ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Mining for bacterial terpene synthases

Παπά Ελισάβετ του Νικολάου

01609

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπουσα: Χατζηγεωργίου Άρτεμις, Καθηγήτρια

Συνεπιβλέποντες: Μακρής Αντώνης, Ερευνητής Α' Βαθμίδας

Ψωμόπουλος Φώτης, Ερευνητής Β' Βαθμίδας

Λαμία, 2022-2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ

ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

INSTITUTE OF APPLIED BIOSCIENCES
ΙΝΣΤΙΤΟΥΤΟ ΕΦΑΡΜΟΣΜΕΝΩΝ ΒΙΟΕΠΙΣΤΗΜΩΝ
CENTRE for RECEARCH and TECHNOLOGY-HELLAS

Εξόρυξη Βακτηριακών Τερπενικών Συνθασών

Παπά Ελισάβετ του Νικολάου

01609

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπουσα: Χατζηγεωργίου Άρτεμις, Καθηγήτρια

Συνεπιβλέποντες: Μακρής Αντώνης, Ερευνητής Α' Βαθμίδας

Ψωμόπουλος Φώτης, Ερευνητής Β' Βαθμίδας

Λαμία, 2022-2023

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [1], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: 15/02/2023

Η Δηλούσα

Παπά Ελισάβετ

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Εξόρυξη Βακτηριακών Τερπενικών Συνθασών

Παπά Ελισάβετ

Τριμελής Επιτροπή:

Χατζηγεωργίου Άρτεμις, Καθηγήτρια (Επιβλέπουσα)

Μπάγκος Παντελής, Καθηγητής

Μπράλιου Γεωργία, Επίκουρος Καθηγήτρια

ΠΡΟΛΟΓΟΣ

Με το πέρας αυτής της εργασίας, κλείνει και ένας κύκλος για τη ζωή μου. Η εκπόνηση και η ολοκλήρωση μιας πτυχιακής εργασίας δεν είναι εύκολη υπόθεση. Απαιτεί θυσίες, συνεχή προσπάθεια και αντοχές. Θα ήθελα να αναφέρω ότι η παρούσα πτυχιακή εργασία πραγματοποιήθηκε στο Ινστιτούτο Εφαρμοσμένων Βιοεπιστημών (ΙΝΕΒ) του Εθνικού Κέντρου Έρευνας και Τεχνολογικής Ανάπτυξης (ΕΚΕΤΑ).

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια μου κ. Χατζηγεωργίου Άρτεμις, για τη μοναδική ευκαιρία που μου έδωσε να συνεργαστώ με το συγκεκριμένο Ινστιτούτο, να αποκτήσω καινούριες γνώσεις και να αναπτύξω επιπλέον δεξιότητες.

Θα ήθελα να ευχαριστήσω θερμά, τον κ. Μακρή Αντώνη, Ερευνητή Α', συνεπιβλέποντα σε αυτή την εργασία και μέντορα μου, που πίστεψε από την πρώτη στιγμή σε εμένα αλλά και για την εξαιρετική συνεργασία και καθοδήγηση.

Θα ήθελα να ευχαριστήσω θερμά, τον κ. Ψωμόπουλο Φώτη, Ερευνητή Β', συνεπιβλέποντα σε αυτή την εργασία, χωρίς τον οποίο δε θα είχα καταφέρει να βγάλω εις πέρας την εργασία. Θα ήθελα να τον ευχαριστήσω για την καθοδήγηση, τη βοήθεια κάθε στιγμή που το χρειαζόμουν αλλά και για την άψογη συνεργασία.

Ξεχωριστό ευχαριστώ στην κ. Ανδρεαδέλλη Αγγελική, Υποψήφια Διδάκτωρ, στον κ. Γιώργο Τσιόλα και κ. Αντιόπη Τσουρέκη για την πολύτιμη βοήθεια αλλά και σε όλα τα μέλη του εργαστηρίου για την άψογη συνεργασία, τη συμπαράσταση στις δυσκολίες και για τις γνώσεις που με βοήθησαν να αποκτήσω.

Τέλος, θέλω να ευχαριστήσω τους δικούς μου ανθρώπους, οικογένεια και φίλους, που με στήριξαν και με βοήθησαν να αντέξω σε όλο το ταξίδι.

ΠΕΡΙΛΗΨΗ

Τα τερπενοειδή και τα ισοπρενοειδή αποτελούν σημαντική ομάδα δευτερογενών μεταβολιτών και την πιο ποικιλόμορφη ομάδα φυσικών προϊόντων. Απομονώνονται από φυτά, μύκητες και βακτήρια και χρησιμοποιούνται ευρέως ως πρόσθετα τροφίμων, αρώματα και φαρμακευτικά προϊόντα. Αν και η μεγαλύτερη ποικιλία τερπενικών συνθασών – που συμβάλλουν στη δημιουργία ενός σκελετού που μπορεί να διαμορφωθεί από επιπλέον ένζυμα για να γίνουν οι δομές αυτές πιο λειτουργικές – προέρχεται από τα φυτά, κάποια τερπενοειδή παράγονται μόνο από βακτηριακές τερπενικές συνθάσες και γι' αυτό η εξόρυξη τους από βακτηριακά γονιδιώματα είναι σημαντική. Οι βακτηριακές τερπενικές συνθάσες ήταν δύσκολο να ερευνηθούν καθώς έχουν χαμηλά ποσοστά ομοιότητας με τις τερπενικές συνθάσες των ευκαρυωτικών οργανισμών. Η βιοπληροφορική είναι μία επιστήμη που μπορεί να λύσει προβλήματα της βιολογίας με τη χρήση υπολογιστικών μέσων, αλγόριθμων και βάσεων δεδομένων. Έτσι λοιπόν, ο σκοπός της παρούσα πτυχιακής εργασίας είναι η εξόρυξη βακτηριακών τερπενικών συνθασών και άλλων πρωτεϊνών που σχετίζονται με τα τερπενοειδή, με την ανάπτυξη μεθοδολογίας που βασίζεται σε υπολογιστικά μέσα και εργαλεία βιοπληροφορικής. Η μεθοδολογία που αναπτύχθηκε έδωσε χρήσιμα αποτελέσματα για την εύρεση γονιδίων βιοσύνθεσης τερπενίων σε γονιδιώματα ακτινοβακτηριδίων όπως και αποτελέσματα που αποτελούν "τροφή για σκέψη" για επόμενα βήματα.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: ΤΕΡΠΕΝΟΕΙΔΗ, ΤΕΡΠΕΝΙΚΕΣ ΣΥΝΘΑΣΕΣ, ΑΚΤΙΝΟΒΑΚΤΗΡΙΑ, ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ, ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΛΟΓΙΑ

ABSTRACT

Terpenoids and isoprenoids are an important group of secondary metabolites and the most diverse group of natural products. They are isolated from plants, fungi and bacteria and are widely used as food additives, perfumes, and pharmaceuticals. Although the largest variety of terpene synthases – which contribute to the creation of a skeleton that can be formed by additional enzymes to make these structures more functional – comes from plants, some terpenoids are produced only by bacterial terpene synthases and therefore their extraction from bacterial genomes is important. Bacterial terpene synthases were difficult to investigate as they have low rates of similarity to the terpene synthases of eukaryotic organisms. Bioinformatics is a science that can solve problems of biology using computational means, algorithms, and databases. Thus, the purpose of this undergraduate thesis is the extraction of bacterial terpene synthases and other proteins related to terpenoids, by developing a methodology based on computational means and bioinformatics tools. The methodology developed gave useful results for identifying terpenoid biosynthetic genes in genomes of *Actinobacteria* as well as results that are "food for thought" for next steps.

KEYWORDS: TERPENOIDS, TERPENE SYNTHASES, ACTINOBACTERIA, BIOINFORMATICS, COMPUTATIONAL BIOLOGY

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ ΕΙΚΟΝΩΝ

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ ΠΙΝΑΚΩΝ

Ακρωνύμια

BGC: Biosynthetic Gene Clusters

DMAPP: dimethylallyl diphosphate

DXP: 1-deoxy-D-xylulose 5-phosphate

ENA: European Nucleotide Archive

FPP: Farnesyl Pyrophosphate

FQ: FASTQ αρχεία

GGPP: Geranylgeranyl Pyrophosphate

GPP: Geranyl Pyrophosphate

HMM: Hidden Markov Models

IGV: Integrative Genomics Viewer

IPP: Isopentenyl Diphosphate

MEP: 2-C-methyl-D-erythritol 4-phosphate

MEP pathway: Methylerythritol Phosphate pathway

MSA: Multiple Sequence Alignment

MVA pathway: Mevalonic Acid pathway

SMs: Secondary Metabolites

TSs: Terpene Synthases

TSS: Transcription Start Site

Κεφάλαιο 1ο

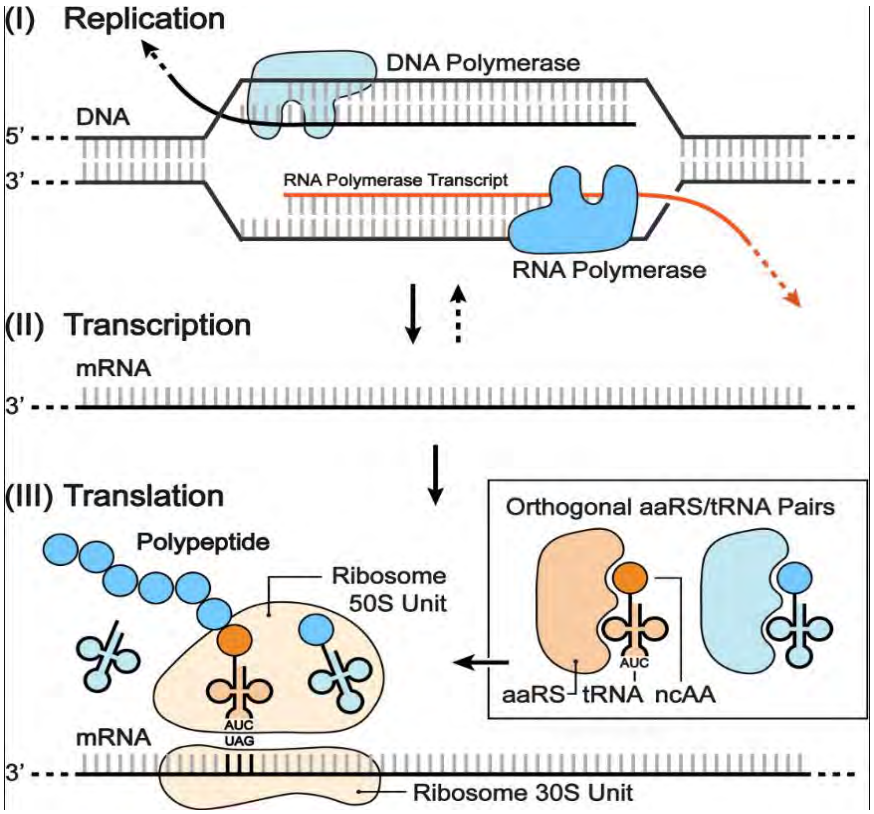
1 ΕΙΣΑΓΩΓΗ

Βιολογικό Υπόβαθρο

Στο παρόν κεφάλαιο θα αναλυθούν βιολογικές έννοιες και διαδικασίες απαραίτητες για την κατανόηση του θέματος της παρούσας εργασίας.

1.1 Κεντρικό δόγμα της μοριακής βιολογίας

Το κεντρικό δόγμα της μοριακής βιολογίας ορίζει τη ροή της γενετικής πληροφορίας από το DNA στην πρωτεΐνη. Αποτελείται από 3 βασικά βήματα την αντιγραφή του DNA, τη μεταγραφή σε mRNA και τέλος την μετάφραση σε πρωτεΐνες (Εικόνα 1) (CRICK, 1970; Diercks, Dik, & Schultz, 2021).



Εικόνα 1. Τα 3 βασικά βήματα του κεντρικού δόγματος της μοριακής βιολογίας. (I) Η αντιγραφή του DNA, (II) Η μεταγραφή του σε mRNA, (III) Η μετάφραση σε πρωτεΐνη (Diercks, Dik, & Schultz, 2021)

1.2 Γονιδιακή Έκφραση

Σύμφωνα με τον ορισμό του National Human Genome Research Institute, «Γονιδιακή έκφραση είναι η διαδικασία κατά την οποία ένα γονίδιο κωδικοποιείται και μετατρέπεται σε μία λειτουργία». Αυτό επιτυγχάνεται μέσω της μεταγραφής μορίων RNA που κωδικοποιούν πρωτεΐνες ή μη-κωδικών μορίων RNA που υπηρετούν άλλες λειτουργίες. Η γονιδιακή έκφραση είναι σαν ένας διακόπτης, που ελέγχει πότε και που μόρια RNA και πρωτεΐνες φτιάχνονται και επίσης καθορίζει πόση ποσότητα θα δημιουργηθεί. Η διαδικασία αυτή αλλάζει ανάλογα με τις συνθήκες και τους τύπους κυττάρων» (National Human Genome Research Institute , 2023). Η γονιδιακή έκφραση παρέχει μία "γέφυρα" μεταξύ της κωδικοποιημένης πληροφορίας που υπάρχει στο γονίδιο και στο τελικό λειτουργικό προϊόν (Volgin, 2014).

1.3 Υποκινητές

Ο υποκινητής είναι μία περιοχή DNA που βρίσκεται πριν το γονίδιο, κοντά στο περιοχή έναρξης της μεταγραφής (Transcription Start Site, TSS), έχει συγκεκριμένη αλληλουχία και δίνει την έναρξη της μεταγραφής (Lin, Liang, Tang, & Chen, 2019). Εκεί προσδένεται η RNA πολυμεράση. Στους ευκαρυωτικούς οργανισμούς με τη βοήθεια γενικών μεταγραφικών παραγόντων που προσδένονται στην RNA πολυμεράση, ξεκινάει η μεταγραφή (Alberts, et al., 2014).

1.3.1 Υποκινητής στα βακτήρια

Στα βακτήρια, υπάρχει μία βασική υπομονάδα που ονομάζεται παράγοντας σίγμα (σ) και αναγνωρίζει την αλληλουχία του υποκινητή. Οι βακτηριακοί υποκινητές

έχουν αλληλουχίες DNA στις θέσεις -35 και -10, ανοδικά του σημείο έναρξης της μεταγραφής(TSS) (Alberts, et al., 2014).

1.4 Μεταγραφικοί Παράγοντες

Όλα τα γονίδια βακτηριακά και ευκαρυωτικά έχουν ρυθμιστικές αλληλουχίες που είναι υπεύθυνες για την ενεργοποίηση ή απενεργοποίησή τους. Στα βακτήρια συνήθως οι αλληλουχίες αυτές δρουν σαν διακόπτες και έχουν μικρό μήκος περίπου 10 ζεύγη βάσεων. Για να λειτουργήσουν αυτές οι ρυθμιστικές αλληλουχίες αναγνωρίζονται από πρωτεΐνες που λέγονται μεταγραφικοί παράγοντες. Και με αυτόν τον τρόπο ελέγχεται η μεταγραφή ενός γονιδίου (Alberts, et al., 2014).

1.5 Οπερόνια

Στα βακτήρια είναι πιθανή η δημιουργία οπερονίου. Τα οπερόνια είναι ομάδες γονιδίων που κωδικοποιούν ένζυμα της ίδιας μεταβολικής οδού, διατάσσονται διαδοχικά και μεταγράφονται από έναν κοινό υποκινητή (Alberts, et al., 2014).

1.6 Ενεργοποιητές και καταστολείς

Ένας άλλος μηχανισμός για τη ρύθμιση της έκφρασης των γονιδίων είναι οι ενεργοποιητές και οι καταστολείς. Οι θέσεις που προσδένονται οι ενεργοποιητές ονομάζονται ενισχυτές (enhancers) και αυξάνουν τον ρυθμό της μεταγραφής ενώ οι καταστολείς μειώνουν τον ρυθμό της μεταγραφής και τη δημιουργία συμπλόκου έναρξης. Είναι ενδιαφέρουσα η παρατήρηση ότι αυτές οι θέσεις μπορεί να βρίσκονται σε μεγάλη απόσταση ανοδικά ή καθοδικά του γονιδίου (Alberts, et al., 2014).

1.7 Δευτερογενείς Μεταβολίτες (Secondary Metabolites, SMs)

Οι δευτερογενείς μεταβολίτες είναι φυσικά προϊόντα που παράγονται από βακτήρια, μύκητες και φυτά και έχουν μεγάλη ποικιλία στη δομή και στις βιολογικές διεργασίες. Αν και χαρακτηρίστηκαν δευτερογενείς καθώς θεωρούνταν πως δεν ήταν απαραίτητοι για την επιβίωση και αναπαραγωγή ενός οργανισμού πλέον θεωρείται πως έχουν σημαντικό ρόλο στην επιβίωση καθώς καθορίζουν την αντίδραση το οργανισμού σε περιβαλλοντικές συνθήκες (Mosunova, Navarro-Muñoz, & Collemare, 2021).

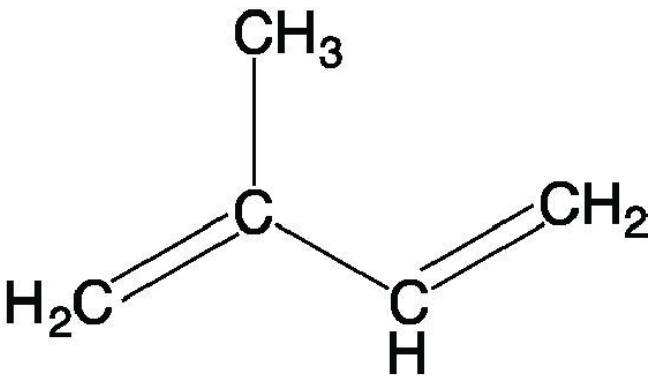
1.8 Βιοσυνθετικές Συστοιχίες Γονιδίων (Biosynthetic Gene Clusters, BGC)

Τα BGC είναι εντοπισμένες ομάδες 3 η περισσοτέρων γονιδίων που βρίσκονται στον ίδιο γενετικό τόπο και κωδικοποιούν ένα κοινό μονοπάτι βιοσύνθεσης για την παραγωγή ενός δευτερογενούς μεταβολίτη (Medema, et al., 2015).

1.9 Τερπενοειδή

Τα τερπενοειδή και τα ισοπρενοειδή αποτελούν σημαντική ομάδα δευτερογενών μεταβολιτών και την πιο ποικιλόμορφη ομάδα φυσικών προϊόντων (Buckingham, 1997). Απομονώνονται από φυτά (Tholl, 2015), μύκητες και βακτήρια (Yamada, et al., 2015) και χρησιμοποιούνται ευρέως ως πρόσθετα τροφίμων, αρώματα και φαρμακευτικά προϊόντα (Leferink & Scrutton, 2022). Τα τερπενοειδή είναι παράγωγα που βασίζονται στο ισοπρένιο (Εικόνα 2), ενός μορίου 5 ατόμων άνθρακα (Jones , Ormondroyd, Curling, Popescu, & Popescu, 2017). Τα τερπένια είναι κυρίως οι υδρογονάνθρακες που παράγονται από το ισοπρένιο, ενώ τα τερπενοειδή είναι οξυγονωμένα παράγωγα αυτών των υδρογονανθράκων (Britannica, 2018). Τα τερπένια σχηματίζονται με την ένωση μορίων ισοπρενίου και δημιουργούν δεσμούς με

διάφορους τρόπους γι' αυτό και υπάρχει μεγάλη ποικιλία στις δομές (Gao, Honzatko, & Peters, 2012). Ο σχηματισμός πολύπλοκων δομών γίνεται με τη βοήθεια ενζύμων, των τερπενικών/ τερπενοειδών συνθασών (Terpene Synthases, TSs) (ή τερπενικών/ τερπενοειδών κυκλασών) που καταλύουν τις αντιδράσεις σχηματισμού δακτυλίων (Baunach, Franke, & Hertweck, 2015) .



Εικόνα 2. Η δομή του ισοπρενίου

1.9.1 Είδη τερπενίων

Διαχωρίζονται σε ομάδες ανάλογα με τον αριθμό μορίων ισοπρενίου (C_5H_8) από τον οποίο αποτελούνται:

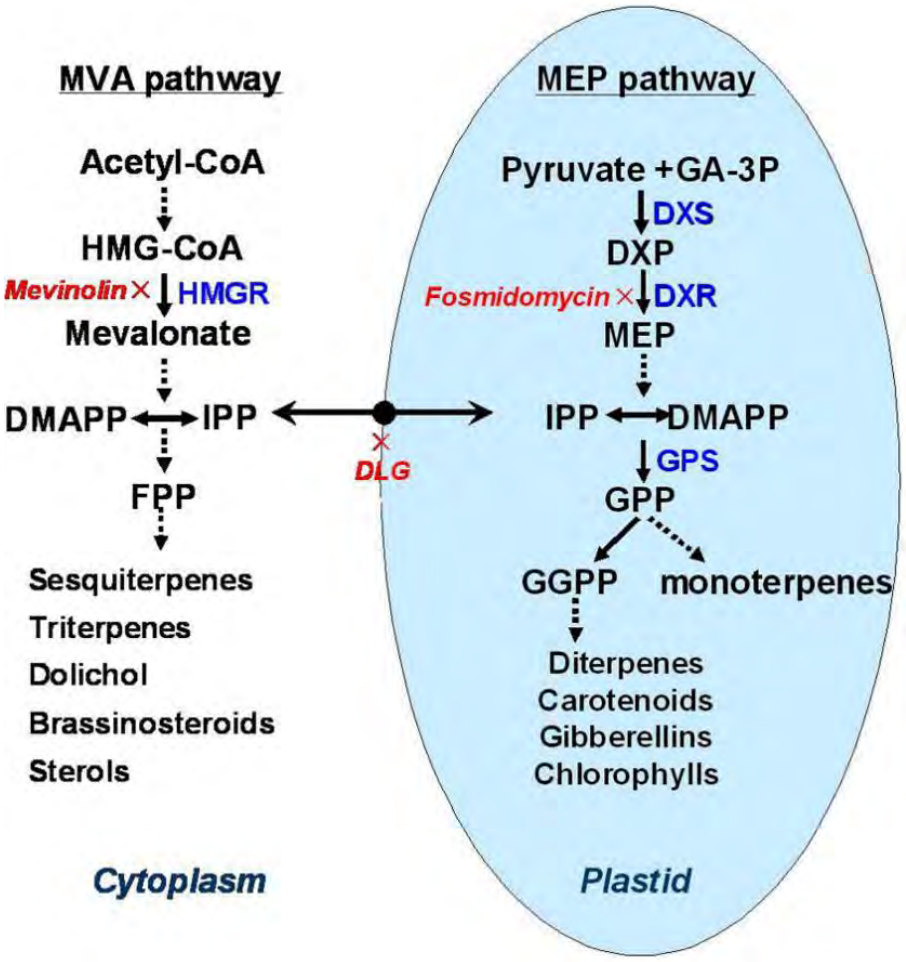
- Ημιτερπένια (C_5H_8): έχουν 1 μόριο ισοπρενίου
- Μονοτερπένια ($C_{10}H_{16}$): έχουν 2 μόρια ισοπρενίου
- Σεσκιτερπένια ($C_{15}H_{24}$): έχουν 3 μόρια ισοπρενίου, πχ. Η βιταμίνη Α
- Διτερπένια ($C_{20}H_{32}$): έχουν 4 μόρια ισοπρενίου
- Τριτερπένια ($C_{30}H_{48}$): έχουν 6 μόρια ισοπρενίου, πχ. σκουαλένιο
- Τετρατερπένια ($C_{40}H_{64}$) : έχουν 8 μόρια ισοπρενίου, πχ. καροτενοειδή (Britannica, 2018)

1.9.2 Μονοπάτια βιοσύνθεσης τερπενοειδών

Η βιοσύνθεση των τερπενοειδών γίνεται από 2 πρόδρομα μόρια C5, το διφοσφωρικό ισοπεντενύλιο (isopentenyl diphosphate, IPP) και το διφοσφωρικό διμεθαλλύλιο (dimethylallyl diphosphate, DMAPP) (McGarvey & Croteau, 1995). Η βασική παραγωγή τους γίνεται μέσω του μονοπατιού του μεβαλονικού οξέος (mevalonic acid - MVA pathway) (Kampranis & Makris, 2012) στους ευκαρυώτες, στα αρχαία και σε κάποια βακτήρια (Heidrun Karlic & Franz Varga, 2019) αλλά και μέσω του μονοπατιού της φωσφορικής μεθυλ-ερυθριτόλης (methylerythritol phosphate - MEP pathway) ή μη μεβαλονικό μονοπάτι (Εικόνα 3). Τα δύο πρόδρομα μόρια IPP και DMAPP στο μονοπάτι MVA συντίθενται από το μεβαλονικό οξύ, το οποίο έχει ως πρόδρομη ουσία το ακετυλο-συνένζυμο Α (Kuzuyama, 2002). Στο MEP μονοπάτι, τα πρόδρομα μόρια αποτελούνται από διφωσφορική γλυκεραλδεΰδη και πυροσταφυλικό και παράγουν διφωσφορική 1-δεοξύ-D-ξυλουλόση (1-deoxy-D-xylulose 5-phosphate, DXP) που ισομερείται σε διφωσφορική μεθυλερυθριτόλη (2-*C*-methyl-D-erythritol 4-phosphate, MEP). Ακολουθεί μία σειρά αντιδράσεων για να καταλήξει σε μόρια IPP και DMAPP. Τα ένζυμα που χρησιμοποιούνται για να καταλυθούν αυτές οι αντιδράσεις είναι τα IspC, IspD, IspE, IspG και IspH (Zhao, Chang, Xiao, Liu, & Liu, 2013). Τα μόρια IPP και DMAPP μέσω της δράσης των πρενυλοτρανσφερών, δημιουργούν σύμπλοκα όπως το διφωσφορικό γερανύλιο (geranyl pyrophosphate GPP), το διφωσφορικό φαρνεσύλιο (farnesyl pyrophosphate FPP) και το διφωσφορικό γερανυλ-γερανύλιο (geranylgeranyl pyrophosphate GGPP) (McGarvey & Croteau, 1995).

Οι ευκαρυώτες χρησιμοποιούν το MVA μονοπάτι για την παραγωγή τερπενίων, τα φυτά χρησιμοποιούν το MVA στο κυτοσόλιο και το MEP μονοπάτι στα πλαστίδια ενώ τα βακτήρια χρησιμοποιούν το MEP μονοπάτι με εξαίρεση κάποια είδη όπως τα Actinobacteria, Bacteroidetes, Chloroflexi, Firmicutes και Proteobacteria που

χρησιμοποιούν και το MVA μονοπάτι (Rinaldi, Ferraz, & Scrutton, 2022; Henry, Gutensohn, Thomas, Noel, & Dudareva, 2015; Thomas, Louie, Lubin, Lundblad, & Noel, 2019).



Εικόνα 3. Το μονοπάτι του μεβαλονικού οξέος (MVA pathway) και το μονοπάτι της φωσφορικής μεθυλ-ερυθριτόλης (MEP pathway) για την παραγωγή τερπενοειδών (Yang, et al., 2012)

1.9.3 Γονίδια τερπενοειδών σε ομάδες

Στη φύση υπάρχουν γονίδια που λειτουργούν και εκφράζονται ως ομάδες. Αυτά σχηματίζουν τα λεγόμενα BGC *1.8 Βιοσυνθετικές Συστοιχίες Γονιδίων (Biosynthetic Gene Clusters, BGC)*. Ένα παράδειγμα στα τερπενοειδή είναι τα γονίδια hpnC, hpnD, hpnE, και hpnF, τα οποία είναι υπεύθυνα για την μετατροπή του FPP σε χοπένιο ή άλλους πεντακυκλικούς μεταβολίτες (Pan, et al., 2015).

1.9.4 Τερπενοειδή στα βακτήρια

Αν και η μεγαλύτερη ποικιλία τερπενικών συνθασών – που συμβάλλουν στη δημιουργία ενός σκελετού που μπορεί να διαμορφωθεί από επιπλέον ένζυμα για να γίνουν οι δομές αυτές πιο λειτουργικές – προέρχεται από τα φυτά, κάποια τερπενοειδή παράγονται μόνο από βακτηριακές τερπενικές συνθάσες και γι' αυτό η εξόρυξη τους από βακτηριδιακά γονιδιώματα είναι σημαντική (Rinaldi, Ferraz, & Scrutton, 2022). Οι βακτηριακές τερπενικές συνθάσες ήταν δύσκολο να ερευνηθούν καθώς έχουν χαμηλά ποσοστά ομοιότητας με τις τερπενικές συνθάσες των ευκαρυωτικών οργανισμών (Rinaldi, Ferraz, & Scrutton, 2022). Παρόλα αυτά μετά από πολλές δοκιμές και χρήση Hidden Markov Models, προέκυψαν περίπου 600 υποψήφια γονίδια τερπενοειδών σε χιλιάδες βακτηριδιακά γονιδιώματα κυρίως του είδους *Streptomyces* (Dickschat, 2016).

Σε αυτό το κεφάλαιο θα παρουσιαστούν κάποιες βασικές πληροφορίες για την επιστήμη της βιοπληροφορικής, τι εφαρμογές έχει και πιο συγκεκριμένα πληροφορίες για το βιοπληροφορικό υπόβαθρο αυτής της πτυχιακής εργασίας.

