ενότητα 3.10).

## **1.1** *Κανονικότητες στη σύσταση του DNA – οι κανόνες του Chargaff*

Η παρούσα μελέτη έχει ως αντικείμενό της την ανάλυση κανονικοτήτων συμμετρίας μεταξύ των συμπληρωματικών κλώνων που εμφανίζονται στο επίπεδο της σύστασης του γενετικού υλικού των βακτηρίων. Φιλοδοξεί δε να συμβάλλει στην κατανόηση των μηχανισμών εκείνων που διαμορφώνουν τη σύσταση των αλληλουχιών DNA σύμφωνα με ή σε αντίθεση προς τις εν λόγω κανονικότητες. Η αναγνώριση και ποσοτικοποίηση προτύπων της νουκλεοτιδικής σύστασης που αποκλίνουν συστηματικά από τα τυχαίως αναμενόμενα παρέχουν χρήσιμες πληροφορίες σχετικά με την εξελικτική δυναμική του DNA.

Η συστηματική εμφάνιση συμμετριών στο DNA είναι παράγωγο ορισμένων θεμελιακών ιδιοτήτων του γενετικού υλικού, όπως είναι οι φυσικοχημικές ιδιότητες των μονομερών που το απαρτίζουν (συμπληρωματικότητα των βάσεων) και ο ημι-συντηρητικός μηχανισμός του αναδιπλασιασμού του. Οι πρώτες ισχυρές ενδείξεις για την ύπαρξη τέτοιων κανονικοτήτων προέκυψαν από πρωτοπόρες βιοχημικές μελέτες που πραγματοποίησαν ο Chargaff και οι συνεργάτες του ήδη από τη δεκαετία του 1950 (Karkas et al. 1968), πριν ακόμα γίνει δυνατή η άμεση ταυτοποίηση κάθε μίας ξεχωριστά από τις νουκλεοτιδικές βάσεις μέσω της αλληλούχησης του γενετικού υλικού. Η πλέον γνωστή κανονικότητα που απαντάται στο επίπεδο της σύστασης του δίκλωνου DNA είναι η ισομοριακότητα των συμπληρωματικών βάσεων [A] = [T] και [G] = [C] (1$^{ος}$ κανόνας του Chargaff) (Zamenhof *et al*. 1952). Η ερμηνεία αυτών των ισοτήτων δόθηκε το 1953, όταν οι Watson και Crick προσδιόρισαν την δομή του DNA (Watson & Crick 1953).

Αντίστοιχες κανονικότητες εμφανίζονται και στο επίπεδο της σύστασης των μεμονωμένων κλώνων του δίκλωνου DNA (ενδοκλωνική σύσταση). Οι πρώτες σχετικές μελέτες (Chargaff 1951, Lin & Chargaff 1967, Karkas et al. 1968,

Rudner et al. 1968a, Rudner et al. 1968b, Karkas et al. 1970, Chargaff 1979, Magasanik & Chargaff 1989) εντόπισαν ότι για κάθε έναν από τους αντιστρόφως συμπληρωματικούς κλώνους του DNA, το άθροισμα των 6-αμινο βάσεων (A + C) ισούται με αυτό των 6-οξο βάσεων (G + T). Ακόλουθα πειράματα κατέδειξαν ότι *σε καθέναν από τους δύο κλώνους του DNA* υπάρχει μια ισχυρή τάση τα κατάλοιπα αδενίνης (A) να ισούνται με εκείνα της θυμίνης (T) και, αντίστοιχα, τα κατάλοιπα γουανίνης (G) να ισούνται με εκείνα της κυτοσίνης (C) (Fickett et al. 1992). Η τάση αυτή, γνωστή και ως ο 2ος κανόνας του Chargaff, συσχετίστηκε με ένα *πρότυπο συμμετρικής εξέλιξης των αντιστρόφως συμπληρωματικών κλώνων* (Lobry 1995). Ας σημειωθεί πως ενώ ο 1ος κανόνας του Chargaff αποτελεί μια αιτιοκρατική σχέση σαφώς καθορισμένη από το ζευγάρωμα των συμπληρωματικών βάσεων της διπλής έλικας, ο 2ος κανόνας του Chargaff αντιστοιχεί στο αποτέλεσμα στοχαστικών διαδικασιών οι οποίες συγκλίνουν στις κατά προσέγγιση ισότητες [A] ~ [T] και [G] ~ [C].

## **1.2** *Οι δύο κανόνες της ισοδυναμίας – παρουσίαση και ερμηνεία*

Οι μεταλλακτικές και επιλεκτικές διαδικασίες που δρουν στο γονιδίωμα καθορίζουν με έναν πολυσύνθετο τρόπο την εξέλιξη της σύστασης του DNA (Lobry & Lobry 1999). Στο επίπεδο του δίκλωνου DNA μπορούν να πραγματοποιηθούν συνολικά 24 διαφορετικοί τύποι υποκατάστασης βάσεων. Συγκεκριμένα, έστω *i* και *j* οι δύο αντιστρόφως συμπληρωματικοί κλώνοι του DNA. Τότε, οι πιθανές υποκαταστάσεις, με ρυθμούς $r_{X \to Y}$ όπου X,Y ∈ (A,T,G,C), είναι οι ακόλουθες:

$$
\begin{aligned}
&(r_{A \to T})_i = (r_{T \to A})_j,\ (r_{T \to A})_i = (r_{A \to T})_j,\ (r_{G \to C})_i = (r_{C \to G})_j,\\
&(r_{C \to G})_i = (r_{G \to C})_j,\ (r_{A \to G})_i = (r_{T \to C})_j,\ (r_{T \to C})_i = (r_{A \to G})_j,\\
&(r_{A \to C})_i = (r_{T \to G})_j,\ (r_{T \to G})_i = (r_{A \to C})_j,\ (r_{G \to A})_i = (r_{C \to T})_i,\\
&(r_{C \to T})_j = (r_{G \to A})_j,\ (r_{G \to T})_i = (r_{C \to A})_i,\ (r_{C \to A})_j = (r_{G \to T})_j \qquad (1)
\end{aligned}
$$

Η ισότητα των ανά ζεύγη αναγραφόμενων ρυθμών υποκατάστασης προκύπτει από τον κανόνα του ζευγαρώματος των συμπληρωματικών βάσεων (base-pairing rule, BPR). Κατά συνέπεια οι συνολικά 24 δυνατοί τύποι υποκατάστασης ανάγονται σε 12 διαφορετικούς ρυθμούς υποκατάστασης. *Όταν δεν υπάρχουν ειδικές ανά κλώνο πολώσεις των μεταλλακτικών ρυθμών και των επιλεκτικών πιέσεων, όταν δηλαδή οι δύο κλώνοι του DNA εξελίσσονται συμμετρικά, τότε ο κάθε συγκεκριμένος τύπος υποκαταστάσεων πραγματοποιείται με τον ίδιο ρυθμό σε κάθε έναν από τους αντιστρόφως συμπληρωματικούς κλώνους του DNA* (βλ. Εικόνα 1a,b). Δηλαδή ισχύουν οι εξής ισότητες:

$$(r_{A\to T})_i = (r_{A\to T})_j,\ (r_{T\to A})_i = (r_{T\to A})_j,\ (r_{G\to C})_i = (r_{G\to C})_j,$$

$$(r_{C\to G})_i = (r_{C\to G})_j,\ (r_{A\to G})_i = (r_{A\to G})_j,\ (r_{T\to C})_i = (r_{T\to C})_j,$$

$$(r_{A\to C})_i = (r_{A\to C})_j,\ (r_{T\to G})_i = (r_{T\to G})_j,\ (r_{G\to A})_i = (r_{G\to A})_j,$$

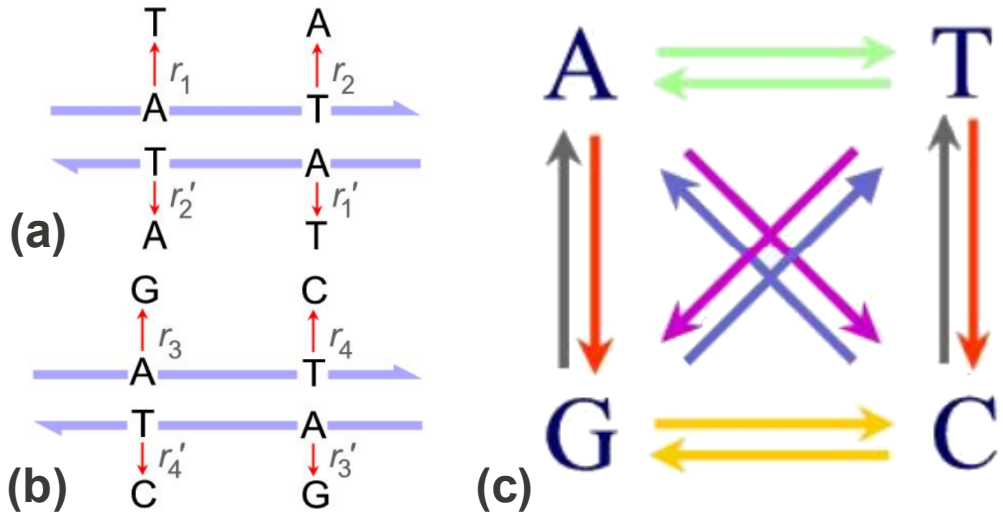
$$(r_{C\to T})_j = (r_{C\to T})_i,\ (r_{G\to T})_i = (r_{G\to T})_j,\ (r_{C\to A})_j = (r_{C\to A})_i \qquad (2)$$

Ο συνδυασμός των σχέσεων (1) και (2) συνεπάγεται πως όταν οι δύο κλώνοι του DNA εξελίσσονται συμμετρικά, οι ακόλουθοι ρυθμοί υποκατάστασης είναι ανά ζεύγη ίσοι, *ανεξαρτήτως του κλώνου στον οποίο λαμβάνουν χώρα*:

$$r_{A\to T} = r_{T\to A},\ r_{G\to C} = r_{C\to G},\ r_{A\to G} = r_{T\to C},$$

$$r_{A\to C} = r_{T\to G},\ r_{G\to A} = r_{C\to T},\ r_{G\to T} = r_{C\to A} \qquad (3)$$

Η σχέση (3) είναι γνωστή ως *ο πρώτος κανόνας της ισοδυναμίας* (parity rule 1, PR1) και συνεπάγεται την εμφάνιση συγκεκριμένων κανονικοτήτων στο επίπεδο στη σύστασης ενός εκάστου των κλώνων του DNA (Lobry 1995, Sueoka 1995). Για μία σχηματική επεξήγηση του PR1, βλ. Εικόνα 1.



**Εικόνα 1.** Ο πρώτος κανόνας της ισοδυναμίας (parity rule 1, PR1). **(a,b)** Σχηματική αναπαράσταση των ρυθμών υποκατάστασης ($r$) στο δίκλωνο μόριο του DNA. Ο κανόνας του ζευγαρώματος των συμπληρωματικών βάσεων (BPR) συνεπάγεται ότι ο ρυθμός υποκατάστασης μίας βάσης, Χ, από μία άλλη, Υ, στον έναν κλώνο του DNA ισούται με το ρυθμό υποκατάστασης της συμπληρωματικής βάσης της Χ από τη συμπληρωματική βάση της Υ, στον άλλο κλώνο. Έτσι, στο παράδειγμά μας, ισχύει ότι $r_1 = r'_2$, $r_2 = r'_1$, $r_3 = r'_4$ και $r_4 = r'_3$ [βλ. σχέσεις (1)]. Όταν οι δύο κλώνοι του DNA εξελίσσονται συμμετρικά, τότε ο κάθε συγκεκριμένος τύπος υποκαταστάσεων πραγματοποιείται με τον ίδιο ρυθμό σε κάθε έναν από τους αντιστρόφως συμπληρωματικούς κλώνους. Συνεπώς, στο παράδειγμά μας, ισχύει ότι $r_1 = r'_1$ , $r_2 = r'_2$, $r_3 = r'_3$ και $r_4 = r'_4$ [βλ. σχέσεις (2)]. Από το σύνολο των παραπάνω ισοτήτων προκύπτει ότι, όταν οι δύο κλώνοι του DNA εξελίσσονται συμμετρικά τότε ισχύει $r_{A \to T} = r_{T \to A}$ και $r_{A \to G} = r_{T \to C}$, για κάθε έναν από τους δύο κλώνους ξεχωριστά [βλ. σχέσεις (3)]. **(c)** Τα βέλη δηλώνουν τους ρυθμούς υποκατάστασης μίας δεδομένης βάσης, Χ, από μία άλλη, Υ, για έναν από τους δύο κλώνους του DNA. Όταν το DNA εξελίσσεται συμμετρικά, οι ρυθμοί των συμπληρωματικών υποκαταστάσεων είναι ίσοι σε κάθε έναν από τους κλώνους του ξεχωριστά. Συνεπώς, οι ρυθμοί υποκατάστασης που δηλώνονται από βέλη του ίδιου χρώματος ισούνται μεταξύ τους. Το σύνολο αυτών των ισοτήτων συγκροτούν τον πρώτο κανόνα της ισοδυναμίας (PR1).

Ο ρυθμός μεταβολής της συχνότητας των τεσσάρων βάσεων προσδιορίζεται από το ακόλουθο σύστημα γραμμικών εξισώσεων:

$$dA/dt = -(r_{A\to G} + r_{A\to C} + r_{A\to T})*f_A + r_{G\to A}*f_G + r_{C\to A}*f_C + r_{T\to A}*f_T$$

$$dG/dt = r_{A\to G}*f_A - (r_{G\to A} + r_{G\to C} + r_{G\to T})*f_G + r_{C\to G}*f_C + r_{T\to G}*f_T$$

$$dC/dt = r_{A\to C}*f_A + r_{G\to C}*f_G - (r_{C\to A} + r_{C\to G} + r_{C\to T})*f_C + r_{T\to C}*f_T$$

$$dT/dt = r_{A\to T}*f_A + r_{G\to T}*f_G + r_{C\to T}*f_C - (r_{T\to A} + r_{T\to G} + r_{T\to C})*f_T \qquad (4)$$

Δυνάμει του PR1 (σχέση 3), η λύση του συστήματος (4) δηλώνει πως η ενδοκλωνική σύσταση του DNA τείνει προς μία κατάσταση ισορροπίας, όπου:

$$[A] = [T] \text{ και } [G] = [C] \qquad (5)$$

Η σχέση (5) είναι γνωστή ως *ο δεύτερος κανόνας της ισοδυναμίας* (parity rule 2, PR2) (Lobry 1995, Sueoka 1995).

Συνοψίζοντας, *όταν ισχύει ο PR1, όταν δηλαδή οι δύο κλώνοι του DNA εξελίσσονται συμμετρικά, η ενδοκλωνική σύσταση του DNA συγκλίνει ασυμπτωτικά στην ισομοριακότητα των συμπληρωματικών βάσεων, όπως περιγράφει ο PR2*, ακόμα και στην περίπτωση που οι ρυθμοί υποκατάστασης μεταβάλλονται

με την πάροδο του χρόνου, κατά την εξελικτική πορεία των οργανισμών. Αντιστρέφοντας το επιχείρημα, ο PR2 αντιστοιχεί σε μια μηδενική υπόθεση σχετικά με την εξέλιξη της αλληλουχίας του δίκλωνου DNA σύμφωνα με την οποία οι δύο κλώνοι του DNA υπόκεινται σε συμμετρικές υποκαταστάσεις, ισχύει δηλαδή ο PR1 (Sueoka 1995, Charneski et al. 2011).

## **1.3** *Αποκλίσεις από τον 2° κανόνα της ισοδυναμίας*

Μελέτες στις οποίες χρησιμοποιήθηκαν σύνολα μεγάλων αλληλουχιών DNA καθώς και ολόκληρα χρωμοσώματα από οργανισμούς διαφορετικής εξελικτικής προέλευσης, επιβεβαίωσαν την ισχύ του PR2 σε αρκούντως μεγάλα τμήματα του γενετικού υλικού (Prabhu 1993, Bell & Forsdyke 1999, Mitchell & Bridge 2006). Ωστόσο, ενδοκλωνικές αποκλίσεις από την ισοδυναμία των συμπληρωματικών βάσεων ([A] ≠ [T] και [G] ≠ [C]) παρατηρούνται σε τοπική κλίμακα. Οι Smithies *et al.* (1981) υπήρξαν οι πρώτοι που κατέγραψαν τέτοιες αποκλίσεις. *Οι παραβιάσεις του PR2 ανοίγουν ένα παράθυρο στη μελέτη θεμελιωδών μοριακών μηχανισμών, όπως η αντιγραφή, η μεταγραφή και η επιδιόρθωση του γενετικού υλικού, που αλληλεπιδρούν άμεσα με το μόριο του DNA αλλά συγχρόνως δρουν με διαφορετικό τρόπο σε κάθε έναν από τους κλώνους του.* Ως εκ τούτου μπορούν και διαμορφώνουν τη σύσταση της αλληλουχίας DNA κατά τρόπο ώστε να αποκλίνει από τη μηδενική υπόθεση της ενδοκλωνικής ισομοριακότητας των συμπληρωματικών βάσεων, οπότε ισχύει [A] ≠ [T] και [G] ≠ [C] (Rocha & Danchin 2001, Lobry & Sueoka 2002, Klasson & Andersson 2006, Necşulea & Lobry 2007, Rocha 2008, Charneski et al. 2011).

Οι αποκλίσεις από τον PR2 είναι ενδεικτικές της ύπαρξης ειδικών ανά κλώνο πολώσεων των ρυθμών υποκατάστασης. Οι Wu και Maeda ήταν από τους πρώτους που πραγματοποίησαν σχετικά πειράματα για τον εντοπισμό ασυμμετριών στους ρυθμούς υποκατάστασης μεταξύ των δύο κλώνων του DNA (Wu & Maeda 1987). Συγκεκριμένα, μέσω φυλογενετικής ανασυγκρότησης μελέτησαν τις υποκαταστάσεις που έχουν λάβει χώρα σε μία περιοχή του συμπλόκου της β-γλοβίνης των πρωτευόντων, αλλά κατέληξαν σε αντικρουόμενα αποτελέσματα, όπως κατέδειξε ακολούθως ο Bulmer (1991a). Χρησιμοποιώντας την ίδια

μεθοδολογία με τους Wu και Maeda, οι Francino *et al*. (1996) κατέγραψαν διαφορές στους ρυθμούς των συμπληρωματικών C→T και G→A υποκαταστάσεων που λαμβάνουν χώρα στις κωδικές περιοχές των βακτηρίων.

Ωστόσο, δεν είναι πάντοτε δυνατή η εφαρμογή μεθόδων βασισμένων στη στοίχιση και σύγκριση ομόλογων αλληλουχιών που προέρχονται από διαφορετικούς οργανισμούς προκειμένου να προσδιοριστούν ασυμμετρίες στα πρότυπα υποκαταστάσεων. Και τούτο διότι δεν είναι εύκολο να προσδιοριστεί για κάθε έναν από τους υπό εξέταση οργανισμούς ο κλώνος του DNA με τον οποίο συνδέονταν αυτές οι ομόλογες περιοχές κατά τη διάρκεια της εξελικτικής τους πορείας (Marín & Xia 2008), καθώς γεγονότα αναστροφών μπορεί να έχουν αλλάξει την τοποθέτησή τους. Ο Lobry εισήγαγε ένα μέτρο που δεν προϋποθέτει τη στοίχιση αλληλουχιών DNA, καθώς οι σχετικοί υπολογισμοί γίνονται εντός του ίδιου χρωμοσώματος (Lobry 1996). Το μέτρο αυτό επεκτείνει τους στοιχειομετρικούς λόγους μεταξύ των νουκλεοτιδικών βάσεων τους οποίους είχαν προσδιορίσει βιοχημικά ο Chargaff και οι συνεργάτες του (Beaven, G. H., Holiday, E. R., & Johnson 1955). Συγκεκριμένα, οι αποκλίσεις από την ισομοριακότητα [A] = [T] ποσοτικοποιήθηκαν ως ο λόγος:

$$([A] - [T]) / ([A] + [T]) \qquad (6)$$

ενώ αντίστοιχα οι αποκλίσεις από την ισομοριακότητα [G] = [C] ποσοτικοποιήθηκαν ως ο λόγος:

$$([G] - [C]) / ([G] + [C]) \qquad (7)$$

Με δεδομένη την αλληλουχία του DNA, οι δύο αυτοί λόγοι μπορούν εύκολα να προσδιοριστούν σε αρκούντως μεγάλες περιοχές του γονιδιώματος, εντός των οποίων σποραδικές αναστροφές δεν αναμένεται να επηρεάζουν τις τιμές τους. Οι πρώτες σχετικές μετρήσεις πραγματοποιήθηκαν κατά μήκος αλληλουχιών DNA οι οποίες χωρίζονταν σε διαδοχικά, μικρότερα τμήματα (κυλιόμενα παράθυρα) εντός των οποίων γίνονταν οι αντίστοιχοι υπολογισμοί (Lobry 1996). Σε μία αναφορά σχετικά με το γονιδίωμα του *Escherichia coli*, οι Blattner *et al*. (1997) παρουσίασαν ένα γράφημα των GC αποκλίσεων κατά μήκος ολόκληρου του χρωμοσώματος του εν λόγω βακτηρίου. Ένα χρόνο αργότερα, ο Grigoriev απεικόνισε το άθροισμα των GC αποκλίσεων, όπως αυτές υπολογίστηκαν σε γειτονικά κυλιόμενα παράθυρα, κατά μήκος διαφόρων βακτηριακών χρωμοσωμάτων και πρότεινε τη χρήση των αθροιστικών αυτών γραφημάτων για τον εντοπισμό των σημείων έναρξης (*ori*) και λήξης (*ter*) της αντιγραφής, καθώς παρατήρησε πως τα ακρότατά τους συμπίπτουν με τα εν λόγω σημεία (*ori* και *ter*)

(Grigoriev 1998). Εκτός από τις GC αποκλίσεις, για τον εντοπισμό του *ori* και του *ter* έχουν εισαχθεί και άλλα μέτρα βασισμένα στη σύσταση του DNA, όπως η διαφορά μεταξύ πουρινών (Pu) και πυριμιδινών (Py), ποσοτικοποιημένη ως *απόκλιση πουρινών*, (Pu – Py) / (Pu + Py) (Mohr et al. 1999).

## **1.4** *Μηχανισμοί που επάγουν αποκλίσεις από τους κανόνες της ισοδυναμίας*

Συμμετρία των ρυθμών υποκατάστασης (PR1) σημαίνει πως η αλλαγή μιας βάσης, έστω $N_1$, προς μία άλλη, έστω $N_2$, στον έναν κλώνο πραγματοποιείται με την ίδια συχνότητα που η συμπληρωματική βάση της $N_1$ υποκαθίσταται από τη συμπληρωματική βάση της $N_2$ στον άλλο κλώνο (Lobry 1995). *Κάθε μηχανισμός που παραβιάζει τον PR1 συνεπάγεται αποκλίσεις από τον PR2*, δηλαδή [A] ≠ [T] και [G] ≠ [C]. Ακολούθως παρουσιάζονται οι βασικοί μηχανισμοί που έχουν συσχετιστεί με την εμφάνιση αποκλίσεων από τον PR1 και PR2. Διακρίνονται σε μηχανισμούς που σχετίζονται με επιλεκτικές ή μεταλλακτικές διαδικασίες.

### **1.4.1** *Επιλεκτικοί μηχανισμοί*

Ασυμμετρίες στους ρυθμούς υποκατάστασης εμφανίζονται στις κωδικές αλληλουχίες (coding sequences, CDS), διακρίνοντας μεταξύ κωδικού (coding) και αντικωδικού (template) κλώνου. Οι *σχετιζόμενες με κωδικές αλληλουχίες αποκλίσεις* από τον PR2 (CDS-αποκλίσεις) εκδηλώνονται σε τοπική κλίμακα. Ωστόσο μπορούν να επηρεάζουν τη συνολική νουκλεοτιδική σύσταση του DNA, ιδίως σε χρωμοσώματα που το μεγαλύτερο μέρος τους καλύπτεται από CDS. Τέτοιες είναι οι περιπτώσεις των προκαρυωτικών και ιϊκών γονιδιωμάτων, καθώς και του οργανιδιακού DNA (μιτοχονδριακού ή χλωροπλαστικού).

**1.4.1.1** Επιλογή χρήσης αμινοξέων - αποκλίσεις στις 1ες και 2ες θέσεις των κωδικονίων

Συνώνυμες υποκαταστάσεις λαμβάνουν χώρα μόνο στις 3ες θέσεις των κωδικονίων, με την εξαίρεση ορισμένων τριπλετών που κωδικοποιούν για Αργινίνη ή Λευκίνη όπου συνώνυμες υποκαταστάσεις μπορούν να συμβούν και στην 1η θέση. Αντίθετα, η νουκλεοτιδική σύσταση της 1ης και 2ης θέσης των κωδικονίων είναι άμεσα συνδεδεμένη με το αμινοξικό περιεχόμενο των πρωτεϊνών. Κατά συνέπεια η προτίμηση συγκεκριμένων αμινοξικών καταλοίπων, όπως της γλυκίνης, της αλανίνης και της βαλίνης, που έχει καταγραφεί στην περίπτωση του *Echerichia coli* (Lobry & Gautier 1994) και άλλων οργανισμών (Karlin et al. 1992), μπορεί να οδηγεί σε συστηματικές παραβιάσεις του PR2 στην 1η και 2η θέση (Nakamura et al. 1999). Υψηλές συχνότητες χρήσης αμινοξικών καταλοίπων γλυκίνης, αλανίνης και βαλίνης συγκλίνουν στον εμπλουτισμό της 1ης θέσης των κωδικονίων σε Gs έναντι Cs. Γενικότερα, η αποκλίνουσα επιλογή (diversifying selection) που ασκείται στο επίπεδο της αμινοξικής σύστασης των πρωτεϊνών οδηγεί σε εμπλουτισμό των *ενσωματωμένων μεμβρανικών* (integral) *πρωτεϊνών* σε υδρόφοβα αμινοξέα (Phe, Leu, Ile, Met, Val, Tyr, Trp) και των *κυτταροπλασματικών πρωτεϊνών* σε πολικά αμινοξέα (Asp, Glu, Lys, Arg, His, Asn, Gln). Δεδομένης της δομής του γενετικού κώδικα, οι επιλεκτικές αυτές πιέσεις που σχετίζονται με τον υποκυτταρικό εντοπισμό των πρωτεϊνικών μορίων συνεπάγονται την υπερεκπροσώπηση στη 2η θέση καταλοίπων θυμίνης (T) και κυτοσίνης (C) σε σχέση με τα κατάλοιπα αδενίνης (A) και γουανίνης (G), αντίστοιχα, στις ενσωματωμένες μεμβρανικές πρωτεΐνες (Lobry & Lobry 1999).

**1.4.1.2** Βελτιστοποίηση της μετάφρασης - αποκλίσεις στις 1ες και 2ες θέσεις των κωδικονίων

Επιλεκτικές πιέσεις που διαμορφώνουν τη σύσταση των κωδικών περιοχών ασκούνται επίσης στο επίπεδο της λειτουργικής αλληλεπίδρασης mRNA – rRNA στο ριβόσωμα κατά τη μετάφραση. Η περιοδική εμφάνιση μοτίβων GHN, η παρουσία δηλαδή καταλοίπων γουανίνης (G) στην 1η θέση των κωδικονίων και η απουσία τους (H) στη 2η, συμβάλει στην ορθή αναγνώριση του πλαισίου ανάγνωσης (Trifonov 1987, Lagunez-Otero & Trifonov 1992). Η τάση αυτή επάγει ασυμμετρίες στην κατανομή των Gs μεταξύ κωδικού και αντικωδικού κλώνου στις 1ες και 2ες κωδικές θέσεις.

**1.4.1.3** Επιλογή χρήσης συνώνυμων κωδικονίων – αποκλίσεις στις 3$^{ες}$ θέσεις των κωδικονίων

Εκτός από τη σύσταση των 1$^{ων}$ και 2$^{ων}$ κωδικών θέσεων, επιλεκτικοί μηχανισμοί καθορίζουν σε σημαντικό βαθμό και τη σύσταση των 3$^{ων}$ κωδικών θέσεων. Πρόκειται για διαδικασίες που επάγουν την επιλεκτική χρήση συνώνυμων κωδικονίων (synonymous codon preference) και *δεν σχετίζονται με περιορισμούς που αντανακλούν στο αμινοξικό περιεχόμενο των πρωτεϊνών*. Οι αποκλίσεις από τον PR2 που οφείλονται στην επιλογή συνώνυμων κωδικονίων ανιχνεύονται πρωτίστως στις 3$^{ες}$ τετραπλά εκφυλισμένες θέσεις των κωδικονίων (Sueoka 1995).

Στις πρώτες σχετικές μελέτες, ο Ikemura κατέδειξε ότι υπάρχει μια ισχυρή, θετική συσχέτιση της κυττοπλασματικής συγκέντρωσης των tRNAs και της συχνότητας των αντίστοιχων κωδικονίων στα mRNAs, η οποία είναι εντονότερη όταν εξετάζονται τα ισοδεκτικά tRNAs (isoaccepting tRNA) σε σχέση με τα αντίστοιχα συνώνυμα κωδικόνια (Ikemura 1981). Οι παρατηρούμενες συσχετίσεις συνδέθηκαν αρχικά με τη βελτιστοποίηση της μετάφρασης των mRNAs που κωδικοποιούν για πρωτεΐνες οι οποίες εμφανίζονται σε υψηλή κυττοπλασματική συγκέντρωση. Ακολούθως, οι Gouy και Gautier (1982) πρότειναν ως καθοριστικό παράγοντα στη συσχέτιση μεταξύ συχνοτήτων tRNAs και κωδικονίων τον ρυθμό έκφρασης των γονιδίων. Υψηλότεροι ρυθμοί έκφρασης συνοδεύονται από εντονότερες συσχετίσεις ανάμεσα στις συχνότητες χρήσης των συνώνυμων κωδικονίων και τη συγκέντρωση των αντίστοιχων tRNAs.

Οι μελέτες αυτές, καθώς και σειρά άλλων που ακολούθησαν, υποστηρίζουν πως η φυσική επιλογή οδηγεί στην χρήση ενός συνόλου *βέλτιστων κωδικονίων* (optimal codons). Το σύνολο αυτό μπορεί να διαφοροποιείται μεταξύ των οργανισμών, αλλά και μεταξύ γονιδίων του ίδιου οργανισμού, ανάλογα με τον κλώνο, οδηγό ή συνοδό, στο οποίο βρίσκονται, όπως παρατηρήθηκε στην περίπτωση του *Borrelia burgdorferi* (McInerney 1998). Η χρήση των *βέλτιστων κωδικονίων* μπορεί να αυξάνει την απόδοση (Kudla et al. 2009) ή την ακρίβεια (Stoletzki & Eyre-Walker 2007) της μετάφρασης. Για μια αναλυτικότερη παρουσίαση των μηχανισμών που έχουν προταθεί για την ερμηνεία του φαινομένου, μπορεί κανείς να ανατρέξει στην επισκόπηση των Sharp *et al.* (2010) και στις εκεί παραπομπές. Αξίζει να σημειωθεί πως, ανεξάρτητα από την αιτία που την προκαλεί, *η επιλογή χρήσης συνώνυμων κωδικονίων οδηγεί σε μια συνολική τροποποίηση της σύστασης των κωδικών αλληλουχιών. Αντικείμενο*

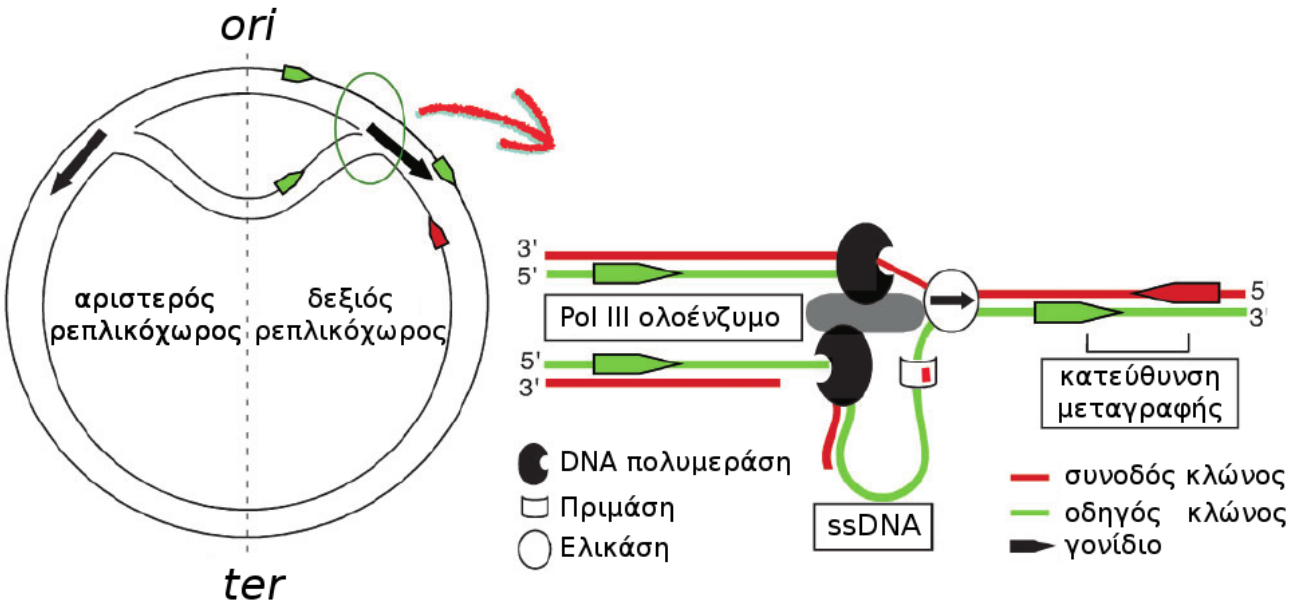
*επιλογής δεν είναι ένα συγκεκριμένο κωδικόνιο, αλλά οι σχετικές συχνότητες όλων των κωδικονίων. Στο πλαίσιο αυτό, η εξέλιξη της σύστασης του DNA εμφανίζεται ως μια διαδικασία σταδιακής συσσώρευσης μικρών μεταβολών.*

**1.4.1.4** Επιλογή του προσανατολισμού των γονιδίων

Οι μηχανισμοί που προαναφέρθηκαν δρουν κατ' αρχήν σε τοπική κλίμακα στις κωδικές περιοχές του DNA (CDS). Οι αποκλίσεις που παράγονται (CDS-αποκλίσεις) συμβάλλουν στη διάκριση μεταξύ του κωδικού (sense) και μεταγραφόμενου (antisense) κλώνου των γονιδίων. Παράλληλα, κατά μήκος του γονιδιώματος δρουν μηχανισμοί που επάγουν την ανακατανομή των κωδικών και μεταγραφόμενων κλώνων μεταξύ της οδηγού και της συνοδού αλυσίδας του DNA. Ως αποτέλεσμα, *σε πολλούς οργανισμούς παρατηρείται μια έντονα πολωμένη (συγγραμμική) διάταξη των κωδικών κλώνων κατά μήκος της οδηγού αλυσίδας, γεγονός που συνεπάγεται την ανάδειξη των CDS-αποκλίσεων από την τοπική κλίμακα στο επίπεδο εκτεταμένων περιοχών του DNA*. Σε οργανισμούς που το μεγαλύτερο τμήμα των χρωμοσωμάτων τους καλύπτεται από κωδικές περιοχές, όπως είναι τα βακτήρια, *οι CDS-αποκλίσεις σε συνδυασμό με την έντονη πόλωση των κωδικών κλώνων μπορεί να διαμορφώνουν πρότυπα αποκλίσεων από τον PR2 στο επίπεδο ολόκληρου του γονιδιώματος*.

Το πλεόνασμα των γονιδίων που κωδικοποιούνται στον οδηγό κλώνο είχε τεκμηριωθεί από σχετικές μελέτες ακόμα και πριν από τη *γονιδιωματική εποχή*, οι οποίες αρχικά αφορούσαν στα γονίδια του rRNA και των ριβοσωμικών πρωτεϊνών του *E.coli* (Nomura & Morgan 1977), ενώ στην συνέχεια επεκτάθηκαν και σε άλλες κλάσεις γονιδίων που κωδικοποιούν για πρωτεΐνες (Brewer 1988). Η διάταξη των κωδικών κλώνων στην οδηγό αλυσίδα του DNA προσφέρει εξελικτικό πλεονέκτημα στους οργανισμούς, καθώς διευκολύνει την αποφυγή της κατά μέτωπο σύγκρουσης (head-on collision) των DNA και RNA πολυμερασών. Το γεγονός αυτό έχει ιδιαίτερη σημασία στους προκαρυωτικούς οργανισμούς όπου οι διαδικασίες της αντιγραφής και της μεταγραφής εξελίσσονται παράλληλα στο χρόνο. Η δομή του ολοενζύμου της DNA πολυμεράσης III προκαλεί τοπική συστροφή της νεοσυντιθέμενης συνοδού αλυσίδας στη διχάλα της αντιγραφής. Ως συνέπεια, παρ' ότι αντιγραφή και μεταγραφή είναι συγγραμμικές διαδικασίες με φορά 5'→3', οι DNA και RNA πολυμεράσες κινούνται προς αντίθετες κατευθύνσεις στον χώρο όταν δρουν πάνω στη συνοδό αλυσίδα. Για το λόγο αυτό ασκείται επιλεκτική πίεση στον προσανατολισμό των γονιδίων ώστε ο κωδικός

τους κλώνος να τοποθετείται στην οδηγό αλυσίδα του DNA (Brewer 1988) (βλ. Εικόνα 2). Δεδομένων των CDS-αποκλίσεων από τον PR2, η πόλωση του προσανατολισμού των γονιδίων προς μία συγκεκριμένη κατεύθυνση κατά μήκος ενός μεγάλου τμήματος του DNA, όπως αυτό ορίζεται από την αντιγραφή, συμβάλλει στην εμφάνιση ευρείας κλίμακας A-T και G-C αποκλίσεων.




**Εικόνα 2.** ***τροποποιημένη από*** **(Rocha 2004).** Ο αναδιπλασιασμός των βακτηριακών χρωμοσωμάτων προχωρά αμφίδρομα από το σημείο έναρξης (*ori*) προς το σημείο λήξης (*ter*) της αντιγραφής, χωρίζοντας το εκάστοτε χρωμόσωμα σε δύο ημίσεα, τους *ρεπλικόχωρους*. Ο σχηματισμός και η περαιτέρω ανάπτυξη της διχάλας της αντιγραφής εξαρτώνται από την δράση της DNA ελικάσης, που ξετυλίγει το δίκλωνο DNA στην μονόκλωνη κατάστασή του (single-stranded DNA, ssDNA). Καθώς στα βακτήρια η αντιγραφή και η μεταγραφή πραγματοποιούνται παράλληλα στον χρόνο, η DNA ελικάση συγκρούεται κατά μέτωπο (head-on collision) με τα ένζυμα της RNA πολυμεράσης που μεταγράφουν γονίδια τα οποία βρίσκονται στο συνοδό κλώνο (κόκκινο βέλος). Προς αποφυγή τέτοιων συγκρούσεων, στα βακτηριακά γονιδιώματα ασκείται επιλεκτική πίεση στον προσανατολισμό των γονιδίων, ώστε αυτά να βρίσκονται κατά προτίμηση στον οδηγό κλώνο.

Η πόλωση της διάταξης των γονιδίων κατά μήκος του DNA δεν περιορίζεται μόνο στους προκαρυωτικούς οργανισμούς, αλλά είναι ένα γενικευμένο φαινόμενο που εμφανίζεται επίσης στο γενετικό υλικό των ιών και των μιτοχονδρίων, ωστόσο η έντασή του ποικίλει (McLean et al. 1998, Frank &

Lobry 1999, Nikolaou & Almirantis 2006). Η επιλεκτική πίεση ώστε οι κωδικοί κλώνοι να κατανέμονται κατά προτίμηση στην οδηγό αλυσίδα είναι εντονότερη σε γονίδια που είτε μεταγράφονται με υψηλούς ρυθμούς (Francino & Ochman 1997) είτε κωδικοποιούν για πρωτεΐνες απαραίτητες για τις βασικές λειτουργίες του οργανισμού (Rocha & Danchin 2003).

**1.4.1.5** Επιλογή της κατανομής αλληλουχιών σηματοδότησης

Στην εμφάνιση αποκλίσεων από τον PR2 συμβάλλει επίσης η άνιση κατανομή αλληλουχιών σηματοδότησης μεταξύ των κλώνων του DNA. Οι αλληλουχίες αυτές μπορεί να συμμετέχουν στη ρύθμιση της αντιγραφής, της γονιδιακής έκφρασης (Karlin et al. 1996, Rocha et al. 1998), του ανασυνδυασμού (Kuzminov 1995), της έναρξης σύνθεσης των τμημάτων Okazaki (Yoda & Okazaki 1991) ή της μετα-αντιγραφικής ανασυγκρότησης των νουκλεοσωμάτων στα άκρα των χρωμοσωμάτων (Cornet et al. 1996). Εξαιτίας των λειτουργιών που επιτελούν, τα ολιγονουκλεοτίδια που αποτελούν αλληλουχίες σηματοδότησης καθώς και η κατανομή τους στους κλώνους του DNA θεωρούνται αντικείμενο επιλογής. Συνεπώς, στο γονιδίωμα ασκούνται επιλεκτικές πιέσεις που οδηγούν εντός του κάθε DNA κλώνου σε έντονες αποκλίσεις των συχνοτήτων συγκεκριμένων ολιγονουκλεοτιδίων σε σχέση με τα αντιστρόφως συμπληρωματικά τους.

**1.4.2** *Μεταλλακτικοί μηχανισμοί*

Στη διαμόρφωση των αποκλίσεων από τον PR2 ιδιαίτερα σημαντική επίδραση ασκούν εκτός από τις επιλεκτικές πιέσεις και *θεμελιώδεις ασυμμετρίες των μοριακών μηχανισμών που αλληλεπιδρούν με το γενετικό υλικό, όπως η αντιγραφή και η μεταγραφή*. Οι ασυμμετρίες αυτές επάγουν ειδικές ανά κλώνο πολώσεις των μεταλλακτικών ρυθμών.

**1.4.2.1** *Διαφορική έκθεση των DNA αλυσίδων στη μονόκλωνη κατάσταση*

Κατά τη διαδικασία της αντιγραφής και της μεταγραφής οι αντιστρόφως συμπληρωματικοί κλώνοι του DNA (οδηγός και συνοδός ή κωδικός και μεταγραφόμενος, αντίστοιχα) διαφέρουν ως προς το χρονικό διάστημα κατά το

οποίο εκτίθενται στη μονόκλωνη κατάσταση. Στο μονόκλωνο DNA οι νουκλεοτιδικές βάσεις υφίστανται χημικές τροποποιήσεις με υψηλότερους ρυθμούς από ότι στο δίκλωνο. Η υδρολυτική απαμίνωση είναι η πλέον χαρακτηριστική από αυτές τις τροποποιήσεις (Frederico et al. 1990). Τα κατάλοιπα κυτοσίνης είναι ασταθή στο μονόκλωνο DNA και κατόπιν απαμίνωσης μετατρέπονται σε κατάλοιπα ουρακίλης (C→U) (Lindahl, 1993, Kreutzer και Essigmann, 1998). Η αντίδραση αυτή πραγματοποιείται περισσότερο από 100 φορές συχνότερα στο μονόκλωνο από ότι στο δίκλωνο DNA. Μετά τη δράση επιδιορθωτικών μηχανισμών, τα κατάλοιπα ουρακίλης αντικαθίστανται από θυμίνη. Μεθυλίωση της κυτοσίνης αυξάνει το ρυθμό απαμίνωσης έως και 4 φορές. Στην περίπτωση αυτή, κατάλοιπα 5-μεθυλ κυτοσίνης μετατρέπονται κατευθείαν σε κατάλοιπα θυμίνης (5mC→T) (Charneski et al. 2011). Συνεπώς, η απαμίνωση καταλοίπων C και 5mC οδηγεί σε μεταβάσεις του τύπου C→T (Echols & Goodman 1991). Στο μονόκλωνο DNA παρατηρείται επίσης, αν και σπανιότερα, απαμίνωση καταλοίπων αδενίνης προς υποξανθίνη, η οποία σχηματίζει ζεύγη με κατάλοιπα κυτοσίνης και οδηγεί σε μεταβάσεις του τύπου A→G (Lindahl 1993). Το τελικό αποτέλεσμα των αντιδράσεων απαμίνωσης είναι ο εμπλουτισμός σε κατάλοιπα T και G εκείνης της DNA αλυσίδας που εκτίθεται περισσότερο χρόνο στη μονόκλωνη κατάσταση.

Η ειδική ανά κλώνο ασυμμετρία στους ρυθμούς απαμινώσεων αναμένεται να επιφέρει αύξηση των καταλοίπων T και G ανάλογη της μείωσης των καταλοίπων C και A που αυτή προκαλεί, και συνεπώς να μεταβάλει την ένταση τόσο των A-T όσο και των G-C αποκλίσεων. Ωστόσο, σχεδόν στο σύνολο των βακτηρίων οι αποκλίσεις G-C είναι εντονότερες από τις αποκλίσεις A-T (McLean et al. 1998). Παρ' ότι λοιπόν οι ασυμμετρίες στους ρυθμούς απαμίνωσης θεωρούνται καθοριστικής σημασίας, πλήθος άλλων παραγόντων συμβάλλει στη διαμόρφωση των PR2 αποκλίσεων.

### **1.4.2.2** Ασύμμετρη δράση της αντιγραφής

Η διχάλα της αντιγραφής εμφανίζει έντονες δομικές και λειτουργικές ασυμμετρίες (Echols & Goodman 1991, Trinh & Sinden 1991, Kunkel 1992, Schaaper 1993, Veaute & Fuchs 1993, Iwaki et al. 1996, Yuzhakov et al. 1996, Mrázek & Karlin 1998, Radman 1998). Διαφορετικοί ρυθμοί υποκατάστασης μεταξύ οδηγού και συνοδού αλυσίδας θεωρούνται ως το πιθανό αποτέλεσμα αυτών των ασυμμετριών και έχουν συσχετιστεί με ειδικές ανά κλώνο πολώσεις της

σύστασης του DNA (Mrázek & Karlin 1998). Οδηγός και συνοδός αλυσίδα αποκλίνουν σημαντικά από το PR2. Οι αποκλίσεις αυτές είναι ιδιαίτερα έντονες στους προκαρυωτικούς οργανισμούς, με την οδηγό αλυσίδα να παρουσιάζει κατά κανόνα υψηλότερες συχνότητες G και T έναντι C και A, αντίστοιχα (Rocha et al. 1999). Ο Lobry πρότεινε ότι οι παρατηρούμενες αποκλίσεις από τον PR2 οφείλονται στις διαφορές μεταξύ οδηγού και συνοδού αλυσίδας στους ρυθμούς των σχετιζόμενων με την αντιγραφή μεταλλάξεων και των επιδιορθώσεων (Lobry 1996). Ευρήματα που παρουσιάστηκαν σε κατοπινές μελέτες συνηγορούν υπέρ αυτού του ερμηνευτικού σχήματος (Rocha et al. 1999, Lobry & Sueoka 2002).

Κατά την αντιγραφή, η οδηγός αλυσίδα εκτίθεται στη μονόκλωνη κατάσταση για μεγαλύτερο χρονικό διάστημα από ότι η συνοδός (Okazaki et al. 1968). Ως συνέπεια, σύμφωνα με τα όσα ειπώθηκαν προηγουμένως (ενότητα 1.4.2.1), στην οδηγό αλυσίδα πραγματοποιούνται συχνότερα C→T μεταβάσεις. Η απαμίνωση της κυτοσίνης έχει προταθεί ως ένας από τους πιο σημαντικούς μηχανισμούς που διαμορφώνουν τις PR2 αποκλίσεις (Frank & Lobry 1999). Ο μηχανισμός αυτός εξηγεί ταυτόχρονα το θετικό πρόσημο των G-C αποκλίσεων και το αρνητικό πρόσημο των A-T αποκλίσεων κατά μήκος του οδηγού κλώνου. Επίσης, συνάδει με τις εντονότερες G-C αποκλίσεις που παρατηρούνται σε χρωμοσώματα με χαμηλό GC περιεχόμενο.

Αποκλίσεις από τον PR2 συσχετίζονται επίσης με την πιστότητα της αντιγραφής, που αφορά στην εισαγωγή των μονονουκλεοτιδίων κατά την επιμήκυνση των νεοσυντιθέμενων κλώνων (insertion step), στην 3'→5' εξωνουκλεολυτική δράση της πολυμεράσης (exonucleolytic proofreading) και στην επιδιόρθωση αταίριαστων ζευγών βάσεων (mismatch repair) (Schaaper 1993). Διαφοροποιήσεις οδηγού και συνοδού αλυσίδας σε αυτά τα τρία στάδια συνεπάγονται ειδικές ανά κλώνο πολώσεις της νουκλεοτιδικής σύστασης του γονιδιώματος.

Στους ευκαρυωτικούς οργανισμούς παρατηρούνται διαφορετικοί ρυθμοί ενσωμάτωσης μονονουκλεοτιδίων κατά τη σύνθεση οδηγού και συνοδού αλυσίδας (Radman 1998), καθώς οι αντιστρόφως συμπληρωματικοί DNA κλώνοι αντιγράφονται από διαφορετικά ενζυμικά σύμπλοκα (DNA πολυμεράση ε, DNA πολυμεράση δ) (Kunkel 1992). Αντίθετα, στα προκαρυωτικά γονιδιώματα και οι δύο κλώνοι αντιγράφονται από το ολοένζυμο της DNA πολυμεράσης III (Baker & Wickner 1992). Η DNA πολυμεράση III (PolIII) διαθέτει δύο ενεργά κέντρα

πολυμερισμού (polymerase cores). Πειράματα που διεξήχθησαν με το ολοένζυμο του *E.coli* κατέδειξαν πως η PolIII εμφανίζει εγγενείς ασυμμετρίες (Maki et al. 1988) που ανταποκρίνονται στις διαφορετικές απαιτήσεις της σύνθεσης οδηγού και συνοδού αλυσίδας (Marians 1992). Συνεπώς, τόσο στους ευκαρυωτικούς όσο και στους προκαρυωτικούς οργανισμούς, η αντιγραφή εμφανίζει κατά το στάδιο της επιμήκυνσης λειτουργικές ασυμμετρίες που μπορούν να οδηγήσουν σε αποκλίσεις από τον PR2.

Στους προκαρυωτικούς, η ασύμμετρη δράση της PolIII αποδίδεται επίσης στις διαφορές των δύο κέντρων πολυμερισμού ως προς την τάση τους να παραμένουν σε στενή επαφή με τους κλώνους που αντιγράφουν (Fijalkowska et al. 1998). Η αποδέσμευση του συμπλόκου της PolIII από το DNA επιτρέπει την επιδιόρθωση σφαλμάτων που εμφανίζονται στα άκρα του νεοσυντιθέμενου κλώνου. Στη διαδικασία αυτή το καθοριστικό βήμα δεν είναι η εισαγωγή ενός λάθος νουκλεοτιδίου, αλλά η δυνατότητα της PolIII να συνεχίσει την επιμήκυνση της νεοσυντιθέμενης αλυσίδας μετά από το σχηματισμό ενός αταίριαστου ζεύγους βάσεων. Η δυνατότητα αυτή περιορίζεται τόσο περισσότερο, όσο μεγαλύτερη είναι η τάση αποσύνδεσης του ολοενζύμου από το υπόστρωμά του (Echols & Goodman 1991). Καθώς η συνοδός αλυσίδα συντίθεται ασυνεχώς, το ενεργό κέντρο της PolIII αποσπάται συχνότερα από τον αντίστοιχο DNA κλώνο. Επιπλέον, ο ασυνεχής τρόπος σύνθεσης της συνοδού αλυσίδας παρέχει πολλά σημεία στα οποία μπορούν να δράσουν μηχανισμοί επιδιόρθωσης μέσω αναγνώρισης των εγκοπών του DNA (nick translation) (Radman 1998), ενώ η δράση των ενζύμων με ενεργότητα λιγάσης, που είναι απαραίτητα για τη σύνδεση των κομματιών Okazaki, παρεμποδίζεται παρουσία αταίριαστων ζευγών βάσεων (Housby & Southern 1998). Ως αποτέλεσμα *οι ρυθμοί επιδιόρθωσης μεταλλάξεων σχετιζόμενων με την αντιγραφή είναι υψηλότεροι στη συνοδό παρά στην οδηγό αλυσίδα*.

Πειραματικές μελέτες σχετικά με τις διαφορές στην πιστότητα της αντιγραφής οδηγού και συνοδού αλυσίδας παρέχουν αντικρουόμενα αποτελέσματα. Ορισμένα από αυτά υποδεικνύουν ότι η συνοδός αλυσίδα συσσωρεύει περισσότερα σφάλματα κατά την αντιγραφή από ότι η οδηγός (Trinh & Sinden 1991, Veaute & Fuchs 1993, Rosche et al. 1995, Iwaki et al. 1996). Αντιθέτως, σύμφωνα με τα πειράματα των Fijalkowska *et al.* (1998) οι ρυθμοί των υποκαταστάσεων είναι υψηλότεροι στην οδηγό αλυσίδα, όπως αναμένεται βάσει και των ασυμμετριών στους ρυθμούς επιδιόρθωσης των DNA κλώνων. Επιπλέον, τα πειράματα αυτά εξηγούν περισσότερο ικανοποιητικά την παρατηρούμενη

νουκλεοτιδική σύσταση των προκαρυωτικών γονιδιωμάτων.

Συγκεκριμένα, στους προκαρυωτικούς οι C→T και A→G μεταβάσεις παρατηρούνται συχνότερα από τις συμπληρωματικές τους, G→A και T→C (Mendelman et al. 1990), ενώ σε ότι αφορά στις μεταστροφές, συχνότερες είναι αυτές του τύπου Py→Pu (Fersht & Knill-Jones 1981). Συνεπώς, αναμένεται πως ο κλώνος εκείνος που εκτίθεται σε εντονότερες μεταλλακτικές πιέσεις θα είναι εμπλουτισμένος σε κατάλοιπα Gs και Ts, ενώ οι πουρίνες θα υπερτερούν των πυριμιδινών. Τα χαρακτηριστικά αυτά ανταποκρίνονται περισσότερο στη νουκλεοτιδική σύσταση του οδηγού κλώνου (Perrière et al. 1996, Freeman 1998).

### **1.4.2.3** Ασύμμετρη δράση της μεταγραφής

Παρ' ότι οι μεταλλάξεις που επάγει η αντιγραφή πραγματοποιούνται με διαφορετικούς ρυθμούς στους αντιστρόφως συμπληρωματικούς DNA κλώνους, η συμβολή αυτών των διαφορών στην εξέλιξη της συνολικής σύστασης του γονιδιώματος έχει αμφισβητηθεί. Οι Francino, Ochman και οι συνεργάτες τους πρότειναν πως δεν είναι η αντιγραφή, αλλά η μεταγραφή και η συζευγμένη με αυτήν επιδιόρθωση του DNA που παράγουν τις ειδικές ανά κλώνο πολώσεις των ρυθμών υποκατάστασης. Οι πρώτες σχετικές μελέτες αφορούσαν σε γονίδια των Εντεροβακτηρίων και υποστήριζαν ότι σημαντικές διαφορές στους ρυθμούς υποκατάστασης παρατηρούνταν μεταξύ κωδικών και μεταγραφόμενων κλώνων και όχι μεταξύ οδηγού και συνοδού αλυσίδας (Francino et al. 1996). Συγκεκριμένα, καταγράφηκαν ιδιαίτερα υψηλές συχνότητες C→T μεταβάσεων στον κωδικό κλώνο σε σχέση με τον μεταγραφόμενο.

Κατά τη μεταγραφή, ο κωδικός κλώνος βρίσκεται εκτεθειμένος στην μονόκλωνη κατάσταση για μεγαλύτερο χρονικό διάστημα απ' ότι ο μεταγραφόμενος. Συνεπώς, περισσότερες C→T μεταβάσεις συμβαίνουν στον κωδικό κλώνο, λόγω απαμίνωσης της κυτοσίνης (Francino & Ochman 1997). Επιπρόσθετα, οι χαμηλότεροι ρυθμοί υποκαταστάσεων στον μεταγραφόμενο κλώνο συσχετίζονται με τη συζευγμένη με τη μεταγραφή επιδιόρθωση του DNA (transcription-coupled repair, TCR) (Francino & Ochman 2001). Οι μεταλλάξεις που πραγματοποιούνται στον μεταγραφόμενο κλώνο αποτελούν το υπόστρωμα της TCR, η οποία κατευθύνει τη δράση της ειδικά σε αυτόν τον κλώνο (Hanawalt 1991). Συνεπώς, *ο μεταγραφόμενος κλώνος υπόκειται σε χαμηλότερους ρυθμούς μεταλλάξεων και υψηλότερους ρυθμούς επιδιορθώσεων, γεγονός που έχει ως αποτέλεσμα τις*

*παρατηρούμενες ασυμμετρίες στους ρυθμούς υποκατάστασης*. Ακόλουθα πειράματα επιβεβαίωσαν την ιδιαίτερη συμβολή της μεταγραφής στη εμφάνιση μεταλλακτικών πολώσεων στο επίπεδο ολόκληρου του γονιδιώματος (Lind & Andersson 2008).

### **1.4.3** *Συνδυασμένη επίδραση επιλεκτικών και μεταλλακτικών πιέσεων*

Πρόκειται για μηχανισμούς που επάγουν διαφοροποιήσεις της νουκλεοτιδικής σύστασης στη βάση της διάκρισης μεταξύ κωδικών και μεταγραφόμενων κλώνων των γονιδίων (Francino et al. 1996, Francino & Ochman 1997). Συγκεκριμένα, οι συστηματικές PR2 αποκλίσεις μεταξύ οδηγού και συνοδού κλώνου αποδίδονται στη συνδυασμένη δράση (*α*) των ασύμμετρων μεταλλακτικών πιέσεων που επάγει η μεταγραφή και (*β*) της έντονα πολωμένης διάταξης των γονιδίων κατά μήκος της οδηγού αλυσίδας, που προκύπτει ως αποτέλεσμα επιλεκτικών πιέσεων.

Μεταξύ κωδικού και μεταγραφόμενου κλώνου εμφανίζονται διαφορές στις συχνότητες των συμπληρωματικών μεταβάσεων C→T και G→A. Οι συνακόλουθες αποκλίσεις από τον PR2 αποδίδονται στους συνδεδεμένους με τη μεταγραφή μηχανισμούς επιδιόρθωσης, οι οποίοι δρουν επιλεκτικά πάνω στον μεταγραφόμενο κλώνο. Οι μηχανισμοί αυτοί έχουν ως υπόστρωμα πρωτίστως διμερή πυριμιδινών (Hanawalt 1991). Τα διμερή αυτά υποβοηθούν την απαμίνωση της C καθώς και την εισαγωγή A απέναντι από C, καταλήγοντας και στις δύο περιπτώσεις στην αντικατάσταση C→T (Hutchinson 1996). Πέραν των μεταλλακτικών αυτών πιέσεων, η υψηλή περιεκτικότητα σε πυριμιδίνες που ανιχνεύεται στους μεταγραφόμενους κλώνους σταθεροποιείται λόγω και της τάσης εμπλουτισμού των αντιστρόφως συμπληρωματικών τους κωδικών κλώνων σε πουρίνες. Η τάση αυτή έχει ως αποτέλεσμα την παραγωγή μεταγράφων πλούσιων σε πουρίνες, τα οποία είναι λιγότερο ευεπίφορα σε επιβλαβείς μεταλλάξεις κατά τη μετάφραση (Szybalski et al. 1966). Οι μεγάλες συγκεντρώσεις πουρινών στα πρώιμα mRNAs είναι ένα φαινόμενο αναγνωρίσιμο σε ευρεία κλίμακα, γνωστό και ως *κανόνας του Szybalski*.

Αντίρροπες δυνάμεις ασκούνται στους κλώνους των γονιδίων λόγω της εγγενούς ασυμμετρίας της φυσαλίδας της μεταγραφής (transcription bubble) και οδηγούν σε μια τάση υπερεκπροσώπησης καταλοίπων T στον κωδικό κλώνο.

Συγκεκριμένα, ο μεταγραφόμενος κλώνος προστατεύεται από μεταλλακτικούς παράγοντες ευρισκόμενους στο κυτταρόπλασμα, καθώς πάνω του προσδένεται τμήμα του νεοσυντιθέμενου mRNA όπως επίσης και τα σύμπλοκα της μεταγραφής και η RNA πολυμεράση. Αντιθέτως, στον κωδικό κλώνο ασκούνται εντονότερες μεταλλακτικές πιέσεις, αφού αυτός παραμένει περισσότερο εκτεθειμένος στο κυτταρόπλασμα (Beletskii & Bhagwat 1996, Beletskii & Bhagwat 1998). Όπως ακριβώς και στην περίπτωση της αντιγραφής, στο μονόκλωνο DNA λαμβάνουν χώρα πολύ συχνότερα απαμινώσεις καταλοίπων κυτοσίνης από ότι στο δίκλωνο, με αποτέλεσμα να γίνονται συχνότερα μεταβάσεις του τύπου C→T στον κωδικό κλώνο (βλ. ενότητα 1.4.2.1 ).

Οι παραπάνω παρατηρήσεις σε συνδυασμό με την τάση των γονιδίων να προσανατολίζονται προς την ίδια κατεύθυνση με αυτή της αντιγραφής, παρέχουν μια πιθανή εξήγηση για τα πρότυπα των αποκλίσεων A-T και G-C, τα οποία αλλάζουν πρόσημο γύρω από τα σημεία έναρξης της αντιγραφής, όπου αλλάζει και η φορά της αντιγραφής και κατά κανόνα ο προσανατολισμός των περισσότερων μεταγραφόμενων τμημάτων του γονιδιώματος. Όπως έχουμε ήδη πει (βλ. ενότητα 1.4.1.4), τα αποτελέσματα των μηχανισμών αυτών ενισχύονται όσο υψηλότερος είναι ο ρυθμός μεταγραφής των αντίστοιχων γονιδίων (Mellon & Hanawalt 1989, Francino & Ochman 1997) και όσο πιο απαραίτητα είναι αυτά τα γονίδια για την επιβίωση του οργανισμού (Rocha & Danchin 2003).

## **1.5** *Αντιγραφή των προκαρυωτικών γονιδιωμάτων — η α-υπομονάδα της* PolIII

Η διαδικασία της αντιγραφής του DNA απαιτεί τη συνεργατική δράση ενός μεγάλου αριθμού πρωτεϊνών. Στους προκαρυώτες, η επιμήκυνση των νεοσυντιθέμενων κλώνων του DNA μεσολαβείται από το ολοένζυμο της DNA πολυμεράσης PolIII (Richardson et al. 1964). Το ολοένζυμο της PolIII αποτελείται από 10 υπομονάδες, οι οποίες οργανώνονται σε 3 λειτουργικές περιοχές (Kelman & O'Donnell 1995). Η α-καταλυτική υπομονάδα συγκροτεί το ενεργό κέντρο της PolIII.

### **1.5.1** *Ισομορφές της α-υπομονάδας*

Πρωτοπόρες μελέτες με μεταλλαγμένα στελέχη του εντεροβακτηρίου *Escherichia coli* κατέδειξαν πως το ενεργό κέντρο της PolIII κωδικοποιείται από το γονίδιο *dna*E (Gefter et al. 1971). Ακολούθησαν πειράματα για τον χαρακτηρισμό του ενεργού κέντρου της PolIII του Gram-θετικού *Bacillus subtilis* (Cozzarelli & Low 1973, Low et al. 1976). Η PolIII των Gram-θετικών βακτηρίων θεωρήθηκε αρχικά ως το προϊόν του *pol*C γονιδίου. Σύμφωνα με κατοπινές μελέτες που πραγματοποιήθηκαν στον *Bacillus subtilis* (Dervyn et al. 2001) και στον *Staphylococcus aureus* (Bruck & O'Donnell 2000, Inoue et al. 2001), τα γονίδια *dna*E και *pol*C είναι και τα δύο απαραίτητα για την αντιγραφή του γονιδιώματος των περισσοτέρων Gram-θετικών βακτηρίων. Στους οργανισμούς που φέρουν και τα δύο αυτά γονίδια, το προϊόν του *pol*C καταλύει την αντιγραφή του οδηγού κλώνου, ενώ το προϊόν του *dna*E αυτήν του συνοδού κλώνου.

Στο γονιδίωμα ορισμένων βακτηρίων εντοπίζονται περισσότερα του ενός γονίδια που είναι ομόλογα του *dna*E, όπως στην περίπτωση του *Mycobacterium tuberculosis*. Εξ αυτών, το *dna*E2 κωδικοποιεί για μία πολυμεράση που μεσολαβεί την επαγόμενη από γενετικές βλάβες μεταλλαξιγένεση (damage-induced mutagenesis), καταλύοντας την αντιγραφή του DNA διαμέσου βλάβης (translesion synthesis, TLS). Η έκφραση του *dna*E2 ρυθμίζεται από το SOS μονοπάτι, στα πλαίσια της απόκρισης των βακτηρίων σε εκτεταμένες βλάβες του γενετικού τους υλικού λόγω περιβαλλοντικών και άλλων καταπονήσεων (Tippin et al. 2004). Η *dnaE2 α*-υπομονάδα είναι ευεπίφορη σε σφάλματα ακόμα και όταν χρησιμοποιεί ως εκμαγείο άθικτους κλώνους DNA (Galhardo et al. 2005).

Βάσει συγκριτικής ανάλυσης της αλληλουχίας των γονιδίων που κωδικοποιούν για την *α*-καταλυτική υπομονάδα της PolIII, διακρίνονται συνολικά τέσσερις ισομορφές: *dna*E1, *dna*E2, *dna*E3 και *pol*C (Zhao et al. 2006). Οι ισομορφές αυτές σχηματίζουν διμερή και συγκροτούν το ενεργό κέντρο της PolIII. Στα βακτήρια που δεν φέρουν το γονίδιο *pol*C, η *α*-υπομονάδα συγκροτείται ως ομοδιμερές DnaE1-DnaE1. Η PolC ισομορφή δεν σχηματίζει ομοδιμερή, αλλά εντοπίζεται σε ετεροδιμερή με την DnaE1 ή την DnaE3, είτε ως PolC-DnaE1 είτε ως PolC-DnaE3. Τα διμερή DnaE-DnaE, PolC-DnaE1 και PolC-DnaE3 καταλύουν τον αναδιπλασιασμό του βακτηριακού

γονιδιώματος. Τέλος, η DnaE2 ισομορφή συγκροτεί με την DnaE1 το ετεροδιμερές DnaE1-DnaE2, το οποίο συμμετέχει στην αντιγραφή του DNA διαμέσου βλάβης (TLS).

Τα βακτήρια μπορούν να ταξινομηθούν σε τρεις ομάδες, σύμφωνα με τον τύπο του διμερούς της *α*-υπομονάδας που φέρουν: σε εκείνα που έχουν (*i*) ομοδιμερή *α*-υπομονάδα DnaE1-DnaE1, 'dnaE', (*ii*) ετεροδιμερή PolC-DnaE1 ή PolC-DnaE3, 'polC', ή (*iii*) ετεροδιμερή DnaE-dnaE2, 'dnaE2' (Zhao et al. 2006). Αυτό το σχήμα ταξινόμησης ('dnaE'/'polC'/'dnaE2') έχει συσχετιστεί με γενικά χαρακτηριστικά της σύστασης των βακτηριακών χρωμοσωμάτων. Συγκεκριμένα, το GC περιεχόμενο των βακτηρίων χωρίζεται σε τρεις διακριτές μεταξύ τους κατανομές, ανάλογα με τον τύπο του διμερούς της *α*-υπομονάδας που καταλύει τον αναδιπλασιασμό του γενετικού τους υλικού. Το περιεχόμενο σε GC των 'dnaE' βακτηρίων ακολουθεί ένα ευρύ φάσμα τιμών, τα 'dnaE2' βακτήρια έχουν υψηλό GC%, ενώ τα 'polC' βακτήρια έχουν χαμηλό GC%. Στη βάση αυτών των ευρημάτων, έχει προταθεί πως οι διαφορετικές ισομορφές της α-υπομονάδας διαμορφώνουν σε σημαντικό βαθμό τη σύσταση των βακτηριακών γονιδιωμάτων (Zhao et al. 2007, Wu et al. 2012).

### **1.5.2** *Ασύμμετρη δράση της α-υπομονάδας σε οδηγό και συνοδό κλώνο*

Η ίδια η αντιγραφή του DNA είναι μια εγγενώς ασύμμετρη διαδικασία, που πραγματοποιείται με διαφορετικό τρόπο σε κάθε έναν από τους αντιπαράλληλους κλώνους, οδηγό και συνοδό. Σύμφωνα με μελέτες στο βακτήριο *E.coli*, που αφορούσαν στο ομοδιμερές της DnaE1-DnaE1 α-υπομονάδας, τα δύο κέντρα πολυμερισμού της PolIII είναι συμμετρικά ως προς τη λειτουργία τους και η ασύμμετρη δράση του ολοενζύμου επάγεται από την DnaB ελικάση (Yuzhakov et al. 1996). Η αλληλεπίδραση μεταξύ DnaB ελικάσης και τ-υπομονάδας της PolIII επιτρέπει την ταυτόχρονη σύνθεση οδηγού και συνοδού αλυσίδας (συνεχής και ασυνεχής αντιγραφής, αντίστοιχα).

*Καθώς η αντιγραφή του γονιδιώματος αποτελεί έναν από τους σημαντικότερους ενδογενείς παράγοντες που επάγουν συστηματικές πολώσεις στους μεταλλακτικούς ρυθμούς, τα μόρια που την καταλύουν πιθανολογείται ότι σχετίζονται με ασυμμετρίες της σύστασης που εκδηλώνονται στην κλίμακα*

*ολόκληρου του γονιδιώματος*. Μετά την ανακάλυψη των DnaE2 και PolC ισομορφών της *α*-υπομονάδας, που διαφέρουν από την DnaE1 ως προς την καταλυτική τους ενεργότητα, εξετάστηκε η πιθανή σύνδεση αυτών των διαφορών με συστηματικές ασυμμετρίες που εμφανίζονται μεταξύ οδηγού και συνοδού κλώνου. Η PolC απαντάται σε πολλά μέλη του φύλου Firmicutes, τα οποία εμφανίζουν μη-τυπικές αποκλίσεις A-T. Συγκεκριμένα, ενώ τα περισσότερα βακτήρια έχουν στον οδηγό κλώνο μεγαλύτερο ποσοστό καταλοίπων θυμίνης έναντι της αδενίνης, στα Firmicutes η σχέση αυτή αντιστρέφεται. Αυτή η αντιστροφή της φοράς των A-T αποκλίσεων έχει συσχετιστεί με την αντιγραφή του οδηγού κλώνου από την PolC *α*-υπομονάδα (Worning et al. 2006, Necşulea & Lobry 2007).

Ωστόσο, η συσχέτιση των νουκλεοτιδικών αποκλίσεων με τς ισομορφές της *α*-υπομονάδας έχει αμφισβητηθεί (Rocha 2002). Στη σχετική μελέτη, τα βακτήρια διαχωρίστηκαν σε δύο μόνο ομάδες, ανάλογα με την παρουσία ή μη της PolC, δίχως να λαμβάνεται υπόψιν η DnaE2 ισομορφή της *α*-υπομονάδας. Στην ίδια μελέτη η παρουσία της PolC συνδέθηκε με την έντονα πολωμένη διάταξη των γονιδίων κατά τη φορά αντιγραφής του οδηγού κλώνου. Σε μια πρόσφατη έρευνα των ασυμμετριών μεταξύ οδηγού και συνοδού κλώνου, τα υπό εξέταση βακτηριακά χρωμοσώματα ταξινομήθηκαν σύμφωνα με το σχήμα 'dnaE'/'polC'/'dnaE2' (Qu et al. 2010). Εν τούτοις, δεν πραγματοποιήθηκαν υπολογισμοί των A-T και G-C αποκλίσεων, αλλά μελετήθηκαν οι συχνότητες των τεσσάρων βάσεων ξεχωριστά, ενώ τα αποτελέσματα ερμηνεύτηκαν στη βάση τόσο επιλεκτικών όσο και μεταλλακτικών πιέσεων.

## **1.6** *Επιδιορθωτικοί μηχανισμοί του* DNA *στους προκαρυωτικούς οργανισμούς*

Οι μεταλλακτικοί ρυθμοί στους οποίους υπόκειται το γονιδίωμα των προκαρυωτικών οργανισμών μεταβάλλονται σημαντικά κατά τη διάρκεια της εξελικτικής τους πορείας (Denamur & Matic 2006). Οι μηχανισμοί τροποποίησης και επιδιόρθωσης του DNA συμβάλλουν στην διαμόρφωση του μεταλλακτικού φάσματος των βακτηρίων. Η οριζόντια μεταφορά των γονιδίων που κωδικοποιούν

για αυτούς τους μηχανισμούς αποτελεί έναν σημαντικό παράγοντα που μπορεί να αλλάζει ριζικά τις λειτουργίες τροποποίησης και επιδιόρθωσης του DNA, οδηγώντας σε *έντονες αποκλίσεις των μεταλλακτικών ρυθμών ακόμα και μεταξύ στελεχών του ίδιου είδους*. Χαρακτηριστικό παράδειγμα είναι τα γονίδια που κωδικοποιούν για συστήματα περιοριστικών ενζύμων (restriction modification systems), τα οποία έχουν επανειλημμένως χαθεί και επανακτηθεί μέσω οριζόντιας μεταφοράς μεταξύ των βακτηρίων (Jeltsch & Pingoud 1996). Στα συστήματα αυτά συμμετέχουν και ένζυμα με ενεργότητα μεθυλάσης, η παρουσία των οποίων αυξάνει τη συχνότητα εμφάνισης μεθυλιωμένων καταλοίπων κυτοσίνης στο DNA, οδηγώντας σε υψηλούς ρυθμούς C→T μεταβάσεων, με συνέπεια την ενίσχυση των νουκλεοτιδικών αποκλίσεων μεταξύ των DNA κλώνων. Συχνή είναι επίσης η οριζόντια μεταφορά των γονιδίων που κωδικοποιούν για το σύστημα επιδιόρθωσης αταίριαστων ζευγών βάσεων (mismatch repair system, MMR), διαδικασία που θεωρείται πως έχει συμβάλλει στη γενετική ποικιλότητα (genomic diversity) των στελεχών του *E.coli* (Denamur et al. 2000). Καθώς το σύστημα MMR επιδιορθώνει αταίριαστα ζεύγη G·T, τα οποία προκύπτουν από την απαμίνωση μεθυλιωμένων καταλοίπων C, ενδέχεται να μειώνει τις νουκλεοτιδικές ασυμμετρίες των DNA κλώνων (Rocha & Danchin 2001).

### **1.6.1** *Η συζευγμένη με τη μεταγραφή επιδιόρθωση του DNA*

Πρώιμα γενετικά πειράματα που εξέταζαν το φαινόμενο της "πτώσης των μεταλλακτικών ρυθμών" ("mutation frequency decline", MFD), κατέδειξαν τη σύζευξη των επιδιορθωτικών συστημάτων του DNA με τον μηχανισμό της μεταγραφής (Bockrath & Cheung 1973, Bockrath & Palmer 1977, Engstrom et al. 1984, Bockrath et al. 1987). Οι πρώτες παρατηρήσεις αφορούσαν στην έντονη μείωση των μεταλλάξεων που επάγονταν από υπεριώδη ακτινοβολία, σε οργανισμούς που είχαν υποστεί συγκεκριμένους πειραματικούς χειρισμούς. Οι Bockrath και Palmer, μελετώντας ένα t-RNA γονίδιο του *E.coli*, κατέληξαν στο συμπέρασμα ότι η μείωση της συχνότητας των επαγόμενων μη-νοηματικών μεταλλάξεων (nonesense mutations) οφειλόταν σε επιδιόρθωση μέσω εκτομής μονάχα εκείνων των προ-μεταλλακτικών βλαβών (premutational lesions) που βρίσκονταν στον μεταγραφόμενο κλώνο (Bockrath & Palmer 1977). Το 1985, ο

Hanawalt και οι συνεργάτες του διαπίστωσαν πως οι ρυθμοί επιδιόρθωσης του DNA είναι έντονα υψηλότεροι στις μεταγραφικώς ενεργές DNA περιοχές σε σχέση με το σύνολο του γονιδιώματος (Bohr et al. 1985). Το προϊόν του *mfd* γονιδίου αναγνωρίστηκε ως ο παράγοντας σύζευξης μεταγραφής-επιδιόρθωσης (transcription-repair coupling factor, TRCF) από τον Selby και τους συνεργάτες του (Selby et al. 1991, C.P. Selby & Sancar 1993, Selby & Sancar 1994). Η συζευγμένη με τη μεταγραφή επιδιόρθωση (transcription coupled repair, TCR) έχει συσχετιστεί με την ύπαρξη ειδικών ανά κλώνο ρυθμών υποκατάστασης, που επηρεάζουν την εξέλιξη του DNA (Francino et al. 1996). Συμπερασματικά, η αλληλεπίδραση των επιδιορθωτικών συστημάτων του DNA με τον μηχανισμό της μεταγραφής επάγει ασυμμετρίες στους μεταλλακτικούς ρυθμούς μεταξύ κωδικών και μεταγραφόμενων κλώνων (Deaconescu 2013).

### **1.6.2** *Μονοπάτια επιδιόρθωσης του DNA, μη συζευγμένα με τη μεταγραφή*

Στα βακτήρια, τα μοριακά μονοπάτια επιδιόρθωσης του DNA οργανώνονται κατά διαφόρους τρόπους, καθώς οι πρωτεΐνες που συμμετέχουν σε ένα συγκεκριμένο μονοπάτι μπορεί να εμφανίζουν ειδικότητα σε ποικίλα υποστρώματα, ενώ άλλες μπορεί να είναι συμπληρωματικές ως προς τις λειτουργίες που επιτελούν (Morita et al. 2010, Mielecki & Grzesiuk 2014). Συνεπώς, η επιδιόρθωση του γενετικού υλικού διενεργείται από πληθώρα μονοπατιών, των οποίων η δράση μπορεί να είναι εν μέρει συμπληρωματική ή και να επικαλύπτεται. Η ποικιλομορφία αυτών των μοριακών συστημάτων οδηγεί σε μια έντονη ετερογένεια των προτύπων (modes) και των ρυθμών υποκατάστασης μεταξύ των βακτηρίων. Οι μηχανισμοί επιδιόρθωσης των βακτηριακών γονιδιωμάτων κατατάσσονται στις εξής κατηγορίες: (*i*) άμεση επιδιόρθωση της βλάβης (direct repair), (*ii*) επιδιόρθωση με εκτομή βάσης (base-excision repair, BER), (*iii*) επιδιόρθωση με εκτομή νουκλεοτιδίου (nucleotide-excision repair, NER), (*iv*) επιδιόρθωση αταίριαστων ζευγών βάσεων (mismatch repair, MMR), (*v*) επιδιόρθωση με ανασυνδυασμό (recombination repair, RR) (Morita et al. 2010, Resende et al. 2011). Ακολούθως, συνοψίζονται οι λειτουργίες των βασικών επιδιορθωτικών συστημάτων που απαντώνται στα βακτήρια.

**I) Άμεση Επιδιόρθωση:** Σε μία αντίδραση ενός βήματος, απομακρύνεται η χημική ομάδα που τροποποιεί τη νουκλεοτιδική βάση, δίχως εκτομή της βάσης (Nieminuszczy & Grzesiuk 2007).

**1)** Τα γονίδια ***ada*** και ***ogt*** κωδικοποιούν δύο μεθυλοτρανσφεράσες που επιδιορθώνουν αλκυλιωμένες βάσεις, όπως η $O^6$-μεθυλγουανίνη ($O^6$meG) και η $O^4$-μεθυλθυμίνη ($O^4$meT). Η $O^6$-μεθυλγουανίνη ζευγαρώνει με κατάλοιπα θυμίνης, οδηγώντας σε **G→A** μεταβάσεις (Rebeck & Samson 1991, Fukui 2010).

**2)** Η διοξυγενάση **AlkB** (Aravind & Koonin 2001) απομακρύνει μεθυλομάδες από τα τοξικά και μεταλλαξογόνα κατάλοιπα $N^1$-μεθυλαδενίνη (1meA) και $N^3$-μεθυλκυτοσίνη (3meC) (Delaney & Essigmann 2004).

**3)** Το γονίδιο ***phr*B** κωδικοποιεί μία φωτολυάση που επιδιορθώνει τα διμερή πυριμιδίνης (Goosen & Moolenaar 2008, Morita et al. 2010), όπως τα κυκλοβουτανικά διμερή πυριμιδίνης (cyclobutane pyrimidine dimers, CPDs) που σχηματίζονται κατόπιν έκθεσης του DNA σε UV ακτινοβολία.

**II) Επιδιόρθωση με εκτομή βάσης (Base Excision Repair, BER):** Πρόκειται για μονοπάτια στα οποία συμμετέχουν πολλές πρωτεΐνες και τα οποία εκκινούν εξειδικευμένες DNA γλυκοζυλάσες που αναγνωρίζουν και εκτέμνουν χημικά τροποποιημένες βάσεις (Lindahl 2001, Dizdaroglu 2005, Huffman et al. 2005, Resende et al. 2011).

**1)** Η ουρακιλ-DNA-γλυκοζυλάση **Ung** εκτέμνει ουρακίλες από τα αταίριαστα ζεύγη U·G και U·A (Ravishankar et al. 1998, Xiao et al. 1999, Mielecki & Grzesiuk 2014). Η απενεργοποίηση του γονιδίου *ung* σε knockout στελέχη του *E.coli* οδηγεί σε αύξηση της συχνότητας των **C→T** μεταβάσεων κατά μία τάξη μεγέθους ή και περισσότερο (Duncan & Weiss 1982).

**2)** Η DNA-γλυκοζυλάση **Mug** εκτέμνει κατάλοιπα ουρακίλης από το δίκλωνο DNA, αλλά εμφανίζει πολύ μεγαλύτερη δραστικότητα εκτομής καταλοίπων 3,$N^4$-εθενοκυτοσίνης (εC) που συναντώνται σε αταίριαστα ζεύγη εC·G (Saparbaev & Laval 1998, Mokkapati et al. 2001).

**3)** Οι DNA-γλυκοζυλάσες **Nth** και **Nei** καταλύουν την εκτομή οξειδωμένων πυριμιδινικών βάσεων (Suzuki et al. 2010), όπως οι προ-μεταλλαξογόνες οξειδωμένες μορφές της κυτοσίνης, οι οποίες ζευγαρώνουν με κατάλοιπα αδενίνης οδηγώντας σε **C→T** μεταβάσεις.

**4)** Οι γλυκοζυλάσες **AlkA** και **Tag** προστατεύουν το DNA από αλκυλιωτικούς παράγοντες. Και οι δύο εκτέμνουν κατάλοιπα $N^3$-μεθυλαδενίνης (3meA), τα οποία είναι έντονα κυτταροτοξικά (cytotoxic). Ωστόσο, η γλυκοζυλάση AlkA αναγνωρίζει και εκτέμνει ένα ευρύτερο φάσμα υποστρωμάτων (Mielecki & Grzesiuk 2014), στα οποία συγκαταλέγονται κατάλοιπα 1meA και 3meC που αποτελούν το βασικό υπόστρωμα της διοξυγενάσης AlkB (Mielecki et al. 2013).

**5)** Οι πρωτεΐνες **MutM** και **MutY**, μαζί με την **MutT** συγκροτούν το **GO** (8-οξο-γουανίνη) **σύστημα πρόληψης σφαλμάτων** (error prevention GO system). Το σύστημα αυτό καταπολεμά βλάβες του DNA που προκαλεί το οξειδωτικό στρες.

**5.*i*)** Τα οξειδωμένα κατάλοιπα γουανίνης (8OG) σχηματίζουν ζεύγη με dATP, εκτός από dCTP, κατά την αντιγραφή του DNA (Cheng et al. 1992), οδηγώντας σε **G→T** μεταπτώσεις (Wood et al. 1990). Η φορμαμιδοπυριμιδιν-DNA-γλυκοζυλάση **MutM** και η A/G-ειδική γλυκοζυλάση της αδενίνης **MutY** εκτέμνουν 8OG και αδενίνη από τα ζεύγη 8OG:C και 8OG:A pairs, αντίστοιχα (Dizdaroglu 2005).

**5.*ii*)** Το ένζυμο **MutT** υδρολύει τα 8-oxo-dGTP μονομερή του DNA που εντοπίζονται στην νουκλεοτιδική δεξαμενή, καταστέλλοντας την μεταλλαξιγένεση (Maki & Sekiguchi 1992). Μεταλλάξεις απενεργοποίησης του γονιδίου *mut*T (defective mutations) στο *E.coli* αυξάνουν τη συχνότητα των **G→T** μεταπτώσεων.

**III) Επιδιόρθωση αταίριαστων ζευγών (Mismatch Repair, MMR)**: επιδιόρθωση αταίριαστων ζευγών βάσεων που προκαλούνται κυρίως από σφάλματα κατά την αντιγραφή του DNA [*βλ.* σχετικές αναφορές στο (Fukui 2010)].

**1) Επιδιόρθωση μακρού τμήματος** (**Long-patch MMR**, εφεξής **MMR**). Πρόκειται για μετα-αντιγραφικό μηχανισμό επιδιόρθωσης, ο οποίος καταστέλλει μεταλλάξεις εξαρτώμενες από τη φάση ανάπτυξης (growth-dependent) των βακτηρίων, οι οποίες επάγονται από οξειδωτικές καταπονήσεις (Wyrzykowski & Volkert 2003), όπως στην περίπτωση σχηματισμού 8OG. Το ένζυμο MutS, που συμμετέχει στο MMR μονοπάτι, εμφανίζει ειδικότητα αναγνώρισης $O^6$meG:T αταίριαστων ζευγών (Rasmussen

& Samson 1996, Taira et al. 2008). Επιδιορθώνοντας κατάλοιπα 8OG και 6OmeG, το σύστημα MMR καταστέλλει **G→T** μεταπτώσεις και **G→A** μεταβάσεις, αντίστοιχα.

**1.*i*) MMR καθοδηγούμενη από μεθυλίωση (methyl-directed MMR):** Το μονοπάτι αυτό είναι ενεργό σε βακτήρια που φέρουν ομόλογα του *mut*H γονίδια (Heinze et al. 2009). Πρόκειται κυρίως για γ-Πρωτεοβακτήρια. Το ομοδιμερές του **MutS** αναγνωρίζει τα αταίριαστα ζεύγη βάσεων και στρατολογεί το **MutL**. Ακολούθως, ενεργοποιείται η λανθάνουσα (latent) ενδονουκλεάση **MutH**, η οποία στοχεύει τον νεοσυντιθέμενο κλώνο του DNA, που δεν έχει ακόμα μεθυλιωθεί, από τον οποίο εκτέμνει ένα ολιγονουκλεοτίδιο που περιλαμβάνει τη βάση του αταίριαστου ζεύγους (Längle-Rouault et al. 1987, Modrich 1989).

**1.*ii*) MMR καθοδηγούμενη από εγκοπή (nick-directed MMR):** Το μονοπάτι αυτό είναι ενεργό σε βακτήρια που στερούνται ομολόγων του *mut*H γονιδίου (Duppatla et al. 2009, Mauris & Evans 2009). Το σύμπλοκο **MutS**-**MutL** αναγνωρίζει παροδικές ασυνέχειες του φωσφοδιεστερικού σκελετού του DNA που εντοπίζονται κατά μήκος του συνοδού κλώνου αμέσως μετά την αντιγραφή (Fang & Modrich 1993, Constantin et al. 2005, Kadyrov et al. 2006, Fukui et al. 2008).

**2) Επιδιόρθωση πολύ βραχέος τμήματος** (**Very-short patch repair system, VSP).** Το σύστημα αυτό διορθώνει αταίριαστα ζεύγη T:G που προκύπτουν κατόπιν απαμίνωσης καταλοίπων 5-μεθυλ-κυτοσίνης. Ως αποτέλεσμα, το VSP σύστημα καταστέλλει **C→T** μεταβάσεις. Οι **MutS** και **MutL** επάγουν τη δράση της ενδονουκλεάσης **Vsr**, η οποία επιτελεί την επιδιόρθωση μέσω VSP (Heinze et al. 2009). Κατά τη διαδικασία αυτή εκτέμνεται μόνο το νουκλεοτίδιο που συμμετέχει στο αταίριαστο ζεύγος.

**IV) Επιδιόρθωση μέσω ανασυνδυασμού (Recombination repair, RR):** Τα μονοπάτια **RecFOR** και **RecBC** αποτελούν μονοπάτια ομόλογου ανασυνδυασμού που επιδιορθώνουν θραύσεις του ενός ή και των δύο κλώνων του DNA, αντίστοιχα (Deaconescu 2013). Στα μονοπάτια αυτά συμμετέχει ένα πλήθος ενζύμων [*βλ.* σχετικές αναφορές στο (Resende et al. 2011)].

**V) Επιδιόρθωση μέσω εκτομής νουκλεοτιδίων (Nucleotide Excision Repair,**

**NER):** Ένα υψηλά συντηρημένο σύστημα επιδιόρθωσης, που αναγνωρίζει ένα μεγάλο εύρος υποστρωμάτων, όπως τα επαγόμενα από UV-ακτινοβολία διμερή πυριμιδίνης ή πιο ογκώδη παράγωγα βάσεων [*βλ.* σχετικές αναφορές στο (Morita et al. 2010)].

**1) καθολική επιδιόρθωση του γονιδιώματος (Global Gemone Repair, GGR):** Η **UvrA** υπομονάδα ανιχνεύει βλάβες του DNA και μαζί με την **UvrB** υπομονάδα σχηματίζουν ένα σύμπλοκο προ-εκτομής (pre-incision complex). Ακολούθως, η **UvrC** εκτέμνει το ολιγονουκλεοτίδιο που περιέχει τη βλάβη. Η ελικάση **UvrD** απομακρύνει το ολιγονουκλεοτίδιο, δημιουργώντας ένα κενό που έρχεται να συμπληρώσει η DNA πολυμεράση. Μία λιγάση ενώνει το νεοσυντιθέμενο τμήμα του DNA με την αλυσίδα του μορίου [*βλ.* σχετικές αναφορές στο (Carvalho et al. 2005)].

**2) επιδιόρθωση συζευγμένη με τη μεταγραφή (Transcription-coupled Repair, TCR):** Ο παράγοντας **Mfd** αναγνωρίζει τα σύμπλοκα της RNA πολυμεράσης των οποίων η δράση παρεμποδίζεται από βλάβες του DNA, τα αναδομεί ή τα απομακρύνει από τον μεταγραφόμενο κλώνο, και ακολούθως κατευθύνει τον UvrABC μηχανισμό στον κλώνο αυτό (Deaconescu 2013). Στη συνέχεια η βλάβη του DNA επιδιορθώνεται μέσω του ίδιου μονοπατιού που δρα στην GGR.

## **1.7** *Κατανομές ολιγονουκλεοτιδίων στους DNA κλώνους*

### **1.7.1** *Συμμετρίες*

#### **1.7.1.1** *Αρχικές παρατηρήσεις*

Ο 2ος κανόνας της ισοδυναμίας (PR2) αφορά ενδοκλωνικές συμμετρίες που εμφανίζονται στο επίπεδο της μονονουκλεοτιδικής σύστασης του DNA. Αντίστοιχες συμμετρίες παρατηρούνται και στο επίπεδο νουκλεοτιδικών ολιγομερών. Μία από τις πρώτες σχετικές μελέτες (Prabhu 1993) κατέγραψε την κατά προσέγγιση ενδοκλωνική ισότητα των παρατηρούμενων συχνοτήτων, για αντιστρόφως συμπληρωματικές ν-αδες νουκλεοτιδίων, όπου ν = 2,3,4,5 και 6. Η συχνότητα εμφάνισης των ολιγονουκλεοτιδίων καθορίζεται τόσο από τη *συχνότητα* των μεμονωμένων βάσεων που τα συγκροτούν (base frequency), όσο και από τη *διάταξη* αυτών των βάσεων (base order).

