ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών

Διπλωματική Εργασία

ΜΕΘΟΔΟΙ ΟΠΤΙΚΟΠΟΙΗΣΗΣ ΓΟΝΙΔΙΑΚΩΝ ΔΕΔΟΜΕΝΩΝ

ΣΚΡΕΤΗ ΓΕΩΡΓΙΑ

Επιβλέπων Καθηγητής: Καθηγητής Μιχάλης Ζερβάκης

Εξεταστική Επιτροπή: Καθηγητής Μιχάλης Ζερβάκης
Καθηγητής Κώστας Μπάλας
Καθηγητής Ευριπίδης Πετράκης

Χανιά, Ιούνιος 2012

Στην οικογένειά μου

Ευχαριστίες

Η στήριξη που έλαβα καθ’ όλη τη διάρκεια των σπουδών μου από ανθρώπους που με ενθάρρυναν και μου προσέφεραν τις γνώσεις τους απλόχερα, αποτέλεσε καταλυτικό παράγοντα της επιτυχούς ολοκλήρωσης της παρούσας εργασίας.

Πρώτα από όλους, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Μιχάλη Ζερβάκη για την καθοδήγηση, την υποστήριξη και τις γνώσεις που μου προσέφερε καθ’ όλη τη διάρκεια της διπλωματικής εργασίας στα πλαίσια μιας εποικοδομητικής συνεργασίας.

Θα ήθελα να εκφράσω τις ευχαριστίες μου στην κ. Αικατερίνη Μπέη για την συνεισφορά της στο βιολογικό μέρος της εργασίας καθώς και για την αστείρευτη ενθάρρυνση που λάμβανα σε όλα τα στάδια της διπλωματικής εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τα υπόλοιπα μέλη της εξεταστικής επιτροπής, τους καθηγητές κ. Κώστα Μπάλα και κ. Ευριπίδη Πετράκη.

Ένα ευχαριστώ, ένα χαμόγελο δεν είναι αρκετά για να εκφράσω όλα όσα νιώθω για τους φίλους μου, οι οποίοι ήταν πάντα στο πλευρό μου όλα τα φοιτητικά χρόνια.

Το μεγάλο ευχαριστώ απονέμεται στην οικογένεια μου, γιατί, απλά της ανήκει.

Περίληψη

Η ανάλυση μικροσυστοιχιών DNA επιτρέπει μελέτες γονιδιακής έκφρασης. Το πρότυπο της γονιδιακής έκφρασης που παράγεται, γνωστό ως προφίλ έκφρασης, απεικονίζει το υποσύνολο των γονιδιακών μεταγράφων που εκφράζονται σε ένα κύτταρο ή σε έναν ιστό. Με τη βοήθεια εξειδικευμένης μεθοδολογίας της επιστήμης της βιοπληροφορικής αναλύονται τα προφίλ γονιδιακής έκφρασης και απαντώνται ερωτήματα που σχετίζονται με γονίδια που εκφράζονται σε παθολογικές καταστάσεις. Στην παρούσα εργασία γίνεται μελέτη και υλοποίηση αλγορίθμων γονιδιακής ανάλυσης με στόχο την ομαδοποίηση και οπτικοποίηση γονιδιακών δεδομένων. Η οπτικοποίηση επιτυγχάνεται μέσω της εκτέλεσης δύο διαδοχικών βημάτων. Αρχικά τα πολυδιάστατα δεδομένα ομαδοποιούνται με διαφορετικές μεθόδους, τόσο με τη χρήση τεχνητών νευρωνικών δικτύων όσο και με τη βοήθεια κλασσικών αλγορίθμων κατηγοριοποίησης. Ένα ενδιάμεσο βήμα συνιστά η αξιολόγηση των αποτελεσμάτων της ομαδοποίησης με την εφαρμογή ποικίλων δεικτών εγκυρότητας. Το δεύτερο και τελικό στάδιο περιλαμβάνει την απεικόνιση των ομαδοποιημένων δεδομένων σε ένα χώρο χαμηλών διαστάσεων. Για το σκοπό αυτό υλοποιούνται γραμμικές και μη γραμμικές τεχνικές μείωσης διαστάσεων. Η πειραματική διαδικασία διεξάγεται πάνω σε δύο γονιδιακές βάσεις δεδομένων. Η πρώτη αφορά τον πρότυπο οργανισμό *Saccharomyces cerevisiae*. Η δεύτερη συσχετίζεται με την ασθένεια της οστεοαρθρίτιδας και περιλαμβάνει δύο τύπους ανθρώπινων κυττάρων, τα μεσεγχυματικά βλαστικά κύτταρα και τα χονδροκύτταρα, όπως αυτά προκύπτουν μετά την διαδικασία της χονδρογενούς διαφοροποίησης σε καλλιέργεια με τη μορφή τρισδιάστατων σφαιριδίων. Στα πλαίσια της επιβεβαίωσης της ερευνητικής διαδικασίας, η μεθοδολογία αναπτύσσεται και εφαρμόζεται σε μία τεχνητή βάση δεδομένων που αποτελείται από ένα σύνολο Gaussian πληθυσμών. Παράλληλα με τις μεθόδους ομαδοποίησης γονιδίων (gene grouping) και οπτικοποίησης που εφαρμόζονται και στα δύο γονιδιακά σύνολα, προτείνεται μία νέα μεθοδολογία ομαδοποίησης και ένα νέος δείκτης αξιολόγησής της, με βάση μέτρα χρονικής εξέλιξης που εφαρμόζεται στο σύνολο δεδομένων της οστεοαρθρίτιδας. Η εν λόγω προσέγγιση στοχεύει στο φιλτράρισμα γονιδίων (gene filtering), γονιδίων που απεικονίζουν διαφορετική συμπεριφορά σε δύο χρονικές στιγμές χρησιμοποιώντας το χρονικό προφίλ. Τέλος τα αποτελέσματα που προκύπτουν κατά την εφαρμογή των αλγορίθμων στο σακχαρομύκητα και στα δύο είδη κυττάρων, ερμηνεύονται βιολογικά με τη χρήση των συστημάτων ταξινόμησης MIPS FunCat και PANTHER, αντίστοιχα.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Εισαγωγή

Ως αρχικός στόχος της παρούσας εργασίας ορίζεται η οπτικοποίηση του γονιδιώματος του σακχαρομύκητα *Saccharomyces cerevisiae* έπειτα από τη διαδικασία της ομαδοποίησης. Επιθυμούμε την κατηγοριοποίηση των γονιδίων που παρουσιάζουν όμοια έκφραση σε ίδιες ομάδες, και την ευδιάκριτη απεικόνιση αυτών σε ένα χώρο χαμηλών διαστάσεων. Η επιλογή της γονιδιακής βάσης δεδομένων του εν λόγω οργανισμού δεν έγινε τυχαία. Το γονιδίωμα του σακχαρομύκητα παρουσιάζει κοινές ιδιότητες και αντιστοιχίες με εκείνο του ανθρώπου, επομένως μια καλή βιολογική ερμηνεία του ομαδοποιημένου αποτελέσματος θα μπορεί να φανεί χρήσιμη σε διάφορους τομείς της επιστήμης της βιολογίας και της ιατρικής για την κατανόηση, την πρόληψη και την αντιμετώπιση ασθενειών.

Μια ποικιλία μεθόδων οπτικοποίησης και ομαδοποίησης έχουν προταθεί κατά καιρούς με στόχο να συμβάλουν στη διαχείριση των γονιδιακών δεδομένων μεγάλης κλίμακας [1, 2, 3]. Η ομαδοποίηση γονιδίων (gene grouping) [4] στην παρούσα εργασία προκύπτει από την υλοποίηση αλγορίθμων από το χώρο των τεχνητών νευρωνικών δικτύων SOM και LVQ καθώς και την ευρέως διαδεδομένη τεχνική ομαδοποίησης k-means. Η οπτικοποίηση επιτυγχάνεται με τη χρήση δύο αλγορίθμων μείωσης διαστάσεων, του PCA που αποτελεί μία γραμμική μέθοδο και του k-PCA, μιας μη γραμμικής τεχνικής κάνοντας χρήση διαφορετικών πυρήνων. Συμπληρωματικά γίνεται αξιολόγηση των ευρημάτων της ομαδοποίησης με χρήση των δεικτών εγκυρότητας Dunn, Davies-Bouldin και Silhouette, οι τιμές των οποίων μας βοηθούν να εκτιμήσουμε αν η ποιότητα της ομαδοποίησης συνάδει με το αποτέλεσμα της οπτικοποίησης των δεδομένων.

Στη συνέχεια εξετάζουμε και την ομαδοποίηση χρονικά μεταβαλλόμενων δεδομένων (μετρήσεις) από γονιδιακούς δείκτες. Η ομαδοποίηση με βάση την ανάλυση του χρονικού προφίλ του γονιδίου που συνιστά το γονιδιακό φιλτράρισμα (gene filtering), μπορεί να προσφέρει νέες γνώσεις σε συγκεκριμένες βιολογικές

διεργασίες. Στην παρούσα διπλωματική εργασία αναπτύσσουμε μια τέτοια ολοκληρωμένη ανάλυση ομαδοποίησης που βασίζεται στα πρότυπα έκφρασης αλλά επηρεάζεται επίσης από τις χρονικές μεταβολές των γονιδίων. Η προτεινόμενη μεθοδολογία παρουσιάζεται με ένα σύνολο γονιδίων, χρονικής έκφρασης, που προέρχεται από τη διαδικασία της χονδρογενούς διαφοροποίησης σε καλλιέργεια με τη μορφή τρισδιάστατων σφαιριδίων, των πρωτογενών ανθρώπινων χονδροκυττάρων και των μεσεγχυματικών βλαστικών κυττάρων που έχουν προέλθει από ανθρώπινο μυελό των οστών. Επίσης, συγκρίνουμε τα αποτελέσματα της μεθοδολογίας που προτείνουμε με εκείνα που προκύπτουν από πρότερες μεθόδους που βασίζονται, είτε στην ομαδοποίηση με βάση την έκφραση των γονιδίων, είτε στην ομαδοποίηση των κωδικοποιημένων μεταβολών, στηριζόμενη μόνο στις χρονικές μεταβολές των δεδομένων [5, 6].

Η προτεινόμενη μεθοδολογία μας παρέχει ένα συνεκτικό εργαλείο που διευκολύνει τόσο τη στατιστική όσο και τη βιολογική επικύρωση των προφίλ χρονικής μεταβολής. Αντικειμενικό στόχο του προβλήματος που αντιμετωπίζουμε, αποτελεί ο έλεγχος για το αν τελικά τα μεσεγχυματικά βλαστικά κύτταρα κατάφεραν μετά από τις συνθήκες χονδρογενούς διαφοροποίησης που υπεβλήθηκαν να μετατραπούν σε χονδροκύτταρα.

Η στατιστική μετάφραση των ευρημάτων ακολουθείται και από βιολογική αξιολόγηση. Στην περίπτωση του σακχαρομύκητα *Saccharomyces cerevisiae*, η ερμηνεία βασίζεται στη χρήση του συστήματος ταξινόμησης MIPS FunCat, ενώ τα αποτελέσματα της οστεοαρθρίτιδας προκύπτουν από το σύστημα ταξινόμησης PANTHER.

Για τον σακχαρομύκητα *Saccharomyces cerevisiae* η ανάλυση της γονιδιακής ομαδοποίησης (gene grouping) καταλήγει σε ομάδες γονιδίων, οι οποίες παρουσιάζουν σημαντικές διαφοροποιήσεις σε επίπεδο βιολογικών διεργασιών. Από τη δεύτερη βάση δεδομένων η γονιδιακή ομαδοποίηση καταλήγει τόσο σε ομοιότητες μεταξύ των δύο τύπων κυττάρων όσο και σε σημαντικές διαφορές, αποτελέσματα που είναι εξίσου ενδιαφέροντα για τον κλάδο της ιατρικής που αφορά την οστεοαρθρίτιδα.

Εν κατακλείδι, για την επιβεβαίωση της μεθοδολογίας που χρησιμοποιήσαμε και στα πλαίσια ενός τυπικού ελέγχου εφαρμόσαμε την πειραματική διαδικασία σε μία

τεχνητή βάση δεδομένων που αποτελείται από Gaussian ανεξάρτητους πληθυσμούς, εξασφαλίζοντας με τον τρόπο αυτό την ορθή λειτουργία της μεθοδολογίας.

Στις ενότητες που ακολουθούν, παρατίθενται κάποιοι βασικοί ορισμοί από την επιστήμη της μοριακής βιολογίας που σχετίζονται με τις βάσεις δεδομένων που χρησιμοποιούνται στην παρούσα εργασία, των οποίων ακολουθεί λεπτομερής περιγραφή.

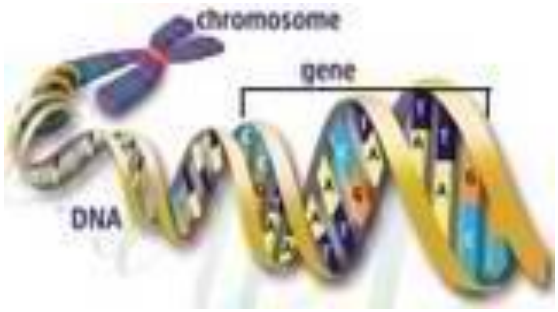
1.2 Βασικές έννοιες Μοριακής Βιολογίας

Μοριακή Βιολογία (Molecular Biology) είναι ο κλάδος της Βιολογίας που μελετά τη δομή, τη σύνθεση και τη λειτουργία της γενετικής πληροφορίας (DNA και RNA) σε μοριακό επίπεδο, καθώς και το πώς αυτή αλληλεπιδρά με πρωτεΐνες. Τα όρια μεταξύ της Μοριακής Βιολογίας με εκείνα ορισμένων κλάδων της Χημείας και άλλους κλάδους της Βιολογίας δεν είναι πάντα ξεκάθαρα, και ιδιαίτερα με εκείνα της βιοχημείας και της γενετικής. Το όνομα <<Μοριακή Βιολογία>> δόθηκε το 1952 από τον Άγγλο μοριακό βιολόγο William Astbury [7].

1.2.1 Γονίδιο

Τα *γονίδια* (Εικόνα 1) είναι τμήματα του DNA που βρίσκονται στα χρωμοσώματα. Τα γονίδια υπάρχουν σε εναλλακτικές μορφές και ονομάζονται αλληλόμορφα. Τα αλληλόμορφα γονίδια βρίσκονται στην ίδια γονιδιακή θέση στα ομόλογα χρωμοσώματα και ελέγχουν με διαφορετικό τρόπο τις ίδιες ιδιότητες που μπορούν να μεταδοθούν από τους γονείς στους απογόνους. Η διαδικασία με την οποία τα γονίδια μεταδίδονται ανακαλύφθηκε από τον Gregor Mendel και διατυπώνεται σε αυτό που είναι γνωστό ως νόμος διαχωρισμού του Mendel.

Τα γονίδια περιέχουν τους κωδικούς για την παραγωγή συγκεκριμένων πρωτεϊνών. Οι πληροφορίες που περιέχονται στο DNA δεν είναι άμεσα μετατρέψιμες σε πρωτεΐνες, αλλά πρέπει πρώτα να μεταγραφούν στο πλαίσιο μιας διαδικασίας που ονομάζεται μεταγραφή του DNA. Η εν λόγω διαδικασία λαμβάνει χώρα στο εσωτερικό του πυρήνα των κυττάρων μας. Πραγματική παραγωγή πρωτεϊνών πραγματοποιείται στο κυτταρόπλασμα των κυττάρων μας, μέσω μιας διαδικασίας που ονομάζεται μετάφραση [8].



Εικόνα 1 : Απεικόνιση ενός γονιδίου. Εικόνα: Genome Management Information System, Oak Ridge National Laboratory.

1.2.2 Μικροσυστοιχίες γονιδίων

Τα τελευταία χρόνια, οι *μικροσυστοιχίες γονιδίων* (DNA microarray, Εικόνα 2) [9] έχουν γίνει η <<state of the art>>[1] μέθοδος για την ανάλυση της γονιδιακής έκφρασης, διαδικασία κατά την οποία η κληρονομήσιμη πληροφορία ενός γονιδίου μετατρέπεται σε ένα λειτουργικό γονιδιακό προϊόν, όπως είναι μια πρωτεΐνη, και για την κατανόηση των υποκείμενων ρυθμιστικών μηχανισμών του κυττάρου. Αυτή η υψηλής απόδοση τεχνολογία επιτρέπει την ταυτόχρονη παρακολούθηση των επιπέδων έκφρασης χιλιάδων τμημάτων των γονιδίων. Ένα προφίλ έκφρασης ενός γονιδίου περιγράφει την εξέλιξη του επιπέδου της έκφρασής του σε μια χρονική σειρά πειραμάτων μικροσυστοιχίας [10].

Η πληροφορία που λαμβάνεται από τη τεχνολογία μικροσυστοιχιών DNA δίνει ένα γενικό στιγμιότυπο από την μεταγραφική κατάσταση, δηλαδή, το επίπεδο έκφρασης όλων των γονιδίων που εκφράζονται σε ένα κυτταρικό πληθυσμό σε κάθε δεδομένη στιγμή. Μια από τις δύσκολες ερωτήσεις που προέκυψαν από την τεράστια ποσότητα των δεδομένων των μικροσυστοιχιών είναι να εντοπίσει ομάδες συν-ρυθμιζόμενων γονιδίων και να κατανοήσει το ρόλο τους στη λειτουργία των κυττάρων [12], με χρήση της γονιδιακής ομαδοποίησης.

[1] State of the art καλείται το υψηλότερο επίπεδο ανάπτυξης μιας συσκευής, μιας τεχνικής ή ενός επιστημονικού πεδίου που έχει επιτευχθεί σε συγκεκριμένο χρόνο.

Εικόνα 2 : Μικροσυστοιχία DNA [11].

1.2.3 Ανάλυση και οπτικοποίηση Γονιδιακών δεδομένων

Η ανάλυση δεδομένων μικροσυστοιχιών έχει ως αντικειμενικούς στόχους: την ομαδοποίηση γονιδίων για την ανακάλυψη προτύπων ευρείας βιολογικής συμπεριφοράς και το φιλτράρισμα των γονιδίων για τον εντοπισμό συγκεκριμένων γονιδίων που παρουσιάζουν ενδιαφέρον. Συγκεκριμένα, η ομαδοποίηση γονιδίων (gene grouping) κατευθύνεται σε μεγάλο βαθμό από την ανάλυση ομάδων ενώ το φιλτράρισμα γονιδίων (gene filtering) βασίζεται κατά κύριο λόγο σε δοκιμές υποθέσεων. Η ομαδοποίηση γονιδίων είναι χρήσιμη για την κατανόηση της κοινής έκφρασης προτύπων ενώ το φιλτράρισμα γονιδίων είναι ειδικό για την αναγνώριση ασυνήθιστων προτύπων έκφρασης. Η δεύτερη προσέγγιση έχει ιδιαίτερη σημασία για την ανακάλυψη και ανάπτυξη φαρμάκων [4].

Οι DNA μικροσυστοιχίες παρέχουν μια ευρεία εικόνα της κατάστασης των κυττάρων κατά τη μέτρηση των επιπέδων έκφρασης χιλιάδων γονιδίων, ταυτόχρονα. Τεχνικές οπτικοποίησης, που επιτυγχάνονται μέσω της μείωσης των διαστάσεων των δεδομένων, μπορούν να επιτρέψουν την διερεύνηση και τον εντοπισμό των προτύπων και των σχέσεων σε ένα σύνθετο σύνολο δεδομένων με την παρουσίαση των δεδομένων σε μια γραφική μορφή στην οποία τα βασικά χαρακτηριστικά γίνονται πιο εμφανή [13].

1.3 Βάσεις δεδομένων

Η παρούσα διπλωματική εργασία βρήκε εφαρμογή σε δύο δημοσιευμένες βάσεις δεδομένων και σε μία άλλη, που αποτελεί μοντέλο ενός τεχνητού προβλήματος. Η πρώτη έχει ως οργανισμό μελέτης τον *σακχαρομύκητα (Saccharomyces cerevisiae)*, ο οποίος αποτελεί ένα από τα πιο εντατικά μελετημένα μοντέλα ευκαρυωτικών οργανισμών στον τομέα της μοριακής και κυτταρικής βιολογίας [14]. Η δεύτερη αναφέρεται σε δυο επιμέρους βάσεις εκείνων των *χονδροκυττάρων* (chondrocytes) και των *μεσεγχυματικών βλαστικών κυττάρων* (Mesenchymal Stem Cells- MSCs*)*. Τέλος, το τεχνητό πρόβλημα που χρησιμοποιήθηκε περιλαμβάνει τρεις 10-διαστάσεων Gaussian πληθυσμούς με διαφορετικές μέσες τιμές.

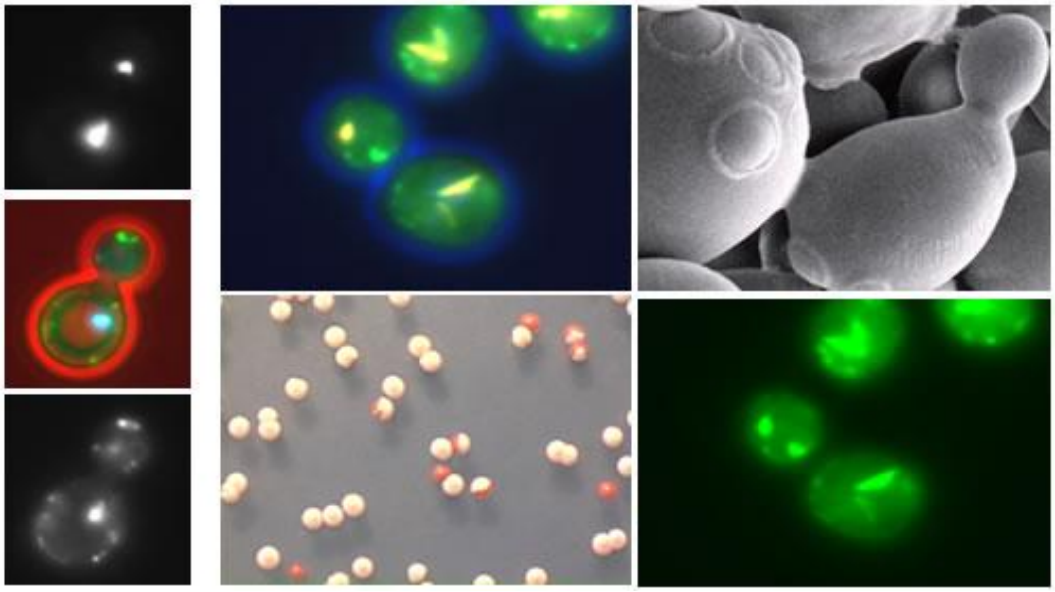
1.3.1 Σακχαρομύκητας *Saccharomyces cerevisiae*

Το μοντέλο του σακχαρομύκητα *Saccharomyces cerevisiae*

Ο σακχαρομύκητας *Saccharomyces cerevisiae* (Εικόνα 3) αποτελεί ένα είδος *ζύμης* (yeast). Οι μοναδικές ιδιότητες του σακχαρομύκητα, μεταξύ των περίπου 700 ειδών ζύμης (μια υποομάδα από 700000 μύκητες), και το τεράστιο κρυμμένο δυναμικό του που έχει γίνει αντικείμενο εκμετάλλευσης για πολλές χιλιάδες χρόνια, τον ανέδειξαν ως πρότυπο οργανισμό στην έρευνα. Επιπλέον, ο *Saccharomyces cerevisiae* και άλλες ζύμες οδήγησαν σε μια συντριπτική πλειονότητα βιομηχανικών και ιατρικών ωφέλιμων εφαρμογών για την ανθρώπινη ζωή [15].

Ορισμένα από τα θεραπευτικά προϊόντα του εν λόγω σακχαρομύκητα είναι η ινσουλίνη, η αυξητική ορμόνη και η χοριακή γοναδοτροπίνη, οι οποίες συγκαταλέγονται στις ανθρώπινες ορμόνες. Στην κατηγορία πρωτεΐνες του ανθρώπινου αίματος εμπεριέχονται η αιμοσφαιρίνη, οι παράγοντες VIII και XIII, η άλφα-1-αντιθρυψίνη, η αντιθρομβίνη III και η λευκωματίνη.

Ξεκινώντας γύρω στο 1960, ο σακχαρομύκητας εισήχθη ως πειραματικό σύστημα για τη μοριακή βιολογία. Το 1996 η ζύμη ήταν ο πρώτος ευκαρυωτικός οργανισμός, της οποίας η πλήρης αλληλουχία γονιδιώματος μπόρεσε να καθοριστεί. Τα χρόνια που ακολούθησαν, ο *Saccharomyces cerevisiae* αποτέλεσε ένα χρήσιμο σημείο αναφοράς βάσει του οποίου ακολουθίες γονιδίων των ανθρώπων, των ζώων ή των

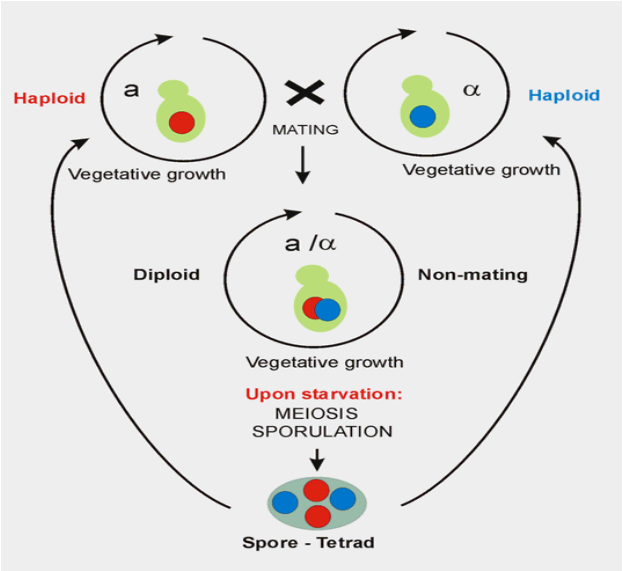


Εικόνα 3 : Ο σακχαρομύκητας *Saccharomyces cerevisiae*[16]

φυτών, και εκείνων από ένα πλήθος μονοκύτταρων οργανισμών υπό μελέτη θα μπορούσαν να συγκριθούν. Επιπλέον, η ευκολία της γενετικής μηχανικής στη ζύμη έδωσε τη δυνατότητα να αναλυθούν, λειτουργικά, γονιδιακά προϊόντα από άλλα ευκαρυωτικά κύτταρα στο σύστημα της ζύμης [15].

<u>Κύκλος ζωής του σακχαρομύκητα</u>

Είναι καλά τεκμηριωμένο ότι η ζύμη αποτελεί ένα ιδανικό σύστημα στο οποίο η αρχιτεκτονική των κυττάρων και οι βασικοί κυτταρικοί μηχανισμοί μπορούν να διερευνηθούν με επιτυχία. Μεταξύ όλων των μοντέλων ευκαρυωτικών οργανισμών, ο σακχαρομύκητας *Saccharomyces cerevisiae* συνδυάζει πολλά πλεονεκτήματα. Πρόκειται για ένα μονοκύτταρο οργανισμό (Εικόνα 4) ο οποίος, σε αντίθεση με πιο πολύπλοκα ευκαρυωτικά κύτταρα, μπορεί να καλλιεργηθεί σε καθορισμένα μέσα δίνοντας στον ερευνητή τον πλήρη έλεγχο των περιβαλλοντικών παραμέτρων. Η ζύμη είναι τιθασεύσιμη στις κλασσικές τεχνικές γενετικής, και λειτουργίες στη ζύμη έχουν μελετηθεί με μεγάλη λεπτομέρεια από βιοχημικές προσεγγίσεις. Στην πραγματικότητα μια μεγάλη ποικιλία από παραδείγματα αποδεικνύουν ότι σημαντικές κυτταρικές λειτουργίες είναι ιδιαίτερα συντηρημένες από τη ζύμη έως τα θηλαστικά και ότι αντίστοιχα γονίδια μπορούν συχνά να συμπληρώσουν το ένα το άλλο [15].



Εικόνα 4 : Ο κύκλος ζωής της ζύμης [15].

Χαρακτηριστικά της βάσης δεδομένων

Το σύνολο δεδομένων το οποίο χρησιμοποιήθηκε στην εργασία προέρχεται από μετρήσεις που έγιναν στον σακχαρομύκητα *Saccharomyces cerevisiae* [9] και αποτελείται από έναν πίνακα δεδομένων. Οι γραμμές του πίνακα αποτελούνται από τα γονίδια για τα οποία συγκεντρώθηκαν τα δεδομένα, ενώ οι στήλες αντιπροσωπεύουν ακολουθίες μεμονωμένων πειραμάτων.

Τα στοιχεία αντλήθηκαν από χρονικές μετρήσεις κατά τη διάρκεια των ακόλουθων διαδικασιών: τον κύκλο της κυτταρικής διαίρεσης [17], μετά τον συγχρονισμό με τον παράγοντα alpha (ALPH: 18 χρονικά σημεία) και το φυγοκεντρικό διαχωρισμό επίπλευσης (Centrifugal elutriation, ELU: 14 χρονικά σημεία), και με το θερμοευαίσθητο μετάλλαγμα cdc 15 (CDC 15: 15 χρονικές στιγμές), τη σπορ(ι)ογένεση [18] (Sporulation-SPO: 7 χρονικές στιγμές προσαυξημένες με 4 επιπλέον δείγματα), το σοκ από υψηλή θερμοκρασία (HT: 6 χρονικά σημεία), τα αναγωγικά μέσα (D: 4 χρονικά σημεία), τη χαμηλή θερμοκρασία (C: 4 χρονικές στιγμές) και τη μετατόπιση δύο φάσεων ανάπτυξης [19] (diauxic shift -DX: 7 χρονικά σημεία). Στην προκείμενη βάση δεδομένων περιλαμβάνονται όλα τα γονίδια των οποίων η λειτουργική περιγραφή ήταν διαθέσιμη στη βάση δεδομένων γονιδιώματος του σακχαρομύκητα [20].

Συνοψίζοντας, η βάση δεδομένων που επεξεργαζόμαστε στην παρούσα εργασία, περιλαμβάνει 2467 γονίδια εκφρασμένα σε 8 παράγοντες, σε διαφορετικές χρονικές στιγμές , έχοντας αθροιστικά 79 μετρήσεις για κάθε ένα από τα γονίδια.

Βιολογική ερμηνεία σύμφωνα με το Σύστημα Ταξινόμησης MIPS FunCat

Το βιολογικό σύνολο για το σακχαρομύκητα *Saccharomyces cerevisae* αναλύθηκε χρησιμοποιώντας το FunCat [21,22], ένα λειτουργικό σύστημα ταξινόμησης για την περιγραφή των πρωτεϊνών.

Το FunCat αποτελεί ένα ιεραρχικά δομημένο, ευέλικτο και κλιμακούμενο ελεγχόμενο σύστημα ταξινόμησης που επιτρέπει τη λειτουργική περιγραφή των πρωτεϊνών από οποιοδήποτε οργανισμό.

Το FunCat αποτελείται από 28 κύριες λειτουργικές κατηγορίες που καλύπτουν γενικούς τομείς όπως η κυτταρική μεταφορά, ο μεταβολισμός και η κυτταρική επικοινωνία. Παρουσιάζει μια ιεραρχική δομή που περιλαμβάνει έξι συνολικά επίπεδα αυξανόμενης εξειδίκευσης. Η έκδοση 2 που χρησιμοποιήθηκε στην παρούσα εργασία, περιλαμβάνει 1307 λειτουργικές κατηγορίες για το γονιδίωμα του σακχαρομύκητα *Saccharomyces cerevisiae* [21,22].

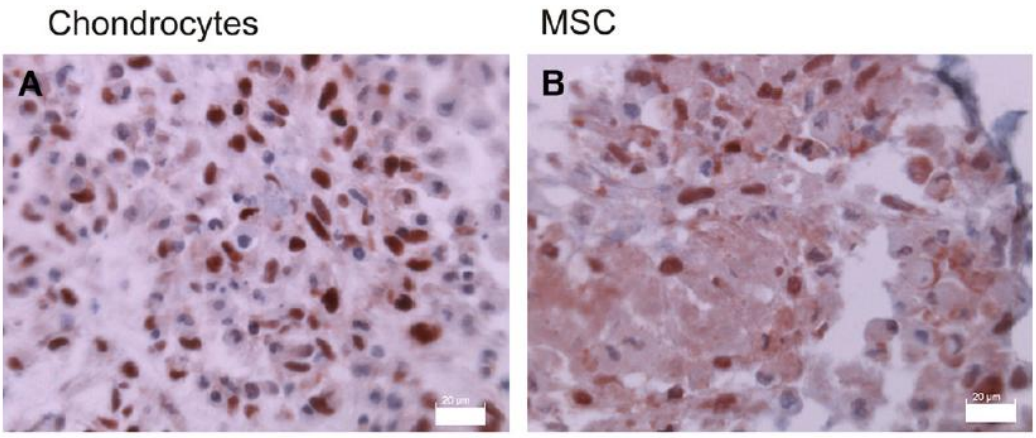
1.3.2 Χονδροκύτταρα και Μεσεγχυματικά βλαστικά κύτταρα

Ορισμός Χονδροκυττάρων και Μεσεγχυματικών βλαστικών κυττάρων

Τα *χονδροκύτταρα* (Chondrocytes -Εικόνα 5(A)) αποτελούν τα μοναδικά κύτταρα που μπορούν να εντοπιστούν στον ώριμο αρθρικό χόνδρο. Τα *μεσεγχυματικά βλαστικά κύτταρα* (Mesenchymal Stem Cell- MSCs-Εικόνα 5(B)) είναι πολυδύναμα στρωματικά κύτταρα που μπορούν να διαφοροποιηθούν σε μια ποικιλία κυτταρικών τύπων (οστεοβλάστες, χονδροκύτταρα, λιποκύτταρα) [23,24].

Ορισμός Οστεοαρθρίτιδας

Η *οστεοαρθρίτιδα* (ΟΑ) είναι ο πιο κοινός τύπος αρθρίτιδας στον κόσμο. Στον δυτικό κόσμο αποτελεί μία από τις πιο συχνές αιτίες πόνου που καταλήγει σε λειτουργική δυσλειτουργία και αρκετά συχνά σε αναπηρία στους ενήλικες [25]. Ο αρθρικός χόνδρος παρέχει μια ζωτική λειτουργία στην ομοιόσταση του περιβάλλοντος της άρθρωσης. Διαθέτει μοναδικές μηχανικές ιδιότητες, επιτρέποντας τη διατήρησή



Εικόνα 5 : Χονδροκύτταρα και Μεσεγχυματικά βλαστικά κύτταρα [27].

σχεδόν χωρίς τριβές κίνησης στη διάρκεια μιας ζωής. Ωστόσο, ο χόνδρος είναι ευάλωτος σε τραυματικές κακώσεις και λόγω του κακού αγγειώματός του και της αδυναμίας πρόσβασης σε αυτόν, θεωρείται ότι τα μεσεγχυματικά βλαστικά κύτταρα μπορούν να διευκολύνουν μια ικανοποιητική απάντηση σε θεραπεία. Ανεξιχνίαστα χόνδρινα ελαττώματα είναι συνεπώς πιθανό να προδιαθέτουν τους ασθενείς για τη ανάπτυξη της οστεοαρθρίτιδας. Η ανασύσταση και επανόρθωση του αρθρικού χόνδρου εξαρτάται από τη σύνθεση νέων ή την εμφύτευση στοιχείων της μεσοκυττάριας ουσίας των χόνδρων [26].

Ενώ ορισμένα από τα θέματα μπορούν να αντιμετωπιστούν χειρουργικά και φαρμακολογικά, μια σημαντική πρόκληση παραμένει η αναγέννηση του καταστραμμένου χόνδρου από στρατηγικές μηχανικής οστών. Αν και τα χονδροκύτταρα προέρχονται από δότες με οστεοαρθρίτιδα μπορούν να χρησιμοποιηθούν για την υποβοήθηση της αναγέννησης του χόνδρου. Ωστόσο η περιορισμένη διαθεσιμότητά τους, ποσότητα και βιωσιμότητα συνιστούν σημαντικά μειονεκτήματα. Η χρήση εναλλακτικών τύπων κυττάρων με χονδρογενή δυναμική διαφοροποίηση μπορεί να προσφέρει μια λύση. Τα μεσεγχυματικά βλαστικά κύτταρα έχουν δείξει να είναι σε θέση να διαφοροποιηθούν σε χόνδρο. Παράλληλα, πρόσφατες μελέτες δίνουν αντικρουόμενα αποτελέσματα σχετικά με τη μηχανική ποιότητα των μεσεγχυματικών βλαστικών κυττάρων που προέρχονται από δομές χόνδρου, γεγονός που υποδηλώνει κατώτερες ιδιότητες από τα χονδρογενή διαφοροποιημένα μεσεγχυματικά βλαστικά κύτταρα. Τα μεσεγχυματικά βλαστικά κύτταρα παρουσιάζουν μόνο κατώτερες συμπιεστικές ιδιότητες, ενώ οι ελαστικές τους ιδιότητες δεν φαίνεται να διαφέρουν από τα χονδροκύτταρα [27].

Χαρακτηριστικά της βάσης δεδομένων

Δεύτερη εφαρμογή βρήκε η παρούσα διπλωματική σε μια βάση δεδομένων που ως κύριο αντικείμενο μελέτης της έχει την οστεοαρθρίτιδα [27].

Η χρήση των μεσεγχυματικών βλαστικών κυττάρων για την αναγέννηση του χόνδρου παρεμποδίζεται από την ελλιπή γνώση σχετικά με τις υποκείμενες βιολογικές διαφορές μεταξύ χονδρογενώς διεγερμένων χονδροκυττάρων και μεσεγχυματικών βλαστικών κυττάρων. Η επιλογή αυτής της βάσης δεδομένων έχει ως στόχο να αξιολογήσει τις διαφορές μεταξύ των πρωτογενών ανθρωπίνων χονδροκυττάρων και των μεσεγχυματικών βλαστικών κυττάρων κατά τη διάρκεια της χονδρογενούς διαφοροποίησης σε καλλιέργεια με τη μορφή τρισδιάστατων σφαιριδίων (3D-pellet culture, PC)[2].

Μέσω ανάλυσης της γονιδιακής έκφρασης με τη χρήση μικροσυστοιχιών, η παρούσα εργασία αποσκοπεί στη σύγκριση της σφαιρικής θεώρησης των χρονικών προτύπων γονιδιακής έκφρασης σε συνθήκες PC, των ανθρώπινων πρωτογενών χονδροκυττάρων έναντι των μεσεγχυματικών βλαστικών κυττάρων, που προέρχονται από το μυελό των οστών.

Στην εργασία των Bernstein και συνεργατών [27], τα χονδροκύτταρα απομονώθηκαν από δείγματα χόνδρου που λαμβάνονται κατά τη διάρκεια επέμβασης ολικής (αλλο)αρθροπλαστικής γόνατος (N=5). Τα μεσεγχυματικά βλαστικά κύτταρα που απομονώθηκαν από τον μυελό των οστών, λήφθηκαν από υγιείς δότες (N=5), τα οποία καλλιεργήθηκαν σε PC κάτω από χονδρογενείς συνθήκες. Για την ανάλυση μικροσυστοιχίας, το υλικό (ριβονουκλεϊκό οξύ, RNA) που απομονώθηκε από τα δείγματα των πέντε ασθενών κάθε ομάδας κυττάρων (χονδροκύτταρα & μεσεγχυματικά βλαστικά κύτταρα) ενοποιήθηκε σε ένα δείγμα για καθεμία από τις παραπάνω ομάδες. Με αυτόν τον τρόπο, οι γονιδιακές εκφράσεις ολόκληρου του γονιδιώματος συγκρίθηκαν μεταξύ ενός δείγματος χονδροκυττάρων και ενός δείγματος μεσεγχυματικών βλαστικών κυττάρων σε 4 χρονικά σημεία (ημέρες 0, 3, 7, 14). Πιο συγκεκριμένα, έχουμε στη διάθεσή μας τις δύο επιμέρους βάσεις δεδομένων (χονδροκύτταρα και μεσεγχυματικά βλαστικά κύτταρα), οι οποίες αποτελούνται από 17589 γονίδια (γραμμές) σε 4 διαφορετικές χρονικές στιγμές (στήλες).

[2] PC είναι μια τρισδιάστατη μέθοδος καλλιέργειας , η οποία έχει αποδειχθεί ότι αποδίδει μια σημαντική χονδρογενή επίδραση στα χονδροκύτταρα και στα μεσεγχυματικά βλαστικά κύτταρα.

Βιολογική ερμηνεία σύμφωνα με το Σύστημα Ταξινόμησης PANTHER

Το βιολογικό σύνολο των δύο κυτταρικών τύπων αναλύθηκε χρησιμοποιώντας το σύστημα ταξινόμησης PANTHER (Protein Analysis Through Evolutionary Relationships, ανάλυση πρωτεϊνών μέσω εξελικτικών σχέσεων) [28] αποτελεί μια μοναδική πηγή ταξινόμησης των γονιδίων σύμφωνα με τις λειτουργίες τους, χρησιμοποιώντας τα δημοσιευμένα επιστημονικά πειραματικά στοιχεία και τις εξελικτικές σχέσεις για να προβλέψει τη λειτουργία τους, ακόμη και απουσία άμεσων πειραματικών ευρημάτων. Οι πρωτεΐνες έχουν ταξινομηθεί από εμπειρογνώμονες βιολόγους σύμφωνα με τις: α) γονιδιακές οικογένειες και υποοικογένειες, β) τάξεις γονιδιακής οντολογίας (μοριακή λειτουργία, βιολογική διεργασία, κυτταρικά συστατικά), γ) κλάσεις πρωτεϊνών PANTHER και τα δ) μονοπάτια (συμπεριλαμβανομένων των διαγραμμάτων) [26].

1.3.3 Μείγμα Gaussian πληθυσμών

Για την καλύτερη αξιοπιστία των μεθόδων που αναπτύσσονται στην παρούσα εργασία παράγουμε έναν τεχνητό πληθυσμό δεδομένων[3] πάνω στον οποίο εφαρμόζεται όλη η μεθοδολογία. Συγκεκριμένα δημιουργούμε τρεις Gaussian πληθυσμούς δέκα διαστάσεων (10-D), ο καθένας με διαφορετικά διανύσματα μέσων τιμών, τα οποία απέχουν μεγάλη απόσταση μεταξύ τους, αλλά ίδιους πίνακες συνδιασποράς. Ο πρώτος πληθυσμός αποτελείται από 500 δείγματα, ο δεύτερος από 600 ενώ ο τρίτος από 700 στοιχεία. Τοποθετώντας τους τρεις πληθυσμούς σε ένα πίνακα, προκύπτει η τεχνητή βάση δεδομένων με μέγεθος 1800 στοιχεία, 10 διαστάσεων το κάθε ένα. Στην πραγματικότητα επιδιώκουμε τη δημιουργία τριών Gaussian πληθυσμών, όπου ο καθένας αποτελεί ένα συμπαγή σύνολο δεδομένων και ανά δύο οι ομάδες είναι επαρκώς διαχωρίσιμες (αρκετά μακριά η μία από την άλλη).

1.4 Συνεισφορά και δομή της εργασίας

Αντικειμενικός σκοπός της παρούσας εργασίας είναι η εφαρμογή και η μελέτη αλγορίθμων ομαδοποίησης και οπτικοποίησης τόσο σε διαφορετικά δεδομένα γονιδιακής ανάλυσης όσο και σε ένα τεχνητό πληθυσμό αντικειμένων. Με τη ανάπτυξη των τεχνικών κατηγοριοποίησης στα δύο διαφορετικά γονιδιακά σύνολα

[3] Τεχνητός πληθυσμός ορίζεται ένα σύνολο δεδομένων του οποίου οι ιδιότητες είναι γνωστές.

επιτυγχάνεται η γονιδιακή ομαδοποίηση. Επικεντρώνοντας το ενδιαφέρον μας στην ενίσχυση του ομαδοποιημένου αποτελέσματος επιδιώκουμε τη χρήση ποικίλων συνδυασμών των αλγορίθμων ομαδοποίησης. Για την καλύτερη αντιμετώπιση των χρονικά μεταβαλλόμενων δεδομένων γονιδιακών βάσεων, στοχεύουμε στην ανάπτυξη εργαλείων για την ομαδοποίηση και την αξιολόγηση της ποιότητάς της με βάση δύο κριτήρια, το προφίλ έκφρασης και τις κωδικοποιημένες τιμές χρονικής μεταβολής, επιδιώκοντας την προσέγγιση του γονιδιακού φιλτραρίσματος. Τα αποτελέσματα και οι συγκρίσεις όλων των διαδικασιών για όλα τα ζητήματα που αντιμετωπίζουμε, κρίνονται τόσο με όρους στατιστικούς όσο και με βιολογικούς.

Η οργάνωση των κεφαλαίων που ακολουθούν, η οποία βασίστηκε στον τρόπο ανάπτυξης της παρούσας διπλωματικής, έχει ως εξής:

Στο Κεφάλαιο 2 ορίζεται η έννοια της ομαδοποίησης και περιγράφονται όλοι οι αλγόριθμοι που χρησιμοποιούνται. Επιπρόσθετα, αναλύονται τα Τεχνητά Νευρωνικά Δίκτυα στα οποία ανήκουν οι δύο από τις μεθόδους που υλοποιούνται ενώ παράλληλα εμπεριέχονται οι δείκτες αξιολόγησης της ομαδοποίησης.

Στο Κεφάλαιο 3 περιγράφεται εκτενώς η μεθοδολογία ομαδοποίησης και η υλοποίηση δεικτών με βάση μέτρα χρονικής εξέλιξης που προτείνεται.

Το Κεφάλαιο 4 εισάγει τη έννοια της οπτικοποίησης των δεδομένων και μια ποικιλία μεθόδων μείωσης διαστάσεων παρουσιάζεται.

Το Κεφάλαιο 5 περιλαμβάνει τη μεθοδολογία που χρησιμοποιείται, η οποία είναι αποτέλεσμα του συνδυασμού των παραπάνω υπολογιστικών μεθόδων. Έπονται όλα τα αποτελέσματα στατιστικού και βιολογικού περιεχομένου για κάθε ένα από τα προβλήματα (βάσεις δεδομένων) που αντιμετωπίζουμε.

Στο Κεφάλαιο 6 παρουσιάζονται τα συμπεράσματα που προέκυψαν και γίνεται αναφορά στις μελλοντικές επεκτάσεις της εργασίας.

Τέλος, το Παράρτημα Α εμπεριέχει ένα σύνολο εικόνων συμπληρωματικών των κεντρικών αποτελεσμάτων, ενώ στο Παράρτημα Β εμφανίζεται σε πίνακες συνοδευτική πληροφορία για συγκεκριμένα κομβικά σημεία της εργασίας.

Συνεισφορά της παρούσας εργασίας αποτελεί η σύγκριση αλγορίθμων ομαδοποίησης και οπτικοποίησης (που εφαρμόζονται σε διαφορετικά προβλήματα) με βάση τη στατιστική τους συμπεριφορά καθώς και η σύγκριση των αποτελεσμάτων

των μεθόδων αλλά και των βιολογικών ευρημάτων με την υπάρχουσα βιβλιογραφία. Επιπλέον συνεισφορά καθιστά η προτεινόμενη μεθοδολογία, η οποία συσχετίζεται με βάσεις δεδομένων που περιέχουν χρονικά μεταβαλλόμενα γονίδια. Η μεθοδολογία που αναπτύσσουμε αφορά: α) την ομαδοποίηση των γονιδίων με βάση το προφίλ της έκφρασής τους, επηρεασμένη επίσης από τη χρονική συμπεριφορά, β) ένα κριτήριο ομοιότητας για την αντιστοίχιση παρόμοιων ομάδων διαφορετικών διαμερίσεων με βάση το χρόνο και γ) ένα δείκτη εγκυρότητας μιας διαμέρισης που αποτυπώνει τη χρονική συμπεριφορά των γονιδιακών προφίλ.

Κεφάλαιο 2

Τεχνικές ομαδοποίησης

2.1 Ομαδοποίηση δεδομένων

2.1.1 Εισαγωγή

Η *ομαδοποίηση* (clustering) είναι μία από τις βασικές τεχνικές ανάλυσης ενός συνόλου δεδομένων [29]. Η ομαδοποίηση δεδομένων (data clustering) αποτελεί μια τεχνική στατιστικής ανάλυσης δεδομένων και βρίσκει εφαρμογή σε πολλούς επιστημονικούς κλάδους. Ανάμεσα σε αυτούς, συγκαταλέγονται η μηχανική μάθηση, η εξόρυξη δεδομένων (data mining), η αναγνώριση προτύπων, η ανάλυση εικόνων, η στατιστική και η βιοπληροφορική. Η ομαδοποίηση έχει επίσης μια πλούσια ιστορία και σε άλλους τομείς, όπως η βιολογία, η γεωλογία, η αρχαιολογία και το marketing. Εφαρμογή βρίσκει ακόμη και σε διάφορες κοινωνικές επιστήμες καθώς και στη ψυχολογία.

Ομαδοποίηση (Εικόνα 6) είναι η ταξινόμηση ομοίων αντικειμένων σε διαφορετικές ομάδες ή αλλιώς ο καταμερισμός των δεδομένων σε υποσύνολα, *clusters* όπως είναι η γνωστή τους ονομασία στη βιβλιογραφία, έτσι ώστε τα δεδομένα να μοιράζονται κοινά χαρακτηριστικά, τα οποία συνήθως σχετίζονται με κάποια μετρική αποστάσεων.

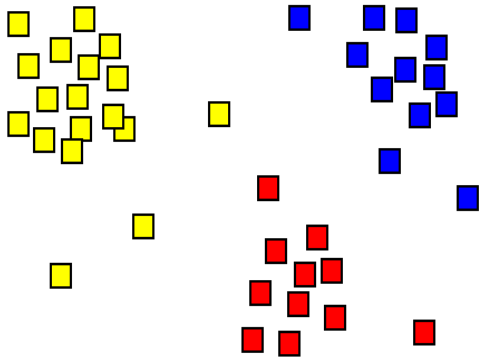
Παρ’ όλο που η βάση της ομαδοποίησης είναι αυστηρά μαθηματική [30], εντούτοις τα κίνητρα ύπαρξής της είναι καθαρά διαισθητικά και προέρχονται από την τάση του ανθρώπου να ομαδοποιεί τα στοιχεία ενός συνόλου πληροφορίας η οποία εισάγεται με οποιοδήποτε τρόπο στο νοητικό του πεδίο. Με την οργάνωση αυτήν, ο άνθρωπος καταφέρνει να διακρίνει γρήγορα τα κυριότερα χαρακτηριστικά της πληροφορίας και να αποκτήσει μια καλή αντίληψη για αυτήν. Τυπικά, η διαδικασία της ομαδοποίησης περιλαμβάνει ή συσχετίζεται με τις εξής περιοχές [31]:

1. Αναπαράσταση: Η αναπαράσταση των προτύπων αναφέρεται στον αριθμό των κλάσεων, το πλήθος των προτύπων καθώς και στο πλήθος, τον τύπο και την κλίμακα των χαρακτηριστικών των προτύπων (επιλογή/ εξαγωγή χαρακτηριστικών) που αντιπροσωπεύουν κάθε κλάση. Επιλογή χαρακτηριστικών είναι η διαδικασία εξεύρεσης του πιο αντιπροσωπευτικού

συνόλου από τα αρχικά χαρακτηριστικά. Εξαγωγή χαρακτηριστικών είναι η χρήση ενός ή περισσοτέρων μετασχηματισμών των αρχικών χαρακτηριστικών σε τρόπο ώστε να παραχθούν καινούρια σημαντικά χαρακτηριστικά. Μια καλή αναπαράσταση των προτύπων μπορεί να οδηγήσει σε «καλή» ομαδοποίηση [30].

2. Ομοιότητα: Η ομοιότητα των προτύπων υπολογίζεται συνήθως με τη βοήθεια κάποιας μετρική απόστασης (αναλυτικά Ενότητα 2.6.1) ή μετρικής ομοιότητας μεταξύ των προτύπων. Η πιο συνηθισμένη μετρική απόσταση είναι η Ευκλείδεια (Euclidean Distance), ενώ μια συνάρτηση ομοιότητας η οποία χρησιμοποιείται ευρέως είναι η Gaussian συνάρτηση ομοιότητας [30].
3. Ομαδοποίηση: Η ομαδοποίηση των δεδομένων μπορεί να γίνει με διάφορους τρόπους, οδηγώντας σε δύο κατηγορίες αποτελεσμάτων. Στην πρώτη κατηγορία τα δεδομένα ανήκουν με απόλυτο τρόπο στις διάφορες ομάδες (Crispy Methods). Στην δεύτερη κατηγορία, τα δεδομένα κατατάσσονται σε ομάδες με βαθμούς συμμετοχής (Fuzzy Methods). Με λίγα λόγια, στην κατηγορία αυτή, με τη βοήθεια της θεωρίας ασαφών συνόλων, ένα πρότυπο δύναται να ανήκει σε περισσότερες από μία ομάδες δεδομένων [30].
4. Περιγραφή: Με τον όρο περιγραφή, εννοούμε την απλή και παράλληλα συμπαγή αναπαράσταση των δεδομένων με βάση τις ομάδες που έχουν προκύψει. Η αναπαράσταση αυτή συνήθως συντελείται με τη βοήθεια κάποιων αντιπροσωπευτικών προτύπων των ομάδων. Τέτοια πρότυπα μπορεί να αποτελέσουν τα κέντρα των ομάδων [30].
5. Αξιολόγηση: Το τελευταίο τμήμα της ομαδοποίησης είναι η αποτίμηση του αποτελέσματος. Το ερώτημα που ανακύπτει είναι πώς μπορεί να εκτιμηθεί η ομαδοποίηση που έχει προκύψει και τί είναι αυτό που την χαρακτηρίζει ως κακή ή αξιόπιστη. Όλοι οι αλγόριθμοι ομαδοποίησης, αφού τους δοθούν τα δεδομένα ως είσοδος, θα παράσχουν κάποια ομαδοποίηση, ανεξάρτητα από το αν όντως υπάρχουν ομάδες στα δεδομένα. Για το λόγο αυτό, θα πρέπει πρώτα να αποτιμηθούν τα ίδια τα δεδομένα και έπειτα ο αλγόριθμος και τα αποτελέσματά του. Αυτό σημαίνει ότι θα πρέπει αρχικά να διαπιστωθεί αν αρμόζει σε ένα συγκεκριμένο σύνολο προτύπων να αποτελέσει την είσοδο ενός αλγορίθμου ομαδοποίησης (cluster tendency). Από τη στιγμή που κριθεί σκόπιμη η ομαδοποίηση των προτύπων, ακολουθεί η αποτίμηση των

αποτελεσμάτων της. Συχνά, γι' αυτόν τον σκοπό, χρησιμοποιείται κάποιο κριτήριο βελτιστοποίησης. Ωστόσο αυτά τα κριτήρια, συνήθως προέρχονται από ευριστικές (heuristics) προσεγγίσεις και αγγίζουν στα πλαίσια του υποκειμενικού [30].



Εικόνα 6 : Μια ομαδοποίηση δεδομένων σε τρεις ομάδες (clusters).

Πηγή : www.wikipedia.org

2.1.2 Ομαδοποίηση δεδομένων γονιδιακής έκφρασης

Η ομαδοποίηση χρησιμοποιείται ευρέως στην ανάλυση δεδομένων γονιδιακής έκφρασης και αποτελεί σύνηθες εργαλείο όταν δεν υπάρχει *a priori* γνώση για τα δεδομένα. Ομαδοποίηση μπορεί να υλοποιηθεί και στις δύο διαστάσεις του πίνακα γονιδιακής έκφρασης, δηλαδή και στα γονίδια και στα πειράματα. Υπό το πρίσμα της μηχανικής μάθησης μια τέτοια ανάλυση θεωρείται μη επιβλεπόμενη (unsupervised) ενώ υπό το πρίσμα της εξόρυξης δεδομένων θεωρείται μέθοδος ερευνητικής ανάλυσης δεδομένων (exploratory data analysis).

Ένα συχνό δίλλημα που προκύπτει στην ομαδοποίηση είναι ποιος αλγόριθμος είναι ο πιο κατάλληλος για τα διαθέσιμα δεδομένα. Το σίγουρο είναι ότι δεν υπάρχουν <<καλοί>> και <<κακοί>> αλγόριθμοι, αλλά η ποιότητα του καθενός μπορεί να κριθεί από τα αποτελέσματα που δίνει σε μια συγκεκριμένη εφαρμογή.

Με την ομαδοποίηση αυτό που αποσαφηνίζεται σε πρώτο στάδιο είναι η ρύθμιση (ή η συν-ρύθμιση) μεμονωμένων γονιδίων, αλλά το θεμιτό είναι η αναζήτηση της ενιαίας εικόνας του δικτύου των ρυθμιστικών αλληλεπιδράσεων.

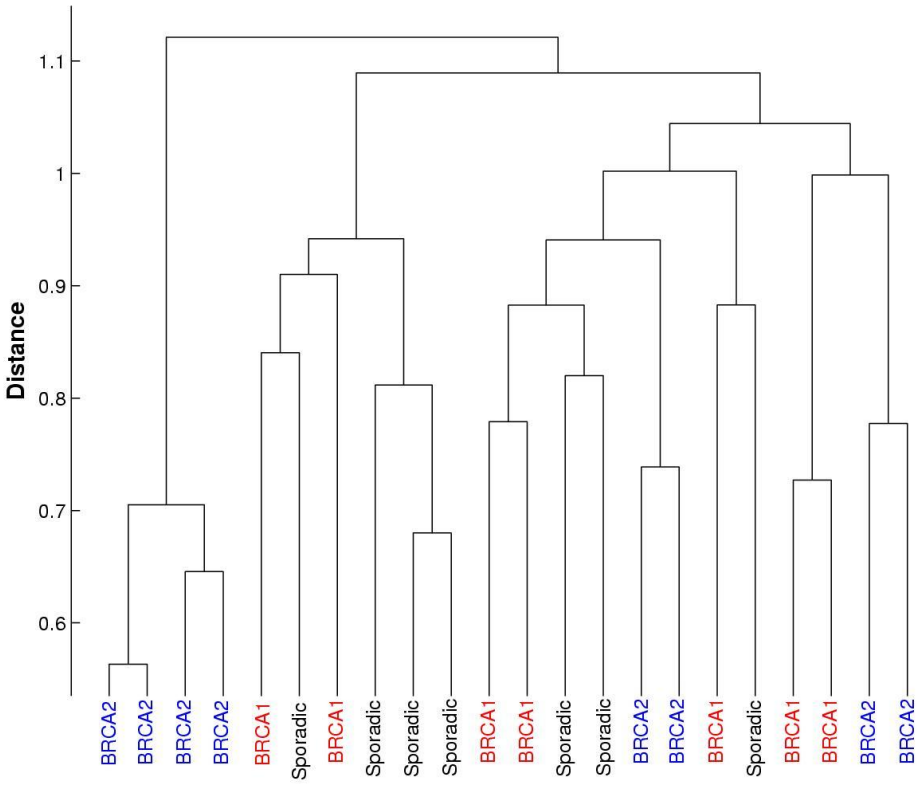
Γονίδια στην ίδια ομάδα έκφρασης έχουν την τάση να μοιράζονται κοινές λειτουργίες αν και υπάρχουν περιπτώσεις όπου λειτουργικά συνδεδεμένα γονίδια έχουν ανόμοια ή και αντίστροφα προφίλ έκφρασης [32]. Ο Tavazoie [33] χρησιμοποίησε τον αλγόριθμο k-means για να χωρίσει τα γονίδια του σακχαρομύκητα σε τριάντα διακριτές ομάδες, παρατηρώντας ότι τα μέλη κάθε ομάδας ήταν εμπλουτισμένα με κοινές λειτουργίες. Οι λειτουργίες των άγνωστων γονιδίων μπορούσαν να πιθανολογηθούν από γονίδια με γνωστή λειτουργία μέσα στην ίδια ομάδα. Επίσης τα επίπεδα παραγωγής mRNA μπορούν να θεωρηθούν μοριακή <<υπογραφή>> του φαινοτύπου του κυττάρου. Έτσι ομαδοποιώντας τα προφίλ έκφρασης είναι δυνατό να διαφοροποιηθούν οι διάφοροι κυτταρικοί τύποι, βοηθώντας στην αναγνώριση των διάφορων υποκατηγοριών καρκίνου.

2.1.3 Αλγόριθμοι ομαδοποίησης

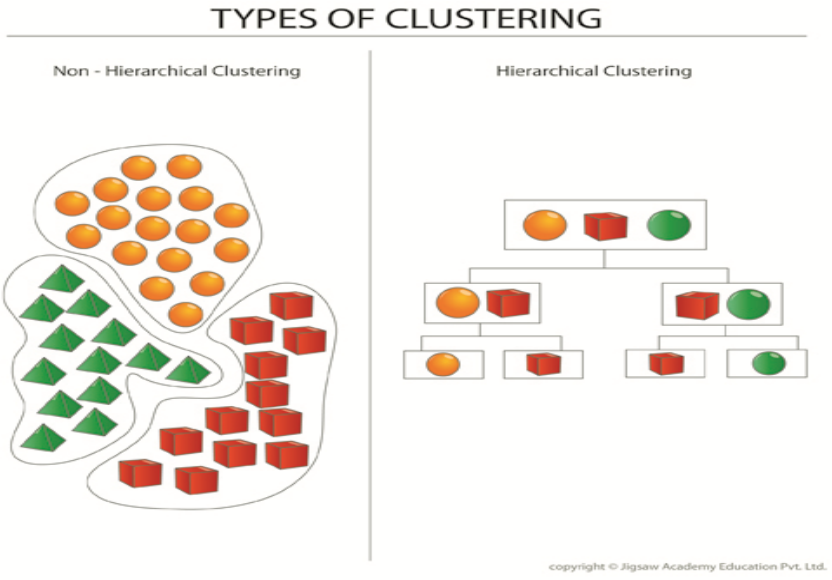
Οι αλγόριθμοι ομαδοποίησης μπορεί να είναι ιεραρχικοί (hierarchical) ή /διαιρετικοί (partitional) [31]. Οι ιεραρχικοί αλγόριθμοι [34, 35] βρίσκουν διαδοχικές ομάδες, χρησιμοποιώντας κάθε φορά ήδη καθιερωμένες ομάδες, ενώ οι μη ιεραρχικοί καθορίζουν τις ομάδες αμέσως. Η πρώτη κατηγορία αλγορίθμων χωρίζονται στους συσσωρευτικούς (agglomerative) και στους διαχωριστικούς (divisive). Οι πρώτοι αντιμετωπίζουν κάθε στοιχείο σαν μια ομάδα από μόνο του και στη συνέχεια συγχωνεύονται σε μεγαλύτερες ομάδες. Οι δεύτεροι ξεκινούν με ολόκληρο το σύνολο και το διασπούν σε μικρότερες ομάδες.

