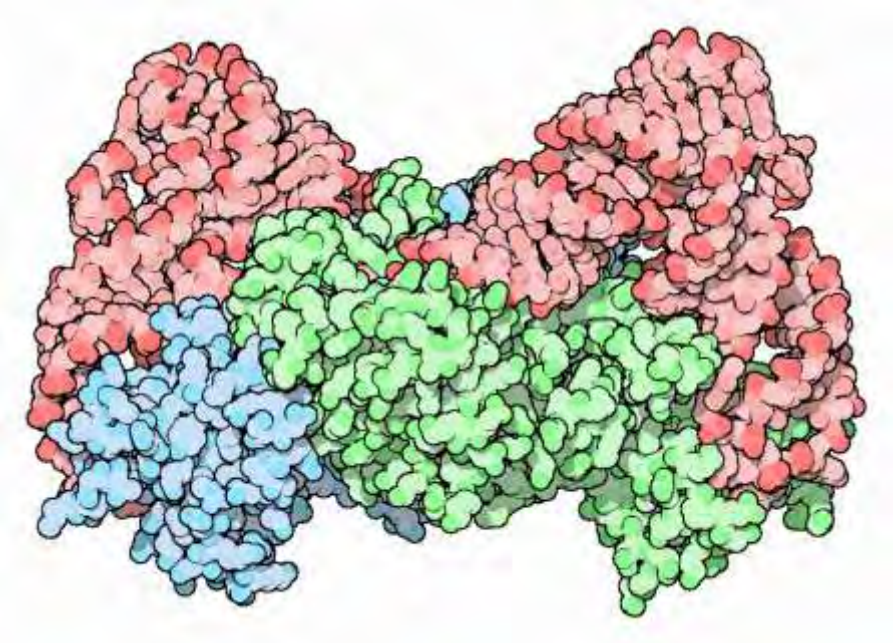
ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΒΙΟΧΗΜΕΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

Ανίχνευση μικροβιακών t-RNA συνθετασών με μεθόδους Βιοπληροφορικής



Διπλωματική Εργασία

Χαλιώτη Ανάργυρου

Λάρισα 2012

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

κ. Γρηγόριος Αμούτζιας (Επιβλέπων).

Λέκτορας Βιοπληροφορικής στη Γονιδιωματική, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας.

κ. Δημήτριος Μόσιαλος.

Επικ. Καθηγητής Βιοτεχνολογίας Μικροβίων, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας.

κ. Κωνσταντίνος Σταθόπουλος.

Αναπ. Καθηγητής Βιοχημείας με έμφαση στη βιοσύνθεση πρωτεϊνών, Τμήμα Ιατρικής, Πανεπιστήμιο Πατρών.

Πρωτίστως θα ήθελα να εκφράσω τις ευχαριστίες μου στον επιβλέπων καθηγητή μου κ. Αμούτζια Γρηγόριο, Λέκτορα Βιοπληροφορικής στη Γονιδιωματική του τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας, για τη βοήθεια του, την εμπιστοσύνη και την υπομονή που έδειξε στο πρόσωπο μου κατά τη διάρκεια εκπόνησης της πτυχιακής μου εργασίας.

Επίσης θα ήθελα να ευχαριστήσω τον κ. Μόσιαλο Δημήτριο, Επίκουρο καθηγητή Βιοτεχνολογίας Μικροβίων και μέλος της τριμελούς επιτροπής καθώς και τον κ. Σταθόπουλο Κωνσταντίνο, Αναπληρωτή καθηγητή Βιοχημείας με έμφαση στη βιοσύνθεση πρωτεϊνών στο τμήμα Ιατρικής του Πανεπιστημίου Πατρών και μέλος της τριμελούς μου επιτροπής για τη βοήθεια και τις συμβουλές που μου έδωσαν, κατά την εκπόνηση της πτυχιακής εργασίας μου.

Περιεχόμενα

Περίληψη

Οι αμινοάκυλο–tRNA συνθετάσες (aminoacyl-tRNA synthetases - AARSs) είναι πολύ συντηρημένες πρωτεΐνες με βασικό ρόλο στην πρωτεΐνοσύνθεση. Είναι υπεύθυνες για τη σύνδεση και ενεργοποίηση του κατάλληλου αμινοξέος στο σωστό μόριο tRNA. Οι προκαρυωτικοί οργανισμοί κάποιες φορές κατασκευάζουν τοξίνες που στοχεύουν και διαταράσσουν την ομαλή λειτουργία αυτών των συνθετασών και έτσι οι τοξίνες αυτές δρουν ως αντιβιοτικά. Οι μικροοργανισμοί που δέχονται την "επίθεση" απο αυτές τις τοξίνες με την σειρά τους μπορεί να αναπτύξουν αντίμετρα, όπως π.χ. μια διπλασιασμένη AARS με τροποιημένη αμινοξική ακολουθία, που δεν είναι ευαίσθητη στην τοξίνη και μπορεί να επιτελέσει την λειτουργία της. Επίσης, ο μικροοργανισμός που παράγει την τοξίνη μπορεί αυτός ο ίδιος να αναπτύξει και το αντίμετρο/αντίδοτο, μέσω του γονιδιακού διπλασιασμού και της εξέλιξης της διπλασιασμένης ακολουθίας. Επομένως, η ανίχνευση αυτών των διπλασιασμένων AARS είναι σημαντική για την έρευνα στην ανάπτυξη νέας γενιάς αντιβιοτικών. Στην παρούσα εργασία, μέσα από απομόνωση των καταλυτικών domain των AARSs δημιουργήσαμε στατιστικά μοντέλα (profile Hidden Markov Models - HMMs), ώστε να εντοπίσουμε με γρήγορες μεθόδους βιοπληροφορικής τις AARSs που έχουν στο πρωτέωμα τους χιλιάδες μικροοργανισμοί. Τα νέα HMM μοντέλα που δημιουργήθηκαν στο εργαστήριο αξιολογήθηκαν και εντόπισαν επιτυχώς το 100% ενός συνόλου γνωστών ακολουθιών. Γενικά, τα νέα μοντέλα που δημιουργήθηκαν στο εργαστήριο λειτούργησαν καλύτερα από ότι τα μοντέλα που δημιουργήθηκαν από τα CDD του NCBI. Στη συνέχεια 'σαρώθηκαν' ~2000 προκαρυωτικά πρωτεώματα από το NCBI. Περίπου 55% των πρωτεωμάτων που σαρώθηκαν είχαν >20 AARS το καθένα, ενώ το 22% των πρωτεωμάτων είχαν <20 AARS το καθένα. Επομένως, η αρχική ιδέα της μίας AARS για κάθε αμινοξύ βλέπουμε ότι δεν ισχύει στην πλειονότητα των μικροοργανισμών που μελετήθηκαν. Τα τρία γονίδια που εμφανίζονται πολύ συχνά διπλασιασμένα είναι η HisRS, LysRS (class II), CysRS. Ο οργανισμός με τις περισσότερες AARS (34) ήταν ο *Kitasatospora setae* KM 6054, όπου εμφανίστηκαν διπλασιασμοί σε 8 διαφορετικές AARS. Άλλοι οργανισμοί που βρέθηκαν με υψηλό αριθμό AARSs είναι στελέχη των ειδών *Bacilus cereus*, *Bacillus thuringiensis* και *Bacillus anthracis*. Στην αντίθετη πλευρά βρίσκονται 3 οργανισμοί στους οποίους εντοπίστηκαν μόλις 9 AARSs στο πρωτέωμά τους. Πρόκειται για τους *Candidatus Sulcia muelleri* DMIN, *Candidatus Sulcia muelleri* GWSS και *Candidatus Hodgkinia cicadicola* Dsem, οι οποίοι είναι και οι 3 συμβιωτικοί.

1. Εισαγωγή

1.1 Σκοπός της εργασίας

Οι αμινοάκυλο–tRNA συνθετάσες (aminoacyl-tRNA synthetases - AARSs) είναι πολύ συντηρημένες πρωτεΐνες με βασικό ρόλο στην πρωτεΐνοσύνθεση. Είναι υπεύθυνες για τη σύνδεση και ενεργοποίηση του κατάλληλου αμινοξέος στο σωστό μόριο tRNA. Οι προκαρυωτικοί οργανισμοί κάποιες φορές κατασκευάζουν τοξίνες που στοχεύουν και διαταράσσουν την ομαλή λειτουργία αυτών των συνθετασών και έτσι οι τοξίνες αυτές δρουν ως αντιβιοτικά. Οι μικροοργανισμοί που δέχονται την "επίθεση" απο αυτές τις τοξίνες με την σειρά τους μπορεί να αναπτύξουν αντίμετρα, όπως π.χ. μια διπλασιασμένη AARS με τροποιημένη αμινοξική ακολουθία, που δεν είναι ευαίσθητη στην τοξίνη και μπορεί να επιτελέσει την λειτουργία της. Επίσης, ο μικροοργανισμός που παράγει την τοξίνη μπορεί αυτός ο ίδιος να αναπτύξει και το αντίμετρο/αντίδοτο, μέσω του γονιδιακού διπλασιασμού και της εξέλιξης της διπλασιασμένης ακολουθίας. Επομένως, η ανίχνευση αυτών των διπλασιασμένων AARS είναι σημαντική για την έρευνα στην ανάπτυξη νέας γενιάς αντιβιοτικών. Στην παρούσα εργασία αναπτύξαμε μια σειρά από ευαίσθητα μοντέλα (Hidden Markov Models) ανίχνευσης αυτών των πρωτεϊνών σε προκαρυωτικά πρωτεώματα και στην συνέχεια σαρώσαμε ~2000 προκαρυωτικά πρωτεώματα και εντοπίσαμε διπλασιασμούς και απώλειες.

1.2 Πρωτεϊνοσύνθεση

Μία από τις σπουδαιότερες διεργασίες των κυττάρων, είναι η δημιουργία-παραγωγή των πρωτεϊνών, που παίζουν ρόλο δομικό, ενζυμικό, ορμονικό, επικοινωνιακό κ.α. Η σύνθεση των πρωτεϊνών είναι παρόμοια σε όλα τα βασίλεια του έμβιου κόσμου και αυτό αποτελεί μία ένδειξη ότι η πρωτεϊνοσύνθεση εμφανίσθηκε πολύ νωρίς κατά την εξέλιξη της ζωής. Ακόμη, λόγω της σπουδαιότητας του ρόλου της πρωτεϊνοσύνθεσης, είναι σημαντικό, η πολύπλοκη αυτή διεργασία να είναι όσο το δυνατόν πιο ακριβής, αλλά και γρήγορη, καθώς έτσι εξασφαλίζεται ότι τα κύτταρα θα παράγουν όταν χρειαστεί λειτουργικές πρωτεΐνες στην απαραίτητη ποσότητα. Η συχνότητα εισαγωγής ενός λάθους αμινοξέους έχει υπολογιστεί ότι είναι $\sim 10^{-4}$ (Edelmann & Gallant, 1977). Η ακρίβεια αυτή, επιτυγχάνεται με 2 στάδια αναγνώρισης:

α) το σωστό αμινοξύ πρέπει να βρεθεί και να συνδεθεί η καρβοξυλική του ομάδα με το 3'- άκρο ενός μορίου μεταφορικού RNA (tRNA)

β) το αντικωδικόνιο του ενεργού $tRNA^{aa}$ πρέπει να αναγνωρίσει το κωδικόνιο του mRNA

Μία ομάδα ενζύμων που ονομάζονται συνθετάσες του αμινοάκυλο-tRNA (aminoacyl-tRNA synthetases: AARS), έχουν πρωταγωνιστικό ρόλο στο πρώτο στάδιο.

1.3 Συνθετάσες των αμινοακυλο-tRNA (AARSs)

Ο ρόλος αυτών των ενζύμων είναι α) να προσάρτουν κάθε αμινοξύ στο αντίστοιχο μόριο tRNA (που να έχει το κατάλληλο αντικωδικόνιο) και β) να καταλύουν την ενεργοποίηση των αμινοξέων, ώστε να είναι θερμοδυναμικά εφικτός ο σχηματισμός του πεπτιδικού δεσμού μεταξύ δύο αμινοξέων.

Η ενεργοποίηση των αμινοξέων, γίνεται μέσα από δύο διαδοχικές αντιδράσεις που καταλύονται από τις AARS, όπου α) αρχικά σχηματίζεται αμινοακυλοαδενυλικό από την αντίδραση ενός αμινοξέος και ενός ATP και β) στη συνέχεια γίνεται η μεταφορά της αμινοακυλικής ομάδας σε ένα συγκεκριμένο μόριο tRNA, για το σχηματισμό του αμινοάκυλο –tRNA (aa-tRNA) με εστεροποίηση. Στη συνέχεια το αμινοάκυλο-tRNA, θα συνεχίσει την πορεία του προς τα ριβοσώματα, για τη συνέχεια της πρωτεϊνοσύνθεσης.

A) Αμινοξύ + ATP -> αμινοακυλο-AMP + PP_j

B) Αμινοάκυλο-AMP + tRNA -> αμινοάκυλο-tRNA +AMP

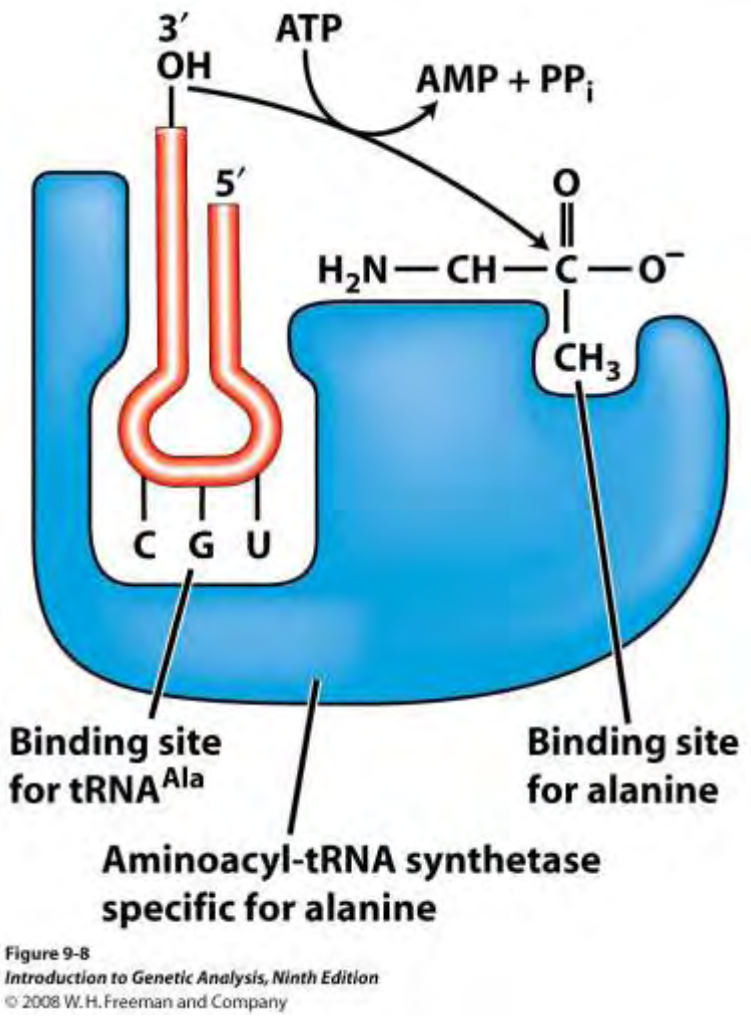


Figure 9-8
Introduction to Genetic Analysis, Ninth Edition

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

Και τα δύο βήματα καταλύονται από την ίδια AARS για το συγκεκριμένο αμινοξύ. Το κάθε ένα από τα 20 αμινοξέα αναγνωρίζεται από μία συγκεκριμένη AARS. Μάλιστα ο Crick (Crick's Adaptor Hypothesis, 1958) είχε προτείνει ότι το κάθε αμινοξύ ενώνεται με το συγκεκριμένο tRNA για το σχηματισμό του αμινοάκυλο-tRNA, μέσω της δικής του AARS. Ωστόσο, έχουν βρεθεί μικροοργανισμοί που διέθεταν αμινοακυλιωμένα tRNA, χωρίς να έχουν βρεθεί οι AARSs (Ibba *et al.*, 2000; Ibba & Soll, 2004).

1.4 Κλάσεις των συνθετασών των αμινοάκυλο- tRNA

Οι AARSs χωρίζονται σε δύο τάξεις , που ονομάζονται τάξη I και τάξη II και η κάθε μία περιέχει από δέκα AARS. Ο διαχωρισμός έγινε με βάση τα δομικά χαρακτηριστικά τους, των συντηρημένων αλληλουχιών τους καθώς και της θέσης πρόσδεσης του αμινοξέος.

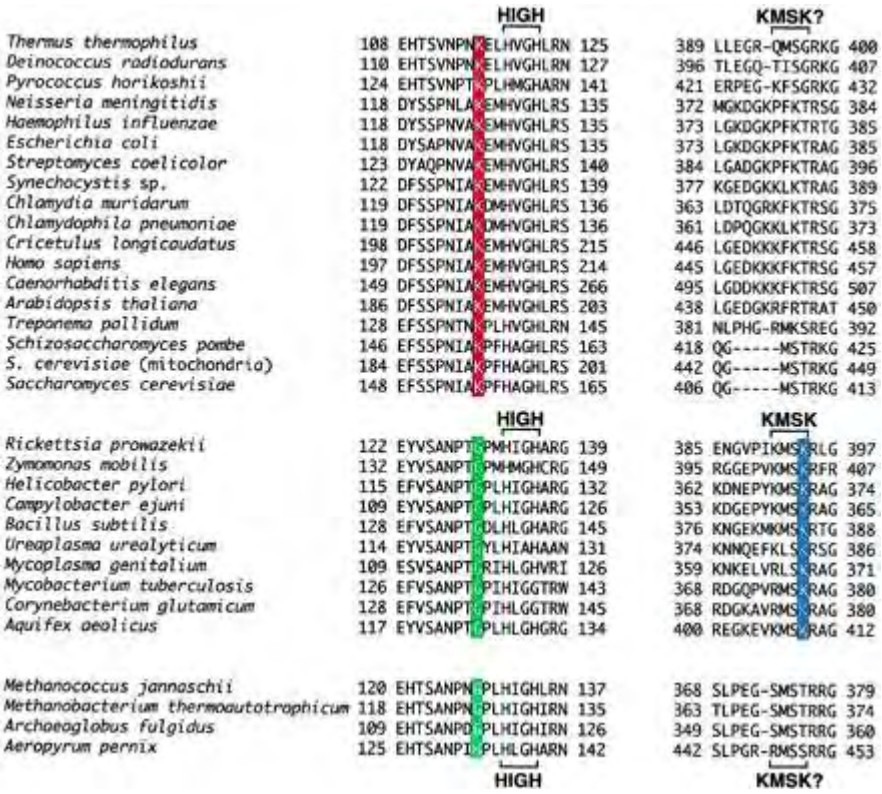
1.4.1 Συνθετάσες τάξης I

Η ομάδα των AARSs της τάξης I περιέχει 10 ένζυμα, με τα περισσότερα από αυτά να είναι μονομερή και δύο να είναι διμερή (Πίνακας 1.1). Ο ρόλος τους είναι να συνδέουν το αμινοξύ με το κατάλληλο μόριο tRNA, μέσω ενός εστερικού δεσμού στο 2΄- ΟΗ άκρο της 3΄- αδενοσίνης. Γενικά υπάρχουν διαφορές μεταξύ των συνθετασών της τάξης I, ωστόσο όλες χαρακτηρίζονται από ένα χαρακτηριστικό N-τελικό domain το οποίο είναι αυτό που περιέχει το ενεργό κέντρο της αδενυλίωσης. Πιο συγκεκριμένα η περιοχή χαρακτηρίζεται από μια συνδεδεμένη με ATP Rossman αγκύλη που περιέχει δύο συντηρημένα μοτίβα (HIGH και KMSKS), τα οποία αποτελούν ένα κοινό χαρακτηριστικό AARSs της τάξης I. Ωστόσο, η συντήρηση των HIGH και KMSKS δεν είναι υψηλή, π.χ. στο HIGH μόνο η πρώτη ιστιδίνη και η γλυκίνη δείχνουν μεγάλη συντήρηση, ενώ στο KMSKS υπάρχει ακόμη μικρότερη συντήρηση, με μόνο την δεύτερη λυσίνη να παρουσιάζει υψηλή συντήρηση (Moras, 1992).