Ο Forsdyke μελέτησε τα τρινουκλεοτίδια και την ιεραρχική τους κατάταξη, σύμφωνα με τη *συχνότητα* εμφάνισής τους, σε γονιδιώματα εξελικτικά απομακρυσμένων οργανισμών, από βακτήρια έως θηλαστικά (Forsdyke 1995). Συγκεκριμένα, εξετάστηκαν τα 32 ζεύγη αντιστρόφως συμπληρωματικών τρινουκλεοτιδίων με σκοπό την κατανόηση της σχέσης ανάμεσα στις ενδοκλωνικές συμμετρίες (single-strand symmetries) των μονονουκλεοτιδίων και των ολιγονουκλεοτιδίων, καθώς και στην εξάρτησή τους από το GC περιεχόμενο του DNA. Προκειμένου να εκτιμηθεί η επίδραση της *διάταξης* των βάσεων στις παρατηρούμενες συμμετρίες, δεν χρησιμοποιήθηκαν σταθμισμένες συχνότητες τρινουκλεοτιδίων, αλλά υπολογίστηκαν οι παρατηρούμενες συχνότητες εμφάνισής τους στο DNA και σε τεχνητές αλληλουχίες που προέκυψαν κατόπιν εκτεταμένης αναδιάταξης (shuffling) των φυσικών αλληλουχιών. Βάσει των σχετικών αποτελεσμάτων, προτάθηκε πως είναι η *συχνότητα* των βάσεων, και όχι η *διάταξή* τους, που καθορίζει την ισότητα των αντιστρόφως συμπληρωματικών τρινουκλεοτιδίων, καθώς αντίστοιχες ισότητες παρατηρήθηκαν τόσο στις φυσικές όσο και στις τεχνητές αλληλουχίες. Συνεπώς, σύμφωνα με την εν λόγω εργασία, οι ισοδυναμίες στο επίπεδο των τρινουκλεοτιδίων είναι το δευτερογενές αποτέλεσμα των αντίστοιχων ισοδυναμιών των νουκλεοτιδικών βάσεων που τα απαρτίζουν. Η *διάταξη* των βάσεων θεωρήθηκε ότι προσδιορίζει

την παρατηρούμενη συχνότητα εμφάνισης των τρινουκλεοτιδίων. Και εδώ, οι παρατηρήσεις ερμηνεύτηκαν ως έμμεσο αποτέλεσμα φαινομένων που εκδηλώνονται στο επίπεδο των δινουκλεοτιδίων. Έτσι, οι χαμηλές συχνότητες των οκτώ τρινουκλεοτίδίων που περιέχουν το δινουκλεοτίδιο TpA αποδόθηκε στην τάση υποεκπροσώπησης του συγκεκριμένου δινουκλεοτιδίου.

**1.7.1.2** *Προταθείσες ερμηνείες*

Το ερμηνευτικό πλαίσιο που προτείνει ο Forsdyke αναφέρεται στο σχηματισμό δομών στελέχους-θηλιάς (stem-loop) κατά μήκος του κάθε κλώνου. Ο εικαζόμενος σχηματισμός τέτοιων δομών θα μπορούσε να μεσολαβεί και να διευκολύνει τον ανασυνδυασμό. Επίσης, δομές στελέχους-θηλιάς ενδέχεται να σταθεροποιούν τα μονόκλωνα μόρια του mRNA. Προϋπόθεση για το σχηματισμό τέτοιων δομών είναι η ύπαρξη γειτονικών, αντιστρόφως συμπληρωματικών ολιγονουκλεοτιδίων κατά μήκος του ίδιου DNA κλώνου. Έτσι, στη μονόκλωνη κατάσταση, τα ολιγονουκλεοτίδια αυτά ή τα αντίστοιχα τμήματα των mRNAs (εάν πρόκειται για κωδικές περιοχές) θα μπορούσαν να σχηματίσουν το στέλεχος των εικαζόμενων δομών.

*Οι σχέσεις ισοδυναμίας που εντοπίζονται κατά μήκος του κάθε DNA κλώνου*, αναδιατυπωμένες στο επίπεδο του δίκλωνου μορίου, *δηλώνουν πως οι αντιστρόφως συμπληρωματικοί κλώνοι έχουν κατά προσέγγιση την ίδια σύσταση και ως εκ τούτου είναι συμμετρικοί*. Τέτοιες συμμετρίες εντοπίστηκαν σε αλληλουχίες μήκους 500bp, που βρίσκονται αναρροϊκά (upstream) των ανοιχτών πλαισίων ανάγνωσης (open reading frames - ORF) στο γονιδίωμα του *Saccharomyces cerevisiae* (Hampson et al. 2000). Εφαρμόστηκαν Μαρκοβιανά μοντέλα (Markov models) διαφόρων τάξεων (από $1^{ης}$ έως $9^{ης}$ τάξης) για την περιγραφή των αλληλουχιών. Τα αποτελέσματα υποδεικνύουν την ύπαρξη ενός συνδυασμού διαφορετικών μηχανισμών που επιβάλλουν πρωτογενώς περιορισμούς, τόσο κατώτερης (lower order) όσο και ανώτερης τάξης (higher order), στη σύσταση του γονιδιώματος. Συνεπώς, σύμφωνα με τη σχετική εργασία και σε αντίθεση με τα συμπεράσματα του Forsdyke (βλ. ενότητα 1.7.1.1), *εκδηλώνονται συσχετίσεις μακράς εμβέλειας* (*long range correlations*) *μεταξύ των βάσεων*, *οι οποίες δεν μπορούν να αναχθούν στην κατανομή των μονονουκλεοτιδίων*.

Σε κατοπινή μελέτη (Baisnée et al. 2002), εξετάστηκε εκτενώς η συμμετρία των κλώνων στο επίπεδο ολόκληρων χρωμοσωμάτων. Μία δεδομένη

κατανομή ολιγομερών *N*-τάξης επιβάλλει περιορισμούς στις πιθανές κατανομές ολιγομερών μικρότερης (έστω *M*) τάξης. Αντιστρόφως, μία δεδομένη κατανομή ολιγομερών *M*-τάξης μπορεί να επηρεάζει την κατανομή ολιγομερών *N*-τάξης, όπου $M < N$. Κατά τον τρόπο αυτό, συμμετρίες που παρατηρούνται στο επίπεδο τρινουκλεοτιδίων μπορεί να καθορίζουν τη συμμετρία στο επίπεδο των δινουκλεοτιδίων και αντιστρόφως. Διακρίνοντας ανάμεσα σε αυτές τις επιδράσεις, οι Baisnée και συνεργάτες επαλήθευσαν την ύπαρξη αυθεντικών (genuine) συμμετριών στη σύσταση των χρωμοσωμάτων, τόσο σε κατώτερο όσο και σε ανώτερο επίπεδο ν-μερών, όπου ν = 1 έως 9. Ως πιθανοί μηχανισμοί που επάγουν τις παρατηρούμενες συμμετρίες, αναφέρονται ο αναδιπλασιασμός μεγάλων τμημάτων του DNA (όπως τα γονίδια), η εισαγωγή μεταθετών στοιχείων (transposons) και ρετροϊών στα ευκαρυωτικά γονιδιώματα, καθώς και η αναστροφή κλώνων (strand invertion) μέσω ανασυνδυασμού, που είχε προηγουμένως προταθεί από τους Fickett *et al*. (1992).

### **1.7.2** *Ασυμμετρίες*

Όπως έχει ήδη αναφερθεί, οι συμμετρίες της ενδοκλωνικής σύστασης του DNA εμφανίζουν χαρακτηριστικές αποκλίσεις. Στα βακτήρια, οι ειδικές ανά κλώνο ασυμμετρίες της σύστασης του DNA αποτελούν το αντικείμενο πλήθους εργασιών. Στην πλούσια σχετική βιβλιογραφία, οι ασυμμετρίες αυτές μελετώνται πρωτίστως με όρους μονονουκλεοτιδικών αποκλίσεων. Αντιθέτως, οι ασυμμετρίες στις κατανομές των ολιγονουκλεοτιδίων δεν έχουν ερευνηθεί εξίσου εκτενώς. Ορισμένες από τις πρώτες συστηματικές αναφορές εντοπίζονται σε τρία άρθρα, όπου παρουσιάζονται τα πλήρως αλληλουχημένα γονιδιώματα ενός εντεροβακτηρίου, του *Escherichia coli* (Blattner et al. 1997), και δύο βακτηρίων που ανήκουν στις Σπειροχαίτες, των *Borrelia burgdorferi* (Fraser et al. 1997) και *Treponema pallidum* (Fraser 1998). Συγκεκριμένα, εντοπίστηκαν νουκλεοτιδικά οκταμερή, ο προσανατολισμός των οποίων είναι έντονα πολωμένος εκατέρωθεν του σημείου έναρξης της αντιγραφής. Ο πολωμένος προσανατολισμός δηλώνει πως τα συγκεκριμένα οκταμερή εμφανίζονται πολύ συχνότερα στον οδηγό από ότι στο συνοδό κλώνο. Επιπλέον, και στα τρία γονιδιώματα που εξετάστηκαν, τα οκταμερή με πολωμένο προσανατολισμό βρέθηκε

πως είναι σημαντικά υπερεκπροσωπημένα, δηλαδή οι συχνότητες εμφάνισής τους είναι σημαντικά υψηλότερες από εκείνες που θα αναμένονταν, με βάση τις υποκείμενες (underlying) συχνότητες των βάσεων που τα απαρτίζουν. Να σημειωθεί, ωστόσο, ότι *ο πολωμένος προσανατολισμός, από τη μία, και η υπερεκπροσώπηση, από την άλλη, είναι δύο ξεχωριστά φαινόμενα, που δεν προϋποθέτουν το ένα το άλλο*.

Στις σχετικές μελέτες, ιδιαίτερη έμφαση δόθηκε στα Chi ενεργά σημεία (hotspots) ανασυνδυασμού. Έτσι, στην περίπτωση του *E.coli* (Blattner et al. 1997), οι Chi αλληλουχίες αποτελούν το τρίτο κατά σειρά αφθονίας οκταμερές κατά μήκος του οδηγού κλώνου. Οι αλληλουχίες αυτές συμμετέχουν στην επιδιόρθωση βλαβών της διχάλας της αντιγραφής, μέσω ανασυνδυασμού (Kuzminov 1995). Επίσης, τα έντονα πολωμένα οκταμερή που εντοπίστηκαν έχουν υψηλό βαθμό ομοιότητας με τις Chi αλληλουχίες και δεν διαφέρουν από αυτές όσον αφορά υποκαταστάσεις οι οποίες να είναι γνωστό ότι απενεργοποιούν την ενεργότητα ανασυνδυασμού των Chi. Έτσι, ο πολωμένος προσανατολισμός αυτών των ολιγονουκλεοτιδίων αποδόθηκε στην δράση τους ως Chi αλληλουχίες. Παράλληλα, ένα τμήμα των Chi αλληλουχιών αντιστοιχεί στα σημεία πρόσδεσης της πριμάσης DnaG, η δράση της οποίας είναι απαραίτητη για τη σύνθεση των τμημάτων Okazaki. Το γεγονός αυτό από μόνο του είναι επαρκές για να εξηγήσει τον πολωμένο προσανατολισμό των Chi.

Μια πιο εκτενής μελέτη κατέδειξε πως οι κατανομές των ολιγονουκλεοτιδίων μεταξύ οδηγού και συνοδού κλώνου εμφανίζουν ποικίλα πρότυπα στα προκαρυωτικά γονιδιώματα (Salzberg et al. 1998). Συγκεκριμένα, εντοπίστηκαν γονιδιώματα στα οποία οι κατανομές των συχνοτήτων εμφάνισης είναι έντονα πολωμένες εκατέρωθεν του σημείου έναρξης της αντιγραφής για δεκάδες ή και εκατοντάδες ολιγονουκλεοτίδια. Αντίθετα, υπάρχουν γονιδιώματα όπου λίγα ολιγονουκλεοτίδια (λιγότερα από δέκα) ακολουθούν πολωμένες κατανομές, ενώ σε ορισμένα γονιδιώματα δεν εμφανίζονται σημαντικές διαφορές στην κατανομή κανενός ολιγονουκλεοτιδίου μεταξύ οδηγού και συνοδού κλώνου. Στην τελευταία αυτή κατηγορία περιλαμβάνονται αρκετά Αρχαία. Στις περιπτώσεις που εξετάστηκαν, *η περιοχή του χρωμοσώματος στην οποία η πόλωση της κατανομής των ολιγονουκλεοτιδίων αλλάζει περιλαμβάνει το σημείο έναρξης της αντιγραφής*.

Η ύπαρξη πολωμένων οκταμερών δεν μπορεί να ερμηνευθεί αποκλειστικά στη βάση διαφορών στους ρυθμούς μεταλλάξεων και επιδιορθώσεων μεταξύ των δύο

κλώνων της αντιγραφής. Οι διαδικασίες αυτές δεν μπορούν να παράγουν συστηματικά ακριβή αντίγραφα νουκλεοτιδικών οκταμερών κατά μήκος του ενός ή του άλλου κλώνου. Αντίθετα, η επίδρασή τους θεωρείται εντονότερη στο επίπεδο των μονονουκλεοτιδικών αποκλίσεων. Για την ερμηνεία του φαινομένου, η μελέτη των Salzberg *et al.* (1998) επικαλείται τη δράση επιλεκτικών μηχανισμών, αντίστοιχων με αυτούς που δρουν στην περίπτωση των Chi αλληλουχιών. Παράλληλα, η ειδική ανά κλώνο κατανομή νουκλεοτιδικών ολιγομερών, ιδίως εξαμερών, μπορεί να οφείλεται στη συνδυαστική δράση δύο επιμέρους μηχανισμών· στην πόλωση της κατανομής των γονιδίων και της χρήσης κωδικονίων. Στην περίπτωση αυτή, τα εξαμερή αντιστοιχούν σε ζεύγη κωδικονίων. Ωστόσο, διαπιστώθηκε πως η πόλωση στην κατανομή ολιγομερών είναι συχνά εντονότερη από αυτή των γονιδίων, ενώ πολωμένα εξαμερή ανιχνεύονται και όταν λαμβάνεται υπόψιν μόνο το μη κωδικό τμήμα των χρωμοσωμάτων (Salzberg et al. 1998).

Σε μια πιο πρόσφατη έρευνα, μελετήθηκαν τα πρότυπα αποκλίσεων από την ενδοκλωνική ισοδυναμία (intra-strand parity) στο επίπεδο των μονο- και τρι-νουκλεοτιδίων, σε γονιδιώματα φυτών και ζώων (Mascher et al. 2013). Σύμφωνα με τη μελέτη, οι τρινουκλεοτιδικές αποκλίσεις μπορούν να ομαδοποιηθούν σε τρεις διακριτές κατηγορίες, που αντιστοιχούν στα μονοκοτυλήδονα φυτά, στα δικοτυλήδονα φυτά και στα θηλαστικά. Επιπλέον, εντός της κάθε κατηγορίας, τα πρότυπα τρινουκλεοτιδικών αποκλίσεων εμφανίζουν μεγαλύτερη ομοιότητα μεταξύ πιο συγγενικών οργανισμών, από ότι μεταξύ οργανισμών που είναι πιο απομακρυσμένοι εξελικτικά. Η παρατηρούμενη ποικιλία των ειδικών ανά κλώνο ασυμμετριών στο επίπεδο των τρινουκλεοτιδίων, υποδεικνύει την *ύπαρξη μηχανισμών που μεταβάλλουν αυτές τις ασυμμετρίες της σύστασης των γονιδιωμάτων, στην κλίμακα του εξελικτικού χρόνου*.

## **1.8** *Σταθμισμένες συχνότητες δινουκλεοτιδίων*

Μεταξύ των νουκλεοτιδικών ολιγομερών, ιδιαίτερο ενδιαφέρον παρουσιάζουν τα δινουκλεοτίδια, καθώς αποτελούν την πρωταρχική μονάδα διάταξης των βάσεων

(primary ordering unit), είναι δηλαδή "τα πιο βασικά συστατικά της διάταξης" των νουκλεοτιδίων (Nussinov 1984a). Τα δινουκλεοτίδια αντιστοιχούν σε αλληλουχίες που απαρτίζονται από τις κοντινότερες γειτονικές βάσεις (nearest neighbor base sequences), και οι συχνότητες εμφάνισής τους μελετώνται ήδη από τη δεκαετία του '60, στο πλαίσιο βιοχημικών πειραμάτων που αποσκοπούσαν στη διαλεύκανση του μηχανισμού της αντιγραφής του DNA (Josse et al. 1961, Swartz et al. 1962).

Ο Subak-Sharp και οι συνεργάτες του, όρισαν τη σχετική αφθονία (relative abundance) των 16 διαφορετικών δινουκλεοτιδίων που απαντώνται στο DNA ως τη συχνότητα εμφάνισής τους, κανονικοποιημένη ως προς τη συχνότητα που θα αναμενόταν εάν τα νουκλεοτίδια κατανέμονταν τυχαία κατά μήκος της υπό εξέταση αλληλουχίας (Russell et al. 1976, Russell & Subak-Sharpe 1977). Πρόκειται, δηλαδή, για την *σταθμισμένη συχνότητα των δινουκλεοτιδίων* (dinucleotide odds-ratio). *Το μέτρο αυτό ποσοτικοποιεί τις προτιμήσεις που εμφανίζει κάθε νουκλεοτίδιο για τα γειτονικά του* (*προτιμήσεις γειτόνων*), *επιτρέποντας τη μελέτη της διάταξης των βάσεων, χωρίς να εμπλέκονται σε αυτή οι παρατηρούμενες συχνότητές τους*. Κατ' αυτό τον τρόπο είναι δυνατόν να ανιχνευθούν υποκείμενες ομοιότητες των αλληλουχιών, οι οποίες συγκαλύπτονται από τις μεταξύ τους διαφορές στη συνολική νουκλεοτιδική τους σύσταση. Όταν ένα δινουκλεοτίδιο υποεκπροσωπείται σε μια αλληλουχία, η σταθμισμένη συχνότητά του είναι μικρότερη της μονάδας. Αντίθετα, υπερεκπροσωπημένα δινουκλεοτίδια έχουν σταθμισμένες συχνότητες μεγαλύτερες της μονάδας.

### **1.8.1** *Πρότυπα και συμμετρίες*

Τα αποτελέσματα βιοχημικών πειραμάτων που πραγματοποιήθηκαν στο γονιδίωμα ιών και θηλαστικών [βλ. (Russell et al. 1976) και σχετικές αναφορές], οδήγησαν στο συμπέρασμα πως οι σταθμισμένες συχνότητες των δινουκλεοτιδίων συγκροτούν ένα σύνολο τιμών που παραμένει στα γενικά του χαρακτηριστικά σταθερό κατά μήκος περιοχών του DNA με διαφορετική λειτουργικότητα. Το σύνολο αυτών των τιμών ονομάστηκε "γενικό σχέδιο" ("general design") του DNA και η σταθερότητά του αποδόθηκε σε κοινές επιλεκτικές πιέσεις που

ασκούνται σε διαφορετικά τμήματα του γενετικού υλικού κατά την εξέλιξή του. Οι πιέσεις αυτές πιθανολογήθηκε πως συνδέονται με την στερεοχημική διαμόρφωση (conformation) των δινουκλεοτιδίων και την ενέργεια πακεταρίσματός τους (stacking energy).

Οι πρώτες εκτενείς περιγραφές των μοτίβων που εμφανίζονται στο επίπεδο της διάταξης των γειτονικών βάσεων πραγματοποιήθηκαν από τη Nussinov στις αρχές του '80 (Nussinov 1980, Nussinov 1981, Nussinov 1984b). Στις εργασίες αυτές αναλύεται η *ιεραρχική κατάταξη των δινουκλεοτιδίων* (dinucleotide hierarchies) βάσει των σταθμισμένων συχνοτήτων τους. Η κατάταξη αυτή εμφανίζει ποικίλες κανονικότητες. Μεταξύ αυτών, η Nussinov παρατήρησε ότι, κατά μήκος του κάθε DNA κλώνου, *η σταθμισμένη συχνότητα κάθε δινουκλεοτιδίου ισούται κατά προσέγγιση με αυτή του αντιστρόφως συμπληρωματικού του* (Nussinov 1984b).

Στο τέλος της δεκαετίας του '80, ο Ohno, στηριζόμενος στις σταθμισμένες συχνότητες δινουκλεοτιδίων, πρότεινε ότι η υποεκπροσώπηση των TA και CG και η υπερεκπροσώπηση των TG και CT είναι ένα καθολικό φαινόμενο στα γονιδιώματα και αποτελεί βασικό κανόνα στον οποίο υπόκειται η οργάνωση των κωδικών περιοχών (Ohno 1988). Ακολούθως, οι Yomo και Ohno παρατήρησαν πως ο κανόνας αυτός δεν περιορίζεται στις κωδικές περιοχές, αλλά επεκτείνεται και στις μη-κωδικές, και κατ' αυτό τον τρόπο συμβάλλει στην εναρμόνιση της εξέλιξης (concordant evolution) των περιοχών αυτών (Yomo & Ohno 1989). Στην ίδια εργασία, διαπίστωσαν πως η συμμετρία των αντιστρόφως συμπληρωματικών κλώνων του DNA διατηρείται στο επίπεδο των δι- και τρι-νουκλεοτιδίων, στις μη-κωδικές περιοχές αλλά και στις κωδικές, αν και σε μικρότερο βαθμό. Τα σχετικά αποτελέσματα αφορούν τόσο τις παρατηρούμενες όσο και τις σταθμισμένες συχνότητες ολιγονουκλεοτιδίων.

Περαιτέρω έρευνες πραγματοποιήθηκαν σχετικά με την υπο- και υπερ-εκπροσώπηση μικρού μήκους ολιγονουκλεοτιδίων (δι-, τρι- και τετρα-νουκλεοτιδίων) στο DNA ιών, βακτηρίων και ευκαρυωτικών οργανισμών, στους οποίους μελετήθηκε τόσο το πυρηνικό DNA όσο και αυτό των οργανιδίων (μιτοχονδρίων και χλωροπλαστών) (Burge et al. 1992). Στη συγκεκριμένη μελέτη καταγράφονται εκτενώς τάσεις υπο-/υπερ-εκπροσώπησης συγκεκριμένων δινουκλεοτιδίων, οι οποίες απαντώνται στο DNA οργανισμών που καλύπτουν ένα ιδιαίτερα ευρύ φυλογενετικό φάσμα.

### **1.8.2** *Υπο- και υπερ-εκπροσωπούμενα δινουκλεοτίδια*

Ακολούθως, καταγράφονται ενδεικτικά παραδείγματα περιπτώσεων υπό- και υπέρ-εκπροσώπησης δινουκλεοτιδίων οι οποίες αποτελούν γενικές τάσεις προτίμησης γειτόνων που παρατηρούνται στα γονιδιώματα πολύ διαφορετικών μεταξύ τους οργανισμών.

Η υποεκπροσώπηση του TA είναι ευρύτατα διαδεδομένη σε πλήθος γονιδιωμάτων. Η συχνή εμφάνιση διαδοχικών TA δινουκλεοτιδίων πιθανώς επηρεάζει αρνητικά την υπερελίκωση (supercoilong) του DNA και/ή τη δομή της χρωματίνης. Επίσης, μεταξύ όλων των δινουκλεοτιδίων, το TA χαρακτηρίζεται θερμοδυναμικά από την μικρότερη ενέργεια πακεταρίσματος. Συγχρόνως, το TA αποτελεί συστατικό ολιγονουκλεοτιδίων τα οποία επιτελούν σημαντικές ρυθμιστικές λειτουργίες στα ευκαρυωτικά γονιδιώματα, όπως είναι οι "TATA box" αλληλουχίες, που εμπλέκονται στην έναρξη της μεταγραφής, και οι σηματοδοτικές αλληλουχίες τερματισμού της μεταγραφής (transcription termination signals). Ως εκ τούτου, το TA μπορεί να είναι αντικείμενο αρνητικής επιλογής, προκειμένου να ελαχιστοποιείται η πιθανότητα ακατάλληλης πρόσδεσης παραγόντων έναρξης και λήξης της μεταγραφής. Η υποεκπροσώπηση του TA στις κωδικές περιοχές συσχετίζεται με την χαμηλή περιεκτικότητα πολλών πρωτεϊνικών μορίων σε κατάλοιπα τυροσίνης (που κωδικοποιούνται από τις τριπλέτες TAY) και με την αποφυγή μη νοηματικών μεταλλάξεων, που θα οδηγούσαν στην πρόωρη εμφάνιση μια TAR τριπλέτας σε ένα ανοιχτό πλαίσιο ανάγνωσης.

Το χαμηλό περιεχόμενο των αλληλουχιών σε CG αποδίδεται στους υψηλούς ρυθμούς μεταλλάξεων που υφίστανται τα κατάλοιπα κυτοσίνης όταν απαντώνται στο δινουκλεοτίδιο CG. Συγκεκριμένα, μέσω μεθυλίωσης (C→5-mC) και/ή απαμίνωσης (5-mC→T ή C→U), κατάλοιπα κυτοσίνης υποκαθίστανται από θυμίνη ή ουρακίλη. Ακολούθως, επιδιορθωτικοί μηχανισμοί του DNA αντικαθιστούν την ουρακίλη με κατάλοιπα θυμίνης. Ως αποτέλεσμα, τα δινουκλεοτίδια CG αντικαθίστανται από TG. Σε συμφωνία με το μηχανισμό αυτό είναι και η έντονη υπερεκπροσώπηση των TG, που παρατηρείται στα γονιδιώματα των σπονδυλωτών. Στους ευκαρυωτικούς, η μεθυλίωση των δινουκλεοτιδίων CG θεωρείται ρυθμιστικό σήμα για την καταστολή της έκφρασης των γονιδίων (Cedar & Razin 1990), ενώ συσχετίζεται με ανώτερης τάξης δομές της χρωματίνης (Tazi & Bird 1990). Η μεθυλίωση σταθεροποιεί τα χρωμοσώματα και μεσολαβεί στην

κληρονομικότητα της κατάστασης της δομής της χρωματίνης (chromatin state). Αξιοσημείωτη είναι η υποεκπροσώπηση του CG και στο μιτοχονδριακό DNA, καθώς εκτιμάται πως στα συγκεκριμένα οργανίδια η ενεργότητα μεθυλάσης απουσιάζει ή εμφανίζεται σε πολύ χαμηλά επίπεδα (Nass 1973, Dawid 1974). Ωστόσο, η εκτίμηση αυτή έχει πρόσφατα ανασκευαστεί από πειράματα που κατέδειξαν ότι η επιγενετική τροποποίηση των καταλοίπων C στο μιτοχονδριακό DNA είναι πολύ πιο εκτεταμένη από ότι παλαιότερα είχε εκτιμηθεί (Shock et al. 2011). Στα βακτήρια, η σχετική αφθονία του CG δεν ακολουθεί ένα τόσο σαφές πρότυπο υποεκπροσώπησης, όπως συμβαίνει στους ευκαρυωτικούς. Αντιθέτως, υπάρχουν περιπτώσεις γονιδιωμάτων όπου το CG στην πραγματικότητα υπερεκπροσωπείται. Στα βακτηριακά χρωμοσώματα, τα πλέον υποεκπροσωπημένα δινουκλεοτίδια είναι τα TA, GT, ενώ ακολουθούν τα AC και AT, με ορισμένες διακυμάνσεις (Shioiri & Takahata 2001).

Διαδοχικές επαναλήψεις ενός μόνο νουκλεοτιδίου εμφανίζονται με υψηλή συχνότητα στα γονιδιώματα πολλών οργανισμών. Τέτοιες επαναλήψεις υπερεκπροσωπούνται πρωτίστως στο επίπεδο των δινουκλεοτιδίων. Ομοδιμερή A ή T προκύπτουν συχνά λόγω γλιστρήματος της πολυμεράσης (polymerase slippage) κατά την αντιγραφή του DNA. Στα μιτοχονδριακά γονιδιώματα, παρά τη χαμηλή συχνότητα εμφάνισης G και C, τα ομοδιμερή CC·GG είναι τα πιο υπερεκπροσωπημένα δινουκλεοτίδια. Η αφθονία καταλοίπων γλυκίνης (που κωδικοποιείται από τριπλέτες GGN) στα πρωτεϊνικά μόρια πιθανώς συνεισφέρει στην υπερεκπροσώπηση CC·GG, καθώς ένα σημαντικό τμήμα των μιτοχονδριακών DNA αντιστοιχεί σε κωδικές περιοχές (Burge et al. 1992).

### **1.8.3** *Γονιδιωματικές υπογραφές.*

#### **1.8.3.1** *Ορισμός και αρχικές παρατηρήσεις*

Οι τιμές των σταθμισμένων συχνοτήτων των δινουκλεοτιδίων παρέχουν ένα μέτρο προκειμένου να εκτιμηθεί ο βαθμός της ομοιογένειας των γονιδιωμάτων. Ο Karlin και οι συνεργάτες του μελέτησαν αυτές τις σταθμισμένες συχνότητες, τροποποιημένες έτσι ώστε οι τιμές τους να είναι συμμετρικές ως προς τους δύο κλώνους του DNA (Karlin, Mocarski, et al. 1994, Karlin, Ladunga, et al. 1994, Karlin & Cardon 1994, Karlin & Ladunga 1994). Οι *συμμετρικές*

σταθμισμένες δινουκλεοτιδικές συχνότητες αναφέρονται με τον συμβολισμό *ρ**. Το σύνολο των τιμών των *ρ** είναι χαρακτηριστικό της κάθε αλληλουχίας και συγκροτεί την "γονιδιωματική υπογραφή" της. Για κάθε δεδομένο ζεύγος αλληλουχιών DNA, η απόσταση Manhattan (rectilinear ή Manhattan distance) μεταξύ των αντίστοιχων γονιδιωματικών υπογραφών καλείται *δ*-απόσταση (βλ. ενότητα 2.4).

Οι τιμές των *δ*-αποστάσεων είναι σημαντικά μικρότερες μεταξύ τμημάτων του ίδιου γονιδιώματος, από ότι μεταξύ διαφορετικών οργανισμών, παρόλο που η μονονουκλεοτιδική σύσταση (GC περιεχόμενο) εντός του ίδιου χρωμοσώματος συχνά παρουσιάζει έντονες διακυμάνσεις, όσον αφορά τους ευκαρυωτικούς οργανισμούς. Οι συγκρίσεις των *δ*-αποστάσεων κατέδειξαν ότι *οι σταθμισμένες συχνότητες των δινουκλεοτιδίων, ενώ είναι παρεμφερείς σε διαφορετικά τμήματα του γονιδιώματος του ίδιου οργανισμού* ("γενικό σχέδιο"), *διαφέρουν μεταξύ των οργανισμών*. Η ιδέα των "γονιδιωματικών υπογραφών", που εισηγήθηκαν οι Karlin και Burge (1995), επεκτείνει τα αποτελέσματα των βιοχημικών πειραμάτων που καθιέρωσαν την έννοια του "γενικού σχεδίου", όπως αυτό εκφράζεται σε όρους προτίμησης γειτονικών βάσεων. Συγκεκριμένα, το DNA συγγενικών οργανισμών εμφανίζει μεγαλύτερη ομοιότητα ως προς το "γενικό σχέδιό" του, σε σύγκριση με το DNA εξελικτικά απομακρυσμένων οργανισμών (Karlin 1998). Έτσι, η έννοια ενός "γενικού σχεδίου" του DNA παραχώρησε την θέση της σε αυτή των καθορισμένων ανά είδος (species-specific) "γονιδιωματικών υπογραφών", οι οποίες είναι χαρακτηριστικές της εξελικτικής πορείας του κάθε οργανισμού. Οι ομοιότητες ή οι διαφορές των γονιδιωματικών υπογραφών διαφορετικών οργανισμών αντανακλώνται στις τιμές των *δ*-αποστάσεων.

Οι *δ*-αποστάσεις χρησιμοποιήθηκαν στη μελέτη της μοριακής εξέλιξης και των φυλογενετικών σχέσεων των ερπητοϊών (Karlin & Ladunga 1994), καθώς και σε συγκρίσεις μεταξύ ευκαρυωτικών γονιδιωμάτων (Karlin & Ladunga 1994). Επίσης, μελετήθηκαν οι γονιδιωματικές υπογραφές των προκαρυωτικών οργανισμών και των οργανιδιακών αλληλουχιών DNA, μιτοχονδριακών και πλαστιδιακών (Campbell et al. 1999). Παρ' ότι, για τον ίδιο οργανισμό, οι γονιδιωματικές υπογραφές του μιτοχονδριακού (Mt) και του πυρηνικού DNA διαφέρουν μεταξύ τους, οι *δ*-αποστάσεις μεταξύ των Mt DNA αλληλουχιών διαφορετικών οργανισμών συσχετίζονται με τις *δ*-αποστάσεις μεταξύ του πυρηνικού DNA των ίδιων οργανισμών. Σε αντίθεση με τις φυλογενετικές μεθόδους που βασίζονται σε συγκρίσεις αμινοξικών αλληλουχιών, οι

*δ*-αποστάσεις παρέχουν τη δυνατότητα φυλογενετικής ανακατασκευής δίχως να απαιτείται ο εντοπισμός ομόλογων αλληλουχιών και η στοίχισή τους. Επίσης, η χρήση των *δ*-αποστάσεων δεν εξαρτάται από την επιλογή συγκεκριμένων DNA περιοχών, αλλά, αντίθετα, λαμβάνει υπόψιν την συνολική σύσταση του γονιδιώματος. Μελετώντας τις *δ*-αποστάσεις μιας μεγάλης συλλογής χρωμοσωμάτων, ο Karlin εξέτασε τις εξελικτικές σχέσεις οργανισμών, όπως οι *Rickettsia prowazekii* και *Helicobacter pylori*, η ταξινομική κατάταξη των οποίων ήταν αμφιλεγόμενη (Karlin 1998). Στην ίδια μελέτη, οι γονιδιωματικές υπογραφές προσέφεραν μια αρκετά συνεκτική περιγραφή των υπό αμφισβήτηση φυλογενετικών σχέσεων των Αρχαίων, σύμφωνα με την οποία τα Αρχαία δεν έχουν μονοφυλετική προέλευση, αλλά αντιπροσωπεύουν μια πολυφυλετική ομάδα.

### **1.8.3.2** *Ενδο-ειδική σταθερότητα και δια-ειδική ετερογένεια*

Το περιορισμένο εύρος τιμών που λαμβάνουν οι δινουκλεοτιδικές σταθμισμένες συχνότητες κατά μήκος διαφορετικών περιοχών του ίδιου χρωμοσώματος ή διαφορετικών χρωμοσωμάτων του ίδιου γονιδιώματος, δηλαδή η σταθερότητα των γονιδιωματικών υπογραφών, συσχετίζεται με συντηρημένα χαρακτηριστικά της δομής του DNA. Στο πλαίσιο αυτό, οι σταθμισμένες συχνότητες δινουκλεοτιδίων εμφανίζονται να εκφράζουν τοπικές ιδιαιτερότητες στο επίπεδο μάλλον της δομής του DNA και όχι της ίδιας της αλληλουχίας των βάσεων. Συγκεκριμένα, αντανακλούν τις φυσικοχημικές αλληλεπιδράσεις των γειτονικών νουκλεοτιδίων, έτσι όπως αυτές καθορίζονται από την ενέργεια πακεταρίσματος (stacking energy) και τις προτιμήσεις διαμόρφωσης (base step conformational preferences) στα ζεύγη διαδοχικών βάσεων. Οι αλληλεπιδράσεις αυτές διαμορφώνουν την καμπυλότητα (curvature), την υπερελίκωση (supercoiling) καθώς και άλλα ανώτερης τάξης δομικά χαρακτηριστικά του DNA.

Οι διαφορές μεταξύ των γονιδιωματικών υπογραφών εξελικτικά απομακρυσμένων οργανισμών πιθανολογείται πως αντιστοιχούν στις συγκεκριμένες ανά είδος (species-specific) ιδιότητες των μοριακών μηχανισμών αντιγραφής, τροποποίησης και επιδιόρθωσης του DNA. Συγκεκριμένα, *έχουν ανιχνευθεί πολώσεις της αντιγραφής προς συγκεκριμένα νουκλεοτίδια, των οποίων ο ρυθμός ενσωμάτωσης στο νεοσυντιθέμενο κλώνο εξαρτάται από τις γειτονικές τους βάσεις*. Επίσης, οι μηχανισμοί τροποποίησης και επιδιόρθωσης, εκτός από την ίδια την αλληλουχία του DNA, αναγνωρίζουν και αλληλεπιδρούν με σχήματα (shapes) και βλάβες (lesions) της στερεοδιάταξής

του, συμβάλλοντας στην παρατηρούμενη ετερογένεια μεταξύ των διαφορετικών γονιδιωματικών υπογραφών. Τέλος, *οι ρυθμοί των μεταλλάξεων μιας βάσης επηρεάζονται έντονα από τους πρώτους γείτονές της*, γεγονός που έχει ως αποτέλεσμα οργανισμοί με διαφορές στο μεταλλακτικό τους φάσμα να διαφέρουν και ως προς το σύνολο των δινουκλεοτιδίων σταθμισμένων συχνοτήτων τους.

### **1.8.3.3** *Γονιδιωματικές υπογραφές και ασυμμετρίες των κλώνων του DNA*

Οι Mrázek και Karlin (1998) εξέτασαν κατά πόσο οι σταθμισμένες συχνότητες των δινουκλεοτιδίων εμφανίζουν ειδικές ανά κλώνο ασυμμετρίες, αντίστοιχες με εκείνες που εντοπίζονται στο επίπεδο της μονονουκλεοτιδικής σύστασης του DNA. Η μελέτη τους αφορούσε 11 προκαρυωτικά γονιδιώματα και 10 γονιδιώματα ερπητοϊών. Συγκρίσεις πραγματοποιήθηκαν μεταξύ των σταθμισμένων συχνοτήτων των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων, όπως αυτές υπολογίζονται κατά μήκος του οδηγού και του συνοδού κλώνου, ξεχωριστά. Βασισμένοι στον περιορισμένο αριθμό των έως τότε πλήρως αλληλουχημένων γονιδιωμάτων, οδηγήθηκαν στο συμπέρασμα ότι "η σχετική αφθονία των δινουκλεοτιδίων τείνει να είναι συμμετρική και ιδιαίτερα σταθερή όσον αφορά τον οδηγό και το συνοδό κλώνο, παρά την ειδική ανά κλώνο ασυμμετρία της σύστασης" του DNA. Οι παρατηρήσεις περί συμμετρίας των σταθμισμένων συχνοτήτων (σχετική αφθονία) των δινουκλεοτιδικών συνδέθηκαν με την σταθερότητα των γονιδιωματικών υπογραφών, την οποία θεωρήθηκε πως ενισχύουν. Έτσι, ενώ ο τύπος και ο ρυθμός των μεταλλάξεων εξαρτώνται από το πλαίσιο εντός του οποίου εμφανίζεται κάθε βάση (context-dependent mutations), οι ειδικές ανά κλώνο μεταλλακτικές πολώσεις που οδηγούν στις ασυμμετρίες της σύστασης του DNA θεωρήθηκε πως δρουν στο επίπεδο των μεμονωμένων βάσεων.

## **1.9** *Το GC περιεχόμενο και η εξελικτική δυναμική του βακτηριακού γονιδιώματος*

Το επί τοις εκατό GC περιεχόμενο (GC%) ποικίλει μεταξύ των βακτηριακών γονιδιωμάτων, λαμβάνοντας μέσες τιμές που κυμαίνονται από ~20% έως ~80%.

Αυτή η δια-ειδική (inter-specific) ετερογένεια συνοδεύεται από μία έντονη ομοιομορφία του GC% εντός του γονιδιώματος ενός εκάστου των βακτηρίων. Η κατανομή και η εξέλιξη του GC% των νουκλεϊκών οξέων αποτελεί αντικείμενο εκτεταμένης έρευνας ήδη από τις αρχές του '60.

### **1.9.1** *Ένα μοντέλο της εξέλιξης του GC%*

Μελετώντας τη διασπορά του GC% εντός του ίδιου DNA μορίου και την ετερογένεια διαφορετικών DNA ως προς το GC%, ο Sueoka πρότεινε ένα εξελικτικό μοντέλο στο οποίο λαμβάνεται υπόψιν μόνο η μετατροπή των ζευγών A·T ή T·A (*α* ζεύγη) προς G·C ή C·G (*γ* ζεύγη), και αντιστρόφως (Sueoka 1962). Οι ρυθμοί μετατροπής (conversion rates) των βάσεων ανά γενεά, έστω *u* για *γ→α* και *v* για *α→γ*, θεωρήθηκε πως κατανέμονται ομοιόμορφα σε όλες τις θέσεις του DNA καθενός οργανισμού. Επίσης, οι τιμές των *u* και *v* λαμβάνονται ως εάν οι ρυθμοί αυτοί να αντιστοιχούν στους ρυθμούς των αλλαγών που επιβιώνουν και μεταφέρονται στην επόμενη γενεά, δηλαδή ως ρυθμοί τελικών υποκαταστάσεων και όχι αρχικών μεταλλάξεων. Στη βάση αυτών των παραδοχών, το GC% εμφανίζεται ως ένα ιδιαίτερα σταθερό χαρακτηριστικό των γονιδιωμάτων, με σημαντικά περιορισμένο εύρος διασποράς κατά μήκος κάθε DNA μορίου. Η κατάσταση ισορροπίας της σύστασης του DNA, σε όρους GC%, ( $\hat{p}$ ) προσδιορίζεται από τη σχέση $\hat{p}=\frac{v}{u+v}$ . Η ομοιογενής κατανομή των ρυθμών *u* και *v* κατά μήκος του γονιδιώματος ενός δεδομένου οργανισμού θεωρήθηκε πως αντιστοιχεί σε μόρια DNA τα οποία είναι εκτεθειμένα σε όμοιο μεταβολικό περιβάλλον και συνεπώς η μεταλλαξογόνος δράση χημικών και άλλων παραγόντων, όπως η ιονίζουσα ακτινοβολία, αναμένεται να επηρεάζει ομοιογενώς τα μόρια αυτά. Σύμφωνα με το εν λόγω μοντέλο, *σε οργανισμούς με παρεμφερές ενδοκυτταρικό περιβάλλον*, ο λόγος *u*/*v* λαμβάνει παρόμοιες τιμές και συνεπώς *η σύσταση του DNA τους συγκλίνει*. Ως εκ τούτου, τα γονιδιώματα εξελικτικά συγγενικών οργανισμών αναμένεται να έχουν όμοια σύσταση.

### **1.9.2** *Μεταλλακτικές πιέσεις που διαμορφώνουν το GC%*

Παρότι οι γνώσεις σχετικά με την οργάνωση του DNA ήταν περιορισμένες όταν προτάθηκε το συγκεκριμένο μοντέλο, η απλότητα και η συνοχή του οδήγησαν σε αποτελέσματα που συμφωνούν εν πολλοίς με τα χαρακτηριστικά των βακτηριακών γονιδιωμάτων. Οι Bernardi και Bernardi (1985) μελετώντας το γονιδίωμα προκαρυωτών, ιών, κατώτερων ευκαρυωτών, ασπόνδυλων και σπονδυλωτών, διαπίστωσαν πως το GC% των $3^{ων}$ κωδικών θέσεων ($GC_3$%) συσχετίζεται γραμμικά με το συνολικό GC% των κωδικών περιοχών. Στην περίπτωση των προκαρυωτικών, των οποίων το γονιδίωμα καλύπτεται κατά το μεγαλύτερο μέρος του από κωδικές περιοχές, γραμμική συσχέτιση υπάρχει και μεταξύ $GC_3$% και ολικού GC% του γονιδιώματος. Επεκτείνοντας αυτές τις συσχετίσεις, οι Muto και Osawa (1987) έδειξαν ότι λειτουργικά διαφορετικές περιοχές του DNA των βακτηρίων έχουν GC% το οποίο συσχετίζεται θετικά με το συνολικό GC% του γονιδιώματός τους, είναι δηλαδή πολωμένο προς την ίδια κατεύθυνση. Τις παρατηρήσεις αυτές τις απέδωσαν σε πολωμένες μεταλλακτικές πιέσεις που ασκούνται στο σύνολο του γονιδιώματος και "σπρώχνουν" το GC% προς χαμηλότερες ή υψηλότερες τιμές. Οι μεταλλακτικές αυτές τάσεις καλούνται A·T/G·C πιέσεις και είναι το ισοδύναμο του λόγου *u*/*v* στο εξελικτικό μοντέλο που πρότεινε ο Sueoka.

Αν και η A·T/G·C πίεση ασκείται ομοιογενώς στο γονιδίωμα, διαφορετικές περιοχές του DNA αποκρίνονται λιγότερο ή περισσότερο έντονα στις συνολικές μεταβολές του GC%, σύμφωνα με επιλεκτικούς περιορισμούς ανάλογους με τη λειτουργική σημασία αυτών των περιοχών. Οι A·T/G·C πιέσεις ποικίλουν τόσο ως προς στην κατεύθυνση όσο και ως προς την έντασή τους μεταξύ διαφορετικών γενεαλογιών, οδηγώντας στην παρατηρούμενη ετερογένεια του GC% διαφορετικών βακτηρίων. *Οι A·T/G·C πιέσεις διαδραματίζουν πρωτεύοντα ρόλο στη διαφοροποίηση* (diversification) *των βακτηριακών γονιδιωμάτων ενώ συσχετίζονται και με τα πρότυπα χρήσης κωδικονίων*. Έτσι, τα χαμηλού GC% βακτήρια χρησιμοποιούν κωδικόνια πολωμένα προς κατάλοιπα A+T (πχ. *Mycoplasma capricolum*), ενώ στα υψηλού GC% βακτήρια παρατηρείται πόλωση της χρήσης κωδικονίων προς τα κατάλοιπα G+C (πχ. *Streptomyces vanaceus*). Επιπλέον, A·T πιέσεις έχουν συνδεθεί με την εναλλακτική χρήση κωδικονίων λήξης, τα οποία σε ορισμένα βακτήρια, όπως τα βλεφαριδοφόρα (ciliates), αποδίδονται για την κωδικοποίηση συγκεκριμένων αμινοξέων (Jukes et al. 1987).

### **1.9.3** *Η εξέλιξη του GC% και το αμινοξικό περιεχόμενο των πρωτεϊνών*

Σε μία πρωτοπόρα μελέτη, ο Sueoka (1961) εξέτασε την επίδραση του γονιδιωματικού GC% στη μέση αμινοξική σύσταση των πρωτεϊνών. Οι μετρήσεις αφορούσαν 11 βακτηριακά είδη, όπου προσδιορίστηκε η συγκέντρωση 14 αμινοξέων. Διαπιστώθηκε πως αυξανομένου του GC% αυξάνονται τα κατάλοιπα Ala, Arg, Gly και Pro, ενώ μειώνονται εκείνα των Ile, Lys, Tyr και Phe. Ακολούθως, ο Lobry (1997) ανέλυσε την αμινοξική σύσταση των πρωτεϊνών ως συνάρτηση του γονιδιωματικού GC% σε 59 βακτηριακά είδη, διακρίνοντας μεταξύ ενσωματωμένων μεμβρανικών (integral membrane proteins, IMP) και μη- (non-IMP) πρωτεϊνών. Θέτοντας ως μοναδική ανεξάρτητη μεταβλητή (predictive variable) το συνολικό GC% του γονιδιώματος, εξήγαγε τις αναμενόμενες συγκεντρώσεις των αμινοξικών καταλοίπων στις κωδικοποιούμενες πρωτεΐνες. Οι πιο σημαντικές αποκλίσεις μεταξύ αναμενόμενων και παρατηρούμενων συγκεντρώσεων αποδόθηκαν σε επιλεκτικές πιέσεις που ασκούνται στη βάση του διαχωρισμού των πρωτεϊνών σε IMP και non-IMP, με τις πρώτες να είναι εμπλουτισμένες σε υδρόφοβα κατάλοιπα και τις δεύτερες σε πολικά. Και στις δύο κατηγορίες παρατηρείται αποφυγή της χρήσης Cys, ώστε να μη σχηματίζονται ανεπιθύμητες δισουλφιδικές γέφυρες (disulfide bridges), ενώ αρνητική επιλογή ασκείται και στα κατάλοιπα Pro, τα οποία τροποποιούν δραστικά τη δομή των πρωτεϊνών, παρεμποδίζοντας το σχηματισμό δομών α-έλικας. Συμπερασματικά, η συσχέτιση του GC% με το αμινοξικό περιεχόμενο των πρωτεϊνών αποδίδεται στην εξέλιξη του γενετικού κώδικα. Στο πλαίσιο αυτό, η δράση της φυσικής επιλογής προσαρμόζει την μέση συγκέντρωση των αμινοξικών καταλοίπων σε ορισμένες βέλτιστες τιμές.

*Σειρά ερευνών υποστηρίζουν την συνεξέλιξη* (coevolution) *του GC% και της συχνότητας των αμινοξέων στα βακτήρια*. Η θεωρία της ουδέτερης εξέλιξης των βιολογικών μορίων προβλέπει την έντονη συσχέτιση των μεταλλακτικών προτύπων του DNA με την αμινοξική σύσταση των πρωτεϊνών. Συγκριτική ανάλυση της αλληλουχίας του *dnaA* και άλλων 14 γονιδίων κατέδειξε πως *η σύσταση των πρωτεϊνών επηρεάζεται δραστικά από μεταλλακτικές πιέσεις, όπως αυτές αποτυπώνονται στο GC%* (Gu et al. 1998). Επιπλέον, αυξανομένου του GC%, εμφανίζεται μια τάση υποκατάστασης των έντονα υδρόφοβων ή υδρόφιλων αμινοξικών καταλοίπων από αμφίφιλα (ambivalent) αμινοξέα, γεγονός που υποδηλώνει ότι η πλειονότητα αυτών των υποκαταστάσεων δεν είναι αποτέλεσμα

θετικής επιλογής.

Ακολουθώντας ένα διαφορετικό τρόπο ανάλυσης, οι D'Onofrio και συνεργάτες οδηγήθηκαν στο συμπέρασμα πως αυξανομένου του $GC_3$% η συχνότητα των υδρόφιλων αμινοξέων μειώνεται, ενώ αντίθετα η συχνότητα των αμφίφιλων αλλά και των υδρόφοβων αμινοξέων αυξάνεται (D'Onofrio et al. 1999). Τα αποτελέσματα διαφέρουν από εκείνα των Gu *et al.* (1998), καθώς εδώ χρησιμοποιήθηκε διαφορετική κλίμακα υδροφοβικότητας για την κατάταξη των αμινοξέων, ενώ επίσης οι συσχετίσεις έγιναν έναντι του GC% των τρίτων θέσεων των κωδικονίων ($GC_3$%) και όχι του συνολικού GC% του γονιδιώματος. Επίσης, τα συμπεράσματα κινήθηκαν προς την αντίθετη κατεύθυνση. Η αύξηση της συχνότητας εμφάνισης αμφίφιλων και υδρόφοβων αμινοξέων σε σχέση με τα υδρόφιλα οδηγεί σε αύξηση της σταθερότητας της δομής των πρωτεϊνικών μορίων και συνοδεύεται από δομικές και πιθανώς λειτουργικές αλλαγές. Έτσι, οι μεταβολές του GC% αποδίδονται στη φυσική επιλογή, με τις αλλαγές στις κωδικοποιούμενες πρωτεΐνες να είναι η κινητήρια δύναμη πίσω από την εξέλιξη της σύστασης του γονιδιώματος.

Στατιστική ανάλυση της κατανομής των αμινοξέων στο γένος Thermus των Αρχαίων και σε σειρά βακτηρίων έδειξε ότι η πόλωση της γονιδιωματικής σύστασης προς υψηλό GC% μπορεί να συνοδεύεται με μια σημαντική ανακατανομή των σχετικών συχνοτήτων μεταξύ των υδρόφοβων αμινοξέων, δίχως να αυξάνεται το συνολικό περιεχόμενο των πρωτεϊνών σε υδρόφοβα κατάλοιπα (Wilquet & Van de Casteele 1999). Έτσι, ενώ το άθροισμα των συχνοτήτων των καταλοίπων Leu, Val, Ile, Met και Phe λαμβάνει παρεμφερείς τιμές, ανεξαρτήτως του GC%, στα υψηλού GC% γονιδιώματα αυξάνεται η συχνότητα των καταλοίπων λευκίνης και βαλίνης. Στα γονιδιώματα που εξετάστηκαν, υψηλό GC% συσχετίζεται με αύξηση των συχνοτήτων των καταλοίπων Leu, Val, Ala, Pro, Arg και Gly, ανεξάρτητα από την εξελικτική προέλευση των οργανισμών ή το ενδιαίτημά τους. Ιδιαίτερη σημασία αποδίδεται στη σύσταση των πρώτων κωδικών θέσεων, καθώς αυξανομένου του GC% παρατηρείται μετατόπιση της κατανομής των αμινοξέων από κατάλοιπα που αντιστοιχίζονται σε κωδικόνια με A ή T στην $1^η$ θέση (Cys, Ser, Thr και Trp) προς κατάλοιπα που αντιστοιχίζονται σε κωδικόνια με G ή C στην $1^η$ θέση (Arg, Gly, Ala και Pro). *Τα αποτελέσματα αυτά συνηγορούν υπέρ της ουδέτερης εξέλιξης που κατευθύνεται από* A·T/G·C *πιέσεις.*

Εάν πρωταρχικά εκδηλωνόταν μια πόλωση της σύστασης των πρωτεϊνών, ως συνέπεια επιλεκτικών περιορισμών που αφορούν τη δομή και λειτουργία τους, η

οποία δευτερευόντως αντανακλώνταν στη σύσταση του DNA, αυτή αναμένεται να επιδρούσε κατά κύριο λόγο στο GC% των μη-εκφυλισμένων θέσεων των κωδικονίων. Ωστόσο, η συσχέτιση μεταξύ του αμινοξικού περιεχομένου των πρωτεϊνών και του GC% των συνώνυμων θέσεων των κωδικονίων υποδηλώνει πως η φορά της αιτιότητας κατευθύνεται από τις A·T/G·C πιέσεις προς τη σύσταση των πρωτεϊνών (Singer & Hickey 2000). *Η κατευθύνουσα μεταλλακτική πίεση στο GC% παρέχει την πιο συνεκτική ερμηνεία των παρατηρούμενων συσχετίσεων.* Εξηγεί επίσης γιατί η ένταση των πολώσεων είναι μεγαλύτερη στις εκφυλισμένες θέσεις από ότι είναι στις μη-εκφυλισμένες. Οι A·T/G·C πιέσεις δρουν στο επίπεδο ολόκληρου του γονιδιώματος, διαμορφώνοντας τη σύσταση όλων των κωδικών θέσεων και κατ' επέκταση τις διαφορές στο αμινοξικό περιεχόμενο των πρωτεϊνικών μορίων. Δευτερευόντως, η φυσική επιλογή έρχεται να εξαλείψει τις επιβλαβείς (deleterious) μεταλλάξεις, θέτοντας περιορισμούς στην εξέλιξη των $1^{ων}$ και $2^{ων}$ κωδικών θέσεων. Οι εν λόγω συσχετίσεις εντοπίζονται τόσο στα Βακτήρια όσο και στα Αρχαία. Επίσης, ισχύουν και όταν εξετάζονται ομόλογα γονίδια που κωδικοποιούν για μόρια τα οποία επιτελούν την ίδια λειτουργία, συνεπώς υπόκεινται σε αντίστοιχους επιλεκτικούς περιορισμούς. Συμπερασματικά, *οι μεταλλακτικές πολώσεις που διαμορφώνουν το συνολικό GC% του γονιδιώματος μπορεί να ασκούν σημαντική επίδραση στη μοριακή εξέλιξη των πρωτεϊνών. Τη διαδικασία αυτή διαμεσολαβεί ο γενετικός κώδικας, η δομή του οποίου εξηγεί ένα πολύ σημαντικό μέρος της δια-γονιδιωματικής ποικιλότητας* (intergenomic variation), *όπως αυτή εκφράζεται στο επίπεδο της πρωτεϊνικής σύστασης.*

### **1.9.4** *Η επίδραση ασύμμετρων μεταλλακτικών πιέσεων στο αμινοξικό περιεχόμενο των πρωτεϊνών*

Οι προαναφερθείσες μελέτες καταδεικνύουν ότι οι συγκεκριμένες ανά είδος (species-specific) μεταλλακτικές πολώσεις καθορίζουν σε μεγάλο βαθμό τη χρήση τόσο των κωδικονίων όσο και των αμινοξέων, παράγοντας την παρατηρούμενη δια-ειδική ποικιλότητα. Οι Lafay *et al.* (1999) ήταν οι πρώτοι που έδειξαν πως οι ειδικές ανά κλώνο πολώσεις των μεταλλακτικών πιέσεων που ασκούνται σε ένα δοσμένο γονιδίωμα επάγουν διακριτά πρότυπα σύστασης στο

αντίστοιχο πρωτέωμα. Συγκρίνοντας το γονιδίωμα δύο βακτηρίων που ανήκουν στις Σπειροχαίτες (*Borrelia burgdorferi*: 28.6 GC%, *Treponema pallidum*: 52.8 GC%) διαπίστωσαν πως η χρήση κωδικονίων και η συχνότητα εμφάνισης των κωδικοποιούμενων αμινοξέων επηρεάζεται δραστικά από τον κλώνο, οδηγό ή συνοδό, στον οποίο εντοπίζονται τα αντίστοιχα γονίδια. Παρά τη μεγάλη διαφορά του GC%, που τροποποιεί τη σύσταση του γονιδιώματος και του αντίστοιχου πρωτεώματος, οι ετερογένεια στη χρήση κωδικονίων μεταξύ γονιδίων που η κωδική τους αλυσίδα βρίσκεται σε διαφορετικό κλώνο, οδηγό ή συνοδό, ακολουθεί παρεμφερή πρότυπα στα δύο αυτά είδη. Τα πρότυπα αυτά αντανακλώνται και στη σύσταση των κωδικοποιούμενων πρωτεϊνών. Επιπλέον, συγκρίνοντας ομόλογα γονίδια που έχουν υποστεί αντιστροφή της φοράς τους, συμπέραναν πως η σύστασή τους εξομοιώνεται με τα γενικότερα χαρακτηριστικά του κλώνου στον οποίο μετατέθηκε η κωδική τους αλυσίδα.

### **1.9.5** *Το περιεχόμενο σε GC και Pu διαμορφώνει την χρήση κωδικονίων και αμινοξέων*

Μεταγενέστερες μελέτες παρουσίασαν μοντέλα πρόβλεψης της χρήσης κωδικονίων και αμινοξέων ως συνάρτηση του γονιδιωματικού GC%. Έτσι, σε δεδομένο GC% αντιστοιχίζεται μονοσήμαντα ένας συγκεκριμένος συνδυασμός χρήσης κωδικονίων ή αμινοξέων. Ένα τέτοιο μοντέλο, βασισμένο αποκλειστικά σε μεταλλακτικές πιέσεις και την αρνητική επιλογή (purifying selection), κατορθώνει να ερμηνεύσει περίπου το 80% της παρατηρούμενη δια-ειδικής διασποράς των συχνοτήτων των κωδικονίων και των κωδικοποιούμενων αμινοξέων (Knight et al. 2001). Αυτό συνηγορεί υπέρ της άποψης πως οι A·T/G·C πιέσεις κατευθύνουν το GC% και τις συχνότητες των κωδικονίων. *Εάν το GC% ήταν το παρεπόμενο αποτέλεσμα των προτύπων χρήσης κωδικονίων, πολλοί διαφορετικοί συνδυασμοί των συχνοτήτων τους θα μπορούσαν να οδηγήσουν στο ίδιο GC%. Αντίθετα, οι* A·T/G·C *πιέσεις προς δεδομένο GC% μπορούν να επηρεάζουν κατά μονοσήμαντο τρόπο τις νουκλεοτιδικές υποκαταστάσεις στις κωδικές θέσεις και να αντιστοιχίζουν το GC% με συγκεκριμένες συχνότητες κωδικονίων και αμινοξέων.* Τα παραπάνω υποδηλώνουν ότι αμινοξικές υποκαταστάσεις μπορεί να είναι ανεκτές στις περισσότερες θέσεις των πρωτεϊνικών μορίων.

Οι συσχετίσεις μεταξύ κωδικονίων και GC% οδηγούν σε χαρακτηριστικά πρότυπα, με τη συχνότητα των κωδικονίων που λήγουν σε G/C να αυξάνεται και εκείνων που λήγουν σε A/T να μειώνεται, αυξανομένου του GC%. *Η διαφορική απόκριση των κωδικονίων στο GC% αποδίδεται σε μεγάλο βαθμό στα πρότυπα σημειακών μεταλλάξεων που μπορούν να πραγματοποιηθούν εντός κάθε ομάδας συνώνυμων κωδικονίων* (Palidwor et al. 2010). Η συχνότητα πολλών κωδικονίων μεταβάλλεται γραμμικά συναρτήσει του GC%. Ωστόσο, τα κωδικόνια της Ισολευκίνης ανταποκρίνονται στο GC% μονότονα, αλλά όχι γραμμικά, ενώ ορισμένα κωδικόνια της Αργινίνης και της Λευκίνης μη-μονότονα. Η Ile είναι το μόνο αμινοξύ που αντιστοιχίζεται σε μονό αριθμό κωδικονίων, και ως εκ τούτου έχει άνισο αριθμό κωδικονίων που λήγουν σε G/C και A/T. Η Arg και η Leu είναι τα μόνα αμινοξέα που επιτρέπουν συνώνυμες υποκαταστάσεις όχι μόνο στην 3$^{η}$, αλλά και στην 1$^{η}$ θέση των κωδικονίων τους. Οι ιδιαιτερότητες των κωδικονίων της Ile, Arg και της Leu οφείλονται στη δομή των συνώνυμων ομάδων που κωδικοποιούν για αυτά τα αμινοξέα (Lafay et al. 1999, Knight et al. 2001, Palidwor et al. 2010).

Εκτός από το GC% του γονιδιώματος, το περιεχόμενο σε πουρίνες ασκεί σημαντική επίδραση στη διαμόρφωση της χρήσης κωδικονίων και αμινοξέων. *Λαμβάνοντας υπόψη τόσο το GC% όσο και το περιεχόμενο σε πουρίνες* (Pu%), *μπορούν να ανασυγκροτηθούν οι εμπειρικές σχέσεις μεταξύ της νουκλεοτιδικής σύστασης του γονιδιώματος και των συχνοτήτων των κωδικονίων*, τόσο στα Βακτήρια και τα Αρχαία, όσο και στους Ευκαρυωτικούς (Zhang & Yu 2010). *Η δομή του γενετικού κώδικα εμφανίζεται ξανά ως ένας καθοριστικός παράγοντας για την εξισορρόπηση των μεταλλακτικών και επιλεκτικών πιέσεων που ασκούνται στις κωδικές περιοχές*. Έτσι, η διάκριση των αμινοξέων σε ευαίσθητα και αδρανή (μη-ευαίσθητα) σε αλλαγές του Pu%, ανάλογα με το εάν τα κωδικόνιά τους επιτρέπουν υποκαταστάσεις του τύπου Pu↔Py στις 3$^{ες}$ θέσεις τους, καθορίζει σε μεγάλο βαθμό την ίδια τη σύσταση των πρωτεϊνικών μορίων. Επιπλέον, το Pu% των 2$^{ων}$ κωδικών θέσεων συνδέεται με τις φυσικοχημικές ιδιότητες των αντίστοιχων αμινοξέων, και συγκεκριμένα με το φορτίο και την υδροφοβικότητά τους. Οι σύσταση των κωδικονίων είναι τέτοια ώστε πολλές μη-συνώνυμες σημειακές μεταλλάξεις να οδηγούν σε υποκαταστάσεις από αμινοξέα με παρεμφερείς φυσικοχημικές ιδιότητες (Yu 2007).

# 2. ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

## **2.1** *Συλλογή βακτηριακών γονιδιωμάτων*

Η παρούσα μελέτη πραγματοποιήθηκε στα βακτηριακά γονιδιώματα της συλλογής που χρησιμοποίησαν οι Necşulea και Lobry (2007) για να μελετήσουν τις ειδικές ανά κλώνο ασυμμετρίες στο DNA των προκαρυωτικών. Για κάθε μία από αυτές τις αλληλουχίες ανακτήσαμε την πιο πρόσφατη από τις εκδόσεις της που είχαν κατατεθεί στις βάσεις δεδομένων του NCBI έως τις 20 Απριλίου 2015. Για κάθε χρωμόσωμα, οι συντεταγμένες του σημείου έναρξης της αντιγραφής (*ori*) ελήφθησαν από την βάση δεδομένων DoriC 5.0 (Gao et al. 2013). Από την μελέτη μας αποκλείσαμε εκείνα τα γραμμικά χρωμοσώματα των οποίων το σημείο έναρξης της αντιγραφής απέχει από τα άκρα τους λιγότερο από το ένα πέμπτο του συνολικού μήκους τους. Έτσι, εξασφαλίζουμε ότι τα τμήματα κάθε χρωμοσώματος που βρίσκονται εκατέρωθεν του *ori* είναι αρκούντως μεγάλα ώστε να μπορεί να εφαρμοστεί η ανάλυση των αποκλίσεων μέσω μοντέλων γραμμικής παλινδρόμησης (βλ. ενότητα 2.6). Από το αρχικό σύνολο, 340 αλληλουχίες DNA ανταποκρίνονται στο ανωτέρω κριτήριο και συγκροτούν την γονιδιωματική μας συλλογή.

Σε όλα τα κυκλικά χρωμοσώματα θέτουμε ως σημείο λήξης της αντιγραφής (*ter*) την θέση που απέχει από το *ori* απόσταση ίση με το μισό του συνολικού τους μήκους, κατά το πρότυπο προγενέστερων εργασιών (Mao et al. 2012, Saha et al. 2014). Οι συντεταγμένες αυτών των χρωμοσωμάτων μετατοπίζονται κατά τρόπο ώστε το *ori* να βρίσκεται στο μέσον της αλληλουχίας του δημοσιευμένου κλώνου. Έτσι, η περιοχή του *ter* τοποθετείται στις άκρες (στο τέλος ή στην αρχή) του δημοσιευμένου κλώνου. Ακολούθως, χρησιμοποιούμε αυτό το μετατοπισμένο σύστημα συντεταγμένων σε όλους τους χειρισμούς που αφορούν τα κυκλικά χρωμοσώματα.

Στα γραμμικά χρωμοσώματα, όπου οι γονιδιωματικές συντεταγμένες τους παραμένουν ως έχουν, η αντιγραφή τερματίζεται στις δύο άκρες τους και συνεπώς τα σημεία λήξης της αντιγραφής βρίσκονται εξ ορισμού στην αρχή και το τέλος της αλληλουχίας του δημοσιευμένου κλώνου. Συνεπώς, είτε πρόκειται για κυκλικό είτε για γραμμικό χρωμόσωμα, το τμήμα από την αρχή του

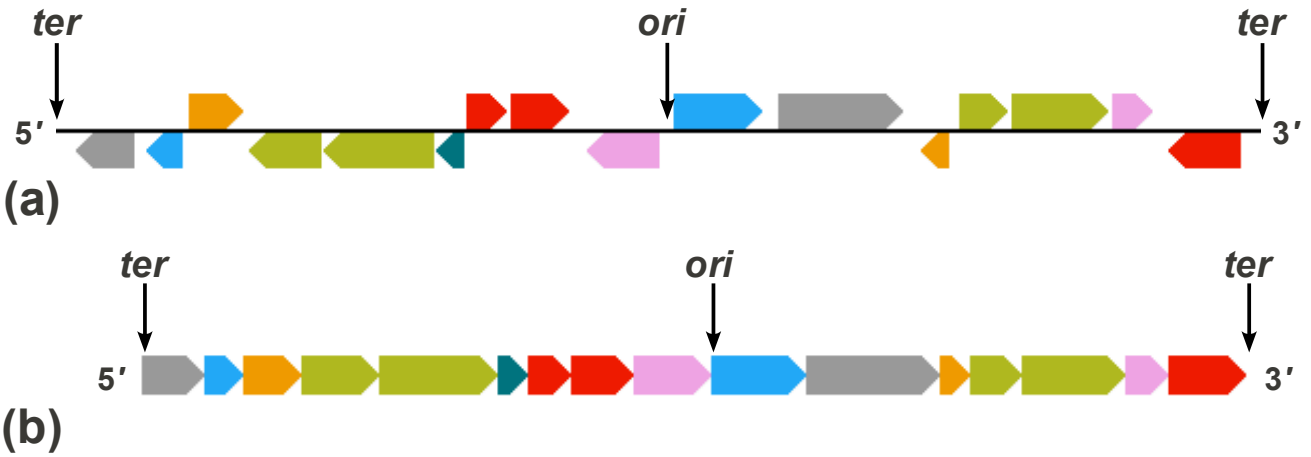
δημοσιευμένου κλώνου έως το *ori* αντιστοιχεί στο συνοδό κλώνο, ενώ το τμήμα από το *ori* έως το τέλος του δημοσιευμένου κλώνο αντιστοιχεί στον οδηγό κλώνο.

## **2.2** *CDS-συρραφές*

Από κάθε αλληλουχία DNA της συλλογής μας λαμβάνουμε τους κωδικούς κλώνους των γονιδίων που κωδικοποιούν για πολυπεπτιδικές αλυσίδες. Συνενώνουμε τους κλώνους αυτούς διαδοχικά, κατά την νοηματική τους κατεύθυνση (sense direction), κατασκευάζοντας μία τεχνητή αλληλουχία, την οποία καλούμε "CDS-συρραφή". Η διάταξη των κωδικών κλώνων στις CDS-συρραφές διατηρεί αμετάβλητη την σχετική τους θέση ως προς το σύστημα συντεταγμένων της αλληλουχίας DNA από την οποία ελήφθησαν (βλ. Εικόνα 3).

Η περιοχή του *ori* κατά κανόνα δεν επικαλύπτεται με γονίδια που κωδικοποιούν για πολυπεπτιδικές αλυσίδες και τότε δεν αντιπροσωπεύεται στις CDS-συρραφές. Εντοπίζουμε την περιοχή των CDS-συρραφών που βρίσκεται πλησιέστερα στο *ori*. Εφεξής, όπου γίνεται λόγος για το *ori* στις CDS-συρραφές, αναφερόμαστε σε αυτή την περιοχή.

Οι τιμές των αποκλίσεων που μετρώνται στις CDS-συρραφές αποτελούν τις CDS-συζευγμένες αποκλίσεις.




**Εικόνα 3.** Κατασκευή των CDS-συρραφών. Τα χρωματισμένα βέλη αντιπροσωπεύουν γονίδια που κωδικοποιούν για πολυπεπτιδικές αλυσίδες. Η φορά κάθε βέλους δηλώνει την νοηματική κατεύθυνση (sense direction) του κωδικού κλώνου του αντίστοιχου

γονιδίου. **(a)** Σχηματική αναπαράσταση του δημοσιευμένου κλώνου ενός χρωμοσώματος. Η οριζόντια μαύρη γραμμή αντιπροσωπεύει το σύνολο της αλληλουχίας του χρωμοσώματος. Στις περιπτώσεις που η νοηματική φορά των κωδικών κλώνων συμπίπτει με την 5′→ 3′ φορά του δημοσιευμένου κλώνου, τα αντίστοιχα γονίδια απεικονίζονται πάνω από την οριζόντια γραμμή. Στην αντίθετη περίπτωση, τα γονίδια απεικονίζονται κάτω από την οριζόντια γραμμή. **(b)** Οι κωδικοί κλώνοι των γονιδίων συνενώνονται διαδοχικά, κατά τη νοηματική τους κατεύθυνση. Η σχετική θέση που έχουν τα γονίδια κατά μήκος του χρωμοσώματος διατηρείται αμετάβλητη στη CDS-συρραφή. Η περιοχή του *ori* δεν επικαλύπτεται με τα εικονιζόμενα γονίδια. Όταν γίνεται λόγος για το *ori* στη CDS-συρραφή, αναφερόμαστε στη θέση της συρραφής που βρίσκεται πλησιέστερα στην αντίστοιχη χρωμοσωμική περιοχή.

## **2.3** *Ορισμός των αποκλίσεων*

Οι μονονουκλεοτιδικές αποκλίσεις υπολογίζονται σύμφωνα με τους λόγους:

$$S^{A-T}=\frac{f^A-f^T}{f^A+f^T} \quad \text{και} \quad S^{G-C}=\frac{f^G-f^C}{f^G+f^C}$$

όπου $f^x$ είναι η παρατηρούμενη συχνότητα του X ∈ (A,T,G,C) σε μία δεδομένη αλληλουχία DNA.

Για κάθε ζεύγος αντιστρόφως συμπληρωματικών δινουκλεοτιδίων οι αποκλίσεις των παρατηρούμενων συχνοτήτων τους υπολογίζονται, όπως και στην μελέτη των Shioiri και Takahata (2001), σύμφωνα με τον λόγο:

$$S^{XY-Y'X'}=\frac{f^{XY}-f^{Y'X'}}{f^{XY}+f^{Y'X'}}$$

όπου Y'X' είναι το αντιστρόφως συμπληρωματικό δινουκλεοτίδιο του XY και $f^{XY}, f^{Y'X'}$ είναι οι παρατηρούμενες συχνότητες των αντίστοιχων δινουκλεοτιδίων.

Με την παρούσα μελέτη εισάγουμε *το μέτρο των αποκλίσεων των σταθμισμένων συχνοτήτων* των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων, το οποίο ορίζεται ως:

$$P^{XY-Y'X'}=\rho^{XY}-\rho^{Y'X'}$$

όπου Y'X' είναι το αντιστρόφως συμπληρωματικό δινουκλεοτίδιο του XY και $\rho^{XY}, \rho^{Y'X'}$ είναι οι σταθμισμένες συχνότητες των αντίστοιχων δινουκλεοτιδίων. Η

σταθμισμένη συχνότητα ενός δινουκλεοτιδίου ορίζεται ως ο λόγος της παρατηρούμενης προς την αναμενόμενη συχνότητά του. Η αναμενόμενη συχνότητα ενός δινουκλεοτιδίου ισούται με το γινόμενο των συχνοτήτων των μονονουκλεοτιδίων που το απαρτίζουν. Συνεπώς, $\rho^{XY} = f^{XY}/f^{X}f^{Y}$.

Ας σημειωθεί ότι 4 από τα συνολικά 16 διαφορετικά δινουκλεοτίδια, τα AT, TA, GC και CG, είναι ταυτόσημα με τα αντιστρόφως συμπληρωματικά τους. Τα υπόλοιπα 12 σχηματίζουν 6 ζεύγη αντιστρόφως συμπληρωματικών δινουκλεοτιδίων, ήτοι τα AG-CT, GA-TC, GG-CC, AA-TT, AC-GT και CA-TG. Οι δινουκλεοτιδικές αποκλίσεις, τόσο των παρατηρούμενων όσο και των σταθμισμένων συχνοτήτων, υπολογίζονται για αυτά τα 6 ζεύγη.

## **2.4** *Γονιδιωματικές υπογραφές*

Ο τύπος $\rho^{XY} = f^{XY}/f^{X}f^{Y}$, όπου X,Y ∈ (A,T,G,C), επιτρέπει κατ' αρχήν τον υπολογισμό των σταθμισμένων δινουκλεοτιδικών συχνοτήτων σε κάθε έναν από τους δύο κλώνους του DNA ξεχωριστά. Οι μελέτες των Karlin, Mocarski, *et al.* (1994), Karlin, Ladunga, *et al.* (1994), Karlin και Cardon (1994) και Karlin και Ladunga (1994) εισήγαγαν την έννοια των σταθμισμένων δινουκλεοτιδικών συχνοτήτων οι οποίες είναι συμμετρικές ως προς τους δύο κλώνους DNA και αναφέρονται με τον συμβολισμό $\rho^*$. Σύμφωνα με την σχετική μεθοδολογία, κάθε αλληλουχία DNA συνενώνεται με την αντιστρόφως συμπληρωματική της. Ακολούθως, για κάθε δινουκλεοτίδιο, έστω XY, τίθεται $\rho^*_{XY}=f^*_{XY}/f^*_{X}f^*_{Y}$, όπου $f^*_{XY}$, $f^*_{X}$ και $f^*_{Y}$ οι συχνότητες του δινουκλεοτιδίου XY και των μονονουκλεοτιδίων X,Y, αντίστοιχα, όπως αυτές υπολογίζονται κατά μήκος της αλληλουχίας που προέκυψε από τη συνένωση των δύο αντιστρόφως συμπληρωματικών κλώνων DNA. Έτσι, αντί των νουκλεοτιδικών συχνοτήτων $f_A$, $f_T$, $f_G$ και $f_C$ της αρχικής αλληλουχίας, λαμβάνονται οι συμμετρικές συχνότητες, $f_A^*= f_T^* = (f_A + f_T)/2$ και $f_G^*= f_C^* = (f_G + f_C)/2$. Παρομοίως, $f_{GT}^*= f_{AC}^* = (f_{GT} + f_{AC})/2$ και ούτω καθ' εξής, για κάθε ζεύγος αντιστρόφως συμπληρωματικών δινουκλεοτιδίων. Συνεπώς, οι *συμμετρικές* δινουκλεοτιδικές σταθμισμένες συχνότητες ($\rho^*$) είναι εκ κατασκευής ίσες μεταξύ των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων. Έτσι, $\rho^*_{CC}=\rho^*_{GG}$, $\rho^*_{TT}=\rho^*_{AA}$, $\rho^*_{TG}=\rho^*_{CA}$, $\rho^*_{AG}=\rho^*_{CT}$, $\rho^*_{AC}=\rho^*_{GT}$, $\rho^*_{GA}=\rho^*_{TC}$.

Το διάνυσμα των τιμών των $\rho^*$ είναι χαρακτηριστικό της κάθε αλληλουχίας και συγκροτεί την *γονιδιωματική υπογραφή* της (Karlin & Mrázek 1997, Karlin 1998, Campbell et al. 1999). Η ετερογένεια δύο αλληλουχιών, *g* και *h*, ποσοτικοποιείται από την απόσταση Manhattan (rectilinear ή Manhattan distance) μεταξύ των αντίστοιχων διανυσμάτων, δηλαδή:

$$\delta(g,h) = (1/10)\Sigma|\rho^*_{XY}(f) - \rho^*_{XY}(g)|$$

όπου XY ∈ (AT, TA, GC, CG, CC ή GG, TT ή AA, TG ή CA, AG ή CT, AC ή GT, GA ή TC).

Στις αναλύσεις που ακολουθούν, οι γονιδιωματικές υπογραφές υπολογίζονται λαμβάνοντας υπόψιν ολόκληρη την αλληλουχία καθενός από τα χρωμοσώματα της συλλογής μας.

## **2.5** *Γραφική αναπαράσταση των αποκλίσεων*

Για κάθε χρωμόσωμα της συλλογής μας, υπολογίζουμε τις αποκλίσεις κατά μήκος του δημοσιευμένου κλώνου και των CDS-συρραφών. Οι μετρήσεις γίνονται εντός διαδοχικών, μη-επικαλυπτόμενων και σταθερού μήκους τμημάτων, τα οποία καλούνται *κυλιόμενα παράθυρα*. Το μήκος των κυλιόμενων παραθύρων ορίζεται ίσο με $10^4$ bps. Για κάθε απόκλιση, σχεδιάζουμε τα αθροιστικά και τα απλά (μη-αθροιστικά) διαγράμματα των αντίστοιχων μετρήσεων κατά μήκος του οδηγού κλώνου. Επίσης, σχεδιάζουμε τα απλά διαγράμματα των αποκλίσεων κατά μήκος των CDS-συρραφών. Τα διαγράμματα αυτά απεικονίζουν τα πρότυπα των αποκλίσεων κατά μήκος των υπό εξέταση αλληλουχιών.

## **2.6** *Εντοπισμός σημείων μεταβολής (breakpoints) στα πρότυπα των αποκλίσεων*

Σε πολλά βακτηριακά χρωμοσώματα οι μονονουκλεοτιδικές αποκλίσεις είναι κατά προσέγγιση σταθερές κατά μήκος του οδηγού και του συνοδού κλώνου, ενώ το πρόσημό τους αλλάζει κοντά στην περιοχή του *ori* και του *ter*. Είναι λοιπόν

εύλογο να υποθέσουμε ότι οι μονονουκλεοτιδικές αποκλίσεις μπορούν να περιγραφούν από ένα μοντέλο γραμμικής παλινδρόμησης σε κάθε έναν από τους δύο κλώνους της αντιγραφής. Στο πλαίσιο αυτό, όταν εξετάζουμε τον δημοσιευμένο κλώνο, το *ori* (ή το *ter*) αντιπροσωπεύει ένα σημείο μεταβολής (breakpoint), όπου οι συντελεστές των μοντέλων γραμμικής παλινδρόμησης μεταβαίνουν από μία σταθερή, γραμμική σχέση μεταξύ των αποκλίσεων και των γονιδιωματικών συντεταγμένων, σε μία άλλη.

Μελετάμε τις τιμές των μονονουκλεοτιδικών αποκλίσεων βάσει των οποίων κατασκευάσαμε τα απλά (μη-αθροιστικά) διαγράμματα κατά μήκος του δημοσιευμένου κλώνου. Προκειμένου να ελέγξουμε εάν οι αλλαγές στην δομή των προτύπων των αποκλίσεων εκατέρωθεν του *ori* είναι στατιστικά σημαντικές, εάν δηλαδή το *ori* αποτελεί στατιστικά σημαντικό σημείο μεταβολής, χρησιμοποιούμε τον αλγόριθμο δυναμικού προγραμματισμού που ανέπτυξαν οι Zeileis *et al*. (2003) και Zeileis *et al*. (2010), έτσι όπως αυτός υλοποιήθηκε στο πακέτο κώδικα strucchange της γλώσσας προγραμματισμού R (Zeileis et al. 2006). Ο αλγόριθμος αυτός επιτρέπει τον ταυτόχρονο υπολογισμό πολλαπλών σημείων μεταβολής κατά μήκος μιας δεδομένης σειράς τιμών που λαμβάνει ένα μέγεθος, εν προκειμένω οι αποκλίσεις. Συγκεκριμένα, υπολογίζεται ο αριθμός και η θέση των βέλτιστων σημείων μεταβολής από τα δεδομένα και μόνο, χωρίς να απαιτείται πρότερη (*a priori*) γνώση. Τα σημεία αυτά συνάγονται βάσει των χαρακτηριστικών ευστάθειας που εμφανίζουν τα μοντέλα γραμμικής παλινδρόμησης τα οποία δοκιμάζει ο αλγόριθμος.

Ακολούθως, εξετάζουμε εάν τα πρότυπα των δινουκλεοτιδικών αποκλίσεων, σε όρους τόσο παρατηρούμενων όσο και σταθμισμένων συχνοτήτων, εμφανίζουν αντίστοιχες δομικές αλλαγές εκατέρωθεν του *ori*. Για το σκοπό αυτό εφαρμόζουμε τον ανωτέρω αλγόριθμο στις τιμές των αντίστοιχων αποκλίσεων βάσει των οποίων κατασκευάσαμε τα απλά (μη-αθροιστικά) διαγράμματα κατά μήκος του δημοσιευμένου κλώνου.

Τέλος, επαναλαμβάνουμε τους ίδιους χειρισμούς για όλες τις αποκλίσεις, όπως αυτές υπολογίστηκαν για την κατασκευή των αντίστοιχων διαγραμμάτων κατά μήκος των CDS-συρραφών.

## **2.7** *Συσχέτιση των αποκλίσεων με την φυλογένεση των βακτηρίων*

Ομαδοποιούμε τις αποκλίσεις κατά μήκος των CDS-συρραφών σε τρεις κλάσεις: (*α*) *μονονουκλεοτιδικές αποκλίσεις*: $V^{MONO}=(S_{CDS}{}^{A-T}, S_{CDS}{}^{G-C})$, (*β*) *αποκλίσεις παρατηρούμενων δινουκλεοτιδικών συχνοτήτων*: $V^{DI}=(S_{CDS}{}^{XY-Y'X'})$, και (*γ*) *αποκλίσεις σταθμισμένων δινουκλεοτιδικών συχνοτήτων*: $V^{RA}=(P_{CDS}{}^{XY-Y'X'})$, όπου XY-Y'X' ∈ (AG-CT, GA-TC, GG-CC, AA-TT, AC-GT, CA-TG). Υπολογίζουμε τις $V^{MONO}$, $V^{DI}$ και $V^{RA}$ κατά μήκος των CDS-συρραφών σε διαδοχικά, μη-επικαλυπτόμενα *παράθυρα* μήκους $10^4$ bps. Κάθε μία από αυτές τις κλάσεις μπορεί να θεωρηθεί ως μία πολυδιάστατη τυχαία μεταβλητή, που αντιστοιχεί σε μία λίστα από διανύσματα τιμών των αποκλίσεων. Κάθε ένα από αυτά τα διανύσματα περιλαμβάνει τις μετρήσεις που πραγματοποιήθηκαν εντός ενός δεδομένου παραθύρου. Προκειμένου να εκτιμήσουμε πόσο όμοια είναι τα πρότυπα αποκλίσεων δύο δεδομένων CDS-συρραφών, συγκρίνουμε ανά ζεύγη τις αντίστοιχες κατανομές που ακολουθεί κάθε μία από τις $V^{MONO}$, $V^{DI}$ και $V^{RA}$, χρησιμοποιώντας την συμμετρική απόκλιση Kullback-Leibler (KL) (Kullback & Leibler 1951). Έτσι, η συμμετρική KL-απόκλιση ποσοτικοποιεί την ανομοιότητα ανάμεσα σε όλα τα ζεύγη των βακτηρίων που αντιπροσωπεύονται στην συλλογή μας, σε όρους αποκλίσεων κατά μήκος των αντίστοιχων CDS-συρραφών.

Ομαδοποιούμε τα βακτήρια που αντιπροσωπεύονται στην συλλογή μας σύμφωνα με το φύλο στο οποίο ανήκουν. Στην περίπτωση των Πρωτεοβακτηρίων, που είναι με διαφορά το μεγαλύτερο φύλο στην συλλογή μας, η ομαδοποίηση γίνεται βάσει των αντίστοιχων κλάσεων. Για κάθε μία από αυτές τις ομάδες (φύλα ή κλάσεις) κατασκευάζουμε το αντίστοιχο κλαδόγραμμα, μέσω ιεραρχικής συσταδοποίησης πλήρους σύνδεσης (complete-linkage hierarchical clustering) των CDS-συρραφών, χρησιμοποιώντας τις τιμές της συμμετρικής KL-απόκλισης των $V^{MONO}$ ή $V^{DI}$ ή $V^{RA}$. Τα κλαδογράμματα αυτά τα ονομάζουμε εφεξής *κλαδογράμματα ασυμμετριών*. Για τους σκοπούς της παρούσας ανάλυσης, εάν ένα βακτηριακό γονιδίωμα αποτελείται από περισσότερα του ενός χρωμοσώματα, κρατάμε εκείνο που έχει το μεγαλύτερο μήκος και αποκλείουμε όλα τα υπόλοιπα. Έτσι, αποκλείουμε από την συλλογή μας συνολικά 29 χρωμοσώματα. Επίσης, για κάθε φύλο ή κλάση κατασκευάζουμε κλαδογράμματα βάσει των αντίστοιχων γονιδιωματικών υπογραφών (Karlin & Burge 1995, Campbell et al. 1999). Συγκεκριμένα, υπολογίζουμε τις δ-αποστάσεις μεταξύ των γονιδιωματικών υπογραφών και, βάσει αυτών, πραγματοποιούμε ιεραρχική

συσταδοποίηση πλήρους σύνδεσης των μελών της κάθε ταξινομικής ομάδας.

Ακολούθως, λαμβάνουμε από την βάση δεδομένων NCBI Taxonomy (Federhen 2012) τα κλαδογράμματα που αναπαριστούν τις εξελικτικές σχέσεις των βακτηριών που αντιπροσωπεύονται στην συλλογή μας, για κάθε φύλο ή κλάση που μελετάμε. Τα κλαδογράμματα αυτά καλούνται *ταξινομικά δέντρα*. Η βάση NCBI Taxonomy παρέχει μία φυλογενετική ταξινόμηση των οργανισμών, βασισμένη σε αλληλουχίες του γονιδιώματος και των πρωτεϊνών. Η ταξινόμηση αυτή τροποποιείται κατάλληλα από τους επιμελητές της βάσης ώστε να συμφωνεί με την τρέχουσα βιβλιογραφία.

Για κάθε φύλο ή κλάση, συγκρίνουμε την τοπολογία των κλαδογραμμάτων ασυμμετριών και των αντίστοιχων ταξινομικών δέντρων, χρησιμοποιώντας το πρόγραμμα Compare2Trees (Ney et al. 2006). Το πρόγραμμα αυτό υπολογίζει την τοπολογική βαθμολογία των υπό σύγκριση κλαδογραμμάτων, η οποία εκφράζει την επί τοις εκατό τοπολογική ομοιότητά τους και εν προκειμένω υποδηλώνει τον βαθμό στον οποίο οι $V^{MONO}$, $V^{DI}$ ή $V^{RA}$ παρακολουθούν τις φυλογενετικές σχέσεις των βακτηρίων. Επίσης, με το ίδιο πρόγραμμα συγκρίνουμε τα κλαδογράμματα των γονιδιωματικών υπογραφών με τα ταξινομικά δέντρα. Καθώς οι γονιδιωματικές υπογραφές θεωρούνται ότι είναι ανά είδος καθορισμένες (species-specific), η τοπολογική βαθμολογία που προκύπτει από αυτές τις συγκρίσεις λαμβάνεται ως μέτρο αναφοράς, ενδεικτικό της απόδοσης των $V^{MONO}$, $V^{DI}$ και $V^{RA}$ στην ανακατασκευή της φυλογένεσης των βακτηρίων.

Προκειμένου να ανιχνεύσουμε πιθανές συσχετίσεις των αποκλίσεων με την φυλογένεση των βακτηρίων, χρησιμοποιούμε τις πολυδιάστατες μεταβλητές $V^{MONO}$, $V^{DI}$ και $V^{RA}$, αντί της κάθε απόκλισης ξεχωριστά. Οι $V^{MONO}$, $V^{DI}$ και $V^{RA}$ είναι καταλληλότερες για την ανάλυσή μας, καθώς εάν επί παραδείγματι οι ρυθμοί των A→G μεταβάσεων διαφέρουν μεταξύ των ειδών, αυτό θα επηρεάζει ταυτόχρονα τόσο τις $S_{CDS}^{A-T}$ όσο και τις $S_{CDS}^{G-C}$. Επιπλέον, στο συγκεκριμένο παράδειγμα, εάν οι ρυθμοί των A→G μεταλλάξεων εξαρτώνται από τις γειτονικές τους βάσεις, τότε οι διαφορές τους ανάμεσα στα είδη θα επηρεάζουν και τις δινουκλεοτιδικές αποκλίσεις, καθώς και τα 6 ζεύγη των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων έχουν τουλάχιστον ένα δινουκλεοτίδιο το οποίο περιέχει κατάλοιπα A ή G.

