Ανάπτυξη και εφαρμογή μεθόδων μηχανικής μάθησης σε δεδομένα ωρίμανσης βρεφικών μικροβιωμάτων για την πρόβλεψη της πορείας τους κατα την ενηλικίωση.

Περίληψη

Η μελέτη του εντερικού μικροβιώματος έχει αναδείξει τον σημαντικό του ρόλο στη λειτουργία του εντέρου και τη συνολική υγεία ανθρώπων και ζώων. Παρόλο που ο ορισμός του υγιούς μικροβιώματος είναι δύσκολος, ειδικά για τον άνθρωπο, έχει καταγραφεί η τάση  των μικροβιακών κοινοτήτων να συγκλίνουν προς συγκεκριμένες συνθέσεις που περιγράφονται ως εντερότυποι. Οι παράγοντες που καθορίζουν την πορεία ενός μικροβιώματος από την αρχική αποίκιση του εντέρου κατά την γέννηση προς κάθε ένα από τους τρεις κύριους εντερότυπους παραμένουν άγνωστοι. Πρόσφατες μελέτες βρεφικών μικροβιωμάτων σε βάθος χρόνου από τη γέννηση και κατά τα πρώτα χρόνια, έχουν καταδείξει τη συσχέτιση του με τον τρόπο γέννησης, το θηλασμό και την χορήγηση ή όχι προβιοτικών. Προτείνουμε την συνδυαστική ανάλυση υπαρχόντων δεδομένων αλληλούχισης βρεφικών και παιδικών μικροβιωμάτων με δεδομένα ώριμων υγιών εθελοντών για την απεικόνιση και την κατανόηση  της πορείας ωρίμανσης των μικροβιωμάτων. Η ανάλυση αυτή θα καταδείξει τα πιθανά μονοπάτια που ακολουθούν αυτές οι κοινότητες και θα προτείνει μικροβιακούς δείκτες με προβλεπτική ικανότητα. Η εφαρμογή υπολογιστικών εργαλείων μηχανικής μάθησης πάνω σε σειρές μικροβιακών προφίλ διαφορετικών ηλικιών, εκτός από τις απαντήσεις που θα δώσει στο πλαίσιο της ωρίμανσης του μικροβιώματος στη βρεφική ηλικία, θα επιτρέψει την ανάπτυξη του πλαισίου εφαρμογής υπολογιστικών/στατιστικών τεχνικών μηχανικής μάθησης σε άλλα πειράματα με παρόμοια μορφή δεδομένων.

Εισαγωγή

Η αποίκηση του βρεφικού γαστρεντερικού συστήματος είναι μια δυναμική διαδικασία που επηρεάζεται απο διατροφικούς (θηλασμός, προβιοτικά) και ιατρικούς/κλινικούς παράγοντες (καισαρική ή πρόωρη γέννηση) (Koenig et al. 2011, Bäckhed et al. 2015, Bokulich et al. 2016, Dahl et al. 2018).  Αυτές οι πρώιμες μικροβιακές κοινότητες ωριμάζουν σταδιακά με την ενηλικίωση του βρέφους, όπου παραμένουν σχετικά σταθερές, για να αποδομηθούν αργότερα κατά την τρίτη ηλικία (O'Toole 2012). Παρόλο που η σύσταση του βρεφικού και παιδικού μικροβιώματος παίζει καθοριστικό ρόλο για τη διαμόρφωση του μικροβιώματος των ενηλίκων, οι διεργασίες που διέπουν αυτή την μετάβαση  δεν είναι ακόμα κατανοητές. Η κατανόηση των μηχανισμών που ελέγχουν και κατευθύνουν την εξέλιξη του μικροβιώματος μπορούν να συμβάλλουν  στην πρόβλεψη και την έγκαιρη αντιμετώπιση με κατάλληλες διατροφικές και προληπτικές αγωγές νόσων που έχουν συνδεθεί με το ανθρώπινο μικροβίωμα (Cho and Blaser 2012, Arrieta et al. 2014). Πρόσφατη μελέτη, κατέγραψε την εξέλιξη του μικροβιώματος νεογνών κατά τους πρώτους 24 μήνες της ζωής τους και συνέκρινε τις αλλαγές που πραγματοποιούνται σε αυτό με βάση παράγοντες όπως ο θηλασμός και η χορήγηση προβιοτικών (Bazanella et al. 2017). Η μελέτη του μικροβιώματος στα πρώιμα στάδια της ζωής είναι σημαντική, καθώς υπάρχουν ενδείξεις ότι η σύσταση του βρεφικού μικροβιώματος είναι σημαντική για τη μετέπειτα υγεία του ατόμου. Παράγοντες, όπως αντιβιοτικά, που χορηγούνται κατά την βρεφική ανάπτυξη και επηρεάζουν τη σύσταση του μικροβιώματος δραστικά, έχει δειχθεί ότι συνδέονται με παθογένειες όπως παχυσαρκία σε μεγαλύτερες ηλικίες, διαβήτη τύπου 1 και τύπου 2 καθώς και κοιλιοκάκη  (Cox and Martin 2015; Angelakis and Raoult 2018). Πιθανότατα η συσχέτιση του μικροβιώματος με παθογένειες πραγματοποιείται μηχανιστικά μέσω της αλληλεπίδρασης του μικροβιώματος με το ανοσοποιητικό σύστημα (Francino 2014) αν και οι λεπτομέρειες της αλληλεπίδρασης αυτής είναι ακόμη άγνωστες. Ακόμη και αν οι μηχανιστικές αλληλεπιδράσεις του μικροβιώματος με άλλους παράγοντες που επιδρούν στην υγεία μας είναι εν πολλοίς άγνωστες, η γνώση του τρόπου μεταβολής του μικροβιώματος στα πρώιμα στάδια ζωής του ανθρώπου και πώς αυτό σχετίζεται με τη σύσταση του μικροβιώματος στους ενήλικες είναι σημαντική, επειδή επιτρέπει την ορθολογική παρέμβαση και την αποκατάσταση του φυσιολογικού μικροβιώματος σε περιπτώσεις απόκλισης.