AARSs τάξης 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 τάξης I και ο διαχωρισμός τους σε 3 ομάδες με βάση τις δομικές διαφορές τους (Eriani *et al.*, 1990; Moras, 1992)

Στην απόσταση που υπάρχει μεταξύ των HIGH και KMSKS περιοχών, υπάρχουν δύο μη συντηρημένες περιοχές, γνωστές ως connective peptide 1 και 2. Το πρώτο βρίσκεται μεταξύ του τέλους του πρώτου μισού του Rossman fold και το δεύτερο εντοπίζεται μεταξύ του strand D και την αρχή του δεύτερου μισού του Rossman fold (Delarue & Moras, 1993). Το μήκος των δύο πεπτιδίων διαφέρει μεταξύ των συνθετασών και για αυτό θεωρείται ότι αυτές οι δύο περιοχές ευθύνονται για τις διαφορές στην τεταρτοταγή δομή των ενζύμων της τάξης I (Fourmy *et al.*, 1995).



Εικόνα 1.2 Τα HIGH και KMSKS μοτίβα που βρέθηκαν με πολλαπλή στοίχιση σε ακολουθίες της ArgRS (Sekine *et al.*, 2001).

1.4.2 Συνθετάσες της τάξης ΙΙ

Η ομάδα των AARSs της τάξης ΙΙ περιέχει και αυτή 10 ένζυμα. Τα εφτά από αυτά είναι διμερή και τα τρία είναι τετραμερή (Πίνακας 1.2).

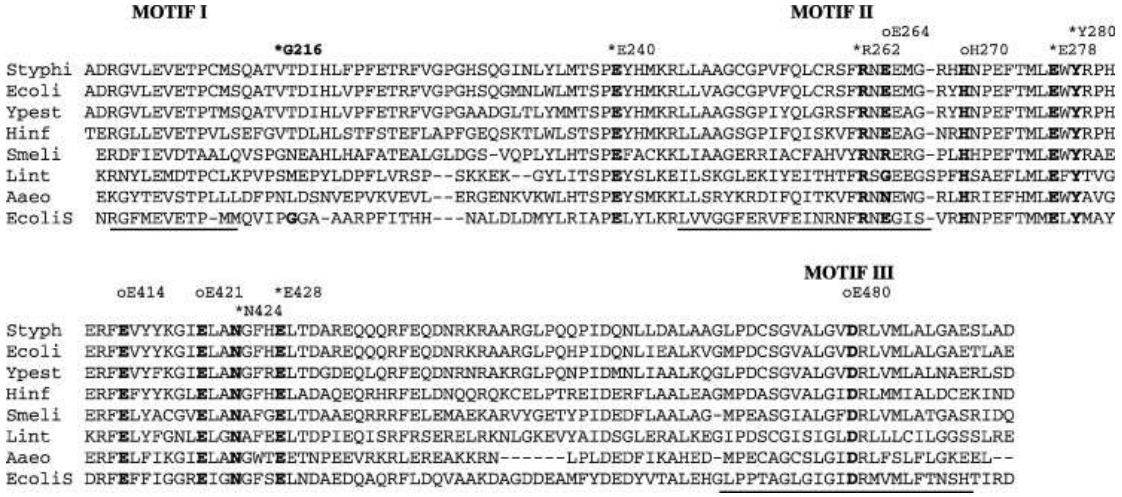
AARSs τάξης ΙΙ		
IIa	IIb	IIc
His α_2	Asp α_2	Gly α_2/b_2
Pro α_2	Asn α_2	Ala α_4
Ser α_2	Lys α_2	Phe α_2/ b_2
Thr α_2		

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

Οι συνθετάσες της τάξης ΙΙ συνδέουν τα αμινοξέα με το κατάλληλο tRNA στο 3΄- υδροξυλικό άκρο της 3΄-αδενοσίνης (με εξαίρεση την PheRS). Επιπλέον, όπως και στις συνθετάσες της τάξης Ι, υπάρχουν διαφορές μεταξύ τους, ωστόσο μοιράζονται τουλάχιστον δύο από τρία συντηρημένα μοτίβα, όπου το ATP συνδέεται και γίνεται η ενεργοποίηση του αμινοξέος. Το ενεργό κέντρο βρίσκεται "χωμένο" βαθιά στο καρβοξυτελικό

άκρο του ενζύμου και λόγω τη τοποθεσίας του, τα αμινοξέα που ενεργοποιούνται από τις συνθετάσες της τάξης ΙΙ τείνουν να είναι μικρότερα από αυτά της τάξης Ι (Delarue & Moras, 1993).

Τα τρία αυτά μοτίβα, σχηματίζουν το domain πρόσδεσης και το ενεργό κέντρο της συνθετάσης. Όλα μαζί τα μοτίβα περιέχουν 250 αμινοξέα και σχηματίζουν ένα αντι-παράλληλο βήτα πτυχωτό φύλλο.



Εικόνα 1.3 Τα τρία μοτίβα που εμφανίζονται στις συνθετάσες της τάξης ΙΙ (Ambrogelly *et al.*, 2010)

Το μοτίβο 1 πάντα τοποθετείται περίπου 50 αμινοξέα πριν το μοτίβο 2. Το μοτίβο 3, συχνά εντοπίζεται κοντά στο τέλος του καρβοξυτελικού άκρου και είναι αρκετά μεταβλητό, καθώς εντοπίζεται 150-250 αμινοξέα μακριά από το μοτίβο 2. Αυτή η απόσταση επιτρέπει στην περιοχή να μην έχει μεγάλη συντήρηση, κάτι το οποίο συμβάλλει στις διαφορετικές δομές των ενζύμων που βρίσκονται στην τάξη ΙΙ (Delarue & Moras, 1993).

1.5 LysRS: μία AARS που βρίσκεται και στις δύο τάξεις

Μετά από πειράματα στον οργανισμό Methanococcus maripaludis, βρέθηκε ένα γονίδιο (lysK) που κωδικοποιούσε μία LysRS που άνηκε στην τάξη Ι, σε αντίθεση με όλες τις προηγούμνες LysRS που είχαν βρεθεί και άνηκαν στην τάξη ΙΙ. Τελικά, μετά από γενομική ανάλυση, βρέθηκε ότι το γονίδιο (lysS) που κωδικοποιούσε την LysRS της τάξης ΙΙ, έλειπε από την πληοψηφία των αρχαίων και κάποιων βακτηρίων και αντί για αυτό είχαν το lysK (Ibba *et al.*, 1997). Έπειτα από μελέτες πάνω στη λειτουργία (Ibba *et al.*, 1999) και τη δομή (Terada *et al.*, 2002) της LysRS, φάνηκε ότι οι δύο συνθετάσες έχουν παρόμοια λειτουργία αλλά διαφορετική δομή.

1.6 Περισσότερες από 20 AARSs

Αν και αρχικά επικρατούσε η άποψη ότι υπάρχουν μόνο 20 AARS, διάφορες μελέτες έδειξαν ότι συχνά σε ένα γονιδίωμα μπορεί να υπάρχουν περισσότερες από 20. Γενικά οι AARS έχουν πολύ υψηλή διακριτική ικανότητα, ώστε να επιλέγουν πάντα το σωστό αμινοξύ για το κατάλληλο tRNA, καθώς η "φόρτιση" ενός

λάθους αμινοξέος μπορεί να γίνει πολύ σπάνια (Ibba & Soll, 1999). Ωστόσο, υπάρχει μία ομάδα AARSs με μικρή εξειδίκευση που είναι αρκετά σημαντική στην πρωτεϊνοσύνθεση (Ibba & Soll, 1999). Πρόκειται για τα μη διακριτικά (non-discriminating) aspartyl-tRNA (ND-AspRS) και glutamyl-tRNA (ND-GluRS), ένζυμα με χαμηλή εξειδίκευση τα οποία είναι απαραίτητα για την δημιουργία των Asn-tRNA^{Asn} και Gln-tRNA^{Gln}. Η διαφορά τους από τα διακριτικά (με υψηλή εξειδίκευση) αντίστοιχα ένζυμα αναλύεται παρακάτω.

Το ND-AspRS μπορεί να δημιουργήσει τόσο τα Asp-tRNA^{Asp} και Asp-tRNA^{Asn} με aspartate, ενώ το διακριτικό AspRS (D-AspRS) δημιουργεί μόνο το Asp-tRNA^{Asp} (Curnow et al., 1998; Becker & Kern, 1998). Το ίδιο κάνει και το ND-GluRS, το οποίο μπορεί να σχηματίσει Glu-tRNA^{Gln} εκτός από Glu-tRNA^{Glu} (Lapointe *et al*., 1986; Wilcox &Nirenberg, 1968). Έπειτα, το γλουταμινικό αντικαθίσταται από τη γλουταμίνη με μία αντίδραση που γίνεται από ένα ένζυμο που ονομαζέται αμιδινοτρανσφεράση του γλουταμινικού (Glu-AdT) (Curnow *et al.*, 1997; Raczniak *et al*., 2001). Τέλος, έχουν βρεθεί ένζυμα με ενεργότητα συνθετάσης της cysteinyl-tRNA, παρόλο που από τον συγκεκριμένο οργανισμό (*Methanococcus jannaschii*) έλειπε η συγκεκριμένη συνθετάση. Ενδιαφέρουσες είναι και οι περιπτώσεις μεταξύ των CysRS και ProRS (Jacquin-Becker *et al*., 2002) π.χ βρέθηκε μία πρωτεΐνη που έμοιαζε με την συνθετάση της prolyl-tRNA και μπορούσε να συνθέσει cysteinyl-tRNA και prolyl-tRNA. (Stathopoulos *et al.*, 2000). Ακόμη, παρατηρήθηκε ότι το πρώτο στάδιο στην ενεργοποίηση των δύο αμινοξέων, ήταν διαφορετικό (Stathopoulos *et al*., 2001).

Αριθμός EC	Όνομα Ενζύμου
6.1.1.1	Tyrosine--tRNA ligase
6.1.1.2	Tryptophan--tRNA ligase
6.1.1.3	Threonine--tRNA ligase
6.1.1.4	Leucine--tRNA ligase
6.1.1.5	Isoleucine--tRNA ligase
6.1.1.6	Lysine--tRNA ligase
6.1.1.7	Alanine--tRNA ligase
6.1.1.9	Valine--tRNA ligase
6.1.1.10	Methionine--tRNA ligase
6.1.1.11	Serine--tRNA ligase
6.1.1.12	Aspartate--tRNA ligase
6.1.1.13	D-alanine--poly(phosphoribitol) ligase
6.1.1.14	Glycine--tRNA ligase
6.1.1.15	Proline--tRNA ligase
6.1.1.16	Cysteine--tRNA ligase
6.1.1.17	Glutamate--tRNA ligase
6.1.1.18	Glutamine--tRNA ligase
6.1.1.19	Arginine--tRNA ligase
6.1.1.20	Phenylalanine--tRNA ligase
6.1.1.21	Histidine--tRNA ligase
6.1.1.22	Asparagine--tRNA ligase
6.1.1.23	Aspartate--tRNA(Asn) ligase
6.1.1.24	Glutamate--tRNA(Gln) ligase
6.1.1.25	Lysine--tRNA(Pyl) ligase
6.1.1.26	Pyrrolysine--tRNA(Pyl) ligase
6.1.1.27	O-phosphoserine--tRNA ligase

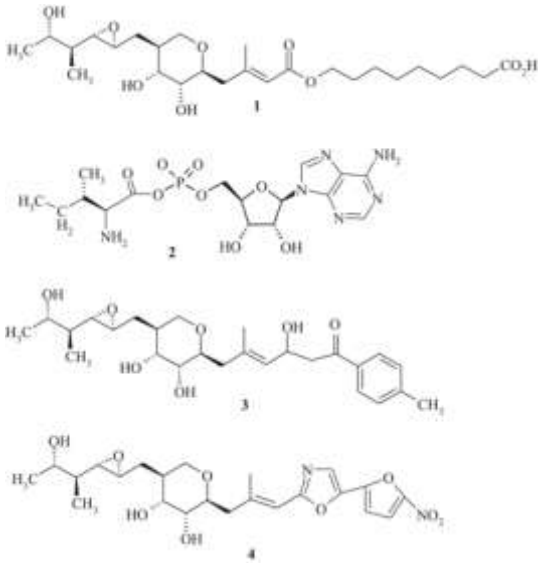
Πίνακας 1.3 Ένζυμα με λειτουργία AARS (http://enzyme.expasy.org/cgi-bin/enzyme/enzyme-search-ec)

1.7 Συνθετάσες και αντιβιοτικά

Η δραματική αύξηση της αντίστασης ορισμένων οργανισμών στα αντιβιοτικά, αποτελεί μία μεγάλη απειλή για την ανθρώπινη κοινωνία και για αυτό το λόγο μεγαλώνει καθημερινά η ανάγκη για την ανακάλυψη/δημιουργία νέας γενιάς αντιβιοτικών. Αρκετοί οργανισμοί ήδη έχουν εμφανίσει αυξημένη αντίσταση απέναντι στα αντιβιοτικά π.χ. *Staphylococcus aureus* (MRSA) στη methicillin, ο *Streptococcus pneumoniae* πενικιλλίνη και ο *Enterococcus* στη vancomycin

Μέσα από έρευνες που έχουν γίνει σε φυσικά προϊόντα που λειτουργούν ως αντιβιοτικά, φάνηκε ένα μοτίβο. Αρκετές ουσίες στόχευαν στο μηχανισμό της μετάφρασης. (Kim *et al.,* 2003). Ουσίες που στόχευαν στην αναστολή των AARSs φάνηκε ότι είχαν πολύ καλή προοπτική στην έρευνα ανάπτυξης νέων αντιβιοτικών, καθώς έχουν αρκετά πλεονεκτήματα. Λόγω της εξελικτικής απόκλισης που υπάρχει μεταξύ των ευκαρυωτικών και προκαρυωτικών AARSs, θα μπορούσαν να δημιουργηθούν αντιβιοτικά που να "χτυπάνε" σε κάθε ένα από τις 20 διαφορετικές βακτηριακές AARSs, χωρίς να επηρεάζουν την φυσιολογική λειτουργία του ανθρώπινου οργανισμού. Ακόμη, έχουν ήδη ανακαλυφθεί φυσικά προϊόντα (θα αναφερθούν παρακάτω) που λειτουργούν ως αναστολείς των βακτηριακών AARSs, και θα μπορούσαν να χρησιμοποιηθούν ως μοντέλα για την δημιουργία τέτοιων ουσιών στο εργαστήριο. Τέλος, λόγω της αυξημένης συντήρησης που παρουσιάζει αυτή η ομάδα πρωτεϊνών, ένα μόνο αντιβιοτικό θα μπορούσε να προσβάλει περισσότερες από μία AARS (Hurdle *et al.,* 2005).

Μία τέτοια ουσία που ήδη υπάρχει στο εμπόριο και λειτουργεί ως αναστολέας των AARS είναι η mupirocin (pseudomonic acid). Πρόκειται για ένα φυσικό προϊόν που απομονώθηκε από τον οργανισμό *Pseudomonas flurescens* και λειτουργεί ως αντιβιοτικό ενάντια στη βακτηριακή IleRS. Μάλιστα, έχει φανεί ότι είναι ιδιαίτερα αποτελεσματική ενάντια σε παθογόνα gram θετικά βακτήρια και χρησιμοποιείται σε παγκόσμια κλίμακα για την καταπολέμηση του MRSA (Hurdle *et al.,* 2005; Boyce, 2001).



Εικόνα 1.4 : Χημική δομή της mupirocin (δομή 1), isoleucyl-AMP (Ile-AMP; δομή 2), και παράγωγα της mupirocins (δομές 3 και 4) (Hurdle *et al.,* 2005).

Μέχρι στιγμής έχουν βρεθεί κι άλλες ουσίες που λειτουργούν ως αναστολείς των AARSs, ωστόσο, κάποιες από αυτές επηρεάζουν και τους ανθρώπους, αλλά υπήρχαν και κάποιες που έδειξαν μεγάλη ακρίβεια απέναντι στις βακτηριακές AARSs (Gallant *et al.,* 2000; Kim *et al.,* 2003; Pohlmann & Brotz-Oesterhelt, 2004; Schimmel *et al.,* 1998). π.χ. η chuangxinmycin και η indolmycin, οι οποίες παράγονται από τους *Actinoplanes tsinanensis* και *Streptomyces griseus* ATCC 12648 αντίστοιχα. Αυτά τα προϊόντα μπορούν να αναστείλουν την λειτουργία της συνθετάσης tryptophanyl-tRNA (Routien, 1966; Brown *et al.,* 2002).

Βέβαια θα πρέπει επίσης να αναφερθεί ότι οι κάποιοι μικροοργανισμοί μπορούν να εμφανίσουν αντίσταση απέναντι σε αυτά τα αντιβιοτικά. Αν κάποια ουσία δρα ως αναστολέας απέναντι σε κάποια AARS, τότε αν ο οργανισμός διαθέτει ένα διπλασιασμένο γονίδιο αυτής της συνθετάσης, μπορεί να ανταπεξέλθει απέναντι στην ουσία αυτή. Ένα παράδειγμα τέτοιας δράσης είναι το παρακάτω. Το διπλασιασμένο γονίδιο μίας AARS (TrpRS2), προσέφερε ένα πλεονέκτημα. Του προσέδιδε αντίσταση τόσο στην chuangxinmycin όσο και στην indolmycin, τα οποία είναι αντιβιοτικά που παράγονται από άλλα είδη και απενεργοποιούν την TrpRS1 (Kitabatake *et al.,* 2002; Vecchione & Sello, 2009). Επιπλέον, έχει φανεί ότι οι διπλασιασμένες/βοηθητικές AARSs προσφέρουν προστασία όχι μόνο ενάντια σε ξένα αντιβιοτικά, αλλά προσφέρουν προστασία και στην τοξίνη που παράγει ο ίδιος ο μικροοργανισμός (Yanagisawa & Kawakami, 2003). π.χ. το *P. fluroescens* έχει την ικανότητα να προστατεύει τον εαυτό του από αντιβιοτικά, μεταφέροντας στο γονιδίωμα του το γονίδιο μίας AARS, η οποία μοιάζει με αυτή των ευκαρυωτικών οργανισμών. Πιο συγκεκριμένα, μεταφέρει το γονίδιο IleRS-R2, μαζί με το IleRS-R1. Το γονίδιο IleRS-R2, το οποίο είναι αυτό που μοιάζει με το γονίδιο των ευκαρυωτικών, του προσδίδει ανθεκτικότητα ενάντια στο pseudomonic acid, μία τοξίνη που παράγει ο ίδιος ο οργανισμός (Yanagisawa & Kawakami, 2003).

Άλλο ένα παράδειγμα ανθεκτικότητας σε τοξίνες είναι αυτό του *Streptomyces sp*. strain ATCC 700974. Ο συγκεκριμένος στρεπτομύκητας διαθέτει δύο διαφοροποιημένα γονίδια για την SerRS. Το ένα από τα δύο γονίδια του προσφέρει αντίσταση στην albomycin (Zeng *et al.,* 2009). Ωστόσο, στο συγκεκριμένο άρθρο, ο συγγραφέας άφησε την πιθανότητα, ότι η το δεύτερο γονίδιο της SerRS, να εμπλέκεται στο στάδιο της βιοσύνθεσης της albomycin.

Όπως αναφέρθηκε και πιο πριν, ο ρόλος των αμινοάκυλο tRNA συνθετάσων είναι πολύ σημαντικός στην πρωτεϊνοσύνθεση και όχι μόνο. Γενικά πιστεύεται ότι η οικογένεια των AARS είναι ανάμεσα στις αρχαιότερες, καθώς χωρίς αυτές θα ήταν δύσκολη η μετάβαση από έναν κόσμο "RNA" σε έναν κόσμο "RNA-πρωτεΐνη" (Schimmel & Ribas De Pouplana, 2000). Το ότι εξελίχθηκαν πολύ νωρίς σημαίνει ότι είναι πάρα πολύ σημαντικά μόρια για την επιβίωση, καθώς χωρίς αυτές δεν ήταν εύκολο να δημιουργηθεί ο πεπτιδικός δεσμός, ούτε θα ήταν δυνατή η σύνδεση με το tRNA και τέλος δεν θα μπορούσαν να παραχθούν λειτουργικές πρωτεΐνες. Αυτό τις έχει βοηθήσει στην ανάπτυξη διακριτικής ικανότητας ανάμεσα σε 20 διαφορετικά αμινοξέα, καθώς επίσης και στην ανάπτυξη διάφορων λειτουργιών (π.χ με τα αντιβιοτικά που αναφέρθηκαν παραπάνω). Δεύτερον, έχει παρατηρηθεί ότι αρκετές πρωτεΐνες οι οποίες δεν έχουν την λειτουργία συνθετάσης, διαθέτουν αρκετά domains που παρατηρούνται στις συνθετάσες. Από αυτό μπορούμε να υποθέσουμε ότι έγιναν αρκετοί γονιδιακοί διπλασιασμοί, όπου τα domains των συνθετασών κρατήθηκαν και προσαρμόστηκαν σε καινούργιες λειτουργίες (Schimmel & Ribas De Pouplana, 2000). Επιπλέον, σύμφωνα με προηγούμενες μελέτες, οι AARS, έχουν την τάση για "οριζόντια μεταφορά" (Brown & Doolittle, 1999; Olendzenski *et al.,* 2000; Woese *et al.,* 2000; Dohm *et al.,* 2006; Luque *et al.,* 2008). Πολύ πιθανό αυτό να συμβαίνει λόγω του ότι έχουν λίγες φυσικές αλληλεπιδράσεις μέσα στο κύτταρο (αμινοξέα, ATP και tRNA) σε σχέση με άλλα ένζυμα που εμπέκονται στη μεταγραφή και στη μετάφραση (Wolf *et al.,* 1999).

