Ατομική Διπλωματική Εργασία

ΑΛΓΟΡΙΘΜΟΙ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ ΚΑΙ ΑΠΕΙΚΟΝΙΣΗΣ ΜΟΡΙΑΚΩΝ ΔΟΜΩΝ

Λουκία Παπαπαύλου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Δεκέμβριος 2013

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ
ΤΜΗΜΑ ΠΗΡΟΦΟΡΙΚΗΣ

Αλγόριθμοι Συσταδοποίησης και Απεικόνισης Μοριακών Δομών

Λουκία Παπαπαύλου

Επιβλέπων Καθηγητής
Κωνσταντίνος Παττίχης

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Δεκέμβριος 2013

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της διπλωματικής μου εργασίας Δρ. Κωνσταντίνο Παττίχη για την ευκαιρία που μου έδωσε να δουλέψω σ' αυτό το θέμα, για την πολύτιμη καθοδήγηση και συμβουλές του και για την υπομονή και κατανόηση του όλους αυτούς του μήνες.

Θα ήθελα επίσης να εκφράσω τις εξαιρετικές μου ευχαριστίες στον κ. Χρήστο Κάννα, διδακτορικό φοιτητή του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου, που χωρίς την βοήθεια του δεν θα ήταν δυνατή η εκπόνηση αυτής τις εργασίας.

Ακόμη θα ήθελα να εκφράσω την ευγνωμοσύνη μου στους γονείς μου για την αμέριστη στήριξη τους καθ' όλη την διάρκεια των σπουδών μου.

Τέλος ένα μεγάλο ευχαριστώ θα ήθελα να πω στην αδερφή μου Αιμιλία και στον φίλο μου Στέλιο για όλα αυτά που έκαναν για 'μένα.

Περίληψη

Στα πλαίσια αυτής της εργασίας μελετήθηκαν τρόποι συσταδοποίησης και απεικόνισης μοριακών δομών και συγκεκριμένα δραστικών ουσιών φαρμάκων που χρησιμοποιούνται για παθήσεις του αναπνευστικού, καρδιαγγειακού και νευρικού συστήματος (ψυχοαναληπτικά, αντικαταθλιπτικά, ψυχοδιεγερτικά, αντιεπιληπτικά, ψυχοληπτικά, αντιψυχωσικά, αγχολυτικά, υπνωτικά και ηρεμιστικά).

Αρχικά μελετήθηκαν δύο κύρια είδη συσταδοποίησης, η ιεραρχική και η διαχωριστική, και ακολούθως οι πιο αντιπροσωπευτικοί τους αλγόριθμοι. Έπειτα κάποιοι από αυτούς τους αλγόριθμους (αυτοί που κρίθηκαν κατάλληλοι για το είδος και το μέγεθος του συνόλου δεδομένων που είχαμε στη διάθεση μας) υλοποιήθηκαν στην γλώσσα προγραμματισμού R, δίνοντας μας διάφορετικές συσταδοποιήσεις του συνόλου δεδομένων ο καθένας, οι οποίες αναπαραστάθηκαν γραφικά.

Ακολούθως επιλέξαμε την καλύτερη συσταδοποίηση του κάθε αλγορίθμου και μελετήσαμε για κάθε μία από τις συστάδες της, ποια κοινά χαρακτηριστικά είχαν οι χημικές ενώσεις που ανήκαν σ’ αυτήν.

Τέλος με την χρήση διάφορων τεχνικών αξιολόγησης των αποτελεσμάτων μιας συσταδοποίησης, υπολογίσαμε τις τιμές των κριτηρίων αξιολόγησης και αναπαραστήσαμε γραφικά την απόδοση του κάθε αλγορίθμου στο κάθε ένα από αυτά.

Όπως φάνηκε από την ανάλυση της απόδοσης του κάθε αλγορίθμου στο καθένα από τα κριτήρια αξιολόγησης καλύτερη επίδοση είχε ο αλγόριθμος DIANA, που αποτελεί χαρακτηριστικό παράδειγμα αλγορίθμου ιεραρχικής διαιρετικής συσταδοποίησης και σε δύο από τα τρία κριτήρια αξιολόγησης είχε καλύτερες τιμές όταν ο αριθμός των συστάδων που του ζητούσαμε να δημιουργήσει ήταν πέντε.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Εισαγωγή

Στα πλαίσια αυτής της εργασίας ασχολήθηκα με την μελέτη και υλοποίηση αλγορίθμων συσταδοποίησης και απεικόνισης μοριακών δεδομένων, και συγκεκριμένα φαρμάκων που δόθηκαν στο Τμήμα Πληροφορικής του Πανεπιστημίου Κύπρου από τα ιδρύματα ZEINCRO Hellas S.A. και CHUV στα πλαίσια του προγράμματος Linked2Safety, στην γλώσσα προγραμματισμού R.

Η *R* είναι γλώσσα προγραμματισμού που χρησιμοποιείται για στατιστικούς υπολογισμούς και γραφικά και είναι ελεύθερα διαθέσιμη στην ιστοσελίδα http://www.r-project.org. Στηρίζεται στην ανάπτυξη προγραμμάτων μέσω πακέτων τα οποία διατίθενται ελεύθερα από χρήστες ανά τον κόσμο. Η R αποτελεί την ταχύτερα αναπτυσσόμενη γλώσσα στατιστικών υπολογισμών και την δημοφιλέστερη γλώσσα έρευνας στα πανεπιστήμια που έχει να κάνει με στατιστικές εφαρμογές [37],[44],[45].

Ο *ZEINCRO Hellas S.A.* είναι ένας ιδιωτικός οργανισμός κλινικών ερευνών που παρέχει συμβουλευτικές υπηρεσίες στην βιοφαρμακευτική κοινότητα στην Ελλάδα αλλά και γενικότερα. Η δράση του καλύπτει ένα ευρύ πεδίο από κλινικές έρευνες στον σχεδιασμό και παρακολούθηση φαρμάκων, βιο-στατιστικές έρευνες, διαχείριση βιο-δεδομένων, ασφάλεια φαρμάκων, φαρμακοοικονομικές μελέτες κ.ά [1].

Το *Κέντρο Ψυχιατρικής Επιδημιολογίας και Ψυχοπαθολογίας CHUV* ασχολείται σε ερευνητικό επίπεδο με α) τις ψυχιατρικές διαταραχές στους κατοίκους της πόλης της Λωζάνης και τη σχέση τους με τους παράγοντες κινδύνου για καρδιαγγειακά νοσήματα, β) την οικογενή συσσώρευση ψυχιατρικών διαταραχών στον γενικότερο πληθυσμό, γ) τις οικογενειακές μελέτες στην διπολική διαταραχή, την υποτροπιάζουσα κατάθλιψη, τον εθισμό σε αλκοόλ και ναρκωτικά, δ) την διαμήκη μελέτη παιδιών των οποίων οι γονείς υποφέρουν από διπολική διαταραχή, υποτροπιάζουσα κατάθλιψη, εθισμό στο αλκοόλ ή τα ναρκωτικά και ε) επικύρωση οργάνων αυτοεκτίμησης και ετερο-διατίμησης για χρήση σε επιδημιολογικές μελέτες [35].

Στόχος του *Linked2Safety* είναι η πρόοδος της κλινικής ιατρικής και η επιτάχυνση της ιατρικής έρευνας για την βελτίωση της ιατρικής περίθαλψης, της δημόσιας υγείας και της ασφάλειας του ασθενούς προσφέροντας στις εταιρείες φαρμάκων, στους επαγγελματίες

υγείας και στους ασθενείς μια καινοτόμα διαλειτουργική πλατφόρμα, ένα βιώσιμο επιχειρηματικό μοντέλο και μια επεκτάσιμη υποδομή για αποδοτική, ομογενοποιημένη πρόσβαση στις ολοένα αυξανόμενες ιατρικές πληροφορίες και για διασύνδεση των αρχείων ασθενών στην Ευρώπη και χρήση τους στην έρευνα [42].
Συνεργαζόμενο πρόγραμμα του Linked2Safety αποτελεί το Granatum, του οποίου στόχος είναι η γεφύρωση του χάσματος των πληροφοριών, της γνώσης και της συνεργασίας μεταξύ των ερευνητών στον τομέα της βιοατρικής διασφαλίζοντας ότι η επιστημονική κοινότητα θα έχει ομογενή και ολοκληρωμένη πρόσβαση στις διαθέσιμες πληροφορίες και τους απαραίτητους πόρους για την εκτέλεση πολύπλοκων πειραμάτων χημειοπροστασίας από τον καρκίνο και την διεξαγωγή μελετών σε σύνολα δεδομένων μεγάλης κλίμακας [36].
Στόχος της παρούσας διπλωματικής εργασίας είναι η συμβολή, σε όποιο βαθμό είναι αυτό δυνατό, στο έργο των δύο αυτών προγραμμάτων και στην επίτευξη των στόχων τους.

1.2 Στόχος Διπλωματικής

Αυτή η διπλωματική εργασία ασχολείται με την μελέτη και υλοποίηση αλγορίθμων συσταδοποίησης και απεικόνισης μοριακών δεδομένων, και συγκεκριμμένα φαρμάκων, στην γλώσσα προγραμματισμού R.
Στόχος είναι η ανάλυση και η σύγκριση των διάφορων τεχνικών και αλγορίθμων συσταδοποίησης και η εύρεση της καταλληλότερης τεχνικής για το συγκεκριμένο σύνολο δεδομένων, αλλά και για σύνολα δεδομένων που υπάγονται στην ίδια κατηγορία.
Ακολούθως στόχο αποτελεί και η εύρεση συσχετίσεων μεταξύ φαρμάκων που ανήκουν σε διαφορετικές κατηγορίες και χρησιμοποιούνται για την θεραπεία διαφορετικών παθήσεων ή/και παθήσεων που αφορούν διαφορετικά όργανα/συστήματα και η παροχή των ευρημάτων αυτών σε ερευνητικά ιδρύματα μέσα από τα προγράμματα Granatum και Linked2Safety με σκοπό την περαιτέρω διερεύνηση και την πιθανή εύρεση κοινής δράσης, άγνωστης μέχρι τώρα.

1.3 Δομή Διπλωματικής Εργασίας

Σ’ αυτό το κεφάλαιο έγινε αρχικά αναφορά στο αντικείμενομε το οποίο ασχολείται η παρούσα διπλωματική εργασία κι ακολούθως ανάλυση των στόχων της.

Στο ***δεύτερο κεφάλαιο*** παρουσιάζονται αναλυτικά οι δύο κατηγορίες στις οποίες χωρίζονται οι αλγόριθμοι συσταδοποίησης και αντιπροσωπευτικά συγκεκριμένα

παραδείγματα των αλγορίθμων αυτών, τόσο στο θεωρητικό τους κομάτι, όσο και στο πως υλοποιούνται στην γλώσσα προγραμματισμού R.

Στο ***τρίτο κεφάλαιο*** γίνεται μια εκτενής περιγραφή του συνόλου δεδομένων όπως αυτό μας δόθηκε από τα δύο ιδρύματα, ZEINCRO Hellas S.A. και CHUV, αλλά και οι επιπρόσθετες πληροφορίες που βρήκαμε γι αυτά τα δεδομένα από την online χημική βάση δεδομένων ChemSpider.

Στο ***τέταρτο κεφάλαιο*** παρουσιάζεται ο τρόπος και η μεθοδολογία που ακολουθήθηκε για την συσταδοποίηση των δεδομένων με κομάτια κώδικα στην γλώσσα προγραμματισμού R.

Στο ***πέμπτο κεφάλαιο*** παρουσιάζονται τα' αποτελέσματα της συσταδοποίησης για κάθε έναν από τους αλγόριθμους που αναφέρθηκαν στα προηγούμενα κεφάλαια, για διάφορες παραμέτρους κάθε φορά.

Τέλος, στο ***έκτο κεφάλαιο*** εξάγονται συμπεράσματα τα οποία βασίζονται στα αποτελέσματα της συσταδοποίησης που παρουσιάστηκαν στο προηγούμενο κεφάλαιο και γίνεται αναφορά σε πιθανή μελλοντική εργασία και βελτιώσεις πάνω στην παρούσα εργασία.

Κεφάλαιο 2

Αλγόριθμοι Συσταδοποίησης και Απεικόνισης

2.1 Ιεραρχική Συσταδοποίηση (hierarchical clustering)

Στόχος των αλγορίθμων ιεραρχικής συσταδοποίησης είναι η δημιουργία μιας ιεραρχίας συστάδων, η οποία μπορεί ν' αναπαρασταθεί με δενδρόγραμμα, κι όχι ένας απλός διαχωρισμός του συνόλου των δεδομένων σε συστάδες. Κριτήριο αυτής της συσταδοποίησης αποτελεί η απόσταση (distance) που τα αντικείμενα έχουν μεταξύ τους.

Ο υπολογισμός της ***απόστασης (distance)*** μεταξύ των αντικειμένων γίνεται με τους παρακάτω τρόπους [6], [19], [21], [24], [29], [31], [41]:

1. **Ευκλείδεια Απόσταση**: $\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
2. **Τετράγωνο Ευκλείδιας Απόστασης:** $\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
3. **Απόσταση Manhattan:** $\|a-b\|_1 = \sum_i |a_i - b_i|$
4. **Μέγιστη Απόσταση:** $\|a-b\|_\infty = \max_i |a_i - b_i|$
5. **Απόσταση Mahalanobis:** $\sqrt{(a-b)^T \cdot S^{-1} \cdot (a-b)}$ όπου S ο πίνακας συνδιακύμανσης
6. **Ομοιότητα Συνημιτόνου:** $\frac{a \cdot b}{\|a\| \cdot \|b\|}$
7. **Απόσταση Hamming/Binary:** Χρησιμοποιείται για τον υπολογισμό απόστασης μεταξύ δύο συμβολοσειρών ίδιου μήκους. Ισούται με τον αριθμό των αντίστοιχων θέσεων που οι δύο συμβολοσειρές έχουν διαφορετικό χαρακτήρα.

8. **Απόσταση Levenshtein:** Χρησιμοποιείται για τον υπολογισμό της απόστασης μεταξύ δύο συμβολοσειρών, χωρίς να είναι απαραίτητο να είναι του ίδιου μήκους. Ισούται με τον αριθμό των εισαγωγών, διαγραφών και αντικαταστάσεων χαρακτήρων ώστε οι δύο συμβολοσειρές να γίνουν ίδιες.
9. **Απόσταση Jaccard:** $1 - \frac{|a \cap b|}{|a \cup b|}$
10. **Απόσταση Soergel:** $\frac{|a|+|b|-2|a \cap b|}{|a|+|b|-|a \cap b|}$
11. **Απόσταση Dice:** Το συμπλήρωμα του λόγου του αριθμού των όμοιων στοιχείων προς τον μέσο όρο του αριθμού των στοιχείων που υπάρχουν συνολικά. $1 - \frac{2|a \cap b|}{|a|+|b|}$

Οι αλγόριθμοι αυτοί, ανάλογα με την προσέγγιση την οποία ακολουθούν, χωρίζονται σε δύο κατηγορίες: τους συσσωρευτικούς (agglomerative) και τους διαιρετικούς (divisive).

Οι ***συσσωρευτικοί αλγόριθμοι (agglomerative)*** ακολουθούν μια προσέγγιση από κάτω προς τα πάνω (bottom up) αφού αρχικά το κάθε αντικείμενο ανήκει σε δική του συστάδα (cluster) και σταδιακά συγχωνεύονται με βάση την απόσταση μεταξύ των συστάδων (linkage) καθώς ανεβαίνουμε τα επίπεδα της ιεραρχίας. Έχουν πολυπλοκότητα $O(n^3)$για τις περισσότερες περιπτώσεις και $O(n^2)$ για συγκεκριμένες περιπτώσεις (όταν χρησιμοποιείται μονή ή ολική σύνδεση) [23].

Ο υπολογισμός της ***απόστασης μεταξύ των συστάδων (linkage)*** γίνεται με τους παρακάτω τρεις διαφορετικούς τρόπους:

1. **Μονή Σύνδεση (single linkage):** η απόσταση μεταξύ των δύο συστάδων είναι ίση με την μικρότερη απόσταση μεταξύ του κάθε αντικειμένου της μιας συστάδας και των αντικειμένων της άλλης [40].
   Απόσταση *= min {distance(a,b), όπου* $a \epsilon A$*,* $b \epsilon B$*} όπου A, B συστάδες*
2. **Ολική Σύνδεση (complete linkage):** η απόσταση μεταξύ των δύο συστάδων είναι ίση με τη μέγιστη απόσταση μεταξύ του κάθε αντικειμένου της μιας συστάδας με τα αντικείμενα της άλλης [12].
   Απόσταση *= max {distance(a,b), όπου* $a \epsilon A$*,* $b \epsilon B$*} όπου A,B συστάδες*
3. **Μέση Σύνδεση (average linkage/UPGMA):** η απόσταση μεταξύ των δύο συστάδων είναι ίση με την μέση απόσταση του κάθε αντικειμένου της μιας συστάδας από τα αντικείμενα της άλλης[43].
   Απόσταση $= \frac{1}{|A| \cdot |B|} \sum_{a \epsilon A, b \epsilon B} diastance(a,b)$ *όπου A,B συστάδες*
4. **Μέθοδος Ward:** στόχος αυτής της μεθόδου είναι η μικρότερη δυνατή αύξηση της συνολικής εσωτερικής διακύμανσης μετά την συγχώνευση δύο

συστάδων, όπου διακύμανση είναι η σταθμισμένη τετραγωνική απόσταση μεταξύ των κέντρων των συστάδων [28].

Διακύμανση = $\frac{n_A n_B}{n_A+n_B}\|m_A - m_B\|^2$ *όπου* m_j *κέντρο συστάδας j και* n_j *αριθμός στοιχείων συστάδας j και A, B συστάδες.*

5. **Μέθοδος McQuitty:** σύμφωνα μ' αυτή τη μέθοδο μετά από κάθε συγχώνευση η απόσταση μεταξύ μιας καινούριας συστάδας *E* και μιας άλλης *C* υπολογίζεται με βάση τις αποστάσεις των δύο παλιών συστάδων *A,B* που έχουν συγχωνευτεί [10].

   Απόσταση = $\frac{n_A D_1+n_B D_2}{n_A+n_B}$ όπου D_1η απόσταση μεταξύ A και C και D_2η απόσταση μεταξύ B και C.

6. **Μέθοδος κέντρων (centroid/UPGMC):** η απόσταση μεταξύ δύο συστάδων είναι η Ευκλείδεια απόσταση μεταξύ των κέντρων τους [7].

   Απόσταση = $d(r,s) = \|\bar{x}_A - \bar{x}_B\|_2$ όπου *A, B συστάδες και* $\bar{x}_A, \bar{x}_B$ τα κέντρα τους. Το κέντρο σ' αυτή την περίπτωση είναι ο αριθμητικός μέσος και υπολογίζεται ως εξής: $\bar{x}_r = \frac{1}{n_r}\sum_{i=1}^{n_r} x_{r_i}$

7. **Μέθοδος μέσων (median/WPGMC):** η απόσταση μεταξύ δύο συστάδων είναι η Ευκλείδεια απόσταση μεταξύ των μέσων τους [30].

   Απόσταση = $\|\bar{x}_A - \bar{x}_B\|_2$ όπου *A, B συστάδες και* $\bar{x}_A, \bar{x}_B$ τα μέσα τους. Το μέσο σ' αυτή την περίπτωση είναι το ***σταθμισμένο κέντρο*** και υπολογίζεται ως εξής: $\bar{x}_C = \frac{1}{2}(\bar{x}_A + \bar{x}_B)$ όπου *C* η καινούρια συστάδα και *A, B* οι παλιές πριν την συγχώνευση συστάδες.

8. **Μέθοδος ευελιξίας (flexible):** κάνει χρήση του αλγορίθμου Lance-Williams ο οποίος καθορίζει την ***απόσταση*** ως εξής: $d(E,C) = a_1 d_{AC} + a_2 d_{BC} + \beta d_{AB} + \gamma|d_{AC} - d_{BC}|$ όπου *E* η καινούρια συστάδα που προέκυψε από την συγχώνευση των *A* και *B, C* μια οποιαδήποτε άλλη συστάδα, $\alpha_1, \alpha_2, \beta, \gamma$ παράμετροι οι οποίες εξαρτώνται από τα μεγέθη των συστάδων και d_{IJ} / d(I,J) η απόσταση μεταξύ των συστάδων I και J. [19]

9. **Μέθοδος σταθμισμένου μέσου όρου (weighted average/WPGMA):** Ο υπολογισμός της απόστασης μεταξύ των συστάδων μ' **α**υτή τη μέθοδο δίνεται από τον τύπο $d(E,C) = \frac{1}{2}(d_{AC} + d_{BC})$ όπου **E** η καινούρια συστάδα που προέκυψε από την συγχώνευση των συστάδων *A* **κ**αι *B, C* μια τυχαία συστάδα και **d**(X,Y) / d_{XY} η απόσταση μεταξύ των συστάδων X και Y [45].

Οι ***διαιρετικοί (divisive):*** Ακολουθούν μια προσέγγιση από πάνω προς τα κάτω (top down) αφού αρχικά όλα τα αντικείμενα ανήκουν σε μία συστάδα (cluster) και σταδιακά διαχωρίζονται καθώς κατεβαίνουμε τα επίπεδα της ιεραρχίας. Ο αλγόριθμος τρέχει αναδρομικά διαιρώντας την παλιά συστάδα σε δύο καινούριες μέχρι η κάθε συστάδα ν' αποτελείται από ένα ακριβώς αντικείμενο. Έχουν πολυπλοκότητα $O(2^n)$ [23].

2.1.1 Αλγόριθμος AGNES (AGlomerative NESting)

Ο αλγόριθμος αυτός ανήκει στους συσσωρευτικούς ιεραρχικούς αλγορίθμους και κατ' επέκταση χρησιμοποιεί μια προσέγγιση από κάτω προς τα πάνω (bottom-up). Αρχικά ο αλγόριθμος αυτός δημιουργεί μία συστάδα για το κάθε στοιχείο και σε κάθε βήμα επιλέγει ζευγάρια συστάδων με την μικρότερη απόσταση όπως αυτή καθορίζεται από το κριτήριο σύνδεσης (linkage criterion), και τα συγχωνεύει. Αυτό επαναλαμβάνεται έως ότου δημιουργηθεί μία συστάδα που περιέχει όλα τα στοιχεία [4].

Περιγραφή Αλγορίθμου [5]:

1. Ανάθεσε το κάθε στοιχείο σε μια συστάδα έτσι ώστε η κάθε συστάδα να έχει ακριβώς ένα στοιχείο.
2. Υπολόγισε την απόσταση μεταξύ των συστάδων και αποθήκευσε τις αποστάσεις στον πίνακα αποστάσεων.
3. Βρες το ζευγάρι συστάδων (r,s) με την μικρότερη απόσταση.
4. Συγχώνευσε τις συστάδες r και s σε μία νέα συστάδα t.
5. Υπολόγισε ξανά την απόσταση μεταξύ των συστάδων και ενημέρωσε τον πίνακα αποστάσεων.
6. Αν όλα τα στοιχεία βρίσκονται σε μία συστάδα τερμάτισε αλλιώς πήγαινε στο βήμα 3.

Περιγραφή του αλγορίθμου στην R βρίσκεται στο Παράρτημα Α τμήμα Α.1.

Ψευδοκώδικας [25]**:**

Input: $D = \{x_1, x_2 \ldots, x_n\}$ **(**στοιχεία που θα συσταδοποιηθούν)
A $n \times n$ πίνακας γειτνίασης

Output: Δενδρόγραμμα

for each $x_i \in D$ **do**
assign x_i to c_i (όπου c_i συστάδα i)
end

$k \leftarrow 0$;
$L(k) \leftarrow 0$; *// επίπεδο απόστασης της κοστής συσταδοποίησης*

Repeat
$c_r \leftarrow findMinPair(A)$ *// επιστρέφει τις συστάδες* c_r και c_s *που έχουν την μικρότερη απόσταση σύμφωνα με τον πίνακα γειτνίασης A*

$c_t \leftarrow merge(c_r, c_s)$;
$k \leftarrow k + 1$;
$L(k) \leftarrow A[r, s]$;

updateA(c_r, c_s, c_t); //*διαγράφει τις γραμμές από τον πίνακα γειτνίασης A που αντιστοιχούν στις συστάδες* c_r και c_s *και εισάγει μια νέα γραμμή που αντιστοιχεί στη νέα συστάδα.*

Until <*all objects are in one cluster*> ;

2.1.2 Αλγόριθμος DIANA (DIvisive ANAlysis)

Ο αλγόριθμος αυτός ανήκει στους διαιρετικούς ιεραρχικούς αλγόριθμους και κατ' επέκταση χρησιμοποιεί μια προσέγγιση από κάτω προς τα πάνω (top-down). Αρχικά ο αλγόριθμος αυτός δημιουργεί μία συστάδα η οποία περιέχει όλα τα στοιχεία του συνόλου δεδομένων και σε κάθε βήμα διαιρείται σε δύο ή περισσότερες καινούριες συστάδες, οι οποίες έχουν την μεγαλύτερη απόσταση μεταξύ τους, μέχρι η κάθε συστάδα ν' αποτελείται από ένα μόνο στοιχείο [16].

Περιγραφή Αλγορίθμου [20]:

1. Δημιούργησε μια συστάδα κι ανάθεσε της όλα τα στοιχεία.
2. Υπολόγισε την διάμετρο της κάθε συστάδας. Διάμετρος είναι η μέγιστη απόσταση μεταξύ των στοιχείων της συστάδας.
3. Επέλεξε την συστάδα C με την μεγαλύτερη διάμετρο.
4. Βρες το στοιχείο x της συστάδας C με την μεγαλύτερη απόσταση από τα υπόλοιπα.
5. Αφαίρεσε το στοιχείο x από τη συστάδα C και δημιούργησε μια συστάδα N με το στοιχείο αυτό.
6. Για κάθε αντικείμενο i της συστάδας C υπολόγισε την $d_C(i)$ (μέση απόσταση του στοιχείου i από τα στοιχεία της συστάδας C) και την $d_N(i)$ (μέση απόσταση του στοιχείου i από τα στοιχεία της συστάδας N). Αν τότε το στοιχείο i αφαιρείται από την συστάδα C και προστίθεται στην συστάδα $Nd_C(i) > d_N(i)$.
7. Επανέλαβε το βήμα 6 μέχρι για όλα τα στοιχεία i της συστάδας C να ισχύει: $d_C(i) < d_N(i)$
8. Επανέλαβε τα βήματα 2-7 μέχρι η κάθε συστάδα να περιέχει μόνο ένα στοιχείο.

Περιγραφή του αλγορίθμου στην R βρίσκεται στο Παράρτημα Α τμήμα Α.2.

2.2 Διαχωριστική Συσταδοποίηση (partitional clustering)

Στη διαχωριστική συσταδοποίηση καθορίζονται εξαρχής ο αριθμός των συστάδων που θα δημιουργηθούν όπως επίσης και τα κέντρα των συστάδων αυτών. Στη συνέχεια γίνεται ανάθεση όλων των αντικειμένων στις συστάδες ανάλογα με την απόσταση που έχουν τα αντικείμενα από τα κέντρα. Η απόσταση αυτή είναι ουσιαστικά μια συνάρτηση κριτηρίου στην οποία βασίζεται ο διαχωρισμός των αντικειμένων. Οι περισσότεροι αλγόριθμοι αυτής της κατηγορίας χρησιμοποιούν το κριτήριο τετραγωνικού σφάλματος [32]:

$$e^2(k, l) = \sum_{j=1}^{k} \sum_{i=1}^{N_j} \left\| x_j^i - c_j \right\|^2$$

όπου x_j^i το i-οστό στοιχείο της συστάδας, c_j το κέντρο της συστάδας j, k ο αριθμός των συστάδων, N_jο αριθμός των στοιχείων της συστάδας j

Περιγραφές των αλγορίθμων k-means, PAM και CLARA που θα αναλυθούν στην συνέχεια στην R βρίσκονται στο Παράρτημα Α τμήματα Α.4, Α.5 και Α.6 αντίστοιχα.

2.2.1 Αλγόριθμος k-means

Σκοπός του αλγορίθμου αυτού είναι ο διαχωρισμός n στοιχείων σε k συστάδες, όπου το κάθε στοιχείο θα βρίσκεται στη συστάδα με το κοντινότερο κέντρο (centroid). Ουσιαστικά στόχος του αλγόριθμου είναι να ελαχιστοποιήσει το άθροισμα των τετραγωνικών αποστάσεων εντός της κάθε συστάδας (WCSS - within cluster sum of squares) όπως δίνεται από τον πιο κάτω τύπο: $\sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2$ όπου $\left\| x_i^j - c_j \right\|^2$ η απόσταση μεταξύ ενός στοιχείου και του κέντρου της συστάδας στην οποία ανήκει [27].

Περιγραφή Αλγορίθμου [11]:

1. Όρισε τα αρχικά κέντρα των συστάδων. Αυτό μπορεί να γίνει ακολουθώντας διάφορες στρατηγικές. Μία εξ' αυτών, η πιο συνήθης, είναι η τυχαία ανάθεση κέντρων.
2. Ανάθεσε το κάθε στοιχείο στη συστάδα που έχει το κοντινότερο κέντρο στο στοιχείο. Αυτό γίνεται αφού υπολογιστεί η απόσταση μεταξύ του κάθε στοιχείου και του κάθε κέντρου ξεχωριστά.
3. Υπολόγισε ξανά τις τιμές των κέντρων. Οι νέες τιμές των κέντρων προκύπτουν από τον υπολογισμό του μέσου όρου των τιμών της κάθε συστάδας.
4. Επανέλαβε τα βήματα 2 και 3 μέχρι να μην παρατηρούνται αλλαγές στις συστάδες.

Ψευδοκώδικας [11]:

Input: E={e_1, e_2,..., e_n} (στοιχεία που θα συσταδοποιηθούν)
k (αριθμός συστάδων)
MaxIters (όριο επαναλήψεων)
Output: C={c_1, c_2,..., c_k} (κέντρα)
L={l(e) | e=1, 2,..., n} (ετικέτες στοιχείων συνόλου E)

for each $c_i \in C$ **do**
$c_i \leftarrow e_j \in E$ (e.g. random selection)
end

changed ←false;
iter ← 0;

repeat
for each $c_i \in C$ **do**
UpdateCluster(c_i);
end
foreach $e_i \in E$**do**
minDist ← argminDistance(e_i, c_j) $j \in \{1 \ldots k\}$;
if minDist ≠ l(e_i)**then**
l(e_i) ← minDist;
changed←true;
end
end
iter++;
until changed=true **and** iter≤MaxIters;

2.2.2 Αλγόριθμος PAM (Partitioning Around Medoids)

Σκοπός του αλγορίθμου αυτού, όπως και του k-means, είναι ο διαχωρισμός n στοιχείων σε k συστάδες, όπου το κάθε στοιχείο θα βρίσκεται στη συστάδα με το κοντινότερο κέντρο (medoid). Η διαφορά του με τον k-means είναι ότι τα κέντρα που ορίζονται σ' αυτόν τον αλγόριθμο αποτελούν μέρους του συνόλου των στοιχείων. Στόχος του αλγορίθμου είναι η ελαχιστοποίηση της απόστασης μεταξύ των στοιχείων της κάθε συστάδας και του κέντρου της και τον πετυχαίνει χρησιμοποιώντας το πιο κάτω μοντέλο [14]:

$$F(x) = minimize \sum_{i=1}^{n} \sum_{j=1}^{n} d(i,j) \cdot z_{ij}$$

όπου d(i,j) είναι η απόσταση μεταξύ των στοιχείων i και j, και z_{ij} είναι η μεταβλητή η οποία διασφαλίζει ότι μόνο στοιχεία που ανήκουν στην ίδια συστάδα θα ληφθούν υπόψη στον υπολογισμό.
εφόσον:

1. $\sum_{i=1}^{n} z_{ij}, j = 1,2, \dots, n$
2. $z_{ij} \leq y_i, i,j = 12, \dots, n$
3. $\sum_{i=1}^{n} y_i = k$, $k =$ αριθμός συστάδων
4. $y_i , z_{ij} \epsilon \{0,1\}, i,j = 1,2, \dots, n$

Το (1) διασφαλίζει ότι το κάθε στοιχείο ανήκει σε μία και μόνο συστάδα.
Το (2) διασφαλίζει ότι το στοιχείο ανατίθεται στο κέντρο το οποίο αντιπροσωπεύει τη συστάδα.
Το (3) ότι ο αριθμός των συστάδων είναι k.
Το (4) διασφαλίζει ότι η μεταβλητή απόφασης είναι είτε 0 είτε 1.
Ο αλγόριθμος αυτός μπορεί να δεχτεί 2 ειδών εισόδους, είτε ένα πίνακα με τις τιμές του κάθε στοιχείου είτε ένα πίνακα αποστάσεων και αποτελείται από δύο φάσεις:

Φάση Οικοδόμησης:

1. Επέλεξε k στοιχεία σαν κέντρα (medoids) αν δεν έχουν δοθεί από το χρήστη.
2. Υπολόγισε τον πίνακα αποστάσεων αν δεν έχει δοθεί.
3. Ανάθεσε κάθε στοιχείο στο κοντινότερο του κέντρο (medoid).

Φάση Ανταλλαγής:

4. Για κάθε συστάδα ψάξε αν κάποιο από τα στοιχεία του ελαχιστοποιεί την μέση απόσταση της συστάδας αν γινόταν κέντρο και κάνε το κέντρο.
5. Αν έχει αλλάξει έστω και ένα κέντρο πήγαινε στο βήμα (3) αλλιώς τερμάτισε.

