Ανάλυση και εξόρυξη δεδομένων μεγάλης κλίμακας στην Ιατρική και την Βιολογία

Βλασιάδης Λοΐζος ΑΜ: 00685 (Ά εξ.)

Στα πλαίσια της εργασίας για το μάθημα θα παρουσιάσω το άρθρο με τίτλο “Clustering of Metagenomic Data by Combining Different Distance Functions” των Isis Bonet, Adriana Escobar, Andrea Mesa-Múnera και Juan Fernando Alzate. Το συγκεκριμένο άρθρο έχει δημοσιευτεί τον Οκτώβριο του 2017 στο Acta Polytechnica Hungarica και μπορεί να βρεθεί στο Researchgate με doi:10.12700/APH.14.3.2017.3.13 αλλά και στο παρακάτω σύνδεσμο.

Εισαγωγή

Το παρόν άρθρο ασχολείται με την ομαδοποίηση δεδομένων που προέρχονται από metagenomics, μεταγονιδιωματική στα ελληνικά, και προτείνει ένα νέο αλγόριθμο ομαδοποίησης. Αυτός ο νέος αλγόριθμος, βασίζεται στον αλγόριθμο k-means++, παρόμοιο με αυτόν που έχουμε δει στο μάθημα, καθώς και σε ένα consensus από ομαδοποιήσεις που βασίζονται σε διαφορετικές συναρτήσεις υπολογισμού αποστάσεων.
Το άρθρο καταλήγει ότι ο προτεινόμενος αλγόριθμος είναι πιο αποδοτικός από τον κλασσικό k-means.

Metagenomics

Οι μικροοργανισμοί βρίσκονται παντού στο φυσικό περιβάλλον και η μελέτη τους είναι σημαντική αφενός για να καταλάβουμε το ρόλο τους στη βιόσφαιρα και αφετέρου για να παράγουμε αντιμικροβιακές θεραπείες καθώς και λύσεις για διάφορα περιβαλλοντικά ζητήματα.
Δεν μπορούμε όμως να απομονώσουμε και να καλλιεργήσουμε όλους τους μικροοργανισμούς σε εργαστηριακό περιβάλλον. Έτσι προέκυψε η μεταγονιδιωματική (metagenomics), παράλληλα με την τεχνολογική εξέλιξη στην αλληλούχιση των ακολουθιών.
Ως μεταγονιδιωματική ορίζουμε την μελέτη γενετικού υλικού κατευθείαν από το φυσικό περιβάλλον, χωρίς να χρειάζεται να τα απομονώσουμε πρώτα. Σκοπός είναι να δούμε την ποικιλομορφία των μικροοργανισμών στο φυσικό τους περιβάλλον, χωρίς να χρειαστεί να τα καλλιεργήσουμε εμείς σε εργαστηριακές συνθήκες.
Παρόλα αυτά τα metagenomics δεν είναι πανάκεια και εξακολουθούν να υπάρχουν προβλήματα. Ενώ πλέον μπορούμε να έχουμε DNA αλληλουχίες, χωρίς όπως είπαμε να τις καλλιεργήσουμε στο εργαστήριο, δεν μπορούμε με αυτή την μέθοδο, να έχουμε πρόσβαση σε όλο το γονιδίωμα παρά μόνο σε κομμάτια (fragments) του DNA. Επίσης η παρουσία δειγμάτων DNA από πολλούς και διαφορετικούς οργανισμούς καθιστά πιο δύσκολη την αναπαράσταση της αλληλουχίας που ζητάμε. Για αυτό στη μεταγονιδιωματική υπάρχει μια διαδικασία που ονομάζεται binning και κατά την διάρκεια της οποίας προσπαθούμε να αναγνωρίσουμε ποια fragments (κομμάτια) ανήκουν σε ένα μικροοργανισμό.

