ΣΧΟΛΗ ΧΗΜΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Διδακτορική διατριβή

Προσδιορισμός νέων μικροβιακών ενζύμων βιοτεχνολογικού ενδιαφέροντος μέσω μεταγενωμικής ανάλυσης

Ευθύμιος Λαδουκάκης

Αθήνα 2017

Τίτλος διδακτορικής διατριβής:

Προσδιορισμός νέων μικροβιακών ενζύμων βιοτεχνολογικού ενδιαφέροντος μέσω μεταγενωμικής ανάλυσης

Υποψήφιος Διδάκτορας:

Ευθύμιος Λαδουκάκης

Επταμελής Εξεταστική Επιτροπή:

Φραγκίσκος Κολίσης, Ομ. Καθηγητής ΕΜΠ (Επιβλέπων)

Δημήτρης Κέκος, Καθηγητής ΕΜΠ

Ανδρέας Μπουντουβής, Καθηγητής ΕΜΠ

Χαράλαμπος Σαρίμβεης, Καθηγητής ΕΜΠ

Γεράσιμος Λυμπεράτος, Καθηγητής ΕΜΠ

Ευάγγελος Τόπακας, Επ. Καθηγητής ΕΜΠ

Αριστοτέλης Χατζηιωάννου, Ερευνητής Β' ΕΙΕ

Η έγκριση της διδακτορικής διατριβής από την Ανώτατη Σχολή Χημικών Μηχανικών του Ε.Μ. Πολυτεχνείου δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα (Ν. 5343/1932, Άρθρο 202).

Ευχαριστίες

Από που να αρχίσω και που να τελειώσω... Ίσως μία από τις πιο κλισέ εκφράσεις που υπάρχουν όπως και ένας πολύ καλός εναλλακτικός τίτλος για το διδακτορικό μου. Παρ' όλα αυτά ήταν η πρώτη φράση που μου ήρθε στο μυαλό όταν ξεκίνησα να γράφω το κομμάτι των ευχαριστιών. Θεωρώ ότι υπήρξα απίστευτα τυχερός που είχα δίπλα μου τόσα πολλά άτομα να με βοηθήσουν, άλλοι στο επίπεδο της δουλειάς, άλλοι με μία συνεχή ψυχολογική υποστήριξη και άλλοι και στα δύο ταυτόχρονα.

Θα ξεκινήσω με τα δύο άτομα χωρίς τα οποία δεν θα μπορούσε ποτέ να ολοκληρωθεί η παρούσα διδακτορική διατριβή, τον επιβλέποντα καθηγητή μου κ. Φραγκίσκο Κολίση ο οποίος με εισήγαγε στο χώρο της έρευνας και τον Ερευνητή Β' από το Εθνικό Ίδρυμα Ερευνών, Δρ. Αριστοτέλη Χατζηϊωάννου, ο οποίος με εισήγαγε στον κόσμο της βιοπληροφορικής απ' το προπτυχιακό κιόλας στάδιο σπουδών μου. Η συνεισφορά και των δύο κατά τη διάρκεια της διδακτορικής μου διατριβής ήταν ύψιστης σημασίας, όχι μόνο με την ουσιαστική καθοδήγηση στη δουλειά μου αλλά και με τη συνεχή ανταλλαγή απόψεων μεταξύ μας η οποία εν τέλει έπαιξε καθοριστικό ρόλο στη τελική διαμόρφωση της δικιάς μου οπτικής γωνίας τόσο σχετικά με τον ερευνητικό όσο και με τον ακαδημαϊκό τομέα. Όταν μιλάω βέβαια για ανταλλαγή απόψεων δεν θα μπορούσα να μην αναφέρω τους πολύ στενούς μου φίλους Κατερίνα Κ., Λεωνίδα Μ. και Παναγιώτη Μ. και να μην τους ευχαριστήσω που, παρ' όλες τις ιδεολογικές διαφωνίες μας, αποτέλεσαν (και αποτελούν) ένα αναπόσπαστο κομμάτι της ζωής μου και ήταν δίπλα μου καθ' όλη τη διάρκεια του διδακτορικού μου. Δεν θα μπορούσα ποτέ να ξεχάσω φυσικά και τους υπόλοιπους καλούς φίλους που έκανα μέσα από το εργαστήριο Βιοτεχνολογίας της σχολής Χημικών Μηχανικών και που με στήριξαν όλα αυτά τα χρόνια, τη Δήμητρα Ζ., τη Μαρία Σ., το Θωμά Π., τον Παναγιώτη Α., τη Μαρία (ΠΠ) Γ. και φυσικά τη Μίκυ Μ. την οποία δεν κατάφερα ποτέ να πείσω να με υιοθετήσει.

Η ενασχόλησή μου με τον τομέα της βιοπληροφορικής όμως με έφερε σε επαφή και με άτομα εκτός Πολυτεχνείου με τα οποία αναπτύχθηκαν συνεργασίες και στενές φιλίες που καθόρισαν την έκβαση αυτού του διδακτορικού. Ο Λευτέρης Π., η Όλγα Π., ο Γιώργος (U) Σ., η Μαριάνθη Λ., ο Στάθης Β. και ο Θοδωρής Κ. είναι μόνο λίγοι από τους καλούς φίλους που απέκτησα μέσα από τη συνεργασία μου με το

Εθνικό Ίδρυμα Ερευνών και την εταιρεία e-NIOS. Επίσης ένα άτομο που με στηρίζει ακατάπαυστα όλα αυτά τα χρόνια είναι η κολλητή μου Βασιλική Β. η οποία, αν και μας χωρίζουν χιλιάδες χιλιόμετρα, ξέρω ότι θα ναι πάντα δίπλα μου. Τέλος, θα ήθελα να ευχαριστήσω ένα άτομο που κατά πάσα πιθανότητα δεν θα μάθει ποτέ ότι μπήκε το όνομά του σε αυτήν την ενότητα, τον Γιάννη Κ. Β. ο οποίος χωρίς να το ξέρει με βοήθησε ίσως στην πιο δύσκολη περίοδο αυτής της διατριβής.

Περιεχόμενα

Περίληψη

Σκοπός της παρούσας διατριβής ήταν η ανάπτυξη ενός αυτοματοποιημένου συστήματος βιοπληροφορικών αναλύσεων, το οποίο θα μπορούσε να διαχειριστεί και να αναλύσει μεταγενωμικά δεδομένα, με τελικό στόχο την εύρεση καινούριων ενζύμων βιοτεχνολογικού ενδιαφέροντος. Η διαδικασία σχεδιασμού του αυτοματοποιημένου συστήματος περιελάμβανε την αξιολόγηση πολυάριθμων βιοπληροφορικών εργαλείων μέσω της εφαρμογής τους σε πραγματικά μεταγενωμικά δεδομένα καθώς και την ανάπτυξη καινούριων αλγορίθμων που καλύπτουν τις αδυναμίες των ήδη υπαρχόντων. Η συλλογή των διαφορετικών εργαλείων και των νέων αλγορίθμων ενοποιήθηκε σε μία διαδικτυακή πλατφόρμα που κατασκευάστηκε με βάση το υπολογιστικό σύστημα Galaxy και ονομάστηκε ANASTASIA (Automated Nucleotide Aminoacid Sequences Translational plAtform for Systemic Interpretation and Analysis). Στην καινούρια πλατφόρμα το κάθε εργαλείο και αλγόριθμος γινόταν διαθέσιμο μέσω ενός φιλικού προς το χρήστη γραφικού περιβάλλοντος ενώ υπήρχε η δυνατότητα αυτοματοποίησης των αναλύσεων που περιλάμβαναν πολλά διαδοχικά εργαλεία μέσα από βιοπληροφορικές γραμμές εργασιών (pipelines).

Ο σχεδιασμός των αυτοματοποιημένων γραμμών εργασιών έγινε μέσω της εφαρμογής των ενσωματωμένων εργαλείων σε μεταγενωμικά δεδομένα που αποκτήθηκαν από δύο ερευνητικά προγράμματα: το HotZyme και το COVERALL. Το πρώτο πρόγραμμα είχε ως σκοπό την εύρεση καινούριων θερμοσταθερών ενζύμων μέσω ανάλυσης των μεταγονιδιωμάτων μικροβιακών πληθυσμών σε θερμές πηγές, ενώ το δεύτερο πραγματευόταν την εύρεση μικροβιακών ειδών και αλληλουχιών που θα μπορούσαν να συσχετιστούν ως βιοδείκτες με την έκθεση σε υψηλές συγκεντρώσεις CO_2. Η βιοπληροφορική ανάλυση και στα δύο ερευνητικά προγράμματα ξεκίνησε από το επίπεδο των δεδομένων αλληλούχισης των μεταγονιδιωμάτων των αντίστοιχων δειγμάτων αλλά εξελίχθηκε σε δύο διαφορετικές μεθοδολογίες από τις οποίες προέκυψαν οι αντίστοιχες αυτοματοποιημένες γραμμές εργασιών. Η τελική έκδοση της πλατφόρμας με τα ενσωματωμένα εργαλεία και τις αντίστοιχες αυτοματοποιημένες γραμμές εργασιών έγινε στη συνέχεια διαθέσιμη διαδικτυακά αξιοποιώντας ένα διακομιστή που ανήκει στη σχολή Χημικών Μηχανικών ΕΜΠ στη διεύθυνση http://motherbox.chemeng.ntua.gr/anastasia_dev/.

Από τα αποτελέσματα της ανάλυσης των δεδομένων του ερευνητικού προγράμματος HotZyme προέκυψε μία λίστα αλληλουχιών πιθανών ενζύμων με

υδρολυτική δράση (αριθμός EC 3.-.-.-) η οποία εξετάστηκε περαιτέρω για την επιλογή των επικρατέστερων υποψηφίων για εργαστηριακή επιβεβαίωση. Κατά τη συγγραφή αυτής της διατριβής ήδη δύο από τις παραπάνω αλληλουχίες έχουν απομονωθεί στο εργαστήριο, έχουν εκφραστεί επιτυχώς, έχουν χαρακτηριστεί πλήρως ως προς την ενζυμική λειτουργία τους και έχουν καταγραφεί σε δημόσιες βάσεις δεδομένων (UniProt) ως καινούριες καταχωρήσεις, επιβεβαιώνοντας τις αρχικές μας προβλέψεις. Αντίστοιχα τα αποτελέσματα του ερευνητικού προγράμματος COVERALL αποκάλυψαν 23 διαφορετικά μικροβιακά είδη των οποίων η παρουσία φαίνεται να συνδέεται στενά με την έκθεση σε υψηλές συγκεντρώσεις CO_2. Οι αντίστοιχες γονιδιακές τους αλληλουχίες έχουν ήδη απομονωθεί υπολογιστικά και ήδη λαμβάνει χώρα περαιτέρω ανάλυση για τον εντοπισμό χαρακτηριστικών αλληλουχιών που θα αποτελέσουν πιθανούς βιοδείκτες έκθεσης για το συγκεκριμένο ρύπο.

Ο σχεδιασμός αυτής της πλατφόρμας, μέσα από τη συνεχή αλληλεπίδραση με πραγματικά μεταγενωμικά δεδομένα, βοήθησε εξαιρετικά στην αξιολόγηση των δυνατοτήτων της, αλλά και στην αναπροσαρμογή των αλγορίθμων της για τη βέλτιστη διαχείριση και ανάλυση των αντίστοιχων αρχείων. Έτσι το ουσιαστικό αποτέλεσμα αυτής της διατριβής δεν αποτελείται μόνο από τα συμπεράσματα των εκάστοτε αναλύσεων, αλλά σαφώς επίσης και από το εύχρηστο και διαρκώς εξελισσόμενο υπολογιστικό σύστημα που προέκυψε. Οι δυνατότητες αυτού του συστήματος ενώ έχουν αποδειχτεί για την περίπτωση μεταγονιδιωματικών δεδομένων μπορούν να επεκταθούν (και ήδη επεκτείνονται) περαιτέρω για όλους τους τομείς της Βιοτεχνολογίας και της Συνθετικής Βιολογίας.

Abstract

Aim of this thesis was the development of an automated bioinformatic framework that could effectively handle and analyze metagenomic data with the final scope being the detection of novel enzymes of industrial interest. The design of the aforementioned framework comprised evaluating various open source bioinformatic tools via running multiple analyses in real metagenomic datasets, as well as developing new algorithms that could tackle any issues derived from these analyses. The selected tools and developed algorithms were integrated in a web-based platform which was developed by exploiting Galaxy's computational framework and was named ANASTASIA (Automated Nucleotide Aminoacid Sequences Translational plAtform for Systemic Interpretation and Analysis). This new platform offered a friendly graphic user interface for all tools incorporated in it, while enabling the automation of each analysis in which they were executed, through the use of appropriate computational pipelines.

The design of the computational pipelines was facilitated by using the integrated tools directly on real metagenomic datasets acquired from two research projects: HotZyme and COVERALL. HotZyme project aimed in discovering novel thermostable enzymes via metagenomic screening of environmental samples from terrestrial hot springs, while COVERALL focused on detecting the taxonomical and functional differences in metagenomic samples from seafloor sediments, that were exposed in high concentrations of CO_2. The bioinformatic analysis in both projects was initiated at the level of metagenomic sequencing data, but was formed in two distinct methodologies of different analytical steps, which were later transformed into the corresponding automated bioinformatic pipelines. The final version of the platform, including the automated pipelines and the integrated tools they consist of, was rendered available online, by using a server owned by the school of Chemical Engineering in National Technical University of Athens, via the following URL: http://motherbox.chemeng.ntua.gr/anastasia_dev/.

The results from HotZyme project consisted of a list of nucleotide sequences of putative hydrolytic activity (EC number: 3.-.-.-) which was further curated in order to select the most promising candidates for experimental validation. During the work for this thesis, two of those sequences were successfully isolated in the lab, expressed, fully annotated and registered in public databases (UniProt) as novel sequences, while confirming our initial prediction regarding their enzymatic activity. On COVERALL

project, the utilization of ANASTASIA resulted in identifying 23 different species that were found present only during exposure to high concentrations of CO_2. The corresponding genomic sequences were parsed and are already under way of further analysis for detecting distinctive biomarker sequences for exposure in that pollutant.

Designing this platform through the constant interaction of real metagenomic data, was of utmost importance for evaluating its potential, as well as for readjusting its algorithms for further optimizing the handling and analyzing of the corresponding files. Thus, the significant outcome of this thesis does not consist solely of the analysis results for each project but also of the user friendly and constantly evolving computational framework that was developed. The potential of that framework has already been proven in the case of metagenomic data but can be further (and already is) expanding for all fields of Biotechnology, -omics technologies and Synthetic Biology.

ΚΕΦΑΛΑΙΟ 1: Εισαγωγή

1.1 Μεταγενωμική

Σε όλους τους ζωντανούς οργανισμούς, μονοκύτταρους ή πολυκύτταρους, προκαρυωτικούς ή ευκαρυωτικούς, κάθε βιολογική λειτουργία τους καθορίζεται από ένα πλήθος «οδηγιών» που είναι αποθηκευμένες, μέσα στα κύτταρα που τους απαρτίζουν, σε βιολογικά μακρομόρια [1] το δεοξυριβονουκλεϊκό οξύ ή DNA (DeoxyriboNucleic Acid) και το ριβονουκλεϊκό οξύ ή RNA (RiboNucleic Acid). Τα μακρομόρια αυτά μέσω της σύστασής τους, εμπεριέχουν το γενετικό πρόγραμμα των οργανισμών που διέπει όλες τις λειτουργίες κάθε κυττάρου, από την σύνθεση μεμονωμένων χημικών ενώσεων που χρειάζεται έως την οργάνωση πολύπλοκων δικτύων αλληλεπίδρασης των ενώσεων αυτών, τα οποία ενεργοποιούνται για την εύρυθμη κυτταρική λειτουργία, την απόκρισή σε εξωτερικά ερεθίσματα και τον κυτταρικό πολλαπλασιασμό. Το σύνολο των μακρομορίων αυτών σε κάθε κύτταρο συνιστά το γενετικό υλικό ή γονιδίωμα του οργανισμού. Η επέκταση αυτού του ορισμού σε έναν μικροβιακό πληθυσμό ονομάζεται μεταγονιδίωμα [2] και αναφέρεται στο σύνολο των γονιδιωμάτων των διαφορετικών οργανισμών, ανεξάρτητα με το αν βρίσκονται σε αφθονία μέσα στο εξεταζόμενο δείγμα ή όχι.

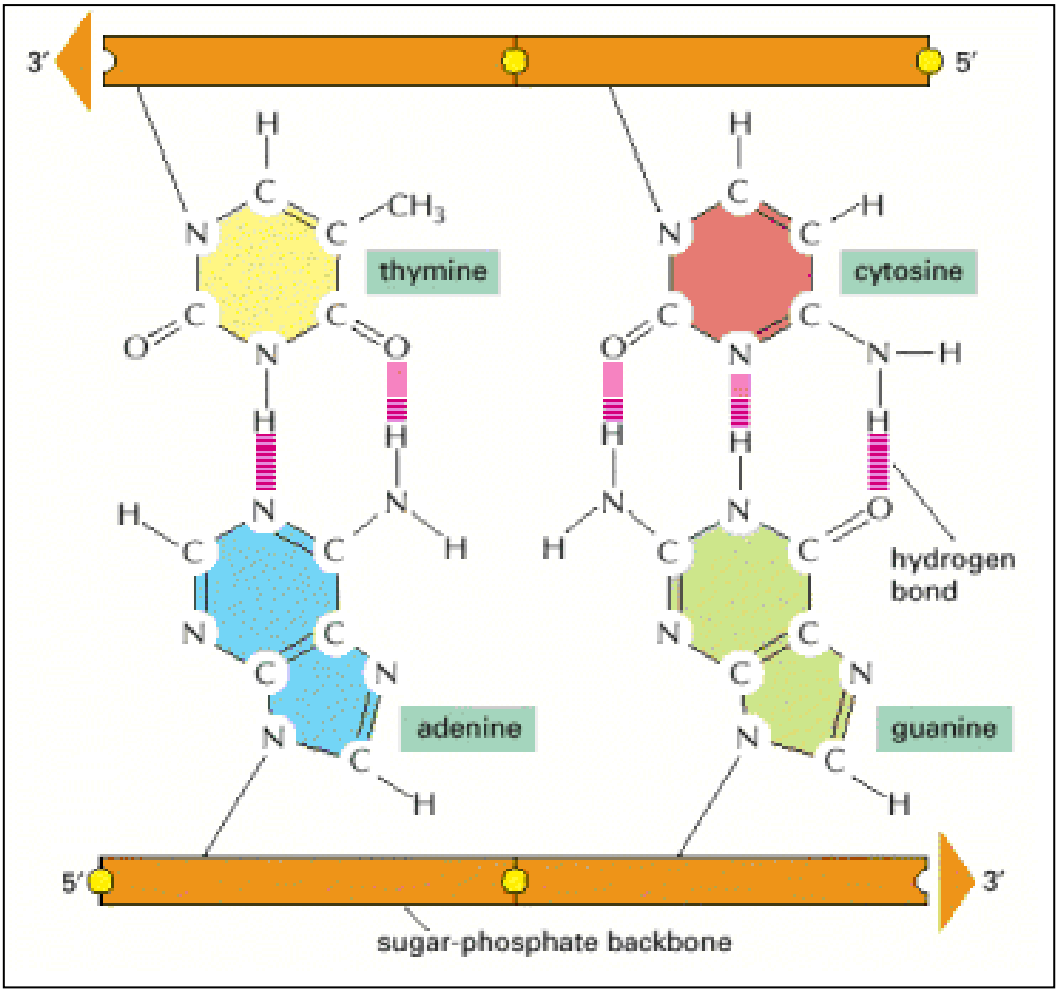
Η διεξοδική μελέτη του μεταγονιδώματος σε έναν οικολογικό θώκο (environmental niche) αποτελεί απαραίτητη προϋπόθεση για την ανάλυση της βιοποικιλότητας του μικροβιακού πληθυσμού που τον αποτελεί [3-5], όπως αυτός έχει διαμορφωθεί λόγω της διαρκούς εξελικτικής πίεσης [6]. Μπορεί μάλιστα να αξιοποιηθεί περαιτέρω για την αξιολόγηση του ενζυμικού δυναμικού του [7] με τον εντοπισμό και καταγραφή των αντίστοιχων γονιδίων που εμπεριέχει. Με τον όρο Μεταγενωμική χαρακτηρίζεται το σύνολο των τεχνικών και τεχνολογιών που χρησιμοποιούνται για την διεξαγωγή του συγκεκριμένου είδους μελέτης, οι οποίες προέρχονται από μία πληθώρα διαφορετικών επιστημονικών τομέων όπως η χημεία, η συνθετική βιολογία, η βιοτεχνολογία, η βιοπληροφορική κ.α. Η μεταγενωμική έχει γνωρίσει ραγδαία εξέλιξη τα τελευταία χρόνια [8, 9] με τις εφαρμογές τις να πληθαίνουν τόσο στη βιομηχανία [10] όσο και στους κλάδους υγείας [11-13]. Η κινητήρια δύναμη πίσω από αυτήν της την εξέλιξη είναι η ανάπτυξη καινούριων τεχνολογιών γονιδιακής αλληλούχισης (αλληλούχιση νέας γενιάς - next generation sequencing) [14] οι οποίες μας επιτρέπουν την καταγραφή των γενετικών πληροφοριών με πολύ μεγαλύτερη απόδοση από ότι στο παρελθόν ενώ ταυτόχρονα απαιτούν εξαιρετικά μικρότερο κόστος [15].

1.2 Μεταγονιδιωματική αλληλούχιση

1.2.1 Δομή γενετικού υλικού - γονιδιωματική αλληλουχία

Η σύσταση των βιοπολυμερικών μορίων που αποτελούν το γενετικό υλικό των οργανισμών, μπορεί να γίνει κατανοητή εάν εξεταστούν τα δομικά στοιχεία (μονομερή) από τα οποία αποτελούνται. Το DNA έχει ως δομικά στοιχεία του οργανικά μόρια που ονομάζονται (δεοξυριβο)νουκλεοτίδια, καθένα από τα οποία αποτελείται από μία πεντόζη, τη δεοξυριβόζη, μία φωσφορική ομάδα ενωμένη στον 5' άνθρακα της πεντόζης, και μία αζωτούχα βάση. Κάθε νουκλεοτίδιο χαρακτηρίζεται από την αζωτούχα βάση που περιέχει η οποία μπορεί να είναι μία εκ των τεσσάρων: αδενίνη (A), θυμίνη (T), γουανίνη (G), και κυτοσίνη (C). Επίσης κάθε νουκλεοτίδιο συνδέεται με το επόμενο μέσω φωσφοδιεστερικού δεσμού μεταξύ της φωσφορικής ομάδας (PO_3-) του πρώτου και του υδροξυλίου (-OH) του 3' άνθρακα της πεντόζης του δευτέρου. Με αυτό τον τρόπο σχηματίζεται μια πολυνουκλεοτιδική αλυσίδα που λέμε ότι έχει προσανατολισμό 5'-3' καθώς το πρώτο νουκλεοτίδιό της έχει ελεύθερη τη φωσφορική ομάδα (5' άκρο) και το τελευταίο έχει ελεύθερη την ομάδα υδροξυλίου (3' άκρο). Η τελική δομή του μορίου DNA προκύπτει από δύο πολυνουκλεοτιδικές αλυσίδες τέτοιου είδους που ενώνονται πλευρικά μέσω δεσμών υδρογόνου μεταξύ των αζωτούχων βάσεων. Οι δεσμοί υδρογόνου σχηματίζονται με βάση τον κανόνα συμπληρωματικότητας των βάσεων - η αδενίνη πάντα ενώνεται με την θυμίνη με δύο δεσμούς υδρογόνου και η γουανίνη πάντα ενώνεται με την κυτοσίνη με τρεις δεσμούς υδρογόνου (Εικόνα 1). Οι δύο συμπληρωματικές αλυσίδες, που ονομάζονται και κλώνοι, έχουν αντιπαράλληλο προσανατολισμό (το 5' άκρο της μίας αντιστοιχεί στο 3' άκρο της άλλης) και σχηματίζουν μία δεξιόστροφη έλικα [16] με τις αζωτούχες βάσεις στο εσωτερικό της (Εικόνα 2).

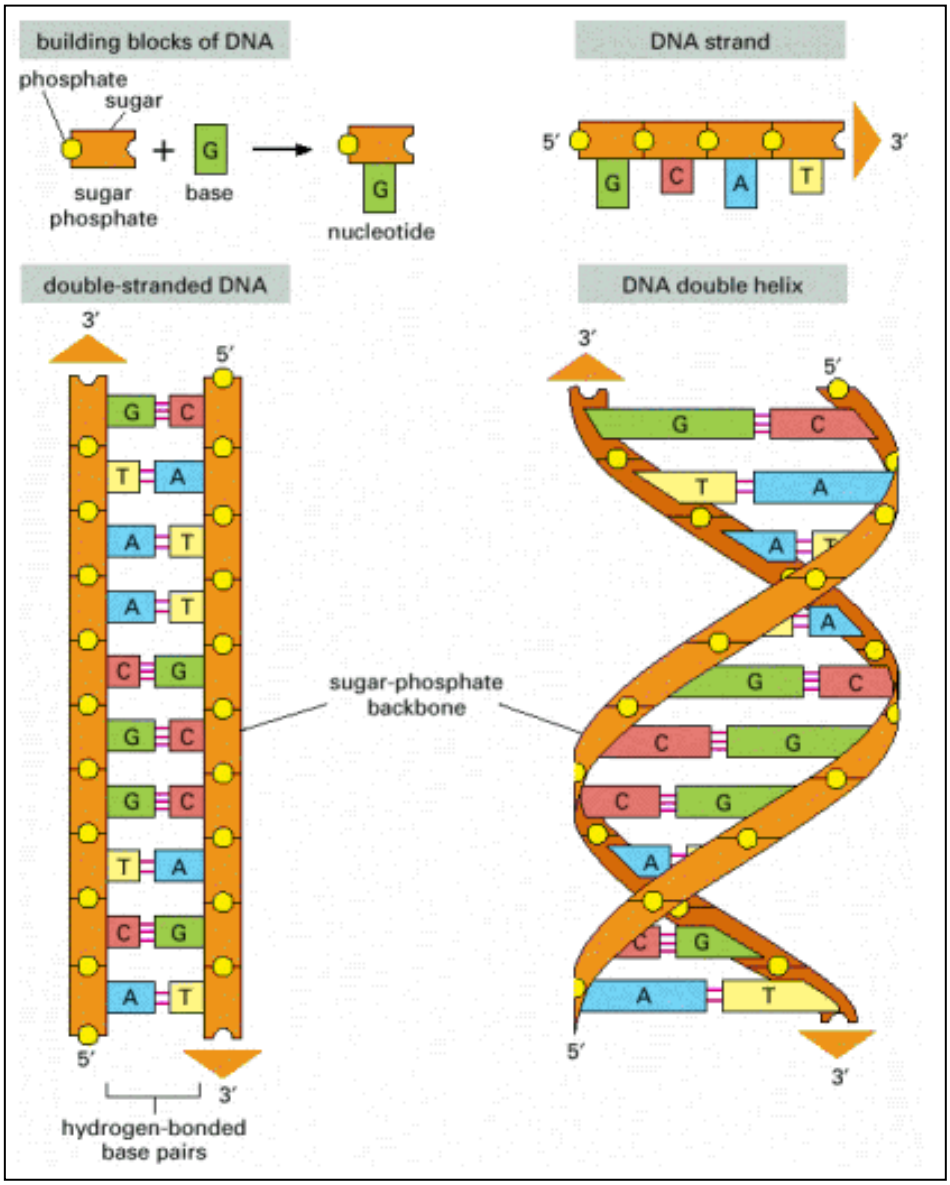
Αντίστοιχα το RNA αποτελείται από ριβονουκλεοτίδια, δηλαδή νουκλεοτίδια των οποίων η πεντόζη είναι η ριβόζη. Η δομή του είναι παρόμοια με το DNA, με τα νουκλεοτιδιά του να συνδέονται μεταξύ τους με φωσφοδιεστερικούς δεσμούς για το σχηματισμό μίας πολυνουκλεοτιδικής αλυσίδας. Επίσης, όπως και στο DNA, κάθε νουκλεοτίδιό του χαρακτηρίζεται από την αζωτούχα βάση που περιέχει, αλλά στη συγκεκριμένη περίπτωση, η τετράδα των βάσεων περιλαμβάνει την ουρακίλη (U) στη θέση της θυμίνης (Εικόνα 3).



Εικόνα 1. Συμπληρωματικά ζεύγη βάσεων στην διπλή έλικα του DNA. Το σχήμα και η χημική δομή των βάσεων επιτρέπουν το σχηματισμό δεσμών υδρογόνου μόνο μεταξύ αδενίνης (A) και (T) όπως και μόνο μεταξύ γουανίνης (G) και κυτοσίνης (C). Αυτό συμβαίνει, καθώς σε κάθε ζεύγος τα άτομα που λαμβάνουν μέρος στο δεσμό υδρογόνου μπορούν να πλησιάσουν το ένα το άλλο χωρίς να υπάρξει παραμόρφωση της διπλής έλικας. Όπως φαίνεται από την εικόνα οι συμπληρωματικές βάσεις μπορούν να συνδεθούν κατά αυτόν τον τρόπο μόνο εάν οι δύο πολυνουκλεοτιδικές αλυσίδες βρίσκονται αντιπαράλληλα μεταξύ τους [17].

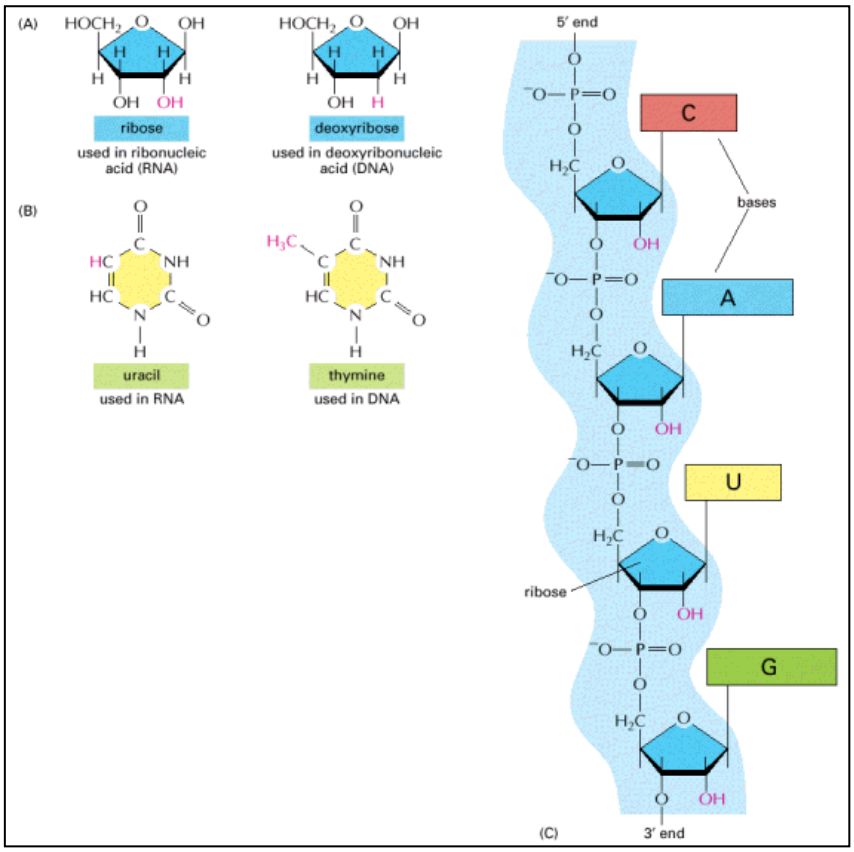
Άλλη διαφορά μεταξύ των δύο μορίων είναι ότι το DNA εμφανίζεται πάντοτε ως δίκλωνη αλυσίδα σε αντίθεση με το RNA που μπορεί να εμφανιστεί ως γενετικό υλικό είτε μονόκλωνο είτε δίκλωνο. Στη δεύτερη περίπτωση, ο αντίστοιχος κανόνας συμπληρωματικότητας για τις βάσεις του είναι η δημιουργία δύο δεσμών υδρογόνου μεταξύ αδενίνης και ουρακίλης, ενώ η γουανίνη και η κυτοσίνη συνδέονται με τρεις δεσμούς υδρογόνου όπως και στο DNA (Εικόνα 4). Εκτός των δομικών/χημικών διαφορών που έχουν τα δύο βιομόρια, η βασικότερη ίσως διάκριση μεταξύ τους αφορά είναι τα είδη στα οποία χρησιμοποιούνται ως γενετικό υλικό. Το RNA έχει βρεθεί να αποτελεί το γενετικό υλικό μόνο σε ιούς [18] ενώ το DNA συναντάται πιο συχνά και σε κάθε φυλογενετικό επίπεδο, από ιούς, βακτήρια και αρχαία έως και τους πιο σύνθετους ευκαρυωτικούς μονοκύτταρους ή πολυκύτταρους οργανισμούς.

Η αποθήκευση της γενετικής πληροφορίας τόσο στο DNA όσο και στο RNA βασίζεται στα τέσσερα διαφορετικά νουκλεοτίδια που χαρακτηρίζονται από την αζωτούχα βάση τους (A, T, G, C στο DNA και A, U, G, C στο RNA). Η διαδοχική εναλλαγή αυτών των βάσεων στο εκάστοτε μόριο ονομάζεται η αλληλουχία του γενετικού υλικού και αποτελεί των κώδικα με τον οποίο είναι αποθηκευμένες όλες οι γενετικές πληροφορίες του οργανισμού.



Εικόνα 2. Το DNA και τα δομικά στοιχεία του. Το DNA αποτελείται από τέσσερα είδη νουκλεοτιδίων που συνδέονται μέσω ομοιοπολικών δεσμών σχηματίζοντας μία πολυνουκλεοτιδική αλυσίδα με σκελετό σακχάρου-φωσφορικού οξέος. Κάθε μόριο DNA αποτελείται από δύο τέτοιες αλυσίδες που συνδέονται μεταξύ τους μέσω των δεσμών υδρογόνου μεταξύ των συμπληρωματικών αζωτούχων βάσεων (A-T και G-C). Τα βέλη υποδεικνύουν τον προσανατολισμό των δύο αλυσίδων (5'-3') οι οποίες είναι αντιπαράλληλες μεταξύ τους μέσα στο μόριο. Αριστερά φαίνεται η διπλή αλυσίδα όπως θα ήταν ξεδιπλωμένη κατά μήκος του μορίου DNA. Δεξιά φαίνεται η πραγματική της δομή όπου σχηματίζει μία δεξιόστροφη έλικα με τον σκελετό σακχάρου-φωσφορικού οξέος να βρίσκεται στο εξωτερικό και τις αζωτούχες βάσεις να βρίσκονται στο εσωτερικό της [17].

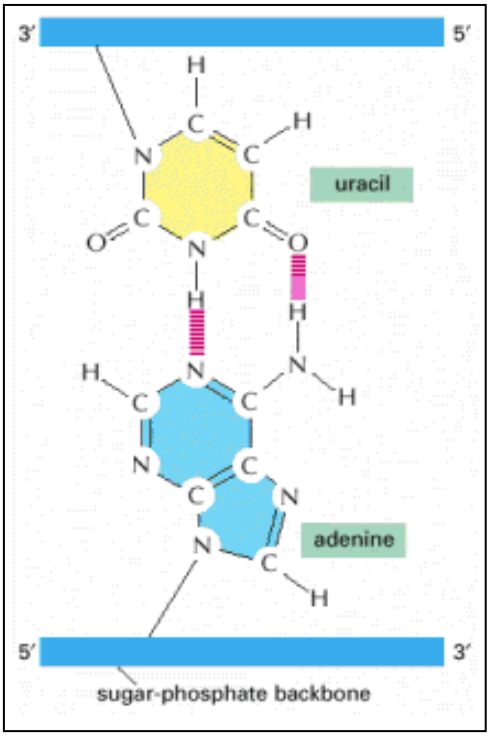
Η αποκρυπτογράφηση αυτής της πληροφορίας από ένα κύτταρο γίνεται με τη διαδικασία της μεταγραφής κατά την οποία συγκεκριμένες αλληλουχίες του DNA, που ονομάζονται γονίδια, χρησιμοποιούνται ως «καλούπι», μέσω της συμπληρωματικότητας των βάσεων, για την παραγωγή μίας αντίστοιχης αλληλουχίας RNA. Η αλληλουχία αυτή μεταφέρει την γονιδιακή πληροφορία στα οργανίδια παραγωγής πρωτεϊνών του κυττάρου (ριβοσώματα) ώστε να ξεκινήσει η σύνθεση πρωτεϊνών με τον μηχανισμό της μετάφρασης κατά την οποία γίνεται αντιστοίχιση ενός αμινοξέος για κάθε διαδοχική τριάδα νουκλεοτιδίων με βάση τον γενετικό κώδικα.



Εικόνα 3. Η χημική δομή του RNA. (A)Το RNA περιέχει μία πεντόζη που ονομάζεται ριβόζη σε αντίθεση με τη δεοξυριβόζη που χρησιμοποιείται στο DNA. Η διαφορά τους έγκειται σε μία επιπλέον υδροξυλομάδα που υπάρχει στη ριβόζη. (B)Το RNA περιέχει τη βάση ουρακίλη που διαφέρει από την αντίστοιχη στο DNA θυμίνη λόγω της έλλειψης μίας μεθυλομάδας. (C)Μία πολυνουκλεοτιδική αλυσίδα του RNA. Τα νουκλεοτίδια συνδέονται κατά τον ίδιο τρόπο με προσανατολισμό 5'-3' όπως και στο DNA[17].

Συνεπώς το πρόβλημα (μετα)γονιδιωματικής μελέτης ενός ή περισσοτέρων κυτταρικών οργανισμών μετατρέπεται σε πρόβλημα αποκωδικοποίησης του (μετα)γονιδιώματός τους, δηλαδή την πλήρη καταγραφή της αλληλουχίας και των γονιδίων που περιέχει. Η διαδικασία αυτή ονομάζεται αλληλούχιση και περιλαμβάνει

πρωτόκολλα απομόνωσης DNA και εισαγωγή τους σε εξειδικευμένα μηχανήματα αλληλούχισης τα οποία όμως απαιτούν μία υπολογιστικά επίπονη βιοπληροφορική ανάλυση των δεδομένων τους [19].



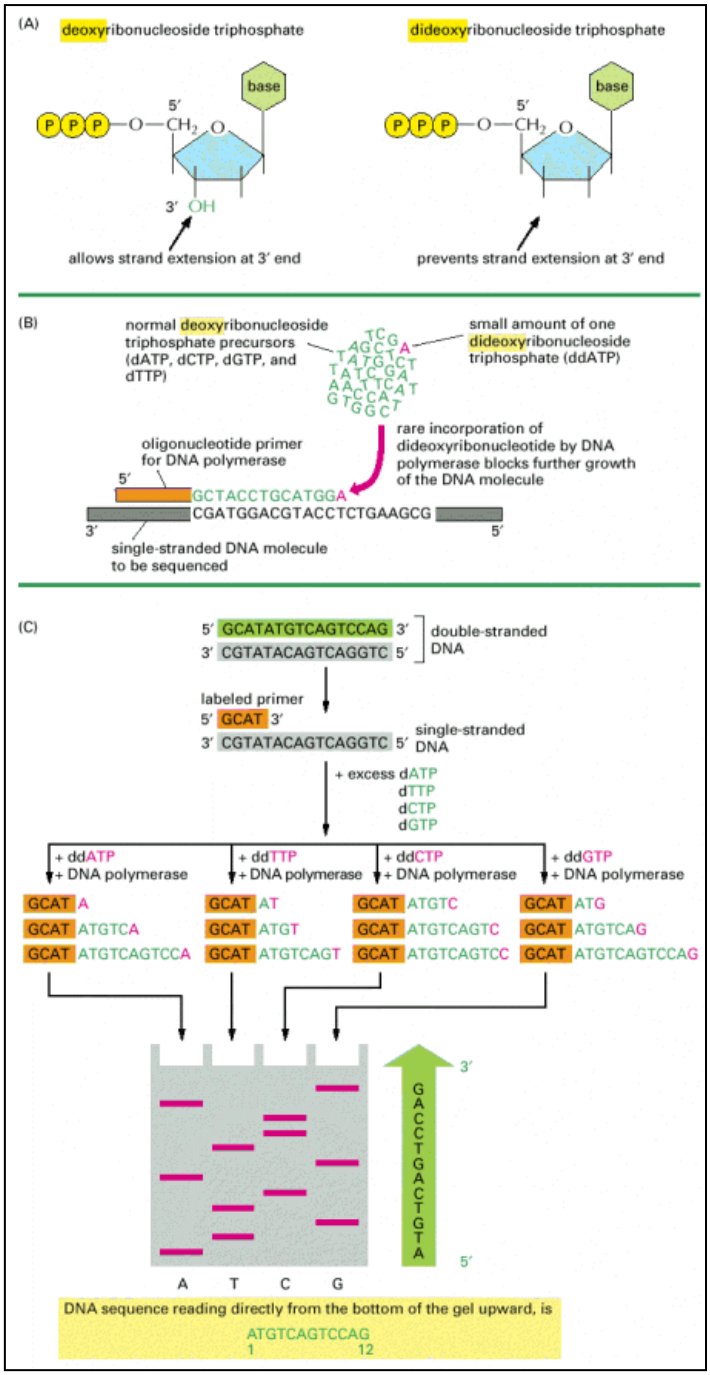
Εικόνα 4. Συμπληρωματικό ζεύγος βάσεων μεταξύ ουρακίλης και αδενίνης. Η απουσία της μεθυλομάδας στην ουρακίλη δεν έχει κάποια επίπτωση στη δημιουργία ζεύγους βάσεων μέσω δεσμών υδρογόνου. Έτσι το ζεύγος που σχηματίζεται είναι παρόμοιο με το αντίστοιχο αδενίνης και θυμίνης[17].

1.2.2 Κλασσικές τεχνικές αλληλούχισης

Η πρώτη τεχνική αλληλούχισης που έδωσε τη δυνατότητα να καταγραφούν ταχύτατα από γονίδια μέχρι και ολόκληρα γονιδιώματα αναπτύχθηκε το 1977 [20] από τον Fred Sanger και Alan R. Coulson και ήταν η εξέλιξη της μεθόδου τους που είχαν αναπτύξει πριν από δύο χρόνια [21], καθώς και μία σαφής βελτίωση της αντίστοιχης μεθόδολογίας των Maxam και Gilbert που είχε εμφανιστεί το ίδιο έτος [22]. Η αρχή λειτουργίας της βασίζεται στην διάνοιξη της δίκλωνης αλυσίδας του δείγματος DNA προς αλληλούχιση και στον διαμοιρασμό των μονόκλωνων αλυσίδων που προέκυπταν σε τέσσερις διαφορετικές αντιδράσεις πολυμερισμού. Κάθεμια από τις αντιδράσεις αυτές, έχει ως προϊόντα συμπληρωματικές (ως προς το αρχικό μόριο) πολυνουκλεοτιδικές αλυσίδες διαφορετικού μεγέθους αλλά ίδιου νουκλεοτιδίου στην τελευταία θέση τους - ένα από τα τέσσερα διαφορετικά νουκλεοτίδια για κάθε μία από τις τέσσερις διαφορετικές αντιδράσεις. Κατά τη μέθοδο αυτή, η επίτευξη της

λήξης κάθε αντίδρασης πολυμερισμού συμπληρωματικής αλυσίδας σε συγκεκριμένο νουκλεοτίδιο επιτυγχάνεται με την ανάμειξη μικρής ποσότητας ενός από τα τέσσερα διαφορετικά τριφωσφορικά διδεοξυνουκλεοτίδια (di-deoxynucleotidetriphosphates - ddNTPs), μαζί με τις απαραίτητες ποσότητες των υπολοίπων αντιδρώντων, δηλαδή των τριφωσφορικών δεοξυνουκλεοτιδίων (deoxynucleotidetriphosphates - dNTPs) που θα χρησιμοποιηθούν ως μονομερή κατά την αντίδραση. Τα υπόλοιπα αντιδρώντα της αντίδρασης είναι μία συγκεκριμένη ολιγονουκλεοτιδική αλληλουχία που δρα ως εκκινητής (primer) καθώς επίσης και το ενζύμο DNA πολυμεράση. Η αντίδραση πολυμερισμού ξεκινάει όταν ο εκκινητής και το ένζυμο DNA πολυμεράση προσκολλούνται πάνω στο μόριο DNA προς αλληλούχιση, το οποίο δρα ως εκμαγείο για την παραγωγή της συμπληρωματικής αλληλουχίας. Η επιμήκυνση της συμπληρωματικής αλυσίδας γίνεται με τη διαδοχική προσθήκη dNTPs (που βρίσκεται σε περίσσεια), ενώ στην περίπτωση προσθήκης του αντίστοιχου ddNTP το ένζυμο αποκολλάται και σταματάει ο πολυμερισμός της αλυσίδας. Τα προϊόντα που προκύπτουν στο τέλος των τεσσάρων αντιδράσεων υποβάλλονται σε τριχοειδή ηλεκτροφόρηση, ώστε η τελική αλληλουχία του αναλυόμενου μορίου DNA να προκύπτει από τα μετρούμενα μεγέθη των αλυσίδων (μετρούμενα σε πλήθος βάσεων) με γνωστό νουκλεοτίδιο στην τελευταία θέση (Εικόνα 5).

Η αξιοποίηση της τεχνικής Sanger για τον προσδιορισμό της αλληλουχίας ενός (μετα)γονιδιώματος βασίζεται επίσης στον κατακερματισμό του αρχικού μορίου DNA σε πλήθος θραυσμάτων καθώς δεν είναι δυνατή η αλληλούχιση ολόκληρων μορίων μεγέθους πάνω από περίπου 1000 βάσεις. Ο κατακερματισμός αυτός γίνεται με μηχανική θραύση ή με μερική χρήση περιοριστικών ενζύμων σε αντίγραφα του ίδιου μορίου DNA. Κατά αυτό τον τρόπο εξασφαλίζεται ότι το κάθε αντίγραφο του μορίου DNA σπάει σε διαφορετικές θέσεις, με τυχαίο τρόπο, παράγοντας έτσι επικαλυπτόμενα θραύσματα που μπορούν στη συνέχεια να αλληλουχηθούν με τη ddNTP προσέγγιση. Λόγω του τυχαίου αυτού κατακερματισμού του μορίου DNA, η τεχνική ονομάστηκε «τυφλή στόχευση» ή «αλληλούχιση τμηματικής ανάλυσης» (shotgun sequencing). Τα δεδομένα που προκύπτουν αποτελούνται από μια λίστα με τις «αναγνωσμένες αλληλουχίες» (sequencing reads) οι οποίες έχουν ίδιο ή μικρότερο μέγεθος από τα αντίστοιχα θραύσματα. Οι επικαλύψεις που περιέχουν μεταξύ τους οι αναγνωσμένες αλληλουχίες μπορούν να προσδιοριστούν με βιοπληροφορικούς αλγορίθμους για τον ανασχηματισμό της αρχικής αλληλουχίας (βλ. κεφάλαιο 1.3.3).

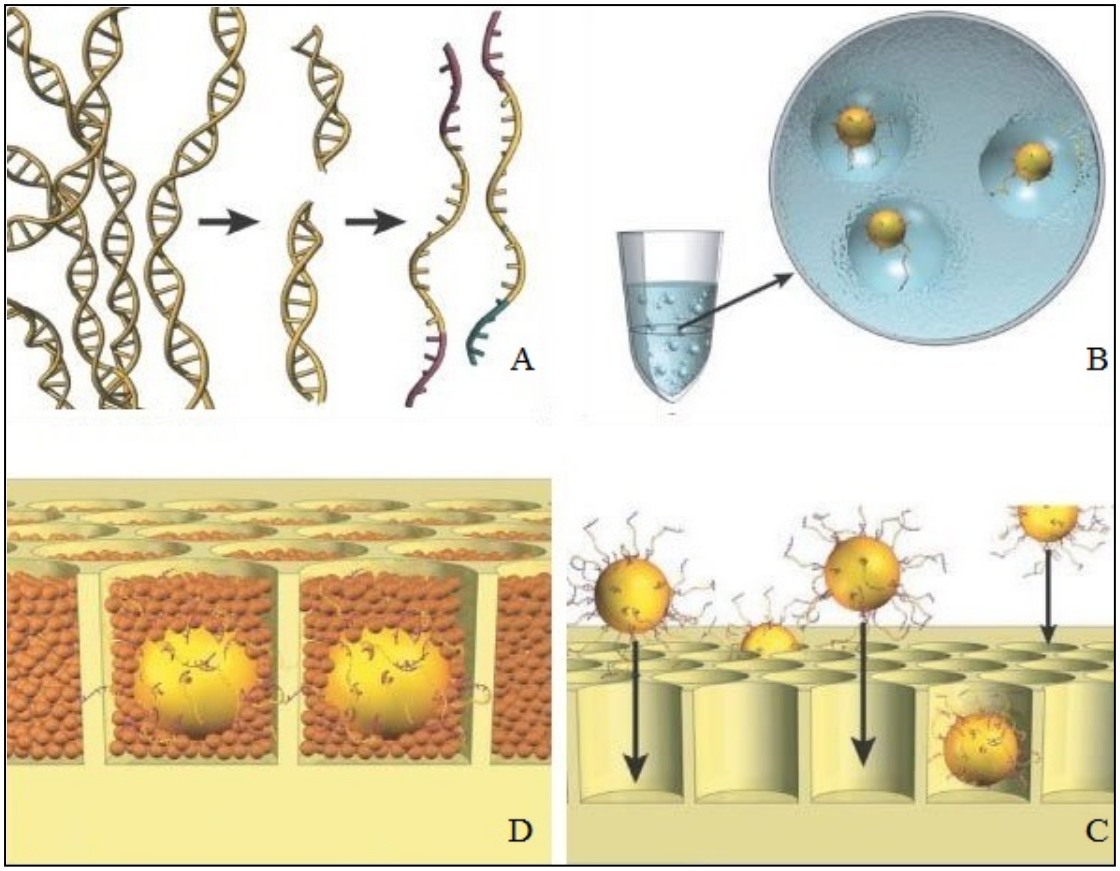


Εικόνα 5. Αλληλούχιση DNA με βάση τα διδεοξυνουκλεοτίδια. (A)Τα τριφωσφορικά διδεοξυνουκλεοτίδια της μεθόδου έχουν προκύψει από τα κανονικά τριφωσφορικά δεοξυνουκλεοτίδια με απομάκρυνση της υδροξυλομάδας στον 3' άνθρακα της πεντόζης. (B)Η σύνθεση μιας συμπληρωματικής αλυσίδας DNA με βάση το αρχικό μόριο(γκρι χρώμα) λαμβάνει χώρα in vitro σε ένα διάλυμα που περιέχει το ένζυμο DNA πολυμεράση, έναν εκκινητή (πορτοκαλί χρώμα), τα τέσσερα τριφωσφορικά δεοξυριβονουκλεοτίδια (πράσινο χρώμα) σε περίσσεια, καθώς και μικρή ποσότητα ενός εκ των τεσσάρων διαφορετικών τριφωσφορικών διδεοξυνουκλεοτιδίων (κόκκινο χρώμα). Η λήξη της αντίδρασης επιμήκυνσης της συμπληρωματικής αλυσίδας από την πολυμεράση συμβαίνει εάν το νουκλεοτίδιο που εισαχθεί είναι ένα τριφωσφορικό διδεοξυνουκλεοτίδιο καθώς η έλλειψη υδροξυλομάδας δεν επιτρέπει τον σχηματισμό νέου δεσμού με επόμενο νουκλεοτίδιο. (C)Κατά την αλληλούχισή του, το δείγμα DNA οδηγείται σε τέσσερις διαφορετικές αντιδράσεις λήξης αλυσίδας (μία για κάθε διαφορετικό διδεοξυνουκλεοτίδιο). Τα συντιθέμενα μόρια από κάθε αντίδραση οδηγούνται στη συνέχεια σε τέσσερα παράλληλα κανάλια ηλεκτροφόρησης όπου διαχωρίζονται με βάση το μήκος τους. Έτσι κάθε σημείο στο gel ηλεκτροφόρησης αντιστοιχεί σε ένα συγκεκριμένο μήκος αλυσίδας με γνωστό το τελευταίο νουκλεοτίδιό της [17].

1.2.3 Τεχνολογίες αλληλούχισης νέας γενιάς (Next generation sequencing)

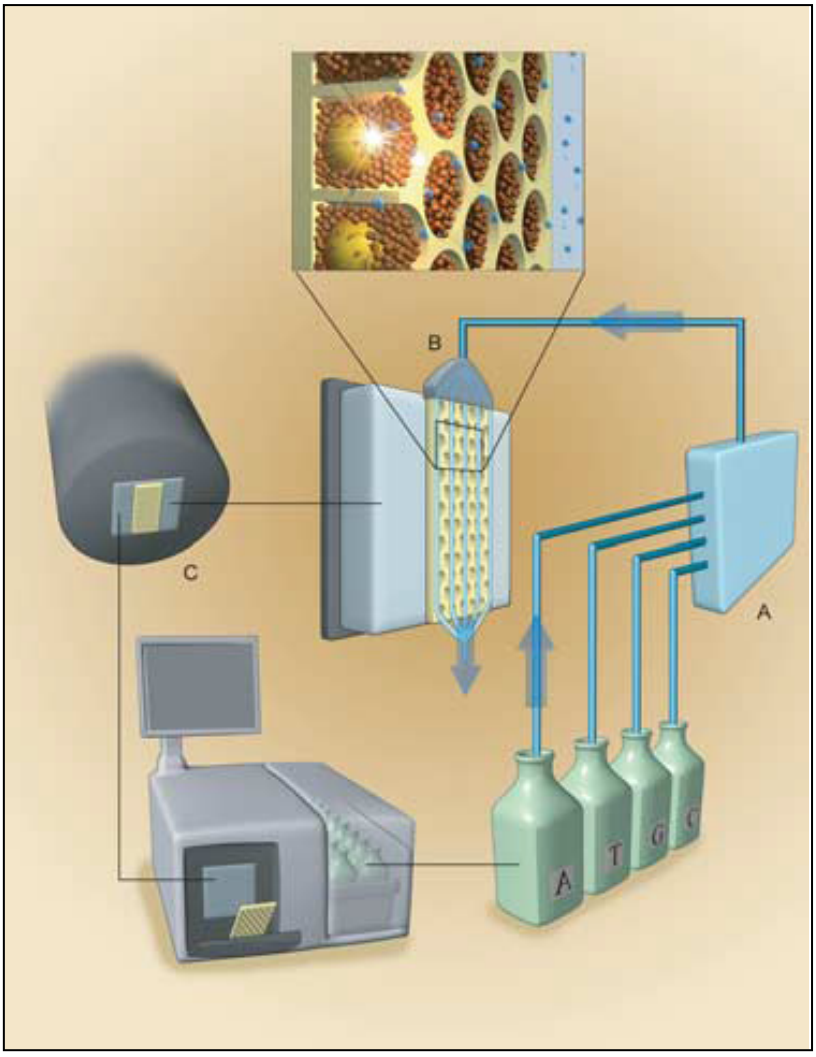
Η αλληλούχιση Sanger παρέμεινε ως η κυρίαρχη μέθοδος αλληλούχισης για τρεις ολόκληρες δεκαετίες [23] μετά την εισαγωγή της στο ρεπερτόριο των πειραματικών τεχνικών, αλλά η ανάγκη για οι απαιτήσεις για μεγαλύτερες αποδόσεις και μικρότερο κόστος αλληλούχισης αυξάνονταν όλο και περισσότερο, καθώς συνιστούσε ουσιαστικά το πρώτο βήμα για τον απώτερο σκοπό της αποκρυπτογράφησης του ανθρώπινου γονιδιώματος. Η μέθοδος βελτιώθηκε αισθητά από την πρώτη εμφάνισή της, με την άμεση καταγραφή των βάσεων σε ηλεκτρονικό υπολογιστή και με τις πιο πρόσφατες παραλλαγές της να έχουν ενοποιήσει τις τέσσερις διαφορετικές αντιδράσεις σε μία χρησιμοποιώντας ιχνηθετημένα ddNTPs που κατέστησαν εφικτή την αυτοματοποίηση της τελικής ανίχνευσης της αλληλουχίας [24].

