ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΒΙΟΧΗΜΕΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΧΑΛΙΩΤΗΣ ΑΝΑΡΓΥΡΟΣ

ΛΑΡΙΣΑ 2017

Η πολύπλοκη εξελικτική πορεία των συνθετασών του αμινοάκυλο-tRNA

The complex evolutionary history of aminoacyl-tRNA synthetases

Η πολύπλοκη εξελικτική πορεία των συνθετασών του αμινοάκυλο-tRNA

Χαλιώτης Ανάργυρος

Αμούτζιας Γρηγόριος (Επιβλέπων Καθηγητής)
Εργαστήριο Βιοπληροφορικής, Τμήμα Βιοχημείας και Βιοτεχνολογίας Πανεπιστήμιο Θεσσαλίας

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ

Αμούτζιας Γρηγόριος, Επίκουρος Καθηγητής Βιοπληροφορικής στη Γενωμική, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας
Μόσιαλος Δημήτριος, Επίκουρος Καθηγητής Βιοτεχνολογίας Μικροβίων, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας
Σταθόπουλος Κωνσταντίνος, Καθηγητής Βιοχημείας με έμφαση στη βιοσύνθεση πρωτεϊνών, Ιατρική Σχολή, Πανεπιστήμιο Πατρών

Πρωτίστως θα ήθελα να εκφράσω τις ευχαριστίες μου στον επιβλέπων καθηγητή μου, κ. Αμούτζια Γρηγόριο, Επίκουρο Καθηγητή Βιοπληροφορικής στη Γενωμική, του τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας, για τη βοήθεια του, την εμπιστοσύνη και κυρίως την υπομονή που έδειξε στο πρόσωπο μου κατά τη διάρκεια εκπόνησης της πτυχιακής μου εργασίας. Μου έδειξε ότι πρέπει να δίνουμε σημασία και στην πιο μικρή λεπτομέρεια για να είμαστε σίγουροι για το αποτέλεσμα. Χωρίς τη βοήθεια του, αυτή η πτυχιακή δεν θα ήταν στο επίπεδο που βρίσκεται αυτή τη στιγμή.

Ακόμη θα ήθελα να ευχαριστήσω τον κ. Μόσιαλο Δημήτριο, Επίκουρο Καθηγητή Βιοτεχνολογίας Μικροβίων, του τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας και μέλος της τριμελούς επιτροπής, καθώς με τις γνώσεις του πάνω στη βιοτεχνολογία των μικροβίων και τις πολυκετιδικές συνθάσες (PKS), με καθοδήγησε στην αναζήτηση μικροοργανισμών που είναι πιθανοί στόχοι για την αναζήτηση αντιβιοτικών νέα γενιάς. Επίσης θα ήθελα να ευχαριστήσω και τον κ. Σταθόπουλο Κωνσταντίνο, Καθηγητή Βιοχημείας με έμφαση στη βιοσύνθεση πρωτεϊνών, στη Ιατρική Σχολή του Πανεπιστημίου Πατρών, επίσης μέλος της τριμελούς μου επιτροπής για τη βοήθεια και τις συμβουλές που μου έδωσε, κατά την εκπόνηση της πτυχιακής εργασίας. Η βαθιά γνώση του πάνω στις AARSs με καθοδήγησαν από την αρχή μέχρι το τέλος αυτής της εργασίας.

Επιπλέον θα ήθελα να ευχαριστήσω τον Prof. Hubert D. Becker., Καθηγητή του τμήματος Génétique moléculaire, génomique, microbiologie του Πανεπιστήμιο του Στρασβούργου καθώς επίσης και τον Prof. Michael Ibba, Καθηγητή του τμήματος Μικροβιολογίας του Πανεπιστήμιου του Οχάιο, στις Ηνωμένες Πολιτείες της Αμερικής. Η βοήθεια τους με τις γνώσεις τους πάνω στην βιοχημεία και την εξέλιξη των AARSs ήταν καίριας σημασίας για την ολοκλήρωση της πτυχιακής μου εργασίας.

Θα ήθελα ακόμη να ευχαριστήσω και τον κ. Βλασταρίδη Παναγιώτη, υποψήφιο διδάκτορα του του τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας, για την μεγάλη βοήθεια του στην δημιουργία της βάσης δεδομένων, καθώς επίσης και του διαδικτυακού υπολογιστικού εργαλείου που δημιουργήσαμε.

Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου και ειδικά τους γονείς μου για την συμπαράσταση και τη βοήθεια τους κατά τη διάρκεια εκπόνησης της πτυχιακής μου εργασίας. Χωρίς την βοήθεια τους δεν θα είχα καταφέρει να ολοκληρώσω τις σπουδές μου.

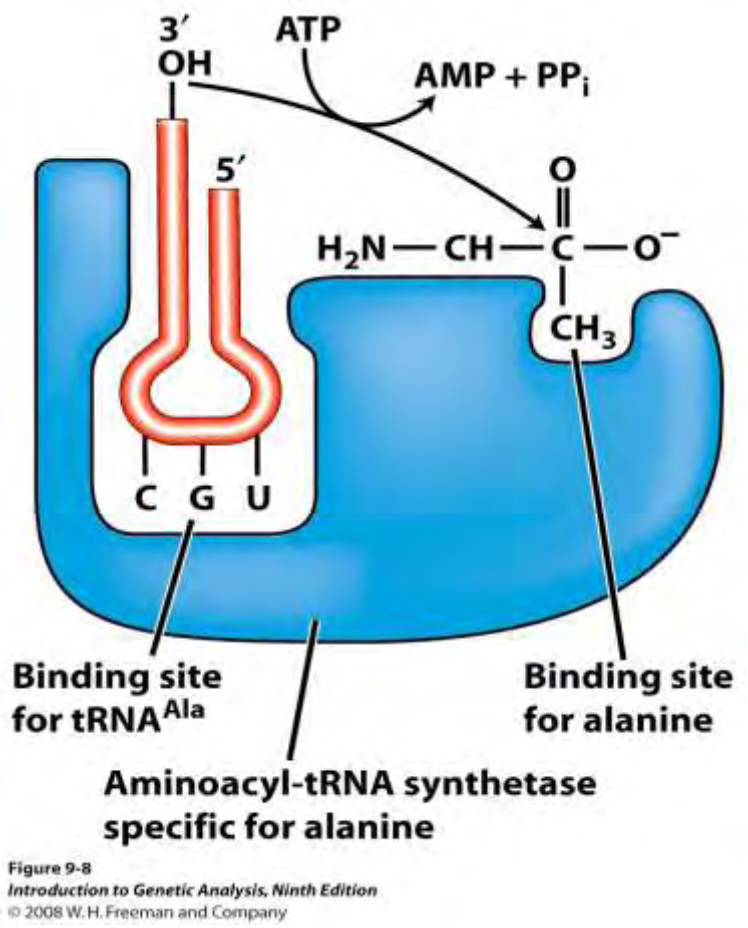
Περιεχόμενα

Περίληψη

Οι συνθετάσες του αμινοάκυλο-tRNA (AARSs) είναι μία υπεροικογένεια ενζύμων, οι οποίες ευθύνονται για την πιστή μετάφραση του γενετικού κώδικα και τα τελευταία χρόνια έχουν γίνει ένας διακεκριμένος στόχος για τους επιστήμονες της συνθετικής βιολογίας. Η μεγάλης κλίμακας ανάλυση μας, σε περισσότερα από 2500 προκαρυωτικά γονιδιώματα, αποκαλύπτει την πολύπλοκη εξελικτική πορεία αυτών των ενζύμων, καθώς και των παραλόγων τους, στην οποία φαίνεται να έπαιξε σημαντικό ρόλο η οριζόντια μεταφορά γονιδίων (HGT). Τα αποτελέσματα δείχνουν ότι μία πολύ διαδεδομένη αντίληψη για την εξελικτική σταθερότητα αυτής της υπεροικογένειας είναι λανθασμένη. Αν και οι AlaRS, GlyRS, LeuRS, lleRS, VaIRS είναι τα πιο σταθερά μέλη της οικογένειας, οι GluRS, LysRS και CysRS έχουν συχνά παράλογα, ενώ οι AsnRS, GlnRS, PylRS και SepRS συχνά λείπουν από πολλά γονιδιώματα. Κατά την διάρκεια αυτής της ανάλυσης, ταυτοποιήθηκαν και χρησιμοποιήθηκαν ιδιαιτέρως καλά συντηρημένα πρωτεϊνικά μοτίβα (motifs) καθώς και πρωτεϊνικές επικράτειες (domains) μέσα στο γονιδιακό τόπο των AARS (loci), με σκοπό να δημιουργήσουμε ένα διαδικτυακό υπολογιστικό εργαλείο για τον εντοπισμό ακολουθιών AARSs σε βακτηριακά πρωτεώματα. Τα μοτίβα και οι επικράτειες που απομονώθηκαν, χρησιμοποιήθηκαν για την δημιουργία στατιστικών μοντέλων. Αυτά είναι βασισμένα στα κρυμμένα Μαρκοβιανά μοντέλα (Hidden Markov Models - HMMs) και είναι διαθέσιμα στο κοινό, μαζί με μία βάση δεδομένων που δημιουργήσαμε, ώστε να είναι δυνατή η χρήση τους σε συγκεκριμένες αναλύσεις. Τα εργαλεία βιοπληροφορικής που έχουμε αναπτύξει, μπορούν επίσης να βοηθήσουν στην ταυτοποίηση πιθανών γονιδίων για την παραγωγή αντιβιοτικών νέας γενιάς, που έχουν ως στόχο τα συγκεκριμένα ένζυμα. Επίσης, αυτά τα εργαλεία θα μπορούσαν να συντελέσουν στην ταυτοποίηση οργανισμών με εναλλακτικά βιοχημικά μονοπάτια που έχουν εξελιχθεί για τη διατήρηση της πιστότητας του γενετικού κώδικα. Η εργασία αυτή δημοσιεύθηκε στο Διεθνές Περιοδικό Nucleic Acids Research το 2017 με τίτλο «The complex evolutionary history of Aminoacyl tRNA synthetases».

1. Εισαγωγή

Οι συνθετάσες του αμινοάκυλο-tRNA (AminoAcyl tRNA Synthetases - AARSs) είναι λειτουργικά ένζυμα που βρίσκονται σε όλους τους ευκαρυωτικούς οργανισμούς, καθώς επίσης στα αρχαία και τα βακτήρια. Είναι το μέσον για την ακριβή εστεροποίηση των αμινοξέων στα αντίστοιχά τους tRNAs (βλ. εικ. 1.1) και για αυτόν τον λόγο, αντιπροσωπεύουν μια σημαντική υπεροικογένεια ενζύμων που είναι υπεύθυνη για την πιστότητα του γενετικού κώδικα.



Εικόνα 1.1 Απεικόνιση της δημιουργίας του αμινοάκυλο-tRNA της Αλανίνης με τη βοήθεια της AlaRS

Εκτός από τις 20 AARSs που είναι υπεύθυνες για την ενσωμάτωση των 20 αμινοξέων, 2 ακόμα AARSs, η PylRS και η SepRS, χρησιμοποιούνται από κάποιους οργανισμούς κατά την διάρκεια της ενσωμάτωσης των σπάνιων αμινοξέων, πυρολυσίνη και φωσφοσερίνη αντίστοιχα.
Οι AARSs χωρίζονται σε 2 μη-ομόλογες κλάσεις: την Κλάση 1 και Κλάση 2 (βλ. πίνακα 1.1 και 1.2), διαχωρισμός που οφείλεται κυρίως σε ξεχωριστές δομές των καταλυτικών domains τους, (Eriani et al. 1990; Ribas de Pouplana and Schimmel 2001).

AARSs κλάσης I		
Ia	Ib	Ic
Leu α	Tyr α_2	Arg α
Ile α	Trp α_2	Gln α
Vla α		Glu α
Cys α_2		
Met α_2		

Πίνακας 1.1 Οι AARSs της Κλάσης I και ο διαχωρισμός τους σε 3 ομάδες με βάση τις δομικές διαφορές τους (Eriani et al. 1990; Moras 1992)

AARSs κλάσης II		
IIa	IIb	IIc
His α_2	Asp α_2	Gly α_2/b_2
Pro α_2	Asn α_2	Ala α_4
Ser α_2	Lys α_2	Phe α_2/ b_2
Thr α_2		

Πίνακας 1.2 Οι AARSs της Κλάσης II και ο διαχωρισμός τους σε 3 ομάδες με βάση τις δομικές διαφορές τους (Eriani et al. 1990; Moras 1992)

Μια κοινή παρανόηση είναι ότι το γονιδίωμα σχεδόν όλων των οργανισμών, περιέχει ένα ολοκληρωμένο σετ των 20 AARSs, με το κάθε ένα γονίδιο να είναι ξεχωριστά υπεύθυνο για την κωδικοποίηση του ενζύμου που ενεργοποιεί 1 από τα 20 γνωστά αμινοξέα, μέσω της ένωσης του με το αντίστοιχο tRNA. Λόγω της συνεχούς αυξανόμενης διαθεσιμότητας ολοκληρωμένων γονιδιωμάτων, γίνεται πιο ξεκάθαρο ότι ο διπλασιασμός γονιδίων, η οριζόντια μεταφορά γονιδίων και η απώλεια γονιδίων είναι πολύ πιο συχνά φαινόμενα μέσα στην υπεροικογένεια των AARSs, από ότι πιστεύαμε μέχρι τώρα.

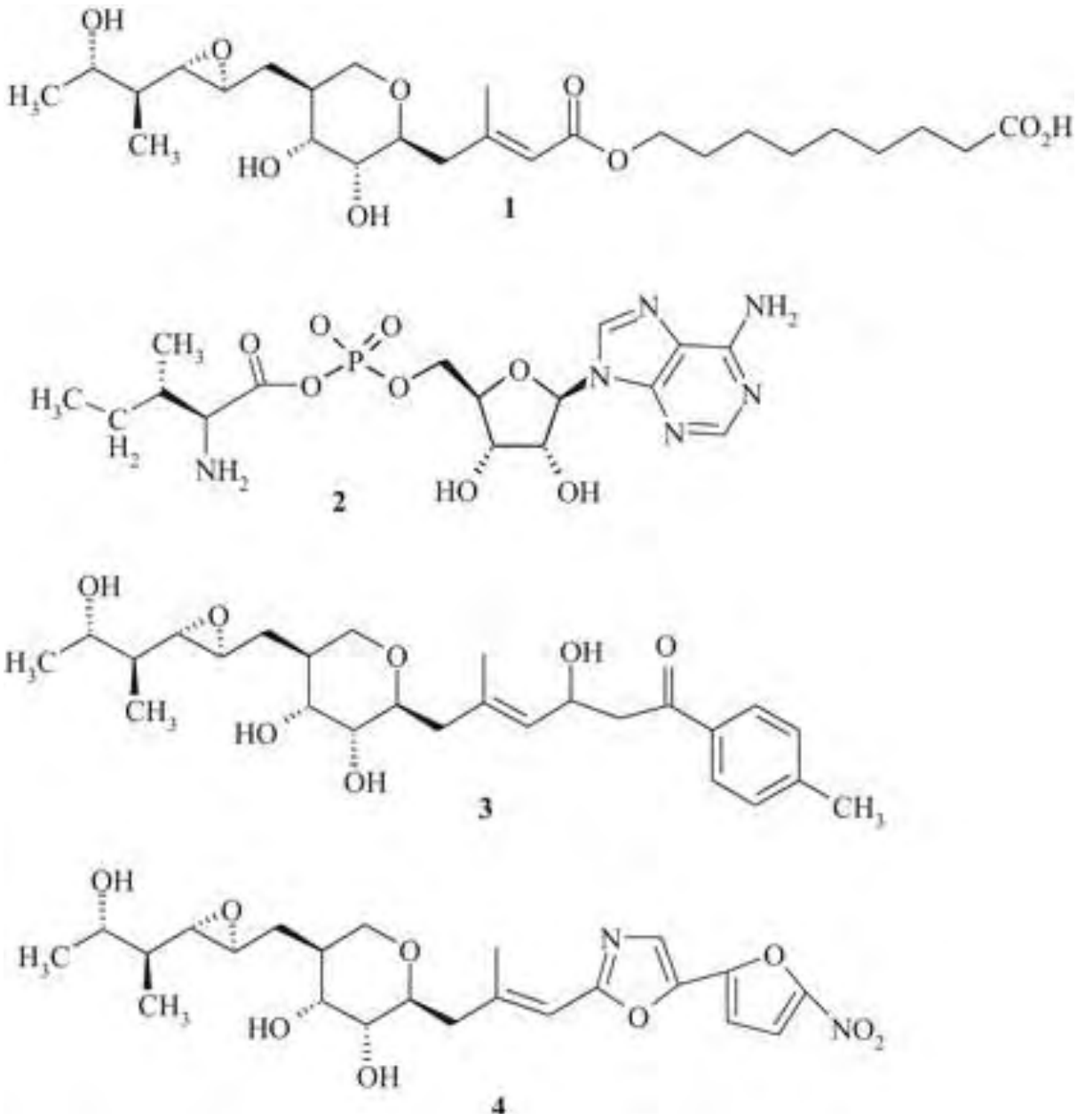
Η απουσία ενός γονιδίου που κωδικοποιεί μία AARS σε ένα γονιδίωμα είναι δυνατή γιατί δεν συμβαδίζει απαραίτητα με την απουσία της αντίστοιχης απαραίτητης βιοχημικής λειτουργίας. Για παράδειγμα, η απουσία της GlnRS καλύπτεται από τη λειτουργία μίας μη-διακριτικής (ND-non discriminated) συνθετάσης του γλουταμικού-tRNA (ND-GluRS), η οποία μπορεί να βάλει λανθασμένα το Glu σε tRNA-Gln. Στη συνέχεια, μετατρέπεται σε Gln-tRNA-Gln από μία tRNA-εξαρτώμενη αμιδοτρανσφεράση. Η ενζυμική μετατροπή ενός αμινοακυλικού-tRNA (aa-tRNA) είναι τεκμηριωμένη για Asn, Gln, Cys, σελενοκυστεΐνη και φορμυλομεθιονίνη (Becker and Kern 1998; Leinfelder et al. 1988; Ibba, Curnow, and Söll 1997; Sheppard et al. 2008; Sauerwald et al. 2005). Για αυτόν τον λόγο, η κατηγοριοποίηση όλων των περιπτώσεων όπου τα κλασσικά γονίδια AARS λείπουν, είναι το πρώτο απαραίτητο βήμα ως προς την ταυτοποίηση εναλλακτικών βιοχημικών μονοπατιών, ικανών να ενεργοποιούν τα tRNAs για τα οποία λείπουν τα συγγενή-αντίστοιχα γονίδια AARSs. Η αποκωδικοποίηση του γενετικού κώδικα είναι ένα πολύ πιο πολύπλοκο βήμα από ότι υπολογίζαμε στην αρχή και θα πρέπει να ποσοτικοποιηθεί (Ling, O'Donoghue, and Söll 2015).

Υπάρχουν πολλές αναφορές γονιδιωμάτων με παραπάνω από ένα γονίδιο για το ίδιο ένζυμο AARS ή ακόμα και για παράλογα θραύσματα που αποτελούνται από επικράτειες διαφορετικές

των AARSs (πχ. καταλυτικές, δέσμευσης αντικωδικωνίων και διαμορφωτικές). Αυτά τα παράλογα και τα θραύσματα παραλόγων, αποτελούν εδώ και καιρό το επίκεντρο ενδιαφέροντος αφού τα παράγωγα των γονιδίων τους παρουσιάζουν μία πληθώρα λειτουργιών, πέρα της μετάφρασης του γενετικού κώδικα π.χ. tRNA εξαρτώμενη σύνθεση αμινοξέων, μετα-μεταγραφική τροποποίηση, επεξεργασία των λανθασμένα-ενεργοποιημένων αμινοξέων, βακτηριακή αντίσταση στα αντιβιοτικά, συμμετοχή σε μοριακά κέντρα ελέγχου σε μονοπάτια ρύθμισης της ογκογένεσης στον άνθρωπο (Ahel et al. 2003; Andam, Fournier, and Gogarten 2011; Blaise et al. 2004; Gilbart, Perry, and Slocombe 1993; Kim, You, and Hwang 2011; Lo et al. 2014; Sissler et al. 1999). Ενδιαφέρουσες αναλύσεις έχουν επισημάνει την σημασία της οριζόντιας μεταφοράς γονιδίων (HGT) στην εξέλιξη της οικογένειας AARS (Lamour et al. 1994) και έχει ήδη βρεθεί ότι αυτή είναι συχνά συνδεδεμένη με την αντίσταση στα αντιβιοτικά, ειδικότερα στα βακτήρια (Andam, Fournier, and Gogarten 2011; Fournier, Andam, and Gogarten 2015; O'Donoghue and Luthey-Schulten 2003; Woese et al. 2000; Wolf et al. 1999). Το γεγονός ότι οι βακτηριακές AARSs δεν συμμετέχουν συχνά σε πολύπλοκες αλληλεπιδράσεις μεταξύ πρωτεϊνών, ενώ συχνά είναι συμβατές με tRNAs από εξελικτικά απομακρυσμένους οργανισμούς, υποδηλώνει τη αυξημένη πιθανότητα σωστής λειτουργίας ακόμα και αν έχουν μεταφερθεί με οριζόντια γονιδιακή μεταφορά σε εξελικτικά απομακρυσμένους οργανισμούς.