Το binning μπορεί να γίνει χρησιμοποιώντας δυο μεθοδολογίες: Την composition-based και την similarity-based. Στην παρούσα έρευνα επέλεξαν να χρησιμοποιήσουν την composition-based μεθοδολογία που δεν είναι τόσο χρονοβόρα και υπολογιστικά πολύπλοκη. Η composition-based βασίζεται στην εύρεση χαρακτηριστικών, που αντικατοπτρίζουν τις αλληλουχίες, και με βάση αυτά μπορούμε να ταξινομήσουμε τους οργανισμούς.
Τα προβλήματα όμως δεν λείπουν και εδώ. Πιο συγκεκριμένα αναφέρουμε: οι βάσεις δεδομένων είναι μεγάλες και ετερογενείς (heterogeneous), ο αριθμός των ειδών σε ένα δείγμα μας είναι άγνωστος, τα fragments DNA μπορούν να διαφέρουν σε ποικιλία αλλά και τα fragments ενός συγκεκριμένου είδους μπορεί να είναι διαφορετικά σε αριθμό από ένα άλλο.
Για να αντιμετωπίσουν τα παραπάνω προβλήματα χρησιμοποίησαν στην έρευνα k-mers σαν αναπαράσταση χαρακτηριστικών. Τα k-mers είναι στην ουσία μια νουκλεοτοδική αλληλουχία συγκεκριμένου μήκους k. Όταν αναφερόμαστε σε k-mers εννοούμε όλες τις πιθανές ακολουθίες με μήκος k. Αν πχ είχαμε την ακολουθία GTAGAGCTGT και το k=2 τότε τα πιθανά k-mers θα ήταν: GT, TA, AG, GA, AG, GC, CT, TG, GT
Συνήθως είναι άνω του 2, οπότε μιλάμε για τετραμερή(4-mers = 256), εξαμερή (6-mers = 4096), κοκ.
Τα αποτελέσματα τα σύγκριναν με τα αποτελέσματα ενός απλού αλγόριθμου ομαδοποίησης χρησιμοποιώντας ένα δείκτη purity συγκρίνοντας τα αποτελέσματα από τα γκρουπ που δημιουργήθηκαν.

Μέθοδοι και Δεδομένα

Τα δεδομένα που χρησιμοποιήθηκαν στην έρευνα που μελετάμε προήλθαν από το Ινστιτούτο Sanger (ftp://ftp.sanger.ac.uk/) Για να επιτευχθεί η ποικιλομορφία, η βάση αποτελείται από εννέα (9) ευκαρυωτικούς οργανισμούς, πέντε (5) ιούς και δύο (2) βακτήρια όπως φαίνεται και στον Πίνακα 1.

Για την μεθοδολογία επιλέχθηκε ως k-mers = 4. Συνολικά η βάση αποτελείται από 872576 contigs. Τα contigs στο dataset διαφέρουν σε μέγεθος από 50 έως 30000 βάσεις.
Τα contigs, συντόμευση από την αγγλική λέξη contiguous = συνεχόμενος, μεταφράζεται ως συνεχόμενες αλληλουχίες. Προκειμένου η διαδικασία του binning να είναι πιο αποδοτική χρειαζόμαστε και μεγαλύτερα fragments DNA, τα contigs τα οποία συνήθως είναι επικαλυπτόμενα κομμάτια.
Επειδή όπως αναφέραμε επιλέχθηκε k=4 θα έχουμε 256 πιθανά τετρανουκλεοτίδια . Το κάθε ένα από αυτά τα τετρανουκλεοτίδια αντιστοιχεί σε ένα χαρακτηριστικό (feature).

Μέθοδος Ομαδοποίησης

Ο αλγόριθμος k-means++, προτάθηκε το 2007 από τους David Arthur και Sergei Vassilvitskii, είναι παρόμοιος με τον κλασσικό k-means μόνο που επιλέγει καλύτερα τα αρχικά κέντρα των ομαδοποιήσεων. Στη συνέχεια η υλοποίηση του είναι ίδια με τον κλασσικό αλγόριθμο k-means.
Ο K-means++ ξεκινάει επιλέγοντας ένα κέντρο τυχαία από τα δεδομένα μας. Στη συνέχεια καθορίζει τα άλλα clusters και τα κέντρα τους, με κατανομή πιθανοτήτων ανάλογα με τη συνάρτηση καθορισμού απόστασης που επιλέξαμε. Πχ η πιθανότητα ενός σημείου x είναι μεγαλύτερη όταν το σημείο αυτό δεν είναι κοντά σε ένα κέντρο. Το “κοντά” υπολογίζεται από τη συνάρτηση της απόστασης που επιλέγουμε να χρησιμοποιήσουμε. Στην εργασία μπορεί να είναι η Ευκλίδεια απόσταση, η Cosine similarity και η Jaccard index.