Παρ' όλες τις βελτιώσεις στην μεθοδολογία Sanger, οι πρώτες πραγματικές ριζοσπαστικές αλλαγές στην τεχνολογία αλληλούχισης προήλθαν το 2005 από την εταιρεία 454 Life Sciences, στην οποία αναπτύχθηκε μία τεχνολογία «αλληλούχισης μέσω σύνθεσης» [25] που βασιζόταν στην ποσοτική ενίσχυση(amplification) των θραυσμάτων DNA πριν την αλληλούχιση τους και στην αξιοποίηση φωτοχημικών φαινομένων για την καταγραφή των βάσεων, που προστίθενται κατά το στάδιο επιμήκυνσης συμπληρωματικών αλυσίδων. Με την τεχνική αυτή, τα θραύσματα στα οποία σπάνε τα μόρια DNA συνδέονται ομοιοπολικά με εξειδικευμένες αλληλουχίες «προσαρμογείς» (adapter sequences) στα δύο άκρα τους ενώ ταυτόχρονα προκαλείται σπάσιμο των πλευρικών δεσμών μεταξύ των βάσεων χωρίζοντας τα έτσι σε μονόκλωνες αλυσίδες. Με τη χρήση μικροσφαιριδίων, με ακινητοποιημένες πάνω τους συμπληρωματικές αλληλουχίες των αντίστοιχων αλληλουχιών προσαρμογέων, επιτυγχάνεται η σύνδεση των θραυσμάτων πάνω τους με τις συνθήκες της αντίδρασης να ευνοούν τη σύνδεση ενός μόνο θραύσματος ανά μικροσφαιρίδιο. Έπειτα πάνω στο κάθε σφαιρίδιο λαμβάνει χώρα ποσοτική ενίσχυση της του θραύσματος με τη μέθοδο της αλυσιδωτής αντίδρασης πολυμεράσης σε γαλάκτωμα (emulsion Polymerase Chain Reaction - emulsion PCR). Στη συνέχεια τα μικροσφαιρίδια που φέρουν τις πολλαπλασιασμένες αλληλουχίες-κλώνους οδηγούνταν σε μικροπηγάδια, με τέτοιο τρόπο ώστε να καταλήγει μόνο ένα μικροσφαρίδιο ανά μικροπηγάδι. Τέλος, στο κάθε μικροπηγάδι προστίθενται μικρότερα μικροσφαιρίδια που φέρουν πάνω τους τα απαραίτητα ένζυμα ώστε να λάβουν χώρα οι αντιδράσεις πυροαλληλούχισης (Εικόνα 6).

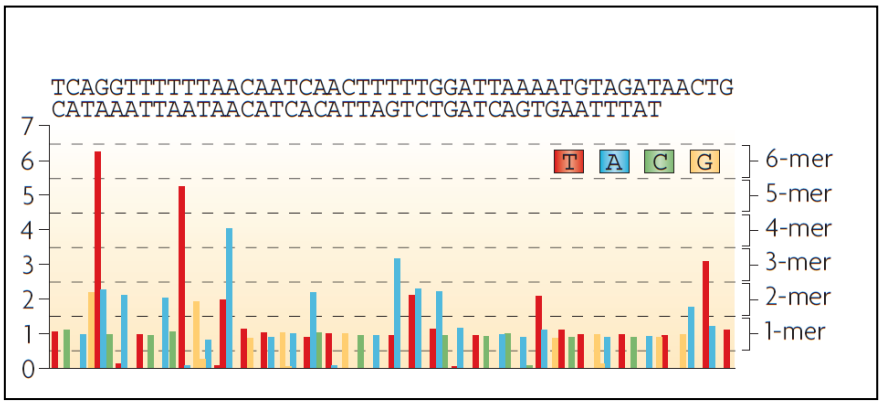


Εικόνα 6. Μέθοδος αλληλούχισης μέσω σύνθεσης (454 πυροαλληλούχιση) - κατακερματισμός DNA και ποσοτική ενίσχυση θραυσμάτων. (A)Το DNA απομονώνεται, σπάει σε θραύσματα, ενώνεται με αλληλουχίες «προσαρμογείς» και διαχωρίζεται σε μονόκλωνες αλυσίδες. (B)Τα θραύσματα προσκολλούνται σε μικροσφαιρίδια υπό συνθήκες που επιτρέπουν μόνο ένα θραύσμα ανά σφαιρίδιο και στη συνέχεια λαμβάνει χώρα αντίδραση PCR σε γαλάκτωμα. (C)Τα μικροσφαιρίδια με τις πολλαπλασιασμένες αλληλουχίες-κλώνους εισάγονται σε μικροπηγάδια. (D)Μικρότερα μικροσφαιρίδια με ακινητοποιημένα ένζυμα για την αντίδραση πυροαλληλούχισης εισάγονται στα μικροπηγάδια [25].

Η κάθε αντίδραση ξεκινάει με την προσθήκη, σε όλα τα μικροπηγάδια ταυτόχρονα, ενός εκ των τεσσάρων νουκλεοτιδίων (A, C, G, T). Αυτό έχει ως αποτέλεσμα να επιμηκύνονται οι συμπληρωματικές αλυσίδες, μόνο σε όσα θραύσματα είχαν την αντίστοιχη συμπληρωματική βάση στην εκάστοτε ελεύθερη θέση. Η προσθήκη νουκλεοτιδίων γίνεται με διαδοχικό τρόπο και κάθε φορά που μία ομάδα αλυσίδων, πάνω σε κάποιο μικροσφαιρίδιο, επιμηκύνεται λόγω συμπληρωματικότητας, γίνεται καταγραφή των φωτονίων που παράγονται από το συγκεκριμένο μικροπηγάδι λόγω απελευθέρωσης ανόργανου πυροφωσφορικού οξέος. Με την καταγραφή αυτή, είναι δυνατός ο προσδιορισμός της εκάστοτε βάσης που προστίθεται σε κάθε θραύσμα και με τις εναλλασσόμενες καταγραφές είναι δυνατή η ταυτόχρονη αλληλούχιση των θραυσμάτων όλων των μικροπηγαδιών (Εικόνες 7-8).



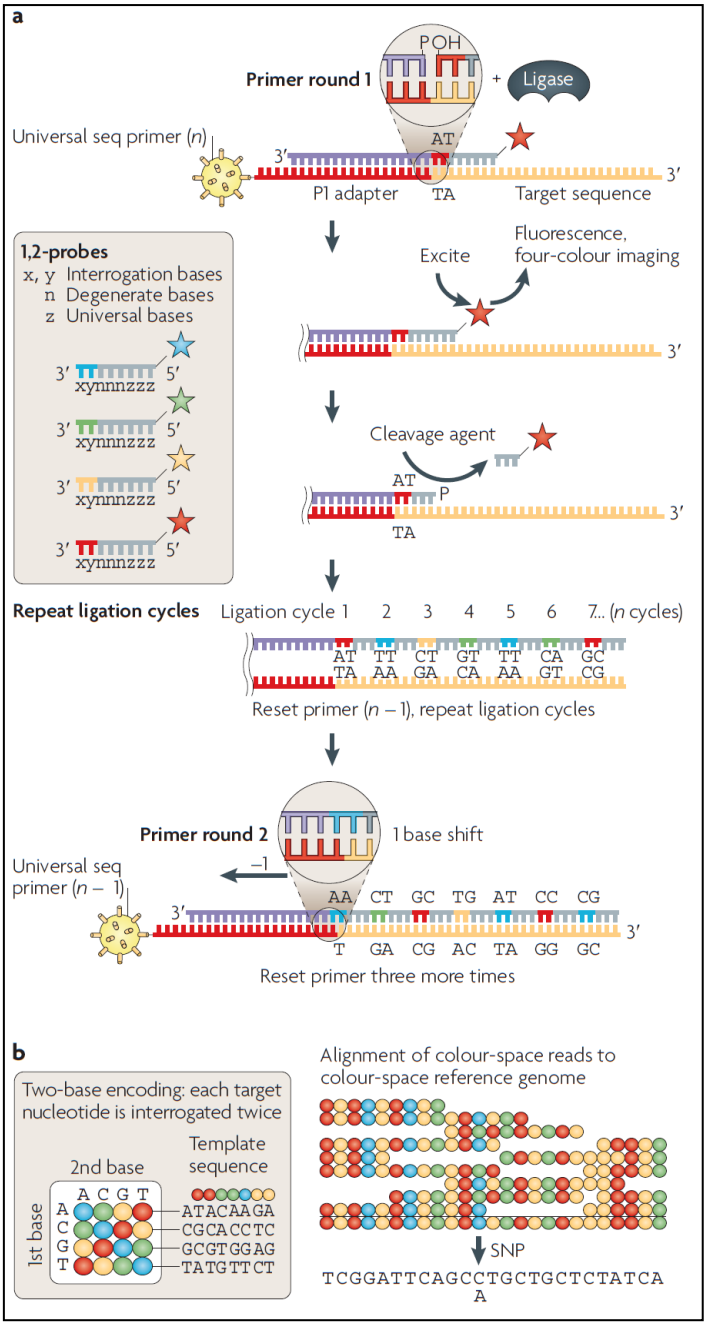
Εικόνα 7. Μέθοδος αλληλούχισης μέσω σύνθεσης (454 πυροαλληλούχιση) - συσκευή αλληλούχισης. Το μηχάνημα αλληλούχισης αποτελείται από τα εξής υποσυστήματα: (A)ένα θάλαμο ροής των τεσσάρων dNTPs που περιλαμβάνει την πλάκα με τα μικροπηγάδια, (B)μία διάταξη καταγραφής εικόνας με ανιχνευτή διάταξης συζευγμένου φορτίου (charged-coupling device - CCD) για τον προσδιορισμό των βάσεων μέσω ανίχνευσης φωτός και (C)έναν υπολογιστή με το κατάλληλο λογισμικό ελέγχου [25].



Εικόνα 8. Μέθοδος αλληλούχισης μέσω σύνθεσης (454 πυροαλληλούχιση) - διάγραμμα ροής αλληλούχισης (flowgram) ενός μικροπηγαδιού. Η συσκευή αλληλούχισης καταγράφει το φως που παράγεται σε κάθε διαδοχική προσθήκη ενός είδους dNTP, ως κορυφές στο διάγραμμα ροής και προσδιορίζει τη βάση που έχει προστεθεί κατά την αντίδραση πολυμερισμού της συμπληρωματικής αλυσίδας. Μεγαλύτερη ένταση φωτός (μεγαλύτερες κορυφές) σημαίνει ότι παραπάνω από μία βάσεις έχουν προστεθεί και με σωστή διαβάθμιση του οργάνου μπορεί να υπολογιστεί ο αριθμός τους [26].

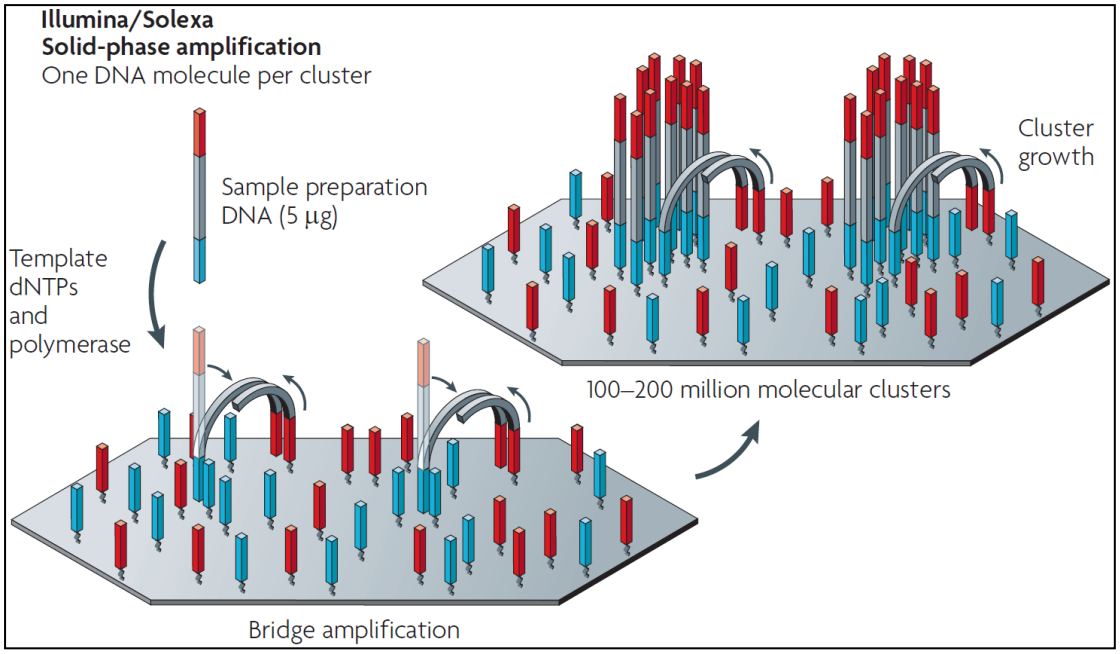
Η νέα μέθοδος επέτρεπε την ταυτόχρονη ανάλυση εκατοντάδων χιλιάδων προτύπων DNA (templates) εν παραλλήλω, αυξάνοντας έτσι τον αριθμό αντιδράσεων αλληλούχισης πάνω από χίλιες φορές σε σχέση με τις ως τότε σύγχρονες μεθόδους Sanger αλληλούχισης. Το παράδειγμα ποσοτικής ενίσχυσης των θραυσμάτων DNA που εισήγαγε η 454 Life Sciences ακολουθήθηκε και από τις υπόλοιπες ανταγωνίστριες εταιρείες, με διάφορες παραλλαγές, με αποτέλεσμα να επικρατήσουν τρεις διαφορετικές τεχνολογίες αλληλούχισης νέας (ή δεύτερης) γενιάς: α)Η τεχνολογία πυροαλληλούχισης της 454 Life Sciences [25], β)η τεχνολογία αντιστρέψιμων κωδικονίων λήξης (reversible terminators) της Illumina/Solexa[27] και γ)η τεχνολογία αλληλούχισης με δέσμευση DNA (ligation) της SOLiD[28].

Η τεχνολογία αλληλούχισης SOLiD διατηρεί τη μέθοδο ποσοτικής ενίσχυσης θραυσμάτων μέσω PCR σε γαλάκτωμα, αλλά εισάγει μία καινούρια μέθοδο καταγραφής των βάσεων. Η μέθοδος αυτή βασίζεται στη χρήση ειδικά κατασκευασμένων ολιγονουκλεοτιδικών αλληλουχιών που είναι ιχνηθετημένες και ονομάζονται 1,2-ανιχνευτές (1,2-probes) καθώς η ιχνηθέτησή τους καθορίζεται από το αρχικό ζεύγος βάσεών τους. Κατά την αντίδραση αλληλούχισης, δεν γίνεται πολυμερισμός της συμπληρωματικής αλληλουχίας, αλλά δέσμευση των 1,2-ανιχνευτών με τη βοήθεια ενζύμων (λιγασών), ξεκινώντας από τον εκκινητή που είναι συζευγμένος στο κάθε θραύσμα, με βάση τη συμπληρωματικότητα του αρχικού ζεύγους βάσεών τους (Εικόνα 9a). Ο προσδιορισμός των βάσεων γίνεται με την καταγραφή των διαφορετικών σημάτων, που λαμβάνονται από τον κάθε διαφορετικά ιχνηθετημένο 1,2-ανιχνευτή κατά τη δέσμευσή του μέσω λιγασών. Η διαδικασία δέσμευσης 1,2-ανιχνευτών επαναλαμβάνεται πέντε φορές, χρησιμοποιώντας διαφορετικό εκκινητή κάθε φορά που διαφέρει από τον προηγούμενο κατά ένα νουκλεοτίδιο έτσι ώστε να επιτευχθεί η «σάρωση» του συνόλου των βάσεων της αλληλουχίας του θραύσματος (Εικόνα 9b).

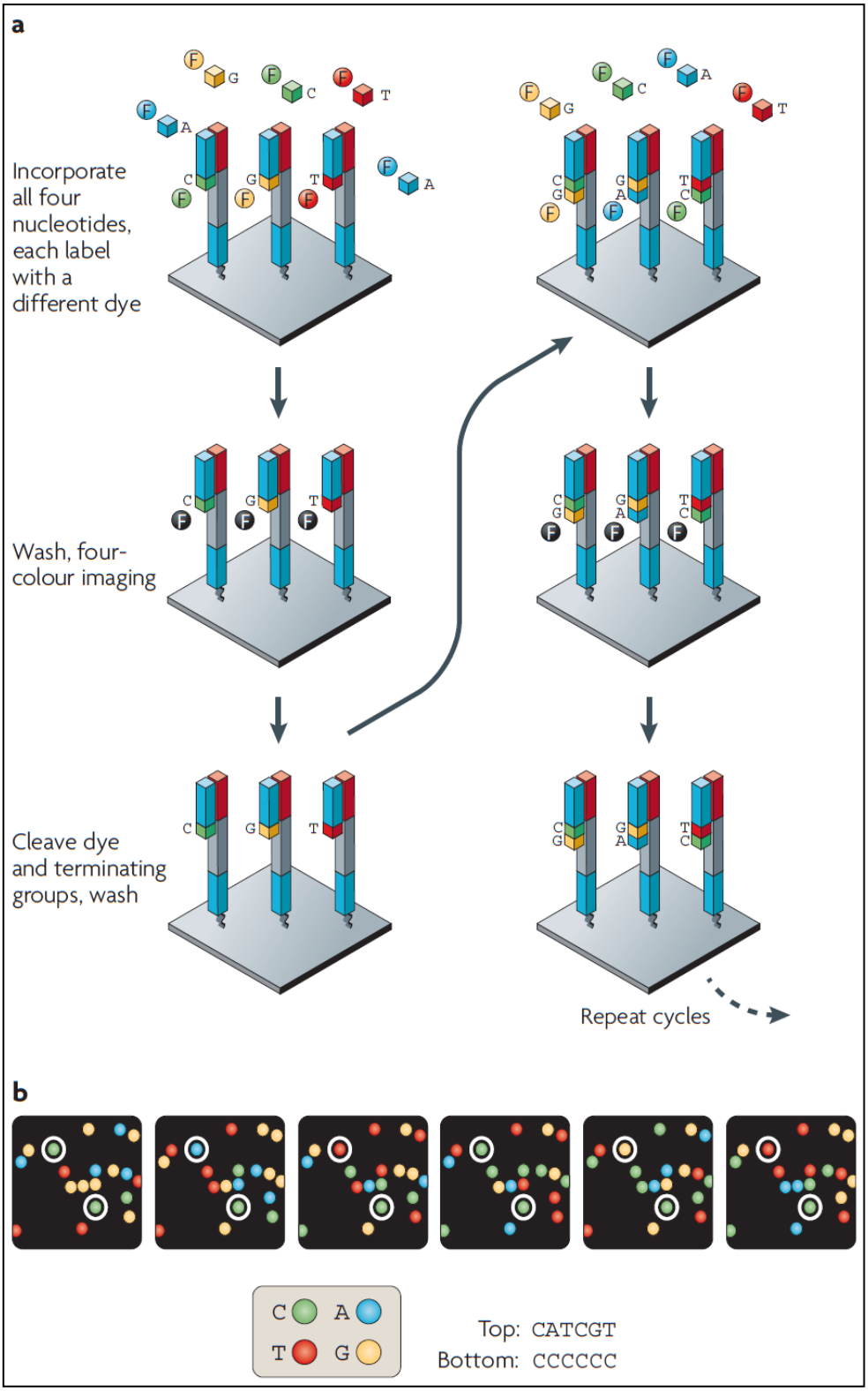


Εικόνα 9. Μέθοδος αλληλούχισης με δέσμευση DNA (SOLiD) - προσδιορισμός βάσεων. a)Μετά την ποσοτική ενίσχυση των θραυσμάτων με PCR με γαλάκτωμα προστίθεται μία βιβλιοθήκη με 1,2-ανιχνευτές σε κάθε μικροσφαιρίδιο. Σε αντίθεση με τις αντιδράσεις πολυμερισμού η σύνδεση, με τη βοήθεια λιγασών, ενός ανιχνευτή με την αλληλουχία του εκκινητή μπορεί να γίνει με οποιαδήποτε κατεύθυνση, είτε από το 5' (-PO_4) είτε από το 3' (-OH) άκρο του αλλά οι συνθήκες της αντίδρασης επιτρέπουν τη σύνδεση μόνο αν ευνοείται από την συμπληρωματικότητα του ζεύγους βάσεων στην αρχή του (3' άκρο). Μετά τη σύνδεση των ανιχνευτών καταγράφεται το φθορίζον σήμα τεσσάρων χρωμάτων από το σύνολο των μικροσφαιριδίων και οι συνδεδεμένοι ανιχνευτές κόβονται χημικά με ιόντα αργύρου ώστε να καταλήξουν με ένα ελεύθερο άκρο φωσφωρικής ομάδας (-PO_4). Ο συγκεκριμένος κύκλος αντιδράσεων επαναλαμβάνεται 9 φορές επιπλέον (σύνολο 10) ή περισσότερες ανάλογα το ζητούμενο μήκος των τελικών αναγνωσμένων αλληλουχιών. Στη συνέχεια ο, επιμηκυμένος πλέον, εκκινητής αφαιρείται και στη θέση του προστίθεται ένας καινούριος μικρότερος κατά ένα νουκλεοτίδιο από τον αρχικό. Η διαδικασία υβριδισμού του εκκινητή (primer round) και των κύκλων σύνδεσης, μέσω λιγασών, των 1,2-ανιχνευτών κατά αυτό τον τρόπο επαναλαμβάνεται άλλες 4 φορές (σύνολο 5). Οι 1,2-ανιχνευτές είναι σχεδιασμένοι έτσι ώστε να εντοπίζουν την πρώτη (x) και δεύτερη (y) θέση μετά τον εκκινητή κατά τέτοιο τρόπο ώστε τα 16 δινουκλεοτίδια xy να κωδικοποιούνται από 4 χρωστικές (χρωματισμένα αστέρια). Οι 1,2-ανιχνευτές περιέχουν επίσης και βάσεις ινοσίνης (z) ώστε να μειώσουν την πολυπλοκότητα της βιβλιοθήκης, ενώ το πέμπτο και έκτο νουκλεοτίδιό τους συνδέονται με έναν φωσφοθειικό (phosphorothiolate) δεσμο (αντί του αντίστοιχο φωσφοδιεστερικού) ο οποίος μπορεί να σπάσει με ιόντα αργύρου. b)Χρησιμοποιείται μία δινουκλεοτιδική κωδικοποίηση για την ερμηνεία των σημάτων από κάθε σύνδεση των 1,2-ανιχνευτών. Κάθε βάση του αρχικού θραύσματος «σαρώνεται» δύο φορές και οι αναγνωσμένες αλληλουχίες αποθηκεύονται ως δεδομένα χρωματικού χώρου (color-space data bits) [26].

Αντίθετα, η τεχνολογία αλληλούχισης Illumina/Solexa χρησιμοποιεί εντελώς νέες μεθόδους, τόσο για το στάδιο ποσοτικής ενίσχυσης των θραυσμάτων DNA (Εικόνα 10), όσο και για τον τελικό προσδιορισμό των βάσεων που προστίθενται κατά την αντίδραση αλληλούχισης (Εικόνα 11). Η μέθοδος αυτή χρησιμοποιεί την ακινητοποίηση όλων των θραυσμάτων σε στερεό υπόστρωμα, όπου ακολουθεί ποσοτική ενίσχυση γέφυρας (bridge amplification). Κατά την αντίδραση αυτή ο πολυμερισμός συμβαίνει με ταυτόχρονη σύνδεση και των δύο άκρων κάθε θραύσματος σε γειτονικές αλληλουχίες-προσαρμογείς με τελικό αποτέλεσμα τον σχηματισμό ενός συσσωματώματος αλληλουχιών-κλώνων ανά θραύσμα. Ο εντοπισμός των βάσεων των θραυσμάτων ανά συσσωμάτωμα γίνεται με χρήση ειδικά κατασκευασμένων dNTPs ώστε να σταματάνε την αντίδραση πολυμερισμού σε κάθε κύκλο, καθώς και ιχνηθετημένων με τέσσερις διαφορετικές φωσφορίζουσες ομάδες ώστε να είναι δυνατή η ταυτόχρονη καταγραφή τους.



Εικόνα 10. Μέθοδος αλληλούχισης Solexa/Illumina - ποσοτική ενίσχυση θραυσμάτων στερεάς φάσης (solid-phase amplification). Η μέθοδος ποσοτικής ενίσχυσης στερεάς φάσης αποτελείται από δύο στάδια. Το πρώτο στάδιο αφορά την ακινητοποίηση των μονόκλωνων θραυσμάτων (γκρι χρώμα) σε στερεό υπόστρωμα χάρις τις αλληλουχίες-προσαρμογείς (κόκκινο και μπλε χρώμα) που έχουν συνδεθεί στα άκρα τους και δρουν και ως εκκινητές. Η ακινητοποίηση των θραυσμάτων γίνεται παράλληλα με την προσθήκη DNA πολυμεράσης και dNTPs ώστε να αρχίσει η αντίδραση πολυμερισμού των συμπληρωματικών τους κλώνων. Στο δεύτερο στάδιο γίνεται ποσοτική ενίσχυση γέφυρας (bridge amplification) του κάθε θραύσματος με την κάθε αλυσίδα-κλώνο να προσκολλάται σε κάποιο γειτονικό ολιγονουκλεοτίδιο-προσαρμογέα κατά τη διάρκεια της αντίδρασης πολυμερισμού. Κατά αυτόν τον τρόπο δημιουργούνται συσσωματώματα (clusters) κλώνων από κάθε θραύσμα πάνω στο στερεό υπόστρωμα που δίνουν ισχυρότερο σήμα κατά την αντίδραση αλληλούχισης [26].



Εικόνα 11. Μέθοδος αλληλούχισης Solexa/Illumina - προσδιορισμός βάσεων. a)Η αντίδραση επιμήκυνσης συμπληρωματικών αλυσίδων χρησιμοποιεί ειδικά κατασκευασμένα dNTPs συνδεδεμένα με μία ομάδα -O-αζιδομεθυλίου (-O-N_3) στο άκρο τους καθώς και ιχνηθετημένα με τέσσερις διαφορετικές φωσφορίζουσες ομάδες (Ⓕ) . Η ομάδα αζιδομεθυλίου σταματά την αντίδραση πολυμερισμού μετά την προσθήκη τους και η φωσφορίζουσες ομάδες επιτρέπουν την καταγραφή σήματος εικόνας για κάθε συσσωμάτωμα θραυσμάτων-κλώνων. Την καταγραφή ακολουθεί ενζυμικό κόψιμο και απομάκρυνση των φωσφοριζουσών ομάδων και των αζιδομεθυλίων αφήνοντας μία ομάδα -OH στην 3' θέση των ελεύθερων νουκλεοτιδίων, ώστε να μπορέσει να συνεχιστεί η αντίδραση πολυμερισμού με την προσθήκη dNTPs στον επόμενο κύκλο αντιδράσεων. b)Ο προσδιορισμός των βάσεων των θραυσμάτων γίνεται με διαδοχική καταγραφή εικόνων για κάθε συσσωμάτωμα. Τα συσσωματώματα των οποίων η αλληλουχία εξετάζεται παραπάνω είναι σημειωμένα με άσπρο κύκλο [26].

Τα πλεονεκτήματα των νέων τεχνολογιών έγιναν εμφανή εξ' αρχής, καθώς οι πολύ υψηλές αποδόσεις τους συνδυάζονταν με σημαντική μείωση του κόστους κάθε πειράματος[23]. Παρ' όλα αυτά, η εξέλιξη των τεχνικών αλληλούχισης σε επίπεδο πειραματικής διαδικασίας, δεν σήμαινε a priori την βελτιστοποίηση εξ ολοκλήρου της αναλυτικής μεθόδου. Τα μηχανήματα αλληλούχισης δεύτερης γενιάς παρήγαγαν δεδομένα που περιελάμβαναν αναγνωσμένες αλληλουχίες μικρότερου μεγέθους από τις κλασσικές τεχνικές. Οι αναγνωσμένες αλληλουχίες των ~1000 βάσεων που προέκυπταν από την αλληλούχιση Sanger είχαν πλέον αντικατασταθεί από αντίστοιχες αλληλουχίες μικρότερες των 100 βάσεων. Ταυτόχρονα, η υπερκάλυψη του μειονεκτήματος αυτού, μέσω της υψηλής απόδοσης των συσκευών αλληλούχισης νέας γενιάς, οδήγησε στην αύξηση του πλήθους των αναγνωσμένων αλληλουχιών κατά πολλές τάξεις μεγέθους με τον τελικό τους αριθμό να αγγίζει πλέον δεκάδες εκατομμύρια ανά πείραμα. Αυτό είχε ως αποτέλεσμα τα δεδομένα αλληλούχισης που προκύπτουν να είναι πολύ μεγαλύτερου όγκου και πολυπλοκότητας, καθιστώντας έτσι τη βιοπληροφορική ανάλυσή τους ένα εξαιρετικά δυσεπίλυτο πρόβλημα.

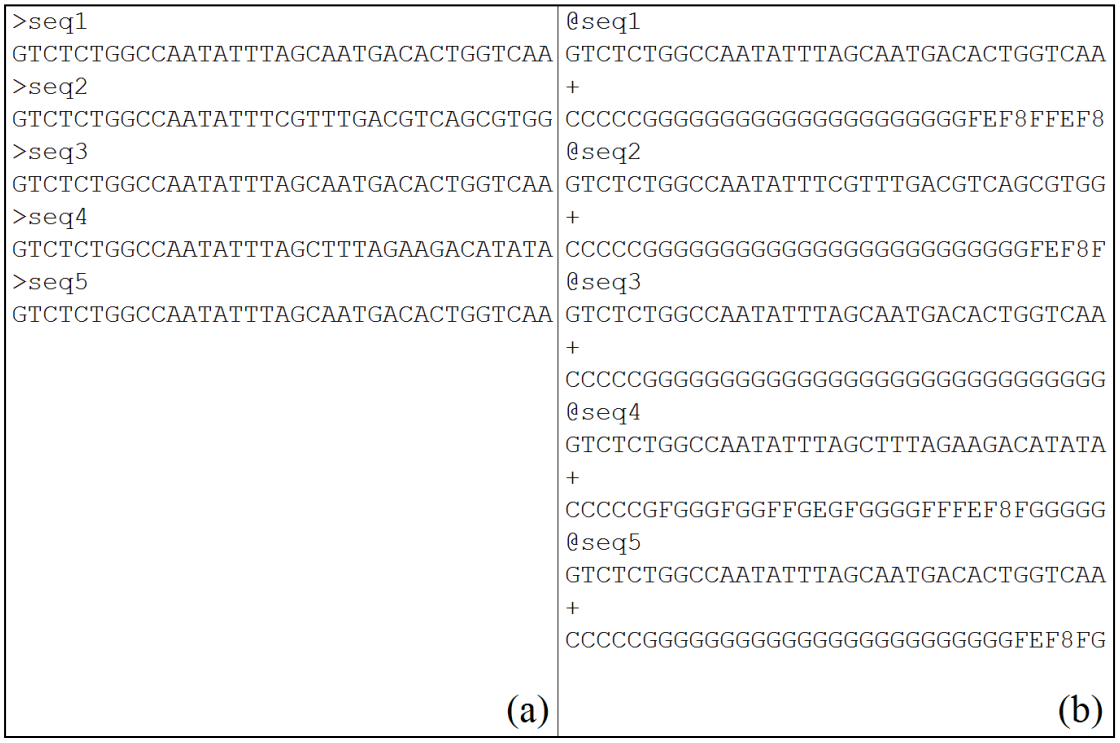
Με την προοπτική να λυθούν τα προβλήματα πολυπλοκότητας των δεδομένων, να επιτευχθούν μεγαλύτερες αποδόσεις αλληλούχισης και να μειωθεί η απαραίτητη ποσότητα αρχικού DNA για την ανάλυση αναπτύχθηκαν μεθοδολογίες μονομοριακής αλληλούχισης (single-molecule DNA sequencing)[29] με πρωτεργάτες τις εταιρείες Helicos[30] και PacBio[31]. Αυτές οι τεχνολογίες ονομάστηκαν τεχνολογίες αλληλούχισης τρίτης γενιάς[32] αλλά η αξιοποίησή τους από την επιστημονική κοινότητα ήταν σχετικά περιορισμένη έως πρόσφατα εξ' αιτίας της μεγαλύτερης προδιάθεσης που έχουν σε λάθη κατά την καταγραφή των βάσεων σε σχέση με τις τεχνολογίες δεύτερης γενιάς[33].

1.3 Βιοπληροφορική ανάλυση μεταγονιδιωματικών δεδομένων

1.3.1 Ποιοτικός έλεγχος δεδομένων

Παρ' όλη τη δραματική εξέλιξη που έχει σημειωθεί στις τεχνολογίες αλληλούχισης και την εμφάνιση νέων μεθοδολογιών, καμία από τις υπάρχουσες τεχνικές ως τώρα δεν έχει τη δυνατότητα απευθείας αλληλούχισης ολόκληρου του γονιδιώματος ενός οργανισμού, κάτι που επεκτείνεται κατά την προσπάθεια αλληλούχισης πολλαπλών γονιδιωμάτων σε ένα μεταγενωμικό δείγμα. Αντ' αυτού, οι μέθοδοι αλληλούχισης περιλαμβάνουν τον κατακερματισμό των μορίων DNA σε τυχαία σημεία και τη δημιουργία θραυσμάτων τα οποία υπόκεινται τελικά σε

ανάλυση αλληλούχισης. Τα δεδομένα που προκύπτουν περιλαμβάνουν εκατοντάδες χιλιάδες, ή ακόμα και εκατομμύρια αναγνωσμένες αλληλουχίες, αποθηκευμένες σε ειδικά διαμορφωμένα αρχεία (Εικόνα 12a), από την βιοπληροφορική επεξεργασία των οποίων μπορεί να προκύψει η αρχική αλληλουχία από την οποία προήλθαν καθώς και λειτουργικές πληροφορίες για τον μικροβιακό πληθυσμό του δείγματος.



Εικόνα 12. Δείγμα δεδομένων αλληλούχισης σε μορφή FASTA και FASTQ. (a)Τα δεδομένα αφορούν πέντε αναγνωσμένες αλληλουχίες μήκους 35 βάσεων σε μορφή αρχείου FASTA. Κάθε γραμμή που αρχίζει από τον χαρακτήρα ">" αποτελεί το αναγνωριστικό κωδικό της αλληλουχίας που βρίσκεται στην επόμενη γραμμή. (b)Τα δεδομένα αφορούν τις ίδιες πέντε αλληλουχίες αλλά σε μορφή αρχείου FASTQ. Σε αυτό το είδος αρχείου δεδομένων ο αναγνωριστικός κωδικός κάθε αλληλουχίες ξεκινάει με τον χαρακτήρα "@". Σε κάθε νουκλεοτιδική αλληλουχία αντιστοιχεί μία αλληλουχία σκορ ποιότητας σε κωδικοποίηση ASCII. Οι γραμμές μεταξύ νουκλεοτιδικής αλληλουχίας και των αντίστοιχων σκορ ποιότητας που αντιστοιχούν σε κάθε βάση της διαχωρίζονται μέσω μιας γραμμής που περιέχει μόνο τον χαρακτήρα "+".

Δεδομένης της πολυπλοκότητας και της άμεσης εξάρτησης κάθε περαιτέρω ανάλυσης από τα αρχικά δεδομένα αλληλούχισης, γίνεται εμφανής η σημαντικότητα της εξασφάλισης της ορθότητας τους. Κατά τον ποιοτικό έλεγχο των δεδομένων αλληλούχισης εξετάζονται συγκεκριμένα χαρακτηριστικά [34] των αναγνωσμένων αλληλουχιών, που υποδεικνύουν κατά πόσο είναι αξιόπιστη η καταγραφή της κάθε βάσης. Υπάρχουν πλέον εξειδικευμένα βιοπληροφορικά εργαλεία ποιοτικού ελέγχου [34-36] των οποίων οι αλγόριθμοι προσφέρουν και δυνατότητες επεξεργασίας των δεδομένων αλληλούχισης, αφαιρώντας είτε ολόκληρες αλληλουχίες για τις οποίες

βρέθηκαν σφάλματα καταγραφής, ή μερικές βάσεις αν τα σφάλματα αυτά είναι περιορισμένα σε συγκεκριμένη περιοχή κάθε αναγνωσμένης αλληλουχίας.

Το κάθε χαρακτηριστικό που εξετάζεται από τα εργαλεία ποιοτικού ελέγχου, επηρεάζει διαφορετικά τόσο την ποιότητα των δεδομένων όσο και το είδος σφαλμάτων που εμφανίζονται σε αυτά και μπορούν να οδηγήσουν σε εσφαλμένα συμπεράσματα κατά την ανάλυσή τους. Για παράδειγμα, ένα σημαντικό χαρακτηριστικό τέτοιου είδους αποτελεί η περιεκτικότητα μίας περιοχής αλληλουχίας σε γουανίνη/κυτοσίνη καθώς συσχετίζεται άμεσα με το πλήθος των αναγνωσμένων αλληλουχιών που προκύπτουν από αυτήν. Αυτό οδηγεί σε εσφαλμένα συμπεράσματα ποσοτικοποίησης των συγκεκριμένων περιοχών και των μικροοργανισμών στους οποίους ανήκουν. Συγκεκριμένα οι περιοχές με αυξημένη περιεκτικότητα σε γουανίνη/κυτοσίνη επιδεικνύουν ασύμμετρη αύξηση στον αριθμό των αναγνωσμένων αλληλουχιών που προκύπτουν από αυτές με το πρόβλημα να επιδεινώνεται όταν κατά την αλληλούχιση περιλαμβάνεται κάποιο PCR βήμα ποσοτικής ενίσχυσης [37]. Άλλο χαρακτηριστικό εξίσου σημαντικό είναι το σκορ ποιότητας (quality score) κάθε βάσης. Τα συγκεκριμένα στατιστικά σκορ προκύπτουν βιοπληροφορικά αξιοποιώντας τα «μεταδεδομένα» (metadata), που προκύπτουν από την εκάστοτε συσκευή αλληλούχισης. Τα μεταδεδομένα των σκορ ποιότητας ανά βάση καταγράφονται συνήθως μαζί με τα δεδομένα αλληλούχισης (Εικόνα 12b) σε κωδικοποιημένη μορφή ASCII (Εικόνα 13). Η διαβάθμισή τους ακολουθεί το λογαριθμικό μοντέλο των σκορ ποιότητας Phred σύμφωνα με το οποίο προσδιορίζεται η πιθανότητα μίας βάσης να έχει καταγραφεί εσφαλμένα (Εικόνα 14). Η ποιότητα των βάσεων είναι συνήθως λίγο πιο χαμηλή στις πρώτες βάσεις κάθε αναγνωσμένης αλληλουχίας και αρκετά πιο χαμηλή στις τελευταίες βάσεις [38] ενώ εξαρτάται από το μέγεθος των αλληλουχιών καθώς και από την τεχνολογία αλληλούχισης.

Άλλα χαρακτηριστικά ποιότητας αποτελούν η κατανομή πολυπλοκότητας αλληλουχιών (sequence complexity distributions), η ύπαρξη πανομοιότυπων αλληλουχιών (sequence duplication) καθώς και η ύπαρξη τεχνητών αλληλουχιών (artifacts) [34]. Η εξέταση όλων των παραπάνω στατιστικών σκορ, που παρέχονται από τα αντίστοιχα βιοπληροφορικά εργαλεία, είναι απαραίτητη τόσο για την αξιολόγηση των δεδομένων, όσο και για την επιλογή της καταλληλότερης μεθόδου περαιτέρω επεξεργασίας τους π.χ. την περικοπή (trimming) των άκρων χαμηλής

ποιότητας, ώστε να θεωρούνται αρκετά αξιόπιστα για οποιαδήποτε μετέπειτα ανάλυση.

ASCII	Symbol	ASCII	Symbol	ASCII	Symbol	ASCII	Symbol
0	NUL	16	DLE	32	(space)	48	0
1	SOH	17	DC1	33	!	49	1
2	STX	18	DC2	34	"	50	2
3	ETX	19	DC3	35	#	51	3
4	EOT	20	DC4	36	$	52	4
5	ENQ	21	NAK	37	%	53	5
6	ACK	22	SYN	38	&	54	6
7	BEL	23	ETB	39	'	55	7
8	BS	24	CAN	40	(	56	8
9	TAB	25	EM	41	)	57	9
10	LF	26	SUB	42	*	58	:
11	VT	27	ESC	43	+	59	;
12	FF	28	FS	44	,	60	<
13	CR	29	GS	45	-	61	=
14	SO	30	RS	46	.	62	>
15	SI	31	US	47	/	63	?
ASCII	**Symbol**	**ASCII**	**Symbol**	**ASCII**	**Symbol**	**ASCII**	**Symbol**
64	@	80	P	96	`	112	p
65	A	81	Q	97	a	113	q
66	B	82	R	98	b	114	r
67	C	83	S	99	c	115	s
68	D	84	T	100	d	116	t
69	E	85	U	101	e	117	u
70	F	86	V	102	f	118	v
71	G	87	W	103	g	119	w
72	H	88	X	104	h	120	x
73	I	89	Y	105	i	121	y
74	J	90	Z	106	j	122	z
75	K	91	[	107	k	123	{
76	L	92	\	108	l	124	|
77	M	93	]	109	m	125	}
78	N	94	^	110	n	126	~
79	O	95	_	111	o	127	▯

Εικόνα 13. Πίνακας μετατροπής συστήματος ASCII. Το σύστημα ASCII κωδικοποιεί 128 χαρακτήρες με ακέραιους αριθμούς από το 0 έως το 127 [39].

Phred quality score	Probability that the base is called wrong	Accuracy of the base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

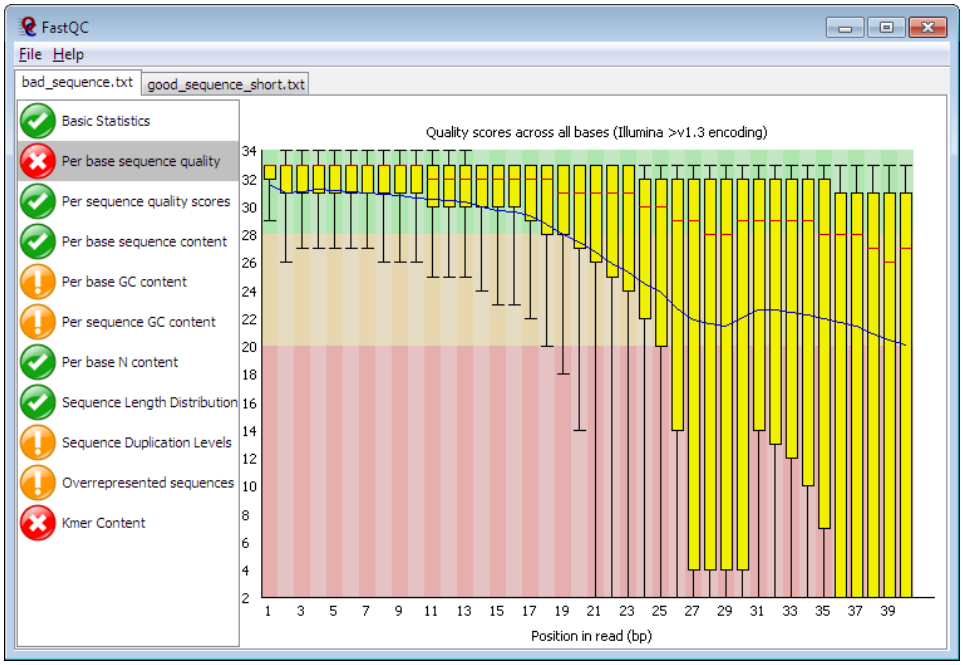
Εικόνα 14. Πίνακας πιθανοτήτων σκορ ποιότητας Phred. Τα σκορ ποιότητας Phred προκύπτουν από τον τύπο $Q=-10\log_{10}P$ όπου Q είναι το σκορ ποιότητας και P η πιθανότητα λάθους καταγραφής βάσης.

Όσον αφορά ήδη υπάρχοντες αλγορίθμους ποιοτικού ελέγχου το εργαλείο FastQC [40] αποτελεί μία ολοκληρωμένη λύση που εξετάζει ένα πλήθος χαρακτηριστικών ποιότητας, ώστε να αξιολογήσει την εγκυρότητα των αναγνωσμένων αλληλουχιών και να παράγει λεπτομερείς αναφορές με τα αποτελέσματα της ανάλυσης. Το εργαλείο δέχεται ως δεδομένα εισόδου αρχεία μορφής FASTQ (Εικόνα 12b) αλλά δεν περιορίζεται εκεί. Μπορεί να επεξεργαστεί και άλλων τύπων αρχεία όπως SAM/BAM [41] τα οποία περιέχουν τις αναγνωσμένες αλληλουχίες και τα σκορ ποιότητας που αντιστοιχούν σε κάθε βάση τους. Κατά την ανάλυση ποιοτικού ελέγχου οι αλγόριθμοι που χρησιμοποιεί εξετάζουν τα εξής χαρακτηριστικά:

- Στατιστική κατανομή βάσεων
- Σκορ ποιότητας ανά βάση και ανά αλληλουχία
- Ποσοστό γουανίνης/κυτοσίνης ανά βάση και ανά αλληλουχία
- Ποσοστό μη καταγεγραμμένων βάσεων
- Κατανομή μηκους αναγνωσμένων αλληλουχιών
- Επίπεδα ύπαρξης πανομοιότυπων αλληλουχιών
- Επίπεδα υπερεκφρασμένων αλληλουχιών
- Κατανομή πολυπλοκότητας αλληλουχιών

Το εργαλείο λειτουργεί πίσω από ένα γραφικό περιβάλλον σχεδιασμένο σε γλώσσα προγραμματισμού Java [42] και περιλαμβάνει οπτικοποίηση των αναφορών αποτελεσμάτων με τη μορφή διαγραμμάτων (Εικόνα 15).
Ένα σημαντικό πλεονέκτημα του εργαλείου FastQC είναι ότι μπορεί να επεξεργαστεί μεγάλου όγκου δεδομένα χωρίς ιδιαίτερες υπολογιστικές απαιτήσεις και είναι συμβατό τόσο με λειτουργικό Linux όσο και με Windows καθιστώντας το εύχρηστο ακόμα και σε προσωπικό υπολογιστή περιορισμένων δυνατοτήτων. Το μόνο του μειονέκτημα είναι ότι δεν περιλαμβάνει εργαλεία επεξεργασίας των δεδομένων αλληλούχισης για απομάκρυνση των προβληματικών αλληλουχιών.

Την ανάγκη επιδιόρθωσης των σφαλμάτων αλληλούχισης μέσω περαιτέρω επεξεργασίας δεδομένων μπορεί να καλύψει επιτυχώς η σουίτα εργαλείων FASTX [43] η οποία μπορεί να χρησιμοποιηθεί συνδυαστικά με τα αποτελέσματα ποιοτικού ελέγχου του FastQC.

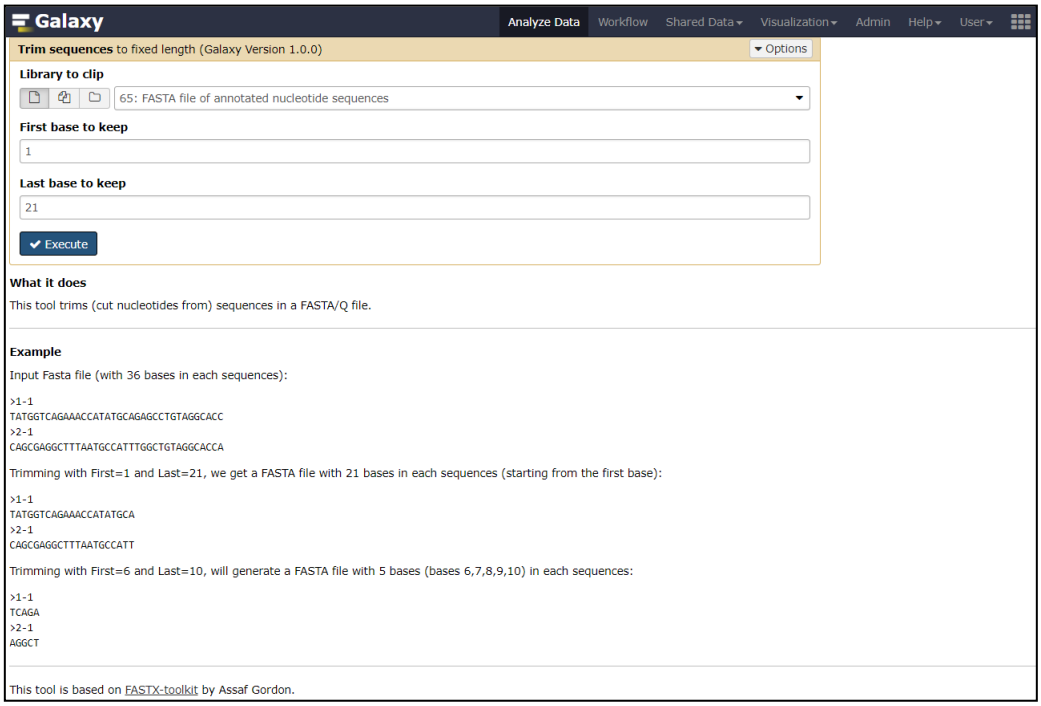


Εικόνα 15. Αναφορά αποτελεσμάτων FastQC. Η αναφορά αποτελεσμάτων του προγράμματος περιλαμβάνει διαγράμματα που περιγράφουν κάθε χαρακτηριστικό ποιότητας με επισήμανση για τα χαρακτηριστικά τα οποία δεν παρουσιάζουν προβλήματα (✓), αυτά που χρήζουν προσοχής (!) και αυτά που παρουσιάζουν σοβαρά προβλήματα ποιότητας (x). Στην παραπάνω εικόνα έχει επιλεγεί το διάγραμμα σκορ ποιότητας ανά βάση από το οποίο βλέπουμε ότι στις αναγνωσμένες αλληλουχίες αυξάνεται σημαντικά η πιθανότητα λανθασμένης καταγραφής βάσης μετά το 21ο νουκλεοτίδιο.

Οι εξειδικευμένοι αλγόριθμοι που περιλαμβάνει το FASTX αναλαμβάνουν ένα μεγάλο εύρος λειτουργιών όπως παρουσιάζονται παρακάτω:

- Μετατροπή αρχείων FASTQ σε FASTA
- Προβολή στατιστικών στοιχείων ποιότητας και κατανομής νουκλεοτιδίων
- Απαλοιφή πανομοιότυπων αναγνωσμένων αλληλουχιών
- Περικοπή αναγνωσμένων αλληλουχιών για την απομάκρυνση περιοχών χαμηλής ποιότητας
- Μετονομασία αναγνωριστικών κωδικών αλληλουχιών
- Απομάκρυνση αλληλουχιών που ανήκουν σε προσαρμογείς από το πείραμα αλληλούχισης
- Μετατροπή αλληλουχιών στις αντίστοιχες συμπληρωματικές
- Διαχωρισμός αρχείων FASTA/Q σε περισσότερα μικρότερου μεγέθους
- Αλλαγή μήκους αλληλουχιών ανά γραμμή σε ένα αρχείο FASTA
- Μετατροπή αλληλουχιών DNA σε/από RNA
- Φιλτράρισμα και αποκοπή αλληλουχιών με βάση το σκορ ποιότητας
- Απόκρυψη (masking) νουκλεοτιδίων, με βάση το σκορ ποιότητας, μέσω αντικατάστασή τους με το χαρακτήρα "N"

Σε αντίθεση με το FastQC το εργαλείο FASTX λειτουργεί μόνο σε περιβάλλον Linux χωρίς κάποιο γραφικό περιβάλλον κάνοντας το σχετικά δύσχρηστο για όσους δεν έχουν κάποια πληροφορική εμπειρία και εξοικείωση με λειτουργία προγραμμάτων μέσω γραμμής εντολών. Παρ' όλα αυτά παρέχει τη δυνατότητα ενσωμάτωσης με την διαδικτυακή πλατφόρμα Galaxy [44] και λειτουργίας μέσω του δικού της γραφικού περιβάλλοντος (Εικόνα 16).



Εικόνα 16. Σουίτα εργαλείων FASTX μέσω γραφικού περιβάλλοντος Galaxy. Στην εικόνα φαίνεται ενσωματωμένο το εργαλείο περικοπής αλληλουχιών της σουίτας FASTX στην πλατφόρμα Galaxy. Οι παράμετροι του εργαλείου, δηλαδή τα δεδομένα εισόδου που θα χρησιμοποιηθούν και οι βάσεις που θα αποκοπούν μπορούν να επιλεγούν από το χρήστη μέσω του διαδικτυακού γραφικού περιβάλλοντος της πλατφόρμας.

1.3.2 Σύσταση μικροβιακού πληθυσμού

Η πλήρης καταγραφή του μικροβιακού πληθυσμού σε ένα μεταγενωμικό δείγμα αποτελεί ένα εξαιρετικά δυσεπίλυτο πρόβλημα, καθώς η μεγάλη πλειοψηφία των οργανισμών που συναντώνται στο περιβάλλον δεν μπορούν να καλλιεργηθούν [45, 46] με τις ήδη υπάρχουσες εργαστηριακές τεχνικές. Το γεγονός αυτό, σε συνδυασμό με τη σημαντική μείωση τους κόστους των τεχνολογιών αλληλούχισης έχει οδηγήσει σε υιοθέτηση μεταγενωμικών προσεγγίσεων για την αποσαφήνηση του ταξονομικού προφίλ ενός μικροβιακού πληθυσμού. Η αξιοποίηση των δεδομένων μεταγενωμικής αλληλούχισης, με την εφαρμογή βιοπληροφορικών τεχνολογιών [47], μας επιτρέπει να παρακάμψουμε την ανάγκη ανάπτυξης καλλιεργειών στο

εργαστήριο για τον ποιοτικό και ποσοτικό προσδιορισμό των διαφορετικών ειδών που περιέχονται στο δείγμα μας. Η μεταγενωμική προσέγγιση μπορεί είτε να στηριχθεί είτε σε χαρακτηριστικές σηματοδοτικές αλληλουχίες για κάθε οργανισμό μέσω βαθειάς αλληλούχισης αμπλικονίων (deep amplicon sequencing - DAS), είτε να συμπεριλάβει ολόκληρο το μεταγονιδίωμα ενός δείγματος μέσω αλληλούχισης τμηματικής ανάλυσης.