Περιγραφή Αλγορίθμου [11]:

Input: $E = \{e_1, e_2, \ldots, e_n\}$ (στοιχεία που θα συσταδοποιηθούν ή πίνακας αποστάσεων)
k (αριθμός συστάδων)
metric (μετρική/είδος απόστασης που θα χρησιμοποιηθεί για την δημιουργία του πίνακα αποστάσεων)
diss (σημαία που υποδεικνύει αν το E είναι πίνακας αποστάσεων ή όχι)

Output: $M = \{m_1, m_2, \ldots, m_k\}$ (πίνακας με κέντρα)
$L = \{I(e) | e = 1,2, \ldots, n\}$ (ετικέτες συνόλου στοιχείων E)

for each $m_i \in M$ **do**
$m_i \leftarrow e_j \in E$ *//e.g. random selection*
end
if diss ≠ true
$Dissimilarity \leftarrow CalculateDissimilarityMatrix(E, metric)$;
else
Dissimilarity ← E;
end
repeat
for each $e_i \in E$ **do**
$I(e_i) \leftarrow$ argminDissimilarity(e_i, Dissimilarity, L);
end
changed ← false;
for each $m_i \in M$ **do**
Mtmp ← SelectBestClusterMedoids(E, Dissimilarity, L);
end
if Mtm ∗ p ≠ M
M ← $Mtmp$;
$changed \leftarrow true$;
end
until changed=true;

2.2.3 Αλγόριθμος CLARA (Clustering for LARge Applications)

Σκοπός αυτού του αλγόριθμου είναι η συσταδοποίηση μεγάλων σετ δεδομένων. Δουλεύει επεκτείνοντας την προσέγγιση k-medoids για μεγάλο αριθμό αντικειμένων. Συσταδοποιεί ένα δείγμα από το σετ δεδομένων χρησιμοποιώντας τον αλγόριθμο PAM και μετά αναθέτει όλα τα αντικείμενα(στοιχεία) του σετ δεδομένων σ' αυτές τις συστάδες. Η ποιότητα των κέντρων(medoids) που βρέθηκαν μετράται από την μέση απόσταση μεταξύ κάθε στοιχείου μέσα σ' ολόκληρο το σετ δεδομένων και το κέντρο της συστάδας του, όπως ορίζεται από την ακόλουθη συνάρτηση [13]:

$$\text{Cost(M, D)} = \frac{\sum_{i=1}^{n} dissimilarity(O_i, rep(M, O_i))}{n}$$

όπου D το σετ δεδομένων, M το σετ των κέντρων (medoids), dissimilarity(X,Y) η απόσταση μεταξύ των στοιχείων X και Y, και rep(M,X) επιστρέφει το κέντρο από το σετ M που είναι πιο κοντά στο X.

Για να ελαχιστοποιηθεί η μεροληψία, ο αλγόριθμος επαναλαμβάνει την διαδικασία δειγματοληψίας για επιλογή κέντρων και τη διαδικασία συσταδοποίησης ένα προκαθορισμένο αριθμό φορών και ακολούθως επιλέγει σαν τελική συσταδοποίηση αυτή που έχει το μικρότερο κόστος. Η ποιότητα των αποτελεσμάτων του CLARA εξαρτάται σε πολύ μεγάλο βαθμό από το μέγεθος του δείγματος που επιλέγει. Όσο πιο μικρό είναι το μέγεθος του δείγματος τόσο χειρότερη είναι η ποιότητα των αποτελεσμάτων.

Επειδή αυτός ο αλγόριθμος απευθύνεται σε μεγάλα σύνολα δεδομένων δεν θα υλοποιηθεί στα πλαίσια αυτής της εργασίας καθώς όπως θα δούμε και στο Κεφάλαιο 3 το σύνολο των δεδομένων που έχω στην διάθεση μου προς συσταδοποίηση είναι περιορισμένο.

Ψευδοκώδικας

Input: $D = \{O_h\}\ h = 1,2,\ldots,n$ (σετ δεδομένων)
q (αριθμός επαναλήψεων)

$f(V)_{max} = 0$;
iteration=0;
Repeat

1. Επέλεξε ένα δείγμα S τυχαία από το D
2. Τρέξε τον αλγόριθμο PAM στο S για να βρεις k κέντρα (medoids)
3. **for each** $O_h \in D$ **do**
    a. Υπολόγισε την απόσταση του O_hαπό το κάθε κέντρο (medoid)
    b. Ανάθεσε το O_h στη συστάδα με το κοντινότερο κέντρο (medoid)
4. Υπολόγισε το f(V)
5. **if** $f(V) < f(V)_{max}$ **then**
    iteration++;
    else
    $f(V)_{max} = f(V)$;
    BestSets = CurrentSets;
    go to step2

Until iteration=q;

Output: BestSets (η καλύτερη συσταδοποίηση του D σε k συστάδες)

Κεφάλαιο 3

Περιγραφή Συνόλου Δεδομένων

3.1 Εισαγωγή

Τα δεδομένα που χρησιμοποιήθηκαν στη συσταδοποίηση ήταν χημικές ενώσεις (compounds) που αποτελούν τις δραστικές ουσίες των φαρμάκων που μας δόθηκαν από τα ιδρύματα ZEINCRO Hellas S.A. και CHUV στα πλαίσια της συνεργασίας του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου για το project Linked2Safety και χρησιμοποιούνται για παθήσεις του αναπνευστικού, καρδιαγγειακού και νευρικού συστήματος (ψυχοαναληπτικά, αντικαταθλιπτικά, ψυχοδιεγερτικά, αντιεπιληπτικά, ψυχοληπτικά, αντιψυχωσικά, αγχολυτικά, υπνωτικά και ηρεμιστικά). Αφού δόθηκαν οι δραστικές ουσίες και τα ATC codes (Anatomical Therapeutic Chemical - χρησιμοποιούνται για την ταξινόμηση των φαρμάκων ανάλογα με το όργανο ή/και το σύστημα στο οποίο δρουν θεραπευτικά), χρησιμοποιήσαμε την online χημική βάση δεδομένων Chemspider για να βρούμε το SMILES, τον χημικό τύπο, την κοινή ονομάσια, το μοριακό βάρος, το ALogP και το XLogP. Ακολούθως υπολογίσαμε το 1024-bit Morgan Fingerprint για το κάθε compound. Τις περισσότερες φορές για την συσταδοποίηση χρησιμοποιήθηκε ο πίνακας αποστάσεων (η μέθοδος υπολογισμού της απόστασης που χρησιμοποιήθηκε ήταν η Soergel) ανάλογα με τις απαιτήσεις του κάθε αλγορίθμου.

3.2 Φάρμακα από ZEINCRO Hellas S.A.

Πίνακας 3.2 *Φάρμακα από ίδρυμα ZEINCRO Hellas S.A. μαζί με πληροφορίες που βρήκαμε από την ChemSpider*

Δραστική Ουσία	ATC	Κατηγορία	Κοινή Ονομασία	ChemSpider ID
Budesonide	R03BA02	1	Budesonide	4444479
Salbutamol	R03AC02 R03CC02	1	Asmol	1999
Fluticasone	R01AD08 R03BA05	1	Fluticasone	4470631
Formoterol + Budeosodine	R03AK07	1	Formoterol + Budeosodine	2340731 + 4444479
Formoterol	R03AC13	1	Formoterol	2340731
Montelukast	R03D C03	1	Montelukast	4444507
Desloratadine	R06AX27	1	Desloratadine	110575
Salmeterol	R03AC12	1	Salmeterol	4968
Theophylline	R03DA04	1	Theophylline	2068
Ciclesonide	R03B A08	1	Ciclesonide	5293368
Candesartan + Hydroxhlorothiazide	C09DA06	2	Candesartan + Hydroxhlorothiazide	2445 + 3513
Ramipril	C09AA05	2	Ramipril	4514937
Valsartan	C09CA03	2	Valsartan	54833
Rosuvastatin	C10A A07	2	Rosuvastatin	393589
Moxonidine	C02AC05	2	Moxonidine	4645
Clopidogrel	B01AC-04	2	Clopidogrel	54632
Glyceryl trinitrate	C01DA02	2	Nitroglycerin	4354
Metoprolol	C07AB02	2	Metoprolol	4027
Simvastatin	C10A A01	2	Lipex	49179
Irbesartan	C09C A04	2	Irbesartan	3618

Losartan	C09CA01	2	Losartan	3824
Amlodipine	C08CA01	2	Amlodipine	2077
Acetylsalicylic acid	B01AC06	2	Aspirin	2157
Enalapril	C09AA02	2	Enalapril	4534998
Enalapril + Hydrochlorothiazide	C09BA02	2	Enalapril + Hydrochlorothiazide	4534998+3513
Bisoprolol	C07AB07	2	Bisoprolol	2312
Digoxin	C01A A05	2	Digon	2006532
Atorvastatin	C10AA05	2	Atorvastatin	54810
Candesartan	C09CA06	2	Candesartan	2445
Glimepiride	A10B B12	3	Glimepiride	16740595
Metformin	A10BA02	3	Metformin	3949
Pioglitazone	A10BG03	3	Pioglitazone	4663
Insulin	A10AB01	3	Insulin	4395710

3.3 Φάρμακα από CHUV

Πίνακας 3.3 *Φάρμακα από ίδρυμα CHUV μαζί με πληροφορίες που βρήκαμε από την ChemSpider*

Δραστική Ουσία	ATC	Κατηγορία	Κοινή Ονομασία	ChemSpider ID
Imipramine	N06AA02	4.1.1	Imipramine	3568
Clomipramine	N06AA04	4.1.1	Clomipramine	2699
Trimipramine	N06AA06	4.1.1	Trimipramine	5382
Dibenzepin	N06AA08	4.1.1	Dibenzepin	9048
Amitriptyline	N06AA09	4.1.1	Amitriptyline	2075
Maprotiline	N06AA21	4.1.1	Maprotiline	3871
Fluoxetine	N06AB03	4.1.2	Fluoxetine	3269
Citalopram	N06AB04	4.1.2	Citalopram	2669

Paroxetin	N06AB05	4.1.2	Paroxetin	39888
Sertraline	N06AB06	4.1.2	Sertraline	61881
Fluvoxamine	N06AB08	4.1.2	Luvox	4481878
Escitalopram	N06AB10	4.1.2	Escitalopram	129277
Moclobémide	N06AG02	4.1.3	Moclobémide	4087
Mianserin	N06AX03	4.1.4	Tolvin	4040
Trazodone	N06AX05	4.1.4	Trazodone	5332
Mirtazapine	N06AX11	4.1.4	Mirtazapine	4060
Bupropion	N06AX12	4.1.4	Bupropion	431
Venlafaxine	N06AX16	4.1.4	Venlafaxine	5454
Reboxetine	N06AX18	4.1.4	Reboxetine	2289101
Duloxetine	N06AX21	4.1.4	Duloxetine	54822
Amfetamine	N06BA01	4.2.1	Amfetamine	13852819
Methylphenidate	N06BA04	4.2.1	Ritalin	4015
Carbamazepine	N03AF01	5.1	Carbamazepine	2457
Valproic acid	N03AG01	5.2	Valproic acid	3009
Lamotrigine	N03AX09	5.3	Lamotrigine	3741
Topiramate	N03AX11	5.3	Topiramate	4447672
Gabapentin	N03AX12	5.3	Neurontin	3328
Levomepromazine	N05AA02	6.1.1	Nomizan	3779
Promazine	N05AA03	6.1.1	Promazine	4757
Periciazine	N05AC01	6.1.2	Neulactil	4585
Thioridazine	N05AC02	6.1.2	Thioridazine	5253
Pipotiazine	N05AC04	6.1.2	Piportil	56598
Haloperidol	N05AD01	6.1.3	Haloperidol	3438
Fluanisone	N05AD09	6.1.3	Metorin	14410

Flupentixol	N05AF01	6.1.4	Flupentixol	4445173
Chlorprothixene	N05AF03	6.1.4	Chlorprothixene	580849
Zuclopenthixol	N05AF05	6.1.4	Zuclopenthixol	4470984
Pimozide	N05AG02	6.1.5	Neoperidole	15520
Penfluridol	N05AG03	6.1.5	Penfluridol	31017
Clozapine	N05AH02	6.1.6	Clozapine	10442628
Olanzapine	N05AH03	6.1.6	Olanzapine	10442212
Quetiapine	N05AH04	6.1.6	Quetiapine	4827
Clotiapine	N05AH06	6.1.6	Clotiapine	15510
Amisulpride	N05AL05	6.1.7	Amisulpride	2074
Lithium	N05AN01	6.1.8	Lithium	2293625
Risperidone	N05AX08	6.1.9	Belivon	4895
Aripiprazole	N05AX12	6.1.9	Aripiprazole	54790
Diazepam	N05BA01	6.2.1	Valium	2908
Oxazepam	N05BA04	6.2.1	Penfluridol	4455
Potassium clorazepate	N05BA05	6.2.1	Azene	25043757
Lorazepam	N05BA06	6.2.1	Lorazepam	3821
Bromazepam	N05BA08	6.2.1	Bromazepam	2347
Alprazolam	N05BA12	6.2.1	Xanax	2034
Buspirone	N05BE01	6.2.2	Buspirone	2383
Flurazepam	N05CD01	6.3.1	Flurazepam	3276
Flunitrazepam	N05CD03	6.3.1	Flunitrazepam	3263
Triazolam	N05CD05	6.3.1	Triazolam	5355
Midazolam	N05CD08	6.3.1	Midazolam Base	4047
Zopiclone	N05CF01	6.3.2	Zopiclone	5533
Zolpidem	N05CF02	6.3.2	Zolpidem	5530

3.4 Κατηγορίες Φαρμάκων

Πίνακας 3.4 *Οι κατηγορίες των φαρμάκων από τα ιδρύματα ZEINCRO Hellas S.A. και CHUV αριθμημένες*

Αριθμός Κατηγορίας	Όνομα Κατηγορίας
1	Respiratory
2	Cardiovascular
3	Endocrinology
4	Psychoanaleptics
4.1	Antidepressants
4.1.1	Non-selective monoamine reuptake inhibitors
4.1.2	Selective serotonin reuptake inhibitors (SSRIs)
4.1.3	Monoamine oxidase A inhibitors
4.1.4	Other antidepressants
4.2	Psychostimulants, agents used for ADHD and Nootropics
4.2.1	Centrally acting sympathomimetics
5	Antiepileptics
5.1	Carboxamide derivatives
5.2	Fatty acid derivatives
5.3	Other antiepileptics
6	Psycholeptics
6.1	Antipsychotics
6.1.1	Phenothiazines with aliphatic side-chain
6.1.2	Phenothiazines with piperidine structure
6.1.3	Butyrophenone derivatives
6.1.4	Thioxanthene derivatives
6.1.5	Diphenylbutylpiperidine derivatives

6.1.6	Diazepines, oxazepines, thiazepines and oxepines
6.1.7	Benzamides
6.1.8	Lithium
6.1.9	Other antipsychotics
6.2	Anxiolytics
6.2.1	Benzodiazepine derivatives
6.2.2	Azaspirodecanedione derivatives
6.3	Hypnotics and Sedatives
6.3.1	Benzodiazepine derivatives
6.3.2	Benzodiazepine related drugs

Κεφάλαιο 4

Αποτελέσματα

4.1 AGNES (Complete Linkage)

Agglomerative Coefficient: 0.39

Οι πίνακες των αποτελεσμάτων για τις 2 πρώτες προσπάθειες βρίσκονται στο Παράρτημα Γ τμήματα Γ.1 και Γ.2 αντίστοιχα.

1η Προσπάθεια:

Αποκοπή: 0.85
Συστάδες που δημιουργήθηκαν: 27

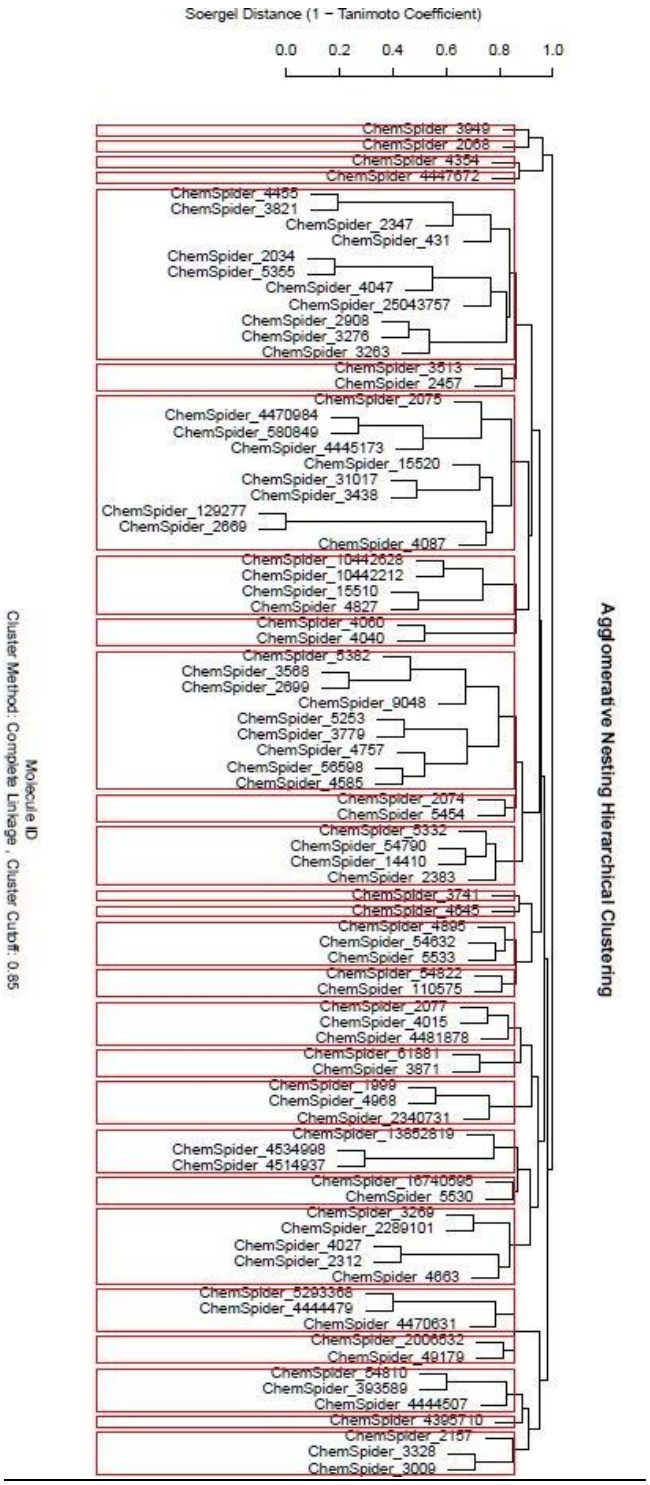
Ο αριθμός των συστάδων που δημιουργήθηκαν ήταν πολύ μεγάλος γι αυτό δοκιμάσαμε διάφορες τιμές για την αποκοπή. Όσο μεγαλύτερη ήταν η τιμή, τόσο μικρότερος ήταν ο αριθμός των συστάδων που δημιουργούνταν.

2η Προσπάθεια:

Αποκοπή: 0.958
Συστάδες που δημιουργήθηκαν: 5

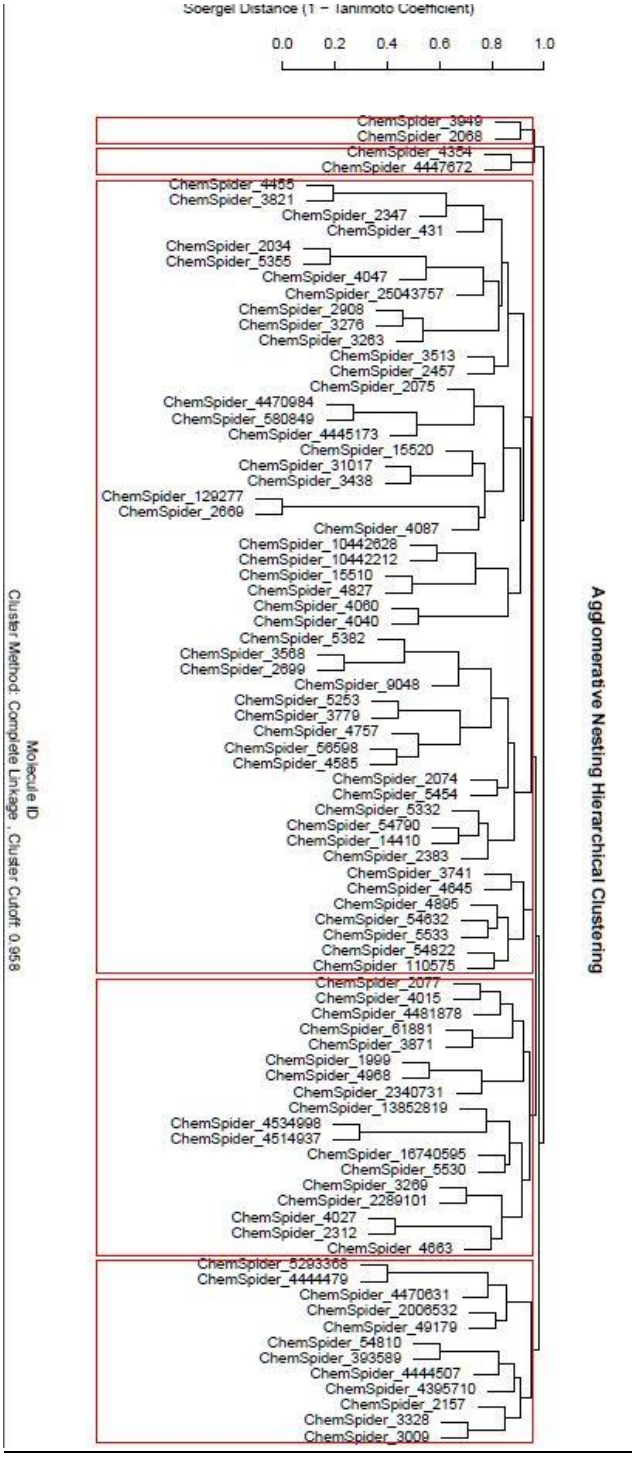
Σ' αυτή την προσπάθεια έγιναν διάφορες δοκιμές με την τιμή της αποκοπής έτσι ώστε να βρω για ποια τιμή σχηματίζονται 5 συστάδες.

Δενδρόγραμμα 1ης προσπάθειας:



Σχήμα 4.1.1 *Δενδρόγραμμα που δείχνει την συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο ANGES με complete linkage και αποκοπή 0.85.*

Δενδρόγραμμα 2^{ης} προσπάθειας:



Σχήμα 4.1.2 *Δενδρόγραμμα που δείχνει την συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο ANGES με complete linkage και αποκοπή 0.958.*

3η Προσπάθεια:

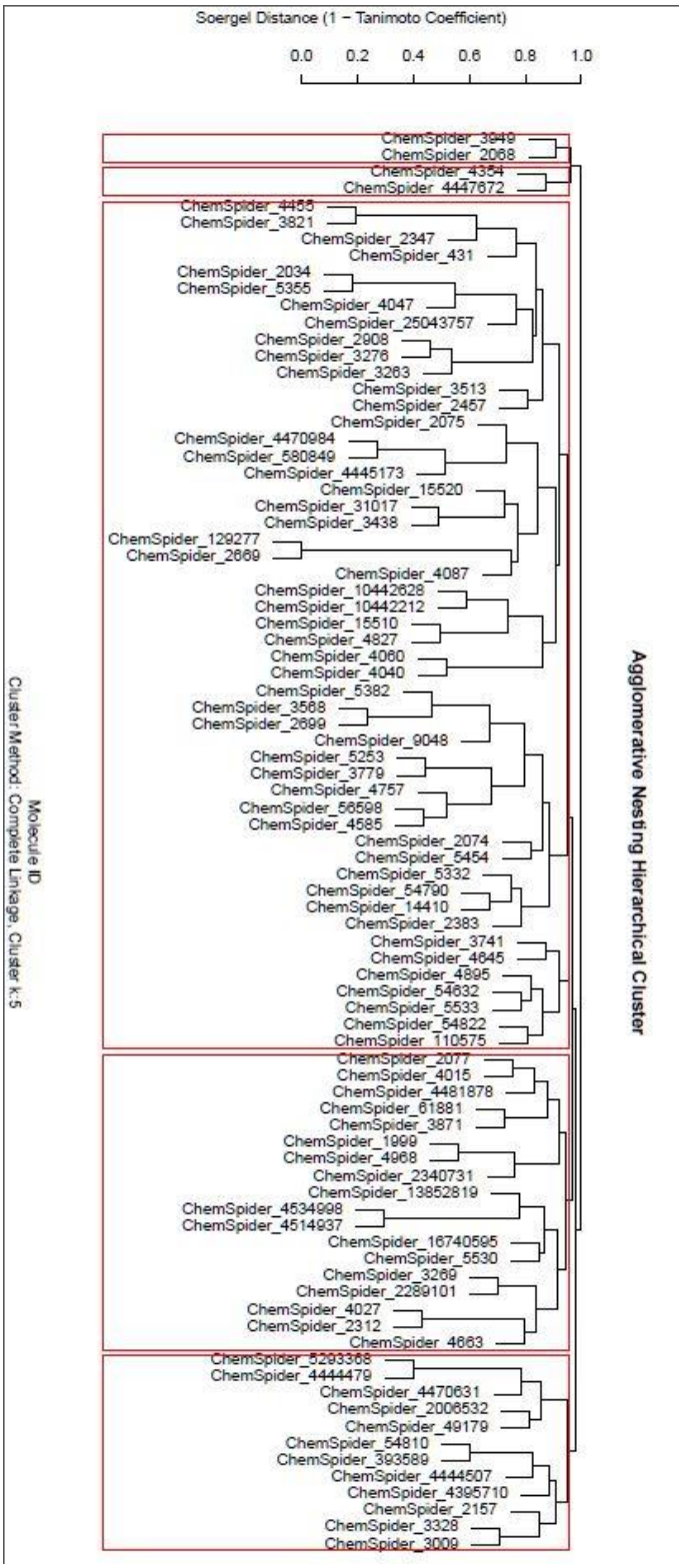
Στην τρίτη προσπάθεια δώσαμε κατευθείαν τον αριθμό των επιθυμητών συστάδων (5) και φυσικά μας επέστρεψε τα ίδια αποτελέσματα με την δεύτερη προσπάθεια.

Πίνακας 4.1 *Συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο AGNES, complete linkage και αριθμό συστάδων ίσο με πέντε.*

Compound ID	Cluster ID	Category ID	Compound ID	Cluster ID	Category ID
ChemSpider_2068	1	1	ChemSpider_5332	2	4.1.4
ChemSpider_3949	1	3	ChemSpider_5355	2	6.3.1
ChemSpider_10442212	2	6.1.6	ChemSpider_5382	2	4.1.1
ChemSpider_10442628	2	6.1.6	ChemSpider_5454	2	4.1.4
ChemSpider_110575	2	1	ChemSpider_54790	2	6.1.9
ChemSpider_129277	2	4.1.2	ChemSpider_54822	2	4.1.4
ChemSpider_14410	2	6.1.3	ChemSpider_5533	2	6.3.2
ChemSpider_15510	2	6.1.6	ChemSpider_56598	2	6.1.2
ChemSpider_15520	2	6.1.5	ChemSpider_580849	2	6.1.4
ChemSpider_2034	2	6.2.1	ChemSpider_9048	2	4.1.1
ChemSpider_2074	2	6.1.7	ChemSpider_2006532	3	2
ChemSpider_2075	2	4.1.1	ChemSpider_2157	3	2
ChemSpider_2347	2	6.2.1	ChemSpider_3009	3	5.2
ChemSpider_2383	2	6.2.2	ChemSpider_3328	3	5.3
ChemSpider_2457	2	5.1	ChemSpider_393589	3	2
ChemSpider_25043757	2	6.2.1	ChemSpider_4395710	3	3
ChemSpider_2669	2	4.1.2	ChemSpider_4444479	3	1
ChemSpider_2699	2	4.1.1	ChemSpider_4444507	3	1
ChemSpider_2908	2	6.2.1	ChemSpider_4470631	3	1

ChemSpider_31017	2	6.1.5	ChemSpider_49179	3	2
ChemSpider_3263	2	6.3.1	ChemSpider_5293368	3	1
ChemSpider_3276	2	6.3.1	ChemSpider_54810	3	2
ChemSpider_3438	2	6.1.3	ChemSpider_13852819	4	4.2.1
ChemSpider_3513	2	2	ChemSpider_16740595	4	3
ChemSpider_3568	2	4.1.1	ChemSpider_1999	4	1
ChemSpider_3741	2	5.3	ChemSpider_2077	4	2
ChemSpider_3779	2	6.1.1	ChemSpider_2289101	4	4.1.4
ChemSpider_3821	2	6.2.1	ChemSpider_2312	4	2
ChemSpider_4040	2	4.1.4	ChemSpider_2340731	4	1
ChemSpider_4047	2	6.3.1	ChemSpider_3269	4	4.1.2
ChemSpider_4060	2	4.1.4	ChemSpider_3871	4	4.1.1
ChemSpider_4087	2	4.1.3	ChemSpider_4015	4	4.2.1
ChemSpider_431	2	4.1.4	ChemSpider_4027	4	2
ChemSpider_4445173	2	6.1.4	ChemSpider_4481878	4	4.1.2
ChemSpider_4455	2	6.2.1	ChemSpider_4514937	4	2
ChemSpider_4470984	2	6.1.4	ChemSpider_4534998	4	2
ChemSpider_4585	2	6.1.2	ChemSpider_4663	4	3
ChemSpider_4645	2	2	ChemSpider_4968	4	1
ChemSpider_4757	2	6.1.1	ChemSpider_5530	4	6.3.2
ChemSpider_4827	2	6.1.6	ChemSpider_61881	4	4.1.2
ChemSpider_4895	2	6.1.9	ChemSpider_4354	5	2
ChemSpider_5253	2	6.1.2	ChemSpider_4447672	5	5.3

Δενδρόγραμμα 3^{ης} προσπάθειας:



Σχήμα 4.1.3 *Δενδρόγραμμα που δείχνει την συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο ANGES με complete linkage και αριθμό συστάδων ίσο με πέντε.*

4.2 AGNES (Ward's Method)

Agglomerative coefficient: 0.69

Οι πίνακες των αποτελεσμάτων για τις 2 πρώτες προσπάθειες βρίσκονται στο Παράρτημα Γ τμήματα Γ.3 και Γ.4 αντίστοιχα.

1η Προσπάθεια:

Αποκοπή: 0.85

Αριθμός συστάδων που δημιουργήθηκαν: 38

Ο αριθμός των συστάδων που δημιουργήθηκαν ήταν πολύ μεγάλος γι αυτό δοκιμάσαμε διάφορες τιμές για την αποκοπή. Όσο μεγαλύτερη ήταν η τιμή, τόσο μικρότερος ήταν ο αριθμός των συστάδων που δημιουργούνταν.

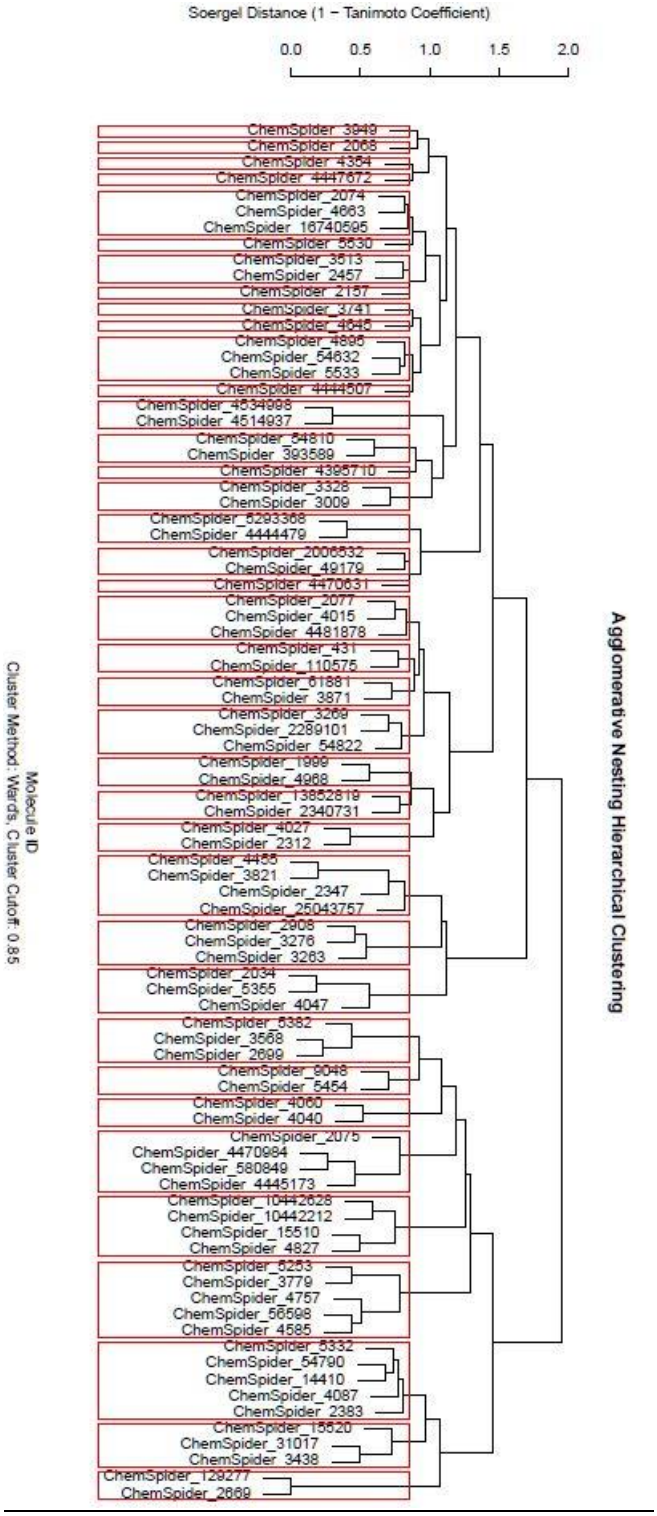
2η Προσπάθεια:

Αποκοπή: 1.37

Αριθμός Συστάδων που δημιουργήθηκαν: 5

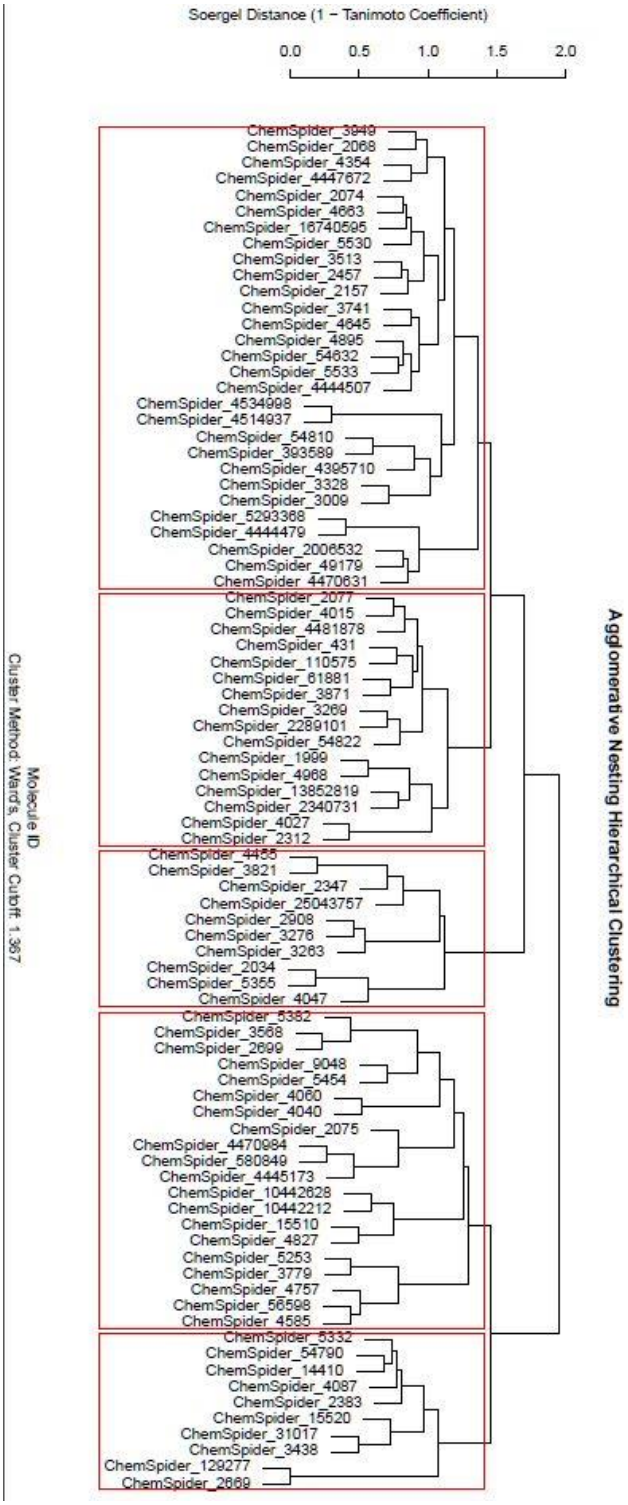
Σ' αυτή την προσπάθεια έγιναν διάφορες δοκιμές με την τιμή της αποκοπής έτσι ώστε να βρούμε για ποια τιμή σχηματίζονται 5 συστάδες.