Για να πραγματοποιηθεί η βιοπληροφορική ανάλυση αυτής της εργασίας, χρησιμοποιήθηκαν μια σειρά από Βάσεις Δεδομένων και εργαλεία Βιοπληροφορικής, για τα οποία θα δωθεί μια σύντομη εισαγωγή (για το κάθε ένα) παρακάτω.

1.8 Το CDD

Το CDD είναι μια βάση δεδομένων που περιέχει πρωτεϊνικές αλληλουχίες, στις οποίες έχει γίνει πολλαπλή στοίχιση που έχει ελεγχθεί και διορθωθεί από επιστήμονες. Έπειτα, αυτές οι στοιχίσεις χρησιμοποιήθηκαν για την δημιουργία στατιστικών μοντέλων είτε για αρχαία domains είτε για ολόκληρες πρωτεΐνες. Μάλιστα, τα μοντέλα αυτά είναι διαθέσιμα ως position specific scoring matrices (PSSMs), ώστε να μπορούν να χρησιμοποιηθούν για εύκολη και γρήγορη αναγνώριση των συντηρημένων δομικών περιοχών (conserved domains) σε πρωτεΐνες, μέσω του προγράμματος RPS-BLAST. Ακόμη, μέσα στις πληροφορίες που βρίσκονται στο CDD, υπάρχουν οι δομικές περιοχές τις οποίες επιμελήθηκε το NCBI και περιέχουν πληροφορίες για την τρισδιάστατη δομή των πρωτεϊνών, ώστε να μπορούν να καθοριστούν με μεγαλύτερη ακρίβεια τα όρια των domains.

1.9 Profile HMMs

Τα profile HMMs είναι στατιστικά μοντέλα που προέρχονται από πολλαπλές στοιχίσεις, αλλά μπορούν να δημιουργηθούν ακόμα και από μία ακολουθία. Διαθέτουν πληροφορίες που είναι ειδικές για τη θέση και αφορούν το πόσο καλά συντηρημένες είναι οι διάφορες θέσεις μίας στοίχισης (Eddy, 1998). Τα profile HMM έγιναν γνωστά στην βιοπληροφορική λόγω των Anders Krogh, David Haussler και των συνεργατών τους στο US Santa Cruz (Krogh *et al.,* 1994), οι οποίοι εφάρμοσαν σε προβλήματα Βιοπληροφορικής τα HMMs, που τότε χρησιμοποιούνταν στην αναγνώριση φωνής. Η αλήθεια είναι ότι τα HMM είχαν χρησιμοποιηθεί και παλαιότερα στην βιολογία, από τον Gary Churchill (Churchill, 1989), ωστόσο, ο Krogh και η ομάδα του, δημιουργώντας τα profile HMM, κατάφεραν να τα κάνουν πολύ πιο ελκυστικά στη βιολογία, καθώς μπορούσαν να χρησιμοποιηθούν με την ίδια μέθοδο των "profile" που ήταν ήδη γνωστή π.χ. για αναζήτηση σε βάσεις δεδομένων χρησιμοποιώντας πολλαπλές στοιχίσεις ακολουθιών, αντί για μόνο μία ακολουθία.

Τα προφιλ έγιναν γνωστά από τον Gribskov και τους συνεργάτες του (Gribskov *et al.,* 1987), καθώς και από άλλες ομάδες που δούλευαν σε παρόμοια projects (π.χ. flexible patterns - Barton, 1990). Αποτελούν και αυτά μία στατιστική περιγραφή της συναίνεσης μίας πολλαπλής στοίχισης και χρησιμοποιούν position-specific-scores για αμινοξέα ή νουκλεοτίδια, καθώς επίσης και αρνητικούς βαθμούς για κενά.

Τα πλεονεκτήματα των HMMs είναι ότι:

- Χρησιμοποιούν επίσημη πιθανολογική βάση (formal probalistic basis). Με τη βοήθεια της θεωρίας των πιθανοτήτων γίνεται η καθοδήγηση για πως θα ρυθμιστεί ο πίνακας βαθμονόμησης.
- Υπάρχει μίας συνεπής θεωρίας για τη βαθμολόγηση των κενών και των εισαγωγών τους.
- Λόγω της συνέπειας των μεθόδων, μπορεί να γίνει αυτοματοποίησή τους και έτσι είναι δυνατή η δημιουργία βιβλιοθηκών που να περιέχουν εκατοντάδες ή και χιλιάδες HMM και έπειτα είναι δυνατή η χρησιμοποίηση τους για την ανάλυση ολόκληρων γονιδιωμάτων. Μία γνωστή βάση δεδομένων για profile HMM είναι η Pfam.

1.10 Pfam

Γενικά οι πρωτεΐνες αποτελούνται από μία ή περισσότερες λειτουργικές περιοχές (domains). Ανάλογα με τον συνδυασμό των domains που περιέχει μία πρωτεΐνη, έχει και διαφορετικές λειτουργίες. Χάρη στην βιοπληροφορική ανίχνευση αυτών των περιοχών, μπορεί να γίνει πιο εύκολη η εύρεση/κατανόηση των λειτουργιών μίας πρωτεΐνης. Το Pfam είναι μία βάση δεδομένων που διαθέτει μία μεγάλη συλλογή από domains, τα οποία μπορούν να βρεθούν είτε με τη μορφή πολλαπλής στοίχισης είτε ως Hidden Markov Models (HMMs) και να χρησιμοποιηθούν για την αναγνώριση άλλων πρωτεϊνών (Punta *et al*., 2012).

1.11 Πως λειτουργούν τα HMM

Τα Μαρκοβιανά μοντέλα (γνωστά και ως Μαρκοβιανές αλυσίδες, Markov chains), είναι στατιστικά μοντέλα τα οποία είναι ικανά να περιγράψουν μία ακολουθία γεγονόνων, τα οποία συμβαίνουν το ένα μετά το άλλο, σαν να βρίσκονται δηλαδή σε μία αλυσίδα. Με αυτόν τον τρόπο λοιπόν, κάθε γεγονός που συμβαίνει επηρεάζει τη πιθανότητα του επόμενου γεγονότος στην αλυσίδα. Επειδή λοιπόν οι βιολογικές ακολουθίες γράφονται σε σειρές γραμμάτων, μπορούν να περιγραφούν μέσω των Μαρκοβιανών μοντέλων, δηλαδή μπορεί να υπολογιστεί η πιθανότητα αλλαγής ενός κατάλοιπου αμινοξέος από ένα άλλο.

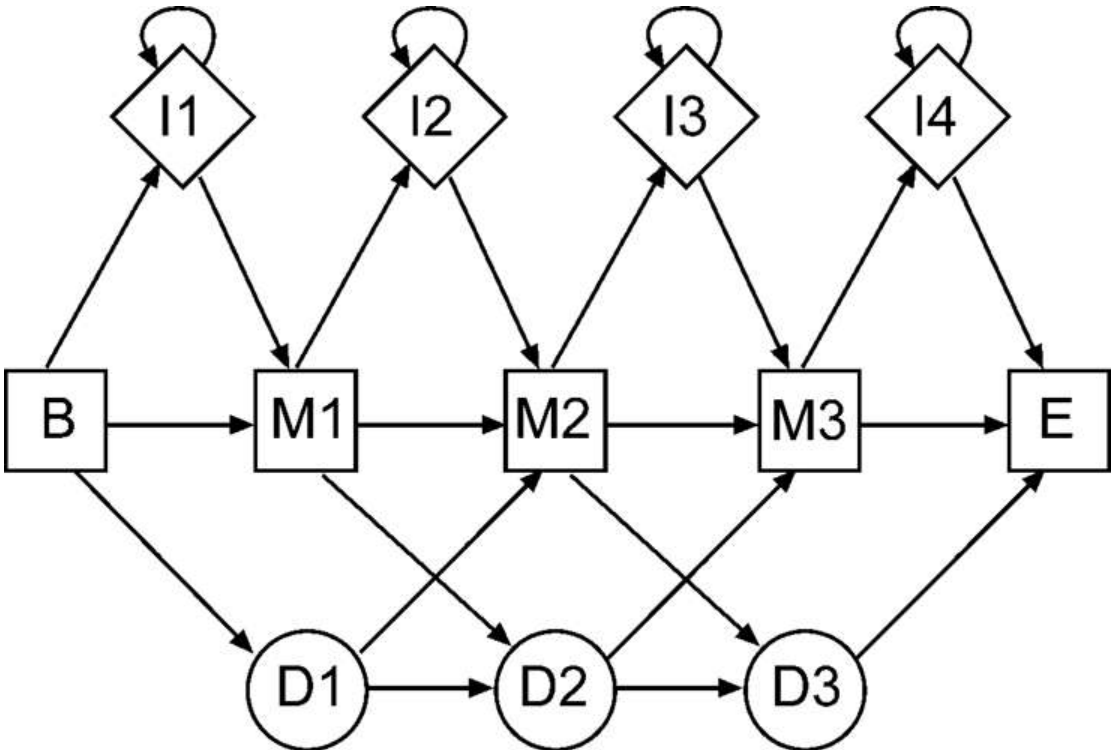
Ο πληθυντικός χρησιμοποιείται γιατί υπάρχουν διαφορετικοί τύποι Μαρκοβιανών μοντέλων, που ο καθένας χρησιμοποιείται για περιγράψει ένα σύνολο δεδομένων διαφορετικής πολυπλοκότητας. Μερικά από αυτά είναι τα εξής:

- Zero_order Markov model: Χρησιμοποιείται να περιγράψει τη πιθανότητα μίας κατάστασης ανεξάρτητα από την προηγούμενη.
- First-order Markov model: Χρησιμοποιείται για να περιγράψει τη πιθανότητα μίας κατάστασης, σε σχέση με το προηγούμενο γεγονός.
- Second-order Markov model: Χρησιμοποιείται για να περιγράψει την πιθανότητα ενός γεγονότος, η οποία καθορίζεται από τις προηγούμενες δύο καταστάσεις (π.χ. μπορεί να περιγράψει ένα κωδικόνιο στις βιολογικές ακολουθίες.

Υπάρχουν και άλλες τάξεις Μαρκοβιανών μοντέλων που μπορούν να περιγράψουν πιο πολύπλοκες καταστάσεις.

Στα Μαρκοβιανά μοντέλα, όλες οι καταστάσεις μίας γραμμικής ακολουθίας, μπορούν να παρατηρήθουν άμεσα, ωστόσο υπάρχουν και περιπτώσεις που δεν μπορούν να παρατηρηθούν όλοι οι παράγοντες που μπορούν να επηρεάσουν την αλλαγή μίας κατάστασης. Για να μπορέσει λοιπόν κάποιος να υπολογίσει και αυτούς του παράγοντες, πρέπει να χρησιμοποιήσει τα Hidden Markov Models (HMMs), τα οποία μπορούν να συνδυάσουν διαφορετικά Μαρκοβιανά μοντέλα. Π.χ. το ένα μοντέλο μπορεί να χρησιμοποιείται για να περιγράψει τα γεγονότα που "φαίνονται" και το άλλο μοντέλο μπορεί να περιγράψει τους "κρυφούς" παράγοντες που μπορούν να επηρεάσουν την μετάβαση από τη μία κατάσταση σε μία άλλη. Ένα παράδειγμα στις βιολογικές ακολουθίες που μπορεί να επηρεάσει τη μετάβαση από μία κατάσταση σε μία άλλη και να θεωρηθεί "κρυφός παράγοντας" είναι τα κενά (gaps).

Για την δημιουργία των HMM, πρέπει να γίνει η εκπαίδευση τους (training), όπου χρησιμοποιούνται οι ακολουθίες κάποιων ομόλογων AARS και υπολογίζονται οι πιθανότητες για την εμφάνιση του κάθε ενός από τα 20 αμινοξέα στην κάθε θέση της πολλαπλής στοίχισης.



Εικόνα 1.5: Ένα τυπικό profile HMM βασίζεται στην πολλαπλή στοίχιση ακολουθιών. Τα τετράγωνα αντιπροσωπεύουν το match (M), τα διαμάντια αντιπροσωπεύουν τα insertions (I) και οι κύκλοι τα deletions (D). Η αρχή και το τέλος χαρακτηρίζονται από τα B και E και το κάθε γεγονός αντιπροσωπεύεται από ένα βέλος με μία τιμή πιθανότητας να συμβεί το κάθε στάδιο (Essential Bioinformatics).

1.12 HMMER

Πρόκειται για ένα πακέτο προγραμμάτων που μπορούν να χρησιμοποιηθούν για την κατασκευή και χρήση των profile HMMs. Δύο από τα γνωστά προγράμματα του HMMER είναι το hmmbuild και το hmmscan. Το hmmbuild χρησιμοποιείται για να δημιουργήσει από ένα αρχείο πολλαπλής στοίχισης ένα αρχείο HMM που περιέχει τις βαθμολογίες για το κάθε αμινοξύ ή νουκλεοτίδιο, ενώ το hmmscan χρησιμοποιείται για να ψάξει μία ακολουθία σε μία βάση δεδομένων με ένα ή περισσότερα profile HMMs (Eddy, 1998; http://hmmer.org/).

Παρακάτω δίδεται μια λίστα με κάποια από τα σημαντικότερα προγράμματα του HMMER

- **Phmmer:** ψάχνει μία ακολουθία σε μία βάση δεδομένων, όπως και το BLASTP.
- **Jackhammer:** (PSIBLAST-like).
- **Hmmbuild:** για τη δημιουργία ενός profile HMM από ένα αρχείο msa.

- **Hmmsearch:** ψάχνει ένα profile HMM σε μία βάση δεδομένων με ακολουθίες.
- **Hmmscan:** ψάχνει μία ακολουθία σε μία βάση δεδομένων με profile HMM.
- **Hmmalign:** κάνει πολλαπλή στοίχιση μίας ή περισσότερων ακολουθιών έχοντας ώς βάση την πολλαπλή στοίχιση ενός profile HMM.
- **Hmmfetch:** παίρνει ένα profile HMM από το όνομα του αρχείου ή από το accession από μία βάση δεδομένων με HMM
- **Hmmpress:** μετατρέπει μία βάση δεδομένων με HMM σε binary format (για το hmmscan)
- **Hmmstat:** εμφανίζει τα στατιστικά του κάθε profile που βρίσκεται σε μία βάση δεδομένων HMM

1.13 Το Blastclust

Πρόκειται για ένα πρόγραμμα του πακέτου προγραμμάτων BLAST, το οποίο χρησιμοποιείται για τη δημιουργία clusters πρωτεϊνικών ή νουκλεοτιδικών ακολουθιών. Αυτό που κάνει είναι να παίρνει τις ακολουθίες και αν βλέπει ότι υπάρχει ομολογία σύμφωνα με αυτήν που του έχει ρυθμίσει ο χρήστης, τότε τοποθετεί τις ακολουθίες στην ίδια ομάδα (ίδιο cluster). Για τις πρωτεϊνικές ακολουθίες χρησιμοποιεί τον αλγόριθμο του blastp για να βρει τον βαθμό ομοιότητας, ενώ για τις νουκλεοτιδικές ακολουθίες χρησιμοποιεί τον αλγόριθμο του Megablast (http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html).

1.14 Πολλαπλή Στοίχιση – Multiple Sequence Alignment (MSA)

Όταν έχουμε μία αμινοξική ακολουθία και θέλουμε να δούμε ποιες είναι οι πιθανές της λειτουργίες, συνήθως προσπαθούμε να βρούμε αν έχει ομοιότητα με κάποια άλλη πρωτεΐνη, για την οποία είναι γνωστή η λειτουργία της. Αν βρεθεί μία ακολουθία η οποία μοιάζει με την δικιά μας, είτε σε όλο το μήκος της είτε σε κάποιες περιοχές μόνο, τότε υποθέτουμε ότι μπορεί να έχουν παρόμοιες λειτουργίες και ότι είναι ομόλογες (έχουν κοινή προέλευση). Το πιο πιθανό βέβαια είναι να μην βρούμε μόνο μία πρωτεΐνη που μοιάζει με τη δικιά μας αλλά αρκετές και μάλιστα όλες αυτές να ανήκουν στην ίδια οικογένεια πρωτεϊνών. Όταν έχουμε πάνω από δύο ακολουθίες, τότε κάνουμε μία πολλαπλή στοίχιση.

Γενικά η πολλαπλή στοίχιση χρησιμοποιείται όταν θέλουμε :

- Να βρούμε domains πρωτεϊνών π.χ. για τη δημιουργία profiles/motifs.
- Να βρούμε συντηρημένα DNA-binding sites σε προαγωγείς γονιδίων.
- Όταν θέλουμε να κάνουμε μία φυλογενετική ανάλυση (στηρίζεται πάνω στη πολλαπλή στοίχιση ομόλογων ακολουθιών).
- Όταν θέλουμε να προβλέψουμε την δευτεροταγή και τριτοταγή δομή μιας ακολουθίας.

Για την πραγματοποίηση μίας πολλαπλής στοίχισης, χρησιμοποιούνται ειδικά προγράμματα. Το κάθε ένα έχει τη δικιά του μέθοδο για την πραγματοποίηση της πολλαπλής στοίχισης. Γενικά μπορούν να χωριστούν στις κατηγορίες i) Δυναμικός προγραμματισμός (dynamic programming) & ii) Ευρετικές μέθοδοι (heuristics).

Μία από τις μεθόδους που χρησιμοποιούνται πιο συχνά για τη δημιουργία μιας πολλαπλής στοίχισης, είναι η προοδευτική πολλαπλή στοίχιση ακολουθιών, η οποία ανήκει στις ευρετικές μεθόδους. Αυτό που συμβαίνει, είναι ότι αρχικά στοιχίζονται οι ακολουθίες σε ζεύγη και για το κάθε ζεύγος υπολογίζεται ένας βαθμός ομοιότητας για το κάθε ζεύγος. Στη συνέχεια γίνεται η δημιουργία ενός δένδρου, το οποίο χρησιμοποιείται ως οδηγός, για τον καθορισμό της σειράς με την οποία θα στοιχιστούν οι ακολουθίες μεταξύ τους. Τέλος, παίρνει το ζεύγος των ακολουθιών με τον μεγαλύτερο βαθμό ομοιότητας και αρχίζει να βάζει τα ζεύγη των

ακολουθιών που είναι πιο απομακρυσμένα μεταξύ τους, μέχρι να τελειώσουν οι ακολουθίες. Αν και είναι η μέθοδος που χρησιμοποιείται πιο συχνά, έχει ένα μεγάλο μειονέκτημα. Αν για κάποιο λόγο εισαχθεί ένα κενό στις στοιχίσεις που έγιναν στην αρχή, αυτό το κενό θα διατηρηθεί και στις επόμενες, ακόμα και αν δεν χρειάζεται (Feng & Doolittle, 1987).

Υπάρχουν όπως αναφέρθηκε και πιο πάνω, αρκετά προγράμματα για τη δημιουργία πολλαπλών στοιχίσεων. Μερικά από τα πιο διαδεδομένα είναι το ClustalW (προοδευτική στοίχιση), το T-coffee (προοδευτική στοίχιση), το MUSCLE (προοδευτική στοίχιση).

Το MUSCLE είναι ένα πρόγραμμα όπου γίνεται ο γρήγορος υπολογισμός των αποστάσεων χρησιμοποιώντας είτε τη μέθοδο kmer, είτε τη μέθοδο UPGMA για τη δημιουργία ενός δένδρου οδηγού, η προοδευτική στοίχιση χρησιμοποιώντας το log-expectation score και τέλος βελτιστοποίηση της στοίχισης μέσω μίας κυκλικής λογικής (Edgar, 2004).