Η Ευκλίδεια απόσταση μετράει την ευθεία γραμμή μεταξύ δυο σημείων στον επίπεδο χώρο. Οι συναρτήσεις τους είναι οι παρακάτω:

$$Euclidean(X,Y)=\sqrt{\sum_{i=0}^{n}(x_i)-(y_i)^2}$$

Η Cosine similarity:

$$Cosine(X,Y)=1-\frac{\sum_{i=1}^{n}(x_i*y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}*\sqrt{\sum_{i=1}^{n}(y_i^2)}}$$

Jaccard index:

$$Jaccard(X,Y)=\frac{\sum_{i=1}^{n}(x_i-y_i)^2}{\sum_{i=1}^{n}(x_i)^2+\sum_{i=1}^{n}(y_i)^2-\sum_{i=1}^{n}(x_i*y_i)}$$

Για την τελική επιλογή της ποιότητας των αλγόριθμων ομαδοποίησης χρησιμοποιήθηκε μια μετρική, ένας δείκτης Purity.

$$Purity(C_j)=\frac{max(n_{ij})}{n_j}$$

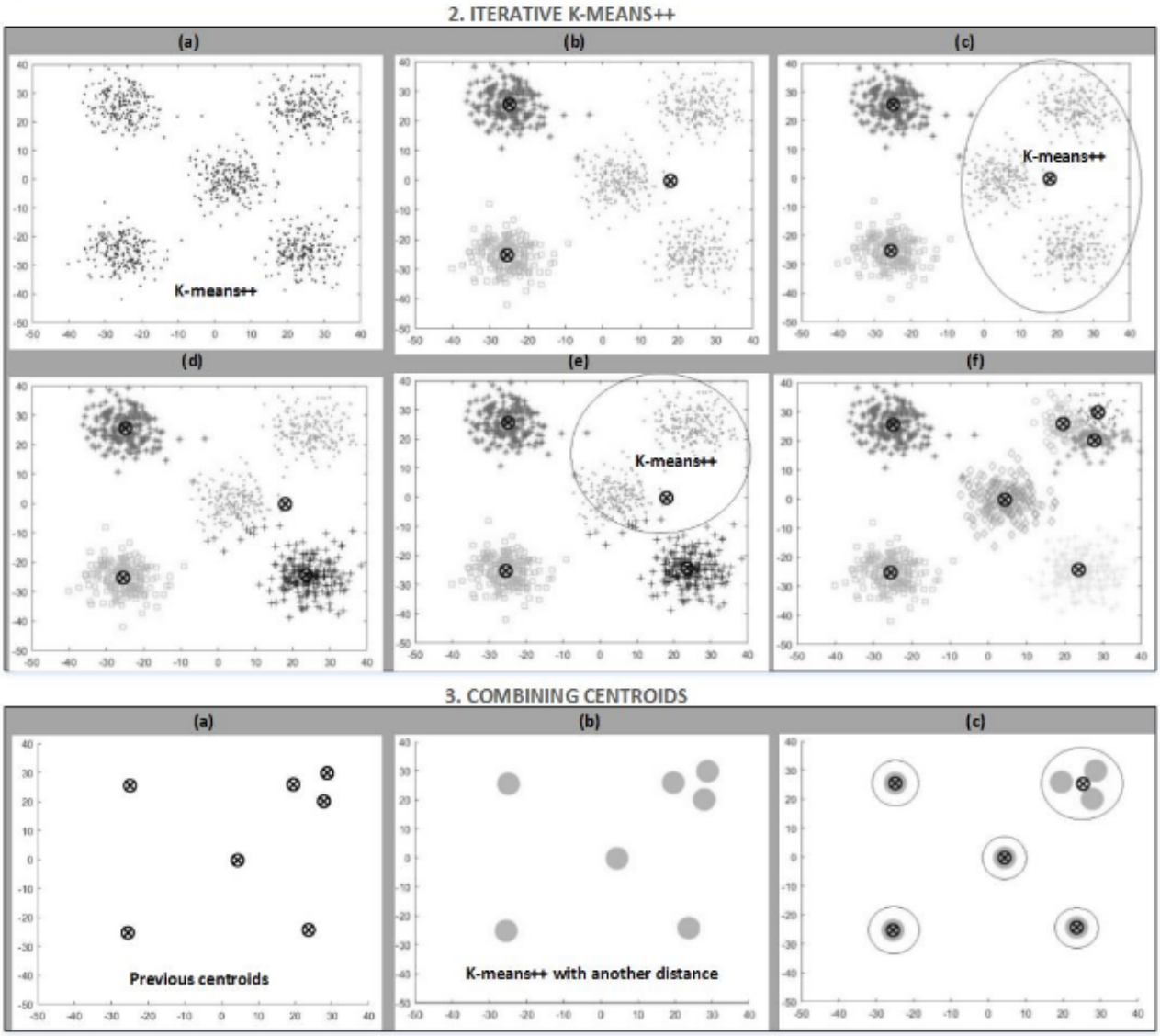
όπου n_j είναι ο αριθμός των οργανισμών στο cluster j (C_j) και n_{ij} είναι ο αριθμός οργανισμών κλάσης i στο cluster j. Δηλαδή θέλουμε να δούμε κατά πόσο ένα cluster περιέχει fragments που ανήκουν σε ένα μόνο οργανισμό.