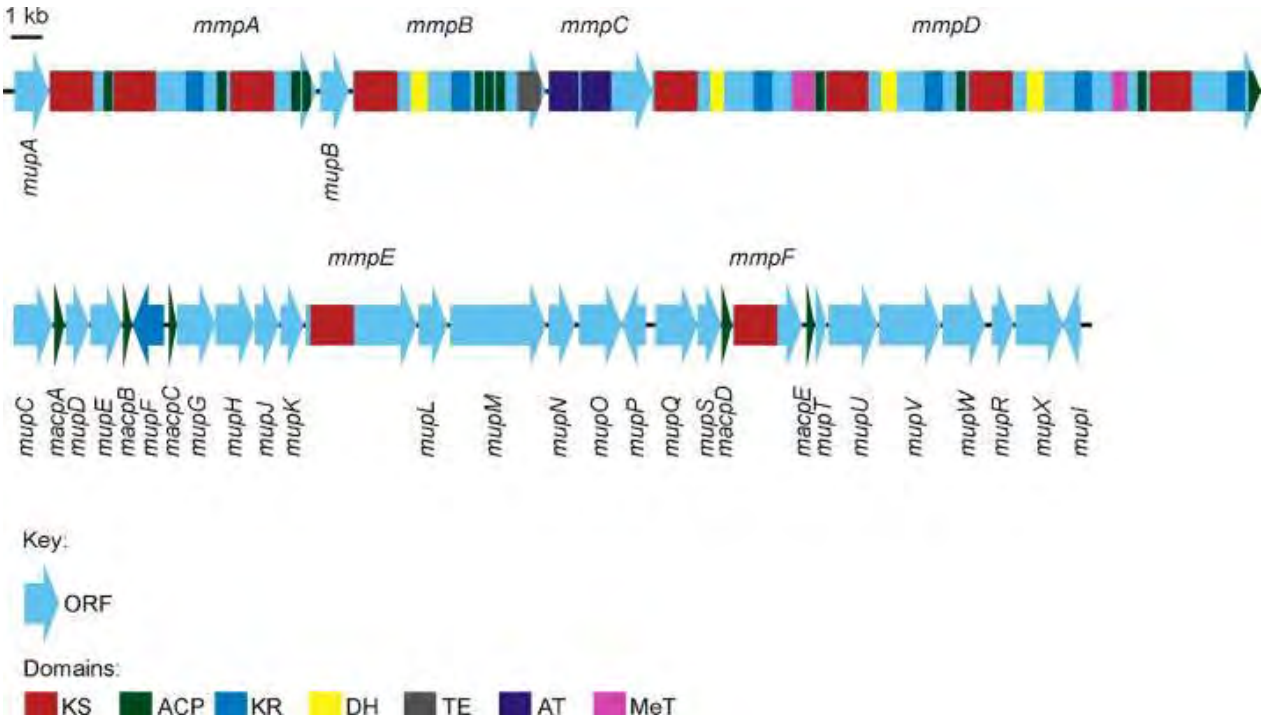
Πολλοί μικροοργανισμοί διαθέτουν τοξίνες μικρού μοριακού βάρους που στοχεύουν αυτά τα σημαντικά ένζυμα σε άλλους μικρο-οργανισμούς. Τέτοιες τοξίνες έχουν ήδη ταυτοποιηθεί για τα AlaRS, AspRS, AsnRS, IleRS, LeuRS, LysRS, MetRS, ProRS, SerRS, ThrRS, TyrRS, TrpRs, PheRS (Andam, Fournier, and Gogarten 2011; Pham et al. 2014). Αντιστοίχως, οι μικροοργανισμοί που δέχονται επίθεση από αυτές τις τοξίνες αποκτούν είτε μικρού επιπέδου αντίσταση μέσω σημειακών-μεταλλάξεων στις υπό στόχευση ενδογενείς AARSs, είτε αποκτούν ένα παράλογο AARS που διαθέτει αντίσταση σε αυτές τις τοξίνες, από άλλους οργανισμούς μέσω των οριζόντιας γονιδιακής μεταφοράς (Antonio, McFerran, and Pallen 2002; Yanagisawa and Kawakami 2003). Το πιο επιφανές παράδειγμα τέτοιων τοξινών είναι το Mupirocin (Gilbart, Perry, and Slocombe 1993), ένα φυσικό αντιβιοτικό το οποίο κυκολοφορεί στην αγορά. Στην εικόνα 1.2. μπορείτε να δείτε τον χημική δομή του Mupirocin σε σύγκριση με τη δομή της Ile-AMP. Η *Pseudomonas fluorescens* NCIMB 10586 περιέχει στο γονιδίωμα της μία συστάδα (cluster) γονιδίων συνθάσης πολυκετιδίων, τα οποία είναι υπεύθυνα για την δημιουργία του ψευδομονικού οξέως (ευρέως γνωστό και ως Mupirocin), ένα φυσικό προϊόν που στοχεύει/αναστέλλει την IleRS. Ενδιαφέρον προκαλεί το γεγονός, ότι στο ίδιο γονιδιακό σύμπλεγμα βρίσκεται και ένα παράλογο της IleRS, το οποίο δεν επηρεάζεται από τη συγκεκριμένη τοξίνη (Yanagisawa and Kawakami 2003; El-Sayed et al. 2003). Έτσι, μέσα στην ίδια συστάδα γονιδίων βρίσκονται και το δηλητήριο και το αντίδοτο.

Η ανακάλυψη ότι τα αποκλίνοντα παράλογα των AARS είναι πιθανόν να παρέχουν αντίσταση σε φυσικούς ανασταλτικούς παράγοντες των AARS, έχει τεκμηριωθεί για παράλογα των MetRS, TrpRS, IleRS και SerRS (Andam, Fournier, and Gogarten 2011; Ochsner et al. 2007). Για αυτόν τον λόγο, ο εντοπισμός των παραλόγων των AARS, μπορεί να οδηγήσει στην ανακάλυψη νέων αντιβιοτικών, που είτε είναι κωδικοποιημένα μέσα σε ένα σύμπλεγμα γονιδίων στην γειτονιά των παραλόγων, είτε σε κάποιο διαφορετικό γονιδιακό τόπο.



Εικόνα 1.2 : Χημική δομή της Mupirocin (δομή 1), isoleucyl-AMP (Ile-AMP; δομή 2), και παράγωγα της mupirocins (δομές 3 και 4) (Hurdle, O'Neill, and Chopra 2005)

Επιπλέον, πιθανοί φυσικοί ανασταλτικοί παράγοντες μπορούν να ταυτοποιηθούν από αναλύσεις συγκριτικής γονιδιωματικής μεταξύ κοντινών βακτηριακών στελεχών. Π.χ. ένα από τα στελέχη μπορεί να διαθέτει ένα παράλογο AARS, ενώ τα άλλα στελέχη έχουν έλλειψη από αντίγραφα AARS. Στη προκειμένη περίπτωση αξίζει να εξετασθεί, για να εξακριβωθεί η πιθανότητα ύπαρξη παραγωγής κάποιου ανασταλτικού παράγοντα για τη συγκεκριμένη AARS. Αν ο οργανισμός παράγει μία ουσία με τη συγκεκριμένη ιδιότητα, τότε πολύ πιθανόν η συγκεκριμένη ουσία να μπορεί να χρησιμοποιηθεί μελλοντικά ως αντιβιοτικό.



Εικόνα 1.3 : Γονιδιακή συστάδα υπεύθυνη για τη βιοσύνθεση της Mupirocin από το βακτήριο *Pseudomonas fluorescens* NCIMB 10586 (El-Sayed et al. 2003)

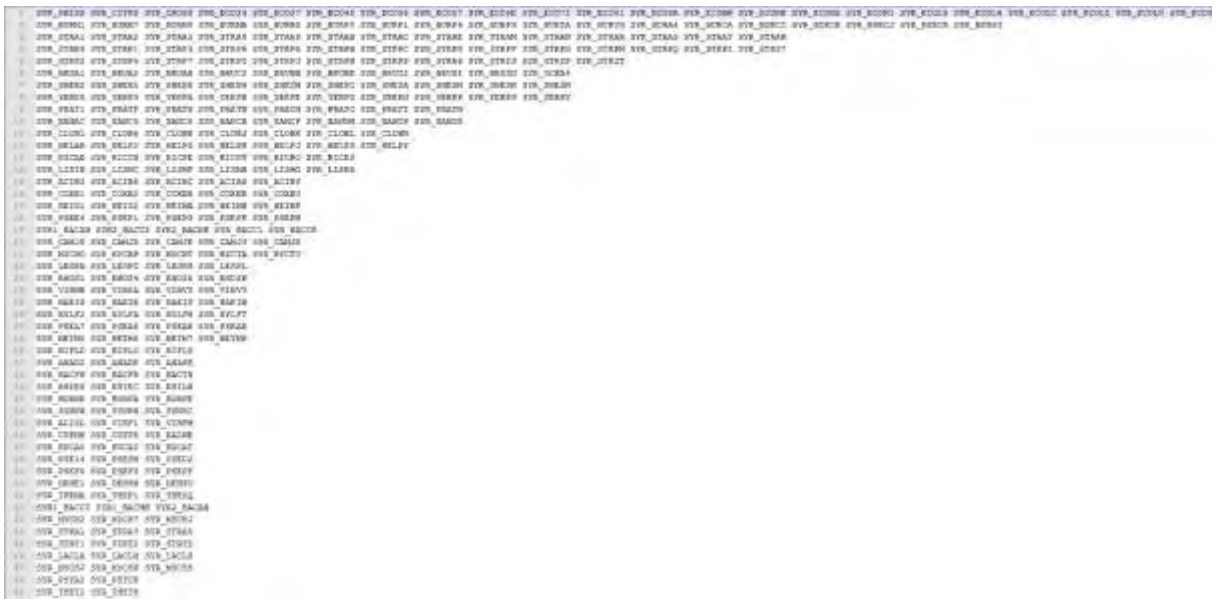
Ο χειρισμός και η επέκταση του γενετικού κώδικα θεωρείται ο ακρογωνιαίος λίθος της συνθετικής βιολογίας και βιοτεχνολογίας. Συγκεκριμένα, η ενεργοποίηση συγκεκριμένων tRNAs με τη βοήθεια συνθετικών αμινοξέων είναι σημαντική για την κατασκευή ασφαλών GMOs, καθώς και για τη κατασκευή-δημιουργία ενζύμων με νέες δυνατότητες (L.-T. Guo et al. 2014; Hadd and Perona 2014; Passioura and Suga 2014; Wang, Xie, and Schultz 2006). Για αυτό το σκοπό, είναι σημαντικό να καταλάβουμε πως προσδιορίζεται η εξειδίκευση των AARSs για τα αμινοξέα και τα tRNAs. Για αυτό το σκοπό, ο εντοπισμός των συντηρημένων μοτίβων που κατηγοριοποιούν κάθε συγκεκριμένο ένζυμο AARS, είναι απαραίτητη προϋπόθεση για την μελλοντική διαχείριση και τις ορθές και ακριβείς μελέτες μεταλλαξιγένεσης.

Οι στόχοι της εργασίας μας ήταν:

1. Ταυτοποίηση, με έναν αμερόληπτο τρόπο, υψηλής συντήρησης μοτίβων και επικρατειών για κάθε AARS.
2. Ανάπτυξη ενός καινούριου και ευαίσθητου διαδικτυακού υπολογιστικού εργαλείου για την ανίχνευση και ταυτοποίηση AARSs (βασισμένο σε αυτά τα μοτίβα και τις επικράτειες).
3. Ανάλυση ~2500 προκαρυωτικών πρωτεωμάτων με σκοπό την εξελικτική μελέτη γονιδίων AARS και των παραλόγων τους.
4. Αποθήκευση και οργάνωση όλων των παραπάνω πληροφοριών σε μία καλά οργανωμένη βάση δεδομένων, στην οποία να υπάρχει ελεύθερη πρόσβαση για μελλοντικές έρευνες.
5. Ταυτοποίηση πιθανών υποψηφίων μικροοργανισμών για την παραγωγή κατασταλτικών παραγόντων (αντιβιοτικών) AARS.

2. Υλικά και Μέθοδοι

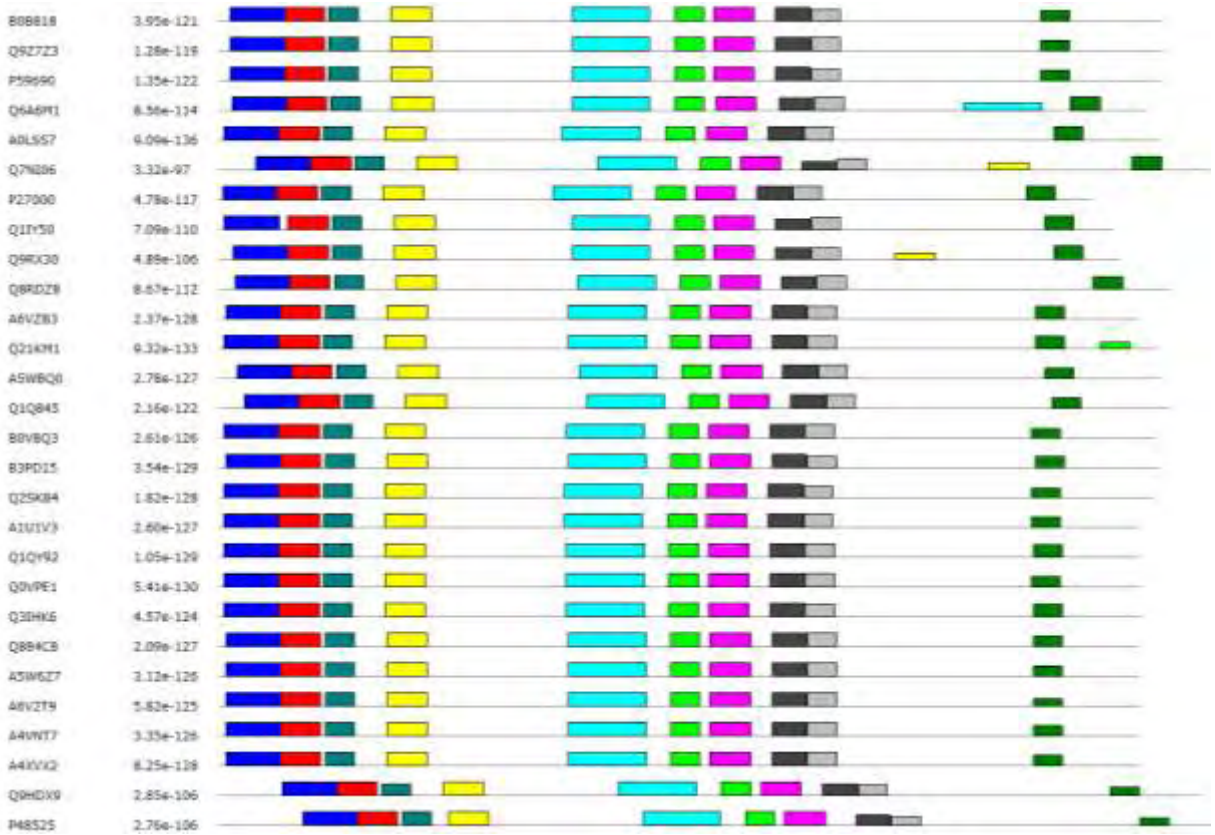
Όλες οι πρωτεϊνικές ακολουθίες που χρησιμοποιήθηκαν για την κατασκευή των κρυφών Μαρκοβιανών μοντέλων (HMMs), είχαν ληφθεί από τη βάση δεδομένων UniProt/SwissProt (UniProt Consortium 2015). Στη συνέχεια, δημιουργήθηκαν 2 σύνολα δεδομένων (datasets) με τις συγκεκριμένες πρωτεΐνες. Το 1^{ο} dataset περιείχε πρωτεΐνες που είχαν έως και 95% ομοιότητα στην πρωτεϊνική του ακολουθία, ενώ στο 2^{ο} dataset, το ποσοστό ομοιότητας ήταν 70%. Η δημιουργία των 2 datasets έγινε με τη χρήση του αλγόριθμου BlastClust. Αυτό είχε ως αποτέλεσμα στο 1^{ο} dataset να έχουμε 7517 πρωτεϊνικές ακολουθίες, ενώ στο 2^{ο} dataset είχαμε 3276 πρωτεϊνικές ακολουθίες. Π.χ. στην εικόνα 2.1, η ομοιότητα των πρωτεϊνικών αλληλουχιών της πρώτης στήλης, δεν ξεπερνάει το 95%.



Εικόνα 2.1: Αποτελέσματα από το blastclust για τις ακολουθίες των επικρατειών της ArgRS. Κάθε σειρά αποτελεί και μία διαφορετική συστάδα, στην οποία οι ακολουθίες έχουν ομοιότητα <95%

Στη συνέχεια, από τον ftp server του NCBI λήφθηκαν ολοκληρωμένα πρωτεώματα από 2588 προκαρυωτικούς οργανισμούς στους οποίους έχει ολοκληρωθεί η αλληλούχιση του γονιδιώματός τους. Οι πρωτεϊνικές ακολουθίες ήταν σε αρχεία με μορφή FASTA, ενώ οι πληροφορίες για τον κάθε οργανισμό και τα γονίδια του, βρίσκονταν σε αρχεία με μορφή ptt.

Συντηρημένα μοτίβα ανιχνεύθηκαν σε κάθε ένζυμο της υπεροικογένειας AARS, χρησιμοποιώντας τη σουίτα προγραμμάτων MEME (Bailey et al. 2015). Έχει προταθεί από προηγούμενες μελέτες, ότι η αρχιτεκτονική των επικρατειών μπορεί να χρησιμοποιηθεί σαν δείκτης για την φυλογενετική ανασυγκρότηση αυτής της υπεροικογένειας ενζύμων (Wolf et al. 1999).



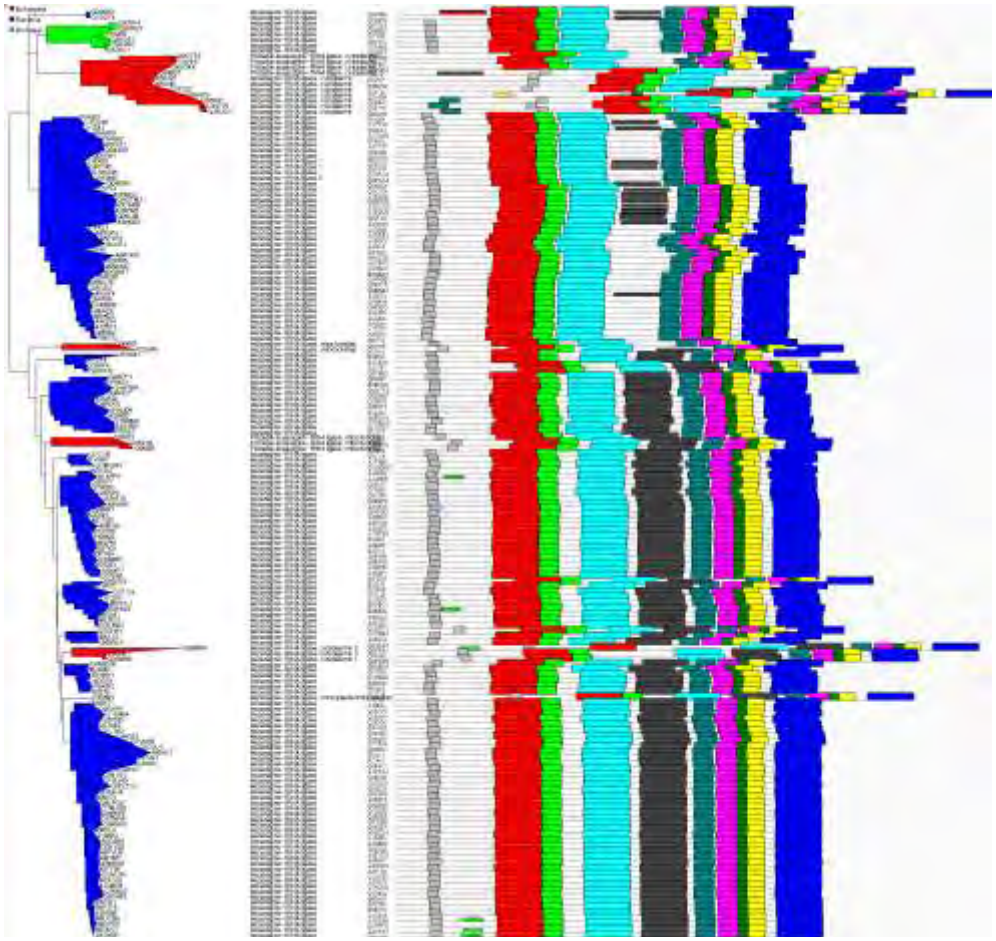
Εικόνα 2.2: Αποτέλεσμα από τη σουίτα προγραμμάτων MEME για την αρχιτεκτονική των μοτίβων σε πρωτεϊνικές ακολουθίες της GluRS. Με διαφορετικό χρώμα φαίνονται τα διαφορετικά συντηρημένα μοτίβα που μπόρεσε να εντοπίσει ο αλγόριθμος του MEME στις συγκεκριμένες ακολουθίες

Σε γενικές γραμμές, η εμφάνιση των επικρατειών έγινε στην αρχή της εξελικτικής πορείας των πρωτεϊνών, ωστόσο η αναδιάρθρωση των επικρατειών είναι ένα πολύ κοινό φαινόμενο (Chothia et al. 2003; Marsh and Teichmann 2010; Teichmann et al. 2001; Weiner, Moore, and Bornberg-Bauer 2008). Η δημιουργία των HMMs και μετέπειτα η σάρωση των πρωτεϊνικών ακολουθιών με τα συγκεκριμένα HMMs, έγινε με τη χρήση του προγράμματος HMMER (Eddy 2011). Οι πολλαπλές στοιχίσεις των ακολουθιών έγινε με το πρόγραμμα MUSCLE (Edgar 2004), ενώ η δημιουργία των φυλογενετικών δέντρων έγινε με τη χρήση του αλγόριθμου BioNJ, μέσω του προγράμματος Seaview (Gouy, Guindon, and Gascuel 2010). Για την απεικόνιση των φυλογενετικών δέντρων χρησιμοποιήθηκε το πρόγραμμα Treedyn (Chevenet et al. 2006). Η σύμπτυξη των φυλογενετικών δέντρων μαζί με την αρχιτεκτονική των μοτίβων (MEME-motifs) που είχαν εντοπισθεί με τη σουίτα προγραμμάτων MEME, έγινε σε σελίδες html με χρήση της γλώσσας προγραμματισμού Javascript.