Πιο συγκεκριμένα οι ιεραρχικοί αλγόριθμοι παράγουν μια σειρά διαμερίσεων σε μορφή δενδρογράμματος (Εικόνα 7). Αναζητούν δομές οι οποίες μπορούν με τη σειρά τους να διαιρεθούν περαιτέρω σε υποδομές.

Οι διαιρετικοί αλγόριθμοι παράγουν ακριβώς μία διαμέριση που βελτιστοποιεί κάποιο κριτήριο (clustering criterion). Αυτό έχει ως αποτέλεσμα τη δημιουργία διαχωριστικών υπέρ-επιφανειών στο χώρο των δεδομένων. Μια εικονική διαφοροποίηση μεταξύ των δύο κεντρικών κατηγοριών ομαδοποίησης παρουσιάζεται στην Εικόνα 8.



Εικόνα 7: Ιεραρχική ομαδοποίηση της BRCA data[36] με χρήση όλων των γονιδίων.



Εικόνα 8 : Τύποι ομαδοποίησης[4].

Ο διαχωρισμός ανάμεσα στις διάφορες ομάδες μπορεί να επιτυγχάνεται με γραμμικό τρόπο [30]. Ωστόσο, υπάρχουν πολλές περιπτώσεις στις οποίες οι

[4] Jigsaw Academy Education Pvt Ltd.

διαχωριστικές επιφάνειες είναι μη γραμμικές. Έχουν προταθεί διάφορες μέθοδοι για την αντιμετώπιση αυτού του προβλήματος. Οι μέθοδοι αυτές μπορούν να χωριστούν σε δύο βασικές κατηγορίες [29].

1. Μέθοδοι πυρήνα
2. Φασματικές με χρήση της θεωρίας των γράφων

Όσον αφορά τις μεθόδους με χρήση πυρήνα , προέρχονται από την τροποποίηση πολλών γνωστών γραμμικών αλγορίθμων με την προσθήκη πυρήνων. Η χρήση των πυρήνων, ουσιαστικά επιτρέπει την απεικόνιση των δεδομένων σε ένα χώρο υψηλής διάστασης στον οποίο εφαρμόζεται ο αντίστοιχος γραμμικός αλγόριθμος με την προσδοκία, τα δεδομένα στο χώρο να διαχωρίζονται γραμμικά σε ομάδες [30].

Από την άλλη, οι φασματικές τεχνικές, προέρχονται από τη φασματική θεωρία των γράφων. Η βασική ιδέα είναι να δημιουργηθεί ο βεβαρυμμένος γράφος των αρχικών δεδομένων με τρόπο ώστε η κάθε κορυφή να παριστά κάποιο δείγμα και κάθε ακμή του γράφου να αντιπροσωπεύει την ομοιότητα μεταξύ των δειγμάτων που ενώνει [30].

Δεδομένου ότι η επιλογή ενός συγκεκριμένου πυρήνα ή ενός συγκεκριμένου μέτρου ομοιότητας είναι κρίσιμης σημασίας, πολλές μέθοδοι έχουν προταθεί ώστε να μαθαίνουν αυτόματα τις παραμέτρους αυτές από τα ίδια τα δεδομένα. Στην πράξη, προκύπτουν κάποια ερωτήματα, τα οποία πρέπει να απαντηθούν ώστε να προχωρήσει κάποιος τη διαδικασία ομαδοποίησης. Μερικά από αυτά είναι τα εξής [30]:

1. Με ποιον τρόπο πρέπει να κανονικοποιηθούν τα δεδομένα προτού περάσουν ως είσοδος στον αλγόριθμο της ομαδοποίησης.
2. Ποια μετρική είναι η καλύτερη για μια συγκεκριμένη περίπτωση.
3. Πως μπορεί να χρησιμοποιηθεί επικουρικά η οποιαδήποτε *a priori* γνώση του συνόλου των προτύπων στη διαδικασία της ομαδοποίησης.
4. Πως μπορούν να αντιμετωπιστούν πρακτικά προβλήματα, όπως για παράδειγμα η ομαδοποίηση ενός μεγάλου συνόλου δεδομένων αποτελεσματικά.

Από τα παραπάνω θα πρέπει να καταστεί σαφές ότι δεν υπάρχει κανένας παγκόσμιος αλγόριθμος ο οποίος να αντιμετωπίζει όλα τα προβλήματα

ομαδοποίησης. Αντιθέτως το κάθε ιδιαίτερο πρόβλημα θα πρέπει να αντιμετωπίζεται ad hoc, λαμβάνοντας υπ' όψιν τις εκάστοτε ιδιαίτερες συνθήκες που το περιβάλλουν.

Το γεγονός ότι η ομοιότητα των προτύπων είναι θεμελιώδης για τον ορισμό της ομαδοποίησης, καθιστά σημαντική της επιλογή της μετρικής που θα χρησιμοποιηθεί. Τα αντικείμενα για τα οποία γίνεται λόγος, μπορεί να είναι οτιδήποτε. Στα πλαίσια όμως της μηχανικής μάθησης και της αναγνώρισης προτύπων τα αντικείμενα περιγράφονται με τη βοήθεια ενός συνόλου πραγματικών αριθμών. Για παράδειγμα, ακόμα και στην περίπτωση που τα χαρακτηριστικά ενός προτύπου είναι κατηγορικά, τότε μπορούν αυτά εύκολα να μετασχηματιστούν σε πραγματικούς αριθμούς. Παρακάτω ως αντικείμενα θα θεωρούμε μαθηματικά διανύσματα τα οποία ανήκουν στον πραγματικό χώρο $\mathbb{R}^m$ μιας ορισμένης διάστασης *m*. Τα διανύσματα αυτά θα τα καλούμε σημεία [30].

Με αυστηρούς μαθηματικούς όρους, το πρόβλημα της ομαδοποίησης ορίζεται ως ένα πρόβλημα μάθησης χωρίς επίβλεψη κατά το οποίο δίνεται ένα σύνολο *n* σημείων $X=\{x_1, x_2, \ldots, x_n\}$ και ζητείται να ομαδοποιηθούν αυτά σε *K* ομάδες, $\{X_1, X_2, \ldots, X_K\}$. Η ομαδοποίηση αυτή πρέπει να γίνει με τρόπο ώστε οι ομάδες να είναι ξένες ανά δύο μεταξύ τους, δηλαδή $\boldsymbol{X_i} \cap \boldsymbol{X_j} = \emptyset \ \boldsymbol{\alpha\nu}\ \boldsymbol{i} \neq \boldsymbol{j}$ και η ένωσή τους να συνιστά ολόκληρο το σύνολο, δηλαδή $\cup_{\boldsymbol{i=1}}^{\boldsymbol{K}} \boldsymbol{X_i} = \boldsymbol{X}$ [30].

Υπάρχουν διάφορες τεχνικές ομαδοποίησης, όπως για παράδειγμα οι K-Μέσοι, η Neural Gas, η single linkage ή τεχνικές που χρησιμοποιούν τη θεωρία των ασαφών συνόλων όπως οι Ασαφείς C-Μέσοι. Για ομαδοποίηση επίσης χρησιμοποιούνται και τεχνικές Expectation-Maximization (EM) [37]. Στις τεχνικές αυτές θεωρούμε ότι το σύνολο των σημείων περιγράφονται με τη βοήθεια ενός αθροίσματος κατανομών (mixture density). Ο στόχος είναι να βρεθούν οι συντελεστές που καθορίζουν τις κατανομές αυτές, αλλά οι μέθοδοι αυτοί πάσχουν από κάποια μειονεκτήματα [38]. Πρώτον, για να γίνει η εκτίμηση των παραμέτρων χρειάζεται να γίνουν κάποιες υποθέσεις, όπως για παράδειγμα ότι η κάθε ομάδα ακολουθεί Gaussian κατανομή (Gaussian density). Δεύτερον, κατά την επίλυση του προβλήματος της μέγιστης πιθανοφάνειας, μπορεί να προκύψουν τοπικά ακρότατα. Αυτό συνεπάγεται χρήση επαναληπτικών αλγορίθμων ώστε να υπερβληθούν τα τοπικά ακρότατα. Από παρόμοια προβλήματα πάσχει και ο αλγόριθμος K-Μέσων [38], ο οποίος υλοποιείται στην παρούσα διπλωματική και αναπτύσσεται σε επόμενη ενότητα.

Στο παρών σύγγραμμα θα ασχοληθούμε επιπρόσθετα, με μία άλλη ευρέως διαδεδομένη τεχνική ομαδοποίησης, εκείνη που βασίζεται στην χρήση Τεχνητών Νευρωνικών Δικτύων. Η τεχνική αυτή στηρίζεται στην ομαδοποίηση των δοσμένων σημείων μιας βάσης δεδομένων, που εξάγεται από την ομαδοποίηση των κόμβων ενός νευρωνικού δικτύου, του οποίου οι παράμετροι (βάρη) εκπαιδεύονται από τα δεδομένα. Τέτοιες μέθοδοι συσταδοποίησης, οι οποίες παραθέτονται στη συνέχεια του κεφαλαίου, αποτελούν τα *Δίκτυα Αυτό-οργανούμενης Απεικόνισης* (Self Organizing Map-SOM) και η *Εκπαιδευόμενη Διανυσματική Κβάντιση* (Learning Vector Quantization-LVQ) .

2.2 Τεχνητά νευρωνικά δίκτυα

Η έρευνα σχετικά με τα Τεχνητά Νευρωνικά Δίκτυα [39] είναι εμπνευσμένη από τη δομή και τη λειτουργία του εγκεφάλου. Βασικό δομικό στοιχείο του εγκεφάλου είναι οι νευρώνες , δηλαδή τα νευρικά κύτταρα τα οποία δημιουργούν ένα πυκνό δίκτυο επικοινωνίας μεταξύ τους. Κίνητρο για τη μελέτη του νευρώνα και των νευρικών δικτύων είναι η ελπίδα ανακάλυψης ενός νέου υπολογιστικού μοντέλου βασισμένου σε μία δικτυακή δομή παρόμοια με αυτή του εγκεφάλου. Αυτή η καινούρια υπολογιστική πλατφόρμα , γνωστή ως Connectionist Model, θα είναι πιο κατάλληλη για ανάπτυξη ευφυών συστημάτων και γενικότερα διαδικασιών σχετιζόμενων με νοημοσύνη [39].

Τα συνήθη Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ) χρησιμοποιούν πολύ απλοποιημένα μοντέλα νευρώνων τέτοια ώστε να διατηρούν μόνο τα πολύ αδρά χαρακτηριστικά των λεπτομερών μοντέλων που χρησιμοποιούνται στη νευρολογία. Θα έλεγε κανείς ότι τα συνήθη τεχνητά νευρωνικά μοντέλα έχουν ελάχιστη σχέση με τα βιολογικά νευρωνικά συστήματα. Ωστόσο πιστεύεται ότι οι λεπτομέρειες δεν έχουν ιδιαίτερη σημασία στην κατανόηση της ευφυούς συμπεριφοράς των βιολογικών νευρωνικών συστημάτων. Ακόμη και αυτά τα απλά μοντέλα νευρώνων μπορούν να δημιουργήσουν ιδιαιτέρως ενδιαφέροντα δίκτυα αρκεί να πληρούν δύο βασικά χαρακτηριστικά [39]:

- Οι νευρώνες να έχουν ρυθμιζόμενες παραμέτρους ώστε να διευκολύνεται η διαδικασία της μάθησης.

- Το δίκτυο να αποτελείται από μεγάλο πλήθος νευρώνων ώστε να επιτυγχάνεται ο παραλληλισμός της επεξεργασίας και κατανομή της πληροφορίας.

Η πρόκληση που αντιμετωπίζει η θεωρία των Τεχνητών Νευρωνικών δικτύων είναι η εύρεση κατάλληλων αλγορίθμων εκπαίδευσης των δικτύων και ανάκλησης της πληροφορίας που αυτά περιέχουν έτσι ώστε να παρουσιάζονται ευφυείς διαδικασίες όπως αυτές που αναφέρθηκαν παραπάνω. Για την επίτευξη αυτού του στόχου απαιτείται ορισμός του κατάλληλου περιβάλλοντος εκπαίδευσης [39].

Οι νευρωνικοί αλγόριθμοι ταξινομούνται ως εξής :

- Εκπαιδευόμενα βάρη
    - Με επίβλεψη
        - ✓ Perceptron
        - ✓ ADALINE
        - ✓ Back-propagation
        - ✓ Αναδρομικά δίκτυα Back-propagation
        - ✓ Δίκτυα RBF
        - ✓ Μοντέλα SVM
        - ✓ Στοχαστικές μηχανές
    - Χωρίς επίβλεψη
        - ✓ Συσχετιστικά μοντέλα(Κανόνας του Hebb)
            - Δίκτυα PCA
            - Δίκτυα ICA
        - ✓ Ανταγωνιστικά μοντέλα
            - Δίκτυο Kohonen(SOM)
            - Learning VQ
            - ART
- Σταθερά βάρη
    - Δίκτυο Hopfield
    - Συσχετισμένες μνήμες
    - Brain State in a Box

Στην παρούσα διπλωματική εργασία αναπτύσσουμε και υλοποιούμε δύο από τους παραπάνω νευρωνικούς αλγορίθμους από την κατηγορία των ανταγωνιστικών μοντέλων, το δίκτυο Kohonen (SOM) και το Learning VQ που αναλύονται ακολούθως.

2.3 Δίκτυα Αυτό-οργανούμενης απεικόνισης

2.3.1 Τοπολογική οργάνωση του εγκεφάλου

Τα περισσότερα μοντέλα μάθησης βασίζονται αποκλειστικά στην έννοια του συναπτικού βάρους για να αναπαραστήσουν την πληροφορία με τη μορφή κάποιας συνάρτησης εισόδου-εξόδου [39]. Η αλλαγή των συναπτικών βαρών μεταξύ των νευρώνων συνεπαγόταν και την τροποποίηση της συνάρτησης αυτής και συνεπώς την αλλαγή της πληροφορίας που φυλάσσεται στο δίκτυο. Η θέση του κάθε νευρώνα μέσα στην αρχιτεκτονική διάταξη του δικτύου δεν έπαιζε κάποιο ρόλο. Σε δίκτυα πολλών στρωμάτων μας ενδιέφερε απλώς ότι ο νευρώνας Α βρίσκεται στο στρώμα Χ, αλλά ακόμη και τότε, η ακριβής θέση του νευρώνα μέσα στο συγκεκριμένο στρώμα δεν είχε ιδιαίτερη σημασία.

Ωστόσο η τοπολογική πληροφορία, δηλαδή η σχετική διάταξη των νευρώνων στο δίκτυο, εμφανίζεται να παίζει σημαντικό ρόλο σε διάφορα τμήματα του εγκεφάλου (Εικόνα 9) που εκτελούν συγκεκριμένες λειτουργίες όπως την αντίληψη του ήχου, της αφής και της εικόνας.

Τα τμήματα αυτά διαθέτουν αυστηρή τοπολογική οργάνωση έτσι ώστε οι νευρώνες που διεγείρονται από συναφή ή γειτονικά εξωτερικά ερεθίσματα να βρίσκονται κοντά ο ένας με τον άλλο. Τέτοιες δομές αποκαλούνται χάρτες (Maps). Οι χάρτες με την αίσθηση στην οποία αντιστοιχούν, αποκαλούνται ακουοτοπικοί χάρτες (για την ακοή), σωματοτοπικοί χάρτες (για την αφή) και ρετινοτοπικοί χάρτες (για την όραση) [39].

2.3.2 Το δίκτυο SOM του Kohonen

Όλη αυτή η διαδικασία της τοπογραφικής πληροφορίας έχει προφανώς σημασία για τη λειτουργία του εγκεφάλου αφού η διάταξη των νευρώνων στα βαθύτερα τμήματα του εγκεφάλου απεικονίζει την τοπογραφική διάταξη των αισθητήρων. Ο πρώτος που εκμεταλλεύτηκε αυτή την ιδέα ήταν ο Teuvo Kohonen [40,41,42] ο οποίος

πρότεινε ένα νευρωνικό δίκτυο που αυτό-οργανώνεται με τρόπο που είναι εμπνευσμένος από την τοπογραφική οργάνωση του εγκεφάλου [43]. Ο χάρτης του Kohonen καλείται *Αυτό-οργανούμενος Τοπογραφικός Χάρτης* (Self Organizing Topographic Map-SOM).

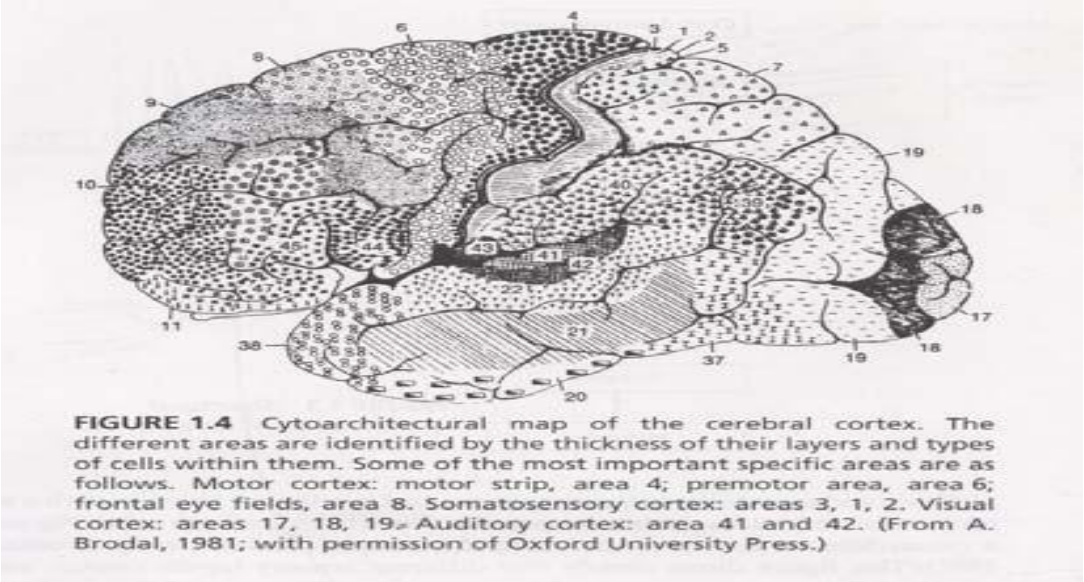


FIGURE 1.4 Cytoarchitectural map of the cerebral cortex. The different areas are identified by the thickness of their layers and types of cells within them. Some of the most important specific areas are as follows. Motor cortex: motor strip, area 4; premotor area, area 6; frontal eye fields, area 8. Somatosensory cortex: areas 3, 1, 2. Visual cortex: areas 17, 18, 19. Auditory cortex: area 41 and 42. (From A. Brodal, 1981; with permission of Oxford University Press.)

Εικόνα 9: Τοπολογική χαρτογράφηση στην επιφάνεια του εγκεφάλου [5].

Το δίκτυο SOM ανήκει στην κατηγορία των τεχνητών νευρωνικών δικτύων που εκπαιδεύονται χωρίς επίβλεψη (unsupervised learning) [44,45] και το οποίο αποτελεί μία ισχυρή μέθοδο για οπτικοποίηση (visualization) υψηλών διαστάσεων δεδομένων [46]. Δηλαδή, μας προσφέρει έναν τρόπο αναπαράστασης πολυδιάστατων δεδομένων, δεδομένων που αναπαριστώνται σε περισσότερες από μία διαστάσεις ,σε πολύ μικρότερες διαστάσεις (συνήθως 1D, 2D ή 3D). Ουσιαστικά αποτελεί έναν τρόπο συμπίεσης δεδομένων [47].

Επιπλέον, ένα τέτοιο δίκτυο μας δίνει τη δυνατότητα ομαδοποίησης δεδομένων. Τα σχέδια (patterns) ,δηλαδή τα δεδομένα εισόδου ,που είναι κοντά το ένα στο άλλο, που συνδέονται με κάποια συγκεκριμένη σχέση πρέπει να είναι στενά συνδεδεμένα στο χάρτη. Ένα SOM δίκτυο αποθηκεύει πληροφορίες με έναν τρόπο έτσι ώστε να διατηρούνται οι τοπολογικές σχέσεις ,στο σύνολο των διανυσμάτων εισόδου με τη χρήση. Συνεπώς και η δυνατότητα ομαδοποίησης.

Ένα Kohonen Network αποτελείται από ένα πλέγμα (grid) από μονάδες εξόδου και από μονάδες εισόδου. Η κάθε τιμή του διανύσματος εισόδου εισάγεται σε μία

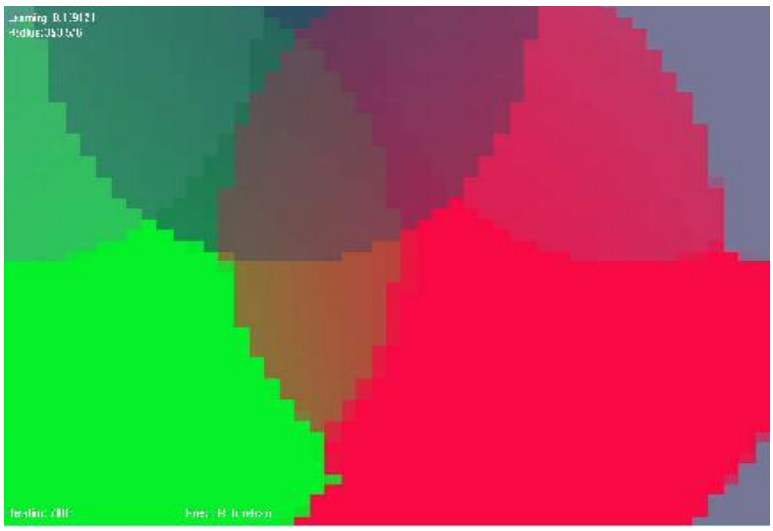
[5] Brodal A., 1981 (Oxford University Press).

μονάδα εισόδου και ακολούθως σε κάθε μονάδα εξόδου. Σε κάθε διάνυσμα εξόδου αντιστοιχεί ένα διάνυσμα ίσου μεγέθους με το διάνυσμα εισόδου που αναπαριστούν τα βάρη κάθε μονάδας εξόδου .

2.3.3 Ομαδοποίηση δεδομένων με χρήση SOM

Όπως έχει ήδη αναφερθεί, η ομαδοποίηση/ ταξινόμηση δεδομένων αποτελεί την κύρια αιτία χρήσης των SOM δικτύων (σε συνδυασμό με τη χαρτογράφηση(mapping) από n-διαστάσεις σε μικρότερες). Έστω ότι υπάρχει ένα σύνολο από n-διαστάσεων διανυσμάτων που περιγράφουν αντικείμενα. Κάθε στοιχείο του κάθε διανύσματος αναπαριστά μία ιδιότητα για το συγκεκριμένο αντικείμενο. Το κάθε διάνυσμα είναι μοναδικό όπως και τα διανύσματα που αναπαριστούν. Η ομαδοποίηση των αντικειμένων αυτών γίνεται πολύ εύκολα με τη χρήση SOM. Τα αντικείμενα αυτά θα ομαδοποιηθούν σε διακριτές περιοχές με τέτοιο τρόπο, έτσι ώστε στην κάθε περιοχή να βρεθούν αντικείμενα με όμοιες ιδιότητες.

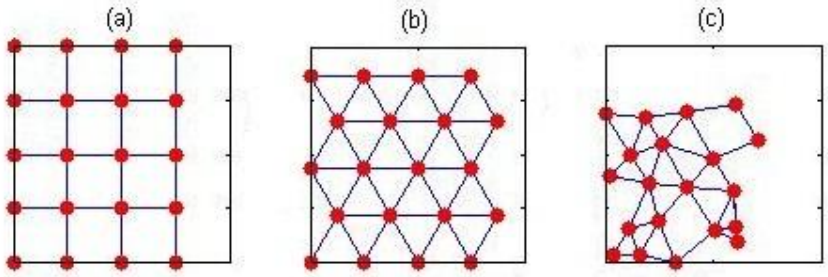
Ένα βασικό παράδειγμα χρήσης είναι η χαρτογράφηση (mapping) των χρωμάτων (που αναπαριστώνται με βάση τα τρία χρώματα που αποτελούν το κάθε χρώμα κόκκινο, πράσινο και μπλε) σε <μήκος ,πλάτος > δηλαδή από 3-D σε 2-D. Με τη χρήση του SOM ,τα χρώματα θα ομαδοποιηθούν σε διάφορες περιοχές ανάλογα με τις τιμές που δίνονται στις παραμέτρους (RGB Coloring) όπως φαίνεται και στην Εικόνα 10 [47].



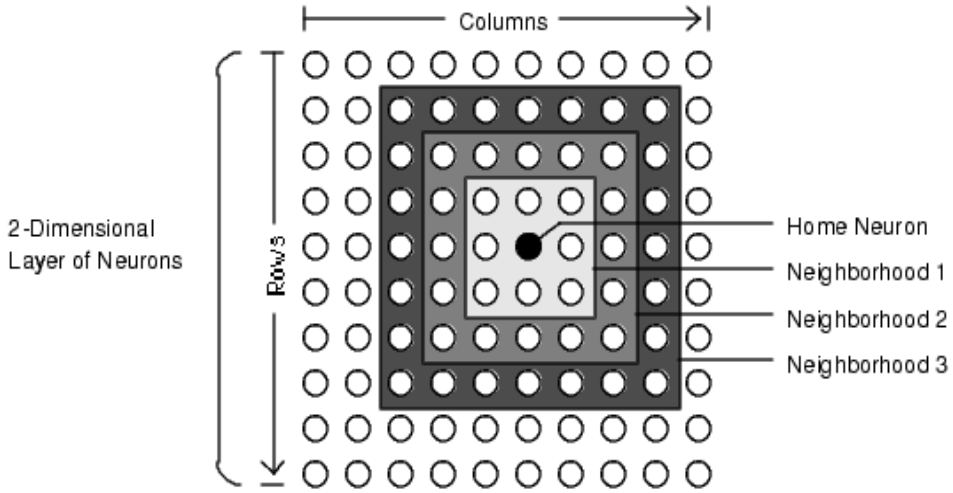
Εικόνα 10: Ένα τυχαίο παράδειγμα με ένα τυχαίο σύνολο στοιχείων σε ένα δίκτυο Kohonen [47].

2.3.4 Αρχιτεκτονική δικτύων SOM

Η διάταξη ενός Kohonen δικτύου, μπορεί να είναι είτε μία μονοδιάστατη είτε μία δισδιάστατη συστοιχία από νευρώνες. Πιο συνηθισμένη διάταξη είναι το δισδιάστατο πλέγμα (Εικόνα 11). Η θέση των νευρώνων στο μονοδιάστατο ή στο δισδιάστατο πλέγμα έχει ιδιαίτερη σημασία καθώς η τοπογραφική οργάνωση χρησιμοποιεί την έννοια της <<γειτονίας>> (Εικόνα 12) ως ένα σύνολο νευρώνων που βρίσκονται σε διπλανές θέσεις στο πλέγμα.



Εικόνα 11: Διάφορες τοπολογίες. (α)Τετραγωνικό πλέγμα, (β) εξαγωνικό πλέγμα και (γ) τυχαίο πλέγμα 20 νευρώνων (Matlab).

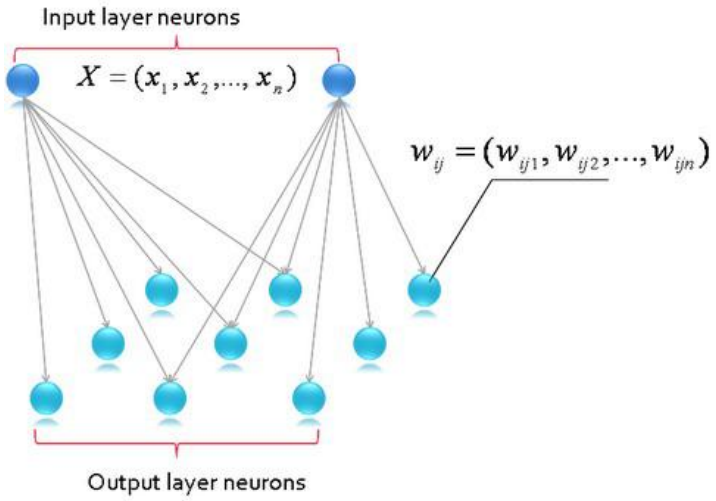


Εικόνα 12: Παραδείγματα διαφόρων γειτονιών σε ένα πλέγμα νευρώνων (Matlab).

Το δίκτυο SOM είναι ένα ανταγωνιστικό δίκτυο χωρίς επίβλεψη. Οι νευρώνες <<ανταγωνίζονται>> μεταξύ τους για το ποιος ταιριάζει καλύτερα στο διάνυσμα εισόδου. Ο νευρώνας που ταιριάζει καλύτερα καλείται νευρώνας *νικητής* για το πρότυπο αυτό. Σκοπός ενός δικτύου SOM είναι η ομαδοποίηση <<ομοίων>> προτύπων εισόδου. Με άλλα λόγια επιθυμία του χρήστη ενός τέτοιου δικτύου είναι

να δημιουργηθούν διαφορετικές <<γειτονιές νευρώνων>> πάνω στον χάρτη τέτοιες ώστε να αντιστοιχούν σε διαφορετικές ομάδες (clusters) εισόδων.

Έχουμε ήδη αναφέρει ότι ένα δίκτυο Kohonen αποτελείται από ένα πλέγμα με μονάδες εισόδου και εξόδου. Συγκεκριμένα υπάρχουν δύο επίπεδα νευρώνων (Εικόνα 13). Το πρώτο επίπεδο αποτελεί το επίπεδο εισόδου και ουσιαστικά αποτελείται από n- νευρώνες –μονάδες εισόδου. Ουσιαστικά ο αριθμός των νευρώνων αυτών καθορίζει το μέγεθος της διάστασης των διανυσμάτων εισόδου (input vectors). Κάθε νευρώνας –μονάδα εισόδου συνδέεται με όλες τις μονάδες εξόδου στο δεύτερο επίπεδο [48,49].



Εικόνα 13: Δομή Kohonen δικτύου [49].

Το δεύτερο επίπεδο είναι συνήθως συνδεδεμένο σαν δισδιάστατο πλέγμα. Υπάρχουν και πιο δύσκολες εφαρμογές που απαιτούν διαφορετικό σχεδιασμό του δεύτερου επιπέδου όμως στις πλείστες εφαρμογές λόγω απλότητας και χρησιμότητας προτιμείται αυτός ο σχεδιασμός. Κάθε νευρώνας εξόδου (ή νευρώνας δεύτερου επιπέδου), όπως έχει γραφτεί παραπάνω, είναι συνδεδεμένος με όλους τους νευρώνες εισόδου. Επιπλέον σε κάθε νευρώνα εξόδου υπάρχει ένα διάνυσμα n-μεγέθους με αριθμούς που αποτελούν τα βάρη για το συγκεκριμένο νευρώνα και έχει το ίδιο μέγεθος με τον αριθμό των νευρώνων εισόδου. Επιπρόσθετα ο κάθε νευρώνας εξόδου συσχετίζεται με άλλους παρακείμενους νευρώνες εξόδου καθορίζοντας την τοπολογία/δομή δικτύου. Η συσχέτιση αυτή καθορίζει γειτνίαση των νευρώνων αυτών και καθορίζεται με βάση μια συνάρτηση που ονομάζεται *topological neighborhood*.

Η τοπολογική ιδιότητα, δηλαδή η διατήρηση των σχέσεων μεταξύ των δεδομένων διανυσμάτων εισόδου διατηρείται με την έννοια της <<γειτονιάς>> (neighborhood), όπως είναι γνωστό, κατά τη διαδικασία εκμάθησης.[49] Ουσιαστικά η <<γειτονιά>> είναι ένα σύνολο νευρώνων εξόδου που σχετίζονται με βάση μια συνάρτηση (topological neighborhood). Συνήθως η συνάρτηση αυτή καθορίζεται με βάση την απόσταση στο τοπολογικό χάρτη δηλαδή στην τοπολογία του δικτύου μας (γραμμική απόσταση).[43] Το μέγεθος της γειτονιάς μειώνεται ανάλογα με τον αριθμό των επαναλήψεων της διαδικασίας εκμάθησης.

2.3.5 Διαδικασία εκμάθησης ενός SOM δικτύου

Τα δίκτυα Kohonen ανήκουν στην κατηγορία των νευρωνικών δικτύων και συγκεκριμένα στην κατηγορία των ανταγωνιστικών δικτύων χωρίς επίβλεψη (competitive networks without supervision). Δηλαδή δεν υπάρχει διάνυσμα στόχου. Αντίθετα η εκμάθηση (learning) του δικτύου γίνεται με διαφορετικό τρόπο. Αρχικά έχουμε μία τυχαία κατανομή των βαρών κάθε κόμβου-νευρώνα εξόδου. Ακολούθως η μονάδα εξόδου, της οποίας τα βάρη ταιριάζουν με το διάνυσμα εισόδου, και η περιοχή του πλέγματος, δηλαδή οι γειτονικοί κόμβοι εξόδου υπολογίζονται με βάση τη συνάρτηση topological neighborhood και αποτελούν τις γειτνιάσεις της μονάδας αυτής, επιλέγονται έτσι ώστε να μοιάσουν με τα στοιχεία της κατηγορίας στην οποία ανήκει το διάνυσμα εισόδου.

Πιο συγκεκριμένα η διαδικασία εκμάθησης που ακολουθείται είναι η εξής:

- Τα βάρη κάθε νευρώνα εξόδου αρχικοποιούνται τυχαία.
- Επιλέγεται τυχαία ένα διάνυσμα εισόδου από το σύνολο διανυσμάτων που θεωρούνται τα δεδομένα που θέλουμε να κατηγοριοποιηθούν.
- Συγκρίνονται τα βάρη κάθε κόμβου με το δεδομένο διάνυσμα εισόδου. Ο κόμβος εξόδου του οποίου τα βάρη <<μοιάζουν>> περισσότερο από κάθε άλλου κόμβου με το δεδομένο εισόδου, επιλέγεται (best matching unit BMU). Ο τρόπος επιλογής γίνεται με βάση μια συνάρτηση, συνήθως γίνεται χρήση της Ευκλείδειας απόστασης.
- Ακολούθως υπολογίζονται οι γειτονικοί κόμβοι ανάλογα με το μέγεθος της ακτίνας. Υπολογίζονται με βάση μια τιμή που ονομάζεται ακτίνα της γειτονιάς

του επιλεγόμενου κόμβου. Η τιμή αυτή μειώνεται σε κάθε επανάληψη και όσοι κόμβοι βρίσκονται μέσα στην τιμή αυτή ανήκουν στην περιοχή αυτή.

- Τα βάρη των κόμβων στην περιοχή αυτή αλλάζουν έτσι ώστε να μοιάζουν στο διάνυσμα εισόδου. Όσο πιο κοντά είναι κάποιος κόμβος στο κόμβο BMU, τόσο μεγαλύτερη αλλαγή υφίστανται τα βάρη του [47].

Η διαδικασία επαναλαμβάνεται (από το δεύτερο βήμα) για ένα αριθμό επαναλήψεων που λέγονται εποχές (epochs) που καθορίζονται από τον χρήστη του δικτύου, σε συνδυασμό με ένα επιπρόσθετο κριτήριο. Πιο συγκεκριμένα υπάρχει επανάληψη έως ότου το τετράγωνο της απολύτου διαφοράς των βαρών γίνει μικρότερο του 0.02, πάνω των 2500 εποχών [50]. Όπως γίνεται αντιληπτό, λόγω της συχνής χρήσης τυχαιοποίησης (randomization), όσο μεγαλύτερος είναι ο αριθμός των εποχών τόσο καλύτερη κατηγοριοποίηση (classification) θα γίνει από το δίκτυο. Όμως είναι σημαντικό να παρατηρηθεί ότι αν ξεπεραστεί ένας αρκετά μεγάλος αριθμός εποχών τότε υπάρχει περίπτωση να πλησιάζει το δίκτυο στο σημείο του over learning δηλαδή στο σημείο όπου το δίκτυο πλέον θα μπορεί να κατηγοριοποιεί μόνο πολύ συγκεκριμένα δεδομένα εισόδου. Συνηθίζεται, από λόγους εμπειρίας, ο αριθμός των εποχών να είναι 500 φορές τον αριθμό των κόμβων του δικτύου SOM. Σκοπός των δικτύων Kohonen είναι η γενική κατηγοριοποίηση δεδομένων και όχι συγκεκριμένων δεδομένων.

Για να γίνει καλύτερα κατανοητός ο αλγόριθμος θα δοθεί βαρύτητα στον τρόπο επιλογής του BMU, τον τρόπο που υπολογίζεται η ακτίνα BMU με την βοήθεια της συνάρτησης topological neighborhood, όπως επίσης και το πώς αλλάζουν τα βάρη στους κόμβους που βρίσκονται μέσα στην ακτίνα του BMU σε κάθε επανάληψη.

Η επιλογή του BMU μπορεί να γίνει με τη βοήθεια πολλών συναρτήσεων. Συνήθως όμως γίνεται η χρήση της Ευκλείδειας απόστασης (Euclidean distance). Δηλαδή για κάθε μονάδα εισόδου **x** υπολογίζεται η Ευκλείδεια απόσταση μεταξύ του διανύσματος εισόδου **x** και κάθε νευρώνα **j** του δικτύου αναπαριστώμενο από το διάνυσμα βάρους (weight vector) $\mathbf{w}_j$. Ο κόμβος με την μικρότερη απόσταση θεωρείται ο BMU. Η ευκλείδεια απόσταση δίνεται από το μαθηματικό τύπο:

$$\sqrt{\sum_{i=1}^{n}(x_i - w_i)^2} \qquad (1)$$

Όπου,

n μέγεθος του διανύσματος εισόδου και βαρών του κόμβου εξόδου που ελέγχεται,

x_i το i-οστό στοιχείο εισόδου,

w_i το i-οστό στοιχείο του διανύσματος των βαρών του κόμβου εξόδου που ελέγχεται.

Για τον υπολογισμό της ακτίνας του BMU και πιο συγκεκριμένα των κόμβων – νευρώνων εξόδου που ανήκουν στη γειτονιά (neighborhood) του BMU χρησιμοποιείται συνήθως η ακόλουθη συνάρτηση :

$$r(t) = r_o \times \exp(-(\frac{t}{\lambda})) \quad (2)$$

με t=0,1,2,…

Όπου,

t ο αριθμός της εποχής που βρίσκεται το δίκτυό μας,

r_0 είναι η αρχική τιμή της ακτίνας (συνήθως έχει το μέγεθος του πλέγματος),

r(t) είναι η τιμή της ακτίνας κατά την εποχή t,

λ μία σταθερά που εξαρτάται από την ακτίνα (αριθμός εποχών που επιλέχθηκε/ log(r0)) [47].

Με βάση την τιμή της συνάρτησης αυτής, αυτό που μένει είναι ο έλεγχος για το ποιοι νευρώνες εξόδου ανήκουν σε αυτή την ακτίνα. Είναι σημαντικό να παρατηρήσουμε ότι κι εδώ μετριέται η ευκλείδεια απόσταση στην τοπολογία μας (π.χ. σε ένα δισδιάστατο πλέγμα η απόσταση ενός κόμβου από τον άλλον μπορεί να υπολογιστεί και με βάση το πυθαγόρειο θεώρημα). Έτσι έχοντας αυτήν την τιμή από τον BMU και την τιμή της ακτίνας μπορούμε εύκολα να προσδιορίσουμε το ποιοι κόμβοι ανήκουν στη <<γειτονιά >> του BMU. Επιπλέον, ένα μοναδικό χαρακτηριστικό των δικτύων Kohonen είναι ότι η ακτίνα σε κάθε εποχή (επανάληψη) μειώνεται και αυτό επιτυγχάνεται προσθέτοντας στον υπολογισμό της ακτίνας το μέρος exp(-t). Αν ο χρήστης επέλεξε αρκετό αριθμό εποχών τότε η ακτίνα θα καταλήξει με μέγεθος 1 που θα είναι μόνο ο νευρώνας BMU, κάτι που θα σημαίνει ότι θα έχουμε πετύχει πολύ καλή κατηγοριοποίηση. Επίσης πρέπει να γίνει ιδιαίτερη αναφορά στη χρήση της

σταθεράς λ. Η σταθερά λ δίνει την δυνατότητα στο χρήστη να ελέγξει το ρυθμό της πτώσης της ακτίνας. Με τον προσδιορισμό της σταθεράς αυτής (μπορεί να οριστεί απευθείας και όχι με βάση τον τύπο) ο χρήστης του δικτύου μπορεί να ελέγξει την πτώση της ακτίνας έτσι ώστε να γίνει με πιο ομαλό τρόπο. Με πολύ μεγάλη τιμή στο λ η ακτίνα μειώνεται πιο αργά , το δίκτυο αργεί να συγκλίνει , χρειάζεται μεγαλύτερος αριθμός εποχών αλλά υπάρχει μεγαλύτερη πιθανότητα εύρεσης της βέλτιστης λύσης, δηλαδή της κατηγοριοποίησης. Μικρότερη τιμή σημαίνει πιο απότομη πτώση στην ακτίνα, λιγότερος υπολογισμός, γρηγορότερη σύγκλιση αλλά πιθανώς να μην βρεθεί η βέλτιστη λύση.

Έχοντας υπολογίσει το σύνολο των κόμβων που βρίσκονται εντός της ακτίνας του BMU (συμπεριλαμβανομένου και του BMU), θα γίνουν οι αλλαγές στα βάρη με βάση την πιο κάτω συνάρτηση:

$$\vec{w}_{t+1} = \vec{w}_t + \Theta(t)L(t)(\vec{x}_t - \vec{w}_t) \quad (3)$$

με t=0,1,2,…

Όπου,

t ο αριθμός των εποχών που βρισκόμαστε,

w_t το παλιό διάνυσμα βαρών,

w_{t+1} το νέο διάνυσμα βαρών,

x_t το δεδομένο διάνυσμα εισόδου,

L(t) που αποτελεί το learning rate,

Θ(t) που αποτελεί το πόσο η απόσταση από το BMU επηρεάζει τη μάθηση του[47].

Το learning rate(ρυθμός μάθησης) L(t) είναι συνήθως μία μικρή τιμή μεταξύ 0.1 - 0.9 που μειώνεται όσο αυξάνεται ο αριθμός των επαναλήψεων. Ουσιαστικά αποτελεί το μέγεθος τροποποίησης των βαρών των διανυσμάτων <<νικητών>>. Με μικρή τιμή στο learning rate το δίκτυο αργεί να συγκλίνει και συνεπώς απαιτείται μεγαλύτερος αριθμός εποχών που κοστίζει αρκετά. Συνήθως αυτή την διαδικασία την επιθυμούμε όταν το δίκτυο κάνει τη <<γενική>> κατηγοριοποίηση του και επιθυμούμε τον πλήρη διαχωρισμό των κατηγοριών και εμφάνιση των κοινών τους χαρακτηριστικών. Γενικά θέτοντας μεγάλη τιμή στο ρυθμό μάθησης επιτρέπουμε στο δίκτυο να

προχωρήσει/συγκλίνει πιο γρήγορα, όμως οι πιθανότητες για να μη βρεθεί η βέλτιστη λύση αυξάνονται, δηλαδή το clustering που θα γίνει δεν θα είναι το ιδανικό αφού θα βρει μεν τις κατηγορίες αλλά δε θα μπορεί να τις ξεκαθαρίσει. Επιπλέον απαιτείται σε κάθε εποχή η μείωση του learning rate έτσι ώστε οι αλλαγές να γίνονται λεπτότερες, οι διακρίσεις μέσα στις περιοχές του χάρτη να γίνουν πιο ξεκάθαρες οδηγώντας σε ένα καλύτερο συντονισμό των νευρώνων και τελικά σε καλύτερες λύσεις λαμβάνοντας έτσι τα πλεονεκτήματα τόσο μιας μεγάλης τιμής στο learning rate (clustering) τόσο και μιας μικρής τιμής (fine tuning).

$$\mathrm{L}(t) = \mathrm{L}_{\mathrm{o}} \times \exp(-\frac{t}{\lambda}) \tag{4}$$

με t=0,1,2…

Όπου,

t ο αριθμός των εποχών που βρισκόμαστε,

λ μια σταθερά που εξαρτάται από την ακτίνα,

L_o η αρχική learning rate.

Η τιμή Θ(t) εξαρτάται από τη γραμμική απόσταση (απόσταση στην τοπολογία) μεταξύ του BMU και του κάθε νευρώνα. Ουσιαστικά καθορίζει το πόσο επηρεάζει η γραμμική απόσταση που έχει ένας κόμβος που ανήκει στην ακτίνα από τον BMU, δηλαδή πόσο πρέπει να επηρεαστεί η εκμάθηση και κατ' επέκταση η αλλαγή των βαρών του συγκεκριμένου κόμβου. Συνήθως πρόκειται για μία Gaussian συνάρτηση γειτονιάς.

$$\Theta(t) = \exp\left(-\frac{\mathrm{d}^2}{2 \times \mathrm{r}^2(t)}\right) \tag{5}$$

με t=0,1,2…

Όπου,

t ο αριθμός των εποχών που βρισκόμαστε,

r(t) η τιμή που υπολογίζεται στη συνάρτηση 3,

d η απόσταση του συγκεκριμένου κόμβου από τον BMU που υπολογίζεται στο βήμα 4 του αλγορίθμου. [42, 47, 49]

2.4 Δίκτυο με εκπαιδευόμενη διανυσματική κβάντιση

2.4.1 Εισαγωγή

Μέχρι τώρα έγινε αναφορά στα Δίκτυα Αυτό-οργανούμενης απεικόνισης, τα οποία αποτελούν τεχνική ομαδοποίησης χωρίς επίβλεψη (unsupervised). Συνεχίζοντας περιγράφουμε την τεχνική ομαδοποίησης LVQ , η οποία επιτυγχάνεται εκτελώντας εκπαίδευση με επίβλεψη ή με μερική επίβλεψη. Με άλλα λόγια η διαδικασία εκτελείται έχοντας επίβλεψη σε όλα τα δείγματα ή μόνο σε λίγα που επηρεάζουν την εκπαίδευση – μάθηση.

Ο αλγόριθμος *Εκπαιδευόμενης Διανυσματικής Κβάντισης (*Learning Vector Quantization-LVQ*)* [44, 45, 51] είναι ένας επιβλέπων, βασιζόμενος σε ομαδοποίηση αλγόριθμος ταξινόμησης. Ανήκει στην κατηγορία της ανταγωνιστικής μάθησης η οποία υλοποιείται με τη χρήση ενός επιπέδου από ανταγωνιστικούς νευρώνες (ανταγωνιστικό επίπεδο). Κάθε φορά που εμφανίζεται κάποιο διάνυσμα εισόδου στο ανταγωνιστικό επίπεδο έχουμε ανταγωνισμό μεταξύ των νευρώνων του ανταγωνιστικού επιπέδου για το ποιος θα βγει νικητής για το συγκεκριμένο πρότυπο με βάση κάποιο κριτήριο επίδοσης. Ο νικητής νευρώνας είναι αυτός ο οποίος συνήθως μεταβάλλει τις τιμές των βαρών του περισσότερο σε σχέση με τους υπόλοιπους νευρώνες. Με τον τρόπο αυτό επιτυγχάνεται αυτό-οργάνωση (self-organization). Το κριτήριο που χρησιμοποιείται για την εύρεση του νικητή νευρώνα είναι η απόσταση (συνήθως Ευκλείδεια) του διανύσματος εισόδου $\vec{x}^{i} = (x_1, \dots, x_d)^T$ από το διάνυσμα των βαρών $w_i = (w_{i1}, \dots, w_{id})^T$ κάθε νευρώνα i. Ο νικητής νευρώνας j είναι αυτός με την μικρότερη απόσταση $\left|\vec{x} - \overrightarrow{w_j}\right|$. Η αυτό-οργάνωση επιτυγχάνεται στη συνέχεια μετακινώντας το διάνυσμα βαρών του νικητή νευρώνα j προς το διάνυσμα εισόδου $\vec{x}^{i}$. Αυτό επαναλαμβάνεται για κάθε νέο διάνυσμα εισόδου. Στην κατηγορία της ανταγωνιστικής μάθησης ανήκει ο αλγόριθμος LVQ που αποτελεί μια προσαρμοσμένη παραλλαγή του αλγορίθμου κ-μέσων. Σύμφωνα με αυτόν τον αλγόριθμο σε κάθε βήμα οι διάφορες ομάδες ανταγωνίζονται για την απόκτηση του πρότυπου εκπαίδευσης που χρησιμοποιείται στο συγκεκριμένο βήμα. Στη συνέχεια έχουμε μετακίνηση των κέντρων των ομάδων ανάλογα με το ποια ομάδα βγαίνει νικήτρια. Ο αλγόριθμος μπορεί εύκολα να υλοποιηθεί με ένα νευρωνικό δίκτυο και

μπορεί να χρησιμοποιηθεί τόσο για ομαδοποίηση (clustering) όσο και για ταξινόμηση (classification).

2.4.2 Βαθμωτή και διανυσματική κβάντιση

Η έννοια της κβάντισης (quantization) είναι πολύ παλιά [36], τουλάχιστον από τότε που υπάρχει ο ψηφιακός υπολογιστής. Σχετίζεται με την προσεγγιστική αναπαράσταση πληροφορίας που προέρχεται από ένα συνεχές σύνολο τιμών χρησιμοποιώντας μόνο λίγες χαρακτηριστικές τιμές. Ο κβαντιστής είναι μια μηχανή που δέχεται σαν είσοδο μια τιμή X και βγάζει στην έξοδο μια προσέγγιση q(X) του X, όπου όμως η τιμή εξόδου προέρχεται από ένα πεπερασμένο σύνολο τιμών. Για παράδειγμα, αν θέλουμε να αναπαραστήσουμε μια τυχαία μεταβλητή X που παίρνει πραγματικές τιμές από 0 έως 1 χρησιμοποιώντας μόνο 10 χαρακτηριστικές τιμές ,θα μπορούσαμε να χρησιμοποιήσουμε τις τιμές

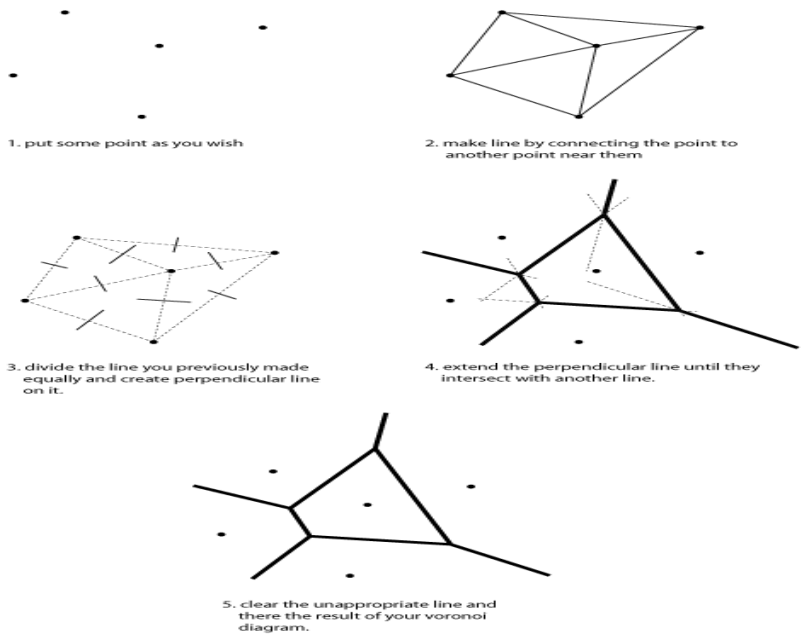
c_o=0.05, c_1=0.15, c_2=0.25, c_3=0.35, c_4=0.45, c_5=0.55, c_6=0.65, c_7=0.75,c_8=0.85, c_9=0.95.

Οι χαρακτηριστικές αυτές τιμές λέγονται κέντρα.

Με την κβάντιση, γενικά, προσεγγίζουμε την τιμή της μεταβλητής με την τιμή του κοντινότερου κέντρου. Έτσι δημιουργούνται περιοχές όπου όλες οι τιμές αντιστοιχούν στο ίδιο κέντρο. Αυτές οι περιοχές λέγονται και γειτονιές. Για παράδειγμα, η τιμή της μεταβλητής X=0.32401 προσεγγίζεται από την τιμή c_3=0.35 ,ενώ η X=0.68920 προσεγγίζεται από την τιμή c_7=0.75.

Προφανώς, η κβάντιση δημιουργεί ένα σφάλμα προσέγγισης. Ο λόγος που κάνουμε την κβάντιση είναι η κωδικοποίηση των τιμών έτσι ώστε αυτές είτε να αποθηκευτούν στη μνήμη ή στο δίσκο χρησιμοποιώντας λίγα bits, είτε να αποσταλούν σε κάποιο παραλήπτη μέσα από ένα κανάλι επικοινωνίας. Προφανώς και στις δύο περιπτώσεις η οικονομία στα bits είναι σημαντική αφού ένα μεγάλο μήνυμα απορροφά μεγάλο χώρο στο δίσκο ή μεγάλο εύρος ζώνης του καναλιού. Για την κωδικοποίηση των δέκα κέντρων αρκούν τέσσερα bits : $c_o \leftrightarrow 0000$, …, $c_9 \leftrightarrow 1001$, ενώ η κωδικοποίηση του πλήρους αριθμού με απεριόριστα δεκαδικά ψηφία θα απαιτούσε άπειρα bits. Δυστυχώς αυτή η οικονομία έρχεται με ένα κόστος, αυτό του σφάλματος προσέγγισης, το οποίο δεν υπάρχει στην τέλεια αναπαράσταση του αριθμού [36].

Η διανυσματική κβάντιση (vector quantization) είναι μια τελείως αντίστοιχη διαδικασία, εξίσου χρήσιμη με την κβάντιση βαθμωτών αριθμών. Για παράδειγμα η κβάντιση ενός διανύσματος $\mathbf{x} = [x_1, x_2]^T$ στο δισδιάστατο χώρο χρησιμοποιώντας δύο κέντρα c_1, c_2 θα δημιουργήσουμε δύο γειτονιές. Οι γειτονιές στην περίπτωση αυτή διαχωρίζονται από τη μεσοκάθετο της ευθείας που ενώνει τα c_1, c_2. Η λογική της κβάντισης είναι ίδια όπως και στη μονοδιάστατη περίπτωση: κάθε τιμή **x** αναπαρίσταται από το κοντινότερο κέντρο. Όταν έχουμε περισσότερα από δύο κέντρα, οι γειτονιές διαχωρίζονται από τεθλασμένες γραμμές που αποτελούνται από τμήματα των μεσοκαθέτων ανάμεσα στα διάφορα ζευγάρια κέντρων. Οι γειτονιές αυτές, που δημιουργούνται γενικά από N κέντρα κβάντισης σε ένα n- διάστατο χώρο καλούνται περιοχές Voronoi (Εικόνα 14) [36].



Εικόνα 14: Ορισμός περιοχών Voronoi.

2.4.3 Το δίκτυο SOM ως κβαντιστής

Η επιλογή των κέντρων κβάντισης είτε στη βαθμωτή είτε στη διανυσματική περίπτωση γίνεται με στόχο την ελαχιστοποίηση του σφάλματος προσέγγισης του αρχικού σήματος, δηλαδή του βαθμωτή μεταβλητής X ή του διανύσματος **x**. Συνήθως το πλήθος των κέντρων N είναι δεδομένο καθώς έχουμε περιορισμένο πλήθος bits για την αναπαράσταση του σήματος. Έτσι λοιπόν, το ζητούμενο είναι να βρεθεί η

καταλληλότερη θέση τους ώστε να ελαχιστοποιηθεί το μέσο σφάλμα προσέγγισης. Έστω το $\mathbf{c}_i$ είναι το κοντινότερο κέντρο στο διάνυσμα **x**, δηλαδή,

$$\|\mathbf{c}_i - \mathbf{x}\| \leq \|\mathbf{c}_j - \mathbf{x}\|, \forall\ j \neq i \tag{6}$$

τότε η αναπαράσταση του **x** γίνεται από το $\mathbf{c}_i$ και η έξοδος του κβαντιστή είναι $\mathbf{q(x)=c}_i$. Το σφάλμα προσέγγισης είναι $\|\mathbf{q}(\mathbf{x}) - \mathbf{x}\|^2$.Αν p(**x**) είναι η πυκνότητα πιθανότητας του **x** τότε η μέση τιμή του σφάλματος είναι

$$J = E\ \|\mathbf{q}(\mathbf{x}) - \mathbf{x}\|^2 = \int_x \|q(\mathbf{x}) - \mathbf{x}\|^2\, p(\mathbf{x}) d_{\mathbf{x}} \tag{7}$$

Σύμφωνα με την (8), εκεί που η πυκνότητα πιθανότητας p(**x**) είναι μικρή το σφάλμα $\|\mathbf{q}(\mathbf{x}) - \mathbf{x}\|^2$ δεν παίζει τόσο ρόλο. Αντίθετα, το σφάλμα πρέπει να είναι μικρό εκεί που η πυκνότητα πιθανότητας p(**x**) είναι μεγάλη. Με άλλα λόγια, εκεί όπου η κατανομή του **x** είναι πυκνή πρέπει να τοποθετηθούν πολλά κέντρα το ένα κοντά στο άλλο ώστε οι γειτονιές τους να είναι μικρές και να έχουμε αντίστοιχα μικρό σφάλμα προσέγγισης. Αντίθετα εκεί που η κατανομή είναι αραιή ή και ανύπαρκτη, τα κέντρα μπορούν να τοποθετηθούν αραιά ή και καθόλου[36].

Υπάρχουν πολλές διαφορετικές, μη-νευρωνικές μέθοδοι διανυσματικής κβάντισης με ποιο γνωστή τη μέθοδο LGB από τα αρχικά των ερευνητών Linde- Buzo- Gray που την πρότειναν [52, 53]. Ο αλγόριθμος αυτός είναι βέλτιστος υπό την έννοια ότι ελαχιστοποιεί το σφάλμα J. Το χαρακτηριστικό είναι ότι για να επιτύχει την βελτιστοποίηση απαιτούνται δύο συνθήκες [36]:

- Για κάθε διάνυσμα εισόδου **x** α αναπαράσταση **q(x)** επιλέγεται να είναι το κέντρο που είναι πιο κοντά στο **x**.
- Κάθε κέντρο **c**, είναι ο μέσος όρος των διανυσμάτων που βρίσκονται πιο κοντά σε αυτό.

Ο αλγόριθμος SOM με μέγεθος γειτονιάς ένα, ικανοποιεί ακριβώς τις ίδιες συνθήκες αφού πρώτον ο νευρώνας που εκπαιδεύεται είναι αυτός του οποίου διάνυσμα βαρών **w** είναι πιο κοντά στο διάνυσμα εισόδου **x** και δεύτερον ο κανόνας εκπαίδευσης (3) συγκλίνει στο μέσο όρο των διανυσμάτων x. Θα μπορούσαμε να πούμε ότι ο αλγόριθμος SOM είναι μία αναδρομική έκδοση του αλγορίθμου LBG ή αλλιώς ότι ο αλγόριθμος LBG είναι μία batch έκδοση του αλγορίθμου SOM [36].

2.4.4 Μαθαίνοντας τη βέλτιστη διανυσματική κβάντιση

Ο αλγόριθμος *Μάθησης της Διανυσματικής Κβάντισης* γνωστός *ως Learning Vector Quantization* ή απλά *LVQ* είναι μία μέθοδος εκτίμησης της βέλτιστης τοποθέτησης των κέντρων για διανυσματική κβάντιση σε n διαστάσεις. Προτάθηκε από τον Kohonen [42, 54] και έχει μαθηματική ομοιότητα με τον αλγόριθμο SOM αλλά λειτουργεί με επίβλεψη. Συγκεκριμένα, γνωρίζουμε τη κλάση i στην οποία ανήκει το κάθε (ή όχι όλα) διάνυσμα εισόδου **x**. Έχουμε M κλάσεις και N κέντρα $w_1,\ldots,w_N$, τα οποία είναι σε πλήθος περισσότερα ή το πολύ ίσα με το πλήθος των κλάσεων ($N \geq M$). Σε κάθε κλάση αντιστοιχούν κάποια συγκεκριμένα κέντρα. Συχνά το N είναι πολλαπλάσιο του M, όπως για παράδειγμα, όταν έχουμε k κέντρα για κάθε κλάση οπότε N=kM. Κάθε πρότυπο εισόδου **x** ταξινομείται με βάση την κλάση στην οποία αντιστοιχεί το κοντινότερο κέντρο w_i. Ζητάμε να βρούμε τη θέση των N κέντρων έτσι ώστε να περιγράφονται όλες οι κλάσεις με τον καλύτερο δυνατό τρόπο. Η επίδοση της μεθόδου είναι μετρήσιμη με βάση το σφάλμα ταξινόμησης, δηλαδή με το ποσοστό των περιπτώσεων που το πρότυπο ταξινομείται σε διαφορετική κλάση από αυτή στην οποία ανήκει.

Ο αλγόριθμος που παραθέτεται στον Πίνακα 1 περιγράφει την αρχική μορφή του LVQ, η οποία είναι γνωστή ως LVQ1. Το αποτέλεσμα του κανόνα LVQ1 είναι να μετακινούνται τα κέντρα προς τα κοντινά τους πρότυπα της ίδιας κλάσης και να απομακρύνονται από κοντινά τους πρότυπα άλλων κλάσεων. Το βήμα εκπαίδευσης β(n) μειώνεται σταδιακά με τις επαναλήψεις (όπως και στο SOM) ξεκινώντας από μία αρχική τιμή και μειούμενο γραμμικά μέχρι το 0. Η αρχικοποίηση των κέντρων γίνεται θέτοντάς τα ίσα με κάποια τυχαία πρότυπα από την κλάση στην οποία αντιστοιχούν.