Για την υλοποίηση του αλγόριθμου χρησιμοποίησαν το Weka 3.9 το οποίο είναι ένα δωρεάν εργαλείο για machine learning.

Μέθοδος Ομαδοποίησης με συνδυασμό διαφορετικών αποστάσεων

Όπως και στον k-means, οι αρχικές παράμετροι που πρέπει να ορίσουμε είναι η τιμή του k και ποια συνάρτηση απόστασης θα επιλέξουμε. Τώρα, σε αυτήν την εκδοχή μπορούμε να επιλέξουμε από δυο και πάνω. Η καλύτερη συνάρτηση απόστασης διαφέρει ανάλογα με το πρόβλημα.

Χρησιμοποιούμε κατά σειρά τη συνάρτηση της Ευκλίδειας απόστασης, του Cosine και της Jaccard για να δοκιμάσουμε την μέθοδο μας στη βάση δεδομένων που έχουμε.
Αρχικά χρησιμοποιούμε την Ευκλίδεια και τη Cosine απόσταση για να καθορίσουμε τις αποστάσεις των contigs. Στο επόμενο βήμα ενώνουμε κάποια cluster χρησιμοποιώντας την Jaccard συνάρτηση.



Εικόνα 1

Στην εικόνα 1 αποτυπώνεται η διαδικασία της ομαδοποίησης που προτείνει η εργασία που εξετάζουμε.

Το πρώτο βήμα, που δεν παρουσιάζεται στην εικόνα είναι να μετατρέψουμε τις ακολουθίες των metagenomics δεδομένων σε ένα αρχείο με βάση k-mers ώστε να μπορεί να το επεξεργαστεί το Weka.

Στο επόμενο δεύτερο βήμα μας, ξεκινάει η επαναληπτική ομαδοποίηση. Στο παράθυρο 2b βλέπουμε πως θα ήταν τα clusters αν είχαμε επιλέξει σαν k=3. Το x μας δείχνει που θα ήταν τα κέντρα των cluster μας. Όπως παρατηρούμε το τέρμα δεξιά cluster δεν είναι πολύ καλά ομαδοποιημένο.

Έτσι εφαρμόζουμε ξανά k-means, σε αυτό το cluster μόνο, με k=2 όπως φαίνεται και από το παράθυρο 2c.
Τα αποτελέσματα με τα νέα κέντρα εμφανίζονται στο 2d. Το cluster πάνω δεξιά δεν είναι συμπαγές και το κέντρο πολύ μακρυά οπότε εφαρμόζουμε ξανά τον αλγόριθμο μας ξανά με k=4 αυτή τη φορά και παίρνουμε τα αποτελέσματα του παραθύρου 2f, τα οποία είναι και τα τελικά μας αποτελέσματα.