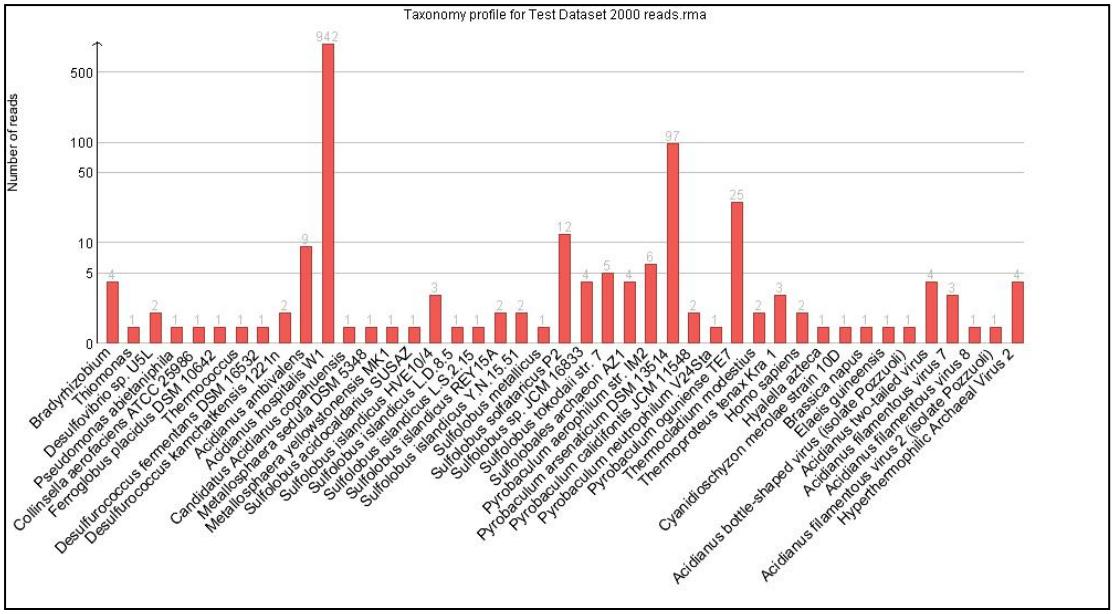
Στην πρώτη περίπτωση γίνεται στοχευόμενη ενίσχυση, μέσω PCR, αλληλουχιών που ανήκουν σε συντηρημένες γονιδιακές οικογένειες όπως ριβοσωμικά γονίδια [48] ή εσωτερικές μεταγραφόμενες περιοχές [49] (internal transcribed spacer regions - ITS). Οι ποσοτικά ενισχυμένες αλληλουχίες που ονομάζονται αμπλικόνια (amplicons) μπορούν να αντιστοιχιστούν σε συγκεκριμένα μικροβιακά είδη, των οποίων το γονιδίωμα περιέχει τις εν λόγω γονιδιακές οικογένειες. Αυτό καθίσταται δυνατόν μέσω βιοπληροφορικής ανάλυσης ομολογίας, κατά την οποία γίνεται σύγκριση των δεδομένων αλληλούχισης των αμπλικονίων με γνωστές βάσεις δεδομένων αντιστοίχων αλληλουχιών [50] ώστε να αποκαλύψουν ομοιότητες με γονιδιώματα ήδη γνωστών μικροβιακών στελεχών και να υποδείξουν την ύπαρξη μη ταυτοποιημένων ειδών. Η ενίσχυση μέσω PCR πριν την τελική αλληλούχιση προσφέρει μία υψηλότερη ευαισθησία στη μέθοδο τον εντοπισμού αλληλουχιών που αντιστοιχούν σε μικροβιακά είδη τα οποία μπορεί να μην βρίσκονται σε περίσσεια αλλά επιφέρει επίσης και πιθανά σφάλματα ποσοτικοποίησης για συγκεκριμένα είδη στον πληθυσμό [51]. Επίσης καθώς η μέθοδος αυτή στηρίζεται σε συγκεκριμένες συντηρημένες περιοχές γονιδίων, που δρουν ως εκκινητές για την ταυτοποίηση των αντίστοιχων μικροοργανισμών μέσω ενίσχυσής τους, μπορεί να παραλείψει εντελώς μικροβιακά είδη τα οποία δεν περιέχουν αυτές τις περιοχές στο γονιδίωμά τους [52].

Στην δεύτερη περίπτωση, οι μέθοδοι που στηρίζονται στην αλληλούχιση τμηματικής ανάλυσης προσφέρουν μία εκτενέστερη απεικόνιση τόσο ποιοτική όσο και ποσοτική των μικροβιακών ειδών που είναι παρόντα, καθώς γίνεται καταγραφή του συνόλου του μεταγονιδιώματος του δείγματος. Η βιοπληροφορική ανάλυση κατά τις μεθόδους αυτές στηρίζεται και πάλι στη σύγκριση αλληλουχιών με βάσεις δεδομένων οι οποίες όμως είναι πολύ μεγαλύτερου μεγέθους από τις αντίστοιχες για τις DAS τεχνικές. Αυτό συμβαίνει γιατί τα δεδομένα αλληλούχισης τμηματικής ανάλυσης δεν προέρχονται από συγκεκριμένες γονιδιακές οικογένειες, αλλά αντιστοιχούν στο σύνολο του μεταγονιδιώματος. Έτσι για να γίνει σωστός προσδιορισμός της ταξονομικής τους ταυτότητας θα πρέπει να συγκριθούν με βάσεις

δεδομένων που περιλαμβάνουν κάθε γνωστή αλληλουχία που έχει καταγραφεί [53] και είναι διαθέσιμη, ανεξάρτητα από το κυτταρικό είδος και γονιδιακή οικογένεια (αν και εφόσον) ανήκει. Η χρήση αυτών των μεθόδων αποφεύγει τα σφάλματα ποσοτικοποίησης της PCR και αποδεικνύεται τελικά πιο αποτελεσματική στην αποσαφήνιση της μικροβιακής ποικιλομορφίας του δείγματος σε σχέση με της DAS τεχνικές [54]. Εκτός από την αξιοποίηση των δεδομένων τμηματικής αλληλούχισης με βάση την ομοιότητα των αναγνωσμένων αλληλουχιών με ήδη γνωστές, υπάρχει και μια επιπλέον προσέγγιση η οποία στηρίζεται στο ότι διαφορετικά γονιδιώματα περιέχουν συντηρημένα στοιχεία νουκλεοτιδικής σύστασης. Αυτά μπορεί να περιλαμβάνουν από συγκεκριμένο ποσοστό γουανίνης/κυτοσίνης έως την υπέρμετρη επανάληψη ορισμένων ολιγονουκλεοτιδικών περιοχών στην αλληλουχία τους. Οι δύο παραπάνω προσεγγίσεις μπορούν να λειτουργήσουν και συνδυαστικά από αλγορίθμους που λαμβάνουν υπόψη τόσο την νουκλεοτιδική σύσταση των αναγνωσμένων αλληλουχιών, όσο και την ομοιότητα τους με ήδη γνωστές [2].

Ανεξάρτητα από τη σύσταση του γενετικού υλικού, την πειραματική διαδικασία απομόνωσής του και τη βάση δεδομένων με την οποία θα συγκριθεί (αν είναι απαραίτητο), ο τελικός σκοπός της βιοπληροφορικής ανάλυσης των μεταγονιδιωματικών δεδομένων παραμένει ο ίδιος και είναι η αντιστοίχιση κάθε αναγνωσμένης αλληλουχίας σε κάποια ταξονομική κατηγορία (bin). Η ποιοτική σύσταση (ποιοι μικροοργανισμοί είναι παρόντες) του δείγματος θα εξαρτηθεί από το σύνολο των μικροβιακών ειδών στα οποία κατηγοριοποιήθηκαν οι αναγνωσμένες αλληλουχίες, ενώ η ποσοτικοποίηση αυτής της κατάταξης εξαρτάται από το πλήθος των αλληλουχιών σε κάθε μία κατηγορία (Εικόνα 17).

Στην περίπτωση που η κατηγοριοποίηση βασίζεται στην ομοιότητα με ήδη γνωστές αλληλουχίες, το στάδιο της ανάλυσης ομολογίας αποτελεί το πιο αργό και καθοριστικό στάδιο όσον αφορά τη διάρκεια της συγκεκριμένης ανάλυσης. Η πιο διαδεδομένη και ίσως πιο αξιόπιστη μεθοδολογία για το στάδιο αυτό αποτελείται από τη χρήση της πλατφόρμα εργαλείων BLAST (Basic Local Alignment Search Tool) [55]. Το BLAST μπορεί να αξιοποιήσει βάσεις δεδομένων που περιέχουν αλληλουχίες αναφοράς (reference sequences) είτε νουκλεοτιδίων ή αμινοξέων για τη σύγκριση ομοιότητας (alignment) των αλληλουχιών «ερωτημάτων» (query sequences), που στην προκειμένη περίπτωση αποτελούνται από τις αναγνωσμένες αλληλουχίες των δεδομένων αλληλούχισης.



Εικόνα 17. Ραβδόγραμμα ταξονομικής ανάλυσης (λογαριθμική κλίμακα). Το παραπάνω δείγμα προέρχεται από την ταξονομική ανάλυση 2000 τυχαίων αναγνωσμένων αλληλουχιών από μεταγενωμικό δείγμα που λήφθηκε από μία θερμοπηγή στο Pisciarelli, Ιταλίας στα πλαίσια του ερευνητικού προγράμματος HotZyme[56]. Οι αλληλουχίες εξετάστηκαν με ανάλυση ομολογίας χρησιμοποιώντας το πρόγραμμα BLAST[55] και η ταξονομική τους κατηγοριοποίηση έγινε με το πρόγραμμα MEGAN[57]. Από τα αποτελέσματα της κατηγοριοποίησης φαίνεται το πλήθος των διαφορετικών ειδών που υπάρχει στο δείγμα καθώς και ότι το κυρίαρχο είδος στον πληθυσμό είναι το Acidianus hospitalis με 942 από τα 2000 αναγνωσμένες αλληλουχίες να έχουν αντιστοιχιστεί σε αυτό. Στο παραπάνω διάγραμμα δεν εμφανίζονται οι αλληλουχίες (σύνολο 840) για τις οποίες δεν ήταν δυνατός ο ταξονομικός προσδιορισμός τους σε κάποιο μικροβιακό είδος.

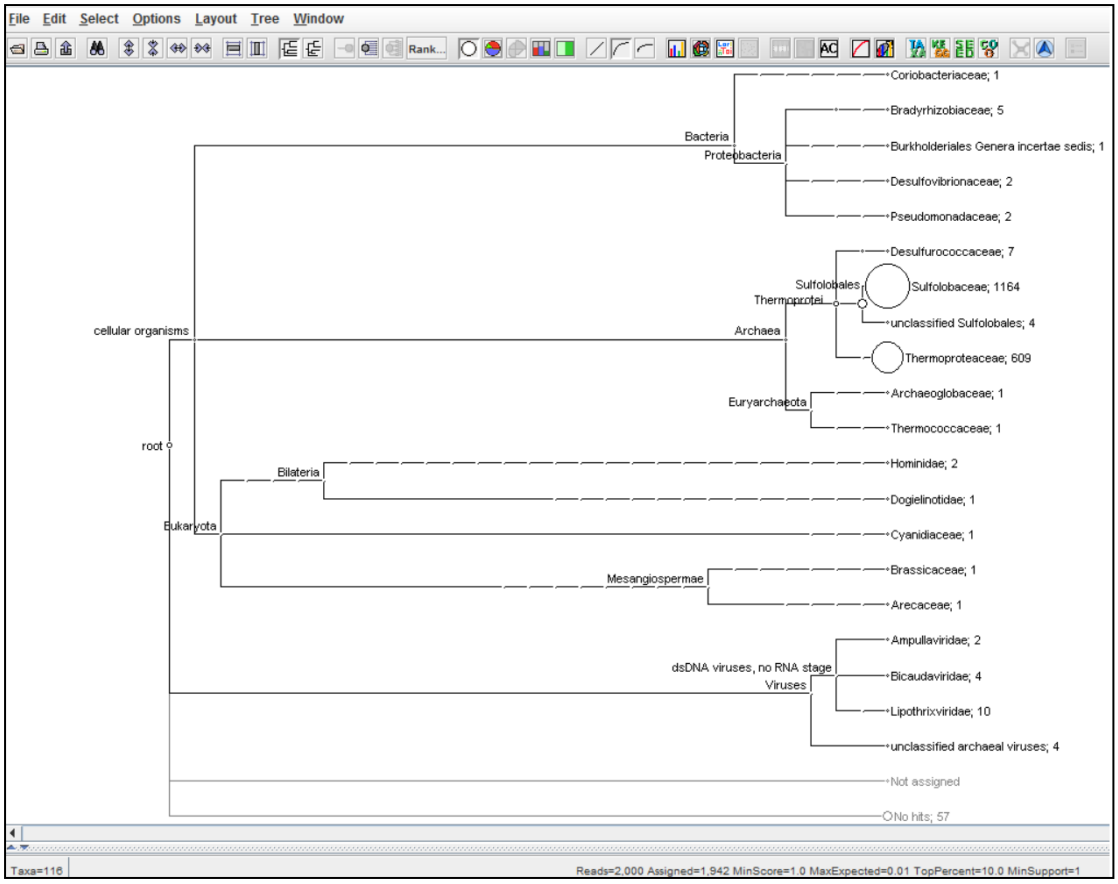
Οι δυνατότητες του BLAST όμως, δεν περιορίζονται μόνο στη σύγκριση νουκλεοτιδικών αλληλουχιών. Μία ανάλυση μπορεί να περιλαμβάνει τη σύγκριση νουκλεοτιδικών αλληλουχιών σε νουκλεοτιδικές βάσεις δεδομένων (εργαλείο BLASTn), τη σύγκριση πεπτιδικών αλληλουχιών σε πεπτιδικές βάσεις δεδομένων (εργαλείο BLASTp), τη μετάφραση νουκλεοτιδικών αλληλουχιών στα έξι πλαίσια ανάγνωσης και σύγκριση με πεπτιδικές βάσεις δεδομένων (εργαλείο BLASTx) ή τη σύγκριση πεπτιδικών αλληλουχιών σε νουκλεοτιδικές βάσεις δεδομένων μεταφρασμένες αντίστοιχα στα έξι πλαίσια ανάγνωσης (εργαλείο tBLASTn). Κάθε εργαλείο της πλατφόρμας BLAST αξιοποιεί ευρετικούς (heuristic) αλγορίθμους οι οποίοι ξεκινάνε την ανάλυση σύγκρισης πάνω σε μικρότερα κομμάτια (words) κάθε αλληλουχίας-ερωτήματος. Για τις εντοπισμένες ομοιότητες κομματιών (seeds) γίνεται επέκταση της ανάλυσης σύγκρισης, με βάση τις αρχικές παραμέτρους του προγράμματος, ώστε να εκτιμηθεί το ποσοστό ομοιότητας στο σύνολο της κάθε αρχικής αλληλουχίας-ερωτήματος με την αντίστοιχη αλληλουχία αναφοράς. Κατά τη διάρκεια αυτής της επέκτασης χρησιμοποιούνται στατιστικά σκορ μέσω εμπειρικών πινάκων υποκαταστάσεων (substitution matrices), όπως ο BLOSUM62 για την εκτίμηση κάθε ομοιότητας. Αν κατά τη διάρκεια της επέκτασης σύγκρισης το

στατιστικό σκορ πέσει κάτω από ένα προκαθορισμένο όριο, η ανάλυση σύγκρισης σταματάει. Από τα στατιστικά τεστ που εφαρμόζονται κατά τη διάρκεια της ανάλυσης προκύπτουν στο τέλος σκορ στατιστικής σημαντικότητας, όπως η «προσδωκόμενη» τιμή (expected value - E-value) και η τιμή πιθανότητας (p-value), που αντιστοιχούν σε κάθε αλληλουχία-ερώτημα για την οποία βρέθηκε ομοιότητα με κάποια αλληλουχία αναφοράς. Κατά αυτόν τον τρόπο μπορεί να εκτιμηθεί η σημαντικότητα των τελικών δεδομένων και να φιλτραριστούν ανάλογα με την ευαισθησία που απαιτείται για την ανάλυσή μας.

Το BLAST εκτός από το εύρος των αναλύσεων που διαθέτει, προσφέρει και μία εξαιρετική ευελιξία στον τρόπο χρήσης του. Διατίθεται ως διαδικτυακή εφαρμογή με ένα φιλικό ως προς το χρήστη γραφικό περιβάλλον καθώς και τη δυνατότητα τοπικής εγκατάστασης σε υπολογιστή ή διακομιστή (server), μαζί με τις αντίστοιχες βάσεις δεδομένων, για τη βέλτιστη διαχείριση των υπολογιστικών πόρων που διατίθενται για τη λειτουργία του. Αυτό είναι εξαιρετικά σημαντικό όταν χρησιμοποιείται για την ανάλυση μεταγενωμικών δεδομένων, καθώς ο όγκος τους απαιτεί τη χρήση πολλαπλών επεξεργαστών (CPUs) όπως και την κατανάλωση μεγάλης ποσότητας εικονικής μνήμης (RAM). Επίσης, λόγω της διαρκούς χρήσης του από την επιστημονική κοινότητα βελτιώνεται συνεχώς, με νέες εκδόσεις [58] να έχουν εμφανιστεί που βελτιστοποιούν την απόδοση και ταχύτητα του προγράμματος.

Για το στάδιο της ταξονομικής κατηγοριοποίησης έχουν αναπτυχθεί πολλά εργαλεία καθένα από τα οποία επιστρατεύει διαφορετικούς αλγορίθμους κατηγοριοποίησης των αναγνωσμένων αλληλουχιών, όπως αυτοοργανωμένοι χάρτες (self organising maps - SOMs) ή ιεραρχική ομαδοποίηση (hierarchical clustering). Τα περισσότερα από αυτά δουλεύουν μόνο σε περιβάλλον Linux μέσα από γραμμή εντολών ενώ λίγα είναι αυτά που έχουν ενσωματώσει κάποιο γραφικό περιβάλλον για χρήστες χωρίς ιδιαίτερες γνώσεις βιοπληροφορικής.

Ένα τέτοιο παράδειγμα αποτελεί το εργαλείο MEGAN [57], το οποίο βασίζεται σε δεδομένα ανάλυσης ομολογίας για την κατηγοριοποίηση των αλληλουχιών και διαθέτει ένα διαδραστικό γραφικό περιβάλλον όπου ο χρήστης μπορεί να επιλέξει συγκεκριμένες λειτουργίες ή ταξονομικές κατηγορίες για την ανάλυσή του (Εικόνα18).



Εικόνα 18. Γραφικό περιβάλλον του εργαλείου MEGAN. Το παραπάνω δείγμα προέρχεται από την ταξονομική ανάλυση 2000 τυχαίων αναγνωσμένων αλληλουχιών από μεταγενωμικό δείγμα που λήφθηκε από μία θερμοπηγή στο Pisciarelli, Ιταλίας στα πλαίσια του ερευνητικού προγράμματος HotZyme[56]. Τα αποτελέσματα της ταξονομικής ανάλυσης του δείγματος φαίνονται μέσα από το γραφικό περιβάλλον του MEGAN όπου έχει κατασκευαστεί το αντίστοιχο ταξονομικό δέντρο μέχρι το φυλογενετικό επίπεδο «οικογένεια» (family). Ο κάθε κύκλος αντιστοιχεί σε μία ταξονομική κατηγορία και το μέγεθος του είναι ανάλογο με το πλήθος των αλληλουχιών που έχουν αντιστοιχιστεί σε αυτήν. Στο συγκεκριμένο παράδειγμα οι επικρατούσες οικογένειες είναι η Sulfolobaceae και η Thermoproteaceae με 1164 και 609 κατηγοριοποιημένες αλληλουχίες σε αυτές αντίστοιχα.

Το MEGAN επιστρατεύει έναν αλγόριθμο εύρεσης κοντινότερου κοινού προγόνου (least common ancestor - LCA) για την κατηγοριοποίηση των αναγνωσμένων αλληλουχιών χρησιμοποιώντας τον ταξονομικό προσδιορισμό του NCBI [53]. Ανάλογα με το πόσο συντηρημένη είναι η κάθε αλληλουχία που εξετάζεται, κατηγοριοποιείται στο κατάλληλο ταξονομικό επίπεδο. Ευρέως συντηρημένες αλληλουχίες αντιστοιχίζονται σε πιο υψηλά επίπεδα (πιο κοντά στο επίπεδο «βασίλειο» - kingdom), ενώ αλληλουχίες συνδεδεμένες με συγκεκριμένα είδη αντιστοιχίζονται σε κατώτερα επίπεδα που βρίσκονται στα «φύλλα» (leaves) του ταξονομικού δέντρου. Το MEGAN επιπλέον προσφέρει και λειτουργικό χαρακτηρισμό των μεταγενωμικών δεδομένων κατηγοριοποιώντας τα επίσης σε πρωτεϊνικές λειτουργίες και τα αντίστοιχα μεταβολικά μονοπάτια τα οποία κατασκευάζονται από αυτές. Ο αλγόριθμος που χρησιμοποιεί υποστηρίζει μεταξύ

άλλων και κατηγοριοποίηση με βάση τις KEGG [59], COG [60] και SEED [61] οντολογίες, προσφέροντας έναν ολοκληρωμένο χαρακτηρισμό για οποιοδήποτε περιβαλλοντικό δείγμα.

1.3.3 Συναρμολόγηση μεταγονιδιώματος

Ανάλογα την τεχνολογία αλληλούχισης που χρησιμοποιείται τα δεδομένα που προκύπτουν μπορεί να είναι μονού άκρου (single-end) ή ζεύγους άκρων (paired-end). Η διαφορά τους είναι ότι στην πρώτη περίπτωση το κάθε θραύσμα αλληλουχείται μία φορά, ξεκινώντας από ένα από τα δύο άκρα του, ενώ στην περίπτωση ζεύγους άκρων κάθε κομμάτι αλληλουχείται και στα δύο άκρα του, ενώ η μέθοδος εγγυάται ότι υπάρχει πάντα η ίδια απόσταση (σε αριθμό νουκλεοτιδίων) μεταξύ κάθε δύο άκρων που διαβάζονται. Σε κάθε περίπτωση, τα δεδομένα που προκύπτουν αποτελούνται από μία τεράστια λίστα αναγνωσμένων αλληλουχιών τα οποία είναι πολύ μικρότερα του αρχικού μορίου από το οποίο προήλθαν. Η διαδικασία ανασχηματισμού της αρχικής αλληλουχίας ονομάζεται συναρμολόγηση (assembly) και μπορεί να παρομοιαστεί με τη διαδικασία συναρμολόγησης ενός παζλ από τα μικρότερα κομμάτια του.

Η μέθοδος της συναρμολόγησης βασίζεται στο γεγονός ότι πανομοιότυπα μόρια DNA σπάνε σε διαφορετικά, μη προκαθορισμένα, σημεία κατά την αλληλούχιση τμηματικής ανάλυσης. Έτσι προκύπτουν δεδομένα αναγνωσμένων αλληλουχιών, που έχουν κάποιο βαθμό επικάλυψης ο οποίος μπορεί να αξιοποιηθεί για τη σωστή διάταξή τους, όπως οι προεξοχές και οπές των κομματιών ενός παζλ χρησιμοποιούνται για την αντίστοιχη διάταξη μεταξύ τους. Η σωστή διάταξη των αναγνωσμένων αλληλουχιών οδηγεί στο σχηματισμό μεγαλύτερων συνεχόμενων (contiguous) αλληλουχιών, που ονομάζονται συναρμολογήματα (contigs). Τα συναρμολογήματα αυτά μπορούν να επεκταθούν περαιτέρω με προσθήκη περισσοτέρων αναγνωσμένων αλληλουχιών, με τελικό σκοπό να αποκαλύψουν την αρχική αλληλουχία από την οποία προήλθαν. Αυτή η διαδικασία ενώ ακούγεται θεωρητικά εύκολη για μικρό αριθμό αναγνωσμένων αλληλουχιών και μικρό μέγεθος αρχικού μορίου DNA, καταλήγει να είναι μία χρονοβόρα και υπολογιστικά επίπονη εργασία, καθώς εκτός από το τεράστιο πλήθος θραυσμάτων και το μεγάλο πλήθος από επαναληπτικές αλληλουχίες που δυσχεραίνουν τη σωστή διάταξη, υπάρχουν και πειραματικά σφάλματα ή μεταλλάξεις γενετικού υλικού όπως πολυμορφισμοί μονού

νουκλεοτιδίου (single nucleotide polymorphism - SNP) που περιπλέκουν ακόμα περισσότερο τη σωστή ανάλυση των δεδομένων.

Για την επίλυση αυτών των προβλημάτων, τα οποία είναι εμφανέστερα στις περιπτώσεις μεταγενωμικών δειγμάτων λόγω των πολλών διαφορετικών γονιδιωμάτων από τα οποία προέρχονται οι αναγνωσμένες αλληλουχίες, έχουν αναπτυχθεί αλγόριθμοι και βιοπληροφορικά εργαλεία που ειδικεύονται στην εκ νεου (de novo) συναρμολόγηση αλληλουχίας, αξιοποιώντας πλήρως τις δυνατότητες των σύγχρονων υπολογιστικών υποδομών [62]. Ο χαρακτηρισμός «εκ νέου» υπάρχει για να τη διαχωρίζει με τη συναρμολόγηση χαρτογράφησης (mapping assembly), που εφαρμόζεται σε αλληλούχιση γονιδιώματος μεμονωμένων οργανισμών. Στην περίπτωση συναρμολόγησης με χαρτογράφηση, το γονιδίωμα του οργανισμού υπάρχει δημοσιευμένο και χρησιμοποιείται ως γονιδίωμα αναφοράς (reference genome) ή εάν δεν είναι διαθέσιμο χρησιμοποιείται γονιδίωμα συγγενούς είδους για τον ίδιο σκοπό. Για την περίπτωση της εκ νέου συναρμολόγησης τα πρώτα εργαλεία που αναπτύχθηκαν [63-65] υιοθετούσαν τη μεθοδολογία της επέκτασης των συναρμολογημάτων μέσω «άπληστων» αλγορίθμων (greedy algorithms) [66] για την τελική ανακατασκευή του αρχικού γονιδιώματος. Η αυξημένη πολυπλοκότητα όμως του προβλήματος διάταξης για τους τεράστιους όγκους δεδομένων και τα μικρότερου μεγέθους θραύσματα που προέκυπταν από τις τεχνολογίες αλληλούχισης νέας γενιάς, οδηγούσε σε σφάλματα της ενσωμάτωσης των επικαλύψεων στα συναρμολογήματα.

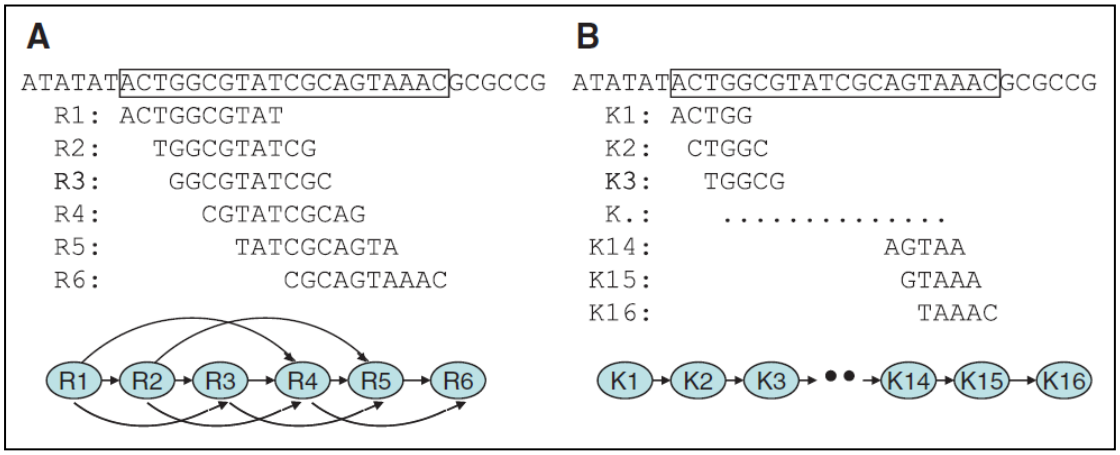
Έτσι αναπτύχθηκαν νέου είδους αλγόριθμοι τους οποίους χρησιμοποιούν μέχρι και σήμερα [67] τα σύγχρονα βιοπληροφορικά εργαλεία για την επίλυση του προβλήματος της συναρμολόγησης. Οι νέοι αυτοί αλγόριθμοι βασίζονται στην απεικόνιση των σχέσεων επικάλυψης μεταξύ των αναγνωσμένων αλληλουχιών ως ένα γράφημα στο οποίο κάθε αλληλουχία αντιπροσωπεύεται από έναν κόμβο και κάθε σχέση από μία ακμή. Κατά αυτόν τον τρόπο το πρόβλημα συναρμολόγησης μετασχηματίζεται σε πρόβλημα εύρεσης ενός μονοπατιού που θα περνάει από όλους τους κόμβους του γραφήματος ακριβώς μία φορά. Τα εργαλεία που βασίζονται σε αυτή τη μεθοδολογία επιστρατεύουν δύο κύρια είδη αλγορίθμων [68]:

- αλγόριθμοι overlap - layout - consensus (OLC)
- αλγόριθμοι γραφημάτων De Bruijn

Με τους OLC αλγόριθμους η κατασκευή του γραφήματος γίνεται σε τρία βήματα. Ξεκινάει με τον εντοπισμό όλων των δυνατών επικαλύψεων (overlap) μεταξύ κάθε πιθανού ζεύγους αναγνωσμένων αλληλουχιών για τον σχεδιασμό ενός πρωταρχικού γραφήματος με όλους τους κόμβους και τις πιθανές ακμές μεταξύ τους. Στη συνέχεια το γράφημα απλοποιείται αφαιρώντας πλεονάζουσες πληροφορίες ακμών με αποτέλεσμα την διάταξη (layout) των κόμβων μεταξύ τους (που αντιστοιχεί σε διάταξη των αλληλουχιών κατά μήκος του γονιδιώματος). Στο τελευταίο στάδιο διορθώνονται τα μονοπάτια, λαμβάνοντας υπόψη τη μεγαλύτερη πιθανότητα ύπαρξης (consensus) κάθε νουκλεοτιδίου στην αντίστοιχη θέση του συναρμολογήματος. Τα σχηματιζόμενα συναρμολογήματα, τις περισσότερες φορές, δεν καλύπτουν το 100% του κάθε γονιδιώματος λόγω της αδυναμίας σωστής διάταξης αλληλουχιών σε όλο το εύρος του, ειδικά σε περιοχές με επαναληπτικές αλληλουχίες, αλλά το τελικό μέγεθός που προκύπτει είναι αρκετά μεγάλο ώστε να μπορούν να αξιοποιηθούν για τον εντοπισμό γονιδίων. Το τελικό μέγεθος των συναρμολογημάτων εξαρτάται από τις ιδιότητες των δεδομένων αλληλούχισης, όπως το μήκος σε ζεύγη βάσεων των αναγνωσμένων αλληλουχιών και τις παραμέτρους του αλγορίθμου συναρμολόγησης, όπως το όριο ελάχιστης επικάλυψης των αναγνωσμένων αλληλουχιών [68]. Στην περίπτωση που τα δεδομένα είναι ζεύγους άκρων, τα δεδομένα της απόστασης μεταξύ των διαφορετικών ζευγών αλληλουχιών μπορούν να αξιοποιηθούν από τους αλγορίθμους συναρμολόγησης, ώστε να γίνει η διάταξη με ακόμα πιο βελτιστοποιημένο τρόπο σχηματίζοντας έτσι υπερ-συναρμολογήματα (supercontigs). Τα υπερ-συναρμολογήματα μπορεί να περιέχουν κενά αλλά προσομοιάζουν σε ακόμα μεγαλύτερο βαθμό την αλληλουχία του αρχικού (μετα)γονιδιώματος.

Οι αλγόριθμοι γραφημάτων De Bruijn ακολουθούν μία διαφορετική μέθοδο κατασκευής του γραφήματος συναρμολόγησης. Κατακερματίζουν κάθε αναγνωσμένη αλληλουχία σε όλα τα πιθανά k-μερή της όπου k: ένας ακέραιος αριθμός που ορίζεται από τον χρήστη π.χ. αν μία αναγνωσμένη αλληλουχία είναι η ACTGGCG και το k=5 τότε όλα τα πιθανά 5μερή που χωρίζεται είναι: ACTGG, CTGGC και TGGCG. Στη συνέχεια κατασκευάζουν το γράφημα με βάση όλες τις πιθανές επικαλύψεις μεταξύ k-μερών που διαφέρουν μεταξύ τους κατά ένα νουκλεοτίδιο π.χ. στο παραπάνω παράδειγμα τα ζεύγη k-μερών που έχουν τέτοια επικάλυψη είναι τα: ACTGG-CTGGC και CTGGC- TGGCG (Εικόνα 19). Όπως και στον προηγούμενο τύπο αλγορίθμου γίνεται διόρθωση των μονοπατιών του γραφήματος, με αποτέλεσμα την

κατασκευή συναρμολογημάτων και υπερ-συναρμολογημάτων στην περίπτωση δεδομένων ζεύγους άκρων.



Εικόνα 19. Κατασκευή γραφημάτων μέσω OLC και DBG αλγορίθμων. (A)Χρήση αλγόριθμου OLC με 6 θραύσματα (R1-R6) για τη συγκεκριμένη αλληλουχία των 20 ζευγών βάσεων(τετράγωνο πλαίσιο). Το μήκος αναγνωσμένων αλληλουχιών (sequencing reads) είναι 10 ζεύγη βάσεων και το όριο ελάχιστης επικάλυψης είναι 5. Οι αναγνωσμένες αλληλουχίες προσανατολίστηκαν κατάλληλα ώστε να φαίνονται οι επικαλύψεις τους και κατασκευάστηκε το αντίστοιχο γράφημα. Στο γράφημα που σχηματίζεται οι κόμβοι αντιστοιχούν σε αναγνωσμένες αλληλουχίες και οι ακμές σε επικαλύψεις μεταξύ τους. Λόγω των πολλαπλών επικαλύψεων ≥5 βάσεων μεταξύ των αλληλουχιών φαίνεται ότι οι περισσότεροι κόμβοι έχουν παραπάνω από μία ακμές που εισέρχονται και εξέρχονται κάτι που υποδεικνύει την ανάγκη απλοποίησης. (B)Με τον αλγόριθμο DBG οι αλληλουχίες αλληλουχιών κατακερματίστηκαν περαιτέρω σε 16 k-μερή μήκους 5 ζευγών βάσεων(K1-K16). Ο προσανατολισμός των 5μερών κατά μήκος των επικαλύψεών τους έδωσε το αντίστοιχο γράφημα όπου ο κάθε κόμβος έχει μόνο μία ακμή που εισέρχεται και μόνο μία που εξέρχεται και μπορεί πλέον να αποκαλύψει την αρχική αλληλουχία χωρίς περαιτέρω επεξεργασία [68].

Τα αποτελέσματα συναρμολόγησης από τις δύο διαφορετικές προσεγγίσεις μπορεί να διαφέρουν σημαντικά ανάλογα την τεχνολογία αλληλούχισης που χρησιμοποιείται και του είδους των γονιδιωμάτων που εξετάζονται. Παρ' όλα αυτά, έχει βρεθεί εμπειρικά ότι η συναρμολόγηση μέσω OLC αλγορίθμων αποδίδει καλύτερα σε μεγαλύτερου μήκους αναγνωσμένες αλληλουχίες ενώ η αντίστοιχη DBG προσέγγιση θα πρέπει να εφαρμόζεται σε δεδομένα μεγάλου πλήθους αλληλουχιών μικρότερου μήκους [68].

Τα πιο ευρέως χρησιμοποιημένα εργαλεία ανοιχτού κώδικα που ανήκουν σε αυτές τις δύο κατηγορίες είναι το Velvet [69] για DBG συναρμολόγηση και το Celera [70] για OLC. Για τον αλγόριθμο του Velvet μάλιστα έχει σχεδιαστεί μία επέκταση, που είναι εξειδικευμένη για την αντιμετώπιση της πολυπλοκότητας των δεδομένων από μεταγενωμικά δείγματα και ονομάζεται MetaVelvet [71]. Το MetaVelvet απλοποιεί το πρόβλημα κατασκευής του γραφήματος με το να αναδιατάσσει τα k-μερή προς κατασκευή υπο-γραφημάτων καθένα από τα οποία αντιστοιχεί σε κάθε ξεχωριστό γονιδίωμα μέσα στο μεταγενωμικό δείγμα. Τα προαναφερθέντα εργαλεία είναι διαθέσιμα μόνο για λειτουργία σε λογισμικό Linux και η χρήση τους απαιτεί

(όπως και τον υπόλοιπων εργαλείων συναρμολόγησης) εξαιρετικά σημαντική κατανομή υπολογιστικών πόρων σε αυτά όμως παραμένουν πολύ αξιόπιστες λύσεις για την εκ νέου συναρμολόγηση μεταγονιδιωμάτων.

1.3.4 Εντοπισμός γονιδίων

Το στάδιο της αποτύπωσης του γονιδιακού περιεχομένου ενός δείγματος μας δίνει την ευκαιρία να εντοπίσουμε πλήθος αλληλουχιών οι οποίες αποτελούν πιθανά γονίδια. Το πρώτο βήμα στον εντοπισμό γονιδίων είναι η εύρεση ανοιχτών πλαισίων ανάγνωσης (open reading frames - ORFs) μέσα στα δεδομένα συναρμολόγησης. Υπάρχουν ήδη δημοσιευμένα εργαλεία [72, 73] που συμπεριλαμβάνουν την εύρεση ORFs στις δυνατότητες που προσφέρουν, αν και λόγω της φύσης του προβλήματος είναι σχετικά εύκολο να σχεδιαστεί και να γραφτεί, σε οποιαδήποτε γλώσσα προγραμματισμού, ένας αλγόριθμος εύρεσης κωδικονίων έναρξης και λήξης λαμβάνοντας υπόψη και τον αριθμό νουκλεοτιδίων μεταξύ τους για την αντιστοίχιση των κωδικονίων στα αντίστοιχα αμινοξέα.

Αυτή η θεωρητικά εύκολη υπολογιστικά διαδικασία δυσκολεύει αισθητά, όταν η ανάλυση μας αφορά δεδομένα προερχόμενα από ένα μεταγενωμικό δείγμα. Τα εν λόγω δεδομένα αποτελούν μία πρόκληση, καθώς παρά το τεράστιο μέγεθός τους (που από μόνο του καθιστά την επεξεργασία τους εξαιρετικά χρονοβόρα), δεν καλύπτουν πάντα το σύνολο της αλληλουχίας όλων των γονιδιωμάτων του δείγματος. Αυτό έχει ως συνέπεια, κατά τη διαδικασία της συναρμολόγησης, να σχηματίζονται συναρμολογήματα μικρότερου μεγέθους αλληλουχίας από εκείνη του γονιδιώματος από το οποίο προέρχονται. Ως αποτέλεσμα, πολλά ανοικτά πλαίσια ανάγνωσης που αντιστοιχούν σε γονίδια να μην αντιπροσωπεύονται επαρκώς από τα δεδομένα. Επίσης, πρέπει να ληφθεί υπόψη και το γεγονός ότι δεν αντιστοιχούν όλες οι αλληλουχίες ανοιχτών πλαισίων ανάγνωσης σε περιοχές που κωδικοποιούν πρωτεΐνες. Την επίλυση αυτών των προβλημάτων που εμφανίζονται σε δεδομένα μεταγενωμικών αλληλουχιών αναλαμβάνουν εργαλεία όπως το Prodigal [74], το MetaGene [75] ή το GeneMark [76]. Τα εργαλεία αυτά βασίζονται σε διαφορετικές μεθόδους εντοπισμού ανοικτών πλαισίων ανάγνωσης, όπως τεχνικές βασισμένες σε πρόβλεψη συχνότητας ολιγονουκλεοτιδίων σε κωδικές περιοχές του γονιδιώματος. Η χρήση τους ενώ είναι δυνατή μόνο σε περιβάλλον Linux δεν απαιτεί ιδιαίτερες υπολογιστικές απαιτήσεις από άποψη εικονικής μνήμης και αριθμού επεξεργαστών και μπορούν να λειτουργήσουν και σε προσωπικούς υπολογιστές.

1.3.5 Χαρακτηρισμός πρωτεϊνών

Η καθολικότητα του γενετικού κώδικα μας επιτρέπει τον μετασχηματισμό κάθε νουκλεοτιδικής αλληλουχίας σε αλληλουχία αμινοξέων και τον καθορισμό της πρωτοταγούς πρωτεϊνικής δομής στην οποία αντιστοιχεί. Παρ' όλα αυτά, η πρόβλεψη της λειτουργίας της απαιτεί ένα πλήθος πειραμάτων τόσο in silico, όσο και in vitro για την πλήρη επισήμανση(annotation) και επεξηγηματική περιγραφή (curation) μιας πρωτεϊνικής αλληλουχίας. Η πρώτη βιοπληροφορική προσέγγιση για τον χαρακτηρισμό μίας πρωτεΐνης είναι μέσω ανάλυσης ομολογίας, κατά την οποία γίνεται σύγκριση αλληλουχίας με βάσεις δεδομένων [53] γνωστών πρωτεϊνών. Η πρόβλεψη της λειτουργίας της κατά αυτόν τον τρόπο, στηρίζεται στην λειτουργία της αντίστοιχης γνωστής αλληλουχίας με τη μεγαλύτερη ομοιότητα σε αυτή.

Η καλύτερη λύση για τέτοιου είδους αναλύσεις ομολογίας αποτελείται και πάλι από την πλατφόρμα εργαλείων BLAST [55], αλλά τα τελευταία χρόνια εμφανίζονται αλγόριθμοι που δίνουν έμφαση περισσότερο στην ταχύτητα των αναλύσεων θυσιάζοντας μικρό μέρος της ακρίβειας όπως το BLAT [77] και το DIAMOND [78] και μπορούν να μειώσουν σημαντικά τους χρόνους ανάλυσης που προκύπτουν από τον τεράστιο όγκο δεδομένων από μεταγενωμικά δείγματα. Τα προγράμματα αυτά ακολουθούν τους ίδιους ευρετικούς αλγορίθμους με το BLAST με μικρές παραλλαγές στον τρόπο εύρεσης ομοιοτήτων μεταξύ των μικρότερων περιοχών (seeds) των αλληλουχιών-ερωτημάτων. Ανάλογα με τα αποτελέσματα της σύγκρισης ομολογίας μπορεί να εκτιμηθεί θεωρητικά η λειτουργία της πρωτεΐνης, με βάση το ποσοστό ομοιότητας με άλλες ήδη χαρακτηρισμένες αλληλουχίες.

Οι μέθοδοι που βασίζονται στην σύγκριση ομοιότητας αλληλουχίας, ενώ αποτελούν χρήσιμα εργαλεία, δεν είναι ικανές να δώσουν έναν ολοκληρωμένο λειτουργικό χαρακτηρισμό, λόγω του υψηλού ποσοστού παράλογων, όπως αποκαλούνται, γονιδίων (paralog genes) που υπάρχουν στη φύση [79] και καθώς περιορίζονται συνήθως από μία παράμετρο: τη βάση δεδομένων που χρησιμοποιούν. Οι σύγχρονες βάσεις δεδομένων ενώ περιλαμβάνουν εκατοντάδες εκατομμύρια αλληλουχίες περιλαμβάνουν μόνο ένα μικρό ποσοστό πλήρως χαρακτηρισμένων αλληλουχιών που μπορούν να χρησιμοποιηθούν αξιόπιστα σε αναλύσεις ομολογίας. Συγκριτικά αναφέρεται ότι από τις $\sim 3 \cdot 10^8$ πρωτεΐνες που υπάρχουν στην NCBI-nr [53] μόνο οι $\sim 6 \cdot 10^5$ είναι πλήρως χαρακτηρισμένες που ανήκουν στην UniProt/SwissProt [80]. Επιπλέον, ακόμα και αν λάβουμε υπόψη όλες τις δημοσιευμένες αλληλουχίες σε όλες τις διαθέσιμες βάσεις δεδομένων, αυτές

αντιπροσωπεύουν μόνο ένα πολύ μικρό ποσοστό των αλληλουχιών που μπορούν να συναντηθούν σε ένα περιβαλλοντικό δείγμα λόγω τις αδυναμίας καλλιέργειας των περισσοτέρων οργανισμών στο εργαστήριο με τα ήδη συμβατικά μέσα [45].

Μία διαφορετική προσέγγιση χαρακτηρισμού πρωτεϊνών είναι η χρήση αλγορίθμων που εξετάζουν την αλληλουχία για χαρακτηριστικά (features), τα οποία έχουν συσχετιστεί με συγκεκριμένη πρωτεϊνική λειτουργικότητα μέσω μεθοδολογιών μηχανικής μάθησης (machine learning). Τέτοιοι αλγόριθμοι υιοθετούνται από εργαλεία εντοπισμού συντηρημένων λειτουργικών περιοχών (conserved protein domains) [81], από εργαλεία πρόβλεψης τρισδιάστατης δομής όπως το HHpred [82] ή από εργαλεία που συνδυάζουν όλα τα παραπάνω σαν το EFICAz [83]. Οι εν λόγω αλγόριθμοι χρησιμοποιούν εξειδικευμένες βάσεις δεδομένων των χαρακτηριστικών στα οποία στηρίζονται για την κατηγοριοποίηση των αλληλουχιών οι οποίες είναι διαθέσιμες είτε ως αυτοτελή αρχεία [84, 85] ή ως κομμάτι του ίδιου του εργαλείου [83]. Οι τεχνικές αυτές μπορούν να αξιοποιηθούν για τον de novo χαρακτηρισμό πρωτεϊνών όταν υπάρχει μικρή ή και καθόλου ομοιότητα με ήδη γνωστές αλληλουχίες, κάτι που τις καθιστά καταλληλότερες για αξιοποίηση στην επεξεργασία μεταγενωμικών δεδομένων.

Ένα σημαντικό μειονέκτημα που παρουσιάζουν τα παραπάνω εργαλεία που βασίζονται σε χαρακτηριστικά αλληλουχιών, είναι ο απαιτούμενος χρόνος λειτουργίας. Συγκεκριμένα το EFICAz, ενώ δεν έχει σημαντικές απαιτήσεις σε εικονική μνήμη και υπολογιστική ισχύ, χρειάζεται αρκετές ώρες για την ταυτοποίηση μερικών δεκάδων αλληλουχιών κάτι που το κάνει ακατάλληλο για ανάλυση μεταγενωμικών δεδομένων. Αντ' αυτού, ο προτεινόμενος τρόπος χρήσης του είναι το φιλτράρισμα από την ανάλυση ομολογίας των πιο «ύποπτων» αλληλουχιών για συγκεκριμένη λειτουργικότητα και να γίνεται επιπλέον πρόβλεψη με τον αλγόριθμό του.

1.3.6 Ανακατασκευή μεταβολικών μονοπατιών

Ο πλήρης χαρακτηρισμός ενός μικροβιακού πληθυσμού σε ένα μεταγενωμικό δείγμα απαιτεί μία ολιστική προσέγγιση, που περιλαμβάνει την μελέτη σε κάθε βιολογικό επίπεδο, από το γονιδιακό έως το λειτουργικό, λαμβάνοντας υπόψη ταυτόχρονα το σύνολο των αλληλεπιδράσεων των διαφορετικών οργανισμών μεταξύ τους και με το περιβάλλον τους. Οι αλληλεπιδράσεις αυτές, καθώς και κάθε στάδιο της λειτουργίας των οργανισμών του εκάστοτε οικολογικού θώκου, καθίστανται

δυνατά μέσω του συνόλου των βιοχημικών αντιδράσεων που λαμβάνουν χώρα μέσα στα κύτταρά τους, ως μέρος του μεταβολισμού τους. Οι εν λόγω αντιδράσεις περιλαμβάνουν είτε οργανικές είτε ανόργανες ενώσεις στα αντιδρώντα ή/και προϊόντα τους, τα οποία ονομάζονται μεταβολίτες. Όλες σχεδόν οι μεταβολικές αντιδράσεις καταλύονται από ένζυμα[86], μία ειδική κατηγορία λειτουργικών πρωτεϊνών που συντίθενται μέσω της γονιδιακής έκφρασής στα κύτταρα, με σκοπό την κατάλυση, αλλά και την αύξηση του ρυθμού των συγκεκριμένων αντιδράσεων συνδέοντας έτσι την ρύθμιση του μεταβολικού δικτύου με το γονιδίωμα του κυττάρου.

Το σύστημα των μεταβολικών αντιδράσεων σε κάθε κύτταρο είναι ιδιαίτερα πολύπλοκο και για να μελετηθεί απαιτείται σε πρώτο στάδιο η καταγραφή κάθε μίας από αυτές. Από αυτή την καταγραφή προκύπτουν γραμμικές αναπαραστάσεις των σχέσεων μεταξύ των μεταβολιτών που ονομάζονται μεταβολικά μονοπάτια (metabolic pathways) που στο σύνολο τους μαζί με τους κυτταρικούς ρυθμιστικούς μηχανισμούς που τα διέπουν αποτελούν ένα μεταβολικό δίκτυο (metabolic network). Υπάρχουν ήδη χαρτογραφημένα τα μεταβολικά δίκτυα πολλών οργανισμών (μεταξύ τους και ανθρώπινων κυττάρων) και είναι διαθέσιμα σε δημόσιες βάσεις δεδομένων [61, 87-89]. Οι έννοιες μεταβολικό μονοπάτι και μεταβολικό δίκτυο ενώ έχουν οριστεί αρχικά για την μελέτη μεμονωμένων κυττάρων μπορούν να επεκταθούν στις περιπτώσεις μικροβιακών πληθυσμών σε μεταγενωμικά δείγματα. Οι σχέσεις αλληλεξάρτησης και ανταλλαγής μεταβολιτών μεταξύ των διαφορετικών μικροβιακών ειδών έχουν ως αποτέλεσμα την δημιουργία δυναμικών μεταβολικών δικτύων τα οποία είναι υψίστης σημασίας για την εύρυθμη λειτουργία ολόκληρης της μικροβιακής κοινότητας [90].Η μελέτη του μεταβολικού δυναμικού σε ένα τέτοιο σύστημα προϋποθέτει, λόγω της πολυπλοκότητας του, την in silico ανακατασκευή κάθε φορά ολόκληρου του μεταβολικού δικτύου του. Τα βιοπληροφορικά εργαλεία που έχουν αναπτυχθεί όπως το MEGAN [57], το MinPath [91] ή το Pathway Tools [92], για να καλύψουν αυτήν την ανάγκη αξιοποιούν αλγορίθμους που βασίζονται στις πληροφορίες για τις ενζυμικές λειτουργίες τροποποίησης των μεταβολιτών στα ήδη γνωστά μεταβολικά μονοπάτια μέσα στις προαναφερθείσες βάσεις δεδομένων. Έτσι, τον χαρακτηρισμό του συνόλου των γονιδίων και πρωτεϊνών ενός μικροβιακού πληθυσμού μέσω βιοπληροφορικής πρόβλεψης ακολουθεί η σωστή χαρτογράφησή τους στο εκτιμώμενο μεταβολικό δίκτυο που αντιπροσωπεύουν. Η χαρτογράφηση αυτή επιτρέπει την εξαγωγή συμπερασμάτων σχετικά με το σύνολο των πρωτεϊνικών

λειτουργιών που υφίστανται στο σύστημα και κατά συνέπεια των μεταβολικών μονοπατιών στα οποία εντάσσονται, αποσαφηνίζοντας τις μεταξύ τους σχέσεις όπως αυτές έχουν διαμορφωθεί λόγω της συνεχούς εξελικτικής πίεσης από το περιβάλλον.

1.4 Εφαρμογές της μεταγενωμικής στην υγεία

1.4.1 Εντοπισμός παθογόνων μικροοργανισμών

Οι κλασσικές τεχνικές προσδιορισμού λοιμωδών ασθενειών και των παθογόνων μικροοργανισμών που τις προκαλούν μπορούν να ταξινομηθούν σε δύο κύριες κατηγορίες: α)στις τεχνικές προσδιορισμού μέσω παρατήρησης συμπτωμάτων και β)στις τεχνικές προσδιορισμού μέσω εργαστηριακών αναλύσεων. Γίνεται αντιληπτό ότι οι εργαστηριακές τεχνικές προσφέρουν μία μεγαλύτερη ακρίβεια στον εντοπισμό και την επιβεβαίωση του εκάστοτε παθογόνου μικροοργανισμού ενώ οι τεχνικές βασισμένες στην παρατήρηση συμπτωμάτων αποτελούν μία οικονομικά και χρονικά βιώσιμη λύση για τις ανάγκες ερευνών σε πληθυσμιακό επίπεδο. Ο εργαστηριακός εντοπισμός αποτελείται από ένα μεγάλο εύρος πρωτοκόλλων ανάλυσης, που περιλαμβάνει από παραδοσιακές αναλύσεις μικροσκοπίας και καλλιεργειών, έως σύγχρονες δοκιμασίες ανάλυσης (assays) που εντοπίζουν αντιγόνα από τον παθογόνο οργανισμό ή προϊόντα της ανοσοποιητικής απόκρισης. Παρόλο το εύρος των διαθέσιμων εργαστηριακών αναλύσεων και της μεγαλύτερης ακρίβειας αυτών δεν εξασφαλίζεται πάντα η σωστή ταυτοποίηση των παθογόνων μικροοργανισμών και κατά συνέπεια των ασθενειών για τις οποίες ευθύνονται. Χαρακτηριστικά αναφέρεται ότι με τις συμβατικές εργαστηριακές τεχνικές ανάλυσης δεν γίνεται επιτυχής εντοπισμός των παθογόνων οργανισμών στο 40% των περιπτώσεων γαστρεντερίτιδας [93] ενώ το ποσοστό αυτό στις περιπτώσεις εγκεφαλίτιδας μπορεί να φτάσει το 60% [94]. Επιπλέον οι ταχύτητα αυτών των μεθόδων είναι πολλές φορές απαγορευτική σε περιπτώσεις όπου υπάρχει ανάγκη άμεσης ταυτοποίησης του παθογόνου που προκαλεί μία ασθένεια.

Το επόμενο βήμα για την αντιμετώπιση των προβλημάτων ταυτοποίησης αιτιατών σχέσεων μεταξύ μικροοργανισμών και των προκαλούμενων ασθενειών, περιλαμβάνει τον εντοπισμό γνωστών διατηρημένων γονιδιακών αλληλουχιών, που χρησιμοποιούνται ως βιοδείκτες, μέσω τεχνικών αλυσιδωτής αντίδρασης πολυμεράσης [95, 96]. Η αρχή αυτής της μεθόδου έχει επεκταθεί περαιτέρω με την υιοθέτηση μεταγενωμικών προσεγγίσεων κάτι που είναι πλέον εφικτό λόγω της σημαντικής μείωσης τους κόστους των τεχνολογιών αλληλούχισης. Ο εντοπισμός

αυτών των αλληλουχιών-βιοδεικτών μπορεί πλέον να γίνει με την εφαρμογή DAS τεχνικών προσφέροντας έτσι ένα νέο, πιο γρήγορο διαγνωστικό εργαλείο υψηλής ακρίβειας. Οι τεχνικές αυτές, εκτός του πλεονεκτήματος ότι δεν εξαρτώνται από την ανάπτυξη καλλιεργειών για τον χαρακτηρισμό μικροβιακών πληθυσμών, προσφέρουν και τη δυνατότητά εντοπισμού καινούριων παθογόνων στελεχών [12].

