ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ Η.Μ.Μ.Υ.

ΤΟΜΕΑΣ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Διπλωματική Εργασία

Αξιολόγηση των μεθόδων επιλογής γονιδίων.

Χατζημαρκάκης Εμμανουήλ

Εξεταστική Επιτροπή:
Καθηγητής Ζερβάκης Μιχαήλ (Επιβλέπων)
Καθηγητής Σταυρακάκης Γεώργιος
Καθηγητής Μπάλας Κωνσταντίνος

ΧΑΝΙΑ 2006

Ευχαριστώ για την πολύτιμη βοήθειά τους,
τον καθηγητή Ζερβάκη Μιχαήλ
και τον διδακτορικό φοιτητή Μπλαζαντωνάκη Μιχαήλ

ΠΕΡΙΕΧΟΜΕΝΑ

Κεφάλαιο 1 - Εισαγωγή

Μετά από την ολοκλήρωση της αποκωδικοποίησης του ανθρώπινου γονιδιώματος οι επιστήμονες έχουν ήδη αρχίσει να προχωρούν στο επόμενο στάδιο το οποίο είναι η μελέτη της λειτουργίας των γονιδίων και του τρόπου με τον οποίο αλληλεπιδρούν μεταξύ τους. Από τα 30.000 περίπου γονίδια τα οποία αποτελούν το ανθρώπινο γονιδίωμα, η λειτουργία των δύο τρίτων σχεδόν απ' αυτά παραμένει ακόμα άγνωστη.

Μέχρι πρόσφατα οι βιολόγοι είχαν στη διάθεση τους τεχνικές που τους επέτρεπαν να μετρούν την έκφραση ενός περιορισμένου αριθμού γονιδίων και για κάθε γονίδιο έπρεπε να πραγματοποιηθεί διαφορετικό πείραμα. Το γεγονός αυτό καθιστούσε την μέτρηση της έκφρασης των χιλιάδων γονιδίων που αποτελούν το DNA αρκετά δύσκολη. Η τεχνολογία των μικροσυστοιχιών DNA (DNA Microarrays) επέτρεψε για πρώτη φορά την παράλληλη μέτρηση της γονιδιακής έκφρασης εκατοντάδων έως και χιλιάδων γονιδίων με την εκτέλεση ενός και μόνο πειράματος. Μία από τις σημαντικότερες δυνατότητες που προσφέρει αυτή η τεχνολογία, είναι η μέτρηση των επιπέδων έκφρασης των χιλιάδων γονιδίων ενός κυττάρου σε διαφορετικές χρονικές στιγμές ή σε διαφορετικές συνθήκες (π.χ σε υγιείς και καρκινικούς ιστούς). Μ' αυτόν τον τρόπο οι βιολόγοι μπορούν, συγκρίνοντας τους κατάλληλους ιστούς, να εντοπίσουν τα γονίδια που ευθύνονται για μια συγκεκριμένη ασθένεια, προσδιορίζοντας τα γονίδια των οποίων η έκφραση διαφοροποιείται στις παθολογικές συνθήκες. Μπορούν επίσης να προβλέψουν με ικανοποιητική ακρίβεια την εξέλιξη μιας ασθένειας, να εκτιμήσουν το προσδόκιμο χρόνο ζωής ασθενών με ανίατες ασθένειες αλλά και να βρουν άγνωστες μέχρι σήμερα υποκατηγορίες διαφόρων ασθενειών.

Ένα από τα θέματα που έχουν προκύψει σχετικά με την τεχνολογία DNA Microarrays είναι το μεγάλο μέγεθος και η πολυπλοκότητα των δεδομένων τα οποία καθιστούν την ανάλυσή τους αρκετά δύσκολη. Τη λύση σ' αυτό το πρόβλημα έρχονται να δώσουν διάφορες υπολογιστικές και στατιστικές μέθοδοι, οι οποίες έχουν ως βασικό στόχο τη μελέτη των συσχετίσεων μεταξύ των γονιδίων και τον εντοπισμό των γονιδίων που συμπεριφέρονται με παρόμοιο τρόπο. Μέχρι σήμερα έχουν προταθεί αρκετές μέθοδοι για την ανάλυση των δεδομένων. Οι περισσότερες απ' αυτές εντάσσονται σε δύο κατηγορίες: στις μεθόδους επιβλέπουσας

κατηγοριοποίησης και στις μεθόδους μη επιβλέπουσας κατηγοριοποίησης. Στις μεθόδους επιβλέπουσας κατηγοριοποίησης, ένα σύνολο από προομαδοποιημένα αντικείμενα (γονίδια) χρησιμοποιούνται για να κατασκευάσουμε μια συνάρτηση απόφασης, η οποία μας δίνει την δυνατότητα να βρούμε σε ποια ομάδα ανήκουν γονίδια άγνωστης ομάδας. Οι μέθοδοι μη επιβλέπουσας κατηγοριοποίησης, οι οποίες ονομάζονται και μέθοδοι ομαδοποίησης, έχουν ως στόχο τη ταξινόμηση των αντικειμένων (γονίδια ή δείγματα ιστών) σε λογικές κλάσεις, χωρίς καμία γνώση για προϋπάρχουσες ομάδες.

Τα αποτελέσματα που προκύπτουν από τις μεθόδους και των δύο κατηγοριών εξαρτώνται από πολλούς παράγοντες (π.χ. από τον καθορισμό διαφόρων παραμέτρων) και δεν είναι πάντα αξιόπιστα. Το γεγονός αυτό οδήγησε στην ανάγκη χρησιμοποίησης διαφόρων τεχνικών αξιολόγησης ούτε ώστε να διασφαλιστεί η αξιοπιστία των αποτελεσμάτων.

Για την αξιολόγηση των μεθόδων μη επιβλέπουσας κατηγοριοποίησης χρησιμοποιούνται κάποιες τεχνικές οι οποίες ονομάζονται δείκτες εγκυρότητας (validity indices). Οι δείκτες εγκυρότητας εξετάζουν ένα σχήμα ομαδοποίησης που προέκυψε από τις παραπάνω μεθόδους ως προς την **πυκνότητα** των ομάδων και τη **διαχωρισιμότητα** μεταξύ τους. Ένα σχήμα ομαδοποίησης θεωρείται ικανοποιητικό όταν τα αντικείμενα μέσα σε κάθε ομάδα είναι αρκετά όμοια μεταξύ τους (μεγάλη πυκνότητα) και οι ομάδες είναι καλά διαχωρισμένες μεταξύ τους (μεγάλος βαθμός διαχωρισιμότητας).

Κάθε δείκτης μετράει την πυκνότητα και τη διαχωρισιμότητα με διαφορετικό τρόπο ενώ υπάρχουν και δείκτες (όπως π.χ. ο δείκτης isolation) οι οποίοι μετρούν μόνο την διαχωρισιμότητα. Για να αξιολογηθεί σωστά ένα σχήμα ομαδοποίησης, χρησιμοποιούνται συνήθως παρά πάνω από ένας δείκτες ούτως ώστε τα αποτελέσματα να είναι όσο το δυνατόν πιο αξιόπιστα. Οι δείκτες εγκυρότητας χρησιμοποιούνται συνήθως είτε για να προσδιορίσουμε τις βέλτιστες τιμές των παραμέτρων των μεθόδων ομαδοποίησης είτε για να συγκρίνουμε μεθόδους ομαδοποίησης μεταξύ τους. Στην πρώτη περίπτωση εντοπίζουμε τις τιμές των παραμέτρων για τις οποίες οι δείκτες εγκυρότητας δίνουν τις καλύτερες τιμές στο σχήμα ομαδοποίησης, ενώ στη δεύτερη συγκρίνουμε τα σχήματα ομαδοποίησης που προκύπτουν από διάφορες μεθόδους, με βάση τις τιμές που δίνουν οι δείκτες εγκυρότητας.

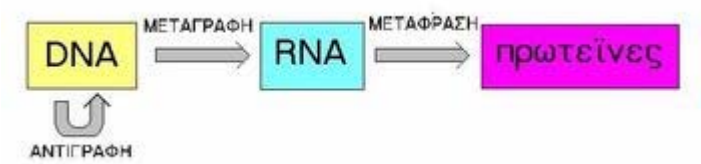
Όσον αφορά τις μεθόδους επιβλέπουσας κατηγοριοποίησης, ένα από τα θέματα που έχουν προκύψει είναι το μεγάλο μέγεθος των δεδομένων το οποίο δημιουργεί αρκετά προβλήματα και δυσκολίες. Το γεγονός αυτό οδήγησε στην ανάγκη χρησιμοποίησης διαφόρων διαδικασιών επιλογής γονιδίων, οι οποίες έχουν ως στόχο τη μείωση των διαστάσεων των δεδομένων. Για να μπορέσουμε να διαπιστώσουμε αν τα αποτελέσματα αυτών των διαδικασιών είναι ικανοποιητικά θα πρέπει να χρησιμοποιήσουμε κάποια κριτήρια αξιολόγησης. Τα κριτήρια αυτά είναι η **ακρίβεια** και η **ποιότητα**. Μία διαδικασία επιλογής γονιδίων έχει υψηλή ακρίβεια όταν τα γονίδια δείκτες (marker genes) που προκύπτουν μπορούν να προβλέψουν σωστά την ομάδα στην οποία ανήκουν κάποια δείγματα άγνωστης ομάδας. Η ποιότητα μιας επιλογής γονιδίων θεωρείται καλή όταν τα γονίδια παρουσιάζουν παραπλήσια συμπεριφορά σε κάθε μία από τις ομάδες που εξετάζουμε, ενώ η συμπεριφορά τους μεταξύ των ομάδων διαφοροποιείται σημαντικά. Μέχρι τώρα, οι περισσότερες μελέτες που ασχολούνται με την αξιολόγηση της επιλογής γονιδίων έχουν επικεντρωθεί σχεδόν αποκλειστικά στην ακρίβεια και ελάχιστα στην ποιότητα των αποτελεσμάτων μιας τέτοιας διαδικασίας.

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι να προτείνει μία τεχνική αξιολόγησης της ποιότητας αυτών των διαδικασιών η οποία βασίζεται σε ένα σύνολο από μεθόδους ομαδοποίησης και δείκτες εγκυρότητας. Η τεχνική αυτή μας δίνει τη δυνατότητα να συγκρίνουμε σύνολα δεδομένων, και κατ' επέκταση τις διαδικασίες επιλογής γονιδίων από τις οποίες προέκυψαν, ως προς την ποιότητα και να εντοπίσουμε τις διαδικασίες εκείνες που δίνουν καλύτερα αποτελέσματα

Κεφάλαιο 2 - Γονιδιακή Έκφραση

2.1 Εισαγωγή

Σύμφωνα με το κεντρικό δόγμα της μοριακής βιολογίας [1] η αποκωδικοποίηση της γενετικής πληροφορίας που βρίσκεται αποθηκευμένη στο DNA ενός οργανισμού γίνεται μέσω των διαδικασιών της μεταγραφής και μετάφρασης. Κατά τη διαδικασία της μεταγραφής η γενετική πληροφορία μετατρέπεται από μορφή DNA σε μορφή αγγελιοφόρου RNA (mRNA) ενώ κατά τη διαδικασία της μετάφρασης η γενετική πληροφορία εκφράζεται στη γλώσσα των αμινοξέων δημιουργώντας πρωτεΐνες. Το τελικό προϊόν αυτών των διαδικασιών, οι πρωτεΐνες, είναι αυτές που καθορίζουν τη δομή και τη λειτουργία των κυττάρων και κατ' επέκταση των οργανισμών. Ο όρος γονιδιακή έκφραση χρησιμοποιείται για να εκφράσει τις δύο αυτές διαδικασίες, κατά τις οποίες η γενετική πληροφορία «ρέει» από τα νουκλεϊκά οξέα (DNA και RNA) προς τις πρωτεΐνες.

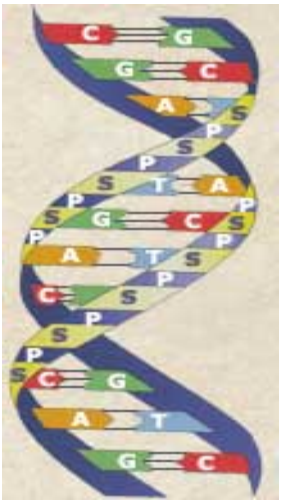


Σχήμα 2.1: Το κεντρικό δόγμα της μοριακής βιολογίας.

Το κεντρικό δόγμα της μοριακής βιολογίας έχει ελαφρώς τροποποιηθεί τα τελευταία χρόνια επειδή η ροή της γενετικής πληροφορίας σε κάποιους ιούς δεν περιγράφεται απόλυτα από το κεντρικό δόγμα όπως είχε εκφραστεί αρχικά. Εμείς για λόγους απλότητας θα αρκεστούμε στην αρχική μορφή του κεντρικού δόγματος, η οποία περιλαμβάνει τις έννοιες που είναι απαραίτητες για την κατανόηση της γονιδιακής έκφρασης.

2.2 DNA

Το DNA είναι ένα νουκλεϊκό οξύ το οποίο μεταφέρει την γενετική πληροφορία που απαιτείται για την βιολογική ανάπτυξη όλων των κυτταρικών μορφών ζωής [2]. Τα μόρια DNA είναι πολυμερή απλούστερων χημικών ενώσεων, των νουκλεοτιδίων. Τα νουκλεοτίδια του DNA σχηματίζονται από μια αζωτούχο βάση (αδενίνη, θυμίνη, γουανίνη ή κυτοσίνη) ενωμένη με ένα σάκχαρο (δεσοξυριβόζη) και ένα μόριο φωσφορικού οξέος. Το DNA αποτελείται από δύο ακολουθίες (αλυσίδες) νουκλεοτιδικών βάσεων μέσα στις οποίες περιλαμβάνονται γονίδια, το καθένα από τα οποία καταλαμβάνει μια έκταση αρκετών χιλιάδων βάσεων. Η συγκεκριμένη σειρά σύμφωνα με την οποία ενώνονται οι βάσεις κατά μήκος του μορίου DNA λειτουργεί σαν κώδικας που περιέχει την γενετική πληροφορία για την σύνθεση των πρωτεϊνών. Με αυτό τον κώδικα τα γονίδια, τα οποία περιέχουν συνεχόμενες τριπλέτες βάσεων, καθορίζουν τη σειρά των αμινοξέων κατά τη σύνθεση των πρωτεϊνών και κατά συνέπεια το είδος των πρωτεϊνών. Τα μόρια DNA έχουν μορφή διπλής έλικας και σχηματίζονται κατά τέτοιο τρόπο, ώστε απέναντι από την αδενίνη να υπάρχει θυμίνη και απέναντι από τη γουανίνη να υπάρχει κυτοσίνη (σχήμα 2.2). Οι αντιστοιχίες αυτές είναι αποτέλεσμα των χημικών δεσμών που δημιουργούνται ανάμεσα στις συμπληρωματικές βάσεις. Πιο συγκεκριμένα, ανάμεσα στις βάσεις θυμίνη και αδενίνη σχηματίζεται διπλός δεσμός υδρογόνου, ενώ ανάμεσα στις βάσεις γουανίνη και κυτοσίνη σχηματίζεται τριπλός δεσμός υδρογόνου. Σαν συνέπεια αυτών των χημικών δεσμών, οι δύο ακολουθίες βάσεων που συνθέτουν τα μόρια DNA είναι συμπληρωματικές η μία ως προς την άλλη. Μονές αλυσίδες DNA, οι οποίες συναντώνται μόνο σε ειδικές συνθήκες (είτε στο εργαστήριο είτε κατά τη φάση της μεταγραφής), έχουν την ιδιότητα να ενώνονται με τις συμπληρωματικές τους αλυσίδες. Αυτή η ιδιότητα ονομάζεται υβριδισμός και αποτελεί το βασικό εργαλείο στις περισσότερες από τις τεχνικές της μοριακής βιολογίας.



Σχήμα 2.2: Η διπλή έλικα του DNA. Όπου A = αδενίνη, T = θυμίνη, G = γουανίνη, C = κυτοσίνη, S = δεσοξυριβόζη, P = φωσφορική ομάδα

2.3 RNA Και Μεταγραφή

Το RNA είναι κι αυτό ένα νουκλεϊκό οξύ το οποίο παράγεται από το DNA κατά τη φάση της μεταγραφής και χρησιμοποιείται για την σύνθεση των πρωτεϊνών [3]. Τα μόρια RNA παρόλο που μοιάζουν αρκετά με τα μόρια DNA έχουν κάποιες σημαντικές διαφορές. Πιο συγκεκριμένα, έχουν μόνο μία αλυσίδα βάσεων και αντί για θυμίνη (T) περιέχουν ουρακίλη (U). Τα μόρια RNA χωρίζονται σε τρεις ομάδες ανάλογα με το ρόλο που έχουν κατά τη σύνθεση των πρωτεϊνών. Οι ομάδες αυτές είναι: το rRNA (ribosomal RNA), το οποίο αποτελεί δομικό συστατικό των ριβοσωμάτων, το tRNA (transport RNA), το οποίο είναι υπεύθυνο για την μεταφορά των αμινοξέων στα ριβοσώματα και το mRNA (messenger RNA), το οποίο είναι υπεύθυνο για την μεταφορά της γενετικής πληροφορίας που περιέχεται στα γονίδια στο ριβόσωμα, όπου θα γίνει η πρωτεϊνοσύνθεση. Το mRNA παράγεται κατά τη φάση της αντιγραφής με την εξής διαδικασία:

Στο τμήμα του DNA που περιέχει τη γενετική πληροφορία την οποία το κύτταρο θέλει να μεταγράψει, σπάνε οι δεσμοί υδρογόνου που συγκρατούν τις αζωτούχες βάσεις και ανοίγει η διπλή έλικα. Αρχίζει στη συνέχεια η σύνθεση ενός μορίου

mRNA, με πρότυπο τον ένα από τους δύο κλώνους του DNA που φέρει την πληροφορία για τη σύνθεση μιας συγκεκριμένης πρωτεΐνης. Απέναντι από κάθε νουκλεοτίδιο αυτού του κλώνου τοποθετείται ένα άλλο νουκλεοτίδιο σύμφωνα με την αρχή της συμπληρωματικότητας των βάσεων. Η μόνη διαφορά σε σχέση με την αντιστοιχία των βάσεων στο DNA, είναι ότι απέναντι από κάθε νουκλεοτίδιο του μεταγραφόμενου κλώνου που περιέχει αδενίνη, τοποθετείται ένα νουκλεοτίδιο που περιέχει ουρακίλη. Το ένζυμο RNA πολυμεράση συνδέει τα νέα νουκλεοτίδια, τα οποία προστίθενται το ένα μετά το άλλο με ομοιοπολικό δεσμό. Όταν ολοκληρωθεί η διαδικασία, έχει πλέον συντεθεί ένα μονόκλωνο μόριο mRNA του οποίου η αλληλουχία των νουκλεοτιδίων «υπαγορεύτηκε» από την αλληλουχία των νουκλεοτιδίων του μεταγραφόμενου τμήματος του DNA, δηλαδή από ένα γονίδιο. Η διαδικασία αυτή ονομάστηκε μεταγραφή επειδή η γενετική πληροφορία που ήταν γραμμένη στη γλώσσα του DNA (A,T,G,C), «μεταγράφεται» στη γλώσσα του RNA στην οποία αντί της θυμίνης χρησιμοποιείται η βάση ουρακίλη.

2.4 Πρωτεΐνες Και Μετάφραση

Μετάφραση είναι η διαδικασία της σύνθεσης των πρωτεϊνών από την γενετική πληροφορία που περιέχεται στο RNA [3]. Κατά τη διαδικασία αυτή, η οποία λαμβάνει χώρα στα ριβοσώματα, η αλληλουχία (σειρά) των νουκλεοτιδίων του mRNA «υπαγορεύει» την παραγωγή μιας πολυπεπτιδικής αλυσίδας με καθορισμένη αλληλουχία αμινοξέων. Με τον τρόπο αυτό η γενετική πληροφορία, που είναι καταγεγραμμένη στα νουκλεϊκά οξέα στη γλώσσα των τεσσάρων γραμμάτων (A,T,G,C για το DNA και A,U,G,C για το RNA), μεταφράζεται σε μια εντελώς διαφορετική γλώσσα με 20 διαφορετικά γράμματα, όσα είναι δηλαδή τα διαφορετικά αμινοξέα που συνθέτουν τις πρωτεΐνες όλων των οργανισμών. Κάθε ένα από τα αμινοξέα κωδικοποιείται από μια τριάδα νουκλεοτιδίων η οποία ονομάζεται κωδικόνιο. Οι κανόνες σύμφωνα με τους οποίους συγκεκριμένες τριάδες νουκλεοτιδίων (κωδικόνια) κωδικοποιούν συγκεκριμένα αμινοξέα, αποτελούν τον γενετικό κώδικα.

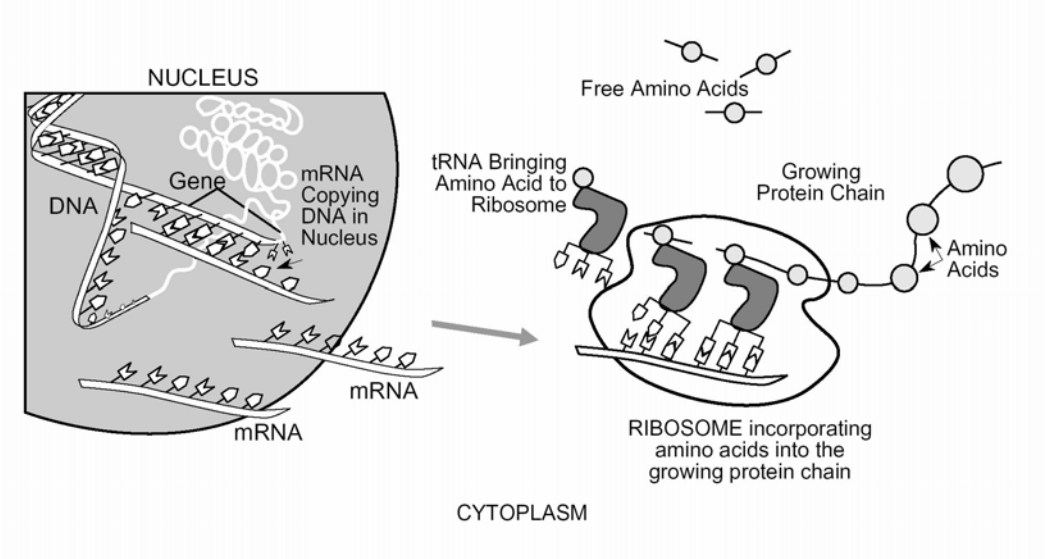
Εκτός από το mRNA, τα ριβοσώματα και τα αμινοξέα, στην πρωτεϊνοσύνθεση μετέχει επίσης και το tRNA. Το tRNA διαθέτει μια χαρακτηριστική τριάδα

νουκλεοτιδίων που λέγεται αντικωδικόνιο και είναι συμπληρωματική με ένα κωδικόνιο του mRNA. Έτσι τα διάφορα είδη tRNA μπορούν να αναγνωρίζουν τα κωδικόνια που είναι συμπληρωματικά των αντικωδικονίων τους, και να συνδέονται μαζί τους με δεσμούς υδρογόνου. Το tRNA διαθέτει επίσης μια θέση σύνδεσης με ένα αμινοξύ. Κάθε μόριο tRNA, ανάλογα με το αντικωδικόνιό του, συνδέεται με ένα συγκεκριμένο είδος αμινοξέος.

Η διαδικασία της μετάφρασης περιλαμβάνει τρία στάδια: την έναρξη, την επιμήκυνση και τη λήξη. Κατά την έναρξη, το mRNA που έχει συντεθεί στον πυρήνα του κυττάρου μεταναστεύει στο κυτταρόπλασμα και συνδέεται με ένα ριβόσωμα σε συγκεκριμένη θέση. Το πρώτο κωδικόνιο που «διαβάζει» το ριβόσωμα είναι το AUG (αδενίνη – ουρακίλη – γουανίνη) το οποίο χαρακτηρίζεται ως κωδικόνιο έναρξης γιατί σηματοδοτεί την έναρξη της πρωτεϊνοσύνθεσης. Ταυτόχρονα μεταφέρεται και συνδέεται στο ριβόσωμα ένα μόριο tRNA, το οποίο φέρει το αμινοξύ μεθειονίνη και έχει αντικωδικόνιο συμπληρωματικό του κωδικονίου έναρξης. Στο στάδιο της επιμήκυνσης, ένα δεύτερο μόριο tRNA με αντικωδικόνιο συμπληρωματικό του δευτέρου κατά σειρά κωδικονίου τοποθετείται στο ριβόσωμα δίπλα στο πρώτο, μεταφέροντας εκεί το δεύτερο αμινοξύ. Ανάμεσα στο δεύτερο αμινοξύ και στη μεθειονίνη δημιουργείται ένας δεσμός που τα συγκρατεί ενωμένα. Στη συνέχεια το πρώτο tRNA αποδεσμεύει τη μεθειονίνη και απελευθερώνεται στο κυτταρόπλασμα. Η ίδια διαδικασία επαναλαμβάνεται για το τρίτο, τέταρτο κτλ. αμινοξύ επιμηκύνοντας έτσι την πεπτιδική αλυσίδα μέχρι την ολοκλήρωση της σύνθεσής της. Κατά το στάδιο της λήξης το ριβόσωμα φτάνει σε ένα από τα τρία κωδικόνια λήξης (UAG,UAA,UGA), η πρωτεϊνοσύνθεση σταματάει και η πολυπεπτιδική αλυσίδα απελευθερώνεται από τα ριβοσώματα

Οι πρωτεΐνες, το τελικό προϊόν της γονιδιακής έκφρασης, είναι υπεύθυνες για τα βασικά δομικά και λειτουργικά χαρακτηριστικά των κυττάρων. Με άλλα λόγια η παρουσία των πρωτεϊνών είναι εκείνη που καθορίζει την λειτουργία του κυττάρου. Είναι γνωστό ότι αν και τα περισσότερα κύτταρα στο σώμα μας περιέχουν τα ίδια γονίδια, δεν χρησιμοποιούνται όλα τα γονίδια σε κάθε κύτταρο. Κάποια απ' αυτά ενεργοποιούνται (εκφράζονται) μόνο σε συγκεκριμένα είδη κυττάρων προσδίδοντας τους έτσι «μοναδικά» χαρακτηριστικά. Έτσι, παρόλο που θεωρητικά κάθε κύτταρο έχει τη δυνατότητα να συνθέσει οποιαδήποτε πρωτεΐνη μπορεί να συνθέσει οποιοδήποτε άλλο κύτταρο, στην πράξη κάθε κύτταρο συνθέτει μόνο κάποιες συγκεκριμένες πρωτεΐνες. Αυτός είναι και ο λόγος για τον οποίον διαφορετικά

κύτταρα, ανάλογα με τις πρωτεΐνες που διαθέτουν, έχουν διαφορετικό ρόλο σε έναν οργανισμό.



Σχήμα 2.3: Η διαδικασία της πρωτεϊνοσύνθεσης.