Το κατώφλι (threshold) βάση του οποίου καθορίστηκε το πόσο συμπαγές είναι ένα cluster προέρχεται από τον τύπο:

$$Threshold = Mean_{i=1}^{n}(intra-cluster\ distance_i) + Std_{i=1}^{n}(intra-cluster\ distance_i)$$

όπου n είναι ο αριθμός των clusters, η intra-cluster απόσταση υπολογίζεται ως το μέσο των αποστάσεων μεταξύ κάθε στιγμιότυπου του cluster και του κέντρου του cluster. Η Std είναι η στάνταρ απόκλιση.

Η εφαρμογή πολλών επαναληπτικών μεθόδων ομαδοποίησης εμφανίζει περισσότερα cluster από τα απαραίτητα. Το σκεπτικό αυτής της λογικής είναι σε κάθε cluster να αντιστοιχούν μέλη ενός μόνο είδους, ακόμη και αν τα είδη είναι χωρισμένα σε διαφορετικά γκρουπ.

Αυτή τη στιγμή έχουμε ένα μεγάλο αριθμό από clusters, επτά (7) συγκεκριμένα, κάποια από τα οποία είναι πολύ κοντά. Θα προχωρήσουμε στο να τα μειώσουμε χρησιμοποιώντας ξανά τον αλγόριθμο k-means++ αλλά με διαφορετική συνάρτηση απόστασης αυτή τη φορά. Πιο συγκεκριμένα ο αλγόριθμος τώρα επιλέγει τη Jaccard συνάρτηση. Τελικά καταλήγει ότι ο αριθμός των clusters θα είναι πέντε (5) όπως αντιλαμβανόμαστε και από το παράθυρο 3c της εικόνας.

Στα metagenomics δεδομένα γενικότερα δημιουργούνται περισσότερα clusters από ότι ο αριθμός των ειδών που έχουμε στο δείγμα μας αλλά χρησιμοποιώντας το δείκτη Purity και την προτεινόμενη μέθοδο καταφέρνουμε τα επιθυμητά αποτελέσματα που είναι κάθε cluster να αντιστοιχεί σε ένα είδος.

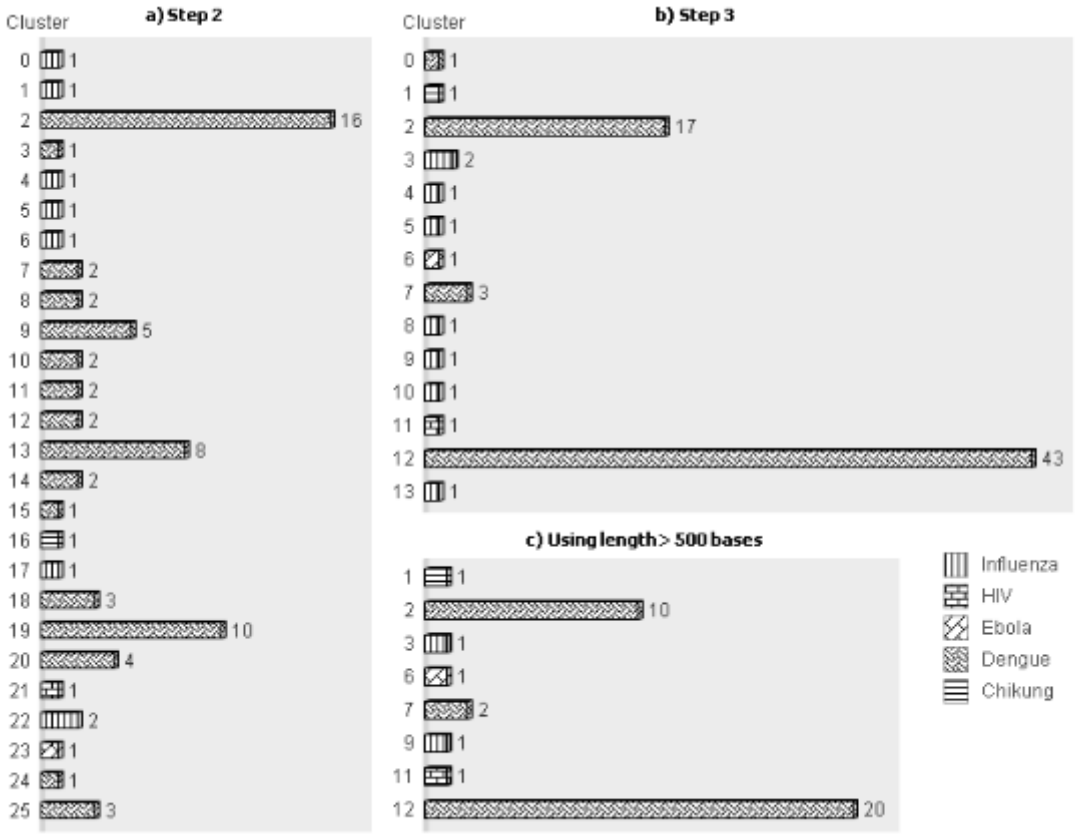
Τέλος σημαντικό είναι ότι υψηλό Purity μπορεί να επιτευχθεί όταν έχουμε πολλά clusters.

