ΠΑΝΕΠΙΣΤΗΜΙΟ ΣΤΕΡΕΑΣ ΕΛΛΑΔΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Εφαρμογές Υπολογιστικής Νοημοσύνης σε Μικροσυστοιχίες DNA

Μαλατράς Απόστολος

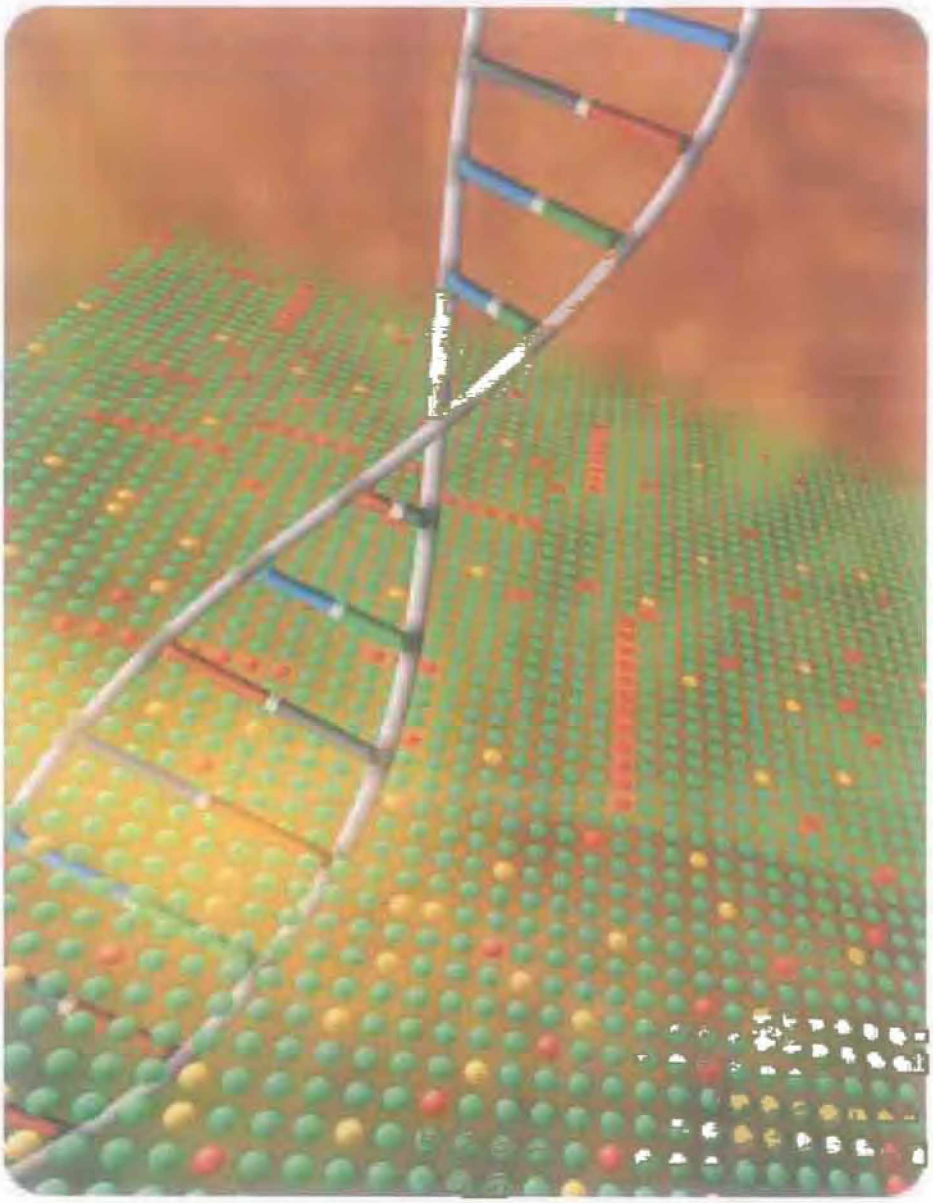
ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνος
Πλαγιανάκος Βασίλειος
Επίκουρος Καθηγητής

Λαμία, 2010

Εφαρμογές Υπολογιστικής Νοημοσύνης σε Μικροσυστοιχίες DNA

Πτυχιακή εργασία



Μαλατράς Απόστολος

Φεβρουάριος 2010

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα πτυχιακή μελέτη εκπονήθηκε από τον φοιτητή Μαλατρά Απόστολο του Τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Στερεάς Ελλάδας κατά το ακαδημαϊκό έτος 2009-2010.

Θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Πλαγιανάκο Βασίλειο για την εμπιστοσύνη που μου έδειξε δίνοντας μου τη δυνατότητα να εκπονήσω την πτυχιακή μου εργασία στον επιστημονικό τομέα που επιθυμούσα. Επίσης, θα ήθελα να τον ευχαριστήσω για τη διάθεση του να με βοηθήσει και να μου λύσει οποιαδήποτε απορία οποιαδήποτε στιγμή το χρειαζόμουν. Ακόμη θα ήθελα να ευχαριστήσω και τα άλλα δύο μέλη της Τριμελούς Επιτροπής κ. Π. Μπάγκο και κ. Ηλ. Μαγκλογιάννη για την υποστήριξή τους.

Ένα μεγάλο ευχαριστώ στους φίλους μου, οι οποίοι μου συμπαραστάθηκαν όλο αυτό τον καιρό. Τέλος, θέλω να ευχαριστήσω θερμά την οικογένειά μου για την ηθική και οικονομική συμπαράσταση όχι μόνο κατά τη διάρκεια της εκπόνησης της πτυχιακής μου εργασίας αλλά και καθ' όλη τη διάρκεια των σπουδών μου.

Περιεχόμενα

Κατάλογος σχημάτων

ABSTRACT

The growth of technology in the past few years has led to very important developments, which have influenced in important degree all the aspects of our life. One of these aspects is also that of biomedicine, the branch that tries to develop the potential of medicine via the development of technology.

Basic branch of biomedicine, are microarrays genes. Microarrays genes (otherwise known as gene chip, DNA chip or provision of genes) are a provision of microscopic points that represents unique genes and are immobilized with covalent bonds in a solid surface (usually glass). They are used for measurement DNA or use DNA for their system of detection. Quantitative or qualitative measurements with microarrays genes exploit the selective nature of the principle of complementarity between nucleic acids DNA-DNA or DNA-RNA or recently between amino-acids of proteins, under strictly controlled conditions of temperature and with the use of fluorescent substances. Microarrays genes are used today most frequently for the examination of gene expression under specific conditions and for the detection of nucleic acids of pathogenic organisms e.g. harmful viruses in test samples. This molecular biology method has the advantage that it can simultaneously examine the expression of thousands of genes, and is suitable for comparative studies of genomes. The disadvantages of this method lie in high cost and in the frequent inaccuracy of the results due to technical problems such as non-specific hybridization fluorescent pigments in the wrong genes. The use of microarrays genes today by the scientific community requires the parallel use of other complementary methods for the verification of the results of the chip such as the Northern Blot and the Quantitative Real Time (reverse transcriptase) Polymerase Chain Reaction).

On the other hand, this technology is based largely on concentration of a large volume of data, which require processing. Thus, as this branch is being developed, it is necessary to use computational intelligence techniques and various relevant algorithms, in order to achieve the best results.

Within this work, the essential characteristics of DNA microarrays will be presented. Subsequently, the main characteristics of the computational intelligence techniques will be presented and some Algorithms used for such applications, such as, the clustering algorithm K-means. Finally, with the use of suitable software, certain examples of treatment of such data will be performed.

Κεφάλαιο 1

Εισαγωγή

Η ανάπτυξη της τεχνολογίας τα τελευταία χρόνια έχει οδηγήσει σε πολύ σημαντικές εξελίξεις, οι οποίες έχουν επηρεάσει σε σημαντικό βαθμό όλες τις πτυχές της ζωής μας. Μία από αυτές τις πτυχές, είναι και αυτή της βιοϊατρικής, δηλαδή του κλάδου που προσπαθεί να αναπτύξει της δυνατότητες της ιατρικής μέσω της εξέλιξης της τεχνολογίας.

Βασικός κλάδος της βιοϊατρικής, είναι οι μικροσυστοιχίες γονιδίων. Οι μικροσυστοιχίες γονιδίων (αλλιώς γνωστές σαν γονιδιακό τσιπ, DNA chip ή διάταξη γονιδίων) είναι μία διάταξη μικροσκοπικών σημείων που αντιπροσωπεύουν μοναδικά γονίδια και ακινητοποιούνται με ομοιοπολικούς δεσμούς σε μία στερεή επιφάνεια (συνήθως γυάλινη). Χρησιμοποιούνται για τη μέτρηση DNA ή χρησιμοποιούν DNA για το σύστημα ανίχνευσής τους. Ποσοτικές ή ποιοτικές μετρήσεις με μικροσυστοιχίες γονιδίων εκμεταλλεύονται την εκλεκτική φύση της αρχής της συμπληρωματικότητας μεταξύ νουκλεϊκών οξέων DNA-DNA ή DNA-RNA ή πρόσφατα και μεταξύ των αμινοξέων των πρωτεϊνών, υπό αυστηρά ελεγχόμενες συνθήκες θερμοκρασίας και με τη χρήση φθορίζουσων ουσιών. Οι μικροσυστοιχίες γονιδίων χρησιμοποιούνται σήμερα κατά κόρον για την εξέταση της γονιδιακής έκφρασης υπό ειδικές συνθήκες και για την ανίχνευση νουκλεϊκών οξέων παθογόνων οργανισμών π.χ. επιβλαβών ιών σε δείγματα ελέγχου. Η μέθοδος μοριακής βιολογίας αυτή έχει το πλεονέκτημα ότι δύναται να εξετάζει ταυτόχρονα την έκφραση χιλιάδων γονιδίων, και ενδείκνυται για συγκριτικές μελέτες γονιδιωμάτων. Τα μειονεκτήματα της μεθόδου έγκειται στο υψηλό κόστος και στη συχνή ανακρίβεια των αποτελεσμάτων λόγω τεχνικών προβλημάτων όπως η μη ειδική υβριδοποίηση φθορίζουσων χρωστικών σε λάθος γονίδια κτλ. Η χρήση μικροσυστοιχιών γονιδίων σήμερα από την επιστημονική κοινότητα απαιτεί την παράλληλη χρήση και άλλων συμπληρωματικών μεθόδων για την επαλήθευση των αποτελεσμάτων του τσιπ όπως η Northern Blot και η Ποσοτική μέθοδος ανάστροφης τσανσκριπτάσης-αλυσιδωτή αντίδραση πολυμεράσης (Quantitative Real time (reverse transcriptase) PCR)).

Από την άλλη μεριά, η τεχνολογία αυτή βασίζεται σε μεγάλο βαθμό σε συγκέντρωση μεγάλου όγκου δεδομένων, τα οποία απαιτούν επεξεργασία. Έτσι, καθώς αναπτύσσεται αυτός ο κλάδος, κρίνεται απαραίτητο να χρησιμοποιηθούν τεχνικές υπολογιστικής νοημοσύνης και διάφοροι σχετικοί αλγόριθμοι, ώστε να είναι εφικτά τα καλύτερα αποτελέσματα.

Στα πλαίσια αυτής της εργασίας, θα παρουσιαστούν καταρχήν τα βασικά χαρακτηριστικά των μικροσυστοιχιών DNA. Εν συνεχεία, θα γίνει αναφορά στα βασικά χαρακτηριστικά των τεχνικών υπολογιστικής νοημοσύνης και κάποιων αλγορίθμων που χρησιμοποιούνται για τέτοιες εφαρμογές, όπως για παράδειγμα ο αλγόριθμος ομαδοποίησης k–means. Τέλος, με τη χρήση του κατάλληλου λογισμικού, θα παρουσιαστούν κάποια παραδείγματα επεξεργασίας τέτοιων δεδομένων.

Κεφάλαιο 2

Μικροσυστοιχίες DNA

Η τεχνολογία των μικροσυστοιχιών αναπτύχθηκε την τελευταία δεκαετία και οδήγησε, από τις μελέτες μεμονωμένων βιολογικών λειτουργιών λίγων συσχετιζόμενων γονιδίων, πρωτεϊνών ή κυτταρικών μονοπατιών σε πιο σφαιρικές προσεγγίσεις της κυτταρικής δραστηριότητας. Η ανάπτυξη της νέας αυτής τεχνολογίας έδωσε νέες, ενδιαφέρουσες πληροφορίες και αύξησε εκθετικά τα διαθέσιμα δεδομένα για την κατανόηση των βιολογικών συστημάτων. Η τεχνολογία των μικροσυστοιχιών, όπως και οι άλλες μέθοδοι υψηλής ανάλυσης, κρίνονται πλέον απαραίτητες για την κατανόηση των βιολογικών διεργασιών που λαμβάνουν χώρα στα βιολογικά συστήματα και συμπληρώνουν τις κοινές μεθόδους [M.03].

Από την αρχική της εφαρμογή σαν καινούρια τεχνική για μεγάλης κλίμακας χαρτογράφηση του DNA και αλληλούχιση και την αρχική επιτυχία σαν εργαλείο ανάλυσης του μεταγράφου, η τεχνολογία των μικροσυστοιχιών εξαπλώθηκε σε πολλές περιοχές προσαρμόζοντας τη βασική επινόηση και συνδυάζοντάς τη με άλλες τεχνικές.

Οι DNA μικροσυστοιχίες είναι μια διάταξη μεγάλου αριθμού ανιχνευτών DNA που αντιπροσωπεύουν συγκεκριμένους γενετικούς τόπους. Οι ανιχνευτές ακινητοποιούνται με ομοιοπολικούς δεσμούς σε μία στερεή επιφάνεια (συνήθως γυάλινη). Με άλλα λόγια, έχουν ακινητοποιηθεί σε γυάλινο πλακάκι, διαστάσεων μικρότερων της ανθρώπινης παλάμης, με τεχνικές της σύγχρονης νανοτεχνολογίας, ανιχνευτές γονιδίων σε συγκεκριμένα σημεία και αυτή η δομή αποτελεί μια μικροσυστοιχία. Στη θέση των ανιχνευτών γονιδίων μπορούν να μπουν ανιχνευτές πρωτεϊνών, τμήματα ιστών, ανιχνευτές μεταβολιτών κ.ά.

Η ανάλυση μικροσυστοιχιών χρησιμοποιεί μικροσυστοιχίες και επιτρέπει την ανάλυση της γονιδιακής έκφρασης, της ποικιλότητας, της αλληλουχίας του DNA, των επιπέδων και τροποποιήσεων των πρωτεϊνών κ.ά. με μαζική και παράλληλη επεξεργασία. Η ανάλυση μικροσυστοιχιών είναι μια εμπνευσμένη τεχνολογία με ευρείες εφαρμογές σε τομείς όπως η ανάλυση γονιδιώματος, η πρωτεομική, η διάγνωση. Στις μέρες μας, δίνουν τη δυνατότητα ανάλυσης ολόκληρου του γονιδιώ-

ματος ενός οργανισμού σε ένα μόνο πείραμα.

Μπορεί κανείς να εξάγει χρήσιμες πληροφορίες για τη βιολογική λειτουργία ενός οργανισμού, βρίσκοντας ποια γονίδια επάγονται ή καταστέλλονται σε κάποια φάση του κυτταρικού κύκλου, σε κάποια αναπτυξιακή στιγμή ή σε απόκριση σε ερεθίσματα του περιβάλλοντος, όπως για παράδειγμα η απόκριση σε ορμόνες ή σε υψηλή θερμοκρασία.

Ομάδες γονιδίων των οποίων η έκφραση αυξάνεται ή μειώνεται υπό τις ίδιες συνθήκες, είναι πιθανό να έχουν συσχετιζόμενη βιολογική λειτουργία και, ίσως, κοινή σχέση ρύθμισης [MWD+00]. Θα μπορούσαν, παραδείγματος χάριν, να έχουν παρόμοιες αλληλουχίες υποκινητών για ίδιους μεταγραφικούς παράγοντες. Επιπρόσθετα, αν μπορούσε κανείς να αποκομίσει ένα πρότυπο έκφρασης για συγκεκριμένες συνθήκες, θα είχε ένα χρήσιμο εργαλείο στη διάθεσή του, ώστε να είναι σε θέση να χαρακτηρίσει παρόμοιες άγνωστες καταστάσεις. Η γονιδιακή έκφραση είναι άμεσα συσχετιζόμενη με τις βιολογικές λειτουργίες και οι μικροσυστοιχίες υπόσχονται τεράστιο μέγεθος δεδομένων πάνω σε ανθρώπινες ασθένειες, γήρανση, φαρμακευτική δράση, ορμονική δράση, διανοητικές ασθένειες, μεταβολισμό και σε αρκετά ακόμα κλινικά θέματα. Ανοίγουν ένα νέο δρόμο σε μεθόδους διάγνωσης καθώς με την εξέλιξή τους γίνονται όλο και πιο διαθέσιμα σε χρήση σε εργαστήρια και χώρους διάγνωσης [M.96].

2.1 Ιστορικά στοιχεία

Τον Οκτώβριο του 1995 γεννήθηκε η τεχνολογία των μικροσυστοιχιών, όπως την αντιλαμβανόμαστε στις μέρες μας, με τη δημοσίευση ενός άρθρου από τους Schena, M. Shalon, D., Davis R.W. και Brown P.O. το οποίο δημοσιεύτηκε στο "Genome Issue" του Science. Προπομποί της τεχνολογίας ήταν οι μακροσυστοιχίες σε μεμβράνες που χρησιμοποιούνταν σαν πλατφόρμες. Οι μικροσυστοιχίες ενθουσιάζουν αλλά και προβληματίζουν την επιστημονική κοινότητα την τελευταία δεκαετία. Ακόμα δεν έχουν φτάσει τα όρια των δυνατοτήτων τους και συνεχίζουν να διεισδύουν σε όλο και περισσότερους τομείς της βιολογίας και της ιατρικής [U.05].

Στην προσπάθεια ανάλυσης του γονιδιώματος πολλών οργανισμών, γεννήθηκε η ανάγκη της λειτουργικής μελέτης χιλιάδων γονιδίων ταυτόχρονα. Ένα βήμα προς αυτήν την κατεύθυνση ήταν η αναγνώριση των γονιδιακών προτύπων έκφρασης υπό φυσιολογικές και παθολογικές καταστάσεις. Εντελώς τυχαία, στο "Genome Issue" του Science το 1995 υπήρχαν δύο άρθρα για ανάλυση προφίλ έκφρασης σε μεγάλη κλίμακα: το άρθρο για τις μικροσυστοιχίες από Schena et al. και το άρθρο για SAGE από Velculescu et al. Φαίνεται ότι οι μικροσυστοιχίες κερδίζουν τον αγώνα.

Από το 1995 η εταιρεία Affymetrix εισάγει το δικό της GeneChip με την ιδιαίτερη τεχνολογία και απαγορεύει τη χρήση παρόμοιας τεχνολογίας από άλλους,

κατόπιν κατοχύρωσης πατέντας. Έτσι αναπτύσσονται δύο παρόμοιες αλλά και διαφορετικές τεχνολογίες κατασκευής μικροσυστοιχιών.

2.2 Περιγραφή πειραματικής διαδικασίας

Σε αυτή την παράγραφο, θα δούμε κάποια βασικά χαρακτηριστικά της πειραματικής διαδικασίας. Η πειραματική διαδικασία αφορά στα στάδια που πρέπει να ακολουθηθούν κατά τη διεξαγωγή ενός πειράματος μικροσυστοιχιών. Εστιάζουμε περισσότερο σε πειράματα ανάλυσης γονιδιακού προφίλ. Ένα πείραμα με μικροσυστοιχίες είναι μια πολύπλοκη ακολουθία διεργασιών που πρέπει να αποπερατωθούν με επιτυχία για να εξασφαλιστεί η αποδεκτή ποιότητα των δεδομένων που θα προκύψουν και των αποτελεσμάτων που θα εξαχθούν. Είναι λογικό να υποθέσει κανείς πως, αφού τα βήματα που ακολουθούνται είναι πολλά και πολύπλοκα, είναι αυξημένες οι πιθανότητες το πείραμα να αποτύχει. Και έτσι είναι. Πρέπει να είμαστε επιφυλακτικοί απέναντι σε δεδομένα μικροσυστοιχιών. Όσο, όμως, τα χρόνια περνούν και οι τεχνικές βελτιώνονται, τόσο πιο εύκολο γίνεται να ολοκληρωθούν σωστά τα πειράματα και τα συμπεράσματα να είναι ασφαλέστερα.

Αν δούμε συνοπτικά τη διαδικασία ανάλυσης γονιδιακού προφίλ με μικροσυστοιχίες, θα λέγαμε ότι, αρχικά, διατυπώνεται ένα βιολογικό ερώτημα το οποίο ελπίζουμε ότι θα απαντηθεί με το τέλος του πειράματος. Στη συνέχεια, προχωράμε στην κατασκευή της μικροσυστοιχίας, πράγμα που σημαίνει ότι επιλέγουμε ποιον τύπο μικροσυστοιχίας θα χρησιμοποιήσουμε για να απαντήσουμε στο βιολογικό ερώτημα. Ο τύπος της μικροσυστοιχίας επιλέγεται ως προς τον τρόπο προετοιμασίας του, αλλά και το ποια είναι η διάταξη και το είδος των ανιχνευτών που ακινητοποιούνται στην επιφάνεια. Οι ανιχνευτές είναι πολύ σημαντικοί, καθώς είναι τα μόρια τα οποία θα ανιχνεύσουν τα μόρια cDNA του δείγματος.

Παράλληλα, ετοιμάζεται το βιολογικό υλικό, δηλαδή απομονώνεται mRNA από δύο τύπους κυττάρων, αν χρειαστεί επεκτείνεται και, τέλος, σημαίνεται με διαφορετικές αποχρώσεις για κάθε δείγμα, σε μικροσυστοιχίες εκτύπωσης. Αν χρησιμοποιείται GeneChip της Affymetrix χρειάζεται μόνο το ένα δείγμα κάθε φορά. Κατόπιν, ακολουθεί υβριδοποίηση των σημασμένων στόχων με τους ανιχνευτές της μικροσυστοιχίας. Το επόμενο βήμα είναι η σάρωση της επιφάνειας της μικροσυστοιχίας, απ' όπου προκύπτει μια ψηφιακή εικόνα, που προέρχεται από τη διέγερση των ετικετών σήμανσης που βρίσκονται στους στόχους και φθορίζουν σε δεδομένα μήκη κύματος.

Σε ένα πείραμα ανάλυσης γονιδιακού προφίλ, οι στόχοι από τα δύο διαφορετικά δείγματα βάφονται ξεχωριστά. Αυτό σημαίνει ότι η εικόνα που θα δώσουν είναι διαφορετική και ανιχνεύεται σε διαφορετικό μήκος κύματος. Αν, για παράδειγμα, το ένα δείγμα έχει σημανθεί με κόκκινο φθορίζον μόριο και το άλλο με πράσινο, δίνεται η δυνατότητα σύγκρισης της έντασης από την εικόνα που θα μας

δώσει το κάθε σημείο μετά την υβριδοποίηση. Λόγω του γεγονότος ότι η υβριδοποίηση γίνεται ταυτόχρονα και για τα δύο δείγματα, υπάρχει σαφής ανταγωνισμός μεταξύ των στόχων για τους ανιχνευτές. Οι στόχοι που βρίσκονται σε περίσσεια από το κάθε δείγμα για κάθε γονίδιο, θα υπερισχύσουν έναντι των λιγότερων άλλων και θα καταλάβουν περισσότερους ανιχνευτές στο σημείο. Θα δούμε κόκκινο σημείο αν το συγκεκριμένο γονίδιο υπερεκφράζεται στα κύτταρα του πρώτου δείγματος, θα δούμε πράσινο σημείο αν ισχύει το αντίστοιχο για το δεύτερο δείγμα και, τέλος, θα δούμε κίτρινο σημείο αν η έκφραση είναι παρόμοια [M.03].

Όταν, τέλος, έχουμε την εικόνα, προχωράμε σε ανάλυσή της, από την οποία προκύπτουν ποσοτικοποιημένα δεδομένα. Τα δεδομένα αυτά επεξεργάζονται στον ηλεκτρονικό υπολογιστή με πληθώρα αλγορίθμων, ώστε να απαλειφθούν τα σφάλματα και να δώσουν συμπεράσματα τα οποία ο άνθρωπος λόγω του μεγάλου όγκου δε θα μπορούσε να εξάγει.

2.2.1 Το «βιολογικό ερώτημα»

Με τον όρο «βιολογικό ερώτημα» εννοούμε κατά κύριο λόγο την ανάγκη να διατυπώνεται με ακρίβεια πριν προχωρήσει κανείς στη διαδικασία ενός πειράματος με μικροσυστοιχίες. Η διαδικασία της διατύπωσης ενός τέτοιου ερωτήματος εστιάζει την έρευνα σε συγκεκριμένο στόχο, βοηθάει στην επιλογή κριτηρίων ελέγχου και καθοδηγεί την ανάλυση των δεδομένων και τη μοντελοποίησή τους. Εξαιτίας της ποσότητας των δεδομένων που εξάγονται από την εκτέλεση ενός πειράματος μικροσυστοιχιών, δεν απαιτείται η διατύπωση μιας σαφούς υπόθεσης εξ' αρχής. Είναι ασφαλέστερο να διατυπωθεί απλά ένα σαφές ερώτημα.

Οι μικροσυστοιχίες δίνουν τη δυνατότητα επαρκούς αποκρυπτογράφησης της γονιδιακής έκφρασης. Η τεχνολογία των μικροσυστοιχιών απαντά σε ερωτήματα γονιδιακού προφίλ με ταυτόχρονη ανάλυση πολλών γονιδίων, ακόμα και ολόκληρου του γονιδιώματος. Ακόμα, προχωρά ταχύτερα η προσπάθεια γονοτύπησης οργανισμών, η αποκρυπτογράφηση πολυμορφισμών. Δίνονται απαντήσεις σε ερωτήματα για το ρόλο των πολλαπλών αντιγράφων γονιδίων στο γονιδίωμα με στοχοποίηση ολόκληρου του γονιδιώματος σε τμήματα πάνω στο πλακίδιο. Πρέπει, λοιπόν, να απαντηθεί αν χρειάζεται να ακινητοποιηθεί ολόκληρο το γονιδίωμα, μέρος του γονιδιώματος ή λίγα συσχετιζόμενα γονίδια.

Σημαντικό ρόλο στις αποφάσεις για την εξέλιξη του πειράματος, παίζει το βιολογικό υλικό που είναι διαθέσιμο. Αν το δείγμα που επιλέξουμε είναι mRNA τότε θα πρέπει να έχουμε κατά νου ότι είναι πιο ασταθές από το DNA, ταυτόχρονα όμως, συλλέγεται με μεγαλύτερη ευκολία. Επίσης, πρέπει να γνωρίζει κανείς αν η ποσότητα του βιολογικού υλικού είναι μικρή ή μεγάλη. Αυτό θα οδηγήσει σε απόφαση για ενίσχυση ή όχι της σήμανσης.

Αρκετοί επιστήμονες εστιάζουν σε ερωτήματα αναπτυξιακής φύσης. Εξετάζοντας τη γονιδιακή έκφραση μεταξύ διαφορετικών ομάδων κυττάρων, γίνεται προ-

σπάθεια να δημιουργηθεί βάση δεδομένων που να περιέχει πληροφορίες για τα επίπεδα γονιδιακής έκφρασης σε κάθε κυτταρικό τύπο. Τέτοιες βάσεις δεδομένων προάγουν τη βαθύτερη κατανόηση των βασικών αναπτυξιακών διαδικασιών.

Κρίσιμα βιολογικά ερωτήματα διατυπώνονται και στον τομέα των ασθενειών που προσβάλλουν τον άνθρωπο. Ελέγχονται διάφοροι παράγοντες και παρατηρούνται οι διαφοροποιήσεις στις αποκρίσεις συγκεκριμένων κυττάρων σε αυτούς. Περίπου 80% των πειραμάτων με μικροσυστοιχίες στις μέρες μας, αφορούν στον καρκίνο.

Στον τομέα ανακάλυψης φαρμάκων διατυπώνονται ερωτήματα για την απόκριση ασθενούντων κυττάρων ή ιστών σε συγκεκριμένες φαρμακευτικές ουσίες. Ένα απλό ερώτημα θα αφορούσε στη διαφοροποίηση της γονιδιακής έκφρασης πριν και μετά τη χορήγηση ενός φαρμάκου. Η προσπάθεια που γίνεται στον τομέα της ιατρικής είναι η ανακάλυψη συγκεκριμένων προτύπων γονιδιακής έκφρασης, για παράδειγμα, τα οποία να μπορούν να χρησιμοποιηθούν για διαγνωστικούς σκοπούς στο μέλλον [M.03].

2.2.2 Η μικροσυστοιχία και η κατασκευή της

Ας μελετήσουμε όμως τα βασικά χαρακτηριστικά μιας συστοιχίας και της κατασκευής της. Οι σημερινές τεχνικές κατασκευής μικροσυστοιχιών βρίσκονται στην αιχμή της νανοτεχνολογίας. Προέρχονται από ένα επιτυχημένο «πάντρεμα» της μηχανικής και της βιολογίας. Χρησιμοποιούνται διάφορες τεχνολογίες, που κάθε μία στοχεύει να κατασκευαστούν μικροσυστοιχίες υψηλής ποιότητας και οικονομικές ταυτόχρονα. Η εξόρυξη δεδομένων απαιτεί πολύ χρόνο και πρέπει να είμαστε σίγουροι ότι τα δεδομένα που θα προκύψουν προς ανάλυση είναι αξιόπιστα. Στην κατασκευή των μικροσυστοιχιών μπαίνουν πολλά κριτήρια όπως η τιμή κατασκευής, το περιεχόμενο, η πυκνότητα, το μέγεθος των χαρακτηριστικών (ανιχνευτών και σημείων), η καθαρότητα των χαρακτηριστικών, η αντιδραστικότητα των ανιχνευτών, η ευκολία εφαρμογής.