## **2.8** *Εξελικτικές σχέσεις των κωδικών περιοχών*

Από την βάση δεδομένων EggNOG v4.0 (Powell et al. 2014) λαμβάνουμε τις ομάδες ορθόλογων γονιδίων, όπως αυτές συνάγονται μέσω τεχνικών μη-εποπτευόμενης μάθησης (nonsupervised orthologous groups, NOGs). Λαμβάνουμε υπόψιν τα γονίδια που κωδικοποιούν για πολυπεπτιδικές αλυσίδες και σχηματίζουν ομάδες ορθολόγων στο ταξινομικό επίπεδο (φύλο ή κλάση) για το οποίο κατασκευάσαμε τα αντίστοιχα κλαδογράμματα (βλ. ενότητα 2.7). Για κάθε φύλο ή κλάση υπολογίζουμε τον αριθμό των χρωμοσωμάτων στα οποία εμφανίζονται τα μέλη μιας δεδομένης ομάδας ορθολόγων. Ο αριθμός αυτός αποτελεί την συχνότητα εμφάνισης κάθε ορθόλογου γονιδίου σε ένα δεδομένο φύλο ή κλάση. Από κάθε χρωμόσωμα λαμβάνουμε τους κωδικούς κλώνους των γονιδίων που ανήκουν σε ομάδες ορθολόγων και τους κατατάσσουμε σε δύο υποσύνολα: (*α*) στα *χαμηλής κάλυψης* ορθόλογα, με συχνότητα εμφάνισης μικρότερη ή ίση του 10%, και (*β*) στα *υπόλοιπα* ορθόλογα, με συχνότητα εμφάνισης μεγαλύτερη του 10%. Τα δύο αυτά υποσύνολα είναι συμπληρωματικά μεταξύ τους και η ένωσή τους αντιστοιχεί στο σύνολο των ορθολόγων που απαντώνται στο εκάστοτε χρωμόσωμα. Στις περιπτώσεις εκείνες όπου μία κλάση ή ένα φύλο αντιπροσωπεύονται από μικρό αριθμό ειδών (λιγότερα από είκοσι) στην συλλογή μας, στα *χαμηλής κάλυψης* ορθόλογα κατατάσσουμε τα γονίδια εκείνα που ανήκουν σε ομάδες ορθολόγων και απαντώνται σε ένα μόνο από τα υπό εξέταση χρωμοσώματα.

Για τα είδη των οποίων το γονιδίωμα αποτελείται από περισσότερα του ενός χρωμοσώματα, κρατάμε το μεγαλύτερο από αυτά και αποκλείουμε από την ανάλυσή μας όλα τα υπόλοιπα, κατ' αναλογία με τους χειρισμούς για την κατασκευή των κλαδογραμμάτων. Επίσης αποκλείουμε από την ανάλυσή μας τα χρωμοσώματα των ειδών που δεν αντιπροσωπεύονται στην βάση δεδομένων EggNOG v4.0. Κατόπιν τούτων, συνολικά λαμβάνουμε υπόψιν 266 αλληλουχίες DNA.

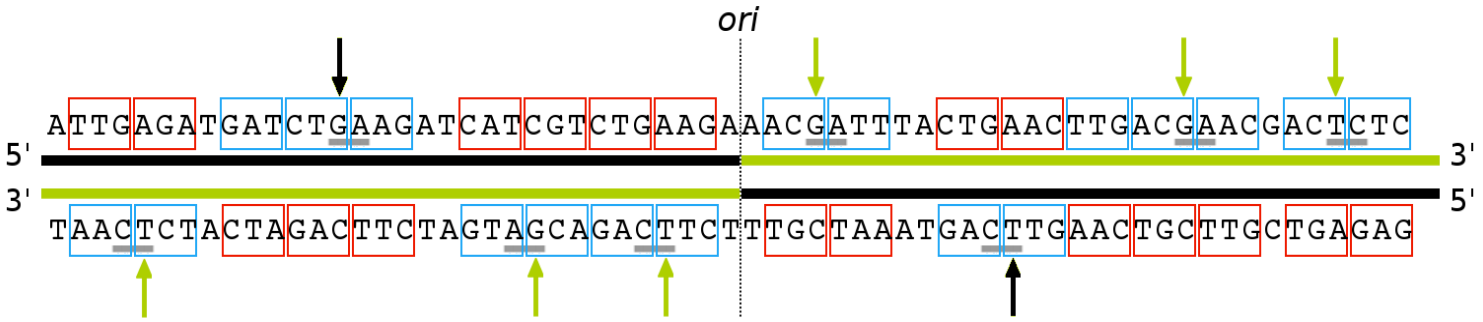
## **2.9** *Αποκλίσεις συζευγμένες με τη μεταγραφή ή την αντιγραφή*

Οι ασυμμετρίες στην σύσταση του DNA προέρχονται από ειδικές ανά κλώνο πιέσεις, που μπορεί να είναι είτε επιλεκτικές είτε μεταλλακτικές. Προκειμένου να ανιχνεύσουμε την συνεισφορά των μεταλλακτικών πιέσεων στις ασυμμετρίες αυτές, μελετάμε τις αποκλίσεις στις τρίτες τετραπλά εκφυλισμένες ($3^{ες}|4$) θέσεις, η σύσταση των οποίων δεν υπόκειται σε επιλεκτικούς περιορισμούς σχετιζόμενους με την ταυτότητα του κωδικοποιούμενου αμινοξέως. Λαμβάνουμε υπόψιν μας την νουκλεοτιδική σύσταση κατά μήκος των κωδικών κλώνων και διακρίνουμε μεταξύ των $3^{ων}|4$ θέσεων των γονιδίων που βρίσκονται στον οδηγό και στο συνοδό κλώνο, όπου η αντιγραφή και η μεταγραφή πραγματοποιούνται κατά την ίδια ή κατά την αντίθετη κατεύθυνση, αντίστοιχα. Έτσι, χωρίζουμε τις $3^{ες}|4$ θέσεις σε δύο υποσύνολα: (*α*) *κωδικός,οδηγός* και (*β*) *κωδικός,συνοδός* (βλ. Εικόνα 4). Για κάθε δεδομένη απόκλιση (SKew), ορίζουμε ως $SK_{κωδικός,οδηγός}$ ($SK_{sense,leading}$) και $SK_{κωδικός,συνοδός}$ ($SK_{sense,lagging}$) τις τιμές της, όπως αυτές υπολογίστηκαν σε κάθε ένα από αυτά τα υποσύνολα. Για επεξηγηματικούς λόγους, έστω *α* η συνιστώσα αυτής της απόκλισης που επάγει η μεταγραφή και *β* η συνιστώσα που επάγει η αντιγραφή. Οι επαγόμενες από τη μεταγραφή αποκλίσεις έχουν αντίθετα πρόσημα στο κωδικό και τον μεταγραφόμενο κλώνο. Θεωρούμε την συνιστώσα *α* σε σχέση προς τον κωδικό κλώνο. Αντίστοιχα, οι επαγόμενες από την αντιγραφή αποκλίσεις έχουν αντίθετα πρόσημα στον οδηγό και το συνοδό κλώνο. Θεωρούμε την συνιστώσα *β* σε σχέση προς τον οδηγό κλώνο.

Ορίζουμε τις συζευγμένες με τη μεταγραφή (Trs) αποκλίσεις ως τον λόγο $\frac{SK_{sense,leading}+SK_{sense,lagging}}{2}$, που θεωρούμε ότι αποτελεί μία εκτίμηση της τιμής του *α*. Κατ' αναλογία, ορίζουμε τις συζευγμένες με την αντιγραφή (Rep) αποκλίσεις ως τον λόγο $\frac{SK_{sense,leading}-SK_{sense,lagging}}{2}$, που θεωρούμε ότι αποτελεί μία εκτίμηση της τιμής του *β*. Οι δύο αυτές σχέσεις αποτελούν έναν ευριστικό τρόπο αποσύζευξης της επίδρασης που ασκούν η αντιγραφή και η μεταγραφή στις ειδικές ανά κλώνο αποκλίσεις.

Καθώς ορισμένες υποκαταστάσεις εξαρτώνται από την ταυτότητα των γειτονικών τους βάσεων (Arndt & Hwa 2005), λαμβάνουμε υπόψιν και τις $1^{ης}$

τάξης γειτονικές βάσεις των 3ων|4 θέσεων. Από την δομή του γενετικού κώδικα προκύπτει ότι δεν υπάρχουν τετραπλά εκφυλισμένα κωδικόνια τα οποία να φέρουν κατάλοιπα Α στην 2η τους θέση. Για τον λόγο αυτό λαμβάνουμε υπόψιν μας μόνο τα δινουκλεοτίδια, $X_3Y_1$, τα οποία απαντώνται σε κάθε 3η|4 θέση και στην 1η θέση του γειτονικού κωδικονίου (βλ. Εικόνα 4), όπως και στην εργασία των Chamary και Hurst (2004). Στο πλαίσιο αυτό, η σχετική συχνότητα του $X_3Y_1$ ισούται με $\rho^{X_3Y_1} = f^{X_3Y_1}/f^{X_3}f^{Y_1}$, όπου $f^{X_3}$ και $f^{Y_1}$ η παρατηρούμενη συχνότητα του Χ και Υ στις 3ες|4 θέσεις και στις γειτονικές τους 1ες, αντίστοιχα, με $X,Y \in (A,T,G,C)$.




**Εικόνα 4.** Υπολογισμός αποκλίσεων σε δεδομένες θέσεις κωδικονίων. Παραθέτουμε μία σχηματική αναπαράσταση τμήματος της αλληλουχίας του δίκλωνου DNA. Το σημείο έναρξης της αντιγραφής (*ori*) ορίζει τον οδηγό και το συνοδό κλώνο (πράσινη και μαύρη γραμμή, αντίστοιχα). Ενδεικτικά, μέσα σε πλαίσια απεικονίζουμε τμήματα κωδικών περιοχών που αντιστοιχούν σε κωδικόνια, κατά μήκος του κωδικού και του μεταγραφόμενου κλώνου των γονιδίων (μπλε και κόκκινα πλαίσια, αντίστοιχα). Τα βέλη δηλώνουν τις 3ες θέσεις των τετραπλώς εκφυλισμένων κωδικονίων (3ες|4). Οι οριζόντιες γκρι γραμμές δηλώνουν τα δινουκλεοτίδια που απαντώνται σε κάθε 3η|4 θέση και στην 1η θέση του γειτονικού κωδικονίου ($X_3Y_1$).

Διακρίνουμε τις 3ες|4 θέσεις καθώς και τις θέσεις των αντίστοιχων δινουκλεοτιδίων $X_3Y_1$ σε δύο υποσύνολα, σύμφωνα με τον κλώνο της αντιγραφής στον οποίο βρίσκεται ο κωδικός κλώνος των αντίστοιχων γονιδίων (πράσινα βέλη: *κωδικός*, *οδηγός*· μαύρα βέλη: *κωδικός*, *συνοδός*). Για τους υπολογισμούς των συζευγμένων με τη μεταγραφή και την αντιγραφή αποκλίσεων, υπολογίζουμε τη σύσταση του DNA σε κάθε ένα από τα δύο αυτά υποσύνολα θέσεων ξεχωριστά.

## **2.10** *Προσδιορισμός μοριακών φαινοτύπων*

**ΠΙΝΑΚΑΣ 1. Μοριακά μονοπάτια επιδιόρθωσης του DNA**

| ***μοριακοί φαινότυποι*** | | **Γενετικοί τόποι που εμπλέκονται στα μονοπάτια επιδιόρθωσης του DNA** |
|---|---|---|
| **άμεση επιδιόρθωση** | **phrB** | phrB |
| | **ogt** | ogt |
| | **alkB** | alkB |
| | **ada** | ada |
| **BER** | **ung** | ung, [xthA or nfo], polA, [ligATP *or* ligNAD] |
| | **mug** | mug, [xthA or nfo], polA, [ligATP *or* ligNAD] |
| | **nth** | nth, [xthA or nfo], polA, [ligATP *or* ligNAD] |
| | **mutM** | mutM, [xthA or nfo], polA, [ligATP *or* ligNAD] |
| | **nei** | nei, [xthA or nfo], polA, [ligATP *or* ligNAD] |
| | **tag** | tag, [xthA or nfo], polA, [ligATP *or* ligNAD] |
| | **alkA** | alkA, [xthA or nfo], polA, [ligATP *or* ligNAD] |
| | **mutY** | mutY, [xthA or nfo], polA, [ligATP *or* ligNAD] |
| | **GO system** | mutM, mutY, mutT, [xthA or nfo], polA, [ligATP *or* ligNAD] |
| **NER** | **GGR** | uvrA, uvrB, uvrC, uvrD.pcrA, polA, [ligATP *or* ligNAD] |
| | **TCR** | mfd, uvrA, uvrB, uvrC, uvrD.pcrA, polA, [ligATP *or* ligNAD] |
| **MMR** | **καθοδηγούμενη από μεθυλίωση** | mutS, mutL, mutH, uvrD.pcrA, dam |
| | **καθοδηγούμενη από εγκοπή** | mutS, mutL, uvrD.pcrA (*****) |
| | **πολύ βραχέως τμήματος (VSP)** | vsr, dam |
| **RR** | **RecFOR** | recJ, ssb, recO, recR, recA, [ruvABC *or* recG] |
| | **RecBC** | recB, recC, recD, recA, recomb, priA, priB, priC, dnaT |

ΣΗΜΕΙΩΣΕΙΣ.- Αντιστοίχηση μοριακών φαινοτύπων με συγκεκριμένα επιδιορθωτικά μονοπάτια. Οι γενετικοί τόποι δίδονται σύμφωνα με τον συμβολισμό που χρησιμοποιείται στην βάση δεδομένων KEGG orthology (Du et al. 2014). Κάθε μοριακός φαινότυπος αντιστοιχεί στο σύνολο των γενετικών τόπων που εμπλέκονται σε ένα δεδομένο μονοπάτι επιδιόρθωσης, όπως αυτό περιγράφεται στην βάση δεδομένων KEGG pathway maps. Για κάθε μονοπάτι, θεωρούμαι ότι ένα βακτηριακό είδος είναι ικανό για επιδιόρθωση εάν το σύνολο των αντίστοιχων γενετικών τόπων εντοπίζεται στο γονιδίωμά του (βλ. ενότητα 1.6.2).
(*): μόνο τα βακτήρια δίχως mut*H* λαμβάνονται υπόψιν, BER: Επιδιόρθωση με εκτομή βάσης, NER: επιδιόρθωση με εκτομή νουκλεοτιδίου, MMR: Επιδιόρθωση αταίριαστων ζευγών, RR: Επιδιόρθωση μέσω ανασυνδυασμού.

Διακρίνουμε τα βακτήρια που αντιπροσωπεύονται στην συλλογή μας σύμφωνα με την παρουσία ή απουσία συγκεκριμένων μοριακών μηχανισμών, οι οποίοι δρουν ή ενδέχεται να δρουν με διακριτό τρόπο στους δύο κλώνους του DNA. Συγκεκριμένα, λαμβάνουμε από την βάση δεδομένων KEGG orthology (Du et al. 2014) τον τύπο της *α*-υπομονάδας της πολυμεράσης III (PolIII) που φέρουν τα βακτήρια. Επίσης, χωρίζουμε την συλλογή μας σε βακτήρια με ή δίχως ικανότητα επιδιόρθωσης του DNA μέσω συγκεκριμένων επιδιορθωτικών μονοπατιών, όπως αυτά προσδιορίζονται στον ανωτέρω πίνακα.

## **2.11** *Μοριακοί μηχανισμοί που σχετίζονται με ασύμμετρα πρότυπα υποκατάστασης*

Προσδιορίζουμε τις συζευγμένες με τη μεταγραφή ή την αντιγραφή αποκλίσεις (βλ. ενότητα 2.9) για το σύνολο των χρωμοσωμάτων της συλλογής μας. Ακολούθως, για κάθε έναν από τους μοριακούς μηχανισμούς που μελετάμε, χωρίζουμε την συλλογή μας σε υποσύνολα, καθένα από τα οποία αντιστοιχεί σε έναν συγκεκριμένο μοριακό φαινότυπο (βλ. ενότητα 2.10). Προκειμένου να ανιχνεύσουμε πιθανές συσχετίσεις των εν λόγω μηχανισμών με ειδικές ανά κλώνο πολώσεις στους ρυθμούς υποκατάστασης, συγκρίνουμε τις κατανομές που ακολουθούν οι αποκλίσεων στα υπό εξέταση υποσύνολα. Για τον σκοπό αυτό εφαρμόζουμε τον *στατιστικό έλεγχο αθροίσματος διατάξεων του Wilcoxon* (two-tailed Wilcoxon rank-sum test). Πρόκειται για έναν μη παραμετρικό

έλεγχο της μηδενικής υπόθεσης ($H_0$) σύμφωνα με την οποία δύο σύνολα τιμών προέρχονται από την ίδια κατανομή. Η εναλλακτική υπόθεση ($H_a$) αυτού του ελέγχου δηλώνει στατιστικά σημαντική διάκριση των υπό σύγκριση συνόλων. Επιλέγουμε το συγκεκριμένο τεστ καθώς δεν προϋποθέτει κανονική κατανομή των δεδομένων, ενώ μπορεί να εφαρμοστεί και όταν τα δύο σύνολα τιμών διαφέρουν σημαντικά ως προς το μέγεθός τους.

Στα πλαίσια της ανάλυσής μας, απόρριψη της μηδενικής υπόθεσης δηλώνει την ύπαρξη στατιστικά σημαντικής συσχέτισης του εκάστοτε μοριακού φαινοτύπου με τις παρατηρούμενες αποκλίσεις, γεγονός που υποδεικνύει ότι ο αντίστοιχος μοριακός μηχανισμός μπορεί να επάγει ασυμμετρίες στα πρότυπα υποκατάστασης μεταξύ των αντιστρόφως συμπληρωματικών κλώνων.

# 3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΥΖΗΤΗΣΗ

Στις ενότητες που ακολουθούν μελετάμε χαρακτηριστικά της σύστασης του DNA που είναι συνυφασμένα με την ασύμμετρη εξέλιξη των κλώνων του. Επιλέγουμε να πραγματοποιήσουμε τις αναλύσεις μας σε βακτηριακά χρωμοσώματα, καθώς η οργάνωσή τους εμφανίζει χαμηλότερο βαθμό πολυπλοκότητας από εκείνον των ευκαρυωτικών γονιδιωμάτων. Το γεγονός αυτό μας επιτρέπει να διακρίνουμε με μεγαλύτερη σαφήνεια μεταξύ των παραγόντων που διαμορφώνουν τις αποκλίσεις από τη συμμετρία και ως εκ τούτου να εξάγουμε ασφαλέστερα συμπεράσματα.

Αρχικά μελετάμε τις αποκλίσεις στο επίπεδο της μονονουκλεοτιδικής σύστασης του DNA και διαπιστώνουμε πως η συλλογή μας είναι αντιπροσωπευτική των γενικών τάσεων ασυμμετρίες που είναι ήδη γνωστές για τα βακτηριακά γονιδιώματα (ενότητα 3.1). Ακολούθως, εξετάζουμε στις κατανομές που ακολουθούν οι τιμές των σταθμισμένων δινουκλεοτιδικών συχνοτήτων στη συλλογή μας (ενότητα 3.2). Εστιάζουμε την ανάλυσή μας στις σχέσεις ανισότητας που εμφανίζονται μεταξύ των σταθμισμένων συχνοτήτων των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων και ανιχνεύουμε την ύπαρξη συστηματικών ασυμμετριών στο επίπεδο αυτό (ενότητα 3.3). Ποσοτικοποιούμε τις αποκλίσεις από τη συμμετρία που εμφανίζουν τα δινουκλεοτίδια, σε όρους τόσο παρατηρούμενων όσο και σταθμισμένων συχνοτήτων, και μελετάμε την ένταση και τη φορά τους (ενότητα 3.4). Ιδιαίτερη έμφαση δίδεται στην κατανομή των αποκλίσεων που παρατηρούνται στον οδηγό κλώνο (ενότητα 3.5). Συνολικά, οι ενότητες 3.3-3.5 στοιχειοθετούν την ύπαρξη ασυμμετριών στο επίπεδο των συσχετίσεων μεταξύ των 1$^{ης}$ τάξης γειτονικών βάσεων, οι οποίες δεν μπορούν να αναχθούν στις ασυμμετρίες της μονο- ή δι-νουκλεοτιδικής σύστασης του DNA.

Οι αποκλίσεις των δινουκλεοτιδίων και των σταθμισμένων τους συχνοτήτων εκδηλώνονται με μεγαλύτερη ένταση όταν εξετάζονται στην τοπική κλίμακα μικρών χρωμοσωμικών περιοχών (ενότητα 3.6). Απεικονίζουμε γραφικά τις τιμές αυτών των αποκλίσεων με τη χρήση αθροιστικών διαγραμμάτων και διαπιστώνουμε ότι οι ασυμμετρίες που μελετάμε οργανώνονται σε χαρακτηριστικά πρότυπα κατά

μήκος του χρωμοσώματος (ενότητα 3.7). Ελέγχουμε τη στατιστική σημαντικότητα αυτών των προτύπων και δείχνουμε ότι οι αποκλίσεις των δινουκλεοτιδίων και των σταθμισμένων τους συχνοτήτων συσχετίζονται ισχυρά με το σημείο έναρξης της αντιγραφής και τη φορά της μεταγραφής (ενότητα 3.8).

Προκειμένου να διαπιστώσουμε εάν οι ασυμμετρίες των εξελικτικών πιέσεων που ασκούνται στο γονιδίωμα είναι κοινές μεταξύ διαφορετικών οργανισμών ή εάν αντίθετα είναι ανά είδος καθορισμένες, μελετάμε τη συσχέτιση των αποκλίσεων με τη φυλογένεση των βακτηρίων (ενότητα 3.9). Σύμφωνα με τα αποτελέσματά μας, οι αποκλίσεις των δινουκλεοτιδίων και των σταθμισμένων τους συχνοτήτων φέρουν πληροφορία που μπορεί να χρησιμοποιηθεί στην ανασυγκρότηση των φυλογενετικών σχέσεων των βακτηρίων και συνεπώς αντανακλούν εξελικτικές ασυμμετρίες που είναι ανά είδος καθορισμένες. Οι ασυμμετρίες αυτές δεν μπορούν να αναχθούν σε ομοιότητα λόγω ομολογίας των χρωμοσωμικών περιοχών στις οποίες υπολογίζονται οι αντίστοιχες αποκλίσεις (ενότητα 3.10). Συνεπώς, οι ρίζες της ασύμμετρης εξέλιξης του γονιδιώματος θα πρέπει να αναζητηθούν σε μοριακούς μηχανισμούς και διαδικασίες των οποίων η δια-ειδική ποικιλότητα είναι ικανή να παράγει ανά είδος καθορισμένες ασυμμετρίες στους ρυθμούς υποκατάστασης (βλ. ενότητες 3.11, 3.12).

Από τη μελέτη των προτύπων που εμφανίζουν οι συσχετίσεις μεταξύ των αποκλίσεων και της χρήσης κωδικονίων, προκύπτει ότι η δια-ειδική ποικιλότητα του GC περιεχομένου συμβάλλει καθοριστικά στην ασύμμετρη εξέλιξη των κωδικών περιοχών (ενότητες 3.11.1, 3.11.2). Εισάγουμε ένα μοντέλο που περιγράφει την χρήση κωδικονίων ως συνάρτηση του GC περιεχομένου των κωδικών περιοχών, λαμβάνοντας υπόψιν την ποικιλότητα σε GC που εμφανίζεται εντός κάθε ομάδας συνωνύμων (ενότητα 3.11). Βάσει αυτού του μοντέλου, συμπεραίνουμε πως όταν οι κωδικές περιοχές υπόκεινται σε GC κατευθύνουσες μεταλλακτικές πιέσεις, η δομή του γενετικού κώδικα επιβάλλει περιορισμούς που οδηγούν στην ασύμμετρη κατανομή των νουκλεοτιδίων μεταξύ κωδικού και μεταγραφόμενου κλώνου, ακόμα και όταν δεν λαμβάνεται υπόψιν η επιλογή στη χρήση κωδικονίων και αμινοξέων.

Η ασύμμετρη εξέλιξη του γενετικού υλικού σχετίζεται επίσης με τους μοριακούς μηχανισμούς αντιγραφής, τροποποίησης και επιδιόρθωσης του DNA (ενότητα 3.12). Τα βακτήρια εμφανίζουν μεγάλη δια-ειδική ποικιλότητα ως προς τους επιδιορθωτικούς μηχανισμούς που διαθέτουν. Η ανάλυση που

εφαρμόζουμε μας επιτρέπει να διακρίνουμε ποιοι εξ αυτών των μηχανισμών μπορούν να επάγουν ειδικές ανά κλώνο πολώσεις των μεταλλακτικών ρυθμών, στο επίπεδο ολόκληρου του γονιδιώματος. Εστιάζοντας στο μηχανισμό της αντιγραφής, διαπιστώνουμε ότι οι διαφορετικές ισομορφές της α-καταλυτικής υπομονάδας της PolIII επάγουν συστηματικές ασυμμετρίες στους ρυθμούς των μεταλλάξεων. Οι ασυμμετρίες αυτές αφορούν και μεταλλάξεις που σχετίζονται με την ταυτότητα των γειτονικών τους βάσεων, όπως υποδεικνύει η ανάλυση των αποκλίσεων των σταθμισμένων δινουκλεοτιδικών συχνοτήτων.

## **3.1** *Μονονουκλεοτιδικές αποκλίσεις*

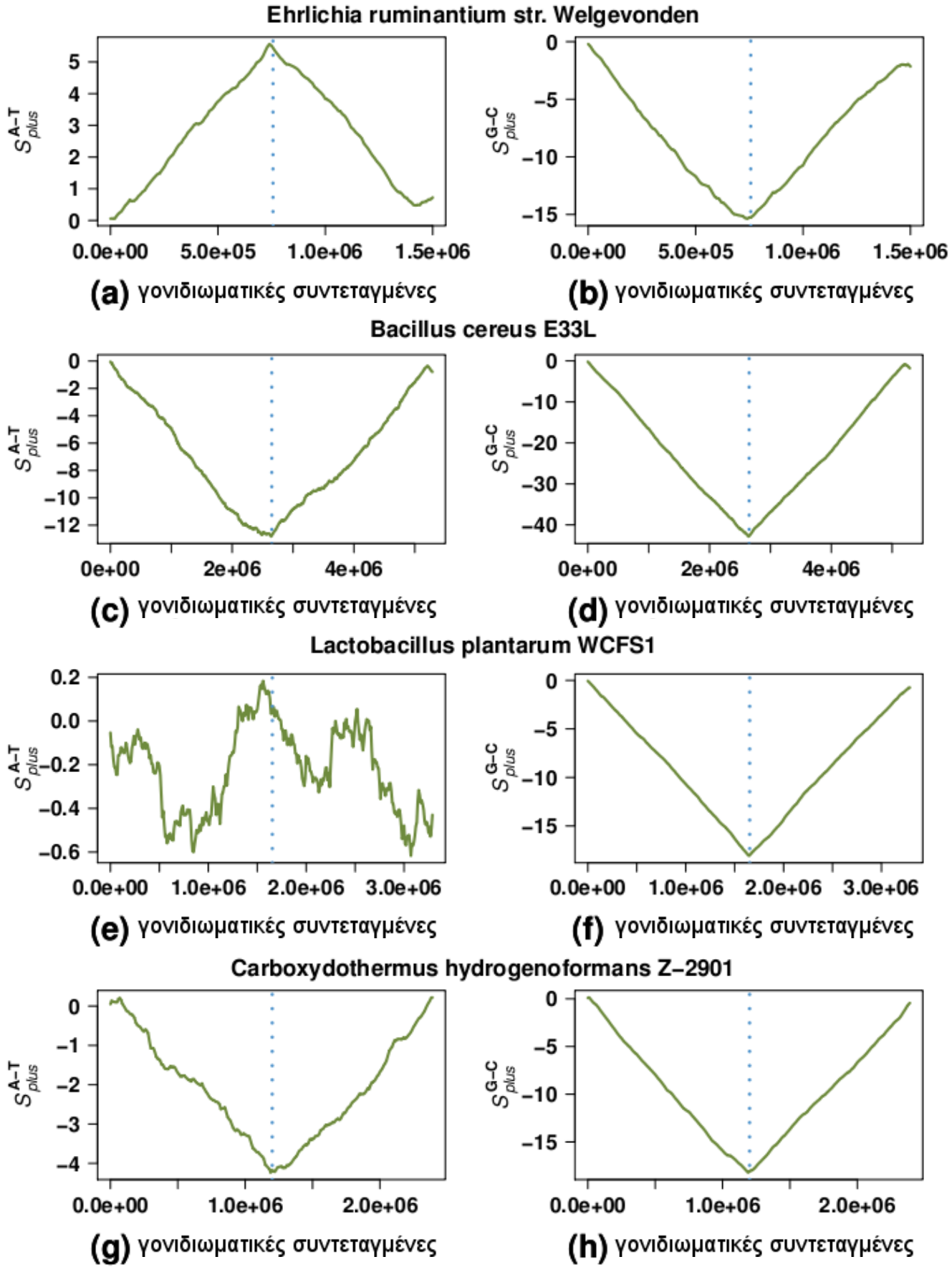
Σύμφωνα με την μεθοδολογία που περιγράφεται αναλυτικά στην ενότητα 2.1, σε όλα τα κυκλικά χρωμοσώματα της συλλογής μας μετατοπίσαμε τις γονιδιωματικές συντεταγμένες κατά τρόπο ώστε το σημείο έναρξης της αντιγραφής (*ori*) να τοποθετείται στο μέσον της δημοσιευμένης αλληλουχίας. Ως σημείο λήξης της αντιγραφής (*ter*) θέσαμε την περιοχή που απέχει από το *ori* απόσταση ίση με το μισό του χρωμοσώματος, σύμφωνα και με προγενέστερες γονιδιωματικές μελέτες (Mao et al. 2012, Saha et al. 2014). Κατά τον τρόπο αυτό το *ter* αντιστοιχεί στο τέλος των μετατοπισμένων αλληλουχιών, ή, πράγμα που είναι ταυτόσημο εφόσον πρόκειται για κυκλικά χρωμοσώματα, στην αρχή τους. Συνεπώς, το πρώτο μισό αυτών των αλληλουχιών αντιστοιχεί στο συνοδό κλώνο και το δεύτερο μισό στον οδηγό. Ακολούθως, για κάθε χρωμόσωμα της συλλογής μας, σχεδιάσαμε τα αθροιστικά γραφήματα των αποκλίσεων για τα δύο ζεύγη των συμπληρωματικών βάσεων ($S_{plus}^{A-T}$, $S_{plus}^{G-C}$).

Στην Εικόνα 5 απεικονίζονται τα αθροιστικά διαγράμματα των $S_{plus}^{A-T}$ και $S_{plus}^{G-C}$, κατά μήκος του δημοσιευμένου κλώνου των χρωμοσωμάτων τεσσάρων βακτηρίων: *Ehrlichia ruminantium* str. Welgevonden, *Bacillus cereus* E33L, *Lactobacillus plantarum* WCFS1 και *Carboxydothermus hydrogenoformans* Z-2901. Εάν οι αποκλίσεις είναι κατά προσέγγιση σταθερές κατά μήκος του οδηγού και του συνοδού κλώνου, ενώ παράλληλα έχουν αντίθετο πρόσημο εκατέρωθεν του σημείου έναρξης της αντιγραφής, τότε τα αθροιστικά διαγράμματα έχουν τη χαρακτηριστική μορφή V ή ανεστραμμένου V (Grigoriev 1998). Στα διαγράμματα

αυτά το *ter* βρίσκεται στην αρχή (ή στο τέλος) της κάθε καμπύλης και το *ori* στο μέσον της, όπου εντοπίζεται και το ακρότατό της. Τα αθροιστικά διαγράμματα V (Εικόνα 5*a*) αντιστοιχούν σε χρωμοσώματα όπου οι αποκλίσεις είναι θετικές στον οδηγό και αρνητικές στο συνοδό, ενώ τα αθροιστικά διαγράμματα ανεστραμμένου V (Εικόνα 5*b*-*d*,*f*-*h*) σε χρωμοσώματα όπου, αντίστροφα, οι αποκλίσεις είναι αρνητικές στον οδηγό και θετικές στο συνοδό.

Και στις τέσσερις εικονιζόμενες περιπτώσεις, τα $S_{plus}^{G-C}$ αθροιστικά διαγράμματα έχουν μορφή V, οπότε το *ori* συμπίπτει με το ελάχιστο της κάθε καμπύλης. Τα διαγράμματα αυτά καταδεικνύουν την περίσσεια καταλοίπων G σε σχέση με τα κατάλοιπα C κατά μήκος του οδηγού κλώνου, μία τάση που θεωρείται σχεδόν καθολική στα βακτήρια. Αντίθετα, τα $S_{plus}^{A-T}$ αθροιστικά διαγράμματα παρουσιάζουν μεγαλύτερη ποικιλομορφία, καθώς η μορφή τους είναι είτε ανεστραμμένο V (Εικόνα 5*a*), είτε V (Εικόνα 5*c*,*g*) είτε έντονα παραμορφωμένη, με ακανόνιστες κορυφές και ελάχιστα (Εικόνα 5*e*). Στο χρωμόσωμα του *Ehrlichia ruminantium* str. Welgevonden υπάρχει περίσσεια Ts έναντι As στον οδηγό κλώνο (Εικόνα 5*a*, ανεστραμμένο V: $S_{plus}^{A-T} > 0$ στο συνοδό, $S_{plus}^{A-T} < 0$ στον οδηγό). Η περίπτωση αυτή είναι η πλέον τυπική για τα βακτηριακά χρωμοσώματα, με την χαρακτηριστική εξαίρεση του φύλλου Firmicutes (Lobry & Sueoka 2002, Morton & Morton 2007, Charneski et al. 2011), πολλά μέλη του οποίου εμφανίζουν την αντίστροφη πόλωση στη σύστασή τους, με περίσσεια As έναντι Ts στον οδηγό (Εικόνα 5*c*,*g*).

Συγκρίνοντας τα ακρότατα των $S_{plus}^{A-T}$ και $S_{plus}^{G-C}$ αθροιστικών διαγραμμάτων της Εικόνας 5, διαπιστώνουμε πως οι αποκλίσεις G-C είναι κατά πολύ εντονότερες από τις A-T, κατά μήκος του ίδιου χρωμοσώματος. Στην περίπτωση μάλιστα του *Lactobacillus plantarum* WCFS1, ενώ οι αποκλίσεις G-C σχηματίζουν σαφή V αθροιστικά διαγράμματα, οι αποκλίσεις A-T είναι τόσο ασθενείς ώστε η αντίστοιχη καμπύλη να έχει εντελώς ακανόνιστη μορφή.


Ehrlichia ruminantium str. Welgevonden
$S^{A-T}_{plus}$
$S^{G-C}_{plus}$
0.0e+00
5.0e+05
1.0e+06
1.5e+06
(a) γονιδιωματικές συντεταγμένες
(b) γονιδιωματικές συντεταγμένες
Bacillus cereus E33L
0e+00
2e+06
4e+06
(c) γονιδιωματικές συντεταγμένες
(d) γονιδιωματικές συντεταγμένες
Lactobacillus plantarum WCFS1
0.0e+00
1.0e+06
2.0e+06
3.0e+06
(e) γονιδιωματικές συντεταγμένες
(f) γονιδιωματικές συντεταγμένες
Carboxydothermus hydrogenoformans Z−2901
(g) γονιδιωματικές συντεταγμένες
(h) γονιδιωματικές συντεταγμένες

**Εικόνα 5.** Αθροιστικά διαγράμματα των μονονουκλεοτιδικών αποκλίσεων κατά μήκος του δημοσιευμένου (plus) κλώνου τεσσάρων βακτηριακών χρωμοσωμάτων. Η κατακόρυφη, στικτή μπλε γραμμή δηλώνει το σημείο έναρξης της αντιγραφής (*ori*). Αριστερά του *ori* οι εικονιζόμενες αποκλίσεις αντιστοιχούν στο συνοδό κλώνο, ενώ δεξιά του *ori* αντιστοιχούν στον οδηγό κλώνο. Στις περιπτώσεις που τα διαγράμματα έχουν τη χαρακτηριστική μορφή V (*a*) ή ανεστραμμένο V(*b-d*, *f-h*), τα ακρότατά τους συμπίπτουν με την περιοχή του *ori*.

Διατρέχοντας τα αθροιστικά διαγράμματα των μονονουκλεοτιδικών αποκλίσεων για το σύνολο της συλλογής μας (βλ. σχετικό link στο τέλος της Βιβλιογραφίας) επιβεβαιώνονται σε μεγάλο βαθμό τα ήδη γνωστά πρότυπα των ειδικών ανά κλώνο ασυμμετριών. Οι μορφές αυτών των διαγραμμάτων ποικίλουν από τις χαρακτηριστικές, V ή ανεστραμμένο V, έως τις έντονα παραμορφωμένες, με τις αντίστοιχες καμπύλες να εμφανίζουν τοπικά ακρότατα με έναν ακανόνιστο τρόπο. Προκειμένου να συνοψίσουμε τις παρατηρήσεις μας με ημι-ποσοτικό τρόπο, υπολογίσαμε σε κάθε χρωμόσωμα την ολική τιμή των $S_{plus}^{A-T}$ και $S_{plus}^{G-C}$, δίχως τη χρήση κυλιόμενων παραθύρων, ξεχωριστά για τον οδηγό και το συνοδό κλώνο. Διακρίναμε μεταξύ τριών διαφορετικών περιπτώσεων: (*α*) οι αποκλίσεις έχουν θετική τιμή στον οδηγό και αρνητική στο συνοδό κλώνο, οπότε τα αθροιστικά διαγράμματά τους είναι τύπου-V, (*β*) οι αποκλίσεις έχουν αρνητική τιμή στον οδηγό και θετική στο συνοδό κλώνο, οπότε τα αθροιστικά διαγράμματά τους είναι τύπου-ανεστραμμένου V, και (*γ*) οι αποκλίσεις έχουν το ίδιο πρόσημο (είτε θετικό είτε αρνητικό) και στους δύο κλώνους, οπότε τα αθροιστικά διαγράμματά τους είναι παραμορφωμένα. Λαμβάνοντας υπόψιν τις ιδιαιτερότητες που εμφανίζουν τα Firmicutes όσον αφορά τις μονονουκλεοτιδικές τους αποκλίσεις, ιδίως τα μη-τυπικά πρότυπα των $S_{plus}^{A-T}$, μελετήσαμε τα μέλη αυτού του φύλου ξεχωριστά. Έτσι, χωρίσαμε τη συλλογή μας σε δύο υποσύνολα, στα γονιδιώματα των Firmicutes (64 τον αριθμό) και στα γονιδιώματα όλων των υπολοίπων βακτηρίων, εκτός-Firmicutes (276 τον αριθμό). Ακολούθως, μετρήσαμε πόσα από τα χρωμοσώματα της συλλογής μας περιλαμβάνονται σε κάθε μία από τις τρεις κατηγορίες, τύπου-V, τύπου-ανεστραμμένου V και παραμορφωμένα. Τα αποτελέσματα παρατίθενται στον Πίνακα 2. Σημειώνουμε πως αυτό το σχήμα ταξινόμησης των αποκλίσεων επιτρέπει μια κατά προσέγγιση σύνοψη των αθροιστικών διαγραμμάτων, καθώς ενδέχεται η τιμή μίας απόκλισης να είναι θετική (ή αρνητική) στον οδηγό και αρνητική (ή θετική, αντιστοίχως) στο συνοδό, χωρίς ωστόσο οι αντίστοιχες καμπύλες να έχουν σαφώς καθορισμένη μορφή.

**ΠΙΝΑΚΑΣ 2. Ποσοστιαία κατάταξη των χρωμοσωμάτων, βάσει των γενικών χαρακτηριστικών της μορφής που έχουν τα αθροιστικά διαγράμματα των μονονουκλεοτιδικών τους αποκλίσεων.**

| | **εκτός-Firmicutes** | | | **Firmicutes** | | |
|---|---|---|---|---|---|---|
| | ***παραμορφωμένα διαγράμματα*** | ***διαγράμματα τύπου-V*** | ***διαγράμματα τύπου ανεστραμμένου V*** | ***παραμορφωμένα διαγράμματα*** | ***διαγράμματα τύπου-V*** | ***διαγράμματα τύπου ανεστραμμένου V*** |
| $S_{plus}^{A-T}$ | 10.9 | 11.6 | 77.5 | 1.56 | 93.7 | 4.7 |
| $S_{plus}^{G-C}$ | 3.99 | 94.2 | 1.8 | 0 | 100.0 | 0.0 |

ΣΗΜΕΙΩΣΕΙΣ.- ***Παραμορφωμένα*** θεωρούνται τα αθροιστικά διαγράμματα που αντιστοιχούν σε αποκλίσεις των οποίων το πρόσημο δεν αλλάζει εκατέρωθεν του σημείου *ori*. Στα χρωμοσώματα των οποίων οι αποκλίσεις αλλάζουν πρόσημο εκατέρωθεν του *ori*, τα αντίστοιχα αθροιστικά διαγράμματα θεωρούνται είτε ως ***τύπου-V*** (θετικές αποκλίσεις στον οδηγό και αρνητικές αποκλίσεις στο συνοδό κλώνο) είτε ως ***τύπου-ανεστραμμένου V*** (αρνητικές αποκλίσεις στον οδηγό και θετικές αποκλίσεις στο συνοδό κλώνο).

Όπως προκύπτει από τον Πίνακα 2, τόσο τα χρωμοσώματα των Firmicutes όσο και εκείνα των εκτός-Firmicutes έχουν σχεδόν στο σύνολό τους αθροιστικά διαγράμματα *τύπου*-V ή *ανεστραμμένου* V. Συγκεκριμένα, στα εκτός-Firmicutes βακτήρια, το 10.9% έχει παραμορφωμένα $S_{plus}^{A-T}$ αθροιστικά διαγράμματα, ενώ μόλις το 3.99% έχει παραμορφωμένα $S_{plus}^{G-C}$ αθροιστικά διαγράμματα. Στα εκτός-Firmicutes, η μεγάλη πλειοψηφία των χρωμοσωμάτων (94.2%) έχει $S_{plus}^{G-C}$ αθροιστικά διαγράμματα *τύπου*-V, γεγονός που συμβαδίζει με την τάση εμπλουτισμού του οδηγού κλώνου σε Gs έναντι Cs. Αντίθετα, στο ίδιο υποσύνολο, το 77.5% των $S_{plus}^{A-T}$ αθροιστικών διαγραμμάτων είναι *τύπου-ανεστραμμένου* V και συνεπώς ο οδηγός κλώνος εμφανίζει περίσσεια Ts έναντι As. Στρεφόμενοι στα Firmicutes, παρατηρούμε πως όλα τα χρωμοσώματα έχουν $S_{plus}^{G-C}$ οι οποίες αλλάζουν πρόσημο εκατέρωθεν του *ori* (0% *παραμορφωμένα* $S_{plus}^{G-C}$ αθροιστικά διαγράμματα), ενώ μόλις ένα χρωμόσωμα, που προέρχεται από

το *Streptococcus mutans* UA159, έχει *παραμορφωμένο* $S_{plus}^{A-T}$ διάγραμμα (1.56% σε σύνολο 64 DNA αλληλουχιών). Ακολουθώντας τη γενική τάση των βακτηριακών χρωμοσωμάτων, τα Firmicutes εμφανίζουν στο σύνολό τους περίσσεια καταλοίπων G έναντι C στον οδηγό κλώνο, με το 100% να έχει $S_{plus}^{G-C}$ αθροιστικά διαγράμματα *τύπου*-V. Αντίθετα, ενώ $S_{plus}^{A-T}$ σχηματίζουν αθροιστικά διαγράμματα *τύπου-ανεστραμμένου* V στην πλειονότητα των υπλοίπων βακτηριακών φύλων, το 93.7% των Firmicutes έχει μη-τυπικά $S_{plus}^{A-T}$ αθροιστικά διαγράμματα *τύπου*-V. Οι παρατηρήσεις αυτές έρχονται σε συμφωνία με ήδη γνωστά πρότυπα των ειδικών ανά κλώνο αποκλίσεων και δείχνουν πως η συλλογή μας είναι αντιπροσωπευτική των τάσεων που εκδηλώνουν οι ασυμμετρίες της νουκλεοτιδικής σύστασης στα γονιδιώματα των βακτηρίων.

## **3.2** *Προφίλ δινουκλεοτιδίων και συσχετίσεις κοντινότερων γειτονικών βάσεων*

Οι συχνότητες εμφάνισης των δινουκλεοτιδίων δεν ακολουθούν τυχαίες διακυμάνσεις κατά μήκος των αλληλουχιών DNA (Beutler et al. 1989, Kozhukhin & Pevzner 1991), όπως άλλωστε έχει ήδη επισημανθεί από προγενέστερες σχετικές μελέτες, αλλά συγκροτούν γενικά μοτίβα τα οποία επαναλαμβάνονται ακόμα και μεταξύ οργανισμών που ανήκουν σε εξελικτικά απομακρυσμένες ταξινομικές ομάδες (Nussinov 1980, Nussinov 1981, Nussinov 1984a, Nussinov 1984b). Τέτοια μοτίβα υποδηλώνουν την ύπαρξη συγκεκριμένων και ευρέως διαδεδομένων ιεραρχήσεων μεταξύ των δινουκλεοτιδίων. Προκειμένου να διαχωριστούν τα πρότυπα υπο- και υπερ-εκπροσώπησης των δινουκλεοτιδίων από τις τάσεις που εκδηλώνονται στο επίπεδο την μονονουκλεοτιδικής σύστασης, υπολογίζονται οι δινουκλεοτιδικές σταθμισμένες συχνότητες (Nussinov 1984b, Burge et al. 1992). Συγκεκριμένα, η παρατηρούμενη συχνότητα εμφάνισης κάθε νουκλεοτιδικού διμερούς κανονικοποιείται ως προς την αναμενόμενη συχνότητά του, όπως αυτή προκύπτει εάν τα μεμονωμένα νουκλεοτίδια κατανέμονταν τυχαία κατά μήκος της DNA αλληλουχίας. Σύμφωνα με την Nussinov (1984b), οι δινουκλεοτιδικές σταθμισμένες συχνότητες κατατάσσονται ιεραρχικά κατά τρόπο παρεμφερή στους διάφορους οργανισμούς, και ειδικότερα στα βακτήρια, που

είναι το αντικείμενο της μελέτης μας, ξεκινώντας από τα πλέον συχνά απαντώμενα, όπως είναι τα AA και TT, και καταλήγοντας σε εκείνα τα οποία με ένταση αποφεύγονται, όπως τα AC και TA.

Καθώς η μελέτη μας πραγματεύεται συμμετρίες, αλλά και αποκλίσεις από την κανονικότητα της σύστασης του DNA τόσο στο επίπεδο των μονονουκλεοτιδίων όσο και σε αυτό των δινουκλεοτιδίων, επιχειρήσαμε να ανιχνεύσουμε τις τάσεις που ακολουθούν οι δινουκλεοτιδικές σταθμισμένες συχνότητες στα γονιδιώματα που εξετάζουμε, ώστε να διαπιστώσουμε σε πιο βαθμό αυτές συμμορφώνονται με τα μοτίβα που έχουν ήδη καταγραφεί (Nussinov 1980, Nussinov 1981, Nussinov 1984b, Ohno 1988, Beutler et al. 1989, Yomo & Ohno 1989, Kozhukhin & Pevzner 1991, Burge et al. 1992). Για το σκοπό αυτό, μελετήσαμε την κατανομή των 16 σταθμισμένων δινουκλεοτιδικών συχνοτήτων στο σύνολο των αλληλουχιών DNA της συλλογής μας. Για κάθε χρωμόσωμα οι σχετικοί υπολογισμοί έγιναν ξεχωριστά κατά μήκος ($\alpha$) του δημοσιευμένου κλώνου (plus strand), ($\beta$) του οδηγού κλώνου (leading strand) και ($\gamma$) των CDS-συρραφών (CDS concatenates). Τα αποτελέσματα παρουσιάζονται στον Πίνακα 3.

Δινουκλεοτίδια με σταθμισμένες συχνότητες πλησίον της μονάδας ($0.99 \leq \rho \leq 1{,}01$) απαντώνται με συχνότητα εμφάνισης η οποία καθορίζεται ευθέως από την μονονουκλεοτιδική σύσταση της αλληλουχίας, δίχως να παρατηρούνται προτιμήσεις μεταξύ των γειτονικών βάσεων. Σταθμισμένες συχνότητες μικρότερες του 0.8 ή μεγαλύτερες του 1.2 δηλώνουν συστηματικά πρότυπα υπο- ή υπερ-εκπροσώπησης των αντίστοιχων δινουκλεοτιδίων. Τιμές σταθμισμένων συχνοτήτων εντός των διαστημάτων [0.8, 0.99) και (1.01, 1.2] μπορεί να προκύπτουν από ασθενείς τάσεις υπό- ή υπερ-εκπροσώπησης, αντίστοιχα, ή απλώς από στοχαστικές διακυμάνσεις γύρω από τις τιμές των δινουκλεοτιδικών συχνοτήτων εμφάνισης που αναμένονται όταν δεν εκδηλώνονται συσχετίσεις μεταξύ των γειτονικών βάσεων. Τα παραπάνω όρια διαστημάτων επιλέγονται ενδεικτικά ώστε να απεικονίζουν τις γενικές τάσεις υπό- ή υπερ-εκπροσώπησης των δινουκλεοτιδίων.

**ΠΙΝΑΚΑΣ 3. Κατανομή των σταθμισμένων δινουκλεοτιδικών συχνοτήτων**

| | **Α. δημοσιευμένος κλώνος** | | | | | **Β. οδηγός κλώνος** | | | | | **Γ. CDS-συρραφές** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **[0,0.80)** | **[0.80,0.99)** | **[0.99,1.01]** | **(1.01,1.20)** | **[1.20,maximum)** | **[0,0.80)** | **[0.80,0.99)** | **[0.99,1.01]** | **(1.01,1.20)** | **[1.20,maximum)** | **[0,0.80)** | **[0.80,0.99)** | **[0.99,1.01]** | **(1.01,1.20)** | **[1.20,maximum)** |
| $\rho^{AA}$ | 0 | 5.29 | 1.76 | 64.41 | 28.53 | 0 | 5.29 | 0.59 | 63.82 | 30.29 | 0 | 4.71 | 1.47 | 61.18 | 32.65 |
| $\rho^{TT}$ | 0 | 5.29 | 1.76 | 62.65 | 30.29 | 0.29 | 6.18 | 0.88 | 65.59 | 27.06 | 2.35 | 12.65 | 2.65 | 59.12 | 23.24 |
| $\rho^{AC}$ | 25.29 | 69.12 | 0.59 | 5 | 0 | 27.65 | 67.65 | 0.59 | 4.12 | 0 | 17.35 | 72.65 | 2.35 | 6.76 | 0.88 |
| $\rho^{GT}$ | 25.59 | 68.82 | 0.59 | 5 | 0 | 25.29 | 68.53 | 1.76 | 4.41 | 0 | 35.29 | 62.06 | 1.18 | 1.47 | 0 |
| $\rho^{AG}$ | 7.65 | 65 | 2.65 | 24.41 | 0.29 | 11.76 | 61.76 | 1.76 | 22.94 | 1.76 | 34.41 | 44.41 | 6.47 | 14.71 | 0 |
| $\rho^{CT}$ | 7.65 | 65 | 2.06 | 25 | 0.29 | 8.82 | 63.24 | 2.94 | 24.12 | 0.88 | 2.35 | 57.94 | 6.76 | 28.24 | 4.71 |
| $\rho^{CA}$ | 0 | 17.35 | 7.06 | 68.53 | 7.06 | 0.29 | 17.35 | 8.53 | 64.41 | 9.41 | 0.29 | 26.47 | 4.12 | 61.18 | 7.94 |
| $\rho^{TG}$ | 0 | 17.35 | 6.47 | 69.41 | 6.76 | 0 | 16.47 | 5.88 | 70 | 7.65 | 0 | 8.24 | 4.12 | 60.88 | 26.76 |
| $\rho^{CC}$ | 11.76 | 44.12 | 4.12 | 33.24 | 6.76 | 11.76 | 44.41 | 2.35 | 34.12 | 7.35 | 15.59 | 50.29 | 4.41 | 24.12 | 5.59 |
| $\rho^{GG}$ | 11.76 | 44.12 | 3.82 | 33.24 | 7.06 | 10.88 | 51.47 | 3.53 | 29.71 | 4.41 | 10.59 | 50.59 | 4.41 | 27.94 | 6.47 |
| $\rho^{GA}$ | 0.88 | 46.47 | 7.06 | 32.65 | 12.94 | 0.88 | 41.18 | 7.06 | 37.65 | 13.24 | 0.88 | 35.59 | 8.53 | 39.41 | 15.59 |
| $\rho^{TC}$ | 1.18 | 47.06 | 6.18 | 32.65 | 12.94 | 1.47 | 51.76 | 4.12 | 29.71 | 12.94 | 4.41 | 50.29 | 3.82 | 28.24 | 13.24 |
| $\rho^{AT}$ | 0.88 | 39.71 | 6.47 | 28.82 | 24.12 | 0.88 | 39.41 | 6.47 | 29.12 | 24.12 | 1.18 | 36.76 | 7.06 | 29.71 | 25.29 |
| $\rho^{CG}$ | 24.41 | 18.53 | 6.47 | 33.24 | 17.35 | 23.53 | 19.12 | 5.59 | 34.41 | 17.35 | 25 | 17.65 | 5.59 | 33.82 | 17.94 |
| $\rho^{GC}$ | 0 | 5 | 0.29 | 35.88 | 58.82 | 0 | 5 | 0 | 30.29 | 64.71 | 0.88 | 3.82 | 0.88 | 29.12 | 65.29 |
| $\rho^{TA}$ | 69.41 | 28.53 | 0.59 | 1.47 | 0 | 69.41 | 28.53 | 0.59 | 1.47 | 0 | 72.94 | 25.29 | 1.18 | 0.59 | 0 |

ΣΗΜΕΙΩΣΕΙΣ.- Το επί τοις εκατό ποσοστό των χρωμοσωμάτων που οι δινουκλεοτιδικές σταθμισμένες συχνότητές τους λαμβάνουν τιμές εντός των καθορισμένων διαστημάτων. Σταθμισμένες συχνότητες με τιμές μικρότερες του 0.80 ή μεγαλύτερες του 1.20 υποδεικνύουν σημαντική υπο- ή υπερ-εκπροσώπηση των αντίστοιχων δινουκλεοτιδίων. Σταθμισμένες συχνότητες εντός του διαστήματος [0.99, 1.01] υποδεικνύουν ότι οι παρατηρούμενες συχνότητες εμφάνισης είναι κατά προσέγγιση ίσες με τις αναμενόμενες. Τα πρότυπα κατανομής των σταθμισμένων συχνοτήτων παρουσιάζονται ξεχωριστά για (*α*) τον δημοσιευμένο κλώνο, (*β*) τις CDS-συρραφές, και (*γ*) τον οδηγό κλώνο.

Σε συμφωνία με προηγούμενες αναφορές, τα AC και GT υποεκπροσωπούνται κατά μήκος του δημοσιευμένου κλώνου σε περισσότερα από το 94% των χρωμοσωμάτων της συλλογής ($\rho^{AC} < 1$, $\rho^{GT} < 1$). Ωστόσο, σε περισσότερα από το 68% των δημοσιευμένων κλώνων που εξετάζονται, η έλλειψη σε AC/GT που εμφανίζεται είναι μάλλον ασθενής, καθώς οι αντίστοιχες σταθμισμένες συχνότητες λαμβάνουν μεν τιμές μικρότερες της μονάδας, αλλά ίσες ή μεγαλύτερες του 0.80. Επίσης, η υπερεκπροσώπηση των AA και TT αποτελεί μια ακόμα γενική τάση των γονιδιωμάτων, όπως αυτή θεμελιώνεται σε προγενέστερες μελέτες, που ισχύει για το σύνολο σχεδόν των εξεταζόμενων περιπτώσεων. Ενώ όμως περισσότερα από το 92% των χρωμοσωμάτων της συλλογής εμφανίζουν συνολικά ένα εμπλουτισμό σε AA και TT κατά μήκος του δημοσιευμένου τους κλώνου, μόνο το ένα τρίτο έχει $\rho^{AA}$ ή $\rho^{TT}$ με τιμές που να ισούνται ή να είναι μεγαλύτερες του 1.20. Συμπερασματικά, ενώ οι σταθμισμένες συχνότητες των AC, GT, AA και TT λαμβάνουν τιμές που συνάδουν με τα πρότυπα υπο- ή υπερ-εκπροσώπησης που έχουν περιγραφεί σε προηγούμενες μελέτες, οι αποκλίσεις αυτών των δινουκλεοτιδίων από τις αναμενόμενες συχνότητες εμφάνισης είναι μάλλον μικρές, με τις αντίστοιχες τιμές των σχετικών συχνοτήτων να κυμαίνονται από 0.80 έως 1.20 στα δύο τρίτα περίπου τς συλλογής. Αξίζει να σημειωθεί ότι *οι σταθμισμένες συχνότητες των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων ακολουθούν παρόμοιες κατανομές κατά μήκος του δημοσιευμένου κλώνου*. Αυτή η τάση παρατηρείται και όταν εξετάζεται ο οδηγός κλώνος, αν και στην περίπτωση αυτή εντοπίζονται ορισμένες χαρακτηριστικές αποκλίσεις.

Όταν εξετάζονται οι CDS-συρραφές, εμφανίζονται ορισμένες σημαντικές διαφοροποιήσεις των προτιμήσεων της δινουκλεοτιδικής τους σύστασης σε σχέση με τα πρότυπα που αντιστοιχούν στον δημοσιευμένο κλώνο. Παρότι η έλλειψη AC/GT ανταποκρίνεται σε μία γενική ροπή της σύστασης του DNA, υπάρχει μια αξιοσημείωτη και συστηματική απόκλιση στον τρόπο με τον οποίο οι τιμές των $\rho^{AC}$ και $\rho^{GT}$ κατανέμονται στο σύνολο των CDS-συρραφών. Συγκεκριμένα, το GT υποεκπροσωπείται έντονα ($\rho^{GT} < 0.80$) σε διπλάσιο αριθμό CDS-συρραφών από ότι το AC ($\rho^{AC} < 0.80$). Αντίστοιχα, ενώ οι περισσότερες κωδικές περιοχές εμφανίζουν προτίμηση υπερεκπροσώπησης των AA/TT (Ohno 1988), οι CDS-συρραφές με $\rho^{AA} >= 1.2$ υπερβαίνουν κατά 1.4 φορές εκείνες με $\rho^{TT} >= 1.2$. Συνολικά, και σε αντίθεση με ό,τι παρατηρείται στον δημοσιευμένο και τον οδηγό κλώνο, στις CDS-συρραφές οι κατανομές των σταθμισμένων συχνοτήτων των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων εμφανίζουν έντονες ασυμμετρίες.

Αυτή η πόλωση των κατανομών μεταξύ αντιστρόφως συμπληρωματικών δινουκλεοτιδίων είναι ιδιαίτερα εμφανής στις περιπτώσεις των AG/CT και CA/TG ζευγών. Το 34% περίπου των CDS-συρραφών έχουν $\rho^{AG} < 0.8$, σε αντίθεση με το μόλις 2% που έχει $\rho^{CT} < 0.8$. Αντίστοιχα, η σχέση $\rho^{CA} >= 1.2$ ισχύει στο κατά προσέγγιση 8% των CDS-συρραφών, εν αντιθέσει με το ~27% που έχει $\rho^{TG} >= 1.2$. *Αυτές οι ασυμμετρίες, που εντοπίζονται στις CDS-συρραφές και συνεπώς χαρακτηρίζουν στους κωδικούς κλώνους των γονιδίων, αποτελούν παραδείγματα της ανά κλώνο εξειδικευμένης φύσης συγκεκριμένων συσχετίσεων μεταξύ των πρώτων γειτονικών βάσεων.*

Ο Ohno είχε προτείνει πως τα πρότυπα της κατανομής των δινουκλεοτιδίων, σε όρους σταθμισμένων συχνοτήτων, καθορίζονται στη βάση καθολικών κανόνων, και συγκεκριμένα σύμφωνα με την προδιάθεση του DNA σε έλλειμμα TA/CG και περίσσεια TG/CT (Ohno 1988, Yomo & Ohno 1989). Οι Shioiri και Takahata (2001) ανασκεύασαν αυτή την γενική αρχή, καθώς έδειξαν ότι ισχύει μόνο μερικώς στα αποτελέσματά τους. Πράγματι, και σε αντίθεση με τον κανόνα του Ohno, στην δική μας συλλογή, το CT υποεκπροσωπείται ($\rho^{CT} < 1$) στο δημοσιευμένο και τον οδηγό κλώνο σε περισσότερα από 72% των χρωμοσωμάτων που εξετάσαμε, καθώς επίσης και σε κατά προσέγγιση 60% των CDS-συρραφών. Τα αποτελέσματά μας διαψεύδουν επίσης την εικαζόμενη ως γενική τάση υποεκπροσώπησης του CG. Η εικόνα δεν αλλάζει ούτε όταν εστιάζουμε στις CDS-συρραφές, παρότι ήταν ειδικά για τις κωδικές περιοχές που είχε αρχικά προταθεί ο κανόνας του Ohno. Το CG είναι αισθητά υπερεκπροσωπημένο σε περισσότερο από το 17% της συλλογής μας, με $\rho^{GC} >= 1.2$, ενώ το ένα τρίτο των χρωμοσωμάτων εμφανίζει μία ήπια τάση εμπλουτισμού σε CG, με $1.01 < \rho^{GC} < 1.2$, είτε εξετάζουμε τον δημοσιευμένο ή τον οδηγό κλώνο, είτε εξετάζουμε τις CDS-συρραφές. Ο κανόνας του Ohno ισχύει πρωτίστως για το TA, αλλά επίσης και για το TG. Το TA υποεκπροσωπείται σχεδόν στο σύνολο των εξεταζόμενων γονιδιωμάτων, και αυτή η έλλειψη σε TA είναι ιδιαίτερα έντονη στο 70% περίπου των χρωμοσωμάτων, με $\rho^{TA} < 0.8$, ανεξαρτήτως του κλώνου που μελετάμε. Το TG υπερεκπροσωπείται στον δημοσιευμένο και τον οδηγό κλώνο σε περισσότερα από 76% των χρωμοσωμάτων, ενώ η τάση αυτή ισχύει για το ~88% των CDS-συρραφών, με το ~27% των περιπτώσεων να έχουν $\rho^{TG} >= 1.2$.

## **3.3** *Ειδικές ανά κλώνο ασυμμετρίες των δινουκλεοτιδικών σταθμισμένων συχνοτήτων*

Δινουκλεοτιδικές σταθμισμένες συχνότητες με τιμές που απέχουν αρκετά από τη μονάδα υποδηλώνουν σημαντικές αποκλίσεις από το τυχαίως αναμενόμενο. Στην προηγούμενη ενότητα εντοπίσαμε τέτοιες αποκλίσεις σε ιδιαίτερα μεγάλο αριθμό χρωμοσωμάτων. Επιπλέον, και πιο σημαντικό, *οι αποκλίσεις αυτές φαίνεται να εμφανίζουν χαρακτηριστικές ασυμμετρίες μεταξύ των ζευγαριών αντιστρόφως συμπληρωματικών δινουκλεοτιδίων, ιδίως στην περίπτωση που εξετάζονται οι CDS-συρραφές* (Πίνακας 3Γ). *Η προδιάθεση ασύμμετρης υπο- ή υπερ-εκπροσώπησης των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων, καθώς ανιχνεύεται σε έναν μεγάλο αριθμό χρωμοσωμάτων, πρέπει να είναι σημαντική, τόσο από στατιστικής όσο και από βιολογικής άποψης*. Ωστόσο, στον Πίνακα 3 παρουσιάζονται οι σταθμισμένες συχνότητες ενός εκάστου των δινουκλεοτιδίων ξεχωριστά. Συνεπώς, δεν προκύπτει ρητά ότι οι ασυμμετρίες μεταξύ των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων που ανιχνεύουμε στο σύνολο της συλλογής ισχύουν και για κάθε χρωμόσωμα ξεχωριστά. Προκειμένου να απαντήσουμε σε αυτό το ζήτημα, και ακολουθώντας τη μεθοδολογία που παρουσίασε η Nussinov (1984a, 1984b), μελετήσαμε τις ανά ζεύγη ιεραρχικές σχέσεις των δινουκλεοτιδίων σε κάθε ένα από τα χρωμοσώματα της συλλογής μας. Καθώς το αντικείμενο της έρευνάς μας είναι οι ειδικές ανά κλώνο ασυμμετρίες, εστιάσαμε στα ζεύγη των αντίστροφων και των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων. *Οι σταθμισμένες συχνότητες των αντίστροφων δινουκλεοτιδίων μας επιτρέπουν να αξιολογήσουμε την επίδραση που ασκεί η διάταξη των βάσεων στις προτιμήσεις υπο- ή υπερ-εκπροσώπησης κατά μήκος του κάθε κλώνου*. Αντίθετα, *οι σταθμισμένες συχνότητες των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων είναι δηλωτικές των τάσεων υπο- ή υπερ-εκπροσώπησης που οφείλονται σε διαφορές μεταξύ των δύο αντιστρόφως συμπληρωματικών κλώνων*. Στον Πίνακα 4 συνοψίζουμε τις ιεραρχικές σχέσεις των δινουκλεοτιδίων για το σύνολο της συλλογής μας, τόσο στον δημοσιευμένο και τον οδηγό κλώνο, όσο και στις CDS-συρραφές.

**ΠΙΝΑΚΑΣ 4. Ιεραρχικές σχέσεις αντίστροφων και αντιστρόφως συμπληρωματικών δινουκλεοτιδίων**

| αντίστροφα δινουκλεοτίδια | δημοσιευμένος κλώνος | οδηγός κλώνος | CDS-συρραφές | αντιστρόφως συμπληρωματικά δινουκλεοτίδια | δημοσιευμένος κλώνος | οδηγός κλώνος | CDS-συρραφές |
|---|---|---|---|---|---|---|---|
| $\rho^{CA} > \rho^{AC}$ | 96 | 94 | 91 | $\rho^{AA} > \rho^{TT}$ | 48 | 59 | 61 |
| $\rho^{TG} > \rho^{GT}$ | 95 | 97 | 100 | $\rho^{AC} > \rho^{GT}$ | 47 | 35 | 94 |
| $\rho^{GA} > \rho^{AG}$ | 70 | 74 | 86 | $\rho^{AG} > \rho^{CT}$ | 51 | 47 | 9 |
| $\rho^{TC} > \rho^{CT}$ | 70 | 68 | 48 | $\rho^{CA} > \rho^{TG}$ | 49 | 37 | 14 |
| $\rho^{AT} > \rho^{TA}$ | 96 | 96 | 97 | $\rho^{GG} > \rho^{CC}$ | 47 | 37 | 71 |
| $\rho^{GC} > \rho^{CG}$ | 79 | 79 | 78 | $\rho^{GA} > \rho^{TC}$ | 51 | 84 | 80 |

ΣΗΜΕΙΩΣΕΙΣ.- Ιεραρχική κατάταξη των δινουκλεοτιδικών προτιμήσεων στον δημοσιευμένο κλώνο, στον οδηγό κλώνο και στις CDS-συρραφές. Σε κάθε γραμμή δίνονται τα ποσοστά των αλληλουχιών DNA της συλλογής μας, στις οποίες ισχύουν οι αναγραφόμενες ανισότητες μεταξύ την δινουκλεοτιδικών σταθμισμένων συχνοτήτων. Οι ανισότητες αφορούν τα ζεύγη των αντίστροφων και των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων.

*Οι σταθμισμένες συχνότητες των αντίστροφων δινουκλεοτιδίων αποκαλύπτουν σαφείς προτιμήσεις στη διάταξη των βάσεων, που παραμένουν σταθερές στις περισσότερες αλληλουχίες DNA της συλλογής μας.* Καθώς τα αντίστροφα δινουκλεοτίδια έχουν ανά δύο την ίδια σύσταση, οι προτιμήσεις

που παρουσιάζονται στον Πίνακα 4 αφορούν την διάταξη του κάθε δεδομένου ζεύγους μονονουκλεοτιδίων. Επί παραδείγματι, στο 96% των εξεταζόμενων αλληλουχιών τα κατάλοιπα A και T απαντώνται κατά προτίμηση σε δινουκλεοτίδια ApT και όχι σε TpA, κατά μήκος του δημοσιευμένου κλώνου. Επιπλέον, *οι προτιμήσεις στο επίπεδο των αντίστροφων δινουκλεοτιδίων ακολουθούν την ίδια τάση, ανεξάρτητα από τον κλώνο τον οποίο εξετάζουμε*. Έτσι, η σχέση $\rho^{AT} > \rho^{TA}$ ισχύει επίσης για το 96% της συλλογής κατά μήκος του οδηγού κλώνου και για το 97% στις CDS-συρραφές. Αντίστοιχες σχέσεις ανισοτήτων, όπως $\rho^{CA} > \rho^{AC}$, $\rho^{TG} > \rho^{GT}$ και $\rho^{GC} > \rho^{CG}$ είναι ευρύτατα διαδεδομένες μεταξύ των βακτηρίων της συλλογής μας. Και σε αυτές τις περιπτώσεις, οι ανισότητες μεταξύ αντίστροφων δινουκλεοτιδίων δεν διακρίνουν μεταξύ διαφορετικών κλώνων, και συνεπώς δεν αποτελούν ένα ανά κλώνο εξειδικευμένο χαρακτηριστικό του DNA. Το ζευγάρι TC/CT αποτελεί την μοναδική εξαίρεση, καθώς για το 70% περίπου των βακτηρίων ισχύει η σχέση $\rho^{TC} > \rho^{CT}$ τόσο στον δημοσιευμένο όσο και στον οδηγό κλώνο, αλλά όχι στις CDS-συρραφές. Ωστόσο, στις CDS-συρραφές δεν εκδηλώνεται μια συστηματική σχέση ανισότητας μεταξύ των TpC και CpT, καθώς στο 48% ισχύει η σχέση $\rho^{TC} > \rho^{CT}$, πράγμα που σημαίνει πως στο υπόλοιπο ~52% ισχύει $\rho^{TC} < \rho^{CT}$.

Σε χαρακτηριστική αντίθεση με τις παρατηρήσεις που αφορούν στα αντίστροφα δινουκλεοτίδια, τα αντιστρόφως συμπληρωματικά δινουκλεοτίδια δεν εμφανίζουν ανά ζεύγη συστηματικές ανισότητες στον δημοσιευμένο κλώνο των χρωμοσωμάτων. Ωστόσο, όταν εστιάζουμε στον οδηγό κλώνο, εμφανίζονται σαφείς σχέσης ανισότητας ανάμεσα στα αντιστρόφως συμπληρωματικά δινουκλεοτίδια. Συγκεκριμένα, στο 84% των οδηγών κλώνων ισχύει ότι $\rho^{GA} > \rho^{TC}$, ενώ σε ένα ποσοστό που κυμαίνεται μεταξύ 63% και 65% ισχύει ότι $\rho^{AC} < \rho^{GT}$, $\rho^{CA} < \rho^{TG}$, και $\rho^{GG} < \rho^{CC}$. Η ιεραρχική κατάταξη των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων είναι ακόμα πιο έντονα πολωμένη στις CDS-συρραφές, αν και για ορισμένα δινουκλεοτιδικά ζεύγη στρέφεται προς την αντίθετη φορά σε σχέση με τα όσα παρατηρούμε στον οδηγό κλώνο. Επί παραδείγματι, περίπου οι μισές από τις εξεταζόμενες αλληλουχίες DNA έχουν $\rho^{AC} > \rho^{GT}$ στον δημοσιευμένο κλώνο. Αντιθέτως, 65% των χρωμοσωμάτων έχουν $\rho^{AC} < \rho^{GT}$ στον οδηγό τους κλώνο, ενώ 94% των CDS-συρραφών έχουν $\rho^{AC} > \rho^{GT}$. Αυτά τα πρότυπα, καθώς επαναλαμβάνονται σε έναν μεγάλο αριθμό γονιδιωμάτων, είναι εξαιρετικά απίθανο να προκύπτουν λόγω τυχαίων διακυμάνσεων της σύστασης του DNA, και συνεπώς *υποδεικνύουν την ύπαρξη ειδικών ανά κλώνο ασυμμετριών στο επίπεδο των συσχετίσεων μεταξύ γειτονικών βάσεων, οι οποίες είναι στατιστικά και*

*βιολογικά σημαντικές*. Πρόκειται για την πρώτη δημοσιοποιημένη καταγραφή αυτού του τύπου των ασυμμετριών.

## **3.4** *Δινουκλεοτιδικές ασυμμετρίες σε όρους παρατηρούμενων και σταθμισμένων συχνοτήτων*

Προγενέστερες μελέτες που εξέταζαν τις συχνότητες εμφάνισης των δινουκλεοτιδίων, αναφέρουν πως οι τιμές τους ακολουθούν παρεμφερείς κατανομές κατά μήκος των αντιστρόφως συμπληρωματικών κλώνων, στην κλίμακα ολόκληρου του χρωμοσώματος (Baisnée et al. 2002). Οι αντίστοιχες δινουκλεοτιδικές αποκλίσεις, όταν υπάρχουν, θεωρήθηκε πως είναι κατ' απόλυτη τιμή πολύ μικρές και ως εκ τούτου αμελητέες (Shioiri & Takahata). Τα συμπεράσματα αυτά συμβαδίζουν σε μεγάλο βαθμό με τα αποτελέσματα του Πίνακα 4, όπου ναι μεν γίνεται σαφές πως υπάρχουν σημαντικές ασυμμετρίες μεταξύ οδηγού και συνοδού ή κωδικού και μεταγραφόμενου κλώνου, στο επίπεδο της δινουκλεοτιδικής σύστασης, ωστόσο κατά μήκος του δημοσιευμένου κλώνου, στην κλίμακα δηλαδή ολόκληρου του χρωμοσώματος, δεν εντοπίζονται συστηματικές σχέσεις ανισότητας μεταξύ των σταθμισμένων συχνοτήτων.

Στην παρούσα ενότητα εξετάζουμε την κατανομή των ειδικών ανά κλώνο δινουκλεοτιδικών ασυμμετριών, σε όρους τόσο παρατηρούμενων όσο και σταθμισμένων συχνοτήτων. Για κάθε αλληλουχία DNA, οι σχετικοί υπολογισμοί έγιναν ($\alpha$) στο δημοσιευμένο κλώνο, ($\beta$) στον οδηγό κλώνο και ($\gamma$) στις CDS-συρραφές. Εκτός από τις αποκλίσεις των δινουκλεοτιδικών συχνοτήτων (εφεξής, *αποκλίσεις δινουκλεοτιδίων*) και των δινουκλεοτιδικών σταθμισμένων συχνοτήτων (εφεξής, *αποκλίσεις σταθμισμένων συχνοτήτων*), υπολογίσαμε επίσης και τις μονονουκλεοτιδικές αποκλίσεις, τις οποίες χρησιμοποιούμε στην ανάλυσή μας ως σημείο αναφοράς των ειδικών ανά κλώνο ασυμμετριών.

### **3.4.1** *Ένταση των ειδικών ανά κλώνο ασυμμετριών*

Όπως και στην ενότητα 3.1, όπου εξετάσαμε τα αθροιστικά γραφήματα των μονονουκλεοτιδικών αποκλίσεων, έτσι και εδώ, χωρίσαμε τη συλλογή μας σε δύο υποσύνολα, στα γονιδιώματα των Firmicutes (64 τον αριθμό) και στα γονιδιώματα όλων των υπολοίπων βακτηρίων, εκτός-Firmicutes (276 τον αριθμό). Στον Πίνακα 5 παρατίθενται τα ποσοστημόρια (quantiles) της κατανομής των απόλυτων τιμών κάθε απόκλισης στη συλλογή μας. Οι απόλυτες τιμές είναι ενδεικτικές της έντασης των ειδικών ανά κλώνο ασυμμετριών.

**ΠΙΝΑΚΑΣ 5. Ποσοστημόρια των απόλυτων τιμών των αποκλίσεων**

**Α. δημοσιευμένος κλώνος**

**εκτός-Firmicutes**

| | | 0% | 10% | 20% | 40% | 50% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|---|---|---|
| αποκλίσεις μονο- και δι- νουκλεοτιδίων | $S^{A-T}$ | 4.7e-06 | 5e-04 | 0.00089 | 0.0018 | 0.0023 | 0.003 | 0.005 | 0.047 |
| | $S^{G-C}$ | 7e-05 | 0.00046 | 0.00092 | 0.002 | 0.0025 | 0.0033 | 0.0064 | 0.053 |
| | $S^{AG-CT}$ | 3.8e-05 | 0.00081 | 0.0014 | 0.0033 | 0.0042 | 0.0053 | 0.0084 | 0.057 |
| | $S^{GA-TC}$ | 6.8e-06 | 0.00061 | 0.0013 | 0.0029 | 0.0039 | 0.0049 | 0.0084 | 0.044 |
| | $S^{GG-CC}$ | 2.8e-05 | 0.0011 | 0.0018 | 0.0046 | 0.0058 | 0.007 | 0.014 | 0.088 |
| | $S^{AA-TT}$ | 8.2e-06 | 0.00079 | 0.002 | 0.0034 | 0.0049 | 0.0064 | 0.011 | 0.088 |
| | $S^{AC-GT}$ | 3.9e-05 | 9e-04 | 0.0016 | 0.0037 | 0.005 | 0.0065 | 0.013 | 0.11 |
| | $S^{CA-TG}$ | 5e-05 | 0.00085 | 0.0016 | 0.0034 | 0.0045 | 0.0058 | 0.0094 | 0.095 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | 1.5e-05 | 0.00091 | 0.0019 | 0.0037 | 0.0049 | 0.0059 | 0.0095 | 0.046 |
| | $P^{GA-TC}$ | 4.4e-05 | 0.00071 | 0.0015 | 0.0028 | 0.0038 | 0.0047 | 0.0083 | 0.049 |
| | $P^{GG-CC}$ | 3.5e-05 | 0.00049 | 0.00095 | 0.0023 | 0.0031 | 0.0043 | 0.0078 | 0.068 |
| | $P^{AA-TT}$ | 8.2e-06 | 0.00084 | 0.0017 | 0.0033 | 0.0043 | 0.005 | 0.0085 | 0.039 |
| | $P^{AC-GT}$ | 2e-05 | 0.00079 | 0.0016 | 0.003 | 0.0041 | 0.0056 | 0.0082 | 0.034 |
| | $P^{CA-TG}$ | 2.1e-05 | 0.00076 | 0.0018 | 0.0039 | 0.0053 | 0.0067 | 0.01 | 0.036 |

**Firmicutes**

| | | 0% | 10% | 20% | 40% | 50% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|---|---|---|
| αποκλίσεις μονο- και δι- νουκλεοτιδίων | $S^{A-T}$ | 0.00029 | 0.00093 | 0.0013 | 0.0021 | 0.0024 | 0.0037 | 0.0056 | 0.021 |
| | $S^{G-C}$ | 0.00023 | 0.00083 | 0.0016 | 0.003 | 0.0042 | 0.0049 | 0.0076 | 0.039 |
| | $S^{AG-CT}$ | 1.5e-05 | 0.00047 | 0.0012 | 0.0049 | 0.0065 | 0.0079 | 0.012 | 0.048 |
| | $S^{GA-TC}$ | 5.7e-05 | 0.0018 | 0.0026 | 0.0054 | 0.0067 | 0.0087 | 0.014 | 0.06 |
| | $S^{GG-CC}$ | 0.00032 | 0.0015 | 0.0042 | 0.0069 | 0.0084 | 0.01 | 0.017 | 0.063 |
| | $S^{AA-TT}$ | 3.2e-05 | 0.0011 | 0.0019 | 0.0036 | 0.0046 | 0.0069 | 0.01 | 0.035 |
| | $S^{AC-GT}$ | 0.00031 | 0.00089 | 0.0018 | 0.0033 | 0.0036 | 0.0041 | 0.0064 | 0.03 |
| | $S^{CA-TG}$ | 5.2e-05 | 0.00071 | 0.0014 | 0.0038 | 0.0058 | 0.0074 | 0.0094 | 0.045 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | 0.00016 | 0.00071 | 0.0014 | 0.0033 | 0.0037 | 0.0047 | 0.0085 | 0.024 |
| | $P^{GA-TC}$ | 0.00016 | 0.00071 | 0.0014 | 0.0032 | 0.0041 | 0.0055 | 0.009 | 0.022 |
| | $P^{GG-CC}$ | 0.00016 | 0.00096 | 0.0025 | 0.005 | 0.0073 | 0.0088 | 0.013 | 0.059 |
| | $P^{AA-TT}$ | 1.4e-06 | 0.00058 | 0.00096 | 0.0021 | 0.0026 | 0.0036 | 0.0058 | 0.015 |
| | $P^{AC-GT}$ | 3.4e-05 | 0.00028 | 0.00094 | 0.0018 | 0.0027 | 0.0039 | 0.0064 | 0.018 |
| | $P^{CA-TG}$ | 5.4e-05 | 0.00059 | 0.0014 | 0.005 | 0.0061 | 0.0073 | 0.0097 | 0.029 |

**Β. οδηγός κλώνος**

**εκτός-Firmicutes**

| | | 0% | 10% | 20% | 40% | 50% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|---|---|---|
| αποκλίσεις μονο- και δι- νουκλεοτιδίων | $S^{A-T}$ | 3.6e-05 | 0.0023 | 0.0059 | 0.01 | 0.014 | 0.018 | 0.029 | 0.11 |
| | $S^{G-C}$ | 0.00025 | 0.01 | 0.019 | 0.03 | 0.035 | 0.041 | 0.062 | 0.22 |
| | $S^{AG-CT}$ | 0.00028 | 0.0039 | 0.0079 | 0.015 | 0.019 | 0.027 | 0.056 | 0.22 |
| | $S^{GA-TC}$ | 0.00014 | 0.0057 | 0.011 | 0.022 | 0.03 | 0.039 | 0.063 | 0.27 |
| | $S^{GG-CC}$ | 0.00018 | 0.024 | 0.041 | 0.061 | 0.071 | 0.084 | 0.11 | 0.33 |
| | $S^{AA-TT}$ | 3.9e-05 | 0.0038 | 0.0072 | 0.016 | 0.023 | 0.029 | 0.06 | 0.21 |
| | $S^{AC-GT}$ | 0.00018 | 0.014 | 0.029 | 0.046 | 0.055 | 0.066 | 0.1 | 0.34 |
| | $S^{CA-TG}$ | 0.00015 | 0.017 | 0.03 | 0.044 | 0.052 | 0.062 | 0.092 | 0.3 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | 0.00029 | 0.0021 | 0.0042 | 0.012 | 0.016 | 0.019 | 0.028 | 0.13 |
| | $P^{GA-TC}$ | 6.3e-05 | 0.0031 | 0.0059 | 0.014 | 0.017 | 0.02 | 0.03 | 0.1 |
| | $P^{GG-CC}$ | 0.00014 | 0.0022 | 0.0052 | 0.0094 | 0.013 | 0.019 | 0.042 | 0.26 |
| | $P^{AA-TT}$ | 0.00049 | 0.0047 | 0.0084 | 0.014 | 0.02 | 0.025 | 0.036 | 0.09 |
| | $P^{AC-GT}$ | 1.8e-05 | 0.0024 | 0.0045 | 0.011 | 0.017 | 0.022 | 0.034 | 0.13 |
| | $P^{CA-TG}$ | 0.00013 | 0.0027 | 0.005 | 0.012 | 0.016 | 0.021 | 0.034 | 0.12 |

**Firmicutes**

| | | 0% | 10% | 20% | 40% | 50% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|---|---|---|
| αποκλίσεις μονο- και δι- νουκλεοτιδίων | $S^{A-T}$ | 0.0012 | 0.012 | 0.018 | 0.03 | 0.037 | 0.047 | 0.055 | 0.11 |
| | $S^{G-C}$ | 0.04 | 0.077 | 0.086 | 0.095 | 0.099 | 0.11 | 0.13 | 0.22 |
| | $S^{AG-CT}$ | 0.0034 | 0.078 | 0.096 | 0.11 | 0.13 | 0.15 | 0.22 | 0.3 |
| | $S^{GA-TC}$ | 0.063 | 0.12 | 0.13 | 0.14 | 0.15 | 0.19 | 0.21 | 0.36 |
| | $S^{GG-CC}$ | 0.058 | 0.15 | 0.15 | 0.18 | 0.2 | 0.21 | 0.24 | 0.36 |
| | $S^{AA-TT}$ | 0.0048 | 0.017 | 0.025 | 0.04 | 0.06 | 0.068 | 0.087 | 0.19 |
| | $S^{AC-GT}$ | 0.002 | 0.009 | 0.039 | 0.057 | 0.059 | 0.061 | 0.092 | 0.12 |
| | $S^{CA-TG}$ | 0.034 | 0.047 | 0.051 | 0.066 | 0.068 | 0.074 | 0.096 | 0.19 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | 0.0036 | 0.0064 | 0.0089 | 0.031 | 0.04 | 0.057 | 0.075 | 0.1 |
| | $P^{GA-TC}$ | 0.0017 | 0.02 | 0.034 | 0.049 | 0.055 | 0.057 | 0.074 | 0.17 |
| | $P^{GG-CC}$ | 8e-06 | 0.0013 | 0.0081 | 0.025 | 0.034 | 0.04 | 0.051 | 0.24 |
| | $P^{AA-TT}$ | 0.00063 | 0.018 | 0.026 | 0.036 | 0.041 | 0.043 | 0.062 | 0.14 |
| | $P^{AC-GT}$ | 2e-04 | 0.0073 | 0.012 | 0.015 | 0.017 | 0.019 | 0.037 | 0.11 |
| | $P^{CA-TG}$ | 0.00048 | 0.003 | 0.01 | 0.033 | 0.035 | 0.036 | 0.038 | 0.13 |

**Γ. CDS-συρραφές**

**εκτός-Firmicutes**

| | | 0% | 10% | 20% | 40% | 50% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|---|---|---|
| αποκλίσεις μονο- και δι- νουκλεοτιδίων | $S^{A-T}$ | 0.00014 | 0.004 | 0.0073 | 0.016 | 0.023 | 0.028 | 0.048 | 0.12 |
| | $S^{G-C}$ | 5.1e-05 | 0.0065 | 0.011 | 0.023 | 0.037 | 0.05 | 0.078 | 0.19 |
| | $S^{AG-CT}$ | 1.4e-05 | 0.021 | 0.044 | 0.074 | 0.09 | 0.1 | 0.14 | 0.25 |
| | $S^{GA-TC}$ | 0.00013 | 0.0083 | 0.016 | 0.055 | 0.078 | 0.093 | 0.14 | 0.34 |
| | $S^{GG-CC}$ | 0.00039 | 0.018 | 0.032 | 0.064 | 0.086 | 0.11 | 0.16 | 0.4 |
| | $S^{AA-TT}$ | 4.7e-06 | 0.014 | 0.031 | 0.057 | 0.066 | 0.077 | 0.12 | 0.2 |
| | $S^{AC-GT}$ | 0.00059 | 0.0071 | 0.019 | 0.036 | 0.044 | 0.051 | 0.078 | 0.21 |
| | $S^{CA-TG}$ | 0.00071 | 0.017 | 0.038 | 0.058 | 0.068 | 0.074 | 0.11 | 0.21 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | 0.0027 | 0.034 | 0.064 | 0.11 | 0.13 | 0.15 | 0.2 | 0.34 |
| | $P^{GA-TC}$ | 4e-05 | 0.011 | 0.02 | 0.048 | 0.06 | 0.068 | 0.091 | 0.19 |
| | $P^{GG-CC}$ | 0.00015 | 0.0078 | 0.017 | 0.031 | 0.042 | 0.05 | 0.075 | 0.24 |
| | $P^{AA-TT}$ | 0.00017 | 0.012 | 0.023 | 0.054 | 0.077 | 0.1 | 0.16 | 0.29 |
| | $P^{AC-GT}$ | 7e-04 | 0.018 | 0.029 | 0.051 | 0.062 | 0.076 | 0.11 | 0.26 |
| | $P^{CA-TG}$ | 0.0046 | 0.023 | 0.036 | 0.065 | 0.084 | 0.098 | 0.16 | 0.27 |

**Firmicutes**

| | | 0% | 10% | 20% | 40% | 50% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|---|---|---|
| αποκλίσεις μονο- και δι- νουκλεοτιδίων | $S^{A-T}$ | 0.00041 | 0.021 | 0.029 | 0.043 | 0.072 | 0.078 | 0.085 | 0.15 |
| | $S^{G-C}$ | 0.016 | 0.071 | 0.072 | 0.089 | 0.1 | 0.12 | 0.15 | 0.28 |
| | $S^{AG-CT}$ | 0.0022 | 0.038 | 0.061 | 0.096 | 0.13 | 0.19 | 0.26 | 0.33 |
| | $S^{GA-TC}$ | 0.071 | 0.11 | 0.14 | 0.15 | 0.21 | 0.24 | 0.26 | 0.45 |
| | $S^{GG-CC}$ | 0.029 | 0.14 | 0.15 | 0.2 | 0.23 | 0.27 | 0.3 | 0.46 |
| | $S^{AA-TT}$ | 0.00099 | 0.04 | 0.048 | 0.063 | 0.11 | 0.12 | 0.13 | 0.26 |
| | $S^{AC-GT}$ | 0.00016 | 0.0018 | 0.0031 | 0.019 | 0.029 | 0.039 | 0.054 | 0.2 |
| | $S^{CA-TG}$ | 0.0064 | 0.015 | 0.02 | 0.066 | 0.067 | 0.069 | 0.11 | 0.2 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | 0.00079 | 0.026 | 0.045 | 0.078 | 0.083 | 0.089 | 0.12 | 0.25 |
| | $P^{GA-TC}$ | 0.0093 | 0.028 | 0.035 | 0.058 | 0.061 | 0.077 | 0.09 | 0.21 |
| | $P^{GG-CC}$ | 0.00089 | 0.0065 | 0.01 | 0.058 | 0.07 | 0.08 | 0.088 | 0.27 |
| | $P^{AA-TT}$ | 0.0034 | 0.024 | 0.033 | 0.053 | 0.073 | 0.077 | 0.088 | 0.17 |
| | $P^{AC-GT}$ | 0.032 | 0.051 | 0.053 | 0.06 | 0.07 | 0.076 | 0.09 | 0.22 |
| | $P^{CA-TG}$ | 0.005 | 0.015 | 0.031 | 0.053 | 0.056 | 0.065 | 0.085 | 0.19 |

ΣΗΜΕΙΩΣΕΙΣ.- Τα ποσοστημόρια των απόλυτων τιμών των αποκλίσεων, για κάθε μία από τις δοσμένες πιθανότητες. Για κάθε απόκλιση, η μικρότερη τιμή αντιστοιχεί σε πιθανότητα που ισούται με το 0% και η μεγαλύτερη τιμή σε πιθανότητα που ισούται με το 100%. *Οι απόλυτες τιμές των αποκλίσεων είναι το μέτρο της έντασης των ειδικών ανά κλώνο ασυμμετριών*. Καταγράφουν, δηλαδή, πόσο έντονα αποκλίνουν τα αντιστρόφως συμπληρωματικά στοιχεία της σύστασης του DNA από τις σχέσεις ενδοκλωνικής ισοδυναμίας (intra-strand compositional parities). Οι υπολογισμοί αφορούν Α: τον δημοσιευμένο κλώνο, Β: τον οδηγό κλώνο και Γ: τις CDS-συρραφές. Σε κάθε μία από αυτές τις περιπτώσεις, η συλλογή των χρωμοσωμάτων χωρίστηκε σε δύο υποσύνολα, όπου το ένα αντιστοιχεί σε όλα τα εξεταζόμενα φύλα εκτός των Firmicutes και το άλλο αντιστοιχεί μόνο στα Firmicutes.