Στην παρούσα πρόταση προτείνουμε τη συγκέντρωση και μελέτη συσσωρευμένων δεδομένων μικροβιακών προφίλ από υπάρχουσες έρευνες του βρεφικού μικροβιώματος (όπως η μελέτη της Bazanella et al. 2017), σε συνδυασμό με αδημοσίευτα δεδομένα μεγάλου αριθμού (περίπου 300) μικροβιακών προφίλ ομάδας ελέγχου (υγιείς ενήλικες) που έχει στην κατοχή της η ερευνητική ομάδα, υπερβαίνοντας στο σύνολο τα 500 δειγμάτα. Συγκεκριμένα, θα χαρακτηρίσουμε κάθε στάδιο της πορείας ωρίμανσης των βρεφικών μικροβιωμάτων και θα μελετήσουμε το κατά πόσο η πορεία αυτή μπορεί να χρησιμοποιηθεί, αφενός για την ταξινόμηση των διαφορετικών μικροβιακών προφίλ και αφ’ ετέρου την προβλεπτική της ικανότητα για την ανάπτυξή τους.

Θεωρητικό πλαίσιο και ερευνητικα ερωτήματα

Το ανθρώπινο μικροβίωμα του εντέρου χαρακτηρίζεται από έντονη μεταβλητότητα κατά τη βρεφική ηλικία του ατόμου εώς την ηλικία των τριών ετών περίπου (Koenig et al. 2011, Bäckhed et al. 2015, Bokulich et al. 2016, Dahl et al. 2018). Επιπλέον, τόσο η σύσταση του βρεφικού μικροβιώματος όσο και η σύσταση του “ενήλικου μικροβιώματος”, μετά την ηλικία των τριών ετών, όπου οι μεταβολές του δεν είναι τόσο ραγδαίες, συνδέονται με παθήσεις όπως ο διαβήτης τύπου 1, η κοιλιοκάκη καθώς και με αρκετά αυτοάνοσα νοσήματα (Cox and Martin 2015; Angelakis and Raoult 2018). Τα ερωτήματα που θα θέσουμε στην ερευνητική πρόταση είναι:

* Μπορούμε να προβλέψουμε το “ενήλικο μικροβίωμα” με βάση τα διάφορα στάδια του βρεφικού;
* Ποιο στάδιο βρεφικού μικροβιώματος είναι κρισιμότερο για τον καθορισμό του ενήλικου μικροβιώματος;
* Υπάρχουν βακτηριακά είδη που καθορίζουν/χαρακτηρίζουν την κάθε φάση του βρεφικού μικροβιώματος;
* Ομαδοποιούνται τα βρέφη με βάση το μικροβίωμα τους;

Η παρούσα πρόταση θα συμβάλλει στην επέκταση των σημερινών γνώσεων, παρέχοντας απαντήσεις στις προαναφερθείσες ερωτήσεις.  Πιο αναλυτικά, με την παρούσα ερευνητική πρόταση θα αξιοποιήσουμε ποιοτικά και ποσοτικά δεδομένα βρεφικών και παιδικών μικροβιώματων ώστε να:

* Χαρακτηρίσουμε τα πρώιμα στάδια στην εξέλιξη του ανθρώπινου μικροβιώματος και να παράγουμε ταξινομητές με τη δυνατότητα κατάταξης εντερικών μικροβιακών προφίλ από άτομα άγνωστης ηλικίας στις κατάλληλες χρονικές περιόδους.
* Κατανοήσουμε αν τα πρώτα χρόνια της ζωής ενός βρέφους είναι καθοριστικά για την διαμόρφωση του μικροβιώματος που το άτομο θα φέρει σε μετέπειτα στάδια ενηλικίωσης.
* Εξακριβώσουμε αν στο σύνολο του φυσιολογικού μικροβιακού φορτίου που το άτομο φέρει ως βρέφος, υπάρχουν μικρόβια *δείκτες* των οποίων τόσο η παρουσία όσο και η ποσότητα θα παίξουν σημαντικό ρόλο και θα έχουν κατά συνέπεια προβλεπτική ικανότητα για τη διαμόρφωση του μικροβιώματος σε μετέπειτα ηλικιακά στάδια.
* Παράγουμε ευέλικτα και εύχρηστα υπολογιστικά εργαλεία για την απάντηση των παραπάνω ερωτημάτων καθώς και για την χρήση σε ανάλογα ερωτήματα σε μικροβιακά μεταγονιδιακά δεδομένα βασισμένα σε χρονικές σειρές.

Τα αναμενόμενα αποτελέσματα της ερευνητικής πρότασης είναι τα εξής:

