Πανεπιστήμιο Θεσσαλίας
Τμήμα Βιοχημείας και Βιοτεχνολογίας

# Βιοπληροφορική ανάλυση Περιοχών Χαμηλής Πολυπλοκότητας σε προκαρυώτες

Ντουντούμη Χρυσούλα

Επιβλέπων καθηγητής: Αμούτζιας Γρηγόριος

University of Thessaly

Department of Biochemistry and Biotechnology

# Bioinformatic analysis of Low Complexity Regions in prokaryotes

Ntountoumi Chrysoula

Supervisor Professor: Amoutzias Grigorios

Η παρούσα διπλωματική εργασία εκπονήθηκε στο εργαστήριο Βιοπληροφορικής, του Τμήματος Βιοχημείας και Βιοτεχνολογίας (ΤΒΒ), της Σχολής Επιστημών Υγείας του Πανεπιστημίου Θεσσαλίας (Π.Θ.).

**Υπεύθυνος Καθηγητής**

**Αμούτζιας Γρηγόριος**, Επίκουρος Καθηγητής Βιοπληροφορικής στη Γενωμική, ΤΒΒ, Π.Θ.

**Τριμελής Επιτροπή**

**Αμούτζιας Γρηγόριος**, Επίκουρος Καθηγητής Βιοπληροφορικής στη Γενωμική, ΤΒΒ, Π.Θ.

**Μαρκουλάτος Παναγιώτης**, Καθηγητής Εφαρμοσμένης Μικροβιολογίας με έμφαση στη Βιοτεχνολογία, ΤΒΒ, Π.Θ.

**Μόσιαλος Δημήτριος**, Επίκουρος Καθηγητής Βιοτεχνολογίας Μικροβίων, ΤΒΒ, Π.Θ.

## ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της πτυχιακής εργασίας μου, Επίκουρο Καθηγητή κ. Αμούτζια Γρηγόριο, για την αμέριστη βοήθεια και καθοδήγησή του, καθώς και για την εμπιστοσύνη που μου έδειξε κατά τη διάρκεια της δουλειάς και παραμονής μου στο εργαστήριο.
Επίσης, είμαι ευγνώμων στον Καθηγητή κ. Μαρκουλάτο Παναγιώτη και τον Επίκουρο Καθηγητή κ. Μόσιαλο Δημήτριο, για τις πολύτιμες υποδείξεις τους και που με τίμησαν με τη συμμετοχή τους στην τριμελή εξεταστική επιτροπή της πτυχιακής εργασίας μου.
Ευχαριστώ τους συναδέλφους και φίλους μου, Φλιάτουρα Βασιλική, Δήμητρα Στεργίου, Μάριο Νικολαΐδη και Παναγιώτη Βλασταρίδη για τη βοήθεια, τη στήριξη και τις όμορφες στιγμές που μου πρόσφεραν εντός και εκτός εργαστηρίου. Τέλος, ευχαριστώ τους γονείς μου, Ηλία Ντουντούμη και Βασιλική Κουρίκα και τον αδερφό μου Γεώργιο Ντουντούμη για την ολόψυχη αγάπη και την υποστήριξη που μου παρέχουν όλα αυτά τα χρόνια.

# Περίληψη

Γενικότερα, επικρατεί η πεποίθηση ότι οι Περιοχές Χαμηλής Πολυπλοκότητας (Low Complexity Regions - LCRs) συναντώνται κατά κύριο λόγο στους ευκαρυώτες, ενώ στους προκαρυώτες εμφανίζονται με πολύ χαμηλότερη συχνότητα. Έτσι, οι μελέτες στην πλειοψηφία τους έχουν επικεντρωθεί σε ευκαρυώτες και λίγοι έχουν μελετήσει συγκεκριμένες περιπτώσεις προκαρυωτών. Σκοπός αυτής της μελέτης είναι να παρέχει την πρώτη, μεγάλης κλίμακας έρευνα σε επίπεδο γονιδιώματος, στους προκαρυώτες και να αποκαλυφθούν α) πόσο σημαντικά είναι τα LCRs β) ποιές είναι οι ιδιότητές τους, γενικότερα, γ) ποιο είναι το αμινοξικό τους περιεχόμενο, δ) εάν αυτό το αμινοξικό περιεχόμενο σχετίζεται με συγκεκριμένες μοριακές λειτουργίες και ε) να κατανοηθεί ο ρόλος των πρώιμων πολυπεπτιδίων, ενώ ο γενετικός κώδικας ακόμη εξελισσόταν. Για το σκοπό αυτό αναλύθηκαν περισσότερα από 1500 πρωτεώματα προκαρυωτών και βακτηριοφάγων. Ανάλογα με την αμινοξική τους σύσταση, αυτές οι σχετικά απλές περιοχές, φαίνεται να σχετίζονται με συγκεκριμένες μοριακές λειτουργίες, όπως πρόσδεση σε μόρια RNA και DNA καθώς και σε συγκεκριμένα ιόντα μετάλλου. Επιπροσθέτως, αυτές οι περιοχές έχουν ταυτοποιηθεί σε αρχέγονες και υψηλά εκφραζόμενες πρωτεΐνες, όπως αυτές του ριβοσώματος, συγκεκριμένες τσαπερόνες (μοριακές πρωτεΐνες-συνοδούς), πρωτεΐνες του κυτταρικού τοιχώματος και της μεμβράνης.

# Abstract

It is generally believed that Low Complexity Regions are found mostly in eukaryotes and that they are not so prominent in prokaryotes. Thus, the vast majority of studies have focused on the evolution and functional role of LCRs found in eukaryotes whereas the prokaryotic world has not been studied thoroughly. The goal of this study is to provide the first large scale genomic investigation in prokaryotes and reveal i) how prominent LCRs are, ii) their general properties, iii) what amino acid content they have, iv) whether this amino acid content is related to certain molecular functions and v) try to understand the role of the early (and by their nature low complexity) polypeptides while the genetic code was still evolving. Towards this goal, we analyzed more than 1500 prokaryotic and bacteriophage proteomes. The specific aminoacid content of these rather simple regions appears to be linked to certain molecular functions, such as RNA, DNA and metal -ion binding. In addition, these regions have been identified in very ancient and usually highly expressed proteins, such as the ribosome, certain chaperones, cell wall and membrane proteins.

# Περιεχόμενα

# Περιεχόμενα Εικόνων

# Περιεχόμενα Πινάκων

# 1. Εισαγωγή

## 1.1 Δομή

Ως Περιοχές Χαμηλής Πολυπλοκότητας (**L**ow **C**omplexity **R**egions, **LCR**s) χαρακτηρίζονται τμήματα πρωτεϊνών με περιορισμένο και ασυνήθιστο αμινοξικό περιεχόμενο, συνήθως εμπλουτισμένα σε ένα ή λίγα αμινοξέα. Αρχικά, θεωρούνταν "άχρηστες" και περιττές περιοχές της πρωτεΐνης, που δεν έχουν κάποια συγκεκριμένη λειτουργία και συνήθως φιλτράρονται κατά την αναζήτηση ομολόγων ακολουθιών (Altschul et al., 1990). Ωστόσο, πειραματικά δεδομένα καταδεικνύουν, όλο και περισσότερο, το σημαντικό και ως τώρα παραμελημένο ρόλο τους (Haerty and Golding, 2010). Οι περιοχές χαμηλής πολυπλοκότητας μπορούν, επίσης, να δημιουργούνται από επαναλήψεις ενός αμινοξέος ή και διαδοχικές ή διάσπαρτες επαναλήψεις ενός μικρού πρωτεϊνικού τμήματος (μήκους 2-5 αμινοξέων περίπου). Σε αυτή τη μελέτη, εστιάζουμε σε περιοχές με μικρή ποικιλία αμινοξέων (ορίζονται ως LCRs) και σε μια ειδική κατηγορία τους, τις επαναλήψεις ενός αμινοξέος (**S**ingle **A**minoacid **R**epeats, ορίζονται ως **SAR**s).

Οι περιοχές χαμηλής πολυπλοκότητας θεωρούνται ότι ανήκουν δομικά στην κατηγορία των περιοχών ενδογενούς δομικής αστάθειας (**I**ntrinsically **D**isordered **R**egions - **IDR**s). Πρωτεΐνες με περιοχές ενδογενούς δομικής αστάθειας (**I**ntrinsically **D**isordered **P**roteins, **IDP**s) δεν έχουν σταθερή ή διαταγμένη τρισδιάστατη δομή (Dunker et al., 2008, 2001; Dyson and Wright, 2005) και ποικίλλουν από πλήρως αδόμητες έως μερικώς δομημένες. Επίσης, σε κάποιες περιπτώσεις, μπορούν να υιοθετήσουν σταθερή τρισδιάστατη δομή μετά τη πρόσδεση άλλων μακρομορίων (van der Lee et al., 2014).

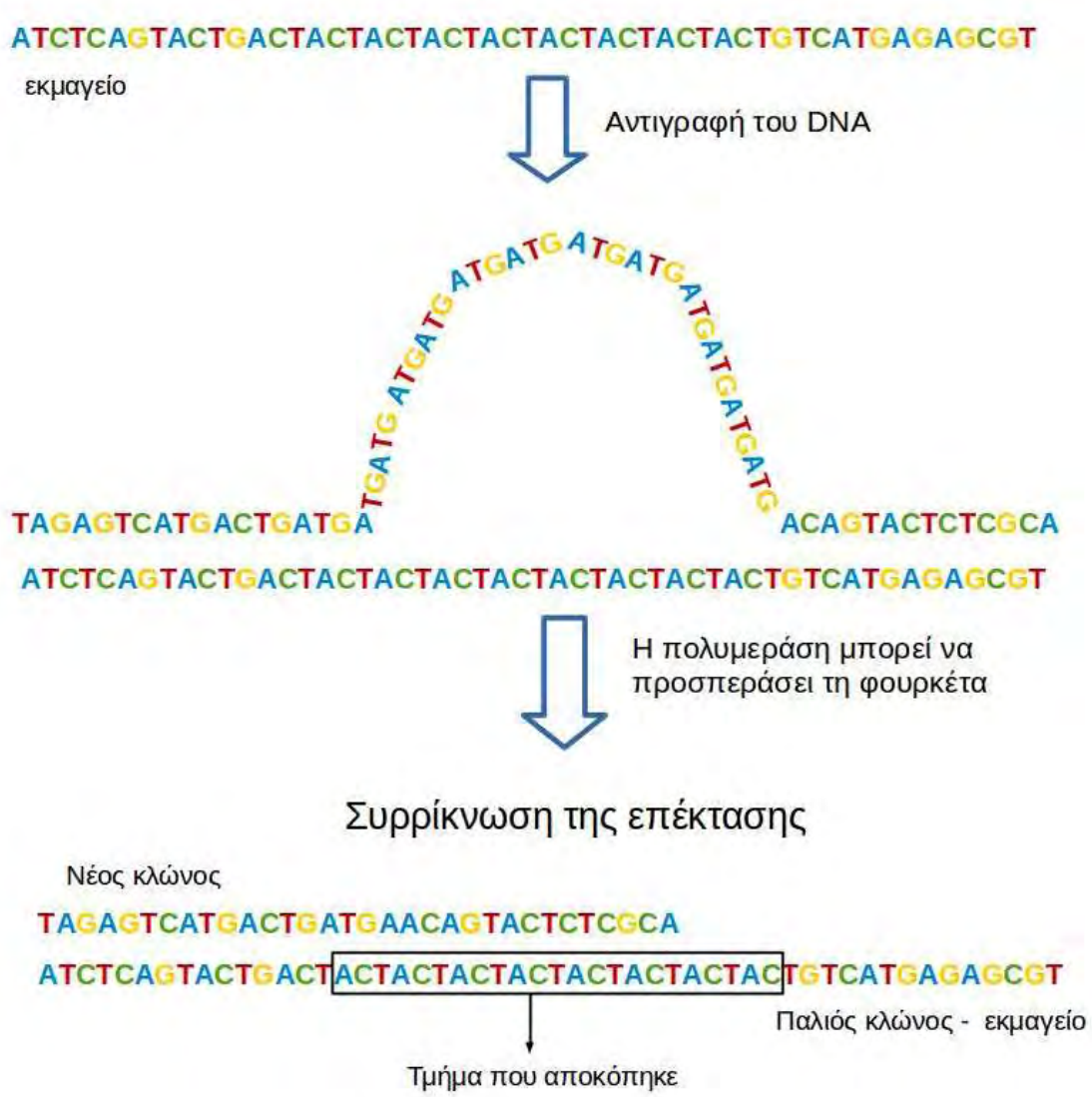
Οι περιοχές χαμηλής πολυπλοκότητας, συνήθως, είναι εύκαμπτες (Coletta et al., 2010) και δεν μπορούν ή είναι δύσκολο να κρυσταλλωθούν. Η παλαιότερη επικρατούσα αντίληψη ήταν ότι τα LCRs είναι περιοχές χωρίς κάποια συγκεκριμένη δομή. Ωστόσο, πρόσφατες μελέτες καταδεικνύουν ότι μπορούν να σχηματίζουν δευτεροταγείς δομές και επιπλέον ότι δεν είναι πάντοτε εκτεθειμένες στην επιφάνεια των πρωτεϊνών. Επίσης, τα LCRs στην πλειοψηφία τους μπορούν να υιοθετήσουν περισσότερες από μία δευτεροταγείς δομές, πράγμα που δείχνει ότι βρίσκονται σε σημεία που αποκτούν μεταβατικές δομές (Kumari et al., 2015). Επιπλέον, έχει δειχθεί ότι πολύ εύκαμπτα LCRs ανήκουν σε πρωτεΐνες που αποκτούν σταθερή δομή μετά την πρόσδεση των μορίων - στόχων τους (Kumari et al., 2015). Τέλος, τα LCRs, πιο συχνά, φαίνεται να δημιουργούν ελικοειδείς δομές (DePristo et al., 2006; Huntley and Golding, 2002; Wootton, 1994). Επιπλέον, ανάλυση της σύνθεσης των αμινοξέων έδειξε πως τα περισσότερα LCRs είναι εμπλουτισμένα σε αμινοξέα που προωθούν την δημιουργία έλικας (Kumari et al., 2015).

## 1.2 Δημιουργία, επέκταση και συρρίκνωση των LCRs

Τα LCRs πιθανότατα δημιουργούνται κατά την αντιγραφή, τον ανασυνδυασμό, και την επιδιόρθωση του DNA (Jentzsch et al., 2013; Marcotte et al., 1999).

Μπορεί να σχηματίζονται από ολίσθηση κατά την αντιγραφή (replication slippage), αλλιώς γνωστή ως ολίσθηση των κλώνων με λάθος ταίριασμα βάσεων (slipped-strand mispairing) που οδηγεί είτε σε επέκταση είτε σε συρρίκνωση ενός τρι- ή δι-νουκλεοτιδίου κατά τη διάρκεια της αντιγραφής του DNA. Μια τέτοια ολίσθηση συμβαίνει, συνήθως, όταν υπάρχουν τέτοιες επαναλαμβανόμενες αλληλουχίες νουκλεοτιδίων. Όταν συναντάται μια τέτοια επανάληψη, μπορεί να αντιγραφεί ή πιο συχνά, να μεταγραφεί, από λάθος, το ίδιο τμήμα περισσότερες από μια φορές (Polymerase stuttering) (Anderson et al., 2007; Kurzynska-Kokorniak et al., 2007; Mauro et al., 2007).

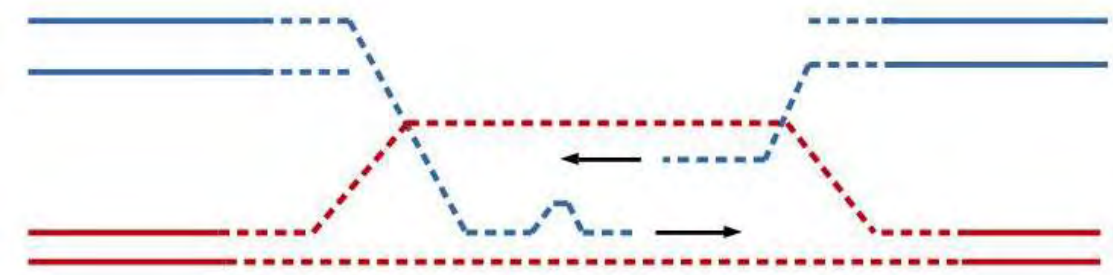
Επιπλέον, αυτές οι επαναλήψεις δημιουργούν σταθερές δομές φουρκέτας στο νεοσυντιθέμενο καθυστερημένο κλώνο του DNA, κατά την αντιγραφή. Αυτό οδηγεί σε επέκταση ή συρρίκνωση των επαναλήψεων (Djian, 1998). Η πολυμεράση μπορεί να προσπεράσει τη δομή φουρκέτας που δημιουργείται κι έτσι, το επαναλαμβανόμενο τμήμα συρρικνώνεται. Η Εικόνα 1 συνοψίζει τη διαδικασία.




***Εικόνα 1.*** Συρρίκνωση επαναλήψεων μέσω παράληψης της δομής φουρκέτας από την πολυμεράση του DNA κατά την αντιγραφή του DNA.

Τέτοιες περιοχές, μπορούν, επιπλέον να σχηματίζονται (και να επεκτείνονται ή να συρρικνώνονται) και κατά την επιδιόρθωση του DNA και έτσι να εντοπίζονται και σε κύτταρα που δεν διαιρούνται. Κατά την επιδιόρθωση ενός κενού ή χάσματος μπορεί να σχηματιστεί δομή φουρκέτας και να αποκοπεί το τμήμα του DNA που έχει το κενό. Στη συνέχεια, μια πολυμεράση του DNA με επιδιορθωτική δράση συνθέτει ξανά το τμήμα που έχει αποκοπεί. Έτσι, επιδιόρθωση του χάσματος μπορεί να οδηγήσει σε επέκταση ή συρρίκνωση της επανάληψης (με τους μηχανισμούς που προαναφέρθηκαν) (Djian, 1998).

Τέλος, οι επαναλήψεις μπορεί να επεκταθούν ή να συρρικνωθούν και μέσω του ανασυνδυασμού, κατά τη σύνθεση του DNA (Djian, 1998; Read et al., 2004; Richard and Pâques, 2000; Wells, 1996). Η διαδικασία φαίνεται στην Εικόνα 2, όπου οι διακεκομμένες γραμμές είναι τα επαναλαμβανόμενα τμήματα του DNA, ενώ τα μαύρα βέλη υποδηλώνουν τη σύνθεση του DNA.



***Εικόνα 2.*** *Επέκταση ή συρρίκνωση των επαναλήψεων κατά τον ανασυνδυασμό του DNA. Οι διακεκομμένες γραμμές είναι τα επαναλαμβανόμενα τμήματα του DNA, ενώ τα μαύρα βέλη υποδηλώνουν τη σύνθεση του DNA.*

## 1.3 Ρόλος και Λειτουργία

### 1.3.1 Ρόλος των LCRs σε ασθένειες

Οι περιοχές χαμηλής πολυπλοκότητας στους ευκαρυώτες βρίσκονται στο επίκεντρο κάποιων ερευνών, καθώς έχουν βρεθεί να εμπλέκονται σε ορισμένες ασθένειες (Mirkin, 2007), ειδικά νευροεκφυλιστικές (ασθένεια του Huntington). Έχουν αναγνωριστεί πολλές νευρολογικές διαταραχές που προκαλούνται από ασταθείς επαναλήψεις τρι-, τετρα- και πεντα- νουκλεοτιδίων. Οι ασθένειες που ανήκουν σε αυτή την κατηγορία έχουν χαρακτηριστεί ως διαταραχές επέκτασης ασταθούς επανάληψης (Gatchel and Zoghbi, 2005).

Επιπλέον, έχουν αναφερθεί εννέα ασθένειες που σχετίζονται με επαναλήψεις πολυ - γλουταμίνης και εννέα που σχετίζονται με επαναλήψεις πολυ - αλανίνης (Albrecht et al., 2004; Amiel et al., 2004; Brown and Brown, 2004; Gatchel and Zoghbi, 2005). Παρόλο που οι ασθένειες, που οφείλονται σε SARs γλουταμίνης και αλανίνης, είναι διαφορετικές, οδηγούν και οι δύο σε λάθος αναδίπλωση πρωτεϊνών. Η λάθος αναδίπλωση του SAR αλανίνης στην πρωτεΐνη PABPN1 έδειξε ότι προκαλεί οφθαλμοφαρυγγική μυϊκή δυστροφία (OPMD), η οποία είναι η μόνη διαταραχή πολυ – αλανίνης, που προκαλείται από μια επέκταση σε μια πρωτεΐνη που δεν είναι μεταγραφικός παράγοντας. Επιπλέον, τα μεγάλα σε μήκος SARs αλανίνης είναι τοξικά για τα κύτταρα (Oma et al., 2005; Oma Yoko et al., 2009; Rankin et al., 2000).

Επίσης, πρωτεΐνες, που διαθέτουν περιοχές χαμηλής πολυπλοκότητας, εντοπίζονται, σε αφθονία, σε παθογόνους μικροοργανισμούς. Αρκετά παράσιτα έχουν ένα ασυνήθιστα υψηλό ποσοστό τέτοιων πρωτεϊνών, όπως τα apicomplexa και τα Plasmodium των θηλαστικών. Συγκεκριμένα το πρωτέωμα του *Plasmodium falciparum sporozoite* έχει πολλές πρωτεΐνες ενδογενούς δομικής αστάθειας και είναι εμπλουτισμένο σε επαναλαμβανόμενες ακολουθίες (Feng et al., 2006). Η δομική πλαστικότητα των πρωτεϊνών αυτών, μπορεί να ευνοήσει την επιβίωση των παρασίτων τόσο μέσω αναστολής της δημιουργίας αποτελεσματικών αντισωμάτων

υψηλής συγγένειας όσο και με το να διευκολύνουν τις αλληλεπιδράσεις με διάφορα μόρια του ξενιστή, που είναι απαραίτητα για την προσκόλληση και την εισβολή στα κύτταρά του.

Επιπλέον, διάφορα επαναλαμβανόμενα μοτίβα έχουν ανοσογονική δράση, που παρατηρήθηκε και μετά από χρήση τους για την ανοσοποίηση ζώων και σε άτομα που εκτέθηκαν σε λοιμώξεις (Feng et al., 2006).

Τέλος, πολλοί επίτοποι των Β-κυττάρων έχουν αναγνωριστεί σε πρωτεΐνες ενδογενούς δομικής αστάθειας και πολλοί από αυτούς φαίνεται να επάγουν λειτουργικές ανοσολογικές αποκρίσεις και επομένως αποτελούν πιθανά υποψήφια εμβόλια (Adda et al., 2012; Foquet et al., 2014; Foucault et al., 2010; Raj et al., 2014; Yagi et al., 2014). Για παράδειγμα, στην προστατευτική δράση του πιο εξελιγμένου εμβολίου κατά της ελονοσίας, του RTS,S, φαίνεται να μεσολαβούν αντισώματα στις δομικά ασταθείς επαναλήψεις της περισποροζωϊτικής (circumsporozoite) πρωτεΐνης (Dyson et al., 1990; Foquet et al., 2014). Επιπλέον, οι δομικά ασταθείς πρωτεΐνες φαίνεται να είναι περισσότερο αντιγονικές από τις δομημένες πρωτεΐνες, πιθανόν γιατί εκτίθενται περισσότερο στον διαλύτη και αυτό επιτρέπει την αναγνώρισή τους από αντισώματα (MacRaild et al., 2016).

### 1.3.2 Δομικός ρόλος των LCRs

Από μια μηχανιστική σκοπιά, τα LCRs αρχικά θεωρούνταν ότι είναι μη δομημένες περιοχές και ότι κυρίως έχουν συνδετικό ρόλο (linkers) με το να παρεμβάλλονται μεταξύ δομημένων επικρατειών και να τις ενώνουν διατηρώντας την απόσταση μεταξύ τους (Huntley and Golding, 2002). Ωστόσο, πολλές έρευνες έχουν καταδείξει το δομικό ρόλο τους σε πρωτεΐνες, όπως το κολλαγόνο, η μυοσίνη, οι κερατίνες, το μετάξι, διάφορες πρωτεΐνες του κυτταρικού τοιχώματος, κ.α. (Luo and Nijveen, 2014), το συγκολλητικό ρόλο τους (So et al., 2016), τη λειτουργία τους σε κολλώδεις πρωτεΐνες που εκκρίνονται και χρησιμοποιούνται για παγίδευση θηραμάτων (Haritos et al., 2010) ή το ρόλο τους ως επαγωγείς της μοριακής κίνησης, όπως για παράδειγμα τα συστήματα TonB/TolA (Brewer et al., 1990).

Φορτισμένες ομάδες αμινοξέων μπορούν να συνεισφέρουν στη δομική σταθερότητα της περιοχής, στην οποία βρίσκονται μέσα στην πρωτεΐνη. Οι γέφυρες αλάτων και οι δεσμοί υδρογόνου σε ομάδες με μικτό (και θετικό και αρνητικό) φορτίο βελτιώνουν τη σταθερότητα της διαμόρφωσης της πρωτεΐνης. Επίσης, υψηλά φορτισμένες περιοχές σε πρωτεΐνες συχνά σχηματίζουν καινούριες έλικες, που μερικές φορές σταθεροποιούνται από ιόντα μετάλλων. Ένα παράδειγμα είναι οι επαναλαμβανόμενες, υπερελικομένες δομές που δεσμεύουν ασβέστιο και βρίσκονται στην αλκαλική ουρά του *Pseudomonas aeruginosa* (Zhu and Karlin, 1996).

### 1.3.3 Λειτουργία των LCRs ως επιφάνειες αλληλεπίδρασης

Έχει δειχθεί, ότι οι πρωτεΐνες που διαθέτουν LCRs τείνουν να κάνουν περισσότερες αλληλεπιδράσεις με άλλα μόρια σε σχέση με τις πρωτεΐνες που δεν έχουν LCRS, καθώς και ότι τα LCRs έχουν διαφορετικές λειτουργίες και ρόλο, που συνήθως

σχετίζεται με τη θέση τους μέσα στην πρωτεΐνη. Πιο συγκεκριμένα, τα LCRs εντοπίζονται περισσότερο στα άκρα της πρωτεϊνικής ακολουθίας και φαίνεται πως το μήκος τους καθορίζει και την ικανότητά τους να προσδένουν διάφορα μόρια, σε αντίθεση με τα LCRs που εντοπίζονται στο μέσο της πρωτεϊνικής ακολουθίας. Όσον αφορά το ρόλο τους, κατά την ανάλυση με οντολογίες, τα LCRs που εντοπίζονται στα άκρα των πρωτεϊνών σχετίζονται με μετάφραση και με απόκριση stress, ενώ τα LCRs στο μέσο της ακολουθίας σχετίζονται με μεταγραφή (Coletta et al., 2010).

Ανάλογα με το περιεχόμενό τους σε αμινοξέα, τα LCRs μπορούν να δημιουργούν επιφάνειες για αλληλεπίδραση με τη διπλοστοιβάδα φωσφολιπιδίων (Robison et al., 2016). Πολλές πρωτεΐνες και πεπτίδια περιέχουν αλληλουχίες αμινοξέων πλούσιες σε αργινίνη και λυσίνη που μπορούν να αλληλεπιδράσουν με αρνητικά φορτισμένα φωσφολιπίδια, όπως η φωσφατιδυλογλυκερόλη (PG). Για παράδειγμα, αντιμικροβιακά πεπτίδια, που εκκρίνονται από ευκαρυωτικά κύτταρα βασίζονται σε ηλεκτροστατικές αλληλεπιδράσεις με τις φορτισμένες βακτηριακές μεμβράνες προκειμένου να προκαλέσουν τη λύση τους. Τα κύτταρα εκμεταλλεύονται την "προτίμηση" των θετικά φορτισμένων μικρών πεπτιδίων να αλληλεπιδρούν με τις βακτηριακές μεμβράνες - στόχους, που είναι αρνητικά φορτισμένες, σε σύγκριση με τις αμφιτεριονικές (zwitterionic) ευκαρυωτικές μεμβράνες.

Επίσης, οι περιοχές χαμηλής πολυπλοκότητας μπορούν να προσδένονται σε μόρια DNA και RNA. Η δέσμευση του RNA και του DNA από πρωτεΐνες μπορεί να επιτευχθεί με διάφορους τρόπους και από διάφορα πρωτεϊνικά μοτίβα ή επικράτειες. Όσον αφορά τα LCRs, μπορούν να προσδένονται στο DNA είτε σαν θετικά φορτισμένες ομάδες είτε με αλληλεπιδράσεις με το phospho-sugar backbone του DNA. Ένα παράδειγμα αποτελεί η πρωτεΐνη Ku. Τόσο τα ευκαρυωτικά όσο και τα προκαρυωτικά ομόλογά της έχει δειχθεί ότι δεσμεύουν τα άκρα του DNA μέσω των καρβοξιτελικών, πλούσιων σε λυσίνη, περιοχών χαμηλής πολυπλοκότητας. Επίσης, και η πρωτεΐνη Hlp (πρωτεΐνη που μοιάζει με ιστόνη, Histone-like protein) των μυκοβακτηρίων διαθέτει παρόμοιες επαναλήψεις πλούσιες σε λυσίνη που προσδένουν DNA (Kushwaha and Grove, 2013). Περιοχές με επαναλαμβανόμενα μοτίβα RGG ή επαναλήψεις γλυκίνης ακολουθούμενες από ένα αρωματικό αμινοξύ (GGY, GGF, GGW) έχουν βρεθεί, επίσης, να εμπλέκονται στην πρόσδεση μορίων DNA και RNA.

Τέλος, τα LCRs μπορούν να σχηματίσουν αρνητικά φορτισμένες ή ακόμα και όξινες-ιστιδινικές ομάδες και με αυτό τον τρόπο να κατευθύνουν ιόντα μετάλλων ($Ca^{+2}$, $Zn^{+2}$, $Mn^{+2}$, κ.α.) (Zhu and Karlin, 1996).

### 1.3.4 Ρόλος των LCRs στη μετάφραση πρωτεϊνών

Τα LCRs μπορούν επίσης να έχουν σημαντικό ρόλο στη μετάφραση των πρωτεϊνών. Για αυτό το λόγο έχουν μεγάλο βιοτεχνολογικό ενδιαφέρον (ετερόλογη έκφραση / μετάφραση γονιδίων).

Μπορούν να λειτουργούν σαν "σφουγγάρια" των tRNAs και να επιβραδύνουν τη μετάφραση δίνοντας χρόνο στις πρωτεΐνες να αναδιπλωθούν σωστά (Frugier et al., 2010). Ειδικά στους προκαρυώτες, ο διαφορετικός βαθμός χρήσης των κωδικονίων που επιτάσσεται από τη διαφορά στη συχνότητα εμφάνισης συνώνυμων κωδικονίων,

έχει εξέχοντα ρόλο στη μετάφραση (codon usage bias) (Lithwick and Margalit, 2003). Μπορούν να λειτουργούν επιπλέον, ως σημεία ελέγχου μέσω της μετατόπισης του αναγνωστικού πλαισίου και να κωδικοποιήσουν για μια διαφορετική επανάληψη που κάνει την πρωτεΐνη εξαιρετικά ασταθή ή αδιάλυτη και ενεργοποιεί την ταχεία ανακύκλωσή της, πριν προκληθεί οποιαδήποτε βλάβη στο κύτταρο (Ling et al., 2012; Tyedmers et al., 2010).

Επιπροσθέτως, οι περιοχές αυτές, λόγω της εγγενούς αστάθειάς τους (μπορούν να επεκτείνονται και να συρρικνώνονται ταχέως), θα μπορούσαν να έχουν μεγάλο αντίκτυπο στο ενεργειακό φορτίο της μετάφρασης, ειδικά στους προκαρυώτες (Akashi and Gojobori, 2002; Barton et al., 2010). Έτσι, τα LCRs που ανιχνεύονται σε υψηλά εκφραζόμενες πρωτεΐνες παρουσιάζουν ιδιαίτερο ενδιαφέρον.

### 1.3.5 Ρόλος των LCRs στην εξέλιξη

Αυτές οι περιοχές είναι επίσης πολύ ενδιαφέρουσες από εξελικτική σκοπιά. Ενδεχομένως, να προάγουν τον ανασυνδυασμό καθώς έχουν συνδεθεί με hotspots ανασυνδυασμού (Siwach et al., 2006). Μπορούν να επιτρέψουν την επέκταση ή συρρίκνωση μιας συγκεκριμένης περιοχής της πρωτεΐνης, παρέχοντας έτσι τις πρώτες ύλες για την εξέλιξη.

Αλλά, το πιο ενδιαφέρον είναι ότι πιθανώς συνδέονται με τα πολύ πρώιμα στάδια εξέλιξης της ζωής. Θεωρείται ότι, κατά τα πρώτα στάδια της ζωής, όταν ο γενετικός κώδικας χρησιμοποιούσε μόνο λίγα αμινοξέα, οι πρώιμες πρωτεΐνες είχαν μικρό μήκος και εξ' ορισμού χαμηλή πολυπλοκότητα (Trifonov, 2009, 2000). Έτσι, αυτές οι περιοχές που συναντάμε σήμερα θα μπορούσαν να προσφέρουν μια ματιά στο πολύ μακρινό παρελθόν και να παρέχουν κάποιες ενδείξεις για τη λειτουργία των πρώιμων πεπτιδίων. Για παράδειγμα, είναι η αμινοξική σύσταση των LCRs παρόμοια με αυτή των πρώιμων πεπτιδίων; Για να δοθεί η απάντηση σε αυτή την ερώτηση είναι σκόπιμο να μελετηθούν τα LCRs σε κυτταρικά συστήματα που είναι όσο το δυνατόν πιο κοντά εξελικτικά στις πολύ πρώιμες μορφές ζωής. Για αυτό το σκοπό, επιλέχθηκαν ευβακτήρια και αρχαία, καθώς η μελέτη τέτοιων επαναλήψεων σε ευκαρυώτες θα ήταν προβληματική, επειδή μπορεί να διαστρεβλώσει την αρχική εικόνα. Τα πολυκυτταρικά συστήματα, η διαμερισματοποίηση των κυττάρων, η πιο περίπλοκη γονιδιακή ρύθμιση μπορούν να αντισταθμίσουν τα επιβλαβή αποτελέσματα που μπορεί να έχουν τέτοιες επαναλήψεις. Στα ευβακτήρια και τα αρχαία, λόγω της γενικότερης έλλειψης διαμερισματοποίησης, των μονοκύτταρων συστημάτων τους, της λιγότερο περίπλοκης γονιδιακής ρύθμισης και του μεγάλου μεγέθους του πληθυσμού που θα δώσει απογόνους (high effective population size), η de-novo εμφάνιση μιας ήπια ή μέτρια επιβλαβούς επανάληψης θα έπρεπε να φιλτράρεται από αυτές τις ισχυρές εξελικτικές πιέσεις πολύ γρήγορα. Επιπλέον, οι προκαρυώτες εξελίσσονται πολύ γρήγορα και πολύ συχνά "ξεφορτώνονται" γονίδια ή γονιδιακές περιοχές, όταν δεν τους είναι πλέον απαραίτητα, όπως φαίνεται στα ενδοκυτταρικά παράσιτα. Έτσι, αυτές οι περιοχές δεν θα έπρεπε να είναι συντηρημένες σε άλλους, σχετικά μακρινούς εξελικτικά, συγγενικούς οργανισμούς, εκτός αν έχουν κάποια λειτουργία. Έτσι, προκαρυωτικά LCRs, που ανιχνεύονται σε πολλά προκαρυωτικά είδη, θα πρέπει να έχουν ένα λειτουργικό ρόλο και όχι να είναι αποτέλεσμα ουδέτερης εξέλιξης.

## 1.4 Ανίχνευση των LCRs

Πολλά εργαλεία έχουν αναπτυχθεί για την ανίχνευση LCRs σε πρωτεΐνες (Luo and Nijveen, 2014; Promponas et al., 2000; Wootton and Federhen, 1996). Ένας από τους καλύτερους τρόπους για την ανίχνευση τους είναι με μέτρηση της εντροπίας του Shannon (Shannon Entropy, SE) (Shannon, 1948; Wootton, 1994) σε ένα πρωτεϊνικό τμήμα. Όσο πιο μικρή είναι η τιμή της εντροπίας του Shannon, τόσο μεγαλύτερη είναι η ομοιογένεια του πρωτεϊνικού τμήματος, όσον αφορά την αμινοξική του σύσταση.