Όταν εξαιρούμε από την ανάλυσή μας τα Firmicutes, το 60% των DNA αλληλουχιών εμφανίζει πολύ ασθενείς αποκλίσεις δινουκλεοτιδίων των οποίων οι απόλυτες τιμές δεν υπερβαίνουν το 0.7% κατά μήκος του δημοσιευμένου κλώνου (plus strand), ενώ έως και το 80% αυτών των αλληλουχιών έχει $|S_{plus}^{AG-CT}|$, $|S_{plus}^{GA-TC}|$ και $|S_{plus}^{CA-TG}|$ μικρότερες του 1% (Πίνακας 5Α). Όσον αφορά τις αποκλίσεις των σταθμισμένων συχνοτήτων, η έντασή τους ακολουθεί πρότυπα αντίστοιχα με εκείνα των δινουκλεοτιδικών αποκλίσεων, με το 60% του εκτός-Firmicutes υποσυνόλου να έχει απόλυτες τιμές αποκλίσεων που δεν υπερβαίνουν το 0.67% στο δημοσιευμένο κλώνο. Οι παρατηρήσεις αυτές θα μπορούσαν να οδηγήσουν στο λανθασμένο συμπέρασμα πως στα βακτηριακά χρωμοσώματα δεν εκδηλώνονται σημαντικές ασυμμετρίες στο επίπεδο των δινουκλεοτιδίων και των σταθμισμένων συχνοτήτων τους. Ωστόσο, και ούτως εχόντων των πραγμάτων, η ένταση των μονονουκλεοτιδικών αποκλίσεων είναι ακόμα μικρότερη, με το 80% των εκτός-Firmicutes να παρουσιάζει $|S_{plus}^{A-T}|$ και $|S_{plus}^{G-T}|$ έως 0.50% και 0.64%, αντίστοιχα, παρότι η ύπαρξη τέτοιων αποκλίσεων είναι καλά τεκμηριωμένη. Στα Firmicutes, οι αποκλίσεις των δινουκλεοτιδίων και των σταθμισμένων συχνοτήτων τους έχουν απόλυτες τιμές που είναι συγκρίσιμες με τις αντίστοιχες των εκτός-Firmicutes.

Όταν οι αποκλίσεις μετρώνται στον οδηγό κλώνο (leading strand), οι απόλυτες τιμές τους είναι εν γένει υψηλότερες από τις αντίστοιχες στον δημοσιευμένο κλώνο κατά μία τάξη μεγέθους. Στο εκτός-Firmicutes υποσύνολο, η διάμεσος των απόλυτων τιμών (το 50° ποσοστημόριο) των αποκλίσεων των δινουκλεοτιδίων κυμαίνεται μεταξύ 1.90%, για τις $|S_{leading}^{AG-CT}|$, και 7.10%, για τις $|S_{leading}^{GG-CC}|$, ενώ η διάμεσος των $|S_{leading}^{A-T}|$ είναι 1.40% και η διάμεσος των $|S_{leading}^{G-C}|$ 3.50% (Πίνακας 5Β). Οι αντίστοιχες τιμές είναι ακόμα μεγαλύτερες

όταν εξετάζουμε αποκλειστικά τα Firmicutes. Επί παραδείγματι, αν εξετάσουμε το ζεύγος των ομο-δινουκλεοτιδίων GG/CC, η διάμεσος των $|S_{leading}^{GG-CC}|$ στον οδηγό κλώνο ισούται με 20%, τιμή που είναι κατά προσέγγιση 24 φορές η διάμεσος των $|S_{plus}^{GG-CC}|$ στον δημοσιευμένο κλώνο. Για να αποκτήσει κάποιος την αίσθηση της έντασης αυτών των αποκλίσεων, οι αντίστοιχες αποκλίσεις των μονονουκλεοτιδικών συστατικών του συγκεκριμένου ζεύγους, δηλαδή οι $S_{leading}^{G-C}$, λαμβάνουν απόλυτες τιμές των οποίων η διάμεσος ισούται με 9.90%. Παρομοίως, η ένταση των αποκλίσεων των σταθμισμένων συχνοτήτων είναι σημαντικά μεγαλύτερη στον οδηγό από ότι στον δημοσιευμένο κλώνο, καταδεικνύοντας την ύπαρξη ειδικών ανά κλώνο ασυμμετριών στο επίπεδο των συσχετίσεων μεταξύ κοντινότερων γειτονικών βάσεων. Επί παραδείγματι, η διάμεσος των $|P_{leading}^{GA-TC}|$ είναι 1.7% στο εκτός-Firmicutes υποσύνολο και 5.5% στα Firmicutes (Πίνακας 5Β).

*Διαπιστώνουμε λοιπόν ότι ενώ στον δημοσιευμένο κλώνο οι συνολικές αποκλίσεις είναι ισχνές, στον οδηγό η έντασή τους είναι, συγκριτικά, πολλαπλάσια.* Οι παρατηρήσεις αυτές έρχονται σε συμφωνία με το ευρέως γνωστό γεγονός ότι σε πλήθος βακτηριακών χρωμοσωμάτων, αν και εμφανίζονται ισχυρές αποκλίσεις σε κάθε ήμισυ του δημοσιευμένου κλώνου τους, όπου το ένα αντιστοιχεί στον οδηγό και το άλλο στο συνοδό κλώνο, οι αποκλίσεις των κλώνων τις αντιγραφής παίρνουν αντίθετες τιμές και στο επίπεδο ολόκληρου του χρωμοσώματος αλληλοαναιρούνται (McLean et al. 1998). Οι παρατηρήσεις μας δείχνουν ότι *οι ασυμμετρίες οδηγού και συνοδού κλώνου συγκροτούν ένα χαρακτηριστικό μοτίβο κατά μήκος των χρωμοσωμάτων, το οποίο ακολουθούν όχι μόνο οι αποκλίσεις των μονονουκλεοτιδίων, στις οποίες αυτό είχε αρχικά περιγραφεί, αλλά και οι δινουκλεοτιδικές αποκλίσεις, σε όρους τόσο παρατηρούμενων όσο και σταθμισμένων συχνοτήτων*. Αυτός είναι και ο λόγος που, μεταξύ των βακτηρίων που εξετάζουμε, η ένταση κάθε δεδομένης δινουκλεοτιδικής απόκλισης ακολουθεί μία κατανομή μετατοπισμένη προς μεγαλύτερες τιμές, όταν οι σχετικοί υπολογισμοί γίνονται στον οδηγό αντί του δημοσιευμένου κλώνου (πρβ. Πίνακα 5Α & Β).

Στις CDS-συρραφές οι ενδοκλωνικες αποκλίσεις είναι ακόμα πιο ισχυρές από ότι στον οδηγό κλώνο. Στο εκτός-Firmicutes υποσύνολο η διάμεσος των $|S^{AG-CT}|$ στις CDS-συρραφές είναι 4.7 φορές η αντίστοιχη διάμεσος στον οδηγό κλώνο (διάμεσος $|S_{CDS}^{AG-CT}|$=9%, διάμεσος $|S_{leading}^{AG-CT}|$=1.9%). Για το ίδιο ζευγάρι δινουκλεοτιδίων, AG/CT, η διάμεσος των αποκλίσεων των σταθμισμένων συχνοτήτων είναι 6.8 φορές μεγαλύτερη στις CDS-συρραφές από ότι στον οδηγό

κλώνο (διάμεσος $|P_{CDS}^{AG-CT}|$=13%, διάμεσος $|P_{leading}^{AG-CT}|$=1.6%). Να σημειωθεί ότι, στο εκτός-Firmicutes υποσύνολο, όλα τα ζεύγη των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων, πλην των GA/TC και GG/CC, εμφανίζουν αποκλίσεις σταθμισμένων συχνοτήτων των οποίων οι απόλυτες τιμές έχουν διάμεσο μεγαλύτερη από εκείνη των αντίστοιχων αποκλίσεων στο επίπεδο των παρατηρούμενων συχνοτήτων. Συνεπώς, *οι συσχετίσεις μεταξύ των 1ης τάξης γειτονικών βάσεων παρουσιάζουν έντονες ασυμμετρίες μεταξύ κωδικών και μεταγραφόμενων κλώνων, οι οποίες σε πολλές περιπτώσεις είναι εντονότερες και από τις ειδικές ανά κλώνο πολώσεις της κατανομής των δινουκλεοτιδίων, σε όρους συχνοτήτων εμφάνισης*. Ισχυρές αποκλίσεις κατά μήκος των CDS-συρραφών παρατηρούνται όχι μόνο στο εκτός-Firmicutes υποσύνολο, αλλά και στα Firmicutes. Ωστόσο, και δεδομένου ότι οι αποκλίσεις του οδηγού κλώνου είναι ήδη ισχυρές στα Firmicutes, η ένταση των αποκλίσεων στο φύλο αυτό δεν διαφέρει σημαντικά μεταξύ των CDS-συρραφών και του οδηγού κλώνου, σε αντίθεση με ό,τι παρατηρείται στο εκτός-Firmicutes υποσύνολο.

Τα ποσοστημόρια του Πίνακα 5 περιγράφουν το εύρος της έντασης των αποκλίσεων στις αλληλουχίες DNA της συλλογής μας και προσφέρουν μια εικόνα της παρουσίας και του μεγέθους των ασυμμετριών μεταξύ των αντιστρόφως συμπληρωματικών κλώνων. Έτσι, το 0% ποσοστημόριο αντιστοιχεί στην αλληλουχία με την πιο ασθενή απόκλιση από την συμμετρία και το 100% ποσοστημόριο σε εκείνη με την πιο έντονη ασυμμετρία, ενώ αλληλουχίες των οποίων οι απόλυτες τιμές των αποκλίσεων είναι της ίδιας τάξης μεγέθους αντιστοιχούν σε κοντινά ποσοστημόρια. Για παράδειγμα, δύο χρωμοσώματα των εκτός-Firmicutes που έχουν $P_{plus}^{CA-TG}$ ίση με -3.6% και 3.2%, κατατάσσονται μεταξύ του 80% και 100% ποσοστημορίου της κατανομής των απόλυτων τιμών των αποκλίσεων (Πίνακας 5A), αφού εδώ κριτήριο είναι η έντασή τους.

### **3.4.2** *Φορά των ειδικών ανά κλώνο ασυμμετριών*

Προκειμένου να εξετάσουμε προς ποιόν κλώνο κινείται η φορά των ασυμμετριών, παραθέτουμε στον Πίνακα 6 τα ποσοστημόρια των τιμών των αποκλίσεων, όπως αυτές υπολογίστηκαν (α) στο δημοσιευμένο κλώνο, (β) στον οδηγό κλώνο και (γ) στις CDS-συρραφές. Κατ' αναλογία με τον Πίνακα 5, χωρίσαμε τη συλλογή μας σε δύο υποσύνολα, στα γονιδιώματα των Firmicutes (συνολικά 64) και στα γονιδιώματα όλων των υπολοίπων βακτηρίων, εκτός-Firmicutes (συνολικά 276).

**ΠΙΝΑΚΑΣ 6. Ποσοστημόρια των αποκλίσεων**

**Α. δημοσιευμένος κλώνος**

**εκτός-Firmicutes**

| | | 0% | 10% | 20% | 40% | 50% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|---|---|---|
| αποκλίσεις μονο- και δι- νουκλεοτιδίων | $S^{A-T}$ | -0.047 | -0.0048 | -0.003 | -0.00075 | 0.00025 | 0.0011 | 0.003 | 0.041 |
| | $S^{G-C}$ | -0.034 | -0.0058 | -0.0033 | -0.00088 | 0.00015 | 0.00092 | 0.0034 | 0.053 |
| | $S^{AG-CT}$ | -0.025 | -0.0073 | -0.0044 | -0.00087 | 0.00069 | 0.0024 | 0.0061 | 0.057 |
| | $S^{GA-TC}$ | -0.043 | -0.008 | -0.0047 | -0.0011 | 3.8e-05 | 0.0016 | 0.0051 | 0.044 |
| | $S^{GG-CC}$ | -0.07 | -0.015 | -0.0069 | -0.0015 | 0.00027 | 0.002 | 0.0079 | 0.088 |
| | $S^{AA-TT}$ | -0.088 | -0.01 | -0.0065 | -0.0021 | 0.00012 | 0.0019 | 0.0064 | 0.086 |
| | $S^{AC-GT}$ | -0.11 | -0.013 | -0.0055 | -0.0016 | 0.00022 | 0.0013 | 0.0067 | 0.082 |
| | $S^{CA-TG}$ | -0.095 | -0.0088 | -0.0062 | -0.0012 | 0.00027 | 0.0018 | 0.0055 | 0.059 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | -0.021 | -0.008 | -0.0054 | -0.0014 | 0.00048 | 0.0025 | 0.0065 | 0.046 |
| | $P^{GA-TC}$ | -0.031 | -0.0083 | -0.0047 | -0.0015 | 0.00017 | 0.0014 | 0.0048 | 0.049 |
| | $P^{GG-CC}$ | -0.068 | -0.0077 | -0.0038 | -0.00096 | -9.5e-05 | 0.00087 | 0.0046 | 0.044 |
| | $P^{AA-TT}$ | -0.039 | -0.0091 | -0.0062 | -0.0025 | -0.00084 | 0.00076 | 0.0045 | 0.028 |
| | $P^{AC-GT}$ | -0.034 | -0.0083 | -0.0048 | -0.0014 | -3.8e-05 | 0.0017 | 0.0058 | 0.023 |
| | $P^{CA-TG}$ | -0.036 | -0.01 | -0.005 | -0.00065 | 0.00076 | 0.0032 | 0.0078 | 0.032 |

**Firmicutes**

| | | 0% | 10% | 20% | 40% | 50% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|---|---|---|
| αποκλίσεις μονο- και δι- νουκλεοτιδίων | $S^{A-T}$ | -0.021 | -0.0071 | -0.0043 | -0.0019 | -0.0012 | 0.00075 | 0.0027 | 0.0078 |
| | $S^{G-C}$ | -0.039 | -0.0072 | -0.0045 | -0.0011 | -0.00056 | 0.0016 | 0.0052 | 0.016 |
| | $S^{AG-CT}$ | -0.048 | -0.013 | -0.0085 | -0.0042 | -0.00084 | -9.4e-05 | 0.0064 | 0.017 |
| | $S^{GA-TC}$ | -0.06 | -0.013 | -0.0077 | -0.0033 | -0.0023 | 0.00079 | 0.0099 | 0.025 |
| | $S^{GG-CC}$ | -0.063 | -0.016 | -0.011 | -0.0048 | -0.0018 | 0.0015 | 0.0084 | 0.029 |
| | $S^{AA-TT}$ | -0.035 | -0.01 | -0.0072 | -0.0031 | -0.001 | 0.00095 | 0.0053 | 0.017 |
| | $S^{AC-GT}$ | -0.015 | -0.0066 | -0.0053 | -0.0032 | -0.0023 | -0.00089 | 0.0027 | 0.03 |
| | $S^{CA-TG}$ | -0.017 | -0.0098 | -0.009 | -0.0054 | -0.0011 | 5.4e-05 | 0.0033 | 0.045 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | -0.02 | -0.01 | -0.0068 | -0.0033 | -0.0011 | 0.00016 | 0.0031 | 0.024 |
| | $P^{GA-TC}$ | -0.022 | -0.0087 | -0.0044 | -0.0011 | 0.00042 | 0.0025 | 0.0064 | 0.015 |
| | $P^{GG-CC}$ | -0.02 | -0.014 | -0.01 | -0.0057 | -0.0037 | -0.00076 | 0.0038 | 0.059 |
| | $P^{AA-TT}$ | -0.0086 | -0.003 | -0.0018 | 0.00065 | 0.0011 | 0.0021 | 0.0054 | 0.015 |
| | $P^{AC-GT}$ | -0.01 | -0.0072 | -0.0047 | -0.0017 | -0.00098 | -9e-05 | 0.0018 | 0.018 |
| | $P^{CA-TG}$ | -0.02 | -0.011 | -0.0092 | -0.0058 | -0.0034 | -0.00051 | 0.0029 | 0.029 |

**Β. οδηγός κλώνος**

**εκτός-Firmicutes**

| | | 0% | 10% | 20% | 40% | 50% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|---|---|---|
| αποκλίσεις μονο- και δι- νουκλεοτιδίων | $S^{A-T}$ | -0.11 | -0.047 | -0.027 | -0.015 | -0.012 | -0.0086 | -0.0014 | 0.098 |
| | $S^{G-C}$ | -0.031 | 0.0091 | 0.019 | 0.03 | 0.035 | 0.041 | 0.062 | 0.22 |
| | $S^{AG-CT}$ | -0.036 | -0.005 | 0.0021 | 0.014 | 0.018 | 0.027 | 0.056 | 0.22 |
| | $S^{GA-TC}$ | -0.038 | 9e-04 | 0.0099 | 0.021 | 0.03 | 0.039 | 0.063 | 0.27 |
| | $S^{GG-CC}$ | -0.066 | 0.018 | 0.04 | 0.061 | 0.071 | 0.084 | 0.11 | 0.33 |
| | $S^{AA-TT}$ | -0.21 | -0.08 | -0.05 | -0.026 | -0.016 | -0.011 | 0.0024 | 0.16 |
| | $S^{AC-GT}$ | -0.34 | -0.14 | -0.1 | -0.066 | -0.055 | -0.046 | -0.028 | 0.038 |
| | $S^{CA-TG}$ | -0.3 | -0.12 | -0.092 | -0.062 | -0.052 | -0.044 | -0.03 | 0.043 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | -0.13 | -0.025 | -0.019 | -0.0042 | -0.00073 | 0.0041 | 0.02 | 0.12 |
| | $P^{GA-TC}$ | -0.029 | -0.006 | 0.00079 | 0.011 | 0.016 | 0.02 | 0.03 | 0.1 |
| | $P^{GG-CC}$ | -0.26 | -0.065 | -0.036 | -0.011 | -0.0047 | 0.00017 | 0.0085 | 0.072 |
| | $P^{AA-TT}$ | -0.09 | -0.02 | -0.0075 | 0.0089 | 0.013 | 0.019 | 0.031 | 0.085 |
| | $P^{AC-GT}$ | -0.13 | -0.047 | -0.033 | -0.019 | -0.012 | -0.0075 | 6e-04 | 0.088 |
| | $P^{CA-TG}$ | -0.084 | -0.04 | -0.027 | -0.012 | -0.0079 | -0.003 | 0.011 | 0.12 |

**Firmicutes**

| | | 0% | 10% | 20% | 40% | 50% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|---|---|---|
| αποκλίσεις μονο- και δι- νουκλεοτιδίων | $S^{A-T}$ | -0.0075 | 0.012 | 0.018 | 0.03 | 0.037 | 0.047 | 0.055 | 0.11 |
| | $S^{G-C}$ | 0.04 | 0.077 | 0.086 | 0.095 | 0.099 | 0.11 | 0.13 | 0.22 |
| | $S^{AG-CT}$ | 0.0034 | 0.078 | 0.096 | 0.11 | 0.13 | 0.15 | 0.22 | 0.3 |
| | $S^{GA-TC}$ | 0.063 | 0.12 | 0.13 | 0.14 | 0.15 | 0.19 | 0.21 | 0.36 |
| | $S^{GG-CC}$ | 0.058 | 0.15 | 0.15 | 0.18 | 0.2 | 0.21 | 0.24 | 0.36 |
| | $S^{AA-TT}$ | -0.017 | 0.016 | 0.025 | 0.04 | 0.06 | 0.068 | 0.087 | 0.19 |
| | $S^{AC-GT}$ | -0.12 | -0.12 | -0.092 | -0.061 | -0.059 | -0.057 | -0.039 | 0.059 |
| | $S^{CA-TG}$ | -0.19 | -0.11 | -0.096 | -0.074 | -0.068 | -0.066 | -0.051 | -0.034 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | -0.1 | -0.06 | -0.041 | -0.0077 | -0.0042 | 0.0096 | 0.074 | 0.092 |
| | $P^{GA-TC}$ | -0.0019 | 0.02 | 0.034 | 0.049 | 0.055 | 0.057 | 0.074 | 0.17 |
| | $P^{GG-CC}$ | -0.24 | -0.064 | -0.051 | -0.038 | -0.027 | -0.016 | 0.00086 | 0.04 |
| | $P^{AA-TT}$ | -0.14 | -0.072 | -0.062 | -0.043 | -0.041 | -0.036 | -0.025 | 0.031 |
| | $P^{AC-GT}$ | -0.015 | -0.0091 | 0.0066 | 0.015 | 0.017 | 0.019 | 0.037 | 0.11 |
| | $P^{CA-TG}$ | -0.13 | -0.041 | -0.036 | -0.0047 | -0.00052 | 0.018 | 0.037 | 0.08 |

## Γ. CDS-συρραφές

### εκτός-Firmicutes

| | | 0% | 10% | 20% | 40% | 50% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|---|---|---|
| αποκλίσεις μονο- και δι- νουκλεοτιδίων | $S^{A-T}$ | -0.063 | -0.025 | -0.0083 | 0.006 | 0.011 | 0.022 | 0.047 | 0.12 |
| | $S^{G-C}$ | -0.059 | -0.012 | 0.0016 | 0.017 | 0.036 | 0.049 | 0.078 | 0.19 |
| | $S^{AG-CT}$ | -0.21 | -0.13 | -0.11 | -0.07 | -0.038 | -0.0026 | 0.091 | 0.25 |
| | $S^{GA-TC}$ | -0.052 | -0.0098 | 0.0066 | 0.055 | 0.078 | 0.093 | 0.14 | 0.34 |
| | $S^{GG-CC}$ | -0.15 | -0.0069 | 0.019 | 0.063 | 0.085 | 0.11 | 0.16 | 0.4 |
| | $S^{AA-TT}$ | -0.13 | -0.019 | 0.0022 | 0.054 | 0.064 | 0.074 | 0.12 | 0.2 |
| | $S^{AC-GT}$ | -0.21 | -0.059 | -0.037 | -0.0026 | 0.015 | 0.033 | 0.06 | 0.2 |
| | $S^{CA-TG}$ | -0.21 | -0.12 | -0.1 | -0.074 | -0.067 | -0.058 | -0.036 | 0.12 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | -0.34 | -0.23 | -0.2 | -0.15 | -0.13 | -0.11 | -0.064 | 0.044 |
| | $P^{GA-TC}$ | -0.12 | -0.043 | -0.0053 | 0.033 | 0.054 | 0.064 | 0.088 | 0.19 |
| | $P^{GG-CC}$ | -0.14 | -0.047 | -0.022 | 0.012 | 0.026 | 0.039 | 0.066 | 0.24 |
| | $P^{AA-TT}$ | -0.18 | -0.068 | -0.027 | 0.022 | 0.045 | 0.087 | 0.16 | 0.29 |
| | $P^{AC-GT}$ | -0.035 | 0.005 | 0.027 | 0.051 | 0.062 | 0.076 | 0.11 | 0.26 |
| | $P^{CA-TG}$ | -0.27 | -0.2 | -0.16 | -0.098 | -0.084 | -0.062 | -0.033 | 0.067 |

### Firmicutes

| | | 0% | 10% | 20% | 40% | 50% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|---|---|---|
| αποκλίσεις μονο- και δι- νουκλεοτιδίων | $S^{A-T}$ | -0.0037 | 0.021 | 0.029 | 0.043 | 0.072 | 0.078 | 0.085 | 0.15 |
| | $S^{G-C}$ | 0.016 | 0.071 | 0.072 | 0.089 | 0.1 | 0.12 | 0.15 | 0.28 |
| | $S^{AG-CT}$ | -0.092 | 0.035 | 0.059 | 0.096 | 0.13 | 0.19 | 0.26 | 0.33 |
| | $S^{GA-TC}$ | 0.071 | 0.11 | 0.14 | 0.15 | 0.21 | 0.24 | 0.26 | 0.45 |
| | $S^{GG-CC}$ | 0.029 | 0.14 | 0.15 | 0.2 | 0.23 | 0.27 | 0.3 | 0.46 |
| | $S^{AA-TT}$ | -0.0046 | 0.04 | 0.048 | 0.063 | 0.11 | 0.12 | 0.13 | 0.26 |
| | $S^{AC-GT}$ | -0.1 | -0.055 | -0.051 | -0.021 | -0.016 | -0.0017 | 0.0069 | 0.2 |
| | $S^{CA-TG}$ | -0.2 | -0.12 | -0.11 | -0.069 | -0.067 | -0.066 | -0.02 | 0.01 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | -0.25 | -0.17 | -0.12 | -0.082 | -0.076 | -0.037 | 0.045 | 0.096 |
| | $P^{GA-TC}$ | -0.028 | 0.023 | 0.035 | 0.058 | 0.061 | 0.077 | 0.09 | 0.21 |
| | $P^{GG-CC}$ | -0.27 | -0.017 | 0.00022 | 0.024 | 0.06 | 0.072 | 0.086 | 0.12 |
| | $P^{AA-TT}$ | -0.17 | -0.09 | -0.088 | -0.077 | -0.069 | -0.037 | -0.03 | 0.1 |
| | $P^{AC-GT}$ | 0.032 | 0.051 | 0.053 | 0.06 | 0.07 | 0.076 | 0.09 | 0.22 |
| | $P^{CA-TG}$ | -0.19 | -0.11 | -0.085 | -0.054 | -0.044 | -0.0085 | 0.037 | 0.072 |

ΣΗΜΕΙΩΣΕΙΣ.- Τα ποσοστημόρια των αποκλίσεων, για κάθε μία από τις δοσμένες πιθανότητες. Για κάθε απόκλιση, η μικρότερη τιμή αντιστοιχεί σε πιθανότητα που ισούται με το 0% και η μεγαλύτερη τιμή σε πιθανότητα που ισούται με το 100%. Αποκλίσεις με τιμές μικρότερες του μηδενός επισημαίνονται με κόκκινο χρώμα. Οι υπολογισμοί αφορούν Α: τον δημοσιευμένο κλώνο, Β: τον οδηγό κλώνο και Γ: τις CDS-συρραφές. Σε κάθε μία από αυτές τις περιπτώσεις, η συλλογή των χρωμοσωμάτων χωρίστηκε σε δύο υποσύνολα, όπου το ένα αντιστοιχεί σε όλα τα εξεταζόμενα φύλα εκτός των Firmicutes και το άλλο αντιστοιχεί μόνο στα Firmicutes.

Για να τονίσουμε τα συμπεράσματα που εξάγονται από τον Πίνακα 6, εστιάζουμε στο ζευγάρι AC/GT. Το 80% του εκτός-Firmicutes υποσυνόλου έχει κατά μήκος του οδηγού κλώνου $S_{leading}^{AC-GT}$ με εύρος τιμών μεταξύ -34% και -2.80% (Πίνακας 6Β). Αντίθετα, λιγότερα από τα μισά χρωμοσώματα του ίδιου υποσυνόλου έχουν $S^{AC-GT}$ με τιμές μικρότερες του μηδενός, τόσο στον δημοσιευμένο κλώνο, όσο και στις CDS-συρραφές (Πίνακας 6Α,Γ). Συνεπώς, *για το ίδιο ζεύγος δινουκλεοτιδίων, οι αποκλίσεις μπορεί να είναι πολωμένες προς διαφορετική κατεύθυνση, ανάλογα με τον κλώνο που εξετάζουμε*. Επιπλέον, στα Firmicutes οι αποκλίσεις δινουκλεοτιδίων για το ζεύγος AC/GT είναι αρνητικές σε περισσότερους από το 80% των οδηγών κλώνων (το 80% ποσοστημόριο των $S_{leading}^{AC-GT}$ ισούται με -3.9%), ενώ ταυτόχρονα τουλάχιστον το 80% των οδηγών κλώνων έχουν θετικές αποκλίσεις σταθμισμένων συχνοτήτων για το ίδιο ζεύγος, AC/GT (το 20% ποσοστημόριο των $P_{leading}^{AC-GT}$ ισούται με 0.66%). *Προκύπτει λοιπόν ότι για δεδομένο ζεύγος αντιστρόφως συμπληρωματικών δινουκλεοτιδίων, οι ειδικές ανά κλώνο ασυμμετρίες είναι δυνατόν να κινούνται προς αντίθετη κατεύθυνση στο επίπεδο των παρατηρούμενων συχνοτήτων εμφάνισης από ότι στο επίπεδο των σταθμισμένων συχνοτήτων.*

### **3.4.3** *Επεξηγηματικά σχόλια και παρατηρήσεις σχετικά με τις δινουκλεοτιδικές αποκλίσεις*

Για να αποσαφηνίσουμε τη σημασία των ευρημάτων μας, σχολιάζουμε ακολούθως ορισμένες πτυχές που αφορούν τις δινουκλεοτιδικές αποκλίσεις, σε όρους παρατηρούμενων συχνοτήτων εμφάνισης και σταθμισμένων συχνοτήτων. Παραμένοντας στο παράδειγμα του ζευγαριού AC/GT, $S_{leading}^{AC-GT} < 0$ σημαίνει ότι

κατά μήκος του οδηγού κλώνου υπάρχουν περισσότερα GT από ότι AC δινουκλεοτίδια. Από την άλλη, $P_{leading}^{AC-GT} > 0$ δηλώνει ότι η σχετική αφθονία των AC είναι μεγαλύτερη από εκείνη των GT, ή, πράγμα που είναι το ίδιο, οι σταθμισμένες συχνότητες των AC (ή των GT) λαμβάνουν τιμές που είναι υψηλότερες (ή, αντίστοιχα, χαμηλότερες) στον οδηγό σε σχέση με το συνοδό κλώνο. Με άλλα λόγια, δεδομένης της μονονουκλεοτιδικής σύστασης του DNA, τα κατάλοιπα A (ή G) και C (ή T) εκδηλώνουν μεγαλύτερη (ή, αντίστοιχα, μικρότερη) τάση να σχηματίζουν AC (ή GT) δινουκλεοτίδια στον οδηγό από ότι στο συνοδό κλώνο. Ωστόσο, η μονονουκλεοτιδική σύσταση του οδηγού και του συνοδού κλώνου είναι μεταξύ τους συμπληρωματικές, και εν γένει διαφορετικές η μία προς την άλλη. Κατά συνέπεια, ακόμα και αν το κλάσμα του συνόλου των A και C που γειτνιάζουν σχηματίζοντας AC δινουκλεοτίδια είναι μεγαλύτερο στον οδηγό από ότι στο συνοδό κλώνο, αυτό δεν συνεπάγεται απαραιτήτως υψηλότερη συχνότητα εμφάνισης των AC που βρίσκονται στον οδηγό έναντι εκείνων που βρίσκονται στο συνοδό κλώνο. Εάν ο οδηγός κλώνος είναι περισσότερο εμπλουτισμένος σε κατάλοιπα G και T από ότι ο συνοδός, μπορεί ταυτόχρονα να ισχύει ότι $S_{leading}^{AC-GT} < 0$ και $P_{leading}^{AC-GT} > 0$. Πράγματι, αυτό είναι που συναντάμε σε πολλά χρωμοσώματα της συλλογής μας.

### **3.5** *Κατανομή των δινουκλεοτιδίων κατά μήκος του οδηγού κλώνου*

Ακολούθως αναλύουμε την σύσταση του οδηγού κλώνου, στη βάση των καταγραφών του Πίνακα 6. Τα δινουκλεοτίδια στα οποία συμμετέχει τουλάχιστον ένα κατάλοιπο γουανίνης τείνουν να εμφανίζονται με υψηλότερες συχνότητες στον οδηγό από ότι στο συνοδό κλώνο. Επί παραδείγματι, στο εκτός-Firmicutes υποσύνολο, τα GA, GG, GT και TG βρίσκονται σε περίσσεια έναντι των αντιστρόφως συμπληρωματικών τους δινουκλεοτιδίων (το 10% ποσοστημόριο των $S_{leading}^{GA-TC}$ και $S_{leading}^{GG-CC}$ είναι θετικό, ενώ το 90% ποσοστημόριο των $S_{leading}^{AC-GT}$ και $S_{leading}^{CA-TG}$ είναι αρνητικό, Πίνακας 6Β). Ακόμα πιο έντονες είναι οι τάσεις αυτές όταν εστιάζουμε την ανάλυσή μας αποκλειστικά στα Firmicutes. Έτσι, στο σύνολο των χρωμοσωμάτων αυτού του φύλου τα οποία εξετάστηκαν, όλα τα δινουκλεοτίδια που περιέχουν κατάλοιπα G εμφανίζονται πιο συχνά από ότι τα

αντιστρόφως συμπληρωματικά τους, με μόνη εξαίρεση τα GT δινουκλεοτίδια. Και πάλι, τουλάχιστον το 80% των Firmicutes έχουν περισσότερα GT από ότι AC στον οδηγό κλώνο ($S_{leading}^{AC-GT} < 0$, Πίνακας 6Β). Οι παρατηρήσεις αυτές συμβαδίζουν με τον εμπλουτισμό του οδηγού κλώνου σε κατάλοιπα γουανίνης, τάση η οποία απαντάται ευρύτατα στα βακτηριακά χρωμοσώματα, όπως έχει ήδη καταγραφεί.

*Παρότι τα δινουκλεοτίδια που περιέχουν κατάλοιπα G είναι πιο άφθονα στον οδηγό κλώνο από εκείνα που περιέχουν C, συμβαίνει την ίδια στιγμή ορισμένα από αυτά να είναι λιγότερο προτιμώμενα*, όπως καταδεικνύουν οι αντίστοιχες αποκλίσεις σταθμισμένων συχνοτήτων. Τέτοια είναι η περίπτωση των AG και GG. Τόσο στο εκτός-Firmicutes υποσύνολο όσο και στα Firmicutes, η διάμεσος των $P_{leading}^{AG-CT}$ και $P_{leading}^{GG-CC}$ είναι αρνητική (Πίνακας 6Β). Συνεπώς, η σχετική αφθονία των AG και GG είναι μικρότερη από εκείνη των CT και CC, αντιστοίχως, στα μισά τουλάχιστον από τα εξεταζόμενα χρωμοσώματα, παρόλο που τα AG και GG απαντώνται συχνότερα στον οδηγό κλώνο από τα αντιστρόφως συμπληρωματικά τους. Αυτές οι τάσεις είναι κοινές και στα δύο υποσύνολα στα οποία χωρίσαμε τη συλλογή μας.

Οι ειδικές ανά κλώνο ασυμμετρίες του ζεύγους AC/GT παρουσιάζουν ιδιαίτερο ενδιαφέρον. Τα δινουκλεοτίδια GT έχουν υψηλότερες συχνότητες εμφάνισης κατά μήκος του οδηγού κλώνου από ότι τα AC για τουλάχιστον το 80% της συλλογής, είτε εξετάζουμε τα Firmicutes είτε το εκτός-Firmicutes υποσύνολο (το 80% ποσοστημόριο των $S_{leading}^{AC-GT}$ ισούται με -0.039 στα Firmicutes και με -0.028 στα εκτός-Firmicutes, Πίνακας 6Β). Όταν μάλιστα εξαιρούνται τα Firmicutes, η τάση αυτή συνοδεύεται και από υψηλότερους σταθμισμένες συχνότητες του GT έναντι του AC στον οδηγό κλώνο. Αντιθέτως, σε περισσότερα από το 80% των Firmicutes, η σταθμισμένη συχνότητα του AC είναι μεγαλύτερη από εκείνη του GT στον οδηγό κλώνο (πρβ. τα ποσοστημόρια των $P_{leading}^{AC-GT}$ στα Firmicutes και τα εκτός-Firmicutes, Πίνακας 6Β). Συνεπώς, *ο εμπλουτισμός του οδηγού κλώνου σε GT προκύπτει ως το αποτέλεσμα ασύμμετρων ανά κλώνο μηχανισμών που είναι διαφορετικοί στα Firmicutes και στο εκτός-Firmicutes υποσύνολο.*

Τα δύο αντιστρόφως συμπληρωματικά δινουκλεοτίδια που δεν περιέχουν κατάλοιπα γουανίνης, δηλαδή τα AA και TT, ακολουθούν πρότυπα αποκλίσεων τα οποία εμφανίζουν μεγαλύτερη ποικιλομορφία κατά μήκος του οδηγού κλώνου, σε σχέση με τα αντίστοιχα πρότυπα των υπλοίπων δινουκλεοτιδίων. Τουλάχιστον το

60% των οδηγών κλώνων στο εκτός-Firmicutes υποσύνολο έχει λιγότερα AA από ότι TT ($S_{leading}^{AA-TT} < 0$, Πίνακας 6B), ενώ ένα άλλο 60% εμφανίζεται να έχει σταθμισμένες συχνότητες μεγαλύτερες για το AA από ότι για το TT ($P_{leading}^{AA-TT} > 0$, Πίνακας 6B). Ωστόσο, όταν εστιάζουμε στα Firmicutes, αναδύεται ένα σαφές πρότυπο περίσσειας των AA έναντι των TT κατά μήκος του οδηγού κλώνου, σε περισσότερα από το 90% των χρωμοσωμάτων. Το πρότυπο αυτό συμβαδίζει με την μη-τυπική κατανομή των $S_{leading}^{A-T}$ στο εν λόγω φύλο (Charneski et al. 2011), όπου ο οδηγός κλώνος, αντί περίσσειας καταλοίπων T, έχει περισσότερα A ($S_{leading}^{A-T} > 0$). Και εδώ, όπως και στο εκτός-Firmicutes υποσύνολο, οι $S_{leading}^{AA-TT}$ είναι πολωμένες προς την αντίθετη κατεύθυνση σε σχέση με τις $P_{leading}^{AA-TT}$. Δηλαδή, *παρόλο που στα Firmicutes ανιχνεύεται η τάση προς υψηλότερες συχνότητες αλλά χαμηλότερη σχετική αφθονία των AA έναντι των TT, ενώ στο εκτός-Firmicutes υποσύνολο εμφανίζεται η αντίθετη τάση, και στις δύο αυτές ομάδες χρωμοσωμάτων η κατεύθυνση των* $S_{leading}^{AA-TT}$ *προκύπτει ως παρεπόμενη συνέπεια των ασυμμετριών που εκδηλώνονται στο επίπεδο των μονονουκλεοτιδίων* ($S_{leading}^{A-T} < 0$ για την πλειοψηφία των εκτός-Firmicutes, $S_{leading}^{A-T} > 0$ σχεδόν για το σύνολο των Firmicutes, Πίνακας 6B).

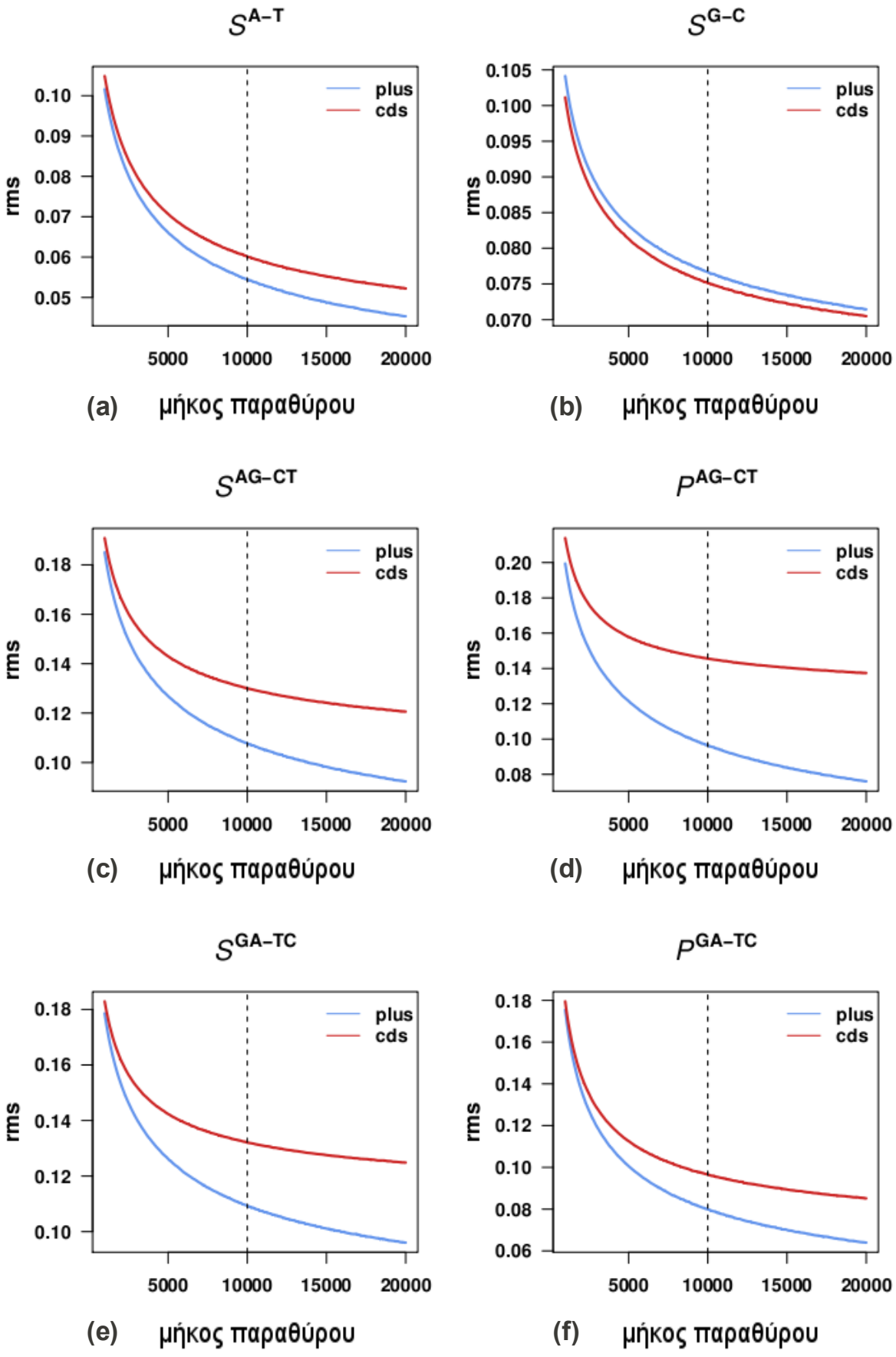
## **3.6** *Οι αποκλίσεις δινουκλεοτιδίων και σταθμισμένων συχνοτήτων εξαρτώνται από την κλίμακα παρατήρησης*

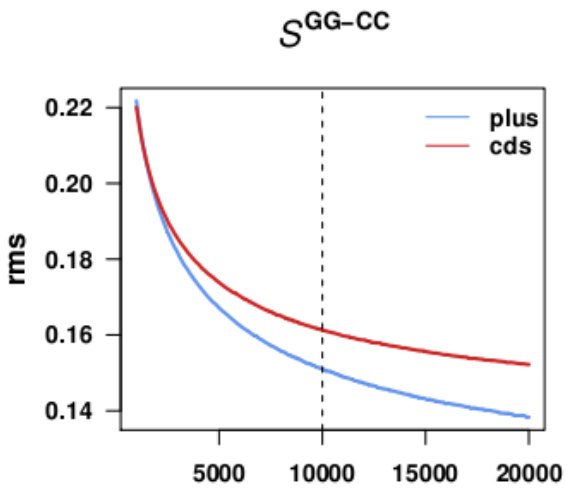
Οι αποκλίσεις από τις ισοδυναμίες [A]=[T] και [G]=[C] (PR2) είναι συνάρτηση του μήκους της χρωμοσωμικής περιοχής στην οποία γίνονται οι σχετικοί υπολογισμοί (Bell & Forsdyke 1999). Ο PR2 ισχύει στην κλίμακα ολόκληρου του χρωμοσώματος για τα περισσότερα από τα βακτηριακά γονιδιώματα (Mitchell & Bridge 2006), δηλαδή οι $S_{plus}^{A-T}$ και $S_{plus}^{G-C}$ τείνουν στο μηδέν, όπως φαίνεται και από τους Πίνακες 5A & 6A. Αντιθέτως, σε τοπική κλίμακα αναδύονται $S_{plus}^{A-T}$ και $S_{plus}^{G-C}$ με μη-μηδενικές τιμές.

Στις προηγούμενες ενότητες της εργασίας μας (3.3-3.5), καταδείχθηκε ρητώς ότι οι παρατηρούμενες συχνότητες των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων εμφανίζουν συστηματικές αποκλίσεις, ενώ παρουσιάστηκαν στοιχεία που τεκμηριώνουν σαφώς την ύπαρξη ειδικών ανά κλώνο ασυμμετριών

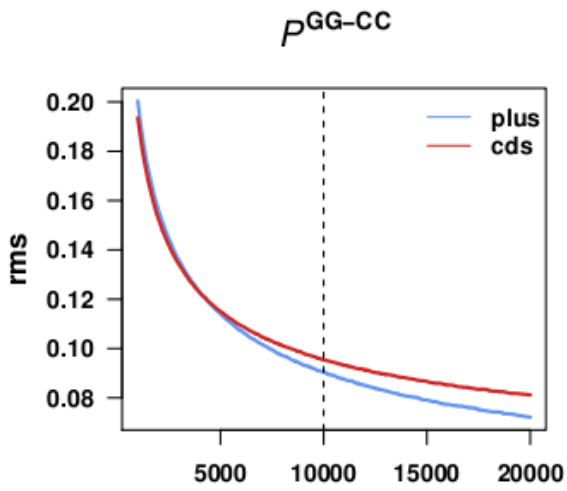
και στο επίπεδο των σταθμισμένων δινουκλεοτιδικών συχνοτήτων (Πίνακες 5 & 6).

Στην παρούσα ενότητα εξετάζουμε κατά πόσον η ένταση αυτών των αποκλίσεων μεταβάλλεται συναρτήσει του μήκους της αλληλουχίας DNA (*κυλιόμενο παράθυρο*) στην οποίο γίνονται οι υπολογισμοί, κατά τρόπο ανάλογο με τα όσα ισχύουν για τις μονονουκλεοτιδικές αποκλίσεις. Συγκεκριμένα, κατά μήκος του δημοσιευμένου κλώνου και των CDS-συρραφών ενός εκάστου χρωμοσώματος, μετρήσαμε κάθε μία από τις εξεταζόμενες αποκλίσεις σε δοσμένου μήκους, διαδοχικά, μη-επικαλυπτόμενα κυλιόμενα παράθυρα. Ακολούθως, υπολογίσαμε τον αντίστοιχο τετραγωνικό μέσο (root mean square, rms). Επί παραδείγματι, έστω οι αποκλίσεις $S^{AG-CT}$ κατά μήκος του δημοσιευμένου κλώνου ή των CDS-συρραφών ενός συγκεκριμένου χρωμοσώματος, όπως αυτές μετρήθηκαν σε $n$ διαδοχικά παράθυρα μήκους $L$ (bps). Υπολογίσαμε τον αντίστοιχο rms σύμφωνα με τον τύπο $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(S_i^{AG-CT}\right)^2}$, όπου $S_i^{AG-CT}$ είναι η τιμή $S^{AG-CT}$ στο κατά σειρά $i$ παράθυρο. Το μήκος $L$ των παραθύρων που χρησιμοποιήσαμε κυμαίνεται από $10^3$ έως $2 \cdot 10^4$ βάσεις. Για κάθε απόκλιση, απεικονίσαμε την μέση τιμή των rms όλων των χρωμοσωμάτων της συλλογής μας ως συνάρτηση του μήκους των παραθύρων που χρησιμοποιήσαμε, τόσο στον δημοσιευμένο κλώνο όσο και στις CDS-συρραφές. Τα αντίστοιχα διαγράμματα παρουσιάζονται ακολούθως, στην Εικόνα 6. Οι υπολογισμοί πραγματοποιήθηκαν τόσο για τις μονονουκλεοτιδικές αποκλίσεις, όσο και για τις αποκλίσεις των δινουκλεοτιδίων, σε όρους παρατηρούμενων και σταθμισμένων συχνοτήτων.

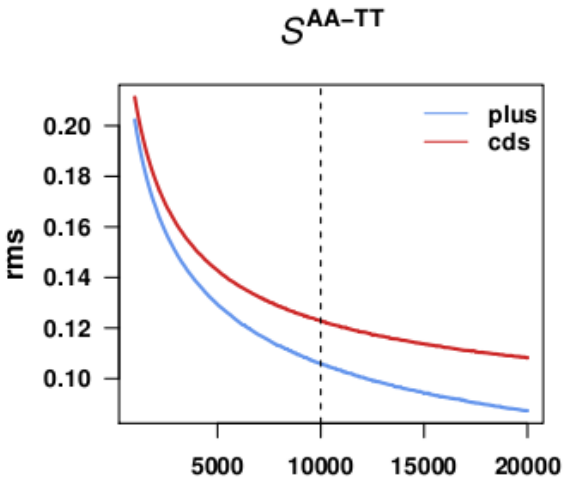

$S^{A-T}$
plus
cds
rms
0.10
0.09
0.08
0.07
0.06
0.05
5000
10000
15000
20000
(a) μήκος παραθύρου
$S^{G-C}$
plus
cds
rms
0.105
0.100
0.095
0.090
0.085
0.080
0.075
0.070
5000
10000
15000
20000
(b) μήκος παραθύρου
$S^{AG-CT}$
plus
cds
rms
0.18
0.16
0.14
0.12
0.10
5000
10000
15000
20000
(c) μήκος παραθύρου
$P^{AG-CT}$
plus
cds
rms
0.20
0.18
0.16
0.14
0.12
0.10
0.08
5000
10000
15000
20000
(d) μήκος παραθύρου
$S^{GA-TC}$
plus
cds
rms
0.18
0.16
0.14
0.12
0.10
5000
10000
15000
20000
(e) μήκος παραθύρου
$P^{GA-TC}$
plus
cds
rms
0.18
0.16
0.14
0.12
0.10
0.08
0.06
5000
10000
15000
20000
(f) μήκος παραθύρου



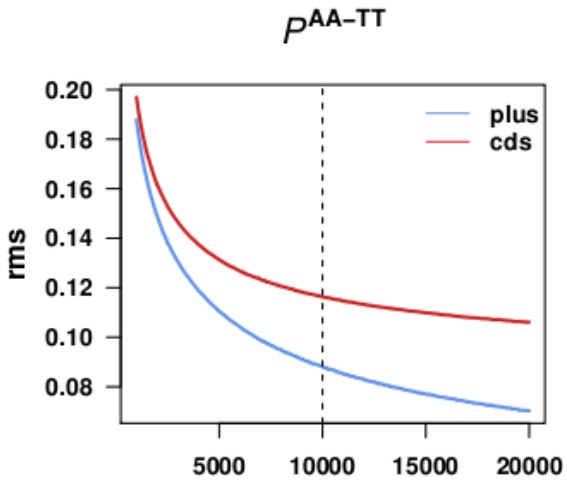

(g) μήκος παραθύρου



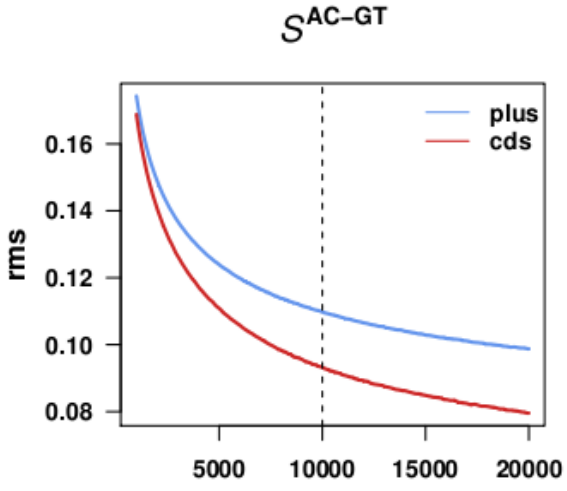

(h) μήκος παραθύρου



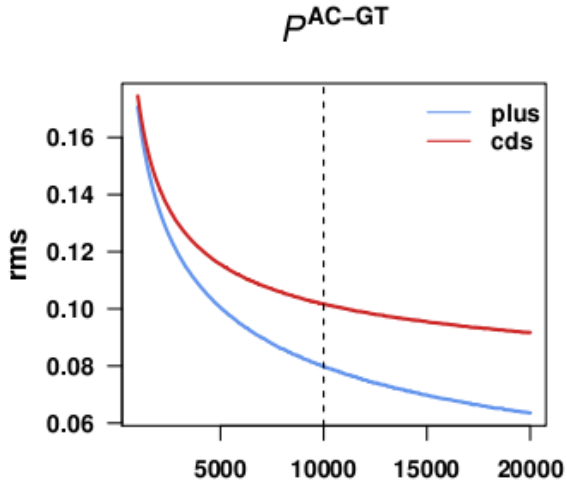

(i) μήκος παραθύρου



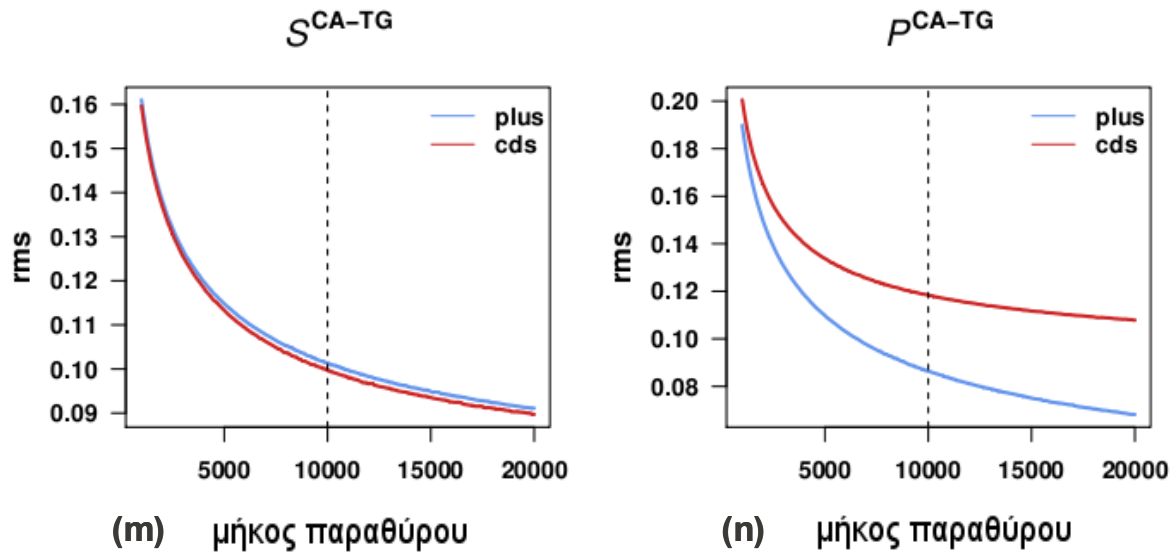

(j) μήκος παραθύρου




(k) μήκος παραθύρου




(l) μήκος παραθύρου




**Εικόνα 6.** Διαγράμματα των μέσων τιμών των rms (root mean squre, τετραγωνικοί μέσοι) των αποκλίσεων των μονο- και δι-νουκλεοτιδίων και των σταθμισμένων συχνοτήτων, ως συνάρτηση του μήκους $L$ των κυλιόμενων παραθύρων, στα οποία έγιναν οι σχετικοί υπολογισμοί. Το μήκος $L$ των παραθύρων κυμαίνεται από $10^3$ έως $2 \cdot 10^4$ βάσεις. Οι μετρήσεις αφορούν τις αποκλίσεις κατά μήκος του δημοσιευμένου κλώνου (μπλε καμπύλη) και των CDS-συρραφών (κόκκινη καμπύλη). Η κατακόρυφη, διακεκομμένη μαύρη γραμμή αντιστοιχεί σε παράθυρο μήκους $L=10^4$ bps, το οποίο χρησιμοποιούμε για τις περαιτέρω αναλύσεις μας.

Όλες οι αποκλίσεις έχουν υψηλότερες μέσες τιμές rms σε μικρότερου μήκους παράθυρα, τόσο στον δημοσιευμένο κλώνο όσο και στις CDS-συρραφές. Όπως συμβαίνει και στην περίπτωση των μονονουκλεοτιδικών αποκλίσεων (Bell & Forsdyke 1999, Nikolaou & Almirantis 2005), οι ειδικές ανά κλώνο ασυμμετρίες των παρατηρούμενων και των σταθμισμένων συχνοτήτων των δινουκλεοτιδικών είναι πιο έντονες όταν εξετάζονται στην τοπική κλίμακα μικρών χρωμοσωμικών περιοχών. Επιπλέον, οι καμπύλες της Εικόνας 6 έχουν κλίση που βαίνει διαρκώς μειούμενη, είτε απότομα είτε σταδιακά, καθώς αυξάνεται το μήκος $L$ του παραθύρου. Οι περιοχές όπου η κλίση των διαγραμμάτων τείνει στο μηδέν είναι ενδεικτικές της ύπαρξης μεγάλων τμημάτων DNA στα οποία οι ασυμμετρίες μεταξύ των αντιστρόφως συμπληρωματικών κλώνων διατηρούν σταθερή την έντασή τους. Τέτοιες περιοχές απαντώνται συχνότερα κατά μήκος των CDS-συρραφών από ότι κατά μήκος του δημοσιευμένου κλώνου (επί παραδείγματι, πρβ. τα σχετικά διαγράμματα των $S^{AG-CT}$ και $P^{AG-CT}$, Εικόνα 6c,d). Οι $S^{AC-GT}$ αποκλίσεις αποτελούν τη μόνη σημαντική

εξαίρεση, όπου ο ρυθμός μείωσης της κλίσης των διαγραμμάτων είναι μεγαλύτερος στον δημοσιευμένο κλώνο από ότι στις CDS-συρραφές, ωστόσο απέχει πολύ από το να τείνει στο μηδέν, τουλάχιστον για το εύρος του μήκους $L$ των παραθύρων που εξετάζουμε. Επίσης, οι ασυμμετρίες των δινουκλεοτιδίων είναι εν γένει πιο ισχυρές, δηλαδή έχουν μεγαλύτερες μέσες τιμές rms για δεδομένο μήκος $L$, στις CDS-συρραφές από ότι στον δημοσιευμένο κλώνο. Ακόμα και σε εκείνες τις περιπτώσεις όπου οι αποκλίσεις των παρατηρούμενων συχνοτήτων είναι εντονότερες στον δημοσιευμένο κλώνο, οι αποκλίσεις των σταθμισμένων συχνοτήτων παραμένουν μεγαλύτερες στις CDS-συρραφές, όπως συμβαίνει για το ζεύγος AC/GT.

Ως συνέπεια των παραπάνω, για τις αναλύσεις μας επιλέξαμε να χρησιμοποιήσουμε διαδοχικά, μη-επικαλυπτόμενα παράθυρα μήκους $L=10^4$ βάσεων. Το μήκος αυτό είναι αρκούντως μεγάλο ώστε να ανταποκρίνεται σε ένα στατιστικά ικανοποιητικό δείγμα για τους υπολογισμούς των δινουκλεοτιδικών αποκλίσεων. Στον Πίνακα 7 παραθέτουμε τις μέσες τιμές και τις τυπικές αποκλίσεις ($\sigma$) των rms που αντιστοιχούν σε κάθε μία από τις εξεταζόμενες αποκλίσεις, όταν οι υπολογισμοί γίνονται σε παράθυρο $10^4$ bps. Καθώς οι τιμές των $\sigma$ είναι ιδιαίτερα μικρές, συμπεραίνουμε ότι σε παράθυρα μήκους $10^4$ bps η μέση τιμή των rms περιγράφει ικανοποιητικά την συμπεριφορά των αποκλίσεων στο σύνολο της συλλογής μας. Συνεπώς *τα παράθυρα που επιλέγουμε αντιστοιχούν σε τμήματα DNA στα οποία η ένταση των αποκλίσεων κατανέμεται ομοιόμορφα*.

**ΠΙΝΑΚΑΣ 7. Στατιστικά στοιχεία της κατανομή των τετραγωνικών μέσων (rms) των αποκλίσεων σε παράθυρα μήκους $10^4$ βάσεων**

| | | A. δημοσιευμένος κλώνος | | B. CDS-συρραφές | |
|---|---|---|---|---|---|
| | | μέση τιμή | τυπική απόκλιση ($\sigma$) | μέση τιμή | τυπική απόκλιση ($\sigma$) |
| αποκλίσεις μόνο- και δι-νουκλεοτιδίων | $S^{A-T}$ | 0.0544 | 0.0176 | 0.0601 | 0.0212 |
| | $S^{G-C}$ | 0.0767 | 0.0455 | 0.0751 | 0.0487 |
| | $S^{AG-CT}$ | 0.108 | 0.0488 | 0.13 | 0.0554 |
| | $S^{GA-TC}$ | 0.109 | 0.0637 | 0.132 | 0.0791 |
| | $S^{GG-CC}$ | 0.151 | 0.0749 | 0.161 | 0.0929 |
| | $S^{AA-TT}$ | 0.106 | 0.0293 | 0.123 | 0.0357 |
| | $S^{AC-GT}$ | 0.11 | 0.0528 | 0.0932 | 0.0351 |
| | $S^{CA-TG}$ | 0.101 | 0.0426 | 0.0997 | 0.0333 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | 0.0964 | 0.0266 | 0.146 | 0.0624 |
| | $P^{GA-TC}$ | 0.08 | 0.0194 | 0.0966 | 0.0316 |
| | $P^{GG-CC}$ | 0.0904 | 0.0502 | 0.0953 | 0.0449 |
| | $P^{AA-TT}$ | 0.0881 | 0.0344 | 0.116 | 0.0633 |
| | $P^{AC-GT}$ | 0.0798 | 0.0201 | 0.102 | 0.0367 |
| | $P^{CA-TG}$ | 0.0864 | 0.0229 | 0.118 | 0.0522 |

ΣΗΜΕΙΩΣΕΙΣ.- Για κάθε χρωμόσωμα, οι αποκλίσεις υπολογίζονται σε διαδοχικά, μη-επικαλυπτόμενα παράθυρα μήκους $L=10^4$ bps και για κάθε απόκλιση υπολογίζεται ο τετραγωνικός μέσος (rms). Στον πίνακα παρατίθενται η μέση τιμή και η τυπική απόκλιση αυτών των rms για το σύνολο της συλλογής μας. Οι υπολογισμοί αφορούν Α: τον δημοσιευμένο κλώνο, και Β: τις CDS-συρραφές.

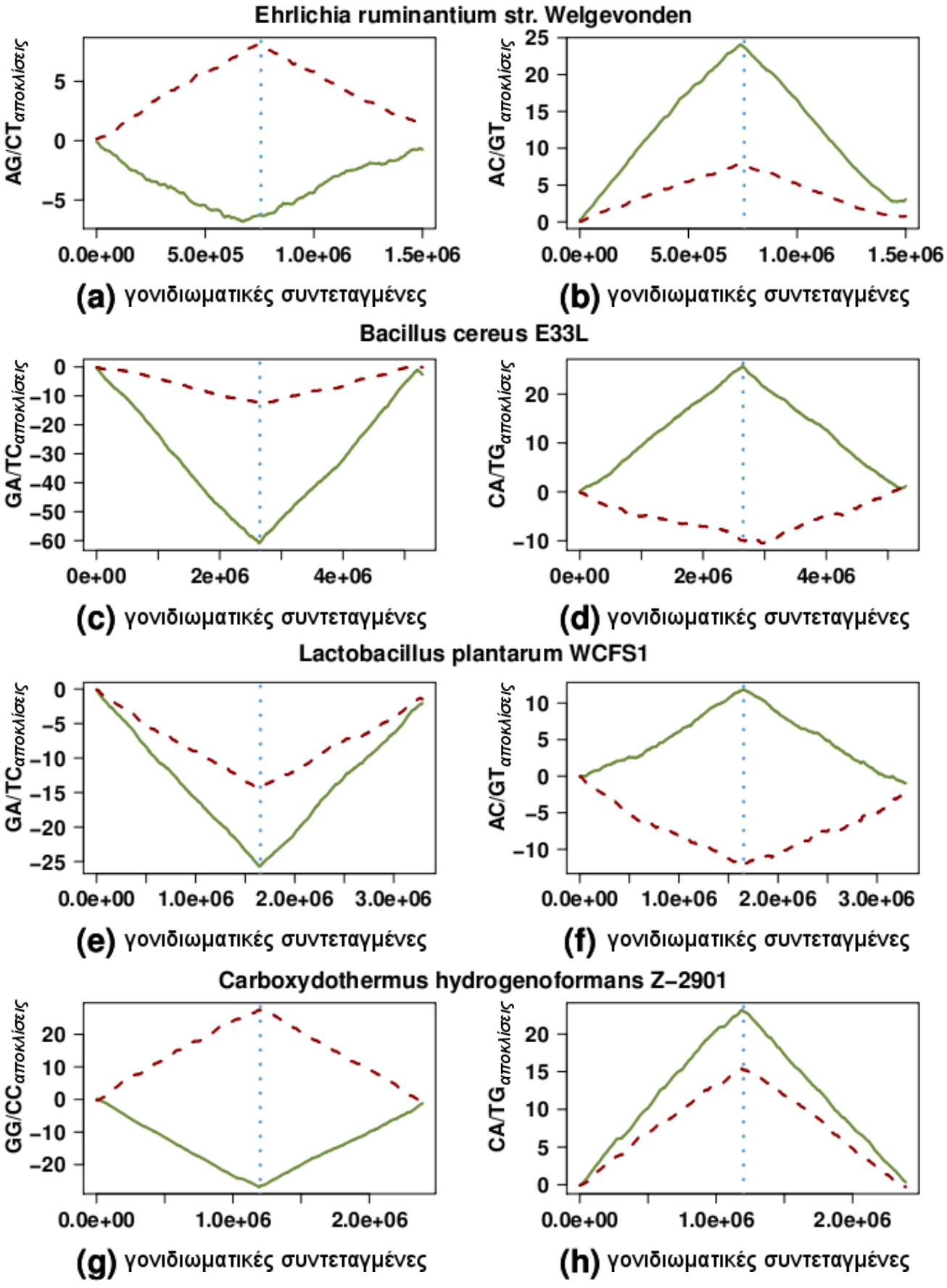
## **3.7** *Αθροιστικά γραφήματα των αποκλίσεων δινουκλεοτιδίων και σταθμισμένων συχνοτήτων*

Για κάθε χρωμόσωμα της συλλογής μας, αναπαριστούμε τα αθροιστικά γραφήματα των αποκλίσεων για τα έξι ζεύγη αντιστρόφως συμπληρωματικών δινουκλεοτιδίων, με όρους τόσο παρατηρούμενων όσο και σταθμισμένων συχνοτήτων (βλ. σχετικό link στο τέλος της Βιβλιογραφίας). Επιλέγουμε τέσσερα από αυτά τα χρωμοσώματα, τα ίδια που εξετάσαμε στην Εικόνα 5 ως προς τις μονονουκλεοτιδικές τους αποκλίσεις, και απεικονίζουμε στην Εικόνα 7 τα αθροιστικά διαγράμματα των δινουκλεοτιδικών αποκλίσεων κατά μήκος του δημοσιευμένου τους κλώνο. Συγκεκριμένα, τα διαγράμματα αυτά αντιστοιχούν στα ζεύγη AG/CT και AC/GT του *Ehrlichia ruminantium* str. Welgevonden, στα ζεύγη GA/TC και CA/TG του *Bacillus cereus* E33L, στα ζεύγη GA/TC και AC/GT του *Lactobacillus plantarum* WCFS1 και στα ζεύγη GG/CC και CA/TG του *Carboxydothermus hydrogenoformans* Z-2901.

Τα αθροιστικά διαγράμματα των απεικονιζόμενων αποκλίσεων, τόσο των δινουκλεοτιδίων όσο και των σταθμισμένων συχνοτήτων, έχουν τη χαρακτηριστική μορφή V ή ανεστραμμένου V, με τα ακρότατά τους να εντοπίζονται στην περιοχή του *ori* ή και να συμπίπτουν με αυτό. Προκύπτει λοιπόν ότι οι δινουκλεοτιδικές αποκλίσεις ακολουθούν πρότυπα κατανομής κατά μήκος του δημοσιευμένου κλώνου αντίστοιχα με εκείνα των μονονουκλεοτιδικών αποκλίσεων. Με άλλα λόγια, *οι αποκλίσεις των δινουκλεοτιδίων και των σταθμισμένων τους συχνοτήτων είναι κατά προσέγγιση σταθερές κατά μήκος του οδηγού και του συνοδού κλώνου, ενώ τα πρόσημά τους είναι αντίθετα εκατέρωθεν του σημείου έναρξης της αντιγραφής*.

Από την Εικόνα 7 γίνεται φανερό ότι για το ίδιο ζεύγος δινουκλεοτιδίων, οι αποκλίσεις των παρατηρούμενων και των σταθμισμένων συχνοτήτων σε ορισμένες περιπτώσεις συμμεταβάλλονται (*b*,*c*,*e*,*h*), ενώ σε άλλες παρουσιάζουν αντίστροφη συσχέτιση (*a*,*d*,*f*,*g*). Συνεπώς, *οι διαδικασίες που παράγουν τις ειδικές ανά κλώνο ασυμμετρίες δρουν με τρόπο διαφορετικό στο επίπεδο των παρατηρούμενων δινουκλεοτιδικών συχνοτήτων από ότι σε εκείνο των σταθμισμένων δινουκλεοτιδικών συχνοτήτων*. Επίσης, για ένα δεδομένο ζεύγος δινουκλεοτιδίων, οι ειδικές ανά κλώνο ασυμμετρίες μπορεί να συγκροτούν πρότυπα (patterns) που διαφέρουν μεταξύ των βακτηριακών γονιδιωμάτων. Έτσι, ενώ οι $S_{plus}^{AC-GT}$ έχουν αθροιστικά διαγράμματα

ανεστραμμένου V τόσο στο *Ehrlichia ruminantium* str. Welgevonden, όσο και στο *Lactobacillus plantarum* WCFS1, τα $P_{plus}^{AC-GT}$ αθροιστικά διαγράμματα έχουν στο μεν πρώτο μορφή ανεστραμμένου V (Εικόνα 7*b*), στο δε δεύτερο μορφή V (Εικόνα 7*f*). Τα δύο αυτά βακτήρια προέρχονται από διαφορετικά φύλα, από τα Πρωτεοβακτήρια το *E.ruminantium* και από τα Firmicutes το *L.plantarum*. Αντίστοιχες περιπτώσεις εντοπίζονται και εντός του ίδιου φύλου. Τα βακτήρια *Bacillus cereus* E33L και *Carboxydothermus hydrogenoformans* Z-2901 ανήκουν στα Firmicutes και έχουν $S_{plus}^{CA-TG}$ αθροιστικά διαγράμματα με μορφή ανεστραμμένου V, αλλά οι $S_{plus}^{CA-TG}$ και $P_{plus}^{CA-TG}$ στο πρώτο κινούνται προς αντίθετες κατευθύνσεις (Εικόνα 7*d*) ενώ στο δεύτερο προς την ίδια κατεύθυνση (Εικόνα 7*h*). Τα παραδείγματα αυτά καταδεικνύουν ότι σε γονιδιώματα όπου η κατανομή των δινουκλεοτιδίων είναι πολωμένη προς τον ίδιο κλώνο, οδηγό ή συνοδό, οι αντίστοιχες σταθμισμένες συχνότητες μπορεί να παρουσιάζουν διαφορετικά πρότυπα ασυμμετριών.


Ehrlichia ruminantium str. Welgevonden
AG/CT αποκλίσεις
AC/GT αποκλίσεις
(a) γονιδιωματικές συντεταγμένες
(b) γονιδιωματικές συντεταγμένες
Bacillus cereus E33L
GA/TC αποκλίσεις
CA/TG αποκλίσεις
(c) γονιδιωματικές συντεταγμένες
(d) γονιδιωματικές συντεταγμένες
Lactobacillus plantarum WCFS1
GA/TC αποκλίσεις
AC/GT αποκλίσεις
(e) γονιδιωματικές συντεταγμένες
(f) γονιδιωματικές συντεταγμένες
Carboxydothermus hydrogenoformans Z−2901
GG/CC αποκλίσεις
CA/TG αποκλίσεις
(g) γονιδιωματικές συντεταγμένες
(h) γονιδιωματικές συντεταγμένες

**Εικόνα 7.** Αθροιστικά διαγράμματα των δινουκλεοτιδικών αποκλίσεων κατά μήκος του δημοσιευμένου (plus) κλώνου τεσσάρων βακτηριακών χρωμοσωμάτων. Σε κάθε πλαίσιο αναπαρίστανται οι αποκλίσεις ενός δεδομένου δινουκλεοτιδικού ζεύγους, σε όρους τόσο παρατηρούμενων συχνοτήτων (συνεχής πράσινη γραμμή) όσο και σταθμισμένων συχνοτήτων (διακεκομμένη κόκκινη γραμμή). Η κατακόρυφη, στικτή μπλε γραμμή δηλώνει το σημείο έναρξης της αντιγραφής (*ori*). Αριστερά του *ori* οι εικονιζόμενες αποκλίσεις αντιστοιχούν στο συνοδό κλώνο, ενώ δεξιά του *ori* αντιστοιχούν στον οδηγό κλώνο. Τα εικονιζόμενα διαγράμματα έχουν τη χαρακτηριστική μορφή V ή ανεστραμμένο V, με τα ακρότατά τους να συμπίπτουν με το *ori* ή να εντοπίζονται πλησίον του. Για το ίδιο δινουκλεοτιδικό ζεύγος, τα αθροιστικά διαγράμματα των αποκλίσεων των παρατηρούμενων και των σταθμισμένων συχνοτήτων μπορεί να κινούνται είτε προς την ίδια κατεύθυνση (b,c,e,h) είτε προς την αντίθετη (a,d,f,g).

### **3.7.1** *Μελέτη των προτύπων ασυμμετρίας βάσει των γενικών χαρακτηριστικών των αθροιστικών διαγραμμάτων*

Όπως και στην περίπτωση των μονονουκλεοτιδικών αποκλίσεων (ενότητα 3.1), έτσι και για τις αποκλίσεις των δινουκλεοτιδίων και των σταθμισμένων συχνοτήτων τους, προκειμένου να συνοψίσουμε τα γενικά χαρακτηριστικά των αθροιστικών διαγραμμάτων τους, κατατάξαμε τα χρωμοσώματα της συλλογής μας σε μία από τις εξής τρεις κατηγορίες: (*α*) τύπου-V, όπου οι αποκλίσεις έχουν θετική τιμή στον οδηγό και αρνητική στο συνοδό κλώνο, (*β*) τύπου-ανεστραμμένου V, όπου οι αποκλίσεις έχουν αρνητική τιμή στον οδηγό και θετική στο συνοδό κλώνο, και (*γ*) παραμορφωμένα, όπου οι αποκλίσεις έχουν το ίδιο πρόσημο (είτε θετικό είτε αρνητικό) και στους δύο κλώνους. Τα αντίστοιχα ποσοστά καταγράφονται στον Πίνακα 8.

**ΠΙΝΑΚΑΣ 8. Ποσοστιαία κατάταξη των χρωμοσωμάτων, βάσει των γενικών χαρακτηριστικών της μορφής που έχουν τα αθροιστικά διαγράμματα των αποκλίσεων των δινουκλεοτιδίων και των σταθμισμένων συχνοτήτων τους.**

| | | **εκτός-Firmicutes** | | | **Firmicutes** | | |
|---|---|---|---|---|---|---|---|
| | | ***παραμορφωμένα διαγράμματα*** | ***διαγράμματα τύπου-V*** | ***διαγράμματα τύπου ανεστραμμένου V*** | ***παραμορφωμένα διαγράμματα*** | ***διαγράμματα τύπου-V*** | ***διαγράμματα τύπου ανεστραμμένου V*** |
| **αποκλίσεις δινουκλεοτιδίων** | $S_{plus}^{AG-CT}$ | 12 | 76.4 | 11.6 | 1.56 | 98.4 | 0.0 |
| | $S_{plus}^{GA-TC}$ | 8.73 | 86.2 | 5.5 | 0 | 100.0 | 0.0 |
| | $S_{plus}^{GG-CC}$ | 3.26 | 93.1 | 3.6 | 0 | 100.0 | 0.0 |
| | $S_{plus}^{AA-TT}$ | 14.5 | 15.9 | 69.6 | 4.69 | 92.2 | 3.1 |
| | $S_{plus}^{AC-GT}$ | 6.88 | 2.9 | 90.2 | 4.69 | 6.3 | 89.0 |
| | $S_{plus}^{CA-TG}$ | 3.99 | 3.3 | 92.7 | 0 | 0.0 | 100.0 |
| **αποκλίσεις σταθμισμένων συχνοτήτων** | $P_{plus}^{AG-CT}$ | 21.4 | 38.4 | 40.2 | 4.69 | 43.7 | 51.6 |
| | $P_{plus}^{GA-TC}$ | 15.6 | 72.8 | 11.6 | 1.56 | 96.8 | 1.6 |
| | $P_{plus}^{GG-CC}$ | 14.9 | 33.7 | 51.4 | 17.2 | 18.7 | 64.1 |
| | $P_{plus}^{AA-TT}$ | 9.78 | 65.2 | 25.0 | 1.56 | 4.7 | 93.7 |
| | $P_{plus}^{AC-GT}$ | 16.3 | 14.8 | 68.9 | 3.12 | 86.0 | 10.9 |
| | $P_{plus}^{CA-TG}$ | 19.9 | 26.1 | 54.0 | 15.6 | 46.9 | 37.5 |

ΣΗΜΕΙΩΣΕΙΣ.- ***Παραμορφωμένα*** θεωρούνται τα αθροιστικά διαγράμματα που αντιστοιχούν σε αποκλίσεις των οποίων το πρόσημο δεν αλλάζει εκατέρωθεν του σημείου *ori*. Στα χρωμοσώματα των οποίων οι αποκλίσεις αλλάζουν πρόσημο εκατέρωθεν του *ori*, τα αντίστοιχα διαγράμματα θεωρούνται είτε ως ***τύπου-V*** (θετικές αποκλίσεις στον οδηγό και αρνητικές αποκλίσεις στο συνοδό κλώνο) είτε ως ***τύπου-ανεστραμμένου V*** (αρνητικές αποκλίσεις στον οδηγό και θετικές αποκλίσεις στο συνοδό κλώνο).

Σε συμφωνία με τα αποτελέσματα που παρουσιάστηκαν στην ενότητα 3.5,

από τον Πίνακα 8 προκύπτει ότι τα δινουκλεοτίδια που περιλαμβάνουν κατάλοιπα γουανίνης εμφανίζονται με υψηλότερες συχνότητες στον οδηγό από ότι στο συνοδό κλώνο. Επί παραδείγματι, το σύνολο των Firmicutes έχει $S_{plus}^{GA-TC}$ και $S_{plus}^{GG-CC}$ αθροιστικά διαγράμματα τύπου-V και $S_{plus}^{CA-TG}$ αθροιστικά διαγράμματα τύπου-ανεστραμμένου V. Παρομοίως, τα $S_{plus}^{GA-TC}$ και $S_{plus}^{GG-CC}$ αθροιστικά διαγράμματα είναι τύπου-V στο 86.2% και 93.1% αντίστοιχα των εκτός-Firmicutes, ενώ τα $S_{plus}^{CA-TG}$ αθροιστικά διαγράμματα είναι τύπου-ανεστραμμένου V στο 92.7% του ίδιου υποσυνόλου. Αυτές οι παρατηρήσεις συμβαδίζουν με τον ευρύτατα απαντώμενο εμπλουτισμό του οδηγού κλώνου σε κατάλοιπα G, που οδηγεί σε $S_{plus}^{G-C}$ αθροιστικά διαγράμματα τύπου-V στο 94.2% των εκτός-Firmicutes και στο 100% των Firmicutes (Πίνακας 2). Από τα εξεταζόμενα δινουκλεοτίδια, τα δύο αντιστρόφως συμπληρωματικά που δεν περιέχουν G, δηλαδή τα AA και TT, παρουσιάζουν ειδικές ανά κλώνο ασυμμετρίες των οποίων τα αθροιστικά διαγράμματα εμφανίζουν μεγαλύτερη ποικιλομορφία. Έτσι, στις αποκλίσεις $S_{plus}^{AA-TT}$ αντιστοιχεί το μεγαλύτερο ποσοστό παραμορφωμένων αθροιστικών διαγραμμάτων για τα εκτός-Firmicutes (14.5%), ενώ παρόμοιο ποσοστό χρωμοσωμάτων (15.9%) έχουν $S_{plus}^{AA-TT}$ αθροιστικά διαγράμματα τύπου-V στο ίδιο υποσύνολο. Ωστόσο, όταν εστιάσουμε την ανάλυσή μας στα Firmicutes, αναδύεται ένα σαφές πρότυπο περίσσειας AA έναντι TT κατά μήκος του οδηγού κλώνου, με το 92.2% των αντίστοιχων χρωμοσωμάτων να έχουν $S_{plus}^{AA-TT}$ αθροιστικά διαγράμματα τύπου-V. Το πρότυπο αυτό συμβαδίζει με τις μη-τυπικές A-T αποκλίσεις των Firmicutes, όπου το 93.7% έχουν $S_{plus}^{A-T}$ αθροιστικά διαγράμματα τύπου-V (Πίνακας 2). Πράγματι, όσον αφορά τα Firmicutes που περιλαμβάνονται στη συλλογή μας, όλα τα χρωμοσώματα που έχουν $S_{plus}^{A-T}$ αθροιστικά διαγράμματα τύπου-V εμφανίζουν παράλληλα και $S_{plus}^{AA-TT}$ αθροιστικά διαγράμματα του ίδιου τύπου, με μόνη εξαίρεση το *Streptococcus agalactiae* A909, του οποίου οι $S_{plus}^{AA-TT}$ αντιστοιχούν σε ένα *παραμορφωμένο* διάγραμμα. Ωστόσο, οι παρατηρήσεις αυτές δεν μας επιτρέπουν να διακρίνουμε εάν και κατά πόσο οι μονονουκλεοτιδικές (1ης τάξης) ασυμμετρίες συμβάλλουν, κατ' επέκταση, στην εμφάνιση ασυμμετριών ανώτερης τάξης (εν προκειμένω, αποκλίσεων των δινουκλεοτιδικών συχνοτήτων), ή, αντιστρόφως, ανώτερης τάξης ασυμμετρίες επιβάλλουν περιορισμούς στην κατανομή των βάσεων, με αποτέλεσμα να επάγουν τις παρατηρούμενες μονονουκλεοτιδικές αποκλίσεις.

Σε αντίθεση με τις παρατηρούμενες συχνότητες, οι σταθμισμένες συχνότητες των δινουκλεοτιδίων απαλείφουν την επίδραση της μονονουκλεοτιδικής σύστασης στις ανώτερης τάξης ασυμμετρίες και έτσι

επιτρέπουν την μελέτη και αξιολόγηση γνήσιων (genuine) ασυμμετριών που εκδηλώνονται πρωτογενώς στο επίπεδο των συσχετίσεων μεταξύ των κοντινότερων γειτονικών βάσεων. Για κάθε δεδομένο ζεύγος αντιστρόφως συμπληρωματικών δινουκλεοτιδίων, οι αποκλίσεις των σταθμισμένων συχνοτήτων εμφανίζουν τοπικές διακυμάνσεις, οι οποίες οδηγούν σε μεγαλύτερο αριθμό *παραμορφωμένων* αθροιστικών διαγραμμάτων σε σύγκριση με τα αντίστοιχα αθροιστικά διαγράμματα των αποκλίσεων των παρατηρούμενων συχνοτήτων. Χαρακτηριστική εξαίρεση αποτελεί το ζευγάρι των AA/TT, το οποίο έχει πιο ακανόνιστα πρότυπα αποκλίσεων στο επίπεδο των παρατηρούμενων συχνοτήτων από ότι σε εκείνο των σταθμισμένων συχνοτήτων. Συγκεκριμένα, οι αποκλίσεις $S_{plus}^{AA-TT}$ έχουν *παραμορφωμένα* αθροιστικά διαγράμματα στο 14.5% των εκτός-Firmicutes και στο 4.69% των Firmicutes, έναντι 9.78% και 1.56%, αντίστοιχα, που έχουν *παραμορφωμένα* $P_{plus}^{AA-TT}$ αθροιστικά διαγράμματα. Ας σημειωθεί ότι ένα μόνο χρωμόσωμα των Firmicutes έχει $P_{plus}^{AA-TT}$ με *παραμορφωμένο* διάγραμμα (1.56% σε σύνολο 64 DNA αλληλουχιών), και είναι αυτό του *Streptococcus mutans* UA159, που συμβαίνει να είναι επίσης το μόνο Firmicutes με *παραμορφωμένο* $S_{plus}^{A-T}$ διάγραμμα. Οι σταθμισμένες συχνότητες ακολουθούν ακανόνιστα πρότυπα κατ' ελάχιστον σε 28 χρωμοσώματα του συνόλου της συλλογής μας, στην περίπτωση των $P_{plus}^{AA-TT}$, ενώ ο αντίστοιχος αριθμός φτάνει τα 65, στην περίπτωση των $P_{plus}^{CA-TG}$.

Όπως έχουμε ήδη αναφέρει, τα δινουκλεοτίδια που περιέχουν G κατανέμονται κατά προτίμηση στον οδηγό κλώνο, συγκρινόμενα με εκείνα που περιέχουν C (Πίνακας 6B, Πίνακας 8). Ωστόσο, αυτή η γενική τάση δεν συνδέεται κατ' ανάγκη με μια υψηλότερη σχετική αφθονία του συνόλου των νουκλεοτιδικών δυάδων με G έναντι εκείνων με C, κατά μήκος του οδηγού κλώνου. Επί παραδείγματι, ενώ οι αποκλίσεις $S_{plus}^{AG-CT}$ έχουν αθροιστικά διαγράμματα τύπου-V για το 76.4% των εκτός-Firmicutes και για το 98.4% των Firmicutes (δηλαδή, για το σύνολο των εξεταζόμενων Firmicutes πλην του *Symbiobacterium thermophilum* IAM 14863), τα αντίστοιχα ποσοστά για τις αποκλίσεις $P_{plus}^{AG-CT}$ είναι μόλις 38.4% και 43.7%. Προκύπτει λοιπόν ότι οι αποκλίσεις $S_{plus}^{AG-CT}$ και $P_{plus}^{AG-CT}$, ανάλογα με τον οργανισμό που εξετάζουμε, ενδέχεται να ακολουθούν διαφορετικά, ασυσχέτιστα ή αρνητικά συσχετιζόμενα πρότυπα, όπως στην περίπτωση του *Ehrlichia ruminantium* str. Welgevonden (Εικόνα 7*a*), όπου οι $S_{plus}^{AG-CT}$ και $P_{plus}^{AG-CT}$ αναπαρίστανται οι μεν πρώτες με τύπου-V αθροιστικά διαγράμματα, οι δε δεύτερες με τύπου-ανεστραμμένου V αθροιστικά διαγράμματα. Στο πλαίσιο αυτής της συζήτησης, σημειώνουμε ότι

παρ' όλο που κανένα από τα Firmicutes της συλλογής μας δεν έχει $S_{plus}^{GG-CC}$ που να δίνουν αθροιστικά διαγράμματα τύπου-ανεστραμμένου V, οι $P_{plus}^{GG-CC}$ έχουν αθροιστικά διαγράμματα αυτού του τύπου στο 64.1% των Firmicutes. Ανάλογη είναι η τάση που εκδηλώνεται και στο εκτός-Firmicutes υποσύνολο.

Το ζεύγος AC/GT παρουσιάζει επίσης ιδιαίτερο ενδιαφέρον. Τα αθροιστικά διαγράμματα των $S_{plus}^{AC-GT}$ είναι του τύπου-ανεστραμμένου V για το 90.2% των εκτός-Firmicutes και το 89% των Firmicutes. Οι αποκλίσεις $P_{plus}^{AC-GT}$ ακολουθούν την ίδια τάση στα εκτός-Firmicutes, αν και σε μικρότερη ένταση, με το 68.9% να έχουν αθροιστικά διαγράμματα τύπου-ανεστραμμένου V. Συνεπώς, στα εκτός-Firmicutes τα αθροιστικά διαγράμματα $S_{plus}^{AC-GT}$ και $P_{plus}^{AC-GT}$ κινούνται προς την ίδια φορά, όπως, επί παραδείγματι, παρατηρούμε στο χρωμόσωμα του *E.ruminantium* (Εικόνα 7*b*). Αντιθέτως, το 86% των Firmicutes παρουσιάζει αθροιστικά διαγράμματα $P_{plus}^{AC-GT}$ τύπου-V. Συνεπώς, στο φύλο αυτό η υψηλότερη συχνότητα εμφάνισης των GT συνοδεύεται κατά κανόνα από τη χαμηλότερη σχετική αφθονία τους, έναντι των AC, στον οδηγό κλώνο, όπως συμβαίνει στην περίπτωση του *L.plantarum* (Εικόνα 7*f*).

Όπως ήδη αναφέραμε, οι ασυμμετρίες των συχνοτήτων εμφάνισης του ζεύγους AA/TT ακολουθούν ανομοιόμορφα πρότυπα που διακρίνουν μεταξύ των Firmicutes και των εκτός-Firmicutes. Το δινουκλεοτιδικό αυτό ζεύγος εμφανίζει χαρακτηριστικές ιδιαιτερότητες και στην κατανομή και στις ειδικές ανά κλώνο πολώσεις των σταθμισμένων του συχνοτήτων. Συγκεκριμένα, ενώ το 69.6% των εκτός-Firmicutes έχει αθροιστικά διαγράμματα $S_{plus}^{AA-TT}$ τύπου-ανεστραμμένου V, οι $P_{plus}^{AA-TT}$ αντιστοιχούν σε αθροιστικά διαγράμματα τύπου-V στο 65.2% του ίδιου υποσυνόλου. Στα Firmicutes ανιχνεύεται η ακριβώς αντίθετη τάση, η οποία μάλιστα εκδηλώνεται με ακόμα μεγαλύτερη ένταση. Έτσι, ενώ το 92.2% αυτού του φύλου έχει αθροιστικά διαγράμματα $S_{plus}^{AA-TT}$ τύπου-V, το 93.7% έχει $P_{plus}^{AA-TT}$ που δίνουν αθροιστικά διαγράμματα τύπου-ανεστραμμένου V. Συμπερασματικά, τόσο στο εκτός-Firmicutes υποσύνολο όσο και στα Firmicutes τα αθροιστικά διαγράμματα των αποκλίσεων των παρατηρούμενων συχνοτήτων και των σταθμισμένων συχνοτήτων του AA/TT κινούνται προς αντίθετες κατευθύνσεις κατά μήκος της πλειοψηφίας των δημοσιευμένων κλώνων που εξετάζουμε. Ωστόσο, η φορά του καθενός από τους δύο τύπους αποκλίσεων διαφέρει στα Firmicutes και στα εκτός-Firmicutes. Οι παρατηρήσεις αυτές επιβεβαιώνουν τα σχετικά αποτελέσματα που έχουμε παρουσιάσει προηγουμένως (Πίνακας 6Β & ενότητα 3.5). Επιπλέον, ενισχύουν την άποψη ότι *η περίσσεια των* AA *έναντι των* TT *που παρατηρείται στον οδηγό*

*κλώνο των Firmicutes, δεν οφείλεται σε τάσεις μεγαλύτερης προτίμησης ή μικρότερης αποφυγής των* AA *σε σχέση με τα* TT*, αλλά είναι απλά το δευτερογενώς παραγόμενο αποτέλεσμα του εμπλουτισμού του οδηγού κλώνου σε κατάλοιπα* A *έναντι* T*, που παρατηρείται ειδικά στο συγκεκριμένο φύλο.*

Ανακεφαλαιώνοντας, στις προηγούμενες ενότητες (3.3-3.7) παρουσιάσαμε ισχυρά στοιχεία που καταδεικνύουν την ύπαρξη ειδικών ανά κλώνο δινουκλεοτιδικών αποκλίσεων, στο επίπεδο τόσο των παρατηρούμενων συχνοτήτων εμφάνισης όσο και των σταθμισμένων συχνοτήτων. Είναι η πρώτη φορά που καταγράφονται ρητά αυτού του τύπου οι ασυμμετρίες της σύστασης του DNA. Προηγούμενες μελέτες είχαν αναζητήσει πιθανές δινουκλεοτιδικές αποκλίσεις, ωστόσο δεν κατάφεραν να τις εντοπίσουν. Οι Mrázek και Karlin (1998), στηριζόμενοι στα περιορισμένης έκτασης δεδομένα της εποχής, κατέληξαν στο συμπέρασμα πως οι σταθμισμένες συχνότητες των αντιστρόφως συμπληρωματικών δινουκλεοτιδίων κατανέμονται ομοιόμορφα στον οδηγό και το συνοδό κλώνο των βακτηριακών χρωμοσωμάτων. Οι τιμές των αντίστοιχων αποκλίσεων θεωρήθηκαν αμελητέες και, ως εκ τούτου, δεν παρουσιάστηκαν εκτενώς. Τρία χρόνια αργότερα, οι Shioiri και Takahata (2001) εξέτασαν τις ειδικές ανά κλώνο αποκλίσεις των δινουκλεοτιδικών συχνοτήτων, εισάγοντας την αντίστοιχη μαθηματική έκφραση που χρησιμοποιούμε και εμείς στην μελέτη μας. Ωστόσο, και σε αυτή την περίπτωση, οι αποκλίσεις που εντοπίστηκαν θεωρήθηκαν ασήμαντες και τα σχετικά αποτελέσματα δεν παρουσιάστηκαν. Αντιθέτως, από τα αποτελέσματα που παρατίθενται στους Πίνακες 4-8 και από τα αθροιστικά διαγράμματα της Εικόνας 7 προκύπτει ότι τόσο *οι παρατηρούμενες δινουκλεοτιδικές συχνότητες όσο και οι σταθμισμένες δινουκλεοτιδικές συχνότητες ακολουθούν πολωμένες ανά κλώνο κατανομές, που διακρίνουν τόσο μεταξύ οδηγού και συνοδού, όσο και μεταξύ κωδικού και μεταγραφόμενου κλώνου. Οι τιμές των αντίστοιχων αποκλίσεων είναι είτε συγκρίσιμες είτε και μεγαλύτερες από τις αποκλίσεις στο επίπεδο της μονονουκλεοτιδικής σύστασης του DNA*. Επιπλέον, γίνεται σαφές ότι *τα πρότυπα που εμφανίζουν οι αποκλίσεις των δινουκλεοτιδίων δεν μπορούν να συναχθούν από εκείνα των αποκλίσεων των σταθμισμένων συχνοτήτων, ούτε και αντιστρόφως*. Στην αμέσως επόμενη ενότητα αξιολογούμε την στατιστική σημαντικότητα των προτύπων που εμφανίζουν οι δινουκλεοτιδικές αποκλίσεις και συμπεραίνουμε ότι *οι αποκλίσεις αυτές απέχουν πολύ από το να είναι απλά αποτέλεσμα τυχαίων διακυμάνσεων, αλλά, αντιθέτως είναι ένα συστηματικά εμφανιζόμενο χαρακτηριστικό της σύστασης του DNA*.