Κεφάλαιο 3 - Μέθοδοι Μέτρησης Γονιδιακής Έκφρασης

3.1 Εισαγωγή

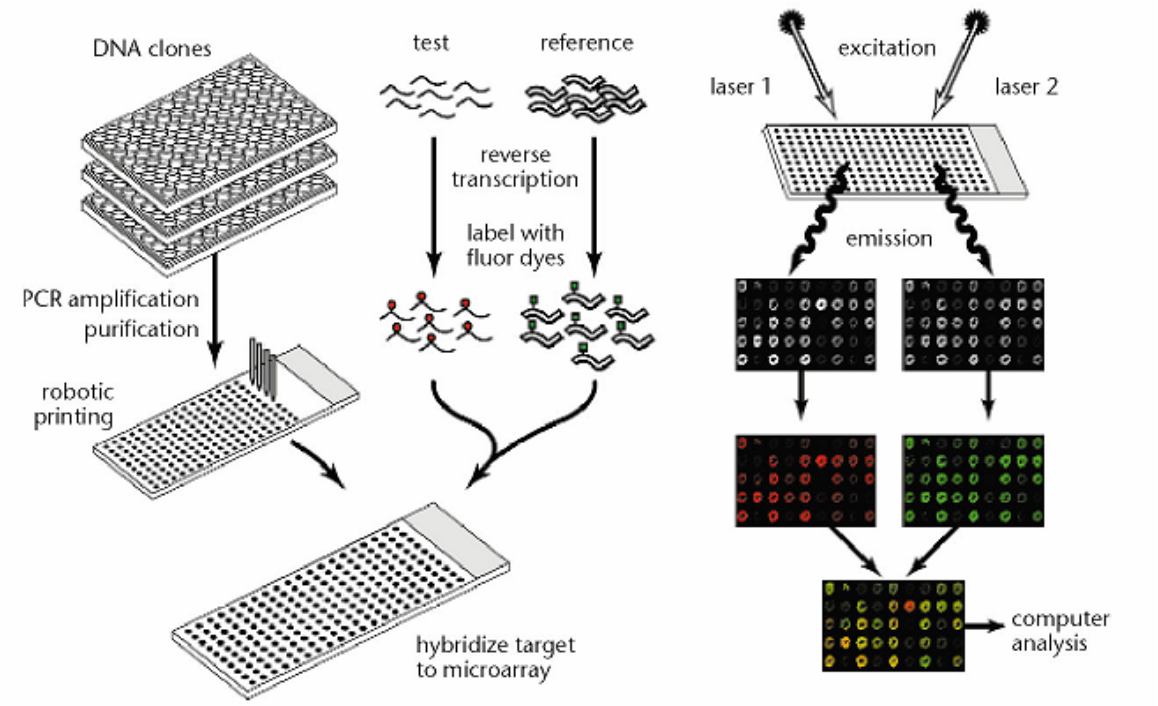
Μέχρι πρόσφατα οι βιολόγοι είχαν στη διάθεση τους τεχνικές που τους επέτρεπαν να μετρούν την έκφραση περιορισμένου αριθμού γονιδίων και για κάθε γονίδιο έπρεπε να πραγματοποιηθεί διαφορετικό πείραμα. Το γεγονός αυτό καθιστούσε πολύ δύσκολη την μέτρηση της έκφρασης των χιλιάδων γονιδίων που αποτελούν το DNA. Ήταν φανερό ότι αυτά τα μεγέθη των δεδομένων απαιτούσαν τεχνικές οι οποίες θα έδιναν μια συνολική εικόνα της συμπεριφοράς των γονιδίων καθώς και της αλληλεπίδρασης μεταξύ τους. Η τεχνολογία των μικροσυστοιχιών DNA (DNA-Microarrays) επέτρεψε για πρώτη φορά την παράλληλη μέτρηση της γονιδιακής έκφρασης εκατοντάδων έως και χιλιάδων γονιδίων με την εκτέλεση ενός και μόνο πειράματος. Υπάρχουν δύο βασικές παραλλαγές αυτής της τεχνολογίας, η complementary DNA (cDNA) microarrays και η oligonucleotide microarrays. Σ' αυτή την εργασία θα ασχοληθούμε με την cDNA Microarrays η οποία είναι και η πιο διαδεδομένη.

3.2 cDNA Microarrays

Μια μικροσυστοιχία DNA είναι ένα πλακίδιο κατασκευασμένο από ειδικό γυαλί πάνω στο οποίο παρατάσσονται σε συγκεκριμένες θέσεις ιχνηθέτες, το πλήθος των οποίων μπορεί να κυμανθεί από μερικές εκατοντάδες έως πολλές χιλιάδες [4 - 6]. Οι ιχνηθέτες αποτελούνται από πολλά αντίγραφα μιας γνωστής αλληλουχίας νουκλεοτιδίων η οποία είναι συμπληρωματική ως προς το τμήμα του DNA που επιθυμούμε να εντοπίσουμε. Πιο συγκεκριμένα, κάθε ένας από τους ιχνηθέτες αντιστοιχεί σε ένα γονίδιο και αποτελείται από το συμπληρωματικό DNA (cDNA) του mRNA του γονιδίου που θέλουμε να μετρηθεί.

Η μέθοδος cDNA microarrays μας δίνει την δυνατότητα να υπολογίσουμε το επίπεδο έκφρασης των γονιδίων σε συγκεκριμένες παθολογικές ή φυσιολογικές συνθήκες. Ο πιο συνηθισμένος σκοπός αυτής της μεθόδου είναι ο εντοπισμός των γονιδίων που εκφράζονται διαφορετικά σε δύο διαφορετικές συνθήκες. Έστω για παράδειγμα ότι θέλουμε να συγκρίνουμε την έκφραση των γονιδίων σε ένα υγιή και σε ένα καρκινικό ιστό. Αρχικά θα απομονώσουμε και στους δύο ιστούς το RNA από το DNA τις πρωτεΐνες και τα υπόλοιπα συστατικά του κυττάρου με κατάλληλη επεξεργασία. Στη συνέχεια θα απομονώσουμε το mRNA από το tRNA και το rRNA, κι αυτό γιατί το mRNA είναι το μοναδικό τμήμα του RNA το οποίο μας δίνει πληροφορίες για την γονιδιακή έκφραση. Στο επόμενο στάδιο το mRNA μεταγράφεται στο συμπληρωματικό του DNA (cDNA) με την βοήθεια ενός ενζύμου που ονομάζεται αντίστροφη μεταγραφάση. Αυτό το ένζυμο έχει την ικανότητα να χρησιμοποιεί το RNA ως μήτρα για να συνθέσει DNA, καθιστώντας έτσι δυνατή την αντίστροφη ροή της γενετικής πληροφορίας από το RNA στο DNA. Ταυτόχρονα τα δείγματα χρωματίζονται με κατάλληλη χρωστική, με πράσινο χρώμα το mRNA αναφοράς και με κόκκινο χρώμα το mRNA μελέτης. Με αυτό τον τρόπο προκύπτουν πράσινα υγιή μόρια cDNA και κόκκινα καρκινικά μόρια cDNA. Η διαδικασία της αντίστροφης μεταγραφής είναι απαραίτητη αφού το mRNA είναι πολύ ασταθές και θα είχε καταστραφεί πριν τελειώσει το πείραμα (σε αντίθεση με το cDNA που είναι σταθερό). Στη συνέχεια το cDNA και των δύο δειγμάτων τοποθετείται πάνω στο πλακίδιο. Τα γονίδιά των δειγμάτων αντιδρούν χημικά με τους αντίστοιχους ιχνηθέτες (υβριδισμός) και ελευθερώνουν στην αντίστοιχη θέση του ιχνηθέτη την χρωστική ουσία. Η ποσότητα της χρωστικής που ελευθερώνεται στην θέση ενός ιχνηθέτη είναι ανάλογη της έκφρασης του αντίστοιχου γονιδίου. Στην συνέχεια ένας οπτικός σαρωτής σαρώνει το πλακίδιο και στην έξοδό του παράγεται μια ψηφιακή εικόνα η οποία αποτελείται από ένα πλήθος κουκκίδων. Κάθε κουκκίδα αντιστοιχεί σε ένα διαφορετικό γονίδιο και η έντασή της αντιστοιχεί στο επίπεδο έκφρασης του. Στις κόκκινες κουκκίδες αντιστοιχούν γονίδια τα οποία εκφράστηκαν περισσότερο στα καρκινικά κύτταρα σε σχέση με τα υγιή κύτταρα (υπερβάλλουσα έκφραση), στις πράσινες κουκκίδες αντιστοιχούν γονίδια τα οποία εκφράστηκαν λιγότερο στα καρκινικά κύτταρα (χαμηλή έκφραση), στις κίτρινες κουκκίδες αντιστοιχούν γονίδια τα οποία εκφράστηκαν με τον ίδιο τρόπο σε καρκινικά και υγιή κύτταρα ενώ στις μαύρες κουκκίδες αντιστοιχούν τα γονίδια που δεν εκφράστηκαν καθόλου. Η εικόνα αναλύεται με κατάλληλο λογισμικό με αποτέλεσμα την παραγωγή ενός διανύσματος,

του οποίου οι μεταβλητές είναι μετρήσεις γονιδιακής έκφρασης των γονιδίων που είχαν επιλεγεί προς μέτρηση μέσω των κατάλληλων ιχνηθετών. Πραγματοποιώντας μια σειρά από τέτοια πειράματα με διαφορετικά δείγματα DNA κατασκευάζεται ένας πίνακας γονιδιακής έκφρασης. Οι στήλες του πίνακα αντιστοιχούν σε διαφορετικά δείγματα ενώ οι γραμμές σε διαφορετικά γονίδια.

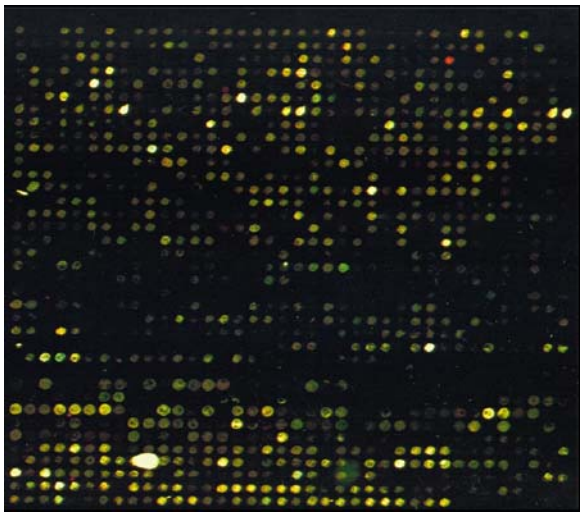


Σχήμα 3.1: Η μέθοδος cDNA Microarrays

Το πιο σημαντικό πλεονέκτημα της cDNA Microarrays είναι ότι μπορεί να μετρήσει την έκφραση ενός πολύ μεγάλου αριθμού γονιδίων ταυτόχρονα. Ο αριθμός αυτός μπορεί να φτάσει μέχρι και τα 10.000 γονίδια σε ένα πλακίδιο. Στην πραγματικότητα μάλιστα δεν υπάρχει κανένας περιορισμός όσον αφορά τον αριθμό των γονιδίων αφού μπορούν να χρησιμοποιηθούν παρά πάνω από ένα πλακίδια για να μετρηθεί το RNA ενός κυττάρου.

Ένα από τα μειονεκτήματα της μεθόδου cDNA Microarrays είναι ότι διαφορετικές μετρήσεις σε δείγματα από το ίδιο κύτταρο μπορεί να οδηγήσουν σε διαφορετικά αποτελέσματα. Αυτό οφείλεται στο γεγονός ότι αρκετά από τα στάδια της διαδικασίας επηρεάζονται από το περιβάλλον και τον τρόπο της εκτέλεσης, προκαλώντας έτσι μεταβλητότητα στα τελικά αποτελέσματα. Το πρόβλημα είναι πιο

έντονο στο στάδιο της επεξεργασίας της εικόνας κατά το οποίο η έλλειψη ακρίβειας του οπτικού σαρωτή σε συνδυασμό με την περιορισμένη ακρίβεια των ρομπότ στη τοποθέτηση των ιχνηθετών δημιουργούν πρόσθετο θόρυβο στα τελικά αποτελέσματα. Αυτό γίνεται φανερό στην παρακάτω εικόνα στην οποία φαίνεται ότι οι κουκκίδες δεν είναι απόλυτα διαχωρίσιμες ενώ κάποιες από τις γειτονικές κουκκίδες έχουν σχεδόν συγχωνευτεί.



Σχήμα 3.2: Εικόνα πλακιδίου μετά από τη σάρωση του οπτικού σαρωτή

Αυτά τα προβλήματα αντιμετωπίζονται μερικώς, όπως ήδη έχει αναφερθεί, με την χρησιμοποίηση RNA από ένα συγκεκριμένο κύτταρο (RNA αναφοράς) σε όλες τις μετρήσεις (σε όλα τα πλακίδια) ενός πειράματος. Το τελικό επίπεδο έκφρασης ενός γονιδίου υπολογίζεται, όπως θα δούμε και παρακάτω, από τον λογάριθμο του λόγου του επιπέδου έκφρασης του γονιδίου στο δείγμα μελέτης (Cy5) προς το επίπεδο έκφρασης του γονιδίου στο δείγμα αναφοράς (Cy3).

$$e = \log(Cy5/Cy3)$$

Ένας άλλος παράγοντας που προκαλεί μεταβλητότητα στα τελικά αποτελέσματα και είναι πολύ συνηθισμένος στην τεχνική cDNA Microarrays λέγεται bleaching. Στην πλειοψηφία των σαρωτών, τα πράσινα και τα κόκκινα σημάδια από την χρωστική σαρώνονται ξεχωριστά. Το πρόβλημα είναι ότι η ένταση της τελικής κουκκίδας επηρεάζεται από την σειρά με την οποία γίνεται η σάρωση. Για να

αντιμετωπιστεί το πρόβλημα χρησιμοποιούνται συγκεκριμένες διαδικασίες κανονικοποίησης των πράσινων και κόκκινων σημάτων.

Το πιο σημαντικό μειονέκτημα αυτής της τεχνικής αυτής αλλά και γενικότερα όλων των τεχνικών microarrays είναι το κόστος το οποίο παραμένει ακόμα πολύ υψηλό, με αποτέλεσμα να τις καθιστά απαγορευτικές για την πλειοψηφία των εργαστηρίων.

Παρόλα τα παραπάνω μειονεκτήματα η τεχνολογία cDNA Microarrays αποτελεί μια επανάσταση για την μοριακή βιολογία και οι δυνατότητες που προσφέρει είναι πολλές και σημαντικές. Ένα από τα πιο συνηθισμένα πεδία εφαρμογής της συγκεκριμένης τεχνολογίας είναι η πρόγνωση και η διάγνωση διαφόρων ασθενειών. Η τεχνολογία αυτή μας επιτρέπει να μετρήσουμε και να συγκρίνουμε τα επίπεδα έκφρασης των γονιδίων σε υγιή και μη κύτταρα και να καταλήξουμε σε χρήσιμα συμπεράσματα. Μπορούμε για παράδειγμα να βρούμε ποια γονίδια ευθύνονται για μια συγκεκριμένη ασθένεια προσδιορίζοντας τα γονίδια των οποίων η έκφραση διαφοροποιείται στις παθολογικές συνθήκες. Μπορούμε επίσης, συγκρίνοντας τους κατάλληλους ιστούς, να βρούμε άγνωστες μέχρι τώρα υποκατηγορίες συγκεκριμένων ασθενειών. Ο προσδιορισμός των υποκατηγοριών μιας ασθένειας μπορεί να έχει ιδιαίτερη πρακτική αξία, αφού μας δίνει την δυνατότητα να χρησιμοποιήσουμε διαφορετικές και πιο εξειδικευμένες θεραπείες για κάθε υποκατηγορία. Μια άλλη δυνατότητα που μας δίνει η τεχνολογία cDNA Microarrays είναι η εκτίμηση για την εξέλιξη που θα έχει μια ασθένεια σε κάποιον ασθενή. Μπορούμε για παράδειγμα να συγκρίνουμε ιστούς από ασθενείς που έζησαν πολλά χρόνια μετά την διάγνωση της ασθένειας με ιστούς από άλλους ασθενείς με την ίδια ασθένεια οι οποίοι έζησαν λιγότερα χρόνια. Με αυτό τον τρόπο μπορούμε να προσδιορίσουμε τα γονίδια τα οποία σχετίζονται με μεγαλύτερο ή μικρότερο προσδόκιμο χρόνο ζωής. Όλα τα παραπάνω, αν και αποτελούν ένα μικρό μέρος μόνο των εφαρμογών της τεχνολογίας DNA Microarrays, είναι ενδεικτικά της χρησιμότητας και των δυνατοτήτων που προσφέρει η συγκεκριμένη τεχνολογία. Το γεγονός μάλιστα ότι οι έρευνες βρίσκονται ακόμα σε σχετικά πρώιμο στάδιο δημιουργεί υψηλές προσδοκίες για την περαιτέρω αξιοποίηση της τεχνολογίας στο μέλλον.

Κεφάλαιο 4 - Κανονικοποίηση

4.1 Εισαγωγή

Μετά από τον υπολογισμό των δεδομένων γονιδιακής έκφρασης με τη μορφή αριθμητικών αποτελεσμάτων και πριν από την ανάλυση αυτών των δεδομένων , είναι απαραίτητο να μεσολαβήσει η κανονικοποίηση τους. Ο όρος κανονικοποίηση χρησιμοποιείται για να περιγράψει μια σειρά από μεθόδους οι οποίες έχουν ως βασικό στόχο την αύξηση της αξιοπιστίας των πειραματικών μετρήσεων που προέκυψαν από τη μέθοδο DNA Microarrays και την μετατροπή των δεδομένων σε μια συγκεκριμένη μορφή, τέτοια ώστε να επιτρέπει την εφαρμογή των μεθόδων ανάλυσης. Αρκετές απ' αυτές τις μεθόδους δεν είναι καθολικά αποδεκτές. Κάποιες ομάδες ερευνητών τις αποδέχονται και τις χρησιμοποιούν και κάποιες άλλες τις αμφισβητούν. Οι πιο σημαντικές μέθοδοι κανονικοποίησης οι οποίες τυγχάνουν ευρύτατης αποδοχής είναι το φιλτράρισμα, η λογαριθμική τροποποίηση καθώς και η ρύθμιση των μέσων όρων και των τυπικών αποκλίσεων των γονιδίων και/ή των δειγμάτων [7-8].

4.2 Λογαριθμική Τροποποίηση

Στο στάδιο της λογαριθμικής τροποποίησης γίνεται αντικατάσταση κάθε τιμής X του πίνακα γονιδιακής έκφρασης με την τιμή $\log_2(X)$. Αιτία αυτής τροποποίησης αποτελεί το γεγονός ότι οι τιμές του πίνακα, οι οποίες αποτελούν λόγο εκφράσεων (RNA μελέτης/RNA αναφοράς), είναι τιμές διαφορετικού μεγέθους στις περιπτώσεις υπερβάλλουσας έκφρασης σε σχέση με τις περιπτώσεις χαμηλής έκφρασης. Πιο συγκεκριμένα, στις περιπτώσεις υπερβάλλουσας έκφρασης οι τιμές των γονιδίων κυμαίνονται μεταξύ του ενός και του $+\infty$ ενώ στις περιπτώσεις χαμηλής έκφρασης οι τιμές περιορίζονται μεταξύ του μηδενός και του ένα. Αν για παράδειγμα το επίπεδο έκφρασης ενός γονιδίου στο δείγμα μελέτης είναι διπλάσιο σε σχέση με το δείγμα αναφοράς, τότε θα έχουμε στον πίνακα την τιμή 2. Αν αντίθετα το επίπεδο έκφρασης

ενός γονιδίου στο δείγμα μελέτης είναι το μισό σε σχέση με το δείγμα αναφοράς, τότε θα έχουμε στον πίνακα την τιμή ½. Αν συγκρίνουμε τις τιμές 2 και ½ με τον αριθμό 1 που εκφράζει την περίπτωση της ίσης έκφρασης του γονιδίου στα δείγματα μελέτης και αναφοράς παρατηρούμε ότι η τιμή 2 απέχει περισσότερο από το 1 σε σχέση με την τιμή ½. Το γεγονός αυτό αποτελεί ένα πρόβλημα που πρέπει να λύσουμε αφού και στις δύο περιπτώσεις η διαφοροποίηση του γονιδίου έχει την ίδια φυσική σπουδαιότητα, το ίδιο φυσικό βάρος, κάτι που παρόλα αυτά δεν αποτυπώνεται στις αριθμητικές τιμές. Για να αντιμετωπίσουμε αυτό το πρόβλημα χρησιμοποιούμε την λογαριθμική τροποποίηση η οποία γίνεται συνήθως με βάση το 2 ($\log_2(X)$). Στο παραπάνω παράδειγμα από τις τιμές 2 και ½ θα προκύψουν οι τιμές 1 και -1 οι οποίες παρέχουν την απαραίτητη συμμετρία σε σχέση με την τιμή 0 που εκφράζει την περίπτωση της μη διαφοροποίησης του γονιδίου στα δείγματα μελέτης και αναφοράς.

4.3 Ρύθμιση Των Μέσων Όρων Και Των Τυπικών Αποκλίσεων Των Γονιδίων Και / Ή Των Δειγμάτων

Οι τιμές γονιδιακής έκφρασης ενός γονιδίου για όλα τα δείγματα ενός πίνακα αποτελούν ένα διάνυσμα γονιδιακής έκφρασης. Αυτά τα διανύσματα χρησιμοποιούνται, όπως θα δούμε παρακάτω, για την ταξινόμηση των γονιδίων σε ομάδες. Ένα από τα προβλήματα που παρατηρούνται σε αρκετές περιπτώσεις κατά την ταξινόμηση των γονιδίων είναι ότι γονίδια με παρόμοια επίπεδα έκφρασης εμφανίζονται να έχουν διαφορετικό συνολικό βαθμό έκφρασης (διαφορετικό expression profile). Αυτό σημαίνει ότι τα διανύσματα των γονιδίων έχουν παρόμοιες διευθύνσεις αλλά διαφορετικά μήκη. Το γεγονός αυτό, το οποίο συνήθως προκύπτει από τις συνθήκες του πειράματος και δεν οφείλεται σε βιολογικούς λόγους, έχει πολλές φορές ως αποτέλεσμα τα γονίδια που έχουν παρόμοια επίπεδα έκφρασης να ταξινομούνται λανθασμένα σε διαφορετικές ομάδες. Αιτία αυτού του προβλήματος αποτελεί το γεγονός ότι για κάθε δείγμα που χρησιμοποιούμε εκτελείται ένα διαφορετικό πείραμα με τη μέθοδο DNA Microarrays. Σε κάθε πείραμα η αντίδραση του υβριδισμού μπορεί να είναι ελαφρώς διαφορετική με αποτέλεσμα πολλές φορές να προκύπτουν πλασματικές διαφορές ανάμεσα στις τιμές των γονιδίων. Για να

μπορέσουμε να συγκρίνουμε αποτελέσματα τα οποία έχουν προκύψει από διαφορετικά πειράματα θα πρέπει να εφαρμόσουμε την παρακάτω διαδικασία:

Υπολογίζουμε τον μέσο όρο των τιμών για κάθε γονίδιο του πίνακα (για κάθε γραμμή) και στη συνέχεια αφαιρούμε από όλες τις τιμές κάθε γονιδίου τον μέσο όρο ούτως ώστε ο νέος μέσος όρος που θα προκύψει να είναι μηδέν. Οι ρυθμίσεις των μέσων όρων γίνονται με βάση τον τύπο:

$$y_{ij} = x_{ij} - \overline{x_i}$$

όπου y_{ij} οι τιμές γονιδιακής έκφρασης μετά τη ρύθμιση των μέσων όρων,

x_{ij} οι τιμές γονιδιακής έκφρασης πριν τη ρύθμιση των μέσων όρων

και $\overline{x_i}$ ο μέσος όρος των τιμών του γονιδίου i για όλα τα δείγματα j ($1 \le j \le n$)

ο οποίος δίνεται από τον τύπο

$$\overline{x_i} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$$

Ομοίως μπορούμε να πράξουμε και για τα δείγματα (τις στήλες του πίνακα) αν αυτό που μας ενδιαφέρει είναι η ταξινόμηση των δειγμάτων, ενώ αν θέλουμε να κάνουμε ταξινόμηση και στα γονίδια και στα δείγματα μπορούμε να εφαρμόσουμε την ίδια πρακτική και στις γραμμές και στις στήλες.

Αφού ρυθμίσουμε τους μέσους όρους προχωράμε στη ρύθμιση των τυπικών αποκλίσεων. Η ρύθμιση αυτή επιτυγχάνεται διαιρώντας όλες τις τιμές για κάθε γονίδιο με την τυπική απόκλιση (standard deviation) του γονιδίου ούτως ώστε η νέα τυπική απόκλιση που προκύπτει να είναι ίση με τη μονάδα. Οι ρυθμίσεις των τυπικών αποκλίσεων γίνονται με βάση τον τύπο:

$$y_{ij} = \frac{x_{ij}}{\sigma_i}$$

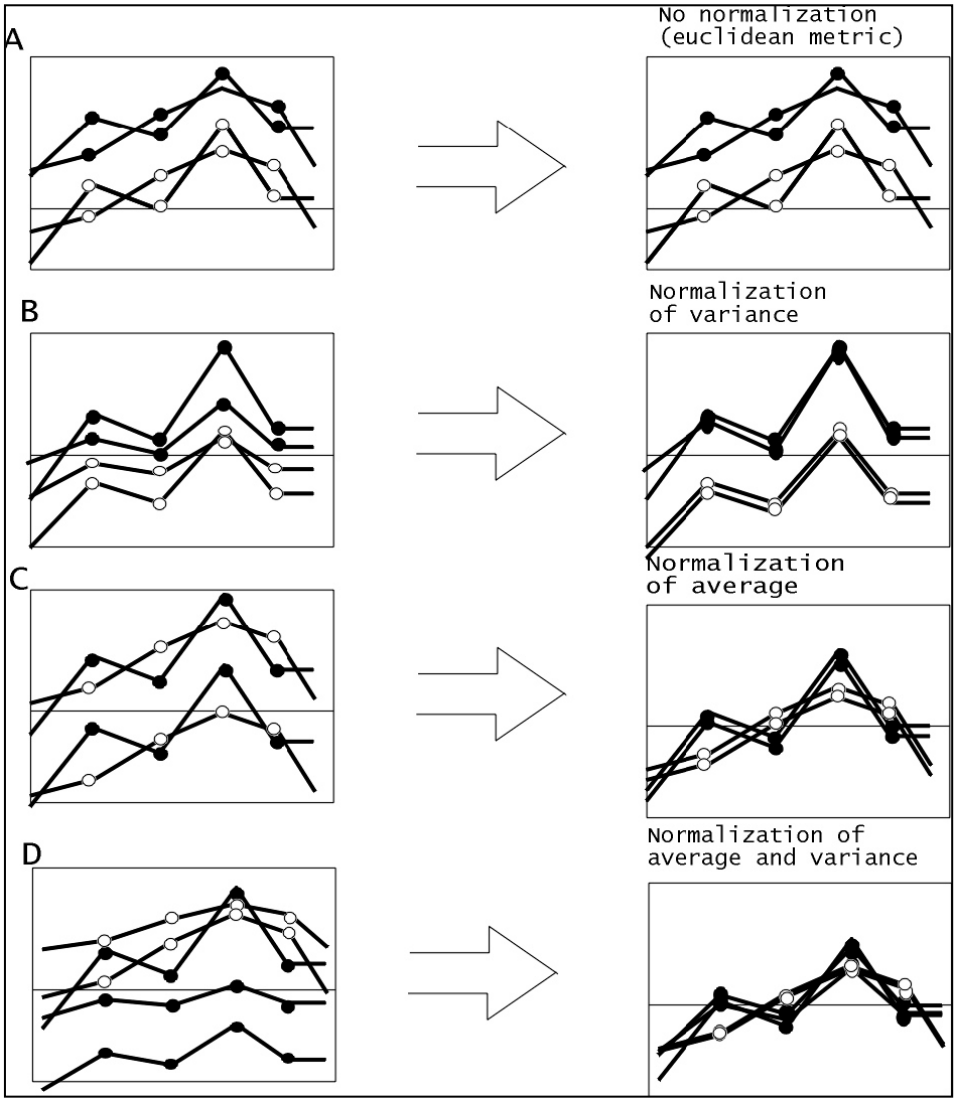
όπου y_{ij} οι τιμές γονιδιακής έκφρασης μετά τη ρύθμιση των τυπικών αποκλίσεων,

x_{ij} οι τιμές γονιδιακής έκφρασης πριν τη ρύθμιση των τυπικών αποκλίσεων

και σ_i η τυπική απόκλιση των τιμών του γονιδίου i η οποία δίνεται από τον τύπο

$$\sigma_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \overline{x_i})^2}$$

Η τυπική απόκλιση αποτελεί στην ουσία μία ένδειξη για την μεταβλητότητα των τιμών σε κάθε γονίδιο. Η ρύθμιση των τυπικών αποκλίσεων έχει ως αποτέλεσμα τη συμπίεση ή την επέκταση των διανυσμάτων των γονιδίων χωρίς όμως να αλλάζει το σχήμα τους. Όπως και προηγουμένως, μπορούμε να εφαρμόσουμε την ίδια πρακτική είτε στις γραμμές είτε στις στήλες ή και στις δύο.