Η εντροπία του Shannon (SE) ορίζεται ως:

$$SE = -\sum_{\alpha \in AA} p(\alpha) * \log_b p(\alpha)$$

όπου:
**α**: Το εκάστοτε αμινοξύ
**AA**: Το πρωτεΐνικό τμήμα που αναλύεται
**p(α)**: συχνότητα του κάθε αμινοξέως (α) στο πρωτεϊνικό κομάτι που αναλύεται (αριθμός του αμινοξέος στο κομμάτι / μήκος κομματιού)
**b**: βάση του λογαρίθμου (συνήθως χρησιμοποιούνται το 2 και το 10). Χρησιμοποιήθηκε ως βάση το 10.

## 1.5 Νευρωνικά Δίκτυα

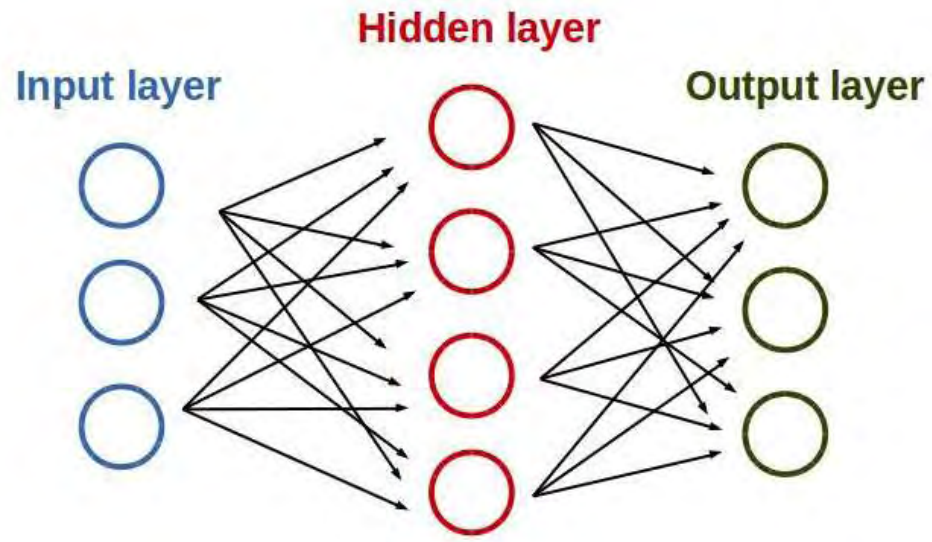
Τα νευρωνικά δίκτυα είναι υπολογιστικά συστήματα εμπνευσμένα από τα βιολογικά νευρωνικά δίκτυα (του εγκεφάλου) (“Artificial Neural Networks as Models of Neural Information Processing | Frontiers Research Topic,” n.d.), που “μαθαίνουν”, δηλαδή προοδευτικά βελτιώνουν την απόδοσή τους, εξετάζοντας παραδείγματα. Πιο συγκεκριμένα, ένα παράδειγμα είναι η αναγνώριση εικόνων. Τα νευρωνικά δίκτυα μπορεί να μάθουν να εντοπίζουν εικόνες που περιέχουν αυτοκίνητα αναλύοντας διάφορα παραδείγματα εικόνων που έχουν επισημανθεί ως "αυτοκίνητο" ή "όχι αυτοκίνητο" και να χρησιμοποιούν τα αποτελέσματα για τον εντοπισμό αυτοκινήτων σε άλλες νέες εικόνες που θα τους δοθούν. Η μάθηση και η αναγνώριση επιτυγχάνεται χωρίς να έχουν από πριν καμία γνώση σχετικά με τα αυτοκίνητα, π.χ. ότι έχουν ρόδες, πόρτες, καθρέφτες ή άλλα χαρακτηριστικά. Αντίθετα, αναπτύσσουν το δικό τους σύνολο χαρακτηριστικών από το μαθησιακό υλικό που επεξεργάζονται.

Ένα νευρωνικό δίκτυο βασίζεται σε ένα σύμπλεγμα από μονάδες ή κόμβους (nodes) που συνδέονται μεταξύ τους και ονομάζονται τεχνητοί νευρώνες. Οι νευρώνες αυτοί

αποτελούν μια απλοποιημένη εκδοχή των βιολογικών νευρώνων του εγκεφάλου. Κάθε σύνδεση μεταξύ των τεχνητών νευρώνων (μια απλοποιημένη εκδοχή των συνάψεων) μπορεί να μεταδώσει ένα σήμα από τον ένα στον άλλο. Ο τεχνητός νευρώνας, που λαμβάνει το σήμα, μπορεί να το επεξεργαστεί και στη συνέχεια να δώσει σήμα στους τεχνητούς νευρώνες που συνδέονται με αυτόν.

Τυπικά, οι τεχνητοί νευρώνες οργανώνονται σε επίπεδα (layers). Τα διαφορετικά επίπεδα μπορούν να εκτελούν διαφορετικές μετατροπές στα σήματα που δέχονται (inputs). Τα σήματα εισέρχονται από το πρώτο επίπεδο (είσοδος), διασχίζουν τα ενδιάμεσα επίπεδα και φτάνουν στο τελευταίο (έξοδος) (“Artificial Neural Networks as Models of Neural Information Processing | Frontiers Research Topic,” n.d.).

Η Εικόνα 3 δείχνει ένα τεχνητό νευρωνικό δίκτυο με τους κόμβους να οργανώνονται σε επίπεδα και τις συνδέσεις μεταξύ τους.




***Εικόνα 3.*** *Τεχνητό νευρωνικό δίκτυο.*

# 2. Υλικά και Μέθοδοι

## 2.1 Λήψη πρωτεωμάτων

Πρωτεώματα από Αρχαία, Βακτήρια και Ιούς λήφθηκαν από το ftp site της Uniprot/Swissprot (Μάρτιος/Απρίλιος 2017) (UniProt Consortium, 2015). Αναλύθηκε μόνο ένα αντιπροσωπευτικό πρωτέωμα από κάθε γένος, οδηγώντας συνολικά σε 1334 πρωτεώματα βακτηρίων, 102 πρωτεώματα αρχαίων και 102 πρωτεώματα βακτηριοφάγων.

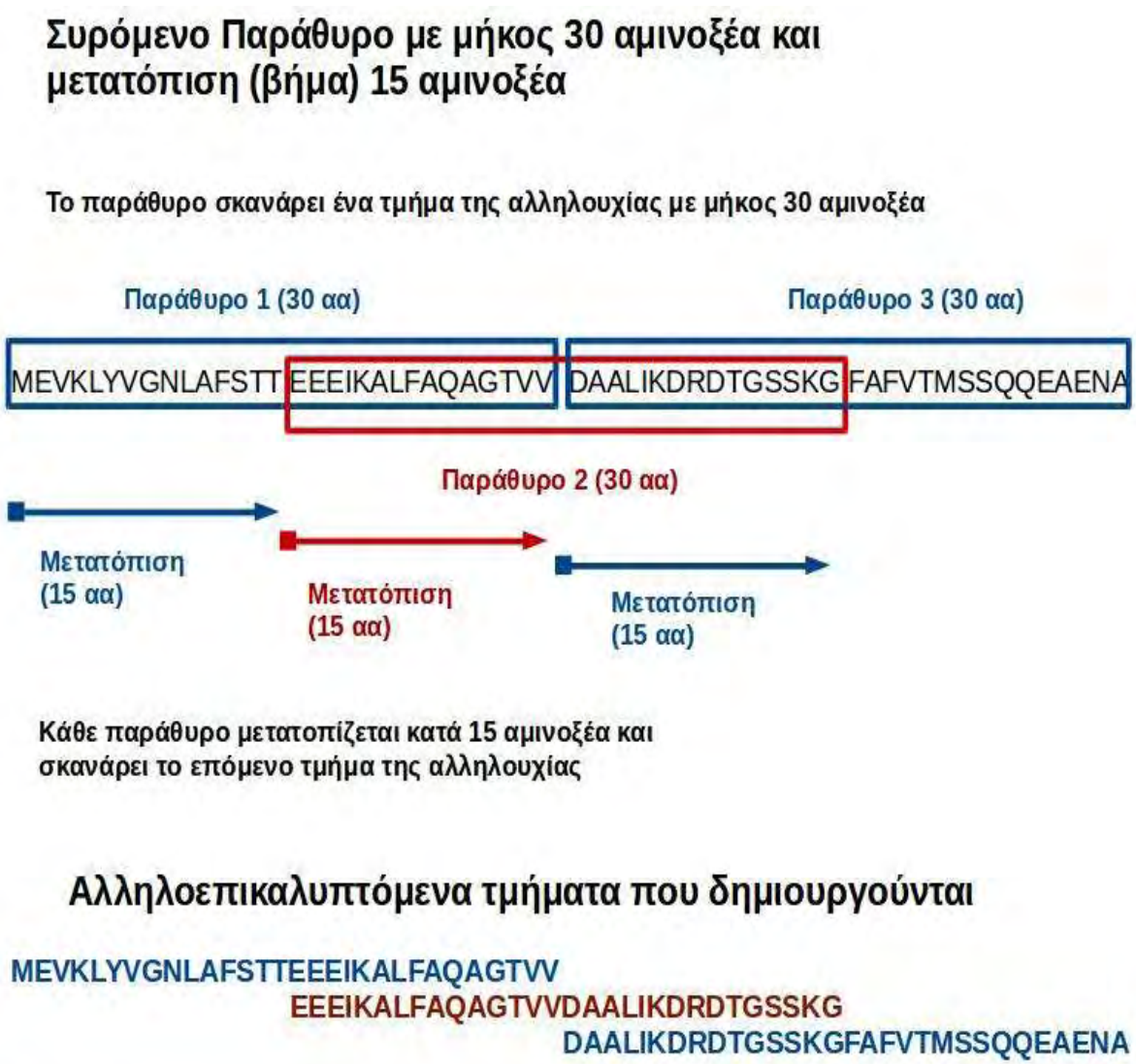
Επιπλέον, αναλύθηκαν και 5 ευκαρυωτικά πρωτεώματα:

- Homo sapiens που λήφθηκε από το ftp://ftp.sanger.ac.uk/pub/vega/human/pep/ (version Homo_sapiens.VEGA68.pep.all.fa),
- Drosophila melanogaster που λήφθηκε από το Flybase,
- Arabidopsis thaliana που λήφθηκε από το ftp://ftp.ensemblgenomes.org/pub/release-26/plants/fasta/arabidopsis_thaliana/pep/,
- Schizosaccharomyces pombe που λήφθηκε από το PomBase και
- Saccharomyces cerevisiae που λήφθηκε από το ftp://ftp.ensembl.org/pub/release-90/fasta/saccharomyces_cerevisiae/pep/.

Στους 3 πρώτους ευκαρυωτικούς οργανισμούς πραγματοποιήθηκε περαιτέρω φιλτράρισμα του πρωτεώματος (λόγω των πολλαπλών ισομόρφων της κάθε πρωτεΐνης) κι έτσι για κάθε γονίδιο επιλέχθηκε και αναλύθηκε η μεγαλύτερη πρωτεΐνη του.

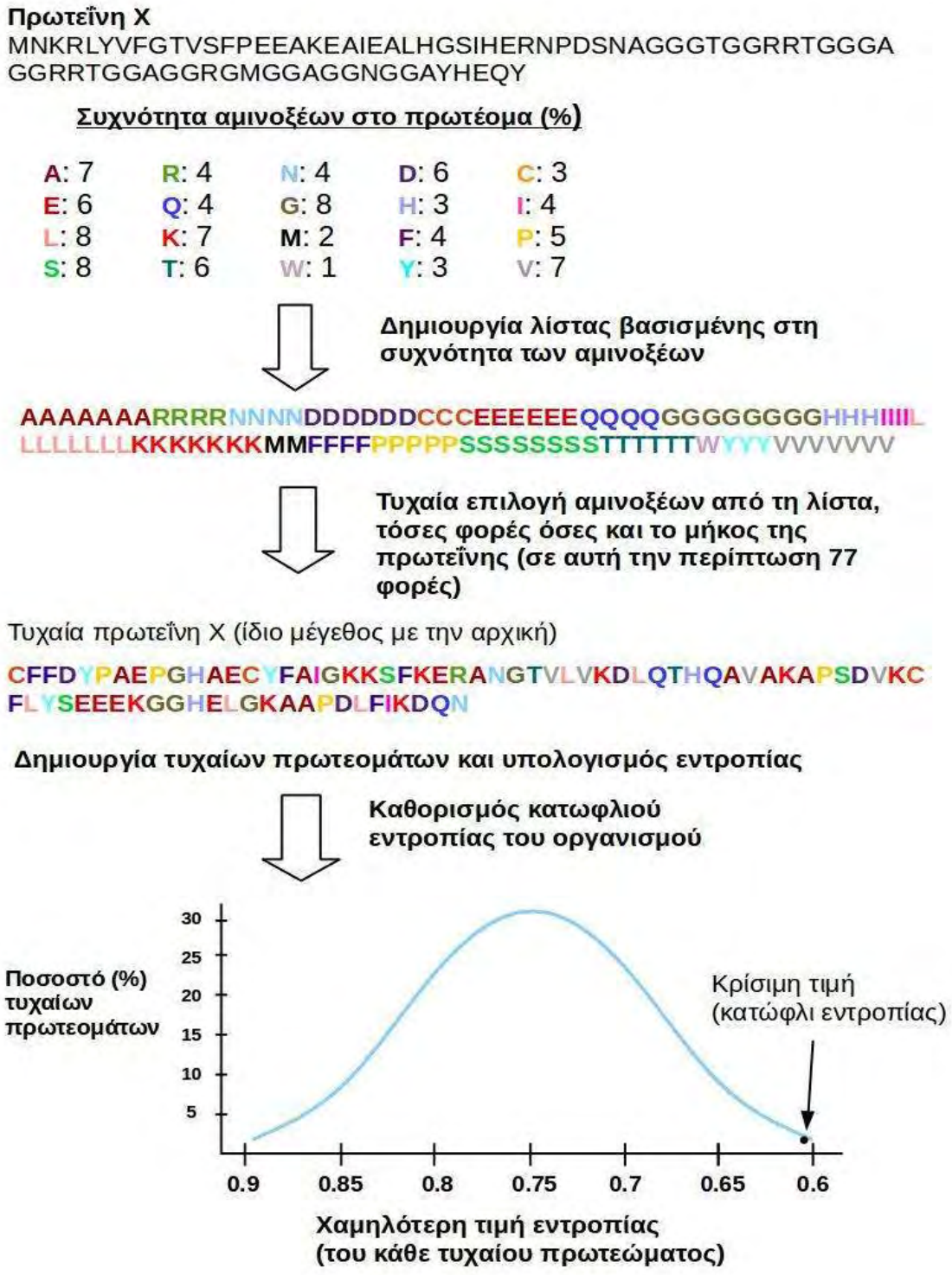
## 2.2 Υπολογισμός της εντροπίας του Shannon με χρήση perl scripts

Σε όλα τα πρωτεώματα, για την ανάλυση της κάθε πρωτεΐνης χρησιμοποιήθηκε η μέθοδος του συρόμενου παραθύρου (sliding window). Πιο αναλυτικά, δημιουργήθηκαν perl scripts που σκάναραν την κάθε πρωτεϊνική ακολουθία σε διαδοχικά αλληλοεπικαλυπτόμενα τμήματα (παράθυρα) με μετατόπιση (βήμα) 25 και 15 αμινοξέα και μήκος 50 και 30 αμινοξέα αντίστοιχα. Για κάθε τμήμα υπολογίστηκε η εντροπία του Shannon.




***Εικόνα 4.*** *Περιγραφή της μεθόδου του συρόμενου παραθύρου με μήκος 30 αμινοξέα και βήμα 15 αμινοξέα*

Για κάθε πρωτέωμα υπολογίστηκε η συχνότητα υποβάθρου των αμινοξέων και στη συνέχεια ανακατεύθηκε κάθε πρωτεΐνη του εκάστοτε πρωτεώματος, χρησιμοποιώντας την συχνότητα υποβάθρου των αμινοξέων του. Αυτή η διαδικασία επαναλήφθηκε 20 φορές και βασίζεται στη μέθοδο Monte Carlo (*Monte Carlo Methods for Applied Scientists*, 2008). Σε κάθε ένα από αυτά τα 20 ανακατασκευασμένα/ανακετεμένα πρωτεώματα, υπολογίστηκε η εντροπία του Shannon για κάθε πρωτεΐνη, με τη μέθοδο του συρόμενου παραθύρου που προαναφέρθηκε. Η χαμηλότερη τιμή εντροπίας που υπολογίστηκε μετά από ανάλυση και των 20 ανακατασκευασμένων πρωτεωμάτων χρησιμοποιήθηκε σαν κατώφλι εντροπίας για το συγκεκριμένο οργανισμό. Με αυτό τον τρόπο εξασφαλίστηκε ότι οποιαδήποτε τιμή εντροπίας χαμηλότερη από αυτήν του κατωφλιού είναι στατιστικά σημαντική και δεν θα μπορούσε να προκύψει καθαρά από τύχη. Κατ' επέκταση όλα τα πρωτεϊνικά τμήματα ενός οργανισμού, με τιμή εντροπίας χαμηλότερη από το κατώφλι που υπολογίστηκε για αυτόν, επιλέχθηκαν, καθώς δεν θα μπορούσαν, τυχαία, να έχουν δημιουργηθεί τμήματα με τόσο χαμηλή πολυπλοκότητα.




***Εικόνα 5.*** *Ανακατασκευή των πρωτεϊνών με τυχαία επιλογή αμινοξέων (βασισμένη στη συχνότητά τους στο πρωτέομα)*

Τα πρωτεϊνικά κομμάτια με πολύ χαμηλή τιμή εντροπίας που λήφθηκαν, αναλύθηκαν περαιτέρω, με μέτρηση του αριθμού και της συχνότητας των αμινοξέων τους, και υπολογίστηκε ένα διάνυσμα συχνότητας αμινοξέων για το καθένα.

Στον Πίνακας 1 φαίνονται τα αποτελέσματα των πρωτεϊνικών τμημάτων που λήφθηκαν για τα αρχαία, τα βακτήρια και τους βακτηριοφάγους μετά από την παραπάνω διαδικασία.

| | **Αρχαία** | **Βακτήρια** | **Φάγοι** |
|---|---|---|---|
| **Παράθυρο 50 αμινοξέων** | 3245 | 60417 | - |
| **Παράθυρο 30 αμινοξέων** | 2466 | 38103 | 170 |

***Πίνακας 1.*** *Αριθμός των πρωτεϊνικών τμημάτων που λήφθηκαν με τη μέθοδο του συρόμενου παραθύρου (μήκους 50 και 30 αα) και το φιλτράρισμα με το κατώφλι εντροπίας του κάθε οργανισμού. Στους φάγους χρησιμοποιήθηκε μόνο το συρόμενο παράθυρο μήκους 30 αα*

## 2.3 Ένωση των αλληλοεπικαλυπτόμενων πρωτεϊνικών τμημάτων

Για τα αλληλοεπικαλυπτόμενα πρωτεϊνικά τμήματα χαμηλής εντροπίας (και πολυπλοκότητας) που λήφθηκαν, κατασκευάστηκαν perl scripts ώστε να συγχωνευτούν και να δημιουργηθούν τα LCRs στο πλήρες μήκος τους. Για τα ενωμένα, ολόκληρα LCRs υπολογίστηκε ξανά η τιμή της εντροπίας και το διάνυσμα συχνότητας των αμινοξέων τους.

| | **Αρχαία** | **Βακτήρια** | **Φάγοι** |
|---|---|---|---|
| **Παράθυρο 50 αμινοξέων** | 2159 | 36517 | - |
| **Παράθυρο 30 αμινοξέων** | 1521 | 22260 | 115 |

***Πίνακας 2.*** *Αριθμός των πρωτεϊνικών κομματιών LCRs που λήφθηκαν μετά την ένωση των αλληλεπικαλυπτόμενων τμημάτων (για παράθυρο μήκους 50 και 30 αα)*

## 2.4 Ανίχνευση επαναλήψεων ενός μοναδικού αμινοξέος (SARs)

Τα προαναφερθέντα ενωμένα πρωτεϊνικά τμήματα χαμηλής πολυπλοκότητας αναλύθηκαν περαιτέρω, για να ανιχνευτούν περιοχές που δημιουργούνται από την επανάληψη του ίδιου αμινοξέος (**S**ingle **A**minoacid **R**epeats, **SAR**s) 10 ή περισσότερες φορές.

## 2.5 Ομαδοποίηση των LCRs

Τα LCRs ομαδοποιήθηκαν με βάση τη συχνότητα (απόλυτος αριθμός, όχι ποσοστό της συχνότητας) των αμινοξέων τους. Τα αμινοξέα με συχνότητα μικρότερη από 3 (για το συρόμενο παράθυρο μήκους 30 αα) φιλτραρίστηκαν. Τα διάνυσματα συχνότητας αμινοξέων (vectors) που προέκυψαν αναλύθηκαν με τη Matlab για ομαδοποίηση και απεικονίστηκαν με τη συνάρτηση clustergram της Matlab.

## 2.6 Ανάλυση με Οντολογίες

Για κάθε πρωτεΐνη που βρέθηκε να περιέχει LCRs ή/και SARs, λήφθηκαν από τη Uniprot οι όροι – οντολογίες (Gene Ontology Consortium, 2015) για βιολογική διεργασία, μοριακή δράση και κυτταρικό διαμέρισμα.

## 2.7 Πολλαπλή στοίχιση και φυλογενετική ανάλυση

Για τις ομόλογες πρωτεΐνες που περιέχουν LCRs, πραγματοποιήθηκε πολλαπλή στοίχιση με χρήση του Muscle (Edgar, 2004) και δημιουργήθηκαν φυλογενετικά δέντρα (με τη μέθοδο BioNJ), που οπτικοποιήθηκαν με το Seaview v4 (Gouy et al., 2010) και το Jalview (Waterhouse et al., 2009).

## 2.8 Δομικές αναλύσεις

Για απεικόνιση πρωτεϊνών που περιέχουν LCRs χρησιμοποιήθηκε το λογισμικό Pymol με δημοσιευμένα δεδομένα δομών 3D (Rose et al., 2015).

## 2.9 Ανάλυση συχνότητας κωδικονίων

Για τις πρωτεΐνες που περιέχουν LCRs, ανακτήθηκαν από την EMBL-Bank (Leinonen et al., 2011) οι ακολουθίες του DNA που τις κωδικοποιούν (CDSs). Αυτό

επιτεύχθηκε με αντιστοίχιση των αναγνωριστικών της πρωτεΐνης (IDs) που δίνονται από τη Uniprot στα αντίστοιχα αναγνωριστικά που δίνονται, για την ίδια πρωτεΐνη, από την EMBL-Bank. Διατηρήθηκαν μόνο οι κωδικές αλληλουχίες που ταίριαζαν ακριβώς στις πρωτεΐνες της Uniprot. Έπειτα, υπολογίστηκε η συχνότητα του κάθε κωδικονίου σε κάθε LCR.

## 2.10 Εργαλείο ανίχνευσης των LCRs και πρόβλεψης της λειτουργίας τους

Με βάση τα αποτελέσματα από την ανάλυση με οντολογίες κατασκευάστηκε, με perl script, ένα εργαλείο που ανιχνεύει LCRs σε ένα δοθέν πρωτεϊνικό τμήμα ή μια πρωτεΐνη και προβλέπει τη λειτουργία τους, με χρήση νευρωνικών δικτύων και με σύγκρισή τους με LCRs που προέρχονται από πρωτεΐνες με γνωστή λειτουργία και που έχει αποδειχθεί πειραματικά ότι εμπλέκονται σε συγκεκριμένες λειτουργίες. Για τη σύγκριση αυτή, χρησιμοποιείται ο συντελεστής συσχέτισης Pearson.

### 2.10.1 Ομαδοποίηση LCRs

Τα LCRs, που ανιχνεύτηκαν τόσο στα προκαρυωτικά όσο και στα ευκαρυωτικά πρωτεώματα, χωρίστηκαν σε 4 μεγάλες ομάδες με βάση τα αποτελέσματα, που προέκυψαν από την ανάλυση με τους όρους - οντολογίες. Για το πρωτέωμα του ανθρώπου οι όροι - οντολογίες λήφθηκαν από τη Uniprot μαζί με το σχολιασμό των πρωτεϊνών στις οποίες ανήκουν τα LCRs. Για το *A.thaliana* λήφθηκαν οι όροι - οντολογίες slim από το Tair, για τη *D. melanogaster* οι όροι - οντολογίες από το Flybase και για τον *S. cerevisiae* οι όροι - οντολογίες slim από το Saccharomyces Genome Database (SGD). Σε κάθε LCR αντιστοιχήθηκαν οι όροι - οντολογίες που χαρακτηρίζουν την πρωτεΐνη, στην οποία ανήκει το LCR. Οι ομάδες δημιουργήθηκαν με χρήση της Matlab και της συνάρτησης clustergram.

### 2.10.2 Κατασκευή διγραμμάτων

Για κάθε LCR των μεγάλων ομάδων που προέκυψαν από την ομαδοποίηση, υπολογίστηκαν τα διγράμματά του με τη μέθοδο του συρόμενου παραθύρου (sliding window) που έχει ήδη αναλυθεί. Δημιουργήθηκαν perl scripts που σκάναραν το κάθε τμήμα LCR με παράθυρο μήκους 2 αμινοξέων και μετατόπιση (βήμα) 1 αμινοξύ. Υπολογίστηκε η συχνότητα των διγραμμάτων και για κάθε LCR δημιουργήθηκε ένα διάνυσμα της συχνότητας του κάθε διγράμματος. Στην Εικόνα 6 συνοψίζεται η διαδικασία κατασκευής των διγραμμάτων.


***Εικόνα 6.*** *Διαδικασία κατασκευής διγραμμάτων, υπολογισμός συνολικού αριθμού και συχνότητας για το καθένα*

### 2.10.3 Κατασκευή Νευρωνικών Δικτύων

Για τις προαναφερθέντες ομάδες, που προέκυψαν από την ομαδοποίηση, κατασκευάστηκαν, επίσης, νευρωνικά δίκτυα με χρήση του γραφικού περιβάλλοντος της Matlab αλλά και με python scripts με χρήση του Tensorflow και του Keras. Δημιουργήθηκαν νευρωνικά δίκτυα με βάση τη συχνότητα των αμινοξέων στα LCRs αλλά και με βάση τη συχνότητα των διγραμμάτων τους.

Οι παράμετροι που χρησιμοποιήθηκαν για τη βέλτιστη λειτουργία του νευρωνικού δικτύου που δημιουργήθηκε με το Tensorflow - Keras ήταν, τόσο για το νευρωνικό δίκτυο που βασίστηκε στη συχνότητα των αμινοξέων όσο και για αυτό που βασίστηκε στη συχνότητα των διγραμμάτων: κόμβοι: 10, dropout: 0.16, epochs: 152, batch size: 60.

Για τα δίκτυα που δημιουργήθηκαν με τη Matlab οι παράμετροι για το νευρωνικό δίκτυο, που κατασκευάστηκε με χρήση της συχνότητας των διγραμμάτων, ήταν: κόμβοι: 10, epochs: 131 και με χρήση της συχνότητας των αμινοξέων: κόμβοι: 10, epochs: 58.

### 2.10.4 Υπολογισμός Συντελεστή Συσχέτισης

Για το λειτουργικό χαρακτηρισμό των LCRs, που θα ανιχνευτούν από το εργαλείο χρησιμοποιείται, επιπλέον, ο συντελεστής συσχέτισης Pearson. Τα LCRs που θα ανιχνευθούν συγκρίνονται, με μέτρηση της τιμής του συντελεστή συσχέτισης, με τα LCRs από καλά χαρακτηρισμένες πρωτεΐνες, που ανήκουν στις παραπάνω ομάδες και έχει βρεθεί πειραματικά ότι επιτελούν μια συγκεκριμένη διεργασία.

### 2.10.5 Εκτίμηση της λειτουργίας του εργαλείου

Τα ευκαρυωτικά πρωτεώματα καθώς και τα πρωτεϊνικά τμήματα ενδογενούς δομικής αστάθειας που δίνονται στις εργασίες των Castello (Castello et al., 2016) και Järvelin (Järvelin et al., 2016) αναλύθηκαν ξανά με το εργαλείο για να εκτιμηθεί η ακρίβεια των προβλέψεών του. Για αυτό το σκοπό, υπολογίστηκε η θετική προγνωστική αξία (**P**ositive **P**redictive **V**alue, **PPV**) και χρησιμοποιήθηκε το υπεργεωμετρικό τεστ.

# 3. Αποτελέσματα - Συζήτηση

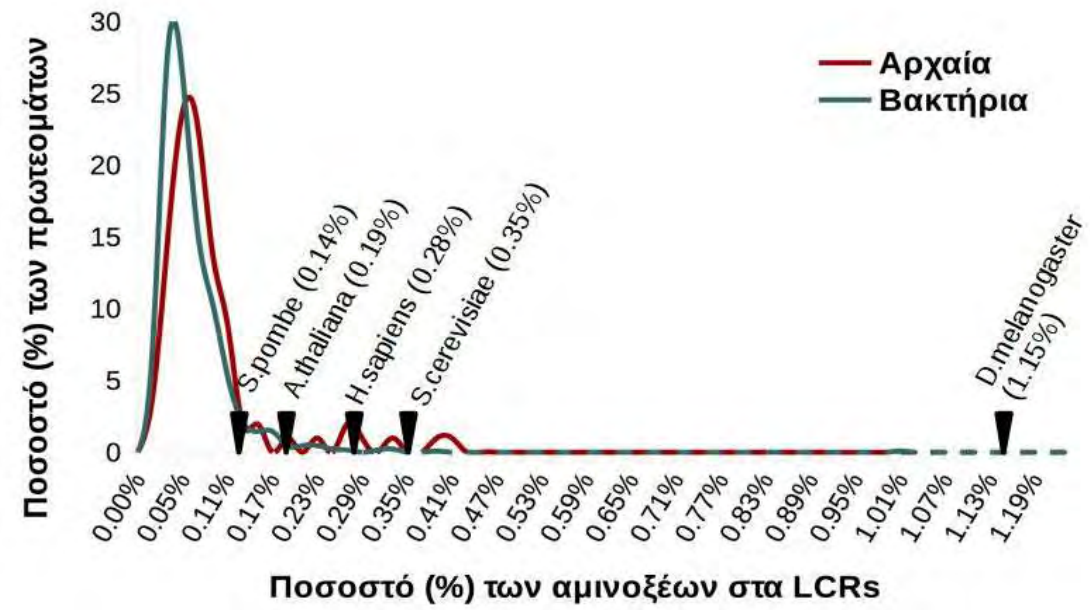
## 3.1 Στατιστικές Αναλύσεις

### 3.1.1 Σύγκριση με τους Ευκαρυώτες

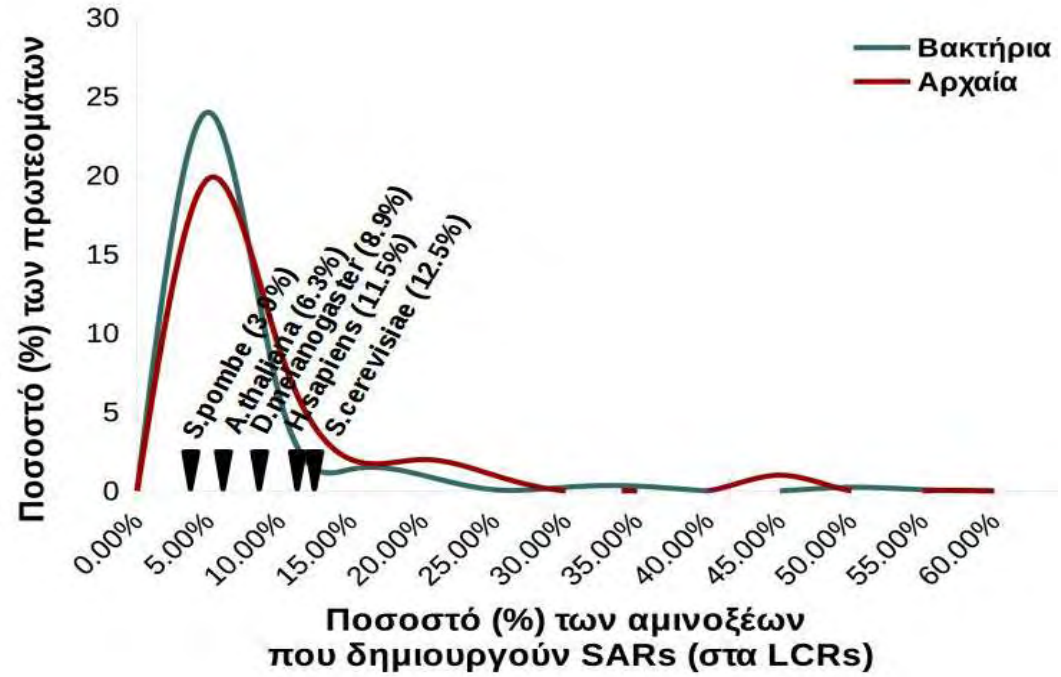
Χρησιμοποιώντας τις μεθόδους που περιγράφηκαν παραπάνω (για συρόμενο παράθυρο μήκους 30 αμινοξέων) ανακτήθηκαν τα πρωτεϊνικά τμήματα LCRs των ευκαρυωτών. Τα αλληλοεπικαλυπτόμενα κομμάτια συγχωνεύθηκαν σε ένα μεγάλο ολοκληρωμένο LCR. Τα διανύσματα για τα αμινοξέα και η εντροπία του Shannon υπολογίστηκαν ξανά για τα συγχωνευμένα LCRs. Ανιχνεύτηκαν επίσης, τα SARs (από τα ενωμένα LCRs).

Έπειτα, υπολογίστηκε ο αριθμός των αμινοξέων σε κάθε πρωτέωμα των ευκαρυωτών και των προκαρυωτών (βακτηρία και αρχαία). Υπολογίστηκε επίσης, ο αριθμός των αμινοξέων στα LCRs και στα SARs για κάθε οργανισμό. Η ποσοστιαία περιεκτικότητα αμινοξέων στα LCRs καθώς και το ποσοστό τους (των αμινοξέων των LCRs) που σχηματίζει SARs στους ευκαρυώτες, συγκρίθηκε με τα αντίστοιχα ποσοστά των προκαρυωτών.

Τα διαγράμματα στην Εικόνα 7 και την Εικόνα 8 δείχνουν την περιεκτικότητα σε αμινοξέα στα LCRs και τα SARs για κάθε κάθε μια μεγάλη εξελικτική γραμμή (τα βέλη αντιπροσωπεύουν τις τιμές κάθε ευκαρυωτικού πρωτεόματος). Όσον αφορά τα SARs, βακτηριακά και αρχαϊκά πρωτεώματα χωρίς SARs (με τιμή 0) δεν φαίνονται στο διάγραμμα (δεν συμπεριλήφθηκαν κατά την κατασκευή του διαγράμματος αλλά συμπεριλήφθηκαν στον υπολογισμό του ποσοστού στα πρωτεώματα).