Δυστυχώς, ακόμα και στα καλύτερα προγράμματα, αν οι ακολουθίες δεν είναι αρκετά όμοιες, τότε οι πολλαπλές στοιχίσεις που δημιουργούν δεν είναι οι βέλτιστες και χρειάζεται να γίνει μία χειρωνακτική βελτίωση της στοίχισης με προγράμματα όπως το Seaview και το Bioedit. Για να είναι όσο το δυνατόν καλύτερα τα αποτελέσματα από τα προγράμματα πολλαπλής στοίχισης, υπάρχουν μερικοί κανόνες που μπορεί να ακολουθήσει κάποιος, όπως η προαιρετική διαγραφή ακολουθιών που είναι πολύ απομακρυσμένες εξελικτικά και που αλλοιώνουν τη στοίχιση και στη συνέχεια επαναστοίχιση των ακολουθιών με πρόγραμμα πολλαπλής στοίχισης.

1.15 Φυλογενετικά δέντρα

Η αναπαράσταση των εξελικτικών σχέσεων γίνεται μέσα από τα φυλογενετικά δένδρα. Οι μέθοδοι που χρησιμοποιούνται για την κατασκευή των δέντρων, χωρίζονται σε δύο βασικές κατηγορίες, τις βασισμένες σε αποστάσεις (distance based methods) και τις βασισμένες σε χαρακτήρες (character based methods).

Οι μέθοδοι που βασίζονται στις αποστάσεις, υπολογίζουν αρχικά τις αποστάσεις για κάθε πιθανό ζεύγος ακολουθιών και έπειτα δημιουργείται ένας πίνακας αποστάσεων. Αφού υπολογιστεί ο πίνακας απόστασεων και γίνει διόρθωση των παρατηρούμενων αποστάσεων σε πραγματικές (λόγω πολλαπλών μεταλλάξεων στην ίδια θέση), χρησιμοποιείται μια από τις πολλές μεθόδους κατασκευής δένδρων, που χρησιμοποιεί τις τιμές του πίνακα. Μια από τις πιο απλές και πιο διαδεδομένες είναι το **Neighbor Joining.** Μοιάζει λίγο με την μέθοδο UPGMA, ωστόσο δε θεωρεί ότι οι ακολουθίες εξελίσσονται με τον ίδιο ρυθμό. Σε αυτή τη μέθοδο, όλα τα taxa συνδέονται σε έναν εσωτερικό κόμβο και δημιουργείται ένα δέντρο που μοιάζει με αστέρι. Στη συνέχεια, για κάθε ζεύγος taxa που φαίνεται να έχουν μικρότερη απόσταση μεταξύ τους, δημιουργούν έναν κόμβο ανάμεσα τους. Όλη αυτή η διαδικασία επαναλαμβάνεται έως τη στιγμή που θα δημιουργηθεί ένα πλήρες δέντρο (Mailund *et al.*, 2006).

Οι μέθοδοι βασίζονται σε χαρακτήρες είναι η **Μέγιστη Φειδωλότητα** (maximum parsimony) και η **Μέγιστη Πιθανοφάνεια** (maximum likelihood). Η κύρια ιδέα πάνω στην οποία στηρίζεται η μέθοδος της Μέγιστης Φειδωλότητας είναι ότι η απλούστερη εξήγηση είναι πιθανόν και η καλύτερη. Με αυτόν τον τρόπο λοιπόν, προσπαθεί να βρει το δέντρο που χρειάζεται τις λιγότερες αντικαταστάσεις χαρακτήρων για την δημιουργία του.

Η Μέγιστη πιθανοφάνεια ψάχνει το δέντρο, στο οποίο το εξελικτικό μονοπάτι που αναπαρίσταται είναι και το πιο πιθανό, σύμφωνα πάντα με τα δεδομένα των ακολουθιών και κάποιες στατιστικές αναλύσεις. Είναι μια μέθοδος υπολογιστικά ακριβή, ωστόσο, τα τελευταία χρόνια, οι πολύ ισχυροί υπολογιστές γραφείου που

έχουν κατασκευαστεί επιτρέπουν την εφαρμογή αυτής της μεθόδου σε πολλά προβλήματα φυλογένεσης, δίνοντας καλύτερα αποτελέσματα από πολλές άλλες μεθόδους.

Μετά την παραγωγή των δέντρων γίνεται και η αξιολόγηση τους, για να φανεί το πόσο σταθερά είναι και αν εκφράζουν σωστά την εξελικτική πορεία. Γενικά οι μέθοδοι που χρησιμοποιούνται συχνότερα, περιλαμβάνουν την επαναδειγματοληψία των θέσεων της πολλαπλής στοίχισης των ακολουθιών που χρησιμοποιήθηκαν για την παραγωγή των δέντρων. Οι μέθοδοι που χρησιμοποιείται συχνότερα για την αξιολόγηση ενός δέντρου, είναι το bootstrap και το jackknife.

1.16 Λογισμικό PHYLIP (PHYlogeny Inference Package)

Πρόκειται για ένα πακέτο 30 περίπου προγραμμάτων που μπορούν να χρησιμοποιηθούν για φυλογενετική ανάλυση (Felsenstein, 1989). Από το PHYLIP εμείς χρησιμοποιήσαμε δύο προγράμματα. Το PROTDIST και το NEIGHBOR για την κατασκευή φυλογενετικών δένδρων neighbor joining από πολλαπλές στοιχίσεις ομόλογων AARS ακολουθιών.

Συγκεκριμένα, το **PROTDIST** υπολογίζει τις αποστάσεις μεταξύ πρωτεϊνικών ακολουθιών στις οποίες έχει γίνει πολλαπλή στοίχιση δημιουργεί πίνακες αποστάσεων. Μάλιστα, δίνεται στον χρήστη η δυνατότητα να επιλέξει πιο εξελικτικό μοντέλο αντικατάστασης αμινοξέων θέλει να χρησιμοποιήσει. Το **NEIGHBOR** είναι ένα πρόγραμμα δημιουργίας δέντρων, το οποίο χρησιμοποιεί σαν δεδομένα του πίνακες αποστάσεων πουέχουν δημιουργηθεί από κάποιο άλλο πρόγραμμα, όπως π.χ. το PROTDIST.

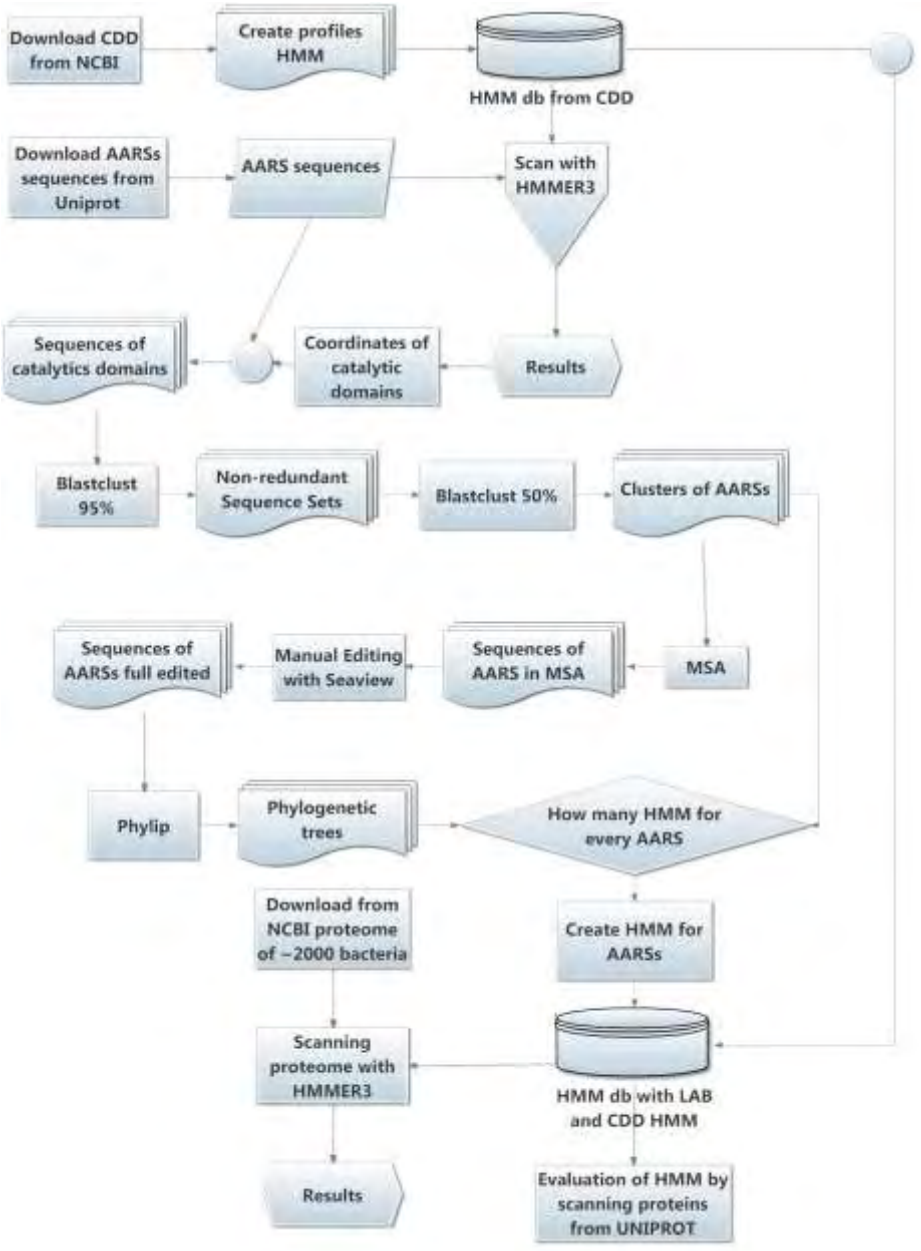
1.17 Λογισμικό Seaview

Πρόκειται για ένα για ένα λογισμικό με γραφικό περιβάλλον για την δημιουργία-επεξεργασία αρχείων πολλαπλής στοίχισης και δημιουργίας φυλογενετικών δέντρων (Gouy *et al*., 2010).

2. Υλικά και Μέθοδοι – Τα βήματα της βιοπληροφορικής ανάλυσης

2.1 Διάγραμμα ροής

Τα βασικά βήματα της ανάλυσης καθώς επίσης και τα δεδομένα που χρησιμοποιήθηκαν συνοψίζονται στην παρακάτω εικόνα 2.1 (διάγραμμα ροής).



Εικόνα 2.1. Διάγραμμα ροής της βιοπληροφορικής ανάλυσης που πραγματοποιήθηκε σε αυτή την εργασία.

Τα βήματα που ακολουθήθηκαν αναπτύσσονται πιο αναλυτικά παρακάτω.

2.2 Συλλογή γνωστών AARSs

Αρχικά έγινε η συλλογή ~13.000 γνωστών αμινοάκυλο tRNA συνθετασών από τη βάση δεδομένων SwissProt της ιστοσελίδας UNIPROT, καθώς επίσης και 71 ακολουθίες από τη βάση δεδομένων TREMBL της ιστοσελίδας UNIPROT και ενώθηκαν όλα μαζί σε ένα αρχείο με διαμόρφωση FASTA (βλ. 1_file_seq_uniprot.fa). Επίσης έγινε μία μετονομασία ορισμένων AARS με EC=6.1.1.23, με βάση την Expasy (http://enzyme.expasy.org/). Ακόμη, πρέπει να αναφερθεί ότι όλα τα αρχεία των scripts και των αποτελεσμάτων βρίσκονται στον ηλεκτρονικό φάκελο του συνοδευτικού CD.

2.3 Εντοπισμός CDD

Στη συνέχεια εντοπίστηκαν και χρησιμοποιήθηκαν τα conserved domains των AARS από την βάση δεδομένων CDD, του NCBI.

2.3.1 Συλλογή των CDD των αμινοάκυλο tRNA συνθετασών από το NCBI

Συλλέξαμε από τον ιστοχώρο του Conserved Domain Databases (CDD) του NCBI τις πολλαπλές στοιχίσεις και τα PSSMs για την κάθε μία AARS που θέλαμε να μελετήσουμε. Πιο συγκεκριμένα κατεβάσαμε τα εξής αρχεία:

Class I		Class II	
CDD	**AARS**	**CDD**	**AARS**
cd00671_ArgRS_core_class_I	Αργινίνη	cd00496_PheRS_alpha_core_class_II	Φαινυλανανίνη α-αλυσίδα
cd00672_CysRS_core_class_I	Κυστενίνη	cd00769_PheRS_beta_core_class_II	Φαινυλανανίνη β-αλυσίδα
cd00674_LysRS_core_class_I	Λυσίνη	cd00770_SerRS_core_class_II	Σερίνη
cd00805_TyrRS_core_class_I	Τυροσίνη	cd00771_ThrRS_core_class_II	Θρεονίνη
cd00806_TrpRS_core_class_I	Τρυπτοφάνη	cd00773_HisRS_like_core_class_II	Ιστιδίνη
cd00807_GlnRS_core_class_I	Γλουταμίνη	cd00774_GlyRS_like_core_class_II	Γλυκίνη
cd00808_GluRS_core_class_I	Γλουταμικό	cd00733_GlyRS_alpha_core_class_II	Γλυκίνη
cd00812_LeuRS_core_class_I	Λευκίνη	cd00776_AsxRS_core_class_II	Ασπαραγίνη
cd00814_MetRS_core_class_I	Μεθειονίνη	cd00777_AspRS_core_class_II	Ασπαραγινικό
cd00817_VaLRS_core_class_I	Βαλίνη	cd00778_ProRS_core_arch_euk_class_II	Προλίνη
cd00818_IleRS_core_class_I	Ισολευκίνη	cd00779_ProRS_core_prok_clas	Προλίνη
-	-	cd00673_AlaRS_core_class_II	Αλανίνη
-	-	cd00645_AsnA_class_II	Ασπαραγίνη
-	-	cd00775_LysRS_core_class_II	Λυσίνη

Πίνακας 2.1 Λίστα με CDD που χρησιμοποιήσαμε από το NCBI (file3.1.txt, folder CDD).

2.4 Δημιουργία profile hidden Markov Models (HMMs)

Για την δημιουργία μοντέλου για την κάθε AARS, έπρεπε να ανοίξουμε το πρόγραμμα HMMER στο directory που βρίσκονταν τα αρχεία με τα CDD και να τρέξουμε την παρακάτω εντολή:

"Hmmbuild <<όνομα αρχείου*>>.hmm ./<<όνομα αρχείου*>>"

Και για τα 20 διαφορετικά αρχεία, έπρεπε να τρέξουμε την παραπάνω εντολή, χωρίς τα <<>>, απλά βάζοντας κάθε φορά το όνομα του κάθε CDD.

Στη συνέχεια δημιουργήθηκαν 20 διαφορετικά αρχεία με την κατάληξη *.hmm.

Τέλος, όλα αυτά τα αρχεία έπρεπε να γίνουν ένα, οπότε τρέξαμε την παρακάτω εντολή:

"cat *.faa > all_cdd.faa"

2.5 Δημιουργία δικών μας profile HMMs

2.5.1 Σάρωση των AARS που ανακτήθηκαν από τη UNIPROT

Για την σάρωση των ααRS χρησιμοποιήσαμε το πρόγραμμα hmmscan του πακέτου HMMER3. Τοποθετήσαμε τα αρχεία . 1_file_seq_uniprot.fa και all_cdd.faa στο κατάλληλο directory και στη συνέχεια πληκτρολογήσαμε την εντολή:

"hmmscan -E 1e-10 -o ./results/raw_cdd.out --tblout ./results/tbl_cdd.out –domtblout ./results/dom_cdd.out ./all_cdd.faa ./ 1_file_seq_uniprot.fa"

Έπειτα, μόλις τελείωσε το πρόγραμμα, πήραμε το αρχείο dom_cdd.out

2.5.2 Ανάλυση του αρχείου dom_cdd.out

Μέσα σε ένα αρχείο dom του hmmscan, βρίσκονται αρκετές πληροφορίες, όπως:

- **target name**: Το όνομα της ακολουθίας στόχος.
- **target accession**: Το Accession της ακολουθίας στόχος, αλλιώς "-" αν δεν υπάρχει.
- **tlen**: το μήκος της ακολουθίας στόχος.
- **query name**: Το όνομα της ακολουθίας επερώτησης.
- **qlen**: το μήκος της ακολουθίας επερώτησης
- **E-value**: Το συνολικό E-value της σύγκρισης για όλα τα domain.
- **score**: Το συνολικό Bit-score της σύγκρισης για όλα τα domain.
- **bias**: Η διόρθωση που έγινε στο score.
- **#**: Ο αριθμός του domain που βρίσκεται στη συγκεκριμένη σειρά
- **of**: Πόσα domain βρέθηκαν στην πρωτεΐνη στόχο.
- **c-Evalue**: Πρόκειται για το "conditional E-value", ένα μέτρο αξιοπιστίας του domain που βρίσκεται σε αυτή τη σειρά.
- **i-Evalue**: Πρόκειται για το "independent E-value", δηλαδή για το E-value που θα είχε το συγκεκριμένο domain, αν ήταν το μοναδικό μέσα στην ακολουθία.
- **score**: το bit score για το συγκεκριμένο domain.
- **bias**: Η διόρθωση που έγινε στο score του συγκεκριμένου domain.
- **from (hmm coord)**: η θέση του αμινοξέος από όπου ξεκινάει η σύγκριση του alignment.

- **to (hmm coord)**: Η θέση που τελειώνει το πιάσιμο.
- **from (ali coord):** Η θέση του αμινοξέος από όπου ξεκινάει το domain που βρέθηκε πάνω στη πρωτεΐνη στόχο.
- **to (ali coord):** Η θέση του αμινοξέος από όπου τελειώνει το domain που βρέθηκε πάνω στην πρωτεΐνη στόχο.
- **from (env coord)**: Η θέση του αμινοξέος που ξεκινάει το match του domain μερικά αμινοξέα περισσότερα
- **to (env coord)**: Το τέλος της θέσης που τελειώνει το match του domain, με μερικά αμινοξέα ακόμα
- **description of target**: Η περιγραφή της πρωτεΐνης στόχου αν υπάρχει

Η ανάλυση του αρχείου έγινε με το perl script parse_dom1, με το οποίο κρατήσαμε τα domain για την κάθε πρωτεΐνη που είχαν το καλύτερο score και δημιουργήσαμε το αρχείο file4.txt

Έπειτα, έγινε ένας χειρονακτικός έλεγχος για να δούμε αν όντως την κάθε πρωτεΐνη την ανίχνευε καλύτερα το κατάλληλο μοντέλο.

2.5.3 Εύρεση συντεταγμένων του κάθε domain

Από το αρχείο file4.txt, κρατήσαμε τις συντεταγμένες του κάθε domain με τη χρήση του perl script keep_coord.pl, το οποίο στη συνέχεια διάβαζε το αρχείο 1_file_seq_uniprot.fa που περιείχε τις ακολουθίες των πρωτεϊνών και κράταγε από τις ακολουθίες μόνο το κομμάτι του domain και τα μετέφερε στη συνέχεια σε καινούργια αρχεία (κάθε ένα αρχείο για κάθε συνθετάση), όπου ήταν σε μορφή FASTA.

2.5.4 Χρήση του Blastclust για μείωση των πρωτεϊνών με μεγάλη ομολογία και δημιουργία clusters

2.5.4.1 Χρήση του Blastclust για απομάκρυνση των πρωτεϊνών που έχουν >95% ταύτιση

Για κάθε μία συνθετάση, πήραμε το αρχείο που περιείχε τις ακολουθίες των domains και το τρέξαμε στο Blastclust με την παρακάτω εντολή:

"blastclust -i domain_AARS_.fa -o cluster_AARS_95.txt -p T –L 0.9 –b T –S 95"

-i: δείνουμε το όνομα του αρχείου με τις ακολουθίες.

-o: το αρχείο που παράγεται από το blastclust και περιέχει τα clusters.

-p: δηλώνουμε αν πρόκειται για πρωτεϊνικές (T) ή νουκλεοτιδικές (F) ακολουθίες.

-L: το ποσοστό του μήκους της ακολουθίας που θέλουμε να υπάρχει ομολογία (0.9 για 90% της ακολουθίας).

-S: το ποσοστό ομοιότητας (95 για 95% ομοιότητα).