1. **Ομαδοποίηση των μικροβιωμάτων κατά την βρεφική ηλικία.** Σε κάθε χρονική στιγμή δειγματοληψίας κατά την διάρκεια της βρεφικής ηλικίας τα δείγματα θα ομαδοποιούνται και ο βέλτιστος αριθμός ομάδων θα προσδιορίζεται. Το αποτέλεσμα αυτό είναι σημαντικό επειδή θα μας επιτρέψει να γνωρίσουμε αν κατά τη διάρκεια της βρεφικής ηλικίας τα μικροβιώματα ομαδοποιούνται με τρόπο παρόμοιο όπως κατά την ωρίμανση. Επίσης με την ομαδοποίηση θα είναι δυνατή η σύγκριση των μικροβιωμάτων σε διαφορετικές χρονικές περιόδους της βρεφικής ηλικίας.
2. **Προβλεπτικό μοντέλο της μικροβιακής σύστασης για κάθε χρονική στιγμή με χρήση Μαρκοβιανών μεθόδων.** Με τη χρήση του συνόλου των μικροβιωματικών δειγμάτων θα αναπτύξουμε μια μεθοδολογία βασιζόμενη στη θεωρία των Μαρκοβιανών αλυσίδων όπου κάθε μελλοντική κατάσταση θα περιγράφεται με βάση τη μικροβιωματική κατάσταση του παρόντος. Θα εξετάσουμε έτσι αν η διαδοχή του μικροβιώματος συνάδει με την Μαρκοβιανή αρχή ή αν πιο πολύπλοκες σχέσεις μικροβιωμάτων προκύπτουν.
3. **Πρόβλεψη ώριμου μικροβιώματος από βρεφικά μικροβιώματα με χρήση μεθόδων μηχανικής μάθησης.** Μέθοδοι μηχανικής μάθησης θα χρησιμοποιηθούν ώστε να προβλεφθεί η τελική κατάσταση του ωρίμου μικροβιώματος από βρεφικό μικροβίωμα. Το αποτέλεσμα της μελέτης αυτής θα είναι ένας ταξινομητής (classifier) ο οποίος με συγκεκριμένα μικροβιωματικά προφίλ που έχουν αποκτηθεί κατά την βρεφική ηλικία ως είσοδο, θα προβλέπει την κατάσταση του ώριμου μικροβιώματος.
4. **Εξαγωγή του πυρήνα των χαρακτηριστικών μικροβίων (core microbial OTUs) σε κάθε χρονική στιγμή κατά τη βρεφική ηλικία.** Το χαρακτηριστικό μικροβίωμα κάθε χρονικής στιγμής στη βρεφική ηλικία που φέρεται από το σύνολο των ατόμων (core microbiome) θα εξαχθεί, εφόσον υπάρχει. Η ύπαρξη και η σύστασή του θα μας βοηθήσει να κατανοήσουμε τη δυναμική του ανθρώπινου μικροβιώματος κατά την πορεία του προς την ωρίμανση.
5. **Εξαγωγή των πληροφοριακών βρεφικών μικροβιακών ομάδων για την ταξινόμηση του ώριμου μικροβιώματος.** Χρησιμοποιώντας μεθόδους μηχανικής μάθησης θα ταξινομήσουμε το μικροβίωμα κάθε χρονικής στιγμής στους εντερότυπους. Θα υλοποιήσουμε μια διαδικασία επιλογής μεταβλητών (feature selection). Η διαδικασία αυτή είναι απαραίτητη επειδή βελτιώνει την απόδοση της μεθόδου. Συγχρόνως όμως, αποδίδει τις μικροβιωματικές ομάδες που έχουν τη βέλτιστη ταξινομική ικανότητα. Έτσι με τη διαδικασία αυτή θα αποκτήσουμε τις μικροβιωματικές ομάδες-κλειδιά για την ταξινόμηση του ώριμου μικροβιώματος.
6. **Προσδιορισμός των χρονικών στιγμών κατά τη βρεφική ηλικία που περιέχουν την μέγιστη  πληροφορία για την πρόβλεψη του ώριμου μικροβιώματος.** Η μέθοδος μηχανικής μάθησης θα περιέχει το σύνολο των χρονικών στιγμών δειγματοληψίας. Δεν είναι όμως όλες οι χρονικές στιγμές το ίδιο πληροφοριακές για την ταξινόμηση του ώριμου μικροβιώματος. Με παρόμοιο τρόπο όπως στο προηγούμενο ερώτημα (feature selection) θα αποκτήσουμε τις χρονικές στιγμές κατά τη βρεφική ηλικία που περιέχουν την μέγιστη πληροφορία για την ταξινόμηση του ώριμου ανθρώπινου μικροβιώματος.

Πέραν των αναμενόμενων αποτελεσμάτων στην αναλυση των βρεφικών μικροβιωμάτων και την εξέταση της προβλεπτικής τους ικανότητας, σημαντική θα είναι η συμβολή των παραγόμενων  υπολογιστικών εργαλείων στην μικροβιωματική έρευνα γενικότερα. Τα νέα εργαλεία που θα υλοποιήσουμε θα δώσουν τα μέσα σε ερευνητές με χρονικά δεδομένα μικροβιωμάτων να εξάγουν νέα πληροφορία που οι κλασικές μέθοδοι ανάλυσης αδυνατούν να παράξουν. Έως σήμερα, η ανάλυση του μεγαλύτερου μέρους των παραγόμενων δεδομένων εξαντλειται στην συγκριση με στατιστικές μεθόδους μεταξύ των διαφορετικών χρονικών προφίλ. Η παραγωγή εργαλείων μηχανικής μάθησης, φιλικών προς μη ειδικούς στη βιοπληροφορική, θα διευκολύνει τους βιολόγους ερευνητες να εξερευνήσουν  πλήθος ερωτημάτων που σχετιζονται με την ανάλυση μικροβιομάτων που έχουν αποκτηθεί σε διαφορετικά χρονικά σημεία.

Η παρούσα πρόταση παρουσιάζει καινοτόμα στοιχεία τόσο ως εννοιολογικό πλαίσιο όσο και ως μεθοδολογική προσέγγιση. Η καταγραφή του μικροβιώματος σε διάφορες περιόδους της ηλικίας του ατόμου και η συσχέτιση του με πιθανές ασθένειες δεν είναι νέα ιδέα (π.χ. η συσχέτιση του μικροβιώματος με ασθένειες ανασκοπείται από τους  Cho and Blaser, 2012). Όμως, η δημιουργία ενός προβλεπτικού μοντέλου του ενήλικου μικροβιώματος από διάφορες περιόδους της βρεφικής ηλικίας αποτελεί μια καινοτόμο προσέγγιση όχι μόνο ως εννοιολογικό πλαίσιο αλλά και ως προσέγγιση. Στην παρούσα μελέτη θα εξετάσουμε κατ’αρχήν αν το Μαρκοβιανό μοντέλο επαρκεί για την πρόβλεψη του μικροβιώματος. Ένα τέτοιο ενδεχόμενο θα ήταν εξαιρετικό, επειδή με μία και μόνο εκτίμηση της σύστασης του ανθρώπινου μικροβιώματος σε κάποια περίοδο της βρεφικής ηλικία θα μπορούσαμε να εκτιμήσουμε (με κάποια σχετική ακρίβεια) το μικροβίωμα σε μεταγενέστερη ηλικία. Αν όμως η Μαρκοβιανή υπόθεση δεν ικανοποιείται τότε για την εκτίμηση και πρόβλεψη του ενήλικου μικροβιώματος θα απαιτούνται πολλαπλά δείγματα από διάφορες χρονικές περιόδους κατά τη βρεφική ηλικία.