***Εικόνα 7.*** *Διάγραμμα που δείχνει το ποσοστό των συνολικών αμινοξέων ενός πρωτεώματος που συμμετέχουν στα LCRs*




***Εικόνα 8****. Διάγραμμα που δείχνει το ποσοστό των συνολικών αμινοξέων ενός πρωτεώματος που συμμετέχουν στα SARs*

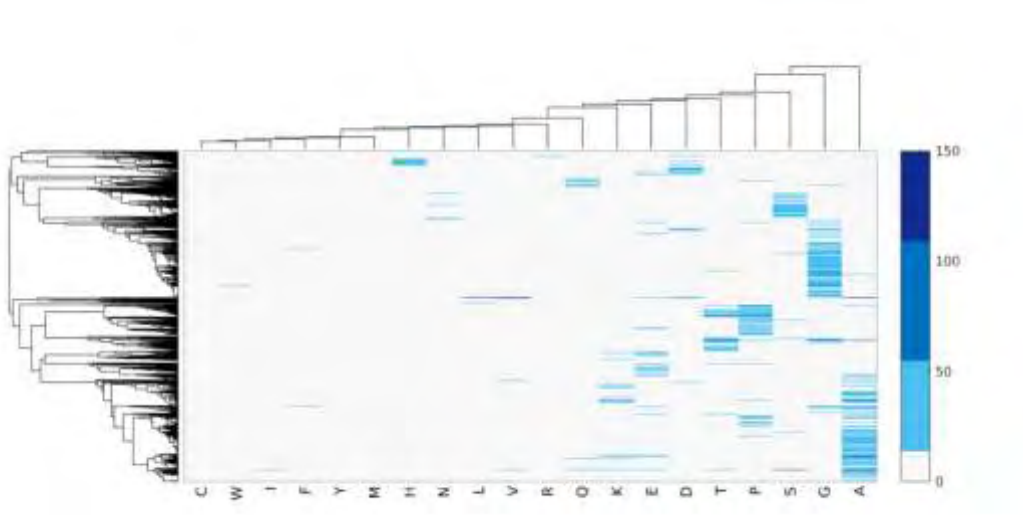
Όσον αφορά τα LCRs, οι ευκαρυώτες, όπως φαίνεται και από τα διαγράμματα στις Εικόνα 7 και Εικόνα 8, έχουν περισσότερα ή και μεγαλύτερα σε μήκος LCRs από τους προκαρυώτες, καθώς κατά μέσο όρο στα βακτηριακά πρωτεώματα τα LCRs αποτελούνται περίπου από το 0.04% των αμινοξέων και στα αρχαϊκά το 0.06%

περίπου. Αντίθετα, στα πρωτεώματα των ευκαρυωτών που αναλύθηκαν, ο οργανισμός με τα λιγότερα LCRs είναι ο *S. pombe*, με το 0.14% των αμινοξέων του να συμμετέχει στη δημιουργία τους. Αν και αυτή η τιμή είναι η μικρότερη μεταξύ των ευκαρυωτικών πρωτεωμάτων είναι κατά πολύ μεγαλύτερη από τους μέσους όρους των προκαρυωτών. Τα SARs των ευκαρυωτών που μελετήθηκαν, είναι κατά μέσο όρο περισσότερα ή και μεγαλύτερα σε μήκος από τα SARs των προκαρυωτών.

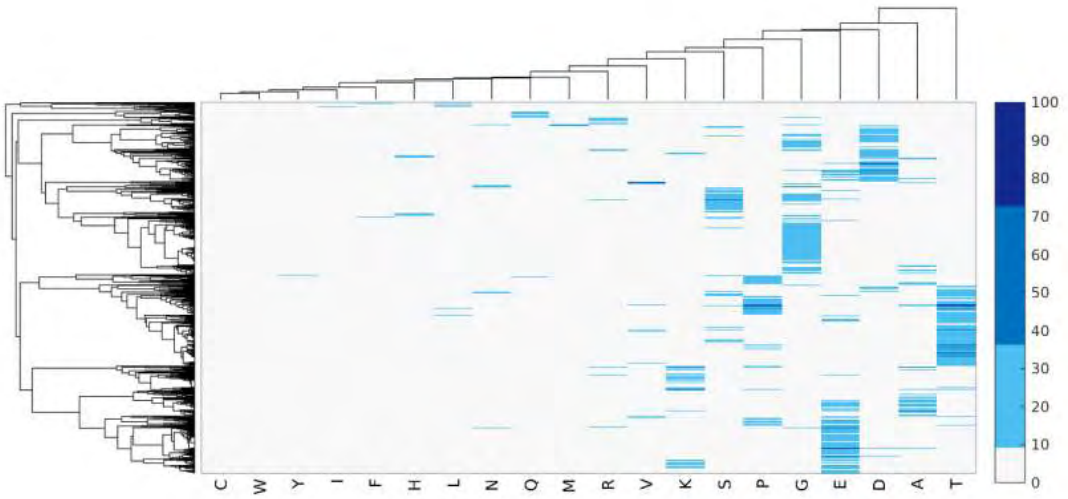
### 3.1.2 Ομαδοποίηση των πρωτεϊνικών τμημάτων

Τα ενωμένα LCRs ομαδοποιήθηκαν για κάθε βασίλειο ξεχωριστά (Βακτήρια, Αρχαία, Βακτηριοφάγοι), χρησιμοποιώντας τα διανύσματα (vectors) για τη συχνότητα των αμινοξέων και τη συνάρτηση clustergram στη Matlab. Πιο συγκεκριμένα, ως μέθοδος απόστασης μεταξύ των διάφορων διανυσμάτων χρησιμοποιήθηκε ο συντελεστή συσχέτισης Pearson. Για την απομάκρυνση του θορύβου, κάθε αμινοξύ εντός της ακολουθίας του LCR, με συχνότητα κάτω από το 10%, έλαβε μηδενική τιμή. Αυτά τα τροποποιημένα διανύσματα χρησιμοποιήθηκαν έπειτα για την ομαδοποίηση.

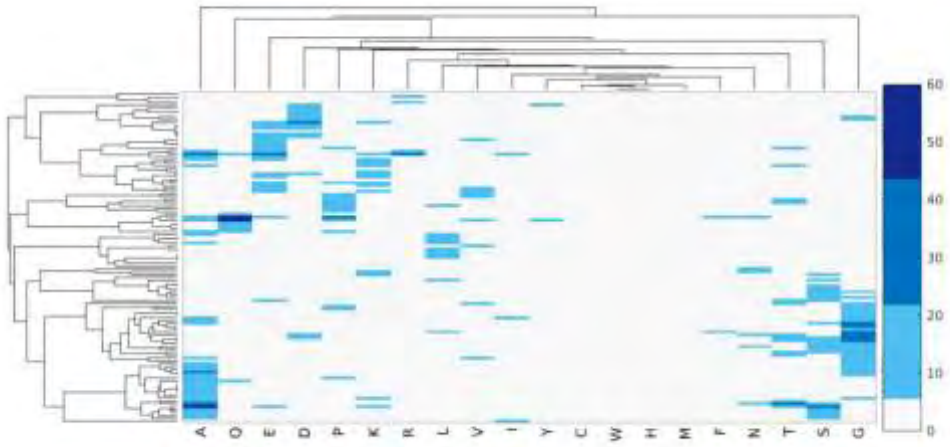
Στα σχήματα της Εικόνα 9, της Εικόνα 10 και Εικόνα 11 φαίνεται η ομαδοποίηση για τα 3 βασίλεια.
Κάθε τελικό κλαδί στο δέντρο που φαίνεται δίπλα από το σχήμα, αντιπροσωπεύει ένα LCR, ενώ οι χρωματισμένες γραμμές που βρίσκονται στην ίδια ευθεία με το κλαδί δείχνουν από ποια αμινοξέα αποτελείται. Η ένταση του χρώματος της κάθε γραμμής είναι ανάλογη με τον αριθμό των αμινοξέων που απαρτίζουν το LCR. Έτσι, αν ένα LCR αποτελείται συνολικά από 70 αλανίνες (A), 160 γλυκίνες (G) και 20 προλίνες (P), στο σχήμα θα φαίνεται, στην ίδια ευθεία, μια μπλε ανοιχτή γραμμή στη στήλη της αλανίνης, μια μπλε σκούρα γραμμή στη στήλη της γλυκίνης και τέλος μια γαλάζια γραμμή στη στήλη της προλίνης.




***Εικόνα 9.*** *Ομαδοποίηση για τα Βακτηριακά LCRs (για συρόμενο παράθυρο μήκους 30 αα)*




***Εικόνα 10.*** *Ομαδοποίηση για τα Αρχαϊκά LCRs (για συρόμενο παράθυρο μήκους 30 αα)*




***Εικόνα 11.*** *Ομαδοποίηση για τα LCRs των Βακτηριοφάγων (για συρόμενο παράθυρο μήκους 30 αα)*

Όπως φαίνεται από τα σχήματα, το αμινοξύ που εμφανίζεται πολύ συχνά στα LCRs και των τριών βασιλείων (βακτήρια, αρχαία, βακτηριοφάγοι) είναι η γλυκίνη (και λιγότερο η αλανίνη).

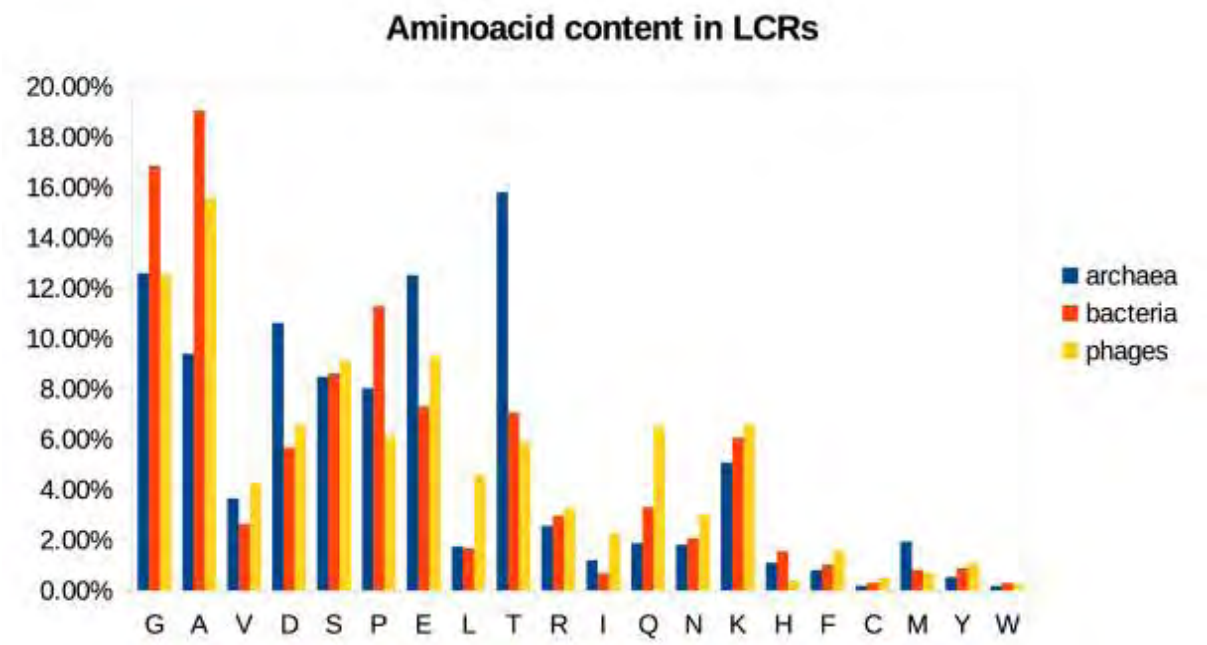
Όσον αφορά τα βακτήρια, τα LCRs στην πλειοψηφία τους αποτελούνται από γλυκίνη και αλανίνη, τα οποία αποτελούν και τις δύο μεγαλύτερες ομάδες.

Στα αρχαία, η πλειοψηφία των LCRs αποτελείται από θρεονίνη, γλυκίνη και γλουταμικό οξύ.

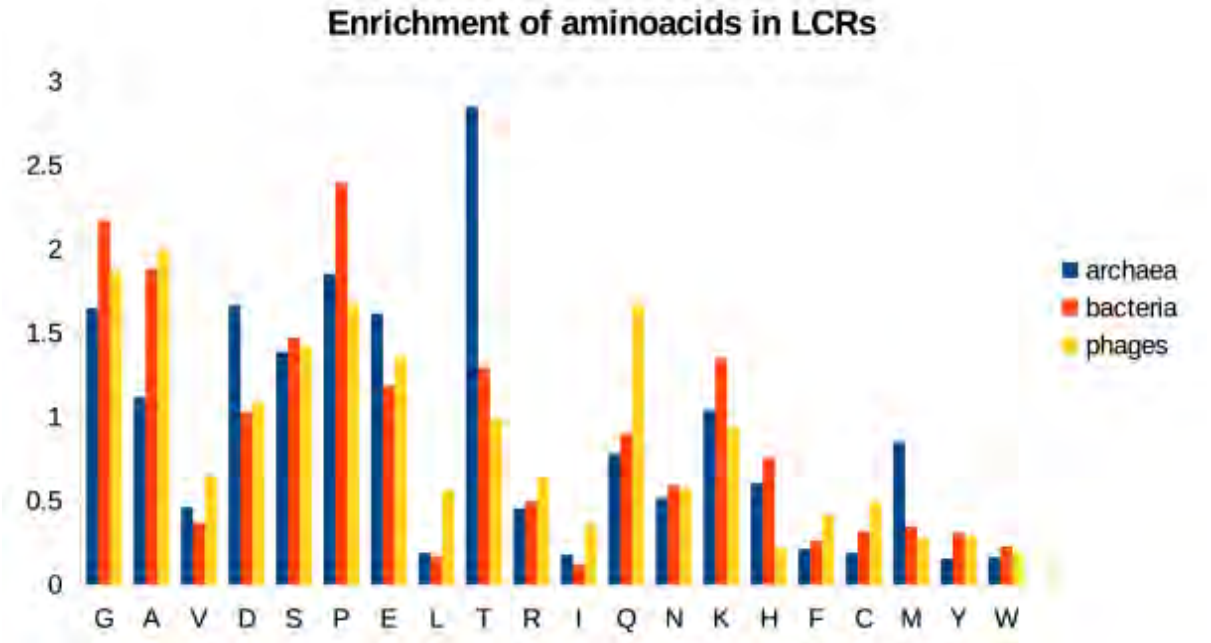
Τέλος, στους βακτηριοφάγους τα περισσότερα LCRs αποτελούνται από αλανίνη και γλυκίνη.

### 3.1.3 Συχνότητα των αμινοξέων και εμπλουτισμός των LCRs σε αμινοξέα

Στη συνέχεια υπολογίστηκε η συχνότητα (ποσοστό %) του κάθε αμινοξέος στα LCRs και συγκρίθηκε με τη συχνότητα υποβάθρου (ποσοστό %) του κάθε αμινοξέος στα πρωτεώματα (π.χ. συχνότητα αλανίνης στα LCRs/ συχνότητα αλανίνης στα πρωτεώματα). Με αυτό τον τρόπο, υπολογίστηκε ο εμπλουτισμός (enrichment) του κάθε αμινοξέος στα LCRs. Το διάγραμμα στην Εικόνα 12 συνοψίζει ποια αμινοξέα εμφανίζονται πιο συχνά στα LCRs και της Εικόνα 13, πόσο πιο συχνά εμφανίζονται στα LCRs σε σχέση με το υπόλοιπο πρωτέωμα (εμπλουτισμός).




***Εικόνα 12.*** *Συχνότητα των αμινοξέων στα LCRs*




***Εικόνα 13.*** *Εμπλουτισμός των LCRs σε αμινοξέα*

Όπως φαίνεται στα διαγράμματα των Εικόνα 12 και Εικόνα 13, η γλυκίνη είναι πολύ συχνό αμινοξύ στα LCRs των βακτηρίων, η αλανίνη στα LCRs των βακτηριοφάγων και η θρεονίνη στα LCRs των αρχαίων. Συγκεκριμένα, στα βακτηριακά LCRs η γλυκίνη είναι πάνω από 2 φορές πιο συχνή σε σχέση με το αναμενόμενο. Η γλυκίνη, η αλανίνη και η προλίνη φαίνεται να είναι και στα 3 βασίλεια, πολύ συχνότερες στα LCRs από ότι αναμένεται καθαρά από τύχη. Τέλος στα αρχαία η θρεονίνη φαίνεται να είναι το συχνότερο και πιο εμπλουτισμένο αμινοξύ (πάνω από 2,5 φορές συχνότερη στα LCRs συγκριτικά με το υπόβαθρο).

Με βάση τις εξελικτικές μελέτες του Trifonov για τον πρώιμο γενετικό κώδικα (Trifonov, 2009, 2000), τα πρώτα γονίδια ήταν μικρά δίκλωνα μόρια, που αποτελούνταν από τριπλέτες GGC και GCC και κωδικοποιούσαν για γλυκίνη (G) και αλανίνη (A) αντίστοιχα. Επομένως, τα ολιγοπεπτίδια που σχηματίστηκαν από το πρώιμο σύστημα μετάφρασης αποτελούνταν από αυτά τα δύο αμινοξέα. Το αμινοξικό περιεχόμενο των LCRs που βρήκαμε στα 3 βασίλεια των προκαρυωτών είναι σύμφωνο με αυτή τη θεωρία του Trifonov.

Επιπλέον, τα αμινοξέα που αποτελούν αυτές τις περιοχές κατά κύριο λόγο, είναι τα αμινοξέα που πιθανολογείται ότι εμφανίστηκαν πρώτα στο γενετικό κώδικα. Στα διαγράμματα, η συχνότητα των αμινοξέων σχετίζεται σε κάποιο βαθμό με τη σειρά, που πιθανολογείται ότι εμφανίστηκαν στο γενετικό κώδικα. Φαίνεται ότι σε γενικές γραμμές, όσο πιο “παλιό” είναι ένα αμινοξύ τόσο πιο συχνό είναι στα LCRs. Αμινοξέα όπως η μεθειονίνη, η τυροσίνη και η τρυπτοφάνη που εμφανίστηκαν πιο πρόσφατα στο γενετικό κώδικα έχουν και χαμηλή συχνότητα στα LCRs.

Οι περιοχές χαμηλής πολυπλοκότητας, λόγω της ικανότητάς τους να επεκτείνονται και να συρρικνώνονται πολύ γρήγορα, θα μπορούσαν να έχουν μεγάλο αντίκτυπο στο ενεργειακό φορτίο της μετάφρασης, ειδικά στους προκαρυώτες (Akashi and

Gojobori, 2002; Barton et al., 2010). Πιο συγκεκριμένα, μια αντικατάσταση ενός αμινοξέως σε ένα υψηλά εκφραζόμενο γονίδιο στο E. coli ή στο B. Subtilis μπορεί να προκαλέσει αύξηση ή μείωση του συνολικού ενεργειακού προϋπολογισμού για τη βιοσύνθεση μακρομορίων περισσότερο από 0,025%. Επομένως, μεγάλες επαναλήψεις, ειδικά όσες ανιχνεύονται σε υψηλά εκφραζόμενες πρωτεΐνες, όπως οι πρωτεΐνες του ριβοσώματος ή όσες επαναλήψεις περιέχουν αμινοξέα με υψηλό κόστος βιοσύνθεσης, όπως τα F, H, W, Y αυξάνουν το ενεργειακό φορτίο της μετάφρασης.

### 3.1.4 Οργανισμοί με μεγάλη περιεκτικότητα σε LCRs

Η πρωτεΐνη με τη μεγαλύτερη επανάληψη μεταξύ των βακτηριακών LCRs ήταν η Αιμαγλουτινίνη από τον οργανισμό *Haemophilus pittmaniae*, με μήκος 1110 αμινοξέα του επαναλαμβανόμενου μοτίβου SATAASTSASAADSSATAASSSASAADSSAX (με το X να είναι K ή Q).

Όσον αφορά τα αρχαία, η πρωτεΐνη με το μεγαλύτερο LCR είναι μια μη χαρακτηρισμένη πρωτεΐνη από τον οργανισμό *Methanocella arvoryzae*, με μια επανάληψη πλούσια σε αλανίνη με μήκος 390 αμινοξέα.

Για τους βακτηριοφάγους, η πρωτεΐνη με το μεγαλύτερο LCR είναι η Gp39 και ανήκει στον οργανισμό *Corynebacterium phage BFK20*. Το LCR έχει μήκος 135 αμινοξέα και είναι πλούσιο σε γλουταμίνη και προλίνη.

Οι οργανισμοί που διαθέτουν τις περισσότερες πρωτεΐνες με LCRs ανήκουν σε μια τάξη βακτηρίων, τα Myxococcales. Η τάξη αυτή ανήκει στα δ-πρωτεοβακτήρια. Τα βακτήρια αυτής της ομάδας ζουν κατά κύριο λόγο στο έδαφος, τρέφονται με αδιάλυτες οργανικές ουσίες και μπορούν να μετακινούνται με ολίσθηση. Οργανώνονται και ταξιδεύουν συνήθως σε συσσωματώματα πολλών κυττάρων, που επικοινωνούν μέσω διακυτταρικών σημάτων. Μέσω της συσσωμάτωσης καθίσταται δυνατή η συσσώρευση εξωκυττάριων ενζύμων που χρησιμοποιούνται για την πέψη των ουσιών που αποτελούν την τροφή τους, κι έτσι η τροφή είναι εύκολα διαθέσιμη σε περισσότερα κύτταρα (Kiskowski et al., 2004). Επίσης, παράγουν και εκκρίνουν πολλές χημικές ουσίες και δευτερογενείς μεταβολίτες, χρήσιμα στη βιοϊατρική και τη βιομηχανία, όπως τα αντιβιοτικά (Reichenbach, 2001).

Στη συνέχεια, έγινε σύγκριση του συνολικού αριθμού των LCRs για κάθε οργανισμό με το συνολικό αριθμό των πρωτεϊνών του, ώστε να υπολογιστεί η συχνότητα των LCRs σε κάθε πρωτέωμα. Έπειτα, η συχνότητα των LCRs σε κάθε πρωτέωμα συγκρίθηκε με το μέσο όρο της συχνότητας (σε όλα τα πρωτεόματα), για να υπολογιστεί ο εμπλουτισμός του κάθε πρωτεόματος σε LCRs. Οι 10 οργανισμοί με τη μεγαλύτερη συχνότητα σε LCRs για τα βακτήρια, τα αρχαία και τους βακτηριοφάγους φαίνονται στον Πίνακας 3, τον Πίνακας 4 και τον Πίνακας 5 μαζί με το Φύλο, την Κλάση, το ταξινομικό τους αναγνωριστικό (ID), το συνολικό αριθμό των πρωτεϊνών τους, τη συχνότητα και τον εμπλουτισμό τους σε LCRs.

Οι οργανισμοί που έχουν επισημανθεί (κίτρινο) βρίσκονται μεταξύ των 10 οργανισμών που έχουν τα περισσότερα LCRs.

| Φύλο | Κλάση | Ταξινομ. ID | Οργανισμός | Συνολικός αριθμός LCRs | Συνολικός αριθμός πρωτεϊνών | Συχνότη-τα | Εμπλουτι-σμός |
|---|---|---|---|---|---|---|---|
| Euryarchaeota | Halobacteria | 547559 | Natrialba magadii | 109 | 4203 | 2.59 | 4.48 |
| Euryarchaeota | X | 1803821 | Euryarchaeota archaeon | 35 | 1509 | 2.32 | 4.01 |
| Thaumarchaeota | Nitrosopumilales | 436308 | Nitrosopumilus maritimus | 38 | 1795 | 2.12 | 3.66 |
| Euryarchaeota | Methanomicrobia | 351160 | Methanocella arvoryzae | 50 | 3071 | 1.63 | 2.81 |
| Crenarchaeota | Thermoprotei | 399550 | Staphylothermus marinus | 25 | 1570 | 1.59 | 2.75 |
| Crenarchaeota | Thermoprotei | 397948 | Caldivigra maquilingensis | 30 | 1962 | 1.53 | 2.64 |
| Thaumarchaeota | Cenarchaeales | 414004 | Cenarchaeum symbiosum | 28 | 2022 | 1.38 | 2.39 |
| X | X | 1579370 | Archaeon GW2011_AR10 | 16 | 1339 | 1.19 | 2.06 |
| Euryarchaeota | Halobacteria | 1333523 | Salinarchaeum sp | 36 | 3013 | 1.19 | 2.06 |
| Crenarchaeota | Thermoprotei | 666510 | Acidilobus saccharovorans | 16 | 1499 | 1.07 | 1.84 |

***Πίνακας 3.*** *Αρχαία: Οι 10 οργανισμοί με τον υψηλότερο εμπλουτισμό σε LCRs*

| Φύλο | Κλάση | Ταξινομ. ID | Οργανισμός | Συνολικός αριθμός LCRs | Συνολικός αριθμός πρωτεϊνών | Συχνότη-τα | Εμπλουτι-σμός |
|---|---|---|---|---|---|---|---|
| Proteobacteria | δ-proteobacteria | 215803 | Enhygromyxa salina | 214 | 8157 | 2.62 | 6.37 |
| Proteobacteria | δ-proteobacteria | 1192034 | Chondromyces apiculatus | 224 | 9037 | 2.48 | 6.02 |
| Proteobacteria | δ-proteobacteria | 391625 | Plesiocystis pacifica | 202 | 8437 | 2.39 | 5.81 |
| Fibrobacteres | Fibrobacterales | 59374 | Fibrobacter succinogenes | 64 | 2871 | 2.23 | 5.41 |
| Proteobacteria | δ-proteobacteria | 502025 | Haliangium ochraceum | 144 | 6684 | 2.15 | 5.23 |
| Candidatus | Berkelbacteria | 1618337 | Berkelbacteria bacterium | 20 | 991 | 2.02 | 4.9 |
| Actinobacteria | Coriobacteriia | 79604 | Denitobacterium detoxificans | 33 | 1731 | 1.91 | 4.63 |
| Proteobacteria | δ-proteobacteria | 1254432 | Sorangium cellulosum | 192 | 10372 | 1.85 | 4.49 |
| Proteobacteria | γ-proteobacteria | 1385625 | Uncultured Termite | 14 | 767 | 1.83 | 4.43 |
| Proteobacteria | β-proteobacteria | 76731 | Roseateles depolymerans | 87 | 4768 | 1.82 | 4.43 |

***Πίνακας 4.*** *Βακτήρια: Οι 10 οργανισμοί με τον υψηλότερο εμπλουτισμό σε LCRs*

| Φύλο | Κλάση | Ταξινομ. ID | Οργανισμός | Συνολικός αριθμός LCRs | Συνολικός αριθμός πρωτεϊνών | Συχνότητα | Εμπλουτι-σμός |
|---|---|---|---|---|---|---|---|
| Microviridae | Gokushovirinae | 10857 | Chlamydia phage 1 | 2 | 8 | 25.0 | 5.91 |
| Leviviridae | Levivirus | 12018 | Enterobacteria phage GA | 1 | 4 | 25.0 | 5.91 |
| Inoviridae | X | 10871 | Pseudomonas phage Pf1 | 3 | 14 | 21.43 | 5.06 |
| Inoviridae | X | 10869 | Enterobacteria phage I2-2 | 1 | 8 | 12.5 | 2.95 |
| Inoviridae | Lineavirus | 10867 | Enterobacteria phage IKe | 1 | 9 | 11.11 | 2.63 |
| Inoviridae | X | 10868 | Enterobacteria phage If1 | 1 | 10 | 10.0 | 2.36 |
| X | Sphaerolipoviridae | 1714272 | Thermus thermophilus bacteriophage | 2 | 37 | 5.41 | 1.28 |
| Caudovirales | Siphoviridae | 10724 | Bacillus phage SPP1 | 5 | 97 | 5.15 | 1.22 |
| Caudovirales | Podoviridae | 10761 | Shigella phage Sf6 | 3 | 66 | 4.55 | 1.07 |
| Caudovirales | Siphoviridae | 467481 | Azospirillum phage | 4 | 95 | 4.21 | 0.99 |

***Πίνακας 5.*** *Βακτηριοφάγοι: Οι 10 οργανισμοί με τον υψηλότερο εμπλουτισμό σε LCRs*

### 3.1.5 Ανάλυση κωδικονίων

Από τη βάση δεδομένων της Uniprot, ανακτήθηκαν τα αναγνωριστικά που δίνονται από την EMBL-Bank (EMBL-Bank IDs) για τις κωδικές ακολουθίες των πρωτεϊνών (CDSs) και χρησιμοποιώντας τα ανακτήθηκαν οι αντίστοιχες γονιδιακές ακολουθίες. Λήφθηκαν συνολικά 40169 κωδικές αλληλουχίες βακτηρίων, 2480 αρχαίων και 106 βακτηριοφάγων, αλλά εξαιτίας μετατοπίσεων αναγνωστικού πλαισίου και σφαλμάτων κατά την ανάκτηση, τελικά χρησιμοποιήθηκαν 39519 κωδικές ακολουθίες βακτηρίων, 2457 αρχαίων και 104 βακτηριοφάγων. Για κάθε LCR υπολογίστηκε η συχνότητα των αμινοξέων και των κωδικονίων του.

Για κάθε αμινοξύ υπολογίστηκε η συχνότητα των κωδικονίων του στα LCRs. Υπολογίστηκε, επιπλέον, η συχνότητα των αμινοξέων και των κωδικονίων για όλους τους οργανισμούς με υψηλό εμπλουτισμό σε LCRs (> 3 φορές), που φαίνονται στους παραπάνω πίνακες και για τα 3 βασίλεια. Όσον αφορά τα βακτήρια, αναλύθηκε επιπλέον και η τάξη των Myxococcales.

Τα αμινοξέα και τα κωδικόνιά τους, μαζί με τις αντίστοιχες συχνότητές τους φαίνονται στον Πίνακας 6. Τα κυρίαρχα κωδικόνια για κάθε βασίλειο έχουν επισημανθεί (κίτρινο).

| | | Βακτήρια | | | | Αρχαία | | | | Φάγοι | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | κωδικόνιο | Συχνότ. κωδικονίου | αμινοξύ | Συχνότ. αμινοξέος | κωδικόνιο | Συχνότ. κωδικονίου | αμινοξύ | Συχνότ αμινοξέος | κωδικόνιο | Συχνότ. κωδικονίου | αμινοξύ | Συχνότ. αμινοξέος |
| A | GCA | 30162 | 3.7 | 155931 | 19.1 | 1462 | 2.7 | 5090 | 9.43 | 134 | 3.7 | 525 | 14.5 |
| | GCC | 51403 | 6.3 | | | 1406 | 2.6 | | | 141 | 3.9 | | |
| | GCG | 38125 | 4.7 | | | 999 | 1.85 | | | 95 | 2.6 | | |
| | GCT | 36241 | 4.4 | | | 1223 | 2.26 | | | 155 | 4.3 | | |
| C | TGC | 1467 | 0.18 | 2351 | 0.3 | 49 | 0.09 | 97 | 0.18 | 7 | 0.19 | 16 | 0.44 |
| | TGT | 884 | 0.1 | | | 48 | 0.09 | | | 9 | 0.25 | | |
| D | GAC | 25556 | 3.1 | 46078 | 5.6 | 4060 | 7.5 | 5761 | 10.7 | 126 | 3.5 | 247 | 6.84 |
| | GAT | 20522 | 2.5 | | | 1701 | 3.15 | | | 121 | 3.35 | | |
| E | GAA | 33677 | 4.1 | 59761 | 7.3 | 3483 | 6.45 | 6738 | 12.5 | 241 | 6.67 | 342 | 9.46 |
| | GAG | 26084 | 3. 2 | | | 3255 | 6.03 | | | 101 | 2.8 | | |
| F | TTC | 4877 | 0.6 | 8155 | 1 | 249 | 0.46 | 405 | 0.75 | 33 | 0.9 | 58 | 1.6 |
| | TTT | 3278 | 0.4 | | | 156 | 0.29 | | | 25 | 0.7 | | |
| G | GGA | 19937 | 2.4 | 138434 | 16.9 | 1267 | 2.3 | 6811 | 12.6 | 86 | 2.4 | 469 | 13 |
| | GGC | 68670 | 8.4 | | | 3044 | 5.6 | | | 126 | 3.49 | | |
| | GGG | 12333 | 1.5 | | | 609 | 1.13 | | | 53 | 1.47 | | |
| | GGT | 37494 | 4.6 | | | 1891 | 3.5 | | | 204 | 5.6 | | |
| H | CAC | 6768 | 0.8 | 12911 | 1.6 | 436 | 0.8 | 591 | 1.1 | 6 | 0.16 | 14 | 0.39 |
| | CAT | 6143 | 0.75 | | | 155 | 0.29 | | | 8 | 0.22 | | |
| I | ATA | 1185 | 0.1 | 5499 | 0.67 | 248 | 0.46 | 643 | 1.2 | 17 | 0.47 | 85 | 2.35 |
| | ATC | 2130 | 0.26 | | | 177 | 0.33 | | | 24 | 0.66 | | |
| | ATT | 2184 | 0.27 | | | 218 | 0.4 | | | 44 | 1.22 | | |
| K | AAA | 21405 | 2.6 | 49454 | 6 | 1262 | 2.3 | 2738 | 5.07 | 142 | 3.93 | 242 | 6.7 |
| | AAG | 28049 | 3.4 | | | 1476 | 2.7 | | | 100 | 2.77 | | |
| L | CTA | 940 | 0.1 | 12861 | 1.6 | 111 | 0.2 | 912 | 1.7 | 14 | 0.39 | 167 | 4.62 |
| | CTC | 2388 | 0.3 | | | 217 | 0.4 | | | 26 | 0.72 | | |
| | CTG | 4083 | 0.5 | | | 235 | 0.4 | | | 55 | 1.52 | | |
| | CTT | 1901 | 0.2 | | | 124 | 0.23 | | | 28 | 0.77 | | |
| | TTA | 1608 | 0.2 | | | 117 | 0.2 | | | 27 | 0.75 | | |
| | TTG | 1941 | 0.2 | | | 108 | 0.2 | | | 17 | 0.47 | | |
| M | ATG | 6538 | 0.8 | 6538 | 0.8 | 1036 | 1.9 | 1036 | 1.92 | 26 | 0.72 | 26 | 0.72 |
| N | AAC | 9466 | 1.1 | 17138 | 2.1 | 578 | 1.07 | 984 | 1.82 | 56 | 1.55 | 111 | 3.07 |
| | AAT | 7672 | 0.9 | | | 406 | 0.75 | | | 55 | 1.52 | | |
| P | CCA | 17382 | 2.1 | 92502 | 11.3 | 1312 | 2.4 | 4349 | 8.05 | 39 | 1.08 | 232 | 6.42 |
| | CCC | 22286 | 2.7 | | | 874 | 1.6 | | | 31 | 0.86 | | |
| | CCG | 36247 | 4.4 | | | 1267 | 2.35 | | | 56 | 1.55 | | |
| | CCT | 16587 | 2 | | | 896 | 1.66 | | | 106 | 2.93 | | |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q | CAA | 10563 | 1.3 | 27175 | 3.3 | 368 | 0.68 | 1016 | 1.9 | 118 | 3.27 | 242 | 6.7 |
| | CAG | 16612 | 2 | | | 648 | 1.2 | | | 124 | 3.43 | | |
| R | AGA | 3414 | 0.4 | 23791 | 2.9 | 365 | 0.68 | 1371 | 2.54 | 30 | 0.83 | 118 | 3.27 |
| | CGA | 1527 | 0.2 | | | 151 | 0.28 | | | 27 | 0.75 | | |
| | CGC | 1726 | 0.2 | | | 209 | 0.39 | | | 26 | 0.72 | | |
| | CGG | 8377 | 1 | | | 194 | 0.36 | | | 7 | 0.19 | | |
| | CGT | 3426 | 0.4 | | | 174 | 0.3 | | | 21 | 0.58 | | |
| | AGG | 5321 | 0.6 | | | 278 | 0.5 | | | 7 | 0.19 | | |
| S | AGC | 17094 | 2.1 | 70815 | 8.7 | 944 | 1.75 | 4574 | 8.47 | 43 | 1.19 | 298 | 8.25 |
| | AGT | 9483 | 1.1 | | | 794 | 1.47 | | | 50 | 1.38 | | |
| | TCA | 8903 | 1.1 | | | 772 | 1.43 | | | 50 | 1.38 | | |
| | TCC | 13087 | 1.6 | | | 684 | 1.27 | | | 34 | 0.94 | | |
| | TCG | 12437 | 1.5 | | | 749 | 1.39 | | | 34 | 0.94 | | |
| | TCT | 9811 | 1.2 | | | 631 | 1.17 | | | 87 | 2.41 | | |
| T | ACA | 12259 | 1.5 | 57521 | 7 | 2471 | 4.57 | 8512 | 15.76 | 58 | 1.6 | 211 | 5.84 |
| | ACC | 18587 | 2.3 | | | 2064 | 3.8 | | | 39 | 1.08 | | |
| | ACG | 16603 | 2 | | | 1829 | 3.39 | | | 37 | 1.02 | | |
| | ACT | 10072 | 1.2 | | | 2148 | 3.98 | | | 77 | 2.13 | | |
| V | GTA | 3971 | 0.5 | 20644 | 2.5 | 433 | 0.8 | 1991 | 3.7 | 42 | 1.16 | 159 | 4.4 |
| | GTC | 5332 | 0.6 | | | 536 | 0.99 | | | 33 | 0.91 | | |
| | GTG | 6240 | 0.76 | | | 499 | 0.92 | | | 31 | 0.86 | | |
| | GTT | 5101 | 0.6 | | | 523 | 0.97 | | | 53 | 1.47 | | |
| W | TGG | 2426 | 0.3 | 2426 | 0.3 | 93 | 0.17 | 93 | 0.17 | 10 | 0.28 | 10 | 0.28 |
| Y | TAC | 3761 | 0.46 | 7120 | 0.9 | 190 | 0.35 | 282 | 0.52 | 19 | 0.53 | 41 | 1.14 |
| | TAT | 3359 | 0.4 | | | 92 | 0.17 | | | 22 | 0.61 | | |
| | | 817105 | | 817105 | | 53994 | | 53994 | | 3613 | | 3613 | |

***Πίνακας 6.*** *Βακτήρια, αρχαία, βακτηριοφάγοι: Αριθμός των αμινοξέων και των κωδικονίων τους στα LCRs, μαζί με τις αντίστοιχες συχνότητές τους. Τα επισημασμένα (κίτρινο) κωδικόνια είναι τα κυρίαρχα σε κάθε βασίλειο*

Όπως έχει προαναφερθεί, τα πρώτα γονίδια, πιθανότατα, ήταν μικρά δίκλωνα μόρια, που αποτελούνταν από τριπλέτες GGC, που κωδικοποιεί για τη γλυκίνη (G) και GCC, που κωδικοποιεί για την αλανίνη (A) (Trifonov, 2009, 2000). Η θεωρία ότι αυτές οι περιοχές ίσως να προέρχονται από αρχέγονες πρωτεΐνες επιβεβαιώνεται και από τη συχνότητα των κωδικονίων που κωδικοποιούν για τη γλυκίνη και την αλανίνη. Η τριπλέτα GGC είναι η συχνότερη τόσο στα βακτηριακά LCRs όσο και στα LCRs των αρχαίων. Επίσης, το κωδικόνιο GCC είναι αυτό που κωδικοποιεί για τις περισσότερες αλανίνες στα LCRs των βακτηρίων αλλά όχι στα LCRs των αρχαίων.