Το κόστος είναι ένας ανασταλτικός παράγοντας εφαρμογής των πειραμάτων μικροσυστοιχιών και απορρέει από το κόστος κατασκευής. Κατά την πορεία της τεχνολογίας των μικροσυστοιχιών και με την πρόοδο της τεχνολογίας, το κόστος συνεχώς μειώνεται. Έτσι αυτός ο σκόπελος δείχνει να ξεπερνιέται. Το περιεχόμενο αφορά στους ανιχνευτές. Είναι σημαντικό να γνωρίζει κανείς πόσοι ανιχνευτές περιέχονται στην επιφάνεια της μικροσυστοιχίας, ποιοι είναι αυτοί, καθώς και ποια είναι η δομή τους. Ακόμα, όταν υπάρχουν πολλαπλά αντίγραφα κάθε σημείου, προτιμάται να βρίσκονται σε γειτονικές θέσεις, ώστε να μπορεί κανείς να ελέγχει αν η μικροσυστοιχία δεν παρουσιάζει κατασκευαστικά προβλήματα, που θα υποδηλώνονταν με διαφορές στην ένταση σήματος μεταξύ αυτών των σημείων [DXGM06].

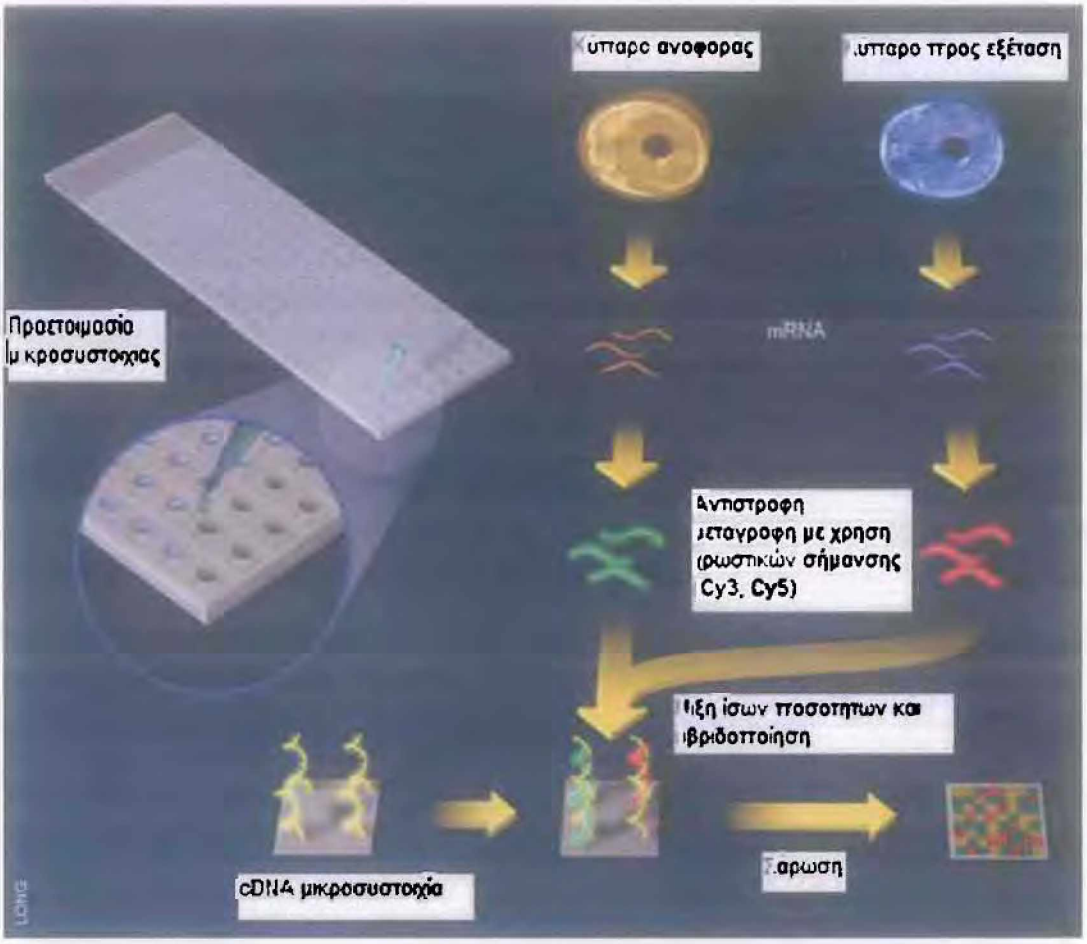
Η πυκνότητα αναφέρεται στον αριθμό των σημείων στην επιφάνεια της μικροσυστοιχίας. Υπολογίζεται απλά, γνωρίζοντας την απόσταση μεταξύ των κέντρων δύο σημείων και τον αριθμό των σημείων. Η πυκνότητα δίνεται πολλές φορές, σε μονάδα των χιλίων, δηλαδή ένα πλακάκι με 25.000 σημεία αναφέρεται σαν πλακάκι 25K. Είναι ένα σημαντικό κριτήριο γιατί καθορίζει την ποσότητα των δεδομένων που θα εξαχθεί από το πλακάκι ανά περιοχή. Ανάλογα με την τεχνολογία κατασκευής μπορούν στις μέρες μας να προκύψουν πλακάκια με 250.000 σημεία.

Το μέγεθος των χαρακτηριστικών αναφέρεται στο μέγεθος των σημείων της μικροσυστοιχίας και εκφράζεται, συνήθως, σαν μέση διάμετρος σημείου. Ανάλογα με την τεχνολογία που χρησιμοποιείται για την κατασκευή της μικροσυστοιχίας, η διάμετρος ποικίλει από τα 10 – 40 μm στα 75 – 300μm. Είναι ένα σημαντικό κριτήριο γιατί καθορίζει την πυκνότητα, για την οποία μιλήσαμε πιο πάνω. Επειδή μικρή διάμετρος σημαίνει και μεγαλύτερη πυκνότητα, άρα και μεγαλύτερος όγκος δεδομένων ανά μονάδα επιφάνειας, οι τεχνολογίες που αποδίδουν μικρότερη πυκνότητα έχουν και μεγαλύτερη αξία. Το μέγεθος των ανιχνευτών είναι μια παράμετρος που θα συζητήσουμε σε άλλο σημείο.

Η καθαρότητα των χαρακτηριστικών αφορά στην ομοιογένεια των ανιχνευτών σε κάθε σημείο. Αποκλίσεις από το 100% οφείλονται, για spotted μικροσυστοιχίες, σε υπολείμματα του παραγόμενου από την PCR υλικού (π.χ. εναπομείναντες εκκινητές) και σε προβλήματα κατά την κατασκευή. Πολλές φορές το αναποτελεσματικό πλύσιμο της εκτυπωτικής συσκευής (spotter) αφήνει κατάλοιπα τα οποία επηρεάζουν την καθαρότητα.

Εδώ βρίσκεται και ένα από τα σημαντικά πλεονεκτήματα των in situ παραγόμενων μικροσυστοιχιών οι οποίες εγγυώνται υψηλή καθαρότητα. Η αντιδραστικότητα αφορά στην ικανότητα αντίδρασης των ανιχνευτών στο σύνολό τους. Μια ιδανική μικροσυστοιχία έχει 100% αντιδραστικότητα, δηλαδή όλοι οι ανιχνευτές είναι σε θέση να αντιδράσουν με τους στόχους. Η πραγματικότητα λέει πως τα επίπεδα αντιδραστικότητας που επιτυγχάνονται συνήθως, είναι της τάξης του 50 – 60%. Ο παράγοντας αυτός καθορίζεται από το στάδιο της κατασκευής της μικροσυστοιχίας αλλά και από τη διαδικασία του πειράματος, όπου είναι πολύ πιθανό να προκληθεί μερική καταστροφή στους ανιχνευτές λόγω ανεπιθύμητης θερμοκρασίας ή άλλων χημικών προβλημάτων.

Τέλος, το κριτήριο που αναφέραμε σαν ευκολία εφαρμογής, αφορά στην ευκολία χρήσης μιας μικροσυστοιχίας στο εργαστήριο. Δηλαδή στο αν μια μικροσυστοιχία απαιτεί ειδική μεταχείριση ή αν είναι συμβατή με αρκετά μηχανήματα και μπορεί να δώσει ικανοποιητικά αποτελέσματα κάτω από πληθώρα συνθηκών. Οι μικροσυστοιχίες που κατασκευάζονται για ευρεία χρήση πρέπει να εξασφαλίζουν ευκολία εφαρμογής και να πληρούν τις απαιτήσεις της νέας τεχνολογίας.



Σχήμα 2.1: *Ένα ολοκληρωμένο παράδειγμα ανάλυσης*

2.2.2.1 Επιλογή τύπου μικροσυστοιχίας και ανιχνευτών που θα ακινητοποιηθούν

Όταν αποφασίζει κανείς να προχωρήσει στη διεξαγωγή ενός πειράματος, πρέπει, εκ των προτέρων, να έχει αποφασίσει τι τύπο μικροσυστοιχίας θα χρησιμοποιήσει. Αυτή η απόφαση εξαρτάται, σε μεγάλο βαθμό, από τα μηχανήματα που διαθέτει, καθώς πάντα είναι απαραίτητο να ελέγχεται η συμβατότητα των συσκευών με την εκάστοτε μικροσυστοιχία που θα επιλεχθεί.

Ο τύπος της μικροσυστοιχίας ελέγχεται ως προς τον τρόπο κατασκευής του και διακρίνεται σε μικροσυστοιχίες που κατασκευάστηκαν με in situ σύνθεση ανιχνευτών και σε μικροσυστοιχίες στις οποίες οι ανιχνευτές εκτυπώθηκαν στην επιφάνεια μετά την παρασκευή τους. Η δεύτερη κατηγορία χωρίζεται, επίσης, σε δύο κατηγορίες σχετικές με το μήκος των ανιχνευτών. Έτσι μιλάμε για μικροσυστοιχίες ολιγονουκλεοτιδίων και για μικροσυστοιχίες cDNA και αναλύονται παρακάτω.

Η επιλογή της μικροσυστοιχίας κατευθύνεται, επίσης, άμεσα από το βιολογικό ερώτημα που έχει τεθεί. Έτσι, αν κάποιος θέλει να ελέγξει την έκφραση ολόκληρου γονιδιώματος, για παράδειγμα, στις μέρες μας έχει πολλές επιλογές, καθώς έχει ταυτοποιηθεί το γονιδίωμα πολλών οργανισμών και οι βάσεις δεδομένων γίνονται όλο και πιο εμπιστεύσιμες και πλούσιες σε υλικό. Οι εταιρείες παρέχουν, επίσης, μεγάλη γκάμα μικροσυστοιχιών προς πώληση, απ' όπου μπορεί κανείς να επιλέξει αυτό που τον ενδιαφέρει. Μπορούμε να διακρίνουμε τις παρακάτω κατηγορίες μικροσυστοιχιών:

- Affymetrix GeneChip
- Μικροσυστοιχίες εκτύπωσης

Οι μικροσυστοιχίες εκτύπωσης, διακρίνονται, ανάλογα με το είδος των ανιχνευτών σε ολιγονουκλεοτιδικές και cDNA μικροσυστοιχίες.

2.2.3 Δημιουργία του κατάλληλου δείγματος

2.2.3.1 Απομόνωση βιολογικού υλικού

Σε ένα πείραμα ανάλυσης του γονιδιακού προφίλ έκφρασης, αλλά και σε οποιοδήποτε πείραμα μικροσυστοιχιών, κρίσιμο σημείο είναι η απομόνωση του βιολογικού υλικού που θα χρησιμοποιηθεί σα στόχος. Κατά την ανάλυση του γονιδιακού προφίλ έκφρασης προχωράει κανείς σε απομόνωση RNA από κύτταρα ή ιστούς. Αν δεν προχωρήσει σε σήμανση του mRNA, το οποίο αντιπροσωπεύει τη γονιδιακή έκφραση, στο διάλυμα του ολικού RNA, μπορεί να απομονώσει το mRNA και να προχωρήσει μεμονωμένα στη σήμανσή του.

Ένα τυπικό κύτταρο θηλαστικού περιέχει περίπου 10-5 μg RNA, 80 – 85% από το οποίο είναι ριβοσωμικό RNA (28S, 18S, 5,8S και 5S). Το περισσότερο από το υπόλοιπο RNA περιέχει μια ποικιλία από tRNA και snRNA. Τα παραπάνω RNA είναι συγκεκριμένου μεγέθους και αλληλουχίας και μπορούν να απομονωθούν επαρκώς με πολλές τεχνικές. Αντιθέτως, το mRNA, το οποίο αποτελεί το 1 – 5% του συνολικού RNA του κυττάρου, είναι ετερογενές στο μέγεθος και στην αλληλουχία. Όμως, το περισσότερο από το mRNA ενός ευκαρυωτικού κυττάρου φέρει στο 3' άκρο μια ουρά από κατάλοιπα πολυαδενυλικού οξέος, η οποία είναι αρκετά μακριά σε μήκος ώστε να επιτρέπει την απομόνωση του mRNA με χρωματογραφία συγγένειας σε oligo(dT) – cellulose. Ένα βασικό πρόβλημα στο χειρισμό του RNA είναι η ύπαρξη RNασών στο χώρο, ένζυμα που βρίσκονται σε όλα τα κύτταρα και απελευθερώνονται κατά τη λύση τους. Τα ένζυμα αυτά σπάνε το δεσμό μεταξύ ριβόζης και φωσφοδιεστερικής ομάδας. Για το λόγο αυτό οι χειρισμοί του RNA είναι ιδιαίτερα λεπτοί.

2.2.3.2 Σήμανση - επέκταση

Υπάρχουν δύο προσεγγίσεις στη διαδικασία της σήμανσης: η άμεση σήμανση και η έμμεση σήμανση. Κατά την πρώτη, τα φθορίζοντα μόρια προσδένονται στους στόχους απ' ευθείας με ενζυμικό ή χημικό μέσο. Κατά τη δεύτερη, οι στόχοι χρειάζονται ένα ενδιάμεσο μόριο – γέφυρα να προσδεθούν, το οποίο, με τη σειρά του, αλληλεπιδρά με τα φθορίζον μόριο. Και οι δύο προσεγγίσεις αποδίδουν στόχους υψηλής ποιότητας, αν και σε κάθε μια υπάρχουν πλεονεκτήματα και μειονεκτήματα. Παρακάτω αναλύονται σύγχρονοι τρόποι σήμανσης που ακολουθούν είτε άμεση είτε έμμεση προσέγγιση σήμανσης.

2.2.3.2.1 Αντίστροφη μεταγραφή

Είναι η πιο κλασσική μέθοδος άμεσης σήμανσης. Κατά τη διαδικασία, δείγματα mRNA απομονώνονται με ολιγο – dT και φθορίζοντα νουκλεοτίδια δημιουργούν cDNA με χρήση αντίστροφης μεταγραφάσης, ένζυμο που ξεκινά τη σύνθεση στα καλούπια mRNA που έχουν ήδη επάνω τους μεταγραφικούς εκκινητές. Στη συνέχεια, αποβάλλονται και πέπτονται με υδροξείδιο του θείου, αποδίδοντας έτσι ένα μονόκλωνο μίγμα cDNA, το οποίο καθαρίζεται για να αποβάλει νουκλεοτίδια που δεν έχουν ενσωματωθεί σε cDNAs, άλατα, ένζυμα και άλλα ανεπιθύμητα στοιχεία. Οι φθορίζοντες στόχοι cDNA που έχουν προκύψει, υβριδοποιούνται στη μικροσυστοιχία και, στη συνέχεια, είμαστε έτοιμοι να παρατηρήσουμε το πλακάκι, αφού προηγηθούν πολλά πλυσίματα, ώστε να απομακρυνθούν οι στόχοι που δεν μπόρεσαν να υβριδοποιηθούν [VRG01]. Από τότε που οι μικροσυστοιχίες μπήκαν στη ζωή μας, η μέθοδος της αντίστροφης μεταγραφής έχει τροποποιηθεί πάρα πολλές φορές και έχει δώσει εκατοντάδες πρωτόκολλα. Κάποια από αυτά χρησιμοποιούν ολικό RNA παρά mRNA, πράγμα που μειώνει λίγο την απόδοση της σήμανσης, αλλά σώζει πολύ χρόνο γιατί το ολικό RNA απομονώνεται πολύ πιο εύκολα από το mRNA. Κάποια άλλα χρησιμοποιούν εκκινητές των 9 νουκλεοτιδίων αντί για 20 ολιγο – dT εκκινητές, για να επιτραπεί η σήμανση mRNAs που δεν διαθέτουν poly-A ουρά, όπως τα βακτηριακά. Ακόμα, έχουν χρησιμοποιηθεί πολλοί διαφορετικοί τύποι αντίστροφης μεταγραφής και φαίνεται ότι κάποιοι από αυτούς αποδίδουν καλύτερη σήμανση από άλλους, ίσως λόγω καλύτερης συγγένειας με τα φθορίζοντα νουκλεοτίδια. Ποικιλία εμφανίζεται και στο στάδιο του καθαρισμού με χρήση είτε καθίζησης με αιθανόλη είτε αποκλεισμό μεγέθους είτε καθαρισμό που βασίζεται σε μεμβράνη. Είναι σημαντικό στάδιο αυτό του καθαρισμού, γιατί τα διαλύματα των στόχων που δεν περιέχουν μη ενσωματωμένα νουκλεοτίδια και άλλες προσμίξεις, δίνουν καλύτερα αποτελέσματα εικόνας που οφείλονται σε μείωση του θορύβου.

Ένα σημαντικό βήμα της άμεσης σήμανσης με αντίστροφη μεταγραφή, ήταν η ανάπτυξη και εξέλιξη ενός πρωτοκόλλου που εισάγει τη χρήση αμινοαλλύλικών

νουκλεοτιδικών αναλόγων, τα οποία επιτρέπουν την άμεση σήμανση των cDNA μορίων μέσω αντίδρασης των αμινοαλλυλικών ομάδων με ενεργές χρωστικές. Οι αμινοαλλύλες είναι πρωταρχικές αμίνες που δρουν σαν νουκλεόφιλα και αλληλεπιδρούν με χρωστικές, όπως παράγωγα κυανινών και Alexa, τα οποία περιέχουν ενεργές ομάδες και ζευγαρώνουν τη χρωστική άμεσα στο παραγώμενο DNA μέσω της αμινοαλλυλικής λειτουργίας. Το βασικό πλεονέκτημα αυτής της διαδικασίας είναι ότι, λόγω του μικρού μεγέθους των αμινοαλλυλικών ομάδων, τα αμινοαλλυλικά νουκλεοτιδικά παράγωγα ενσωματώνονται ευκολότερα στο DNA από τις ογκώδεις φθορίζουσες βάσεις. Η πιο αποτελεσματική ενσωμάτωση επιτρέπει πιο ομοιόμορφη σήμανση και, σε πολλές περιπτώσεις, στόχους που φθορίζουν εντονότερα.

Το βασικό πλεονέκτημα της άμεσης σήμανσης με αντίστροφη μεταγραφή είναι η απλότητα της μεθόδου. Μόρια στόχοι που σημαίνονται άμεσα εξαλείφουν την ανάγκη αντιδράσεων προ της υβριδοποίησης, πράγμα που μπορεί να είναι επίπονο και χρονοβόρο. Οι μέθοδοι αντίστροφης μεταγραφής είναι, ακόμα, πολύ ευπροσάρμοστες, επιτρέποντας τη χρήση πολλών διαφορετικών φθοριζουσών χρωστικών, όπως φλουροσείνη, Λισσαμίνη, κυανίνη, Alexa, BODIPY, αλλά και αμινοαλλυλικών παραγώγων. Όλες οι παραπάνω έχουν παρουσιάσει καλά αποτελέσματα σε πειράματα μικροσυστοιχιών. Το βασικό μειονέκτημα της άμεσης σήμανσης με αντίστροφη μεταγραφή είναι η μειωμένη ικανότητα απόδοσης σήματος σε σύγκριση με έμμεσες μεθόδους που παρέχουν τη δυνατότητα επέκτασης σήματος.

2.2.3.2.2 Χρήση RNA πολυμεράσης

Είναι ένα ένζυμο που χρησιμοποιείται πολύ στην ετοιμασία στόχων για πειράματα μικροσυστοιχιών. Οι RNA πολυμεράσες που χρησιμοποιούνται είναι μια οικογένεια ενζύμων από φάγους που καταλύουν τη σύνθεση RNA από δίκλωνο DNA – καλούπι που περιέχει ειδικούς υποκινητές φάγων. Η κατάλυση από T3 ή T7 RNA πολυμεράση είναι εξαιρετικά αποδοτική και παράγονται δεκάδες μικρογραμμαρίων RNA με λιγότερο από ένα μικρογραμμάριο αρχικής ποσότητας DNA. Οι χρήσεις στα πειράματα των μικροσυστοιχιών αφορούν σε: σύνθεση παρουσία rNTPs ώστε να παραχθούν μεγάλες ποσότητες RNA για μεταγενέστερες αντιδράσεις σήμανσης, επέκταση RNA με της μέθοδο Eberwine και σύνθεση παρουσία φθοριζόντων ή μη φθοριζόντων νουκλεοτιδικών αναλόγων για άμεση ή έμμεση σήμανση. Η ενσωμάτωση rNTPs που έχουν σημανθεί με βιοτίνη για χρήση σε μεταγενέστερη έμμεση διαδικασία σήμανσης φανερώνει την ενζυμολογία της RNA πολυμεράσης [LW01].

Υπάρχει ένας αριθμός πλεονεκτημάτων στη χρήση RNA πολυμεράσης για παραγωγή μορίων στόχων. Το προϊόν είναι RNA και όχι DNA, κάνοντάς το πιο εύκολο να τμηματοποιήσει κανείς το RNA σε μικρά τμήματα για υβριδοποίηση σε ολιγονουκλεοτιδικές μικροσυστοιχίες. Λόγω της υψηλής απόδοση της αντίδρασης

της αντίστροφης μεταγραφής, προκύπτει μια επέκταση με εκατονταπλάσιο προϊόν του RNA σε σύγκριση με το DNA – καλούπι, πράγμα που αυξάνει την πιθανότητα να παρατηρήσει κανείς σπάνια κομμάτια του δείγματος. Το ένζυμο είναι ενεργό σε κάθε δίκλωνο καλούπι που περιέχει τον κατάλληλο υποκινητή, κάνοντάς πιθανό να συντεθούν RNA – καλούπια από δίκλωνα μίγματα DNA (π.χ. cDNA) ή κλωνοποιημένες αλληλουχίες (π.χ. cDNA βιβλιοθήκες). Είναι, επίσης, πιθανό να προκύψει μεγάλο μέγεθος in vitro μεταγράφου που αφορά σε συγκεκριμένο γονίδιο που μας ενδιαφέρει, και αυτά τα in vitro μετάγραφα παρέχουν εξαιρετικά δείγματα ελέγχου (controls) για πειράματα σήμανσης mRNA.

Το κύριο μειονέκτημα των παραγώγων της RNA πολυμεράσης είναι ότι τα RNA μόρια που προκύπτουν είναι πολύ πιο ευάλωτα σε χημική και ενζυμική πέψη (ριβονουκλεάσες) από τα cDNA μίγματα που προκύπτουν από αντίστροφη μεταγραφή. Για να αποφευχθεί η αποδιάταξη του συντιθέμενου RNA, τα πλακάκια δεν πρέπει να έχουν την παραμικρή πρόσμιξη ριβονουκλεασών. Απαιτείται, σε τέτοιες περιπτώσεις, αποστειρωμένο δωμάτιο υβριδοποίησης και χρήση γαντιών.

2.2.3.2.3 Διαδικασία Eberwine

Η μέθοδος μετατρέπει το mRNA σε cDNA και έπειτα σε επεκταμένο RNA (aRNA), με τρόπο που επιτρέπει την επέκταση του αρχικού δείγματος κατά 106 φορές! Η μέθοδος βασίζεται στην γραμμική επέκταση των κλωνοποιημένων αλληλουχιών από την RNA ολυμεράση. Έτσι προκύπτει aRNA ανάλογο του αρχικού δείγματος. Η μέθοδος εφαρμόζεται όταν το mRNA που έχει απομονωθεί είναι πολύ μικρής ποσότητας. Το απομονωμένο mRNA υβριδοποιείται με ολιγο – dT εκκινητή που περιέχει επιπρόσθετες αλληλουχίες υποκινητή της T7 RNA πολυμεράσης. Το mRNA παράγει ένα μονόκλωνο cDNA μέσω αντίστροφης μεταγραφής και, στη συνέχεια, σε δίκλωνο cDNA χρησιμοποιώντας DNA πολυμεράση. Κάθε δίκλωνο cDNA διαθέτει έναν T7 υποκινητή. Έτσι μπορούμε να εξάγουμε τεράστιες ποσότητες aRNA από το διάλυμα cDNA με RNA πολυμεράση. Κάθε κύκλος της διαδικασίας δίνει εκατονταπλάσια ποσότητα από το αρχικό υλικό και τρεις κύκλοι δίνουν 106 φορές (100*100*100) περισσότερο. Δεδομένου ότι ένα ανθρώπινο κύτταρο έχει 0,5 pg mRNA, εφαρμόζοντας τη μέθοδο παίρνουμε 0,5 μg από ένα μόνο κύτταρο, αρκετό υλικό για την προετοιμασία φθοριζόντων στόχων ενός πειράματος μικροσυστοιχιών. Η διαδικασία Eberwine χρησιμοποιείται για την απάντηση βιολογικών ερωτημάτων που δε θα μπορούσαν να απαντηθούν με άλλα μέσα.

Το πλεονέκτημα της μεθόδου είναι ότι επιτρέπει τη φυσική επέκταση του RNA, με τους στόχους να παράγονται άμεσα από το επεκταμένο υλικό. Δίνει τη δυνατότητα διεξαγωγής πειράματος μικροσυστοιχιών με υλικό από ένα μόνο κύτταρο. Η διαδικασία είναι, μάλλον ποσοτική, λόγω του γεγονότος ότι η επέκταση γίνεται με γραμμικό τρόπο από δίκλωνα καλούπια, που κάθε ένα διαθέτει έναν πανομοιότυπο T7 υποκινητή. Μπορούμε να συγκρίνουμε τη μέθοδο με αυτήν της PCR. Τα

βασικά μειονεκτήματα είναι η δυσκολία της μεθόδου και ο χρόνος που απαιτείται για τη διεξαγωγή της. Τρεις κύκλοι απαιτούν 2 – 3 ημέρες πειραμάτων με αόρατο υλικό, όπως είναι τα νουκλεϊκά οξέα [R.03].

2.2.3.2.4 Επέκταση σήματος με τυραμίδιο

Η μέθοδος αυτή χρησιμοποιεί την ενζυμική δραστηριότητα του ενζύμου HRP να καταλύει την τοποθέτηση των φθοριζόντων μορίων τυραμιδίου στα σημεία των μικροσυστοιχιών που υβριδοποιούνται με μόρια που περιέχουν ετικέτες για τα αντισώματα φέρουν την HRP. Η μέθοδος μπορεί να παράγει εκατό φορές εντονότερο φθορισμό από τις μεθόδους άμεσης σήμανσης, χωρίς καθόλου αύξηση στη συγκέντρωση των στόχων. Κατά τη διαδικασία, η αντίστροφη μεταγραφή χρησιμοποιείται για να ενσωματώσει τη βιοτίνη, το DNP ή κάποιο άλλο μικρό μόριο στο cDNA. Το cDNA με τις ενσωματωμένες ετικέτες υβριδοποιείται στη μικροσυστοιχία και επωάζεται με αντίσωμα που φέρει την HRP. Το αντίσωμα προσδένεται στις ετικέτες του cDNA, φέρνοντας την ενσωματωμένη HRP σε γειτνίαση με την επιφάνεια της μικροσυστοιχίας. Το πλακάκι, στη συνέχεια, επωάζεται σε υπεροξείδιο του υδρογόνου, με την HRP να το χρησιμοποιεί για να οξειδώσει φθορίζοντα μόρια τυραμιδίου (π.χ Cy3 – tyramide). Το οξειδωμένο τυραμίδιο αντιδρά γρήγορα και προσδένεται αμέσως στην επιφάνεια της μικροσυστοιχίας.

Το βασικό πλεονέκτημα της επέκτασης σήματος με τυραμίδιο είναι ότι κατέχει την αποκλειστικότητα της ενζυμικής επέκτασης σήματος στην ανάλυση μικροσυστοιχιών. Εξαιτίας του γεγονότος ότι τα αντιδρώντα μόρια τυραμιδίου έχουν μικρό χρόνο ημιζωής, η τοποθέτηση των φθοριζόντων τυραμιδίων γίνεται πάνω ή πολύ κοντά στα σημεία της μικροσυστοιχίας, παρέχοντας έναν υψηλό βαθμό χωρικής ανάλυσης. Δεν υπάρχει ανεπιθύμητη εξάπλωση του σήματος, πράγμα που καθιστά τη μέθοδο να ταιριάζει απόλυτα σε διαμορφώσεις μικροσυστοιχιών υψηλής πυκνότητας. Υπάρχουν διαθέσιμα παράγωγα τυραμιδίου για φλουροσείνη, κυανίνη 3 και 5 και άλλων χρωστικών ουσιών, κάτι που εξασφαλίζει ποικιλία χρωμάτων. Το σήμα επεκτείνεται ανεξάρτητα από τον αριθμό των μορίων στόχων, αποτρέποντας τον κορεσμό των ανιχνευτών στην επιφάνεια της μικροσυστοιχίας και διατηρώντας τη γραμμικότητα σε πολλές τάξεις μεγέθους. Η μέθοδος επιτρέπει τη χρήση μικρών ποσοτήτων ολικού RNA (π.χ. 2-5 μg), κάτι που ελαχιστοποιεί την κατανάλωση πολύτιμου βιολογικού υλικού και επιτρέπει την πρόσβαση σε δείγματα αίματος και άλλων ιστών διαγνωστικής σημασίας. Η εκατονταπλάσια επέκταση σήματος επιτρέπει την ανίχνευση ανθρωπίνων μεταγράφων που θα ήταν αδύνατο να μετρήσουμε με μεθόδους άμεσης σήμανσης. Το κύριο μειονέκτημα της μεθόδου είναι ότι το πρωτόκολλό της είναι απαιτητικό σε χρόνο και επίπονο σε σχέση με αυτά της άμεσης σήμανσης, αλλά η ποιότητα των αποτελεσμάτων που προκύπτει επιβραβεύει την επιπλέον προσπάθεια [G.A02].