Η προσέγγιση της αρχικής ομαδοποίησης των δειγμάτων σε κάθε χρονική στιγμή  με την χρήση μεθόδων ταξινόμησης με τεχνικές μηχανικής μάθησης ( k-medians, SVMs, Random Forest, Neural Networks) για την προσαρμογή των τροχιών ωρίμανσης των μικροβιωμάτων σε μορφή μεταβάσεων μεταξύ διακριτών κατηγοριών, αποτελεί καινοτομία της πρότασης. Εκτός του ότι επιτρέπει την καλύτερη περιγραφή των φάσεων εξέλιξης του μικροβιώματος, με τον τρόπο αυτό, τα δεδομένα αποκτούν μορφή συμβατή για την χρήση Μαρκοβιανών μοντέλων. Αυτό είναι σημαντικό επειδή η χρήση των Μαρκοβιανών μοντέλων έχει εφαρμοστεί σε πολλές ερευνητικές περιοχές με σημαντικά αποτελέσματα, για παράδειγμα, στην Πληθυσμιακή Γενετική (Kimura 1971), στην Εξελικτική Βιολογία (Hobolth et al. 2007), στην Μαθηματική Βιολογία (Gibson et al. 2006) και σε πεδία πέραν της Βιολογίας, όπως στην Φυσική (Greiner et al. 2012), στην Χημεία (Kopp et al. 2012) ή στην Μουσική (McAlpine et al. 1999). Διατυπώνοντας το πρόβλημα ως Μαρκοβιανό μοντέλο θα είμαστε σε θέση να χρησιμοποιήσουμε το σύνολο της θεωρίας και των εργαλείων που έχει αναπτυχθεί για την μελέτη Μαρκοβιανών μοντέλων. Η χρήση μεθόδων μηχανικής μάθησης σε χρονικές σειρές δεν έχει χρησιμοποιηθεί μέχρι σήμερα στην ανάλυση του μικροβιώματος. Σίγουρα η εξαγωγή ενός μαθηματικού μοντέλου (για παράδειγμα με την χρήση Μαρκοβιανών Αλυσίδων) θα ήταν εξαιρετικό αποτέλεσμα επειδή θα προσφέρει μια ακριβή περιγραφή της ενήλικης μικροβιωματικής σύστασης με βάση την βρεφική ηλικία, καθώς και τους παράγοντες (μικροβιακά τάξα, OTUs) που έχουν προβλεπτική ικανότητα. Όμως, σε περίπτωση που η χρονική σχέση των μικροβιωμάτων δεν είναι Μαρκοβιανή, τότε η εξαγωγή μαθηματικού μοντέλου θα είναι αρκετά δύσκολη. Στην περίπτωση αυτή, μέθοδοι μηχανικής μάθησης μπορούν μέσω της χρήσης μεγάλου όγκου δεδομένων να ανακαλύψουν τα χαρακτηριστικά της μικροβιακής σύστασης που μπορούν να έχουν προβλεπτικό χαρακτήρα (αν και τα χαρακτηριστικά αυτά δεν μπορούν να αναχθούν σε απλές αφθονίες από μικροβιακά τάξα).

Μεθοδολογία

Συλλογή, Ενοποίηση και Ανάλυση Δεδομένων (Πακέτο Εργασιών 1)

Περιγραφή και στόχοι του πακέτου εργασιών 1

Το πρώτο πακέτο εργασιών συγκεντρώνει όλες τις εργασίες που αφορούν τη συλλογή, ενοποίηση και ανάλυση των δεδομένων από τεχνολογίες αλληλούχισης νέας γενιάς. Η ερευνητική ομάδα έχει πολυετή εμπειρία στην επεξεργασία τέτοιων δεδομένων, που συνοδεύεται με παραγωγή σχετικών υπολογιστικών εργαλείων και μεθόδων. Οι στόχοι του πακέτου 1 έχουν χωριστεί σε τέσσερις επιμέρους εργασίες που περιγράφονται αναλυτικά ακολούθως. Το πακέτο εργασίας 1 αποτελείται από τις ακόλουθες Ερευνητικές Ενότητες (ΕΕ).

Συλλογή  μεταγονιδιακών δεδομένων (ΕΕ 1.1)

Δημοσιευμένα πρωτογενή μεταγονιδιακά δεδομένα (αμπλικόνια) από έρευνες σε βρεφικά και παιδικά εντερικά μικροβιώματα θα συγκεντρωθούν από δημόσιες βάσεις όπως η Sequence Read Archive (SRA) και η European Nucleotide Archive (ENA). Ενδεικτικά, πέραν των δεδομένων από την εργασία της Bazanella και των συνεργατών της, που προσφέρουν πυκνές δειγματοληψίες σε διάστημα 2 χρόνων για 40 βρέφη, θα προστεθούν δεδομένα από τις αναγνωρισμένες εργασίες των  Bäckhed et al. και Koenig et al. (Koenig et al. 2011, Bäckhed et al. 2015) καθώς και δεδομένα από όποια άλλη εργασία με συμβατή μεθοδολογία είναι διαθέσιμη. Μεγάλη προσοχή θα δοθεί στην εξαγωγή όσων περισσότερων μεταδεδομένων υπάρχουν πέραν των αλληλουχιών (φύλο, βάρος, διατροφή κλπ) για κάθε ομάδα δειγμάτων είτε από τις σχετικές δημοσιεύσεις είτε από απευθείας επικοινωνία με τους συγγραφείς. Στα δεδομένα αυτά θα προστεθούν αδημοσίευτα πρωτογενή μεταγονιδιακά δεδομένα εντερικών μικροβιώματων απο 300 υγιείς ενήλικες, που έχουν συλλεχθεί την τελευταία 3ετια απο την ερευνητική ομάδα, σαν ομάδα ελέγχου και για σύγκριση με τα βρεφικά και παιδικά μικροβιώματα.