### 3.1.6 Επαναλήψεις ενός αμινοξέος (Single Aminoacid Repeats, SARs)

Κατασκευάστηκαν perl scripts για την ανίχνευση περιοχών χαμηλής πολυπλοκότητας που δημιουργούνται από την επανάληψη του ίδιου αμινοξέος 10 ή περισσότερες φορές. Συνολικά ανιχνεύτηκαν 1052 SARs στα πρωτεώματα των βακτηρίων, 107 στα πρωτεώματα των αρχαίων και 2 στων βακτηριοφάγων. Για κάθε SAR δημιουργήθηκε ένα διάνυσμα για τα κωδικόνια που το κωδικοποιούν.

Στη συνέχεια, για κάθε αμινοξύ ανιχνεύτηκε η μεγαλύτερη σε μήκος επανάληψη που δημιουργεί, η πρωτεΐνη στην οποία βρίσκεται και ο οργανισμός στον οποίο ανήκει αυτή η πρωτεΐνη. Υπολογίστηκε, επίσης, ο συνολικός αριθμός των SARs σε κάθε οργανισμό. Τέλος, για κάθε αμινοξύ, υπολογίστηκε η συχνότητα των κωδικονίων του, πόσες επαναλήψεις δημιουργεί και σε πόσες πρωτεΐνες και οργανισμούς έχουν ανιχνευτεί αυτές οι επαναλήψεις.

Μερικά αμινοξέα φαίνεται πως δεν σχηματίζουν SARs σε κανένα από τα 3 βασίλεια. Αυτά τα αμινοξέα είναι η μεθειονίνη (M), η κυστεΐνη (C), η τρυπτοφάνη (W) και η τυροσίνη (Y).

Τα 5 μεγαλύτερα σε μήκος SARs για τα βακτήρια, τα αρχαία και τους φάγους, μαζί με το μέγεθός τους, το όνομα του οργανισμού και της πρωτεΐνης στην οποία ανήκουν φαίνονται στους Πίνακας 7, Πίνακας 8 και Πίνακας 9.

| **Βακτήρια** | | | |
|---|---|---|---|
| **αα** | **Μέγεθος** | **Οργανισμός** | **Πρωτεΐνη** |
| D | 256 | Desulfocapsa sulfexigens | Uncharacterized protein |
| G | 119 | Tepidimonas fonticaldi | Uncharacterized protein |
| P | 101 | Acidithiobacillus ferrivorans | Uncharacterized protein |
| S | 56 | Teredinibacter turnerae | Putative lipoprotein |
| Q | 48 | Coxiellaceae bacterium | RNase III inhibitor |

***Πίνακας 7.*** *Βακτήρια: Τα 5 μεγαλύτερα σε μήκος SARs, το όνομα του οργανισμού και της πρωτεΐνης στην οποία βρέθηκαν*

| **Αρχαία** | | | |
|---|---|---|---|
| **αα** | **Μέγεθος** | **Οργανισμός** | **Πρωτεΐνη** |
| E | 36 | Salinarchaeum sp. | Uncharacterized protein |
| T | 32 | Staphylothermus marinus | Uncharacterized protein |

| | | | |
|---|---|---|---|
| D | 27 | Natronococcus occultus | Uncharacterized protein |
| G | 21 | Haloarcula marismortui | Putative lipoprotein |
| S | 21 | Methanobrevibacter millerae | Putative oligopeptide transport system substrate-binding protein/Membrane-anchored protein predicted to be involved in regulation of amylopullulanase-like protein |

***Πίνακας 8**. Αρχαία: Τα 5 μεγαλύτερα σε μήκος SARs, το όνομα του οργανισμού και της πρωτεΐνης στην οποία βρέθηκαν*

| **Βακτηριοφάγοι** | | | |
|---|---|---|---|
| **αα** | **Μέγεθος** | **Οργανισμός** | **Πρωτεΐνη** |
| S | 13 | Synechococcus phage syn9 | Tail length tape measure protein |
| G | 10 | Phage phiJL001 | Gp185 |

***Πίνακας 9.** Φάγοι: Τα μεγαλύτερα σε μήκος SARs, το όνομα του οργανισμού και της πρωτεΐνης στην οποία βρέθηκαν*

Ο συνολικός αριθμός των SARs σε κάθε οργανισμό φαίνεται στον Πίνακας 10 που ακολουθεί. Μετρήθηκε τόσο ο συνολικός αριθμός των πρωτεϊνών, στις οποίες ανιχνεύονται SARs, όσο και των πρωτεϊνικών τμημάτων, καθώς μια πρωτεΐνη μπορεί να έχει περισσότερα από ένα SARs, τα οποία μπορεί να βρίσκονται διάσπαρτα κατά μήκος της ακολουθίας ή να παρεμβάλλεται μεταξύ τους ένα μικρό πρωτεϊνικό κομμάτι (και να μην είναι συνεχόμενα) είτε να είναι συνεχόμενα αλλά να αποτελούνται από διαφορετικά αμινοξέα (π.χ. μια επανάληψη πολυ-σερίνης ακολουθούμενη από μια επανάληψη πολυ-προλίνης: ..SSSSSPPPPPP...).

| **Βακτήρια** | | | **Αρχαία** | | | **Φάγοι** | | |
|---|---|---|---|---|---|---|---|---|
| Οργανισμός | Πρωτεΐνες | SARs | Οργανισμός | Πρωτεΐνες | SARs | Οργανισμός | Πρωτεΐνες | SARs |
| Teredinibacter turnerae | 45 | 136 | Staphylothermus marinus | 7 | 12 | Synechococcus phage syn9 | 1 | 1 |
| Saccharophagus degradans | 40 | 88 | Thermogladius cellulolyticus | 6 | 7 | Phage phiJL001 | 1 | 1 |
| Ilumatobacter coccineus | 16 | 17 | Desulfurococcus kamchatkensis | 5 | 12 | - | - | - |

***Πίνακας 10.** Βακτήρια, Αρχαία, Βακτηριοφάγοι: Οι 3 οργανισμοί με τα περισσότερα SARs, ο αριθμός των πρωτεϊνών και των πρωτεϊνικών τμημάτων τους, που περιέχουν SARs*

Για κάθε αμινοξύ, η συχνότητα του κυρίαρχου (πιο συχνού) και του δεύτερου πιο συχνού κωδικονίου, ο αριθμός των SARs που δημιουργεί και ο αριθμός των ορθόλογων πρωτεϊνών και των οργανισμών που ανιχνεύτηκαν τα SARs,

παρουσιάζεται στους Πίνακας 11, Πίνακας 12 και Πίνακας 13. Τα αμινοξέα που χρησιμοποιούν 2 κωδικόνια για τις επαναλήψεις έχουν επισημανθεί.

| Βακτήρια | | | | | |
|---|---|---|---|---|---|
| **AA** | **Επαναλήψεις** | **Οργανισμοί** | **Ορθόλογες πρωτεΐνες** | **Κυρίαρχο κωδικόνιο** | **Κυρίαρχο κωδικόν. % (2ο συχνότερο κωδικ.%)** |
| A | **51** | **44** | **24** | GCC | 29.1 (24.9) |
| D | 73 | 53 | 27 | GAC | 52 (48) |
| E | 20 | 17 | 11 | GAA | 61 (39) |
| F | 1 | 1 | 1 | TTC | 97 (3) |
| G | 260 | 197 | 76 | GGC | 50 (29) |
| H | 12 | 10 | 7 | CAT | 50.3 (49.7) |
| I | 2 | 2 | 1 | ATT | 74 (22) |
| K | 6 | 4 | 4 | AAG | 77 (23) |
| L | 1 | 1 | 1 | TTG | 100 (0) |
| N | 20 | 16 | 6 | AAC | 59 (41) |
| P | 136 | 104 | 50 | CCG | 47 (23) |
| Q | 20 | 19 | 14 | CAA | 52.2 (47.8) |
| S | 368 | 102 | 109 | AGC | 24.9 (16.2) |
| T | 81 | 24 | 32 | ACG | 40 (38) |
| V | 1 | 1 | 1 | GTT | 55 (27) |

***Πίνακας 11.*** *Βακτήρια: Τα αμινοξέα με τον αριθμό των επαναλήψεων που δημιουργούν, τον αριθμό των οργανισμών και των πρωτεϊνών που ανιχνεύτηκαν, το κυρίαρχο και το δεύτερο πιο συχνό κωδικόνιο στην παρένθεση, και τα ποσοστά τους*

| Αρχαία | | | | | |
|---|---|---|---|---|---|
| **AA** | **Επαναλήψεις** | **Οργανισμοί** | **Ορθόλογες πρωτεΐνες** | **Κυρίαρχο κωδικόνιο** | **Κυρίαρχο κωδικόν. % (2ο συχνότερο κωδικ.%)** |
| A | 18 | 1 | 1 | GCT | 64.7 (26.9) |
| D | 4 | 3 | 2 | GAC | 62 (38) |
| E | 12 | 12 | 4 | GAA | 61.9 (38.1) |
| G | 12 | 9 | 7 | GGC | 53 (28) |
| L | 1 | 1 | 1 | CTT | 80 (20) |
| P | 1 | 1 | 1 | CCA/CCT | 33/33 (25) |
| Q | 1 | 1 | 1 | CAG | 58 (42) |
| R | 1 | 1 | 1 | AGA | 52.6 (36.8) |
| S | 8 | 6 | 6 | TCA | 24.1 (22.4) |
| T | 49 | 11 | 16 | ACT | 28.2 (26.1) |

***Πίνακας 12.*** *Αρχαία: Τα αμινοξέα με τον αριθμό των επαναλήψεων που δημιουργούν, τον αριθμό των οργανισμών και των πρωτεϊνών που ανιχνεύτηκαν, το κυρίαρχο και το δεύτερο πιο συχνό κωδικόνιο στην παρένθεση, και τα ποσοστά τους*

| Βακτηριοφάγοι | | | | | |
|---|---|---|---|---|---|
| **AA** | **Επαναλήψεις** | **Οργανισμοί** | **Ορθόλογες πρωτεΐνες** | **Κυρίαρχο κωδικόνιο** | **Κυρίαρχο κωδικόν. % (2ο συχνότερο κωδικ.%)** |
| G | 1 | 1 | 1 | GGT | 40 (30) |
| S | 1 | 1 | 1 | TCC | 30.8 (23.1) |

***Πίνακας 13**. Βακτηριοφάγοι: Τα αμινοξέα με τον αριθμό των επαναλήψεων που δημιουργούν, τον αριθμό των οργανισμών και των πρωτεϊνών που ανιχνεύτηκαν, το κυρίαρχο και το δεύτερο πιο συχνό κωδικόνιο στην παρένθεση, και τα ποσοστά τους*

Τα SARs είναι εμπλουτισμένα σε ορισμένα αμινοξέα, όπως η γλυκίνη, η σερίνη, θρεονίνη, αλανίνη, κ.α., τόσο στα αρχαία όσο και στα βακτήρια και στους βακτηριοφάγους και εντοπίζονται σε πολλά διαφορετικά είδη σε κάθε βασίλειο (Sampath Kumar et al., 2016).

Έχει δειχθεί ότι, πολλές επαναλήψεις του ίδιου αμινοξέος, είναι υπεύθυνες για τη λανθασμένη αναδίπλωση της πρωτεΐνης, στην οποία έχουν ανιχνευθεί, με αποτέλεσμα να δημιουργούνται αδιάλυτα συσσωματώματα, τα οποία είναι τοξικά για τα κύτταρα. Ορισμένες πρωτεΐνες που περιέχουν SARs γίνονται παθολογικές όταν οι επαναλήψεις επεκτείνονται πέρα από ένα ορισμένο μήκος (Oma Yoko et al., 2009).

Κατά γενική ομολογία, τα SARs που δημιουργούνται από υδρόφοβα αμινοξέα παρουσιάζουν ισχυρή τάση προς συσσωμάτωση και είναι πιο τοξικά για τα κύτταρα (Dorsman et al., 2002; Faux et al., 2005; Oma et al., 2004). Αυτό έχει σαν αποτέλεσμα, σπάνια να παρατηρούνται περιοχές με υδρόφοβα SARs σε φυσικές πρωτεΐνες (Dorsman et al., 2002; Oma et al., 2005, 2004). Αντίθετα, τα λιγότερο τοξικά SARs όπως η πολυαλανίνη, η πολυγλουταμίνη, η πολυπρολίνη, η πολυσερίνη και η πολυγλυκίνη είναι πιο “ανεκτά” για τα κύτταρα και σχετικά άφθονα, ιδιαίτερα μεταξύ των μεταγραφικών παραγόντων (Dorsman et al., 2002; Oma et al., 2005, 2004). Ένα παράδειγμα είναι οι επαναλήψεις πολυ-λευκίνης, που είναι πολύ πιο τοξικές από τις επαναλήψεις πολυ-γλουταμίνης (Dorsman et al., 2002).

Επιπλέον, έχει δειχθεί ότι όσο πιο υδρόφοβο είναι ένα SAR τόσο πιο ισχυρή είναι και η επαγωγή κυτταρικού θανάτου και η δραστικότητα της κασπάσης-3. Έτσι με παρουσία υδρόφοβων SARs επάγεται η απόπτωση. Φυσιολογικά, τα τοξικά SARs αποβάλλονται, καθώς οι πρωτεΐνες που τα έχουν αποικοδομούνται. Κάποια SARs (poly-Q) επάγουν απόκριση στρες (στο ενδοπλασματικό δίκτυο, ER) και προκαλούνται ελαττώματα στο μηχανισμό αποικοδόμησης. Έτσι τοξικές, λάθος αναδιπλωμένες πρωτεΐνες συσσωρεύονται στο κύτταρο (Uchio Naohiro et al., 2007).

Τα αποτελέσματα κάποιων άλλων ερευνών υποδεικνύουν ότι τα υδρόφοβα SARs μπορούν να αλληλεπιδρούν με τον εαυτό τους και με άλλα υδρόφοβα SARs. Επιπλέον, ως επαναλαμβανόμενα τμήματα όμοιων αμινοξέων, τα SARs μπορούν να υιοθετήσουν χαρακτηριστικές διαμορφώσεις με σημαντικό αντίκτυπο στις αλληλεπιδράσεις πρωτεΐνης - πρωτεΐνης (Oma Yoko et al., 2009).

### 3.1.7 Επαναλήψεις Πολυ - σερίνης (Poly - serine tracts)

Οι πρωτεΐνες με επαναλήψεις πολυ - σερίνης (poly - serine tracts) στα βακτήρια αναλύθηκαν περαιτέρω. Το 42% (90/214 πρωτεΐνες - 55 διαφορετικοί σχολιασμοί πρωτεϊνών) των πρωτεϊνών που διαθέτουν επαναλήψεις πολυ – σερίνης εμπλέκονται στην αποικοδόμηση πολυσακχαριτών / υδατανθράκων και στο μεταβολισμό. Σε αντίθεση, η συχνότητα υποβάθρου των πρωτεϊνών που περιέχουν SARs και εμπλέκονται στο μεταβολισμό ήταν 14% (117/856), μια στατιστικά σημαντική διαφορά (p-value του υπεργεωμετρικού τεστ: 4e-38). Οι περισσότερες από αυτές τις πρωτεΐνες (90% - 198 πρωτεΐνες) προέρχονται από τα *Teredinibacter turnerae* και *Saccharophagus degradans*, δύο θαλάσσιους μικροοργανισμούς που κωδικοποιούν για πολλά ένζυμα αποικοδόμησης πολυσακχαριτών (Weiner et al., 2008; Yang et al., 2009). Αυτές οι επαναλήψεις είτε έχουν κάποιο ρόλο στην αποικοδόμηση πολυσακχαριτών είτε προκαλούν γενετική αστάθεια που πυροδότησε τεράστια εξάπλωση των γονιδίων τους.

Επίσης, υπάρχουν αναφορές ότι σε πρωτεΐνες του *Microbulbifer degradans* τέτοια SARs πολυ - σερίνης εντοπίζονται σε πρωτεΐνες, που στη μεγάλη πλειοψηφία τους εμπλέκονται σε αποικοδόμηση υδατανθράκων. Αυτές οι περιοχές παρεμβάλλονται μεταξύ των διάφορων επικρατειών και τις χωρίζουν, πιθανότατα λειτουργώντας ως εύκαμπτοι σύνδεσμοι, που ενισχύουν την προσβασιμότητα του υποστρώματος (Howard et al., 2004).

## 3.2 LCRs και λειτουργία

### 3.2.1 Λειτουργικός εμπλουτισμός

Για να έχουμε μια πρώτη εικόνα των πρωτεϊνών στις οποίες εντοπίζονται τα LCRs, αρχικά, προσδιορίστηκε ποιες πρωτεΐνες που περιέχουν LCRs εμφανίζονται σε πολλούς οργανισμούς.

Στον Πίνακας 14 και τον Πίνακας 15 φαίνονται οι 20 πιο συχνές πρωτεΐνες στα βακτήρια και οι 10 πιο συχνές πρωτεΐνες των αρχαίων αντίστοιχα, μαζί με το συνολικό αριθμό των τμημάτων τους.

| **Βακτήρια** | |
|---|---|
| **Αριθμός τμημάτων LCRs** | **Πρωτεΐνη** |
| 9751 | Uncharacterized protein |
| 321 | Translation initiation factor IF-2 |
| 281 | DNA topoisomerase 1 |
| 220 | 60 kDa chaperonin |
| 208 | Acetyltransferase component of pyruvate dehydrogenase complex |
| 186 | 30S ribosomal protein S16 |
| 167 | Dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase complex |
| 166 | Protein TonB |
| 152 | Single-stranded DNA-binding protein |
| 127 | 50S ribosomal protein L25 |
| 120 | Protein TolA |
| 102 | 30S ribosomal protein S2 |
| 135 | Serine/threonine protein kinase |
| 90 | Signal recognition particle protein |
| 81 | Ribonuclease E |
| 69 | Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex |
| 146 | RNA-binding protein |
| 68 | Polyribonucleotide nucleotidyltransferase |
| 65 | Endoglucanase |
| 63 | Membrane protein |

***Πίνακας 14.*** *Βακτήρια: Οι κορυφαίες 20 πρωτεΐνες που περιέχουν LCRs και εμφανίζονται στους περισσότερους οργανισμούς*

| Αρχαία | |
|---|---|
| **Αριθμός τμημάτων LCRs** | **Πρωτεΐνη** |
| 662 | Uncharacterized protein |
| 60 | Thermosome |
| 40 | 50S ribosomal protein L12 |
| 23 | Extracellular solute-binding protein family 5 |
| 18 | Chaperone protein DnaK |
| 13 | 50S ribosomal protein L10 |
| 13 | 30S ribosomal protein S24e |
| 11 | Prefoldin subunit alpha |
| 11 | Carbohydrate binding family 6 |
| 11 | 30S ribosomal protein S3 |
| 7 | Signal recognition particle receptor FtsY |

***Πίνακας 15.*** *Αρχαία: Οι κορυφαίες 10 πρωτεΐνες που περιέχουν LCRs και εμφανίζονται στους περισσότερους οργανισμούς*

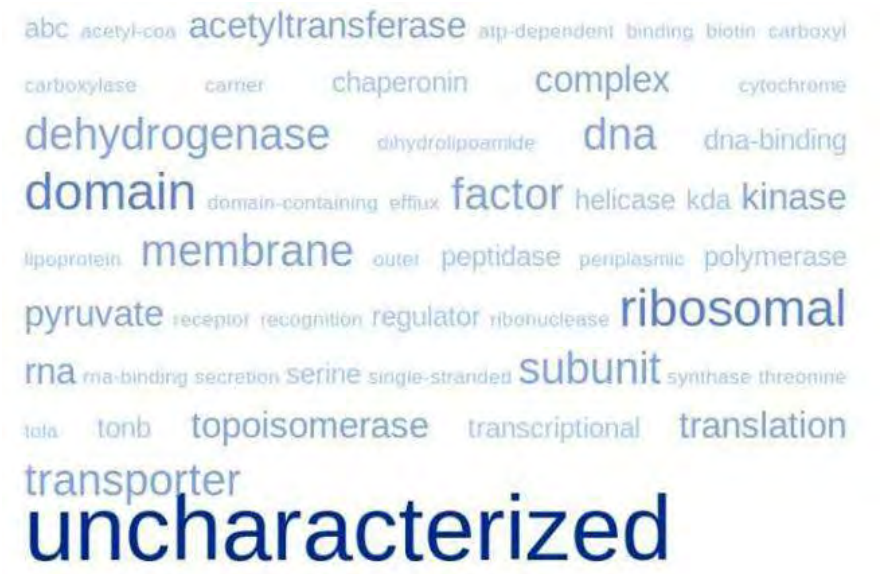
Τόσο στα αρχαία όσο και στα βακτήρια οι περισσότερες πρωτεΐνες με LCRs είναι μη χαρακτηρισμένες.

Όσον αφορά τα βακτήρια, οι πρωτεΐνες με LCRs που εντοπίζονται σε πολλούς οργανισμούς φαίνεται να είναι ο παράγοντας έναρξης της μετάφρασης IF-2, η τσαπερονίνη 60 kDa, ριβοσωμικές πρωτεΐνες καθώς και πρωτεΐνες σχετικές με πρόσδεση / επεξεργασία του DNA και του RNA και πρωτεΐνες του συμπλόκου της αφυδρογονάσης του πυρουβικού.

Στα αρχαία οι πρωτεΐνες με LCRs που εντοπίζονται στους περισσότερους οργανισμούς είναι το θερμόσωμα και τσαπερόνες καθώς και ριβοσωμικές πρωτεΐνες.

Στη συνέχεια, προκειμένου να πραγματοποιηθεί μια εκτίμηση των συχνότερων λειτουργικών κατηγοριών στις οποίες ανήκουν οι πρωτεΐνες που περιέχουν LCRs, συγκεντρώθηκαν οι γονιδιακοί σχολιασμοί (gene annotations) τους και η συχνότητα των λειτουργικών λέξεων - κλειδιών απεικονίστηκε με σύννεφα λέξεων (word clouds) από το https://tagcrowd.com/.

Η ένταση του χρώματος και το μέγεθος των λέξεων είναι ανάλογο της συχνότητάς τους. Πιο συγκεκριμένα, όσες περισσότερες φορές συναντάται μια λέξη - κλειδί τόσο μεγαλύτερο μέγεθος και πιο σκούρο χρώμα έχει.




***Εικόνα 14.*** *Wordclouds των πρωτεϊνών που περιέχουν LCRs στα Βακτήρια*


***Εικόνα 15.*** *Word clouds των πρωτεϊνών που περιέχουν LCRs στα Αρχαία*


***Εικόνα 16.*** *Word clouds των πρωτεϊνών που περιέχουν LCRs στους Βακτηριοφάγους*

Φαίνεται επίσης, πως κάποιες από τις μεγάλες κατηγορίες πρωτεϊνών με LCRs είναι κοινές μεταξύ αρχαίων και βακτηρίων. Τέτοιες κατηγορίες είναι οι τσαπερόνες, οι ριβοσωμικές πρωτεΐνες, οι μεταφορείς, και οι πρωτεΐνες με σχολιασμούς σχετικούς με DNA καθώς και RNA και μεταγραφή.

### 3.2.2 Ανάλυση με Οντολογίες (Gene Ontology, GO)

Με χρήση του εργαλείου χαρτογράφησης της Uniprot ανακτήθηκαν οι όροι – οντολογίες των πρωτεϊνών που περιέχουν LCRs, χρησιμοποιώντας το αναγνωριστικό που τους έχει δοθεί από τη Uniprot (Uniprot ID). Ανακτήθηκαν οι όροι - οντολογίες για βιολογική διεργασία, μοριακή λειτουργία και κυτταρικό διαμέρισμα.
Κάθε τμήμα LCR και SAR χαρακτηρίστηκε με τους αντίστοιχους όρους – οντολογίες και στη συνέχεια, για κάθε όρο – οντολογία υπολογίστηκε ο αριθμός των πρωτεϊνών (για τα LCRs και τα SARs) στις οποίες εκχωρήθηκε. Τα αποτελέσματα για τα 3 βασίλεια φαίνονται στους Πίνακας 16, Πίνακας 17 και Πίνακας 18.

| Βακτήρια | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LCRs | | SARs | | LCRs | | SARs | | LCRs | | SARs | |
| Βιολογική διεργασία | | | | Μοριακή λειτουργία | | | | Κυτταρικό διαμέρισμα | | | |
| 830 | translation [GO:0006412] | 56 | carbohydrate metabolic process [GO:0005975] | 1213 | ATP binding [GO:0005524] | 53 | hydrolase activity, hydrolyzing O-glycosyl compounds [GO:0004553] | 4802 | integral component of membrane [GO:0016021] | 281 | integral component of membrane [GO:0016021] |
| 336 | carbohydrate metabolic process [GO:0005975] | 15 | transport [GO:0006810] | 949 | DNA binding [GO:0003677] | 51 | carbohydrate binding [GO:0030246] | 1258 | cytoplasm [GO:0005737] | 35 | extracellular region [GO:0005576] |
| 327 | DNA topological change [GO:0006265] | 14 | cellulose catabolic process [GO:0030245] | 820 | structural constituent of ribosome [GO:0003735] | 33 | ATP binding [GO:0005524] | 653 | ribosome [GO:0005840] | 26 | cytoplasm [GO:0005737] |
| 281 | DNA replication [GO:0006260] | 12 | xylan catabolic process [GO:0045493] | 627 | metal ion binding [GO:0046872] | 31 | cellulose binding [GO:0030248] | 375 | plasma membrane [GO:0005886] | 23 | cell outer membrane [GO:0009279] |
| 274 | transport [GO:0006810] | 11 | cell adhesion [GO:0007155] | 422 | GTP binding [GO:0005525] | 19 | lyase activity [GO:0016829] | 276 | membrane [GO:0016020] | 14 | plasma membrane [GO:0005886] |
| 220 | protein refolding [GO:0042026] | 10 | polysaccharide catabolic process [GO:0000272] | 410 | GTPase activity [GO:0003924] | 18 | nucleic acid binding [GO:0003676] | 177 | extracellular region [GO:0005576] | 9 | cell wall [GO:0005618] |
| 212 | metabolic process [GO:0008152] | 10 | translation [GO:0006412] | 374 | nucleic acid binding [GO:0003676] | 15 | penicillin binding [GO:0008658] | 173 | pyruvate dehydrogenase complex [GO:0045254] | 9 | membrane [GO:0016020] |
| 186 | protein folding [GO:0006457] | 8 | protein transport [GO:0015031] | 347 | RNA binding [GO:0003723] | 15 | transferase activity [GO:0016740] | 170 | cell outer membrane [GO:0009279] | 9 | ribosome [GO:0005840] |
| 156 | regulation of transcription, DNA-templated [GO:0006355] | 7 | transmembrane transport [GO:0055085] | 316 | unfolded protein binding [GO:0051082] | 14 | translation initiation factor activity [GO:0003743] | 164 | small ribosomal subunit [GO:0015935] | 8 | outer membrane-bounded periplasmic space [GO:0030288] |
| 153 | transcription, DNA-templated [GO:0006351] | 6 | pseudouridine synthesis [GO:0001522] | 311 | DNA topoisomerase type I activity [GO:0003917] | 13 | metal ion binding [GO:0046872] | 141 | chromosome [GO:0005694] | 6 | viral capsid [GO:0019028] |

***Πίνακας 16.*** *Βακτήρια: Αριθμός πρωτεϊνών για κάθε όρο - οντολογία*

| Αρχαία | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **LCRs** | | **SARs** | | **LCRs** | | **SARs** | | **LCRs** | | **SARs** | |
| **Βιολογική διεργασία** | | | | **Μοριακή λειτουργία** | | | | **Κυτταρικό διαμέρισμα** | | | |
| 101 | protein folding [GO:0006457] | 7 | translational elongation [GO:0006414] | 127 | ATP binding [GO:0005524] | 7 | structural constituent of ribosome [GO:0003735] | 467 | integral component of membrane [GO:0016021] | 39 | integral component of membrane [GO:0016021] |
| 78 | translation [GO:0006412] | 4 | carbohydrate metabolic process [GO:0005975] | 109 | structural constituent of ribosome [GO:0003735] | 6 | hydrolase activity [GO:0016787] | 97 | ribosome [GO:0005840] | 9 | ribosome [GO:0005840] |
| 40 | translational elongation [GO:0006414] | 2 | ribosome biogenesis [GO:0042254] | 100 | unfolded protein binding [GO:0051082] | 3 | calcium ion binding [GO:0005509] | 29 | cytoplasm [GO:0005737] | 1 | integral component of plasma membrane [GO:0005887] |
| 17 | ribosome biogenesis [GO:0042254] | 1 | carbohydrate transport [GO:0008643] | 32 | rRNA binding [GO:0019843] | 3 | carbohydrate binding [GO:0030246] | 16 | small ribosomal subunit [GO:0015935] | 1 | mucus layer [GO:0070701] |
| 14 | carbohydrate metabolic process [GO:0005975] | 1 | cobalamin biosynthetic process [GO:0009236] | 25 | DNA binding [GO:0003677] | 2 | catalytic activity [GO:0003824] | 12 | prefoldin complex [GO:0016272] | 0 | X |
| 14 | SRP-dependent cotranslational protein targeting to membrane [GO:0006614] | 1 | D-amino acid catabolic process [GO:0019478] | 22 | metal ion binding [GO:0046872] | 2 | sequence-specific DNA binding [GO:0043565] | 10 | plasma membrane [GO:0005886] | 0 | X |
| 13 | amino acid transport [GO:0006865] | 1 | homophilic cell adhesion via plasma membrane adhesion molecules [GO:0007156] | 21 | copper ion binding [GO:0005507] | 1 | large ribosomal subunit rRNA binding [GO:0070180] | 8 | large ribosomal subunit [GO:0015934] | 0 | X |
| 13 | DNA replication [GO:0006260] | 1 | phosphate ion transmembrane transport [GO:0035435] | 21 | serine-type endopeptidase activity [GO:0004252] | 1 | metal ion binding [GO:0046872] | 7 | intrinsic component of plasma membrane [GO:0031226] | 0 | X |
| 7 | archaeal or bacterial-type flagellum-dependent cell motility [GO:0097588] | 1 | polysaccharide catabolic process [GO:0000272] | 21 | zinc ion binding [GO:0008270] | 1 | precorrin-3B C17-methyltransferase activity [GO:0030789] | 7 | signal recognition particle [GO:0048500] | 0 | X |
| 5 | regulation of transcription, DNA-templated [GO:0006355] | 1 | translation [GO:0006412] | 17 | carbohydrate binding [GO:0030246] | 1 | transporter activity [GO:0005215] | 7 | viral nucleocapsid [GO:0019013] | 0 | X |

***Πίνακας 17.*** *Αρχαία: Αριθμός πρωτεϊνών για κάθε όρο - οντολογία*

| Βακτηριοφάγοι | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **LCRs** | | **SARs** | | **LCRs** | | **SARs** | | **LCRs** | | **SARs** | | |
| **Βιολογική διεργασία** | | | | **Μοριακή λειτουργία** | | | | **Κυτταρικό διαμέρισμα** | | | | |
| 5 | adhesion receptor-mediated virion attachment to host cell [GO:0098671] | 0 | X | 5 | DNA binding [GO:0003677] | 0 | X | 20 | integral component of membrane [GO:0016021] | 1 | integral component of membrane [GO:0016021] |
| 5 | virion attachment to host cell pilus [GO:0039666] | 0 | X | 1 | amidase activity [GO:0004040] | 0 | X | 6 | viral capsid [GO:0019028] | 0 | X |
| 4 | entry receptor-mediated virion attachment to host cell [GO:0098670] | 0 | X | 1 | exonuclease activity [GO:0004527] | 0 | X | 5 | host cell membrane [GO:0033644] | 0 | X |
| 4 | viral entry into host cell via pilus retraction [GO:0039667] | 0 | X | 1 | metal ion binding [GO:0046872] | 0 | X | 3 | virion [GO:0019012] | 0 | X |
| 4 | viral extrusion [GO:0099045] | 0 | X | 1 | RNA binding [GO:0003723] | 0 | X | 2 | host cell cytoplasm [GO:0030430] | 0 | X |
| 3 | viral release from host cell [GO:0019076] | 0 | X | 1 | serine-type peptidase activity [GO:0008236] | 0 | X | 1 | virus tail, fiber [GO:0098024] | 0 | X |
| 2 | DNA repair [GO:0006281] | 0 | X | 1 | single-stranded DNA binding [GO:0003697] | 0 | X | 1 | virus tail, tube [GO:0098026] | 0 | X |
| 2 | DNA replication [GO:0006260] | 0 | X | 0 | X | 0 | X | 0 | X | 0 | X |
| 2 | viral entry into host cell [GO:0046718] | 0 | X | 0 | X | 0 | X | 0 | X | 0 | X |
| 2 | viral tail assembly [GO:0098003] | 0 | X | 0 | X | 0 | X | 0 | X | 0 | X |

***Πίνακας 18.*** *Βακτηριοφάγοι: Αριθμός πρωτεϊνών για κάθε όρο – οντολογία*

Τόσο στα αρχαία όσο και στα βακτήρια εντοπίζονται κάποιοι κοινοί όροι - οντολογίες και σχετίζονται με μετάφραση, μεταβολισμό υδατανθράκων, DNA και RNA, αναδίπλωση πρωτεϊνών, πρόσδεση μετάλλων, μεμβράνη, ριβόσωμα και κυτταρόπλασμα.

Για κάθε όρο – οντολογία υπολογίστηκε, επιπλέον, ο αριθμός τμημάτων LCRs και SARs (καθώς μια πρωτεΐνη μπορεί να έχει περισσότερα από ένα LCRs) και το αντίστοιχο διάνυσμα για τα αμινοξέα. Υπολογίστηκε, επίσης, η συχνότητα και ο εμπλουτισμός του κάθε αμινοξέος στους όρους – οντολογίες (για τα αρχαία, τα βακτήρια και τους φάγους).