2.2.4 Το στάδιο της υβριδοποίησης

Η επόμενη διαδικασία, αφορά στο στάδιο της υβριδοποίησης. Τα μόρια των μονόκλωνων στόχων θα υβριδοποιηθούν με τα μονόκλωνα μόρια των ανιχνευτών. Η υβριδοποίηση αφορά στον σχηματισμό δεσμών υδρογόνου μεταξύ συμπληρωματικών νουκλεοτιδίων. Υπάρχουν ειδικά μηχανήματα, τα οποία ονομάζονται υβριδοποιητές και επιτελούν τη διαδικασία της υβριδοποίησης. Κατά τη διαδικασία τοποθετούμε τη μικροσυστοιχία σε ειδική θέση του μηχανήματος, το οποίο με τη βοήθεια ρυθμιστικού διαλύματος κρατά σταθερές τις επιθυμητές συνθήκες υβριδοποίησης [YD03].

Η σύσταση της αλληλουχίας, το μήκος του ανιχνευτών και στόχων, η θερμοκρασία υβριδοποίησης, η δευτεροταγής δομή, η συγκέντρωση αλάτων, το pH και αρκετοί άλλοι παράγοντες επηρεάζουν την αποτελεσματικότητα της υβριδοποίησης και τη συνοχή των διμερών που προκύπτουν. Ας εξετάσουμε αυτούς τους παράγοντες. Η αλληλουχία είναι σημαντική ως προς το ποσοστό των ζευγαριών GC που περιέχει, αφού σχηματίζουν τρεις δεσμούς υδρογόνου ανά ζευγάρι και όχι δύο, όπως τα AT ζευγάρια. Όσο, λοιπόν αυξάνεται το ποσοστό GC η συγγένεια υβριδοποίησης αυξάνεται και αντίστροφα. Η διαφορά στη δύναμη των δεσμών μεταξύ GC και AT μπορεί να μειωθεί με τη χρήση διαλύματος που περιέχει TMAC (τετραμεθυλαμμώνιο χλωρίδιο). Το τελευταίο αποσταθεροποιεί το ζευγάρωμα των GC και το κάνει σχετικό με αυτό των AT. Με αυτόν τον τρόπο μειώνονται οι διαφορές δεσμών μεταξύ GC – πλουσίων και AT – πλουσίων αλληλουχιών.

Στις περισσότερες αναλύσεις γονιδιακής έκφρασης, η ομολογία βρίσκεται στο 100%, καθώς οι ανιχνευτές υβριδοποιούνται με τους στόχους που προέρχονται από τα ίδια ομοειδή τους γονίδια. Σε αυτές τις περιπτώσεις, τα υβρίδια που προκύπτουν είναι πολύ δυνατά συνδεδεμένα και η υβριδοποίηση συντελείται με αυστηρούς όρους. Όμως η υβριδοποίηση, είναι μια δυναμική διαδικασία με όρους «ή όλα ή τίποτα». Αλληλουχίες με μικρότερη του 100% ομολογία μπορούν να σχηματίσουν υβρίδια, παρόλο που κάποιες από τις βάσεις μεταξύ ανιχνευτή και στόχου δεν έχουν ζευγαρώσει. Τέτοια υβρίδια πρέπει να λαμβάνονται σοβαρά υπ' όψιν. Οι γονιδιακές οικογένειες που είναι διατηρημένες σε μεγάλο βαθμό ή έχουν μεγάλο αριθμό μελών, θα διασταυρωθούν με μη ομοειδείς στόχους και αυτή η διασταυρούμενη υβριδοποίηση θα επηρεάσει τα αποτελέσματά μας. Η διασταυρούμενη υβριδοποίηση εμφανίζεται όταν η ομολογία μεταξύ ανιχνευτή και στόχου ξεπερνά το 70%, αλλά μπορεί να παρατηρηθεί και σε μικρότερη ομολογία κάποιες φορές. Έχει παρατηρηθεί ότι όσο μικραίνει το μέγεθος των ανιχνευτών και γίνεται πιο ειδικό, αυξάνεται η αποτελεσματικότητα της υβριδοποίησης, καθώς εξαλείφεται η διασταυρούμενη υβριδοποίηση. Σημαντικός ανασταλτικός παράγοντας είναι κάποιες φορές, το μέγεθος του στόχου, αφού όταν είναι μεγάλο συμβαίνει αναδίπλωση και εμφάνιση δευτεροταγών δομών, πράγμα που εμποδίζει την υβριδοποίηση. Όταν μικραίνουμε το μέγεθος του στόχου, πράγμα που πετυχαίνουμε με ενζυμική πέψη

ή χημική διάσπαση, εξασφαλίζουμε επιτυχέστερη υβριδοποίηση.

Είναι σημαντικό να έχουμε κατά νου ότι το κυρίαρχο κριτήριο είναι η ταυτότητα της νουκλεοτιδικής αλληλουχίας και όχι της αμινοξικής. Λόγω του εκφυλισμού του γενετικού κώδικα, δύο πρωτεΐνες είναι δυνατό να έχουν πανομοιότυπη αμινοξική αλληλουχία, αλλά τα γονίδιά τους να εμφανίζουν μικρότερη του 70% ομολογία. Η ιδιότητα να διασταυρώνονται παρόμοιες αλληλουχίες έχει χρησιμοποιηθεί σε εξελικτικές μελέτες.

Κατά την υβριδοποίηση πρέπει να λαμβάνονται υπ' όψιν και άλλα σημαντικά στοιχεία, όπως η θερμοκρασία και το pH. Η αυξημένη θερμοκρασία διευκολύνει τη διαδικασία γιατί αυξάνει τη θερμική ενέργεια, πράγμα που επιτρέπει αυξάνει την ταχύτητα διάχυσης και αποτρέπει το σχηματισμό δευτεροταγών δομών σε ανιχνευτές και στόχους. Η ενδεδειγμένη θερμοκρασία υβριδοποίησης υπολογίζεται στους 10°C περισσότερο από τη θερμοκρασία αναδιάταξης των κλώνων. Όσο για το pH, θεωρείται καλό να είναι ουδέτερο (από 5,5 ως 8,5) γιατί σε τέτοιες τιμές είναι πιο εύκολος ο σχηματισμός των δεσμών υδρογόνου μεταξύ των κλώνων. Το χαμηλό pH προκαλεί αποπουρινώσεις και το υψηλό διασπά τους δεσμούς υδρογόνου.

Στα ρυθμιστικά διαλύματα υβριδοποίησης προστίθενται πολλές φορές και συγκεκριμένες συγκεντρώσεις αλάτων. Ο σκοπός των αλάτων είναι να βελτιώσουν την αποτελεσματικότητα της υβριδοποίησης θωρακίζοντας αρνητικά τις φορτισμένες φωσφορικές ομάδες και ελαχιστοποιώντας τις ηλεκτροστατικές απωθήσεις.

2.2.5 Το στάδιο της σάρωσης

Είναι το στάδιο του πειράματος στο οποίο εισάγουμε τη μικροσυστοιχία σε ειδικό μηχάνημα που λέγεται σαρωτής (scanner). Το μηχάνημα σαρώνει την επιφάνεια της μικροσυστοιχίας και μετατρέπει την ένταση σήματος συγκεκριμένων μηκών κύματος, σε ψηφιακή εικόνα. Στην εικόνα που αποδίδεται περιλαμβάνονται όλα τα δεδομένα που μπορούμε να αποκομίσουμε από το πείραμα.

Ο όρος σαρωτής προέρχεται από το γεγονός ότι η ανίχνευση του φωτός (φθορισμού) γίνεται με κίνηση εμπρός και πίσω, δηλαδή σαρώνοντας την επιφάνεια της μικροσυστοιχίας. Πρακτικά, ένας σαρωτής μικροσυστοιχιών μπορεί να χαρακτηριστεί σαν τύπος συνεστιακού μικροσκοπίου που κινείται γρήγορα στο επίπεδο που ορίζουν οι άξονες x, y και εξάγει μεγάλες εικόνες φθορισμού.

Οι σαρωτές μικροσυστοιχιών είναι εντυπωσιακές συσκευές που στα βασικά σημεία δε διαφέρουν μεταξύ τους, παρά τη μεγάλη ποικιλία που υπάρχει στο εμπόριο. Αποτελούνται από λέιζερ (lazer), φίλτρο διέγερσης, διαχωριστή ακτίνας, φακούς αντικειμένου, αντικειμενοφόρο επιφάνεια, καθρέφτη ανάκλασης, φίλτρο εκπομπής, τρύπα ακίδας, φωτοπολλαπλασιαστικό σωλήνα (PMT), μετατροπέα αναλογικού σήματος σε ψηφιακό και έναν υπολογιστή [D.03].Το φως του λέιζερ κα-

τευθύνεται διαμέσου του φίλτρου διέγερσης το οποίο επιτρέπει να περάσει το φως με μήκος κύματος της επιλογής μας. Το συγκεκριμένο μήκος κύματος πρέπει να βρίσκεται κάτω από το μέγιστο μήκος κύματος της διέγερσης που επιτρέπει το φθορίζον μόριο με το οποίο έχουμε σημάνει τους στόχους μας. Το φως του λέιζερ ανακλάται από το διαχωριστή ακτίνας, ο οποίος το εστιάζει στην επιφάνεια της μικροσυστοιχίας. Η διέγερση που προκαλείται από το φως κάνει τη μικροσυστοιχία να φθορίζει φως, το οποίο επιστρέφει με σποραδικό τρόπο και συγκεντρώνεται σε παράλληλη ακτίνα με την πρώτη με μια διαδικασία γνωστή ως παράλληλη σκόπευση (collimation). Ένας κεκλιμένος καθρέφτης οδηγεί την τελευταία ακτίνα φθορισμού διαμέσου ενός φίλτρου εκπομπής, το οποίο επιτρέπει σε να περάσει μόνο αυστηρά συγκεκριμένο μήκος κύματος και αναχαιτίζει τα υπόλοιπα, συμπεριλαμβανομένου και του μήκους κύματος του φωτός διέγερσης. Η ακτίνα που πέρασε, οδηγείται μέσα από φακούς και την τρύπα ακίδας, στο φωτοπολλαπλασιαστικό σωλήνα. Τα αναλογικά σήματα που προέρχονται από το σωλήνα μετατρέπονται σε ψηφιακά σήματα από ειδικό μετατροπέα και μεταφέρονται σε ηλεκτρονικό υπολογιστή σε μορφή αρχείου εικόνας [VBH07].

Οι σαρωτές μικροσυστοιχιών εξάγουν δεδομένα με ρυθμό ένα στοιχείο (pixel) σε κάθε βήμα και για το λόγο αυτό, το να προκύψει μια μεγάλη εικόνα απαιτείται φυσική κίνηση του υποστρώματος ή του οπτικού αντικειμένου κατά τη διάρκεια της ανίχνευσης. Υπάρχουν, λοιπόν, δύο τύποι αρχιτεκτονικής σαρωτών, γνωστές σαν κινητού υποστρώματος, όπου το μηχάνημα μετακινεί τη μικροσυστοιχία, και κινητού οπτικού αντικειμένου, όπου η μικροσυστοιχία παραμένει σταθερή και κινείται η ακτίνα διέγερσης. Ένα πείραμα απαιτεί, συνήθως, ανίχνευση σε περισσότερα του ενός μήκους κύματος. Αυτό επιτυγχάνεται με χρήση ξεχωριστών λέιζερ, φίλτρων και φωτοπολλαπλασιαστικών σωλήνων. Η ανίχνευση των Cy3 και Cy5, που ανιχνεύονται στα 550nm και 649nm αντίστοιχα, συνήθως επιτυγχάνεται με δύο ξεχωριστά λέιζερ, φίλτρα και φωτοπολλαπλασιαστικούς σωλήνες, που είναι ρυθμισμένα να διεγείρουν την κάθε απόχρωση ξεχωριστά και να εξάγουν ξεχωριστά δεδομένα φθορισμού. Οι σημερινοί σαρωτές είναι ικανοί να ανιχνεύσουν μέχρι και 11 αποχρώσεις ξεχωριστά στο εύρος 488 – 652nm. Για την εξαγωγή εικόνας από μικροσυστοιχίες που περιέχουν περισσότερες από μια αποχρώσεις, ο σαρωτής πρέπει να έχει σχεδιαστεί με τέτοιο τρόπο ώστε να αποτρέπεται η εισχώρηση της μιας ακτίνας στο κανάλι μιας άλλης. Ιδιαίτερα όταν τα μήκη κύματος βρίσκονται κοντά, πρέπει να λαμβάνεται μεγάλη φροντίδα στη λειτουργία του σαρωτή.

Μια συλλογή ψηφιακών δεδομένων στον υπολογιστή, που προέρχονται από αναλογικά σήματα φωτονίων από την επιφάνεια της μικροσυστοιχίας χαρακτηρίζονται ως αρχείο. Τέτοια αρχεία εικόνας μικροσυστοιχίας έχουν, συνήθως, TIFF διαμόρφωση (.tif επέκταση). Σε κάποιες περιπτώσεις συναντάμε GIF (.gif) και JPEG (.jpg), οι οποίες είναι, επίσης, διαμορφώσεις αρχείων εικόνας σε πιο συμπιεσμένη μορφή και μπορούν εύκολα να μετατραπούν σε TIFF και αντίστροφα.

Μια ακτίνα λέιζερ χρησιμοποιείται για να διεγείρει τις χρωστικές στην επιφά-

νεια της μικροσυστοιχίας. Στη συνέχεια, η ακτινοβολία που εκπέμπει η χρωστική συλλέγεται από τον φωτοπολλαπλασιαστή (PMT) και μετατρέπεται σε ψηφιακό σήμα. Για να αποθηκευτεί σαν αρχείο γραφικών, η εικόνα της μικροσυστοιχίας σπάει σε διακριτά κομμάτια, τα στοιχεία (pixels), τα οποία αντιπροσωπεύουν το μέσο σήμα φθορισμού κάθε περιοχής στοιχείου. Όσο μεγαλύτερη είναι η ανάλυση, η επιφάνεια που καλύπτει το κάθε ένα μικραίνει, ο αριθμός των στοιχείων αυξάνεται και η απόδοση της πραγματικής αναλογικής εικόνας γίνεται πιο αντιπροσωπευτική. Κάθε σημείο της μικροσυστοιχίας αναλύεται πλέον σε πολύ μεγάλο αριθμό στοιχείων (ως 2μm διάμετρος στοιχείου). Η αύξηση της ανάλυσης βελτιώνει την ποιότητα της εικόνας και των δεδομένων, κατ' επέκταση, αλλά επιβραδύνει την ταχύτητα σάρωσης και παράγει μεγαλύτερα αρχεία που απαιτούν μεγαλύτερο αποθηκευτικό χώρο. Θεωρείται ότι μπορούμε να υπολογίσουμε τη βέλτιστη ανάλυση από τον τύπο:

Προτεινόμενη ανάλυση σάρωσης = διάμετρος σημείου (σε μm) / 10

Τέλος, με τα σύγχρονα προγράμματα, είναι εύκολο να συνδυάσει κανείς δύο διαφορετικές εικόνες από την ίδια μικροσυστοιχία, οι οποίες έχουν καταγραφεί σε διαφορετικές αποχρώσεις, αφού αφορούν σε διαφορετική σήμανση. Αυτό είναι πολύ σημαντικό, κυρίως σε πειράματα ανάλυσης του μεταγραφικού προφίλ, όπου χρησιμοποιούνται διαφορετικές ετικέτες σήμανσης για τα ξεχωριστά δείγματα. Η συνδυασμένη εικόνα παρέχει ένα εύκολο τρόπο να δει κανείς και να αναγνωρίσει τα γονίδια που βρίσκονται σε μεγαλύτερη έκφραση στο δείγμα σε σύγκριση με το δείγμα αναφοράς. Σε έναν πράσινο – κόκκινο χρωματισμό, τα πράσινα σημεία αφορούν σε γονίδια τα οποία βρίσκονται σε αφθονία στο εξεταζόμενο δείγμα και τα κόκκινα σε γονίδια τα οποία βρίσκονται σε αφθονία στο δείγμα αναφοράς.

2.2.6 Ανάλυση δεδομένων

Παρακάτω θα ασχοληθούμε με μεθόδους ανάλυσης δεδομένων από μικροσυστοιχίες DNA. Πιο συγκεκριμένα θα ασχοληθούμε με τις μεθόδους τις ποσοτικοποίησης, της εξαγωγής αναλογιών, της κανονικοποίησης και της πολυπαραμετρικής στατιστικής ανάλυσης.

2.2.6.1 Ποσοτικοποίηση

Η επιφάνεια της μικροσυστοιχίας βρίσκεται σε αρχείο εικόνας (tiff) στον υπολογιστή, όπως έχει προαναφερθεί. Η εικόνα μοιάζει με πολύχρωμο χάρτη και είναι χωρισμένη σε μικρά ψηφία (pixels). Κάθε σημείο της μικροσυστοιχίας αναπαριστά ένα γονίδιο και εμπεριέχει πολλαπλά ψηφία, από τα οποία κάθε ένα περιέχει ενσωματωμένες ποσοτικές πληροφορίες για το γονίδιο ή το παράγωγό του στη δεδομένη τοποθεσία της μικροσυστοιχίας. Η διαδικασία κατά την οποία παίρνουμε πο-

σοτικοποιημένες πληροφορίες από τα αρχεία εικόνας των μικροσυστοιχιών, ονομάζεται ποσοτικοποίηση.

Η ποσοτικοποίηση επιτρέπει στον ερευνητή να εξάγει αριθμητικές τιμές για κάθε σημείο της μικροσυστοιχίας από το αρχείο εικόνας. Οι τιμές δίνουν πληροφορίες που μπορούν να χρησιμοποιηθούν για να προσδιοριστούν οι συγκεντρώσεις των δειγμάτων σε κάθε σημείο. Η απόλυτη ποσοτικοποίηση είναι η διαδικασία κατά την οποία παίρνουμε αριθμητικές τιμές για κάθε σημείο βασισμένες στην ένταση σήματος κάθε τοποθεσίας. Τα απόλυτα αυτά νούμερα που προκύπτουν μας βοηθούν να υπολογίσουμε την αφθονία δείγματος που εμφανίζεται σε κάθε σημείο. Η σχετική ποσοτικοποίηση αναφέρεται στη σύγκριση των απόλυτων σημάτων από δύο διαφορετικά δείγματα μιας μικροσυστοιχίας και δίνεται με τη μορφή πηλίκου διαίρεσης του απόλυτου σήματος του ενός με το άλλο. Δίνει συγκριτικά αποτελέσματα και χρησιμοποιείται σε διαδικασίες σήμανσης με περισσότερα από ένα χρώματα [GT03].

Όλες οι μορφές ποσοτικοποίησης πραγματοποιούνται με χρήση υπολογιστή. Μια ακόμη διάκριση ποσοτικοποίησης αφορά στον αριθμό των σημείων που ποσοτικοποιούνται σε κάθε βήμα. Κατά την ποσοτικοποίηση με το χέρι, ο ερευνητής ποσοτικοποιεί κάθε σημείο ξεχωριστά σε κάθε βήμα. Αυτό γίνεται με επιλογή μέσα από το πρόγραμμα επεξεργασίας που χρησιμοποιείται κάθε φορά. Κάθε σημείο έχει δύο αριθμητικά στοιχεία γνωστά σα σήμα και θόρυβος (background). Το σήμα αναφέρεται σε πραγματικά δεδομένα τιμών έντασης μικροσυστοιχιών και δίνει πληροφορίες για την έκφραση γονιδίων, για αλληλεπιδράσεις πρωτεϊνών κτλ. Ο θόρυβος αφορά σε μη επιθυμητές τιμές έντασης σχετιζόμενες με βιοχημικά γεγονότα όπως μη ειδική πρόσδεση στους ανιχνευτές, φωσφορισμός υποστρώματος κ.ά. Ο θόρυβος πρέπει να αφαιρεθεί. Όταν πραγματοποιούμε ποσοτικοποίηση με το χέρι, ανοίγουμε το αρχείο εικόνας με το πρόγραμμα επεξεργασίας, προσδιορίζουμε μια περιοχή ποσοτικοποίησης με χρήση ενός πλαισίου (συνήθως κύκλου) γύρω από κάθε σημείο και εξάγουμε τα δεδομένα. Θα πρέπει, κατά την τοποθέτηση του πλαισίου, να συμπεριλαμβάνεται μαζί με το κάθε σημείο και κάποιος χώρος έξω από αυτό, για να μπορεί να προσδιοριστεί ο θόρυβος και να αφαιρεθεί από το σήμα που εξάγεται από το σημείο. Κάποια προγράμματα δίνουν τη δυνατότητα να ορίσει κανείς πλαίσιο υπολογισμού σήματος και θορύβου ξεχωριστά. Μπορεί κανείς να υπολογίσει τη μέση τιμή σήματος και θορύβου από το άθροισμα των εντάσεων κάθε ψηφίου εντός του πλαισίου διαιρεμένο με το συνολικό αριθμό των ψηφίων. Η διαδικασία κατά την οποία τα σήματα των μικροσυστοιχιών περιχαρακώνονται από το θόρυβο της εικόνας ονομάζεται κατάτμηση σήματος (signal segmentation). Ο χρήστης μπορεί να αυξήσει την ακρίβεια της κατάτμησης σήματος ορίζοντας τις φυσικές και στατιστικές παραμέτρους που θα χρησιμοποιηθούν για να ξεχωρίσουν το σήμα από το θόρυβο. Οι δύο βασικές μορφές της κατάτμησης σήματος είναι η διαστηματική κατάτμηση και η κατάτμηση έντασης. Οι μέθοδοι που χρησιμοποιούν διαστηματική κατάτμηση βασίζονται σε προσεκτική τοπο-

θέτηση του πλαισίου ποσοτικοποίησης για το σήμα και το θόρυβο ξεχωριστά. Η κατάτμηση έντασης χρησιμοποιεί τιμές έντασης κάθε ψηφίου και βασίζεται σε μαθηματικές και στατιστικές προσεγγίσεις ώστε να αποκλείσει ανώμαλες εντάσεις ψηφίων.

Κατά την ποσοτικοποίηση των τιμών έντασης μπαίνουν κάποια όρια έντασης. Όταν ξεπεραστούν αυτά τα όρια τα σήματα δε γίνονται αντιληπτά από το όργανο ανίχνευσης και ονομάζονται κορεσμένα σήματα. Η αναπαράσταση των δεδομένων των μικροσυστοιχιών σε παλέτες χρωμάτων ουράνιου τόξου είναι ένα στοιχείο που βοηθά πολύ το χρήστη να επιλέξει τις ρυθμίσεις ανίχνευσης και να ορίσει μεγαλύτερα όρια ώστε να ανιχνευτούν και τα κορεσμένα σήματα. Τα κορεσμένα σήματα συνήθως αναπαρίστανται με άσπρο χρώμα και αγνοούνται κατά την ποσοτικοποίηση.

Κάτι παρόμοιο γίνεται και με τα αδύναμα σήματα τα οποία αντιστοιχούν σε σπάνια μετάγραφα και μπορεί να είναι πολύ σημαντικό να ανιχνευτούν. Ο χρήστης αυξάνει την ευαισθησία και ανιχνεύει σήματα που μόλις ξεπερνούν το θόρυβο αν και τα περισσότερα από τα υπόλοιπα σήματα φαίνονται κορεσμένα [LM04].

Οι μικροσυστοιχίες συνήθως, περιέχουν χιλιάδες ή ακόμα και δεκάδες χιλιάδες στοιχεία που αντιστοιχούν σε σημεία και απαιτούν ισχυρά προγράμματα επεξεργασίας για ποσοτικοποίηση. Η διαδικασία κατά την οποία πολλά στοιχεία της μικροσυστοιχίας ποσοτικοποιούνται ταυτόχρονα λέγεται αυτοματοποιημένη ποσοτικοποίηση. Με αυτή τη διαδικασία παράγεται μεγάλος όγκος δεδομένων. Χρησιμοποιεί τις βασικές αρχές που αναφέρθηκαν στην ποσοτικοποίηση με το χέρι.

Η αυτοματοποιημένη ποσοτικοποίηση κάνει χρήση ενός συγκεκριμένου γραφικού στοιχείου, του καλουπιού ποσοτικοποίησης. Το τελευταίο τοποθετείται πάνω στην εικόνα σε συγκεκριμένη θέση ώστε να συμπέσουν οι γωνίες του με αυτές της εικόνας της μικροσυστοιχίας. Υπολογίζονται και καθορίζονται εξ' αρχής οι αριθμοί των στηλών και των γραμμών των σημείων, ώστε να συμπεριληφθεί κάθε σημείο σε συγκεκριμένο πλαίσιο ποσοτικοποίησης. Μόλις το καλούπι τοποθετηθεί στη σωστή θέση, ο υπολογιστής ξεκινά την ποσοτικοποίηση για κάθε σημείο ξεχωριστά, αλλά και ταυτόχρονα. Παράγονται έτσι ακόμα και γιγαμπάιτ (GB) δεδομένων από ένα μόνο πείραμα. Τα δεδομένα που παράγονται από την αυτοματοποιημένη ποσοτικοποίηση δίνονται από το πρόγραμμα σε μορφή αρχείων κειμένου (.txt) ή και άλλων μορφών (.csv, .gpr).

2.2.6.2 Εξαγωγή αναλογιών

Είναι η πρώτη επεξεργασία που γίνεται στα δεδομένα για την ανίχνευση διαφορικής έκφρασης [Yan02]. Παρόλο που η αναλογία T δίνει μια μέτρηση των αλλαγών έκφρασης, έχει το μειονέκτημα ότι συμπεριφέρεται διαφορετικά στα θετικά και στα αρνητικά ρυθμιζόμενα γονίδια. Τα θετικά ρυθμιζόμενα γονίδια σε διπλάσιο βαθμό δίνουν μια αναλογία της τάξης του 2, ενώ τα αρνητικά ρυθμιζόμενα γονί-

δια στον ίδιο βαθμό δίνουν μια αναλογία της τάξης του 0,5. Για το λόγο αυτό, μια συνήθης τακτική που εφαρμόζεται, είναι η λογαρίθμιση των δεδομένων στη βάση του 2, κάτι που δίνει το πλεονέκτημα να εξάγεται ένα συνεχές φάσμα τιμών και να υπάρχει παρόμοια συμπεριφορά στην επεξεργασία των θετικά και αρνητικά ρυθμιζόμενων γονιδίων. Οι λογάριθμοι των αναλογιών, επίσης, εμφανίζουν συμμετρία. Έτσι, ένα γονίδιο θετικά ρυθμιζόμενο σε διπλάσιο βαθμό έχει log2(αναλογία) = 1, ένα αρνητικά ρυθμιζόμενο γονίδιο στον ίδιο βαθμό έχει log2(αναλογία) = -1 και ένα γονίδιο με σταθερή έκφραση έχει log2(αναλογία) = 0.

2.2.6.3 Κανονικοποίηση

Με την κανονικοποίηση, δεδομένα από διαφορετικές συστοιχίες εξισορροπούνται πριν την ανάλυση. Αν εκτελεστεί σωστά, η κανονικοποίηση δε μεταβάλλει τα δεδομένα, αντιθέτως εξαλείφει ανισορροπίες που προκύπτουν κατά τη διαδικασία απεικόνισης και που προέρχονται από διαφορές στην αποτελεσματικότητα της σήμανσης και της υβριδοποίησης, των εκπλύσεων, διαφοροποιήσεις στην ένταση του λέιζερ και την ευαισθησία του ρομπότ ανίχνευσης. Είναι επίσης, απαραίτητη για την ακριβή εξαγωγή αναλογιών [JD06].

Είναι καλό να ελέγχονται εκ των προτέρων οι ρυθμίσεις των οργάνων ανίχνευσης και απεικόνισης καθώς και να τηρούνται συγκεκριμένα πρωτόκολλα, ώστε ο ρόλος της κανονικοποίησης να περιορίζεται και να προκύπτουν ισορροπημένα δεδομένα [SCSF00]. Πολλά σύγχρονα προγράμματα επεξεργασίας, απεικόνισης και ελέγχου οργάνων, παρέχουν δυνατότητες ακριβούς ελέγχου των συνθηκών του πειράματος και μειώνουν εξ' αρχής την ανάγκη κανονικοποίησης. Μερικές συνηθισμένες τακτικές κανονικοποίησης είναι οι παρακάτω:

- Κανονικοποίηση ολικής έντασης (total intensity normalization)
- Lowess (locally weighted linear regression) κανονικοποίηση
- Τυπική απόκλιση (Standard deviation)
- Φιλτράρισμα αντιγράφων (replicates)

2.2.6.4 Πολυπαραμετρική στατιστική ανάλυση

Η πολυπαραμετρική στατιστική ανάλυση αφορά σε αλγορίθμους ανάλυσης δεδομένων που χρησιμοποιούν περισσότερες από μια μεταβλητές σε κάθε βήμα. Για την ανάλυση δεδομένων μικροσυστοιχιών μπορεί να χρησιμοποιηθεί μια πληθώρα στατιστικών μεθόδων, αλγορίθμων ταξινόμησης και διάκρισης των στοιχείων, που εντάσσονται στα πλαίσια της πολυπαραμετρικής ανάλυσης [A.02, KL.07]. Παρακάτω θα περιγράψουμε κάποιες από τις πιο διαδεδομένες μεθόδους, χωρίζοντάς

τες σε δύο μεγάλες κατηγορίες: αυτές που αφορούν στην ομαδοποίηση των δεδομένων και σε αυτές που αφορούν στην οπτικοποίηση τους.