Παρ' όλη την ακρίβεια, την ταχύτητα και τη δυνατότητά εντοπισμού καινούριων παθογόνων που προσφέρουν, οι DAS τεχνικές υπόκεινται στα σφάλματα ποσοτικοποίησης της PCR, που έχουν ως αποτέλεσμα τη φαινομενική υπερ-αντιπροσώπευση ορισμένων στελεχών στα τελικά δεδομένα. Η μεταγενωμική προσέγγιση που μπορεί να ξεπεράσει αυτό το εμπόδιο, προσφέροντας ένα πιο αξιόπιστο ταξονομικό προφίλ για τη διάγνωση ενός βιολογικού δείγματος, στηρίζεται στην αξιοποίηση της αλληλούχισης τμηματικής ανάλυσης. Η πιο ολοκληρωμένη καταγραφή του μεταγονιδιωματικού περιεχομένου μέσω της αλληλούχισης τμηματικής ανάλυσης οδηγεί σε έναν πιο ακριβή ποσοτικό προσδιορισμό των στελεχών που υφίστανται εν ενεργεία, επιτρέποντας την διεξαγωγή πιο έγκυρων διαγνωστικών συμπερασμάτων. Τα δεδομένα που προκύπτουν από μία τέτοιου είδους ανάλυση μπορούν να οδηγήσουν και στον εντοπισμό βοηθητικών γονιδίων (accessory genes) [97], τα οποία αποτελούν την ουσιαστική διαφοροποιό δύναμη όσον αφορά την παθογένεια μεταξύ πολύ κοντινών μικροβιακών στελεχών, επιτρέποντας έτσι έναν πιο εκτεταμένο ταξονομικό διαχωρισμό σε αντίθεση με τις υπόλοιπες τεχνικές. Το κύριο μειονέκτημα τους σε σχέση με τις DAS τεχνικές όμως, παραμένουν οι υψηλές απαιτήσεις της βιοπληροφορικής ανάλυσης καθώς ο όγκος των δεδομένων που παράγουν είναι πολύ μεγαλύτερος κάτι που αυξάνει παράλληλα την πολυπλοκότητα και το χρόνο των αναλύσεων. Ένα επιπλέον μειονέκτημα, που μοιράζονται με τις DAS τεχνικές, έναντι των παραδοσιακών εργαστηριακών τεχνικών είναι το κόστος λειτουργίας τους [98], το οποίο όμως ενδέχεται να μειωθεί με τη συνεχή εμφάνιση καινούριων και βελτιστοποιημένων τεχνολογιών αλληλούχισης.

1.4.2 Ανθρώπινο μικροβίωμα του γαστρεντερικού συστήματος

Εκτός από τον εντοπισμό παθογόνων μικροοργανισμών που αποτελούν κίνδυνο για τη δημόσια υγεία, η μεταγενωμική προσφέρει εργαλεία για την μελέτη των μικροβιακών στελεχών από τα οποία αποτελείται η φυσιολογική χλωρίδα του ανθρώπινου σώματος και των οποίων ο πληθυσμός είναι τόσο υψηλός που είναι της ίδιας τάξης μεγέθους με τον αριθμό του συνόλου των ανθρωπίνων κυττάρων [99]. Το

πιο χαρακτηριστικό παράδειγμα εφαρμογής της μεταγενωμικής προς αυτό τον σκοπό, είναι η εκτεταμένη μελέτη του μικροβιώματος του ανθρώπινου γαστρεντερικού συστήματος.

Το μικροβίωμα του γαστρεντερικού συστήματος αποτελείται από περισσότερα από 1000 διαφορετικά μικροβιακά είδη τα οποία περιλαμβάνουν από ευκαρυωτικούς οργανισμούς έως αρχαία και βακτήρια [100] με τα τελευταία μάλιστα να βρίσκονται σε πολύ μεγαλύτερο ποσοστό. Η σημαντικότητα του για την έρρυθμη λειτουργία του οργανισμού είναι τεράστια, καθώς τον προστατεύει από άλλους παθογόνους μικροοργανισμούς [101], συμμετέχει στην ρύθμιση του ανοσοποιητικού συστήματος [102], όπως και πολλών μεταβολικών διαδικασιών [103]. Η ολιστική μελέτη λοιπόν του εν λόγω οικοσυστήματος αποδεικνύεται εξαιρετικής σημασίας για την κατανόηση της συνεχούς αλληλεπίδρασης του με τον υπόλοιπο οργανισμό και την επίδρασή του στην ενδεχόμενη εμφάνιση νοσημάτων.

Η πρώτες προσπάθειες αποκάλυψης του ταξονομικού προφίλ του μικροβιώματος του γαστρεντερικού συστήματος βασίστηκαν σε κλασσικές τεχνικές ανάπτυξης αποικιών στο εργαστήριο. Παρ' όλα αυτά αποδείχτηκε ότι με τις τεχνικές αυτές μπορούσαν να καλλιεργηθούν επιτυχώς το 10%-30% του συνόλου των μικροβιακών ειδών [104]. Η εφαρμογή των εργαλείων της μεταγενωμικής σε αυτήν την περίπτωση έχει οδηγήσει στο να αντιμετωπίζεται αυτό το πρόβλημα, με τις τεχνικές DAS και αλληλούχισης τμηματικής ανάλυσης να είναι σε θέση πλέον, να δώσουν μία ολοκληρωμένη εικόνα των διαφορετικών μικροβιακών ειδών που συνυπάρχουν στο γαστρεντερικό σύστημα [105]. Με την περαιτέρω ανάλυση των δεδομένων αλληλούχισης τμηματικής ανάλυσης μέσω των βιοπληροφορικών εργαλείων, μπορούμε να εμβαθύνουμε επιπλέον την ανάλυση μας, είτε με την καταγραφή των γονιδίων που υπάρχουν [106] ή που εκφράζονται [107] στο μικροβίωμα κάτω από συγκεκριμένες συνθήκες π.χ. λόγω του προκαλούμενου στρες από κάποια ασθένεια. Τα ίδια αυτά εργαλεία μας επιτρέπουν επίσης να ανακατασκευάσουμε in silico τα μεταβολικά μονοπάτια που διέπουν ολόκληρο το βιολογικό σύστημα [108] του μικροβιώματος του γαστρεντερικού συστήματος και να διεξάγουμε συμπεράσματα για το πλήρες λειτουργικό δυναμικό του.

1.5 Εφαρμογές της μεταγενωμικής στη βιομηχανία και περιβαλλοντική μηχανική

1.5.1 Εντοπισμός και απομόνωση βιοκαταλυτών

Η μεταγενωμική ανάλυση μικροβιακών πληθυσμών δεν περιορίζεται μόνο στις περιπτώσεις που υπάρχει άμεση επίδραση τους στην ανθρώπινη υγεία. Οι μικροοργανισμοί που υπάρχουν ελεύθεροι στη φύση μπορούν να αξιοποιηθούν μέσω της απομόνωσης των ενζύμων που παράγουν κατά το μεταβολισμό τους και της αξιοποίησής τους σε βιομηχανικές διεργασίες. Τα μικροβιακά ένζυμα αποτελούν ένα εξαιρετικά μεγάλο σύνολο βιοκαταλυτών, με αντίστοιχα τεράστιο εύρος χημικών αντιδράσεων που είναι σε θέση να καταλύσουν. Αυτό οφείλεται στη μικροβιακή ποικιλομορφία και στις αντίστοιχες πολύπλοκες μεταβολικές ανάγκες όπως αυτές έχουν διαμορφωθεί μέσω της εξελικτικής πίεσης που υφίστανται οι οργανισμοί σε κάθε διαφορετικό οικολογικό θώκο.

Η θερμοκρασία, η οξύτητα, τα διαθέσιμα θρεπτικά συστατικά και η πρόσβαση σε οξυγόνο είναι μόνο μερικοί από τους περιβαλλοντικούς παράγοντες που επηρεάζουν την ανάπτυξη του μικροβιακού πληθυσμού και οδηγούν την επικράτηση συγκεκριμένων στελεχών. Η επικράτηση τους αυτή, συνδέεται άμεσα με την δυνατότητά τους να φέρουν εξειδικευμένα ένζυμα, μέσω των αντίστοιχων γονιδίων τους, για την βέλτιστη αξιοποίηση αυτών των παραγόντων (ή προστασία από αυτούς). Αυτό έχει ως αποτέλεσμα, το μεταγονιδίωμα τους να περιέχει τελικά πληθώρα ενζύμων με τις λειτουργίες τους προσαρμοσμένες στις συνθήκες που τους περιβάλλουν π.χ. μεταγονιδιώματα πληθυσμών σε πηγές πλούσιες σε ανόργανο άνθρακα περιλαμβάνουν γονίδια αφομοίωσης αυτού του άνθρακα ως πηγή ενέργειας [109] ή μεταγονιδιώματα μικροβίων από θερμές πηγές φέρουν ένζυμα υψηλής θερμοσταθερότητας [110].

Η τεράστια βιοποικιλότητα που υπάρχει στη φύση λοιπόν, οδηγεί σε ένα ουσιαστικά τεράστιο δυναμικό ενζυμικών λειτουργιών, οι οποίες μπορούν να ανακαλυφθούν μόνο μέσω των εργαλείων μεταγενωμικής για περαιτέρω αξιοποίηση από την βιομηχανία σε βιοτεχνολογικές εφαρμογές [111]. Υπάρχουν ήδη τεχνικές εντοπισμού και απομόνωσης γονιδίων γνωστής λειτουργικότητας από μικροοργανισμούς μέσα σε ένα μεταγονιδίωμα όπως η επαγωγή γονιδιακής έκφρασης μέσω υποστρώματος (substrate-induced gene expression screening - SIGEX). Οι τεχνικές αυτές βασίζονται στην επικράτηση των οργανισμών, είτε σε εργαστηριακές καλλιέργειες είτε in situ, που περιλαμβάνουν τα απαραίτητα γονίδια ώστε να καταβολίσουν συγκεκριμένα υποστρώματα. Όμως με αυτόν τον τρόπο υπάρχει πάντα

ένα σημαντικό μέρος της γονιδιακής πληροφορίας το οποίο χάνεται. Μικροοργανισμοί που υπήρχαν σε μικρότερο ποσοστό λόγω αδυναμίας ανταγωνιστικής ανάπτυξης δεν εντοπίζονται και τα αντίστοιχα γονίδια που πιθανόν να ήταν μέχρι και υψηλότερου βιομηχανικού ενδιαφέροντος παραλείπονται. Τη λύση σε αυτό το πρόβλημα προσφέρει η αλληλούχιση τμηματικής ανάλυσης, καθώς σε συνδυασμό με τη βιοπληροφορική ανάλυση των δεδομένων της, καθιστά πλέον εφικτή την καταγραφή ολόκληρου του μεταγονιδιώματος ενός οικολογικού θώκου και τον εντοπισμό όλων των γονιδίων που υπάγονται σε αυτό. Με την πληροφορία που προκύπτει για τις γονιδιακές αλληλουχίες που αντιστοιχούν σε ένζυμα υψηλού βιομηχανικού ενδιαφέροντος, προερχόμενα από κάθε στέλεχος του μεταγενωμικού δείγματος, μπορεί στη συνέχεια να γίνει έκφραση είτε εργαστηριακά, είτε σε μεγάλη κλίμακα των εν λόγω ενζύμων και η παραγωγή και απομόνωσή τους για μετέπειτα χρήση.

1.5.2 Βιομηχανία τροφίμων

Η αντικατάσταση χημικών διεργασιών από αντιδράσεις που καταλύονται από ένζυμα έχει επικρατήσει σε όλους τους κλάδους της βιομηχανίας τα τελευταία χρόνια, με τον τομέα τροφίμων να ξεχωρίζει αισθητά στην προσπάθεια αυτή π.χ. η υδρόλυση του αμύλου. Για τη βιομηχανία τροφίμων οι ενζυμικές διεργασίες δεν είναι κάτι καινούριο, καθώς αξιοποιούνται ένζυμα από φυσικές πηγές εδώ και χιλιάδες χρόνια (έστω και ακούσια) για την παραγωγή ή/και την βελτιστοποίηση των προϊόντων της. Αυτό οφείλεται στην τεράστια διαθεσιμότητα των ενζύμων στη φύση και το αντίστοιχα μεγάλο εύρος δυνατοτήτων τους που καθιστά τη χρησιμότητά τους αδιαμφισβήτητη σε διεργασίες επεξεργασίας τροφίμων. Επιπλέον πλεονέκτημα αποτελεί και η αξιοσημείωτη βιοσυμβατότητά τους καθώς πολλά από αυτά τα ένζυμα προέρχονται από βιολογικές διεργασίες αποδόμησης των θρεπτικών συστατικών που λαμβάνονται μέσω της τροφής. Ζυμώσεις αρτοπαρασκευασμάτων, αλκοολούχων και γαλακτοκομικών προϊόντων, αντιδράσεις επεξεργασίας χρώματος, υφής και γεύσης τροφίμων, καθώς και παραγωγής φυσικών χυμών, είναι μόνο η κορυφή του παγόβουνου για τις διεργασίες που έχουν ωφεληθεί από την χρήση ενζύμων και για τις οποίες διεξάγεται συνεχής έρευνα για περαιτέρω βελτιστοποίηση.

Πρωταγωνιστικό ρόλο σε πολλές από αυτές τις διεργασίες κατέχουν οι υδρολάσες, μία κατηγορία ενζύμων που ειδικεύονται στην υδρόλυση ομοιοπολικών δεσμών. Χαρακτηριστικά αναφέρεται ότι οι λιπάσες, μία υποκατηγορία των

υδρολασών, χρησιμοποιούνται κυρίως στην γαλακτοβιομηχανία για την βελτίωση των δειγματοληπτικών χαρακτηριστικών του γάλακτος, στην επεξεργασία φυτικών ελαίων και στην συντήρηση αρτοπαρασκευασμάτων [112]. Άλλα υδρολυτικά ένζυμα που χρησιμοποιούνται στη γαλακτοβιομηχανία είναι οι β-γαλακτοζιδάσες οι οποίες υδρολύουν την λακτόζη σε γαλακτόζη και γλυκόζη, με τελικό σκοπό την παραγωγή προϊόντων καταλλήλων για άτομα με δυσανεξία στη λακτόζη, καθώς και την βελτίωση της γεύσης τους [113]. Για τη βελτίωση της γεύσης των γαλακτοκομικών προϊόντων χρησιμοποιούνται επίσης και οι πρωτεάσες, οι οποίες υδρολύουν πεπτιδικούς δεσμούς με αποτέλεσμα την αποδόμηση πρωτεϊνών. Οι πρωτεάσες χρησιμοποιούνται επίσης σε διεργασίες τεχνητής μαλάκωσης κρεάτων [114], καθώς επίσης και σε διεργασίες πρωτεόλυσης της γλουτένης για την δημιουργία αρτοπαρασκευασμάτων χωρίς την εν λόγω ουσία [115]. Μία άλλη υποκατηγορία υδρολασών, οι εστεράσες χρησιμοποιούνται για την μεταποίηση λιπών και ελαίων και για την παραγωγή ορισμένων αρωμάτων και γεύσεων σε χυμούς φρούτων και αλκοολικά ροφήματα [116].

Παρόλα τα πλεονεκτήματα που παρέχουν (ταχύτεροι χρόνοι αντίδρασης, δυνατότητα του βιοκαταλύτη να απομακρυνθεί και να επαναχρησιμοποιηθεί κτλ) οι ενζυμικές διεργασίες, πολύ συχνά περιορίζονται λόγω των δυνατοτήτων χρήσης των ίδιων των ενζύμων στις συνθήκες που απαιτούνται για την χημική αντίδραση. Υψηλές θερμοκρασίες, υψηλή αλατότητα και ακραίες τιμές pH είναι μόνο μερικές από τις συνθήκες που μπορούν να καταστήσουν ανενεργό ένα ένζυμο ακόμα και μόνιμα, καθώς διαταράσσουν την τεταρτοταγή του δομή (μετουσίωση). Οι ιδιότητες των ενζύμων και η ανθεκτικότητά τους σε διάφορες συνθήκες καθορίζεται σε μεγάλο βαθμό από την προέλευση του ενζύμου και τις εξελικτικές πιέσεις που οδήγησαν στην παρούσα μορφή του. Τα περισσότερα ένζυμα που χρησιμοποιούνταν στη βιομηχανία τροφίμων μέχρι πρόσφατα ήταν ζωικής προέλευσης και επομένως έτσι διαμορφωμένα ώστε να λειτουργούν σε μικρό εύρος συνθηκών αντίδρασης. Παρ' όλα αυτά, η τεράστια ποικιλία διαθέσιμων ενζύμων που έχουν εντοπιστεί από μεταγονιδιώματα έχει οδηγήσει στην αυξανόμενη εφαρμογή τους για την κατάλυση των ίδιων διεργασιών [117].

Σε αντίθεση με τα ένζυμα ζωικής προέλευσης, τα αντίστοιχα που προέρχονται από μεταγενωμικές πηγές, ακολουθώντας τις μεταβολικές ανάγκες των μικροβιακών στελεχών στα οποία ανήκουν, επιδεικνύουν ένα πολύ μεγαλύτερο εύρος συνθηκών στις οποίες είναι λειτουργικά. Έτσι τα νέα ένζυμα που εντοπίζονται από

μεταγονιδιώματα έχουν συχνότερα βελτιωμένες ιδιότητες, όπως υψηλότερη ενεργότητα, θερμοσταθερότητα, λειτουργικότητα σε ακραίες τιμές οξύτητας κτλ. Ιδιότητες που έδιναν την δυνατότητα στους μικροοργανισμούς να τα χρησιμοποιήσουν για την εξασφάλιση της επιβίωσής τους και την προσαρμογή τους στις εκάστοτε περιβαλλοντικές συνθήκες πλέον αξιοποιούνται για την ρύθμιση και βελτιστοποίηση διεργασιών στην παρασκευή και επεξεργασία τροφίμων.

1.5.3 Φαρμακοβιομηχανία

Η αντικατάσταση χημικών διεργασιών με αντίστοιχες ενζυμικές είναι μία τάση που έχει υιοθετηθεί και από τις βιομηχανίες φαρμάκων τα τελευταία χρόνια είτε για την άμεση σύνθεση βιοενεργών μορίων (bioactives), είτε για την έμμεση παραγωγή τους μέσω της σύνθεσης ενδιάμεσων ουσιών. Τα ένζυμα αυτά προκύπτουν μέσω τις απομόνωσης γονιδίων από μεταγονιδιώματα και την ετερόλογη έκφρασή τους, με σκοπό την μαζική παραγωγή τους σε βιομηχανική κλίμακα. Ένα χαρακτηριστικό παράδειγμα αποτελεί η ετερόλογη έκφραση γονιδίων που αποτελούν ένα μεταβολικό μονοπάτι σύνθεσης της βιοτίνης [118] (βιταμίνη Η), ενός συνένζυμου για καρβοξυλικά ένζυμα που ανήκει στο σύμπλεγμα βιταμινών Β. Παραδείγματα άλλων βιοενεργών μορίων που αξιοποιούνται από την βιομηχανία φαρμάκων και έχουν απομονωθεί από μεταγονιδιώματα μικροβιακών πληθυσμών αποτελούν η σερπίνη [119] (serpin), ένας αναστολέας πρωτεασών και η μπορεγκομυκίνη [120] (borremycin) μία ουσία με αντιαυξητική(antiproliferative) και αντιβιωτική δράση.

Η εφαρμογή της μεταγενωμικής για την απομόνωση χρήσιμων βιομορίων δεν σταματάει στην αξιοποίηση μόνο περιβαλλοντικών δειγμάτων. Γονίδια που αντιστοιχούν σε χρήσιμα βιομόρια έχουν απομονωθεί από μεταγονιδιώματα της φυσικής χλωρίδας που συνυπάρχει σε όργανα ζώων. Η πεδερίνη (pederin) για παράδειγμα, μία ουσία με αντινεοπλαστική δράση, έχει απομονωθεί από βακτήρια που συμβιώνουν σε σκαθάρια του γένους Paederus [121]. Αντίστοιχα, ένζυμα με λειτουργία συνθάσης πολυκετιδίων (PKS), έχουν απομονωθεί από μικροβιακούς πληθυσμούς πάνω σε θαλάσσια σφουγγάρια του γένους Discodermia [122]. Αξίζει να σημειωθεί ότι τα πολυκετίδια αποτελούν δευτερογενείς μεταβολίτες με μεγάλο εύρος ιδιοτήτων που συμπεριλαμβάνουν εκτός των άλλων αντιβιωτικές, αντιμυκητιακές έως και αντικαρκινικές δράσεις [123]. Το εύρος των ζωικών ειδών τα οποία έχουν μελετηθεί μέσω μεταγενωμικής ανάλυσης για τον εντοπισμό και απομόνωση χρήσιμων βιομορίων εκτείνεται πολύ περαιτέρω και τα τελευταία χρόνια έχει φτάσει

μέχρι και τον άνθρωπο με αποτέλεσμα την απομόνωση πλήθους βιομορίων από το μεταγονιδίωμα μικροβιώματος του γαστρεντερικού συστήματος [117].

1.5.4 Βιομηχανία βιοκαυσίμων

Η εκθετική αύξηση της χρήσης ορυκτών καυσίμων από την βιομηχανική επανάσταση του 19^{ου} αιώνα και μετά, έχει οδηγήσει σε σημαντική μείωση των κοιτασμάτων τους και στην εστίαση της έρευνας προς εναλλακτικές πηγές ενέργειας. Μία από τις κύριες λύσεις στο ενεργειακό πρόβλημα δόθηκε με την εμφάνιση των βιοκαυσίμων, δηλαδή καυσίμων με παραπλήσιες ιδιότητες με τα αντίστοιχα ορυκτά καύσιμα, αλλά που η παραγωγή τους λαμβάνει χώρα μέσω βιολογικών διεργασιών (σε αντίθεση με τις γεωχημικές που είναι υπεύθυνες για τα ήδη υπάρχοντα αποθέματα ορυκτών καυσίμων). Οι ουσίες που χρησιμοποιούνται ως βιοκαύσιμα περιλαμβάνουν βιοαιθανόλη, βιοντίζελ, βιοβουτανόλη και βιοαέριο και η παραγωγή τους βασίζεται στην βιολογική επεξεργασία υποστρωμάτων όπως αμυλούχα και ελαιοπαραγωγικά φυτά, κτηνοτροφικά και αγροτικά απόβλητα, διαφόρων ειδών σάκχαρα και λιγνινοκυτταρική βιομάζα [124].

Η επεξεργασία των παραπάνω υποστρωμάτων περιλαμβάνει επίπονες διεργασίες από άποψη των αντίστοιχων χημικών αντιδράσεων, οι οποίες στηρίζονται κατά κύριο λόγο στη χρήση ενζύμων. Για την παραγωγή βιοαιθανόλης, για παράδειγμα, είναι αναγκαία η κατεργασία πολυσακχαριτών (άμυλο, κυτταρίνη ή ημικυτταρίνη), ώστε να διασπαστούν σε σάκχαρα μικρότερου μοριακού βάρους που θα μπορούν να λάβουν μέρος σε αντιδράσεις αλκοολικής ζύμωσης. Η κατεργασία αυτή βασίζεται κυρίως στη δράση υδρολυτικών ενζύμων όπως κυτταρινάσες και ημικυτταρινάσες, η οποία όμως παρεμποδίζεται αισθητά από την παρουσία λιγνίνης στο υπόστρωμα [125]. Ομοίως, η παραγωγή βιοντίζελ, ως μη τοξικό και βιοδιασπώμενο υποκατάστατο του συμβατικού ντίζελ, απαιτεί ενζυμικές αντιδράσεις τρανσ-εστεροποίησης μεταξύ ελαίων και αλκοολών για την παραγωγή μεθυλεστέρων. Τα ένζυμα που καταλύουν τις αντιδράσεις αυτές είναι λιπάσες ή εστεράσες και έχουν αντικαταστήσει τις παραδοσιακές τεχνικές βασισμένες σε ισχυρές βάσεις ως καταλύτες, που παρουσίαζαν σημαντικά μειονεκτήματα, όπως τη δυσκολία απομάκρυνσής τους από το τελικό προϊόν κάτι που είχε ως τελικό αποτέλεσμα, αλκαλικά υδατικά διαλύματα να αποτελούν μέρος των αποβλήτων [126].

Η αδιαμφισβήτητη ανάγκη αποδοτικότερης και πιο μαζικής χρήσης βιοκαυσίμων έχει οδηγήσει στην χρήση μεταγενωμικών τεχνικών για τον εντοπισμό νέων ενζύμων, που μπορούν να βελτιστοποιήσουν τις αντιδράσεις παραγωγής τους. Η εκτεταμένη έρευνα σε μεταγονιδιώματα τόσο σε περιβαλλοντικά όσο και σε βιολογικά δείγματα τα τελευταία χρόνια, έχει οδηγήσει στον εντοπισμό πληθώρας ενζύμων με κατάλληλες ιδιότητες για την βελτιστοποίηση των διεργασιών παραγωγής βιοκαυσίμων [127]. Εκτός όμως από την άμεση απομόνωση ενζύμων για την κατάλυση των προαναφερθέντων διεργασιών, η μεταγενωμική προσέγγιση σε συνδυασμό με τις τεχνικές ανασυνδυασμού γονιδίων, που έχουν αναπτυχθεί από τον κλάδο της βιοτεχνολογίας, μας προσφέρει τη δυνατότητα αξιοποίησης ολόκληρων μικροοργανισμών για τον ίδιο σκοπό. Ένα μεγάλο εύρος μικροβιακών ειδών, από κοινούς σακχαρομύκητες [128], έως τα πιο ανθεκτικά εξτρεμόφιλα [124] έχουν ανακαλυφθεί μέσω μεταγενωμικής ανάλυσης, και είναι πλέον στη διάθεση της βιομηχανίας βιοκαυσίμων, με αποτέλεσμα την αξιοποίηση των εξειδικευμένων μεταβολικών μονοπατιών που διαθέτουν για την παραγωγή ουσιών υψηλής ενεργειακής περιεκτικότητας.

1.5.5 Περιβαλλοντική μηχανική

Η κατανόηση του λειτουργικού δυναμικού των μικροβιακών πληθυσμών μέσω των μεταγενωμικών τεχνικών, μας επιτρέπει την στοχευμένη χρήση τους για τη διαχείριση περιπτώσεων εκτεταμένης περιβαλλοντικής ρύπανσης. Αυτού του είδους η προσέγγιση μπορεί να διακριθεί σε δύο κύριες κατηγορίες: α)στη χρήση μικροβιακών πληθυσμών για τον εντοπισμό περιβαλλοντικών ρύπων και β)στη χρήση τους για τη βιολογική αποκατάσταση (bioremediation) οικοσυστημάτων που έχουν υποστεί μόλυνση.

Στην πρώτη περίπτωση, ακολουθούνται μέθοδοι ταξονομικής ανάλυσης, βασισμένοι σε μεταγενωμικά δεδομένα, για τον προσδιορισμό της μικροβιακής σύστασης του πληθυσμού ενός δείγματος που προέρχεται από περιβαλλοντική πηγή εκτεθειμένη σε κάποιο ρύπο. Με τη σύγκριση των δεδομένων ανάλυσης μεταξύ δειγμάτων εκτεθειμένων σε περιβαλλοντικούς ρυπαντές και δειγμάτων μη εκτεθειμένων, αλλά παρόμοιας γεωχημικής προέλευσης, μπορούμε να διακρίνουμε στελέχη που ευνοούνται σε κάθε μία από τις δύο περιπτώσεις. Αυτό μπορεί να παρατηρηθεί μέσω της αύξησης του πληθυσμού τους, η οποία συμβαίνει λόγω της διαφορετικής ικανότητας του εκάστοτε στελέχους να μεταβολίσει την ουσία-ρυπαντή,

χρησιμοποιώντας την για την ανάπτυξή του. Κατά αυτό τον τρόπο η εύρεση χαρακτηριστικών αλληλουχιών για τα αναπτυχθέντα στελέχη μπορεί να μας δώσει έγκυρους βιοανιχνευτές για τον εντοπισμό περιβαλλοντικής μόλυνσης μέσω μεταγενωμικής ανάλυσης. Παραδείγματα τέτοιων μεθόδων περιλαμβάνουν τον εντοπισμό ουρανίου και νιτρικών ιόντων σε μέρη ταφής πυρηνικών αποβλήτων [129] καθώς και τον εντοπισμό διαρροών αποθηκευμένου CO_2 (carbon capture and storage - CCS) σε υποθαλάσσια υποστρώματα [130].

Στην περίπτωση βιολογικής αποκατάστασης, ο στόχος είναι η αξιοποίηση οργανισμών που μπορούν να μεταβολίσουν κάποιον ρυπαντή ώστε να επιτευχθεί η οριστική απομάκρυνσή του. Ο τρόπος εντοπισμού των εν λόγω οργανισμών και των αντίστοιχων γονιδίων που συνεισφέρουν στην βιοδιάσπαση (biodegration) ενός ρυπαντή περιλαμβάνει και πάλι της μεθόδους ανάλυσης μεταγενωμικών δεδομένων. Τον εντοπισμό ακολουθεί η βιοδιέγερση (biostimulation) του πληθυσμού με τέτοιο τρόπο ώστε να ευνοηθεί η υπέρμετρη αύξηση των επιθυμητών στελεχών που θα απομακρύνουν τις ρυπαντικές ουσίες [131].

1.6 Αδυναμίες των σύγχρονων μέσων μεταγενωμικής ανάλυσης

Η διαρκής βελτίωση των μεθόδων αλληλούχισης DNA, καθώς και η εμφάνιση της αλληλούχισης επόμενης γενιάς, έχει μετατρέψει την μεταγενωμική σε έναν ταχύτατα εξελισσόμενο κλάδο με σχεδόν απεριόριστες προοπτικές. Η ταχύτητα, η αυξημένη απόδοση και η υψηλή ακρίβεια των τεχνικών αυτών σε συνδυασμό με το σημαντικά μειωμένο τους κόστος, τις καθιστούν ένα εύχρηστο εργαλείο για την πλήρη καταγραφή του (μετα)γονιδιακού περιεχομένου σε ένα δείγμα που περιλαμβάνει κάποιον μικροβιακό πληθυσμό. Παρ' όλα αυτά, οι νέες τεχνικές εμφανίζουν ορισμένα μειονεκτήματα που δυσχεραίνουν σημαντικά την ερμηνεία των αποτελεσμάτων τους.

Το πρώτο από αυτά είναι ο όγκος δεδομένων που παράγουν. Υψηλότερη ακρίβεια στα αποτελέσματα αλληλούχισης προϋποθέτει υπερκάλυψη κάθε νουκλεοτιδίου του μεταγονιδιώματος αρκετές φορές από αντίστοιχα θραύσματα αλληλούχισης, ώστε να θεωρείται αξιόπιστη η καταγραφή του. Αυτή η υπερκάλυψη που ονομάζεται και βάθος της αλληλούχισης (sequencing depth) αποτελεί και ένα μέτρο της ποιότητας των δεδομένων ενός πειράματος αλληλούχισης π.χ. έχει προκύψει εμπειρικά ότι κάθε νουκλεοτίδιο πρέπει να είναι πάνω από 30 φορές επικαλυπτόμενο κατά μέσο όρο ώστε να θεωρούνται αξιόπιστα τα αποτελέσματα

στην περίπτωση εκ νέου συναρμολόγησης ενός ανθρώπινου γονιδιώματος [132]. Η υψηλή απόδοση των τεχνικών αλληλούχισης δεύτερης γενιάς, μας επιτρέπει την παραγωγή αρκετών αναγνωσμένων αλληλουχιών ώστε να είναι εφικτή η παραπάνω ποιοτική προδιαγραφή αλλά ως συνέπεια τα παραγόμενα δεδομένα είναι μεγάλου όγκου ακόμα και μικρού μεγέθους γονιδιώματα. Η τάξη μεγέθους των δεδομένων για παράδειγμα από ένα απλό πείραμα αλληλούχισης ξεκινάει από μερικά GB και μπορεί να φτάσει μερικές δεκάδες, έως και εκατοντάδες GB ανάλογα και με την τεχνολογία που χρησιμοποιείται, το βάθος της αλληλούχισης και το ίδιο το μέγεθος του δείγματος. Αυτός ο όγκος δεδομένων μπορεί μάλιστα να αυξηθεί αρκετές τάξεις μεγέθους με τις μετέπειτα αναλύσεις, καθώς για κάθε αλληλουχία θα προκύπτουν τα αντίστοιχα δεδομένα αποτελεσμάτων από τα βιοπληροφορικά εργαλεία. Για να δοθεί ένα σημείο αναφοράς σχετικά με το μέγεθος των δεδομένων αυτών αναφέρεται συγκριτικά, ότι ένα αρχείο κειμένου 1000 σελίδων είναι μικρότερο από 1 MB. Γίνεται κατανοητό λοιπόν ότι τέτοιου μεγέθους αρχεία παρουσιάζουν δυσκολίες στην αποθήκευση και απαιτούν εξειδικευμένες γνώσεις πληροφορικής για τη σωστή διαχείριση και επεξεργασία τους. Μέχρι και το απλό, υπό κανονικές συνθήκες, άνοιγμα και διάβασμα ενός αρχείου, καταντά υπολογιστικά επίπονη διαδικασία και ανθρωπίνως αδύνατο αν λάβουμε υπόψη τα δισεκατομμύρια γραμμών που το αποτελούν.

Το επόμενο πρόβλημα που παρουσιάζεται αφορά την βελτιστοποίηση της ανάλυσης των μεταγενωμικών δεδομένων. Η πλήρη κατανόηση του μεταγενωμικού περιεχομένου ενός δείγματος απαιτεί μία σειρά πολλαπλών διαφορετικών αναλύσεων από τα αντίστοιχα βιοπληροφορικά εργαλεία, τα οποία λειτουργούν είτε διαδοχικά (δηλαδή το κάθε εργαλείο εφαρμόζει την ανάλυσή του στα δεδομένα αποτελεσμάτων που προκύπτουν από το προηγούμενο) ξεκινώντας από τα δεδομένα αλληλούχισης, είτε εν παραλλήλω (δηλαδή πολλαπλά εργαλεία λειτουργούν πάνω στα ίδια σετ δεδομένων) [19]. Αυτού του είδους η διαδικασία παρουσιάζει σημαντικές δυσκολίες τόσο στην εποπτεία της, όσο και στη σωστή οργάνωσή της λόγω του όγκου των δεδομένων, αλλά και της μακρόχρονης διάρκειας ανάλυσης που χαρακτηρίζει πολλά από τα υπολογιστικά εργαλεία [133]. Επιπλέον το κάθε διαφορετικό εργαλείο απαιτεί τη σωστή ρύθμιση των παραμέτρων λειτουργίας του ώστε να επιτευχθεί ή βέλτιστη απόδοσή του. Οι παράμετροι αυτοί δεν είναι πάντα εύκολο να επιλεγούν καθώς εξαρτώνται από το μέγεθος και τύπο των δεδομένων, τις δυνατότητες της υπολογιστικής υποδομής και την ευαισθησία της ανάλυσης που απαιτείται. Η

σημαντικότητα των παραμέτρων αυτών είναι τέτοια που μπορούν να προκαλέσουν σημαντικές διαφορές όχι μόνο στην ποιότητα των αποτελεσμάτων, αλλά και στους χρόνους λειτουργίας των ίδιων των εργαλείων. Μία ανάλυση ομολογίας, για παράδειγμα, σε μεταγενωμικά δεδομένα με το εργαλείο BLAST θα μπορούσε ανάλογα τις παραμέτρους που θα εισάγουμε και την υπολογιστική υποδομή που χρησιμοποιούμε, να διαρκέσει από μερικές ώρες έως και δεκάδες μέρες. Για κάθε βήμα της ανάλυσης των δεδομένων λοιπόν, καθίσταται απαραίτητη όχι μόνο μία βαθιά γνώση βιοπληροφορικής από τον χρήστη αλλά και η αντίστοιχη εμπειρία πάνω στο εκάστοτε εργαλείο για την βελτιστοποίηση της λειτουργίας του, τόσο από θέμα ταχύτητας όσο και από θέμα έγκυρων αποτελεσμάτων. Αυτό βέβαια αποκτά ιδιαίτερη δυσκολία, αν αναλογιστεί κανείς την πληθώρα εργαλείων που είναι απαραίτητα για μία βιοπληροφορική ανάλυση και τον διαφορετικό τρόπο λειτουργίας του καθενός.

Οι προαναφερθείσες αδυναμίες που παρουσιάζουν τα διαθέσιμα βιοπληροφορικά εργαλεία τα καθιστούν δύσχρηστα και κάποιες φορές μη αξιόπιστα, καθώς η αξιολόγηση των αποτελεσμάτων τους πρέπει να λαμβάνει υπόψη αν οι αρχικές παράμετροι που χρησιμοποιήθηκαν ήταν οι κατάλληλες. Επίσης, ο υπερβολικά μεγάλος χρόνος λειτουργίας τους, σε συνδυασμό με την ανάγκη διαδοχικής ή παράλληλης ανάλυσης των δεδομένων οδηγεί σε σημαντικές απώλειες χρόνου λειτουργίας της υπολογιστικής υποδομής εάν δεν έχει αυτοματοποιηθεί με κάποιο τρόπο η διαδικασία. Η ανάγκη αυτή της σωστής οργάνωσης και αυτοματοποίησης γίνεται ακόμα πιο εμφανής, αν λάβουμε υπόψη ότι ορισμένα εργαλεία απαιτούν πολύ μεγαλύτερους πόρους υπολογιστικού συστήματος από άλλα και μπορούν να αποτελέσουν το κυρίαρχο εμπόδιο (bottleneck) στην όλη ανάλυση. Η σωστή διαχείριση των εργαλείων με σκοπό την αυτοματοποίησή τους, προϋποθέτει εξειδικευμένες γνώσεις πληροφορικής κάτι το οποίο στερούνται οι περισσότεροι ερευνητές βιολογικών επιστημών. Αντίστοιχα οι ερευνητές πληροφορικής συνήθως στερούνται τις απαραίτητες βιολογικές γνώσεις ώστε να ερμηνεύσουν τα αποτελέσματα των εργαλείων αυτών με σκοπό τη βελτιστοποίηση της λειτουργίας τους.

Έχουν γίνει ήδη στο παρελθόν [134] αρκετές προσπάθειες για τη γεφύρωση των γνώσεων αυτών με την κατασκευή αυτοματοποιημένων πλατφόρμων ανάλυσης μεταγενωμικών δεδομένων. Παρ' όλα αυτά, καμία από τις υπάρχουσες πλατφόρμες ως τώρα δεν προσφέρει μία αυτοματοποιημένη λύση, που να συνδυάζει τις δυνατότητες μίας πλήρους ανάλυσης μεταγενωμικών δεδομένων με την αναγκαία

ευκολία στη χρήση. Οι περισσότερες από αυτές δεν περιλαμβάνουν ένα ολοκληρωμένο «οπλοστάσιο» εργαλείων ώστε να αντιμετωπίσουν όλες τις ανάγκες ανάλυσης που περιλαμβάνονται σε ένα μεταγενωμικό δείγμα και περιορίζονται σε συγκεκριμένες λειτουργίες όπως εύρεση και χαρακτηρισμό γονιδίων [135] ή ταξονομική ανάλυση [136]. Ακόμα και οι βιοπληροφορικές πλατφόρμες που καταφέρνουν να συμπεριλάβουν ένα πλήρες σετ εργαλείων για την διεξοδική ανάλυση των δεδομένων [137], υστερούν στη φιλικότητα προς το χρήστη απαιτώντας εξειδικευμένες γνώσεις πληροφορικής για την εγκατάσταση ή/και για τη σωστή λειτουργία τους. Έχοντας υπόψη τις παραπάνω ελλείψεις στον τομέα αυτόν, η παρούσα διδακτορική διατριβή εστιάστηκε στην κάλυψη αυτών των αναγκών με τον σχεδιασμό μίας αυτοματοποιημένης πλατφόρμας εργαλείων ανάλυσης μεταγενωμικών δεδομένων μέσω ενός φιλικού προς το χρήστη διαδικτυακού περιβάλλοντος και την εφαρμογή της για τον εντοπισμό νέων ενζύμων βιομηχανικού ενδιαφέροντος.

ΚΕΦΑΛΑΙΟ 2: Παρουσίαση Μεθοδολογίας

2.1 Ανάπτυξη υπολογιστικής πλατφόρμας

2.1.1 Υπολογιστική υποδομή και προγράμματα ανάλυσης

Η ανάπτυξη ενός εξειδικευμένου υπολογιστικού συστήματος για μεταγενωμικά δεδομένα ξεκίνησε με την αξιολόγηση των διαθέσιμων βιοπληροφορικών εργαλείων ανοιχτού κώδικα που έχουν δημοσιευθεί και μπορούν να αναλάβουν το πλήθος των διαφορετικών αναλύσεων. Η αξιολόγηση και η επικείμενη διαλογή έγινε τόσο μέσω βιβλιογραφικής έρευνας, όσο και με πρακτική εφαρμογή του κάθε εργαλείου για τις ανάγκες της οποίας αξιοποιήθηκαν προσωπικοί υπολογιστές, καθώς και διακομιστές για όσα προγράμματα υπήρχαν υψηλές υπολογιστικές απαιτήσεις. Οι διακομιστές που χρησιμοποιήθηκαν για τις πρακτικές εφαρμογές των εργαλείων ήταν οι εξής:

- Ένα σετ πέντε διακομιστών με ονόματα «grissomdevweb.ekt.gr», «grissomdevweb.vima.ekt.gr», «mebioinfo.ekt.gr», «grissom.vima.ekt.gr» και «grissomweb.vima.ekt.gr», οι οποίοι ανήκουν στο Εθνικό Κέντρο Τεκμηρίωσης του Εθνικού Ιδρύματος Ερευνών και στους οποίους η πρόσβαση χορηγήθηκε από το Ινστιτούτο Βιολογίας, Φαρμακευτικής Χημείας και Βιοτεχνολογίας του ίδιου ιδρύματος. Το σετ των παραπάνω διακομιστών χρησιμοποιήθηκε για τις πρώτες δοκιμαστικές αναλύσεις δεδομένων, καθώς και για τον προσδιορισμό των προαπαιτούμενων πληροφορικών πακέτων για την εγκατάσταση και σωστή λειτουργία των επιλεγμένων εργαλείων. Καθένας από τους παραπάνω διακομιστές είχε λειτουργικό σύστημα Ubuntu 14.04.3 με μνήμη RAM 2GB και 2 επεξεργαστές (CPUs).
- Ένας διακομιστής με όνομα «Helios» που άνηκε στο πανεπιστήμιο της Κοπεγχάγης και στον οποίο η πρόσβαση εξασφαλίστηκε στα πλαίσια του ερευνητικού προγράμματος «HotZyme». Στον διακομιστή αυτό αναπτύχθηκε η δοκιμαστική (beta) έκδοση μίας αυτοματοποιημένης πλατφόρμας μεταγενωμικών αναλύσεων (βλ. κεφάλαιο 2.1.4), στην οποία ενσωματώθηκαν τα εργαλεία που επιλέχθηκαν, ώστε να χρησιμοποιηθούν για την ανάλυση των μεταγενωμικών δεδομένων που προέκυψαν από το προαναφερθέν πρόγραμμα. Ο διακομιστής αυτός είχε λειτουργικό σύστημα Ubuntu 14.04.3 με μνήμη RAM 256GB και 32 επεξεργαστές.

- Ένα σετ δύο διακομιστών με ονόματα «motherbox» και «tyranistar» που άνηκε στο ΕΜΠ και αποκτήθηκε στα πλαίσια του ερευνητικού προγράμματος «COVERALL» [138]. Η αξιοποίηση του σετ των δύο αυτών διακομιστών οδήγησε στον εντοπισμό και επιδιόρθωση υπολογιστικών σφαλμάτων (bugs) και τη διαμόρφωση της τελικής μορφής της αυτοματοποιημένης πλατφόρμας αναλύσεων, η οποία κατέστη διαθέσιμη διαδικτυακά μέσω του motherbox. Καθένας από τους παραπάνω διακομιστές είχε λειτουργικό σύστημα CentOS 7-3.1611, μνήμη RAM 512GB και 64 επεξεργαστές.

Τα εργαλεία που εγκαταστάθηκαν στους παραπάνω διακομιστές αξιολογήθηκαν ως τα πιο κατάλληλα από άποψη απόδοσης και αξιοπιστίας αποτελεσμάτων κατά τη διάρκεια εκπόνησης αυτής της διατριβής, ενώ η διαλογή τους περιορίστηκε σε ένα ή δύο ανά στάδιο ανάλυσης ώστε να μειωθεί η πολυπλοκότητα του τελικού υπολογιστικού συστήματος.

Η αρχική επεξεργασία των δεδομένων αλληλούχισης εξασφαλίστηκε επιλέγοντας ένα πλήθος εργαλείων τόσο για τον έλεγχο, όσο και για την απομάκρυνση βάσεων ή ολόκληρων αναγνωσμένων αλληλουχιών χαμηλής ποιότητας. Καθώς η ανάλυση ποιοτικού ελέγχου είναι μία διαδικασία που δεν απαιτεί ιδιαίτερη υπολογιστική ισχύ, η επιλογή στηρίχθηκε περισσότερο στην ευχρηστία του κάθε εργαλείου. Το εργαλείο ποιοτικού ελέγχου που επιλέχθηκε ήταν το FastQC όχι μόνο για την ολοκληρωμένη ανάλυση που προσφέρει (βλ. κεφάλαιο 1.3.1) αλλά και για τις δυνατότητες οπτικοποίησης των αποτελεσμάτων που προσφέρει, καθώς και την ταυτόχρονη ενσωμάτωσή τους σε ευανάγνωστα αρχεία αναφορών τύπου html. Αντίστοιχα για την επεξεργασία των αλληλουχιών χαμηλής ποιότητας επιλέχθηκε η σουίτα εργαλείων FASTX καθώς το εύρος αναλύσεων που μπορεί να αναλάβει καλύπτει πλήρως τις ανάγκες επεξεργασίας των δεδομένων αλληλούχισης και έτσι μπορεί να χρησιμοποιηθεί συνδυαστικά με το FastQC, στοχεύοντας τις προβληματικές περιοχές των αναγνωσμένων αλληλουχιών. Άλλος ένας λόγος για την επιλογή των δύο παραπάνω προγραμμάτων, ήταν ότι υπάρχουν για αυτά διαθέσιμοι αλγόριθμοι «επικάλυψης» (wrapper scripts) οι οποίοι επιτρέπουν την ενσωμάτωσή τους στην πλατφόρμα Galaxy (βλ. κεφάλαιο 2.1.4) και τη λειτουργία τους ως μέρος μίας αυτοματοποιημένης γραμμής πληροφορικών εργασιών (pipeline).

Τα δύο στάδια της βιοπληροφορικής ανάλυσης για τον προσδιορισμό της μικροβιακής σύστασης ενός δείγματος διακρίνονται στην ανάλυση ομοιότητας των δεδομένων αλληλούχισης με γνωστές αλληλουχίες και στην αντιστοίχιση των

αναγνωσμένων αλληλουχιών σε ταξονομικές κατηγορίες. Για το πρώτο στάδιο τα εργαλεία με τις περισσότερες προοπτικές ήταν η σουίτα εργαλείων BLAST και το πρόγραμμα DIAMOND. Το BLAST αποτέλεσε την πρωταρχική επιλογή λόγω της αξιοπιστίας των αποτελεσμάτων του, ενώ το DIAMOND εξετάστηκε λόγω της, συγκριτικά με το BLAST, πολύ υψηλής ταχύτητας του, η οποία όμως δεν έχει τόσο υψηλό αντίκτυπο στην απώλεια πληροφορίας [78]. Τελικά αποφασίστηκε μόνο ένα εκ των δύο να ενσωματωθεί στο τελικό υπολογιστικό σύστημα για χάριν απλούστευσης του, αλλά το DIAMOND παρέμεινε εγκατεστημένο στον ίδιο διακομιστή, ώστε να είναι διαθέσιμο για προσθήκη σε επόμενες εκδόσεις. Για το δεύτερο στάδιο επιλέχθηκε το πρόγραμμα MEGAN, καθώς εκτός από την ανάλυση ταξονομικού χαρακτηρισμού που προσφέρει, διαθέτει και αλγορίθμους εντοπισμού μεταβολικών μονοπατιών αξιοποιώντας οντολογίες όπως η KEGG, η SEED κτλ. Επίσης το πρόγραμμα αυτό έχει ευρείες δυνατότητες οπτικοποίησης και επεξεργασίας των δεδομένων αποτελεσμάτων του τόσο μέσω γραμμής εντολών όσο και μέσω γραφικού περιβάλλοντος.

Για την ανάλυση συναρμολόγησης εξετάστηκαν προγράμματα βασισμένα σε DBG αλγορίθμους καθώς από συγκριτικές έρευνες [139] βρέθηκε ότι μπορούν να διαχειριστούν καλύτερα αρχεία δεδομένων αλληλούχισης μεγαλύτερου μεγέθους ($>10^7$ αναγνωσμένες αλληλουχίες) όπως αυτά που προκύπτουν από μεταγενωμικά δείγματα. Από τα διάφορα DBG προγράμματα συναρμολόγησης που εξετάστηκαν, επιλέχθηκαν το Velvet [69] και το Megahit [140]. Το Velvet αποτέλεσε την αρχική επιλογή καθώς η σύγκριση με άλλα προγράμματα συναρμολόγησης έδειξε ότι έχει σχετικά μικρότερες απαιτήσεις σε εικονική μνήμη RAM καθώς και σε χρόνους ανάλυσης [139]. Επίσης διαθέτει ενσωματωμένους αλγορίθμους τόσο για τη διαχείριση των δεδομένων αλληλούχισης (shuffleSequences.pl), όσο και για τον αυτόματο προσδιορισμό των βέλτιστων παραμέτρων του (VelvetOptimiser) με βάση τα χαρακτηριστικά (π.χ. N50) των τελικών αποτελεσμάτων συναρμολόγησης. Τέλος, η προαναφερθείσα επέκτασή του, MetaVelvet [71], το καθιστά μία πιο εξειδικευμένη λύση για μεταγενωμικά δεδομένα αλληλούχισης, από την οποία προκύπτουν βελτιωμένα αποτελέσματα κατά την ανάλυση συναρμολόγησης. Παρ' όλα αυτά κατά την αξιολόγηση του προγράμματος και της επέκτασής του, μέσω πρακτικής εφαρμογής, βρέθηκαν προβλήματα συμβατότητας μεταξύ τους που καθιστούσαν αδύνατη την ανάλυση των δεδομένων αλληλούχισης. Η αιτία αυτής της ασυμβατότητας εντοπίστηκε να είναι το g++ [141], ένα από τα προαπαιτούμενα

προγράμματα για τη «μεταγλώττιση» (compiling) των εργαλείων Velvet και MetaVelvet. Μετά από επικοινωνία με την προγραμματιστική ομάδα του MetaVelvet αναγνωρίστηκε το πρόβλημα ως υπολογιστικό σφάλμα (bug) και προτάθηκε, ως προσωρινή λύση, η υποβάθμιση του g++ σε παλαιότερη έκδοση (4.7.2) με την οποία το εργαλείο θα ήταν πλέον συμβατό. Η πρόταση αυτή θεωρήθηκε μη βιώσιμη για ένα ενιαίο υπολογιστικό σύστημα, καθώς για την εγκατάσταση των υπόλοιπων εργαλείων ανάλυσης ήταν απαραίτητη μία από τις πιο πρόσφατες εκδόσεις του g++. Το πρόγραμμα Velvet, που δεν εμφάνιζε κάποια ασυμβατότητα, διατηρήθηκε στη λίστα διαθέσιμων προγραμμάτων, καθώς η υπολογιστική ομάδα του MetaVelvet ενημέρωσε ότι το πρόβλημα θα διορθωθεί σε επόμενες εκδόσεις, ενώ ως εναλλακτική λύση για την ανάλυση μεταγενωμικών δεδομένων αλληλούχισης επιλέχθηκε το Megahit. Το Megahit αποδείχτηκε μία ακόμα πιο συμφέρουσα επιλογή από άποψη υπολογιστικών απαιτήσεων και ταχύτητας κατά την αξιολόγηση μέσω πρακτικής εφαρμογής, ενώ διαθέτει επιλογές αυτοματοποιημένης ρύθμισης των παραμέτρων του για προσαρμογή και βελτιστοποίηση της ανάλυσης στην περίπτωση μεταγενωμικών δεδομένων.

Τα εργαλεία που επιλέχθηκαν για τον εντοπισμό γονιδίων ήταν το Getorf [72] και το Prodigal[74]. Το Getorf επιλέχθηκε ως μέρος της σουίτας EMBOSS [72], η οποία περιλαμβάνει και άλλα χρήσιμα εργαλεία για αναλύσεις δεδομένων νουκλεοτιδικών ή πρωτεϊνικών αλληλουχιών. Το Getorf προσφέρει την δυνατότητα εντοπισμού όλων των πιθανών ανοικτών πλαισίων ανάγνωσης στα δεδομένα συναρμολόγησης ενός δείγματος, αφήνοντας τη διαδικασία χαρακτηρισμού τους εξολοκλήρου σε μετέπειτα αναλύσεις. Αντίθετα το Prodigal χρησιμοποιεί αλγορίθμους απόρριψης ανοικτών πλαισίων ανάγνωσης ως μη κωδικές περιοχές και μπορεί να χρησιμοποιηθεί για πιο στοχευμένες αναλύσεις μειώνοντας σημαντικά τον χρόνο και τον όγκο δεδομένων τους. Ο εντοπισμός γονιδίων γίνεται κατά μήκος όλων των συναρμολογημάτων του δείγματος που μπορεί να προέρχονται από οργανισμούς με πολλαπλά αντίγραφα κάποιων γονιδίων ή από διαφορετικούς οργανισμούς με όμοια γονίδια. Αυτό έχει ως αποτέλεσμα οι λίστες γονιδίων που προκύπτουν από τα παραπάνω εργαλεία να έχουν έχουν πολλές πανομοιότυπες (ή σε πολύ υψηλό ποσοστό όμοιες) αλληλουχίες. Η απομάκρυνση του προβλήματος πλεονασμού (redundancy) των δεδομένων εξασφαλίσθηκε με την επιλογή του CD-HIT [142] ως εργαλείο ομαδοποίησης (clustering) όμοιων αλληλουχιών.

Ο χαρακτηρισμός των αλληλουχιών που προκύπτουν από την προηγούμενη ανάλυση μπορεί να επιβεβαιώσει την πρόβλεψή τους ως πιθανά γονίδια και να προσφέρει μία επιπλέον υπόθεση για την λειτουργική τους υπόσταση. Την πιο συνηθισμένη λύση για μία τέτοιου είδους ανάλυση αποτελεί η ανάλυση ομολογίας μέσω της σουίτας εργαλείων BLAST, η οποία είχε ήδη εγκατασταθεί, μαζί με τις απαραίτητες βάσεις δεδομένων (NCBI-nr, NCBI-nt, UniProt/SwissProt), για τις ανάγκες του ταξονομικού προσδιορισμού. Για την συγκριτική ανάλυση αλληλουχιών μπορούν να χρησιμοποιηθούν, εκτός από βάσεις δεδομένων ολόκληρων γονιδίων και βάσεις δεδομένων από αλληλουχίες που αποτελούν χαρακτηριστικές λειτουργικές περιοχές (domains) όπως η Pfam [84] η οποία και εγκαταστάθηκε. Το πρόγραμμα HMMER [143] επιλέχθηκε για την αποτελεσματική αξιοποίηση της βάσης δεδομένων Pfam-A καθώς προσφέρει τη δυνατότητα εύρεσης ομοιοτήτων μεταξύ αλληλουχιών, μέσω αξιοποίησης της μεθοδολογίας των κρυμμένων μοντέλων Markov (Hidden Markov Models). Το HMMER όπως και το BLAST είναι εργαλεία που μπορούν να παραμετροποιηθούν κατάλληλα ώστε να αξιοποιούν πολλαπλούς επεξεργαστές στο υπολογιστικό σύστημα στο οποίο είναι εγκατεστημένα, κάτι που τα καθιστά αρκετά ευέλικτα για την ανάλυση μεταγενωμικών δεδομένων μεγάλου μεγέθους. Δεδομένου όμως ότι ένα μεταγονιδίωμα μπορεί να αποτελείται από πλήθος γονιδιωμάτων από οργανισμούς οι οποίοι δεν έχουν παρατηρηθεί ποτέ, έπρεπε να βρεθούν εργαλεία που δεν βασίζονται μόνο σε βάσεις δεδομένων ήδη γνωστών αλληλουχιών. Το πρώτο εργαλείο λειτουργικής πρόβλεψης που εγκαταστάθηκε ήταν το EFICAz [83] το οποίο στηρίζεται σε μεθόδους μηχανικής μάθησης και μπορεί να προβλέψει την λειτουργία αλληλουχιών με ακρίβεια τεσσάρων ψηφίων της αριθμητικής κατηγοριοποίησης Ενζυμικής Επιτροπής (Enzyme Commision - EC number). Κατά τη διάρκεια όμως της πρακτικής εφαρμογής του παρατηρήθηκε ότι απαιτεί τεράστιους χρόνους ανάλυσης, οι οποίοι ήταν απαγορευτικοί για τις τεράστιες λίστες γονιδίων που προέκυπταν από μεταγενωμικά δεδομένα. Αντ' αυτού προτιμήθηκε η λύση της υπολογιστικής γραμμής εργασιών PROKKA [144], η οποία συνδυάζει μεταξύ άλλων το Prodigal ως εργαλείο εντοπισμού κωδικών περιοχών και βάσεις δεδομένων κρυφών μοντέλων Markov για λειτουργικές περιοχές, μέσω του HMMER, ώστε να διεξάγει συμπεράσματα σχετικά με το λειτουργικό χαρακτηρισμό κάθε αλληλουχίας. Ταυτόχρονα η ανάλυση με το PROKKA παράγει αρχεία που μπορούν να χρησιμοποιηθούν από άλλα εργαλεία για την οπτικοποίηση των αποτελεσμάτων.