Προκειμένου να εντοπιστεί ποιό/ά αμινοξέα είναι κυρίαρχα σε κάθε όρο – οντολογία, συλλέχθηκαν τα διανύσματα κάθε όρου που είχε εκχωρηθεί σε περισσότερα από 50 τμήματα LCRs και φιλτραρίστηκαν όλα τα αμινοξέα με εμπλουτισμό μικρότερο από 2,5 φορές, δηλαδή κάθε αμινοξύ με εμπλουτισμό μικρότερο του 2,5 έλαβε την τιμή 0.

Στους Πίνακας 19, Πίνακας 20, Πίνακας 21, Πίνακας 22, Πίνακας 23, Πίνακας 24, Πίνακας 25, Πίνακας 26 και Πίνακας 27 φαίνονται οι όροι - οντολογίες, το αναγνωριστικό του κάθε όρου και τα αμινοξέα στα οποία είναι περισσότερο εμπλουτισμένος.

| Βακτηριακά LCRs | | | | | | |
|---|---|---|---|---|---|---|
| **Περιγραφή οντολογίας** | **ID οντολογίας** | **C** | **P** | **Q** | **S** | **T** |
| chitin binding | GO:0008061 | - | - | - | 2.5 | 3.9 |
| carbohydrate binding | GO:0030246 | - | - | - | 2.9 | 3.1 |
| carbohydrate metabolic process | GO:0005975 | - | - | - | 3.2 | 2.8 |
| hydrolase activity, hydrolyzing O-glycosyl compounds | GO:0004553 | - | - | - | 3.6 | 2.5 |
| cellulose catabolic process | GO:0030245 | - | - | - | 4.2 | 2.5 |
| cellulase activity | GO:0008810 | - | - | - | 4.2 | - |
| peptidoglycan binding | GO:0042834 | - | - | 2.7 | - | - |
| chitinase activity | GO:0004568 | - | 2.9 | - | - | 4.2 |
| xylan catabolic process | GO:0045493 | 2.8 | - | - | 3.8 | - |
| cellulose binding | GO:0030248 | 3.8 | - | - | 6.8 | - |
| endo-1,4-beta-xylanase activity | GO:0031176 | 3.9 | - | - | 4.6 | - |

***Πίνακας 19.*** *Βακτηριακά LCRs: Κυρίαρχα αμινοξέα σε κάθε όρο-οντολογία που σχετίζεται με πολυσακχαρίτες*

| **Περιγραφή οντολογίας** | **ID οντολογίας** | **D** | **H** | **Q** | **R** | **S** | **T** | **V** |
|---|---|---|---|---|---|---|---|---|
| cytoplasmic side of plasma membrane | GO:0009898 | - | - | - | 2.7 | - | - | 3.2 |
| integral component of plasma membrane | GO:0005887 | - | - | 3.6 | - | - | - | - |
| plasma membrane | GO:0005886 | - | 3.7 | - | - | - | - | - |
| cell wall | GO:0005618 | 2.9 | - | - | - | 2.8 | 3 | - |

***Πίνακας 20.*** *Βακτηριακά LCRs: Κυρίαρχα αμινοξέα σε κάθε όρο-οντολογία που σχετίζεται με μεμβράνη*

| **Περιγραφή οντολογίας** | **ID οντολογίας** | **A** | **V** |
|---|---|---|---|
| Dihydrolopoyllysine-residue acetyltransferase activity | GO:0004742 | 2.9 | - |
| Pyruvate dehydrogenase complex | GO:0045254 | 2.9 | - |
| Glycolytic process | GO:0006096 | 2.9 | - |
| Fatty acid biosynthetic process | GO:0006633 | 2.7 | 2.5 |
| Acetyl-CoA cardoxylase activity | GO:0003989 | 2.8 | 2.7 |
| Acetyl-CoA cardoxylase complex | GO:0009317 | 2.8 | 2.7 |
| Tricarboxylic acid cycle | GO:0006099 | 2.7 | - |
| Dihydrolopoyllysine-residue succinyltransferase activity | GO:0004149 | 2.8 | - |
| Oxoglutarate dehydrogenase complex | GO:0045252 | 2.8 | - |
| L-lysine catabolic process to acetyl-CoA via saccharopine | GO:0033512 | 2.8 | - |
| Pyruvate metabolic process | GO:0006090 | 2.6 | - |
| Flavin adenine dinucleotide binding | GO:0050660 | 2.5 | - |

***Πίνακας 21.*** *Βακτηριακά LCRs: Κυρίαρχα αμινοξέα σε κάθε όρο-οντολογία που σχετίζεται με μεταβολισμό*

| Περιγραφή οντολογίας | ID οντολογίας | D | E | F | I | L | M | N | R | V |
|---|---|---|---|---|---|---|---|---|---|---|
| 7S RNA binding | GO:0008312 | - | - | 2.8 | - | 3.9 | 22 | - | - | - |
| DNA-directed 5'-3' RNA polymerase activity | GO:0003899 | 5.9 | 4.5 | 3.3 | 4.3 | 3.9 | - | - | - | - |
| polyribonucleotide nucleotidyltransferase activity | GO:0004654 | 2.9 | - | - | - | - | - | - | 10.4 | - |
| 3'-5'-exoribonuclease activity | GO:0000175 | 2.97 | - | - | - | - | - | - | 10.4 | - |
| RNA processing | GO:0006396 | - | - | - | - | - | - | - | 7.5 | - |
| helicase activity | GO:0004386 | - | - | - | - | - | - | 4.3 | 6.8 | - |
| mRNA catabolic process | GO:0006402 | - | - | - | - | - | - | - | 5.9 | - |
| RNA binding | GO:0003723 | - | - | - | - | - | - | 2.5 | 4.9 | - |
| small ribosomal subunit | GO:0015935 | - | - | - | - | - | - | - | 3.2 | - |
| translation initiation factor activity | GO:0003743 | - | - | - | - | - | - | - | 2.9 | - |
| rRNA binding | GO:0019843 | - | - | - | - | - | - | - | 2.8 | - |
| endoribonuclease activity | GO:0004521 | - | - | - | - | - | - | - | 2.7 | 3.2 |
| rRNA processing | GO:0006364 | - | - | - | - | - | - | - | 2.6 | 3 |
| tRNA processing | GO:0008033 | - | - | - | - | - | - | - | 2.6 | 3.2 |
| ribonuclease E activity | GO:0008995 | - | - | - | - | - | - | - | 2.5 | 3.6 |
| translation | GO:0006412 | - | 2.8 | - | - | - | - | - | - | - |
| structural constituent of ribosome | GO:0003735 | - | 2.8 | - | - | - | - | - | - | - |
| ribosome | GO:0005840 | - | 3 | - | - | - | - | - | - | - |
| 5S rRNA binding | GO:0008097 | - | 3.2 | - | - | - | - | - | - | - |
| transcription, DNA-templated | GO:0006351 | 3.5 | 2.5 | - | - | 2.5 | - | - | - | - |

***Πίνακας 22.*** *Βακτηριακά LCRs: Κυρίαρχα αμινοξέα σε κάθε όρο-οντολογία που σχετίζεται με RNA*

| Περιγραφή οντολογίας | ID οντολογίας | F | G | H | K | L | N | P | Q | Y |
|---|---|---|---|---|---|---|---|---|---|---|
| DNA recombination | GO:0006310 | - | - | - | - | - | - | - | - | 5.1 |
| DNA polymerase III complex | GO:0009360 | - | - | - | - | - | - | 2.6 | 2.5 | - |
| regulation of transcription, DNA-templated | GO:0006355 | - | - | - | - | - | 2.9 | - | - | - |
| DNA-templated transcription, initiation | GO:0006352 | - | - | - | - | 3.6 | - | - | - | - |
| DNA binding | GO:0003677 | - | - | - | 4.2 | - | - | - | - | - |
| chromosome condensation | GO:0030261 | - | - | - | 5.2 | - | - | - | - | - |
| chromosome | GO:0005694 | - | - | - | 5.8 | - | - | - | - | - |
| DNA topological change | GO:0006265 | - | - | - | 6.1 | - | - | - | - | - |
| DNA topoisomerase type I activity | GO:0003917 | - | - | - | 6.3 | - | - | - | - | - |
| nucleosome | GO:0000786 | - | - | - | 6.5 | - | - | - | - | - |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| nucleosome assembly | GO:0006334 | - | - | - | 6.5 | - | - | - | - | - |
| nucleotide binding | GO:0000166 | - | - | 3.1 | - | - | - | - | - | - |
| DNA repair | GO:0006281 | - | 2.6 | - | - | - | - | - | - | 4.8 |
| single-stranded DNA binding | GO:0003697 | 2.6 | 3.2 | - | - | - | - | - | 3 | 5.4 |
| DNA replication | GO:0006260 | 5.4 | 2.5 | - | - | - | - | - | 2.6 | 3.2 |
| DNA-templated transcription, termination | GO:0006353 | - | - | - | - | - | 7.8 | - | 2.5 | 2.9 |

***Πίνακας 23.*** *Βακτηριακά LCRs: Κυρίαρχα αμινοξέα σε κάθε όρο-οντολογία που σχετίζεται με DNA*

| Περιγραφή οντολογίας | ID οντολογίας | F | G | H | I | M |
|---|---|---|---|---|---|---|
| Unfolded protein binding | GO:0051082 | 5.7 | 2.6 | 4 | - | 21.6 |
| Protein refolding | GO:0042026 | - | 2.9 | - | - | 37.2 |
| Protein folding | GO:0006457 | 7.4 | - | 6.3 | - | - |
| Heat shock protein binding | GO:0031072 | 21.5 | 3 | - | 5.2 | - |

***Πίνακας 24.*** *Βακτηριακά LCRs: Κυρίαρχα αμινοξέα σε κάθε όρο-οντολογία που σχετίζεται με τσαπερόνες*

| Περιγραφή οντολογίας | ID οντολογίας | D | F | H | K |
|---|---|---|---|---|---|
| Metal ion binding | GO:0046872 | - | - | 4.9 | 3.1 |
| Zinc ion binding | GO:0008270 | - | 4.5 | - | - |
| Nickel cation binding | GO:0016151 | 2.7 | - | 33 | - |
| Metal ion transport | GO:0030001 | 5.3 | - | 21.6 | - |
| Cobalamin biosynthetic process | GO:0009236 | 2.7 | - | 29 | - |

***Πίνακας 25.*** *Βακτηριακά LCRs: Κυρίαρχα αμινοξέα σε κάθε όρο-οντολογία που σχετίζεται με πρόσδεση μετάλλων*

Για τα βακτήρια, από τα κυρίαρχα αμινοξέα σε κάθε όρο - οντολογία που σχετίζεται με πολυσακχαρίτες, αυτά που εμφανίζονται στους περισσότερους όρους είναι τα S και T. Σε κάθε όρο - οντολογία που σχετίζεται με μεταβολισμό τα πιο συχνά είναι τα A και V. Στους περισσότερους όρους - οντολογίες που σχετίζονται με RNA εμφανίζεται το R, στους περισσότερους όρους - οντολογίες που σχετίζονται με DNA εμφανίζεται το K και τα G, F, Q και Y. Σε αυτούς που σχετίζονται με τσαπερόνες τα G, M και F και τέλος στους σχετικούς με πρόσδεση μετάλλων τα H και D.

Όσον αφορά τους όρους - οντολογίες που σχετίζονται με μεμβράνη, δεν υπάρχει κάποιο αμινοξύ που να κυριαρχεί στους περισσότερους.

| Βακτηριακά SARs | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Περιγραφή οντολογίας** | **ID οντολογίας** | **F** | **I** | **K** | **L** | **S** | **V** |
| integral component of membrane | GO:0016021 | 4.3 | 4.3 | 3.7 | 4.3 | - | 4.3 |
| cellulose catabolic process | GO:0030245 | - | - | - | - | 2.5 | - |
| cellulose binding | GO:0030248 | - | - | - | - | 2.6 | - |

***Πίνακας 26.*** *Βακτηριακά SARs: Κυρίαρχα αμινοξέα σε κάθε όρο-οντολογία*

| LCRs Αρχαίων | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Περιγραφή οντολογίας** | **ID οντολογίας** | **A** | **E** | **G** | **K** | **M** | **Q** |
| translation | GO:0006412 | - | - | - | 3.15 | - | - |
| ATP binding | GO:0005524 | - | - | 3 | - | 6.5 | - |
| protein folding | GO:0006457 | - | - | 3.5 | - | 9.2 | 3 |
| unfolded protein binding | GO:0051082 | - | - | 3.6 | - | 9.5 | 3.1 |
| ribosome | GO:0005840 | 2.8 | 2.6 | - | - | - | - |
| structural constituent of ribosome | GO:0003735 | 2.8 | 2.5 | - | - | - | - |

***Πίνακας 27.*** *LCRs Αρχαίων: Κυρίαρχα αμινοξέα σε κάθε όρο-οντολογία*

Για τα αρχαία, σε όλους τους όρους που σχετίζονται με τσαπερόνες, τα αμινοξέα, που εμφανίζονται συχνά είναι τα G και M (λιγότερο Q), σε όλους τους όρους σχετικούς με το ριβόσωμα τα αμινοξέα A και E και στη μετάφραση το K.

## 3.3 Ομαδοποίηση και ανάλυση συχνότερων λειτουργικών κατηγοριών

### 3.3.1 Πρωτεΐνες που προσδένουν σε RNA και DNA

Μια μεγάλη κατηγορία πρωτεϊνών που διαθέτουν LCRs και ανιχνεύθηκαν στην ανάλυση ήταν οι πρωτεΐνες που προσδένουν σε μόρια DNA και RNA.

Από τις μεγάλες κατηγορίες πρωτεϊνών (που βρέθηκαν σε πολλούς οργανισμούς) που περιέχουν LCRs, συγκεντρώθηκαν όλες όσες είχαν χαρακτηρισμούς ή όρους – οντολογίες που να σχετίζονται με πρόσδεση μορίων DNA και RNA και λήφθηκε από τη Uniprot η αλληλουχία τους.

Οι πρωτεΐνες της κάθε κατηγορίας στοιχήθηκαν με χρήση του muscle και τα αποτελέσματα οπτικοποιήθηκαν με το Seaview και το Jalview. Στη στοίχιση δεν συμπεριλήφθηκαν κάποιες πρωτεΐνες που σχολιάζονταν υποθετικές (hypothetical) ή ήταν θραύσματα (fragments). Επίσης, οι πρωτεϊνικές αλληλουχίες του Παράγοντα έναρξης της μετάφρασης IF2 χωρίστηκαν σε 3 διαφορετικές ομάδες με βάση τις C-τελικές συντηρημένες περιοχές τους (για εξελικτικούς λόγους, επειδή οι αλληλουχίες αποκλίνουν αρκετά). Επιπλέον, τα LCRs που ανιχνεύθηκαν στις πρωτεΐνες της κάθε κατηγορίας ομαδοποιήθηκαν με τη Matlab, όπως έχει προαναφερθεί.

Στον Πίνακας 28 φαίνονται οι μεγάλες κατηγορίες πρωτεϊνών που προσδένουν DNA ή RNA, ο αριθμός των ομολόγων που στοιχήθηκαν και μια σύντομη περιγραφή των LCRs τους.

| **Βακτήρια** | | |
|---|---|---|
| **Σχολιασμός πρωτεΐνης** | **Αριθμός ομολόγων** | **Περιγραφή των LCRs** |
| Πολυριβονουκλεοτιδική νουκλεοτιδυλοτρανσφεράση | 63 | Πλούσια σε G-R-D. Τα LCR βρίσκονται στο C-τελικό άκρο |
| Τοποϊσομεράση του DNA | 299 | Πλούσια σε K-A-T (μερικά έχουν R, P or S). Τα LCR βρίσκονται στο C-τελικό άκρο |
| Παράγοντας έναρξης της μετάφρασης IF2 (όλες οι ομάδες) | 234 | Διακρίνονται 2 ομάδες: 1)Η πλούσια σε G, 2)η πλούσια σε A-P |
| Παράγοντας έναρξης της μετάφρασης IF2 (ομάδα 1) | 173 | Μια περιοχή πλούσια σε A-P ή μια περιοχή πλούσια σε E με Rs και Ks προς το N-τελικό άκρο, που ακολουθείται από μια περιοχή πλούσια σε G (με Ns, και Qs) στη μέση της πρωτεΐνης |
| Παράγοντας έναρξης της μετάφρασης IF2 (ομάδα 2) | 38 | 1)Μια ομάδα πλούσια σε A με Ps και Es και 2) και μια ομάδα πλούσια σε G με Rs. Τα LCRs βρίσκονται στο μέσο της πρωτεΐνης |

| | | |
|---|---|---|
| Παράγοντας έναρξης της μετάφρασης IF2 (ομάδα 3) | 23 | Πλούσια σε A με E, K και Rs. Τα LCRs βρίσκονται στο μέσο της πρωτεΐνης |
| Πρωτεΐνη πρόσδεσης μονόκλωνου DNA | 150 | Πλούσια σε G και λιγότερα από αυτά πλούσια σε Q. Στις επαναλήψεις του G παρεμβάλονται S,N και λιγότερο Y (εκτός από το Q). Τα LCR βρίσκονται προς το C-τελικό άκρο ακολουθούμενα από ένα μοτίβο 6-7 αμινοξέων. |
| Ριβονουκλεάση E | 64 | 1) Μια ομάδα πλούσια σε A με P (μερικά E, Vs), 2) μια ομάδα πλούσια σε G-R (μερικά N και D). Μερικά LCRs βρίσκονται προς το N-τελικό άκρο (κυρίως πλούσια σε D-E). Τα περισσότερα από τα LCR βρίσκονται στη μέση της πρωτεΐνης (A-D-E και A-P ακολουθούμενα από G-R που ακολουθείται από E-A-D που ακολουθείται με τη σειρά του από περιοχή πλούσια σε R με Gs ακολουθούμενη τελικά από μια περιοχή A,E,P,V προς το C-τελικό άκρο) |
| Πρωτεΐνη πρόσδεσης RNA | 131 | Πλούσια σε G με Rs (και μερικά Y, S). Τα LCR βρίσκονται στο C-τελικό άκρο (πλούσια σε G με μερικά Rs και ένα μόνο Y ή S μεταξύ των επαναλήψεων του G) |

***Πίνακας 28.*** *Βακτήρια: Οι πιο συχνές κατηγορίες πρωτεϊνών που προσδένουν DNA/RNA και στις οποίες ανιχνεύτηκαν LCRs, αριθμός των ομολόγων τους που στοιχήθηκαν και περιγραφή των LCRs τους*

Χαρακτηριστικό παράδειγμα περιοχών χαμηλής πολυπλοκότητας που εμπλέκονται στην πρόσδεση μορίων DNA και RNA, αποτελούν τα μοτίβα RGG, τα οποία προσδένουν είτε RNAs, είτε πρωτεΐνες (όπως τον eIF4, ιικές πρωτεΐνες κ.α) και παίζουν καθοριστικό ρόλο στη μετα- μεταγραφική ρύθμιση (Rajyaguru and Parker, 2012; Thandapani et al., 2013). Οι πρωτεϊνικές αλληλεπιδράσεις των μοτίβων RGG μπορούν, επίσης, να ρυθμιστούν με μεθυλίωση της αργινίνης από μη - ιστονικές μεθυλτρανσφεράσες (Thandapani et al., 2013).

Η ικανότητά τους να προσδένουν RNA αναγνωρίστηκε για πρώτη φορά το 1992 και βρέθηκαν επίσης αρωματικά αμινοξέα, υποδεικνύοντας ότι συμβάλλουν στην αλληλεπίδραση (stacking) με τις βάσεις του RNA (Kiledjian and Dreyfuss, 1992). Τέτοιες επαναλήψεις γλυκίνης ακολουθούμενες από αρωματικά αμινοξέα εντοπίστηκαν σε πολλά από τα προκαρυωτικά LCRs που ανιχνεύθηκαν σε αυτή την ανάλυση και περιλαμβάνουν πιο συχνά επαναλήψεις των μοτίβων GGY και GGF και λιγότερο συχνά του μοτίβου GGW. Οι πρωτεΐνες στις οποίες ανήκουν, έχουν σχολιασμούς ή όρους - οντολογίες σχετικούς με πρόσδεση, μεταφορά ή επεξεργασία μορίων RNA και DNA.

Επιπλέον, το μοτίβο RGG βρέθηκε ότι προσδένει δευτεροταγείς δομές του RNA, γνωστές ως G-quartets (Hanakahi et al., 1999). Μεγάλης κλίμακας πειράματα αλληλεπίδρασης RNA-πρωτεΐνης έχουν αποκαλύψει την ύπαρξη μοτίβων RGG και YGG σε πολλές και ποικίλες πρωτεΐνες που προσδένουν RNA (Castello et al., 2016, 2012). Δομικές αναλύσεις που βασίζονται στον κυκλικό διχρωισμό και στον πυρηνικό μαγνητικό συντονισμό έδειξαν ότι δύο από τις αργινίνες σχηματίζουν ενδομοριακές επαφές με τον διπλό-τετραπλό κόμβο του RNA (RNA duplex-quadruplex junction), ενώ οι γλυκίνες μεταξύ των αργινινών λειτουργούν σαν εύκαμπτοι σύνδεσμοι (Phan et al., 2011). Το μοτίβο έχει αναφερθεί ότι εντοπίζεται τόσο στα βακτήρια όσο και στα αρχαία, αλλά δεν πραγματοποιήθηκε περαιτέρω μελέτη (Corley and Gready, 2008).

Επικράτειες με ενδογενή δομική αστάθεια, που αποτελούνται από μοτίβα RGG/RG, παρέχουν σε πρωτεΐνες “εκφυλισμένη” εξειδίκευση στην πρόσδεση RNA (Ozdilek et al., 2017). Αλληλεπιδράσεις πρωτεϊνών με το RNA in vitro μπορεί να προκύψουν μη φυσιολογικά από τις ηλεκτροστατικές ιδιότητες του RNA (Oma et al., 2004; Oma Yoko et al., 2009).

Δύο φαίνεται να είναι οι κυριότεροι τρόποι δέσμευσης του RNA από αυτές τις περιοχές: αλληλεπίδραση με βάσεις (πιθανότατα με αρωματικά αμινοξέα) ή με το phosphor-sugar backbone του DNA (πιθανότατα με θετικά φορτισμένες ομάδες αμινοξέων).

Περιοχές πρόσδεσης του RNA που ανιχνεύθηκαν στην εργασία του Castello (Castello et al., 2016) καθώς και περιοχές χαμηλής πολυπλοκότητας που προσδένουν RNA και παρέχονται στην εργασία (Järvelin et al., 2016) έχουν πολύ παρόμοιο αμινοξικό περιεχόμενο με τα LCRs που βρέθηκαν να προσδένουν RNA σε αυτή την εργασία.

Όσον αφορά τη δέσμευση μορίων DNA έχει δειχθεί ότι η πλούσια σε λυσίνη περιοχή χαμηλής πολυπλοκότητας καρβοξιτελικό άκρο τόσο της τοποϊσομεράσης του DNA (Strzalka et al., 2017) όσο και των πρωτεϊνών KU (Kushwaha and Grove, 2013) μπορούν να προσδένουν το DNA.

Ένα μοναδικό χαρακτηριστικό της πρωτεΐνης Ku του *Mycobacterium smegmatis* είναι η καρβοξυτελική της ουρά, που περιέχει μια περιοχή χαμηλής πολυπλοκότητας με πολλές, πλούσιες σε λυσίνη, επαναλήψεις PAKKA. Αυτές οι επαναλήψεις απουσιάζουν από ομόλογα που κωδικοποιούνται από υποχρεωτικά παρασιτικά μυκοβακτήρια. Τέτοιες επαναλήψεις PAKKA εντοπίζονται επίσης στην Hlp (πρωτεΐνη που μοιάζει με ιστόνη, Histone-like protein) των μυκοβακτηρίων, στην οποία έχει αποδειχθεί ότι προσδίδουν την ικανότητα να ενώνει τα άκρα του DNA. Οι επαναλήψεις PAKKA συμβάλλουν επίσης στην αποτελεσματική ένωση των άκρων του DNA (Kushwaha and Grove, 2013).

Τέλος, έχει δειχθεί ότι και ριβοσωμικές πρωτεΐνες μπορούν να προσδένουν DNA μέσω ηλεκτροστατικών αλληλεπιδράσεων των περιοχών χαμηλής πολυπλοκότητας, που είναι πλούσιες σε K ή R και έχουν θετικό φορτίο με τα αρνητικά φορτισμένα μόρια DNA και RNA (Klein et al., 2004; Peng et al., 2014; Wool, 1996).

### 3.3.2 Ριβοσωμικές πρωτεΐνες

Συγκεντρώθηκαν 822 βακτηριακά και 117 αρχαϊκά LCRs που ανήκουν σε ριβοσωμικές πρωτεΐνες και λήφθηκε η πρωτεϊνική τους ακολουθία από τη Uniprot. Τα ορθόλογα ομαδοποιήθηκαν και κρατήθηκαν για ανάλυση οι ομάδες που είχαν περισσότερες από 20 πρωτεϊνικές ακολουθίες, για τα βακτήρια και περισσότερες από 10 πρωτεϊνικές ακολουθίες, για τα αρχαία, με αποτέλεσμα να αναλυθούν τελικά 11 ομάδες βακτηριακών και 4 ομάδες αρχαϊκών ορθόλογων ριβοσωμικών πρωτεϊνών.

Οι βακτηριακές ομάδες που αναλύθηκαν ήταν:

Για την 30s υπομονάδα:

1. S2 (102 ορθόλογες πρωτεϊνικές ακολουθίες)
2. S3 (47 ορθόλογες πρωτεϊνικές ακολουθίες)
3. S6 (46 ορθόλογες πρωτεϊνικές ακολουθίες)
4. S16 (186 ορθόλογες πρωτεϊνικές ακολουθίες)

Για την 50s υπομονάδα:

5. L3 (30 ορθόλογες πρωτεϊνικές ακολουθίες)
6. L10 (32 ορθόλογες πρωτεϊνικές ακολουθίες)
7. L17 (40 ορθόλογες πρωτεϊνικές ακολουθίες)
8. L19 (32 ορθόλογες πρωτεϊνικές ακολουθίες)
9. L21 (45 ορθόλογες πρωτεϊνικές ακολουθίες)
10. L25 (128 ορθόλογες πρωτεϊνικές ακολουθίες)
11. L31 (21 ορθόλογες πρωτεϊνικές ακολουθίες)

Η 50S ριβοσωμική πρωτεΐνη L25 διαιρέθηκε αργότερα σε 2 μικρότερες ομάδες (για εξελικτικούς λόγους) την L25_1 (75 ορθόλογες πρωτεϊνικές ακολουθίες) και την L25_2 (53 ορθόλογες πρωτεϊνικές ακολουθίες).

Οι αρχαϊκές ομάδες που αναλύθηκαν ήταν:

Για την 30s υπομονάδα:

1. S3 (10 ορθόλογες πρωτεϊνικές ακολουθίες)
2. S24 (13 ορθόλογες πρωτεϊνικές ακολουθίες)

Για την 50s υπομονάδα:

3. L10(14 ορθόλογες πρωτεϊνικές ακολουθίες)
4. L12(40 ορθόλογες πρωτεϊνικές ακολουθίες)

Για τα κομμάτια LCR της κάθε ριβοσωμικής πρωτεΐνης (από τις μεγαλύτερες, επιλεγμένες ομάδες) υπολογίστηκε το αντίστοιχο διάνυσμα συχνότητας για τα αμινοξέα. Οι Πίνακας 29 και Πίνακας 30 δείχνουν κάθε ριβοσωμική πρωτεΐνη, μαζί με ένα διάνυσμα των 3 αμινοξέων που απαντώνται πιο συχνά στα LCRs όλων των ριβοσωμικών πρωτεϊνών.

| **Βακτήρια** | | | |
|---|---|---|---|
| **Πρωτεΐνη** | **A** | **E** | **K** |
| Όλες οι βακτηριακές ριβοσωμικές πρωτεΐνες | 9279 | 4901 | 2075 |
| S16 | 2726 | 1614 | 384 |
| S2 | 1203 | 711 | 116 |
| S3 | 63 | 37 | 45 |
| S6 | 296 | 179 | 39 |
| L10 | 364 | 197 | 10 |
| L17 | 385 | 293 | 115 |

| | | | |
|---|---|---|---|
| L19 | 452 | 163 | 99 |
| L21 | 617 | 110 | 474 |
| L25_1 | 386 | 679 | 14 |
| L25_2 | 972 | 64 | 225 |
| L3 | 642 | 291 | 49 |
| L31 | 329 | 137 | 143 |

***Πίνακας 29.*** *Βακτήρια: Μεγαλύτερες ομάδες ριβοσωμικών πρωτεϊνών με LCRs μαζί με τα 3 αμινοξέα που εμφανίζονται πιο συχνά στα ριβοσωμικά LCRs*

| **Αρχαία** | | | |
|---|---|---|---|
| **Πρωτεΐνη** | **A** | **E** | **K** |
| Όλες οι αρχαϊκές ριβοσωμικές πρωτεΐνες | 925 | 1119 | 337 |
| S24 | 44 | 91 | 45 |
| S3 | 59 | 118 | 4 |
| L10 | 90 | 170 | 74 |
| L12 | 465 | 445 | 62 |

***Πίνακας 30.*** *Αρχαία: Μεγαλύτερες ομάδες ριβοσωμικών πρωτεϊνών με LCRs μαζί με τα 3 αμινοξέα που εμφανίζονται πιο συχνά στα ριβοσωμικά LCRs*

Οι ριβοσωμικές πρωτεΐνες της κάθε ομάδας στοιχίστηκαν με χρήση του muscle. Στη συνέχεια υπολογίστηκε το δέντρο για κάθε στοιχισμένη ομάδα χρησιμοποιώντας το Jalview (Μέθοδος μέσης απόστασης με χρήση του BLOSUM62) και σε κάθε στοίχιση οι ακολουθίες ταξινομήθηκαν με βάση το αντίστοιχο δέντρο. Οι στοιχίσεις οπτικοποιήθηκαν με το SeaView. Επιπλέον, τα LCRs που ανιχνεύθηκαν στις πρωτεΐνες της κάθε ομάδας ριβοσωμικών πρωτεϊνών, ομαδοποιήθηκαν με τη Matlab, όπως έχει προαναφερθεί.

Στους Πίνακας 31 και Πίνακας 32 φαίνεται κάθε ομάδα των ριβοσωμικών πρωτεϊνών (για τα βακτήρια και τα αρχαία) και μια σύντομη περιγραφή της φύσης των LCRs που περιέχονται στις πρωτεΐνες της.

| **Βακτήρια** | |
|---|---|
| **Πρωτεΐνη** | **Περιγραφή των LCRs** |
| **Μικρή ριβοσωμική υπομονάδα (30S)** | |
| S2 | C-τελικό άκρο, μερικά πλούσια σε A με Es, μερικά πολύ όξινα, μερικά έχουν πολλά Ks |
| S3 | C-τελικό άκρο, πλούσια σε G με βασικά |
| S6 | C-τελικό άκρο, μερικά πλούσια σε G με όξινα και βασικά, μερικά πλούσια σε A με όξινα, μερικά πολύ όξινα |

| | |
|---|---|
| S16 | C-τελικό άκρο, πλούσια σε A με όξινα |
| **Μεγάλη ριβοσωμική υπομονάδα (50S)** | |
| L3 | C-τελικό άκρο, πλούσια σε A με πολλά Es και με Ks στο τέλος |
| L10 | C-τελικό άκρο, πλούσια σε A με πολλά Es |
| L17 | C-τελικό άκρο, πλούσια σε A με πολλά Es. Πολλά έχουν μια βασική περιοχή στην αρχή του LCR |
| L19 | C-τελικό άκρο, πλούσια σε A με πολλά Es |
| L21 | C-τελικό άκρο, πλούσια σε A με πολλά Ks, μερικά έχουν μια περιοχή που μοιάζει με επικράτεια μετά το LCR |
| L25 | C-τελικό άκρο, μια ομάδα με όξινη ουρά, μια 2η ομάδα πλούσια σε A με K στο τέλος (έχει G, P) |
| L31 | C-τελικό άκρο, πλούσια σε A με μια βασική περιοχή στην αρχή και στα μισά από αυτά ακολουθεί μια όξινη περιοχή |

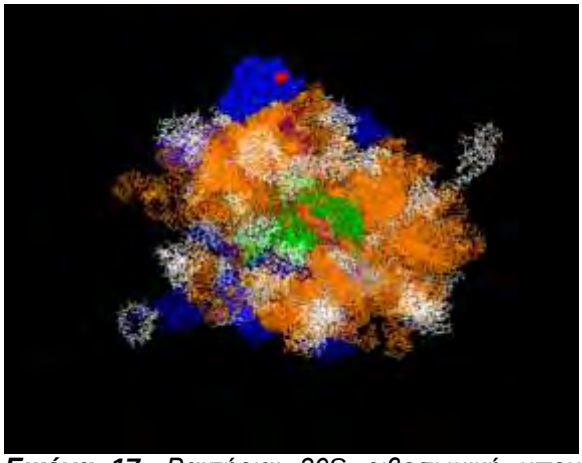
***Πίνακας 31.*** *Βακτήρια: Ομάδες των ριβοσωμικών πρωτεϊνών και περιγραφή των LCRs τους*

| **Αρχαία** | |
|---|---|
| **Πρωτεΐνη** | **Περιγραφή των LCRs** |
| **Μικρή ριβοσωμική υπομονάδα (30S)** | |
| S3 | C-τελικό άκρο, πολύ όξινη ουρά |
| S24 | C-τελικό άκρο, μερικά όξινα, μερικά βασικά |
| **Μεγάλη ριβοσωμική υπομονάδα (50S)** | |
| L10 | C-τελικό άκρο, πολύ όξινη ουρά, μερικά έχουν μια πολύ βασική ουρά |
| L12 | Προς το C-τελικό άκρο, πλούσιο σε A ακολουθούμενο από μια πολύ όξινη περιοχή ακολουθούμενη από ένα συντηρημένο μοτίβο 7 αμινοξέων |

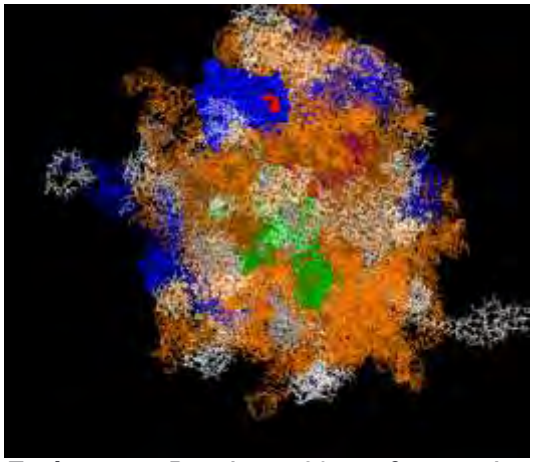
***Πίνακας 32.*** *Αρχαία: Ομάδες των ριβοσωμικών πρωτεϊνών και περιγραφή των LCRs τους*

Για την οπτικοποίηση της τρισδιάστατης δομής του ριβοσώματος των αρχαίων και των βακτηρίων χρησιμοποιήθηκε το PyMOL. Για αυτό το σκοπό, λήφθηκαν από την RCSB Protein Data Bank (PDB), οι καταχωρίσεις για την 50S (3I8I) και την 30S (3I8H) ριβοσωμικές υπομονάδες του *Thermus Thermophilus* και 50S(3J21), 30S(3J20) ριβοσωμικές υπομονάδες και του 50S ριβοσωμικού RNA(3J2L) του *Pyrococcus furiosus* (η πρωτεΐνη L12 από τον *Pyrococcus horikoshii*, 3WY9).