Μια παραλλαγή του LVQ1, γνωστή ως OLVQ1, ορίζει ξεχωριστά βήματα εκπαίδευσης για κάθε κέντρο. Για το κέντρο $\mathbf{w}_i$ που είναι πιο κοντά στο πρότυπο εισόδου **x** το βήμα εκπαίδευσης $\beta_i(n)$ είναι σταθερό αλλά ρυθμίζεται αναδρομικά με τον παρακάτω κανόνα:

$$\beta_i(n) = \begin{cases} \dfrac{\beta_i(n-1)}{1+\beta_i(n-1)} \\ \dfrac{\beta_i(n-1)}{1-\beta_i(n-1)} \end{cases}$$

Το πάνω κλάσμα αναφέρεται στην περίπτωση της σωστής ταξινόμησης και το κάτω κλάσμα του παραπάνω κανόνα αφορά τη λάθος ταξινόμησης.

Σύμφωνα με τον προαναφερθέντα κανόνα, το βήμα εκπαίδευσης μειώνεται κάθε φορά που έχουμε σωστή ταξινόμηση και αυξάνεται στην αντίθετη περίπτωση. Έτσι τα κέντρα που βρίσκονται στο κέντρο της κλάσης και κοντά σε άλλα πρότυπα τείνουν να χάνουν το ρυθμό εκπαίδευσής τους (γίνονται πιο δυσκίνητα) ενώ, αντίθετα, αυτά που βρίσκονται στα περιθώρια της κλάσης και συνεπώς βρίσκονται μακριά από τα πρότυπα της κλάσης τους και είναι υπαίτια για πολλές λάθος ταξινομήσεις, τείνουν να αυξάνουν το ρυθμό εκπαίδευσης (γίνονται πιο ευκίνητα).

Πίνακας 1: LVQ1 αλγόριθμος [36].

Είσοδος:

- Τα διανυσματικά πρότυπα $\mathbf{x}^{(1)},\ldots, \mathbf{x}^{(P)}$
- Οι δείκτες label($\mathbf{x}^{(1)}$),…, label($\mathbf{x}^{(P)}$) των κλάσεων στις οποίες ανήκουν τα διανύσματα αυτά
- Οι αρχικές τιμές των κέντρων $\mathbf{w}_1,\ldots,\mathbf{w}_N$

Έξοδος:

- Οι εκπαιδευμένες τιμές των κέντρων $\mathbf{w}_1,\ldots,\mathbf{w}_N$

Μέθοδος:

n=0

Για κάθε εποχή {

Για κάθε πρότυπο p=1,…,P {

$n \leftarrow n + 1$

j=label($\mathbf{x}^{(P)}$)

$\mathbf{w}_i$ είναι το κοντινότερο κέντρο στο $\mathbf{x}^{(P)}$

Εκπαίδευσε το $\mathbf{w}_i$, και μόνο αυτό, ως εξής:

(α) φέρε το $\mathbf{w}_i$ πιο κοντά στο **x** αν έγινε σωστή ταξινόμηση,

δηλαδή το $\mathbf{w}_i$ αντιστοιχεί στην κλάση *j*:

$$\mathbf{w}_i(n+1) = \mathbf{w}_i(n) + \beta(n)[\mathbf{x} - \mathbf{w}_i(n)] \quad (8)$$

(β) αλλιώς απομάκρυνε το $\mathbf{w}_i$ από το **x** αν έγινε λάθος ταξινόμηση:

$$\mathbf{w}_i(n+1) = \mathbf{w}_i(n) - \beta(n)[\mathbf{x} - \mathbf{w}_i(n)] \quad (9)$$

}

}

Ο αλγόριθμος LVQ1 είναι η πρώτη από τις τρεις εκδόσεις που προτάθηκαν από τον Kohonen. Δεν έχει πάντα καλή επίδοση ιδίως όταν τα δεδομένα των κλάσεων δε συσσωρεύονται γύρω από ένα μόνο σημείο αλλά από περισσότερα. Για το λόγο αυτό ο Kohonen πρότεινε δύο πιο επιτηδευμένες και πιο πολύπλοκες εκδόσεις: LVQ2 και LVQ3 [54]. Σύμφωνα με την έκδοση LVQ2, γίνεται η παρακάτω διόρθωση στα δύο πιο κοντινά κέντρα $\mathbf{w}_i$, $\mathbf{w}_j$,και μόνο σε αυτά:

$$\mathbf{w}_i(n+1) = \mathbf{w}_i(n) - \beta(n)[\mathbf{x} - \mathbf{w}_i(n)] \tag{10}$$

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \beta(n)[\mathbf{x} - \mathbf{w}_j(n)] \tag{11}$$

υπό τις εξής προϋποθέσεις:

(α) τα κέντρα $\mathbf{w}_i$, $\mathbf{w}_j$, αντιστοιχούν σε διαφορετικές κλάσεις X_i ,X_j ,

(β) το πρότυπο **x** ανήκει στη κλάση X_j αλλά το κοντινότερο κέντρο είναι το $\mathbf{w}_i$ της κλάσης X_i ,

(γ) το πρότυπο **x** ανήκει σε μία περιοχή πλάτους *w* στο κέντρο της απόστασης μεταξύ $\mathbf{w}_i$ και $\mathbf{w}_j$.

Τα αποτελέσματα όλων αυτών των προϋποθέσεων είναι να γίνεται διόρθωση αρκετά σπάνια και πάντα σε ζευγάρια κέντρων. Ο κανόνας τείνει να επιδιορθώνει τα κέντρα απομακρύνοντάς τα από τη διαχωριστική επιφάνεια μεταξύ των κλάσεων και γι' αυτό συνιστάται να μη χρησιμοποιείται για πολλές επαναλήψεις (για παράδειγμα πλήθος εποχών ≤ 100 φορές το πλήθος των κέντρων).

Μια επιπλέον βελτίωση της μεθόδου γίνεται με τον κανόνα LVQ3. Σύμφωνα με αυτόν, αν τα δύο κοντινότερα κέντρα $\mathbf{w}_i$, $\mathbf{w}_j$, αντιστοιχούν σε διαφορετικές κλάσεις εφαρμόζουμε τη διόρθωση που προτείνει ο LVQ2, δηλαδή τις Εξισώσεις (10), (11). Αν όμως τα $\mathbf{w}_i$, $\mathbf{w}_j$, αντιστοιχούν στην ίδια κλάση στην οποία ανήκει και το **x**, τότε διορθώνονται σύμφωνα με τον τύπο:

$$\mathbf{w}_i(n+1) = \mathbf{w}_i(n) + \varepsilon\beta(n)[\mathbf{x} - \mathbf{w}_i(n)] \tag{12}$$

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \varepsilon\beta(n)[\mathbf{x} - \mathbf{w}_j(n)] \tag{13}$$

όπου ε είναι μια μικρή σταθερά με $0.1 < \varepsilon < 0.5$. Στην περίπτωση αυτή δε χρησιμοποιείται το παράθυρο πλάτους *w*. Αυτό χρησιμοποιείται μόνο στην περίπτωση των διαφορετικών κλάσεων [36].

2.5 Αλγόριθμος K-means ομαδοποίησης

Ο αλγόριθμος k-means (Κ-μέσων) είναι ένας αλγόριθμος (MacQueen, 1967) [55, 56] που ομαδοποιεί αντικείμενα βάσει των χαρακτηριστικών των k-μεριδίων. Συνιστά έναν από τους απλούστερους αλγορίθμους εκμάθησης χωρίς επίβλεψη που επιλύει το γνωστό πρόβλημα της ομαδοποίησης.

Η διαδικασία ακολουθεί έναν απλό κι εύκολο τρόπο ταξινόμησης ενός συνόλου δεδομένων σε ένα συγκεκριμένο αριθμό ομάδων(k clusters), ο οποίος ορίζεται εκ των προτέρων [57]. Βασικός στόχος είναι ο ορισμός k κέντρων, ένα για κάθε μία από τις k ομάδες. Τα κέντρα αυτά πρέπει να τοποθετηθούν με ένα έξυπνο τρόπο διότι μια διαφορετική θέση προκαλεί διαφορετικό αποτέλεσμα. Έτσι, η καλύτερη επιλογή είναι η τοποθέτηση κάθε κέντρου , όσο το δυνατόν πιο μακριά από το άλλο. Επόμενο βήμα είναι να ληφθεί ένα σημείο που ανήκει στο δοσμένο σύνολο δεδομένων και να τοποθετηθεί στο πλησιέστερο κέντρο. Όταν δεν εκκρεμούν άλλα σημεία, το πρώτο βήμα έχει ολοκληρωθεί και μία αρχική ομαδοποίηση έχει προκύψει.

Στο σημείο αυτό, απαιτείται ο υπολογισμός εκ νέου k νέων κέντρων ως τα κέντρα βάρους των ομάδων που προέκυψαν από το προηγούμενο βήμα. Έπειτα έχουμε τα προαναφερθέντα νέα κέντρα, και νέες επανατοποθετήσεις λαμβάνουν χώρα έπειτα από συγκρίσεις που γίνονται μεταξύ των ίδιων αρχικών δεδομένων με τα νέα πλέον κοντινότερα κέντρα. Με τον τρόπο αυτό ένας βρόγχος έχει δημιουργηθεί. Ως αποτέλεσμα αυτών των επαναλήψεων παρατηρείται ότι τα k κέντρα αλλάζουν τη θέση τους βήμα προς βήμα, μέχρις ότου άλλες αλλαγές να μην πραγματοποιούνται. Με άλλα λόγια τα κέντρα των ομάδων δεν μετακινούνται πια.

Ο αλγόριθμος k-means αποτελεί μεταβλητή του αλγορίθμου πρόβλεψης/μεγιστοποίησης (expectation-maximization algorithm-EM), όπου σκοπός είναι να οριστεί ο k-means δεδομένων που προήλθαν από Gaussian κατανομές. Ο αλγόριθμος υποθέτει ότι τα χαρακτηριστικά του αντικειμένου δημιουργούν ένα χώρο διανυσμάτων και ο σκοπός του είναι να ελαχιστοποιεί τη συνολική διακύμανση της ομάδας ή τη συνάρτηση τετραγωνικού σφάλματος:

$$\mathrm{V} = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - m_i \right\|^2 \qquad (14)$$

όπου υπάρχουν k ομάδες S_i , i= 1,2,…,k και m_i είναι το κεντροειδές ή το μεσαίο σημείο από όλα τα στοιχεία του συνόλου.

Τα βασικά βήματα του αλγορίθμου είναι τα εξής :

Πίνακας 2: K-means αλγόριθμος.

1. Επιλογή του αριθμού ομάδων.
2. Τυχαία δημιουργία k ομάδων και ορισμός των κεντροειδών των ομάδων.
3. Μεταβίβαση του κάθε σημείου στο κεντροειδές της κοντινότερης ομάδας.
4. Υπολογισμός των νέων κεντροειδών των ομάδων.
5. Επανάληψη μέχρι να συγκλίνει ο αλγόριθμος σε κάποιο κριτήριο.

Ο αλγόριθμος αυτός, όπως προαναφέρθηκε ξεκινά διαχωρίζοντας τα αρχικά σημεία σε k αρχικά σύνολα είτε τυχαία είτε χρησιμοποιώντας ευρετικά δεδομένα. Στη συνέχεια υπολογίζεται το μέσο στοιχείο για κάθε ένα από τα k σύνολα, το οποίο αποτελεί και το κέντρο του εκάστοτε συνόλου, σύμφωνα με την σχέση:

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j \qquad (15)$$

όπου i=1,…,k, t ο αριθμός της επανάληψης και x_j το j στοιχείο της S_i ομάδας [58].

Ο νέος διαχωρισμός που υλοποιείται, επιτυγχάνεται τοποθετώντας κάθε ένα από τα αρχικά στοιχεία, στην ομάδα εκείνη που το κέντρο της απέχει τη μικρότερη απόσταση (Ευκλείδεια απόσταση) από αυτό και τα νέα σύνολα στοιχείων προκύπτουν ως εξής:

$$\mathrm{S_i^{(t)}} = \left\{ \mathrm{x_p} : \left\| \mathrm{x_p} - \mathrm{m_i^{(t)}} \right\| \le \left\| \mathrm{x_p} - \mathrm{m_j^{(t)}} \right\| \quad \forall 1 \le \mathrm{j} \le \mathrm{k} \right\} \qquad (16)$$

όπου κάθε $\mathrm{x_p}$ αντιστοιχεί ακριβώς σε ένα $\mathrm{S_i^{(t)}}$.

Ο αλγόριθμος αυτός παραμένει διάσημος επειδή τείνει σε κάποιο όριο πολύ γρήγορα. Όσον αφορά την απόδοση ο k-means δεν εγγυάται ότι θα αγγίξει το βέλτιστο. Η ποιότητα της τελικής λύσης εξαρτάται, σε μεγάλο βαθμό, από το αρχικό σύνολο ομάδων και ενδέχεται να είναι πολύ χαμηλότερη από το συνολικό βέλτιστο.

Επιπλέον μειονέκτημα του αλγορίθμου, αποτελεί το γεγονός ότι ο αριθμός των ομάδων απαιτείται να ορισθεί εξαρχής.

2.6 Αξιολόγηση ομαδοποίησης

Λαμβάνοντας υπόψη ένα σύνολο δεδομένων, κάθε αλγόριθμος μπορεί να παράγει μία διαμέριση ή να μην μπορεί να υπάρχει πραγματικά μια συγκεκριμένη δομή στα δεδομένα. Επιπλέον, διαφορετικές προσεγγίσεις ομαδοποίησης συνήθως οδηγούν σε διαφορετικές ομάδες δεδομένων, ακόμα και για τον ίδιο αλγόριθμο, η επιλογή μιας παραμέτρου ή η σειρά παρουσίασης των προτύπων εισόδου μπορεί να επηρεάσει τα τελικά αποτελέσματα. Ως εκ τούτου, αποτελεσματικά πρότυπα αξιολόγησης και κριτήρια είναι κρίσιμης σημασίας να παρέχονται στους χρήστες με ένα βαθμό εμπιστοσύνης για τα αποτελέσματα ομαδοποίησης. Μία διαδικασία μάθησης χωρίς επίβλεψη είναι πιο δύσκολο να εκτιμηθεί σε σχέση με μία επιβλέπουσα διεργασία.

Η διαδικασία για την εκτίμηση των αποτελεσμάτων ενός αλγόριθμου ομαδοποίησης είναι γνωστή ως αξιολόγηση/επικύρωση ομαδοποίησης (clustering evaluation/validation). Αν και μια δομή ομαδοποίησης που προκύπτει από ένα συγκεκριμένο αλγόριθμο μπορεί να αξιολογηθεί από την καθολική γνώση και την εμπειρία ειδικών, η επικύρωση ομάδων δίνει έμφαση στην εκτίμηση του αποτελέσματος ομαδοποίησης με ένα τρόπο αντικειμενικό και ποσοτικό , ο οποίος είναι συνήθως βασιζόμενος στην στατιστική. Αυτές οι αξιολογήσεις θα πρέπει να είναι αντικειμενικές και δεν θα πρέπει να έχουν προτιμήσεις σε κανέναν αλγόριθμο. Θα πρέπει να είναι σε θέση να παρέχουν ουσιαστική αντίληψη, απαντώντας σε ερωτήσεις όπως πόσες ομάδες είναι κρυμμένες στα δεδομένα , κατά πόσον οι ομάδες που λαμβάνονται έχουν νόημα από μία συγκεκριμένη άποψη ή απλά αντικείμενα των αλγορίθμων, ή γιατί επιλέγουμε έναν αλγόριθμο αντί ενός άλλου. Το πρώτο ερώτημα αφορά την τάση διασποράς των δεδομένων και θα πρέπει αρχικά να απαντηθεί πριν γίνει προσπάθεια εφαρμογής της ομαδοποίησης, χρησιμοποιώντας συγκεκριμένες στατιστικές δοκιμές. Δυστυχώς, οι δοκιμές αυτές δεν παρουσιάζουν πάντα μεγάλη χρησιμότητα και απαιτούν τη διατύπωση εξειδικευμένων υποθέσεων ελέγχου. Τα υπόλοιπα ερωτήματα αφορούν την ανάλυση της εγκυρότητας ομαδοποίησης και μπορούν να απαντηθούν μόνο μετά την εφαρμογή της μεθόδου ομαδοποίησης στα δεδομένα.

Γενικά, οι δείκτες (indices) εκτίμησης των ομάδων (clusters) που υπάρχουν στη βιβλιογραφία είναι πολλοί και χωρίζονται σε τρεις κατηγορίες [59, 60, 61]:

- *Internal indices*: Ο στόχος τους είναι να εκτιμηθούν τα αποτελέσματα του αλγορίθμου clustering χρησιμοποιώντας μόνο ποσότητες που εμπλέκουν τα δεδομένα. Τέτοιοι δείκτες βασίζουν τους υπολογισμούς τους μόνο στο αποτέλεσμα της ομαδοποίησης που προκύπτει και πρέπει να εκτιμηθεί. Συνεπώς είναι κατάλληλοι για συσταδοποίηση χωρίς επίβλεψη.
- *External indices*: Οι συγκεκριμένοι δείκτες εκτιμούν το αποτέλεσμα μέσω αναφορών από προηγούμενη εξωτερική γνώση, όπως για παράδειγμα μία προ-ορισμένη δομή που δίνει τη δυνατότητα γνώσης διαισθητικά για τη δομή που θα προκύψει από ένα σύνολο δεδομένων. Αξίζει να σημειωθεί ότι αυτοί οι δείκτες δεν είναι εφαρμόσιμοι σε αληθινά δεδομένα του πραγματικού κόσμου όπου εφαρμόζονται ομαδοποιήσεις χωρίς επίβλεψη. Γενικά το κύριο μειονέκτημα τους σε σχέση με τους internal δείκτες, είναι το υπολογιστικό τους κόστος, συμπεριλαμβανομένου ότι μετρούν σε ποιο βαθμό τα clusters που προέκυψαν ταυτίζονται με ένα συγκεκριμένο και προκαθορισμένο σχήμα.
- *Relative indices-Αξιολόγηση δεικτών*: Σκοπός τους είναι η ανάκτηση του καλύτερου clustering που μπορεί να επιτευχθεί από έναν αλγόριθμο ομαδοποίησης, υπό κάποιες προϋποθέσεις και παραμέτρους. Η κύρια μέθοδος που εφαρμόζεται σε αυτούς τους δείκτες είναι η εκτίμηση του αποτελέσματος μέσω συγκρίσεων με άλλα αποτελέσματα ομαδοποίησης του ίδιου ωστόσο αλγορίθμου αλλά με διαφορετικές τιμές παραμέτρων. Όταν εφαρμόζονται κατάλληλα οι internal, φαίνεται να δρουν σαν συσχέτισης [62]. Για παράδειγμα, ψάχνοντας για τον βέλτιστο αριθμό ομάδων, ουσιαστικά εκτιμώνται διαφορετικές εκτελέσεις του ίδιου αλγορίθμου για διαφορετικό αριθμό clusters.

Οι τρεις πιο διαδεδομένοι δείκτες της πρώτης κατηγορίας είναι ο Dunn index, ο Davies-Bouldin index και ο Silhouette index, οι οποίοι είναι και αυτοί που υλοποιούνται σε αυτήν τη διπλωματική εργασία. Οι προαναφερθέντες δείκτες είναι οι πιο διαδεδομένοι για την εκτίμηση του αριθμού των ομάδων σε ένα σύνολο δεδομένων και για την αξιολόγηση των αποτελεσμάτων ενός αλγόριθμου ομαδοποίησης σε δεδομένα ανάλυσης γονιδιακής έκφρασης [63, 64, 65, 66, 67].

2.6.1 Μέτρα ομοιότητας

Το πόσο επιτυχημένο θεωρείται το αποτέλεσμα της ομαδοποίησης δεδομένων [68, 69], εξαρτάται από τα κριτήρια που θα χρησιμοποιηθούν για τον διαχωρισμό των στοιχείων σε ομάδες. Η σωστή επιλογή των κριτηρίων αυτών είναι ένα πολύ σημαντικό ζήτημα. Το πιο σύνηθες κριτήριο το οποίο χρησιμοποιείται είναι η απόσταση μεταξύ των στοιχείων του συνόλου δεδομένων.

Τα στοιχεία που ανήκουν σε κάθε cluster παρουσιάζουν ομοιότητα μεταξύ τους. Αυτή η ιδιότητα εξάλλου, είναι αναγκαία προκειμένου να ορισθεί ένα ξεχωριστό cluster κατά τη διαδικασία της ομαδοποίησης. Έτσι για όλες τις τεχνικές είναι σημαντικό να ορίζεται ένα μέτρο ομοιότητας μεταξύ δύο στοιχείων από το χώρο δεδομένων. Δεδομένης της μεγάλης ποικιλίας στα χαρακτηριστικά των στοιχείων, η επιλογή των μέτρων ομοιότητας , θα πρέπει να χρήζει υψίστης προσοχής. Η αναπαράσταση των προτύπων συνιστά σπουδαίο ρόλο στην επιλογή του εν λόγω μέτρου [70]. Σε πολλές περιπτώσεις, με το μέτρο ομοιότητας αυτό που συνήθως μετράται δεν είναι η ομοιότητα αλλά η διαφορετικότητα ή ανομοιότητα δύο τυχαίων στοιχείων. Στη συνέχεια θα αναφερθούν μέτρα ομοιότητας, τα οποία είναι ευρέως διαδεδομένα και χρησιμοποιούνται για τη σύγκριση στοιχείων των οποίων τα χαρακτηριστικά περιγράφονται από διακριτές τιμές. Το μέτρο ομοιότητας καλείται και απόσταση και συμβολίζεται με d και ικανοποιεί τα παρακάτω για δύο στοιχεία x,y:

- d(x,y)>=0 (μη αρνητικό)
- d(y,x)=d(x,y) (συμμετρία)
- d(x,x)=0 (ένδειξη ταυτοποίησης)
- d(x,y**)=**0 εάν x=y (απολυτότητα)
- d(x,y)<= d(x,z)+ d(z,y) (τριγωνική ανισότητα)

Το πιο γνωστό μέτρο ομοιότητας που χρησιμοποιείται κατά κόρον και στην παρούσα διπλωματική είναι η *Ευκλείδεια απόσταση*(Euclidean Distance) [71] η οποία ορίζεται ως εξής:

Έστω $x=(x_1,\ldots,x_n)$ και $y=(y_1,\ldots,y_n)$ δύο διανύσματα του χώρου $\mathbb{R}^n$,τότε η Ευκλείδεια απόσταση είναι,

- $$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \quad (17)$$

Γίνεται εύκολα κατανοητό ότι στην περίπτωση αυτή αλλά και στις επόμενες που θα εισαχθούν, όσο μεγαλύτερη είναι η τιμή *d(x,y)* τόσο αυξάνει η πιθανότητα τα δύο διανύσματα x και y να τοποθετηθούν σε διαφορετικές ομάδες (clusters).

Άλλος τρόπος για να υπολογιστεί η απόσταση (ανομοιότητα) μεταξύ δύο στοιχείων είναι η *απόσταση Manhattan* (Manhattan Distance) [71]:

- $$d(x,y)=\sum_{i=1}^{n}\left|x_i-y_i\right| \quad (18)$$

Τέλος, σημαντικό μέτρο ομοιότητας αποτελεί η *απόσταση Chebychev* (Chebychev Distance) [71]:

- $$d(x,y)=\max_{i=1}^{n}\left|x_i-y_i\right| \quad (19)$$

Η Ευκλείδεια απόσταση, χρησιμοποιείται ευρέως σε περιπτώσεις λίγων διαστάσεων και έχει καλά αποτελέσματα όταν τα δεδομένα κατηγοριοποιούνται σε συμπαγή και αρκετά απομονωμένα clusters. Ένα πρόβλημα που παρουσιάζει, είναι ότι στις πολλές διαστάσεις το χαρακτηριστικό το οποίο παρουσιάζει την μεγαλύτερη διαφοροποίηση από τα άλλα κυριαρχεί και αποπροσανατολίζει το τελικό αποτέλεσμα. Εδώ πρόκειται, για αυτό που συνήθως αναφέρεται ως *κατάρα των πολλών διαστάσεων (curse of dimensionality)*.

2.6.2 Δείκτες αξιολόγησης

Η πρόβλεψη του σωστού αριθμού των ομάδων αποτελεί θεμελιώδες τροχοπέδη στα ανεξέλεγκτα προβλήματα ταξινόμησης. Πολλοί αλγόριθμοι απαιτούν τον εκ των προτέρων ορισμό του αριθμού των ομάδων. Με στόχο να υπερβληθεί το πρόβλημα αυτό, ποικιλία δεικτών αξιολόγησης ομαδοποίησης έχουν προταθεί για την εκτίμηση της ποιότητας ενός διαμερισμού. Αυτή η προσέγγιση επιβάλει την εκτέλεση ενός αλγορίθμου αρκετές φορές για τη λήψη διαφορετικών διαμερίσεων. Ο διαμερισμός που βελτιστοποιεί έναν δείκτη εγκυρότητας επιλέγεται ως η καλύτερη διαμέριση (partition). Έτσι, σκοπός μιας τεχνικής εκτίμησης ομαδοποίησης είναι να αναγνωρίσει

εκείνον το διαμερισμό των ομάδων για τον οποίο ένα μέτρο ποιότητας βελτιστοποιείται.

Τα μέτρα εκτίμησης ομάδων χρησιμοποιούνται για να συγκρίνουν διαφορετικούς διαμερισμούς που δημιουργούνται από διαφορετικούς αλγορίθμους, ή από τον ίδιο αλγόριθμο χρησιμοποιώντας διαφορετικές τιμές παραμέτρων. Επικύρωση ενός συμπλέγματος αποτελεί πολύ σημαντικό ζήτημα στην ανάλυση ομαδοποίησης, επειδή το αποτέλεσμα αυτής πρέπει να αξιολογηθεί στο περισσότερο των περιπτώσεων. Στους περισσότερους αλγορίθμους, ο αριθμός των ομάδων είναι μια παράμετρος που ο χρήστης τοποθετεί. Πληθώρα προσεγγίσεων υπάρχουν για την εύρεση του καλύτερου αριθμού ομάδων. Μια ποικιλία δεικτών εκτίμησης είναι διαθέσιμοι και έτσι είναι δυνατόν να μάθουμε : πρώτον, πόσο καλά οι αλγόριθμοι ομαδοποίησης έχουν εργασθεί και πως οι μεταβλητές παράμετροι επηρεάζουν την ομαδοποίηση και δεύτερον, η ομοιότητα μεταξύ των δεικτών.

Σε αυτήν τη μελέτη, τρεις δείκτες έχουν υλοποιηθεί ο Dunn index, ο Davies-Bouldin index και ο Silhouette index. Οι μέθοδοι αυτές έχει αποδειχθεί ότι αποτελούν αποτελεσματικές εκτιμητές αξιολόγησης ομαδοποίησης για διαφορετικούς τύπους εφαρμογών. Επιπρόσθετα, έχουν επιλεγεί για την υποστήριξη της έρευνας τεχνικών επικύρωσης ομάδων για ταξινόμηση δεδομένων γονιδιακής έκφρασης. Παρ' όλα αυτά, οι υλοποιήσεις αυτών των μεθόδων παρουσιάζουν πλεονεκτήματα αλλά συνοδεύονται και από αναπόφευκτους περιορισμούς.

2.6.3 Dunn index

Η τεχνική αυτή (Dunn,1974) [72] βασίζεται στην ιδέα προσδιορισμού συνόλων ομάδων που είναι συμπαγή και καλά διαχωρίσιμα. Για κάθε διαμερισμό των ομάδων, όπου c_i αντιπροσωπεύει την i-οστή ομάδα κάθε διαμέρισης , ο Dunn δείκτης αξιολόγησης, *D* , μπορεί να υπολογιστεί από τον παρακάτω τύπο:

$$\min_{1\le i\le n}\left\{\min_{\substack{1\le j\le n\\ i\neq j}}\left\{\frac{d(c_i,c_j)}{\max_{1\le k\le n}\{d'(c_k)\}}\right\}\right\} \qquad (20)$$

όπου $d(c_i,c_j)$, η απόσταση μεταξύ των ομάδων c_i και c_j (intercluster distance), $d'(c_k)$ η *intracluster* απόσταση της ομάδας c_k και *n* ο αριθμός των ομάδων.

Ο στόχος του δείκτη είναι να μεγιστοποιήσει τις intercluster αποστάσεις και να ελαχιστοποιήσει τις intracluster αποστάσεις. Ως εκ τούτου, ο αριθμός των ομάδων που μεγιστοποιεί τον Dunn δείκτη, λαμβάνεται ως ο βέλτιστος. Έτσι λοιπόν, καθώς $D \in [0, \infty]$, όσο πιο μεγάλες είναι οι τιμές του μέτρου τόσο καλύτερη είναι η ομαδοποίηση. Ο συγκεκριμένος δείκτης, θεωρείται εύκολος αλλά είναι ασταθής όταν υπάρχουν απομακρυσμένα στοιχεία, επειδή μετριούνται μόνο δύο αποστάσεις.

Intercluster αποστάσεις

Στην ενότητα αυτή, θα παρουσιαστούν τα εσωτερικά μέτρα που χρησιμοποιούνται για την υλοποίηση του δείκτη εγκυρότητας Dunn. Όσον αφορά τις μεταξύ-ομάδων αποστάσεις (*intercluster distances*), υπάρχουν έξι τύποι τέτοιων αποστάσεων που βρίσκουν εφαρμογή στον συγκεκριμένο δείκτη [63]:

- *Single Linkage (ενιαίος σύνδεσμος*): Είναι η κοντινότερη απόσταση μεταξύ δύο δειγμάτων, τα οποία ανήκουν σε δύο διαφορετικές ομάδες.
- *Comlete Linkage (πλήρης σύνδεσμος):* Αντιπροσωπεύει την απόσταση μεταξύ των πιο απομακρυσμένων δειγμάτων, τα οποία ανήκουν σε δύο ξένες ομάδες.
- *Average Linkage (μέσος σύνδεσμος):* Καθορίζει την μέση απόσταση μεταξύ όλων των δειγμάτων που ανήκουν σε διαφορετικές ομάδες.
- *Centroid Linkage (κέντρο βάρους συνδέσμου):* Η συγκεκριμένη χρησιμοποιείται μόνο στην Ευκλείδεια απόσταση. Αποτελεί την Ευκλείδεια απόσταση μεταξύ των κέντρων δύο ομάδων , όπως υπολογίζεται από τον αριθμητικό μέσο όρο.
- *Average of Centroids Linkage (μέσος όρος κέντρων βάρους συνδέσμου)*: Ανταποκρίνεται στην απόσταση μεταξύ του κέντρου μιας ομάδας και όλων των στοιχείων που ανήκουν σε διαφορετική ομάδα.
- *Handsoff Metrics:* Τα πρότυπα αυτά βασίζονται στην ανακάλυψη μιας μέγιστης απόστασης μεταξύ των δειγμάτων μιας ομάδας και του κοντινότερου αντικειμένου ενός άλλου συνόλου του διαμερισμού.

Intracluster αποστάσεις

Όσον αφορά, την κατηγορία των intracluster αποστάσεων, στον υπολογισμό του Dunn δείκτη εκτίμησης, υπάρχουν τρεις τύποι που χρησιμοποιούνται [63]:

- *Complete Diameter (πλήρης διάμετρος)*: Ο τύπος αυτός ορίζει την απόσταση μεταξύ των πιο απομακρυσμένων δειγμάτων που ανήκουν στην ίδια cluster.
- *Average Diameter (μέση διάμετρος)*: Αναπαριστά τη μέση απόσταση μεταξύ του συνόλου των στοιχείων που απαρτίζουν όμοια ομάδα.
- *Centroid Diameter (κεντροειδές διαμέτρου)*: Η κατηγορία αυτή αντανακλά τη μέση απόσταση μεταξύ όλων των αντικειμένων και του κέντρου μιας ομάδας.

2.6.4 Davies-Bouldin index

Ο δείκτης αυτός (Davies and Bouldin, 1979) [73], αποτελεί μια συνάρτηση της σχέσης μεταξύ του αθροίσματος της διασποράς μέσα στην ίδια την ομάδα και του διαχωρισμού μεταξύ των διαφορετικών συστάδων. Ο *DB* δείκτης , όπως διαφορετικά συμβολίζεται, ορίζεται από την επόμενη σχέση:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left\{\frac{d_n(Q_i)+d_n(Q_j)}{d(Q_i,Q_j)}\right\} \tag{21}$$

όπου *n* είναι ο αριθμός των ομάδων, το d_n αποτελεί το μέσο όρο των αποστάσεων όλων των στοιχείων μιας ομάδας από το κέντρο αυτής και το $d(Q_i , Q_j)$ είναι η απόσταση μεταξύ των κέντρων των clusters. Συνεπώς, ο όρος $\frac{d_n(Q_i)+d_n(Q_j)}{d(Q_i,Q_j)}$ θα είναι μικρός αν οι ομάδες i και j είναι συμπαγείς και τα κέντρα τους είναι μακριά το ένα από το άλλο. Άρα ένας δείκτης Davies-Bouldin θα παρουσιάζει μια μικρή τιμή, $DB \in [0,\infty]$ για μια καλή ομαδοποίηση. Είναι χρήσιμο να αναφερθεί ότι το παρών μέτρο παρουσιάζει μικρή πολυπλοκότητα.

2.6.5 Silhouette index

Η τεχνική επικύρωσης Silhouette (Rousseeuw, 1987) [74], υπολογίζει το πλάτος περιγράμματος κάθε δείγματος, το μέσο όρο του πλάτους σχεδιαγράμματος για κάθε ομάδα κι όλη τη μέση τιμή του περιγράμματος για το σύνολο των δεδομένων. Με τη προσέγγιση αυτή κάθε ομάδα αναπαριστάται από αυτό που ονομάζεται περίγραμμα (*silhouette*), το οποίο βασίζεται στην σύγκριση της συγκέντρωσής της και του διαχωρισμού. Ο μέσος όρος του πλάτους περιγράμματος εφαρμόζεται για την αξιολόγησης της εγκυρότητας της ομαδοποίησης και επίσης χρησιμοποιείται για να

ληφθεί απόφαση για το πόσο καλός είναι ο αριθμός των ομάδων που δημιουργήθηκαν.

Για τον υπολογισμό των Silhouettes τιμών *S(i)* χρησιμοποιείται ο τύπος που έπεται:

$$S(i) = \frac{(\min(d_e(i,k)) - d_a(i))}{\max(d_a(i), \min(d_e(i,k)))} \quad (22)$$

όπου d_a *(i)* είναι ο μέσος όρος ανομοιότητας του i-οστού στοιχείου προς όλα τα άλλα αντικείμενα της ίδιας ομάδας και d_e *(i)* είναι η μέση τιμή της απόσταση του i-οστού σημείου προς όλα τα αντικείμενα μιας άλλης συστάδας k.

Ο δείκτης αυτός ακολουθείται από την σχέση $-1 \leq S(i) \leq 1$. Αν η Silhouette τιμή πλησιάζει το 1, αυτό σημαίνει ότι το δείγμα είναι καλώς ομαδοποιημένο (well-clustered) και αυτό έχει ανατεθεί σε μια κατάλληλη ομάδα. Αν ο Silhouette δείκτης κινείται κοντά στο 0, αυτό συνεπάγεται την εκχώρησή του σε ένα πλησίον cluster και το δείγμα βρίσκεται εξίσου μακριά από τις δύο ομάδες. Αν η εν λόγω τιμή πλησιάζει το -1, το σημείο είναι εσφαλμένα ταξινομημένο (misclassified) και βρίσκεται απλώς κάπου μεταξύ των ομάδων. Ο συνολικός μέσος όρος του πλάτους του περιγράμματος για ολόκληρο το γράφημα είναι απλά η μέση τιμή των ποσοτήτων $S(i)$ για όλα τα αντικείμενα στο σύνολο της βάσης δεδομένων.

Η μεγαλύτερη μέση συνολική Silhouette τιμή δείχνει την καλύτερη ομαδοποίηση (πλήθος ομάδων). Ως απόρροια, ο αριθμός των συστάδων με τη μέγιστη μέση τιμή Silhouette λαμβάνεται ως ο βέλτιστος αριθμός των clusters.

Κεφάλαιο 3

Ομαδοποίηση γονιδίων με βάση μέτρα χρονικής εξέλιξης

3.1 Εισαγωγή

Το ενδιαφέρον για τη χρονική δυναμική των προφίλ γονιδιακής έκφρασης αυξάνει δραματικά με την ανάπτυξη μεθόδων που χειρίζονται μεγάλης κλίμακας γονιδιακά δεδομένα. Η γενετική εξέλιξη ενός οργανισμού ή μιας ασθένειας δίνει τη δυνατότητα της μελέτης δύσκολων βιολογικών προβλημάτων και διευκολύνει την αξιολόγηση των θεραπευτικών πρωτοκόλλων που βασίζονται στην εξέταση της δυναμικής των μοριακών μηχανισμών και στην αντίδραση των φαρμάκων. Εν προκειμένω, η εξέταση του χρονικού προφίλ των γονιδίων σε ένα πείραμα μικροσυστοιχιών αποκτά ιδιαίτερη σημασία και πολλές ερευνητικές προσπάθειες έχουν αναπτυχθεί με στόχο να αποτυπώσουν τη χρονική δυναμική της γονιδιακής έκφρασης. Ειδικότερα, η ανάπτυξη αλγορίθμων ομαδοποίησης γονιδίων, που παρατηρούν και χρονικά προφίλ, γίνεται ολοένα και σημαντικότερη. Οι μέθοδοι παλινδρόμησης (regression fitting methods) έχουν προταθεί για τη περιγραφή των χρονικών προφίλ [75,76] και οι στατιστικές μέθοδοι τυχαίας εκκίνησης (bootstrap) έχουν αναπτυχθεί για την ανάθεση σε γονίδια υποψήφιων προφίλ [77]. Μελέτες βασισμένες σε ταξινόμηση (ranked-based) του χρονικού προφίλ έχουν προταθεί [4,5], όπου οι χρονικές ακολουθίες ταξινομούνται με βάση τις αντίστοιχες τιμές έκφρασης και χρησιμοποιούνται για να περιγράψουν το χρονικό προφίλ. Επιπλέον, μέθοδοι αμφίδρομης ταξινόμησης (*biclustering*[6]) έχουν αναπτυχθεί για να ανακαλύψουν τοπικά πρότυπα έκφρασης που συνάδουν σε μια υποομάδα συνθηκών ή χρονικών στιγμών [78,79].

Οι περισσότερες από τις ανεπτυγμένες μεθόδους μελετούν κριτήρια ομαδοποίησης που βασίζονται σε κάποια μορφή κωδικοποιημένης χρονικής συμπεριφοράς σε αντίθεση με τους παραδοσιακούς αλγορίθμους που στηρίζονται σε μεγάλο βαθμό στην έκφραση γονιδίων.

[6] Biclustering αποτελεί μια τεχνική εξόρυξης δεδομένων που επιτρέπει την ταυτόχρονη ομαδοποίηση των γραμμών και των στηλών ενός πίνακα.

Στις μεθόδους βασισμένες σε ταξινόμηση, τα υποψήφια προφίλ αποτυπώνουν αποκλειστικά το χρόνο μέσω του χαρακτηρισμού τόσο του τύπου της χρονικής εξέλιξης (αύξηση ή μείωση) όσο και των σημείων της ελάχιστης ή μέγιστης έκφρασης [5, 6]. Στους biclustering αλγορίθμους, η χρονική συμπεριφορά των γονιδίων κωδικοποιείται από σύμβολα και τα συνδεδεμένα πρότυπα έκφρασης των υποομάδων των γονιδίων σε συγκεκριμένες χρονικές στιγμές (biclusters) χαρακτηρίζονται ως ακολουθίες (strings) συμβόλων [78, 79].

Σε πολλές βιολογικές διεργασίες, ωστόσο, όχι μόνο η χρονική συμπεριφορά αλλά και το προφίλ της έκφρασης είναι το ίδιο σημαντικό για την ομαδοποίηση των γονιδίων για μια συγκεκριμένη διεργασία ή μια συνθήκη. Στην ομαδοποίηση γονιδιακών χαρακτηριστικών, ως εκ τούτου, τόσο η χρονική συμπεριφορά όσο και η έκφραση του προφίλ πρέπει να εξεταστούν. Στην εργασία μας προσπαθούμε να αναπτύξουμε κατάλληλα εργαλεία για τέτοιους στόχους ομαδοποίησης με βάση ένα διπλό κριτήριο. Πιο συγκεκριμένα, αναπτύσσουμε ένα κριτήριο βασισμένο σε προφίλ έκφρασης, το οποίο επηρεάζεται και από τη χρονική τάση. Μετά την διαδικασία της ομαδοποίησης, εξετάζουμε δύο ζητήματα που σχετίζονται με την σύγκριση των χωρικών διαμερίσεων σε ομάδες. Το πρώτο αφορά την εξαγωγή αντιστοιχιών στις δύο διαμερίσεις, ενώ το δεύτερο εξετάζει το κριτήριο εγκυρότητας ομάδων. Σκοπός της παρούσας μελέτης είναι η ανάπτυξη εργαλείων για την ομαδοποίηση και την αξιολόγηση της ποιότητας των ομάδων στηριζόμενη σε δύο κριτήρια: προφίλ έκφρασης και κωδικοποιημένες τιμές χρονικής μεταβολής . Συνολικά, η συνεισφορά της εργασίας αφορά τα εξής τρία ζητήματα:

* Ομαδοποίηση των γονιδίων με βάση το προφίλ της έκφρασης τους (γονιδιακή ομαδοποίηση), επηρεασμένη επίσης από τη χρονική συμπεριφορά του προφίλ (γονιδιακό φιλτράρισμα).
* Κριτήριο ομοιότητας για την αντιστοίχιση παρόμοιων ομάδων διαφορετικών διαμερίσεων με βάση το χρόνο.
* Δείκτης εγκυρότητας ενός διαμερισμού που αποτυπώνει τη χρονική συμπεριφορά των γονιδιακών προφίλ.

Στις επόμενες ενότητες παρέχουμε τον συμβολισμό που χρησιμοποιείται, παρουσιάζουμε τη νέα προσέγγιση ομαδοποίησης και τέλος εισάγουμε τον προτεινόμενο δείκτη εγκυρότητας.

3.2 Προφίλ έκφρασης και κωδικοποιημένη μορφή χρονικής συμπεριφοράς

Για κάθε γονίδιο, το πρότυπο έκφρασης στο χρόνο κωδικοποιείται σε ένα *N* διαστάσεων διάνυσμα $\underline{x}$, όπου *N* αντιπροσωπεύει την έκταση της χρονικής εξέλιξης ενδιαφέροντος. Η γραφική αναπαράσταση των τιμών της έκφρασης x_i,με *i=1,…, N* καθ' όλη τη χρονική εξέλιξη ενδιαφέροντος καθορίζει το χρονικό προφίλ για κάθε γονίδιο. Επιπλέον, η διαφορά των εκφράσεων μεταξύ δύο διαδοχικών χρονικών σημείων ορίζει τη χρονική μεταβολή-εξέλιξη για κάθε γονίδιο. Εμείς χρησιμοποιούμε μια κωδικοποιημένη μορφή της χρονικής μεταβολής που κωδικοποιεί, σε μια τριαδική μορφή, τις αλλαγές από το ένα στο επόμενο χρονικό σημείο. Θέτουμε $\delta_i = x_{i+1} - x_i$ με $i = 1, \dots, N-1$. Στη συνέχεια, αντιστοιχίζοντας τις διαφορετικές τιμές στο σύνολο {-1,0,1} μέσω ενός κατωφλίου (*threshold*) Δ, δημιουργούμε το κωδικοποιημένο χρονικό διάνυσμα $\underline{v}$, χρονικής μεταβολής, με διάσταση *N-1*. Πιο συγκεκριμένα, ορίζουμε τη συνάρτηση κωδικοποίησης $c\{.\}$ από την οποία προκύπτουν οι κωδικοποιημένες, χρονικής μεταβολής, τιμές:

$$v_i = c\{\delta_i\} = \begin{cases} -1 & if\ \delta_i \le -\Delta \\ 0 & if\ -\Delta \le \delta_i \le +\Delta \\ 1 & if\ \delta_i \ge +\Delta \end{cases} \qquad (23).$$

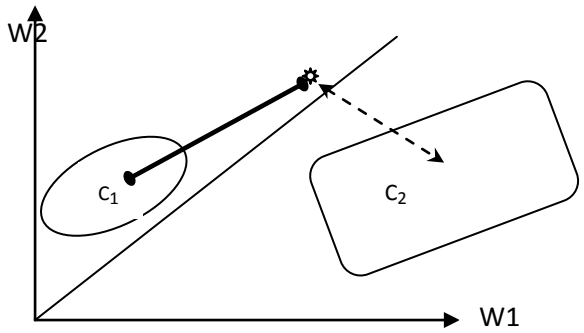
Η διαδικασία κωδικοποίησης είναι παρόμοια με εκείνη στα [80,81], αλλά με βάση ενός κατωφλίου που ορίζει όρια γύρω του μηδενός, και επίσης ομοιάζει με τη διαδικασία διακριτοποίησης στα [78, 79], με τη διαφορά ότι χρησιμοποιεί τριαδικές τιμές αντί των συμβόλων. Έτσι, στην περίπτωσή μας μία κωδικοποιημένη συμβολοσειρά χρονικής μεταβολής, αποτελείται από ψηφία-αριθμούς που μπορούν να χρησιμοποιηθούν σε αριθμητικούς υπολογισμούς σε αντίθεση με σύμβολα.

Ας υποθέσουμε ότι στοχεύουμε σε ένα διαμερισμό *S* με *C* ομάδες S_t , με *t=1,…,C*, με το μέσο διάνυσμα της S_t ομάδας να συμβολίζεται με $\underline{\mu_t}$. Η κωδικοποιημένη χρονική μεταβολή του διανύσματος αυτού προκύπτει επίσης από τον τελεστή κωδικοποίησης $c\{.\}$ και συμβολίζεται ως $\lambda_t = c\{\mu_t\}$. Ένα μέλος *j* της S με πρότυπο έκφρασης $\underline{x_j}$ $\{S_j \in S\}$, έχει ένα κωδικοποιημένο-χρονικό πρότυπο $\underline{v_j}$. Το μέγεθος N-1 του προτύπου με βάση το χρόνο , καθώς και η τριαδική διακριτοποίηση των τιμών συνεπάγονται την ύπαρξη μόνο ενός περιορισμένου αριθμού διανυσμάτων κωδικών

$M = 3^{N-1}$. Ιδανικά, κάθε ομάδα χαρακτηρίζεται από ένα πρότυπο χρονικής εξέλιξης. Στην πραγματικότητα όμως, κάθε ομάδα περιέχει έναν αριθμό από κωδικοποιημένα πρότυπα χρονικής εξέλιξης, αριθμημένα από *1* έως *M*. Η προτεινόμενη διαδικασία ομαδοποίησης δίνει *C* ομάδες από τα διανύσματα έκφρασης με βάση τόσο της έκφρασης όσο και των κωδικοποιημένων τιμών χρονικής μεταβολής . Στη μελέτη μας, η πυκνότητα μιας ομάδας εξαρτάται από τη συνοχή τόσο των εκφράσεων όσο και των κωδικών χρονικής μεταβολής.

3.3 Μεθοδολογία ομαδοποίησης

Η προτεινόμενη μεθοδολογία για την ομαδοποίηση προφίλ έκφρασης με την επίδραση του προφίλ χρονικής μεταβολής τους βασίζεται σε μια τροποποίηση της μεθοδολογίας SOM που αποτελείται από τρία μέρη. Το πρώτο βήμα εξάγει έναν αριθμό από ομάδες με βάση τη SOM οργάνωση των διανυσμάτων έκφρασης σε *QxQ* κόμβους. Στη συνέχεια τα κομβικά πρότυπα-βάρη του SOM οργανώνονται σε *C* ομάδες με βάση την επαναληπτική προσέγγιση K-means με ένα κριτήριο το οποίο επηρεάζεται από τη χρονική μεταβολή. Τέλος, κάθε αρχικό δείγμα αναθέτεται σε μία από τις *C* ομάδες με βάση την απόσταση του διανύσματος έκφρασης από το κέντρο κάθε ομάδας. Η συμβολή μας σε αυτήν τη μεθοδολογία είναι το κριτήριο που χρησιμοποιείται στη δεύτερη φάση της οργάνωσης των κόμβων του SOM. Πιο συγκεκριμένα, το πρώτο βήμα εξάγει έναν αριθμό (*QxQ*) ομάδων βασιζόμενο στα προφίλ έκφρασης. Προχωρούμε σε ένα δεύτερο επίπεδο ομαδοποίησης όπου οργανώνουμε τους κόμβους περισσότερο, εξάγοντας ένα μικρό αριθμό ομάδων. Στο σημείο αυτό προτείνουμε τη χρήση πληροφορίας σχετικά με τη χρονική τάση προκειμένου να ευνοηθεί η συνένωση των ομάδων που παρουσιάζουν παρόμοιες επιδόσεις. Αυτό απεικονίζεται γραφικά στην Εικόνα 15, όπου ένα νέο πρότυπο πρέπει να ταξινομηθεί σε μία από τις δύο υπάρχουσες ομάδες. Στο παράδειγμα αυτό, το διάνυσμα έκφρασης είναι δύο διαστάσεων (δύο χρονικά σημεία), που συνεπάγεται την ύπαρξη μίας χρονικής διαφοράς. Ο χρονικός παράγοντας είναι θετικός πάνω από τη διαγώνιο και αρνητικός κάτω από αυτήν. Σύμφωνα με το προτεινόμενο κριτήριο, προσπαθούμε να επηρεάσουμε θετικά την κατηγορία στην ίδια πλευρά (*C1*) με το πρότυπο δοκιμής ως προς το χρονικό παράγοντα, ή να πιέσουμε το πρότυπο μακριά από την κατηγορία (*C2*) της ανάποδης χρονικά πλευράς.



Εικόνα 15: Στρατηγική της επιρροής μέσω του προφίλ χρονικής μεταβολής: ομάδες που συμπεριφέρονται διαφορετικά από το πρότυπο δοκιμής στο χώρο «χρονικής μεταβολής» ωθούνται μακριά από το πρότυπο αυτό.

Η συνολική προσέγγιση της ομαδοποίησης συνοψίζεται παρακάτω:

Βήμα 1^{ov} : Αυτό-οργάνωση του χάρτη χαρακτηριστικών(SOM)

Επανάληψη των βημάτων του SOM μέχρι σύγκλισης, όταν το τετράγωνο της απόλυτης διαφοράς της μεταβολής των βαρών γίνει μικρότερο του 0.02 πάνω από 2500 εποχές. Ένα επιπλέον κριτήριο που χρησιμοποιείται περιορίζει των αριθμό των επαναλήψεων σε 500 φορές του αριθμού των κόμβων που χρησιμοποιούνται.

Βήμα 2^{ov} : K-means ομαδοποίηση των SOM κόμβων

Ο k-means αλγόριθμος στοχεύει στο να οργανώσει περαιτέρω τις ομάδες που σχηματίζονται από τον SOM σε *C* ομάδες βασισμένος στην ελαχιστοποίηση ενός κριτηρίου απόστασης μεταξύ των δειγμάτων (κόμβοι SOM) και των κέντρων της ομάδας, έτσι ώστε να ελαχιστοποιήσει τη συνολική διακύμανση κάθε ομάδας. Το k-means σχήμα λειτουργεί με μία επαναληπτική φόρμα που βασίζεται στην λύση της μεγιστοποίησης της πρόβλεψης/μεγιστοποίησης (EM). Το προτεινόμενο κριτήριο στο βήμα αυτό ορίζεται ως η συνολική l_2 νόρμα των διαφορών των εκφρασμένων-διανυσμάτων:

$$Q = \sum_{t=1}^{C} \sum_{w_j \in S_t} \alpha_j \| \underline{w}_j - \underline{\mu}_t \|_2^2 \qquad (24)$$

όπου $\underline{w_j}$ είναι ένα δείγμα (διάνυσμα έκφρασης ενός κόμβου του SOM) το οποίο ανήκει στην ομάδα t με κέντρο $\underline{\mu_t}$ και a_j είναι ένας σταθμικός παράγοντας που εξαρτάται από την κωδικοποιημένη χρονική μεταβολή του $\underline{w_j}$ σε σύγκριση με εκείνη του κέντρου της ομάδας $\underline{\mu_t}$.

Πιο συγκεκριμένα, το

$$a_j = \frac{1}{1+e^{-6(r_j-1)}} + 1 \tag{25}$$

είναι μια λογιστική συνάρτηση (logistic function) από 1 έως 2 {[1,2]} εκτιμώμενη από την παράμετρο r_j. Ο παράγοντας αυτός αντικατοπτρίζει την κωδικοποιημένη διαφορά $r_j = \left\| \underline{v}_j - \underline{m}_t \right\|_1$ ως την l_1 νόρμα των κωδικοποιημένων-χρονικής μεταβολής-διαφορών που αντιστοιχούν στο δείγμα και στο κέντρο της ομάδας, όπου $\underline{v}_j = c\{\underline{w}_j\}$ και $\underline{m}_t = c\{\underline{\mu}_t\}$. Σημειώνεται ότι το r_j κυμαίνεται μεταξύ του εύρους [0,2], με 0 και 2 να είναι οι περιπτώσεις της μηδενικής διαφοροποίησης και της μέγιστης, αντίστοιχα, σε όλα τα ψηφία του κωδικοποιημένου διανύσματος μεγέθους τριών χρονικών στιγμών. Στην περίπτωση της μηδενικής διαφοροποίησης, ο χρονικός παράγοντας a_j γίνεται 1, ενώ στη μέγιστη διαφορά μεταξύ όλων των ψηφίων ο παράγοντας γίνεται 2, αυξάνοντας με τον τρόπο αυτό την απόσταση στην (24) με βάση τις εκφράσεις. Λαμβάνοντας υπόψη το κριτήριο αυτό, η ΕΜ βελτιστοποίηση προχωρεί σε διακριτά βήματα προς την κατεύθυνση του υπολογισμού των νέων κέντρων των ομάδων και της ανάθεσης εκ νέου των δειγμάτων σε ομάδες. Σε κάθε βήμα, για ορισμένα δείγματα που τοποθετούνται σε μία ομάδα, το νέο κέντρο υπολογίζεται ως εξής:

$$\underline{\mu}_t = \frac{1}{|S_t|} \sum_{w_j \in S_t} \alpha_j \, \underline{w}_j \tag{26}$$

όπου $|S_t|$ η διάσταση της ομάδας S_t.

Στη συνέχεια τα δείγματα αξιολογούνται και επανατοποθετούνται με βάση την ελάχιστη απόσταση από τα κέντρα των κλάσεων, δηλαδή $\min_t \ \hat{\alpha}_{j,t} \| \underline{w}_j - \underline{\hat{\mu}}_t \|_2^2$, με τους εκ νέου χρονικούς παράγοντες $\hat{a}_{j,t}$ για το δείγμα j και το υπό δοκιμή κεντρικό διάνυσμα της κλάσης t.

Τα βασικά βήματα είναι τα εξής:

1. Ορισμός αριθμού των ομάδων.
2. Πραγματοποίηση τυχαίας ανάθεσης των δειγμάτων σε C ομάδες με αρχική τιμή για το χρονικό παράγοντα μεταβολής $\hat{a}_{j,t} = 1$.

3. Υπολογισμός του κέντρου μ_t της ομάδας με βάση το σταθμισμένο χρονικά μέσο όρο.
4. Υπολογισμός του νέου χρονικού παράγοντα μεταβολής $\hat{a}_{j,t}$ και επανατοποθέτηση των δειγμάτων σε ομάδες με βάση την ελάχιστη σταθμισμένη απόσταση για τα κέντρα των ομάδων.
5. Επανάληψη των βημάτων (3-4) μέχρι να επιτευχθεί το κριτήριο σύγκλισης.

Βήμα 3^{ov} : Ανάθεση των αρχικών τιμών έκφρασης σε ομάδες

Το βήμα δύο οργανώνει τους κόμβους του SOM δικτύου, αλλά τα αρχικά διανύσματα έκφρασης εξακολουθούν να πρέπει να τοποθετηθούν σε κάποια από τις C ομάδες. Η ανάθεση αυτή γίνεται με βάση την ελάχιστη l_2 νόρμα της διαφοράς μεταξύ κάθε διανύσματος έκφρασης $\underline{x_i}$ και του κέντρου κάθε ομάδας $\underline{\mu_t}$ που υπολογίζονται από το βήμα 2.

3.4 Κριτήρια για την αξιολόγηση της κατανομής χρονικών μεταβολών

Στην υπολογιστική βιολογία υπάρχει συχνά η ανάγκη να συγκριθούν δύο ή περισσότερες διαμερίσεις. Ειδικότερα, χρειάζεται να ανακαλύψουμε αντιστοιχίες μεταξύ των διαμερίσεων αλλά επίσης να συγκρίνουμε την ποιότητα κάθε διαμέρισης από την άποψη της πυκνότητας και της διάκρισης των ομάδων του. Για το πρώτο σκέλος χρειάζεται να συγκρίνουμε τη μία ομάδα της μία διαμέρισης με όλες τις ομάδες της άλλης διαμέρισης. Για το δεύτερο πρόβλημα πρέπει να εξετάσουμε την κατανομή των προτύπων εντός και μεταξύ των ομάδων κάθε διαμέρισης. Για τη σύγκριση αυτή έχουμε δύο διαθέσιμες τιμές , δηλαδή τις τιμές των δειγμάτων και τις πιθανότητες αυτών. Στη δική μας εφαρμογή όπου ενδιαφερόμαστε για χρονικές ομοιότητες/διαφορές, έχουμε δείγματα κωδικοποιημένα με τους κωδικούς χρονικής μεταβολής ως τριαδικές συμβολοσειρές, με μέγεθος κάθε συμβολοσειράς N-1, και με αριθμό δειγμάτων από 1 έως M=3^{N-1} .

3.4.1 Αντιστοίχιση ομάδων με βάση το προφίλ χρονικής μεταβολής

Σε κάθε ομάδα C_i έχει ανατεθεί ένας αριθμός L_i από δείγματα, το κάθε ένα με ένα αντίστοιχο κωδικοποιημένο διάνυσμα χρονικής μεταβολής. Η κατανομή των χρονικών αυτών διανυσμάτων συνοψίζεται σε ένα ιστόγραμμα κωδικών με αριθμημένους κωδικούς *1,…,M* με πιθανότητες *p[1],…,p[M]*. Το προτεινόμενο μέτρο ομοιότητας βασίζεται στη διαφορά των ιστογραμμάτων αυτών των κωδικών. Κατά τη διαδικασία ταύτισης, κάθε ομάδα από ένα διαμερισμό συσχετίζεται με την καλύτερα αντιστοιχιζόμενη ομάδα της άλλης διαμέρισης. Συνεπώς, ας ορίσουμε δύο ομάδες με πιθανότητες αριθμημένων κωδικών *p[1],…,p[M]* και *q[1],…,q[M]* . Η ομοιότητα των δύο ομάδων μπορεί να ορισθεί σαν το άθροισμα της απόλυτης απόστασης των δύο κατανομών:

$$D_c = \sum_{i=1}^{M} | \, p[i] - q[i] \, | \qquad (27).$$

3.4.2 Δείκτης εγκυρότητας με βάση το προφίλ χρονικής μεταβολής

Ο δείκτης που προτείνεται βασίζεται στην πυκνότητα του ιστογράμματος κωδικών καθώς και στην απόσταση μεταξύ ζευγαριών των κωδικών του ιστογράμματος. Εξαιτίας της κατανομής των πιθανοτήτων ή της σταθμισμένης διαμόρφωσης του ιστογράμματος, το μεγαλύτερο κουτί (bin) του ιστογράμματος παίζει πολύ πιο σημαντικό ρόλο στο μέτρο εγκυρότητας σε σύγκριση με μικρότερα bins. Κατά συνέπεια, κατατάσσουμε τις πιθανότητες σε φθίνουσα σειρά για να καταλήξουμε στις ταξινομημένες πιθανότητες $p_1,\ldots,p_M$. Να σημειωθεί ότι ισοψηφίες στην κατάταξη με βάση την πιθανότητα λύονται προς όφελος της ελάχιστης απόστασης Hamming[7] από την προηγούμενη συμβολοσειρά. Κατά συνέπεια, αναθέτουμε την ακόλουθη ορολογία:

Αριθμημένες πιθανότητες *p[1] ,p[2],…,p[M]*

[7] Hamming απόσταση μεταξύ δύο συμβολοσειρών ίδιου μήκους είναι ο αριθμός των θέσεων στις οποίες τα αντίστοιχα σύμβολα είναι διαφορετικά.

Ταξινομημένες πιθανότητες $p_1, p_2,\ldots,p_M$ όπου $p_1 = \max_i \{p[i]\}$ και $p_k = \max_i^k \{p[i]\}$ είναι η *k-οστή* υψηλότερη καταταγμένη πιθανότητα.

Ας υποθέσουμε ότι έχουμε δύο ομάδες *C1* και *C2* με μέγεθος *L1* και *L2*, αντίστοιχα. Η πρώτη έχει ταξινομημένες πιθανότητες p_i που αντιστοιχούν σε συμβολοσειρές s_i, ενώ η δεύτερη έχει q_i ταξινομημένες πιθανότητες που αντιστοιχούν σε συμβολοσειρές t_i, με $i=1,\ldots,M$. Για να υπολογιστεί ο εσωτερικός δείκτης πυκνότητας της ομάδας, χρησιμοποιούμε την Hamming απόσταση μεταξύ δύο διαδοχικών συμβολοσειρών (κωδικών) κατά φθίνουσα βαθμολογική διάταξη, σταθμισμένη με την πιθανότητα της δεύτερης συμβολοσειράς. Αυτό το μέτρο εισάγει την απόσταση ανάμεσα σε δύο διαδοχικές συμβολοσειρές μιας συγκεκριμένης ομάδας από την σημαντικότερη στη λιγότερο σημαντική. Ξεκινώντας το μέτρημα των αποστάσεων για την ομάδα αυτή, η εντός-ομάδας απόσταση που καθορίζεται από τη πιο σημαντική συμβολοσειρά είναι μηδέν. Λαμβάνοντας υπόψη την επόμενη σε σειρά κατάταξης συμβολοσειρά, η απόσταση που εισάγεται αντιστοιχεί στη διαφορά των κωδικών αλλά σταθμισμένη και με την πιθανότητα της νέας συμβολοσειράς. Γενικά, για την *i-οστή* ταξινομημένη συμβολοσειρά, η απόσταση που εισάγεται μπορεί να εκφραστεί από την απόστασή της (codeword) από την προηγούμενη συμβολοσειρά σταθμισμένη από την πιθανότητα της i-οστής συμβολοσειράς. Επομένως, για την εντός ομάδας απόσταση έχουμε:

$$Q(C_1) = \sum_{i=1}^{M-1} d(s_i, s_{i+1}) p_{i+1} \tag{28}$$

$$Q(C_2) = \sum_{i=1}^{M-1} d(t_i, t_{i+1}) q_{i+1} \tag{29}$$

όπου d(.,.) υποδηλώνει τη Hamming απόσταση μεταξύ δύο συμβολοσειρών.