1.10 Τι είναι η Βιοπληροφορική

Η βιοπληροφορική είναι η επιστήμη που συνδυάζει τα βιολογικά δεδομένα με την ανάλυση τους με υπολογιστικά μέσα (Εικόνα 4). Είναι η χρήση αλγόριθμων και βάσεων δεδομένων για την ανάλυση βιολογικών δεδομένων όπως πρωτεϊνών, γονιδίων κ.α. (Pevsner, 2015). Βοηθάει στον χειρισμό βιολογικών δεδομένων μεγάλου όγκου και περιέχει διαδικασίες ανάλυσης και αποθήκευσης αυτών (Blackshaw, 2022). Τα δεδομένα που χρησιμοποιούνται για μία βιοπληροφορική ανάλυση είναι βιολογικά και η μεθοδολογία αποτελείται από υπολογιστικά μέσα (Benton, 1996).

Εικόνα 4. Ροή Βιοπληροφορικής Ανάλυσης

Η βιοπληροφορική έχει γνωρίσει μεγάλη ανάπτυξη τα τελευταία χρόνια καθώς είναι όλο και πιο επιτακτική ανάγκη η ανάλυση μεγάλου όγκου δεδομένων και όχι μόνο. Κάποιες άλλες εφαρμογές της βιοπληροφορικής μπορεί να είναι οι εξής:

- Αναλύσεις που αφορούν την έκφραση γονιδίων
- Πρόβλεψη δομής και λειτουργίας γονιδίων
- Πρόβλεψη δομής και λειτουργίας πρωτεϊνών
- Ανάλυση μοριακών μονοπατιών κ.α. (Tsoka & Ouzounis, 2000)

Υπάρχουν 2 λογισμικά στη βιοπληροφορική, το διαδικτυακό και της γραμμής εντολών, αλλά υπάρχουν και εργαλεία που γεφυρώνουν τις 2 προσεγγίσεις (Pevsner, 2015).

1.11 Hidden Markov Models (HMM)

Η μεθοδολογία αυτής της εργασίας βασίστηκε στη λογική των πιθανολογικών μοντέλων Hidden Markov Models και συγκεκριμένα στα profile Hidden Markov Models.

Το Hidden Markov Model είναι ένα στατιστικό, πιθανοκρατικό μοντέλο που χρησιμοποιείται για την περιγραφή της εξέλιξης γεγονότων που παρατηρούνται, τα οποία εξαρτώνται από εσωτερικούς παράγοντες που δεν γίνονται απευθείας αντιληπτά. Αποτελείται από 2 διαδικασίες, μία κρυφή κατάσταση και μία παρατήρηση, ένα γεγονός (Rabiner, 1989).

1.12 Profile HMM

Ένα προφίλ HMM είναι ένας πίνακας που μετατρέπει μία πολλαπλή στοίχιση αλληλουχιών (Multiple Sequence Alignment), σε ένα σύστημα που δείχνει το πόσο πιθανό είναι να βρίσκεται ένα νουκλεοτίδιο σε συγκεκριμένη θέση και είναι κατάλληλα για αναζήτηση ομόλογων αλληλουχιών και στοίχισης (Eddy, 1998). Έχουν χρησιμοποιηθεί στη στοίχιση αλληλουχιών, στην πρόβλεψη πρωτεϊνικής δομής, στην πρόβλεψη γονιδίων κ.α. Τα προφίλ HMM μετατρέπουν μία πολλαπλή στοίχιση σε πίνακα βαθμολόγησης ανά θέση και αυτό βοηθά στον εντοπισμό συγκεκριμένων περιοχών (Pevsner, 2015).

1.13 Γονιδιακή Ανάλυση

Η γονιδιακή ανάλυση αποτελείται από 2 βασικά μέρη, την διαδικασία συναρμολόγησης της πλήρης αλληλούχιας των αναγνώσεων (Assembly) και την διαδικασία ανίχνευσης των γονιδίων και σε ποιες περιοχές βρίσκονται στην αλληλουχία, όπως και ο καθορισμός της δομή και της λειτουργίας αυτών των γονιδίων (Annotation) (Kong, et al., 2019)

Για να μπορούν να εφαρμοστούν αυτά τα βήματα θα πρέπει να υπάρχουν διαθέσιμα δεδομένα αλληλούχισης DNA/RNA. Η αλληλούχιση είναι η διαδικασία κατά την οποία καθορίζεται η αλληλουχία, η σειρά δηλαδή με την οποία εμφανίζονται τα νουκλεοτίδια. Οι πιο γνωστές τεχνικές αλληλούχισης είναι:

- Αλληλούχιση κατά Sanger: Το μόριο DNA που θα αλληλουχηθεί αποδιατάσσεται και στους κλώνους προστίθεται εκκινητής μήκους ~20 βάσεων. Ξεκινάει η σύνθεση με παρουσία DNA πολυμεράσης και δεοξυνουκλεοτιδίων. Η σύνθεση σταματά κάθε φορά που συνδέεται ένα

διδεοξυνουκλεοτίδιο καθώς η ριβόζη του δεν έχει την 3’ υδροξυλομάδα. Έτσι δημιουργούνται πολλά κομμάτια διαφορετικού μήκους τα οποία ηλεκτροφορούνται σε τζελ και διαβάζονται, καθώς τα διαπερνάει μία ακτίνα λέιζερ, ανάλογα με το φθορισμό τους (Pevsner, 2015).

- Αλληλούχιση Επόμενης Γενιάς: είναι νέες τεχνολογίες που έχουν αναπτυχθεί τα τελευταία χρόνια και έχουν μειώσει το χρόνο και το κόστος της αλληλούχισης καθώς και παράγουν μεγαλύτερο αριθμό αναγνώσεων. Υπάρχουν διάφορες τεχνολογίες, κάποιες από αυτές είναι οι εξής: Illumina, SOLiD, Ion Torrent, Pacific Biosciences κ.α (Pevsner, 2015).

1.14 Πολλαπλή Στοίχιση Αλληλουχιών (MSA)

Κατά την μελέτη μιας πρωτεΐνης αρκετό ενδιαφέρον έχει η συσχέτιση της με άλλες πρωτεΐνες. Με μία πολλαπλή στοίχιση αλληλουχιών στοιχίζονται 3 ή περισσότερες αλληλουχίες πρωτεϊνών ή νουκλεϊκών οξέων οι οποίες θεωρείται πως έχουν προέλθει από έναν κοινό εξελικτικά πρόγονο. Βοηθάει στην εύρεση λειτουργίας, δομής και εξέλιξης της πρωτεΐνης που μελετάται (Pevsner, 2015).

1.15 E-value και Bit-score στο BLAST

Το λεγόμενο e-value είναι ο αριθμός των αναμενόμενων επιτυχιών (hits) με παρόμοια βαθμολογία ποιότητας που θα μπορούσαν να εμφανιστούν τυχαία. Π.χ. ένα e-value 10, σημαίνει πως αναμένεται περίπου 10 επιτυχίες που θα προκύψουν να έχουν βρεθεί τυχαία. Μπορεί να χρησιμοποιηθεί για έλεγχο της ποιότητας των αποτελεσμάτων. Όσο πιο μικρό το e-value τόσο χαμηλότερη η πιθανότητα μια επιτυχία

να έχει εμφανιστεί τυχαία, άρα τόσο καλύτερη η αντιστοιχία. (Altschul, Gish, Miller, Myers, & Lipman, 1990; Scholz, n.d.)

Το bit-score είναι το απαιτούμενο μέγεθος μιας βάσης αλληλουχιών στην οποία η τρέχουσα επιτυχία μπορεί να έχει εμφανιστεί τυχαία. Δεν εξαρτάται από το μέγεθος της βάσης. Όσο πιο υψηλό το Bit-score τόσο καλύτερο το ποσοστό ομοιότητας των αλληλουχιών (Altschul, Gish, Miller, Myers, & Lipman, 1990; Scholz, n.d.).

1.16 Χρήσιμες μορφές αρχείων

1.16.1 FASTQ

Τα FASTQ (FQ) αρχεία περιέχουν τα κομμάτια που διαβάζονται από την αλληλούχιση και κάθε τέτοιο κομμάτι αποτελείται από 4 γραμμές. Η πληροφορία που παρέχεται από κάθε γραμμή είναι η εξής:

- Γραμμή 1: ξεκινάει με το σύμβολο "@" και προσδιορίζει τον καταχωρητή. Περιλαμβάνει το ID της ανάγνωσης, το μήκος της, το μηχάνημα που χρησιμοποιήθηκε κ.α.
- Γραμμή 2: περιλαμβάνει την ακολουθία
- Γραμμή 3: αρχίζει με το σύμβολο "+" και περιλαμβάνει ή μόνο αυτό ή και κάποιες άλλες πληροφορίες
- Γραμμή 4: περιλαμβάνει τη βαθμολογία ποιότητας (quality score) για κάθε βάση και γι' αυτό το μέγεθος αυτής της γραμμής είναι ίδιο με το μέγεθος της ακολουθίας στη γραμμή 2

  (Pevsner, 2015)

1.16.2 FASTA

Ένα αρχείο σε μορφή FASTA (Εικόνα 5) είναι ένα αρχείο κειμένου που περιλαμβάνει τις αλληλουχίες νουκλεοτιδίων ή πρωτεϊνών. Η πρώτη γραμμή περιλαμβάνει την επικεφαλίδα, η οποία παρέχει χρήσιμες περιγραφικές πληροφορίες για την αλληλουχία. Η επικεφαλίδα ξεκινά πάντα με το χαρακτηριστικό σύμβολο “>”, και παρέχει πληροφορίες όπως το ID, το όνομα της αλληλουχίας, το όνομα του γονιδίου ή και του οργανισμού.

```
>ENA|LCZK01000006|LCZK01000006.1 Actinobacteria bacterium IMCC26207 IMCC26207_106, whole genome shotgun sequence.
GTCAGCGGAACACGGTCGAGAACCGAGTACGACGCCCAAGGCCGCGCTGTCACAGCTATT
GGGCCGTCGAGGCGTGCGATGACAACCGGTTCGGGTGCGCCTGTTGTCGAAACGCGTTAC
GACACCTATGACGCCGACCCGGGTGCGCGGATCCGTATGGAACCCATCACCGGTCTCGTT
GGCATGGTCTACAACAACACCGGGTTTACTGGTGTGCCGTCTGATCGTTCGATCGGCCCC
CGCGTCAATGGCCGCCCGACACAAACCATGGCGTTCGGGTACAACGCGAACCCGTCGGGT
GCTGGTGGTGTCTGGTCGATGCGACTCGCCGGCAAATTCATAGCCCCAACTGAGGGCACC
TACCAGTTCAGAAGCACCACAAACGCCCGGCTCACGATCAACGCAATGACCTGCACGGTT
CAAGCACCGTGTCGTCAACAACTTCGCCGTAATGCTGCAGCGGACATCCAAGCCGAAATT
GTCTCCGGAACCAACGGTGTAGCAAATGTTTCGTTGGAAGTAGCCGGAGCAGACAACCGC
TGGCAGCCAATCAACGCAGCACAAACCGCACCTGGGCTCAACGAGGTCACCGAACAAGAA
ACCACCGATCAGTTAGCCACCGGCCGCGGCGCTGTATCACTGAAACAAACCTCTGAGTAC
GACCTCGAAACCGGGTCTGCCAACTTGTTGAAACAAACCTCATCTTCAGGTGCGACCGTG
TCGCATACCTGGGCTGAAAACACCGGTGTTGATGGCCAGTACGGCCAACACCTCACATGG
ACTGCACCAGATGGGCACACCACCTATTACGGATACCACGGTGCAACCGCACAGGCTTCT
GGTTGTGACGGACCGCCCGCAAACCAAGGCGGCCAACTCGCAATCACCACCCGCCCCGGT
GCAATCACCACAGAACAAACCTACGACGCTGCCGGCAGCACCAACAAAGCAACCGGCCAA
GGCACCGAAACCTGTGTCAACTACCGCCCCGACGGCTCAGCAGAATCAGGTGCTGTCACC
GGAGACGGCCCCTCCTACGCGGTGTCATCAACCCCGATGCTGTTCAACAACCCGCTCGTC
TCATCCGGGACAACCACCACACAAGGTGTCACAAATACTGCTACTCAGTCCGTGTCAATC
ACCGGGCAACTCTTCCAATCAACAGACTCCCACGGAACCGTCACCACTCACGCGTACGAC
CCGAACACCGGGCACCCAACCGTTACCAGAACCGTCACCGCGAACGGCGAAACCCGCGTC
TACACCTACAACTACGACCGGTTCGGCCAACCAACCTCACAAGCAGTCAACGGCAAAACC
TTAAGCACCACGACCTACCGCAACGACGGCACCATCGCGAATGTTCGTTACAAAAACGGC
ACCACCTCAGCGATCACACTCGACGCGAACAACAACGCCGACGCCATCACCTACGCCGGA
TTCGCAGGCGGCGCAACCATCGGCGAAACCAACACCTTCTCACGAGACAACGCGATCCTG
TCGCGCACCTTGCGAGGAACCGACGGAACAGCCACCACCCAAGCCACCTACAACCAAGAC
CACCGGATCATCTCCGCTGTTAACACCGGCACCATAGACCTGACCACCAGATCCAAAACT
GTTGCGTTCGAAGGCCAGGCCGGATCGAACGGCAACCGGACATCACAAACCACCACCACA
ACCAACAACAAACAACAAACCGCAACATTTAGCTACAACGAAGCAAACCAACTCACCTCA
ACCACCCAAGAAAACATTGGCACACCCACCTACGACTCCCACGGCCGCGCCACCAGCATC
```

Εικόνα 5. Παράδειγμα μορφής FASTA αρχείου

1.16.3 GenBank

Η GenBank μορφή αρχείου (Εικόνα 6) αποθηκεύει πληροφορίες για DNA/ πρωτεϊνικές αλληλουχίες. Περιέχει πληροφορίες για τις συντεταγμένες που βρίσκεται το κάθε γονίδιο στο γονιδίωμα, το όνομα του γονιδίου, το προϊόν που παράγει η πρωτεΐνη που κωδικοποιεί, σε ποιο στέλεχος είναι, σε ποιον οργανισμό, την πρωτεϊνική αλληλουχία, κωδικοί της InterPro, κ.α.

```
source          1..89983
                /organism="Actinobacteria bacterium IMCC26207"
                /strain="IMCC26207"
                /mol_type="genomic DNA"
                /country="South Korea:Lake Soyang"
                /isolation_source="freshwater"
                /collected_by="Suhyun Kim"
                /collection_date="14-Apr-2014"
                /db_xref="taxon:1641811"
gene            239..1486
                /locus_tag="IMCC26207_1011"
                /note="IMG reference gene:2606506155"
CDS             239..1486
                /codon_start=1
                /transl_table=11
                /locus_tag="IMCC26207_1011"
                /product="Glycosyl hydrolases family 6"
                /EC_number="3.2.1.4"
                /note="PFAM: Glycosyl hydrolases family 6"
                /db_xref="EnsemblGenomes-Gn:IMCC26207_1011"
                /db_xref="EnsemblGenomes-Tr:KLR62664"
                /db_xref="GOA:A0A0J0V0V4"
                /db_xref="InterPro:IPR016288"
                /db_xref="InterPro:IPR036434"
                /db_xref="UniProtKB/TrEMBL:A0A0J0V0V4"
                /protein_id="KLR62664.1"
                /translation="MASFQLTRSTQTRSTPKTLSALAVALLCAGSGITLGCSSNSDIAG
                AEVAAPAGVESKGSEATSGTYQAGIDQPFNFNWGGNRLWVRWKSQGDCANGIKVTTDGK
                SSEVPVNPEWSSLVLSDSVREVTVDPSGCVLDVDHLSDTTRPFIGMQITDTSSAPAIAP
                KSISTTPTAVWLTDADTAPGAVESALSEAERNDSVPVFVLYLRPQRDCGGFSSNTDSSS
                SSSSNESAAYLGAVQDLATRLGDTPAIVVAEPDASSQDCGDPALIASAVQALKVNPSTY
                VYVDGGHPGWKSVSDQASRLRAAGVEQADGFALNVSNFVHLNEVQIYGDDLSNVLGKDY
                VVDISRSGGTVPAGEWCNPTGARIGTEPTFDTPSINMVAELWVKQPGESDGECNGGPSA
                GTWWQEGAEELASNQN"
sig_peptide     239..346
                /locus_tag="IMCC26207_1011"
                /note="Signal predicted by SignalP 3.0 HMM; IMG reference
                gene:2606506167_SP"
gene            1548..2117
```

Εικόνα 6. Μορφή Genbank αρχείου

1.17 Χρήσιμα εργαλεία Βιοπληροφορικής

Σε αυτό το κεφάλαιο θα γίνει η παρουσίαση των εργαλείων και των βάσεων δεδομένων που χρησιμοποιήθηκαν για την παρούσα εργασία.

1.17.1 InterPro

Η InterPro[1], είναι μία βάση δεδομένων που παρέχει δεδομένα ανάλυσης των λειτουργιών των πρωτεϊνών με την κατηγοριοποίηση των πρωτεϊνών σε οικογένειες καθώς και πρόβλεψη σημαντικών περιοχών (Blum M, 2020).

[1] Σημείωση: Η αρχική βάση που θα χρησιμοποιούνταν ήταν η Pfam, η οποία πλέον κάνει ανακατεύθυνση στην ιστοσελίδα της InterPro.

1.17.2 UniProt

Η UniProt είναι μία βάση δεδομένων που παρέχει πληροφορίες για αλληλουχίες πρωτεϊνών και για τις λειτουργίες τους (Consortium, 2020).

1.17.3 R

Η R είναι μία γλώσσα προγραμματισμού με ευρεία χρήση σε βιολογικά δεδομένα καθώς έχει και πολλά πακέτα με εξειδίκευση σε αυτά (Team, 2021).

Τα πακέτα που χρησιμοποιήθηκαν σε αυτή την εργασία είναι τα εξής:

- Bioseq (Keck, 2020)
- Seqinr (Charif D, 2007)
- Biostrings (DebRoy, 2022)
- Genbank (Lawrence, 2022)
- GenomicFeatures (Carey, 2013)

1.17.4 FastQC

Είναι ένα εργαλείο που εκτελεί ποιοτικό έλεγχο σε δεδομένα αλληλουχιών που έχουν προκύψει από αλληλούχιση. Παρέχει χρήσιμες πληροφορίες και συνοπτικά αποτελέσματα κυρίως με γραφήματα για την ποιότητα των δεδομένων (Andrews, 2010). Βοηθάει στο να παρθούν αποφάσεις για τις διαδικασίες που θα ακολουθήσουν και είναι ένας τρόπος να γίνει μία πρώτη αναγνώριση των δεδομένων.

1.17.5 Trim-Galore

Είναι ένα εργαλείο το οποίο κάνει καθαρισμό των αναγνώσεων, κόβει κομμάτια χαμηλής ποιότητας, αντάπτορες αλλά και άλλες περιοχές που καθορίζει ο χρήστης με ειδικές παραμέτρους που επιλέγει (Felix Krueger, Frankie James, Phil Ewels, Ebrahim Afyounian, & Benjamin Schuster-Boeckler, 2021).

1.17.6 SPAdes

Είναι ένα εργαλείο που χρησιμοποιείται για τη διαδικασία ανακατασκευής του γονιδιώματος (assembly). Αποτελείται από πολλά εσωτερικά βήματα και υποστηρίζει δεδομένα από Illumina και IonTorrent. Είναι κατάλληλο για βακτηριακά γονιδιωμάτα (Nurk, Sergey, et al., 2013).

1.17.7 QUAST

Είναι ένα εργαλείο αξιολόγησης της ανακατασκευής του γονιδιώματος. Λειτουργεί με και χωρίς γονιδίωμα αναφοράς και δίνει αποτελέσματα ως προς το πόσο βέλτιστη είναι η ανακατασκευή του γονιδιώματος. Παράγει αποτελέσματα με γραφήματα αλλά και κάποιες συνοπτικές μετρήσεις σε πίνακα (Mikheenko, 2016). Σε αυτούς τους δείκτες υπάρχουν οι δείκτες N50, N75, L50, L75 κ.α. Για να υπολογιστούν αυτά οι μετυρήσεις, πρέπει να τοποθετηθούν τα συναρμολογήματα (contig) στη σειρά, από το μεγαλύτερο στο μικρότερο και προστίθενται μέχρι να αγγίξουν το 50% του συνολικού μήκους του assembly. Το μήκος του συναρμολογήματος που σταμάτησε δίνει το N50 ενώ ο αριθμός του συναρμολογήματος δίνει το L50.

Ένα παράδειγμα είναι το εξής:

Αριθμός contig	1	2	3	4	5	6	7
Μέγεθος	100	70	60	50	50	40	30

Το 50% του 400 είναι 200, άρα προστίθενται μεγέθη μέχρι να φτάσουν στο 200, δηλαδή 100+70+30=200. Το συναρμολόγημα στο οποίο σταμάτησε έχει μήκος 60 άρα N50 = 60 και είναι το 3^{ο} άρα L50 = 3. Όσο πιο μικρό το L50 και όσο πιο μεγάλο το N50, τόσο καλύτερο καθώς αυτό σημαίνει πως δεν υπάρχουν μεγάλα κενά και το γονιδίωμα που ανακατασκευάστηκε αποτελείται από μεγάλα κομμάτια και άρα είναι πλήρες. Αντίστοιχα προκύπτουν και τα N75, L75 με το 75%.

1.17.8 BUSCO

Είναι ένα εργαλείο που αξιολογεί την ακεραιότητα του συναρμολογήματος (assembly), και το πόσο πλήρες είναι αυτό σε σχέση με μία βάση δεδομένων οργανισμών (Manni, Berkeley, Seppey, Simão, & Zdobnov, 2021).

1.17.9 Minimap2

Είναι ένα εργαλείο για χαρτογράφηση και στοίχιση αλληλουχιών σε ένα γονιδίωμα αναφοράς. Βοηθάει στην εύρεση κενών ή επικαλύψεων και συνεπώς είναι ένας τρόπος αξιολόγησης του συναρμολογήματος assembly (Li, 2018).

1.17.10 Samtools

Είναι ένα εργαλείο που περιέχει πολλά προγράμματα που μπορούν να εκτελεστούν. Βοηθάει στην επεξεργασία των δεδομένων ώστε να έχουν κατάλληλη μορφή για τις διαδικασίες που μπορεί να επέλθουν (Danecek, et al., 2021).

1.17.11 Integrative Genomics Viewer (IGV)

Είναι ένα διαδραστικό εργαλείο που βοηθάει στην οπτικοποίηση δεδομένων γονιδιώματος και στην ανάλυσή τους (Robinson, James T, et al., 2011).

1.17.12 Unicycler

Είναι ένα εργαλείο για ανακατασκευή γονιδιώματος σε βακτηριδιακά γονιδιώματα. Είναι κατάλληλο για δεδομένα από πλατφόρμα Illumina και λειτουργεί ως βελτίωση της κατασκευής ως προς το SPAdes (Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, & Kathryn E. Holt, 2017).

1.17.13 RagTag

Αποτελείται από μία συλλογή εργαλείων για βελτίωση ενός assembly βάσει ενός γονιδιώματος αναφοράς (Alonge, 2021). Περιέχει 4 προγράμματα για βελτίωση τα οποία είναι:

- Διόρθωση τυχόν λαθών που προέκυψαν
- Απαλοιφή των κενών χωρίς να αλλάξει την αλληλουχία
- Συμπλήρωση των κενών που έχουν προκύψει
- Ένωση διαφορετικών κομματιών που μπορεί να προέκυψαν

1.17.14 Prokka

Είναι ένα εργαλείο που κάνει το annotation ενός ανακατασκευασμένου γονιδιώματος ειδικό για προκαρυωτικούς οργανισμούς. Αναγνωρίζει τα χαρακτηριστικά που ενδιαφέρουν τον χρήστη και δίνει χρήσιμες πληροφορίες (Seemann, 2014).

1.17.15 Diamond blast

Είναι ένα εργαλείο στοίχισης μεγάλων αλληλουχιών με βάσεις δεδομένων. Είναι γρήγορο, δεν έχει πολλές απαιτήσεις και δίνει πολλές επιλογές στην έξοδο (Buchfink, Benjamin, Reuter, Klaus, & Drost, Hajk-Georg, 2021).

1.17.16 MEGAN

Σε συνδυασμό με το DIAMOND χρησιμοποιείται για οπτικοποιήσεις δεδομένων και αποτελεσμάτων στοίχισης (Huson, Auch, Qi, & Schuster, 2007).

1.17.17 BLAST

Ένα από τα πιο χρήσιμα εργαλεία της βιοπληροφορικής είναι το blast, το οποίο είναι κατάλληλο για στοίχιση γονιδιωμάτων σε βάσεις δεδομένων.

Υπάρχουν 2 εκδοχές:

- online (Altschul, Gish, Miller, Myers, & Lipman, 1990)
- Γραμμή εντολών (Camacho, et al., 2009)

1.17.18 Clustal Omega

Είναι ένα πρόγραμμα που εκτελεί πολλαπλή στοίχιση αλληλουχιών χρησιμοποιώντας HMM profile-profile τεχνικές για να δημιουργήσει στοιχίσεις μεταξύ 3 και παραπάνω αλληλουχιών (Madeira, 2022).

1.17.19 Microsoft Excel

Είναι ένα πρόγραμμα υπολογιστικών φύλλων που βοηθάει στην οργάνωση και επεξεργασία δεδομένων καθώς και στην παραγωγή γραφημάτων (Corporation, 2018).

1.18 Σκοπός

Βάσει όλων των παραπάνω, η εύρεση τερπενίων σε βακτηριακά γονίδια θα είχε μεγάλο όφελος σε πολλούς τομείς, αλλά και μεγάλο ενδιαφέρον στην εύρεση γονιδίων που ίσως λειτουργούν σε ομάδες. Έτσι, ο σκοπός αυτής της εργασίας είναι η εξόρυξη τερπενικών συνθασών και άλλων πρωτεϊνών σχετικών με τερπενοειδή μέσω της ανάπτυξης μιας μεθοδολογίας, τη χρήση υπολογιστικών μέσων και εργαλείων βιοπληροφορικής, αλλά και με βασική προϋπόθεση την ερμηνεία αποτελεσμάτων και την εξόρυξη χρήσιμων πληροφοριών.

Κεφάλαιο 2ο

2 ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

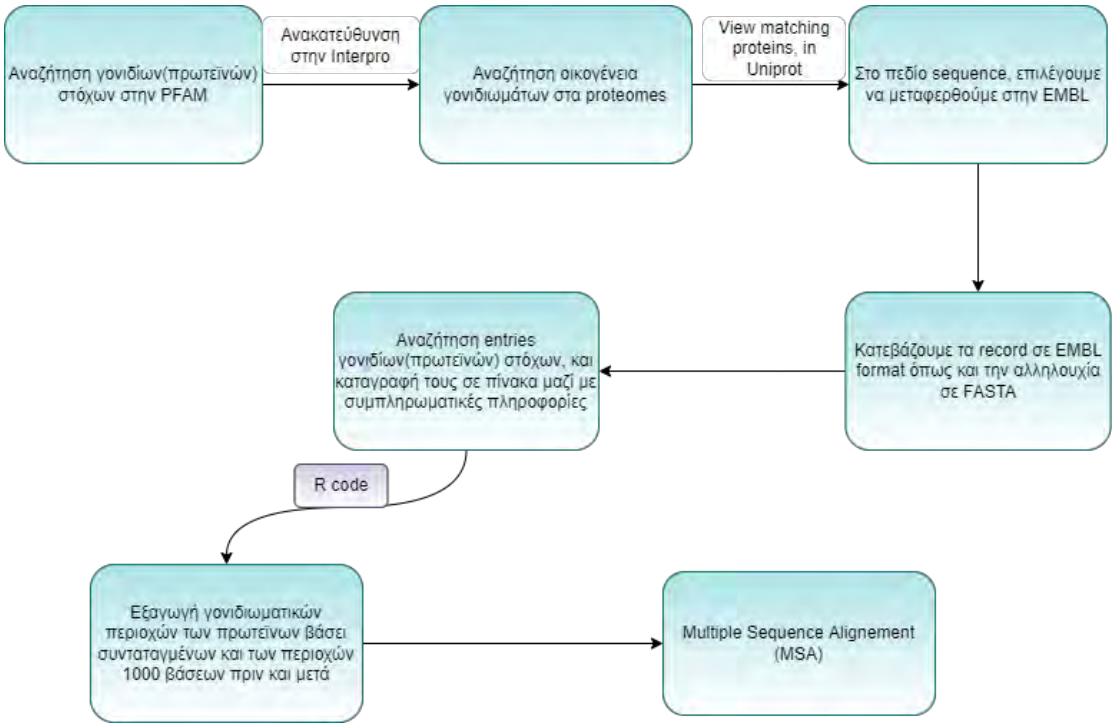
2.1 Περιγραφή Μεθοδολογίας

Η μεθοδολογία αποτελείται από 3 βασικά μέρη:

- Προετοιμασία των δεδομένων της οικογένειας γονιδίων-στόχων, όπου προετοιμάστηκαν τα δεδομένα πρωτεϊνών για να αναζητηθούν στα γονιδιώματα ενδιαφέροντος.
- Προετοιμασία των δεδομένων των γονιδιωμάτων από τους οργανισμούς-στόχους, όπου δεδομένα που προκύπτουν από αλληλούχιση έπρεπε να προσαρμοστούν ώστε να είναι κατάλληλα για την αναζήτηση
- Εύρεση των γονιδίων-στόχων στους οργανισμούς ενδιαφέροντος, όπου εφαρμόστηκε το τελικό βήμα και αναζητήθηκε σε ποια από τα γονιδιώματα-στόχους υπάρχουν οι πρωτεΐνες, σε ποιες θέσεις, με τι ποσοστό επιτυχίας, καθώς και τη συσχέτιση τους με γειτονικά γονίδια.