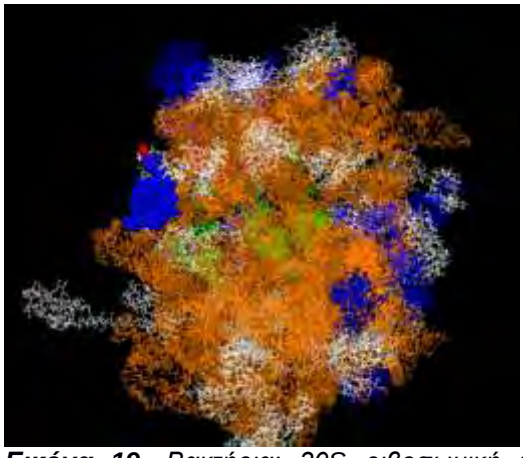
Οι πρωτεΐνες των 11 μεγαλύτερων ομάδων, που προαναφέρθηκαν, χρωματίστηκαν μπλε, το ριβοσωμικό RNA και των 2 υπομονάδων χρωματίστηκε πορτοκαλί, τα tRNAs χρωματίστηκαν πράσινα και το mRNA ρόζ. Στις Εικόνα 17, Εικόνα 18, Εικόνα 19, Εικόνα 20, Εικόνα 21, Εικόνα 22, Εικόνα 23, Εικόνα 24, Εικόνα 25, Εικόνα 26, Εικόνα 27, Εικόνα 28, Εικόνα 29, Εικόνα 30 και Εικόνα 31, που απεικονίζουν το ριβόσωμα, κάθε πρωτεΐνη μαζί με τα κατάλοιπα του καρβοξυτελικού άκρου (τα οποία χρωματίστηκαν κόκκινα) παρουσιάζονται σαν σφαίρες.



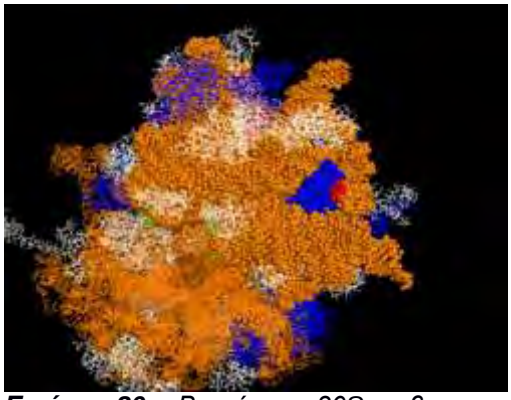
***Εικόνα 17***. *Βακτήρια: 30S ριβοσωμική υπομονάδα, S2, κατάλοιπα του C-τελικού άκρου στην επιφάνεια*



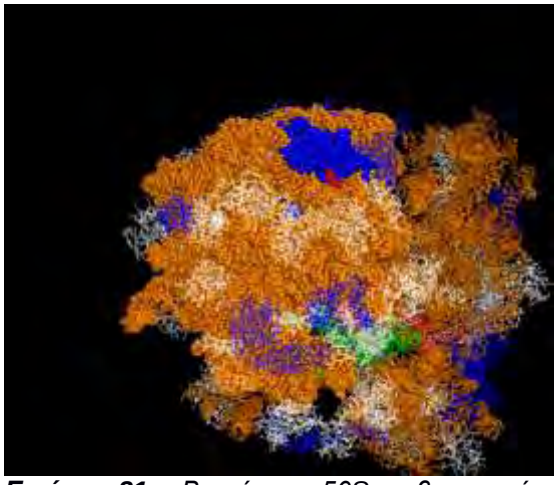
***Εικόνα 18.*** *Βακτήρια: 30S ριβοσωμική υπομονάδα, S3, κατάλοιπα του C-τελικού άκρου στην επιφάνεια*



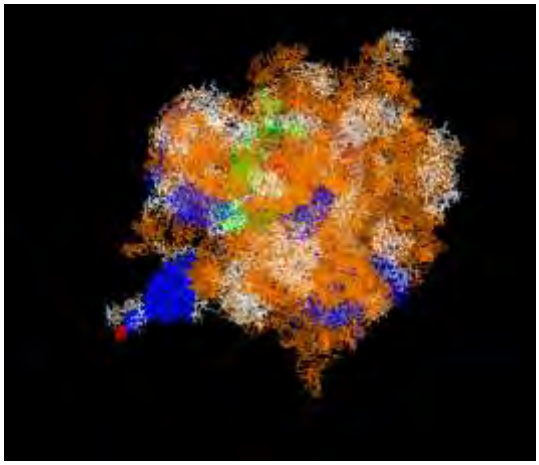
***Εικόνα 19.*** *Βακτήρια: 30S ριβοσωμική υπομονάδα, S6, κατάλοιπα του C-τελικού άκρου στην επιφάνεια*



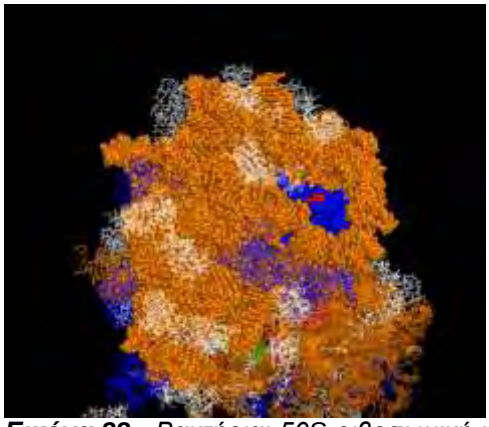
***Εικόνα 20.*** *Βακτήρια: 30S ριβοσωμική υπομονάδα, S16, κατάλοιπα του C-τελικού άκρου αλληλεπιδρούν με το rRNA*



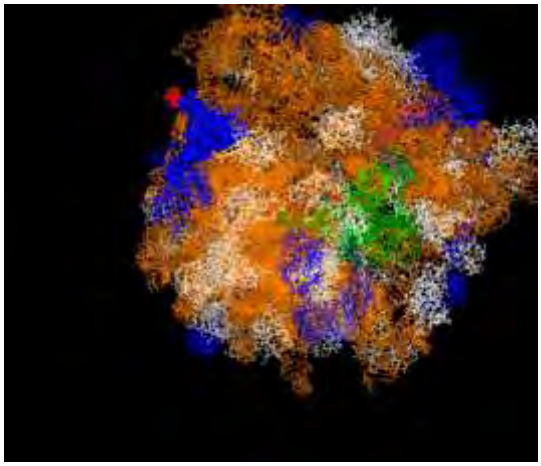
***Εικόνα 21.*** *Βακτήρια: 50S ριβοσωμική υπομονάδα, L3, κατάλοιπα του C-τελικού άκρου αλληλεπιδρούν με το rRNA*



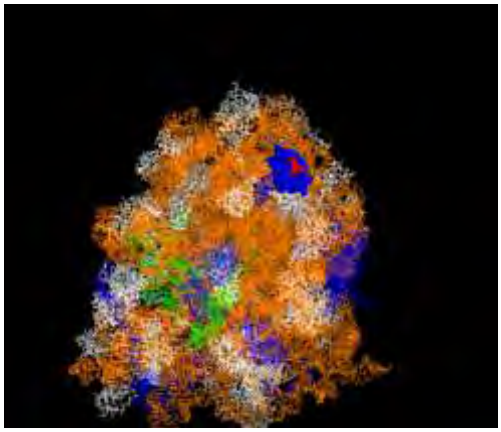
***Εικόνα 22.*** *Βακτήρια: 50S ριβοσωμική υπομονάδα, L10, κατάλοιπα του C-τελικού άκρου στην επιφάνεια*



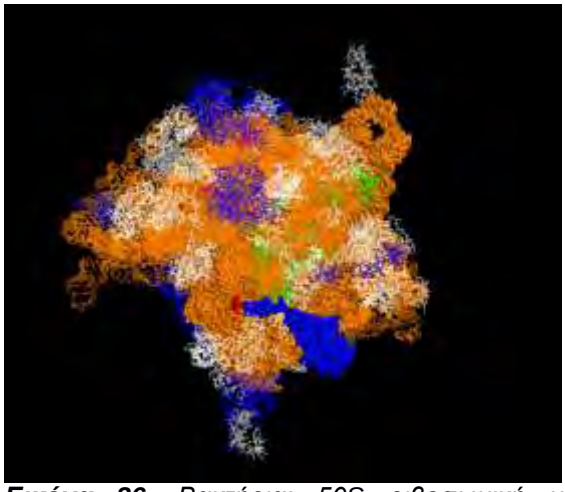
***Εικόνα 23.*** *Βακτήρια: 50S ριβοσωμική υπομονάδα, L17, κατάλοιπα του C-τελικού άκρου πολύ κοντά στο rRNA*



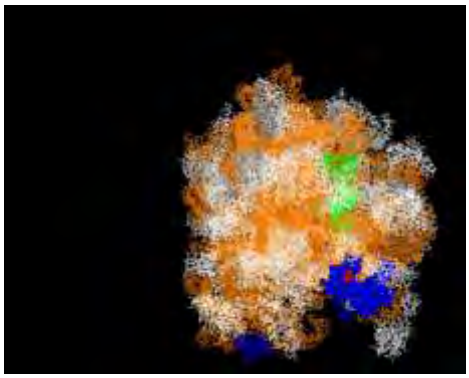
***Εικόνα 24.*** *Βακτήρια: 50S ριβοσωμική υπομονάδα, L19, κατάλοιπα του C-τελικού άκρου στην επιφάνεια*



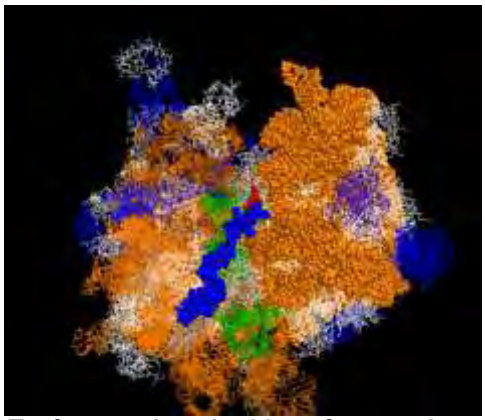
***Εικόνα 25.*** *Βακτήρια: 50S ριβοσωμική υπομονάδα, L21, κατάλοιπα του C-τελικού άκρου στην επιφάνεια*



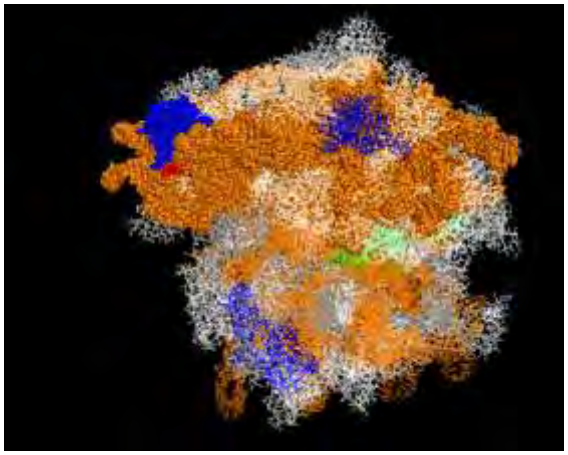
***Εικόνα 26.*** *Βακτήρια: 50S ριβοσωμική υπομονάδα, L25, κατάλοιπα του C-τελικού άκρου αλληλεπιδρούν με το rRNA*



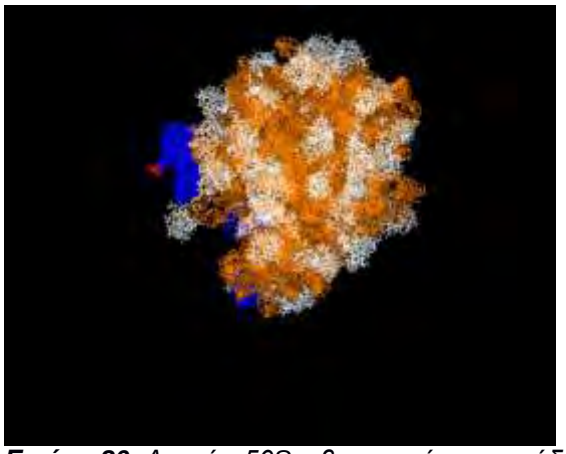
***Εικόνα 27.*** *Βακτήρια: 50S ριβοσωμική υπομονάδα, L31, κατάλοιπα του C-τελικού άκρου στην επιφάνεια*



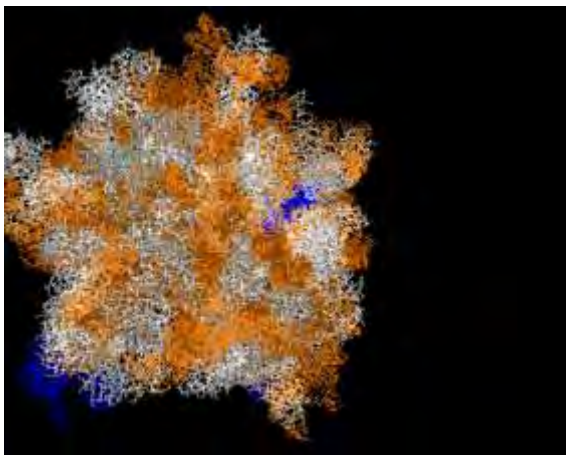
***Εικόνα 28.*** *Αρχαία: 30S ριβοσωμική υπομονάδα, S3, κατάλοιπα του C-τελικού άκρου στην επιφάνεια*



***Εικόνα 29.*** *Αρχαία: 30S ριβοσωμική υπομονάδα, S24, κατάλοιπα του C-τελικού άκρου αλληλεπιδρούν με το rRNA*



***Εικόνα 30.*** *Αρχαία: 50S ριβοσωμική υπομονάδα, L10, κατάλοιπα του C-τελικού άκρου στην επιφάνεια*



***Εικόνα 31.*** *Αρχαία: 50S ριβοσωμική υπομονάδα, L12, κατάλοιπα του C-τελικού άκρου αλληλεπιδρούν με το rRNA*

Φαίνεται, επομένως ότι τα LCRs των ριβοσωμικών πρωτεϊνών εντοπίζονται στο καρβοξυτελικό άκρο.

Από διάφορες μελέτες αποδεικνύεται ότι τα LCRs των ριβοσωμικών πρωτεϊνών έχουν σημαντικό ρόλο σε ένα πλήθος αλληλεπιδράσεων. Μπορούν να διευκολύνουν τις αλληλεπιδράσεις με άλλες πρωτεΐνες, με το DNA και το RNA (Brodersen Ditlev E. and Nissen Poul, 2005; Ciriello et al., 2010; Klein et al., 2004; Lanier et al., 2017; Perederina et al., 2002), καθώς και άλλων προσδετών. Μπορούν επίσης να συμμετέχουν στην πρόσδεση μετάλλων (Peng et al., 2014).

Υπάρχουν, επίσης, αναφορές ότι αυτές οι περιοχές διαδραματίζουν βασικό ρόλο στη συναρμολόγηση του ριβοσώματος (Brodersen et al., 2002; Garrett, 1983; Klein et al., 2004; Timsit et al., 2009). Έχει προταθεί ότι, οι μεγάλες βασικές επαναλήψεις των ριβοσωμικών πρωτεϊνών μπορούν να εισχωρήσουν στις ριβοσωμικές υπομονάδες, να αποκτήσουν σταθερή διαμόρφωση ή να αναδιπλωθούν μαζί με το RNA, διευκολύνοντας, έτσι, τη σωστή αναδίπλωση του rRNA (Timsit et al., 2009).

Σε γενικές γραμμές, τόσο τα μικρά όσο και τα μεγάλα σε μήκος LCRs των πρωτεϊνών του ριβοσώματος εμπλέκονται σε πολλές διεργασίες όπως αλληλεπιδράσεις με πρωτεΐνες, DNA και άλλους προσδέτες. Ωστόσο, τα πιο μικρά σε μήκος LCRs έχει δειχθεί ότι εμπλέκονται σε περισσότερες λειτουργίες, που περιλαμβάνουν την πρόσδεση RNA και ιόντων μετάλλου, λειτουργίες αυτο - ρύθμισης, πολυμερισμό κ.α. Τα μεγάλα σε μήκος LCRs χρησιμοποιούνται πιο συχνά σαν συνδέτες και παίζουν σημαντικό ρόλο στις ενδο-πρωτεϊνικές αλληλεπιδράσεις (Peng et al., 2014).

### 3.3.3 Πρωτεΐνες πλούσιες σε Ιστιδίνη (H)

Περιοχές πλούσιες σε ιστιδίνη, σε πρωτεΐνες, έχουν βρεθεί να εμπλέκονται σε διεργασίας πρόσδεσης μετάλλων. Έτσι, όλα τα βακτηριακά τμήματα LCRs που περιείχαν 5 ή περισσότερες ιστιδίνες συλλέχθηκαν για περαιτέρω ανάλυσή τους.

Για τα βακτήρια, από τις αρχικές 726 πρωτεΐνες (749 τμήματα LCRs), όλες οι μη χαρακτηρισμένες πρωτεΐνες (232 πρωτεΐνες) απομακρύνθηκαν, καταλήγοντας σε 494 πρωτεΐνες (509 τμήματα LCRs) με 207 διαφορετικούς σχολιασμούς, από 454 διαφορετικούς οργανισμούς.

Το μεγαλύτερο SAR ιστιδίνης είχε μήκος 22 αμινοξέα από τον οργανισμό *Acinetobacter guillouiae* και ήταν μια μη χαρακτηρισμένη πρωτεΐνη. Μετά την απομάκρυνση των μη χαρακτηρισμένων πρωτεϊνών, το μεγαλύτερο SAR είχε μήκος 13 αμινοξέα, ανήκε στον οργανισμό *Thermoactinomyces vulgaris* και ήταν ένας ZitB μεταφορέας ψευδαργύρου.

Οι περισσότερες από τις πρωτεΐνες που περιείχαν LCRs πλούσια σε ιστιδίνη (69%, 353/509 κομμάτια LCRs) βρέθηκαν να εμπλέκονται στην πρόσδεση ιόντων μετάλλου (με βάση τους σχολιασμούς και τους όρους – οντολογίες που τους είχαν αποδοθεί). Στον Πίνακας 33 παρουσιάζονται οι έξι κορυφαίοι, πιο συχνοί σχολιασμοί των πρωτεϊνών με LCRs πλούσια σε H (μετά το φιλτράρισμα με την απομάκρυνση των μη χαρακτηρισμένων).

| Σχολιασμός πρωτεΐνης | |
|---|---|
| 35 | Urease accessory protein UreE |
| 27 | Cobalamin biosynthesis protein CobW |
| 26 | Nickel/cobalt efflux system |
| 24 | Sirohydrochlorin cobaltochelatase |
| 20 | Protoporphyrin ferrochelatase |
| 16 | Cobalamin synthesis protein P47K |

***Πίνακας 33.*** *Βακτήρια: Κορυφαίοι 6 πιο συχνοί σχολιασμοί πρωτεϊνών που έχουν LCRs πλούσια σε Η, μαζί με τον αντίστοιχο αριθμό των πρωτεϊνών τους*

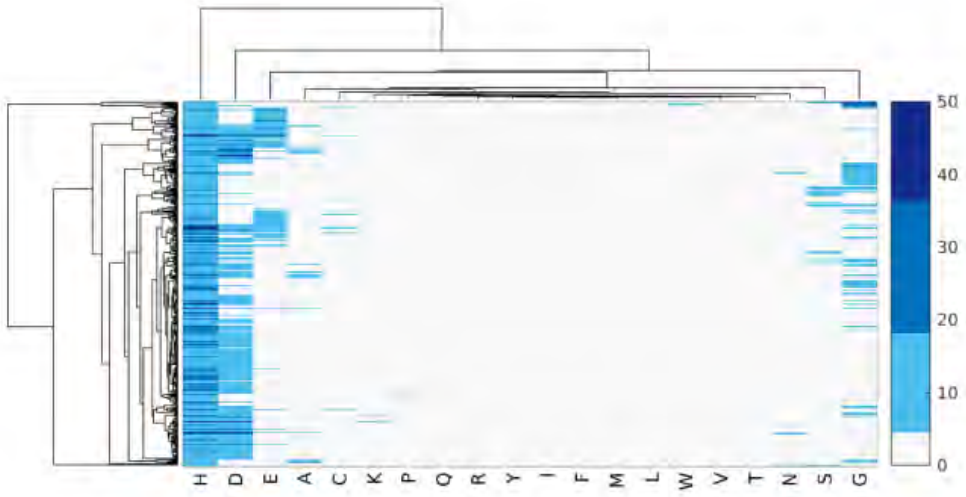
Από τις πρωτεΐνών με LCRs πλούσια σε Η η μεγάλη πλειοψηφία εμπλεκόταν στην πρόσδεση νικελίου και κοβαλτίου (28% και 37% αντίστοιχα, των πλούσιων σε Η LCRs που προσδένουν μέταλλα). Πιο συγκεκριμένα, πολλά από αυτά (23%, 82/353 τμήματα LCRs) εμπλέκονταν στη βιοσύνθεση και μεταφορά κοβαλαμίνης/βιταμίνης B12. Αυτές οι πρωτεΐνες ανακτήθηκαν και στοιχήθηκαν με χρήση των SeaView και Jalview.

Μεγάλο ενδιαφέρον παρουσιάζουν οι ετικέτες πολύ - Η, που αποτελούνται από 6 ή περισσότερες διαδοχικές ιστιδίνες και χρησιμοποιούνται για τον καθαρισμό πρωτεϊνών, με πρόσδεση σε στήλες με νικέλιο ή κοβάλτιο με μικρομοριακή συγγένεια (Bornhorst and Falke, 2000). Τα LCRs, που είναι πλούσια σε Η και βρίσκονται σε αυτή την κατηγορία πεπτιδίων, έχουν κατά μέσο όρο 34 αμινοξέα (μήκος) από τα οποία το 50% είναι ιστιδίνη, το 21% ασπαρτικό οξύ, το 8% γλουταμικό οξύ και 10% γλυκίνη.

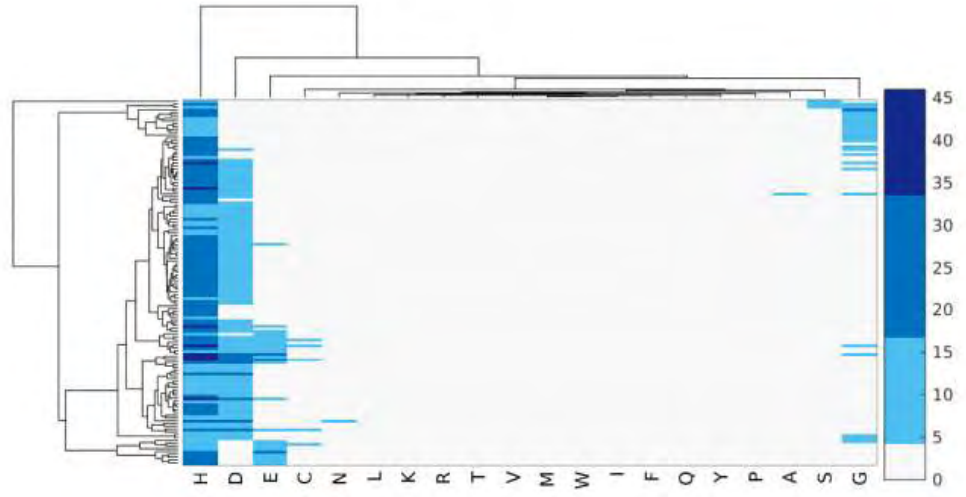
Στη συνέχεια τα διανύσματα συχνότητας για τα αμινοξέα:

- των πλούσιων σε Η LCRs που εμπλέκονται γενικότερα στην πρόσδεση μετάλλων,
- των πλούσιων σε Η LCRs που εμπλέκονται στην πρόσδεση κοβαλτίου/κοβαλαμίνης,
- των πλούσιων σε Η LCRs που εμπλέκονται στην πρόσδεση νικελίου και
- των πλούσιων σε Η LCRs που εμπλέκονται στην πρόσδεση κοβαλτίου/νικελίου,

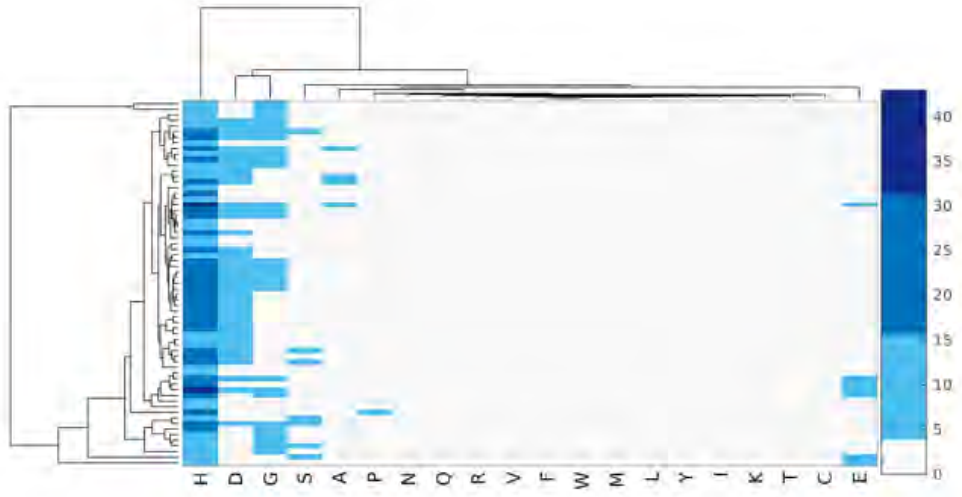
ομαδοποιήθηκαν και οπτικοποιήθηκαν με χρήση της Matlab και της συνάρτησης clustergram. Στα σχήματα των Εικόνα 32, Εικόνα 33, Εικόνα 34 και Εικόνα 35 παρουσιάζεται η ομαδοποίηση για τις 4 κατηγορίες.



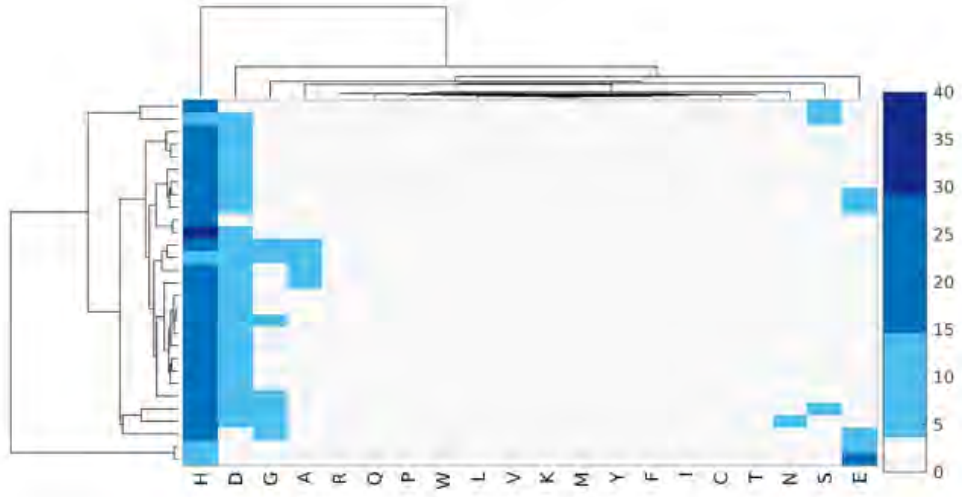

***Εικόνα 32.*** *Ομαδοποίηση των πλούσιων σε H LCRs που εμπλέκονται σε πρόσδεση μετάλλων γενικότερα*




***Εικόνα 33.*** *Ομαδοποίηση των πλούσιων σε H LCRs που εμπλέκονται σε πρόσδεση κοβαλτίου/κοβαλαμίνης*




***Εικόνα 34.*** *Ομαδοποίηση των πλούσιων σε H LCRs που εμπλέκονται σε πρόσδεση νικελίου*




***Εικόνα 35.*** *Ομαδοποίηση των πλούσιων σε H LCRs που εμπλέκονται σε πρόσδεση κοβαλτίου / νικελίου*

Είναι φανερό, ότι η ιστιδίνη σε συνδυασμό με το ασπαρτικό οξύ (και λιγότερο συχνά με το γλουταμικό οξύ και τη γλυκίνη) συμμετέχουν στη δημιουργία περιοχών χαμηλής πολυπλοκότητας που προσδένουν μέταλλα.

Αυτά τα φυσικά δημιουργημένα LCRs που προσδένουν νικέλιο-κοβάλτιο / κοβαλαμίνη, πιθανών έχουν βελτιστοποιηθεί από τη φυσική επιλογή για εκατομμύρια χρόνια και θα μπορούσαν να χρησιμοποιηθούν ως αφετηρία για διάφορες βιοτεχνολογικές εφαρμογές, ως ετικέτες που προσδένουν μέταλλα με υψηλή συγγένεια κατά τον καθαρισμό πρωτεϊνών, ή ως περιβαλλοντικοί βιοαισθητήρες.

LCRs πλούσια σε Η βρέθηκαν, επίσης και σε μερικούς υποδοχείς εξαρτώμενους από το TonB, που ανήκουν σε αρνητικά κατά gram βακτήρια (Flavobacteria, γ-proteobacteria). Οι υποδοχείς TonB, σε αυτό τον τύπο βακτηρίων, έχουν συνδεθεί μεταξύ άλλων και με την πρόσληψη και μεταφορά κοβαλαμίνης και συμπλόκων σιδεροφόρων - σιδήρου (Koebnik et al., 2000). Αυτό υποδεικνύει ότι τα πλούσια σε ιστιδίνη LCRs που περιέχονται στους υποδοχείς TonB, πιθανότατα έχουν κάποιο ρόλο στη μεσολάβηση της δέσμευσης αυτών των υποστρωμάτων.

Τα αποτελέσματα από τα πλούσια σε Η LCRs των Αρχαίων, δείχνουν ότι το 50% (11/21) των πλούσιων σε Η πρωτεϊνών εμπλέκονται στην πρόσδεση ιόντων μετάλλου.

Έχει δειχθεί ότι, σε επικράτειες που έχουν την ιδιότητα να προσδένουν μέταλλα, τα αμινοξέα που εμπλέκονται στον προσανατολισμό του μετάλλου προσομοιάζουν κατά πολύ στα παραπάνω ευρήματα (Dokmanić et al., 2008). Για παράδειγμα, τα 3 συχνότερα αμινοξέα που βρέθηκαν να εμπλέκονται στον προσανατολισμό κοβαλτίου είναι η ιστιδίνη (Η), το ασπαρτικό (D) και το γλουταμικό οξύ (E). Επομένως, αυτές οι επικράτειες θα μπορούσαν να έχουν εξελιχθεί από περιοχές χαμηλής πολυπλοκότητας που αποτελούνταν από τα αμινοξέα που κατευθύνουν μέταλλα.

### 3.3.4 Τσαπερόνες (Πρωτεΐνες συνοδοί)

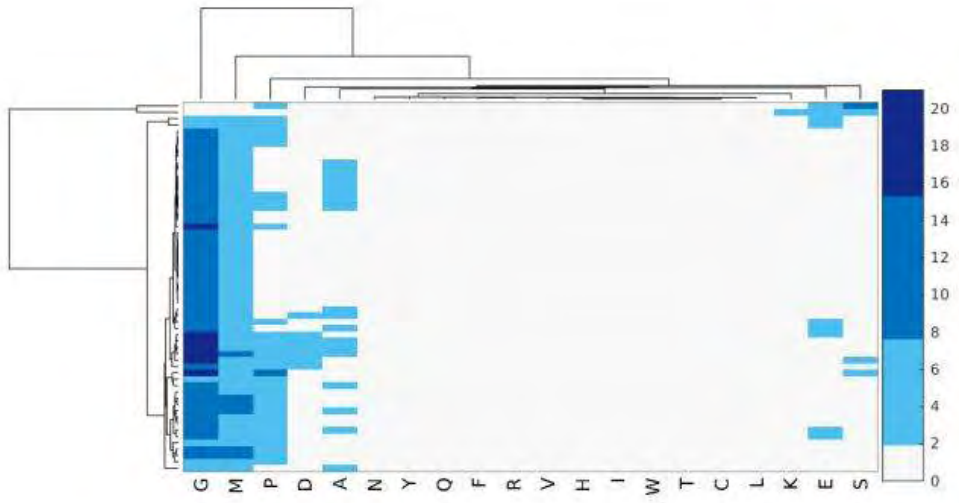
Οι τσαπερόνες (πρωτεΐνες συνοδοί) τόσο των βακτηρίων όσο και των αρχαίων που βρέθηκαν να έχουν LCRs συγκεντρώθηκαν, καταλήγοντας σε 434 τμήματα LCRs (340 βακτηριακά και 94 αρχαϊκά LCRs). Τα LCRs των βακτηρίων στη συντριπτική πλειοψηφία τους ανήκαν σε 3 μεγάλες ομάδες πρωτεϊνών συνοδών: την 60κDa τσαπερόνη, την τσαπερόνη DnaJ και την DnaK. Αντίστοιχα, τα LCRs των αρχαίων ανήκαν σε 4 μεγάλες ομάδες: το θερμόσωμα, την τσαπερόνη DnaK, την DnaJ και την GroEL.

Το θερμόσωμα είναι μια τσαπερονίνη της ομάδας II με δομή διπλού οκταμερούς δακτυλίου, που εντοπίζεται στα αρχαία (Zhang et al., 2013). Η GroEL (Hsp60) είναι μια μοριακή πρωτεΐνη συνοδός της ομάδας I με δομή διπλού ομοεπταμερούς δακτυλίου, που εντοπίζεται στα βακτήρια (Zhang et al., 2013).

Οι πρωτεϊνικές ακολουθίες, για τα LCRs της κάθε ομάδας, λήφθηκαν από τη Uniprot χρησιμοποιώντας το αναγνωριστικό της πρωτεΐνης στην οποία ανήκαν και ακολούθησε στοίχισή τους με χρήση του Muscle και του Seaview για οπτικοποίηση των αποτελεσμάτων. Τόσο στα αρχαία όσο και στα βακτήρια παρατηρήθηκε στο θερμόσωμα και στην GroEL (και λιγότερο στην DnaK) αλλά και στην 60kDa τσαπερόνη αντίστοιχα μια αρκετά συντηρημένη καρβοξυ - τελική περιοχή, χαμηλής πολυπλοκότητας πλούσια σε διαδοχικές επαναλήψεις GGM. Από τη βιβλιογραφία είναι γνωστό ότι η καρβοξυ - τελική περιοχή της GroEL είναι πλούσια σε μοτίβα GGM που βοηθούν τον εγκλεισμό της πρωτεΐνης-υποστρώματος στο εσωτερικό του δακτυλίου κι επίσης, συμμετέχουν άμεσα στην αναδίπλωση της πρωτεΐνης (Weaver and Rye, 2014). Επίσης, η καρβοξυ - τελική περιοχή της υπομονάδας του θερμοσώματος βρέθηκε ότι επηρεάζει τη συναρμολόγηση και τη θερμική σταθερότητα του συμπλόκου (Zhang et al., 2013). Όσον αφορά την τσαπερόνη

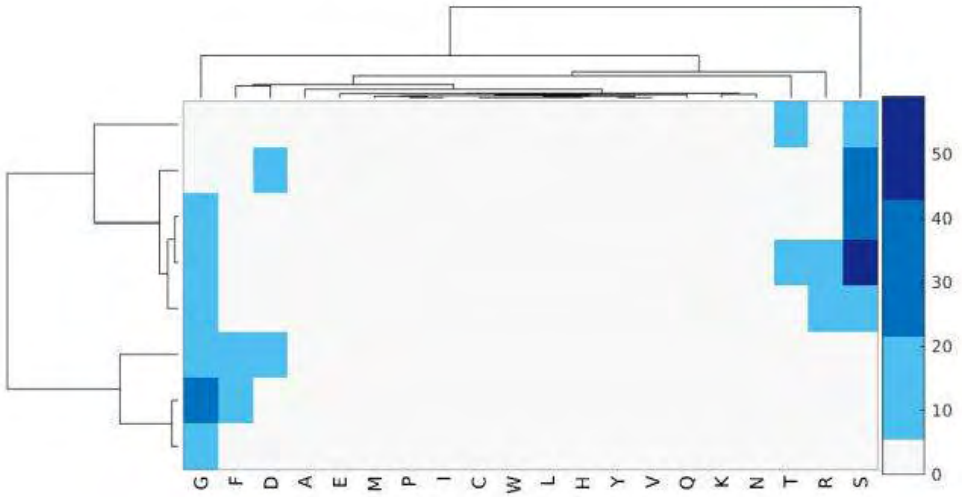
DnaJ, μεταλλάξεις στην πρωτεΐνη Sis1, την ομόλογη της DnaJ στο ζυμομύκητα, έδειξαν ότι η πλούσια σε GGF περιοχή, αμέσως μετά την επικράτεια J, απαιτείται για τη σωστή λειτουργία της πρωτεΐνης και δεν είναι απλώς μια συνδετική επικράτεια, όπως θεωρούνταν αρχικά (Yan and Craig, 1999).