2.2.6.4.1 Ομαδοποίηση: Ιεραρχική και k-means

Οι αλγόριθμοι ομαδοποίησης ταξινομούν τα δεδομένα και τις ομάδες γονιδίων μαζί, με βάση το διαχωρισμό τους στην έκφραση. Έχουν αναπτυχθεί διάφορες τεχνικές ομαδοποίησης για την εξαγωγή προτύπων γονιδιακής έκφρασης [AJZ02]. Οι περισσότερες από αυτές είναι ιεραρχικές (hierarchical), δηλαδή η ταξινόμηση που προκύπτει έχει έναν αυξανόμενο αριθμό εμφωλευμένων κλάσεων και το αποτέλεσμα μοιάζει με φυλογενετική ταξινόμηση. Υπάρχουν επίσης, μη ιεραρχικοί αλγόριθμοι (non – hierarchical), όπως ο k-means, οι οποίοι απλά διαχωρίζουν τα αντικείμενα σε διαφορετικές ομάδες, χωρίς να προσπαθούν να βρουν τη σχέση μεταξύ των ξεχωριστών αντικειμένων, πράγμα που είναι το ζητούμενο στις ιεραρχικές ομαδοποιήσεις [CCJP06].

Οι τεχνικές ομαδοποίησης μπορούν να χωριστούν περαιτέρω σε διαχωριστικές (divisive) και συνενωτικές (agglomerative). Μια διαχωριστική μέθοδος ξεκινά με όλα τα δεδομένα σε μια ομάδα, η οποία σταδιακά χωρίζεται σε όλο και μικρότερες. Αντίστροφα, οι συνενωτικές μέθοδοι ξεκινούν από ομάδες του ενός μόνο στοιχείου και σταδιακά δημιουργούν ομάδες με περισσότερα μέλη.

Τέλος, η ομαδοποίηση μπορεί να είναι επιβλεπόμενη (supervised) ή μη επιβλεπόμενη (unsupervised). Οι επιβλεπόμενες μέθοδοι χρησιμοποιούν υπάρχουσες βιολογικές πληροφορίες για συγκεκριμένα γονίδια που μπορούν να λειτουργήσουν δίνοντας κατευθύνσεις στον αλγόριθμο ομαδοποίησης. Οι περισσότερες, όμως από τις μεθόδους είναι μη επιβλεπόμενες [J.01].

Κατά την εφαρμογή αλγορίθμων ομαδοποίησης πρέπει να λαμβάνεται σοβαρά υπ' όψιν η απόσταση μεταξύ των αντικειμένων που ομαδοποιούνται και, ακριβέστερα, ο τρόπος υπολογισμού της, ο οποίος είναι σημαντικός και δίνει διαφορετικά αποτελέσματα ανά περίσταση. Έτσι, πρέπει να είναι γνωστό εκ των προτέρων ποιος τύπος απόστασης θα χρησιμοποιηθεί και γιατί.

Ένας άλλος τρόπος ομαδοποίησης είναι η ιεραρχική. Η ιεραρχική ομαδοποίηση έχει το πλεονέκτημα ότι είναι σχετικά απλή και μπορεί να γίνει εύκολα ορατή [DXGM06]. Είναι η πιο διαδεδομένη μέθοδος ανάλυσης δεδομένων γονιδιακής έκφρασης. Αποτελεί μια συνενωτική μέθοδο κατά την οποία το κάθε δεδομένο εντάσσεται σε ομάδες, οι οποίες ομαδοποιούνται και μεταξύ τους μέχρι να ολοκληρωθεί η διαδικασία που θα δώσει ένα ιεραρχικό δένδρο [A.02, KL.07]. Η διαδικασία προχωρά με απλό τρόπο. Αρχικά, υπολογίζεται ένας πίνακας αποστάσεων για όλα τα γονίδια που πρέπει να ομαδοποιηθούν. Στη συνέχεια, ο πίνακας αποστάσεων εξερευνάται για την εύρεση των δύο πιο όμοιων γονιδίων ή ομάδων. Θυμίζουμε ότι αρχικά κάθε ομάδα περιέχει ένα μόνο γονίδιο, δηλαδή έχουμε τόσες ομάδες όσα και τα γονίδια. Αν υπάρχουν περισσότερα του ενός ζευγάρια με ίσες

αποστάσεις, καλό θα ήταν να έχει καθοριστεί εκ των προτέρων ένας κανόνας απόφασης για τέτοιες περιπτώσεις. Τώρα, τα δύο γονίδια – ομάδες ενώνονται και παράγουν μια ομάδα με δύο μέλη το λιγότερο. Συνεχίζοντας, επαναϋπολογίζονται οι αποστάσεις μεταξύ της νέας ομάδας και των υπολοίπων. Δεν είναι απαραίτητο να υπολογιστούν εκ νέου όλες οι αποστάσεις, παρά μόνο αυτές που συμπεριλαμβάνουν την ομάδα που σχηματίστηκε. Τέλος, επαναλαμβάνονται οι διαδικασίες σχηματισμού ομάδων με βάση τις αποστάσεις μέχρι να ομαδοποιηθούν όλα τα δεδομένα σε μια και μόνο ομάδα [MPPD98].

Υπάρχουν αρκετοί αλγόριθμοι ιεραρχικής ομαδοποίησης που χρησιμοποιούνται σε δεδομένα μικροσυστοιχιών. Διαφέρουν στον τρόπο που υπολογίζονται οι αποστάσεις μεταξύ των αναπτυσσόμενων ομάδων και των εναπομεινάντων μελών του πακέτου των δεδομένων. Τέτοιοι αλγόριθμοι [J.01] είναι οι:

Ομαδοποίησης απλής συνδετικής διάταξης (single linkage clustering): η απόσταση μεταξύ δύο ομάδων, i και j, υπολογίζεται σαν η ελάχιστη απόσταση μεταξύ ενός μέλους της ομάδας i και ενός μέλους της ομάδας j. Η συγκεκριμένη τεχνική αναφέρεται και σαν μέθοδος του ελαχίστου ή σαν μέθοδος του κοντινότερου γείτονα. Παράγει «χαλαρές» ομάδες και δένδρα με πολλούς και μακρούς κλάδους.

Ομαδοποίησης ολικής συνδετικής διάταξης (complete linkage clustering): είναι γνωστή σαν μέθοδος του μεγίστου ή μέθοδος του απόμακρου γείτονα. Η απόσταση υπολογίζεται σαν η μέγιστη απόσταση μεταξύ των μελών των συσχετιζόμενων ομάδων. Παράγει πολύ συμπιεσμένες ομάδες που συχνά είναι όμοιες και στο μέγεθος.

Ομαδοποίησης μέσης συνδετικής διάταξης (average linkage clustering): η απόσταση που χρησιμοποιείται είναι η μέση απόσταση μεταξύ των μελών. Υπάρχουν πολλοί τρόποι να υπολογιστεί η μέση απόσταση. Ένας από αυτούς είναι η UPGMA που μετράει τις αποστάσεις μεταξύ ενός μέλους από μια ομάδα και όλων των μελών μιας άλλης και εξάγει τη μέση απόσταση.

Ένας ιδιαίτερα διαδεδομένος τρόπος ομαδοποίησης, είναι αυτός που βασίζεται στον αλγόριθμο k–means. Αν διαθέτουμε προχωρημένη γνώση για τον αριθμό των ομάδων που πρέπει να διαχωριστούν τα δεδομένα μας, ο αλγόριθμος k–means είναι μια καλή εναλλακτική πρόταση στις ιεραρχικές μεθόδους. Εδώ τα δεδομένα χωρίζονται σε προκαθορισμένο αριθμό (k) ομάδων, έτσι ώστε τα δεδομένα να είναι παρόμοια εντός μιας ομάδας και ανόμοια μεταξύ των ομάδων. Ο αλγόριθμος δεν παράγει δενδρογράμματα, αλλά κάποιος θα μπορούσε μετά το τέλος της διαδικασίας, να χρησιμοποιήσει ιεραρχική ταξινόμηση για τις ομάδες που προέκυψαν από τον k–means [J.01].

Η διαδικασία που ακολουθεί ο αλγόριθμος είναι απλή, αλλά μπορεί να γίνει και πολύ απαιτητική υπολογιστικά [PDJ+99]. Αρχικά, όλα τα δεδομένα τοποθετούνται τυχαία σε κάποια από τις k ομάδες που έχουμε ορίσει. Δεύτερον, υπολογίζεται ένα διάνυσμα μέσης έκφρασης για κάθε ομάδα και χρησιμοποιείται για να υπολογίσει τις αποστάσεις μεταξύ των ομάδων. Τρίτον, με χρήση επαναληπτικής μεθόδου, τα αντικείμενα μετακινούνται μεταξύ των ομάδων και επαναϋπολογίζονται οι αποστάσεις μεταξύ αντικειμένων και ομάδων αλλά και μεταξύ των ομάδων μετά από κάθε μετακίνηση.

Τα αντικείμενα επιτρέπεται να μετακινηθούν σε άλλη ομάδα μόνο όταν βρίσκονται πιο κοντά σε αυτήν από την προηγούμενη. Τέταρτον, μετά από κάθε μετακίνηση, τα διανύσματα έκφρασης για κάθε ομάδα επαναϋπολογίζονται. Τέλος, το ανακάτεμα συνεχίζεται μέχρι μια νέα μετακίνηση να χαλάει την ομοιογένεια των ομάδων.

2.2.6.4.2 Οπτικοποίηση: Αλγόριθμοι PCA, SAM και SVM.

Οπτικοποίηση είναι η τεχνική κατά την οποία δημιουργούνται γραφήματα που μας επιτρέπουν να εξάγουμε συμπεράσματα για τα δεδομένα μας. Ο ανθρώπινος νους μπορεί να αντιλαμβάνεται εικόνες σε τρισδιάστατο χώρο. Όταν τα δεδομένα μας ξεπερνούν τις τρεις διαστάσεις δε μπορούμε να κάνουμε διάκριση χαρακτηριστικών και προτύπων. Για το λόγο αυτό έχουν εξελιχθεί αλγόριθμοι που μας επιτρέπουν να αντιληφθούμε τα δεδομένα. Γίνεται, λοιπόν, ειδική μορφοποίηση και τελική παρουσίαση των δεδομένων και, πλέον, γινόμαστε ικανοί να διακρίνουμε στοιχεία που μας ενδιαφέρουν. Παρακάτω περιγράφονται μέθοδοι οπτικοποίησης που χρησιμοποιούνται συχνά κατά την ανάλυση δεδομένων μικροσυστοιχιών.

Ας μελετήσουμε καταρχήν τον αλγόριθμο PCA (principal component analysis – πρωταρχική ανάλυση συνιστωσών). Πολλές από τις τεχνικές μέτρησης, που χρησιμοποιούνται στις επιστήμες ζωής, συγκεντρώνουν δεδομένα για πάρα πολλές περισσότερες μεταβλητές ανά δείγμα, από το βασικό αριθμό των δειγμάτων. Οι μικροσυστοιχίες DNA μετρούν τα επίπεδα χιλιάδων mRNAs από εκατοντάδες δείγματα. Αυτή η πολυδιάστατη συλλογή δεδομένων κάνει δύσκολη την οπτικοποίηση των δειγμάτων και περιορίζει την απλή εξερεύνηση των δεδομένων [PD99].

Η πρωταρχική ανάλυση συνιστωσών είναι ένας μαθηματικός αλγόριθμος που μειώνει τις διαστάσεις των δεδομένων, ενώ διατηρεί το μεγαλύτερο μέρος της απόκλισης των δεδομένων. Το πετυχαίνει με το να αναγνωρίζει κατευθύνσεις που ονομάζονται πρωταρχικές συνιστώσες, κατά μήκος της μέγιστης απόκλισης. Χρησιμοποιώντας λίγες συνιστώσες, κάθε δείγμα αναπαρίσταται από σχετικά λίγα νούμερα σε σχέση με τις τιμές χιλιάδων μεταβλητών που έχουμε στην πραγματικότητα. Τα δείγματα τώρα, μπορούν να αναπαρασταθούν σε διάγραμμα και να μπορέσουμε οπτικά, να δούμε ομοιότητες και διαφορές μεταξύ τους, αλλά και που ομαδοποιούνται. Ας δούμε ένα παράδειγμα. Υποθέτουμε ότι μετράμε 10000 γο-

νίδια από 8 διαφορετικούς ασθενείς. Αυτές οι τιμές θα σχημάτιζαν ένα πίνακα 8 x 10000. Στη συνέχεια, αν τοποθετούσαμε τις τιμές αυτών των 10000 γονιδίων σε διάγραμμα θα χρειαζόμασταν 8 άξονες και θα προέκυπτε ένα νέφος σημείων οκτώ διαστάσεων, πράγμα που δε μπορούμε να το φανταστούμε, άρα ούτε να το δούμε. Για να εξαχθούν συμπεράσματα από τα δεδομένα, η πρωταρχική ανάλυση συνιστωσών εξάγει κατευθύνσεις όπου το νέφος μοιάζει πιο εκτεταμένο. Αν, για παράδειγμα, το νέφος έχει σχήμα μακρόστενου πεπονιού, η πρώτη συνιστώσα θα προκύψει από τον άξονα που βρίσκεται κατά μήκος του σχήματος. Ο αλγόριθμος, στη συνέχεια, θα αναζητήσει τη δεύτερη συνιστώσα με παρόμοιες ιδιότητες και με αυτόν τον τρόπο μειώνονται οι διαστάσεις του πολυδιάστατου νέφους σε δύο. Η δεύτερη συνιστώσα θα είναι κάθετη στην πρώτη.

Μπορούμε να μειώσουμε ακόμα περισσότερο τις διαστάσεις, αν προβάλουμε τα δείγματα στην πρώτη συνιστώσα. Αυτή η αναπαράσταση μιας διάστασης που προκύπτει μπορεί να θεωρηθεί σαν ένα γονιδιακό πρότυπο έκφρασης όλων των δειγμάτων. Έτσι για κάθε γονιδιακή συνιστώσα, ο αλγόριθμος αποκαλύπτει ένα ταιριαστό γονιδιακό πρότυπο το οποίο παρουσιάζει την ίδια απόκλιση με τη συνιστώσα, στα δεδομένα που διαθέτουμε.

Ακόμα, ένα πολύ σημαντικό στοιχείο του αλγορίθμου είναι ότι μπορεί να μειώσει τον αριθμό των διαστάσεων στο νούμερο του αριθμού των δειγμάτων, όταν τα δείγματα είναι λιγότερα από τις μεταβλητές, δηλαδή τα γονίδια [Μ.08].

Μπορεί κανείς, στη συνέχεια, να προβάλει τα δείγματά του σε συγκεκριμένες από τις συνιστώσες που έχουν οριστεί για να ελέγξει αν οι αποκλίσεις που προκύπτουν εμπεριέχουν σχετικά ενδιαφέρουσες πληροφορίες. Ο αλγόριθμος, όμως, πολύ δύσκολα προχωρά σε ομαδοποίηση γιατί είναι σχεδιασμένος να ανακαλύπτει τις κατευθύνσεις με τη μεγαλύτερη απόκλιση και όχι κατευθύνσεις διαχωρισμού των δεδομένων. Πρέπει πάντα να έχουμε στο μυαλό ότι μεγάλο ποσοστό αποκλίσεων οφείλεται σε συστηματικά λανθασμένους πειραματικούς χειρισμούς, πράγμα που ενδέχεται να προκαλέσει μεγάλο πρόβλημα κατά την πρωταρχική ανάλυση συνιστωσών, δίνοντας κυρίαρχες συνιστώσες που σχετίζονται με σφάλματα.

Μια προφανής εφαρμογή του αλγορίθμου είναι να εξερευνά δεδομένα πολλών διαστάσεων. Τις περισσότερες φορές, χρησιμοποιείται τρισδιάστατη οπτικοποίηση και τα δείγματα είτε προβάλλονται στις συνιστώσες είτε μπαίνουν σε διαγράμματα σύμφωνα με η συσχέτιση με τις συνιστώσες. Η απόφαση για το πόσες και ποιες θα είναι οι συνιστώσες που θα οριστούν είναι μεγάλη πρόκληση και καθορίζεται με διάφορους τρόπους. Για παράδειγμα, κάποιος μπορεί να χρησιμοποιήσει συνιστώσες που σχετίζονται με ένα φαινότυπο που τον ενδιαφέρει ή να χρησιμοποιήσει πολλές συνιστώσες, ώστε να συμπεριλάβει τις περισσότερες από τις αποκλίσεις στα δεδομένα. Γενικά όμως, τα αποτελέσματα της πρωταρχικής ανάλυσης συνιστωσών εξαρτώνται από την προεπεξεργασία των δεδομένων και την επιλογή των μεταβλητών.

Γενικά, μπορούμε να φανταστούμε τη χρησιμότητα του αλγορίθμου ως εξής:

αν είχαμε ένα τρισδιάστατο σύννεφο από κουκίδες και μπορούσαμε να το πάρουμε στα χέρια μας και το μετακινούσαμε σε διάφορες κατευθύνσεις, ίσως να μπορούσαμε σε κάποια από τις πολλές όψεις που θα παίρναμε, να δούμε ξεχωριστές ομάδες δεδομένων. Η PCA μπορεί να βοηθήσει στο να καθορίσουμε, περίπου, τον αριθμό των ομάδων που χρειάζονται αλγόριθμοι όπως οι k–means, SOMs.

Η SAM (Significance Analysis of Microarrays) από την άλλη μεριά, είναι μια στατιστική μέθοδος που αναγνωρίζει γονίδια με στατιστικά σημαντικές αλλαγές στην έκφραση συνταιριάζοντας ένα σύνολο t–tests που αφορούν συγκεκριμένα γονίδια. Κάθε γονίδιο παίρνει μια συγκεκριμένη τιμή που βασίζεται στην αλλαγή της έκφρασής του σχετικά με την τυπική απόκλιση επαναλαμβανόμενων μετρήσεων για το γονίδιο. Τα γονίδια με τιμές μεγαλύτερες από ένα δεδομένο όριο καθορίζονται ως δυναμικώς σημαντικά. Το ποσοστό αυτών των γονιδίων που θα προκύψουν τυχαία αποτελεί το δείκτη λανθασμένης ανακάλυψης (FDR), ο οποίος προσδιορίζεται από ανάλυση πολλαπλών συνδυασμών των μετρήσεων [IEJ+02].

Τέλος, οι SVM (Support Vector Machine) είναι αλγόριθμοι οι οποίοι μπορούν να μαθαίνουν από παραδείγματα και εκπαίδευση να αναγνωρίζουν και να κατηγοριοποιούν αντικείμενα [WS.06]. Μια βασική βιοϊατρική τους εφαρμογή είναι να κατηγοριοποιούν δεδομένα γονιδιακής έκφρασης από πειράματα μικροσυστοιχιών. Αποτελούν επιβλεπόμενους αλγορίθμους οι οποίοι χρησιμοποιούν έναν αρχικό αριθμό δεδομένων για εκπαίδευση της συνάρτησής τους και, στη συνέχεια, είναι σε θέση να αναγνωρίσουν και να ξεχωρίσουν τα υπόλοιπα δεδομένα με βάση την έκφρασή τους. Με λίγα λόγια, ένα βασικό τους πλεονέκτημα είναι ότι χρησιμοποιούν βιολογικές πληροφορίες για να καθορίσουν τα στοιχεία έκφρασης που χρειάζονται για να χαρακτηρίσουν τις ομάδες και, μετά, είναι ικανοί να κατηγοριοποιήσουν οποιοδήποτε άλλο αντικείμενο.

Ένα βασικό τους εργαλείο είναι το διαχωριστικό υπερεπίπεδο (separating hyperplane) το οποίο διαχωρίζει τα δεδομένα με βάση τα διαφορετικά πρότυπα έκφρασης. Ακόμα ένα εργαλείο είναι το υπερεπίπεδο μέγιστου περιθωρίου (maximum – margin hyperplane). Είναι στοιχείο που βοηθά στην επιλογή του καταλληλότερου διαχωριστικού υπερεπιπέδου καθώς τα υπερεπίπεδα που διαχωρίζουν επαρκώς μπορεί να είναι πολλά.

Επιλέγεται, συνήθως, αυτό με τη μέση απόσταση μεταξύ των δεδομένων των κλάσεων. Ένα ακόμα βοηθητικό στοιχείο είναι το χαλαρό περιθώριο (soft margin). Χρησιμοποιείται όταν ο διαχωρισμός δεν είναι τόσο καθαρός για να συντελεστεί από ένα υπερεπίπεδο και στοιχεία παραμένουν εκατέρωθεν του υπερεπιπέδου ενώ ανήκουν σε άλλη ομάδα. Ο χρήστης μπορεί να καθορίσει τη χαλαρότητα των περιθωρίων που δίνει για να πάρει έστω και έναν όχι τόσο αυστηρό διαχωρισμό.

Πολλές φορές ένας διαχωρισμός είναι ευκολότερος όταν αυξήσουμε τις διαστάσεις και δούμε τα δεδομένα σε έναν διαφορετικό χώρο περισσότερων διαστάσεων. Αυτό επιτυγχάνεται με χρήση της συνάρτησης Kernel, αλλά πολλές φορές προκύπτουν προβλήματα από την «κατάρα των πολλών διαστάσεων», ζήτημα που

βασανίζει τους μαθηματικούς.

Έχοντας ολοκληρώσει την διατύπωση των βασικών χαρακτηριστικών που αφορούν στις μικροσυστοιχίες DNA, και έχοντας κάνει μια αναφορά σε μεθόδους επεξεργασίας των δεδομένων αυτών, στο επόμενο κεφάλαιο θα παρουσιαστούν τα βασικά χαρακτηριστικά μεθόδων Υπολογιστικής Νοημοσύνης και των αντίστοιχων τεχνικών.

Κεφάλαιο 3

Υπολογιστική Νοημοσύνη και Τεχνικές

Έχοντας μελετήσει τα βασικά χαρακτηριστικά των μικροσυστοιχιών DNA, στα πλαίσια αυτού του κεφαλαίου θα δούμε μερικά βασικά χαρακτηριστικά που αφορούν στην Υπολογιστική Νοημοσύνη, καθώς και μερικές συγκεκριμένες σχετικές τεχνικές όπως αυτή του αλγορίθμου k–means και άλλων.

Έτσι, θα γίνει πρώτα μια αναφορά σε βασικές έννοιες της κλασικής Τεχνιτής Νοημοσύνης, όπως είναι τα προβλήματα που καλείται να αντιμετωπίσει και οι τρόποι με τους οποίους τα αντιμετωπίζει με τη βοήθεια αλγορίθμων αναζήτησης, για τους οποίους θα αναφερθούν αρκετά παραδείγματα. Στη συνέχεια θα δούμε χαρακτηριστικά παραδείγματα μεθόδων Υπολογιστικής Νοημοσύνης. Τέλος, θα γίνει πιο συγκεκριμένη αναφορά στον αλγόριθμο k-means αλλά και σε άλλους παρόμοιους αλγορίθμους ομαδοποίησης.

3.1 Προβλήματα

Ένα πρόβλημα είναι ένα σύνολο αντικειμένων, ιδιοτήτων και σχέσεων το οποίο ορίζεται από μία αρχική κατάσταση, μία επιθυμητή τελική κατάσταση και τις επιτρεπτές ενέργειες στα αντικείμενα του προβλήματος. Στόχος είναι, ξεκινώντας από την αρχική κατάσταση, να γίνει μία κατάλληλη ακολουθία ενεργειών η οποία να καταλήγει στην τελική κατάσταση. Αυτή η διαδικασία ονομάζεται επίλυση του προβλήματος (π.χ. η διεξαγωγή μίας παρτίδας σκάκι) και αποτέλεσε στόχο της ΤΝ από τη δεκαετία του '50. Η επίλυση προβλημάτων έχει προφανώς σπουδαίες εφαρμογές στην απόδειξη θεωρημάτων, στον χρονοπρογραμματισμό ενεργειών, στη διεξαγωγή παιγνίων κλπ, ενώ θεωρείται κεντρικό χαρακτηριστικό της ευφυΐας [W.S79]. Βασικό στοιχείο στην επίλυση προβλημάτων είναι η αναπαράσταση τους και γι' αυτό το σκοπό υπάρχουν δύο μεθοδολογίες: η αναπαράσταση με χώρο

καταστάσεων και η αναπαράσταση με αναγωγή. Στη μέθοδο της αναγωγής δομική μονάδα περιγραφής του προβλήματος είναι η ίδια η περιγραφή αναλυόμενη σε πολλαπλές, απλούστερες εκδοχές. Αυτή η ανάλυση συμβαίνει διαδοχικά ώσπου να καταλήξει σε αρχέγονα προβλήματα επιλυόμενα με προφανή τρόπο. Προγραμματιστικά η αναγωγή υλοποιείται με αναδρομή και κεντρική έννοια σε αυτήν αποτελούν οι τελεστές αναγωγής, διαδικασίες οι οποίες ανάγουν ένα πρόβλημα σε υποπροβλήματα.

Στη μέθοδο χώρου καταστάσεων βασική δομική μονάδα είναι η κατάσταση, το σύνολο δηλαδή των αντικειμένων που εμπλέκονται στο πρόβλημα συν τις ιδιότητες τους και τις μεταξύ τους σχέσεις. Η κατάσταση ορίζεται σε ένα απλουστευμένο, αφαιρετικό μοντέλο του κόσμου και το σύνολο των καταστάσεων (στιγμιοτύπων) στις οποίες μπορεί να βρεθεί αυτός ο κόσμος του προβλήματος ονομάζεται χώρος καταστάσεων. Το ίδιο το πρόβλημα ορίζεται με βάση την αρχική κατάσταση από την οποία ξεκινάμε, την επιθυμητή τελική κατάσταση στην οποία πρέπει να καταλήξουμε (ή πολλές δυνατές τελικές καταστάσεις) και το σύνολο των τελεστών μετάβασης, δηλαδή επιτρεπτών πράξεων, που μπορούν να εκτελεστούν στα αντικείμενα μίας κατάστασης οδηγώντας σε μια άλλη (π.χ. στην αναπαράσταση μίας παρτίδας σκάκι τελεστής είναι η έγκυρη μετακίνηση ενός πιονιού). Λύση του προβλήματος είναι μία ακολουθία διαδοχικών τελεστών μετάβασης και καταστάσεων που ξεκινά από μία αρχική κατάσταση και καταλήγει σε μία τελική.

Ο πιο κατάλληλος τρόπος γραφικής αναπαράστασης του προβλήματος είναι ένα δένδρο με ρίζα την αρχική κατάσταση, φύλλα τις τελικές καταστάσεις και τα αδιέξοδα, κόμβους τις ενδιάμεσες καταστάσεις και κλαδιά τους τελεστές μετάβασης. Επέκταση ενός κόμβου ονομάζεται η εύρεση όλων των παιδιών του στο δένδρο μέσω της εφαρμογής σε αυτόν όλων των πιθανών τελεστών. Οι λύσεις του προβλήματος είναι μονοπάτια από τη ρίζα σε κάποιο φύλλο του δένδρου που αντιστοιχεί σε τελική κατάσταση. Σε πραγματικά προβλήματα το μέγεθος αυτού του δένδρου γίνεται εξαιρετικά μεγάλο μετά την επέκταση λίγων μόλις κόμβων και επομένως η αναζήτηση λύσεων σε ένα τέτοιο δένδρο καθίσταται εξαιρετικά χρονοβόρα. Αυτό το ζήτημα στη βιβλιογραφία της ΤΝ αναφέρεται ως συνδυαστική έκρηξη.

3.2 Αλγόριθμοι αναζήτησης

Υπάρχει μία πλειάδα πραγματικών ή συνθετικών, απλών ή πολύπλοκων προβλημάτων που μπορούν να αναπαρασταθούν με χώρο καταστάσεων. Όλα τα προηγούμενα βρίσκουν εφαρμογή και το μόνο που αλλάζει σε κάθε πρόβλημα είναι οι λεπτομέρειες (οι ιδιότητες των αντικειμένων, οι επιτρεπτοί τελεστές κλπ). Προκειμένου ένα πρόγραμμα να επιλύσει ένα τέτοιο πρόβλημα πρέπει να αναπαραστήσει κατάλληλα και να κατασκευάσει το δένδρο των καταστάσεων, ξεκινώντας από τη

ρίζα και επεκτείνοντας τους κόμβους μέχρι να φτάσει σε κάποια τελική κατάσταση. Αν το ζητούμενο είναι να βρεθεί μία οποιαδήποτε λύση τότε το πρόγραμμα μπορεί τότε να τερματίσει επιστρέφοντας το μονοπάτι που οδηγεί στο τρέχον φύλλο, διαφορετικά (εξαντλητική αναζήτηση) μπορεί να αποθηκεύσει έναν δείκτη προς αυτό το φύλλο και να συνεχίσει την κατασκευή του δένδρου μέχρι να ανακαλύψει όλες τις πιθανές καταστάσεις που είναι προσβάσιμες από την αρχική, με τους διαθέσιμους τελεστές μετάβασης, και όλες τις πιθανές λύσεις.

Υπάρχει ένας γενικός αλγόριθμος αναζήτησης που εκτελεί αυτήν τη διερεύνηση και οι πραγματικοί αλγόριθμοι που χρησιμοποιούνται είναι παραλλαγές του που διαφέρουν στα βήματα 7 και 8. Στον αλγόριθμο αυτό Μέτωπο Αναζήτησης (Μ.Α.) είναι το σύνολο των καταστάσεων που έχουμε επισκεφθεί αλλά δεν έχουμε επεκτείνει και Κλειστό Σύνολο (Κ.Σ.) το σύνολο των καταστάσεων που και έχουμε επισκεφθεί και έχουμε επεκτείνει. Το Κ.Σ. είναι απαραίτητο μόνο αν υπάρχει κίνδυνος παγίδευσης του αλγορίθμου σε ατέρμονα βρόχο λόγω απείρου μήκους κλαδιών στο δένδρο [Koh92].