Όλη η διαχείριση των δεδομένων έγινε με τη χρήση μικρών προγραμμάτων που είχαν δημιουργηθεί στο εργαστήριο, χρησιμοποιώντας τη γλώσσα προγραμματισμού PERL. Η στοίχιση ανά ζεύγη και ο υπολογισμός της ομοιότητας των ομόλογων πρωτεϊνών μέσα στο ίδιο γονιδίωμα, έγινε με τη χρήση του προγράμματος EMBOSSWater.

Η κατασκευή των HMMs για τις καταλυτικές επικράτειες, επικράτειες διαμόρφωσης και επικράτειες πρόσθεσης tRNA, βασίστηκε σε ενσωμάτωση πληροφοριών από πολλές πηγές, όπως οι πολλαπλές στοιχίσεις, η αρχιτεκτονική των MEME-motif, από την βιβλιογραφία (Woese et al. 2000; Wolf et al. 1999), από τον χαρακτηρισμό των διαφορετικών επικράτειων από τις βάσεις δεδομένων PDB (Rose et al. 2015), Astral (Fox, Brenner, and Chandonia 2014), InterPro (Mitchell et al. 2015) και CDD (Marchler-Bauer et al. 2015).

Η οργάνωση και η αποθήκευση των δεδομένων έγινε σε μία βάση δεδομένων MySQL. Επίσης, αναπτύχθηκε ένα διαδικτυακό γραφικό περιβάλλον, το οποίο είναι χωρισμένο σε 2 μέρη. Στο κομμάτι που τρέχει στον εξυπηρετητή (server) και ονομάζεται back-end και στο κομμάτι του προγράμματος που δουλεύει στον περιηγητή διαδικτύου (browser) και ονομάζεται front-end. Το back-end αναπτύχθηκε χρησιμοποιώντας τη γλώσσα προγραμματισμού Java σε συνδυασμό με δομές Spring, ενώ το front-end αναπτύχθηκε με χρήση της γλώσσας Javascript σε συνδυασμό με δομές Angular JS. Η επικοινωνία του back-end με το front-end γίνεται μέσω ενός πιστοποιημένου RESTful API. Η ανάπτυξη της βάσης δεδομένων και του γραφικού περιβάλλοντος έγινε κυρίως από τον υποψήφιο διδάκτορα Παναγιώτη Βλασταρίδη.

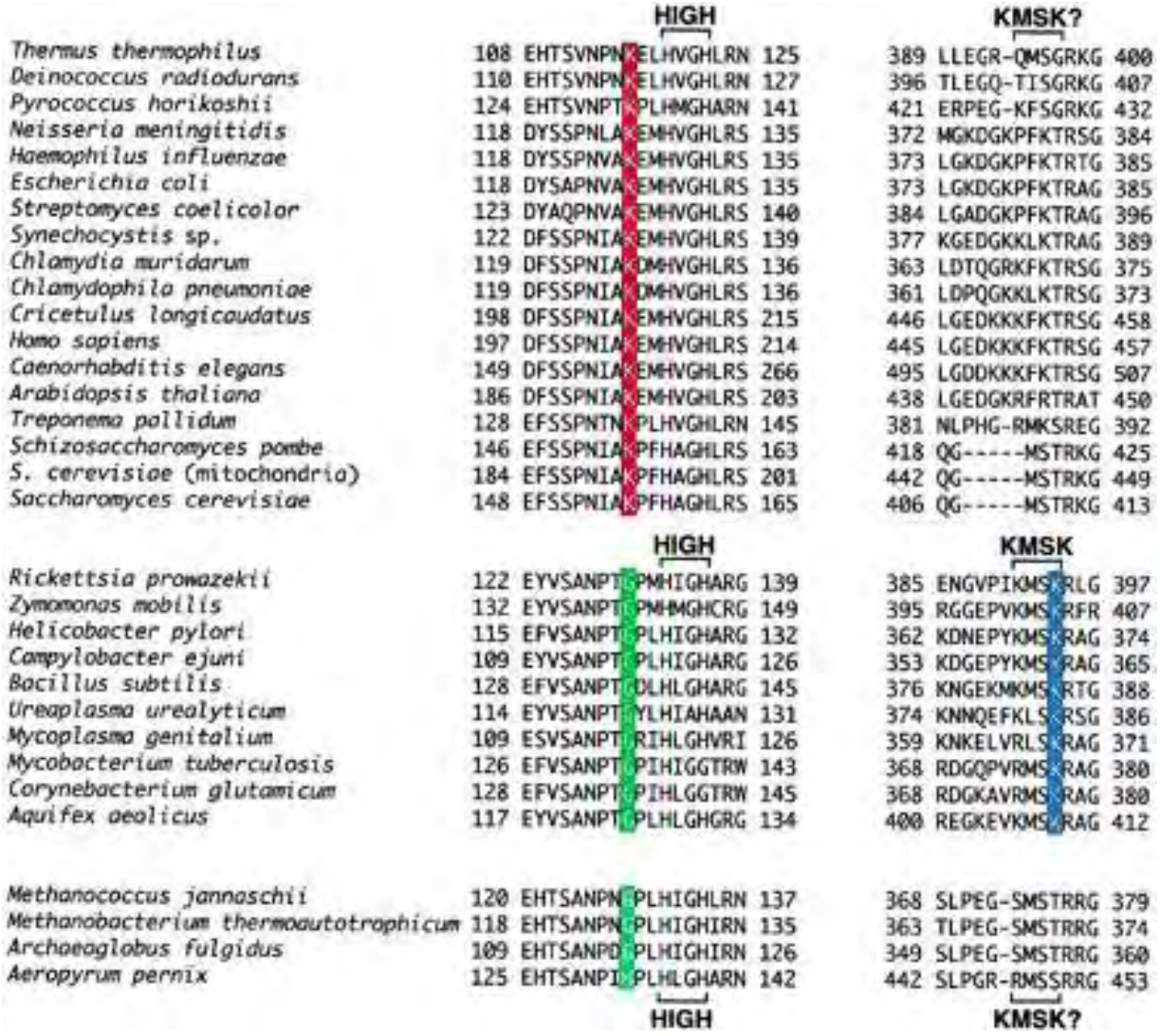


Εικόνα 2.3: Μικρογραφία του φυλογενετικού δέντρου της AsnRS με σύμπτυξη των πληροφοριών της αρχιτεκτονικής των MEME-motifs με χρήση javascript

3. Αποτελέσματα και Συζήτηση

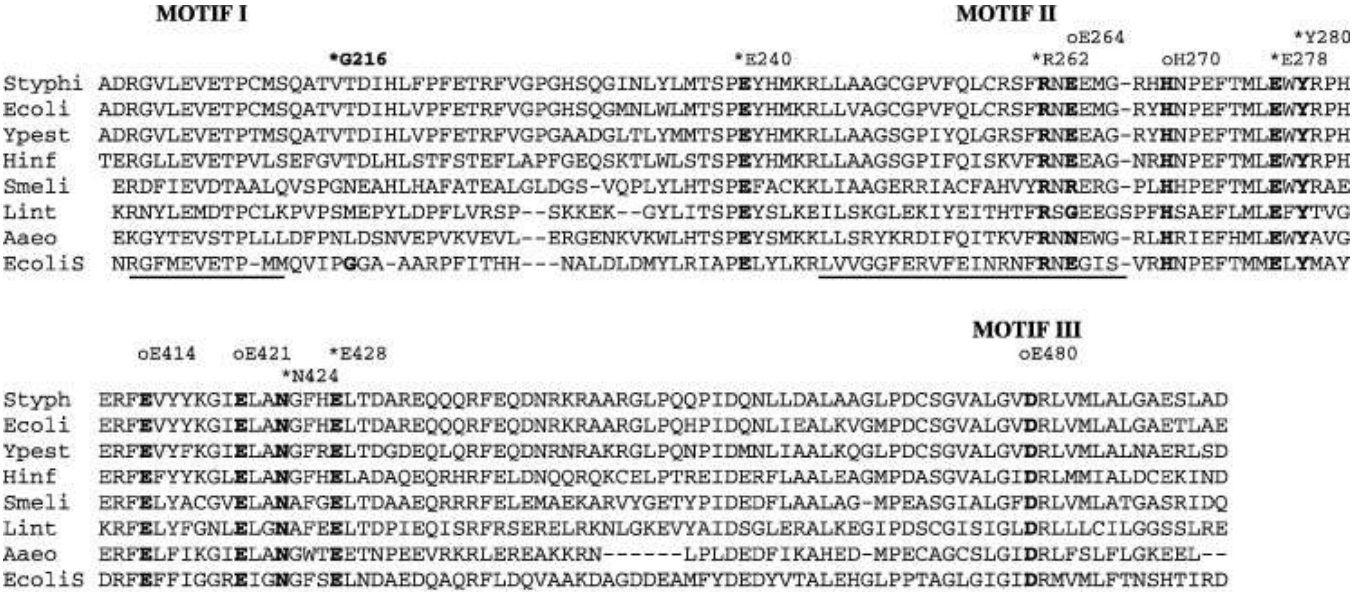
3.1 Εντοπισμός συντηρημένων μοτίβων σε κάθε ένα από τα ένζυμα AARS

Αρχικά χρησιμοποιήθηκαν 3276 ακολουθίες από τα βάση δεδομένων SwissProt (αφαίρεση πρωτεϊνών με ομολογία άνω του 70%) σε κάθε μία από τις δύο κλάσεις (1739 I 1537 II) και αναλύθηκαν με το πρόγραμμα MEME, με σκοπό τον εντοπισμό κοινών συντηρημένων μοτίβων μέσα σε κάθε μία από τις 2 ομόλογες κλάσεις. Είναι ήδη γνωστό από τη βιβλιογραφία ότι τα ένζυμα της τάξης I περιέχουν μία πτυχή Rossman με 2 χαρακτηριστικά μοτίβα, το HIGH και το KMSKS (βλ. εικ. 3.1), ενώ τα ένζυμα της τάξης 2 περιέχουν 3 συντηρημένα μοτίβα (βλ. εικ. 3.2) (Eriani et al. 1990).

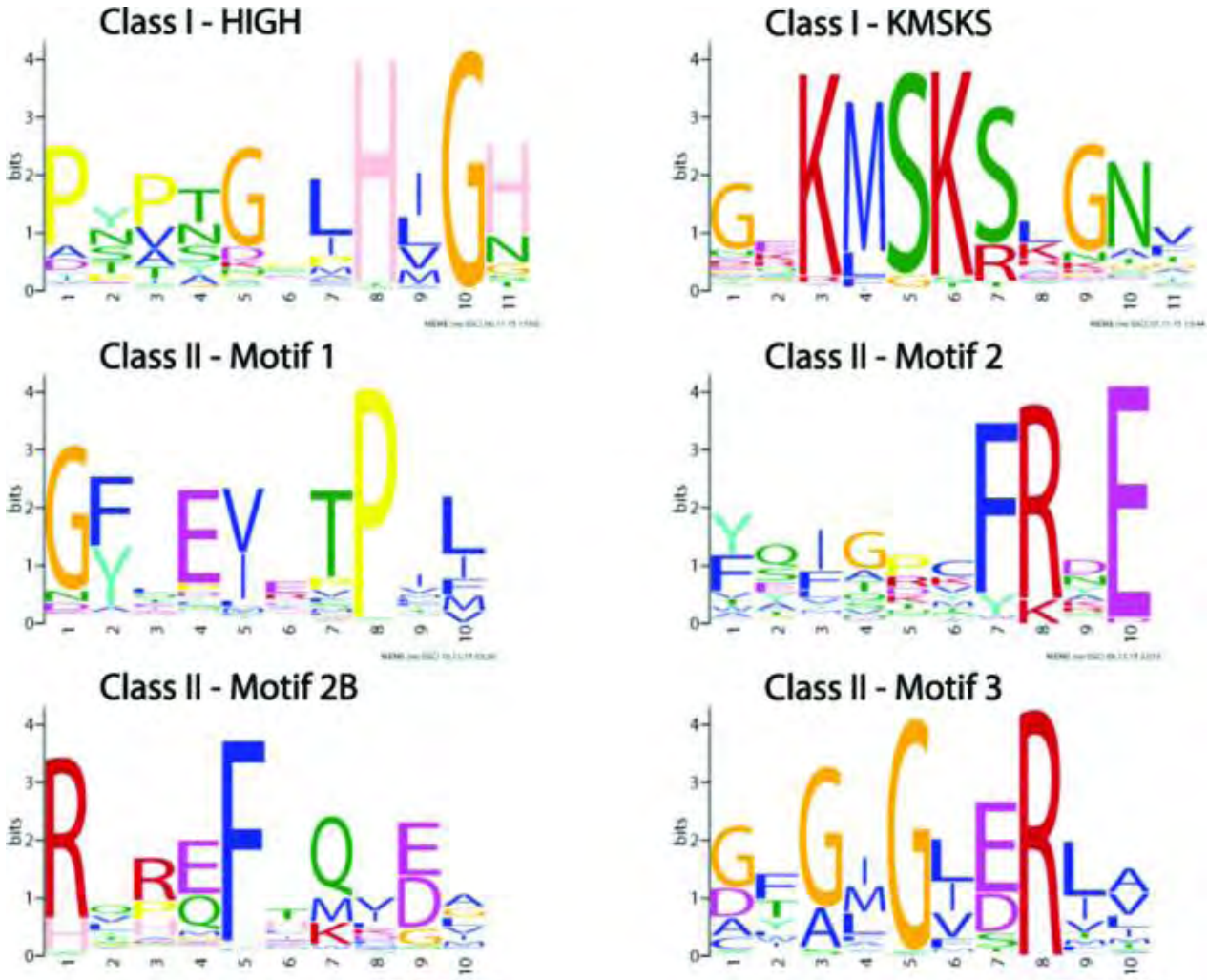


Εικόνα 3.1 Τα HIGH και KMSKS μοτίβα που βρέθηκαν με πολλαπλή στοίχιση σε ακολουθίες της ArgRS (Sekine et al. 2001).

Η ανάλυση με το πρόγραμμα MEME αναγνώρισε επιτυχώς τα HIGH και KMSKS μοτίβα, στο 98% (1708/1739) και στο 80% (1394/1739) των πρωτεϊνών της κλάσης I. Επίσης τα μοτίβα 1,2 και 3 αναγνωρίστηκαν στο 59% (901/1537), στο 73% (1121/1537) και στο 70% (1069/1537) των πρωτεϊνών της κλάσης II. Ο λόγος είναι ότι αυτά τα 3 μοτίβα και κυρίως το μοτίβο 1 δεν έχουν συντηρηθεί τόσο έντονα στα ένζυμα κλάσης II. Είναι ενδιαφέρον ότι για πρώτη φορά, ένα τέταρτο μοτίβο, το οποίο ονομάστηκε μοτίβο 2B, λόγω του εντοπισμού του στο C-terminal του μοτίβου 2, επίσης αναγνωρίστηκε στο 73% (1124/1537)) των πρωτεϊνών της κλάσης II. Από αυτά τα LOGO (βλ. εικ. 3.3) φαίνεται ότι τα συγκεκριμένα μοτίβα έχουν πολύ μικρό μήκος, λίγων μόνο αμινοξέων. Όταν ο αλγόριθμος MEME αναλύει τις ομόλογες ακολουθίες των 2 κλάσεων, είναι ικανός να αναγνωρίσει τα πολύ μικρά μοτίβα που εμπλουτίζονται σε κάθε κλάση αυτής της υπεροικογένειας. Ωστόσο, τα προφίλ HMMs, λόγω του μικρού μήκους αυτών των μοτίβων, έχουν πολύ περιορισμένη ικανότητα ανίχνευσης. Ως εκ τούτου, η ανίχνευση ομόλογων AARSs που βασίζεται αποκλειστικά σε αυτά τα πολύ συντηρημένα αλλά ταυτόχρονα μικρού μήκους μοτίβων, είναι αρκετά προβληματική. Για αυτό το λόγο, είναι αναγκαία η χρήση διαφορετικής προσέγγισης, όπου μία σειρά από συντηρημένα μοτίβα, χρησιμοποιούνται στο επίπεδο του ενζύμου, και όχι στο επίπεδο της κλάσης.



Εικόνα 3.2 Τα τρία μοτίβα που εμφανίζονται στις συνθετάσες της κλάσης II (Ambrogelly et al. 2010)



Εικόνα 3.3: Τα χαρακτηριστικά μοτίβα HIGH και KMSKS της κλάσης I, και τα χαρακτηριστικά μοτίβα της κλάσης II Motif 1, Motif 2 και Motif 3 καθώς επίσης και το Motif 2B, όπως εντοπίστηκαν από τον αλγόριθμό του MEME

3.2 Ανακάλυψη συντηρημένων μοτίβων σε κάθε ένζυμο AARS και φυλογενετική ανάλυση

Για κάθε ένα από τα 20 ένζυμα AARSs, καθώς και για τα ένζυμα PylRS και SepRS, οι αντίστοιχες αναγνωρισμένες πρωτεΐνες από τη βάση δεδομένων SwissProt αναλύθηκαν με το λογισμικό MEME, για τον εντοπισμό συντηρημένων μοτίβων εντός του κάθε ενζύμου. Η αναγνώριση και η αρχιτεκτονική των μοτίβων MEME μας επέτρεψε να διαχωρίσουμε τη κάθε ενζυμική ομάδα σε ξεχωριστές φυλογενετικές υποομάδες. Επιπλέον, μέσω αυτής της προσέγγισης, είχαμε μία καλύτερη κατανόηση για το βαθμό διαφοροποίησης του κάθε

ορθόλογου, αλλά ακόμα πιο σημαντικό είναι ότι καταφέραμε να ξεχωρίσουμε παράλογα ή θραύσματα παραλόγων. Για την επίτευξη αυτού του στόχου, το πρόγραμμα MEME έτρεξε με τις συγκεκριμένες παραμέτρους (10 μοτίβα για κάθε ένζυμο εκτός των Asp-Asn-AARS και Glu-Gln-AARS, όπου το MEME ρυθμίστηκε να εντοπίσει 20 μοτίβα).