Δενδρόγραμμα 1ης προσπάθειας:



Σχήμα 4.2.1 *Δενδρόγραμμα που δείχνει την συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο ANGES με Ward's method και αποκοπή 0.85.*

Δενδρόγραμμα 2ης προσπάθειας:



Σχήμα 4.2.2 *Δενδρόγραμμα που δείχνει την συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο ANGES με Ward's method και αποκοπή 1.367*

3η Προσπάθεια:

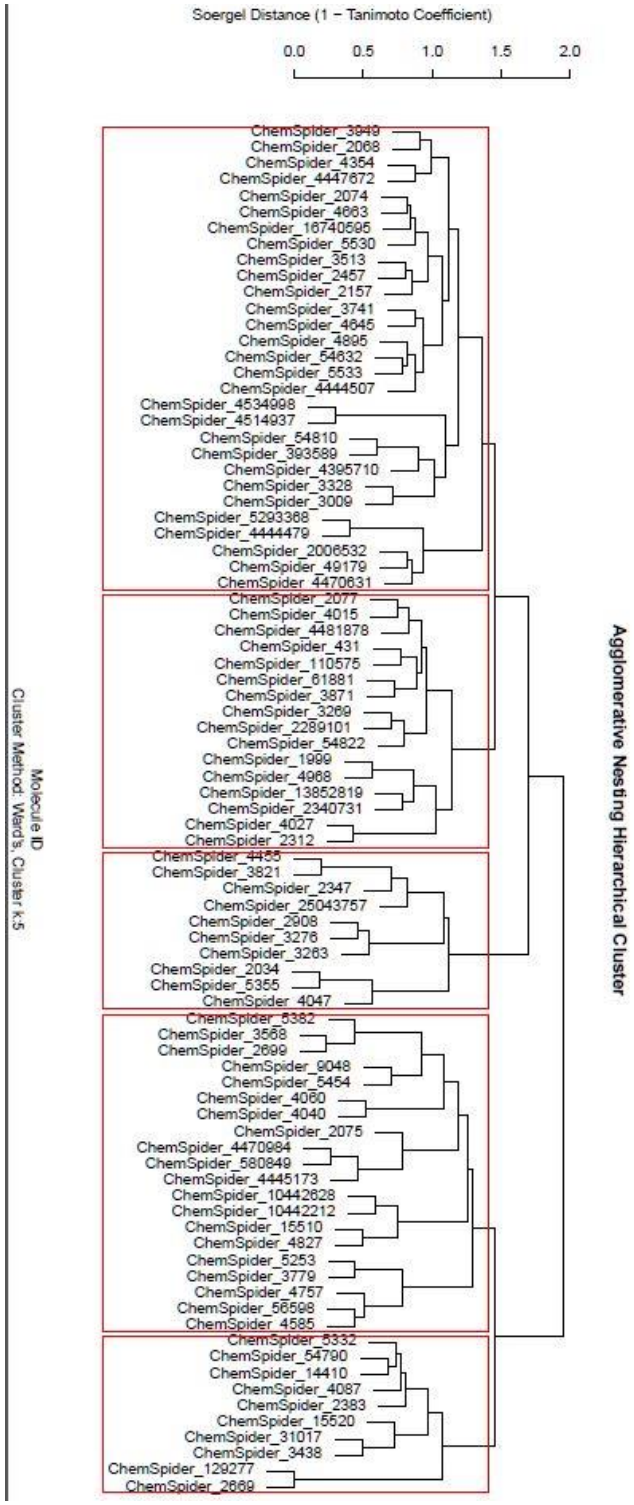
Στην τρίτη προσπάθεια δώσαμε κατευθείαν τον αριθμό των επιθυμητών συστάδων (5) και φυσικά μας επέστρεψε τα ίδια αποτελέσματα με την δεύτερη προσπάθεια.

Πίνακας 4.2 *Συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο AGNES, Ward's method και αριθμό συστάδων ίσο με πέντε.*

ChemSpider ID	Cluster ID	Category ID	ChemSpider ID	Cluster ID	Category ID
ChemSpider_16740595	1	3	ChemSpider_2699	3	4.1.1
ChemSpider_2006532	1	2	ChemSpider_3568	3	4.1.1
ChemSpider_2068	1	1	ChemSpider_3779	3	6.1.1
ChemSpider_2074	1	6.1.7	ChemSpider_4040	3	4.1.4
ChemSpider_2157	1	2	ChemSpider_4060	3	4.1.4
ChemSpider_2457	1	5.1	ChemSpider_4445173	3	6.1.4
ChemSpider_3009	1	5.2	ChemSpider_4470984	3	6.1.4
ChemSpider_3328	1	5.3	ChemSpider_4585	3	6.1.2
ChemSpider_3513	1	2	ChemSpider_4757	3	6.1.1
ChemSpider_3741	1	5.3	ChemSpider_4827	3	6.1.6
ChemSpider_393589	1	2	ChemSpider_5253	3	6.1.2
ChemSpider_3949	1	3	ChemSpider_5382	3	4.1.1
ChemSpider_4354	1	2	ChemSpider_5454	3	4.1.4
ChemSpider_4395710	1	3	ChemSpider_56598	3	6.1.2
ChemSpider_4444479	1	1	ChemSpider_580849	3	6.1.4
ChemSpider_4444507	1	1	ChemSpider_9048	3	4.1.1
ChemSpider_4447672	1	5.3	ChemSpider_110575	4	1
ChemSpider_4470631	1	1	ChemSpider_13852819	4	4.2.1
ChemSpider_4514937	1	2	ChemSpider_1999	4	1

ChemSpider_4534998	1	2	ChemSpider_2077	4	2
ChemSpider_4645	1	2	ChemSpider_2289101	4	4.1.4
ChemSpider_4663	1	3	ChemSpider_2312	4	2
ChemSpider_4895	1	6.1.9	ChemSpider_2340731	4	1
ChemSpider_49179	1	2	ChemSpider_3269	4	4.1.2
ChemSpider_5293368	1	1	ChemSpider_3871	4	4.1.1
ChemSpider_54632	1	2	ChemSpider_4015	4	4.2.1
ChemSpider_54810	1	2	ChemSpider_4027	4	2
ChemSpider_5530	1	6.3.2	ChemSpider_431	4	4.1.4
ChemSpider_5533	1	6.3.2	ChemSpider_4481878	4	4.1.2
ChemSpider_2034	2	6.2.1	ChemSpider_4968	4	1
ChemSpider_2347	2	6.2.1	ChemSpider_54822	4	4.1.4
ChemSpider_25043757	2	6.2.1	ChemSpider_61881	4	4.1.2
ChemSpider_2908	2	6.2.1	ChemSpider_129277	5	4.1.2
ChemSpider_3263	2	6.3.1	ChemSpider_14410	5	6.1.3
ChemSpider_3276	2	6.3.1	ChemSpider_15520	5	6.1.5
ChemSpider_3821	2	6.2.1	ChemSpider_2383	5	6.2.2
ChemSpider_4047	2	6.3.1	ChemSpider_2669	5	4.1.2
ChemSpider_4455	2	6.2.1	ChemSpider_31017	5	6.1.5
ChemSpider_5355	2	6.3.1	ChemSpider_3438	5	6.1.3
ChemSpider_10442212	3	6.1.6	ChemSpider_4087	5	4.1.3
ChemSpider_10442628	3	6.1.6	ChemSpider_5332	5	4.1.4
ChemSpider_15510	3	6.1.6	ChemSpider_54790	5	6.1.9
ChemSpider_2075	3	4.1.1			

Δενδρόγραμμα 3ης προσπάθειας:



Σχήμα 4.2.3 *Δενδρόγραμμα που δείχνει την συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο ANGES με Ward's method και αριθμό συστάδων ίσο με πέντε.*

4.3 DIANA

Divisive Coefficient: 0.38

Οι πίνακες των αποτελεσμάτων για τις 2 πρώτες προσπάθειες βρίσκονται στο Παράρτημα Γ τμήματα Γ.5 και Γ.6 αντίστοιχα.

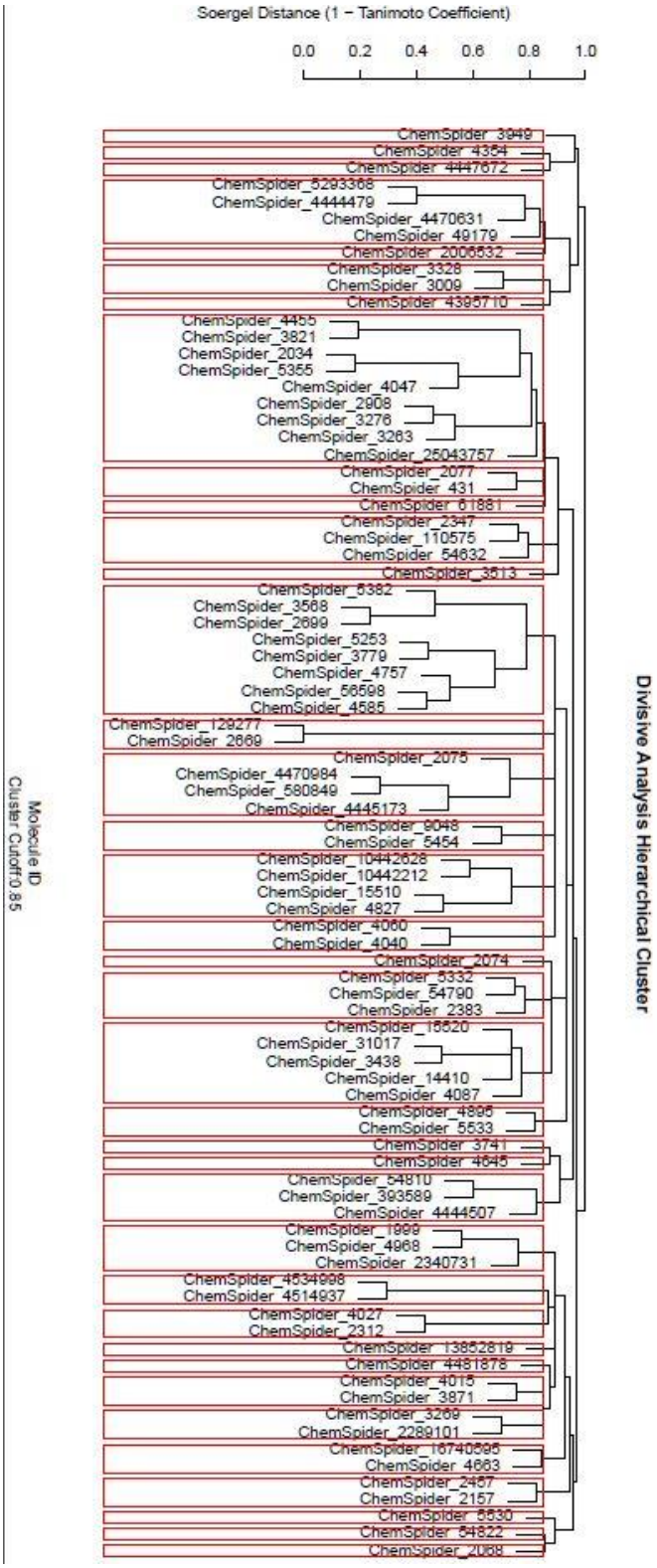
1η Προσπάθεια:

Αποκοπή: 0.85
Συστάδες που δημιοργήθηκαν: 37

2η Προσπάθεια:

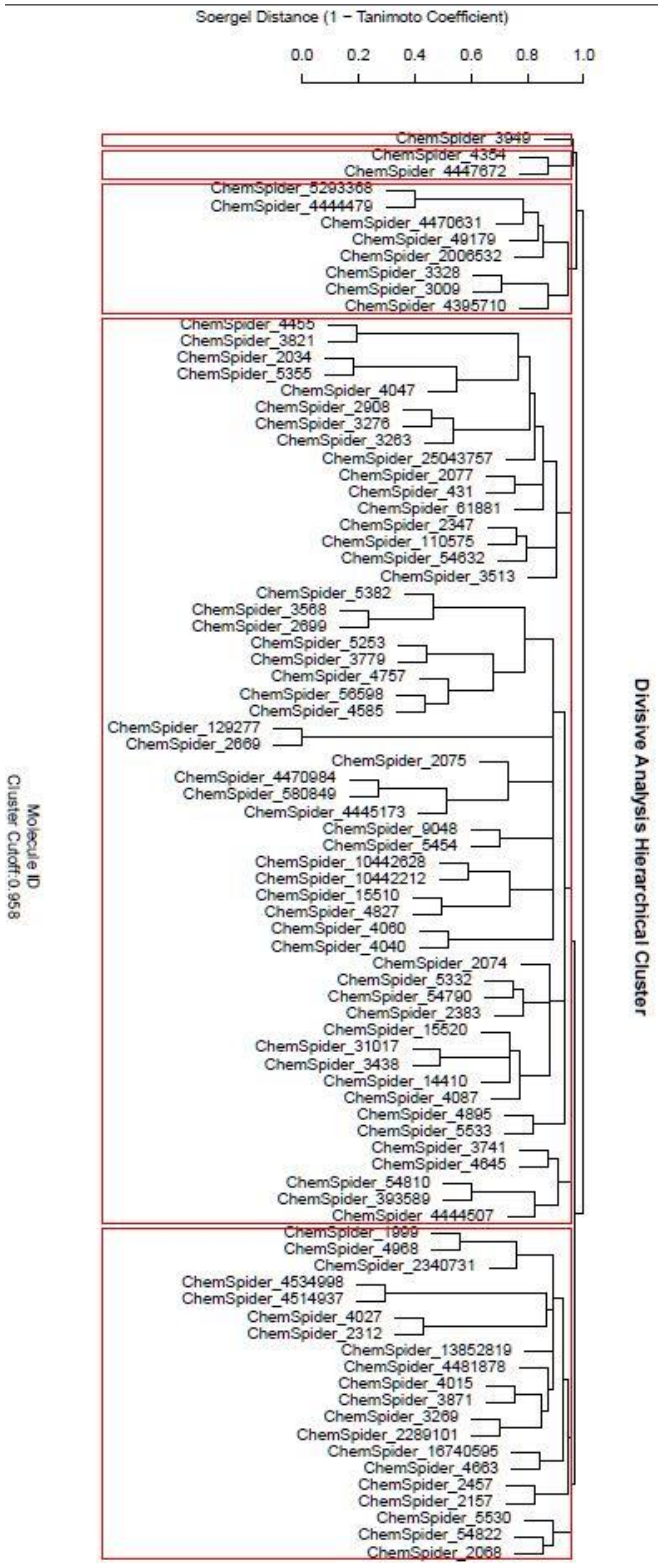
Αποκοπή: 0.958
Αριθμός συστάδων που δημιουργήθηκαν: 5

Δενδρόγραμμα 1ης προσπάθειας:



Σχήμα 4.3.1 *Δενδρόγραμμα που δείχνει την συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο DIANA και αποκοπή ίση με 0.85*

Δενδρόγραμμα 2ης προσπάθειας:



Σχήμα 4.3.2 *Δενδρόγραμμα που δείχνει την συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο DIANA και αποκοπή ίση με 0.958*

3η Προσπάθεια:

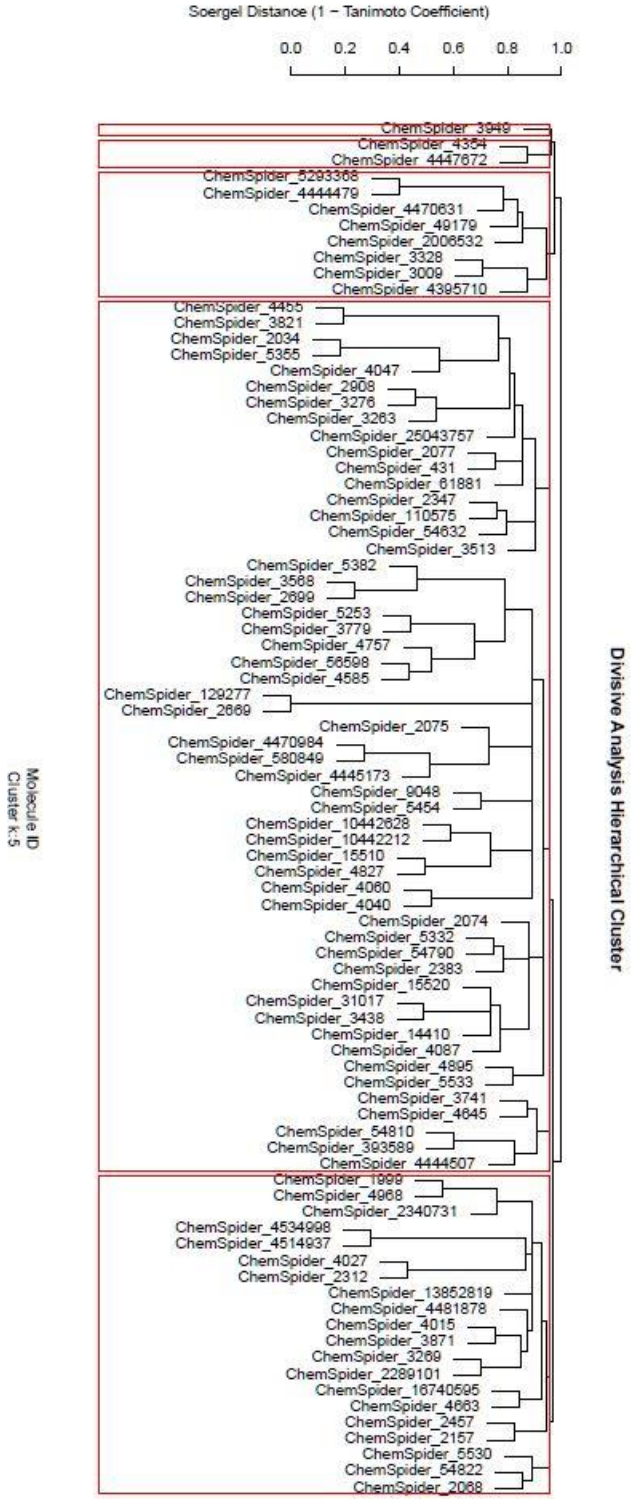
Στην τρίτη προσπάθεια δώσαμε κατευθείαν τον αριθμό των επιθυμητών συστάδων (5) και φυσικά μας επέστρεψε τα ίδια αποτελέσματα με την δεύτερη προσπάθεια.

Πίνακας 4.3 *Συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο DIANA, με αριθμό συστάδων ίσο με πέντε.*

Compound ID	Cluster ID	Category ID	Compound ID	Cluster ID	Category ID
ChemSpider_3949	1	3	ChemSpider_5332	2	4.1.4
ChemSpider_10442212	2	6.1.6	ChemSpider_5355	2	6.3.1
ChemSpider_10442628	2	6.1.6	ChemSpider_5382	2	4.1.1
ChemSpider_110575	2	1	ChemSpider_5454	2	4.1.4
ChemSpider_129277	2	4.1.2	ChemSpider_54632	2	2
ChemSpider_14410	2	6.1.3	ChemSpider_54790	2	6.1.9
ChemSpider_15510	2	6.1.6	ChemSpider_54810	2	2
ChemSpider_15520	2	6.1.5	ChemSpider_5533	2	6.3.2
ChemSpider_2034	2	6.2.1	ChemSpider_56598	2	6.1.2
ChemSpider_2074	2	6.1.7	ChemSpider_580849	2	6.1.4
ChemSpider_2075	2	4.1.1	ChemSpider_61881	2	4.1.2
ChemSpider_2077	2	2	ChemSpider_9048	2	4.1.1
ChemSpider_2347	2	6.2.1	ChemSpider_2006532	3	2
ChemSpider_2383	2	6.2.2	ChemSpider_3009	3	5.2
ChemSpider_25043757	2	6.2.1	ChemSpider_3328	3	5.3
ChemSpider_2669	2	4.1.2	ChemSpider_4395710	3	3
ChemSpider_2699	2	4.1.1	ChemSpider_4444479	3	1
ChemSpider_2908	2	6.2.1	ChemSpider_4470631	3	1
ChemSpider_31017	2	6.1.5	ChemSpider_49179	3	2

ChemSpider_3263	2	6.3.1	ChemSpider_5293368	3	1
ChemSpider_3276	2	6.3.1	ChemSpider_13852819	4	4.2.1
ChemSpider_3438	2	6.1.3	ChemSpider_16740595	4	3
ChemSpider_3513	2	2	ChemSpider_1999	4	1
ChemSpider_3568	2	4.1.1	ChemSpider_2068	4	1
ChemSpider_3741	2	5.3	ChemSpider_2157	4	2
ChemSpider_3779	2	6.1.1	ChemSpider_2289101	4	4.1.4
ChemSpider_3821	2	6.2.1	ChemSpider_2312	4	2
ChemSpider_393589	2	2	ChemSpider_2340731	4	1
ChemSpider_4040	2	4.1.4	ChemSpider_2457	4	5.1
ChemSpider_4047	2	6.3.1	ChemSpider_3269	4	4.1.2
ChemSpider_4060	2	4.1.4	ChemSpider_3871	4	4.1.1
ChemSpider_4087	2	4.1.3	ChemSpider_4015	4	4.2.1
ChemSpider_431	2	4.1.4	ChemSpider_4027	4	2
ChemSpider_4444507	2	1	ChemSpider_4481878	4	4.1.2
ChemSpider_4445173	2	6.1.4	ChemSpider_4514937	4	2
ChemSpider_4455	2	6.2.1	ChemSpider_4534998	4	2
ChemSpider_4470984	2	6.1.4	ChemSpider_4663	4	3
ChemSpider_4585	2	6.1.2	ChemSpider_4968	4	1
ChemSpider_4645	2	2	ChemSpider_54822	4	4.1.4
ChemSpider_4757	2	6.1.1	ChemSpider_5530	4	6.3.2
ChemSpider_4827	2	6.1.6	ChemSpider_4354	5	2
ChemSpider_4895	2	6.1.9	ChemSpider_4447672	5	5.3
ChemSpider_5253	2	6.1.2			

Δενδρόγραμμα 3^{ης} προσπάθειας:



Σχήμα 4.3.3 *Δενδρόγραμμα που δείχνει την συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο DIANA και αριθμό συστάδων ίσο με πέντε.*

4.4 k-means

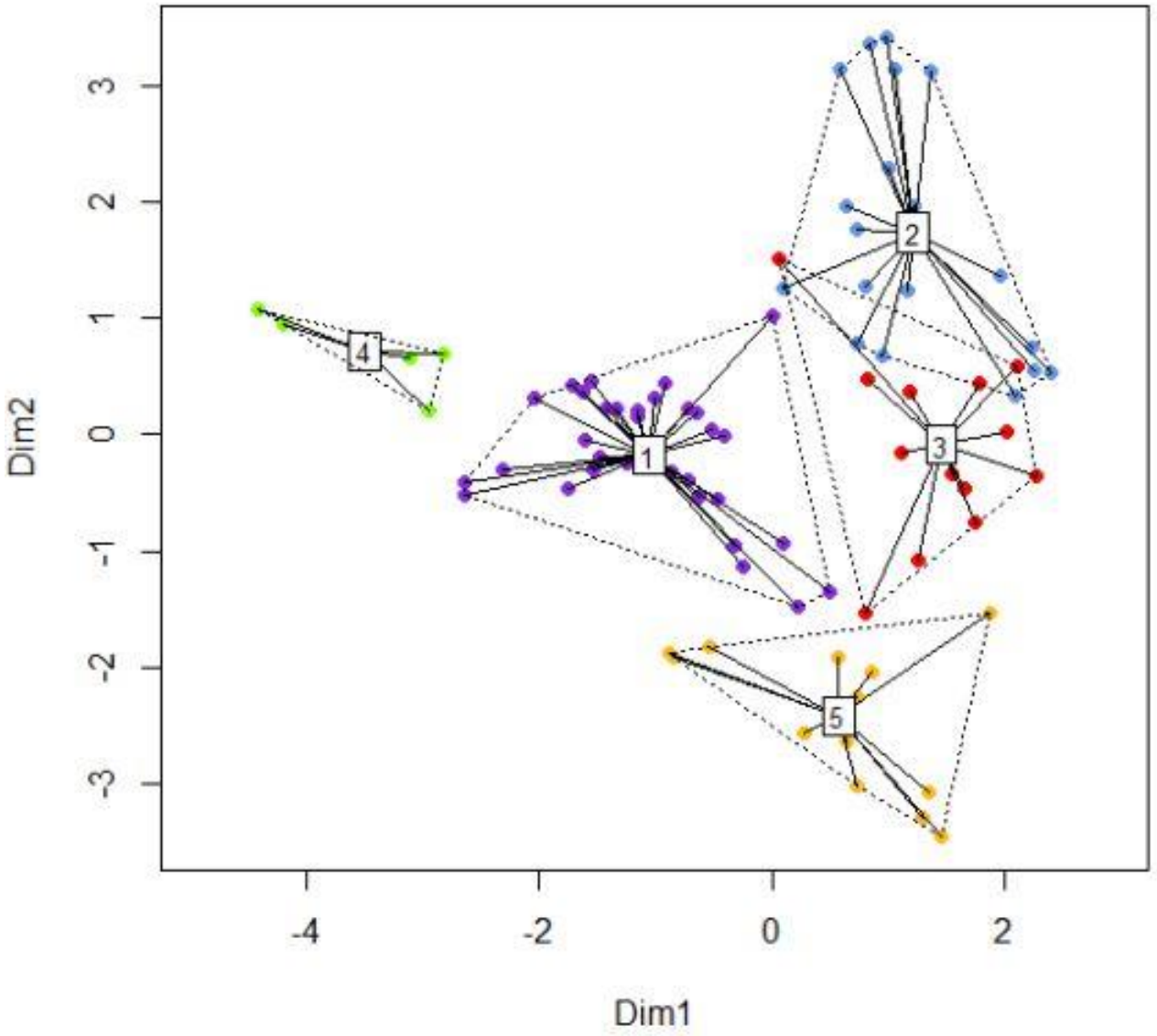
Αριθμός Συστάδων κ: 5
Ολική Τετραγωνική Απόσταση μεταξύ στοιχείων μιας συστάδας: 475.15

Πίνακας 4.4 *Συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο k-means, με αριθμό συστάδων ίσο με πέντε.*

Compound ID	Cluster ID	Category ID	Compound ID	Cluster ID	Category ID
ChemSpider_13852819	1	4.2.1	ChemSpider_4470984	2	6.1.4
ChemSpider_16740595	1	3	ChemSpider_4585	2	6.1.2
ChemSpider_1999	1	1	ChemSpider_4757	2	6.1.1
ChemSpider_2068	1	1	ChemSpider_4827	2	6.1.6
ChemSpider_2074	1	6.1.7	ChemSpider_5253	2	6.1.2
ChemSpider_2077	1	2	ChemSpider_5382	2	4.1.1
ChemSpider_2157	1	2	ChemSpider_5454	2	4.1.4
ChemSpider_2289101	1	4.1.4	ChemSpider_56598	2	6.1.2
ChemSpider_2312	1	2	ChemSpider_580849	2	6.1.4
ChemSpider_2340731	1	1	ChemSpider_9048	2	4.1.1
ChemSpider_2457	1	5.1	ChemSpider_110575	3	1
ChemSpider_3009	1	5.2	ChemSpider_129277	3	4.1.2
ChemSpider_3269	1	4.1.2	ChemSpider_14410	3	6.1.3
ChemSpider_3328	1	5.3	ChemSpider_15520	3	6.1.5
ChemSpider_3513	1	2	ChemSpider_2383	3	6.2.2
ChemSpider_3741	1	5.3	ChemSpider_2669	3	4.1.2
ChemSpider_3871	1	4.1.1	ChemSpider_31017	3	6.1.5
ChemSpider_3949	1	3	ChemSpider_3438	3	6.1.3
ChemSpider_4015	1	4.2.1	ChemSpider_4087	3	4.1.3

ChemSpider_4027	1	2	ChemSpider_4895	3	6.1.9
ChemSpider_431	1	4.1.4	ChemSpider_5332	3	4.1.4
ChemSpider_4354	1	2	ChemSpider_54632	3	2
ChemSpider_4395710	1	3	ChemSpider_54790	3	6.1.9
ChemSpider_4447672	1	5.3	ChemSpider_5533	3	6.3.2
ChemSpider_4481878	1	4.1.2	ChemSpider_2006532	4	2
ChemSpider_4514937	1	2	ChemSpider_4444479	4	1
ChemSpider_4534998	1	2	ChemSpider_4470631	4	1
ChemSpider_4645	1	2	ChemSpider_49179	4	2
ChemSpider_4663	1	3	ChemSpider_5293368	4	1
ChemSpider_4968	1	1	ChemSpider_2034	5	6.2.1
ChemSpider_54822	1	4.1.4	ChemSpider_2347	5	6.2.1
ChemSpider_5530	1	6.3.2	ChemSpider_25043757	5	6.2.1
ChemSpider_61881	1	4.1.2	ChemSpider_2908	5	6.2.1
ChemSpider_10442212	2	6.1.6	ChemSpider_3263	5	6.3.1
ChemSpider_10442628	2	6.1.6	ChemSpider_3276	5	6.3.1
ChemSpider_15510	2	6.1.6	ChemSpider_3821	5	6.2.1
ChemSpider_2075	2	4.1.1	ChemSpider_393589	5	2
ChemSpider_2699	2	4.1.1	ChemSpider_4047	5	6.3.1
ChemSpider_3568	2	4.1.1	ChemSpider_4444507	5	1
ChemSpider_3779	2	6.1.1	ChemSpider_4455	5	6.2.1
ChemSpider_4040	2	4.1.4	ChemSpider_5355	5	6.3.1
ChemSpider_4060	2	4.1.4	ChemSpider_54810	5	2
ChemSpider_4445173	2	6.1.4			

Γραφική αναπαράσταση k-means:



Σχήμα 4.4 *Γραφική αναπαράσταση συσταδοποίησης του συνόλου δεδομένων με τον αλγόριθμο k-means και με αριθμό συστάδων ίσο με πέντε.*

4.5 PAM

average silhouette width: 0.049 , **Αριθμός medoids:** 5
Medoids:

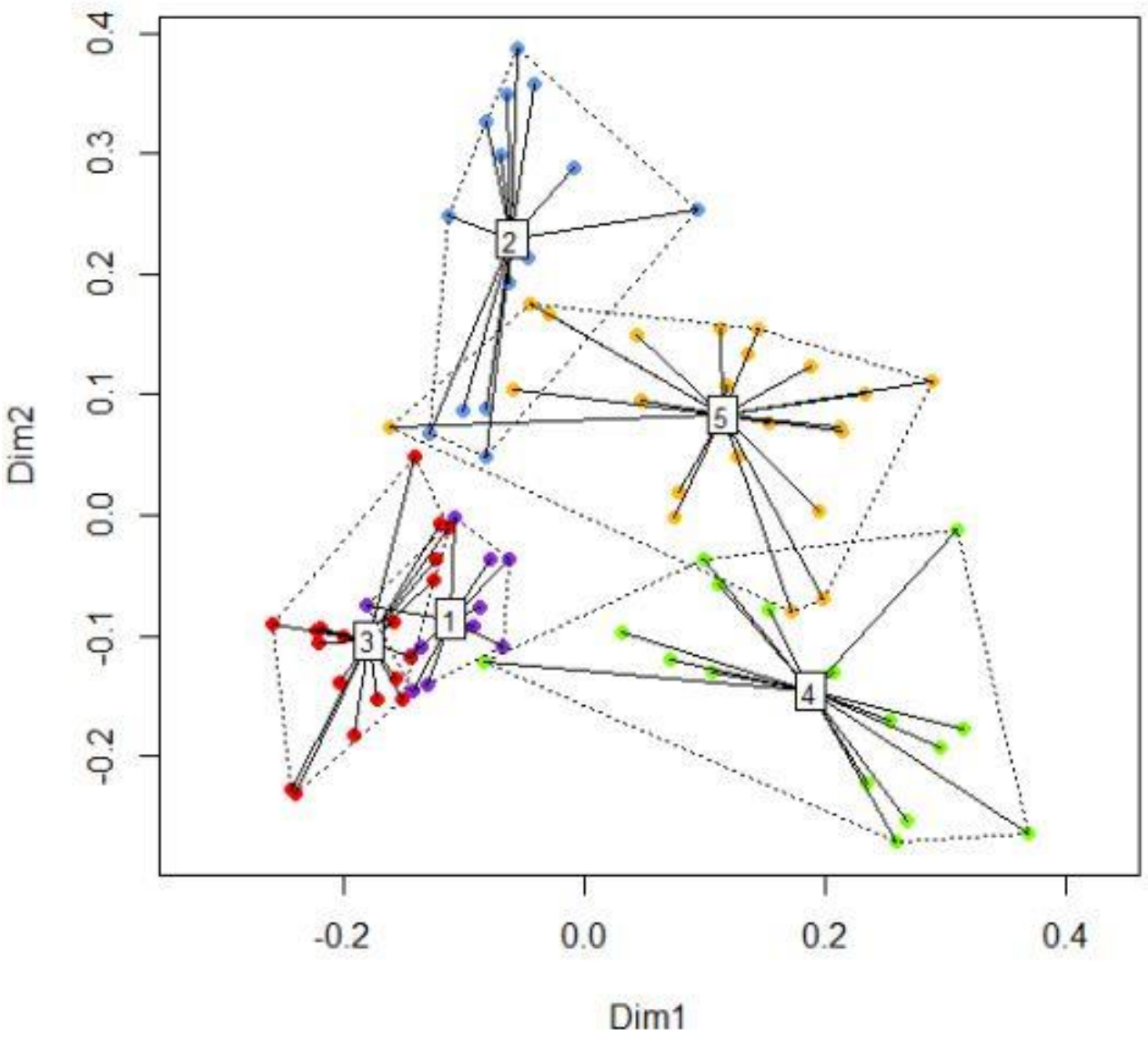
Compound ID	Category ID
ChemSpider_4027	2
ChemSpider_3821	6.2.1
ChemSpider_4534998	2
ChemSpider_3568	4.1.1
ChemSpider_580849	6.1.4