Έχοντας μια λίστα χαρακτηρισμένων αλληλουχιών με βάση τη λειτουργία τους, το επόμενο βήμα είναι να μελετήσουμε τα μεταβολικά μονοπάτια στα οποία ανήκουν. Για την μελέτη αυτή επιλέχθηκε το πρόγραμμα MinPath [91] καθώς αποτελεί ένα συντηρητικό, όσον αφορά τα αποτελέσματα, αλλά αξιόπιστο βιοπληροφορικό εργαλείο για την ανακατασκευή μεταβολικών μονοπατιών τόσο σε μεμονωμένα γονιδιώματα όσο και σε μεταγονιδιώματα. Το MinPath χρησιμοποιήθηκε σε συνδυασμό με τη βάση δεδομένων KEGG, στην οποία γίνεται αντιστοίχιση κάθε μεταβολικού μονοπατιού με τις λειτουργίες, εκφρασμένες σε αριθμούς EC, από τις οποίες αποτελείται. Εκτός από το MinPath, ο εντοπισμός των μεταβολικών μονοπατιών ενός δείγματος είναι δυνατός και μέσω του προγράμματος MEGAN, που είχε ήδη εγκατασταθεί για τις ανάγκες τις ταξονομικής ανάλυσης. Το MEGAN μάλιστα δεν έχει ως προαπαιτούμενο τη συναρμολόγηση και τον εντοπισμό γονιδίων. Αντ' αυτού, μπορεί να χρησιμοποιηθεί άμεσα στα δεδομένα αλληλούχισης χρησιμοποιώντας αλγορίθμους ταξινόμησης των αναγνωσμένων αλληλουχιών, αλλά αυτή τη φορά κάθε χρησιμοποιούμενη κατηγορία αντιστοιχίζεται σε ένα μεταβολικό μονοπάτι. Κατ' αυτόν τον τρόπο προσφέρει επιπλέον και πληροφορίες ποσοτικοποίησης για κάθε μεταβολικό μονοπάτι, βασισμένες στον αριθμό αναγνωσμένων αλληλουχιών που έχουν κατηγοριοποιηθεί σε κάθε μεταβολικό μονοπάτι. Με τη χρήση των παραπάνω εργαλείων είτε ξεκινώντας από το επίπεδο των δεδομένων αλληλούχισης ή από τη λίστα πρωτεϊνών, που χαρακτηρίστηκαν με βάση τη λειτουργία τους, προκύπτει μία αντίστοιχη λίστα μεταβολικών μονοπατιών επιτρέποντας έτσι την αποσαφήνιση του λειτουργικού δυναμικού ενός μεταγενωμικού δείγματος.

Η εφαρμογή των παραπάνω εργαλείων παράγει δεδομένα των οποίων το μέγεθος μπορεί να διαφέρει από μερικά MB έως και αρκετά GB καθιστώντας τη διαχείρισή τους ένα εξαιρετικά δύσκολο έργο υπολογιστικά. Για την αντιμετώπιση αυτού του προβλήματος αυτό εγκαταστάθηκε ο διακομιστής σχεσιακών βάσεων δεδομένων MySQL. Ο διακομιστής αυτός, έχοντας τη δικιά του γλώσσα προγραμματισμού, διαθέτει δυνατότητες ανάπτυξης αλγορίθμων για την αυτοματοποίηση της εισαγωγής και διαχείρισης μεγάλων αρχείων σε εξειδικευμένες βάσεις δεδομένων. Επίσης υπάρχουν εργαλεία οπτικοποίησης (π.χ. MySQL Workbench) των δεδομένων, τα οποία μπορούν να εγκατασταθούν σε προσωπικούς υπολογιστές και να προσφέρουν πρόσβαση στα δεδομένα μέσω ενός εύχρηστου γραφικού περιβάλλοντος χρήστη. Για την οπτικοποίηση των βάσεων δεδομένων,

εγκαταστάθηκε το περιβάλλον σχεδιασμού και ανάπτυξης διαδικτυακών εφαρμογών, Web2py. Οι εφαρμογές που μπορούσαν να αναπτυχθούν μέσω του περιβάλλοντος αυτού, είχαν δυνατότητες σύνδεσης με τη βάση δεδομένων και παρουσίασης των δεδομένων τους στο χρήστη μέσω ενός διαδικτυακού γραφικού περιβάλλοντος.

2.1.2 Ανάπτυξη εξειδικευμένων αλγορίθμων ανάλυσης

Τα διαθέσιμα βιοπληροφορικά εργαλεία που εξετάστηκαν και επιλέχθηκαν, δεν θα μπορούσαν από μόνα τους να καλύψουν εξ ολοκλήρου τις ανάγκες μίας πλήρους ανάλυσης μεταγενωμικών δεδομένων. Ανάγκες όπως η αυτοματοποίηση της διαδοχικής λειτουργίας των εργαλείων, η βέλτιστη παραμετροποίηση τους πριν από κάθε αναλυτικό βήμα και η διαχείριση δεδομένων εισόδου και εξόδου από αυτά ήταν μερικά από τα προβλήματα που αντιμετωπίστηκαν με την ανάπτυξη εξειδικευμένων αλγορίθμων ως εκτελέσιμα αρχεία που ονομάζονται «σενάρια» (scripts). Παρακάτω αναφέρονται οι διαφορετικοί αλγόριθμοι που αναπτύχθηκαν και περιγράφεται με ποια εργαλεία ήταν συμβατοί και σε ποια σημεία της ανάλυσης μπορούσαν να λειτουργήσουν βοηθητικά ή/και συμπληρωματικά.

Ο ταξονομικός προσδιορισμός ενός δείγματος ξεκινάει με την ανάλυση ομολογίας των αναγνωσμένων αλληλουχιών μέσω του εργαλείου BLAST. Κατά τη διάρκεια της πρακτικής εφαρμογής του εργαλείου όμως, παρατηρήθηκε ότι οι αλγόριθμοι που το αποτελούν δεν μπορούν να διαχειριστούν δεδομένα FASTA στα οποία οι αναγνωριστικοί κωδικοί των αλληλουχιών περιέχουν κενά, όταν ζητείται συγκεκριμένη μορφή δεδομένων εξόδου (π.χ. tabular). Λόγω αυτής της αδυναμίας των εργαλείων, τα τελικά αποτελέσματα της ανάλυσης περιλαμβάνουν τις εκάστοτε αλληλουχίες με διαφορετικούς αναγνωριστικούς κωδικούς από εκείνους στο αρχικό αρχείο εισόδου, οι οποίοι μάλιστα έχουν υποστεί περικοπή των χαρακτήρων τους μετά το πρώτο κενό. Αυτό έχει ως αποτέλεσμα αλληλουχίες με παρόμοιους αναγνωριστικούς κωδικούς, όπως π.χ. στην περίπτωση δεδομένων ζεύγους άκρων, να αντιμετωπίζονται ως πανομοιότυπες κατά την καταγραφή των αποτελεσμάτων. Για παράδειγμα δύο αλληλουχίες στο αρχείο FASTA με αναγνωριστικούς κωδικούς «seq_paired 1» και «seq_paired 2» αντίστοιχα, θα μετατραπούν σε αλληλουχίες με αναγνωριστικό κωδικό «seq_paired» προκαλώντας σύγχυση στην ερμηνεία των αποτελεσμάτων τους. Για το λόγο αυτό αναπτύχθηκε ένας αλγόριθμος ως σενάριο (fasta_names.py) σε γλώσσα Python [145], ο οποίος μετατρέπει το αρχείο αναγνωσμένων αλληλουχιών μορφής FASTA σε ένα παρόμοιο, με τη μόνη διαφορά

μεταξύ των δύο να είναι ότι τα κενά στους αναγνωριστικούς κωδικούς έχουν αντικατασταθεί με τον χαρακτήρα κάτω παύλα «_». Έτσι στο παραπάνω παράδειγμα, μετά την εφαρμογή του σεναρίου, οι δύο αλληλουχίες αποκτούν τους νέους αναγνωριστικούς κωδικούς «seq_paired_1» και «seq_paired_2» αντίστοιχα.

Για την καλύτερη διαχείριση του προγράμματος MEGAN κατασκευάστηκαν τρία σενάρια «συντακτικής ανάλυσης» (parser scripts) σε γλώσσα Python τα οποία καλούσαν το πρόγραμμα εισάγοντας τις απαραίτητες παραμέτρους και αρχεία εισόδου για τη σωστή λειτουργία του. Το πρώτο σενάριο (blastsub.py), αξιοποιούσε τα δεδομένα εξόδου της ανάλυσης ομολογίας (είτε με το πρόγραμμα BLAST ή με το πρόγραμμα DIAMOND) και ανέλυε τα δεδομένα αλληλούχισης, κατασκευάζοντας ένα αρχείο μορφής .rma με τα χαρακτηριστικά κάθε αναγνωσμένης αλληλουχίας. Το αρχείο .rma που προέκυπτε μπορούσε να αξιοποιηθεί από το MEGAN μέσω των δύο επόμενων σεναρίων (megansub.py και functionalsub.py), για την κατασκευή του ταξονομικού και λειτουργικού προφίλ αντίστοιχα των δεδομένων. Οι αλγόριθμοι ταξινόμησης αναγνωσμένων αλληλουχιών των δύο σεναρίων εξήγαγαν τα δεδομένα αποτελεσμάτων τους, είτε ως αρχεία κειμένου είτε ως εικόνες ενώ ταυτόχρονα τα αποθήκευαν και σε συμπιεσμένη μορφή (.tar.gz) για ευκολία στη διαχείρισή τους. Οι αλγόριθμοι αυτοί αναπτύχθηκαν, ώστε να παρακάμψουν το διαδραστικό περιβάλλον εργασίας γραμμής εντολών του MEGAN και να προσφέρουν μία άμεση μέθοδο τροφοδότησης του με τα δεδομένα και παραμέτρους, η οποία μπορούσε να αξιοποιηθεί αργότερα για την αυτοματοποίηση της λειτουργίας του. Επίσης το αρχείο .rma που προέκυπτε από την εφαρμογή του πρώτου σεναρίου μπορούσε να μεταφερθεί σε προσωπικό υπολογιστή και να επαναχρησιμοποιηθεί από μία τοπική εγκατάσταση του MEGAN, ώστε να επαναληφθούν οι αναλύσεις μέσα από γραφικό περιβάλλον και να εξαχθούν τα αποτελέσματα σε οπτικοποιημένη μορφή.

Τα δεδομένα αλληλούχισης, εκτός από τα εργαλεία ταξονομικής ανάλυσης, αξιοποιούνται ταυτόχρονα και από εργαλεία εκ νέου συναρμολόγησης. Για το πρόγραμμα συναρμολόγησης Megahit αναπτύχθηκε ένας αλγόριθμος ως σενάριο συντακτικής ανάλυσης (megahit.py) σε γλώσσα Python με τον οποίο γίνεται η εισαγωγή των παραμέτρων και δεδομένων εισόδου ενώ ταυτόχρονα γίνεται διαχείριση των δεδομένων εξόδου. Με την εφαρμογή του συγκεκριμένου αλγορίθμου, η λίστα συναρμολογημάτων αποθηκεύεται ως αρχείο FASTA, ενώ τα υπόλοιπα αρχεία συμπιέζονται σε ένα ξεχωριστό αρχείο .tar.gz, ώστε να καθίσταται εύκολη η μεταφορά και η περαιτέρω επεξεργασία τους. Για έναν επιπλέον ποιοτικό

έλεγχο (εκτός αυτού που εφαρμόζεται από το εργαλείο Megahit μετά το πέρας της ανάλυσης) των δεδομένων συναρμολογημάτων που προκύπτουν αναπτύχθηκε ένας ακόμα αλγόριθμος σε γλώσσα προγραμματισμού Python (count_mapped.py) ο οποίος, καλώντας το εργαλείο Samtools, καταγράφει τον αριθμό των αναγνωσμένων αλληλουχιών που μπορούν να χαρτογραφηθούν πίσω στα συναρμολογήματα.

Από την επεξεργασία των δεδομένων συναρμολόγησης από τα εργαλεία εύρεσης γονιδίων προκύπτουν λίστες γονιδίων αποθηκευμένες σε αρχεία FASTA. Τα αρχεία αυτά, στην περίπτωση μεταγενωμικών δεδομένων, αποτελούνται από μερικές χιλιάδες έως αρκετές εκατοντάδες χιλιάδες αλληλουχίες πιθανών γονιδίων. Για την ευκολότερη διαχείριση των δεδομένων αυτών αναπτύχθηκε ένας αλγόριθμος ως σενάριο σε γλώσσα προγραμματισμού Python (sequence_parser.py), ο οποίος αποθηκεύει τα δεδομένα αυτά ως πίνακα σε μία βάση δεδομένων MySQL και στη συνέχεια τα εξάγει ως αρχείο τύπου .sql. Το αρχείο .sql που προκύπτει μπορεί να επανεισαχθεί σε οποιαδήποτε τοπική εγκατάσταση ενός διακομιστή βάσης δεδομένων MySQL, όπου μπορούν είτε να οπτικοποιηθούν μέσω των αντίστοιχων συμβατών προγραμμάτων, είτε να χωριστούν σε υποσύνολα υψηλού ενδιαφέροντος μέσω των υψηλών δυνατοτήτων εξειδικευμένης αναζήτησης που προσφέρει η ίδια η γλώσσα προγραμματισμού.

Ακολουθώντας την ίδια μεθοδολογία για την αποθήκευση αλληλουχιών σε βάση δεδομένων MySQL, αναπτύχθηκαν δύο επιπλέον αλγόριθμοι (blast_parser.py και hmmer_parser.py) σε γλώσσα προγραμματισμού Python για την αποθήκευση των αποτελεσμάτων ανάλυσης από τα εργαλεία σύγκρισης ομοιότητας, BLAST και HMMER, σε βάση δεδομένων MySQL. Όπως και στην προηγούμενη περίπτωση, τα δεδομένα αποθηκεύονται σε μορφή πίνακα ο οποίος στη συνέχεια εξάγεται ως αρχείο τύπου .sql. Η αντίστοιχη επεξεργασία των δεδομένων μέσω διακομιστή MySQL προσφέρει έναν εύκολο τρόπο φιλτραρίσματος των αποτελεσμάτων, που επιτρέπει να εστιάσουμε σε αλληλουχίες πιθανών λειτουργιών υψηλού ενδιαφέροντος οι οποίες θα μπορούν να ταξινομηθούν, μεταξύ άλλων και με βάση το στατιστικό σκορ ομοιότητας τους.

Για την αυτοματοποίηση της υπολογιστικής γραμμής εργασιών PROKKA αναπτύχθηκε ένας αλγόριθμος συντακτικής ανάλυσης ως σενάριο σε γλώσσα προγραμματισμού Python (prokka.py) ο οποίος τροφοδοτεί το πρόγραμμα με τα δεδομένα συναρμολόγησης και τις απαραίτητες παραμέτρους για τη λειτουργία του. Τα αποτελέσματα που προκύπτουν από την εφαρμογή του PROKKA αναλύονται

περαιτέρω (parsing) ώστε να εξαχθεί η λίστα με τους αριθμούς EC για τους οποίους έχει βρεθεί κάποια αλληλουχία και καταγράφονται σε αρχείο κειμένου .txt ενώ οι εν λόγω αλληλουχίες καταγράφονται σε αρχεία FASTA στη νουκλεοτιδική και στην αμινοξική τους μορφή. Τα υπόλοιπα συμπληρωματικά αρχεία αποτελεσμάτων συμπιέζονται όλα μαζί σε αρχείο τύπου .tar.gz για την εύκολη διαχείριση τους. Αναπτύχθηκε επίσης ένας αλγόριθμος συντακτικής ανάλυσης και για το δεύτερο εργαλείο πρόβλεψης πρωτεϊνικής λειτουργίας, το EFICAZ. Όπως και στο PROKKA, ο αλγόριθμος στοχεύει στην αυτοματοποίηση της λειτουργίας του εργαλείου τροφοδοτώντας το με τα απαραίτητα δεδομένα και παραμέτρους ενώ ταυτόχρονα καταγράφει τα αποτελέσματα των αριθμών EC που βρέθηκαν σε αρχείο κειμένου .txt.

Για τη σωστή λειτουργία του MinPath σχεδιάστηκαν αλγόριθμοι σε γλώσσα Bash [146] με την αξιοποίηση των οποίων χαρτογραφήθηκαν όλες οι λειτουργίες γνωστών πρωτεϊνών ανά μεταβολικό μονοπάτι. Για να επιτευχθεί αυτό χρησιμοποιήθηκε η βάση δεδομένων KEGG στην οποία τα μεταβολικά μονοπάτια εμφανίζονται ως οντολογίες και στα οποία έχουν αντιστοιχιστεί πρωτεϊνικές λειτουργίες μέσω του αριθμού EC τους. Η εφαρμογή των αλγορίθμων αυτών οδήγησε στην καταγραφή της λίστας αριθμών EC για κάθε μεταβολικό μονοπάτι μέσω της ιστοσελίδας της KEGG (http://rest.kegg.jp) και στην αποθήκευσή τους σε αρχεία κειμένου που μπορούσαν να χρησιμοποιηθούν ως δεδομένα αναφοράς από το MinPath.

2.1.3 Σχεδιασμός και υλοποίηση βάσης δεδομένων

Για την εύκολη αποθήκευση και διαχείριση των δεδομένων αναλύσεων, αναπτύχθηκε ένας αλγόριθμος ως σενάριο σε γλώσσα προγραμματισμού Python (knowledgebase_parser.py) και μία εφαρμογή σε περιβάλλον Web2py (anastasia_knowledgebase), που σε συνδυασμό μεταξύ τους αξιοποιούσαν τον εγκατεστημένο διακομιστή MySQL. Ο Python αλγόριθμος χρησιμοποιούσε τα αρχεία εξαγωγής (dump files) από βάσεις δεδομένων MySQL, που προέκυπταν από τα εργαλεία διαχείρισης αποτελεσμάτων (π.χ. sequence_parser.py). Με την εφαρμογή του συγκεκριμένου σεναρίου γινόταν επανεισαγωγή των δεδομένων σε καινούρια βάση MySQL και σύνδεσή τους με τη Web2py εφαρμογή, ώστε να είναι προσβάσιμα μέσω ενός διαδικτυακού γραφικού περιβάλλοντος χρήστη. Ο τρόπος με τον οποίον το σενάριο επιτύγχανε αυτή τη σύνδεση, ήταν μέσω της αυτοματοποιημένης επεξεργασίας, ή εκ νέου κατασκευής των απαραίτητων αρχείων παραμετροποίησης

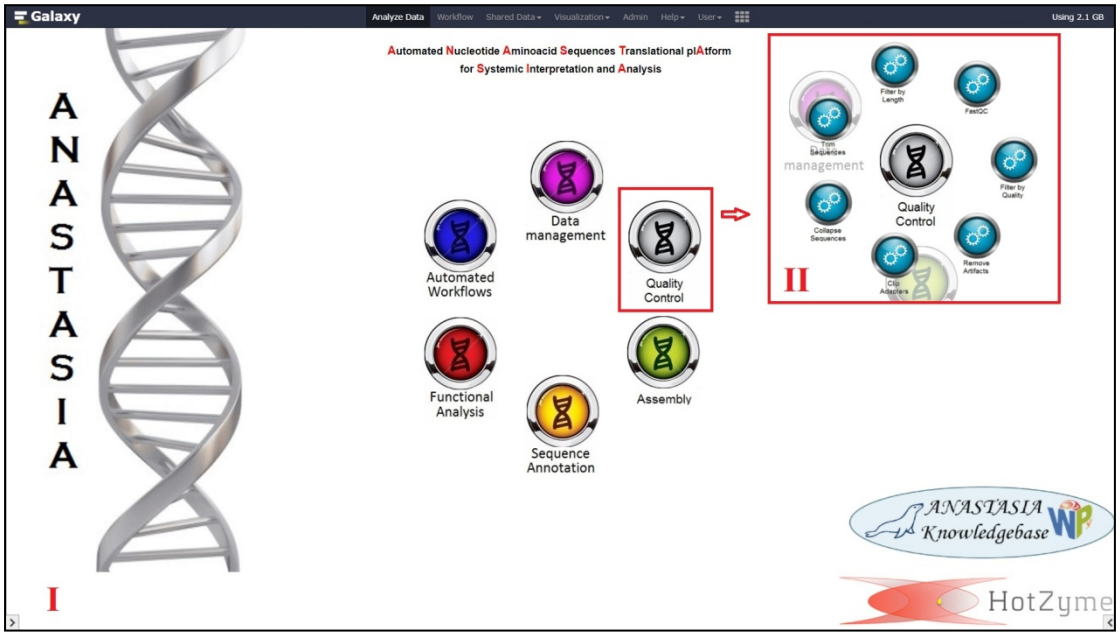
σε γλώσσα Python και HTML της ίδιας της εφαρμογής. Το αποτέλεσμα της παραπάνω σύνδεσης ήταν η ενσωμάτωση των εξελιγμένων δυνατοτήτων αναζήτησης δεδομένων του διακομιστή MySQL στο γραφικό περιβάλλον που προσφέρει η Web2py εφαρμογή. Παράλληλα, η πρόσβαση στα δεδομένα γινόταν με τμηματικό τρόπο, επιτρέποντας την παρουσίαση συγκεκριμένου αριθμού καταχωρήσεων τη φορά, έτσι ώστε να μην προκαλούνται προβλήματα απόδοσης κατά την οπτικοποίηση μεγάλου όγκου αρχείων αποτελεσμάτων.

Παρ' όλη την ευκολία με την οποία γίνονταν οι λειτουργίες αποθήκευσης και επεξεργασίας των δεδομένων, θεωρήθηκε αναγκαίο να καθιερωθεί και ένα αξιόπιστο σύστημα ασφαλείας για την πρόσβαση σε αυτά κατά τη διάρκεια των αναλύσεων. Αυτό εξασφαλίστηκε με δύο διαφορετικούς τρόπους. Πρώτον μέσω του περιβάλλοντος Web2py το οποίο προσφέρει τη δυνατότητα ανάπτυξης ξεχωριστού συστήματος ασφαλείας για κάθε εφαρμογή που αναπτύσσεται σε αυτό. Η δυνατότητα αυτή μπορούσε να ενεργοποιηθεί με τις κατάλληλες εντολές στα αρχεία παραμετροποίησης της εφαρμογής, η οποία πλέον ήταν προσβάσιμη μόνο μέσω μίας αρχικής σελίδας ταυτοποίησης. Έτσι ο κάθε χρήστης πρέπει να έχει τα κατάλληλα διαπιστευτήρια (όνομα χρήστη/κωδικός) ώστε να έχει πρόσβαση στο γραφικό περιβάλλον της εφαρμογής, τα οποία μπορεί να προμηθευτεί μόνο από τον διαχειριστή του συστήματος. Δεύτερον, σε κάθε σετ δεδομένων που εισαγόταν στην εφαρμογή/βάση δεδομένων, αντιστοιχιζόταν ένας μοναδικός αλφαριθμητικός κωδικός ταυτοποίησης. Ο κωδικός ταυτοποίησης εξαγόταν μέσω του σεναρίου Python και μόνο με την εισαγωγή του στη Web2py εφαρμογή μπορούσαν να εμφανιστούν τα αντίστοιχα δεδομένα. Έτσι διασφαλιζόταν ότι ακόμα και με τα σωστά διαπιστευτήρια εισόδου στην εφαρμογή, ο κάθε χρήστης θα είχε πρόσβαση μόνο στα σετ δεδομένων τα οποία είχε εισάγει ο ίδιος και συνεπώς είχε προμηθευτεί τον αντίστοιχο κωδικό ταυτοποίησης. Λόγω του τρόπου σχεδιασμού της εφαρμογής, ο κωδικός ταυτοποίησης αποτελούσε και μέρος του διαδικτυακού υπερσυνδέσμου (hyperlink) στον οποίο εμφανιζόταν το συγκεκριμένο σετ δεδομένων. Έτσι, ο κάθε χρήστης είχε δύο επιλογές για την πρόσβαση στα δεδομένα: είτε με την είσοδο στην εφαρμογή και την εισαγωγή του κωδικού ταυτοποίησης μέσω της αντίστοιχης φόρμας αναζήτησης, ή άμεσα με τη χρήση του υπερσυνδέσμου που είχε προμηθευτεί από το εργαλείο Python.

2.1.4 Σχεδιασμός πλατφόρμας αυτοματοποίησης

Τα προγράμματα που επιλέχθηκαν, μαζί με τους νέους αλγορίθμους που αναπτύχθηκαν, αποτέλεσαν τμήματα αυτοματοποιημένων γραμμών υπολογιστικών εργασιών, οι οποίες κάλυπταν πλήρως τις ανάγκες ανάλυσης μεταγονιδωματικών δεδομένων, ξεκινώντας είτε από το επίπεδο των δεδομένων αλληλούχισης είτε από μετέπειτα επίπεδα επεξεργασίας. Ο σχεδιασμός των συγκεκριμένων γραμμών υπολογιστικών εργασιών και η ενσωμάτωση των απαραίτητων εργαλείων σε αυτές έγινε στα πλαίσια ανάπτυξης μίας διαδικτυακής εφαρμογής για αναλύσεις μεταγενωμικών δεδομένων, η οποία ονομάστηκε ANASTASIA (Automated Nucleotide Aminoacid Sequences Translational plAtform for Systemic Interpretation and Analysis). Το ANASTASIA (Εικόνα 20) περιελάμβανε όλα τα ενσωματωμένα σε αυτό εργαλεία και αλγορίθμους που ήταν απαραίτητα για μεταγενωμικές αναλύσεις και καθιστούσε εφικτή την πρόσβαση σε αυτά μέσω ενός φιλικού προς το χρήστη γραφικού περιβάλλοντος εργασίας. Ο σχεδιασμός του εν λόγω γραφικού περιβάλλοντος βασίστηκε πάνω στην διαδικτυακή πλατφόρμα ανοικτού λογισμικού Galaxy και διαμορφώθηκε αξιοποιώντας έναν αλγόριθμο σε μορφή σεναρίου (jQuery.radmenu.js) σε γλώσσα Java, σε συνδυασμό με εξειδικευμένες αλλαγές στο αρχείο HTML (welcome.html) που χρησιμοποιούταν από το Galaxy για να ορίσει την αρχική σελίδα του. Επίσης, έγιναν παρεμβάσεις στον πηγαίο κώδικα του ίδιου του Galaxy και συγκεκριμένα στο σενάριο Java που ήταν υπεύθυνο για τη λειτουργία της αρχικής σελίδας (galaxy/client/galaxy/scripts/apps/analysis.js). Αυτό είχε ως αποτέλεσμα να απλοποιηθεί αρκετά το περιβάλλον εργασίας, τόσο λόγω της νέας εμφάνισης που περιελάμβανε διαδραστικά κουμπιά συνεδεμένα με κάθε εργαλείο, όσο και χάρη στην κατηγοριοποίηση των εργαλείων ανάλογα με τη λειτουργία τους. Η διαθεσιμότητα του ANASTASIA μέσω διαδικτύου εξασφαλίστηκε με την εγκατάσταση και χρήση του διαδικτυακού διακομιστή (web server) Apache, στον οποίον έγιναν οι κατάλληλες τροποποιήσεις ώστε να παρέχει πρόσβαση στη θύρα από την οποία λειτουργούσε η πλατφόρμα. Αντίστοιχες ρυθμίσεις, ώστε να προσδιοριστεί η θύρα λειτουργίας, έγιναν και στο αρχείο παραμετροποίησης του Galaxy (galaxy.ini), μαζί με επιπλέον αλλαγές (Παράρτημα Ι) για την περαιτέρω προσαρμογή της πλατφόρμας στις ανάγκες της υπολογιστικής υποδομής. Οι κυριότερες αλλαγές που έγιναν είχαν ως στόχο να καθοριστεί η διαδικτυακή διεύθυνση πίσω από την οποία θα φαίνεται η πλατφόρμα, να οριστεί ο χρήστης-διαχειριστής (administrator) της και να επιτραπεί η άμεση πρόσβασή της στις θέσεις των αρχείων δεδομένων που

υπήρχαν στο διακομιστή όπου ήταν εγκατεστημένη. Η τελευταία ρύθμιση ήταν εξέχουσας σημασίας για την ανάλυση δεδομένων που καταλάμβαναν τεράστιο όγκο (όπως π.χ. από αλληλούχιση μεταγονιδιώματος), καθώς η αποθήκευσή τους στον διακομιστή μέσω πρωτοκόλλου μεταφοράς αρχείων (File Transfer Protocol - FTP) εξασφάλιζε την άμεση πρόσβαση της πλατφόρμας σε αυτά, δίχως την ανάγκη περαιτέρω αντιγραφής ή μεταφόρτωσής (upload) τους. Τα αρχεία που εισάγονταν στην πλατφόρμα, είτε μέσω μεταφόρτωσης είτε μέσω άμεσης σύνδεσης με τη θέση τους στον διακομιστή, μπορούσαν να οργανωθούν σε βιβλιοθήκες δεδομένων (data libraries) για καθεμία από τις οποίες να παρέχεται ξεχωριστό σύστημα ασφαλείας, ώστε η πρόσβαση να είναι δυνατή μόνο από χρήστες οι οποίοι έχουν κάνει εγγραφή στην πλατφόρμα και έχουν τα κατάλληλα διαπιστευτήρια (όνομα/κωδικός χρήστη). Μέσα από τις βιβλιοθήκες αυτές, τα δεδομένα μπορούσαν να χρησιμοποιηθούν επανειλημμένως από τα κατάλληλα ενσωματωμένα εργαλεία για τις αντίστοιχες αναλύσεις. Σημαντική αλλαγή επίσης, που έγινε μέσω της επεξεργασίας του αρχείου παραμετροποίησης, ήταν η σύνδεση της πλατφόρμας με το διακομιστή MySQL που είχε εγκατασταθεί, για τη βέλτιστη διαχείριση των δεδομένων πρόσβασης, ταυτοποίησης και ιστορικού των χρηστών σε ξεχωριστή βάση δεδομένων. Συνολικά σχεδιάστηκαν και εγκαταστάθηκαν δύο εκδόσεις της πλατφόρμας.



Εικόνα 20. Γραφικό περιβάλλον της πλατφόρμας ANASTASIA. I)Το γραφικό περιβάλλον της διαδικτυακής πλατφόρμας περιλαμβάνει διαδραστικά κουμπιά για κάθε κύριο τομέα της ανάλυσης μεταγενωμικών δεδομένων όπως π.χ. ποιοτική ανάλυση (κουμπί γκρι χρώματος - Quality Control). II)Με την επιλογή τομέα ανάλυσης εμφανίζονται τα διαθέσιμα εργαλεία (κουμπιά μπλε χρώματος) που μπορεί να χρησιμοποιήσει ο χρήστης.

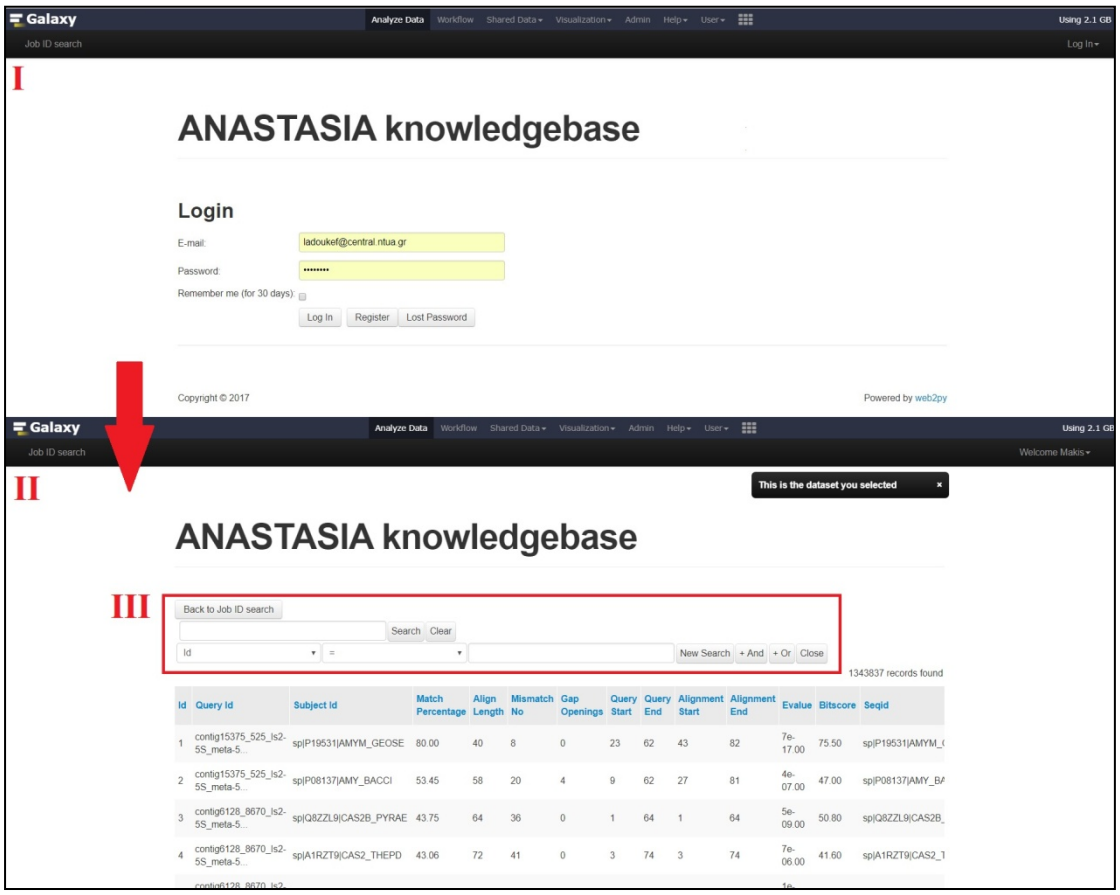
Η δοκιμαστική έκδοση του ANASTASIA εγκαταστάθηκε στο διακομιστή Helios του πανεπιστημίου της Κοπεγχάγης (http://galaxy.hotzyme.binf.ku.dk), ενώ η τελική έκδοση εγκαταστάθηκε και έγινε διαθέσιμη μέσω του διακομιστή motherbox του ΕΜΠ (motherbox.chemeng.ntua.gr/anastasia_dev/). Η ενσωμάτωση των εργαλείων σε όλες τις εκδόσεις του ANASTASIA έγινε μέσω αλλαγών στο αρχείο παραμέτρων του Galaxy «tool_conf.xml» και με την ανάπτυξη νέων αλγορίθμων ως σενάρια «επικάλυψης» (wrapper scripts) σε γλώσσα προγραμματισμού XML. Οι αλγόριθμοι αυτοί επέτρεπαν τη σύνδεση της πλατφόρμας με τα εκτελέσιμα σενάρια συντακτικής ανάλυσης για κάθε εγκατεστημένο εργαλείο, ενώ ταυτόχρονα προσέφεραν ένα φιλικό προς το χρήστη γραφικό περιβάλλον. Το γραφικό περιβάλλον κάθε εργαλείου ήταν ορατό μέσω της κεντρικής σελίδας του ANASTASIA προσφέροντας στο χρήστη τη δυνατότητα επιλογής των αρχείων προς ανάλυση καθώς και των τιμών των απαραίτητων παραμέτρων του εκάστοτε προγράμματος. Συνολικά αναπτύχθηκαν 15 αλγόριθμοι XML (Παράρτημα ΙΙ) για τα εργαλεία που παρουσιάζονται παρακάτω στον Πίνακα 1:

Αλγόριθμος επικάλυψης XML	Εκτελέσιμο σενάριο συντακτικής ανάλυσης	Εργαλείο που καλείται
rma_builder.xml	blastsub.py	MEGAN
megan_analysis.xml	megansub.py	MEGAN
megan_analysisf.xml	functionalsub.py	MEGAN
blast_parser.xml	blast_parser.py	MySQL
hmmer_parser.xml	hmmer_parser.py	MySQL
count_mapped.xml	count_mapped.py	Samtools
eficaz.xml	eficaz.py	EFICAZ
fasta_names.xml	fasta_names.py	-
getorf.xml	-	GETORF
megahit.xml	megahit.py	Megahit
sequence_parser.xml	sequence_parser.py	MySQL
knowledgebase_parser.xml	knowledgebase_parser.py	Web2py / MySQL
minpath.xml	-	Minpath
transeq.xml	-	Emboss/Transeq
fasta_joiner2.xml	shuffleSequences_fasta.pl ή shuffleSequences_fastq.pl	-

Πίνακας 1. Λίστα εξειδικευμένων αλγορίθμων επικάλυψης XML που αναπτύχθηκαν για την ενσωμάτωση εργαλείων στο ANASTASIA. Ο κάθε αλγόριθμος σε γλώσσα XML συνδέει το ANASTASIA με το αντίστοιχο εκτελέσιμο σενάριο συντακτικής ανάλυσης παραθέτοντας ένα γραφικό περιβάλλον χρήστη. Το κάθε σενάριο συντακτικής ανάλυσης καλεί με τη σειρά του το αντίστοιχο εγκατεστημένο εργαλείο ώστε να γίνει η ανάλυση και να εμφανιστούν τα δεδομένα αποτελεσμάτων στο περιβάλλον εργασίας του ANASTASIA. Για ορισμένους αλγορίθμους XML δεν υπάρχει εκτελέσιμο σενάριο συντακτικής ανάλυσης, καθώς καλούν τα ίδια με άμεσο τρόπο το εργαλείο με τις απαραίτητες παραμέτρους. Αντίστοιχα κάποια εκτελέσιμα σενάρια δεν καλούν κάποιο επιπλέον εργαλείο καθώς η ανάλυση γίνεται άμεσα από τα ίδια.

Λόγω της εκτεταμένης χρήσης της πλατφόρμας Galaxy για βιοπληροφορικές αναλύσεις από πολλές ερευνητικές ομάδες, έχουν αναπτυχθεί και δημοσιευτεί αλγόριθμοι επικάλυψης για πολλά από τα πιο διάσημα εργαλεία όπως το BLAST και το HMMER. Η προσθήκη αυτών των αλγορίθμων στη λίστα των ήδη ανεπτυγμένων για το ANASTASIA, επέτρεψε την ενσωμάτωση όλων των απαραίτητων εργαλείων στην πλατφόρμα. Στην περίπτωση του εξειδικευμένου εργαλείου διαχείρισης δεδομένων μέσω Web2py εφαρμογής, η ενσωμάτωσή του περιελάμβανε εκτός από τη χρήση του αντίστοιχου αλγορίθμου επικάλυψης και την ένωση των δύο διαδικτυακών γραφικών περιβαλλόντων. Η ένωση αυτή επιτεύχθηκε καθώς οι υπερσύνδεσμοι που παράγονταν από το εργαλείο, για την παρουσίαση των αντίστοιχων σετ δεδομένων,

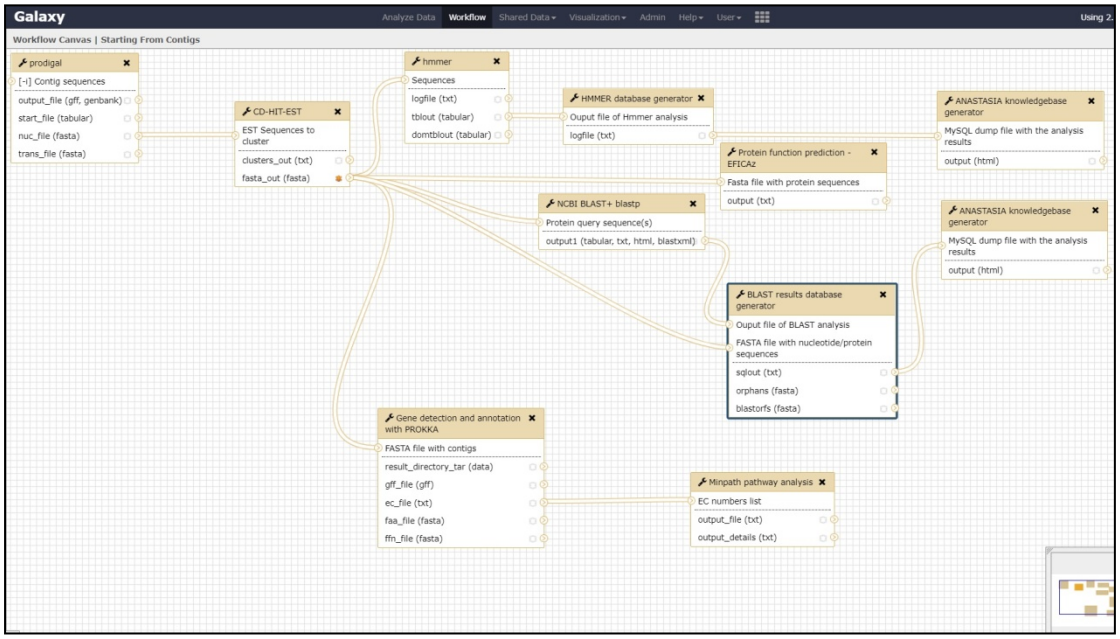
άνοιγαν μέσα από το περιβάλλον του ίδιου ANASTASIA (Εικόνα 21), διατηρώντας έτσι την πρόσβαση στις υπόλοιπες λειτουργίες του.



Εικόνα 21. Γραφικό περιβάλλον της web2py εφαρμογής αποθήκευσης δεδομένων (ANASTASIA knowledgebase). I)Το γραφικό περιβάλλον της εφαρμογής εμφανίζεται μέσα από το περιβάλλον της πλατφόρμας ANASTASIA και για την είσοδο απαιτείται όνομα και κωδικός χρήστη. II)Μετά την είσοδο χρήστη τα δεδομένα εμφανίζονται σε μορφή πλέγματος (grid). III)Η εφαρμογή περιλαμβάνει εξειδικευμένα διαδραστικά κουμπιά για την αναζήτηση και το φιλτράρισμα των δεδομένων αξιοποιώντας τις δυνατότητες της MySQL υποδομής.

Η αυτοματοποίηση των αναλύσεων έγινε κατασκευάζοντας υπολογιστικές γραμμές εργασιών μέσα στο περιβάλλον του Galaxy (Εικόνα 22) και ενσωματώνοντας τις στο ANASTASIA. Οι συγκεκριμένες γραμμές εργασιών αποτελούνταν από εργαλεία και αλγορίθμους που λειτουργούσαν είτε σε διαδοχική σειρά, είτε παράλληλα αλλά με προκαθορισμένο τρόπο, έτσι ώστε τα δεδομένα που προέκυπταν από την ανάλυση του ενός να μεταφέρονταν αυτόματα για την επόμενη ανάλυση, που μπορούσε να περιλαμβάνει ένα ή περισσότερα εργαλεία. Κατ' αυτόν τον τρόπο τα δεδομένα αλληλούχισης π.χ. μπορούσαν να τροφοδοτηθούν στο εργαλείο συναρμολόγησης και την ανάλυση αυτή να ακολουθήσει η ανάλυση εύρεσης γονιδίων στα συναρμολογήματα, χωρίς να απαιτείται κάποια ενδιάμεση

παρέμβαση του χρήστη. Συνολικά αναπτύχθηκαν τέσσερις διαφορετικές αυτοματοποιημένες γραμμές εργασιών με εξειδίκευση σε μεταγενωμικά δεδομένα, με την πρώτη να αποτελεί τη λύση για μία ολοκληρωμένη ανάλυση ξεκινώντας από τα δεδομένα αλληλούχισης, ενώ οι υπόλοιπες εστιάζουν σε κάποιο μικρότερο κομμάτι της ανάλυσης.



Εικόνα 22. Σχεδιασμός υπολογιστικής γραμμής εργασιών μέσα από το περιβάλλον του Galaxy. Κάθε πλαίσιο αντιστοιχεί σε ένα εργαλείο με τα σημεία εισόδου δεδομένων να φαίνονται αριστερά του και τα σημεία εξόδου δεδομένων να φαίνονται δεξιά. Οι γραμμές που ενώνουν διαφορετικά πλαίσια μεταξύ τους δηλώνουν τη μεταφορά δεδομένων ανάμεσά τους. Π.χ. από το εργαλείο Prodigal (πάνω αριστερά πλαίσιο) παράγονται 4 αρχεία (output_file, start_file, nuc_file και trans_file) ένα εκ των οποίων (nuc_file) μεταφέρεται αυτόματα στο εργαλείο CD-HIT-EST για την επόμενη ανάλυση. Η σχεδιασμός και ενσωμάτωση στο ANASTASIA μίας γραμμής εργασιών εξασφαλίζει την αυτόματη εκτέλεση όλων των επιμέρων εργαλείων της.

Η πρώτη γραμμή εργασίας ονομάζεται «Starting From Reads» (Πίνακας 2) και εφαρμόζεται σε δεδομένα αλληλούχισης μορφής FASTQ ώστε να καταλήξει σε μία πλήρως ολοκληρωμένη ανάλυση του μεταγονιδιώματος. Τα διαδοχικά στάδια ανάλυσης από τα οποία αποτελείται είναι τα εξής:

ΣΤΑΔΙΟ ΑΝΑΛΥΣΗΣ	ΕΡΓΑΛΕΙΟ	ΔΕΔΟΜΕΝΑ ΕΙΣΟΔΟΥ	ΔΕΔΟΜΕΝΑ ΕΞΟΔΟΥ
Ποιοτικός έλεγχος	FASTX	Δεδομένα αλληλούχισης μορφής FASTQ	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTQ
Μετατροπή δεδομένων αλληλούχισης	FASTX	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTQ	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTA
Επεξεργασία αρχείων FASTA	fasta_names.py	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTA	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTA με διορθωμένους αναγνωριστικούς κωδικούς
Ανάλυση ομολογίας αναγνωσμένων αλληλουχιών	BLAST	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTA με διορθωμένους αναγνωριστικούς κωδικούς	Δεδομένα σύγκρισης ομολογίας αναγνωσμένων αλληλουχιών σε μορφή πίνακα (tabular)
Ταξονομικός και λειτουργικός προσδιορισμός	MEGAN	Δεδομένα σύγκρισης ομολογίας αναγνωσμένων αλληλουχιών σε μορφή πίνακα (tabular) και δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTA	Δεδομένα ταξονομικού και λειτουργικού προσδιορισμού σε μορφή κειμένου (.txt), εικόνων (.jpg) και αρχείων επεξεργάσιμων από το MEGAN (.rma)
Συναρμολόγηση	Megahit	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTA με διορθωμένους αναγνωριστικούς κωδικούς	Δεδομένα συναρμολογημάτων μορφής FASTA
Εντοπισμός γονιδίων	Prodigal	Δεδομένα συναρμολογημάτων μορφής FASTA	Δεδομένα πιθανών γονιδίων μορφής FASTA
Ομαδοποίηση αλληλουχιών	CD-HIT	Δεδομένα πιθανών γονιδίων μορφής FASTA	Ομαδοποιημένα δεδομένα πιθανών γονιδίων μορφής FASTA
Χαρακτηρισμός γονιδίων μέσω σύγκρισης ομοιότητας	BLAST και HMMER	Ομαδοποιημένα δεδομένα πιθανών γονιδίων μορφής FASTA	Δεδομένα γνωστών γονιδίων και λειτουργικών περιοχών υψηλής ομοιότητας σε μορφή πίνακα (tabular)
Χαρακτηρισμός γονιδίων μέσω αλγορίθμων μηχανικής μάθησης	EFICAZ και PROKKA	Ομαδοποιημένα δεδομένα πιθανών γονιδίων μορφής FASTA	Δεδομένα πιθανών λειτουργιών αντιστοιχισμένων με αλληλουχίες
Ανακατασκευή μεταβολικών μονοπατιών	Minpath	Δεδομένα πιθανών λειτουργιών αντιστοιχισμένων με αλληλουχίες	Λίστα μεταβολικών μονοπατιών που ανακατεσκευάστηκαν
Αποθήκευση δεδομένων σε βάση MySQL	hmmer_parser.py και blast_parser.py	Δεδομένα γνωστών γονιδίων και λειτουργικών περιοχών υψηλής ομοιότητας σε μορφή πίνακα (tabular)	Δεδομένα εξαγωγής βάσης δεδομένων μορφής .sql
Οπτικοποίηση βάσης δεδομένων	knowledgebase_parser.py	Δεδομένα εξαγωγής βάσης δεδομένων μορφής .sql	Αρχείο καταγραφής αναγνωριστικών κωδικών και υπερσυνδέσμων

Πίνακας 2. Στάδια ανάλυσης της γραμμής εργασιών «Starting From Reads» του ANASTASIA. Κάθε στάδιο ανάλυσης χρησιμοποιεί ένα η περισσότερα εργαλεία για την ανάλυση των δεδομένων εισόδου και την εξαγωγή αποτελεσμάτων (δεδομένα εξόδου) που θα αξιοποιηθούν από το επόμενο στάδιο.

Η δεύτερη γραμμή εργασίας ονομάζεται «Starting From Contigs» (Πίνακας 3) και εφαρμόζεται σε ήδη συναρμολογημένα δεδομένα αλληλούχισης, εστιάζοντας στον εντοπισμό και χαρακτηρισμό νέων γονιδίων και λειτουργιών:

ΣΤΑΔΙΟ ΑΝΑΛΥΣΗΣ	ΕΡΓΑΛΕΙΟ	ΔΕΔΟΜΕΝΑ ΕΙΣΟΔΟΥ	ΔΕΔΟΜΕΝΑ ΕΞΟΔΟΥ
Εντοπισμός γονιδίων	Prodigal	Δεδομένα συναρμολογημάτων μορφής FASTA	Δεδομένα πιθανών γονιδίων μορφής FASTA
Ομαδοποίηση αλληλουχιών	CD-HIT	Δεδομένα πιθανών γονιδίων μορφής FASTA	Ομαδοποιημένα δεδομένα πιθανών γονιδίων μορφής FASTA
Χαρακτηρισμός γονιδίων μέσω σύγκρισης ομοιότητας	BLAST και HMMER	Ομαδοποιημένα δεδομένα πιθανών γονιδίων μορφής FASTA	Δεδομένα γνωστών γονιδίων και λειτουργικών περιοχών υψηλής ομοιότητας σε μορφή πίνακα (tabular)
Χαρακτηρισμός γονιδίων μέσω αλγορίθμων μηχανικής μάθησης	EFICAZ και PROKKA	Ομαδοποιημένα δεδομένα πιθανών γονιδίων μορφής FASTA	Δεδομένα πιθανών λειτουργιών αντιστοιχισμένων με αλληλουχίες
Ανακατασκευή μεταβολικών μονοπατιών	Minpath	Δεδομένα πιθανών λειτουργιών αντιστοιχισμένων με αλληλουχίες	Λίστα μεταβολικών μονοπατιών που ανακατεσκευάστηκαν
Αποθήκευση δεδομένων σε βάση MySQL	hmmer_parser.py και blast_parser.py	Δεδομένα γνωστών γονιδίων και λειτουργικών περιοχών υψηλής ομοιότητας σε μορφή πίνακα (tabular)	Δεδομένα εξαγωγής βάσης δεδομένων μορφής .sql
Οπτικοποίηση βάσης δεδομένων	knowledgebase_parser.py	Δεδομένα εξαγωγής βάσης δεδομένων μορφής .sql	Αρχείο καταγραφής αναγνωριστικών κωδικών και υπερσυνδέσμων

Πίνακας 3. Στάδια ανάλυσης της γραμμής εργασιών «Starting From Contigs» του ANASTASIA. Κάθε στάδιο ανάλυσης χρησιμοποιεί ένα η περισσότερα εργαλεία για την ανάλυση των δεδομένων εισόδου και την εξαγωγή αποτελεσμάτων (δεδομένα εξόδου) που θα αξιοποιηθούν από το επόμενο στάδιο.

Για την περίπτωση που μας ενδιαφέρει μόνο ο ταξονομικός και λειτουργικός προσδιορισμός, αναπτύχθηκε η γραμμή εργασίας με το όνομα «Taxonomic & functional analysis» (Πίνακας 4) της οποίας η ανάλυση ξεκινάει πάλι από τα δεδομένα αλληλούχισης:

ΣΤΑΔΙΟ ΑΝΑΛΥΣΗΣ	ΕΡΓΑΛΕΙΟ	ΔΕΔΟΜΕΝΑ ΕΙΣΟΔΟΥ	ΔΕΔΟΜΕΝΑ ΕΞΟΔΟΥ
Ποιοτικός έλεγχος	FASTX	Δεδομένα αλληλούχισης μορφής FASTQ	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTQ
Μετατροπή δεδομένων αλληλούχισης	FASTX	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTQ	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTA
Επεξεργασία αρχείων FASTA	fasta_names.py	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTA	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTA με διορθωμένους αναγνωριστικούς κωδικούς
Ανάλυση ομολογίας αναγνωσμένων αλληλουχιών	BLAST	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTA με διορθωμένους αναγνωριστικούς κωδικούς	Δεδομένα σύγκρισης ομολογίας αναγνωσμένων αλληλουχιών σε μορφή πίνακα (tabular)
Ταξονομικός και λειτουργικός προσδιορισμός	MEGAN	Δεδομένα σύγκρισης ομολογίας αναγνωσμένων αλληλουχιών σε μορφή πίνακα (tabular) και δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTA	Δεδομένα ταξονομικού και λειτουργικού προσδιορισμού σε μορφή κειμένου (.txt), εικόνων (.jpg) και αρχείων επεξεργάσιμων από το MEGAN (.rma)
Συναρμολόγηση	Megahit	Δεδομένα αλληλούχισης υψηλής ποιότητας μορφής FASTA με διορθωμένους αναγνωριστικούς κωδικούς	Δεδομένα συναρμολογημάτων μορφής FASTA
Χαρακτηρισμός γονιδίων μέσω αλγορίθμων μηχανικής μάθησης	PROKKA	Δεδομένα πιθανών γονιδίων μορφής FASTA	Δεδομένα πιθανών λειτουργιών αντιστοιχισμένων με αλληλουχίες
Ανακατασκευή μεταβολικών μονοπατιών	Minpath	Δεδομένα πιθανών λειτουργιών αντιστοιχισμένων με αλληλουχίες	Λίστα μεταβολικών μονοπατιών που ανακατεσκευάστηκαν

Πίνακας 4.. Στάδια ανάλυσης της γραμμής εργασιών «Taxonomic & functional analysis» του ANASTASIA. Κάθε στάδιο ανάλυσης χρησιμοποιεί ένα η περισσότερα εργαλεία για την ανάλυση των δεδομένων εισόδου και την εξαγωγή αποτελεσμάτων (δεδομένα εξόδου) που θα αξιοποιηθούν από το επόμενο στάδιο.