Σχήμα 4.1: Ρύθμιση των μέσων όρων και των τυπικών αποκλίσεων. Στην εικόνα φαίνεται με ποιο τρόπο η ρύθμιση των μέσων όρων και των τυπικών αποκλίσεων βοηθάει στην σωστή ομαδοποίηση των γονιδίων. Οι γραμμές με τους μαύρους και τους άσπρους κύκλους αντιπροσωπεύουν γονίδια με παρόμοια επίπεδα έκφρασης πριν και μετά την επεξεργασία των δεδομένων. Στην πρώτη περίπτωση δεν γίνεται επεξεργασία των δεδομένων με αποτέλεσμα, όπως φαίνεται στο σχήμα, γονίδια που φαίνεται να έχουν παρόμοια έκφραση να έχουν στην εικόνα expression profiles με διαφορετικό σχήμα. Στη δεύτερη περίπτωση γίνεται κατάλληλη επεξεργασία των δεδομένων ούτως ώστε οι τυπικές αποκλίσεις να γίνουν ίσες με την μονάδα. Τα αποτελέσματα δείχνουν δύο ομάδες με σημαντικές ομοιότητες οι οποίες έχουν το ίδιο σχήμα αλλά διαφορετικό μέσο όρο. Στην τρίτη περίπτωση γίνεται κατάλληλη επεξεργασία των δεδομένων ούτως ώστε ο μέσος όρος των expression profiles να είναι μηδέν. Οι ομάδες που προκύπτουν αποτελούνται από expression profiles που βρίσκονται στο ίδιο επίπεδο. Στη τελευταία περίπτωση γίνεται κανονικοποίηση και στον μέσο όρο και στην τυπική απόκλιση. Όλα τα γονίδια παρουσιάζουν παραπλήσια επίπεδα έκφρασης.

4.4 Φιλτράρισμα

Στο στάδιο του φιλτραρίσματος απομακρύνονται τα γονίδια των οποίων η έκφραση είναι πολύ χαμηλή, καθώς και τα γονίδια εκείνα τα οποία παίρνουν περίπου τις ίδιες τιμές σε όλα τα δείγματα. Τα γονίδια που παρουσιάζουν τις παραπάνω συμπεριφορές όχι μόνο δεν μας παρέχουν χρήσιμες πληροφορίες αλλά αποτελούν και εμπόδιο στην σωστή ταξινόμηση των γονιδίων.

Μέχρι σήμερα έχουν προταθεί αρκετές διαφορετικές διαδικασίες φιλτραρίσματος. Μια από τις πιο συνηθισμένες είναι η απομάκρυνση των γονιδίων των οποίων η τιμή έκφρασης είναι μικρότερη από μία συγκεκριμένη τιμή την οποία την καθορίζουμε εμείς. Με αυτό τον τρόπο επιτυγχάνουμε την απομάκρυνση των γονιδίων εκείνων τα οποία παίρνουν τιμές κοντά στο μηδέν. Μια άλλη διαδικασία φιλτραρίσματος που αποτελεί παραλλαγή της προηγούμενης, είναι η απομάκρυνση των γονιδίων που παίρνουν μη μηδενικές τιμές σε ένα ποσοστό επί της εκατό των δειγμάτων το οποίο είναι χαμηλότερο απ’ αυτό που έχουμε καθορίσει εμείς σαν όριο.

Ο πιο συνηθισμένος τρόπος για να απομακρύνουμε τα “flat genes” δηλαδή τα γονίδια τα οποία έχουν περίπου τις ίδιες τιμές σε όλα τα δείγματα είναι να κατατάξουμε όλα τα γονίδια με βάση την τυπική τους απόκλιση. Στη συνέχεια μπορούμε να επιλέξουμε ένα από τα παρακάτω κριτήρια.

α) Να καθορίσουμε ένα συγκεκριμένο ποσοστό γονιδίων που θα κρατήσουμε. Αν έχουμε για παράδειγμα 1000 γονίδια και επιλέξουμε ποσοστό 20% θα κρατηθούν τα 200 πρώτα γονίδια της λίστας τα οποία είναι αυτά που μεταβάλλονται περισσότερο από δείγμα σε δείγμα.

β) Να καθορίσουμε ένα συγκεκριμένο όριο για την τυπική απόκλιση των γονιδίων. Όσα γονίδια έχουν μεγαλύτερη τυπική απόκλιση πάνω από το όριο που καθορίσαμε θα κρατηθούν.

Όπως μπορούμε να παρατηρήσουμε, σε όλες τις διαδικασίες φιλτραρίσματος τα ποσοτικά κριτήρια τα καθορίζουμε εμείς. Επειδή δεν υπάρχουν κάποιες προφανείς τιμές γι’ αυτά τα κριτήρια οι οποίες θα αποφέρουν αποδοτικό φιλτράρισμα, οι διαδικασίες φιλτραρίσματος θα πρέπει να χρησιμοποιούνται με προσοχή ούτως ώστε να μην αφαιρούνται βιολογικά χρήσιμα δεδομένα.

Κεφάλαιο 5 - Ανάλυση Των Δεδομένων

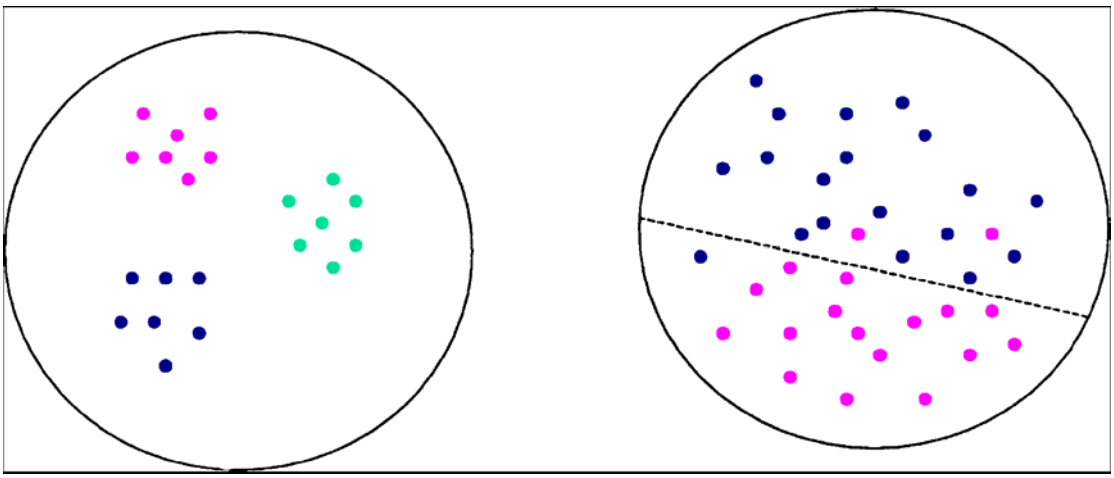
5.1 Εισαγωγή

Όπως έχουμε ήδη προαναφέρει, το μέγεθος των δεδομένων που προκύπτουν από την εφαρμογή της μεθόδου DNA Microarrays είναι αρκετά μεγάλο. Η ανάλυση των δεδομένων αυτού του μεγέθους καθίσταται δυνατή μόνο με τη βοήθεια υπολογιστικών και στατιστικών μεθόδων. Βασικός στόχος αυτής της ανάλυσης είναι η μελέτη των συσχετίσεων μεταξύ των γονιδίων και ο εντοπισμός των γονιδίων που συμπεριφέρονται με παρόμοιο τρόπο. Μέχρι σήμερα έχουν προταθεί αρκετές μέθοδοι για την ανάλυση των δεδομένων. Η μεγάλη πλειοψηφία των μεθόδων αυτών έχουν ως στόχο την κατηγοριοποίηση των δεδομένων γι' αυτό και ονομάζονται μέθοδοι κατηγοριοποίησης (classification methods). Σ' αυτήν την εργασία θα χρησιμοποιήσουμε τις μεθόδους Hierarchical Clustering, K-Means και Self-Organising Maps.

5.2 Μέθοδοι Κατηγοριοποίησης

Οι μέθοδοι κατηγοριοποίησης έχουν ως βασικό στόχο την τοποθέτηση των γονιδίων ή των δειγμάτων σε ομάδες με βάση κάποιο μέτρο ομοιότητας. Η κατηγοριοποίηση μπορεί να είναι είτε μη επιβλέπουσα (unsupervised classification) είτε επιβλέπουσα (supervised classification) [9-10]. Στην επιβλέπουσα κατηγοριοποίηση ένα σύνολο από προ-ομαδοποιημένα στοιχεία (γονίδια ή δείγματα) είναι διαθέσιμο, και αυτό που μας ζητείται είναι να εντάξουμε ένα νέο στοιχείο σε κάποια από τις ήδη υπάρχουσες ομάδες. Τα προ-ομαδοποιημένα στοιχεία χρησιμοποιούνται για να περιγράψουν τις διαφορετικές ομάδες – κλάσεις στις οποίες θα εντάξουμε τα νέα στοιχεία. Αντίθετα στην μη επιβλέπουσα κατηγοριοποίηση (unsupervised classification), η οποία συναντάται συνήθως με τον όρο ομαδοποίηση (clustering), σκοπός μας είναι να ομαδοποιήσουμε σε λογικές κλάσεις τα στοιχεία μας, χωρίς καμιά γνώση για προϋπάρχουσες ομάδες. Η κατηγοριοποίηση σ' αυτή τη

περίπτωση είναι απόλυτα οδηγούμενη από τα δεδομένα (data driven) και παράγεται μόνο από αυτά.



Σχήμα 5.1: Επιβλέπουσα και μη επιβλέπουσα κατηγοριοποίηση. Στην μη επιβλέπουσα κατηγοριοποίηση (αριστερά) έχουμε σημεία δεδομένων σε ένα χώρο n διαστάσεων (στο σχήμα μας n=2) και προσπαθούμε να εντάξουμε σε ομάδες τα σημεία που έχουν παρόμοια χαρακτηριστικά. Στο παράδειγμα μας διακρίνουμε τρεις διαφορετικές ομάδες οι οποίες περιέχουν σημεία δεδομένων που βρίσκονται πολύ κοντά μεταξύ τους. Στόχος ενός αλγορίθμου μη επιβλέπουσας κατηγοριοποίησης αποτελεί ο εντοπισμός αυτών των ομάδων. Στην επιβλέπουσα κατηγοριοποίηση (δεξιά) γνωρίζουμε από πριν τις ομάδες στις οποίες ανήκουν τα σημεία δεδομένων. Στο σχήμα μας οι δύο ομάδες εμφανίζονται με το μωβ και το μπλε χρώμα. Στόχος μας είναι να βρούμε ένα σύνολο κανόνων κατηγοριοποίησης που θα μας επιτρέψει να διαχωρίσουμε τα σημεία δεδομένων με όσον το δυνατόν μεγαλύτερη ακρίβεια. Στο σχήμα μας αυτό φαίνεται από τη μαύρη γραμμή.

5.3 Μέτρα Ομοιότητας

Όπως έχει ήδη προαναφερθεί, κάθε γραμμή στον πίνακα γονιδιακής έκφρασης αντιπροσωπεύει ένα γονίδιο και κάθε στήλη αντιπροσωπεύει ένα δείγμα. Αν θεωρήσουμε ότι ο αριθμός των γονιδίων είναι n και ο αριθμός των δειγμάτων είναι m, τότε κάθε γονίδιο αντιπροσωπεύεται από ένα διάνυσμα που αποτελείται από m στοιχεία και κάθε δείγμα αντιπροσωπεύεται από ένα διάνυσμα που αποτελείται από

n στοιχεία. Τα διανύσματα αυτά μπορούν να θεωρηθούν σαν σημεία σε ένα χώρο πολλών διαστάσεων (m διαστάσεων αν πρόκειται για διανύσματα γονιδίων ή n διαστάσεων αν πρόκειται για διανύσματα δειγμάτων). Για να μπορέσουμε να ταξινομήσουμε τα δείγματα ή τα γονίδια σε κατηγορίες θα πρέπει να χρησιμοποιήσουμε μεθόδους οι οποίες μετρούν την απόσταση δυο σημείων σε ένα χώρο πολλών διαστάσεων. Το μέτρο ομοιότητας λοιπόν μεταξύ των στοιχείων καθορίζεται συνήθως από μια συνάρτηση απόστασης. Αυτό που μετράει στην ουσία μια συνάρτηση απόστασης δεν είναι η ομοιότητα αλλά η ανομοιότητα (απόσταση) των στοιχείων. Υπάρχουν όμως και κάποιες συναρτήσεις οι οποίες ποσοτικοποιούν την ομοιότητα [11].

Μερικά από τα πιο διαδεδομένα μέτρα ομοιότητας είναι:

1. Euclidian Distance

Η πιο γνωστή συνάρτηση απόστασης που χρησιμοποιείται είναι η Ευκλείδεια απόσταση η οποία ορίζεται ως εξής:

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

Η ευκλείδεια απόσταση χρησιμοποιείται ευρέως σε περιπτώσεις λίγων διαστάσεων και έχει καλά αποτελέσματα όταν τα δεδομένα κατηγοριοποιούνται σε συμπαγείς και αρκετά απομονωμένες ομάδες. Το σημαντικότερο μειονέκτημα που παρουσιάζει η ευκλείδια απόσταση είναι ότι στις περιπτώσεις πολλών διαστάσεων το χαρακτηριστικό το οποίο παρουσιάζει την μεγαλύτερη διαφοροποίηση από τα άλλα κυριαρχεί και αποπροσανατολίζει το τελικό αποτέλεσμα.

2. Manhattan Distance:

$$D(x,y) = \sum_{i=1}^{n} | x_i - y_i |$$

Η απόσταση Manhattan είναι το άθροισμα των απόλυτων διαφορών των τιμών (x_i, y_i) των διανυσμάτων και αποτελεί στην ουσία μια παραλλαγή της ευκλείδιας απόστασης διατηρώντας τα ίδια πλεονεκτήματα και μειονεκτήματα.

3. Minkowski Distance:

Η απόσταση Minkowski, η οποία αποτελεί μια γενίκευση της Ευκλείδιας απόστασης και της απόστασης Manhattan, δίνεται από τον τύπο:

$$\mathrm{D(x,y)} = \left\{ \sum_{i=1}^{n} (x_i - y_i)^m \right\}^{1/m}$$

όπου m ο παράγοντας Minkowski.

Παρατηρούμε ότι για m = 1 η απόσταση Minkowski ταυτίζεται με την απόσταση Manhattan ενώ για m = 2 ταυτίζεται με την ευκλείδια απόσταση.

4. Chebyshev Distance:

$$\mathrm{D(x,y)} = \max_{i=1}^{n} | x_i - y_i |$$

Η απόσταση Chebyshev καθορίζεται από τα στοιχεία των διανυσμάτων που έχουν την μεγαλύτερη απόσταση. Ένα από τα πλεονεκτήματα αυτής της συνάρτησης είναι οι μικροί χρόνοι των υπολογισμών.

5. Pearson Correlation Coefficient:

Ο γραμμικός συντελεστής συσχέτισης ή συντελεστής συσχέτισης του Pearson, αποτελεί το πιο διαδεδομένο μέτρο συσχέτισης μεταξύ δύο διανυσμάτων. Ο συντελεστής αυτός μας επιτρέπει να ομαδοποιήσουμε τα διανύσματα χωρίς να λαμβάνουμε υπόψιν το συνολικό επίπεδο των τιμών τους. Αν για παράδειγμα δύο γονίδια έχουν διαφορετικά επίπεδα έκφρασης (διαφορετικές τιμές) αλλά έχουν «παράλληλους» τρόπους έκφρασης (όταν αυξάνεται ή μειώνεται το διάνυσμα του ενός αυξάνεται ή μειώνεται ανάλογα και το διάνυσμα του άλλου) τότε ο βαθμός ομοιότητας που θα δείξει ο συντελεστής θα είναι μεγάλος.

Έστω ότι x και y είναι δύο διανύσματα n στοιχείων για τα οποία θέλουμε να υπολογίσουμε το βαθμό συσχέτισης. Αν (x_i, y_i) είναι τα ζευγάρια τιμών των διανυσμάτων τότε ο συντελεστής συσχέτισης του Pearson δίνεται από τη σχέση:

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

όπου $\bar{x}$ η μέση τιμή του διανύσματος x και $\bar{y}$ η μέση τιμή του διανύσματος y

Η τιμή του συντελεστή κυμαίνεται από -1 μέχρι 1. Πιο συγκεκριμένα, ο συντελεστής παίρνει τη τιμή ένα όταν τα διανύσματα x και y είναι πανομοιότυπα, τη τιμή μηδέν όταν δεν έχουν καμία ομοιότητα και τη τιμή -1 όταν είναι ακριβώς αντίθετα.

6. Uncentered Pearson Correlation Coefficient:

$$r_{uc} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

Ο παραπάνω τύπος αποτελεί μια παραλλαγή του συντελεστή συσχέτισης του Pearson. Η διαφορά είναι ότι σ' αυτόν τον συντελεστή παίζει ρόλο και το συνολικό επίπεδο των τιμών των διανυσμάτων. Αν για παράδειγμα τα διανύσματα δύο γονιδίων έχουν πανομοιότυπους τρόπους έκφρασης αλλά διαφορετικές τιμές ο Uncentered Pearson Correlation Coefficient θα μας δώσει τιμή <1 ενώ ο Pearson Correlation Coefficient θα μας έδινε την τιμή 1. Αν λοιπόν θεωρούμε ότι το μέγεθος των τιμών είναι σημαντικό, είναι προτιμότερο να χρησιμοποιήσουμε αυτόν τον συντελεστή.

7. Squared Pearson Correlation Coefficient

$$r_{sq} = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \right)^2$$

Ο συντελεστής Squared Pearson Correlation Coefficient υπολογίζει το τετράγωνο του συντελεστή συσχέτισης του Pearson γεγονός που έχει ως αποτέλεσμα την μετατροπή των αρνητικών τιμών σε θετικές. Αν για παράδειγμα υποθέσουμε ότι έχουμε δυο γονίδια x,y με εντελώς αντίθετα επίπεδα έκφρασης (όταν αυξάνεται η έκφραση του ενός μειώνεται ανάλογα η έκφραση του άλλου) τότε ο Pearson Correlation Coefficient θα μας έδινε τιμές που πλησιάζουν το -1 , ενώ ο Squared Pearson Correlation Coefficient θα μας έδινε τιμές που πλησιάζουν το 1. Αν εφαρμόσουμε κάποιον αλγόριθμο ομαδοποίησης και χρησιμοποιήσουμε ως μέτρο ομοιότητας τον Pearson Correlation Coefficient, τα γονίδια x και y θα τοποθετηθούν σε ομάδες που βρίσκονται πολύ μακριά μεταξύ τους. Όμως, τα γονίδια x,y από βιολογική σκοπιά μπορεί να έχουν μια στενή σχέση μεταξύ τους παρόλο που τα διανύσματά τους είναι αντίθετα. Το πρόβλημα αυτό μπορούμε να το λύσουμε χρησιμοποιώντας τον Squared Pearson Correlation Coefficient. Αυτό θα είχε ως αποτέλεσμα να τοποθετηθούν τα γονίδια πολύ κοντά το ένα στο άλλο.

8. Averaged Dot Product

Το πιο απλό μέτρο συσχέτισης μεταξύ δύο διανυσμάτων είναι το εσωτερικό τους γινόμενο. Το εσωτερικό γινόμενο δύο διανυσμάτων x και y καθορίζεται από το άθροισμα των γινομένων των στοιχείων τους. Αν πάρουμε την μέση τιμή του εσωτερικού γινομένου, τότε η τιμή που προκύπτει είναι ανεξάρτητη από τον αριθμό των στοιχείων των διανυσμάτων. Ο μέσος όρος του εσωτερικού γινομένου δίνεται από τον τύπο:

$$d = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

9. Cosine Correlation Coefficient

Στις περιπτώσεις στις οποίες μας ενδιαφέρει η συμπεριφορά των διανυσμάτων (η αυξομείωση τους) και όχι τα συγκεκριμένα μεγέθη των τιμών τους, μπορούμε τροποποιήσουμε τον τύπο του εσωτερικού γινομένου ως εξής:

$$\cos\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Οι τιμές του συντελεστή κυμαίνονται από -1 μέχρι 1. Οι τιμές κοντά στο 1 δείχνουν μια έντονη θετική συσχέτιση μεταξύ των διανυσμάτων ενώ οι τιμές κοντά στο -1 δείχνουν μια έντονη αρνητική συσχέτιση (αντίθετα διανύσματα).

Όλα τα παραπάνω μέτρα ομοιότητας βασίζονται στην υπόθεση ότι η σχέση μεταξύ δύο διανυσμάτων x και y είναι γραμμική. Είναι πιθανόν όμως, μία μη γραμμική σχέση μεταξύ των διανυσμάτων να είναι πιο ακριβής σε σχέση με μια γραμμική σχέση. Ο υπολογισμός μη γραμμικών σχέσεων σε μεγάλα μεγέθη δεδομένων δεν είναι κάτι συνηθισμένο και θα απαιτούσε πολλά χρόνια έρευνας στα μαθηματικά που χρησιμοποιούνται για την ανάλυση των δεδομένων έκφρασης των γονιδίων.

Παρόλα αυτά, τα αποτελέσματα που παίρνουμε χρησιμοποιώντας μεθόδους κατηγοριοποίησης με βάση τα παραπάνω μέτρα ομοιότητας, καταδεικνύουν ότι αυτές οι διαδικασίες αποτελούν σχετικά ασφαλείς μεθόδους για τον εντοπισμό των συσχετίσεων σε μεγάλες βάσεις δεδομένων.

5.4 Μέθοδοι Μη Επιβλέπουσας Κατηγοριοποίησης

5.4.1 Γενικά

Οι μέθοδοι μη επιβλέπουσας κατηγοριοποίησης (clustering methods) χωρίζονται σε κάποιες κατηγορίες ανάλογα με τον τρόπο με τον οποίο εκτελείται η ομαδοποίηση [10]. Πιο συγκεκριμένα οι αλγόριθμοι μπορεί να είναι:

- **Ιεραρχικοί ή Διαμεριστικοί (Hierarchical or Partitioning).** Οι αλγόριθμοι ιεραρχικής ομαδοποίησης παράγουν μια σειρά από εμφωλευμένες κλάσεις οι οποίες προκύπτουν από διαδικασίες διαχωρισμού ή συγχώνευσης που πραγματοποιούνται με βάση κάποιο συγκεκριμένο μέτρο ομοιότητας. Οι αλγόριθμοι διαμέρισης στοχεύουν στο να διαχωρίσουν τα δεδομένα με τέτοιο τρόπο ούτως ώστε να βελτιστοποιείται κάποιο συγκεκριμένο κριτήριο (π.χ. η συνάρτηση τετραγωνικού λάθους).
- **Συγκεντρωτικοί ή Διαχωριστικοί (Agglomerative or Divisive).** Η διαφοροποίηση αυτών των δύο κατηγοριών σχετίζεται με τη λειτουργία και τις δομές του αλγορίθμου. Στην περίπτωση των συγκεντρωτικών αλγορίθμων, ο αλγόριθμος ξεκινά θεωρώντας κάθε στοιχείο σαν μία ξεχωριστή ομάδα και προχωρά συγχωνεύοντας στοιχεία και ομάδες μέχρις ότου ικανοποιηθεί μία συγκεκριμένη συνθήκη. Στους διαχωριστικούς αλγόριθμους αντίθετα, όλα τα στοιχεία θεωρούνται ότι ανήκουν σε μία ομάδα και ακολουθεί μια συνεχής διάσπαση της ομάδας αυτής σε υποομάδες μέχρις ότου ικανοποιηθεί η συνθήκη τερματισμού.
- **Σκληροί ή Ασαφείς (Hard or Fuzzy).** Ένας σκληρός αλγόριθμος τοποθετεί κάθε στοιχείο σε μία και μόνο ομάδα, σε αντίθεση με τους ασαφείς αλγορίθμους οι οποίοι δίνουν σε κάθε στοιχείο μιας ομάδας ένα βαθμό ο οποίος δείχνει κατά πόσο το στοιχείο αυτό ανήκει στην ομάδα.

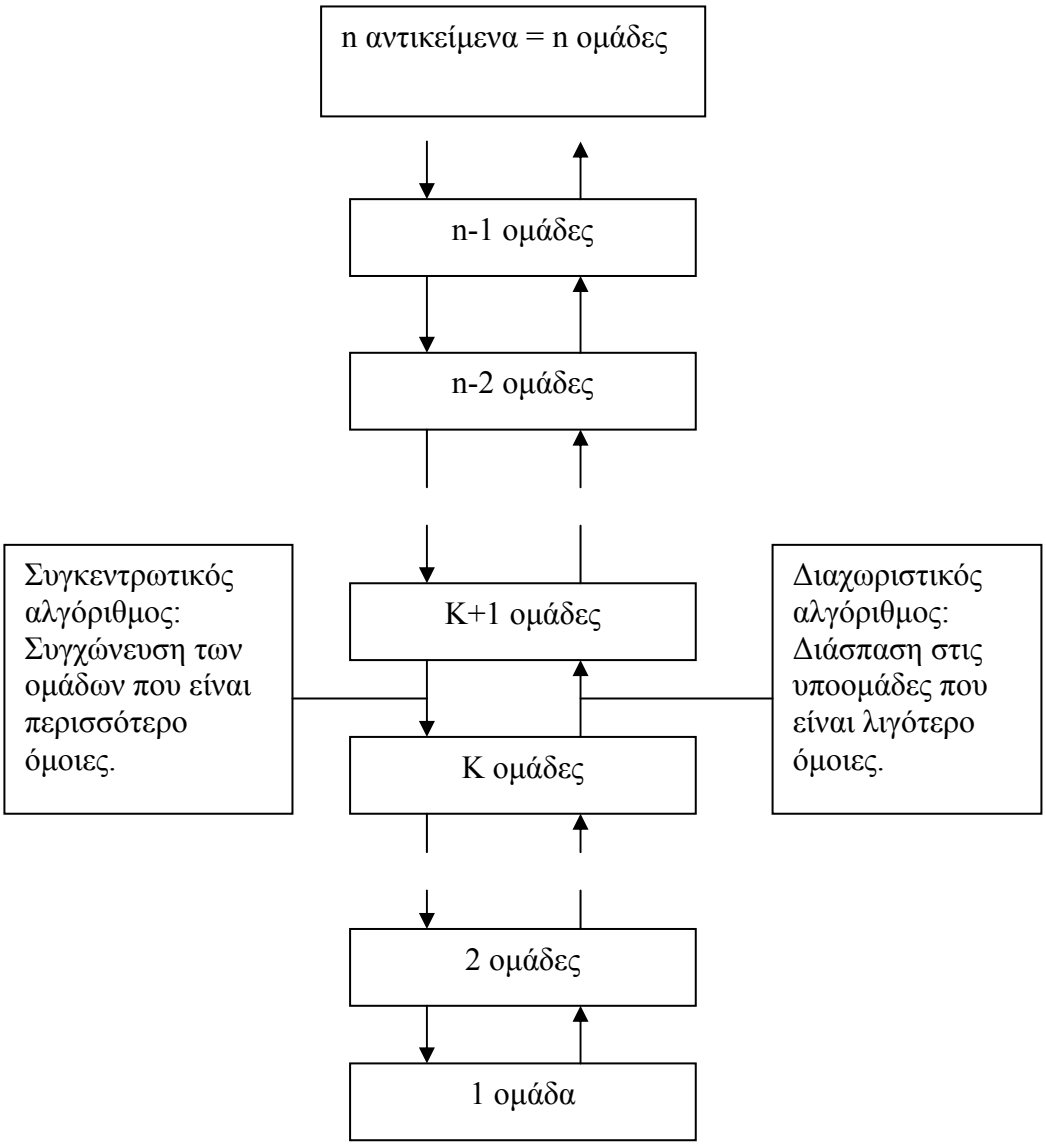
5.4.2 Αλγόριθμοι Ιεραρχικής Ομαδοποίησης (Hierarchical Clustering)

5.4.2.1 Γενικά

Οι αλγόριθμοι ιεραρχικής ομαδοποίησης ανήκουν όπως προαναφέραμε στις μεθόδους μη επιβλέπουσας κατηγοριοποίησης και αποτελούν την πιο διαδεδομένη μέθοδο αυτής της κατηγορίας. Οι αλγόριθμοι αυτοί παράγουν ένα δενδροδιάγραμμα το οποίο αναπαριστά τις συσχετίσεις μεταξύ των γονιδίων ή των δειγμάτων [8,12]. Πιο συγκεκριμένα, τα γονίδια ή τα δείγματα που παίρνουν παρόμοιες τιμές

αναπαριστώνται το ένα δίπλα στο άλλο και το μήκος των κλάδων αναπαριστά τον βαθμό ομοιότητάς τους.