Algorithm 1 Ο αλγόριθμος Μετώπου Αναζήτησης Μ.Α

```
1:  M.A. = NULL
2:  Κ.Σ. = NULL
3:  Εισαγωγή της ρίζας στο Μ.Α.
4:  while Κεφαλή του Μ.Α. ≠ NULL do
5:    Κ = Κεφαλή του Μ.Α.
6:    if Κ περιέχεται στο Κ.Σ. then
7:      goto 4
8:    end if
9:    if Κ είναι τελική κατάσταση then
10:     return Κ,μη εξαντλητική αναζήτηση ή
11:     insertToSolutions Κ,εξαντλητική αναζήτηση then goto 4
12:   end if
13:   Επέκταση του Κ, Εισαγωγή των παιδιών του στο Μ.Α., Εισαγωγή του Κ στο
      Κ.Σ.
14:   Αφαίρεση κάποιων καταστάσεων από το Μ.Α. με κάποιο κριτήριο,
15:   κλάδεμα του δένδρου, γίνεται για εξοικονόμηση χρόνου όταν θεωρείται
      απίθανο το υποδένδρο που ξεκινά από κάποια κατάσταση να οδηγεί σε λύση
16:   Αναδιοργάνωση του Μ.Α. με κάποιο κριτήριο
17: end while
```

Οι διάφορες πραγματικές παραλλαγές αυτού του αλγορίθμου διακρίνονται σε αλγορίθμους τυφλής αναζήτησης, που διατάσσουν το Μ.Α. αποκλειστικά με βάση το χρόνο δημιουργίας κάθε κόμβου κατά την κατασκευή του δένδρου, και σε αλγορίθμους ευρετικής αναζήτησης (heuristic search), όπου τα βήματα 7 και 8 εξαρτώ-

νται από μία επιπλέον πληροφορία που υπολογίζεται σε πραγματικό χρόνο και που στις περισσότερες περιπτώσεις, αλλά όχι πάντα, είναι σχετικά ακριβής και αξιολογεί προσεγγιστικά τις καταστάσεις σε «καλές» και «κακές». Ένα παράδειγμα ευρετικής πληροφορίας που μπορεί να αντιστοιχιστεί σε κάθε ενδιάμεση κατάσταση είναι η εκτιμώμενη «απόστασή» της (με βάση ένα μέτρο που εξαρτάται από το πρόβλημα και την υλοποίηση) από την τελική. Έτσι μπορούμε, φερ' ειπείν, να κλαδεύουμε τα υποδένδρα με ρίζα «κακή» κατάσταση, αφαιρώντας τη ρίζα τους από το Μ.Α. προτού την επεκτείνουμε (βήμα 7). Προφανώς αυτή η τακτική συμβάλλει στην αντιμετώπιση του φαινομένου της συνδυαστικής έκρηξης.

Μία άλλη κατηγοριοποίηση των αλγορίθμων γίνεται ανάλογα με τον τύπο του προβλήματος που επιλύουν: εκτός από τα συνηθισμένα που προαναφέρθηκαν, υπάρχουν και προβλήματα βελτιστοποίησης (όπου σε κάθε τελεστή μετάβασης αντιστοιχίζεται μία τιμή κόστους και αναζητούμε τη λύση με το μονοπάτι που αθροιστικά έχει το ελάχιστο κόστος) ή προβλήματα ικανοποίησης περιορισμών (όπου η τελική κατάσταση δεν είναι πλήρως γνωστή, γνωρίζουμε όμως κάποιες ιδιότητες της και επιθυμούμε να καταλήξουμε σε μία κατάσταση που να τις διαθέτει). Πληρότητα ενός αλγορίθμου αναζήτησης ονομάζεται το κατά πόσον βρίσκει πάντα μία λύση, εφ' όσον τέτοια υπάρχει.

3.2.1 Τυφλή αναζήτηση

Οι σπουδαιότεροι αλγόριθμοι τυφλής αναζήτησης είναι ο DFS (Depth-First Search ή αναζήτηση κατά βάθος) και ο BFS (Breadth-First Search ή αναζήτηση κατά πλάτος), οι οποίοι κατασκευάζουν το δένδρο ξεκινώντας από τη ρίζα και παράγοντας κόμβους, ο μεν DFS (ακολουθεί ένα κλαδί μέχρι να φτάσει σε φύλλο και μετά επεκτείνει έναν κόμβο προηγούμενου επιπέδου; αυτή η μέθοδος ονομάζεται «οπισθοδρόμηση»), ο δε BFS (επεκτείνει πρώτα όλους τους κόμβους ενός επιπέδου, οι οποίοι έχουν το ίδιο βάθος, και μετά προχωρά στους κόμβους του επόμενου επιπέδου). Προγραμματιστικά είναι σχεδόν ίδιοι μεταξύ τους, αλλά και με το γενικό αλγόριθμο που περιγράφηκε προηγουμένως, μόνο που διαφέρουν στο βήμα 8 (το βήμα 7 δεν υπάρχει αφού δε γίνεται κλάδεμα): ο DFS τοποθετεί τους νέους κόμβους που προστίθενται στο Μ.Α. στην αρχή της λίστας (LIFO στοίβα), ώστε στην επόμενη επανάληψη του βρόχου να επεκταθεί ένας από αυτούς, ενώ ο BFS τους τοποθετεί στο τέλος της λίστας (FIFO ουρά), ώστε στην επόμενη επανάληψη του βρόχου να επεκταθεί ένας «αδελφός» του γονέα τους αν υπάρχει.

Ο BFS εγγυάται ότι θα βρει πρώτα τη λύση με την ελάχιστη απόσταση από τη ρίζα (οπότε είναι ιδανικός και για προβλήματα βελτιστοποίησης όπου όλοι οι τελεστές έχουν ίσο κόστος) και είναι πλήρης, το Μ.Α. όμως μπορεί να γιγαντωθεί για μεγάλους χώρους αναζήτησης και άρα έχει μεγάλες απαιτήσεις σε μνήμη. Από την άλλη ο DFS είναι τυχαίο το ποια λύση θα βρει πρώτα και δεν είναι πλήρης, καθώς αν δε χρησιμοποιείται Κλειστό Σύνολο μπορεί να παγιδευτεί σε κλαδιά απείρου

μήκους (αφού ακολουθεί ένα κλαδί μέχρι να καταλήξει σε φύλλο). Από την άλλη έχει μικρές απαιτήσεις σε μνήμη διατηρώντας πάντα μικρό το Μ.Α.

Συμβιβασμό μεταξύ αυτών των δύο αποτελεί ο αλγόριθμος ID (Iterative Deepening ή επαναληπτική εκβάθυνση), ο οποίος είναι κατά βάση DFS αλλά προχωρά μέχρι ένα προκαθορισμένο βάθος, ενώ στη συνέχεια το επιτρεπτό βάθος αυξάνεται και ο αλγόριθμος ξεκινά από την αρχή χωρίς να διατηρεί δεδομένα από την προηγούμενη αναζήτηση. Το δένδρο δηλαδή κατασκευάζεται διαρκώς από τη ρίζα, ξανά και ξανά, αλλά σε όλο και μεγαλύτερο βάθος. Παρ' όλο που ο ID εκτελεί πολλή περιττή εργασία αυτό δεν παίζει ρόλο σε μεγάλους χώρους αναζήτησης όσον αφορά την αλγοριθμική πολυπλοκότητα. Ο αλγόριθμος είναι πλήρης γιατί δεν μπορεί να παγιδευτεί σε άπειρα κλαδιά, αφού το βάθος αναζήτησης είναι προκαθορισμένο, έχει τις μικρές απαιτήσεις μνήμης του DFS, ενώ αν το επιτρεπτό βάθος σε κάθε επανάληψη αυξάνεται κατά 1 εγγυάται ότι θα βρει πρώτα τη λύση με την ελάχιστη απόσταση από τη ρίζα (όπως ο BFS, αφού αν υπήρχε καλύτερη λύση θα βρισκόταν σε προηγούμενη επανάληψη).

Οποιοσδήποτε από αυτούς τους αλγορίθμους μπορεί να χρησιμοποιηθεί με τη μέθοδο BiS (Bidirectional Search ή αμφίδρομη αναζήτηση), η οποία μπορεί να εφαρμοστεί σε υπολογιστικό σύστημα με δύο επεξεργαστές όταν η τελική κατάσταση είναι πλήρως γνωστή και οι τελεστές μετάβασης είναι αντιστρέψιμοι: ο ένας επεξεργαστής εκτελεί αναζήτηση από την αρχική προς την τελική κατάσταση και ο άλλος από την τελική προς την αρχική. Όταν βρεθεί μία κοινή κατάσταση το πρόγραμμα ενώνει τα δύο μονοπάτια και επιστρέφει την τελική λύση; ιδανικά στο 1/2 του χρόνου που θα απαιτούσε μία μονόδρομη αναζήτηση.

Σε προβλήματα βελτιστοποίησης με τελεστές διαφορετικού (αλλά πάντα θετικού) κόστους μπορεί να εφαρμοστεί ο αλγόριθμος τυφλής αναζήτησης B&B (Branch and Bound ή επέκταση και οριοθέτηση), ο οποίος μπορεί να βασιστεί είτε στον DFS είτε στον BFS προσφέροντας όμως επιπλέον κλάδεμα των καταστάσεων -και των αντίστοιχων υποδένδρων που θα προέκυπταν από την επέκταση τους- που αποκλείεται να οδηγούν σε λύση καλύτερη από την τρέχουσα. Για να το πετύχει αυτό κρατά σε μία μεταβλητή Β το ολικό κόστος του μονοπατιού της βέλτιστης λύσης που έχει βρεθεί ως τώρα και, αν το μονοπάτι του τρέχοντος ενδιάμεσου κόμβου έχει κόστος μεγαλύτερο του Β, δεν τον αναπτύσσει και τον αφαιρεί από το Μέτωπο Αναζήτησης. Στη χειρότερη περίπτωση δε θα γίνει κανένα κλάδεμα, αφού είναι θέμα τύχης η σειρά με την οποία θα ανακαλυφθούν οι λύσεις, και ο B&B λειτουργεί όπως ο DFS ή ο BFS.

3.2.2 Ευρετική αναζήτηση

Προκειμένου να μειωθεί ο γιγάντιος για ρεαλιστικά προβλήματα, χώρος αναζήτησης και ο απαιτούμενος για την εύρεση της λύσης χρόνος, μπορούν να χρησιμοποιηθούν αλγόριθμοι που εκμεταλλεύονται ευρετικούς μηχανισμούς, δηλαδή

στρατηγικές (συνήθως συναρτήσεις που εξαρτώνται από το εκάστοτε πρόβλημα) οι οποίες αξιολογούν προσεγγιστικά τις ενδιάμεσες καταστάσεις ως προς την εκτιμώμενη απόσταση τους από μία τελική κατάσταση, επεκτείνουν πρώτα αυτές με τη βέλτιστη ευρετική τιμή (οι οποίες αναμένεται να οδηγήσουν συντομότερα σε λύση) ή/και κλαδεύουν τις υπόλοιπες. Οι ευρετικοί μηχανισμοί δεν είναι αντικειμενικοί και, παρόλο που κωδικοποιούνται αλγοριθμικά υπό τη μορφή της ευρετικής συνάρτησης, δεν μπορούν να θεωρηθούν αλγόριθμοι. Αυτό γιατί, προκειμένου να μειώσουν το χώρο αναζήτησης ή να επιταχύνουν την εύρεση της λύσης, λειτουργούν προσεγγιστικά και «διαισθητικά» (περίπου όπως οι άνθρωποι), ενώ οι αλγόριθμοι είναι ακριβείς και λειτουργούν πάντα ορθά. Στην πλειονότητα των περιπτώσεων πάντως οι ευρετικές στρατηγικές οδηγούν σε πολύ καλά αποτελέσματα (αναλόγως βέβαια του προβλήματος), ωστόσο απέχουν πολύ από το να προσομοιώνουν τους μηχανισμούς της ανθρώπινης σκέψης: η τελευταία χρησιμοποιεί επίσης ευρετικές μεθόδους οι οποίες όμως είναι ποιοτικές, όχι ποσοτικές / αριθμητικές όπως η ευρετική συνάρτηση, και φαίνεται να αποδίδουν καλύτερα [SK06].

Ένας βασικός ευρετικός αλγόριθμος είναι ο HC (Hill Climbing ή αναρρίχηση λόφων), ο οποίος μοιάζει με τον DFS αλλά σε κάθε επανάληψη κλαδεύει όλες τις καταστάσεις που προκύπτουν από μία επέκταση εκτός από την ευρετικά βέλτιστη (δηλαδή κάθε στιγμή το Μ.Α. έχει μία κατάσταση) και μεταβαίνει στην τελευταία μόνο αν έχει καλύτερη ευρετική τιμή από το γονέα της; διαφορετικά τερματίζει έχοντας βρει μία τοπικά βέλτιστη λύση. Προφανώς ο HC δεν είναι πλήρης αλλά είναι πολύ γρήγορος και καθόλου μνημοβόρος. Υπάρχουν διάφορες παραλλαγές του που θυσιάζουν λίγη από την ταχύτητα του προκειμένου να αυξήσουν την πιθανότητα του να βρει λύση. Μία παραλλαγή είναι ο EHC (Enforced Hill Climbing ή εξαναγκασμένη αναρρίχηση λόφων), στον οποίον διατηρούνται στο Μ.Α. τα αδέρφια του τρέχοντος κόμβου και, αν η επέκταση του τελευταίου δεν οδηγήσει σε μετάβαση, αντί ο αλγόριθμος να τερματίσει εκτελεί μία αναζήτηση κατά πλάτος στα αδέρφια του μέχρι να βρεθεί μία καλύτερη κατάσταση οπότε και συνεχίζεται η αναρρίχηση από εκεί. Επίσης δημοφιλής είναι και ο SA (Simulated Annealing ή προσομοιωμένη ανόπτηση), ο οποίος δίνει μία πιθανότητα μετάβασης σε χειρότερες καταστάσεις (p), αφήνοντας έτσι περιθώριο στην αναζήτηση να ξεφύγει από τοπικά βέλτιστα. Αν η πιθανότητα p τείνει στο 0 ο SA λειτουργεί όπως ο HC. Επίσης υπάρχει ο TS (Taboo Search ή αναζήτηση με απαγορεύσεις), όπου σε κάθε επέκταση γίνεται πάντα μετάβαση στο καλύτερο παιδί, ακόμα και αν είναι χειρότερη κατάσταση από την τρέχουσα, και η αναζήτηση συμβουλεύεται μία λίστα απαγορευμένων καταστάσεων (παρόμοιας λειτουργικότητας με το Κλειστό Σύνολο αλλά σταθερού μεγέθους). Ο BS (Beam Search ή ακτινωτή αναζήτηση), όπου ένας σταθερός αριθμός εκ των καλύτερων καταστάσεων παραμένει στο Μ.Α. δίνοντας τη δυνατότητα οπισθοδρόμησης αν χρειαστεί, είναι μία ακόμα επέκταση του κεντρικού αλγορίθμου αναρρίχησης λόφων.

Άλλος δημοφιλής ευρετικός αλγόριθμος είναι ο BestFS (αναζήτηση πρώτα στο

καλύτερο) ο οποίος κρατά όλες τις καταστάσεις στο Μ.Α. και μοιάζει με τον BFS, μόνο που σε κάθε επέκταση εφαρμόζει τον ευρετικό μηχανισμό και στην επόμενη επανάληψη μεταβαίνει στο ευρετικά βέλτιστο παιδί. Είναι πλήρης, μνημοβόρος και δεν εγγυάται ότι θα βρει τη βέλτιστη λύση αφού εξαρτάται απόλυτα από την εγκυρότητα των εκτιμήσεων της ευρετικής συνάρτησης. Τροποποίηση του BestFS αποτελεί ο πλήρης και βέλτιστος αλγόριθμος A*, στον οποίον η ευρετική τιμή που αντιστοιχίζεται σε κάθε νέα κατάσταση Κ για να την αξιολογήσει ο μηχανισμός δεν είναι μόνο μία εκτίμηση Α της απόστασης της από μία τελική κατάσταση, αλλά το άθροισμα Α συν την ακριβή απόσταση της Κ από τη ρίζα. Ο A* εγγυάται ότι θα βρει τη βέλτιστη λύση αρκεί η ευρετική συνάρτηση να είναι πάντα υποεκτίμηση της πραγματικής απόστασης από τη λύση και ποτέ υπερεκτίμηση («αποδεκτή συνάρτηση»). Σε περίπτωση που είναι σπουδαιότερη η ταχύτητα παρά η βελτιστότητα δε χρειάζεται η ευρετική συνάρτηση να είναι αποδεκτή. Παραλλαγή του A* αποτελεί ο IDA* (A* με επαναληπτική εκβάθυνση) ο οποίος αναπτύσσει το δένδρο αναζήτησης κατά βάθος σε διαδοχικές επαναλήψεις, αξιοποιώντας την ευρετική συνάρτηση του A* για να επιλέξει την επεκτεινόμενη κάθε φορά κατάσταση, αλλά όταν η ευρετική τιμή μίας νέας κατάστασης ξεπερνά το όριο που έχει τεθεί για την τρέχουσα επανάληψη όλο το υποδένδρο το οποίο ξεκινά από αυτήν κλαδεύεται. Στην επόμενη επανάληψη, όπου όπως στον ID το δένδρο αναζήτησης κατασκευάζεται από την αρχή, το νέο ευρετικό όριο τίθεται στη μικρότερη τιμή που εμφανίστηκε στις καταστάσεις οι οποίες κλαδεύτηκαν κατά την προηγούμενη επανάληψη.

Όλοι αυτοί οι ευρετικοί αλγόριθμοι κατάγονται από τη θεωρία μαθηματικής βελτιστοποίησης, όπου αναπτύχθηκαν για να εντοπίζουν το ελάχιστο ή το μέγιστο μίας πραγματικής συνάρτησης διακριτής μεταβλητής. Στην επίλυση προβλημάτων τον ρόλο της τελευταίας προφανώς τον παίζει η ευρετική συνάρτηση και ο χώρος των λύσεων οπτικοποιείται ως ένα γεωγραφικό «τοπίο»: όσο περισσότερο δύο λύσεις διαφέρουν τόσο απέχουν μεταξύ τους σε αυτό το τοπίο, ενώ όσο καλύτερη ευρετική τιμή έχει μία λύση τόσο υψηλότερα από το επίπεδο του «εδάφους» τοποθετείται σε αυτό το τοπίο. Το τελευταίο, καθώς οι υποψήφιες καταστάσεις είναι διακριτές μεταξύ τους, ουσιαστικά είναι ένας γράφος με κορυφές τις καταστάσεις και ακμές τους τελεστές μετάβασης. Η παραλλαγή της αναρρίχησης λόφων σε συνεχή χώρο, με στόχο την εύρεση ακρότατου μιας συνάρτησης συνεχούς μεταβλητής, ονομάζεται μέθοδος μέγιστης ανόδου (gradient ascent, αν η συνάρτηση εκφράζει βελτιστότητα και αναζητείται το μέγιστό της) ή μέθοδος μέγιστης καθόδου (gradient descent, αν η συνάρτηση εκφράζει σφάλμα / απόκλιση από το βέλτιστο και αναζητείται το ελάχιστό της) υλοποιείται με μεθόδους του απειροστικού λογισμού και απαιτεί το διάνυσμα μερικών παραγώγων της αντικειμενικής συνάρτησης.

3.2.3 Παίγνια

Μία ειδική κατηγορία προβλημάτων ΤΝ είναι τα παίγνια δύο αντιπάλων (π.χ. σκάκι). Εδώ υπάρχουν δύο διαφορετικά σύνολα τελεστών μετάβασης που εφαρμόζονται εναλλάξ από δύο ανταγωνιστικά ενεργά συστήματα (τους παίκτες). Οι τελικές καταστάσεις δεν είναι πλήρως γνωστές, έχουν όμως συγκεκριμένα γνωστά χαρακτηριστικά και αντιστοιχούν σε ισοπαλία ή σε νίκη ενός εκ των δύο αντιπάλων / ήττα του άλλου. Στα προβλήματα αυτά ξεκινώντας από μία κατάσταση μπορούμε με συνεχείς επεκτάσεις να δημιουργήσουμε το δένδρο του παιχνιδιού, με όλα τα εναλλακτικά μονοπάτια που πηγάζουν από την τρέχουσα κατάσταση, στο οποίο οι κινήσεις δύο διαδοχικών επιπέδων (όπου κίνηση είναι ένα κλαδί του δένδρου, δηλαδή η μετάβαση από μία κατάσταση σε άλλη) αντιστοιχούν σε διαφορετικό παίκτη. Η ανάπτυξη αυτού του δένδρου μπορεί να γίνει πριν την πραγματοποίηση μια κίνησης μέχρι κάποιο βάθος (όσο μεγαλύτερο είναι αυτό το βάθος τόσο καλύτερα αποδίδει ο αλγόριθμος). Σε αυτό το σημείο ο παίκτης μπορεί να αξιολογήσει ευρετικά ποια από τις εναλλακτικές κινήσεις οδηγεί στην ευνοϊκότερη γι' αυτόν εξέλιξη. Στα περισσότερα πραγματικά παίγνια το βάθος του κατασκευαζόμενου σε κάθε εκτέλεση δένδρου δεν μπορεί να είναι πολύ μεγάλο λόγω του προβλήματος της συνδυαστική έκρηξης.

Ο γνωστότερος αλγόριθμος για επίλυση παιγνίων είναι ο Minimax (αναζήτηση ελαχίστου-μεγίστου), ο οποίος καλείται να αποφασίσει ποια θα είναι η επόμενη κίνηση του εναντίον του αντιπάλου. Κάθε φορά που εκτελείται αναπτύσσει το δένδρο του παιχνιδιού με ρίζα την τρέχουσα κατάσταση και ως ένα προκαθορισμένο βάθος, στη συνέχεια αποδίδει στα φύλλα ευρετικές τιμές οι οποίες αξιολογούν τις αντίστοιχες καταστάσεις ως ευνοϊκές γι' αυτόν (μεγάλες τιμές) ή για τον αντίπαλο (μικρές τιμές), και αποδίδει τιμές στους ενδιάμεσους κόμβους από τα φύλλα προς τη ρίζα του δένδρου ώσπου να αποκτήσει τιμή η τελευταία, θεωρώντας ότι σε κάθε επίπεδο του δένδρου κάθε παίκτης (κατάσταση-γονέας) θα επιλέξει τη συμφερότερη γι' αυτόν κίνηση (μετάβαση) από όλες τις πιθανές (καταστάσεις-παιδιά). Έτσι η τιμή κάθε κόμβου Max (κόμβου που στο δένδρο προηγείται κίνησης του ίδιου του παίκτη) είναι η μέγιστη των τιμών των παιδιών του και η τιμή κάθε κόμβου Min (κόμβου που προηγείται κίνησης του αντιπάλου) είναι η ελάχιστη των τιμών των παιδιών του. Τελικά καλύτερη κίνηση είναι αυτή που οδηγεί στον κόμβο που έδωσε τη μεγαλύτερη τιμή για τη ρίζα.

Ο αλγόριθμος ουσιαστικά εγγυάται πάντα την πιο συμφέρουσα εξέλιξη του παιχνιδιού, υποθέτοντας ότι ο αντίπαλος επιλέγει διαρκώς τις καλύτερες για εκείνον κινήσεις και δεν κάνει λάθη. Με τον Minimax υπάρχουν πάντα δυνατές κινήσεις οι οποίες φαίνεται να συμφέρουν περισσότερο, αγνοούνται όμως όταν μπορεί να οδηγήσουν σε καταστάσεις όπου ο αντίπαλος -αν κάνει την καλύτερη επιλογή- αποκτά προβάδισμα. Ο αλγόριθμος ΑΒ (Άλφα Βήτα) είναι παραλλαγή του εξαντλητικού Minimax με κλάδεμα προκειμένου να εξοικονομηθεί χρόνος, όπου αγνο-

ούνται καταστάσεις-παιδιά οι οποίες αποκλείεται να δώσουν την τιμή τους στον γονέα τους. Στη χειρότερη περίπτωση αποδίδει σαν τον Minimax. Το κλάδεμα γίνεται με αντικειμενικά κριτήρια (όπως στον B&B), ενώ προκειμένου να βελτιωθεί η απόδοση μπορούν επιπλέον να επιστρατευθούν το ευρετικό κλάδεμα, μια δυναμική συνάρτηση αξιολόγησης η οποία εξαρτάται από την εξέλιξη της παρτίδας, η κρυφή μνήμη τιμών αξιολόγησης και άλλες μέθοδοι. Το σπουδαιότερο πρόβλημα όμως τόσο του Minimax όσο και του AB, πέρα από την απόδοση, είναι το «πρόβλημα του ορίζοντα», κατά το οποίο η άσχημη εξέλιξη της παρτίδας δεν μπορεί τελικώς να αποφευχθεί λόγω του περιορισμένου βάθους κατασκευής του δένδρου σε κάθε στάδιο. Ένας τρόπος να μειωθεί η επίδραση του προβλήματος του ορίζοντα είναι η χρήση ειδικών αλγορίθμων ανίχνευσης οι οποίοι εντοπίζουν χρονικά σημεία της παρτίδας κατά τα οποία η κατασκευή δένδρου με μεγαλύτερο βάθος από το συνηθισμένο ενδέχεται να βελτιώσει το τελικό αποτέλεσμα.

3.2.4 Ικανοποίηση περιορισμών

Μία ακόμα σημαντική κατηγορία προβλημάτων είναι τα προβλήματα ικανοποίησης περιορισμών. Και σε αυτά δεν είναι πλήρως γνωστές οι τελικές καταστάσεις, μόνο κάποιες ιδιότητες τους και οι περιορισμοί που πρέπει να ικανοποιούν. Επίσης δε μας ενδιαφέρει και το μονοπάτι που οδηγεί στη λύση: το μόνο που επιθυμούμε να βρούμε είναι η πλήρης μορφή μίας τελικής κατάστασης. Στα προβλήματα αυτά κωδικοποιούμε τους περιορισμούς με τη βοήθεια μεταβλητών και ενός πεδίου τιμών (διακριτών ή συνεχών, ανάλογα με το πρόβλημα) από το οποίο αυτές λαμβάνουν τιμή. Οι περιορισμοί αφορούν τις πιθανές τιμές που μπορεί να ανατεθούν σε κάθε μεταβλητή (ή σε κάποιες από τις μεταβλητές) και κάθε λύση που τους ικανοποιεί είναι αποδεκτή. Στην ικανοποίηση περιορισμών μπορούν να χρησιμοποιηθούν παραλλαγές του HC προσαρμοσμένες στην αναζήτηση σε χώρο μεταβλητών, όπου εκκινούμε από μία τυχαία ανάθεση τιμών στις μεταβλητές και σταδιακά τη διορθώνουμε (με τη βοήθεια ευρετικής αξιολόγησης) ώστε να παραβιάζει λιγότερους περιορισμούς, πάντα με τον εγγενή στην αναρρίχηση λόφων κίνδυνο να παγιδευτούμε σε τοπικό ελάχιστο του σφάλματος (τοπικά βέλτιστη λύση). Συνήθως σε κάθε επανάληψη του αλγορίθμου είτε γίνεται αναζήτηση, σε όλες τις μεταβλητές, μίας τιμής που ελαχιστοποιεί περαιτέρω τις διενέξεις με τους περιορισμούς, είτε επιλέγεται τυχαία μία μεταβλητή και αναζητώνται μόνο δικές της τιμές που μειώνουν τις διενέξεις.

Εναλλακτικά μπορούμε να κωδικοποιήσουμε περαιτέρω τις μεταβλητές σε καταστάσεις, όπου ο μόνος πιθανός τελεστής είναι η ανάθεση τιμής σε μία μεταβλητή, και να εφαρμόσουμε τυφλή αναζήτηση στο χώρο καταστάσεων που προκύπτει. Η τελευταία μέθοδος όμως δεν αποδίδει καλά σε μεγάλους χώρους λόγω της συνδυαστικής έκρηξης, οπότε μπορεί να συνδυαστεί με αλγορίθμους συνέπειας. Οι τελευταίοι ελέγχουν διαδοχικά όλους τους περιορισμούς του προβλήματος και για

τον καθέναν εξετάζουν τα πεδία τιμών των μεταβλητών ώστε να αφαιρέσουν τις τιμές που τον παραβιάζουν. Κάθε αφαίρεση απαιτεί επανέλεγχο των περιορισμών αφού μπορεί να επηρεάζονται και άλλες μεταβλητές μέσω αυτών. Οι αλγόριθμοι συνέπειας τόξου (AC) αναπαριστούν τις μεταβλητές ως κόμβους ενός γράφου και τους περιορισμούς ως τόξα που συνδέουν τους κατάλληλους κόμβους.

Ο απλούστερος αλγόριθμος της κατηγορίας είναι ο AC-3, ο οποίος μπορεί να χειριστεί μόνο δυαδικούς περιορισμούς (στους οποίους συμμετέχουν το πολύ δύο μεταβλητές). Αποδεικνύεται ότι κάθε περιορισμός ανώτερης τάξης μπορεί να αναχθεί σε ένα σύνολο δυαδικών περιορισμών. Όταν ο αλγόριθμος ολοκληρώνεται μπορεί να έχει αφήσει μόνο λίγες δυνατές τιμές στα πεδία, μειώνοντας κατά πολύ το χώρο αναζήτησης, ωστόσο σχεδόν πάντα παραμένουν ασυνεπείς τιμές κάνοντας αναγκαία τη χρήση και ενός κλασικού αλγορίθμου αναζήτησης. Στην τελευταία περίπτωση ο αλγόριθμος συνέπειας χρησιμοποιείται για κλάδεμα κατά τη διάρκεια της αναζήτησης. Το κλάδεμα αυτό συμβαίνει σε κάθε επανάληψη του βρόχου οπότε αντί για την, ακριβή υπολογιστικά, πλήρη εφαρμογή ενός αλγορίθμου συνέπειας τόξου («έγκαιρη πλήρης εξέταση») μπορεί να εφαρμοστεί μία απλουστευμένη «έγκαιρη μερική εξέταση», κατά την οποία κάθε περιορισμός ελέγχεται μόνο μία φορά ακόμα και αν αφαιρεθεί κάποια τιμή από κάποιο πεδίο. Ακόμα πιο απλουστευμένος είναι ο «προοπτικός έλεγχος» κατά τον οποίον εξετάζονται μόνο τα πεδία τιμών των μεταβλητών που συνδέονται, μέσω περιορισμών, με τη μεταβλητή στην οποία ανατέθηκε τιμή κατά την προηγούμενη επανάληψη. Κατά την εκτέλεση του αλγορίθμου, αν το πεδίο τιμών μίας μεταβλητής γίνει κενό τότε η συγκεκριμένη κατάσταση απορρίπτεται, ενώ αν όλες οι μεταβλητές έχουν λάβει τιμή και δεν παραβιάζεται κανένας περιορισμός τότε ο αλγόριθμος έχει καταλήξει σε τελική κατάσταση.