2.5.4.2 Ανάλυση των αρχείων cluster_AARS_95.txt

Για την ανάλυση των αρχείων cluster_AARS_95.txt, χρησιμοποιήσαμε ένα perl script, με το οποίο παίρναμε μία μόνο πρωτεΐνη από το κάθε cluster για την κάθε tRNA συνθετάση και στη συνέχεια το perl script δημιουργούσε νέα αρχεία nr_AARS_domain.fa που περιείχαν για την κάθε συνθετάση τα domains των πρωτεϊνών που κρατήσαμε από τα clusters.

2.5.4.3 Blastclust για τη δημιουργία clusters

Έπειτα, ξαναχρησιμοποιήσαμε το πρόγραμμα Blastclust, για να δημιουργήσουμε clusters των domain των πρωτεϊνών για την κάθε μία συνθετάση, με την εντολή:

"blustclust -i domain_AARS_.fa -o cluster_AARS_95.txt -p T –L 0.5 –b F –S 50"

2.5.4.4 Ανάλυση των αρχείων clusters_nr_AARS.txt

Από τα αρχεία που δημιουργήθηκαν με την ονομασία clusters_nr_AARS.txt, με τη χρήση του perl script parse_clusters_nr_AARS.pl κρατήσαμε για κάθε συνθετάση μόνο τα clusters που είχαν παραπάνω από 20 ακολουθίες και δημιουργήσαμε αρχεία με την ονομασία list_cl_nr_AARS.txt που περιείχαν τις λίστες με τις πρωτεϊνες που υπήρχαν στα clusters που κρατήσαμε και έπειτα με τη χρήση του perl script get_seq_for_cl_nr_AARS.pl, δημιουργήσαμε νέα αρχεία με την ονομασία seq_cl_nr_AARS.fa, που περιείχαν τις ακολουθίες των πρωτεϊνών που κρατήσαμε για την κάθε συνθετάση σε μορφή FASTA.

2.5.5 Χρήση του MUSCLE για πολλαπλή στοίχιση των ακολουθιών

Έγινε χρήση του προγράμματος MUSCLE για την πολλαπλή στοίχιση των ακολουθιών που βρίσκονταν στα αρχεία seq_cl_nr_AARS.fa. Η εντολή που χρησιμοποιήθηκε ήταν η εξής:

"muscle –in seq_cl_nr_AARS.fa –phyiout msa_AARS.phyi"

-in: αρχείο με τις ακολουθίες που προορίζονται για πολλαπλή στοίχιση

-phyiout: αρχείο με ακολουθίες σε πολλαπλή στοίχιση σε διαμόρφωση phyi για χρήση από το πρόγραμμα phylip

2.5.6 Χειρωνακτική διόρθωση (manual editing) των αρχείων phyiout_msa_AARS.phyi

Η βελτίωση των αρχείων πολλαπλής στοίχισης έγινε χειρωνακτικά, με τη χρήση του προγράμματος Seaview.

2.5.7 Δημιουργία φυλογενετικών δέντρων με το πρόγραμμα PHYLIP

2.5.7.1 Χρήση του PHYLIP για την δημιουργία των φυλογενετικών δένδρων

Αρχικά έπρεπε να γίνει η δημιουργία των πινάκων απόστασης (χρησιμοποιήθηκε ο πίνακας JTT) και για αυτό χρησιμοποιήσαμε την παρακάτω εντολή στο PHYLIP για τα αρχεία msa_AARS.phyi.

"phylip protdist"

Εδώ πρέπει να σημειωθεί ότι το protdist διαβάζει τα αρχεία με την ονομασία infile, και για αυτό το λόγο, για τη δημιουργία των πινάκων απόστασης της κάθε συνθετάσης, κάθε αρχείο μετονομάστηκε σε infile και μετά ξανά στην ονομασία msa_AARS.phyi. Έπειτα οι πίνακες απόστασης που δημιουργήθηκαν από το protdist είχαν την ονομασία outfile.

Στη συνέχεια γινόταν μετονομασία του outfile σε infile και με την εντολή

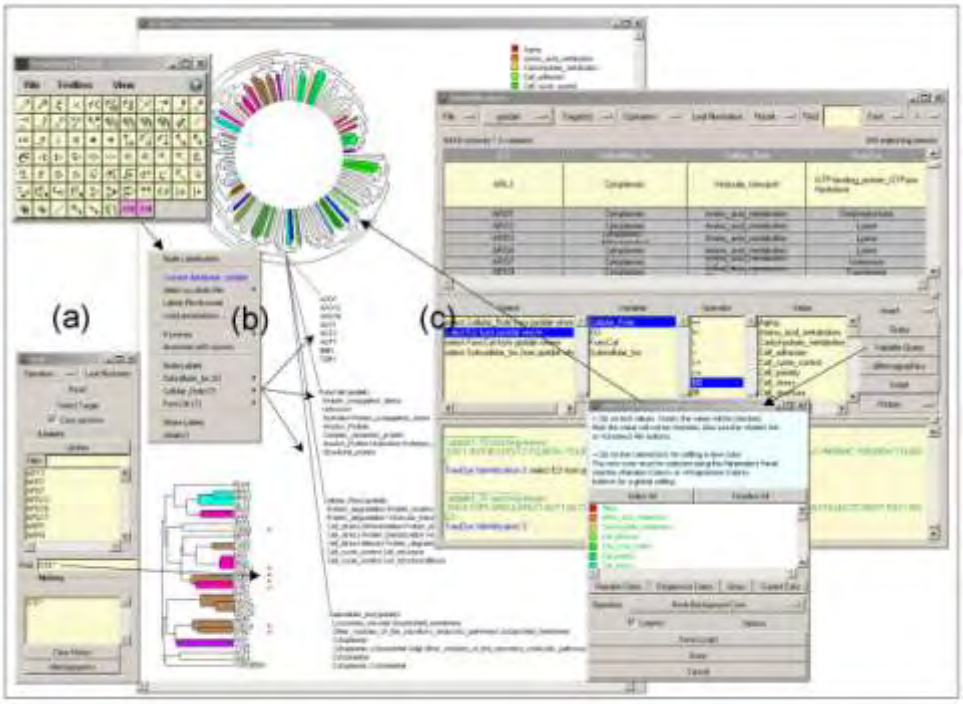
"phylip neighbor",

Δημιουργήθηκαν τα δέντρα με τη μέθοδο Neighbor-Joining και είχαν την ονομασία outfile.

Τέλος γινόταν η μετονομασία τους σε tree_AARS.

2.6 Επεξεργασία δένδρων

Για την επεξεργασία των δέντρων χρησιμοποιήθηκε το πρόγραμμα Treedyn, το οποίο χρησιμοποιείται για την απεικόνιση, την γραφική επεξεργασία, τον σχολιασμό και την ανάλυσητων δέντρων που έχουν παραχθεί από κάποιο πρόγραμμα π.χ. το PHYLIP.



Εικόνα 2.2: Στην παραπάνω εικόνα φαίνεται το γραφικό περιβάλλον του Treedyn (Chevenet *et al.*, 2006)

2.6.1 Δημιουργία αρχείου για σχολιασμό πρωτεΐνών (annotation file)

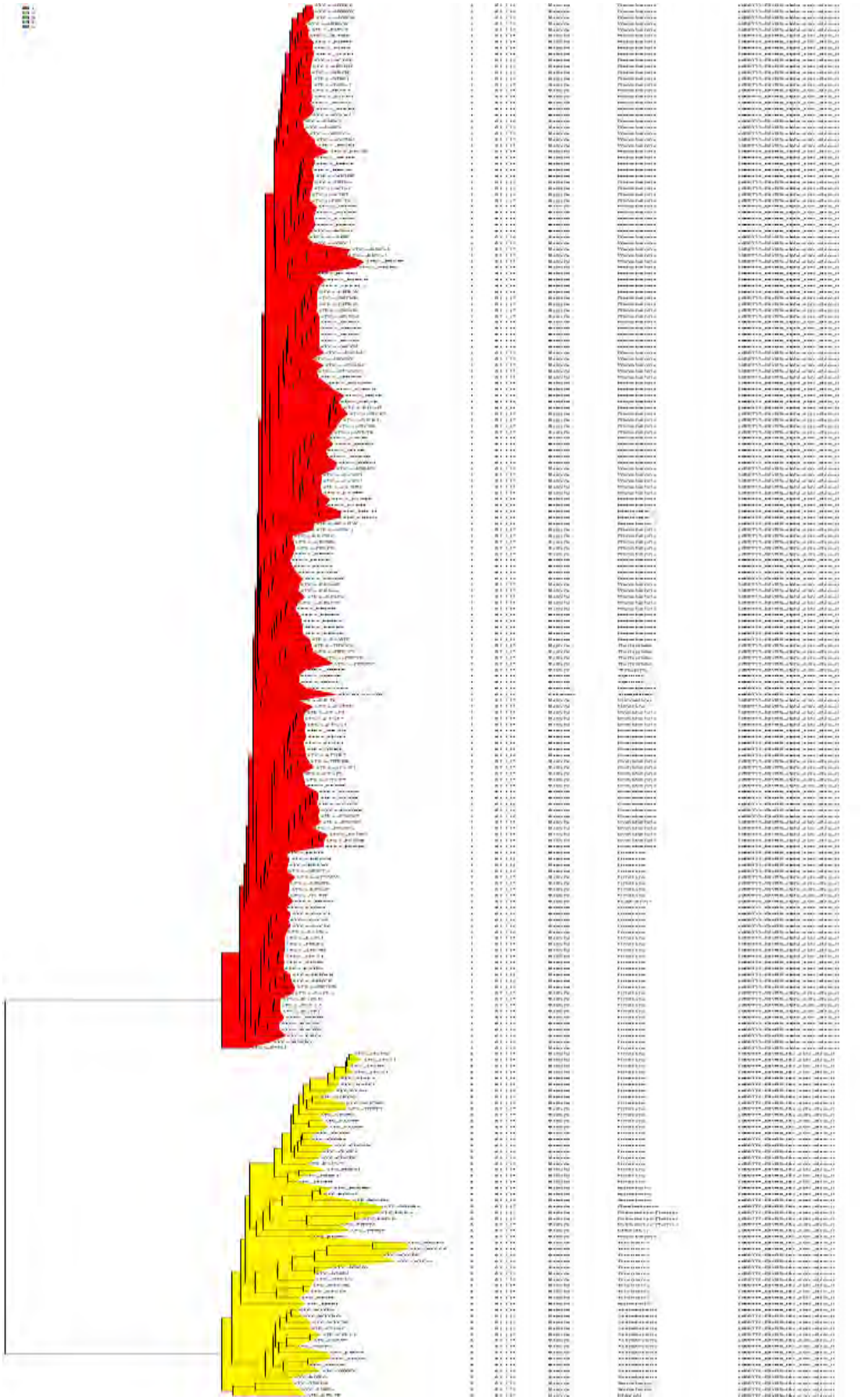
Για την επεξεργασία των δένδρων με το treedyn, αρχικά δημιουργήσαμε ένα annotion file με την ονομασία annot_file.tlf. Η δημιουργία του αρχείου έγινε με τη χρήση του perl script create_annot_file.pl, το οποίο χρησιμοποίησε το αρχείο table.txt που περιείχε της πληροφορίες για την κάθε πρωτεΐνη. Οι πληροφορίες που υπήρχαν μέσα στο annot_file.tlf ήταν οι εξής:

Πληροφορίες του annot_file.tlf	
PROT_ID	Το ID της κάθε πρωτεΐνης
EC	Ο αριθμός EC της πρωτεΐνης
Core_model	Το μοντέλο HMM που χτύπαγε τη πρωτεϊνη
E_Value	Το E-Value της
Organism	Το όνομα του οργανισμού που ανήκει η πρωτεΐνη
Group	Το group στο οποίο ανήκει ο οργανισμός
Tax_2	Το 2^{ο} group που ανήκει ο οργανισμός
Lineage	Το lineage του οργανισμού
Cluster	Το cluster στο οποίο ανήκει η πρωτεΐνη

Πίνακας 2.2: Στον παραπάνω πίνακα συνοψίζονται οι πληροφορίες που βρίσκονται μέσα στο αρχείο annot_file.tlf, που χρησιμοποιήθηκε για τον σχολιασμό των δέντρων

2.6.2 Χρήση του προγράμματος treedyn για επεξεργασία

Στη συνέχεια χρησιμοποιήθηκε το treedyn για τη γραφική παράσταση και επεξεργασία των δένδρων που πήραμε από το PHYLIP. Έγινε σχολιασμός των δέντρων με τη βοήθεια του annot_file.tlf και μετέπειτα ο χρωματισμός των groups των πρωτεϊνών, ανάλογα με το cluster [2.5.4.3] στο οποίο βρισκόντουσαν.



Εικόνα 2.3: Φυλογενετικό δέντρο της GlyRS, στο οποίο εμφανίζονται 2 μονοφυλετικές ομάδες

2.6.3 Έλεγχος των δέντρων

Έγινε έλεγχος των δέντρων, για την ύπαρξη μονοφυλετικών ομάδων ανάλογα με το cluster [2.5.4.3] στο οποίο ανήκαν οι πρωτεΐνες της κάθε ομάδας.

2.7 Δημιουργία profile HMM

2.7.1 Συλλογή των μονοφυλετικών group για τη δημιουργία συγγενικών ομάδων HMMs που ανιχνεύουν την ίδια AARS.

Με βάση την μονοφυλετικότητα των clusters για την κάθε συνθετάση, φτιάξαμε από ένα profile HMM για το κάθε cluster. Άρα, μία συνθετάση μπορεί να αποτελείται από 1 ή περισσότερα profile HMMs που δημιουργήθηκαν στο εργαστήριο. Για να δημιουργήσουμε τα HMM έγινε συλλογή των πρωτεϊνών, που ανήκαν στα μονοφυλετικά group, σε αρχεία με την ονομασία list_prot_AARS_hmm.txt και έπειτα, με τη χρήση του perl script get_seq_for_hmm.pl δημιουργήσαμε αρχεία με την ονομασία seq_AARS_for_hmm_(x).fa, όπου περιείχαν τις ακολουθίες των πρωτεϊνών σε μορφή FASTA.

2.7.2 Χρήση του προγράμματος MUSCLE για πολλαπλή στοίχιση των ακολουθιών

Για κάθε αρχείο που δημιουργήθηκε με την ονομασία seq_AARS_for_hmm_(x).fa, έγινε πολλαπλή στοίχιση των ακολουθιών με την χρήση του προγράμματος MUSCLE, με τον τρόπο που περιγράφεται στο [2.5.5] με τη διαφορά ότι αντί για –phyiout χρησιμοποιήσαμε το -fastaout και δημιουργήθηκαν τα αρχεία msa_AARS_for_hmm_(x).fa, τα οποία είναι σε μορφή FASTA.

2.7.3 Χειροκίνητη διόρθωση των αρχείων msa_AARS_for_hmm_(x).fa με το Seaview

Έγινε βελτιστοποίηση της πολλαπλής στοίχισης των αρχείων msa_AARS_for_hmm_(x).fa με το Seaview και έπειτα τα αρχεία αποθηκεύτηκαν με την ονομασία edited_msa_AARS_for_hmm_(x).fa.

2.7.4 Χρήση του προγράμματος HMMER3 για τη δημιουργία των profile HMM

Με τη χρήση του προγράμματος HMMER3 δημιουργήσαμε τα profile HMM όπως περιγράφεται στο [2.4] και τα αρχεία που δημιουργήθηκαν ονομάστηκαν AARS_(x).hmm.

2.7.5 Δημιουργία βάσης δεδομένων profile HMM

Όλα τα αρχεία που δημιουργήθηκαν από το [2.7.4] μετατράπηκαν σε μία βάση δεδομένων profile HMMs με την εντολή

"cat *.hmm > all_lab_models.hmm"

Έπειτα πήραμε το αρχείο all_lab_models.hmm που περιείχε τα μοντέλα που δημιουργήθηκαν στο εργαστήριο, καθώς επίσης και το αρχείο all_cdd.hmm[2.4] και τα ενώσαμε σε ένα αρχείο με την εντολή:

"cat *.hmm > all_models.hmm"

Με αυτόν τον τρόπο δημιουργήθηκε μία βάση δεδομένων profile HMM που περιείχε τα μοντέλα HMM που δημιουργήθηκαν στο εργαστήριο μαζί με τα μοντέλα HMM που δημιουργήθηκαν από τα αρχεία CDD.

2.8 Αξιολόγηση των μοντέλων που δημιουργήθηκαν στο εργαστήριο

2.8.1 Συλλογή γνωστών πρωτεϊνών του UNIPROT που δεν χρησιμοποιήθηκαν για την δημιουργία των μοντέλων

Αρχικά με τη βοήθεια του perl script get_seq_2.pl, πήραμε τις πρωτεΐνες που δεν χρησιμοποιήθηκαν για την κατασκευή των μοντέλων HMM που δημιουργήθηκαν στο εργαστήριο και τις τοποθετήσαμε στο αρχείο prot_nu.fa με τη μορφή FASTA. Οι πρωτεΐνες που δεν χρησιμοποιήθηκαν, ήταν αυτές που στο [2.5.4.1] είχαν ταύτιση >95% με κάποια πρωτεΐνη που χρησιμοποιήθηκε για την δημιουργία των μοντέλων HMM.

2.8.2 Σάρωση (με το hmmscan) των πρωτεϊνών που δεν χρησιμοποιήθηκαν για την δημιουργία των μοντέλων HMM

Με τη χρήση του προγράμματος hmmscan του HMMER3 έγινε σάρωση των πρωτεϊνών του αρχείου prot_nu.fa με τη βάση δεδομένων all_models.hmm και δημιουργήθηκε το αρχείο dom_eval_hmm.hmm.

2.8.3 Επεξεργασία του αρχείου dom_eval_hmm.hmm

Με τη χρήση του perl script parse_dom_eval_hmm.pl , πήραμε τις πρωτεΐνες οι οποίες είχαν εντοπισθεί από κάποιο μοντέλο με το μεγαλύτερο score και τοποθετήθηκαν στο αρχείο eval_dom.txt.

2.8.4 Ανάλυση αρχείου eval_dom.txt

Το αρχείο eval_dom.txt τοποθετήθηκε στο excel και έγινε έλεγχος χειροκίνητα, για να δούμε αν το κάθε profile έβρισκε τις πρωτεΐνες που έπρεπε και ποιο ποσοστό των γνωστών πρωτεϊνών βρέθηκαν.

2.9 Δημιουργία πίνακα διπλασιασμών γονιδίων AARS σε προκαρυωτικά γονιδιώματα

2.9.1 Συλλογή βακτηριακών γονιδιωμάτων από το NCBI

Από τη βάση δεδομένων του NCBI πήραμε το αρχείο all.faa.tar.gz (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) το οποίο περιείχε όλες τις πρωτεΐνες από περίπου 2000 βακτήρια, καθώς επίσης και το αρχείο l_proks0.txt, το οποίο περιείχε πληροφορίες για τα βακτήρια.

2.9.2 Σάρωση των πρωτεϊνών του αρχείου all.faa.tar.gz

Έγινε σάρωση των πρωτεϊνών με το hmmscan, χρησιμοποιώντας τη βάση δεδομένων all_models.hmm [2.7.5] και δημιουργήθηκε το αρχείο dom_ncbi.txt

2.9.3 Ανάλυση του αρχείου dom_ncbi.txt

Έγινε ανάλυση του αρχείου dom_ncbi.txt με το perl script parse_dom_ncbi.txt, με το οποίο πήραμε τις πρωτεΐνες που βρέθηκαν με το μεγαλύτερο score από τα μοντέλα και τοποθετήθηκαν στο αρχείο parse_dom_ncbi.txt.

2.9.4 Ανάλυση δεδομένων στο excel.

Τα δεδομένα του αρχείου parse_dom_ncbi.txt αναλύθηκαν στο excel.

2.9.5 Δημιουργία πίνακα διπλασιασμών από το αρχείο

Η δημιουργία του πίνακα διπλασιασμών έγινε με τη χρήση του perl script create_mega_table.pl, το οποίο δημιούργησε τον πίνακα mega_table.xls. Αυτό που έκανε το script ήταν ότι μέτραγε για κάθε οργανισμό πόσες φορές βρισκόταν στο πρωτέωμά του η κάθε AARS.

3. Αποτελέσματα

3.1 Ανεύρεση των συντεταγμένων των domain των AARSs

Μετά την σάρωση των σχολιασμένων ακολουθιών των πρωτεϊνών που λήφθηκαν από την ιστοσελίδα του UNIPROT, έγινε η επεξεργασία των αρχείων και βρέθηκαν οι συντεταγμένες των domain των AARS και δημιουργήθηκαν αρχεία.