## **3.8** *Στατιστική αξιολόγηση των αλλαγών της δομής στα πρότυπα των αποκλίσεων*

### **3.8.1** *Δημοσιευμένος κλώνος*

#### **3.8.1.1** *Στατιστική σημαντικότητα της συσχέτισης των αποκλίσεων με το σημείο έναρξης της αντιγραφής*

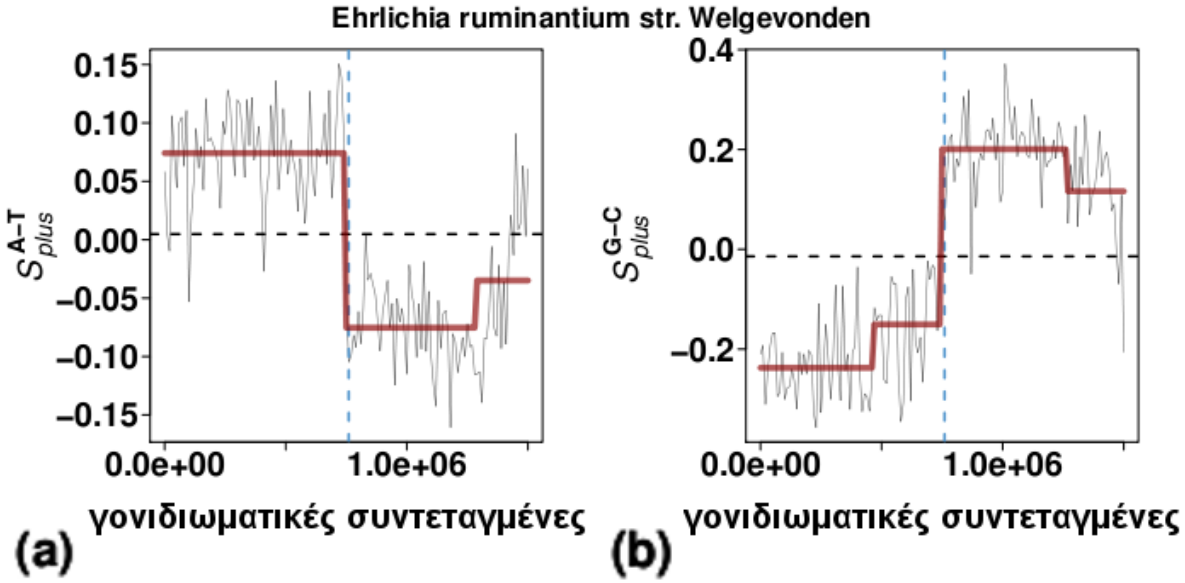
Η χρήση των αθροιστικών διαγραμμάτων των αποκλίσεων, όπως αυτές υπολογίζονται κατά μήκος του δημοσιευμένου κλώνου, αποτελεί έναν πρόσφορο μέσο για την ανάλυση των ειδικών ανά κλώνο ασυμμετριών που μελετάμε (ενότητα 3.7). Αθροιστικά διαγράμματα τύπου V ή ανεστραμμένου-V, τα οποία έχουν τα ακρότατά τους στην περιοχή του *ori* και του *ter*, φανερώνουν την ύπαρξη ισχυρών ασυμμετριών μεταξύ οδηγού και συνοδού κλώνου. Στις περιπτώσεις αυτές, *οι απεικονιζόμενες αποκλίσεις έχουν αντίθετα πρόσημα στους δύο κλώνους της αντιγραφής, ενώ κατά μήκος του κάθε ενός από τους δύο αυτούς κλώνους οι αποκλίσεις εμφανίζουν μικρή διασπορά, καθώς συγκεντρώνονται γύρω από μια σταθερή τιμή*. Συνεπώς, οι αποκλίσεις που έχουν αθροιστικά διαγράμματα τύπου V ή ανεστραμμένου-V μπορούν να περιγραφούν από μοντέλα γραμμικής παλινδρόμησης τα οποία αλλάζουν συντελεστές στα σημεία έναρξης (*ori*) και λήξης (*ter*) της αντιγραφής. Σε αυτό το πλαίσιο, το *ori* (ή το *ter*) αντιστοιχεί σε ένα σημείο μεταβολής (breakpoint) των συντελεστές που περιγράφουν τα μοντέλα.

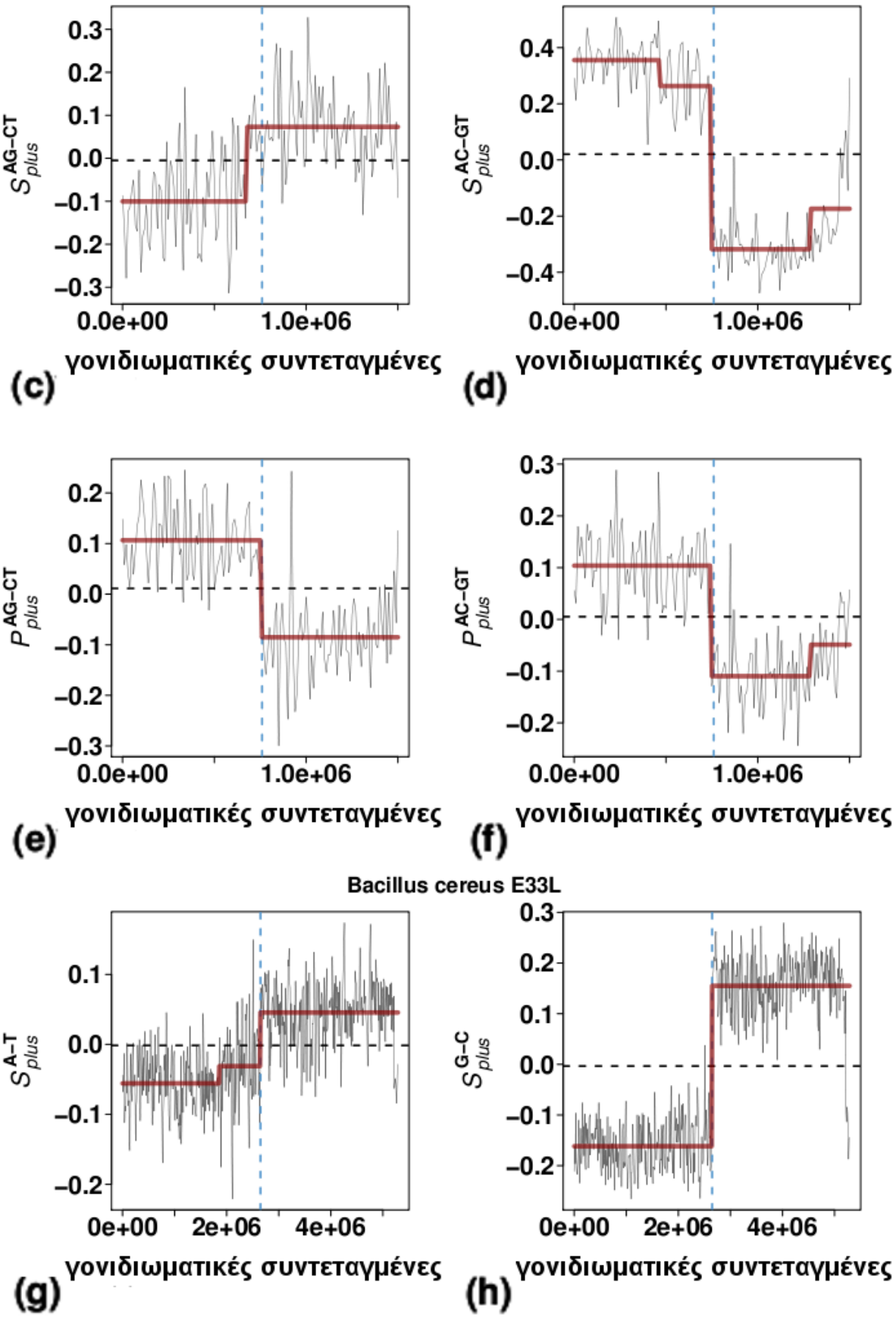
Η παραπάνω συλλογιστική βρίσκει εφαρμογή στην αξιολόγηση της στατιστικής σημαντικότητας των αλλαγών που εμφανίζει η πόλωση (κατεύθυνση) των αποκλίσεων εκατέρωθεν του *ori*. Προκειμένου να εκτιμήσουμε την ευθύγραμμη παλινδρόμηση των αποκλίσεων πάνω στις γονιδιωματικές συντεταγμένες, χρησιμοποιήσαμε μία κατάλληλη στατιστική κατασκευή (setup) που επιτρέπει τον έλεγχο έναντι πιθανών σημείων μεταβολής (breakpoints) των συντελεστών γραμμικής παλινδρόμησης. Συγκεκριμένα, για το σύνολο των δημοσιευμένων κλώνων της συλλογής μας εφαρμόσαμε τον αλγόριθμο δυναμικού προγραμματισμού που ανέπτυξαν οι Zeileis *et al*. (2003, 2010) και υλοποίησαν με τη γλώσσα προγραμματισμού R, στο πακέτο strucchange (Zeileis et al. 2002). Ο αλγόριθμος αυτός εντοπίζει το σύνολο των βέλτιστων σημείων

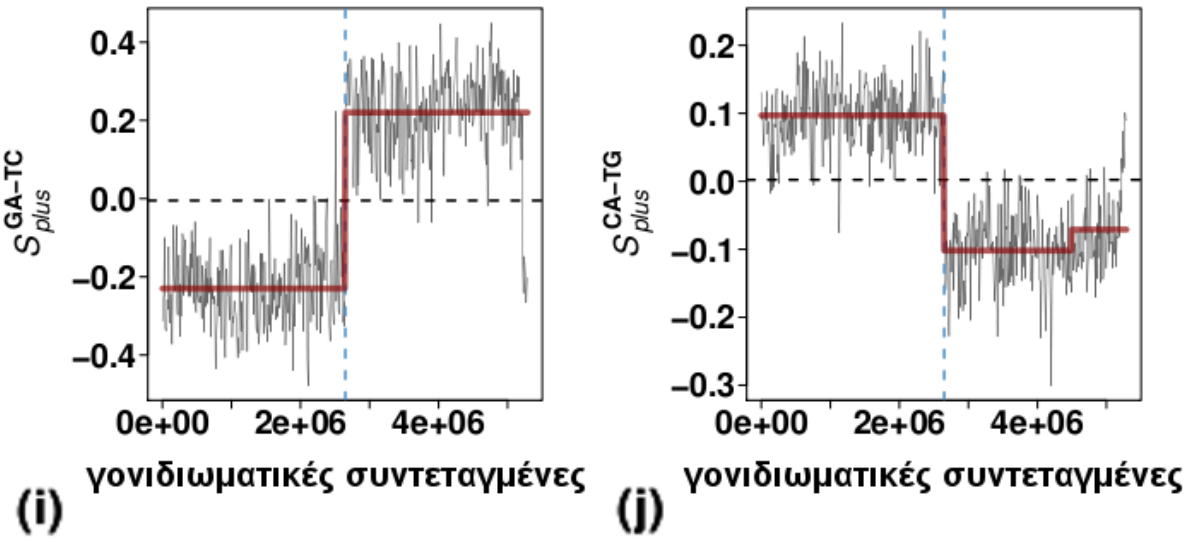
μεταβολής, στα οποία οι συντελεστές των μοντέλων παλινδρόμησης μεταβαίνουν από μία σταθερή, γραμμική σχέση μεταξύ αποκλίσεων και γονιδιωματικών συντεταγμένων, σε μία άλλη. Με αυτό τον τρόπο, κάθε δεδομένο χρωμόσωμα χωρίζεται σε επιμέρους διακριτά τμήματα, καθένα από τα οποία αντιστοιχίζεται σε ένα συγκεκριμένο μοντέλο γραμμικής παλινδρόμησης. Για μία αναλυτικότερη παρουσίαση της σχετικής μεθοδολογίας, βλ. ενότητα 2.6.

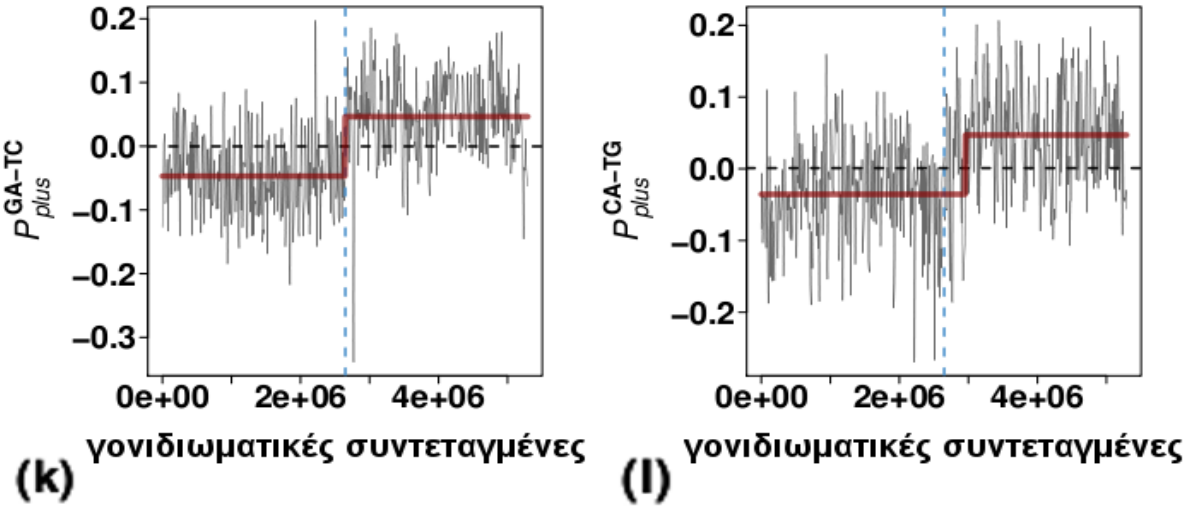
#### **3.8.1.2** *Γραφική απεικόνιση*

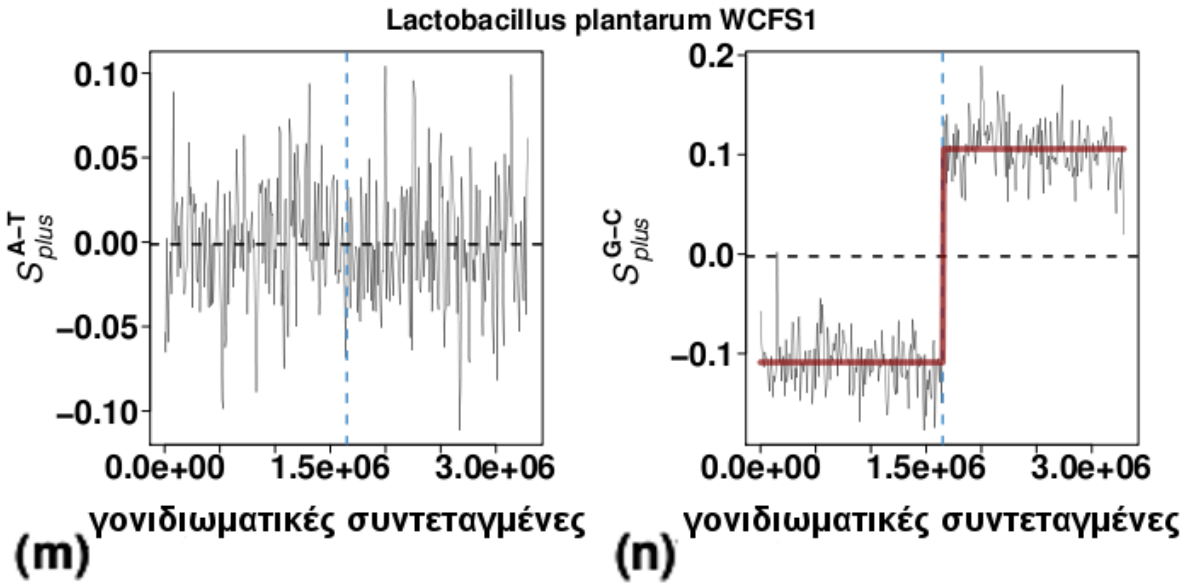
Στην Εικόνα 8 παρουσιάζουμε τα απλά (μη-αθροιστικά) διαγράμματα των μονονουκλεοτιδικών αποκλίσεων και των αποκλίσεων των δινουκλεοτιδίων και των σταθμισμένων τους συχνοτήτων, για το σύνολο των περιπτώσεων που εμφανίζονται στις Εικόνες 5 και 7, στις οποίες απεικονίζονται τα αντίστοιχα αθροιστικά διαγράμματα. Εκτός από τα διαγράμματα των αποκλίσεων, σε κάθε γράφημα της Εικόνας 8 σχεδιάσαμε και τα διαγράμματα των μοντέλων γραμμικής παλινδρόμησης που περιγράφουν αυτές τις αποκλίσεις (τεθλασμένες κόκκινες γραμμές).

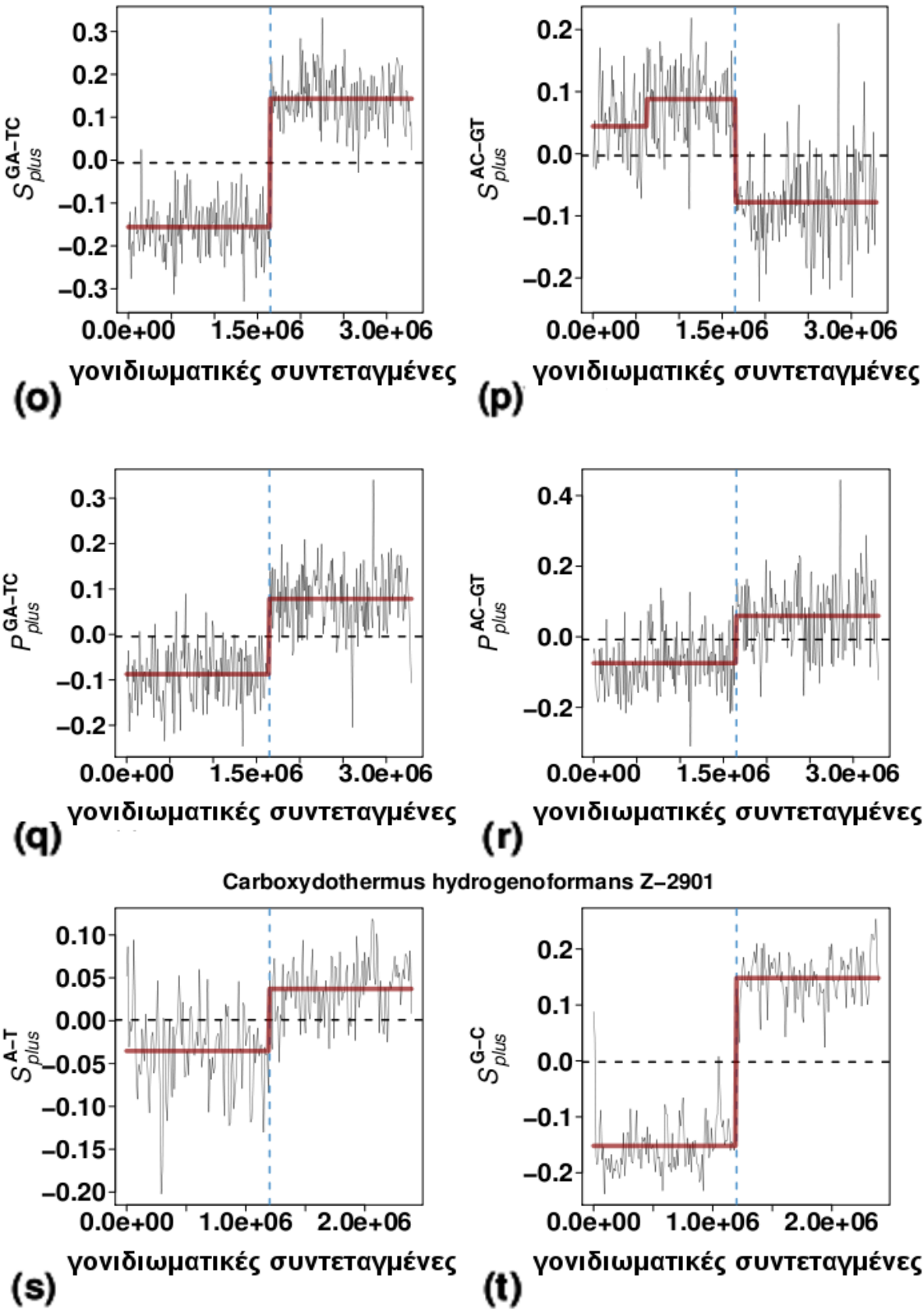


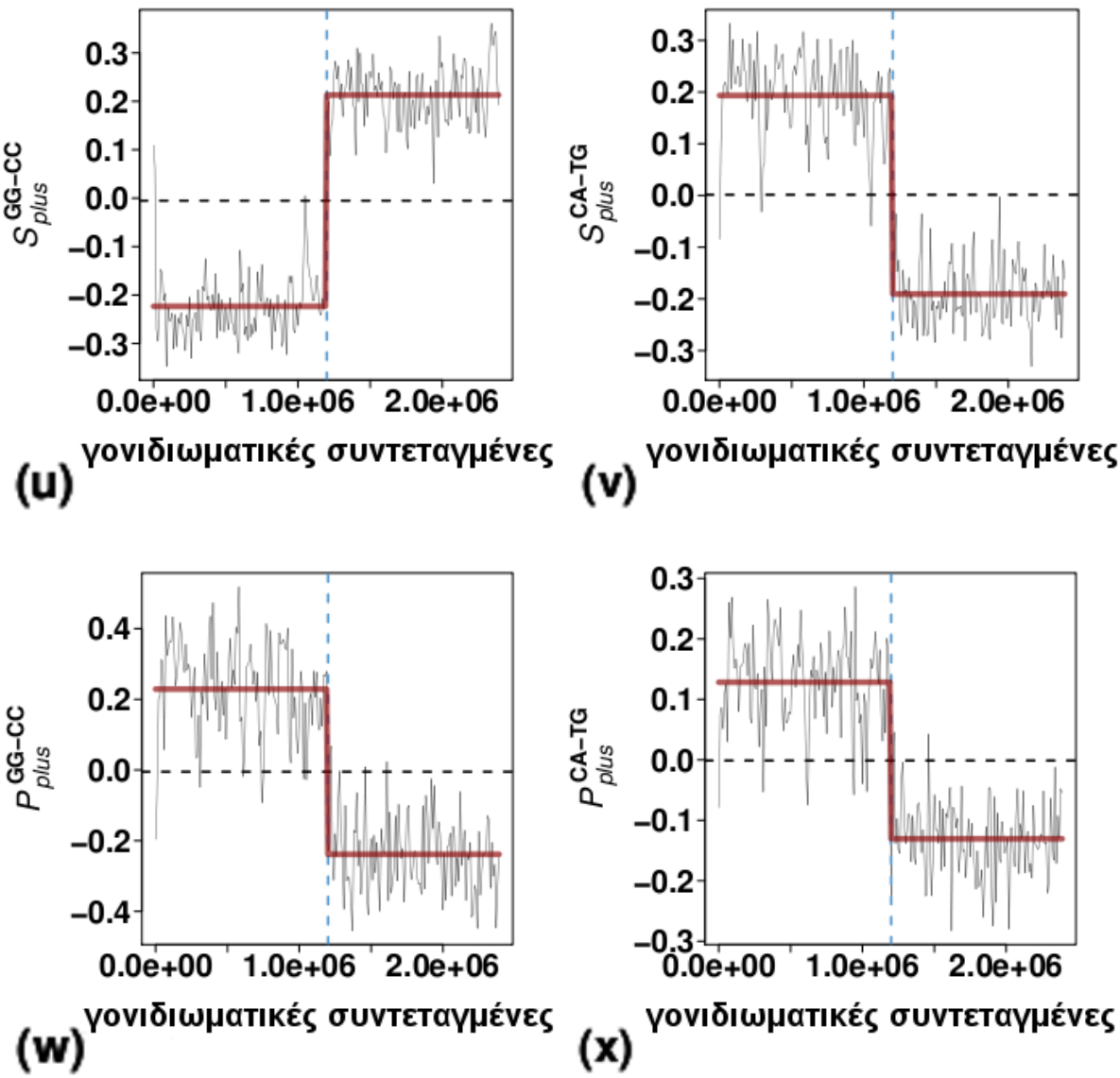

$S^{AG-CT}_{plus}$
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.0e+00
1.0e+06
γονιδιωματικές συντεταγμένες
(c)
$S^{AC-GT}_{plus}$
0.4
0.2
0.0
−0.2
−0.4
0.0e+00
1.0e+06
γονιδιωματικές συντεταγμένες
(d)
$P^{AG-CT}_{plus}$
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.0e+00
1.0e+06
γονιδιωματικές συντεταγμένες
(e)
$P^{AC-GT}_{plus}$
0.3
0.2
0.1
0.0
−0.1
−0.2
0.0e+00
1.0e+06
γονιδιωματικές συντεταγμένες
(f)
Bacillus cereus E33L
$S^{A-T}_{plus}$
0.1
0.0
−0.1
−0.2
0e+00
2e+06
4e+06
γονιδιωματικές συντεταγμένες
(g)
$S^{G-C}_{plus}$
0.3
0.2
0.1
0.0
−0.1
−0.2
0e+00
2e+06
4e+06
γονιδιωματικές συντεταγμένες
(h)


$S^{GA-TC}_{plus}$
0.4
0.2
0.0
−0.2
−0.4
0e+00
2e+06
4e+06
γονιδιωματικές συντεταγμένες
(i)
$S^{CA-TG}_{plus}$
0.2
0.1
0.0
−0.1
−0.2
−0.3
0e+00
2e+06
4e+06
γονιδιωματικές συντεταγμένες
(j)


$P^{GA-TC}_{plus}$
0.2
0.1
0.0
−0.1
−0.2
−0.3
0e+00
2e+06
4e+06
γονιδιωματικές συντεταγμένες
(k)
$P^{CA-TG}_{plus}$
0.2
0.1
0.0
−0.1
−0.2
0e+00
2e+06
4e+06
γονιδιωματικές συντεταγμένες
(l)


Lactobacillus plantarum WCFS1
$S^{A-T}_{plus}$
0.10
0.05
0.00
−0.05
−0.10
0.0e+00
1.5e+06
3.0e+06
γονιδιωματικές συντεταγμένες
(m)
$S^{G-C}_{plus}$
0.2
0.1
0.0
−0.1
0.0e+00
1.5e+06
3.0e+06
γονιδιωματικές συντεταγμένες
(n)


$S^{GA-TC}_{plus}$
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.0e+00
1.5e+06
3.0e+06
γονιδιωματικές συντεταγμένες
(o)
$S^{AC-GT}_{plus}$
0.2
0.1
0.0
−0.1
−0.2
0.0e+00
1.5e+06
3.0e+06
γονιδιωματικές συντεταγμένες
(p)
$P^{GA-TC}_{plus}$
0.3
0.2
0.1
0.0
−0.1
−0.2
0.0e+00
1.5e+06
3.0e+06
γονιδιωματικές συντεταγμένες
(q)
$P^{AC-GT}_{plus}$
0.4
0.2
0.0
−0.2
0.0e+00
1.5e+06
3.0e+06
γονιδιωματικές συντεταγμένες
(r)
Carboxydothermus hydrogenoformans Z−2901
$S^{A-T}_{plus}$
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
0.0e+00
1.0e+06
2.0e+06
γονιδιωματικές συντεταγμένες
(s)
$S^{G-C}_{plus}$
0.2
0.1
0.0
−0.1
−0.2
0.0e+00
1.0e+06
2.0e+06
γονιδιωματικές συντεταγμένες
(t)




**Εικόνα 8.** Ασυμμετρίες κατά μήκος του δημοσιευμένου κλώνου. Διαγράμματα των μονονουκλεοτιδικών αποκλίσεων και επιλεγμένων δινουκλεοτιδικών αποκλίσεων, σε όρους τόσο παρατηρούμενων όσο και σταθμισμένων συχνοτήτων, (κυματιστές γκρι γραμμές), μαζί με τα αντίστοιχα μοντέλα γραμμικής παλινδρόμησης, που περιγράφουν αυτές τις αποκλίσεις (τεθλασμένες κόκκινες γραμμές). Στο γράφημα (*m*) δεν εμφανίζεται το διάγραμμα του μοντέλου, καθώς δεν εντοπίζεται κανένα στατιστικά σημαντικό σημείο μεταβολής. Επίσης, στα γραφήματα (*c*, *l*) ενώ το διάγραμμα του μοντέλου εν πολλοίς διακρίνει μεταξύ οδηγού και συνοδού κλώνου, το σημείο μεταβολής, αν και σαφές, είναι σχετικά απομακρυσμένο από το *ori*.

Οι αποκλίσεις υπολογίζονται κατά μήκος του δημοσιευμένου κλώνου τεσσάρων βακτηρίων, εντός διαδοχικών, μη-επικαλυπτόμενων παραθύρων μήκους $10^4$ bps. Η κατακόρυφη, διακεκομμένη μπλε γραμμή δηλώνει το σημείο έναρξης της αντιγραφής (*ori*). Αριστερά

του *ori* οι εικονιζόμενες αποκλίσεις αντιστοιχούν στο συνοδό κλώνο, ενώ δεξιά του *ori* αντιστοιχούν στον οδηγό κλώνο. Η οριζόντια, διακεκομμένη μαύρη γραμμή δηλώνει την μέση τιμή των αποκλίσεων, που είναι κατά προσέγγιση μηδενική στον δημοσιευμένο κλώνο, καθώς οι αποκλίσεις του οδηγού και του συνοδού κλώνου αλληλοεξουδετερώνονται στην κλίμακα ολόκληρου του γονιδιώματος.

Από τα εικοσιτέσσερα διαγράμματα που παρουσιάζονται στην Εικόνα 8, στα εικοσιένα εντοπίζεται τουλάχιστον ένα σαφές σημείο μεταβολής στο *ori*. Σε ορισμένα από αυτά η περιοχή του *ori* συμπίπτει με το μοναδικό, βέλτιστο σημείο μεταβολής, όπως παρατηρούμε στο σύνολο των απεικονιζόμενων αποκλίσεων κατά μήκος του δημοσιευμένου κλώνου του *C.hydrogenoformans* (Εικόνα 8*s-x*), ενώ σε άλλα υπάρχουν περισσότερα του ενός σαφή σημεία μεταβολής, όπως συμβαίνει, επί παραδείγματι, στα διαγράμματα των $S_{plus}^{A-T}$, $S_{plus}^{G-C}$, $S_{plus}^{AC-GT}$ και $P_{plus}^{AC-GT}$ του *E.ruminantium* (Εικόνα 8*a*,*b*,*d*,*f*). Ωστόσο, ακόμα και όταν εντοπίζονται περισσότερα του ενός βέλτιστα σημεία μεταβολής, χωρίζοντας έτσι σε περισσότερα από δύο διακριτά τμήματα το εξεταζόμενο χρωμόσωμα, τα αντίστοιχα μοντέλα γραμμικής παλινδρόμησης έχουν συντελεστές του ίδιου προσήμου εκατέρωθεν του *ori*. Συνεπώς, σε κάθε περίπτωση, *τα μοντέλα γραμμικής παλινδρόμησης που περιγράφουν αυτές τις αποκλίσεις, διακρίνουν ανάμεσα στον οδηγό και το συνοδό κλώνο*.

Ορισμένα από τα διαγράμματα της Εικόνας 8, τρία τον αριθμό, παρουσιάζουν διαφορετικού τύπου πρότυπα κατανομής των αποκλίσεών τους, σε σύγκριση με τα προαναφερθέντα. Οι $S_{plus}^{A-T}$ αποκλίσεις του *L.plantarum* (Εικόνα 8*m*) δεν εμφανίζουν κανένα στατιστικά σημαντικό σημείο μεταβολής κατά μήκος του οδηγού κλώνου, γεγονός που έρχεται σε συμφωνία με το πολύ περιορισμένο εύρος τιμών και την ακανόνιστη μορφή του αντίστοιχου αθροιστικού διαγράμματος (Εικόνα 5*e*). Οι αποκλίσεις $S_{plus}^{AG-CT}$ του *E.ruminantium* (Εικόνα 8*c*) και $P_{plus}^{CA-TG}$ του *B.cereus* (Εικόνα 8*l*), ενώ εκδηλώνουν σαφή αλλαγή στη δομή των προτύπων τους σε ένα μόνο σημείο κατά μήκος του δημοσιευμένου κλώνου, το σημείο αυτό δεν είναι αρκούντως κοντά στο *ori*. Κατ' αντιστοιχία, παρατηρούμε ότι τα ακρότατα των αθροιστικών διαγραμμάτων για αυτές τις αποκλίσεις είναι μετατοπισμένα σε σχέση με το *ori* ($S_{plus}^{AG-CT}$: συνεχής πράσινη γραμμή στην Εικόνα 7*a*, $P_{plus}^{CA-TG}$: διακεκομμένη κόκκινη γραμμή στην Εικόνα 7*d*).

Αντιπαραβάλλοντας την Εικόνα 8 με τις Εικόνες 5 και 7, συμπεραίνουμε ότι *τα αθροιστικά διαγράμματα με μορφή V ή ανεστραμμένου V αντιστοιχούν σε αποκλίσεις των οποίων η μεταβολή της πόλωσης είναι, κατά κανόνα, στατιστικά*

*σημαντική*. Στα αθροιστικά γραφήματα το σημείο μεταβολής εμφανίζεται ως ακρότατο. Παρότι το σημείο αυτό διακρίνει, κατά προσέγγιση, μεταξύ οδηγού και συνοδού κλώνου, δεν ταυτίζεται απαραιτήτως με το *ori*. Επιπλέον, είναι δυνατόν να υπάρχουν και άλλα σημεία μεταβολής κατά μήκος του δημοσιευμένου κλώνου, που να μην εντοπίζονται ευκρινώς στα αθροιστικά διαγράμματα. Τέλος, τα πλέον ακανόνιστα στη μορφή τους αθροιστικά διαγράμματα ανταποκρίνονται στην απουσία σημείων μεταβολής, ιδίως όταν οι αντίστοιχες αποκλίσεις έχουν μικρή ένταση. Για το σύνολο της συλλογής μας, σχεδιάσαμε τα απλά (μη-αθροιστικά) διαγράμματα των μονονουκλεοτιδικών αποκλίσεων και όλων των δινουκλεοτιδικών αποκλίσεων, σε όρους τόσο παρατηρούμενων όσο και σταθμισμένων συχνοτήτων, μαζί με τα διαγράμματα των αντίστοιχων μοντέλων γραμμικής παλινδρόμησης (βλ. σχετικό link στο τέλος της Βιβλιογραφίας). Στην ακόλουθη ενότητα συνοψίζουμε τις σχετικές παρατηρήσεις.

#### **3.8.1.3** *Μελέτη των προτύπων ασυμμετρίας βάσει της στατιστικής σημαντικότητας των σημείων μεταβολής*

Για κάθε χρωμόσωμα, κατατάσσουμε το πρότυπο μίας δοσμένης απόκλισης σύμφωνα με τον αριθμό των βέλτιστων σημείων μεταβολής και τη θέση τους πάνω στον δημοσιευμένο κλώνο. Όταν δεν εντοπίζεται κανένα σημείο μεταβολής, τα μοτίβα (patterns) των αποκλίσεων θεωρούνται ως "ισόπεδα" (Εικόνα 8*m*). Στις περιπτώσεις εκείνες που εντοπίζονται ένα ή περισσότερα σημεία μεταβολής, διακρίνουμε τα μοτίβα των αποκλίσεων σε δύο διαφορετικούς τύπους· εάν τουλάχιστον ένα σημείο μεταβολής απέχει από το *ori* όχι περισσότερο από το 5% του μήκους του χρωμοσώματος, τα μοτίβα των αποκλίσεων χαρακτηρίζονται ως "σαφή" (Εικόνα 8*a*,*b*,*d*–*k*,*n*–*x*), ειδάλλως ταξινομούνται ως "ασαφή" (Εικόνα 8*c*,*l*). Τα αποτελέσματα παρατίθενται τον Πίνακα 9.

**ΠΙΝΑΚΑΣ 9. Ποσοστιαία κατάταξη των χρωμοσωμάτων, βάσει των προτύπων που εμφανίζουν τα μοντέλα γραμμικής παλινδρόμησης που περιγράφουν τις αποκλίσεις τους κατά μήκος του δημοσιευμένου κλώνου.**

| | | *σαφή* | *ασαφή* | *ισόπεδα* |
|---|---|---|---|---|
| **αποκλίσεις μόνο- και δι-νουκλεοτιδίων** | $S_{plus}^{A-T}$ | 57.35 | 21.47 | 21.18 |
| | $S_{plus}^{G-C}$ | 85.88 | 9.12 | 5.00 |
| | $S_{plus}^{AG-CT}$ | 56.76 | 18.53 | 24.71 |
| | $S_{plus}^{GA-TC}$ | 71.47 | 13.82 | 14.71 |
| | $S_{plus}^{GG-CC}$ | 80.88 | 13.53 | 5.59 |
| | $S_{plus}^{AA-TT}$ | 48.53 | 23.53 | 27.94 |
| | $S_{plus}^{AC-GT}$ | 81.18 | 10.29 | 8.53 |
| | $S_{plus}^{CA-TG}$ | 85.00 | 9.41 | 5.59 |
| **αποκλίσεις σταθμισμένων συχνοτήτων** | $P_{plus}^{AG-CT}$ | 33.24 | 21.47 | 45.29 |
| | $P_{plus}^{GA-TC}$ | 49.12 | 22.94 | 27.94 |
| | $P_{plus}^{GG-CC}$ | 41.47 | 22.06 | 36.47 |
| | $P_{plus}^{AA-TT}$ | 53.24 | 22.35 | 24.41 |
| | $P_{plus}^{AC-GT}$ | 44.71 | 17.35 | 37.94 |
| | $P_{plus}^{CA-TG}$ | 42.06 | 21.18 | 36.76 |

ΣΗΜΕΙΩΣΕΙΣ.- ***σαφή*** πρότυπα: τουλάχιστον ένα στατιστικώς σημαντικό σημείο μεταβολής (breakpoint) εντοπίζεται σε απόσταση από το *ori* ίση ή μικρότερη από το 5% του μήκους του χρωμοσώματος. ***ασαφή*** πρότυπα: εντοπίζεται τουλάχιστον ένα στατιστικώς σημαντικό σημείο μεταβολής (breakpoint), αλλά σε απόσταση από το *ori* μεγαλύτερη από το 5% του μήκους του χρωμοσώματος. ***ισόπεδα*** πρότυπα: δεν εντοπίζεται κανένα στατιστικώς σημαντικό σημείο μεταβολής.

Στον οδηγό κλώνο, η αλλαγή των συντελεστών γραμμικής παλινδρόμησης εκατέρωθεν του *ori* ισοδυναμεί με στατιστικά σημαντικές ασυμμετρίες μεταξύ οδηγού και συνοδού κλώνου, υποδηλώνοντας ότι *ο μηχανισμός της αντιγραφής επιδρά ισχυρά στη διαμόρφωση των ειδικών ανά κλώνο αποκλίσεων της σύστασης*

*του DNA*. Όπως προκύπτει από τον Πίνακα 9, περισσότεροι από το 85% των δημοσιευμένων κλώνων που εξετάζονται εμφανίζουν στατιστικώς σημαντικά σημεία μεταβολής των $S_{plus}^{G-C}$, τα οποία συμπίπτουν με το σημείο έναρξης της αντιγραφής (*σαφή* πρότυπα). Οι αποκλίσεις $S_{plus}^{A-T}$, εμφανίζουν μεγαλύτερη ετερογένεια ως προ τα πρότυπά τους. Το 21.47% των δημοσιευμένων κλώνων έχουν *ασαφή* πρότυπα $S_{plus}^{A-T}$, ποσοστό σημαντικά υψηλότερο από το 9.12% των κλώνων που έχουν *ασαφή* πρότυπα $S_{plus}^{G-C}$ αποκλίσεων. Οι $S_{plus}^{A-T}$ έχουν *ισόπεδα* πρότυπα στο 21.18% των κλώνων, ενώ περίπου το 57% αυτών εμφανίζουν *σαφή* πρότυπα, ποσοστό σημαντικά μικρότερο συγκριτικά με το αντίστοιχο των $S_{plus}^{G-C}$. Συνεπώς, οι αποκλίσεις G-C καθορίζονται σε μεγάλο βαθμό από ασύμμετρες μεταλλακτικές πιέσεις που επάγονται από την αντιγραφή. Η επίδραση της αντιγραφής είναι σημαντική και στη διαμόρφωση των αποκλίσεων A-T, αν και σε αυτή την περίπτωση ο αριθμός των χρωμοσωμάτων με *σαφή* πρότυπα ασυμμετριών μεταξύ οδηγού και συνοδού κλώνου είναι αισθητά μικρότερος. Τα ευρήματά μας έρχονται σε συμφωνία με προγενέστερες μελέτες, οι οποίες αποδίδουν τον εμπλουτισμό του οδηγού κλώνου σε κατάλοιπα G, πρωτίστως, και T, δευτερευόντως, στην ασύμμετρη δράση της αντιγραφής (Lobry & Sueoka 2002).

Σύμφωνα με τον Πίνακα 9, η αντιστροφή της πόλωσης των ειδικών ανά κλώνο ασυμμετριών εκατέρωθεν του *ori* είναι στατιστικά σημαντική και στο επίπεδο των δινουκλεοτιδικών συχνοτήτων. Συγκεκριμένα, ο προσανατολισμός των δινουκλεοτιδίων που περιέχουν κατάλοιπα G ή C είναι έντονα πολωμένος μεταξύ του οδηγού και του συνοδού κλώνου, με την εξαίρεση του ζεύγους AG/CT. Οι αντίστοιχες αποκλίσεις ακολουθούν *σαφή* πρότυπα κατανομής στον δημοσιευμένο κλώνο, με ποσοστά που κυμαίνονται από 71.47% έως 85.00% της συλλογής μας. Σε αντίθεση με αυτές τις καλά καθορισμένες κατανομές των αποκλίσεων, οι $S_{plus}^{AG-CT}$ και $S_{plus}^{AT-TT}$ εμφανίζουν *σαφή* πρότυπα αλλαγής της διεύθυνσής τους εκατέρωθεν του *ori* μόλις στο 56.76% και 48.53%, αντίστοιχα, των εξεταζόμενων χρωμοσωμάτων. Επί πλέον, οι $S_{plus}^{AG-CT}$ και $S_{plus}^{AT-TT}$ έχουν *ισόπεδα* πρότυπα στο 24.71% και 27.94%, αντίστοιχα, της συλλογής, ενώ, αντίθετα, για τις υπόλοιπες αποκλίσεις δινουκλεοτιδίων *ισόπεδα* πρότυπα παρατηρούνται σε ποσοστά που κυμαίνονται από 5.59% έως το πολύ 14.71% της συλλογής. Συνολικά, η ποσοστιαία κατανομή των αποκλίσεων στα τρία πρότυπα ασυμμετριών (*σαφή*, *ασαφή*, *ισόπεδα*) δεν διαφοροποιείται σημαντικά όταν εστιάζουμε από τα μονονουκλεοτίδια στα δινουκλεοτίδια. Συνεπώς, *οι ασυμμετρίες μεταξύ οδηγού και συνοδού κλώνου στο επίπεδο των δινουκλεοτιδίων αποτελούν γενικό χαρακτηριστικό των βακτηριακών*

*γονιδιωμάτων, όπως συμβαίνει και με τις αποκλίσεις των μονονουκλεοτιδίων.*

Οι δινουκλεοτιδικές αποκλίσεις ακολουθούν *σαφή* πρότυπα ασυμμετρίας σε περισσότερα χρωμοσώματα, όταν μελετώνται σε όρους παρατηρούμενων αντί σταθμισμένων συχνοτήτων. Εξαίρεση αποτελεί το ζεύγος AA/TT, το μόνο που δεν περιέχει κατάλοιπα G ή C, όπου παρατηρείται η αντίστροφη τάση. Κατά προσέγγιση, ένα στα δύο χρωμοσώματα έχει *σαφή* πρότυπα αποκλίσεων σταθμισμένων συχνοτήτων, και συνεπώς, σε αυτές τις αλληλουχίες οι συσχετίσεις των 1ης τάξης γειτονικών βάσεων είναι έντονα πολωμένες μεταξύ οδηγού και συνοδού κλώνου. Όπως και στην περίπτωση των παρατηρούμενων συχνοτήτων, η συμπεριφορά των σταθμισμένων συχνοτήτων του ζεύγους AG/CT δεν ευθυγραμμίζεται με την γενικότερη τάση, καθώς μόνο το 33.24% των εξεταζόμενων αλληλουχιών έχει *σαφή* πρότυπα $P_{plus}^{AG-CT}$, ενώ το 45.29% έχει *ισόπεδα* πρότυπα.

Όταν εστιάζουμε αποκλειστικά στα γονιδιώματα εκείνα που ανήκουν στα Firmicutes (βλ. Παράρτημα, Πίνακας Ι) τα *σαφή* πρότυπα αποκλίσεων είναι ακόμα πιο διαδεδομένα, σε σχέση με το σύνολο της συλλογής. Έτσι, το 100% των Fimicutes που εξετάσαμε έχουν *σαφή* πρότυπα $S_{plus}^{G-C}$, $S_{plus}^{GA-TC}$ και $S_{plus}^{GG-CC}$. Επίσης, όλες οι αποκλίσεις σταθμισμένων συχνοτήτων, εκτός του ζεύγους AC/GT, έχουν *σαφή* πρότυπα σε περισσότερα από τα μισά χρωμοσώματα των Firmicutes, με τα σχετικά ποσοστά να κυμαίνονται μεταξύ 57.81% και 71.88%. Όπως ήδη αναφέραμε (βλ. Πίνακες 5,6,8), το συγκεκριμένο φύλο εμφανίζει ιδιαιτερότητες τόσο ως προς τη διεύθυνση της πόλωσης των ασυμμετριών μεταξύ οδηγού και συνοδού κλώνου όσο και ως προς την έντασή τους. Αξίζει να σημειωθεί ότι οι $P_{plus}^{AC-GT}$, οι οποίες οργανώνονται σε *σαφή* πρότυπα μόλις στο 35.94% αυτού του φύλου, είναι πολωμένες προς την αντίθετη κατεύθυνση σε σχέση με τις αντίστοιχες αποκλίσεις των εκτός-Firmicutes, με το 86.00% των Firmicutes να έχουν διαγράμματα $P_{plus}^{AC-GT}$ τύπου-V (Πίνακας 8).

### **3.8.2** CDS-συρραφές

Στην αμέσως προηγούμενη ενότητα (3.8.1) καταδείξαμε την στατιστικά σημαντική συσχέτιση των ειδικών ανά κλώνο συμμετριών με τον μηχανισμό της

αντιγραφής. Η επίδραση που ασκεί η αντιγραφή του DNA στην οργάνωση *σαφών* προτύπων ασυμμετρίας μεταξύ οδηγού και συνοδού κλώνου πηγάζει από δύο βασικές διαδικασίες, των οποίων τα αποτελέσματα επικαλύπτονται. *Αφενός, ο ίδιος ο μηχανισμός της αντιγραφής μπορεί να επάγει ασύμμετρα πρότυπα υποκατάστασης σε οδηγό και συνοδό κλώνο*. Ως συνέπεια, στην κλίμακα ολόκληρου του γονιδιώματος σχηματίζονται πρότυπα αποκλίσεων τα οποία συσχετίζονται με τον αναδιπλασιασμό του DNA και αλλάζουν φορά στην περιοχή του *ori* και του *ter*. *Αφετέρου, η ασύμμετρη δομή της διχάλας της αντιγραφής προκαλεί πόλωση της διάταξης των κωδικών κλώνων των γονιδίων κατά μήκος του οδηγού κλώνου του χρωμοσώματος*, προσφέροντας εξελικτικό πλεονέκτημα στα βακτήρια, καθώς έτσι αντιγραφή και μεταγραφή πραγματοποιούνται συγγραμμικά και αποφεύγεται η κατά μέτωπο σύγκρουση των DNA- και RNA-πολυμερασών. Στο πλαίσιο αυτό, αποκλίσεις οι οποίες συγκροτούνται αρχικά στην κλίμακα των κωδικών κλώνων, αναδεικνύονται στην κλίμακα ολόκληρου του χρωμοσώματος. Οι δύο αυτές διαδικασίες δρουν παράλληλα, με την αντιγραφή να διαμορφώνει τις αποκλίσεις με τρόπο άμεσο στην πρώτη περίπτωση και έμμεσο στη δεύτερη (Necşulea & Lobry 2007). Για μία αναλυτικότερη παρουσίαση των σχετικών μηχανισμών, βλ. ενότητα 1.4.2.2 και 1.4.1.4.

Σύμφωνα με τα παραπάνω, *σαφή* πρότυπα αποκλίσεων που οργανώνονται γύρω από το *ori* μπορούν να ειδωθούν ως το αποτέλεσμα δύο κύριων συνιστωσών: (*α*) των ασύμμετρων υποκαταστάσεων που επάγονται από την αντιγραφή, και (*β*) των ασύμμετρων υποκαταστάσεων που σχετίζονται ειδικά με τις κωδικές αλληλουχίες (Morton & Morton 2007).

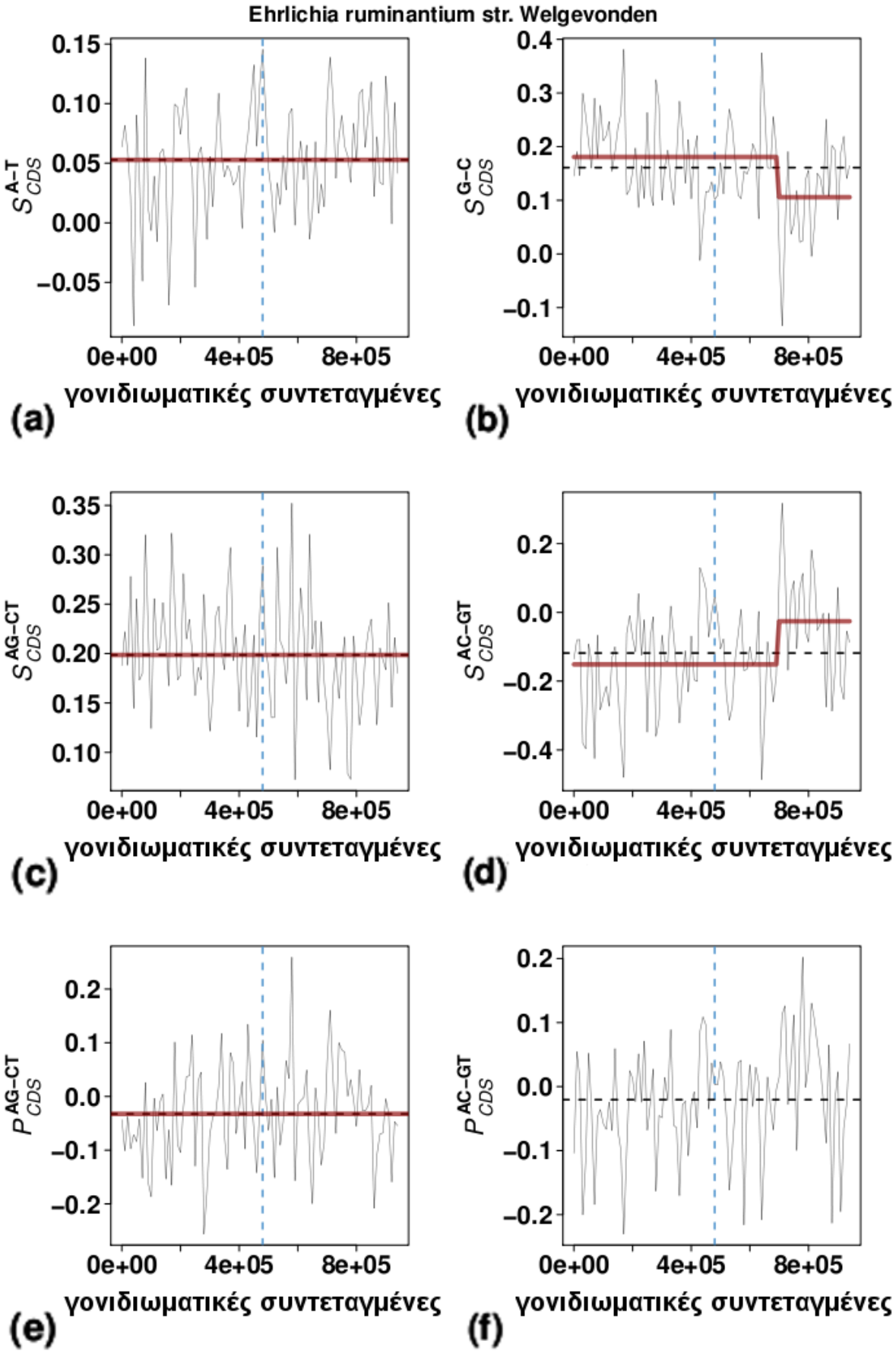
### **3.8.2.1** *Διαχωρισμός των CDS-συζευγμένων αποκλίσεων από τις αποκλίσεις που σχετίζονται με την αντιγραφή*

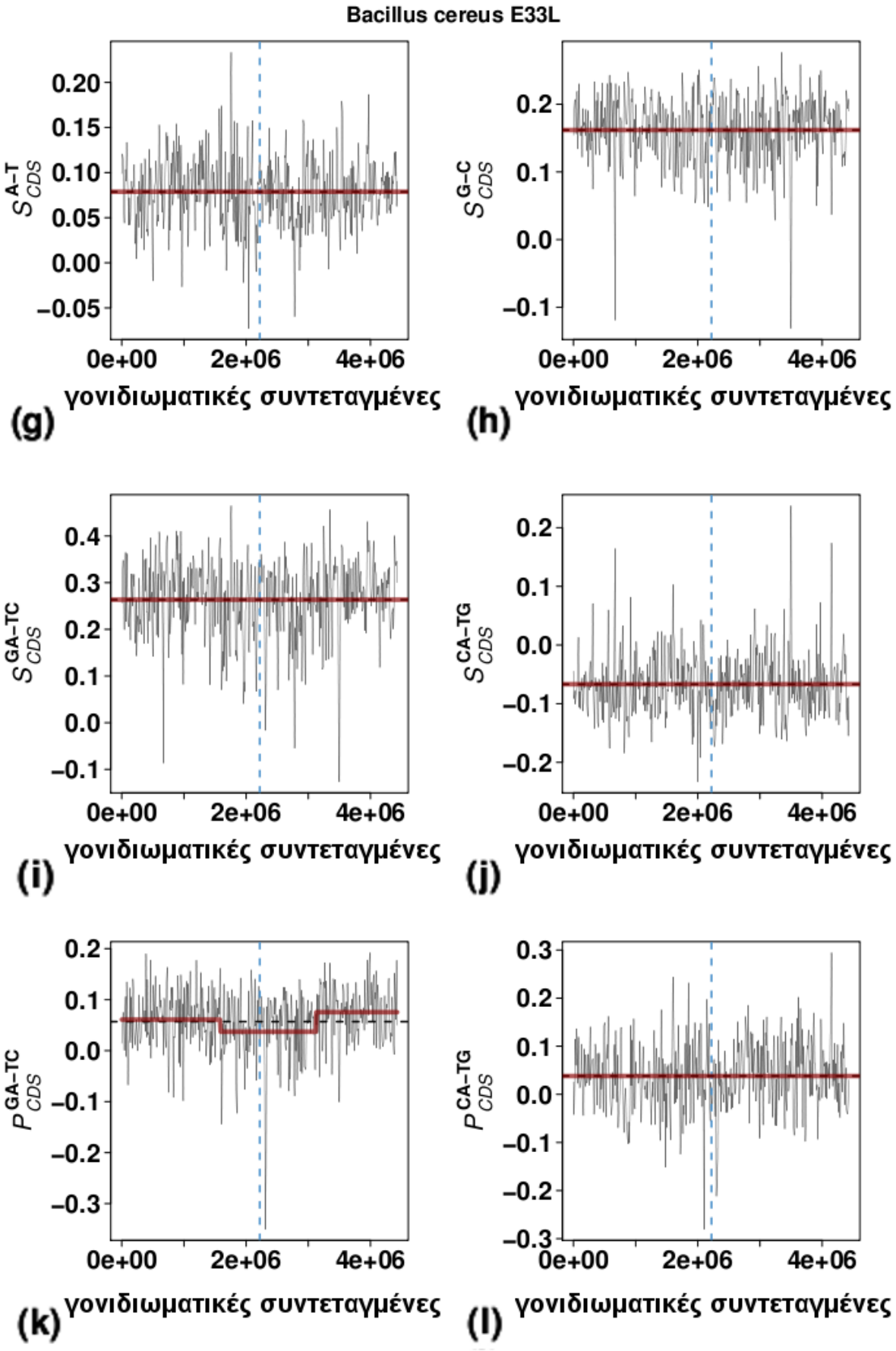
Προκειμένου να αξιολογήσουμε τη στατιστική σημαντικότητα των CDS-συζευγμένων αποκλίσεων, μελετήσαμε τη σύσταση των CDS-συρραφών (βλ. ενότητα 2.2). Οι αλληλουχίες αυτές κατασκευάζονται με τη διαδοχική συνένωση των κωδικών αλληλουχιών. Συγκεκριμένα, οι κωδικοί κλώνοι όλων των γονιδίων λαμβάνονται κατά τη φορά της μεταγραφής τους, ανεξαρτήτως του εάν βρίσκονται στον οδηγό ή το συνοδό κλώνο. Συνεπώς, κατά μήκος των CDS-συρραφών, τμήματα του οδηγού κλώνου, τα οποία συντίθενται κατά τη φορά ανάπτυξης της διχάλας της αντιγραφής, διαδέχονται εναλλάξ τμήματα του συνοδού κλώνου, τα οποία συντίθενται κατά την αντίθετη φορά. Έστω ότι οι

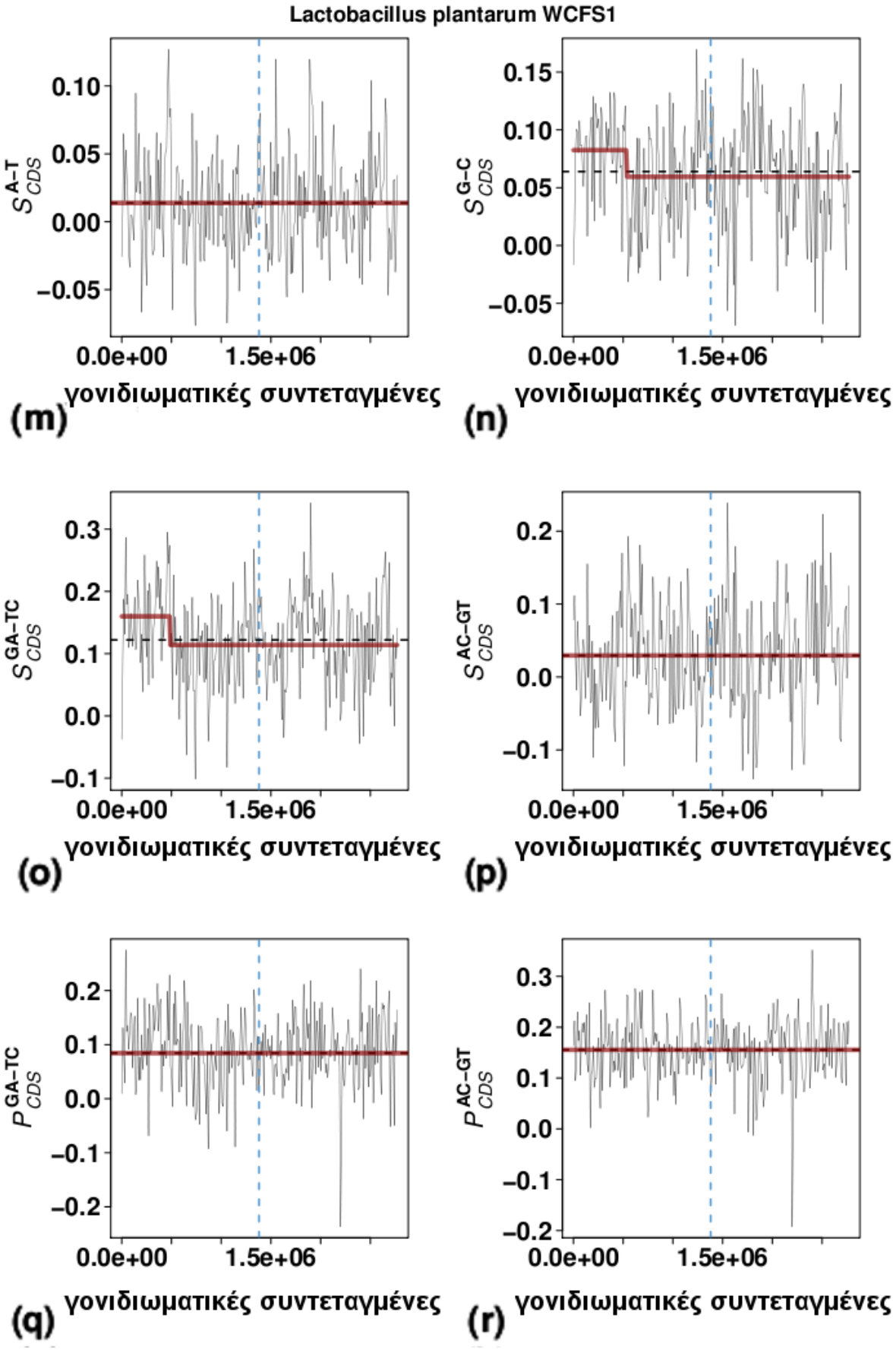
επαγόμενες από την αντιγραφή υποκαταστάσεις οδηγούν σε περίσσεια καταλοίπων G έναντι C στον οδηγό κλώνο, και συνεπώς σε έλλειμμα G έναντι C στο συνοδό κλώνο. Όταν συνενώνονται διαδοχικά τμήματα των δύο αυτών κλώνων σε μία τεχνητή αλληλουχία, όπως συμβαίνει στην περίπτωση των CDS-συρραφών, οι αποκλίσεις G-C που επάγονται από την αντιγραφή αναμένεται να αλληλοαναιρούνται εντός ενός παραθύρου μερικών χιλιάδων βάσεων. Το ίδιο αναμένεται να συμβεί για το σύνολο των αποκλίσεων που επάγει η αντιγραφή. Συγχρόνως, η αναδιάταξη των γονιδίων στις CDS-συρραφές αναδεικνύει και τονίζει τα πρότυπα των CDS-συζευγμένων αποκλίσεων, δηλαδή των ασυμμετριών μεταξύ κωδικού και μεταγραφόμενου κλώνου. Οι ασυμμετρίες των CDS-συρραφών δεν διακρίνουν οδηγό και συνοδό κλώνο, και συνεπώς δεν περιμένουμε να οργανώνονται γύρω από το *ori*.

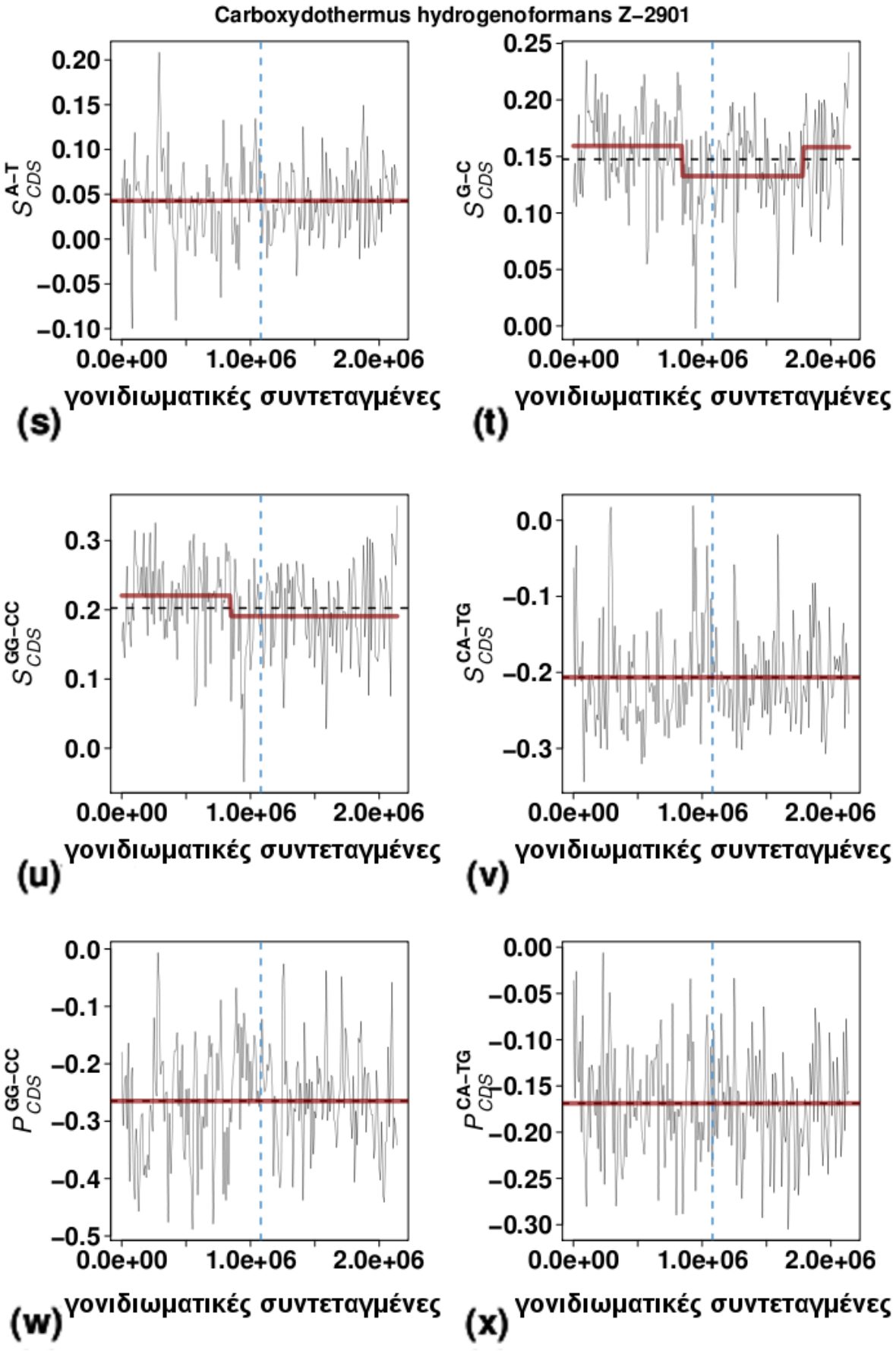
#### **3.8.2.2** *Γραφική απεικόνιση*

Εφαρμόσαμε τον αλγόριθμο των Zeileis *et al.* (2003, 2010) στις CDS-συζευγμένες αποκλίσεις, όπως προηγουμένως κάναμε στην περίπτωση των αποκλίσεων του δημοσιευμένου κλώνου. Κατ' αναλογία με την Εικόνα 8 της προηγούμενης ενότητας (3.8.1), στην Εικόνα 9 παρουσιάζουμε τα απλά (μη-αθροιστικά) διαγράμματα των μονονουκλεοτιδικών αποκλίσεων και των αποκλίσεων των δινουκλεοτιδίων και των σταθμισμένων τους συχνοτήτων κατά μήκος των CDS-συρραφών, για το σύνολο των περιπτώσεων που εμφανίζονται στις Εικόνες 5 και 7, στις οποίες απεικονίζονται τα αθροιστικά διαγράμματα των αντίστοιχων αποκλίσεων κατά μήκος του δημοσιευμένου κλώνου. Εκτός από τα διαγράμματα των αποκλίσεων, σε κάθε γράφημα της Εικόνας 9 σχεδιάσαμε και τα διαγράμματα των μοντέλων γραμμικής παλινδρόμησης που περιγράφουν αυτές τις αποκλίσεις (οριζόντιες και τεθλασμένες κόκκινες γραμμές).


Ehrlichia ruminantium str. Welgevonden
$S^{A-T}_{CDS}$
$S^{G-C}_{CDS}$
$S^{AG-CT}_{CDS}$
$S^{AC-GT}_{CDS}$
$P^{AG-CT}_{CDS}$
$P^{AC-GT}_{CDS}$
0e+00
4e+05
8e+05
γονιδιωματικές συντεταγμένες
(a)
(b)
(c)
(d)
(e)
(f)


Bacillus cereus E33L
$S^{A-T}_{CDS}$
0.20
0.15
0.10
0.05
0.00
−0.05
0e+00
2e+06
4e+06
γονιδιωματικές συντεταγμένες
(g)
$S^{G-C}_{CDS}$
0.2
0.1
0.0
−0.1
0e+00
2e+06
4e+06
γονιδιωματικές συντεταγμένες
(h)
$S^{GA-TC}_{CDS}$
0.4
0.3
0.2
0.1
0.0
−0.1
0e+00
2e+06
4e+06
γονιδιωματικές συντεταγμένες
(i)
$S^{CA-TG}_{CDS}$
0.2
0.1
0.0
−0.1
−0.2
0e+00
2e+06
4e+06
γονιδιωματικές συντεταγμένες
(j)
$P^{GA-TC}_{CDS}$
0.2
0.1
0.0
−0.1
−0.2
−0.3
0e+00
2e+06
4e+06
γονιδιωματικές συντεταγμένες
(k)
$P^{CA-TG}_{CDS}$
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0e+00
2e+06
4e+06
γονιδιωματικές συντεταγμένες
(l)


Lactobacillus plantarum WCFS1
$S^{A-T}_{CDS}$
0.10
0.05
0.00
−0.05
0.0e+00
1.5e+06
γονιδιωματικές συντεταγμένες
(m)
$S^{G-C}_{CDS}$
0.15
0.10
0.05
0.00
−0.05
0.0e+00
1.5e+06
γονιδιωματικές συντεταγμένες
(n)
$S^{GA-TC}_{CDS}$
0.3
0.2
0.1
0.0
−0.1
0.0e+00
1.5e+06
γονιδιωματικές συντεταγμένες
(o)
$S^{AC-GT}_{CDS}$
0.2
0.1
0.0
−0.1
0.0e+00
1.5e+06
γονιδιωματικές συντεταγμένες
(p)
$P^{GA-TC}_{CDS}$
0.2
0.1
0.0
−0.1
−0.2
0.0e+00
1.5e+06
γονιδιωματικές συντεταγμένες
(q)
$P^{AC-GT}_{CDS}$
0.3
0.2
0.1
0.0
−0.1
−0.2
0.0e+00
1.5e+06
γονιδιωματικές συντεταγμένες
(r)


Carboxydothermus hydrogenoformans Z−2901
$S^{A-T}_{CDS}$
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
0.0e+00
1.0e+06
2.0e+06
γονιδιωματικές συντεταγμένες
(s)
$S^{G-C}_{CDS}$
0.25
0.20
0.15
0.10
0.05
0.00
0.0e+00
1.0e+06
2.0e+06
γονιδιωματικές συντεταγμένες
(t)
$S^{GG-CC}_{CDS}$
0.3
0.2
0.1
0.0
0.0e+00
1.0e+06
2.0e+06
γονιδιωματικές συντεταγμένες
(u)
$S^{CA-TG}_{CDS}$
0.0
−0.1
−0.2
−0.3
0.0e+00
1.0e+06
2.0e+06
γονιδιωματικές συντεταγμένες
(v)
$P^{GG-CC}_{CDS}$
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
0.0e+00
1.0e+06
2.0e+06
γονιδιωματικές συντεταγμένες
(w)
$P^{CA-TG}_{CDS}$
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
−0.30
0.0e+00
1.0e+06
2.0e+06
γονιδιωματικές συντεταγμένες
(x)

**Εικόνα 9.** Ασυμμετρίες κατά μήκος των CDS-συρραφών. Διαγράμματα των μονονουκλεοτιδικών αποκλίσεων και επιλεγμένων δινουκλεοτιδικών αποκλίσεων, σε όρους τόσο παρατηρούμενων όσο και σταθμισμένων συχνοτήτων (κυματιστές γκρι γραμμές), μαζί με τα αντίστοιχα μοντέλα γραμμικής παλινδρόμησης, που περιγράφουν αυτές τις αποκλίσεις (οριζόντιες και τεθλασμένες κόκκινες γραμμές). Στο γράφημα (*f*) το μοντέλο που περιγράφει τις αποκλίσεις δεν είναι στατιστικά σημαντικό (p-value > 0.01, όπως υπολογίζεται από τον σχετικό αλγόριθμο) και για το λόγο αυτό δεν παρουσιάζεται το διάγραμμά του.
Οι αποκλίσεις υπολογίζονται κατά μήκος των CDS-συρραφών τεσσάρων βακτηρίων, εντός διαδοχικών, μη-επικαλυπτόμενων κυλιόμενων παραθύρων μήκους $10^4$ bps. Η κατακόρυφη, διακεκομμένη μπλε γραμμή δηλώνει το σημείο έναρξης της αντιγραφής (*ori*). Τόσο αριστερά όσο και δεξιά του *ori*, τμήματα του οδηγού κλώνου διαδέχονται εναλλάξ τμήματα του συνοδού κλώνου. Η οριζόντια, διακεκομμένη μαύρη γραμμή δηλώνει την μέση τιμή των αποκλίσεων. Σε αντίθεση με ότι παρατηρούμε στον δημοσιευμένο κλώνο, στις CDS-συρραφές η μέση απόκλιση είναι κατά κανόνα διάφορη του μηδενός.

Δεκάξι από τα εικοσιτέσσερα γραφήματα της Εικόνας 9 δεν εμφανίζουν κανένα σημείο μεταβολής, καθώς οι αντίστοιχες αποκλίσεις συγκροτούν *ισόπεδα* πρότυπα. Μεταξύ αυτών συγκαταλέγονται γραφήματα όπου απεικονίζονται τόσο μονονουκλεοτιδικές αποκλίσεις, όσο και αποκλίσεις των δινουκλεοτιδίων και των σταθμισμένων τους συχνοτήτων. Με την εξαίρεση τριών μόνο περιπτώσεων (*E.ruminantium*: $S^{AG-CT}$, *B.cereus*: $P^{CA-TG}$, *L.plantarum*: $S^{A-T}$), όλες οι αποκλίσεις με *ισόπεδα* πρότυπα κατά μήκος των CDS-συρραφών ακολουθούν *σαφή* πρότυπα στον δημοσιευμένο κλώνο (πρβ. Εικόνα 8 και 9). Συνεπώς, ακόμα και όταν οι ασύμμετρες υποκαταστάσεις μεταξύ οδηγού και συνοδού κλώνου συμβάλλουν καθοριστικά στη διαμόρφωση του προτύπου των αποκλίσεων (βλ. Εικόνα 8 και Πίνακα 9), τα προφίλ αυτών των αποκλίσεων μπορεί να παραμένουν εν πολλοίς αμετάβλητα κατά μήκος των κωδικών κλώνων, είτε τα αντίστοιχα γονίδια βρίσκονται στον οδηγό είτε στο συνοδό κλώνο του χρωμοσώματος. Συμπερασματικά, *ασύμμετρες υποκαταστάσεις μπορεί να επάγονται εν μέρει από την αντιγραφή και παράλληλα να είναι εν μέρει συζευγμένες με τις CDSs. Οι αποκλίσεις που εκδηλώνονται στην κλίμακα του χρωμοσώματος προκύπτουν ως συνισταμένη των δύο αυτών επί μέρους ασυμμετριών.*

Επτά από τα γραφήματα που παρουσιάζονται στην Εικόνα 9 απεικονίζουν *ασαφή* πρότυπα αποκλίσεων. Τα αντίστοιχα σημεία μεταβολής εντοπίζονται σε θέσεις απομακρυσμένες από την περιοχή του *ori*. Οι CDS-συζευγμένες αποκλίσεις δεν έχουν *σαφή* πρότυπα, οργανωμένα γύρω από την περιοχή του *ori*,

σε κανένα από τα γράφημα της Εικόνας 9.

Ιδιαίτερο ενδιαφέρον παρουσιάζει το γράφημα (*f*), Εικόνα 9, όπου αναπαριστώνται οι αποκλίσεις $P_{CDS}^{AC-GT}$ του *Ehrlichia ruminantium*. Η μέση τιμή των αποκλίσεων του ζεύγους AC/GT, σε όρους σταθμισμένων συχνοτήτων, είναι περίπου μηδενική (βλ. οριζόντια, διακεκομμένη μαύρη γραμμή), ενώ δεν εμφανίζεται κανένα σημείο μεταβολής (*ισόπεδο* πρότυπο). Ωστόσο, το μοντέλο που περιγράφει αυτές τις αποκλίσεις δεν είναι στατιστικά σημαντικό (*p*-value > 0.01, όπως υπολογίζεται από τον σχετικό αλγόριθμο) και συνεπώς οι υπό μελέτη $P_{CDS}^{AC-GT}$ έχουν ακανόνιστο προφίλ στις CDS-συρραφές. Στο αντίστοιχο γράφημα (*f*) της Εικόνας 8 οι $P_{CDS}^{AC-GT}$ ακολουθούν ένα *σαφές* πρότυπο αποκλίσεων κατά μήκος του δημοσιευμένου κλώνου, διακρίνοντας μεταξύ οδηγού και συνοδού κλώνου. Λαμβάνοντας μαζί αυτές τις παρατηρήσεις, συμπεραίνουμε ότι οι αποκλίσεις των σταθμισμένων συχνοτήτων των AC/GT στο χρωμόσωμα του *Ehrlichia ruminantium* διαμορφώνονται από τις ασυμμετρίες μεταξύ οδηγού και συνοδού κλώνου και όχι μεταξύ κωδικών και μεταγραφόμενων κλώνων. Συνεπώς, *σε ορισμένα γονιδιώματα η ασύμμετρη δράση της αντιγραφής αρκεί για την ανάδειξη ειδικών ανά κλώνο αποκλίσεων στο επίπεδο των υποκαταστάσεων που εξαρτώνται από τις γειτονικές βάσεις.*

### **3.8.2.3** *Μελέτη των προτύπων ασυμμετρίας βάσει της στατιστικής σημαντικότητας των σημείων μεταβολής*

Όπως και στην ενότητα 3.8.1.3, κατατάσσουμε τις αποκλίσεις κατά μήκος των CDS-συρραφών σε μία από τις τρεις κατηγορίες προτύπων (*σαφή*, *ασαφή*, *ισόπεδα*), για το σύνολο της συλλογής μας. Στον ακόλουθο πίνακα συνοψίζουμε τα σχετικά αποτελέσματα.

**ΠΙΝΑΚΑΣ 10. Ποσοστιαία κατάταξη των χρωμοσωμάτων, βάσει των προτύπων που εμφανίζουν τα μοντέλα γραμμικής παλινδρόμησης που περιγράφουν τις αποκλίσεις τους κατά μήκος των CDS-συρραφών.**

| | | *σαφή* | *ασαφή* | *ισόπεδα* |
|---|---|---|---|---|
| **αποκλίσεις μόνο- και δι-νουκλεοτιδίων** | $S_{CDS}^{A-T}$ | 0.88 | 21.47 | 77.65 |
| | $S_{CDS}^{G-C}$ | 1.18 | 25.88 | 72.94 |
| | $S_{CDS}^{AG-CT}$ | 0.88 | 23.82 | 75.29 |
| | $S_{CDS}^{GA-TC}$ | 0.88 | 20.29 | 78.82 |
| | $S_{CDS}^{GG-CC}$ | 1.47 | 21.18 | 77.35 |
| | $S_{CDS}^{AA-TT}$ | 1.76 | 26.18 | 72.06 |
| | $S_{CDS}^{AC-GT}$ | 1.18 | 16.18 | 82.65 |
| | $S_{CDS}^{CA-TG}$ | 0.88 | 18.24 | 80.88 |
| **αποκλίσεις σταθμισμένων συχνοτήτων** | $P_{CDS}^{AG-CT}$ | 1.76 | 15.29 | 82.94 |
| | $P_{CDS}^{GA-TC}$ | 0 | 15.29 | 84.71 |
| | $P_{CDS}^{GG-CC}$ | 0 | 20.29 | 79.71 |
| | $P_{CDS}^{AA-TT}$ | 1.47 | 25.00 | 73.53 |
| | $P_{CDS}^{AC-GT}$ | 1.47 | 23.53 | 75.00 |
| | $P_{CDS}^{CA-TG}$ | 0.88 | 22.94 | 76.18 |

ΣΗΜΕΙΩΣΕΙΣ.- ***σαφή*** πρότυπα: τουλάχιστον ένα στατιστικώς σημαντικό σημείο μεταβολής (breakpoint) εντοπίζεται σε απόσταση από το *ori* ίση ή μικρότερη από το 5% του μήκους του χρωμοσώματος. ***ασαφή*** πρότυπα: εντοπίζεται τουλάχιστον ένα στατιστικώς σημαντικό σημείο μεταβολής (breakpoint), αλλά σε απόσταση από το *ori* μεγαλύτερη από το 5% του μήκους του χρωμοσώματος. ***ισόπεδα*** πρότυπα: δεν εντοπίζεται κανένα στατιστικώς σημαντικό σημείο μεταβολής.

Σύμφωνα με τον Πίνακα 10, όλες οι CDS-συζευγμένες αποκλίσεις ακολουθούν *ισόπεδα* πρότυπα σε περισσότερες από το 72% των εξεταζόμενων περιπτώσεων. Αντίθετα, *σαφή* πρότυπα γύρω από την περιοχή του *ori* εμφανίζονται το πολύ σε 6 CDS-συρραφές, ανάλογα με τον τύπο της εκάστοτε

απόκλισης. Συνεπώς, στις περισσότερες CDS-συρραφές οι αποκλίσεις μπορούν να περιγραφούν από μία (και μοναδική) σταθερή, γραμμική σχέση έναντι των συντεταγμένων των αλληλουχιών (βλ. Εικόνα 9). Το γεγονός αυτό υποδηλώνει ότι *η κατεύθυνση της μεταγραφής των γονιδίων αποτελεί καθοριστικό παράγοντα, επαρκή για τον προσδιορισμό των ασυμμετριών στις κωδικές περιοχές (CDSs)* (Nikolaou & Almirantis 2005). Οι ασυμμετρίες αυτές αποδίδονται, μεταξύ άλλων, σε μεταλλάξεις σχετιζόμενες με τη μεταγραφή (transcription-associated mutations, TAM) (Francino et al. 1996, Francino & Ochman 1997) καθώς επίσης και στις προτιμήσεις στη χρήση κωδικονίων (codon usage preferences) (Ikemura 1981, Gouy & Gautier 1982, Ikemura 1982, Bulmer 1991b, Xia 1998). Σε ακόλουθες ενότητες (3.11, 3.12) της παρούσας μελέτης εξετάζουμε τη συμβολή τέτοιων παραγόντων στη διαμόρφωση των CDS-συζευγμένων αποκλίσεων.

Κάθε μεταλλακτική ή επιλεκτική διαδικασία που διακρίνει μεταξύ του κωδικού και του μεταγραφόμενου κλώνου των γονιδίων μπορεί να επιδράσει καθοριστικά στην εξέλιξη των αλληλουχιών DNA, εισάγοντας ειδικές ανά κλώνο αποκλίσεις στη σύσταση των χρωμοσωμάτων. Όπως προκύπτει από την ανάλυση της κατανομής των απόλυτων τιμών των αποκλίσεων (Πίνακας 5) *η ένταση των ασυμμετριών είναι εν γένει μεγαλύτερη μεταξύ κωδικών και μεταγραφόμενων κλώνων, από ότι μεταξύ οδηγού και συνοδού κλώνου*. Προκειμένου να αξιολογήσουμε την συμβολή των CDS-συζευγμένων αποκλίσεων στη διαμόρφωση του κωδικού περιεχομένου (coding content) των βακτηριακών χρωμοσωμάτων, στις ακόλουθες ενότητες εστιάζουμε τη μελέτη μας στις CDS-συρραφές.

## **3.9** *Συσχέτιση των ειδικών ανά κλώνο ασυμμετριών με τη φυλογένεση των βακτηρίων*

Στα μέσα της δεκαετίας του '90 οι Karlin και Burge (1995) διαπίστωσαν ότι οι σταθμισμένες συχνότητες των δινουκλεοτιδίων συγκροτούν ένα σύνολο τιμών το οποίο επιτρέπει τη διάκριση μεταξύ αλληλουχιών DNA που προέρχονται από διαφορετικούς οργανισμούς. Από αυτή την άποψη, το σύνολο των

δινουκλεοτιδικών σταθμισμένων συχνοτήτων ενός χρωμοσώματος μπορεί να θεωρηθεί ως μία χαρακτηριστική *γονιδιωματική υπογραφή*. Στη σχετική βιβλιογραφία οι γονιδιωματικές υπογραφές ορίζονται ως οι συμμετρικές προς τους δύο κλώνους DNA σταθμισμένες συχνότητες δινουκλεοτιδίων ($\rho^*$) (Karlin & Mrázek 1997, Karlin 1998, Campbell et al. 1999). Συνεπώς, οι αναλύσεις στις οποίες χρησιμοποιούνται οι γονιδιωματικές υπογραφές δεν λαμβάνουν υπόψη την ασυμμετρία των κλώνων, σε όρους δινουκλεοτιδικών σταθμισμένων συχνοτήτων (βλ. ενότητα 2.4). Μάλιστα, στη μελέτη των Mrázek και Karlin (1998) αναφέρεται πως οι τιμές των μη-συμμετρικών σταθμισμένων συχνοτήτων ($\rho$) παραμένουν σταθερές στους δύο κλώνους του DNA και συνεπώς οι συμμετρικοί $\rho^*$ περιγράφουν ικανοποιητικά τις συσχετίσεις $1^{ης}$ τάξης γειτονικών βάσεων, τόσο στο επίπεδο του δίκλωνου μορίου όσο και στον κάθε κλώνο ξεχωριστά (βλ. ενότητα 1.8.3.3).

### **3.9.1** *Τοπολογική ανάλυση των κλαδογραμμάτων αποκλίσεων*

Στις προηγούμενες ενότητες (3.3–3.8) της παρούσας μελέτης αποδείξαμε ότι οι $\rho$ εμφανίζουν στατιστικά σημαντικές ασυμμετρίες μεταξύ των αντιστρόφως συμπληρωματικών κλώνων, σε αντίθεση με την έως τώρα καθιερωμένη αντίληψη. Δεδομένου ότι οι συμμετρικοί $\rho^*$ περιέχουν φυλογενετική πληροφορία, εξετάσαμε σε τι βαθμό οι ασυμμετρίες των $\rho$ μπορούν επίσης να ανιχνεύσουν την εξελικτική πορεία διαφορετικών οργανισμών.

Για κάθε ζεύγος οργανισμών της συλλογής μας, συγκρίναμε την κατανομή του συνόλου των αποκλίσεων των $\rho$ ($V^{RA}$) κατά μήκος των αντίστοιχων CDS-συρραφών. Για τις συγκρίσεις μας χρησιμοποιήσαμε την συμμετρική Kullback–Leibler (KL) απόκλιση (Kullback & Leibler 1951). Βάσει των συμμετρικών KL-αποκλίσεων, κατασκευάσαμε κλαδογράμματα για κάθε φύλο ή κλάση που αντιπροσωπεύονται στη συλλογή μας (*κλαδογράμματα αποκλίσεων*, Εικόνα 10). Ακολούθως, συγκρίναμε τα κλαδογράμματα αποκλίσεων των σταθμισμένων συχνοτήτων με τα αντίστοιχα ταξινομικά δέντρα και λάβαμε τα επί τοις εκατό ποσοστά τοπολογικής ομοιότητας των δέντρων ("τοπολογική βαθμολογία"). Για λόγους σύγκρισης με ένα ήδη εν χρήσει μέτρο, περιλάβαμε και την τοπολογική βαθμολογία που αντιστοιχεί στα κλαδογράμματα των γονιδιωματικών υπογραφών. Επίσης, για κάθε φύλο ή κλάση που αντιπροσωπεύονται στη συλλογή μας, κατασκευάσαμε κλαδογράμματα βάσει της

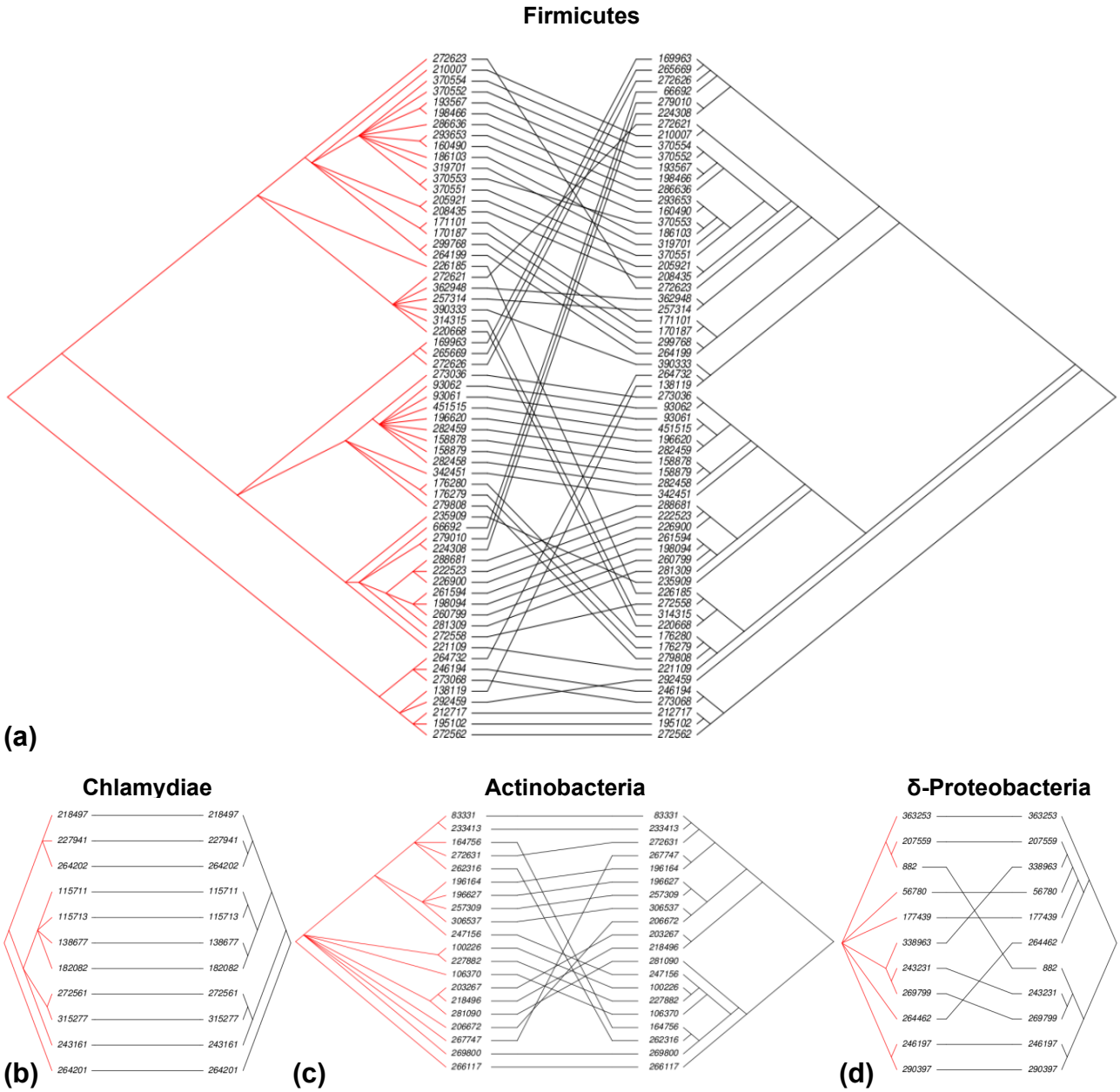
κατανομής των μονονουκλεοτιδικών αποκλίσεων ($V^{MONO}$) και των αποκλίσεων των δινουκλεοτιδίων ($V^{DI}$) κατά μήκος των CDS-συρραφών και προσδιορίσαμε την τοπολογική τους βαθμολογία. Στον Πίνακα 11 παραθέτουμε την τοπολογική βαθμολογία για το σύνολο των σχετικών κλαδογραμμάτων. Για μια λεπτομερή περιγραφή της μεθοδολογίας που ακολουθήσαμε, βλ. ενότητα 2.7.