Αποτελέσματα και συζήτηση

Μια βάση δεδομένων με metagenomics δεδομένα από 16 οργανισμούς κατασκευάστηκε για να αξιολογηθεί η προτεινόμενη μέθοδος. Όπως είδαμε επιλέχθηκε k=4 για τα k-mers ως χαρακτηριστικό για να περιγράψουμε τις ακολουθίες. Οι Ευκλείδεια και Cosine συναρτήσεις αποστάσεων επιλέχθηκαν για την αρχικά βήματα του αλγόριθμου ενώ η Jaccard συνάρτηση στο τελικό στάδιο.
Τα δεδομένα από τη βάση χωρίστηκαν σε τρία (3) διαφορετικά σύνολα δεδομένων ανάλογα με τον οργανισμό. Ο αλγόριθμος βέβαια δοκιμάστηκε και έναντι ολόκληρης της βάσης. Τα αποτελέσματα που πήραμε όταν τρέξαμε τον αλγόριθμο στα ξεχωριστά dataset ήταν πολύ καλά για το κάθε πεδίο.

Οι ιοί παρόλο που είναι πιο δύσκολο να διαχωριστούν από τους άλλους οργανισμούς έδωσαν τα καλύτερα αποτελέσματα.

Στην παρακάτω εικόνα βλέπουμε τα αποτελέσματα από το dataset των ιών. Στην αριστερή στήλη έχουμε 26 clusters που προέκυψαν από τον αλγόριθμο με την Cosine μέθοδο και μετά από δύο φάσεις του k-means++ με k=10 και k=25 αντίστοιχα. Πάνω δεξιά παρατηρούμε τα αποτελέσματα από την τελευταία φάση του k-means++ χρησιμοποιώντας τη Jaccard συνάρτηση και k=14.



Εικόνα 2

Το purity στα cluster των ιών έφτασε το 100%. Οι Chikungunya, Ebola and HIV διαχωρίζονται σε ένα cluster όσο και ο αριθμός των contigs που περιέχουν. Από την άλλη, αρχικά οι Dengue και Influenza κατηγοριοποιούνται σε 16 και 7 clusters ενώ στο τελικό βήμα ομαδοποιούνται σε 4 και 7 clusters.

Στα βακτήρια τα καλύτερα αποτελέσματα τα είχαμε με k=10 και k=20. Το Purity στα βακτήρια ήταν γύρω στο 97.5% και αυτό παρόλο που είχαμε μόνο δυο κατηγορίες βακτηρίων να διαχωρίσουμε. Αυτό αποδίδεται στο γεγονός ότι τα contigs των βακτηρίων είναι περισσότερα και μεγαλύτερα σε μέγεθος.

Οι ευκαρυωτικοί οργανισμοί που είναι οι περισσότεροι επιτυγχάνουν ένα ποσοστό Purity 99.7% Οι οργανισμοί Candida parasilopsis και Aspergillus fumigatus δεν καταφέραμε να τους ομαδοποιήσουμε μεταξύ τους και να τους κατηγοριοποιήσουμε ξεχωριστά.