3.2 Blastclust

Για την δημιουργία των profile HMM για την κάθε AARS, έπρεπε να μειώσουμε τον αριθμό των αρκετά όμοιων ακολουθιών, ώστε το μοντέλο της κάθε AARS να μπορεί να "πιάνει" όσο το δυνατόν πιο απομακρυσμένες ακολουθίες. Για αυτό τον λόγο επιλέχθηκε μία ακολουθία από κάθε cluster με ομολογία 95%, για την κάθε AARS. Στην εικόνα 3.1, μπορεί να δει κανείς τα διαφορετικά cluster που δημιουργήθηκαν για τις ακολουθίες των domain της ArgRS.

Εικόνα 3.1: Αποτελέσματα από το blastclust [2.5.4.1] για τις ακολουθίες των domain της ArgRS. Κάθε σειρά αποτελεί και ένα διαφορετικό cluster, στο οποίο οι ακολουθίες έχουν ομοιότητα 95%

Έπειτα, μετά την συλλογή των ακολουθιών, πραγματοποιήθηκε και το δεύτερο blastclust, ώστε να γίνει η σωστή επιλογή του αριθμού των profile HMM που θα έπρεπε να δημιουργηθούν για την κάθε AARS. Στην εικόνα 3.2 μπορεί να δει κανείς τα διαφορετικά cluster των ακολουθιών που συλλέχθηκαν σύμφωνα με την διαδικασία στο [2.5.4.1], οι οποίες είχαν ομοιότητα 50%. Έπειτα, για την δημιουργία των διαφορετικών profile HMM, επιλέχθηκαν τα clusters που είχαν περισσότερες από 20 ακολουθίες, ώστε να υπάρχει ένας ικανοποιητικός αριθμός ακολουθιών για τη δημιουργία του κάθε μοντέλου. Π.χ. για την ArgRS, επιλέχθηκαν τα 3 πρώτα clusters, όπου οι ακολουθίες του κάθε cluster χρησιμοποιήθηκαν για την δημιουργία διαφορετικών μοντέλων για την ArgRS.

Εικόνα 3.2: Αποτελέσματα από το blastclust [2.5.4.3] για τις ακολουθίες των domain της ArgRS. Κάθε σειρά αποτελεί και ένα διαφορετικό cluster, στο οποίο οι ακολουθίες έχουν ομοιότητα 50%.

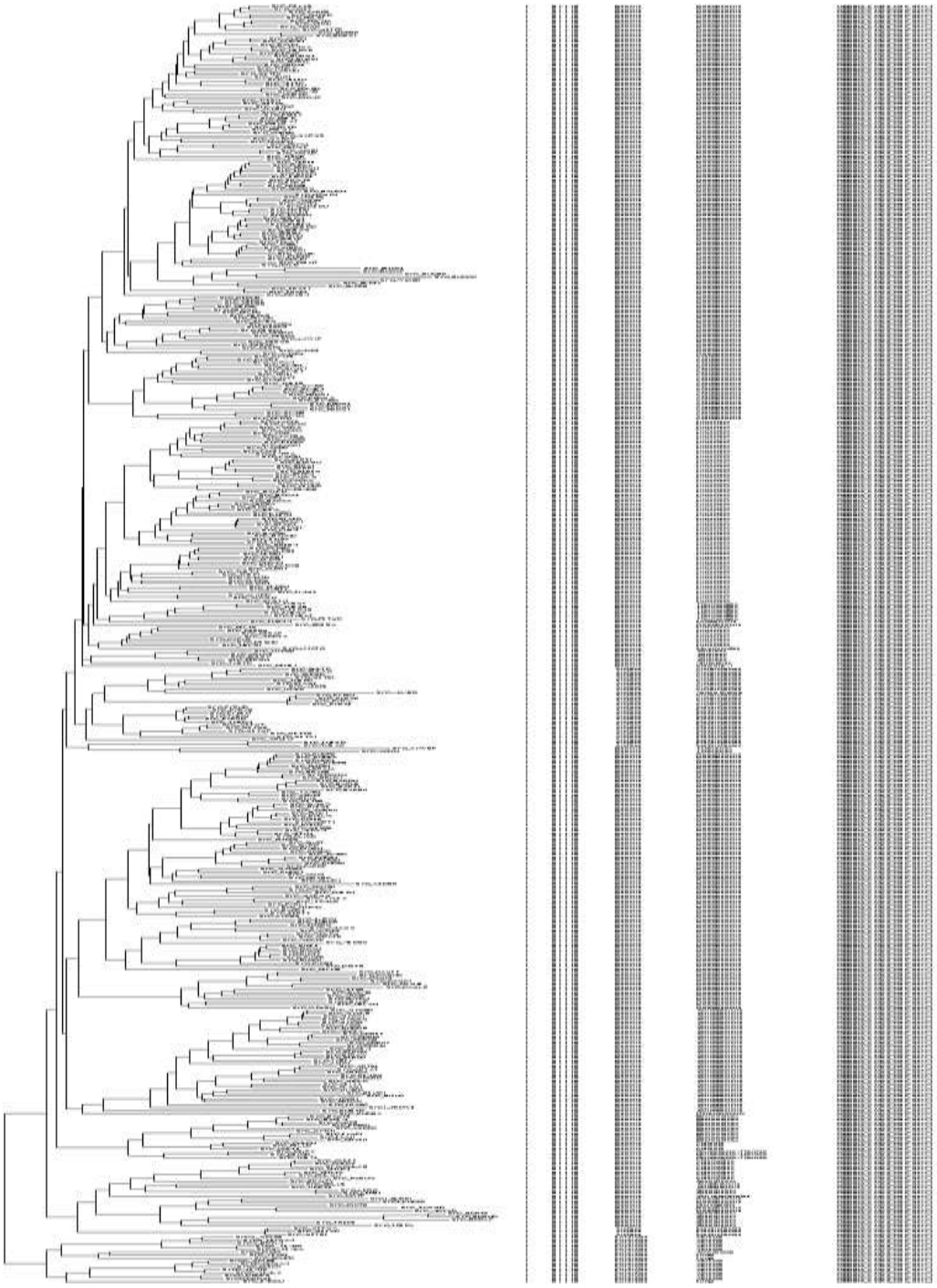
3.3 Πολλαπλή στοίχιση των ακολουθιών του κάθε μοντέλου και χειρωνακτική βελτιστοποίηση της στοίχισης

Στις ακολουθίες του κάθε μοντέλου έγινε πολλαπλή στοίχιση. Στην εικόνα 3.3, φαίνεται το αποτέλεσμα της πολλαπλής στοίχισης του προγράμματος MUSCLE για την AARS. Εύκολα μπορεί να παρατηρήσει κανείς, ότι η αυτόματη πολλαπλή στοίχιση που γίνεται δεν είναι βέλτιστη. Για αυτό το λόγο, έγινε χειρωνακτική επεξεργασία (manual editing) των πολλαπλών στοιχίσεων του κάθε μοντέλου με το πρόγραμμα Seaview.

Εικόνα 3.3: Το αρχείο της πολλαπλής στοίχισης για τις ακολουθίες της SerRS όπως δημιουργήθηκε από το πρόγραμμα MUSCLE. Η απεικόνιση του αρχείου έγινε με το πρόγραμμα Seaview.

3.4 Ανάλυση των δέντρων

Μετά την βελτιστοποίηση που έγινε στις πολλαπλές στοιχίσεις, τα αρχεία χρησιμοποιήθηκαν για την δημιουργία των φυλογενετικών δέντρων όπως περιγράφεται στο [2.5.7.1]. Στη συνέχεια, τα δέντρα που δημιουργήθηκαν, επεξεργαστήκαν με το Treedyn, όπως περιγράφεται στο [2.6.2]. Η τελική μορφή των δέντρων φαίνεται στις παρακάτω εικόνες. Δυστυχώς, λόγω του μεγάλου μεγέθους των φυλογενετικών δέντρων, δεν ήταν δυνατή μία ευδιάκριτη απεικόνιση τους στο παρόν έγγραφο.



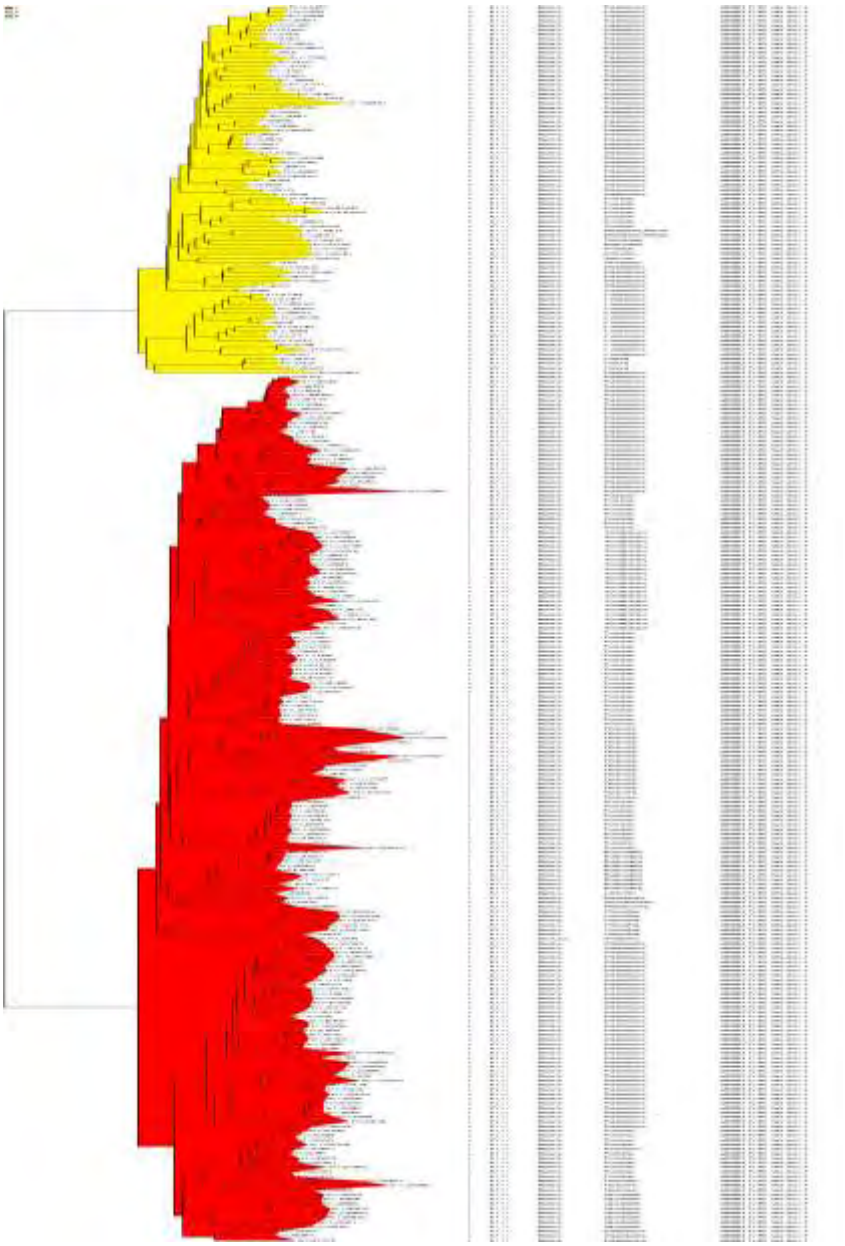
Εικόνα 3.4: Φυλογενετικό δέντρο της CysRS.

Εύκολα μπορεί να παρατηρήσει κανείς στην εικόνα 3.4 ότι δεν σχηματίζεται κάποιο μεγάλο μονοφυλετικό group ακολουθιών, παρά μόνο μία μικρή ομάδα στη κάτω πλευρά του δέντρου. Ωστόσο, λόγω του μικρού αριθμού των ακολουθιών που σχηματίζουν το outgroup, δημιουργήσαμε μόνο ένα profile HMM για την CysRS. Στην εικόνα 3.5, απεικονίζεται το φυλογενετικό δέντρο της TyrRS. Εύκολα και εδώ μπορεί να παρατηρήσει κανείς 2 μονοφυλετικές ομάδες ακολουθιών. Λόγων των δύο group που σχηματίστηκαν, δημιουργήσαμε δύο profile HMMs για την TyrRS. Το πρώτο profile είχε τις ακολουθίες του κόκκινου group, ενώ το δεύτερο profile είχε τις ακολουθίες τους κίτρινου group. Με τον ίδιο ακριβώς τρόπο, δουλέψαμε με τα υπόλοιπα δέντρα. Ανάλογα με τα group πρωτεϊνών που υπήρχαν, δημιουργήσαμε και τον ανάλογο αριθμό των profiles HMM για την κάθε AARS. Στον πίνακα 3.1 αναγράφεται ο συνολικός αριθμός profile HMMs που δημιουργήθηκαν για την κάθε συνθετάση.

AARSs	Models	#
AlaRS	AlaRS_core_class_II_a	2
	AlaRS_core_class_II_b	
ArgRS	ArgRS_core_class_II_a	3
	ArgRS_core_class_II_b	
	ArgRS_core_class_II_c	
AsnRS	AsnRS_core_class_II_a	2
	AsnRS_core_class_II_b	
AspRS	AspRS_core_class_II_archaea	3
	AspRS_core_class_II_bact	
	Asx_core_class_II_61123	
GlnRS	GlnRS_core_class_I	1
GluRS	GluRS_core_class_I_17_24	4
	GluRS_core_class_I_17	
	GluRS_GLUQ	
	GluRS_non_discr_archaea	
GlyRS	GlyRS_core_class_II_a	2
	GlyRS_core_class_II_b	
IleRS	IleRS_core_class_I_a	2
	IleRS_core_class_I_b	
LeuRS	LeuRS_core_class_I_a	2
	LeuRS_core_class_I_b	
MetRS	MetRS_core_class_I_a	3
	MetRS_core_class_I_b	
	MetRS_core_class_I_c	
PheRS	PheRS_alpha_core_class_II_a	2
	PheRS_alpha_core_class_II_b	
ProRS	ProRS_core_class_II_a	3
	ProRS_core_class_II_b	
	ProRS_core_class_II_c	
PSerRS	PSerRS_6.1.1.27	1
PyrLysRS	PyrLysRS_6.1.1.26	1
TyrRS	TyrRS_core_class_I_a	2
	TyrRS_core_class_I_b	
SerRS	SerRS_cII	1

LysRS	LysRS_cII	1
CysRS	CysRS_cI	1
HisRS	HistRS_cII	1
ThrRS	ThrRS_cII	1
TrpRS	TrpRS_cI	1
VlaRS	ValRS_cI	1
LysRS	LysRS_cI	1

Πίνακας 3.1 Αριθμός μοντέλων που δημιουργήθηκαν για την κάθε AARS



Εικόνα 3.5: Φυλογενετικό δέντρο της TyrRS, όπου φαίνεται χαρακτηριστικά η δημιουργία δύο μονοφυλετικών ομάδων. Η πρώτη ομάδα έχει κόκκινο χρώμα και η δεύτερη κίτρινο.

3.5 Αξιολόγηση των profiles HMM που δημιουργήθηκαν.

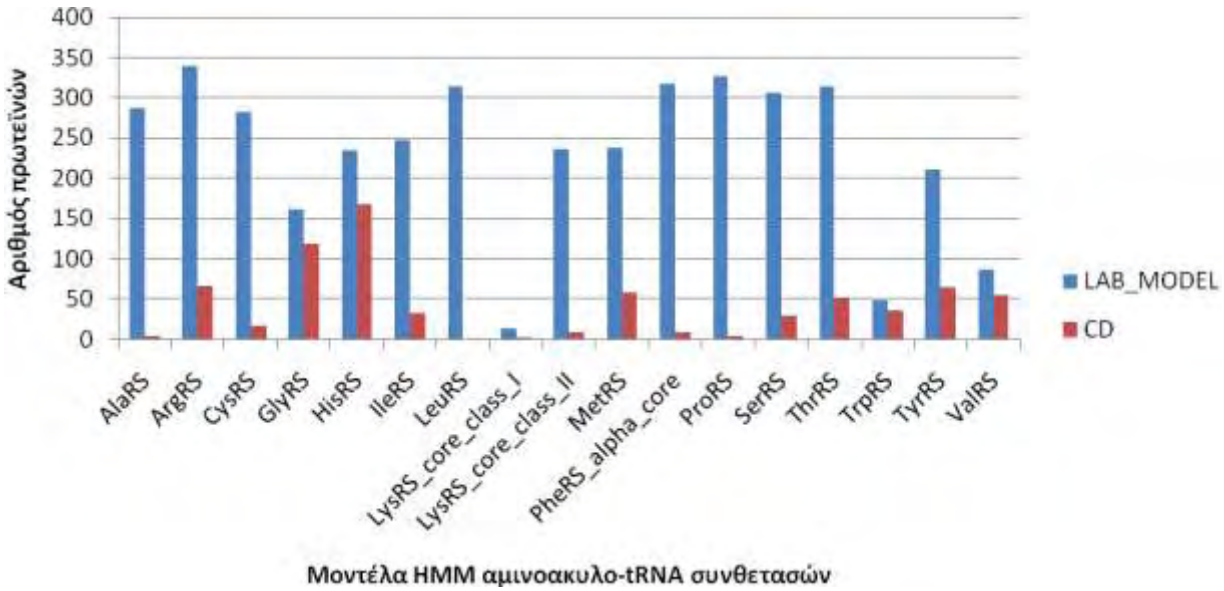
Για την αξιολόγηση των μοντέλων που δημιουργήθηκαν, χρησιμοποιήθηκαν οι ακολουθίες των domains από το blastclust [2.5.4.1] που δεν χρησιμοποιήθηκαν στην περαιτέρω διαδικασία της δημιουργίας των μοντέλων. Στον πίνακα που ακολουθεί, αναγράφεται ο αριθμός των ακολουθιών των domains της κάθε AARS, που χρησιμοποιήθηκε στην διαδικασία της αξιολόγησης. Η αξιολόγιση έγινε επίσης και για τα profile HMMs που δημιουργήθηκαν για τις οικογένειες των AspRS-AsnRS & GluRS – GlnRS, αν και τα αποτελέσματα αυτά δεν αναλύονται στην παρούσα διπλωματική εργασία. Από την αξιολόγιση όλων των μοντέλων προέκυψε ότι ανακτήσαν σωστά το 100% των γνωστών AARS. Στους πίνακες (3.2, 3.3) και γραφήματα (3.6, 3.7) που ακολουθούν εμφανίζονται τα δεδομένα και αποτελέσματα της αξιολόγισης των μοντέλων. Είναι προφανές ότι τα νέα μοντέλα που δημιουργήσαμε είναι πιο αποτελεσματικά.