Στοχεύοντας τώρα στο να εισάγουμε αποστάσεις μεταξύ ομάδων (across-cluster), πρώτα μελετάμε τις διαφορές δύο ομάδων για κάθε συμβολοσειρά. Έτσι, για κάθε αριθμημένο κωδικό υπολογίζουμε τη διαφορά πιθανοτήτων: σε

αριθμημένες πιθανότητες $r[i] = |p[i] - q[i]|$ και έπειτα εκφράζουμε τις

ταξινομημένες πιθανότητες $r_1, r_2, \ldots, r_M$ όπου $r_1 = \max_i \{r[i]\}$ και $r_k = \max_i^k \{r[i]\}$ είναι η *k-οστή* υψηλότερη καταταγμένη πιθανότητα. Για πιθανές συγκρίσεις των ορίων σημειώνουμε ότι $r_1 \leq \max\{p_1, q_1\}$ και $r_1 \geq |p_1 - q_1|$.

Έχουμε ορίσει στο σημείο αυτό ένα ιστόγραμμα διαφορών ή τομής πιθανοτήτων των δύο ομάδων *C1* και *C2* , απ' όπου ο δείκτης απόστασης μεταξύ ομάδων μπορεί να ορισθεί όπως προηγουμένως:

$$Q(C_1, C_2) = \sum_{i=1}^{M-1} d(z_i, z_{i+1}) r_{i+1} \quad \{\geq 1 - r_1\} \tag{30}.$$

Οι δείκτες αυτοί αντικατοπτρίζουν την συνολική απόσταση των κωδικών-λέξεων (συμβολοσειρών) σταθμισμένη με τις αντίστοιχες πιθανότητές τους. Οι τιμές τους κυμαίνονται σε ένα εύρος [*1,K*] με κλίμακα (*1-r_1*). Η ελάχιστη τιμή του δείκτη είναι μηδέν και επιτυγχάνεται όταν το ιστόγραμμα έχει μόνο ένα σημείο με πιθανότητα ένα. Η μέγιστη τιμή *K* προκύπτει όταν το ιστόγραμμα περιλαμβάνει δύο κωδικούς με μέγιστη απόσταση *K* και με πιθανότητα ½ ο καθένας. Κατά συνέπεια, ο λόγος μετρικών εντός-ομάδων προς ζευγών ομάδων, καθορίζει ένα κατάλληλο δείκτη εγκυρότητας για κάθε διαμερισμό *P* των C_p ομάδων που αναφέρεται ως Ταξινομημένος Δείκτης Χρονικής Μεταβολής (Ranked Shape Index) και ερμηνεύει τις χρονικές ομοιότητες ,ως εξής:

$$RSI\{P\} = \frac{1}{C_P} \sum_{i=1}^{C_P} \max_{j \neq i} \left\{ \frac{Q(C_i) + Q(C_j)}{Q(C_i, C_j)} \right\}, \; i, j = 1, \ldots, C_P. \tag{31}$$

Κεφάλαιο 4

Αλγόριθμοι οπτικοποίησης

4.1 Οπτικοποίηση δεδομένων

Η οπτικοποίηση των δεδομένων (data visualization), δίνει μία παρουσίαση της βάσης δεδομένων χρησιμοποιώντας τους πιο σημαντικούς βαθμούς ελευθερίας (διαστάσεις), συνήθως μετά τη διαδικασία της μείωσης διαστάσεων (dimensionality reduction) των δεδομένων [82]. Αποτελεί μία μελέτη της οπτικής αναπαράστασης των δεδομένων, παρουσιάζοντας πληροφορίες που έχουν αντληθεί από κάποια σχηματική μορφή, συμπεριλαμβάνοντας γενικά χαρακτηριστικά ή μεταβλητές των δεδομένων.

Σύμφωνα με το Friedman [83], ο κύριος στόχος της οπτικοποίησης δεδομένων είναι η κοινοποίηση πληροφορίας καθαρά και αποτελεσματικά, χρησιμοποιώντας γραφικά μέσα. Αυτό δε σημαίνει ότι η διαδικασία αυτή πρέπει να δείχνει βαρετή για να είναι λειτουργική, ή να είναι ιδιαίτερα πολύπλοκη για να εμφανίζεται ευπαρουσίαστη. Για να αναπτύξουμε τις ιδέες με επιτυχία, τόσο η αισθητική μορφή όσο και η λειτουργικότητα θα πρέπει να συμπορεύονται, ώστε να παρέχει το αποτέλεσμα γνώσεις καλού επιπέδου.

Η οπτικοποίηση των δεδομένων είναι στενά συνδεδεμένη με την γραφική πληροφορία (information graphics), την απεικόνιση πληροφορίας (information visualization), την επιστημονική οπτικοποίηση (scientific visualization) και τη γραφική στατιστική (statistical graphics). Η οπτικοποίηση των δεδομένων αποτελεί έναν ενεργό τομέα της έρευνας, της διδασκαλίας και της ανάπτυξης. Έχει ενοποιήσει τα πεδία της πληροφοριακής και επιστημονικής απεικόνισης [84]. Η KPI βιβλιοθήκη έχει αναπτύξει τον Περιοδικό Πίνακα Μεθόδων Οπτικοποίησης [85], ένα διαδραστικό γράφημα που εμφανίζει μια ποικιλία μεθόδων για απεικόνιση της πληροφορίας.

Υπάρχουν διαφορετικές προσεγγίσεις στο πεδίο της οπτικοποίησης δεδομένων. Στην παρούσα διπλωματική, η οπτικοποίηση γονιδιακών δεδομένων (visualization of genomic data), που αποτελεί και τον κύριο στόχο αυτής, επιτυγχάνεται μέσω της διαδικασίας της μείωσης διαστάσεων (Ενότητα 4.2) της αρχικής πληροφορίας με διάφορες μεθόδους. Έτσι, η ομαδοποίηση που γίνεται στα δεδομένα, σε συνδυασμό

με την τεχνική της οπτικοποίησης , μας δίνει τη δυνατότητα να έχουμε σε δύο μόνο διαστάσεις τις διάφορες ομάδες που έχουν προκύψει και να εξάγουμε από την προβολή αυτών, σημαντικές πληροφορίες και συμπεράσματα για τη φύση των δεδομένων που επεξεργαζόμαστε.

4.2 Μείωση διαστάσεων

Τα δεδομένα του πραγματικού κόσμου όπως τα σήματα ομιλίας, οι ψηφιακές φωτογραφίες, οι γονιδιακές βάσεις δεδομένων, που αποτελούν αντικείμενο μελέτης της εν λόγω διπλωματικής, έχουν μεγάλες διαστάσεις [86]. Για την ικανοποιητική διαχείριση αυτών των πραγματικών δεδομένων, η διάσταση τους πρέπει να μειωθεί. Ως μείωση διαστάσεων (dimensionality reduction) ορίζεται η μετατροπή των μεγάλων διαστάσεων δεδομένων (high-dimensional data) σε μια ουσιαστική εκπροσώπηση των ελαττωμένων διαστάσεων. Με άλλα λόγια, παράγει μια συμπαγή, χαμηλών διαστάσεων (low dimensional), κωδικοποίηση, δοσμένου ενός συνόλου δεδομένων υψηλών διαστάσεων [82].

Ιδανικά, η μειωμένη αναπαράσταση θα πρέπει να έχει μία μορφή που να ανταποκρίνεται στην πραγματική κατανομή των δεδομένων. Η εγγενής αναπαράσταση των στοιχείων επιτυγχάνεται με τον ελάχιστο αριθμό των παραμέτρων που απαιτούνται για τον υπολογισμό των παρατηρούμενων ιδιοτήτων της πληροφορίας [87].

Η μείωση διαστάσεων παρουσιάζει μεγάλη σπουδαιότητα σε διάφορους τομείς, δεδομένου ότι μετριάζει το φαινόμενο της διάστασης (curse of dimensionality) και άλλες ανεπιθύμητες ιδιότητες των χώρων υψηλών διαστάσεων [88]. Ως εκ τούτου, διευκολύνει μεταξύ άλλων την ταξινόμηση (classification), την οπτικοποίηση (visualization) και τη συμπίεση μεγάλων διαστάσεων δεδομένων. Παραδοσιακά, πραγματοποιείται με τη χρήση τόσο γραμμικών μεθόδων όσο και μη γραμμικών διαδικασιών.

Στη μηχανική μάθηση (machine learning), η μείωση διαστάσεων αποτελεί τη μείωση του αριθμού των τυχαίων μεταβλητών υπό εξέταση και μπορεί να διακριθεί στην *επιλογή χαρακτηριστικών* (feature selection) και στην *εξαγωγή χαρακτηριστικών* (feature extraction) [89, 90, 91].

4.2.1 Επιλογή χαρακτηριστικών

Η *επιλογή χαρακτηριστικών* ή επιλογή μεταβλητών (variable selection) ή μείωση χαρακτηριστικών (*feature reduction*) όπως αλλιώς είναι γνωστή, είναι η επιλογή ενός υποσυνόλου των αρχικών μεταβλητών για την οικοδόμηση ισχυρών μοντέλων μάθησης [92]. Όταν εφαρμόζεται στον τομέα της βιολογίας, καλείται διακριτική επιλογή γονιδίων (discriminative gene selection), η οποία ανιχνεύει αλληλεπιδράσεις γονιδίων, ορμώμενες από πειράματα σε μικροσυστοιχίες DNA. Με την απομάκρυνση των άνευ σημασίας και περιττών αντικειμένων της βάσης δεδομένων, η επιλογή χαρακτηριστικών συμβάλλει στη βελτίωση της απόδοσης των μαθησιακών μοντέλων με:

* Άμβλυνση της επίδρασης από την κατάρα της διαστατικότητας.
* Ενίσχυση της ικανότητας γενίκευσης.
* Επιτάχυνση της διαδικασίας μάθησης.
* Βελτίωση της ποιότητας του μοντέλου.

Η επιλογή χαρακτηριστικών βοηθάει επιπρόσθετα τους ανθρώπους να αποκτήσουν καλύτερη κατανόηση, υποδεικνύοντάς τους ποια είναι τα σημαντικά χαρακτηριστικά και πως αυτά σχετίζονται μεταξύ τους.

4.2.2 Εξαγωγή χαρακτηριστικών

Στην αναγνώριση προτύπων (pattern recognition) και στην επεξεργασία εικόνας (image processing), η *εξαγωγή χαρακτηριστικών* (feature extraction*)* αποτελεί μια ιδιαίτερη μορφή της μείωσης διαστάσεων .

Η εξαγωγή χαρακτηριστικών μετατρέπει τα δεδομένα από ένα χώρο υψηλών διαστάσεων σε ένα επίπεδο χαμηλών διαστάσεων [93]. Η μετατροπή των δεδομένων μπορεί να είναι γραμμική (linear), όπως η Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis-PCA), αλλά και ποικίλες μη γραμμικές τεχνικές (non linear) υφίστανται [89, 90].

Η κύρια γραμμική μέθοδος μείωσης διαστάσεων, η ανάλυση σε κύριες συνιστώσες, εφαρμόζει μια γραμμική χαρτογράφηση των δεδομένων σε ένα χαμηλό διαστάσεων χώρο, με τέτοιο τρόπο ώστε η διακύμανση/διασπορά των δεδομένων στην χαμηλών διαστάσεων αναπαράσταση να μεγιστοποιείται. Στην πράξη, ο πίνακας συσχέτισης

(correlation matrix) των δεδομένων κατασκευάζεται και τα ιδιοδιανύσματα στον εν λόγω πίνακα υπολογίζονται. Τα ιδιοδιανύσματα που αντιστοιχούν στις υψηλότερες ιδιοτιμές (κύριες συνιστώσες), μπορούν να χρησιμοποιηθούν για να αναπαραστήσουν ένα μεγάλο μέρος της διασποράς των αρχικών δεδομένων. Επιπλέον, τα πρώτα ιδιοδιανύσματα μπορούν συχνά να ερμηνευτούν υπό την άποψη της μεγάλης κλίμακας συμπεριφοράς του συστήματος. Ο αρχικός χώρος (με διάσταση του αριθμού των στοιχείων) έχει μειωθεί (με την απώλεια δεδομένων, αλλά ελπίζοντας στη διατήρηση της πιο σημαντικής διασποράς) στο χώρο που παράγεται από λίγα ιδιοδιανύσματα.

Η ανάλυση κύριων συνιστωσών μπορεί να εφαρμοστεί σε ένα μη γραμμικό τρόπο με το κόλπο του πυρήνα (kernel trick). Η τεχνική που προκύπτει είναι ικανή να κατασκευάσει μη γραμμικές απεικονίσεις που μεγιστοποιούν τη διασπορά των δεδομένων. Η προκύπτουσα μέθοδος αποκαλείται Ανάλυση Κύριων Συνιστωσών με χρήση Πυρήνα (Kernel Principal Component Analysis, k-PCA). Άλλες μη γραμμικές μέθοδοι περιλαμβάνουν πολλαπλές τεχνικές μάθησης (manifold learning) όπως η γραμμική τοπική ενσωμάτωση (locally linear embedding- LLE), Hessian LLE, Λαπλασιανοί πίνακες (Laplacian eigenmaps) και LTSA [86]. Οι τεχνικές αυτές δημιουργούν μια χαμηλής διάστασης δεδομένα, χρησιμοποιώντας μια συνάρτηση κόστους που διατηρεί τις τοπικές ιδιότητες των στοιχείων και μπορεί να παρατηρηθεί ορίζοντας ένα γράφο βασισμένο σε ένα πυρήνα, στην περίπτωση του k-PCA. Τα τελευταία χρόνια, τεχνικές έχουν προταθεί, οι οποίες αντί να καθορίζουν ένα σταθερό πυρήνα, προσπαθούν να εκπαιδεύσουν τον πυρήνα χρησιμοποιώντας *semidefinite προγραμματισμό*. Το πιο χαρακτηριστικό παράδειγμα αυτής της κατηγορίας είναι η εξέλιξη της μέγιστης διακύμανσης (maximum variance unfolding-MVU) [86]. Η κεντρική ιδέα της MVU είναι η επακριβώς διατήρηση όλων των ζευγών αποστάσεων μεταξύ των πλησιέστερων γειτόνων, ενώ παράλληλα να μεγιστοποιήσει τις αποστάσεις μεταξύ σημείων που δεν αποτελούν κοντινούς γείτονες.

Μια εναλλακτική προσέγγιση της διατήρησης της γειτονίας , εμφανίζεται μέσω της ελαχιστοποίησης του κόστους λειτουργίας που υπολογίζει τις διαφορές μεταξύ του χώρου εισόδου και εξόδου. Χαρακτηριστικές τεχνικές αυτού του τύπου είναι οι Isomap, diffusion Maps και οι t-SNE.

Στην παρούσα διπλωματική εργασία, για την οπτικοποίηση των δεδομένων γονιδιακής έκφρασης, επιλέχθηκαν δύο μέθοδοι μείωσης διαστάσεων. Η μία είναι η Ανάλυση Κύριων Συνιστωσών που έρχεται από το χώρο των γραμμικών τεχνικών. Η δεύτερη είναι η Ανάλυση Κύριων Συνιστωσών σε Πυρήνα που ανήκει στην κατηγορία των μη γραμμικών μεθόδων. Η αναλυτική περιγραφή αυτών και ο τρόπος υλοποίησή τους παρουσιάζεται στις επόμενες ενότητες.

4.3 Γνωστικό υπόβαθρο

Στην προκειμένη ενότητα, περιλαμβάνεται η θεμελίωση του γνωστικού υποβάθρου που αφορά τις διαδικασίες μείωσης διαστάσεων που υλοποιούνται. Συγκεκριμένα αναφέρονται και εξηγούνται βασικές μαθηματικές έννοιες όπως τυπική απόκλιση, διασπορά, ενώ η ενότητα ολοκληρώνεται με ανάπτυξη εννοιών από τον τομέα της άλγεβρας πινάκων.

4.3.1 Στατιστικές έννοιες

Στατιστική [94] είναι ο κλάδος των εφαρμοσμένων μαθηματικών, ο οποίος βασίζεται σε ένα σύνολο αρχών και μεθοδολογιών για:

- Το σχεδιασμό της διαδικασίας συλλογής δεδομένων.
- Τη συνοπτική και αποτελεσματική παρουσίαση δεδομένων.
- Την ανάλυση και εξαγωγή συμπερασμάτων από τα δεδομένα.

Η στατιστική είναι μια ατελής επαγωγή. Από τις ιδιότητες του μέρους εξάγει συμπεράσματα για το όλον.

Το όλο θέμα των στατιστικών στοιχείων βασίζεται στην ιδέα ότι υπάρχει ένα μεγάλο σύνολο δεδομένων και επιθυμούμε να αναλύσουμε το σύνολο αυτό, υπό την οπτική των σχέσεων μεταξύ των επιμέρους στοιχείων της βάσης δεδομένων. Γίνεται, επομένως εστίαση, σε μερικά από τα μέτρα τα οποία μπορούμε να εφαρμόσουμε στα αντικείμενά μας, καθώς και στην πληροφορία που μπορεί να αντληθεί από αυτά για τα ίδια τα δεδομένα.

4.3.1.1 Τυπική απόκλιση

Στη στατιστική και στη θεωρία πιθανοτήτων, η *τυπική απόκλιση* (standard deviation) δείχνει πόση διασπορά ή διακύμανση υπάρχει από το μέσο όρο(μέση ή αναμενόμενη τιμή). Δηλαδή, αποτελεί ένα μέτρο που υπολογίζει το μέσο όρο της απόστασης, από το μέσο στοιχείο της βάσης δεδομένων, για κάθε σημείο αυτής. Μια χαμηλή τυπική απόκλιση δείχνει ότι τα σημεία των δεδομένων τείνουν να είναι πολύ κοντά στο μέσο όρο, ενώ η υψηλή τυπική απόκλιση υποδεικνύει ότι τα στοιχεία κατανέμονται γύρω από ένα μεγάλο εύρος τιμών.

Η τυπική απόκλιση τυχαίας μεταβλητής, στατιστικής πληθυσμού, συνόλου δεδομένων ή κατανομή πιθανοτήτων είναι η τετραγωνική ρίζα της διακύμανσής της. Είναι αλγεβρικά απλή, αν και πρακτικά είναι λιγότερη ισχυρή από τη μέση απόλυτη απόκλιση. Μια χρήσιμη ιδιότητα της τυπικής απόκλισης, είναι ότι σε αντίθεση με τη διασπορά, αυτή εκφράζεται στις ίδιες μονάδες με τα δεδομένα.

Για τον υπολογισμό της τιμής αυτής, υπολογίζονται αρχικά τα τετράγωνα των αποστάσεων μεταξύ κάθε στοιχείου και του μέσου στοιχείου και στη συνέχεια αθροίζονται. Όλη η προηγούμενη ποσότητα διαιρείται με την τιμή n, όπου το πλήθος των στοιχείων, και εν συνεχεία λαμβάνεται η θετική τιμή της τετραγωνικής ρίζας. Ο μαθηματικός τύπος της τυπικής απόκλισης, s, είναι ο εξής:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}} \tag{32}$$

όπου, X_i είναι το i-οστό στοιχείο του συνόλου, και $\bar{X}$ μέση τιμή του συνόλου των δεδομένων που δίνεται από τον τύπο $\frac{\sum_{i=1}^{n} X_i}{n}$.

4.3.1.2 Διακύμανση

Η *διακύμανση* (variance) αποτελεί ένα άλλο μέτρο της διασποράς των στοιχείων σε ένα σύνολο δεδομένων. Στην πραγματικότητα είναι σχεδόν ταυτόσημη με την τυπική απόκλιση. Ο τύπος υπολογισμού είναι ο εξής:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} \quad (33)$$

Η διακύμανση ή διασπορά δεν είναι τίποτα άλλο από το τετράγωνο της τυπική απόκλισης. Ο συμβολισμός s^2 είναι ο πιο συνηθισμένος για την αναφορά στη διακύμανση ενός δείγματος. Μαζί τα δύο αυτά μέτρα εφαρμόζονται για την εκτίμηση της διασποράς των δεδομένων. Η τυπική απόκλιση είναι το πιο συχνά χρησιμοποιούμενο μέτρο αλλά και η διακύμανση χρησιμοποιείται.

4.3.1.3 Συνδιακύμανση

Η *συνδιακύμανση (*covariance*)* αποτελεί ένα μέτρο της σχέσης μεταξύ δύο περιοχών δεδομένων. Σαν εργαλείο ανάλυσης, η συνδιακύμανση αποδίδει το μέσο όρο του γινομένου των αποκλίσεων των δεδομένων από τις αντίστοιχες μέσες τιμές τους, βάση του τύπου που έπεται:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n} \quad (34)$$

όπου $\bar{X}, \bar{Y}$ οι μέσες τιμές των στοιχείων των X και Y. Μπορεί να χρησιμοποιηθεί για να καθορισθεί κατά πόσο μεγάλες τιμές του ενός συνόλου σχετίζονται με μεγάλες τιμές του άλλου (θετική συνδιακύμανση) ή κατά πόσο μικρές τιμές του ενός συνόλου σχετίζονται με μεγάλες τιμές του άλλου (αρνητική συνδιακύμανση) ή τέλος κατά πόσο οι τιμές και των δύο συνόλων είναι ασυσχέτιστες (σχεδόν μηδενική συνδιακύμανση).

4.3.1.4 Πίνακας συνδιακύμανσης

Στο τομέα της στατιστικής, ο πίνακας συνδιακύμανσης είναι ένας πίνακας το στοιχείο του οποίου στην *i, j* θέση είναι η τιμή της συνδιακύμανσης μεταξύ του i-οστού και j-οστού στοιχείο ενός τυχαίου διανύσματος. Κάθε στοιχείο του διανύσματος είναι μια μονοδιάστατη τυχαία μεταβλητή, είτε με έναν πεπερασμένο αριθμό των παρατηρούμενων εμπειρικών τιμών ή με ένα πεπερασμένο ή άπειρο αριθμό πιθανών τιμών που καθορίζονται από μία κοινή θεωρητική κατανομή πιθανότητας όλων των τυχαίων μεταβλητών.

Διαισθητικά, ο πίνακας συνδιασποράς όπως αλλιώς ονομάζεται, γενικεύει την έννοια της διακύμανσης σε πολλαπλές διαστάσεις. Για παράδειγμα, αν έχουμε ένα σύνολο δεδομένων με διαστάσεις περισσότερων των δύο, υπάρχει παραπάνω από ένα μέτρο συνδιακύμανσης που πρέπει να υπολογιστούν. Στην περίπτωση μιας βάσης δεδομένων με τρεις διαστάσεις (*διαστάσεις x, y, z*) πρέπει να υπολογιστούν οι ποσότητες *cov(x,y), cov(x,z)* και *cov(y,z).* Στην πραγματικότητα, για ένα n-διαστάσεων σύνολο δεδομένων θα πρέπει να υπολογιστούν $\frac{n!}{(n-2)!*2}$, διαφορετικές τιμές συνδιακύμανσης.

Ένας εύκολος τρόπος για να παρθούν όλες οι πιθανές τιμές συνδιασποράς μεταξύ όλων των διαφορετικών διαστάσεων είναι να υπολογιστούν αυτές και έπειτα να τοποθετηθούν σε ένα πίνακα. Επομένως, ο ορισμός ενός πίνακα συνδιακύμανσης για ένα σύνολο δεδομένων με n-διαστάσεις είναι:

$$C^{n\times n} = (c_{i,j}, c_{i,j} = cov(Dim_i, Dim_j)) \quad (35)$$

όπου C^{nxn} είναι ένα πίνακας με *n* γραμμές και *n* στήλες και Dim_x είναι η *x*-οστή διάσταση.

Ένα παραστατικό παράδειγμα πίνακα, που αναφέρεται στην περίπτωση των τριών διαστάσεων, όπως αναφέρθηκε προηγουμένως είναι το εξής:

$$C = \begin{pmatrix} cov(x,y) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix} \quad (36)$$

Αξίζει να σημειωθεί, ότι πάνω στην κύρια διαγώνιο, υπάρχουν οι τιμές συνδιακύμανσης της κάθε διάστασης με τον εαυτό της. Αυτές αποτελούν τη διασπορά για κάθε μία διάσταση. Επίσης, αφού οι τιμές *cov(a,b), cov(b,a)* είναι ίσες, ο πίνακας παρουσιάζει συμμετρία γύρω από τη κύρια διαγώνιο.

4.3.2 Διανυσματική Ανάλυση/Άλγεβρα πινάκων

Στην ενότητα αυτή γίνεται αναφορά σε δύο έννοιες, θεμελιώδεις στην άλγεβρα πινάκων (*matrix algebra*) [95]. Η πρώτη είναι τα *ιδιοδιανύσματα* (eigenvectors) ενός τετραγωνικού πίνακα και η δεύτερη είναι οι *ιδιοτιμές* (eigenvalues).

Έστω ένας τετραγωνικός πίνακας *A*, ο οποίος δρα σε ένα διάνυσμα (πίνακα στήλη) $\vec{x}$. Το αποτέλεσμα θα είναι ένα νέο διάνυσμα (πίνακας στήλη) $\vec{y}$, δηλαδή:

$$A\vec{x} = \vec{y} \tag{37}$$

Για κάθε πίνακα *A*, υπάρχουν κάποια χαρακτηριστικά διανύσματα, τέτοια ώστε η δράση του *A* πάνω σε αυτά να αφήνει αναλλοίωτη τη διεύθυνσή του και απλώς να τα πολλαπλασιάζει με ένα αριθμό (δηλαδή αλλάζει μόνο το μέτρο τους, τα <<διαστέλλει>> ή τα <<συστέλλει>>, χωρίς να τους αλλάζει τη διεύθυνση), δηλαδή:

$$A\vec{x} = \lambda\vec{x} \tag{38}$$

Τα διανύσματα αυτά, $\vec{x}$, για τα οποία ισχύει η Εξίσωση (38) λέγονται ι*διοδιανύσματα* (eigenvectors) του πίνακα *A*, και οι αριθμοί *λ*, οι οποίοι στην ουσία ισοδυναμούν με τον *A* όσον αφορά τη δράση πάνω στα ιδιοδιανύσματα, λέγονται *ιδιοτιμές* (eigenvalues)του *A*. Μπορούμε να πούμε ότι ο *A* πάνω στα ιδιοδιανύσματα εκφυλίζεται πάνω στον αριθμό *λ* ή ότι ο αριθμός *λ* πάνω στο ιδιοδιάνυσμα αντικαθιστά τη δράση του *A*.

Για κάθε ιδιοδιάνυσμα υπάρχει μία ιδιοτιμή, που ικανοποιεί την Εξίσωση (38). Τα ιδιοδιανύσματα όπως ορίζονται στη σχέση αυτή, ορίζονται με απροσδιοριστία μιας πολλαπλασιαστικής σταθεράς. Δηλαδή αν το $\vec{x}$ είναι ιδιοδιάνυσμα με ιδιοτιμή *λ*, τότε και το $\kappa\vec{x}$,όπου κ αριθμός, είναι επίσης ιδιοδιάνυσμα, με την ίδια ιδιοτιμή. Θα μπορούσαμε να πούμε δηλαδή ότι η σχέση ορισμού των ιδιοδιανυσμάτων, ορίζει μόνο τη διεύθυνση των ιδιοδιανυσμάτων και όχι το μέτρο τους.

Η εύρεση των ιδιοδιανυσμάτων και των ιδιοτιμών ενός πίνακα λέγεται πρόβλημα ιδιοτιμής ή ιδιοτιμών. Έστω το πρόβλημα ιδιοτιμής:

$$A\vec{x} = \lambda\vec{x} \tag{39}$$

όπου *A* είναι ένας γνωστός τετραγωνικός πίνακας *nxn*, και ζητούνται οι αριθμοί *λ* και τα διανύσματα $\vec{x}$ (μη μηδενικά). Το σύστημα της Εξίσωσης (39), που μπορεί να γραφτεί στη μορφή $A\vec{x} = \lambda I\vec{x}$ ή ισοδύναμα $(A - \lambda I)\vec{x} = 0$ επιλύεται με την εύρεση των ριζών του χαρακτηριστικού πολυωνύμου $\varphi_A(\lambda)$ του *A*:

$$\varphi_A(\lambda) = \det[A - \lambda I] \tag{40}$$

όπου *det* είναι η ορίζουσα (determinant) ενός πίνακα και *I* είναι ο μοναδιαίος πίνακας (identity matrix), καθώς είναι γνωστό ότι για να έχει το ομογενές σύστημα μη τετριμμένες λύσεις θα πρέπει $\det[A - \lambda I] = 0$. Για τον υπολογισμό του ιδιοδιανύσματος που αντιστοιχεί σε κάθε ιδιοτιμή θέτουμε στην Εξίσωση (39) τη συγκεκριμένη ιδιοτιμή και λύνουμε το ομογενές σύστημα που προκύπτει.

4.4 Ανάλυση κύριων συνιστωσών

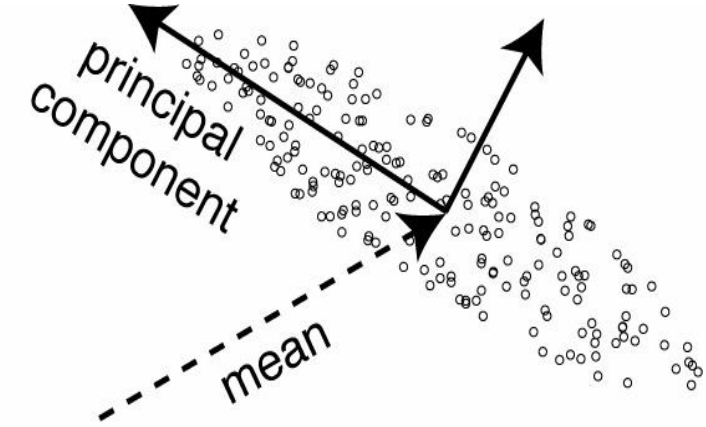
4.4.1 Εισαγωγή

Η *Ανάλυση Κύριων/Πρωτευουσών Συνιστωσών* (Principal Components Analysis-PCA) , Εικόνα 16, αποτελεί μία γραμμική τεχνική μείωσης διαστάσεων, που σημαίνει ότι εκτελεί ελάττωση αυτών με την ενσωμάτωση των δεδομένων σε ένα γραμμικό υποχώρο χαμηλής διάστασης [86]. Στόχος είναι η αύξηση της διασποράς της προβάλλουσας πληροφορίας στη χαμηλή διάσταση και ελαχιστοποίηση του σφάλματος προβολής. Αν και υπάρχουν διάφορες τεχνικές που κάνουν αυτό, η μέθοδος που αναλύουμε, είναι μακράν η πιο δημοφιλής (μη επιβλέπουσα) γραμμική τεχνική.

Η ανάλυση κυρίων συνιστωσών είναι μια μαθηματική διαδικασία που χρησιμοποιεί έναν ορθογώνιο μετασχηματισμό για να μετατρέψει ένα σύνολο παρατηρήσεων από πιθανές συσχετισμένες μεταβλητές σε ένα σύνολο τιμών από γραμμικά ασυσχέτιστες τιμές που αποκαλούνται *κύριες συνιστώσες* (principal components). Ο αριθμός των πρωτευουσών μεταβλητών είναι μικρότερος ή ίσος του αριθμού των αρχικών μεταβλητών. Ο μετασχηματισμός αυτός ορίζεται κατά τέτοιο τρόπο ώστε η πρώτη τέτοια συνιστώσα να έχει τη μεγαλύτερη δυνατή διασπορά και κάθε διαδεχόμενη συνιστώσα να έχει με τη σειρά της τη μεγαλύτερη δυνατή διακύμανση, υπό τον περιορισμό να είναι ορθογώνια (ή ασυσχέτιστη) με τα στοιχεία που προηγούνται. Το PCA παρουσιάζει ευαισθησία στη σχετική κλιμάκωση των αρχικών μεταβλητών.

Το PCA εφευρέθηκε το 1901 από τον Karl Pearson [96]. Χρησιμοποιείται κυρίως ως εργαλείο για την ερευνητική ανάλυση δεδομένων και για τη κατασκευή μοντέλων πρόβλεψης. Η τεχνική αυτή μπορεί να γίνει με τη διάσπαση ιδιοτιμής ενός πίνακα

συνδιακύμανσης μιας βάσης δεδομένων, μετά συνήθως από το κεντράρισμα στη μέση τιμή (εξομάλυνση) του πίνακα δεδομένων για κάθε χαρακτηριστικό [97].



Εικόνα 16: Κύρια συνιστώσα σε δύο διαστάσεις.

4.4.2 Βήματα μεθοδολογίας

Στο σημείο αυτό θα γίνει αναφορά στα επακριβή βήματα που εφαρμόζονται σε μια βάση δεδομένων, για να υπολογιστούν οι κύριες συνιστώσες αυτής ώστε να έχουμε την τελική προβολή αυτών σε δύο-τρεις διαστάσεις [98].

Βήμα 1: Επιλογή δεδομένων

Το αρχικό στάδιο υλοποίησης του PCA περιλαμβάνει την επιλογή των δεδομένων, πάνω στα οποία θα εφαρμόσουμε τη μέθοδο. Τα δεδομένα αυτά βρίσκονται τοποθετημένα σε ένα πίνακα, η μία διάσταση του οποίου αποτελεί τα στοιχεία του συνόλου (column) και η άλλη (row) το μέγεθος της διάστασης των μεταβλητών.

Ας ορίσουμε ένα σύνολο *N* δεδομένων $x_i, i \in 1,...,N$, με $x_i \in \mathbb{R}^n$.

Βήμα 2: Αφαίρεση της μέσης τιμής

Στο βήμα αυτό δημιουργείται μία προσαρμοσμένη βάση δεδομένων *A* (*nxN*) (43) (adjusted data set). Για να επιτύχει το PCA αυτήν την κατασκευή, αφαιρεί από κάθε αντικείμενο (θέση πίνακα (n ,N)) την μέση τιμή (42), που υπολογίζεται από τον μέσο όρο των τιμών στην *n* διάσταση (41). Οι μέσες τιμές των διαστάσεων του προσαρμοσμένου πίνακα είναι μηδέν.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (41)$$

$$\hat{\mathrm{x}}_i = x_i - \overline{x} \tag{42}$$

$$A = [\hat{\mathrm{x}}_1\ \hat{\mathrm{x}}_2 ... \hat{\mathrm{x}}_N] \tag{43}$$

Βήμα 3: Υπολογισμός πίνακα συνδιακύμανσης

Από το προηγούμενο βήμα έχουμε λάβει ένα προσαρμοσμένο πίνακα δεδομένων *A*, διαστάσεων ίδιων με την αρχική βάση δεδομένων. Υπολογίζουμε στο επόμενο στάδιο τον πίνακα συνδιακύμανσής του, *C* (44) και έτσι παράγεται ένας πίνακας διαστάσεων *NxN* , όπου *N* είναι η διάσταση των αρχικών δεδομένων. Πρόκειται για ένα τετραγωνικό πίνακα, ο οποίος παρουσιάζει συμμετρία ως προς την κύρια διαγώνιό του. Σημειώνεται ότι η μέση τιμή των διαστάσεων του πίνακα *A* είναι 0 [99].

$$C = \frac{1}{N} AA^{\mathrm{T}} \tag{44}$$

$$C = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathrm{x}}_i \hat{\mathrm{x}}_i^{\mathrm{T}} \tag{45}$$

Βήμα 4: Υπολογισμός των ιδιοδιανυσμάτων και ιδιοτιμών του πίνακα συνδιακύμανσης

Αφού ο πίνακας συνδιακύμανσης που προέκυψε από το προηγούμενο βήμα είναι τετραγωνικός, μπορούν να υπολογιστούν τα ιδιοδιανύσματα και οι ιδιοτιμές του εν λόγω πίνακα σύμφωνα με τη διαδικασία που αναλύθηκε στην Ενότητα 4.3.2. Αυτές οι ποσότητες παρουσιάζουν μεγάλο ενδιαφέρον όσο και σπουδαιότητα, καθώς παράγουν χρήσιμη πληροφορία σχετικά με τα δεδομένα.

Έτσι, ο PCA διαγωνοποιεί τον πίνακα συνδιασποράς μέσω της επίλυσης του προβλήματος των ιδιοτιμών:

$$C\mathrm{v} = \lambda \mathrm{v} \tag{46}$$

υπολογίζοντας :

- Τις ιδιοτιμές $\lambda_1, \lambda_2, ..., \lambda_n \quad \mu\varepsilon \quad \lambda \geq 0$ (47)
- Τα ιδιοδιανύσματα $\mathrm{v}_1, \mathrm{v}_2, ..., \mathrm{v}_n \quad \mu\varepsilon \quad \mathrm{v} \in \mathbb{R}^n \setminus \{0\}$ (48)

Βήμα 5: Επιλογή συνιστωσών και σχηματισμός ενός διανύσματος χαρακτηριστικών γνωρισμάτων

Στο σημείο αυτό, η έννοια της συμπίεσης δεδομένων και της μείωσης διαστάσεων εμφανίζονται. Στο βήμα αυτό μελετούμε τις ιδιοτιμές και τα ιδιοδιανύσματα που έχουν υπολογισθεί. Στην πραγματικότητα αποδεικνύεται ότι το ιδιοδιάνυσμα με τη μεγαλύτερη ιδιοτιμή αποτελεί την κύρια συνιστώσα του συνόλου δεδομένων.

Σε γενικές γραμμές, μιας και τα ιδιοδιανύσματα προκύπτουν από τον πίνακα συνδιακύμανσης, το επόμενο βήμα είναι να ταξινομηθούν αυτά σύμφωνα με τις ιδιοτιμές, από την υψηλότερη στη χαμηλότερη (οι σχέσεις 48 και 49 παρουσιάζουν τα ταξινομημένα πλέον ιδιοδιανύσματα σύμφωνα με τις ιδιοτιμές). Αυτή η διαδικασία παρουσιάζει τις συνιστώσες σε *σειρά σπουδαιότητας* (order of significant).

$$\lambda_1 > \lambda_2 > .... > \lambda_n \tag{49}$$

$$\mathrm{v}_1 > \mathrm{v}_2 > .... > \mathrm{v}_n \tag{50}$$

Δεδομένου ότι ο πίνακας συνδιασποράς *C* είναι συμμετρικός, τα v_1, v_2, …, v_n σχηματίζουν μια βάση (κάθε διάνυσμα *x* ή συγκεκριμένα ($x_i - \bar{x}$) , μπορεί να γραφτεί σαν ένας γραμμικός συνδυασμός των ιδιοδιανυσμάτων):

$$x - \bar{x} = b_1\mathrm{v}_1 + b_2\mathrm{v}_2 ... + b_n\mathrm{v}_n = \sum_{i=1}^{n} b_i\mathrm{v}_i \quad \text{όπου} \quad b_i = \mathrm{v}_i^T(\hat{\mathrm{x}}) \tag{51}$$

Στο σημείο αυτό, δίνεται η δυνατότητα σε όποιον επιθυμεί να αγνοήσει τα στοιχεία με χαμηλή σπουδαιότητα. Με την ενέργεια αυτή θα χαθεί πληροφορία, αλλά αν οι ιδιοτιμές είναι μικρές, η απώλειά της θα είναι ασήμαντη. Αν απαλειφθούν συνιστώσες, η τελική βάση δεδομένων θα έχει λιγότερες διαστάσεις από την αρχική. Προς αποφυγή ασάφειας, αν η ακριβής διάσταση των δεδομένων είναι *n*, και υπολογίζονται *n* ιδιοδιανύσματα και ιδιοτιμές και επιλέγονται μόνο τα *k* πρώτα ιδιοδιανύσματα, τότε το τελικό σύνολο δεδομένων θα έχει *k* διάσταση:

$$x - \bar{x} = \sum_{i=1}^{k} b_i\mathrm{v}_i \quad \text{όπου k<<n} \tag{52}$$

Υπολείπεται, ο σχηματισμός ενός *διανύσματος χαρακτηριστικών* (feature vector), ο οποίος αποτελεί ένα φανταστικό όνομα ενός πίνακα διανυσμάτων. Αυτό δημιουργείται παίρνοντας τα ιδιοδιανύσματα που θέλουμε να διατηρήσουμε από την λίστα των ιδιοδιανυσμάτων και κατασκευάζοντας ένα πίνακα με τα επιλεχθέντα ιδιοδιανύσματα, στις στήλες του.

$$\mathrm{v}^k = (\mathrm{v}_1, \mathrm{v}_2, \mathrm{v}_3, \ldots, \mathrm{v}_k) \tag{53}$$

Βήμα 6: Εξαγωγή της νέας βάσης δεδομένων

Αυτό είναι το τελευταίο βήμα της γραμμικής τεχνικής μείωσης διαστάσεων, PCA. Καθώς έχουν επιλεγεί οι συνιστώσες (ιδιοδιανύσματα), εκείνες οι οποίες επιθυμούνται να διατηρηθούν στα δεδομένα και κατασκεύασαν ένα διάνυσμα χαρακτηριστικών, υπολογίζεται ο ανάστροφος του διανύσματος αυτού και πολλαπλασιάζεται από αριστερά με την ανάστροφη αρχική βάση δεδομένων.

$$\hat{\mathbf{x}}_{pc}^{k} = \mathbf{v}^k \cdot \hat{\mathbf{x}} \tag{54}$$

όπου $\mathbf{v}^k$ είναι ο πίνακας με τα ιδιοδιανύσματα στις στήλες, αναστραμμένος, έτσι τα ιδιοδιανύσματα βρίσκονται πλέον στις γραμμές, με τα πιο σημαντικά στην κορυφή, και όπου $\hat{\mathbf{x}}$ είναι μια συνιστώσα της προσαρμοσμένης βάση δεδομένων A . Στον πίνακα $\hat{\mathbf{x}}_{pc}^{k}$ τα στοιχεία των δεδομένων βρίσκονται σε κάθε στήλη, ενώ κάθε γραμμή εμπεριέχει μια ξεχωριστή διάσταση.

4.5 Ανάλυση σε κύριες συνιστώσες με χρήση πυρήνα

4.5.1 Συναρτήσεις πυρήνα για εφαρμογές Μηχανικής Μάθησης

Τα τελευταία χρόνια οι *μέθοδοι πυρήνα* (kernel functions) έχουν λάβει μεγάλη προσοχή, ιδιαίτερα λόγω της αυξημένης δημοτικότητας που παρουσιάζουν οι μηχανές υποστήριξης διανυσμάτων(SVM) [100]. Οι συναρτήσεις πυρήνα μπορούν να χρησιμοποιηθούν σε πολλές εφαρμογές καθώς παρέχουν μια απλή γέφυρα από τη γραμμικότητα στη μη γραμμικότητα για τους αλγορίθμους που μπορούν να εκφραστούν με όρους εσωτερικού γινομένου.

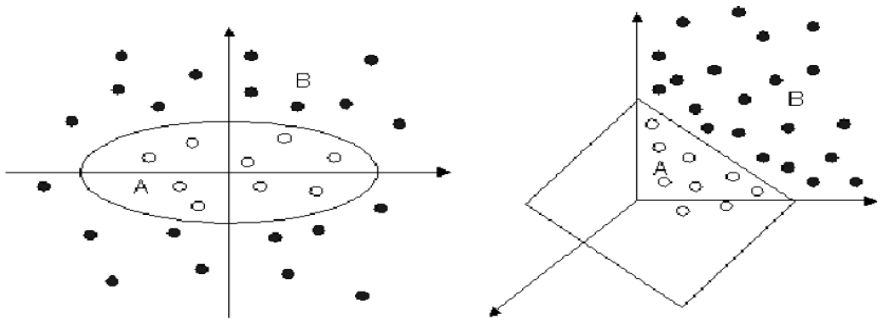
Μέθοδοι πυρήνα

Οι *μέθοδοι πυρήνα* (kernel methods) είναι μια κατηγορία αλγορίθμων για την ανάλυση ή αναγνώριση προτύπων. Η γενική αρμοδιότητα της ανάλυσης προτύπων είναι να βρει και να μελετήσει γενικούς τύπους σχέσεων (όπως ομάδες, ταξινομήσεις, συσχετίσεις, κύριες συνιστώσες) σε διάφορους τύπους δεδομένων (όπως διανύσματα, γραφήματα, σύνολα σημείων, εικόνες).

Το κύριο χαρακτηριστικό των μεθόδων πυρήνα, όμως, είναι η διαφορετική προσέγγιση σε αυτό το πρόβλημα. Οι μέθοδοι αυτές χαρτογραφούν τα δεδομένα σε χώρους υψηλότερης διάστασης με την ελπίδα, ότι τα δεδομένα σε αυτό τον χώρο θα μπορούν να γίνουν πιο εύκολα διαχωρίσιμα ή να είναι πιο καλά δομημένα (Εικόνα 17). Επίσης δεν υπάρχουν περιορισμοί για τη μορφή αυτής της χαρτογράφησης, η οποία θα μπορούσε να οδηγήσει ακόμη σε χώρους άπειρης διάστασης. Αυτή η λειτουργία χαρτογράφησης, ελάχιστα θα πρέπει να υπολογιστεί, λόγω της ύπαρξης ενός εργαλείου που ονομάζεται *Τέχνασμα του Πυρήνα* (kernel trick) [100].

Kernel Trick

Το τέχνασμα του πυρήνα αποτελεί ένα μαθηματικό εργαλείο που μπορεί να εφαρμοστεί σε οποιοδήποτε αλγόριθμο που εξαρτάται απλώς από το εσωτερικό γινόμενο δύο διανυσμάτων. Όταν ένα εσωτερικό γινόμενο χρησιμοποιείται, αυτό αντικαθίσταται από μία συνάρτηση πυρήνα. Όταν εφαρμόζεται σωστά, οι εν λόγω



Εικόνα 17: Κατηγοριοποίηση προτύπων σε δύο κλάσεις [101]. Αριστερά: Δείγματα στον δισδιάστατο χώρο εισόδου, όπου απαιτείται ένα μη γραμμικό ελλειψοειδές όριο απόφασης για τον διαχωρισμό των κλάσεων Α και Β. Δεξιά: Προβολή των δειγμάτων σε τρισδιάστατο χώρο όπου ένα γραμμικό υπέρ-επίπεδο μπορεί να διαχωρίσει τις κλάσεις.

υποψήφιοι γραμμικοί αλγόριθμοι μετατρέπονται σε μη γραμμικούς (μερικές φορές με λίγη προσπάθεια ή αναδιατύπωση). Οι μη γραμμικοί αλγόριθμοι είναι ισοδύναμοι με τα γραμμικά πρωτότυπά τους λειτουργώντας στο χώρο της εμβέλειας ενός χώρου χαρακτηριστικών φ. Επειδή πυρήνες χρησιμοποιούνται, η λειτουργία φ δεν χρειάζεται να υπολογιστεί ποτέ ρητά. Αυτό είναι ιδιαίτερα επιθυμητό, γιατί όπως σημειώθηκε προηγουμένως, ο χώρος των χαρακτηριστικών υψηλών διαστάσεων μπορεί να γίνει άπειρων διαστάσεων και έτσι θα είναι ανέφικτο να υπολογιστεί. Δεν υπάρχουν επίσης περιορισμοί στη φύση των διανυσμάτων εισόδου. Εσωτερικό γινόμενο θα μπορούσε να οριστεί μεταξύ οποιοδήποτε είδους δομής όπως δέντρα ή συμβολοσειρές.

Συναρτήσεις πυρήνα

Ακολουθεί η περιγραφή ορισμένων συναρτήσεων πυρήνα, από το σύνολο που υπάρχουν στη βιβλιογραφία, και οι οποίοι έλαβαν χώρα στη συγκεκριμένη διπλωματική εργασία.

- *Linear Kernel*: Αποτελεί την απλούστερη συνάρτηση πυρήνα. Αυτή δίνεται από το εσωτερικό γινόμενο <*x,y*> προσαυξημένο μιας προαιρετικής σταθεράς *c*. Αλγόριθμοι πυρήνα που χρησιμοποιούν linear kernel συχνά ισοδυναμούν με τους non-kernel ομολόγους τους. Για παράδειγμα ο k-PCA με linear kernel είναι ακριβώς ο ίδιος με τον τυπικό PCA.

$$k(x, y) = x^T y + c \tag{55}$$

- *Polynomial Kernel*: Ο πολυωνυμικός πυρήνας αποτελεί έναν μη στάσιμο πυρήνα. Η κατηγορία αυτή είναι κατάλληλη για προβλήματα όπου όλα τα δεδομένα εκπαίδευσης είναι κανονικοποιημένα.

$$k(x, y) = (ax^T y + c)^d \tag{56}$$

Ρυθμιζόμενοι παράγοντες είναι η κλίση α, ο σταθερός όρος *c* και ο βαθμός του πολυωνύμου *d*.

- *Gaussian Kernel*: Αποτελεί ένα παράδειγμα ακτινικής βάσης συνάρτηση πυρήνα.

$$k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{57}$$

Εναλλακτικά θα μπορούσε να υλοποιηθεί με τη χρήση :

$$k(x, y) = \exp\left(-\gamma\|x - y\|^2\right) \tag{58}$$

Η προσαρμόσιμη παράμετρος *σ*, διαδραματίζει σημαντικό ρόλο στην απόδοση του πυρήνα, και πρέπει να συντονιστεί με προσοχή το πρόβλημα στο χέρι. Αν είναι υπερτιμημένο, το εκθετικό θα συμπεριφέρεται σχεδόν γραμμικά και η υψηλή διαστάσεων προβολή θα αρχίσει να χάνει τη μη γραμμική δύναμή της. Από την άλλη πλευρά, αν υποτιμηθεί, η συνάρτηση θα στερείται διευθέτησης και το όριο απόφασης θα είναι πολύ ευαίσθητο στο θόρυβο στα δεδομένα εκπαίδευσης.

4.5.2 Υλοποίηση αλγορίθμου k-PCA

Δοθέντος ενός συνόλου από *N* κεντρικά (centered) δείγματα $x_i = [x_1, x_2, \dots, x_N] \in \mathbb{R}^n$, ο αλγόριθμος PCA στοχεύει στη εύρεση εκείνων των διευθύνσεων προβολής που μεγιστοποιούν τη διακύμανση C, των x_i κάτι που είναι ισοδύναμο με την εύρεση των ιδιοτιμών του πίνακα συνδιακύμανσης των δειγμάτων [102, 103]:

$$\begin{aligned} \lambda \mathbf{v} &= C\mathbf{v} \\ \|\mathbf{v}\|_1 &= 1 \end{aligned} \tag{59}$$

Όπου λ≥0 οι ιδιοτιμές και $\mathbf{v} \in \mathbb{R}^n$ τα ιδιοδιανύσματα.

Στην ανάλυση Κύριων Συνιστωσών του πυρήνα, κάθε διάνυσμα *x* προβάλλεται από τον αρχικό χώρο, σε έναν χώρο χαρακτηριστικών υψηλής τάξης, F, με τη βοήθεια μίας μη γραμμική συνάρτησης αντιστοίχισης (mapping function) ϕ :

$$\begin{aligned} \boldsymbol{\phi} : \mathbb{R}^n &\rightarrow \mathrm{F} \\ \mathbf{x} &\rightarrow \boldsymbol{\phi}(\mathbf{x}) \end{aligned} \tag{60}$$

όπου ϕ είναι μια μη γραμμική συνάρτηση και F μπορεί να έχει μια άπειρη διάσταση.

Το PCA μπορεί να εκτελεστεί στο χώρο χαρακτηριστικών υψηλής διάστασης F με την ίδια διαδικασία όπως προηγουμένως: τα δεδομένα κεντράρονται και ο πίνακας συνδιασποράς ορίζεται ως εξής:

$$C_{\boldsymbol{\phi}(\mathbf{x})} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\phi}(\mathbf{x}_i) \boldsymbol{\phi}(\mathbf{x}_i)^T \tag{61}$$

Όμοια με το PCA, πρέπει να επιλυθεί:

$$\begin{aligned}\lambda \mathbf{v}_{\phi} &= C_{\phi(\mathbf{x})}\mathbf{v}_{\phi} = \left(\frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^T\right)\mathbf{v}_{\phi} \\ &= \frac{1}{N}\sum_{i=1}^{N}(\phi(\mathbf{x}_i)\cdot\mathbf{v}_{\phi})\phi(\mathbf{x}_i)\end{aligned} \tag{62}$$

Από τη σχέση (62) γίνεται σαφές ότι τα V_{ϕ} εξαπλώνονται στην έκταση των $\phi(\mathbf{x}_1)...\phi(\mathbf{x}_N)$, έτσι κάθε ιδιοδιάνυσμα μπορεί να γραφτεί ως εξής:

$$\mathbf{v}_{\phi} = \sum_{i=1}^{N} a_i\phi(\mathbf{x}_i) \tag{63}$$

όπου a_i $(i = 1,...,N)$ συντελεστές.

Πολλαπλασιάζοντας την (62) με το $\phi(\mathbf{x}_k)$ από τα αριστερά και αντικαθιστώντας την (63) σε αυτό, παίρνουμε:

$$\begin{aligned}&\lambda\sum_{i=1}^{N}a_i(\phi(\mathbf{x}_k)\phi(\mathbf{x}_i)) = \\ &\frac{1}{N}\sum_{i=1}^{N}a_i\left(\phi(\mathbf{x}_k)\cdot\sum_{j=1}^{N}\left(\phi(\mathbf{x}_j)\cdot\phi(\mathbf{x}_i)\right)\phi(\mathbf{x}_j)\right)\end{aligned} \tag{64}$$

για $k \in [1, N]$.

Ορίζοντας τον $N \times N$ πίνακα K με:

$$K_{i,j} := \left(\phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j)\right) \tag{65}$$

η παραπάνω εξίσωση γίνεται:

$$\lambda K a = \frac{1}{N}K^2 a \tag{66}$$

όπου a ορίζει το διάνυσμα στήλη με καταχωρήσεις $a_1,...,a_N$ και K είναι ένας συμμετρικός θετικά ημιορισμένος πίνακας. Η λύση της (66) έρχεται από τη επίλυση του προβλήματος των ιδιοτιμών:

$$N\lambda a = Ka \tag{67}$$

για μη μηδενικές ιδιοτιμές. Προφανώς, όλες οι λύσεις της (67) πληρούν και την (66). Ωστόσο, αυτό δεν δίνει όλες τις λύσεις, καθώς ιδιοδιανύσματα που συσχετίζονται με μηδενικές ιδιοτιμές αποτελούν λύση για την (66) που δεν είναι όμως λύση για την (67). Αλλά, μπορεί να αποδειχθεί ότι αυτές οι λύσεις οδηγούν σε κενή επέκταση της (63) και ως εκ τούτου είναι άνευ σημασίας για το εξεταζόμενο πρόβλημα. Τελικά, η επίλυση της εξίσωσης ιδιοτιμών του $C_{\phi(\mathbf{x})}$ ισοδυναμεί με την επίλυση της εξίσωσης ιδιοτιμών του πίνακα K.

Η μοναδιαία νόρμα που αποτελεί συνθήκη της (59) μεταφράζεται στο χώρο χαρακτηριστικών υψηλών διαστάσεων F σε $\lambda_k(a^k a^k)=1$ [104]. Η προβολή στο χώρο F γίνεται απλά:

$$\boldsymbol{\phi}(\mathbf{x})^k_{kpc} = \mathrm{v}^k_{\boldsymbol{\phi}} \cdot \boldsymbol{\phi}(\mathbf{x}) = \sum_{i=1}^{N} a_i^k (\boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x})) \quad . \tag{68}$$

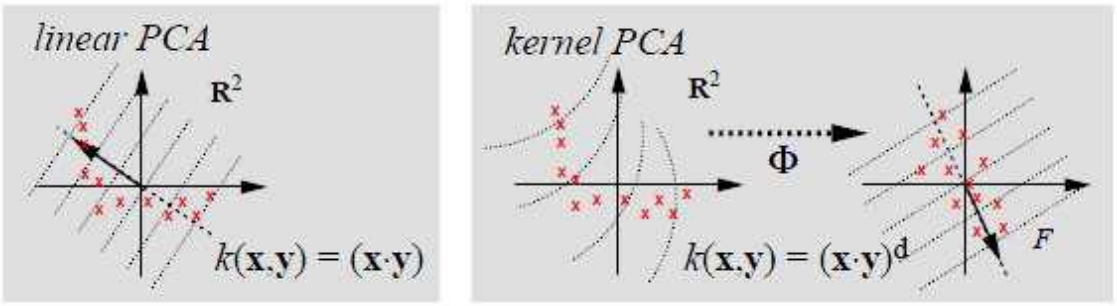
Ωστόσο, ο υπολογισμός του PCA στο F έχει μεγάλο υπολογιστικό κόστος. Το τέχνασμα του πυρήνα, είναι δυνατόν να λειτουργήσει έμμεσα στο F καθώς όλοι οι υπολογισμοί γίνονται στο χώρο εισόδου. Χρησιμοποιώντας συνάρτηση πυρήνα, το εσωτερικό γινόμενο στο χώρο των χαρακτηριστικών μειώνεται σε μια (πιθανά μη γραμμική) συνάρτηση στο χώρο εισόδου:

$$\boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}_j) = \mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) \, . \tag{69}$$

Έτσι αν αντικαταστήσουμε όλες τις εμφανίσεις του εσωτερικού γινομένου $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ με συναρτήσεις πυρήνα λαμβάνουμε τον αλγόριθμο Kernel PCA η κύρια ιδέα του οποίου φαίνεται στην Εικόνα 18 που ακολουθεί.

Η συνάρτηση πυρήνα πρέπει να πληροί τη θεωρία του Mercer, δηλαδή να εξασφαλίσει ότι είναι δυνατόν να κατασκευαστεί μια αντιστοίχιση σε ένα χώρο όπου το K λειτουργεί ως ένα εσωτερικό γινόμενο.

Τέλος, το k-PCA γίνεται στον αρχικό χώρο ως εξής:



Εικόνα 18: Βασική ιδέα του αλγορίθμου Kernel PCA. Με τη χρήση μια συνάρτησης πυρήνα, ο k-PCA υλοποιεί ένα γραμμικό PCA σε έναν πολυδιάστατο χώρο χαρακτηριστικών και σχετίζεται μη γραμμικά με τον αρχικό χώρο των δεδομένων. (Αριστερά): Ο γραμμικός PCA στον αρχικό χώρο δεν είναι ικανός να δώσει μια καλή περιγραφή για την κατεύθυνση των δεδομένων σε αυτό το απλό παράδειγμα. (Δεξιά): Με τη χρήση της κατάλληλης μη γραμμικής απεικόνισης Φ και εφαρμόζοντας τον γραμμικό PCA στα πρότυπα που έχουν προκύψει μετά την απεικόνιση (kernel PCA), η μη γραμμική κατεύθυνση στον αρχικό χώρο, η οποία είναι το αποτέλεσμα της εφαρμογής του αλγορίθμου, είναι σε θέση να μας πληροφορήσει για την κατεύθυνση που μας ενδιαφέρει.

Υπολογισμός του πίνακα πυρήνα (Kernel Matrix): $K_{i,j} = \mathrm{K}(\mathbf{x}_i, \mathbf{x}_j)$.

1. Κεντράρισμα του K [102,104]:

$$K_c = K - 1_N K - K 1_N + 1_N K 1_N$$

όπου 1_N είναι ένας τετραγωνικός πίνακας για τον οποίο $(1_N)_{i,j} = \frac{1}{N}$, για όλα τα (i, j) στο $[1,\ldots,N]$.

2. Ο K_c γίνεται διαγωνοποιήσιμος και κανονικοποιούνται τα ιδιοδιανύσματα:

$$\lambda_k (a^k a^k) = 1 .$$

3. Εξαγωγή των k κύριων συνιστωσών:

$$\boldsymbol{\phi}(\mathbf{x})^k_{kpc} = \sum_{i=1}^{N} a_i^k (\boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x})) .$$

Κεφάλαιο 5

Προτεινόμενη Μεθοδολογία και Αποτελέσματα

5.1 Ανάπτυξη μεθοδολογίας

Οι μέθοδοι ομαδοποίησης, οι δείκτες αξιολόγησης αυτών καθώς και οι τεχνικές οπτικοποίησης συνδυάζονται και προσαρμόζονται στο κεφάλαιο αυτό, με στόχο την βέλτιστη εφαρμογή τους και την λήψη χρήσιμων αποτελεσμάτων σε διαφορετικούς τύπους δεδομένων.

Ακολουθείται μια ροή συγκεκριμένων βημάτων, η οποία ξεκινάει με την ομαδοποίηση των πολυδιάστατων δεδομένων μας με διαφορετικές προσεγγίσεις. Οι προσεγγίσεις αυτές περιλαμβάνουν συνδυασμούς των τεχνικών ομαδοποίησης που αναλύθηκαν σε προηγούμενο κεφάλαιο, προς ενίσχυση του ομαδοποιημένου αποτελέσματος. Στη συνέχεια, για κάθε μια από τις διαφορετικές ομαδοποιήσεις που προκύπτουν (για το ίδιο πρόβλημα), κρίνεται η ποιότητα κάθε αποτελέσματος με χρήση των δεικτών εγκυρότητας, αλλά και τις μεταξύ τους σχέσεις. Το τελικό βήμα της διπλωματικής εργασίας συνιστά η οπτικοποίηση των ομαδοποιημένων πλέον δεδομένων σε δύο διαστάσεις, με μία ποικιλία μεθόδων. Η εν λόγω απεικόνιση, μας δίνει τη δυνατότητα πρώτον να παρατηρήσουμε αν η ομαδοποίηση των αρχικών δεδομένων μπορεί να γίνει <<ορατή>> σε μικρές διαστάσεις και δεύτερον να αποφανθούμε αν οι τιμές των δεικτών αξιολόγησης συνάδουν με το οπτικό αποτέλεσμα. Αντικείμενο μελέτης αποτελεί και η απόδοση των διαφορετικών τεχνικών οπτικοποίησης που αναπτύσσονται.