Με βάση τον αριθμό και τις συντεταγμένες πάνω στην ακολουθία των αναγνωρισμένων μοτίβων που υπήρχαν σε όλες τις πρωτεϊνικές αλληλουχίες ενός συγκεκριμένου ενζύμου (αριθμός που κυμαίνεται μεταξύ 2-10 μοτίβων), κατασκευάστηκαν BioNJ φυλογενετικά δέντρα και στη συνέχεια απεικονίστηκαν με το λογισμικό Treedyn. Πιο συγκεκριμένα, οι ακολουθίες των μοτίβων που ήταν παρούσες σε όλες τις πρωτεϊνικές ακολουθίες του κάθε ενζύμου, συνενώθηκαν σε μία ακολουθία, και στη συνέχεια αυτή η ακολουθία χρησιμοποιήθηκε για να δημιουργηθεί το φυλογενετικό δέντρο. Στη συνέχεια, σε κάθε ένα δέντρο, για κάθε ένα ένζυμο της οικογένειας AARS ενσωματώθηκε η αρχιτεκτονική των μοτίβων, σύμφωνα με το πρόγραμμα MEME, χρησιμοποιώντας javascript, ενώ η απεικόνιση των δέντρων έγινε σε ιστοσελίδες HTML. Με βάση το φυλογενετικό δέντρο, και την αρχιτεκτονική των μοτίβων, όποτε ήταν εφαρμόσιμο, εντοπίστηκαν φυλογενετικές υποομάδες με αυτοψία εντός ενός συγκεκριμένου ενζύμου. Σε κάθε μία από αυτές τις υποομάδες, ξανακάναμε την ίδια διαδικασία με εντοπισμό μοτίβων από το πρόγραμμα MEME και δημιουργία φυλογενετικών δέντρων με βάση τα κοινά συντηρημένα μοτίβα. Τα νέα φυλογενετικά δέντρα απεικονίστηκαν ξανά με την παραπάνω διαδικασία.

Με αυτήν την επαναληπτική τεχνική, καταφέραμε να αντιμετωπίσουμε αποτελεσματικά τις εξαιρετικά αποκλίνουσες υποομάδες μέσα σε ένα ένζυμο AARS, ή ακόμα και σε ομάδες που είναι παραφυλετικές, όπως στη περίπτωση του ενζύμου GlyRS (Valencia-Sánchez et al. 2016).

Ένα παράδειγμα της παραπάνω ανάλυσης είναι το ακόλουθο. Για το ένζυμο ProRS, υπήρχαν 500 ακολουθίες στη βάση SwissProt. Αυτές οι ακολουθίες υποβλήθηκαν σε έναν πρώτο γύρο εντοπισμού 10 μοτίβων, με τη χρήση του προγράμματος MEME. Από τα 10 μοτίβα που εντοπίστηκαν, μόνο 2 ήταν κοινά σε όλες τις ακολουθίες (που μαζί αποτελούνταν από 36 αμινοξέα). Στη συνέχεια αυτά τα 2 μοτίβα, ενώθηκαν σε μία ακολουθία με συνολικό μήκος 36 αμινοξέα, και αυτή η ακολουθία χρησιμοποιήθηκε για την κατασκευή ενός φυλογενετικού δέντρου (Supplementary Figure S1). Έπειτα, σε αυτό το φυλογενετικό δέντρο ενσωματώθηκαν οι πληροφορίες για την αρχιτεκτονική των μοτίβων, χρησιμοποιώντας τα αποτελέσματα του MEME. Με βάση λοιπόν το φυλογενετικό δέντρο και την αρχιτεκτονική των μοτίβων, αποφασίστηκε να αντιμετωπιστεί το ένζυμο ProRS ως δύο διαφορετικές φυλογενετικές υποομάδες.

Η υποομάδα 1 περιελάβανε 386 ακολουθίες και η υποομάδα 2 περιελάβανε 114 ακολουθίες. Στη συνέχεια, η κάθε μία υποομάδα υποβλήθηκε σε νέο γύρο με το λογισμικό MEME για τον εντοπισμό 10 μοτίβων. Αυτή τη φορά, στη πρώτη υποομάδα εντοπίστηκαν 8 μοτίβα που ήταν κοινά και στις 386 ακολουθίες. Με βάση αυτά τα 8 μοτίβα, κατασκευάστηκε μια ακολουθία με μήκος 241 αμινοξέα, και διεξήχθη μία δεύτερη φυλογενετική ανάλυση. Στο νέο δέντρο ενσωματώθηκαν πάλι πληροφορίες για την αρχιτεκτονική των μοτίβων. Η διαδικασία ακολουθήθηκε και για τη 2η φυλογενετική υποομάδα της ProRS. Αυτή τη φορά 9 μοτίβα ήταν κοινά και στις 114 ακολουθίες. Με βάση τα 9 μοτίβα, δημιουργήθηκε μία ακολουθία συνολικού

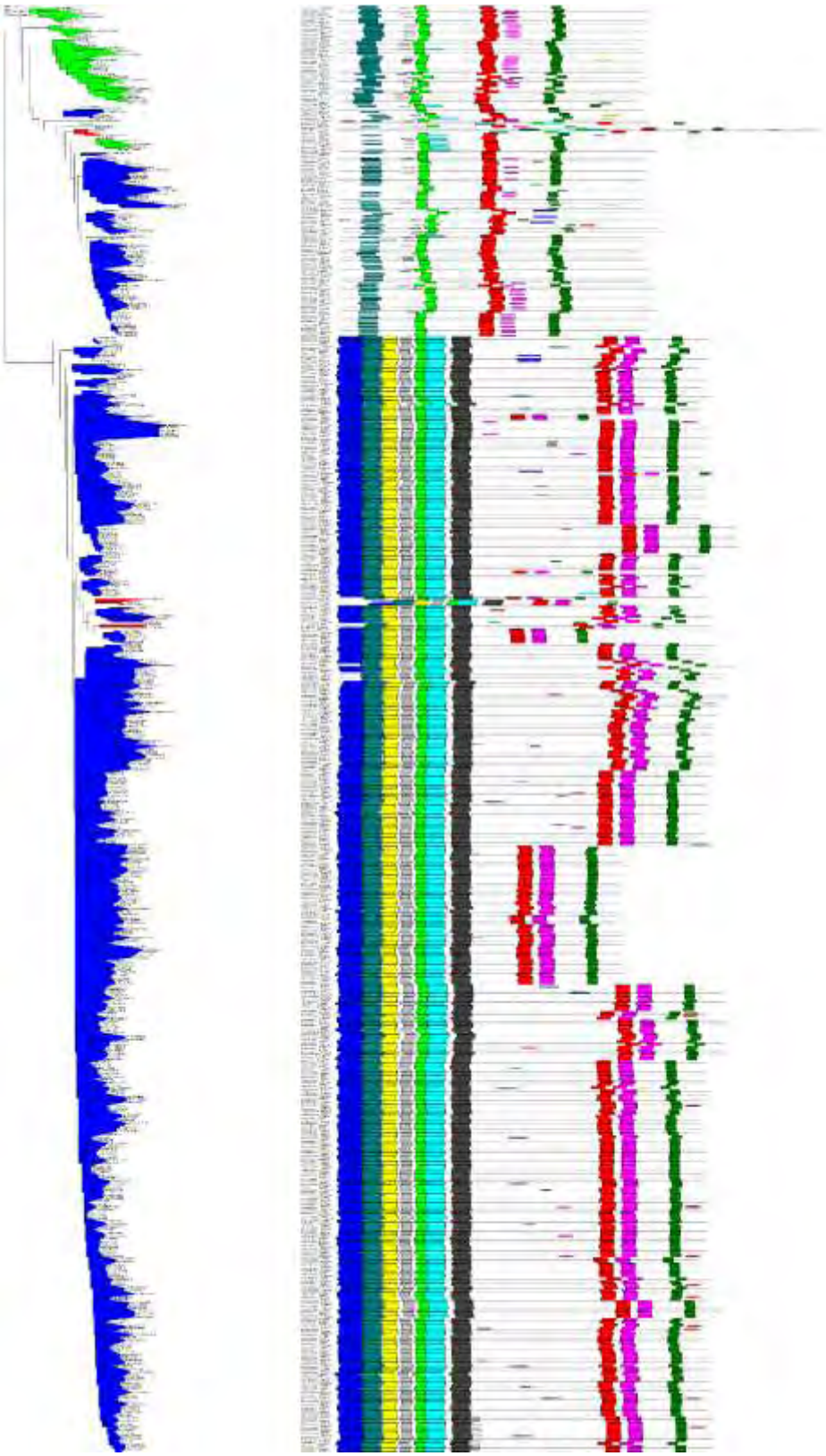
μήκους 229 αμινοξέα. Αυτή η ακολουθία χρησιμοποιήθηκε για να την δημιουργία ενός νέου φυλογενετικού δέντρου. Έπειτα σε αυτό το δέντρο προστέθηκαν οι πληροφορίες από το MEME για την αρχιτεκτονική των μοτίβων και η απεικόνιση του έγινε πάλι σε HTML ιστοσελίδα με χρήση javascripts. Στις εικόνες 3.4, 3.5 και 3.6 μπορείτε να δείτε τα 3 διαφορετικά φυλογενετικά δέντρα που δημιουργήθηκαν για τη ProRS σε συνδυασμό με την αρχιτεκτονική των συντηρημένων μοτίβων.

Για όλα τα ένζυμα AARSs απο τη βάση δεδομένων SwissProt, τα φυλογενετικά δέντρα μαζί με τα αντίστοιχα μοτίβα που εντοπίστηκαν , είναι διαθέσιμα σε μορφή HTML στην ιστοσελίδα http://bioinf.bio.uth.gr .

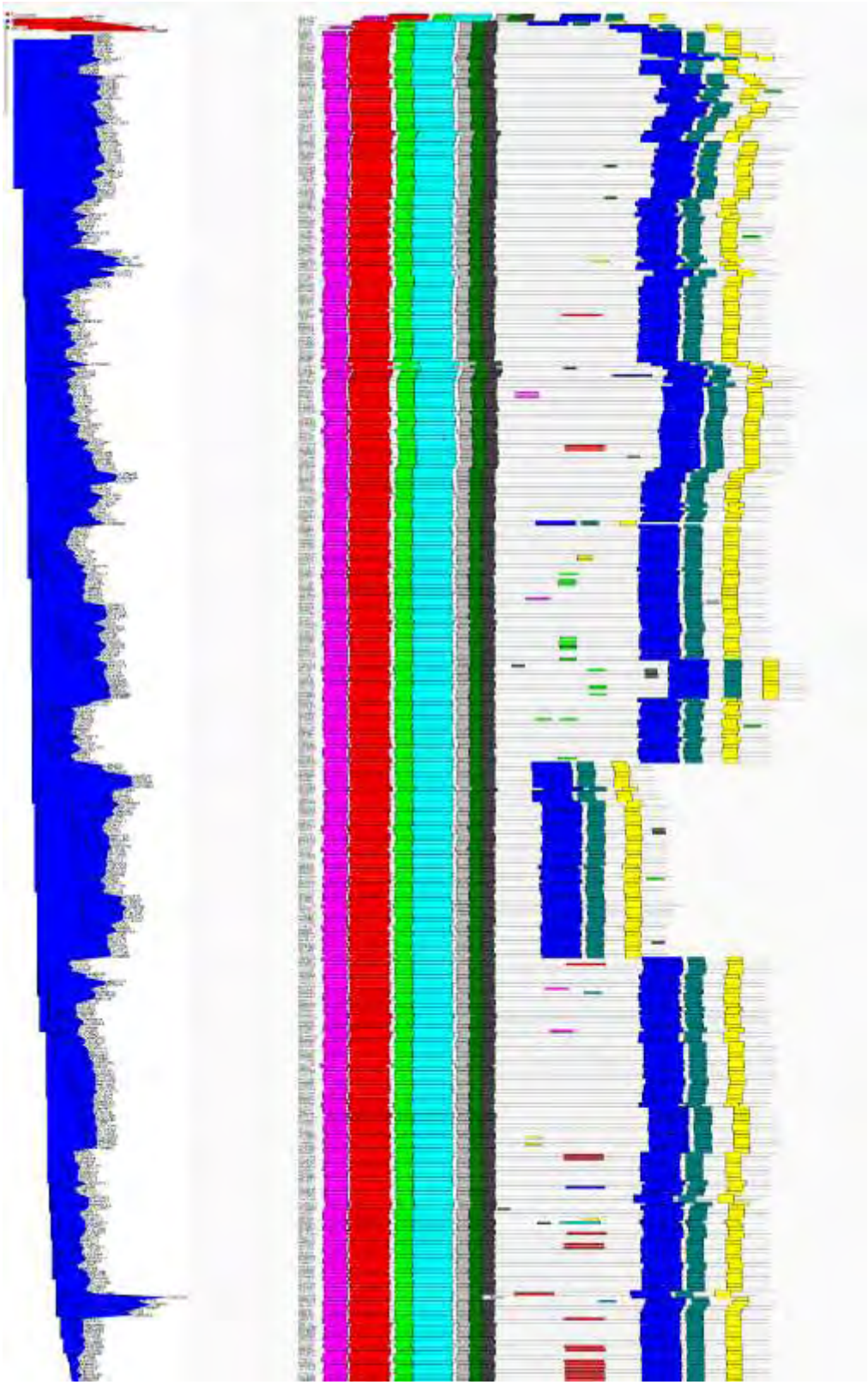
Η παραπάνω ανάλυση είχε ως αποτέλεσμα την δημιουργία 33 διαφορετικών φυλογενετικών δέντρων, για 22 ένζυμα AARSs. Αυτό συνέβη, γιατί ένα ένζυμο AARS μπορεί να χρειαστεί να σπάσει σε περισσότερες φυλογενετικές υποομάδες, έτσι για κάθε νέα υποομάδα υπάρχει ένα νέο φυλογενετικό δέντρο και μία νέα αρχιτεκτονική για τα μοτίβα. Τα μέλη κάθε βασίλειου (βακτήρια, αρχαία, ευκαρυωτικά) διακρίνονται με διαφορετικό χρώμα.

Για κάθε ένα από τα 10 μοτίβα, από κάθε φυλογενετική ομάδα ή υποομάδα, εξάγαμε τις ακολουθίες των μοτίβων από όλα τα μέλη της συγκεκριμένης υποομάδα και κάναμε πολλαπλή στοίχιση χρησιμοποιώντας τον αλγόριθμο MUSCLE (Edgar 2004). Στη συνέχεια, κάθε μία από αυτές τις στοιχισμένες ακολουθίες, μετατράπηκε σε profile - HMM, χρησιμοποιώντας το πρόγραμμα HMMER (Eddy 2011). Συνολικά δημιουργήθηκαν 330 profile - HMMs, που αντιστοιχούν με τα MEME-motifs. Στη συνέχεια αυτά τα 330 profile - HMMs τα χρησιμοποιήσαμε για να σαρώσουμε τις πρωτεΐνες που ήταν επισημασμένες ως AARS από την βάση δεδομένων Swissprot. Το αρχείο στη συνέχεια φιλτραρίστηκε με την χρήση ενός Perl-script που δημιουργήσαμε στο εργαστήριο, το οποίο αντιστοιχεί μία πρωτεΐνη σε ένα συγκεκριμένο ένζυμο AARS, χρησιμοποιώντας τον κανόνα της πλειοψηφίας. Πιο συγκεκριμένα, αν δύο HMMs χτυπούν στην ίδια περιοχή και το ένα αλληλεπικαλύπτει το άλλο, τότε ο αλγόριθμος του perl script θα διαλέξει το μοτίβο που έχει το καλύτερο bit score. Με αυτόν τον τρόπο, HMMs από διαφορετικά ένζυμα μπορούν να ανταγωνίζονται μεταξύ τους για την ίδια πρωτεϊνική περιοχή. Τέλος, όταν ο αλγόριθμος αναλύσει όλα τα HMMs που "χτύπησαν" την πρωτεΐνη, θα βρει ποια είναι αυτά με το καλύτερο bit score και ταυτόχρονα δεν αλληλεπικαλύπτονται. Ανάλογα με το πόσα HMMs από ένα συγκεκριμένο ένζυμο "χτύπησαν" μία πρωτεΐνη, τότε ο αλγόριθμος θα υποθέσει ότι η πρωτεΐνη πιθανόν να ανήκει στο συγκεκριμένο ένζυμο, στο οποίο ανήκει η πλειοψηφία των μοτίβων που χτύπησαν την πρωτεΐνη. Π.χ αν μία πρωτεΐνη χτυπηθεί από 5 HMMs που ανήκουν στο ένζυμο της ProRS και 2 HMMs που ανήκουν στο ένζυμο της CysRS, τότε ο αλγόριθμος θα υποθέσει ότι η συγκεκριμένη πρωτεΐνη ανήκει στην οικογένεια της ProRS.

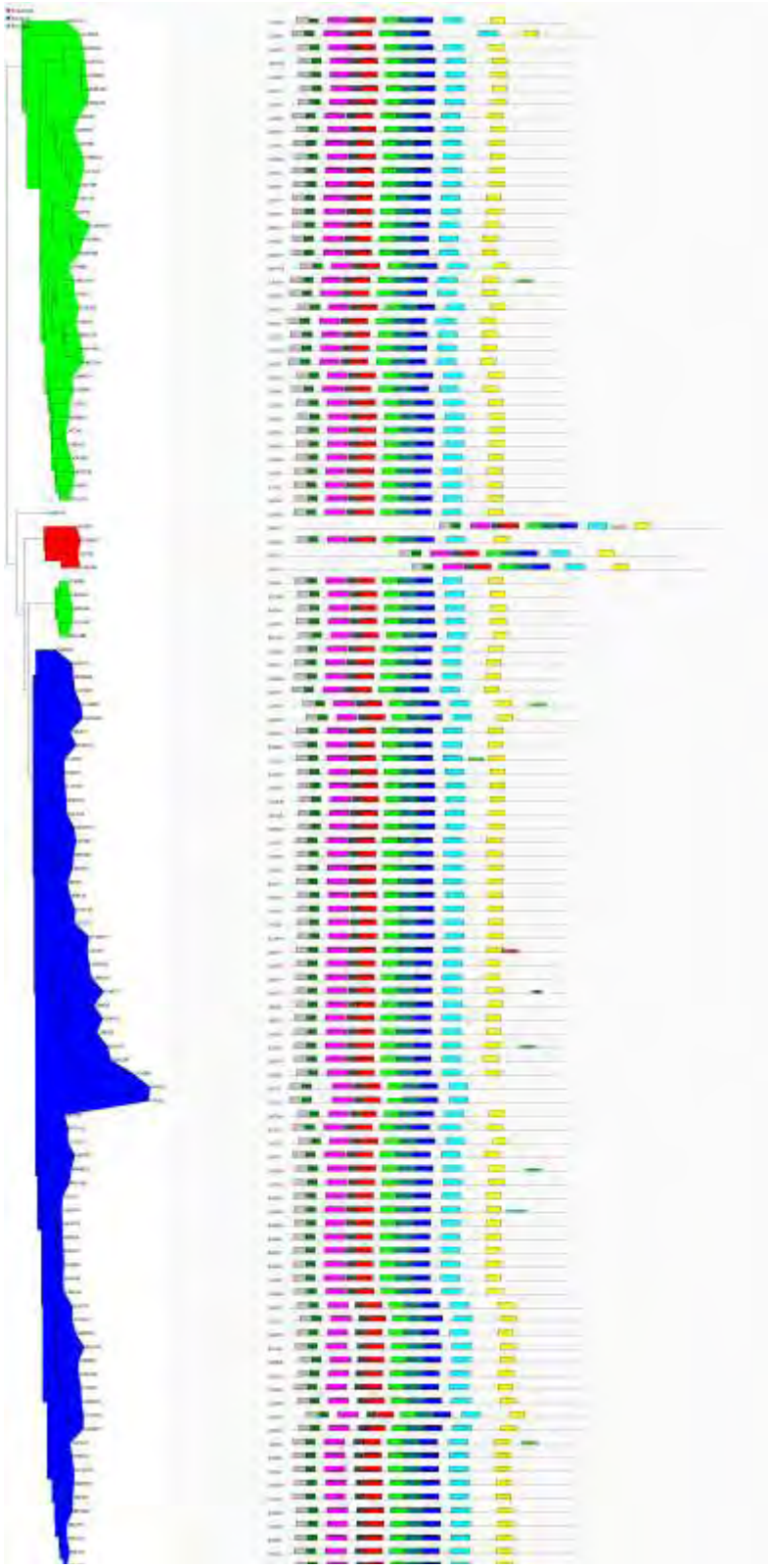
Εφαρμόζοντας ένα ελάχιστο όριο τα 2 ή τα 5 μοτίβα και χρησιμοποιώντας ως φίλτρο τον κανόνα της πλειοψηφίας, δοκιμάσαμε τα HMMs σε ήδη γνωστές ακολουθίες από τη βάση δεδομένων SwissProt. Τα HMMs κατάφεραν να εντοπίσουν σωστά το 99,9% των πρωτεϊνών, στη πρώτη περίπτωση (cut-off 2 motifs), ενώ στη δεύτερη περίπτωση το 99,8%(cut-off 5 motifs) (βλ. διάγραμμα 3.1). Το συγκεκριμένο πρωτόκολλο είχε μεγάλη ακρίβεια. Παρόλο που το όριο των τουλάχιστον 5 HMMs είναι πολύ αυστηρό, η απόδοση του είναι παραπάνω από ικανοποιητική.