Οι αλγόριθμοι ιεραρχικής ομαδοποίησης μπορεί να είναι είτε συγκεντρωτικοί (Agglomerative Hierarchical Clustering) είτε διαχωριστικοί (Divisive Hierarchical Clustering). Ο συγκεντρωτικός ιεραρχικός αλγόριθμος ξεκινά θεωρώντας κάθε αντικείμενο (γονίδιο ή δείγμα) σαν μια ξεχωριστή ομάδα και στη συνέχεια δημιουργεί όλο και μεγαλύτερες ομάδες ομαδοποιώντας τα δύο πιο όμοια αντικείμενα ή ομάδες αντικειμένων μέχρις ότου το σύνολο των δεδομένων να αποτελεί μία και μόνο ομάδα. Οι διαχωριστικοί ιεραρχικοί αλγόριθμοι ξεκινούν θεωρώντας το σύνολο των δεδομένων ως μία ομάδα η οποία στη συνέχεια διαχωρίζεται σε όλο και μικρότερες υποομάδες μέχρις ότου κάθε υποομάδα να αποτελείται από ένα και μόνο αντικείμενο. Οι συγκεντρωτικοί αλγόριθμοι απαιτούν λιγότερη υπολογιστική ισχύ σε σχέση με τους διαχωριστικούς γι' αυτό και χρησιμοποιούνται πολύ περισσότερο. Αυτό οφείλεται στο γεγονός ότι για να διασπάσουν οι διαχωριστικοί αλγόριθμοι μια ομάδα σε δύο μικρότερες, θα πρέπει να εξετάσουν όλους τους πιθανούς τρόπους διάσπασης αναλύοντας έτσι όλες τις δυνατές υποομάδες. Στους συγκεντρωτικούς αλγόριθμους αντίθετα, σε κάθε βήμα ενώνονται απλώς τα δύο πιο όμοια αντικείμενα κάτι που απαιτεί πολύ λιγότερη υπολογιστική ισχύ. Σ' αυτήν την εργασία θα χρησιμοποιήσουμε τον συγκεντρωτικό ιεραρχικό αλγόριθμο γι' αυτό και θα τον αναλύσουμε με περισσότερες λεπτομέρειες παρακάτω.



Σχήμα 5.2: Οι δύο κατηγορίες των ιεραρχικών αλγορίθμων.

5.4.2.2 Ο Συγκεντρωτικός Ιεραρχικός Αλγόριθμος

Η διαδικασία εκτέλεσης του συγκεντρωτικού ιεραρχικού αλγορίθμου περιλαμβάνει τα εξής βασικά στάδια [8,12,13]:

1) Για κάθε αντικείμενο (γονίδιο ή δείγμα) παίρνουμε ένα διάνυσμα από τις τιμές του πίνακα γονιδιακής έκφρασης. Αν το αντικείμενο είναι γονίδιο, το διάνυσμα αποτελείται από τις τιμές του γονιδίου για κάθε πείραμα ενώ αν το αντικείμενο είναι δείγμα, το διάνυσμα αποτελείται από τις τιμές των γονιδίων στο συγκεκριμένο δείγμα.

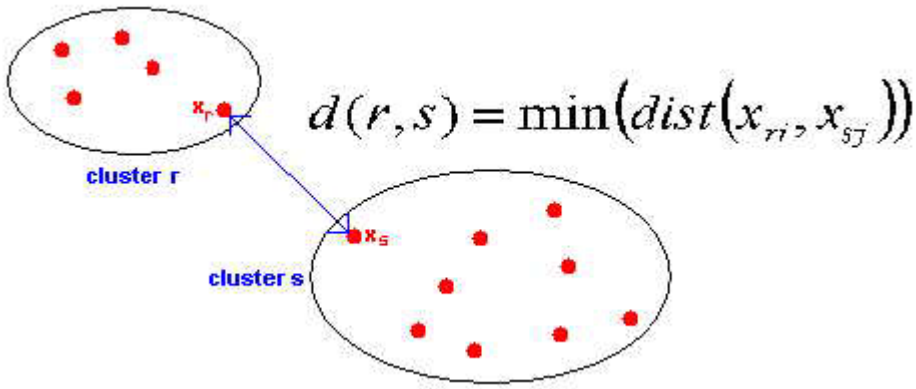
2) Υπολογίζονται οι αποστάσεις κάθε αντικειμένου σε σχέση με όλα τα άλλα αντικείμενα. Οι αποστάσεις τοποθετούνται σε ένα πίνακα.
3) Εντοπίζονται τα αντικείμενα που έχουν την μικρότερη απόσταση μεταξύ τους.
4) Τα αντικείμενα που εντοπίστηκαν τοποθετούνται στην ίδια ομάδα. Γίνεται επιστροφή στο βήμα 2 και υπολογίζονται ξανά οι αποστάσεις με την καινούρια ομάδα να θεωρείται πλέον σαν ένα αντικείμενο.
5) Τα βήματα 2-4 επαναλαμβάνονται μέχρις ότου όλα τα αντικείμενα να τοποθετηθούν σε μία μεγάλη ομάδα.

Μέθοδοι Διασύνδεσης των Ομάδων (Linkage Methods)

Όταν ο αλγόριθμος εκτελείται για πρώτη φορά, κάθε ομάδα αποτελείται από ένα μόνο αντικείμενο οπότε η απόσταση μεταξύ των ομάδων καθορίζεται από το μέτρο ομοιότητας που έχουμε επιλέξει. Όταν όμως τα πρώτα αντικείμενα τοποθετηθούν σε ομάδες, το μέτρο ομοιότητας από μόνο του δεν αρκεί. Θα πρέπει να καθορίσουμε κάποιο κριτήριο σύμφωνα με το οποίο ο αλγόριθμος θα υπολογίζει τις αποστάσεις μεταξύ των ομάδων που περιέχουν περισσότερα από ένα αντικείμενα. Το κριτήριο αυτό λέγεται μέθοδος διασύνδεσης (linkage method) [12,14,15]. Οι πιο γνωστές μέθοδοι διασύνδεσης είναι:

Single Linkage:

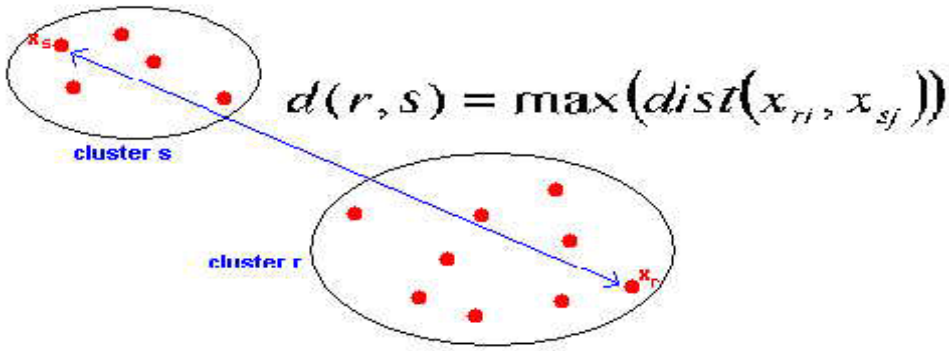
Η απόσταση μεταξύ δύο ομάδων καθορίζεται από την απόσταση μεταξύ των δύο πιο κοντινών αντικειμένων τους. Αν υπάρχουν πολλές μικρές αποστάσεις μεταξύ των αντικειμένων των δύο ομάδων τότε η συγκεκριμένη μέθοδος δίνει αρκετά καλά αποτελέσματα. Το σοβαρότερο μειονέκτημα αυτής της μεθόδου είναι ότι μία και μόνο τυχαία μικρή απόσταση μεταξύ δύο αντικειμένων των ομάδων είναι αρκετή για να ενωθούν ομάδες που κατά τ' άλλα είναι εντελώς διαφορετικές μεταξύ τους. Το πρόβλημα αυτό ονομάζεται «φαινόμενο chaining» και έχει σαν αποτέλεσμα να ενώνονται το ένα μετά το άλλο μοναδικά αντικείμενα με ομάδες αντικειμένων σχηματίζοντας «αλυσίδες» (σχήμα 5.6).



Σχήμα 5.3: Single Linkage. Η απόσταση των ομάδων r και s ισούται με την απόσταση των πιο κοντινών αντικειμένων x_r και x_s.

Complete Linkage:

Η απόσταση μεταξύ δύο ομάδων καθορίζεται από την απόσταση μεταξύ των δύο πιο μακρινών αντικειμένων τους. Η μέθοδος αυτή δίνει καλά αποτελέσματα και πολύ συμπαγείς ομάδες με την προϋπόθεση ότι στα δεδομένα μας δεν υπάρχει θόρυβος. Στις περιπτώσεις στις οποίες υποπτευόμαστε ότι τα δεδομένα μας περιέχουν σημαντικό θόρυβο, η μέθοδος αυτή δεν ενδείκνυται καθώς η παρουσία του θορύβου θα συμβάλλει καθοριστικά στη διαμόρφωση των αποτελεσμάτων.

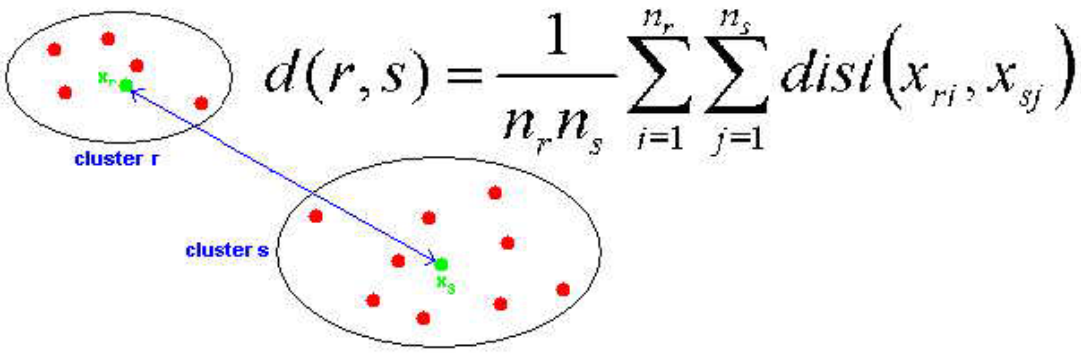


Σχήμα 5.4: Complete Linkage. Η απόσταση των ομάδων r και s ισούται με την απόσταση των πιο μακρινών αντικειμένων x_r και x_s.

Unweighted Pair-Group Average Linkage:

Η μέθοδος αυτή υπολογίζει την μέση απόσταση μεταξύ όλων των ζευγαριών των αντικειμένων των δύο ομάδων. Οι ομάδες που έχουν την μικρότερη μέση απόσταση

μεταξύ τους συγχωνεύονται. Το πλήθος των υπολογισμών που απαιτούνται για την εύρεση της μέσης απόστασης, καθιστά αυτή τη μέθοδο πιο δαπανηρή σε υπολογιστική ισχύ σε σχέση με τις προηγούμενες. Παρόλα αυτά, το γεγονός ότι η μέθοδος αυτή αποτελεί μια «ενδιάμεση περίπτωση» των δύο προηγουμένων, την απαλλάσει από τα «ακραία φαινόμενα» και την καθιστά πιο ασφαλή όσον αφορά το φαινόμενο της αλυσίδας και την ευαισθησία στο θόρυβο.



$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist\left(x_{ri}, x_{sj}\right)$$

Σχήμα 5.5: Unweighted Pair-Group Average Linkage. Η απόσταση των ομάδων r και s ισούται με την μέση απόσταση όλων των δυνατών ζευγαριών των αντικειμένων των ομάδων.

Weighted Pair-Group Average Linkage:

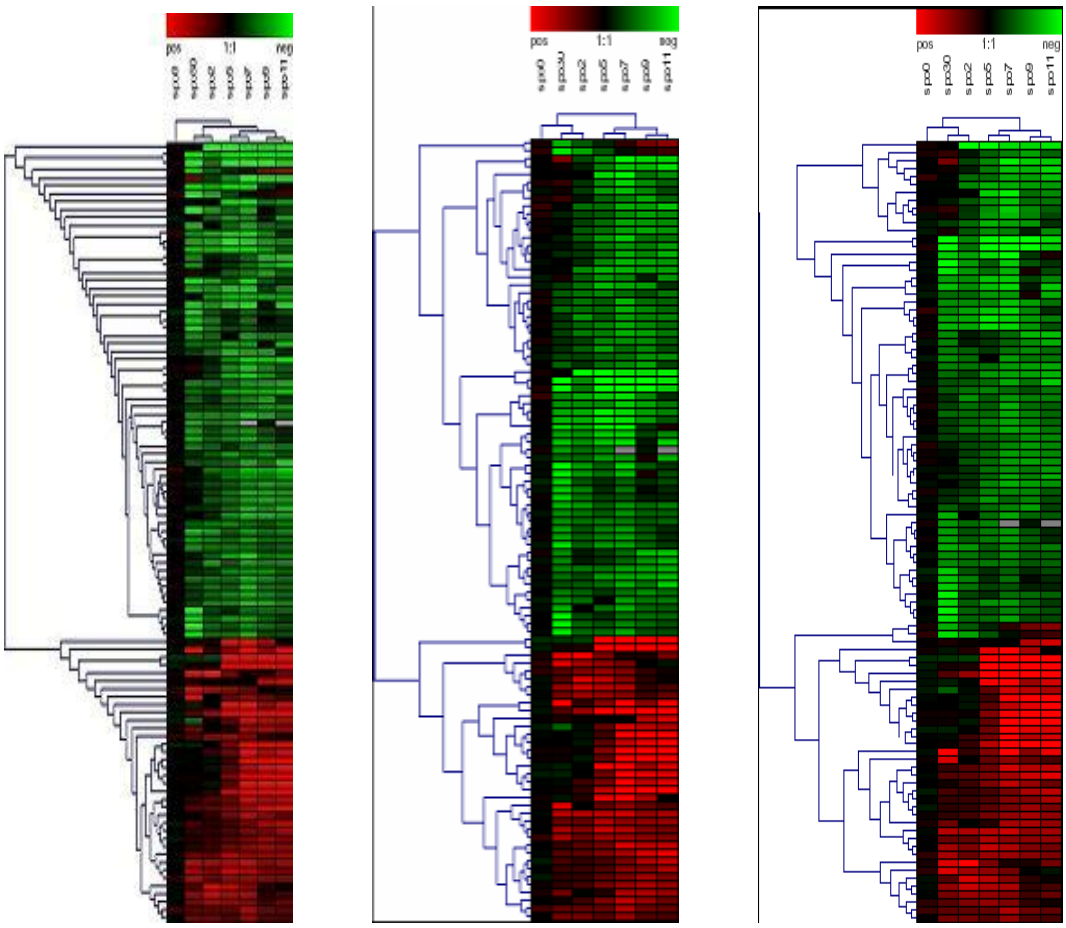
Η μέθοδος αυτή είναι ίδια με την προηγούμενη με τη διαφορά ότι τα μεγέθη των ομάδων (δηλαδή ο αριθμός των αντικειμένων που περιέχουν) χρησιμοποιούνται στους υπολογισμούς σαν βάρη. Η μέθοδος αυτή χρησιμοποιείται όταν αναμένουμε πως τα μεγέθη των ομάδων θα είναι αρκετά ανόμοια.

Unweighted Pair-Group Centroid Linkage

Η απόσταση μεταξύ δύο ομάδων καθορίζεται από την απόσταση μεταξύ των κεντροειδών τους. Το κεντροειδές (centroid) μιας ομάδας είναι ένα φανταστικό αντικείμενο που ανήκει στην ομάδα, του οποίου οι συντεταγμένες προκύπτουν από το μέσο όρο των συντεταγμένων όλων των αντικειμένων της ομάδας. Το σημαντικότερο μειονέκτημα αυτής της μεθόδου είναι ότι όταν μια πολύ μικρή ομάδα συγχωνεύεται με μια μεγαλύτερη τότε τα χαρακτηριστικά της σχεδόν εξαφανίζονται στη νέα ομάδα που προκύπτει.

Weighted Pair-Group Centroid Linkage

Αυτή η μέθοδος είναι ίδια με την προηγούμενη με τη διαφορά ότι στους υπολογισμούς εισάγονται βάρη ούτως ώστε να λαμβάνονται υπόψιν τα διαφορετικά μεγέθη μεταξύ των ομάδων.



(α) (β) (γ)

Σχήμα 5.6: Στην εικόνα φαίνονται παραδείγματα δενδροδιαγραμμάτων που προκύπτουν από τον συγκεντρωτικό ιεραρχικό αλγόριθμο. Στο σχήμα (α) έχει χρησιμοποιηθεί single linkage, στο (β) complete linkage και στο (γ) average linkage.

Περιγραφή του Αλγορίθμου

Όπως μπορούμε να διαπιστώσουμε από τα προαναφερόμενα, η υλοποίηση του δεύτερου βήματος του αλγορίθμου, ο υπολογισμός δηλαδή των αποστάσεων μεταξύ των ομάδων, προϋποθέτει τον προκαθορισμό:

α) Του μέτρου ομοιότητας που θα χρησιμοποιήσουμε (π.χ. Ευκλείδια απόσταση, Pearson correlation coefficient κτλ.)

β) Της μέθοδου διασύνδεσης μεταξύ των ομάδων (π.χ. Single Linkage, Complete Linkage κτλ.)

Οι αποστάσεις που υπολογίζονται σχηματίζουν ένα άνω διαγώνιο πίνακα. Σ' αυτόν τον πίνακα εντοπίζεται:

α) Η μεγαλύτερη τιμή (ο μεγαλύτερος βαθμός ομοιότητας μεταξύ των ομάδων) στην περίπτωση που επιλέξουμε κάποιο συντελεστή συσχέτισης (π.χ. Pearson correlation coefficient)

β) Η μικρότερη τιμή (η μικρότερη απόσταση μεταξύ των ομάδων) στην περίπτωση που επιλέξουμε κάποια συνάρτηση απόστασης (π.χ. Ευκλείδια απόσταση)

Σε κάθε περίπτωση λοιπόν εντοπίζονται οι ομάδες που έχουν τον μεγαλύτερο βαθμό ομοιότητας και στην συνέχεια συγχωνεύονται δημιουργώντας μια καινούρια ομάδα. Στη συνέχεια, ο πίνακας αποστάσεων υπολογίζεται από την αρχή και η διαδικασία συνεχίζεται μέχρις ότου απομείνουν δύο ομάδες. Σ' αυτό το σημείο δε χρειάζεται να γίνουν υπολογισμοί αφού είναι προφανές πως οι εναπομείναντες ομάδες θα ενωθούν μεταξύ τους και θα δημιουργήσουν την ρίζα του δενδροδιαγράμματος (super cluster).

Είναι σημαντικό να συνειδητοποιήσουμε ότι το τελικό δενδροδιάγραμμα που παράγεται εξαρτάται από την επιλογή του μέτρου ομοιότητας και της μεθόδου διασύνδεσης. Διαφορετικοί συνδυασμοί μεθόδων διασύνδεσης και μέτρων ομοιότητας έχουν σαν αποτέλεσμα διαφορετικά δενδροδιαγράμματα. Άρα, όταν μελετάμε μια συγκεκριμένη ομάδα σε ένα δενδροδιάγραμμα θα πρέπει να γνωρίζουμε ότι μπορεί απλώς να είναι αποτέλεσμα των μεθόδων που έχουμε επιλέξει για τον υπολογισμό των αποστάσεων και να μην έχει κάποια σημαντική βιολογική αξία. Παρόλο που κάποιες από τις μεθόδους έχουν περισσότερα πλεονεκτήματα από κάποιες άλλες, δεν υπάρχουν γενικοί κανόνες επιλογής των μεθόδων. Η επιλογή μας εξαρτάται από τα χαρακτηριστικά του συνόλου των δεδομένων μας.

Πλεονεκτήματα και Μειονεκτήματα

Η ιεραρχική ομαδοποίηση αποτελεί αυτή τη στιγμή τη πιο διαδεδομένη μέθοδο ομαδοποίησης για την ανάλυση δεδομένων γονιδιακής έκφρασης. Το πλεονέκτημα αυτής της μεθόδου είναι ότι με εξαίρεση την επιλογή του μέτρου ομοιότητας και της μεθόδου διασύνδεσης δεν χρειάζεται να προκαθορίσουμε άλλες παραμέτρους και κυρίως δεν χρειάζεται να προκαθορίσουμε τον αριθμό των ομάδων. Ένα άλλο εξίσου σημαντικό πλεονέκτημα είναι ότι το μήκος των κλάδων στο δενδροδιάγραμμα είναι ανάλογο της ομοιότητας μεταξύ των ομάδων. Αυτός ο τρόπος με τον οποίο αναπαριστώνται οι συσχετίσεις σε ένα δενδροδιάγραμμα είναι πολύ χρήσιμος γιατί έχει την ικανότητα να αποκαλύπτει α) διαφορετικούς βαθμούς ομοιότητας μεταξύ των αντικειμένων και β) μακρινές σχέσεις μεταξύ ομάδων αντικειμένων.

Το σημαντικότερο μειονέκτημα του ιεραρχικού αλγορίθμου είναι ότι απαιτεί πολύ μεγάλη υπολογιστική ισχύ. Αυτό οφείλεται στο γεγονός ότι για να κατασκευαστεί ένας πίνακας ομοιοτήτων σε ένα μεγάλο σύνολο δεδομένων απαιτείται πολύ μεγάλη ποσότητα διαθέσιμης μνήμης. Ένα άλλο μειονέκτημα είναι ότι ο αλγόριθμος δεν δίνει από μόνος του συγκεκριμένες ομάδες στο τέλος των υπολογισμών αλλά ανατοποθετεί απλώς τα αντικείμενα βάζοντας τα όμοια το ένα κοντά στο άλλο. Έτσι, ο χρήστης καλείται ο ίδιος να θεωρήσει ως τελικές ομάδες κάποια συγκεκριμένα κομμάτια του δενδροδιαγράμματος. Τέλος, ο αλγόριθμος δεν μπορεί να διορθώσει λανθασμένες αποφάσεις που έχουν γίνει προηγουμένως με αποτέλεσμα πολλές φορές να κλειδώνει σε μια συγκεκριμένη μορφή η οποία μπορεί να είναι αποτέλεσμα τοπικών ομοιοτήτων.

5.4.3 Αλγόριθμοι Διαμέρισης

5.4.3.1 Γενικά

Οι διαμεριστικοί αλγόριθμοι δίνουν ως αποτέλεσμα μια διαμέριση του χώρου των δεδομένων σε ένα συγκεκριμένο αριθμό από ομάδες ο οποίος προκαθορίζεται.από την αρχή. Οι αλγόριθμοι αυτοί υπερτερούν σε σχέση με τους ιεραρχικούς στις περιπτώσεις στις οποίες τα δεδομένα είναι πάρα πολλά και η δημιουργία δενδροδιαγραμμάτων είναι αρκετά δύσκολη. Στόχος αυτών των αλγόριθμων είναι να

βελτιστοποιήσουν κάποιο συγκεκριμένο κριτήριο (π.χ. να ελαχιστοποιήσουν την συνάρτηση τετραγωνικού λάθους) γι' αυτό και μπορούν να θεωρηθούν ως προβλήματα βελτιστοποίησης. Οι πιο διαδεδομένοι αλγόριθμοι διαμέρισης είναι ο αλγόριθμος K-Means και ο αλγόριθμος SOM (Self-Organising Maps).

5.4.3.2 K - Means

Η διαδικασία εκτέλεσης του αλγορίθμου K-Means περιλαμβάνει τα εξής βασικά στάδια [8,12,13]:

1) Καθορίζεται ο αριθμός των ομάδων που θα δημιουργηθούν (κ).

2) Προσδιορίζεται τυχαία το κέντρο κάθε ομάδας. Το κέντρο θα πρέπει να είναι ένα σημείο δεδομένων το οποίο αναπαριστά ένα διάνυσμα που έχει τον ίδιο αριθμό στοιχείων με τα διανύσματα των αντικειμένων που θέλουμε να ομαδοποιήσουμε (same dimensionality).

3) Υπολογίζεται η απόσταση κάθε αντικειμένου από όλα τα κέντρα των ομάδων και εντοπίζεται η ομάδα που έχει το κέντρο της πιο κοντά στο αντικείμενο που εξετάζεται. Το αντικείμενο εντάσσεται σ' αυτήν την ομάδα.

4) Προσδιορίζεται ξανά το κέντρο κάθε ομάδας. Αυτή τη φορά το κέντρο προσδιορίζεται από το κεντροειδές της ομάδας και όχι τυχαία.

5) Τα αντικείμενα ομαδοποιούνται από την αρχή μετρώντας και πάλι τις αποστάσεις από τα καινούρια πλέον κέντρα των ομάδων. Σ' αυτό το βήμα είναι πιθανόν κάποια αντικείμενα να αλλάξουν ομάδα.

6) Υπολογίζεται η συνάρτηση τετραγωνικού λάθους. Αν η συνάρτηση παραμένει σταθερή ή δεν παρουσιάζει σημαντικές αλλαγές για ένα συγκεκριμένο αριθμό επαναλήψεων τότε ο αλγόριθμος σταματάει. Αν δεν παραμένει σταθερή, τότε εκτελούνται ξανά τα βήματα 3-6.

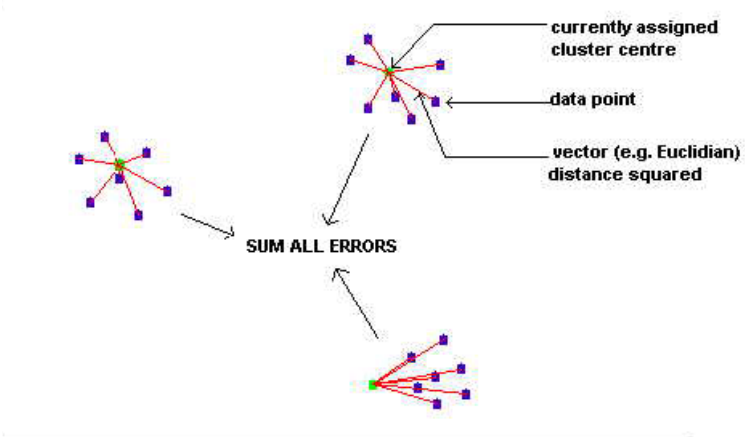
Ο προσδιορισμός του κεντροειδούς κάθε ομάδας (βήμα 4) γίνεται υπολογίζοντας το γεωμετρικό μέσο όρο όλων των αντικειμένων της ομάδας. Αν $Datapo\,\mathrm{int}_i^c$ είναι το διάνυσμα του αντικειμένου i της ομάδας c και n_c είναι ο αριθμός των αντικειμένων της ομάδας c τότε το κεντροειδές της ομάδας c προκύπτει από τον τύπο:

$$\overline{x_c} = \frac{\sum_{i=1}^{n_c} Datapo\,\mathrm{int}_i^{(c)}}{n_c}$$

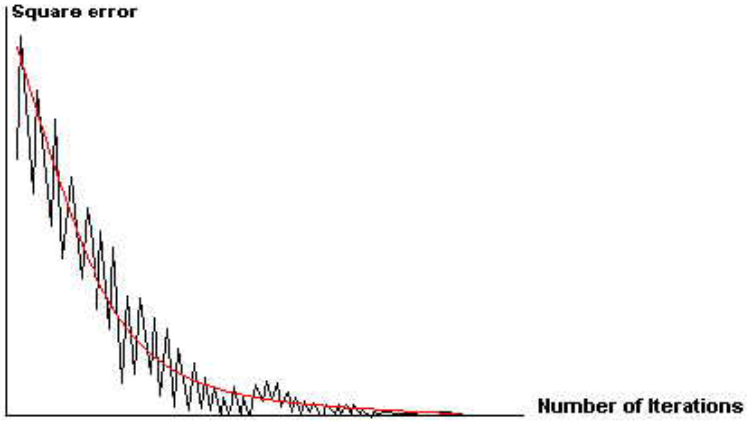
Ο υπολογισμός της συνάρτησης τετραγωνικού λάθους για το αποτέλεσμα μιας ομαδοποίησης δίνεται από τον τύπο

$$E = \sum_{c=1}^{\kappa} \sum_{i=1}^{n_c} d_i^{\,2}$$

όπου κ ο συνολικός αριθμός των ομάδων, n_c ο αριθμός των αντικειμένων της ομάδας c και d_i η απόσταση ανάμεσα στο διάνυσμα του αντικειμένου i και του κεντροειδούς της ομάδας στην οποία ανήκει.