Στο σημείο αυτό παρουσιάζονται τα βήματα που ακολουθούνται για τη διαδικασία της ομαδοποίησης. Υπάρχουν δύο κατευθύνσεις που ακολουθούνται:

- Η πρώτη αποτελεί μία τροποποίηση του Αυτό- Οργανούμενου χάρτη (SOM). Αρχικά εκτελεί εκπαίδευση SOM στα αρχικά δείγματα, εν συνεχεία εφαρμόζει k-means ομαδοποίηση στους τελικούς ανανεωμένους κόμβους του SOM, για διάφορους πιθανούς αριθμούς ομάδων, επιλέγοντας τελικά τον αριθμό των k-ομάδων στον οποίο ο δείκτης Silhouette παρουσιάζει τη βέλτιστη τιμή. Η διαδικασία ολοκληρώνεται υπολογίζοντας τις αποστάσεις των αρχικών

δειγμάτων από τα κέντρα που επιστρέφει ο k-means και με κριτήριο την ελάχιστη απόσταση κατηγοριοποιούνται τα αρχικά μας δείγματα σε ομάδες (Πίνακας 3).

- Η δεύτερη ακολουθεί αντίστοιχη πορεία, έχοντας ως αφετηρία όμως την LVQ τεχνική (Πίνακας 4).

Πίνακας 3: Βήματα τροποποιημένης ομαδοποίησης SOM.

Βήμα 1: Αυτό- Οργανούμενοι Χάρτες απεικόνισης Το πρώτο βήμα εξάγει έναν αριθμό ομάδων βασισμένο στην SOM οργάνωση των δεδομένων εισόδου σε $Q \times Q$ κόμβους (Ενότητα 2.3.5). *Βήμα 2:* K-means ομαδοποίηση αυτό- οργανούμενου χάρτη απεικόνισης Τα κομβικά-βάρη πρότυπα, πλήθους $Q \times Q$, οργανώνονται σε *K* ομάδες με βάση την επαναληπτική προσέγγιση του k-means (Ενότητα 2.5). *Βήμα 3*: Ανάθεση των δεδομένων εισόδου σε *K* ομάδες Η ανάθεση αυτή γίνεται με βάση την ελάχιστη l_2 νόρμας της διαφοράς μεταξύ κάθε δείγματος εισόδου και του κέντρου κάθε ομάδας.

Πίνακας 4: Βήματα τροποποιημένης ομαδοποίησης LVQ.

Βήμα 1: Διανυσματική εκπαιδευόμενη μάθηση Το πρώτο βήμα εξάγει έναν αριθμό ομάδων βασισμένο στην LVQ οργάνωση των δεδομένων εισόδου σε $Q \times Q$ κόμβους (Ενότητα 2.4.4). *Βήμα 2:* K-means ομαδοποίηση του LVQ Τα κομβικά-βάρη πρότυπα, πλήθους $Q \times Q$, οργανώνονται σε K ομάδες με βάση την επαναληπτική προσέγγιση του k-means (Ενότητα 2.5). *Βήμα 3*: Ανάθεση των δεδομένων εισόδου σε K ομάδες Η ανάθεση αυτή γίνεται με βάση την ελάχιστη l_2 νόρμας της διαφοράς μεταξύ κάθε δείγματος εισόδου και του κέντρου κάθε ομάδας.

Παρακινούμενοι από την ομαδοποίηση των κόμβων ενός SOM δικτύου με τη χρήση μιας πιθανής συνάρτησης ευρωστίας (robust potential function) [50], επεκταθήκαμε και αναπτύξαμε την προτεινόμενη μεθοδολογία, στην οποία οι έξοδοι του δικτύου υφίστανται k-means ομαδοποίηση. Για το λόγο αυτό αναφερόμαστε σε τροποποιημένες SOM και LVQ μεθόδους.

Αφού ολοκληρωθεί η ομαδοποίηση, το επόμενο βήμα είναι να υπολογιστούν οι δείκτες Dunn Index, Silhouette και Davies Bouldin (Ενότητα 2.6.2) για να γίνει μία γενική εκτίμηση της ποιότητας της ομαδοποίησης. Για κάθε λοιπόν βάση δεδομένων καταλήγουμε σε δύο διαφορετικές ομαδοποιήσεις. Βασιζόμενοι σε αυτές, και με χρήση των μεθόδων οπτικοποίησης (Κεφάλαιο 4) καταλήγουμε στην τελική απεικόνιση των ομάδων που προκύπτουν σε δύο διαστάσεις. Ένα συνοπτικό διάγραμμα όλων των βημάτων που ακολουθούνται παρουσιάζεται στην Εικόνα 19.

Η μεθοδολογία της παρούσας διπλωματικής εργασίας εφαρμόζεται σε τρεις διαφορετικές βάσεις δεδομένων. Η πρώτη αποτελεί προϊόν ενός τεχνητού προβλήματος που αναπτύσσουμε, έχοντας ως απώτερο στόχο την απόδειξη της ορθής λογικής και ευσταθούς λειτουργίας των τεχνικών που υλοποιούμε. Η δεύτερη βάση δεδομένων εμπεριέχει γονίδια του σακχαρομύκητα *Saccharomyces cerevisiae* [9] ενώ η τρίτη αναφέρεται στα *χονδροκύτταρα* και στα *μεσεγχυματικά βλαστικά κύτταρα* [27], τα οποία είναι άρρηκτα δεμένα με την πάθηση της *οστεοαρθρίτιδας*. Την τελευταία βάση δεδομένων, που στην πραγματικότητα αποτελείται από δύο μέρη, χρησιμοποιούμε για να εξετάσουμε την ομαδοποίηση με βάση τα μέτρα χρονικής εξέλιξης (Κεφάλαιο 3).

Τα ακριβή βήματα που ακολουθούνται σε όλες τις περιπτώσεις, τα αποτελέσματα και τα συμπεράσματα που προκύπτουν από κάθε στάδιο εφαρμογής παρατίθενται στις επόμενες ενότητες. Εκτός από τα στατιστικά αποτελέσματα προχωρούμε (στην περίπτωση των δύο υπαρκτών βάσεων) και σε βιολογική ερμηνεία των αποτελεσμάτων, καθώς μια ικανοποιητική ομαδοποίηση γονιδιακών δεδομένων και μια επιτυχής οπτικοποίηση αυτών θα πρέπει να συνοδεύεται και από μία καλή βιολογική περιγραφή.

Εικόνα 19: Γενική ροή μεθοδολογίας.

5.2 Εφαρμογή σε Τεχνητό Πρόβλημα

Για την καλύτερη αξιοπιστία των μεθόδων που αναπτύσσονται στην παρούσα εργασία και την εγκυρότητα της απόδοσης αυτών, παράγουμε έναν τεχνητό πληθυσμό δεδομένων πάνω στον οποίο εφαρμόζεται και αξιολογείται αρχικά όλη η μεθοδολογία.

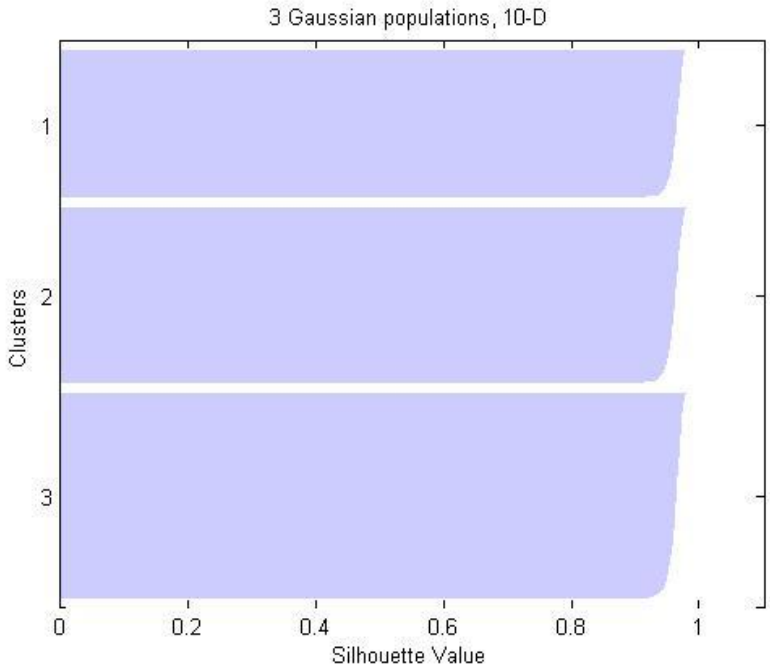
Συγκεκριμένα δημιουργούμε τρεις Gaussian πληθυσμούς δέκα διαστάσεων (10-D), ο καθένας με διαφορετικά mean vectors, τα οποία έχουν μεγάλη απόσταση μεταξύ τους. Ο πρώτος πληθυσμός αποτελείται από 500 δείγματα, ο δεύτερος από 600 ενώ ο τρίτος από 700 στοιχεία. Στην πραγματικότητα γνωρίζουμε ότι ο κάθε πληθυσμός

αποτελεί ένα συμπαγή σύνολο δεδομένων και ανά δύο οι ομάδες είναι επαρκώς διαχωρίσιμες (αρκετά μακριά η μία από την άλλη). Άρα, αναμένουμε τόσο η ομαδοποίηση των δεδομένων όσο και η οπτικοποίηση αυτών να μας δώσει πολύ ικανοποιητικά αποτελέσματα, εξαιτίας της φύσης των δεδομένων.

SOM & ΟΠΤΙΚΟΠΟΙΗΣΗ

ΔΙΑΔΙΚΑΣΙΑ ΟΜΑΔΟΠΟΙΗΣΗΣ

Μετά τη δημιουργία των τριών πληθυσμών, πριν προβούμε σε κάποια άλλη κίνηση, εκτελούμε το δείκτη εγκυρότητας Silhouette (Ενότητα 2.6.5). Παρουσιάζουμε γραφικά τις Silhouette τιμές που προκύπτουν από όλα τα αρχικά δεδομένα, καθώς γνωρίζουμε εκ των προτέρων το κάθε ένα από τα αρχικά δείγματα εισόδου που ανήκει. Ο μέσος όρος των τιμών αυτών υπολογίζεται *0.9596* και το διάγραμμα αυτών παραθέτεται στην Εικόνα 20. Ο οριζόντιος άξονας αναπαριστά τις Silhouette τιμές, ενώ ο κατακόρυφος δηλώνει τα στοιχεία κάθε ομάδας. Εύκολα παρατηρείται τόσο από το μέσο όρο των τιμών (ο οποίος πλησιάζει το βέλτιστο) όσο και από το διάγραμμα ότι πρόκειται για μια πολύ καλή ομαδοποίηση δεδομένων, με αυξημένη την πυκνότητα δεδομένων εσωτερικά των ομάδων αλλά και με έντονο διαχωρισμό των ομάδων μεταξύ τους.



Εικόνα 20: Αρχικό Silhouette διάγραμμα των 10-D Gaussian πληθυσμών .

Στο σημείο αυτό συνεχίζουμε την διαδικασία ομαδοποίησης με χρήση της τροποποιημένης μεθόδου SOM που αναπτύσσεται στον Πίνακα 3.

1. Μετά τον τυπικό έλεγχο που προηγήθηκε, αντιμετωπίζουμε τα δεδομένα μας (1800 δείγματα τριών ανεξάρτητων Gaussian πληθυσμών) σαν ένα σύνολο στοιχείων, αγνοώντας την κατηγοριοποίησή τους. Εφαρμόζουμε την τεχνική SOM (Ενότητα 2.2.5) στα αρχικά δεδομένα για να επιτύχουμε ομαδοποίηση των 1800 διανυσμάτων εισόδου. Ως εκ τούτου, δημιουργούμε ένα νευρωνικό δίκτυο SOM. Το δίκτυο αποτελείται από 25 κόμβους (nodes) έχοντας βάρος ο καθένας διάστασης 10 (ίσο με τη διάσταση κάθε εισόδου) και το πλέγμα των νευρώνων είναι εξαγωνικό. Αρχικοποιούνται τα βάρη κάθε κόμβου τυχαία, ο ρυθμός μάθησης (learning rate) αρχικοποιείται στην τιμή 0.9 και η εκπαίδευση του δικτύου ξεκινά σύμφωνα με τη γνωστή διαδικασία ανανέωσης των βαρών των κόμβων. Η εκπαίδευση ολοκληρώνεται όταν επιτευχθεί ένα από τα δύο κριτήρια τερματισμού, δηλαδή όταν οι εποχές φτάσουν τον αριθμό 25x500=12500 ή όταν το τετράγωνο της απολύτου διαφοράς των βαρών γίνει μικρότερο του 0.02 πάνω από 2500 εποχές. Το αποτέλεσμα της διαδικασίας αυτής είναι οι 25 κόμβοι ανανεωμένοι σε βάρη.
2. Δεύτερο βήμα της τροποποιημένης ομαδοποίησης SOM αποτελεί η ομαδοποίηση των παραπάνω 25 κόμβων σε ομάδες. Επιλέγουμε την δημιουργία 3 ομάδων (clusters). Με k-means (2.3) τεχνική χωρίζουμε τους κόμβους σε 3 ομάδες και βρίσκουμε τα κέντρα κάθε μίας, χωρίς να αποτελούν απαραίτητα κάποιο υπαρκτό κόμβο.
3. Στην συνέχεια για να ολοκληρωθεί η ομαδοποίηση, υπολογίζουμε για καθένα από τα αρχικά στοιχεία την ευκλείδεια απόσταση από τα 3 SOM κέντρα, και κατηγοριοποιούμε το εκάστοτε στοιχείο στην ομάδα εκείνη που το κέντρο της απέχει τη μικρότερη απόσταση.

Από το βήμα 3 προκύπτει ένα διάνυσμα στήλη (1800x1) το οποίο προβλέπει σε ποια ομάδα ανήκει κάθε ένα από τα δείγματα που δώσαμε για ομαδοποίηση. Συγκρίνοντας την προβλεπόμενη ομαδοποίηση με την *a priori* γνωστή κατάταξη παρατηρείται ότι υπάρχει 100% ταύτιση στην τοποθέτηση των στοιχείων σε ομάδες.

Συγκεκριμένα τα πρώτα 500 αποτελούν μια ομάδα, τα επόμενα 700 μια άλλη ομάδα και τέλος τα υπόλοιπα 800 κατηγοριοποιούνται μαζί.

Στον πίνακα 5, παραθέτουμε τις τιμές όλων των δεικτών εγκυρότητας που αναπτύχθηκαν αναλυτικά στο Κεφάλαιο 3, έπειτα από την ομαδοποίηση της τεχνητής βάσης δεδομένων.

Πίνακας 5: Δείκτες εγκυρότητας για τους Gaussian 10-D πληθυσμούς.

Silhouette	**Dunn**	**Davies-Bouldin**
0.9596	7.1367	0.1162

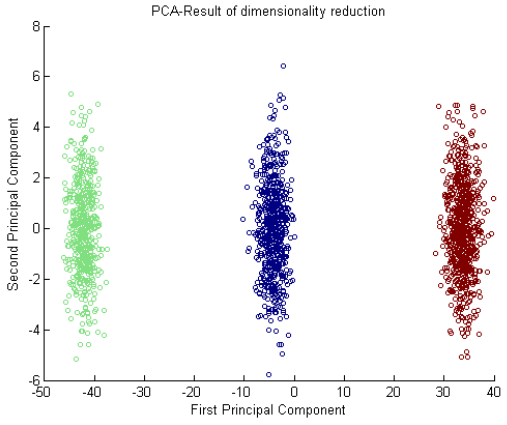
Παρατηρώντας τις τιμές των δεικτών συμπεραίνουμε ότι η ομαδοποίηση η οποία προέκυψε είναι η αναμενόμενη ικανοποιητική. Ο δείκτης Silhouette πλησιάζει τα όρια του βέλτιστου 1, ο DB τείνει στο μηδέν υποδεικνύοντας ένα καλό αποτέλεσμα ενώ η τιμή του DI είναι αρκετά υψηλή ορίζοντας την καλή ποιότητα της ομαδοποίησης. Και τα τρία συνεπώς μέτρα αξιολόγησης συνυπογράφουν για το καλό επίπεδο ομαδοποίησης που προέκυψε.

ΔΙΑΔΙΚΑΣΙΑ ΟΠΤΙΚΟΠΟΙΗΣΗΣ

Έχοντας τα δεδομένα μας και την κατηγοριοποίηση αυτών σε ομάδες προχωρούμε στην οπτικοποίηση των ομαδοποιημένων πλέον δεδομένων σε δύο διαστάσεις με τις τεχνικές μείωσης διαστάσεων που αναπτύχθηκαν στο Κεφάλαιο 5.

PCA

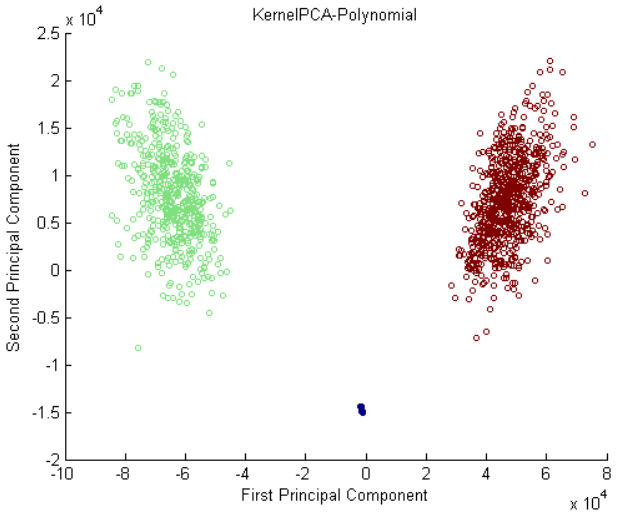
Στην Εικόνα 21 παρουσιάζεται η οπτικοποίηση των δεδομένων σε δύο διαστάσεις, προβάλλοντας τις δύο κύριες συνιστώσες.



Εικόνα 21: Οπτικοποίηση PCA. Τεχνητό πρόβλημα μετά από ομαδοποίηση SOM.Οι ομάδες που προέκυψαν παρουσιάζονται με διαφορετικό χρώμα.

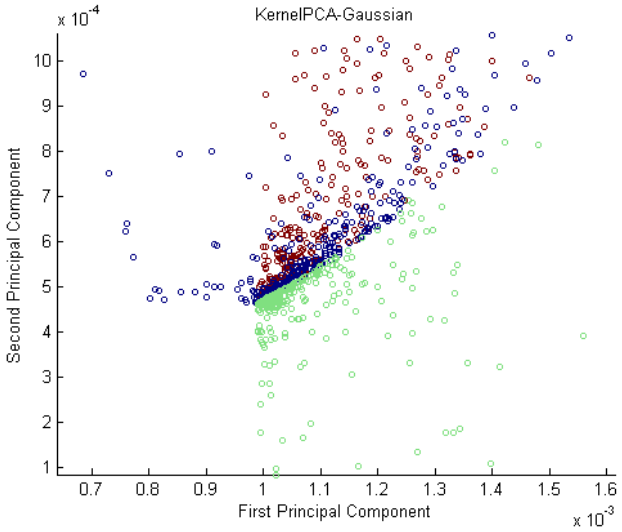
K-PCA

Στην Εικόνα 22 εμφανίζεται το αποτέλεσμα της ανάλυσης σε κύριες συνιστώσες με χρήση ενός πολυωνυμικού πυρήνα τρίτου βαθμού(degree=3).



Εικόνα 22: Οπτικοποίηση K-PCA polynomial. Τεχνητό πρόβλημα μετά από ομαδοποίηση SOM.

Στην Εικόνα 23 φαίνεται το αποτέλεσμα του k-pca με χρήση ενός Gaussian πυρήνα (σ=1).



Εικόνα 23: Οπτικοποίηση K-PCA Gaussian. Τεχνητό πρόβλημα μετά από ομαδοποίηση SOM.

Παρατηρώντας τις απεικονίσεις που καταλήξαμε παρατηρούμε ότι το αποτέλεσμα συνάδει με τον αναμενόμενο, που οι δείκτες μας όρισαν. Δηλαδή, η καλή ομαδοποίηση μεταφέρεται και στις δύο διαστάσεις των δεδομένων όπου η κάθε ομάδα φαίνεται να παρουσιάζει συμπαγή συμπεριφορά και να διακρίνεται εύκολα από τις άλλες. Η οπτικοποίηση εκείνη που μπορεί να χαρακτηριστεί ως η χειρότερη είναι του k-PCA με χρήση Gaussian πυρήνα.

LVQ & ΟΠΤΙΚΟΠΟΙΗΣΗ

Στο σημείο αυτό συνεχίζουμε την διαδικασία ομαδοποίησης με χρήση της μεθοδολογίας που αναπτύσσεται στον Πίνακα 4, δηλαδή τον τροποποιημένο LVQ αλγόριθμο. Ο αριθμός των κόμβων του νευρωνικού δικτύου που χρησιμοποιούμε είναι 25. Η εκπαίδευση εκτελείται μέχρι σύγκλισης του αλγορίθμου και έπειτα ακολουθεί ο αλγόριθμος k-means για να ομαδοποιήσει τους ανανεωμένους κόμβους του LVQ σε 3 ομάδες, για να τοποθετηθούν τελικά τα αρχικά δείγματα με το γνωστό κριτήριο της απόστασης στις 3 ομάδες. Όμοια με την SOM και η εν λόγω ομαδοποίηση τοποθετεί τα αρχικά στοιχεία στις ίδιες ομάδες. Με άλλα λόγια τόσο η SOM όσο και η LVQ τεχνικές καταλήγουν σε δύο διαμερίσεις πανομοιότυπες μεταξύ τους. Όλα λοιπόν τα αποτελέσματα της δεύτερης προσέγγισης ομαδοποίησης, συμπεριλαμβανομένων των δεικτών εγκυρότητας και των γραφημάτων οπτικοποίησης είναι ίδια με εκείνα της SOM τεχνικής και δεν επαναλαμβάνουμε την εμφάνισή τους.

Εφαρμόζοντας στον τεχνητό πληθυσμό δεδομένων όλη τη μεθοδολογία που έχουμε αναπτύξει στην εργασία, συμπεραίνουμε ότι παράγονται ικανοποιητικά αποτελέσματα, δηλαδή τα αποτελέσματα στα οποία καταλήγει συνάδουν με τα αναμενόμενα, τα οποία η φύση των δεδομένων μας επιτρέπει να γνωρίζουμε.

5.3 Εφαρμογή σε δεδομένα Σακχαρομύκητα

Οι παραπάνω συνδυαστικές τεχνικές της τροποποιημένης SOM και LVQ (Πίνακας 3 και 4 αντίστοιχα) ομαδοποίησης και οι μέθοδοι οπτικοποίησης εφαρμόζονται στη βάση δεδομένων του σακχαρομύκητα *Saccharomyces cerevisiae*. Οι εν λόγω διαδικασίες εκτελούνται στην κατεύθυνση των γονιδίων (γραμμές) όσο και των παραγόντων/μετρήσεις όπου τα γονίδια εκφράζονται (στήλες). Η βάση του σακχαρομύκητα που περιγράψαμε στην Ενότητα 1.3.1 πριν χρησιμοποιηθεί υπέστη κάποια τροποποίηση διατηρώντας τη βιολογική της ακεραιότητα. Για την καλύτερη ομαδοποίηση, αφαιρούμε κάποιους από τους παράγοντες (στήλες) και έτσι έχουμε στην διάθεσή μας 2467 γονίδια και 56 μετρήσεις (ALPH (8), ELU (8), CDC15 (10), SPO (11), HT (4), D(4), C (4) και DX(7)- Παράρτημα Β). Συγκεκριμένα, πραγματοποιήθηκε η αφαίρεση κάποιων χρονικών στιγμών των δειγμάτων με απώτερο στόχο τη χρονική εξομάλυνση των παραγόντων, ενώ παράλληλα γίνεται μία ενοποίηση κάποιων

ομάδων που παρουσιάζουν κοινή βιολογική συμπεριφορά, καθιστώντας το πλήθος των ομάδων 6.

SOM και ΟΠΤΙΚΟΠΟΙΗΣΗ των 56 παραγόντων

Η εφαρμογή της μεθοδολογίας ξεκινά από την περίπτωση των 56 πλέον παραγόντων καθώς στην κατεύθυνση αυτή γνωρίζουμε εξαρχής την κατηγοριοποίηση αυτών και μπορούμε να συγκρίνουμε αν η ομαδοποίηση που θα προκύψει συνάδει με την αναμενόμενη ή μπορεί να ερμηνευτεί βιολογικά. Να σημειωθεί στο σημείο αυτό ότι εξαιτίας της έλλειψης κάποιων μετρήσεων στη βάση δεδομένων εφαρμόστηκε μία τεχνική τοποθέτησης των τιμών που λείπουν που ονομάζεται μέθοδος k-πλησιέστερου γείτονα (knn- k nearest neighbor)[8].

Εφαρμόζουμε την τροποποιημένη SOM ομαδοποίηση με εισόδους τα 56 δείγματα διάστασης 2467 το καθένα. Η εκπαίδευση ξεκινά με χρήση 36 κόμβων και ολοκληρώνεται μετά το πέρας 18000 εποχών. Ο k-means αλγόριθμος εκτελείται με είσοδο τους ανανεωμένους κόμβους, τέσσερις φορές για 6, 7, 8 και 9 ομάδες (πλήθος ομάδων που δεν απέχει πολύ από το αρχικό). Για κάθε περίπτωση δημιουργούμε ένα διάνυσμα στήλη 56x1, που παρουσιάζει πως τα αρχικά δεδομένα κατηγοριοποιήθηκαν μετά την ολοκλήρωση της τροποποιημένης μεθοδολογίας SOM. Τα 5 συνολικά διανύσματα (4 διαμερίσεις συν το αρχικό) παρουσιάζονται στον Παράρτημα Β.

Παρατηρώντας τις διαμερίσεις έτσι όπως έχουν προκύψει στο Παράρτημα Β συμπεραίνουμε τα εξής: α)σε όλες τις περιπτώσεις οι δύο πρώτες ομάδες παραγόντων (alpha και elutriation) ομαδοποιούνται, β) οι ομάδες cdc και sporulation εξακολουθούν να αποτελούν ακέραιες ομάδες, γ) οι δύο εναπομείνασες 5 και 6 παρουσιάζουν μια γενική διαταραχή και δ) τέλος κάποια δείγματα στις ομάδες 4 και 5 σε όλες τις διαμερίσεις προκαλούν σημαντικό πρόβλημα στην ομαδοποίηση.

Σύμφωνα με τις παραπάνω παρατηρήσεις και με τη βοήθεια βιολογικής γνώσης αποφασίζουμε ότι οι ομάδες 1 και 2 θα ενσωματωθούν, οι ομάδες 3 και 4 θα αποτελούν χωριστές οντότητες κάτι που γίνεται και με τις 5 και 6, ενώ παράλληλα προχωρούμε στην αφαίρεση 10 δειγμάτων τα οποία επηρεάζουν αρνητικά την

[8] Η knn μέθοδος αντικαθιστά τις NaNs τιμές στα δεδομένα με την αντίστοιχη τιμή από τη στήλη του k πλησιέστερου γείτονα. Αν η k κοντινότερη στήλη είναι NaN τότε χρησιμοποιείται η αμέσως επόμενη στήλη.

ομαδοποίηση. Άρα οι 56 παράγοντες γίνονται 46 και η τελική διαμόρφωση των ομάδων έχει ως εξής:

- Ομάδα 1: Όλοι οι παράγοντες alpha και elutriation (16).
- Ομάδα 2:Τα 10 δείγματα της αρχικής 3 ομάδας.
- Ομάδα 3: Τα 7 δείγματα της αρχικής 4 ομάδας εκτός των spo 0,spo 5, spo5 7 και spo5 11.
- Ομάδα 4: Οι 6 παράγοντες της αρχικής 5 ομάδας εκτός των heat 0, dtt 15, dtt 30, dtt 60, dtt 120 και cold 0.
- Ομάδα 5: Τα 7 δείγματα της αρχικής 6 ομάδας.

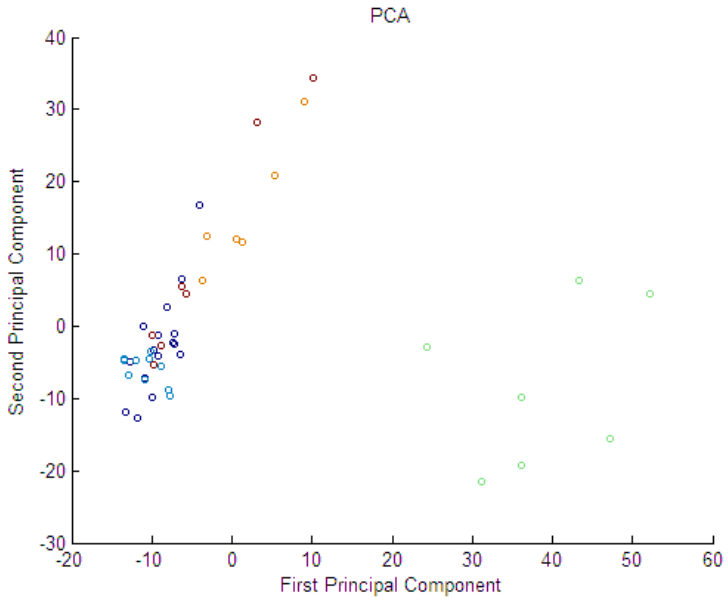
Στο σημείο αυτό παρουσιάζουμε στον Πίνακα 6 πως κινούνται οι δείκτες εγκυρότητας στην ομαδοποίηση με 56 δείγματα στις αρχικές 6 ομάδες (διαμέριση 1), με 56 δείγματα στις 6 ομάδες μετά την εφαρμογή της τροποποιημένης μεθοδολογίας SOM (διαμέριση 2) και τέλος με την δημιουργία των 5 ομάδων με 46 παράγοντες (διαμέριση 3).

Πίνακας 6: Δείκτες εγκυρότητας για διαφορετικές διαμερίσεις των παραγόντων του σακχαρομύκητα(τροποποιημένη SOM μεθοδολογία).

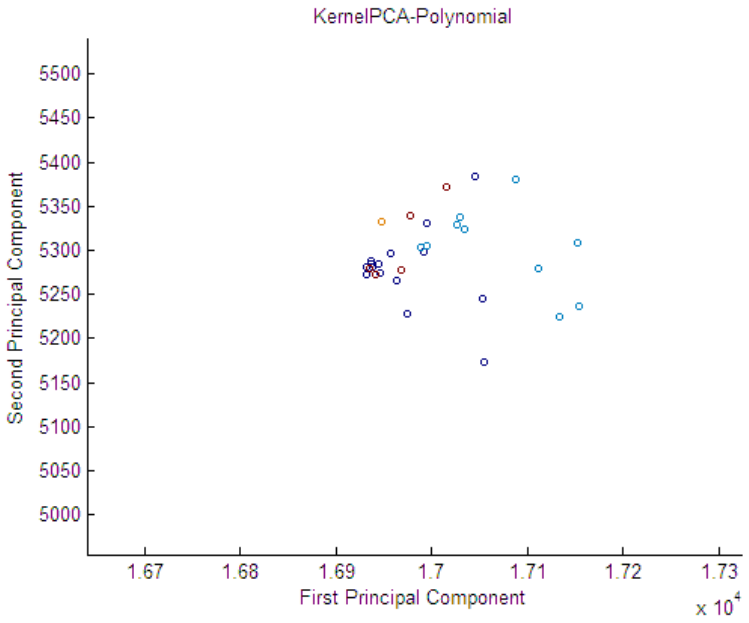
	Διαμέριση 1	**Διαμέριση 2**	**Διαμέριση 3**
Silhouette	0.1112	0.1244	0.2342
Dunn	0.4109	0.4591	0.6892
Davies Bouldin	1.9873	1.8933	1.3423

Εύκολα διαπιστώνουμε ότι με τις επιτρεπόμενες επεμβάσεις που έχουμε κάνει στην ομαδοποίηση καταφέρνουμε να βελτιώσουμε τους δείκτες αξιολόγησης όπως φαίνεται από τον παραπάνω πίνακα, οι οποίοι φαίνεται να κινούνται παράλληλα σε επίπεδο απόδοσης. Τα αποτελέσματα εντούτοις δεν είναι αρκετά ικανοποιητικά, καθώς όλοι οι δείκτες κινούνται μακριά από τα βέλτιστα όρια τους. Η μη επιθυμητή ομαδοποίηση σε επίπεδο δεικτών, μπορεί εύκολα να ερμηνευτεί από τη φύση των δεδομένων που προσπαθούμε να ομαδοποιήσουμε.

Αφού έχουμε καταλήξει στην κατηγοριοποίηση των 46 δειγμάτων σε 5 ομάδες, εν συνεχεία παρουσιάζουμε την οπτικοποίηση των ομαδοποιημένων δεδομένων σε 2 διαστάσεις, κάνοντας χρήση των τεχνικών ομαδοποίησης (PCA και k-PCA).



Εικόνα 24: Οπτικοποίηση PCA. Παράγοντες σακχαρομύκητα μετά από ομαδοποίηση SOM.



Εικόνα 25: Οπτικοποίηση K-PCA polynomial. Παράγοντες σακχαρομύκητα μετά από ομαδοποίηση SOM.

Από τις Εικόνες 24 και 25 βλέπουμε ότι τα 46 δείγματα-παράγοντες παρουσιάζουν μια σχετικά καλή οπτικοποίηση. Πιο καλή διάκριση των πέντε ομάδων έχουμε με τη μέθοδο PCA όπου η ομάδα 3 (sporulation με πράσινο χρώμα), φαίνεται να διαχωρίζεται καλύτερα σε σχέση με τις υπόλοιπες, οι οποίες μπορούμε να πούμε ότι παρουσιάζουν μια επικάλυψη. Συνοψίζοντας, διαπιστώνουμε ότι η μη ικανοποιητική συμπεριφορά των δεικτών εγκυρότητας στην ομαδοποίηση δεδομένων υψηλών διαστάσεων αντικατοπτρίζεται και στην τελική τους απεικόνιση σε δύο διαστάσεις.

LVQ και ΟΠΤΙΚΟΠΟΙΗΣΗ των 56 παραγόντων

Εφαρμόζουμε ομοίως την τροποποιημένη LVQ ομαδοποίηση με εισόδους τα 56 δείγματα διάστασης 2467 το καθένα. Όλα τα δείγματα χρησιμοποιούνται με την πληροφορία της κατηγοριοποίησής τους για καλύτερη εκπαίδευση του LVQ . Μετά από τη εκτέλεση δύο διαμερίσεων με αριθμό ομάδων 6 και 8 αντίστοιχα, δημιουργούμε και παρουσιάζουμε στο Παράρτημα Β δύο διανύσματα (56x1) τα οποία μας παρέχουν πληροφορία για το πώς τα 56 αρχικά δεδομένα κατηγοριοποιήθηκαν μετά την ολοκλήρωση της τροποποιημένης μεθοδολογίας LVQ.

Παρατηρώντας τις διαμερίσεις έτσι όπως έχουν προκύψει συμπεραίνουμε τα εξής: α) οι δύο πρώτες ομάδες παραγόντων (alpha και elutriation) διαχωρίζονται καλά, β) οι ομάδες cdc και sporulation εξακολουθούν να αποτελούν ακέραιες ομάδες, γ) οι δύο που υπολείπονται (5 και 6) παρουσιάζουν μια γενική διαταραχή και δ) τέλος κάποια δείγματα στις ομάδες 4 και 5 σε όλες τις διαμερίσεις προκαλούν σημαντικό πρόβλημα στην ομαδοποίηση.

Σύμφωνα με τα συμπεράσματα στα οποία καταλήξαμε και λαμβάνοντας υπ' όψιν τις βιολογικές πληροφορίες αποφασίζουμε ότι όλες οι ομάδες θα παραμείνουν ως έχουν κάνοντας μόνο μία μικρή ελεγχόμενη βιολογικά παρέμβαση αφαιρώντας 10 δείγματα, η παρουσία των οποίων επιδρά αρνητικά στην ομαδοποίηση. Άρα οι 56 παράγοντες γίνονται 46 και η τελική μορφή των ομάδων παρουσιάζεται παρακάτω:

- Ομάδα 1: Οι 8 παράγοντες alpha.
- Ομάδα 2: Οι 8 παράγοντες elutriation.
- Ομάδα 3: Τα 10 cdc15 δείγματα.
- Ομάδα 4: Τα 7 δείγματα της αρχικής 4 ομάδας εκτός των spo 0,spo 5, spo5 7 και spo5 11.
- Ομάδα 5: Τα 6 δείγματα της αρχικής 5 ομάδας εκτός των heat 0, dtt 15, dtt 30, dtt 60, dtt 120 και cold 0.
- Ομάδα 6: Τα 7 δείγματα της αρχικής 6 ομάδας.

Στον Πίνακα 7 εμφανίζονται οι τιμές των δεικτών εγκυρότητας στην ομαδοποίηση με 56 δείγματα στις αρχικές 6 ομάδες (διαμέριση 1), με 56 δείγματα στις 6 ομάδες μετά την εφαρμογή της τροποποιημένης μεθοδολογίαςLVQ (διαμέριση 2) και τέλος τι τιμές αυτοί παίρνουν μετά την δημιουργία των 6 ομάδων με 46 παράγοντες (διαμέριση 3).

Πίνακας 7: Δείκτες εγκυρότητας για διαφορετικές διαμερίσεις των παραγόντων του σακχαρομύκητα(τροποποιημένη LVQ μεθοδολογία).

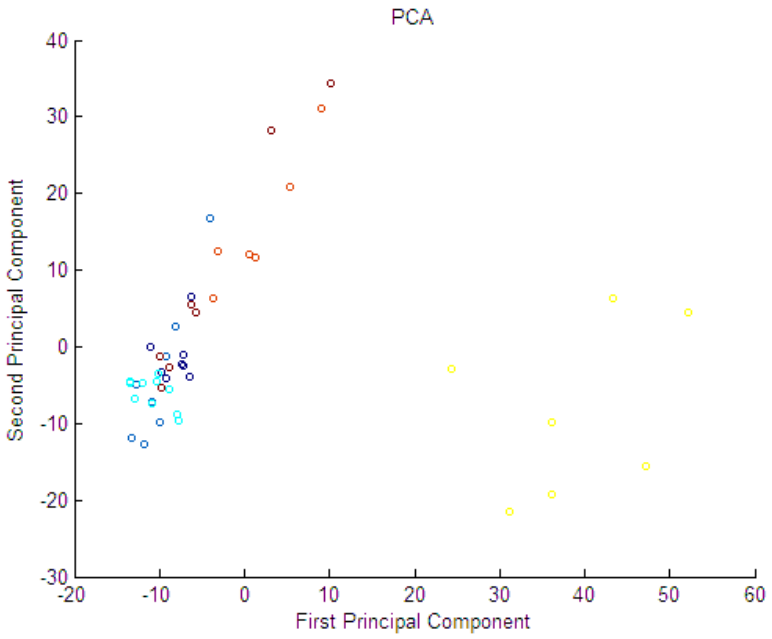
	Διαμέριση 1	**Διαμέριση 2**	**Διαμέριση 3**
Silhouette	0.1112	0.1194	0.1338
Dunn	0.4109	0.4295	0.4736
Davies Bouldin	1.9873	1.9333	1.7760

Διαπιστώνουμε ότι με τις επεμβάσεις που έχουμε κάνει στα όρια του επιτρεπτού, επιτυγχάνουμε ελάχιστη βελτίωση των δεικτών αξιολόγησης όπως φαίνεται από τον παραπάνω πίνακα (από αριστερά προς δεξιά) και ότι όλοι οι υπολογισμένοι δείκτες ακολουθούν μια ανάλογη πορεία. Τα αποτελέσματα όμως δεν είναι ιδιαίτερα ενθαρρυντικά, καθώς η απόδοση των δεικτών παραμένει σταθερά χαμηλή.

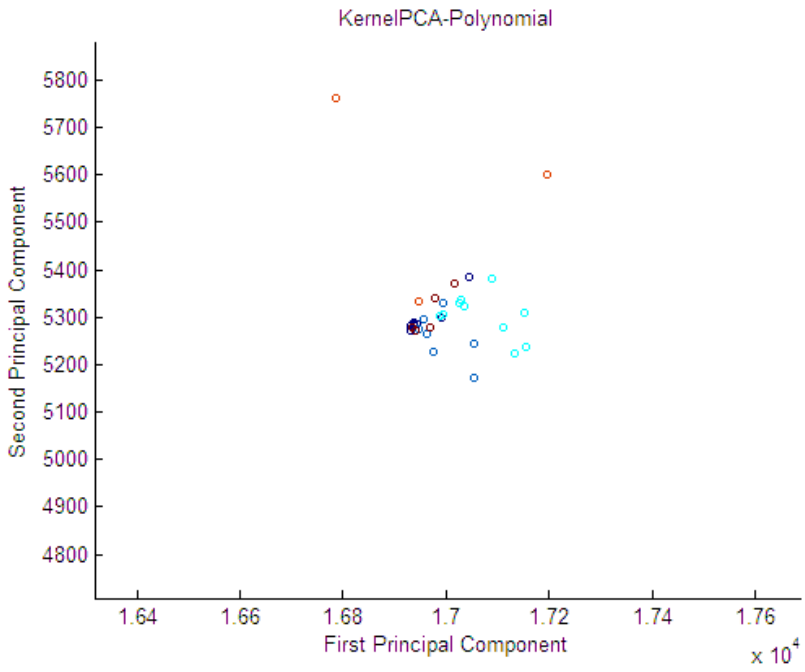
Αν αποπειραθούμε τη σύγκριση των αποτελεσμάτων των δύο μεθοδολογιών παρατηρούμε ότι η απόδοση της SOM υπερτερεί σε σχέση με εκείνη της LVQ, έκβαση αναμενόμενη αφού η δεύτερη αποτελεί στο πρώτο βήμα της εφαρμογής της μια επιβλέπουσα διαδικασία. Άρα οι ομάδες που προκύπτουν δεν μπορούν να απέχουν αρκετά από την αρχική κατηγοριοποίηση, διατηρώντας με τον τρόπο αυτό χαμηλά την ποιότητα της ομαδοποίησης. Με άλλα λόγια ο LVQ, δίνοντας για εκπαίδευση όλα τα δεδομένα εισόδου, επηρεάζεται από την αρχική βιολογική ομαδοποίηση και όχι από τη τάση των διανυσμάτων έκφρασης (στατιστική). Μια άλλη αντιμετώπιση αποτελεί ο συνδυασμός της βιολογίας και της στατιστικής, χρησιμοποιώντας ένα μέρος των δεδομένων για εκπαίδευση. Αυτό εφαρμόζεται στις ενότητες που ακολουθούν.

Μετά την τελική απόφαση της ομαδοποίησης των 46 δειγμάτων σε 6 ομάδες παρουσιάζουμε παρακάτω την οπτικοποίηση των ομαδοποιημένων δεδομένων σε 2 διαστάσεις με την υλοποίηση των αλγορίθμων μείωσης διαστάσεων.

Από τις Εικόνες 26 και 27 διαπιστώνουμε ότι υπάρχει μια σχετικά καλή απεικόνιση. Το καλύτερο οπτικό αποτέλεσμα προκύπτει με τη μέθοδο PCA όπου η ομάδα 4 (sporulation με κίτρινο χρώμα), φαίνεται να διαχωρίζεται καλύτερα σε σχέση με τις υπόλοιπες, οι οποίες μπορούμε να πούμε ότι παρουσιάζουν μια επικάλυψη. Συμπεραίνουμε λοιπόν ότι τα χαμηλά επίπεδα ομαδοποίησης που υπέδειξαν οι δείκτες εγκυρότητας , στο χώρο των υψηλών διαστάσεων γίνονται αντιληπτά και στην τελική απεικόνιση των δεδομένων σε δύο διαστάσεις.



Εικόνα 26: Οπτικοποίηση PCA. Παράγοντες σακχαρομύκητα μετά από ομαδοποίηση LVQ.

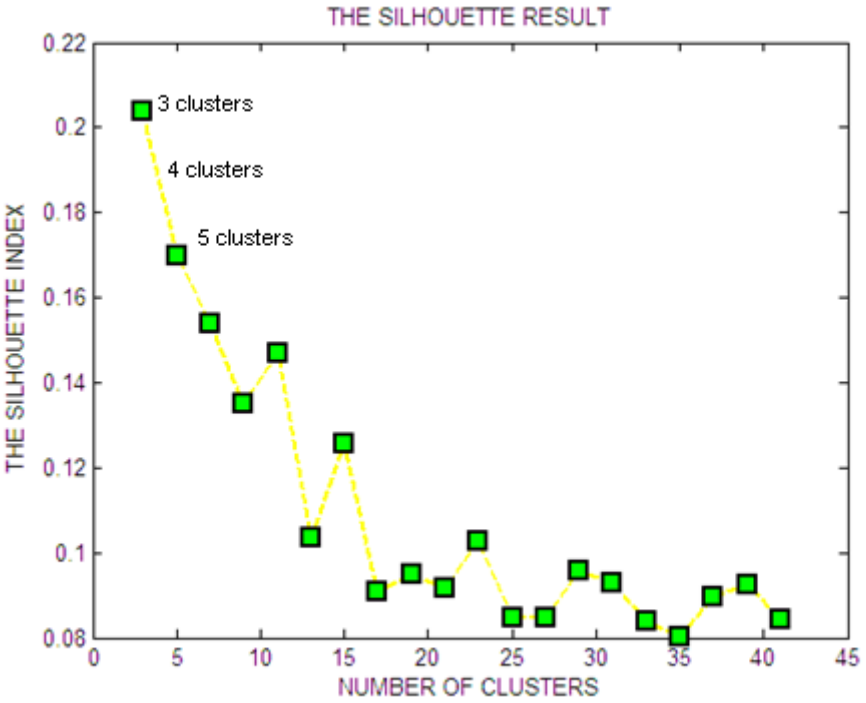


Εικόνα 27: Οπτικοποίηση K-PCA polynomial. Παράγοντες σακχαρομύκητα μετά από ομαδοποίηση LVQ.

<u>SOM και ΟΠΤΙΚΟΠΟΙΗΣΗ των 2467 γονιδίων</u>

Στο σημείο αυτό παρουσιάζουμε τα αποτελέσματα της ομαδοποίησης και οπτικοποίησης των 2467 γονιδίων του σακχαρομύκητα με βάση την τροποποιημένη μεθοδολογία ομαδοποίησης SOM. Το τεχνητό νευρωνικό δίκτυο SOM που χρησιμοποιείται έχει 100 κόμβους (10x10) και εκπαιδεύεται μετά από 50000 εποχές.

Λόγω της ιδιαιτερότητας του αλγορίθμου k-means, δηλαδή την απαίτηση που έχει να γνωρίζει εκ των προτέρων τον αριθμό των ομάδων, και της δικής μας αδυναμίας να έχουμε κάποια γνώση για τον αναμενόμενο αριθμό κατηγοριών, επαναλαμβάνουμε την ομαδοποίηση των κόμβων και κατ' επέκταση την κατηγοριοποίηση των γονιδίων σε ομάδες, αρκετές φορές. Ο λόγος που λαμβάνει χώρα αυτή η επανάληψη είναι για να αποφανθούμε ποιος είναι ο καλύτερος διαμερισμός πάνω στον οποίο θα βασιστεί η οπτικοποίηση. Τα μέτρα που μας βοηθούν να καταλήξουμε στον αριθμό των ομάδων που θα χρησιμοποιήσουμε είναι η συμπεριφορά του δείκτη Silhouette (Εικόνα 28) στις διάφορες διαμερίσεις καθώς και οι τιμές των δεικτών Davies Bouldin και Dunn όπως εμφανίζονται στον Πίνακα 8.

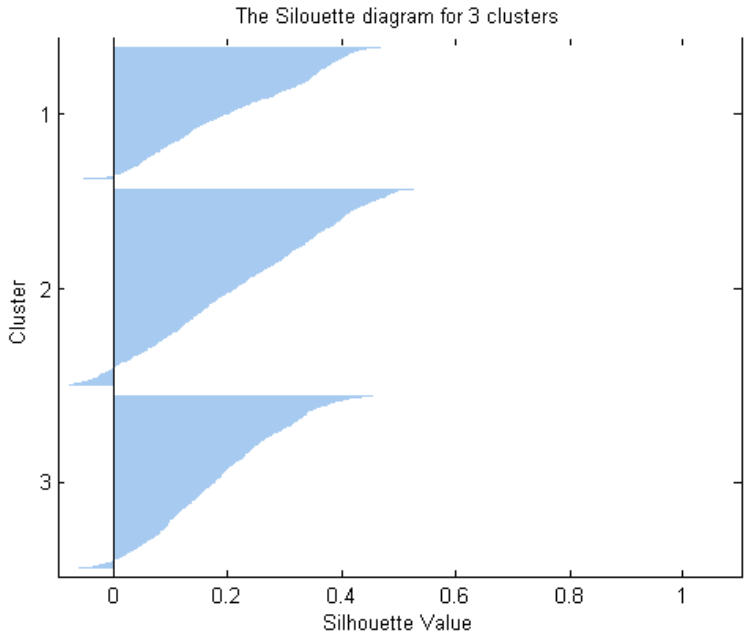


Εικόνα 28: Δείκτης Silhouette σε διάφορες διαμερίσεις των γονιδίων του σακχαρομύκητα.

Πίνακας 8: Δείκτες εγκυρότητας για την ομαδοποίηση των γονιδίων του σακχαρομύκητα σε διαφορετικές διαμερίσεις (τροποποιημένη SOM μεθοδολογία).

	Ομάδες 3	**Ομάδες 4**	**Ομάδες 5**
Silhouette	0.2039	0.1721	0.1674
Dunn	0.6435	0.5603	0.6144
Davies Bouldin	2.3963	2.4803	2.4047

Παρατηρώντας την Εικόνα 28 διαπιστώνουμε ότι ο δείκτης Silhouette εμφανίζει τη μέγιστη τιμή του στις 3 ομάδες, ενώ με την αύξηση των αριθμών των ομάδων ο δείκτης γνωρίζει μεγάλη πτώση, γεγονός που συνεπάγεται τη μείωση της ποιότητας της ομαδοποίησης. Σε συνδυασμό με τα παραπάνω, την ομαδοποίηση σε 3 ομάδες έρχεται να ενισχύσει η καλή απόδοση των Dunn και Davies Bouldin δεικτών. Ο Πίνακας 8 δείχνει τη ομόφωνη σύγκλιση όλων των δεικτών στις 3 ομάδες. Αν και οι τρεις δείκτες συνυπογράφουν για την δημιουργία τριών ομάδων, παρόλο αυτά η ομαδοποίηση δεν μπορεί να χαρακτηριστεί αρκετά ικανοποιητική. Στην Εικόνα 29 παρουσιάζουμε το Silhouette διάγραμμα όπως αυτό ορίζεται για 3 σύνολα γονιδίων.

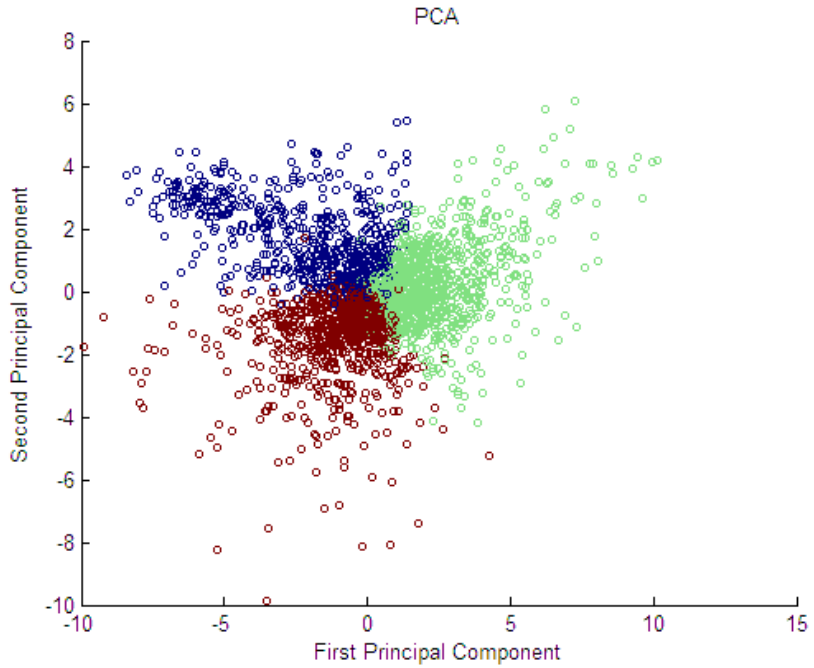


Εικόνα 29: Διάγραμμα Silhouette για τις 3 ομάδες γονιδίων του σακχαρομύκητα.

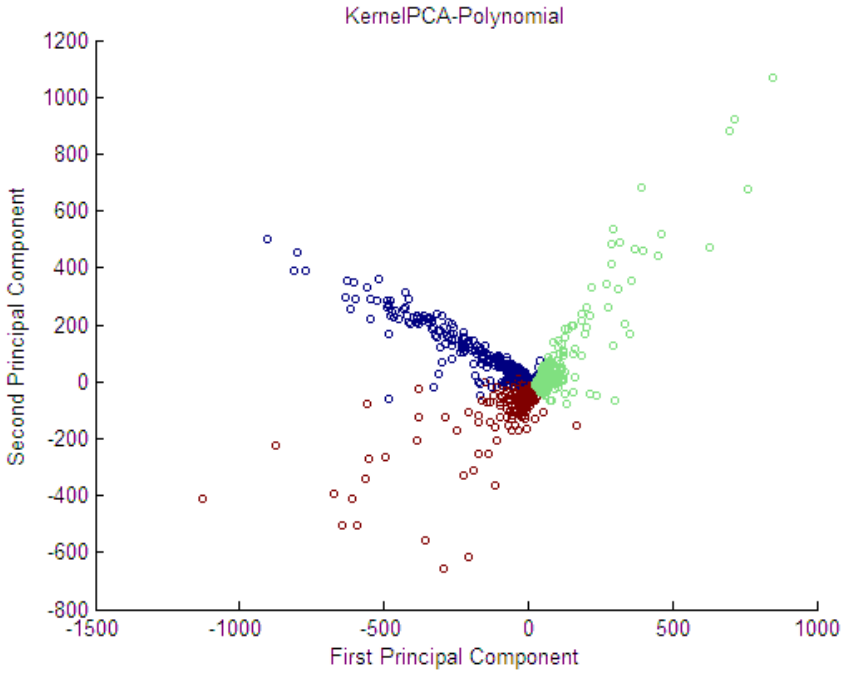
Χρήσιμο είναι να αναφέρουμε και το πλήθος των γονιδίων που απαρτίζουν κάθε ομάδα. Συγκεκριμένα έχουμε:

Ομάδα 1: 648 γονίδια. *Ομάδα 2*: 970 γονίδια *Ομάδα 3*: 849 γονίδια.

Στη συνέχεια παραθέτονται τα αποτελέσματα της οπτικοποίησης των τριών ομάδων γονιδίων.



Εικόνα 30: Οπτικοποίηση PCA. Γονίδια σακχαρομύκητα μετά από ομαδοποίηση SOM.



Εικόνα 31: Οπτικοποίηση K-PCA-polynomial (3d). Γονίδια σακχαρομύκητα μετά από ομαδοποίηση SOM.

Παρατηρώντας όλα τα γραφήματα διαπιστώνουμε ότι στις δύο διαστάσεις οι ομάδες διακρίνονται καθαρά. Κάθε ομάδα φαίνεται αρκετά συμπαγής αλλά δυστυχώς ο μεταξύ τους διαχωρισμός δεν είναι εφικτός σε μεγάλο βαθμό. Τα όρια της μιας εισέρχονται στα όρια της άλλης δημιουργώντας κάποια επικάλυψη των ομάδων σε ορισμένα σημεία. Οι σχετικά χαμηλές τιμές των δεικτών έρχονται και επιβεβαιώνουν

το οπτικό αποτέλεσμα, καθώς μπορεί η κάθε ομάδα να παρουσιάζει ομοιογένεια αλλά είναι οριακά διαχωρίσιμη από τις υπόλοιπες.

LVQ και ΟΠΤΙΚΟΠΟΙΗΣΗ των 2467 γονιδίων

Σε ένα δίκτυο LVQ με χαρακτηριστικά 100 κόμβους με 56 διάσταση ο καθένας και περάτωση του αλγορίθμου μετά από 50000 εποχές δίνουμε για εκπαίδευση 747 γονίδια από τα αρχικά 2467 γονίδια. Καθώς ο αλγόριθμος LVQ γνωρίζει την κατηγοριοποίηση κάποιων δεδομένων εισόδου για εκπαίδευση, εξάγουμε με βάση τα αποτελέσματα της τροποποιημένης SOM ομαδοποίησης τα γονίδια εκείνα που έχουν silhouette τιμή μεγαλύτερη ή ίση του 0.3 (Εικόνα 29). Αφού ολοκληρωθεί η εκπαίδευση των 747 δειγμάτων, εφαρμόζουμε k-means ομαδοποίηση στους κόμβους και έπειτα κατηγοριοποιούμε τα αρχικά δεδομένα σύμφωνα με τις αποστάσεις αυτών από τα κέντρα που ο k-means ορίζει. Όμοια με πριν παράγουμε διαφορετικές διαμερίσεις και προσπαθούμε να αποφασίσουμε για τα βέλτιστο αριθμό ομάδων. Στην περίπτωση αυτή καταλήγουμε επίσης σε 3 ομάδες, καθώς οι δείκτες εγκυρότητας το επιδεικνύουν (Πίνακας 9). Το γεγονός ότι και στην ομαδοποίηση αυτή καταλήξαμε σε 3 ομάδες δεν προκαλεί έκπληξη, καθώς η ομαδοποίηση των κόμβων έγινε με δείγματα εκπαίδευσης ομαδοποιημένα σε τρεις ομάδες επηρεάζοντας το τελικό αποτέλεσμα. Οι τιμές των δεικτών αξιολόγησης σε σχέση με την τροποποιημένη μεθοδολογία SOM εμφανίζονται βελτιωμένοι.

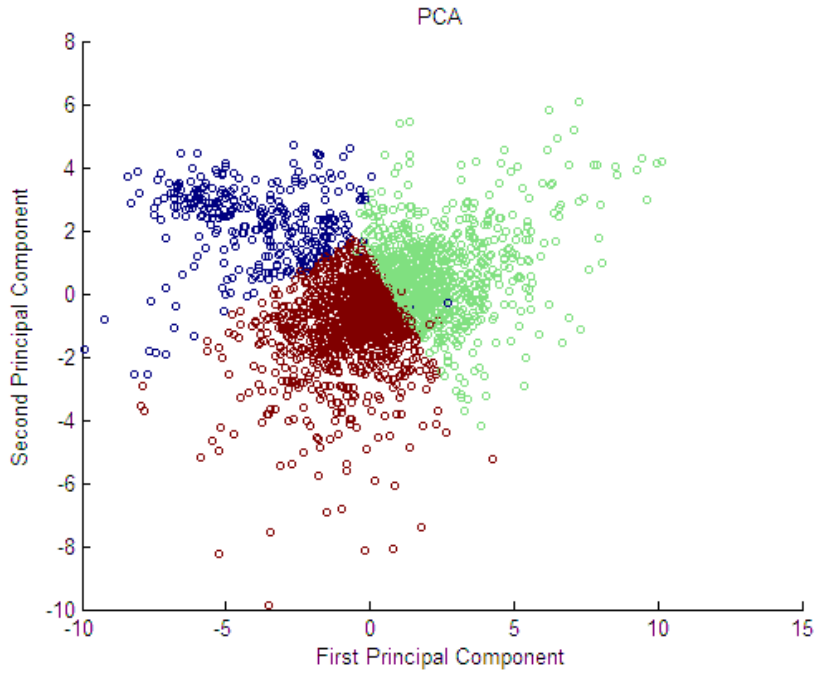
Πίνακας 9: Δείκτες εγκυρότητας για την ομαδοποίηση των γονιδίων του σακχαρομύκητα σε 3 ομάδες (τροποποιημένη LVQ μεθοδολογία).

Silhouette	**Dunn**	**Davies-Bouldin**
0.3112	0.8432	1.9054

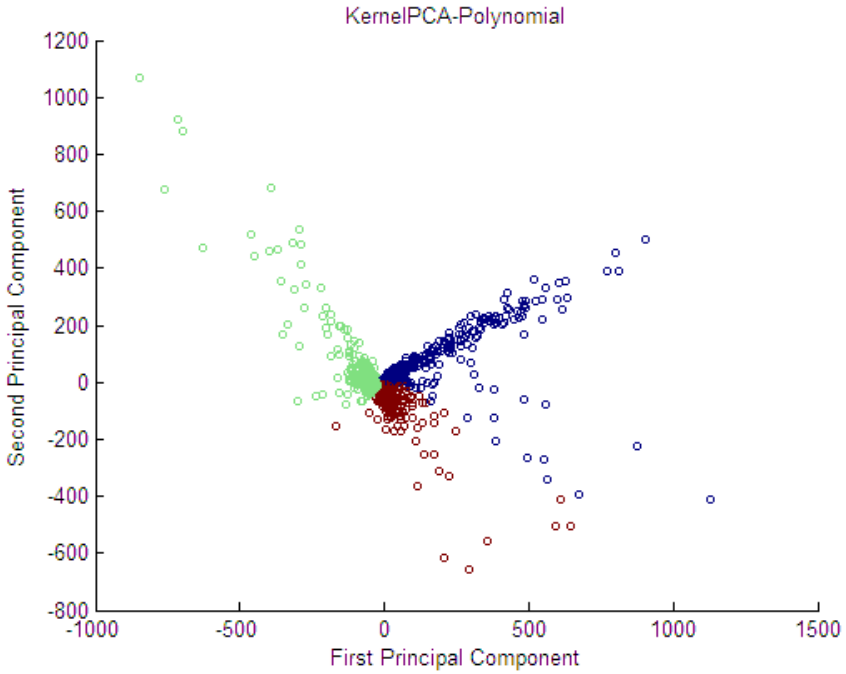
Το πλήθος των γονιδίων που κάθε ομάδα εμπεριέχει φαίνονται παρακάτω:

Ομάδα 1: 317 γονίδια. *Ομάδα 2*: 844 γονίδια *Ομάδα 3*: 1306 γονίδια.