Εικόνα 3.4: Φυλογενετικό δέντρο της ProRS (500 ακολουθίες) πριν την διάσπαση του σε δύο υποομάδες, σε συνδυασμό με την αρχιτεκτονική των MEME-motifs



Εικόνα 3.5: Φυλογενετικό δέντρο της υποομάδας 1 της ProRS (386 ακολουθίες), σε συνδυασμό με την αρχιτεκτονική των MEME-motifs



Εικόνα 3.6: Φυλογενετικό δέντρο της υποομάδας 2 της ProRS (114 ακολουθίες), σε συνδυασμό με την αρχιτεκτονική των MEME-motifs

Επιπλέον, αυτή η ιεραρχική προσέγγιση που ακολουθήσαμε με τα MEME-motifs, είναι πολύ ανθεκτική ακόμα και σε περιπτώσεις που έχουμε αναδιάταξη επικρατειών (domain rearrangement), ή ακόμα και σε περιπτώσεις που έχουμε την εισαγωγή μεγάλων εισδοχών, που πολλές φορές έχουν ως αποτέλεσμα την μειωμένη ανίχνευση από άλλες τεχνικές. Η αρχή που εφαρμόζεται στην ανάλυση αυτή, είναι παρόμοια με αυτή της βάσης δεδομένων PRINTS (Attwood et al. 2012).

Παρόλα αυτά, ένας μικρός αριθμός απομακρυσμένων παραλόγων, όπως είναι το HisZ, δε μπόρεσαν να ανιχνευθούν με την τεχνική των MEME-motifs. Για να ξεπεράσουμε αυτό το πρόβλημα, δημιουργήσαμε άλλο ένα σύνολο από HMMs, τα οποία όμως βασίζονταν σε ολόκληρη την καταλυτική επικράτεια (μερικές φορές συμπεριλαμβανομένων και μιας εισδοχής). Επίσης, για την κατασκευή των συγκεκριμένων HMMs χρησιμοποιήσαμε πάλι πρωτεΐνες από τη Swissprot, που να μην είχαν βαθμό ομοιότητας μεταξύ τους, πάνω από 90% και ήταν σχολιασμένες (annotated) ως AARSs. Ακόμη, λάβαμε υπόψη τις εξελικτικές υποομάδες μέσα σε κάθε AARS. Τέλος, δημιουργήσαμε και HMMs για τις περιοχές που είχαν κάποια γνωστή λειτουργία π.χ. τομέας δέσμευσης tRNA.

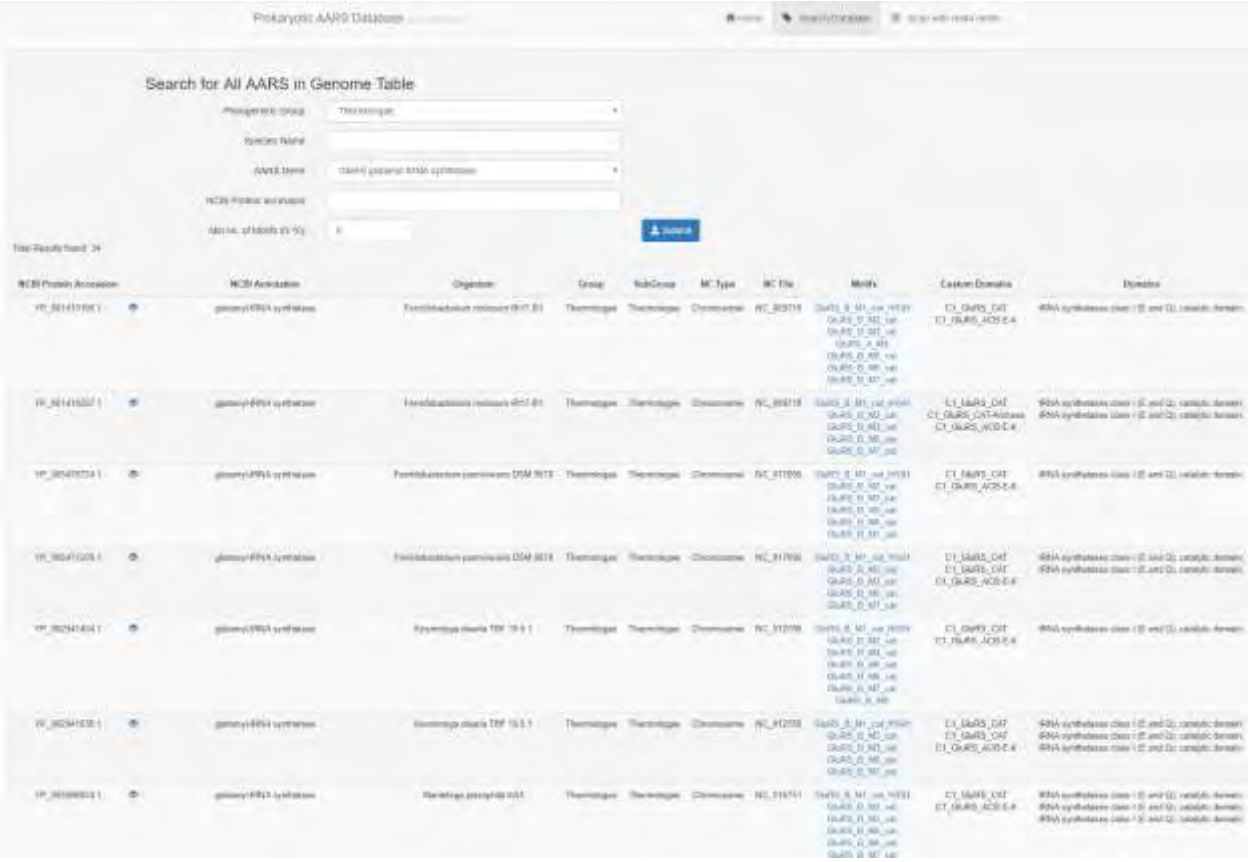
Έπειτα για την αξιολόγηση των μοντέλων, χρησιμοποιήσαμε πάλι πρωτεΐνες με γνωστή λειτουργία από τη βάση Swissprot, και δοκιμάσαμε τα συγκεκριμένα HMMs για το αν εντοπίζουν σωστά τις αντίστοιχες πρωτεΐνες. Το ποσοστό της θετικής αναγνώρισης άγγιξε το 99,9%. Επιπλέον, απομακρυσμένα παράλογα τα οποίο δεν μπορούσαν να εντοπιστούν με την τεχνική των MEME-motifs, πλέον εντοπίζονταν. Μόνο 9 από τις 883 πρωτεΐνες που ήταν σημασμένες ως GluRS, αναγνωρίστηκαν λανθασμένα ως Gln-GltxRS. Για τους λόγους που προαναφέραμε, επιλέξαμε να χρησιμοποιήσουμε αυτά τα 2 σύνολα από HMMs (MEME-motifs και catalytic-domain-motifs) στο υπολογιστικό μας εργαλείο και στην βιοπληροφορική ανάλυση που κάναμε στη συνέχεια.

3.3 Κατασκευή βάσης δεδομένων και ανάπτυξη διαδικτυακής εφαρμογής-εργαλείου

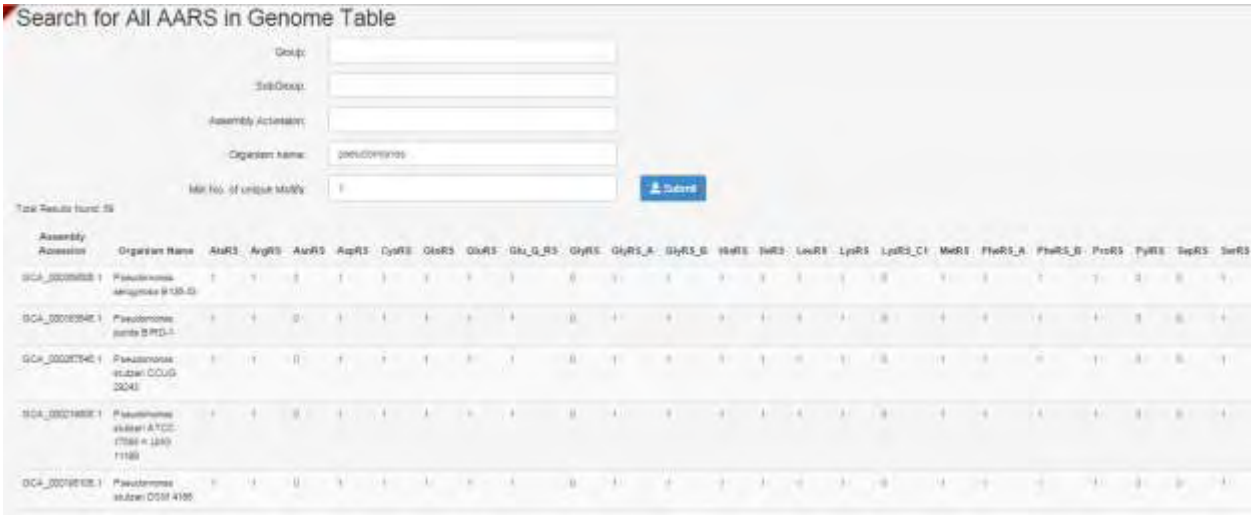
Προκειμένου να διερευνήσουμε το εξελικτικό προφίλ των AARSs στους προκαρυωτικούς οργανισμούς, το νέο υπολογιστικό εργαλείο που αναπτύξαμε, χρησιμοποιήθηκε για την σάρωση-ανάλυση ~8.000.000 πρωτεϊνών από 2588 προκαρυωτικά πρωτεώματα, τα οποίο βρέθηκαν στη βάση δεδομένων του NCBI. Οι συγκεκριμένες πρωτεΐνες αρχικά σαρώθηκαν από το σύνολο των HMMs που δημιουργήθηκαν από τη καταλυτική επικράτεια των AARSs, τα οποία είναι εξαιρετικά ευαίσθητα αλλά όχι τόσο εξειδικευμένα, ενώ σε έναν δεύτερο γύρο, σαρώθηκαν από τα πιο εξειδικευμένα MEME-motifs. Συνολικά, από τις δύο σαρώσεις εντοπίστηκαν 56.469 και 52.595 πρωτεΐνες αντιστοίχως. Στην περίπτωση των MEME-motifs, όταν χρησιμοποιήσαμε το πιο αυστηρό όριο των τουλάχιστον 5 motifs ανά πρωτεΐνη, τότε ο αριθμός των πρωτεϊνών που προέκυψαν από τη σάρωση, μειώθηκε στις 49.788 (περισσότερες πληροφορίες στο Supplementary file 1: NCBI_protein_info).

Τα αποτελέσματα από τη μεγάλης κλίμακας σάρωση των προκαρυωτικών πρωτεωμάτων από το NCBI, έχουν οργανωθεί και αποθηκευτεί σε μία βάση δεδομένων MySQL(βλ εικόνες 3.7-3.10), την οποία μπορείτε να επισκεφτείτε στην ιστοσελίδα http://bioinf.bio.uth.gr. Ο εκάστοτε χρήστης μπορεί να αναζητήσει στη βάση δεδομένων πρωτεΐνες οι οποίες εντοπίστηκαν από την παραπάνω ανάλυση, χρησιμοποιώντας ως όριο τον αριθμό των MEME-motifs τον οποίο ο ίδιος

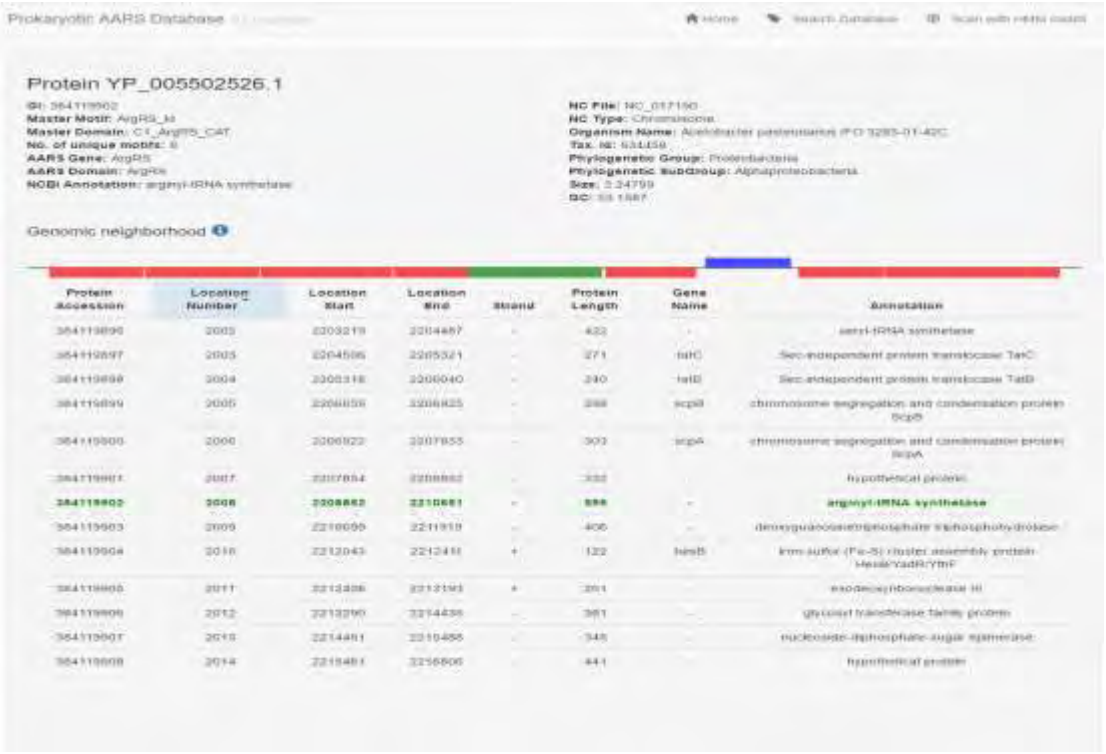
επιθυμεί. Μπορεί επίσης να χρησιμοποιήσει και άλλου είδους φίλτρα, όπως π.χ. αν θέλει να δει πρωτεΐνες



Εικόνα 3.7: Αποτελέσματα στη βάση δεδομένων (http://bioinf.bio.uth.gr/aars) μετά την επιλογή συγκεκριμένης φυλογενετικής ομάδας και συγκεκριμένης AARS



Εικόνα 3.8: Συγκεντρωτικός πίνακας από τη βάση δεδομένων (http://bioinf.bio.uth.gr/aars) για τον συνολικό αριθμό AARSs μετά την επιλογή συγκεκριμένου οργανισμού



Εικόνα 3.9: Απεικόνιση των 10 πιο κοντινών γονιδίων στον γονιδιακό τόπο της ArgRS στον οργανισμό *Acetobacter pasteurianus*, στη βάση δεδομένων (http://bioinf.bio.uth.gr/aars)

Εικόνα 3.10: Απεικόνιση των HMMs με το υψηλότερο bitscore που χτύπησαν την ArgRS στον οργανισμό *Acetobacter pasteurianus*, στη βάση δεδομένων (http://bioinf.bio.uth.gr/aars)

που ανήκουν σε μία συγκεκριμένη οικογένεια AARS ή και σε ένα συγκεκριμένο φυλογενετικό γκρουπ. Επίσης, τα αποτελέσματα οργανώνονται σε έναν συγκεντρωτικό πίνακα, στον οποίο φαίνονται για κάθε γονιδίωμα-πρωτέωμα ο αριθμός των AARS που έχουν εντοπιστεί από την ανάλυσή μας. Εμφανίζεται επίσης, η επιλογή αν θέλει ο χρήστης να δει συγκεκριμένες πληροφορίες για μία συγκεκριμένη AARS. Κάποιες από αυτές τις πληροφορίες είναι τα γειτονικά γονίδια, σχολιασμός από το NCBI, γραφιστική απεικόνιση των γειτονικών γονιδίων καθώς επίσης και της αρχιτεκτονικής των μοτίβων και από τα 2 σύνολα που δημιουργήσαμε (MEME-motifs, catalytic domain-motifs), καθώς και HMMs από τη βάση δεδομένων PFAM. Με ότι γνωρίζουμε μέχρι σήμερα, η μόνη άλλη βάση δεδομένων που έχει φτιαχτεί αποκλειστικά για AARSs, δημιουργήθηκε το 2001 (Szymanski, Deniziak, and Barciszewski 2001).

Στη βάση δεδομένων, ενσωματώθηκε επίσης ένα διαδικτυακό εργαλείο ανίχνευσης motifs/domains για AARSs. Ένας χρήστης μπορεί να ανεβάσει ένα αρχείο με μορφή FASTA, που περιέχει από μία πρωτεΐνη έως και ένα ολόκληρο πρωτέωμα (ανώτατο όριο μεγέθους αρχείου είναι τα 20mb). Μέσα σε μικρό χρονικό διάστημα, το αρχείο θα σαρωθεί από τα HMMs που δημιουργήσαμε και θα εμφανιστούν τα αποτελέσματα. Ένα τέτοιο εργαλείο είναι χρήσιμο όχι μόνο για την σήμανση μίας πρωτεΐνης ως AARS, αλλά επίσης και στον εντοπισμό θραυσμάτων από παράλογα, ή καθώς επίσης για τον εντοπισμό γονιδιωμάτων που πιθανόν έχουν κάποιο εναλλακτικό βιοχημικό μονοπάτι, παρατηρώντας για το αν υπάρχει κάποια έλλειψη ή διπλασιασμός σε AARS. Πρόσφατα υπήρξε ένα έντονο ενδιαφέρον για τον εντοπισμό ανθρώπινων θραυσμάτων από AARSs, ακόμα και θραύσματα που δεν είχαν την καταλυτική λειτουργία των AARSs, καθώς φαίνεται να εμπλέκονται σε διάφορες λειτουργίες του οργανισμού. Τέτοια παράλογα έχουν βρεθεί τα τελευταία χρόνια σχεδόν σε όλους τους οργανισμούς και φαίνεται να εμπλέκονται σε διάφορες βιολογικές λειτουργίες. Οι νέες αυτές λειτουργίες από τα θραύσματα των AARSs, ή ακόμα και πρωτεϊνών που προέρχονται από διπλασιασμένα γονίδια των AARSs, σχετίζονται κυρίως με λειτουργίες που δεν έχουν σχέση με την πρωτεϊνοσύνθεση. Τέτοιες λειτουργίες μπορεί να είναι η επεξεργασία και η αντίσταση απέναντι σε αντιβιοτικά, έως και μοριακούς κεντρικούς κόμβους μέσα σε μονοπάτια σηματοδότησης που ρυθμίζουν την ογκογένεση στους ανθρώπους (Lo et al. 2014; L.-T. Guo et al. 2014; M. Guo et al. 2009; Han et al. 2012). Η αρχιτεκτονική των domains στις AARSs φαίνεται να έχει μεγάλη σημασία, για την ύπαρξη ή όχι άλλων λειτουργιών. Ωστόσο, το ακριβές ρεπερτόριο αυτών των λειτουργιών από αυτά τα domains παραμένει ένα μεγάλο μυστήριο, ειδικά όταν αυτά δρουν in-trans. Είναι πολύ ενδιαφέρον ότι πρόσφατα αποδείχθηκε η άμεση συμμετοχή τους σε πολλά μεταβολικά δίκτυα τα οποία ελέγχουν πολλές κυτταρικές διεργασίες (M. Guo, Schimmel, and Yang 2010).