Ενοποίηση και καθαρισμός δεδομένων (ΕΕ 1.2)

Τα αρχικά συλλεχθέντα δεδομένα έχουν τη μορφή αρχείων FASTQ συνήθως σε επικαλυπτόμενα ζεύγη (Paired-End Illumina sequencing). Για την ενοποίηση τους είναι σημαντικό να καθαριστούν οι αλληλουχίες από λάθη, να εξεταστεί η ποιότητά τους και να αφαιρεθούν τυχόν τεχνικές αλληλουχίες (εκκινητές/primers, προσαρμοστές/adaptors). Για τον λόγο αυτό θα χρησιμοποιηθούν τα εργαλεία fastQC (Andrews 2010) και  trimmomatic (Bolger 2014) καθώς και κώδικας που έχει ήδη αναπτυχθεί από την ομάδα σε Perl και Python. Οι αλληλουχίες κάθε ομάδας δεδομένων θα προσαρμόζονται ώστε το εύρος τους να είναι κοινό μεταξύ όλων των ομάδων. Δεδομένου ότι τα περισσότερα δεδομένα αμπλικόνιων στην μικροβιολογία εντέρου στοχεύουν στην πολυμορφική περιοχή V4-V6 του γονιδίου της υπομονάδας 16 του ριβοσώματος (16S rRNA gene), αναμένεται μεγάλη αλληλοκάληψη με μικρές τροποποιήσεις λόγω διαφορετικών εκκινητών.

Δημιουργία μικροβιακών προφίλ (ΕΕ 1.3)

Η αρχικη επεξεργασία των αλληλουχιών θα γίνει με τη χρήση της μεθόδου UPARSE (Edgar 2013) όπως αυτή υλοποιείται στην πλατφόρμα IMNGS (Lagkouvardos 2016). Συνοπτικά μέσω αυτών των εργαλείων θα γίνει ενοποίηση των επικαλυπτόμενων ζευγών αλληλουχιών, αφαίρεση των χιμαιρικών αλληλουχιών, ομαδοποίηση τους σε ταξονομικές μονάδες και η παραγωγή ενός ενιαίου πίνακα μικροβιακών συχνοτήτων ανά δείγμα (OTU table).  Στη συνέχεια τα μικροβιακά προφίλ θα αναλυθούν με χρήση του πακέτου εργαλείων Rhea (Lagkouvardos et al. 2017) για την πρώτη επεξεργασία, τον υπολογισμό και την απεικόνιση των σχετικών αποστάσεων μεταξύ των δειγμάτων.

Εφαρμογή εργαλείων και ανάλυση αποτελεσμάτων (ΕΕ 1.4)

Με την χρήση του πρώτου εργαλείου που θα αναπτύξουμε (Εργ. 2.1) θα γίνει κατηγοριοποίηση των μικροβιωμάτων της ομάδας ελέγχου σε ομάδες ομοειδών προφίλ (εντερότυπων) και θα εκπαιδεύσουμε έναν ταξινομητή για την ομάδα αυτή. Στη συνέχεια θα εφαρμόσουμε το πρώτο εργαλείο για τα προφίλ κάθε χρονικής περιόδου, αναγνωρίζοντας έτσι τις όποιες εσωτερικές δομές υπάρχουν για κάθε χρονική ομάδα. Η αναπαράσταση της σχετικής απόστασης των προφίλ ανά χρονική περίοδο με τις εσωτερικές ομάδες κάθε περιόδου, φαίνεται στο Γράφημα 2Α. Με την εφαρμογή του δεύτερου εργαλείου θα υπολογίσουμε το βαθμό ωριμότητας κάθε προφίλ ως το αθροιστικό βαθμό κάλυψης των ταξονομικών πυρήνων που ορίζονται για κάθε ομάδα στο δείγμα των ενήλικων προφίλ. Αναλυτική περιγραφή της μεθόδου για το δεύτερο εργαλείο στην αντίστοιχη ενότητα. Τέλος, θα χρησιμοποιήσουμε το τρίτο εργαλείο για την καταγραφή της πορείας κάθε βρεφικού μικροβιώματος διαμέσου των ομάδων κάθε περιόδου (Γράφημα 2Β) και τη δημιουργία προβλεπτικών μοντέλων για την εξέλιξη του.

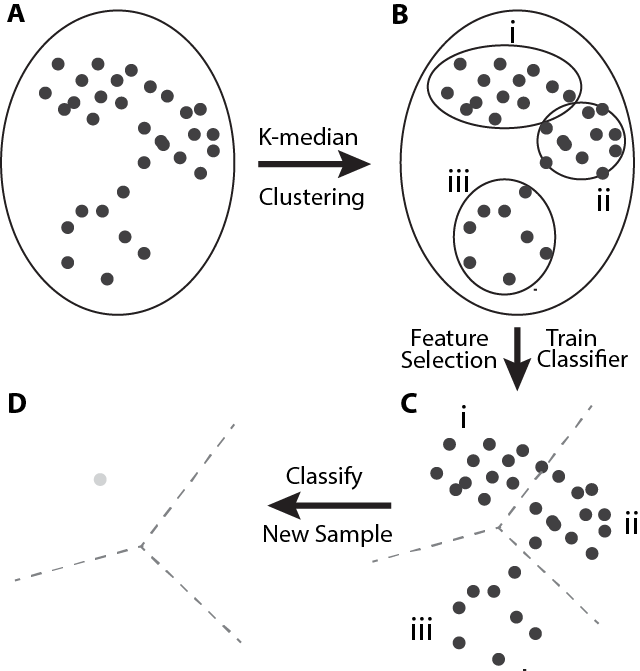
Ανάπτυξη αναλυτικών εργαλείων για μικροβιακά προφίλ (Πακέτο Εργασιών 2)

Περιγραφή και στόχοι του πακέτου εργασιών 2

Στο δεύτερο πακέτο εργασιών περιλαμβάνονται όλες οι εργασίες για την ανάπτυξη των νέων υπολογιστικών μεθόδων και εργαλείων για την ανάλυση των βρεφικών μικροβιωμάτων. Στόχος είναι η ανάπτυξη απλού και εύχρηστου κώδικα στη γλώσσα R για χρήση σε παρόμοια προβλήματα ανάλυσης χρονικών μικροβιακών δεδομένων και της αξιοποίησης υπαρχόντων δεδομένων για σύγκριση. Τα παραχθέντα προγράμματα θα προσφέρονται σαν επέκταση της υπάρχουσας πλατφόρμας Rhea για τυποποιημένη ανάλυση μικροβιωμάτων που συντηρεί η ερευνητική ομάδα (Lagkouvardos et al. 2017). Το πακέτο εργασίας 2 χωρίζεται στις ακόλουθες Ερευνητικές Ενότητες (ΕΕ).