Στη συνέχεια, πραγματοποιήθηκε ομαδοποίηση των LCRs που βρέθηκαν στις τσαπερόνες των αρχαίων και των βακτηρίων, με τον τρόπο που έχει προαναφερθεί, και τα αποτελέσματα φαίνονται στα σχήματα των Εικόνα 36, Εικόνα 37, Εικόνα 38, Εικόνα 39, Εικόνα 40, Εικόνα 41 και Εικόνα 42.

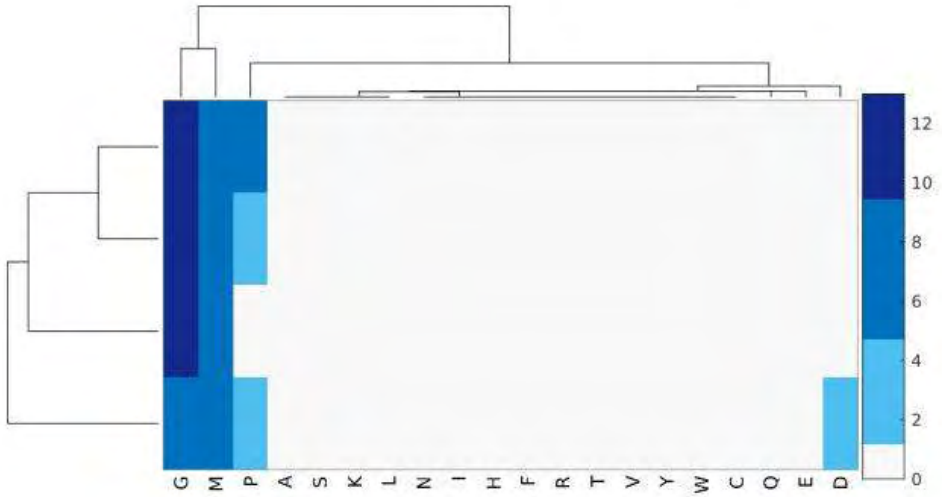



***Εικόνα 36.*** *Αρχαία: Ομαδοποίηση των LCRs του θερμοσώματος*

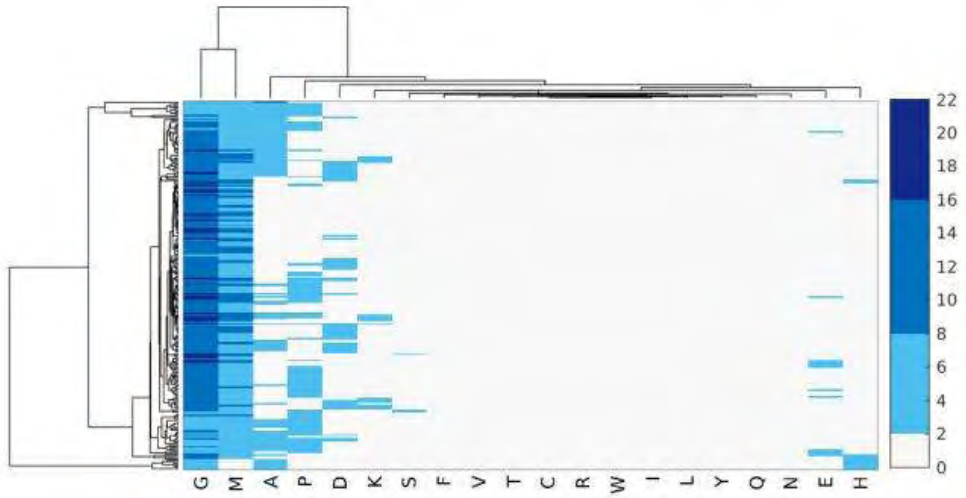

***Εικόνα 37.*** *Αρχαία: Ομαδοποίηση των LCRs της τσαπερόνης DnaK*



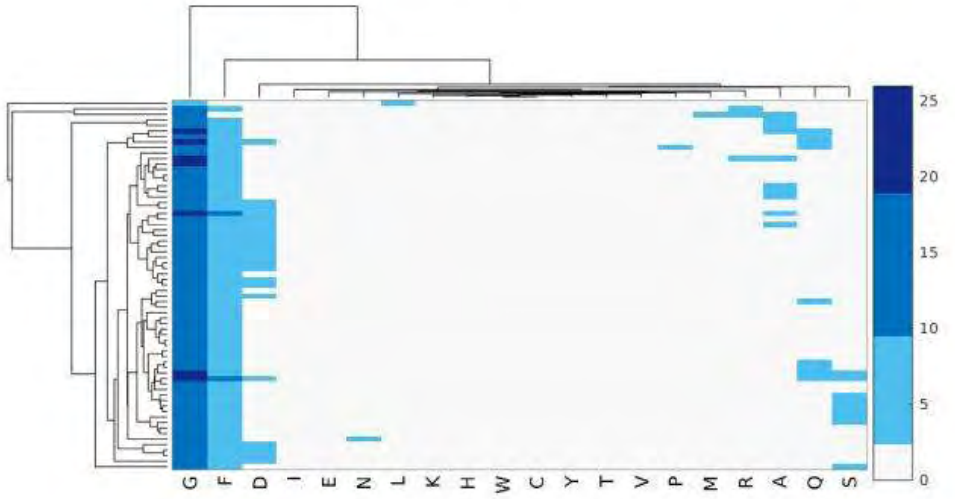

***Εικόνα 38.*** *Αρχαία: Ομαδοποίηση των LCRs της τσαπερόνης DnaJ*



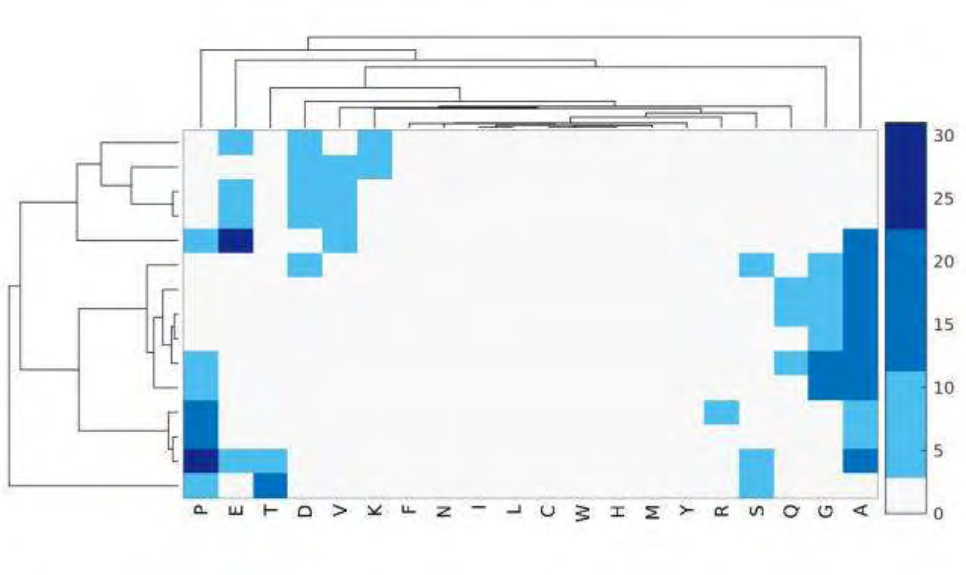

***Εικόνα 39.*** *Αρχαία: Ομαδοποίηση των LCRs της τσαπερόνης GroEL*




***Εικόνα 40.*** *Βακτήρια: Ομαδοποίηση των LCRs της τσαπερόνης 60kDa*




***Εικόνα 41.*** *Βακτήρια: Ομαδοποίηση των LCRs της τσαπερόνης DnaJ*




***Εικόνα 42.*** *Βακτήρια: Ομαδοποίηση των LCRs της τσαπερόνης DnaK*

Με βάση τα αποτελέσματα που παρατηρήθηκαν μετά τη στοίχιση των ομολόγων και την ομαδοποίηση των LCRs για κάθε κατηγορία τσαπερονών, τόσο για τα αρχαία όσο και για τα βακτήρια, δημιουργήθηκαν οι Πίνακας 34 και Πίνακας 35 με το όνομα της πρωτεΐνης, τον αριθμό των ομολόγων που στοιχήθηκαν και μια σύντομη περιγραφή των LCRs της.

| **Αρχαία** | | |
|---|---|---|
| **Σχολιασμός πρωτεΐνης** | **Αριθμός ομολόγων** | **Περιγραφή των LCRs** |
| Θερμόσωμα | 58 | Πλούσια σε G-M-P (με D,E,As). Τα LCRs βρίσκονται στο C-τελικό άκρο. Ένα μοτίβο D-E motif (στα περισσότερα) ακολουθούμενο από περιοχή πλούσιες σε G με Ms, Ps ή As (σε σχεδόν όλα). |
| DnaK τσαπερόνη | 18 | 1) Ομάδα πλούσια σε G-A (με P και M) και 2) Ομάδα D-E-Q. Τα LCRs βρίσκονται στο C-τελικό άκρο |
| DnaJ τσαπερόνη | 7 | 1) Ομάδα πλούσια σε G και 2) Ομάδα πλούσια σε S. Τα LCRs βρίσκονται στο μέσο της ακολουθίας |
| GroEL τσαπερόνη | 4 | Πλούσια σε G-M (με P και D). Τα LCRs βρίσκονται στο C-τελικό άκρο |

***Πίνακας 34.*** *Αρχαία: Πιο συχνές κατηγορίες τσαπερονών που βρέθηκαν να έχουν LCRs, αριθμός ομολόγων που στοιχήθηκαν και περιγραφή των LCRs της καθεμίας*

| Βακτήρια | | |
|---|---|---|
| **Σχολιασμός πρωτεΐνης** | **Αριθμός ομολόγων** | **Περιγραφή των LCRs** |
| Τσαπερόνη 60kDa | 221 | Πλούσια σε G-M με μερικά A, P. Τα LCR βρίσκονται στο C-τελικό άκρο |
| DnaJ τσαπερόνη | 67 | Πλούσια σε G με F (μερικά άλλα αα που συμμετέχουν λιγότερο: D, R, S, Q, A). Τα LCR στο μέσο της ακολουθίας (πιο πολύ προς το N-τελικό άκρο). Περιοχές πλούσιες σε G: μικρές επαναλήψεις του G (2 ή περισσότερα Gs) που διακόπτονται από 1 αα ή ένα μικρό μοτίβο (~ 2-6 αα) |
| DnaK τσαπερόνη | 15 | 2 μεγάλες ομάδες: 1) Πλούσια σε A (με G ή Ps) 2) Πλούσια σε E-D-V. Από τη στοίχηση οι 2 ομάδες φαίνονται πιο καθαρά: 1) Τα LCRs που γρίσκονται προς το N-τελικό άκρο (πλούσια σε A-P) και 2) τα LCRs που βρίσκονται προς το C-τελικό άκρο: περιοχές πλούσιες σε A-G και D-E-V |

***Πίνακας 35.*** *Βακτήρια: Πιο συχνές κατηγορίες τσαπερονών που βρέθηκαν να έχουν LCRs, αριθμός ομολόγων που στοιχήθηκαν και περιγραφή των LCRs της καθεμίας*

### 3.3.5 Υπόλοιπες μεγάλες κατηγορίες πρωτεϊνών

Οι πρωτεΐνες που βρέθηκαν να περιέχουν LCRs και εμφανίζονται στους περισσότερους οργανισμούς (για τα βακτήρια), συγκεντρώθηκαν και λήφθηκαν από τη Uniprot οι ακολουθίες τους. Στη συνέχεια στοιχήθηκαν με χρήση του Muscle και τα αποτελέσματα οπτικοποιήθηκαν με το Seaview και το Jalview. Στις στοιχίσεις που πραγματοποιήθηκαν, προστέθηκε σαν σημείο αναφοράς και η αντίστοιχη πρωτεϊνική ακολουθία του *E.coli* (αν δεν είχε LCRs και δεν βρισκόταν ήδη μεταξύ των πρωτεϊνών προς στοίχιση). Κάποιες πρωτεΐνες που σχολιάζονταν υποθετικές (putative, hypothetical, κ.α) ή ήταν θραύσματα (fragment) δεν συμπεριλήφθηκαν στις στοιχίσεις. Ο αριθμός των ομολόγων που στοιχήθηκαν για κάθε πρωτεΐνη, μαζί με το χαρακτηρισμό της και μια σύντομη περιγραφή της φύσης των LCRs που περιέχει, φαίνονται στον Πίνακας 36.

Οι κατηγορίες πρωτεϊνών, που έχουν βρεθεί σε πολλούς οργανισμούς να περιέχουν LCRs και ανήκουν σε κάποια από τις κατηγορίες που έχουν ήδη αναφερθεί (πρωτεΐνες που προσδένουν RNA/DNA, ριβοσωμικές, τσαπερόνες, κ.λ.π) δεν συμπεριλήφθηκαν στον Πίνακας 36.

Για τα αρχαία, οι μεγάλες κατηγορίες πρωτεϊνών ήταν οι ριβοσωμικές πρωτεΐνες και οι τσαπερόνες, που προαναφέρθηκαν και δεν κρίθηκε σκόπιμο να αναφερθούν ξανά.

| Βακτήρια | | |
|---|---|---|
| **Σχολιασμός πρωτεΐνης** | **Αριθμός ομολόγων** | **Περιγραφή των LCRs** |
| Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex | 64 | Πλούσια σε A-P. Τα LCRs βρίσκονται στο μέσο της ακολουθίας (περισσότερο προς το N-τελικό άκρο). |
| Dihydrolipoamide /Acetyltransferase component of pyruvate dehydrogenase complex | 287 | Πλούσια σε A-P. Τα Τα LCRs βρίσκονται στο μέσο της ακολουθίας (περισσότερο προς το N-τελικό άκρο). Σε μερικά από αυτά παρεμβάλλεται ένα συντηρημένο μοτίβο στο μέσο του LCR χωρίζοντάς το σε 2 τμήματα |
| Signal recognition particle receptor protein FtsY | 39 | Πλούσια σε A-E-P rich. Τα LCRs βρίσκονταιπρος το N-τελικό άκρο |
| Signal recognition particle protein | 91 | 1) Πλούσια σε G-P-M (με L, Ks) 2) Πλούσια σε K (με μερικά R, G). Τα LCRs βρίσκονται στο C-τελικό άκρο (τα περισσότερα έχουν στο τέλος ένα K ή R ή μια μικρή επανάληψή τους) |
| chitinase | 83 | Οι ακολουθίες αποκλίνουν αρκετά αλλά μερικά μοτίβα υπερισχύουν. Τα G-D και G-S, τα μεγάλα SARs του Sκαι οι επαναλήψεις P-T (με μερικές από αυτές να έχουν As). Τα LCRs είναι διάσπαρτα σε όλο το μήκος της ακολουθίας |
| protein tola | 108 | 1) Μεγάλη ομάδα πλούσια σε A, E, K, και δυο μικρότερες: 2) Πλούσια σε P και 3) Πλούσια σε Q. Τα LCRs βρίσκονται στο μέσο της ακολουθίας (μερικά πιο πολύ προς το N-τελικό άκρο) |
| serine threonine kinase | 128 | 1) Η μεγαλύτερη ομάδα είναι πλούσια σε P (με T, E) και 3 μικρότερες ομάδες 2) πλούσια σε S, 3) πλούσια σε G και 4) πλούσια σε A. Τα LCRs είναι διάσπαρτα σε όλο το μήκος της ακολουθίας |
| tonb protein | 138 | Η μεγαλύτερη ομάδα είναι πλούσια σε P με E, A και K. Μια μικρότερη ομάδα είναι πλούσια σε G-S. Τα LCRs βρίσκονται στο μέσο της ακολουθίας |

***Πίνακας 36.*** *Βακτήρια: Λοιπές πιο συχνές κατηγορίες πρωτεϊνών που βρέθηκαν να έχουν LCRs, αριθμός ομολόγων που στοιχήθηκαν και περιγραφή των LCRs της καθεμίας*

## 3.4 Εργαλείο ανίχνευσης των LCRs και πρόβλεψης της λειτουργίας τους

### 3.4.1 Ομαδοποίηση των LCRs με βάση τους όρους - οντολογίες

Οι πιο εμπλουτισμένοι όροι - οντολογίες των προκαρυωτικών LCRs συγκεντρώθηκαν και ομαδοποιήθηκαν σε 4 μεγάλες κατηγορίες:

1. Τα LCRs που ανήκουν σε **τσαπερόνες**
2. Τα LCRs που ανήκουν σε **πρωτεΐνες που προσδένουν DNA/RNA**
3. Τα LCRs που ανήκουν σε **πρωτεΐνες που προσδένουν μέταλλα**
4. Τα LCRs που ανήκουν στις **υπόλοιπες πρωτεΐνες**

Η κατηγορία 4 περιλαμβάνει όλα τα υπόλοιπα LCRs που χαρακτηρίζονται από τους υπόλοιπους όρους - οντολογίες αλλά δεν ήταν εμπλουτισμένοι σε κάποιο αμινοξύ και δεν είχαν κάποιο χαρακτηριστικό αμινοξικό μοτίβο.

Από την ανάλυση των ευκαρυωτικών LCRs προέκυψε μια επιπλέον κατηγορία που εντάχθηκε στην ανάλυση:

5. Τα LCRs που ανήκουν σε **δομικές πρωτεΐνες** (κολλαγόνα, κερατίνες) και εντάχθηκαν στην κατηγορία με τις υπόλοιπες πρωτεΐνες

Τα πρωτεϊνικά τμήματα LCRs συγκεντρώθηκαν και ομαδοποιήθηκαν, για κάθε κατηγορία ξεχωριστά, με χρήση της Matlab και της συνάρτησης clustergram, όπως έχει προαναφερθεί. Στη συνέχεια, από κάθε ομαδοποίηση λήφθηκαν τα LCRs των μεγάλων κλάδων που κατηγοριοποιούνται μαζί, για κάθε μια από τις παραπάνω κατηγορίες. Με αυτό τον τρόπο επιτεύχθηκε ένας καθαρισμός των αρχικών κατηγοριών, ώστε να απομακρυνθούν LCRs που δεν σχετίζονται με την εκάστοτε λειτουργία.

Κάθε ομάδα έχει το δικό της χαρακτηριστικό αμινοξικό αποτύπωμα. Τα LCRs που ανήκουν σε τσαπερόνες είναι πλούσια σε G (γλυκίνη) και M (μεθειονίνη) και διαθέτουν το μοτίβο GGM, τα LCRs που προσδένουν DNA/RNA χωρίζονται σε δύο μεγάλες κατηγορίες: 1) όσα είναι πλούσια σε G (γλυκίνη), και R (αργινίνη) με κάποια να διαθέτουν τα μοτίβα GGR, GGY, GGF και GGQ, και 2) όσα είναι πλούσια σε K (λυσίνη) και A (αλανίνη), ενώ τα LCRs που ανήκουν στις πρωτεΐνες που προσδένουν μέταλλα διαθέτουν LCRs πλούσια σε H (ιστιδίνη).

Με βάση αυτά τα ευρήματα, τα LCRs σε κάθε μια από τις παραπάνω καθαρές κατηγορίες έλαβαν ένα κοινό λειτουργικό σχολιασμό. Έτσι, κάθε LCR που ανήκει στην κατηγορία των τσαπερονών έλαβε το λειτουργικό σχολιασμό CHAP_GM (από το **CHAP**eron και τα δύο αμινοξέα που είναι κυρίαρχα στην κατηγορία αυτή), τα LCRs που ανήκουν στην κατηγορία των πρωτεϊνών που σχετίζονται με πρόσδεση του DNA και του RNA έλαβαν το λειτουργικό σχολιασμό DRB_GR (από το **D**NA/**R**NA **B**inding, για όσα ανήκαν στην 1η κατηγορία και ήταν πλούσια σε G και R) και

DRB_KA (για όσα ανήκαν στην 2η κατηγορία και ήταν πλούσια σε K και A). Τα LCRs της κατηγορίας των πρωτεϊνών που σχετίζονται με πρόσδεση μετάλλων έλαβαν το λειτουργικό σχολιασμό MI_H (από το **M**etall **I**on και το κυρίαρχο αμινοξύ τους). Τέλος, όλα τα υπόλοιπα LCRs που δεν εμπίπτουν σε καμία από τις παραπάνω κατηγορίες αλλά και αυτά που ανήκουν σε δομικές πρωτεΐνες των ευκαρυωτών, έλαβαν το σχολιασμό Other.

### 3.4.2 Κατασκευή Διγραμμάτων

Σε όλες τις κατηγορίες που προαναφέρθηκαν, για την ανάλυση του κάθε πρωτεϊνικού τμήματος LCR υπολογίστηκε ο συνολικός αριθμός και η συχνότητα των διγραμμάτων του και κατασκευάστηκε ένα διάνυσμα συχνότητας για 400 διαφορετικά διγράμματα, που χρησιμοποιήθηκε, όπως αναλύεται παρακάτω, για τη σύγκριση των LCRs από καλά χαρακτηρισμένες πρωτεΐνες, με τα LCRs που θα ανιχνευθούν από το πρόγραμμα (με χρήση του συντελεστή συσχέτισης Pearson) και για την κατασκευή των νευρωνικών δικτύων .

### 3.4.3 Κατασκευή Νευρωνικών δικτύων

Οι 4 μεγάλες κατηγορίες χρησιμοποιήθηκαν για τη δημιουργία και εκπαίδευση νευρωνικών δικτύων με χρήση του TensorFlow και του Keras αλλά και με χρήση της Matlab. Κατασκευάστηκαν δύο νευρωνικά δίκτυα, ένα βασισμένο στις συχνότητες των αμινοξέων σε κάθε τμήμα LCR και ένα που βασίστηκε στη συχνότητα των διγραμμάτων του. Το 70% των LCRs από την κάθε κατηγορία χρησιμοποιήθηκε για την εκπαίδευση του νευρωνικού δικτύου, ενώ το υπόλοιπο 30% των LCRs χρησιμοποιήθηκε για την εκτίμηση της λειτουργίας του και τον υπολογισμό της ακρίβειας των προβλέψεών του.

Υπολογίστηκε επίσης η μήτρα σύγχυσης (Confusion Matrix) τόσο για τα νευρωνικά δίκτυα που κατασκευάστηκαν με το Tensorflow - Keras όσο και για αυτά που κατασκευάστηκαν με τη Matlab.

Η ακρίβεια (accuracy) των νευρωνικών δικτύων με το Tensorflow - Keras ήταν: για αυτό που κατασκευάστηκε με χρήση της συχνότητας των διγραμμάτων: 0.924 και για αυτό που κατασκευάστηκε με χρήση της συχνότητας των αμινοξέων: 0.921.

Για τη Matlab: αυτό που κατασκευάστηκε με χρήση των διγραμμάτων: 0.925 και για αυτό που κατασκευάστηκε με χρήση της συχνότητας των αμινοξέων: 0.921.

Στους Πίνακας 37, Πίνακας 38, Πίνακας 39 και Πίνακας 40 φαίνονται οι μήτρες σύγχυσης που υπολογίστηκαν από το σύνολο των δεδομένων που χρησιμοποιήθηκαν για την εκτίμηση της λειτουργίας του νευρωνικού δικτύου (evaluation dataset), για τα νευρωνικά δίκτυα που κατασκευάστηκαν με το Tensorflow - Keras και με τη Matlab. Τα μπλε κελιά σε κάθε κατηγορία δείχνουν τον αριθμό των LCRs που προβλέφθηκαν σωστά (αληθώς θετικά, True Positive). Τα ποσοστά κάτω από κάθε κατηγορία αποτελούν την ευαισθησία (sensitivity) για την κάθε κατηγορία, ενώ τα ποσοστά

οριζόντια δείχουν την ευστοχία (precision) ή αλλιώς θετική προγνωστική αξία (**P**ositive **P**redictive **V**alue, **PPV**) για την κάθε κατηγορία.

| | **Chaperones** | **DNA / RNA binding** | **Metal ion binding** | **Other** | |
|---|---|---|---|---|---|
| **Chaperones** | 27 | 1 | 0 | 3 | **87.1%** |
| **DNA / RNA binding** | 0 | 127 | 0 | 45 | **73.84%** |
| **Metal ion binding** | 0 | 0 | 14 | 3 | **82.35%** |
| **Other** | 1 | 54 | 0 | 1085 | **95.18%** |
| | **96.43%** | **69.78%** | **100%** | **95.51%** | |

***Πίνακας 37.*** *Μήτρα σύγχυσης για το νευρωνικό δίκτυο που κατασκευάστηκε με τη Matlab και βασίστηκε στη συχνότητα των αμινοξέων στα LCRs*

| | **Chaperones** | **DNA / RNA binding** | **Metal ion binding** | **Other** | |
|---|---|---|---|---|---|
| **Chaperones** | 38 | 0 | 0 | 2 | **95%** |
| **DNA / RNA binding** | 0 | 105 | 0 | 43 | **70.95%** |
| **Metal ion binding** | 0 | 0 | 19 | 4 | **82.61%** |
| **Other** | 1 | 52 | 0 | 1096 | **95.39%** |
| | **97.44%** | **66.88%** | **100%** | **95.72%** | |

***Πίνακας 38.*** *Μήτρα σύγχυσης για το νευρωνικό που κατασκευάστηκε με τη Matlab και βασίστηκε στη συχνότητα των διγραμμάτων στα LCRs*

| | **Chaperones** | **DNA / RNA binding** | **Metal ion binding** | **Other** | |
|---|---|---|---|---|---|
| **Chaperones** | 34 | 1 | 0 | 3 | **89.47%** |
| **DNA / RNA binding** | 0 | 119 | 0 | 41 | **74.38%** |
| **Metal ion binding** | 0 | 0 | 19 | 2 | **90.48%** |
| **Other** | 0 | 58 | 2 | 1080 | **94.74%** |
| | **100%** | **66.85%** | **90.48%** | **95.91%** | |

***Πίνακας 39.*** *Μήτρα σύγχυσης για το νευρωνικό δίκτυο που κατασκευάστηκε με το Keras και το Tensorflow και βασίστηκε στη συχνότητα των αμινοξέων στα LCRs*

| | Chaperones | DNA / RNA binding | Metal ion binding | Other | |
|---|---|---|---|---|---|
| **Chaperones** | 34 | 1 | 0 | 4 | **87.18%** |
| **DNA / RNA binding** | 0 | 120 | 0 | 39 | **75.47%** |
| **Metal ion binding** | 0 | 0 | 20 | 1 | **95.24%** |
| **Other** | 0 | 57 | 1 | 1082 | **94.91%** |
| | **100%** | **67.42%** | **95.24%** | **96.09%** | |

***Πίνακας 40.*** *Μήτρα σύγχυσης για το νευρωνικό δίκτυο που κατασκευάστηκε με το Keras και το Tensorflow και βασίστηκε στη συχνότητα των διγραμμάτων στα LCRs*

### 3.4.4 Συντελεστής Συσχέτισης Pearson

Για την πρόβλεψη της λειτουργίας των LCRs χρησιμοποιήθηκε, επιπλέον, ο συντελεστής συσχέτισης Pearson (Pearson correlation coefficient). Οι 4 κατηγορίες που προαναφέρθηκαν χρησιμοποιούνται για σύγκριση με ένα νέο LCR που θα ανιχνευθεί στην αλληλουχία προς ανάλυση. Υψηλή τιμή του συντελεστή συσχέτισης υποδηλώνει μεγάλη ομοιότητα στο αμινοξικό περιεχόμενο των δύο συγκρινόμενων πρωτεϊνικών τμημάτων (μέγιστη τιμή: 1, ελάχιστη τιμή: -1). Η ομοιότητα στο αμινοξικό περιεχόμενο σχετίζεται με πιθανή ομοιότητα στη λειτουργία (Aravind and Koonin, 1999; Gough et al., 2001; Morais et al., 2011). Έτσι, με σύγκριση των LCRs με γνωστή λειτουργία και του προς ανάλυση LCR επιτυγχάνεται η πρόβλεψη της λειτουργίας, στην οποία πιθανότατα εμπλέκεται το πρωτεϊνικό τμήμα χαμηλής πολυπλοκότητας.

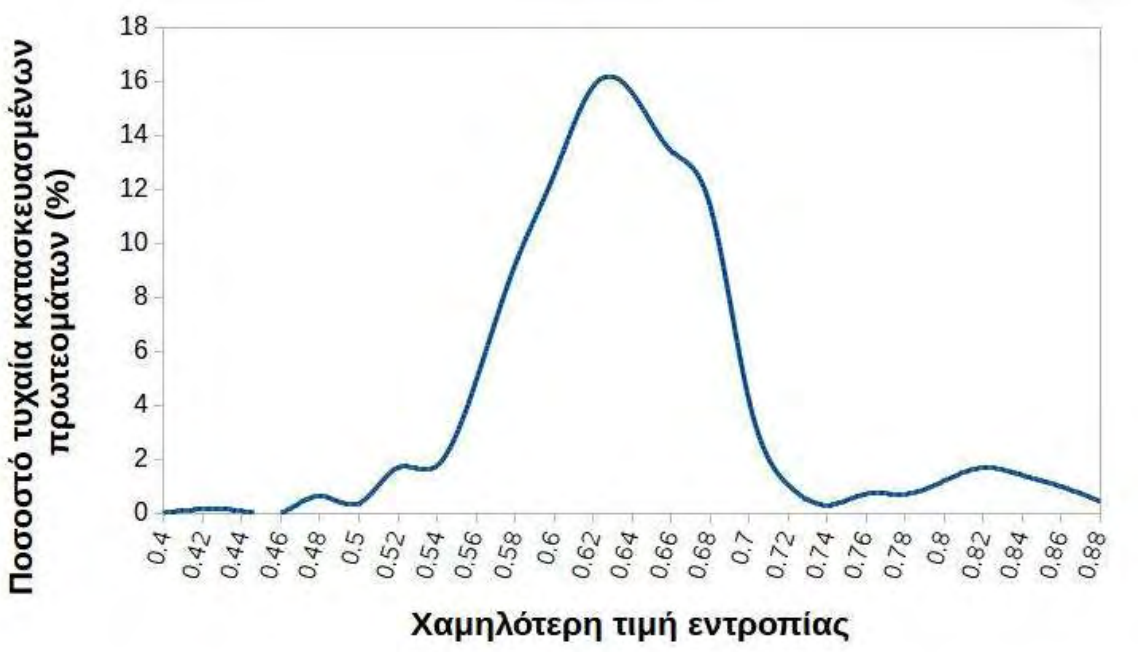
Για τον υπολογισμό του συντελεστή συσχέτισης χρησιμοποιούνται τόσο οι συχνότητες των αμινοξέων στα δύο, προς σύγκριση, LCRs όσο και οι συχνότητες των διγραμμάτων τους.

### 3.4.5 Κατασκευή του εργαλείου

Κατασκευάστηκε perl script για την ανίχνευση των LCRs σε ένα δοθέν πρωτεϊνικό τμήμα, το οποίο πρέπει να δοθεί σε fasta format. Επίσης ο χρήστης μπορεί να ρυθμίσει το μήκος του συρόμενου παραθύρου και το κατώφλι εντροπίας για την επιλογή των LCRs, δηλαδή να καθορίσει την κατώτερη τιμή εντροπίας για επιλογή των τμημάτων. Έτσι, αν το μήκος του παραθύρου οριστεί στο 20, η ακολουθία θα σκανάρεται σε διαδοχικά αλληλοεπικαλυπτόμενα τμήματα με μήκος 20 αμινοξέων και θα υπολογίζεται η τιμή της εντροπίας τους. Η μετατόπιση του παραθύρου είναι καθορισμένη στο μισό του μήκους που θα οριστεί. Επίσης, αν η τιμή του κατωφλιού οριστεί στο 0.6, θα εμφανιστούν μόνο τα τμήματα με τιμή εντροπίας από 0.6 και κάτω (π.χ. ένα τμήμα με τιμή εντροπίας 0.65 δεν θα εμφανιστεί).

Η προεπιλεγμένη ρύθμιση μήκους παραθύρου είναι το 30, καθώς αυτό το μήκος παραθύρου χρησιμοποιήθηκε στις αναλύσεις μας. Συγκεντρώθηκαν όλες οι χαμηλότερες τιμές εντροπίας, που υπολογίστηκαν κατά την τυχαία ανακατασκευή των πρωτεϊνών (Monte Carlo), από όλα τα πρωτεώματα των αρχαίων και των

βακτηρίων που αναλύθηκαν, και κατασκευάστηκε η καμπύλη κατανομής. Με βάση την (κανονική) κατανομή, όπως φαίνεται και στην Εικόνα 43, στα περισσότερα τυχαία κατασκευασμένα πρωτεώματα οι χαμηλότερες τιμές εντροπίας, που υπολογίστηκαν, κυμαίνονται από 0.6 έως 0.68. Έτσι, ως προεπιλεγμένη ρύθμιση κατωφλιού εντροπίας ορίστηκε η τιμή 0.65.




***Εικόνα 43.*** *Κατανομή των χαμηλότερων τιμών εντροπίας που υπολογίστηκαν στα τυχαία πρωτεώματα*

Επίσης, για να διευκολυνθεί και να βελτιστοποιηθεί η ανάλυση πρωτεωμάτων με τη χρήση του εργαλείου, υπάρχει η επιλογή να υπολογιστεί το κατώφλι εντροπίας του δοθέντος πρωτεώματος με τη μέθοδο Monte Carlo (ανακατασκευή των πρωτεϊνών με βάση τη συχνότητα των αμινοξέων στο πρωτέωμα), με τον τρόπο που έχει προαναφερθεί.

Αφού ανιχνευθούν τα LCRs με τιμή ίση ή μικρότερη από αυτή του κατωφλιού που έχει οριστεί, τα αλληλοεπικαλυπτόμενα τμήματα - αν υπάρχουν - συγχωνεύονται και υπολογίζονται οι συχνότητες των αμινοξέων και των διγραμμάτων του κάθε LCR. Αυτές οι συχνότητες συγκρίνονται, με το συντελεστή συσχέτισης Pearson, με τις συχνότητες των αμινοξέων και των διγραμμάτων, που έχουν υπολογιστεί για τα LCRs από πρωτεΐνες με γνωστή λειτουργία, που ανήκουν στις 4 κατηγορίες και εμφανίζεται το LCR (από τις 4 κατηγορίες), με την υψηλότερη τιμή συντελεστή συσχέτισης (ξεχωριστά για τα διγράμματα και για το αμινοξικό περιεχόμενο) μαζί με το όνομα της πρωτεΐνης στην οποία ανήκει, το λειτουργικό σχολιασμό που του έχει δοθεί και την τιμή του συντελεστή συσχέτισης. Επίσης, υπολογίζεται και εμφανίζεται η συχνότητα των δύο κυρίαρχων αμινοξέων στο κάθε LCR που ανιχνεύθηκε, καθώς και η θέση του μέσα στην πρωτεΐνη ή στο πρωτεϊνικό τμήμα που δόθηκε προς ανάλυση. Τέλος, εμφανίζεται και η κατηγοριοποίησή τους με βάση το νευρωνικό δίκτυο, που κατασκευάστηκε με βάση τη συχνότητα των διγραμμάτων στα LCRs. Δίνεται η πιθανότητα με την οποία το νευρωνικό δίκτυο έχει κατατάξει το LCR σε

κάθε κατηγορία (για μια πιο πλήρη εικόνα αλλά και για να είναι δυνατή η σύγκριση). Η κατηγορία με τη μεγαλύτερη πιθανότητα αποτελεί και την πρόβλεψη του νευρωνικού δικτύου.

Τέλος, κατασκευάστηκε, από τον Παναγιώτη Βλασταρίδη, ο web server, που βασίζεται στο εργαλείο και είναι διαθέσιμος στη διεύθυνση: http://bioinf.bio.uth.gr/lcr/#/.

Στην Εικόνα 44 παρουσιάζονται τα αποτελέσματα, όπως δίνονται από το web server, μετά από ανάλυση της δοκιμαστικής αλληλουχίας (test sequence). Η συγκεκριμένη αλληλουχία περιέχει 1 LCR πλούσιο σε λυσίνη (K).