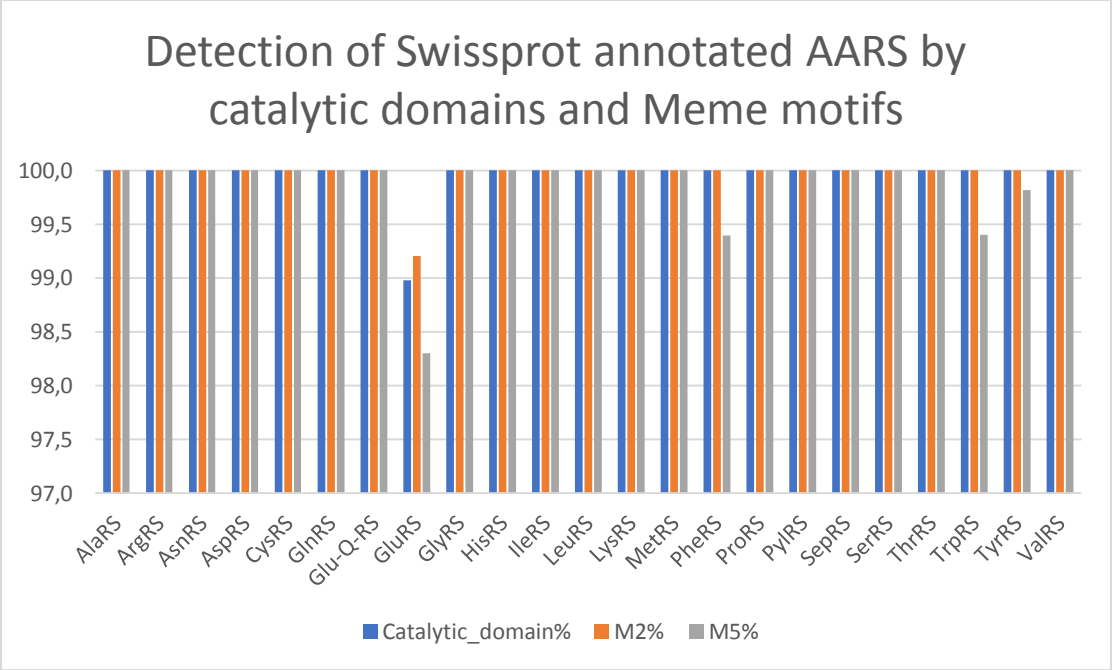
3.4 Επισκόπηση του εξελικτικού προφίλ των προκαρυωτικών AARSs

Αναλύοντας 2588 προκαρυωτικά πρωτεώματα, ήταν δυνατό να αποκτήσουμε ένα ολοκληρωμένο εξελικτικό προφίλ για αυτήν την πολύπλοκη υπεροικογένεια, που διαδραματίζει κεντρικό ρόλο στην εξασφάλιση της πιστής μετάφρασης του γενετικού κώδικα, καθώς επίσης και στη φυσική αντίσταση απέναντι σε τοξίνες. Για πολλές από τις επόμενες αναλύσεις εφαρμόσαμε διαφορετικά κριτήρια. Στο πρώτο και αυστηρότερο κριτήριο, ένα γονίδιο θεωρήθηκε ως μέλος της υπεροικογένειας των AARSs, αν εμφάνιζε τουλάχιστον 5 μοτίβα. Η ανάλυση η οποία προηγήθηκε, όταν κάναμε την αξιολόγηση των HMMs, έδειξε ότι χρησιμοποιώντας ένα όριο των 5 motifs, μπορούμε να εντοπίσουμε το 99,8% των πρωτεϊνών. Σε ένα δεύτερο και πιο χαλαρό κριτήριο, μία πρωτεΐνη θεωρήθηκε ως μέλος της οικογένειας AARSs, όταν την

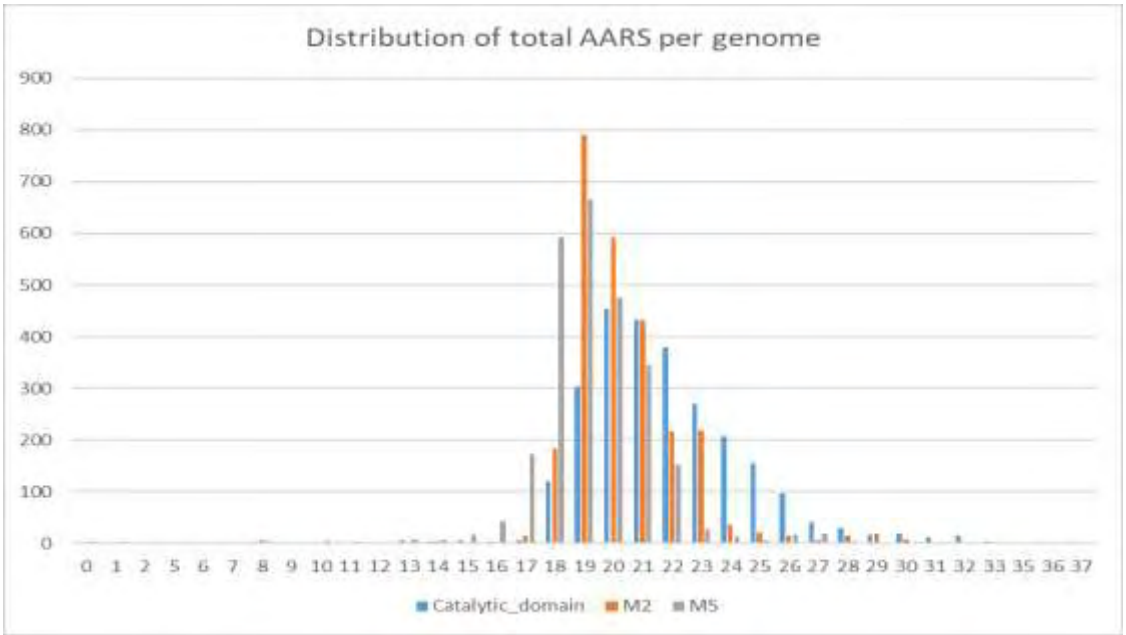
χτυπούσαν τουλάχιστον 2 motifs από το ίδιο ένζυμο. Τέλος σε ένα ακόμα πιο χαλαρό κριτήριο, για τον λόγο ότι δεν μπορούσαμε να εντοπίσουμε παράλογα που ήταν αρκετά απομακρυσμένα, όπως π.χ. το παράλογο HisZ, με τα MEME-motifs, μία πρωτεΐνη θεωρήθηκε ως AARS, όταν την "χτυπούσε" ένα catalytic-HMM. Με Αυτόν τον τρόπο ακόμα και τα πολύ απομακρυσμένα παράλογα μπορούσαν να εντοπιστούν. Τα αποτελέσματα από τα MEME-motifs καθώς και από τα catalytic-HMMs, παρουσιάζονται στο αρχείο supplementary file 1. Εδώ θα πρέπει να σημειώσουμε ότι τα catalytic-HMMs έχουν περίπου την ίδια συμπεριφορά - απόδοση με τα MEME-motifs όταν χρησιμοποιούμε ως όριο τα 2 motifs, εκτός από τις περιπτώσεις των ProRs και HisRS. Στο ίδιο συμπέρασμα σχετικά με την μειωμένη απόδοση των MEME-motifs, αλλά μόνο για τα απομακρυσμένα ομόλογα των ProRS και HisRS, φαίνεται και από την ανάλυση των πρωτεωμάτων από το NCBI (δείτε στον πίνακα 1). Στην περίπτωση του HisRS το καταλυτικό κομμάτι μπορεί να εντοπίσει τα παράλογα του HisZ, τα οποία είναι γνωστό ότι δεν έχουν την λειτουργία της αμινοακυλίωσης. (Sissler et al. 1999). Σε αυτή τη περίπτωση τα MEME-motifs δεν μπόρεσαν να εντοπίσουν την πλειοψηφία αυτών των παραλόγων. Το NCBI έχει σχολιάσει την πλειοψηφία των παραλόγων του HisZ ως ATP phosphoribosyltransferases.

Στην περίπτωση της ProRS, το catalytic-HMM μπορεί να εντοπίσει τις πρωτεΐνες ybaK/ebsC/ProX, οι οποίες αποτελούν την επικράτεια διαμόρφωσης για μία υποομάδα των ProRS, η οποία δρα in-trans, ενώ τα MEME-motifs δεν μπορούσαν να εντοπίσουν την πλειοψηφία αυτών των πρωτεϊνών, πιθανόν επειδή τα μοτίβα ήταν σχεδιασμένα για το καταλυτικό domain αυτών των πρωτεϊνών. Οπτική εξέταση των σχολιασμών από το NCBI για τις συγκεκριμένες πρωτεΐνες που εντοπίστηκαν με τα παραπάνω 3 κριτήρια, έδειξε ότι το 93,2%(52662/56531), το 98.7%(51975/52653) και το 99.1% (49380/49846) αυτών, ήταν σχολιασμένες ως AARSs ή ως κάποιο γνωστό παράλογο των AARSs. Αυτό αποτελεί μία ισχυρή ένδειξη ότι η μέθοδος με τα MEME-motifs δεν είναι μόνο πολύ ειδική αλλά και πολύ ευαίσθητη.

Αν και είναι κοινή πεποίθηση ότι τα περισσότερα προκαρυωτικά γονιδιώματα έχουν 20 γονίδια AARSs, η ανάλυση μεγάλης κλίμακας που κάναμε έδειξε σαφώς ότι αυτό δε συμβαίνει. Αντίθετα, ο συχνότερος αριθμός γονιδίων AARSs σε ένα προκαρυωτικό γονιδίωμα ήταν 19, αν και υπάρχουν επίσης πολλά γονιδιώματα με 20 AARSs, ωστόσο η συχνότητα εμφάνισής τους είναι μικρότερη από αυτή των 19. Μία κατανομή του αριθμού των γονιδίων AARSs ανά γονιδίωμα παρουσιάζεται στην εικόνα 3.4. Αυτός ο μειωμένος αριθμός οφείλεται κυρίως στην συχνή απουσία GlnRS (62% των σαρωμένων πρωτεωμάτων, βλ. Πίνακα 1). Αυτό το εύρημα είναι σύμφωνο με ένα εξελικτικό σενάριο όπου η GlnRS εμφανίστηκε αργότερα στην εξέλιξη της ευκαρυωτικής γραμμή, μέσω του διπλασιασμού της GluRS, και μεταφέρθηκε στα βακτήρια μέσω οριζόντιας γονιδιακής μεταφοράς (Lamour et al. 1994; Brown and Doolittle 1999; Koonin, Makarova, and Aravind 2001). Σε όλους τους προκαρυώτες που λείπει το γονίδιο της GlnRS, η ενσωμάτωση του Gln διαμεσολαβείτε δια μέσου ενός μονοπατιού τρανσαμίδωσης (Curnow et al. 1997; Tumbula et al. 2000). Τα Gamma-Proteobacteria θεωρούνται ως το σημείο εισόδου της GlnRS στον βακτηριακό κόσμο. Με την ανάλυση της παρούσας κατανομής GlnRS στα γονιδιώματα που έχουν αλληλουχηθεί και συγκρίνοντας τα με ένα σενάριο που προέρχεται καθαρά από τύχη, η πιο εμπλουτισμένη φυλογενετική σειρά για την GlnRS, είναι όντως τα Proteobacteria (hypergeometric test: $1.5e^{-169}$). Επίσης, θα πρέπει να αναφέρουμε ό,τι δεν ανιχνεύτηκε Gln-GlxRS στα γονιδιώματα των αρχαίων που αναλύθηκαν.



Γράφημα 3.1: Αξιολόγηση των HMMs σε σχολιασμένες ακολουθίες από τη βάση δεδομένων Uniprot/Swissprot. Μπλε μπάρες: catalytic-domain HMM, Πορτοκαλί μπάρες: MEME-motifs με όριο τα 2 μοτίβα. Γκρι μπάρες: MEME-motifs με όριο τα 5 μοτίβα



Γράφημα 3.2: Κατανομή των συνολικών AARSs ανά γονιδίωμα (συμπεριλαμβανομένων των SepRS και PylRS). Μπλε μπάρες: ανίχνευση βασίστηκε στο catalytic-HMM, Πορτοκαλί μπάρες: ανίχνευση βασίστηκε στα MEME-motifs με όριο τα 2 μοτίβα, Γκρι μπάρες: ανίχνευση βασίστηκε στα MEME-motifs με όριο τα 5 μοτίβα

Σύμφωνα με τον ίδιο συλλογισμό, είναι πολύ πιθανόν και η AsnRS να ακολούθησε παρόμοια εξελικτική πορεία με τη GlnRS, αφού λείπει στο 47%-48% των προκαρυωτικών γονιδιωμάτων που αναλύθηκαν (λείπει και από το 69%-74% των γονιδιωμάτων των αρχαίων) (βλ. Πίνακα 1 και supplementary file 1: Genomes_tables). Σε αυτή τη περίπτωση, αναλύοντας την παρούσα κατανομή της AsnRS στα 2588 γονιδιώματα και συγκρίνοντας την με ένα τυχαίο σενάριο, η πιο εμπλουτισμένη φυλογενετική σειρά για την AsnRS προέρχεται από τα Firmicutes (hypergeometric test: $2.7e^{-116}$).

Όσον αφορά τις TrpRS και TyrRS, έχει προταθεί (βασισμένη σε δομικές στοιχίσεις) ότι το πιο αρχαίο γονίδιο είναι αυτό της TyrRS, ενώ το γονίδιο της TrpRS προήλθε από διπλασιασμό στην γενεαλογική σειρά των αρχαίων και αργότερα μεταφέρθηκε στα βακτήρια με ορίζοντια γονιδιακή μεταφορά (Dong et al. 2010). Όντως αυτό το σενάριο επίσης υποστηρίζεται από το εξελικτικό προφίλ που έχουμε από τα δεδομένα μας, αλλά μόνο στη περίπτωση που χρησιμοποιούμε ως φίλτρο τα 5 MEME-motifs., όπου το 16% των προκαρυωτικών οργανισμών έχουν χάσει την TrpRS, ενώ για την TyrRS το ποσοστό αυτό είναι 1,5%., Ενδιαφέρον έχει το γεγονός, επίσης όταν χρησιμοποιούμε ως φίλτρο τα 5 MEME-motifs, η HisRS λείπει από το 9% των συνολικών γονιδιωμάτων που μελετήθηκαν. Η ιστιδίνη επίσης θεωρείται "νεό" αμινοξύ. Η σειρά με την οποία τα αμινοξέα εισήχθησαν στον γενετικό κώδικα, όπως έχει προταθεί από την θεωρία της συν-εξέλιξης του κώδικα (co-evolution code theory) και βρίσκεται σε συμφωνία με το εξελικτικό προφίλ των ενζύμων της υπεροικογένειας AARSs, ωστόσο μόνο για όταν χρησιμοποιούμε το φίλτρο με το όριο των 5 MEME-motifs (βλ. Πίνακα 1) (Francklyn 2003; Wong 1975)

Συγκεκριμένα, φαίνεται να υπάρχει κάποια ευρεία συσχετίση μεταξύ της εξελικτικής ηλικίας και της αντιδραστικότητας των αμινοξέων, καθώς και με την συχνότητα διαγραφής/διπλασιασμού της αντίστοιχης AARS. Για τα αμινοξέα Arg, Cys, Met, Gln, Tyr, Trp, His, Asn, Phe, έχει προταθεί ότι εισήχθησαν στον γενετικό κώδικα αργότερα (Higgs and Pudritz 2009). Χρησιμοποιώντας το αρκετά αυστηρό φίλτρο των τουλάχιστον 5 MEME-motifs, οι AARSs που ενεργοποιούν αυτά τα "νέα" αμινοξέα (με την εξαίρεση των Cys και Tyr), επίσης φαίνεται να "λείπουν" από τα προκαρυωτικά γονιδιώματα πιο συχνά σε σχέση με την ομάδα των AARSs που "ενεργοποιούν" τα "παλαιότερα αμινοξέα. Επίσης έχει πολύ ενδιαφέρον, ό,τι πολλές από τις AARSs που ενεργοποιούν "παλαιότερα" και υδρόφοβα/όχι δραστικά αμινοξέα (Leu, Ile, Val, Ala, Gly), φαίνεται να είναι εξαιρετικά σταθερές σύμφωνα με το εξελικτικό προφίλ που έχουμε (μόνο ένα αντίγραφο από το κάθε γονίδιο σε κάθε γονιδίωμα).

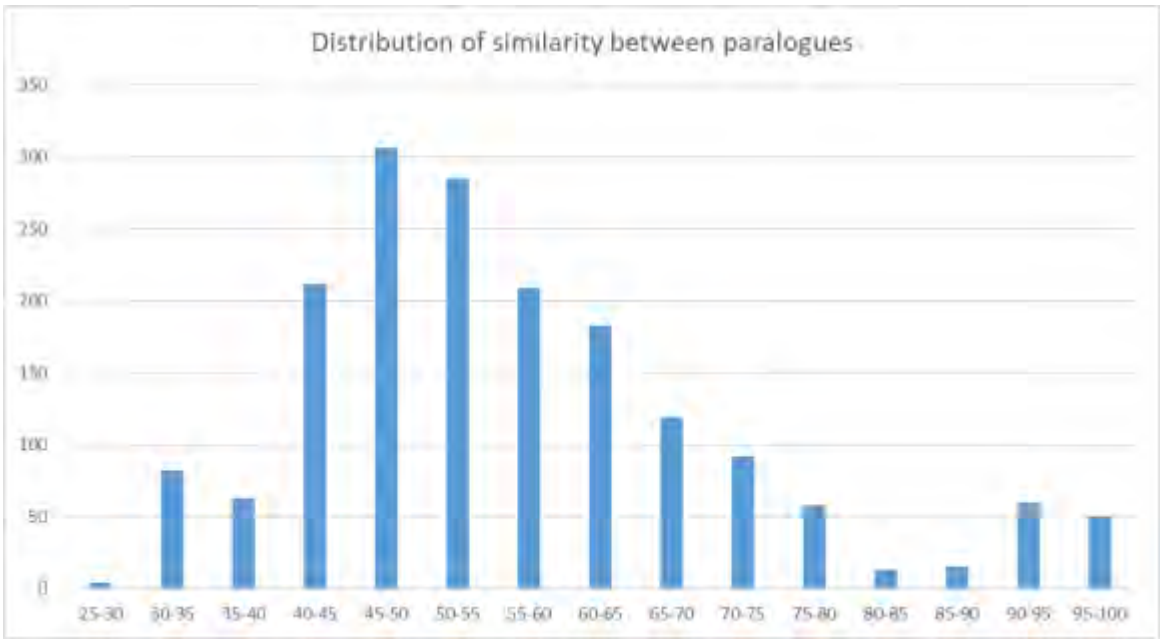
3.5 Η παρουσία των παραλόγων είναι πολύ συχνή

Για να διασφαλιστεί μία συντηρητική προσέγγιση, τα αποτελέσματα από αυτό το τμήμα βασίζονται στο φίλτρο με το όριο των 5 MEME-motifs, εκτός αν αναφέρεται κάτι συγκεκριμένο. Η συγκεκριμένη ανάλυση αποκάλυψε ότι στο 40%-61% (για 5 MEME-motifs και για 2 MEME-motifs) των πρωτεωμάτων που σαρώθηκαν, ανιχνεύθηκε τουλάχιστον 1 παράλογο AARS.

Ο μεγαλύτερος αριθμός γονιδίων AARSs που ανιχνεύτηκαν σε ένα γονιδίωμα ήταν 29 (*Ketasatospora setae*) και 28 (*Bacillus cereus*). Γενικά, το γένος *Bacillus* είχε αρκετά γονιδιώματα με υψηλό αριθμό γονιδίων/παραλόγων AARSs. Το 22% (581/2588) και το 3%

(85/2588) των γονιδιωμάτων που λήφθηκαν από το NCBI και σαρώθηκαν είχαν ≥ 21 και ≥23 AARSs αντιστοίχως. Επιπλέον η ομάδα των γονιδιωμάτων με ≥23 AARSs ήταν εμπλουτισμένη για Firmicutes (64%, αντί για 21% που ήταν το υπόβαθρο). Από την άλλη μεριά, η *Nasuia deltocephalinicola str. NAS-ALF* ήταν το γονιδίωμα με τον μικρότερο αριθμό AARSs (συγκεκριμένα ανιχνεύτηκε μόλις ένα γονίδιο AARS, της LeuRS, με φίλτρο τα 5 MEME-motifs, ενώ με φίλτρο τα 2 MEME-motifs ανιχνεύτηκαν 5 γονίδια AARS). Αυτό πιθανώς να οφείλεται στον ενδοσυμβιωτικό τρόπο ζωής, του μικρότερου προκαρυωτικού γονιδιώματος που έχει αλληλουχηθεί (διαθέτει συνολικά 169 γονίδια) (Bennett and Moran 2013). Αν κάποιος θεωρεί ότι το 20 είναι ο αναμενόμενος αριθμός των διαφορετικών γονιδίων AARSs που θα περίμενε να βρει σε ένα γονιδίωμα, η ανάλυση μας δείχνει ότι 59% (1531/2588) από τα γονιδιώματα που αναλύσαμε, είχαν λιγότερα από 20 γονίδια AARSs. Σε σχέση με τα παράλογα, η πιο ακραία περίπτωση που εντοπίσαμε, ήταν της *Kitasatospora setae*, η οποία διέθετε 4 παράλογα για το γονίδιο της SerRS. Κάθε ένα από τα 4 παράλογα, είχε εντοπιστεί τουλάχιστον από 7 MEME-motifs και είχαν όλα σχολιαστεί από το NCBI ως SerRS. Ομοίως, υπήρχαν 3 αλληλόμορφα του ίδιου γονιδίου για τα συγκεκριμένα AARSs (spRS, CysRS, GluRS, LeuRS, LysRS, ThrRS and ValRS) σε διαφορετικά είδη.

Η LysRS είναι μία πολύ ενδιαφέρουσα περίπτωση, καθώς είναι το μοναδικό AARS που έχει εντοπιστεί να ανήκει είτε στη κλάση I είτε στη κλάση II (Ibba et al. 1997). Συνολικά 2143 γονιδιώματα είχαν μόνο LysRS που να ανήκει στη κλάση II, ενώ 377 γονιδιώματα είχαν LysRS που άνηκε στην κλάση I. Μόνο 39 γονιδιώματα (κυρίως στελέχη από *Bacillus thuringiensis* και *Bacillus cereus*, καθώς και μερικά μέλη από το γένος Streptomyces) είχαν γονίδια και από τις 2 κλάσεις. Είναι πιο κοινό να υπάρχουν παράλογα ενός γονιδίου και να προέρχονται από την ίδια κλάση (218 γονιδιώματα με παράλογα που ανήκαν στην κλάση II και μόλις μία περίπτωση με παράλογο από τη κλάση I) αντί να έχουμε γονίδια και από τις δύο κλάσεις (39 περιπτώσεις).