AARS	Σύνολο Ακολουθιών
AlaRS	291
ArgRS	406
CysRS	299
GlyRS	279
HisRS	403
IleRS	280
LeuRS	314
LysRS_core_class_I	15
LysRS_core_class_II	245
MetRS	296
PheRS_alpha_core	325
ProRS	331
SerRS	335
ThrRS	366
TrpRS	84
TyrRS	274
ValRS	142
Σύνολο	4685

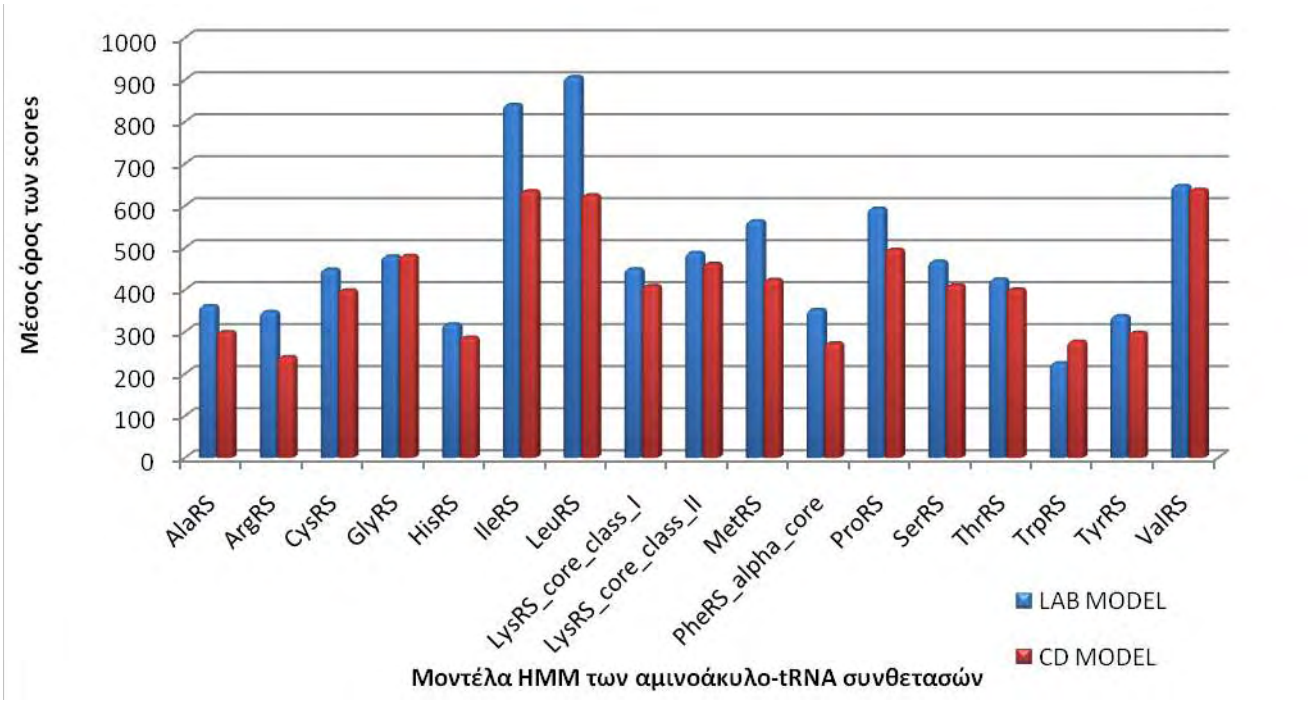
Πίνακας 3.2 Αριθμός ακολουθιών που χρησιμοποιήθηκαν για την αξιολόγηση της κάθε AARS

Μοντέλα	LAB_MODEL	CD	Σύνολο
AlaRS	287	4	291
ArgRS	340	66	406
CysRS	283	16	299
GlyRS	161	118	279
HisRS	235	168	403
IleRS	247	33	280
LeuRS	314	0	314
LysRS_core_class_I	13	2	15
LysRS_core_class_II	237	8	245
MetRS	238	58	296
PheRS_alpha_core	317	8	325
ProRS	327	4	331
SerRS	306	29	335
ThrRS	315	51	366
TrpRS	48	36	84
TyrRS	210	64	274
ValRS	87	55	142
Σύνολο	3965	720	4685

Πίνακας 3.3 Αριθμός ακολουθιών που βρέθηκαν με μεγαλύτερο score από τα μοντέλα που δημιουργήθηκαν στο εργαστήριο (LAB_MODEL) σε σχέση με τον αριθμό των ακολουθιών που βρέθηκαν με μεγαλύτερο score από μοντέλα που δημιουργήθηκαν από τα CDD (CD).



Εικόνα 3.6: Διαγραμματική απεικόνιση του αριθμού των ακολουθιών που εντοπίστηκαν από το κάθε μοντέλο. Φαίνεται χαρακτηριστικά ότι τα μοντέλα που αναπτύχθηκαν στο εργαστήριο εντόπιζαν τις ακολουθίες με μεγαλύτερο score σε σχέση με τα μοντέλα του CDD.



Εικόνα 3.7: Διαγραμματική απεικόνιση του μέσου όρου των scores για τις πρωτεΐνες (από τη UNIPROT) που βρέθηκαν από τα μοντέλα του εργαστηρίου, σε σχέση με το score που είχαν οι αντίστοιχες ακολουθίες από τα μοντέλα των CDD.

3.6 Αξιολόγηση των μοντέλων με βάση τη σάρωση ~ 2000 προκαρυωτικών πρωτεωμάτων

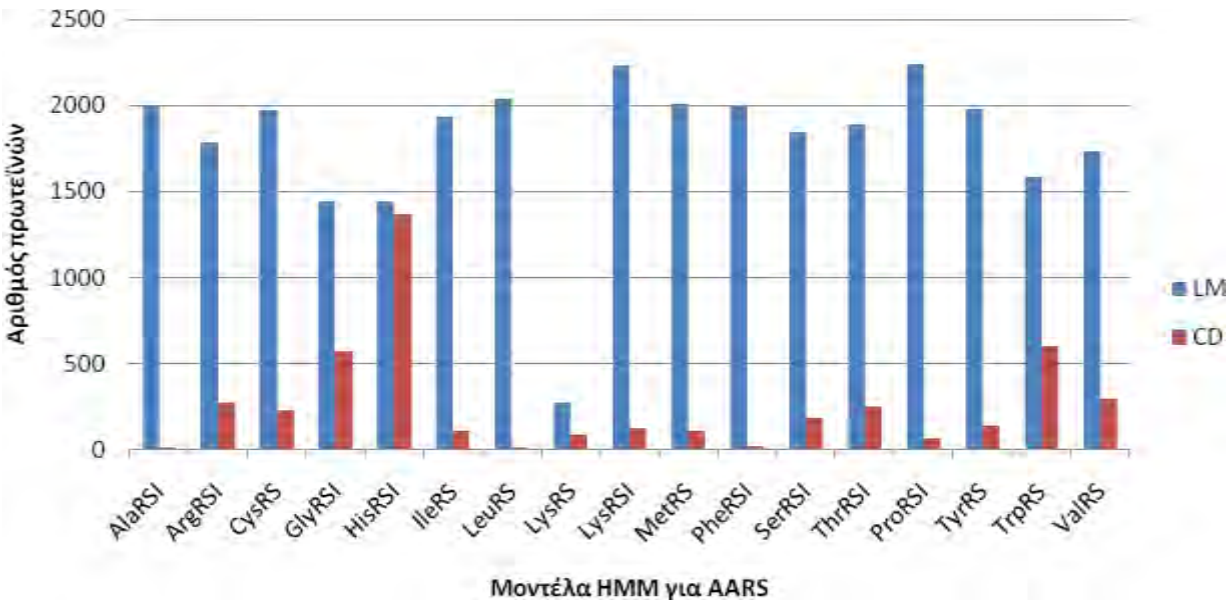
Συνολικά αναλύθηκε το πρωτέωμα από 2006 προκαρυώτες, ωστόσο αφαιρέθηκαν τα αποτελέσματα από 17 οργανισμούς, καθώς δεν είχε ολοκληρωθεί η αλληλούχιση του γονιδιώματός τους.

Μοντέλα	LM	CD	Σύνολο
AlaRS	2002	10	2012
ArgRS	1789	273	2062
CysRS	1975	228	2203
GlyRS	1446	569	2015
HisRS	1442	1369	2811
IleRS	1936	107	2043
LeuRS	2042	9	2051
LysRS	275	89	364
LysRS	2235	126	2361
MetRS	2007	106	2113
PheRS	1996	22	2018
SerRS	1845	185	2030
ThrRS	1889	253	2142

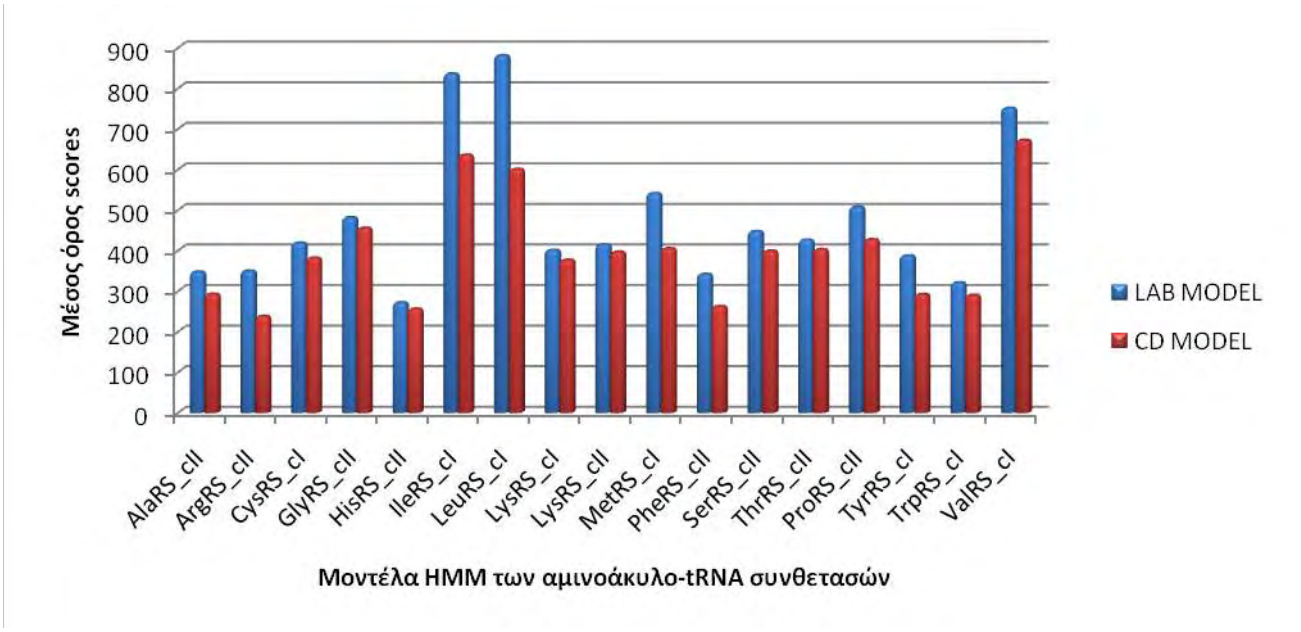
ProRS	2237	67	2304
TyrRS	1980	139	2119
TrpRS	1584	603	2187
ValRS	1736	294	2030
PSerRS	44	0	44
PyrLysRS	23	0	23
Σύνολο	30483	4449	34932

Πίνακας 3.4 AARS που εντοπίστηκαν στο πρωτέωμα 1989 οργανισμών. Συνολικά εντοπίστηκαν 34932 πρωτεΐνες, από τις οποίες οι 30483 εντοπίστηκαν με μεγαλύτερο score από τα μοντέλα που δημιουργήθηκαν στο εργαστήριο (LM – LAB MODELS) και οι 4449 εντοπίστηκαν με μεγαλύτερο score από τα μοντέλα που δημιουργήθηκαν με βάση τα CDD (CD)

Στην παρακάτω εικόνα (Εικόνα 3.8), μπορεί κανείς να παρατηρήσει ότι γενικά τα μοντέλα που δημιουργήθηκαν στο εργαστήριο, λειτουργούν αρκετά καλύτερα από τα μοντέλα που δημιουργήθηκαν από τα CDD, εκτός μόνο από τη περίπτωση της HisRS.



Εικόνα 3.8: Διαγραμματική απεικόνιση του αριθμού των πρωτεϊνών που εντοπίστηκαν με μεγαλύτερο score από τα μοντέλα που δημιουργήθηκαν στο εργαστήριο, σε σχέση με τον αριθμό των πρωτεϊνών που εντοπίστηκαν με μεγαλύτερο score από τα μοντέλα που δημιουργήθηκαν από τα CDD.



Εικόνα 3.9: Διαγραμματική απεικόνιση του μέσου όρου των scores για τις πρωτεΐνες (που κατεβάσαμε από το NCBI) και που βρέθηκαν από τα μοντέλα του εργαστηρίου (LAB MODEL) σε σχέση με το score που είχαν οι αντίστοιχες πρωτεΐνες από τα μοντέλα των CDD (CD).

Σύμφωνα με τους παραπάνω πίνακες και εικόνες, φαίνεται ότι τα μοντέλα που δημιουργήθηκαν στο εργαστήριο (LM) λειτουργούν καλύτερα σε σχέση με τα μοντέλα που δημιουργήθηκαν από τα CDD (CD).

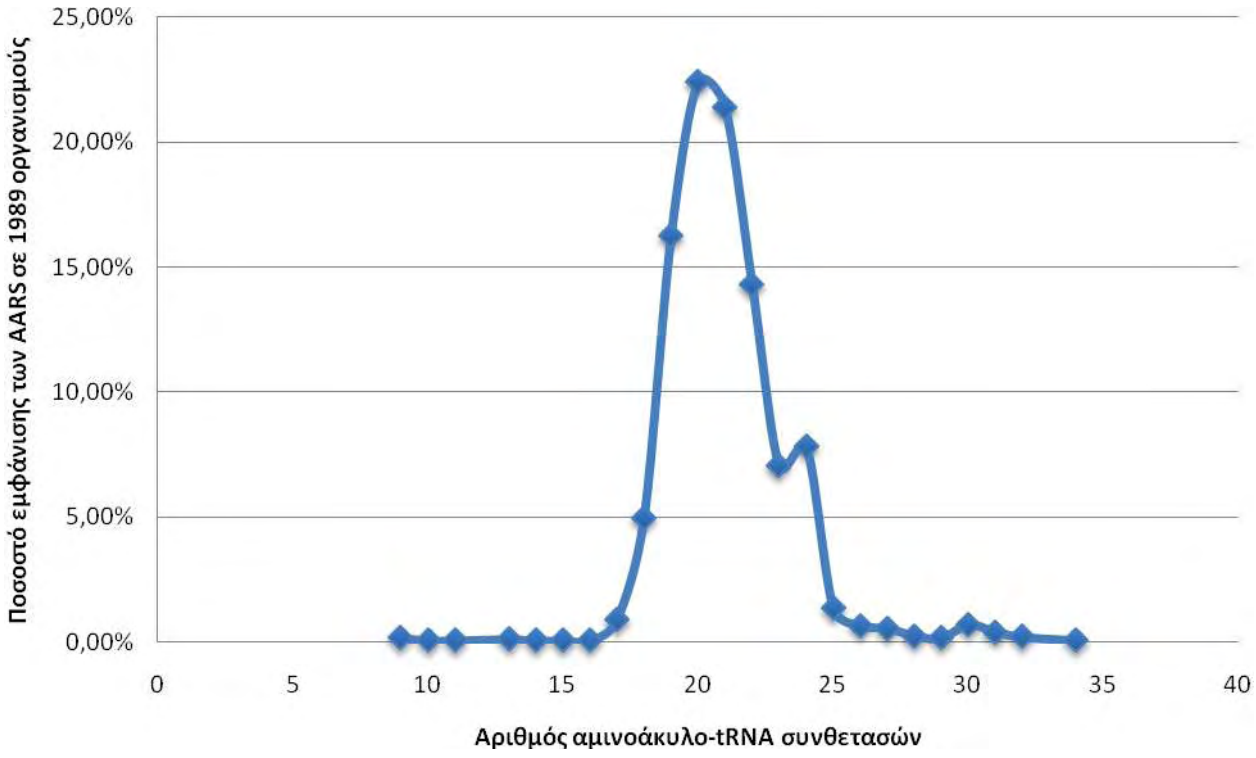
3.7 Η πλειοψηφία των προκαρυωτικών οργανισμών έχουν >20 AARS ανά γονιδίωμα

Περίπου 55% των προκαρυωτικών γονιδιωμάτων που σαρώθηκαν είχαν >20 AARS το καθένα, ενώ το 22% των γονιδιωμάτων είχαν <20 AARS το καθένα, όπως φαίνεται και αναλυτικά στον πίνακα 3.5 και εικόνα 3.10. Επομένως, η αρχική ιδέα της μιας AARS για κάθε αμινοξύ βλέπουμε ότι δεν ισχύει. Επιπλέον, σε ένα γονιδίωμα με 20 συνολικά AARS είναι δυνατόν να υπάρχουν κάποια/ες AARS που είναι διπλασιασμένες και κάποια/ες που έχουν απωλεσθεί.

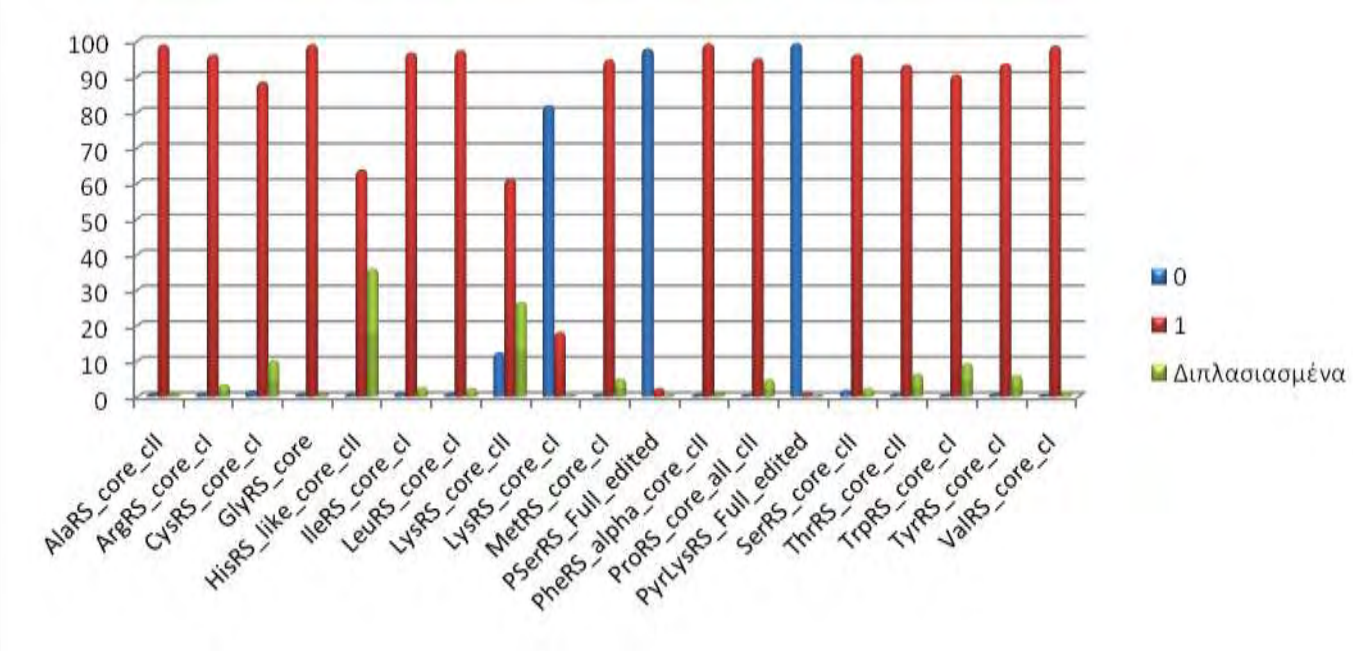
Αριθμός AARS	Συχνότητα Εμφάνισης	Σχετική Συχνότητα	Ποσοστό
34	1	0,000502765	0,05%
32	4	0,002011061	0,20%
31	8	0,004022122	0,40%
30	14	0,007038713	0,70%
29	3	0,001508296	0,15%
28	5	0,002513826	0,25%
27	11	0,005530417	0,55%
26	13	0,006535948	0,65%
25	27	0,013574661	1,36%
24	156	0,078431373	7,84%
23	140	0,070387129	7,04%

22	285	0,143288084	14,33%
21	426	0,214177979	21,42%
20	446	0,224233283	22,42%
19	323	0,162393162	16,24%
18	99	0,049773756	4,98%
17	18	0,009049774	0,90%
16	1	0,000502765	0,05%
15	1	0,000502765	0,05%
14	1	0,000502765	0,05%
13	2	0,00100553	0,10%
11	1	0,000502765	0,05%
10	1	0,000502765	0,05%
9	3	0,001508296	0,15%
Σύνολο	1989	1	100,00%

Πίνακας 3.5 Κατανομή συχνοτήτων του συνολικού αριθμού των AARS που εμφανίζονται σε 1989 οργανισμούς.



Εικόνα 3.10 Διαγραμματική απεικόνιση του πίνακα 4.7, όπου απεικονίζεται το ποσοστό των οργανισμών, ανάλογα με τον αριθμό των συνθετασών που έχουν στο γονιδίωμά τους



Εικόνα 3.11 Διαγραμματική απεικόνιση του ποσοστού % των γονιδιωμάτων που έχουν διπλασιασμό, ή όχι ή έχουν απώλεια ενός τύπου AARS. Μπλε χρώμα για τα γονίδια που έχουν απωλεσθεί, κόκκινο για τα γονίδια που υπάρχουν μία φορά στο γονιδίωμα των οργανισμών και πράσινο για τα γονίδια που υπάρχουν ≥2 στο γονιδίωμα των οργανισμών.

Στους παρακάτω πίνακες εμφανίζονται οι οργανισμοί με τους περισσότερους διπλασιασμούς καθώς και οι οργανισμοί με τις περισσότερες απώλειες. Οι οργανισμοί με τους περισσότερους διπλασιασμούς θα πρέπει να αποτελέσουν στόχους για μελλοντικές έρευνες ανακάλυψης αντιβιοτικών.