2.1.1 Προετοιμασία δεδομένων Γονιδίων-Στόχων

Η προετοιμασία των δεδομένων των γονιδίων-στόχων είναι η βασική μεθοδολογία που διαμορφώθηκε όπως φαίνεται συνοπτικά στο παρακάτω διάγραμμα ροής (Εικόνα 7):



Εικόνα 7. Διάγραμμα Ροής για την προετοιμασία των δεδομένων για τα γονίδια-στόχους.

Η βασική μεθοδολογία βασίζεται στην λογική της αναζήτησης μέσω HMM (Hidden Markov Models) και συγκεκριμένα σε profile HMM, όπου γίνεται η αναζήτηση και η εύρεση των αρχικών κωδικών (entry) των γονιδίων-στόχων. Ξεκινώντας την αναζήτηση για τους κωδικούς της οικογένειας πρωτεϊνών-στόχο στην ιστοσελίδα της βάσης δεδομένων PFAM, η οποία πλέον ανακατευθύνει στην ιστοσελίδα της βάσης InterPro (Blum M, 2020). Εκεί είναι εύκολο να βρεθούν κι άλλοι κωδικοί σχετικοί με την οικογένεια-στόχο για περαιτέρω αναζήτηση. Σε κάθε κωδικό,

στον τομέα πρωτεώματα, όπου δίνεται το πλήρες σετ των πρωτεϊνών που εκφράζονται σε έναν οργανισμό, γίνεται αναζήτηση για τον οργανισμό στόχο που έχει τεθεί. Από τους οργανισμούς που θα προκύψουν θα επιλεχθεί ένας για να φανούν οι πρωτεΐνες που ταιριάζουν και θα προβληθούν στη βάση UniProt (Consortium, 2020), από όπου δίνεται η επιλογή να ανακτηθεί τόσο το EMBL αρχείο όσο και η FASTA αλληλουχία που αντιστοιχεί είτε στο πλήρες γονιδίωμα (αν είναι πλήρες), είτε στο κομμάτι που υπάρχει η πρωτεΐνη στόχος που αναζητείται. Το EMBL αρχείο παρέχει τις απαραίτητες πληροφορίες για τις πρωτεϊνικές αλληλουχίες αλλά και για τις συντεταγμένες που βρίσκονται οι αντίστοιχες DNA αλληλουχίες στο γονιδίωμα του εκάστοτε οργανισμού. Αυτή η διαδικασία θα επαναληφθεί για όλους τους κωδικούς που σχετίζονται με τα γονίδια-στόχους. Έτσι, μπορούν να συλλεχθούν τα δεδομένα για τις DNA αλληλουχίες, τις θέσεις που βρίσκονται, δηλαδή τις ακριβείς συντεταγμένες στο γονιδίωμα του συγκεκριμένου οργανισμού, την πρωτεϊνική αλληλουχία όπως και ποιο είναι το προϊόν που δίνουν, δηλαδή το όνομα ή μία περιγραφή της πρωτεΐνης που παράγουν και να οργανωθούν σε υπολογιστικά φύλλα Excel (Corporation, 2018).

2.1.1.1 Κώδικας Εξαγωγής Γονιδίων

Το επόμενο βήμα είναι να εξαχθούν οι DNA αλληλουχίες των πρωτεϊνών αυτών. Αρχικά, συντάχθηκε ένα κομμάτι κώδικα σε γλώσσα προγραμματισμού R (Team, 2021) 7.1.1 *Κώδικας 1*], χρησιμοποιώντας τα πακέτα της R, bioseq (Keck, 2020), seqinr (Charif D, 2007) και Biostrings (DebRoy, 2022), με το οποίο βάσει συντεταγμένων εξάγονται οι DNA αλληλουχίες της κάθε πρωτεΐνης. Για να λειτουργήσει ο κώδικας θα πρέπει να δίνονται:

- οι συντεταγμένες στις οποίες τοποθετείται η κάθε πρωτεΐνη στο γονιδίωμα του οργανισμού,
- το μονοπάτι του FASTA αρχείου του γονιδιώματος
- η επικεφαλίδα του FASTA αρχείου π.χ. ">ENA |CP011489| CP011489.1 Actinobacteria bacterium IMCC26256, complete genome.", καθώς πρέπει να αφαιρεθεί από το αρχείο.

Ο κώδικας διαβάζει το FASTA αρχείο, αφαιρεί την επικεφαλίδα όπως και τους χαρακτήρες καινούριας γραμμής (new line), δηλαδή η αλληλουχία βρίσκεται πλέον σε μία ενιαία γραμμή. Στη συνέχεια, τοποθετεί τον κάθε χαρακτήρα, δηλαδή το κάθε νουκλεοτίδιο σε ένα δικό του διάνυσμα, και αποθηκεύει την αλληλουχία από τις συντεταγμένες που έχουν οριστεί σε ένα FASTA αρχείο. Τελικά, η έξοδος αντιστοιχεί σε ένα FASTA αρχείο με την DNA αλληλουχία της πρωτεΐνης-στόχου. Αν στο EMBL αρχείο δίνονται οι συντεταγμένες στον συμπληρωματικό κλώνο, δηλαδή αναφέρονται ως complement, πρέπει η αλληλουχία αντιστραφεί αφού εξαχθεί με τη χρήση ενός εργαλείου online (Stothard, 2000).

Η διαδικασία επαναλαμβάνεται για όλες τις πρωτεΐνες από όλους τους οργανισμούς που προέκυψαν από την αρχική αναζήτηση και το τελικό αρχείο θα είναι ένα FASTA αρχείο που περιέχει όλες τις αλληλουχίες DNA των πρωτεϊνών-στόχων.

Επιπλέον, εφαρμόζοντας Blastx online (Altschul, Gish, Miller, Myers, & Lipman, 1990) στις αλληλουχίες που προέκυψαν μπορεί να επιβεβαιωθεί πως η πρωτεΐνη που προκύπτει από την κάθε αλληλουχία που έχει εξαχθεί είναι σχετική με την οικογένεια-στόχο.

2.1.2 Εξαγωγή γειτονικών γονιδίων

Το επόμενο βήμα είναι να βρεθούν οι περιοχές ή/και τα γονίδια που βρίσκονται πριν και μετά από τα γονίδια ενδιαφέροντος, καθώς είναι σημαντική η συσχέτιση τους με τα γειτονικά γονίδια, και κατά πόσο σχηματίζονται συστοιχίες γονιδίων που ανήκουν στο ίδιο βιοσυνθετικό μονοπάτι (BGC – Biosynthetic Gene Clusters) ή εντοπίζονται συντηρημένες αλληλουχίες που μπορεί να είναι ρυθμιστικοί παράγοντες. Γι' αυτό τον λόγο, διαμορφώθηκε ο κώδικας που εξηγήθηκε παραπάνω, έτσι ώστε να υπολογίζει τις συντεταγμένες για +/- 1000 βάσεις ώστε να εξάγει και να αποθηκεύει την αλληλουχία του γονιδίου μαζί με τις γειτονικές της περιοχές [*7.1.2 Κώδικας 2*].

Επιπλέον, χρησιμοποιώντας τα πακέτα της R, genbankr (Lawrence, 2022) και GenomicFeatures (Carey, 2013), συντάχθηκε κώδικας που εξάγει όλες τις DNA αλληλουχίες γονιδίων [*7.1.3 Κώδικας 3*]. Αυτό γίνεται δίνοντας ως είσοδο το GenBank αρχείο του κάθε οργανισμού στον οποίο βρέθηκαν τα γονίδια-στόχοι. Στην έξοδο, ο κώδικας δίνει ένα FASTA αρχείο με όλες τις DNA αλληλουχίες των γονιδίων που περιέχει ο οργανισμός και επιπλέον ένα tabular αρχείο με όλα τα γονίδια, τις θέσεις που βρίσκονται, σε ποιο κλώνο, τι μήκος έχουν κλπ.

Εφόσον τα γονίδια-στόχοι είναι γνωστά, αναζητούνται τα γειτονικά τους γονίδια, και στους δύο κλώνους. Είτε με blastx, είτε μέσω του EMBL αρχείου, είναι δυνατόν να αναγνωριστεί η παραγόμενη πρωτεΐνη και με βιβλιογραφική αναζήτηση να επιλέγονται αυτά που πιθανά σχετίζονται με τα γονίδια-στόχους.

2.1.3 Τα δεδομένα που προέκυψαν

Τα τελικά δεδομένα μετά από αυτό το βήμα είναι:

- ένα FASTA αρχείο με αλληλουχίες γονιδίων-στόχων
- ένα αρχείο FASTA με τις αλληλουχίες των γειτονικών γονιδίων που βρέθηκε ότι σχετίζονται με τα γονίδια-στόχους
- υπολογιστικά φύλλα Excel με συγκεντρωτικές πληροφορίες σχετικά με τις αλληλουχίες

2.1.4 Πολλαπλή Στοίχιση για εύρεση ρυθμιστικών παραγόντων

Ένα ακόμη βήμα που πραγματοποιήθηκε σε αυτό το στάδιο είναι μία πολλαπλή στοίχιση (MSA - Multiple Sequence Alignment), μεταξύ των αλληλουχιών των 1000 βάσεων πριν το κάθε γονίδιο-στόχο. Αυτή η στοίχιση είναι δυνητικά αρκετά ενδιαφέρουσα γιατί δείχνει αν υπάρχουν συντηρημένες περιοχές πριν τα γονίδια και αν προκύπτουν μοτίβα που ίσως λειτουργούν ως ρυθμιστικοί παράγοντες που ελέγχουν την έκφραση των γονιδίων που τα ακολουθούν, δηλαδή συμβάλουν είτε στην αποσιώπηση των γονιδίων είτε στην ενίσχυση της έκφρασης.

Με τη χρήση του Clustal Omega (Madeira, 2022) εφαρμόστηκε MSA στις ακολουθίες των γονιδίων από όπου προκύπτει ένα φυλογενετικό δέντρο στο οποίο διαχωρίστηκαν/ομαδοποιήθηκαν οι αλληλουχίες και εφαρμόστηκε πάλι πολλαπλή στοίχιση μεταξύ των ομαδοποιημένων αλληλουχιών στις 1000 βάσεις πριν, ώστε να προβληθούν στο Jalview (Waterhouse, 2009) και εκεί παρατηρείται κατά πόσο υπάρχουν συντηρημένες περιοχές οι οποίες μπορεί να λειτουργούν ως ρυθμιστικά στοιχεία για την ενίσχυση ή την αποσιώπηση της έκφρασης των γονιδίων.

2.1.5 Προετοιμασία δεδομένων των οργανισμών-στόχων

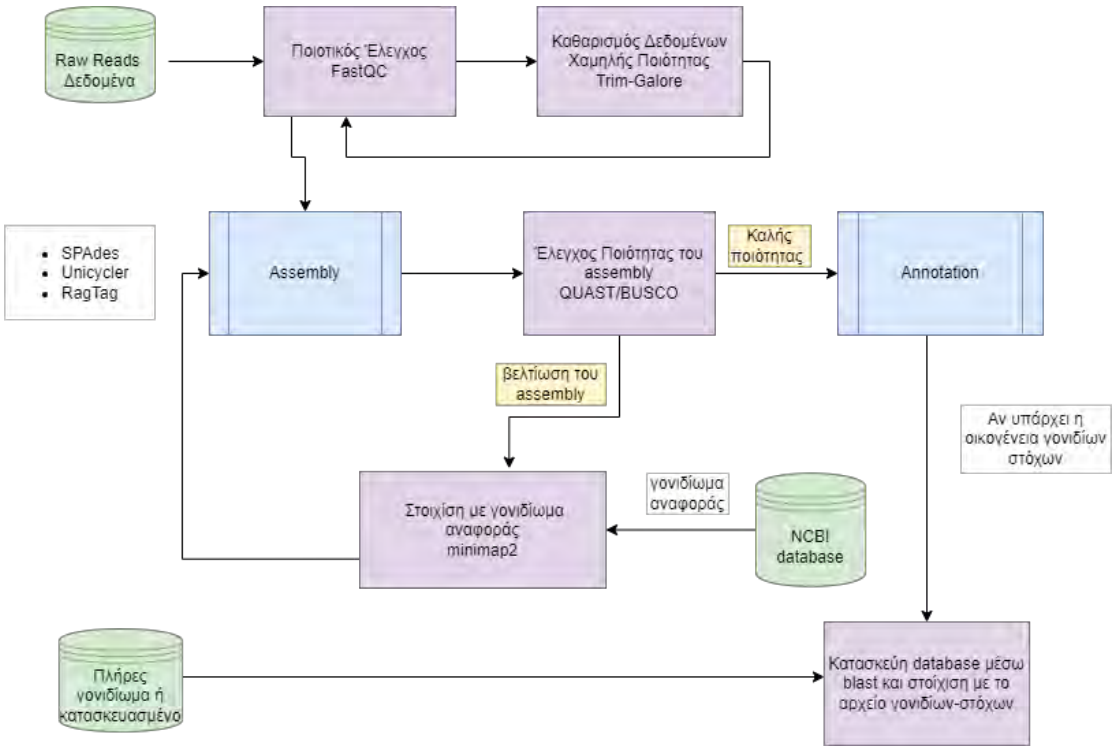
Σε αυτό το κεφάλαιο θα αναλυθεί το δεύτερο βασικό μέρος της μεθοδολογίας που είναι η προετοιμασία των δεδομένων τον οργανισμών-στόχων.

Αν τα δεδομένα των γονιδιωμάτων των οργανισμών που έχουν τεθεί ως στόχοι είναι σε FASTA μορφή, είτε πλήρες γονιδιώματα. είτε σε μορφή κομματιών (contigs), κατασκευασμένου γονιδιώματος, μπορεί να ξεκινήσει απευθείας η αναζήτηση όπως περιγράφεται στο επόμενο κεφάλαιο *2.1.6 Εύρεση γονιδίων-στόχων στα γονιδιώματα*.

Αν όμως τα διαθέσιμα αρχεία είναι FASTQ με δεδομένα που έχουν προκύψει απευθείας από αλληλούχιση (raw reads), αυτά πρέπει να υποστούν μία επεξεργασία ώστε να έρθουν σε μορφή κατάλληλη για αναζήτηση. Ο κώδικας της μεθοδολογίας αυτής περιγράφεται πλήρως και αναλυτικά στο *7.2 Κώδικας στη γραμμή εντολών*. Η μεθοδολογία που ακολουθείται είναι μία γονιδιωματική ανάλυση και αποτελείται από δύο βασικά μέρη:

- Assembly – η διαδικασία ανακατασκευής της πλήρης αλληλούχιας των κομματιών που προκύπτουν από την αλληλούχιση (Kong, et al., 2019).
- Annotation – η διαδικασία ανίχνευσης των γονιδίων, σε ποιες περιοχές βρίσκονται στην αλληλουχία, όπως και ο καθορισμός της δομής και της λειτουργίας αυτών των γονιδίων (Kong, et al., 2019)

Τα μέρη αυτά αποτελούνται από μία σειρά βημάτων που παρουσιάζεται συνοπτικά στο παρακάτω διάγραμμα ροής (Εικόνα 8):



Εικόνα 8. Διάγραμμα ροής για την προετοιμασία των δεδομένων των οργανισμών-στόχων

2.1.5.1 Ανάλυση και προετοιμασία δεδομένων

Αρχικά, εφαρμόζεται ένας ποιοτικός έλεγχος στα δεδομένα των FASTQ αρχείων (reads) με το εργαλείο FastQC (Andrews, 2010), για να αξιολογηθεί η ποιότητα τους, αν υπάρχουν αντάπτορες ή πολλά κενά ή μολύνσεις κλπ. Το εργαλείο βοηθάει ώστε να ανιχνευθούν προβλήματα που έχουν προκύψει είτε στην αλληλούχιση είτε στην κατασκευή της βιβλιοθήκης. Η εικόνα που θα προκύψει βοηθάει στην εκτίμηση της ποιότητας και με βάση αυτή, κόβονται και καθαρίζονται τα κομμάτια από τυχόν αντάπτορες που βρέθηκαν, ή πολλά κενά (Ns) και θα αφαιρεθούν επίσης, χαμηλής ποιότητας ή μικρά κομμάτια που υπάρχουν, με τη χρήση του εργαλείου trim-galore (Felix Krueger, Frankie James, Phil Ewels, Ebrahim Afyounian, & Benjamin

Schuster-Boeckler, 2021). Επαναλαμβάνεται ο ποιοτικός έλεγχος με το εργαλείο FastQC, για να επιβεβαιωθεί πως πλέον τα δεδομένα είναι καλής ποιότητας. Αυτή είναι η ανάλυση και προετοιμασία των δεδομένων για την ανακατασκευή του γονιδιώματος.

2.1.5.2 Ανακατασκευή του γονιδιώματος και Αξιολόγηση

Η διαδικασία που ακολουθείται σε αυτό το βήμα είναι εκ νέου κατασκευή, de novo assembly του γονιδιώματος καθώς δεν υπάρχει κάποιο γονιδίωμα αναφοράς. Το εργαλείο που χρησιμοποιείται είναι το SPAdes (Nurk, Sergey, et al., 2013), το οποίο είναι κατάλληλο για βακτήρια, δέχεται δεδομένα από αλληλούχιση με πλατφόρμα Illumina ή IonTorrent και υποστηρίζει δεδομένα paired-end, mate-pairs και unpaired.

Για την αξιολόγηση των αρχείων που προέκυψαν από την ανακατασκευή του γονιδιώματος θα χρησιμοποιηθούν 2 εργαλεία, το QUAST (Mikheenko, 2016), το οποίο παρέχει διάφορα μέτρα για να ελεγχθεί η ποιότητα, όπως το N50 και L50, και το BUSCO (Manni, Berkeley, Seppey, Simão, & Zdobnov, 2021), το οποίο δίνει πληροφορίες για την πληρότητα της ανακατασκευής σε σχέση με μία βάση δεδομένων που επιλέγει αυτόματα. Το επόμενο βήμα είναι η στοίχιση του γονιδιώματος που κατασκευάστηκε με ένα γονιδίωμα αναφοράς, που είναι πιο κοντά σε αυτό, για να παρατηρηθεί πόσα κενά ή πόσες επικαλύψεις υπάρχουν. Η εύρεση του πιο κοντινού γονιδιώματος αναφοράς, γίνεται με το BLAST online (Altschul, Gish, Miller, Myers, & Lipman, 1990) για κάθε FASTA αρχείο που περιέχει τα κομμάτια του γονιδιώματος που κατασκευάστηκε. Μόλις το βρεθεί το γονιδίωμα, εφαρμόζεται το εργαλείο Minimap2 (Li, 2018) για τη στοίχιση του FASTA αρχείου απέναντι στο αντίστοιχο γονιδίωμα αναφοράς του. Για την οπτικοποίηση των αποτελεσμάτων, ταξινομούνται το αποτέλεσμα του minimap2 και κατασκευάζεται το Index με το εργαλείο samtools

(Danecek, et al., 2021) για φορτωθούν ως είσοδο σε ένα πρόγραμμα, το IGV (Robinson, James T, et al., 2011). Αν από τις μεθόδους αξιολόγησης προκύψει ότι η ποιότητα του γονιδιώματος που κατασκευάστηκε δεν είναι καλή, αν υπάρχουν επικαλύψεις ή πολλά κενά, θα πρέπει αυτό να βελτιωθεί. Αρχικά, χρησιμοποιείται ένα ακόμη εργαλείο για την κατασκευή του, το unicycler (Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, & Kathryn E. Holt, 2017), το οποίο είναι κατάλληλο για βακτηριδιακά γονιδιώματα και ακολουθεί μία σειρά από βήματα χρησιμοποιώντας το SPAdes οπότε πιθανώς θα ανακτηθεί μία βελτιστοποιημένη εικόνα. Επαναλαμβάνεται η αξιολόγηση και τα αποτελέσματα συγκρίνονται σε σχέση με αυτά του SPAdes, χρησιμοποιώντας το QUAST και το BUSCO αλλά και ένα ακόμη εργαλείο, το BANDAGE (Wick, 2015), με το οποίο μπορούν να οπτικοποιηθούν τα κομμάτια που προέκυψαν, και να φανούν οι συνδέσεις μεταξύ τους. Στη συνέχεια, μπορεί να βελτιωθεί το ήδη υπάρχον γονιδίωμα που κατασκευάστηκε χρησιμοποιώντας το RagTag (Alonge, 2021), που είναι κατάλληλο για διόρθωση και βελτιστοποίηση των κατασκευασμένων γονιδιωμάτων καθώς δεν αφήνει μεγάλα κενά και διορθώνει πιθανές αποτυχίες που υπήρξαν στην προηγούμενη προσπάθεια. Συγκεκριμένα οι εντολές correct, όπου γίνεται η διόρθωση του γονιδιώματος αφού εντοπίζονται τυχόν αποτυχίες, σύμφωνα με ένα γονιδίωμα αναφοράς, και scaffold, όπου γίνεται αντιληπτό που υπάρχουν τα κενά ώστε να ενωθούν και να δημιουργηθούν μεγαλύτερα κομμάτια. Κατόπιν επαναλαμβάνεται το βήμα της αξιολόγησης ώστε να επιβεβαιωθεί πως έχει κατασκευαστεί καλής ποιότητας γονιδίωμα.

2.1.5.3 *Ανίχνευση των γονιδίων*

Το επόμενο μέρος, είναι το Annotation. Στόχος είναι η ανίχνευση γονιδίων που υπάρχουν στο γονιδίωμα που κατασκευάστηκε και πιο συγκεκριμένα, να υπάρξει μία πρώτη εικόνα για το κατά πόσο η οικογένεια πρωτεϊνών-στόχων υπάρχει στο γονιδίωμα του οργανισμού. Χρησιμοποιείται το εργαλείο PROKKA (Seemann, 2014), που είναι κατάλληλο για προκαρυωτικούς οργανισμούς. Έτσι, στα αρχεία που προκύπτουν παρατηρείται αν υπάρχει η οικογένεια γονιδίων που μας ενδιαφέρει.

2.1.5.4 *Εύρεση γειτονικών οργανισμών για κάθε κατασκευασμένο γονιδίωμα*

Ένα επιπλέον βήμα είναι η εύρεση της συσχέτισης του γονιδιώματος που κατασκευάστηκε με άλλους οργανισμούς. Για να γίνει αυτό, το γονιδίωμα στοιχίζεται με μία βάση δεδομένων όπως είναι η nr, με το εργαλείο DIAMOND (Buchfink, Benjamin, Reuter, Klaus, & Drost, Hajk-Georg, 2021), και συγκεκριμένα με την εντολή blastx. Για την οπτικοποίηση αυτών των αποτελεσμάτων χρησιμοποιείται το MEGAN (Huson, Auch, Qi, & Schuster, 2007), το οποίο είναι κατάλληλο για την ανάλυση μικροβιακών δεδομένων και από αυτό θα προκύψουν τα φυλογενετικά δέντρα από τα οποία φαίνεται το κυρίαρχο είδος του γονιδιώματος που κατασκευάστηκε.

2.1.6 Εύρεση γονιδίων-στόχων στα γονιδιώματα

Εφόσον έχουν ολοκληρωθεί τα παραπάνω βήματα, πλέον υπάρχουν όλα τα απαραίτητα δεδομένα για να απαντηθεί το ερώτημα «ποια από τα γονίδια-στόχους που έχουν συγκεντρωθεί σε FASTA αρχείο υπάρχουν στα γονιδιώματα των οργανισμών-στόχων που είτε κατασκευάστηκε το γονιδίωμά, είτε υπήρχε διαθέσιμο σε κάποια βάση δεδομένων, με τη μορφή κομματιών ή πλήρες.

Για κάθε γονιδίωμα κατασκευάζεται μία βάση με χρήση του εργαλείου NCBI BLAST στη γραμμή εντολών (Camacho, et al., 2009), χρησιμοποιώντας την εντολή makeblastdb και στη συνέχεια στοιχίζεται με το FASTA αρχείο που περιέχει τόσο τα κύρια γονίδια-στόχους όσο και τα γειτονικά τους, με την εντολή blastx. Σε αυτό το βήμα χρησιμοποιήθηκαν οι καθορισμένες από το πρόγραμμα παράμετροι.

Από αυτή τη διαδικασία προκύπτει ένα TEXT αρχείο που δείχνει τη στοίχιση των γονιδίων-στόχων με το γονιδίωμα, το ποσοστό επιτυχίας, πόσα κενά υπάρχουν, σε πόσα σημεία έχει αποτύχει η στοίχιση, σε ποιες συντεταγμένες του γονιδιώματος έγινε η στοίχιση κ.α., αλλά και ένα TABULAR αρχείο που περιέχει συγκεντρωτικές πληροφορίες σχετικά με τα στατιστικά της στοίχισης. Συγκεντρώνοντας και οργώνοντας αυτές τις πληροφορίες σε φύλλα Excel (Corporation, 2018), κατασκευάζονται πίνακες που δείχνουν ποια γονίδια-στόχοι υπάρχουν στα γονιδιώματα ενδιαφέροντος.

2.2 Εφαρμογή Μεθοδολογίας

Σε αυτή την περίπτωση εφαρμόστηκε η παραπάνω μεθοδολογία στην οικογένεια των τερπενικών συνθασών που θεωρούνται ως τα γονίδια-στόχοι, και ως οργανισμούς-στόχους θεωρούνται τα ακτινοβακτήρια.

2.2.1 Προετοιμασία δεδομένων τερπενίων

Το πρώτο βήμα ήταν η εφαρμογή της μεθοδολογίας που περιεγράφηκε στο κεφάλαιο 2.1.1 *Προετοιμασία δεδομένων Γονιδίων-Στόχων*, με μία πρώτη αναζήτηση για την οικογένεια YgbB, που είναι γονίδια σχετικά με τη βιοσύνθεση των τερπενιών. Μέσω της InterPro (Blum M, 2020) και της Pfam ταυτοποιήθηκαν και άλλοι κωδικοί (entries) που σχετίζονται με την παραγωγή τερπενίων. Επιλέχθηκαν οι κωδικοί (entries) που αντιστοιχούν σε πρωτεΐνες σχετικές με τερπενικές συνθάσες ή με τη βιοσύνθεσή τους και οργανώθηκαν συγκεντρωτικά σε έναν πίνακα (Πίνακας 1):

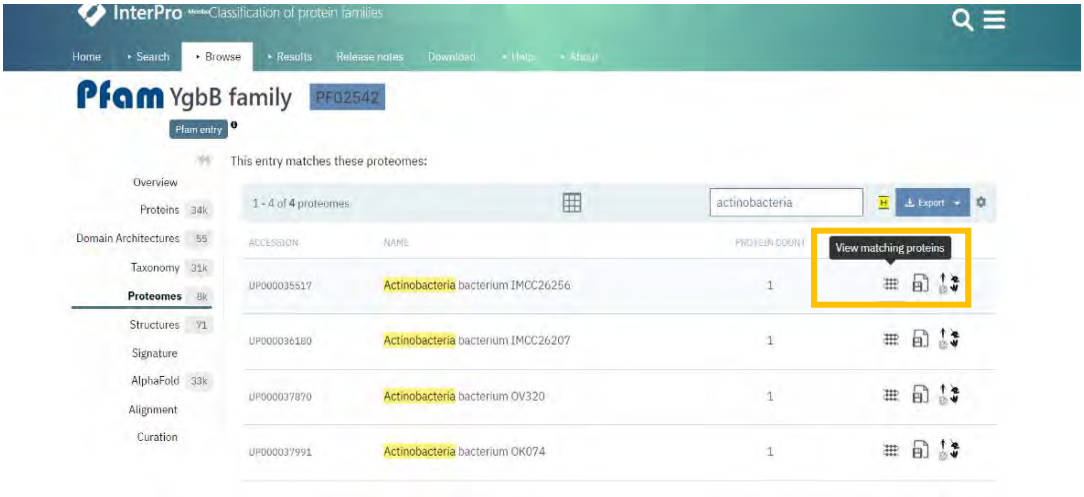
Terpene synthase	**Entry**
Isoprenoid synthase domain superfamily	IPR008949
Terpene cyclase-like 2	IPR034686
Terpene synthase family, metal binding domain	PF03936
Terpene synthase, N-terminal domain superfamily	IPR036965
Terpene synthase family 2, C-terminal metal binding	PF19086
Terpene synthase, N-terminal domain	PF01397
Terpene synthase, conserved site	IPR002365
Squalene hopene cyclase	IPR006400
Terpenoid cyclases/protein prenyltransferase alpha-alpha toroid	IPR008930
Terpene synthase, N-terminal domain	PF01397
Terpene synthase family 2, C-terminal metal binding	PF19086
Polyprenyl synthetase	PF00348
YgbB family	PF02542

Πίνακας 1. Πίνακας με ονόματα και κωδικούς για πρωτεΐνες σχετικές με τερπενοειδή

Σε κάθε κωδικό, στον τομέα πρωτεώματα, αναζητήθηκε ο οργανισμός ενδιαφέροντος, τα ακτινοβακτήρια.