Σχήμα 5.7: Ο υπολογισμός της συνάρτησης τετραγωνικού λάθους. Η τιμή της συνάρτησης υπολογίζεται από το άθροισμα των τετραγώνων των αποστάσεων (κόκκινες γραμμές) μεταξύ των κεντροειδών (πράσινες κουκκίδες) των ομάδων και των αντικειμένων των ομάδων (μπλε κουκκίδες).



Σχήμα 5.8: Μείωση της τιμής της συνάρτησης τετραγωνικού λάθους. Η τιμή της συνάρτησης μειώνεται συνεχώς μέχρι να ελαχιστοποιηθεί μετά από κάποιο αριθμό επαναλήψεων.

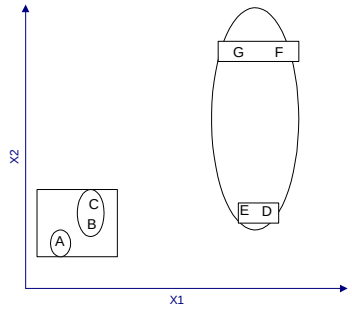
Πλεονεκτήματα - Μειονεκτήματα

Το σημαντικότερο πλεονέκτημα του αλγορίθμου K-Means, το οποίο τον καθιστά ιδιαίτερα δημοφιλή, είναι η απλότητά του και κατά συνέπεια η ευκολία στην υλοποίησή του. Ένα άλλο πλεονέκτημα αυτού του αλγορίθμου είναι οι μικρές απαιτήσεις σε υπολογιστική ισχύ. Αυτόν τον καθιστά ιδιαίτερα αποτελεσματικό σε μεγάλα μεγέθη δεδομένων (πάνω από 10000 γονίδια) στα οποία οι ιεραρχικοί αλγόριθμοι αντιμετωπίζουν προβλήματα.

Το πιο σημαντικό μειονέκτημα του αλγορίθμου K-Mean είναι ο προκαθορισμός του αριθμού των ομάδων. Παρόλα αυτά, αυτή η παράμετρος μαζί με κάποιο κριτήριο που χρησιμοποιούμε για να αποφύγουμε την επ' αόριστο επανάληψη του αλγορίθμου είναι οι μοναδικές παράμετροι που χρειάζονται να προκαθοριστούν. Το κριτήριο που χρησιμοποιείται συνήθως είναι είτε η συνάρτηση του τετραγωνικού λάθους είτε ο αριθμός των μέγιστων επαναλήψεων.

Το πρόβλημα με τον προκαθορισμό του αριθμού των ομάδων μπορεί να λυθεί χρησιμοποιώντας μια παραλλαγή του αλγορίθμου K-Means που ονομάζεται ISODATA. Σ' αυτόν τον αλγόριθμο περιλαμβάνεται μια διαδικασία εύρεσης του αριθμού των ομάδων η οποία υπολογίζει τη συνάρτηση λάθους που προκύπτει για κάθε αριθμό και στη συνέχεια προσδιορίζει τον αριθμό των ομάδων για τον οποίο η τιμή της συνάρτησης βελτιστοποιείται.

Ένα άλλο μειονέκτημα του αλγορίθμου είναι ότι το τελικό αποτέλεσμα εξαρτάται από τον αρχική τοποθέτηση των κέντρων των ομάδων. Το πρόβλημα αυτό φαίνεται στην παρακάτω εικόνα:



Σχήμα 5.9: Η ευαισθησία του αλγόριθμου k-means στην αρχική επιλογή των κέντρων των ομάδων. Στην εικόνα φαίνονται 7 αντικείμενα τα οποία ομαδοποιούνται με δύο διαφορετικούς τρόπους. Η πρώτη ομαδοποίηση, η οποία παριστάνεται με τις ελλείψεις, προκύπτει από την αρχική επιλογή τριών ομάδων με κέντρα τα αντικείμενα A,B,C. Η δεύτερη ομαδοποίηση φαίνεται από τα παραλληλόγραμμα και προκύπτει από την αρχική επιλογή τριών ομάδων με κέντρα τα αντικείμενα A,D,F. Από το σχήμα γίνεται φανερό ότι στην πρώτη περίπτωση η τιμή της συνάρτησης τεραγωνικού λάθους είναι μεγαλύτερη και η ομαδοποίηση είναι μη-αποτελεσματική.

Ο Αλγόριθμος Fuzzy C-Means

Ο αλγόριθμος Fuzzy C-Means Algorithm ανήκει στην κατηγορία των fuzzy αλγορίθμων και αποτελεί μια παραλλαγή του αλγορίθμου K-Means. Το πλεονέκτημα σ' αυτό τον αλγόριθμο είναι ότι ένα αντικείμενο μπορεί να ανήκει σε παραπάνω από μία ομάδες και να έχει ένα βαθμό συσχέτισης με κάθε ομάδα. Από βιολογική σκοπιά αυτή η δυνατότητα έχει σημαντική πρακτική αξία αφού ένα γονίδιο μπορεί να σχετίζεται και να επηρεάζει με διάφορους τρόπους αρκετά από τα υπόλοιπα γονίδια.

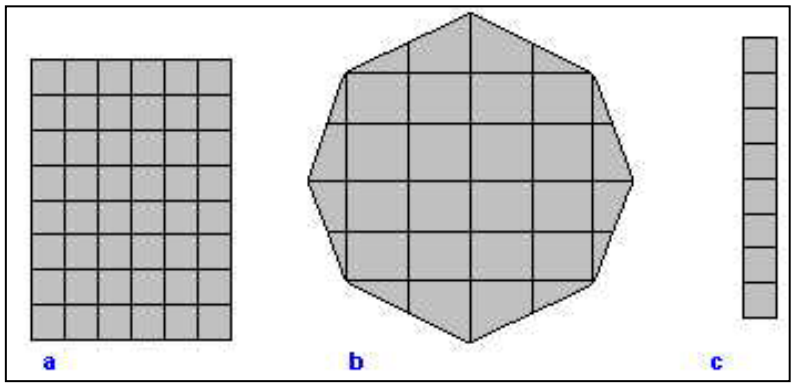
5.4.3.3 Self - Organising Maps (SOM)

Ένας από τους πιο δημοφιλείς διαμεριστικούς αλγόριθμους είναι ο αλγόριθμος Self-Organising Maps [8,12,13]. Οι βασικοί στόχοι αυτού του αλγορίθμου είναι:

α) Η αναπαράσταση δεδομένων πολλών διαστάσεων σε μια μορφή λιγότερων διαστάσεων χωρίς να χάνεται η «ουσία» των δεδομένων.

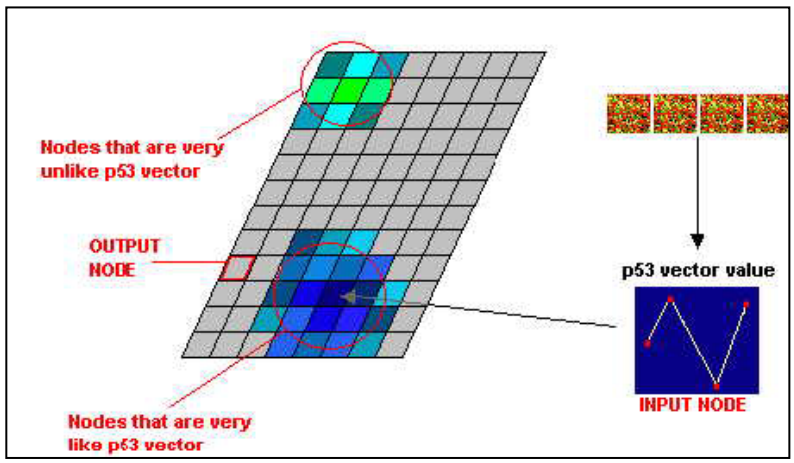
β) Η οργάνωση των δεδομένων με τέτοιο τρόπο ούτως ώστε τα όμοια αντικείμενα να βρίσκονται κοντά το ένα με το άλλο.

Η υλοποίηση ενός τέτοιου αλγορίθμου έχει ως αποτέλεσμα την παραγωγή «χαρτών» (πλεγματοειδείς πίνακες) οι οποίοι μπορεί να είναι είτε μίας είτε δύο διαστάσεων. Το σχήμα των χαρτών, το οποίο είναι συνήθως εξάγωνο ή παραλληλόγραμμο, καθορίζεται από την προτίμηση του χρήστη.



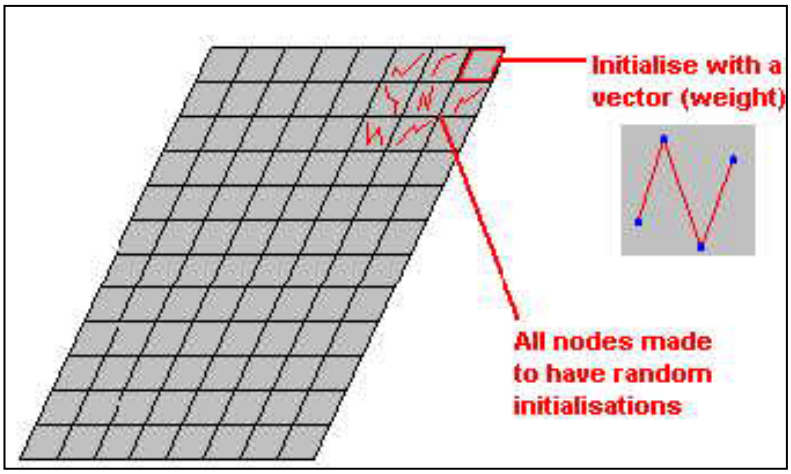
Σχήμα 5.10: Οι χάρτες μπορεί να είναι είτε δύο διαστάσεων (a και b) είτε μίας διάστασης (c).

Κάθε ένα από τα τετραγωνάκια στους χάρτες αναπαριστά ένα output node. Το output node αποτελεί την εξωτερική αναπαράσταση της τιμής ενός διανύσματος. Η τιμή αυτή ονομάζεται weight vector. Τα δεδομένα τα οποία εισέρχονται στον αλγόριθμο χαρακτηρίζονται ως input nodes και αποτελούν κι αυτά εξωτερικές αναπαραστάσεις διανυσμάτων και πιο συγκεκριμένα των input vectors, δηλαδή των διανυσμάτων που προκύπτουν από την εφαρμογή της μεθόδου DNA Microarrays. Οι έννοιες input node, output node, weight vector και input vector γίνονται πιο ξεκάθαρες στο παρακάτω σχήμα:



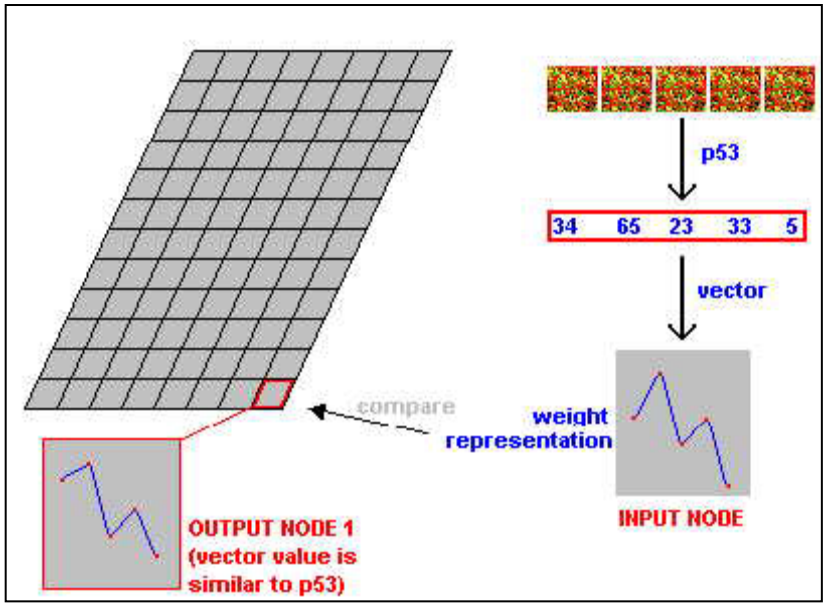
Σχήμα 5.11: Στο σχήμα βλέπουμε το διάνυσμα ενός γονιδίου το οποίο εισέρχεται στο χάρτη με μορφή input node. Το διάνυσμα τοποθετείται σε συγκεκριμένη θέση στο χάρτη έτσι ώστε τα γειτονικά output nodes να έχουν τιμές (weight vectors) παραπλήσιες με την τιμή του διανύσματος του γονιδίου (input vector).

Το πρώτο βήμα που απαιτείται για την υλοποίηση ενός αλγορίθμου Self-Organising Map είναι η αρχικοποίηση. Με τον όρο αρχικοποίηση εννοούμε τον καθορισμό της τοπολογίας του χάρτη, δηλαδή το μέγεθος του, το σχήμα του και το αρχικό του περιεχόμενο. Το αρχικό περιεχόμενο καθορίζεται δίνοντας τιμές στα weight vectors των output nodes. Οι τιμές αυτές μπορεί να είναι είτε τυχαίες είτε τιμές κάποιων διανυσμάτων των δεδομένων μας. Σε κάθε περίπτωση τα διανύσματα των weight vectors θα πρέπει να περιέχουν τον ίδιο αριθμό στοιχείων (τις ίδιες διαστάσεις) με τα input vectors (same dimensionality). Ο ρόλος της αρχικοποίησης είναι ιδιαίτερα σημαντικός αφού διαφορετικές αρχικοποιήσεις μπορούν να δώσουν αποτελέσματα με σημαντικές διαφορές. Για το λόγο αυτό έχουν προταθεί κάποιες πειραματικές διαδικασίες οι οποίες επιλέγουν κατάλληλες αρχικοποιήσεις. Μια καλή αρχικοποίηση, εκτός από καλύτερα αποτελέσματα αποφέρει και μικρότερο υπολογιστικό κόστος.



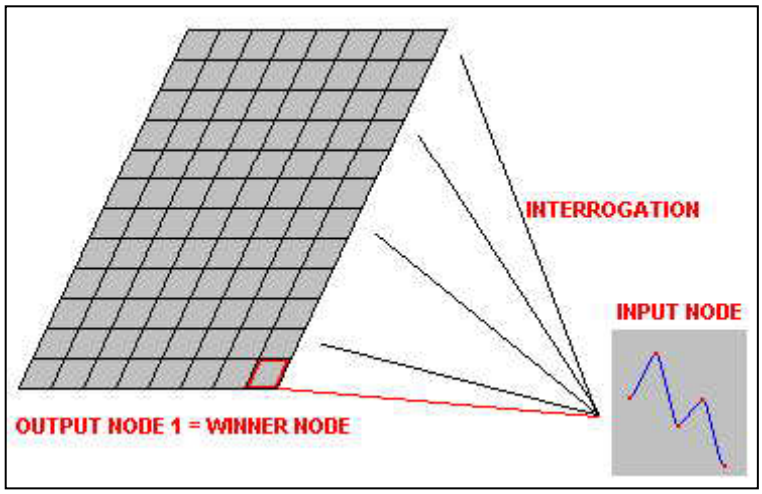
Σχήμα 5.12: Αρχικοποίηση του αλγορίθμου Self-Organising Maps. Τα weight vectors όλων των output nodes αρχικοποιούνται σε μία τυχαία ή καθορισμένη αρχική τιμή.

Για κάθε input node που εισέρχεται στον αλγόριθμο υπολογίζονται οι αποστάσεις μεταξύ του input vector και των weight vectors των output nodes. Οι αποστάσεις υπολογίζονται με βάση το μέτρο ομοιότητας που έχουμε καθορίσει (π.χ. Ευκλείδια απόσταση, συντελεστής συσχέτισης του Pearson κτλ.).



Σχήμα 5.13: Σύγκριση του input node με κάθε ένα από τα output nodes.

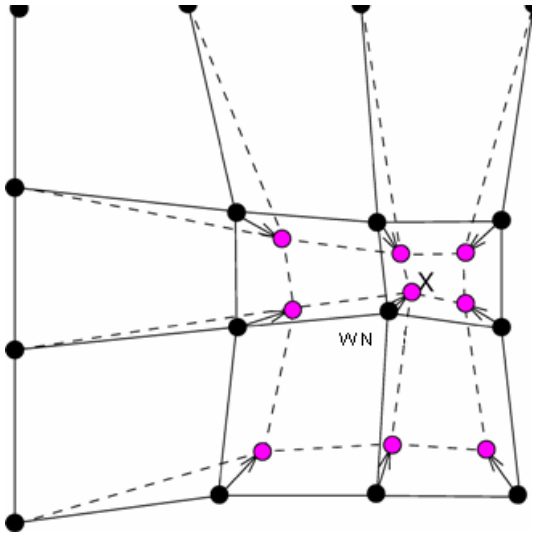
Ο αλγόριθμος εντοπίζει το output node το οποίο έχει την μεγαλύτερη ομοιότητα με το input node. Αυτό το output node το οποίο έχει την μεγαλύτερη ομοιότητα ονομάζεται winner node.



Σχήμα 5.14: Σύγκριση του input node με όλα τα output nodes. Winner node είναι το output node 1 το οποίο έχει την μεγαλύτερη ομοιότητα με το input node.

Μετά από τον καθορισμό του winner node, το weight vector του winner node αλλά και τα weight vectors των output nodes που βρίσκονται δίπλα στο winner node

τροποποιούνται με τέτοιο τρόπο ούτως ώστε να γίνουν περισσότερο όμοια με το input vector.



Σχήμα 5.15: Τροποποίηση του winner node και των γειτονικών output nodes. Στην εικόνα φαίνονται οι θέσεις του winner node (WN) και των γειτονικών output nodes πριν την τροποποίηση (μαύρες κουκκίδες) και μετά την τροποποίηση (μωβ κουκκίδες). Το input node βρίσκεται στη θέση x.

Οι τροποποιήσεις γίνονται με βάση την παρακάτω εξισώση:

$$w_j(t+1) = w_j(t) + \eta(t) h_{j,i}(t)\left(x - w_j(t)\right)$$

όπου t η επανάληψη του αλγορίθμου,

x το input vector,

$w_j(t+1)$ το weight vector μετά την τροποποίηση,

$w_j(t)$ το weight vector πριν την τροποποίηση,

$\eta(t)$ η συνάρτηση learning rate,

$h_{j,i}(t)$ η συνάρτηση neighborhood

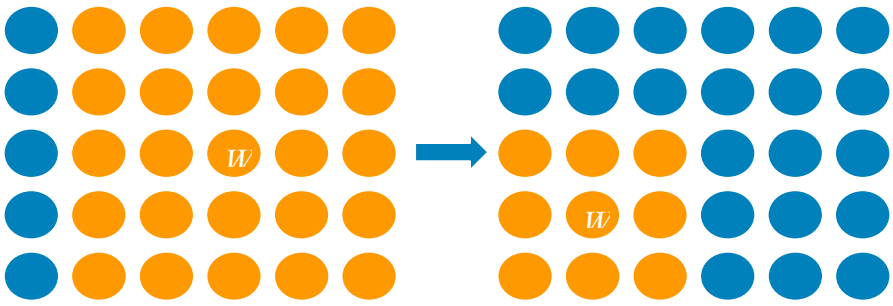
Η συνάρτηση learning rate καθορίζεται από το χρήστη και εξαρτάται από την επανάληψη t. Συνήθως είναι μια γραμμική συνάρτηση της μορφής:

$$a(t)=a(0)\ (1-t/T)$$

όπου α(0) η αρχική τιμή της συνάρτησης (συνήθως δίνεται η τιμή 0.1)
και T ο συνολικός αριθμός των επαναλήψεων.

Η συνάρτηση neighborhood καθορίζει:
α) Την ακτίνα της γειτονιάς (neighborhood radius), δηλαδή την περιοχή γύρω από το winner node στην οποία τα output nodes τροποποιούνται.
β) Το βαθμό τροποποίησης των output nodes. Όσο πιο μακριά βρίσκεται ένα output node από το winner node τόσο λιγότερο τροποποιείται το weight vector του.

Η πιο απλή μορφή που μπορεί να πάρει μια συνάρτηση neighborhood είναι η συνάρτηση φυσσαλίδα (bubble). Η συνάρτηση bubble παραμένει σταθερή μέσα στη περιοχή που καθορίζεται από την ακτίνα της γειτονιάς και παίρνει την τιμή μηδέν έξω από αυτή τη περιοχή.



10 επαναλήψεις 500 επαναλήψεις

Σχήμα 5.16: Η συνάρτηση φυσαλίδα παίρνει την ίδια τιμή για όλα τα output nodes που βρίσκονται γύρω από το winner node (W) και εντός της ακτίνας της γειτονιάς. Η ακτίνα της γειτονιάς μειώνεται με τις επαναλήψεις.

Μια άλλη μορφή συνάρτησης που χρησιμοποιείται συχνά για την συνάρτηση neighborhood είναι η συνάρτηση guassian. Σ' αυτήν τη περίπτωση η συνάρτηση neighborhood δίνεται από τον τύπο:

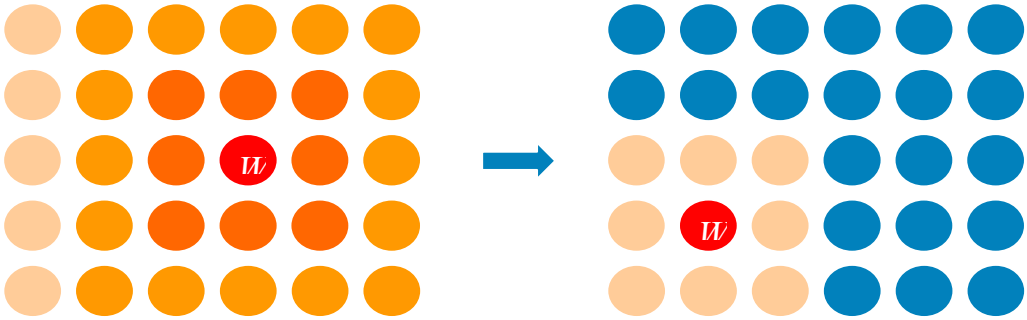
$$h_{j,i} = e^{\left(-\frac{d_{j,i}^2}{2\sigma(t)^2}\right)}$$

όπου $h_{j,i}$ η τιμή της συνάρτησης neighborhood για το output node j σε σχέση με το winner node i,

$d^2_{j,i}$ η απόσταση ανάμεσα στο output node j και το winner node i και

σ(t) η ακτίνα της γειτονιάς κατά την επανάληψη t του αλγορίθμου,

Όπως φαίνεται από την εξίσωση η τιμή της συνάρτησης μειώνεται όσο αυξάνεται η απόσταση από το winner node. Πρακτικά αυτό σημαίνει ότι όσο πιο μακριά βρίσκεται το output node από το winner node τόσο λιγότερο τροποποιείται, τόσο λιγότερο δηλαδή πλησιάζει το input node.

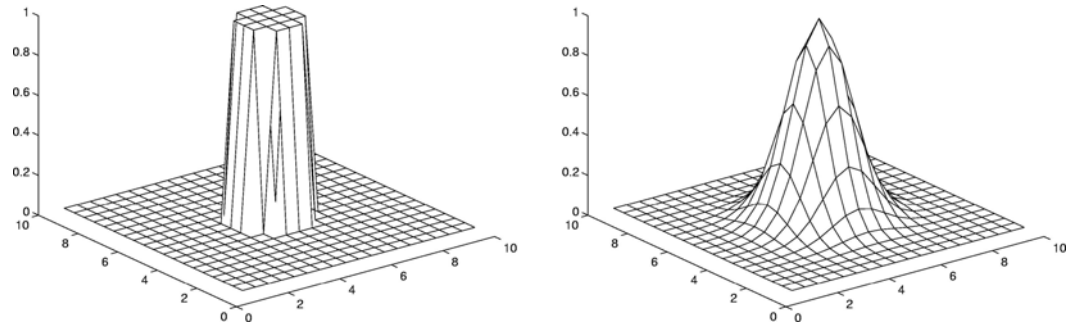


Σχήμα 5.17: Η τιμή της συνάρτησης guassian μειώνεται όσο αυξάνεται η απόσταση από το winner node. Η ένταση του χρώματος στην εικόνα δείχνει το βαθμό της μείωσης. Η ακτίνα της γειτονιάς μειώνεται με τον αριθμό των επαναλήψεων.

Για να βρούμε πόσο τροποποιείται το winner node θέτουμε $d_{j,i}=0$ οπότε $h_{j,i}=1$

Ο γενικός τύπος της τροποποίησης γίνεται:

$$w_j(t+1) = w_j(t) + \eta(t)\left(x - w_j(t)\right)$$

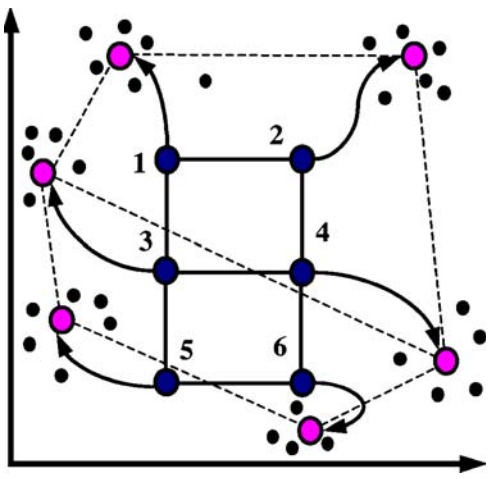


(α) (β)

Σχήμα 5.18: Οι δυο βασικές μορφές της συνάρτησης neighborhood: (α) η συνάρτηση φυσαλίδα και (β) η συνάρτηση guassian

Όταν οι τροποποιήσεις τελειώσουν, η διαδικασία επαναλαμβάνεται για τα επόμενα input node μέχρις ότου όλα τα input node να έχουν εισέλθει στον αλγόριθμο από μία φορά. Όταν εισέλθει κάποιο input node το οποίο έχει ήδη ξαναμπεί, τότε οι συναρτήσεις learning rate και neighborhood μειώνονται (εξαρτώνται από την επανάληψη t) και διατηρούν τις καινούριες τους τιμές σε όλη τη διάρκεια αυτής της επανάληψης.

Η διαδικασία σταματάει όταν οι μεταβολές στις τιμές των weight vectors γίνουν πλέον μηδαμινές ή όταν ολοκληρωθεί ένας αριθμός επαναλήψεων τον οποίο έχουμε απ' την αρχή προκαθορίσει. Στο τέλος της διαδικασίας το πρόγραμμα του υπολογιστή καταχωρεί τα input nodes πάνω στο χάρτη. Κάθε output node i προσδιορίζει μια ομάδα η οποία αποτελείται από εκείνα τα input nodes τα οποία έχουν σαν πλησιέστερο output node το output node i. Τα output nodes έχουν τροποποιηθεί με τέτοιο τρόπο ούτως ώστε να αποτελούν τα κεντροειδή των ομάδων.



Σχήμα 5.19: Τα output nodes ξεκινούν με μια αυθαίρετη τοπολογία και τροποποιούνται κατά τη διάρκεια κάθε επανάληψης. Μετά από αρκετές επαναλήψεις τα output nodes τοποθετούνται (ιδανικά) στο κέντρο κάθε ομάδας.

Ο Αλγόριθμος

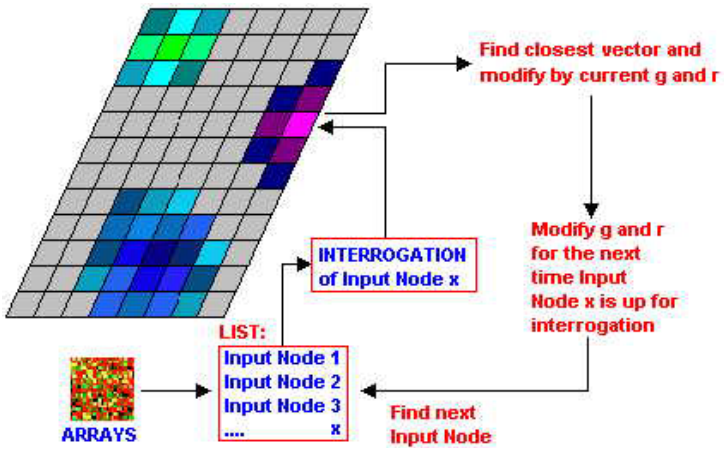
Συνοψίζοντας όλα τα παραπάνω, τα βασικά στάδια του αλγορίθμου Self-Organising Maps είναι τα εξής:

1) Αρχικοποίηση. Καθορίζεται από τον χρήστη το σχήμα του χάρτη, το μέγεθός του και το αρχικό του περιεχόμενο. Το περιεχόμενο αρχικοποιείται με τυχαίες ή καθορισμένες τιμές στα weight vectors όλων των output nodes. Στη συνέχεια καθορίζεται η μορφή των συναρτήσεων learning rate και neighborhood καθώς και οι αρχικές τους τιμές.