3.2.5 Εξελικτικός υπολογισμός

Σε προβλήματα βελτιστοποίησης μπορεί εναλλακτικά να χρησιμοποιηθεί κάποιος τύπους εξελικτικού υπολογισμού όπως οι εξελικτικοί αλγόριθμοι. Ο εξελικτικός υπολογισμός είναι μία κατηγορία εργαλείων της υπολογιστικής νοημοσύνης ο οποίος βασίζεται στην ιδέα της σταδιακής ανάπτυξης, με μια επαναληπτική διαδικασία, πιθανών λύσεων για ένα πρόβλημα μέσω πολλαπλών παράλληλων αναζητήσεων στο χώρο των λύσεων. Η διαφορά από άλλες μεθόδους βελτιστοποίησης είναι ότι οι υποψήφιες λύσεις αλληλεπιδρούν και αλληλεπηρεάζονται ώσπου να τερματίσει η διαδικασία. Όταν αυτή η αλληλεπίδραση συμβαίνει με βάση τις αρχές της βιολογικής εξέλιξης των ειδών τότε μιλάμε για εξελικτικούς αλγορίθμους. Στον τομέα της βέλτιστης επίλυσης προβλημάτων συνήθως χρησιμοποιούνται οι γενετικοί αλγόριθμοι, μία υποκατηγορία εξελικτικών αλγορίθμων η οποία μοιάζει με την αναρρίχηση λόφων αλλά δεν παγιδεύεται εύκολα σε τοπικά βέλτιστες λύσεις.

Στους γενετικούς αλγορίθμους αρχικά παράγεται ένα σύνολο Ν υποψήφιων λύσεων για το εκάστοτε πρόβλημα (πληθυσμός n), το οποίο κατασκευάζεται τυχαία και επομένως τα περισσότερα μέλη του είναι άκυρα ή μη βέλτιστα ως λύσεις. Αυτοί οι υποψήφιοι αξιολογούνται με μία καθοριζόμενη από τον χειριστή (συνήθως ευρευτική) πραγματική συνάρτηση διακριτής μεταβλητής, τη συνάρτηση καταλληλότητας, η οποία επιχειρεί να βαθμολογήσει κάθε υποψήφιο ανάλογα με το πόσο κοντά βρίσκεται σε κάποια ιδανική λύση. Ακολούθως από τον αρχικό πληθυσμό σχηματίζονται N/2 ζεύγη υποψηφίων («γονέων»), με μεγαλύτερη προτεραιότητα στις πιο θετικά αξιολογημένες λύσεις, όπου κάθε υποψήφιος μπορεί να συμμετέχει σε περισσότερα από ένα ζεύγη. Τα μέλη κάθε ζεύγους συνδυάζονται με κάποιον τρόπο μεταξύ τους και το αποτέλεσμα είναι δύο νέες υποψήφιες λύσεις («απόγονοι»). Ο νέος πληθυσμός n+1 αποτελείται από το σύνολο αυτών των απογόνων (πλήρης ανανέωση). Εναλλακτικά οι απόγονοι μπορούν να συνυπάρχουν με μέλη του αμέσως προηγούμενου πληθυσμού n (μερική ανανέωση), σε κάθε περίπτωση όμως ο πληθάριθμος Ν παραμένει σταθερός σε κάθε «γενιά».

Το ποσοστό υποψηφίων που αντικαθίσταται από απογόνους ονομάζεται «χάσμα γενεών» και στην πλήρη ανανέωση είναι 100%, ενώ στη μερική ανανέωση η πιθανότητα αντικατάστασης μίας λύσης της γενιάς n από απόγονο της γενιάς n+1 είναι αντιστρόφως ανάλογη της καταλληλότητας της. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να ικανοποιηθεί κάποιο κριτήριο τερματισμού, δηλαδή συνήθως να βρεθεί μία λύση που αξιολογείται ως βέλτιστη από τη συνάρτηση καταλληλότητας ή ο μέσος όρος των λύσεων του τρέχοντος πληθυσμού να τείνει να συγκλίνει σε μία μόνο λύση (ή μικρές παραλλαγές μίας). Αυτή η μεθοδολογία επιχειρεί να μιμηθεί τη βιολογική ιδέα της γενετικής διαφοροποίησης και της φυσικής επιλογής, των πυλώνων της εξέλιξης των ειδών, αλλά ουσιαστικώς η φυσική επιλογή αντικαθίσταται από μία τεχνητή επιλογή η οποία γίνεται μέσω της συνάρτησης καταλληλότητας. Η τελευταία αποτελεί και το υπέρτατο κριτήριο για την πραγματική απόδοση του αλγορίθμου.

3.3 Τεχνητά Νευρωνικά Δίκτυα

3.3.1 Εισαγωγή

Τα Τεχνητά Νευρωνικά Δίκτυα (Neural Networks) επιχειρούν να προσομοιώσουν με μαθηματικό τρόπο το νευρικό σύστημα του ανθρώπου και να μιμηθούν την κατανεμημένη λειτουργία του εγκεφάλου. Αποτελούνται από απλά δομικά στοιχεία, τα νευρόνια, τα οποία διαθέτουν προσαρμοζόμενες παραμέτρους και είναι ικανά να "μαθαίνουν" και να αποκρίνονται με "εξυπνάδα" σε νέα ερεθίσματα. Η ικανότητα μάθησης που διαθέτουν τα Νευρωνικά Δίκτυα ακολουθεί παρόμοιες διαδικασίες με αυτές που βιώνει ένας άνθρωπος στα πρώτα στάδια της ζωής του.

Όπως ο οργανισμός ενός μικρού παιδιού μαθαίνει πως π.χ. το μάτι της κουζίνας είναι επικίνδυνο όταν ειναι κόκκινο και εκπέμπει θερμότητα, ενώ είναι ασφαλές όταν είναι μάυρο και κρύο, με την ίδια συλλογιστική και τα Νευρωνικά Δίκτυα εκπαιδεύουν τον "οργανισμό" τους να διαχωρίζει καταστάσεις και αντικείμενα ανάλογα με τα χαρακτηριστικά τους και την παρελθούσα γνώση που υπάρχει για αυτά.

3.3.2 Διαφορετικά είδη Νευρωνικών δικτύων

Τα Τεχνητά Νευρωνικά Δίκτυα Απλού Στρώματος Μολονότι ένα μόνο του νευρώνιο μπορεί να εκτελέσει μερικές απλές λειτουργίες που αφορούν την ανίχνευση απλών προτύπων , η δύναμη της υπολογιστικότητας των νευρωνίων ανοίγετε μπροστά μας μόνο όταν αυτά συνδεθούν σε δίκτυο . Το πιο απλό δίκτυο αποτελείτε από μια ομάδα από νευρώνια διατεταγμένους σε στρώμα , όπως φαίνετε στην δεξιά πλευρά του παρακάτω Σχήματος 4. Σημειώνουμε ότι , οι κύκλοι που φαίνονται στα αριστερά εργάζονται μόνο για το μίραζμα των τιμών εισόδου , δεν εκτελούν υπολογισμούς , και έτσι δεν θεωρούνται ότι αποτελούν στρώμα αντίθετα τα νευρώνια που εκτελούν υπολογισμούς έχουν την μορφή τετραγώνων . Το σετ των τιμών εισόδου X έχει το κάθε του στοιχείο συνδεδεμένο σε κάθε ΤΝ διαμέσου διαφορετικών βαρών . Τα πρώτα ΤΝΔ δεν ήταν περισσότερο πολύπλοκα από ότι αυτό εδώ . Το κάθε νευρώνιο απλά παράγει ένα άθροισμα από τις τιμές εισόδου του δικτύου που έχουν πολλαπλασιαστεί με τα αντίστοιχα βάρη .Στην πραγματικότητα στα τεχνητά και βιολογικά δίκτυα πολλές από τις συνδέσεις τους μπορεί να μην υπάρχουν , όμως φαίνονται όλες οι για λόγους γενικότητας .

Πολύστρωμα Νευρωνικά Δίκτυα Μεγαλύτερα , περισσότερο πολύπλοκα δίκτυα , γενικά , προσφέρουν μεγαλύτερη ικανότητα υπολογισμών . Μολονότι τα δίκτυα έχουν κατασκευαστή με κάθε δυνατό τρόπο διάταξής τους , διατάσσοντας τα νευρώνια σε στρώμα μιμούνται την στρωματική δομή των διάφορων τμημάτων του εγκεφάλου .Τα πολύστρωμα δίκτυα , έχει αποδειχθεί , ότι έχουν ικανότητες πέρα από αυτές των μονόστρωμων δικτύων και στα πρόσφατα χρόνια αναπτύχθηκαν αλγόριθμοι για να τα εκπαιδεύσουν . Τα πολύστρωμα δίκτυα μπορούν να σχηματιστούν από ομάδες μονόστρωμων δικτύων , η έξοδος ενός στρώματος αποτελεί την είσοδο του απομένουν στρώματος .

3.3.3 Διαφορετικοί Αλγόριθμοι Εκπαίδευσης

Ανατρέχοντας στη βιβλιογραφία μπορεί κανείς να βρει πολλούς αλγορίθμους εκπαίδευσης. Στην παρούσα εργασία γίνεται αναφορά μόνο στους πιο γνωστούς από αυτούς. Αυτοί είναι:

- Support Vector Machines - SVM
- K-Nearest Neighbor – kNN

- Neural Network – Nnet
- Linear Least-squares Fit (LLSF) mapping

Ο αλγόριθμος SVM βασίζεται στα διανύσματα υποστήριξης (support vectors). Το πρόβλημα έγκειται στην εύρεση μιας επιφάνειας η οποία χωρίζει τα δεδομένα σε δύο κλάσεις με τον καλύτερο δυνατό τρόπο. Η επιφάνεια αυτή ονομάζεται επιφάνεια απόφασης (best decision surface). Ένα παράδειγμα εφαρμογής του αλγορίθμου SVM περιέχει συνήθως έναν γραμμικό διαχωρισμό των δεδομένων. Υπάρχει μια επιφάνεια που περιγράφει θετικά και αρνητικά παραδείγματα αντίστοιχα, καθώς και γραμμές αναπαριστούν τις επιφάνειες απόφασης. Διακεκομμένες γραμμές δείχνουν τα δύο άκρα στα οποία μπορεί κανείς να μεταφέρει τις επιφάνειες απόφασης χωρίς να αλλάξει η κατηγοριοποίηση. Τα στοιχεία που βρίσκονται πάνω στις διακεκομμένες γραμμές ονομάζονται υποστηρικτικά διανύσματα (support vectors). Η επιφάνεια απόφασης που αναπαρίσταται με την μεσαία γραμμή είναι η καλύτερη δυνατή δεδομένου ότι είναι το μέσο στοιχείο μεταξύ των προηγούμενων επιφανειών απόφασης. Αξίζει να αναφερθεί ότι η επιφάνεια απόφασης που επιλέγεται κάθε φορά είναι η καλύτερη από ένα μικρό σύνολο εκπαιδευτικών παραδειγμάτων.

Στόχος του αλγορίθμου KNN είναι να αποφασίσει εάν ένα έγγραφο ανήκει σε μια κατηγορία . Η διαδικασία που ακολουθείται περιλαμβάνει την αναζήτηση του k πιο κοντινού (όσον αφορά την ομοιότητα) γείτονα ανάμεσα στο αρχικό σετ εγγράφων. Συγκεκριμένα, ελέγχεται εάν τα k πιο κοντινά έγγραφα ανήκουν επίσης στην κατηγορία. Εάν η απάντηση είναι θετική για ένα αρκετά μεγάλο ποσοστό αυτών τότε όντως το έγγραφο ανήκει σ' αυτήν την κατηγορία.

Οι τεχνικές που βασίζονται στα νευρωνικά δίκτυα χρησιμοποιούνται κυρίως στον τομέα της Τεχνητής Νοημοσύνης. Η κατηγοριοποίηση με νευρωνικά δίκτυα αποτελείται από ένα δίκτυο ομάδων όπου οι μονάδες εισόδου αναπαριστούν όρους ενώ οι μονάδες εξόδου αναπαριστούν την κατηγορία ή τις κατηγορίες στις οποίες ανήκει ένα αντικείμενο. Τα βάρη που αποδίδονται στις ακμές του δικτύου, και συνδέουν τις μονάδες, αναπαριστούν σχέσεις εξάρτησης. Εάν για παράδειγμα, ζητείται να κατηγοριοποιηθεί ένα έγγραφο η διαδικασία που ακολουθείται περιλαμβάνει την φόρτωση των βαρών των όρων του σαν είσοδο, την ενεργοποίηση τους και τη διάδοσή τους μέσα στο δίκτυο. Η έξοδος του δικτύου αποτελεί τις τελικές κατηγορίες. Ένας τρόπος για την εκπαίδευση ενός νευρωνικού δικτύου είναι η «προς τα πίσω διάδοση των λαθών» (backpropagation) όπου τα βάρη των όρων φορτώνονται στην είσοδο και τα λάθη που προκύπτουν διαδίδονται προς τα πίσω προκειμένου να γίνουν αλλαγές στις παραμέτρους του δικτύου με στόχο την ελαχιστοποίησή τους.

Η προσέγγιση LLSF είναι μια μέθοδος χαρτογράφησης που αναπτύχθηκε το 1992. Στην προσέγγιση αυτή ένα μοντέλο οπισθοδρόμησης πολλών μεταβλητών μαθαίνει αυτόματα από ένα σύνολο κατάρτισης εγγράφων και των κατηγοριών

στις οποίες ανήκουν. Τα δεδομένα αναπαριστώνται από ζεύγη διανυσμάτων εισόδου / εξόδου με το διάνυσμα εισόδου να αποτελείται από τους όρους και τα βάρη τους για κάθε έγγραφο (όπως στον αλγόριθμο SVM) ενώ το διάνυσμα εξόδου αποτελείται από τις κατηγορίες στις οποίες κατηγοριοποιήθηκε το έγγραφο. Λύνοντας το πρόβλημα LLSF παράγεται ένας πίνακας με τις λέξεις και τις κατηγορίες στις οποίες αυτές αντιστοιχούν.

3.4 Τεχνικές ομαδοποίησης: Ο αλγόριθμος k-means

Το clustering (ομαδοποίηση) είναι μια τεχνική data mining που ορίζεται ως η διαδικασία της ομαδοποίησης ενός αρχικού συνόλου δεδομένων στον πολυδιάστατο χώρο σε clusters (ομάδες), έτσι ώστε τα αντικείμενα που ανήκουν σε μια ομάδα να έχουν μεγάλη ομοιότητα μεταξύ τους, αλλά να διαφέρουν πολύ από τα αντικείμενα που ανήκουν σε άλλες ομάδες.

Πρόκειται για μια μέθοδο που έχει μελετηθεί εντατικά και χρησιμοποιείται κατά κόρο στις περιπτώσεις που χρειάζεται να γίνει διαχείριση μεγάλου όγκου πληροφορίας, με τρόπο ώστε η απαιτούμενη γνώση να μπορεί να εξορυχθεί με σχετική ευκολία. Το ονομαζόμενο 'cluster analysis' (ανάλυση ομαδοποίησης) στοχεύει στον εντοπισμό αντικειμένων που έχουν αντιπροσωπευτική συμπεριφορά στη συλλογή. Η βασική ιδέα είναι ότι εάν ένας κανόνας ισχύει για ένα αντικείμενο μιας ομάδας, είναι πολύ πιθανό ότι ισχύει επίσης και για όλα τα υπόλοιπα αντικείμενα που του 'μοιάζουν', δηλαδή τα υπόλοιπα αντικείμενα εντός της ίδιας ομάδας. Είναι ουσιαστικά μια μορφή unsupervised classification με την έννοια ότι οι κατηγορίες στις οποίες θα πρέπει να διαμεριστεί η αρχική συλλογή δεν είναι γνωστές εκ των προτέρων, και η ανακάλυψή τους είναι ένας από τους στόχους της ομαδοποίησης.

Για την ομαδοποίηση μιας αρχικής συλλογής αντικειμένων σε ομάδες, πρέπει να έχει επιλεγεί ο τύπος των χαρακτηριστικών γνωρισμάτων των αντικειμένων στα οποία θα βασιστεί η ομαδοποίηση, καθώς και ο τρόπος αναπαράστασής τους. Για την περίπτωση που τα αντικείμενα που εξετάζονται είναι κείμενα (documents), ο πιο διαδεδομένος τρόπος αναπαράστασής τους είναι το μοντέλο Vector Space, σύμφωνα με το οποίο κάθε κείμενο αντιπροσωπεύεται από ένα διάνυσμα χαρακτηριστικών (συνήθως, συγκεκριμένοι όροι ή λέξεις που αναφέρονται συχνά). Το μήκος του διανύσματος ισούται με τον αριθμό των μοναδικών διακριτών ιδιοτήτων στη συλλογή (δηλ. τον αριθμό των χαρακτηριστικών λέξεων). Σε κάθε συστατικό αυτού του διανύσματος αποδίδεται ένα βάρος που δηλώνει τη σημασία του συγκεκριμένου χαρακτηριστικού για το συγκεκριμένο κείμενο. Οι επιτρεπόμενες τιμές του βάρους είναι 0 ή 1 (ανάλογα με το αν η συγκεκριμένη ιδιότητα είναι χαρακτηριστική για το document, ή αλλιώς ανάλογα με το εάν εμφανίζεται η λέξη στο κείμενο ή όχι), ή ενδέχεται το βάρος να παίρνει τιμές που προκύπτουν

από τη συχνότητα εμφάνισης του χαρακτηριστικού στο κείμενο και τη συχνότητα εμφάνισης του χαρακτηριστικού συνολικά στη συλλογή.

Αφού αποφασιστεί ο τρόπος αναπαράστασης και τα χαρακτηριστικά στα οποία θα βασιστεί η ομαδοποίηση, πρέπει να επιλεγεί η κατάλληλη μετρική για τον υπολογισμό της ομοιότητας δύο αντικειμένων. Στις ευρέως χρησιμοποιούμενες μετρικές ομοιότητας περιλαμβάνονται η Cosine Coefficient, η Jaccard Coefficient και η Dice Coefficient.

Η ομαδοποίηση έχει μελετηθεί και χρησιμοποιηθεί σε πολλά επιστημονικά πεδία συμπεριλαμβανομένου της Μηχανικής Μάθησης, των Τεχνητών Νευρωνικών Δικτύων, των Βάσεων Δεδομένων και της Στατιστικής. Με βάση τα χαρακτηριστικά των αντικειμένων στα οποία βασίζεται ο αλγόριθμος ομαδοποίησης, διακρίνονται τρεις βασικές κατηγορίες:

- text-based ομαδοποίηση που εξετάζει το περιεχόμενο των web αντικειμένων,
- link-based ομαδοποίησης που εξετάσει τη δομή των συνδέσμων μεταξύ των web αντικειμένων της αρχικής συλλογής και
- hybrid ομαδοποίηση που λαμβάνει υπόψη του, τόσο το περιεχόμενο, όσο και τη δομή των συνδέσμων.

3.4.1 Ο αλγόριθμος k-means

Ο k-means ανήκει στην κατηγορία των επιμεριστικών αλγόριθμων και είναι κατάλληλος για μη-ιεραρχική ομαδοποίηση μεγάλων συνόλων δεδομένων σε κατάλληλες ομάδες. Είναι ο απλούστερος και ο πιο διαδεδομένος αλγόριθμος ομαδοποίησης και χρησιμοποιεί ένα κριτήριο τετραγωνικού σφάλματος. Ο αλγόριθμος λαμβάνει σαν είσοδο ένα σύνολο από n αριθμητικά αντικείμενα και έναν ακέραιο k (k<n), και υπολογίζει μια διαμέριση του αρχικού συνόλου σε k ομάδες, με τρόπο ώστε η ομοιότητα μεταξύ των αντικειμένων στην ίδια ομάδα να είναι μεγάλη και η ομοιότητα μεταξύ των αντικειμένων διαφορετικών ομάδων να είναι μικρή (ή αλλιώς, να μεγιστοποιείται η ομοιότητα εντός ομάδας και να ελαχιστοποιείται η ομοιότητα μεταξύ των ομάδων). Η ομοιότητα των ομάδων υπολογίζεται με βάση τη μέση τιμή των αντικειμένων του και μπορεί συνεπώς να θεωρηθεί ότι συμπίπτει με το κέντρο βάρους της ομάδας.

Ο k-means εκτελείται επαναληπτικά σε δύο φάσεις. Αρχικά επιλέγονται αυθαίρετα k από τα αντικείμενα του αρχικού συνόλου και το καθένα από αυτά θεωρείται ότι αντιπροσωπεύει το μέσο ή κέντρο μιας ομάδας (mean ή center). Για καθένα από τα υπόλοιπα αντικείμενα του αρχικού συνόλου, υπολογίζεται η απόστασή του από τα κέντρα των ομάδων και αποδίδεται στην ομάδα από την οποία απέχει τη μικρότερη απόσταση.

Αφού αντιστοιχηθούν όλα τα αντικείμενα επανα-υπολογίζονται οι μέσες τιμές των ομάδων. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να επιτευχθεί σύγκλιση ελαχιστοποιώντας το κριτήριο τετραγωνικού σφάλματος που ορίζεται ως εξής:

$$E = \sum_{i=1}^{k} \sum_{i_l \in C} q(i_l, i_{mj}) \tag{3.1}$$

Στη σχέση αυτή, το ορίζει το άθροισμα των τετραγωνικών σφαλμάτων για όλα τα αντικείμενα στο αρχικό σύνολο, το i_l είναι το σημείο στο χώρο που αντιπροσωπεύει ένα δεδομένο αντικείμενο, και i_{mj} είναι το μέσο της ομάδας C_i (τόσο το i_l, όσο και το i_{mj} είναι πολυσδιάστατα). Το συγκεκριμένο κριτήριο έχει σα στόχο,οι τελικές ομάδες να είναι όσο το δυνατό πιο συμπαγείς και απομακρυσμένες, δηλ. με διακριτά όρια.

Στον απευθείας k-means το q ορίζεται ως η τετραγωνική Ευκλείδεια απόσταση και συνεπώς $q(x, y) = \|x - y\|^2$. Η συνολική διαδικασία του k-means παρουσιάζεται στον πίνακα που ακολουθεί.

Ο k-means, αν και σχετικά κλιμακωτός (επεκτάσιμος), είναι υπολογιστικά δαπανηρός για την επεξεργασία μεγάλων συνόλων δεδομένων. Πιο συγκεκριμένα, η εκτέλεση του 1ου βρόγχου στον πίνακα που ακολουθεί, χρειάζεται για κάθε επανάληψη χρόνο O(ndk), του 2ου O(nd) και η συνάρτηση ποιότητας O(nd). Συνεπώς, η συνολική πολυπλοκότητα του αλγορίθμου είναι O(ndkt), όπου n είναι το πλήθος των αντικειμένων στο αρχικό σύνολο, k είναι ο αριθμός των ομάδων, t είναι ο αριθμός των επαναληπτικών εκτελέσεων που θα χρειαστούν μέχρι να επιτευχθεί σύγκλιση και d είναι ο αριθμός των διαστάσεων (συνήθως ισχύει ότι k«n και t«n).

Για την εφαρμογή του k-means, χρειάζεται να είναι καθορισμένο το μέσο μιας ομάδας, κάτι που ενδέχεται να μη συμβαίνει, όπως για παράδειγμα όταν πρόκειται για δεδομένα με τιμές τύπου κατηγορήματος. Επιπλέον, μπορεί να θεωρηθεί μειονέκτημα και το γεγονός ότι ο χρήστης του αλγορίθμου θα πρέπει να καθορίσει εκ των προτέρων τον αριθμό των ομάδων, k. Γενικότερα, ο k-means δεν είναι κατάλληλος για τον εντοπισμό ομάδων με ανομοιόμορφα σχήματα και με πολύ διαφορετικά μεγέθη.

Επιπλέον, είναι ευαίσθητος στο θόρυβο και σε παρεκκλίνουσες τιμές (outliers), καθώς ακόμα και ένας μικρός αριθμός από τέτοια σημεία μπορεί να επηρεάσει αισθητά τη μέση τιμή.

Στη σχετική βιβλιογραφία έχουν προταθεί ένα σύνολο από παραλλαγές του απευθείας k-means που βελτιώνουν είτε την υπολογιστική πολυπλοκότητα, είτε την εφαρμοσιμότητα του αλγορίθμου, ώστε να μπορεί να χρησιμοποιηθεί σε κατηγορικά δεδομένα ή μεικτού τύπου. Μια άλλη κατηγορία παραλλαγών βασίζεται στη χρήση διαφορετικών μέτρων ομοιότητας. Οι προσπάθειες για βελτίωση της υπολογιστικής πολυπλοκότητας εστιάζονται γύρω από δύο βασικές προσεγγίσεις: ή χρησιμοποιούν πιο εκλεπτυσμένες μεθόδους αρχικοποίησης, ή μειώνουν

Algorithm 2 Ο αλγόριθμος K-means αναλυτικά

Require: k - το πλήθος των clusters που θα υπολογιστούν
Require: n - το πλήθος των αντικειμένων που θα διαμεριστούν σε clusters
1: Διάλεξε τυχαία k objects $i_{m1}, i_{m2}, ..., i_{mk}$, ως αρχικά κέντρα των clusters
2: **repeat**
3: **for all** object i_l with $1 \leq l \leq N$ **do**
4: Αντιστοίχισε το i_l στο C_j με το κοντινότερο μέσο i_{mj}, έτσι ώστε
5: $\|i_l - i_{mj}\|^2 \leq \|i_l - i_{mu}\|^2$ όπου $l \leq j$ και $u \leq k$
6: **end for**
7: **for all** cluster $C_j \in C$ with $1 \leq j \leq k$ **do**
8: ξαναυπολόγισε το μέσο των σημείων $i_l \in C_j, i_{mj} = \frac{1}{|C_j|} \sum_{i_l \in C_j} i_l$
9: όπου το C_j δηλώνει το cardinality του C_j
10: **end for**
11: Υπολόγισε τη συνάρτηση ποιότητας q
12: **until** κανένα αντικείμενο να μην αλλάζει cluster ή να μην αλλάζει η τιμή του q

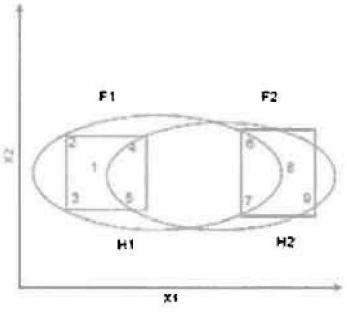
τον αριθμό των απαιτούμενων υπολογισμών ομοιότητας

3.5 Ασαφής Ομαδοποίηση

Μέχρι τώρα έχουμε δει ότι όλες οι τεχνικές και οι αλγόριθμοι ομαδοποίησης τοποθετούν ένα στοιχείο σε μια και μόνο ομάδα, σε αυτό που τελικά ανήκει. Και αυτό συνεπάγεται ότι οι ομάδες σε αυτές τις περιπτώσεις είναι ξένα μεταξύ τους σύνολα. Η ασαφής ομαδοποίησης επεκτείνει την έννοια του «ένα στοιχειό ανήκει σε μια ομάδα και συνδέει κάθε στοιχείο με όλα τις ομάδες χρησιμοποιώντας μια συνάρτηση μέλους. Το αποτέλεσμα είναι κάποια σύνολα από στοιχεία αλλά όχι μια απόλυτη διαμέριση του χώρου δεδομένων. Ένας αλγόριθμος ασαφούς ομαδοποίησης κάνει τα εξής σε γενικές γραμμές:

- Επιλογή μιας ασαφούς διάμεσης των N στοιχείων σε K ομάδες. Καθορισμός του πίνακα U=NxK του οποίου κάθε στοιχείο u_{ij} δηλώνει τον βαθμό συμμετοχής του στοιχείου i στην ομάδα j. Η τιμές των u είναι μεταξύ 0 και 1.
- Χρησιμοποιώντας τον πίνακα U βρίσκεται η τιμή κάποιας συνάρτησης που αποτελεί και το κριτήριο τερματισμού, και η οποία πρέπει να βελτιστοποιηθεί. Συνεχώς επανατοποθετούμε στοιχεία στις ομάδες με νέες τιμές συμμετοχής και επαναπροσδιορίζουμε τον πίνακα U και την τιμή της συνάρτησης.
- Επαναλαμβάνουμε το βήμα 2 μέχρι να μην επέρχονται σημαντικές αλλαγές στον πίνακα U και την τιμή της συνάρτησης.

Στο παρακάτω σχήμα φαίνεται η ιδέα της ασαφούς ομαδοποίησης. Τα παραλληλόγραμμα H1 και H2 εκφράζουν τις ομάδες μετά από κάποιον σκληρό αλγόριθμο, ενώ οι ελλείψεις δείχνουν δύο ασαφείς ομάδες, το F1 και F2. Τα στοιχεία των ελλείψεων συνοδεύονται από τιμές συμμετοχής σε κάθε cluster. Έτσι η ομάδα F1 περιγράφεται από το σύνολο τιμών (1,0.9), (2,0.8), (3,0.7), (4,0.6), (5,0.55), (6,0.4), (7,0.35), (8,0.0), (9,0.0). Αντίστοιχα με ένα τέτοιο σύνολο περιγράφεται και η ομάδα F2. Κάθε ζευγάρι τιμών (i, μ) σε αυτό το σύνολο εκφράζει το ποσοστό συμμετοχής μ του στοιχείου i στην ομάδα αυτή.