2.2 Εφαρμογή της πλατφόρμας σε μεταγενωμικά δεδομένα

2.2.1 Μεταγονιδιωματικά δείγματα

Η ανάπτυξη του ANASTASIA, καθώς και ο έλεγχος σωστής λειτουργίας του δεν θα ήταν δυνατόν να γίνουν χωρίς τη συνεχή αλληλεπίδραση με πραγματικά δεδομένα μεταγενωμικών δειγμάτων. Η απόκτηση τέτοιων δεδομένων έγινε μέσω του ευρωπαϊκού προγράμματος HotZyme [56], στα πλαίσια του οποίου κατασκευάστηκε η πλατφόρμα. Το πρόγραμμα HotZyme ήταν μία πολυεθνική συνεργασία από δέκα ερευνητικές ομάδες πανεπιστήμιων και ερευνητικών κέντρων από Ευρώπη, Ασία, Αμερική και Μέση Ανατολή μαζί με τρεις εταιρείες παραγωγής ενζύμων (Microdish, Novozymes και Sigma Aldrich) και είχε ως στόχο την απομόνωση νέων υδρολασών υψηλής θερμοσταθερότητας, μέσω εντοπισμού των αντίστοιχων γονιδιακών αλληλουχιών από μεταγονιδιώματα. Για την εύρεση θερμοσταθερών ενζύμων, επιλέχθηκαν μεταγενωμικά δείγματα από θερμές πηγές σε όλο τον κόσμο και η δειγματοληψία έγινε είτε μέσω άμεσης απομόνωσης (Πίνακας 5) ή μέσω in situ εμπλουτισμού αποικιών, χρησιμοποιώντας υποστρώματα υψηλού ενδιαφέροντος (Πίνακας 6).

Όνομα δείγματος	Τοποθεσία δειγματοληψίας	T, pH
CH1102	Θερμές πηγές στο εθνικό πάρκο Yellowstone	79,3^{o}C / pH1,8
It-3	Θερμές πηγές στο Pisciarelli, Ιταλία	85,0^{o}C / pH3,5
It-6	Θερμές πηγές στο Pozzuolli, Ιταλία	76,0^{o}C / pH3,0-3,5
MW-2	Θερμές πηγές στο εθνικό πάρκο Yellowstone	84,0^{o}C / pH8,0
NL-100808	Εθνικό πάρκο Yellowstone, ΗΠΑ	89-92^{o}C / pH3,0-5,5
NL-100908	Εθνικό πάρκο Yellowstone, ΗΠΑ	89-92^{o}C / pH3,0-5,5
Sunspring	Θερμές πηγές στην Ρωσσία	61,0-64,0^{o}C / pH5,8-6,0
Tomsk-sample	Θερμές πηγές στο Parabel, στην περιοχή Tomsk, Ρωσσία	46,0^{o}C / pH7,3

Πίνακας 5. Περιβαλλοντικά δείγματα που λήφθηκαν για μεταγενωμική ανάλυση κατά το πρόγραμμα HotZyme. Στον πίνακα σημειώνεται το όνομα δείγματος, η τοποθεσία στην οποία έγινε η δειγματοληψία καθώς και οι τιμές θερμοκρασίες και pH (όπου αυτές είναι διαθέσιμες) τη στιγμή της δειγματοληψίας.

Όνομα δείγματος	Τοποθεσία δειγματοληψίας	T/pH	Υπόστρωμα
A5-7T	Θερμές πηγές στην Ισλανδία	55°C / pH7,0	Ξυλάνη
A6-2319x	Θερμές πηγές στην Ρωσσία	85°C / pH6,0	Ξυλάνη
A7-Loc1F3	Θερμές πηγές στην Ισλανδία	78°C / pH6,0	Πολυβινυλική αλκοόλη - PVA
B1-7Tnr1	Θερμές πηγές στην Ισλανδία	55°C / pH7,0	Κόμμι ξανθάνης
B3-6Tnr4	Θερμές πηγές στην Ισλανδία	85°C / pH7,0	Κόμμι ξανθάνης
B4-12Tnr1	Θερμές πηγές στην Ισλανδία	85°C / pH7,0	Κόμμι ξανθάνης
DG#14,1xg	Θερμές πηγές στην Δανία	55°C / pH7,0	Κόμμι ξανθάνης
Planctomycetes	Θερμές πηγές στην Ρωσσία	60°C / pH6,0	Κόμμι ξανθάνης και ξυλάνη

Πίνακας 6. Περιβαλλοντικά δείγματα που λήφθηκαν με τη μέθοδο εμπλουτισμού αποικιών χρησιμοποιώντας συγκεκριμένα υποστρώματα κατά το πρόγραμμα HotZyme. Στον πίνακα σημειώνεται το όνομα δείγματος, η τοποθεσία στην οποία έγινε η δειγματοληψία καθώς και οι τιμές θερμοκρασίες και pH τη στιγμή της δειγματοληψίας.

Κατ' αυτό τον τρόπο συλλέχθηκαν συνολικά 16 περιβαλλοντικά δείγματα των οποίων τα μεταγονιδιώματα αναλύθηκαν με τεχνολογίες αλληλούχισης νέας γενιάς.

Εκτός από το πρόγραμμα HotZyme, αποκτήθηκαν δεδομένα και από το πρόγραμμα COVERALL [138], μία συνεργασία τεσσάρων ερευνητικών ομάδων από ευρωπαϊκά πανεπιστήμια, με σκοπό την μελέτη μικροβιακών πληθυσμών σε υποθαλάσσια συστήματα που έχουν εκτεθεί σε υψηλές ποσότητες CO_2 και την εύρεση ενδεικτικών αλληλουχιών (βιοδεικτών) που μπορούν να συσχετιστούν με αυτή την έκθεση. Στα πλαίσια του προγράμματος λήφθηκαν δείγματα υποθαλάσσιων ιζημάτων από τον κόλπο Ardmucknish κοντά στο Oban της Σκωτίας [147] τα οποία στη συνέχεια μεταφέρθηκαν σε έξι ειδικά κατασκευασμένες διατάξεις ανακυκλοφορίας νερού (flumes), στις εγκαταστάσεις του Σκωτσέζικου Συλλόγου Θαλάσσιων Επιστημών (Scottish Association for Marine Science - SAMS) ώστε να γίνει προσομοίωση έκθεσης θαλάσσιου οικοσυστήματος σε CO2. Οι διατάξεις χωρίστηκαν σε τρία ζεύγη, καθένα από τα οποία αποτελούσε και ένα ξεχωριστό σύστημα μελέτης συνδεδεμένο με μία παροχή αέριας τροφοδοσίας. Οι παροχές αερίου στα τρία συστήματα μελέτης αποτελούνταν από 1)ατμοσφαιρικό αέρα (0% CO_2), 2)ατμοσφαιρικό αέρα και CO_2 σε αναλογία 1:1 κ.ο. (50% CO_2) και 3)καθαρό

(100%) CO_2. Η πειραματική διαδικασία κράτησε συνολικά 12 εβδομάδες με τις πρώτες 4 να αποτελούν την περίοδο ηρεμίας, τις επόμενες 4 να αποτελούν την περίοδο έκθεσης στην αέρια τροφοδοσία και τις τελευταίες 4 να αποτελούν την περίοδο «επαναφοράς» έχοντας διακόψει πλέον την τροφοδοσία. Κατά τη διάρκεια των διαφορετικών συνθηκών έκθεσης λήφθηκαν δείγματα ιζήματος από τυχαία σημεία των δύο διατάξεων παροχής 100% CO_2, από τα οποία έγινε απομόνωση μεταγενωμικού DNA και καταγραφή του μέσω αλληλούχισης νέας γενιάς. Συγκεκριμένα έγινε λήψη 6 δειγμάτων στο τέλος της πρώτης εβδομάδας έκθεσης, άλλων 4 στο τέλος της τέταρτης εβδομάδας έκθεσης και 2 επιπλέον κατά την πρώτη βδομάδα διακοπής της τροφοδοσίας.

2.2.2 Διαχείριση αρχικών δεδομένων

Τα δεδομένα αλληλούχισης από το πρόγραμμα HotZyme μεταφέρθηκαν μέσω FTP στον διακομιστή Helios του πανεπιστημίου της Κοπεγχάγης, πάνω στον οποίο είχε εγκατασταθεί η δοκιμαστική έκδοση του ANASTASIA. Το συνολικό μέγεθος των δεδομένων ήταν ~100GB και αποτελούταν από πολλαπλά αρχεία τύπου SFF ή FASTQ, είτε ένα ή δύο ανά δείγμα, ανάλογα αν το πείραμα αλληλούχισης ήταν μονού άκρου ή ζεύγους άκρων. Τα αντίστοιχα αρχεία εισήχθηκαν στο ANASTASIA μέσα σε μία βιβλιοθήκη δεδομένων με το όνομα «Hotzyme raw data», στην οποία η πρόσβαση επιτράπηκε μόνο στους συνεργάτες του προγράμματος. Τα αρχεία αυτά αποφασίστηκε να μεταφερθούν και σε ξεχωριστό διακομιστή από την ερευνητική ομάδα του πανεπιστημίου της Κοπεγχάγης η οποία ανέλαβε και την ανάλυση συναρμολόγησης λόγω των τεράστιων υπολογιστικών απαιτήσεων σε μνήμη και υπολογιστική ισχύ. Οι αναλύσεις γίνανε σε ξεχωριστό διακομιστή, αλλά τα αποτελέσματα των συναρμολογημάτων εισάχθηκαν στο διακομιστή Helios και περιλήφθηκαν σε μια νέα βιβλιοθήκη με το όνομα «Hotzyme assemblies». Μέσα από τις βιβλιοθήκες αυτές υπήρχε δυνατότητα λήψης και αποθήκευσης των δεδομένων σε τοπικό υπολογιστή, ενώ ταυτόχρονα τα ενσωματωμένα εργαλεία μπορούσαν να τα αξιοποιήσουν για περαιτέρω αναλύσεις. Αντίστοιχα τα δεδομένα αλληλούχισης από το πρόγραμμα COVERALL εισήχθησαν στην τελική έκδοση του ANASTASIA στον διακομιστή του ΕΜΠ, σε αντίστοιχη βιβλιοθήκη δεδομένων με το όνομα «COVERALL data». Το συνολικό μέγεθος των δεδομένων ήταν ~500GB και αποτελούνταν από δύο αρχεία FASTQ ανά δείγμα τα οποία περιείχαν αλληλουχίες ζεύγους άκρων.

2.2.3 Ανάλυση μεταγονιδιωματικών δεδομένων

Οι δυνατότητες αυτοματοποιημένης ανάλυσης δεν ήταν εξ' αρχής διαθέσιμες στην πλατφόρμα που σχεδιάστηκε, αλλά αναπτύχθηκαν σταδιακά κατά τη διάρκεια των αναλύσεων διαφορετικών δεδομένων. Αυτό έγινε εφικτό με την ενσωμάτωση των διαφόρων εργαλείων που χρησιμοποιήθηκαν, στις υπολογιστικές γραμμές εργασιών που αναφέρθηκαν προηγουμένως. Οι αρχικές γραμμές εργασιών που αναπτύχθηκαν και εφαρμόστηκαν διέφεραν από αυτές της τελικής έκδοσης του ANASTASIA καθώς σχεδιάστηκαν με γνώμονα τις εξειδικευμένες απαιτήσεις των εκάστοτε δεδομένων. Για την περίπτωση των δεδομένων αλληλούχισης από το ερευνητικό πρόγραμμα HotZyme αναπτύχθηκε η γραμμή εργασιών «Starting From Contigs» της οποία η αρχική μορφή είχε τις εξής τροποποιήσεις:

- Αντί του μεμονωμένου εργαλείο Prodigal, χρησιμοποιούταν ένας αλγόριθμος για την εύρεση νέων γονιδίων, σε μορφή σεναρίου σε γλώσσα BASH, που αναπτύχθηκε από την ερευνητική ομάδα του πανεπιστημίου της Κοπεγχάγης και ο οποίος καλούσε 3 διαφορετικά εργαλεία εύρεσης γονιδίων: α)το Prodigal, β)το MetaGeneMark και γ)το MetaGeneAnnotator, καθώς και εξειδικευμένους αλγόριθμους σε γλώσσα R. Το καθένα από τα τρία εργαλεία εντοπισμού γονιδίων εκτελούταν πάνω στα δεδομένα συναρμολόγησης και τα αποτελέσματά τους αποθηκεύονταν σε προσωρινούς φακέλους μέσα στο διακομιστή. Τα αποτελέσματα αυτά στη συνέχεια αξιοποιούνταν από τους αλγορίθμους σε γλώσσα R, οι οποίοι κατασκεύαζαν ένα αρχείο FASTA με τις αλληλουχίες γονιδίων οι οποίες χαρακτηρίστηκαν ως πιθανά γονίδια και από τα τρία εργαλεία και ένα αρχείο FASTA με τις αλληλουχίες γονιδίων οι οποίες χαρακτηρίστηκαν ως πιθανά γονίδια από τουλάχιστον ένα από τα τρία εργαλεία.
- Η γραμμή εργασιών περιελάμβανε ένα επιπλέον εξειδικευμένο εργαλείο διαχείρισης δεδομένων BLAST και HMMER, το οποίο είχε αναπτυχθεί σε γλώσσα Python συγκεκριμένα για τις ανάγκες του προγράμματος. Το συγκεκριμένο εργαλείο φιλτράριζε τα αποτελέσματα ομολογίας με βάση ποιες αλληλουχίες από αυτές είχαν κάποιο αποτέλεσμα στην αντίστοιχη ανάλυση με το εργαλείο HMMER. Έτσι προέκυπτε μία λίστα που αποτελούταν μόνο από τα πιθανά γονίδια για τα οποία είχε βρεθεί υψηλή ομοιότητα με κάποιο ήδη

γνωστό, ενώ ταυτόχρονα περιείχαν και κάποια συντηρημένη λειτουργική ομάδα στην αλληλουχία τους.

- Αναπτύχθηκε σε συνεργασία με το Εθνικό Ίδρυμα Ερευνών και προστέθηκε στη γραμμή εργασιών ένα επιπλέον εξειδικευμένο εργαλείο σε γλώσσα Perl, το οποίο αντιστοίχιζε σε κάθε αλληλουχία ένα πιθανό αριθμό EC με βάση το καλύτερο αποτέλεσμα ομολογίας (υψηλότερο ποσοστό ομοιότητας) της από την ανάλυση BLAST στη βάση δεδομένων UniProt/SwissProt. Για τις ανάγκες του συγκεκριμένου εργαλείου είχε κατασκευαστεί επίσης μία βάση δεδομένων MySQL, στην οποία αντιστοιχίζονταν οι αναγνωριστικοί κωδικοί αλληλουχιών της UniProt/SwissProt με τους αντίστοιχους EC αριθμούς. Κατά την εφαρμογή του σε δεδομένα ομολογίας, ο αλγόριθμος αποκτούσε πρόσβαση στη βάση δεδομένων και αξιοποιώντας τις δυνατότητες εξειδικευμένων αναζητήσεων της MySQL, απέδιδε τον αντίστοιχο αριθμό EC σε κάθε αλληλουχία.
- Δεν χρησιμοποιήθηκε η γραμμή εργασιών PROKKA καθώς κατά το χρονικό διάστημα που έλαβαν χώρα οι αναλύσεις των δεδομένων δεν υπήρχε κάποια σταθερή έκδοσή της. Η έλλειψη του εργαλείου PROKKA οδήγησε στην παράβλεψη και του εργαλείου Minpath καθώς το δεύτερο χρειαζόταν τα δεδομένα εξόδου του πρώτου για τη λειτουργία του.
- Η ανάλυση με το εργαλείο EFICAz δεν πραγματοποιήθηκε στο σύνολο των πιθανών γονιδίων από κάθε δείγμα λόγω των τεράστιων χρόνων λειτουργίας που απαιτούσε. Αντ' αυτού αξιοποιούταν κατά περίπτωση σε μικρότερα σετ δεδομένων τα οποία είχαν εκτιμηθεί από τις αναλύσεις ομολογίας ότι περιέχουν αλληλουχίες υψηλού ενδιαφέροντος.

Η τροποποιημένη πλέον βιοπληροφορική γραμμή εργασιών εκτελέστηκε με τις προκαθορισμένες τιμές παραμέτρων για κάθε εργαλείο της και για τα 16 δείγματα και η αυτοματοποίηση της εξασφάλισε την ανταλλαγή των παραγόμενων δεδομένων μεταξύ των εργαλείων ανάλυσης, χωρίς περαιτέρω επέμβαση. Οι βάσεις δεδομένων που εγκαταστάθηκαν για χρήση των εργαλείων σύγκριση με γνωστές αλληλουχίες ήταν η NCBI-nr, για χρήση με το αντιστοιχο εργαλείο BLASTp και η Pfam-A με το αντίστοιχο εργαλείο HMMER. Τα δεδομένα από κάθε ξεχωριστό δείγμα αποθηκεύονταν μέσω του ANASTASIA σε ξεχωριστό ιστορικό ανάλυσης και ήταν διαθέσιμα για λήψη από όλους τους συνεργάτες του ερευνητικού προγράμματος.

Αντίστοιχα, για την περίπτωση του ερευνητικού προγράμματος COVERALL, αναπτύχθηκε η γραμμή εργασιών «Taxonomic & functional analysis» αλλά η αρχική της μορφή εφαρμόστηκε με τις εξής τροποποιήσεις:

- Απομακρύνθηκε το εργαλείο ποιοτικού ελέγχου καθώς το βήμα αυτό είχε εκτελεστεί από την εταιρεία που παρείχε την αλληλούχιση μέσω του εργαλείου FASTQC
- Έγινε αντικατάσταση του εργαλείου ανάλυσης ομολογίας από το BLAST στο DIAMOND. Ο λόγος που έγινε αυτή η αλλαγή ήταν η διαφορά σε ταχύτητα των δύο εργαλείων για την ίδια διάθεση υπολογιστικών πόρων σε αυτά, ενώ διατηρούσαν σχεδόν το ίδιο επίπεδο αξιοπιστίας των αποτελεσμάτων ανάλογα και με τις τιμές των παραμέτρων ευαισθησίας του δεύτερου.

Επίσης τα βήματα της γραμμής εργασιών χωρίστηκαν σε δύο κατηγορίες που έγιναν σε διαφορετικά δεδομένα. Για το πρώτο μέρος της ανάλυσης, που περιελάμβανε την ανάλυση ομολογίας και την ταξονομική ανάλυση, τα 12 δείγματα δεν αναλύθηκαν ξεχωριστά αλλά έγινε συνένωση (pooling) των αποτελεσμάτων αλληλούχισης από κάθε περίοδο έκθεσης. Έτσι το σύνολο των 6 δειγμάτων από την πρώτη εβδομάδα έκθεσης αποτέλεσε το πρώτο σετ δεδομένων, το σύνολο των 4 δειγμάτων από την τελευταία εβδομάδα έκθεσης αποτέλεσε το δεύτερο σετ δεδομένων ενώ τα 2 δείγματα της πρώτης βδομάδας επαναφοράς αποτέλεσαν τη βάση αναφοράς (baseline) για τη σύγκριση των μικροβιακών πληθυσμών κατά τις διαφορετικές περιόδους έκθεσης. Στα 3 καινούρια σετ δεδομένων εφαρμόστηκε η προαναφερθείσα γραμμή εργασιών μέχρι το σημείο της ταξονομικής ανάλυσης με τις προκαθορισμένες παραμέτρους εργαλείων κάθε φορά και από τα δεδομένα που προέκυψαν χρησιμοποιήθηκαν τα αρχεία .rma από το πρόγραμμα MEGAN εκτός γραμμής εργασιών αυτή τη φορά. Η επιμέρους λειτουργία που χρησιμοποιήθηκε από το MEGAN περιελάμβανε την κανονικοποίηση των σετ δεδομένων, καθώς περιείχαν πολύ διαφορετικό και μη συγκρίσιμο αριθμό αλληλουχιών και τη σύγκριση, των κανονικοποιημένων πλέον δεδομένων μεταξύ τους, για την διεξαγωγή συμπερασμάτων σχετικά με το ταξονομικό και λειτουργικό προφίλ των μεταγενωμικών δειγμάτων. Για την ανακατασκευή των μεταβολικών μονοπατιών μέσω του MinPath χρησιμοποιήθηκαν ξεχωριστά τα δεδομένα αλληλούχισης κάθε δείγματος για να αποφευχθούν χειμερικά συναρμολογήματα που θα προέρχονταν από αναγνωσμένες αλληλουχίες διαφορετικών δειγμάτων.

ΚΕΦΑΛΑΙΟ 3: Αποτελέσματα και Συζήτηση

3.1 Αποτελέσματα μεταγενωμικών δεδομένων από θερμές πηγές

3.1.1 Εντοπισμός πιθανών γονιδίων

Ο εντοπισμός πιθανών γονιδίων ήταν το πρώτο στάδιο της ανάλυσης μέσω της υπολογιστικής γραμμής εργασιών του ANASTASIA και έγινε πάνω στα δεδομένα συναρμολόγησης που είχαν παραχθεί από την ερευνητική ομάδα του πανεπιστημίου της Κοπεγχάγης. Για κάθε ένα από τα 16 δείγματα τα αντίστοιχα δεδομένα εισόδου αποτελούνταν από συναρμολογήματα μεγέθους μεγαλύτερου από 500 βάσεις, καθώς μικρότερες αλληλουχίες θεωρήθηκαν ότι προέρχονται από εσφαλμένες απόπειρες συναρμολόγησης και απορρίφθηκαν. Η ανάλυση των συναρμολογημάτων παρήγαγε λίστες πιθανών γονιδίων, αλλά για λόγους εξασφάλισης της αξιοπιστίας των αποτελεσμάτων η υπολογιστική γραμμή εργασιών είχε ρυθμιστεί κατά τέτοιο τρόπο, ώστε να κρατήσει μόνο όσες αλληλουχίες είχαν προβλεφθεί ως πιθανά γονίδια και από τα τρία εργαλεία (Prodigal, MetaGeneMark, MetaGeneAnnotator). Ο αριθμός των εντοπισμένων πιθανών γονιδίων ανά δείγμα φαίνεται στον παρακάτω πίνακα:

Μεταγενωμικό δείγμα	Πλήθος εντοπισμένων αλληλουχιών γονιδίων
CH1102	40328
It-3	20889
It-6	29910
MW-2	8222
NL-100808	31526
NL-100908	37530
RC-2	3663
Sunspring	32525
Tomsk-sample	11366
A5-7T	2825
A6-2319x	2176
A7-Loc1F3	2644
B1-7Tnr1	6419
B3-6Tnr4	10217
B4-12Tnr1	2462
DG#14,1xg	16221
Planctomycetes	3233

Πίνακας 7. Εντοπισμένα γονίδια ανά μεταγενωμικό δείγμα από το ερευνητικό πρόγραμμα HotZyme. Στον πίνακα σημειώνεται το όνομα δείγματος και το πλήθος των εντοπισμένων πιθανών γονιδίων που προέκυψαν από την συνδυαστική ανάλυση των εργαλείων Prodigal, MetaGeneMark και MetaGeneAnnotator.

3.1.2 Λειτουργικός χαρακτηρισμός γονιδίων

Τα αρχεία δεδομένων που περιείχαν τις λίστες των αλληλουχιών που χαρακτηρίστηκαν ως πιθανά γονίδια, μεταφέρθηκαν αυτόματα ως δεδομένα εισόδου για τα εργαλεία BLAST και HMMER. Για τα δεδομένα της ανάλυσης BLAST θεωρήθηκε ως επικρατέστερο αποτέλεσμα (αλληλουχία-αναφοράς) που αντιστοιχεί σε κάθε άγνωστη αλληλουχία (αλληλουχία-ερώτημα), εκείνο με το μεγαλύτερο ποσοστό ομοιότητας. Έτσι σε κάθε αλληλουχία-ερώτημα αντιστοιχούσαν δύο αλληλουχίες αναφοράς: μία με το μεγαλύτερο ποσοστό ομοιότητας από τη βάση δεδομένων NCBI-nr και μία με το μεγαλύτερο ποσοστό ομοιότητας από τη βάση δεδομένων UniProt/SwissProt. Επειδή για τις ανάγκες του προγράμματος HotZyme αποφασίστηκε να διερευνηθεί η ύπαρξη ενζύμων με συγκεκριμένες υδρολυτικές δράσεις (Πίνακας 8.) απομονώθηκαν τα αποτελέσματα για τα οποία η επικρατέστερη αλληλουχία-αναφοράς αντιστοιχούσε στους αντίστοιχους αριθμούς EC.

Ενζυμικές δράσεις τύπου α'	Ενζυμικές δράσεις τύπου β'
3.1.1.4,3.1.1.15, 3.1.1.17, 3.1.1.19, 3.1.1.25, 3.1.1.27, 3.1.1.30, 3.1.1.65, 3.1.1.68, 3.1.1.81, 3.1.1.83, 3.2.1.111, 3.2.1.127, 3.2.1.129, 3.2.1.130, 3.2.1.169, 3.3.2.8, 3.3.2.10, 3.4.21.4, 3.4.21.5	3.1.1.1, 3.1.1.2, 3.1.1.3, 3.1.1.7, 3.1.1.8, 3.1.1.14, 3.1.1.24, 3.1.1.31, 3.1.1.32, 3.1.1.36, 3.1.1.37, 3.1.1.57, 3.1.1.74, 3.1.4.3, 3.1.4.4, 3.2.1.1, 3.2.1.2, 3.2.1.4, 3.2.1.7, 3.2.1.8, 3.2.1.14, 3.2.1.18, 3.2.1.33, 3.2.1.35, 3.2.1.36, 3.2.1.39, 3.2.1.44, 3.2.1.50, 3.2.1.51, 3.2.1.52, 3.2.1.53, 3.2.1.55, 3.2.1.63, 3.2.1.76, 3.2.1.78, 3.2.1.92, 3.2.1.102, 3.2.1.109, 3.2.1.124, 3.2.1.144, 3.2.1.166, 3.3.2.9, 3.3.2.11, 3.4.21.1, 3.5.1.3, 3.5.1.4, 3.5.1.5, 3.5.1.8, 3.5.1.10, 3.5.1.11, 3.5.1.14, 3.5.1.23, 3.5.1.31, 3.5.1.68, 3.5.1.69, 3.5.4.1, 3.5.4.2, 3.5.4.3, 3.5.4.4, 3.5.4.5, 3.5.4.14, 3.5.4.15, 3.5.4.16, 3.5.5.1, 3.5.5.5, 3.5.5.7, 3.5.99.1, 3.5.99.2

Πίνακας 8. Αριθμοί EC των υδρολυτικών ενζύμων υψηλής προτεραιότητας για το ερευνητικό πρόγραμμα HotZyme. Στον πίνακα φαίνονται δύο λίστες αριθμών EC, τύπου α' και β'. Οι δράσεις που θα αποτελούσαν τη λίστα τύπου α' αντιστοιχούν σε δράσεις που στο πλαίσιο του προγράμματος θεωρήθηκαν ύψιστης σημασίας και για τις οποίες δόθηκε προτεραιότητα στην ανάλυση για την εύρεση νέων ενζύμων.

Οι τιμές στα στατιστικά τεστ του BLAST αποτέλεσαν επιπλέον κριτήρια που εφαρμόστηκαν κατά την απομόνωση αποτελεσμάτων, ώστε να διατηρηθούν μόνο δεδομένα υψηλότερης ομοιότητας. Τα κατώφλια των τιμών αυτών ήταν α)≥30%

ποσοστό πανομοιότυπων αντιστοιχίσεων (percentage of identical matches) και β)≥50% ποσοστό θετικών αντιστοιχίσεων (percentage of positive-scoring matches). Ο αριθμός των αλληλουχιών ανά δείγμα για τις οποίες βρέθηκαν αποτελέσματα ομοιότητας με υδρολάσες, έχοντας υπόψη και τα παραπάνω κριτήρια, φαίνονται στον παρακάτων πίνακα:

Μεταγενωμικό δείγμα	Πλήθος αλληλουχιών με ομοιότητα με ενζυμικές δράσεις τύπου α'	Πλήθος αλληλουχιών με ομοιότητα με ενζυμικές δράσεις τύπου β'
CH1102	-	11
It-3	-	3
It-6	2	30
MW-2	-	3
NL-100808	-	2
NL-100908	-	1
Sunspring	-	99
Tomsk-sample	-	9
A5-7T	-	5
A6-2319x	-	3
A7-Loc1F3	-	8
B1-7Tnr1	-	22
B3-6Tnr4	-	18
B4-12Tnr1	-	8
DG#14,1xg	19	26
Planctomycetes	-	2

Πίνακας 9. Αριθμός αλληλουχιών πιθανών γονιδίων ανά μεταγενωμικό δείγμα από το ερευνητικό πρόγραμμα HotZyme για τις οποίες βρέθηκε ομοιότητα με γνωστές υδρολάσες. Στον πίνακα σημειώνεται το όνομα δείγματος και το πλήθος των εντοπισμένων πιθανών γονιδίων για τα οποία βρέθηκε ομοιότητα με ενζυμικές δράσεις τύπου α' ή β' (βλ. Πίνακα 8.). Για τη σύγκριση ομοιότητας χρησιμοποιήθηκε το πρόγραμμα BLASTp με τις προεπιλεγμένες παραμέτρους του και με χρήση της βάσης δεδομένων UniProt/SwissProt. Τα τελικά δεδομένα που κρατήθηκαν ήταν αυτά για τα οποία ίσχυε α) ≥30% ποσοστό πανομοιότυπων αντιστοιχίσεων (percentage of identical matches) και β) ≥50% ποσοστό θετικών αντιστοιχίσεων (percentage of positive-scoring matches).

Αποφασίστηκε επίσης η εκτενής αναζήτηση στόχων που επιδεικνύουν ενζυμική λειτουργία λιπάσης ή κυτταρινάσης, μέσω της αναζήτησης συγκεκριμένων λειτουργικών περιοχών στην αλληλουχία τους. Για το λόγο αυτό, κατά την επεξεργασία των αποτελεσμάτων της ανάλυσης HMMER, απομονώθηκαν οι αλληλουχίες για τις οποίες βρέθηκαν οι λειτουργικές περιοχές που φαίνονται στον Πίνακα 10 καθώς είναι συνδεδεμένες με τις εν λόγω ενζυμικές δράσεις.

Λειτουργικές περιοχές συνδεδεμένες με δράση λιπάσης	Λειτουργικές περιοχές συνδεδεμένες με δράση κυτταρινάσης
PF00388, PF01764, PF00151, PF00657, PF01477, PF01674, PF03893, PF04083, PF13472, PF03583	PF02927, PF00150, PF02013, PF09478, PF12876, PF12891, PF03443, PF01670, PF03424, PF03425, PF00331, PF00734, PF00759, PF00840, PF02011, PF00942, PF01341, PF02018, PF02015, PF02156, PF00553, PF01270, PF00232, PF00704, PF00722

Πίνακας 10. Αναγνωριστικοί κωδικοί αλληλουχιών λειτουργικών περιοχών που σχετίζονται με δράση κυτταρινάσης και λιπάσης από τη βάση δεδομένων Pfam-A.

Τα αποτελέσματα αυτής της αναζήτησης απέδωσαν τις εξής αλληλουχίες ανα δείγμα:

Μεταγενωμικό δείγμα	Πλήθος αλληλουχιών που περιέχουν τουλάχιστον μία λειτουργική περιοχή συσχετισμένη με δράση λιπάσης	Πλήθος αλληλουχιών που περιέχουν τουλάχιστον μία λειτουργική περιοχή συσχετισμένη με δράση κυτταρινάσης
CH1102	8	25
It-3	3	9
It-6	25	50
MW-2	1	1
NL-100808	12	21
NL-100908	12	18
Sunspring	9	41
Tomsk-sample	23	8
A5-7T	2	19
A6-2319x	4	6
A7-Loc1F3	9	2
B1-7Tnr1	17	17
B3-6Tnr4	25	19
B4-12Tnr1	8	1
DG#14,1xg	60	55
Planctomycetes	13	31

Πίνακας 11. Αριθμός αλληλουχιών πιθανών γονιδίων ανά μεταγενωμικό δείγμα από το ερευνητικό πρόγραμμα HotZyme με μία τουλάχιστον μία λειτουργική περιοχή λιπάσης ή κυτταρινάσης. Στον πίνακα σημειώνεται το όνομα δείγματος και το πλήθος των εντοπισμένων πιθανών γονιδίων για τα οποία βρέθηκε ότι περιέχουν τουλάχιστον μία λειτουργική περιοχή λιπάσης (αριστερή στήλη) ή κυτταρινάσης (δεξιά στήλη). Για την ανάλυση εντοπισμού λειτουργικών περιοχών χρησιμοποιήθηκε το πρόγραμμα HMMER με τις προεπιλεγμένες παραμέτρους του και με χρήση της βάσης δεδομένων Pfam-A.

Η απομόνωση όλων των ανωτέρω αποτελεσμάτων έγινε αξιοποιώντας το εργαλείο εισαγωγής των δεδομένων σε βάση δεδομένων MySQL και των αρχείων εξαγωγής

δεδομένων μορφής .sql που προέκυπταν από αυτό. Μετά την επανεισαγωγή των αρχείων αυτών σε τοπικό υπολογιστή έγιναν οι απαραίτητες αναζητήσεις και φιλτράρισμα δεδομένων, χρησιμοποιώντας τις κατάλληλες εντολές σε γλώσσα MySQL, ώστε να προκύψουν λίστες αλληλουχιών με το επικρατέστερο αποτέλεσμα κάθε φορά είτε από τις αλληλουχίες-αναφοράς είτε από τις λειτουργικές περιοχές. Οι αλληλουχίες-στόχοι που προέκυψαν, καταγράφηκαν σε αρχεία υπολογιστικών φύλλων και διαμοιράστηκαν στους υπόλοιπους συνεργάτες του προγράμματος για περαιτέρω ανάλυση. Η λίστα αυτή επανεξετάστηκε με το εργαλείο EFICAz (και πάλι μέσω της πλατφόρμας ANASTASIA) και όσες από αυτές αναγνωρίστηκαν ως ένζυμα υψηλού ενδιαφέροντος οδηγήθηκαν προς εργαστηριακή απομόνωση και επιβεβαίωση της ενζυμικής λειτουργίας τους.

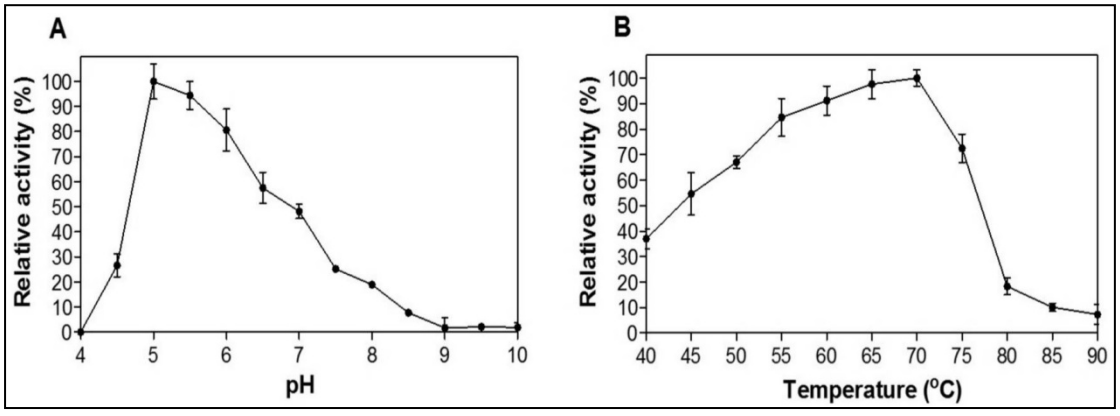
3.1.4 Εργαστηριακή επιβεβαίωση αποτελεσμάτων

Από τη λίστα των αλληλουχιών που προέκυψαν μέσω της ανάλυσης των διαφόρων δειγμάτων, επιλέχθηκαν 2 για εργαστηριακή επιβεβαίωση σε συνεργασία με το Εθνικό Ίδρυμα Ερευνών. Η πρώτη αλληλουχία προήλθε από το δείγμα A5-7T και επιλέχθηκε αρχικά λόγω της υψηλής ομοιότητάς (59%) της αλληλουχίας της με μία ενδο-1,4-βήτα-γλουκανάση (αριθμός EC 3.2.1.4) από το μικροβιακό είδος Bacillus akibai (JCM 9157), όταν εξετάστηκε στη βάση UniProt/SwissProt (αναγνωριστικός κωδικός: P06564.1). Η ίδια αλληλουχία έδειξε 95% ομοιότητα με ένα ένζυμο που έχει αναγνωριστεί ως μέλος της οικογένειας γλυκοζυλικών υδρολασών 5 (glycosyl hydrolase family 5) από το μικροβιακό είδος Thermoanaerobacterium aotearoense όταν εξετάστηκε με κριτήριο τη βάση NCBI-nr (αναγνωριστικός κωδικός: WP_014757289.1). Επίσης βρέθηκε, ότι περιέχει 2 λειτουργικές περιοχές που είναι συνδεδεμένες με ενζυμική δράση κυτταρινάσης, όταν εξετάστηκε με κριτήριο τη βάση Pfam-A: α)την περιοχή με αναγνωριστικό κωδικό PF00150.13 που αντιστοιχεί στην οικογένεια γλυκοζυλικών υδρολασών 5 και β)την περιοχή με αναγνωριστικό κωδικό PF03424.9 που αντιστοιχεί σε οικογένεια δέσμευσης με υδρογονάνθρακες (carbohydrate-binding domain). Η περαιτέρω ανάλυσή της με το εργαλείο EFICAz προέβλεψε ως πιθανή EC κατηγορία στην οποία ανήκει την 3.1.2.4 ισχυροποιώντας τα αρχικά αποτελέσματα πρόβλεψης.

Για την εργαστηριακή απομόνωση της πρώτης αλληλουχίας, η οποία ονομάστηκε CelDZ1, λήφθηκε βιολογικό υλικό του μεταγενωμικού δείγματος A5-7T από την ερευνητική ομάδα του Εθνικού Ιδρύματος Ερευνών και σχεδιάστηκαν οι

κατάλληλοι εκκινητές για την ποσοτική ενίσχυσή της μέσω PCR. Οι αλληλουχίες κλώνοι που δημιουργήθηκαν εισάχθηκαν σε πλασμίδια pET 28a(+) (Novagen) τα οποία χρησιμοποιήθηκαν για το μετασχηματισμό κυττάρων E. coli BL21(DE3). Τα κύτταρα στη συνέχεια αναπτύχθηκαν σε θρεπτικό μέσο που περιλάμβανε το κατάλληλο υπόστρωμα (ισοπρόπυλο-θείο-β-γαλακτοζίδιο - IPTG) ώστε να προκληθεί η έκφραση του ενζύμου, που αντιστοιχούσε στην εισαχθείσα αλληλουχία. Την έκφραση ακολούθησαν μεθοδολογίες απομόνωσης και καθαρισμού (purification) του ενζύμου ενώ ταυτόχρονα σχεδιάστηκαν δοκιμασίες ανάλυσης για το βιοχημικό χαρακτηρισμό του και τον προσδιορισμό της ενεργότητάς του [148].

Από τα αποτελέσματα των δοκιμασιών ανάλυσης προέκυψε ότι η CelDZ1 αποτελούσε μία καινούρια ενδο-γλουκανάση (EC 3.2.1.4) που δρα σε υδατοδιαλυτή κυτταρίνη με βέλτιστη ενεργότητα στους 70^{o} C και pH 5.0 (Εικόνα 23). Οι επιπλέον ιδιότητες του ενζύμου αυτού, όπως παρατηρήθηκαν στο εργαστήριο, περιλαμβάνουν εξαιρετική ανθεκτικότητα σε υψηλές συγκεντρώσεις αλάτων, μεταλλικών ιόντων και άλλων ουσιών που προκαλούν μετουσίωση όπως οργανικοί διαλύτες και τασιενεργές ουσίες [148]. Η αλληλουχία κατατέθηκε στην βάση δεδομένων GenBank [149] και είναι πλέον διαθέσιμη στη βάση δεδομένων UniProt/SwissProt ως πλήρως χαρακτηρισμένη πρωτεΐνη με αναγνωριστικό κωδικό A0A0U4EBH5.



Εικόνα 23. Επίδραση του pH και της θερμοκρασίας στη σχετική ενεργότητα του ενζύμου CelDZ1. (Α) Η ενεργότητα του ενζύμου μετρήθηκε σε αντίδραση υδρόλυσης του υποστρώματος καρβοξυλομεθυλο-κυτταρίνη (CMC) σε θερμοκρασία 40ºC για 5 λεπτά σε τιμές pH που κυμαίνονταν από 4-10. (Β) Η ενεργότητα του ενζύμου μετρήθηκε σε αντίδραση υδρόλυσης του ίδιου υποστρώματος σε σταθερό pH τιμής 5 για 5 λεπτά και σε θερμοκρασιακό εύρος μεταξύ 40 και 90ºC. Οι αναγραφόμενες τιμές αντιστοιχούν στο μέσο όρο μεταξύ τριών ανεξάρτητων πειραμάτων και οι γραμμές σφαλμάτων σε μία τυπική απόκλιση από το μέσο όρο [148].

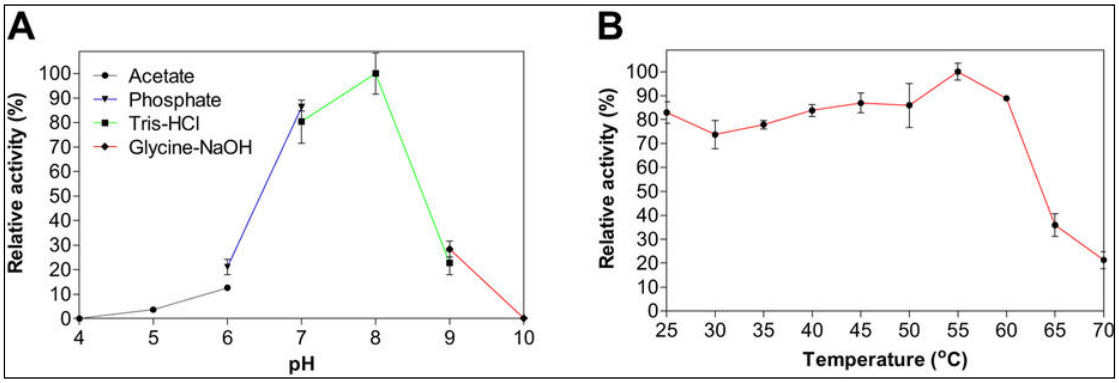
Αντίστοιχα η δεύτερη αλληλουχία προερχόταν από το δείγμα Sunspring και η αλληλουχία της είχε ομοιότητα (23%) με μία ισοπρενοκυστεΐνη μεθυλεστεράση (EC

3.1.1.n2) μέσα στη βάση UniProt/SwissProt (αναγνωριστικός κωδικός: Q94AS5.1). Η έλλειψη υψηλότερης ομοιότητας με οποιαδήποτε πρωτεΐνη μέσα στη βάση δεδομένων ήταν άλλος ένας λόγος που επιλέχθηκε για περαιτέρω ανάλυση, καθώς το γεγονός αυτό την καθιστούσε πιθανό υποψήφιο για να αντιστοιχεί σε ένα εντελώς καινούριο ένζυμο με νέες ιδιότητες. Την υπόθεση αυτή υποστήριξε και η εξέτασή της με κριτήριο τη βάση NCBI-nr, η οποία έδειξε ως υψηλότερο ποσοστό ομοιότητας 82% με μία αλληλουχία χαρακτηρισμένη ως υποθετική λιπάση από το μη καλλιεργήσιμο είδος Acetothermia bacterium (αναγνωριστικός κωδικός: BAL56305.1). Επίσης βρέθηκε ότι περιέχει μία λειτουργική περιοχή που είναι συσχετισμένη με την οικογένεια των α/β υδρολασών (αναγνωριστικός κωδικός: PF07859), όταν εξετάστηκε με κριτήριο τη βάση δεδομένων Pfam-A. Η περαιτέρω ανάλυσή της με το εργαλείο EFICAz έδωσε ως πιθανή EC κατηγορία στην οποία ανήκει, την 3.1.1.- η οποία αντιστοιχεί σε ένζυμα εστερολυτικής δράσης.

Για την εργαστηριακή απομόνωση της δεύτερης αλληλουχίας, η οποία ονομάστηκε estDZ2α, λήφθηκε βιολογικό υλικό του μεταγενωμικού δείγματος Sunspring από την ερευνητική ομάδα του Εθνικού Ιδρύματος Ερευνών και όπως και στην προηγούμενη περίπτωση σχεδιάστηκαν οι κατάλληλοι εκκινητές για την ποσοτική ενίσχυσή της μέσω PCR. Τα πλασμίδια στα οποία εισάχθηκε η νέα αλληλουχία και τα κύτταρα που μετασχηματίστηκαν με αυτά παρέμειναν τα ίδια όπως και στην περίπτωση της celDZ1, ενώ και σε αυτή την περίπτωση σχεδιάστηκαν δοκιμασίες ανάλυσης για τον προσδιορισμό των βιοχημικών χαρακτηριστικών και της ενεργότητάς της.

Από τα αποτελέσματα των δοκιμασιών ανάλυσης προέκυψε ότι η estDZ2α αποτελούσε μία καινούρια καρβοξυλεστεράση (EC 3.1.1.1) με υψηλή ενεργότητα σε εστέρες μεσαίου μήκους αλυσίδας για θερμοκρασίες μεταξύ 25 και 60^{o} C και για τιμές pH 7,0-8,0 (Εικόνα 24). Η συγκεκριμένη αλληλουχία μάλιστα, φαίνεται να ανήκει σε μία νέα οικογένεια βακτηριακών εστερασών η οποία δεν είχε παρατηρηθεί ξανά στο παρελθόν αλλά φέρει πολλές ομοιότητες με την οικογένεια εστερασών IV (HSLs). Η συγκεκριμένη οικογένεια χαρακτηρίζεται από την ύπαρξη του καταλυτικού αμινοξικού μοτίβου GHSAG και για την κατηγοριοποίησή της προτάθηκε ο κωδικός XV [150]. Οι επιπλέον ιδιότητες του ενζύμου αυτού, όπως παρατηρήθηκαν στο εργαστήριο, περιλαμβάνουν εξαιρετική ανθεκτικότητα σε ένα πλήθος διαφορετικών οργανικών διαλυτών όπως ακετόνη, ισοπροπανόλη, ισοοκτάνιο, n-εξάνιο κ.α. Η αλληλουχία κατατέθηκε στην βάση δεδομένων GenBank

και είναι πλέον διαθέσιμη στη βάση δεδομένων UniProt/TrEMBL ως πρωτεΐνη με αναγνωριστικό κωδικό A0A1L2DXZ2.

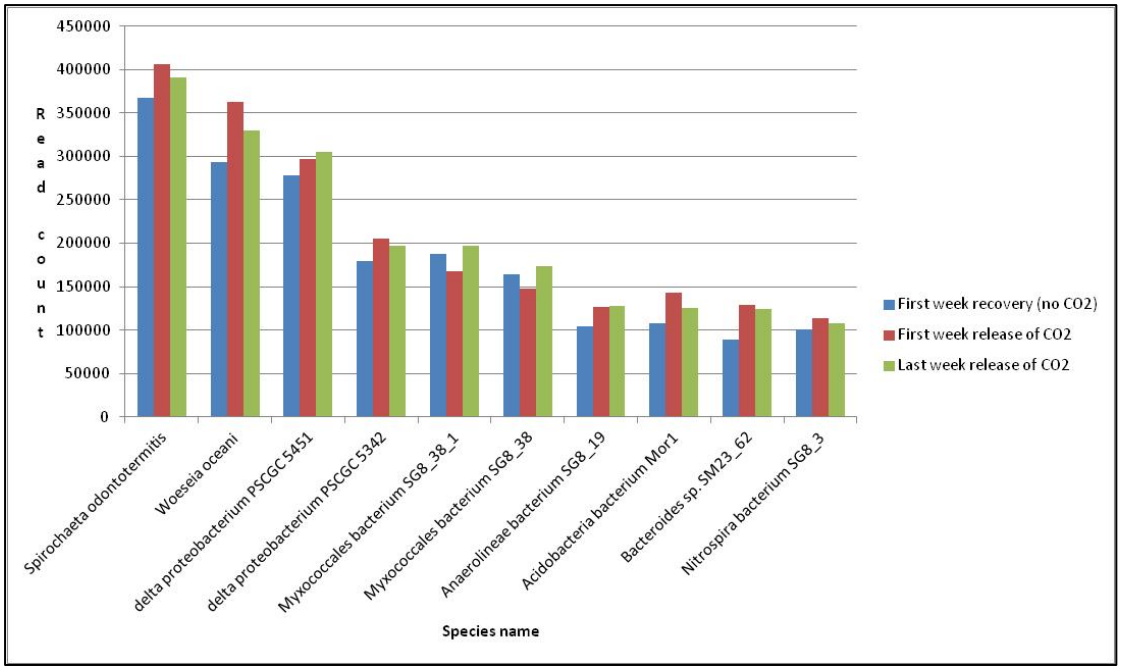


Εικόνα 24. Επίδραση του pH και της θερμοκρασίας στη σχετική ενεργότητα του ενζύμου estDZ2α. (A) Η ενεργότητα του ενζύμου μετρήθηκε σε αντίδραση υδρόλυσης του υποστρώματος καρβοξυλομεθυλο-κυτταρίνη σε θερμοκρασία 40°C για 5 λεπτά σε τιμές pH που κυμαίνονταν από 4-10. (B) Η ενεργότητα του ενζύμου μετρήθηκε σε αντίδραση υδρόλυσης του ίδιου υποστρώματος σε σταθερό pH τιμής 8 για 5 λεπτά και σε θερμοκρασιακό εύρος μεταξύ 25 και 70°C. Οι αναγραφόμενες τιμές αντιστοιχούν στο μέσο όρο μεταξύ τριών ανεξάρτητων πειραμάτων και οι γραμμές σφαλμάτων σε μία τυπική απόκλιση από το μέσο όρο [150].

3.2 Αποτελέσματα μεταγενωμικών δεδομένων πληθυσμών εκτεθειμένων σε υψηλές συγκεντρώσεις CO_2

Η ανάλυση των δεδομένων του προγράμματος COVERALL στηρίχθηκε κυρίως σε δύο εργαλεία ενσωματωμένα στο ANASTASIA, το MEGAN και το DIAMOND (το οποίο αντικαταστάθηκε από το BLAST στην τελική έκδοση της πλατφόρμας). Το πρώτο σετ δεδομένων (πρώτη εβδομάδα έκθεσης CO_2) αποτελούταν από $\sim 13 \cdot 10^7$ αναγνωσμένες αλληλουχίες, το δεύτερο σετ δεδομένων (τελευταία εβδομάδα έκθεσης σε CO_2) αποτελούταν από $\sim 55 \cdot 10^6$ αναγνωσμένες αλληλουχίες ενώ για την πρώτη βδομάδα επαναφοράς τα δεδομένα αποτελούνταν από $\sim 50 \cdot 10^6$. Η σύγκριση των τριών σετ δεδομένων κατέστη εφικτή με την κανονικοποίηση του αριθμού των αναγνωσμένων αλληλουχιών σε $\sim 50 \cdot 10^6$ (49841372) ανά δείγμα η οποία έγινε αυτόματα μέσω του προγράμματος MEGAN.

Η ταξονομική ανάλυση έδειξε περίπου το ίδιο προφίλ για τα επικρατέστερα είδη στις 3 διαφορετικές περιόδους (Εικόνα 25) με το επικρατέστερο όλων να είναι το Spirochaeta odontotermitis.



Εικόνα 25. Δέκα επικρατέστερα μικροβιακά είδη σε διαφορετικές συνθήκες έκθεσης σε CO_2. Το ποσοτικό μέτρο με το οποίο καθορίστηκε η παρουσία ενός είδους ήταν ο αριθμός αναγνωσμένων αλληλουχιών που αντιστοιχίστηκαν σε αυτό με βάση την ταξονομική ανάλυση. Στο ραβδόγραμμα παρουσιάζονται με μπλε χρώμα τα επικρατέστερα είδη κατά τη διάρκεια διακοπής του CO_2, με κόκκινο χρώμα τα είδη κατά τη διάρκεια της πρώτης εβδομάδας έκθεσης και με πράσινο τα είδη της τελευταίας εβδομάδας έκθεσης. Όλες οι τιμές περιγράφουν τον αριθμό αναγνωσμένων αλληλουχιών όπως προέκυψαν από τα κανονικοποιημένα δεδομένα.

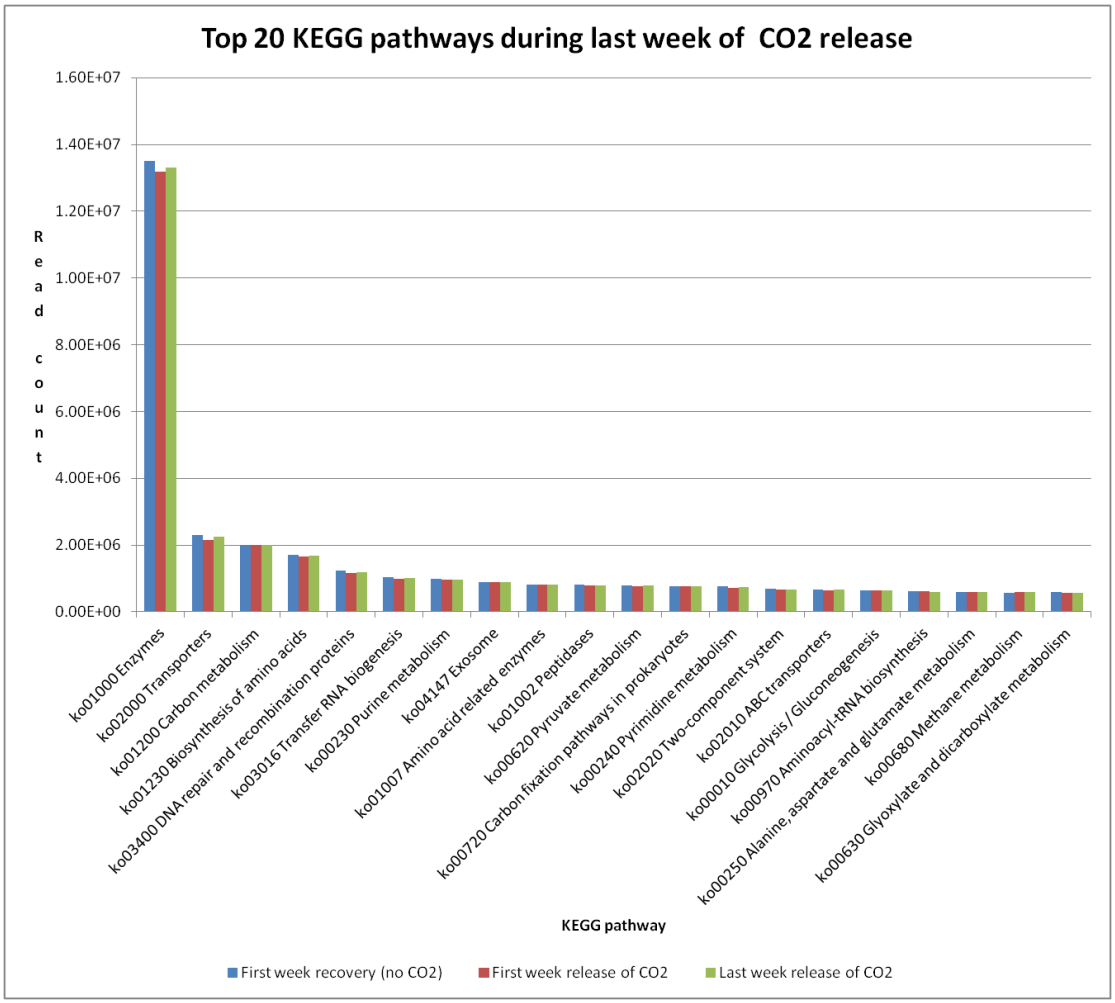
Παρ' όλα αυτά, όταν εξετάστηκε το ταξονομικό προφίλ στο σύνολό του για τις διαφορετικές περιόδους εντοπίστηκαν 23 διαφορετικά είδη για τα οποία υπήρχαν καταχωρήσεις μόνο κατά τη διάρκεια έκθεσης σε CO_2. Από τα είδη αυτά φαίνεται να ξεχωρίζει το Magnetovibrio blakemorei, το οποίο όχι μόνο είναι το επικρατέστερο εξ' αυτών αλλά φαίνεται να διπλασιάζει σχεδόν τον πληθυσμό του μεταξύ της πρώτης και τελευταίας εβδομάδας έκθεσης, σε αντίθεση με τα υπόλοιπα, των οποίων το επίπεδο διατηρείται σταθερό. Τα είδη αυτά φαίνονται στον Πίνακα 12. Η απομόνωση των αναγνωσμένων αλληλουχιών τους, έδειξε ότι οι καταχωρήσεις αυτές προέρχονται από όλα τα δείγματα που αποτελούν τα συνενωμένα σετ δεδομένων από κάθε περίοδο έκθεσης (6 δείγματα για την πρώτη εβδομάδα έκθεσης και 4 δείγματα για την τελευταία εβδομάδα έκθεσης). Οι αναλύσεις για τα συγκεκριμένα σετ δεδομένων συνεχίζονται κατά τη διάρκεια συγγραφής αυτής της διατριβής με τα επόμενα βήματα να αποτελούνται από: α)την ξεχωριστή συναρμολόγηση των αναγνωσμένων αλληλουχιών που αντιστοιχίστηκαν σε κάθε είδος και β)τον εντοπισμό χαρακτηριστικών αλληλουχιών ενζύμων στα συναρμολογήματα που προκύπτουν, οι οποίες θα μπορέσουν να αποτελέσουν βιοδείκτες έκθεσης σε υψηλές συγκεντρώσεις CO_2. Χρησιμοποιώντας το ίδιο εργαλείο, σε συνδυασμό και πάλι με τα αποτελέσματα

της ανάλυσης ομολογίας, έγινε ανάλυση του λειτουργικού δυναμικού των μεταγενωμικών δειγμάτων. Κατά την ανάλυση αυτή αντιστοιχίστηκαν οι αναγνωσμένες αλληλουχίες σε μεταβολικά μονοπάτια από τη βάση δεδομένων KEGG. Από την κατηγοριοποίηση των αναγνωσμένων αλληλουχιών παρατηρήθηκε παρόμοιο προφίλ στα επικρατέστερα μεταβολικά μονοπάτια KEGG και για τις τρεις περιόδους (Εικόνα 26).