Εργαλείο περιγραφής μικροβιωμάτων (ΕΕ 2.1)

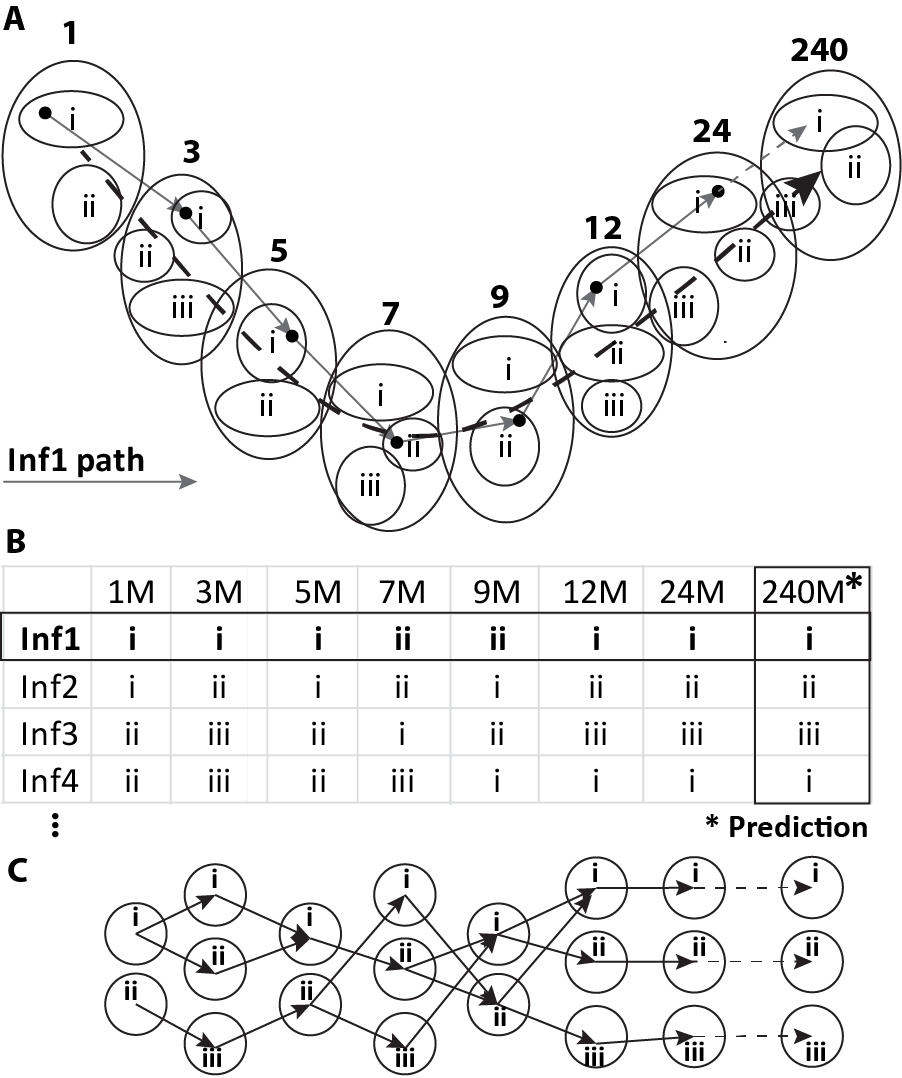
Το πρώτο εργαλείο θα χαρακτηρίζει (de novo) τα δεδομένα χρησιμοποιώντας k-medians ομαδοποιητές (Jain and Dubes 1988) όπου το βέλτιστο K θα επιλέγεται από τον υπολογισμό του δείκτη Calinski-Harabasz (CH), (Calinski and Harabasz 1974) έτσι ώστε οι αντίστοιχες αποστάσεις μεταξύ των ομάδων να είναι οι μέγιστες ενώ εντός των ομάδων οι ελάχιστες. Οι συνεπαγόμενες αυτές ομάδες, στη συνέχεια θα χαρακτηρίζονται βάσει της ταξινομικής τους σύστασης χρησιμοποιώντας κοινούς ταξινομητές τύπου Random forest και SVM, αφού πρώτα προηγηθεί διαδικασία ανακατασκευής των μεταβλητών λαμβάνοντας υπόψην την ιεραρχική δομή των δεδομένων (Hierarchical Feature Engineering) (Oudah and Henschel 2018).  Τα πιο ακριβή μοντέλα των ταξινομητών αυτών θα εφαρμοστούν σε κάθε χρονικό δείγμα των βρεφικών και ώριμων μικροβιωμάτων με σκοπό τον χαρακτηρισμό τους.



**Γράφημα 1.** Σχηματική αναπαράσταση της λειτουργίας του εργαλείου 1. Α) Απεικόνιση ομάδας μικροβιακών προφίλ με βάση την μεταξύ τους απόσταση (beta-diversity PCoA plot). Β) Αναγνώριση της εσωτερικής οργάνωσης των δεδομένων με την εφαρμογη του βέλτιστου k.  C) Εκπαίδευση ταξινομητών με βάση τα χαρακτηριστικά κάθε ομάδας. D) Χρήση ταξινομητών για την κατάταξη νέων προφίλ στην αντίστοιχη ομάδα.