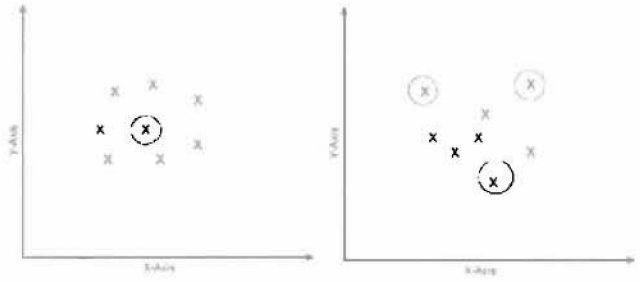


Σχήμα 3.1: *Ασαφής ομαδοποίηση*

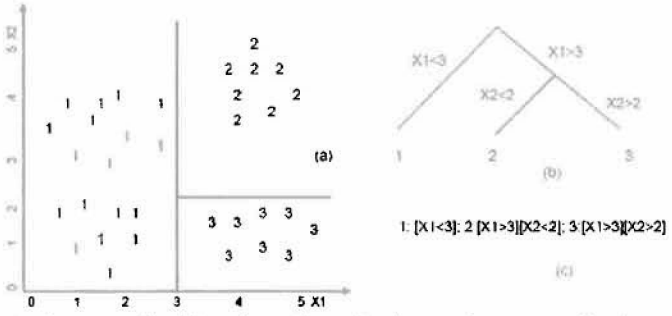
3.6 Αναπαράσταση των ομάδων

Το αποτέλεσμα της ομαδοποίησης είναι μια διαμέριση των δεδομένων σε ομάδες. Η διαμέριση αυτή δίνει μια ιδέα για το πως μπορούμε να ομαδοποιήσουμε τα δεδομένα μας σε έναν συγκεκριμένο αριθμό από κλάσεις. Σε πολλές εφαρμογές και κυρίως σε εκείνες λήψεως αποφάσεων είναι ανάγκη να υπάρξει μια συμπαγής και κατανοητή αναπαράσταση των ομάδων. Το ζήτημα της αναπαράστασης των ομάδων που είναι άμεσα συνδεδεμένο με την αφαίρεση δεδομένων είναι πολύ σημαντικό για την λήψη αποφάσεων. Η αναπαράσταση μιας ομάδας μπορεί να γίνει με διάφορους τρόπους. Μερικοί από τους προτεινόμενους είναι και οι παρακάτω:

- Μια ομάδα μπορεί να αναπαρασταθεί από το βαρύκεντρο σημείο του (ή κέντρο βάρους) ή από έναν αριθμό σημείων που είναι τα πιο απομακρυσμένα στην ομάδα.
- Οι κόμβοι ενός δένδρου κατηγοριοποίησης μπορούν να αναπαραστήσουν μια ομάδα.
- Επίσης μια ομάδα μπορεί να αναπαρασταθεί χρησιμοποιώντας συζευκτικούς λογικούς τελεστές.



Αναπαράσταση clusters από σημεία



(a)clusters σε χώρο 2 διαστάσεων, (b) ομαδοποίηση με δένδρο, (c) ομαδοποίηση με λογικούς συζευκτικούς τελεστές.

Σχήμα 3.2: *Αναπάρασταση ομάδων*

Ο πρώτος τρόπος αναπαράστασης με την χρήση ενός κέντρου βάρους είναι ο πιο συνήθης και έχει καλα αποτελέσματα όταν οι ομάδες είναι συμπαγείς και τα στοιχεία κατανέμονται ομοιόμορφα γύρω από το κέντρο βάρους. Σε αντίθετη περίπτωση ο τρόπος αυτός αναπαράστασης δεν είναι ο πιο κατάλληλος. Στην περίπτωση αυτή η αναπαράσταση μιας ομάδας από συνοριακά σημεία είναι μια πολύ καλή λύση υπάρχουν αρκετοί αλγόριθμοι που ακολουθούν αυτή την τεχνική όπως ο αλγόριθμος CURE.

Όσο το σχήμα της ομάδας αλλάζει και παίρνει διάφορα σχήματα στο χώρο, η επιλογή των σημείων πρέπει να είναι τέτοια ώστε να περιγράφεται η ομάδα όσο το δυνατόν καλύτερα. Η αναπαράσταση με ένα δένδρο κατηγοριοποίησης είναι ισοδύναμη με την αναπαράσταση μιας ομάδας με λογικούς συζευκτικούς τελεστές.

Η αναπαράσταση των ομάδων και η αφαίρεση δεδομένων που αυτή συνεπάγεται είναι πολύ σημαντική γιατί

- δίνει μια απλή και ανθρωπίνως κατανοητή αναπαράσταση των δεδομένων,
- Επιτυγχάνεται συμπίεση των δεδομένων η οποία μπορεί να αξιοποιηθεί από άλλες υπολογιστικές εφαρμογές και

- βοηθάει και επιταχύνει την διαδικασία λήψεως αποφάσεων.

3.7 Ομαδοποιώντας μεγάλα σύνολα δεδομένων

Σήμερα υπάρχουν εφαρμογές που απαιτούν την ομαδοποίηση πολύ μεγάλων συνόλων δεδομένων. Η έννοια του μεγάλου αλλάζει από εποχή σε εποχή και οι αλλαγές αυτές ακολουθούν τις τεχνολογικές εξελίξεις. Πριν χρόνια πολλά δεδομένα θεωρούνταν μερικές χιλιάδες από στοιχεία, ενώ σήμερα τα δεδομένα προς ομαδοποίηση όχι μόνο μετρώνται σε εκατομμύρια αλλά και διασπασιμότητα τους είναι πολύ μεγάλη. Κλασσικό παράδειγμα μεγάλης ποσότητας δεδομένων και διαστάσεων αποτελούν και τα πειράματα μικροσυστοιχιών DNA.

Τα τελευταία χρόνια νέες τεχνικές και αλγόριθμοι έχουν αναπτυχθεί, προταθεί για την ομαδοποίησης μεγάλων συνόλων δεδομένων, στα πλαίσια και νέων εφαρμογών όπως είναι η εξόρυξη δεδομένων. Μερικοί από τους πρώτους αλγορίθμους για ομαδοποίηση μεγάλων δεδομένων είναι ο CLARANS και ο BIRCH. Οι αλγόριθμοι αυτοί ομαδοποιούν μεγάλα σύνολα δεδομένων εφόσον όλα τα δεδομένα μπορούν να χωρέσουν στην μνήμη. Όμως αυτό είναι και το πρόβλημα των περισσοτέρων εφαρμογών σήμερα.

Τα δεδομένα είναι τόσα πολλά που η τοποθέτηση τους είναι αδύνατη στην κύρια μνήμη για συνολική επεξεργασία. Το πρόβλημα αυτό αντιμετωπίζεται με διάφορες τεχνικές όπως:

- Το σύνολο των στοιχείων αποθηκεύεται στη δευτερεύουσα μνήμη και υποσύνολα των δεδομένων ομαδοποιούνται ανεξάρτητα μεταξύ τους. Στη συνέχεια ακολουθείται μια διαδικασία συγχώνευσης ώστε να ομαδοποιηθούν συνολικά τα δεδομένα. Αυτή η τεχνική ακολουθεί την μέθοδο «διαίρει και βασίλευε».

- Όλα τα δεδομένα τοποθετούνται στην δευτερεύουσα μνήμη. Τα δεδομένα μεταφέρονται στην κύρια μνήμη ένα-ένα για ομαδοποίηση. Στην κύρια μνήμη αποθηκεύονται μόνο οι αναπαραστάσεις των ομάδων.

- Μια παράλληλη υλοποίηση μπορεί επίσης να χρησιμοποιηθεί.

Κεφάλαιο 4

Πρακτική εφαρμογή

Στα πλαίσια της πρακτικής εφαρμογής της εργασίας, θα χρησιμοποιηθεί το λογισμικό MEV, για την παρουσίαση παραδειγμάτων υλοποίησης αλγορίθμων υπολογιστικής νοημοσύνης σε μικροσυστοιχίες DNA. Έτσι, γίνεται καταρχήν μια αναφορά στα βασικά χαρακτηριστικά του προγράμματος MEV, ενώ στη συνέχεια θα παρουσιαστούν κάποιες ενδεικτικές υλοποιήσεις.

4.1 Το πρόγραμμα MEV

Το MEV είναι ένα αρκετά εύχρηστο πρόγραμμα, το οποίο είναι δυνατόν να δουλέψει με ένα πλήθος τύπων αρχείων. Τα κανονικοποιημένα και φιλτραρισμένα αρχεία έκφρασης μπορούν να αναλυθούν χρησιμοποιώντας την εμφάνιση TIGR Multiexperiment (MEV). Το MEV είναι ένα ευπροσάρμοστο εργαλείο ανάλυσης στοιχείων, ενσωματώνοντας περίπλοκους αλγορίθμους για την απεικόνιση, την ταξινόμηση, τη στατιστική ανάλυση και τη βιολογική ανακάλυψη θέματος. Το συγκεκριμένο λογισμικό μπορεί να χειριστεί διάφορες μορφές αρχείου εισόδου. Αυτοί περιλαμβάνουν .mev και .tav αρχεία που παράγονται τα αρχεία Affymetrix (.txt) και Genepix από TIGR, και επίσης (.gpr).

Το MEV παράγει τις πληροφοριακές και αλληλένδετες απεικονίσεις των διάφορων δεδομένων από ένα ή και πολλαπλά πειράματα. Σε αυτήν την τελική φάση της TM4 διαδικασίας, η ευελιξία και η ποικιλία των τεχνικών ανάλυσης είναι κρίσιμες παράμετροι, δεδομένου ότι κάθε αλγόριθμος έχει δυνατότητες που μπορούμε να εκμεταλλευτούμε, όταν χρησιμοποιείται σε ορισμένα σύνολα δεδομένων και πειραματικά σχέδια. Οι δυνατότητες ρύθμισης του εργαλείου λειτουργούν ιδιαίτερα καλά σε αυτό το σύστημα, δεδομένου ότι οι νέοι αλγόριθμοι και οι αντίστοιχες βάσεις δεδομένων από μικροσυστοιχίες μπορούν να ενσωματωθούν με τη χρήση του java – based MeV χρησιμοποιώντας μια καθορισμένη με σαφήνεια ενότητα API.

Πολλοί από τους αλγορίθμους που εφαρμόζονται στα πλαίσια του λογισμικού, έχουν ήδη αναφερθεί. Για παράδειγμα, από το MEV χρησιμοποιείται ο αλγόριθμος K-means, ο οποίος λαμβάνει ως δείγμα εκ νέου το σύνολο δεδομένων για να παραγάγει τους τομείς συναίνεσης. Οι διάφορες γραφικές παραστάσεις απαιτούν και αντίστοιχες παραμέτρους εισόδου για τους αλγορίθμους όπως ο k – means. Τα εργαλεία που χρησιμοποιούνται για να επιτρέψουν στις μεταβολικές διαβάσεις και τους χρωμοσωμικούς χάρτες να εμφανιστούν με αντίστοιχα δεδομένα είναι αυτή τη στιγμή υπό ανάπτυξη και δοκιμές. Ένα εργαλείο για να χειριστεί τις συνδέσεις με τους ιστότοπους βάσεων δεδομένων αναπτύσσεται επίσης. Οι ομάδες (clusters) που προσδιορίζονται μέσω οποιασδήποτε μεθόδου ανάλυσης μπορούν να ονομαστούν και να ακολουθηθούν μέσω άλλων αναλύσεων, που παρέχουν στο χρήστη τη δυνατότητα να συγκριθούν τα αποτελέσματα αρκετά που συγκεντρώνονται τους αλγορίθμους για να καθορίσουν τη συναίνεση και να εστιάσουν στα γονίδια με τα προσδιορισμένα πρότυπα έκφρασης και τα βιολογικά σχεδιαγράμματα.

Οι παρακάτω εικόνες, δείχνουν το περιβάλλον του MEV, που έχει χρησιμοποιηθεί για την επεξεργασία ενός πλήθους δεδομένων σχετικά με μικροσυστοιχίες DNA.

Στο γράφημα 1, φαίνεται ένα πρώτο παράδειγμα αναφορικά με το περιβάλλον εργασίας του λογισμικού MEV. Αντίστοιχα, ένα δεύτερο τέτοιο παράδειγμα φαίνεται και στο γράφημα 2.

4.2 Παρουσίαση πειραμάτων στο λογισμικό MEV

4.2.1 Δεδομένα που χρησιμοποιήθηκαν

Για τις υλοποιήσεις στο MEV, χρησιμοποιήθηκαν δεδομένα διαθέσιμα μέσω του προγράμματος, μέσω του αρχείου sample data. Στις παρακάτω εικόνες 4.3 και 4.4, φαίνεται ένα δείγμα των δεδομένων, όπως αυτά φαίνονται στο interface του MEV.

Το γράφημα 3 απεικονίζει το παράθυρο διαλόγουμ που εμφανίζεται όταν ο χρήστης θέλει να εισάγει δεδομένα προς μελέτη στο πρόγραμμα. Φαίνεται από την εικόνα ότι δίνεται στον χρήστη η δυνατότητα επιλογής συγκεκριμένου αρχείου προς εξέταση, ενώ με την επιλογή αυτόματα εμφανίζονται τα δεδομένα στον πίνακα στο κάτω μέρος του παραθύρου.

Το γράφημα 4 αποτελεί ένα παράδειγμα πίνακα που περιέχει δεδομένα προς εξέταση από το πρόγραμμα MEV.

4.2.2 Αποτελέσματα - Μείωση διάστασης με χρήση PCA αλγόριθμου κατά 50 τοις εκατό.

Τα δεδομένα στο πρόγραμμα φορτώνονται με την επιλογή "File - Load Data", με την οποία ανοίγει ένα παράθυρο διαλόγου. Στο συγκεκριμένο παράθυρο, δίνεται από το πρόγραμμα η δυνατότητα στον χρήστη να επιλέξει το αρχείο των δεδομένων. Για τη συγκεκριμένη εργασία χρησιμοποιήθηκε το αρχείο CGH_Sample Data. Αυτόματα φορτώνονται και τα Annotation Data. Μετά τη φόρτωση των δεδομένων, αυτά υπέστησαν επεξεργασία με τη χρήση του αλγορίθμου PCA, με χρήση τριών διαστάσεων. Αυτό, στο πρόγραμμα MEV, γίνεται μέσω της επιλογής "Data Reduction - Principal Component Analysis". Η αντίστοιχη απεικόνιση, φαίνεται στα δύο γραφήματα 4.5 και 4.6 που ακολουθούν.

Το γράφημα 5, δείχνει το πώς κατανέμονται τα διάφορα δεδομένα στις δύο διαστάσεις, μετά τη χρήση του αλγορίθμου PCA. Ο αλγόριθμος PCA (Principal Component Analysis) είναι αυτός που χρησιμοποιείται για τη μείωση της διάστασης των δεδομένων κατά 50

Δεδομένου του ότι η χρήση του αλγορίθμου PCA γίνεται σε 3 διαστάσεις, χρησιμοποιείται και το γράφημα 6 για να απεικονίσει την κατανομή των δεδομένων μετά τη χρήση του αλγορίθμου και για τη 3η διάσταση.

Στο σχήμα 4.7, φαίνονται αναλυτικά τα αποτελέσματα της διαδικασίας υλοποίησης του αλγορίθμου PCA.

4.3 K–means ομαδοποίηση

Εν συνεχεία, τα δεδομένα κατηγοριοποιήθηκαν σε ομάδες με τη χρήση του αλγορίθμου k–means. Η διαδικασία αυτή, στο πρόγραμμα MEV, έγινε μέσω των εντολών "Clustering - K means clustering". Με την επιλογή αυτή, ανοίγει ένα παράθυρο διαλόγου μέσω του οποίου ο χρήστης συμπληρώνει τα δεδομένα που επιθυμεί. Χρησιμοποιήθηκαν 10 ομάδες. Παρακάτω, στα 4.8 και 4.9, φαίνεται η κατηγοριοποίησή τους όπως αυτή παριστάνεται από τα κέντρα βάρους και τα γραφήματα έκφρασης.

Το γράφημα 8 απεικονίζει την κατηγοριοποίηση των δεδομένων μέσω των εικόνων κέντρου βάρους. Ουσιαστικά, πρόκειται για ένα σύνολο δέκα γραφημάτων, κάθε ένα από τα οποία αντιστοιχεί και σε διαφορετική ομάδα.

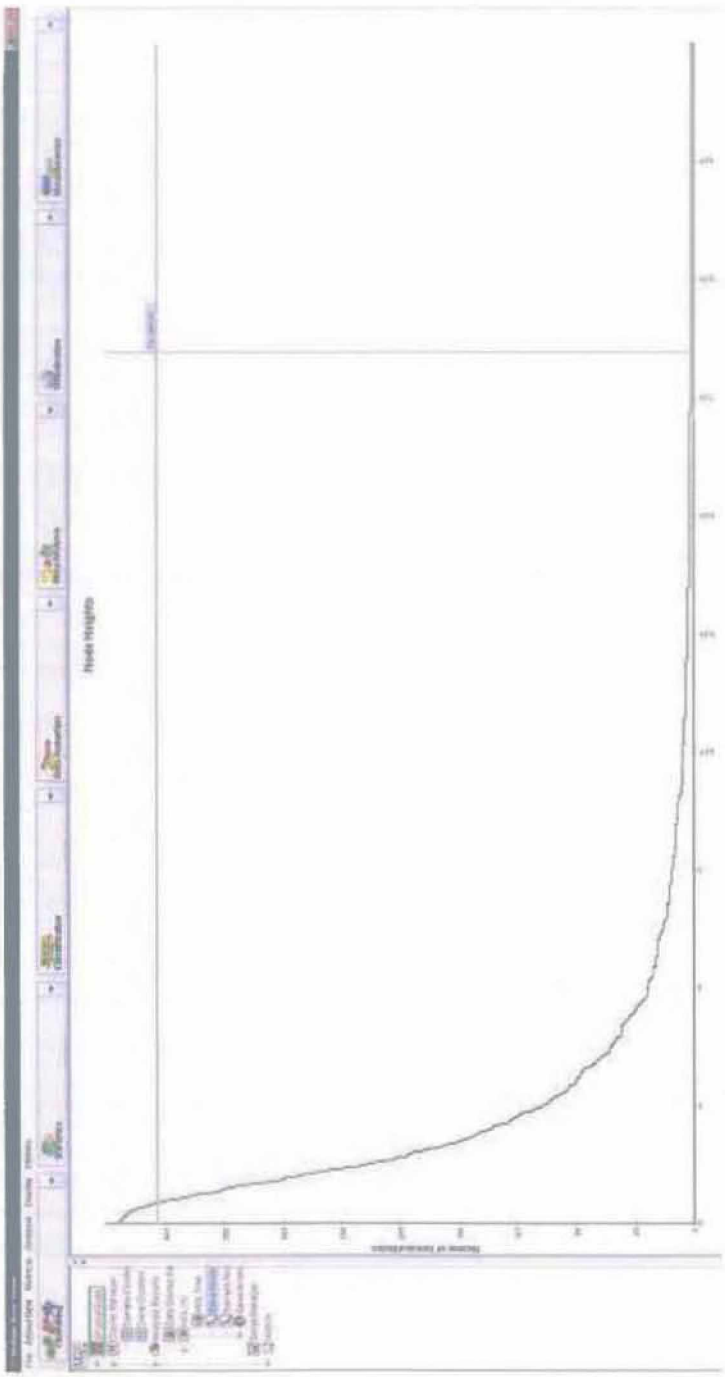
Αντίστοιχα, το γράφημα 9 απεικονίζει την κατηγοριοποίηση των δεδομένων μέσω των εικόνων έκφρασης. Και πάλι, φαίνεται ένα σύνολο δέκα γραφημάτων, ένα για κάθε διαφορετική ομάδα. Τέλος, στο παρακάτω 4.10 φαίνεται το πώς έχει γίνει ο διαχωρισμός στις 10 ομάδες, με αριθμούς.

4.4 Ταξινόμηση SVM

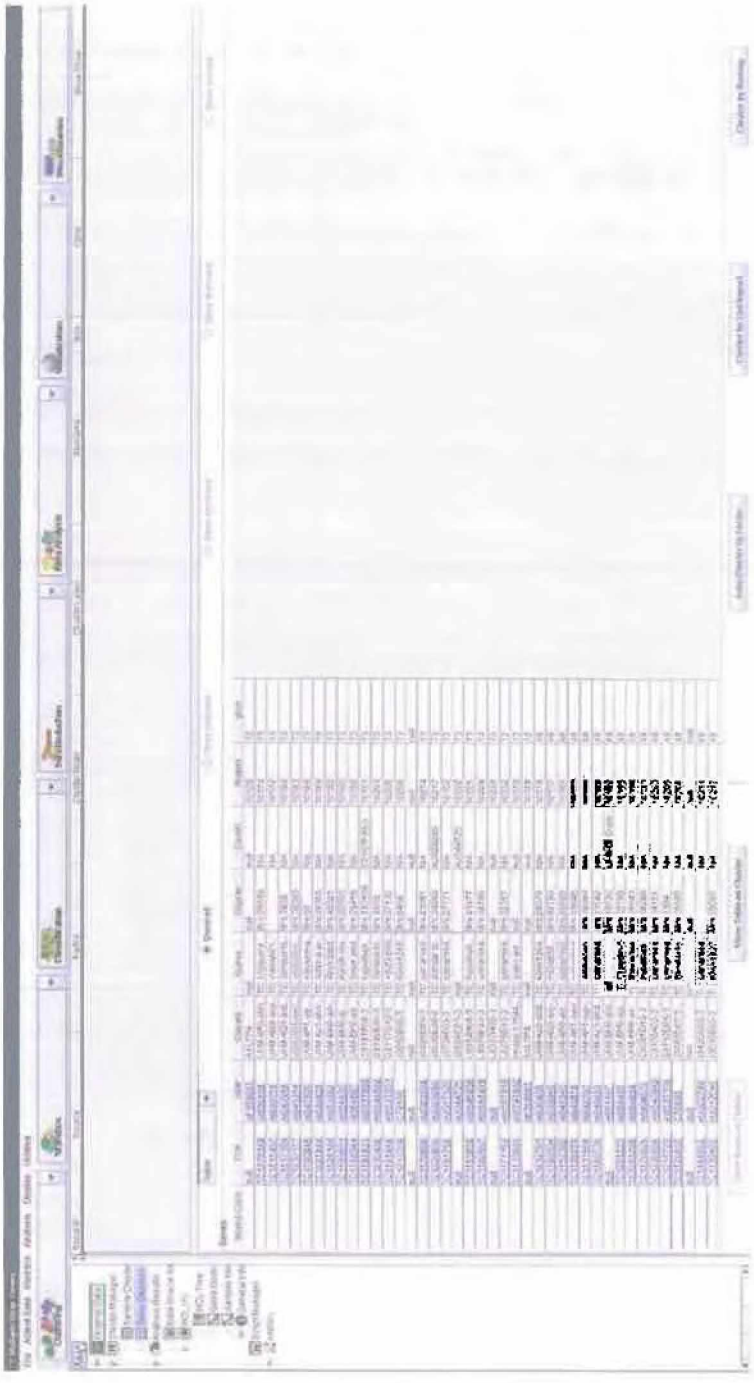
Τέλος, στα πλαίσια της επεξεργασίας των δεδομένων στο πρόγραμμα MEV, έγινε η διαδικασία της ταξινόμησης με τη χρήση της τεχνικής Support Vector Machine. Παρακάτω στα 4.11 και 4.12, φαίνονται τα αντίστοιχα γραφήματα κέντρου βάρους και έκφρασης.

Στο γράφημα 11, φαίνεται η ομαδοποίηση που πραγματοποιείται με βάση την έκφραση, για τα δεκα εννέα διαφορετικά πειράματα. Αντίστοιχα βλέπουμε στο γράφημα 12, την ομαδοποίηση βάσει του κέντρου βάρους.

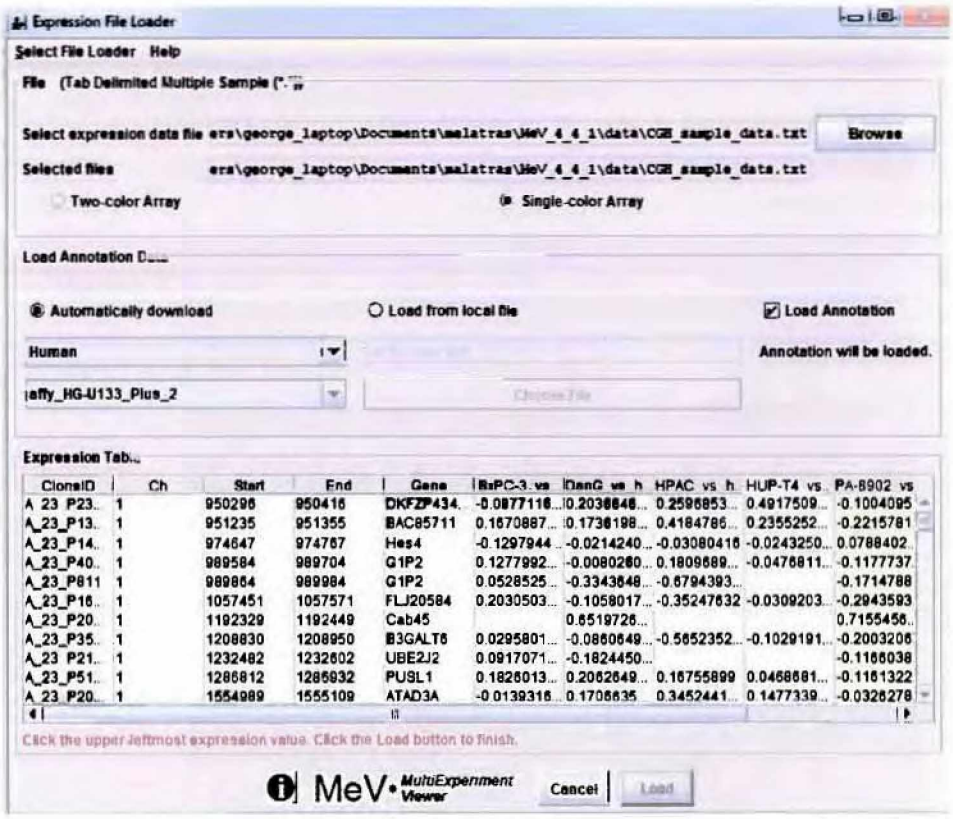
Τέλος, η διαδικασία με τη βοήθεια αριθμών συνοψίζεται στο παρακάτω σχήμα 4.13. Απότι φαίνεται, ο αλγόριθμος στην διαδικασία ταξινόμησης έβαλε όλα τα δείγματα στην 1η κατηγορία, κάτι που ενδέχεται να οφείλεται στο είδος των δεδομένων που χρησιμοποιήθηκαν.