Πίνακας 4.5 *Συσταδοποίηση του συνόλου δεδομένων μας με τον αλγόριθμο PAM, με αριθμό συστάδων ίσο με πέντε.*

Compound ID	Cluster ID	Category ID	Compound ID	Cluster ID	Category ID
ChemSpider_1999	1	1	ChemSpider_4968	3	1
ChemSpider_2068	1	1	ChemSpider_5293368	3	1
ChemSpider_2289101	1	4.1.4	ChemSpider_5530	3	6.3.2
ChemSpider_2312	1	2	ChemSpider_10442212	4	6.1.6
ChemSpider_2340731	1	1	ChemSpider_2074	4	6.1.7
ChemSpider_3269	1	4.1.2	ChemSpider_2157	4	2
ChemSpider_3949	1	3	ChemSpider_2383	4	6.2.2
ChemSpider_4027	**1**	**2**	ChemSpider_2699	4	4.1.1
ChemSpider_4481878	1	4.1.2	**ChemSpider_3568**	**4**	**4.1.1**
ChemSpider_4645	1	2	ChemSpider_3779	4	6.1.1
ChemSpider_4663	1	3	ChemSpider_4040	4	4.1.4
ChemSpider_54822	1	4.1.4	ChemSpider_4060	4	4.1.4
ChemSpider_2034	2	6.2.1	ChemSpider_4585	4	6.1.2
ChemSpider_2077	2	2	ChemSpider_4757	4	6.1.1
ChemSpider_2347	2	6.2.1	ChemSpider_5253	4	6.1.2
ChemSpider_2457	2	5.1	ChemSpider_5382	4	4.1.1
ChemSpider_25043757	2	6.2.1	ChemSpider_5454	4	4.1.4

ChemSpider_2908	2	6.2.1	ChemSpider_56598	4	6.1.2
ChemSpider_3263	2	6.3.1	ChemSpider_9048	4	4.1.1
ChemSpider_3276	2	6.3.1	ChemSpider_10442628	5	6.1.6
ChemSpider_3513	2	2	ChemSpider_110575	5	1
ChemSpider_3821	**2**	**6.2.1**	ChemSpider_129277	5	4.1.2
ChemSpider_4047	2	6.3.1	ChemSpider_14410	5	6.1.3
ChemSpider_431	2	4.1.4	ChemSpider_15510	5	6.1.6
ChemSpider_4455	2	6.2.1	ChemSpider_15520	5	6.1.5
ChemSpider_5355	2	6.3.1	ChemSpider_2075	5	4.1.1
ChemSpider_54810	2	2	ChemSpider_2669	5	4.1.2
ChemSpider_13852819	3	4.2.1	ChemSpider_31017	5	6.1.5
ChemSpider_16740595	3	3	ChemSpider_3438	5	6.1.3
ChemSpider_2006532	3	2	ChemSpider_3741	5	5.3
ChemSpider_3009	3	5.2	ChemSpider_393589	5	2
ChemSpider_3328	3	5.3	ChemSpider_4087	5	4.1.3
ChemSpider_3871	3	4.1.1	ChemSpider_4445173	5	6.1.4
ChemSpider_4015	3	4.2.1	ChemSpider_4470984	5	6.1.4
ChemSpider_4354	3	2	ChemSpider_4827	5	6.1.6
ChemSpider_4395710	3	3	ChemSpider_4895	5	6.1.9
ChemSpider_4444479	3	1	ChemSpider_5332	5	4.1.4
ChemSpider_4444507	3	1	ChemSpider_54632	5	2
ChemSpider_4447672	3	5.3	ChemSpider_54790	5	6.1.9
ChemSpider_4470631	3	1	ChemSpider_5533	5	6.3.2
ChemSpider_4514937	3	2	**ChemSpider_580849**	**5**	**6.1.4**
ChemSpider_4534998	**3**	**2**	ChemSpider_61881	5	4.1.2
ChemSpider_49179	3	2			

Γραφική Αναπαράσταση PAM:



Σχήμα 4.5 *Γραφική αναπαράσταση συσταδοποίησης του συνόλου δεδομένων με τον αλγόριθμο PAM και με αριθμό συστάδων ίσο με πέντε.*

4.6 Κύρια Χαρακτηριστικά Συστάδων

Επιλέξαμε τις 5 καλύτερες συσταδοποιήσεις και βρήκαμε με την βοήθεια γραφικών παραστάσεων, ποια χαρακτηριστικά (από τα 1024) είχαν τα περισσότερα compounds (>=50% των compounds) σε κάθε συστάδα. Τα αποτελέσματα παρουσιάζονται στον πιο κάτω πίνακα:
Οι γραφικές παραστάσεις των αποτελεσμάτων βρίσκονται στο Παράρτημα Δ

Πίνακας 4.6 *Κοινά χαρακτηριστικά compounds που βρίσκονται στην ίδια συστάδα*

ΑΛΓΟΡΙΘΜΟΣ	ΣΥΣΤΑΔΑ	ΜΕΓΕΘΟΣ	ΚΟΙΝΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ
AGNES (Complete)	1	2	F1, F3, F6, F412
AGNES (Complete)	2	51	F1, F3, F5, F19, F366, F533, F547, F599, F700, F793, F823, F826, F892, F944, F1017
AGNES (Complete)	3	12	F2, F3, F4, F5, F26, F34, F63, F351, F501, F628, F646, F665, F794, F828, F859, F928, F1018, F1019
AGNES (Complete)	4	18	F1
AGNES (Complete)	5	2	F1, F2, F613, F651, F807
AGNES (Ward's)	1	29	F1, F2, F3, F5, F62, F429, F793, F994, F1017
AGNES (Ward's)	2	10	F1,F3, F5, F9, F52, F86, F153, F350, F429, F499, F533, F571, F599, F700, F826, F892, F937, F994, F1017, F1018
AGNES (Ward's)	3	20	F1, F3, F5, F19, F60, F212, F429, F533, F547, F522, F599, F823, F994, F1017
AGNES (Ward's)	4	16	F1, F3, F5, F20, F371, F429, F533, F599, F793, F822, F994, F1017
AGNES (Ward's)	5	10	F1, F3, F5, F9, F19, F224, F277, F350, F366, F429, F533, F547, F567, F672, F700, F793, F823, F826, F892, F921, F994, F1017
DIANA	1	1	-

DIANA	2	54	F1, F3, F5, F9,F19, F350, F429, F533, F547, F599, F700, F793, F823, F892, F997, F1017
DIANA	3	8	F1, F3, F4, F62, F380, F500, F627, F645, F664, F708, F758, F793, F813, F827, F858, F1017, F1018
DIANA	4	20	F1, F3, F5, F20, F62, F371, F429, F533, F599, F822, F827, F994, F1017
DIANA	5	2	F1, F2, F3, F613, F651, F807
k-means	1	33	F1, F3, F5,F62, F371, F429, F533, F599, F822, F994, F1017
k-means	2	20	F1, F3, F5, F19, F60, F212, F429, F533, F547, F552, F599, F823, F994, F1017
k-means	3	14	F1, F3, F5, F9, F19, F350, F366, F429, F533, F547, F567, F672, F700, F793, F723, F826, F892, F1017
k-means	4	5	F1, F3,F4, F62, F114, F380, F419, F500, F627, F645, F664, F668, F708, F714, F758, F793, F813, F815, F817, F827, F858, F912, F1017, F1018
k-means	5	13	F1, F3, F5, F7, F9, F62, F52, F86, F153, F350, F429, F499, F533, F571, F599, F700, F892, F937, F994, F1017, F1018
PAM	1	12	F1, F3, F5, F20, F60, F62, F371, F429, F533, F807, F822, F967, F994, F1017
PAM	2	15	F1,F3, F5, F9, F52, F86, F153, F350, F429, F499, F533, F571, F599, F700, F709, F826, F892, F937, F994, F1017, F1018
PAM	3	19	F1, F3, F5, F62, F793, F827, F1017
PAM	4	16	F1, F3, F5, F19, F60, F62, F212, F366, F392, F429, F547, F552, F599, F793, F807, F823, F994, F1017, F1022
PAM	5	23	F1, F3, F5, F7, F9, F19, F350, F366, F429, F533, F547, F599, F672, F700, F793, F823, F826, F892, F994, F1017

4.7 Αξιολόγηση Συσταδοποίησης

Υπάρχουν διάφορες μετρικές για την διατίμηση και την επικύρωση των αποτελεσμάτων αλλά εδώ θα επικεντρωθούμε στις ***εσωτερικές μετρήσεις*** και σε μια ξεχωριστή εκδοχή τους τις ***μετρήσεις σταθερότητας.***

4.7.1 Εσωτερικές Μετρήσεις

Οι εσωτερικές μετρήσεις (internal measures) χρησιμοποιούν μόνο το σύνολο δεδομένων και τη συσταδοποίηση του σαν είσοδο και χρησιμοποιούν εγγενείς πληροφορίες για να εκτιμήσουν την ποιότητα των δεδομένων ως προς την συμπύκνωση (compactness), τη διαχωρισιμότητα (separation)και την συνεκτικότητα (connectedness). Η συνεκτικότα σχετίζεται με το αν τα στοιχεία είναι τοποθετημένα στην ίδια συστάδα με τους κοντινότερους «γείτονες» τους και αυτό μετριέται με την συνδεσιμότητα (**connectivity**). Η συμπύκνωση εκτιμά την ομοιογένεια μιας συστάδας χρησιμοποιώντας την εσωτερική διασπορά ενώ η διαχωρισιμότητα ποσοτικοποιεί τον βαθμό διαχωρισμού μεταξύ των συστάδων (συνήθως μετρώντας την απόσταση μεταξύ των κέντρων των συστάδων). Η συμπύκνωση και η διαχωρισιμότητα αντιπροσωπεύουν αντίθετες τάσεις καθώς η συμπύκνωση αυξάνεται με τον αριθμό των συστάδων ενώ η διαχωρισιμότητα μειώνεται. Οι σύγχρονες μεθόδοι όπως είναι ο δείκτης Dunn (**Dunn index**) και το πλάτος σιλουέτας (**silhouette width**) συνδυάζουν αυτές τις δύο μετρήσεις [18].

Συνδεσιμότητα:

Ας υποθέσουμε ότι το N είναι ο συνολικός αριθμός των στοιχείων (γραμμές) σε ένα σύνολο δεδομένων και M είναι ο ο αριθμός των χαρακτηριστικών (στήλες) τους. Ορίζουμε το $nn_i(j)$ ως τοω j-οστό κοντινότερο γείτονα του στοιχείου i, και δίνουμε στο $x_{i,nn_{i(j)}}$ την τιμή 0 αν το i και j βρίσκονται στην ίδια συστάδα ή αλλιώς την τιμή $\frac{1}{j}$. Τότε η ***σύνδεση*** της συσταδοποίηση $\mathrm{C}= \{C_1, C_2, \dots, C_k\}$ με N στοιχεία και k συστάδες ορίζεται ως εξής:

$$Conn(\mathrm{C}) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}}$$

Όπου L είναι η παράμετρος που δίνει τον αριθμό των κοντινότερων γειτόνων. Η τιμές που παίρνει η συνδεσιμότητα κυμαίνονται αππό το 0 μέχρι το ∞ και ιδανικά η τιμή της θα πρέπει να είναι όσο το δυνατόν μικρότερη [18].

Πλάτος σιλουέτας:

Το πλάτος σιλουέτας είναι ο μέσος όρος της κάθε τιμής της σιλουέτας όλω των στοιχείων και μετράει τον βαθμό εμπιστοσύνης της ανάθεσης ενός στοιχείου σε μία συστάδα. Ορίζεται ως εξής [39]:

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)}$$

Όπου:

$\boldsymbol{a_i}$ η μέση απόσταση μεταξύ του στοιχείου i και των άλλων στοιχείων της συστάδας στην οποία ανήκει

$\boldsymbol{b_i}$ η μέση απόσταση μεταξύ του στοιχείου i και των στοιχείων της συστάδας που αποτελεί τον κοντινότερο γείτονα του i. Ο κοντινότερος γείτονας του i υπολογίζεται ως εξής:

1. Ορίζουμε C μια οποιαδήποτε συστάδα και d(i,C) την μέση απόσταση του i από όλα τα στοιχεία της συστάδας C.
2. Υπολογίζουμε το d(i,C) για όλες τις συστάδες εκτός αυτής που ανήκει το i και βρίσκουμε αυτή τη συστάδα για την οποία το d(i,C) είναι ελάχιστο. Αυτή η συστάδα αποτελεί τον κοντινότερο γείτονα του i.

Το s(i) παίρνει τιμές από το 1 μέχρι το -1 και μεταφράζεται ως ακολούθως:

s(i)=1 το $\boldsymbol{a_i}$ είναι πολύ μικρότερο από το$\boldsymbol{b_i}$ άρα το i έχει ανατεθεί στη σωστή συστάδα. Δηλαδή η δεύτερη καλύτερη συστάδα (ο κοντινότερος γείτονας του i) δεν είναι τόσο καλή όσο η συστάδα που έχει ανατεθεί το i.

s(i)=0 το $\boldsymbol{a_i}$ κσι το $\boldsymbol{b_i}$ είναι περίπου ίσα. Άρα δεν είναι ξεκάθαρο κατά πόσο το i θα έπρεπε να ανατεθεί στον κοντινότερο του γείτονα ή στη συστάδα που βρίκεται**.**

s(i)=-1 το i έχει λανθασμένα ανατεθεί στην συστάδα στην οποία βρίσκεται καθώς το $\boldsymbol{a_i}$ είναι πολύ μεγαλύτερο από το$\boldsymbol{b_i}$ γεγονός που δείχνει ότι η μέση απόσταση του στοιχείου i από τα στοιχεία της δικής του συστάδας είναι μεγαλύτερη από τα στοιχεία του κοντινότερου του γείτονα.

Δείκτης Dunn

Ο δείκτης Dunn είναι ο λόγος της μικρότερης απόστασης μεταξύ των στοιχείων που δεν ανήκουν στην ίδια συστάδα προς την μεγαλύτερη απόσταση μεταξύ των στοιχείων που ανήκουν στην ίδια συστάδα. Υπολογίζεται ως εξής [17]:

$$D(\mathrm{C}) = \frac{\min\limits_{C_k, C_l \in \mathrm{C}, C_k \neq C_l} \left(\min\limits_{i \in C_k, j \in C_l} dist(i,j) \right)}{\max\limits_{C_m \in \mathrm{C}} diam(C_m)}$$

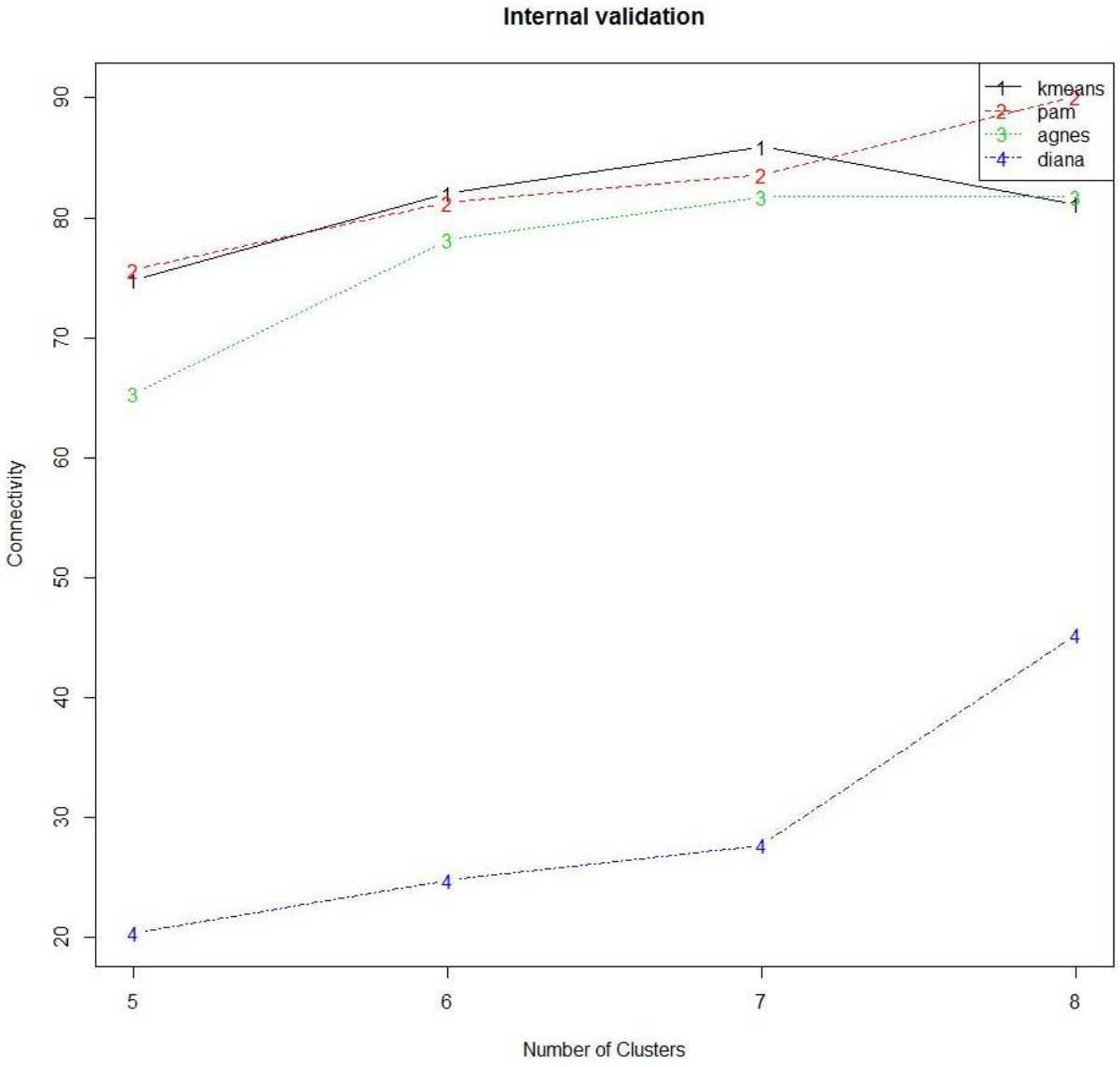
Όπου:

$diam(C_m)$ η μέγιστη απόσταση μεταξύ των στοιχείων της συστάδας C_m

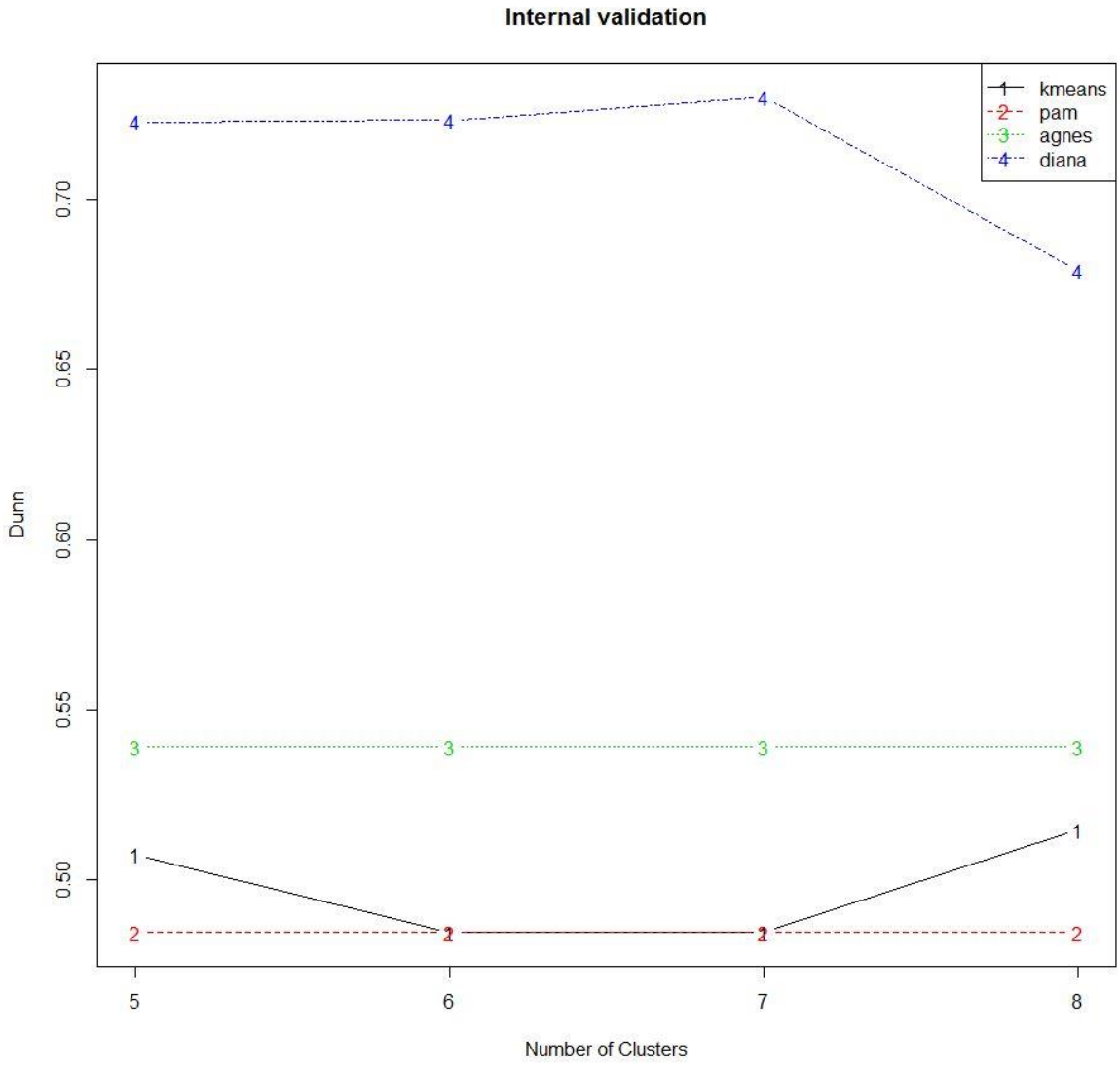
Παίρνει τιμές από μηδέν μέχρι ∞ και όσο μεγαλύτερη είναι η τιμή που παίρνει, τόσο καλύτερη είναι η συσταδοποίηση.

Γραφική Αναπαράσταση απόδοσης αλγορίθμων ως προς τις εσωτερικές μετρήσεις

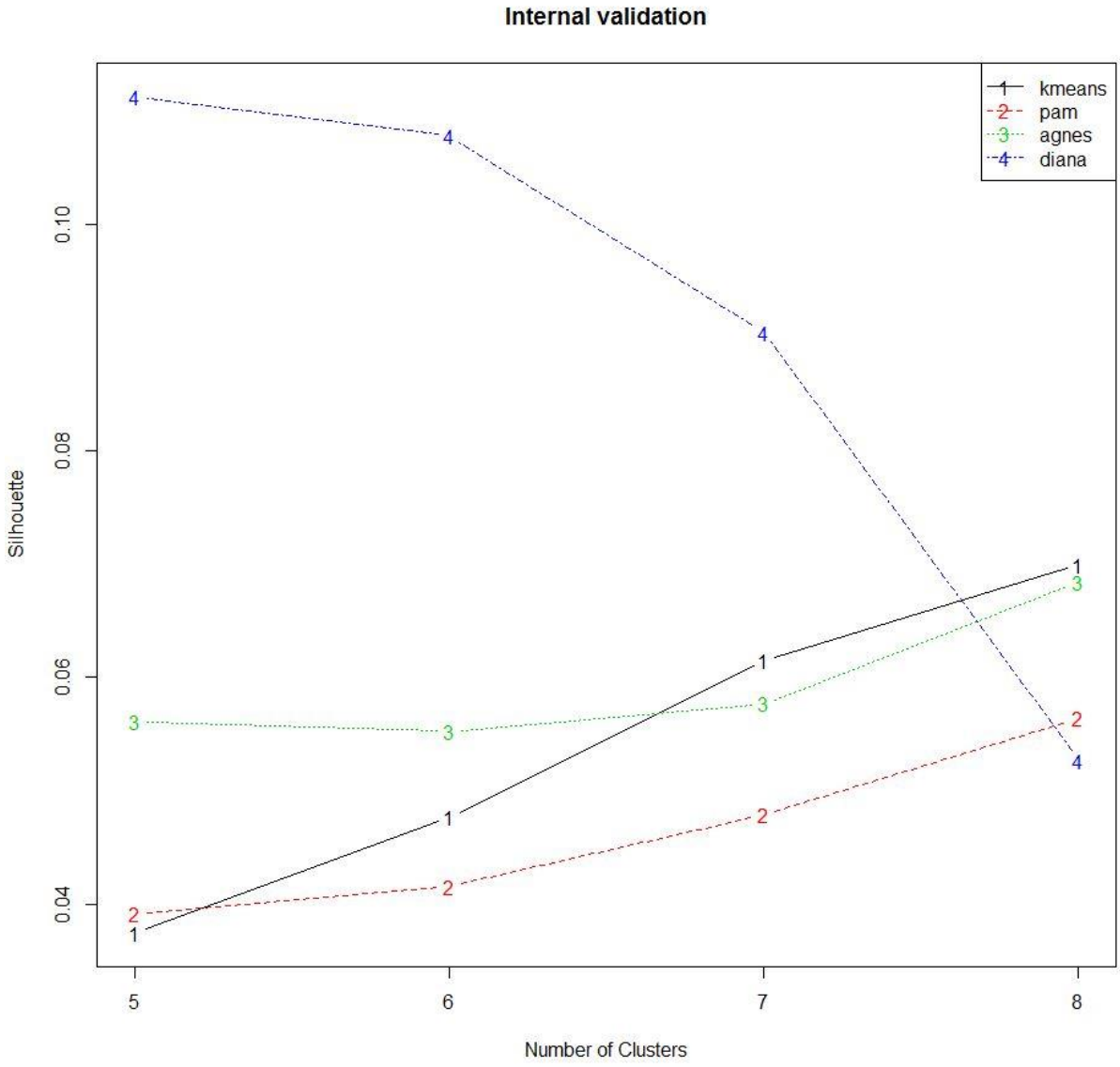
Τρέχοντας κάποια R scripts (*Παράρτημα Β τμήμα Β.6*) βλέπουμε γραφικά πως ο καθένας από τους τέσσερις αλγόριθμους αποδίδει ως προς τις τρεις προαναφερθείσες εσωτερικές μετρήσεις για 5, 6, 7 και 8 συστάδες.



Σχήμα 4.7.1.1 *Γραφική αναπαράσταση της απόδοσης των 4 αλγορίθμων ως προς την συνδεσιμότητα για αριθμό συστάδων από 5 μέχρι 8.*

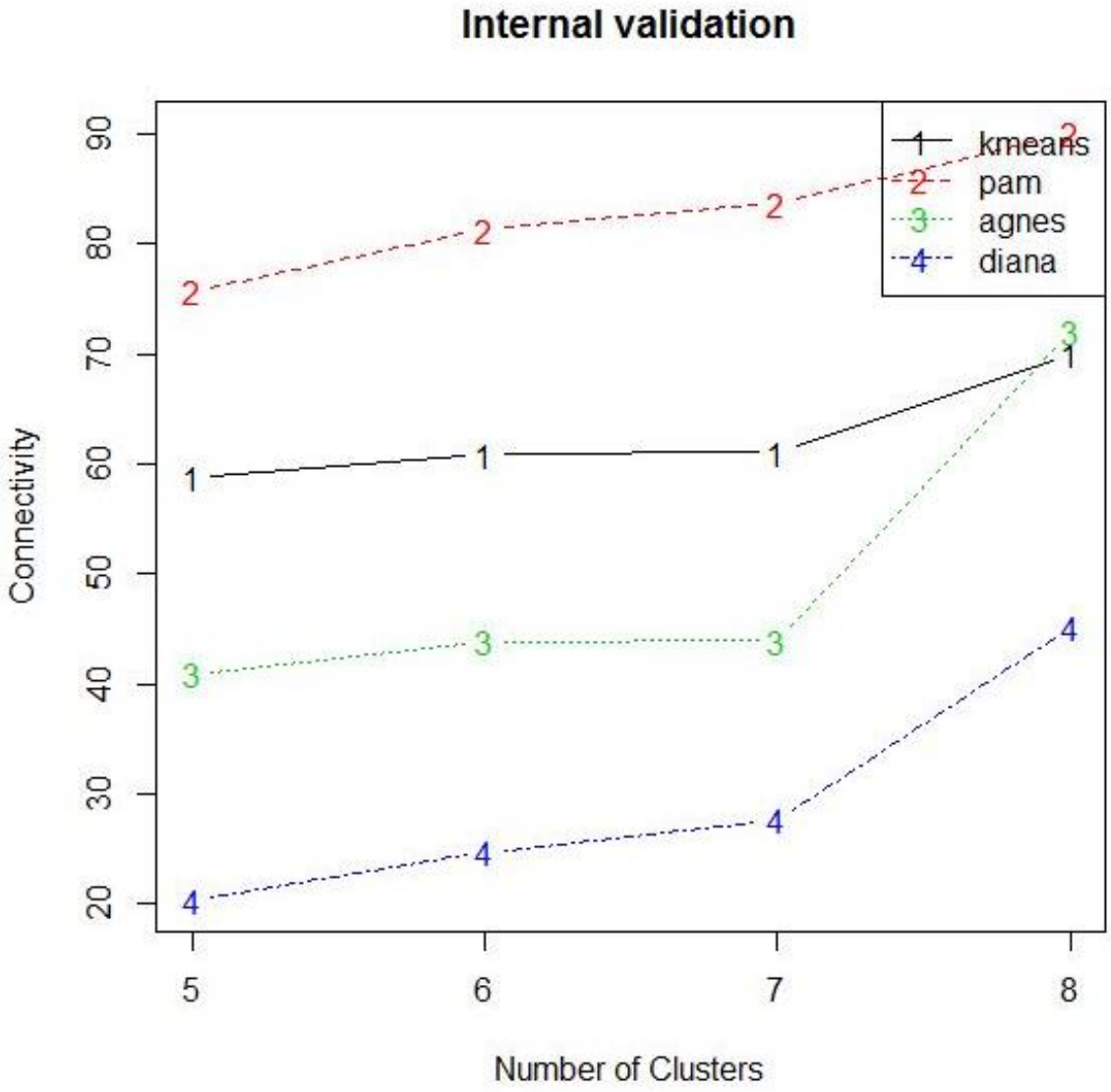


Σχήμα 4.7.1.2 *Γραφική αναπαράσταση της απόδοσης των 4 αλγορίθμων ως προς τον δείκτη Dunn για αριθμό συστάδων από 5 μέχρι 8.*

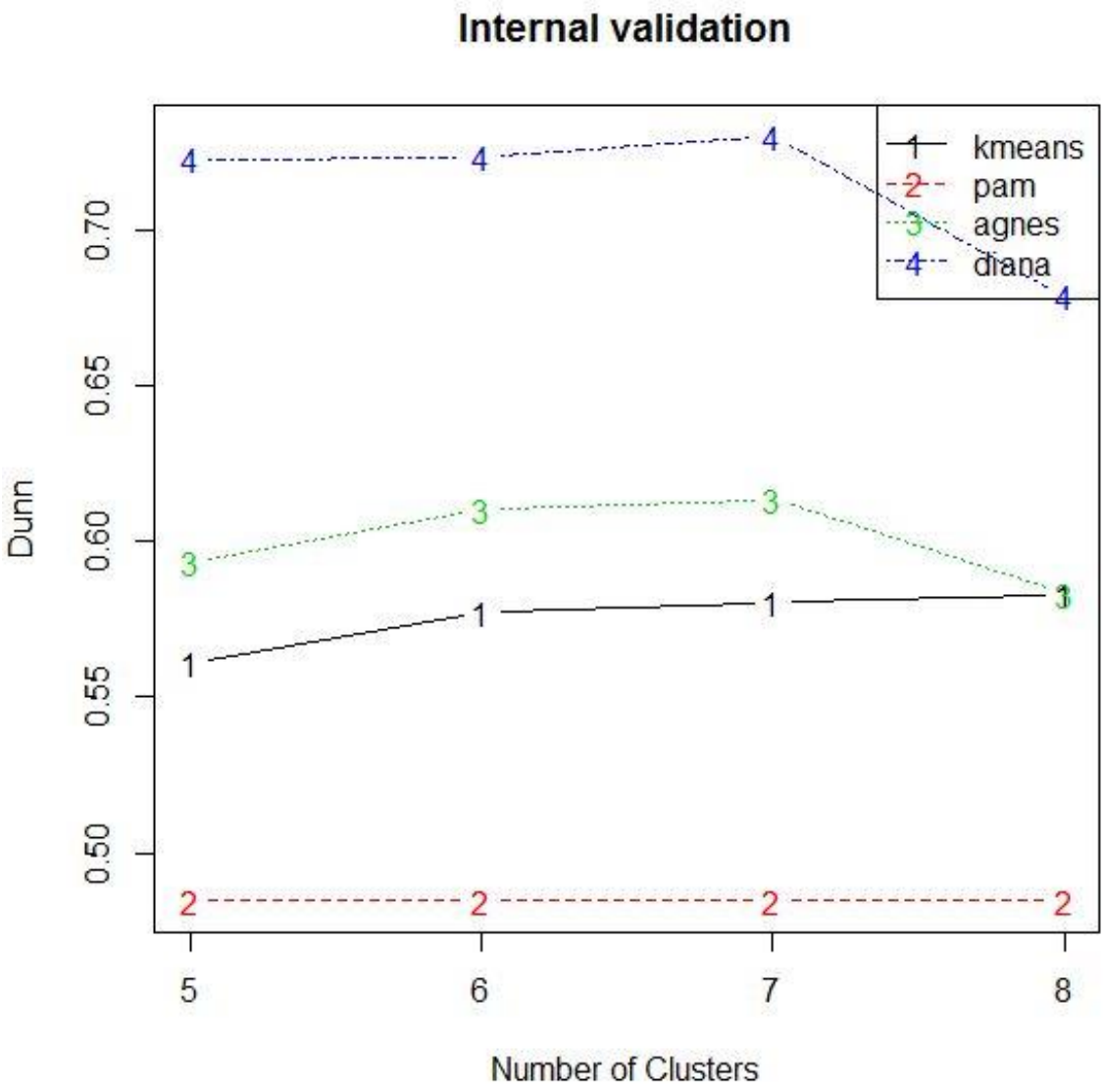


Σχήμα 4.7.1.3 *Γραφική αναπαράσταση της απόδοσης των 4 αλγορίθμων ως προς το πλάτος σιλουέτας για αριθμό συστάδων από 5 μέχρι 8.*