2) Για κάθε input node που εισέρχεται στον αλγόριθμο προσδιορίζεται το winner node, δηλαδή το output node το οποίο έχει τη μικρότερη απόσταση από το input node.

3) Γίνεται τροποποίηση στο weigh vector του winner node αλλά και στα weigh vectors των γειτονικών output nodes ούτως ώστε να γίνουν περισσότερο όμοια με το input vector.

4) Όταν όλα τα input nodes εισέλθουν στον αλγόριθμο τότε η επανάληψη του αλγόριθμου ολοκληρώνεται, οι τιμές των συναρτήσεων learning rate και neighborhood μειώνονται και ο αλγόριθμος επανέρχεται στο βήμα 3. Η διαδικασία ολοκληρώνεται όταν οι μεταβολές στις τιμές των weight vectors γίνουν μηδαμινές ή όταν ολοκληρωθεί ένας προκαθορισμένος αριθμός επαναλήψεων.

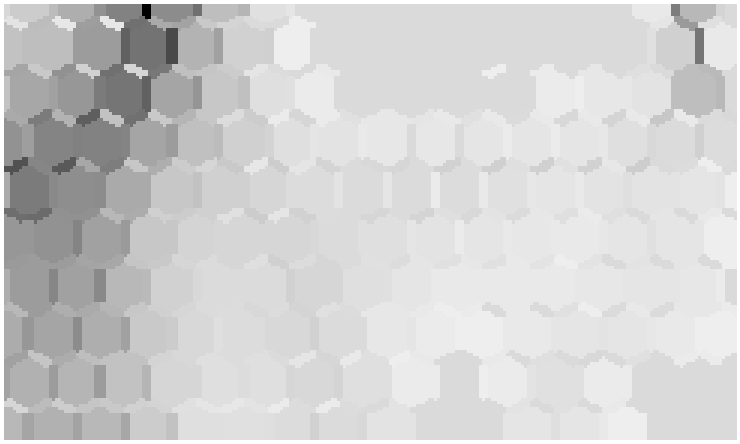


Σχήμα 5.20: Ο αλγόριθμος Self-Organining Map

Οπτική Αναπαράσταση των Αποτελεσμάτων

Οι ομάδες που προκύπτουν μετά την εφαρμογή του αλγόριθμου αποτελούνται από σύνολα διανυσμάτων τα οποία βρίσκονται αρκετά κοντά μεταξύ τους σε σχέση με τα διανύσματα των άλλων ομάδων. Το γεγονός αυτό μας δίνει την δυνατότητα να οπτικοποιήσουμε τις ομάδες που έχουν προκύψει χρησιμοποιώντας τις αποστάσεις μεταξύ των διανυσμάτων. Η πιο γνωστή μέθοδος που χρησιμοποιείται γι' αυτό το σκοπό λέγεται unified distance matrix (u-matrix).

Στη μέθοδο u-matrix υπολογίζεται ένας πίνακας αποστάσεων ανάμεσα στα διανύσματα μιας ομάδας και στα διανύσματα των γειτονικών της ομάδων. Χρησιμοποιώντας αποχρώσεις του γκρι μπορούμε να αναπαραστήσουμε τις αποστάσεις μεταξύ των γειτονικών ομάδων σε ολόκληρο το χάρτη.



Σχήμα 5.21: Αναπαράσταση των ομάδων ενός χάρτη με τη βοήθεια της μεθόδου u-matrix. Τα πιο έντονα (σκούρα) χρώματα αναπαριστούν μεγάλες αποστάσεις μεταξύ των ομάδων ενώ τα πιο ελαφριά χρώματα αναπαριστούν μικρότερες αποστάσεις.

Πλεονεκτήματα – Μειονεκτήματα

Το πιο σημαντικό πλεονέκτημα του αλγορίθμου Self-Organising Maps είναι η δυνατότητα οπτικοποίησης των αποτελεσμάτων η οποία μας προσφέρει επιπλέον πληροφορίες για την ομαδοποίηση (π.χ πληροφορίες για την απόσταση μεταξύ των γειτονικών ομάδων αν χρησιμοποιήσουμε την μέθοδο u-matrix). Αυτές οι επιπλέον δυνατότητες αποτελούν το πλεονέκτημα αυτού του αλγορίθμου σε σχέση με άλλους διαμεριστικούς αλγορίθμους όπως ο K-Means.

Ένα άλλο πλεονέκτημα του αλγορίθμου είναι οι μικρές απαιτήσεις σε υπολογιστική ισχύ. Στον αλγόριθμο Self-Organising Maps δε δημιουργείται κάποιος πίνακας αποστάσεων όπως συμβαίνει για παράδειγμα στον αλγόριθμο Hierarchical Clustering. Το γεγονός αυτό έχει ως αποτέλεσμα το υπολογιστικό κόστος να διατηρείται σε χαμηλά επίπεδα. Ο χρόνος υπολογισμού εξαρτάται από τον αριθμό των επαναλήψεων και το μέγεθος του χάρτη και συνήθως δεν ξεπερνάει τα μερικά δευτερόλεπτα.

Το μεγαλύτερο μειονέκτημα του αλγορίθμου είναι η εξάρτηση του αποτελέσματος από τον καθορισμό των αρχικών παραμέτρων. Όπως έχει ήδη προαναφερθεί για να υλοποιήσουμε τον αλγόριθμο θα πρέπει να καθορίσουμε:

- Το σχήμα και το μέγεθος του χάρτη.
- Τον αριθμό των επαναλήψεων.
- Τον τύπο της συνάρτησης learning rate και την αρχική της τιμή.
- Τον τύπο συνάρτησης neighborhood και την ακτίνα της γειτονιάς (neighborhood radius).
- Τον τρόπο με τον οποίο θα γίνει η αρχικοποίηση των διανυσμάτων (weight vectors) του χάρτη (είτε τυχαία είτε με βάση κάποιες τιμές από τα διανύσματα των δεδομένων).

Έχει παρατηρηθεί ότι οι τιμές των συναρτήσεων learning rate και neighborhood δεν επηρεάζουν σημαντικά το αποτέλεσμα. Αντίθετα το αποτέλεσμα εξαρτάται σε μεγάλο βαθμό από τις διαστάσεις του χάρτη (δηλαδή από τον αριθμό των ομάδων) και από τον αριθμό των επαναλήψεων.

Ένα άλλο μειονέκτημα αυτού του αλγορίθμου είναι ότι το αποτέλεσμα εξαρτάται από την σειρά με την οποία εισέρχονται τα δεδομένα στον αλγόριθμο. Για να περιοριστούν οι επιπτώσεις αυτού του προβλήματος τα δεδομένα εισέρχονται στον αλγόριθμο με τυχαίο τρόπο σε κάθε επανάληψη και όχι με κάποια καθορισμένη σειρά.

Κεφάλαιο 6 - Τεχνικές Αξιολόγησης Ομάδας (Cluster Validation Techniques)

6.1 Εισαγωγή

Ένα από τα σημαντικότερα προβλήματα της μη επιβλεπόμενης κατηγοριοποίησης είναι η αξιολόγηση των αποτελεσμάτων. Όπως έχουμε ήδη προαναφέρει το αποτέλεσμα των μεθόδων ομαδοποίησης εξαρτάται άμεσα από τις αρχικές υποθέσεις. Διαφορετικές τιμές στις αρχικές παραμέτρους ενός αλγορίθμου μπορεί να οδηγήσουν σε διαφορετικά σχήματα ομαδοποίησης του ίδιου συνόλου δεδομένων. Κατά συνέπεια, στις περισσότερες περιπτώσεις, το τελικό σχήμα ομαδοποίησης απαιτεί κάποιο είδος αξιολόγησης. Τη λύση σ' αυτό το πρόβλημα ήρθαν να δώσουν κάποιοι δείκτες οι οποίοι μας δίνουν πληροφορίες για τον βαθμό εγκυρότητας των αποτελεσμάτων. Οι δείκτες αυτοί ονομάζονται δείκτες εγκυρότητας ομάδας.

6.2 Δείκτες Εγκυρότητας Ομάδας (Cluster Validity Indices)

6.2.1 Γενικά

Οι δείκτες εγκυρότητας ομάδας έχουν ως βασικό στόχο τον προσδιορισμό των βέλτιστων τιμών των παραμέτρων ενός αλγορίθμου και ιδιαίτερα τον προσδιορισμό του βέλτιστου αριθμού των ομάδων [16,17,18]. Όπως έχει ήδη προαναφερθεί, στις μεθόδους μη επιβλεπόμενης κατηγοριοποίησης, συνήθως δεν υπάρχει κάποιος προφανής αριθμός των ομάδων που θα προκύψουν. Ο αριθμός καθορίζεται αυθαίρετα από μας και κατά συνέπεια οδηγεί σε αποτελέσματα περιορισμένης αξιοπιστίας. Για να προσδιορίσουμε τον βέλτιστο αριθμό των ομάδων αλλά και τις βέλτιστες τιμές των άλλων παραμέτρων μπορούμε να εφαρμόσουμε τον αλγόριθμο αρκετές φορές με διαφορετικούς συνδυασμούς στις τιμές των παραμέτρων και να αξιολογήσουμε τα αποτελέσματα με βάση τους δείκτες εγκυρότητας που έχουμε επιλέξει. Οι βέλτιστες

τιμές των παραμέτρων είναι αυτές που επιφέρουν τις βέλτιστες τιμές στους δείκτες εγκυρότητας και κατά συνέπεια το βέλτιστο σχήμα ομαδοποίησης.

Μία άλλη δυνατότητα που μας παρέχουν οι δείκτες εγκυρότητας ομάδας είναι η σύγκριση των μεθόδων ομαδοποίησης. Ο καλύτερος τρόπος για να συγκρίνουμε αλγορίθμους ομαδοποίησης είναι να προσδιορίσουμε τις βέλτιστες τιμές των παραμέτρων (όπως περιγράψαμε προηγουμένως) για κάθε αλγόριθμο και στη συνέχεια να τους συγκρίνουμε με βάση αυτές τις τιμές. Δηλαδή με βάση το καλύτερο σχήμα ομαδοποίησης που μπορεί να δώσει κάθε αλγόριθμος.

Σ' αυτήν την εργασία οι δείκτες αξιολόγησης θα χρησιμοποιηθούν με τρόπο διαφορετικό σε σχέση μ' αυτόν που συναντάμε μέχρι σήμερα στην βιβλιογραφία. Πιο συγκεκριμένα, δε θα χρησιμοποιήσουμε τους δείκτες για να αξιολογήσουμε σχήματα ομαδοποίησης που προέκυψαν από διάφορους αλγόριθμους, αλλά για να αξιολογήσουμε σύνολα δεδομένων που προέκυψαν από διαδικασίες επιλογής γονιδίων. Περισσότερες λεπτομέρειες γι' αυτό το θέμα θα αναφέρουμε στο επόμενο κεφάλαιο.

Τα κριτήρια που χρησιμοποιούν οι δείκτες εγκυρότητας ομάδας για να αξιολογήσουν ένα σχήμα ομαδοποίησης είναι:

α) Η πυκνότητα των ομάδων. Τα αντικείμενα σε κάθε ομάδα θα πρέπει να είναι όσο το δυνατόν πιο όμοια μεταξύ τους.

β) Η διαχωρισιμότητα μεταξύ των ομάδων. Οι ομάδες θα πρέπει να είναι καλά διαχωρισμένες μεταξύ τους.

Για να μετρηθεί η πυκνότητα μιας ομάδας θα πρέπει να υπολογισθεί η εσωτερική της απόσταση (intracluster distance) [19]. Οι πιο συνηθισμένες μέθοδοι που χρησιμοποιούνται για τον υπολογισμό της εσωτερικής απόστασης είναι:

α) Η πλήρης διάμετρος (complete diameter). Η εσωτερική απόσταση μιας ομάδας καθορίζεται από την απόσταση μεταξύ των δύο πιο μακρινών αντικειμένων της ομάδας και δίνεται από τον τύπο:

$$d(c) = \max_{x,y \in c} \{d(x,y)\}$$

β) Η μέση διάμετρος (average diameter). Η εσωτερική απόσταση μιας ομάδας καθορίζεται από τη μέση απόσταση μεταξύ όλων των αντικειμένων που ανήκουν στην ομάδα.

$$d(c) = \frac{1}{N} \sum_{\substack{x,yec \\ x \neq y}} d(x, y)$$

όπου N είναι ο αριθμός των αποστάσεων μεταξύ των αντικειμένων.

γ) Η κεντροειδής διάμετρος (centroid diameter). Η εσωτερική απόσταση μιας ομάδας καθορίζεται από το διπλάσιο της μέσης απόστασης όλων των αντικειμένων από το κεντροειδές της ομάδας.

$$d(c) = 2\left(\frac{\sum_{xec} d(x, v)}{n}\right)$$

όπου n ο αριθμός των αντικειμένων της ομάδας και v το κεντροειδές το οποίο καθορίζεται από τον τύπο:

$$v = \frac{1}{n} \sum_{xec} x$$

Για να μετρηθεί η διαχωρισιμότητα μεταξύ των ομάδων χρησιμοποιούνται οι μέθοδοι διασύνδεσης (linkage methods). Οι μέθοδοι που χρησιμοποιούνται περισσότερο είναι η single linkage, η complete linkage η average linkage και η centroid linkage τις οποίες τις έχουμε περιγράψει στο προηγούμενο κεφάλαιο, στην παράγραφο με τους ιεραρχικούς αλγόριθμους.

6.2.2 Ο Δείκτης C (C-Index)

Ο δείκτης C δίνεται από τον τύπο:

$$C = \frac{S - S_{\min}}{S_{\max} - S_{\min}}$$

Αν θεωρήσουμε ότι p είναι ο αριθμός όλων των ζευγαριών των αντικειμένων στα οποία και τα δύο αντικείμενα βρίσκονται στην ίδια ομάδα τότε το S ορίζεται ως το

άθροισμα των αποστάσεων μεταξύ των αντικειμένων σ' αυτά τα p ζευγάρια. Έστω P ο αριθμός όλων των δυνατών συνδυασμών ζευγαριών μέσα στο σύνολο δεδομένων. Ταξινομούμε τα P ζευγάρια με βάση τις αποστάσεις μεταξύ των αντικειμένων και επιλέγουμε τα p ζευγάρια με τις μικρότερες αποστάσεις και τα p ζευγάρια με τις μεγαλύτερες αποστάσεις. Το $S_{\min}$ ισούται με το άθροισμα των p μικρότερων αποστάσεων και το $S_{\max}$ με το άθροισμα των p μεγαλύτερων αποστάσεων. Από τον τύπο φαίνεται ότι όσο πιο μικρό είναι το C τόσο περισσότερο το άθροισμα των αποστάσεων των ζευγαριών που βρίσκονται στην ίδια ομάδα πλησιάζει το άθροισμα των αποστάσεων των ζευγαριών με τη μικρότερη απόσταση. Αυτό πρακτικά σημαίνει ότι τα ζευγάρια των αντικειμένων που βρίσκονται στην ίδια ομάδα ταυτίζονται σχεδόν με τα ζευγάρια που έχουν τη μικρότερη απόσταση. Άρα, όσο περισσότερο πλησιάζει το C στο μηδέν τόσο καλύτερο είναι το σχήμα ομαδοποίησης. Αντίθετα, όταν το C παίρνει τιμές κοντά στο ένα η ομαδοποίηση κρίνεται μη ικανοποιητική.

6.2.3 Ο Δείκτης του Dunn (Dunn's Index)

Ο δείκτης του Dunn δίνεται από τον τύπο:

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

Το $d(c_i, c_j)$ είναι η απόσταση μεταξύ των ομάδων i και j (intercluster distance), το $d'(c_k)$ είναι η εσωτερική απόσταση της ομάδας c_k (intracluster distance) και το n είναι ο αριθμός των ομάδων. Όπως φαίνεται από τον τύπο ο δείκτης Dunn καθορίζεται με βάση τις αδυναμίες της ομαδοποίησης. Πιο συγκεκριμένα, καθορίζεται από την απόσταση των ομάδων που έχουν την μικρότερη απόσταση μεταξύ τους (αριθμητής) και από την εσωτερική απόσταση της ομάδας που έχει την μεγαλύτερη εσωτερική απόσταση (παρονομαστής). Είναι φανερό λοιπόν ότι όταν ο δείκτης Dunn παίρνει μεγάλες τιμές η ομαδοποίηση κρίνεται ικανοποιητική αφού οι αποστάσεις μεταξύ των ομάδων είναι αρκετά μεγαλύτερες σε σχέση με τις

εσωτερικές τους αποστάσεις. Τα σημαντικότερα μειονεκτήματα του δείκτη Dunn είναι οι μεγάλες απαιτήσεις σε υπολογιστική ισχύ και η ευαισθησία του στο θόρυβο.

6.2.4 Ο Δείκτης Davies-Bouldin (Davies-Bouldin Index)

Ο δείκτης Davies-Bouldin δίνεται από τον τύπο:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j} \left\{ \frac{S_n(Q_i) + S_n(Q_j)}{S(Q_i, Q_j)} \right\}$$

Τα $S_n(Q_i)$ και $S_n(Q_j)$ είναι οι εσωτερικές αποστάσεις των ομάδων Q_i και Q_j (intracluster distance) ενώ το $S(Q_i,Q_j)$ ισούται με την απόσταση μεταξύ των ομάδων Q_i και Q_j (intercluster distance). Ο αριθμός των ομάδων είναι n. Από τον τύπο παρατηρούμε ότι ο δείκτης DB παίρνει μικρές τιμές όταν οι εσωτερικές αποστάσεις των ομάδων είναι μικρότερες σε σχέση με τις αποστάσεις που έχουν μεταξύ τους. Συνεπώς, όσο μικρότερος είναι ο δείκτης Davies-Bouldin τόσο καλύτερη είναι η ομαδοποίηση που εξετάζουμε. Παρόλο που ο δείκτης Davies-Bouldin έχει εμφανείς ομοιότητες με τον δείκτη Dunn, θεωρείται καλύτερος απ' αυτόν κυρίως επειδή είναι πιο ανθεκτικός στην παρουσία θορύβου.

6.2.5 Ο Δείκτης Silhouette (Silhouette Index)

Ο δείκτης εγκυρότητας Silhouette μπορεί να υπολογισθεί είτε για ένα συγκεκριμένο αντικείμενο μιας ομάδας, είτε για μία ολόκληρη ομάδα είτε για ένα ολόκληρο σχήμα ομαδοποίησης [20]. Αν υποθέσουμε ότι i είναι ένα οποιοδήποτε αντικείμενο που ανήκει στην ομάδα A τότε ο δείκτης Silhouette γι' αυτό το αντικείμενο δίνεται από τον τύπο:

$$S(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}}$$

Το α(i) είναι η μέση απόσταση του i απ' όλα τα υπόλοιπα αντικείμενα που ανήκουν στην ομάδα A. Έστω C μια οποιαδήποτε ομάδα διαφορετική από την A και d(i,C) η

μέση απόσταση του αντικειμένου i από όλα τα αντικείμενα που ανήκουν στην ομάδα C. Υπολογίζουμε την μέση απόσταση του αντικειμένου i από όλες τις ομάδες $C \neq A$ και ορίζουμε ως b(i) την μικρότερη απ' αυτές τις τιμές.

$$b(i) = \min_{C \neq A} d(i, C)$$

Η ομάδα Β για την οποία ισχύει d(i,B)=b(i) ονομάζεται «γείτονας» (neighbor) του αντικειμένου i και αποτελεί τη δεύτερη καλύτερη επιλογή για την τοποθέτηση του αντικειμένου μετά την ομάδα Α. Ο υπολογισμός του δείκτη Silhouette βασίζεται στην ουσία στη σύγκριση ανάμεσα στην μέση απόσταση του αντικειμένου από την ομάδα του Α και τη μέση απόσταση του αντικειμένου από την πλησιέστερη προς αυτό ομάδα Β.

Από τον τύπο παρατηρούμε ότι το S(i) παίρνει τιμές ανάμεσα στο -1 και το 1. Όταν η τιμή του δείκτη είναι κοντά στο 1 τότε ομαδοποίηση θεωρείται ικανοποιητική αφού η μέση απόσταση του αντικειμένου από την ομάδα του είναι πολύ μικρότερη σε σχέση με την μέση απόσταση από την πλησιέστερη προς αυτό ομάδα. Στην αντίθετη περίπτωση που το S(i) παίρνει τιμές κοντά στο -1 η ομαδοποίηση θεωρείται μη ικανοποιητική αφού υπάρχει κάποια ομάδα που είναι πιο κοντά στο αντικείμενο σε σχέση με την ομάδα στην οποία ανήκει. Όταν το S(i) παίρνει τιμές κοντά στο μηδέν τότε το a(i) και το b(i) είναι περίπου ίσα. Αυτό σημαίνει ότι δεν είναι ξεκάθαρο σε ποια ομάδα θα πρέπει να τοποθετηθεί το αντικείμενο αφού απέχει περίπου το ίδιο από τις ομάδες Α και Β. Αν η ομάδα Α περιέχει μόνο ένα αντικείμενο τότε δεν είναι δυνατόν να υπολογίσουμε το a(i) και θεωρούμε ότι ο δείκτης S(i) παίρνει τη τιμή μηδέν.

Η τιμή του δείκτη Silhouette για μία ολόκληρη ομάδα C_j ισούται με το μέσο όρο της τιμής του δείκτη για όλα τα αντικείμενα της ομάδας και δίνεται από τον τύπο:

$$S_j = \frac{1}{m} \sum_{i=1}^{m} S(i)$$

όπου m είναι ο αριθμός των αντικειμένων της ομάδας C_j.

Η τιμή του δείκτη για ένα ολόκληρο σχήμα ομαδοποίησης C ισούται με το μέσο όρο της τιμής του δείκτη για όλες τις ομάδες C_j όπου j=1,2,…,n και δίνεται από τον τύπο:

$$S(C) = \frac{1}{n}\sum_{i=1}^{n} S_j$$

6.2.6 Ο Δείκτης Goodman-Kruskal (Goodman-Kruskal Index)

Ο υπολογισμός του δείκτη Goodman-Kruskal βασίζεται στη σύγκριση όλων των δυνατών τετράδων ενός συνόλου δεδομένων. Έστω ένα σύνολο δεδομένων X_j (j=1,2,…,κ, όπου κ ο συνολικός αριθμός των αντικειμένων j) και d η απόσταση μεταξύ δύο οποιωνδήποτε αντικειμένων a και b ή c και d. Μία τετράδα λέγεται αρμονική αν μία από τις παρακάτω συνθήκες είναι αληθής:

- d(a,b)<d(c,d) όπου τα a,b ανήκουν στην ίδια ομάδα και τα c,d ανήκουν σε διαφορετικές.
- d(a,b)>d(c,d) όπου τα a,b ανήκουν σε διαφορετικές ομάδες και τα c,d ανήκουν στην ίδια ομάδα.

Αντίστοιχα μία τετράδα λέγεται μη αρμονική όταν ισχύει μία από τις παρακάτω συνθήκες:

- d(a,b)<d(c,d) όπου τα a,b ανήκουν σε διαφορετικές ομάδες και τα c,d ανήκουν στην ίδια ομάδα.
- d(a,b)>d(c,d) όπου τα a,b ανήκουν στην ίδια ομάδα και τα c,d ανήκουν σε διαφορετικές.

Ένα καλό σχήμα ομαδοποίησης θα πρέπει να περιέχει πολλές αρμονικές και λίγες μη αρμονικές ομάδες. Αν υποθέσουμε ότι N_c και N_d είναι οι αριθμοί των αρμονικών και των μη αρμονικών τετράδων αντίστοιχα, τότε ο δείκτης εγκυρότητας Goodman-Kruskal δίνεται από τον τύπο:

$$GK = \frac{N_c - N_d}{N_c + N_d}$$

Από τον παραπάνω τύπο παρατηρούμε ότι όσο μεγαλύτερη είναι η τιμή του δείκτη Goodman-Kruskal τόσο καλύτερο είναι το σχήμα ομαδοποίησης που εξετάζεται.

6.2.7 Ο Δείκτης Isolation (Isolation Index)

Ο δείκτης Isolation υπολογίζεται με βάση τη σταθερά «κ πλησιέστεροι γείτονες» (k-nearest neighbour norm – k-nn). Για μία συγκεκριμένη τιμή του κ (η ακριβής τιμή του δεν έχει μεγάλη σημασία), η σταθερά «κ πλησιέστεροι γείτονες» $\nu_k(x_i)$ για ένα σημείο δεδομένων x_i ισούται με το κλάσμα των k πλησιέστερων γειτόνων που ανήκουν στην ίδια ομάδα με το x_i. Ο δείκτης Isolation για ένα ολόκληρο σχήμα ομαδοποίησης ισούται με το μέσο όρο της σταθεράς «k πλησιέστεροι γείτονες» για όλα τα αντικείμενα που ανήκουν στο σύνολο δεδομένων και δίνεται από τον παρακάτω τύπο:

$$I_k = \frac{1}{n}\sum_{i=1}^{n} \nu_k(x_i)$$

Είναι προφανές ότι όταν η ομαδοποίηση είναι καλή, όλα σχεδόν τα αντικείμενα που βρίσκονται δίπλα στο x (οι πλησιέστεροι γείτονες) θα ανήκουν στην ίδια ομάδα με αυτό. Σε μια τέτοια περίπτωση η τιμή της σταθεράς θα είναι περίπου 1 ($\nu_k(x) \approx 1$). Αντίθετα στη περίπτωση που η ομαδοποίηση δεν είναι καλή η τιμή της σταθεράς θα είναι αρκετά κάτω από το 1.

Ο δείκτης Isolation εξετάζει στην ουσία το κατά πόσο μία ομάδα είναι αρκετά απομονωμένη σε σχέση με το υπόλοιπο σύνολο δεδομένων. Το σημαντικότερο μειονέκτημα αυτού του δείκτη είναι ότι δε μπορεί να αναγνωρίσει το λάθος της συγχώνευσης των ομάδων το οποίο συμβαίνει όταν δύο καλά διαχωρισμένες ομάδες εμφανίζονται στο σχήμα ομαδοποίησης ως μία. Ακόμα και στην ακραία περίπτωση στην οποία όλα τα αντικείμενα του συνόλου δεδομένων έχουν τοποθετηθεί σε μία και μόνο ομάδα ο δείκτης Isolation θα έδινε την καλύτερη τιμή ($\nu_k(x) = 1$).

Κεφάλαιο 7 - Ο Σκοπός Της Εργασίας, Τα Σύνολα Δεδομένων Και Οι Μέθοδοι Που Χρησιμοποιήθηκαν

7.1 Εισαγωγή

Σ' αυτό το κεφάλαιο θα περιγράψουμε το σκοπό αυτής της εργασίας καθώς και τα σύνολα δεδομένων και τις μεθόδους που χρησιμοποιήθηκαν για την εξαγωγή των αποτελεσμάτων. Ο βασικός στόχος αυτής της εργασίας είναι η αξιολόγηση συνόλων δεδομένων τα οποία έχουν προκύψει από την εφαρμογή μεθόδων επιλογής γονιδίων. Η αξιολόγηση γίνεται με βάση κάποιους μεθόδους ομαδοποίησης και κάποιους δείκτες εγκυρότητας.

7.2 Ο Σκοπός Της Εργασίας

Όπως έχουμε ήδη προαναφέρει η ανάλυση των δεδομένων που προκύπτουν απ' την εφαρμογή της μεθόδου DNA Microarrays μπορεί να γίνει με πολλούς διαφορετικούς τρόπους ανάλογα με τους στόχους που θέλουμε να επιτύχουμε. Ένας απ' αυτούς τους τρόπους είναι η εφαρμογή επιβλέπουσας κατηγοριοποίησης (supervised classification) την οποία έχουμε περιγράψει με συντομία στο τέταρτο κεφάλαιο. Στην επιβλέπουσα κατηγοριοποίηση, ένα σύνολο από προομαδοποιημένα αντικείμενα (training data) χρησιμοποιούνται για να κατασκευάσουμε μια συνάρτηση απόφασης (decision function). Η συνάρτηση απόφασης μας δίνει την δυνατότητα να εντοπίσουμε την ομάδα στην οποία ανήκουν αντικείμενα των οποίων η ομάδα μας είναι άγνωστη (test data). Τα σύνολα δεδομένων που θα χρησιμοποιηθούν σ' αυτήν την εργασία περιέχουν αντικείμενα (δείγματα) τα οποία ανήκουν σε δύο ομάδες. Αν συμβολίσουμε τις δύο ομάδες με τα σύμβολα (+) και (-) τότε ισχύει:

$$D(x) > 0 \Rightarrow x \in (+)$$

$$D(x) < 0 \Rightarrow x \in (-)$$

όπου x ένα δείγμα άγνωστης ομάδας και D(x) η συνάρτηση απόφασης.