Εργαλείο σύγκρισης μικροβιωμάτων (ΕΕ 2.2)

Σκοπός του δευτερου εργαλείου είναι η δημιουργία μέτρου ωριμότητας μικροβιακών προφίλ σε σύγκριση με μια ομάδα ελέγχου (βαθμός επικάλυψης). Δεδομένης μιας ομάδας δειγμάτων αναφοράς (π.χ. υγιεις ενήλικες) που έχει χαρακτηριστεί με το εργαλείο 1 (ΕΕ 2.1), το πρόγραμμα αρχικά θα διακρίνει τον ταξονομικό πυρήνα κάθε ομάδας. Στη συνέχεια για κάθε υπό διερεύνηση προφίλ θα υπολογίζεται ο βαθμός επικάλυψης ή ωρίμανσης για τα βρεφικά προφίλ ως το συσσωρευτικό ποσοστό των μικροβιακών ειδών του προφίλ που αντιστοιχούν στις ταξονομικές ομάδες του πυρήνα κάθε ομάδας. Με την εφαρμογή του εργαλείου αυτού σε χρονικές σειρές μικροβιακών προφίλ, θα δίνει τροχιές σύγκλισης ή απόκλισης προς τα κέντρα του δείγματος αναφοράς.



**Γράφημα 2.** Αναπαράσταση των βημάτων για τη δημιουργία προβλεπτικών μοντέλων σε χρονικές σειρές μικροβιακών προφίλ. Α) Απεικόνιση της σχετικής απόστασης των προφίλ ανά χρονική περίοδο ( beta diversity PCoA plot). Τα προφίλ σε κάθε χρονικό σημείο έχουν ομαδοποιηθεί και χαρακτηριστεί με τη χρήση του εργαλείου 1. Τα βέλη αντιστοιχούν στην υποθετική εξέλιξη του μικροβιώματος του βρέφους Inf1. Β) Ενδεικτικός πίνακας των μεταβάσεων στο χρόνο των μικροβιωμάτων κάθε βρέφους διαμέσου των διαμορφωμένων υποκατηγοριών.  Η αντιστοίχιση από το μήνα 24 στο μήνα 240 αποτελεί πρόβλεψη με βάση τα αποτελέσματα του εργαλείου 2. C) Σχηματική αναπαράσταση των παρατηρούμενων τροχιών των μικροβιωμάτων στο χρόνο. Τα δεδομένα αυτά θα χρησιμοποιηθούν για την εκπαίδευση μοντέλων Markov.

Εργαλείο ανάλυσης χρονικών σειρών (ΕΕ 2.3)

**Χρήση Μαρκοβιανού μοντέλου**

Ο στόχος του τρίτου εργαλείου είναι η δυνατότητα πρόβλεψης της κατάστασης του μικροβιώματος τη χρονική στιγμή *tn* δεδομένης της κατάστασης του μικροβιώματος την χρονική στιγμή *tm*, όπου *tm* < *tn*. Σε κάθε χρονική στιγμή *tm*στην οποία έχει πραγματοποιηθεί δειγματοληψία θα ομαδοποιήσουμε τα δείγματα βασιζόμενοι στο μικροβιωματικό τους προφίλ με το Εργαλείο 1. Για το βέλτιστο πλήθος ομάδων *k* (Εργαλείο 1), κάθε δείγμα *p* ανήκει στην ομάδα *i, 0 < i < k.* Επομένως, λαμβάνοντας υπόψη το σύνολο των δειγματοληψιών για κάθε άτομο, το άτομο χαρακτηρίζεται από μια ακολουθία ομαδοποιήσεων S = (Si,t1, Si,t2, …, Si,tn), όπου Si,t1 παριστάνει την ομάδα *i* την χρονική στιγμή t1 (Γράφημα 2Β). Η χρονική αυτή ακολουθία ομαδοποιήσεων θα χρησιμοποιηθεί ώστε να κατασκευαστεί ένα Μαρκοβιανό μοντέλο το οποίο προβλέπει την ομάδα του ατόμου σε κάποια χρονική στιγμή με βάση την κατάσταση του κάποια προηγούμενη χρονική στιγμή (Γράφημα 2C). Επειδή συχνά τα βιολογικά δεδομένα αποκλίνουν από την Μαρκοβιανή υπόθεση (με δεδομένη μια χρονική στιγμή κάθε πρότερη κατάσταση δεν δίνει επιπλέον πληροφορία), για κάθε χρονική στιγμή *tn* θα ελεγχθούν όλες οι προηγούμενες χρονικές στιγμές ώστε να κατανοήσουμε την προβλεπτική αξία κάθε χρονικής στιγμής. Σε περίπτωση που μεμονωμένες χρονικές στιγμές έχουν μειωμένη προβλεπτική ικανότητα, θα εξετάσουμε πολλαπλές χρονικές στιγμές (ζεύγος ή τριπλέτες).

**Χρήση μεθόδων μηχανικής μάθησης στην ανάλυση χρονικών δεδομένων (temporal data)**

Τα δεδομένα που θα αναλυθούν στην πρόταση συλλέγονται σε διάφορες χρονικές στιγμές, όχι όμως κατ’ ανάγκη σε ίσα χρονικά διαστήματα. Αυτό, ενδεχομένως, καταστήσει δύσκολη την Μαρκοβιανή ανάλυση των δεδομένων. Σε αυτήν την περίπτωση,  θα χρησιμοποιήσουμε μεθόδους μηχανικής μάθησης (Random Forests,Support Vector Machines και Neural Networks) ώστε να προβλέψουμε την ομαδοποίηση του μικροβιώματος στην ενήλικη φάση από προγενέστερες βρεφικές μικροβιωματικές καταστάσεις. Μέθοδοι μηχανικής μάθησης για ανάλυση χρονοσειρών υπάρχουν διαθέσιμες (για παράδειγμα Sapankevych and Sankar 2009), όμως οι χρονοσειρές προϋποθέτουν δειγματοληψία σε κανονικά χρονικά διαστήματα, γεγονός που δεν ισχύει απαραίτητα σε μελέτες ανθρώπινου μικροβιώματος. Θα εξετάσουμε αν το πληθος όλων των χρονικών δειγμάτων απαιτούνται ή αν ένα υποσύνολο αυτών (σε δεδομένες χρονικές στιγμές) μπορεί αξιόπιστα να προβλέψει το “ενήλικο μικροβίωμα”. Θα χρησιμοποιήσουμε κυρίως Support Vector Machines (SVMs; Vapnik et al. 1995) εξαιτίας της πολύ υψηλής απόδοσης ως μέθοδος ταξινόμησης. Σαν μεταβλητές θα χρησιμοποιήσουμε την ομαδοποίηση του μικροβιώματος (Εργαλείο 1) σε όλες τις περιόδους της βρεφικής ηλικίας. Ως κατηγορία δεδομένων θα χρησιμοποιήσουμε την ομαδοποίηση του μικροβιώματος στην ενήλικη φάση. Το σύνολο δεδομένων εκπαίδευσης (training set) θα αποτελείται από υποσύνολο όλων των δεδομένων, με τα οποία θα εκπαιδευτεί η μέθοδος. Μέρος αυτού θα χρησιμοποιηθεί και ως σύνολο εκτίμησης (validation set) ώστε να εκτιμηθούν οι βέλτιστες τιμές των παραμέτρων της μεθόδου. Τέλος, η τελική εκτίμηση της απόδοσης θα πραγματοποιηθεί σε ένα υποσύνολο των δεδομένων άγνωστο στη μέθοδο εκπαίδευσης και εκτίμησης (test dataset).