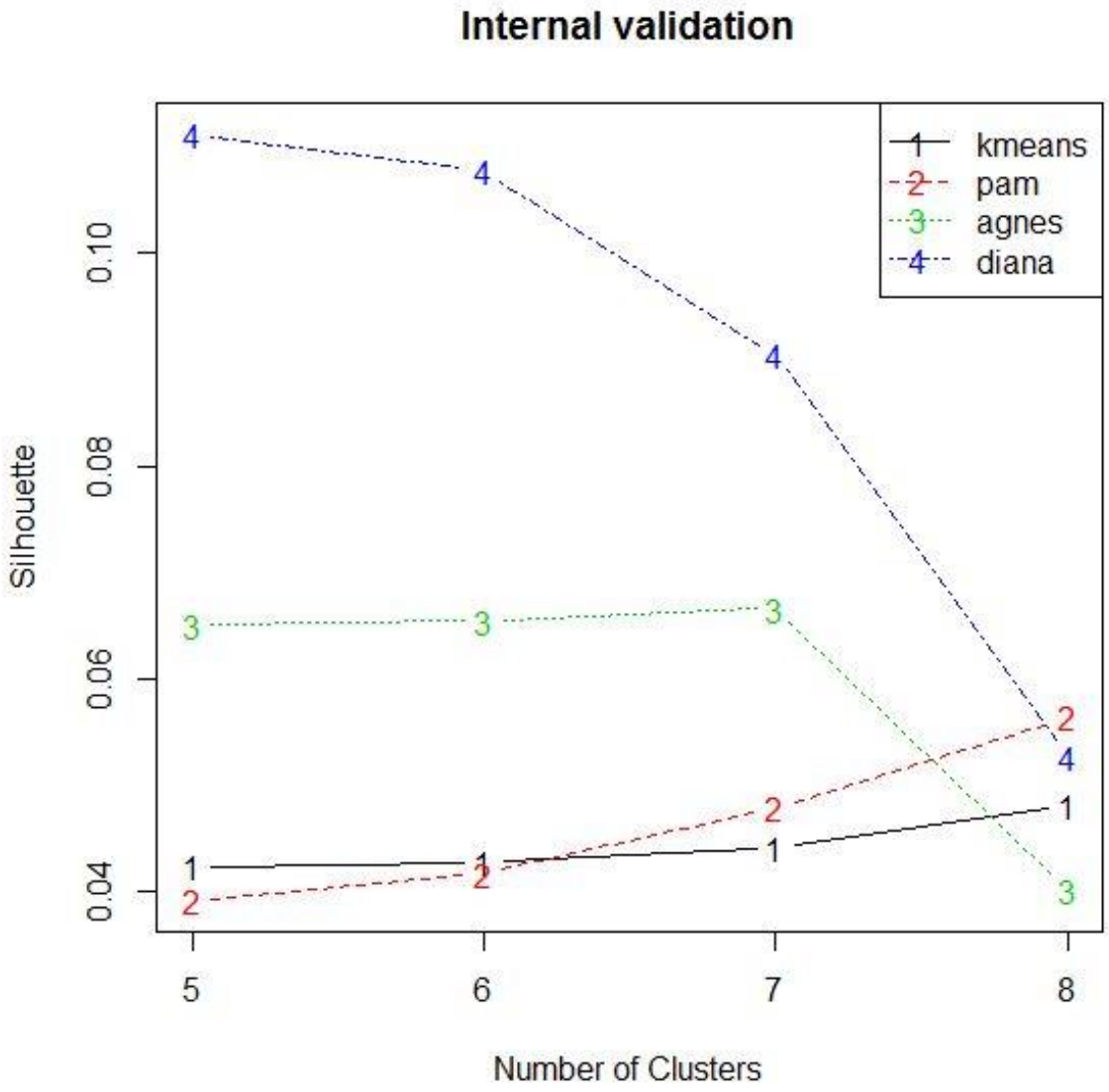
Όταν η σύγκριση αυτή γίνεται με τον αλγόριθμο AGNES (complete) η σειρά απόδοσης στις τρεις μετρήσεις δεν αλλάζει, απλώς αλλάζουν οι τιμές που παίρνει ο αλγόριθμος AGNES στις τρεις αυτές μετρήσεις όπως φαίνεται στις πιο κάτω γραφικές παραστάσεις: Έχοντας αυτά τα στοιχεία μπορούμε να καταλήξουμε στα ακόλουθα συμπεράσματα:



Σχήμα 4.7.1.4 *Γραφική αναπαράσταση της απόδοσης των 4 αλγορίθμων ως προς την συνδεσιμότητα για αριθμό συστάδων από 5 μέχρι 8.*



Σχήμα 4.7.1.5 *Γραφική αναπαράσταση της απόδοσης των 4 αλγορίθμων ως προς τον δείκτη Dunn για αριθμό συστάδων από 5 μέχρι 8.*



Σχήμα 4.7.1.6 *Γραφική αναπαράσταση της απόδοσης των 4 αλγορίθμων ως προς το πλάτος σιλουέτας για αριθμό συστάδων από 5 μέχρι 8.*

4.7.2 Μετρήσεις Σταθερότητας

Οι μετρήσεις σταθερότητας (stability measures) αξιολογούν την συνοχή της συσταδοποίησης συγκρίνοντας τα αποτελέσματα από την συσταδοποίηση ολόκληρου του συνόλου δεδομένων με τις συσταδοποιήσεις του συνόλου δεδομένων αφαιρώντας μία στήλη κάθε φορά. Αυτές οι μετρήσεις λειτουργούν πολύ καλά σε περιπτώσεις όπου τα δεδομένα έχουν υψηλή συσχέτιση. Οι μετρήσεις αυτές είναι: το μέσο ποσοστό μη επικάλυψης (**APN** – Average Proportion of Non-overlap), η μέση απόσταση (**AD** – Average Distance), η μέση απόσταση μεταξύ των μέσων (**ADM** – Average Distance between Means) και ο συντελεστής κέρδους (**FOM** – Figure of Merit).
Όσο πιο μικρές τιμές έχουν οι πιο πάνω μετρήσεις τόσο πιο καλή είναι η συσταδοποίηση που έχει γίνει. Επειδή όμως για να ληφθούν αυτές οι μετρήσεις πρέπει να γίνει συσταδοποίηση του συνόλου δεδομένων μας πολλές φορές (μία για κάθε στήλη που αφαιρείται), απαιτείται μεγάλη υπολογιστική ισχύς. Γι αυτό τον λόγο δεν ήταν δυνατή η λήψη αυτών των μετρήσεων για τους σκοπούς της εργασίας αυτής [18].

4.7.3 Ομοιότητα Μεταξύ Συσταδοποιήσεων

Για 2 συσταδοποιήσεις του ίδιου συνόλου δεδομένων υπολογίζουμε με μια συνάρτηση της R (*Παράρτημα Β, Β.6*) την ομοιότητα μεταξύ τους υπολογίζοντας πρώτα έναν 2x2 πίνακα με τα ακόλουθα 4 κελιά [38]:
n_11: ο αριθμός των ζευγών των στοιχείων που βρίσκονται στην ίδια συστάδα και στις δύο συσταδοποιήσεις
n_10: ο αριθμός των ζευγών των στοιχείων που βρίσκονται στην ίδια συστάδα στην πρώτη συσταδοποίηση αλλά σε διαφορετική συστάδα στην δεύτερη συσταδοποίηση
n_01: ο αριθμός των ζευγών των στοιχείων που βρίσκονται σε διαφορετική συστάδα στην πρώτη συσταδοποίηση αλλά στην ίδια συστάδα στην δεύτερη συσταδοποίηση
n_00: ο αριθμός τν ζευγών των στοιχείων όπου δεν βρίσκονται στην ίδια συστάδα σε καμία από τις δύο συσταδοποιήσεις.

Η ομοιότητα (με την χρήση αυτού του πίνακα) μπορεί να υπολογιστεί με τον δείκτη *Rand* και *Jaccard*.

$$\text{Δείκτης Rand} = \frac{n\_11+n\_00}{n\_11+n\_10+n\_01+n\_00}$$

$$\text{Δείκτης Jaccard} = \frac{n\_11}{n\_11+n\_10+n\_01}$$

Ομοιότητα αποτελέσματος αλγορίθμων σύμφωνα με τον δείκτη ***Jaccard***

Πίνακας 4.7.3.1 *Ποσοστό ομοιότητας συσταδοποιήσεων μεταξύ αλγορίθμων με την μέθοδο Jaccard*

	AGNES(complete)	AGNES(ward)	DIANA	k-means	PAM
AGNES(comp)	1	0.25	0.73	0.22	0.28
AGNES(ward)	0.25	1	0.23	0.38	0.29
DIANA2	0.73	0.23	1	0.28	0.3
k-means	0.22	0.38	0.28	1	0.39
PAM	0.28	0.29	0.3	0.39	1

Ομοιότητα αποτελέσματος αλγορίθμων σύμφωνα με τον δείκτη ***Rand***

Πίνακας 4.7.3.2 *Ποσοστό ομοιότητας συσταδοποιήσεων μεταξύ αλγορίθμων με την μέθδο Rand*

	AGNES(complete)	AGNES(ward)	DIANA	k-means	PAM
AGNES(comp)	1	0.61	0.86	0.58	0.65
AGNES(ward)	0.61	1	0.57	0.79	0.76
DIANA2	0.86	0.57	1	0.6	0.65
k-means	0.58	0.79	0.6	1	0.8
PAM	0.65	0.76	0.65	0.8	1

Κεφάλαιο 5

Συμπεράσματα και Μελλοντική Εργασία

5.1 Συμπεράσματα

Σύμφωνα με τα αποτελέσματα (Σχήματα 4.7.1.1-4.7.1.6) που είχαμε αναφορικά με τις μετρήσεις σταθερότητας συμπεραίνουμε ότι:
Ως προς την ***συνδεσιμότητα***, όταν ο αριθμός των συστάδων είναι 5 οι αλγόριθμοι αποδίδουν καλύτερα με την ακόλουθη (φθίνουσα) σειρά:

1. DIANA
2. AGNES (complete)
3. AGNES (ward)
4. k-means
5. PAM

Ως προς τον ***δείκτη Dunn***, όταν ο αριθμός των συστάδων είναι 5 οι αλγόριθμοι αποδίδουν καλύτερα με την ακόλουθη (φθίνουσα) σειρά:

1. DIANA
2. AGNES (complete)
3. AGNES (ward)
4. K-means
5. PAM

Ως προς το ***πλάτος σιλουέτας***, όταν ο αριθμός των συστάδων είναι 5 οι αλγόριθμοι αποδίδουν καλύτερα με την ακόλουθη (φθίνουσα) σειρά:

1. DIANA
2. AGNES (complete)
3. AGNES (ward)
4. PAM
5. K-means

Πιο κάτω βλέπουμε ποιος αλγόριθμος και με ποιο αριθμό συστάδων είχε την καλύτερη απόδοση ως προς τις τρεις μετρήσεις

ΜΕΤΡΗΣΗ	**ΣΚΟΡ**	**ΑΛΓΟΡΙΘΜΟΣ**	**ΣΥΣΤΑΔΕΣ**
Συνδεσιμότητα	20.36	DIANA	5
Δείκτης Dunn	0.73	DIANA	7
Πλάτος Σιλουέτας	0.11	DIANA	5

Βλέπουμε ότι και στις τρεις μετρήσεις καλύτερη απόδοση έχει ο αλγόριθμος DIANA και γιαδύο εξ αυτών ο ιδανικό αριθμός συστάδων είναι πέντε. .

Σύμφωνα με τα αποτελέσματα (Πίνακας 4.7.3.1) οι συσταδοποιήσεις που έγιναν από τους 5 αλγορίθμους είναι όμοιες με την ακόλουθη φθίνουσα σειρά, όταν η ομοιότητα υπολογίζεται με την μέθοδο Jaccard:

1. DIANA - AGNES(complete) 0.73
2. k-means - AGNES(ward) 0.38
3. DIANA – PAM 0.3
4. PAM - AGNES(ward) 0.29
5. PAM - AGNES(complete) 0.28 + k-means - DIANA 0.28
6. AGNES(ward) - AGNES(complete) 0.25
7. DIANA - AGNES(ward) 0.23
8. k-means - AGNES(complete) 0.22

Σύμφωνα με τα αποτελέσματα (Πίνακας 4.7.3.2) οι συσταδοποιήσεις που έγιναν από τους 5 αλγορίθμους είναι όμοιες με την ακόλουθη φθίνουσα σειρά, όταν η ομοιότητα υπολογίζεται με την μέθοδο Rand:

1. DIANA - AGNES(complete) 0.86
2. k-means - AGNES(ward) 0.79
3. PAM - AGNES(ward) 0.76
4. PAM - AGNES(complete) 0.65 + DIANA – PAM 0.65
5. AGNES(ward) - AGNES(complete) 0.61
6. k-means - DIANA 0.6
7. k-means - AGNES(complete) 0.58
8. DIANA - AGNES(ward) 0.57

5.2 Μελλοντική Εργασία

Η παρούσα μελλοντική εργασία θα μπορούσε να επεκταθεί στο μέλλον με τους εξής τρόπους:

Να γίνει κατ' αρχάς **εξωτερική αξιολόγηση.** Στην εξωτερική αξιολόγηση τα αποτελέσματα της συσταδοποίησης αξιολογούνται με βάση δεδομένα τα οποία δεν χρησιμοποιήθηκαν στην συσταδοποίηση και είναι προταξινομημένα συνήθως από ειδικούς επιστήμονες.
Επιπλέον, καθώς ένας από τους σκοπούς της εργασίας αυτής ήταν η συμβολή στα πρόγραμμα Granatum, θα μπορούσαν τα R scripts τα οποία χρησιμοποιήθηκαν για την εξαγωγή των αποτελεσμάτων να ενσωματωθούν στην πλατφόρμα **LiSIs** υπό την μορφή εργαλείου έτσι ώστε να δώσει την δυνατότητα στους ειδικούς επιστήμονες να εκτελούν αυτές τις εργασίες εύκολα, γρήγορα και αξιόπιστα.
Επίσης μια ακόμα μελλοντική επέκταση, με τεράστια χρησιμότητα, θα ήταν η **δισδιάστατη και τρισδιάστατη γραφική αναπαράσταση** των χημικών ενώσεων των δραστικών ουσιών των φαρμάκων, η οποία θα έδινε την δυνατότητα στους ειδικούς επιστήμονες που θα έκαναν χρήση αυτού του εργαλείου να δουν και γραφικά τα περιεχόμενα της κάθε συστάδας.
Τέλος στο εργαλείο αυτό θα μπορούσε να **ενσωματωθεί οποιαδήποτε πληροφορία** για το κάθε στοιχείο, εκτός από την δισδιάστατη και τρισδιάστατη γραφική αναπαράσταση των χημικών ενώσεων, όπως είναι ο χημικός τους τύπος, το μοριακό τους βάρος, το drug-likeness τους (ALogP,XLogP), SMILES κ.λ.π.

Βιβλιογραφία

[1] About ZEINCRO, *ZEINCRO*, http://www.zeincro.com/about-zeincro.

[2] Agglomerative Nesting (Hierarchical Clustering), *R manual*, http://stat.ethz.ch/R-manual/R-devel/library/cluster/html/agnes.html.

[3] Agglomerative Nesting Object, *R manual*, http://stat.ethz.ch/R-manual/R-devel/library/cluster/html/agnes.object.html.

[4] Agglomerative Nesting, *Unesco*, http://www.unesco.org/webworld/idams/advguide/Chapt7_1_4.htm.

[5] Applied Data Mining, *PennState University*, https://onlinecourses.science.psu.edu/stat857/node/108.

[6] Cambridge Crystallographic Data Centre, http://www.ccdc.cam.ac.uk/SupportandResources/Support/Pages/SupportSolution.aspx?supportsolutionid=228.

[7] Centroid Linkage, *Stanford University*, http://www.stanford.edu/~maureenh/quals/html/ml/node78.html.

[8] Clustering Large Applications (CLARA) Object, *R manual*, http://stat.ethz.ch/R-manual/R-devel/library/cluster/html/clara.object.html.

[9] Clustering Large Applications, *R manual*, http://stat.ethz.ch/R-manual/R-devel/library/cluster/html/clara.html.

[10] Clustering Methods, *SAS OnlineDoc*, http://v8doc.sas.com/sashtml/stat/chap23/sect12.htm.

[11] Clustering: A Survey http://www.slideshare.net/rcapaldo/cluster-analysis-presentation.

[12] Complete Linkage clustering, *Wikipedia*, http://en.wikipedia.org/wiki/Complete-linkage_clustering.

[13] Data Mining Algorithms In R/Clustering/CLARA, *Wikibooks*, http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/CLARA.

[14] Data Mining Algorithms In R/Clustering/Partitioning Around Medoids (PAM), *Wikibooks*, http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Partitioning_Around_Medoids_(PAM)

[15] DIvisive ANAlysis Clustering, *R manual*, http://stat.ethz.ch/R-manual/R-devel/library/cluster/html/diana.html

[16] Divisive Analysis, *Unesco*, http://www.unesco.org/webworld/idams/advguide/Chapt7_1_5.htm

[17] Dunn index, *Wikipedia*, http://en.wikipedia.org/wiki/Dunn_index

[18] Guy Brock, Vasyl Pihur, Susmita Datta, and Somnath Datta, "clValid, an R package for cluster validation". Department of Bioinformatics and Biostatistics, University of Louisville, 17 October 2011
[19] Hamming Distance, *Wikipedia*, http://en.wikipedia.org/wiki/Hamming_distance
[20] Hierarchical Clustering Algorithms, The Computational Geometry Lab at McGill, http://cgm.cs.mcgill.ca/~soss/cs644/projects/siourbas/sect5.html
[21] Hierarchical Clustering Metrics, *Wikipedia*, http://en.wikipedia.org/wiki/Hierarchical_clustering
[22] Hierarchical Clustering, *R manual*, http://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html
[23] Hierarchical Clustering, *Wikipedia*, http://en.wikipedia.org/wiki/Hierarchical_clustering#Example_for_Agglomerative_Clustering
[24] Jaccard Index, *Wikipedia*, http://en.wikipedia.org/wiki/Jaccard_index#Tanimoto_Similarity_and_Distance
[25] Johnson, S. C. Hierarchial Clustering Schemes, Psychometrika, 1967.
[26] K-Means Clustering, *R manual*, http://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html
[27] k-means Clustering, *Wikipedia*, http://en.wikipedia.org/wiki/K-means_clustering
[28] Legendre, Fionn Murtagh and Pierre, "Ward's Hierarchical Clustering Method", Science Foundation Ireland, Wilton Park House and Departement de sciences biologiques, Universite de Montreal, C.P. 6128 succursale Centre-ville, Montreal, Quebec, Canada H3C 3J7
[29] M. Shahriar Hossain, Michael Narayan, and Naren Ramakrishnan, "Efficiently Discovering Hammock Paths from Induced Similarity Networks", Department of Computer Science, Virginia Tech, Blacksburg, VA 24061.
[30] Median linkage, *Stanford University*, http://www.stanford.edu/~maureenh/quals/html/ml/node79.html
[31] Mining Similarity http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/mvoget/similarity/similarity.html
[32] Partitional Clustering, *Springer Reference*, http://www.springerreference.com/docs/html/chapterdbid/179343.html
[33] Partitioning Around Medoids (PAM) Object, *R manual*, http://stat.ethz.ch/R-manual/R-devel/library/cluster/html/pam.object.html
[34] Partitioning Around Medoids, *R manual*, http://stat.ethz.ch/R-manual/R-devel/library/cluster/html/pam.html
[35] Project Partners, *Linked2Safety*, http://www.linked2safety-project.eu/node/28
[36] Project Vision, *Granatum*, http://www.granatum.org/pub/vision.html

[37] R (programming language), *Wikipedia*, http://en.wikipedia.org/wiki/R_(programming_language)
[38] R (Υπολογιστική στατιστική) http://stavrakoudis.econ.uoi.gr/stavrakoudis/?iid=169
[39] Ramey, John A. "Evaluation of Clustering Algorithms", 31 August 2012
[40] Silhouette (clustering), *Wikipedia*, http://en.wikipedia.org/wiki/Silhouette_(clustering)
[41] Single Linkage Clustering, *Wikipedia*, http://en.wikipedia.org/wiki/Single-linkage_clustering
[42] Sorensen-Dice Coefficient, *Wikipedia*, http://en.wikipedia.org/wiki/Dice's_coefficient
[43] The project, *Linked2Safety*, http://www.linked2safety-project.eu/node/23
[44] UPGMA, *Wikipedia*, http://en.wikipedia.org/wiki/UPGMA
[45] WPGMA http://stn.spotfire.com/spotfire_client_help/hc/hc_wpgma.htm
[46] Εισαγωγή στην R http://www2.ucy.ac.cy/~fokianos/GreekRbook/indexRbook.htm

Παράρτημα Α

Αλγόριθμοι Συσταδοποίησης στην R

Α.1 Ο Agnes στην R

Ανήκει στην βιβλιοθήκη ***cluster*** και καλείται ως ακολούθως [2]:

agnes(x, diss = inherits(x, "dist"), metric = "euclidean", stand = FALSE, method = "average", par.method, keep.diss = n < 100, keep.data = !diss)

x: πίνακας δεδομένων ή αποστάσεων αναλόγως της τιμής που θα δοθεί στην παράμετρο diss

diss: αν είναι TRUE τότε το x χρησιμοποιείται σαν πίνακας αποστάσεων. Αν είναι FALSE το x χρησιμοποιείται σαν πίνακας στοιχείων.

metric: καθορίζει ποια μέθοδος θα χρησιμοποιηθεί για το υπολογισμό των αποστάσεων μεταξύ των στοιχείων. Δίνεται επιλογή μεταξύ Ευκλείδειας (Euclidean) και Manhattan. Αν το x είναι πίνακας αποστάσεων τότε αυτή η παράμετρος αγνοείται.

stand: αν είναι true τότε οι μετρήσεις στο **x** τυποποιούνται πριν υπολογιστούν οι αποστάσεις. Οι μετρήσεις τυποποιούνται σε κάθε στήλη αφαιρώντας τη μέση τιμή κάθε στήλης και διαιρώντας με με την μέση απόλυτη απόκλιση της μεταβλητής. Αν το **x** είναι πίνακας αποστάσεων τότε αυτή η παράμετρος αγνοείται.

method: η μέθοδος με την οποία γίνεται ο υπολογισμός της απόστασης μεταξύ των συστάδων (linkage). Οι επιλογές που υπάρχουν είναι οι εξής: “average” (UPGMA), “single” (single linkage), “complete” (complete linkage), “ward” (ward’s method), “weighted” (weighted average linkage), “flexible” (flexible linkage).

par.method: πίνακας αριθμών μεγέθους 1, 2, 3 ή 4 υφίσταται μόνο στην περίπτωση που το method είναι flexible.

keep.diss: υποδεικνύει αν οι αποστάσεις επιστρέφονται (TRUE) ή όχι (FALSE) στο αποτέλεσμα.

keep.data: υποδεικνύει αν τα δεδομένα επιστρέφονται (TRUE) ή όχι (FALSE) στο αποτέλεσμα.

Μετά την ορθή εκτέλεση της πιο πάνω γραμμής κώδικα στην R δημιουργείται ένα αντικείμενο ***agnes*** *με τα ακόλουθα μέρη [3]:*

order: διάνυσμα που περιέχει τα στοιχεία μετατιθέμενα έτσι ώστε να είναι δυνατή η σχεδίαση τους.

order.lab: διάνυσμα παρόμοιο με το orders με τη διαφορά ότι αντί για τα πραγματικά στοιχεία περιέχει ετικέτες. Αυτό το μέρος είναι διαθέσιμο εφόσον τα αρχικά στοιχεία είχαν ετικέτες.

height: διάνυσμα με τις αποστάσεις μεταξύ των συγχωνευμένων συστάδων στα διαδοχικά επίπεδα.

ac: agglomerative coefficient η οποία μετράει την δομή του συνόλου δεδομένων και η οποία ορίζεται ως εξής: $\frac{1}{n}\sum_{i=1}^{n} 1 - \frac{d_{f_i}}{d_{l_i}}$ όπου d_{f_i} η απόσταση του στοιχείου i από την πρώτη συστάδα με την οποία συγχωνεύτηκε και d_{l_i} η απόσταση του στοιχείου i από την τελευταία συστάδα με την οποία συγχωνεύτηκε. Επειδή το agglomerative coefficient μεγαλώνει όσο μεγαλώνει και ο αριθμός των στοιχείων δεν μπορεί να χρησιμοποιηθεί για να συγκρίνει συσταδοποιήσεις με μεγάλη διαφορά στο μέγεθος.

merge: πίνακας $n-1 \times 2$ όπου n ο αριθμός των στοιχείων. Η γραμμή i του πίνακα περιγράφη την συγχώνευση των συστάδων στο βήμα i της συσταδοποίησης. Αν το merge(i) είναι αρνητικό τότε το στοιχείο i συγχωνεύεται σ' αυτό το στάδιο. Αν το merge(i) είναι θετικό τότε το στοιχείο ανήκει στην συστάδα που συγχωνεύτηκε στο στάδιο merge(i).

diss: ο πίνακας αποστάσεων

data: πίνακας που περιέχει τις αρχικές ή τις τυποποιημένες μετρήσεις. Εξαρτάται από την παράμετρο stand της συνάρτησης agnes. Αν έχει δοθεί πίνακας αποστάσεων σαν είσοδος τότε αυτό το μέρος δεν είναι διαθέσιμο.

A.2 Ο Diana στην R

Ανήκει στην βιβλιοθήκη **cluster** και καλείται ως ακολούθως [15]:

diana(x, diss = inherits(x, "dist"), metric = "euclidean", stand = FALSE,
keep.diss = n < 100, keep.data = !diss, trace.lev = 0)

x: πίνακας δεδομένων (όλες οι τιμές πρέπει να είναι αριθμητικές) ή πίνακας αποστάσεων ανάλογα με την τιμή της παραμέτρου diss.

diss: λογική σημαία. αν είναι TRUE τότε το x χρησιμοποιείται σαν πίνακας αποστάσεων. Αν είναι FALSE το x χρησιμοποιείται σαν πίνακας στοιχείων.

stand: αν είναι true τότε οι μετρήσεις στο **x** τυποποιούνται πριν υπολογιστούν οι αποστάσεις. Οι μετρήσεις τυποποιούνται σε κάθε στήλη αφαιρώντας τη μέση τιμή κάθε στήλης και διαιρώντας με την μέση απόλυτη απόκλιση της μεταβλητής. Αν το **x** είναι πίνακας αποστάσεων τότε αυτή η παράμετρος αγνοείται.

keep.diss: υποδεικνύει αν οι αποστάσεις επιστρέφονται (TRUE) ή όχι (FALSE) στο αποτέλεσμα.

keep.data: υποδεικνύει αν τα δεδομένα επιστρέφονται (TRUE) ή όχι (FALSE) στο αποτέλεσμα.

trace.lev: ακέραιος αριθμός που προσδιορίζει το trace level για την εκτύπωση diagnostics κατά τη διάρκεια του αλγορίθμου. Το προκαθορισμένο 0 υποδεικνύει μηδενική εκτύπωση. Μεγαλύτερες τιμές υποδεικνύουν και μεγαλύτερο ποσοστό εκτύπωσης diagnostics.

Μετά την ορθή εκτέλεση της πιο πάνω γραμμής κώδικα στην R δημιουργείται ένα αντικείμενο ***diana*** *με τα ακόλουθα μέρη [15]:*

order: διάνυσμα που περιέχει τα στοιχεία μετατιθέμενα έτσι ώστε να είναι δυνατή η σχεδίαση τους.

order.lab: διάνυσμα παρόμοιο με το orders με τη διαφορά ότι αντί για τα πραγματικά στοιχεία περιέχει ετικέτες. Αυτό το μέρος είναι διαθέσιμο εφόσον τα αρχικά στοιχεία είχαν ετικέτες.

height: διάνυσμα με τις αποστάσεις μεταξύ των συγχωνευμένων συστάδων στα διαδοχικά επίπεδα πριν την συγχώνευση.

dc: divisive coefficient η οποία μετράει την δομή του συνόλου δεδομένων και η οποία ορίζεται ως εξής: $\frac{1}{n}\sum_{i=1}^{n} 1 - d_i - d$ όπου d_i η διάμετρος της τελευταίας συστάδας στην οποία ανήκει το στοιχείο i και d η διάμετρος του συνόλου δεδομένων. Επειδή το divisive coefficient μεγαλώνει όσο μεγαλώνει και ο αριθμός των στοιχείων δεν μπορεί να χρησιμοποιηθεί για να συγκρίνει συσταδοποιήσεις με μεγάλη διαφορά στο μέγεθος.

A.3 Ο hclust στην R

Για σκοπούς απλοποίησης και καλύτερης γραφικής αναπαράστασης τα αντικείμενα ***agnes*** *και* ***diana*** *μετατρέπονται σε αντικείμενα* ***hclust*** [22]:

merge: πίνακας $n-1 \times 2$ όπου n ο αριθμός των στοιχείων. Η γραμμή i του πίνακα περιγράφει την συγχώνευση των συστάδων στο βήμα i της συσταδοποίησης. Αν το merge(i) είναι αρνητικό τότε το στοιχείο i συγχωνεύεται σ' αυτό το στάδιο. Αν το merge(i) είναι θετικό τότε το στοιχείο ανήκει στην συστάδα που συγχωνεύτηκε στο στάδιο merge(i). Άρα οι αρνητικοί αριθμοί στον πίνακα δείχνουν συσταδοποίηση μονών στοιχείων ενώ οι θετικοί
αριθμοί δείχνουν συσταδοποιήσεις ομάδων στοιχείων.

height: διάνυσμα με τις αποστάσεις μεταξύ των συγχωνευμένων συστάδων στα διαδοχικά επίπεδα.

order: διάνυσμα που περιέχει τα στοιχεία μετατιθέμενα έτσι ώστε να είναι δυνατή η σχεδίαση τους, δηλαδή να μην υπάρχουν διασταυρώσεις κλάδων.

labels: ετικέτες για το κάθε στοιχείο.

call: η κλήση η οποία παρήξε το αποτέλεσμα.

method: η μέθοδος συσταδοποίησης που χρησιμοποιήθηκε.

dist.method: η απόσταση που χρησιμοποιήθηκε για να δημιουργηθεί ο πίνακας αποστάσεων. Επιστρέφεται εφόσον είναι διαθέσιμο.

A.4 Ο k-means στην R

Ανήκει στη βιβλιοθήκη ***stats*** και καλείται ως ακολούθως [26]:

kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"))

x: Ένας πίνακας αριθμών

centers: ο αριθμός των συστάδων ή ένα σετ από αρχικά ξεχωριστά κέντρα. Αν δοθεί αριθμός τότε το τυχαίο σετ από ξεχωριστές γραμμές στο x επιλέγεται για τα κέντρα των συστάδων

iter.max: Ο μέγιστος επιτρεπόμενος αριθμός επαναλήψεων

nstart: Αν το centers είναι αριθμός τότε το nstart δίνει τον αριθμό των τυχαίων σετ που θα επιλεγεί

algorithm: ο αλγόριθμος που θα χρησιμοποιηθεί. Δίνεται επιλογή μέσα από 4 υλοποιήσεις "Hartigan-Wong", "Lloyd", "Forgy" και "MacQueen". Default τιμή είναι το "Hartigan-Wong".

Μετά την ορθή εκτέλεση της πιο πάνω γραμμής κώδικα στην R δημιουργείται ένα αντικείμενο ***kmeans*** *με τα ακόλουθα μέρη:*

cluster: πίνακας ακεραίων που δηλώνει σε ποια συστάδα ανήκει το κάθε στοιχείο

centers: πίνακας με τα κέντρα των συστάδων.

withnss: το άθροισμα των τετραγωνικών αποστάσεων για την κάθε συστάδα.

size: ο αριθμός των στοιχείων στην κάθε συστάδα.