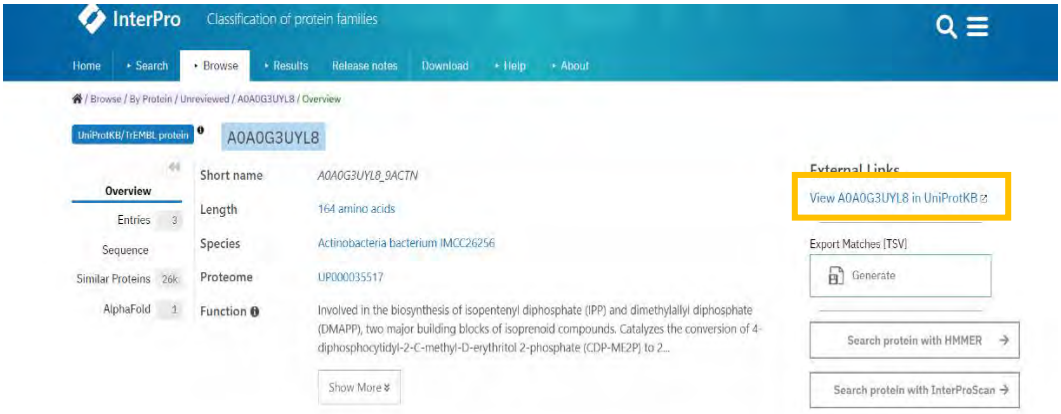
Προέκυψαν δεδομένα για 4 ακτινοβακτήρια (Εικόνα 9):

1. Actinobacteria bacterium IMCC26256
2. Actinobacteria bacterium IMCC26207
3. Actinobacteria bacterium OV320
4. Actinobacteria bacterium OK074



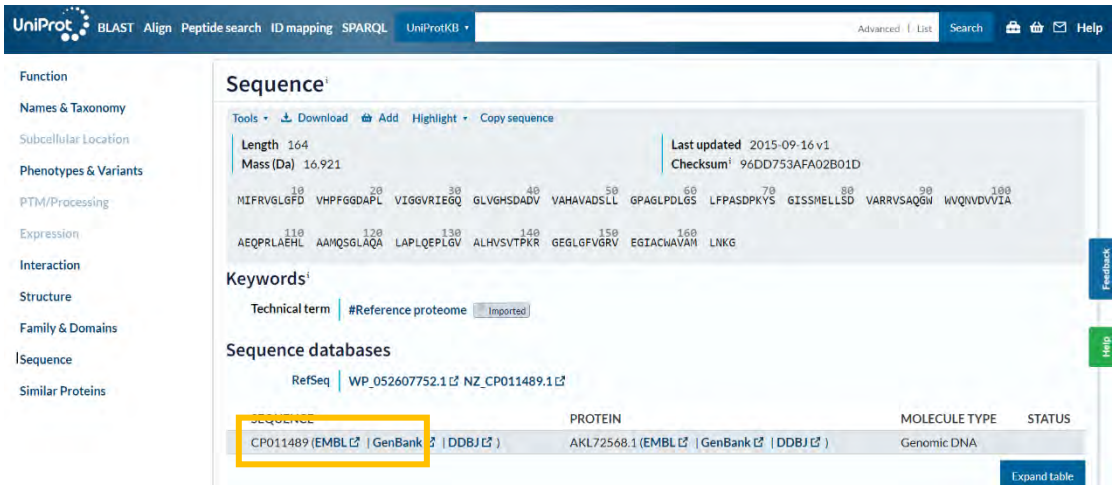
Εικόνα 9. Αναζήτηση του όρου actinobacteria στον κωδικό της οικογένειας YgbB στην InterPro

Επιλέχθηκε ένας από τους οργανισμούς που προέκυψαν για να βρεθούν οι πρωτεΐνες που ταιριάζουν στον χαρακτηριστικό κωδικό που δόθηκε, στην βάση δεδομένων UniProt (Εικόνα 10) (Consortium, 2020):



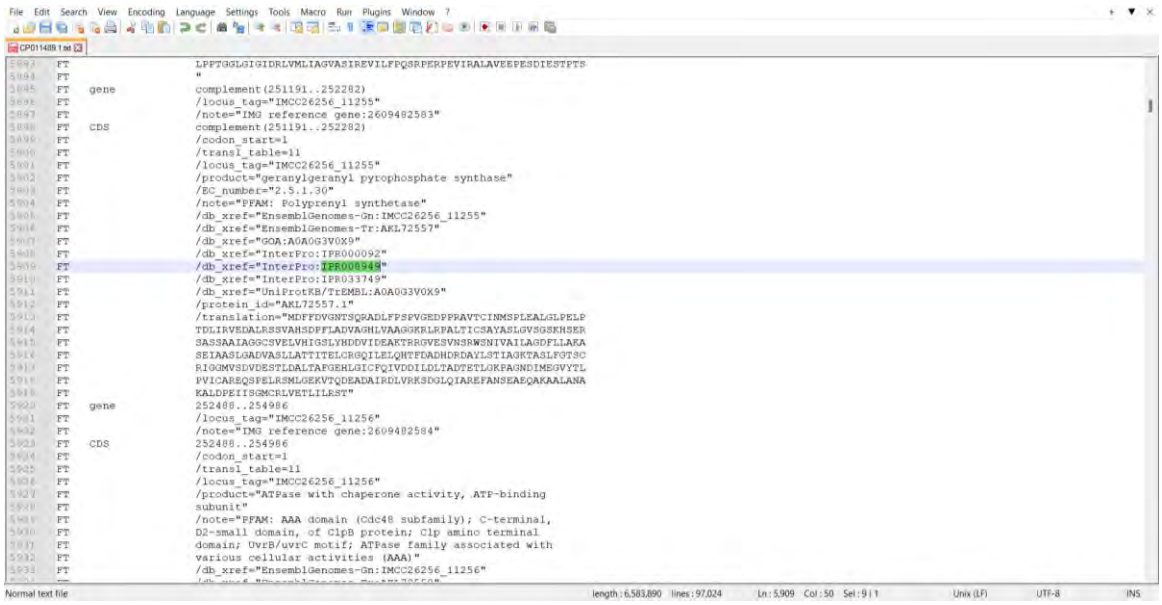
Εικόνα 10. Πρωτεΐνη που ταιριάζει στην αναζήτηση του κωδικού και ανακατεύθυνση στη βάση δεδομένων της Uniprot (πορτοκαλί πλαίσιο)

Από εκεί ανακτήθηκε το EMBL και το FASTA αρχείο (Εικόνα 11) όπως και τα απαραίτητα δεδομένα. Αυτή η διαδικασία επαναλήφθηκε για όλους τους κωδικούς σχετικούς με τερπενικές συνθάσες (Πίνακας 1).



Εικόνα 11. Απόκτηση των αρχείων στις μορφές EMBL,FASTA και Genbank

Η μορφή του αρχείου EMBL (Εικόνα 12):

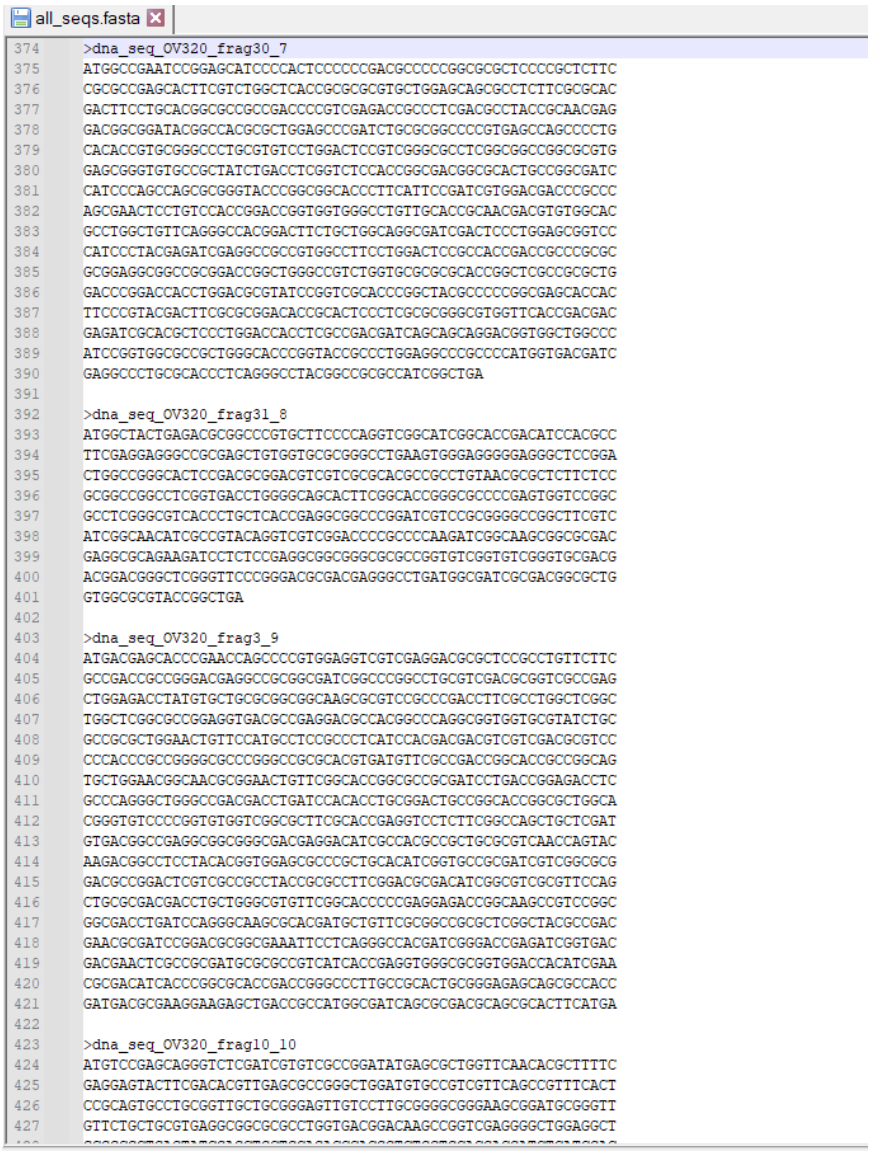


Εικόνα 12. Παράδειγμα αναζήτησης του κωδικού IPR008949 στο EMBL αρχείο και ανάκτηση πληροφοριών

Βάσει αυτού του αρχείου συλλέχθηκαν και οργανώθηκαν τα δεδομένα για τις DNA αλληλουχίες, τις θέσεις που βρίσκονται και ποια πρωτεΐνη παράγουν. Όλες αυτές οι πληροφορίες υπάρχουν σε ένα υπολογιστικό φύλλο EXCEL όπου τους δόθηκε ένα κωδικό-όνομα βασισμένο σε ποιο ακτινοβακτήριο βρισκόταν, σε ποιο κομμάτι του γονιδιώματος του οργανισμού και έναν αριθμό για να δείχνει τη σειρά του. Ένα παράδειγμα είναι το εξής: dna_seq_OV320_frag26_1, δηλαδή η συγκεκριμένη αλληλουχία ανήκει σε τερπενική συνθάση του ακτινοβακτηριδίου OV320, βρίσκεται το κομμάτι 26 και την χαρακτηρίσαμε με τον αριθμό 1. Η σειρά των αριθμών είναι τυχαία και προστέθηκε σε περίπτωση που υπάρξουν πάνω από μία πρωτεΐνη στο ίδιο κομμάτι γονιδιώματος για να μην δημιουργηθεί σύγχυση.

Στη συνέχεια εξάχθηκαν οι DNA αλληλουχίες των πρωτεϊνών αυτών με τον κώδικα στην R, όπως περιεγράφηκε στο κεφάλαιο *2.1.1.1 Κώδικας Εξαγωγής Γονιδίων*, και επαναλήφθηκε η διαδικασία και για τα 4 ακτινοβακτήρια. Προέκυψε

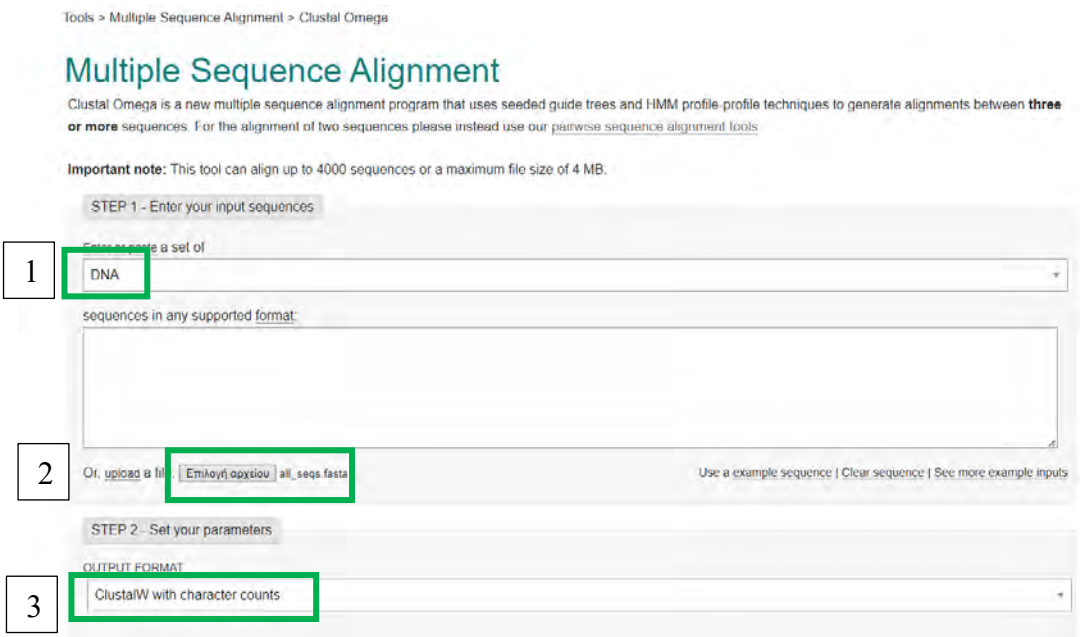
λοιπόν, ένα FASTA αρχείο με 44 τερπενικές συνθάσες (Εικόνα 13) και ένα αρχείο Excel με πληροφορίες σχετικά με το γονίδιο που βρίσκονται, από ποιο entry προέρχονται, ποια πρωτεΐνη παράγουν, σε ποιες θέσεις βρίσκονται κλπ. Εφαρμόστηκε ένα τελικό Blastx για να επιβεβαιωθεί πως σχετίζονται οι αλληλουχίες με την οικογένεια των τερπενίων. Επιπλέον, όπως περιεγράφηκε στο κεφάλαιο *2.1.2 Εξαγωγή γειτονικών γονιδίων*, μέσω του GenBank αρχείου εξάχθηκαν και τα γονίδια που υπάρχουν στο κάθε ακτινοβακτήριο και βρέθηκε ποια από αυτά είναι γειτονικά και ποια συσχετίζονται με τις τερπενικές συνθάσες.



Εικόνα 13. Αρχείο με αλληλουχίες τερπενικών συνθασών

Και τέλος, εξάχθηκαν και οι αλληλουχίες +/- 1000 βάσεων για να είναι δυνατή η ανάλυση πολλαπλής στοίχισης όπως περιγράφεται στο κεφάλαιο *2.1.4 Πολλαπλή Στοίχιση για εύρεση ρυθμιστικών παραγόντων* με τη χρήση του Clustal omega (Madeira, 2022) (Εικόνα 14) για εύρεση συντηρημένων περιοχών που μπορεί να είναι ρυθμιστικές αλληλουχίες. Εκεί, επιλέχθηκε:

1. να στοιχηθεί ένα σετ από DNA αλληλουχίες
2. εισάχθηκε το FASTA αρχείο με τις αλληλουχίες τερπενικών συνθασών και
3. στις παραμέτρους επιλέχθηκε για το αρχείο που θα προκύψει στην έξοδο, Clustal W with character counts για να φανεί και η απόκλιση στις θέσεις



Εικόνα 14. Παράδειγμα Multiple Sequence Alignment

Από εκεί προέκυψαν αρχεία με την στοίχιση τα οποία προβλήθηκαν στο Jalview για καλύτερη οπτικοποίηση και φυλογενετικό δέντρο που δίνει χρήσιμες πληροφορίες.

2.2.2 Προετοιμασία Δεδομένων Ακτινοβακτηριδίων

Όπως έχει ήδη αναφερθεί στο κεφάλαιο *2.1.5 Προετοιμασία δεδομένων των οργανισμών-στόχων*, αν τα γονιδιώματα που έχουν δοθεί είναι σε FASTA μορφή, είτε ως πλήρες γονιδίωμα είτε σε μορφή συναρμολογημάτων, τότε συνεχίζεται η εφαρμογή της μεθοδολογίας στο επόμενο βήμα που είναι η αναζήτηση των γονιδίων στόχων.

Αν όμως τα δεδομένα που έχουν προκύψει από την αλληλούχιση είναι σε FASTQ μορφή τότε πρέπει να εφαρμοστεί η μεθοδολογία του κεφαλαίου *2.1.5 Προετοιμασία δεδομένων των οργανισμών-στόχων.*

Στο εργαστήριο όπου πραγματοποιήθηκε η παρούσα πτυχιακή εργασία, υπάρχουν κάποια στελέχη ακτινοβακτηριδίων που προέκυψαν από την αλληλούχιση, σε FASTQ μορφή, επομένως έπρεπε να ακολουθηθεί η αντίστοιχη μεθοδολογία για να βρεθούν αυτά τα στελέχη σε μορφή κατάλληλη για αναζήτηση. Αρχικά έπρεπε να γίνει μία ανάλυση των δεδομένων της αλληλούχισης και να επεξεργαστούν σύμφωνα με τη μεθοδολογία του *2.1.5.1 Ανάλυση και προετοιμασία δεδομένων*. Στα fastq αρχεία εφαρμόστηκε ένας ποιοτικός έλεγχος με το FastQC (Andrews, 2010). Οι εικόνες που προέκυψαν στην αρχή δείχνουν πως υπάρχουν κάποια δεδομένα που είναι κακής ποιότητας, και στους δύο κλώνους ειδικά προς το τέλος τις κάθε ακολουθίας, επομένως, οι αλληλουχίες κόπηκαν και καθαρίστηκαν χρησιμοποιώντας κατάλληλες παραμέτρους. Αφαιρέθηκαν τα κομμάτια που είχαν χαμηλή ποιότητα($\leq$30), αλλά και κομμάτια που είχαν μικρό μήκος ($\leq$40). Επιπλέον, χρειάστηκε ένα επιπλέον καθάρισμα στο τέλους του 5’ άκρου του αντίστροφου κλώνου καθώς εκεί η ποιότητα ήταν πολύ πιο χαμηλή. Εφαρμόζοντας και πάλι το FastQC, φάνηκε πως τα δεδομένα είναι πλέον καλής ποιότητας επομένως, και μπορούν να χρησιμοποιηθούν για την κατασκευή του γονιδιώματος όπως περιγράφεται στο *2.1.5.2 Ανακατασκευή του γονιδιώματος και Αξιολόγηση*. Στη συνέχεια της διαδικασίας εκτελέστηκε ο έλεγχος και η αξιολόγηση με

το QUAST (Mikheenko, 2016) όπως και με το BUSCO (Manni, Berkeley, Seppey, Simão, & Zdobnov, 2021). Επιπλέον, πραγματοποιήθηκε στοίχιση με ένα γονίδωμα αναφοράς για κάθε ένα από τα δύο στελέχη, το οποίο βρέθηκε με αναζήτηση μέσω Blast (Altschul, Gish, Miller, Myers, & Lipman, 1990), και τα αποτελέσματα αφού προσαρμόστηκαν στην κατάλληλη μορφή, με το SAMTOOLS (Danecek, et al., 2021), φορτώθηκαν στο IGV (Robinson, James T, et al., 2011), και παρατηρήθηκε πως δεν υπάρχουν πολλά κενά, αλλά υπάρχουν μερικές επικαλύψεις, γι' αυτό και αποφασίστηκε να γίνει άλλη μία δοκιμή για κατασκευή γονιδιώματος με άλλο εργαλείο αυτή τη φορά, το Unicycler (Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, & Kathryn E. Holt, 2017), το οποίο με τις κατάλληλες παραμέτρους, έδωσε λίγο καλύτερα αποτελέσματα στο QUAST.

Από αυτά τα αποτελέσματα έγινε αντιληπτό πως χρειάζεται να βελτιωθεί το γονιδίωμα που κατασκευάστηκε. Σύμφωνα με τη μεθοδολογία εφαρμόστηκε το εργαλείο RagTag (Alonge, 2021), στα γονιδιώματα που προέκυψαν από το unicycler, και έγινε μία δοκιμή στο γονιδίωμα του ακτινοβακτηριδίου 1 που κατασκευάστηκε με το SPAdes, καθώς έδωσε λιγότερα κομμάτια. Οι εντολές που επιλέχθηκαν να χρησιμοποιηθούν ήταν correct και scaffold, αλλά έγινε και μία δοκιμή να χρησιμοποιηθεί και η εντολή patch, με την οποία όμως το αποτέλεσμα έβγαζε ένα κομμάτι, που θα ήταν ιδανικό, αλλά άλλαζε αρκετά την αυθεντική αλληλουχία, επομένως αυτό το αποτέλεσμα απορρίφθηκε.

Τα αποτελέσματα ήταν ικανοποιητικά και επιλέχθηκε το γονιδίωμα που προέκυψε από το Unicycler+RagTag καθώς έδωσε λιγότερα κομμάτια.

Αφού ολοκληρώθηκε αυτό το βήμα, ξεκίνησε η εκτέλεση της μεθοδολογίας του κεφαλαίου *2.1.5.3 Ανίχνευση των γονιδίων* για να παρατηρηθεί αν υπάρχουν οι

πρωτεΐνες-στόχοι που σε αυτή την περίπτωση ήταν οι τερπενικές συνθάσες και πρωτεΐνες σχετικές με την παραγωγή και βιοσύνθεσή τους. Με τη χρήση του εργαλείου PROKKA (Seemann, 2014), φάνηκε στα αρχεία που προέκυψαν πως υπάρχουν τερπενικές συνθάσες που μπορεί να σχηματίζουν ομάδες και να συνεκφράζονται αλλά υπήρξαν και πολλές υποθετικές πρωτεΐνες που ίσως ανήκουν στην ίδια κατηγορία.

Το τελευταίο βήμα για αυτή την ανάλυση ήταν να στοιχηθούν τα αρχικά δεδομένα της αλληλούχισης με μία βάση δεδομένων όπως η nr για να παρατηρηθούν συγγενείς οργανισμοί που ίσως είναι πλούσιοι σε τερπενοειδή. Τα αποτελέσματα της στοίχισης και των φυλογενετικών δέντρων έδωσαν τους συγγενείς οργανισμούς.

2.2.3 Εύρεση τερπενοειδών σε γονιδιώματα ακτινοβακτηριδίων

Στο τέλος, εφαρμόστηκε η μεθοδολογία που αναλύθηκε στο κεφάλαιο *2.1.6 Εύρεση γονιδίων-στόχων στα γονιδιώματα*, με βάση το FASTA αρχείο με τις αλληλουχίες τερπενικών συνθασών που δημιουργήθηκε στο κεφάλαιο *2.2.1 Προετοιμασία δεδομένων τερπενίων* στα γονιδιώματα που κατασκευάστηκαν στο κεφάλαιο *2.2.2 Προετοιμασία Δεδομένων Ακτινοβακτηριδίων* και σε γονιδιώματα που υπάρχουν διαθέσιμα σε δημόσιες βάσεις δεδομένων όπως η ENA (ENA - European Nucleotide Archive, n.d.). Ελέγχθηκαν γονιδιώματα που είναι κοντινά στα 2 που κατασκευάστηκαν.

Κατασκευάστηκε βάση με τη χρήση του BLAST στη γραμμή εντολών (Camacho, et al., 2009) για κάθε γονιδίωμα είτε κατασκευασμένο είτε πλήρες, και πραγματοποιήθηκε η στοίχιση με τα αρχεία τερπενοειδών.

Έτσι, βρέθηκε σε ποια σημεία του κάθε γονιδιώματος υπάρχουν τερπενικές συνθάσες, ποιες είναι αυτές, τι προϊόν παράγουν, σε τι ποσοστό επιτυχίας έγινε η

στοίχιση και άλλες πληροφορίες που παρουσιάζονται αναλυτικά στο κεφάλαιο *ΑΠΟΤΕΛΕΣΜΑΤΑ*.

Επιπλέον, κατασκευάστηκε ένας πίνακας όπου στις στήλες έχει τα ακτινοβακτήρια και στις γραμμές τις πρωτεΐνες με τους κωδικούς τους.

- Αν το γονίδιο υπάρχει στο ακτινοβακτήριο τότε παίρνει την τιμή 1
- Αν το γονίδιο δεν υπάρχει στο ακτινοβακτήριο τότε μένει κενό

Βάσει αυτού του πίνακα κατασκευάστηκε ένα heatmap, βάσει του κώδικα [*7.1.4 Κώδικας 4*], όπου φαίνεται ποια γονίδια βρέθηκαν σε ποια ακτινοβακτήρια.

Το heatmap με κόκκινο χρώμα δείχνει που υπάρχει 1, δηλαδή που υπάρχει το γονίδιο, και με γαλάζιο χρώμα δείχνει ότι το γονίδιο δεν υπάρχει στο γονιδίωμα.

2.3 Χρήσιμες Πληροφορίες

Τα κομμάτια κώδικα σε γλώσσα προγραμματισμού R, οι εντολές που εκτελέστηκαν για να εφαρμοστούν τα εργαλεία στη γραμμή εντολών, καθώς και πίνακες με δεδομένα υπάρχουν αναλυτικά στο κεφάλαιο *7 ΠΑΡΑΡΤΗΜΑ*.

Οι εντολές, λόγω μεγάλου όγκου εκτελέστηκαν στους servers του Ινστιτούτου Εφαρμοσμένων Βιοεπιστημών (INEB) του Εθνικού Κέντρου Έρευνας και Τεχνολογικής Ανάπτυξης (ΕΚΕΤΑ).

Κεφάλαιο 3ο

3 ΑΠΟΤΕΛΕΣΜΑΤΑ

Στο παρόν κεφάλαιο θα αναλυθούν τα αποτελέσματα που προέκυψαν εφαρμόζοντας τις μεθοδολογίες που επεξηγήθηκαν στο κεφάλαιο *2 ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ*, σε δεδομένα τερπενίων ως γονίδια-στόχοι και ακτινοβακτηριδίων ως οργανισμοί ενδιαφέροντος.

3.1 Αποτελέσματα Προετοιμασίας γονιδίων-στόχων

Η προετοιμασία των γονιδίων-στόχων, σε αυτή την περίπτωση των γονιδίων που παράγουν τερπενικές συνθάσες ή πρωτεΐνες σχετικές με τη βιοσύνθεσή τους, πραγματοποιήθηκε σύμφωνα με τη μεθοδολογία στο κεφάλαιο *2.1.1 Προετοιμασία δεδομένων Γονιδίων-Στόχων*.

Από την αναζήτηση στα 4 αρχικά ακτινοβακτήρια, βρέθηκαν και συγκεντρώθηκαν πληροφορίες για:

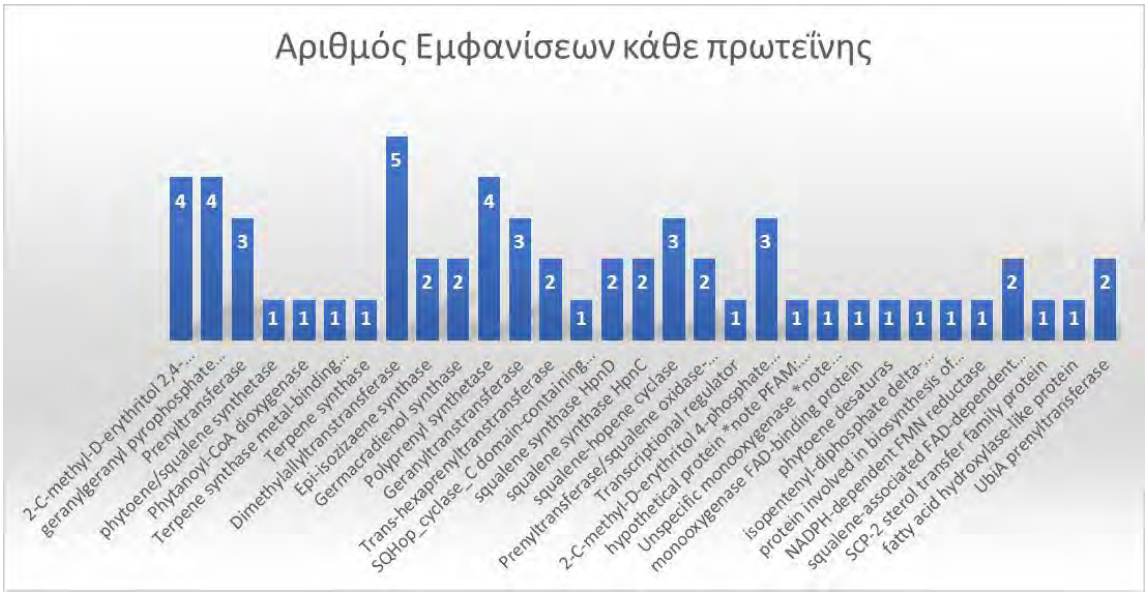
- 44 αλληλουχίες γονιδίων που παράγουν πρωτεΐνες σχετικές με τερπενοειδή
- 12 αλληλουχίες γονιδίων που είναι γειτονικά με τις πρωτεΐνες στόχους και είναι αρκετά σχετικά με τα τερπένια
- 4 αλληλουχίες γονιδίων που είναι γειτονικά με τις πρωτεΐνες στόχους και είναι λιγότερο σχετικά με τα τερπένια

Οι πρωτεΐνες έχουν διαφορές στην αλληλουχία, αλλά κάποιες παράγουν το ίδιο προϊόν. Έτσι μπορεί να υπάρχουν 4 γονίδια που παράγουν την πρωτεΐνη διφωσφορική συνθάση του γερανυλ-γερανυλίου, 3 γονίδια που παράγουν την πρωτεΐνη πρενυλοτρανσφεράση κ.ο.κ.