Έπονται οι απεικονίσεις των τριών ομάδων που καταλήξαμε σε δύο διαστάσεις.



Εικόνα 32: Οπτικοποίηση PCA. Γονίδια σακχαρομύκητα μετά από ομαδοποίηση LVQ.

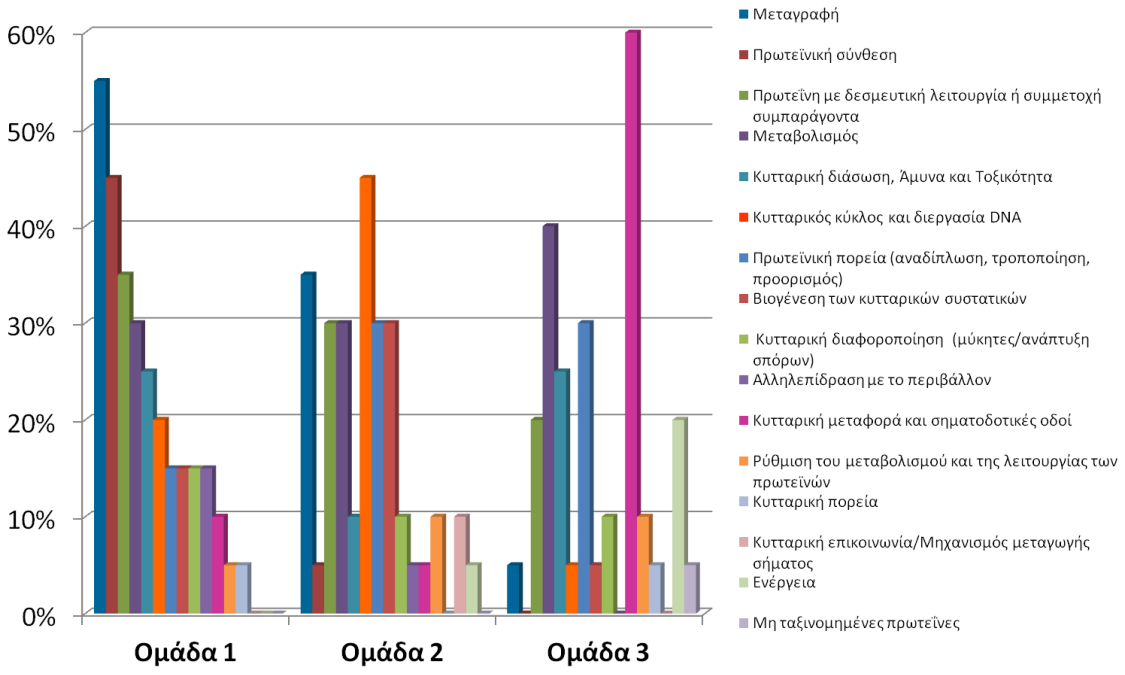


Εικόνα 33: Οπτικοποίηση K-PCA-polynomial (3d). Γονίδια σακχαρομύκητα μετά από ομαδοποίηση LVQ.

Η ομαδοποίηση με βάση την LVQ μεθοδολογία εμφανίζει καλύτερα αποτελέσματα τόσο σε επίπεδο δεικτών όσο και στο οπτικό αποτέλεσμα. Μεγάλη βελτίωση στην οπτικοποίηση παρουσιάζεται στη τεχνική PCA. Οι ομάδες είναι συμπαγείς και δεν υπάρχει έντονη επικάλυψη μεταξύ τους, σε αντίθεση με την οπτικοποίηση των ομάδων που ακολούθησε την SOM ομαδοποίηση.

Βιολογική αξιολόγηση

Βασιζόμενοι στα αποτελέσματα ομαδοποίησης του LVQ (λόγω καλύτερης ομαδοποίησης των γονιδίων) εξάγουμε τα είκοσι κεντρικά γονίδια (Παράρτημα Β) από καθεμία ομάδα. Ο τρεις ομάδες, των είκοσι γονίδιων, αξιολογούνται βιολογικά με το Σύστημα Ταξινόμησης MIPS FunCat και τα αποτελέσματα της βιολογικής ερμηνείας παρουσιάζονται στην Εικόνα 34. Σε κάθε ομάδα (Παράρτημα Α) υπάρχει μία διεργασία η οποία υπερτερεί, με αποτέλεσμα η στατιστική ομαδοποίηση να συνοδεύεται και από διαφορές στη βιολογική ερμηνεία. Αναλυτικότερα, τα 20 κεντρικά γονίδια της καθεμία από τις 3 ομάδες συμμετέχουν σε 11 κοινές βιολογικές διεργασίες συμπεριλαμβάνοντας τη μεταγραφή, τον μεταβολισμό, την κυτταρική διάσωση-άμυνα-τοξικότητα. Η πρωτεϊνική σύνθεση και η αλληλεπίδραση με το περιβάλλον αποτελούν κοινές βιολογικές διεργασίες για τα 20 γονίδια των ομάδων 1 και 2, ενώ αντίστοιχα η κυτταρική πορεία για τα γονίδια των ομάδων 1 και 3, και η ενέργεια για τα γονίδια των ομάδων 2 και 3. Η διακριτότητα των 3 ομάδων ανακλάται από την υπεροχή των τριών βιολογικών διεργασιών σε κάθε μία από αυτές. Συνεπώς, η επικράτηση της μεταγραφής (55%), του κυτταρικού κύκλου και της διεργασίας DNA (45%) και της κυτταρικής μεταφοράς και των σηματοδοτικών οδών (60%) σε καθεμία από τις 3 ομάδες παρέχει την επιβεβαίωση της στατιστικής ομαδοποίησης και οπτικοποίησης γονιδίων.



Εικόνα 34: Βιολογικές Διεργασίες. Συγκριτικά αποτελέσματα από τις 3 ομάδες Saccharomyces cerevisiae.

5.4 Εφαρμογή σε δεδομένα Οστεοαρθρίτιδας

Οι προτεινόμενες μεθοδολογίες ομαδοποίησης SOM και LVQ εφαρμόζονται σε δύο διαφορετικούς τύπους κυττάρων (μεσεγχυματικά βλαστικά κύτταρα και χονδροκύτταρα) [27] με σκοπό να παραχθούν χρήσιμα αποτελέσματα, που θα βοηθήσουν στην αντιμετώπιση της ασθένειας της οστεοαρθρίτιδας.

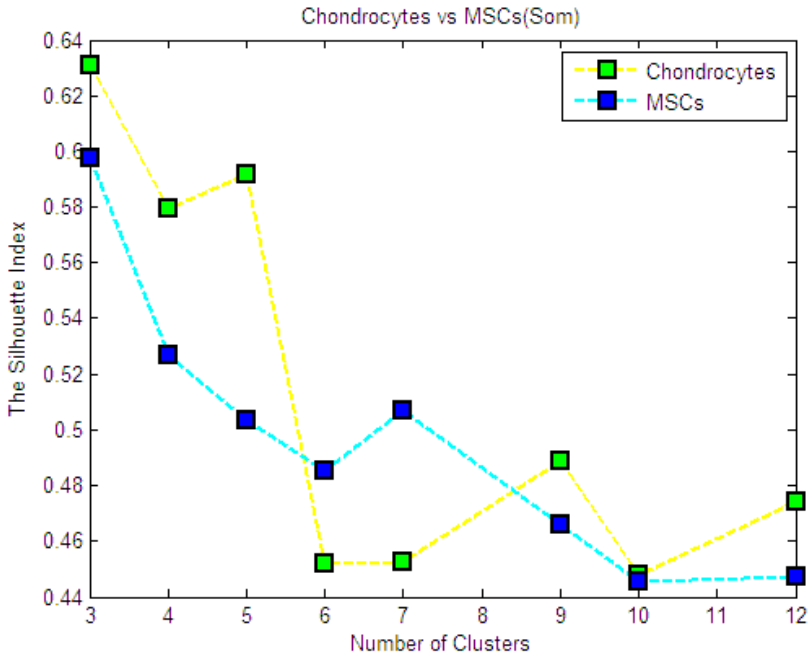
Στοχεύουμε σε μια καλή ομαδοποίηση των γονιδίων και στις δύο περιπτώσεις κάνοντας χρήση και των δύο προτεινόμενων μεθόδων, ενώ παράλληλα επιθυμούμε να εξάγουμε και ένα καλό οπτικό αποτέλεσμα των γονιδίων. Θέλουμε να επιτύχουμε ένα διαμερισμό όσο το δυνατόν συμπαγή ώστε να μπορούμε να προβούμε στην βιολογική αξιολόγηση της ομαδοποίησης. Η εκτίμηση θα γίνει τόσο σε επίπεδο του διαμερισμού κάθε ομάδας κυττάρων όσο και μεταξύ των δύο επικρατέστερων διαμερίσεων μεσεγχυματικών βλαστικών κυττάρων και χονδροκυττάρων. Στην πρώτη περίπτωση θα αποφανθούμε για το αν οι ομάδες που θα συσταθούν χαρακτηρίζονται από διαφορετικά βιολογικά χαρακτηριστικά (γονίδια), ενώ στην δεύτερη θα συγκρίνουμε αν οι ομάδες μεταξύ μεσεγχυματικών βλαστικών κυττάρων και χονδροκυττάρων παρουσιάζουν ομοιότητες. Μια αυξημένη ομοιότητα μεταξύ των δύο κυτταρικών τύπων υποδηλώνει την επιτυχία μετατροπής των μεσεγχυματικών βλαστικών κυττάρων σε χονδροκύτταρα κατά τη διάρκεια της χονδρογενούς διαφοροποίησης.

SOM και ΟΠΤΙΚΟΠΟΙΗΣΗ των 17589 γονιδίων Chondocytes και MSCs

Τα γονίδια των δύο βάσεων δεδομένων (μεσεγχυματικά βλαστικά κύτταρα και χονδροκύτταρα) ομαδοποιούνται με την τροποποιημένη SOM μεθοδολογία (100 κόμβοι -50000 εποχές) και παράγονται για κάθε περίπτωση διαφορετικές διαμερίσεις για να μπορέσουμε να αποφασίσουμε ποια είναι η πιο κατάλληλη, με τη βοήθεια του δείκτη εγκυρότητας Silhouette. Στην Εικόνα 35 παρουσιάζονται οι τιμές του μέτρου εκτίμησης σε ένα πλήθος διαφορετικών διαμερίσεων ταυτόχρονα και για τις δύο κατηγορίες κυττάρων.

Παρατηρώντας τη συμπεριφορά του δείκτη βλέπουμε ότι σε γενικές γραμμές οι διαμερίσεις για λίγες ομάδες παρουσιάζουν ένα καλό επίπεδο ομαδοποίησης και στις δύο κατηγορίες. Η αύξηση των αριθμών των ομάδων γενικά οδηγεί σε μείωση της απόδοσης του δείκτη εγκυρότητας. Η ομαδοποίηση των χονδροκυττάρων, παρά της

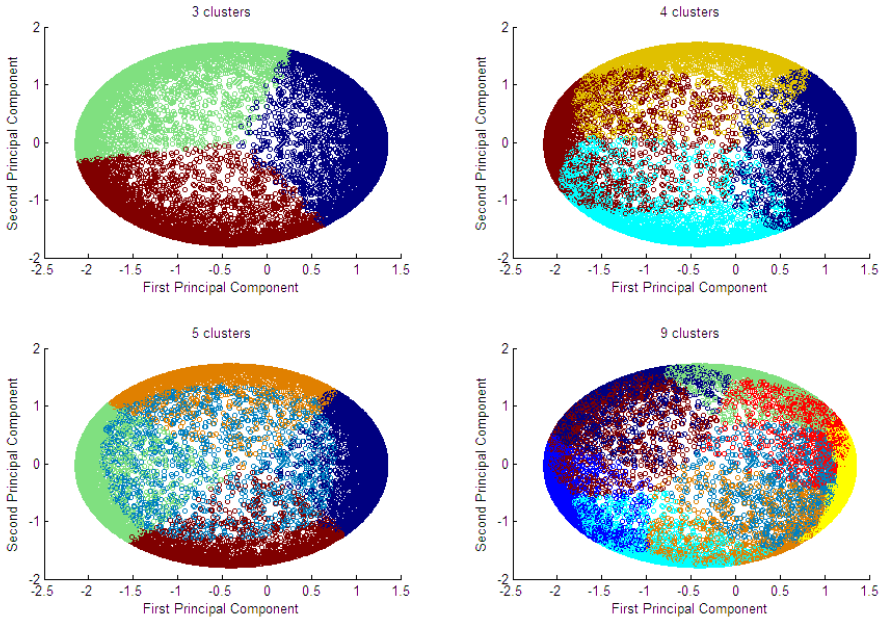
έντονες μεταβολές εμφανίζει μια καλύτερη απόδοση συγκριτικά με τα μεσεγχυματικά βλαστικά κύτταρα. Και οι δύο τύποι παρουσιάζουν τη μέγιστη τιμή τους (0.6307 και 0.5978 αντίστοιχα) για αριθμό ομάδων ίσο με τρία, χωρίς όμως οι άλλες διαμερίσεις να υπολείπονται τιμής αισθητά.



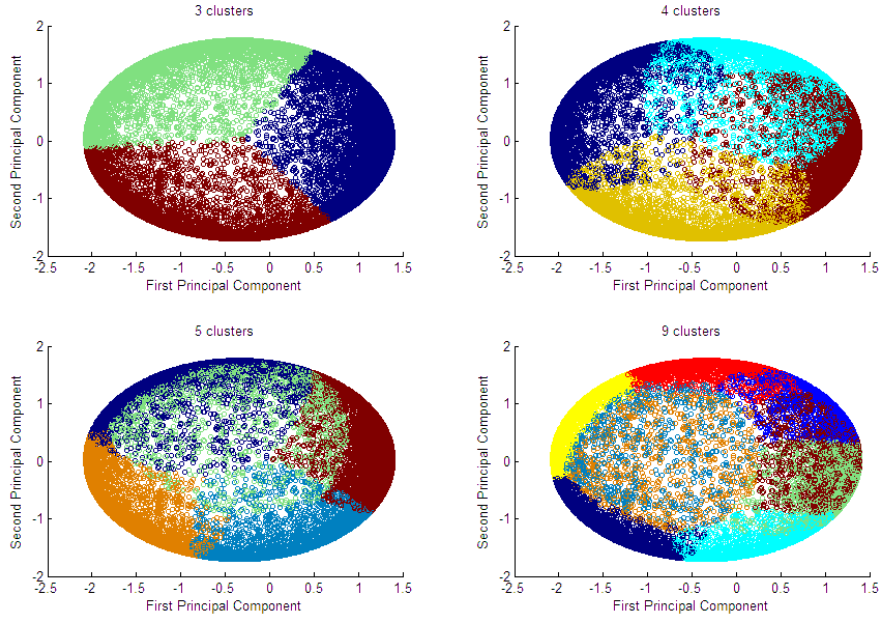
Εικόνα 35: Δείκτης Silhouette σε διάφορες διαμερίσεις των MSCs και Χονδροκυττάρων μετά από ομαδοποίηση SOM.

Θέλοντας να δούμε πως εξελίσσεται η οπτικοποίηση των δύο τύπων κυττάρων σε διαφορετικό αριθμό ομάδων υλοποιούμε την τεχνική PCA και προβάλουμε τα αποτελέσματα για τα χονδροκύτταρα και τα μεσεγχυματικά βλαστικά κύτταρα στις Εικόνες 36 και 37 αντίστοιχα (αποτελέσματα από άλλες τεχνικές μείωσης διαστάσεων στο Παράρτημα Α).

Εύκολα παρατηρείται ποιοτικά ότι τα αποτελέσματα της οπτικοποίησης με την αύξηση των αριθμού των ομάδων φθίνουν από άποψη απόδοσης, καθώς οι ομάδες επικαλύπτονται ολοένα και περισσότερο. Η τιμή του δείκτη αξιολόγησης της εκάστοτε ομαδοποίησης συνάδει με το οπτικό αποτέλεσμα στις δύο διαστάσεις. Συγκεκριμένα, στην περίπτωση των τριών ομάδων όπου ο δείκτης έχει αρκετά υψηλή τιμή, η οπτικοποίηση είναι πολύ ικανοποιητική. Επιπρόσθετα, αξίζει να σημειωθεί ότι οι αντίστοιχες, σε αριθμό ομάδων, διαμερίσεις των μεσεγχυματικών βλαστικών κυττάρων και χονδροκυττάρων παρουσιάζουν μια παράλληλη εικόνα σε επίπεδο οπτικοποίησης, η οποία και αναμενόμενη από το δείκτη Silhouette.



Εικόνα 36: Οπτικοποίηση PCA των χονδροκυττάρων σε 2-D για διαφορετικό αριθμό ομάδων μετά από ομαδοποίηση SOM.



Εικόνα 37: Οπτικοποίηση PCA των μεσεγχυματικών βλαστικών κυττάρων σε 2-D για διαφορετικό αριθμό ομάδων μετά από ομαδοποίηση SOM.

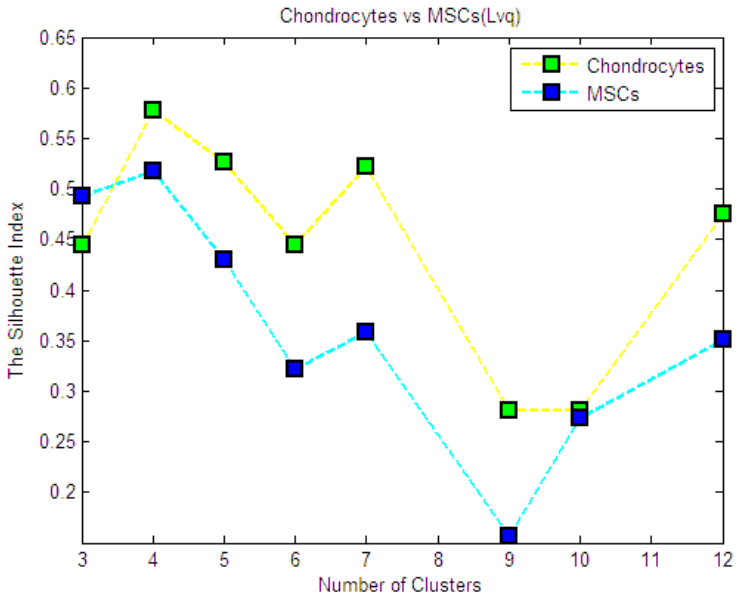
LVQ και ΟΠΤΙΚΟΠΟΙΗΣΗ των 17589 γονιδίων Chondrocytes και MSCs

Συνεχίζουμε την ομαδοποίηση των δύο επιμέρους βάσεων δεδομένων κάνοντας χρήση τώρα της τροποποιημένης μεθόδου LVQ. Έχουμε ήδη αναφέρει ότι το πρώτο βήμα της εν λόγω μεθόδου χρειάζεται τη στοιχειώδη κατηγοριοποίηση των στοιχείων της εισόδου της. Για το λόγο αυτό η εκπαίδευση στην περίπτωση των

χονδροκυττάρων ξεκινάει με τα 10533 γονίδια τα οποία έχουν Silhouette τιμή (Παράρτημα Α) μεγαλύτερη ή ίση του 0.58 κατά την ομαδοποίηση SOM σε τρείς ομάδες. Ανάλογα, τα γονίδια της δεύτερης βάσης που δίνονται για ομαδοποίηση είναι 9979 και χαρακτηρίζονται από Silhouette τιμή (Παράρτημα Α) μεγαλύτερη ή ίση του 0.5 σύμφωνα με την ομαδοποίηση SOM σε τρείς ομάδες. Έπειτα συνεχίζουμε και με την ομαδοποίηση και των υπολοίπων γονιδίων.

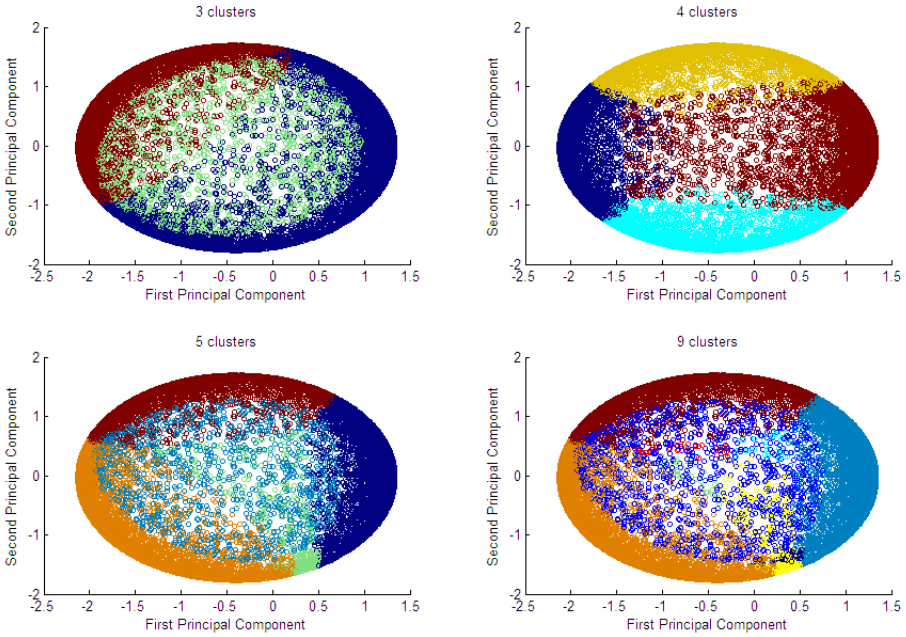
Για κάθε μία των περιπτώσεων δημιουργούμε διαφορετικές LVQ διαμερίσεις, με τη γνωστή διαδικασία, ώστε να επιλέξουμε εκείνη που έχει μεγαλύτερη απόδοση. Στην Εικόνα 38 προβάλουμε τη συμπεριφορά του δείκτη Silhouette, όπως αυτός διαμορφώνεται με την αύξηση τον αριθμό των ομάδων και στους δύο τύπους κυττάρων.

Στην Εικόνα 38 παρατηρούμε ότι το επίπεδο ομαδοποίησης των χονδροκυττάρων εξακολουθεί να υπερτερεί σχεδόν σε όλες τις αντίστοιχες διαμερίσεις. Συγκριτικά , οι δύο τύποι κυττάρων εμφανίζουν τη βέλτιστη ομαδοποίησης τους στη διαμέριση με τέσσερις ομάδες. Στα χονδροκύτταρα η τιμή του δείκτη είναι 0.5775 ενώ στα μεσεγχυματικά βλαστικά κύτταρα το μέτρο αγγίζει το 0.5172. Γενικά, οι τέσσερις ομάδες ορίζουν μία πολύ καλή ομαδοποίηση καθώς ο δείκτης αξιολόγησης κινείται σε ικανοποιητικά επίπεδα.

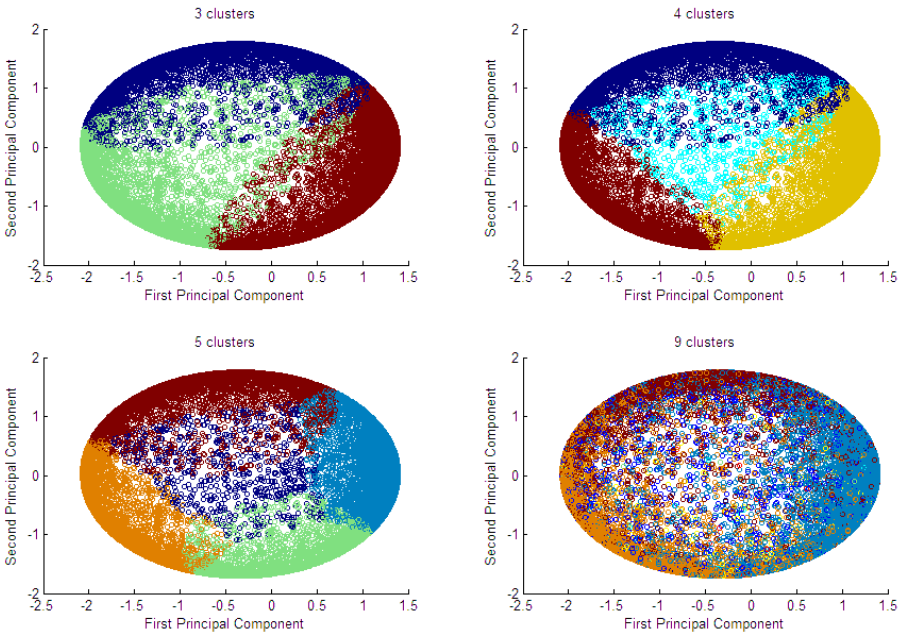


Εικόνα 38: Δείκτης Silhouette σε διάφορες διαμερίσεις των MSCs και Χονδροκυττάρων μετά από ομαδοποίηση LVQ.

Στη συνέχεια (Εικόνα 39 και 40) παραθέτουμε τα αποτελέσματα της οπτικοποίησης με χρήση της τεχνικής PCA και για τις δύο κατηγορίες, όπως διαμορφώνονται με την αύξηση του αριθμού των ομάδων (αποτελέσματα άλλων τεχνικών στο Παράρτημα Α).



Εικόνα 39: Οπτικοποίηση PCA των χονδροκυττάρων σε 2-D για διαφορετικό αριθμό ομάδων μετά από ομαδοποίηση LVQ.



Εικόνα 40: Οπτικοποίηση PCA των μεσεγχυματικών βλαστικών κυττάρων σε 2-D για διαφορετικό αριθμό ομάδων μετά από ομαδοποίηση LVQ.

Η εξέλιξη των απεικονίσεων με την αύξηση των ομάδων, που παρατίθεται παραπάνω, είναι συγκρίσιμη με την πορεία που ακολουθεί ο δείκτης Silhouette. Πιο

συγκεκριμένα, η αύξηση της τιμής του δείκτη από της τρεις ομάδες στις τέσσερις συνάδει με το οπτικό αποτέλεσμα των δύο αυτών διαμερίσεων ,και στους δύο τύπους κυττάρων. Η απεικόνιση των εννιά ομάδων φαίνεται κακής ποιότητας, γεγονός που οι δείκτες έχουν προδιαθέσει, αφού οι τιμές τους κυμαίνονται σε χαμηλά επίπεδα (Εικόνα 38).

Σύγκριση διαμερίσεων και επιλογή ομάδων για βιολογική αξιολόγηση

Στο σημείο αυτό καλούμαστε να αποφασίσουμε τον αριθμό των ομάδων που θα δοθούν για βιολογική ερμηνεία. Αν η απόφαση μας επηρεαστεί από τα αποτελέσματα της πρώτης ομαδοποίησης (SOM) θα καταλήξουμε σε διαμερισμό τριών ομάδων, ενώ αν στραφούμε στην δεύτερη (LVQ) θα υπερισχύσουν οι τέσσερις ομάδες. Οι επιδόσεις των δύο αυτών ενδεχόμενων επιλογών είναι σχετικά όμοιες , παρουσιάζοντας οι τρεις ομάδες ένα μικρό προβάδισμα (όπου ο δείκτης Silhouette έχει υψηλότερη τιμή). Παρόλα αυτά επιλέγουμε να προχωρήσουμε σε βιολογική αξιολόγηση με τις τέσσερις ομάδες (μεσεγχυματικά βλαστικά κύτταρα και χονδροκύτταρα), αυξάνοντας έτσι την διερεύνηση του αποτελέσματος.

Στα πλαίσια της σύγκρισης των δύο διαμερίσεων (μεσεγχυματικά βλαστικά κύτταρα και χονδροκύτταρα), ελέγχουμε τον ποσοστό ομοιότητας κάθε ομάδας των μεσεγχυματικών βλαστικών κυττάρων με όλες τις ομάδες των χονδροκυττάρων. Το ποσοστό αυτό υπολογίζεται από το λόγο των κοινών γονιδίων κάθε συνδυασμού προς το πλήθος της εκάστοτε συγκρινόμενης ομάδας μεσεγχυματικών βλαστικών κύτταρων. Στο Πίνακα 10 παρουσιάζεται το πλήθος των τεσσάρων ομάδων κάθε διαμέρισης ενώ στον Πίνακα 11 παραθέτουμε τα ποσοστά ομοιότητας που προκύπτουν από κάθε συνδυασμό.

Πίνακας 10: Πλήθος τεσσάρων ομάδων χονδροκυττάρων και μεσεγχυματικών κυττάρων.

Αριθμός ομάδας	Πλήθος MSCs	Πλήθος Chondrocytes
1	4485	3210
2	865	2701
3	8319	3039
4	3920	8639
Σύνολο	17589	17589

Πίνακας 11: Ποσοστά ομοιότητας για όλους τους συνδυασμούς των ομάδων (Chondocytes και MSCs).

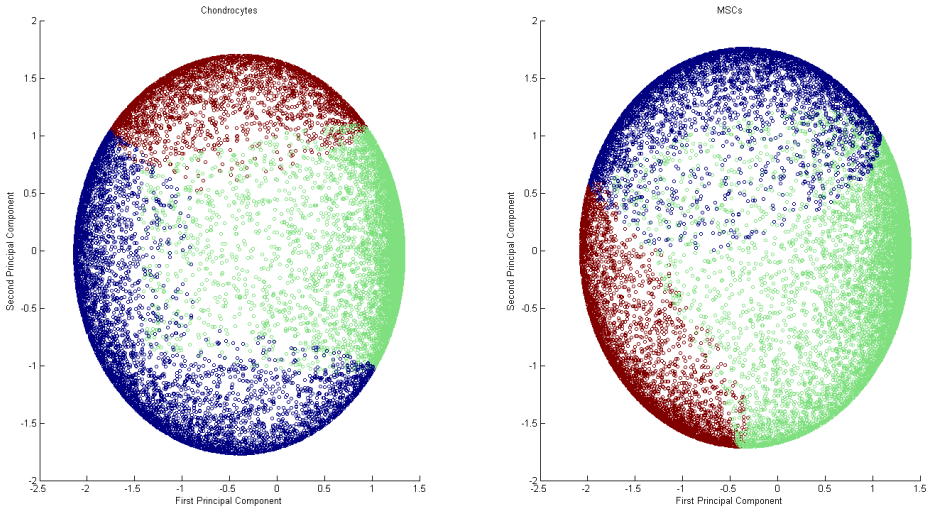
Ομάδα MSCs-Ομάδα Chondrocytes	**Πλήθος κοινών γονιδίων**	**Ποσοστό ομοιότητας**
1-1	1032	23.01%
1-2	131	2.92%
1-3	2266	50.52%
1-4	1056	23.55%
Σύνολο		**100%**
2-1	181	20.92%
2-2	90	10.40%
2-3	170	19.65%
2-4	424	49.05%
Σύνολο		**100%**
3-1	203	2.44%
3-2	1027	12.35%
3-3	377	4.53%
3-4	6712	80.68%
Σύνολο		**100%**
4-1	1794	45.77%
4-2	1453	37.07%
4-3	226	5.77%
4-4	447	11.40%
Σύνολο		**100%**

Τα ποσοστά ομοιότητας του παραπάνω πίνακα μας υποδεικνύουν την ένωση κάποιων ομάδων, τόσο των μεσεγχυματικών βλαστικών κυττάρων όσο και των

χονδροκυττάρων ώστε να αυξήσουμε την ομοιότητα. Έτσι, προχωρούμε στην ένωση των ομάδων 2 και 3 όσον αφορά την πρώτη κατηγορία και στην ένωση των ομάδων 1 και 2 για τη δεύτερη, διαμορφώνοντας τις ομάδες στον Πίνακα 12. Παράλληλα στην Εικόνα 41 παρουσιάζουμε το αποτέλεσμα της οπτικοποίησης των νέων ομάδων.

Πίνακας 12: Πλήθος τριών ομάδων χονδροκυττάρων και μεσεγχυματικών κυττάρων μετά την ένωση.

Αριθμός ομάδας	Πλήθος MSCs	Πλήθος Chondrocytes
1	4485	5911
2	9184	8639
3	3920	3039
Σύνολο	17589	17589



Εικόνα 41: Οπτικοποίηση PCA των μεσεγχυματικών βλαστικών κυττάρων και των χονδροκυττάρων σε 2-D μετά την ένωση.

Στον Πίνακα 13 παραθέτουμε τα αποτελέσματα των ποσοστών μεταξύ όλων των δυνατών συνδυασμών των ομάδων όπως αυτά προέκυψαν από την ένωση των συστάδων του προηγούμενου βήματος. Με έντονο συμβολισμό έχουμε τις καλύτερες αντιστοιχίες που προκύπτουν.

Οι αντιστοιχίσεις οι οποίες προκύπτουν είναι αυτές που κατέχουν το μεγαλύτερο ποσοστό ομοιότητας. Έχοντας αυτό το στοιχείο μπορούμε να προχωρήσουμε στην βιολογική ερμηνεία των αποτελεσμάτων όπως φαίνεται ακολούθως.

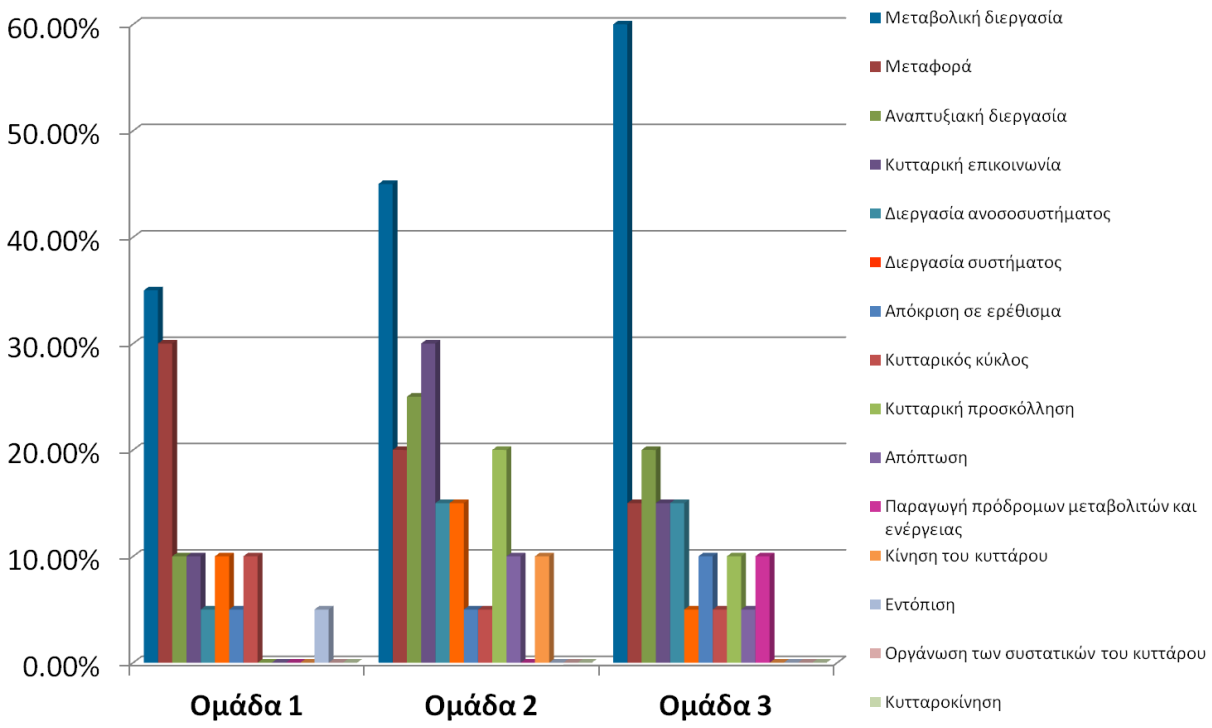
Βιολογική αξιολόγηση

Στο βήμα αυτό, έχοντας στη διάθεσή μας τρείς ομάδες από κάθε κατηγορία κυττάρων, εξάγουμε τα 20 κεντρικά γονίδια (Παράρτημα Β) καθεμιάς. Πληροφορίες για τα γονίδια αυτά που βασίζονται στην LVQ μέθοδο (καλύτερος δείκτης αξιολόγησης) και συμμετέχουν σε γνωστές βιολογικές διεργασίες και μονοπάτια, προέρχονται από τη χρήση του συστήματος ταξινόμησης PANTHER.

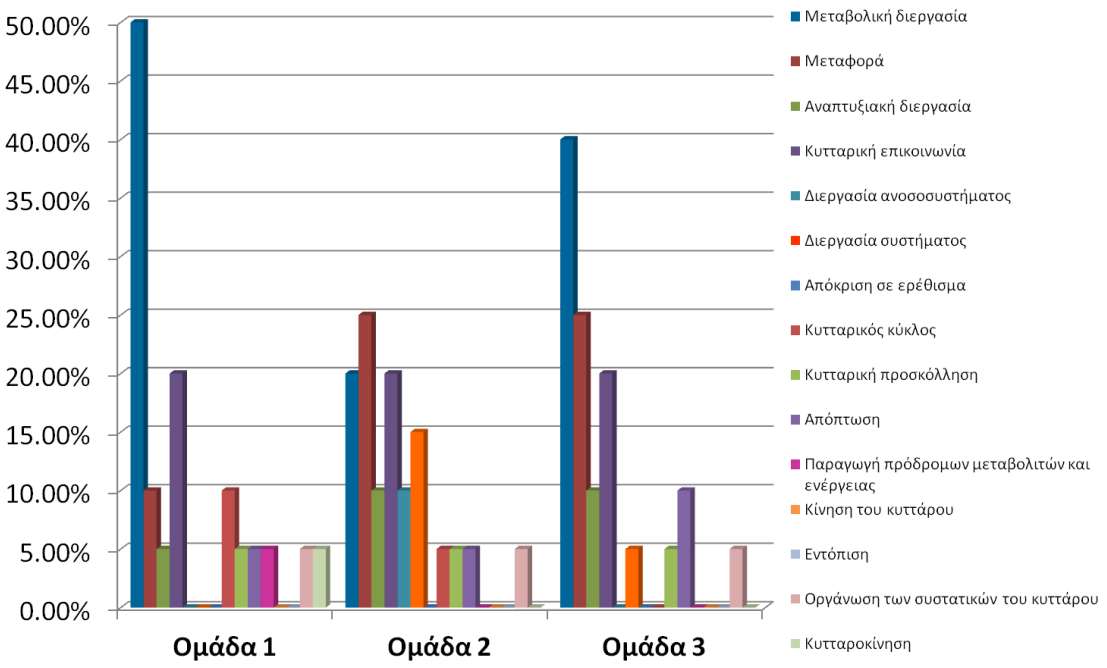
Πίνακας 13: Ποσοστά ομοιότητας για όλους τους συνδυασμούς των ομάδων (Chondocytes και MSCs), μετά την ένωση.

Ομάδα MSC-Ομάδα Chondrocytes	**Πλήθος κοινών γονιδίων**	**Ποσοστό ομοιότητας**
1-1	1163	25.39%
1-2	1056	23.55%
1-3	**2266**	**50.52%**
Σύνολο		**100%**
2-1	1501	16.34%
2-2	**7136**	**77.70%**
2-3	547	5.96%
Σύνολο		**100%**
3-1	**3247**	**82.83%**
3-2	447	11.40%
3-3	226	5.77%
Σύνολο		**100%**

Αρχικά μελετάται η κάθε διαμέριση κυττάρων (χονδροκύτταρα και μεσεγχυματικά βλαστικά κύτταρα) χωριστά. Ο στόχος είναι να ελέγξουμε τις ομάδες που τις απαρτίζουν σαν διακριτές οντότητες και να διερευνήσουμε αν αυτές χαρακτηρίζονται από κοινές ιδιότητες. Τα αποτελέσματα της μελέτης αυτής παρουσιάζονται στις Εικόνες 42 και 43.



Εικόνα 42: Βιολογικές Διεργασίες. Συγκριτικά αποτελέσματα από τις 3 ομάδες των χονδροκυττάρων.



Εικόνα 43: Βιολογικές Διεργασίες. Συγκριτικά αποτελέσματα από τις 3 ομάδες των μεσεγχυματικών βλαστικών κυττάρων.

Σε αυτό το σημείο θα πρέπει να τονίσουμε, ότι στις Εικόνες 42 και 43 απεικονίζονται συνολικά 15 βιολογικές διεργασίες στις οποίες συμμετέχουν ή όχι τα γονίδια των επιμέρους ομάδων των δυο κυτταρικών τύπων, προκειμένου να παρουσιάσουμε τις ομοιότητες και τις διαφορές των ομάδων του κάθε κυτταρικού τύπου ξεχωριστά αλλά και των δυο κυτταρικών τύπων μεταξύ τους. Αναλυτικότερα, στις Εικόνες 42 και 43 παρατηρούμε βιολογικές διεργασίες, από τις οποίες: α) 4 είναι κοινές για τα 20 γονίδια των δύο κυτταρικών τύπων, β) 4 είναι κοινές για τα 20 γονίδια των χονδροκυττάρων, γ) 3 είναι κοινές για τα 20 γονίδια των μεσεγχυματικών βλαστικών κυττάρων, δ) 3 διέπουν αποκλειστικά τα χονδροκύτταρα, ε) 2 διέπουν αποκλειστικά τα μεσεγχυματικά, στ) 5 είναι κοινές για τα 20 γονίδια μιας ή δυο ομάδων των χονδροκυττάρων ή μεσεγχυματικών κυττάρων, αντίστοιχα. Παράλληλα, στον Πίνακα 14 διακρίνονται με ευκρίνεια οι ομοιότητες και οι διαφορές του κάθε κυτταρικού τύπου ξεχωριστά αλλά και των δυο κυτταρικών τύπων μεταξύ τους.

Πίνακας 14: Ομοιότητες των βιολογικών διεργασιών μεταξύ των μεσεγχυματικών βλαστικών κυττάρων και των χονδροκυττάρων.

ΟΜΟΙΟΤΗΤΕΣ ΤΩΝ ΒΙΟΛΟΓΙΚΩΝ ΔΙΕΡΓΑΣΙΩΝ ΜΕΤΑΞΥ ΤΩΝ ΜΕΣΕΓΧΥΜΑΤΙΚΩΝ ΒΛΑΣΤΙΚΩΝ ΚΥΤΤΑΡΩΝ ΚΑΙ ΤΩΝ ΧΟΝΔΡΟΚΥΤΤΑΡΩΝ															
ΜΕΣΕΓΧΥΜΑΤΙΚΑ ΚΥΤΤΑΡΑ															
	Μεταβολική διεργασία	Μεταφορά	Αναπτυξιακή διεργασία	Κυτταρική επικοινωνία	Διεργασία ανοσο-συστήματος	Διεργασία συστήματος	Απόκριση σε ερέθισμα	Κυτταρικός κύκλος	Κυτταρική προσκόλληση	Απόπτωση	Παραγωγή πρόδρομων μεταβολιτών και ενέργειας	Κίνηση του κυττάρου	Εντόπιση	Οργάνωση των συστατικών του κυττάρου	Κυτταρο-κίνηση
Ομάδα 1	50,00%	10,00%	5,00%	20,00%	0,00%	0,00%	0,00%	10,00%	5,00%	5,00%	5,00%	0,00%	0,00%	5,00%	5,00%
Ομάδα 2	20,00%	25,00%	10,00%	20,00%	10,00%	15,00%	0,00%	5,00%	5,00%	5,00%	0,00%	0,00%	0,00%	5,00%	0,00%
Ομάδα 3	40,00%	25,00%	10,00%	20,00%	0,00%	5,00%	0,00%	0,00%	5,00%	10,00%	0,00%	0,00%	0,00%	5,00%	0,00%
ΧΟΝΔΡΟΚΥΤΤΑΡΑ															
	Μεταβολική διεργασία	Μεταφορά	Αναπτυξιακή διεργασία	Κυτταρική επικοινωνία	Διεργασία ανοσο-συστήματος	Διεργασία συστήματος	Απόκριση σε ερέθισμα	Κυτταρικός κύκλος	Κυτταρική προσκόλληση	Απόπτωση	Παραγωγή πρόδρομων μεταβολιτών και ενέργειας	Κίνηση του κυττάρου	Εντόπιση	Οργάνωση των συστατικών του κυττάρου	Κυτταρο-κίνηση
Ομάδα 1	35,00%	30,00%	10,00%	10,00%	5,00%	10,00%	5,00%	10,00%	0,00%	0,00%	0,00%	0,00%	5,00%	0,00%	0,00%
Ομάδα 2	45,00%	20,00%	25,00%	30,00%	15,00%	15,00%	5,00%	5,00%	20,00%	10,00%	0,00%	10,00%	0,00%	0,00%	0,00%
Ομάδα 3	60,00%	15,00%	20,00%	15,00%	15,00%	5,00%	10,00%	5,00%	10,00%	5,00%	10,00%	0,00%	0,00%	0,00%	0,00%

Ειδικότερα, παρατηρούμε ότι διεργασίες όπως ο μεταβολισμός, η μεταφορά, η αναπτυξιακή διεργασία, και η κυτταρική επικοινωνία αποτελούν τις 4 κοινές βιολογικές διεργασίες για τα 20 γονίδια όλων των επιμέρους ομάδων των δυο κυτταρικών τύπων (Εικόνες 42 και 43, Πίνακας 14). Ταυτόχρονα παρατηρούμε ότι τα 20 γονίδια κάθε ομάδας μεσεγχυματικών βλαστικών κυττάρων συμμετέχουν σε τρεις κοινές βιολογικές διεργασίες (κυτταρική προσκόλληση, απόπτωση, οργάνωση των συστατικών του κυττάρου), και τα 20 γονίδια κάθε ομάδας χονδροκυττάρων συμμετέχουν σε τέσσερις κοινές διεργασίες (διεργασία ανοσοσυστήματος, διεργασία συστήματος, απόκριση σε ερέθισμα, κυτταρικός κύκλος) (Εικόνες 42 και 43, Πίνακας 14). Επιπλέον, διεργασίες όπως η απόκριση σε ερέθισμα, η κίνηση του κυττάρου και η εντόπιση, συνιστούν τις τρεις βιολογικές διεργασίες που δεν εμφανίζονται σε καμία ομάδα των μεσεγχυματικών βλαστικών κυττάρων και εμφανίζονται σε κάποιες ομάδες των χονδροκυττάρων, ενώ διεργασίες όπως η οργάνωση των συστατικών του κυττάρου, και η κυτταροκίνηση συνιστούν τις δυο διεργασίες που δεν εμφανίζονται σε καμία ομάδα των χονδροκυττάρων και εμφανίζονται σε κάποιες ομάδες των μεσεγχυματικών κυττάρων (Εικόνες 42 και 43, Πίνακας 14). Συνοψίζοντας προκύπτει ότι: 1) τα 20 γονίδια όλων των ομάδων κάθε κυτταρικού τύπου συμμετέχουν σε 4 κοινές βιολογικές διεργασίες (βλέπε κόκκινο χρώμα, Πίνακας 14), 2) τα 20 γονίδια κάθε ομάδας μεσεγχυματικών βλαστικών κυττάρων συμμετέχουν σε 3 κοινές βιολογικές διεργασίες (βλέπε μπλε χρώμα, Πίνακας 14), 3) τα 20 γονίδια κάθε ομάδας χονδροκυττάρων συμμετέχουν σε 4 κοινές βιολογικές διεργασίες (βλέπε ροζ χρώμα, Πίνακας 14), και 4) τρεις βιολογικές διεργασίες εμφανίζονται αποκλειστικά στα χονδροκύτταρα και δυο βιολογικές διεργασίες εμφανίζονται αποκλειστικά στα μεσεγχυματικά βλαστικά κύτταρα (Εικόνες 42 και 43, βλέπε σκιαγράφηση στον Πίνακα 14).

Έτσι, παρατηρούμε ένα μεγάλο ποσοστό ομοιότητας μεταξύ των επιμέρους ομάδων για κάθε κυτταρικό τύπο αντίστοιχα, γεγονός που δεν επιτρέπει τη σαφή διάκριση των εκάστοτε ομάδων του κάθε κυτταρικού τύπου. Πιθανά να μπορούσε κανείς να διαχωρίσει τις επιμέρους ομάδες από την παρουσία ή απουσία συγκεκριμένων βιολογικών διεργασιών, όπως για παράδειγμα η ομάδα 1 των μεσεγχυματικών βλαστικών κυττάρων διαφοροποιείται από τις ομάδες 2 και 3 λόγω α) της παρουσίας των διεργασιών της κυτταροκίνησης (τελευταία σε σειρά εμφάνισης

στην Εικόνα 43 και τον Πίνακα 14) και της παραγωγής των πρόδρομων μεταβολιτών και ενέργειας καθώς και β) της απουσίας της διεργασίας συστήματος (ενδέκατη και έκτη σε σειρά εμφάνισης στην Εικόνα 43 και τον Πίνακα 14, αντίστοιχα).

Στη συνέχεια διερευνούμε τις σχέσεις που υπάρχουν μεταξύ των μεσεγχυματικών βλαστικών κυττάρων και των χονδροκυττάρων. Έπειτα από μελέτη καταλήγουμε σε ομοιότητες των βιολογικών διεργασιών μεταξύ των προαναφερθέντων κυττάρων (Πίνακας 14). Στον Πίνακα 14, παρατηρούμε ένα μεγάλο ποσοστό ομοιότητας μεταξύ των ομάδων των δυο κυτταρικών τύπων που οφείλεται στις 4 κοινές επικρατέστερες βιολογικές διεργασίες αλλά και σε 6 δευτερεύουσες υπολειπόμενες βιολογικές διεργασίες (διεργασία ανοσοσυστήματος, διεργασία συστήματος, κυτταρικός κύκλος, η κυτταρική προσκόλληση, απόπτωση, παραγωγή πρόδρομων μεταβολιτών και ενέργειας) στις οποίες συμμετέχουν τα γονίδια των περισσοτέρων ομάδων των δυο κυτταρικών τύπων και που ανακλάται και από τα ποσοστά ομοιότητας που έχουν προκύψει για όλους τους συνδυασμούς των ομάδων των δυο κυτταρικών τύπων (Πίνακας 13). Για παράδειγμα, η ομοιότητα της ομάδας 3 των μεσεγχυματικών βλαστικών κυττάρων με την ομάδα 1 των χονδροκυττάρων οφείλεται σε παρόμοια ποσοστά συμμετοχής των 4 επικρατέστερων βιολογικών διεργασιών όπως του μεταβολισμού (40% με 35%), της μεταφοράς (25% με 30%), της αναπτυξιακής διεργασίας (10% με 10%) αλλά και της δευτερεύουσας βιολογικής διεργασίας όπως της διεργασίας συστήματος (5% με 10%). Αξίζει να αναφερθεί, ότι η ομοιότητα σε ποσοστιαία αναλογία των μεσεγχυματικών βλαστικών κυττάρων της ομάδας 3 με τα χονδροκύτταρα της ομάδας 1 ανέρχεται σε 82.83% (3247 κοινά γονίδια) (Πίνακας 13).

Παράλληλα, οι διαφορές που αποτυπώνονται στον Πίνακα 13 οφείλονται γενικότερα στην παρουσία ή/και απουσία συγκεκριμένων βιολογικών διεργασιών (π.χ. απόκριση σε ερέθισμα, κυτταροκίνηση), διαφοροποιώντας με αυτό τον τρόπο τα χονδροκύτταρα από τα μεσεγχυματικά βλαστικά κύτταρα (Πίνακας 14).

Συμπερασματικά προκύπτει ότι οι βιολογικές διεργασίες αναδεικνύουν σε μεγάλο βαθμό τις ομοιότητες τόσο μεταξύ των ομάδων κάθε κυτταρικού τύπου όσο και μεταξύ των ομάδων των δυο κυτταρικών τύπων, ενώ φαίνεται να αποτυπώνουν σε μικρότερο βαθμό τη μεταξύ τους διαφοροποίηση.

Γι’ αυτό το λόγο θελήσαμε να διερευνήσουμε τα μονοπάτια (pathways) που διέπουν τις 3 ομάδες των 20 γονιδίων των μεσεγχυματικών βλαστικών κυττάρων και

των χονδροκυττάρων, αντίστοιχα. Στον Πίνακα 15 αναφέρουμε τις σχέσεις που εντοπίστηκαν στα μονοπάτια ανάμεσα στις 3 ομάδες του κάθε κυτταρικού τύπου και μεταξύ των μεσεγχυματικών βλαστικών κυττάρων και των χονδροκυττάρων. Η μελέτη των μοριακών μονοπατιών ανέδειξε την σαφή διάκριση τόσο μεταξύ των ομάδων κάθε κυτταρικού τύπου χωριστά όσο και μεταξύ των ομάδων των δυο κυτταρικών τύπων. Αξίζει να τονιστεί ότι ανάμεσα στα μονοπάτια που εντοπίστηκαν περιλαμβάνονται η βιοσύνθεση της χοληστερόλης, η διακίνηση συναπτικών κυστιδίων και το σηματοδοτικό μονοπάτι των ιντεγκρινών, μονοπάτια που τονίζονται και στην εργασία των Bernstein και συνεργατών [27]. Στην πρωτότυπη εργασία των Bernstein και συνεργατών [27], τα περισσότερα γονίδια που εκφράζονται με άνισο τρόπο (inequal) μεταξύ των μεσεγχυματικών βλαστικών κυττάρων και των χονδροκυττάρων κατά την χρονική διάρκεια των 3 πρώτων ημερών (π.χ. μεσεγχυματικά=υπερέκφραση, χονδροκύτταρα=μη σταθερή έκφραση) ανήκουν στο μονοπάτι λιπιδίων/χοληστερόλης, το μονοπάτι ακτίνης και ιντεγκρίνης καθώς και σε συστήματα μεταφοράς κυστιδίων. Επίσης, γονίδια που εμφανίζουν σταθερή ανομοιομορφία (constant dissimilarity) στην έκφρασή τους κατά την χρονική διάρκεια των 3 πρώτων ημερών μεταξύ των δυο κυτταρικών τύπων (π.χ. μεσεγχυματικά= υπερέκφραση, χονδροκύτταρα=υποέκφραση) ανήκουν στο μονοπάτι των ιντεγκρινών και το μεταβολισμού των λιπιδίων.

Σύμφωνα με τα παραπάνω και παρόλο που δεν έχει εισαχθεί η έννοια του χρόνου στη μέχρι τώρα μελέτη μας, φαίνεται ότι τόσο οι ομοιότητες μεταξύ των ομάδων των δυο κυτταρικών τύπων όπως φαίνεται από τα στατιστικά στοιχεία (π.χ. ομάδες 3-1, Πίνακας 13) και αποκαλύπτεται από τις εμπλεκόμενες βιολογικές διεργασίες (π.χ. ομάδες 2-2, Πίνακας 14), όσο και οι σαφείς διαφορές που παρατηρούνται στα μοριακά μονοπάτια μεταξύ των ομάδων των δυο κυτταρικών τύπων (π.χ. ομάδες 3-1, πίνακας 15), επιβεβαιώνουν εν μέρει τα πειραματικά δεδομένα από την αρχική μελέτη [27] συνιστώντας επιπρόσθετα στοιχεία, τα οποία θα μπορούσαν να χρησιμοποιηθούν για περαιτέρω διερεύνηση ή/και βελτίωση των συνθηκών που θα εξασφάλιζαν τη χονδρογενή διαφοροποίηση των μεσεγχυματικών βλαστικών κυττάρων σε χονδροκύτταρα .

Πίνακας 15: Διαφορές στα μονοπάτια ανάμεσα στις 3 ομάδες του κάθε κυτταρικού τύπου και μεταξύ των μεσεγχυματικών βλαστικών κυττάρων και των χονδροκυττάρων.

ΔΙΑΦΟΡΕΣ ΣΤΑ ΜΟΝΟΠΑΤΙΑ ΑΝΑΜΕΣΑ ΣΤΙΣ 3 ΟΜΑΔΕΣ ΤΟΥ ΚΑΘΕ ΚΥΤΤΑΡΙΚΟΥ ΤΥΠΟΥ ΚΑΙ ΜΕΤΑΞΥ ΤΩΝ ΜΕΣΕΓΧΥΜΑΤΙΚΩΝ ΒΛΑΣΤΙΚΩΝ ΚΥΤΤΑΡΩΝ ΚΑΙ ΤΩΝ ΧΟΝΔΡΟΚΥΤΤΑΡΩΝ		
ΜΕΣΕΓΧΥΜΑΤΙΚΑ ΚΥΤΤΑΡΑ		
Ομάδα 1	**Ομάδα 2**	**Ομάδα 3**
Αγγειογένεση	Βιοσύνθεση της αδρεναλίνης και της νοραδρεναλίνης	Σηματοδοτικό μονοπάτι FAS
Βιοσύνθεση χοληστερόλης	*Αγγειογένεση*	*νόσος του Huntington*
Βιοσύνθεση της αίμης	Σηματοδότηση του GABA-B υποδοχέα της ομάδας II	**Σηματοδοτικό μονοπάτι των ιντεγκρινών**
	Μονοπάτι σηματοδότησης της ετεροτριμερούς πρωτεΐνης-G -Gi άλφα και Gs άλφα μεσολαβούμενο μονοπάτι	
	Μονοπάτι σηματοδότησης της ετεροτριμερούς πρωτεΐνης-G -Gq άλφα και Go άλφα μεσολαβούμενο μονοπάτι	
	Σηματοδοτικό μονοπάτι της ιντερφερόνης-γάμμα	
	Σηματοδοτικό μονοπάτι JAK/STAT	
	Μονοπάτι σηματοδότησης των μουσκαρινικών υποδοχέων 2 και 4 της ακετυλοχολίνης	
	Βιοσύνθεση S-αδενοσυλο-μεθειονίνης	
	Διακίνηση συναπτικών κυστιδίων	
ΧΟΝΔΡΟΚΥΤΤΑΡΑ		
Ομάδα 1	**Ομάδα 2**	**Ομάδα 3**
Διακίνηση συναπτικών κυστιδίων	Μονοπάτι της προδυνορφίνης οπιοειδών	Σηματοδοτικό μονοπάτι απόπτωσης
	Καταρράκτης ενεργοποίησης του πλασμινογόνου	Καθοδήγηση των νευραξόνων μεσολαβούμενη από Slit/Robo
	Σύνθεση βαζοπρεσίνης	*Νόσος του Huntington*
		Αποικοδόμηση νικοτίνης
		Μονοπάτι p53 βρόχος ανάδρασης 1
		Σύνθεση βαζοπρεσίνης
		Μεταβολισμός και μεταβολικό μονοπάτι της βιταμίνης D
		Σηματοδοτικό μονοπάτι Wnt
		Μονοπάτι p53 με στέρηση γλυκόζης
		Μονοπάτι p53 βρόχος ανάδρασης 2
		Μονοπάτι p53

Τα μοριακά μονοπάτια που απαντώνται σε περισσότερες από μια ομάδες τόσο στα μεσεγχυματικά βλαστικά κύτταρα ή/και στα χονδροκύτταρα όσο και μεταξύ των δυο αυτών κυτταρικών τύπων έχουν σημειωθεί με πλάγια γραφή. Η έντονη γραφή καταδεικνύει τα μοριακά μονοπάτια που αναφέρονται στην εργασία των Bernstein και συνεργατών [27].

5.5 Εφαρμογή σε Οστεοαρθρίτιδα των μέτρων χρονικής εξέλιξης

Η μεθοδολογία που προτείνουμε με βάση μέτρα χρονικής εξέλιξης στο Κεφάλαιο 3, εφαρμόζεται στα δεδομένα του Bernstein [27] όπως και στην Ενότητα 5.4. Αντιμετωπίζουμε τα χονδροκύτταρα και τα μεσεγχυματικά βλαστικά κύτταρα ξεχωριστά και καταλήγουμε σε μία διαμέριση για κάθε μία ομάδα κυττάρων. Η φύση των δεδομένων (μετρήσεις σε 4 χρονικές στιγμές)και το πρόβλημα που έχουμε να διερευνήσουμε μας οδήγησαν να αντιμετωπίσουμε διαφορετικά την ομαδοποίηση των δεδομένων και να αναπτύξουμε την προτεινόμενη μεθοδολογία.

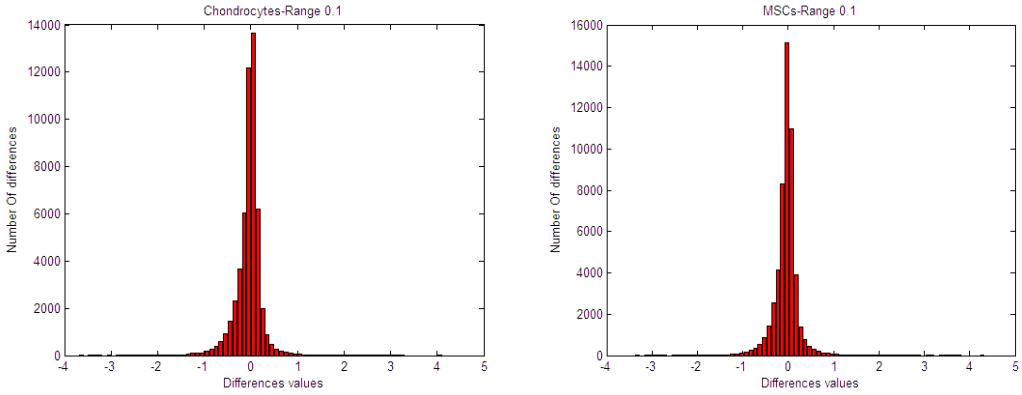
Περιγραφή της κωδικοποίησης των χρονικών μεταβολών

Έχουμε στη διάθεσή μας 17589 εκφρασμένα γονίδια σε 4 χρονικές στιγμές, με άλλα λόγια διαθέτουμε 17589 γονίδια και τα χρονικά προφίλ αυτών, τόσο για τα χονδροκύτταρα όσο και για τα μεσεγχυματικά βλαστικά κύτταρα. Το πρώτο βήμα που κάνουμε είναι να βρούμε τη κωδικοποιημένη χρονική μεταβολή του κάθε γονιδίου και στις δύο περιπτώσεις. Όπως περιγράφεται αναλυτικά στην Ενότητα 4.2 το μέγεθος των προτύπων χρονικής μεταβολής θα πρέπει να είναι N-1, άρα στη δική μας περίπτωση θα έχουμε κωδικούς μήκους 3 και το συνολικό πλήθος αυτών, λόγω της τριαδικής αναπαράστασης θα είναι 3^3=27. Τα 27 διανύσματα κωδικών που χρησιμοποιούμε στην εφαρμογή μας βρίσκονται στο Παράρτημα Β, και η σειρά εμφάνισης τους είναι ίδια με την σειρά τοποθέτησής του στους οριζόντιους άξονες των γραφημάτων που θα ακολουθήσουν.