**ΠΙΝΑΚΑΣ 11. Τοπολογική βαθμολογία κλαδογραμμάτων**

| | **αποκλίσεις μονονουκλεοτιδίων** | **αποκλίσεις δινουκλεοτιδίων** | **αποκλίσεις σταθμισμένων συχνοτήτων** | **γονιδιωματικές υπογραφές** |
|---|---|---|---|---|
| **Actinobacteria** | 59.5 | 72.1 | 64.1 | 67.7 |
| **Bacteroidetes** | 77.8 | 83.3 | 100 | 100 |
| **Chlamydiae** | 90 | 83.3 | 90 | 100 |
| **Cyanobacteria** | 83.5 | 80.8 | 81.8 | 77.6 |
| **Firmicutes** | 68.8 | 77.1 | 80.8 | 81.1 |
| **α-Proteobacteria** | 69.8 | 79.5 | 72.3 | 72.5 |
| **β-Proteobacteria** | 70.6 | 72 | 72.8 | 79.1 |
| **δ-Proteobacteria** | 49.9 | 53.2 | 62.4 | 55.8 |
| **ε-Proteobacteria** | 100 | 93.3 | 93.3 | 93.3 |
| **γ-Proteobacteria** | 67.6 | 75.3 | 74 | 77.9 |
| **Spirochaetes** | 75 | 75 | 100 | 100 |
| **Tenericutes** | 71 | 72.6 | 83.9 | 74.2 |
| **μέσος όρος** | 73.6 | 76.5 | 81.3 | 81.6 |
| **διάμεσος** | 70.8 | 76.2 | 81.3 | 78.5 |

ΣΗΜΕΙΩΣΕΙΣ.- Επί τοις εκατό ποσοστιαία τοπολογική ομοιότητα μεταξύ των κλαδογραμμάτων των αποκλίσεων και των ταξινομικών δέντρων. Οι οργανισμοί που αντιπροσωπεύονται στη συλλογή μας ομαδοποιούνται σύμφωνα με το φύλο ή την κλάση στη οποία ανήκουν. Για κάθε ταξινομική βαθμίδα (φύλο ή κλάση) συγκρίνουμε το κλαδόγραμμα αποκλίσεων με το αντίστοιχο ταξινομικό δέντρο και λαμβάνουμε την τοπολογική βαθμολογία, χρησιμοποιώντας το πρόγραμμα Compare2Trees (Nye et al. 2006). Οι αποκλίσεις ομαδοποιούνται σε τρεις κλάσεις: (α) $V^{MONO}$, αποκλίσεις μονονουκλεοτιδίων, (β) $V^{DI}$, αποκλίσεις δινουκλεοτιδίων, και (γ) $V^{RA}$, αποκλίσεις σταθμισμένων συχνοτήτων. Επιπλέον, κατασκευάζουμε κλαδογράμματα βάσει των δ-αποστάσεων των γονιδιωματικών υπογραφών και χρησιμοποιούμε την τοπολογική τους βαθμολογία ως σημείο αναφοράς. Ο μέσος όρος και η διάμεσος της τοπολογικής βαθμολογίας υπολογίζεται για κάθε κλάση αποκλίσεων και για τις γονιδιωματικές υπογραφές.

### **3.9.2** *Φυλογενετική συσχέτιση των αποκλίσεων των σταθμισμένων δινουκλεοτιδικών συχνοτήτων*




**Εικόνα 10.** Σχεδιαγράμματα σύγκρισης κλαδογραμμάτων. Για κάθε φύλο ή κλάση, οι συγκρίσεις γίνονται μεταξύ των κλαδογραμμάτων που λαμβάνονται από τη βάση NCBI taxonomy (*ταξινομικά δέντρα*· κόκκινο χρώμα) και των κλαδογραμμάτων που κατασκευάσαμε βάσει των $V^{RA}$ (*κλαδογράμματα αποκλίσεων* των σταθμισμένων συχνοτήτων· μαύρο χρώμα). Τα βακτήρια αναφέρονται σύμφωνα με τον ταξινομικό κωδικό του NCBI. Σε κάθε σχεδιάγραμμα, οι άκρες των κλαδογραμμάτων που αντιπροσωπεύουν τον ίδιο οργανισμό ενώνονται ανά ζεύγη με μαύρες γραμμές. Κατά τον τρόπο αυτό, τα σχεδιαγράμματα σύγκρισης παρέχουν μια γενική εικόνα της τοπολογικής ομοιότητας και

των διαφορών δύο κλαδογραμμάτων. Το μήκος των κλαδιών δεν λαμβάνεται υπόψιν. Για τη δημιουργία των σχεδιαγραμμάτων σύγκρισης χρησιμοποιήσαμε την εντολή cophyloplot, που περιλαμβάνεται στο πακέτο ape 3.1-2 (Paradis et al. 2004) της γλώσσας προγραμματισμού R.

Στην Εικόνα 10 παρουσιάζονται διαγράμματα σύγκρισης μεταξύ των κλαδογραμμάτων αποκλίσεων των σταθμισμένων συχνοτήτων και των ταξινομικών δέντρων, για τα Firmicutes, τα Χλαμύδια, τα Ακτινοβακτήρια και τα δ-Πρωτεοβακτήρια που αντιπροσωπεύονται στη συλλογή μας. Η αντίστοιχη τοπολογική βαθμολογία είναι: (α) Firmicutes, 80.8% (β) Χλαμύδια, 90% (γ) Ακτινοβακτήρια, 64.1% και (δ) δ-Πρωτεοβακτήρια, 62.4%. Παρά τις διαφορές στην τοπολογική βαθμολογία, η δομή των διαγραμμάτων σύγκρισης εμφανίζει σημαντικές ομοιότητες στις τέσσερις υπό εξέταση περιπτώσεις, όπως προκύπτει από τις γραμμές που ενώνουν τις άκρες των κλαδογραμμάτων που αντιπροσωπεύουν τον ίδιο οργανισμό. *Στα κλαδογράμματα αποκλίσεων κάθε κόμβος προσδιορίζεται κατά τρόπο ώστε όλα τα κλαδιά που εκκινούν από αυτόν να καταλήγουν σε taxa τα οποία έχουν προφίλ αποκλίσεων σταθμισμένων συχνοτήτων πιο όμοια μεταξύ τους από ότι μεταξύ οποιουδήποτε άλλου taxon που αντιπροσωπεύεται στο κλαδόγραμμα.* Τα ταξινομικά δέντρα προσφέρουν μια απεικόνιση των εξελικτικών σχέσεων των οργανισμών. Από τα σχεδιαγράμματα σύγκρισης προκύπτει ότι οι οργανισμοί που ομαδοποιούνται πάνω από τον ίδιο κόμβο στα κλαδογράμματα αποκλίσεων είναι, εν γένει, εξελικτικά πιο συγγενικοί μεταξύ τους από ότι μεταξύ των υπολοίπων. Συνεπώς, *τα πρότυπα αποκλίσεων σταθμισμένων συχνοτήτων είναι πιο όμοια μεταξύ συγγενικών οργανισμών, ενώ εμφανίζουν περισσότερες διαφορές μεταξύ οργανισμών που έχουν αποκλίνει σε προγενέστερες φάσεις τις εξελικτικής τους πορείας.*

Αντιπαραβάλλοντας την τοπολογική βαθμολογία των κλαδογραμμάτων που κατασκευάσαμε βάσει των αποκλίσεων των σταθμισμένων συχνοτήτων, ($\rho_{XY} - \rho_{Y'X'}$), και βάσει των γονιδιωματικών υπογραφών, ($\rho^*$), (βλ. Πίνακα 11) γίνεται φανερό ότι τόσο οι τιμές των ($\rho_{XY} - \rho_{Y'X'}$) όσο και εκείνες των $\rho^*$ έχουν παρεμφερείς αποδόσεις όταν χρησιμοποιούνται για φυλογενετική ανακατασκευή δέντρων. Επιπλέον, είτε χρησιμοποιούμε τις αποκλίσεις των σταθμισμένων συχνοτήτων είτε τις γονιδιωματικές υπογραφές, η τοπολογία των παραγόμενων κλαδογραμμάτων βρίσκεται εν πολλοίς σε συμφωνία με τις φυλογενετικές σχέσεις των βακτηρίων. Εστιάζοντας στα κλαδογράμματα αποκλίσεων των σταθμισμένων συχνοτήτων, η τοπολογική βαθμολογία είναι μεγαλύτερη από 72%

για όλες τις ταξινομικές βαθμίδες (φύλα ή κλάσεις), εκτός των Ακτινοβακτηρίων και των δ-Πρωτεοβακτηρίων. Ωστόσο, τα σχεδιαγράμματα σύγκρισης (Εικόνα 10) υποδηλώνουν ότι και στις περιπτώσεις των Ακτινοβακτηρίων και των δ-Πρωτεοβακτηρίων, τα κλαδογράμματα αποκλίσεων δεν διαφέρουν ριζικά από τα αντίστοιχα ταξινομικά δέντρα ως προς την τοπολογία τους, παρά την σχετικά χαμηλή βαθμολογία που λαμβάνουν.

Προκειμένου να εξετάσουμε περαιτέρω εάν τα Ακτινοβακτήρια και τα δ-Πρωτεοβακτήρια έχουν ασυνήθιστα πρότυπα αποκλίσεων ($\rho_{XY} - \rho_{Y'X'}$), τα οποία δεν παρακολουθούν την εξελικτική πορεία των αντίστοιχων βακτηρίων, εφαρμόζουμε το τεστ του Grubbs για ακραίες αποκλίνουσες τιμές. Η μηδενική υπόθεση ($H_0$) αυτού του τεστ δηλώνει ότι δεν υπάρχουν αποκλίνουσες τιμές στο υπό εξέταση σύνολο (εν προκειμένω, στο σύνολο της τοπολογικής βαθμολογίας των κλαδογραμμάτων αποκλίσεων σταθμισμένων συχνοτήτων). Στην εκδοχή του τεστ που εφαρμόζουμε, η εναλλακτική υπόθεση ($H_a$) δηλώνει ότι οι δύο μικρότερες τιμές (62.4% και 64.1%, για τα δ-Πρωτεοβακτήρια και τα Ακτινοβακτήρια, αντίστοιχα) είναι συγχρόνως και αποκλίνουσες τιμές. Το τεστ Grubbs δίνει *p*-value ίση με 0.5347, οπότε η εναλλακτική υπόθεση απορρίπτεται ως στατιστικά μη σημαντική. Συνεπώς, τα Ακτινοβακτήρια και τα δ-Πρωτεοβακτήρια δεν αποτελούν εξαιρέσεις από τη γενικότερη εικόνα, σύμφωνα με την οποία οι αποκλίσεις ($\rho_{XY} - \rho_{Y'X'}$) συσχετίζονται με τη φυλογένεση.

Συμπερασματικά, οι ασυμμετρίες των $\rho$ έχουν προφίλ τα οποία είναι ανά είδος καθορισμένα (species-specific). Η μέση τοπολογική βαθμολογία των κλαδογραμμάτων που κατασκευάζονται είτε βάσει των αποκλίσεων των σταθμισμένων συχνοτήτων είτε βάσει των γονιδιωματικών υπογραφών λαμβάνει κατά προσέγγιση ίσες τιμές (81.3% και 81.6%, αντίστοιχα). Μάλιστα, τα κλαδογράμματα αποκλίσεων των σταθμισμένων συχνοτήτων εμφανίζονται να είναι κατά κανόνα πιο ακριβή ως προς την ανασυγκρότηση των εξελικτικών σχέσεων των βακτηρίων, σε σχέση με τα κλαδογράμματα των γονιδιωματικών υπογραφών, όπως προκύπτει από την διάμεσο της τοπολογικής τους βαθμολογίας, που είναι 81.3% και 78.5%, αντίστοιχα. Τα παραπάνω συνηγορούν υπέρ του ισχυρισμού ότι *οι ειδικές ανά κλώνο ασυμμετρίες των συσχετίσεων μεταξύ των 1ης τάξης γειτονικών βάσεων αποτελούν ένα ιδιοσυγκρασιακό* (*idiosyncratic*) *γνώρισμα του γονιδιώματος, βαθιά ριζωμένο στην εξελικτική δυναμική των κλώνων του DNA*.

### **3.9.3** *Φυλογενετική συσχέτιση των μονονουκλεοτιδικών αποκλίσεων*

Ακολούθως, εξετάσαμε εάν οι ασυμμετρίες της μονονουκλεοτιδικής σύστασης των κλώνων του DNA είναι και αυτές ανά είδος καθορισμένες ή, αντίθετα, διαμορφώνονται από εξελικτικά πρότυπα κοινά μεταξύ διαφορετικών οργανισμών. Προγενέστερες μελέτες υποστηρίζουν ότι οι ασυμμετρίες στη σύσταση των κλώνων των βακτηριακών χρωμοσωμάτων προκύπτουν ως το αποτέλεσμα μεταλλακτικών πολώσεων, καθολικά απαντώμενων, οι οποίες συνδέονται με τη γενική αρχιτεκτονική του γονιδιώματος (Rocha et al. 1999).

#### **3.9.3.1** *Καθολικά απαντώμενα πρότυπα ασυμμετριών – οι έως τώρα μελέτες και αντιλήψεις*

Οι Rocha και Danchin (2001) κατέδειξαν ότι η συγκρότηση των ειδικών ανά κλώνο αποκλίσεων παραμένει μια εν εξελίξει διαδικασία, όπως συμβαίνει στην περίπτωση των *Chlamydia trachomatis* και *C.muridarum*. Τα χρωμοσώματα αυτών των βακτηρίων έχουν σύσταση διαφορετική από την αναμενόμενη στην κατάσταση ισορροπίας προς την οποία κινείται η αλληλουχία του DNA βάσει των ρυθμών υποκατάστασης. Ωστόσο, οι ασυμμετρίες των κλώνων του DNA θεωρείται πως είναι της ίδιας φύσης σε όλα τα βακτήρια, είτε η σύστασή τους βρίσκεται σε κατάσταση ισορροπίας είτε όχι. Συγκεκριμένα, αποτελεί κοινή παραδοχή ότι ο οδηγός κλώνος εμφανίζει περίσσεια των κετο- έναντι των αμινο-βάσεων (Lobry 1996, Danchin 2003). Όπως έχουμε ήδη αναφέρει, τα Firmicutes με χαμηλό περιεχόμενο σε GC%, όπως ο *Staphylococcus aureus*, αποτελούν εξαιρέσεις αυτού του γενικού κανόνα, καθώς ο οδηγός κλώνος τους έχει περισσότερα κατάλοιπα A από ότι T. Ωστόσο, σε σχετική μελέτη οι Charneski *et al*. (2011) προσδιόρισαν το προφίλ των μεταλλακτικών ρυθμών στο *S.aureus* και, βασιζόμενοι σε αυτό, συμπέραναν ότι ο οδηγός κλώνος αναμένεται να έχει υψηλότερες συχνότητες T έναντι A στην κατάσταση ισορροπίας, σε αντίθεση με την παρατηρούμενη απόκλιση, όπου [T] < [A]. Έτσι, οι μη-τυπικές A-T αποκλίσεις στον οδηγό κλώνο των Firmicutes αποδόθηκαν σε επιλεκτικές πιέσεις, οι οποίες ασκούνται στη σύσταση των κωδικών περιοχών και στον προσανατολισμό των γονιδίων, τα οποία διατάσσουν τους κωδικούς τους κλώνους κατά προτίμηση στον οδηγό κλώνο του χρωμοσώματος. Τα παραπάνω ευρήματα έρχονται σε συμφωνία και ενισχύουν την εκτίμηση ότι οι ειδικές ανά κλώνο ασυμμετρίες των μεταλλάξεων είναι πολωμένες προς την ίδια κατεύθυνση, στο

σύνολο σχεδόν των βακτηριακών γονιδιωμάτων. Κατά συνέπεια, σύμφωνα με αυτή τη θεώρηση, οι ασυμμετρίες της μονονουκλεοτιδικής σύστασης των βακτηριακών χρωμοσωμάτων παράγονται από εξελικτικές τάσεις που είναι κοινές σε αποκλίνοντα είδη (divergent species) και ως εκ τούτου δεν συνδέονται με την εξελικτική τους πορεία.

Ωστόσο, μελέτες που αφορούν το DNA των ευκαρυωτικών οργανισμών καταδεικνύουν ότι οι ειδικές ανά κλώνο αποκλίσεις μπορεί να συσχετίζονται με την φυλογένεση. Συγκεκριμένα, σύμφωνα με τις σχετικές αναφορές, η ασύμμετρη κατανομή των πουρινών μεταξύ των κλώνων του μιτοχονδριακού DNA συγκροτεί προφίλ τα οποία παρακολουθούν τις εξελικτικές σχέσεις των ευκαρυωτικών οργανισμών (Mohr et al. 1999, Barral P et al. 2005). Οι παρατηρήσεις αυτές μας παρακίνησαν να αναζητήσουμε πιθανές συσχετίσεις των ασυμμετριών, σε όρους μονονουκλεοτιδικών αποκλίσεων, με τη φυλογένεση εν προκειμένω των βακτηρίων, επανεκτιμώντας έτσι την καθιερωμένη αντίληψη που αμφισβητεί την ύπαρξη τέτοιων συσχετίσεων.

### **3.9.3.2** *Εξελικτικά πρότυπα ασυμμετριών – Ανάλυση κλαδογραμμάτων*

Στον Πίνακα 11 παρουσιάζεται η τοπολογική βαθμολογία των κλαδογραμμάτων που κατασκευάσαμε βάσει των KL-αποκλίσεων μεταξύ των $V^{MONO}$ κατά μήκος των CDS-συρραφών. *Τα αποτελέσματά μας αποκαλύπτουν μια ποικιλία εξελικτικών προτύπων, στο επίπεδο της εμφάνισης των μονονουκλεοτιδικών αποκλίσεων, ανάλογα με το φύλο ή την κλάση που εξετάζουμε*. Στα Χλαμύδια, στα Κυανοβακτήρια και στα ε-Πρωτεοβακτήρια η εμφάνιση διακριτών μεταξύ τους μονονουκλεοτιδικών αποκλίσεων συμβαδίζει με την διαφοροποίηση των ειδών. Σε αυτές τις ταξινομικές ομάδες, η τοπολογική ομοιότητα μεταξύ των κλαδογραμμάτων μονονουκλεοτιδικών αποκλίσεων και των ταξινομικών δέντρων κυμαίνεται από 83.5% έως 100%. Αντιθέτως, τα κλαδογράμματα μονονουκλεοτιδικών αποκλίσεων για τα δ-Πρωτεοβακτήρια και τα Ακτινοβακτήρια διαφέρουν έντονα από τα αντίστοιχα ταξινομικά δέντρα, με την τοπολογική βαθμολογία τους να ισούται με 49.9% και 59.5%, αντίστοιχα. Τα προφίλ των μονονουκλεοτιδικών αποκλίσεων στα Βακτηριοειδή, στα β-Πρωτεοβακτήρια, στις Σπειροχαίτες και στα Tenericutes παρακολουθούν, εν γένει, τις εξελικτικές σχέσεις των βακτηρίων, με τα αντίστοιχα κλαδογράμματα να λαμβάνουν τοπολογική βαθμολογία από 71% έως 77.8%. Σε ό,τι αφορά τα Firmicutes, των οποίων οι μονονουκλεοτιδικές αποκλίσεις θεωρούνται ως μη-τυπικές και ως το

αποτέλεσμα πρωτίστως επιλεκτικών παρά μεταλλακτικών πιέσεων, η αντίστοιχη τοπολογική βαθμολογία είναι 68.8%, κατά τι μικρότερη από τη διάμεσο (70.8%).

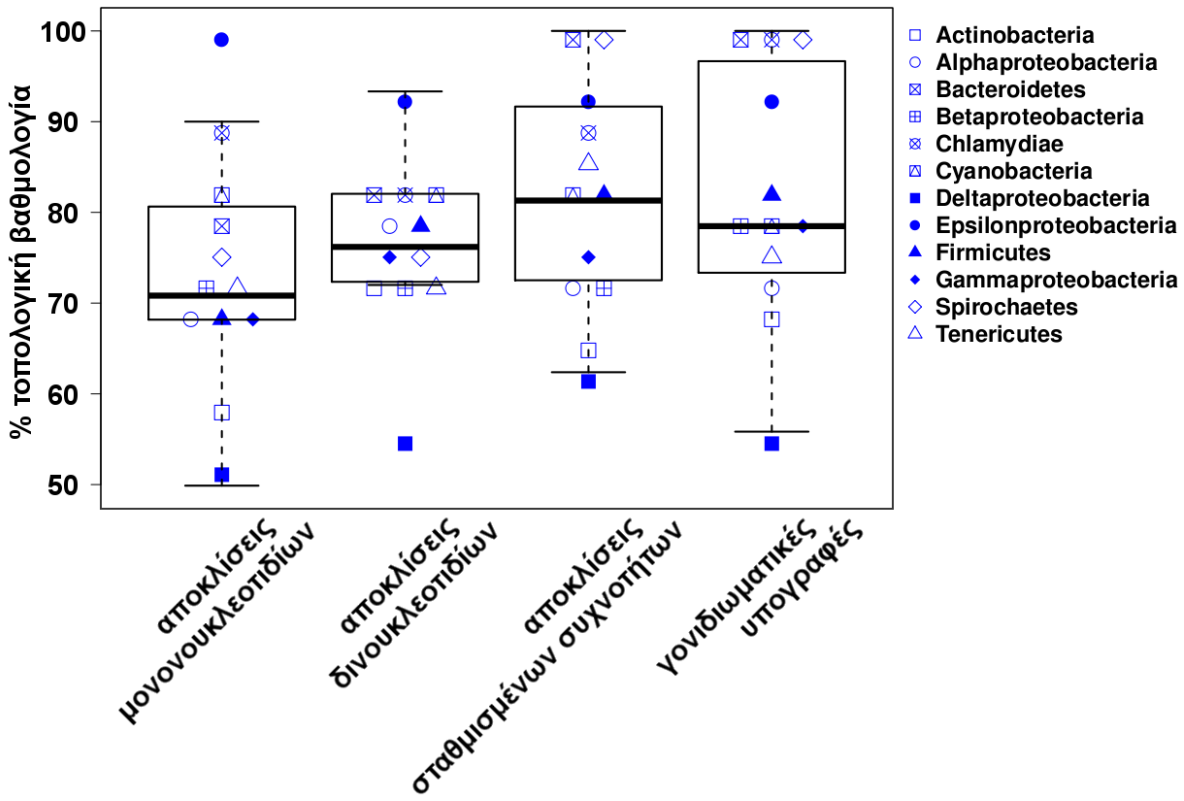
Τα παραπάνω αποτελέσματα υποδεικνύουν ότι τα πρότυπα των μονονουκλεοτιδικών αποκλίσεων απέχουν από το να είναι η συνισταμένη κοινών μεταξύ των βακτηρίων πολώσεων στους ρυθμούς υποκατάστασης. Αντί να προκύπτουν από ασυμμετρίες με καθολική ισχύ, *οι μονονουκλεοτιδικές αποκλίσεις φαίνεται ότι αντανακλούν επί μέρους πτυχές της αντιγραφής και επιδιόρθωσης του DNA, καθώς επίσης και των επιλεκτικών περιορισμών που ασκούνται στο γονιδίωμα*. Η ανάλυσή μας δείχνει ότι οι αποκλίσεις αυτές περιέχουν φυλογενετική πληροφορία που επιτρέπει την ανασυγκρότηση των εξελικτικών σχέσεων των βακτηρίων, άλλοτε με μικρότερη και άλλοτε με μεγαλύτερη πιστότητα, ανάλογα με το φύλο ή την κλάση που μελετάμε.

### **3.9.3** *Φυλογενετική συσχέτιση των δινουκλεοτιδικών αποκλίσεων*

Για λόγους πληρότητας, στον Πίνακα 11 παρουσιάζουμε την τοπολογική βαθμολογία που αντιστοιχεί στα κλαδογράμματα αποκλίσεων των δινουκλεοτιδίων. Εκτός των δ-Πρωτεοβακτηρίων, το σύνολο των υπλοίπων φύλων ή κλάσεων δίνουν κλαδογράμματα με τοπολογική βαθμολογία από 72% έως 93.3%. Συνεπώς, οι ασυμμετρίες των δινουκλεοτιδίων συσχετίζονται με τη φυλογένεση των βακτηρίων. Οι παρατηρούμενες συχνότητες των δινουκλεοτιδίων είναι το συνδυαστικό αποτέλεσμα των μονονουκλεοτιδικών συχνοτήτων και των συσχετίσεων μεταξύ των 1$^{ης}$ τάξης γειτονικών βάσεων. *Οι μονονουκλεοτιδικές αποκλίσεις και οι αποκλίσεις των σταθμισμένων συχνοτήτων συγχωνεύονται στο επίπεδο των αποκλίσεων των δινουκλεοτιδίων*. Έτσι, η ασύμμετρη κατανομή των δινουκλεοτιδίων στους κλώνους του DNA μπορεί να ιδωθεί ως το δευτερογενές αποτέλεσμα επί μέρους ασυμμετριών, δίχως ωστόσο να επιτρέπει να εστιάσουμε σε κάθε μία από αυτές ξεχωριστά.

### **3.9.4** *Η ανά είδος καθορισμένη φύση των ειδικών ανά κλώνο αποκλίσεων*

Απεικονίσαμε με θηκογράμματα (boxplots) την κατανομή της τοπολογικής βαθμολογίας για κάθε μία από τις τρεις κλάσεις αποκλίσεων ($V^{MONO}$, $V^{DI}$, $V^{RA}$) καθώς επίσης και τις γονιδιωματικές υπογραφές (Εικόνα 11). Παρατηρούμε ότι *οι ασυμμετρίες στην εξέλιξη των κλώνων φέρουν πληροφορία που επιτρέπει την ανασυγκρότηση των φυλογενετικών σχέσεων με ολοένα μεγαλύτερη ακρίβεια, καθώς μετατοπιζόμαστε από το επίπεδο των μονονουκλεοτιδικών αποκλίσεων στις αποκλίσεις των δινουκλεοτιδίων και, τελικώς, των σταθμισμένων τους συχνοτήτων*. Η διάμεσος της τοπολογικής βαθμολογίας των κλαδογραμμάτων αποκλίσεων σταθμισμένων συχνοτήτων είναι μετατοπισμένη προς μεγαλύτερες τιμές σε σχέση με τις κατανομές που αντιστοιχούν στα κλαδογράμματα που έχουν κατασκευαστεί όχι μόνο βάσει των αποκλίσεων μονο- ή δι-νουκλεοτιδίων, αλλά και βάσει των γονιδιωματικών υπογραφών (Εικόνα 11). Η παρατήρηση αυτή καταδεικνύει την *ανά είδος καθορισμένη* (*species specific*) *φύση των αποκλίσεων των σταθμισμένων συχνοτήτων*, δεδομένου ότι οι γονιδιωματικές υπογραφές έχουν επιτυχώς χρησιμοποιηθεί ώστε να διακρίνουν μεταξύ χρωμοσωμάτων που ανήκουν σε αποκλίνοντα είδη (divergent species) (van Passel et al. 2006, Phillippy et al. 2007, Bohlin et al. 2009, Bohlin & Skjerve 2009). Συνεπώς, *τόσο οι συσχετίσεις των 1$^{ης}$ τάξης γειτονικών βάσεων, καθαυτές, όσο και οι ειδικές ανά κλώνο ασυμμετρίες τους, αποτελούν χαρακτηριστικά του γονιδιώματος συνυφασμένα με την εξελικτική πορεία των οργανισμών*.




**Εικόνα 11.** Θηκογράμματα που απεικονίζουν την κατανομή της % τοπολογικής βαθμολογίας (topological scores) των κλαδογραμμάτων. Η βαθμολογία κάθε κλαδογράμματος δηλώνει την τοπολογική ομοιότητά του με το αντίστοιχο ταξινομικό δέντρο. Οι συγκρίσεις αφορούν κλαδογράμματα που κατασκευάστηκαν βάσει της κατανομής των αποκλίσεων (*α*) των μονονουκλεοτιδίων, (*β*) των δινουκλεοτιδίων και (*γ*) των σταθμισμένων συχνοτήτων των δινουκλεοτιδίων. Επίσης, παρουσιάζουμε την % τοπολογική βαθμολογία των κλαδογραμμάτων που κατασκευάστηκαν βάσει των δ-αποστάσεων των γονιδιωματικών υπογραφών (genomic signatures). Κάθε κλαδόγραμμα περιλαμβάνει βακτηριακά είδη που ανήκουν στο ίδιο φύλο ή την ίδια κλάση.

Τα αποτελέσματα της παρούσας ενότητας υποδηλώνουν ότι οι αποκλίσεις που μελετάμε θα μπορούσαν να χρησιμοποιηθούν σε φυλογενετικές αναλύσεις, βελτιστοποιώντας την απόδοσή τους. Ωστόσο, σκοπός της παρούσας μελέτης δεν είναι να παρουσιάσουμε το πρόπλασμα μιας νέας μεθόδου για την κατασκευή φυλογενετικών δέντρων. Άλλωστε, όποιο κριτήριο και αν εφαρμόσουμε, η κατασκευή κλαδογραμμάτων δεν μας επιτρέπει να συλλάβουμε πλήρως τις εξελικτικές διαδικασίες που οδήγησαν στην δημιουργία των παρατηρούμενων χαρακτηριστικών του γονιδιώματος. Σε κάθε κλαδόγραμμα αντιστοιχεί ένας απροσδιόριστος αριθμός πιθανών εξελικτικών σεναρίων σχετικά με την διαμόρφωση των χαρακτηριστικών που μελετάμε (Podani 2013). Στόχος μας ήταν

να εντοπίσουμε πιθανές συσχετίσεις της ασύμμετρης εξέλιξης των κλώνων του DNA με τη φυλογένεση των βακτηρίων, συσχέτιση που μέχρι πρότινος αμφισβητείτο. Όπως δείξαμε, η συσχέτιση αυτή υπάρχει και, ειδικά για την περίπτωση των αποκλίσεων των σταθμισμένων συχνοτήτων, είναι ιδιαίτερα έντονη. Στις ενότητες 3.11 και 3.12, θα επιχειρήσουμε να προσδιορίσουμε συγκεκριμένους μηχανισμούς και διαδικασίες που εμπλέκονται στην εμφάνιση των ειδικών ανά κλώνο αποκλίσεων.

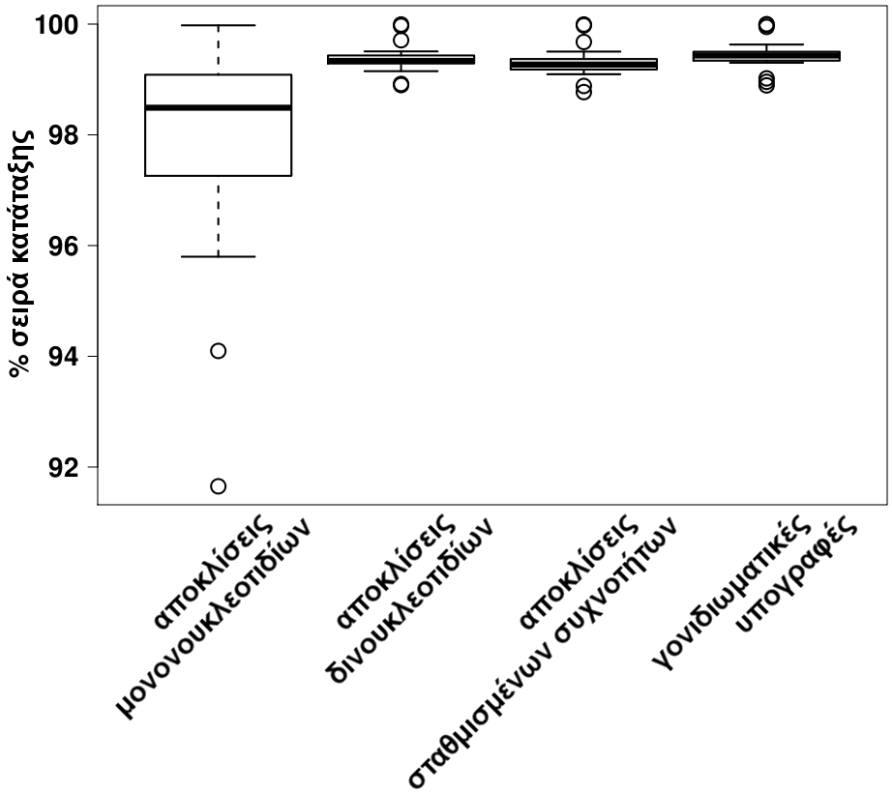
## **3.10** *Ασυμμετρίες της σύστασης και εξελικτικές σχέσεις των κωδικών περιοχών*

Για την κατασκευή των κλαδογραμμάτων χρησιμοποιήσαμε ως χαρακτηριστικά των βακτηριακών ειδών τις κατανομές των αποκλίσεων τους κατά μήκος των CDS-συρραφών. Σύμφωνα με τα αποτελέσματά μας, οι ασυμμετρίες της σύστασης των κωδικών περιοχών, ιδιαίτερα σε όρους αποκλίσεων σταθμισμένων συχνοτήτων, παρέχουν σημαντική πληροφορία για την ανασυγκρότηση των εξελικτικών σχέσεων των βακτηρίων. Ωστόσο, οι CDS-συρραφές αποτελούνται από κωδικές περιοχές, μέρος των οποίων μοιράζεται εκτεταμένες ομολογίες μεταξύ διαφορετικών βακτηρίων. Ως εκ τούτου, οι ομοιότητες και οι διαφορές μεταξύ των προφίλ των αποκλίσεων, βάσει των οποίων κατασκευάστηκαν τα κλαδογράμματα, ενδέχεται να αντανακλούν απλώς την μεγαλύτερη ή μικρότερη ομοιότητα της αλληλουχίας (sequence similarity) μεταξύ περιοχών DNA που κατάγονται από έναν πιο πρόσφατο ή πιο μακρινό κοινό πρόγονο, αντίστοιχα. Προκειμένου να αντιμετωπίσουμε αυτό το ενδεχόμενο εφαρμόσαμε δύο διαφορετικές μεθοδολογίες, οι οποίες βασίζονται στη σύγκριση των αποκλίσεων μη ομόλογων περιοχών DNA που ανήκουν είτε σε διαφορετικά χρωμοσώματα του ίδιου γονιδιώματος είτε στο ίδιο χρωμόσωμα.

### **3.10.1** *Ασυμμετρίες των κωδικών περιοχών στα χρωμοσώματα που ανήκουν στο ίδιο γονιδίωμα*

Αρχικά, ξεχωρίσαμε εκείνα τα βακτήρια των οποίων το γονιδίωμα αποτελείται από περισσότερα του ενός χρωμοσώματα. Στη συλλογή μας αντιπροσωπεύονται 26 τέτοια βακτηριακά είδη. Τα χρωμοσώματα που ανήκουν στο ίδιο γονιδίωμα δεν είναι μεταξύ τους ομόλογα, τουλάχιστον κατά το μεγαλύτερο μέρος τους. Συνεπώς, ο βαθμός ομοιότητας μεταξύ των ειδικών ανά κλώνο ασυμμετριών τους δεν μπορεί να αποδοθεί στον βαθμό ομοιότητας των αλληλουχιών τους λόγω ομολογίας. Το γεγονός αυτό αποτελεί τη βάση για τον πρώτο έλεγχο που εφαρμόσαμε.

Συγκεκριμένα, για κάθε τύπο αποκλίσεων πραγματοποιούμε ανά ζεύγη συγκρίσεις των αντίστοιχων κατανομών ($V^{MONO}$, $V^{DI}$, $V^{RA}$) κατά μήκος των CDS-συρραφών, όπως ακριβώς κάναμε και για την κατασκευή των κλαδογραμμάτων. Όσο μικρότερη είναι η KL-απόκλιση μεταξύ δύο κατανομών αποκλίσεων, τόσο μεγαλύτερη είναι η ομοιότητά τους. Διατάσσουμε όλα τα ζεύγη χρωμοσωμάτων κατά αύξουσα σειρά σύμφωνα με την ομοιότητα των $V^{MONO}$, των $V^{DI}$ και των $V^{RA}$. Έτσι, προκύπτουν τρεις διατεταγμένες σειρές από ζεύγη χρωμοσωμάτων. Σε κάθε μία από αυτές εντοπίζουμε την σειρά κατάταξης των ζευγαριών που αντιστοιχούν σε συγκρίσεις μεταξύ χρωμοσωμάτων του ίδιου γονιδιώματος. Στη συλλογή μας υπάρχουν 55 τέτοια ζεύγη. Όσο μεγαλύτερη είναι η σειρά κατάταξής τους, τόσο πιο όμοια είναι τα προφίλ των αποκλίσεων μεταξύ των χρωμοσωμάτων του ίδιου γονιδιώματος από ότι μεταξύ των χρωμοσωμάτων που ανήκουν σε διαφορετικούς οργανισμούς. Επίσης, κατ' αναλογία με τα παραπάνω, δημιουργήσαμε μία διατεταγμένη σειρά από όλα τα δυνατά ζεύγη χρωμοσωμάτων, χρησιμοποιώντας τις $\delta$-αποστάσεις των γονιδιωματικών τους υπογραφών. Έτσι, όσο μικρότερη είναι η $\delta$-απόσταση δύο χρωμοσώματα, τόσο μεγαλύτερη είναι η σειρά κατάταξης του αντίστοιχου ζεύγους.




**Εικόνα 12.** Θηκογράμματα που απεικονίζουν την κατανομή της % σειράς κατάταξης για κάθε ζεύγος χρωμοσωμάτων που ανήκει στο ίδιο γονιδίωμα. Όσο μεγαλύτερη είναι η σειρά κατάταξης, τόσο μεγαλύτερη είναι η ομοιότητα μεταξύ των χρωμοσωμάτων του ίδιου γονιδιώματος από ότι μεταξύ χρωμοσωμάτων που ανήκουν σε διαφορετικούς οργανισμούς. Η ομοιότητα αφορά τα προφίλ των αποκλίσεων (*α*) των μονονουκλεοτιδίων, (*β*) των δινουκλεοτιδίων και (*γ*) των σταθμισμένων συχνοτήτων των δινουκλεοτιδίων, καθώς επίσης και (*δ*) τις γονιδιωματικές υπογραφές των χρωμοσωμάτων (genomic signatures).

**ΠΙΝΑΚΑΣ 12. Σειρά κατάταξης βάσει ομοιότητας**

| | **αποκλίσεις μονονουκλεοτιδίων** | **αποκλίσεις δινουκλεοτιδίων** | **αποκλίσεις σταθμισμένων συχνοτήτων** | **γονιδιωματικές υπογραφές** |
|---|---|---|---|---|
| ελάχιστη τιμή | 91.65 | 98.9 | 98.77 | 98.89 |
| μέγιστη τιμή | 99.98 | 99.99 | 99.99 | 100 |
| μέση τιμή | 98.04 | 99.37 | 99.3 | 99.44 |
| διάμεσος | 98.49 | 99.34 | 99.27 | 99.43 |
| τυπική απόκλιση | 1.75 | 0.22 | 0.25 | 0.24 |

ΣΗΜΕΙΩΣΕΙΣ.- Περιγραφικά στατιστικά στοιχεία (descriptive statistics) των κατανομών της επί τοις εκατό ποσοστιαίας σειράς κατάταξης για όλα τα ζεύγη χρωμοσωμάτων που ανήκουν στο ίδιο γονιδίωμα. Η κατάταξη γίνεται κατά αύξουσα σειρά ομοιότητας, βάσει των προφίλ των αποκλίσεων (*α*) των μονονουκλεοτιδίων, (*β*) των δινουκλεοτιδίων και (*γ*) των σταθμισμένων συχνοτήτων των δινουκλεοτιδίων, καθώς επίσης και βάσει (*δ*) των γονιδιωματικών υπογραφών των χρωμοσωμάτων.

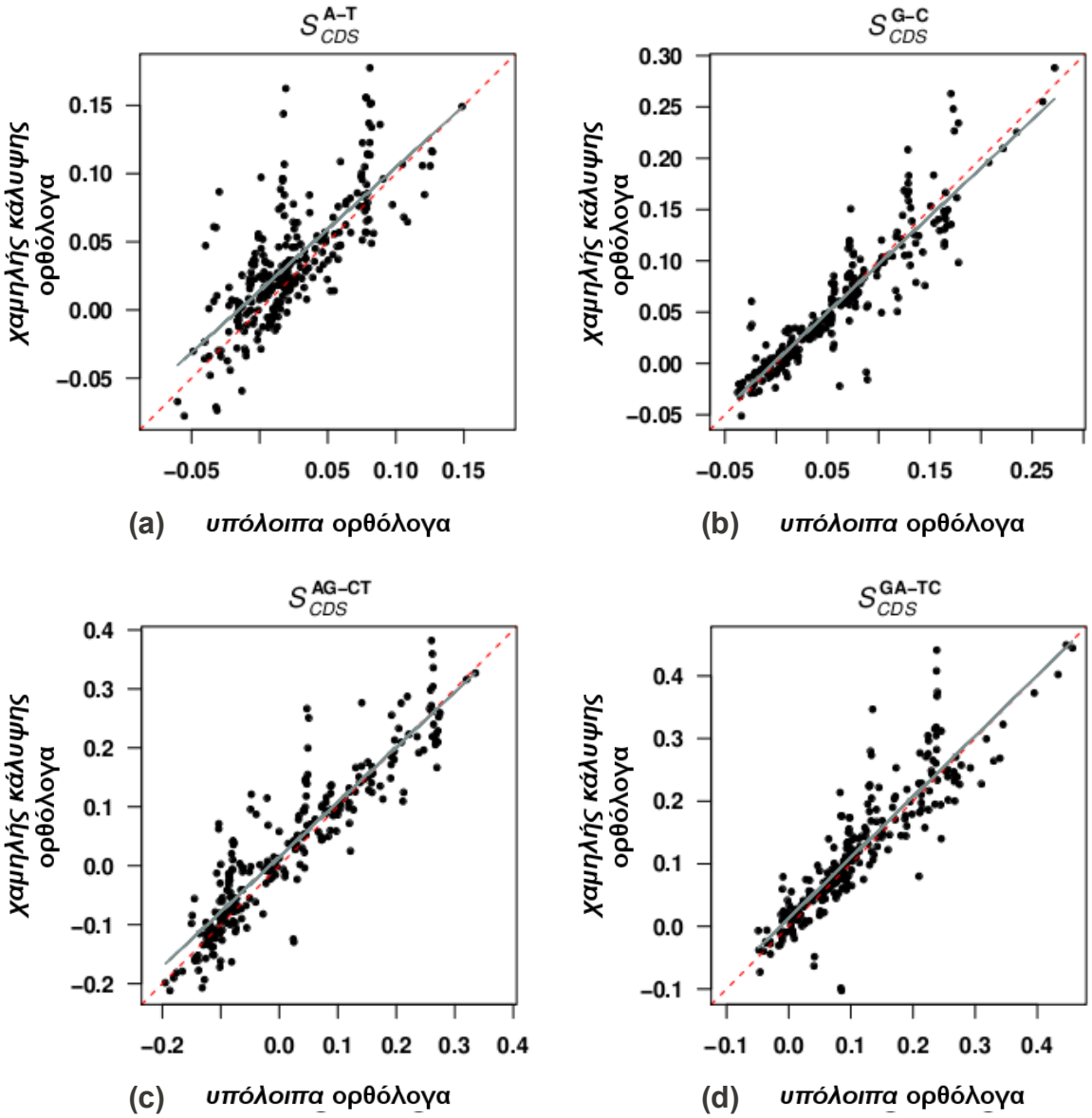
Από την Εικόνα 12 προκύπτει ότι τα ζεύγη των χρωμοσωμάτων που ανήκουν στον ίδιο οργανισμό καταλαμβάνουν % σειρά κατάταξης με μικρό εύρος κατανομής, όταν συγκρίνουμε τις αποκλίσεις των δινουκλεοτιδίων και των σταθμισμένων τους συχνοτήτων, καθώς επίσης και τις γονιδιωματικές τους υπογραφές. Η % σειρά κατάταξης βάσει των προφίλ των μονονουκλεοτιδικών αποκλίσεων ακολουθεί μία κατανομή με μεγαλύτερη διασπορά τιμών. Για κάθε μία από αυτές τις κατανομές, ωστόσο, η διάμεσος λαμβάνει τιμή μεγαλύτερη του 98% (Πίνακας 12). Τα ζεύγη χρωμοσωμάτων του ίδιου γονιδιώματος κατατάσσονται στο άνω άκρο των διατεταγμένων σειρών που εξετάζουμε. Μάλιστα, στην περίπτωση των αποκλίσεων των δινουκλεοτιδίων και σταθμισμένων τους συχνοτήτων η ελάχιστη τιμή της σειράς κατάταξης ισούται με 98.90% και 98.77%, αντίστοιχα. Συνεπώς, *τα χρωμοσώματα έχουν προφίλ αποκλίσεων τα οποία είναι χαρακτηριστικά του είδους στο οποίο ανήκουν, γεγονός που, σύμφωνα με τα παραπάνω, δεν μπορεί να αναχθεί σε εκτεταμένη μεταξύ τους ομοιότητα αλληλουχιών λόγω ομολογίας*.

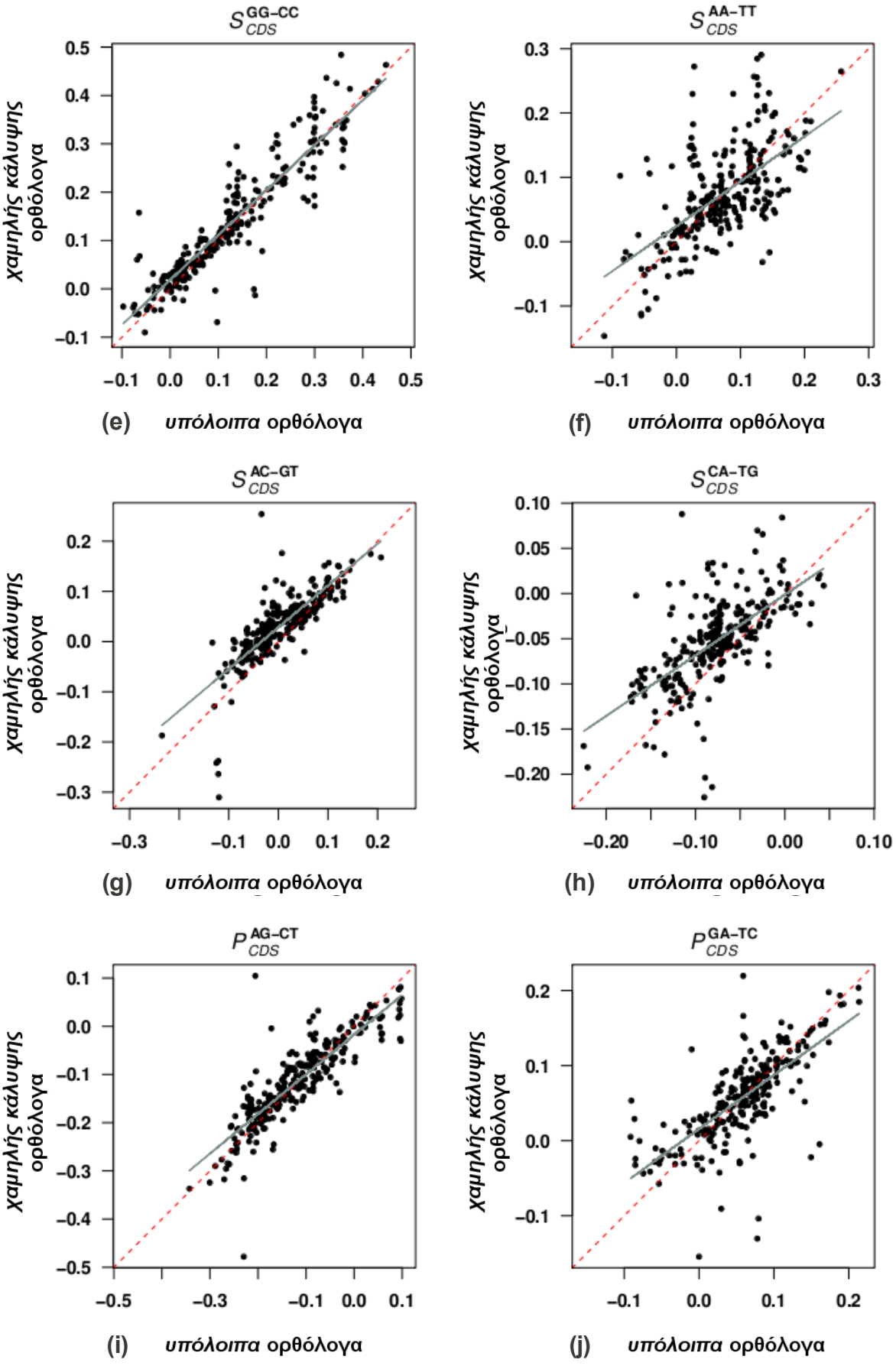
### **3.10.2** *Ασυμμετρίες γονιδίων με διαφορετική εξελικτική προέλευση*

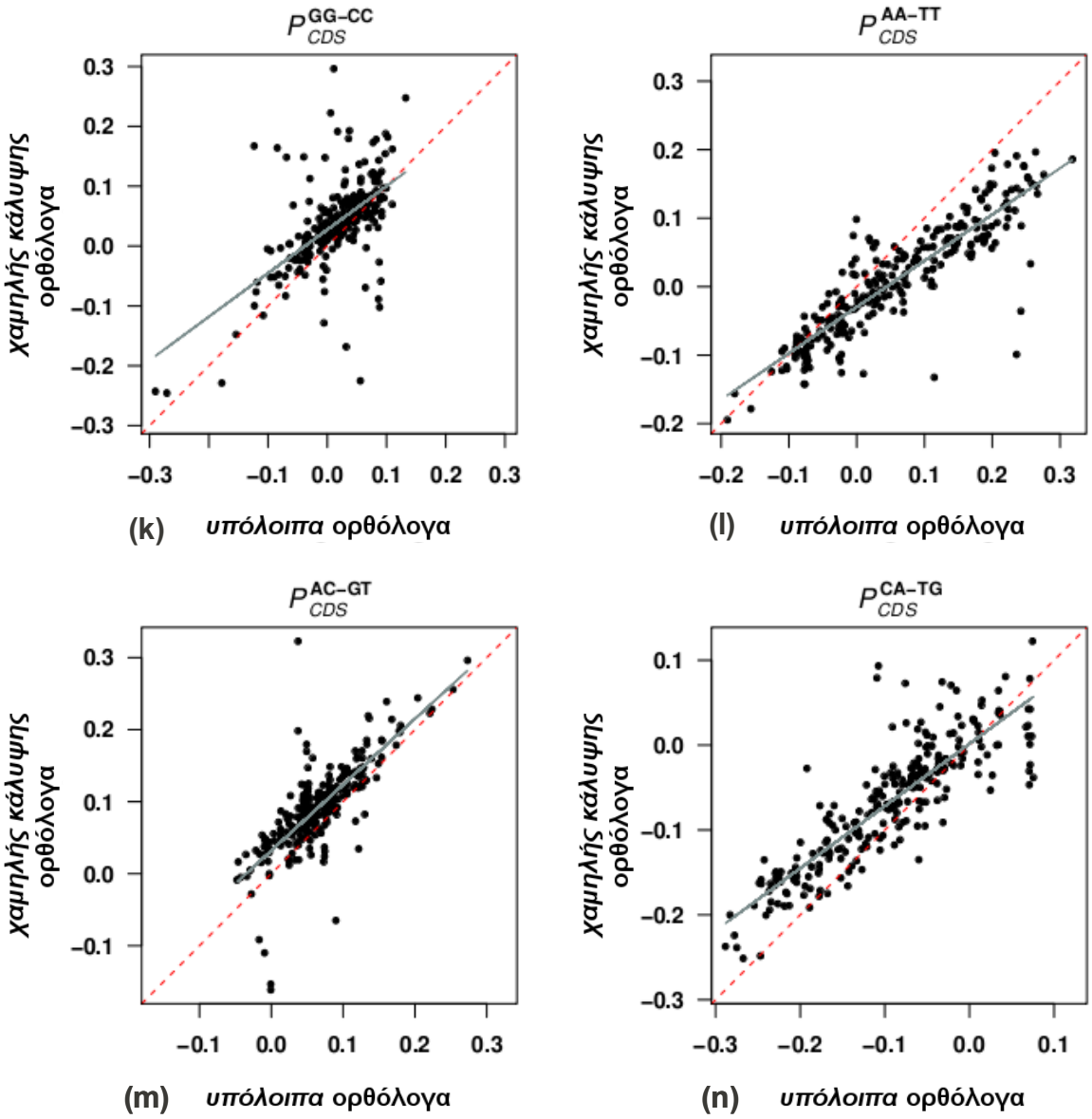
Για να ελέγξουμε περαιτέρω το επιχείρημά μας, συγκρίνουμε τις αποκλίσεις γονιδίων που έχουν διαφορετική εξελικτική προέλευση. Λαμβάνουμε τις ορθόλογες ομάδες (orthologous groups) των γονιδίων (Powell et al. 2014), όπως αυτές προσδιορίζονται στο επίπεδο κάθε φύλου ή κλάσης που αναλύσαμε με τα κλαδογράμματα αποκλίσεων. Για κάθε γονίδιο υπολογίζουμε την συχνότητα εμφάνισης των ορθολόγων του στο σύνολο των χρωμοσωμάτων κάθε φύλου ή κλάσης. Από κάθε χρωμόσωμα ξεχωρίζουμε τα γονίδια των οποίων τα ορθόλογα έχουν χαμηλή συχνότητα εμφάνισης. Τα γονίδια αυτά συγκροτούν το *χαμηλής κάλυψης* υποσύνολο ορθολόγων. Ακολούθως, για κάθε χρωμόσωμα της συλλογής

μας, αναπαριστούμε τις αποκλίσεις του *χαμηλής κάλυψης* υποσυνόλου ορθολόγων συναρτήσει των αποκλίσεων στο υποσύνολο των *υπλοίπων* ορθολόγων που εντοπίζονται στο χρωμόσωμα αυτό. Οι αποκλίσεις υπολογίζονται κατά μήκος των κωδικών κλώνων των γονιδίων. Για μία αναλυτικότερη παρουσίαση της κατάταξης των γονιδίων στο *χαμηλής κάλυψης* υποσύνολο ορθολόγων και στο υποσύνολο των *υπλοίπων* ορθολόγων, βλ. ενότητα 2.8.

Εκτιμάμε την ευθεία παλινδρόμησης των αποκλίσεων στο *χαμηλής κάλυψης* υποσυνόλου ορθολόγων πάνω στις αντίστοιχες αποκλίσεις του συνόλου των *υπλοίπων* ορθολόγων. Στην Εικόνα 13 παρουσιάζουμε τα σχετικά γραφήματα για κάθε τύπο απόκλισης, μαζί με τα αντίστοιχα μοντέλα γραμμικής παλινδρόμησης. Οι συντελεστές αυτών των μοντέλων παρατίθενται στον Πίνακα 13.




$S_{CDS}^{GG-CC}$
χαμηλής κάλυψης ορθόλογα
υπόλοιπα ορθόλογα
(e)
$S_{CDS}^{AA-TT}$
χαμηλής κάλυψης ορθόλογα
υπόλοιπα ορθόλογα
(f)
$S_{CDS}^{AC-GT}$
χαμηλής κάλυψης ορθόλογα
υπόλοιπα ορθόλογα
(g)
$S_{CDS}^{CA-TG}$
χαμηλής κάλυψης ορθόλογα
υπόλοιπα ορθόλογα
(h)
$P_{CDS}^{AG-CT}$
χαμηλής κάλυψης ορθόλογα
υπόλοιπα ορθόλογα
(i)
$P_{CDS}^{GA-TC}$
χαμηλής κάλυψης ορθόλογα
υπόλοιπα ορθόλογα
(j)




**Εικόνα 13.** Διαγράμματα διασποράς που απεικονίζουν τις αποκλίσεις των *χαμηλής κάλυψης* ορθολόγων συναρτήσει των αποκλίσεων των *υπολοίπων* ορθολόγων. Κάθε σημείο των γραφημάτων αντιπροσωπεύει και ένα χρωμόσωμα της συλλογής μας. Για κάθε τύπο απόκλισης δίδεται και το αντίστοιχο μοντέλο γραμμικής παλινδρόμησης (συνεχής γκρι γραμμή). Σε κάθε γράφημα σχεδιάζεται η διαγώνιος (κλίση=1, σημείο τομής με τον y-άξονα=0· διακεκομμένη κόκκινη γραμμή). Τα σημεία που κείνται πάνω στη διαγώνιο δηλώνουν ότι οι αποκλίσεις στα δύο υποσύνολα ορθολόγων είναι ταυτόσημες. Για κάθε τύπο απόκλισης, εάν η γραμμή παλινδρόμησης έχει κλήση κοντά στη μονάδα και σημείο τομής με τον y-άξονα κοντά στο μηδέν, δηλαδή τείνει να συμπέσει με τη διαγώνιο, οι αντίστοιχες ασυμμετρίες ακολουθούν πολύ όμοια πρότυπα στα δύο υποσύνολα ορθολόγων, για τα περισσότερα από τα χρωμοσώματα της συλλογής μας.

**ΠΙΝΑΚΑΣ 13. Γραμμική παλινδρόμηση των αποκλίσεων των *χαμηλής κάλυψης* ορθόλογων πάνω στις αποκλίσεις των *υπολοίπων* ορθόλογων**

| | | κλίση | συντελεστής διεύθυνσης | $r^2$ | *p*-value | Pearson's *r* |
|---|---|---|---|---|---|---|
| αποκλίσεις μόνο- και δι-νουκλεοτιδίων | $S^{A-T}$ | 0.905 | 0.014 | 0.565 | 1.21e-49 | 0.752 |
| | $S^{G-C}$ | 0.939 | 0.00275 | 0.84 | 4.47e-107 | 0.917 |
| | $S^{AG-CT}$ | 0.931 | 0.0149 | 0.834 | 8.39e-105 | 0.913 |
| | $S^{GA-TC}$ | 0.97 | 0.0128 | 0.798 | 1.09e-93 | 0.893 |
| | $S^{GG-CC}$ | 0.929 | 0.0185 | 0.851 | 5.13e-111 | 0.922 |
| | $S^{AA-TT}$ | 0.698 | 0.024 | 0.375 | 9.99e-29 | 0.612 |
| | $S^{AC-GT}$ | 0.832 | 0.0284 | 0.639 | 2.77e-60 | 0.799 |
| | $S^{CA-TG}$ | 0.671 | -0.00135 | 0.424 | 1.7e-33 | 0.651 |
| αποκλίσεις σταθμισμένων συχνοτήτων | $P^{AG-CT}$ | 0.823 | -0.0174 | 0.747 | 7.52e-81 | 0.865 |
| | $P^{GA-TC}$ | 0.719 | 0.0156 | 0.503 | 6.46e-42 | 0.709 |
| | $P^{GG-CC}$ | 0.727 | 0.0274 | 0.358 | 3.07e-27 | 0.599 |
| | $P^{AA-TT}$ | 0.676 | -0.0298 | 0.79 | 2.52e-91 | 0.889 |
| | $P^{AC-GT}$ | 0.915 | 0.0321 | 0.585 | 2.55e-52 | 0.765 |
| | $P^{CA-TG}$ | 0.733 | 0.00153 | 0.747 | 9.38e-81 | 0.864 |

ΣΗΜΕΙΩΣΕΙΣ.- Για κάθε τύπο απόκλισης δίδονται η κλίση, ο συντελεστής διεύθυνσης, ο συντελεστής προσδιορισμού ($r^2$) και η τιμή *p* του αντίστοιχου γραμμικού μοντέλου. Επίσης, υπολογίζεται ο δειγματικός συντελεστής γραμμικής συσχέτισης του *Pearson* (*r*) μεταξύ των αποκλίσεων στα *χαμηλής κάλυψης* ορθόλογα και στα *υπόλοιπα* ορθόλογα. Όταν η κλίση είναι κοντά στην μονάδα και ο συντελεστής διεύθυνσης κοντά στο μηδέν, τα δύο υποσύνολα ορθολόγων έχουν πολύ όμοια πρότυπα αποκλίσεων.

Οι συγκρίσεις των *χαμηλής κάλυψης* ορθολόγων με το υποσύνολο των *υπολοίπων* ορθολόγων δίνουν τη δυνατότητα να εκτιμήσουμε εάν η ανά είδος καθορισμένη φύση των αποκλίσεων συναρτάται από την εξελικτική προέλευση των αλληλουχιών του DNA. Συγκεκριμένα, η εξελικτική ιστορία των *χαμηλής κάλυψης*

ορθολόγων δεν αντανακλά τις φυλογενετικές σχέσεις των βακτηρίων στο επίπεδο του φύλου ή της κλάσης, καθώς οι κωδικές περιοχές που συγκροτούν αυτό το υποσύνολο δεν έχουν ορθόλογα παρά μονάχα σε ένα μικρό αριθμό μελών της κάθε ταξινομικής ομάδας του εξετάζουμε. Συνεπώς, εάν αυτές οι κωδικές αλληλουχίες εμφανίζουν πρότυπα αποκλίσεων όμοια με εκείνα των *υπολοίπων* ορθολόγων του χρωμοσώματος στο οποίο ανήκουν, τότε η συσχέτιση των ειδικών ανά κλώνο ασυμμετριών με την εξελικτική πορεία των βακτηρίων δεν μπορεί να αποδοθεί στον βαθμό ομοιότητας της σύστασης των χρωμοσωμάτων τους λόγω εκτεταμένης ομολογίας.

Όπως προκύπτει από τον Πίνακα 13, η σχέση που συνδέει τις αποκλίσεις των *χαμηλής κάλυψης* ορθολόγων με τις αντίστοιχες αποκλίσεις των *υπολοίπων* ορθολόγων μπορεί να περιγραφεί από μοντέλα γραμμικής παλινδρόμησης, τα οποία στις περισσότερες των περιπτώσεων χαρακτηρίζονται από μεγάλη στατιστική σημαντικότητα (βλ. $r^2$ και τιμή $p$). Στις περισσότερες περιπτώσεις τα μοντέλα αυτά αντιστοιχούν σε ευθείες των οποίων η κλίση τείνει στη μονάδα και ο συντελεστής διεύθυνσης στο μηδέν. Ο βαθμός συσχέτιση των ασυμμετριών στα δύο υποσύνολα ορθολόγων είναι υψηλός, τόσο για τις μονο- και δι-νουκλεοτιδικές αποκλίσεις όσο και για τις αποκλίσεις των σταθμισμένων δινουκλεοτιδικών συχνοτήτων (βλ. Pearson's $r$). Επί παραδείγματι, οι αποκλίσεις G-C είναι σχεδόν ταυτόσημες στα δύο υποσύνολα (Εικόνα 13*b*). Η μεταξύ τους σχέση δίδεται από την εξίσωση $y = 0.939 \cdot x + 0.00275$ ($r^2 = 0.84$), όπου y οι αποκλίσεις G-C στα *χαμηλής κάλυψης* ορθόλογα και x οι αποκλίσεις G-C στο υποσύνολο των *υπολοίπων* ορθολόγων. Αντίστοιχα, η εξίσωση $y = 0.823 \cdot x - 0.0174$ ($r^2 = 0.747$) περιγράφει την σχέση μεταξύ των αποκλίσεων $\rho_{AG} - \rho_{CT}$ στα *χαμηλής κάλυψης* ορθόλογα (y) και στο υποσύνολο των *υπολοίπων* ορθολόγων (x). Συνεπώς, οι αποκλίσεις $\rho_{AG} - \rho_{CT}$ λαμβάνουν παραπλήσιες τιμές κατά μήκος του κάθε χρωμοσώματος, ανεξάρτητα από την εξελικτική προέλευση των γονιδίων στα οποία μετρώνται (Εικόνα 13*i*).

Το γραμμικό μοντέλο που περιγράφει την σχέση μεταξύ των αποκλίσεων $\rho_{GG} - \rho_{CC}$ στα *χαμηλής κάλυψης* ορθόλογα και στο υποσύνολο των *υπολοίπων* ορθολόγων (Εικόνα 13*k*) έχει συντελεστή προσδιορισμού ο οποίος λαμβάνει την χαμηλότερη τιμή ($r^2$=0.358) συγκριτικά με όλες τις υπόλοιπες περιπτώσεις που εξετάζουμε. Συνεπώς, ένα σημαντικό ποσοστό της διασποράς των τιμών $\rho_{GG} - \rho_{CC}$ στα *χαμηλής κάλυψης* ορθόλογα δεν μπορεί να συσχετιστεί με τις αντίστοιχες τιμές στο υποσύνολο των *υπολοίπων* ορθολόγων. Και σε αυτή την περίπτωση, ωστόσο, η γραμμική σχέση που συνδέει τις εν λόγω ασυμμετρίες στα δύο

υποσύνολα ορθολόγων ($y = 0.727 \cdot x + 0.0274$) εμφανίζει μεγάλη στατιστική σημαντικότητα, καθώς η τιμή *p* ισούται με $3.07 \cdot 10^{-27}$, είναι δηλαδή τάξεις μεγέθους μικρότερη του 0.001. Παρά λοιπόν τις διαφοροποιήσεις που εμφανίζουν οι αποκλίσεις, ανάλογα με την εξελικτική καταγωγή των κωδικών περιοχών στις οποίες υπολογίζονται, κατά κανόνα ακολουθούν παρόμοια γενικά πρότυπα σε κάθε δεδομένο χρωμόσωμα.

Συνολικά, τα αποτελέσματά μας υποδεικνύουν ότι τα *χαμηλής κάλυψης* ορθόλογα συνεισφέρουν στη διαμόρφωση του συνολικού προφίλ των αποκλίσεων με τρόπο παρόμοιο της συνεισφοράς των *υπολοίπων* ορθολόγων. Παρότι λοιπόν η εξελικτική ιστορία των *χαμηλής κάλυψης* ορθολόγων δεν παρακολουθεί τα γεγονότα διάσπασης της φυλογενετικής γραμμής (radiation events) που οδήγησαν στην γέννηση των ειδών που φέρουν τα εν λόγω γονίδια, οι ασυμμετρίες αυτών των γονιδίων συσχετίζονται με τις εξελικτικές σχέσεις των βακτηρίων. Συμπερασματικά, *οι ασυμμετρίες των κωδικών περιοχών συγκροτούν ανά είδος καθορισμένα πρότυπα, τα οποία αντανακλούν τις φυλογενετικές σχέσεις των βακτηρίων ακόμα και όταν μελετάμε αλληλουχίες οι οποίες δεν αποκλίνουν από έναν κοινό εξελικτικό πρόγονο*.

Μια πιθανή ερμηνεία των αποτελεσμάτων μας μπορεί να αναζητηθεί σε εκείνους τους μοριακούς μηχανισμούς που δρουν στην κλίμακα ολόκληρου του χρωμοσώματος, όπως είναι η αντιγραφή, η τροποποίηση και η επιδιόρθωση του DNA. Η δια-ειδική ποικιλότητα αυτών των μηχανισμών μπορεί να προκαλέσει μία συνολική αύξηση/μείωση των μεταλλακτικών ρυθμών ή ακόμα και να οδηγήσει σε διαφορετικά μεταλλακτικά προφίλ (Klasson & Andersson 2006), κατά τρόπο που είναι ανά είδος καθορισμένος. Το γεγονός αυτό, σε συνδυασμό με τις ασυμμετρίες των μεταλλακτικών ρυθμών μεταξύ κωδικού και μεταγραφόμενου κλώνου (Beletskii & Bhagwat 1996, Francino et al. 1996) μπορεί να οδηγήσει σε *πρότυπα αποκλίσεων κατά μήκος των κωδικών περιοχών τα οποία είναι διακριτά μεταξύ διαφορετικών ειδών αλλά παρόμοια μεταξύ γονιδίων του ίδιου οργανισμού*. Δεδομένου ότι οι ασυμμετρίες της σύστασης του DNA διαμορφώνονται σταδιακά με την πάροδο του χρόνου, τα πρότυπα των αποκλίσεων μπορεί να φέρουν πληροφορία σχετικά με την εξελικτική ιστορία των οργανισμών, όπως προκύπτει από την σύγκριση των κατανομών των αποκλίσεων μεταξύ διαφορετικών γονιδιωμάτων.

## **3.11** *Ασυμμετρίες της σύστασης ανά θέση κωδικονίων και η GC-πολωμένη δομή του γενετικού κώδικα*

### **3.11.1** *Γενικές παρατηρήσεις και αρχικές υποθέσεις*

Για την περαιτέρω κατανόηση των αιτιών που διέπουν την εμφάνιση των CDS-συζευγμένων αποκλίσεων, μελετήσαμε τις ασυμμετρίες της σύστασης ανά θέση κωδικονίων. Η ένταση της αρνητικής επιλογής (purifying selection) που ασκείται στις κωδικές περιοχές διαφέρει ανάλογα με την θέση των κωδικονίων που εξετάζουμε. Δεδομένου ότι 18 από τα 20 αμινοξέα κωδικοποιούνται από 2 έως 6 διαφορετικά κωδικόνια, διακρίνουμε μεταξύ συνώνυμων και μη-συνώνυμων θέσεων κωδικονίων. Οι πρώτες θέσεις των κωδικονίων είναι μη-συνώνυμες, με την εξαίρεση ορισμένων από τα κωδικόνια της Λευκίνης (T<u>TA</u>/C<u>TA</u>, T<u>TG</u>/C<u>TG</u>) και της Αργινίνης (C<u>GA</u>/A<u>GA</u>, C<u>GG</u>/A<u>GG</u>). Οι δεύτερες θέσεις των κωδικονίων είναι όλες μη-συνώνυμες. Ο εκφυλισμός του γενετικού κώδικα εντοπίζεται πρωτίστως στις τρίτες θέσεις των κωδικονίων και είναι είτε (*α*) τετραπλός, οπότε όλες οι υποκαταστάσεις είναι συνώνυμες, είτε (*β*) διπλός, οπότε μόνο οι μεταβάσεις (A↔G ή C↔T) είναι συνώνυμες. Έτσι, οι συνώνυμες τρίτες θέσεις διακρίνονται σε (*α*) $3^{ες}$ τετραπλά εκφυλισμένες ($3^{ες}|4$) και (*β*) $3^{ες}$ διπλά εκφυλισμένες ($3^{ες}|2$). Μοναδική εξαίρεση αποτελούν οι τρίτες θέσεις των κωδικονίων της Ισολευκίνης, οι οποίες είναι τριπλά εκφυλισμένες. Για τους υπολογισμούς που ακολουθούν, στα τετραπλά εκφυλισμένα κωδικόνια περιλαμβάνονται και τα CTN, CGN, και TCN που κωδικοποιούν για Leu, Arg και Ser, αντίστοιχα, ενώ στα διπλά εκφυλισμένα κωδικόνια περιλαμβάνονται και τα TTR, ATY, AGY και AGR που κωδικοποιούν για Leu, Ile, Ser και Arg, αντίστοιχα.

Η σύσταση των συνώνυμων θέσεων λαμβάνεται συχνά ως σημείο αναφοράς προκειμένου να εκτιμηθεί η επιλεκτική πίεση που ασκείται ενάντια στις μη-συνώνυμες μεταλλάξεις (Koonin & Wolf 2010). Εάν οι αποκλίσεις που μελετάμε προκύπτουν πρωτίστως ως το αποτέλεσμα ασύμμετρων μεταλλάξεων, θα πρέπει να εκδηλώνονται πιο έντονα στις συνώνυμες θέσεις των κωδικονίων, όπου δεν ασκείται αρνητική επιλογή στο επίπεδο του κωδικοποιούμενου αμινοξέος (Lobry & Sueoka 2002, Klasson & Andersson 2006, Necşulea & Lobry

2007). Εάν, από την άλλη, οι αποκλίσεις αυτές είναι το συνολικό αποτέλεσμα λειτουργικών περιορισμών που επιβάλλονται στην σύσταση των κωδικών περιοχών, θα αποτυπώνονται πιο έντονα στις μη-συνώνυμες θέσεις των κωδικονίων (Charneski et al. 2011) ή, όταν μετρώνται στις συνώνυμες θέσεις, θα συσχετίζονται με πρότυπα χρήσης συνώνυμων κωδικονίων (βλ. ενότητα 1.4.1.3).

Η σύσταση των $1^{ων}$ και $2^{ων}$ θέσεων είναι στενά συνδεδεμένη με την δομή και την λειτουργία των κωδικοποιούμενων πρωτεϊνών. Αντίθετα, η σύσταση των $3^{ων}|4$ και $3^{ων}|2$ θέσεων μπορεί να τείνει προς τις ισοδυναμίες [A] = [T] και [G] = [C] ή να αποκλίνει από αυτές δίχως να μεταβάλλονται οι συχνότητες των κωδικοποιούμενων αμινοξέων. Ωστόσο, και σε αντίθεση με ό,τι συμβαίνει στις $3^{ων}|4$ θέσεις, στην περίπτωση των $3^{ων}|2$ θέσεων συνώνυμες μεταλλάξεις είναι μόνο οι μεταβάσεις από τις *α*-βάσεις (A ή T) στις *γ*-βάσεις (G ή C) και αντίστροφα, οι οποίες τροποποιούν το GC περιεχόμενο του DNA. Οι μεταλλάξεις του τύπου *α*↔*γ* ασκούν έντονες GC κατευθύνουσες μεταλλακτικές πιέσεις (GC directional mutation pressure) και θεωρούνται ως η κύρια αιτία της ποικιλότητας του GC%, όπως προέκυψε αρχικά από βιοχημικά πειράματα (Cox & Yanofsky 1967) και επιβεβαιώθηκε περαιτέρω από ακόλουθες μελέτες (Jukes et al. 1987, Muto & Osawa 1987, Sueoka 1995, Palidwor et al. 2010). Συνεπώς, η σύσταση των $3^{ων}|4$ θέσεων μας επιτρέπει να εκτιμήσουμε τυχόν ασυμμετρίες στο σύνολο των μεταλλακτικών ρυθμών, ενώ η μελέτη των $3^{ων}|2$ εστιάζει στο υποσύνολο των GC κατευθυνουσών μεταλλάξεων.

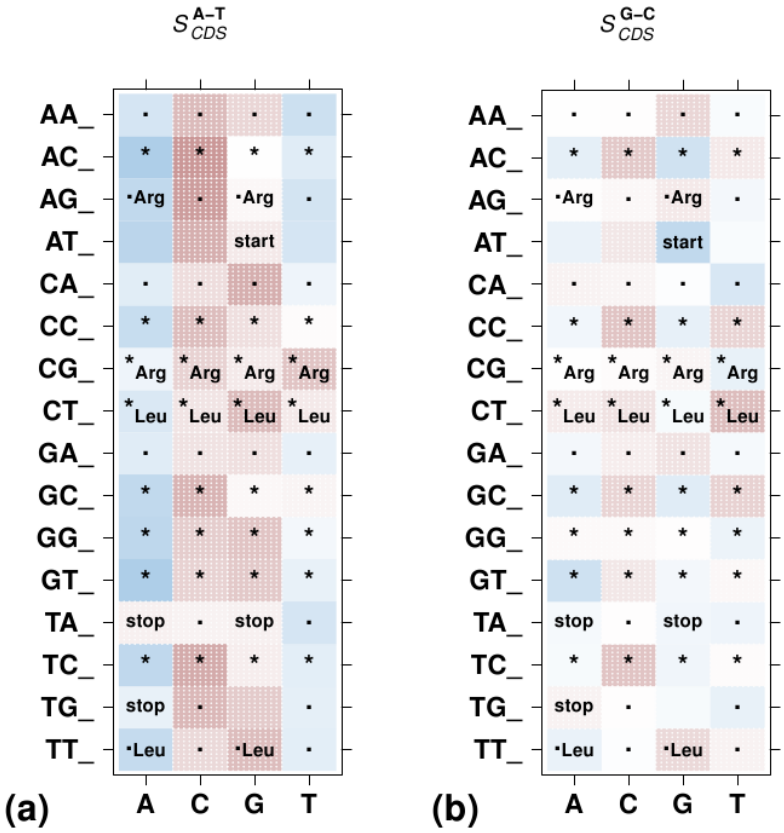
### **3.11.2** *Συσχετίσεις των αποκλίσεων με τα πρότυπα χρήσης κωδικονίων στα βακτηριακά χρωμοσώματα*

Προκειμένου να αξιολογήσουμε τα πιθανά αίτια που διαμορφώνουν τις CDS-συζευγμένες ασυμμετρίες, υπολογίσαμε τις αποκλίσεις A-T και G-C στις $1^{ες}$, $2^{ες}$, $3^{ες}|4$ και $3^{ες}|2$ θέσεις των κωδικονίων, για κάθε ένα από τα χρωμοσώματα της συλλογής μας. Οι αντίστοιχες μετρήσεις έγιναν κατά μήκος των κωδικών κλώνων. Ακολούθως, υπολογίσαμε τον δειγματικό συντελεστή γραμμικής συσχέτισης του Pearson (*r*) μεταξύ των αποκλίσεων (A-T ή G-C) σε κάθε θέση κωδικονίων και της συχνότητας εμφάνισης καθενός κωδικονίου, για

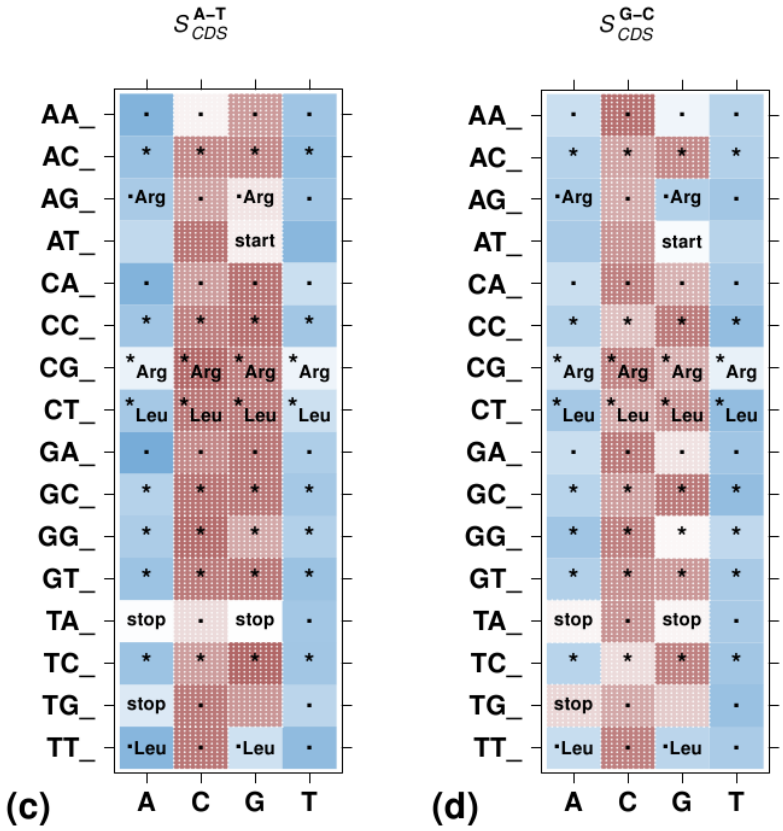
το σύνολο των χρωμοσωμάτων της συλλογής μας. Στην Εικόνα 14 παρουσιάζουμε τις αντίστοιχες τιμές των *r* του Pearson, χρησιμοποιώντας χάρτες θερμότητας (heatmaps).

Σύμφωνα με τα όσα αναφέραμε παραπάνω, στον βαθμό που οι προτιμήσεις στη χρήση συνώνυμων κωδικονίων εμπλέκονται στο φαινόμενο των ειδικών ανά κλώνο ασυμμετριών, αναμένουμε να εμφανίζονται ισχυρές συσχετίσεις μεταξύ των αποκλίσεων στις 3$^{ες}$|4 ή στις 3$^{ες}$|2 θέσεις και των προτύπων χρήσης κωδικονίων που ανήκουν στην ίδια ομάδα συνωνύμων. Εάν οι ασυμμετρίες των κωδικών περιοχών διαμορφώνονται πρωτίστως από επιλεκτικούς περιορισμούς, τότε οι αποκλίσεις στις 1$^{ες}$ και 2$^{ες}$ θέσεις των κωδικονίων αναμένεται να συσχετίζονται με την ταυτότητα των κωδικοποιούμενων αμινοξέων και συνεπώς να διακρίνουν μεταξύ διαφορετικών ομάδων συνωνύμων. Τυχόν συσχετίσεις των αποκλίσεων με ολόκληρες ομάδες συνωνύμων δεν μπορούν να αποδοθούν σε μεταλλακτικές ασυμμετρίες, καθώς δεν νοείται να συμβαίνουν συστηματικά μεταλλάξεις κατά τρόπο καθορισμένο από την ταυτότητα του κωδικοποιούμενου αμινοξέος (Sueoka 1995).

**3ες τεταπλά εκφυλισμένες θέσεις κωδικονίων**

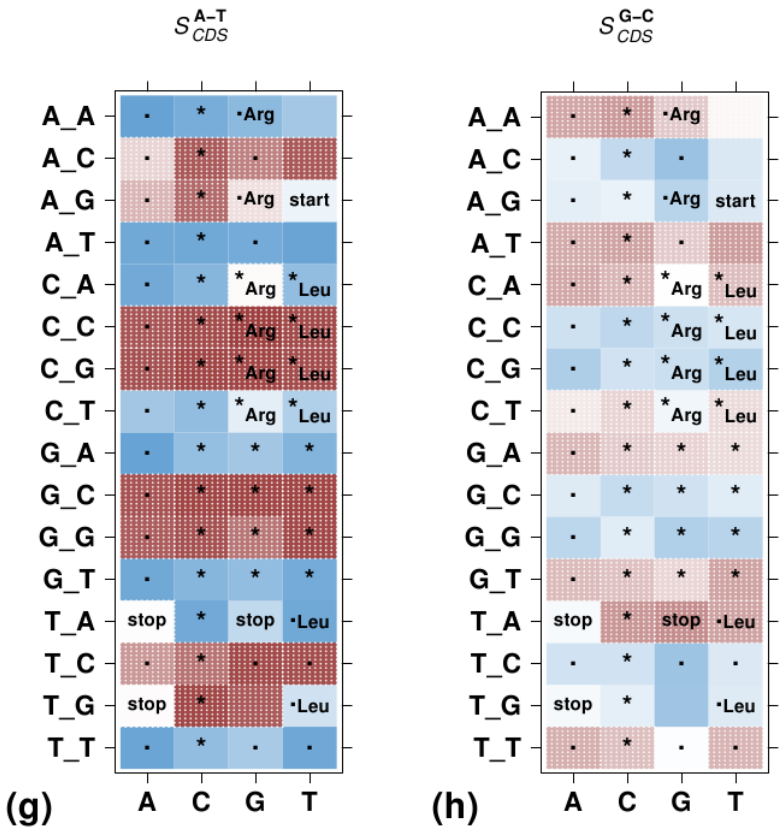


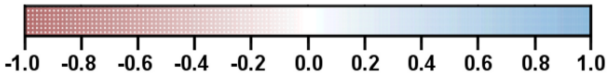

**3ες διπλά εκφυλισμένες θέσεις κωδικονίων**



**1ες θέσεις κωδικονίων**


**2ες θέσεις κωδικονίων**






**Εικόνα 14.** Χάρτες θερμότητας (heatmaps) του δειγματικού συντελεστή γραμμικής συσχέτισης του Pearson (r) μεταξύ της χρήσης κωδικονίων και των αποκλίσεων (A–T ή G–C), για το σύνολο των χρωμοσωμάτων της συλλογής μας. Οι αποκλίσεις υπολογίζονται σε συγκεκριμένες θέσεις κωδικονίων ($3^{ες}$ τετραπλά εκφυλισμένες: $3^{ες}|4$, $3^{ες}$ διπλά εκφυλισμένες: $3^{ες}|2$, $1^{ες}$ και $2^{ες}$). Σε κάθε χάρτη θερμότητας οι στήλες δηλώνουν την σύσταση των θέσεων των κωδικονίων στις οποίες μετρήθηκαν οι αποκλίσεις. Τα διάστικτα κόκκινα κελιά αντιστοιχούν σε αρνητικές τιμές του *r*, ενώ τα μπλε κελιά αντιστοιχούν σε θετικές τιμές του *r*. Η ένταση του χρωματισμού των κελιών υποδεικνύει το βαθμό συσχέτισης των αποκλίσεων με τη συχνότητα εμφάνισης του αντίστοιχου κωδικονίου, σύμφωνα με τον χρωματικό κώδικα που παρατίθεται στην λεζάντα.
Τα κωδικόνια έναρξης και λήξης δεν λαμβάνονται υπόψιν στους υπολογισμούς. Το ATG, που λειτουργεί και ως κωδικόνιο έναρξης, και σε ορισμένες περιπτώσεις το TGA, που λειτουργεί κατά κανόνα ως κωδικόνιο λήξης, εντοπίζονται επίσης εντός των κωδικών περιοχών, με το πρώτο να κωδικοποιεί για Μεθειονίνη και το δεύτερο για Τρυπτοφάνη ή Σεληνοκυστεΐνη. Ως αποτέλεσμα εμφανίζονται συσχετίσεις μεταξύ των αποκλίσεων και των [ATG] και [TGA]. Σε κάθε χάρτη θερμότητας επισημαίνονται τα κελιά που αντιστοιχούν στα εξαπλά εκφυλισμένα κωδικόνια της Λευκίνης και της Αργινίνης. Τα κωδικόνια που λαμβάνονται υπόψιν για τους υπολογισμούς των αποκλίσεων στις $3^{ες}|4$ και $3^{ες}|2$ θέσεις επισημαίνονται, αντίστοιχα, με τα σύμβολα "*****" (τετραπλά εκφυλισμένα κωδικόνια) και "·" (διπλά εκφυλισμένα κωδικόνια). "start": κωδικόνιο έναρξης, "stop": κωδικόνια λήξης.

### **3.11.2.1** *Αποκλίσεις στις $3^{ες}$ τετραπλά εκφυλισμένες θέσεις των κωδικονίων*

Τα αποτελέσματά μας δείχνουν πως οι συσχετίσεις ανάμεσα στην συχνότητα χρήσης των κωδικονίων και στις ασυμμετρίες της σύστασης των $3^{ων}|4$ θέσεων είναι ιδιαίτερα ασθενείς (Εικόνα 14*a*,*b*). Τα πρότυπα αυτών των συσχετίσεων εν μέρει αντιβαίνουν τα όσα αναμένονταν. Στο επίπεδο μιας πρώτης προσέγγισης, θα περιμέναμε ότι οι συχνότητες των τετραπλά εκφυλισμένων κωδικονίων που λήγουν σε A ή T θα σχετίζονταν, αντιστοίχως, θετικά ή αρνητικά με τις τιμές των A–T στις $3^{ες}|4$ θέσεις ($S_{3|4}^{A-T}$). Παρομοίως, οι συχνότητες των τετραπλά εκφυλισμένων κωδικονίων που λήγουν σε G ή C αναμενόταν ότι θα εμφανίζουν, αντιστοίχως, θετική ή αρνητική συσχέτιση με τις τιμές των G–C στις $3^{ες}|4$ θέσεις ($S_{3|4}^{G-C}$). Οι προβλέψεις αυτές επαληθεύονται μόνο εν μέρει, κυρίως σε ότι αφορά τις $S_{3|4}^{A-T}$ (Εικόνα 14*a*), ενώ

στην περίπτωση των $S_{3|4}^{G-C}$ τα πρότυπα των συσχετίσεων είναι μάλλον ασαφή (Εικόνα 14*b*).

Συγκεκριμένα, οι $S_{3|4}^{A-T}$ συσχετίζονται θετικά με τις συχνότητες των κωδικονίων που λήγουν σε A. Ωστόσο, και σε αντίθεση με τα αναμενόμενα, θετικές συσχετίσεις υπάρχουν και με τα κωδικόνια που λήγουν σε T, παρότι αυτές είναι συγκριτικά πιο ασθενείς. Επιπλέον, παρατηρούμε ότι οι τιμές των $S_{3|4}^{A-T}$ συσχετίζονται αρνητικά με τις συχνότητες των κωδικονίων που λήγουν σε C. Τα κατάλοιπα κυτοσίνης υφίστανται απαμίνωση με ρυθμούς που είναι υψηλότεροι στον κωδικό από ότι στον μεταγραφόμενο κλώνο των γονιδίων (Beletskii & Bhagwat 1996, Beletskii & Bhagwat 1998, Beletskii et al. 2000). Παρ' ότι οι αντιδράσεις απαμίνωσης εμπλέκονται στην εμφάνιση ασυμμετριών της σύστασης του DNA, υπάρχει διχογνωμία σχετικά με την έκταση των επιπτώσεων που επιφέρουν (Rocha & Danchin 2001). Συγκεκριμένα, η απαμίνωση της κυτοσίνης οδηγεί τελικά σε μεταβάσεις του τύπου C→T (Lindahl & Nyberg 1974, Frederico et al. 1990, Lindahl 1993) και συνεπώς, στο βαθμό που αυτές οι αντιδράσεις επιδρούν καθοριστικά στην διαμόρφωση των αποκλίσεις, η μείωση των συχνοτήτων των κωδικονίων που λήγουν σε C θα έπρεπε να συνοδεύεται από μία μετατόπιση προς χαμηλότερες τιμές $S_{3|4}^{A-T}$, σε αντίθεση με ό,τι παρατηρούμε (Εικόνα 14a).

Καθώς όλες οι υποκαταστάσεις στις $3^{ες}|4$ θέσεις είναι συνώνυμες και συνεπώς δεν ασκούνται επιλεκτικοί περιορισμοί στο επίπεδο των κωδικοποιούμενων αμινοξέων, εάν οι πολώσεις στην χρήση συνώνυμων κωδικονίων συνδέονταν συστηματικά με ασυμμετρίες της σύστασης των κωδικών περιοχών οι συσχετίσεις που εμφανίζονται στους αντίστοιχους χάρτες θερμότητας (Εικόνα 14*a*,*b*) θα ήταν πιο έντονες από τις παρατηρούμενες. Επίσης, οι αποκλίσεις των $3^{ων}|4$ θέσεων θα αποκρίνονταν στις μεταβολές των συχνοτήτων των κωδικονίων κατά τρόπο που θα διέκρινε μεταξύ των μελών της ίδιας ομάδας συνωνύμων. Ωστόσο, τέτοια πρότυπα συσχετίσεων δεν εντοπίζονται.