## LCR Fragments for Protein SEQ2


## Tables for LCR Fragments Found on Protein SEQ2

| A/A | Fragment Id | Fragment Sequence | Start | End | Entropy | Dominant AA (%) | 2nd Dominant AA (%) |
|---|---|---|---|---|---|---|---|
| 1 | SEQ2__1 | KKKKKKAKAKAKAAAAAKKKKKKKKKKKAAAKKKKKKKAAAAAAA | 1 | 46 | 0.29 | K:60 | A:40 |

### Similarity with functionally determined LCRs based on aminoacid content (Pearson correlation coefficient)

| Closest LCR | Protein annotation of closest LCR relative | LCR functional category | LCR fragment of closest relative | Pearson correlation coefficient |
|---|---|---|---|---|
| ENA\|ABI67137\|ABI67137.1_1 | Transcriptional regulator, MarR family | DRB_KA | KKKAKKKAKEKAKAKAKKAARKSKKVSADA | 0.99 |

### Similarity with functionally determined LCRs based on Bigrams (Pearson correlation coefficient)

| Closest LCR | Protein annotation of closest LCR relative | LCR functional category | LCR fragment of closest relative | Pearson correlation coefficient |
|---|---|---|---|---|
| ENA\|KPN62101\|KPN62101.1_1 | DNA topoisomerase 1 | DRB_KA | KKKAVAKKAAPKKKAAAKK | 0.95 |

### Neural Network Classification of the LCR found is **DNA/RNA binding**

DNA/RNA binding Score: 8.72E-1 | Chaperone Score: 1.02E-7 | Nickel-Cobalt Binding Score: 1.86E-8 | Other Category Score: 1.28E-1

***Εικόνα 44.*** *Αποτελέσματα από το web server*

### 3.4.6 Εκτίμηση της λειτουργίας του εργαλείου

Τα ευκαρυωτικά πρωτεώματα αναλύθηκαν ξανά με το εργαλείο, για ανίχνευση LCRs και πρόβλεψη της λειτουργίας τους τόσο με το συντελεστή συσχέτισης (με βάση τα διγράμματα και το αμινοξικό περιεχόμενο) όσο και με τα νευρωνικά δίκτυα. Στα κομμάτια LCRs που ανιχνεύτηκαν για κάθε ευκαρυωτικό οργανισμό, οι λειτουργικές προβλέψεις που είχαν δοθεί από το εργαλείο συγκρίθηκαν χειροκίνητα με το σχολιασμό της πρωτεΐνης και τους όρους - οντολογίες του κάθε LCR για να εκτιμηθεί η ακρίβεια των προβλέψεων, όσον αφορά τις πρωτεΐνες που προσδένουν DNA και RNA, δηλαδή την κατηγορία DRB. Για το σκοπό αυτό χρησιμοποιήθηκε η θετική προγνωστική αξία και ελέγχθηκε η αξιοπιστία των αποτελεσμάτων, που λαμβάνονται από τις προβλέψεις του νευρωνικού δικτύου (που κατασκευάστηκε με βάση τη συχνότητα των διγραμμάτων).

Στην *D. melanogaster*, συνολικά 34 τμήματα προβλέφθηκαν ότι προσδένουν DNA/RNA. Τα 6 τμήματα ήταν ψευδώς θετικά (προβλέφθηκαν ότι προσδένουν δηλαδή DNA/RNA, ενώ οι σχολιασμοί των πρωτεϊνών και οι όροι - οντολογίες τους δεν σχετίζονται με πρόσδεση DNA/RNA) και 28 τμήματα αληθώς θετικά (ότι προσδένουν δηλαδή DNA/RNA και οι σχολιασμοί των πρωτεϊνών και οι όροι - οντολογίες τους σχετίζονται με πρόσδεση DNA/RNA). Η θετική τιμή πρόβλεψης που υπολογίστηκε ήταν 0.82

Στον *S. cerevisiae* συνολικά 7 τμήματα προβλέφθηκαν ότι προσδένουν DNA/RNA και ήταν όλα αληθώς θετικά. Έτσι, η θετική τιμή πρόβλεψης που υπολογίστηκε ήταν 1.

Στην *A. thaliana* προβλέφθηκαν ότι προσδένουν DNA/RNA συνολικά 55 τμήματα, από τα οποία τα 17 ήταν ψευδώς θετικά και τα 38 αληθώς θετικά. Η θετική τιμή πρόβλεψης που υπολογίστηκε ήταν 0.69.

Τέλος, στον άνθρωπο συνολικά 41 τμήματα προβλέφθηκαν ότι προσδένουν DNA/RNA και από αυτά το 1 ήταν ψευδώς θετικό και τα 40 αληθώς θετικά. Η θετική τιμή πρόβλεψης που υπολογίστηκε ήταν 0.97.

Επιπλέον, τα πρωτεϊνικά τμήματα ενδογενούς δομικής αστάθειας, που δίνονται στις εργασίες των Castello (Castello et al., 2016) και Järvelin (Järvelin et al., 2016) και δείχθηκε ότι προσδένουν RNA, λήφθηκαν και αναλύθηκαν για να εκτιμηθεί η ακρίβεια του εργαλείου όσον αφορά την πρόβλεψη των LCRs που προσδένουν DNA και RNA.

Αρχικά, τα τμήματα LCRs που παρέχονται στην εργασία του Järvelin αναλύθηκαν με το εργαλείο. Παρείχε 45 πρωτεΐνες με 57 τμήματα με ενδογενώς ασταθείς περιοχές που προσδένουν RNA. Τα 44 από αυτά είχαν μήκος πάνω από 15 αμινοξέα και τα 20 είχαν τμήματα που ανιχνεύθηκαν ως περιοχές χαμηλής πολυπλοκότητας από το εργαλείο (με βάση τα κριτήρια που τέθηκαν για το κατώφλι εντροπίας). Από αυτά, 7 προβλέφθηκαν σωστά (ότι προσδένουν DNA/RNA) με το συντελεστής συσχέτισης που βασίστηκε στη συχνότητα των αμινοξέων, 10 προβλέφθηκαν σωστά με το συντελεστής συσχέτισης που βασίστηκε στη συχνότητα των διγραμμάτων, 8 προβλέφθηκαν σωστά με τα νευρωνικά δίκτυα που βασίστηκαν στη συχνότητα των

αμινοξέων και 10 με τα νευρωνικά δίκτυα που βασίστηκαν στη συχνότητα των διγραμμάτων.

# 4. Συμπεράσματα

Όπως φαίνεται τα LCRs πιθανότατα προέρχονται από εξελικτικά αρχέγονες περιοχές. Αυτό αποδεικνύεται τόσο από το αμινοξικό τους περιεχόμενο, που είναι πολύ παρόμοιο με τα πρώιμα ολιγοπεπτίδια, όσο και από τα κωδικόνια που κωδικοποιούν για τα LCRs. Τα αμινοξέα που αποτελούν αυτές τις περιοχές κατά κύριο λόγο, είναι τα αμινοξέα που πιθανολογείται ότι εμφανίστηκαν πρώτα στο γενετικό κώδικα. Όσον αφορά τα κωδικόνια, αυτά που πιθανολογείται ότι εμφανίστηκαν πρώτα είναι τα GGC (για τη γλυκίνη) και GCC (για την αλανίνη). Προς επιβεβαίωση των παραπάνω παρατηρήσεων, στα LCRs το πιο συχνό κωδικόνιο για τη γλυκίνη είναι το GGC (στα βακτήρια και στα αρχαία) και για την αλανίνη το GCC (στα βακτήρια).

Τα LCRs μπορεί να έχουν την τάση να επεκτείνονται ή να συρρικνώνονται με γρήγορους ρυθμούς και με αυτό τον τρόπο να συνεισφέρουν στην εξέλιξη, παρέχοντας καινούριες ιδιότητες στις πρωτεΐνες, επιτρέποντας τη μη–ειδική αλληλεπίδρασή τους με διάφορα άλλα μόρια, όπως το DNA και το RNA, άλλες πρωτεΐνες, μέταλλα, κ.α. Πολλές από τις περιοχές χαμηλής πολυπλοκότητας είναι συντηρημένες στις ορθόλογες πρωτεΐνες των διάφορων ειδών βακτηρίων και μάλιστα αρκετές εντοπίζονται και σε υψηλά εκφραζόμενες πρωτεΐνες, όπως οι πρωτεΐνες του ριβοσώματος, αυξάνοντας έτσι το ενεργειακό κόστος της μετάφρασης. Αυτό υποδεικνύει μια εξελικτική πίεση να διατηρηθούν και υποδηλώνει ότι πιθανότατα έχουν κάποιο ρόλο στην αντίστοιχη πρωτεΐνη, είτε δομικό είτε λειτουργικό.

Μια άλλη παρατήρηση είναι ότι τα LCRs έχουν συγκεκριμένο αμινοξικό περιεχόμενο. Ανάλογα με αυτό το αμινοξικό περιεχόμενο εμπλέκονται και επιτελούν συγκεκριμένες λειτουργίες στις πρωτεΐνες. Φαίνεται, ότι όσα LCRs είναι πλούσια σε K, R, και σε επαναλαμβανόμενα μοτίβα GGR, GGX (όπου X, αρωματικό αμινοξύ), εμπλέκονται σε πρόσδεση μορίων DNA και RNA. Οι επαναλήψεις του μοτίβου GGM είναι χαρακτηριστικό κάποιων κατηγοριών τσαπερονών και εμπλέκονται στην αναδίπλωση πρωτεϊνών. Τέλος, LCRs πλούσια σε H, D, (και λιγότερο E) φαίνεται να εμπλέκονται σε πρόσδεση ιόντων μετάλλων.

Με βάση αυτά τα ευρήματα, είναι δυνατή η ανίχνευση τέτοιων περιοχών σε πρωτεΐνες και η πρόβλεψη της λειτουργίας τους με βάση το αμινοξικό τους περιεχόμενο. Έτσι παρέχεται η δυνατότητα χαρακτηρισμού πρωτεϊνικών τμημάτων χαμηλής πολυπλοκότητας με άγνωστη λειτουργία και δίνονται πληροφορίες σχετικά με τη λειτουργία αυτών των περιοχών, που ως τώρα θεωρούνται κατά γενική ομολογία περιττές και ότι πολλές από αυτές δεν έχουν κάποιο συγκεκριμένο ρόλο.

# ΒΙΒΛΙΟΓΡΑΦΙΑ


Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anderson, E.C., Hunt, S.L., Jackson, R.J., 2007. Internal initiation of translation from the human rhinovirus-2 internal ribosome entry site requires the binding of Unr to two distinct sites on the 5' untranslated region. Journal of General Virology 88, 3043–3052. https://doi.org/10.1099/vir.0.82463-0

Artificial Neural Networks as Models of Neural Information Processing | Frontiers Research Topic [WWW Document], n.d. URL https://www.frontiersin.org/research-topics/4817/artificial-neural-networks-as-models-of-neural-information-processing#authors (accessed 6.14.18).

Dyson, H.J., Wright, P.E., 2005. Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. 6, 197–208. https://doi.org/10.1038/nrm1589

Huntley, M.A., Golding, G.B., 2002. Simple sequences are rare in the Protein Data Bank. Proteins 48, 134–140. https://doi.org/10.1002/prot.10150

Kumari, B., Kumar, R., Kumar, M., 2015. Low complexity and disordered regions of proteins have different structural and amino acid preferences. Mol. Biosyst. 11, 585–594. https://doi.org/10.1039/C4MB00425F

Wootton, J.C., 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput. Chem. 18, 269–285.

Adda, C.G., MacRaild, C.A., Reiling, L., Wycherley, K., Boyle, M.J., Kienzle, V., Masendycz, P., Foley, M., Beeson, J.G., Norton, R.S., Anders, R.F., 2012. Antigenic characterization of an intrinsically unstructured protein, Plasmodium falciparum merozoite surface protein 2. Infect. Immun. 80, 4177–4185. https://doi.org/10.1128/IAI.00665-12

Akashi, H., Gojobori, T., 2002. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc. Natl. Acad. Sci. U. S. A. 99, 3695–3700. https://doi.org/10.1073/pnas.062526999

Albrecht, A.N., Kornak, U., Böddrich, A., Süring, K., Robinson, P.N., Stiege, A.C., Lurz, R., Stricker, S., Wanker, E.E., Mundlos, S., 2004. A molecular pathogenesis for transcription factor associated poly-alanine tract expansions. Hum. Mol. Genet. 13, 2351–2359. https://doi.org/10.1093/hmg/ddh277

Amiel, J., Trochet, D., Clément-Ziza, M., Munnich, A., Lyonnet, S., 2004. Polyalanine expansions in human. Hum. Mol. Genet. 13, R235–R243. https://doi.org/10.1093/hmg/ddh251

Aravind, L., Koonin, E.V., 1999. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches11Edited by J. M. Thornton. J. Mol. Biol. 287, 1023–1040. https://doi.org/10.1006/jmbi.1999.2653

Barton, M.D., Delneri, D., Oliver, S.G., Rattray, M., Bergman, C.M., 2010. Evolutionary systems biology of amino acid biosynthetic cost in yeast. PloS One 5, e11935. https://doi.org/10.1371/journal.pone.0011935

Bornhorst, J.A., Falke, J.J., 2000. [16] Purification of Proteins Using Polyhistidine Affinity Tags. Methods Enzymol. 326, 245–254.

Brewer, S., Tolley, M., Trayer, I.P., Barr, G.C., Dorman, C.J., Hannavy, K., Higgins, C.F., Evans, J.S., Levine, B.A., Wormald, M.R., 1990. Structure and function of X-Pro dipeptide repeats in the TonB proteins of Salmonella typhimurium and Escherichia coli. J. Mol. Biol. 216, 883–895. https://doi.org/10.1016/S0022-2836(99)80008-4

Brodersen, D.E., Clemons, W.M., Carter, A.P., Wimberly, B.T., Ramakrishnan, V., 2002. Crystal structure of the 30 S ribosomal subunit from Thermus thermophilus: structure

of the proteins and their interactions with 16 S RNA. J. Mol. Biol. 316, 725–768. https://doi.org/10.1006/jmbi.2001.5359
Brodersen Ditlev E., Nissen Poul, 2005. The social life of ribosomal proteins. FEBS J. 272, 2098–2108. https://doi.org/10.1111/j.1742-4658.2005.04651.x
Brown, L.Y., Brown, S.A., 2004. Alanine tracts: the expanding story of human illness and trinucleotide repeats. Trends Genet. 20, 51–58. https://doi.org/10.1016/j.tig.2003.11.002
Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., Krijgsveld, J., Hentze, M.W., 2012. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell 149, 1393–1406. https://doi.org/10.1016/j.cell.2012.04.031
Castello, A., Fischer, B., Frese, C.K., Horos, R., Alleaume, A.-M., Foehr, S., Curk, T., Krijgsveld, J., Hentze, M.W., 2016. Comprehensive Identification of RNA-Binding Domains in Human Cells. Mol. Cell 63, 696–710. https://doi.org/10.1016/j.molcel.2016.06.029
Ciriello, G., Gallina, C., Guerra, C., 2010. Analysis of interactions between ribosomal proteins and RNA structural motifs. BMC Bioinformatics 11, S41. https://doi.org/10.1186/1471-2105-11-S1-S41
Coletta, A., Pinney, J.W., Solís, D.Y.W., Marsh, J., Pettifer, S.R., Attwood, T.K., 2010. Low-complexity regions within protein sequences have position-dependent roles. BMC Syst. Biol. 4. https://doi.org/10.1186/1752-0509-4-43
Corley, S.M., Gready, J.E., 2008. Identification of the RGG box motif in Shadoo: RNA-binding and signaling roles? Bioinforma. Biol. Insights 2, 383–400.
DePristo, M.A., Zilversmit, M.M., Hartl, D.L., 2006. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. Gene 378, 19–30. https://doi.org/10.1016/j.gene.2006.03.023
Djian, P., 1998. Evolution of Simple Repeats in DNA and Their Relation to Human Disease. Cell 94, 155–160. https://doi.org/10.1016/S0092-8674(00)81415-4
Dokmanić, I., Sikić, M., Tomić, S., 2008. Metals in proteins: correlation between the metal-ion type, coordination number and the amino-acid residues involved in the coordination. Acta Crystallogr. D Biol. Crystallogr. 64, 257–263. https://doi.org/10.1107/S090744490706595X
Dorsman, J.C., Pepers, B., Langenberg, D., Kerkdijk, H., Ijszenga, M., den Dunnen, J.T., Roos, R. a. C., van Ommen, G.-J.B., 2002. Strong aggregation and increased toxicity of polyleucine over polyglutamine stretches in mammalian cells. Hum. Mol. Genet. 11, 1487–1496.
Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C., Obradovic, Z., 2001. Intrinsically disordered protein. J. Mol. Graph. Model. 19, 26–59. https://doi.org/10.1016/S1093-3263(00)00138-8
Dunker, A.K., Silman, I., Uversky, V.N., Sussman, J.L., 2008. Function and structure of inherently disordered proteins. Curr. Opin. Struct. Biol., Catalysis and regulation / Proteins 18, 756–764. https://doi.org/10.1016/j.sbi.2008.10.002
Dyson, H.J., Satterthwait, A.C., Lerner, R.A., Wright, P.E., 1990. Conformational preferences of synthetic peptides derived from the immunodominant site of the circumsporozoite protein of Plasmodium falciparum by 1H NMR. Biochemistry (Mosc.) 29, 7828–7837.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340
Faux, N.G., Bottomley, S.P., Lesk, A.M., Irving, J.A., Morrison, J.R., Banda, M.G. de la, Whisstock, J.C., 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res. 15, 537–551. https://doi.org/10.1101/gr.3096505
Feng, Z.-P., Zhang, X., Han, P., Arora, N., Anders, R.F., Norton, R.S., 2006. Abundance of intrinsically unstructured proteins in P. falciparum and other apicomplexan parasite

proteomes. Mol. Biochem. Parasitol. 150, 256–267. https://doi.org/10.1016/j.molbiopara.2006.08.011
Foquet, L., Hermsen, C.C., van Gemert, G.-J., Van Braeckel, E., Weening, K.E., Sauerwein, R., Meuleman, P., Leroux-Roels, G., 2014. Vaccine-induced monoclonal antibodies targeting circumsporozoite protein prevent Plasmodium falciparum infection. J. Clin. Invest. 124, 140–144. https://doi.org/10.1172/JCI70349
Foucault, M., Mayol, K., Receveur-Bréchot, V., Bussat, M.-C., Klinguer-Hamour, C., Verrier, B., Beck, A., Haser, R., Gouet, P., Guillon, C., 2010. UV and X-ray structural studies of a 101-residue long Tat protein from a HIV-1 primary isolate and of its mutated, detoxified, vaccine candidate. Proteins 78, 1441–1456. https://doi.org/10.1002/prot.22661
Frugier, M., Bour, T., Ayach, M., Santos, M. a. S., Rudinger-Thirion, J., Théobald-Dietrich, A., Pizzi, E., 2010. Low complexity regions behave as tRNA sponges to help co-translational folding of plasmodial proteins. FEBS Lett. 584, 448–454.
Garrett, R.A., 1983. Structure and role of eubacterial ribosomal proteins. Horiz. Biochem. Biophys. 7, 101–138.
Gatchel, J.R., Zoghbi, H.Y., 2005. Diseases of Unstable Repeat Expansion: Mechanisms and Common Principles. Nat. Rev. Genet. 6, 743–755. https://doi.org/10.1038/nrg1691
Gene Ontology Consortium, 2015. Gene Ontology Consortium: going forward. Nucleic Acids Res. 43, D1049-1056. https://doi.org/10.1093/nar/gku1179
Gough, J., Karplus, K., Hughey, R., Chothia, C., 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure11Edited by G. Von Heijne. J. Mol. Biol. 313, 903–919. https://doi.org/10.1006/jmbi.2001.5080
Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27, 221–224. https://doi.org/10.1093/molbev/msp259
Haerty, W., Golding, G.B., 2010. Low-complexity sequences and single amino acid repeats: not just “junk” peptide sequences. Genome 53, 753–762. https://doi.org/10.1139/g10-063
Hanakahi, L.A., Sun, H., Maizels, N., 1999. High affinity interactions of nucleolin with G-G-paired rDNA. J. Biol. Chem. 274, 15908–15912.
Haritos, V.S., Niranjane, A., Weisman, S., Trueman, H.E., Sriskantha, A., Sutherland, T.D., 2010. Harnessing disorder: onychophorans use highly unstructured proteins, not silks, for prey capture. Proc. R. Soc. Lond. B Biol. Sci. 277, 3255–3263. https://doi.org/10.1098/rspb.2010.0604
Howard, M.B., Ekborg, N.A., Taylor, L.E., Hutcheson, S.W., Weiner, R.M., 2004. Identification and analysis of polyserine linker domains in prokaryotic proteins with emphasis on the marine bacterium Microbulbifer degradans. Protein Sci. 13, 1422–1425. https://doi.org/10.1110/ps.03511604
Järvelin, A.I., Noerenberg, M., Davis, I., Castello, A., 2016. The new (dis)order in RNA regulation. Cell Commun. Signal. CCS 14. https://doi.org/10.1186/s12964-016-0132-3
Jentzsch, I.M.V., Bagshaw, A.T., Buschiazzo, E., Merkel, A., Gemmell, N.J., 2013. Evolution of Microsatellite DNA, in: ELS. American Cancer Society. https://doi.org/10.1002/9780470015902.a0020847.pub2
Kiledjian, M., Dreyfuss, G., 1992. Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. EMBO J. 11, 2655–2664.
Kiskowski, M.A., Jiang, Y., Alber, M.S., 2004. Role of streams in myxobacteria aggregate formation. Phys. Biol. 1, 173. https://doi.org/10.1088/1478-3967/1/3/005
Klein, D.J., Moore, P.B., Steitz, T.A., 2004. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. J. Mol. Biol. 340, 141–177. https://doi.org/10.1016/j.jmb.2004.03.076

Koebnik, R., Locher, K.P., Van Gelder, P., 2000. Structure and function of bacterial outer membrane proteins: barrels in a nutshell. Mol. Microbiol. 37, 239–253.
Kurzynska-Kokorniak, A., Jamburuthugoda, V.K., Bibillo, A., Eickbush, T.H., 2007. DNA-directed DNA Polymerase and Strand Displacement Activity of the Reverse Transcriptase Encoded by the R2 Retrotransposon. Journal of Molecular Biology 374, 322–333. https://doi.org/10.1016/j.jmb.2007.09.047
Kushwaha, A.K., Grove, A., 2013. C-terminal low-complexity sequence repeats of Mycobacterium smegmatis Ku modulate DNA binding. Biosci. Rep. 33, 175–184. https://doi.org/10.1042/BSR20120105
Lanier, K.A., Roy, P., Schneider, D.M., Williams, L.D., 2017. Ancestral Interactions of Ribosomal RNA and Ribosomal Proteins. Biophys. J. 113, 268–276. https://doi.org/10.1016/j.bpj.2017.04.007
Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V., Cochrane, G., 2011. The European Nucleotide Archive. Nucleic Acids Res. 39, D28-31. https://doi.org/10.1093/nar/gkq967
Ling, J., Cho, C., Guo, L.-T., Aerni, H.R., Rinehart, J., Söll, D., 2012. Protein aggregation caused by aminoglycoside action is prevented by a hydrogen peroxide scavenger. Mol. Cell 48, 713–722. https://doi.org/10.1016/j.molcel.2012.10.001
Lithwick, G., Margalit, H., 2003. Hierarchy of sequence-dependent features associated with prokaryotic translation. Genome Res. 13, 2665–2673. https://doi.org/10.1101/gr.1485203
Luo, H., Nijveen, H., 2014. Understanding and identifying amino acid repeats. Brief. Bioinform. 15, 582–591. https://doi.org/10.1093/bib/bbt003
MacRaild, C.A., Richards, J.S., Anders, R.F., Norton, R.S., 2016. Antibody Recognition of Disordered Antigens. Structure 24, 148–157. https://doi.org/10.1016/j.str.2015.10.028
Marcotte, E.M., Pellegrini, M., Yeates, T.O., Eisenberg, D., 1999. A census of protein repeats. J. Mol. Biol. 293, 151–160. https://doi.org/10.1006/jmbi.1999.3136
Mauro, V.P., Chappell, S.A., Dresios, J., 2007. Chapter Fifteen - Analysis of Ribosomal Shunting During Translation Initiation in Eukaryotic mRNAs, in: Lorsch, J. (Ed.), Methods in Enzymology, Translation Initiation: Extract Systems and Molecular Genetics. Academic Press, pp. 323–354. https://doi.org/10.1016/S0076-6879(07)29015-9
Mirkin, S.M., 2007. Expandable DNA repeats and human disease [WWW Document]. Nature. https://doi.org/10.1038/nature05977
Monte Carlo Methods for Applied Scientists, 2008. World Scientific.
Morais, de L., A, D., Fang, H., Rackham, O.J.L., Wilson, D., Pethica, R., Chothia, C., Gough, J., 2011. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res. 39, D427–D434. https://doi.org/10.1093/nar/gkq1130
Oma, Y., Kino, Y., Sasagawa, N., Ishiura, S., 2005. Comparative analysis of the cytotoxicity of homopolymeric amino acids. Biochim. Biophys. Acta 1748, 174–179. https://doi.org/10.1016/j.bbapap.2004.12.017
Oma, Y., Kino, Y., Sasagawa, N., Ishiura, S., 2004. Intracellular localization of homopolymeric amino acid-containing proteins expressed in mammalian cells. J. Biol. Chem. 279, 21217–21222. https://doi.org/10.1074/jbc.M309887200
Oma Yoko, Kino Yoshihiro, Toriumi Kazuya, Sasagawa Noboru, Ishiura Shoichi, 2009. Interactions between homopolymeric amino acids (HPAAs). Protein Sci. 16, 2195–2204. https://doi.org/10.1110/ps.072955307
Ozdilek, B.A., Thompson, V.F., Ahmed, N.S., White, C.I., Batey, R.T., Schwartz, J.C., 2017. Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. Nucleic Acids Res. 45, 7984–7996. https://doi.org/10.1093/nar/gkx460
Peng, Z., Oldfield, C.J., Xue, B., Mizianty, M.J., Dunker, A.K., Kurgan, L., Uversky, V.N., 2014. A creature with a hundred waggly tails: intrinsically disordered proteins in the

ribosome. Cell. Mol. Life Sci. 71, 1477–1504. https://doi.org/10.1007/s00018-013-1446-6
Perederina, A., Nevskaya, N., Nikonov, O., Nikulin, A., Dumas, P., Yao, M., Tanaka, I., Garber, M., Gongadze, G., Nikonov, S., 2002. Detailed analysis of RNA–protein interactions within the bacterial ribosomal protein L5/5S rRNA complex. RNA 8, 1548–1557. https://doi.org/10.1017/S1355838202029953
Phan, A.T., Kuryavyi, V., Darnell, J.C., Serganov, A., Majumdar, A., Ilin, S., Raslin, T., Polonskaia, A., Chen, C., Clain, D., Darnell, R.B., Patel, D.J., 2011. Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction. Nat. Struct. Mol. Biol. 18, 796–804. https://doi.org/10.1038/nsmb.2064
Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Sander, C., Ouzounis, C.A., 2000. CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. Bioinforma. Oxf. Engl. 16, 915–922.
Raj, D.K., Nixon, C.P., Nixon, C.E., Dvorin, J.D., DiPetrillo, C.G., Pond-Tor, S., Wu, H.-W., Jolly, G., Pischel, L., Lu, A., Michelow, I.C., Cheng, L., Conteh, S., McDonald, E.A., Absalon, S., Holte, S.E., Friedman, J.F., Fried, M., Duffy, P.E., Kurtis, J.D., 2014. Antibodies to PfSEA-1 block parasite egress from RBCs and protect against malaria infection. Science 344, 871–877. https://doi.org/10.1126/science.1254417
Rajyaguru, P., Parker, R., 2012. RGG motif proteins: modulators of mRNA functional states. Cell Cycle Georget. Tex 11, 2594–2599. https://doi.org/10.4161/cc.20716
Rankin, J., Wyttenbach, A., Rubinsztein, D.C., 2000. Intracellular green fluorescent protein-polyalanine aggregates are associated with cell death. Biochem. J. 348, 15–19.
Read, L.R., Raynard, S.J., Rukść, A., Baker, M.D., 2004. Gene repeat expansion and contraction by spontaneous intrachromosomal homologous recombination in mammalian cells. Nucleic Acids Res 32, 1184–1196. https://doi.org/10.1093/nar/gkh280
Reichenbach, H., 2001. Myxobacteria, producers of novel bioactive substances. J. Ind. Microbiol. Biotechnol. 27, 149–156. https://doi.org/10.1038/sj.jim.7000025
Richard, G.-F., Pâques, F., 2000. Mini- and microsatellite expansions: the recombination connection. EMBO Rep 1, 122–126. https://doi.org/10.1093/embo-reports/kvd031
Robison, A.D., Sun, S., Poyton, M.F., Johnson, G.A., Pellois, J.-P., Jungwirth, P., Vazdar, M., Cremer, P.S., 2016. Polyarginine Interacts More Strongly and Cooperatively than Polylysine with Phospholipid Bilayers. J. Phys. Chem. B 120, 9287–9296. https://doi.org/10.1021/acs.jpcb.6b05604
Rose, P.W., Prlić, A., Bi, C., Bluhm, W.F., Christie, C.H., Dutta, S., Green, R.K., Goodsell, D.S., Westbrook, J.D., Woo, J., Young, J., Zardecki, C., Berman, H.M., Bourne, P.E., Burley, S.K., 2015. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res. 43, D345-356. https://doi.org/10.1093/nar/gku1214
Sampath Kumar, A., Tej Sowpati, D., Mishra, R., 2016. Single Amino Acid Repeats in the Proteome World: Structural, Functional, and Evolutionary Insights. PLOS ONE 11, e0166854. https://doi.org/10.1371/journal.pone.0166854
Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x
Siwach, P., Pophaly, S.D., Ganesh, S., 2006. Genomic and evolutionary insights into genes encoding proteins with single amino acid repeats. Mol. Biol. Evol. 23, 1357–1369. https://doi.org/10.1093/molbev/msk022
So, C.R., Fears, K.P., Leary, D.H., Scancella, J.M., Wang, Z., Liu, J.L., Orihuela, B., Rittschof, D., Spillmann, C.M., Wahl, K.J., 2016. Sequence basis of Barnacle Cement Nanostructure is Defined by Proteins with Silk Homology. Sci. Rep. 6, 36219. https://doi.org/10.1038/srep36219
Strzalka, A., Szafran, M.J., Strick, T., Jakimowicz, D., 2017. C-terminal lysine repeats in Streptomyces topoisomerase I stabilize the enzyme-DNA complex and confer high

enzyme processivity. Nucleic Acids Res. 45, 11908–11924. https://doi.org/10.1093/nar/gkx827
Thandapani, P., O'Connor, T.R., Bailey, T.L., Richard, S., 2013. Defining the RGG/RG motif. Mol. Cell 50, 613–623. https://doi.org/10.1016/j.molcel.2013.05.021
Timsit, Y., Acosta, Z., Allemand, F., Chiaruttini, C., Springer, M., 2009. The Role of Disordered Ribosomal Protein Extensions in the Early Steps of Eubacterial 50 S Ribosomal Subunit Assembly. Int. J. Mol. Sci. 10, 817–834. https://doi.org/10.3390/ijms10030817
Trifonov, E.N., 2009. The origin of the genetic code and of the earliest oligopeptides. Res. Microbiol. 160, 481–486. https://doi.org/10.1016/j.resmic.2009.05.004
Trifonov, E.N., 2000. Consensus temporal order of amino acids and evolution of the triplet code. Gene 261, 139–151.
Tyedmers, J., Mogk, A., Bukau, B., 2010. Cellular strategies for controlling protein aggregation. Nat. Rev. Mol. Cell Biol. 11, 777–788. https://doi.org/10.1038/nrm2993
Uchio Naohiro, Oma Yoko, Toriumi Kazuya, Sasagawa Noboru, Tanida Isei, Fujita Eriko, Kouroku Yoriko, Kuroda Reiko, Momoi Takashi, Ishiura Shoichi, 2007. Endoplasmic reticulum stress caused by aggregate- prone proteins containing homopolymeric amino acids. FEBS J. 274, 5619–5627. https://doi.org/10.1111/j.1742-4658.2007.06085.x
UniProt Consortium, 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204-212. https://doi.org/10.1093/nar/gku989
van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., Kim, P.M., Kriwacki, R.W., Oldfield, C.J., Pappu, R.V., Tompa, P., Uversky, V.N., Wright, P.E., Babu, M.M., 2014. Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. 114, 6589–6631. https://doi.org/10.1021/cr400525m
Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J., 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinforma. Oxf. Engl. 25, 1189–1191. https://doi.org/10.1093/bioinformatics/btp033
Weaver, J., Rye, H.S., 2014. The C-terminal tails of the bacterial chaperonin GroEL stimulate protein folding by directly altering the conformation of a substrate protein. J. Biol. Chem. 289, 23219–23232. https://doi.org/10.1074/jbc.M114.577205
Weiner, R.M., Taylor, L.E., Henrissat, B., Hauser, L., Land, M., Coutinho, P.M., Rancurel, C., Saunders, E.H., Longmire, A.G., Zhang, H., Bayer, E.A., Gilbert, H.J., Larimer, F., Zhulin, I.B., Ekborg, N.A., Lamed, R., Richardson, P.M., Borovok, I., Hutcheson, S., 2008. Complete Genome Sequence of the Complex Carbohydrate-Degrading Marine Bacterium, Saccharophagus degradans Strain 2-40T. PLoS Genet. 4. https://doi.org/10.1371/journal.pgen.1000087
Wells, R.D., 1996. Molecular Basis of Genetic Instability of Triplet Repeats. J. Biol. Chem. 271, 2875–2878. https://doi.org/10.1074/jbc.271.6.2875
Wool, I.G., 1996. Extraribosomal functions of ribosomal proteins. Trends Biochem. Sci. 21, 164–165.
Wootton, J.C., Federhen, S., 1996. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 266, 554–571.
Yagi, M., Bang, G., Tougan, T., Palacpac, N.M.Q., Arisue, N., Aoshi, T., Matsumoto, Y., Ishii, K.J., Egwang, T.G., Druilhe, P., Horii, T., 2014. Protective epitopes of the Plasmodium falciparum SERA5 malaria vaccine reside in intrinsically unstructured N-terminal repetitive sequences. PloS One 9, e98460. https://doi.org/10.1371/journal.pone.0098460
Yan, W., Craig, E.A., 1999. The glycine-phenylalanine-rich region determines the specificity of the yeast Hsp40 Sis1. Mol. Cell. Biol. 19, 7751–7758.
Yang, J.C., Madupu, R., Durkin, A.S., Ekborg, N.A., Pedamallu, C.S., Hostetler, J.B., Radune, D., Toms, B.S., Henrissat, B., Coutinho, P.M., Schwarz, S., Field, L., Trindade-Silva, A.E., Soares, C.A.G., Elshahawi, S., Hanora, A., Schmidt, E.W., Haygood, M.G., Posfai, J., Benner, J., Madinger, C., Nove, J., Anton, B., Chaudhary,

K., Foster, J., Holman, A., Kumar, S., Lessard, P.A., Luyten, Y.A., Slatko, B., Wood, N., Wu, B., Teplitski, M., Mougous, J.D., Ward, N., Eisen, J.A., Badger, J.H., Distel, D.L., 2009. The complete genome of Teredinibacter turnerae T7901: an intracellular endosymbiont of marine wood-boring bivalves (shipworms). PloS One 4, e6085. https://doi.org/10.1371/journal.pone.0006085

Zhang, K., Wang, L., Liu, Y., Chan, K.-Y., Pang, X., Schulten, K., Dong, Z., Sun, F., 2013. Flexible interwoven termini determine the thermal stability of thermosomes. Protein Cell 4, 432–444. https://doi.org/10.1007/s13238-013-3026-9

Zhu, Z.Y., Karlin, S., 1996. Clusters of charged residues in protein three-dimensional structures. Proc. Natl. Acad. Sci. U. S. A. 93, 8350–8355.