Οι τιμές που θα φέρουν πλέον τα γονίδια σε κάθε bit της κωδικοποιημένης χρονικής μεταβολής προκύπτουν με τον τρόπο που περιγράφεται παρακάτω. Πρώτον υπολογίζουμε τις διαφορές μεταξύ δύο συνεχόμενων χρονικών στιγμών και δεύτερον αφού ολοκληρωθούν όλοι οι υπολογισμοί που αφορούν το βήμα αυτό δημιουργούμε ένα ιστόγραμμα (Εικόνα 44) που στον οριζόντιο άξονα παρουσιάζονται όλες αυτές οι διαφορές ξεκινώντας αριστερά με την μικρότερη μεταβολή και τελειώνοντας με την μεγαλύτερη στα δεξιά του άξονα, αυξανόμενες με ένα βήμα της τάξης του 0.1. Ο κατακόρυφος άξονας συμβολίζει το πλήθος των διαφορών που βρίσκονται σε κάθε bin. Παρατηρώντας λοιπόν τα διαγράμματα και στις δύο περιπτώσεις αποφασίζουμε

ότι μεταβολές κοντά στο 0 θα ισοδυναμούν με μηδενική αλλαγή ενώ διαφορές μεγαλύτερες του 0.1 υποδηλώνουν αύξηση, σε αντίθεση με εκείνες που είναι μικρότερες του -0.1 που σημαίνουν μείωση της χρονικής συμπεριφοράς. Συγκεκριμένα, με threshold στο 0.1:

$$c\{\delta_i\} = \begin{Bmatrix} -1 & if\ \delta_i \leq -0.1 \\ 0 & if -0.1 \leq \delta_i \leq +0.1 \\ 1 & if\ \delta_i \geq +0.1 \end{Bmatrix} \quad .$$



Εικόνα 44: Κατανομή των διαφορών. (αριστερά) χονδροκύτταρα,(δεξιά) μεσεγχυματικά.

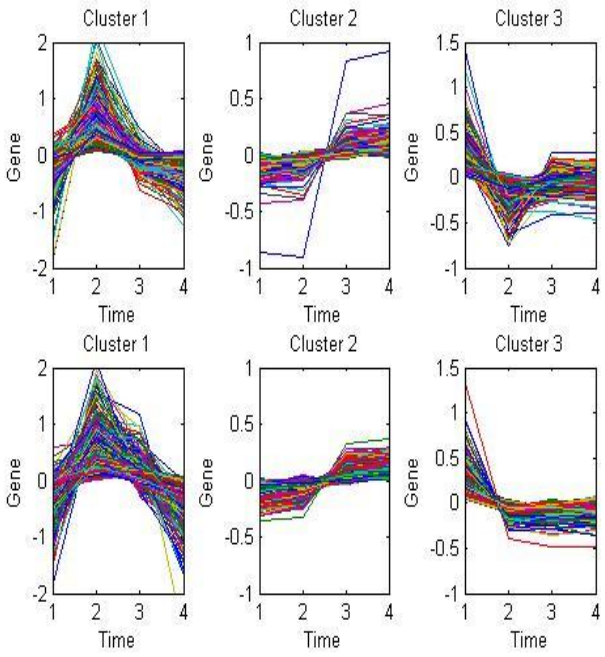
Πρακτικά μετά από την παραπάνω διαδικασία η βάση δεδομένων (μεσεγχυματικά βλαστικά κύτταρα και των χονδροκύτταρα) εμπλουτίζεται με ένα πίνακα διαστάσεων 17589x3, περιέχοντας για κάθε γονίδιο όχι το προφίλ έκφρασής του αλλά το κωδικοποιημένο διάνυσμα χρονικής μεταβολής του.

Μελέτη των δεικτών εγκυρότητας

Ο κατάλληλος αριθμός των ομάδων έχει μελετηθεί τόσο με μη αυτόματο (εποπτικό) τρόπο όσο και με τη χρήση της έννοιας των δεικτών εγκυρότητας (η ομαδοποίηση πραγματοποιήθηκε με τη μεθοδολογία που προτείνουμε στην Ενότητα 3.3, με χρήση 10x10=100 κόμβων στο βήμα 1). Στην εποπτική μελέτη εξετάζουμε τη χρονική μεταβολή του χρονικού προφίλ καθώς και τις ιδιότητες των ομάδων που έχουν συσταθεί από την άποψη των τιμών της έκφρασης και των χρονικών αλλαγών των δειγμάτων τους. Καταλήξαμε στο συμπέρασμα ότι το υπό εξέταση πρόβλημα συμφωνεί με τη δημιουργία 3 τύπων γονιδίων (clusters), τα οποία απεικονίζονται από τα προφίλ χρονικής έκφρασής τους στην Εικόνα 45, για τις περιπτώσεις των μεσεγχυματικών βλαστικών κυττάρων (άνω μέρος) και των χονδροκυττάρων (κάτω μέρος). Η ομαδοποίηση στην εικόνα αυτή παρουσιάζει τις δύο πιο σημαντικές

κωδικοποιημένες χρονικές μεταβολές (που εκφράζουν τις μεγαλύτερες πιθανότητες κωδικών λέξεων-συμβολοσειρών), από κάθε ομάδα, αλλά είναι ενδεικτική των τάσεων επίδοσης που παρατηρήθηκαν στις βάσεις δεδομένων. Σαν πιθανότητα κάθε κωδικοποιημένου διανύσματος σε κάθε ομάδα, ορίζουμε το λόγο του πλήθους των γονιδίων που αντιπροσωπεύονται από τον εκάστοτε κωδικό προς το συνολικό αριθμό γονιδίων ολόκληρης της ομάδας.

Οι παραδοσιακοί δείκτες εγκυρότητας , όπως ο Davies Bouldin ή ο Dunn δείκτης , αδυνατούν να παρέχουν ένα συνεκτικό επιχείρημα σχετικά με τον αριθμό ομάδων. Συγκεκριμένα ο DB δείκτης που μοιάζει με τη φιλοσοφία του RSI στο επίπεδο των εκφράσεων, επιτυγχάνει τις χαμηλότερες τιμές για 2 ομάδες στην προτεινόμενη μεικτή ομαδοποίηση με βάση τη χρονική εξέλιξη (proposed mixed clustering scheme), τόσο για τα μεσεγχυματικά βλαστικά κύτταρα όσο και για τα χονδροκύτταρα. Για την ομαδοποίηση με βάση τα διανύσματα κωδικοποιημένων μεταβολών (rank-based clustering scheme), ο DB παρουσιάζει διακυμάνσεις και ελαχιστοποιείται για αρκετά μεγάλο αριθμό ομάδων. Οι διακυμάνσεις του δείκτη συνεχίζονται για αριθμό ομάδων μεγαλύτερο του 3, περιπτώσεις για τις οποίες εμφανίζονται αποτελέσματα στον Πίνακα 16. Τέλος, για την ομαδοποίηση που βασίζεται μόνο στις τιμές έκφρασης, ο DB ελαχιστοποιείται για 3 και 4 ομάδες στα μεσεγχυματικά βλαστικά κύτταρα και στα χονδροκύτταρα, αντίστοιχα.



Εικόνα 45: Προφίλ έκφρασης σε τρεις ομάδες για τις περιπτώσεις των MSCs(άνω) και Chondrocytes (κάτω) ιστών. Προτεινόμενο μεικτό κριτήριο ομαδοποίησης.

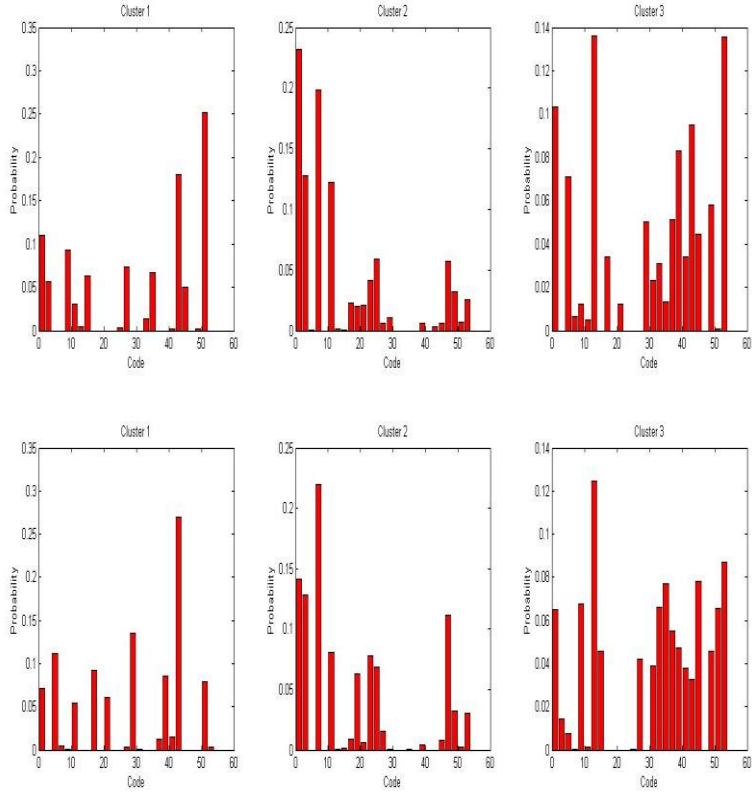
Αντίθετα, ο προτεινόμενος δείκτης RSI παρουσιάζει ισχυρή προτίμηση και υποδεικνύει έναν ιδανικό αριθμό από 3 ομάδες στις περισσότερες μελετώμενες προσεγγίσεις. Πιο συγκεκριμένα, για την ομαδοποίηση που βασίζεται στις τιμές έκφρασης των γονιδίων και την προτεινόμενη συνδυαστική ομαδοποίηση, ο RSI δείκτης ελαχιστοποιείται σε 3 ομάδες και για τις δύο βάσεις δεδομένων (μεσεγχυματικά βλαστικά κύτταρα και χονδροκύτταρα). Η κατάσταση διαφοροποιείται ελαφρώς για την rank-based ομαδοποίηση, με τον δείκτη RSI να αποκτά μικρές τιμές στις 2 ομάδες. Στην περίπτωση αυτή ο RSI δείκτης ελαχιστοποιείται για ένα μεγάλο αριθμό ομάδων (ίσο με 8), στην οποία περίπτωση 8 κύρια διανύσματα κωδικοποιημένων χρονικών μεταβολών εμφανίζονται χωριστά σε μία μόνο ομάδα το καθένα. Ο RSI δείκτης φαίνεται στον Πίνακα 16, μαζί με τον αντίστοιχο δείκτη DB (σε παρενθέσεις).

Αξίζει να σημειωθεί ότι ο RSI δείκτης συνάδει περισσότερο με τη συνδυαστική ομαδοποίηση με βάση το χρόνο, ενώ ο DB δείκτης βασίζεται στην “παραδοσιακή ομαδοποίηση” που στηρίζεται στις τιμές έκφρασης. Με τις κυριότερες χρονικές μεταβολές που εμφανίζονται στις ομάδες της προτεινόμενης μεθοδολογίας στην Εικόνα 45 και της παραδοσιακής ομαδοποίησης με βάση την έκφραση στην Εικόνα 47, μπορούμε οπτικά να διαπιστώσουμε ότι η προτεινόμενη μεθοδολογία έχει ως αποτέλεσμα πιο συνεκτικές ομάδες από την παραδοσιακή ομαδοποίηση και αυτό αντικατοπτρίζεται καλύτερα στον RSI δείκτη από ότι στον δείκτη DB. Στις εικόνες αυτές σχεδιάζουμε το πρότυπο έκφρασης στην πάροδο του χρόνου (4 ημέρες) για τις δύο πιο συχνά εμφανιζόμενες τάσεις σε κάθε ομάδα (σύμφωνα με τα αντίστοιχα ιστογράμματα στις Εικόνες 46 και 48). Παρουσιάζουμε τα γραφήματα για την προτεινόμενη ομαδοποίηση και για την παραδοσιακή προσέγγιση βασισμένη στις εκφράσεις. Παρατηρούμε ότι στην ομαδοποίηση εκφράσεων (Εικόνα 47), η χρήση μιας συμμετρικής συνάρτησης απόστασης (l_2 norm) οδηγεί στο να συμπεριληφθούν συμμετρικά πρότυπα σε μια ομάδα. Από την άλλη πλευρά, η προτεινόμενη προσέγγιση (Εικόνα 45) που υπολογίζει τη χρονική τάση, ευνοεί μόνο τις περιπτώσεις που έχουν παρόμοιες χρονικές συμπεριφορές στην πάροδο του χρόνου. Το ζήτημα της συμμετρίας επίσης μεταφέρεται και στους παραδοσιακούς δείκτες εγκυρότητας, όπως ο DB, που ευνοεί την τελευταία αυτή περίπτωση (συμμετρία) χωρίς να εξετάζει

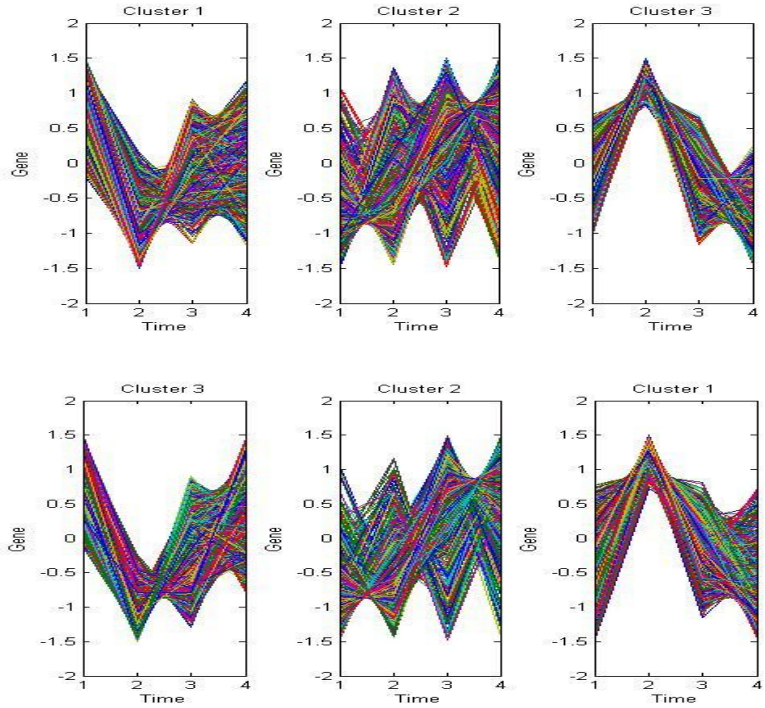
τη βιολογική διαφορά που συνεπάγεται από τις αντίθετες χρονικές τάσεις, οι οποίες προκαλούν τις συμμετρικές αποστάσεις από το κέντρο βάρους της ομάδας.

Πίνακας 16: RSI δείκτης για διάφορες προσεγγίσεις και τύπους κυττάρων: ο αντίστοιχος δείκτης DB εμφανίζεται σε παρενθέσεις.

Δείκτης RSI(DB)για κάθε μεθοδολογία ομαδοποίησης		**Αριθμός ομάδων**	
Μεικτή ομαδοποίηση (mixed clustering)	***2 ομάδες***	***3 ομάδες***	***4 ομάδες***
MSC	1.1546 (1.1839)	0.9118 (1.3646)	1.3065 (1.2334)
Chondrocytes	1.0058 (0.9997)	0.8906 (1.0553)	1.1141 (1.1794)
Ομαδοποίηση κωδικοποιημένων μεταβολών (rank-based clustering)	***2 ομάδες***	***3 ομάδες***	***4 ομάδες***
MSC	1.0648 (1.7126)	1.5489 (1.9712)	1.6220 (1.7621)
Chondrocytes	0.9066 (1.3281)	0.9104 (2.1713)	1.3674 (1.5124)
Ομαδοποίηση εκφράσεων (expression clustering)	***2 ομάδες***	***3 ομάδες***	***4 ομάδες***
MSC	1.2793 (1.1818)	1.1119 (0.9257)	1.4488 (1.4667)
Chondrocytes	1.1602 (0.9906)	1.1089 (0.9780)	1.3014 (0.9065)



Εικόνα 46: Ιστογράμματα ομάδων κωδικοποιημένων συμβολοσειρών (mixed clustering), αντιστοιχισμένα για τις περιπτώσεις των MSCs (άνω) και Chondrocytes (κάτω).



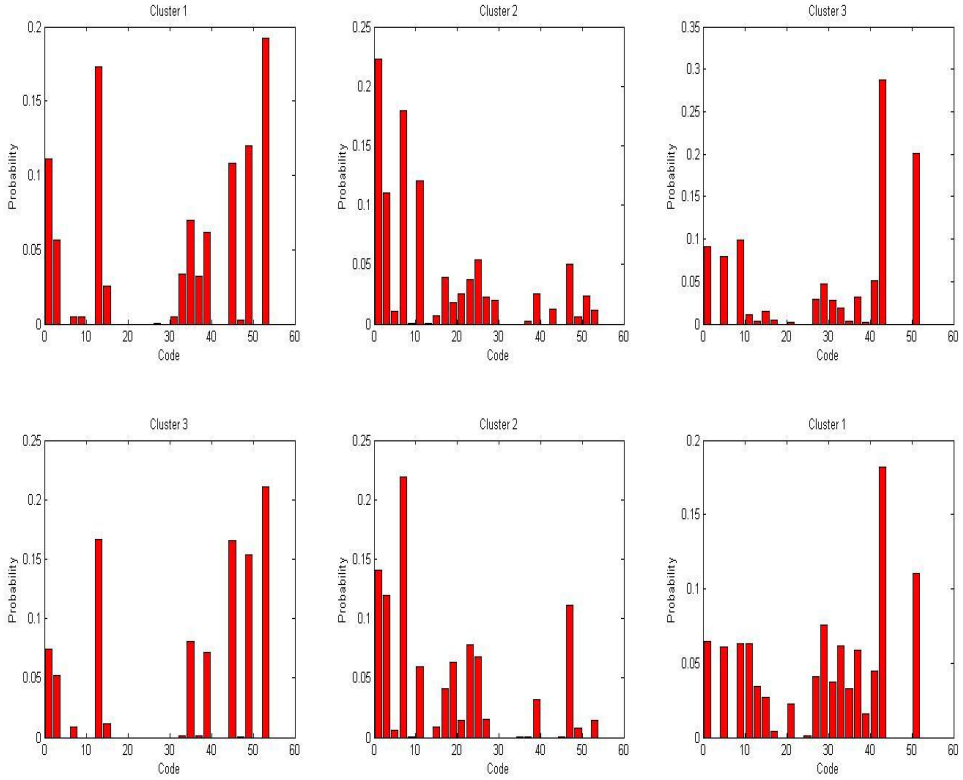
Εικόνα 47: Προφίλ έκφρασης σε τρεις ομάδες για τις περιπτώσεις των MSCs(άνω) και Chondrocytes (κάτω) ιστών. Ομαδοποίηση με βάση τις εκφράσεις.

Σύγκριση των τριών τύπων ομαδοποίησης

Στο σημείο αυτό παρουσιάζουμε και συζητάμε τα αποτελέσματα των τριών προσεγγίσεων ομαδοποίησης που μελετήσαμε.

Α) Ομαδοποίηση με βάση τα αρχικά εκφρασμένα πρότυπα(clustering based on initial expression patterns)

- Η ομαδοποίηση αυτή έχει προκύψει με τον τρόπο που εργαστήκαμε στην ενότητα 5.4. Αυτό το σύστημα ομαδοποίησης εξάγει τρεις ομάδες με βάση μόνο τα πρότυπα έκφρασης των δειγμάτων. Με σκοπό να συγκριθεί με τις άλλες μεθοδολογίες , εξάγουμε τις χρονικά κωδικοποιημένες συμβολοσειρές των ήδη ομαδοποιημένων δειγμάτων και παρουσιάζουμε τις ομάδες (Εικόνα 48) σε σχέση με τις συχνότητες ιστογράμματος (πιθανότητες) των κωδικοποιημένων συμβολοσειρών που συμμετέχουν σε κάθε ομάδα. Παράλληλα, στον Πίνακα 17 εμφανίζουμε τον δείκτη αντιστοίχισης μεταξύ των δύο διαμερίσεων (όπως αναπτύχθηκε στην Ενότητα 3.4.1), τονίζοντας τις αντιστοιχίσεις με τον μικρότερο δείκτη RSI, άρα τη μεγαλύτερη ομοιότητα. Στο Παράρτημα Β παραθέτουμε έναν πίνακα ο οποίος περιέχει το ποσοστό κοινών γονιδίων μεταξύ όλων των δυνατών συνδυασμών των ομάδων. Διαπιστώνεται από την Εικόνα 48 ότι κάθε ομάδα απεικονίζει μία κυρίαρχη μορφή παράστασης, αλλά σε ορισμένες περιπτώσεις με τις αντίθετες χρονικές τάσεις (αντίθετο πρόσημο χρονικής προβολής). Κάθε ομάδα εκφράζει μεγάλες πιθανότητες σε πολλά bins που αντιπροσωπεύουν συμβολοσειρές που διαφέρουν επί το πλείστον σε ένα ψηφίο, αλλά σε πολλές περιπτώσεις με ψηφία αντίθετων πρόσημων. Επιπλέον, κάθε κωδικοποιημένη συμβολοσειρά έχει διαχωριστεί και διανεμηθεί σε διάφορες ομάδες, με αποτέλεσμα να δημιουργούνται ομάδες με μεικτό περιεχόμενο. Συγκρίνοντας τις αντίστοιχες ομάδες των μεσεγχυματικών βλαστικών κυττάρων και των χονδροκυττάρων, δείχνουν παρόμοιο περιεχόμενο αλλά με διαφορετική αναλογία. Υπενθυμίζουμε ότι τα χρονικά προφίλ των επικρατέστερων προτύπων φαίνονται στην Εικόνα 47.



Εικόνα 48: Ιστογράμματα ομάδων κωδικοποιημένων συμβολοσειρών (expression clustering), αντιστοιχισμένα για τις περιπτώσεις των MSCs (άνω) και Chondrocytes (κάτω).

Πίνακας 17: Δείκτες αντιστοίχισης ομάδων των διαμερίσεων (expression clustering).

Ομάδα MSC-Ομάδα Chondrocytes	Τιμή δείκτη ομοιότητας
1-1	1.5039
1-2	1.5232
1-3	**0.2696**
2-1	1.4593
2-2	**0.4472**
2-3	1.6217
3-1	**0.5632**
3-2	1.7165
3-3	1.8057

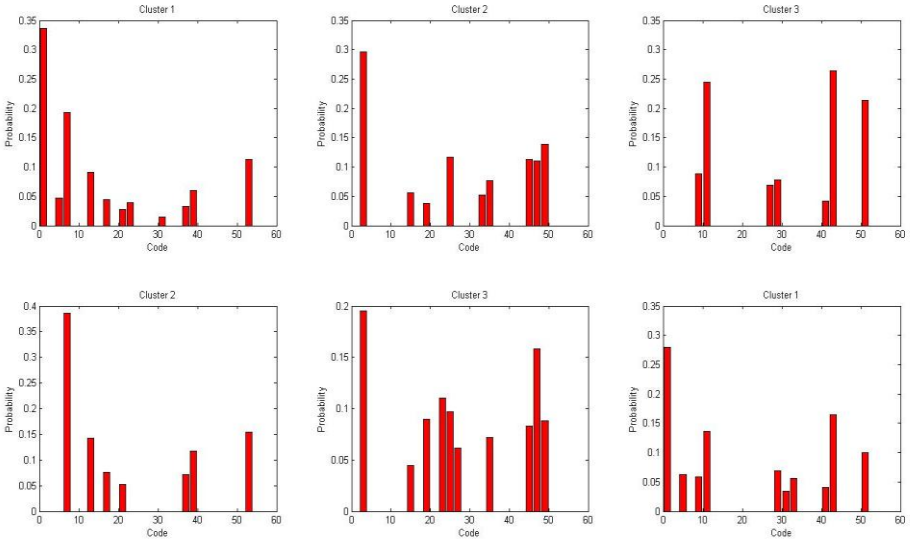
Β) Ομαδοποίηση κωδικοποιημένων μεταβολών (ranked-based clustering)

- Στην περίπτωση αυτή το σύστημα ομαδοποίησης λειτουργεί αποκλειστικά με τις χρονικά κωδικοποιημένες συμβολοσειρές. Οι τρείς ομάδες που προκύπτουν για κάθε βάση δεδομένων περιλαμβάνουν συμβολοσειρές των οποίων τα ιστογράμματα παρουσιάζονται στην Εικόνα 49. Κάθε συγκεκριμένη συμβολοσειρά τοποθετείται σε μία μόνο ομάδα, αλλά κάθε ομάδα περιλαμβάνει διαφορετικούς τύπους από ανόμοιες παραστάσεις ακόμη και αν χαρακτηρίζεται από μία (ή δύο) κυρίαρχες συμβολοσειρές. Η ομαδοποίηση με βάση τις κωδικοποιημένες μεταβολές αποτυγχάνει να οργανώσει μαζί δείγματα με μάλλον παρόμοια έκφραση και λίγο διαφορετική εξέλιξη (profile), δίνοντας μεγαλύτερη βαρύτητα στην κωδικοποιημένη χρονική εξέλιξη. Λόγω της ψηφιοποίησης με χρήση threshold, μικρές διαφοροποιήσεις στο προφίλ μεταβολής μπορεί να καταλήξουν σε μεγάλες διαφορές κωδικοποιημένων συμβολοσειρών. Αυτή η κβάντιση τιμών μπορεί να καταλήξει στην παράσταση παρόμοιων δειγμάτων με διαφορετικές συμβολοσειρές και στην τοποθέτησή τους σε διαφορετικές ομάδες. Το χαρακτηριστικό αυτό επηρεάζει και τον δείκτη RSI, ο οποίος λαμβάνει μικρότερες τιμές σε 2 ομάδες και για τους δύο τύπους κυττάρων (μεσεγχυματικά βλαστικά κύτταρα και χονδροκύτταρα). Μια επιπλέον πτυχή που σχετίζεται με την ομοιότητα των ομάδων στους δύο τύπους ιστών είναι ο άνισος αριθμός των δειγμάτων στις αντίστοιχες ομάδες, γεγονός που συνεπάγεται μια άνιση διάταξη των δειγμάτων στους δύο τύπους των ιστών. Στον Πίνακα 18 παραθέτουμε τον δείκτη αντιστοίχισης των ζευγαριών μεταξύ των δύο διαμερίσεων με βάση τη χρονική μεταβολή. Στο Παράρτημα Β επίσης παραθέτουμε έναν πίνακα ο οποίος περιέχει το ποσοστό κοινών γονιδίων μεταξύ όλων των δυνατών συνδυασμών των ομάδων.

Γ) Προτεινόμενη μεικτή ομαδοποίηση επηρεασμένη από χρονικές μεταβολές (proposed mixed clustering influenced by shape)

- Η προτεινόμενη μέθοδος επιτυγχάνει το μικρότερο δείκτη RSI και για τις δύο βάσεις δεδομένων (μεσεγχυματικά βλαστικά κύτταρα και χονδροκύτταρα) που επιτυγχάνεται για 3 ομάδες. Τα ιστογράμματα των ομάδων που προέκυψαν παρουσιάζονται στην Εικόνα 46. Οι ομάδες αποδεικνύονται

συμπαγείς τόσο από τα ιστογράμματα των χρονικά κωδικοποιημένων συμβολοσειρών όσο και τα προφίλ των προτύπων έκφρασης (Εικόνα 46 και 45 αντίστοιχα). Κάθε ομάδα περιλαμβάνει συμβολοσειρές με παρόμοια μορφή



Εικόνα 49: Ιστογράμματα ομάδων κωδικοποιημένων συμβολοσειρών (ranked-based clustering), αντιστοιχισμένα για τις περιπτώσεις των MSCs (άνω) και Chondrocytes (κάτω).

Πίνακας 18: Δείκτες αντιστοίχισης ομάδων των διαμερίσεων (ranked-based clustering).

Ομάδα MSC-Ομάδα Chondrocytes	Τιμή δείκτη ομοιότητας
1-1	1.3178
1-2	**0.8752**
1-3	1.9210
2-1	1.8948
2-2	2
2-3	**0.5415**
3-1	**0.8657**
3-2	2
3-3	1.8765

(τουλάχιστον κατά τις πρώτες μέρες), ενώ παρόμοιες συμβολοσειρές εμφανίζεται να συγκεντρώνονται σε μία ομάδα. Με την ενσωμάτωση της χρονικής κωδικοποίησης, ο αλγόριθμος ομαδοποίησης καταφέρνει να αποφύγει τη διάσπαση μιας συμβολοσειράς κωδικών σε πολλές ομάδες , ενώ

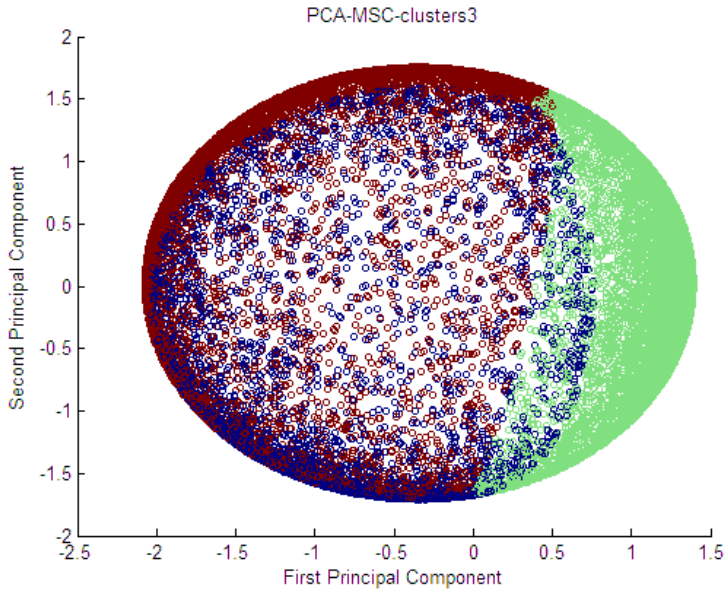
επίσης διατηρεί τις αρχικές σχέσεις με βάση την έκφραση των δειγμάτων. Οι αντίστοιχες ομάδες έχουν παρόμοιο περιεχόμενο όσο αφορά τους κωδικούς αλλά επίσης περιέχουν συγκρίσιμο αριθμό δειγμάτων για κάθε συμβολοσειρά. Στον Πίνακα 19 παραθέτουμε τον δείκτη αντιστοίχισης, με βάση τη χρονική μεταβολή, μεταξύ των δύο διαμερισμάτων. Στο Παράρτημα Β παραθέτουμε έναν πίνακα ο οποίος περιέχει το ποσοστό κοινών γονιδίων μεταξύ όλων των δυνατών συνδυασμών των ομάδων.

Πίνακας 19: Δείκτες αντιστοίχισης ομάδων των διαμερίσεων (mixed clustering).

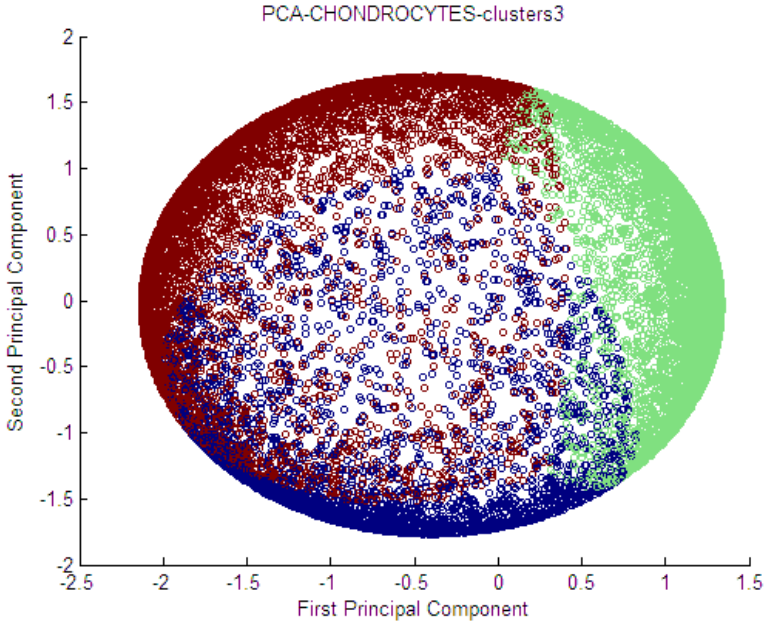
Ομάδα MSC-Ομάδα Chondrocytes	Τιμή δείκτη ομοιότητας
1-1	**1.2661**
1-2	1.5379
1-3	1.0556
2-1	1.5831
2-2	**0.3651**
2-3	1.6631
3-1	1.0810
3-2	1.5863
3-3	**0.7577**

Βιολογική αξιολόγηση των Αποτελεσμάτων

Σε αυτό το σημείο είναι ενδιαφέρον να συγκρίνουμε τις ομάδες που προκύπτουν για τα μεσεγχυματικά βλαστικά κύτταρα και τα χονδροκύτταρα με τη χρήση της προτεινόμενης προσέγγισης ομαδοποίησης. Ένας οπτικός χάρτης των τριών ομάδων σε κάθε περίπτωση, με βάση τις δύο κύριες συνιστώσες από την τεχνική PCA, απεικονίζεται στην Εικόνα 50 και 51.



Εικόνα 50: Οπτικοποίηση των τριών ομάδων των MSCs σε 2-D (mixed clustering).



Εικόνα 51: Οπτικοποίηση των τριών ομάδων των Chondrocytes σε 2-D (mixed clustering).

Φαίνεται ότι ο δύο περιπτώσεις επιφέρουν παρόμοιες συγκεντρώσεις γονιδίων στις τρεις ομάδες. Η ανάμειξη των ομάδων υπάρχει και στα μεσεγχυματικά βλαστικά κύτταρα και στα χονδροκύτταρα, ακόμα και αν φαίνεται σε διαφορετικές αναλογίες. Έτσι, η μελέτη μας δείχνει μια καλή ομοιότητα των ομάδων των (μεσεγχυματικά βλαστικά κύτταρα και χονδροκύτταρα) δεδομένων, η οποία επαληθεύει τα αποτελέσματα της αρχικής μελέτης του Bernstein. Παρόλα αυτά, οι μικρές περιοχές των ορίων των ομάδων χρειάζονται περαιτέρω εξέταση, δεδομένου ότι η ανάμειξη ομάδων είναι αρκετά σημαντική, πέρα από την ανοχή των σφαλμάτων που

οφείλονται στον θόρυβο μέτρησης. Σε μια περαιτέρω προσπάθεια να συγκρίνουμε τους δύο τύπους κυττάρων σε μια βαθύτερη βιολογική βάση, συγκρίνουμε τρία κεντρικά γονίδια (Παράρτημα Β) από τα μεσεγχυματικά βλαστικά κύτταρα και τα χονδροκύτταρα, από κάθε ζεύγος συνδυασμένων ομάδων, μέσω του συστήματος ταξινόμησης PANTHER [28]. Δεν προχωρούμε σε εκτεταμένη βιολογική ανάλυση λόγω του ότι οι ομάδες εμφανίζονται συμπαγείς οπότε αντί για 20 γονίδια εξετάζουμε τα 3 πιο κεντρικά γονίδια από κάθε ομάδα.

Με βάση τα ευρήματα της πρώτης ομάδας, τόσο τα μεσεγχυματικά βλαστικά κύτταρα όσο και τα χονδροκύτταρα περικλείουν τις βιολογικές διεργασίες του μεταβολισμού και της κυτταρικής επικοινωνίας. Πιο συγκεκριμένα, τα μεσεγχυματικά βλαστικά κύτταρα χαρακτηρίζονται από περαιτέρω αναπτυξιακές διεργασίες, όπως την ανάπτυξη έξω βλαστικού δέρματος, μεσοβλάστη και νευρικού συστήματος. Από την άλλη πλευρά, τα χονδροκύτταρα εμπλέκονται στον κυτταρικό κύκλο, στην διεργασία νευρολογικού συστήματος και στην μεταφορά. Όλες αυτές οι βιολογικές διεργασίες εξελίσσονται με το χρόνο, γεγονός που εξηγεί και την αλλαγή της συμπεριφοράς αυτής της ομάδας, όπως φαίνεται στην Εικόνα 45 (αριστερά). Σημειώνεται εντούτοις ότι οι υπεύθυνες βιολογικές διεργασίες για αυτήν την συμπεριφορά είναι διαφορετικές στους δύο τύπους κυττάρων. Όσον αφορά τη δεύτερη κατηγορία, τόσο τα μεσεγχυματικά βλαστικά κύτταρα όσο και τα χονδροκύτταρα εμπλέκονται στον μεταβολισμό, αλλά τα πρώτα είναι πιο συσχετισμένα με τη μεταβολική διεργασία των υδατανθράκων και των πρωτεϊνών, ενώ τα χονδροκύτταρα σχετίζονται με το μεταβολισμό των νουκλεοτιδικών βάσεων καθώς και με άλλες βιολογικές διεργασίες που δεν είναι κοινές με τα μεσεγχυματικά βλαστικά κύτταρα. Αυτές οι κύριες βιολογικές διεργασίες παρατηρούμε ότι δεν εκφράζουν στις συγκεκριμένες συνθήκες, σημαντική αλλαγή με την πάροδο του χρόνου, γεγονός που εξηγεί και τις μικρές αλλαγές στα χρονικά προφίλ που παρατηρήθηκαν στην Εικόνα 45 (μεσαία στήλη). Τέλος, η τρίτη κατηγορία περιλαμβάνει μεταβολικές διεργασίες, αναπτυξιακές και διεργασίες μεταφοράς, κοινές για τους δύο τύπους κυττάρων. Επιπλέον, τα μεσεγχυματικά βλαστικά κύτταρα περιλαμβάνουν την ομοιοστατική διεργασία, ενώ τα χονδροκύτταρα περιλαμβάνουν την οργάνωση κυτταρικών συστατικών. Τα τελευταία στοιχεία σταθεροποιούνται με το χρόνο και είναι περισσότερο υπεύθυνα για τη σύγκλιση των τάσεων

(σταθεροποίηση έκφρασης) των δύο τύπων κυττάρων που παρατηρήθηκαν στην Εικόνα 45 (δεξιά στήλη). Συνολικά, οι αντιστοιχιζόμενες ομάδες των δύο τύπων κυττάρων (MSCs και Chondrocytes) εκφράζουν παρόμοιες χρονικές τάσεις, αλλά οι υπεύθυνες βιολογικές διεργασίες για την συμπεριφορά αυτή μπορεί να είναι διαφορετικές σε κάθε μία, γεγονός που απαιτεί βαθύτερη βιολογική ερμηνεία.

Κεφάλαιο 6

Συμπεράσματα και Μελλοντικές Επεκτάσεις

6.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήσαμε διάφορες μεθοδολογίες ομαδοποίησης και οπτικοποίησης δεδομένων. Η κάθε μεθοδολογία ομαδοποίησης που αποτελούσε συνδυασμό αλγορίθμων, μας παρείχε διαφορετικά αποτελέσματα για το εκάστοτε πρόβλημα τα οποία αξιολογήσαμε στατιστικά αλλά και ερμηνεύσαμε βιολογικά.

Εφαρμόσαμε την πειραματική διαδικασία σε ένα τεχνητό πληθυσμό που αποτελούταν από τρεις ανεξάρτητους Gaussian πληθυσμούς. Τα αποτελέσματα που λάβαμε ήταν τα αναμενόμενα επιθυμητά. Τόσο η ομαδοποίηση αυτών όσο και η οπτικοποίηση άγγιξε τα όρια του βέλτιστου, κρίνοντας την απόδοση των δεικτών εγκυρότητας σε συνδυασμό με την τελική απεικόνιση των πληθυσμών σε δύο διαστάσεις.

Ικανοποιητικά ήταν και τα στατιστικά αποτελέσματα που προέκυψαν μετά την εφαρμογή των μεθοδολογιών στη βάση δεδομένων του σακχαρομύκητα. Τα αποτελέσματα αυτά συνάδουν με εκείνα της λειτουργικής ταξινόμησης των 20 αντιπροσωπευτικών γονιδίων για κάθε μία από τις τρείς ομάδες που καταλήξαμε. Παρόλο που τα γονίδια της κάθε ομάδας συμμετέχουν σε παρόμοιες βιολογικές διεργασίες, επικρατεί μία βιολογική διεργασία σε κάθε ομάδα που την καθιστά διαχωρίσιμη από τις υπόλοιπες.

Η εφαρμογή της μεθοδολογίας στους δύο τύπους κυττάρων (χονδροκύτταρα και μεσεγχυματικά βλαστικά κύτταρα) μας έδωσαν αποτελέσματα υψηλής απόδοσης τόσο στον τομέα της ομαδοποίησης όσο και στην οπτικοποίηση. Προχωρώντας στη βιολογική αξιολόγηση των δύο διαμερίσεων που προέκυψαν, με τρεις ομάδες η καθεμιά, παρατηρήσαμε: α) ομοιότητα σε βιολογικές διεργασίες μεταξύ των χονδροκυττάρων και των μεσεγχυματικών βλαστικών κυττάρων και β) διαφορές στα μονοπάτια (pathways) τόσο μεταξύ των χονδροκυττάρων και των μεσεγχυματικών βλαστικών κυττάρων όσο και μεταξύ των ομάδων κάθε κυτταρικού τύπου.

Η στατιστική αξιολόγηση των προφίλ χρονικής μεταβολής με βάση τις τιμές έκφρασης του γονιδίου διαδραματίζει ένα σημαντικό ρόλο σε συγκεκριμένες διεργασίες. Προτείναμε μία μέθοδο ομαδοποίησης για αυτό το σκοπό η οποία βασίζεται στα πρότυπα έκφρασης αλλά επίσης επηρεάζεται από τις χρονικές μεταβολές των δεδομένων. Συμπεράναμε ότι οι χρονικές αλλαγές μπορούν να υποστηρίξουν τη βιολογική απόδοση και τη σύγκριση των δύο τύπων κυττάρων με βάση το βιολογικό ρόλο τους. Παραγάγαμε τρείς ομάδες για κάθε τύπο κυττάρου και συγκρίναμε το περιεχόμενο καθεμιάς υπό όρους χρονικών μεταβολών. Για την στατιστική αξιολόγηση προτείναμε ένα μέτρο εγκυρότητας το οποίο λαμβάνει υπόψη τις χρονικές αλλαγές και πραγματοποιήσαμε μια ανάλυση (PANTHER) των τριών κεντρικών γονιδίων κάθε ομάδας. Εντοπίσαμε πολλές στατιστικές ομοιότητες οι οποίες συνοδεύτηκαν από βιολογική ερμηνεία.

6.2 Μελλοντικές Επεκτάσεις

Ενδιαφέρον θα αποτελούσε η εφαρμογή διαφορετικών μεθόδων ομαδοποίησης και οπτικοποίησης στις ίδιες γονιδιακές βάσεις δεδομένων, με απώτερο στόχο την σύγκριση των αποτελεσμάτων με αυτά που εμείς καταλήξαμε και την ενίσχυση των συμπερασμάτων. Μια σημαντική ενδεχόμενη προοπτική θα ήταν η μελέτη της προτεινόμενης μεθοδολογίας σε άλλες γονιδιακές βάσεις, με σκοπό τον περαιτέρω έλεγχο της εγκυρότητας των τεχνικών.

Αναφορικά με τη νέα μεθοδολογία που προτείνουμε με βάση μέτρα χρονικής εξέλιξης θα άξιζε της προσπάθειας η αναζήτηση ενός εναλλακτικού τρόπου κωδικοποίησης των χρονικών μεταβολών. Μια διαφορετική προσέγγιση ίσως να οδηγούσε σε ποιοτικά αποτελέσματα που θα κρίνονταν τόσο με όρους στατιστικούς όσο και βιολογικούς.

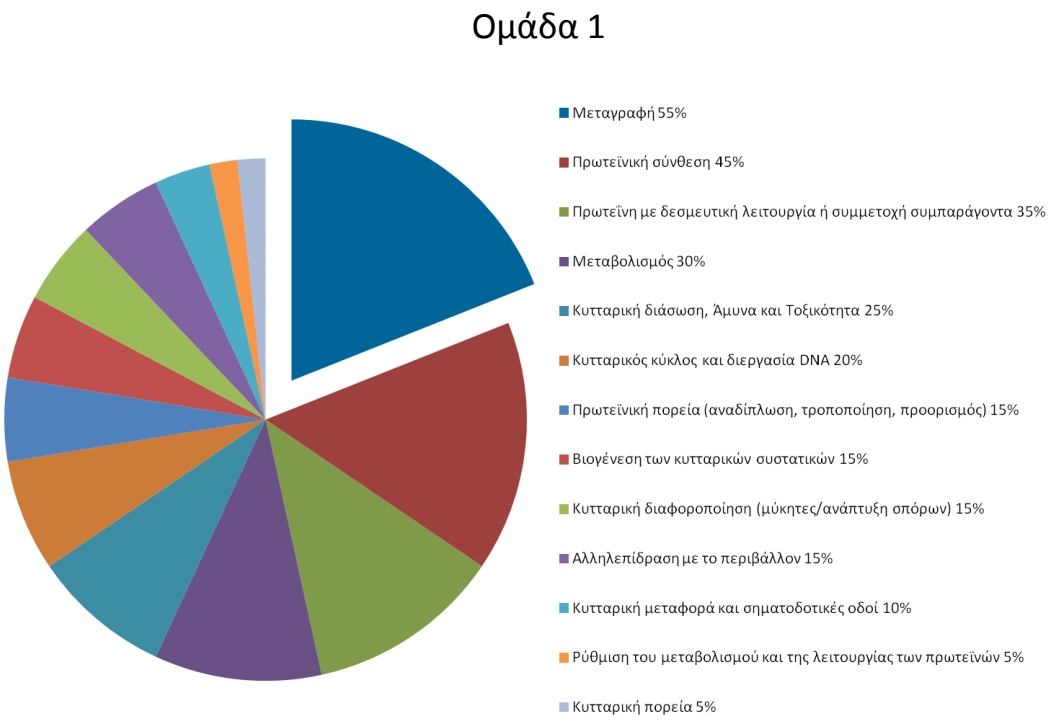
Τέλος, ως μελλοντική επέκταση συνιστάται η χρήση διαφορετικών μετρικών απόστασης τόσο στις προτεινόμενες τροποποιημένες μεθοδολογίες, SOM και LVQ, όσο και στην μεθοδολογία ομαδοποίησης με βάση τις χρονικές μεταβολές. Προτείνουμε στη δεύτερη περίπτωση, συγκεκριμένα στον υπολογισμό του δείκτη RSI την αντικατάσταση της Hamming distance με την εύρεση της πραγματικής απόστασης μεταξύ των εκάστοτε συγκρινόμενων συμβολοσειρών. Μια τέτοιου είδους

αντιμετώπιση ενδέχεται να οδηγήσει τα αποτελέσματα προς ενδιαφέρουσες κατευθύνσεις.

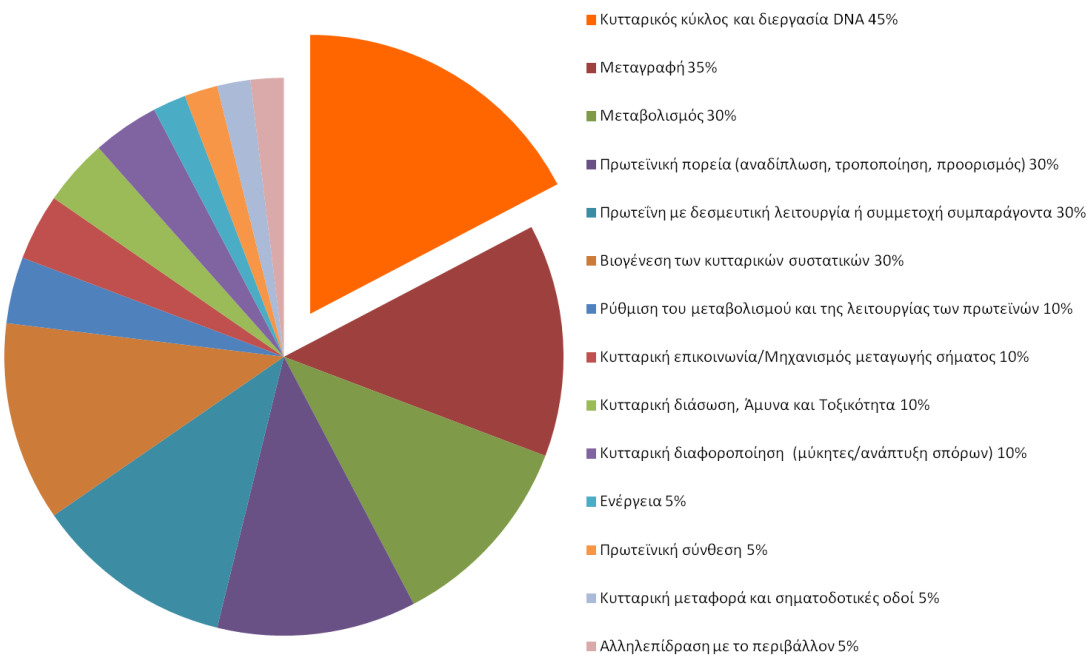
Παράρτημα Α

Στο παράρτημα αυτό ενισχύονται τα αποτελέσματα που δόθηκαν κατά τη διαδικασία παρουσίασης τους, στην εν λόγω διπλωματική εργασία. Συγκεκριμένα παρουσιάζονται τα ακόλουθα θέματα: οπτική απεικόνιση των βιολογικών διεργασιών για κάθε μια από τις τρεις ομάδες του *Saccharomyces cerevisiae* όπως προέκυψαν από την τροποποιημένη LVQ ομαδοποίηση, αποτελέσματα οπτικοποίησης των ομάδων των μεσεγχυματικών βλαστικών κυττάρων και χονδροκυττάρων με την τεχνική k-PCA (με πυρήνες πολυωνυμικούς και Gaussian) για διαφορετικό αριθμό ομάδων και τέλος διαγράμματα Silhouette για τους ομαδοποιημένους τύπους κυττάρων μετά την τροποποιημένη SOM ομαδοποίηση. Πριν από την τοποθέτηση κάθε εικόνας δίνεται ακριβής προσδιορισμός της ενότητας στην οποία αναφέρεται καθώς και περιγράφεται αναλυτικά το περιεχόμενό της.

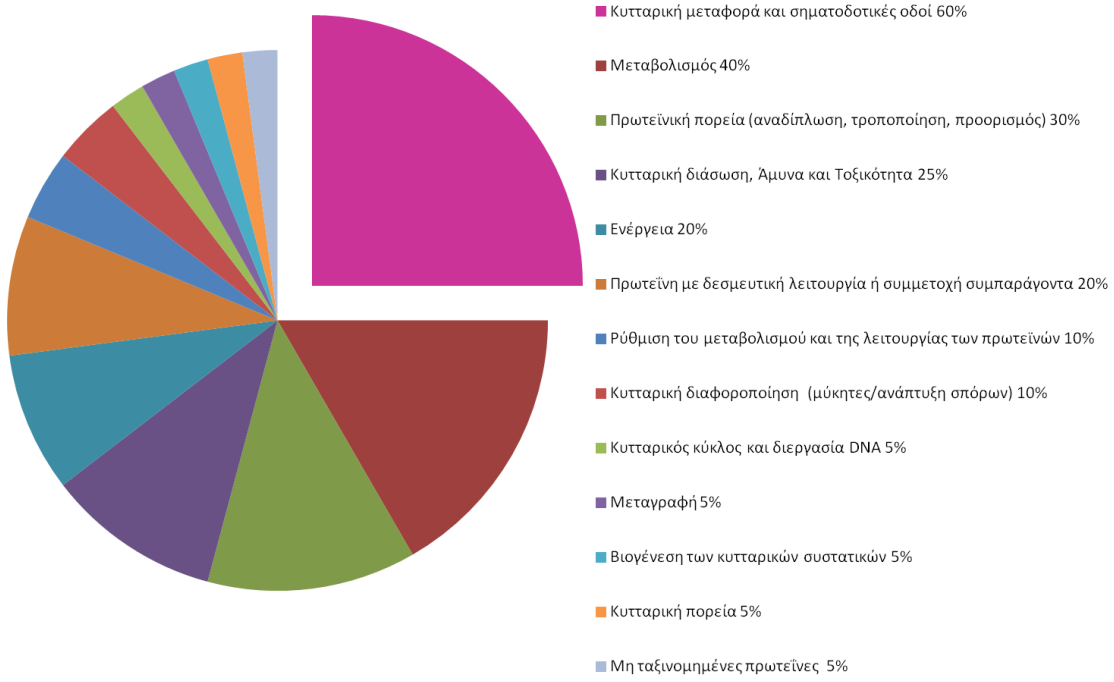
Ενότητα 5.3: Οι παρακάτω εικόνες υποδεικνύουν πως οι βιολογικές διεργασίες κατανέμονται σε κάθε μία από τις 3 ομάδες του σακχαρομύκητα *Saccharomyces cerevisiae* , σύμφωνα με τα αποτελέσματα της τροποποιημένης ομαδοποίησης LVQ.



Ομάδα 2



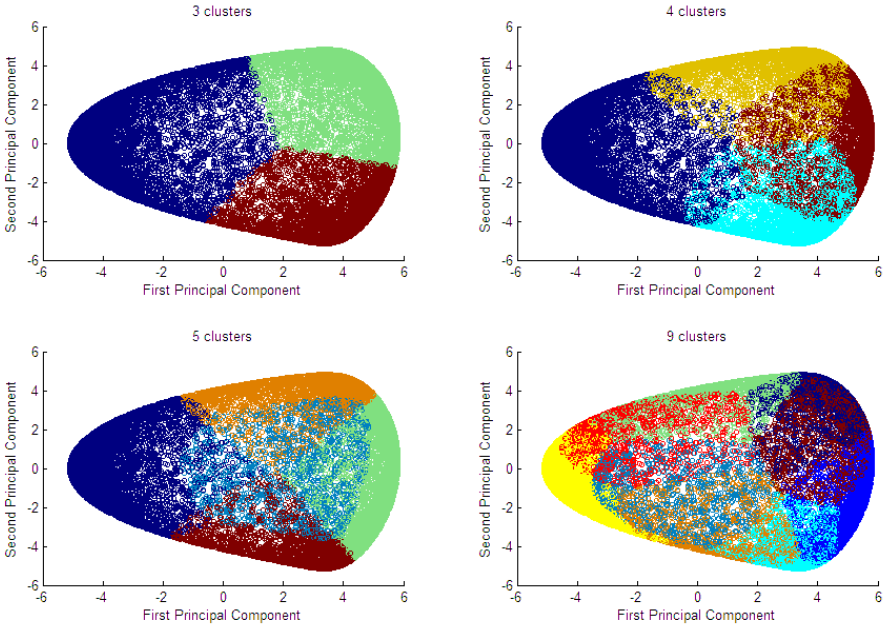
Ομάδα 3



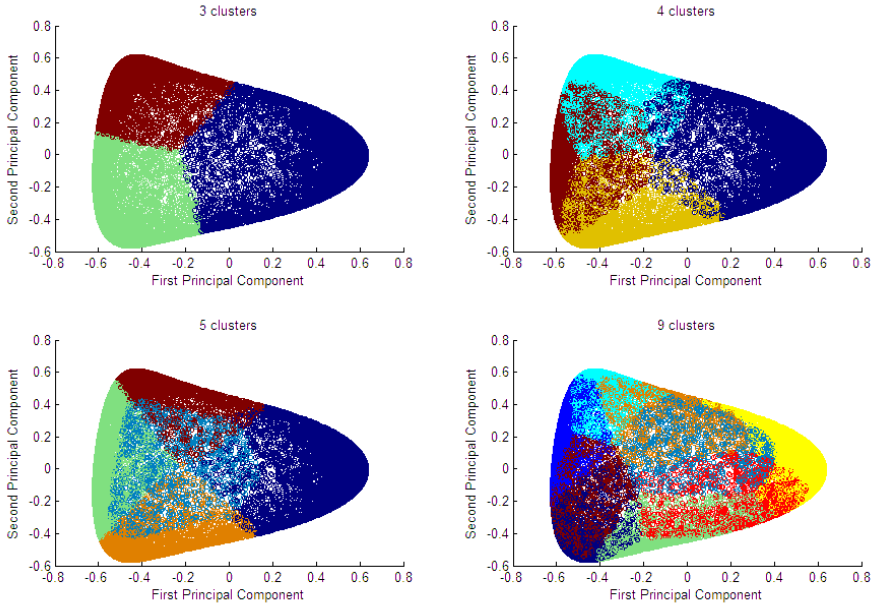
Ενότητα 5.4: Οι εικόνες που ακολουθούν δείχνουν τα αποτελέσματα της οπτικοποίησης των χονδροκυττάρων, μετά την ομαδοποίηση SOM, με k-

pca(polynomial-3d (σχήμα 1) & gaussian(σχήμα 2)) τεχνική, για διαφορετικό αριθμό ομάδων.

Σχήμα 1



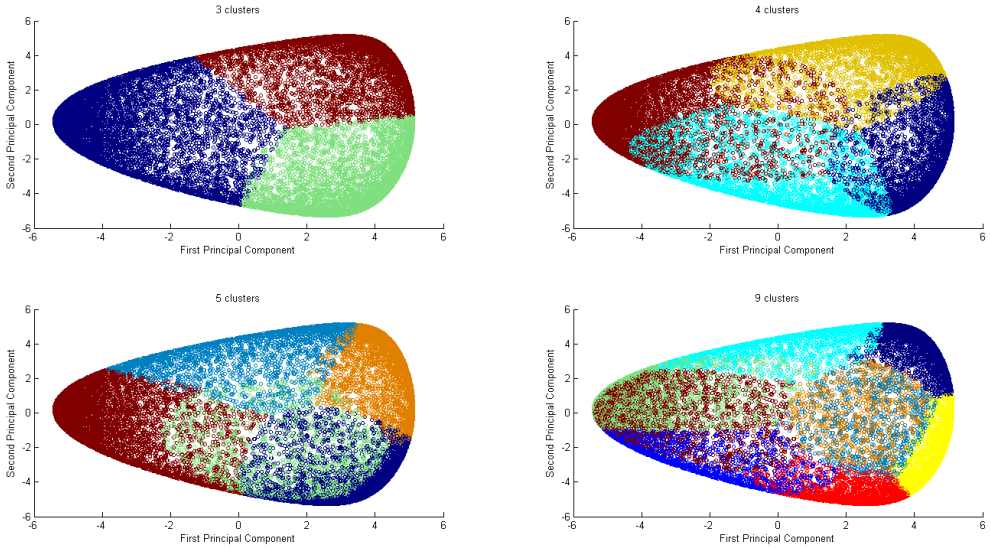
Σχήμα 2



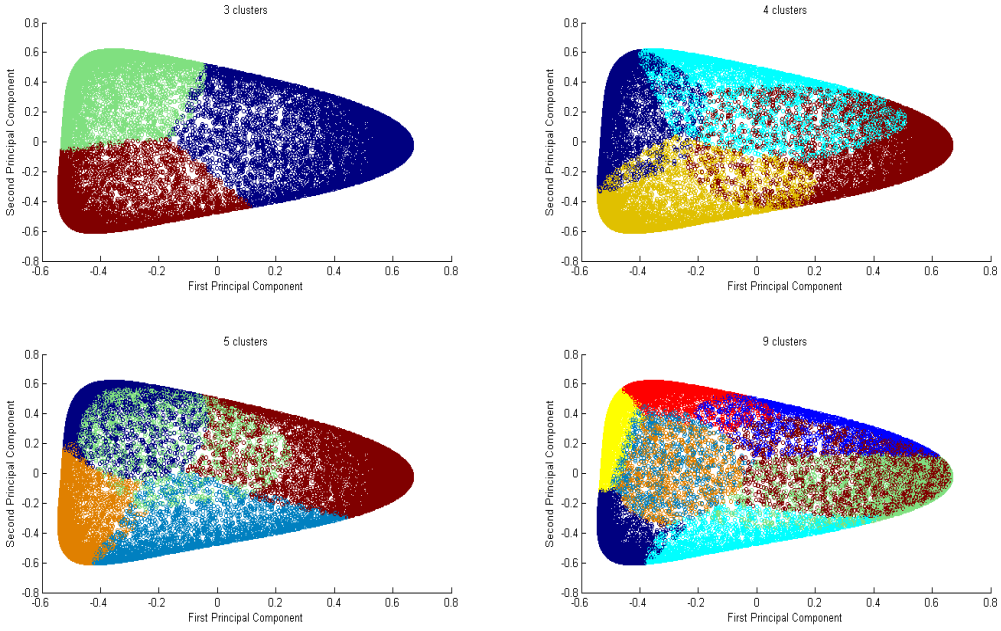
Ενότητα 5.4: Οι εικόνες που ακολουθούν δείχνουν τα αποτελέσματα της οπτικοποίησης των μεσεγχυματικών βλαστικών κυττάρων, μετά την ομαδοποίηση

SOM, με k-pca(polynomial-3d (σχήμα 1) & gaussian(σχήμα 2)), τεχνική για διαφορετικό αριθμό ομάδων.