A.5 Ο pam στην R

Ανήκει στην βιβλιοθήκη ***cluster*** και καλείται ως ακολούθως [34]:

pam(x, k, diss, metric, medoids, stand, cluster.only, do.swap, keep.diss, keep.data, trace.lev)

x: ανάλογα με την τιμή της μεταβλητής ***diss*** μπορεί να είναι πίνακας είτε στοιχείων είτε αποστάσεων. Αν είναι πίνακας στοιχείων η κάθε γραμμή είναι ένα στοιχείο και η κάθε στήλη είναι μια μεταβλητή.

k: ο αριθμός των συστάδων που θα δημιουργηθούν, 0<k<n όπου n ο αρθμός των στοιχείων

diss: αν είναι TRUE το ***x*** χρησιμοποιείται σαν πίνακας αποστάσεων,αν είναι FALSE τότε χρησιμοποιείται σαν πίνακας στοιχείων.

metric: καθορίζει ποια μέθοδος θα χρησιμοποιηθεί για το υπολογισμό των αποστάσεων μεταξύ των στοιχείων. Δίνεται επιλογή μεταξύ Ευκλείδειας (Euclidean) και Manhattan. Αν το x είναι πίνακας αποστάσεων τότε αυτή η παράμετρος αγνοείται.

stand: αν είναι true τότε οι μετρήσεις στο **x** τυποποιούνται πριν υπολογιστούν οι αποστάσεις. Οι μετρήσεις τυποποιούνται σε κάθε στήλη αφαιρώντας τη μέση τιμή κάθε στήλης και διαιρώντας με την μέση απόλυτη απόκλιση της μεταβλητής. Αν το **x** είναι πίνακας αποστάσεων τότε αυτή η παράμετρος αγνοείται.

cluster.only: αν είναι true μόνο η συσταδοποίηση θα υπολογιστεί και θα επιστραφεί.

do.swap: αν είναι true τότε θα εκτελεστεί η φάση της ανταλλαγής (swap phase) αλλιώς δεν θα εκτελεστεί.

keep.diss: αν είναι true τότε οι αποστάσεις θα επιστραφούν σαν αποτέλεσμα αλλιώς όχι.

keep.data: αν είναι true τότε θα επιστραφεί σαν αποτέλεσμα ο πίνακας x αλλιώς όχι.

trace.lev: αριθμητική παράμετρος που καθορίζει το επίπεδο εκτύπωσης diagnostics. Default είναι το 0 όπου δεν τυπώνεται τίποτα.

Μετά την ορθή εκτέλεση της πιο πάνω γραμμής κώδικα στην R δημιουργείται ένα αντικείμενο ***pam*** *με τα ακόλουθα μέρη* [33]:

medoids: τα κέντρα (medoids) των συστάδων. Αν έχει δοθεί σαν είσοδος πίνακας αποστάσεων τότε το medoids είναι ένα διάνυσμα με τα στοιχεία (αριθμοί ή ετικέτες) αλλιώς είναι ένα διάνυσμα με συντεταγμένες των στοιχείων.

id.med: διάνυσμα ακεραίων δεικτών που αντιπροσωπεύουν τους αριθμούς των medoids.

clustering: διάνυσμα συσταδοποίησης μεγέθους n (αριθμός στοιχείων) που περιέχει για κάθε στοιχείο τον αριθμό (id) της συστάδας στην οποία ανήκει.

objective: η αντικειμενική συνάρτηση μετά το πρώτο και δεύτερο βήμα του αλγορίθμου.

isolation: διάνυσμα με μέγεθος τον αριθμό των συστάδων που δείχνει ποιες συστάδες είναι απομονωμένες (L ή L*)και ποιες όχι. Μια συστάδα είναι L* αν η διάμετρος της είναι μικρότερη από τον διαχωρισμό της (πόσο καλά διαχωρισμένη είναι μια συστάδα από τις υπόλοιπες). Μια συστάδα είναι L αν για κάθε στοιχείο i η μέγιστη απόσταση μεταξύ του i και οποιουδήποτε άλλου στοιχείου της συστάδας είναι μικρότερη από την μικρότερη απόσταση μεταξύ του i και οποιουδήποτε άλλου στοιχείου που δεν ανήκει στην συστάδα. Προφανώς κάθε συστάδα L* είναι και L.

clussinfo: πίνακας του οποίου η κάθε γραμμή περιέχει πληροφορίες για μια συστάδα: αριθμό στοιχείων, μέγιστη και μέση απόσταση μεταξύ των στοιχείων, medoid, διάμετρο, διαχωρισμό.

silinfo: μια λίστα με αντικείμενα silhoutte που κρατούν διάφορες πληροφορίες για το σύνολο δεδομένων.

diss: ο πίνακας αποστάσεων
call: παράγει κλήση της συνάρτησης

data: πίνακας που περιέχει τα αρχικά δεδομένα

A.6 Ο Clara στην R

Ανήκει στην βιβλιοθήκη ***cluster*** και καλείται ως ακολούθως [9]:

clara(x, k, metric = "euclidean", stand = FALSE, samples = 5, sampsize = min(n, 40 + 2 * k), trace = 0, medoids.x = TRUE, keep.data = medoids.x, rngR = FALSE)

x: πίνακας στοιχείων, κάθε γραμμή είναι ένα στοιχείο και κάθε στήλη μια μεταβλητή

k: ο αριθμός των συστάδων που θα δημιουργηθούν, 0<k<n όπου n ο αριθμός των στοιχείων

metric: καθορίζει ποια μέθοδος θα χρησιμοποιηθεί για το υπολογισμό των αποστάσεων. Δίνεται επιλογή μεταξύ Ευκλείδειας (Euclidean) και Manhattan.

stand: αν είναι true τότε οι μετρήσεις στο **x** τυποποιούνται πριν υπολογιστούν οι αποστάσεις. Οι μετρήσεις τυποποιούνται σε κάθε στήλη αφαιρώντας τη μέση τιμή κάθε στήλης και διαιρώντας με την μέση απόλυτη απόκλιση της μεταβλητής. Αν το **x** είναι πίνακας αποστάσεων τότε αυτή η παράμετρος αγνοείται.

samples: ο αριθμός των δειγμάτων που θα παρθούν από το σετ δεδομένων.

sampsize: ο αριθμός των στοιχείων σε κάθε δείγμα ο οποίος πρέπει να υπερβαίνει τον αριθμό των συστάδων (k) αλλά να μην ξεπερνά τον αριθμό των στοιχείων.

trace: αριθμητική παράμετρος που καθορίζει το επίπεδο εκτύπωσης διαγνωστικών.

medoids.x: υποδεικνύει κατά πόσο τα κέντρα θα επιστραφούν. Αν πάρει την τιμή false τότε και το keep.data θα πρέπει να είναι false.

keep.data: αν είναι true τότε θα επισταφεί σαν αποτέλεσμα ο πίνακας x αλλιώς όχι.

rngR: αν είναι true τότε θα χρησιμοποιηθεί ο random number generator της R αλλιώς θα χρησιμοποιηθεί αυτός του clara. Αν είναι true σημαίνει επίσης ότι κάθε κλήση του αλγορίθμου clara θα επιστρέφει διαφορετικό αποτέλεσμα αν και οι διαφορές είναι συνήθως μικρές.

Μετά την ορθή εκτέλεση της πιο πάνω γραμμής κώδικα στην R δημιουργείται ένα αντικείμενο ***clara*** *με τα ακόλουθα μέρη* [8]*:*

sample: ετικέτες ή αριθμοί των στοιχείων στο καλύτερο δείγμα, το οποίο χρησιμοποιείται από τον αλγόριθμο για τον τελευταίο διαχωρισμό.

medoids: τα κέντρα (medoids) των συστάδων. Αν έχει δοθεί σαν είσοδος πίνακας αποστάσεων τότε το medoids είναι ένα διάνυσμα με τα στοιχεία (αριθμοί ή ετικέτες) αλλιώς είναι ένα διάνυσμα με συντεταγμένες των στοιχείων.

i.med: διάνυσμα ακεραίων δεικτών που αντιπροσωπεύουν τους αριθμούς των medoids.

clustering: διάνυσμα συσταδοποίησης μεγέθους n (αριθμός στοιχείων) που περιέχει για κάθε στοιχείο τον αριθμό (id) της συστάδας στην οποία ανήκει.

objective: η αντικειμενική συνάρτηση μετά το πρώτο και δεύτερο βήμα του αλγορίθμου.

clussinfo: πίνακας του οποίου η κάθε γραμμή περιέχει πληροφορίες για μια συστάδα: αριθμό στοιχείων, μέγιστη και μέση απόσταση μεταξύ των στοιχείων, medoid, διάμετρο, διαχωρισμό.

diss: ο πίνακας αποστάσεων

silinfo: μια λίστα με αντικείμενα silhouette που κρατούν διάφορες πληροφορίες για το καλύτερο δείγμα.

call: παράγει κλήση της συνάρτησης

data: πίνακας που περιέχει τα αρχικά δεδομένα

Παράρτημα Β

Κώδικας στην R

Β.1 Αλγόριθμος AGNES (complete linkage)

```
#set working directory
setwd("C:/Drugs_l2s/")

#load necessary packages
library(cluster)
library(rjson)

#read distance matrix
sim_matrix    <-    read.table("C:/Drugs_l2s/distance_matrix.dat",    sep=,
header=TRUE, row.names=1)
dmat<- as.dist(sim_matrix, diag=TRUE)

#run agnes
agn <- agnes(dmat, diss=TRUE, method="complete")

#save agglomerative coefficient to txt file
write.table("agnes_complete_h",file="C:/Drugs_l2s/results/agglomerative_c
oefficient.txt",append=TRUE,sep=" ")
write.table(agn$ac,file="C:/Drugs_l2s/results/agglomerative_coefficient.t
xt",append=TRUE,sep=" ")

#convert agnes object to hclust object
hagn <- as.hclust(agn)



# ============================ CUTOFF=0.85 ============================ #
#create pdf file
pdf(file='C:/Drugs_l2s/results/agnes_results_complete_h.pdf',  width=17.0,
height=7)

#create dendrogram
plot(hagn,         main='Agglomerative         Nesting         Hierarchical
Clustering',sub='Cluster  Method:  Complete  Linkage  ,  Cluster  Cutoff:
0.85',xlab='Molecule   ID',   ylab='Soergel   Distance   (1   -   Tanimoto
Coefficient)')
x <- rect.hclust(hagn, h=0.85, border='red')
dev.off()
cluster_groups <- cutree(hagn, k=length(x))

#save clustering results to csv file
write.table(cluster_groups,file="C:/Drugs_l2s/results/agnes_results_compl
ete_h.csv",sep=",")
```

```
#save dendrogram to pdf file
sink('C:/Drugs_l2s/json/agnes_results_complete_h.json')
cat(toJSON(cluster_groups))
sink()



# ============================ CUTOFF=0.958 ============================ #
#create pdf file
pdf(file='C:/Drugs_l2s/results/agnes_results_complete_h2.pdf',
width=17.0, height=7)

#create dendrogram
plot(hagn,         main='Agglomerative         Nesting         Hierarchical
Clustering',sub='Cluster  Method:  Complete  Linkage  ,  Cluster  Cutoff:
0.958',xlab='Molecule   ID',   ylab='Soergel   Distance   (1   -   Tanimoto
Coefficient)')
x <- rect.hclust(hagn, h=0.958, border='red')
dev.off()
cluster_groups <- cutree(hagn, k=length(x))

#save clustering results to csv file
write.table(cluster_groups,file="C:/Drugs_l2s/results/agnes_results_compl
ete_h2.csv",sep=",")

#save dendrogram to pdf file
sink('C:/Drugs_l2s/json/agnes_results_complete_h2.json')
cat(toJSON(cluster_groups))
sink()



# ============================ No of Clusters=5 ======================== #
#create pdf file
pdf(file='C:/Drugs_l2s/results/agnes_results_complete_k.pdf',  width=17.0,
height=7)

#create dendrogram
plot(hagn, main='Agglomerative Nesting Hierarchical Cluster',sub='Cluster
Method: Complete Linkage, Cluster k:5',xlab='Molecule ID', ylab='Soergel
Distance (1 - Tanimoto Coefficient)')
x <- rect.hclust(hagn, k=5, border='red')
dev.off()
cluster_groups <- cutree(hagn, k=length(x))

#save clustering results to csv file
write.table(cluster_groups,file="C:/Drugs_l2s/results/agnes_results_compl
ete_k.csv",sep=",")

#save dendrogram to pdf file
sink('C:/Drugs_l2s/json/agnes_results_complete_k.json')
cat(toJSON(cluster_groups))
sink()
```

B.2 Αλγόριθμος AGNES (Ward's Method)

```
#set working directory
setwd("C:/Drugs_l2s/")

#load necessary packages
library(cluster)
library(rjson)

#read distance matrix
sim_matrix    <-    read.table("C:/Drugs_l2s/distance_matrix.dat",    sep=,
header=TRUE, row.names=1)
dmat<- as.dist(sim_matrix, diag=TRUE)

#run agnes
agn <- agnes(dmat, diss=TRUE, method="ward")

#save agglomerative coefficient to txt file
write.table("agnes_wards_h",file="C:/Drugs_l2s/results/agglomerative_coef
ficient.txt",append=TRUE,sep=" ")
write.table(agn$ac,file="C:/Drugs_l2s/results/agglomerative_coefficient.t
xt",append=TRUE,sep=" ")

#convert agnes object to hclust object
hagn <- as.hclust(agn)



# ============================ CUTOFF=0.85 ============================ #
#create pdf file
pdf(file='C:/Drugs_l2s/results/agnes_results_ward_h.pdf',      width=17.0,
height=7)

#create dendrogram
plot(hagn,         main='Agglomerative         Nesting         Hierarchical
Clustering',sub='Cluster     Method:     Ward\'s,     Cluster     Cutoff:
0.85',xlab='Molecule   ID',   ylab='Soergel   Distance   (1   -   Tanimoto
Coefficient)')
x <- rect.hclust(hagn, h=0.85, border='red')
dev.off()
cluster_groups <- cutree(hagn, k=length(x))

#save clustering results to csv file
write.table(cluster_groups,file="C:/Drugs_l2s/results/agnes_results_ward_
h.csv",sep=",")

#save dendrogram to pdf file
sink('C:/Drugs_l2s/json/agnes_results_ward_h.json')
cat(toJSON(cluster_groups))
sink()



# ============================ CUTOFF=0.958 ============================ #
```

```
#create pdf file
pdf(file='C:/Drugs_l2s/results/agnes_results_ward_h2.pdf',     width=17.0,
height=7)

#create dendrogram
plot(hagn,         main='Agglomerative         Nesting         Hierarchical
Clustering',sub='Cluster     Method:     Ward\'s,     Cluster     Cutoff:
1.367',xlab='Molecule   ID',   ylab='Soergel   Distance   (1   -   Tanimoto
Coefficient)')
x <- rect.hclust(hagn, h=1.367, border='red')
dev.off()
cluster_groups <- cutree(hagn, k=length(x))

#save clustering results to csv file
write.table(cluster_groups,file="C:/Drugs_l2s/results/agnes_results_ward_
h2.csv",sep=",")

#save dendrogram to pdf file
sink('C:/Drugs_l2s/json/agnes_results_ward_h2.json')
cat(toJSON(cluster_groups))
sink()



# ============================ No of Clusters=5 ======================== #
#create pdf file
pdf(file='C:/Drugs_l2s/results/agnes_results_ward_k.pdf',      width=17.0,
height=7)

#create dendrogram
plot(hagn, main='Agglomerative Nesting Hierarchical Cluster',sub='Cluster
Method: Ward\'s, Cluster k:5',xlab='Molecule ID', ylab='Soergel Distance
(1 - Tanimoto Coefficient)')
x <- rect.hclust(hagn, k=5, border='red')
dev.off()
cluster_groups <- cutree(hagn, k=length(x))

#save clustering results to csv file
write.table(cluster_groups,file="C:/Drugs_l2s/results/agnes_results_ward_
k.csv",sep=",")

#save dendrogram to pdf file
sink('C:/Drugs_l2s/json/agnes_results_ward_k.json')
cat(toJSON(cluster_groups))
sink()
```

B.3 Αλγόριθμος DIANA

```
#set working directory
setwd("C:/Drugs_l2s/")

#load necessary packages
library(cluster)
library(rjson)

#read distance matrix
sim_matrix    <-    read.table("C:/Drugs_l2s/distance_matrix.dat",    sep=,
header=TRUE, row.names=1)
dmat<- as.dist(sim_matrix, diag=TRUE)

#run diana
dian <- diana(dmat, diss=TRUE)

#save divisive coefficient to txt file
write.table("diana_h",file="C:/Drugs_l2s/results/divisive_coefficient.txt
",append=TRUE,sep=" ")
write.table(dian$dc,file="C:/Drugs_l2s/results/divisive_coefficient.txt",
append=TRUE,sep=" ")

#convert diana object to hclust object
hagn <- as.hclust(dian)



# ============================ CUTOFF=0.85 ============================ #
#create pdf file
pdf(file='C:/Drugs_l2s/results/diana_results_h.pdf',           width=17.0,
height=7)

#create dendrogram
plot(hagn,  main='Divisive  Analysis  Hierarchical  Cluster',sub='Cluster
Cutoff:0.85',xlab='Molecule  ID',  ylab='Soergel  Distance  (1  -  Tanimoto
Coefficient)')
x <- rect.hclust(hagn, h=0.85, border='red')
dev.off()
cluster_groups <- cutree(hagn, k=length(x))

#save clustering results to csv file
write.table(cluster_groups,file="C:/Drugs_l2s/results/diana_results_h.csv
",sep=",")

#save dendrogram to pdf file
sink('C:/Drugs_l2s/json/diana_results_h.json')
cat(toJSON(cluster_groups))
sink()



# ============================ CUTOFF=0.958 ============================ #
```

```
#create pdf file
pdf(file='C:/Drugs_l2s/results/diana_results_h2.pdf',          width=17.0,
height=7)

#create dendrogram
plot(hagn,  main='Divisive  Analysis  Hierarchical  Cluster',sub='Cluster
Cutoff:0.958',xlab='Molecule  ID',  ylab='Soergel  Distance  (1  -  Tanimoto
Coefficient)')
x <- rect.hclust(hagn, h=0.958, border='red')
dev.off()
cluster_groups <- cutree(hagn, k=length(x))

#save clustering results to csv file
write.table(cluster_groups,file="C:/Drugs_l2s/results/diana_results_h2.cs
v",sep=",")

#save dendrogram to pdf file
sink('C:/Drugs_l2s/json/diana_results_h2.json')
cat(toJSON(cluster_groups))
sink()



# ============================ No of Clusters=5 ======================== #
#create pdf file
pdf(file='C:/Drugs_l2s/results/diana_results_k.pdf',          width=17.0,
height=7)

#create dendrogram
plot(hagn,  main='Divisive  Analysis  Hierarchical  Cluster',sub='Cluster
k:5',xlab='Molecule   ID',   ylab='Soergel   Distance   (1   -   Tanimoto
Coefficient)')
x <- rect.hclust(hagn, k=5, border='red')
dev.off()
cluster_groups <- cutree(hagn, k=length(x))

#save clustering results to csv file
write.table(cluster_groups,file="C:/Drugs_l2s/results/diana_results_k.csv
",sep=",")


#save dendrogram to pdf file
sink('C:/Drugs_l2s/json/diana_results_k.json')
cat(toJSON(cluster_groups))
sink()
```

B.4 Αλγόριθμος k-means

```
#set working directory
setwd("C:/Drugs_l2s/")

#load necessary packages
library(cluster)
library(stats)
library(rjson)
library(vegan)


#read compounds(data) matrix
x        <-        read.table("C:/Drugs_l2s/L2S_compounds_morgan_fps.dat",
sep=,header=FALSE,row.names=1)


#run kmeans with number of clusters=5
clus <- kmeans(x,5)

#save total within-cluster sum of squares and between-cluster sum of
squares to txt
write.table("tot.withinss",file="C:/Drugs_l2s/results/kmeans_evaluation.t
xt",append=TRUE,sep=" ")
write.table(clus$tot.withinss,file="C:/Drugs_l2s/results/kmeans_evaluatio
n.txt",append=TRUE,sep=" ")
write.table("betweenss",file="C:/Drugs_l2s/results/kmeans_evaluation.txt"
,append=TRUE,sep=" ")
write.table(clus$betweenss,file="C:/Drugs_l2s/results/kmeans_evaluation.t
xt",append=TRUE,sep=" ")


#compute distance matrix
diss<-dist(x)


# Multidimensional scaling of data matrix
cmd<-cmdscale(diss)


# plot MDS, with colors by groups from kmeans
groups<-levels(factor(clus$cluster))
ordiplot(cmd,type="n")
cols<-
c("blueviolet","cornflowerblue","red","chartreuse","darkgoldenrod1")
for(i in seq_along(groups)){
    points(cmd[factor(clus$cluster)==groups[i],],col=cols[i],pch=16)
}


# add spider and hull
ordispider(cmd, factor(clus$cluster), label = TRUE)
ordihull(cmd, factor(clus$cluster), lty = "dotted")
```

```
#save clustering results to csv file
write.table(clus$cluster,file="C:/Drugs_l2s/results/kmeans_results.csv",s
ep=",")
```

B.5 Αλγόριθμος PAM

```
#set working directory
setwd("C:/Drugs_l2s/")

#load necessary packages
library(cluster)
library(rjson)
library(vegan)


#read distance matrix
x  <-  read.table("C:/Drugs_l2s/distance_matrix.dat",  sep=,  header=TRUE,
row.names=1)
dmat<- as.dist(x, diag=TRUE)


#run pam
clus <- pam(dmat,k=5,diss=TRUE)


#save clustering validation results to csv file
write.table(clus$silinfo,file="C:/Drugs_l2s/results/pam_validation.csv",s
ep=",")

#save medoids to csv file
write.table(clus$medoids,file="C:/Drugs_l2s/results/pam_medoids.csv",sep=
",")


# Multidimensional scaling of data matrix
cmd<-cmdscale(x)


# plot MDS, with colors by groups from kmeans
groups<-levels(factor(clus$clustering))
ordiplot(cmd,type="n")
cols<-
c("blueviolet","cornflowerblue","red","chartreuse","darkgoldenrod1")
for(i in seq_along(groups)){
    points(cmd[factor(clus$cluster)==groups[i],],col=cols[i],pch=16)
}


# add spider and hull
ordispider(cmd, factor(clus$clustering), label = TRUE)
ordihull(cmd, factor(clus$clustering), lty = "dotted")
```

```
#save clustering results to csv file
write.table(clus$clustering,file="C:/Drugs_l2s/results/pam_results3.csv",
sep=",")
```

B.6 Σύγκριση Αποτελεσμάτων Αλγορίθμων

```
#set working directory
setwd("C:/Drugs_l2s/")

#load necessary packages
library(clValid)

#read compounds(data) matrix
x        <-        read.table("C:/Drugs_l2s/L2S_compounds_morgan_fps.dat",
sep=,header=FALSE,row.names=1)

#run clValid to validate cluster results
myVal<-clValid(x,         5:8,clMethods=c("kmeans","pam","agnes","diana"),
metric="euclidean",method="ward",validation="internal")

#plot validation results
plot(myVal)

#display algorithms with optimal scores
optimalScores(myVal)

#find similarity betwween clusterings
cluster_similarity(agnes_ward,diana,similarity=c("jaccard"))
cluster_similarity(agnes_ward,diana,similarity=c("rand"))
cluster_similarity(agnes_ward,agnes_complete,similarity=c("jaccard"))
cluster_similarity(agnes_ward,agnes_complete,similarity=c("rand"))
cluster_similarity(agnes_ward,kmeans,similarity=c("jaccard"))
cluster_similarity(agnes_ward,kmeans,similarity=c("rand"))
cluster_similarity(agnes_ward,pam,similarity=c("jaccard"))
cluster_similarity(agnes_ward,pam,similarity=c("rand"))

cluster_similarity(diana,agnes_complete,similarity=c("jaccard"))
cluster_similarity(diana,agnes_complete,similarity=c("rand"))
cluster_similarity(diana,kmeans,similarity=c("jaccard"))
cluster_similarity(diana,kmeans,similarity=c("rand"))
cluster_similarity(diana,pam,similarity=c("jaccard"))
cluster_similarity(diana,pam,similarity=c("rand"))
cluster_similarity(agnes_complete,kmeans,similarity=c("jaccard"))
cluster_similarity(agnes_complete,kmeans,similarity=c("rand"))
cluster_similarity(agnes_complete,pam,similarity=c("jaccard"))
cluster_similarity(agnes_complete,pam,similarity=c("rand"))

cluster_similarity(kmeans,pam,similarity=c("jaccard"))
cluster_similarity(kmeans,pam,similarity=c("rand"))
```

Παράρτημα Γ

Πίνακες Αποτελεσμάτων

Γ.1 AGNES (complete, cutoff=0.85)

Compound ID	Cluster ID	Compound ID	Cluster ID
ChemSpider_3949	1	ChemSpider_13852819	9
ChemSpider_4455	2	ChemSpider_4534998	9
ChemSpider_2034	2	ChemSpider_4514937	9
ChemSpider_2908	2	ChemSpider_10442628	10
ChemSpider_2347	2	ChemSpider_15510	10
ChemSpider_3263	2	ChemSpider_4827	10
ChemSpider_431	2	ChemSpider_10442212	10
ChemSpider_25043757	2	ChemSpider_5332	11
ChemSpider_3821	2	ChemSpider_2383	11
ChemSpider_4047	2	ChemSpider_54790	11
ChemSpider_3276	2	ChemSpider_14410	11
ChemSpider_5355	2	ChemSpider_16740595	12
ChemSpider_5293368	3	ChemSpider_5530	12
ChemSpider_4444479	3	ChemSpider_3513	13
ChemSpider_4470631	3	ChemSpider_2457	13
ChemSpider_5382	4	ChemSpider_3741	14
ChemSpider_5253	4	ChemSpider_2006532	15
ChemSpider_4757	4	ChemSpider_49179	15
ChemSpider_9048	4	ChemSpider_4895	16
ChemSpider_56598	4	ChemSpider_54632	16

ChemSpider_3779	4	ChemSpider_5533	16
ChemSpider_4585	4	ChemSpider_3269	17
ChemSpider_3568	4	ChemSpider_4027	17
ChemSpider_2699	4	ChemSpider_4663	17
ChemSpider_2077	5	ChemSpider_2289101	17
ChemSpider_4481878	5	ChemSpider_2312	17
ChemSpider_4015	5	ChemSpider_4060	18
ChemSpider_1999	6	ChemSpider_4040	18
ChemSpider_2340731	6	ChemSpider_54810	19
ChemSpider_4968	6	ChemSpider_4444507	19
ChemSpider_2075	7	ChemSpider_393589	19
ChemSpider_15520	7	ChemSpider_54822	20
ChemSpider_31017	7	ChemSpider_110575	20
ChemSpider_129277	7	ChemSpider_61881	21
ChemSpider_2669	7	ChemSpider_3871	21
ChemSpider_4470984	7	ChemSpider_2157	22
ChemSpider_3438	7	ChemSpider_3328	22
ChemSpider_4087	7	ChemSpider_3009	22
ChemSpider_4445173	7	ChemSpider_4354	23
ChemSpider_580849	7	ChemSpider_4645	24
ChemSpider_2074	8	ChemSpider_4395710	25
ChemSpider_5454	8	ChemSpider_2068	26
		ChemSpider_4447672	27

Γ.2 AGNES (complete, cutoff=0.958)

Compound ID	Cluster ID	Compound ID	Cluster ID
ChemSpider_3949	1	ChemSpider_3568	2
ChemSpider_2068	1	ChemSpider_14410	2
ChemSpider_4455	2	ChemSpider_110575	2
ChemSpider_5382	2	ChemSpider_4087	2
ChemSpider_2034	2	ChemSpider_10442212	2
ChemSpider_2075	2	ChemSpider_4445173	2
ChemSpider_2074	2	ChemSpider_3276	2
ChemSpider_10442628	2	ChemSpider_580849	2
ChemSpider_5253	2	ChemSpider_5355	2
ChemSpider_2908	2	ChemSpider_2699	2
ChemSpider_15510	2	ChemSpider_5293368	3
ChemSpider_4757	2	ChemSpider_4444479	3
ChemSpider_5332	2	ChemSpider_2006532	3
ChemSpider_15520	2	ChemSpider_54810	3
ChemSpider_31017	2	ChemSpider_2157	3
ChemSpider_2347	2	ChemSpider_3328	3
ChemSpider_9048	2	ChemSpider_4444507	3
ChemSpider_3513	2	ChemSpider_3009	3
ChemSpider_3741	2	ChemSpider_4470631	3
ChemSpider_129277	2	ChemSpider_4395710	3
ChemSpider_2457	2	ChemSpider_49179	3
ChemSpider_2669	2	ChemSpider_393589	3
ChemSpider_56598	2	ChemSpider_2077	4

ChemSpider_3263	2	ChemSpider_1999	4
ChemSpider_431	2	ChemSpider_13852819	4
ChemSpider_4895	2	ChemSpider_4481878	4
ChemSpider_4060	2	ChemSpider_4015	4
ChemSpider_54822	2	ChemSpider_16740595	4
ChemSpider_25043757	2	ChemSpider_4534998	4
ChemSpider_4470984	2	ChemSpider_4514937	4
ChemSpider_3779	2	ChemSpider_3269	4
ChemSpider_2383	2	ChemSpider_5530	4
ChemSpider_3821	2	ChemSpider_61881	4
ChemSpider_54790	2	ChemSpider_4027	4
ChemSpider_4645	2	ChemSpider_2340731	4
ChemSpider_5454	2	ChemSpider_4663	4
ChemSpider_4047	2	ChemSpider_2289101	4
ChemSpider_4827	2	ChemSpider_2312	4
ChemSpider_4585	2	ChemSpider_3871	4
ChemSpider_54632	2	ChemSpider_4968	4
ChemSpider_4040	2	ChemSpider_4354	5
ChemSpider_3438	2	ChemSpider_4447672	5
ChemSpider_5533	2		

Γ.3 AGNES (Ward's, cutoff=0.85)

Compound ID	Cluster ID	Compound ID	Cluster ID
ChemSpider_3949	1	ChemSpider_4087	14
ChemSpider_4455	2	ChemSpider_15520	15
ChemSpider_2347	2	ChemSpider_31017	15
ChemSpider_25043757	2	ChemSpider_3438	15
ChemSpider_3821	2	ChemSpider_4534998	16
ChemSpider_5293368	3	ChemSpider_4514937	16
ChemSpider_4444479	3	ChemSpider_9048	17
ChemSpider_5382	4	ChemSpider_5454	17
ChemSpider_3568	4	ChemSpider_3513	18
ChemSpider_2699	4	ChemSpider_2457	18
ChemSpider_2034	5	ChemSpider_3741	19
ChemSpider_4047	5	ChemSpider_129277	20
ChemSpider_5355	5	ChemSpider_2669	20
ChemSpider_2077	6	ChemSpider_2006532	21
ChemSpider_4481878	6	ChemSpider_49179	21
ChemSpider_4015	6	ChemSpider_431	22
ChemSpider_1999	7	ChemSpider_110575	22
ChemSpider_4968	7	ChemSpider_4895	23
ChemSpider_2075	8	ChemSpider_54632	23
ChemSpider_4470984	8	ChemSpider_5533	23
ChemSpider_4445173	8	ChemSpider_3269	24
ChemSpider_580849	8	ChemSpider_54822	24
ChemSpider_2074	9	ChemSpider_2289101	24
ChemSpider_16740595	9	ChemSpider_4060	25

ChemSpider_4663	9	ChemSpider_4040	25
ChemSpider_13852819	10	ChemSpider_5530	26
ChemSpider_2340731	10	ChemSpider_54810	27
ChemSpider_10442628	11	ChemSpider_393589	27
ChemSpider_15510	11	ChemSpider_61881	28
ChemSpider_4827	11	ChemSpider_3871	28
ChemSpider_10442212	11	ChemSpider_2157	29
ChemSpider_5253	12	ChemSpider_4354	30
ChemSpider_4757	12	ChemSpider_3328	31
ChemSpider_56598	12	ChemSpider_3009	31
ChemSpider_3779	12	ChemSpider_4444507	32
ChemSpider_4585	12	ChemSpider_4027	33
ChemSpider_2908	13	ChemSpider_2312	33
ChemSpider_3263	13	ChemSpider_4645	34
ChemSpider_3276	13	ChemSpider_4470631	35
ChemSpider_5332	14	ChemSpider_4395710	36
ChemSpider_2383	14	ChemSpider_2068	37
ChemSpider_54790	14	ChemSpider_4447672	38
ChemSpider_14410	14		

Γ.4 AGNES (Ward's, cutoff=1.37)

Compound ID	Cluster ID	Compound ID	Cluster ID
ChemSpider_3949	1	ChemSpider_15510	3
ChemSpider_5293368	1	ChemSpider_4757	3
ChemSpider_2074	1	ChemSpider_9048	3
ChemSpider_4444479	1	ChemSpider_56598	3
ChemSpider_16740595	1	ChemSpider_4060	3
ChemSpider_4534998	1	ChemSpider_4470984	3
ChemSpider_4514937	1	ChemSpider_3779	3
ChemSpider_3513	1	ChemSpider_5454	3
ChemSpider_3741	1	ChemSpider_4827	3
ChemSpider_2006532	1	ChemSpider_4585	3
ChemSpider_2457	1	ChemSpider_4040	3
ChemSpider_4895	1	ChemSpider_3568	3
ChemSpider_5530	1	ChemSpider_10442212	3
ChemSpider_54810	1	ChemSpider_4445173	3
ChemSpider_2157	1	ChemSpider_580849	3
ChemSpider_4354	1	ChemSpider_2699	3
ChemSpider_3328	1	ChemSpider_2077	4
ChemSpider_4444507	1	ChemSpider_1999	4
ChemSpider_4645	1	ChemSpider_13852819	4
ChemSpider_3009	1	ChemSpider_4481878	4
ChemSpider_4470631	1	ChemSpider_4015	4
ChemSpider_4395710	1	ChemSpider_431	4

ChemSpider_54632	1	ChemSpider_3269	4
ChemSpider_4663	1	ChemSpider_54822	4
ChemSpider_49179	1	ChemSpider_61881	4
ChemSpider_393589	1	ChemSpider_4027	4
ChemSpider_5533	1	ChemSpider_2340731	4
ChemSpider_2068	1	ChemSpider_2289101	4
ChemSpider_4447672	1	ChemSpider_110575	4
ChemSpider_4455	2	ChemSpider_2312	4
ChemSpider_2034	2	ChemSpider_3871	4
ChemSpider_2908	2	ChemSpider_4968	4
ChemSpider_2347	2	ChemSpider_5332	5
ChemSpider_3263	2	ChemSpider_15520	5
ChemSpider_25043757	2	ChemSpider_31017	5
ChemSpider_3821	2	ChemSpider_129277	5
ChemSpider_4047	2	ChemSpider_2669	5
ChemSpider_3276	2	ChemSpider_2383	5
ChemSpider_5355	2	ChemSpider_54790	5
ChemSpider_5382	3	ChemSpider_3438	5
ChemSpider_2075	3	ChemSpider_14410	5
ChemSpider_10442628	3	ChemSpider_4087	5
ChemSpider_5253	3		