Τέλος γίνεται και ένας έλεγχος όπου “τρέχει” ο αλγόριθμος έναντι ολόκληρης της βάσης δεδομένων. Όταν τα contigs έχουν μέγεθος κάτω από 10000 και με k=15 και στη συνέχεια k=2500 έχουμε σαν αποτέλεσμα 2483 clusters ενώ όταν τα contigs έχουν μέγεθος μεγαλύτερο ή ίσο του 10000 και με k=15 και k=230 τα clusters που δημιουργούνται είναι 232.
Χρησιμοποιώντας contigs με μέγεθος άνω του 10000 έχουμε 100% purity ενώ σε contigs μεγέθους μικρότερου του 10000 το purity πέφτει στο 98.11%.

Συνοψίζοντας, η προτεινόμενη μέθοδος υλοποιώντας αρχικά μια επαναληπτική ομαδοποίηση και στη συνέχεια άλλη μια με βάση τα κέντρα της πρώτης καταλήγουμε ότι βελτιστοποιεί την διαδικασία του binning στα metagenomics. Η ουσία είναι στο να επιλέγουμε και να χρησιμοποιούμε διαφορετικές συναρτήσεις αποστάσεων.

Επιπλέον είναι σημαντικό να λαμβάνουμε υπόψη το μέγεθος των ακολουθιών. Προκειμένου να ελαχιστοποιήσουμε το σφάλμα μπορούμε να διαιρούμε το πρόβλημα και να δημιουργούμε μοντέλα που να εστιάζουν σε μικρές ακολουθίες και μοντέλα που εστιάζουν σε μεγάλες ακολουθίες αντίστοιχα.

Συμπεράσματα

Η συγκεκριμένη έρευνα προτείνει μια μέθοδο ομαδοποίησης με δυο φάσεις εκπαίδευσης. Η πρώτη είναι μια επαναληπτική διαδικασία ομαδοποίησης βασιζόμενη στον k-means++ ενώ η δεύτερη είναι άλλη μια ομαδοποίηση βασιζόμενη στα κέντρα της πρώτης φάσης. Σε κάθε βήμα επιλέγουμε διαφορετική συνάρτηση απόστασης.
Η μέθοδος εφαρμόστηκε σε μια βάση με metagenomic δεδομένα, αποτελούμενη από 16 οργανισμούς που χωρίζονταν σε τρεις κατηγορίες : Βακτήρια, Ιοί και Ευκαρυωτικά.
Τα αποτελέσματα, σε σύγκριση με έναν δείκτη Purity, ήταν πολύ καλύτερα από ένα κλασσικό k-means++ αλγόριθμο.
Μπορούμε να υποθέσουμε ότι μεγαλύτερα DNA fragments θα βελτιώσουν την απόδοση. Παρόλο που ο αριθμός των ομαδοποιήσεων είναι μεγαλύτερος από τον αντίστοιχο αριθμό των οργανισμών, η προτεινόμενη μέθοδος μας εξασφαλίζει 100% purity όταν τα fragments είναι μεγαλύτερα από 10.000 και 99% στις άλλες περιπτώσεις.
Αν και η μέθοδος έχει εφαρμοστεί μόνο σε μια βάση δεδομένων φαίνεται πολλά υποσχόμενη για ομαδοποιήσεις μεγάλων ακολουθιών ή ως το πρώτο βήμα για την διαδικασία της ταξινόμησης.

Πίνακας 1. Οργανισμοί στη βάση δεδομένων

Organism	**Domain**	**Contig**	**Min Length**	**Max Length**
Ascaris suum	Eukaryote	137650	50	30000
Aspergillus fumigatus	Eukaryote	295	1001	29660
Bacteroides dorei	Bacteria	1928	500	29906
Bifidobacterium longum	Bacteria	18	540	26797
Bos taurus	Eukaryote	315841	101	5000
Candida parasilopsis	Eukaryote	1540	1003	29956
Chikungunya	Virus	1	11826	11826
Dengue	Virus	64	10392	10785
Ebola	Virus	1	18957	18957
Glossina morsitans	Eukaryote	20334	101	29996
HIV	Virus	1	9181	9181
Influenza	Virus	8	853	2309
Malus domestica	Eukaryote	66739	102	5000
Manihot esculenta	Eukaryote	7192	1998	4998
Pantholops hodgsonii	Eukaryote	159729	50	5000
Zea mays	Eukaryote	161235	102	5000
		872576	**50**	**3000**