Τέτοιο είδους αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα (Πίνακας 2) και στην εικόνα (Εικόνα 15):

	Πρωτεΐνη	*Αριθμός Εμφανίσεων*
Γονίδια-στόχοι	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	4
	geranylgeranyl pyrophosphate synthase	4
	Prenyltransferase	3
	phytoene/squalene synthetase	1
	Phytanoyl-CoA dioxygenase	1
	Terpene synthase metal-binding domain-containing protein	1
	Terpene synthase	1
	Dimethylallyltranstransferase	5
	Epi-isozizaene synthase	2
	Germacradienol synthase	2
	Polyprenyl synthetase	4
	Geranyltranstransferase	3
	Trans-hexaprenyltranstransferase	2
	SQHop_cyclase_C domain-containing protein	1
	squalene synthase HpnD	2
	squalene synthase HpnC	2
	squalene-hopene cyclase	3
	Prenyltransferase/squalene oxidase-like repeat protein	2
	Transcriptional regulator	1
Γειτονικά γονίδια αρκετά σχετικά	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	3
	hypothetical protein *note PFAM: Antibiotic biosynthesis monooxygenase	1
	A0A0N1FIJ2-Unspecific monooxygenase *note PFAM: cytochrome P450 074	1
	monooxygenase FAD-binding protein	1
	phytoene desaturase	1
	isopentenyl-diphosphate delta-isomerase	1
	protein involved in biosynthesis of mitomycin antibiotics/polyketide fumonisin *note PFAM: Phytanoyl-CoA dioxygenase (PhyH)	1
	NADPH-dependent FMN reductase	1
	squalene-associated FAD-dependent desaturase	2
Γειτονικά γονίδια λιγότερο σχετικά	SCP-2 sterol transfer family protein	1
	fatty acid hydroxylase-like protein	1
	UbiA prenyltransferase	2

Πίνακας 2. Πίνακας που παρουσιάζει τον αριθμό εμφανίσεων κάθε πρωτεΐνης στα 4 βασικά ακτινοβακτήρια



Εικόνα 15. Διαγραμματική παρουσίαση του αριθμού εμφανίσεων της κάθε πρωτεΐνης στα 4 βασικά ακτινοβακτήρια

Αυτές είναι οι πρωτεΐνες σχετικές με τερπενικές συνθάσες που υπάρχουν στα 4 ακτινοβακτήρια και βρέθηκαν μέσω του EMBL αρχείου με την αναζήτηση των κωδικών της PFAM ή της InterPro. Κάθε μία από αυτές της πρωτεΐνες έχει ένα σημαντικό ρόλο στην παραγωγή τερπενίων και είτε παράγει η ίδια τερπενική συνθάση είτε συμμετέχει στη βιοσύνθεση τερπενοειδών. Η κάθε μία έχει διαφορετικό ρόλο αν και κάποιες μπορεί να εκφράζονται στο ίδιο μεταβολικό μονοπάτι ή να δημιουργούν κάποια ομάδα και να συνεκφράζονται.

3.2 Γειτονικά Γονίδια

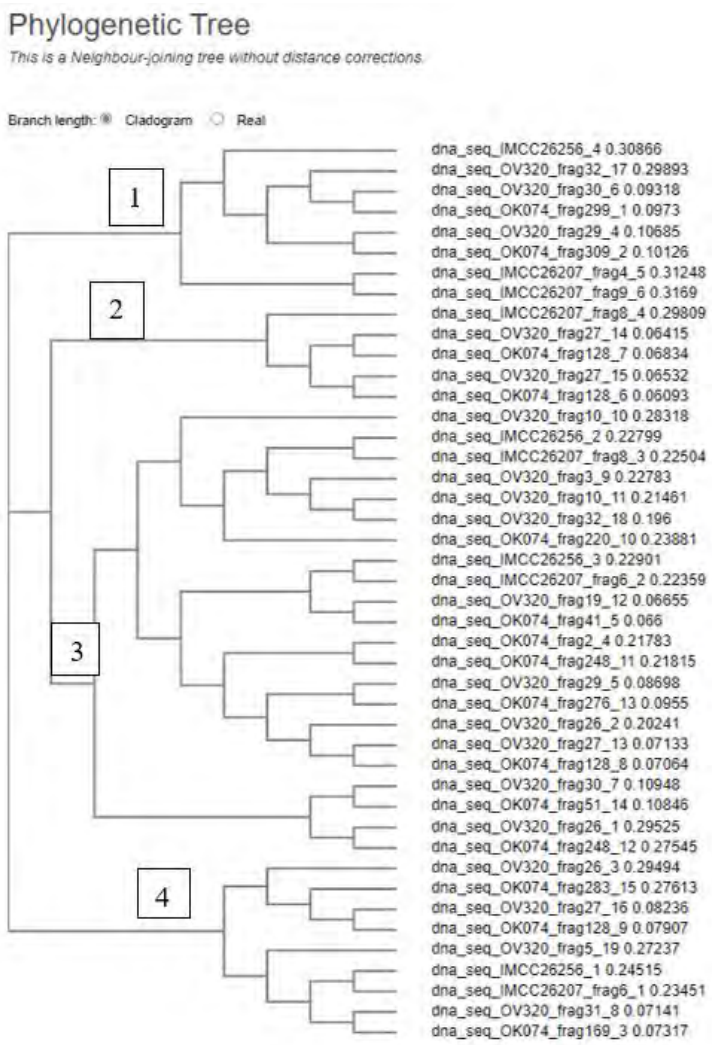
Με αυτή τη διαδικασία βρέθηκαν γονίδια γειτονικά στα γονίδια-στόχους που είναι πιθανό να σχετίζονται με την βιοσύνθεση και την παραγωγή τερπενίων. Κάποια από αυτά είναι πιο σχετικά με αυτές τις διαδικασίες και άλλα λιγότερο.

Είναι πιθανό κάποια από τα γονίδια που βρέθηκαν να είναι δίπλα-δίπλα να συνεκφράζονται ή να δημιουργούν ομάδες γονιδίων που συμμετέχουν στο ίδιο μεταβολικό μονοπάτι. Μία τέτοια ομάδα είναι τα εξής γονίδια: συνθάση σκουαλενίου (squalene synthase) HpnD, συνθάση σκουαλενίου (squalene synthase) HpnC, κυκλάση σκουαλενίου-χοπενίου (squalene- hopene cyclase) και ίσως η τρανσφεράση του διμεθυλλαλίου (Dimethylallyltranstransferase) που βρίσκονται μαζί στα γονιδιώματα που υπάρχουν και βάσει βιβλιογραφίας είναι γονίδια που εμπλέκονται στο ίδιο μεταβολικό μονοπάτι βιοσύνθεσης (*1.9.3 Γονίδια τερπενοειδών σε ομάδες*).

Αυτή η θεωρία θα φανεί αν έχει βάση και σε επόμενο κεφάλαιο, στην *2.2.3 Εύρεση τερπενοειδών σε γονιδιώματα ακτινοβακτηριδίων*, όπου θα παρατηρηθεί αν βρίσκονται και σε άλλα γονιδιώματα μαζί.

3.3 Πολλαπλή Στοίχιση

Για να ολοκληρωθεί η ανάλυση εξάχθηκαν και οι αλληλουχίες πριν τα γονίδια-στόχους, και συγκεντρώθηκαν οι αλληλουχίες των 1000 βάσεων πριν από κάθε γονίδιο. Αρχικά, πραγματοποιήθηκε η ανάλυση πολλαπλής στοίχισης στις αλληλουχίες των γονιδίων, με τη χρήση του CLUSTAL W (Madeira, 2022) για να φανεί ποια γονίδια είναι πιο συγγενικά. Προέκυψε το παρακάτω φυλογενετικό δέντρο (Εικόνα 16) στο οποίο ομαδοποιήθηκαν οι αλληλουχίες βάσει κοντινότερων κλαδιών, όπως στις αναλύσεις κοντινότερων γειτόνων:

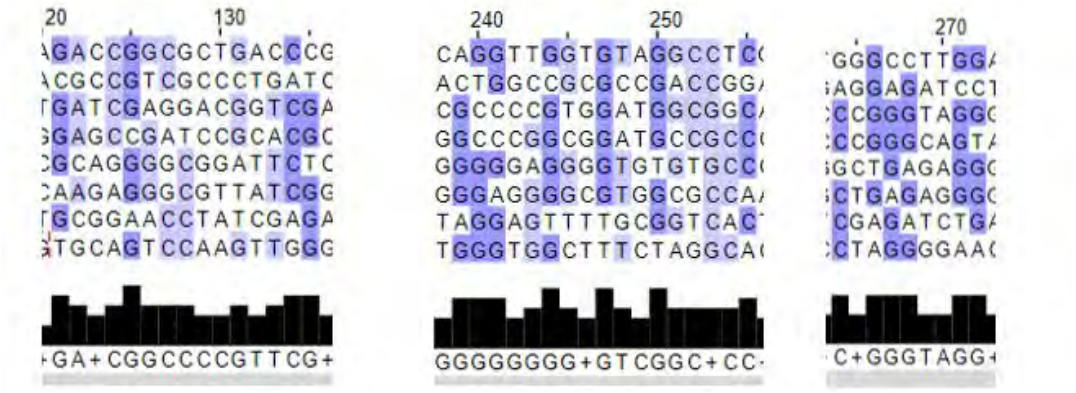


Εικόνα 16. Αποτελέσματα Πολλαπλής Στοίχισης των αλληλουχιών 1000 βάσεων πριν τα γονίδια-στόχους

Οι ομάδες που σχηματίστηκαν ήταν ξεκάθαρες και έτσι σε κάθε μία από αυτές εφαρμόστηκε εκ νέου πολλαπλή στοίχιση με το ίδιο εργαλείο, πλέον στις 1000 βάσεις πριν το γονίδιο. Τα αποτελέσματα προβλήθηκαν στο Jalview (Waterhouse, 2009), ένα πρόγραμμα οπτικοποίησης. Εκεί φάνηκαν να σχηματίζονται κάποιες περιοχές οι οποίες μπορούν να χαρακτηριστούν ως συντηρημένες στα γονίδια και επιπλέον εμφανίζουν αυτόματα μία συνεκτική αλληλουχία (consensus sequence) για κάθε ομάδα.

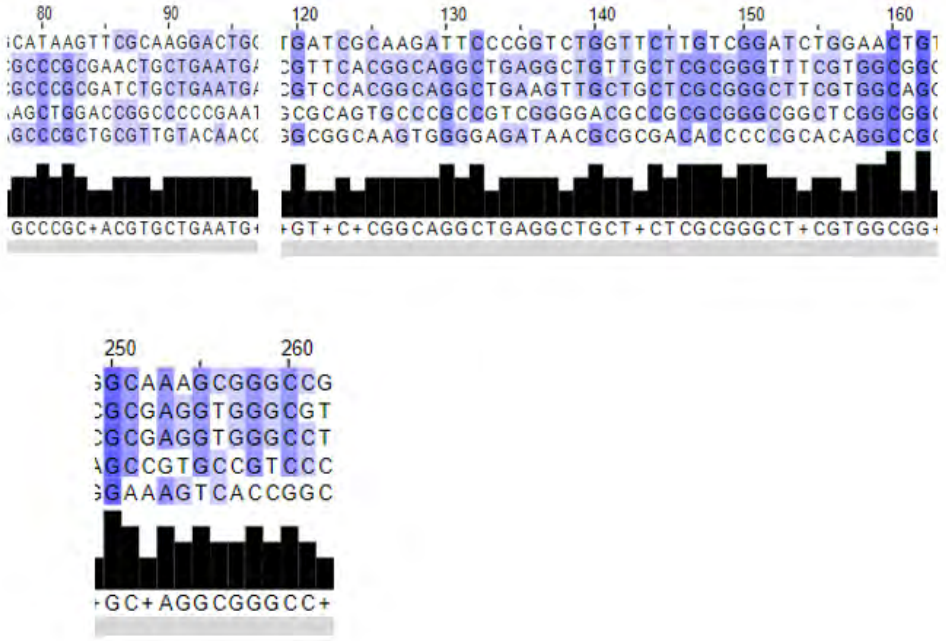
Στις παρακάτω εικόνες παρουσιάζονται κάποια σημεία τα οποία θα μπορούσαν να χαρακτηριστούν συντηρημένες περιοχές:

- Για την πρώτη ομάδα (Εικόνα 17):



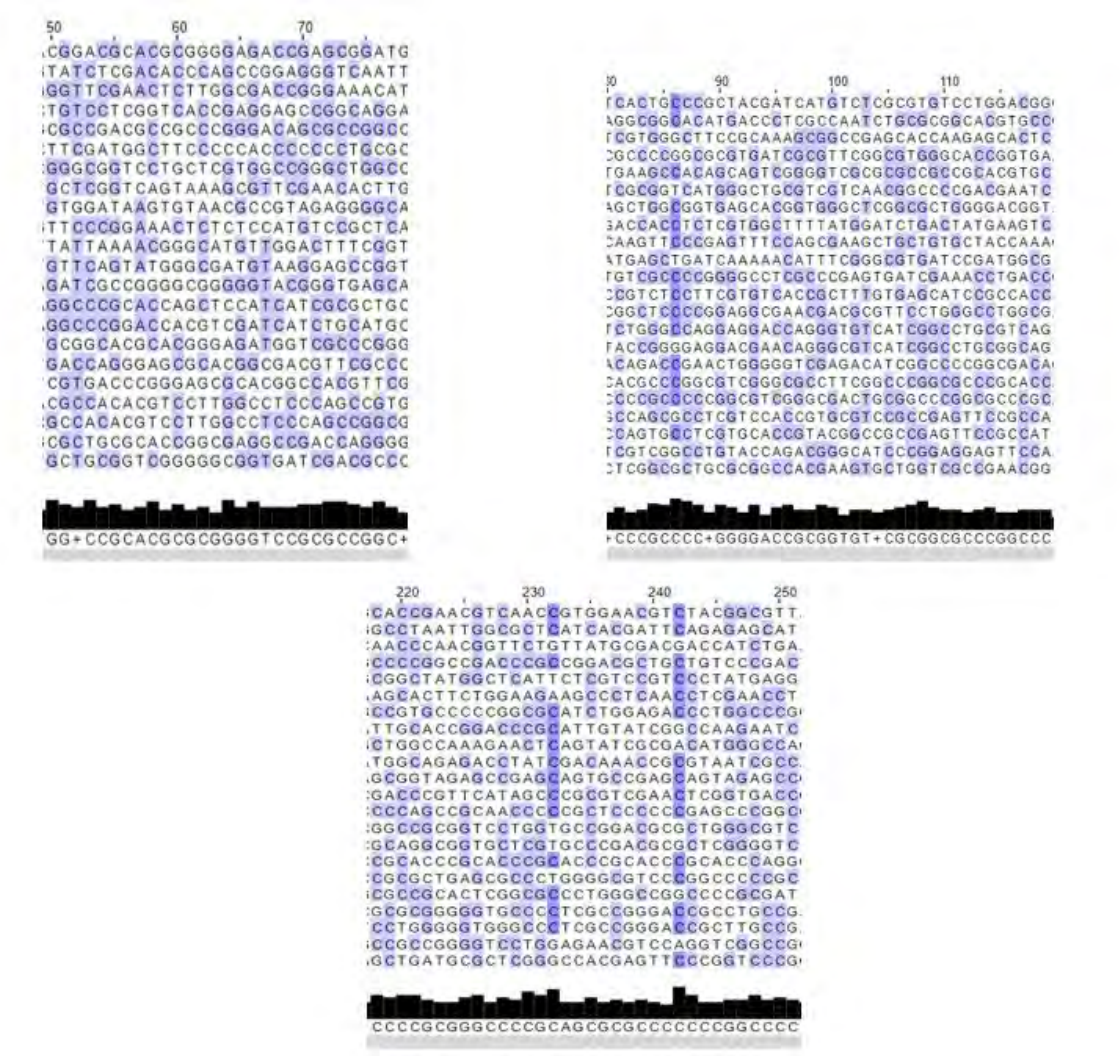
Εικόνα 17. Πιθανές συντηρημένες περιοχές στην πρώτη ομάδα του φυλογενετικού δέντρου

- Για την δεύτερη ομάδα (Εικόνα 18):



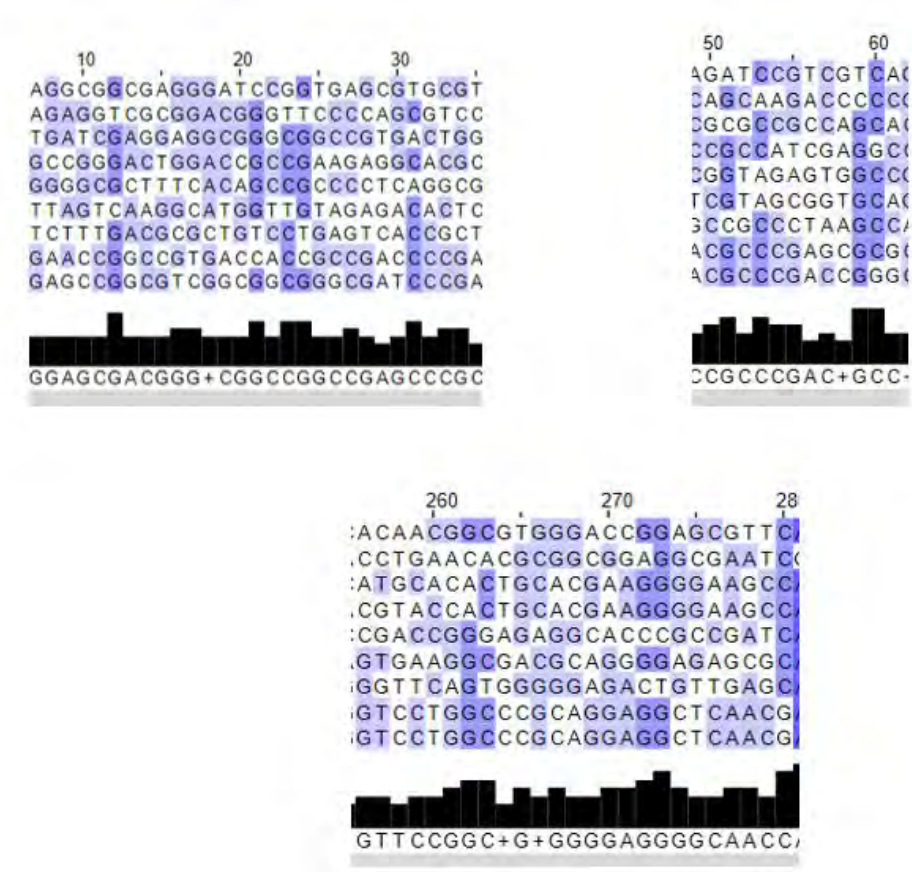
Εικόνα 18 Πιθανές συντηρημένες περιοχές στην δεύτερη ομάδα του φυλογενετικού δέντρου

- Για την Τρίτη ομάδα (Εικόνα 19):



Εικόνα 19. Πιθανές συντηρημένες περιοχές στην τρίτη ομάδα του φυλογενετικού δέντρου

- Για την τέταρτη ομάδα (Εικόνα 20):



Εικόνα 20. Πιθανές συντηρημένες περιοχές στην τέταρτη ομάδα του φυλογενετικού δέντρου

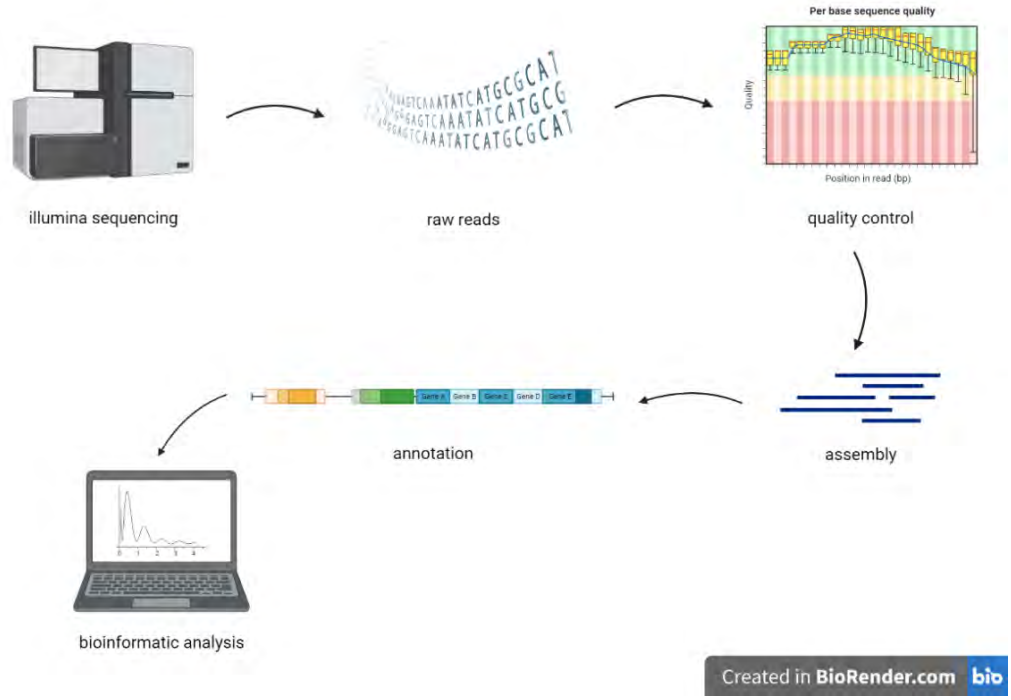
Όσο πιο έντονο είναι το μωβ χρώμα, τόσο καλύτερη η στοίχιση και άρα τόσο πιο ισχυρά συντηρημένη είναι μία περιοχή. Στην ομάδα 3 είναι φυσιολογικό να μην υπάρχουν τόσο ισχυρά συντηρημένες περιοχές καθώς είναι η ομάδα με τις περισσότερες αλληλουχίες επομένως έχει διευρυνθεί λίγο και η αναζήτηση συντηρημένων περιοχών. Αυτές οι περιοχές μπορεί να είναι ρυθμιστικά στοιχεία που ελέγχουν την έκφραση των γονιδίων που τα ακολουθούν, δηλαδή συμβάλουν είτε στην αποσιώπηση των γονιδίων είτε στην ενίσχυση της έκφρασης τους, κάτι που δεν έχει ελεγχθεί στην παρούσα εργασία αλλά είναι ένας από τους μελλοντικούς στόχους [*5.1 Μελλοντικοί Στόχοι*].

3.4 Κατασκευή εκ νέου γονιδιωμάτων

Πραγματοποιήθηκαν περαιτέρω αναλύσεις για να προκύψουν συμπεράσματα σχετικά με το ποιες από τις πρωτεΐνες που βρέθηκαν στο *3.1 Αποτελέσματα Προετοιμασίας γονιδίων-στόχων*, βρίσκονται σε άλλους οργανισμούς ακτινοβακτηριδίων, ποιες υπάρχουν σε μεγαλύτερο ποσοστό και ποιες καθόλου.

Για να προκύψουν τέτοια αποτελέσματα έπρεπε να εφαρμοστεί η μεθοδολογία και σε άλλα γονιδιώματα όπως και έγινε.

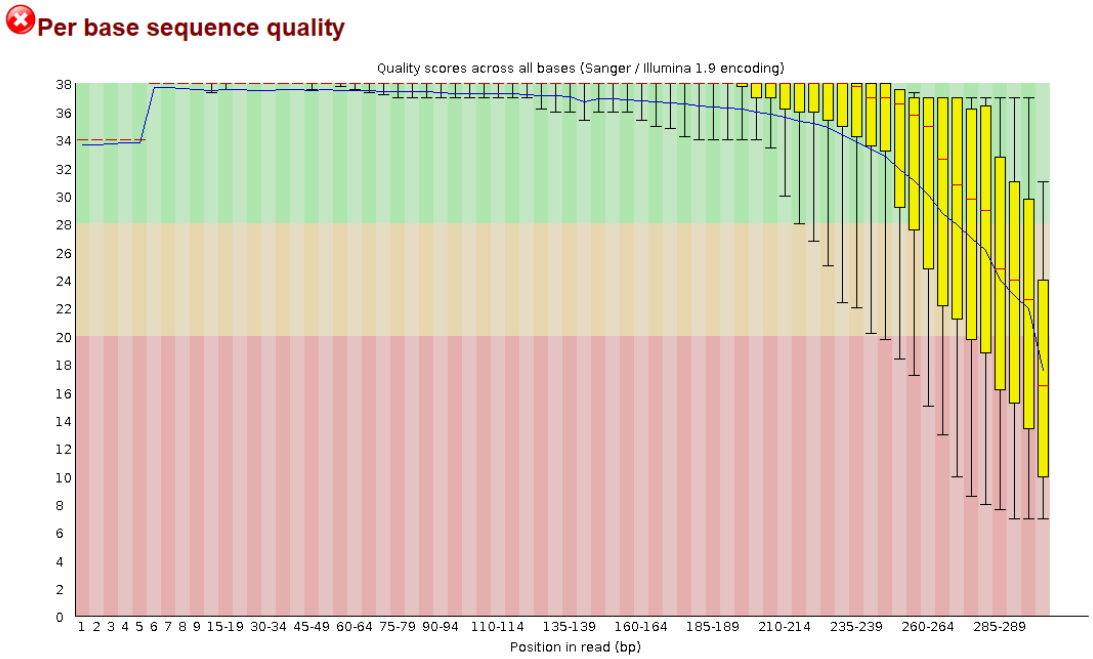
Για 2 στελέχη του εργαστηρίου εφαρμόστηκε η μεθοδολογία κατασκευής του γονιδιώματος (Εικόνα 21), καθώς αυτά υπήρχαν ως αρχεία από αλληλούχιση.



Εικόνα 21. Οπτικοποίηση διαδικασίας εξαγωγής αποτελεσμάτων για την κατασκευή του γονιδιώματος

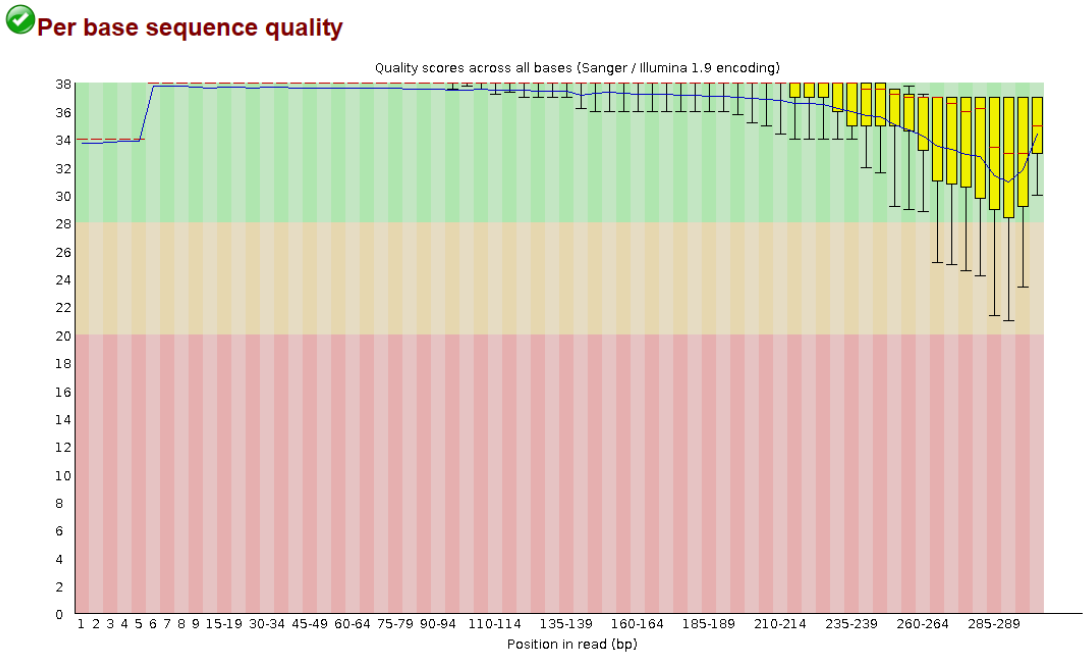
Τα δύο αυτά στελέχη υποβλήθηκαν στη διαδικασία και τα αποτελέσματα της αναφοράς του FastQC (Andrews, 2010), πριν και μετά το κόψιμο και τον καθαρισμό παρουσιάζονται στις παρακάτω εικόνες. Οι εικόνες είναι ενδεικτικές για ένα από τα

στελέχη, αλλά παρόμοιες ήταν και οι εικόνες του 2ου. Οι εικόνες είναι από ένα διάγραμμα στην FastQC αναφορά που προκύπτει με την εκτέλεση του συνώνυμου εργαλείου, και δείχνει πόσο καλή ήταν η ποιότητα των κομματιών ανά βάση.