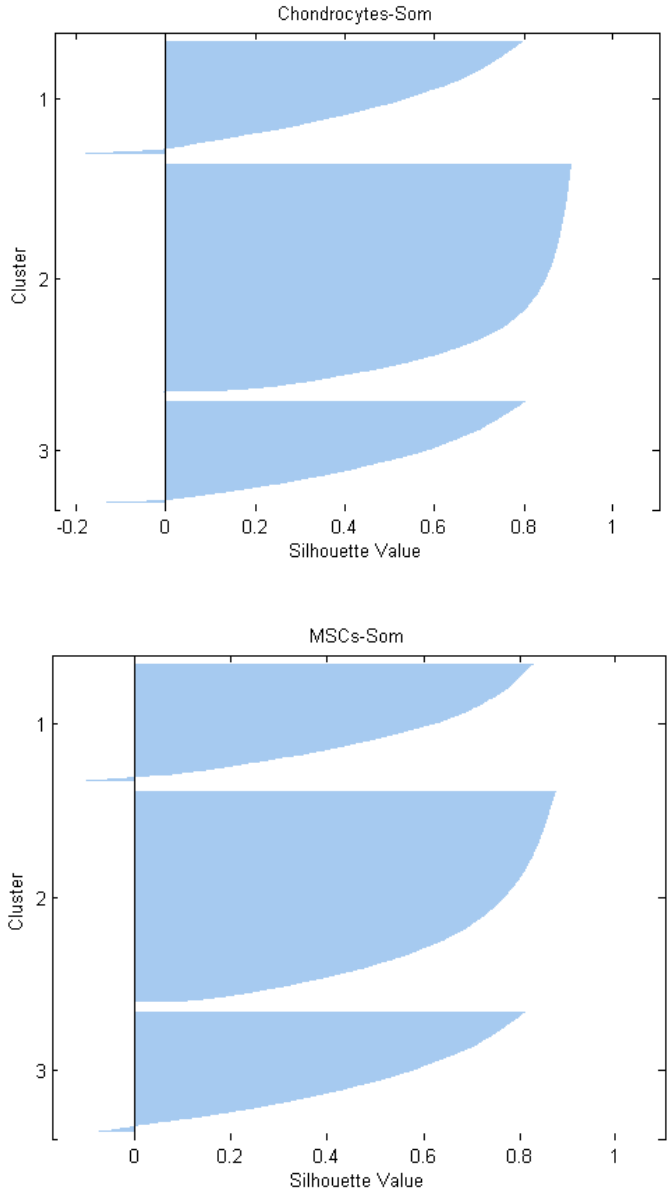
Σχήμα 1



Σχήμα 2

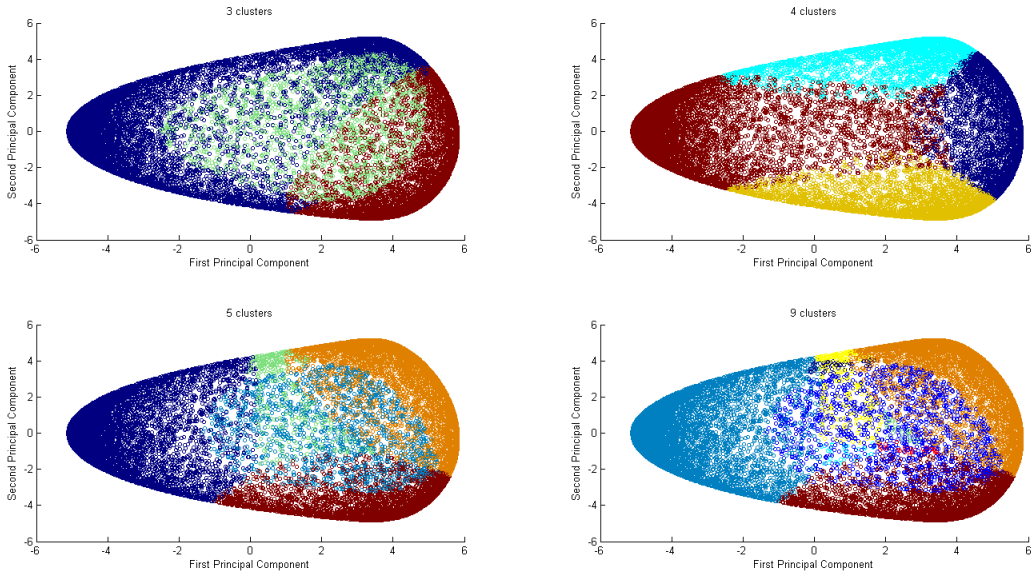


Ενότητα 5.4: Στα δύο εικόνες που ακολουθούν δίνονται τα διαγράμματα Silhouette πάνω στα οποία στηρίχθηκε η ομαδοποίηση LVQ τόσο για τα χονδροκύτταρα (1) όσο και για τα μεσεγχυματικά βλαστικά κύτταρα (2).

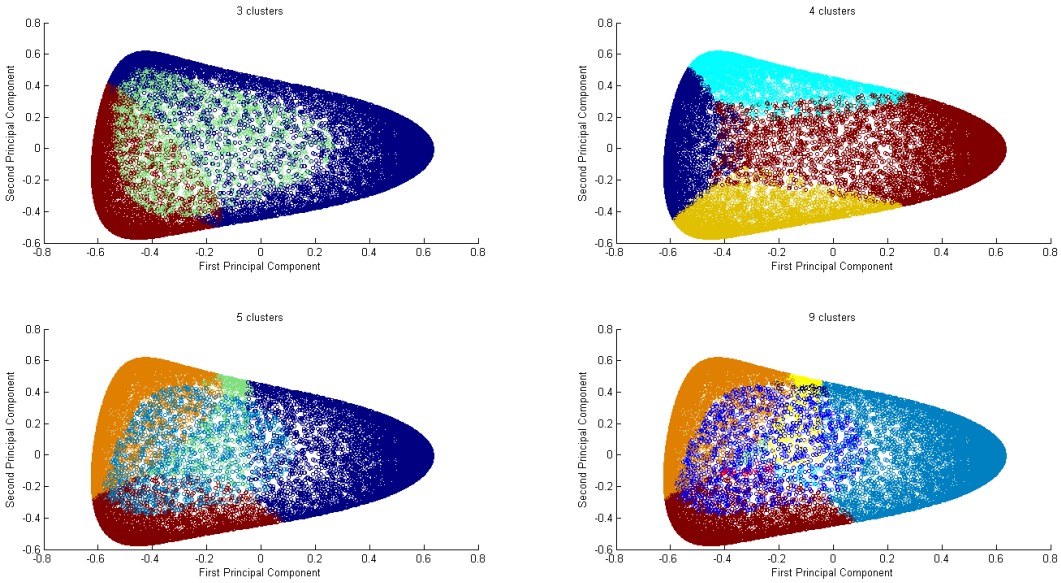


Ενότητα 5.4: Οι εικόνες που ακολουθούν δείχνουν τα αποτελέσματα της οπτικοποίησης των χονδροκυττάρων, μετά την ομαδοποίηση LVQ, με k-pca(polynomial-3d (σχήμα 1) & gaussian(σχήμα 2)) τεχνική, για διαφορετικό αριθμό ομάδων.

Σχήμα 1

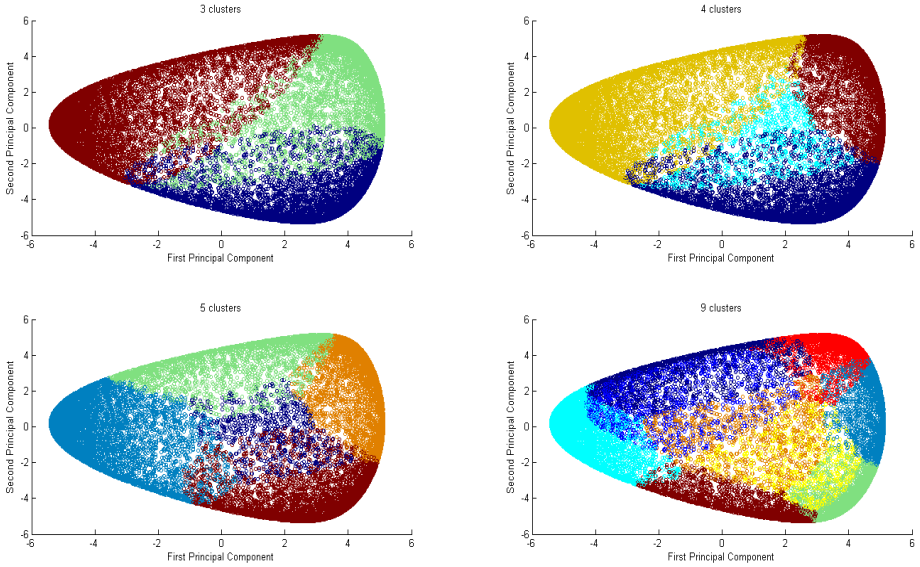


Σχήμα2

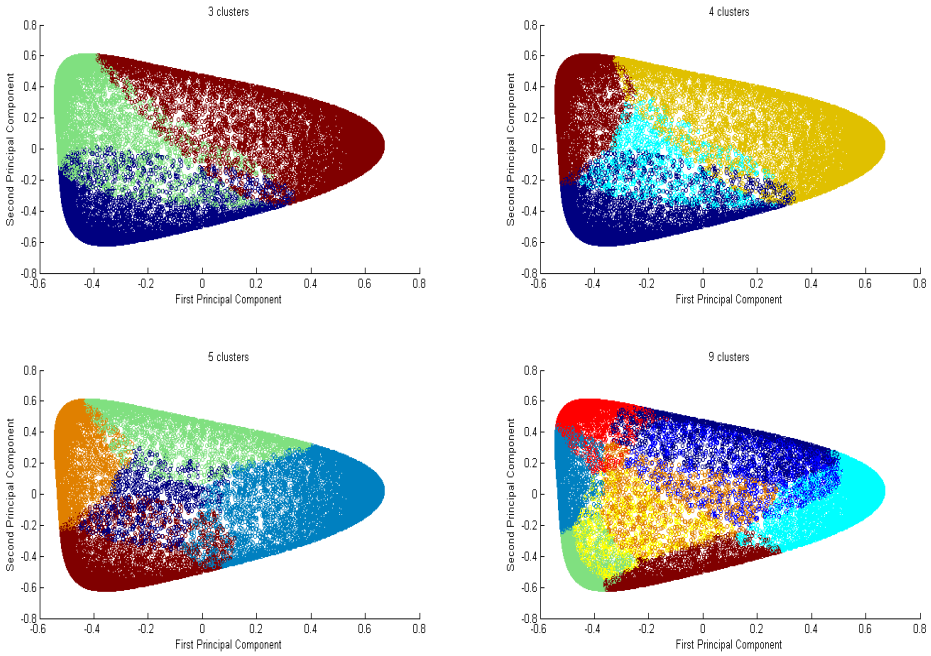


Ενότητα 5.4: Οι εικόνες που ακολουθούν δείχνουν τα αποτελέσματα της οπτικοποίησης των μεσεγχυματικών βλαστικών κυττάρων, μετά την ομαδοποίηση LVQ, με k-pca(polynomial-3d (σχήμα 1) & gaussian(σχήμα 2)), τεχνική για διαφορετικό αριθμό ομάδων.

Σχήμα1



Σχήμα2



Παράρτημα Β

Στο παράρτημα αυτό παρουσιάζεται περεταίρω πληροφορία, σε μορφή πινάκων, σχετικά με μεταβατικά στάδια στα διάφορα προβλήματα που αντιμετωπίσαμε. Συγκεκριμένα παρουσιάζονται τα ακόλουθα θέματα: οι 56 παράγοντες της βάσης του *Saccharomyces cerevisiae* όπως διαμορφώθηκαν μετά την αφαίρεση ορισμένων από τους αρχικούς παράγοντες, οι διαφορετικές διαμερίσεις των παραγόντων του *Saccharomyces cerevisiae* μετά από τις SOM και LVQ τροποποιημένες μεθοδολογίες, πληροφορίες για τα 20 ενδεικτικά γονίδια από τις 3 ομάδες του *Saccharomyces cerevisiae* που δόθηκαν για βιολογική αξιολόγηση, τα γονίδια που δόθηκαν για βιολογική ερμηνεία για τους δύο τύπους κυττάρων μετά από την LVQ τροποποιημένη μεθοδολογία και τέλος τα 3 κεντρικά γονίδια από κάθε ομάδα των μεσεγχυματικών βλαστικών κυττάρων και χονδροκυττάρων όπως προέκυψαν μετά την ομαδοποίηση με βάση τα μέτρα χρονικής εξέλιξης . Πριν από την τοποθέτηση κάθε πίνακα δίνεται ακριβής προσδιορισμός της ενότητας στην οποία αναφέρεται καθώς και περιγράφεται αναλυτικά το περιεχόμενό του.

Ενότητα 5.3: Ο παρακάτω πίνακας εμπεριέχει τους 56 παράγοντες της γονιδιακής βάσης του σακχαρομύκητα *Saccharomyces cerevisiae*, όπως αυτοί διαμορφώθηκαν μετά την αφαίρεση ορισμένων παραγόντων.

ΑΥΞΟΥΣΑ ΘΕΣΗ(ΣΤΗΛΕΣ) ΣΤΗ ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ	ΟΝΟΜΑ ΠΑΡΑΓΟΝΤΑ
Ομάδα 1	
1	alpha 0
2	alpha 7
3	alpha 28
4	alpha 49
5	alpha 70
6	alpha 91
7	alpha 112
8	alpha 119
Ομάδα 2	
9	elu 0
10	elu 30

11	elu 60
12	elu 90
13	elu 120
14	elu 150
15	elu 180
16	elu 210
Ομάδα 3	
17	cdc15 10
18	cdc15 30
19	cdc15 50
20	cdc15 70
21	cdc15 90
22	cdc15 110
23	cdc15 130
24	cdc15 150
25	cdc15 170
26	cdc15 190
Ομάδα 4	
27	spo 0
27	spo 2
29	spo 5
30	spo 7
31	spo 9
32	spo 11
33	spo5 2
34	spo5 7
35	spo5 11
36	spo- early
37	spo- mid
Ομάδα 5	
38	heat 0
39	heat 20
40	heat 40
41	heat 160
42	dtt 15

43	dtt 30
44	dtt 60
45	dtt 120
46	cold 0
47	cold 20
48	cold 40
49	cold 160
Ομάδα 6	
50	diau a
51	diau b
52	diau c
53	diau d
54	diau e
55	diau f
56	diau g

Ενότητα 5.3: Παραθέτονται οι τέσσερις διαμερίσεις των 56 παραγόντων του σακχαρομύκητα *cerevisiae* μετά από την τροποποιημένη SOM ομαδοποίηση.

Αρχική ομαδοποίηση	**6 ομάδες**	**7 ομάδες**	**8 ομάδες**	**9 ομάδες**
1	3	2	3	5
1	3	2	3	5
1	3	2	3	5
1	3	2	3	5
1	3	2	3	5
1	3	2	3	5
1	3	2	3	5
1	3	2	3	5
2	5	1	7	7
2	3	2	3	5
2	3	2	3	5

2	3	2	3	5
2	3	2	3	5
2	3	2	3	5
2	3	2	3	5
2	3	2	3	5
3	6	2	1	3
3	6	2	1	3
3	6	2	1	3
3	6	2	1	3
3	6	2	1	3
3	6	2	1	3
3	6	2	1	3
3	2	2	8	6
3	2	2	8	6
3	6	2	1	3
4	3	2	3	5
4	1	3	2	4
4	1	3	2	4
4	1	3	2	4
4	1	3	2	4
4	1	3	2	4
4	2	2	8	6
4	3	4	3	5
4	3	4	3	5
4	1	7	6	9
4	1	7	6	9
5	3	2	3	5
5	4	5	4	8
5	4	5	4	8
5	4	6	4	1
5	3	2	7	7
5	3	2	7	7

5	3	2	7	7
5	3	2	7	7
5	3	2	3	5
5	2	6	8	6
5	4	6	4	1
5	4	6	4	1
6	3	2	3	5
6	2	2	8	6
6	2	2	8	6
6	3	2	7	7
6	3	2	7	7
6	5	1	2	2
6	5	1	2	2

Ενότητα 5.3: Παρουσιάζονται οι δύο διαμερίσεις των 56 παραγόντων του σακχαρομύκητα *cerevisiae* μετά από την τροποποιημένη LVQ ομαδοποίηση.

Αρχική ομαδοποίηση	**6 ομάδες**	**8 ομάδες**
1	6	5
1	6	5
1	6	5
1	6	5
1	6	5
1	6	5
1	6	5
1	6	5
2	1	7
2	2	2
2	2	2
2	2	2
2	2	2

2	2	2
2	2	2
2	2	2
3	3	8
3	3	8
3	3	8
3	3	8
3	3	8
3	3	8
3	3	8
3	3	8
3	3	8
3	3	8
4	6	5
4	5	3
4	5	3
4	5	3
4	5	3
4	5	3
4	3	6
4	4	4
4	6	2
4	5	3
4	5	3
5	6	5
5	2	1
5	2	1
5	2	1
5	1	7
5	1	7
5	1	7
5	1	7

5	6	5
5	2	1
5	2	1
5	2	1
6	6	2
6	3	6
6	3	6
6	3	6
6	1	6
6	3	6
6	4	6

Ενότητα 5.3: Στον πίνακα που ακολουθεί δίνονται τα γονίδια τα οποία επιλέχθηκαν για βιολογική ερμηνεία από τις 3 ομάδες του σακχαρομύκητα *cerevisiae*, σύμφωνα με τα αποτελέσματα της τροποποιημένης ομαδοποίησης LVQ.

20 ΕΝΔΕΙΚΤΙΚΑ ΓΟΝΙΔΙΑ* ΓΙΑ ΚΑΘΕΜΙΑ ΑΠΟ ΤΙΣ 3 ΟΜΑΔΕΣ ΤΟΥ *SACCHAROMYCES CEREVISIAE*				
Αύξοντας αριθμός	**Κωδικός γονιδίου**	**ORF**	**Σύμβολο γονιδίου**	**Περιγραφή**
Ομάδα 1 - *Saccharomyces cerevisiae*				
1	1704	YLR293C	GSP1	NUCLEAR PROTEIN TARGETIN GTP-BINDING PROTEIN, RAS SUPERFAMILY
2	1638	YLR172C	DPH5	DIPHTHAMIDE BIOSYNTHESIS DIPHTHAMIDE METHYLTRANSFERASE
3	1496	YNL062C	GCD10	PROTEIN SYNTHESIS TRANSLATION INITIATION FACTOR EIF3 RNA-BINDING SUBUNIT
4	1687	YBR079C	RPG1	PROTEIN SYNTHESIS TRANSLATION INITIATION FACTOR EIF3
5	1683	YKR092C	SRP40	TRANSCRIPTION (PUTATIVE) SUPPRESSOR OF MUTANT AC40 SUBUNIT OF RNA POLYMERASE I AND III
6	1435	YBR154C	RPB5	TRANSCRIPTION SHARED SUBUNIT OF RNA POLYMERASES I, II, AND III
7	1752	YMR199W	CLN1	CELL CYCLE G1/S CYCLIN
8	1519	YPL037C	EGD1	TRANSCRIPTION REGULATOR OF POL II TRANSCRIBED GENES; ENHANCES GAL4 DNA BINDING
9	1451	YOR361C	PRT1	PROTEIN SYNTHESIS TRANSLATION INITIATION FACTOR EIF3 SUBUNIT
10	1698	YDR037W	KRS1	PROTEIN SYNTHESIS LYSYL-TRNA SYNTHETASE
11	1492	YGL120C	PRP43	MRNA SPLICING SPLICEOSOME DISASSEMBLY FACTOR; RNA HELICASE
12	1685	YMR229C	RRP5	RRNA PROCESSING UNKNOWN; REQUIRED FOR PRE-RRNA CLEAVAGE
13	1639	YHR193C	EGD2	PROTEIN SYNTHESIS (PUTAT HOMOLOG OF HUMAN NASCENT-POLYPEPTIDE-ASSOCIATED COMPLEX SUBUNIT)
14	760	YDR502C	SAM2	METHIONINE BIOSYNTHESIS REGULATOR; S-ADENOSYLMETHIONINE SYNTHETASE
15	1463	YBR121C	GRS1	PROTEIN SYNTHESIS GLYCYL-TRNA SYNTHASE
16	1692	YCL037C	SRO9	CYTOSKELETON ACTIN FILAMENT ORGANIZATION
17	1433	YDR023W	SES1	PROTEIN SYNTHESIS SERYL-TRNA SYNTHETAS
18	1703	YDR184C	ATC1	CELL POLARITY MEMBER OF BUD6P COMPLEX
19	1647	YIL035C	CKA1	CELL CYCLE (PUTATIVE) CASEIN KINASE II, CATALYTIC SUBUNIT
20	1705	YER043C	SAH1	METHIONINE BIOSYNTHESIS S-ADENOSYL-L-HOMOCYSTEINE HYDROLASE
Ομάδα 2 - *Saccharomyces cerevisiae*				
1	309	YDL170W	UGA3	TRANSCRIPTION ACTIVATOR OF GABA CATABOLIC GENES
2	113	YGL205W	POX1	FATTY ACID METABOLISM ACYL-COA OXIDASE
3	258	YMR127C	SAS2	SILENCING ZINC-FINGER PROTEIN
4	128	YLR102C	APC9	CELL CYCLE ANAPHASE-PROMOTING COMPLEX SUBUNIT
5	316	YPL122C	TFB2	TRANSCRIPTION TFIIH 55 KD SUBUNIT
6	92	YBR257W	POP4	RRNA AND TRNA PROCESSING RNASE P AND RNASE MRP SUBUNIT
7	156	YDR356W	NUF1	CYTOSKELETON SPINDLE POLE BODY COMPONENT
8	90	YNL270C	ALP1	TRANSPORT BASIC AMINO ACID PERMEASE
9	124	YKR019C	IRS4	SILENCING (RDNA) UNKNOWN
10	279	YHR172W	SPC97	CYTOSKELETON SPINDLE POLE BODY COMPONENT
11	262	YJL074C	SMC3	CHROMATIN STRUCTURE COHESIN
12	310	YDR256C	CTA1	OXIDATIVE STRESS RESPONS CATALASE A
13	485	YKR034W	DAL80	NITROGEN CATABOLISM TRANSCRIPTION FACTOR
14	336	YGR188C	BUB1	CELL CYCLE, CHECKPOINT PROTEIN KINASE
15	399	YOR261C	RPN8	PROTEIN DEGRADATION 26S PROTEASOME REGULATORY SUBUNIT
16	313	YML091C	RPM2	TRNA PROCESSING, MITOCHO RNASE P SUBUNIT
17	263	YLR313C	SPH1	BUD SITE SELECTION, BIPO INTERACTS WITH MAPKKS
18	84	YPR106W	ISR1	STAUROSPORINE RESISTANCE PROTEIN KINASE
19	151	YGR099W	TEL2	TELOMERE LENGTH REGULATI TELOMERE BINDING PROTEIN
20	335	YGL240W	DOC1	CELL CYCLE ANAPHASE-PROMOTING COMPLEX SUBUNIT
Ομάδα 3 - *Saccharomyces cerevisiae*				
1	2406	YLR120C	YPS1	PROTEIN PROCESSING GPI-ANCHORED ASPARTIC PROTEASE
2	2354	YDR342C	HXT7	TRANSPORT HEXOSE PERMEASE
3	641	YJR095W	ACR1	TRANSPORT MITOCHONDRIAL CARRIER
4	2399	YDR513W	TTR1	ELECTRON CARRIER GLUTAREDOXIN
5	2305	YNL036W	NCE103	SECRETION, NON-CLASSICAL UNKNOWN
6	909	YJR019C	TES1	FATTY ACID METABOLISM PEROXISOMAL ACYL-COA THIOESTERASE
7	2355	YDR343C	HXT6	TRANSPORT HEXOSE PERMEASE
8	2353	YIL162W	SUC2	SUCROSE UTILIZATION INVERTASE
9	2376	YIL101C	XBP1	STRESS RESPONSE TRANSCRIPTIONAL ACTIVATOR
10	643	YKL217W	JEN1	TRANSPORT LACTATE TRANSPORTER
11	2328	YLL024C	SSA2	ER AND MITOCHONDRIAL TRA CYTOSOLIC HSP70
12	974	YKL193C	SDS22	GLUCOSE REPRESSION GLC7P REGULATORY SUBUNIT
13	2397	YIL136W	OM45	(PUTATIVE) MITOCHONDRIAL OUTER MITOCHONDRIAL MEMBRANE PROTEIN
14	2329	YAL005C	SSA1	ER AND MITOCHONDRIAL TRA CYTOSOLIC HSP70
15	976	YGL191W	COX13	OXIDATIVE PHOSPHORYLATIO CYTOCHROME-C OXIDASE SUBUNIT VIA
16	2373	YLR259C	HSP60	PROTEIN FOLDING MITOCHONDRIAL CHAPERONIN
17	913	YCR088W	ABP1	CYTOSKELETON ACTIN BINDING PROTEIN
18	2374	YBR132C	AGP2	TRANSPORT GENERAL AMINO ACID PERMEASE
19	765	YLR028C	ADE16	PURINE BIOSYNTHESIS 5-AMINOIMIDAZOLE-4-CARBOXAMIDE RIBONUCLEOTIDE (AICAR) TRANSFORMYLASE/IMP
20	642	YNL117W	MLS1	GLYOXYLATE CYCLE MALATE SYNTHASE

*Τα ενδεικτικά γονίδια για καθεμία από τις τρεις ομάδες του *Saccharomyces cerevisiae* αναφέρονται στα 20 πρώτα γονίδια που είναι πλησιέστερα στο κέντρο.

Ενότητα 5.4: Στους πίνακες που έπονται δίνονται τα 20 γονίδια τα οποία επιλέχθηκαν για βιολογική αξιολόγηση από τις 3 ομάδες των μεσεγχυματικών βλαστικών κυττάρων και των χονδροκυττάρων σύμφωνα με τα αποτελέσματα της τροποποιημένης ομαδοποίησης LVQ.

20 ΕΝΔΕΙΚΤΙΚΑ ΓΟΝΙΔΙΑ* ΓΙΑ ΚΑΘΕΜΙΑ ΑΠΟ ΤΙΣ 3 ΟΜΑΔΕΣ ΤΩΝ ΜΕΣΕΓΧΥΜΑΤΙΚΩΝ ΒΛΑΣΤΙΚΩΝ ΚΥΤΤΑΡΩΝ

	ΟΜΑΔΑ 1				ΟΜΑΔΑ 2				ΟΜΑΔΑ 3			
Αύξοντας αριθμός	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (ENTREZ_GENE_ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Περιγραφή	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (ENTREZ_GENE_ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Περιγραφή	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (ENTREZ_GENE_ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Περιγραφή
1	1948_at	1948	NM_004093	ephrin-B2	11123_at	11123	BU947377	RCAN family member 3	10838_at	10838	DB306278	zinc finger protein 275
2	210_at	210	BC000977	aminolevulinate, delta-, dehydratase	115761_at	115761	NM_138450	ADP-ribosylation factor-like 11	11235_at	11235	BG201813	programmed cell death 10
3	2134_at	2134	U67191	exostoses (multiple)-like 1	140462_at	140462	CA776490	ankyrin repeat and SOCS box-containing 9	115548_at	115548	AL831971	FCH domain only 2
4	22883_at	22883	BG027254	calsyntenin 1	255189_at	255189	BM986790	phospholipase A2, group IVF	115950_at	115950	AI768492	zinc finger protein 653
5	23169_at	23169	BC093786	solute carrier family 35 (UDP-glucuronic acid/UDP-N-acetylgalactosamine dual transporter), member D1	30010_at	30010	BX107562	neurexophilin 1	151987_at	151987	AA576701	protein phosphatase 4, regulatory subunit 2
6	266812_at	266812	AI816263	nucleosome assembly protein 1-like 5	4143_at	4143	BX119626	methionine adenosyltransferase I, alpha	25948_at	25948	BC032367	kelch repeat and BTB (POZ) domain containing 2
7	27235_at	27235	BM126596	coenzyme Q2 homolog, prenyltransferase (yeast)	56891_at	56891	BG572129	lectin, galactoside-binding, soluble, 14	285282_at	285282	BX647559	RAB, member of RAS oncogene family-like 3
8	29102_at	29102	CK902469	ribonuclease III, nuclear	6708_at	6708	M61877	spectrin, alpha, erythrocytic 1 (elliptocytosis 2)	29889_at	29889	AA722289	guanine nucleotide binding protein-like 2 (nucleolar)
9	340481_at	340481	BC042558	zinc finger, DHHC-type containing 21	79935_at	79935	AK023327	cyclin N-terminal domain containing 2	3638_at	3638	NM_005542	insulin induced gene 1
10	4338_at	4338	EL952437	molybdenum cofactor synthesis 2	80063_at	80063	AK022730	activating transcription factor 7 interacting protein 2	375_at	375	NM_001024226	ADP-ribosylation factor 1
11	4731_at	4731	CR593887	NADH dehydrogenase (ubiquinone) flavoprotein 3, 10kDa	83849_at	83849	AJ303363	synaptotagmin XV	50515_at	50515	BU622688	carbohydrate (chondroitin 4) sulfotransferase 11
12	55610_at	55610	AL832393	coiled-coil domain containing 132	84626_at	84626	DB311253	KRAB-A domain containing 1	51026_at	51026	NM_016072	golgi transport 1 homolog B (S. cerevisiae)
13	55752_at	55752	NM_018243	septin 11	9362_at	9362	CR598214	copine VI (neuronal)	51596_at	51596	CN363654	cutA divalent cation tolerance homolog (E. coli)
14	60494_at	60494	BE352845	coiled-coil domain containing 81	114805_at	114805	AB067505	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 13 (GalNAc-T13)	55596_at	55596	BF062671	zinc finger, CCHC domain containing 8
15	6713_at	6713	BX647400	squalene epoxidase	25928_at	25928	BI494438	sclerostin domain containing 1	6617_at	6617	BX346864	small nuclear RNA activating complex, polypeptide 1, 43kDa
16	83452_at	83452	BM893798	RAB33B, member RAS oncogene family	3760_at	3760	BC022495	potassium inwardly-rectifying channel, subfamily J, member 3	79850_at	79850	BX101022	family with sequence similarity 57, member A
17	83932_at	83932	DB367526	chromosome 1 open reading frame 124	51588_at	51588	BM981828	protein inhibitor of activated STAT, 4	85365_at	85365	AI693613	asparagine-linked glycosylation 2 homolog (S. cerevisiae, alpha-1,3-mannosyltransferase)
18	8434_at	8434	BU742733	reversion-inducing-cysteine-rich protein with kazal motifs	6530_at	6530	BC039309	solute carrier family 6 (neurotransmitter transporter, noradrenalin), member 2	9867_at	9867	BU181103	praja 2, RING-H2 motif containing
19	9162_at	9162	AK091081	diacylglycerol kinase, iota	6693_at	6693	NM_003123	sialophorin (leukosialin, CD43)	9908_at	9908	AB014560	GTPase activating protein (SH3 domain) binding protein 2
20	9831_at	9831	AA479038	zinc finger protein 623	83873_at	83873	CD633289	G protein-coupled receptor 61	11124_at	11124	BC004970	Fas (TNFRSF6) associated factor 1

*Τα ενδεικτικά γονίδια για καθεμία από τις τρεις ομάδες των μεσεγχυματικών βλαστικών κυττάρων αναφέρονται στα 20 πρώτα γονίδια που είναι πλησιέστερα στο κέντρο. Πλατφόρμα της βάσης δεδομένων: GPL9828 (GEO Accession viewer_PlatformGPL9828).

20 ΕΝΔΕΙΚΤΙΚΑ ΓΟΝΙΔΙΑ* ΓΙΑ ΚΑΘΕΜΙΑ ΑΠΟ ΤΙΣ 3 ΟΜΑΔΕΣ ΤΩΝ ΧΟΝΔΡΟΚΥΤΤΑΡΩΝ

	ΟΜΑΔΑ 1				ΟΜΑΔΑ 2				ΟΜΑΔΑ 3			
Αύξοντας αριθμός	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (ENTREZ_GENE_ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Περιγραφή	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (ENTREZ_GENE_ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Περιγραφή	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (ENTREZ_GENE_ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Περιγραφή
1	10312_at	10312	BU070886	T-cell, immune regulator 1, ATPase, H+ transporting, lysosomal V0 subunit A3	11085_at	11085	AY358734	ADAM metallopeptidase domain 30	10099_at	10099	CR622572	tetraspanin 3
2	149041_at	149041	D60030	ring finger and CCCH-type zinc finger domains 1	11086_at	11086	AF171930	ADAM metallopeptidase domain 29	112812_at	112812	DA806538	similar to RIKEN cDNA B230118G17 gene
3	29982_at	29982	CR615835	nuclear receptor binding factor 2	138649_at	138649	AL041075	ankyrin repeat domain 19	22845_at	22845	BU539639	dolichol kinase
4	4803_at	4803	BT019733	nerve growth factor, beta polypeptide	170261_at	170261	BX422751	zinc finger, CCHC domain containing 12	2329_at	2329	AW340441	flavin containing monooxygenase 4
5	55005_at	55005	AK000634	required for meiotic nuclear division 1 homolog (S. cerevisiae)	2041_at	2041	Z27409	EPH receptor A1	375346_at	375346	BM669322	transmembrane protein 110
6	55122_at	55122	AK222701	chromosome 6 open reading frame 166	284100_at	284100	XM_375443	hypothetical protein LOC284100	387921_at	387921	BX647872	similar to RIKEN cDNA 8030451K01
7	55621_at	55621	BM984899	TRM1 tRNA methyltransferase 1 homolog (S. cerevisiae)	386608_at	386608	0	AYP1 pseudogene 1	5031_at	5031	AA290864	pyrimidinergic receptor P2Y, G-protein coupled, 6
8	55791_at	55791	BX089321	chromosome 1 open reading frame 103	389763_at	389763	DB341555	FLJ46321 protein	5264_at	5264	AW316781	phytanoyl-CoA 2-hydroxylase
9	55861_at	55861	AL591714	dysbindin (dystrobrevin binding protein 1) domain containing 2	400831_at	400831	CA440818	hypothetical LOC400831	5272_at	5272	BX640606	serpin peptidase inhibitor, clade B (ovalbumin), member 9
10	56907_at	56907	AJ277587	spire homolog 1 (Drosophila)	4018_at	4018	NM_005577	lipoprotein, Lp(a)	57088_at	57088	AF199023	phospholipid scramblase 4
11	64087_at	64087	DA602390	methylcrotonoyl-Coenzyme A carboxylase 2 (beta)	440072_at	440072	AK055956	chromosome 11 open reading frame 37	57134_at	57134	EL945850	mannosidase, alpha, class 1C, member 1
12	64853_at	64853	EC442595	chromosome 1 open reading frame 80	51151_at	51151	AF172849	solute carrier family 45, member 2	6387_at	6387	EL948719	chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)
13	6812_at	6812	CR602337	syntaxin binding protein 1	551_at	551	BC126224	arginine vasopressin (neurophysin II, antidiuretic hormone, diabetes insipidus, neurohypophyseal)	64924_at	64924	AF461760	solute carrier family 30 (zinc transporter), member 5
14	7779_at	7779	DA405288	solute carrier family 30 (zinc transporter), member 1	56547_at	56547	BI831623	matrix metallopeptidase 26	7157_at	7157	DQ186650	tumor protein p53 (Li-Fraumeni syndrome)
15	7874_at	7874	CA868275	ubiquitin specific peptidase 7 (herpes virus-associated)	6490_at	6490	BU161268	silver homolog (mouse)	7284_at	7284	L38995	Tu translation elongation factor, mitochondrial
16	8662_at	8662	AK127885	eukaryotic translation initiation factor 3, subunit B	6900_at	6900	NM_005076	contactin 2 (axonal)	79671_at	79671	AK225223	NLR family member X1
17	8674_at	8674	AL035296	vesicle-associated membrane protein 4	7113_at	7113	BC051839	transmembrane protease, serine 2	79834_at	79834	AB082533	NKF3 kinase family member
18	9792_at	9792	D50917	SERTA domain containing 2	91752_at	91752	CA417236	zinc finger protein 804A	8607_at	8607	CN389613	RuvB-like 1 (E. coli)
19	9818_at	9818	AK025800	nucleoporin like 1	9580_at	9580	CN293464	SRY (sex determining region Y)-box 13	8828_at	8828	NM_201279	neuropilin 2
20	9973_at	9973	AV718445	copper chaperone for superoxide dismutase	9609_at	9609	BU607791	RAB36, member RAS oncogene family	9706_at	9706	NM_014683	unc-51-like kinase 2 (C. elegans)

*Τα ενδεικτικά γονίδια για καθεμία από τις τρεις ομάδες των χονδροκυττάρων αναφέρονται στα 20 πρώτα γονίδια που είναι πλησιέστερα στο κέντρο. Πλατφόρμα της βάσης δεδομένων: GPL9828 (GEO Accession viewer_PlatformGPL9828).

Ενότητα 5.5: Παρουσιάζονται οι 27 συνδυασμοί των χρονικών μεταβολών των δεδομένων , δηλαδή οι κωδικοποιημένες μορφές των γονιδίων.

ΑΥΞΩΝ ΑΡΙΘΜΟΣ	Bit 1	Bit 2	Bit 3
1	0	0	0
2	0	0	1
3	0	0	-1
4	0	1	0
5	0	-1	0
6	1	0	0
7	-1	0	0
8	0	-1	1
9	0	1	-1
10	1	1	1
11	1	1	-1
12	1	1	0
13	1	0	1
14	1	-1	1
15	1	0	-1
16	-1	-1	-1
17	-1	-1	0
18	-1	-1	1
19	-1	0	-1
20	-1	1	-1
21	0	-1	-1
22	1	-1	-1
23	-1	0	1
24	0	1	1
25	-1	1	1
26	1	-1	0
27	-1	1	0

Ενότητα 5.5: Στους 3 πίνακες που ακολουθούν δίνεται λεπτομερής περιγραφή των ομάδων των δύο κυτταρικών τύπων (Chondrocytes και MSC), καθώς και της μεταξύ τους σχέσης όπως προέκυψε από την ομαδοποίηση γονιδίων με βάση τις εκφράσεις (Expression clustering). Τα ποσοστά που αναφέρονται στον τελευταίο σε σειρά πίνακα εκφράζουν το λόγο του πλήθους των κοινών γονιδίων, σε κάθε συνδυασμό, προς τον αριθμό γονιδίων της εκάστοτε MSC ομάδας.

Αριθμός Chondrocytes ομάδας	**Πλήθος γονιδίων**
1	5911
2	8639
3	3039
Σύνολο	**17589**

Αριθμός MSC ομάδας	**Πλήθος γονιδίων**
1	4485
2	9184
3	3920
Σύνολο	**17589**

Ομάδα MSC-Ομάδα Chondrocytes	**Πλήθος κοινών γονιδίων**	**Ποσοστό ομοιότητας**
1-1	1163	25.39%
1-2	1056	23.55%
1-3	**2266**	**50.52%**
Σύνολο		**100%**
2-1	1501	16.34%
2-2	**7136**	**77.70%**

2-3	547	5.96%
Σύνολο		**100%**
3-1	**3247**	**82.83%**
3-2	447	11.40%
3-3	226	5.77%
Σύνολο		**100%**

Ενότητα 5.5: Έπεται λεπτομερής περιγραφή των ομάδων των δύο κυτταρικών τύπων (Chondrocytes και MSC), καθώς και της μεταξύ τους σχέσης όπως προέκυψε από την ομαδοποίηση γονιδίων με βάση την κατάταξη (Rank-based clustering).

Αριθμός Chondrocytes ομάδας	**Πλήθος γονιδίων**
1	6538
2	4960
3	6091
Σύνολο	**17589**

Αριθμός MSC ομάδας	**Πλήθος γονιδίων**
1	8638
2	4259
3	4692
Σύνολο	**17589**

Ομάδα MSC-Ομάδα Chondrocytes	**Πλήθος κοινών γονιδίων**	**Ποσοστό ομοιότητας**
1-1	2601	30.11%
1-2	**3320**	**38.43%**

1-3	2717	31.45%
Σύνολο		**100%**
2-1	825	19.37%
2-2	1079	25.33%
2-3	**2355**	**55.29%**
Σύνολο		**100%**
3-1	**3112**	**66.33%**
3-2	561	11.96%
3-3	1019	21.72%
Σύνολο		**100%**

Ενότητα 5.5: Στους 3 πίνακες που ακολουθούν δίνεται λεπτομερής περιγραφή των ομάδων των δύο κυτταρικών τύπων (Chondrocytes και MSC), καθώς και της μεταξύ τους σχέσης όπως προέκυψε από την ομαδοποίηση γονιδίων με βάση τα μέτρα χρονικής εξέλιξης (Combined clustering).

Αριθμός Chondrocytes ομάδας	**Πλήθος γονιδίων**
1	3309
2	8656
3	5624
Σύνολο	**17589**

Αριθμός MSC ομάδας	**Πλήθος γονιδίων**
1	3744
2	8223
3	5622

Σύνολο	**17589**

Ομάδα MSC-Ομάδα Chondrocytes	**Πλήθος κοινών γονιδίων**	**Ποσοστό ομοιότητας**
1-1	**1670**	**44.60%**
1-2	848	22.65%
1-3	1226	32.75%
Σύνολο		**100%**
2-1	651	7.52%
2-2	**6935**	**84.34%**
2-3	637	7.75%
Σύνολο		**100%**
3-1	1432	25.47%
3-2	873	15.53%
3-3	**3317**	**59%**
Σύνολο		**100%**

Ενότητα 5.5: Στον πίνακα που έπεται δίνονται τα 3 γονίδια τα οποία επιλέχθηκαν για βιολογική αξιολόγηση από τις 3 ομάδες των μεσεγχυματικών βλαστικών κυττάρων και των χονδροκυττάρων σύμφωνα με τα αποτελέσματα της προτεινόμενης μεθοδολογίας ομαδοποίησης με βάση μέτρα χρονικής εξέλιξης.

3 ΕΝΔΕΙΚΤΙΚΑ ΓΟΝΙΔΙΑ* ΓΙΑ ΚΑΘΕΜΙΑ ΑΠΟ ΤΙΣ 3 ΟΜΑΔΕΣ ΤΩΝ ΜΕΣΕΓΧΥΜΑΤΙΚΩΝ ΒΛΑΣΤΙΚΩΝ ΚΥΤΤΑΡΩΝ

	ΟΜΑΔΑ 1				ΟΜΑΔΑ 2				ΟΜΑΔΑ 3			
Αύξοντας αριθμός	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (ENTREZ_GENE_ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Περιγραφή	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (ENTREZ_GENE_ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Περιγραφή	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (ENTREZ_GENE_ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Περιγραφή
1	26974_at	26974	AW513227	zinc finger protein 285A	1369_at	1369	CR603435	carboxypeptidase N, polypeptide 1	23141_at	23141	AW604816	KIAA0692
2	8260_at	8260	AA847551	ARD1 homolog A, N-acetyltransferase (S. cerevisiae)	23302_at	23302	AB011095	WSC domain containing 1	388585_at	388585	DB573581	hairy and enhancer of split 5 (Drosophila)
3	223117_at	223117	BC029590	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3D	1821_at	1821	U43519	dystrophin related protein 2	476_at	476	BM041435	ATPase, Na+/K+ transporting, alpha 1 polypeptide

3 ΕΝΔΕΙΚΤΙΚΑ ΓΟΝΙΔΙΑ* ΓΙΑ ΚΑΘΕΜΙΑ ΑΠΟ ΤΙΣ 3 ΟΜΑΔΕΣ ΤΩΝ ΧΟΝΔΡΟΚΥΤΤΑΡΩΝ

	ΟΜΑΔΑ 1				ΟΜΑΔΑ 2				ΟΜΑΔΑ 3			
Αύξοντας αριθμός	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (ENTREZ_GENE_ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Περιγραφή	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (ENTREZ_GENE_ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Περιγραφή	Κωδικός καταχώρησης στην GEO (GEO Accession viewer ID)	Κωδικός γονιδίου (ENTREZ_GENE_ID)	Αριθμός καταχώρησης στην GenBank (GenBank Accession number)	Περιγραφή
1	29935_at	29935	BC104013	replication protein A4, 34kDa	3226_at	3226	CA432500	homeobox C10	79001_at	79001	AI969210	vitamin K epoxide reductase complex, subunit 1
2	254394_at	254394	BX108374	minichromosome maintenance complex component 9	137196_at	137196	BC026098	coiled-coil domain containing 26	6508_at	6508	CK825914	solute carrier family 4, anion exchanger, member 3
3	6813_at	6813	CR598164	syntaxin binding protein 2	85019_at	85019	CR610638	chromosome 18 open reading frame 45	144097_at	144097	CR749697	hypothetical protein BC007540

*Τα ενδεικτικά γονίδια για καθεμία από τις τρεις ομάδες των μεσεγχυματικών βλαστικών κυττάρων ή των χονδροκυττάρων αναφέρονται στα 3 πρώτα γονίδια που είναι πλησιέστερα στο κέντρο. Πλατφόρμα της βάσης δεδομένων: GPL9828 (GEO Accession viewer_PlatformGPL9828).

Βιβλιογραφία

[1] Bushati N., Smith J. ,Briscoe J., Watkins C., 2011. "An intuitive graphical visualization technique for the interrogation of transcriptome data". *Nucleic Acids Research* 1-10.

[2] Saraiya P., Lee P., North C., 2005. "Visualization of Graphs with Associated Timeseries Data". *Virginia Polytechnic Institute and State University BlacksBurg.*

[3] Westenberg A. M., Hijum S., Lulko T. A., Kuipers P. O., Roerdink B. T. M., 2006. "Interactive Visualization of Gene Regulatory Networks with Associated Gene Expression Time Series Data". *University of Groningen*.

[4] Wu D. Thomas, 2001. "Analysing gene expression data from DNA microarrays to identify candidate genes". *Department of Bioinformatics, Genentech, Inc., South San Francisco, Journal of Pathology*; 195: 53–65.

[5] Yi S-G., Joo Y-J. , Park T., 2007. "Rank Based Analysis for the Time-course Microarray Data", VLDB '07, September 23-28, Vienna, Austria.

[6] Yi S-G., Joo Y-J., Park T.,2009. "Rank Based Clustering Analysis for the Time-course Microarray Data". *Bioinformatics and Comp. Biology*, vol. 7, no.1, pp. 75-91.

[7] http://en.wikipedia.org/wiki/Molecular_biology ,[Accessed: May. 07, 2012].

[8] http://biology.about.com/od/geneticsglossary/g/Genes.htm , [Accessed: May. 07, 2012].

[9] Eisen MB, Spellman PT, Brown PO, Botstein D, 1998. "Cluster analysis and display of genome-wide expression patterns". *PNAS Genet.*95:14863–14868.

[10] Brameier M, Wiuf C, 2007. "Co-clustering and visualization of gene expression data and gene ontology terms for Saccharomyces cerevisiae using self-organizing maps". *J Biomed Inform,*40(2):160-173.

[11] http://learn.genetics.utah.edu/content/labs/microarray/analysis/ , [Accessed: May. 07, 2012].

[12] Lelandais G., Marc P., Vincens P., Jacq C., Vialette S., 2004. " MiCoViTo: a tool for gene-centric comparison and visualization of yeast transcriptome states". *BMC Bioinformatics,* 5:20.

[13] Li Zhang, Aidong Zhang, Murali Ramanathan, 2004. "VizStruct: exploratory visualization for gene expression profiling". *Bioinformatics, Vol. 20 no. 1, pages*

85–92.

[14] http://en.wikipedia.org/wiki/Saccharomyces_cerevisiae , [Accessed: May. 07, 2012].

[15] http://biochemie.web.med.uni-muenchen.de/Yeast_Biol/01%20Introduction.pdf [Accessed: May. 07, 2012].

[16] http://www.google.gr/imgres?imgurl=http://www.bio.davidson.edu/courses/genomics/20 04/Bossie/yeastimages.jpg , [Accessed: May. 07, 2012].

[17] Spellman P. T., Sherlock G., Iyer V. R., Zhang M., Anders K., Eisen M. B., Brown P. O., Botstein D. , Futcher B., 1998. "Mol. Biol. Cell*", in press*.

[18] Chu S., DeRisi J., Eisen M., Mulholland J., Botstein D., Brown P. O. , Herskowitz I. , 1998. *Science* 282**,** 699–705.

[19] DeRisi, J. L., Iyer, V. R. & Brown, P. O. ,1997. *Science* 278**,** 680–686.

[20] Cherry J. M., Ball C., Weng S., Juvik G., Schmidt R., Adler C., Dunn B., Dwight S., RilesL., MortimerR. K., *et al*. ,1997. *Nature (London)* 387**,** 67–73.

[21] Mewes HW., Albermann K., Heumann K., Liebl S., Pfeiffer F., 1997. "MIPS: a database for protein sequences, homology data and yeast genome information". *Nucleic Acids Res,* 25(1):28-30.

[22] Ruepp A., Zollner A., Maier D., Albermann K., Hani J., Mokrejs M., Tetko I., Güldener U., Mannhaupt G., Münsterkötter M., Mewes HW., 2004. "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes." *Nucleic Acids Res,* 32(18):5539-45.

[23] http://en.wikipedia.org/wiki/Chondrocyte , [Accessed: May. 08, 2012].

[24] http://en.wikipedia.org/wiki/Mesenchymal_stem_cell , [Accessed: May. 08, 2012].

[25] Arden N., Nevitt MC., 2006. "Osteoarthritis: epidemiology." *Best Pract Res Clin Rheumatol,* 20(1):3-25.

[26] Schindler OS. ,2011. "Current concepts of articular cartilage repair". *Acta Orthop Belg.* ,77(6):709-26.

[27] Bernstein P., Strict C., Jacobi A., Liebers C., Manthey S., Stiehler M., 2010. "Expression pattern differences between osteoarthritic chondrocytes and mesenchymal stem cells during chondrogenic differentiation". *Osteoarthritis and Cartilage*, 18 , 1596-1607.

[28] Thomas P. D., Campbell M. J., Kejariwal A., Mi H., Karlak B., Daverman R., Diemer K., Muruganujan A., Narechania A., 2003. "PANTHER: a library of protein

families and ubfamilies indexed by function".*Genome Res*, vol. 13, pp. 2129-2141.

[29] Filippone M. , Camastra F. , Masulli F. , Rovetta S. , 2008. "A survey of kernel and spectral methods for clustering". *Pattern Recognition (PR),* vol. 41, no. 1, pp. 176–190.

[30] Μαρωνίδης Α., 2010. "Εμβύθιση Γράφων με χρήση Υποκλάσεων στην Κατηγοριοποίηση Δεδομένων".

[31] Jain A. , Murty M. , Flynn P. , 1999. "Data Clustering: A Review". *Association for Computing Machinery (ACM) Computing Survey*, vol. 31, no. 3, pp. 264–323.

[32] Eisen MB, Spellman PT, Brown PO, Botstein D., 1998. "Cluster analysis and display of genome-wide expression patterns*". Proc Natl Acad Sci USA* 95(25):14863–8.

[33] Tavazoie S., Hughes J.D., Campbell M.J., Cho R.J., Church J.M., 1999. "Systematic determination of genetic network architecture". *Nature Genetic*, 22, 281-285.

[34] Hastie T., Tibshirani R., Friedman J., 2009. "14.3.12 Hierarchical Clustering". *The Elements of Statistical Learning* (2nd ed.). New York: Springer. pp. 520–528. ISBN 0-387-84857-6.

[35] Press WH., Teukolsky SA., Vetterling WT., Flannery BP., 2007. "Section 16.4. Hierarchical Clustering by Phylogenetic Trees". *Numerical Recipes: The Art of Scientific Computing (3rd ed.).* New York: Cambridge University Press. ISBN 978-0-521-88068-8.

[36] Sudhir Varma, Richard Simon, "Iterative class discovery and feature selection using Minimal Spanning Trees", *Biometric Research Branch, National Cancer Institute, Rockville, USA, BMC Bioinformatics* 2004, 5:126 doi:10.1186/1471-2105-5-126.

[37] McLachlan G. J , Krishnan T., 2008. "The EM algorithm and extensions*", Wiley series in probability and statistics*. Wiley, Hoboken, NJ, 2nd edition.

[38] Ng A. Y. , Jordan M. I. , Weiss Y. ,2001. "On spectral clustering: Analysis and an algorithm". *In Advances in Neural Information Processing Systems*, pp. 849–856, MIT Press.

[39] Διαμαντάρας Κ., 2007."Τεχνητά Νευρωνικά Δίκτυα". Κλειδάριθμος.

[40] Kohonen T., 1982. "Analysis of a simple self-organizing process*". Biological Cybernetics*, vol. 44, no. 2, pp. 135-140.

[41] Kohonen T. , 1982. "Self-organizing formation of topologically correct feature

maps". *Biological Cybernetics*, vol. 43, no. 1, pp. 59-69.

[42] Kohonen T., 2001. "Self-Organizing Maps". 3rd edition, Springer Verlag.

[43] http://www.statsoft.com/textbook/neural-networks/#sofm [Accessed: April. 09, 2012].

[44] Kohonen T., 1984. "Self-Organization and Associative Memory". Berlin: Springer-Verlag.

[45] Kohonen T., 1990. "The self-organizing map".*Proc. IEEE*, vol. 78, no. 9, pp. 1464-1480.

[46] Aroui T., Koubaa Y., Toumi A., 2009. "Clustering of the Self-Organizing Map based Approach in Induction Machine Rotor Faults Diagnostics". *Leonardo Journal of Sciences,* 15, p. 1-14.

[47] http://www.ai-junkie.com/ann/som/som1.html , [Accessed: April. 10, 2012].

[48] http://www.cs.bham.ac.uk/~jlw/sem2a2/Web/Kohonen.htm , [Accessed: April. 11, 2012].

[49] http://www.codeproject.com/KB/recipes/sofm.aspx ,[Accessed: April. 11, 2012].

[50] Tang B., Heywood M. I., Stepherd M., 2002. "Input Partitioning to Mixtures of Experts". *Neural Networks. IJCNN '02. Proceedings of the 2002 International Joint Conference on*, vol. 1, pp. 227-232.

[51] Yu Hen Hu, 1997. "A patient-Adaptable EGG Beat Classifier Using a Mixture of Experts Approach". *IEEE Transactions of Biomedical Engineering*, vol. 44, no. 9.

[52] Gray R. M. , 1984. "Vector Quantization". *IEEE ASSP Magazine*, pp. 4-29.

[53] Linde Y. , Buzo A. , Gray R. M. , 1980. "An algorithm for vector quantization design". *IEEE trans. On Communications*, vol COM-28, no. 1, pp. 84-95.

[54] Kohonen T. , 1986 . "Learning Vector Quantization for pattern recognition". *Technical Report TKK-F-A601*, Helsinki University of Technology.

[55] MacQueen J. B. , 1967. "Some Methods for classification and Analysis of Multivariate Observations". *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability,* Berkeley, University of California Press, 1:281-297.

[56] http://www-2.cs.cmu.edu/~awm/tutorials/kmeans.html ,[Accessed: April 20,2012].

[57] http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html [Accessed: April 20,2012].

[58] MacKay D. ,2003. "Chapter 20. An Example Inference Task: Clustering". Information Theory. *Inference and Learning Algorithms*. Cambridge University

Press. pp. 284–292. ISBN 0-521-64298-1. MR2012999.

[59] Stein B., Meyer S. zu Eissen, Wibrock F., 2003."On Cluster Validity and the Information Need of Users".*3rd IASTED Int. Conf. on Artificial Intelligence and Applications (AIA 03)*, pp. 216-221.

[60] Halkidi M., Batistakis Y., Vazirgiannis M., 2002. "Cluster Validity Methods: Part I". *SIGMOD Record*, Vol. 31, No. 12, pp. 40-45.

[61] Xu R., Wunsch II D. C., 2009. "Clustering". *IEEE Press Series on Computational Intelligence*, Wiley, New York.

[62] Jain A.K., DUBES R.C., 1998. "Algorithms for Clustering Data". *Prentice-Hall advanced reference series*, Prentice-Hall, Inc., Upper Saddle River, NJ.

[63] Bolshakova N. , Azuaje F., 2003. "Cluster Validation Techniques for Genome Expression Data". *Signal Processing,* vol. 83, no. 4, pp. 825–833.

[64] Bolshakova N. , Azuaje F. , 2003. "Improving Expression Data Mining through Cluster Validation". *4th International IEEE EMBS Special Topic Conference Information Technology Applications in Biomedicine*, pp. 19- 22, 24-26.

[65] Rousseeuw P. J. , , 1987. "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *J. Comp Appl.Math*., vol. 20.

[66] Hubert L. , Arabie P., , 1985. "Comparing partitions". *Journal of Classification*, vol. 2, no. 1, pp. 193-218.

[67] Albatineh A. N., Niewiadomska-Bugaj M. , Mihalko D. , 2006. "On Similarity Indices and Correction for Chance Agreement". *Journal of Classification*, vol. 23, pp. 301- 313.

[68] Mirkin B., 1996. "Mathematical Classification and Clustering". *Kluwer Academic Publishers Group.*

[69] Halgamuge S. K., Lipo W., , 2005. "Classification and Clustering for Knowledge Discovery". Springer-Verlag Berlin and Heidelberg GmbH & Co.

[70] Jain A. K., Murty M. N., Flynn P. J., 1999. "*Data Clustering: A Review*".*ACM Computing Surveys*, vol. 31, no. 3.

[71] Everitt B., , 1993. ."Cluster Analysis". Halsted Press, New York.

[72] Dunn J. C., , 1974.. "Well-Separated Clusters and Optimal Fuzzy Partitions".*Journal of Cybernetics*, vol. 4, pp. 95-104.

[73] Davies D., Bouldin D., , 1979. "A cluster separation measure". *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 1(2), pp. 224–227.

[74] Rousseeuw P. J., , 1987. “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis”. *J. Comp Appl.Math*., vol. 20.

[75] Hoon D., Imoto S., Miyano S. ,2002. “Statistical analysis of a small set of time-ordered gene expression data using linear splines”. *Bioinformatics*, vol. 18, pp. 1477– 1485.

[76] Liu H., Tarima S., Borders A.S., Getchell T.V., 2005. “Quadratic regression analysis for gene discovery and pattern experiments”. *BMC Bioinformatics*, vol. 6:106.

[77] Peddada S., Lobenhofer E., Li L., Afshari C., Weinberg C., Umbach D.M., 2003. “Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference”. *Bioinformatics*, vol. 19, , pp. 834– 841.

[78] Madeira S., Teixeira M., Sa´ -Correia I., Oliveira A., 2010. “Identification of Regulatory Modules in Time Series Gene Expression Data Using a Linear Time Biclustering Algorithm”. *IEEE/ACM Trans. On Comp. Biology & Bioinformatics*, vol. 7, no. 1.

[79] Carreiro A., Anunciac O. ¸ Carric J. ¸ Madeira S.,2011. “Prognostic Prediction Through Biclustering-Based Classification of Clinical Gene Expression Time Series”. *Journal of Integrative Bioinformatics*, vol. 8, no. 3.

[80] Jiand L., Tan K.,2004. “Mining Gene Expression Data for Positive and Negative Co-Regulated Gene Clusters”. *Bioinformatics*, vol. 20, no. 16, pp. 2711-2718.

[81] L. Jiand, K. Tan, 2005. “Identifying Time-Lagged Gene Clusters Using Gene Expression Data”. *Bioinformatics*, vol. 21, no. 4, pp. 509-516.

[82] Ghodsi A.,2006. “Dimensionality Reduction A Short Tutorial”.Waterloo, Ontario, Canada.

[83] Friedman V.,2008. "Data Visualization and Infographics" in: *Graphics*, Monday Inspiration.

[84] Post Frits H., Nielson Gregory M. , Bonneau Georges-Pierre, 2002. “ Data Visualization: The State of the Art”. *Research paper TU delft*.

[85] http://www.visual-literacy.org/periodic_table/periodic_table.html , [Accessed: May 1,2012].

[86] Laurens van der Maaten, Eric Postma, Jaap van den Heric, 2009. “Dimensionality Reduction: A Comparative Review”.*Tilburg centre for Creative Computing*, Tilburg University.

[87] Fukunaga K., 1990. “Introduction to Statistical Pattern Recognition”. *Academic*

Press Professional, Inc., San Diego, CA, USA.

[88] Jimenez L.O., Landgrebe D.A., 1997. "Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data".*IEEE Transactions on Systems, Man and Cybernetics*, 28(1):39–54.

[89] Samet H., 2006. "Foundations of Multidimensional and Metric Data Structures". Morgan Kaufmann.

[90] Ding C., He X., Zha H. , Simon H.D., 2002. "Adaptive Dimension Reduction for *Mining*, Clustering High Dimensional Data".*Proceedings of International Conference on Data.*

[91] Hu ., Zahorian S. A.2010. "Dimensionality Reduction Methods for HMM Phonetic Recognition" . ICASSP , Dallas, TX.

[92] Garcia-Lopez F.C., Garcia-Torres M., Melian B., Moreno-Perez J.A., Moreno-Vega J.M.,2006. "Solving feature subset selection problem by a Parallel Scatter Search". *European Journal of Operational Research*, vol. 169, no. 2, pp. 477-489.

[93] Fodor I.K., 2002. "A survey of dimension reduction techniques ". *US DOE Office of Scientific and Technical Information*.

[94] http://en.wikipedia.org/wiki/Statistics ,[Accessed: May. 04, 2012].

[95] Αβδελάς Γ., Σίμος Η.Θ.,2003. " Αριθμητική Γραμμική Άλγεβρα Θεωρία". Συμεών.

[96] Pearson K., "On Lines and Planes of Closest Fit to Systems of Points in Space" . *Philosophical Magazine* 2 (6): 559–572.

[97] Abdi H., Williams L.J. ,2010. "Principal component analysis". *Wiley Interdisciplinary Reviews: Computational Statistics,* 2: 433–459.

[98] http://www.sccg.sk/~haladova/principal_components.pdf , [Accessed: May. 04, 2012].

[99] http://mathcentral.uregina.ca/RR/database/RR.09.95/weston2.html, [Accessed: May. 04, 2012].

[100] http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html#kernel_methods , [Accessed: May. 08, 2012].

[101] Scholkopt B. , Smola A. J., 2001. Cambridge: MIT Press.

[102] Scholkopf B., Smola A., Muller K. R.,1996. "Nonlinear component analysis as a kernel eigenvalue problem". *Neural Computation*, vol. 10, no. 5, pp. 1299-1319.

[103] Scholkopf B., Smola A. J., Muller K. R.,1999. “Kernel Principal Component Analysis”. *Neural Information Processing Systems*, vol. 1327, no. 3, pp. 327-352.

[104] Shawe-Taylor J., Cristianini N., 2004. “Kernel methods for pattern analysis”. Cambridge University Press.