Γράφημα 3.2: Κατανομή του ποσοστού ομοιότητας μεταξύ των ζευγαριών των παραλόγων

			Absent			One			Duplicates				
E/L	**E.G.**	**AARSS**	**CD**	**M2**	**M5**	**CD**	**M2**	**M5**	**CD**	**M2**	**M5**	**T.b.t**	**R**
L	Ia	ArgRS	1.2	1.3	4.9	94.7	95.5	92.1	4.0	3.2	3.0		
L	Ia	C1-LysRS	83.5	83.6	83.9	16.3	16.3	16.1	0.2	0.1	0.0		
L	Ia	CysRS	1.5	1.5	1.8	87.9	88.5	88.9	10.6	10.0	9.4		
E	Ia	IleRS	0.8	0.8	1.3	96.1	96.4	96.0	3.1	2.9	2.7	T	R
E	Ia	LeuRS	0.4	0.4	0.8	96.5	97.4	97.6	3.1	2.2	1.6	T	
L	Ia	MetRS	0.7	1.0	7.8	93.7	96.2	90.0	5.6	2.9	2.2	T	R
E	Ia	ValRS	0.2	0.2	0.5	98.3	98.5	98.5	1.5	1.4	1.0		
L	Ib	GlnRS	61.9	62.1	62.4	37.9	37.8	37.4	0.2	0.1	0.1		
E	Ib	Glu-Q-RS	63.4	63.9	72.6	36.6	36.1	27.4	0.1	0.0	0.0		
E	Ib	GluRS	0.7	0.9	3.5	83.8	84.7	85.5	15.5	14.4	10.9		
L	Ic	TrpRS	0.4	1.0	16.3	89.4	88.9	81.9	10.2	10.1	1.8	T	R
L	Ic	TyrRS	0.5	0.5	1.5	92.9	93.3	92.3	6.6	6.2	6.2	T	
L		C1_C2-LysRS	0.5	0.6	1.1	67.9	68.7	89.0	31.5	30.7	9.9	T	
E	IIa	AlaRS	0.8	1.0	1.4	98.3	98.4	98.4	1.0	0.6	0.2	T	
E	IIa	GlyRS	0.4	0.9	1.0	98.4	98.5	98.4	1.2	0.7	0.5		
L	IIa	HisRS	0.5	0.7	9.3	60.8	95.3	89.9	38.6	3.9	0.8		
E	IIa	ProRS	0.4	0.5	0.9	41.0	95.1	95.7	58.5	4.4	3.4	T	
E	IIa	SerRS	0.4	0.7	2.1	93.1	97.1	96.3	6.5	2.1	1.6	T	R
E	IIa	ThrRS	0.9	0.9	1.5	92.2	92.6	92.5	7.0	6.5	6.0	T	
L	IIb	AsnRS	46.8	47.5	48.4	50.0	50.6	51.2	3.3	1.9	0.4	T	
E	IIb	AspRS	0.4	0.4	1.1	89.4	93.5	93.2	10.2	6.1	5.7	T	
L	IIb	C2-LysRS	11.7	11.8	15.7	62.1	62.6	75.9	26.2	25.6	8.4		
L	IIc	PheRS	0.3	0.6	7.0	98.8	99.3	93.0	0.9	0.1	0.0	T	
L	IIc	PylRS	99.1	99.1	99.7	0.9	0.9	0.3	0.0	0.0	0.0		
L	IIc	SepRS	97.6	98.1	98.1	2.4	1.9	1.9	0.0	0.0	0.0		

Πίνακας 3.1: Προφίλ εξελικτικής μεταβλητότητας. Η συχνότητα (%) έλλειψης-παρουσίας ενός γονιδίου – παρουσίας διπλασιασμένου γονιδίου (2 και πάνω). Η συχνότητα υπολογίστηκε με 3 μεθόδους ανίχνευσης. i) HMM καταλυτικής επικράτειας (CD), ii) τουλάχιστον 2 MEME-motifs (M2) iii) τουλάχιστον 5 MEME-motifs (M5). Πορτοκαλί χρώμα για γονίδια που απουσιάζουν τουλάχιστον στο 5% των γονιδιωμάτων. Μπλε χρώμα για γονίδια που υπάρχουν ως ένα αντίγραφο σε τουλάχιστον 95% των γονιδιωμάτων. Πράσινο χρώμα για γονίδια που έχουν ένα και πάνω αντίγραφα ενός γονιδίου σε τουλάχιστον 5% των γονιδιωμάτων. Η πρώτη στήλη δηλώνει εάν το αμινοξύ θεωρείται "νέο" ή "παλαιό" ως προς την εισαγωγή του στον γενετικό κώδικα. Η δεύτερη στήλη υποδηλώνει την εξελικτική ομάδα της συνθετάσης. Η προ-τελευταία στήλη υποδηλώνει με T εάν η AARS στοχεύεται από κάποια τοξίνη. Η τελευταία στήλη υποδηλώνει με το R εάν η AARS έχει παράλογο με αντίσταση σε φυσικούς αναστολείς.

Μία κρίσιμη ερώτηση για να καταλάβουμε τους μοριακούς μηχανισμούς της εξέλιξης των παράλογων των AARSs που βρίσκονται στο ίδιο γονιδίωμα, είναι αν προκύπτουν είτε από γονιδιακό διπλασιασμό είτε από οριζόντια γονιδιακή μεταφορά. Για να καταφέρουμε να απαντήσουμε σε αυτή την ερώτηση, έγινε μία βασική παραδοχή. Αν δύο ομόλογα του ίδιου ενζύμου έχουν πολύ υψηλή ομοιότητα πρωτεϊνικής αλληλουχίας, τότε πιθανότατα να προέρχονται από πρόσφατο γονιδιακό διπλασιασμό. Από την άλλη πλευρά, αν δύο ομόλογα έχουν μάλλον χαμηλή ομοιότητα στις πρωτεϊνικές τους αλληλουχίες, τότε είτε είναι αποτέλεσμα οριζόντιας γονιδιακής μεταφοράς, είτε είναι αποτέλεσμα ενός πολύ παλαιού γονιδιακού διπλασιασμού, το οποίο μάλιστα οδήγησε και σε ταχεία απόκλιση. Χρησιμοποιώντας 1751 ζευγάρια ομολόγων πρωτεϊνών, υπολογίστηκε η κατανομή του ποσοστού ομοιότητας των πρωτεϊνών (βλ. γράφημα 3.2). Από το γράφημα είναι προφανές ότι (με βάση τις παραπάνω παραδοχές) <10% των ομόλογων είναι αποτέλεσμα πρόσφατου γονιδιακού διπλασιασμού μέσα στα γονιδιώματα. Αυτός ο υπολογισμός είναι σύμφωνος με προηγούμενη μεγάλης κλίμακας ανάλυσης σε άλλες οικογένειες πρωτεϊνών, όπου η οριζόντια γονιδιακή μεταφορά σε σχέση με τον γονιδιακό διπλασιασμό, θεωρείται η κινητήριος δύναμη πίσω από την επέκταση μίας οικογένειας πρωτεϊνών στους προκαρυωτικούς οργανισμούς, με μια εκτιμώμενη συμβολή 88%-98% (Treangen and Rocha 2011). Επιπλέον μία προηγούμενη ανάλυση με πολύ λιγότερα γονιδιωματικά δεδομένα διαθέσιμα εκείνη την εποχή (1999), επίσης υποστηρίζει την οριζόντια γονιδιακή μεταφορά ως κινητήριο δύναμη στην πρώιμη εξέλιξη της συγκεκριμένης πρωτεϊνικής οικογένειας (Wolf et al. 1999). Κατά μέσο όρο, δύο ομόλογα του ίδιου ενζύμου AARS στο ίδιο γονιδίωμα, είχαν 57% ομοιότητα στην πρωτεϊνική τους αλληλουχία. Το 60% των γονιδιωμάτων που αναλύθηκαν δεν είχαν καμία ένδειξη για ύπαρξη παραλόγων σε κάποιο γονίδιο AARS. Ενδιαφέρον έχει το γεγονός ότι η συντριπτική πλειοψηφία των AARSs ανιχνεύτηκε σε βακτηριακά χρωμοσώματα και μόλις το 4% των AARSs βρέθηκε σε πλασμίδια (βλ. supplementary file 1: NCBI_protein_info). Προφανώς, η οριζόντια γονιδιακή μεταφορά παίζει σημαντικό ρόλο στην επέκταση των παράλογων AARSs, αλλά θα πρέπει επίσης να εξεταστεί η περίπτωση σύνθετων εξελικτικών σεναρίων όπως είναι ο γονιδιακός διπλασιασμός ακολουθούμενος από ταχεία απόκλιση, ανακατάταξη επικρατειών και απώλεια γονιδίων (Ribas de Pouplana and Schimmel 2001; O'Donoghue and Luthey-Schulten 2003; Kunin and Ouzounis 2003; Kyrpides, Overbeek, and Ouzounis 1999).

Είναι δελεαστικό να υποθέτουμε ό,τι η εκτεταμένη παρουσία παραλόγων οδηγείται κυρίως από ανθεκτικότητα σε αναστολείς, όπως για παράδειγμα στην περίπτωση των βακτηριακών πρωτεϊνών MprF, που είναι συντηγμένες στην LysRS (Roy and Ibba 2008). Παρόλα αυτά, η ενσωμάτωση της διαθέσιμης γνώσης για τα γνωστά παράλογα των AARSs που έχουν τέτοιες ιδιότητες, μαζί με παρουσία των παραλόγων στους προκαρυωτικά γονιδιώματα, (βλ. πίνακα 1) δεν παρέχει κάποια ισχυρή υποστήριξη για έναν τέτοιο ισχυρισμό. Μία εναλλακτική εξήγηση είναι ότι αυτά τα παράλογα συμμετέχουν σε άλλες βιοχημικές λειτουργίες (Giegé and Springer 2016), πολλές από οποίες είναι ακόμα άγνωστες. Π.χ. υπάρχουν παρατηρήσεις όπου οι AARSs σχηματίζουν σύμπλοκα με άλλους τύπους πρωτεϊνών, συμμετέχοντας έτσι σε άλλες λειτουργίες πέρα από τη μετάφραση (Laporte et al. 2014; Rubio et al. 2015). Είναι επίσης ενδιαφέρον το γεγονός ότι άτυπα παράλογα των AARSs έχουν εντοπισθεί να συμμετέχουν στον σχηματισμό πεπτιδίων μέσω μη-ριβοσωμικού μηχανισμού σύνθεσης πεπτιδίων (Mocibob et al. 2010). Με μία βασική αυτοψία της αρχιτεκτονικής των MEME-motifs και των επικρατείων, μέσα από το υπολογιστικό μας εργαλείο, κάποιος θα μπορούσε σχετικά εύκολα να έχει μία καλύτερη εικόνα για το ποιο είναι το τυπικό αλληλόμορφο και ποιο το διαφοροποιημένο.

3.6 Υπολογιστική ανίχνευση πιθανών αντιβιοτικών/φυσικών αναστολέων των AARSs

Πρόσφατα υπάρχει μία αυξανόμενη ανησυχία, ότι η κακή χρήση των αντιβιοτικών προκαλεί την εμφάνιση ανθεκτικών στελεχών που θα μπορούσαν ενδεχομένων να οδηγήσουν σε ένα κόσμο χωρίς αποτελεσματικά αντιβιοτικά (Liu et al. 2015). Ως εκ τούτου, η ανακάλυψη νέων αντιβιοτικών, αν και έχει παραμεληθεί για αρκετό καιρό, είναι πλέον και πάλι στο επίκεντρο. Τα AARSs έχουν προσελκύσει ιδιαίτερη προσοχή τα τελευταία χρόνια, ως στόχος νέων αντιβακτηριακών αναστολέων.

Το Mupirocin είναι προϊόν μιας συστάδας γονιδίων PKS στον οργανισμό *Pseudomonas fluorenscens* και δρα ως αναστολέας της IleRS. Μάλιστα χρησιμοποιείται ως αντιβιοτικό απέναντι στον *Staphylococcus aureus*, που είναι ανθεκτικός στην μεθικιλλίνη. Το τροποποιημένο παράγωγο IleRS mupA που βρίσκεται εντός της βιοσυνθετικής συστάδας του mupirocin (βλ. εικ. 1.3), δεν καταστέλεται από το mupirocin , δηλαδή δρα ως αντίδοτο (El-Sayed et al. 2003; Seah et al. 2012). Για να ακυρώσει τον αναστολέα του mupirocin ο *S. aureus* έχει αναπτύξει ένα άλλο παράλογο, το mupB (Seah et al. 2012). Με σκοπό να εντοπιστούν και άλλα πιθανά αντιβιοτικά που να παράγονται με τον ίδιο μηχανισμό, αναζητήσαμε παράλογα/ξενόλογα AARSs που έχουν στην γονιδιακή περιοχή τους (±10 γειτονικά γονίδια) γονίδια που έχουν σχολιαστεί από το NCBI ως πολυκετιδικές συνθάσες (PKS). Τα αποτελέσματα αυτής της *in silico* γονιδιακής αναζήτησης συνοψίζονται στην εικόνα 3.2 και θα πρέπει να επαληθευτούν πειραματικά στο μέλλον. Το cluster του mupirocin δεν εντοπίστηκε, καθώς το γονιδίωμα του συγκεκριμένου στελέχους του *P. fluorescens* δεν ήταν διαθέσιμο στο NCBI. Με την μείωση του κόστους της αλληλούχισης, πολύ σύντομα θα χρησιμοποιούνται μεταγονιδιοματικές προσεγγίσεις για τη σάρωση και τον εντοπισμό περιβαλλοντικών περιοχών που φιλοξενούν τόσο γονίδια PKS όσο και γονίδια AARSs στην ίδια γονιδιωματική γειτονιά.

Organism	Phylogenetic group	AARS paralogue/ xenologue	Organismal characteristics	Number of PKS within vicinity
Paenibacillus polymyxa E681	Firmicutes	AspRS	Produces compounds with antifungal or antibacterial activity.	7
Azospirillum sp. B510	Proteobacteria	PheRS	Plant growth promoting organism.	2
Mycobacterium abscessus	Actinobacteria	CysRS	Causes a chronic lung infection, similar to tuberculosis, in patients with cystic fibrosis. Very resistant to many commonly used antibiotics.	6
Mycobacterium smegmatis JS623	Actinobacteria	CysRS	Associated with soft tissue lesions following trauma or surgery.A possible factor in penile carcinogenesis.	4

Εικόνα 3.2: Μικροοργανισμοί που πιθανόν να διαθέτουν συστάδα γονιδίων για την βιοσύνθεση φυσικών αναστολέων AARSs (Chaliotis et al. 2017)

4. Συμπεράσματα

Ένας από τους βασικούς στόχους αυτής της έρευνας ήταν η ανάπτυξη ενός υπολογιστικού εργαλείου για την γρήγορη και ευαίσθητη ανίχνευση πρωτεϊνών AARS σε ολόκληρα πρωτεώματα οργανισμών. Αυτό επίτευχθηκε με την αυτόματη και αμερόληπτη ανίχνευση motifs μέσα σε κάθε ένα από τα 22 ένζυμα AARS, χρησιμοποιώντας τον αλγόριθμο του MEME σε συνδυασμό με φυλογενετική ανάλυση. Στη συνέχεια τα motfs τα οποία βρέθηκαν, χρησιμοποιήθηκαν για την δημιουργία HMMs. Επίσης, για να εξασφαλιστεί η ανίχνευση απομακρυσμένων ομολόγων πρωτεϊνών, δημιουργήθηκαν HMMs από όλο το καταλυτικό domain, κάθε ενός από τα ένζυμα της οικογένειας AARS. Στη συνέχεια αυτά τα HMMs ομαδοποιήθηκαν σε μία βάση δεδομένων HMMs, και έπειτα δημιουργήσαμε ένα διαδικτυακό εργαλείο που χρησιμοποιεί αυτά τα HMMs για να σαρώνει ολόκληρα πρωτεώματα οργανισμών.

Ο δεύτερος μεγάλος στόχος αυτής της μελέτης ήταν η ανάλυση περισσότερων από 2500 προκαρυωτικών πρωτεωμάτων, για τον εντοπισμό της παρουσίας ή απουσίας ακολουθιών AARSs. Τα αποτελέσματα αυτής της ανάλυσης αποθηκεύτηκαν και οργανώθηκαν σε μία βάση δεδομένων η οποία είναι ανοιχτή στο κοινό. Αυτή η μεγάλη σε μέγεθος και ταυτόχρονα ολοκληρωμένη εξελικτική ανάλυση ποσοτικοποίησε για πρώτη φορά τη μεγάλη μεταβλητότητα που υπάρχει μέσα σε ένα συγκεκριμένο αλλά ταυτόχρονα απαραίτητο στοιχείο για την λειτουργία του μηχανισμού μετάφρασης του γενετικού κώδικα. Η ακυλίωση του tRNAμε το συγγενές αμινοξύ του. Επίσης, από ότι φαίνεται η παρουσία περισσότερων του ενός ομολόγων για το ίδιο ένζυμο, είναι συνήθως το αποτέλεσμα οριζόντιας γονιδιακής μεταφοράς ή γονιδιακός διπλασιασμός ακολουθούμενος από ταχεία απόκλιση της ακολουθίας, πιθανότατα συνδεδεμένη με εναλλακτικές βιοχημικές λειτουργίες. Ωστόσο δεν μπορεί να αποκλιστεί και η πιθανότητα ανάπτυξης αντίστασης απέναντι σε τοξίνες που στοχεύουν τα συγκεκριμένα ένζυμα.

Η εργασία αυτή δημοσιεύθηκε στο Διεθνές Επιστημονικό Περιοδικό Nucleic Acids Research (https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1182) και ήταν το αποτέλεσμα έρευνας 5 ετών που ξεκίνησα ως προπτυχιακός φοιτητής στο εργαστήριο του κ. Αμούτζια Γρηγόριου, Επίκουρου Καθηγητή Βιοπληροφορικής στη Γενωμική, στο Τμήμα Βιοχημείας και Βιοτεχνολογίας, στο Πανεπιστήμιο Θεσσαλίας και συνέχισα ως μεταπτυχιακός φοιτητής.

Published online 28 November 2016 Nucleic Acids Research, 2017, Vol. 45, No. 3 1059–1068
doi: 10.1093/nar/gkw1182

The complex evolutionary history of aminoacyl-tRNA synthetases

Anargyros Chaliotis[1], Panayotis Vlastaridis[1], Dimitris Mossialos[2], Michael Ibba[3], Hubert D. Becker[4], Constantinos Stathopoulos[5,*] and Grigorios D. Amoutzias[1,*]

[1]Bioinformatics Laboratory, Department of Biochemistry and Biotechnology, University of Thessaly, 41500 Larissa, Greece, [2]Molecular Microbiology Laboratory, Department of Biochemistry and Biotechnology, University of Thessaly, 41500 Larissa, Greece, [3]Department of Microbiology, The Ohio State University, Columbus, OH 43210, USA, [4]Génétique Moléculaire, Génomique, Microbiologie, UMR 7156, CNRS, Université de Strasbourg, 4 allée Konrad Röntgen, 67084 Strasbourg Cedex, France and [5]Department of Biochemistry, School of Medicine, University of Patras, 26504 Patras, Greece

Received August 25, 2016; Revised October 20, 2016; Editorial Decision November 14, 2016; Accepted November 16, 2016

Βιβλιογραφία

Ahel, Ivan, Dragana Korencic, Michael Ibba, and Dieter Söll. 2003. "Trans-Editing of Mischarged TRNAs." *Proceedings of the National Academy of Sciences of the United States of America* 100 (26): 15422–27. doi:10.1073/pnas.2136934100.

Ambrogelly, Alexandre, Patrick O'Donoghue, Dieter Söll, and Sarath Moses. 2010. "A Bacterial Ortholog of Class II Lysyl-TRNA Synthetase Activates Lysine." *FEBS Letters* 584 (14): 3055–60. doi:10.1016/j.febslet.2010.05.036.

Andam, Cheryl P, Gregory P Fournier, and Johann Peter Gogarten. 2011. "Multilevel Populations and the Evolution of Antibiotic Resistance through Horizontal Gene Transfer." *FEMS Microbiology Reviews* 35 (5): 756–67. doi:10.1111/j.1574-6976.2011.00274.x.

Antonio, Martin, Neil McFerran, and Mark J. Pallen. 2002. "Mutations Affecting the Rossman Fold of Isoleucyl-TRNA Synthetase Are Correlated with Low-Level Mupirocin Resistance in Staphylococcus Aureus." *Antimicrobial Agents and Chemotherapy* 46 (2): 438–42.