Ένα από τα σημαντικότερα θέματα που έχουν προκύψει σχετικά με την επιβλέπουσα κατηγοριοποίηση είναι η ανάγκη για μείωση των διαστάσεων των δεδομένων. Όπως έχουμε ήδη προαναφέρει οι πίνακες γονιδιακής έκφρασης που προκύπτουν από την εφαρμογή της μεθόδου DNA Microarrays περιέχουν αρκετές χιλιάδες γονιδίων και μερικές δεκάδες δειγμάτων. Τα σύνολα δεδομένων αυτών των διαστάσεων, στα οποία ο αριθμός των γονιδίων είναι πολύ μεγαλύτερος σε σχέση με τον αριθμό των δειγμάτων, αντιμετωπίζουν ένα πρόβλημα το οποίο συναντάται στη βιβλιογραφία με τον όρο «overfitting». Σαν συνέπεια αυτού του προβλήματος μπορεί κάποιος να βρει εύκολα μία συνάρτηση απόφασης η οποία να διαχωρίζει τα προομαδοποιημένα δείγματα αλλά η συνάρτηση αυτή έχει πολύ χαμηλή απόδοση όταν χρησιμοποιείται για να κατηγοριοποιήσει δείγματα άγνωστης ομάδας. Για να ξεπεραστεί αυτό το πρόβλημα απαιτείται μία διαδικασία επιλογής γονιδίων η οποία θα καταλήγει σε ένα σύνολο δεδομένων πολύ μικρότερο από το αρχικό. Τα γονίδια που επιλέγονται θα πρέπει να διαφοροποιούνται όσο το δυνατόν περισσότερο μεταξύ των ομάδων που μας ενδιαφέρουν. Η διαδικασία αυτή συναντάται στη βιβλιογραφία με τον όρο «επιλογή γονιδίων» (gene selection) ενώ τα γονίδια στα οποία καταλήγει μια τέτοια διαδικασία ονομάζονται «γονίδια δείκτες» (marker genes). Εκτός από την αντιμετώπιση του προβλήματος «overfitting», μια διαδικασία επιλογής γονιδίων παρουσιάζει και άλλα σημαντικά πλεονεκτήματα. Μερικά απ' αυτά είναι:

(1) Η ακρίβεια της κατηγοριοποίησης βελτιώνεται. Αρκετές μελέτες έχουν δείξει ότι η ακρίβεια της κατηγοριοποίησης που βασίζεται σε μικρά σύνολα με γονίδια δείκτες είναι μεγαλύτερη σε σχέση με την ακρίβεια που δίνουν τα αρχικά σύνολα δεδομένων.

(2) Οι κλινικές μελέτες των βιολόγων γίνονται πιο εύκολες και το κόστος τους μειώνεται σημαντικά. Αντίθετα στα μεγάλα σύνολα δεδομένων ο βαθμός δυσκολίας των μελετών είναι πολύ μεγαλύτερος και το οικονομικό κόστος δυσβάσταχτο για τα περισσότερα εργαστήρια.

(3) Δίνει την δυνατότητα στους βιολόγους να μελετήσουν σε βάθος συγκεκριμένα γονίδια τα οποία φέρονται να έχουν σχέση με κάποια ασθένεια. Η δυνατότητα αυτή επιτρέπει στους βιολόγους να κατανοήσουν τη γενετική δομή και τους μηχανισμούς που σχετίζονται με την ασθένεια, γεγονός που μπορεί να οδηγήσει στην έγκαιρη διάγνωσή της και στην ανακάλυψη νέων φαρμάκων.

Όλα τα παραπάνω καταδεικνύουν τη σημαντική πρακτική αξία που έχει η επιλογή γονιδίων καθώς και την αναγκαιότητα η επιλογή να είναι όσο το δυνατόν καλύτερη

ούτως ώστε να αποφέρει τα προσδοκώμενα αποτελέσματα. Τα κριτήρια τα οποία χρησιμοποιούνται για να αξιολογήσουμε μία επιλογή γονιδίων είναι η **ακρίβεια** (accuracy) και η **ποιότητα** (quality). Μία διαδικασία επιλογής γονιδίων έχει υψηλή ακρίβεια όταν τα γονίδια δείκτες που προκύπτουν μπορούν να προβλέψουν σωστά την ομάδα στην οποία ανήκουν κάποια δείγματα άγνωστης ομάδας. Η ποιότητα μιας επιλογής γονιδίων θεωρείται καλή όταν τα γονίδια παρουσιάζουν παραπλήσια συμπεριφορά σε κάθε μία από τις ομάδες που εξετάζουμε, ενώ η συμπεριφορά τους μεταξύ των ομάδων διαφοροποιείται σημαντικά [23]. Παρόλο που τα κριτήρια αυτά φαίνονται με μια πρώτη ματιά να σχετίζονται μεταξύ τους, υπάρχουν περιπτώσεις στις οποίες τα γονίδια δείκτες παρουσιάζουν υψηλή ακρίβεια αλλά χαμηλή ποιότητα. Η αξιοπιστία ενός συνόλου γονιδίων που παρουσιάζει μια τέτοια συμπεριφορά είναι περιορισμένη.

Μέχρι τώρα οι περισσότερες μελέτες που ασχολούνται με την αξιολόγηση της επιλογής γονιδίων έχουν επικεντρωθεί, σχεδόν αποκλειστικά, στην ακρίβεια και ελάχιστα στην ποιότητα των αποτελεσμάτων μιας τέτοιας διαδικασίας. Ο σκοπός αυτής της εργασίας είναι να προτείνει ένα νέο τρόπο αξιολόγησης της ποιότητας ενός συνόλου δεδομένων ο οποίος θα βασίζεται σε μεθόδους ομαδοποίησης (μη επιβλέπουσας κατηγοριοποίησης) και σε δείκτες εγκυρότητας.

Οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται συνήθως με σκοπό την ανακάλυψη ομάδων μέσα σε ένα σύνολο δεδομένων οι οποίες ήταν προηγουμένως άγνωστες σε μας. Η διαδικασία αυτή, η οποία είναι γνωστή με τον όρο **ανακάλυψη κλάσεων** (class discovery), χρησιμοποιείται συνήθως για να εντοπιστούν άγνωστες έως τώρα υποκατηγορίες διαφόρων ασθενειών. Σ' αυτή την εργασία οι αλγόριθμοι ομαδοποίησης θα χρησιμοποιηθούν με τρόπο διαφορετικό απ' αυτόν που χρησιμοποιούνται συνήθως. Θα χρησιμοποιηθούν σε σύνολα δεδομένων στα οποία οι ομάδες είναι γνωστές από πριν, με σκοπό να βρούμε κατά πόσο (σε ποιο ποσοστό) μπορούν να τις εντοπίσουν. Όσο καλύτερα μπορεί ένας αλγόριθμος ομαδοποίησης να εντοπίσει αυτές τις ομάδες, τόσο πιο «ενδεικτικά» είναι τα γονίδια που χρησιμοποιούμε όσον αφορά τη σχέση τους με τις ομάδες που μας ενδιαφέρουν. Αυτό συμβαίνει επειδή όταν τα γονίδια παίρνουν τιμές παραπλήσιες εντός των ομάδων και αρκετά διαφορετικές μεταξύ των ομάδων, ο αλγόριθμος ομαδοποίησης μπορεί να εντοπίσει τις διαφορετικές ομάδες σχετικά εύκολα και με μεγάλο ποσοστό επιτυχίας. Αντίθετα αν το σύνολο δεδομένων περιέχει γονίδια τα οποία παίρνουν περίπου τις

ίδιες τιμές σε όλες τις ομάδες ο εντοπισμός των ομάδων γίνεται δυσκολότερος και το ποσοστό επιτυχίας μειώνεται σημαντικά.

Το πρώτο κριτήριο λοιπόν που θα χρησιμοποιήσουμε για να αξιολογήσουμε ένα σύνολο δεδομένων είναι το ποσοστό επιτυχίας με το οποίο μπορεί ένας αλγόριθμος ομαδοποίησης να εντοπίσει τις ομάδες που μας ενδιαφέρουν. Το δεύτερο κριτήριο βασίζεται στους δείκτες εγκυρότητας. Όπως έχουμε ήδη προαναφέρει στο προηγούμενο κεφάλαιο, οι δείκτες εγκυρότητας χρησιμοποιούνται συνήθως είτε για να προσδιορίσουμε τις βέλτιστες τιμές των παραμέτρων ενός αλγορίθμου ομαδοποίησης είτε για να συγκρίνουμε δύο αλγόριθμους μεταξύ τους. Και στις δύο περιπτώσεις, οι δείκτες χρησιμοποιούνται για να αξιολογήσουν το σχήμα ομαδοποίησης που προέκυψε από την εφαρμογή ενός ή περισσοτέρων αλγορίθμων. Σ' αυτή την εργασία οι δείκτες εγκυρότητας δεν χρησιμοποιούνται σε ομάδες που προέκυψαν από αλγόριθμους ομαδοποίησης. Χρησιμοποιούνται σε σύνολα δεδομένων στα οποία οι ομάδες, μας είναι ήδη γνωστές. Ο στόχος μας είναι να υπολογίσουμε την ποιότητα των συνόλων δεδομένων που έχουν προκύψει από διάφορες διαδικασίες επιλογής γονιδίων. Οι δείκτες εγκυρότητας δίνουν καλά αποτελέσματα όταν οι ομάδες που μας ενδιαφέρουν είναι συμπαγείς και αρκετά διαφορετικές μεταξύ τους. Άρα, όσο πιο «ενδεικτικά» είναι τα γονίδια δείκτες που μελετάμε τόσο καλύτερα είναι τα αποτελέσματα που δίνει ένας δείκτης εγκυρότητας.

Συμπερασματικά λοιπόν θα λέγαμε, ότι οι μέθοδοι ομαδοποίησης και οι δείκτες εγκυρότητας μπορούν, αν χρησιμοποιηθούν με τον τρόπο που περιγράψαμε, να αξιολογήσουν σύνολα δεδομένων με γονίδια δείκτες και κατ' επέκταση να τα συγκρίνουν μεταξύ τους. Τόσο οι αλγόριθμοι ομαδοποίησης όσο και οι δείκτες εγκυρότητας μετρούν το ίδιο πράγμα με διαφορετικό τρόπο, την ποιότητα των γονιδίων. Για να μετρήσουμε λοιπόν την ποιότητα με όσο το δυνατόν καλύτερο τρόπο, θα χρησιμοποιήσουμε ένα συνδυασμό από αλγόριθμους ομαδοποίησης και δείκτες εγκυρότητας. Πιο συγκεκριμένα, θα χρησιμοποιήσουμε τρεις μεθόδους ομαδοποίησης και έξι δείκτες εγκυρότητας. Από τα πειραματικά αποτελέσματα παρατηρήσαμε ότι οι δείκτες εγκυρότητας παρουσιάζουν σημαντικές διακυμάνσεις στα αποτελέσματα που δίνουν γι' αυτό κι αποφασίσαμε να χρησιμοποιήσουμε έξι απ' αυτούς ούτως ώστε τα αποτελέσματα να είναι όσο το δυνατόν πιο αξιόπιστα. Αντίθετα οι αλγόριθμοι ομαδοποίησης δεν παρουσιάζουν μεγάλες διακυμάνσεις στα αποτελέσματα γι' αυτό και αρκεστήκαμε σε τρεις. Τα σύνολα δεδομένων που θα αξιολογηθούν και στη συνέχεια θα συγκριθούν μεταξύ τους είναι οκτώ και έχουν όλα

προκύψει από ένα μεγαλύτερο σύνολο δεδομένων με διάφορες διαδικασίες επιλογής γονιδίων.

7.3 Τα Σύνολα Δεδομένων Που Χρησιμοποιήθηκαν

Όλα τα σύνολα δεδομένων που έχουν χρησιμοποιηθεί σ' αυτή την εργασία αποτελούν τμήματα του συνόλου δεδομένων της Van't Veer [21]. Το σύνολο αυτό αποτελείται από 25000 γονίδια και 78 δείγματα ιστών από ισάριθμους ασθενείς με καρκίνο του στήθους. Οι 34 απ' αυτούς παρουσίασαν μετάσταση μέσα σε 5 χρόνια μετά από τη θεραπεία τους ενώ οι υπόλοιποι 44 δεν παρουσίασαν μετάσταση για τουλάχιστον 5 χρόνια. Για να υπολογιστεί ο πίνακας γονιδιακής έκφρασης, απομονώθηκε από κάθε ασθενή ίδια ποσότητα RNA και χρησιμοποιήθηκε για την παρασκευή συμπληρωματικού RNA (cRNA). Στην συνέχεια χρησιμοποιήθηκε η μέθοδος DNA Microarrays και στα αποτελέσματα που προέκυψαν εφαρμόστηκαν δύο μέθοδοι κανονικοποίησης: Η λογαριθμική τροποποίηση και η ρύθμιση των μέσων όρων και των τυπικών αποκλίσεων σε γονίδια και δείγματα. Και τις δύο μεθόδους τις έχουμε περιγράψει αναλυτικά στο κεφάλαιο με τις μεθόδους κανονικοποίησης.

Τα εφτά από τα οκτώ σύνολα δεδομένων που χρησιμοποιούνται στην εργασία έχουν προκύψει από διάφορες μεθόδους επιλογής γονιδίων. Το ένα απ' αυτά αποτελείται από 67 γονίδια ενώ τα υπόλοιπα 6 αποτελούνται από 64. Το όγδοο σύνολο δεδομένων αποτελείται από 4958 γονίδια και έχει προκύψει από τη μέθοδο του φιλτραρίσματος την οποία επίσης έχουμε περιγράψει στο κεφάλαιο με τις μεθόδους κανονικοποίησης.

Τα εφτά σύνολα δεδομένων τα οποία προέκυψαν με διαδικασίες επιλογής γονιδίων είναι τα εξής: RFE , NNW, LK, RBF, 2 DK, 4 DK, Van't Veer Markers. Με εξαίρεση το Van't Veer Markers όλα τα υπόλοιπα έχουν προκύψει από δουλειά που έχει γίνει στο εργαστήριο μας [23]. Το RFE έχει προκύψει από την μέθοδο επιλογής γονιδίων SVM Recursive Feature Elimination [24] ενώ τα υπόλοιπα αποτελούν παραλλαγές αυτής της μεθόδου. Το Van't veer markers είναι ένα σύνολο δεδομένων το οποίο έχει προκύψει από μία διαφορετική μέθοδο επιλογής γονιδίων που εφάρμοσε η Van't Veer [21].

Ο σκοπός αυτής της εργασίας δεν είναι να αξιολογήσουμε κάποιες συγκεκριμένες διαδικασίες επιλογής γονιδίων, αλλά να προτείνουμε μία γενική μέθοδο αξιολόγησης αυτών των διαδικασιών. Γι' αυτό ακριβώς το λόγο, θα χρησιμοποιήσουμε τα σύνολα δεδομένων που περιγράψαμε παραπάνω απλώς για να εφαρμόσουμε τη μέθοδο που προτείνουμε, χωρίς να αναλύσουμε περαιτέρω τις διαδικασίες οι οποίες χρησιμοποιήθηκαν για να εξαχθούν τα γονίδια δείκτες.

Το όγδοο σύνολο δεδομένων (5000 significant genes) αποτελείται από πολύ περισσότερα γονίδια (5000 περίπου) και έχει προκύψει, όπως προαναφέραμε, από μία διαδικασία φιλτραρίσματος του αρχικού συνόλου δεδομένων. Το σύνολο αυτό χρησιμοποιείται στην εργασία σαν ένα μέτρο σύγκρισης για τα υπόλοιπα σύνολα δεδομένων. Τα υπόλοιπα σύνολα, τα οποία είναι αρκετά μικρότερα και περιέχουν πιο ενδεικτικά γονίδια, έχουν όπως είναι λογικό αρκετά καλύτερη ποιότητα. Αν επαληθευτεί κάτι τέτοιο από την εφαρμογή της μεθόδου που προτείνουμε, αν δηλαδή όλα τα μικρότερα σύνολα δεδομένων αποδειχτούν καλύτερα όσον αφορά την ποιότητα από το μεγαλύτερο, τότε θα έχουμε μια ένδειξη για την αξιοπιστία της μεθόδου.

Το σύνολο δεδομένων 5000 significant genes αποτελεί στην ουσία ένα πρώτο βήμα πριν τη διαδικασία επιλογής γονιδίων που εφαρμόζει η Van't Veer. Πριν ξεκινήσει η διαδικασία επιλογής γονιδίων γίνεται ένα πρώτο φιλτράρισμα για να απομακρυνθούν τα γονίδια τα οποία δεν μας ενδιαφέρουν. Τα κριτήρια σύμφωνα με τα οποία γίνεται το φιλτράρισμα είναι:

α) Διπλάσια διαφορά (two fold difference). Αυτό το κριτήριο ικανοποιείται όταν για όλες τις τιμές ενός γονιδίου, ο λόγος του επιπέδου έκφρασης του γονιδίου στο δείγμα μελέτης (Cy5) προς το επίπεδο έκφρασης του στο δείγμα αναφοράς (Cy3) είναι είτε μεγαλύτερος του 2 είτε μικρότερος από ½. Θα πρέπει λοιπόν να ισχύει:

$Cy5/Cy3 > 2$ ή $Cy5/Cy3 < ½$. $\Leftrightarrow$ ratio >2 ή ratio $< ½ \Leftrightarrow$ $\log10(ratio) > 0.3$ ή $\log10(ratio) < -0.3$

Άρα τα γονίδια που θα παραμείνουν μετά το φιλτράρισμα είναι αυτά που παίρνουν τιμές μεγαλύτερες από 0,3 ή μικρότερες από -0,3 στον πίνακα γονιδιακής έκφρασης. Το κριτήριο αυτό μας εξασφαλίζει ότι στα γονίδια που απομένουν, δεν υπάρχουν γονίδια τα οποία εκφράζονται περίπου το ίδιο στα δείγματα μελέτης και αναφοράς

και παίρνουν τιμές κοντά στο μηδέν. Τα γονίδια που παρουσιάζουν αυτή τη συμπεριφορά αποτελούν στην ουσία θόρυβο σε ένα σύνολο δεδομένων και δυσκολεύουν οποιαδήποτε προσπάθεια για ανάλυση των δεδομένων, είτε αυτή έχει σαν στόχο την ομαδοποίηση είτε την επιλογή γονιδίων.

β) P- Value μικρότερο από 0,01 σε τουλάχιστον 6 περιπτώσεις ασθενών. Το P-value είναι μία τιμή η οποία μας παρέχεται στον πίνακα γονιδιακής έκφρασης της Van't Veer μαζί με τον λογάριθμο του λόγου των εκφράσεων. Όσο πιο μικρή είναι η τιμή του P-value τόσο περισσότερο ο λόγος των εκφράσεων απέχει από το 1. Θέτοντας λοιπόν αυτό το κριτήριο, εξασφαλίζουμε ότι το γονίδιο εκφράζεται αρκετά διαφορετικά στα δείγματα μελέτης απ' ότι στα δείγματα αναφοράς (υπερβάλουσα ή χαμηλή έκφραση) σε τουλάχιστον έξι από τους ασθενείς.

7.4 Οι Μέθοδοι Που Χρησιμοποιήθηκαν

Σ' αυτήν εργασία έχουμε χρησιμοποιήσει τρεις μεθόδους ομαδοποίησης και έξι δείκτες εγκυρότητας. Οι μέθοδοι ομαδοποίησης είναι οι: α) Hierarchical Clustering β) K-Means και γ) Self-Organising Maps. Οι δείκτες εγκυρότητας είναι οι: α) C-Index β) Dunn's Index γ) Davies-Bouldin Index δ) Silhouette Index ε) Goodman-Kruskal Index και στ) Isolation Index. Όλοι οι αλγόριθμοι ομαδοποίησης εκτελέστηκαν με το πρόγραμμα TIGR MultiExperiment Viewer (MeV) ενώ για τους δείκτες εγκυρότητας χρησιμοποιήθηκε το πρόγραμμα Machaon Clustering Validation Environment (Machaon CVE).

Οι παράμετροι στις μεθόδους που χρησιμοποιήσαμε καθορίστηκαν ως εξής:

Μέθοδοι Ομαδοποίησης:

α) Hierarchical Clustering

- Επιλέξαμε ομαδοποίηση στα δείγματα (στους ασθενείς).
- Χρησιμοποιήσαμε ως μέτρο ομοιότητας τον συντελεστή συσχέτισης του Pearson.
- Χρησιμοποιήσαμε ως μέθοδο υπολογισμού των αποστάσεων μεταξύ των ομάδων τη μέθοδο διασύνδεσης complete linkage.

β) K-Means

- Επιλέξαμε ομαδοποίηση στα δείγματα (στους ασθενείς).
- Επιλέξαμε να δημιουργηθούν δύο ομάδες.
- Χρησιμοποιήσαμε ως μέτρο ομοιότητας τον συντελεστή συσχέτισης του Pearson (Pearson correlation coefficient) σε όλα τα σύνολα δεδομένων εκτός από το NNWs στο οποίο χρησιμοποιήθηκε ο μέσος όρος του εσωτερικού γινομένου.
- Ορίσαμε τον αριθμό 50 ως τον μέγιστο αριθμό επαναλήψεων του αλγορίθμου.

γ) Self-Organising Maps

- Επιλέξαμε ομαδοποίηση στα δείγματα (στους ασθενείς).
- Δώσαμε τις τιμές 1 και 2 στις διαστάσεις του χάρτη X και Y ούτως ώστε να δημιουργηθούν δύο ομάδες.
- Επιλέξαμε το σχήμα εξάγωνο για το σχήμα του χάρτη.
- Χρησιμοποιήσαμε ως μέτρο ομοιότητας τον συντελεστή συσχέτισης του Pearson σε όλα τα σύνολα δεδομένων εκτός από το NNWs στο οποίο χρησιμοποιήθηκε ο μέσος όρος του εσωτερικού γινομένου (averaged dot product).
- Ορίσαμε τον αριθμό 2000 ως τον μέγιστο αριθμό επαναλήψεων του αλγορίθμου.
- Δώσαμε στην συνάρτηση learning rate αρχική τιμή 0,05.
- Επιλέξαμε τη συνάρτηση guassian ως συνάρτηση γειτονιάς (neighborhood function) και τον αριθμό 3 ως αρχική τιμή για την ακτίνα της γειτονιάς (neighborhood radius).

Δείκτες Εγκυρότητας:

α) C-index

- Επιλέξαμε ως μέτρο ομοιότητας την ευκλείδια απόσταση.

β) Dunn’s index

- Επιλέξαμε ως μέτρο ομοιότητας την ευκλείδια απόσταση.
- Χρησιμοποιήσαμε ως μέθοδο υπολογισμού των αποστάσεων μεταξύ των ομάδων τη μέθοδο διασύνδεσης complete linkage.

- Ορίσαμε ως μέθοδο υπολογισμού της εσωτερικής απόστασης των ομάδων τη μέθοδο complete diameter.

γ) Davies – Bouldin index
- Επιλέξαμε ως μέτρο ομοιότητας την ευκλείδια απόσταση.
- Χρησιμοποιήσαμε ως μέθοδο υπολογισμού των αποστάσεων μεταξύ των ομάδων τη μέθοδο διασύνδεσης complete linkage.
- Ορίσαμε ως μέθοδο υπολογισμού της εσωτερικής απόστασης των ομάδων τη μέθοδο complete diameter.

δ) Silhouette index
- Επιλέξαμε ως μέτρο ομοιότητας την ευκλείδια απόσταση.

ε) Goodman-Kruskal index
- Επιλέξαμε ως μέτρο ομοιότητας την ευκλείδια απόσταση.

στ) Isolation index
- Επιλέξαμε ως μέτρο ομοιότητας την ευκλείδια απόσταση.
- Δώσαμε στη σταθερά neighborhood size τη τιμή 0,1. Αυτό σημαίνει ότι η σταθερά κ που καθορίζει τον αριθμό των πλησιέστερων γειτόνων ισούται με το 10% των αντικειμένων που αποτελούν το σύνολο δεδομένων.

Στους αλγόριθμους ομαδοποίησης K-Means και Self-Organising Maps επιλέξαμε να δημιουργηθούν δύο ομάδες, ούτως ώστε να δούμε κατά πόσο μπορούν να διαχωρίσουν την ομάδα των ασθενών που πέθαναν μέσα σε 5 χρόνια από την ομάδα των ασθενών που επέζησαν για τουλάχιστον 5 χρόνια. Στις περισσότερες από τις μετρήσεις μας έχουμε χρησιμοποιήσει το συντελεστή συσχέτισης του Pearson αφού μετά από αρκετές δοκιμές παρατηρήσαμε ότι δίνει καλύτερα αποτελέσματα σε σχέση με τα υπόλοιπα μέτρα ομοιότητας. Ο μέσος όρος του εσωτερικού γινομένου (averaged dot product) χρησιμοποιήθηκε μόνο για το σύνολο δεδομένων NNWs στους αλγορίθμους K-Means και Self-Organising Maps. Αυτό έγινε γιατί ο συντελεστής συσχέτισης του Pearson δίνει, χωρίς κάποιον προφανή λόγο, πολύ χαμηλά αποτελέσματα (πολύ χαμηλότερα από τα άλλα μέτρα ομοιότητας) στις συγκεκριμένες μετρήσεις.

Στον αλγόριθμο Hierarchical Clustering, επιλέξαμε για την μέτρηση των αποστάσεων μεταξύ των ομάδων τη μέθοδο διασύνδεσης complete linkage παρόλο που η μέθοδος average linkage θεωρείται γενικώς καλύτερη. Ο λόγος που μας οδήγησε σ' αυτήν την επιλογή είναι ότι η μέθοδος average linkage δίνει ως αποτέλεσμα, στις περισσότερες από τις μετρήσεις, μία πολύ μεγάλη ομάδα με 66-76 ασθενείς και μία πολύ μικρότερη με 2-10 ασθενείς. Αντίθετα, η μέθοδος complete linkage δίνει σε όλες τις περιπτώσεις ομάδες με συγκρίσιμα μεγέθη. Με δεδομένο το γεγονός ότι ο αλγόριθμος θα πρέπει, στην ιδανική περίπτωση, να εντοπίσει τις δύο ομάδες των 44 και των 34 ασθενών, είναι φανερό ότι όταν οι ομάδες έχουν τόσο διαφορετικά μεγέθη τα ποσοστά επιτυχίας των αλγορίθμων είναι πολύ χαμηλά. Η μέθοδος single linkage απορρίφθηκε εξαρχής καθώς σε όλες τις μετρήσεις παρουσιάστηκε το φαινόμενο chaining το οποίο έχουμε περιγράψει σε προηγούμενο κεφάλαιο.

Στους δείκτες εγκυρότητας παρατηρήσαμε ότι οι τιμές σε όλα τα σύνολα δεδομένων βελτιώνονται ή χειροτερεύουν περίπου στον ίδιο βαθμό όταν αλλάζει το μέτρο ομοιότητας που χρησιμοποιούμε, με αποτέλεσμα η τελική κατάταξη των συνόλων με βάση την ποιότητα να παραμένει η ίδια. Για το λόγο αυτό χρησιμοποιήσαμε την ευκλείδια απόσταση, η οποία αποτελεί και την προεπιλογή, σε όλους τους δείκτες εγκυρότητας. Για τον ίδιο ακριβώς λόγο, στους δείκτες Dunn's και Davies - Bouldin, στους οποίους απαιτείται καθορισμός της μεθόδου διασύνδεσης των ομάδων και της μεθόδου υπολογισμού της εσωτερικής τους απόστασης, επιλέξαμε τις προεπιλογές complete linkage και complete diameter.