Εικόνα 22. Παράδειγμα αναφοράς FastQC πριν το κόψιμο και τον καθαρισμό

Είναι φανερό πως η ποιότητα ειδικά προς το τέλος δεν είναι καλή (Εικόνα 22) και αυτά τα κομμάτια δεν μπορούν να χρησιμοποιηθούν για να κατασκευαστεί ένα γονιδίωμα.



Εικόνα 23. Παράδειγμα αναφοράς FastQC μετά το κόψιμο και τον καθαρισμό

Αντιθέτως η εικόνα μετά τον καθαρισμό φαίνεται πολύ πιο καθαρή (Εικόνα 23) και τα αποτελέσματα της αναφοράς είναι αποδεκτά για να συνεχιστεί η διαδικασία.

Επομένως εκτελέστηκε και η κατασκευή του γονιδώματος με τη χρήση του SPAdes (Nurk, Sergey, et al., 2013) και στα αποτελέσματα πάρθηκαν αλληλουχίες σε κομμάτια. Μετά την εφαρμογή του QUAST (Mikheenko, 2016) προέκυψαν τα εξής μέτρα (Πίνακας 3):

	Ακτινοβακτήριο 1	*Ακτινοβακτήριο 2*
Αριθμός κομματιών	87	114
N50	163584	139710
N75	112649	87087
L50	15	16
L75	27	34
BUSCO	98-99%	98-99%

Πίνακας 3. Αποτελέσματα Αξιολόγησης της κατασκευής του γονιδιώματος με SPADES

Όπως μπορεί να παρατηρηθεί, ο αριθμός των κομματιών είναι σχετικά μεγάλος. Οι μετρήσεις N50, N75 είναι ικανοποιητικά και στις 2 περιπτώσεις, λίγο καλύτερη απόδοση έχει δώσει το ακτινοβακτήριο 1, όπως και στα μέτρα L50 και L75.

Από την εφαρμογή του BUSCO προέκυψε ποσοστό πληρότητας γύρω στο 98.4% στη βάση των βακτηριδίων και 99% στη βάση των *Streptomycetales* που είναι υψηλά ποσοστά. Στη συνέχεια έγινε η σύγκρισή τους με το γονιδίωμα αναφοράς για κάθε στέλεχος με τη χρήση του IGV (Robinson, James T, et al., 2011) και τα αποτελέσματα έδειξαν πως υπάρχουν μερικά κενά, όχι ανησυχητικά πολλά, και κάποιες επικαλύψεις. Γι' αυτό έγινε η κατασκευή και με άλλο εργαλείο, το Unicycler (Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, & Kathryn E. Holt, 2017), και αξιολογήθηκε με το QUAST, του οποίο τα αποτελέσματα εμφανίζονται παρακάτω (Πίνακας 4):

	Ακτινοβακτήριο 1	*Ακτινοβακτήριο 2*
Αριθμός κομματιών	89	109
N50	154644	120033
N75	111258	78737
L50	14	17
L75	27	36

Πίνακας 4. Αποτελέσματα Αξιολόγησης της κατασκευής του γονιδιώματος με UNICYCLER

Εδώ μπορεί να παρατηρηθεί πως έχουμε καλύτερα αποτελέσματα στον αριθμό των κομματιών που προέκυψαν και μικρές διαφορές στα υπόλοιπα μέτρα.

Αποφασίστηκε όπως αναφέρθηκε σε παραπάνω κεφάλαιο, να βελτιωθεί το γονιδίωμα που είχε κατασκευαστεί με την εφαρμογή του RagTag (Alonge, 2021). Με την αξιολόγηση αυτού του εργαλείου ήταν φανερό πως τα αποτελέσματα που προέκυψαν ήταν ποιοτικά και τα γονιδιώματα που προέκυψαν μπορούσαν να χρησιμοποιηθούν.

Τα αποτελέσματα αναλύονται παρακάτω (Πίνακας 5):

	Ακτινοβακτήριο 1 (unicycler)	*Ακτινοβακτήριο 2*	*Ακτινοβακτήριο 1 (SPAdes)*
Αριθμός κομματιών	10	14	11
N50	6466871	7434270	6752555
N75	6466871	7434270	6752555
L50	1	1	1
L75	1	1	1
BUSCO	96.8-98.4%	98.4-99.1%	

Πίνακας 5. Αποτελέσματα Αξιολόγησης της κατασκευής του γονιδιώματος με RagTag

- ✓ Ο αριθμός των κομματιών που προέκυψαν είναι μικρός
- ✓ Τα μέτρα N50, N75 ταυτίζονται και είναι μεγάλος αριθμός, κάτι που δείχνει πως το μεγαλύτερο μέρος του γονιδιώματος που κατασκευάστηκε είναι ενιαίο
- ✓ Το παραπάνω συμπέρασμα επιβεβαιώνεται με τα μέτρα L50, L75 που σε κάθε περίπτωση είναι 1, άρα τουλάχιστον το 75% του γονιδιώματος βρίσκεται στο πρώτο κομμάτι.
- ✓ Τα αποτελέσματα του BUSCO ήταν ικανοποιητικά σε ποσοστά μεγαλύτερα από 95%

Από τα δύο γονιδιώματα για το ακτινοβακτήριο 1 επιλέχθηκε αυτό που προέκυψε από το Unicycler+RagTag, καθώς έδωσε λιγότερα κομμάτια.

Το επόμενο βήμα ήταν η ανίχνευση της ομάδας γονιδίων ενδιαφέροντος, και αυτά ήταν τα τερπένια. Έγινε λοιπόν η εφαρμογή του εργαλείου PROKKA (Seemann, 2014), το οποίο στα αποτελέσματά του έδειξε πως υπάρχουν τερπένια και στα δύο ακτινοβακτήρια. Έδειξε επιπλέον και την ύπαρξη πολλών υποθετικών πρωτεϊνών, που

σχημάτιζαν και κάποιες ομάδες γονιδίων. Επομένως ήταν επιτακτική ανάγκη να ελεγχθεί η ύπαρξη κι άλλων τερπενίων και πιο συγκεκριμένων πρωτεϊνών.

Ένα επιπλέον βήμα που εφαρμόστηκε είναι η στοίχιση των στελεχών με μία βάση δεδομένων όπως η nr, για να δοθούν συγγενείς οργανισμοί. Τα αποτελέσματα που προέκυψαν από τα φυλογενετικά δέντρα φαίνονται εδώ (Πίνακας 6):

Ακτινοβακτήριο 1	*Streptomyces albidoflavus*
Ακτινοβακτήριο 2	Κυρίως unclassified *Streptomyces*

Πίνακας 6. Συγγενείς οργανισμοί των στελεχών

3.5 Εύρεση τερπενίων σε γονιδιώματα ακτινοβακτηριδίων

Στο παρόν κεφάλαιο θα αναλυθούν τα αποτελέσματα της τελικής εφαρμογής της μεθοδολογίας για την εύρεση τερπενίων σε γονιδιώματα ακτινοβακτηριδίων που είτε κατασκευάστηκαν είτε υπήρχαν σε κάποια δημόσια βάση δεδομένων.

Τα αποτελέσματα που θα αναλυθούν παρακάτω περιέχουν πληροφορίες σχετικά με:

- ✓ ποια γονίδια υπάρχουν στα ακτινοβακτήρια που μελετήθηκαν
- ✓ ποιες πρωτεΐνες φαίνεται να εκφράζονται σε ομάδες

3.5.1 Γονίδια τερπενίων που υπάρχουν στα γονιδιώματα ακτινοβακτηριδίων

Μετά τις αναζητήσεις που πραγματοποιήθηκαν σύμφωνα με το κεφάλαιο *2.2.3 Εύρεση τερπενοειδών σε γονιδιώματα ακτινοβακτηριδίων*, προέκυψαν πίνακες οι οποίοι περιέχουν λίστες με γονίδια που υπάρχουν στο κάθε ακτινοβακτήριο, σε ποια θέση, σε τι ποσοστό βρέθηκε να υπάρχει ομοιότητα αλλά και τι e-value και bit-score είχαν.

Ενδεικτικά στους παρακάτω πίνακες παρουσιάζονται κάποια από αυτά τα αποτελέσματα. Περισσότερα αποτελέσματα φαίνονται στο παράρτημα [*7.4 Πίνακες Αναλυτικών Αποτελεσμάτων*].

Για το ακτινοβακτήριο 1, του οποίου το γονιδίωμα κατασκευάστηκε με τη μεθοδολογία του κεφαλαίου *2.1.5 Προετοιμασία δεδομένων των οργανισμών-στόχων* βρέθηκαν οι εξής πρωτεΐνες (Πίνακας 7):

Προϊόν που παράγεται	**Κωδικός InterPro**	**%Ομοιότητα**	**e-value**	**Bit score**
Γονίδια Στόχοι				
Epi-isozizaene synthase	A0A0N1NPH7	86.139	5.14e-24	110
Dimethylallyltranstransferase	A0A0N1H381	77.585	1.97e-167	586
Germacradienol synthase	A0A0N1GNM8	80.428	0.0	728
MULTISPECIES: polyprenyl synthetase family protein [unclassified Streptomyces]	A0A0N1NU03	83.645	0.0	948
Dimethylallyltranstransferase	A0A0N1GTW2	84.732	0.0	1103
Dimethylallyltranstransferase	A0A0N1GTW2	77.876	2.45e-32	137
squalene synthase HpnD	A0A0N1GM64	82.545	0.0	773
squalene-hopene cyclase	A0A0N1H9Z9	82.311	0.0	1550
Germacradienol synthase	A0A0N1FJV3	81.682	0.0	811
Trans-hexaprenyltranstransferase	A0A0N0N5M1	84.539	0.0	996
squalene synthase HpnC	A0A0N1GKQ9	78.477	2.69e-160	562
squalene synthase HpnD	A0A0N1NIL0	83.69	0.0	828
Dimethylallyltranstransferase	A0A0N1G9M3	86.945	0.0	1223
squalene-hopene cyclase	A0A0N1G4W1	82.432	0.0	1598
Γειτονικές πρωτεΐνες				
squalene-associated FAD-dependent desaturase	A0A0N1GSC6	76.749	0.0	708
squalene-associated FAD-dependent desaturase	A0A0N1GAB7	78.034	0.0	798

Πίνακας 7. Πίνακας με τα αποτελέσματα των γονιδίων που βρέθηκαν στο ακτινοβακτήριο 1

Στο ακτινοβακτήριο 2, του οποίου το γονιδίωμα κατασκευάστηκε με την ίδια μεθοδολογία βρέθηκαν οι εξής πρωτεΐνες (Πίνακας 8):

Προϊόν που παράγεται	Κωδικός InterPro	%Ομοιότητα	e-value	Bit score
Γονίδια Στόχοι				
Epi-isozizaene synthase	A0A0N1NPH7	82.108	0.0	874
Dimethylallyltranstransferase	A0A0N1H381	81.391	0.0	822
Germacradienol synthase	A0A0N1GNM8	87.884	0.0	1256
2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	A0A0N1NPP5	88.106	7.90e-154	540
MULTISPECIES: polyprenyl synthetase family protein [unclassified Streptomyces]	A0A0N1NU03	86.601	0.0	1114
Dimethylallyltranstransferase	A0A0N1GTW2	88.632	0.0	1367
squalene synthase HpnD	A0A0N1GM64	86.378	0.0	1027
squalene synthase HpnC	A0A0N0NBI2	83.025	0.0	821
squalene-hopene cyclase	A0A0N1H9Z9	86.146	0.0	2119
Germacradienol synthase	A0A0N1FJV3	82.571	0.0	1797
Epi-isozizaene synthase	A0A0N0MKU1	79.962	0.0	760
2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	A0A0N1GHD8	87.364	8.50e-149	523
Trans-hexaprenyltranstransferase	A0A0N0N5M1	87.786	0.0	1173
squalene synthase HpnC	A0A0N1GKQ9	84.447	0.0	846
squalene synthase HpnD	A0A0N1NIL0	86.868	0.0	1038
Dimethylallyltranstransferase	A0A0N1G9M3	88.759	0.0	1358
squalene-hopene cyclase	A0A0N1G4W1	85.262	0.0	2010
Dimethylallyltranstransferase	A0A0N1FQ50	81.099	0.0	809
Γειτονικές πρωτεΐνες				
Unspecific monooxygenase	A0A0N1H714	82.992	0.0	1177
NADPH-dependent FMN reductase	A0A0N0NEF3	84.882	7.25e-155	544
squalene-associated FAD-dependent desaturase	A0A0N1GSC6	85.255	0.0	1402
squalene-associated FAD-dependent desaturase	A0A0N1GAB7	82.963	0.0	1195

Πίνακας 8. Πίνακας με τα αποτελέσματα των γονιδίων που βρέθηκαν στο ακτινοβακτήριο 2

Για το ακτινοβακτήρια των οποίων τα γονιδιώματα βρέθηκαν στη βάση ENA (ENA - European Nucleotide Archive, n.d.).

Για το GCA_001704195 (Πίνακας 9):

Προϊόν που παράγεται	Κωδικός InterPro	%Ομοιότητα	e-value	Bit score
Γονίδια Στόχοι				
Epi-isozizaene synthase	A0A0N1NPH7	72.871	8.56E-82	302
Dimethylallyltranstransferase	A0A0N1H381	77.444	1.29E-164	577
Germacradienol synthase	A0A0N1GNM8	80.448	0	728
Trans-hexaprenyltranstransferase	A0A0N1NU03	83.448	0	937
Dimethylallyltranstransferase	A0A0N1GTW2	84.492	0	1086
		83.459	7.43E-28	122
squalene-hopene cyclase	A0A0N1H9Z9	82.262	0	1550
Germacradienol synthase	A0A0N1FJV3	81.001	0	761
		74.297	2.23E-80	298
Trans-hexaprenyltranstransferase	A0A0N0N5M1	84.325	0	985
squalene synthase HpnC	A0A0N1GKQ9	78.429	2.91E-160	562
squalene synthase HpnD	A0A0N1NIL0	82.958	0	819
Dimethylallyltranstransferase	A0A0N1G9M3	86.57	0	1201
squalene-hopene cyclase	A0A0N1G4W1	82.727	0	1633
Γειτονικές πρωτεΐνες				
squalene-associated FAD-dependent desaturase	A0A0N1GSC6	77.241	0	723
squalene-associated FAD-dependent desaturase	A0A0N1GAB7	77.879	0	787

Πίνακας 9. Πίνακας με τα αποτελέσματα των γονιδίων που βρέθηκαν στο ακτινοβακτήριο GCA_001704195

Για το GCA_001865315 (Πίνακας 10):

Προϊόν που παράγεται	Κωδικός InterPro	%Ομοιότητα	e-value	Bit score
Γονίδια Στόχοι				
Epi-isozizaene synthase	A0A0N1NPH7	72.871	1.75E-81	302
Dimethylallyltranstransferase	A0A0N1H381	77.541	5.66E-166	582
Germacradienol synthase	A0A0N1GNM8	80.448	0	728
Trans-hexaprenyltranstransferase	A0A0N1NU03	83.448	0	937
Dimethylallyltranstransferase	A0A0N1GTW2	84.492	0	1086
		83.459	1.52E-27	122
squalene-hopene cyclase	A0A0N1H9Z9	82.262	0	1550
Germacradienol synthase	A0A0N1FJV3	81.001	0	761
		74.297	4.56E-80	298
Trans-hexaprenyltranstransferase	A0A0N0N5M1	84.325	0	985
squalene synthase HpnC	A0A0N1GKQ9	78.429	5.95E-160	562
squalene synthase HpnD	A0A0N1NIL0	82.851	0	813
Dimethylallyltranstransferase	A0A0N1G9M3	86.57	0	1201
squalene-hopene cyclase	A0A0N1G4W1	82.727	0	1633
Γειτονικές πρωτεΐνες				
squalene-associated FAD-dependent desaturase	A0A0N1GSC6	77.343	0	730
squalene-associated FAD-dependent desaturase	A0A0N1GAB7	77.875	0	809

Πίνακας 10. Πίνακας με τα αποτελέσματα των γονιδίων που βρέθηκαν στο ακτινοβακτήριο GCA_001865315

Για το GCA_017377825 (Πίνακας 11):

Προϊόν που παράγεται	Κωδικός InterPro	%Ομοιότητα	e-value	Bit score
Γονίδια Στόχοι				
Epi-isozizaene synthase	A0A0N1NPH7	72.944	1.85E-83	307
Dimethylallyltranstransferase	A0A0N1H381	77.67	2.79E-166	582
Germacradienol synthase	A0A0N1GNM8	80.918	0	754
Trans-hexaprenyltranstransferase	A0A0N1NU03	83.251	0	926
Dimethylallyltranstransferase	A0A0N1GTW2	84.684	0	1098
		84.496	5.78E-29	126
squalene-hopene cyclase	A0A0N1H9Z9	82.317	0	1555
Germacradienol synthase	A0A0N1FJV3	81.325	0	793
		74.169	1.04E-78	292
Trans-hexaprenyltranstransferase	A0A0N0N5M1	84.127	0	974
squalene synthase HpnC	A0A0N1GKQ9	77.987	1.37E-153	540
squalene synthase HpnD	A0A0N1NIL0	83.296	0	806
Dimethylallyltranstransferase	A0A0N1G9M3	86.854	0	1218
squalene-hopene cyclase	A0A0N1G4W1	82.674	0	1628
Γειτονικές πρωτεΐνες				
squalene-associated FAD-dependent desaturase	A0A0N1GSC6	76.94	0	730
squalene-associated FAD-dependent desaturase	A0A0N1GAB7	78.102	0	809

Πίνακας 11. Πίνακας με τα αποτελέσματα των γονιδίων που βρέθηκαν στο ακτινοβακτήριο GCA_017377825

Για το GCA_023702435 (Πίνακας 12):

Προϊόν που παράγεται	Κωδικός InterPro	%Ομοιότητα	e-value	Bit score
Γονίδια Στόχοι				
Epi-isozizaene synthase	A0A0N1NPH7	81.966	0	878
Dimethylallyltranstransferase	A0A0N1H381	81.489	0	828
Germacradienol synthase	A0A0N1GNM8	87.907	0	1256
2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	A0A0N1NPP5	88.72	1.72E-160	562
Trans-hexaprenyltranstransferase	A0A0N1NU03	86.969	0	1136
Dimethylallyltranstransferase	A0A0N1GTW2	88.789	0	1360
squalene synthase HpnD	A0A0N1GM64	86.272	0	1022
squalene-hopene cyclase	A0A0N1H9Z9	86.247	0	2130
Germacradienol synthase	A0A0N1FJV3	82.427	0	1781
Epi-isozizaene synthase	A0A0N0MKU1	79.866	0	754
2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	A0A0N1GHD8	88.121	1.85E-155	545
Trans-hexaprenyltranstransferase	A0A0N0N5M1	87.921	0	1184
squalene synthase HpnD	A0A0N1NIL0	86.975	0	1044
Dimethylallyltranstransferase	A0A0N1G9M3	88.52	0	1347
squalene-hopene cyclase	A0A0N1G4W1	85.204	0	2004
Dimethylallyltranstransferase	A0A0N1FQ50	81.292	0	821
Γειτονικές πρωτεΐνες				
Unspecific monooxygenase	A0A0N1H714	82.992	0	1177
NADPH-dependent FMN reductase	A0A0N0NEF3	84.787	1.23E-152	536
squalene-associated FAD-dependent desaturase	A0A0N1GSC6	84.964	0	1380
Unspecific monooxygenase *note PFAM: cytochrome P450	A0A0N1FIJ2	76.169	0	676
squalene-associated FAD-dependent desaturase	A0A0N1GAB7	82.634	0	1194

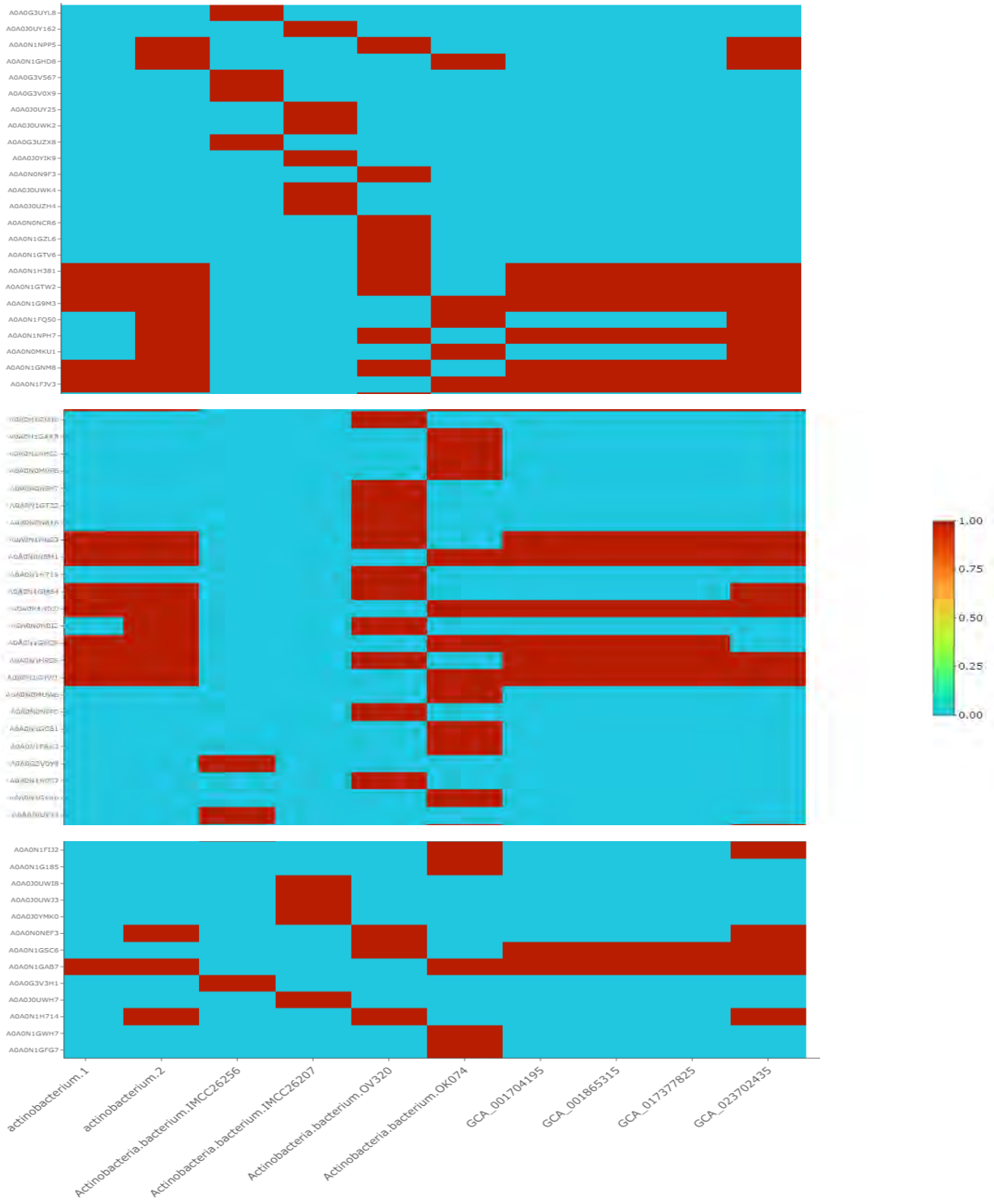
Πίνακας 12. Πίνακας με τα αποτελέσματα των γονιδίων που βρέθηκαν στο ακτινοβακτήριο GCA_023702435

Παρατηρείται πως κάποιες πρωτεΐνες όπως η συνθάση επι-ισοζιζαενίου (Epi-isozizaene synthase), η τρανσφεράση διμεθυλλαλίου (Dimethylallyltranstransferase), η συνθάση γερμακραδιενολίου (Germacradienol synthase), η τρανς-εξαπρενυλοτρανσφεράση (Trans-hexaprenyltranstransferase), κυκλάση σκουαλένιου-χοπένιου (squalene-hopene cyclase), η συνθάση σκουαλένιου (squalene synthase) HpnD και συνθάση σκουαλένιου (squalene synthase) HpnC, εμφανίζονται αρκετά συχνά στα ακτινοβακτήρια.

Για να υπάρχει συγκεντρωτικά η πληροφορία σχετικά με ποια από τα γονίδια-στόχους, δηλαδή ποια τερπενοειδή υπάρχουν σε ποια ακτινοβακτήρια, κατασκευάστηκε ένας πίνακας όπου στις στήλες έχει τα ακτινοβακτήρια και στις γραμμές τις πρωτεΐνες με τους κωδικούς τους.

- Αν το γονίδιο υπάρχει στο ακρινοβακτήριο τότε παίρνει την τιμή 1
- Αν το γονίδιο δεν υπάρχει στο ακτινοβακτήριο τότε μένει κενό

Βάση αυτού του πίνακα κατασκευάστηκε ένα heatmap, το οποίο με ποιο θερμό χρώμα δείχνει που υπάρχει 1, δηλαδή που υπάρχει το γονίδιο (στην περίπτωση αυτή υπάρχει κόκκινο χρώμα), και με ποιο ψυχρό χρώμα δείχνει ότι το γονίδιο δεν υπάρχει στο γονιδίωμα (στην περίπτωση αυτή υπάρχει γαλάζιο χρώμα).



Εικόνα 24. Heatmap που δείχνει ποια γονίδια τερπενίων βρίσκονται στα γονιδιώματα ακτινοβακτηριδίων

Ο αντίστοιχος κώδικα δίνεται στο [*7.1.4 Κώδικας 4*].

Μία επιπλέον παρατήρηση που αναφέρθηκε σε παραπάνω κεφάλαιο είναι ότι τα εξής γονίδια φαίνεται και σε άλλα γονιδιώματα να βρίσκονται μαζί και ίσως να συνεκφράζονται. Τα γονίδια αυτά είναι: squalene synthase HpnD, squalene synthase HpnC, squalene-hopene cyclase και Dimethylallyltranstransferase.

Κεφάλαιο 4ο

4 ΣΥΖΗΤΗΣΗ

Τα πιο σημαντικά αποτελέσματα που αξίζει να αναφερθούν είναι πως τα ακτινοβακτήρια φαίνεται να περιέχουν αρκετά γονίδια σχετικά με τη βιοσύνθεση και την παραγωγή τερπενίων αλλά και ότι περιέχουν αρκετά γονίδια τερπενικών συνθασών. Είναι σημαντική η παρατήρηση αυτή καθώς φαίνεται πως και τα βακτήρια είναι σημαντική πηγή τερπενίων παρόλο που τα τερπένια έχουν συνδεθεί κυρίως με τα φυτά. Όπως έχει αναφερθεί τα βακτήρια δεν έχουν αναλυθεί τόσο για την παραγωγή των τερπενικών συνθασών αν και είναι γνωστό πως είναι πλούσια πηγή δευτερογενών μεταβολιτών.

Πιο συγκεκριμένα, από την παρούσα εργασία προκύπτει πως τα τερπενοειδή που παράγονται πιο συχνά στα ακτινοβακτήρια φαίνεται να είναι τα εξής:

- Συνθάση Επι-Ισοζιζαενίου
- Τρανσφεράση Διμεθαλλυλίου
- Συνθάση Γερμακραδιενόλης
- Τρανς- Εξαπρενυλοτρανσφεράση
- Κυκλάση Σκουαλένιου-Χοπένιου
- Συνθάση Σκουαλένιου HpnD
- Συνθάση Σκουαλένιου HpnC
- FAD εξαρτώμενη δεσατουράση σχετική με Σκουαλένιο

Παρ ’όλα αυτά είναι σχεδόν αδύνατο να γίνει η πρόβλεψη για το τελικό προϊόν που θα παραχθεί, και επομένως η πρόβλεψη γενικεύεται στις κατηγορίες τερπενίων που ανήκουν, όπως εδώ παρατηρούνται σεσκιτερπένια, τριτερπένια κ.α.

Τα αποτελέσματα τις στοίχισης έδωσαν πολύ καλά ποσοστά στα γονίδια που βρέθηκαν, πάνω από 70%, όπως και χαμηλό e-value (κοντά στο 0), που είναι ένδειξη καλής στοίχισης. Το bit-score ήταν υψηλό επομένως από τα δύο αυτά μέτρα φαίνεται πως τα αποτελέσματα δεν οφείλονται σε τυχαίο γεγονός, η στοίχιση δεν έγινε κατά τύχη.