Attwood, Teresa K., Alain Coletta, Gareth Muirhead, Athanasia Pavlopoulou, Peter B. Philippou, Ivan Popov, Carlos Romá-Mateo, Athina Theodosiou, and Alex L. Mitchell. 2012. "The PRINTS Database: A Fine-Grained Protein Sequence Annotation and Analysis Resource--Its Status in 2012." *Database: The Journal of Biological Databases and Curation* 2012: bas019. doi:10.1093/database/bas019.

Bailey, Timothy L., James Johnson, Charles E. Grant, and William S. Noble. 2015. "The MEME Suite." *Nucleic Acids Research*, May. doi:10.1093/nar/gkv416.

Becker, H. D., and D. Kern. 1998. "Thermus Thermophilus: A Link in Evolution of the TRNA-Dependent Amino Acid Amidation Pathways." *Proceedings of the National Academy of Sciences of the United States of America* 95 (22): 12832–37.

Bennett, Gordon M., and Nancy A. Moran. 2013. "Small, Smaller, Smallest: The Origins and Evolution of Ancient Dual Symbioses in a Phloem-Feeding Insect." *Genome Biology and Evolution* 5 (9): 1675–88. doi:10.1093/gbe/evt118.

Blaise, Mickaël, Hubert Dominique Becker, Gérard Keith, Christian Cambillau, Jacques Lapointe, Richard Giegé, and Daniel Kern. 2004. "A Minimalist Glutamyl-TRNA Synthetase Dedicated to Aminoacylation of the TRNAAsp QUC Anticodon." *Nucleic*

Acids Research 32 (9): 2768–75. doi:10.1093/nar/gkh608.
Brown, J. R., and W. F. Doolittle. 1999. “Gene Descent, Duplication, and Horizontal Transfer in the Evolution of Glutamyl- and Glutaminyl-TRNA Synthetases.” *Journal of Molecular Evolution* 49 (4): 485–95.
Chaliotis, Anargyros, Panayotis Vlastaridis, Dimitris Mossialos, Michael Ibba, Hubert D. Becker, Constantinos Stathopoulos, and Grigorios D. Amoutzias. 2017. “The Complex Evolutionary History of Aminoacyl-TRNA Synthetases.” *Nucleic Acids Research* 45 (3): 1059–68. doi:10.1093/nar/gkw1182.
Chevenet, François, Christine Brun, Anne-Laure Bañuls, Bernard Jacq, and Richard Christen. 2006. “TreeDyn: Towards Dynamic Graphics and Annotations for Analyses of Trees.” *BMC Bioinformatics* 7: 439. doi:10.1186/1471-2105-7-439.
Chothia, Cyrus, Julian Gough, Christine Vogel, and Sarah A. Teichmann. 2003. “Evolution of the Protein Repertoire.” *Science (New York, N.Y.)* 300 (5626): 1701–3. doi:10.1126/science.1085371.
Curnow, A. W., K. w Hong, R. Yuan, S. i Kim, O. Martins, W. Winkler, T. M. Henkin, and D. Söll. 1997. “Glu-TRNAGln Amidotransferase: A Novel Heterotrimeric Enzyme Required for Correct Decoding of Glutamine Codons during Translation.” *Proceedings of the National Academy of Sciences of the United States of America* 94 (22): 11819–26.
Dong, Xianchi, Minyun Zhou, Chen Zhong, Bei Yang, Ning Shen, and Jianping Ding. 2010. “Crystal Structure of Pyrococcus Horikoshii Tryptophanyl-TRNA Synthetase and Structure-Based Phylogenetic Analysis Suggest an Archaeal Origin of Tryptophanyl-TRNA Synthetase.” *Nucleic Acids Research* 38 (4): 1401–12. doi:10.1093/nar/gkp1053.
Eddy, Sean R. 2011. “Accelerated Profile HMM Searches.” *PLoS Computational Biology* 7 (10): e1002195. doi:10.1371/journal.pcbi.1002195.
Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97. doi:10.1093/nar/gkh340.
El-Sayed, A. Kassem, Joanne Hothersall, Sian M. Cooper, Elton Stephens, Thomas J. Simpson, and Christopher M. Thomas. 2003. “Characterization of the Mupirocin Biosynthesis Gene Cluster from Pseudomonas Fluorescens NCIMB 10586.” *Chemistry & Biology* 10 (5): 419–30.
Eriani, G., M. Delarue, O. Poch, J. Gangloff, and D. Moras. 1990. “Partition of TRNA Synthetases into Two Classes Based on Mutually Exclusive Sets of Sequence Motifs.” *Nature* 347 (6289): 203–6. doi:10.1038/347203a0.
Fournier, Gregory P., Cheryl P. Andam, and Johann Peter Gogarten. 2015. “Ancient Horizontal Gene Transfer and the Last Common Ancestors.” *BMC Evolutionary Biology* 15: 70. doi:10.1186/s12862-015-0350-0.
Fox, Naomi K., Steven E. Brenner, and John-Marc Chandonia. 2014. “SCOPe: Structural Classification of Proteins--Extended, Integrating SCOP and ASTRAL Data and Classification of New Structures.” *Nucleic Acids Research* 42 (Database issue): D304-309. doi:10.1093/nar/gkt1240.
Francklyn, Christopher. 2003. “TRNA Synthetase Paralogs: Evolutionary Links in the Transition from TRNA-Dependent Amino Acid Biosynthesis to de Novo Biosynthesis.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (17): 9650–52. doi:10.1073/pnas.1934245100.
Giegé, Richard, and Mathias Springer. 2016. “Aminoacyl-TRNA Synthetases in the Bacterial World.” *EcoSal Plus* 7 (1). doi:10.1128/ecosalplus.ESP-0002-2016.

Gilbart, J., C. R. Perry, and B. Slocombe. 1993. “High-Level Mupirocin Resistance in Staphylococcus Aureus: Evidence for Two Distinct Isoleucyl-TRNA Synthetases.” *Antimicrobial Agents and Chemotherapy* 37 (1): 32–38.

Gouy, Manolo, Stéphane Guindon, and Olivier Gascuel. 2010. “SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building.” *Molecular Biology and Evolution* 27 (2): 221–24. doi:10.1093/molbev/msp259.

Guo, Li-Tao, Yane-Shih Wang, Akiyoshi Nakamura, Daniel Eiler, Jennifer M. Kavran, Margaret Wong, Laura L. Kiessling, Thomas A. Steitz, Patrick O’Donoghue, and Dieter Söll. 2014. “Polyspecific Pyrrolysyl-TRNA Synthetases from Directed Evolution.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (47): 16724–29. doi:10.1073/pnas.1419737111.

Guo, Min, Yeeting E. Chong, Ryan Shapiro, Kirk Beebe, Xiang-Lei Yang, and Paul Schimmel. 2009. “Paradox of Mistranslation of Serine for Alanine Caused by AlaRS Recognition Dilemma.” *Nature* 462 (7274): 808–12. doi:10.1038/nature08612.

Guo, Min, Paul Schimmel, and Xiang-Lei Yang. 2010. “Functional Expansion of Human TRNA Synthetases Achieved by Structural Inventions.” *FEBS Letters* 584 (2): 434–42. doi:10.1016/j.febslet.2009.11.064.

Hadd, Andrew, and John J. Perona. 2014. “Recoding Aminoacyl-TRNA Synthetases for Synthetic Biology by Rational Protein-RNA Engineering.” *ACS Chemical Biology* 9 (12): 2761–66. doi:10.1021/cb5006596.

Han, Jung Min, Seung Jae Jeong, Min Chul Park, Gyuyoup Kim, Nam Hoon Kwon, Hoi Kyoung Kim, Sang Hoon Ha, Sung Ho Ryu, and Sunghoon Kim. 2012. “Leucyl-TRNA Synthetase Is an Intracellular Leucine Sensor for the MTORC1-Signaling Pathway.” *Cell* 149 (2): 410–24. doi:10.1016/j.cell.2012.02.044.

Higgs, Paul G., and Ralph E. Pudritz. 2009. “A Thermodynamic Basis for Prebiotic Amino Acid Synthesis and the Nature of the First Genetic Code.” *Astrobiology* 9 (5): 483–90. doi:10.1089/ast.2008.0280.

Hurdle, Julian Gregston, Alexander John O’Neill, and Ian Chopra. 2005. “Prospects for Aminoacyl-TRNA Synthetase Inhibitors as New Antimicrobial Agents.” *Antimicrobial Agents and Chemotherapy* 49 (12): 4821–33. doi:10.1128/AAC.49.12.4821-4833.2005.

Ibba, M., A. W. Curnow, and D. Söll. 1997. “Aminoacyl-TRNA Synthesis: Divergent Routes to a Common Goal.” *Trends in Biochemical Sciences* 22 (2): 39–42.

Ibba, M., S. Morgan, A. W. Curnow, D. R. Pridmore, U. C. Vothknecht, W. Gardner, W. Lin, C. R. Woese, and D. Söll. 1997. “A Euryarchaeal Lysyl-TRNA Synthetase: Resemblance to Class I Synthetases.” *Science (New York, N.Y.)* 278 (5340): 1119–22.

Kim, Sunghoon, Sungyong You, and Daehee Hwang. 2011. “Aminoacyl-TRNA Synthetases and Tumorigenesis: More than Housekeeping.” *Nature Reviews. Cancer* 11 (10): 708–18. doi:10.1038/nrc3124.

Koonin, E. V., K. S. Makarova, and L. Aravind. 2001. “Horizontal Gene Transfer in Prokaryotes: Quantification and Classification.” *Annual Review of Microbiology* 55: 709–42. doi:10.1146/annurev.micro.55.1.709.

Kunin, Victor, and Christos A. Ouzounis. 2003. “The Balance of Driving Forces during Genome Evolution in Prokaryotes.” *Genome Research* 13 (7): 1589–94. doi:10.1101/gr.1092603.

Kyrpides, N., R. Overbeek, and C. Ouzounis. 1999. “Universal Protein Families and the Functional Content of the Last Universal Common Ancestor.” *Journal of Molecular*

Evolution 49 (4): 413–23.
Lamour, V., S. Quevillon, S. Diriong, V. C. N'Guyen, M. Lipinski, and M. Mirande. 1994. "Evolution of the Glx-TRNA Synthetase Family: The Glutaminyl Enzyme as a Case of Horizontal Gene Transfer." *Proceedings of the National Academy of Sciences of the United States of America* 91 (18): 8670–74.
Laporte, Daphné, Jonathan L. Huot, Gaétan Bader, Ludovic Enkler, Bruno Senger, and Hubert Dominique Becker. 2014. "Exploring the Evolutionary Diversity and Assembly Modes of Multi-Aminoacyl-TRNA Synthetase Complexes: Lessons from Unicellular Organisms." *FEBS Letters* 588 (23): 4268–78. doi:10.1016/j.febslet.2014.10.007.
Leinfelder, W., E. Zehelein, M. A. Mandrand-Berthelot, and A. Böck. 1988. "Gene for a Novel TRNA Species That Accepts L-Serine and Cotranslationally Inserts Selenocysteine." *Nature* 331 (6158): 723–25. doi:10.1038/331723a0.
Ling, Jiqiang, Patrick O'Donoghue, and Dieter Söll. 2015. "Genetic Code Flexibility in Microorganisms: Novel Mechanisms and Impact on Physiology." *Nature Reviews. Microbiology* 13 (11): 707–21. doi:10.1038/nrmicro3568.
Liu, Yi-Yun, Yang Wang, Timothy R Walsh, Ling-Xian Yi, Rong Zhang, James Spencer, Yohei Doi, et al. 2015. "Emergence of Plasmid-Mediated Colistin Resistance Mechanism MCR-1 in Animals and Human Beings in China: A Microbiological and Molecular Biological Study." *The Lancet Infectious Diseases*, November. doi:10.1016/S1473-3099(15)00424-7.
Lo, Wing-Sze, Elisabeth Gardiner, Zhiwen Xu, Ching-Fun Lau, Feng Wang, Jie J. Zhou, John D. Mendlein, et al. 2014. "Human TRNA Synthetase Catalytic Nulls with Diverse Functions." *Science (New York, N.Y.)* 345 (6194): 328–32. doi:10.1126/science.1252943.
Marchler-Bauer, Aron, Myra K. Derbyshire, Noreen R. Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y. Geer, Renata C. Geer, et al. 2015. "CDD: NCBI's Conserved Domain Database." *Nucleic Acids Research* 43 (Database issue): D222-226. doi:10.1093/nar/gku1221.
Marsh, Joseph A., and Sarah A. Teichmann. 2010. "How Do Proteins Gain New Domains?" *Genome Biology* 11 (7): 126. doi:10.1186/gb-2010-11-7-126.
Mitchell, Alex, Hsin-Yu Chang, Louise Daugherty, Matthew Fraser, Sarah Hunter, Rodrigo Lopez, Craig McAnulla, et al. 2015. "The InterPro Protein Families Database: The Classification Resource after 15 Years." *Nucleic Acids Research* 43 (Database issue): D213-221. doi:10.1093/nar/gku1243.
Mocibob, Marko, Nives Ivic, Silvija Bilokapic, Timm Maier, Marija Luic, Nenad Ban, and Ivana Weygand-Durasevic. 2010. "Homologs of Aminoacyl-TRNA Synthetases Acylate Carrier Proteins and Provide a Link between Ribosomal and Nonribosomal Peptide Synthesis." *Proceedings of the National Academy of Sciences of the United States of America* 107 (33): 14585–90. doi:10.1073/pnas.1007470107.
Moras, D. 1992. "Structural and Functional Relationships between Aminoacyl-TRNA Synthetases." *Trends in Biochemical Sciences* 17 (4): 159–64.
Ochsner, Urs A., Xicheng Sun, Thale Jarvis, Ian Critchley, and Nebojsa Janjic. 2007. "Aminoacyl-TRNA Synthetases: Essential and Still Promising Targets for New Anti-Infective Agents." *Expert Opinion on Investigational Drugs* 16 (5): 573–93. doi:10.1517/13543784.16.5.573.
O'Donoghue, Patrick, and Zaida Luthey-Schulten. 2003. "On the Evolution of Structure in Aminoacyl-TRNA Synthetases." *Microbiology and Molecular Biology Reviews: MMBR*

67 (4): 550–73.
Passioura, Toby, and Hiroaki Suga. 2014. “Reprogramming the Genetic Code in Vitro.” *Trends in Biochemical Sciences* 39 (9): 400–408. doi:10.1016/j.tibs.2014.07.005.
Pham, James S., Karen L. Dawson, Katherine E. Jackson, Erin E. Lim, Charisse Flerida A. Pasaje, Kelsey E. C. Turner, and Stuart A. Ralph. 2014. “Aminoacyl-TRNA Synthetases as Drug Targets in Eukaryotic Parasites.” *International Journal for Parasitology. Drugs and Drug Resistance* 4 (1): 1–13. doi:10.1016/j.ijpddr.2013.10.001.
Ribas de Pouplana, L., and P. Schimmel. 2001. “Two Classes of TRNA Synthetases Suggested by Sterically Compatible Dockings on TRNA Acceptor Stem.” *Cell* 104 (2): 191–93.
Rose, Peter W., Andreas Prlić, Chunxiao Bi, Wolfgang F. Bluhm, Cole H. Christie, Shuchismita Dutta, Rachel Kramer Green, et al. 2015. “The RCSB Protein Data Bank: Views of Structural Biology for Basic and Applied Research and Education.” *Nucleic Acids Research* 43 (Database issue): D345-356. doi:10.1093/nar/gku1214.
Roy, Hervé, and Michael Ibba. 2008. “RNA-Dependent Lipid Remodeling by Bacterial Multiple Peptide Resistance Factors.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (12): 4667–72. doi:10.1073/pnas.0800006105.
Rubio, Miguel Ángel, Mauro Napolitano, Jesús A. G. Ochoa de Alda, Javier Santamaría-Gómez, Carl J. Patterson, Andrew W. Foster, Roque Bru-Martínez, Nigel J. Robinson, and Ignacio Luque. 2015. “Trans-Oligomerization of Duplicated Aminoacyl-TRNA Synthetases Maintains Genetic Code Fidelity under Stress.” *Nucleic Acids Research* 43 (20): 9905–17. doi:10.1093/nar/gkv1020.
Sauerwald, Anselm, Wenhong Zhu, Tiffany A. Major, Hervé Roy, Sotiria Palioura, Dieter Jahn, William B. Whitman, John R. Yates, Michael Ibba, and Dieter Söll. 2005. “RNA-Dependent Cysteine Biosynthesis in Archaea.” *Science (New York, N.Y.)* 307 (5717): 1969–72. doi:10.1126/science.1108329.
Seah, Christine, David C Alexander, Lisa Louie, Andrew Simor, Donald E Low, Jean Longtin, and Roberto G Melano. 2012. “MupB, a New High-Level Mupirocin Resistance Mechanism in Staphylococcus Aureus.” *Antimicrobial Agents and Chemotherapy* 56 (4): 1916–20. doi:10.1128/AAC.05325-11.
Sekine, S., A. Shimada, O. Nureki, J. Cavarelli, D. Moras, D. G. Vassylyev, and S. Yokoyama. 2001. “Crucial Role of the High-Loop Lysine for the Catalytic Activity of Arginyl-TRNA Synthetase.” *The Journal of Biological Chemistry* 276 (6): 3723–26. doi:10.1074/jbc.C000756200.
Sheppard, Kelly, Jing Yuan, Michael J. Hohn, Brian Jester, Kevin M. Devine, and Dieter Söll. 2008. “From One Amino Acid to Another: TRNA-Dependent Amino Acid Biosynthesis.” *Nucleic Acids Research* 36 (6): 1813–25. doi:10.1093/nar/gkn015.
Sissler, M., C. Delorme, J. Bond, S. D. Ehrlich, P. Renault, and C. Francklyn. 1999. “An Aminoacyl-TRNA Synthetase Paralog with a Catalytic Role in Histidine Biosynthesis.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (16): 8985–90.
Szymanski, M., M. A. Deniziak, and J. Barciszewski. 2001. “Aminoacyl-TRNA Synthetases Database.” *Nucleic Acids Research* 29 (1): 288–90.
Teichmann, S. A., S. C. Rison, J. M. Thornton, M. Riley, J. Gough, and C. Chothia. 2001. “The Evolution and Structural Anatomy of the Small Molecule Metabolic Pathways in Escherichia Coli.” *Journal of Molecular Biology* 311 (4): 693–708. doi:10.1006/jmbi.2001.4912.

Treangen, Todd J., and Eduardo P. C. Rocha. 2011. “Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes.” *PLoS Genetics* 7 (1): e1001284. doi:10.1371/journal.pgen.1001284.

Tumbula, D. L., H. D. Becker, W. Z. Chang, and D. Söll. 2000. “Domain-Specific Recruitment of Amide Amino Acids for Protein Synthesis.” *Nature* 407 (6800): 106–10. doi:10.1038/35024120.

UniProt Consortium. 2015. “UniProt: A Hub for Protein Information.” *Nucleic Acids Research* 43 (Database issue): D204-212. doi:10.1093/nar/gku989.

Valencia-Sánchez, Marco Igor, Annia Rodríguez-Hernández, Ruben Ferreira, Hugo Aníbal Santamaria-Suárez, Marcelino Arciniega, Anne-Catherine Dock-Bregeon, Dino Moras, et al. 2016. “Structural Insights into the Polyphyletic Origins of Glycyl TRNA Synthetases.” *The Journal of Biological Chemistry*, May. doi:10.1074/jbc.M116.730382.

Wang, Lei, Jianming Xie, and Peter G. Schultz. 2006. “Expanding the Genetic Code.” *Annual Review of Biophysics and Biomolecular Structure* 35: 225–49. doi:10.1146/annurev.biophys.35.101105.121507.

Weiner, January, Andrew D. Moore, and Erich Bornberg-Bauer. 2008. “Just How Versatile Are Domains?” *BMC Evolutionary Biology* 8: 285. doi:10.1186/1471-2148-8-285.

Woese, C R, G J Olsen, M Ibba, and D Söll. 2000. “Aminoacyl-TRNA Synthetases, the Genetic Code, and the Evolutionary Process.” *Microbiology and Molecular Biology Reviews: MMBR* 64 (1): 202–36.

Wolf, Y. I., L. Aravind, N. V. Grishin, and E. V. Koonin. 1999. “Evolution of Aminoacyl-TRNA Synthetases--Analysis of Unique Domain Architectures and Phylogenetic Trees Reveals a Complex History of Horizontal Gene Transfer Events.” *Genome Research* 9 (8): 689–710.

Wong, J. T. 1975. “A Co-Evolution Theory of the Genetic Code.” *Proceedings of the National Academy of Sciences of the United States of America* 72 (5): 1909–12.

Yanagisawa, Tatsuo, and Makoto Kawakami. 2003. “How Does Pseudomonas Fluorescens Avoid Suicide from Its Antibiotic Pseudomonic Acid?: Evidence for Two Evolutionarily Distinct Isoleucyl-TRNA Synthetases Conferring Self-Defense.” *The Journal of Biological Chemistry* 278 (28): 25887–94. doi:10.1074/jbc.M302633200.