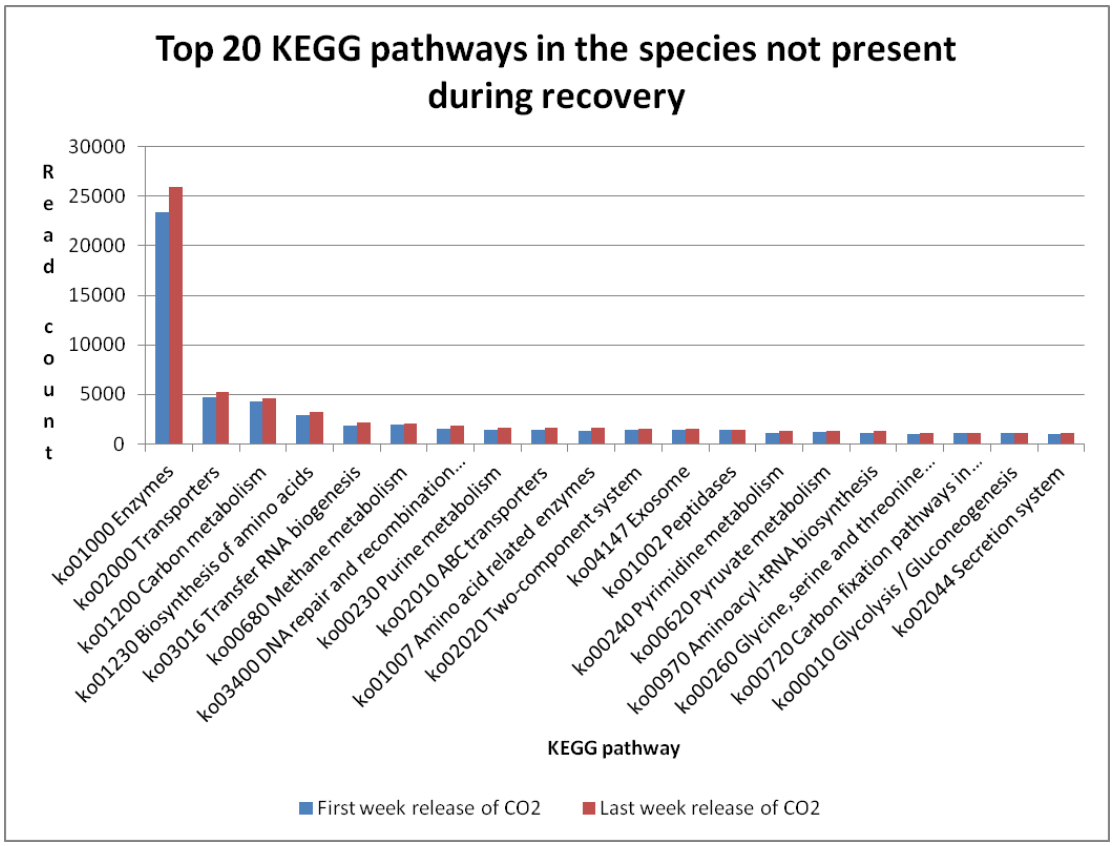
Species name	First week recovery (no CO2	First week release of CO2	Last week release
Magnetovibrio blakemorei	0	12054	23393
Candidatus Bathyarchaeota archaeon RBG_16_57_9	0	7536	6368
Candidatus Latescibacteria bacterium 4484_7	0	7300	6254
Saccharicrinis fermentans	0	7826	6192
Desulfobacterales bacterium C00003060	0	5770	5958
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_02_FULL_52_10	0	7332	5813
Desulfococcus sp. 4484_242	0	5889	5794
Geothermobacter sp. EPR-M	0	6825	5743
miscellaneous Crenarchaeota group-1 archaeon SG8-32-1	0	5868	5738
miscellaneous Crenarchaeota group-6 archaeon AD8-1	0	5463	5697
candidate division KSB1 4572_119	0	6471	5665
Desulfobacterales bacterium RIFOXYA12_FULL_46_15	0	6603	5640
Candidatus Cloacimonas sp. 4484_143	0	7793	5634
Deltaproteobacteria bacterium RBG_16_49_23	0	5698	5629
Gammaproteobacteria bacterium RIFCSPLOWO2_02_FULL_61_13	0	6246	5522
Desulfatirhabdium butyrativorans	0	5866	5510
Coxiella sp. DG_40	0	6483	5467
Bacteroidetes bacterium GWF2_38_335	0	7642	5449
Chloroflexi bacterium OLB15	0	6434	5434
Deltaproteobacteria bacterium RBG_16_48_10	0	5633	5363
Bacteroidetes bacterium GWA2_33_15	0	6959	5360
candidate division Zixibacteria bacterium SM23_81	0	6233	5337
Planctomycetes bacterium GWF2_42_9	0	6536	5310
candidate division Zixibacteria bacterium 4484_95	0	6276	5300
Gammaproteobacteria bacterium RIFCSPLOWO2_02_FULL_56_15	0	6094	5233
Planctomycetes bacterium RBG_13_44_8b	0	6599	5205
Steroidobacter denitrificans	0	5789	5197
Labilibacter aurantiacus	0	6342	5157
Sunxiuqinia elliptica	0	6187	5133
Lentisphaerae bacterium GWF2_44_16	0	5890	5074
Candidatus Schekmanbacteria bacterium RBG_13_48_7	0	5212	5044

Πίνακας 12. Αριθμός αναγνωσμένων αλληλουχιών ανά είδος σε διαφορετικές συνθήκες έκθεσης CO_2. Στον πίνακα σημειώνονται μόνο τα μικροβιακά είδη για τα οποία δεν βρέθηκαν αντιστοιχισμένες αλληλουχίες κατά τη διάρκεια της διακοπής της έκθεσης σε CO_2. Όλες οι τιμές περιγράφουν τον αριθμό αναγνωσμένων αλληλουχιών όπως προέκυψαν από τα κανονικοποιημένα δεδομένα.

Στη συνέχεια έγινε απομόνωση των αναγνωσμένων αλληλουχιών που είχαν αντιστοιχιστεί στα 23 μη παρόντα είδη κατά την πρώτη εβδομάδα επαναφοράς και επανεξετάστηκαν τα μεταβολικά μονοπάτια στα οποία ανήκουν. Οι πιο σημαντικές διαφορές που παρατηρήθηκαν στα επικρατέστερα μονοπάτια KEGG ήταν η εντονότερη παρουσία του μεταβολικού μονοπατιού για τον μεταβολισμό μεθανίου (ko00680) και η αντίστοιχη μείωση του ποσοστού του μονοπατιού για αφομοίωση άνθρακα σε προκαρυώτες (ko00720). Παρ' όλα αυτά και τα δύο μεταβολικά μονοπάτια παραμένουν στα 20 επικρατέστερα που εντοπίζονται και στην περίπτωση μελέτης των 23 ειδών και στην περίπτωση μελέτης του συνόλου του μεταγονιδιώματος.



Εικόνα 26. Είκοσι επικρατέστερα μεταβολικά μονοπάτια KEGG σε διαφορετικές συνθήκες έκθεσης σε CO_2. Το ποσοτικό μέτρο με το οποίο καθορίστηκε η παρουσία ενός μεταβολικού μονοπατιού ήταν ο αριθμός αναγνωσμένων αλληλουχιών που αντιστοιχίστηκαν σε αυτό με βάση τη λειτουργική ανάλυση. Στο ραβδόγραμμα παρουσιάζονται με μπλε χρώμα τα επικρατέστερα μεταβολικά μονοπάτια κατά τη διάρκεια διακοπής του CO_2, με κόκκινο χρώμα τα μεταβολικά μονοπάτια κατά τη διάρκεια της πρώτης εβδομάδας έκθεσης και με πράσινο τα μεταβολικά μονοπάτια της τελευταίας εβδομάδας έκθεσης. Όλες οι τιμές περιγράφουν τον αριθμό αναγνωσμένων αλληλουχιών όπως προέκυψαν από τα κανονικοποιημένα δεδομένα.



Εικόνα 27. Είκοσι επικρατέστερα μεταβολικά μονοπάτια KEGG σε διαφορετικές συνθήκες έκθεσης σε CO_2 για τα 23 μικροβιακά είδη που δεν είναι παρόντα κατά την περίοδο διακοπής της έκθεσης σε CO_2. Το ποσοτικό μέτρο με το οποίο καθορίστηκε η παρουσία ενός μεταβολικού μονοπατιού ήταν ο αριθμός αναγνωσμένων αλληλουχιών που αντιστοιχίστηκαν σε αυτό με βάση τη λειτουργική ανάλυση. Στο ραβδόγραμμα παρουσιάζονται με μπλε χρώμα τα μεταβολικά μονοπάτια κατά τη διάρκεια της πρώτης εβδομάδας έκθεσης και με κόκκινο τα μεταβολικά μονοπάτια της τελευταίας εβδομάδας έκθεσης. Όλες οι τιμές περιγράφουν τον αριθμό αναγνωσμένων αλληλουχιών όπως προέκυψαν από τα κανονικοποιημένα δεδομένα.

ΚΕΦΑΛΑΙΟ 4: Συμπεράσματα

4.1 Ανάπτυξη αυτοματοποιημένης υπολογιστικής πλατφόρμας και εφαρμογή της για την ανάλυση μεταγενωμικών δειγμάτων

Στην παρούσα διδακτορική διατριβή αναπτύχθηκε μία διαδικτυακή πλατφόρμα που ονομάστηκε ANASTASIA (Automated Nucleotide Aminoacid Sequences Translational plAtform for Systemic Interpretation and Analysis) με σκοπό την εύκολη αποθήκευση και αυτοματοποιημένη ανάλυση μεταγενωμικών δεδομένων, από τα οποία θα εντοπιστούν ένζυμα βιομηχανικού ενδιαφέροντος. Στο ANASTASIA έχει ενσωματωθεί ένα πλήθος βιοπληροφορικών εργαλείων που μπορούν να αναλάβουν ένα πολύ μεγάλο εύρος αναλύσεων από το πρωταρχικό επίπεδο των δεδομένων αλληλούχισης, έως το τελικό στάδιο του εντοπισμού γονιδίων και των μεταβολικών μονοπατιών στα οποία εμπλέκονται. Στην εν λόγω πλατφόρμα έχουν ενσωματωθεί ήδη δημοσιευμένα εργαλεία καθώς και εξειδικευμένοι αλγόριθμοι γραμμένοι σε διάφορες γλώσσες προγραμματισμού, οι οποίοι αναπτύχθηκαν στο πλαίσιο της παρούσας διδακτορικής διατριβής. Ο σχεδιασμός της βασίστηκε στην πλατφόρμα ανοιχτού κώδικα Galaxy της οποίας ο πηγαίος κώδικας τροποποιήθηκε κατά τέτοιο τρόπο ώστε να απλουστευτεί το γραφικό περιβάλλον εργασίας της. Ταυτόχρονα, η πλατφόρμα εμπλουτίστηκε με σχεδιασμένες βιοπληροφορικές γραμμές εργασιών, ειδικά προσαρμοσμένες για την ανάλυση μεταγενωμικών δεδομένων, με τις οποίες επιτυγχάνεται η αυτοματοποιημένη διαδοχική λειτουργία των απαραίτητων εργαλείων. Η ανάπτυξη αυτών των γραμμών εργασιών έγινε κατά την αξιοποίηση της πλατφόρμας και την εφαρμογή των εργαλείων της για την ανάλυση πραγματικών μεταγενωμικών δεδομένων.

Τα δεδομένα που αναλύθηκαν προήλθαν από μεταγενωμικά δείγματα που συλλέχθηκαν στο πλαίσιο δύο διαφορετικών ερευνητικών προγραμμάτων, το πρόγραμμα HotZyme και το πρόγραμμα COVERALL. Τα δείγματα από το πρόγραμμα HotZyme προήλθαν από θερμά περιβάλλοντα και η ανάλυση τους, στοχεύει στον εντοπισμό νέων ενζύμων υδρολυτικής δράσης και υψηλής θερμοσταθερότητας. Αντίστοιχα το πρόγραμμα COVERALL περιέλαβε δείγματα από μικροβιακούς πληθυσμούς εκτεθειμένους σε υψηλές συγκεντρώσεις CO_2 και ο στόχος του είναι η μελέτη των ταξονομικών και λειτουργικών τους προφίλ στις συνθήκες αυτές. Οι βιοπληροφορικές γραμμές εργασιών που αναπτύχθηκαν για την ανάλυση των παραπάνω δεδομένων εφαρμόστηκαν με επιτυχία και ήδη ορισμένες από τις προβλέψεις τους σχετικά με τον εντοπισμό και τον λειτουργικό χαρακτηρισμό

πρωτεϊνών έχουν επιβεβαιωθεί εργαστηριακά οδηγώντας στην ανακάλυψη δύο εντελώς νέων ενζύμων υψηλής θερμοσταθερότητας.

Εδώ θα πρέπει να σημειωθεί ότι η σημαντικότητα αυτού του εγχειρήματος δεν υφίσταται μόνο στην τελική ανακάλυψη των νέων ενζύμων, αλλά και στη συνολική και μαζική διαχείριση του όγκου δεδομένων, με το βέλτιστο πιθανό τρόπο, τόσο προς την ποιότητα των αποτελεσμάτων όσο και προς τη διεξοδικότητα των αναλύσεων. Η πλατφόρμα αυτή, μαζί με τις αυτοματοποιημένες γραμμές εργασιών της, έχει σχεδιαστεί κατά τέτοιον τρόπο ώστε να μπορεί να μειώνει στο ελάχιστο την ανάγκη παρέμβασης του χρήστη και αλληλεπίδρασής του με τα δύσχρηστα δεδομένα που προκύπτουν από τις τεχνολογίες αλληλούχισης νέας γενιάς και από τα προγράμματα ανάλυσής τους. Αυτό σημαίνει ότι η εφαρμογή της πλατφόρμας ANASTASIA μπορεί να προσφέρει πρακτικά τη δυνατότητα σε οποιονδήποτε ερευνητή, δίχως εξειδικευμένες γνώσεις βιοπληροφορικής, να αναλύσει δεδομένα που προκύπτουν από μεταγενωμικά πειράματα και να εντοπίσει μέσα από εκατομμύρια αλληλουχίες τα γονίδια που τον ενδιαφέρουν. Οι εν λόγω βιοπληροφορικές γραμμές εργασιών που προσφέρουν αυτή τη δυνατότητα βελτιστοποιήθηκαν περαιτέρω και μετά τη λήξη των ερευνητικών προγραμμάτων στα οποία χρησιμοποιούνταν ώστε να είναι έτοιμες για την ενσωμάτωσή τους στην τελική έκδοση της πλατφόρμας. Η έκδοση αυτή είναι διαθέσιμη διαδικτυακά μέσω ενός διακομιστή της σχολής Χημικών Μηχανικών ΕΜΠ στη διεύθυνση http://motherbox.chemeng.ntua.gr/anastasia_dev/.

4.2 Εντοπισμός ενζύμων υδρολυτικής δράσης από θερμές πηγές

Η βιοπληροφορική ανάλυση των δεδομένων μεταγενωμικών δειγμάτων από θερμά περιβάλλοντα (πρόγραμμα HotZyme) είχε ως αποτέλεσμα την απόκτηση μίας λίστας εκατοντάδων αλληλουχιών, που προβλέφθηκαν να αντιστοιχούν σε ένζυμα υδρολυτικής δράσης με ταυτόχρονη πιθανή θερμοσταθερότητα. Η αναζήτηση υδρολασών περιορίστηκε σε δύο κατηγορίες «υψηλής» και «χαμηλής» προτεραιότητας με την κάθε μία να αποτελείται από συγκεκριμένες ενζυμικές δράσεις κατηγοριοποιημένες με βάση τον αντίστοιχο αριθμό EC τους. Η πρόβλεψη του αριθμού EC τους έγινε με αναλύσεις ομολογίας έναντι των βάσεων δεδομένων γνωστών αλληλουχιών NCBI-nr και UniProt/SwissProt, καθώς και με μεθοδολογίες μηχανικής μάθησης (EFICAz). Διερευνήθηκε επιπλέον η ύπαρξη ενζύμων με δράση κυτταρινάσης ή λιπάσης με βάση την ύπαρξη συντηρημένων περιοχών (domains) μέσα στην αλληλουχία τους, οι οποίες έχουν συσχετιστεί με αυτές τις λειτουργίες. Η

διερεύνηση αυτή έγινε με χρήση αλγορίθμων βασισμένων σε κρυφά μοντέλα Markov (HMMER), ή αξιοποιώντας τη βάση δεδομένων Pfam-A που αποτελείται από ήδη γνωστές συντηρημένες περιοχές.

Από το σύνολο των αλληλουχιών πιθανής υδρολυτικής δράσης επιλέχθηκαν δύο για εργαστηριακή επιβεβαίωση σε συνεργασία με το Εθνικό Ίδρυμα Ερευνών. Οι συγκεκριμένες αλληλουχίες είχανε δείξει υψηλή ομολογία με αντίστοιχες αλληλουχίες ενζύμων αριθμού EC 3.2.1.4 (ενδογλουκανάση) και 3.1.1.- (καρβοξυλεστεράση) αντίστοιχα και η ανάλυσή τους με τις μεθοδολογίες μηχανικής μάθησης υποστήριζε αυτή την πρόβλεψη. Η πρόβλεψη αυτή στηριζόταν επιπλέον και από τις συντηρημένες λειτουργικές περιοχές που εντοπίστηκαν στην αλληλουχία τους. Συγκεκριμένα η αλληλουχία με αρχική πρόβλεψη ως ένζυμο αριθμού EC 3.2.1.4, που ονομάστηκε CelDZ1, βρέθηκε να περιέχει 2 συντηρημένες λειτουργικές περιοχές οι οποίες είναι συσχετισμένες με δράση κυτταρινάσης. Αντίστοιχα η δεύτερη αλληλουχία, με αρχική πρόβλεψη ως ένζυμο αριθμού EC 3.1.1.-, βρέθηκε να περιέχει 1 συντηρημένη λειτουργική περιοχή η οποία είναι συσχετισμένη με την οικογένεια των α/β υδρολασών. Η επιτυχής απομόνωσή τους στο εργαστήριο και οι δοκιμασίες ανάλυσης επιβεβαίωσαν τόσο την εκ των προτέρων προβλεπόμενη λειτουργία τους, όσο και την αναμενόμενη θερμοσταθερότητά τους λόγω των περιβαλλόντων από τα οποία προήλθαν.

Παρ' όλο που τα τελικά αποτελέσματα περιλαμβάνουν μία μεγάλη λίστα αλληλουχιών η οποία μπορεί να καταλήξει στην εργαστηριακή απομόνωση νέων θερμοσταθερών ενζύμων υδρολυτικής δράσης, πρέπει να τονιστεί ότι η σημαντικότητα των δεδομένων αυτών είναι πολύ μεγαλύτερη. Η ίδια ακριβώς μεθοδολογία που χρησιμοποιήθηκε για τον εντοπισμό νέων υδρολασών μπορεί να χρησιμοποιηθεί για την εύρεση οποιασδήποτε άλλης κατηγορίας ενζύμων. Δεδομένου μάλιστα την προέλευση των μεταγενωμικών δεδομένων, για τα καινούρια ένζυμα που θα ανακαλυφθούν, υπάρχει πολύ μεγάλη πιθανότητα να χαρακτηρίζονται από υψηλή θερμοσταθερότητα. Επιπλέον στη μεθοδολογία αυτή μπορεί να μειωθεί η αυστηρότητα με την οποία απορρίπταμε αποτελέσματα αναλύσεων. Για παράδειγμα η μείωση στατιστικών κατωφλιών στην ανάλυση ομολογίας για τα οποία δεχόμασταν αποτελέσματα ή η χρήση αλληλουχιών που προέκυψαν από τουλάχιστον ένα από τα τρία εργαλεία εντοπισμού γονιδίων θα ήταν τρόποι με τους οποίους θα προέκυπτε ένας μεγαλύτερος αριθμός αλληλουχιών πιθανών ενζύμων.

4.2 Μελέτη ταξονομικού προφίλ και μεταβολικών λειτουργιών πληθυσμών υπό την επήρεια έκθεσης σε CO_2 για την εύρεση αλληλουχιών/βιοδεικτών έκθεσης

Για την ανάπτυξη μίας αυτοματοποιημένης γραμμής εργασιών για τον προσδιορισμό του ταξονομικού και λειτουργικού δυναμικού ενός μεταγενωμικού δείγματος χρησιμοποιήθηκαν τα δεδομένα αλληλούχισης από το πρόγραμμα COVERALL. Για την ανάλυση των εν λόγω δεδομένων χρησιμοποιήθηκε ανάλυση ομολογίας με το πρόγραμμα DIAMOND, ταξονομική ανάλυση με το πρόγραμμα MEGAN και λειτουργική ανάλυση με τα προγράμματα MEGAN και MinPath. Η τελική γραμμή εργασιών που αναπτύχθηκε περιέλαβε και το βήμα ποιοτικού ελέγχου τον αναγνωσμένων αλληλουχιών που στην προκειμένη περίπτωση παραλείφθηκε, καθώς είχε εφαρμοστεί από την εταιρεία που ανέλαβε την αλληλούχιση.

Η μελέτη των περιβαλλοντικών δειγμάτων, που έγινε κάτω από ελεγμένες συνθήκες έκθεσης σε CO_2, έδειξε ότι δεν υπάρχει κάποια διαφορά στα επικρατέστερα είδη μεταξύ των περιόδων έκθεσης και επαναφοράς. Παρ' όλα αυτά βρέθηκαν 23 διαφορετικά είδη που εμφανίζονται μόνο κατά τις περιόδους έκθεσης σε CO_2. Μάλιστα ένα από αυτά το Magnetovibrio blakemorei φαίνεται να αυξάνει το ποσοστό του κατά τη διάρκεια της περιόδου (4 εβδομάδες) κατά την οποία τα δείγματα εκτέθηκαν σε CO_2, κάτι που το καθιστά εξαιρετικό υποψήφιο για συσχετισμό χαρακτηριστικών αλληλουχιών του με την έκθεση στον εν λόγω ρύπο. Η μελέτη των αντίστοιχων KEGG μεταβολικών μονοπατιών έδειξε και πάλι ακριβώς το ίδιο προφίλ στα επικρατέστερα (αυτά στα οποία αντιστοιχίστηκαν οι περισσότερες αναγνωσμένες αλληλουχίες) μεταβολικά μονοπάτια. Παρόμοιο προφίλ παρατηρήθηκε όταν απομονώθηκαν οι αλληλουχίες που κατηγοριοποιήθηκαν σε ένα από τα 23 είδη που ήταν παρόντα κατά την έκθεση του CO_2.

Κατά τη διάρκεια συγγραφής της παρούσας διατριβής διεξάγεται περαιτέρω ανάλυση των δεδομένων ώστε να απομονωθούν αλληλουχίες από αυτά τα είδη που θα είναι μοναδικές σε σχέση με αυτές του υπόλοιπου μικροβιακού πληθυσμού. Στην περίπτωση εντοπισμού τέτοιων αλληλουχιών, είτε από τα δεδομένα αναγνωσμένων αλληλουχιών είτε από τα δεδομένα μιας ενδεχόμενης συναρμολόγησης, θα μπορούν να χρησιμοποιηθούν ως αξιόπιστοι βιοδείκτες για το γρήγορο εντοπισμό υψηλών συγκεντρώσεων CO_2 σε ένα υποθαλάσσιο περιβάλλον. Μία επιπλέον προσέγγιση που θα ήταν ενδιαφέρουσα είναι η αντίστοιχη μελέτη του μετα-μεταγραφώματος αυτών των δειγμάτων, για την αποσαφήνιση του αριθμού των γονιδίων που πραγματικά εκφράζονται κάτω από τέτοιες συνθήκες. Τέτοιου είδους δεδομένα έχουν ήδη

αποκτηθεί από το πρόγραμμα COVERALL και αναμένεται η περαιτέρω επεξεργασία τους με ταυτόχρονη ανάπτυξη επιπλέον γραμμών εργασιών στην πλατφόρμα ANASTASIA.

4.3 Προοπτικές εξέλιξης της υπολογιστικής υποδομής

Το ANASTASIA αποτελεί μία ολοκληρωμένη υπολογιστική υποδομή για την ανάλυση μεταγενωμικών δεδομένων αλλά βρίσκεται σε συνεχή εξέλιξη με την προσθήκη νέων εργαλείων και τη βελτιστοποίηση των αλγορίθμων του. Οι τωρινές δυνατότητες ανάλυσης της πλατφόρμας περιλαμβάνουν μεν όλα τα επίπεδα δεδομένων αλλά σε καθένα από αυτά έχουν προγραμματιστεί αλλαγές για τις επόμενες εκδόσεις:

Στο επίπεδο ποιοτικού ελέγχου, η απομάκρυνση κάθε αναγνωσμένης αλληλουχίας βασίζεται στη διάμεσο του σκορ ποιότητας του συνόλου των βάσεών της μέσω του εργαλείου FASTX. Αυτό όμως αφήνει ανοιχτό το ενδεχόμενο να κρατηθούν αναγνωσμένες αλληλουχίες στις οποίες ένα αρκετά μεγάλο ποσοστό των βάσεων είναι χαμηλής ποιότητας. Εμπειρικά έχει παρατηρηθεί, ότι οι βάσεις χαμηλής ποιότητας βρίσκονται συνήθως στην αρχή ή στις τελευταίες θέσεις της αναγνωσμένης αλληλουχίας. Ένας συνδυασμός του εργαλείου FASTQC και του FASTX θα έλυνε αυτό το πρόβλημα καθώς το πρώτο θα εντόπιζε τις θέσεις σε κάθε αναγνωσμένη αλληλουχία όπου υπάρχει πρόβλημα ποιότητας και το δεύτερο περιλαμβάνει εργαλεία αποκοπής συγκεκριμένου αριθμού βάσεων από την αρχή ή το τέλος.

Στο επίπεδο συναρμολόγησης, το οποίο προς το παρόν χειρίζεται το εργαλείο Megahit, έχει ήδη σχεδιαστεί αλγόριθμος ενσωμάτωσης του MetaVelvet (όταν είναι διαθέσιμη η νέα έκδοσή του με την απομάκρυνση των τωρινών του σφαλμάτων) στο ANASTASIA. Επίσης εξετάζεται η περίπτωση ενσωμάτωσης προγραμμάτων μετα-συναρμολόγησης [151], δηλαδή της χρήσης συναρμολογημάτων από διαφορετικά εργαλεία για την εκ νέου συναρμολόγηση και διόρθωση πιθανών σφαλμάτων.

Για τον εντοπισμό γονιδίων το Prodigal είναι ένα πολύ αξιόπιστο εργαλείο αλλά εξειδικεύεται μόνο σε γονίδια από προκαρυωτικούς οργανισμούς. Θα πρέπει να συνδυαστεί με κάποιο αντίστοιχο για ευκαρυωτικούς οργανισμούς [152] ώστε να σχηματίζεται μία πιο ολοκληρωμένη εικόνα για τον μικροβιακό πληθυσμό ενός μεταγενωμικού δείγματος.

Ο χαρακτηρισμός των αλληλουχιών γονιδίων είναι άλλο ένα σημείο στο οποίο πρέπει να επιστρατευτούν νέοι αλγόριθμοι και εξειδικευμένες μεθοδολογίες. Το

σχετικά μικρό ποσοστό των οργανισμών που έχει καλλιεργηθεί στο εργαστήριο και ως συνέπεια το μικρό ποσοστό των αντίστοιχων γονιδίων που έχουν καταγραφεί, καθιστά λιγότερο αποτελεσματικές τις τεχνικές σύγκρισης ομολογίας με βάσεις δεδομένων γνωστών αλληλουχιών. Αντίθετα οι μεθοδολογίες μηχανικής μάθησης στηρίζονται σε μοτίβα που παρουσιάζονται σε γονίδια ίδιας ενζυμικής δράσης και μπορούν να αποτελέσουν έναν πιο αποδοτικό τρόπο πρόβλεψης άγνωστων αλληλουχιών. Ο σχεδιασμός ενός τέτοιου εργαλείου έχει ήδη αρχίσει σε συνεργασία με την ερευνητική ομάδα του Εθνικού Ιδρύματος Ερευνών και την εταιρεία βιοπληροφορικής e-NIOS στοχεύοντας στην βελτιστοποίηση της ταχύτητας των αναλύσεων και της ικανότητας ορθής πρόβλεψης των αντίστοιχων ενζυμικών λειτουργιών. Την ολοκλήρωση του εργαλείου θα ακολουθήσει η ενσωμάτωσή του στην πλατφόρμα και στις αυτοματοποιημένες γραμμές εργασιών της.

Κατά τη έναρξη της εκπόνησης αυτής της διατριβής οι βάσεις δεδομένων KEGG ήταν πλήρως διαθέσιμες για την επιστημονική κοινότητα, αλλά αυτό έχει πλέον αλλάξει με τη χρήση τους να προϋποθέτει μία ετήσια συνδρομή για τον εκάστοτε χρήστη. Για τον λειτουργικό χαρακτηρισμό ενός μεταγενωμικού δείγματος μία εναλλακτική λύση είναι η χρήση της βάσης δεδομένων SEED, η οποία προσφέρει παρόμοιες οντολογίες και μεταβολικά μονοπάτια και μπορεί να αξιοποιηθεί εξίσου αποδοτικά από τα προγράμματα MEGAN και MinPath. Μία επιπλέον επιλογή που εξετάζεται είναι η βάση δεδομένων Reactome [153] της EBI καθώς αποτελείται από μία πιο εκτεταμένη συλλογή μεταβολικών μονοπατιών και χαρακτηρίζεται από μία πολύ πιο ενεργή και δραστήρια κοινότητα χρηστών.

Σχεδιάζονται ήδη επιπλέον εργαλεία αποθήκευσης σε βάση MySQL των αποτελεσμάτων των υπόλοιπων εργαλείων π.χ. των εργαλείων πρόβλεψης ενζυμικής δράσης. Για την ενσωματωμένη Web2py πλατφόρμα οπτικοποίησης των δεδομένων αποτελεσμάτων ανάλυσης σχεδιάζονται επίσης αλλαγές στον πηγαίο κώδικά της, με σκοπό τη δυνατότητα συνένωσης διαφορετικών βάσεων δεδομένων μεταξύ τους βάση του αναγνωριστικού κωδικού της εκάστοτε αλληλουχίας. Το τελικό αποτέλεσμα θα είναι ο χρήστης να μπορεί να εξάγει εξειδικευμένες αναφορές με τα αντίστοιχα αποτελέσματα των αναλύσεων των αλληλουχιών που τον ενδιαφέρουν όπως π.χ. τα αποτελέσματα από το BLAST, το HMMER και το EFICAz και που πληρούν συγκεκριμένα κριτήρια αναζήτησης.

Η ενδεχόμενη εξέλιξη της πλατφόρμας δεν χρειάζεται να γίνει μόνο πάνω στην ανάλυση των μεταγενωμικών δεδομένων. Υπάρχουν πολλοί τομείς (μετα-

μεταγραφωμική, επιγενωμική κτλ) που χρήζουν της ίδιας αναγκαιότητας αυτοματοποίησης των αναλύσεων τους, με ταυτόχρονη δυνατότητα πρόσβασης σε μέλη της επιστημονικής κοινότητας με λίγες ή και καθόλου γνώσεις βιοπληροφορικής. Ήδη από το πρόγραμμα COVERALL υπάρχουν μετα-μεταγραφωμικά δεδομένα τα οποία θα αξιοποιηθούν τόσο για την διερεύνηση νέων αλληλουχιών συσχετισμένων με την έκθεση σε CO_2 όσο και για την ενσωμάτωση νέων εργαλείων και υλοποίηση νέων αυτοματοποιημένων γραμμών εργασιών.

Έχουνε γίνει επίσης τα πρώτα βήματα για την αποδοτικότερη χρήση της πλατφόρμας ANASTASIA από τους χρήστες ανεξάρτητα με τη διαδικτυακή της εφαρμογή. Έχουν ήδη αναπτυχθεί τα πρώτα σενάρια σε γλώσσα BASH τα οποία επιτρέπουν τη λήψη και εγκατάσταση της πλατφόρμας, μαζί με τα απαραίτητα εργαλεία και βάσεις δεδομένων, σε τοπικό διακομιστή και έχουν ήδη ελεγχθεί με επιτυχία σε περιβάλλον Linux-Ubuntu και Linux-CentOS. Επιπλέον εξετάζεται η ανακατασκευή των σεναρίων εγκατάστασης μέσω της χρήση του περιβάλλοντος Docker [154], έναν υπολογιστικό φλοιό που προσφέρει μεγάλη προσαρμοστικότητα σε κάθε υπολογιστικό σύστημα, καθώς παρακάμπτει οποιαδήποτε εξειδικευμένη απαίτηση σε προαπαιτούμενα προγράμματα, πακέτα ακόμα και λογισμικό λειτουργίας.

Είναι σαφές ότι παρ' όλες τις δυνατότητές της, η πλατφόρμα ANASTASIA έχει πολλές δυνατότητες εξέλιξης τόσο με την ανάπτυξη νέων εργαλείων όσο και με την βελτιστοποίηση των ήδη υπαρχόντων αλγορίθμων της. Η συνεχής αξιοποίηση της για την ανάλυση νέων δειγμάτων (είτε μεταγενωμικών, είτε από άλλους τομείς) θα επιτρέψει επίσης τον προσδιορισμό των βέλτιστων παραμέτρων ανά περίπτωση και συνεπώς την δημιουργία νέων πιο εξειδικευμένων γραμμών εργασιών, που θα προσαρμόζονται στις ανάγκες κάθε μορφής δεδομένων. Αυτό σε συνδυασμό με το ειδικά σχεδιασμένο γραφικό περιβάλλον χρήστη, καθιστά την πλατφόρμα ένα εξαιρετικά εύχρηστο εργαλείο στη διάθεση κάθε ερευνητή, με η χωρίς εξειδικευμένες γνώσεις βιοπληροφορικής, που ασχολείται με την βιοτεχνολογία, τη μεταγενωμική ή και με τη συνθετική βιολογία.

Βιβλιογραφία

1. Avery, O.T., C.M. Macleod, and M. McCarty, *Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii.* J Exp Med, 1944. **79**(2): p. 137-58.
2. Thomas, T., J. Gilbert, and F. Meyer, *Metagenomics - a guide from sampling to data analysis.* Microb Inform Exp, 2012. **2**(1): p. 3.
3. Biddle, J.F., et al., *Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment.* Proc Natl Acad Sci U S A, 2008. **105**(30): p. 10583-8.
4. Jiang, B., et al., *Comparison of metagenomic samples using sequence signatures.* BMC Genomics, 2012. **13**: p. 730.
5. Willner, D., R.V. Thurber, and F. Rohwer, *Metagenomic signatures of 86 microbial and viral metagenomes.* Environ Microbiol, 2009. **11**(7): p. 1752-66.
6. Finkel, O.M., et al., *Metagenomic Signatures of Bacterial Adaptation to Life in the Phyllosphere of a Salt-Secreting Desert Tree.* Appl Environ Microbiol, 2016. **82**(9): p. 2854-61.
7. Popovic, A., et al., *Metagenomics as a Tool for Enzyme Discovery: Hydrolytic Enzymes from Marine-Related Metagenomes.* Adv Exp Med Biol, 2015. **883**: p. 1-20.
8. Chistoserdovai, L., *Functional metagenomics: recent advances and future challenges.* Biotechnol Genet Eng Rev, 2010. **26**: p. 335-52.
9. Singh, J., et al., *Metagenomics: Concept, methodology, ecological inference and recent advances.* Biotechnol J, 2009. **4**(4): p. 480-94.
10. Lorenz, P. and J. Eck, *Metagenomics and industrial applications.* Nat Rev Microbiol, 2005. **3**(6): p. 510-6.
11. Martin, R., et al., *The role of metagenomics in understanding the human microbiome in health and disease.* Virulence, 2014. **5**(3): p. 413-23.
12. Miller, R.R., et al., *Metagenomics for pathogen detection in public health.* Genome Med, 2013. **5**(9): p. 81.
13. Puehler, A. and J. Kalinowski, *Microbial genomics and metagenomics in human health and disease.* J Biotechnol, 2017.
14. Zhou, X., et al., *The next-generation sequencing technology: a technology review and future perspective.* Sci China Life Sci, 2010. **53**(1): p. 44-57.
15. van Nimwegen, K.J., et al., *Is the $1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing.* Clin Chem, 2016. **62**(11): p. 1458-1464.
16. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-8.
17. Alberts B, J.A., Lewis J, et al. , *Molecular Biology of the Cell*. 4th ed. 2002: New York: Garland Science.
18. Holland, J., et al., *Rapid evolution of RNA genomes.* Science, 1982. **215**(4540): p. 1577-85.
19. Kunin, V., et al., *A bioinformatician's guide to metagenomics.* Microbiol Mol Biol Rev, 2008. **72**(4): p. 557-78, Table of Contents.
20. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
21. Sanger, F. and A.R. Coulson, *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.* J Mol Biol, 1975. **94**(3): p. 441-8.
22. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA.* Proc Natl Acad Sci U S A, 1977. **74**(2): p. 560-4.

23. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nat Methods, 2008. **5**(1): p. 16-8.
24. Tipu, H.N. and A. Shabbir, *Evolution of DNA sequencing.* J Coll Physicians Surg Pak, 2015. **25**(3): p. 210-5.
25. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
26. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
27. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
28. Pandey, V., R.C. Nutter, and E. Prediger, *Applied Biosystems SOLiD™ System: Ligation-Based Sequencing*, in *Next Generation Genome Sequencing*. 2008, Wiley-VCH Verlag GmbH & Co. KGaA. p. 29-42.
29. Gupta, P.K., *Single-molecule DNA sequencing technologies for future genomics research.* Trends Biotechnol, 2008. **26**(11): p. 602-11.
30. Check Hayden, E., *Genome sequencing: the third generation.* Nature, 2009. **457**(7231): p. 768-9.
31. Au, K.F., et al., *Improving PacBio long read accuracy by short read alignment.* PLoS One, 2012. **7**(10): p. e46679.
32. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Hum Mol Genet, 2010. **19**(R2): p. R227-40.
33. Bleidorn, C., *Third generation sequencing: technology and its potential impact on evolutionary biodiversity research.* Systematics and Biodiversity, 2016. **14**(1): p. 1-8.
34. Schmieder, R. and R. Edwards, *Quality control and preprocessing of metagenomic datasets.* Bioinformatics, 2011. **27**(6): p. 863-4.
35. Brown, J., M. Pirrung, and L.A. McCue, *FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool.* Bioinformatics, 2017.
36. Anvar, S.Y., et al., *Determining the quality and complexity of next-generation sequencing data without a reference genome.* Genome Biol, 2014. **15**(12): p. 555.
37. Benjamini, Y. and T.P. Speed, *Summarizing and correcting the GC content bias in high-throughput sequencing.* Nucleic Acids Res, 2012. **40**(10): p. e72.
38. Dohm, J.C., et al., *Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.* Nucleic Acids Res, 2008. **36**(16): p. e105.
39. *ASCII Codes Table*. Available from: http://ascii.cl/.
40. *Babraham Bioinformatics*. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
41. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
42. *Java*. Available from: https://java.com/.
43. *FASTX-Toolkit FASTQ/A short-reads pre-processing tools*. Available from: http://hannonlab.cshl.edu/fastx_toolkit/index.html.
44. Giardine, B., et al., *Galaxy: a platform for interactive large-scale genome analysis.* Genome Res, 2005. **15**(10): p. 1451-5.
45. Achtman, M. and M. Wagner, *Microbial diversity and the genetic nature of microbial species.* Nat Rev Microbiol, 2008. **6**(6): p. 431-40.
46. Rappe, M.S. and S.J. Giovannoni, *The uncultured microbial majority.* Annu Rev Microbiol, 2003. **57**: p. 369-94.
47. Droge, J. and A.C. McHardy, *Taxonomic binning of metagenome samples generated by next-generation sequencing technologies.* Brief Bioinform, 2012. **13**(6): p. 646-55.

48. Huse, S.M., et al., *Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing.* PLoS Genet, 2008. **4**(11): p. e1000255.
49. Jeewon, R. and K.D. Hyde, *Detection and Diversity of Fungi from Environmental Samples: Traditional Versus Molecular Approaches*, in *Advanced Techniques in Soil Microbiology*, A. Varma and R. Oelmüller, Editors. 2007, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 1-15.
50. Quast, C., et al., *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.* Nucleic Acids Res, 2013. **41**(Database issue): p. D590-6.
51. Gonzalez, J.M., et al., *Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities.* PLoS One, 2012. **7**(1): p. e29973.
52. Forney, L.J., X. Zhou, and C.J. Brown, *Molecular microbial ecology: land of the one-eyed king.* Curr Opin Microbiol, 2004. **7**(3): p. 210-20.
53. Coordinators, N.R., *Database Resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2017. **45**(D1): p. D12-D17.
54. Shakya, M., et al., *Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities.* Environ Microbiol, 2013. **15**(6): p. 1882-99.
55. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
56. Menzel, P., et al., *Comparative Metagenomics of Eight Geographically Remote Terrestrial Hot Springs.* Microb Ecol, 2015. **70**(2): p. 411-24.
57. Huson, D.H. and N. Weber, *Microbial community analysis using MEGAN.* Methods Enzymol, 2013. **531**: p. 465-85.
58. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.
59. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
60. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic Acids Res, 2000. **28**(1): p. 33-6.
61. Overbeek, R., et al., *The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).* Nucleic Acids Res, 2014. **42**(Database issue): p. D206-14.
62. Stein, L.D., *The case for cloud computing in genome informatics.* Genome Biol, 2010. **11**(5): p. 207.
63. Warren, R.L., et al., *Assembling millions of short DNA sequences using SSAKE.* Bioinformatics, 2007. **23**(4): p. 500-1.
64. Jeck, W.R., et al., *Extending assembly of short DNA sequences to handle error.* Bioinformatics, 2007. **23**(21): p. 2942-4.
65. Dohm, J.C., et al., *SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing.* Genome Res, 2007. **17**(11): p. 1697-706.
66. Curtis, S.A., *The classification of greedy algorithms.* Science of Computer Programming, 2003. **49**(1): p. 125-157.
67. Miller, J.R., S. Koren, and G. Sutton, *Assembly algorithms for next-generation sequencing data.* Genomics, 2010. **95**(6): p. 315-27.
68. Li, Z., et al., *Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph.* Brief Funct Genomics, 2012. **11**(1): p. 25-37.
69. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.
70. Myers, E.W., et al., *A whole-genome assembly of Drosophila.* Science, 2000. **287**(5461): p. 2196-204.

71. Namiki, T., et al., *MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads.* Nucleic Acids Res, 2012. **40**(20): p. e155.
72. Olson, S.A., *EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite.* Brief Bioinform, 2002. **3**(1): p. 87-91.
73. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology.* Nucleic Acids Res, 2003. **31**(1): p. 28-33.
74. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC Bioinformatics, 2010. **11**: p. 119.
75. Noguchi, H., J. Park, and T. Takagi, *MetaGene: prokaryotic gene finding from environmental genome shotgun sequences.* Nucleic Acids Res, 2006. **34**(19): p. 5623-30.
76. Zhu, W., A. Lomsadze, and M. Borodovsky, *Ab initio gene identification in metagenomic sequences.* Nucleic Acids Res, 2010. **38**(12): p. e132.
77. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
78. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nat Methods, 2015. **12**(1): p. 59-60.
79. Gevers, D., et al., *Gene duplication and biased functional retention of paralogs in bacterial genomes.* Trends Microbiol, 2004. **12**(4): p. 148-54.
80. The UniProt, C., *UniProt: the universal protein knowledgebase.* Nucleic Acids Res, 2017. **45**(D1): p. D158-D169.
81. Eddy, S.R., *Multiple alignment using hidden Markov models.* Proc Int Conf Intell Syst Mol Biol, 1995. **3**: p. 114-20.
82. Soding, J., A. Biegert, and A.N. Lupas, *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W244-8.
83. Kumar, N. and J. Skolnick, *EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes.* Bioinformatics, 2012. **28**(20): p. 2687-8.
84. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D211-22.
85. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.
86. *Cooper GM. The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates; 2000. The Central Role of Enzymes as Biological Catalysts.*; Available from: https://www.ncbi.nlm.nih.gov/books/NBK9921/.
87. Ogata, H., et al., *Computation with the KEGG pathway database.* Biosystems, 1998. **47**(1-2): p. 119-28.
88. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**: p. 41.
89. Krieger, C.J., et al., *MetaCyc: a multiorganism database of metabolic pathways and enzymes.* Nucleic Acids Res, 2004. **32**(Database issue): p. D438-42.
90. Hanson, N.W., et al., *Metabolic pathways for the whole community.* BMC Genomics, 2014. **15**: p. 619.
91. Ye, Y. and T.G. Doak, *A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes.* PLoS Comput Biol, 2009. **5**(8): p. e1000465.
92. Karp, P.D., S. Paley, and P. Romero, *The Pathway Tools software.* Bioinformatics, 2002. **18 Suppl 1**: p. S225-32.
93. Finkbeiner, S.R., et al., *Metagenomic analysis of human diarrhea: viral detection and discovery.* PLoS Pathog, 2008. **4**(2): p. e1000011.

94. Ambrose, H.E., et al., *Diagnostic strategy used to establish etiologies of encephalitis in a prospective cohort of patients in England.* J Clin Microbiol, 2011. **49**(10): p. 3576-83.
95. Mahony, J.B., et al., *Multiplex PCR tests sentinel the appearance of pandemic influenza viruses including H1N1 swine influenza.* J Clin Virol, 2009. **45**(3): p. 200-2.
96. Rohayem, J., et al., *A simple and rapid single-step multiplex RT-PCR to detect Norovirus, Astrovirus and Adenovirus in clinical stool samples.* J Virol Methods, 2004. **118**(1): p. 49-59.
97. Segerman, B., *The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories.* Front Cell Infect Microbiol, 2012. **2**: p. 116.
98. Hilton, S.K., et al., *Metataxonomic and Metagenomic Approaches vs. Culture-Based Techniques for Clinical Pathology.* Front Microbiol, 2016. **7**: p. 484.
99. Sender, R., S. Fuchs, and R. Milo, *Revised Estimates for the Number of Human and Bacteria Cells in the Body.* PLoS Biol, 2016. **14**(8): p. e1002533.
100. Weinstock, G.M., *Genomic approaches to studying the human microbiota.* Nature, 2012. **489**(7415): p. 250-6.
101. Fukuda, S., et al., *Acetate-producing bifidobacteria protect the host from enteropathogenic infection via carbohydrate transporters.* Gut Microbes, 2012. **3**(5): p. 449-54.
102. Maynard, C.L., et al., *Reciprocal interactions of the intestinal microbiota and immune system.* Nature, 2012. **489**(7415): p. 231-41.
103. Tremaroli, V. and F. Backhed, *Functional interactions between the gut microbiota and host metabolism.* Nature, 2012. **489**(7415): p. 242-9.
104. Wang, W.L., et al., *Application of metagenomics in the human gut microbiome.* World J Gastroenterol, 2015. **21**(3): p. 803-14.
105. Jovel, J., et al., *Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics.* Front Microbiol, 2016. **7**: p. 459.
106. Qin, J., et al., *A human gut microbial gene catalogue established by metagenomic sequencing.* Nature, 2010. **464**(7285): p. 59-65.
107. Bashiardes, S., G. Zilberman-Schapira, and E. Elinav, *Use of Metatranscriptomics in Microbiome Research.* Bioinform Biol Insights, 2016. **10**: p. 19-25.
108. Abubucker, S., et al., *Metabolic reconstruction for metagenomic data and its application to the human microbiome.* PLoS Comput Biol, 2012. **8**(6): p. e1002358.
109. Ortiz, M., et al., *Making a living while starving in the dark: metagenomic insights into the energy dynamics of a carbonate cave.* ISME J, 2014. **8**(2): p. 478-91.
110. Lopez-Lopez, O., M.E. Cerdan, and M.I. Gonzalez Siso, *New extremophilic lipases and esterases from metagenomics.* Curr Protein Pept Sci, 2014. **15**(5): p. 445-55.
111. Nigam, P.S., *Microbial enzymes with special characteristics for biotechnological applications.* Biomolecules, 2013. **3**(3): p. 597-611.
112. Hasan, F., A.A. Shah, and A. Hameed, *Industrial applications of microbial lipases.* Enzyme and Microbial Technology, 2006. **39**(2): p. 235-251.
113. Panesar, P.S., S. Kumari, and R. Panesar, *Potential Applications of Immobilized beta-Galactosidase in Food Processing Industries.* Enzyme Res, 2010. **2010**: p. 473137.
114. Ashie, I.N.A., T.L. Sorensen, and P.M. Nielsen, *Effects of Papain and a Microbial Enzyme on Meat Proteins and Beef Tenderness.* Journal of Food Science, 2002. **67**(6): p. 2138-2142.
115. Hamada, S., et al., *Improvements in the qualities of gluten-free bread after using a protease obtained from Aspergillus oryzae.* Journal of Cereal Science, 2013. **57**(1): p. 91-97.
116. Panda, T. and B.S. Gowrishankar, *Production and applications of esterases.* Applied Microbiology and Biotechnology, 2005. **67**(2): p. 160-169.

117. Coughlan, L.M., et al., *Biotechnological applications of functional metagenomics in the food and pharmaceutical industries.* Front Microbiol, 2015. **6**: p. 672.
118. Streit, W.R. and P. Entcheva, *Biotin in microbes, the genes involved in its biosynthesis, its biochemical role and perspectives for biotechnological production.* Appl Microbiol Biotechnol, 2003. **61**(1): p. 21-31.
119. Jiang, C.J., et al., *Characterization of a novel serine protease inhibitor gene from a marine metagenome.* Mar Drugs, 2011. **9**(9): p. 1487-501.
120. Chang, F.Y. and S.F. Brady, *Discovery of indolotryptoline antiproliferative agents by homology-guided metagenomic screening.* Proc Natl Acad Sci U S A, 2013. **110**(7): p. 2478-83.
121. Piel, J., *A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14002-7.
122. Schirmer, A., et al., *Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge Discodermia dissoluta.* Appl Environ Microbiol, 2005. **71**(8): p. 4840-9.
123. Gomes, E.S., V. Schuch, and E.G. de Macedo Lemos, *Biotechnology of polyketides: new breath of life for the novel antibiotic genetic pathways discovery through metagenomics.* Braz J Microbiol, 2013. **44**(4): p. 1007-34.
124. Barnard, D., et al., *Extremophiles in biofuel synthesis.* Environ Technol, 2010. **31**(8-9): p. 871-88.
125. Gray, K.A., L. Zhao, and M. Emptage, *Bioethanol.* Curr Opin Chem Biol, 2006. **10**(2): p. 141-6.
126. Xu, Y., et al., *A novel enzymatic route for biodiesel production from renewable oils in a solvent-free medium.* Biotechnol Lett, 2003. **25**(15): p. 1239-41.
127. Xing, M.N., X.Z. Zhang, and H. Huang, *Application of metagenomic techniques in mining enzymes from microbial communities for biofuel synthesis.* Biotechnol Adv, 2012. **30**(4): p. 920-9.
128. Buijs, N.A., V. Siewers, and J. Nielsen, *Advanced biofuel production by the yeast Saccharomyces cerevisiae.* Curr Opin Chem Biol, 2013. **17**(3): p. 480-8.
129. Smith, M.B., et al., *Natural bacterial communities serve as quantitative geochemical biosensors.* MBio, 2015. **6**(3): p. e00326-15.
130. Håvelsrud, O.E., et al., *Metagenomics in CO2 Monitoring.* Energy Procedia, 2013. **37**: p. 4215-4233.
131. Techtmann, S.M. and T.C. Hazen, *Metagenomic applications in environmental monitoring and bioremediation.* J Ind Microbiol Biotechnol, 2016. **43**(10): p. 1345-54.
132. Sims, D., et al., *Sequencing depth and coverage: key considerations in genomic analyses.* Nat Rev Genet, 2014. **15**(2): p. 121-32.
133. Muir, P., et al., *The real cost of sequencing: scaling computation to keep pace with data generation.* Genome Biol, 2016. **17**: p. 53.
134. Ladoukakis, E., F.N. Kolisis, and A.A. Chatziioannou, *Integrative workflows for metagenomic analysis.* Front Cell Dev Biol, 2014. **2**: p. 70.
135. Treangen, T.J., et al., *MetAMOS: a modular and open source metagenomic assembly and analysis pipeline.* Genome Biol, 2013. **14**(1): p. R2.
136. Angiuoli, S.V., et al., *CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing.* BMC Bioinformatics, 2011. **12**: p. 356.
137. Arumugam, M., et al., *SmashCommunity: a metagenomic annotation and analysis tool.* Bioinformatics, 2010. **26**(23): p. 2977-8.
138. *COVERALL*. Available from: http://www.mn.uio.no/cees/english/research/projects/143935/.