Κεφάλαιο 8 - Αποτελέσματα Και Συμπεράσματα

8.1 Εισαγωγή

Σ' αυτό το τελευταίο κεφάλαιο θα δώσουμε τα αποτελέσματα που προέκυψαν από τις μετρήσεις μας και στη συνέχεια θα αναπτύξουμε τα συμπεράσματα που προκύπτουν από την εφαρμογή της μεθόδου που χρησιμοποιήσαμε. Στα αποτελέσματα περιλαμβάνονται τα ποσοστά επιτυχίας των αλγορίθμων ομαδοποίησης καθώς και οι τιμές των δεικτών εγκυρότητας για κάθε ένα από τα σύνολα δεδομένων. Στη συνέχεια, τα σύνολα δεδομένων συγκρίνονται με βάση τα παραπάνω αποτελέσματα χρησιμοποιώντας την τεχνική της σταθμισμένης ψηφοφορίας (weighed voting strategy).

8.2 Αποτελέσματα

Τα αποτελέσματα της εργασίας δίνονται, όπως θα δούμε παρακάτω, σε μορφή πινάκων. Ο πρώτος πίνακας περιέχει τα ποσοστά επιτυχίας των αλγορίθμων ομαδοποίησης για κάθε ένα από τα σύνολα δεδομένων. Στις γραμμές του πίνακα είναι τα σύνολα δεδομένων και στις στήλες οι αλγόριθμοι ομαδοποίησης που χρησιμοποιήσαμε. Για να κάνουμε πιο κατανοητό τι εννοούμε με τον όρο ποσοστό επιτυχίας θα αναφερθούμε σε ένα συγκεκριμένο παράδειγμα. Η εφαρμογή του αλγορίθμου K-Means στο σύνολο δεδομένων 4DK δίνει ως αποτέλεσμα μία ομάδα με 44 ασθενείς και μία με 34. Στην ομάδα με τους 44 ασθενείς, οι 39 ανήκουν σ' αυτούς που δεν παρουσίασαν μετάσταση για τουλάχιστον 5 χρόνια και οι υπόλοιποι 5 σ' αυτούς που παρουσίασαν μετάσταση. Παρατηρούμε λοιπόν, ότι ο αλγόριθμος εντοπίζει την ομάδα των ασθενών που δεν παρουσίασαν μετάσταση με ποσοστό επιτυχίας 39/44 = 88,6 %. Στην ομάδα με τους 34 ασθενείς, οι 29 ανήκουν σ' αυτούς που παρουσίασαν μετάσταση μέσα σε 5 χρόνια ενώ οι υπόλοιποι 5 ανήκουν σ' αυτούς που δεν παρουσίασαν μετάσταση στο ίδιο χρονικό διάστημα. Άρα ο αλγόριθμος εντοπίζει την ομάδα των ασθενών που παρουσίασαν μετάσταση με ποσοστό επιτυχίας 29/34 = 85,3 %. Το συνολικό ποσοστό επιτυχίας του αλγορίθμου K-Means για το σύνολο δεδομένων 4DK είναι ο μέσος όρος των ποσοστών επιτυχίας

για κάθε ομάδα.. Ο μέσος όρος με στρογγυλοποίηση ενός δεκαδικού ψηφίου είναι 87,0 %. Με τον ίδιο τρόπο υπολογίζονται και οι υπόλοιπες τιμές του πίνακα.

Στον δεύτερο πίνακα περιέχονται οι τιμές που παίρνουν οι δείκτες εγκυρότητας για κάθε ένα από τα σύνολα δεδομένων. Στις γραμμές του πίνακα είναι τα σύνολα δεδομένων και στις στήλες οι δείκτες εγκυρότητας που χρησιμοποιήσαμε.

Στον τρίτο πίνακα, τέλος, φαίνονται τα αποτελέσματα της τεχνικής της σταθμισμένης ψηφοφορίας (weighed voting strategy) που χρησιμοποιήθηκε για να συγκρίνουμε τα σύνολα δεδομένων μεταξύ τους [18,22]. Η τεχνική αυτή μας δίνει την δυνατότητα να συνδυάσουμε τα αποτελέσματα από τους αλγόριθμους ομαδοποίησης και τους δείκτες εγκυρότητας, θεωρώντας τα ως κριτήρια ίσης βαρύτητας για την αξιολόγηση της ποιότητας ενός συνόλου δεδομένων. Πιο συγκεκριμένα, δίνουμε σε όλα τα σύνολα δεδομένων ένα βαθμό από το 1 μέχρι το 8 για κάθε αλγόριθμο ομαδοποίησης και για κάθε δείκτη εγκυρότητας. Ο βαθμός 8 δίνεται στο σύνολο δεδομένων που παίρνει το μεγαλύτερο ποσοστό επιτυχίας του αλγορίθμου ομαδοποίησης ή την καλύτερη τιμή του δείκτη εγκυρότητας. Συνεχίζοντας, δίνουμε την τιμή 7 στο δεύτερο κατά σειρά καλύτερο σύνολο δεδομένων, την τιμή 6 στο τρίτο κατά σειρά καλύτερο κ.ο.κ. Το σύνολο δεδομένων που παίρνει το χαμηλότερο ποσοστό επιτυχίας του αλγορίθμου ή τη χειρότερη τιμή του δείκτη εγκυρότητας παίρνει 1 βαθμό. Στο τέλος υπολογίζεται, για κάθε σύνολο δεδομένων, ο μέσος όρος των βαθμών από τους αλγόριθμους, ο μέσος όρος των βαθμών από τους δείκτες και ο τελικός μέσος όρος ο οποίος προκύπτει από τους άλλους δύο και χρησιμοποιείται για να συγκρίνουμε τα σύνολα δεδομένων μεταξύ τους. Στις γραμμές του πίνακα είναι τοποθετημένοι κατά σειρά οι αλγόριθμοι ομαδοποίησης, ο μέσος όρος για τους αλγόριθμους ομαδοποίησης, οι δείκτες εγκυρότητας, ο μέσος όρος για τους δείκτες εγκυρότητας και ο τελικός μέσος όρος της βαθμολογίας. Στις στήλες βρίσκονται τα σύνολα δεδομένων.

Οι τρεις πίνακες με τα αποτελέσματα φαίνονται παρακάτω:

	HCL	K-Means	SOMs
5000 Significant Genes	65%	64%	64,9%
2DK	74,6%	83,4%	84,8%
4DK	83,4%	87,0%	80,7%
LK	79,1%	76,0%	79,4%
NNWs	93,1%	83,7%	83,7%
RBF	73,3%	75,3%	78,2%
RFE	75,3%	74,2%	73,1%
Van't Veer Markers	77,1%	78%	79,7%

Πίνακας 1. Ο πίνακας περιέχει τα ποσοστά επιτυχίας των αλγορίθμων ομαδοποίησης για κάθε ένα από τα σύνολα δεδομένων. Στις γραμμές του πίνακα είναι τα σύνολα δεδομένων και στις στήλες οι αλγόριθμοι ομαδοποίησης που χρησιμοποιήσαμε.

	C-index	Davies Boulding index	Dunn's index	Goodman-Kruskal index	Silhouettes Index	Isolation Index
5000 Significant Genes	0,449	1,606	0,995	0,098	0,024	0,545
2DK	0,339	1,797	0,837	0,331	0,122	0,774
4DK	0,305	1,677	0,952	0,366	0,136	0,765
LK	0.368	1,674	0,939	0,269	0,116	0,716
NNWs	0,43	1,930	0,848	0,129	0,076	0,780
RBF	0,331	1,687	1,086	0,292	0,129	0,686
RFE	0,448	1,869	0,999	0,083	0,041	0,670
Van't Veer Markers	0,340	1,779	0,915	0,312	0,121	0,671

Πίνακας 2. Ο πίνακας περιέχει τις τιμές που παίρνουν οι δείκτες εγκυρότητας για κάθε ένα από τα σύνολα δεδομένων. Στις γραμμές του πίνακα είναι τα σύνολα δεδομένων και στις στήλες οι δείκτες εγκυρότητας που χρησιμοποιήσαμε.

	5000 Significant Genes	2DK	4DK	LK	NNWs	RBF	RFE	Van't Veer Markers
HCL	1	3	7	6	8	2	4	5
K-Means	1	6	8	4	7	3	2	5
SOMs	1	8	6	4	7	3	2	5
Average 1	1	5,66	7	4,66	7,33	2,66	2,66	5
C-index	1	6	8	4	3	7	2	5
Davies-Bouldin index	8	3	6	7	1	5	2	4
Dunn's index	6	1	5	4	2	8	7	3
Goodman-Kruskal index	2	7	8	4	3	5	1	6
Silhouettes index	1	6	8	4	3	7	2	5
Isolation index	1	7	6	5	8	4	2	3
Average 2	3,16	5	6,83	4,66	3,33	6	2,66	4,33
Total Average	2.08	5,33	6,92	4,66	5,33	4,33	2,66	4,67

Πίνακας 3. Στο πίνακα φαίνονται οι βαθμοί που παίρνουν τα σύνολα δεδομένων από κάθε αλγόριθμο και κάθε δείκτη εγκυρότητας καθώς και οι μέσοι όροι της βαθμολογίας. Στις γραμμές του πίνακα είναι τοποθετημένοι κατά σειρά οι αλγόριθμοι ομαδοποίησης, ο μέσος όρος για τους αλγόριθμους ομαδοποίησης (average 1), οι δείκτες εγκυρότητας, ο μέσος όρος για τους δείκτες εγκυρότητας (average 2) και ο τελικός μέσος όρος της βαθμολογίας. Στις στήλες βρίσκονται τα σύνολα δεδομένων.

8.3 Συμπεράσματα

Στον πίνακα 1 βλέπουμε ότι οι τιμές που δίνουν οι αλγόριθμοι ομαδοποίησης σε κάθε ένα από τα σύνολα δεδομένων δεν παρουσιάζουν μεγάλες διακυμάνσεις μεταξύ τους. Τα καλύτερα ποσοστά επιτυχίας εμφανίζονται στο σύνολο δεδομένων NNWs ενώ πολύ καλά ποσοστά παίρνουν και τα σύνολα 4DK και 2DK. Τα σύνολα LK, RBF, RFE, και Van't Veer Markers εμφανίζουν πιο μέτριες επιδόσεις ενώ το σύνολο με τα 5000 γονίδια παρουσιάζει, όπως αναμενόταν, τα χαμηλότερα ποσοστά ακρίβειας και στους τρεις αλγορίθμους ομαδοποίησης.

Στον πίνακα 2 παρατηρούμε ότι οι διακυμάνσεις στα αποτελέσματα που δίνουν οι δείκτες εγκυρότητας είναι πολύ μεγαλύτερες. Ενδεικτική περίπτωση αποτελεί το σύνολο με τα 5000 γονίδια, το οποίο εμφανίζεται καλύτερο απ' όλα με βάση το δείκτη Davies – Bouldin, τρίτο κατά σειρά καλύτερο με βάση το δείκτη του Dunn και χειρότερο απ' όλα με βάση τους δείκτες C-index, Silhouettes και Isolation. Αυτές οι σημαντικές διακυμάνσεις που παρατηρήσαμε ήταν η αιτία που χρησιμοποιήσαμε αρκετούς δείκτες εγκυρότητας ούτως ώστε τα αποτελέσματα να είναι όσο το δυνατόν πιο αξιόπιστα. Το σύνολο δεδομένων που εμφανίζει τις καλύτερες τιμές στους δείκτες εγκυρότητας είναι το 4DK ενώ αρκετά καλές τιμές παίρνουν και τα σύνολα RBF και 2DK. Τα σύνολα LK, NNWs, RFE, Van't Veer Markers παρουσιάζουν πιο μέτριες επιδόσεις ενώ το σύνολο με τα 5000 γονίδια, παρόλο που εμφανίζει πολύ καλές τιμές στους δείκτες Davies-Bouldin και Dunn, έχει συνολικά τη χειρότερη επίδοση.

Τέλος, στον πίνακα 3, οι συγκρίσεις με βάση τους αλγόριθμους και τους δείκτες εγκυρότητας γίνονται πλέον πιο εύκολες με τη βοήθεια των μέσων όρων average 1 και average 2. Επιπλέον, ο συνολικός μέσος όρος μας δίνει την δυνατότητα να συγκρίνουμε τα σύνολα δεδομένων με βάση το σύνολο των κριτηρίων που χρησιμοποιήσαμε και να καταλήξουμε σε κάποια τελικά συμπεράσματα.

Όπως φαίνεται από τον πίνακα, το σύνολο δεδομένων που έχει την καλύτερη ποιότητα με βάση το σύνολο των κριτηρίων είναι το 4DK το οποίο έχει συνολικό μέσο όρο 6,92 και παίρνει πολύ καλές τιμές και στους αλγόριθμους και στους δείκτες. Στην δεύτερη θέση βρίσκονται μαζί τα σύνολα 2DK και NNWs με συνολικό μέσο όρο 5,33. Παρόλο που και τα δύο σύνολα εμφανίζουν τον ίδιο μέσο όρο, παρατηρούμε ότι το 2DK παρουσιάζει παρόμοια συμπεριφορά στους αλγόριθμους και στους δείκτες εγκυρότητας ενώ το NNWs παρουσιάζει σημαντικά καλύτερες τιμές στους αλγόριθμους σε σχέση με τους δείκτες. Πιο συγκεκριμένα, το 2DK

παίρνει καλές τιμές και στους δύο μέσους όρους (average1 = 5,66 και average2 = 5) ενώ το NNWs εμφανίζει πολύ καλό μέσο όρο στους αλγόριθμους ομαδοποίησης (average1 = 7,33) και πολύ χαμηλό μέσο όρο στους δείκτες εγκυρότητας (average2 = 2,33). Μπορούμε επίσης να παρατηρήσουμε ότι εκτός από τους αλγόριθμους ομαδοποίησης το σύνολο NNWs παίρνει πολύ καλή τιμή στο δείκτη isolation, μέτριες τιμές στους δείκτες C-index και Goodman–Kruskal και πολύ χαμηλές τιμές τους δείκτες Davies-Bouldin και Dunn. Ο δείκτης isolation, όπως έχουμε ήδη προαναφέρει, παρουσιάζει μία ιδιαιτερότητα σε σχέση με τους υπόλοιπους δείκτες. Εξετάζει μόνο τη διαχωρισιμότητα των ομάδων και όχι το βαθμό πυκνότητας. Ακόμα και αν οι ομάδες δεν είναι καθόλου συμπαγείς και αποτελούν στην ουσία συγχωνεύσεις μικρότερων ομάδων ο δείκτης θα δώσει πολύ καλά αποτελέσματα αρκεί οι ομάδες να είναι καλά διαχωρισμένες. Οι δείκτες C-index και Goodman-Kruskal παρουσιάζουν έχουν ένα κοινό χαρακτηριστικό. Δεν χρησιμοποιούν τιμές εσωτερικής απόστασης (είτε με τη μορφή της εσωτερικής απόστασης μιας ολόκληρης ομάδας είτε με τη μορφή της απόστασης μεταξύ συγκεκριμένων αντικειμένων της ίδιας ομάδας) στον υπολογισμό του τελικού αποτελέσματος. Οι τιμές εσωτερικές απόστασης υπολογίζονται απλώς για να συγκριθούν με τιμές εξωτερικής απόστασης. Αν οι τιμές εσωτερικής απόστασης είναι μικρότερες σε σχέση με τις αντίστοιχες τιμές εξωτερικής απόστασης, τότε οι δείκτες αυτοί παίρνουν καλές τιμές, χωρίς να εξετάζουν αν οι τιμές εσωτερικής απόστασης είναι αρκετά μικρές. Με λίγα λόγια, οι καλές τιμές σ’ αυτούς τους δείκτες, αποδεικνύουν ότι οι ομάδες είναι καλά διαχωρισμένες χωρίς να εξετάζουν αν είναι και αρκετά συμπαγείς. Ο βαθμός πυκνότητας δε παίζει καθοριστικό ρόλο στο τελικό αποτέλεσμα. Αντίθετα στους δείκτες Dunn και Davies–Bouldin οι τιμές της εσωτερικής απόστασης των ομάδων χρησιμοποιούνται στο τελικό αποτέλεσμα αφού αποτελούν στη μεν περίπτωση του Dunn τον αριθμητή του κλάσματος, στη δε περίπτωση του Davies–Bouldin τον παρονομαστή του κλάσματος. Κατά συνέπεια, στους δείκτες αυτούς, ο βαθμός πυκνότητας των ομάδων έχει πιο σημαντικό ρόλο αφού οι τιμές εσωτερικής απόστασης δεν χρησιμοποιούνται απλώς για να συγκριθούν με τις τιμές εξωτερικής απόστασης αλλά συνδιαμορφώνουν μαζί τους το τελικό αποτέλεσμα Οι αλγόριθμοι ομαδοποίησης, τέλος, δεν χρησιμοποιούν τιμές εσωτερικής απόστασης των ομάδων. Απλώς συγκρίνουν τις αποστάσεις κάθε αντικειμένου με όλες τις ομάδες και τοποθετούν το αντικείμενο στην πιο κοντινή απ’ αυτές. Κατά συνέπεια, τα μεγάλα ποσοστά επιτυχίας των αλγορίθμων ομαδοποίησης αποδεικνύουν απλώς ότι οι ομάδες

είναι εύκολα διαχωρίσιμες και όχι ότι είναι κατ' ανάγκη αρκετά συμπαγείς. Συμπερασματικά μπορούμε να πούμε λοιπόν, ότι το σύνολο NNWs παίρνει πολύ καλές τιμές στα κριτήρια εκείνα τα οποία δεν εξετάζουν αν οι ομάδες είναι αρκετά συμπαγείς (αλγόριθμοι ομαδοποίησης και δείκτης isolation), παίρνει πολύ χαμηλές τιμές στα κριτήρια στα οποία ο βαθμός πυκνότητας παίζει σημαντικό ρόλο (δείκτες Davies-Bouldin, Dunn), και μέτριες τιμές στα κριτήρια τα οποία προϋποθέτουν ότι οι ομάδες είναι σε κάποιο βαθμό συμπαγείς αλλά δεν αποδεικνύουν αρκετά μεγάλο βαθμό πυκνότητας (δείκτες C-index και Goodman-Kruskal). Το γεγονός αυτό μας οδηγεί στο συμπέρασμα ότι τα γονίδια στο σύνολο NNWs είναι μεν αρκετά ενδεικτικά ούτως ώστε να καθιστούν τις δύο ομάδες ασθενών που μας ενδιαφέρουν διαχωρίσιμες αλλά όχι τόσο πολύ ούτως ώστε να τις καθιστούν αρκετά συμπαγείς. Έχουμε ήδη προαναφέρει στο προηγούμενο κεφάλαιο ότι η ποιότητα μιας επιλογής γονιδίων θεωρείται καλή όταν τα γονίδια παρουσιάζουν παραπλήσια συμπεριφορά σε κάθε μία από τις ομάδες που εξετάζουμε, ενώ η συμπεριφορά τους μεταξύ των ομάδων διαφοροποιείται σημαντικά. Η ποιότητα εξαρτάται λοιπόν τόσο από τη διαχωρισιμότητα όσο και από την πυκνότητα των ομάδων. Όλα τα παραπάνω μας οδηγούν στο συμπέρασμα ότι παρόλο που οι τελικές τιμές της ποιότητας στα σύνολα NNWs και 2DK είναι ίσες, το σύνολο NNWs περιέχει γονίδια τα οποία καθιστούν τις ομάδες που μας ενδιαφέρουν περισσότερο διαχωρίσιμες και λιγότερο συμπαγείς σε σχέση με το σύνολο 2DK.

Παρατηρούμε λοιπόν ότι εκτός από το τελικό μέσο όρο ο οποίος προσφέρει μία τελική κατάταξη των συνόλων δεδομένων και τα υπόλοιπα κομμάτια του πίνακα μπορούν να χρησιμοποιηθούν για να προκύψουν κάποια χρήσιμα συμπεράσματα. Αυτό μπορεί να γίνει, όπως είδαμε, κατανοώντας με ποιο τρόπο μετράει το καθένα από τα κριτήρια που χρησιμοποιήσαμε τη ποιότητα ενός συνόλου δεδομένων και ερμηνεύοντας ανάλογα τις τιμές που μας δίνουν.

Συνεχίζοντας την παρατήρηση των τελικών μέσων όρων διαπιστώνουμε ότι το τέταρτο καλύτερο σύνολο δεδομένων είναι το σύνολο LK το οποίο έχει ακριβώς την ίδια επίδοση (4,66) και στους αλγόριθμους και στους δείκτες. Στη συνέχεια ακολουθεί το σύνολο Van't Veer Markers το οποίο εμφανίζει ανάλογη σταθερότητα παίρνοντας μέτριες τιμές σε όλα σχεδόν τα κριτήρια.

Στην έκτη θέση βρίσκεται το σύνολο RBF το οποίο παίρνει μέτριες και καλές τιμές στους δείκτες εγκυρότητας και αρκετά χαμηλές τιμές στους αλγόριθμους ομαδοποίησης. Έχει δηλαδή ακριβώς αντίθετη συμπεριφορά σε σχέση με το σύνολο

NNWs. Παρατηρούμε επίσης ότι το σύνολο RBF παίρνει τη χαμηλότερη του τιμή στο δείκτη isolation σε σχέση με όλους τους υπόλοιπους δείκτες στους οποίους οι τιμές του κυμαίνονται από καλές έως πολύ καλές. Το γεγονός λοιπόν ότι το σύνολο RBF παίρνει τις χαμηλότερες τιμές του στα κριτήρια που εξετάζουν μόνο τη διαχωρισιμότητα (στους αλγόριθμους ομαδοποίησης και στο δείκτη isolation) μας οδηγεί στο συμπέρασμα ότι τα γονίδια που αποτελούν το σύνολο καθιστούν τις ομάδες αρκετά συμπαγείς αλλά όχι πολύ καλά διαχωρίσιμες.

Μετά το σύνολο RBF, στην έβδομη θέση κατά σειρά βρίσκεται το σύνολο RFE το οποίο παίρνει χαμηλές τιμές στα περισσότερα από τα κριτήρια. Τέλος, στη τελευταία θέση της κατάταξης κατατάσσεται το σύνολο 5000 significant genes το οποίο παίρνει όπως αναμενόταν πολύ χαμηλές τιμές σχεδόν σε όλα τα κριτήρια και ειδικά σε αυτά που εξετάζουν μόνο τη διαχωρισιμότητα (αλγόριθμοι ομαδοποίησης και δείκτης isolation). Μπορούμε επίσης να παρατηρήσουμε, ότι οι πολύ καλές τιμές στους δείκτες Dunn και Davies-Bouldin δεν είναι αρκετές για να αποφύγει το σύνολο αυτό τη τελευταία θέση της κατάταξης. Το γεγονός αυτό καταδεικνύει ότι η χρησιμοποίηση μεγάλου αριθμού κριτηρίων αποτελεί εμπόδιο σε μια τυχόν λανθασμένη αξιολόγηση της ποιότητας που μπορεί να προκύψει με βάση κάποια συγκεκριμένα κριτήρια.

Επίλογος

Ένας από τους πιο διαδεδομένους τρόπους ανάλυσης των δεδομένων που προκύπτουν από την εφαρμογή της μεθόδου DNA Microarrays, είναι η εφαρμογή επιβλέπουσας κατηγοριοποίησης. Στην επιβλέπουσα κατηγοριοποίηση, ένα σύνολο από προομαδοποιημένα αντικείμενα χρησιμοποιούνται για να κατασκευάσουμε μια συνάρτηση απόφασης. Η συνάρτηση αυτή μας δίνει την δυνατότητα να εντοπίσουμε την ομάδα στην οποία ανήκουν αντικείμενα, των οποίων η ομάδα μας είναι άγνωστη. Ένα από τα θέματα που έχουν προκύψει σχετικά με τη μη επιβλέπουσα κατηγοριοποίηση είναι το μεγάλο μέγεθος των δεδομένων το οποίο δημιουργεί αρκετά προβλήματα και δυσκολίες. Το γεγονός αυτό οδήγησε στην ανάγκη χρησιμοποίησης διαφόρων διαδικασιών επιλογής γονιδίων που έχουν ως στόχο τη μείωση των διαστάσεων των δεδομένων. Για να μπορέσουμε να διαπιστώσουμε αν τα αποτελέσματα αυτών των διαδικασιών είναι ικανοποιητικά θα πρέπει να χρησιμοποιήσουμε κάποια κριτήρια αξιολόγησης. Τα κριτήρια αυτά είναι η ακρίβεια (accuracy) και η ποιότητα (quality). Μέχρι τώρα οι περισσότερες μελέτες που ασχολούνται με την αξιολόγηση της επιλογής γονιδίων έχουν επικεντρωθεί σχεδόν αποκλειστικά στην ακρίβεια και ελάχιστα στην ποιότητα των αποτελεσμάτων μιας τέτοιας διαδικασίας.

Σ’ αυτήν την εργασία δείξαμε ότι οι αλγόριθμοι ομαδοποίησης και οι δείκτες εγκυρότητας, εκτός από τις συνήθεις τους χρήσεις, μπορούν, αν χρησιμοποιηθούν με τον τρόπο που περιγράψαμε, να αξιολογήσουν την ποιότητα ενός συνόλου γονιδίων και κατ’ επέκταση την ποιότητα μιας διαδικασίας επιλογής γονιδίων. Μπορούν επίσης, σε συνδυασμό με την τεχνική της σταθμισμένης ψηφοφορίας, να συγκρίνουν τα σύνολα δεδομένων ως προς την ποιότητα και να ανακαλύψουν ποιες διαδικασίες επιλογής γονιδίων δίνουν καλύτερα αποτελέσματα.

Βιβλιογραφία

[1] Francis Crick. Central Dogma of Molecular Biology. Nature. 1970, August 8, Vol. 227.

[2] Francis Crick. The Structure Of DNA. Medical Research Council Unit for the Study of Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge.1953.

[3] Καψάλης, Μπουρμπουχάκης, Περάκη, Σαλαμαστράκης. Βιολογία Γενικής Παιδείας Β' Ενιαίου Λυκείου. Έκδοση 2004.

[4] Virginie Mittard-Runte. Introduction to Microarray. 2005.

[5] Samuel D. Conzone, Carlo G. Pantano. Glass slides to DNA microarrays. 2004.

[6] A Basic Introduction to the Science Underlying NCBI Resources. http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html#microarrays

[7] M. Rattray, N. Morrison, D. Hoyleand A. Brass. DNA microarray normalisation, PCA and a related latent variable model. 2001.

[8] Alexander Sturn. Cluster Analysis for Large Scale Gene Expression Studies. Master Thesis. 2000.

[9] Sandrine Dudoit and Robert Gentleman.Classification in microarray experiments. Department of Biostatistics Harvard School of Public Health. 2002.

[10] Συσταδοποίηση δεδομένων (Data clustering). www.softlab.ece.ntua.gr/facilities/public/AD/DM/Clustering_CW.doc

[11] Calculating Distances of Vectors. http://www.ucl.ac.uk/oncology/MicroCore/HTML_resource/distances_popup.htm

[12] Microarray Tutorial http://www.ucl.ac.uk/oncology/MicroCore/MAA_Tutorial.pdf

[13] Hwangmin Ki. Microarray Data Analysis Methods Comparison : A Review. Biochemistry 218 Project.

[14] Classical Clustering Methods. http://globin.cse.psu.edu/courses/spring2002/9_cluster.pdf

[15] John Quackenbush. Computational Analysis Of Microarray Data. Nature. June 2001.

[16] Nadia Bolshakova, Francisco Azuaje and Pádraig Cunningham. An integrated tool for microarray data clustering and cluster validity assessment. Bioinformatics. 2004.

[17] Cluster validity algorithms.
https://www.cs.tcd.ie/Nadia.Bolshakova/validation_algorithms.html
[18] N. Bolshakova* and F. Azuaje. Cluster validation techniques for genome expression data. Department of Computer Science, Trinity College Dublin.
[19] Distances.
https://www.cs.tcd.ie/Nadia.Bolshakova/distances.html#INTRA
[20] Peter Rousseeuw. Silhouettes a graphical aid to the interpretation and validation of cluster analysis. Journal Of computational And Applied Mathemetics. 1987.
[21] Van't Veer et.al. Gene Expression Profiling Predicts Clinical Outcome Of Breast Cancer. Nature. 2002.
[22] N. Bolshakova, F. Azuaje.Improving Expression Data Mining Through Cluster Validation. Fourth Annual IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine, 2003.
[23] Michalis E. Blazadonakis, Michalis Zervakis. Improving the Performance of the RFE Gene Selection Method us-ing Neural Networks and non-Linear Kernels. Submitted for publication in Bioinformatics international journal.
[24] Guyon, Weston, Barnhill, Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 2002, Vol. 46.