Οργανισμοί	Σύνολο AARS στο γονιδίωμα των οργανισμών	Αριθμός πρωτεϊνών στον οργανισμό
***Kitasatospora setae* KM 6054**	34	7.566
***Bacillus cereus* biovar anthracis CI**	32	5.558
***Bacillus cereus* Q1**	32	5.489
***Bacillus cereus* AH187**	32	5.783
***Bacillus cereus* NC7401**	32	5.754
***Bacillus thuringiensis* serovar chinensis CT 43**	31	6.206
***Bacillus thuringiensis* BMB171**	31	5.352
***Bacillus thuringiensis* serovar konkukian 97 27**	31	5.197
***Bacillus weihenstephanensis* KBAB4**	31	5.653
***Clostridium phytofermentans* ISDg**	31	3.902
***Bacillus cereus* B4264**	31	5.398
***Bacillus cereus* G9842**	31	5.857
***Bacillus cereus* 03BB102**	31	5.606
***Bacillus thuringiensis* serovar finitimus YBT 020**	30	5.782
***Bacillus anthracis* H9401**	30	5.791
***Actinoplanes* SE50 110**	30	8.247

***Paenibacillus* Y412MC10**	30	6.238
***Kribbella flavida* DSM 17836**	30	6.943
***Bacillus anthracis* Ames**	30	5.328
***Bacillus anthracis* Ames Ancestor**	30	5.484

Πίνακας 3.6 Οι 20 οργανισμοί που εντοπίστηκαν με τον μεγαλύτερο αριθμό AARSs στο γονιδίωμά τους.

Οργανισμοί	Σύνολο AARS στο γονιδίωμα των οργανισμών	Αριθμός πρωτεΐνών στον οργανισμό
***Borrelia burgdorferi* B31**	17	1.388
***Methanocaldococcus jannaschii* DSM 2661**	17	1.771
***Methanothermobacter thermautotrophicus* Delta H**	17	1.873
***Methanopyrus kandleri* AV19**	17	1.687
***Methanosaeta thermophila* PT**	17	1.696
***Methanocaldococcus fervens* AG86**	17	1.581
***Methanothermus fervidus DSM* 2088**	17	1.283
***Aerococcus urinae* ACS 120 V Col10a**	17	1.726
***Methanotorris igneus* Kol 5**	17	1.772
***Methanosaeta harundinacea* 6Ac**	17	2.371
***Rickettsia bellii* OSU 85 389**	16	1.475
***Candidatus Sulcia muelleri* SMDSEM**	15	242
***Paenibacillus polymyxa* M1**	14	3.508
***Mycobacterium tuberculosis* RGTB423**	13	3.622
***Advenella kashmirensis* WT001**	13	3.933
***Buchnera aphidicola* JF98 Acyrthosiphon pisum**	11	477
Candidatus Carsonella ruddii	10	192
***Candidatus Sulcia muelleri* DMIN**	9	226
***Candidatus Sulcia muelleri* GWSS**	9	227
***Candidatus Hodgkinia cicadicola* Dsem**	9	169

Πίνακας 3.7 Οι 20 οργανισμοί που εντοπίστηκαν με τον μικρότερο αριθμό AARSs στο γονιδίωμά τους

4. Συζήτηση –Συμπεράσματα

Τα νέα HMM μοντέλα που δημιουργήθηκαν στο εργαστήριο αξιολογήθηκαν και εντόπισαν επιτυχώς το 100% ενός συνόλου γνωστών ακολουθιών. Γενικά, τα νέα μοντέλα που δημιουργήθηκαν στο εργαστήριο λειτούργησαν καλύτερα από ότι τα μοντέλα που δημιουργήθηκαν από τα CDD. Ο λόγος είναι ότι χρησιμοποιήθηκαν περισσότερες ακολουθίες για την εκπαίδευση των νέων HMMs. Ωστόσο, υπάρχουν ακόμα περιθώρια βελτίωσης, καθώς όπως φαίνεται και από την εικόνα 4.9, τα CDD HMM μοντέλα των GlyRS, HisRS και TrpRS σε αρκετές περιπτώσεις βρίσκουν το στόχο τους με μεγαλύτερο score από ότι τα νέα HMM μοντέλα. Είτε τα νέα αυτά μοντέλα θα πρέπει να βελτιωθούν, χρησιμοποιώντας περισσότερες και πιο απομακρυσμένες εξελικτικά ακολουθίες για την εκπαίδευσή τους, είτε θα πρέπει να λειτουργούν σε συνδυασμό με τα αντίστοιχα CDD HMMs. Γενικά, και οι δύο τύποι μοντέλων λειτούργησαν πολύ ικανοποιητικά.

Περίπου 55% των προκαρυωτικών γονιδιωμάτων που σαρώθηκαν είχαν >20 AARS το καθένα, ενώ το 22% των γονιδιωμάτων είχαν <20 AARS το καθένα, όπως φαίνεται και αναλυτικά στον πίνακα 3.6 και εικόνα 3.10. Επομένως, η αρχική ιδέα της μίας AARS για κάθε αμινοξύ βλέπουμε ότι δεν ισχύει στην πλειονότητα των γονιδιωμάτων που μελετήθηκαν. Τα τρία γονίδια που εμφανίζονται πολύ συχνά διπλασιασμένα είναι η HisRS, LysRS (class II), CysRS.

Ο οργανισμός με τις περισσότερες AARS (34) ήταν ο *Kitasatospora setae* KM 6054, όπου εμφανίστηκαν διπλασιασμοί σε 8 διαφορετικές AARS. Ο οργανισμός αυτός είναι γνωστό ότι παράγει την setamycin (bafilomycin B1) που διαθέτει αντι-τριχομοναδική δράση (Ichikawa *et al.*, 2010). Το γένος στο οποίο ανήκει είναι μορφολογικά όμοιο με τους Streptomyces. Πιθανόν κωδικοποιεί πολύ περισσότερα αντιβιοτικά. Άλλοι οργανισμοί που βρέθηκαν με υψηλό αριθμό AARSs είναι στελέχη των ειδών *Bacilus cereus*, *Bacillus thuringiensis* και *Bacillus anthracis*. Ο αριθμός των AARSs που βρέθηκαν στο πρωτέωμα τους, κυμαίνεται μεταξύ 30-32. Θα πρέπει επίσης να αναφερθεί ότι αρκετοί από αυτούς τους οργανισμούς είναι υπεύθυνοι για δηλητηριάσεις π.χ. ο *Bacillus cereus* NC7401 είναι υπεύθυνος για δηλητηρίαση από φαγητό. Ο *Bacillus anthracis str. 'Ames Ancestor'*, είναι υπεύθυνος για την ασθένεια του άνθρακα (διαθέτει τα γονίδια για την παραγωγή της τοξίνης του άνθρακα). Στην αντίθετη πλευρά βρίσκονται 3 οργανισμοί στους οποίους εντοπίστηκαν μόλις 9 AARSs στο πρωτέωμα τους. Πρόκειται για τους *Candidatus Sulcia muelleri* DMIN, *Candidatus Sulcia muelleri* GWSS και *Candidatus Hodgkinia cicadicola* Dsem, οι οποίοι εμφανίζουν κάποια κοινά χαρακτηριστικά. Ο μέσος όρος των συνολικών πρωτεϊνών που παράγουν είναι 207 πρωτεΐνες, ενώ και οι 3 οργανισμοί είναι συμβιωτικοί. Φαίνεται ότι ο πολύ μικρός αριθμός των AARSs οφείλεται στην συμβιωτική σχέση, ενώ ο υψηλός αριθμός των AARSs μπορεί να συσχετίζεται με κάποιον μηχανισμό άμυνας-επίθεσης. Δηλαδή κάποιοι προκαρυώτες μπορεί να εμφάνισαν πολύ υψηλό αριθμό AARSs, λόγω της συνεχής επαφής τους με κάποιο φυσικό αντιβιοτικό (που στοχεύει σε AARSs), είτε οι ίδιοι παράγουν πολλά αντιβιοτικά (που στοχεύουν σε AARSs), είτε μπορεί να συμβαίνουν και τα 2 ταυτόχρονα.

Μελλοντικές μελέτες θα πρέπει να επικεντρωθούν στην φυλογενετική ανάλυση των διπλασιασμένων AARS, για να καθορισθεί κατά πόσο συνεισφέρει στον αυξημένο αριθμό των AARS ανά γονιδίωμα η οριζόντια μεταφορά γονιδίων και αν αυτά τα οριζόντια μεταφερόμενα γονίδια προσφέρουν προστασία από φυσικά

αντιβιοτικά. Επίσης, θα είναι πολύ ενδιαφέρον να μελετηθεί κατά πόσο οι διπλασιασμοί συγκεκριμένων AARS συσχετίζονται, π.χ., αν ο διπλασιασμός μιας AARS εμφανίζεται ταυτόχρονα με τον διπλασιασμό μιας άλλης AARS στο ίδιο γονιδίωμα. Ακόμα, θα πρέπει να μελετηθεί εάν ο διπλασιασμός στις AARS συσχετίζεται με διπλασιασμό σε άλλες οικογένειες γονιδίων (που δεν είναι συνθετάσες), π.χ. οι NRPS.

Αυτή η πληθώρα γονιδιωμάτων από διάφορα προκαρυωτικά είδη εξελικτικά απομακρυσμένα, αλλά και από πολλά στελέχη του ίδιου είδους με ελαφρά τροποποιημένες λειτουργίες θα επιτρέψει εξελικτικές μελέτες σε πολλά επίπεδα και υπόσχεται πολύ ενδιαφέροντα αποτελέσματα ως προς την βασική έρευνα και πολύ πιθανόν και ως προς την ανάπτυξη νέας γενιάς αντιβιοτικών.

Βιβλιογραφία:

Ambrogelly A, O'Donoghue P, Söll D, Moses S. (2010) A bacterial ortholog of class II lysyl-tRNA synthetase activates lysine. FEBS Lett. Jul 16;584(14):3055-60. Epub 2010 May 24.

Barton GJ, Sternberg MJ. (1990) Flexible protein sequence patterns. A sensitive method to detect weak structural similarities. J Mol Biol. Mar 20;212(2):389-402.

Barton GJ.(1990) Protein multiple sequence alignment and flexible pattern matching. Methods Enzymol.;183:403-28.

Becker, H. D. & Kern, D. (1998) Thermus thermophilus: a link in evolution of the tRNA-dependent amino acid amidation pathways Proc Natl Acad Sci U S A. Oct 27;95(22):12832-7.

Boyce. (2001). MRSA patients: proven methods to treat colonization and infection. J. Hosp. Infect. 48(Suppl. A):S9-S14.

Brown JR, Doolittle WF. (1999) Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases. J Mol Evol;49:485-495.

Chevenet F, Brun C, Bañuls AL, Jacq B, Christen R. (2006) TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics. Oct 10;7:439.

Churchill GA. (1989) Stochastic models for heterogeneous DNA sequences. Bull Math Biol.51(1):79-94.

Curnow AW, Hong K, Yuan R, Kim S, Martins O, Winkler W, Henkin TM, Söll D. (1997) Glu-tRNAGln amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. Proc Natl Acad Sci U S A. Oct 28;94(22):11819-26.

Curnow AW, Tumbula DL, Pelaschier JT, Min B, Söll D. (1998)Glutamyl-tRNA(Gln) amidotransferase in Deinococcus radiodurans may be confined to asparagine biosynthesis. Proc. Natl. Acad. Sci USA. Oct 27;95(22):12838-43.

Delarue. M. and Moras D. (1993) The aminoacyl-tRNA synthetase family: modules at work.Bioessays. Oct;15(10):675-87.

Dohm JC, Vingron M, Staub E. (2006) Horizontal gene transfer in aminoacyl-tRNA synthetases including leucine-specific subtypes. J Mol Evol. Oct;63(4):437-47.

Eddy SR. (1998) Profile hidden Markov models. Bioinformatics.14(9):755-63.

Edelmann P, Gallant J. (1977) Mistranslation in E. coli. Cell. Jan;10(1):131-7.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput.Nucleic Acids Res. 32(5):1792-1797.

Eriani G, Delarue M, Poch O, Gangloff J, Moras D. (1990) Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. Nature 347,203-206.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.

Feng DF, Doolittle RF. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol. 25(4):351-60.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A. (2010) The Pfam protein families database. Nucleic Acids Res. Jan;38(Database issue):D211-22.

Fourmy D, Mechulam Y, Blanquet S. (1995) Crucial Role of an Idiosyncratic Insertion in the Rossman Fold of Class I Aminoacyl-tRNA Synthetases: The Case of Methionyl-tRNA Synthetase. Biochemistry. Dec 5;34(48):15681-8.

Gallant P, Finn J, Keith D, Wendler P. (2000) The identification of quality antibacterial drug discovery targets: a case study with aminoacyl-tRNA synthetases February 0, Vol. 4, No. 1 , Pages 1-9

Gouy M., Guindon S. & Gascuel O. (2010) SeaView version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution 27(2):221-224.

Gribskov M, McLachlan AD, Eisenberg D. (1987) Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A. Jul;84(13):4355-8.

Hurdle JG, O'Neill AJ, Chopra I. (2005) Prospects for aminoacyl-tRNA synthetase inhibitors as new antimicrobial agents. Antimicrob Agents Chemother. Dec;49(12):4821-33.

Ibba M, Becker HD, Stathopoulos C, Tumbula DL, Söll D. The adaptor hypothesis revisited. Trends Biochem Sci. Jul;25(7):311-6.

Ibba M, Bono JL, Rosa PA, Söll D. (1997) Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete Borrelia burgdorferi. Proc Natl Acad Sci U S A. Dec 23;94(26):14383-8.

Ibba M, Losey HC, Kawarabayasi Y, Kikuchi H, Bunjun S, Söll D. (1999) Substrate recognition by class I lysyl-tRNA synthetases: a molecular basis for gene displacement. Proc Natl Acad Sci U S A. Jan 19;96(2):418-23.

Ibba M, Söll D. (1999) Quality control mechanisms during translation. Science. Dec 3;286(5446):1893-7.

Ibba M, Söll D. (2004) Aminoacyl-tRNAs: setting the limits of the genetic code. Genes Dev. Apr 1;18(7):731-8.

Ichikawa N, Oguchi A, Ikeda H, Ishikawa J, Kitani S, Watanabe Y, Nakamura S, Katano Y, Kishi E, Sasagawa M, Ankai A, Fukui S, Hashimoto Y, Kamata S, Otoguro M, Tanikawa S, Nihira T, Horinouchi S, Ohnishi Y, Hayakawa M, Kuzuyama T, Arisawa A, Nomoto F, Miura H,Takahashi Y, Fujita N. (2010) Genome sequence of Kitasatospora setae NBRC 14216T: an evolutionary snapshot of the family Streptomycetaceae. DNA Res. Dec;17(6):393-406.

J B Routien (1966) Identity of streptomycete producing antibiotic PA155A. J Bacteriol. April; 91(4): 1663.

Jacquin-Becker C, Ahel I, Ambrogelly A, Ruan B, Söll D, Stathopoulos C. (2002) Cysteinyl-tRNA formation and prolyl-tRNA synthetase. FEBS Lett. 2002 Mar 6;514(1):34-6.

Kim S, Lee SW, Choi EC, Choi SY. (2003) Aminoacyl-tRNA synthetases and their inhibitors as a novel family of antibiotics. Appl Microbiol Biotechnol. May;61(4):278-88.

Kitabatake M, Ali K, Demain A, Sakamoto K, Yokoyama S, Söll D. Indolmycin resistance of Streptomyces coelicolor A3(2) by induced expression of one of its two tryptophanyl-tRNA synthetases. J Biol Chem. Jun 28;277(26):23882-7.

Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol. Feb 4;235(5):1501-31.

Lapointe J, Duplain L, Proulx M. ()A single glutamyl-tRNA synthetase aminoacylates tRNAGlu and tRNAGln in Bacillus subtilis and efficiently misacylates Escherichia coli tRNAGln1 in vitro. J Bacteriol. Jan;165(1):88-93.

Loddenkemper, R., D. Sagebiel, and A. Brendel. 2002. Strategies against multidrug-resistant tuberculosis. Eur. Respir. J. Suppl. 36:66S-77S.

Luque I, Riera-Alberola ML, Andújar A, Ochoa de Alda JA. (2008) Intraphylum diversity and complex evolution of cyanobacterial aminoacyl-tRNA synthetases. Mol Biol Evol. Nov;25(11):2369-89.

Mailund T, Brodal GS, Fagerberg R, Pedersen CN, Phillips D. (2006) Recrafting the neighbor-joining method. BMC Bioinformatics. Jan 19;7:29.

Menichetti, F. 2005. Current and emerging serious gram-positive infections. Clin. Microbiol. Infect. 11(Suppl. 3):22-28.

Moras, D. (1992) Structural and functional relationships between aminoacyl-tRNA synthetases. Trends in Biochemical Sciences 17, 159-164

Olendzenski L, Liu L, Zhaxybayeva O, Murphey R, Shin DG, Gogarten JP. (2000) Horizontal transfer of archaeal genes into the deinococcaceae: detection by molecular and computer-based approaches. J Mol Evol. Dec;51(6):587-99.

Pohlmann J, Brötz-Oesterhelt H. (2004) New aminoacyl-tRNA synthetase inhibitors as antibacterial agents. Curr Drug Targets Infect Disord. Dec;4(4):261-72.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. (2012) The Pfam protein families database. Nucleic Acids Res. Jan;40(Database issue):D290-301.

Raczniak G, Becker HD, Min B, Söll D. (2001) A single amidotransferase forms asparaginyl-tRNA and glutaminyl-tRNA in Chlamydia trachomatis. J Biol Chem. Dec 7;276(49):45862-7. Epub 2001 Oct 3.

Ribas de Pouplana L, Schimmel P. (2000) A view into the origin of life: aminoacyl-tRNA synthetases.Cell Mol Life Sci. Jun;57(6):865-70.

Schimmel P, Ribas De Pouplana L. (2000) Footprints of aminoacyl-tRNA synthetases are everywhere. Trends Biochem Sci. May;25(5):207-9.

Schimmel P, Tao J, Hill J. (1998) Aminoacyl tRNA synthetases as targets for new anti-infectives. FASEB J. Dec;12(15):1599-609. Review.

Sekine S, Shimada A, Nureki O, Cavarelli J, Moras D, Vassylyev DG, Yokoyama S. (2001) Crucial role of the high-loop lysine for the catalytic activity of arginyl-tRNA synthetase. J Biol Chem. Feb 9;276(6):3723-6.

Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R.(1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains Nucleic Acids Res. Jan 1;26(1):320-2.

Sonnhammer EL, Eddy SR, Durbin R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. Jul;28(3):405-20.

Stathopoulos C, Jacquin-Becker C, Becker HD, Li T, Ambrogelly A, Longman R, Söll D. (2001)Methanococcus jannaschii prolyl-cysteinyl-tRNA synthetase possesses overlapping amino acid binding sites. Biochemistry. Jan 9;40(1):46-52.

Stathopoulos C, Li T, Longman R, Vothknecht UC, Becker HD, Ibba M, Söll D. (2000) One polypeptide with two aminoacyl-tRNA synthetase activities. Science. Jan 21;287(5452):479-82.

Terada T, Nureki O, Ishitani R, Ambrogelly A, Ibba M, Söll D, Yokoyama S (2002) Functional convergence of two lysyl-tRNA synthetases with unrelated topologies. Nat Struct Biol. Apr;9(4):257-62.

Vecchione JJ, Sello JK. (2008) Characterization of an inducible, antibiotic-resistant aminoacyl-tRNA synthetase gene in Streptomyces coelicolor. J Bacteriol. Sep;190(18):6253-7. Epub 2008 Jul 11.

Wilcox M, Nirenberg M. (1968) Transfer RNA as a cofactor coupling amino acid synthesis with that of protein. Proc Natl Acad Sci U S A. Sep;61(1):229-36.

Woese CR, Olsen GJ, Ibba M, Söll D. (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev. Mar;64(1):202-36.

Wolf YI, Aravind L, Grishin NV, Koonin EV. (1999) Evolution of aminoacyl-tRNA synthetases--analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res. Aug;9(8):689-710.

Yanagisawa & Kawakami, (2003) How does Pseudomonas fluorescens avoid suicide from its antibiotic pseudomonic acid?: Evidence for two evolutionarily distinct isoleucyl-tRNA synthetases conferring self-defense.J Biol Chem. Jul 11;278(28):25887-94.

Zeng Y, Roy H, Patil PB, Ibba M, Chen S. (2009) Characterization of two seryl-tRNA synthetases in albomycin-producing Streptomyces sp. strain ATCC 700974. Antimicrob Agents Chemother. Nov;53(11):4619-27.