Ενοποίηση εργαλειων σε πλατφόρμα  (ΕΕ. 2.4)

Τα Εργαλεία 1, 2, και 3 θα ενοποιηθούν σε μια ενιαία ροή εργασίας με σκοπό να γίνουν διαθέσιμά προς χρήση από την κοινότητα.  Θα περιγραφούν αναλυτικά οι μέθοδοι που υλοποιούν και ο τρόπος ορθής χρήσης για κάθε εργαλείο και ο κώδικας των προγραμματων θα οργανωθεί έτσι ώστε να να ενισχύεται η διαυγεια κάθε λειτουργίας και η απλότητα στις ρυθμίσεις από την πλευρά του χρήστη. Θα ακολουθήσουμε την δομή των προγραμμάτων του πακέτου Rhea (Lagkouvardos et al. 2017) με σκοπό να υπάρχει συμβατότητα μεταξύ των δύο πλατφόρμων.

Διάχυση αποτελεσμάτων (Πακέτο Εργασιών 3)

Περιγραφή και στόχοι του πακέτου εργασιών 3

Η ανάλυση μικροβιακών κοινοτήτων είναι ένα πεδίο έρευνας με μεγάλη δυναμική με πλήθος ερευνητικών ομάδων που δραστηριοποιούνται στο συγκεκριμένο αντικείμενο. Αυτό από την μία δίνει ένα μεγάλο κοινό με άμεσο ενδιαφέρον στα αναμενόμενα αποτελέσματα και τα παραχθέντα εργαλεία,  αλλά από την άλλη δημιουργεί μια πρόκληση στην ανάδειξη των αποτελεσμάτων μέσα στην πληθώρα των δημοσιεύσεων στο χώρο. Για την αποτελεσματικότερη ανάδειξη των αποτελεσμάτων έχουμε προγραμματίσει τις παρακάτω εργασίες.

Παρουσίαση σε συνέδριο (ΕΕ 3.1)

Η παρουσίαση σε συνέδριο με την μορφή ομιλίας ή ανάρτησης δίνει την ευκαιρία της ταυτόχρονης προβολής των αποτελεσμάτων της παρούσας εργασίας σε ένα ευρύ κοινό επιστημόνων. Ανάμεσα στα εθνικά συνέδρια με σχετική θεματολογία που προσφέρονται για παρουσίαση, είναι αυτά της Ελληνικής Εταιρείας Βιοπληροφορικής και του Μικροβιόκοσμου για την διάδοση των αποτελεσμάτων της έρευνας στην Εθνική ερευνητική κοινότητα. Περαν των Εθνικών συνεδρίων, στόχος της ομάδας είναι να πραγματοποιήσει δύο παρουσιάσεις σε διεθνή συνέδρια με συναφές αντικείμενο (μικροβιώματος και βιοπληροφορικής) ώστε να επιτευχθεί η μέγιστη διάχυση των αποτελεσμάτων.

Συγγραφή και παρουσίαση αποτελεσμάτων (ΕΕ 3.2)

Η διεθνή κοινοποίηση και αναγνώριση των αποτελεσμάτων της έρευνας και η μεγιστοποίηση της χρήσης των παραγόμενων εργαλείων θα επιτευχθεί με την δημοσίευση τους σε έγκριτα επιστημονικά περιοδικά ανοικτής πρόσβασης. Η δημοσίευση σε περιοδικά ανοικτής πρόσβασης κρίνεται αναγκαία για την άμεση επικοινωνία με την κοινότητα. Ανάμεσα στα περιοδικά με σχετικό περιεχόμενο είναι τα  Bioinformatics, BMC Bioinformatics και Journal of Biomedical Informatics. Η δημοσίευση σε περιοδικά με πιο γενική θεματολογία όπως τα Scientific Reports, Plos One και PeerJ θα αξιολογηθεί ανάλογα με τα αποτελέσματα.

Χρονοδιάγραμμα

Για τον έλεγχο της ροής και την εξασφάλιση της επιτυχημένης ολοκλήρωσης των στόχων του έργου, οι προγραμματισμένες εργασίες έχουν χωριστεί σε τρία πακέτα εργασιών που καλύπτουν μια χρονική περίοδο 24 μηνών.  Συνοπτικά, το πρώτο πακέτο περιλαμβάνει τη συλλογή και επεξεργασία των μεταγωνιδιακών δεδομένων. Το δεύτερο πακέτο αφορά την ανάπτυξη των περιγεγραμμένων υπολογιστικών μεθόδων και εργαλειοθηκών. Τέλος, το τρίτο πακέτο συνδυάζει όλες τις εργασίες για την διάχυση τον αποτελεσμάτων της έρευνας όπως της παρουσίασης σε συνέδρια και τη συγγραφή και δημοσίευση σε επιστημονικά περιοδικά. Η κατανομή των εργασιών κάθε πακέτου στο συνολικό χρόνο του έργου και η μεταξύ τους εξάρτηση και ροή δεδομένων παρουσιάζεται αναλυτικά στο Διάγραμμα 3.



Διάγραμμα 3. Χρονική και θεματική κατανομή των προγραμματισμένων εργασιών με σήμανση των ροών δεδομένων.