Γ.5 DIANA (cutoff=0.85)

Compound ID	Cluster ID	Compound ID	Cluster ID
ChemSpider_3949	1	ChemSpider_15520	14
ChemSpider_4455	2	ChemSpider_31017	14
ChemSpider_2034	2	ChemSpider_3438	14
ChemSpider_2908	2	ChemSpider_14410	14
ChemSpider_3263	2	ChemSpider_4087	14
ChemSpider_25043757	2	ChemSpider_2347	15
ChemSpider_3821	2	ChemSpider_54632	15
ChemSpider_4047	2	ChemSpider_110575	15
ChemSpider_3276	2	ChemSpider_16740595	16
ChemSpider_5355	2	ChemSpider_4663	16
ChemSpider_5293368	3	ChemSpider_4534998	17
ChemSpider_4444479	3	ChemSpider_4514937	17
ChemSpider_4470631	3	ChemSpider_9048	18
ChemSpider_49179	3	ChemSpider_5454	18
ChemSpider_5382	4	ChemSpider_3513	19
ChemSpider_5253	4	ChemSpider_3741	20
ChemSpider_4757	4	ChemSpider_129277	21
ChemSpider_56598	4	ChemSpider_2669	21
ChemSpider_3779	4	ChemSpider_2006532	22
ChemSpider_4585	4	ChemSpider_2457	23
ChemSpider_3568	4	ChemSpider_2157	23
ChemSpider_2699	4	ChemSpider_4895	24
ChemSpider_2077	5	ChemSpider_5533	24
ChemSpider_431	5	ChemSpider_3269	25

ChemSpider_1999	6	ChemSpider_2289101	25
ChemSpider_2340731	6	ChemSpider_4060	26
ChemSpider_4968	6	ChemSpider_4040	26
ChemSpider_2075	7	ChemSpider_5530	27
ChemSpider_4470984	7	ChemSpider_54810	28
ChemSpider_4445173	7	ChemSpider_4444507	28
ChemSpider_580849	7	ChemSpider_393589	28
ChemSpider_2074	8	ChemSpider_54822	29
ChemSpider_13852819	9	ChemSpider_61881	30
ChemSpider_10442628	10	ChemSpider_4354	31
ChemSpider_15510	10	ChemSpider_3328	32
ChemSpider_4827	10	ChemSpider_3009	32
ChemSpider_10442212	10	ChemSpider_4027	33
ChemSpider_4481878	11	ChemSpider_2312	33
ChemSpider_4015	12	ChemSpider_4645	34
ChemSpider_3871	12	ChemSpider_4395710	35
ChemSpider_5332	13	ChemSpider_2068	36
ChemSpider_2383	13	ChemSpider_4447672	37
ChemSpider_54790	13		

Γ.6 DIANA (cutoff=0.958)

Compound ID	Cluster ID	Compound ID	Cluster ID
ChemSpider_3949	1	ChemSpider_393589	2
ChemSpider_4455	2	ChemSpider_3438	2
ChemSpider_5382	2	ChemSpider_5533	2
ChemSpider_2034	2	ChemSpider_3568	2
ChemSpider_2077	2	ChemSpider_14410	2
ChemSpider_2075	2	ChemSpider_110575	2
ChemSpider_2074	2	ChemSpider_4087	2
ChemSpider_10442628	2	ChemSpider_10442212	2
ChemSpider_5253	2	ChemSpider_4445173	2
ChemSpider_2908	2	ChemSpider_3276	2
ChemSpider_15510	2	ChemSpider_580849	2
ChemSpider_4757	2	ChemSpider_5355	2
ChemSpider_5332	2	ChemSpider_2699	2
ChemSpider_15520	2	ChemSpider_5293368	3
ChemSpider_31017	2	ChemSpider_4444479	3
ChemSpider_2347	2	ChemSpider_2006532	3
ChemSpider_9048	2	ChemSpider_3328	3
ChemSpider_3513	2	ChemSpider_3009	3
ChemSpider_3741	2	ChemSpider_4470631	3
ChemSpider_129277	2	ChemSpider_4395710	3
ChemSpider_2669	2	ChemSpider_49179	3
ChemSpider_56598	2	ChemSpider_1999	4
ChemSpider_3263	2	ChemSpider_13852819	4

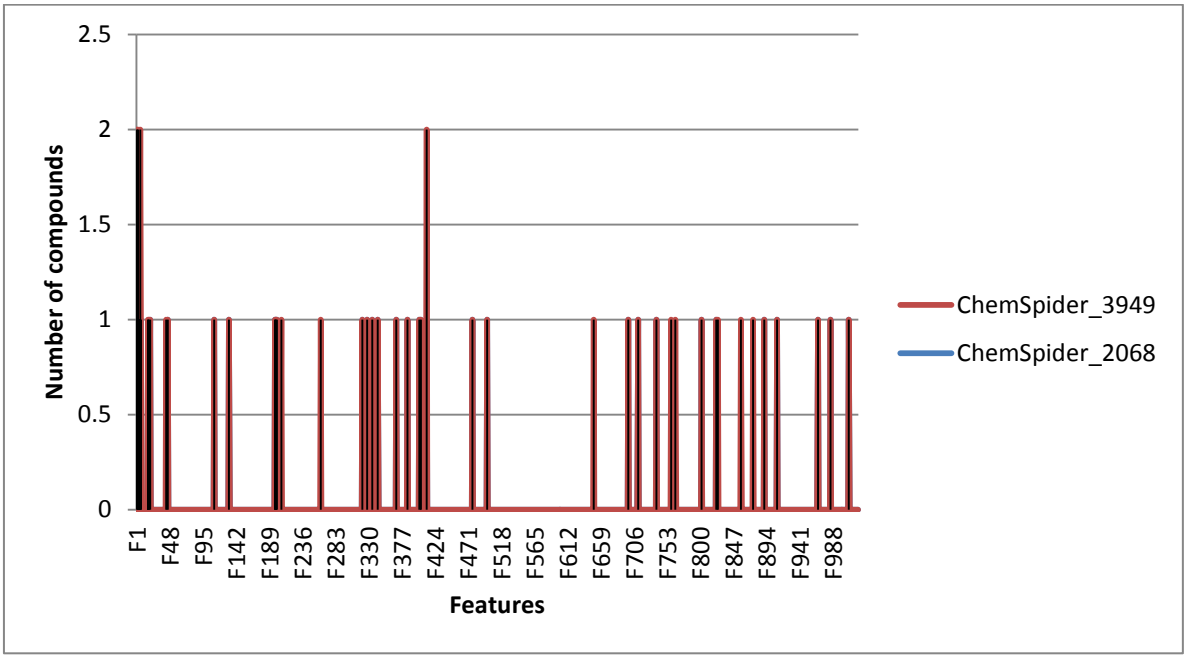
ChemSpider_431	2	ChemSpider_4481878	4
ChemSpider_4895	2	ChemSpider_4015	4
ChemSpider_4060	2	ChemSpider_16740595	4
ChemSpider_54810	2	ChemSpider_4534998	4
ChemSpider_61881	2	ChemSpider_4514937	4
ChemSpider_25043757	2	ChemSpider_2457	4
ChemSpider_4470984	2	ChemSpider_3269	4
ChemSpider_3779	2	ChemSpider_5530	4
ChemSpider_2383	2	ChemSpider_54822	4
ChemSpider_3821	2	ChemSpider_2157	4
ChemSpider_4444507	2	ChemSpider_4027	4
ChemSpider_54790	2	ChemSpider_2340731	4
ChemSpider_4645	2	ChemSpider_4663	4
ChemSpider_5454	2	ChemSpider_2289101	4
ChemSpider_4047	2	ChemSpider_2068	4
ChemSpider_4827	2	ChemSpider_2312	4
ChemSpider_4585	2	ChemSpider_3871	4
ChemSpider_54632	2	ChemSpider_4968	4
ChemSpider_4040	2	ChemSpider_4354	5
		ChemSpider_4447672	5

Παράρτημα Δ

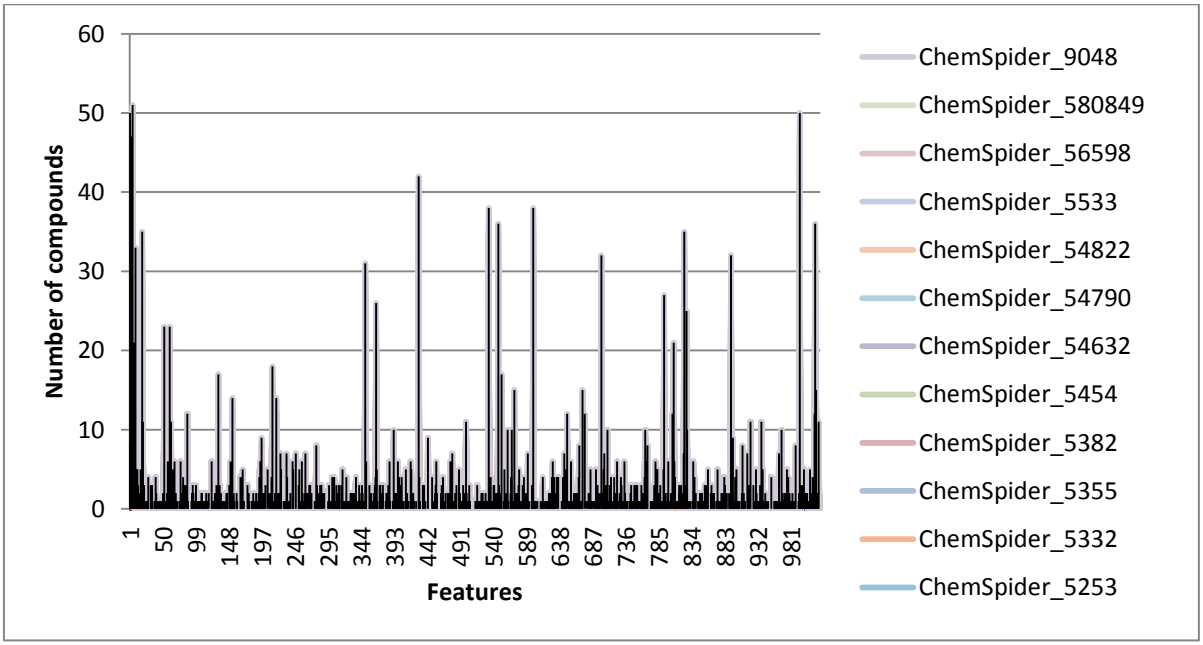
Γραφικές Παραστάσεις Χαρακτηριστικών Συστάδων

Δ.1 Agnes (Complete)

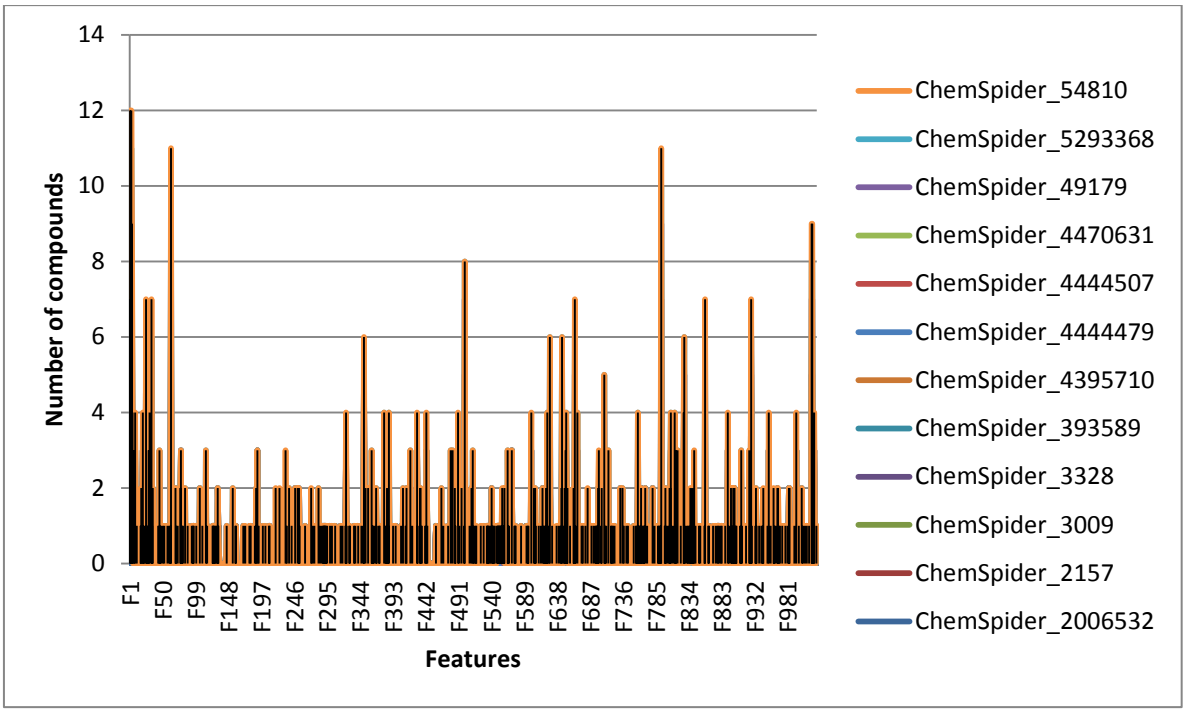
Cluster 1



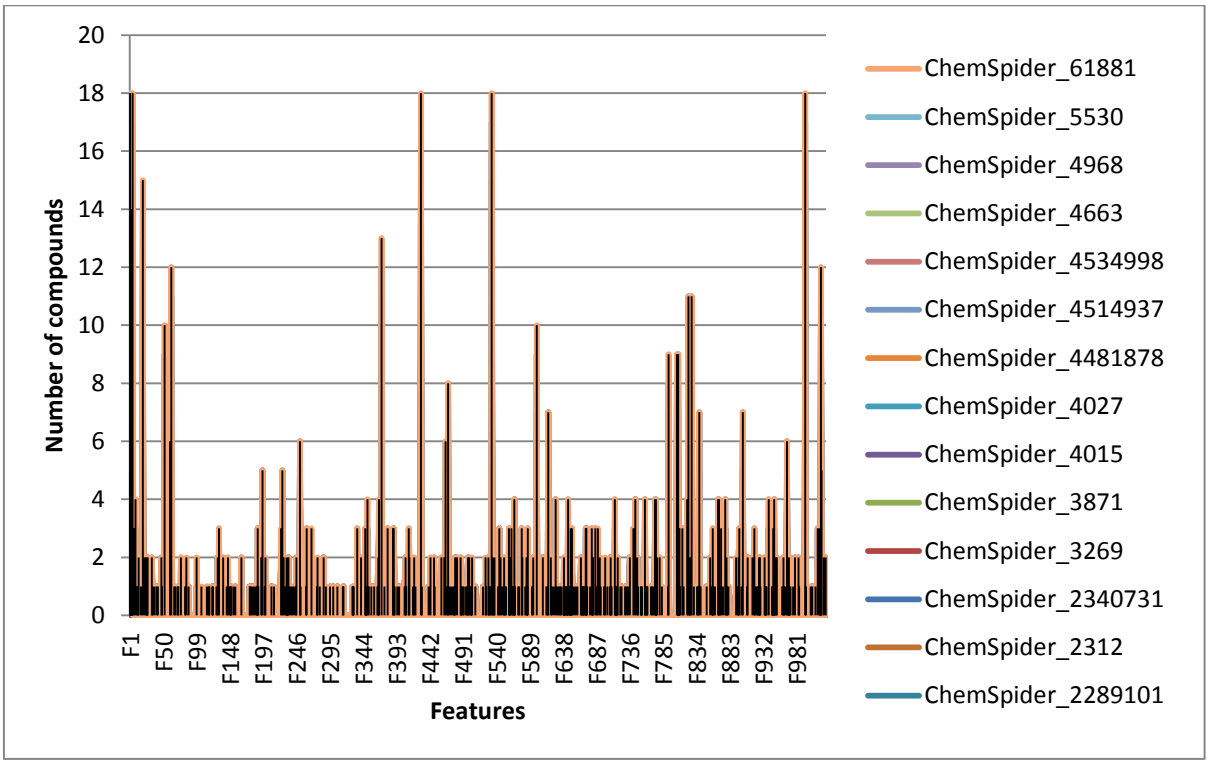
Cluster2



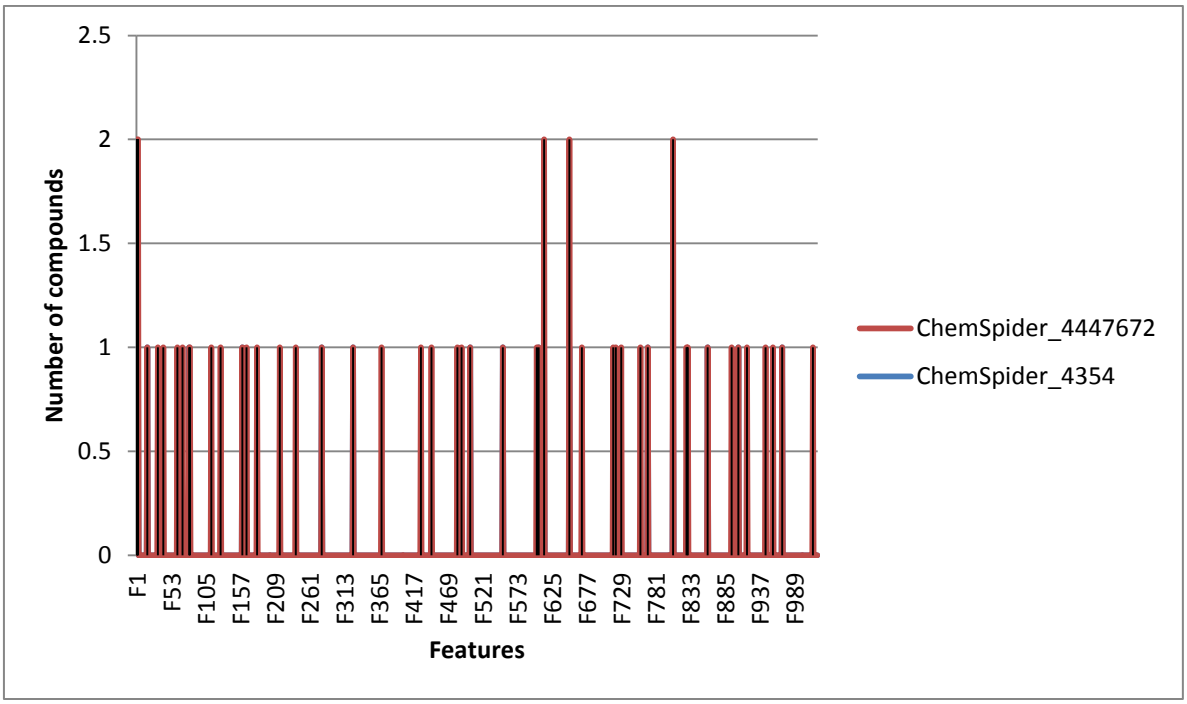
Cluster 3


Number of compounds
Features
0
2
4
6
8
10
12
14
F1
F50
F99
F148
F197
F246
F295
F344
F393
F442
F491
F540
F589
F638
F687
F736
F785
F834
F883
F932
F981
ChemSpider_54810
ChemSpider_5293368
ChemSpider_49179
ChemSpider_4470631
ChemSpider_4444507
ChemSpider_4444479
ChemSpider_4395710
ChemSpider_393589
ChemSpider_3328
ChemSpider_3009
ChemSpider_2157
ChemSpider_2006532

Cluster 4

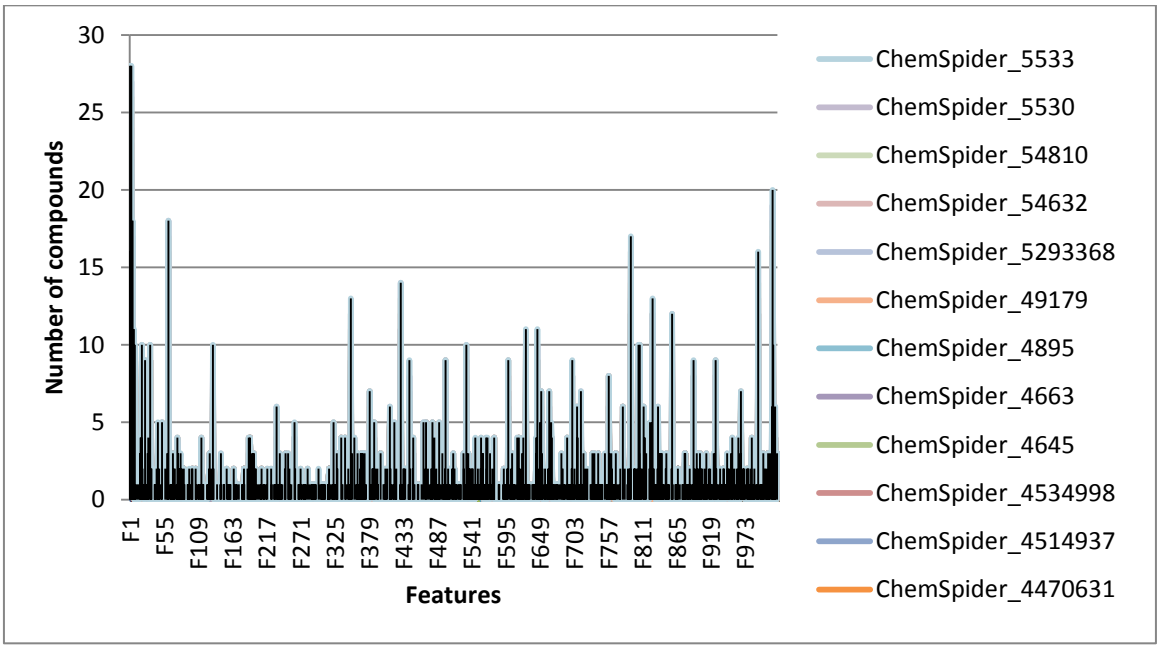

Number of compounds
Features
0
2
4
6
8
10
12
14
16
18
20
F1
F50
F99
F148
F197
F246
F295
F344
F393
F442
F491
F540
F589
F638
F687
F736
F785
F834
F883
F932
F981
ChemSpider_61881
ChemSpider_5530
ChemSpider_4968
ChemSpider_4663
ChemSpider_4534998
ChemSpider_4514937
ChemSpider_4481878
ChemSpider_4027
ChemSpider_4015
ChemSpider_3871
ChemSpider_3269
ChemSpider_2340731
ChemSpider_2312
ChemSpider_2289101

Cluster 5

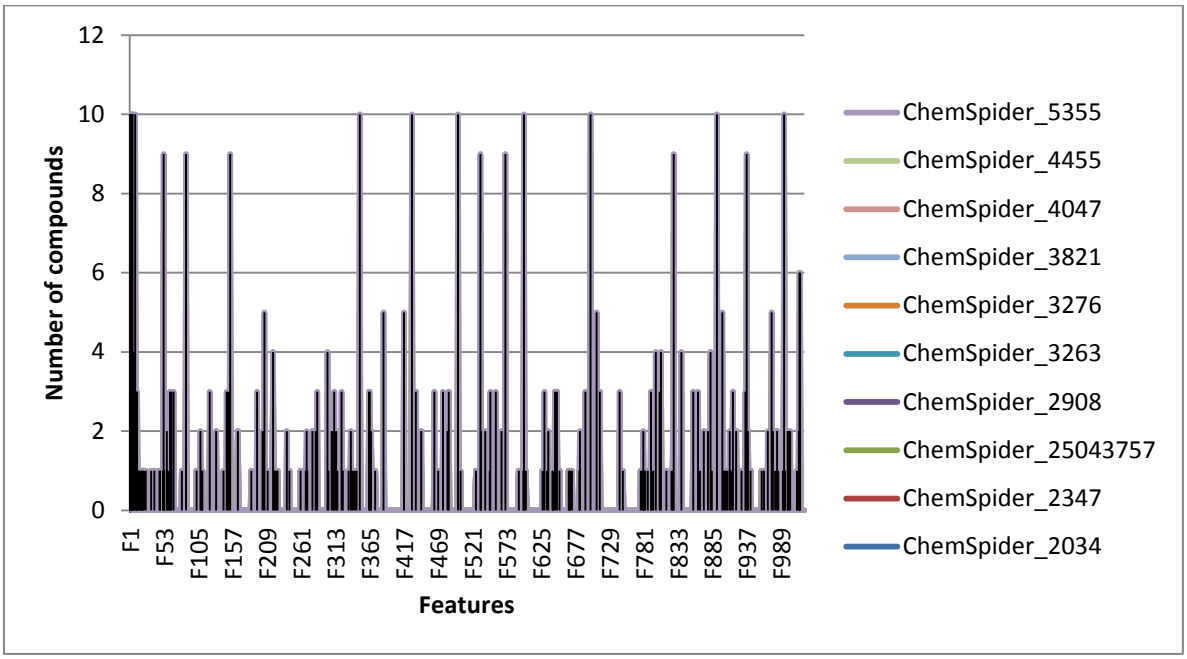


Δ.2 Agnes (Ward's)