139. Zhang, W., et al., *A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies.* PLoS One, 2011. **6**(3): p. e17915.
140. Li, D., et al., *MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices.* Methods, 2016. **102**: p. 3-11.
141. *g++*. Available from: http://www.cprogramming.com/g++.html.
142. Fu, L., et al., *CD-HIT: accelerated for clustering the next-generation sequencing data.* Bioinformatics, 2012. **28**(23): p. 3150-2.
143. Finn, R.D., J. Clements, and S.R. Eddy, *HMMER web server: interactive sequence similarity searching.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W29-37.
144. Seemann, T., *Prokka: rapid prokaryotic genome annotation.* Bioinformatics, 2014. **30**(14): p. 2068-9.
145. *Python Software Foundation*. Available from: https://www.python.org/.
146. *GNU Bash*. Available from: https://www.gnu.org/software/bash/.
147. Blackford, J., et al., *Detection and impacts of leakage from sub-seafloor deep geological carbon dioxide storage.* Nature Clim. Change, 2014. **4**(11): p. 1011-1016.
148. Zarafeta, D., et al., *Discovery and Characterization of a Thermostable and Highly Halotolerant GH5 Cellulase from an Icelandic Hot Spring Isolate.* PLoS One, 2016. **11**(1): p. e0146454.
149. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2017. **45**(D1): p. D37-D42.
150. Zarafeta, D., et al., *Metagenomic mining for thermostable esterolytic enzymes uncovers a new family of bacterial esterases.* Sci Rep, 2016. **6**: p. 38886.
151. Wences, A.H. and M.C. Schatz, *Metassembler: merging and optimizing de novo genome assemblies.* Genome Biol, 2015. **16**: p. 207.
152. Majoros, W.H., M. Pertea, and S.L. Salzberg, *TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.* Bioinformatics, 2004. **20**(16): p. 2878-9.
153. Fabregat, A., et al., *Reactome pathway analysis: a high-performance in-memory approach.* BMC Bioinformatics, 2017. **18**(1): p. 142.
154. O'Connor, B.D., et al., *The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows.* F1000Res, 2017. **6**: p. 52.

Παράρτημα Ι: Σενάρια παραμετροποίησης για το Galaxy/ANASTASIA

galaxy/static/welcome.html

```
<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml">
	<head>
	<title>jQuery Radmenu Example</title>
	<style>

	body {
		margin:0 auto;
		text-align:center;
		position:relative;
		width:950px;
	}

	.container {
		margin-top:100px;
		position:relative;
		top:50px;
	}

	.my_class {
		font-size:1.5em;
		color:#abc;
		background-color:#def;
		-moz-border-radius:30px;
		width:30px;
		height:30px;
		-webkit-border-radius:30px;
	}

#nested_container {
	position:relative;
	left:170px;
	top: 25px;
	width:100px;
	height:100px;
	text-align:center;
}
.list .item .sublist,
.sublist {
	display:none;
}
.my_class_main,
.my_class_sub {
	font-size:1.5em;
	color:#123;
	background-color:transparent;
	-moz-border-radius:30px;
	width:30px;
	height:30px;
	-webkit-border-radius:30px;
}

.my_class_sub {
	background-color:transparent;
	color:#123;
}

#nested_container .radial_div_item {
	background-color:none;
	height:30px;
	padding:10px;
```

```
        color:#234;
        -moz-border-radius:10px;
        -webkit-border-radius:10px;
        cursor:pointer;
}
#nested_container .radial_div_item.active {
        background-color:transparent; color:white;
        padding: 0px;
        -moz-border-radius:40px;
        z-index:100;
}

        </style>
    <script type="text/javascript" src="feimg/jquery.min.js"></script>
    <script type='text/javascript' src='feimg/jQuery.radmenu.js'></script>
    <script type="text/javascript">
          jQuery(document).ready(function(){
        jQuery("#nested_container").radmenu({
                listClass: 'list', // the list class to look within for items
                itemClass: 'item', // the items
                radius: 120, // radius in pixels
                animSpeed:400, // animation speed in millis
                centerX: -184, // the center x axis offset -184
                centerY: 20, // the center y axis offset 20
                selectEvent: "click", // the select event (click)
                onSelect: nestedSelection,
                angleOffset: -30 // in degrees
        });
});

var jlastSelected;
var build = true;

// the onSelect method for the main radmenu
function nestedSelection(jselected){
// make sure that there are elements
        if(jselected.length){
                // only do this if a parent menu item has been selected
                if(jlastSelected) {
                        if(jlastSelected.is(":visible")){  //  if  theres  an  active
radmenu item, i.e it's visible
                                 // check to see if its the parent of the nested menu
item
                                if(jlastSelected.parents(".radial_div").length){
                                        jlastSelected.radmenu("hide");   //   hide   the
radmenu
                                        build = false; // we don't want to build a new
radmenu
                                }
                        }else { // this is a new selection
                                jlastSelected = null;
                                build  =  true;  //  we  want  to  build  the  sublist  if
avail.
                        }
                }
                if(build){
                        // build the sublist radmenu and show it
                        jselected.radmenu(selectedRadmenuOptions).radmenu("show");
                        // add some effects for the menu selections
                        jselected.siblings().fadeTo("slow", 0.32);
                        jselected.fadeTo("slow", 1);
                        // store the last item so for nested elements
                        jlastSelected = jselected;
                }else{
                        // toggle  :  show  the  parent  radmenu  (i.e.  'reset'  the  main
radmenu)
```

```
                        jselected.parents(".radial_div").radmenu("show");
                }
        }
};

// sublist menu options
var selectedRadmenuOptions = {
        listClass: "sublist",
        itemClass: "subitem",
        select: "click",
        rotate:false,
        onSelect: function(jselected){
                //update the box in the top right corner with the selected item's
HTML
                jQuery("#nested_selection").html(jselected.html());
                jselected.siblings().fadeTo("slow", 0.32);
                jselected.fadeTo("slow", 1);
        },
        radius: 150,
        centerX: 10,
        centerY: -20,
        angleOffset: 0
};
        </script>

  </head>
  <body
onload='jQuery("#nested_container").radmenu("show");jQuery("#nested_container").rad
menu("scale","1.5")' >
<style>
body
{
background:url(feimg/back2.gif);
background-size:30% 240%;
background-repeat:no-repeat;
}
</style>

 <h2><font size="3" face="helvetica" color="black"><font size="4" face="helvetica"
color="red">A</font>utomated         <font         size="4"         face="helvetica"
color="red">N</font>ucleotide         <font         size="4"         face="helvetica"
color="red">A</font>minoacid         <font         size="4"         face="helvetica"
color="red">S</font>equences         <font         size="4"         face="helvetica"
color="red">T</font>ranslational       pl<font       size="4"       face="helvetica"
color="red">A</font>tform     <br     />for     <font     size="4"     face="helvetica"
color="red">S</font>ystemic          <font          size="4"          face="helvetica"
color="red">I</font>nterpretation     and     <font     size="4"     face="helvetica"
color="red">A</font>nalysis</font></h2>
<br />
<div align="center" class="container">


        <br/>
                <div id='nested_container'>
                        <ul class='list'>
                                <li class='item'>
                                <div                          class='my_class_main'><img
src="feimg/quality.gif"/></div>
                                <ul class="sublist">
                                        <li                          class="subitem"><div
class='my_class_sub'><a      href="../tool_runner?tool_id=cshl_fastq_quality_filter"
target="_parent"><img           src="feimg/quality_filter.gif"           width=250%
height=320%/></div></li>
                                        <li                          class="subitem"><div
class='my_class_sub'><a    href="../tool_runner?tool_id=cshl_fastx_artifacts_filter"
```

```
target="_parent"><img               src="feimg/seq_art.gif"               width=250%
height=320%/></a></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a          href="../tool_runner?tool_id=cshl_fastx_clipper_ng"
target="_parent"><img src="feimg/clip.gif" width=250% height=320%/></a></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a           href="../tool_runner?tool_id=cshl_fastx_collapser"
target="_parent"><img              src="feimg/collapse.gif"              width=250%
height=320%/></a></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a             href="../tool_runner?tool_id=cshl_fastx_trimmer"
target="_parent"><img src="feimg/trim.gif" width=250% height=320%/></a></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a         href="../tool_runner?tool_id=fasta_filter_by_length"
target="_parent"><img            src="feimg/filter_length.gif"            width=250%
height=320%/></a></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a  href="../tool_runner?tool_id=fastqc"  target="_parent"><img
src="feimg/fastqc.gif" width=250% height=300%/></a></div></li>
                                </ul>
                        </li>

                                <li class='item'>
                                <div                          class='my_class_main'><img
src="feimg/Assembly.gif"/></div>
                                <ul class="sublist">
                                    <li                          class="subitem"><div
class='my_class_sub'><a  href="../tool_runner?tool_id=velvetg"  target="_parent"><img
src="feimg/velvetg.gif" width=250% height=300%/></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a  href="../tool_runner?tool_id=velveth"  target="_parent"><img
src="feimg/velveth.gif" width=250% height=300%/></a></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a  href="../tool_runner?tool_id=megahit"  target="_parent"><img
src="feimg/megahit.gif" width=250% height=300%/></a></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a               href="../tool_runner?tool_id=velvetoptimiser"
target="_parent"><img           src="feimg/velvetoptimiser.gif"           width=250%
height=290%/></a></div></li>
                                </ul>
                        </li>



                        <li class='item'>
                                <div                          class='my_class_main'><img
src="feimg/protein.gif"/></div>
                                <ul class="sublist">
                                    <li                          class="subitem"><div
class='my_class_sub'><a            href="../tool_runner?tool_id=ncbi_blastn_wrapper"
target="_parent"><img              src="feimg/blastn.gif"               width=250%
height=300%/></a></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a            href="../tool_runner?tool_id=ncbi_blastp_wrapper"
target="_parent"><img              src="feimg/blastp.gif"               width=250%
height=300%/></a></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a   href="../tool_runner?tool_id=hmmer"   target="_parent"><img
src="feimg/hmmer2.gif" width=250% height=300%/></a></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a                       href="../tool_runner?tool_id=eficaz25"
target="_parent"><img              src="feimg/eficaz.gif"               width=250%
height=300%/></a></div></li>
                                        <li                  class="subitem"><div
class='my_class_sub'><a                    href="../tool_runner?tool_id=rapidminer"
```

```
target="_parent"><img             src="feimg/classifier.gif"             width=250%
height=300%/></a></div></li>
                                <li                          class="subitem"><div
class='my_class_sub'><a                        href="../tool_runner?tool_id=prodigal"
target="_parent"><img src="feimg/ORF.gif" width=250% height=300%/></a></div></li>
                                <li                          class="subitem"><div
class='my_class_sub'><a  href="../tool_runner?tool_id=prokka"  target="_parent"><img
src="feimg/prokka.gif" width=250% height=320%/></a></div></li>
                                </ul>
                        </li>

                        <li class='item'>
                                <div                         class='my_class_main'><img
src="feimg/RunWork.gif"/></div>
                                <ul class="sublist">
                                <li                          class="subitem"><div
class='my_class_sub'><a                 href="../tool_runner?tool_id=megan_analysis"
target="_parent"><img               src="feimg/megan_t.gif"               width=250%
height=320%/></a></div></li>
                                <li                          class="subitem"><div
class='my_class_sub'><a                href="../tool_runner?tool_id=megan_analysisf"
target="_parent"><img               src="feimg/megan_f.gif"               width=250%
height=350%/></a></div></li>
                                <li                          class="subitem"><div
class='my_class_sub'><a                       href="../tool_runner?tool_id=rma_builder"
target="_parent"><img              src="feimg/megan_rma.gif"              width=250%
height=300%/></a></div></li>
                                <li                          class="subitem"><div
class='my_class_sub'><a                       href="../tool_runner?tool_id=cluster_otu"
target="_parent"><img              src="feimg/otu_clust.gif"              width=250%
height=300%/></a></div></li>
                                <li                          class="subitem"><div
class='my_class_sub'><a                       href="../tool_runner?tool_id=usearch_der"
target="_parent"><img            src="feimg/dereplication.gif"            width=250%
height=300%/></a></div></li>
                                <li                          class="subitem"><div
class='my_class_sub'><a  href="../tool_runner?tool_id=minpath"  target="_parent"><img
src="feimg/minpath.gif" width=250% height=320%/></a></div></li>
                                </ul>
                                </li>

                                <li class='item'>
                                <div                         class='my_class_main'><img
src="feimg/EditWork.gif"/></div>
                                <ul class="sublist">
                                <li                          class="subitem"><div
class='my_class_sub'><a             href="../u/makis/w/copy-of-starting-from-reads-1"
target="_parent"><img           src="feimg/start_taxonomic.gif"           width=260%
height=390%/></a></div></li>
                                <li                          class="subitem"><div
class='my_class_sub'><a               href="../workflow/copy-of-starting-from-reads"
target="_parent"><img             src="feimg/start_contigs.gif"             width=265%
height=350%/></a></div></li>
                                <li                          class="subitem"><div
class='my_class_sub'><a  href="../u/makis/w/start-from-reads"  target="_parent"><img
src="feimg/start_reads.gif" width=270% height=350%/></a></div></li>
                                </ul>
                        </li>

                                <li class='item'>
                                <div                         class='my_class_main'><img
src="feimg/GetData.gif"/></div>
                                <ul class="sublist">
                                <li                          class="subitem"><div
class='my_class_sub'><a                 href="../tool_runner?tool_id=sequence_parser"
target="_parent"><img src="feimg/seq_db.gif" width=250% height=320%/></div></li>
```

```
                                    <li                          class="subitem"><div
class='my_class_sub'><a                  href="../tool_runner?tool_id=blast_parser"
target="_parent"><img src="feimg/blast_db.gif" width=250% height=320%/></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a                  href="../tool_runner?tool_id=hmmer_parser"
target="_parent"><img              src="feimg/hmmer_db.gif"              width=250%
height=320%/></a></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a           href="../tool_runner?tool_id=knowledgebase_parser"
target="_parent"><img src="feimg/know_db.gif" width=250% height=320%/></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a         href="../library/index"         target="_parent"><img
src="feimg/Libraries.gif" width=250% height=300%/></a></div></li>
                                    <li                          class="subitem"><div
class='my_class_sub'><a  href="../tool_runner?tool_id=upload1"  target="_parent"><img
src="feimg/Upload.gif" width=250% height=300%/></div></li>
                                </ul>
                                </li>
                            </ul>
                    </div>
             </div>

<a   href="http://hotzyme.com"><img   style="position:fixed;   bottom:0%;   right:0%;
width:25%; height:10%" src="feimg/X-logo.gif"/></a>


<a             href="http://motherbox.chemeng.ntua.gr/anastasia_knowledgebase/"><img
style="position:fixed;     bottom:12%;     right:0%;     width:24%;     height:12%"
src="feimg/logo_mysql.png"/></a>
  </body>

</html>
```

---

galaxy/config/tool_conf.xml

```
<?xml version='1.0' encoding='utf-8'?>
<toolbox>
  <section name="ANASTASIA" id="anastasia">
    <label text="Data management" id="get_dat" />
    <tool file="data_source/upload.xml"/>
    <tool file="myTools/sequence_parser/sequence_parser.xml" />
    <tool file="myTools/hmmer_parser/hmmer_parser.xml" />
    <tool file="myTools/blast_parser/blast_parser.xml" />
    <tool file="myTools/knowledgebase_parser/knowledgebase_parser.xml"/>
    <label text="File manipulation" id="data_manp" />
    <tool file="myTools/shuffle_fasta/fasta_joiner.xml" />
    <tool file="myTools/fasta_names/fasta_names.xml" />
    <tool file="fasta_tools/fasta_concatenate_by_species.xml" />
    <tool file="fasta_tools/fasta_to_tabular.xml" />
    <tool file="fastx_toolkit/fasta_formatter.xml" />
    <tool file="fastx_toolkit/fasta_nucleotide_changer.xml" />
    <tool file="fasta_tools/fasta_compute_length.xml" />
    <tool file="fastx_toolkit/fastx_reverse_complement.xml" />
    <tool file="fastx_toolkit/fastx_renamer.xml" />
    <tool file="fastx_toolkit/fastx_barcode_splitter.xml" />
    <tool file="myTools/cdhit/cd_hit_est.xml" />
    <tool file="myTools/cdhit/cd_hit.xml" />
    <tool file="myTools/transeq/transeq.xml" />
    <tool file="fastx_toolkit/fastq_to_fasta.xml" />
    <tool file="fasta_tools/tabular_to_fasta.xml" />
    <label text="Quality control" id="qc_an" />
    <tool file="myTools/fastqc/rgFastQC.xml" />
    <tool file="fastx_toolkit/fastq_quality_filter.xml" />
    <tool file="fastx_toolkit/fastx_artifacts_filter.xml" />
```

```
    <tool file="fastx_toolkit/fastx_clipper.xml" />
    <tool file="fastx_toolkit/fastx_collapser.xml" />
    <tool file="fastx_toolkit/fastx_trimmer.xml" />
    <tool file="fasta_tools/fasta_filter_by_length.xml" />
    <tool file="fastx_toolkit/fastq_quality_converter.xml" />
    <tool file="fastx_toolkit/fastx_quality_statistics.xml" />
    <tool file="fastx_toolkit/fastq_quality_boxplot.xml" />
    <tool file="fastx_toolkit/fastx_nucleotides_distribution.xml" />
    <tool file="myTools/count_mapped/count_mapped.xml" />
    <tool file="myTools/example_gc_content/example.xml" />
    <label text="Sequencing assembly" id="dnv_as" />
    <tool file="myTools/velvetoptimiser/velvetoptimiser.xml" />
    <tool file="myTools/Velvet/velveth.xml" />
    <tool file="myTools/Velvet/velvetg.xml" />
    <tool file="myTools/megahit/megahit.xml" />
    <tool file="myTools/gene_coverage/gene_coverage.xml" />
    <tool file="myTools/bowtie2/bowtie2_wrapper.xml" />
    <label text="Sequence annotation" id="seq_annot" />
    <tool file="myTools/prodigal/prodigal.xml" />
    <tool file="myTools/getorf/getorf.xml" />
    <tool file="myTools/prokka/prokka.xml" />
    <tool file="myTools/hmmer/hmmer.xml" />
    <tool file="myTools/eficaz/eficaz.xml" />
    <tool file="myTools/hydrolase_classifier/rapidminer.xml" />
    <tool file="ncbi_blast_plus/ncbi_blastn_wrapper.xml" />
    <tool file="ncbi_blast_plus/ncbi_blastp_wrapper.xml" />
    <tool file="ncbi_blast_plus/ncbi_blastx_wrapper.xml" />
    <tool file="ncbi_blast_plus/ncbi_tblastn_wrapper.xml" />
    <tool file="ncbi_blast_plus/ncbi_tblastx_wrapper.xml" />
    <tool file="ncbi_blast_plus/blastxml_to_tabular.xml" />
     <label text="Taxonomic and functional analysis" id="tax_fun_an" />
    <tool file="myTools/megan_parser/megan_analysis.xml" />
    <tool file="myTools/megan_parser/megan_analysisf.xml" />
    <tool file="myTools/megan_parser/rma_builder.xml" />
    <tool file="myTools/minpath/minpath.xml" />
    <tool file="myTools/usearch_cluster_otus/otu_clustering/cluster_otu.xml" />
    <tool    file="myTools/usearch_cluster_otus/usearch_dereplicate/usearch_der.xml"
/>
  </section>
</toolbox>
```

---

apache_mod_anastasia.conf (αρχείο με αλλαγές που εφαρμόζονται στον Apache web server)

```
<VirtualHost *:80>
        ServerAdmin anastasia@email.com

        DocumentRoot /var/www/html
        <Directory />
                Options FollowSymLinks
                AllowOverride None
        </Directory>
        <Directory /var/www/html>
                Options Indexes FollowSymLinks MultiViews
                AllowOverride None
                Order allow,deny
                allow from all
        </Directory>

        ScriptAlias /cgi-bin/ /usr/lib/cgi-bin/
        <Directory "/usr/lib/cgi-bin">
                AllowOverride None
                Options +ExecCGI -MultiViews +SymLinksIfOwnerMatch
```

```
                Order allow,deny
                Allow from all
        </Directory>


        #ANASTASIA application
        RewriteEngine on
        RewriteRule ^/anastasia_web$ /anastasia_web/ [R]
        RewriteRule                                  ^/anastasia_web/static/style/(.*)
/home/anastasia_web/galaxy/static/june_2007_style/blue/$1 [L]
        RewriteRule                                ^/anastasia_web/static/scripts/(.*)
/home/anastasia_web/galaxy/static/scripts/packed/$1 [L]
        RewriteRule  ^/anastasia_web/static/(.*)  /home/anastasia_web/galaxy/static/$1
[L]
        RewriteRule                                          ^/anastasia_web/favicon.ico
/home/anastasia_web/galaxy/static/favicon.ico [L]
        RewriteRule                                           ^/anastasia_web/robots.txt
/home/anastasia_web/galaxy/static/robots.txt [L]
        RewriteRule ^/anastasia_web(.*) http://localhost:8092$1 [P]

        #ANASTASIA knowledgebase
        WSGIScriptAlias /anastasia_web /home/anastasia_web/web2py/wsgihandler.py
        WSGIDaemonProcess anastasia_web user=apache group=apache
        <Directory /home/anastasia_web/web2py>
                WSGIProcessGroup anastasia_web
                AllowOverride None
                Order Allow,Deny
                Deny from all
                <Files wsgihandler.py>
                        Allow from all
                </Files>
        </Directory>
        AliasMatch                                  ^/anastasia_web/([^/]+)/static/(.*)
/home/anastasia_web/web2py/applications/$1/static/$2
        <Directory /home/anastasia_web/web2py/applications/*/static/>
                Order Allow,Deny
                Allow from all
        </Directory>
        <Location /anastasia_web/admin>
                Deny from all
        </Location>
        <LocationMatch ^/anastasia_web/([^/]+)/appadmin>
                Deny from all
        </LocationMatch>




        ErrorLog /home/anastasia_web/error.log

        # Possible values include: debug, info, notice, warn, error, crit,
        # alert, emerg.
        LogLevel warn

        CustomLog /home/anastasia_web/access.log combined

</VirtualHost>
-----------------------------------------------------------------------------------------------
```

rma_builder.xml

```
<tool id="rma_builder" name="RMA builder">
 <description> Produce a MEGAN .rma file from a BLAST output</description>
  <command interpreter="python">blastsub.py $query $blastfile $meganfile
   #if $type_option.opts_selector == "advanced":
    ${type_option.useseed}
    ${type_option.usecog}
    ${type_option.usekegg}
    ${type_option.maxmatches}
   #else:
    true
    true
    true
    100
   #end if
   2> $output2
  </command>
  <inputs>
   <param format="fasta" name="query" type="data" label="Fasta file of reads"/>
   <param   format="txt"   name="blastfile"   type="data"   label="BLAST   file   with
results"/>
   <conditional name="type_option">
   <param name="opts_selector" type="select" label="Advanced options">
        <option value="standard">Standard options</option>
        <option value="advanced">Advanced options</option>
       </param>
       <when value="standard" />
       <when value="advanced">
       <param  name="maxmatches"  type="integer"  value="100"  label="Max  number  of
matches per read"/>
       <param name="useseed" type="select" label="Enable SEED analysis">
        <option value="true" selected="yes">Yes</option>
        <option value="false">No</option>
       </param>
       <param name="usecog" type="select" label="Enable COG analysis">
        <option value="true" selected="yes">Yes</option>
        <option value="false">No</option>
       </param>
       <param name="usekegg" type="select" label="Enable KEGG analysis">
        <option value="true" selected="yes">Yes</option>
        <option value="false" >No</option>
       </param>
       </when>
       </conditional>
  </inputs>
  <outputs>
  <data format="data" name="meganfile"/>
  <data label="Additional output and errors" name="output2" format="txt"/>
  </outputs>
  <help>

**RMA builder Overview**

The RMA builder tool uses the results from a BLAST analysis of a dataset containing
nucleotide sequences andd produces a MEGAN rma file

-----

**Input formats**

This tool uses as input a BLASTX or BLASTN tab delimited file and the corresponding
nucleotide sequences FASTA file
```

```
-----

**Outputs**

The tool provides a txt file carrying all the information of a MEGAN rma file. In
order to open this file using MEGAN, add the extension .rma

-----

.. class:: infomark

**Info:**
SEED, KEGG and COG analysis are enabled by default

.. class:: infomark

**Info:**
If functional analysis is desired in a following step, BLASTX search is the only
appropriate option


  </help>
 </tool>
```

---

megan_analysis.xml

```
 <tool id="megan_analysis" name="MEGAN t-analysis">
 <description>for taxonomic classification</description>
  <command  interpreter="python">megansub.py  $meganfile  $rank  $output1  $output2
$output3 $pick_format
  #if $more_options.opts_selector == "advanced":
   ${more_options.minscore}
   ${more_options.maxexpected}
   ${more_options.toppercent}
   ${more_options.minsupport}
   ${more_options.mincomplexity}
   ${more_options.use_minimal_coverage_heuristic}
   ${more_options.use_identity_filter}
   ${more_options.minsupportpercent}
   ${more_options.lcapercent}
   ${more_options.paired}
  #else:
  50.0
  0.01
  10.0
  1
  0.0
  false
  false
  0.0
  100.0
  false
  #end if
  </command>
  <inputs>
   <param format="data" name="meganfile" type="data" label="MEGAN .rma file"/>
   <param name="rank" type="select" label="Select rank">
    <option value="all">All</option>
    <option value="SuperKingdom">SuperKingdom</option>
    <option value="Kingdom">Kingdom</option>
    <option value="Phylum">Phylum</option>
    <option value="Class">Class</option>
    <option value="Order">Order</option>
    <option value="Family">Family</option>
```

```
    <option value="Varietas">Varietas</option>
    <option value="Genus">Genus</option>
    <option value="Species_group">Species_group</option>
    <option value="Subspecies">Subspecies</option>
    <option value="Species" selected="yes">Species</option>
   </param>
   <param name="pick_format" type="select" label="Select image format">
    <option value="eps">eps</option>
    <option value="svg">svg</option>
    <option value="gif">gif</option>
    <option value="png">png</option>
    <option value="jpg">jpg,jpeg</option>
    <option value="pdf">pdf</option>
    <option value="bmp">bmp</option>
   </param>
   <conditional name="more_options">
      <param name="opts_selector" type="select" label="Advanced options">
         <option value="standard">Standard options</option>
         <option value="advanced">Advanced options</option>
      </param>
      <when value="standard" />
      <when value="advanced">
       <param name="minscore" type="float" value="50.0" label=" Minscore - Set a
minimum threshold for the bit score of hits"/>
       <param name="maxexpected" type="float" value="0.01" label="Max Expected -
Maximum threshold of E-value of hits"/>
       <param name="toppercent" type="float" value="10.0" label="Top percent"/>
       <param name="minsupport" type="integer" value="1" label=" Minsupport - Set a
threshold for the minimum support that a taxon
requires"/>
       <param name="mincomplexity" type="float" value="0.0" label="Mincomplexity -
Identify low complexity
reads"/>
       <param name="minsupportpercent" type="float" value="0.0" label="Min Support
Percent"/>
       <param name="lcapercent" type="float" value="100.0" label="LCA Percent"/>
       <param    name="use_minimal_coverage_heuristic"    type="select"    label="Use
Minimal Coverage Heuristic">
        <option value="false" selected="yes">No</option>
        <option value="true">Yes</option>
       </param>
       <param name="paired" type="select" label="Enable paired analysis">
        <option value="false" selected="yes">No</option>
        <option value="true">Yes</option>
       </param>
       <param name="use_identity_filter" type="select" label="Use Identity Filter
(enable only for 16S rRNA sequences)">
        <option value="false" selected="yes">No</option>
        <option value="true">Yes</option>
       </param>
      </when>
   </conditional>
  </inputs>
  <outputs>
  <data format="jpg" name="output1" label="Chart image"/>
  <data format="txt" name="output2" label=" Chart data txt "/>
  <data format="data" name="output3" label=" tar.gz file "/>
  </outputs>
  <help>

**MEGAN t-analysis Overview**

The MEGAN t-analysis tool is attempting to provide a first insight into the
taxonomic content of a metagenomic sample through taxonomic binning

-----
```

Input formats

This tool uses MEGAN rma files as input. This is the case of the RMA builder tool output

---

Outputs

The tool provides a txt file of the selected taxonomic nodes and the number of reads assigned or summarized to them, a similar chart image file and a tar.gz file containing an image of the taxonomic tree and plain text files of all selected nodes along with the specific sequences of the reads assigned or summarized to them

---

.. class:: warningmark

WARNING: Rank options and default values are provided by MEGAN v.5.5.3 official manual.

.. class:: warningmark

WARNING: In case of a specific rank search, summarized reads are included in the results. Option "All" presents the assigned reads of each node, on a fully collapsed tree

.. class:: warningmark

WARNING: For the downloaded tar.gz file please add the .tar.gz extension

.. class:: infomark

Info:
The Min Support item can be used to set a threshold for the minimum support that a taxon requires, that is, the number of reads that must be assigned to it so that it appears in the result.Any read that is assigned to a taxon that does not have the required support is pushed up the taxonomy until a node is found that has sufficient support.

.. class:: infomark

Info:
The Min Support Percent item is used to set a threshold for the minimum support that a taxon requires, as a percentage of assigned reads. This feature is turned off by setting the value to 0. If a value greater than 0 (and at most 100) is given, then the program will set the Min Support threshold appropriately.

.. class:: infomark

Info:
The Min Score item can be used to set a minimum threshold for the bit score of hits.Any hit in the input data that scores less than the given threshold is ignored.

.. class:: infomark

Info:
The Max Expected item can be used to set a maximum threshold for the expected value of hits.Any hit in the input data whose E-value exceeds this value is ignored.

.. class:: infomark

Info:

```
The Top Percentage item can be used to set a threshold for the maximum percentage
by which the score of a hit may fall below the best score achieved for a given
read.Any hit that falls below this threshold is discarded.The Min Complexity item
can be used to identify low complexity reads.These are placed on a special Low
Complexity node.To turn this filter off, set the value to 0. A value of 0.3 catches
most low complexity short reads.

.. class:: infomark

**Info:**
The Paired Reads item can be used to turn paired-read awareness of MEGAN on and
off.In paired-read mode, MEGAN utilities read-pairing information to enhance the
taxonomic assignment of reads.

.. class:: infomark

**Info:**
The Use 16S Percent Identity Filter item can be used to turn on an additional
filter for assigning reads to a specific taxonomic level.When this is active, the
percent identity of a match must exceed the given value of percent identity to be
assigned at the given rank:
Species 99%, Genus 97%, Family 95%, Order 90%, Class 85%, Phylum 80%.This should
only be used when analyzing 16S rRNA sequences.

.. class:: infomark

**Info:**
Minimal Coverage Heuristic,use a minimum set of taxa that cover all reads.Increases
the specificity of the LCA algorithm.

.. class:: infomark

**Info:**
The LCA Percent item is used to set the percent of matches that the LCA of a read
must cover, in the range 50-100. When a value of less than 100 is specified then
the LCA of a fixed percent is used.

  </help>

 </tool>
```

---

megan_analysisf.xml

```
<tool id="megan_analysisf" name="MEGAN f-analysis">
 <description>for functional annotation</description>
  <command   interpreter="python">functionalsub.py   $meganfile   $output1   $output2
$output3 $output4
  #if $more_options.opts_selector == "advanced":
   ${more_options.minscore}
   ${more_options.maxexpected}
   ${more_options.toppercent}
   ${more_options.minsupport}
   ${more_options.mincomplexity}
   ${more_options.use_minimal_coverage_heuristic}
   ${more_options.use_identity_filter}
   ${more_options.minsupportpercent}
   ${more_options.lcapercent}
   ${more_options.paired}
  #else:
  50.0
  0.01
  10.0
  1
```

```
  0.0
  false
  false
  0.0
  100.0
  false
  #end if
  </command>
  <inputs>
   <param format="data" name="meganfile" type="data" label="MEGAN .rma file"/>
   <conditional name="more_options">
      <param name="opts_selector" type="select" label="Advanced options">
         <option value="standard">Standard options</option>
         <option value="advanced">Advanced options</option>
      </param>
      <when value="standard" />
      <when value="advanced">
       <param name="minscore" type="float" value="5.0" label=" Minscore - Set a
minimum threshold for the bit score of hits"/>
       <param name="maxexpected" type="float" value="0.01" label="Max Expected -
Maximum threshold of E-value of hits"/>
       <param name="toppercent" type="float" value="10.0" label="Top percent"/>
       <param name="minsupport" type="integer" value="1" label=" Minsupport - Set a
threshold for the minimum support that a taxon
requires"/>
       <param name="mincomplexity" type="float" value="0.44" label="Mincomplexity -
Identify low complexity
reads"/>
       <param name="minsupportpercent" type="float" value="0.1" label="Min Support
Percent"/>
       <param name="lcapercent" type="float" value="100.0" label="LCA Percent"/>
       <param    name="use_minimal_coverage_heuristic"    type="select"    label="Use
Minimal Coverage Heuristic">
        <option value="false" selected="yes">No</option>
        <option value="true">Yes</option>
       </param>
       <param name="paired" type="select" label="Enable paired analysis">
        <option value="false" selected="yes">No</option>
        <option value="true">Yes</option>
       </param>
       <param name="use_identity_filter" type="select" label="Use Identity Filter
(enable only for 16S rRNA sequences)">
        <option value="false">No</option>
        <option value="true">Yes</option>
       </param>
      </when>
   </conditional>
  </inputs>
  <outputs>
  <data format="txt" name="output1" label=" tar.gz file (functional analysis)"/>
  <data format="jpg" name="output2" label=" SEED Chart image "/>
  <data format="jpg" name="output3" label="COG Chart image "/>
  <data format="jpg" name="output4" label="KEGG Chart image "/>
  </outputs>
  <help>

**MEGAN f-analysis Overview**

The MEGAN f-analysis tool is providing information considering the functional
profile of a metagenomic sample. Three different approaches are enabled for such a
functional annotation, as MEGAN activates SEED, COG and KEGG classifications

-----

**Input formats**
```

This tool uses MEGAN rma files as input. This is the case of the RMA builder tool output

---

Outputs

The tool provides three different chart images , one for each type of classification, visualizing the number of reads assigned to the top nodes of each classification. Moreover, a tar.gz is created containing three different folders (SEED, COG, KEGG). In every one of them , MEGAN creates a DSV file with paths and readnames assigned to them and multiple files of each leave node of a fully uncollapsed tree with the actual sequences of the reads assigned to them

---

.. class:: warningmark

WARNING: Default values are provided by MEGAN v.5.5.3 official manual.

.. class:: warningmark

WARNING: Charts are visualizations of the number of assigned reads at the top nodes of each classification tree

.. class:: warningmark

WARNING: For the downloaded tar.gz file please add the .tar.gz extension

.. class:: infomark

Info:
The Min Support item can be used to set a threshold for the minimum support that a taxon requires, that is, the number of reads that must be assigned to it so that it appears in the result.Any read that is assigned to a taxon that does not have the required support is pushed up the taxonomy until a node is found that has sufficient support.

.. class:: infomark

Info:
The Min Support Percent item is used to set a threshold for the minimum support that a taxon requires, as a percentage of assigned reads. This feature is turned off by setting the value to 0. If a value greater than 0 (and at most 100) is given, then the program will set the Min Support threshold appropriately.

.. class:: infomark

Info:
The Min Score item can be used to set a minimum threshold for the bit score of hits.Any hit in the input data that scores less than the given threshold is ignored.

.. class:: infomark

Info:
The Max Expected item can be used to set a maximum threshold for the expected value of hits.Any hit in the input data whose E-value exceeds this value is ignored.

.. class:: infomark

Info:
The Top Percentage item can be used to set a threshold for the maximum percentage by which the score of a hit may fall below the best score achieved for a given read.Any hit that falls below this threshold is discarded.The Min Complexity item can be used to identify low complexity reads.These are placed on a special Low

```
Complexity node.To turn this filter off, set the value to 0. A value of 0.3 catches
most low complexity short reads.

.. class:: infomark

**Info:**
The Paired Reads item can be used to turn paired-read awareness of MEGAN on and
off.In paired-read mode, MEGAN utilities read-pairing information to enhance the
taxonomic assignment of reads.

.. class:: infomark

**Info:**
The Use 16S Percent Identity Filter item can be used to turn on an additional
filter for assigning reads to a specific taxonomic level.When this is active, the
percent identity of a match must exceed the given value of percent identity to be
assigned at the given rank:
Species 99%, Genus 97%, Family 95%, Order 90%, Class 85%, Phylum 80%.This should
only be used when analyzing 16S rRNA sequences.

.. class:: infomark

**Info:**
Minimal Coverage Heuristic,use a minimum set of taxa that cover all reads.Increases
the specificity of the LCA algorithm.

.. class:: infomark

**Info:**
The LCA Percent item is used to set the percent of matches that the LCA of a read
must cover, in the range 50-100. When a value of less than 100 is specified then
the LCA of a fixed percent is used.

  </help>

 </tool>
```

---

blast_parser.xml

```
<tool id="blast_parser" name="BLAST results database generator" version="1.0.0">
       <description>Import BLAST results into a MySQL database</description>
  <command interpreter="python">blast_parser.py '$blastpfile' '$orfile' '$sqlout'
/home/anastasia_dev/galaxy/tools/myTools/blastp_parser/database_datafile '$orphans'
'$blastorfs' 2>'$output2'</command>
  <inputs>
    <param name="blastpfile" type="data" format="tabular" label="Ouput file of
BLAST analysis" />
    <param name="orfile" type="data" format="fasta" label="FASTA file with
nucleotide/protein sequences" />

  </inputs>
  <outputs>
    <data name="sqlout" label="MySQL dump of the results" format="txt" />
    <data name="orphans" label="List of sequences without BLAST hits"
format="fasta" />
    <data name="blastorfs" label="List of sequences having BLAST hits"
format="fasta" />
       <data name="output2" label="Additional output and errors" format="txt" />
  </outputs>
    <tests>
      <test>
        <param name="blastn_out" value="blastn.out" ftype="fasta" />
        <output name="logfile" file="logfile.log"/>
```

```
      </test>
    </tests>
    <help>

This tool imports the results of a BLAST analysis into a MySQL database and
produces the corresponding dumpfile. The format of the results needs to be in tab-
delimited format as
taken from the NCBI BLAST+ Galaxy tool for 12 columns (BLAST output parameter -
outfmt 6) or for 25 columns (BLAST output parameter -outfmt "6 std sallseqid score
nident positive gaps ppos qframe sframe qseq sseq qlen slen salltitles")
It works for BLASTn, BLASTp, BLASTx, tBLASTn and tBLASTx

  </help>

</tool>
```

---

hmmer_parser.xml

```
<tool id="hmmer_parser" name="HMMER database generator" version="1.0.0">
       <description>Import HMMER analysis results into MySQL database</description>
  <command  interpreter="/usr/bin/python">hmmer_parser.py  '$hmmer_out'  '$logfile'
/home/anastasia_dev/galaxy/tools/myTools/hmmer_parser/database_datafile
2>'$output2'</command>
  <inputs>
    <param  name="hmmer_out"  type="data"  format="txt"  label="Ouput  file  of  Hmmer
analysis" />
  </inputs>
  <outputs>
    <data name="logfile" format="txt" label="HMMER output" />
       <data name="output2" label="Additional output and errors" format="txt" />

  </outputs>
    <tests>
      <test>
       <param name="hmmer_out" value="hmmer.out" ftype="txt" />
       <output name="logfile" file="logfile.log"/>
      </test>
    </tests>
    <help>

This tool imports the results of a Hmmer analysis into a MySQL database and returns
the appropriate dumpfile

  </help>

</tool>
```

---

count_mapped.xml

```
<tool id="count_mapped" name="Count mapped reads" version="0.0.1">
       <description>Tool  that  displays  the  number  of  mapped  or  unmapped
reads</description>
  <command  interpreter="python">count_mapped.py  $bam_file  $mapped  $count_file  2>
$output2</command>
  <inputs>
    <param name="bam_file" type="data" label="BAM file from mapping analysis (i.e.
Bowtie2)" />
    <param  name="mapped"  type="select"  label="Select  which  kind  of  reads  you  want
to count" >
              <option value="mapped" selected="yes">Mapped reads</option>
              <option value="unmapped">Unmapped reads</option>
              </param>
```

```
  </inputs>
  <outputs>
       <data  name="count_file"  label="Number  of  mapped  or  unmapped  reads"
format="txt" />
       <data name="output2" label="Additional output and errors" format="txt" />
  </outputs>
    <tests>
      <test>
       <param name="orf_out" value="orf.fasta" ftype="fasta" />
       <output name="out_sql" file="out.sql"/>
      </test>
    </tests>
    <help>

    </help>

</tool>
```

---

eficaz.xml

```
<tool id="eficaz25" name="Protein function prediction - EFICAz">
  <description>EFICAz software for protein function prediction</description>
  <command interpreter="/usr/bin/python">eficaz.py $input $output > /tmp/eficaz.log
2>$output2</command>
  <inputs>
    <param format="fasta" name="input" type="data" label="Fasta file with protein
sequences"/>

  </inputs>
  <outputs>
    <data format="txt" name="output" label="EC predictions"/>
       <data name="output2" label="Additional output and errors" format="txt" />

  </outputs>

  <tests>
    <test>
      <param name="input" value="fa_gc_content_input.fa"/>
      <output name="out_file1" file="fa_gc_content_output.txt"/>
    </test>
  </tests>

  <help>
EFICAz2.5 (Enzyme Function Inference by a Combined Approach) is an automatic engine
for large-scale enzyme function inference that combines predictions from six
different methods developed and optimized to achieve high prediction accuracy: (i)
recognition of functionally discriminating residues (FDRs) in enzyme families
obtained by a Conservation-controlled HMM Iterative procedure for Enzyme Family
classification (CHIEFc), (ii) pairwise sequence comparison using a family specific
Sequence Identity Threshold, (iii) recognition of FDRs in Multiple Pfam enzyme
families, (iv) recognition of multiple Prosite patterns of high specificity, (v)
SVM evaluation of CHIEFc families, and (vi) SVM evaluation of Multiple Pfam enzyme
families.
  </help>

</tool>
```

---

fasta_names.xml

```
<tool id="fasta_names" name="Replace blanks in FASTA files" version="1.0.0">
```

```
       <description>Replace blanks in description lines inside FASTA files
</description>
  <command interpreter="python">fasta_names.py '$fastafile' '$newfile' > /dev/null
2 > '$output2'</command>
  <inputs>
    <param name="fastafile" type="data" format="fasta" label="FASTA file with
blanks in description lines" />
  </inputs>
  <outputs>
    <data name="newfile" label="Fasta file without blanks" format="fasta" />
       <data name="output2" label="Additional output and errors" format="txt" />
  </outputs>
    <tests>
      <test>
       <param name="orf_in" value="orf.fasta" ftype="fasta" />
       <output name="out_fasta" file="out.fasta"/>
      </test>
    </tests>
    <help>

This tool takes a fasta file with blanks in description lines and replaces them
with underscores.
Useful in cases where you want your fasta file to be the input for BLAST.
  </help>

</tool>
```

---

getorf.xml

```
<tool id="getorf" name="Getorf" version="1.0.0">
       <description>Open reading frame analysis</description>
  <command>getorf -sequence '$contig_file' -minsize '$minsize' -maxsize '$maxsize'
-outseq '$result' -auto
  #if $advanced.select=="1":
  -table       '$advanced.table'       -find       '$advanced.find'       -methionine
'$advanced.methionine' -circular '$advanced.circular' -reverse '$advanced.reverse'
-flanking '$advanced.flanking'
  #end if</command>
  <inputs>
    <param name="contig_file" type="data" format="fasta" label="Contig fasta file"
/>
    <param name="minsize" type="text" size="40" value="50" label="Minimum size of
ORF" />
    <param name="maxsize" type="text" size="40" value="1000000" label="Maximum size
of ORF" />
       <conditional name="advanced">
              <param name="select" type="select" label="Advanced options">
           <option value="0">Use defaults</option>
           <option value="1">Define options</option>
        </param>
          <when value="0">
        </when>
        <when value="1">
                    <param name="table" type="select" label="Code to use">
                           <option value="0" selected="True">Standard</option>
                           <option value="1">Standard with alternative initiation
codons</option>
                           <option value="2">Vertebrate Mitochondrial</option>
                           <option value="3">Yeast Mitochondrial</option>
                           <option    value="4">Mold,    Protozoan,    Coelenterate
Mitochondrial and Mycoplasma-Spiroplasma</option>
                           <option value="5">Invertebrate Mitochondrial</option>
                           <option     value="6">Ciliate     Macronuclear     and
Dasycladacean</option>
```

```
                            <option value="9">Echinoderm Mitochondrial</option>
                            <option value="10">Euplotid Nuclear</option>
                            <option value="11">Bacterial</option>
                            <option value="12">Alternative Yeast Nuclear</option>
                            <option value="13">Ascidian Mitochondrial</option>
                            <option value="14">Flatworm Mitochondrial</option>
                            <option value="15">Blepharisma Macronuclear</option>
                            <option                          value="16">Chlorophycean
Mitochondrial</option>
                            <option value="21">Trematode Mitochondrial</option>
                            <option value="22">Scenedesmus obliquus</option>
                            <option                      value="23">Thraustochytrium
Mitochondrial</option>
                        </param>
                        <param   name="find"   type="select"   label="Possible   output
options">
                            <option   value="0"   selected="True">Translation   of
regions between STOP codons</option>
                            <option value="1">Translation of regions between START
and STOP codons</option>
                            <option   value="2">Nucleic   sequences   between   STOP
codons</option>
                            <option value="3">Nucleic sequences between START and
STOP codons</option>
                            <option     value="4">Nucleotides     flanking     START
codons</option>
                            <option  value="5">Nucleotides  flanking  initial  STOP
codons</option>
                            <option  value="6">Nucleotides  flanking  ending  STOP
codons</option>
                        </param>
                        <param  name="flanking"  type="text"  size="40"  value="100"
label="Number of nucleotides flanking START or STOP codons" />
                        <param name="methionine" type="select" label="START codons to
code for Methionine">
                            <option value="yes" selected="True">Yes</option>
                            <option value="no">No</option>
                        </param>
                        <param  name="circular"  type="select"  label="Is  the  sequence
circular?">
                            <option value="yes">Yes</option>
                            <option value="no" selected="True">No</option>
                        </param>
                        <param  name="reverse"  type="select"  label="Find  ORFs  in  the
reverse complement of the sequence">
                            <option value="yes" selected="True">Yes</option>
                            <option value="no">No</option>
                        </param>
                </when>
        </conditional>
    </inputs>
    <outputs>
      <data name="result" format="fasta" />
    </outputs>
      <tests>
        <test>
          <param name="contig_test_file" value="fasta_contigs.fa" ftype="fasta" />
          <output name="out_file1" file="run1.fa"/>
        </test>
      </tests>
      <help>

This program finds and outputs the sequences of open reading frames (ORFs) in
one or more nucleotide sequences. An ORF may be defined as a region of a
specified minimum size between two STOP codons, or between a START and a
STOP codon. The ORFs can be output as the nucleotide sequence or as the
```

```
protein translation. Optionally, the program will output the region around the
START codon, the first STOP codon, or the final STOP codon of an ORF. The
START and STOP codons are defined in a Genetic Code table; a suitable table
can be selected for the organism you are investigating. The output is a
sequence file containing predicted open reading frames longer than the
minimum size, which defaults to 30 bases (i.e. 10 amino acids).

  </help>

</tool>
```

---

megahit.xml

```
<tool id="megahit" name="Megahit assembler" version="0.0.1">
       <description>Use Megahit to assemble reads into contigs </description>
  <command     interpreter="python">megahit.py     $contigs     $result_directory_tar
$min_contig_lgth 4 $presets
        #if $kind_of_reads.opts_selector == "paired":
                  paired
            ${kind_of_reads.right}
            ${kind_of_reads.left}
              #elif $kind_of_reads.opts_selector == "inter":
                  inter
            ${kind_of_reads.input}
              #elif $kind_of_reads.opts_selector == "single":
                  single
            ${kind_of_reads.input2}
              #end if

              2>$output2 </command>
  <inputs>
       <conditional name="kind_of_reads">
              <param name="opts_selector" type="select" label="Type of reads">
                     <option value="paired">Paired-end reads</option>
                     <option   value="inter">Interleaved   paired-end   reads   (one
file)</option>
                     <option value="single">Single-end reads</option>
              </param>
              <when value="paired" >
               <param  name="right"  type="data"  format="fasta,fastq"  label="Right
reads"/>
               <param   name="left"   type="data"   format="fasta,fastq"   label="Left
reads"/>
              </when>
              <when value="paired-inter">
               <param      name="input"      type="data"      format="fasta,fastq"
label="Interleaved paired-end reads"/>
              </when>
              <when value="single">
               <param name="input2" type="data" format="fasta,fastq" label="Single-
end reads"/>
              </when>
    </conditional>
    <param  name="presets"  type="select"  label="Presets  parameters"  help="Choose
presets parameters for Megahit according to your data">
        <option value="meta" selected="yes">For metagenomes</option>
        <option   value="meta-sensitive"   >For   metagenomes   (more   sensitive   but
slower)</option>
        <option   value="meta-large">For   large   and   complex   metagenomes,   like
soil</option>
        <option  value="meta-bulk">Experimental,  standard  bulk  sequencing  with  >=
30x depth</option>
        <option value="single-cell">Experimental, single cell data</option>
```

```
    </param>
       <param  name="min_contig_lgth"  type="integer"  value="500"  label="Minimum
contig length"/>

  </inputs>
  <outputs>
       <data name="contigs" label="Contig file" format="fasta"/>
       <data name="result_directory_tar" label="tar.gz file of Megahit results" />
       <data  name="output2"  label="Additional  output  and  errors"  format="txt"  />

  </outputs>
    <tests>
      <test>
        <param name="orf_out" value="orf.fasta" ftype="fasta" />
        <output name="out_sql" file="out.sql"/>
      </test>
    </tests>
    <help>
Megahit is a NGS de novo assembler for assembling large and complex metagenomics
data in a time- and cost-efficient manner

  </help>

</tool>
```

---

sequence_parser.xml

```
<tool id="sequence_parser" name="Sequence database generator" version="1.0.0">
  <command   interpreter="python">sequence_parser.py   '$contig_file'   '$dumpfile'
/home/anastasia_dev/galaxy/tools/myTools/sequence_parser/database_datafile
2>/dev/null</command>
  <inputs>
    <param  name="contig_file"  type="data"  format="fasta"  label="Fasta  file  of
sequences" />
  </inputs>
  <outputs>
    <data name="dumpfile" format="txt" />
  </outputs>
  <tests>
    <test>
      <param name="contig_test_file" value="fasta_contigs.fa" ftype="fasta" />
      <output name="out_file1" file="logfile.txt"/>
    </test>
  </tests>
  <help>

This tool takes a FASTA containing nucleotide or protein sequences and returns a
MySQL database dumpfile containing all the sequence data.

  </help>

</tool>
```

---

knowledgebase_parser.xml

```
<tool    id="knowledgebase_parser"    name="ANASTASIA    knowledgebase    generator"
version="1.0.0">
       <description>Import       analysis       results       into       ANASTASIA
knowledgebase</description>
```

```
  <command    interpreter="python">knowledgebase_parser.py    '$dump_sql'    '$title'
/home/anastasia_dev/galaxy/tools/myTools/knowledgebase_parser/database_datafile
'$output' 2> '$output2'</command>
  <inputs>
    <param name="title" type="text" size="40" label="Title of your project" />
    <param name="dump_sql" type="data" format="txt" label="MySQL dump file with the
analysis results" />
  </inputs>
  <outputs>
    <data name="output" label="Resulting database info" format="html" />
    <data name="output2" label="Additional output or errors" format="txt" />
  </outputs>


    <tests>
      <test>
        <param name="blastn_out" value="blastn.out" ftype="fasta" />
        <output name="logfile" file="logfile.log"/>
      </test>
    </tests>
    <help>

This tool imports the results of an analysis into ANASTASIA knowledgebase. The
results must be in a MySQL/MariaDB dump format. Several of the tools have the
corresponding parsers to produce this kind of MySQL dumps.
The tool takes imports the data into the ANASTASIA knowledgebase and returns the ID
which is necessary for accessing the data.

  </help>

</tool>
```

---

minpath.xml

```
<tool id="minpath" name="Minpath pathway analysis" version="0.0.1">
       <description>Use Minpath to find KEGG pathways that can be reconstructed
from the annotated genes of your data</description>
  <command
interpreter="python">/home/ladoukef/programs_various/MinPath/MinPath1.2.py      -any
$ec_file    -map    $ec_to_pwy    -report    $output_file    -details    $output_details
>/dev/null</command>
  <inputs>
    <param   name="ec_file"   type="data"   format="txt"   label="EC   numbers   list"
help="tab delimited file with the first column being the gene names and the second
the putative EC number" />
    <param name="ec_to_pwy" type="select" label="Choose type of pathways" >
               <option
value="/home/anastasia_dev/tmp/ec_kegg/kegg_pathways_to_ec_name.txt"
selected="yes">KEGG pathways by name</option>
               <option
value="/home/anastasia_dev/tmp/ec_kegg/kegg_pathways_to_ec_id.txt"   >KEGG   pathways
by ID</option>
               <option
value="/home/anastasia_dev/tmp/ec_kegg/kegg_modules_to_ec_name.txt"   >KEGG   modules
by name</option>
               <option
value="/home/anastasia_dev/tmp/ec_kegg/kegg_modules_to_ec_id.txt"  >KEGG  modules  by
ID</option>
       </param>

  </inputs>
  <outputs>
       <data name="output_file" label="Results file" format="txt" />
```

```
       <data name="output_details" label="Details file" format="txt" />
  </outputs>
    <tests>
      <test>
        <param name="orf_out" value="orf.fasta" ftype="fasta" />
        <output name="out_sql" file="out.sql"/>
      </test>
    </tests>
    <help>


  </help>

</tool>
```

---

transeq.xml

```
<tool id="transeq" name="Nucleotide Sequence Translation" version="1.0.0">
       <description>Translate genomic sequences</description>
  <command>transeq -sequence '$contig_file' -outseq '$outfile' -frame $frame_value
2>/dev/null </command>
  <inputs>
    <param  name="contig_file"  type="data"  format="fasta"  label="FASTA  file  of
nucleotide sequences" />
    <param name="frame_value" size="40" type="text" value="6" label="Frame" />
  </inputs>
  <outputs>
    <data name="outfile" format="fasta" />
  </outputs>
    <tests>
      <test>
        <param name="contig_test_file" value="fasta_contigs.fa" ftype="fasta" />
        <output name="out_file1" file="run1.out"/>
      </test>
    </tests>
    <help>

EMBOSS  Transeq  translates  nucleic  acid  sequences  to  the  corresponding  peptide
sequences.

  </help>

</tool>
```

---

fasta_joiner2.xml

```
<tool id="fasta_shuffler" name="FASTA shuffler" version="1.0.0">
  <command    interpreter="perl">$script_to_run    '$input_file1'    '$input_file2'
'$shuffled_seqs'</command>
  <inputs>
       <conditional name="filetype">
              <param  name="filetype_select"  label="File  type  of  paired-end  raw
data" type='select'>
                     <option value="FASTA">FASTA</option>
                     <option value="FASTQ">FASTQ</option>
              </param>
              <when value="FASTA">
                     $script_to_run=shuffleSequences_fasta.pl
                         <param   name="input_file1"   type="data"   format="fasta"
label="Left-hand Reads" />
```

```
                              <param  name="input_file2"  type="data"  format="fasta"
label="Right-hand Reads" />
               </when>
               <when value="FASTQ">
                      $script_to_run=shuffleSequences_fastq.pl
                              <param  name="input_file1"  type="data"  format="fastq"
label="Left-hand Reads" />
                              <param  name="input_file2"  type="data"  format="fastq"
label="Right-hand Reads" />
               </when>
       </conditional>
  </inputs>
  <outputs>
    <data name="shuffled_seqs" format="input" />
  </outputs>
  <tests>
    <test>
      <param name="input_file1" value="fasta_seqs1.txt" ftype="fasta" />
      <param name="input_file2" value="fasta_seqs2.txt" ftype="fasta" />
      <output name="shuffled_seqs" file="shuffled_seqs_fasta.txt" />
    </test>
  </tests>
  <help>

This tool joins paired end FASTA reads from two separate files into a single read
in one file. The script used is from the /contrib folder of Velvet assembler.

  </help>

</tool>
```