Σύμφωνα και με τη βιβλιογραφία που αναλύθηκε στο κεφάλαιο *1.9.3 Γονίδια τερπενοειδών σε ομάδες*, τα εξής γονίδια- squalene-hopene cyclase, squalene synthase HpnD, squalene synthase HpnC και Dimethylallyltranstransferase- ανήκουν στο ίδιο βιοσυνθετικό μονοπάτι. Η υπόθεση, λοιπόν, που βασίστηκε σε αυτό ήταν η εξής: ίσως, τα γονίδια αυτά εκφράζονται σε ομάδα και η παραγωγή του ενός επηρεάζει της παραγωγή του άλλου. Αυτή η υπόθεση φαίνεται να έχει βάση καθώς στα ακτινοβακτήρια που μελετήθηκαν, αυτές οι πρωτεΐνες βρίσκονται δίπλα-δίπλα. Αυτό ενισχύει τη θεωρία πως μπορεί να υπάρχουν και άλλα γονίδια που ανήκουν στο ίδιο μονοπάτι βιοσύνθεσης, σχηματίζουν ομάδες και η έκφραση ή όχι του ενός επηρεάζει την έκφραση ή όχι του άλλου.

Επιπλέον, η πολλαπλή στοίχιση στις περιοχές πριν τα γονίδια έδωσε αξιοσημείωτα αποτελέσματα για πιθανές συντηρημένες περιοχές που μπορεί να αποτελούν ρυθμιστικά στοιχεία για την έκφραση και την τελική παραγωγή ή όχι των πρωτεϊνών.

Τέλος, στην κατασκευή του γονιδιώματος, ο ποιοτικός έλεγχος αλλά και η αξιολόγηση φαίνεται να είναι πολύ βασικά βήματα για την πορεία της διαδικασίας καθώς δίνουν χρήσιμες πληροφορίες για την ποιότητα του αποτελέσματος που έχει προκύψει από το κάθε εργαλείο και μπορούν να βοηθήσουν στα επόμενα βήματα της διαδικασίας. Εδώ, τα αποτελέσματα της αξιολόγησης ήταν θετικά σε κάθε βήμα και

υπήρχε σημαντική βελτίωση όσο προχωρούσε η διαδικασία, το κάθε εργαλείο έδινε καλύτερα αποτελέσματα. Επιπλέον, τα 2 γονιδιώματα που κατασκευάστηκαν είναι συγγενικά με *Streptomyces* όπως έδειξαν τα αποτελέσματα των διαδικασιών, οργανισμός που σύμφωνα με τη βιβλιογραφία είναι πλούσιος σε τερπενικές συνθάσες (Dickschat, 2016).

Κεφάλαιο 5ο

5 ΣΥΜΠΕΡΑΣΜΑΤΑ

Συνοψίζοντας, από την παρούσα εργασία και τα αποτελέσματα που παρουσιάστηκαν μπορεί να προκύψει πως τα ακτινοβακτήρια είναι πλούσια στην παραγωγή τερπενίων αν και δεν έχουν μελετηθεί τόσο. Είναι σημαντικό επίσης να αναφερθεί πως περαιτέρω μελέτη σχετικά με τη γονίδια που βρίσκονται κοντά στο γονιδιώματα ενός οργανισμού και ίσως συνεκφράζονται έχει πολύ μεγάλο ενδιαφέρον καθώς ήδη φαίνονται πως κάποια γονίδια συνθέτουν ομάδες γονιδίων και ίσως επηρεάζουν την παραγωγή ή όχι πρωτεϊνών. Τέλος, οι συντηρημένες περιοχές που μπορεί να υπάρχουν πριν από κάθε γονίδιο είναι πιθανό να συμβάλλουν είτε ως ενισχυτές είτε ως αποσιωπητές στην παραγωγή του τελικού προϊόντος και συνεπώς να έχουν βασικό ρόλο στην παραγωγή ή όχι τερπενίων.

5.1 Μελλοντικοί Στόχοι

Τα αποτελέσματα τις εργασίας δημιούργησαν περισσότερα ερωτήματα για επιπλέον αναλύσεις. Στους μελλοντικούς στόχους της μελέτης είναι η πειραματική επιβεβαίωση παραγωγής και έκφρασης τερπενίων σε ακτινοβακτήρια, αλλά και η περαιτέρω ανάλυση των βιοσυνθετικών ομάδων που μπορεί να σχηματίζουν κάποια γονίδια καθώς και το αν εκφράζονται ή όχι στους οργανισμούς, αν παράγουν δηλαδή κάποιο προϊόν και ποια είναι η συλλογική λειτουργία τους. Όπως επίσης και η μελέτη των περιοχών που φαίνεται να είναι συντηρημένες και ίσως αποτελούν ρυθμιστικά στοιχεία των γονιδίων. Επίσης, η μελέτη περισσότερων γονιδιωμάτων ακτινοβακτηριδίων για επιπλέον επιβεβαίωση της μεθόδου είναι ένας από τους μελλοντικούς στόχους. Τέλος, η δοκιμή της μεθόδου σε άλλα γονίδια-στόχους με διαφορετικά γονιδιώματα ενδιαφέροντος θα ήταν μία πρόκληση.

Κεφάλαιο 6ο

6 ΒΙΒΛΙΟΓΡΑΦΙΑ

Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., . . . Walter, P. (2014). *Βασικές Αρχές Κυτταρικής Βιολογίας (ελληνική έκδοση)* (4 ed.). (K. Σταματόπουλος, N. Μοσχονάς, A. Ηλιόπουλος, Π. Παπαζαφείρη , N. Π. Ανάγνου, Eds., K. Σταματόπουλος, A. Γ. Κριεμπάρδης, A. Σ. Παπουτσή, & Σ. Σπυριδωνίδου, Trans.) Broken Hill Publishers LTD.

Alonge, M. a. (2021). Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. *bioRxiv*. doi:10.1101/2021.11.18.469135

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*(3), pp. 403-410. Retrieved from https://doi.org/10.1016/S0022-2836(05)80360-2

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Baunach, M., Franke, J., & Hertweck, C. (2015). Terpenoid Biosynthesis Off the Beaten Track: Unconventional Cyclases and Their Impact on Biomimetic Synthesis. *Angewandte Chemie International Edition, 54*(9), pp. 2604-2626. doi:https://doi.org/10.1002/anie.201407883

Benton, D. (1996). Bioinformatics — principles and potential of a new multidisciplinary tool. *Trends in Biotechnology, 14*(8), pp. 261-272. doi:https://doi.org/10.1016/0167-7799(96)10037-8

Blackshaw, J. (2022). What is bioinformatics and how do we use it? Retrieved from https://www.yourgenome.org/facts/what-is-bioinformatics-and-how-do-we-use-it/

Blum M, C. H.-L.-B. (2020, Nov). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*. doi:10.1093/nar/gkaa977

Britannica, T. E. (2018). *Encyclopedia Britannica.* Retrieved from terpene: https://www.britannica.com/science/terpene

Buchfink, Benjamin, Reuter, Klaus, & Drost, Hajk-Georg. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods, 18*(4). doi:10.1038/s41592-021-01101-x

Buckingham, J. (1997). *Dictionary of natural products, supplement 4* (Vol. 11). CRC press.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*. doi:10.1186/1471-2105-10-421

Carey, M. L. (2013). Software for Computing and Annotating Genomic Ranges. *Computational Biology*(8). doi:10.1371/journal.pcbi.1003118

Charif D, L. J. (2007). *SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. in Structural approaches to sequence evolution: Molecules, networks, populations* (Vols. Biological and Medical Physics, Biomedical Engineering). (U. B. Vendruscolo, Ed.) New York: Springer Verlag.

Consortium, T. U. (2020, Nov). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*. doi:10.1093/nar/gkaa1100

Corporation, M. (2018, Sept 24). Microsoft Excel. *2019 (16.0)*. Retrieved from https://office.microsoft.com/excel

CRICK, F. (1970). Central Dogma of Molecular Biology. *Nature, 227*, pp. 561–563. doi:https://doi.org/10.1038/227561a0

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., . . . Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*(2, giab008). doi:10.1093/gigascience/giab008

DebRoy, H. P. (2022). *Biostrings: Efficient manipulation of biological strings.* Retrieved from https://bioconductor.org/packages/Biostrings

Dickschat, J. S. (2016). Bacterial terpene cyclases. *Nat. Prod. Rep., 33*(1), pp. 87-110. doi:10.1039/C5NP00102A

Diercks, C. S., Dik, D. A., & Schultz, P. G. (2021). Adding new chemistries to the central dogma of molecular biology. *Chem, 7*(11), pp. 2883-2895. doi:https://doi.org/10.1016/j.chempr.2021.09.014

Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics, 14*(9), pp. 755–763. doi:https://doi.org/10.1093/bioinformatics/14.9.755

ENA - European Nucleotide Archive. (n.d.). Retrieved from https://www.ebi.ac.uk/ena/browser/home

Felix Krueger, Frankie James, Phil Ewels, Ebrahim Afyounian, & Benjamin Schuster-Boeckler. (2021, July 23). *FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo (Version 0.6.7). Version 0.6.7. Zenodo*. doi:10.5281/zenodo.5127899

Gao, Y., Honzatko, R. B., & Peters, R. J. (2012). Terpenoid synthase structures: a so far incomplete view of complex catalysis. *Natural product reports, 29*(10), σσ. 1153–1175. doi:https://doi.org/10.1039/c2np20059g

Heidrun Karlic, & Franz Varga. (2019). Mevalonate Pathway. In *Encyclopedia of Cancer (Third Edition)* (pp. 445-457). Academic Press. doi:https://doi.org/10.1016/B978-0-12-801238-3.65000-6

Henry, L. K., Gutensohn, M., Thomas, S. T., Noel, J. P., & Dudareva, N. (2015). Orthologs of the archaeal isopentenyl phosphate kinase regulate terpenoid production in plants. *Proceedings of the National Academy of Sciences, 112*(32), pp. 10050-10055. doi:10.1073/pnas.1504798112

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome research, 17*(3), pp. 377–386. doi:10.1101/gr.5969107

Jones , D., Ormondroyd, G., Curling, S., Popescu, C.-M., & Popescu, M.-C. (2017). 2 - Chemical compositions of natural fibres. Στο *Advanced High Strength Natural Fibre Composites in Construction* (σσ. 23-58). Woodhead Publishing. doi:https://doi.org/10.1016/B978-0-08-100411-1.00002-9

Kampranis, S. C., & Makris, A. M. (2012). Developing a yeast cell factory for the production of terpenoids. *Computational and structural biotechnology journal, 3*, p. e201210006. doi:https://doi.org/10.5936/csbj.201210006

Keck, F. (2020). Handling biological sequences in R with the bioseq package. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.13490

Kong, J., Huh, S., Won, J.-I., Yoon, J., Kim, B., & Kim, K. (2019). GAAP: A Genome Assembly + Annotation Pipeline. *BioMed Research International*. doi:10.1155/2019/4767354

Kuzuyama, T. (2002). Mevalonate and nonmevalonate pathways for the biosynthesis of isoprene units. *Bioscience, biotechnology, and biochemistry, 66*(8), pp. 1619-1627. doi:https://doi.org/10.1271/bbb.66.1619

Lawrence, G. B. (2022). *genbankr: Parsing GenBank files into semantically useful objects.*

Leferink, N. G., & Scrutton, N. S. (2022). Predictive Engineering of Class I Terpene Synthases Using Experimental and Computational Approaches. *ChemBioChem, 23*(5), p. e202100484. doi:https://doi.org/10.1002/cbic.202100484

Li, H. (2018, May). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), pp. 3094-3100. doi:10.1093/bioinformatics/bty191

Lin, H., Liang, Z.-Y., Tang, H., & Chen, W. (2019). Identifying Sigma70 Promoters with Novel Pseudo Nucleotide Composition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16*(4), pp. 1316-1321. doi:10.1109/TCBB.2017.2666141

Madeira, F. a. (2022, April). Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic acids research*. doi:10.1093/nar/gkac240

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021, October). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic,

Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*(10), pp. 4647-4654. doi:10.1093/molbev/msab199

McGarvey, D., & Croteau, R. (1995). Terpenoid metabolism. *The Plant cell, 7*(7), pp. 1015–1026. doi:https://doi.org/10.1105/tpc.7.7.1015

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., . . . Glöckner, F. (2015). Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology, 11*(9), pp. 625–631. doi:10.1038/nchembio.1890

Mikheenko, A. a. (2016, July). Icarus: visualizer for de novo assembly evaluation. *Bioinformatics*(32), pp. 3321-3323. doi:10.1093/bioinformatics/btw379

Mosunova, O., Navarro-Muñoz, J. C., & Collemare, J. (2021). The Biosynthesis of Fungal Secondary Metabolites: From Fundamentals to Biotechnological Applications. In Ó. Z. Casadevall (Ed.), *Encyclopedia of Mycology* (pp. 458-476). Elsevier. doi:https://doi.org/10.1016/B978-0-12-809633-8.21072-8

National Human Genome Research Institute . (2023). Retrieved from GENE EXPRESSION: https://www.genome.gov/genetics-glossary/Gene-Expression

Nurk, Sergey, Bankevich, Anton, Antipov, Dmitry, Gurevich, Alexey, Korobeynikov, Anton, Lapidus, Alla, . . . Pevzner, Pavel A. (2013). Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. In M. Deng, Jiang, Rui, Sun, Fengzhu, & Zhang, Xuegong (Eds.), *Research in Computational Molecular Biology* (pp. 158--170). Berlin, Heidelberg: Springer Berlin Heidelberg.

Pan, J.-J., Solbiati, J., Ramamoorthy, G., Hillerich, B., Seidel, R., Cronan, J., . . . Poulter, C. (2015). Biosynthesis of Squalene from Farnesyl Diphosphate in

Bacteria: Three Steps Catalyzed by Three Enzymes. *ACS Central Science*. doi:10.1021/acscentsci.5b00115

Pevsner, J. (2015). *Βιοπληροφορική και Λειτουργική Γονιδιωματική (ελληνική μετάφραση αγγλικής έκδοσης)* (3 ed.). Wiley Blackwell.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), pp. 257-286. doi:10.1109/5.18626

Rinaldi, M. A., Ferraz, C. A., & Scrutton, N. S. (2022). Alternative metabolic pathways and strategies to high-titre terpenoid production in Escherichia coli. *Nat. Prod. Rep., 39*(1), pp. 90-118. doi:10.1039/D1NP00025J

Robinson, James T, Thorvaldsdóttir, Helga, Winckler, Wendy, Guttman, Mitchell, Lander, Eric S, Getz, Gad, & Mesirov, Jill P. (2011). Integrative genomics viewer. *Nature Biotechnology, 29 (24-26)*. doi:10.1038/nbt.1754

Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, & Kathryn E. Holt. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *Plos Computational Biology*. Retrieved from https://doi.org/10.1371/journal.pcbi.1005595

Scholz, M. (n.d.). *Metagenomics*. Retrieved from E-value & Bit-score: https://www.metagenomics.wiki/tools/blast/evalue

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England), 30*(14). doi:10.1093/bioinformatics/btu153

Stothard, P. (2000). *The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences.* (Vol. BioTechniques 28).

Team, R. C. (2021). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Tholl, D. (2015). Biosynthesis and biological functions of terpenoids in plants. *Advances in biochemical engineering/biotechnology, 148*, pp. 63–106. doi:https://doi.org/10.1007/10_2014_295

Thomas, S. T., Louie, G. V., Lubin, J. W., Lundblad, V., & Noel, J. P. (2019). Substrate Specificity and Engineering of Mevalonate 5-Phosphate Decarboxylase. *ACS chemical biology, 14*(8), pp. 1767-1779. doi:https://doi.org/10.1021/acschembio.9b00322

Tsoka, S., & Ouzounis, C. (2000). Recent developments and future directions in computational genomics. *FEBS letters, 480*(1), pp. 42-48. doi:10.1016/s0014-5793(00)01776-2

Volgin, D. V. (2014). Chapter 17 - Gene Expression: Analysis and Quantitation. In *Animal Biotechnology* (pp. 307-325). Academic Press. doi:https://doi.org/10.1016/B978-0-12-416002-6.00017-1

Waterhouse, A. P. (2009). Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics, 25*(9), pp. 1189-1191. doi:10.1093/bioinformatics/btp033

Wick, R. R. (2015, June). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics, 31*(20), pp. 3350-3352. doi:10.1093/bioinformatics/btv383

Yamada, Y., Kuzuyama, T., Komatsu, M., Shin-Ya, K., Omura, S., Cane, D. E., & Ikeda, H. (2015). Terpene synthases are widely distributed in bacteria. *Proceedings of the National Academy of Sciences of the United States of America, 112*(3), pp. 857–862. doi:https://doi.org/10.1073/pnas.1422108112

Yang, D., Du, X., Liang, X., Han, R., Liang, Z., Liu, Y., . . . Zhao, J. (2012). Different Roles of the Mevalonate and Methylerythritol Phosphate Pathways in Cell Growth and Tanshinone Production of Salvia miltiorrhiza Hairy Roots. *PLOS ONE, 7*(11), pp. 1-9. doi:10.1371/journal.pone.0046797

Zhao, L., Chang, W.-c., Xiao, Y., Liu, H.-w., & Liu, P. (2013). Methylerythritol phosphate pathway of isoprenoid biosynthesis. *Annual review of biochemistry, 82*, pp. 497-530. doi:https://doi.org/10.1146/annurev-biochem-052010-100934

Κεφάλαιο 7ο

7 ΠΑΡΑΡΤΗΜΑ

Στο παρόν κεφάλαιο παρουσιάζονται αναλυτικά οι κώδικες που γράφηκαν, οι εντολές που εκτελέστηκαν στη γραμμή εντολών καθώς και επιπλέον πληροφορίες και αποτελέσματα.

7.1 Κώδικας σε R

7.1.1 Κώδικας 1

sequence extraction- code 1

Extract a specific part of the sequence based on coordinates

```
library(bioseq)

library(seqinr)

library(Biostrings)
```

set the coordinates

```
start <-  start coordinate
end <-  end coordinate
```

path to the genome

```
path <-  'absolute path to FASTA file that contains the DNA sequence
'
```

set the header that needs to be removed

```
seq.header <- "header of the FASTA file that contains the DNA sequen
ce"
```

read the FASTA sequence

```
seq <- read.fasta(path, as.string = TRUE)
```

remove the new line

```
seq <- gsub(pattern = "\\n",
            replacement = "",
            x = seq)
```

remove the header > from the FASTA sequence

```
seq <- gsub(pattern = seq.header,
            replacement = "",
            x = seq)
```

view the first 80 chars of the sequence to make sure it worked

```
strtrim(seq, 80)
```

put each nucleotide of the sequence into its own spot in a vector

```
seq_split <- stringr::str_split(seq,
                                pattern = "",
                                simplify = FALSE)
```

see what type the sequence is

```
str(seq_split)
```

we want to extract the DNA seq from 263381, to 263875 the protein, terpene synthase is found there

```
DNA_protein_seq <- seq_split[start:end]
```

convert the letters a -> A

```
DNA_uppercase_seq <- toupper(DNA_protein_seq)
```

write the final FASTA sequence in a new file

```
write.fasta(DNA_uppercase_seq, names = "dna_seq", file.out = "dna.fa
sta", open = "w")
```

7.1.2 Κώδικας 2

sequence extraction- extended seqs- code 2

Extract a specific part of the sequence based on coordinates

```
library(bioseq)

library(seqinr)

library(Biostrings)
```

set the coordinates

```
start <- start coordinate
end <- end coordinate
```

find seq, 1000 bases upstream and downstream for promoter identification

```
up_1000 <- start - 1000
down_1000 <- end + 1000
```

path to the genome

```
path <-  'absolute path to FASTA file that contains the DNA sequence
'
```

set the header that needs to be removed

```
seq.header <-  "header of the FASTA file that contains the DNA seque
nce"
```

read the FASTA sequence

```
seq <- read.fasta(path, as.string = TRUE)
```

remove the new line

```
seq <- gsub(pattern = "\\n",
          #  replacement = "",
           # x = seq)
```

remove the header > from the FASTA sequence

```
seq <- gsub(pattern = seq.header,
            replacement = "",
            x = seq)
```

view the first 80 chars of the sequence to make sure it worked

```
strtrim(seq, 80)
```

put each nucleotide of the sequence into its own spot in a vector

```
seq_split <- stringr::str_split(seq,
                                pattern = "",
                                simplify = FALSE)
```

see what type the sequence is

```
str(seq_split)
```

we want to extract the DNA seq from 263381, to 263875 the protein, terpene synthase is found there

```
DNA_protein_seq <- seq_split[start:end]
DNA_extend_1000_seq <- seq_split[up_1000 : down_1000]
```

convert the letters a -> A

```
DNA_uppercase_seq <- toupper(DNA_protein_seq)
DNA_uppercase_seq_1000 <- toupper(DNA_extend_1000_seq)
```

write the final FASTA sequence in a new file

```
write.fasta(DNA_uppercase_seq, names = "dna_seq", file.out = "dna.fa
sta", open = "w")
write.fasta(DNA_uppercase_seq_1000, names = "dna_ext_seq_1000", file
.out = "dna_ext_1000.fasta", open = "w")
```

7.1.3 Κώδικας 3

gene extraction- code 3

gene extraction from the GenBank file

```
library(genbankr)

library(Biostrings)

library(GenomicFeatures)
```

set the path to the Genank file

```
IMCC26256 <-  "path/filename.gb"
```

read the GenBank file

```
IMCC26256_gb <- genbankr::readGenBank(IMCC26256)
```

store the genes from the GenBank file

```
GENES_IMCC26256 <- GenomicFeatures::genes(IMCC26256_gb)
```

create a dataframe of genes and further information

```
GenesDF_IMCC26256 <- data.frame(GENES_IMCC26256)
```

get the sequences

```
gene_seq <- getSeq(IMCC26256_gb@sequence, setNames(GENES_IMCC26256,
GENES_IMCC26256$gene_id))
```

write all sequences in FASTA file

```
writeXStringSet(gene_seq, "genes_IMCC26256.fa")
```

create a CSV file with information about genes, coordinates,strand, etc.

```
write.csv(GenesDF_IMCC26256,'GenesDF_IMCC26256.csv')
```

Same for the other actinobacteria

```
IMCC26207 <-  "path/filename.gb"

IMCC26207_gb <- genbankr::readGenBank(IMCC26207)
GENES_IMCC26207 <- GenomicFeatures::genes(IMCC26207_gb)
GenesDF_IMCC26207 <- data.frame(GENES_IMCC26207)


res <- getSeq(IMCC26207_gb@sequence, setNames(GENES_IMCC26207, GENES
_IMCC26207$gene_id))
writeXStringSet(res, "genes_IMCC26207.fa")
write.csv(GenesDF_IMCC26207,'GenesDF_IMCC26207.csv')

OV320 <-  "path/filename.gb"
```

```
OV320_gb <- genbankr::readGenBank(OV320)
GENES_OV320 <- GenomicFeatures::genes(OV320_gb)
GenesDF_OV320 <- data.frame(GENES_OV320)


res <- getSeq(OV320_gb@sequence, setNames(GENES_OV320, GENES_OV320$g
ene_id))
writeXStringSet(res, "genes_OV320.fa")
write.csv(GenesDF_OV320,'GenesDF_OV320.csv')


OK074 <-  "path/filename.gb"

OK074_gb <- genbankr::readGenBank(OK074)
GENES_OK074 <- GenomicFeatures::genes(OK074_gb)
GenesDF_OK074 <- data.frame(GENES_OK074)


res <- getSeq(OK074_gb@sequence, setNames(GENES_OK074, GENES_OK074$g
ene_id))
writeXStringSet(res, "genes_OK074.fa")
write.csv(GenesDF_OK074,'GenesDF_OK074.csv')
```

7.1.4 Κώδικας 4

Heatmap – code 4

This part of code is for producing a heatmap to see which protein exists in the actinobacteria we wanted to study

Provide the appropriate library

```
library("heatmaply")
```

Read the CSV files from the absolute path

```
proteinXgenome <- read.csv(('{absolute-path-to-file/proteinxgenome.csv'), header = TRUE)
```

If there are NA values, convert them to 0

```
proteinXgenome[is.na(proteinXgenome)] <- 0
```

Keep the part of data that contain the values as a matrix

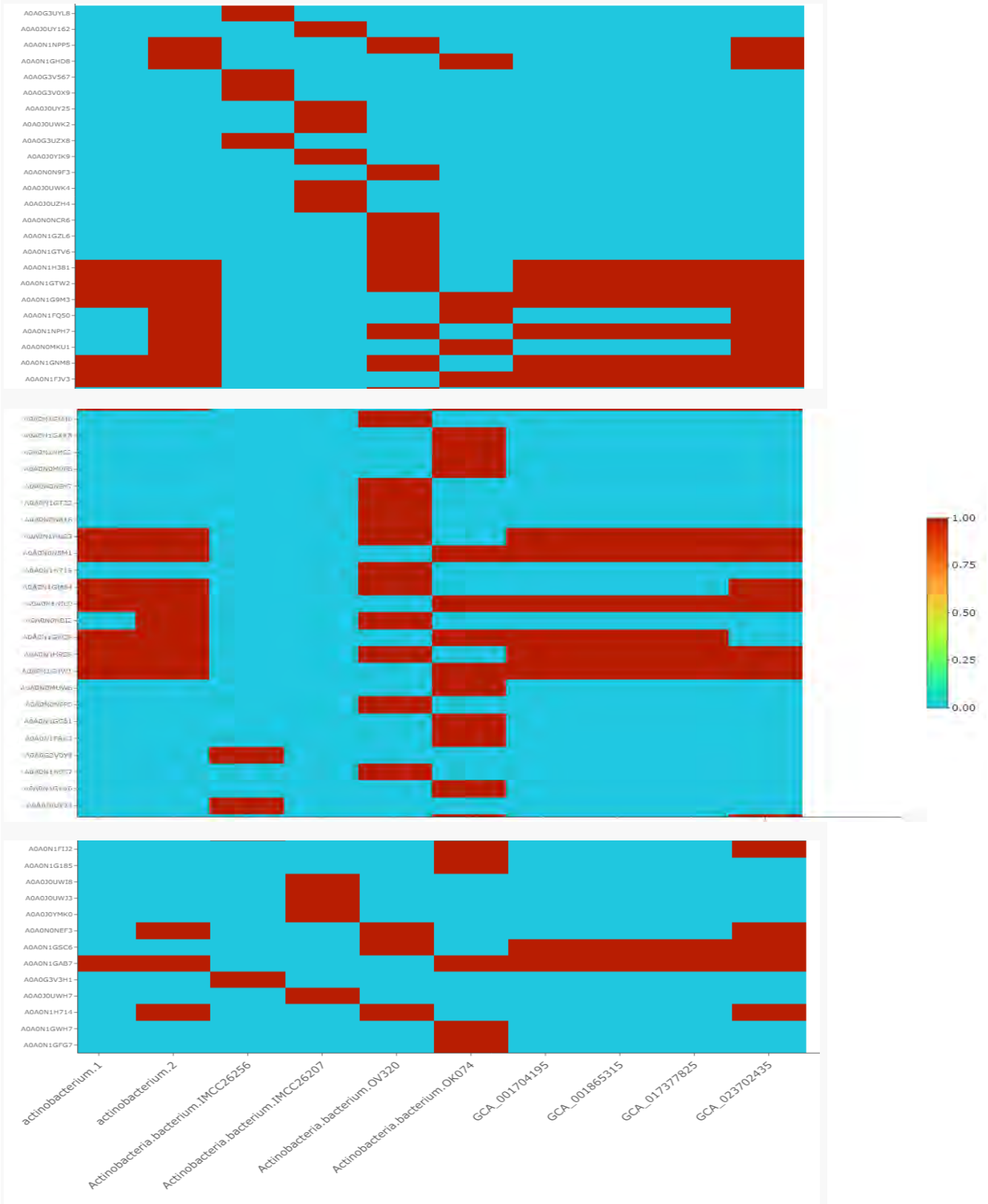
```
data <- as.matrix(proteinXgenome[, c(3:8)])
```

As row names keep the code of protein

```
row.names <- proteinXgenome[ , 2]
rownames(data) <- row.names
```

Make the heatmap

```
heatmaply(data, Rowv = FALSE, Colv = FALSE, fontsize_row = 6, colors
= viridis(n = 256, begin = 0.27, end = 0.92, option = "turbo"))
```



7.2 Κώδικας στη γραμμή εντολών

Assembly and Annotation methodology

The commands are written without providing an absolute path for tools and files, but it is required for the execution.

Preperation of raw data

First of all we need to run a Quality Control using the tool FastQC for our raw reads data. We have data R1.fastq for forward strand and R2.fastq for reverse strand We run the command with a * for every fastq file in our path

```
$ fastqc *.fastq
```

Based on the QCreport files that FastQC produces we have to deside where trim our data and how to clean them up. If there are adaptors, low quality reads, gaps(Ns) or short reads, we have to clean them. We will use the tool trim-galore.

```
$ trim-galore --quality 30 --length 40 --trim-n --paired --three_pri
me_clip_R2 50 R1.fastq R2.fastq
```

The we will re-run the FastQC tool for our trimmed data to check the quality again. If the quality is not satisfying we have to change our attributes and run again until we have a proper set of data.