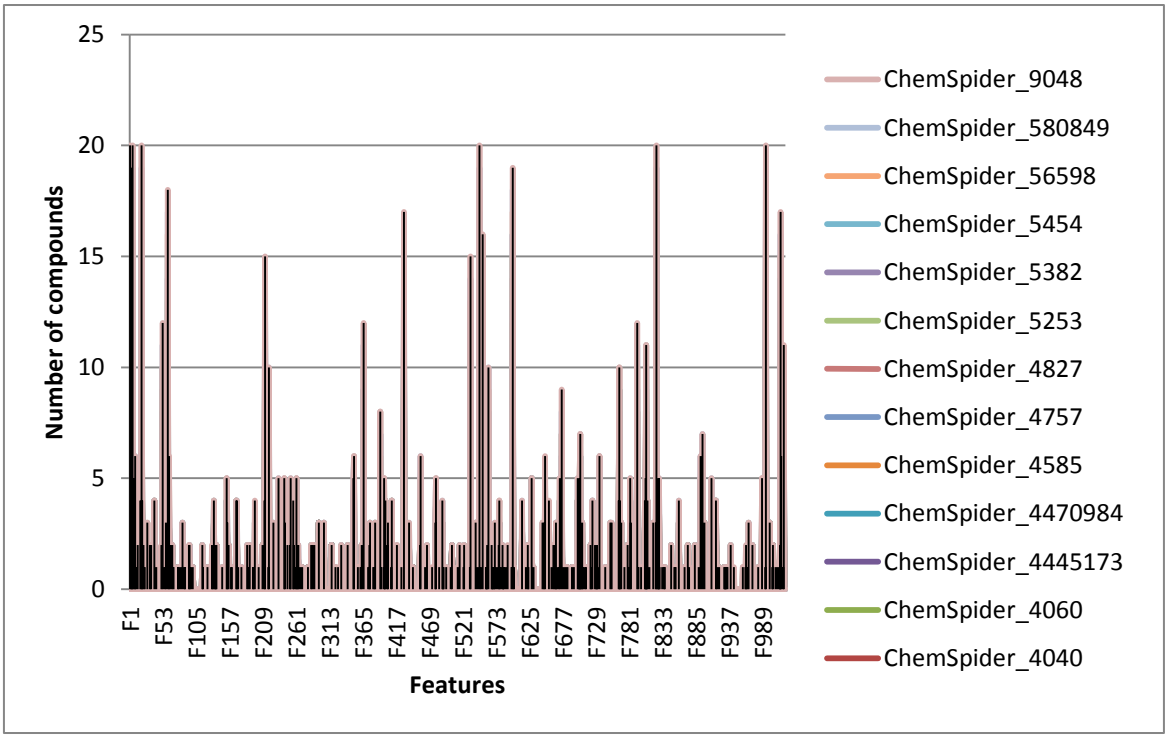
Cluster 1



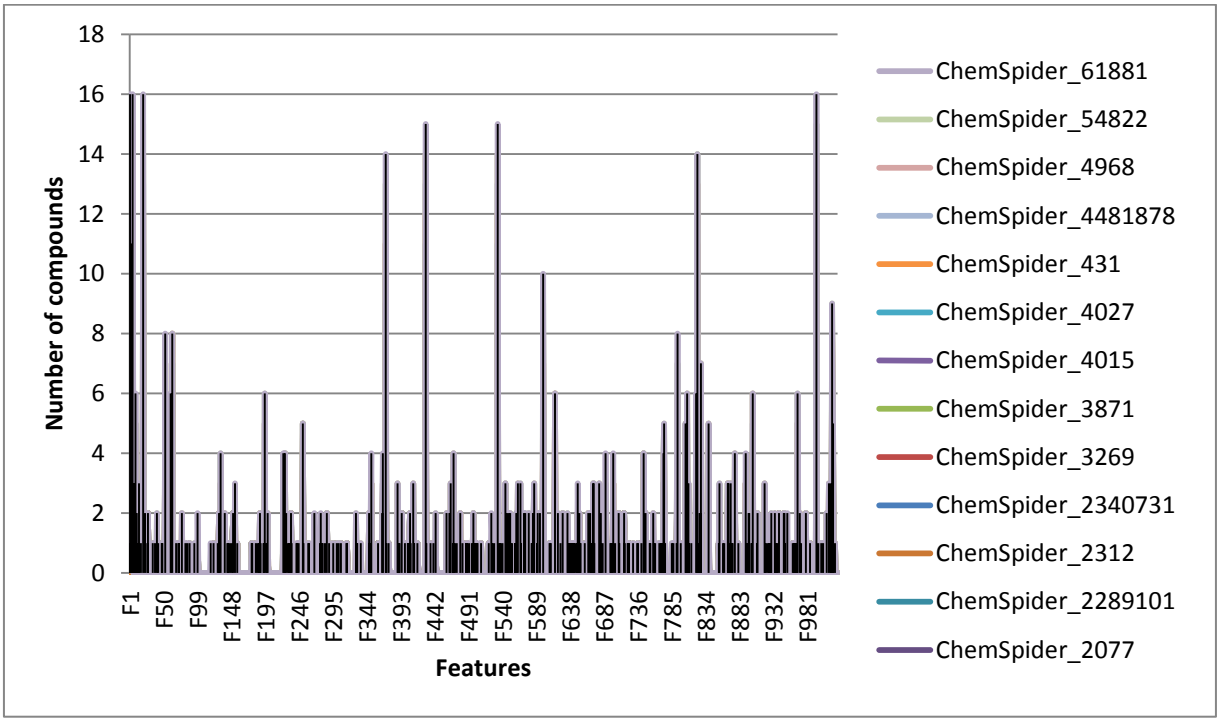
Cluster 2



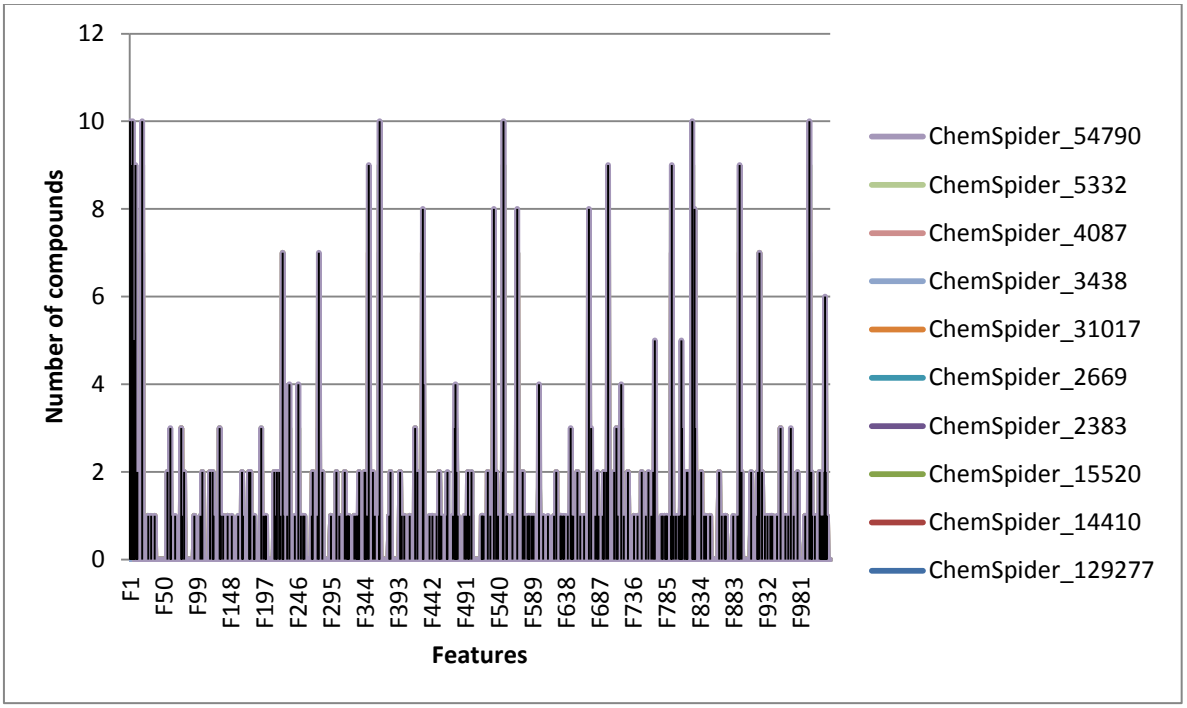
Cluster 3



Cluster 4

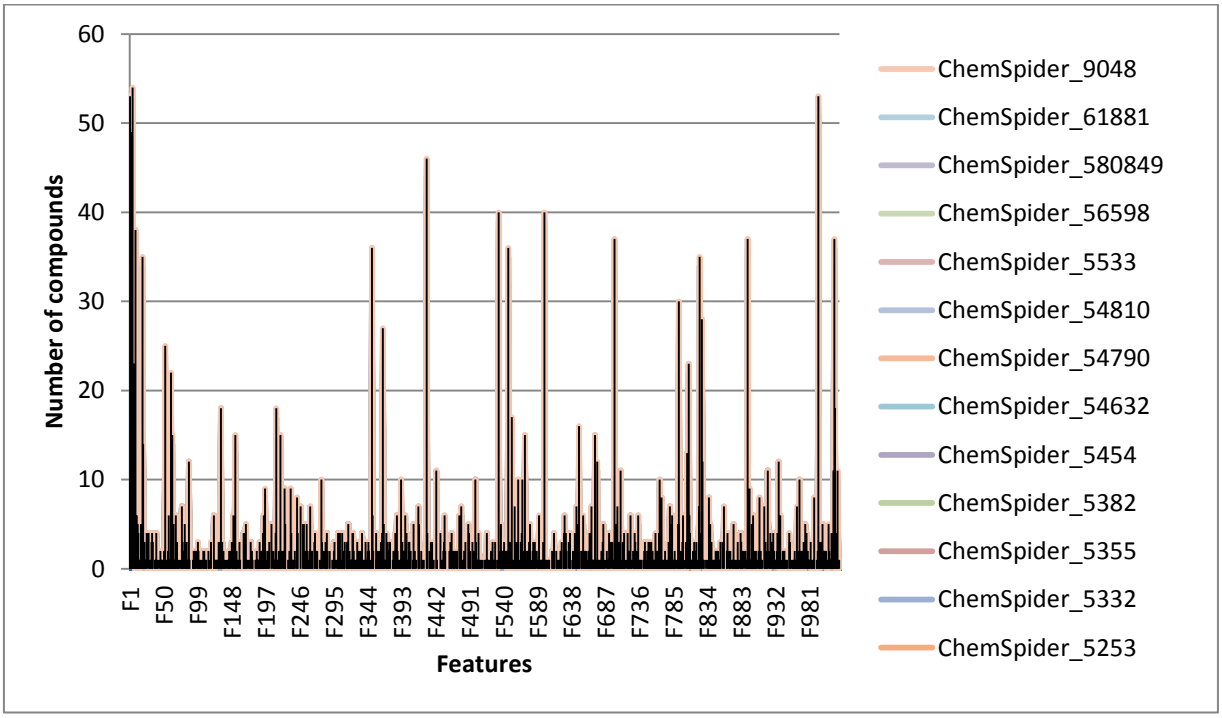


Cluster 5

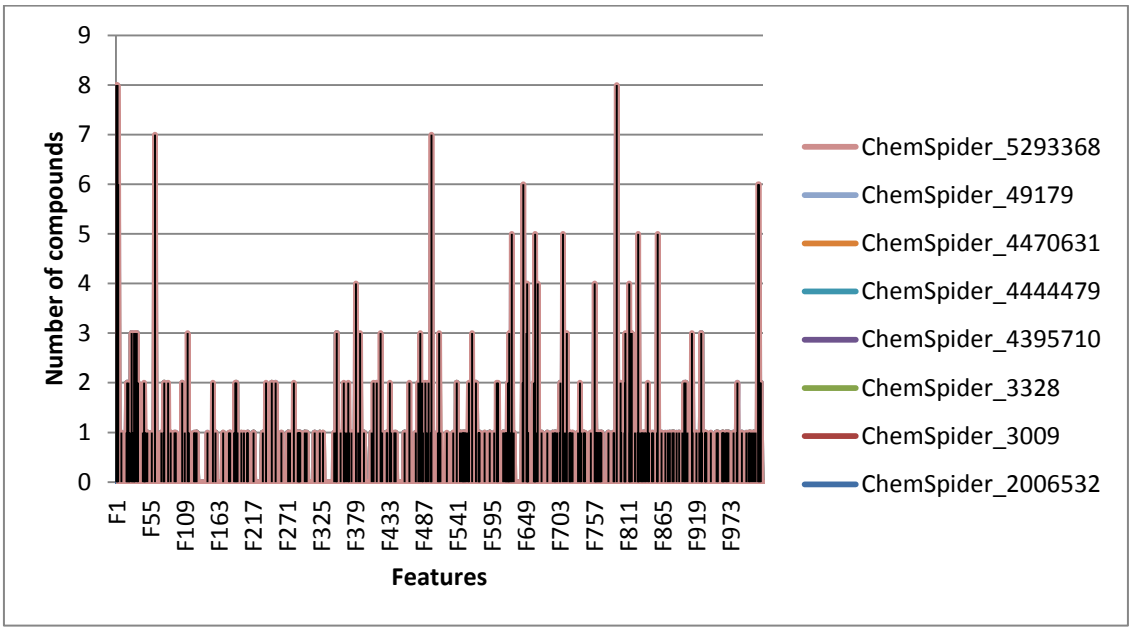


Δ.3 DIANA

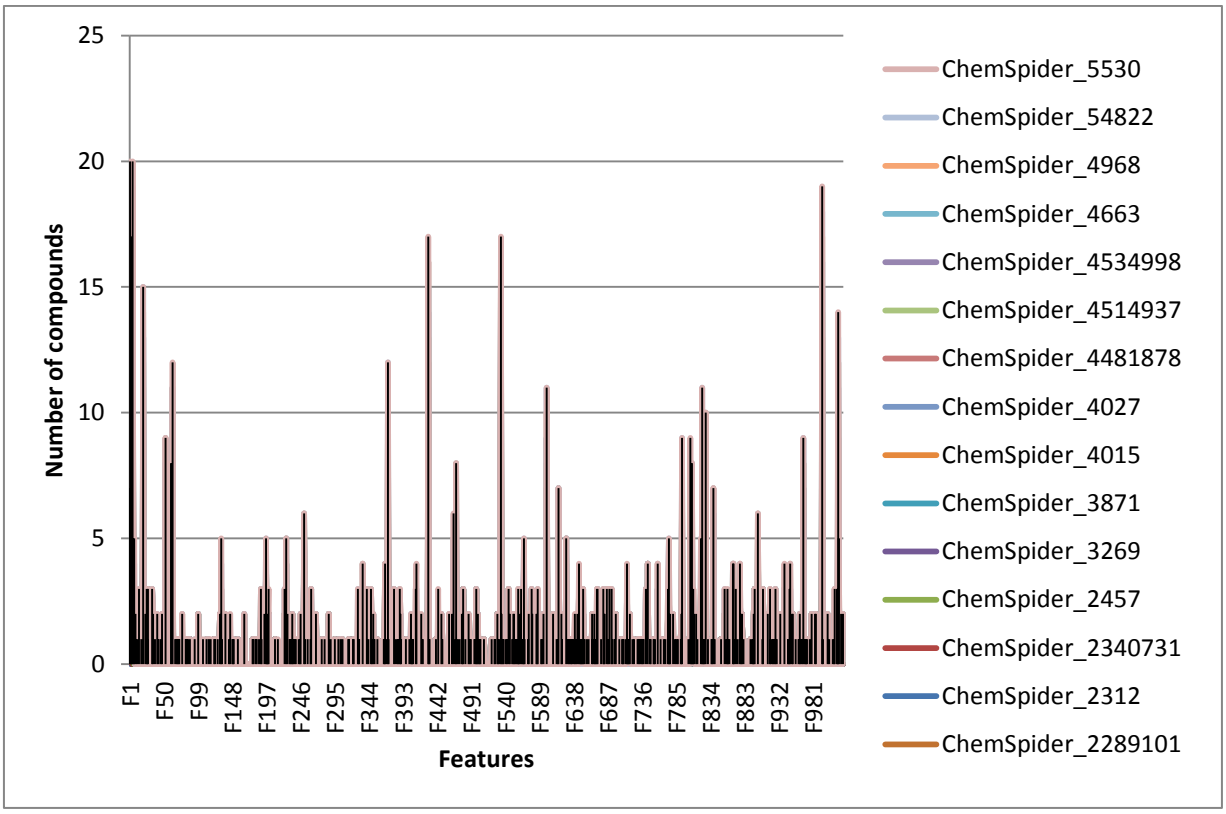
Cluster 2



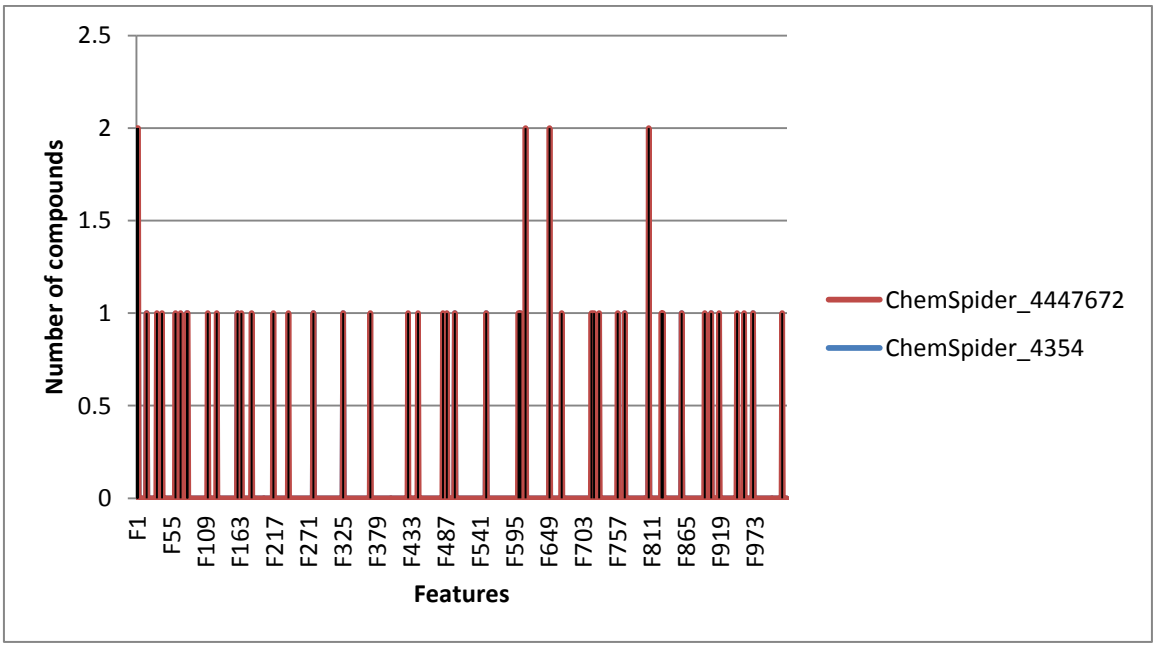
Cluster 3



Cluster 4

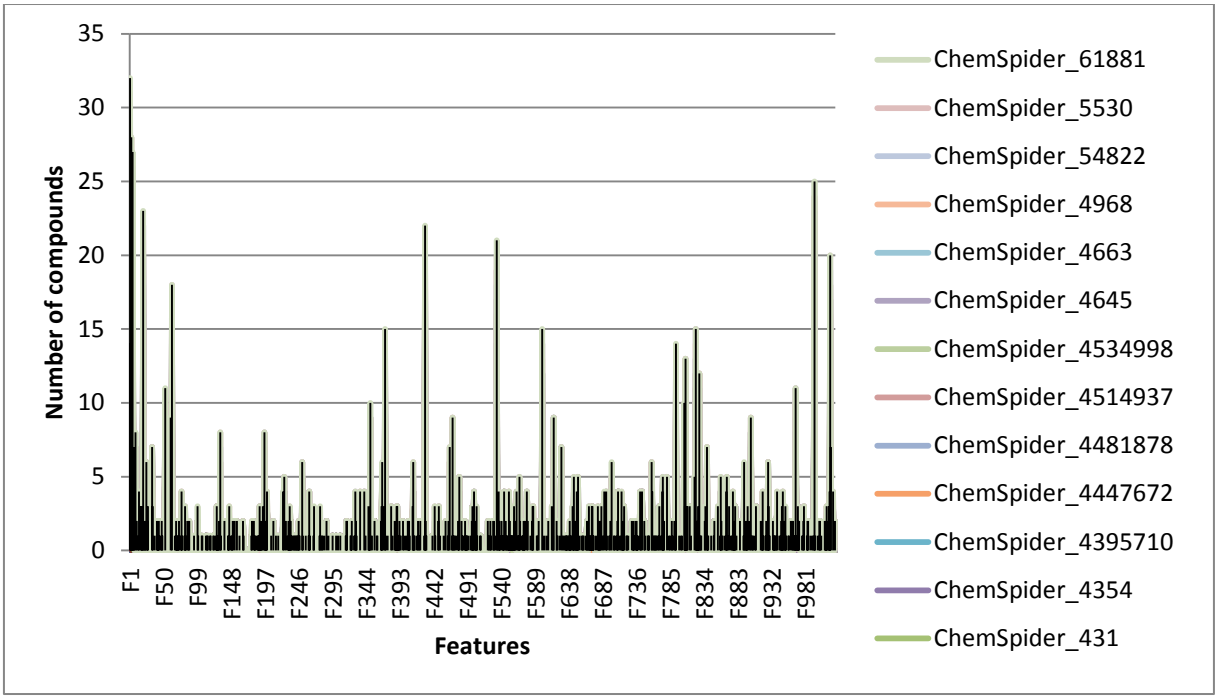


Cluster 5

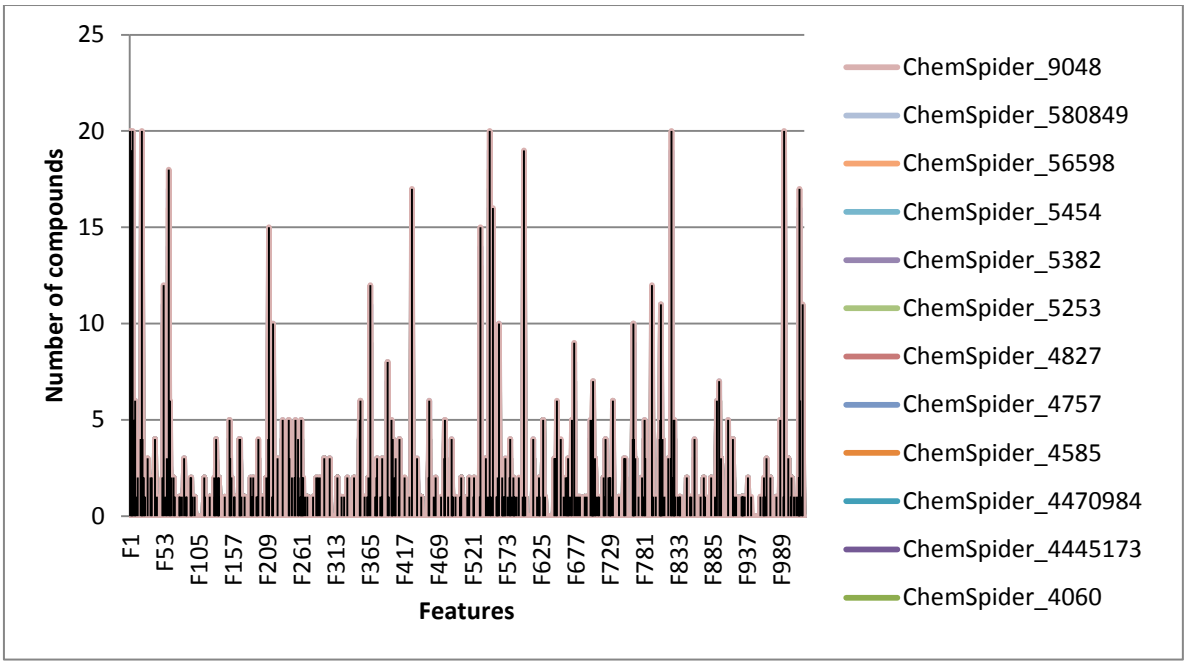


Δ.4 k-means

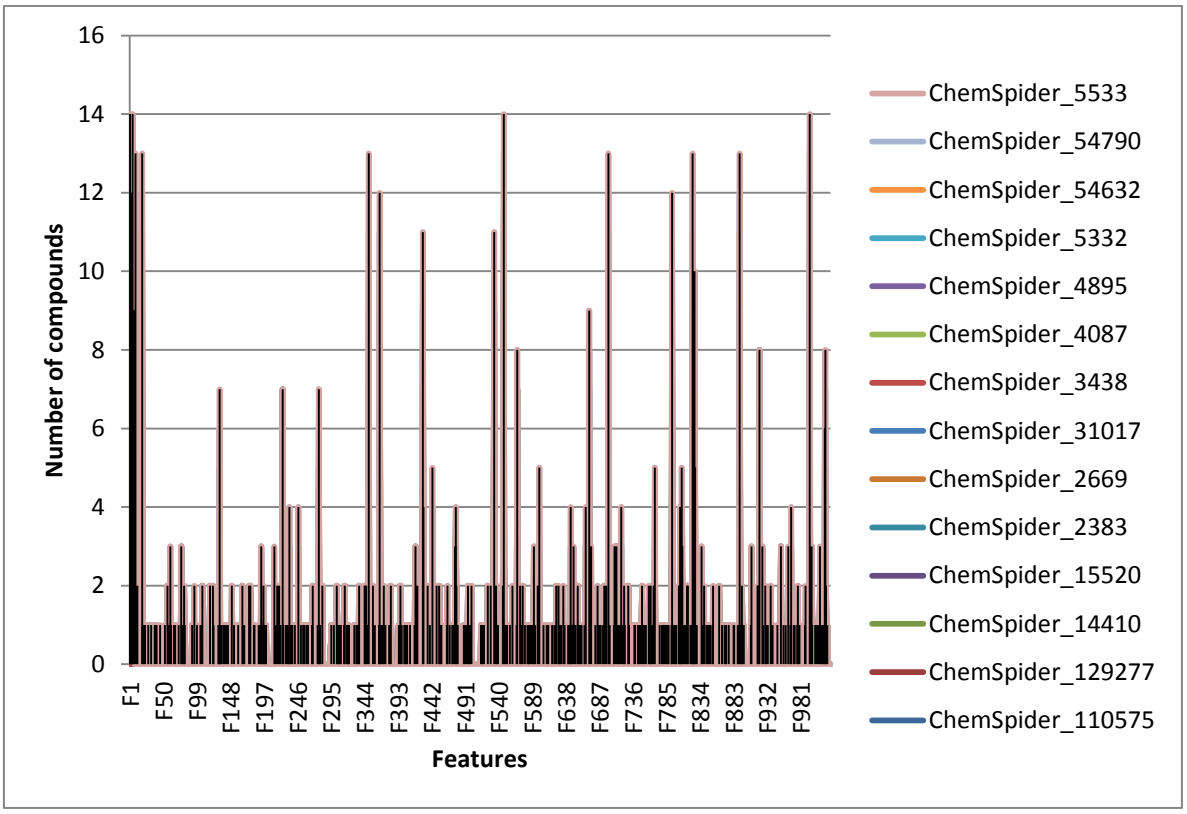
Cluster 1



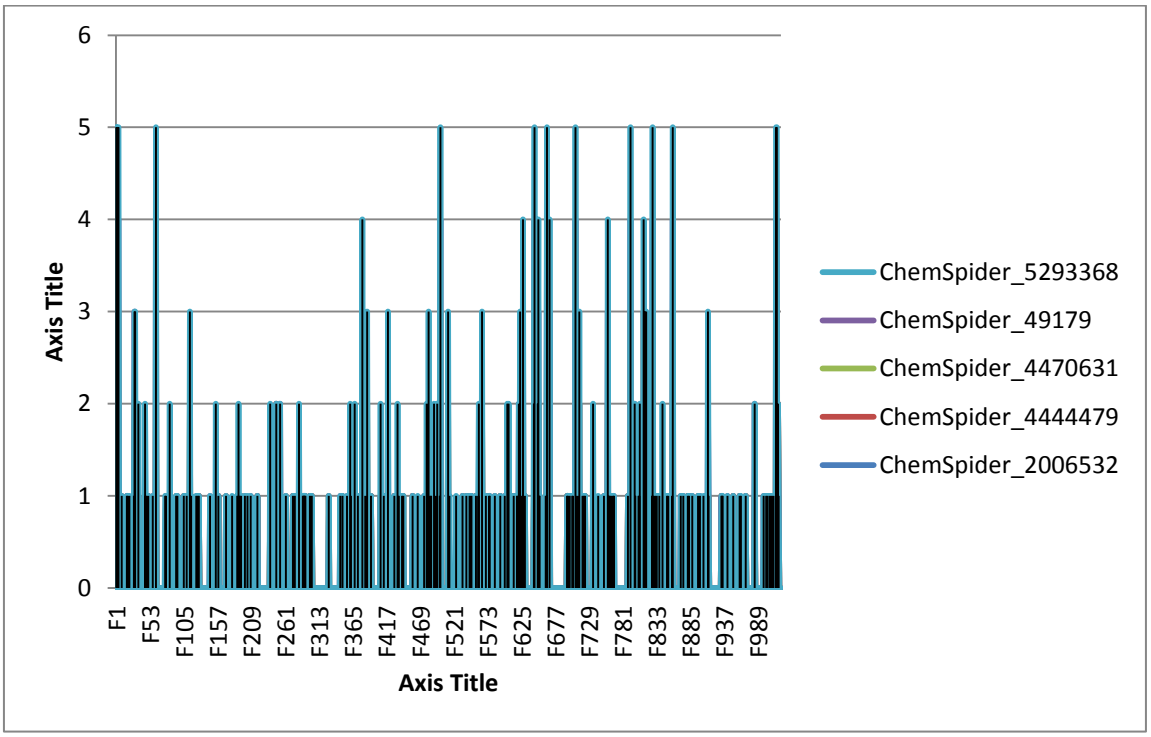
Cluster 2



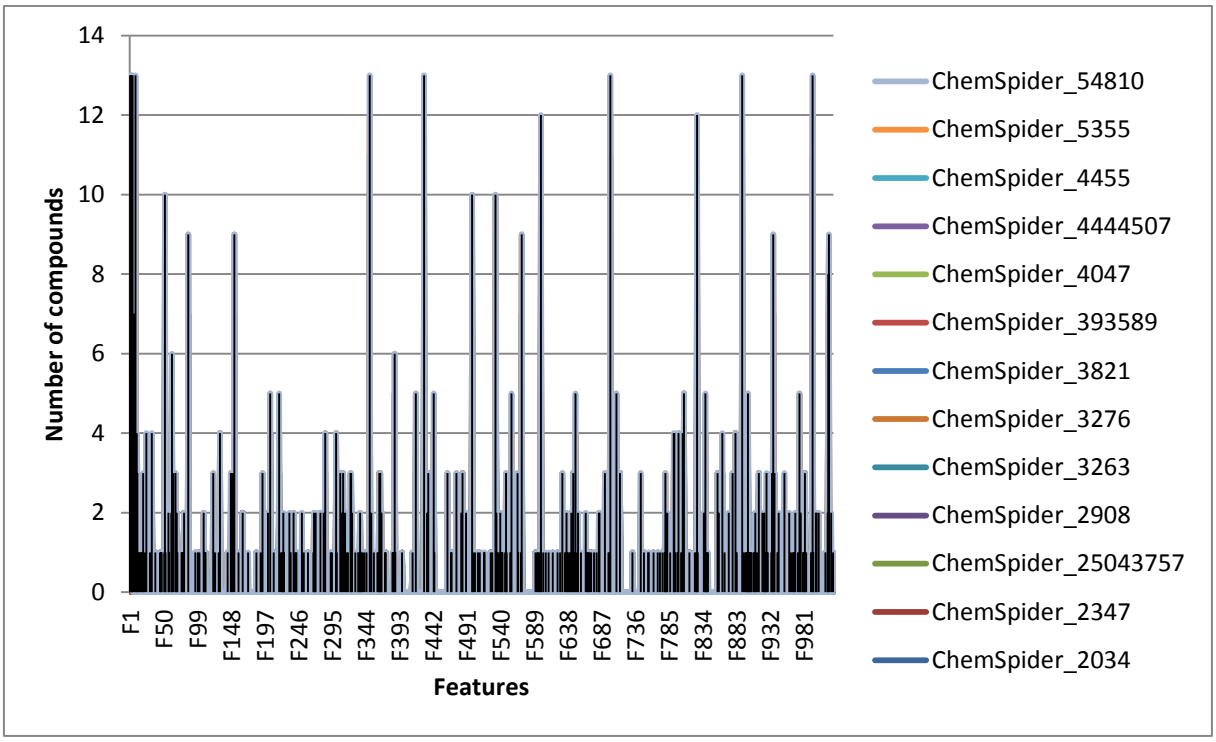
Cluster 3



Cluster 4

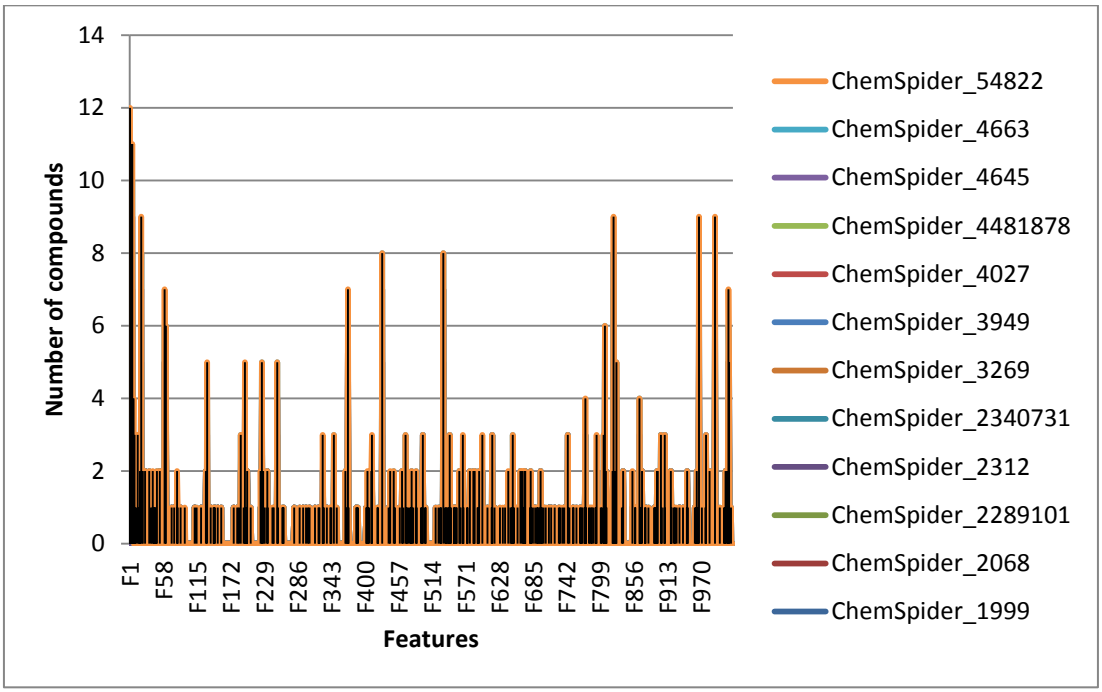


Cluster 5

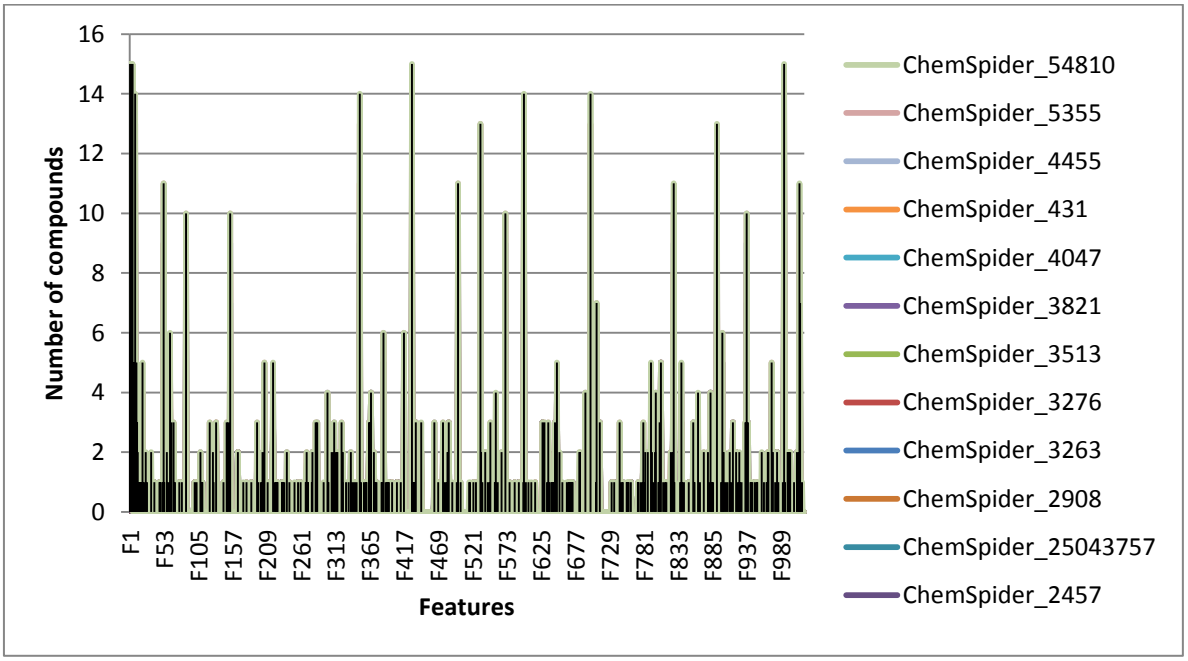


Δ.5 PAM

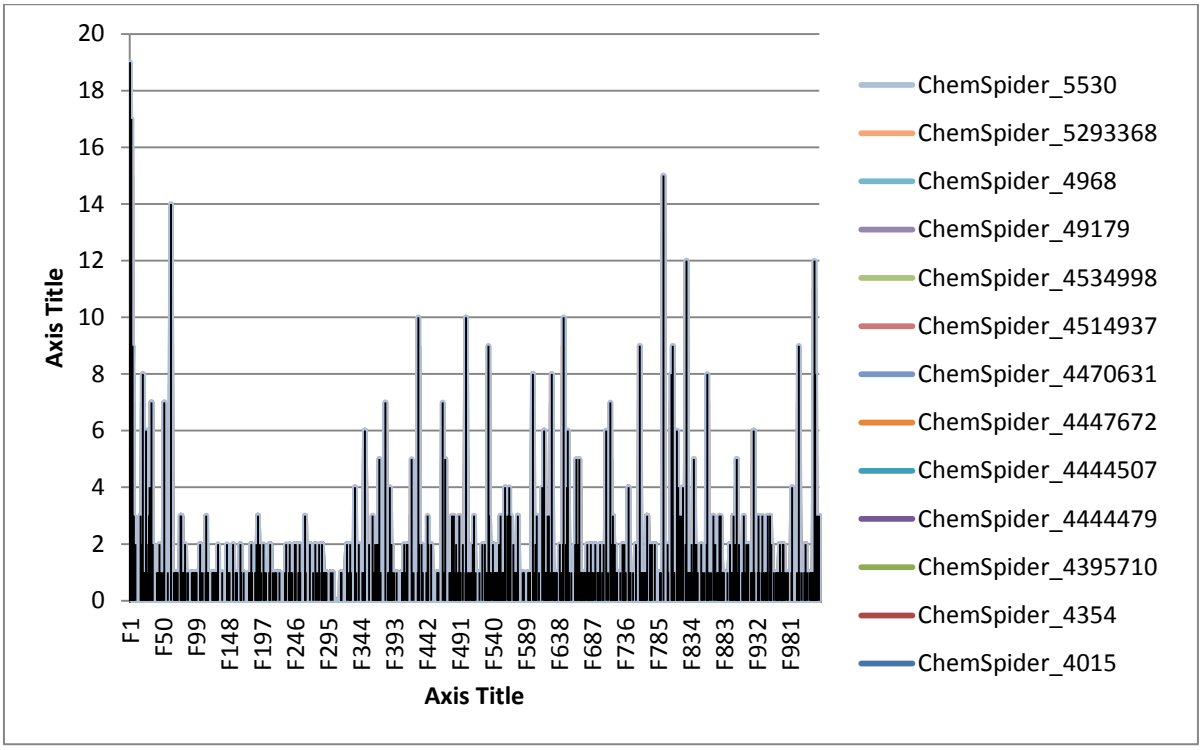
Cluster 1



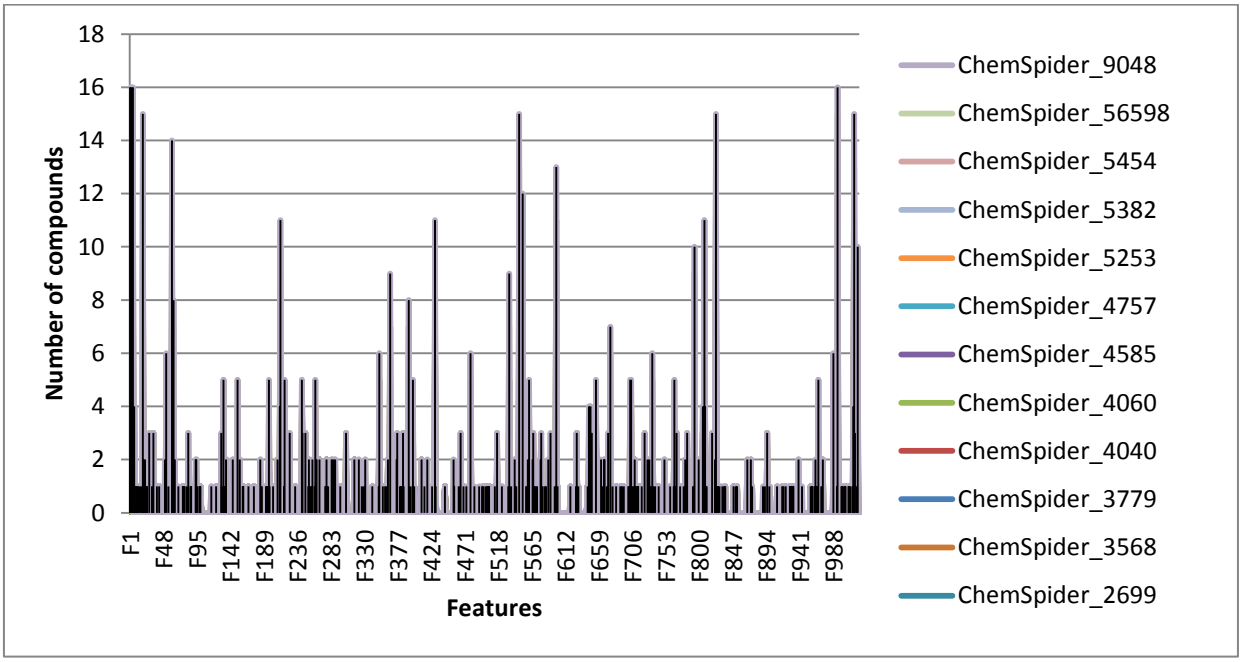
Cluster 2



Cluster 3



Cluster 4



Cluster 5