*Τα αποτελέσματά μας υποδηλώνουν ότι οι ασυμμετρίες των κωδικών περιοχών δεν συνδέονται άμεσα με τις προτιμήσεις συνώνυμων κωδικονίων*. Στο ίδιο συμπέρασμα συγκλίνει και το γεγονός ότι οι συσχετίσεις που παρατηρούμε, παρότι ασθενείς, δεν περιορίζονται στις περιπτώσεις των τετραπλά εκφυλισμένων κωδικονίων. Ωστόσο, μονάχα οι τρίτες θέσεις αυτών των κωδικονίων λαμβάνονται υπόψιν για τον υπολογισμό των αποκλίσεων των $3^{ων}|4$ θέσεων. Έτσι, η συσχέτιση των $S_{3|4}^{A-T}$ με κωδικόνια που λήγουν σε A ή C και δεν

είναι τετραπλά εκφυλισμένα θέτει υπό περαιτέρω αμφισβήτηση την ύπαρξη αιτιακής σχέσης ανάμεσα στις αποκλίσεις που μελετάμε και στην χρήση κωδικονίων.

### **3.11.2.2** *Αποκλίσεις στις* $3^{ες}$ *διπλά εκφυλισμένες θέσεις των κωδικονίων*

Στις $3^{ες}|2$ θέσεις αναδεικνύονται πιο σαφή πρότυπα συσχέτισης μεταξύ των ασυμμετριών της νουκλεοτιδικής σύστασης και των συχνοτήτων εμφάνισης των κωδικονίων (Εικόνα 14*c*,*d*). Τα πρότυπα αυτά οργανώνονται σύμφωνα με το GC περιεχόμενο των τρίτων θέσεων των κωδικονίων. Συγκεκριμένα, τόσο οι αποκλίσεις A-T ($S_{3|2}^{A-T}$) όσο και οι G-C ($S_{3|2}^{G-C}$) συσχετίζονται θετικά με τα κωδικόνια που λήγουν σε A/T και αρνητικά με εκείνα που λήγουν σε G/C. Οι $S_{3|2}^{A-T}$ και $S_{3|2}^{G-C}$ αποκρίνονται στις μεταβολές των συχνοτήτων σχεδόν όλων των κωδικονίων και όχι μόνο των διπλά εκφυλισμένων, στα οποία έγιναν και οι σχετικές μετρήσεις. Συνεπώς, όπως και στην περίπτωση των $3^{ων}|4$ θέσεων, οι προτιμήσεις στην χρήση συνώνυμων κωδικονίων δεν προσφέρονται για την ερμηνεία του συνόλου των παρατηρούμενων συσχετίσεων.

Ο διπλός εκφυλισμός των τρίτων θέσεων των κωδικονίων επιτρέπει μόνο υποκαταστάσεις που μεταβάλουν το GC περιεχόμενο του DNA χωρίς να αλλάζει το κωδικοποιούμενο αμινοξύ. Σύμφωνα με μελέτες που αφορούσαν την ποικιλότητα και την ετερογένεια της σύστασης του DNA προτάθηκε ότι οι GC κατευθύνουσες μεταλλακτικές πιέσεις, όπως εκφράζονται από τον λόγο των μεταλλάξεων A:T→G:C προς G:C→A:T, μπορούν να επηρεάζουν αποφασιστικά τα συνολικά χαρακτηριστικά της σύστασης των κωδικών περιοχών (Sueoka 1961, Sueoka 1962, Sueoka 1988, Sueoka 1992, Sueoka 1999) (βλ. Ενότητα 1.9). Αυτό συνεπάγεται ότι, *δεδομένης της δομής του γενετικού κώδικα, οι μεταλλάξεις που μεταβάλλουν το GC% κατευθύνουν την χρήση των κωδικονίων και ως εκ τούτου διαμορφώνουν τα χαρακτηριστικά της σύστασης των κωδικών περιοχών, συμπεριλαμβανομένων και των A-T και G-C αποκλίσεων*. Αναδιατυπώνοντας, η απόκριση των $S_{3|2}^{A-T}$ και $S_{3|2}^{G-C}$ στις συχνότητες εμφάνισης των κωδικονίων δίνει σαφή πρότυπα συσχετίσεων, όπως περιγράφεται από τις υψηλές κατ' απόλυτο τιμές του *r*, τα οποία μπορούν να ερμηνευθούν ως απλή αντανάκλαση των GC κατευθυνουσών μεταλλακτικών πιέσεων στο επίπεδο της σύστασης των κωδικών περιοχών.

**3.11.2.3** *Αποκλίσεις στις 1ες και 2ες θέσεις των κωδικονίων*

Ακολούθως, εξετάζουμε τις αποκλίσεις A-T και G-C στις 1ες ($S_1^{A-T}$, $S_1^{G-C}$) και 2ες ($S_2^{A-T}$, $S_2^{G-C}$) θέσεις των κωδικονίων (Εικόνα 14e,f,g,h). Η απόκριση των αποκλίσεων στις συχνότητες εμφάνισης των κωδικονίων οδηγεί σε σαφώς δομημένα πρότυπα συσχετίσεων, τα οποία ωστόσο αντιβαίνουν εκ πρώτης τα αναμενόμενα. Δεν υπάρχει προφανής συσχέτιση ανάμεσα στις ασυμμετρίες της σύστασης και στην ταυτότητα των βάσεων που καταλαμβάνουν τις θέσεις των κωδικονίων τις οποίες μελετάμε. Αντιθέτως, *για δεδομένη θέση κωδικονίων, οι αποκλίσεις αποκρίνονται με παρόμοιο τρόπο στα κωδικόνια που φέρουν την ίδια βάση στην τρίτη τους θέση, ανεξαρτήτως της σύστασής τους στην 1η και 2η θέση*. Οι συσχετίσεις αυτές είναι πιο έντονες στην 1η θέση για τις αποκλίσεις G-C (Εικόνα 14*f*) και στην 2η θέση για τις αποκλίσεις A-T (Εικόνα 14*g*). Συγκεκριμένα, οι αποκλίσεις G-C στην 1η θέση και οι αποκλίσεις A-T στην 1η και 2η θέση των κωδικονίων μετατοπίζονται προς υψηλότερες τιμές όσο αυξάνονται οι συχνότητες των κωδικονίων που λήγουν σε A/T και μειώνονται οι συχνότητες των κωδικονίων που λήγουν σε G/C. Το αντίστροφο πρότυπο παρατηρείται στην περίπτωση των $S_2^{G-C}$, που εμφανίζουν θετική συσχέτιση με τα κωδικόνια που λήγουν σε G/C και αρνητική συσχέτιση με όσα λήγουν σε A/T (Εικόνα 14*h*). Συνολικά, η συσχέτιση των ασυμμετριών της σύστασης των κωδικών περιοχών με το GC% των τρίτων θέσεων των κωδικονίων, την ύπαρξη της οποίας διαπιστώσαμε προηγουμένως στις 3ες|2 θέσεις, επανεμφανίζεται όταν μελετάμε τις 1ες και 2ες θέσεις των κωδικονίων, παρόλο που οι θέσεις αυτές υφίστανται επιλεκτικούς περιορισμούς, καθώς η σύστασή τους καθορίζει τα κωδικοποιούμενα αμινοξέα.

**3.11.2.4** *Γενικά πρότυπα συσχέτισης μεταξύ αποκλίσεων και χρήσης κωδικονίων*

Συνοψίζοντας τα αποτελέσματα που αναφέρονται στα γονιδιώματα της συλλογής μας (Εικόνα 14), στις περιπτώσεις όπου υπάρχει σημαντική συσχέτιση μεταξύ της χρήσης κωδικονίων και των ασυμμετριών της σύστασης μιας δεδομένης θέσης (εν προκειμένω της 1ης, 2ης και 3ης|2 θέσης), τα παρατηρούμενα πρότυπα διακρίνουν μεταξύ των κωδικονίων που λήγουν σε A/T και εκείνων που λήγουν σε G/C. Τότε, οι αποκλίσεις συσχετίζονται θετικά με την μία ομάδα κωδικονίων και αρνητικά με την άλλη. Εξαιρέσεις, που παραβιάζουν αυτή την εμπειρική σχέση, αποτελούν το οπάλ κωδικόνιο λήξης (opal stop codon), TGA, (βλ. Εικόνα 14d) και ορισμένα από τα κωδικόνια της Λευκίνης και της

Αργινίνης. Σε ορισμένα βακτήρια το TGA μπορεί να βρίσκεται εντός του πλαισίου ανάγνωσης των γονιδίων, οπότε κωδικοποιεί για Τρυπτοφάνη ή Σεληνοκυστεΐνη. Ωστόσο, το TGA σηματοδοτεί πρωτίστως την λήξη της μετάφρασης και συνεπώς η συχνότητα εμφάνισής του υπόκειται σε αυστηρούς επιλεκτικούς περιορισμούς, καθώς οι μη νοηματικές μεταλλάξεις (nonsense mutations) που οδηγούν σε πρώιμα κωδικόνια λήξης αποφεύγονται έντονα. Η Λευκίνη και η Αργινίνη αποτελούν τα μόνα αμινοξέα που κωδικοποιούνται από κωδικόνια τα οποία επιτρέπουν συνώνυμες μεταλλάξεις που μεταβάλλουν το GC% όχι μόνο στην τρίτη θέση τους, αλλά επίσης και στην 1$^{η}$ θέση. Οι Palidwor *et al*. (2010) έδειξαν ότι αυτή η ιδιαιτερότητα της δομής του γενετικού κώδικα οδηγεί σε μη τυπικά πρότυπα χρήσης των κωδικονίων της Λευκίνης και της Αργινίνης, συνεπεία της απόκρισής τους σε GC κατευθύνουσες μεταλλακτικές πιέσεις.

Η εμπειρική σχέση που περιγράψαμε υποδεικνύει μία *έμμεση σύνδεση ανάμεσα στις ασυμμετρίες των κωδικών περιοχών και στην χρήση κωδικονίων*. Ένα συνεκτικό ερμηνευτικό σχήμα μπορεί να αναζητηθεί σε προηγούμενες μελέτες σχετικά με τις μεταλλακτικές πολώσεις που διαμορφώνουν το GC% του DNA (Sueoka 1961, Muto & Osawa 1987, Knight et al. 2001, Hu et al. 2007, Goodarzi et al. 2008, Sorimachi & Okayasu 2008). Μελέτες όπως αυτές τεκμηρίωσαν την συσχέτιση του GC περιεχομένου των κωδικών περιοχών με τις συχνότητες των κωδικονίων καθώς επίσης και με την αμινοξική σύσταση των κωδικοποιούμενων πρωτεϊνών (βλ. ενότητα 1.9.3). Διαπιστώθηκε ότι είδη με παραπλήσιο GC περιεχόμενο εμφανίζουν όμοια πρότυπα χρήσης κωδικονίων και αμινοξέων. Η παρατήρηση αυτή σε συνδυασμό με το γεγονός ότι οι συχνότητες των κωδικονίων μπορούν να οργανωθούν με πολλούς διαφορετικούς συνδυασμούς που αντιστοιχούν στο ίδιο GC%, συνεπάγεται μια συγκεκριμένη φορά της αιτιότητας που διέπει τις συσχετίσεις του GC περιεχομένου και της χρήσης κωδικονίων (Knight et al. 2001). Συγκεκριμένα, *δεν είναι οι επιλεκτικές προτιμήσεις σε συγκεκριμένα κωδικόνια ή ακόμα και σε συγκεκριμένα αμινοξέα που οδηγούν σε ένα, ορισμένο κάθε φορά, GC περιεχόμενο· αντίθετα, είναι οι GC κατευθύνουσες μεταλλακτικές πιέσεις οι οποίες, καθώς συνυφαίνονται με την δομή του γενετικού κώδικα και τροποποιούνται από επιλεκτικούς περιορισμούς που δρουν στο επίπεδο της πρωτεϊνικής δομής και λειτουργίας* (Yu 2007), *διαμορφώνουν τις συχνότητες των κωδικονίων και κατ' επέκταση την νουκλεοτιδική σύσταση των κωδικών περιοχών* (Zhang & Yu 2010). Στο πλαίσιο αυτό, οι συχνότητες των κωδικονίων και οι ασυμμετρίες στην σύσταση των 1$^{ων}$,

$2^{ων}$ και $3^{ων}|2$ θέσεών τους δεν συνδέονται αιτιακά και οι μεταξύ τους συσχετίσεις μπορούν να αναχθούν στην κοινή τους εξελικτική ρίζα, δηλαδή στις GC κατευθύνουσες μεταλλακτικές πιέσεις. Υπό αυτό το πρίσμα είναι δυνατόν να ερμηνευθούν και τα πρότυπα των παρατηρούμενων συσχετίσεων, τα οποία οργανώνονται σύμφωνα με την παρουσία G/C ή A/T στην τρίτη θέση των κωδικονίων, ακόμα και όταν οι αποκλίσεις μετρώνται στις $1^{ες}$ και $2^{ες}$ θέσεις. *Στην χρονική κλίμακα της εξέλιξης οι αλλαγές στο GC% πραγματοποιούνται ταχύτερα στις $3^{ες}$ από ότι στις $1^{ες}$ και $2^{ες}$ θέσεις των κωδικονίων και συνεπώς αντανακλούν με μεγαλύτερη ακρίβεια τις πολώσεις των μεταλλλακτικών ρυθμών που οδηγούν προς ένα συγκεκριμένο GC περιεχόμενο στο επίπεδο ολόκληρου του χρωμοσώματος* (Knight et al. 2001).

**3.11.3** *Ένα μοντέλο χρήσης κωδικονίων που ενσωματώνει τις GC-πολώσεις εντός κάθε ομάδας συνωνύμων*

Προκειμένου να ελέγξουμε την ορθότητα της συλλογιστικής που αναπτύξαμε στην αμέσως προηγούμενη ενότητα (3.11.2.4), κατασκευάσαμε ένα απλό μοντέλο βάσει του οποίου η χρήση κωδικονίων μίας τεχνητής αλληλουχίας προσδιορίζεται αποκλειστικά από το GC περιεχόμενό της, λαμβάνοντας υπόψιν την ποικιλότητα σε GC που εμφανίζεται εντός κάθε ομάδας συνωνύμων. Στο μοντέλο αυτό η απόκριση ενός κωδικονίου, *i*, στο GC περιεχόμενο της αλληλουχίας εκφράζεται ως συνάρτηση τριών παραμέτρων:

*α*) του $n_i$, που είναι ο αριθμός των G και C του κωδικονίου *i*

*β*) του $R_i^{GC}$, που δίδεται από τον τύπο

$$R_i^{GC} = \frac{GC_i}{\frac{1}{a_i}\sum_{j=1}^{a_i} GC_j}$$

όπου $GC_i$ το GC περιεχόμενο του κωδικονίου *i* και $a_i$ ο αριθμός των κωδικονίων της αντίστοιχης ομάδας συνωνύμων, και

*γ*) του $R_i^{AT}$, που δίδεται από τον τύπο

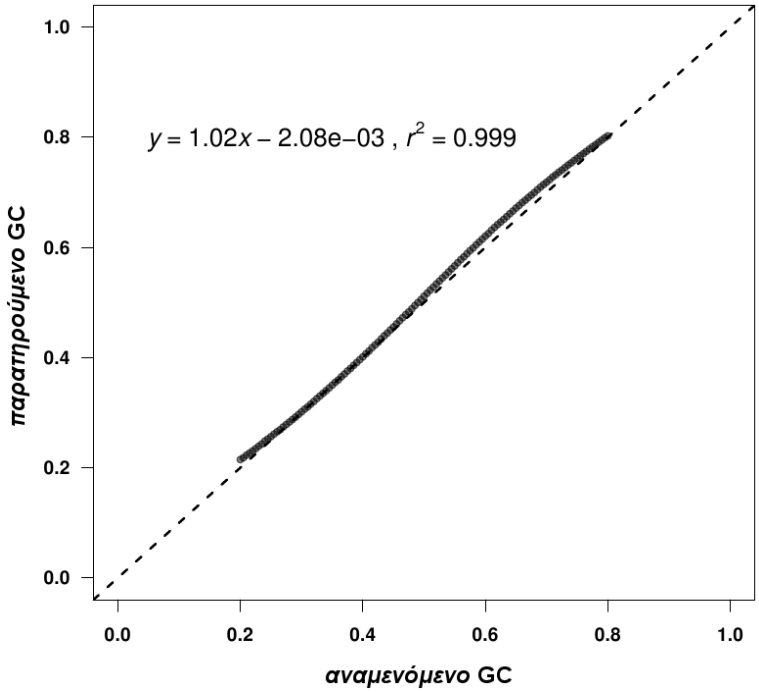
$$R_i^{AT} = \frac{AT_i}{\frac{1}{a_i}\sum_{j=1}^{a_i} AT_j}$$

όπου $AT_i = 1 - GC_i$. Οι παράμετροι $R_i^{GC}$ και $R_i^{AT}$ αποδίδουν την διαφορική απόκριση των συνώνυμων κωδικονίων στο GC περιεχόμενο της αλληλουχίας, δεδομένου του μέσου GC περιεχομένου της ομάδας στην οποία ανήκουν. Συνεπώς, *τα* $R_i^{GC}$ *και* $R_i^{AT}$ *εισάγουν στο μοντέλο μας GC-πολώσεις μεταξύ των συνώνυμων κωδικονίων* (*synonymous-codon GC biases*). Κάθε τεχνητή αλληλουχία αντιστοιχεί σε ένα δεδομένο GC περιεχόμενο (*GC*) και απαρτίζεται από $10^6$ κωδικόνια. Η πιθανότητα εμφάνισης ενός κωδικονίου *i* ($p_i$) υπολογίζεται ως ακολούθως:

$$p_i = (GC)^{n_i R_i^{GC}} (1 - GC)^{(3 - n_i) R_i^{AT}}$$

κανονικοποιημένη ώστε όλα τα $p_i$ να αθροίζονται στην μονάδα. Η ανωτέρω σχέση περιγράφει το GC-πολωμένο μοντέλο μας (GC-biased model).

Κατασκευάσαμε ένα σύνολο τεχνητών αλληλουχιών, εισάγοντας στην συνάρτηση των $p_i$ τιμές για το GC περιεχόμενο που κυμαίνονται από 0.2 έως 0.8, με βήμα ίσο με 0.005. Προκειμένου να αξιολογήσουμε την ευστάθεια του μοντέλου μας, εξετάζουμε κατά πόσον το GC που εισάγουμε στην συνάρτηση των $p_i$ (*αναμενόμενο* GC) συμπίπτει με το GC που μετράμε τελικά στις αντίστοιχες τεχνητές αλληλουχίες (*παρατηρούμενο* GC). Στην Εικόνα 15 αναπαριστούμε το *παρατηρούμενο* GC συναρτήσει του *αναμενόμενου* GC. Τα σημεία του διαγράμματος βρίσκονται πάνω ή πολύ κοντά στην διαγώνιο. Η αντίστοιχη σχέση γραμμικής παλινδρόμησης εμφανίζει κλίση περίπου ίση με την μονάδα (1.02), συντελεστή διεύθυνσης που προσεγγίζει το μηδέν ($-2.08 \times 10^{-3}$) και συντελεστής προσδιορισμού $r^2=0.999$. Συνεπώς, το *παρατηρούμενο* GC προσεγγίζει με ιδιαίτερα ικανοποιητική ακρίβεια το *αναμενόμενο* GC, καταδεικνύοντας ότι το μοντέλο μας χαρακτηρίζεται από εσωτερική συνοχή.




**Εικόνα 15.** Διάγραμμα διασποράς του GC περιεχομένου, όπως προκύπτει από μετρήσεις στις τεχνητές αλληλουχίες (*παρατηρούμενο* GC), έναντι του αντίστοιχου GC περιεχομένου που εισάγεται στην συνάρτηση της πιθανότητας εμφάνισης των κωδικονίων, $p_i$, (*αναμενόμενο* GC). Τα εικονιζόμενα σημεία κείνται πάνω ή πολύ κοντά στην διαγώνιο (διακεκομμένη γραμμή), ήτοι το GC περιεχόμενο που εισάγεται στην συνάρτηση $p_i$ ισούται κατά προσέγγιση με το GC περιεχόμενο της προκύπτουσας τεχνητής αλληλουχίας ($r^2$=0.999). Δίδεται επίσης η σχέση που περιγράφει την γραμμική παλινδρόμηση του *παρατηρούμενου* GC πάνω στο αναμενόμενο GC.

Το μοντέλο που περιγράφουμε ενσωματώνει στην διατύπωσή του έναν ελάχιστο αριθμό παραμέτρων και δεν αποσκοπεί στην πρόβλεψη της χρήσης κωδικονίων που πραγματικά παρατηρείται στα βακτηριακά χρωμοσώματα. Αντ' αυτού, παρέχει την βάση ώστε να εκτιμήσουμε εάν υπάρχει μία αιτιακή σύνδεση μεταξύ των ειδικών ανά κλώνο ασυμμετριών στις κωδικές περιοχές και της ποικιλότητας του GC περιεχομένου αυτών των περιοχών σε συνδυασμό με την δομή του γενετικού κώδικα. Οι αριθμητικές τιμές των παραμέτρων του μοντέλου για κάθε κωδικόνιο *i* δίδονται στον Πίνακα 14. Οι παράμετροι $R_i^{GC}$, $R_i^{AT}$ και $n_i$ είναι *συμμετρικές όσον αφορά τις συμπληρωματικές νουκλεοτιδικές βάσεις*. Κωδικόνια που ανήκουν σε ομάδες συνωνύμων με το ίδιο μέσο GC αντιστοιχίζονται σε παραμέτρους με τις ίδιες αριθμητικές τιμές εάν έχουν τον ίδιο αριθμό S (G ή C) και W (A ή T) βάσεων, χωρίς να διακρίνουν μεταξύ G και C ή A και T. Συνεπώς, *τυχόν ασυμμετρίες στην σύσταση των τεχνητών*

*αλληλουχιών καθώς επίσης και οι συσχετίσεις τους με την χρήση κωδικονίων, εάν υπάρχουν, δεν μπορεί παρά να απορρέουν μονάχα από τον διαμερισμό των κωδικονίων σε ομάδες συνωνύμων και από την ποικιλότητα του GC σε αυτές τις ομάδες.*

**ΠΙΝΑΚΑΣ 14. Το μέσο GC% για κάθε ομάδα συνωνύμων και οι παράμετροι $R_i^{GC}$, $R_i^{AT}$ και $n_i$ του μοντέλου για κάθε κωδικόνιο *i***

| | | | |
|---|---|---|---|
| **TTT F (16.7, 0.000, 1.200, 0)** | **TCT S (50.0, 0.667, 1.333, 1)** | **TAT Y (16.7, 0.000, 1.200, 0)** | **TGT C (50.0, 0.667, 1.333, 1)** |
| **TTC F (16.7, 2.000, 0.800, 1)** | **TCC S (50.0, 1.333, 0.667, 2)** | **TAC Y (16.7, 2.000, 0.800, 1)** | **TGC C (50.0, 1.333, 0.667, 2)** |
| **TTA L (38.9, 0.000, 1.636, 0)** | **TCA S (50.0, 0.667, 1.333, 1)** | **TAA * Ter** | **TGA * Ter** |
| **TTG L (38.9, 0.857, 1.091, 1)** | **TCG S (50.0, 1.333, 0.667, 2)** | **TAG * Ter** | **TGG W (66.7, 1.000, 1.000, 2)** |
| | | | |
| **CTT L (38.9, 0.857, 1.091, 1)** | **CCT P (83.3, 0.800, 2.000, 2)** | **CAT H (50.0, 0.667, 1.333, 1)** | **CGT R (72.2, 0.923, 1.200, 2)** |
| **CTC L (38.9, 1.714, 0.545, 2)** | **CCC P (83.3, 1.200, 0.000, 3)** | **CAC H (50.0, 1.333, 0.667, 2)** | **CGC R (72.2, 1.385, 0.000, 3)** |
| **CTA L (38.9, 0.857, 1.091, 1)** | **CCA P (83.3, 0.800, 2.000, 2)** | **CAA Q (50.0, 0.667, 1.333, 1)** | **CGA R (72.2, 0.923, 1.200, 2)** |
| **CTG L (38.9, 1.714, 0.545, 2)** | **CCG P (83.3, 1.200, 0.000, 3)** | **CAG Q (50.0, 1.333, 0.667, 2)** | **CGG R (72.2, 1.385, 0.000, 3)** |
| | | | |
| **ATT I (11.1, 0.000, 1.125, 0)** | **ACT T (50.0, 0.667, 1.333, 1)** | **AAT N (16.7, 0.000, 1.200, 0)** | **AGT S (50.0, 0.667, 1.333, 1)** |
| **ATC I (11.1, 3.000, 0.750, 1)** | **ACC T (50.0, 1.333, 0.667, 2)** | **AAC N (16.7, 2.000, 0.800, 1)** | **AGC S (50.0, 1.333, 0.667, 2)** |
| **ATA I (11.1, 0.000, 1.125, 0)** | **ACA T (50.0, 0.667, 1.333, 1)** | **AAA K (16.7, 0.000, 1.200, 0)** | **AGA R (72.2, 0.462, 2.400, 1)** |
| **ATG M (33.3, 1.000, 1.000, 1)** | **ACG T (50.0, 1.333, 0.667, 2)** | **AAG K (16.7, 2.000, 0.800, 1)** | **AGG R (72.2, 0.923, 1.200, 2)** |
| | | | |
| **GTT V (50.0, 0.667, 1.333, 1)** | **GCT A (83.3, 0.800, 2.000, 2)** | **GAT D (50.0, 0.667, 1.333, 1)** | **GGT G (83.3, 0.800, 2.000, 2)** |
| **GTC V (50.0, 1.333, 0.667, 2)** | **GCC A (83.3, 1.200, 0.000, 3)** | **GAC D (50.0, 1.333, 0.667, 2)** | **GGC G (83.3, 1.200, 0.000, 3)** |
| **GTA V (50.0, 0.667, 1.333, 1)** | **GCA A (83.3, 0.800, 2.000, 2)** | **GAA E (50.0, 0.667, 1.333, 1)** | **GGA G (83.3, 0.800, 2.000, 2)** |
| **GTG V (50.0, 1.333, 0.667, 2)** | **GCG A (83.3, 1.200, 0.000, 3)** | **GAG E (50.0, 1.333, 0.667, 2)** | **GGG G (83.3, 1.200, 0.000, 3)** |

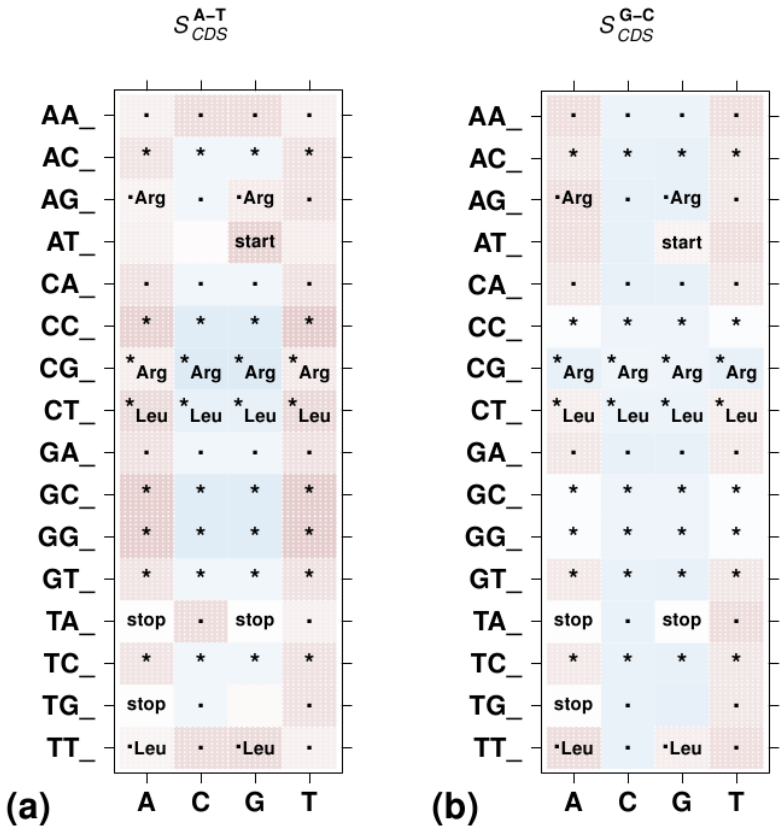
ΣΗΜΕΙΩΣΕΙΣ.– Σε κάθε κωδικόνιο *i* αντιστοιχεί ένα διάνυσμα (*x*, *y*, *z*, *n*), όπου *x* το μέσο GC% της ομάδας των συνωνύμων του και *y*, *z*, *n* οι παράμετροι $R_i^{GC}$, $R_i^{AT}$ και $n_i$ του μοντέλου για το κωδικόνιο *i*. $n_i$: ο αριθμός των G και C του κωδικονίου *i*. $R_i^{GC}$: ο λόγος του GC περιεχομένου του κωδικονίου *i* προς το μέσο GC περιεχόμενο της ομάδας των συνωνύμων του. $R_i^{AT}$: ο λόγος του AT περιεχομένου του κωδικονίου *i* προς το μέσο AT περιεχόμενο της ομάδας των συνωνύμων του. Τα $R_i^{GC}$ και $R_i^{AT}$ δηλώνουν, αντιστοίχως, εάν το GC και AT περιεχόμενο του κωδικονίου *i* είναι μεγαλύτερο ($R_i^{GC} > 1$, $R_i^{AT} > 1$) ή μικρότερο ($R_i^{GC} < 1$, $R_i^{AT} < 1$) του μέσου GC και AT περιεχομένου της ομάδας των συνωνύμων του. Έτσι, *τα $R_i^{GC}$ και $R_i^{AT}$ εγγράφουν στο μοντέλο τις πολώσεις στην ποικιλότητα του GC μεταξύ συνώνυμων κωδικονίων.*

**3.11.3.1** *Συσχετίσεις των αποκλίσεων με τα πρότυπα χρήσης κωδικονίων στις τεχνητές αλληλουχίες*

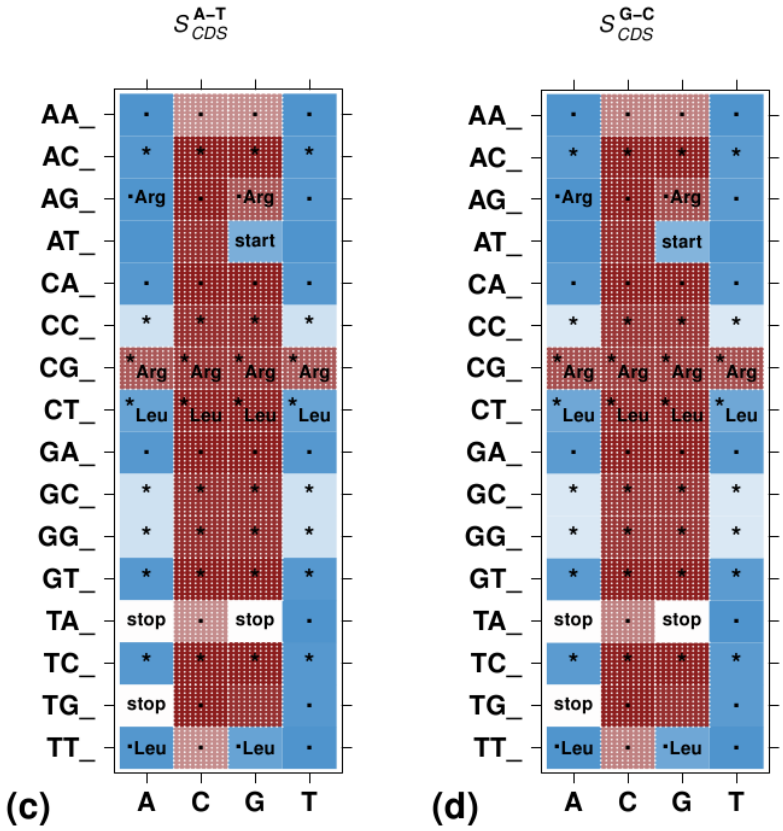
Ακολούθως, μετρήσαμε τις αποκλίσεις A-T και G-C στις 1ες, 2ες, 3ες|4 και 3ες|2 θέσεις των κωδικονίων, για κάθε μία από τις τεχνητές αλληλουχίες που κατασκευάσαμε. Όπως και στην ανάλυση που αφορούσε στα γονιδιώματα της συλλογής μας, υπολογίσαμε τον δειγματικό συντελεστή γραμμικής συσχέτισης του Pearson (*r*) μεταξύ των αποκλίσεων (A-T ή G-C) και της συχνότητας εμφάνισης καθενός κωδικονίου, για το σύνολο των τεχνητών αλληλουχιών (Εικόνα 16).

Στους αντίστοιχους χάρτες θερμότητας δεν εμφανίζονται σημαντικές συσχετίσεις μεταξύ της χρήσης κωδικονίων και των αποκλίσεων στις 3ες|4 θέσεις (Εικόνα 16*a*,*b*). Αντίθετα, *όταν εξετάζουμε τις* 3ες|2, *τις* 1ες *και τις* 2ες *θέσεις, το GC-πολωμένο μοντέλο αναπαράγει εν πολλοίς τα πρότυπα των συσχετίσεων που εντοπίσαμε στα βακτηριακά γονιδιώματα ανάμεσα στις αποκλίσεις και στις συχνότητες εμφάνισης των κωδικονίων που λήγουν σε A/T ή G/C*. Συγκεκριμένα, στις 3ες|2 και στις 1ες θέσεις (Εικόνα 16*c*,*d* και *e*,*f*, αντίστοιχα) των τεχνητών αλληλουχιών οι τιμές των αποκλίσεων A-T και G-C σχετίζονται θετικά με τις συχνότητες των κωδικονίων που λήγουν σε A/T και αρνητικά με τις συχνότητες των κωδικονίων που λήγουν σε G/C, όπως συμβαίνει και στα βακτηριακά χρωμοσώματα (πρβ. Εικόνα 16*c*-*f* και Εικόνα 14*c*-*f*). Τα πρότυπα των συσχετίσεων αντιστρέφονται στις 2ες θέσεις των τεχνητών αλληλουχιών (Εικόνα 16*g*,*h*).

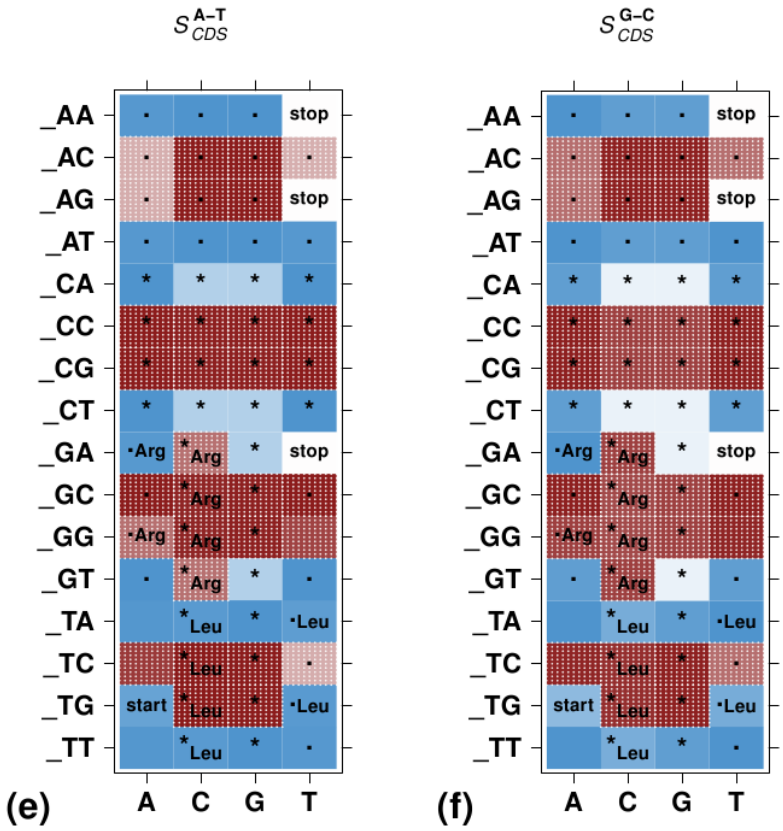
**3ες τετραπλά εκφυλισμένες θέσεις κωδικονίων**



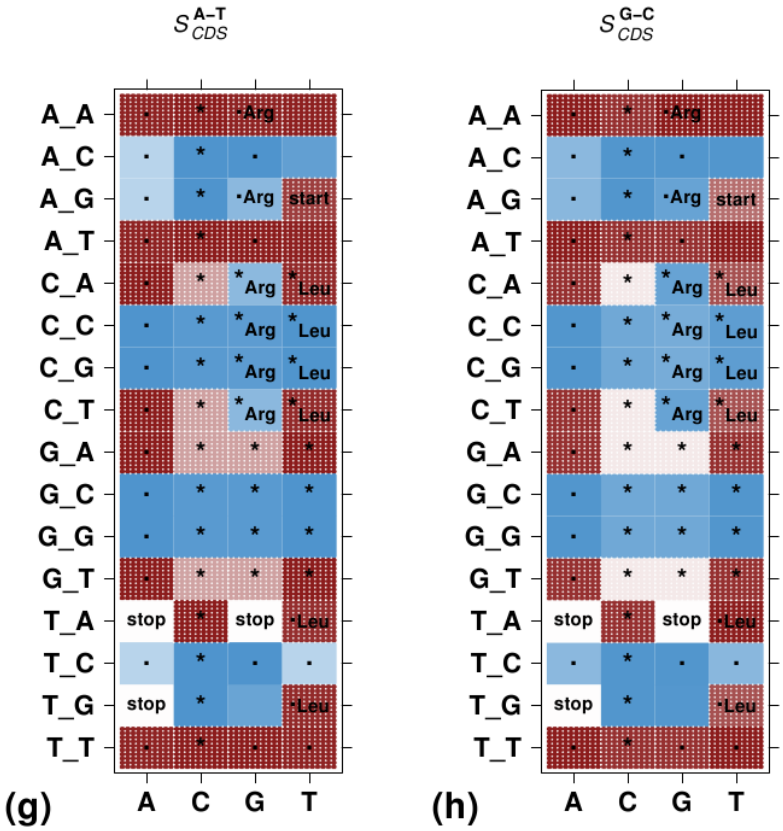

**3ες διπλά εκφυλισμένες θέσεις κωδικονίων**

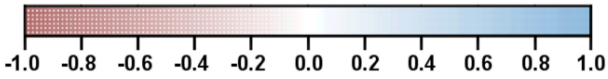


**1ες θέσεις κωδικονίων**




**2ες θέσεις κωδικονίων**





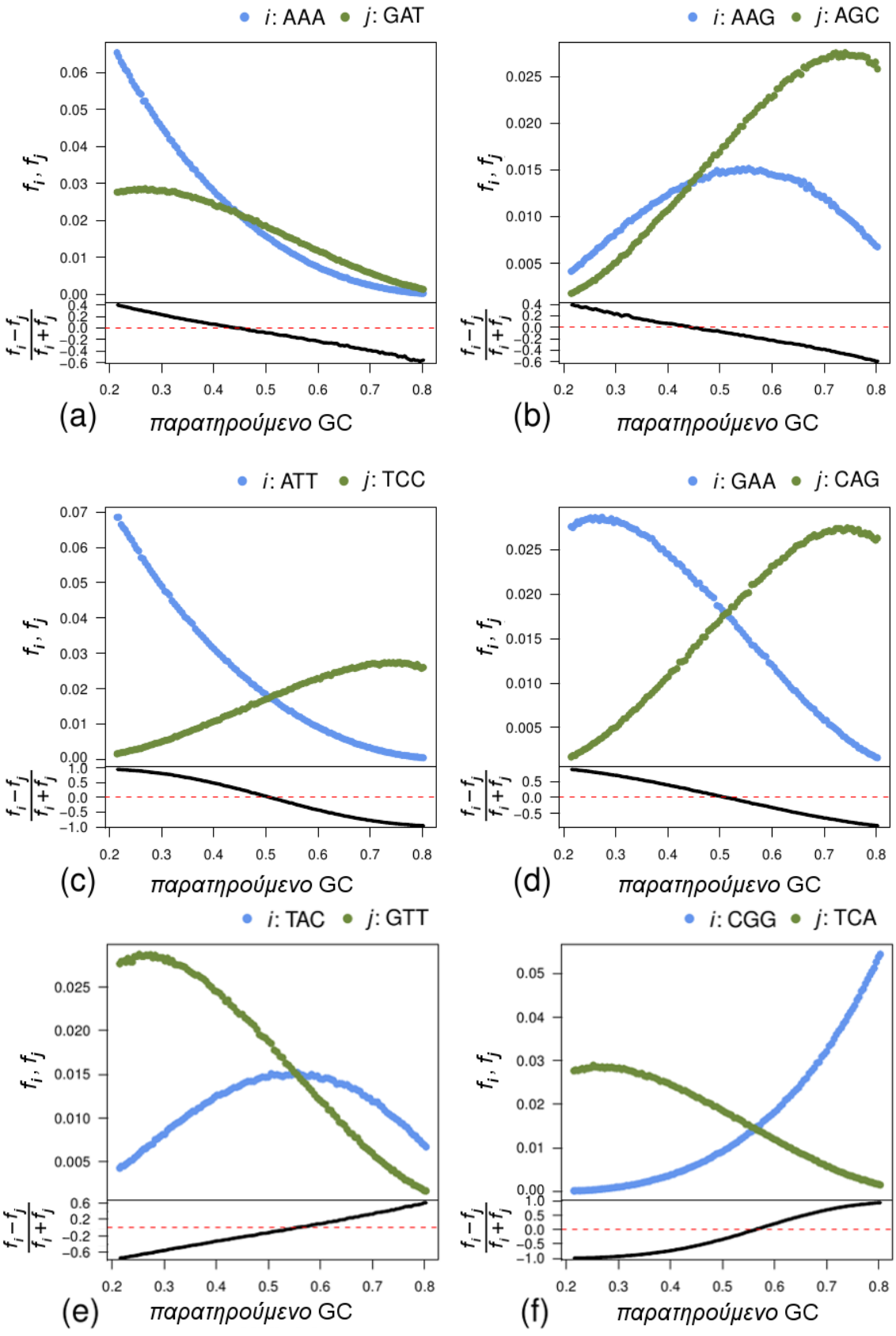

**Εικόνα 16.** Χάρτες θερμότητας (heatmaps) του δειγματικού συντελεστή γραμμικής συσχέτισης του Pearson (*r*) μεταξύ της χρήσης κωδικονίων και των αποκλίσεων (A-T ή G-C), για το σύνολο των τεχνητών αλληλουχιών που κατασκευάστηκαν σύμφωνα με το GC-πολωμένο μοντέλο. Οι αποκλίσεις υπολογίζονται σε συγκεκριμένες θέσεις κωδικονίων ($3^{ες}$ τετραπλά εκφυλισμένες: $3^{ες}|4$, $3^{ες}$ διπλά εκφυλισμένες: $3^{ες}|2$, $1^{ες}$ και $2^{ες}$). Σε κάθε χάρτη θερμότητας οι στήλες δηλώνουν την σύσταση των θέσεων των κωδικονίων στις οποίες μετρήθηκαν οι αποκλίσεις. Τα διάστικτα κόκκινα κελιά αντιστοιχούν σε αρνητικές τιμές του *r*, ενώ τα μπλε κελιά αντιστοιχούν σε θετικές τιμές του *r*. Η ένταση του χρωματισμού των κελιών υποδεικνύει το βαθμό συσχέτισης των αποκλίσεων με τη συχνότητα εμφάνισης του αντίστοιχου κωδικονίου, σύμφωνα με τον χρωματικό κώδικα που παρατίθεται στην λεζάντα.
Τα κωδικόνια έναρξης και λήξης δεν λαμβάνονται υπόψιν στους υπολογισμούς. Το ATG, που λειτουργεί και ως κωδικόνιο έναρξης, και σε ορισμένες περιπτώσεις το TGA, που λειτουργεί κατά κανόνα ως κωδικόνιο λήξης, εντοπίζονται επίσης εντός των κωδικών περιοχών, με το πρώτο να κωδικοποιεί για Μεθειονίνη και το δεύτερο για Τρυπτοφάνη ή Σεληνοκυστεΐνη. Ως αποτέλεσμα εμφανίζονται συσχετίσεις μεταξύ των αποκλίσεων και των [ATG] και [TGA]. Σε κάθε χάρτη θερμότητας επισημαίνονται τα κελιά που αντιστοιχούν στα εξαπλά εκφυλισμένα κωδικόνια της Λευκίνης και της Αργινίνης. Τα κωδικόνια που λαμβάνονται υπόψιν για τους υπολογισμούς των αποκλίσεων στις $3^{ες}|4$ και $3^{ες}|2$ θέσεις επισημαίνονται, αντίστοιχα, με τα σύμβολα "*****" (τετραπλά εκφυλισμένα κωδικόνια) και "·" (διπλά εκφυλισμένα κωδικόνια). "start": κωδικόνιο έναρξης, "stop": κωδικόνια λήξης.

Η μόνη σημαντική ασυμφωνία ανάμεσα στα πρότυπα συσχετίσεων που παρατηρούμε στα γονιδιωματικά δεδομένα και σε εκείνα που παράγονται βάσει του GC-πολωμένου μοντέλου αφορούν τις αποκλίσεις A-T στις $2^{ες}$ θέσεις των κωδικονίων. Ενώ στα χρωμοσώματα τις συλλογής μας οι $S_2^{A-T}$ σχετίζονται θετικά με τις συχνότητες των κωδικονίων που λήγουν σε A/T και αρνητικά με τις συχνότητες όσων λήγουν σε G/C (Εικόνα 14*g*), οι ακριβώς αντίθετες συσχετίσεις παρατηρούνται στις τεχνητές αλληλουχίες (Εικόνα 16*g*). Αυτή η αναντιστοιχία μπορεί να οφείλεται, μεταξύ άλλων, στην εμφάνιση μη-τυπικών αποκλίσεων A-T (Charneski et al. 2011) λόγω επιλεκτικών πιέσεων, τις οποίες το μοντέλο μας δεν ενσωματώνει. Η σύσταση των $2^{ων}$ θέσεων των κωδικονίων συνδέεται στενά με τις φυσικοχημικές ιδιότητες των κωδικοποιούμενων αμινοξέων. Κωδικόνια με κατάλοιπα πουρινών στην $2^{η}$ θέση τους αντιστοιχούν σε φορτισμένα ή πολικά αμινοξέα (Copley et al. 2005, Xiao & Yu 2007, Yu

2007). Στην περίπτωση των ενσωματωμένων μεμβρανικών πρωτεϊνών (integral membrane proteins) και προκειμένου αυτές να διατηρήσουν τον υποκυτταρικό τους εντοπισμό (subcellular location), ασκούνται επιλεκτικές πιέσεις προς αποφυγή τέτοιων κωδικονίων (Lobry & Lobry 1999). Οι πιέσεις αυτές συνοδεύονται από αύξηση της συχνότητας των C έναντι των G στις 2$^{ες}$ θέσεις (Zhang & Yu 2010). Σε ότι αφορά στην ανάλυσή μας, η επιλογή μπορεί να διαμορφώνει τις ιδιαιτερότητες των CDS-συζευγμένων A-T αποκλίσεων στις 2$^{ες}$ θέσεις, επιβάλλοντας συγκεκριμένους περιορισμούς στις GC-πολωμένες υποκαταστάσεις.

**3.11.3.2** *Εμφάνιση αποκλίσεων στις τεχνητές αλληλουχίες – παραδείγματα και επεξηγήσεις*

Προκειμένου να διασαφηνισθεί ο τρόπος με τον οποίο μπορούν να προκύψουν αποκλίσεις A-T και G-C βάσει του μοντέλου μας, *παρόλο που αυτό δεν διακρίνει μεταξύ G και C ή A και T*, παρουσιάζουμε τα διαγράμματα συχνοτήτων τεσσάρων ζευγαριών κωδικονίων συναρτήσει του *παρατηρούμενου* GC για το σύνολο των τεχνητών αλληλουχιών (Εικόνα 17). Κάθε ζεύγος κωδικονίων *i* και *j* επιλέγεται έτσι ώστε ο λόγος $\frac{f_i - f_j}{f_i + f_j}$ να κατευθύνει την σύσταση των τεχνητών αλληλουχιών προς συγκεκριμένες αποκλίσεις. Και στις τέσσερις περιπτώσεις, η φορά της ανισότητας μεταξύ των συχνοτήτων $f_i$ και $f_j$ αλλάζει σε μια περιοχή του *παρατηρούμενου* GC κοντά στο 50%. Εκατέρωθεν αυτής της περιοχής, η διαφορά $f_i - f_j$ παράγει αποκλίσεις A-T στις 3$^{ες}$|2 θέσεις για *i*=AAA και *j*=GAT (Εικόνα 17*a*), στις 1$^{ες}$ θέσεις για *i*=ATT και *j*=TCC (Εικόνα 17*c*), και στις 2$^{ες}$ θέσεις για *i*=TAC και *j*=GTT (Εικόνα 17*e*). Αντίστοιχα, η διαφορά $f_i - f_j$ παράγει αποκλίσεις G-C στις 3$^{ες}$|2 θέσεις για *i*=AAG και *j*=AGC (Εικόνα 17*b*), στις 1$^{ες}$ θέσεις για *i*=GAA και *j*=CAG (Εικόνα 17*d*), και στις 2$^{ες}$ θέσεις για *i*=CGG και *j*=TCA (Εικόνα 17*f*). Στην Εικόνα 17 δεν παρουσιάζουμε ζεύγη τετραπλά εκφυλισμένων κωδικονίων που να δίνουν αποκλίσεις στις τρίτες θέσεις τους, καθώς οι συσχετίσεις των $S_{3|4}^{A-T}$ και $S_{3|4}^{G-C}$ με την χρήση κωδικονίων είναι ιδιαίτερα ασθενείς (Εικόνα 16*a*,*b*).


i: AAA
j: GAT
i: AAG
j: AGC
i: ATT
j: TCC
i: GAA
j: CAG
i: TAC
j: GTT
i: CGG
j: TCA
παρατηρούμενο GC
(a)
(b)
(c)
(d)
(e)
(f)

**Εικόνα 17.** Απόκριση της συχνότητας των κωδικονίων στο συνολικό *παρατηρούμενο* GC περιεχόμενο των τεχνητών αλληλουχιών, σύμφωνα με το προτεινόμενο μοντέλο που ενσωματώνει την ποικιλότητα του GC εντός κάθε ομάδας συνωνύμων (GC-πολωμένο μοντέλο). Σε κάθε διάγραμμα απεικονίζονται οι *παρατηρούμενες* συχνότητες ενός δεδομένου ζεύγους κωδικονίων ($f_i$,$f_j$) καθώς και ο λόγος της διαφοράς αυτών των συχνοτήτων προς το άθροισμά τους $\frac{f_i-f_j}{f_i+f_j}$ . Η οριζόντια διακεκομμένη κόκκινη γραμμή αντιστοιχεί σε $f_i-f_j=0$. Για κάθε ζεύγος κωδικονίων *i* και *j*, ο λόγος $\frac{f_i-f_j}{f_i+f_j}$ οδηγεί προς συγκεκριμένες A-T ή G-C αποκλίσεις. Αυξανομένου του *παρατηρούμενου* GC ο λόγος $\frac{f_i-f_j}{f_i+f_j}$ συνεπάγεται: (*a*) μείωση των $S_{3|2}^{A-T}$ για *i*=AAA και *j*=GAT, (*b*) μείωση των $S_{3|2}^{G-C}$ για *i*=AAG και *j*=AGC, (*c*) μείωση των $S_1^{A-T}$ για *i*=ATT και *j*=TCC, (*d*) μείωση των $S_1^{G-C}$ για *i*=GAA και *j*=CAG, (*e*) αύξηση των $S_2^{A-T}$ για *i*=TAC και *j*=GTT, (*f*) αύξηση των $S_2^{G-C}$ για *i*=CGG και *j*=TCA.

Συμβολίζουμε ως $(\frac{f_i-f_j}{f_i+f_j})_{\alpha^\beta}$ τον λόγο $\frac{f_i-f_j}{f_i+f_j}$ που συνεπάγεται αποκλίσεις τύπου $\beta$ στις θέσεις $\alpha$ των κωδικονίων, όπου $\alpha$ είναι οι 1ες, 2ες ή 3ες|2 θέσεις και $\beta$ οι αποκλίσεις A-T ή G-C. Ο λόγος $(\frac{f_{AAA}-f_{GAT}}{f_{AAA}+f_{GAT}})_{3|2^{A-T}}$ συσχετίζεται θετικά με την $f_{AAA}$ ($r$=0.961) και την $f_{GAT}$ ($r$=0.983) ενώ ο λόγος $(\frac{f_{AAG}-f_{AGC}}{f_{AAG}+f_{AGC}})_{3|2^{G-C}}$ συσχετίζεται αρνητικά με την $f_{AAG}$ ($r$=-0.347) και την $f_{AGC}$ ($r$=-0.979), κατ' αναλογία με τις συσχετίσεις στις 3ες|2 θέσεις των τεχνητών αλληλουχιών, που είναι θετικές για τα κωδικόνια που λήγουν σε A/T και αρνητικές για τα κωδικόνια που λήγουν σε G/C (Εικόνα 16*c*,*d*). Αντίστοιχα, ο λόγος $(\frac{f_{ATT}-f_{TCC}}{f_{ATT}+f_{TCC}})_{1^{A-T}}$ συσχετίζεται θετικά με την $f_{ATT}$ ($r$=0.960) και αρνητικά με

την $f_{TCC}$ ($r$=-0.992) ενώ ο λόγος $(\frac{f_{GAA}-f_{CAG}}{f_{GAA}+f_{CAG}})_{1^{G-C}}$ συσχετίζεται θετικά με την $f_{GAA}$ ($r$=0.993) και αρνητικά με την $f_{CAG}$ ($r$=-0.989), σε συμφωνία με τα πρότυπα των συσχετίσεων στις 1ες θέσεις των τεχνητών αλληλουχιών που διακρίνουν μεταξύ κωδικονίων που λήγουν σε A/T και G/C (Εικόνα 16*e*,*f*).

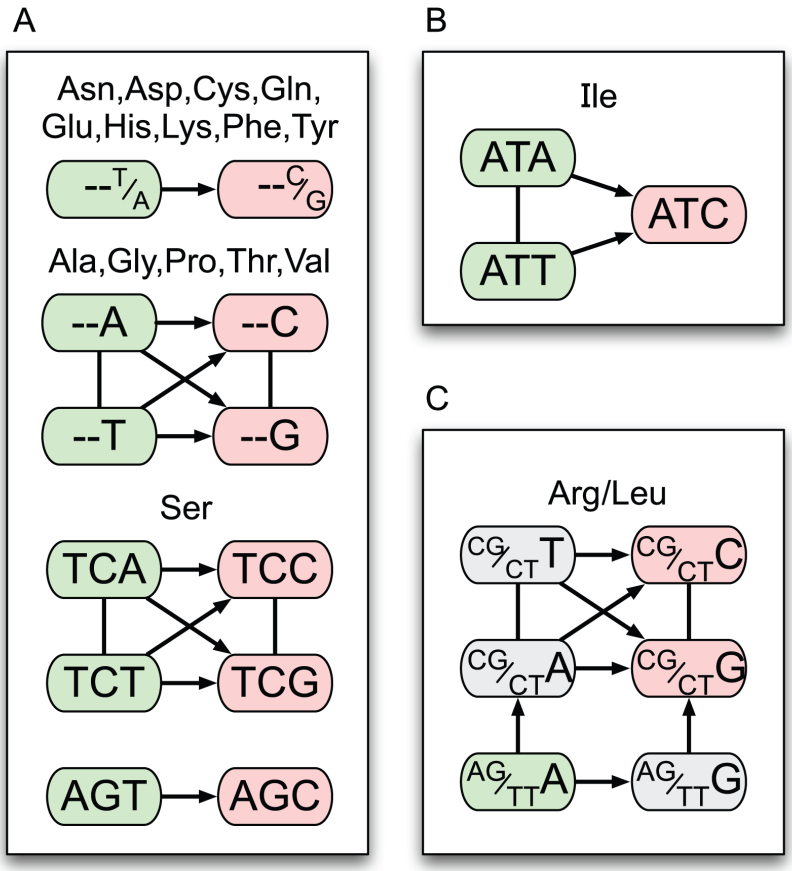
Τέλος, ο λόγος $(\frac{f_{TAC}-f_{GTT}}{f_{TAC}+f_{GTT}})_{2^{A-T}}$ συσχετίζεται θετικά με την $f_{TAC}$ ($r$=0.328) και αρνητικά με την $f_{GTT}$ ($r$=-0.989) ενώ ο λόγος $(\frac{f_{CGG}-f_{TCA}}{f_{CGG}+f_{TCA}})_{2^{G-C}}$ συσχετίζεται θετικά με την $f_{CGG}$ ($r$=0.963) και αρνητικά με την $f_{TCA}$ ($r$=-0.999), κατ' αναλογία με τα πρότυπα συσχετίσεων στις 2ες θέσεις των τεχνητών αλληλουχιών, όπου, σε αντίθεση με ό,τι συμβαίνει στις 3|2ες και 1ες θέσεις, οι αποκλίσεις μετατοπίζονται από μικρότερες σε μεγαλύτερες προσημασμένες τιμές καθώς μειώνεται η συχνότητα των κωδικονίων που λήγουν σε A/T και καθώς αυξάνεται η συχνότητα των κωδικονίων που λήγουν σε G/C (Εικόνα 16*g*,*h*).

Ας σημειωθεί ότι τα παραδείγματα της Εικόνας 17 είναι απλώς ενδεικτικά του μηχανισμού με τον οποίο παράγονται αποκλίσεις σύμφωνα με το GC-πολωμένο μοντέλο. Στην πράξη, ένα κωδικόνιο μπορεί να συνεισφέρει στις αποκλίσεις περισσοτέρων της μίας θέσεων κατά μήκος των τεχνητών αλληλουχιών. Επίσης, οι χάρτες θερμότητας της Εικόνας 16 αφορούν αποκλίσεις που έχουν υπολογιστεί στο σύνολο των αντίστοιχων θέσεων (3|4ων, 3|2ων, 1ων ή 2ων). Συνεπώς λαμβάνουν υπόψιν όλα τα κωδικόνια και όχι μόνο ένα συγκεκριμένο ζεύγος *i*,*j* όπως στην Εικόνα 17.

**3.11.3.3** *Οι GC-πολώσεις των συνώνυμων κωδικονίων διαμορφώνουν τις CDS-συζευγμένες αποκλίσεις*

Συνοψίζοντας τα παραπάνω, προκύπτει ότι *η οργάνωση των κωδικονίων σε ομάδες συνωνύμων είναι ικανή, για δεδομένο ολικό GC περιεχόμενο, να παράγει ειδικές ανά θέση ασυμμετρίες στην σύσταση των κωδικών περιοχών* (Εικόνα 17), *οι οποίες ακολουθούν πρότυπα συσχέτισης με την χρήση κωδικονίων όμοια με αυτά που παρατηρούνται στα χρωμοσώματα των βακτηρίων* (πρβ. Εικόνα 14 και

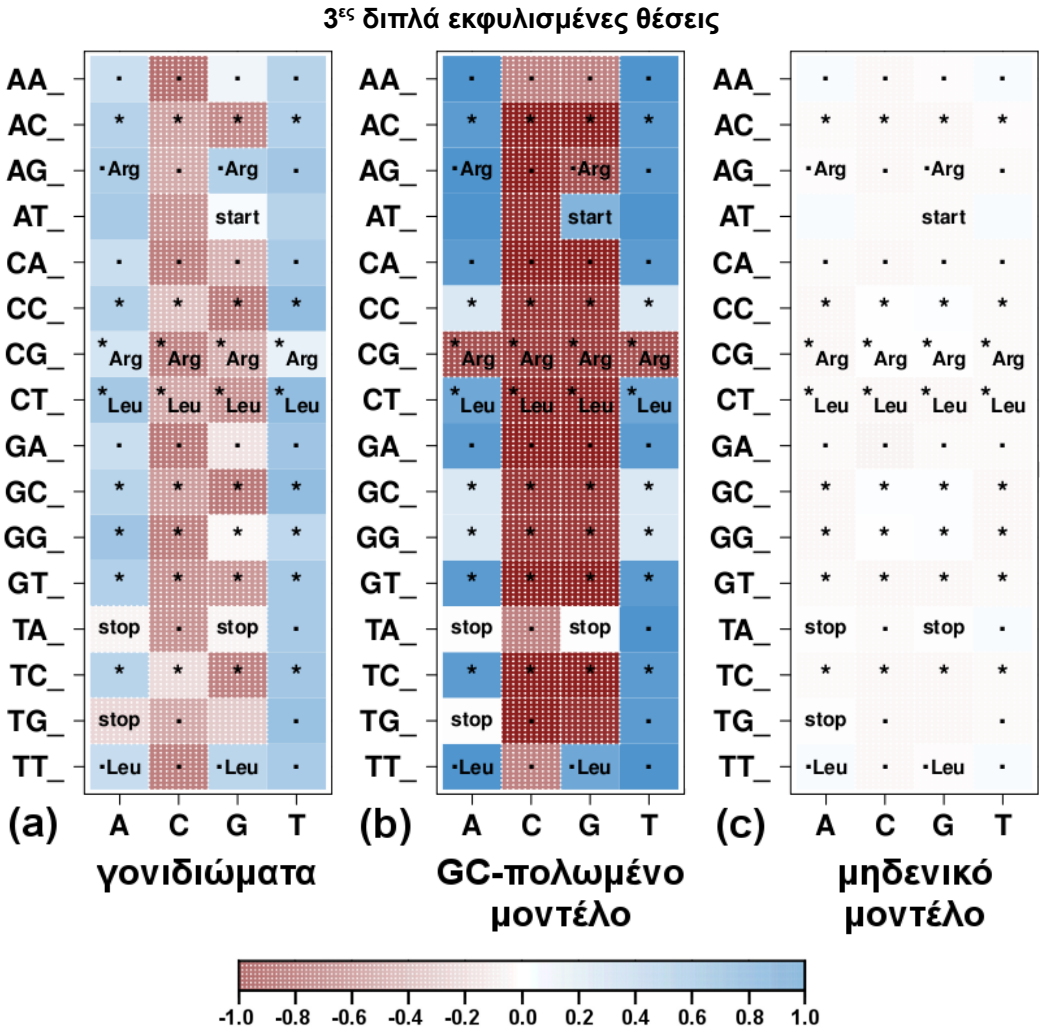
16). Όπως ήδη αναφέραμε, η γενική μορφή αυτών των προτύπων χωρίζει τα κωδικόνια σε δύο ομάδες, ανάλογα με το εάν λήγουν σε A/T ή G/C. Ωστόσο, υπάρχουν κάποιες χαρακτηριστικές εξαιρέσεις. Όπως ήδη αναφέραμε (βλ. ενότητα 3.11.2.4), ορισμένα κωδικόνια της Λευκίνης και της Αργινίνης εμφανίζουν πρότυπα συσχέτισης με στις A-T και G-C αποκλίσεις, τα οποία είναι διαφορετικά ή και αντίθετα από τα πρότυπα που ακολουθούν τα υπόλοιπα κωδικόνια με το ίδιο GC περιεχόμενο στην 3η θέση (Εικόνα 14*c-h*). Οι (Palidwor et al. 2010) έδειξαν ότι τα κωδικόνια των δύο αυτών αμινοξέων αποκρίνονται στις GC κατευθύνουσες μεταλλακτικές πιέσεις κατά τρόπο που τα διαφοροποιεί από όσα άλλα λήγουν σε A/T ή G/C. Η παρατήρηση αυτή αποδόθηκε στο γεγονός ότι *η Λευκίνη και η Αργινίνη είναι τα μόνα αμινοξέα των οποίων τα κωδικόνια επιτρέπουν συνώνυμες υποκαταστάσεις που μεταβάλλουν το GC περιεχόμενο τόσο στην 3η όσο και στην 1η θέση τους* (Εικόνα 18). Τα δίκτυα των συνώνυμων υποκαταστάσεων εντός των ομάδων που κωδικοποιούν για Λευκίνη ή Αργινίνη μπορεί επίσης να ευθύνονται για τις παρατηρούμενες διαφορές στα πρότυπα συσχέτισης μεταξύ των συχνοτήτων των αντίστοιχων κωδικονίων και των A-T ή G-C αποκλίσεων.



**Εικόνα 18,** ***τροποποιημένη από*** **(Palidwor et al. 2010).** Δίκτυα συνώνυμων σημειακών υποκαταστάσεων, για το σύνολο των κωδικοποιούμενων αμινοξέων. Τα βέλη δηλώνουν συνώνυμες υποκαταστάσεις που αυξάνουν το GC περιεχόμενο των κωδικών περιοχών. Οι απλές γραμμές αντιστοιχούν σε συνώνυμες υποκαταστάσεις που δεν μεταβάλλουν το GC περιεχόμενο των κωδικών περιοχών. Για όλα τα αμινοξέα πλην Λευκίνης και Αργινίνης, τα αντίστοιχα κωδικόνια χρωματίζονται κόκκινα εάν λήγουν σε G/C και πράσινα εάν λήγουν σε A/T. Για την Λευκίνη και την Αργινίνη, τα κωδικόνια που έχουν G/C και στις δύο συνώνυμες θέσεις τους ($1^{η}$ και $3^{η}$) χρωματίζονται κόκκινα, εκείνα που έχουν G/C σε μία από τις δύο συνώνυμες θέσεις ($1^{η}$ ή $3^{η}$) χρωματίζονται γκρι, ενώ εκείνα που έχουν A/T και στις δύο συνώνυμες θέσεις ($1^{η}$ και $3^{η}$) χρωματίζονται πράσινα. **(A)** Διπλά και τετραπλά εκφυλισμένα κωδικόνια. **(B)** Τριπλά εκφυλισμένα κωδικόνια. **(C)** Εξαπλά εκφυλισμένα κωδικόνια.

Το GC-πολωμένο μοντέλο οδηγεί επίσης σε μη-τυπικές συσχετίσεις, όσον αφορά τα κωδικόνια της Leu και της Arg. Τόσο στα χρωμοσώματα όσο και στις τεχνητές αλληλουχίες τα κωδικόνια CGA/CGT (Arg) συσχετίζονται αρνητικά με τις $S_1^{A-T}$ και $S_1^{G-C}$ (πρβ. Εικόνα 14*e*,*f* και 16*e*,*f*), ενώ το TTG (Leu) έχει θετική συσχέτιση με τις $S_{3|2}^{A-T}$ και $S_{3|2}^{G-C}$ (πρβ. Εικόνα 14*c*,*d* και 16*c*,*d*). Αυτή η παρατήρηση ενισχύει περαιτέρω το επιχείρημά μας, σύμφωνα με το οποίο *οι GC-πολώσεις εντός της κάθε ομάδας συνωνύμων αποτελούν έναν αιτιώδη σύνδεσμο ανάμεσα στην χρήση κωδικονίων και στις CDS-συζευγμένες αποκλίσεις*. Και τούτο διότι το μοντέλο μας, εγγράφοντας στις παραμέτρους $R_i^{A-T}$ και $R_i^{G-C}$ την δομή του γενετικού κώδικα βάσει της οποίας οργανώνονται τα δίκτυα των συνώνυμων υποκαταστάσεων, *επιτυγχάνει να ανασυγκροτήσει όχι μόνο το γενικό πρότυπο συσχετίσεων που διακρίνει μεταξύ κωδικονίων με A/T ή G/C στην* $3^{η}$ *θέση, αλλά αναπαράγει και χαρακτηριστικές εξαιρέσεις από τον γενικό κανόνα*. Σημειώνουμε επίσης πως ορισμένα κωδικόνια της Leu και Arg τα οποία εκδηλώνουν ασθενείς συσχετίσεις με τις αποκλίσεις των χρωμοσωμάτων, εμφανίζουν στις τεχνητές αλληλουχίες συσχετίσεις αντίστροφες εκείνων που κατά κανόνα ακολουθούν οι ομάδες των κωδικονίων που λήγουν σε A/T ή G/C (πρβ. Εικόνα 14*c*,*d* και 16*c*,*d* για το CGT· Εικόνα 14*f* και 16*f* για το TTG· Εικόνα 14*h* και 16*h* για τα CGA/T). Παρότι λοιπόν σε αυτές τις περιπτώσεις τα αποτελέσματα στις τεχνητές αλληλουχίες δεν συμπίπτουν με τα όσα παρατηρούμε στα βακτηριακά χρωμοσώματα, υποδηλώνουν ωστόσο ότι η εξασθένιση των συγκεκριμένων συσχετίσεων στα γονιδιωματικά δεδομένα συνδέεται με την GC ποικιλότητα μεταξύ των εξαπλά εκφυλισμένων κωδικονίων.

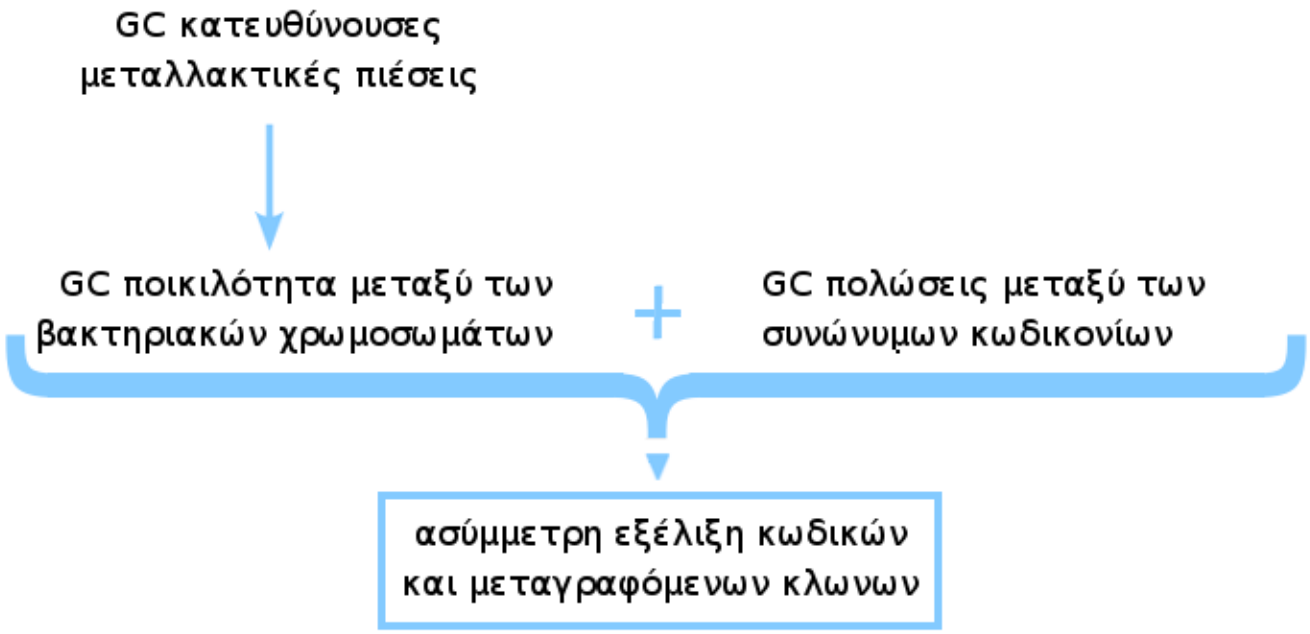
Προκειμένου να ελέγξουμε εάν οι GC-πολώσεις των συνώνυμων κωδικονίων αποτελούν πράγματι τον καθοριστικό παράγοντα που διαμορφώνει τα πρότυπα συσχετίσεων στις τεχνητές αλληλουχίες, θεωρούμε ένα μηδενικό μοντέλο (null-hypothesis model) στο οποίο παραλείπονται οι παράμετροι $R_i^{A-T}$ και $R_i^{G-C}$ και η πιθανότητα εμφάνισης των κωδικονίων δίδεται από τον τύπο $p_i=(GC)^{n_i}(1-GC)^{(3-n_i)}$. Το μηδενικό μοντέλο καταδεικνύει ότι *εάν δεν λάβουμε υπόψιν τις παραμέτρους $R_i^{A-T}$ και $R_i^{G-C}$, τα πρότυπα των συσχετίσεων που παράγει το GC-πολωμένο μοντέλο καταρρέουν* (Εικόνα 19c). Συνεπώς, οι GC-πολώσεις των συνώνυμων κωδικονίων συμπλέκονται στενά με τις ασυμμετρίες των κωδικών περιοχών.




**Εικόνα 19.** Χάρτες θερμότητας (heatmaps) του δειγματικού συντελεστή γραμμικής συσχέτισης του Pearson ($r$) μεταξύ της χρήσης κωδικονίων και των αποκλίσεων (A-T ή G-C), για το σύνολο **(a)** των χρωμοσωμάτων της συλλογής μας, **(b)** των τεχνητών αλληλουχιών που κατασκευάστηκαν σύμφωνα με το GC-πολωμένο μοντέλο, **(c)** των τεχνητών αλληλουχιών που κατασκευάστηκαν σύμφωνα με το μηδενικό μοντέλο. Οι αποκλίσεις

υπολογίζονται στις 3ες διπλά εκφυλισμένες θέσεις των κωδικονίων (3ες|2). Τα πρότυπα συσχέτισης που ανιχνεύονται στα γονιδιωματικά δεδομένα αναπαράγονται από το GC-πολωμένο μοντέλο, ενώ καταρρέουν όταν δεν λαμβάνονται υπόψιν οι GC-πολώσεις των συνώνυμων κωδικονίων, όπως συμβαίνει στο μηδενικό μοντέλο. Για περαιτέρω διευκρινίσεις σχετικά με την γραφική απεικόνιση, βλ. το επεξηγηματικό κείμενο στην Εικόνα 14 και 16.

Συμπερασματικά, *η ποικιλότητα του GC περιεχομένου μεταξύ των βακτηριακών χρωμοσωμάτων που αποδίδεται σε GC κατευθύνουσες μεταλλακτικές πιέσεις, παρότι καθεαυτή είναι συμμετρική ως προς τους κλώνους του DNA, σε συνδυασμό με τις GC-πολώσεις εντός των ομάδων συνωνύμων μπορεί να κατευθύνει την εμφάνιση ασυμμετριών μεταξύ κωδικών και μεταγραφόμενων κλώνων* (βλ. Εικόνα 20). Αυτό ισχύει ακόμα και όταν δεν λαμβάνεται υπόψιν η δράση της φυσικής επιλογής στον καθορισμό της χρήσης κωδικονίων ή αμινοξέων. Οι αποκλίσεις που προκύπτουν κατ' αυτόν τον τρόπο είναι εντονότερες στις 1ες, 2ες και 3|2ες θέσεις των κωδικονίων, όπου οι ασυμμετρίες της σύστασης επιβάλλονται εγγενώς από την ίδια την δομή του γενετικού κώδικα. Τα ευρήματά μας υποδηλώνουν ότι, *δεδομένου του συνολικού GC περιεχομένου των κωδικών περιοχών, ο διαμερισμός των κωδικονίων σε ομάδες συνωνύμων προσφέρει την βάση για μία a priori εκτίμηση των CDS-συζευγμένων αποκλίσεων ανά θέση κωδικονίων*. Οι αποκλίσεις αυτές ανταποκρίνονται σε μία ασύμμετρη κατανομή αναφοράς (baseline) των νουκλεοτιδικών βάσεων ανάμεσα στις θέσεις των κωδικονίων, η οποία μπορεί να χρησιμοποιηθεί σε μελέτες συγκριτικής γονιδιωματικής ανάλυσης ως σημείο εκκίνησης κατά την εκτίμηση του αναμενόμενου αριθμού των ανά θέση υποκαταστάσεων.




**Εικόνα 20.** Μηχανισμός σύζευξης των GC κατευθυνουσών μεταλλακτικών πιέσεων με τις ασυμμετρίες της σύστασης των κωδικών περιοχών. Οι GC κατευθύνουσες μεταλλακτικές πιέσεις διαμορφώνουν καθοριστικά την ποικιλότητα που εμφανίζουν τα βακτηριακά γονιδιώματα ως προς το GC περιεχόμενό τους. Το GC% του DNA είναι καθαυτό συμμετρικό ως προς τους δύο κλώνους. Ωστόσο, δεδομένων των GC πολώσεων που εμφανίζονται εντός κάθε ομάδας συνωνύμων, το GC περιεχόμενο των κωδικών περιοχών διαμορφώνει τις ασυμμετρίες μεταξύ κωδικών και μεταγραφόμενων κλώνων. Έτσι, η επαγόμενη από μεταλλακτικές πιέσεις δια-γονιδιωματική ποικιλότητα του GC περιεχομένου μπορεί να καθοδηγεί την ασύμμετρη εξέλιξη των κωδικών περιοχών.

Όπως προαναφέραμε, οι αποκλίσεις στις 3|4$^{ες}$ θέσεις εμφανίζουν μία σαφώς ασθενέστερη συσχέτιση με τα πρότυπα χρήσης κωδικονίων, σε σύγκριση με τις αποκλίσεις στις 1$^{ες}$, 2$^{ες}$ και 3|2$^{ες}$ θέσεις. Στην επόμενη ενότητα διερευνούμε τον ρόλο που διαδραματίζουν συγκεκριμένοι μοριακοί μηχανισμοί στην διαμόρφωση των αποκλίσεων στις 3|4$^{ες}$ θέσεις. Εστιάζουμε σε μηχανισμούς που από την φύση τους μπορούν να επάγουν ειδικές ανά κλώνο πολώσεις, συζευγμένες με την αντιγραφή, τη μεταγραφή και την επιδιόρθωση του DNA.

## **3.12** *Μεταλλακτικές πολώσεις στο επίπεδο ολόκληρου του γονιδιώματος ως αποτέλεσμα συγκεκριμένων μοριακών μηχανισμών*

*Οι ασυμμετρίες στην σύσταση του γενετικού υλικού προσφέρουν χρήσιμες πληροφορίες σχετικά με μοριακούς μηχανισμούς που δρουν με διακριτό τρόπο σε κάθε έναν από τους δύο κλώνους του DNA. Οι αποκλίσεις της νουκλεοτιδικής σύστασης μας επιτρέπουν να συνάγουμε ποιοι από αυτούς τους μηχανισμούς συνδέονται με συγκεκριμένους τύπους ασυμμετριών στους ρυθμούς υποκατάστασης*. Στην παρούσα ενότητα μελετάμε την κατανομή των αποκλίσεων, για το σύνολο των χρωμοσωμάτων της συλλογής μας, σε σχέση με την παρουσία ή μη (*α*) του συστήματος της συζευγμένης με τη μεταγραφή επιδιόρθωσης (transcription coupled-repair, TCR), (*β*) μονοπατιών επιδιόρθωσης του DNA, μη συζευγμένων με τη μεταγραφή και (*γ*) διαφορετικών ισομορφών της α-υπομονάδας της πολυμεράσης PolIII. Όλες οι αντίστοιχες συγκρίσεις των κατανομών των αποκλίσεων πραγματοποιούνται με την χρήση του *στατιστικού ελέγχου αθροίσματος διατάξεων του Wilcoxon* (τεστ Wilcoxon, two-tailed Wilcoxon rank-sum test).

### **3.12.1** *Η συζευγμένη με τη μεταγραφή επιδιόρθωση του DNA*

Η σύζευξη της μεταγραφής με την επιδιόρθωση του DNA είχε διαφανεί ήδη από τις αρχές του '70, βάσει σχετικών γενετικών πειραμάτων (Bockrath & Cheung 1973, Bockrath & Palmer 1977, Engstrom et al. 1984, Bockrath et al. 1987). Οι Bockrath και Palmer διαπίστωσαν ότι τα βακτηριακά κύτταρα που έχουν εκτεθεί σε υπεριώδη ακτινοβολία, υπό κατάλληλες συνθήκες εμφανίζουν σημαντική μείωση της συχνότητας των παρατηρούμενων μεταλλάξεων (“mutation frequency decline” ή MFD). Σε αυτή την μείωση των μεταλλακτικών ρυθμών εμπλέκεται ένα σύστημα επιδιόρθωσης που εκτέμνει εκείνες μόνο τις προ-μεταλλακτικές βλάβες (premutational lesions) που εντοπίζονται στον μεταγραφόμενο κλώνο (Bockrath & Palmer 1977) Οι Bohr *et al.* (1985) κατέδειξαν ότι υπάρχει μια ισχυρή προτίμηση επιδιόρθωσης του DNA στις μεταγραφικά ενεργές περιοχές, σε σύγκριση με το σύνολο του

γονιδιώματος. Το προϊόν του γονιδίου *mfd* ταυτοποιήθηκε ως ο παράγοντας σύζευξης μεταγραφής-επιδιόρθωσης (transcription-repair coupling factor, TRCF) (Selby et al. 1991, C. Selby & Sancar 1993, Selby & Sancar 1994). Για μια αναλυτικότερη παρουσίαση των σχετικών ευρημάτων, βλ. ενότητα 1.6.1.

Η συζευγμένη με τη μεταγραφή επιδιόρθωση του DNA (transcription-coupled repair, TCR) αποτελεί μία χαρακτηριστική περίπτωση μοριακού μηχανισμού με δράση ασύμμετρη ως προς τους κλώνους του DNA. Προκειμένου να αξιολογήσουμε την επίδραση της TCR στις ειδικές ανά κλώνο ασυμμετρίες των υποκαταστάσεων, συγκρίνουμε την κατανομή των συζευγμένων με τη μεταγραφή αποκλίσεων (Trs, βλ. ενότητα 2.9), συναρτήσει της παρουσίας ή μη του *mfd* γονιδίου ($TCR^+$: παρουσία *mfd*, $TCR^-$: απουσία *mfd*). Στο βαθμό που υπάρχει στατιστικά σημαντική διαφορά της κατανομής των αποκλίσεων μεταξύ των στελεχών που φέρουν το *mfd* και εκείνων στα οποία το *mfd* απουσιάζει (βλ. Πίνακας 15), καταδεικνύεται ότι οι ασυμμετρίες της σύστασης του DNA συνδέονται με την TCR επιδιόρθωση.

### **3.12.1.1** *Συσχέτιση της TCR με τις μονονουκλεοτιδικές αποκλίσεις*

Η διάμεσος των $S_{Trs}^{A-T}$ στα βακτήρια με ικανότητα TCR επιδιόρθωσης ισούται με -0.156, ενώ σε εκείνα δίχως ικανότητα TCR επιδιόρθωσης είναι -0.0642. Η κατανομή των $S_{Trs}^{A-T}$ στα $TCR^+$ βακτήρια ($TCR^+$-$S_{Trs}^{A-T}$) είναι μετατοπισμένη προς χαμηλότερες τιμές σε σχέση με την αντίστοιχη κατανομή στα $TCR^-$ βακτήρια ($TCR^-$-$S_{Trs}^{A-T}$). Η μετατόπιση αυτή είναι στατιστικά πολύ σημαντική ($p$-τιμή < $10^{-5}$). Συνεπώς, η TCR επιδιόρθωση συσχετίζεται με μια έντονη τάση υπερεκπροσώπησης καταλοίπων T έναντι A. Η τάση αυτή ([T] > [A]), παρότι ανιχνεύεται και στα $TCR^-$ βακτήρια, είναι σαφώς πιο μετριασμένη. Αντίθετα με ότι παρατηρείται στις $S_{Trs}^{A-T}$, η TCR επιδιόρθωση δεν σχετίζεται με τις $S_{Trs}^{G-C}$ ($p$-τιμή=0.239).

**ΠΙΝΑΚΑΣ 15. Ανάλυση της κατανομής των αποκλίσεων συναρτήσει της συζευγμένης με τη μεταγραφή επιδιόρθωσης (TCR)**

| διάμεσος | $S_{Trs}^{A-T}$ | $S_{Trs}^{G-C}$ | $P_{Trs}^{AG-CT}$ | $P_{Trs}^{GA-TC}$ | $P_{Trs}^{GG-CC}$ | $P_{Trs}^{AA-TT}$ | $P_{Trs}^{AC-GT}$ | $P_{Trs}^{CA-TG}$ |
|---|---|---|---|---|---|---|---|---|
| $TCR^+$ | -0.156 | -0.0412 | -0.191 | 0.0358 | 0.173 | -0.186 | 0.313 | 0.0912 |
| $TCR^-$ | -0.0642 | -0.00305 | -0.178 | -0.0193 | -0.205 | 0.0333 | -0.0806 | 0.0604 |
| *p*-τιμή | 9.98e-06 *** | 0.239 - | 0.372 - | 0.027 * | 1.55e-11 *** | 1.07e-08 *** | 1.86e-09 *** | 0.36 - |

ΣΗΜΕΙΩΣΕΙΣ.- Υπολογίζουμε τις συζευγμένες με τη μεταγραφή αποκλίσεις, για το σύνολο των χρωμοσωμάτων της συλλογής μας. Χωρίζουμε την συλλογή μας σε δύο ομάδες, που αντιστοιχούν σε βακτήρια με ή δίχως ικανότητα TCR επιδιόρθωσης ($TCR^+$ ή $TCR^-$, αντίστοιχα). Για κάθε ομάδα δίδεται η διάμεσος των αποκλίσεων. Για κάθε απόκλιση συγκρίνουμε τις κατανομές των αντίστοιχων τιμών στις δύο ομάδες, $TCR^+$ και $TCR^-$, χρησιμοποιώντας το *στατιστικό έλεγχο αθροίσματος διατάξεων του Wilcoxon* (two-tailed Wilcoxon rank-sum test). Οι *p*-τιμές του τεστ Wilcoxon δηλώνουν την στατιστική σημαντικότητα της διαφοράς των υπό σύγκριση κατανομών. "-": *p*-τιμή ≥ 0.05, "*": 0.05 > *p*-τιμή ≥ 0.01, "**": 0.01 > *p*-τιμή ≥ 0.001, "***": *p*-τιμή < 0.001

Προκειμένου να διευκρινίσουμε με ποιον τρόπο η TCR επιδιόρθωση μπορεί να επιδρά στις ειδικές ανά κλώνο ασυμμετρίες του DNA, πρέπει να διακρίνουμε δύο καθοριστικά βήματα στην διαδικασία υποκατάστασης βάσεων· από την μία *τις τροποποιήσεις των βάσεων και τους αντίστοιχους ρυθμούς μεταλλάξεων*, και από την άλλη *τους ρυθμούς επιδιόρθωσης των μεταλλαγμένων βάσεων*. Τόσο οι ρυθμοί μεταλλάξεων όσο και οι ρυθμοί επιδιόρθωσης μπορεί να είναι διαφορετικοί ανάμεσα στους δύο κλώνους του DNA. Ο κωδικός κλώνος είναι πιο επιρρεπής σε μεταλλάξεις από ότι ο μεταγραφόμενος (Beletskii & Bhagwat 1996, Green et al. 2003), καθώς είναι εκτεθειμένος για μεγαλύτερο χρονικό διάστημα στην μονόκλωνη κατάσταση κατά τη μεταγραφή (Francino & Ochman 1997, Francino & Ochman 2001, Sekine et al. 2012, Deaconescu 2013). Αυτό έχει ως αποτέλεσμα η σχετιζόμενη με τη μεταγραφή μεταλλαξιγένεση (Transcription-Associated Mutagenesis, TAM) να αυξάνει τον ρυθμό των μεταλλάξεων C→T κατά 100 και πλέον φορές στις κωδικές περιοχές (Lindahl &

Nyberg 1974, Frederico et al. 1990, Beletskii et al. 2000). Συγχρόνως, η παρουσία του μηχανισμού TCR συνεπάγεται την κατά προτίμηση επιδιόρθωση του μεταγραφόμενου κλώνου έναντι του κωδικού (Ganesan et al. 2012). Ο παράγοντας Mfd κατευθύνει το σύστημα επιδιόρθωσης εκτομής νουκλεοτιδίων *uvr*ABC στον μεταγραφόμενο κλώνο, με δύο αλληλένδετες συνέπειες. Αρχικά, οι προ-μεταλλακτικές βλάβες απομακρύνονται από τον μεταγραφόμενο κλώνο, με την εκτομή της αλληλουχίας DNA που βρίσκεται εκατέρωθεν των βλαβών. Ακολούθως, η πολυμεράση PolI συμπληρώνει το κενό της εκτομής, χρησιμοποιώντας ως εκμαγείο το απέναντι τμήμα του κωδικού κλώνου, με αποτέλεσμα την μονιμοποίηση (fixation) των μεταλλάξεων που βρίσκονται στον κλώνο αυτό (Selby et al. 1991). Ως εκ τούτου, *οι ασυμμετρίες που προκαλούνται από την σχετιζόμενη με τη μεταγραφή μεταλλαξιγένεση* (TAM) *αυξάνονται ακόμα περισσότερο στα βακτήρια με ικανότητα TCR επιδιόρθωσης* ($TCR^+$). Αντιθέτως, στα βακτήρια δίχως ικανότητα TCR επιδιόρθωσης ($TCR^-$), όταν η παρουσία κακώσεων στον μεταγραφόμενο κλώνο σταματά προσωρινά τη μεταγραφή, το τριμερές σύμπλοκο επιμήκυνσης (ternary elongation complex, TEC) καλύπτει την βλάβη και παρεμποδίζει την επιδιόρθωσή της. Ως αποτέλεσμα, *στα $TCR^-$ βακτήρια η επιδιόρθωση του DNA διεξάγεται ταχύτερα στον κωδικό από ότι στον μεταγραφόμενο κλώνο*. Συνεπώς, απουσία του *mfd* γονιδίου αναμένουμε ότι οι αποκλίσεις που οφείλονται στην σχετιζόμενη με τη μεταγραφή μεταλλαξιγένεση (TAM) θα μετριάζονται (Oller et al. 1992).

Η παραπάνω συλλογιστική συνηγορεί υπέρ των αποτελεσμάτων μας. Η μετατόπιση της κατανομής $TCR^+$-$S_{Trs}^{A-T}$ προς χαμηλότερες τιμές σε σχέση με την $TCR^+$-$S_{Trs}^{A-T}$ βρίσκεται σε συμφωνία με μια ισχυρότερη πόλωση των μεταβάσεων C→T προς τον κωδικό κλώνο στα $TCR^+$ βακτήρια. Ωστόσο, ο παραπάνω μηχανισμός αναμένεται να οδηγήσει και σε συσχετίσεις της TCR επιδιόρθωσης με τις αποκλίσεις $S_{Trs}^{G-C}$. Η απουσία μιας τέτοιας συσχέτισης υποδεικνύει ότι οι αποκλίσεις $S_{Trs}^{G-C}$ προέρχονται κατά μεγάλο μέρος από άλλες αιτίες, πέραν των C→T ασυμμετριών, σύμφωνα και με προγενέστερες μελέτες (Rocha & Danchin 2001).

### **3.12.1.2** *Συσχέτιση της TCR με τις αποκλίσεις των σταθμισμένων δινουκλεοτιδικών συχνοτήτων*

Ακολούθως εξετάσαμε εάν και κατά πόσον η TCR επιδιόρθωση επάγει ασυμμετρίες στα πρότυπα υποκαταστάσεων οι οποίες εξαρτώνται από τις γειτονικές βάσεις.