Assembly and Evaluation

Now we are ready to procede to the Assembly step. We will use a tool called SPAdes because it's suitable for bacterial genomes.

```
$ spades.py -1 trimmed_R1.fq –2 trimmed_R2.fq --careful --memory 80
--threads 48
```

And now we have a FASTA file that contains contigs.

Then we have to evaluate our assembly and for this we will use 2 tools. QUAST, a Quality Assessment tool that evaluates a genome assembly with or without a reference genome and BUSCO which provides measures for quantitative assessment of genome assembly completeness

```
# quast
$ quast.py contigs.fasta --output quast_results
# busco
$ busco --in contigs.fasta --out busco_results --mode genome --out_p
ath results
```

We can also align the FASTA file to a reference genome (found with BLAST), closest to ours, to check for gaps or overlaps. We will use the tool minimap2 for the alignment and then we will visualize it using IGV locally on our desktop.

```
$ minimap2 -a -o FILENAME -x asm10 reference_genome.fasta contigs.fa
sta
```

For importing the data to IGV we have to sort the results from minimap2 and make an index

```
$ samtools sort mapping_result_file > sorted_file.bam
$ samtools index sorted_file.bam
```

After this we may observe a low quality assembly, or many gaps or overlaps, so we may consider improve our assembly or producve another one. We can use the Unicycler assembler which is apropriate for bacterial genomes and produce another assembly. Unicycler uses SPAdes in its pipeline.

```
$ unicycler –1 trimmed_R1.fq -2 trimmed_R2.fq -o output_dir --keep 1
--threads 32 --mode bold --spades_options '-m 80000'
```

To evaluate, we can run QUAST and BUSCO again and we can import our results locally in bandage to see the connection between the nodes.

If we want to improve our assembly, we can run a tool called RagTag which is a collection of software tools for scaffolding and improving genome assemblies. We can run it in our data from unicycler assembly because it gave better results than SPAdes. We will use the commands correct, which uses a reference genome to identify and corect potential misassmeblies, and scaffold, which ordering and orienting draft assembly sequence into longer and joingn gaps without altering the sequence

```
$ ragtag.py correct reference_genome.fasta unicycler_assembly.fasta
-o ragtag_results_correct
$ ragtag.py scaffold reference_genome.fasta ragtag_result_correct/ra
gtag.correct.file.fasta -o ragtag_results_scaffold
```

We do not use the other commands of ragtag because we may alter the sequence based on the reference genome and lose important differences.

Again, we evaluate using QUAST and BUSCO, and we decide which assembly is better based on the metrics produced, number of contigs, genome completeness, etc.

Annotation

Now we have our constructed genome, and we are ready to continue with the annotation step. We will use PROKKA, a software tool for rapidly annotate genes and identify coding seqs in prokaryotic genomes

```
$ prokka ragtag.scaffold.fasta --outdir prokka_annotation_dir
```

We can navigate through the files and check the TSV file produced, to see clusters of specific genes in our interest.

If we want to see a phylogenetic tree, with organisms close to our we can align our trimmed data to a database such as nr with DIAMOND tool and then use MEGAN locally to observe the phylogenetic trees.

```
$ diamond blastx -p 60 -d path_to_nr -f 100 -k 1 --max-hsps 1 --stra
nd both -e 10E-5 --outfmt 100 -q trimmed_R1.fq -o path_to_output_fil
e
```

Then, we will use MEGAN locally.
We will use meganizer for our DAA file produced by DIAMOND and then import it in MEGAN to see the phylogenetic trees.

Searching for terpene synthases

For the last step of finding the terpene synthases in the genomes either produced by assembly or in complete genomes we have to use command line BLAST tool. We have to construct a database for the assembled genome from ragtag

```
$ makeblastdb -in ragtag.scaffold.fasta -dbtype nucl -out path_to_ou
tput_database
```

And then we have to align the FASTA file containing all the sequences for genes related to tepene synthases (work done in Desktop and described in thesis) to the database we just constructed

```
# text format
$ blastn -db path_to_output_database -query file_with_genes.fasta -o
ut filename.txt
# tabular format
$ blastn -db path_to_output_database -query file_with_genes.fasta -o
utfmt 7 -out filename_tabular
```

And now that we have our data and our results, we can continue our analysis in the desktop!

7.3 Γονίδια Στόχοι με τους αντίστοιχους κωδικούς στην InterPro

2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	A0A0G3UYL8
	A0A0J0UY162
	A0A0N1NPP5
	A0A0N1GHD8
geranylgeranyl pyrophosphate synthase	A0A0G3V567
	A0A0G3V0X9
	A0A0J0UY25
	A0A0J0UWK2
Prenyltransferase	A0A0G3UZX8
	A0A0J0YIK9
	A0A0N0N9F3
phytoene/squalene synthetase	A0A0J0UWK4
Phytanoyl-CoA dioxygenase	A0A0J0UZH4
Terpene synthase metal-binding domain-containing protein	A0A0N0NCR6
Terpene synthase	A0A0N1GZL6
Dimethylallyltranstransferase	A0A0N1GTV6
	A0A0N1H381
	A0A0N1GTW2
	A0A0N1G9M3
	A0A0N1FQ50
Epi-isozizaene synthase	A0A0N1NPH7
	A0A0N0MKU1
Germacradienol synthase	A0A0N1GNM8
	A0A0N1FJV3
Polyprenyl synthetase	A0A0N1GU40
	A0A0N1GXK8
	A0A0N1NHS2
	A0A0N0MVF6
Geranyltranstransferase	A0A0N0NEY7
	A0A0N1GT32
	A0A0N0N816
Trans-hexaprenyltranstransferase	A0A0N1NU03
	A0A0N0N5M1
SQHop_cyclase_C domain-containing protein	A0A0N1H719
squalene synthase HpnD	A0A0N1GM64
	A0A0N1NIL0
squalene synthase HpnC	A0A0N0NBI2
	A0A0N1GKQ9
squalene-hopene cyclase	A0A0N1H9Z9
	A0A0N1G4W1
	A0A0N0MUW6
Prenyltransferase/squalene oxidase-like repeat protein	A0A0N0NFP0
	A0A0N1GC61
Transcriptional regulator	A0A0N1FRK3

2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	A0A0G3V0Y9
	A0A0N1H057
	A0A0N1G1V0
hypothetical protein *note PFAM: Antibiotic biosynthesis monooxygenase	A0A0J0UY24
A0A0N1FIJ2-Unspecific monooxygenase *note PFAM: cytochrome P450 074	A0A0N1FIJ2
monooxygenase FAD-binding protein	A0A0N1G185
phytoene desaturas	A0A0J0UWI8
isopentenyl-diphosphate delta-isomerase	A0A0J0UWJ3
protein involved in biosynthesis of mitomycin antibiotics/polyketide fumonisin *note PFAM: Phytanoyl-CoA dioxygenase (PhyH)	A0A0J0YMK0
NADPH-dependent FMN reductase	A0A0N0NEF3
squalene-associated FAD-dependent desaturase	A0A0N1GSC6
	A0A0N1GAB7
SCP-2 sterol transfer family protein	A0A0G3V3H1
fatty acid hydroxylase-like protein	A0A0J0UWH7
Unspecific monooxygenase	A0A0N1H714
UbiA prenyltransferase	A0A0N1GWH7
	A0A0N1GFG7

7.4 Πίνακες Αναλυτικών Αποτελεσμάτων

7.4.1 Ακτινοβακτήριο 1

contig	% identity	alignment length	mismatches	gap opens	gene start coordinate	gene end coordinate	start coordinate in genomes	end coordinate in genome	evalue	bit score	product	protein name in InterPro
CP029377.1_RagTag	86.139	101	14	0	972	1072	2039991	2039891	5.14e-24	110	Epi-isozizaene	A0A0N1NPH7
CP029377.1_RagTag	77.585	1035	188	35	89	1104	2010030	2009021	1.97e-167	586	Dimethylallyltranstransferase	A0A0N1H381
CP029377.1_RagTag	80.428	981	170	18	21	989	1707730	1708700	0.0	728	Germacradienol synthase	A0A0N1GNM8
CP029377.1_RagTag	83.645	1015	158	8	1	1011	3931262	3932272	0.0	948	MULTISPECIES: polyprenyl synthetase family protein [unclassified Streptomyces]	A0A0N1NU03
CP029377.1_RagTag	84.732	1120	152	16	26	1140	655006	653901	0.0	1103	Dimethylallyltranstransferase	A0A0N1GTW2
CP029377.1_RagTag	77.876	226	46	4	652	875	6232444	6232667	2.45e-32	137	Dimethylallyltranstransferase	A0A0N1GTW2
CP029377.1_RagTag	82.545	888	145	9	35	917	657330	656448	0.0	773	squalene synthase	A0A0N1GM64
CP029377.1_RagTag	82.311	1826	287	32	158	1966	653536	651730	0.0	1550	squalene-hopene cyclase	A0A0N1H9Z9
CP029377.1_RagTag	81.682	999	161	16	6	992	1707712	1708700	0.0	811	Germacradienol synthase	A0A0N1FJV3
CP029377.1_RagTag	84.539	1009	152	4	1	1007	3931262	3932268	0.0	996	Trans-hexaprenyltranstransferase	A0A0N0N5M1
CP029377.1_RagTag	78.477	906	162	31	19	903	658254	657361	2.69e-160	562	squalene synthase	A0A0N1GKQ9
CP029377.1_RagTag	83.69	889	133	12	35	917	657330	656448	0.0	828	squalene synthase	A0A0N1NIL0
CP029377.1_RagTag	86.945	1103	126	15	41	1137	654991	653901	0.0	1223	Dimethylallyltranstransferase	A0A0N1G9M3
CP029377.1_RagTag	82.432	1867	292	28	111	1959	653592	651744	0.0	1598	squalene-hopene cyclase	A0A0N1G4W1
CP029377.1_RagTag	76.749	1372	255	56	61	1403	656390	655054	0.0	708	squalene-associated FAD-dependent desaturase	A0A0N1GSC6
CP029377.1_RagTag	78.034	1343	244	43	79	1400	656384	655072	0.0	798	squalene-associated FAD-dependent desaturase	A0A0N1GAB7

7.4.2 Ακτινοβακτήριο 2

contig	% identity	alignment length	mismatches	gap opens	gene start coordinate	gene end coordinate	start coordinate in genomes	end coordinate in genome	evalue	bit score	protein	protein name in InterPr
CP094918.1_RagTag	82108	1034	173	12	71	1098	1178269	1179296	0.0	874	Epi-isozizaene synthase	A0A0N1NPH7
CP094918.1_RagTag	81391	1021	178	12	91	1105	1201759	1202773	0.0	822	Dimethylallyltranstransferase	A0A0N1H381
CP094918.1_RagTag	87884	1073	124	6	1	1070	2134497	2135566	0.0	1256	Germacradienol synthase	A0A0N1GNM8
CP094918.1_RagTag	88106	454	54	0	31	484	7070882	7070429	7.90e-154	540	2-C-methyl-D-erythritol 2,4-cyclodiphosphate	A0A0N1NPP5
CP094918.1_RagTag	86601	1015	128	8	1	1011	560278	561288	0.0	1114	MULTISPECIES: polyprenyl synthetase family protein [unclassified	A0A0N1NU03
CP094918.1_RagTag	88632	1126	125	3	15	1140	2775488	2776610	0.0	1367	Dimethylallyltranstransferase	A0A0N1GTW2
CP094918.1_RagTag	86378	947	121	8	1	943	2772978	2773920	0.0	1027	squalene synthase HpnD	A0A0N1GM64
CP094918.1_RagTag	83025	919	142	10	1	912	2772070	2772981	0.0	821	squalene synthase HpnC	A0A0N0NBI2
CP094918.1_RagTag	86146	1985	250	21	1	1971	2776727	2778700	0.0	2119	squalene-hopene cyclase	A0A0N1H9Z9
CP094918.1_RagTag	82571	2077	326	32	9	2067	2134502	2136560	0.0	1797	Germacradienol synthase	A0A0N1FJV3
CP094918.1_RagTag	79962	1048	196	13	46	1086	1178256	1179296	0.0	760	Epi-isozizaene synthase	A0A0N0MKU1
CP094918.1_RagTag	87364	459	54	4	61	517	7070885	7070429	8.50e-149	523	2-C-methyl-D-erythritol 2,4-cyclodiphosphate	A0A0N1GHD8
CP094918.1_RagTag	87786	1007	117	6	1	1004	560278	561281	0.0	1173	Trans-hexaprenyltranstransferas	A0A0N0N5M1
CP094918.1_RagTag	84447	868	126	4	45	903	2772114	2772981	0.0	846	squalene synthase HpnC	A0A0N1GKQ9
CP094918.1_RagTag	86868	929	120	2	1	928	2772978	2773905	0.0	1038	squalene synthase HpnD	A0A0N1NIL0
CP094918.1_RagTag	88759	1112	121	4	28	1137	2775501	2776610	0.0	1358	Dimethylallyltranstransferase	A0A0N1G9M3
CP094918.1_RagTag	85262	1988	252	33	1	1969	2776727	2778692	0.0	2010	squalene-hopene cyclase	A0A0N1G4W1
CP094918.1_RagTag	81099	1037	174	18	175	1203	1201759	1202781	0.0	809	Dimethylallyltranstransferase	A0A0N1FQ50
CP094918.1_RagTag	82992	1317	209	12	52	1359	1179338	1180648	0.0	1177	Unspecific	A0A0N1H714
CP094918.1_RagTag	84882	549	72	11	1	545	565245	564704	7.25e-155	544	NADPH-dependent FMN reductase	A0A0N0NEF3
CP094918.1_RagTag	85255	1370	192	10	58	1422	2774110	2775474	0.0	1402	squalene-associated FAD-dependent desaturase	A0A0N1GSC6
CP094918.1_RagTag	82963	1350	204	24	63	1401	2774104	2775438	0.0	1195	squalene-associated FAD-dependent desaturase	A0A0N1GAB7

7.4.3 GCA_001704195

contig	% identity	alignment length	mismatches	gap opens	gene start coordinate	gene end coordinate	start coordinate in genomes	end coordinate in genome	evalue	bit score	protein	protein name in InterPro
ENA|CP016824|CP016824.1	72.871	1010	224	40	97	1072	5022094	5023087	8.56E-82	302	Epi-isozizaene synthase	A0A0N1NPH7
ENA|CP016824|CP016824.1	77.444	1033	189	36	91	1104	5052941	5053948	1.29E-164	577	Dimethylallyltranstransferase	A0A0N1H381
ENA|CP016824|CP016824.1	80.448	982	168	19	21	989	5357783	5356813	0	728	Germacradienol synthase	A0A0N1GNM8
ENA|CP016824|CP016824.1	83.448	1015	160	8	1	1011	2799853	2798843	0	937	Trans-hexaprenyltranstransferase	A0A0N1NU03
ENA|CP016824|CP016824.1	84.492	1122	151	18	26	1140	6526819	6527924	0	1086	Dimethylallyltranstransferase	A0A0N1GTW2
ENA|CP016824|CP016824.1	83.459	133	20	2	747	878	373601	373470	7.43E-28	122		
ENA|CP016824|CP016824.1	82.262	1821	297	24	158	1966	6528289	6530095	0	1550	squalene-hopene cyclase	A0A0N1H9Z9
ENA|CP016824|CP016824.1	81.001	979	168	15	24	992	5357783	5356813	0	761	Germacradienol synthase	A0A0N1FJV3
ENA|CP016824|CP016824.1	74.297	782	166	26	1226	1988	5356516	5355751	2.23E-80	298		
ENA|CP016824|CP016824.1	84.325	1008	156	2	1	1007	2799853	2798847	0	985	Trans-hexaprenyltranstransferase	A0A0N0N5M1
ENA|CP016824|CP016824.1	78.429	904	166	28	19	903	6523560	6524453	2.91E-160	562	squalene synthase HpnC	A0A0N1GKQ9
ENA|CP016824|CP016824.1	82.958	933	133	23	1	917	6524450	6525372	0	819	squalene synthase HpnD	A0A0N1NIL0
ENA|CP016824|CP016824.1	86.57	1102	132	13	41	1137	6526834	6527924	0	1201	Dimethylallyltranstransferase	A0A0N1G9M3
ENA|CP016824|CP016824.1	82.727	1870	289	29	107	1959	6528229	6530081	0	1633	squalene-hopene cyclase	A0A0N1G4W1
ENA|CP016824|CP016824.1	77.241	1305	252	41	58	1340	6525427	6526708	0	723	squalene-associated FAD-dependent desaturase	A0A0N1GSC6
ENA|CP016824|CP016824.1	77.879	1320	256	33	52	1355	6525409	6526708	0	787	squalene-associated FAD-dependent desaturase	A0A0N1GAB7

7.4.4 GCA_001865315

protein coding name	contig	% identity	alignment length	mismatches	gap opens	gene start coordinate	gene end coordinate	start coordinate in genomes	end coordinate in genome	evalue	bit score	protein	protein name in InterPro
dna_seq_OV320_frag29_4	ENA|KV861362|KV861362.1	72.871	1010	224	40	97	1072	5007767	5008760	1.75E-81	302	Epi-isozizaene synthase	A0A0N1NPH7
dna_seq_OV320_frag29_4	ENA|LORI01000411|LORI01000411.1	72.871	1010	224	40	97	1072	11276	12269	1.75E-81	302		
dna_seq_OV320_frag29_5	ENA|KV861362|KV861362.1	77.541	1033	188	36	91	1104	5038833	5039840	5.66E-166	582	Dimethylallyltranstransferase	A0A0N1H381
dna_seq_OV320_frag29_5	ENA|LORI01000413|LORI01000413.1	77.541	1033	188	36	91	1104	16333	17340	5.66E-166	582		
dna_seq_OV320_frag30_6	ENA|KV861362|KV861362.1	80.448	982	168	19	21	989	5347007	5346037	0	728	Germacradienol synthase	A0A0N1GNM8
dna_seq_OV320_frag30_6	ENA|LORI01000433|LORI01000433.1	80.448	982	168	19	21	989	2627	1657	0	728		
dna_seq_OV320_frag19_12	ENA|KV861362|KV861362.1	83.448	1015	160	8	1	1011	2748700	2747690	0	937	Trans-hexaprenyltranstransferase	A0A0N1NU03
dna_seq_OV320_frag19_12	ENA|LORI01000231|LORI01000231.1	83.448	1015	160	8	1	1011	13099	12089	0	937		
dna_seq_OV320_frag27_13	ENA|KV861362|KV861362.1	84.492	1122	151	18	26	1140	6528545	6529650	0	1086	Dimethylallyltranstransferase	A0A0N1GTW2
dna_seq_OV320_frag27_13	ENA|KV861362|KV861362.1	83.459	133	20	2	747	878	275796	275665	1.52E-27	122		
dna_seq_OV320_frag27_13	ENA|LORI01000512|LORI01000512.1	84.492	1122	151	18	26	1140	29476	30581	0	1086		
dna_seq_OV320_frag27_13	ENA|LORI01000032|LORI01000032.1	83.459	133	20	2	747	878	8317	8186	1.52E-27	122		
dna_seq_OV320_frag27_16	ENA|KV861362|KV861362.1	82.262	1821	297	24	158	1966	6530015	6531821	0	1550	squalene-hopene cyclase	A0A0N1H9Z9
dna_seq_OV320_frag27_16	ENA|LORI01000512|LORI01000512.1	82.262	1821	297	24	158	1966	30946	32752	0	1550		
dna_seq_OK074_frag299_1	ENA|KV861362|KV861362.1	81.001	979	168	15	24	992	5347007	5346037	0	761	Germacradienol synthase	A0A0N1FJV3
dna_seq_OK074_frag299_1	ENA|KV861362|KV861362.1	74.297	782	166	26	1226	1988	5345740	5344975	4.56E-80	298		
dna_seq_OK074_frag299_1	ENA|LORI01000433|LORI01000433.1	81.001	979	168	15	24	992	2627	1657	0	761		
dna_seq_OK074_frag299_1	ENA|LORI01000433|LORI01000433.1	74.297	782	166	26	1226	1988	1360	595	4.56E-80	298		
dna_seq_OK074_frag41_5	ENA|KV861362|KV861362.1	84.325	1008	156	2	1	1007	2748700	2747694	0	985	Trans-hexaprenyltranstransferase	A0A0N0N5M1
dna_seq_OK074_frag41_5	ENA|LORI01000231|LORI01000231.1	84.325	1008	156	2	1	1007	13099	12093	0	985		
dna_seq_OK074_frag128_6	ENA|KV861362|KV861362.1	78.429	904	166	28	19	903	6525291	6526184	5.95E-160	562	squalene synthase HpnC	A0A0N1GKQ9
dna_seq_OK074_frag128_6	ENA|LORI01000512|LORI01000512.1	78.429	904	166	28	19	903	26222	27115	5.95E-160	562		
dna_seq_OK074_frag128_7	ENA|KV861362|KV861362.1	82.851	933	134	23	1	917	6526181	6527103	0	813	squalene synthase HpnD	A0A0N1NIL0
dna_seq_OK074_frag128_7	ENA|LORI01000512|LORI01000512.1	82.851	933	134	23	1	917	27112	28034	0	813		
dna_seq_OK074_frag128_8	ENA|KV861362|KV861362.1	86.57	1102	132	13	41	1137	6528560	6529650	0	1201	Dimethylallyltranstransferase	A0A0N1G9M3
dna_seq_OK074_frag128_8	ENA|LORI01000512|LORI01000512.1	86.57	1102	132	13	41	1137	29491	30581	0	1201		
dna_seq_OK074_frag128_9	ENA|KV861362|KV861362.1	82.727	1870	289	29	107	1959	6529955	6531807	0	1633	squalene-hopene cyclase	A0A0N1G4W1
dna_seq_OK074_frag128_9	ENA|LORI01000512|LORI01000512.1	82.727	1870	289	29	107	1959	30886	32738	0	1633		
OV320_5724_A0A0N1GSC6	ENA|KV861362|KV861362.1	77.343	1302	251	40	58	1340	6527158	6528434	0	730	squalene-associated FAD-dependent desaturase	A0A0N1GSC6
OV320_5724_A0A0N1GSC6	ENA|LORI01000512|LORI01000512.1	77.343	1302	251	40	58	1340	28089	29365	0	730		
OK074_8671_A0A0N1GAB7	ENA|KV861362|KV861362.1	77.875	1365	261	35	52	1400	6527140	6528479	0	809	squalene-associated FAD-dependent desaturase	A0A0N1GAB7
OK074_8671_A0A0N1GAB7	ENA|LORI01000512|LORI01000512.1	77.875	1365	261	35	52	1400	28071	29410	0	809		

7.4.5 GCA_017377825

contig	% identity	alignment length	mismatches	gap opens	gene start coordinate	gene end coordinate	start coordinate in genomes	end coordinate in genome	evalue	bit score	protein	protein name in InterP
ENA|JAFRUC010000004|JAFRUC010000004.1	72.944	1009	225	41	97	1072	189751	188758	1.85E-83	307	Epi-isozizaene synthase	A0A0N1NPH7
ENA|JAFRUC010000004|JAFRUC010000004.1	77.67	1030	180	39	97	1104	158811	157810	2.79E-166	582	Dimethylallyltranstransferase	A0A0N1H381
ENA|JAFRUC010000055|JAFRUC010000055.1	80.918	980	165	17	21	988	16278	17247	0	754	Germacradienol synthase	A0A0N1GNM8
ENA|JAFRUC010000001|JAFRUC010000001.1	83.251	1015	162	8	1	1011	537759	536749	0	926	Trans-hexaprenyltranstransferase	A0A0N1NU03
ENA|JAFRUC010000001|JAFRUC010000001.1	84.684	1123	147	20	26	1140	261993	260888	0	1098	Dimethylallyltranstransferase	A0A0N1GTW2
ENA|JAFRUC010000007|JAFRUC010000007.1	84.496	129	18	2	748	875	59601	59728	5.78E-29	126		
ENA|JAFRUC010000001|JAFRUC010000001.1	82.317	1821	296	25	158	1966	260523	258717	0	1555	squalene-hopene cyclase	A0A0N1H9Z9
ENA|JAFRUC010000055|JAFRUC010000055.1	81.325	996	168	13	6	991	16260	17247	0	793	Germacradienol synthase	A0A0N1FJV3
ENA|JAFRUC010000055|JAFRUC010000055.1	74.169	782	167	28	1226	1988	17545	18310	1.04E-78	292		
ENA|JAFRUC010000001|JAFRUC010000001.1	84.127	1008	158	2	1	1007	537759	536753	0	974	Trans-hexaprenyltranstransferase	A0A0N0N5M1
ENA|JAFRUC010000001|JAFRUC010000001.1	77.987	904	170	28	19	903	265242	264349	1.37E-153	540	squalene synthase HpnC	A0A0N1GKQ9
ENA|JAFRUC010000001|JAFRUC010000001.1	83.296	898	126	21	35	917	264318	263430	0	806	squalene synthase HpnD	A0A0N1NIL0
ENA|JAFRUC010000001|JAFRUC010000001.1	86.854	1103	127	15	41	1137	261978	260888	0	1218	Dimethylallyltranstransferase	A0A0N1G9M3
ENA|JAFRUC010000001|JAFRUC010000001.1	82.674	1870	290	29	107	1959	260583	258731	0	1628	squalene-hopene cyclase	A0A0N1G4W1
ENA|JAFRUC010000001|JAFRUC010000001.1	76.94	1366	263	45	61	1403	263372	262036	0	730	squalene-associated FAD-dependent desaturase	A0A0N1GSC6
ENA|JAFRUC010000001|JAFRUC010000001.1	78.102	1338	252	35	79	1400	263366	262054	0	809	squalene-associated FAD-dependent desaturase	A0A0N1GAB7

7.4.6 GCA_023702435

contig	% identity	alignment length	mismatches	gap opens	gene start coordinate	gene end coordinate	start coordinate in genomes	end coordinate in genome	evalue	bit score	protein	protein name in InterPro
JAMQBH010000003.1	81.966	1048	177	12	57	1098	309583	310624	0	878	Epi-isozizaene synthase	A0A0N1NPH7
JAMQBH010000003.1	81.489	1021	177	12	91	1105	333084	334098	0	828	Dimethylallyltran stransferase	A0A0N1H381
JAMQBH010000001.1	87.907	1075	120	9	1	1070	440367	441436	0	1256	Germacradienol synthase	A0A0N1GNM8
JAMQBH010000005.1	88.72	461	51	1	25	484	102875	103335	1.72E-160	562	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	A0A0N1NPP5
JAMQBH010000016.1	86.969	1013	128	4	1	1011	12638	11628	0	1136	Trans-hexaprenyltranstr ansferase	A0A0N1NU03
JAMQBH010000043.1	88.789	1115	118	7	28	1140	5154	4045	0	1360	Dimethylallyltran stransferase	A0A0N1GTW2
JAMQBH010000043.1	86.272	947	122	8	1	943	7677	6735	0	1022	squalene synthase HpnD	A0A0N1GM64
JAMQBH010000043.1	86.247	1985	248	21	1	1971	3928	1955	0	2130	squalene-hopene cyclase	A0A0N1H9Z9
JAMQBH010000001.1	82.427	2077	329	32	9	2067	440372	442430	0	1781	Germacradienol synthase	A0A0N1FJV3
JAMQBH010000003.1	79.866	1048	197	13	46	1086	309584	310624	0	754	Epi-isozizaene synthase	A0A0N0MKU1
JAMQBH010000005.1	88.121	463	50	5	58	517	102875	103335	1.85E-155	545	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	A0A0N1GHD8
JAMQBH010000016.1	87.921	1010	116	6	1	1007	12638	11632	0	1184	Trans-hexaprenyltranstr ansferase	A0A0N0N5M1
JAMQBH010000043.1	86.975	929	119	2	1	928	7677	6750	0	1044	squalene synthase HpnD	A0A0N1NIL0
JAMQBH010000043.1	88.52	1115	124	4	25	1137	5157	4045	0	1347	Dimethylallyltran stransferase	A0A0N1G9M3
JAMQBH010000043.1	85.204	1987	255	30	1	1969	3928	1963	0	2004	squalene-hopene cyclase	A0A0N1G4W1
JAMQBH010000003.1	81.292	1037	172	18	175	1203	333084	334106	0	821	Dimethylallyltran stransferase	A0A0N1FQ50
JAMQBH010000003.1	82.992	1317	209	12	52	1359	310666	311976	0	1177	Unspecific monooxygenase	A0A0N1H714
JAMQBH010000016.1	84.787	539	76	6	10	545	7676	8211	1.23E-152	536	NADPH-dependent FMN reductase	A0A0N0NEF3
JAMQBH010000043.1	84.964	1370	196	10	58	1422	6545	5181	0	1380	squalene-associated FAD-dependent desaturase	A0A0N1GSC6
JAMQBH010000003.1	76.169	1326	292	22	102	1413	310661	311976	0	676	Unspecific monooxygenase *note PFAM: cytochrome P450	A0A0N1FIJ2
JAMQBH010000043.1	82.634	1382	209	28	70	1437	6545	5181	0	1194	squalene-associated FAD-dependent desaturase	A0A0N1GAB7