Η παρουσία του *mfd* εμφανίζει πολύ ισχυρή συσχέτιση (*p*-τιμή < $10^{-7}$) με τρεις από τους έξι διαφορετικούς τύπους αποκλίσεων των σταθμισμένων συχνοτήτων ($P_{Trs}^{GG-CC}$, $P_{Trs}^{AA-TT}$ και $P_{Trs}^{AC-GT}$· Πίνακας 15). Επίσης, ανιχνεύουμε στατιστική διάκριση (*p*-τιμή < 0.05) ανάμεσα στις κατανομές των $P_{Trs}^{GA-TC}$ στα $TCR^+$ και στα $TCR^-$ βακτήρια. Εκ πρώτης όψεως, από αυτές τις παρατηρήσεις δεν μπορούμε να συνάγουμε το συμπέρασμα ότι η ίδια η ενεργότητα της TCR μεταβάλλεται ανάλογα με την ταυτότητα των βάσεων που γειτονεύουν με την προς επιδιόρθωση βλάβη. Και τούτο διότι οι ρυθμοί της σχετιζόμενης με τη μεταγραφή μεταλλαξιγένεσης (TAM) είναι γνωστό πως εξαρτώνται από τις γειτονικές βάσεις, με τους ρυθμούς των υποκαταστάσεων C→T να αυξάνονται ραγδαία όταν διαδοχικά κατάλοιπα C σχηματίζουν διμερή πυριμιδίνης (Peng & Shaw 1996). Συνεπώς, η επιτάχυνση των CC-->TT μεταβάσεων οδηγεί από μόνη της σε πρότυπα υποκατάστασης που είναι *πλαίσιο-εξαρτημένα* (context-dependent). Η δράση της TCR, ενισχύοντας τα αποτελέσματα της TAM στον κωδικό κλώνο, μετατοπίζει την κατανομή των $P_{Trs}^{GG-CC}$ και $P_{Trs}^{AA-TT}$ προς υψηλότερες ή χαμηλότερες τιμές, αντιστοίχως, γεγονός που μπορεί να εξηγήσει την συσχέτιση αυτών των αποκλίσεων με την παρουσία του *mfd*. Ωστόσο, η TCR επιδιόρθωση συσχετίζεται ισχυρά και με τις αποκλίσεις των σταθμισμένων συχνοτήτων άλλων δινουκλεοτιδίων, τα οποία δεν μπορούν να σχηματίσουν διμερή πυριμιδίνης, όπως παρατηρούμε στην περίπτωση των $P_{Trs}^{AC-GT}$. Συνεπώς, *τα αποτελέσματά μας υποδεικνύουν ότι η δραστικότητα της TCR μπορεί καθαυτή να είναι πλαίσιο-εξαρτημένη*.

Συμπερασματικά, *η συζευγμένη με τη μεταγραφή επιδιόρθωση (TCR) πολώνει τους ρυθμούς υποκατάστασης προς τον κωδικό κλώνο και η ανάλυση της κατανομής των αποκλίσεων είναι ικανή να ανιχνεύσει τέτοιες ασυμμετρίες*. Επίσης, *η μελέτη μας υποδηλώνει ότι η ενεργότητα της TCR εξαρτάται από τους* $1^{ης}$ *τάξης γείτονες της προς επιδιόρθωση βλάβης του DNA*.

### **3.12.2** *Μονοπάτια επιδιόρθωσης του DNA, μη συζευγμένα με τη μεταγραφή*

Όλοι οι οργανισμοί έχουν αναπτύξει πολύπλοκα δίκτυα μοριακών μηχανισμών τροποποίησης και επιδιόρθωσης του DNA, προκειμένου να διαφυλάττουν την

ακεραιότητα του γονιδιώματός τους. Πολλά από τα στοιχεία που συγκροτούν τα επιδιορθωτικά μονοπάτια παραμένουν συντηρημένα μεταξύ εξελικτικά απομακρυσμένων βακτηριακών ειδών. Ωστόσο, τα στοιχεία αυτά μπορούν να οργανώνονται με πολλούς διαφορετικούς τρόπους, καθώς ορισμένες πρωτεΐνες που συμμετέχουν σε ένα δεδομένο επιδιορθωτικό μονοπάτι αναγνωρίζουν ποικιλία υποστρωμάτων, ενώ άλλες είναι μεταξύ τους συμπληρωματικές ως προς τις λειτουργίες που επιτελούν (Morita et al. 2010, Mielecki & Grzesiuk 2014). Συνεπώς, οι βλάβες του DNA επιδιορθώνονται μέσω ενός πλήθους μονοπατιών, τα οποία είναι εν μέρει συμπληρωματικά μεταξύ τους και σε πολλές περιπτώσεις αλληλοεπικαλύπτονται. Ως αποτέλεσμα της ποικιλότητας αυτών των μονοπατιών, μεταξύ των βακτηρίων υπάρχει μεγάλη ετερογένεια ως προς τους τύπους (modes) και την συχνότητα βλάβης και επιδιόρθωσης του DNA. Για μια αναλυτικότερη περιγραφή των εν λόγω επιδιορθωτικών μονοπατιών, βλ. ενότητα 1.6.2.

Εστιάζουμε σε συστήματα (*α*) άμεσης επιδιόρθωσης (direct reversal), (*β*) επιδιόρθωσης εκτομής βάσης (base excision repair, BER), (*γ*) επιδιόρθωσης εκτομής νουκλεοτιδίων (nucleotide excision repair, NER), (*δ*) επιδιόρθωσης αταίριαστων ζευγών (mismatch repair, MMR) και (*ε*) επιδιόρθωσης μέσω ανασυνδυασμού (recombination repair, RR)· βλ. ενότητα 1.6.2. Τα περισσότερα από αυτά τα μονοπάτια απαιτούν την αλληλεπίδραση σειράς πρωτεϊνών, οι οποίες κωδικοποιούνται από ένα αντίστοιχο πλήθος γενετικών τόπων (genomic loci). Βασιζόμενοι στην παρουσία ή απουσία αυτών των γενετικών τόπων, προσδιορίζουμε τον μοριακό φαινότυπο των βακτηρίων της συλλογής μας, αναφορικά με τα επιδιορθωτικά μονοπάτια που εξετάζουμε (βλ. ενότητα 2.10).

### **3.12.2.1** *Φυλογενετική διασπορά και ποικιλότητα των επιδιορθωτικών μονοπατιών*

Η ποικιλότητα των μοριακών φαινοτύπων διαφοροποιεί την ικανότητα επιδιόρθωσης του DNA στα βακτήρια που αντιπροσωπεύονται στην συλλογή μας. Ο Πίνακας 16 δείχνει πώς κατανέμονται μεταξύ των φύλων ή κλάσεων οι μοριακοί φαινότυποι που εξετάζουμε. Κάθε γραμμή αντιστοιχεί σε έναν δεδομένο μοριακό φαινότυπο. Συνεπώς, οι γραμμές του Πίνακα 16 είναι ενδεικτικές της φυλογενετικής διασποράς (phylogenetic dispersion) του αντίστοιχου επιδιορθωτικού μονοπατιού. Όπως προκύπτει, τα βακτήρια με ή δίχως ικανότητα επιδιόρθωσης μέσω ενός δεδομένου μοριακού μονοπατιού δεν αποτελούν

μονοφυλετικές ομάδες. Τα πιο πολλά από τα επιδιορθωτικά μονοπάτια που εξετάζουμε απαντώνται σε εξελικτικά απομακρυσμένα είδη, τα οποία ανήκουν σε διαφορετικά φύλα ή κλάσεις. Στην συλλογή μας, μόνο τα μονοπάτια της καθοδηγούμενης από μεθυλίωση επιδιόρθωσης αταίριαστων ζευγών (methyl-directed MMR) και της επιδιόρθωσης μέσω ομόλογου ανασυνδυασμού RecBC απαντώνται αποκλειστικά σε μία κλάση βακτηρίων, στα γ-Πρωτεοβακτήρια. Οι στήλες του Πίνακα 16 αντιστοιχούν σε φύλα ή κλάσεις, και δίνουν μια γενική εικόνα της ποικιλότητας των επιδιορθωτικών μονοπατιών που απαντώνται εντός αυτών των φυλογενετικών βαθμίδων. Παρατηρούμε ότι τα μέλη κάθε φύλου ή κλάσης έχουν διαφορετικούς μοριακούς φαινοτύπους όσον αφορά τέσσερα, κατ' ελάχιστον, από τα υπό εξέταση επιδιορθωτικά μονοπάτια. Μάλιστα, στα γ-Πρωτεοβακτήρια της συλλογής μας αντιπροσωπεύεται το σύνολο των διαφορετικών μοριακών φαινοτύπων.

**ΠΙΝΑΚΑΣ 16. Ποσοστά των βακτηρίων με συγκεκριμένους μοριακούς φαινοτύπους ανά φύλα ή κλάσεις**

| | *μοριακοί φαινότυποι* | Actinobacteria | Bacteroidetes | Chlamydiae | Cyanobacteria | Firmicutes | α-Proteobacteria | β-Proteobacteria | γ-Proteobacteria | δ-Proteobacteria | ε-Proteobacteria | Spirochaetes | Tenericutes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| άμεση επιδιόρθωση | **phrB** | | | | | | | | | | | | |
| | *proficient* | 6.2 | 1.1 | 0.6 | 8.5 | 18.6 | 11.9 | 8.5 | 37.3 | 5.1 | 1.1 | 1.1 | 0 |
| | *deficient* | 6.7 | 3 | 7.5 | 1.5 | 22.4 | 18.7 | 7.5 | 11.2 | 1.5 | 5.2 | 3 | 11.9 |
| | **ogt** | | | | | | | | | | | | |
| | *proficient* | 8.2 | 2.2 | 4.7 | 1.3 | 20.7 | 13.4 | 10.3 | 29.3 | 4.3 | 3.9 | 0.9 | 0.9 |
| | *deficient* | 1.3 | 1.3 | 0 | 17.7 | 19 | 19 | 1.3 | 16.5 | 1.3 | 0 | 5.1 | 17.7 |
| | **alkB** | | | | | | | | | | | | |
| | *proficient* | 28 | 0 | 0 | 0 | 0 | 12 | 28 | 28 | 4 | 0 | 0 | 0 |
| | *deficient* | 4.5 | 2.1 | 3.8 | 5.9 | 22 | 15 | 6.3 | 25.9 | 3.5 | 3.1 | 2.1 | 5.6 |
| | **ada** | | | | | | | | | | | | |
| | *proficient* | 0 | 1.1 | 0 | 1.1 | 3.3 | 28.6 | 14.3 | 46.2 | 1.1 | 2.2 | 2.2 | 0 |
| | *deficient* | 9.1 | 2.3 | 5 | 7.3 | 27.3 | 9.1 | 5.5 | 17.7 | 4.5 | 3.2 | 1.8 | 7.3 |
| BER | **ung** | | | | | | | | | | | | |
| | *proficient* | 7.7 | 2.1 | 4.7 | 0 | 24.5 | 3.4 | 8.6 | 33.9 | 2.6 | 3.9 | 1.7 | 6.9 |
| | *deficient* | 2.6 | 1.3 | 0 | 21.8 | 7.7 | 48.7 | 6.4 | 2.6 | 6.4 | 0 | 2.6 | 0 |
| | **mug** | | | | | | | | | | | | |
| | *proficient* | 18.2 | 0 | 0 | 0 | 6.1 | 6.1 | 15.2 | 51.5 | 3 | 0 | 0 | 0 |
| | *deficient* | 5 | 2.2 | 4 | 6.1 | 21.9 | 15.8 | 7.2 | 23 | 3.6 | 3.2 | 2.2 | 5.8 |
| | **nth** | | | | | | | | | | | | |
| | *proficient* | 5.9 | 2.1 | 3.8 | 5.9 | 21.7 | 15.9 | 8.6 | 27.6 | 3.4 | 3.1 | 2.1 | 0 |
| | *deficient* | 14.3 | 0 | 0 | 0 | 0 | 0 | 0 | 4.8 | 4.8 | 0 | 0 | 76.2 |

| μοριακοί φαινότυποι | | Actinobacteria | Bacteroidetes | Chlamydiae | Cyanobacteria | Firmicutes | α-Proteobacteria | β-Proteobacteria | γ-Proteobacteria | δ-Proteobacteria | ε-Proteobacteria | Spirochaetes | Tenericutes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **BER** | **mutM** | | | | | | | | | | | | |
| | *proficient* | 7.1 | 0.8 | 0.4 | 6.4 | 22.2 | 15.8 | 9.4 | 27.8 | 4.1 | 0 | 0 | 6 |
| | *deficient* | 2.2 | 8.9 | 22.2 | 0 | 8.9 | 8.9 | 0 | 15.6 | 0 | 20 | 13.3 | 0 |
| | **nei** | | | | | | | | | | | | |
| | *proficient* | 34.2 | 0 | 0 | 2.6 | 0 | 0 | 0 | 60.5 | 2.6 | 0 | 0 | 0 |
| | *deficient* | 2.6 | 2.2 | 4 | 5.9 | 23.1 | 16.8 | 9.2 | 21.2 | 3.7 | 3.3 | 2.2 | 5.9 |
| | **tag** | | | | | | | | | | | | |
| | *proficient* | 8.5 | 1.6 | 0.5 | 0 | 24.9 | 13.8 | 11.1 | 33.9 | 3.2 | 0.5 | 1.1 | 1.1 |
| | *deficient* | 3.3 | 2.5 | 8.2 | 13.9 | 13.1 | 16.4 | 3.3 | 13.9 | 4.1 | 6.6 | 3.3 | 11.5 |
| | **alkA** | | | | | | | | | | | | |
| | *proficient* | 1.7 | 0 | 0.8 | 1.7 | 12.6 | 23.5 | 18.5 | 37.8 | 0.8 | 0 | 1.7 | 0.8 |
| | *deficient* | 9.4 | 3.1 | 5.2 | 7.8 | 25 | 9.4 | 1.6 | 18.8 | 5.2 | 4.7 | 2.1 | 7.8 |
| | **mutY** | | | | | | | | | | | | |
| | *proficient* | 7 | 2.3 | 4.3 | 4.7 | 21 | 12.5 | 9.7 | 30 | 4.3 | 3.1 | 1.2 | 0 |
| | *deficient* | 3.7 | 0 | 0 | 9.3 | 16.7 | 25.9 | 0 | 7.4 | 0 | 1.9 | 5.6 | 29.6 |
| | **GO system** | | | | | | | | | | | | |
| | *proficient* | 7.8 | 0 | 0.5 | 2.4 | 24.3 | 13.6 | 12.1 | 34.5 | 4.9 | 0 | 0 | 0 |
| | *deficient* | 3.8 | 5.7 | 9.5 | 11.4 | 12.4 | 17.1 | 0 | 9.5 | 1 | 8.6 | 5.7 | 15.2 |
| **NER** | **GGR** | | | | | | | | | | | | |
| | *proficient* | 6.8 | 2 | 3.7 | 5.8 | 21.4 | 13.6 | 8.5 | 24.5 | 3.7 | 3.1 | 1.7 | 5.1 |
| | *deficient* | 0 | 0 | 0 | 0 | 0 | 35.3 | 0 | 52.9 | 0 | 0 | 5.9 | 5.9 |
| | **TCR** | | | | | | | | | | | | |
| | *proficient* | 7.2 | 2.2 | 4 | 6.2 | 22.8 | 13.4 | 9.1 | 26.1 | 4 | 3.3 | 1.8 | 0 |
| | *deficient* | 0 | 0 | 0 | 0 | 0 | 25.7 | 0 | 25.7 | 0 | 0 | 2.9 | 45.7 |

| | μοριακοί φαινότυποι | Actinobacteria | Bacteroidetes | Chlamydiae | Cyanobacteria | Firmicutes | α-Proteobacteria | β-Proteobacteria | γ-Proteobacteria | δ-Proteobacteria | ε-Proteobacteria | Spirochaetes | Tenericutes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MMR** | **methyl-directed** | | | | | | | | | | | | |
| | *proficient* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 |
| | *deficient* | 7.6 | 2.3 | 4.2 | 6.4 | 23.9 | 17.4 | 9.5 | 12.9 | 4.2 | 3.4 | 2.3 | 6.1 |
| | **nick-directed** | | | | | | | | | | | | |
| | *proficient* | 0.5 | 3 | 5.6 | 4.6 | 30.5 | 22.8 | 12.7 | 13.2 | 5.6 | 0 | 1.5 | 0 |
| | *deficient* | 16.7 | 0 | 0 | 7 | 2.6 | 0.9 | 0 | 48.2 | 0 | 7.9 | 2.6 | 14 |
| | **VSR-patch** | | | | | | | | | | | | |
| | *proficient* | 0 | 0 | 0 | 0 | 0 | 3.4 | 13.8 | 79.3 | 3.4 | 0 | 0 | 0 |
| | *deficient* | 7.1 | 2.1 | 3.9 | 6 | 22.3 | 16 | 7.4 | 20.6 | 3.5 | 3.2 | 2.1 | 5.7 |
| **RR** | **RecFOR** | | | | | | | | | | | | |
| | *proficient* | 0.8 | 2.4 | 4.4 | 6 | 24.6 | 17.9 | 9.5 | 29.4 | 4 | 0 | 1.2 | 0 |
| | *deficient* | 30.5 | 0 | 0 | 3.4 | 1.7 | 1.7 | 1.7 | 11.9 | 1.7 | 15.3 | 5.1 | 27.1 |
| | **RecBC** | | | | | | | | | | | | |
| | *proficient* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 |
| | *deficient* | 6.7 | 2 | 3.7 | 5.7 | 21.2 | 15.5 | 8.4 | 22.6 | 3.7 | 3 | 2 | 5.4 |

ΣΗΜΕΙΩΣΕΙΣ.- Η επί τοις εκατό συμμετοχή κάθε φύλου ή κλάσης ανά ομάδα βακτηρίων με συγκεκριμένο μοριακό φαινότυπο. Οι ταξινομικές βαθμίδες που λαμβάνουμε υπόψιν είναι εκείνες για τις οποίες πραγματοποιήσαμε την κλαδιστική μας ανάλυση (βλ. ενότητα 3.9). Οι μοριακοί φαινότυποι δηλώνουν την παρουσία ή την έλλειψη των αντίστοιχων επιδιορθωτικών μονοπατιών. "*proficient*": βακτήρια με ικανότητα επιδιόρθωσης, "*deficient*": βακτήρια δίχως ικανότητα επιδιόρθωσης.

Γενικά, στα βακτήρια υπάρχουν διαφορές στην ικανότητα επιδιόρθωσης του DNA όχι μόνον μεταξύ διαφορετικών ειδών, αλλά και μεταξύ στελεχών ή φυσικών δειγμάτων (natural isolates) του ίδιου είδους, ως αποτέλεσμα εκτεταμένης

οριζόντιας μεταφοράς γονιδίων που εμπλέκονται σε διαφορετικά επιδιορθωτικά μονοπάτια (Denamur et al. 2000). Το γονιδίωμα των βακτηρίων που στερούνται ένα συγκεκριμένο επιδιορθωτικό ένζυμο είναι πιθανό να υφίσταται ταχύτερα εκείνου του τύπου τις μεταλλάξεις οι οποίες συνδέονται με τις βλάβες του DNA που αποτελούν το υπόστρωμα του αντίστοιχου μηχανισμού επιδιόρθωσης. Καθώς ο κωδικός κλώνος είναι πιο εκτεθειμένος σε αλλοιώσεις από ότι ο μεταγραφόμενος, η αύξηση των μεταλλακτικών ρυθμών απουσία επιδιορθωτικών μηχανισμών αναμένεται να είναι πολωμένη προς τον κωδικό κλώνο.

### **3.12.2.2** *Μονοπάτια επιδιόρθωσης του DNA που συσχετίζονται με τις ειδικές ανά κλώνο ασυμμετρίες*

Στην ενότητα 3.12.1 διαπιστώσαμε ότι η TCR επιδιόρθωση συσχετίζεται με διακριτά πρότυπα $S_{Trs}^{A-T}$. Από την άλλη, δεν εμφανίζεται αντίστοιχη συσχέτιση στην περίπτωση των $S_{Trs}^{G-C}$. Συνεπώς, άλλοι επιδιορθωτικοί μηχανισμοί, όπως αυτοί που αναφέρονται στον Πίνακα 16, ενδέχεται να εμπλέκονται στην εμφάνιση των $S_{Trs}^{G-C}$ αποκλίσεων. Προκειμένου να εντοπίσουμε τέτοιους πιθανούς συσχετισμούς, για κάθε επιδιορθωτικό μονοπάτι συγκρίνουμε τις κατανομές των $S_{Trs}^{G-C}$ στα βακτήρια με ή δίχως ικανότητα επιδιόρθωσης (Πίνακας 17).

**ΠΙΝΑΚΑΣ 17. Ανάλυση της κατανομής των $S_{Trs}^{G-C}$ αποκλίσεων συναρτήσει συγκεκριμένων επιδιορθωτικών μονοπατιών**

**τετραπλά εκφυλισμένες θέσεις**

| | | $S_{Trs}^{G-C}$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | **διάμεσος** | | ***p*-τιμή** | | | |
| | | *δίχως ικανότητα επιδιόρθωσης* | *με ικανότητα επιδιόρθωσης* | *αρχική* | | *διορθωμένη* | |
| | ***μοριακοί φαινότυποι*** | | | | | | |
| **άμεση επιδιόρθωση** | **phrB** | -0.0635 | -0.0206 | 0.00189 | ** | 0.0378 | * |
| | **ogt** | -0.0638 | -0.028 | 0.0289 | * | 0.578 | - |
| | **alkB** | -0.0439 | 0.0103 | 0.00163 | ** | 0.0326 | * |
| | **ada** | -0.058 | -0.008 | 9.77e-07 | *** | 1.954e-05 | *** |
| **BER** | **ung** | -0.0838 | -0.0212 | 0.0361 | * | 0.722 | - |
| | **mug** | -0.0458 | 0.0773 | 1.89e-08 | *** | 3.78e-07 | *** |
| | **nth** | -0.113 | -0.0331 | 0.049 | * | 0.98 | - |
| | **mutM** | -0.0515 | -0.0307 | 0.433 | - | 1 | - |
| | **nei** | -0.0444 | 0.00169 | 0.000488 | *** | 0.00976 | ** |
| | **tag** | -0.0654 | -0.0229 | 0.251 | - | 1 | - |
| | **alkA** | -0.095 | -0.00831 | 4.06e-11 | *** | 8.12e-10 | *** |
| | **mutY** | -0.0969 | -0.0323 | 0.0963 | - | 1 | - |
| | **GO system** | -0.0617 | -0.0229 | 0.0194 | * | 0.388 | - |
| **NER** | **GGR** | 0.0345 | -0.0412 | 0.0299 | * | 0.598 | - |
| | **TCR** | -0.00305 | -0.0412 | 0.239 | - | 1 | - |
| **MMR** | **καθοδηγούμενη από μεθυλίωση** | -0.0531 | 0.0221 | 5.29e-07 | *** | 1.058e-05 | *** |
| | **καθοδηγούμενη από εγκοπή** | -0.0242 | -0.0448 | 0.439 | - | 1 | - |
| | **πολύ βραχέως τμήματος (VSP)** | -0.0451 | 0.0906 | 1.75e-07 | *** | 3.5e-06 | *** |
| **RR** | **RecFOR** | -0.0762 | -0.0298 | 0.0578 | - | 1 | - |
| | **RecBC** | -0.0437 | 0.106 | 4.94e-09 | *** | 9.88e-08 | *** |

ΣΗΜΕΙΩΣΕΙΣ.- Για κάθε μοριακό φαινότυπο, υπολογίζουμε την διάμεσο των $S_{Trs}^{G-C}$ για τις ομάδα των βακτηρίων με ή δίχως ικανότητα επιδιόρθωσης, και συγκρίνουμε τις κατανομές των $S_{Trs}^{G-C}$ στις δύο αυτές ομάδες, χρησιμοποιώντας τον *στατιστικό έλεγχο*

*αθροίσματος διατάξεων του Wilcoxon* (two-tailed Wilcoxon rank-sum test). Οι *p*-τιμές του τεστ Wilcoxon δηλώνουν την στατιστική σημαντικότητα της διαφοράς των υπό σύγκριση κατανομών. Καθώς πραγματοποιούμε πολλαπλές συγκρίσεις και προκειμένου να ελέγξουμε το *αθροιστικό σφάλμα* Τύπου I (false positive), σταθμίζουμε τις *p*-τιμές χρησιμοποιώντας την *διόρθωση κατά Bonferroni*. "-": *p*-τιμή ≥ 0.05, "*": 0.05 > *p*-τιμή ≥ 0.01, "**": 0.01 > *p*-τιμή ≥ 0.001, "***": *p*-τιμή < 0.001

BER: σύστημα επιδιόρθωσης εκτομής βάσης, NER: σύστημα επιδιόρθωσης εκτομής νουκλεοτιδίων, MMR: σύστημα επιδιόρθωσης αταίριαστων ζευγών βάσεων, RR: σύστημα επιδιόρθωσης μέσω ανασυνδυασμού

Σύμφωνα με τις διορθωμένες *p*-τιμές (Πίνακας 17), υπάρχει πολύ σημαντική στατιστικά συσχέτιση (*p*-τιμή < 0.01) ανάμεσα στις $S_{Trs}^{G-C}$ και τα συστήματα (*α*) επιδιόρθωσης αλκυλιωμένων βάσεων (*ada*, *alk*A), (*β*) εκτομής οξειδωμένων πυριμιδινών (*nei*) (*γ*) εκτομής 3,$N^4$-εθενοκυτοσίνης (εC) και ουρακίλης (U) από αταίριαστα ζεύγη εC:G ή U:G, αντίστοιχα (*mug*), (*δ*) επιδιόρθωσης MMR κατευθυνόμενης από μεθυλίωση, (*ε*) επιδιόρθωσης πολύ βραχέως τμήματος (VSP), και (*στ*) επιδιόρθωσης μέσω ανασυνδυασμού (RecBC). Σε όλες τις προαναφερθείσες περιπτώσεις, πλην των *ada* και *alk*A, παρατηρούμε μια μετατόπιση από αρνητικές σε θετικές τιμές των $S_{Trs}^{G-C}$ στα βακτήρια με ικανότητα επιδιόρθωσης (βλ. αντίστοιχες τιμές των διαμέσων, Πίνακας 17). *Το γεγονός αυτό υποδεικνύει μια αντιστροφή στην πόλωση των ρυθμών υποκατάστασης στα βακτήρια όπου τα εν λόγω συστήματα επιδιόρθωσης είναι ενεργά*. Επιπλέον, σε όλες τις περιπτώσεις όπου εμφανίζεται στατιστικά σημαντική διαφορά των κατανομών $S_{Trs}^{G-C}$ (*p*-τιμή < 0.05), οι αποκλίσεις μετατοπίζονται από χαμηλότερες σε υψηλότερες προσημασμένες (signed) τιμές στα βακτήρια με ικανότητα επιδιόρθωσης, σε σύγκριση με εκείνα στα οποία το αντίστοιχο επιδιορθωτικό μονοπάτι απουσιάζει.

Σε αντίθεση με τα παραπάνω, τα βακτήρια που διαθέτουν το σύστημα καθολικής επιδιόρθωσης του γονιδιώματος (GGR) εμφανίζουν μικρότερες προσημασμένες τιμές από ότι τα βακτήρια δίχως GGR. Η ίδια τάση παρατηρείται και όταν συγκρίνουμε τα $TCR^+$ και $TCR^-$ βακτήρια. Δηλαδή, *τόσο η GGR όσο και η TCR φαίνεται να πολώνουν τους ρυθμούς υποκατάστασης προς την ίδια κατεύθυνση*, πράγμα που εξηγείται από το γεγονός ότι στην TCR οι βλάβες του DNA επιδιορθώνονται μέσω του ίδιου μονοπατιού που δρα και στην GGR (Deaconescu 2013). Ωστόσο, και στις δύο περιπτώσεις (GGR και TCR) οι κατανομές των $S_{Trs}^{G-C}$ δεν εμφανίζουν στατιστικά σημαντική διαφορά μεταξύ των

βακτηρίων με ή δίχως ικανότητα υποκατάστασης (διορθωμένες *p*-τιμές > 0.05).

Όπως αναφέραμε, οι $S_{Trs}^{G-C}$ συσχετίζονται ισχυρά με τα συστήματα επιδιόρθωσης MMR κατευθυνόμενης από μεθυλίωση και επιδιόρθωσης πολύ βραχέως τμήματος (VSP). Ωστόσο, και τα δύο αυτά συστήματα πρωτίστως επιδιορθώνουν σφάλματα που εντοπίζουν στον νεοσυντιθέμενο κλώνο του χρωμοσώματος. Ως εκ τούτου θα αναμέναμε πως δεν θα συσχετίζονταν με τις συζευγμένες με τη μεταγραφή αποκλίσεις (Trs), σε αντίθεση με τα όσα παρατηρούμε (διορθωμένη *p*-τιμή < 0.001). Η αντίθεση αυτή αίρεται εάν λάβουμε υπόψιν ότι στα βακτήρια η αντιγραφή και η μεταγραφή πολύ συχνά πραγματοποιούνται ταυτόχρονα στον ίδιο κλώνο του DNA. Μάλιστα, σε περισσότερα από το 81% των χρωμοσωμάτων της συλλογής μας, η κάλυψη του οδηγού κλώνου από κωδικούς κλώνους υπερβαίνει το 50%. Συνεπώς, οι παρατηρούμενες συσχετίσεις των $S_{Trs}^{G-C}$ με τα συστήματα επιδιόρθωσης MMR κατευθυνόμενης από μεθυλίωση και VSP μπορούν να εκληφθούν ως παρεπόμενα του εμπλουτισμού του οδηγού κλώνου σε κωδικούς κλώνους.

Σε προγενέστερες μελέτες αναφέρεται ότι βακτήρια που στερούνται τα γονίδια *rec*A και *pri*A εμφανίζουν εντονότερες ασυμμετρίες στην σύσταση των χρωμοσωμάτων τους (Klasson & Andersson 2006). Τα δύο αυτά γονίδια συμμετέχουν στο σύστημα επιδιόρθωσης RecBC μέσω ομόλογου ανασυνδυασμού. Σύμφωνα με τα αποτελέσματά μας (Πίνακας 17) τα γονίδια *rec*A και *pri*A συσχετίζονται έντονα με τις $S_{Trs}^{G-C}$ αποκλίσεις (RecBC: διορθωμένη *p*-τιμή < 0.001). Ωστόσο, τα βακτήρια δίχως *rec*A και *pri*A ($RecBC^-$) εμφανίζουν πολύ ασθενείς αποκλίσεις (διάμεσος $S_{Trs}^{G-C}$: -0.0437) συγκρινόμενα με εκείνα που φέρουν και τα δύο αυτά γονίδια ($RecBC^+$, διάμεσος $S_{Trs}^{G-C}$: 0.106). Συνεπώς, *η ανάλυσή μας υποδεικνύει ότι ο ανασυνδυασμός μπορεί να ενισχύει τις ειδικές ανά κλώνο ασυμμετρίες*, σε συμφωνία και με τα σχετικά ευρήματα άλλων εργασιών (Rocha et al. 2005).

Συμπερασματικά, η ανάλυση των συζευγμένων με τη μεταγραφή αποκλίσεων προσφέρει έναν πλούτο στοιχείων σχετικά με μοριακούς μηχανισμούς τροποποίησης και επιδιόρθωσης του γενετικού υλικού, διακρίνοντας ποιοι από αυτούς τους μηχανισμούς εμπλέκονται στην ασύμμετρη εξέλιξη των κλώνων του DNA.

### **3.12.3** *Οι ισομορφές της α-υπομονάδας της πολυμεράσης PolIII*

Η αντιγραφή του γονιδιώματος είναι ένας από τους βασικούς μηχανισμούς που επάγουν ασυμμετρίες στους ρυθμούς υποκατάστασης, στην κλίμακα ολόκληρου του γονιδιώματος (Rocha et al. 1999). Στα βακτήρια η αντιγραφή και των δύο νεοσυντιθέμενων κλώνων του DNA καταλύεται από το διμερές της α-υπομονάδας της DNA πολυμεράσης III (PolIII) (Zhao et al. 2006). Τα μονομερή της *α*-υπομονάδας απαντώνται σε τέσσερις διαφορετικές ισομορφές, που κωδικοποιούνται από τα γονίδια *dna*E1, *dna*E2, *dna*E3 και *pol*C. Ο συνδυασμός αυτών των ισομορφών ανά δύο δίνει τρεις διαφορετικούς τύπους διμερών: (*α*) το ομοδιμερές της DnaE1 (dnaE α-υπομονάδα), (*β*) το ετεροδιμερές της DnaE1 με την DnaE2 (dnaE2 α-υπομονάδα) και (*γ*) το ετεροδιμερές της PolC με την DnaE1 ή με την DnaE3 (PolC α-υπομονάδα) (Hu et al. 2007). Τα διμερή αυτά διαφέρουν ως προς την καταλυτική τους ενεργότητα. Για μια εκτενέστερη παρουσίαση σχετικά με τις ισομορφές της *α*-υπομονάδας, βλ. ενότητα 1.5.

Χωρίζουμε την συλλογή μας σε τρεις ομάδες ('dnaE', 'dnaE2' και 'polC'), σύμφωνα με τον τύπο του διμερούς της *α*-υπομονάδας που αντιγράφει το κάθε γονιδίωμα. Η ομάδα polC αποτελείται από το σύνολο των Firmicutes και των Tenericutes που περιλαμβάνονται στην συλλογή μας. Ακολούθως, για κάθε χρωμόσωμα υπολογίζουμε τις συζευγμένες με την αντιγραφή αποκλίσεις (Rep) (βλ. ενότητα 2.9) και συγκρίνουμε ανά ζεύγη τις αντίστοιχες κατανομές στις ομάδες 'dnaE', 'dnaE2' και 'polC' (Πίνακας 18).

**ΠΙΝΑΚΑΣ 18. Ανάλυση της κατανομής των αποκλίσεων συναρτήσει των ισομορφών της α-υπομονάδας της πολυμεράσης PolIII**

| | | $S_{Rep}^{A-T}$ | $S_{Rep}^{G-C}$ | $P_{Rep}^{AG-CT}$ | $P_{Rep}^{GA-TC}$ | $P_{Rep}^{GG-CC}$ | $P_{Rep}^{AA-TT}$ | $P_{Rep}^{AC-GT}$ | $P_{Rep}^{CA-TG}$ |
|---|---|---|---|---|---|---|---|---|---|
| διάμεσος | 'dnaE' | -0.0357 | 0.0806 | 0.00521 | 0.0137 | 0.000299 | 0.0161 | -0.0309 | 0.00614 |
| | 'dnaE2' | -0.07 | 0.0386 | 0.017 | 0.0152 | 0.0168 | 0.0371 | -0.0532 | 0.0116 |
| | 'polC' | 0.0241 | 0.132 | 0.000189 | 0.012 | -0.0222 | -0.029 | -0.0376 | 0.0203 |
| *p*-τιμή | 'dnaE' 'dnaE2' | *** | *** | * | – | *** | *** | *** | – |
| | 'dnaE' 'polC' | *** | *** | – | – | ** | *** | – | ** |
| | 'polC' 'dnaE2' | *** | *** | *** | – | *** | *** | ** | ** |

ΣΗΜΕΙΩΣΕΙΣ.- Υπολογίζουμε τις συζευγμένες με την αντιγραφή αποκλίσεις, για το σύνολο των χρωμοσωμάτων της συλλογής μας. Ανάλογα με τον τύπο της *α*-υπομονάδας της PolIII των βακτηρίων που αντιπροσωπεύονται στην συλλογή μας, διακρίνουμε τρεις ομάδες γονιδιωμάτων ('dnaE', 'dnaE2', 'polC'). Για κάθε ομάδα δίδεται η διάμεσος των αποκλίσεων. Για κάθε απόκλιση συγκρίνουμε ανά ζεύγη τις κατανομές των αντίστοιχων τιμών στις ομάδες 'dnaE', 'dnaE2' και 'polC', χρησιμοποιώντας το *στατιστικό έλεγχο αθροίσματος διατάξεων του Wilcoxon* (two-tailed Wilcoxon rank-sum test). Οι *p*-τιμές του τεστ Wilcoxon δηλώνουν την στατιστική σημαντικότητα της διαφοράς των υπό σύγκριση κατανομών. "–": *p*-τιμή ≥ 0.05, "*": 0.05 > *p*-τιμή ≥ 0.01, "**": 0.01 > *p*-τιμή ≥ 0.001, "***": *p*-τιμή < 0.001

Μια προγενέστερη μελέτη (Rocha 2002) με αντικείμενο την επίδραση της αντιγραφής στις ειδικές ανά κλώνο ασυμμετρίες του DNA, χρησιμοποιώντας τον ίδιο στατιστικό έλεγχο (two-tailed Wilcoxon rank-sum test) που εφαρμόζουμε στην παρούσα ανάλυση, κατέληξε στο συμπέρασμα ότι οι αποκλίσεις G-C δεν συσχετίζονται με τους διαφορετικούς τύπους της *α*-υπομονάδας. Ωστόσο, σε εκείνη την εργασία τα βακτήρια διακρίθηκαν σε δύο ομάδες, σύμφωνα με την παρουσία ή μη του γονιδίου *pol*C, χωρίς να λαμβάνεται υπόψιν το *dna*E2. Επιπλέον, χρησιμοποιήθηκε διαφορετική μεθοδολογία προκειμένου να

υπολογιστούν οι αποκλίσεις G-C που σχετίζονται με την αντιγραφή, ενώ η ανάλυση αφορούσε μόνο 64 χρωμοσώματα. Αντίθετα, όταν χωρίσαμε την δική μας συλλογή χρωμοσωμάτων (340 τον αριθμό) σε εκείνα που φέρουν το *polC* και σε εκείνα από τα οποία το *polC* απουσιάζει, και συγκρίναμε τις συζευγμένες με την αντιγραφή αποκλίσεις G-C ($S_{Rep}^{G-C}$) στις δύο αυτές ομάδες, εντοπίσαμε στατιστικά σημαντική διαφορά μεταξύ τους (*p*-τιμή = $2.201 \cdot 10^{-11}$). Περαιτέρω, σύμφωνα με το σχήμα κατάταξης που επιλέγουμε για την ανάλυσή μας ('dnaE'/'dnaE2'/'polC'), το οποίο λαμβάνει υπόψιν όλους τους γνωστούς συνδυασμούς των ισομορφών της *α*-υπομονάδας, οι $S_{Rep}^{G-C}$ ομαδοποιούνται σε τρία υποσύνολα τα οποία διακρίνονται μεταξύ τους με επίσης πολύ μεγάλη στατιστική σημαντικότητα, όπως δηλώνουν οι αντίστοιχες *p*-τιμές, που είναι: $1.738 \cdot 10^{-15}$, μεταξύ 'polC' και 'dnaE2', (β) $3.082 \cdot 10^{-6}$, μεταξύ 'polC' και 'dnaE', και (γ) $1.664 \cdot 10^{-10}$, μεταξύ 'dnaE2' και 'dnaE'.

Όπως προκύπτει από τον Πίνακα 18, *οι συζευγμένες με την αντιγραφή αποκλίσεις A-T και G-C συσχετίζονται πολύ ισχυρά με την α-υπομονάδα της PolIII*. Οι κατανομές που ακολουθούν τόσο οι $S_{Rep}^{A-T}$ όσο και οι $S_{Rep}^{G-C}$ στις ομάδες 'dnaE', 'dnaE2' και 'polC' είναι σαφώς διακριτές μεταξύ τους (*p*-τιμές << $10^4$). Επίσης, οι αποκλίσεις των σταθμισμένων δινουκλεοτιδικών συχνοτήτων εμφανίζουν στατιστικά σημαντικές συσχετίσεις με τον τύπο της *α*-υπομονάδας (πχ. $P_{Rep}^{GG-CC}$, $P_{Rep}^{AA-TT}$ και $P_{Rep}^{AC-GT}$, Πίνακας 18). Μόνη εξαίρεση αποτελούν οι $P_{Rep}^{GA-TC}$, που δεν εμφανίζουν σημαντική συσχέτιση με καμία από τις ομάδες 'dnaE', 'dnaE2' και 'polC'. Μάλιστα, η παρουσία της PolC φαίνεται να καθορίζει σε μεγάλο βαθμό το πρόσημο των αποκλίσεων, σε συμφωνία και με προηγούμενες δημοσιευμένες εργασίες (Worning et al. 2006). Η διάμεσος των $S_{Rep}^{A-T}$ είναι αρνητική στις ομάδες 'dnaE' και 'dnaE2' και θετική στην ομάδα 'polC'. Επίσης, *η ανάλυσή μας δείχνει για πρώτη φορά ότι η PolC ενδέχεται να διαμορφώνει την φορά των ασυμμετριών στα πρότυπα υποκατάστασης που εξαρτώνται από την ταυτότητα των 1$^{ης}$ τάξης γειτονικών βάσεων*. Συγκεκριμένα, η διάμεσος των $P_{Rep}^{GG-CC}$ και $P_{Rep}^{AA-TT}$ είναι θετική στις ομάδες 'dnaE' και 'dnaE2' και αρνητική στην ομάδα 'polC'.

Τόσο η PolC όσο και η DnaE2 σχηματίζουν ετεροδιμερή αποτελούμενα από ισομορφές της *α*-υπομονάδας οι οποίες διαφέρουν μεταξύ τους ως προς την ενεργότητα ενσωμάτωσης νουκλεοτιδίων και την επιδιορθωτική ενεργότητά τους (proofreading activity). Η PolC ισομορφή αντιγράφει ειδικά τον οδηγό κλώνο και όχι το συνοδό (Dervyn et al. 2001) εισάγοντας μία συστηματική ασυμμετρία στον αναδιπλασιασμό του DNA των βακτηρίων της ομάδας polC. Η

DnaE2 καταλύει την επαγόμενη από το σύστημα SOS αντιγραφή του DNA διαμέσου βλάβης (SOS-induced translesion synthesis, TLS) (Boshoff et al. 2003, Galhardo et al. 2005), που έχει ως αποτέλεσμα την αύξηση των μεταλλακτικών ρυθμών. Τα ευρήματά μας (Πίνακας 18) υποδεικνύουν ότι *οι ασυμμετρίες των ετεροδιμερών της α-υπομονάδας επάγουν μεταλλακτικές πολώσεις στην κλίμακα ολόκληρου του γονιδιώματος. Οι πολώσεις αυτές αφορούν σημειακές μεταλλάξεις οι ρυθμοί των οποίων μπορεί να τροποποιούνται ανάλογα με την ταυτότητα των γειτονικών τους βάσεων*, όπως αποκαλύπτουν οι ισχυρές συσχετίσεις της α-υπομονάδας με τις $P_{Rep}^{AG-CT}$, $P_{Rep}^{GG-CC}$, $P_{Rep}^{AA-TT}$ και $P_{Rep}^{AC-GT}$. Στην ομάδα 'dnaE' η αντιγραφή των δύο κλώνων του DNA καταλύεται από την ίδια ισομορφή της α-υπομονάδας. Συνεπώς, οι συσχετίσεις που παρουσιάζονται στον Πίνακα 18 θα πρέπει να αποδοθούν στις εγγενείς ασυμμετρίες της ίδιας της διχάλας της αντιγραφής (Lobry 1996).

Η *μελέτη της αντιγραφής και των επιδιορθωτικών μονοπατιών του DNA με τα εργαλεία που παρέχει η ανάλυση των αποκλίσεων, προσφέρει χρήσιμες ενδείξεις σχετικά με συγκεκριμένες πτυχές αυτών των μηχανισμών και μπορεί να κατευθύνει την περαιτέρω πειραματική τους διερεύνηση.*

# **4.** ΣΥΜΠΕΡΑΣΜΑΤΑ

Οι φυσικοχημικές ιδιότητες του DNA συνεπάγονται συγκεκριμένου τύπου κανονικότητες στην σύστασή του, εφόσον οι ρυθμοί υποκατάστασης είναι συμμετρικοί ως προς τους δύο κλώνους του. Η παρούσα μελέτη έχει ως αντικείμενό της την διερεύνηση των αποκλίσεων από αυτές τις κανονικότητες. Η ασυμμετρία στην εξέλιξη των δύο κλώνων του DNA είναι βαθιά ριζωμένη στις μεταλλακτικές και επιλεκτικές διαδικασίες που διαμορφώνουν το γενετικό υλικό των οργανισμών. Για τους σκοπούς της εργασίας μας, εστιάζουμε στο DNA των βακτηρίων. Ο χαμηλότερος βαθμός πολυπλοκότητας της οργάνωσης του γενετικού υλικού των βακτηρίων, σε σχέση με εκείνο των ευκαρυωτικών οργανισμών, το καθιστά ιδιαίτερα πρόσφορο για την ανάλυση των ειδικών ανά κλώνο ασυμμετριών.

Οι αποκλίσεις από τις κανονικότητες, θεωρούμενες ως αποτέλεσμα των ασυμμετριών μεταξύ των δύο κλώνων, αποτελούν την βάση για τον ορισμό απλών ποσοτήτων οι οποίες προσφέρουν πολύτιμες πληροφορίες για την εξελικτική δυναμική του γονιδιώματος. Εξ όσων γνωρίζουμε, η εργασία μας είναι η πρώτη που αποδεικνύει ότι οι συσχετίσεις των $1^{ης}$ τάξης γειτονικών βάσεων εμφανίζουν συστηματικές πολώσεις μεταξύ των κλώνων του DNA. Για την ποσοτικοποίηση αυτών των πολώσεων εισάγουμε το μέτρο των *αποκλίσεων των σταθμισμένων δινουκλεοτιδικών συχνοτήτων*, που μας επιτρέπει να εκτιμήσουμε την ύπαρξη ασυμμετριών στους ρυθμούς των υποκαταστάσεων οι οποίες εξαρτώνται από την ταυτότητα των παρακείμενων βάσεων (context-dependend substitutions). Οι αποκλίσεις των σταθμισμένων συχνοτήτων είναι ένα ιδιαίτερα εύχρηστο μέτρο, καθώς δίνει την δυνατότητα να εξάγουμε σημαντικά συμπεράσματα για τους ρυθμούς υποκατάστασης χωρίς να προϋποθέτει την στοίχιση ομόλογων αλληλουχιών.

Οι ασυμμετρίες της σύστασης του DNA κατά μήκος των κωδικών περιοχών (CDS) επηρεάζουν σημαντικά την χρήση κωδικονίων, ενώ μπορούν να καθορίσουν ακόμα και την ταυτότητα των κωδικοποιούμενων αμινοξέων. Τα αποτελέσματά μας καταδεικνύουν ότι οι CDS-αποκλίσεις τείνουν να είναι παρεμφερείς στα

γονίδια του ίδιου οργανισμού, αλλά συγκροτούν πρότυπα που είναι διακριτά μεταξύ διαφορετικών οργανισμών. Τα πρότυπα αυτά παρακολουθούν τις εξελικτικές σχέσεις των βακτηρίων και η πληροφορία που περιέχουν μπορεί να χρησιμοποιηθεί σε μεθόδους φυλογενετικής ανακατασκευής. Συνεπώς, και σε αντίθεση με προγενέστερες εκτιμήσεις, οι ασύμμετρες υποκαταστάσεις δεν μπορούν να αποδοθούν σε εξελικτικές τάσεις κοινές μεταξύ διαφορετικών βακτηριακών ειδών, καθώς είναι σε μεγάλο βαθμό καθορισμένες ανά είδος. Μάλιστα, όταν εξετάζουμε τις σταθμισμένες συχνότητες των δινουκλεοτιδίων, οι ειδικές ανά κλώνο αποκλίσεις τους συσχετίζονται εντονότερα με την φυλογένεση των βακτηρίων από ότι η συμμετρική ως προς τους δύο κλώνους εκδοχή τους, η οποία θεωρείται ως μία χαρακτηριστική *γονιδιωματική υπογραφή*. Το γεγονός αυτό είναι ενδεικτικό της ιδιαίτερης σημασίας που διαδραματίζουν οι ασυμμετρίες στην εξέλιξη του DNA, καθώς αφήνουν ένα είδος *δακτυλικού αποτυπώματος* στο γονιδίωμα κάθε οργανισμού.

Οι αποκλίσεις, ως εργαλεία ανάλυσης του γονιδιώματος, μπορούν να διαλευκάνουν ποικίλα ερωτήματα σχετικά με την προέλευση των ασυμμετριών στους ρυθμούς υποκατάστασης. Παρουσιάζουμε ένα απλό μοντέλο στο οποίο η πιθανότητα εμφάνισης ενός κωδικονίου μεταβάλλεται συναρτήσει του συνολικού GC% των κωδικών περιοχών, δεδομένων των πολώσεων της σύστασης εντός κάθε ομάδας συνώνυμων κωδικονίων. Παρότι οι παράμετροι που ενσωματώνουμε είναι συμμετρικές ως προς τους δύο κλώνους, το μοντέλο μας αναπαράγει χαρακτηριστικά πρότυπα συσχετίσεων μεταξύ των αποκλίσεων και της χρήσης κωδικονίων, όπως αυτά εντοπίζονται στα βακτηριακά γονιδιώματα. Συνεπώς, προκύπτει ότι οι μεταλλακτικές πιέσεις που κατευθύνουν την σύσταση των κωδικών περιοχών προς ένα συγκεκριμένο GC περιεχόμενο, αν και είναι *per se* συμμετρικές ως προς τους αντιστρόφως συμπληρωματικούς κλώνους, λόγω της σύζευξής τους με την δομή του γενετικού κώδικα μπορούν να επάγουν χαρακτηριστικές ασυμμετρίες κατά μήκος των CDSs. Είναι λοιπόν η ίδια η δομή του γενετικού κώδικα που επιβάλλει ασύμμετρα πρότυπα υποκατάστασης, ακόμα και όταν η ταυτότητα του κωδικοποιούμενου αμινοξέως δεν λαμβάνεται υπόψιν. Το γεγονός αυτό οδηγεί σε μία ασύμμετρη κατανομή αναφοράς (baseline) των βάσεων στις διάφορες θέσεις των κωδικονίων, η οποία θα πρέπει να συνυπολογίζεται προκειμένου να εκτιμηθεί ο αναμενόμενος αριθμός υποκαταστάσεων, και εξ αυτού η επίδραση που έχουν οι κατευθύνουσες μεταλλάξεις (directional mutation) και η αρνητική επιλογή (purifying selection) στην σύσταση των κωδικών περιοχών.

Στην ασύμμετρη εξέλιξη του DNA εμπλέκονται μοριακοί μηχανισμοί που δρουν με διακριτό τρόπο κατά μήκος του κάθε κλώνου. Ο αναδιπλασιασμός του γενετικού υλικού επάγει μεταλλακτικές πολώσεις που ανιχνεύονται στην κλίμακα ολόκληρου του χρωμοσώματος. Έτσι, οι δύο κλώνοι της αντιγραφής υφίστανται ασύμμετρες υποκαταστάσεις. Ωστόσο, παλαιότερες μελέτες υποστηρίζουν ότι οι ασυμμετρίες αυτές οργανώνονται σε πρότυπα καθολικά απαντώμενα στους διάφορους οργανισμούς, τα οποία δεν συσχετίζονται με τις ισομορφές της α-υπομονάδας της PolIII, που καταλύει τον πολυμερισμό των νεοσυντιθέμενων αλυσίδων. Αντίθετα, η παρούσα μελέτη, αναλύοντας την κατανομή των αποκλίσεων μεταξύ διαφορετικών ειδών, αποκαλύπτει ισχυρή συσχέτιση ανάμεσα στα πρότυπα των αποκλίσεων και τις ισομορφές της α-υπομονάδας. Το γεγονός αυτό αποδίδεται στις εγγενείς ασυμμετρίες του καταλυτικού κέντρου της PolIII, οι οποίες διαφοροποιούν την ενεργότητα ενσωμάτωσης των νουκλεοτιδίων και την επιδιορθωτική ενεργότητα της α-υπομονάδας κατά μήκος του οδηγού και του συνοδού κλώνου. Τα αποτελέσματά μας δείχνουν, με υψηλό βαθμό στατιστικής σημαντικότητας, ότι οι διαφορετικές ισομορφές του εν λόγω μορίου οδηγούν σε ασύμμετρα πρότυπα υποκαταστάσεων, τα οποία είναι πλαίσιο-εξαρτημένα (context-dependent).

Επίσης, μελετάμε το αποτύπωμα ποικίλων επιδιορθωτικών μονοπατιών στις ασυμμετρίες του γενετικού υλικού. Η συζευγμένη με τη μεταγραφή επιδιόρθωση (TCR) πολώνει τους ρυθμούς υποκατάστασης προς τον κωδικό κλώνο των γονιδίων, γεγονός που ανιχνεύεται επιτυχώς μέσω των αποκλίσεων της νουκλεοτιδικής τους σύστασης. Επεκτείνουμε την σχετική ανάλυση για να αξιολογήσουμε ενδεχόμενες ασυμμετρίες που επάγουν επιδιορθωτικά μονοπάτια, μη-συζευγμένα με τη μεταγραφή. Έτσι, εξάγουμε ενδιαφέροντα συμπεράσματα για μοριακούς μηχανισμούς των οποίων η δράση δεν έχει εξακριβωθεί εάν είναι πολωμένη προς έναν από τους δύο κλώνους του DNA. Τα σχετικά αποτελέσματα μπορούν να αξιοποιηθούν ως πρώτες ενδείξεις ώστε να κατευθύνουν την πειραματική διερεύνηση της ειδικής ανά κλώνο δράσης συγκεκριμένων ενζύμων και μοριακών μονοπατιών.

# **5.** ΒΙΒΛΙΟΓΡΑΦΙΑ


Aravind L, Koonin E V. 2001. The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. Genome Biol. 2:RESEARCH0007.

Arndt PF, Hwa T. 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. Bioinformatics 21:2322–2328.

Baisnée P-F, Hampson S, Baldi P. 2002. Why are complementary DNA strands symmetric? Bioinformatics 18:1021–1033.

Baker TA, Wickner SH. 1992. Genetics and enzymology of DNA replication in Escherichia coli. Annu. Rev. Genet. 26:447–477.

Barral P J, Cantini L, Hasmy A, Jiménez J, Marcano A. 2005. Correlation between strand asymmetry and phylogeny in mitochondrial DNA. J. Theor. Biol. 236:422–426.

Beaven, G. H., Holiday, E. R., & Johnson EA. 1955. The Nucleic Acids, ed. by E. Chargaff and JN Davidson.

Beletskii A, Bhagwat AS. 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 93:13919–13924.

Beletskii A, Bhagwat AS. 1998. Correlation between transcription and C to T mutations in the non-transcribed DNA strand. Biol. Chem. 379:549–551.

Beletskii A, Grigoriev A, Joyce S, Bhagwat AS. 2000. Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. J. Mol. Biol. 300:1057–1065.

Bell SJ, Forsdyke DR. 1999. Accounting units in DNA. J. Theor. Biol. 197:51–61.

Bernardi G, Bernardi G. 1985. Codon usage and genome composition. J. Mol. Evol. 22:363–365.

Beutler E, Gelbart T, Han JH, Koziol JA, Beutler B. 1989. Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc. Natl. Acad. Sci. U. S. A. 86:192–196.

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et al. 1997. The complete genome sequence of Escherichia coli K-12. Science 277:1453–1462.

Bockrath R, Barlow A, Engstrom J. 1987. Mutation frequency decline in Escherichia coli B/r after mutagenesis with ethyl methanesulfonate. Mutat. Res. 183:241–247.

Bockrath R, Cheung MK. 1973. The role of nutrient broth supplementation in UV mutagenesis of E. coli. Mutat. Res. 19:23–32.

Bockrath RC, Palmer JE. 1977. Differential repair of premutational UV-lesions at tRNA genes in E. coli. Mol. Gen. Genet. 156:133–140.

Bohlin J, Skjerve E. 2009. Examination of genome homogeneity in prokaryotes using genomic signatures. PLoS One 4:e8113.

Bohlin J, Skjerve E, Ussery DW. 2009. Analysis of genomic signatures in prokaryotes using multinomial regression and hierarchical clustering. BMC Genomics 10:487.

Bohr VA, Smith CA, Okumoto DS, Hanawalt PC. 1985. DNA repair in an active gene: removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than in the genome overall. Cell 40:359–369.

Boshoff HIM, Reed MB, Barry CE, Mizrahi V. 2003. DnaE2 polymerase contributes to in vivo survival and the emergence of drug resistance in Mycobacterium tuberculosis. Cell 113:183–193.

Brewer BJ. 1988. When polymerases collide: replication and the transcriptional organization of the E. coli chromosome. Cell 53:679–686.

Bruck I, O'Donnell M. 2000. The DNA Replication Machine of a Gram-positive Organism. J. Biol. Chem. 275:28971–28983.

Bulmer M. 1991a. Strand symmetry of mutation rates in the beta-globin region. J. Mol. Evol. 33:305–310.

Bulmer M. 1991b. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907.

Burge C, Campbell AM, Karlin S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. U. S. A. 89:1358–1362.

Campbell A, Mrázek J, Karlin S. 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc. Natl. Acad. Sci. U. S. A. 96:9184–9189.

Carvalho FM, Fonseca MM, Batistuzzo De Medeiros S, Scortecci KC, Blaha CAG, Agnez-Lima LF. 2005. DNA repair in reduced genome: the Mycoplasma model. Gene 360:111–119.

Cedar H, Razin A. 1990. DNA methylation and development. Biochim. Biophys. Acta 1049:1–8.

Chamary J-V, Hurst LD. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Mol. Biol. Evol. 21:1014–1023.

Chargaff E. 1951. Structure and function of nucleic acids as cell constituents. Fed. Proc. 10:654–659.

Chargaff E. 1979. How genetics got a chemical education. Ann. N. Y. Acad. Sci. 325:344–360.

Charneski CA, Honti F, Bryant JM, Hurst LD, Feil EJ. 2011. Atypical AT skew in Firmicute genomes results from selection and not from mutation. PLoS Genet. 7:e1002283.

Constantin N, Dzantiev L, Kadyrov FA, Modrich P. 2005. Human mismatch repair: reconstitution of a nick-directed bidirectional reaction. J. Biol. Chem. 280:39752–39761.

Copley SD, Smith E, Morowitz HJ. 2005. A mechanism for the association of amino acids with their codons and the origin of the genetic code. Proc. Natl. Acad. Sci. U. S. A. 102:4442–4447.

Cornet F, Louarn J, Patte J, Louarn JM. 1996. Restriction of the activity

of the recombination site dif to a small zone of the Escherichia coli chromosome. Genes Dev. 10:1152–1161.

Cox EC, Yanofsky C. 1967. Altered base ratios in the DNA of an Escherichia coli mutator strain. Proc. Natl. Acad. Sci. U. S. A. 58:1895–1902.

Cozzarelli NR, Low RL. 1973. Mutational alteration of Bacillus subtilis DNA polymerase 3 to hydroxyphenylazopyrimidine resistance: polymerase 3 is necessary for DNA replication. Biochem. Biophys. Res. Commun. 51:151–157.

Danchin A. 2003. Genomes and evolution. Curr. Issues Mol. Biol. 5:37–42.

Dawid IB. 1974. 5-methylcytidylic acid: absence from mitochondrial DNA of frogs and HeLa cells. Science 184:80–81.

Deaconescu AM. 2013. RNA polymerase between lesion bypass and DNA repair. Cell. Mol. Life Sci. 70:4495–4509.

Delaney JC, Essigmann JM. 2004. Mutagenesis, genotoxicity, and repair of 1-methyladenine, 3-alkylcytosines, 1-methylguanine, and 3-methylthymine in alkB Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 101:14051–14056.

Denamur E, Lecointre G, Darlu P, Tenaillon O, Acquaviva C, Sayada C, Sunjevaric I, Rothstein R, Elion J, Taddei F, et al. 2000. Evolutionary Implications of the Frequent Horizontal Transfer of Mismatch Repair Genes. Cell 103:711–721.

Denamur E, Matic I. 2006. Evolution of mutation rates in bacteria. Mol. Microbiol. 60:820–827.

Dervyn E, Suski C, Daniel R, Bruand C, Chapuis J, Errington J, Jannière L, Ehrlich SD. 2001. Two essential DNA polymerases at the bacterial replication fork. Science 294:1716–1719.

Dizdaroglu M. 2005. Base-excision repair of oxidative DNA damage by DNA glycosylases. Mutat. Res. 591:45–59.

D'Onofrio G, Jabbari K, Musto H, Bernardi G. 1999. The correlation of protein hydropathy with the base composition of coding sequences. Gene 238:3–14.

Du J, Yuan Z, Ma Z, Song J, Xie X, Chen Y. 2014. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. Mol. Biosyst. 10:2441–2447.

Duncan BK, Weiss B. 1982. Specific mutator effects of ung (uracil-DNA glycosylase) mutations in Escherichia coli. J. Bacteriol. 151:750–755.

Duppatla V, Bodda C, Urbanke C, Friedhoff P, Rao DN. 2009. The C-terminal domain is sufficient for endonuclease activity of Neisseria gonorrhoeae MutL. Biochem. J. 423:265–277.

Echols H, Goodman MF. 1991. Fidelity mechanisms in DNA replication. Annu. Rev. Biochem. 60:477–511.

Engstrom J, Larsen S, Rogers S, Bockrath R. 1984. UV-mutagenesis at a cloned target sequence: converted suppressor mutation is insensitive to mutation frequency decline regardless of the gene orientation. Mutat. Res. 132:143–152.

Fang WH, Modrich P. 1993. Human strand-specific mismatch repair occurs by a

bidirectional mechanism similar to that of the bacterial reaction. J. Biol. Chem. 268:11838–11844.

Fersht AR, Knill-Jones JW. 1981. DNA polymerase accuracy and spontaneous mutation rates: frequencies of purine.purine, purine.pyrimidine, and pyrimidine.pyrimidine mismatches during DNA replication. Proc. Natl. Acad. Sci. U. S. A. 78:4251–4255.

Fickett JW, Torney DC, Wolf DR. 1992. Base compositional structure of genomes. Genomics 13:1056–1064.

Fijalkowska IJ, Jonczyk P, Tkaczyk MM, Bialoskorska M, Schaaper RM. 1998. Unequal fidelity of leading strand and lagging strand DNA replication on the Escherichia coli chromosome. Proc. Natl. Acad. Sci. U. S. A. 95:10020–10025.

Forsdyke DR. 1995. Relative roles of primary sequence and (G + C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. J. Mol. Evol. 41:573–581.

Francino MP, Chao L, Riley MA, Ochman H. 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science 272:107–109.

Francino MP, Ochman H. 1997. Strand asymmetries in DNA evolution. Trends Genet. 13:240–245.

Francino MP, Ochman H. 2001. Deamination as the basis of strand-asymmetric evolution in transcribed Escherichia coli sequences. Mol. Biol. Evol. 18:1147–1150.

Frank AC, Lobry JR. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238:65–77.

Fraser CM. 1998. Complete Genome Sequence of Treponema pallidum, the Syphilis Spirochete. Science (80-. ). 281:375–388.

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, et al. 1997. Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature 390:580–586.

Frederico LA, Kunkel TA, Shaw BR. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry 29:2532–2537.

Freeman JM. 1998. Patterns of Genome Organization in Bacteria. Science (80-. ). 279:1827a – 1827.

Fukui K. 2010. DNA mismatch repair in eukaryotes and bacteria. J. Nucleic Acids 2010.

Fukui K, Nishida M, Nakagawa N, Masui R, Kuramitsu S. 2008. Bound nucleotide controls the endonuclease activity of mismatch repair enzyme MutL. J. Biol. Chem. 283:12136–12145.

Galhardo RS, Rocha RP, Marques M V, Menck CFM. 2005. An SOS-regulated operon involved in damage-inducible mutagenesis in Caulobacter crescentus. Nucleic Acids Res. 33:2603–2614.

Ganesan A, Spivak G, Hanawalt PC. 2012. Transcription-coupled DNA repair in prokaryotes. Prog. Mol. Biol. Transl. Sci. 110:25–40

Gefter ML, Hirota Y, Kornberg T, Wechsler JA, Barnoux C. 1971. Analysis of

DNA polymerases II and 3 in mutants of Escherichia coli thermosensitive for DNA synthesis. Proc. Natl. Acad. Sci. U. S. A. 68:3150–3153.

Goodarzi H, Torabi N, Najafabadi HS, Archetti M. 2008. Amino acid and codon usage profiles: adaptive changes in the frequency of amino acids and codons. Gene 407:30–41.

Goosen N, Moolenaar GF. 2008. Repair of UV damage in bacteria. DNA Repair (Amst). 7:353–379.

Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10:7055–7074.

Green P, Ewing B, Miller W, Thomas PJ, Green ED. 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet. 33:514–517.

Grigoriev A. 1998. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. 26:2286–2290.

Gu X, Hewett-Emmett D, Li WH. 1998. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. Genetica 102-103:383–391.

Hampson S, Baldi P, Kibler D, Sandmeyer SB. 2000. Analysis of yeast's ORF upstream regions by parallel processing, microarrays, and computational methods. Proc. Int. Conf. Intell. Syst. Mol. Biol. 8:190–201.

Hanawalt PC. 1991. Heterogeneity of DNA repair at the gene level. Mutat. Res. Mol. Mech. Mutagen. 247:203–211.

Heinze RJ, Giron-Monzon L, Solovyova A, Elliot SL, Geisler S, Cupples CG, Connolly BA, Friedhoff P. 2009. Physical and functional interactions between Escherichia coli MutL and the Vsr repair endonuclease. Nucleic Acids Res. 37:4453–4463.

Housby JN, Southern EM. 1998. Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. Nucleic Acids Res. 26:4259–4266.

Huffman JL, Sundheim O, Tainer JA. 2005. DNA base damage recognition and removal: new twists and grooves. Mutat. Res. 577:55–76.

Hu J, Zhao X, Zhang Z, Yu J. 2007. Compositional dynamics of guanine and cytosine content in prokaryotic genomes. Res. Microbiol. 158:363–370.

Hutchinson F. 1996. Escherichia coli and Salmonella. Cellular and Molecular BiologyNo Title. In: Neidhardt F., editor. 2nd ed. Washington, DC: ASM press. p. 749–763.

Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. 151:389–409.

Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer R. J. Mol. Biol. 158:573–597.

Inoue R, Kaito C, Tanabe M, Kamura K, Akimitsu N SK. 2001. Genetic identification of two distinct DNA polymerases, DnaE and PolC, that are

essential for chromosomal DNA replication in Staphylococcus aureus. Mol Genet Genomics 266:564–571.

Iwaki T, Kawamura A, Ishino Y, Kohno K, Kano Y, Goshima N, Yara M, Furusawa M, Doi H, Imamoto F. 1996. Preferential replication-dependent mutagenesis in the lagging DNA strand in Escherichia coli. Mol. Gen. Genet. 251:657–664.

Jeltsch A, Pingoud A. 1996. Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. J. Mol. Evol. 42:91–96.

Josse J, Kaiser AD, Kornberg A. 1961. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J. Biol. Chem. 236:864–875.

Jukes TH, Osawa S, Muto A. 1987. Divergence and directional mutation pressures. Nature 325:668.

Kadyrov FA, Dzantiev L, Constantin N, Modrich P. 2006. Endonucleolytic function of MutLalpha in human mismatch repair. Cell 126:297–308.

Karkas JD, Rudner R, Chargaff E. 1968. Seapration of B. subtilis DNA into complementary strands. II. Template functions and composition as determined by transcription with RNA polymerase. Proc. Natl. Acad. Sci. U. S. A. 60:915–920.

Karkas JD, Rudner R, Chargaff E. 1970. Template properties of complementary fractions of denatured microbial deoxyribonucleic acids. Proc. Natl. Acad. Sci. U. S. A. 65:1049–1056.

Karlin S. 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. Curr. Opin. Microbiol. 1:598–610.

Karlin S, Blaisdell BE, Bucher P. 1992. Quantile distributions of amino acid usage in protein classes. "Protein Eng. Des. Sel. 5:729–738.

Karlin S, Burge C. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11:283–290.

Karlin S, Cardon LR. 1994. Computational DNA sequence analysis. Annu. Rev. Microbiol. 48:619–654.

Karlin S, Ladunga I. 1994. Comparisons of eukaryotic genomic sequences. Proc. Natl. Acad. Sci. U. S. A. 91:12832–12836.

Karlin S, Ladunga I, Blaisdell BE. 1994. Heterogeneity of genomes: measures and values. Proc. Natl. Acad. Sci. U. S. A. 91:12837–12841.

Karlin S, Mocarski ES, Schachtel GA. 1994. Molecular evolution of herpesviruses: genomic and protein sequence comparisons. J. Virol. 68:1886–1902.

Karlin S, Mrázek J. 1997. Compositional differences within and between eukaryotic genomes. Proc. Natl. Acad. Sci. U. S. A. 94:10227–10232.

Karlin S, Mrázek J, Campbell AM. 1996. Frequent oligonucleotides and peptides of the Haemophilus influenzae genome. Nucleic Acids Res. 24:4263–4272.

Kelman Z, O'Donnell M. 1995. DNA polymerase III holoenzyme: structure and function of a chromosomal replicating machine. Annu. Rev. Biochem. 64:171–200.

Klasson L, Andersson SGE. 2006. Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways. Mol. Biol. Evol. 23:1031–1039.

Knight RD, Freeland SJ, Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. 2:RESEARCH0010.

Koonin E V, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. Nat. Rev. Genet. 11:487–498.

Kozhukhin CG, Pevzner PA. 1991. Genome inhomogeneity is determined mainly by WW and SS dinucleotides. Bioinformatics 7:39–49.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science 324:255–258.

Kullback S, Leibler RA. 1951. On Information and Sufficiency. Ann. Math. Stat. 22:79–86.

Kunkel TA. 1992. DNA replication fidelity. J. Biol. Chem. 267:18251–18254.

Kuzminov A. 1995. Collapse and repair of replication forks in Escherichia coli. Mol. Microbiol. 16:373–384.

Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res. 27:1642–1649.

Lagunez-Otero J, Trifonov EN. 1992. mRNA periodical infrastructure complementary to the proof-reading site in the ribosome. J. Biomol. Struct. Dyn. 10:455–464.

Längle-Rouault F, Maenhaut-Michel G, Radman M. 1987. GATC sequences, DNA nicks and the MutH function in Escherichia coli mismatch repair. EMBO J. 6:1121–1127.

Lindahl T. 1993. Instability and decay of the primary structure of DNA. Nature 362:709–715.

Lindahl T. 2001. Keynote: past, present, and future aspects of base excision repair. Prog. Nucleic Acid Res. Mol. Biol. 68:xvii – xxx.

Lindahl T, Nyberg B. 1974. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. Biochemistry 13:3405–3410.

Lind PA, Andersson DI. 2008. Whole-genome mutational biases in bacteria. Proc. Natl. Acad. Sci. U. S. A. 105:17878–17883.

Lin HJ, Chargaff E. 1967. On the denaturation of deoxyribonucleic acid. II. Effects of concentration. Biochim. Biophys. Acta 145:398–409.

Lobry JR. 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. 40:326–330.

Lobry JR. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13:660–665.

Lobry JR. 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205:309–316.

Lobry JR, Gautier C. 1994. Hydrophobicity, expressivity and aromaticity are

the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res. 22:3174–3180.

Lobry JR, Lobry C. 1999. Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. Mol. Biol. Evol. 16:719–723.

Lobry JR, Sueoka N. 2002. Asymmetric directional mutation pressures in bacteria. Genome Biol. 3:RESEARCH0058.

Low RL, Rashbaum SA, Cozzarelli NR. 1976. Purification and characterization of DNA polymerase III from Bacillus subtilis. J. Biol. Chem. 251:1311–1325.

Magasanik B, Chargaff E. 1989. Studies on the structure of ribonucleic acids. 1951. Biochim. Biophys. Acta 1000:17–33.

Maki H, Maki S, Kornberg A. 1988. DNA Polymerase III holoenzyme of Escherichia coli. IV. The holoenzyme is an asymmetric dimer with twin active sites. J. Biol. Chem. 263:6570–6578.

Maki H, Sekiguchi M. 1992. MutT protein specifically hydrolyses a potent mutagenic substrate for DNA synthesis. Nature 355:273–275.

Mao X, Zhang H, Yin Y, Xu Y. 2012. The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. Nucleic Acids Res. 40:8210–8218.

Marians KJ. 1992. Prokaryotic DNA replication. Annu. Rev. Biochem. 61:673–719.

Marín A, Xia X. 2008. GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. J. Theor. Biol. 253:508–513.

Mascher M, Schubert I, Scholz U, Friedel S. 2013. Patterns of nucleotide asymmetries in plant and animal genomes. Biosystems. 111:181–189.

Mauris J, Evans TC. 2009. Adenosine triphosphate stimulates Aquifex aeolicus MutL endonuclease activity. PLoS One 4:e7175.

McInerney JO. 1998. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proc. Natl. Acad. Sci. U. S. A. 95:10698–10703.

McLean MJ, Wolfe KH, Devine KM. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. 47:691–696.

Mellon I, Hanawalt PC. 1989. Induction of the Escherichia coli lactose operon selectively increases repair of its transcribed DNA strand. Nature 342:95–98.

Mendelman L V, Petruska J, Goodman MF. 1990. Base mispair extension kinetics. Comparison of DNA polymerase alpha and reverse transcriptase. J. Biol. Chem. 265:2338–2346.

Mielecki D, Grzesiuk E. 2014. Ada response - a strategy for repair of alkylated DNA in bacteria. FEMS Microbiol. Lett. 355:1–11.

Mielecki D, Saumaa S, Wrzesiński M, Maciejewska AM, Żuchniewicz K, Sikora A, Piwowarski J, Nieminuszczy J, Kivisaar M, Grzesiuk E. 2013. Pseudomonas putida AlkA and AlkB proteins comprise different defense

systems for the repair of alkylation damage to DNA - in vivo, in vitro, and in silico studies. PLoS One 8:e76198.

Mitchell D, Bridge R. 2006. A test of Chargaff's second rule. Biochem. Biophys. Res. Commun. 340:90–94.

Modrich P. 1989. Methyl-directed DNA mismatch correction. J. Biol. Chem. 264:6597–6600.

Mohr S, Freeman J, Plasterer T, Smith T. 1999. Patterns of Mitochondrial DNA Strand Asymmetry Correlate with Phylogeny. Biol. Bull. 196:411.

Mokkapati SK, Fernández de Henestrosa AR, Bhagwat AS. 2001. Escherichia coli DNA glycosylase Mug: a growth-regulated enzyme required for mutation avoidance in stationary-phase cells. Mol. Microbiol. 41:1101–1111.

Morita R, Nakane S, Shimada A, Inoue M, Iino H, Wakamatsu T, Fukui K, Nakagawa N, Masui R, Kuramitsu S. 2010. Molecular mechanisms of the whole DNA repair system: a comparison of bacterial and eukaryotic systems. J. Nucleic Acids 2010:179594.

Morton RA, Morton BR. 2007. Separating the effects of mutation and selection in producing DNA skew in bacterial chromosomes. BMC Genomics 8:369.

Mrázek J, Karlin S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. U. S. A. 95:3720–3725.

Muto A, Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. U. S. A. 84:166–169.

Nakamura Y, Gojobori T, Ikemura T. 1999. Codon usage tabulated from the international DNA sequence databases; its status 1999. Nucleic Acids Res. 27:292–292.

Nass MM. 1973. Differential methylation of mitochondrial and nuclear DNA in cultured mouse, hamster and virus-transformed hamster cells. In vivo and in vitro methylation. J. Mol. Biol. 80:155–175.

Necşulea A, Lobry JR. 2007. A new method for assessing the effect of replication on DNA base composition asymmetry. Mol. Biol. Evol. 24:2169–2179.

Nieminuszczy J, Grzesiuk E. 2007. Bacterial DNA repair genes and their eukaryotic homologues: 3. AlkB dioxygenase and Ada methyltransferase in the direct repair of alkylated DNA. Acta Biochim. Pol. 54:459–468.

Nikolaou C, Almirantis Y. 2005. A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species. Nucleic Acids Res. 33:6816–6822.

Nikolaou C, Almirantis Y. 2006. Deviations from Chargaff's second parity rule in organellar DNA Insights into the evolution of organellar genomes. Gene 381:34–41.

Nomura M, Morgan EA. 1977. Genetics of bacterial ribosomes. Annu. Rev. Genet. 11:297–347.

Nussinov R. 1980. Some rules in the ordering of nucleotides in the DNA. Nucleic Acids Res. 8:4545–4562.

Nussinov R. 1981. Nearest neighbor nucleotide patterns. Structural and

biological implications. J. Biol. Chem. 256:8458–8462.

Nussinov R. 1984a. Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res. 12:1749–1763.

Nussinov R. 1984b. Strong doublet preferences in nucleotide sequences and DNA geometry. J. Mol. Evol. 20:111–119.

Nye TMW, Liò P, Gilks WR. 2006. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics 22:117–119.

Ohno S. 1988. Universal rule for coding sequence construction: TA/CG deficiency-TG/CT excess. Proc. Natl. Acad. Sci. U. S. A. 85:9630–9634.

Okazaki R, Okazaki T, Sakabe K, Sugimoto K, Sugino A. 1968. Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. Proc. Natl. Acad. Sci. U. S. A. 59:598–605.

Oller AR, Fijalkowska IJ, Dunn RL, Schaaper RM. 1992. Transcription-repair coupling determines the strandedness of ultraviolet mutagenesis in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 89:11036–11040.

Palidwor GA, Perkins TJ, Xia X. 2010. A general model of codon bias due to GC mutational bias. PLoS One 5:e13431.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–290.

Van Passel MWJ, Kuramae EE, Luyf ACM, Bart A, Boekhout T. 2006. The reach of the genome signature in prokaryotes. BMC Evol. Biol. 6:84.

Peng W, Shaw BR. 1996. Accelerated deamination of cytosine residues in UV-induced cyclobutane pyrimidine dimers leads to CC-->TT transitions. Biochemistry 35:10172–10181.

Perrière G, Lobry JR, Thioulouse J. 1996. Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences. Bioinformatics 12:519–524.

Phillippy AM, Mason JA, Ayanbule K, Sommer DD, Taviani E, Huq A, Colwell RR, Knight IT, Salzberg SL. 2007. Comprehensive DNA signature discovery and validation. PLoS Comput. Biol. 3:e98.

Podani J. 2013. Tree thinking, time and topology: comments on the interpretation of tree diagrams in evolutionary/phylogenetic systematics. Cladistics 29:315–327.

Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldón T, Rattei T, Creevey C, Kuhn M, et al. 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res. 42:D231–D239.

Prabhu V V. 1993. Symmetry observations in long nucleotide sequences. Nucleic Acids Res. 21:2797–2800.

Qu H, Wu H, Zhang T, Zhang Z, Hu S, Yu J. 2010. Nucleotide compositional asymmetry between the leading and lagging strands of eubacterial genomes. Res. Microbiol. 161:838–846.

Radman M. 1998. DNA replication: one strand may be more equal. Proc. Natl. Acad. Sci. U. S. A. 95:9718–9719.

Rasmussen LJ, Samson L. 1996. The Escherichia coli MutS DNA mismatch binding protein specifically binds O(6)-methylguanine DNA lesions. Carcinogenesis 17:2085–2088.

Ravishankar R, Bidya Sagar M, Roy S, Purnapatre K, Handa P, Varshney U, Vijayan M. 1998. X-ray analysis of a complex of Escherichia coli uracil DNA glycosylase (EcUDG) with a proteinaceous inhibitor. The structure elucidation of a prokaryotic UDG. Nucleic Acids Res. 26:4880–4887.

Rebeck GW, Samson L. 1991. Increased spontaneous mutation and alkylation sensitivity of Escherichia coli strains lacking the ogt O6-methylguanine DNA repair methyltransferase. J. Bacteriol. 173:2068–2076.

Resende BC, Rebelato AB, D'Afonseca V, Santos AR, Stutzman T, Azevedo VA, Santos LL, Miyoshi A, Lopes DO. 2011. DNA repair in Corynebacterium model. Gene 482:1–7.

Richardson CC, Inman RB, Kornberg A. 1964. Enzymic synthesis of deoxyribonucleic acid. J. Mol. Biol. 9:46–IN4.

Rocha EP. 2002. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol. 10:393–395.

Rocha EPC. 2004. The replication-related organization of bacterial genomes. Microbiology 150:1609–1627.

Rocha EPC. 2008. The organization of the bacterial genome. Annu. Rev. Genet. 42:211–233.

Rocha EPC, Cornet E, Michel B. 2005. Comparative and Evolutionary Analysis of the Bacterial Homologous Recombination Systems. PLoS Genet. 1:e15.

Rocha EPC, Danchin A. 2003. Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res. 31:6570–6577.

Rocha EP, Danchin A. 2001. Ongoing evolution of strand composition in bacterial genomes. Mol. Biol. Evol. 18:1789–1799.

Rocha EP, Danchin A, Viari A. 1999. Universal replication biases in bacteria. Mol. Microbiol. 32:11–16.

Rocha EP, Viari A, Danchin A. 1998. Oligonucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons. Nucleic Acids Res. 26:2971–2980.

Rosche WA, Trinh TQ, Sinden RR. 1995. Differential DNA secondary structure-mediated deletion mutation in the leading and lagging strands. J. Bacteriol. 177:4385–4391.

Rudner R, Karkas JD, Chargaff E. 1968a. Separation of B. subtilis DNA into complementary strands, I. Biological properties. Proc. Natl. Acad. Sci. U. S. A. 60:630–635.

Rudner R, Karkas JD, Chargaff E. 1968b. Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. Proc. Natl. Acad. Sci. U. S. A. 60:921–922.

Russell GJ, Subak-Sharpe JH. 1977. Similarity of the general designs of protochordates and invertebrates. Nature 266:533–536.

Russell GJ, Walker PM, Elton RA, Subak-Sharpe JH. 1976. Doublet frequency

analysis of fractionated vertebrate nuclear DNA. J. Mol. Biol. 108:1–23.

Saha SK, Goswami A, Dutta C. 2014. Association of purine asymmetry, strand-biased gene distribution and PolC within Firmicutes and beyond: a new appraisal. BMC Genomics 15:430.

Salzberg SL, Salzberg AJ, Kerlavage AR, Tomb JF. 1998. Skewed oligomers and origins of replication. Gene 217:57–67.

Saparbaev M, Laval J. 1998. 3,N4-ethenocytosine, a highly mutagenic adduct, is a primary substrate for Escherichia coli double-stranded uracil-DNA glycosylase and human mismatch-specific thymine-DNA glycosylase. Proc. Natl. Acad. Sci. U. S. A. 95:8508–8513.

Schaaper RM. 1993. Base selection, proofreading, and mismatch repair during DNA replication in Escherichia coli. J. Biol. Chem. 268:23762–23765.

Sekine S, Tagami S, Yokoyama S. 2012. Structural basis of transcription by bacterial and eukaryotic RNA polymerases. Curr. Opin. Struct. Biol. 22:110–118.

Selby CP, Sancar A. 1993. Molecular mechanism of transcription-repair coupling. Science 260:53–58.

Selby CP, Sancar A. 1994. Mechanisms of transcription-repair coupling and mutation frequency decline. Microbiol. Rev. 58:317–329.

Selby CP, Witkin EM, Sancar A. 1991. Escherichia coli mfd mutant deficient in “mutation frequency decline” lacks strand-specific repair: in vitro complementation with purified coupling factor. Proc. Natl. Acad. Sci. U. S. A. 88:11574–11578.

Selby C, Sancar A. 1993. Molecular mechanism of transcription-repair coupling. Science 260:53–58.

Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 365:1203–1212.

Shioiri C, Takahata N. 2001. Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. J. Mol. Evol. 53:364–376.

Shock LS, Thakkar P V, Peterson EJ, Moran RG, Taylor SM. 2011. DNA methyltransferase 1, cytosine methylation, and cytosine hydroxymethylation in mammalian mitochondria. Proc. Natl. Acad. Sci. U. S. A. 108:3630–3635.

Singer GA, Hickey DA. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol. Biol. Evol. 17:1581–1588.

Smithies O, Engels WR, Devereux JR, Slightom JL, Shen S. 1981. Base substitutions, length differences and DNA strand asymmetries in the human G gamma and A gamma fetal globin gene region. Cell 26:345–353.

Sorimachi K, Okayasu T. 2008. Codon evolution is governed by linear formulas. Amino Acids 34:661–668.

Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in Escherichia coli: selection for translational accuracy. Mol. Biol. Evol. 24:374–381.

Sueoka N. 1961. Compositional correlation between deoxyribonucleic acid and

protein. Cold Spring Harb. Symp. Quant. Biol. 26:35–43.

Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. U. S. A. 48:582–592.

Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. U. S. A. 85:2653–2657.

Sueoka N. 1992. Directional mutation pressure, selective constraints, and genetic equilibria. J. Mol. Evol. 34:95–114.

Sueoka N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. 40:318–325.

Sueoka N. 1999. Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C. J. Mol. Evol. 49:49–62.

Suzuki T, Harashima H, Kamiya H. 2010. Effects of base excision repair proteins on mutagenesis by 8-oxo-7,8-dihydroguanine (8-hydroxyguanine) paired with cytosine and adenine. DNA Repair (Amst). 9:542–550.

Swartz MN, Trautner TA, Kornberg A. 1962. Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. J. Biol. Chem. 237:1961–1967.

Szybalski W, Kubinski H, Sheldrick P. 1966. Pyrimidine Clusters on the Transcribing Strand of DNA and Their Possible Role in the Initiation of RNA Synthesis. Cold Spring Harb. Symp. Quant. Biol. 31:123–127.

Taira K, Nakamura S, Nakano K, Maehara D, Okamoto K, Arimoto S, Loakes D, Worth L, Schaaper RM, Seio K, et al. 2008. Binding of MutS protein to oligonucleotides containing a methylated or an ethylated guanine residue, and correlation with mutation frequency. Mutat. Res. 640:107–112.

Tazi J, Bird A. 1990. Alternative chromatin structure at CpG islands. Cell 60:909–920.

Tippin B, Pham P, Goodman MF. 2004. Error-prone replication for better or worse. Trends Microbiol. 12:288–295.

Trifonov EN. 1987. Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. J. Mol. Biol. 194:643–652.

Trinh TQ, Sinden RR. 1991. Preferential DNA secondary structure mutagenesis in the lagging strand of replication in E. coli. Nature 352:544–547.

Veaute X, Fuchs RP. 1993. Greater susceptibility to mutations in lagging strand of DNA replication in Escherichia coli than in leading strand. Science 261:598–600.

Watson JD, Crick FHC. 1953. Molecular Structure of Nucleic Acids; A Structure for Deoxyribose Nucleic Acid. Nature 171, 737–738

Wilquet V, Van de Casteele M. 1999. The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. Res. Microbiol. 150:21–32.

Wood ML, Dizdaroglu M, Gajewski E, Essigmann JM. 1990. Mechanistic studies of ionizing radiation and oxidative mutagenesis: genetic effects of a single 8-hydroxyguanine (7-hydro-8-oxoguanine) residue inserted at a

unique site in a viral genome. Biochemistry 29:7024–7032.

Worning P, Jensen LJ, Hallin PF, Staerfeldt H-H, Ussery DW. 2006. Origin of replication in circular prokaryotic chromosomes. Environ. Microbiol. 8:353–361.

Wu CI, Maeda N. 1987. Inequality in mutation rates of the two strands of DNA. Nature 327:169–170.

Wu H, Zhang Z, Hu S, Yu J. 2012. On the molecular mechanism of GC content variation among eubacterial genomes. Biol. Direct 7:2.

Wyrzykowski J, Volkert MR. 2003. The Escherichia coli methyl-directed mismatch repair system repairs base pairs containing oxidative lesions. J. Bacteriol. 185:1701–1704.

Xiao G, Tordova M, Jagadeesh J, Drohat AC, Stivers JT, Gilliland GL. 1999. Crystal structure of Escherichia coli uracil DNA glycosylase and its complexes with uracil and glycerol: structure and glycosylase mechanism revisited. Proteins 35:13–24.

Xiao J-F, Yu J. 2007. A scenario on the stepwise evolution of the genetic code. Genomics. Proteomics Bioinformatics 5:143–151.

Xia X. 1998. How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? Genetics 149:37–44.

Yoda K, Okazaki T. 1991. Specificity of recognition sequence forEscherichia coli primase. MGG Mol. Gen. Genet. 227:1–8.

Yomo T, Ohno S. 1989. Concordant evolution of coding and noncoding regions of DNA made possible by the universal rule of TA/CG deficiency-TG/CT excess. Proc. Natl. Acad. Sci. U. S. A. 86:8452–8456.

Yu J. 2007. A content-centric organization of the genetic code. Genomics. Proteomics Bioinformatics 5:1–6.

Yuzhakov A, Turner J, O'Donnell M. 1996. Replisome Assembly Reveals the Basis for Asymmetric Function in Leading and Lagging Strand Replication. Cell 86:877–886.

Zamenhof S, Brawerman G, Chargaff E. 1952. On the desoxypentose nucleic acids from several microorganisms. Biochim. Biophys. Acta 9: 402-405.

Zeileis A, Kleiber C, Krämer W, Hornik K. 2003. Testing and dating of structural changes in practice. Comput. Stat. Data Anal. 44:109–123.

Zeileis A, Leisch F, Hornik K, Kleiber C. 2002. strucchange : An R Package for Testing for Structural Change in Linear Regression Models. J. Stat. Softw. 7:1–38.

Zeileis A, Shah A, Patnaik I. 2010. Testing, monitoring, and dating structural changes in exchange rate regimes. Comput. Stat. Data Anal. 54:1696–1706.

Zhang Z, Yu J. 2010. Modeling compositional dynamics based on GC and purine contents of protein-coding sequences. Biol. Direct 5:63.

Zhao X-Q, Hu J-F, Yu J. 2006. Comparative analysis of eubacterial DNA polymerase III alpha subunits. Genomics. Proteomics Bioinformatics 4:203–211.

Zhao X, Zhang Z, Yan J, Yu J. 2007. GC content variability of eubacteria is

governed by the pol III alpha subunit. Biochem. Biophys. Res. Commun. 356:20–25.

*Επιπλέον Εικόνες και Πίνακες*

link: http://bio.demokritos.gr/index.php?option=com_content&view=article&id=165&Itemid=167&lang=el

# ΠΑΡΑΡΤΗΜΑ

**ΠΙΝΑΚΑΣ Ι. Ποσοστιαία κατάταξη των χρωμοσωμάτων των Firmicutes, βάσει των προτύπων που εμφανίζουν τα μοντέλα γραμμικής παλινδρόμησης που περιγράφουν τις αποκλίσεις τους κατά μήκος του δημοσιευμένου κλώνου.**

| | | *σαφή* | *ασαφή* | *ισόπεδα* |
|---|---|---|---|---|
| **αποκλίσεις μόνο- και δι-νουκλεοτιδίων** | $S_{plus}^{A-T}$ | 71.88 | 23.44 | 4.69 |
| | $S_{plus}^{G-C}$ | 100 | 0 | 0 |
| | $S_{plus}^{AG-CT}$ | 98.44 | 0 | 1.56 |
| | $S_{plus}^{GA-TC}$ | 100 | 0 | 0 |
| | $S_{plus}^{GG-CC}$ | 100 | 0 | 0 |
| | $S_{plus}^{AA-TT}$ | 70.31 | 21.88 | 7.81 |
| | $S_{plus}^{AC-GT}$ | 87.5 | 4.69 | 7.81 |
| | $S_{plus}^{CA-TG}$ | 100 | 0 | 0 |
| **αποκλίσεις σταθμισμένων συχνοτήτων** | $P_{plus}^{AG-CT}$ | 64.06 | 10.94 | 25 |
| | $P_{plus}^{GA-TC}$ | 68.75 | 26.56 | 4.69 |
| | $P_{plus}^{GG-CC}$ | 60.94 | 12.5 | 26.56 |
| | $P_{plus}^{AA-TT}$ | 71.88 | 20.31 | 7.81 |
| | $P_{plus}^{AC-GT}$ | 35.94 | 50 | 14.06 |
| | $P_{plus}^{CA-TG}$ | 57.81 | 21.88 | 20.31 |

ΣΗΜΕΙΩΣΕΙΣ.– ***σαφή*** πρότυπα: τουλάχιστον ένα στατιστικώς σημαντικό σημείο μεταβολής (breakpoint) εντοπίζεται σε απόσταση από το *ori* ίση ή μικρότερη από το 5% του μήκους του χρωμοσώματος. ***ασαφή*** πρότυπα: εντοπίζεται τουλάχιστον ένα στατιστικώς σημαντικό σημείο μεταβολής (breakpoint), αλλά σε απόσταση από το *ori* μεγαλύτερη από το 5% του μήκους του χρωμοσώματος. ***ισόπεδα*** πρότυπα: δεν εντοπίζεται κανένα στατιστικώς σημαντικό σημείο μεταβολής.

Στον Πίνακα Ι λαμβάνονται υπόψιν μόνο τα χρωμοσώματα που ανήκουν στο φύλο των Firmicutes. Για τα αποτελέσματα που αντιστοιχούν στο σύνολο των χρωμοσωμάτων της συλλογής μας, βλ. Πίνακα 9.