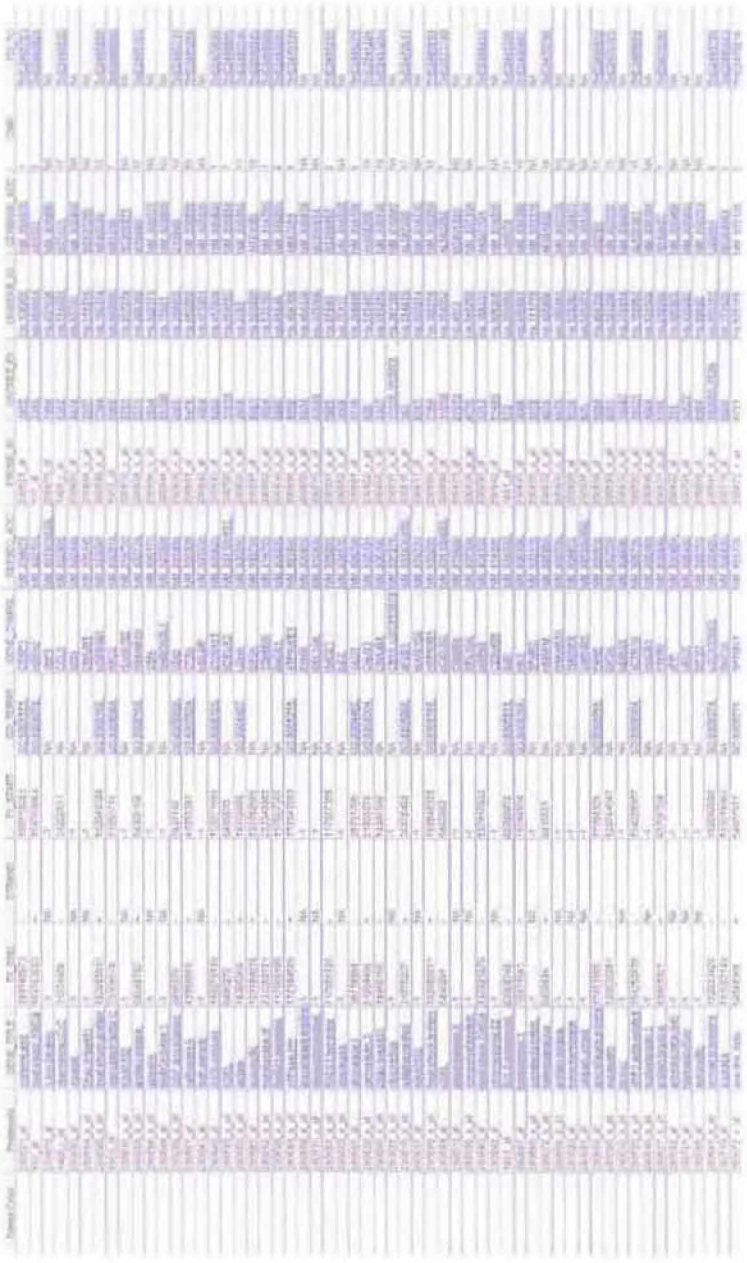
Σχήμα 4.1: *Το πρόγραμμα MEV - Παράδειγμα 1*



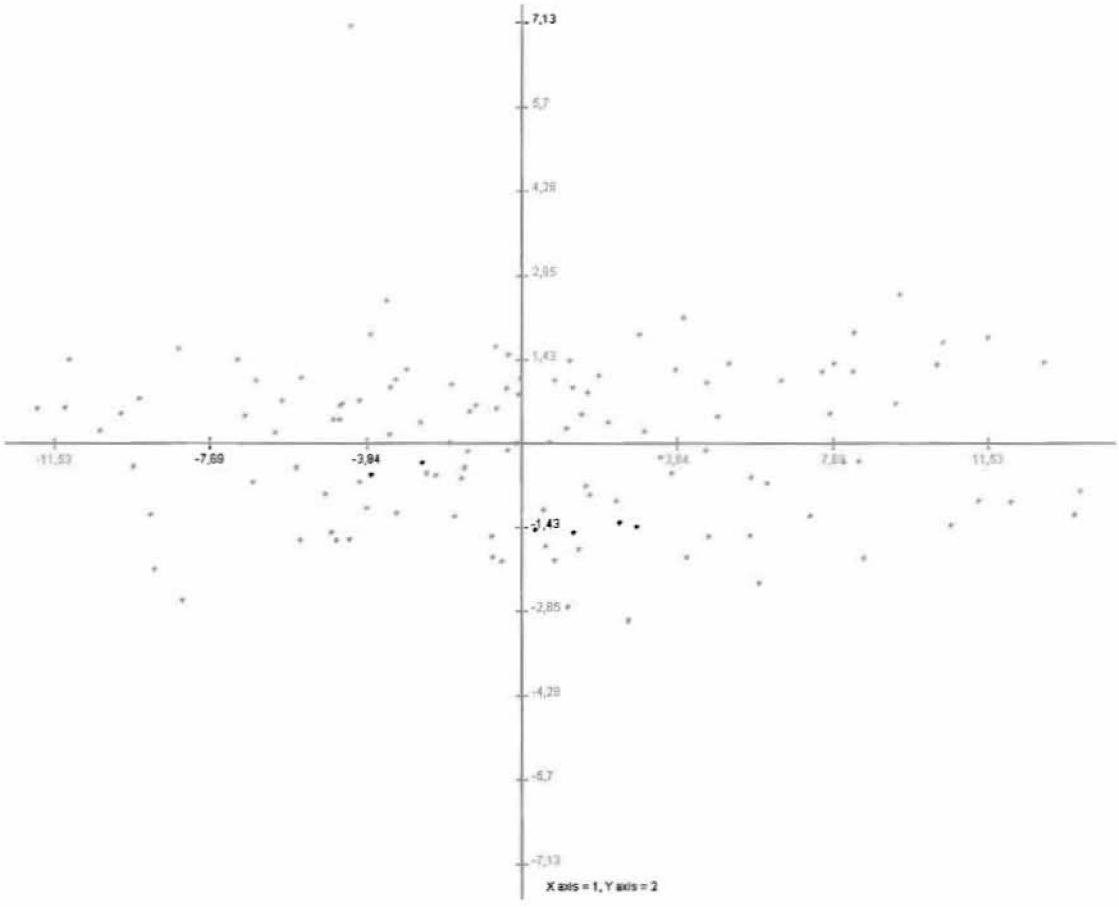
Σχήμα 4.2: *Το πρόγραμμα MEV - Παράδειγμα 2*



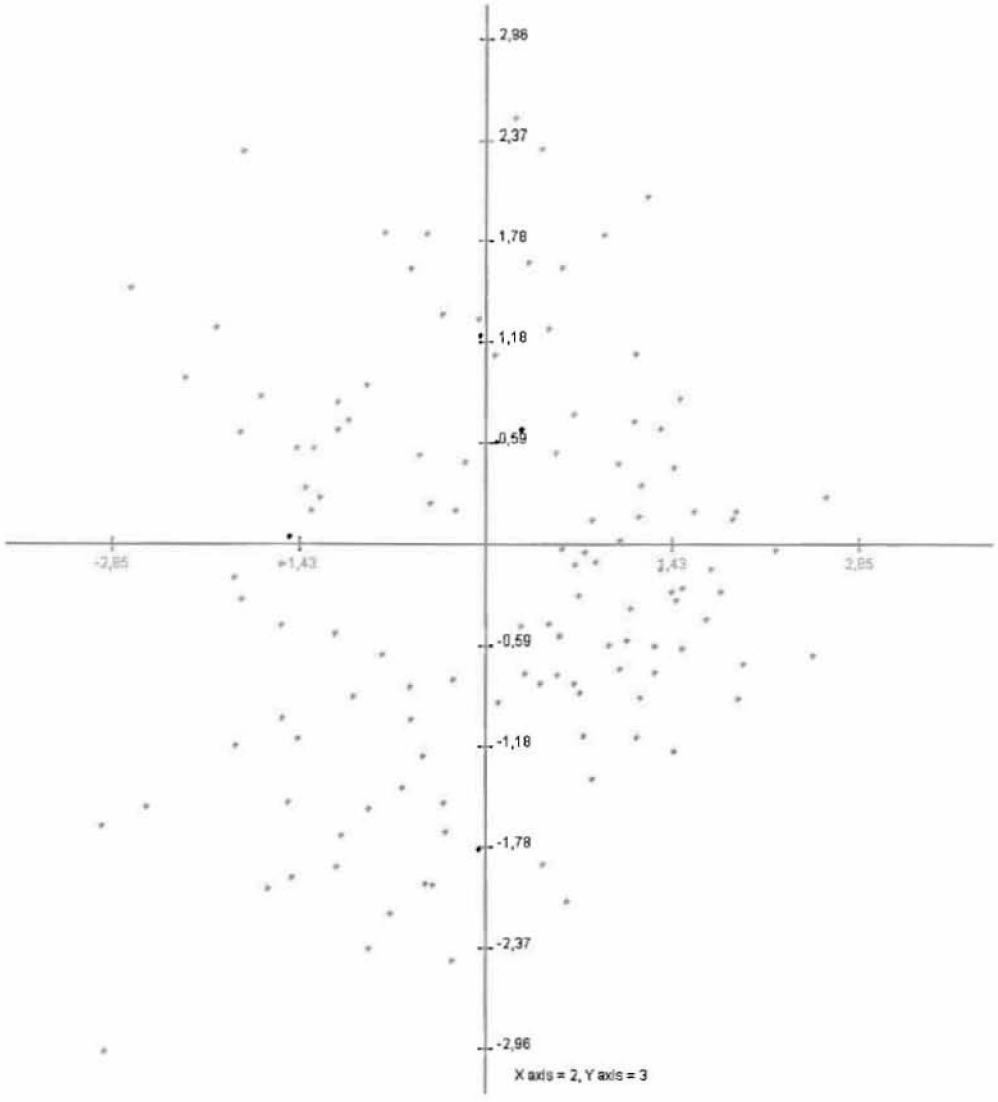
Σχήμα 4.3: *Εισαγωγή δεδομένων*



Σχήμα 4.4: *Τα δεδομένα*



Σχήμα 4.5: *Απεικόνιση αποτελέσματος PCA αλγορίθμου*

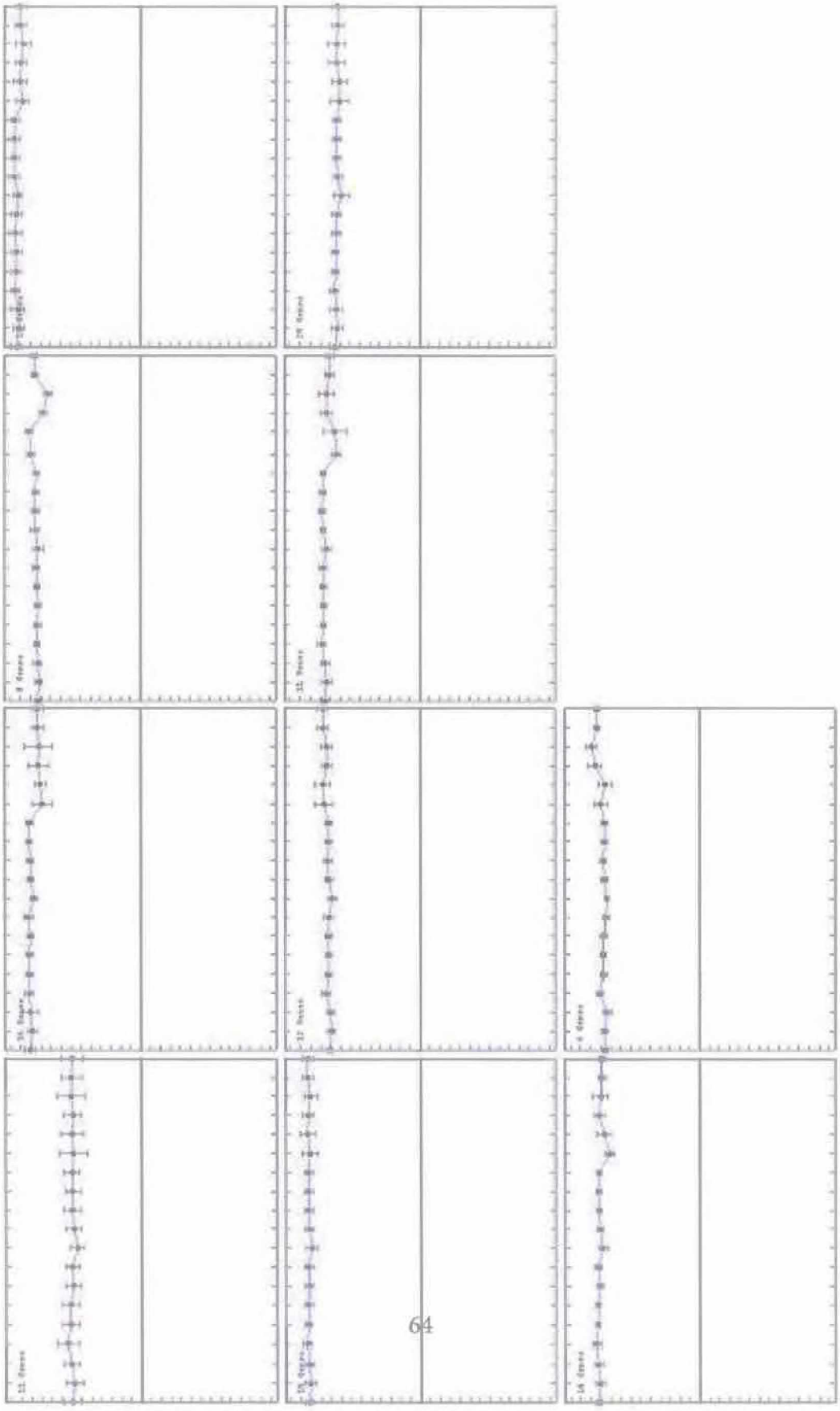


Σχήμα 4.6: *Απεικόνιση αποτελέσματος PCA αλγορίθμου*

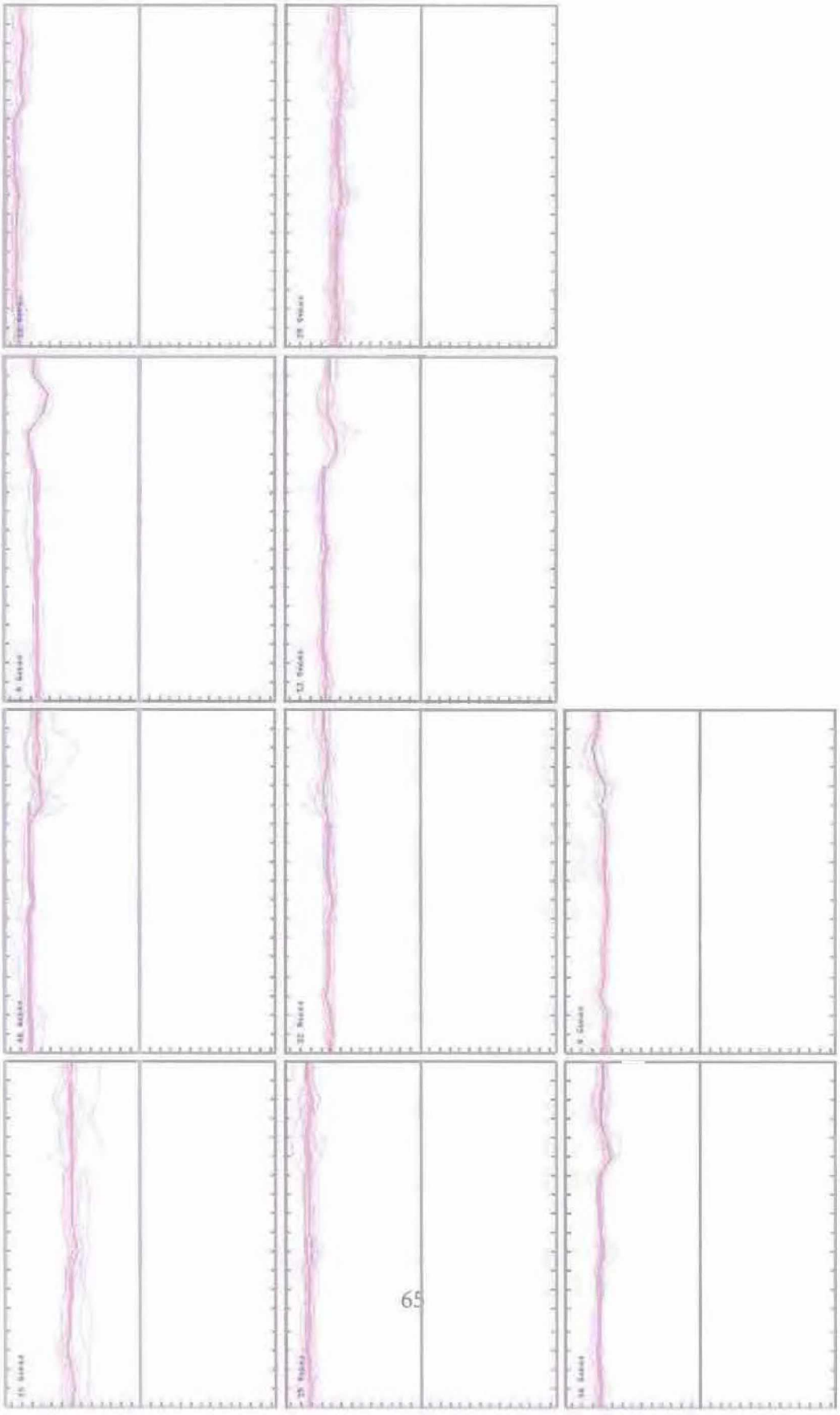
Principal Component 1	41,292	89,465 %
Principal Component 2	02,144	04,645 %
Principal Component 3	01,347	02,919 %
Principal Component 4	00,357	00,774 %
Principal Component 5	00,277	00,601 %
Principal Component 6	00,239	00,517 %
Principal Component 7	00,200	00,433 %
Principal Component 8	00,080	00,174 %
Principal Component 9	00,077	00,166 %
Principal Component 10	00,045	00,097 %
Principal Component 11	00,034	00,074 %
Principal Component 12	00,019	00,041 %
Principal Component 13	00,014	00,031 %
Principal Component 14	00,011	00,023 %
Principal Component 15	00,007	00,014 %
Principal Component 16	00,004	00,009 %
Principal Component 17	00,004	00,008 %
Principal Component 18	00,002	00,004 %
Principal Component 19	00,002	00,003 %

First 2 components: 94,110 %
First 3 components: 97,029 %

Σχήμα 4.7: *Διαδικασία υλοποίησης του αλγορίθμου PCA*



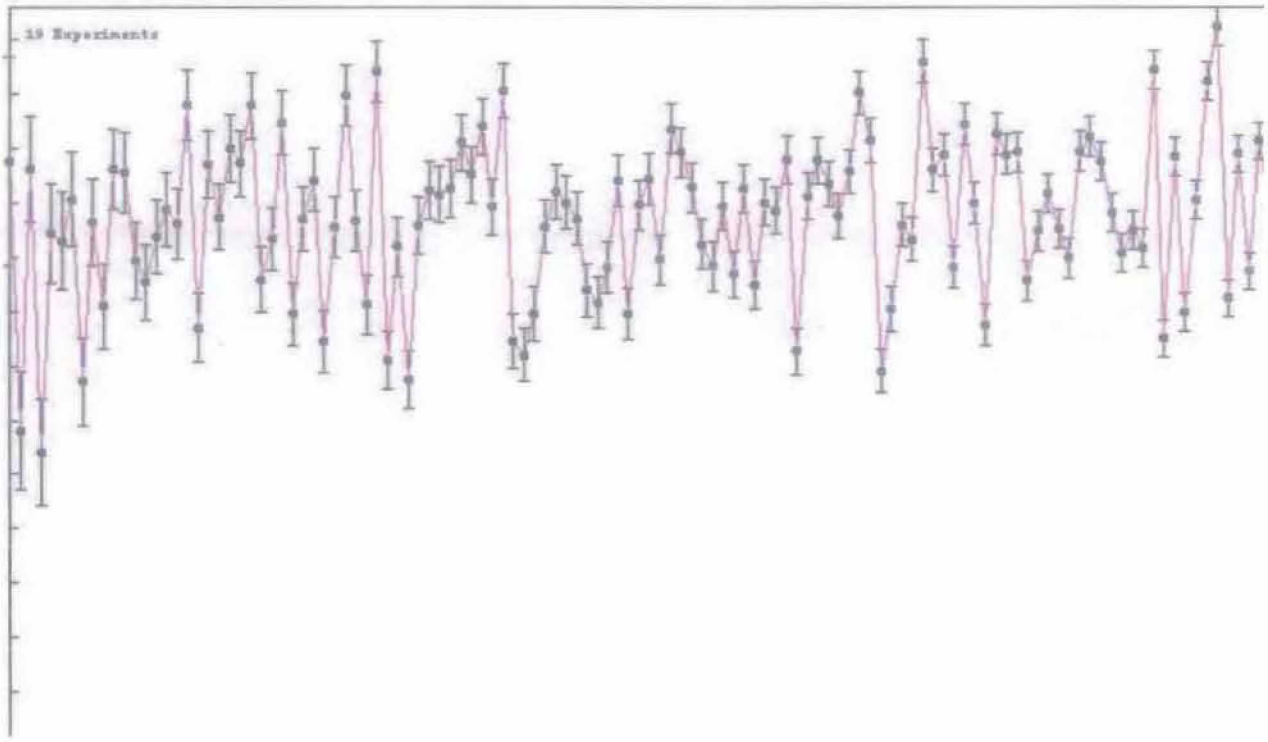
Σχήμα 4.8: *Γράφημα ομαδοποίησης με κέντρο βάρους*



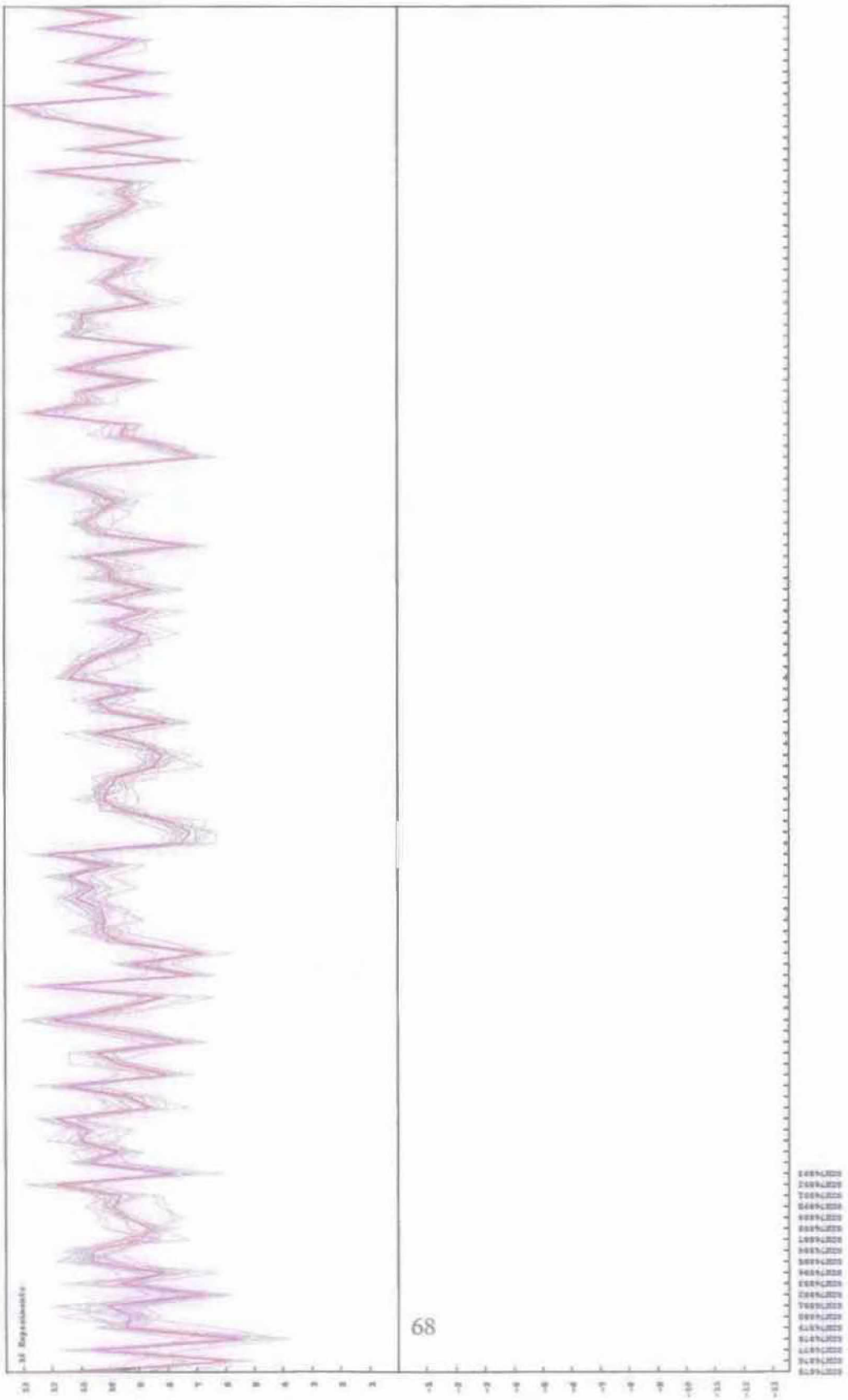
Σχημα 4.9: *Γράφημα ομαδοποίησης με έκφραση*

Cluster 1	# of Genes in Cluster: 11 % of Genes in Cluster: 9% Last iteration with Gene Exchange: 1
Cluster 2	# of Genes in Cluster: 14 % of Genes in Cluster: 11% Last Iteration with Gene Exchange: 2
Cluster 3	# of Genes in Cluster: 8 % of Genes in Cluster: 6% Last Iteration with Gene Exchange: 2
Cluster 4	# of Genes in Cluster: 11 % of Genes in Cluster: 9% Last Iteration with Gene Exchange: 3
Cluster 5	# of Genes in Cluster: 15 % of Genes in Cluster: 12% Last Iteration with Gene Exchange: 4
Cluster 6	# of Genes in Cluster: 12 % of Genes in Cluster: 10% Last Iteration with Gene Exchange: 5
Cluster 7	# of Genes in Cluster: 11 % of Genes in Cluster: 9% Last Iteration with Gene Exchange: 5
Cluster 8	# of Genes in Cluster: 20 % of Genes in Cluster: 16% Last Iteration with Gene Exchange: 6
Cluster 9	# of Genes in Cluster: 14 % of Genes in Cluster: 11% Last Iteration with Gene Exchange: 6
Cluster 10	# of Genes in Cluster: 9 % of Genes in Cluster: 7% Last Iteration with Gene Exchange: 7

Σχήμα 4.10: *Διαχωρισμός στα 10 clusters*



Σχήμα 4.11: *Γράφημα ομαδοποίησης με έκφραση*



Σχήμα 4.12: *Γράφημα ομαδοποίησης με κέντρο βάρους*

SVM Mode: Training and Classification

Total Number of Experiments: 19

Positive Experiments
# of Experiments initially selected as positive examples: 10
# of Experiments classified as positive (Total Positives): 19
# of Experiments retained in positive class (True Positives): 10
# of Experiments recruited into positive class from Negatives (False Negatives): 9

Negative Experiments
# of Experiments initially selected as negative examples: 9
# of Experiments classified as negative (Total Negatives): 0
# of Experiments retained in negative class (True Negatives): 0
# of Experiments recruited into negative class from Positives (False Positives): 0

Σχήμα 4.13: *Διαδικασία Ταξινόμησης*

Κεφάλαιο 5

Επίλογος

Είναι γεγονός, πως η ραγδαία ανάπτυξη της τεχνολογίας τις τελευταίες δεκαετίες έχει οδηγήσει σε πολύ σημαντικές εξελίξεις, οι οποίες έχουν επηρεάσει σε σημαντικό βαθμό πλήθος πτυχών της κοινωνίας. Μία από αυτές τις πτυχές, είναι και αυτή της βιοϊατρικής, δηλαδή του κλάδου που προσπαθεί να αναπτύξει της δυνατότητες της ιατρικής μέσω της εξέλιξης της τεχνολογίας.

Η μελέτη μικροσυστοιχιών γονιδίων αποτελεί ένα βασικό κομμάτι του κλάδου αυτού, που αναπτύσσεται πολύ τα τελευταία χρόνια. Οι μικροσυστοιχίες γονιδίων αυτές είναι μία διάταξη μικροσκοπικών σημείων που αντιπροσωπεύουν μοναδικά γονίδια και ακινητοποιούνται με ομοιοπολικούς δεσμούς σε μία στερεή επιφάνεια (συνήθως γυάλινη). Χρησιμοποιούνται για τη μέτρηση DNA ή χρησιμοποιούν DNA για το σύστημα ανίχνευσής τους. Ποσοτικές ή ποιοτικές μετρήσεις με μικροσυστοιχίες γονιδίων εκμεταλλεύονται την εκλεκτική φύση της αρχής της συμπληρωματικότητας μεταξύ νουκλεϊκών οξέων DNA-DNA ή DNA-RNA ή πρόσφατα και μεταξύ των αμινοξέων των πρωτεϊνών, υπό αυστηρά ελεγχόμενες συνθήκες θερμοκρασίας και με τη χρήση φθορίζουσων ουσιών. Οι μικροσυστοιχίες γονιδίων χρησιμοποιούνται σήμερα κατά κόρον για την εξέταση της γονιδιακής έκφρασης υπό ειδικές συνθήκες και για την ανίχνευση νουκλεϊκών οξέων παθογόνων οργανισμών π.χ. επιβλαβών ιών σε δείγματα ελέγχου. Πρέπει βέβαια να σημειωθεί, όπως είδαμε και στα προηγούμενα, ότι η τεχνολογία αυτή βασίζεται σε μεγάλο βαθμό σε συγκέντρωση μεγάλου όγκου δεδομένων, τα οποία απαιτούν επεξεργασία. Έτσι, καθώς αναπτύσσεται αυτός ο κλάδος, έχει κριθεί απαραίτητο να χρησιμοποιηθούν τεχνικές υπολογιστικής νοημοσύνης και διάφοροι σχετικοί αλγόριθμοι, ώστε να είναι εφικτά τα καλύτερα αποτελέσματα.

Έτσι λοιπόν, στα πλαίσια αυτής της εργασίας έγινε μια σχετική μελέτη, παρουσιάζοντας καταρχήν τα βασικά χαρακτηριστικά των μικροσυστοιχιών DNA. Εν συνεχεία, έγινε αναφορά στα κυριότερα σημεία που αφορούν στις τεχνικές υπολογιστικής νοημοσύνης και σε κάποιους αλγορίθμους αλγορίθμων που χρησιμοποιούνται για τέτοιες εφαρμογές, όπως για παράδειγμα ο αλγόριθμος k-means.

Ολοκληρώνοντας το θεωρητικό κομμάτι, παρουσιάστηκε το πακέτο λογισμικού MEV, που χρησιμοποιείται για τέτοιες διαδικασίες. Τέλος, μέσω του προγράμματος αυτού, υλοποιήθηκαν κάποια παραδείγματα μελέτης μικροσυστοιχιών DNA, με δεδομένα τα οποία υπέστησαν την κατάλληλη επεξεργασία, από μείωση όγκου με χρήση του αλγορίθμου PCA μέχρι clustering και classification.

Στο πειραματικό μέρος, αυτό που έγινε ήταν να χρησιμοποιηθούν δεδομένα που βρίσκονται έτοιμα σε αρχεία στο πρόγραμμα, και να μελετηθεί η λειτουργία των ανωτέρω διαδικασιών. Έτσι, καταρχήν τα δεδομένα εισήχθησαν στο πρόγραμμα, μέσω του αντίστοιχου παραθύρου διαλόγου, και στη συνέχεια έγινε εφαρμογή του PCA αλγόριθμου. Τα αποτελέσματα της εφαρμογής αυτής, παρουσιάστηκαν αναλυτικά στην προηγούμενη ενότητα, με πληθώρα εικόνων που παράγει το πρόγραμμα MEV. Εν συνεχεία, χρησιμοποιήθηκε η διαδικασία ομαδοποίησης μέσω της τεχνικής K - means clustering. Για την υλοποίηση της τεχνικής αυτής, επιλέχθηκε να χρησιμοποιηθούν δέκα ομάδες (clusters). Η ομαδοποίηση των δεδομένων βάσει των ομάδων αυτών, παρουσιάστηκε τόσο με τη βοήθεια γραφημάτων έκφρασης, όσο και με τη βοήθεια γραφημάτων κέντρου βάρους. Τέλος, στο πειραματικό μέρος πραγματοποιήθηκε η διαδικασία της ταξινόμησης των δεδομένων. Η ταξινόμηση αυτή, έγινε με τη χρήση της τεχνικής Support Vector Machine (SVM). Όπως και στα προηγούμενα, έτσι και εδώ χρησιμοποιήθηκαν για την απεικόνιση των αποτελεσμάτων τόσο γραφήματα κέντρου βάρους, όσο και γραφήματα έκφρασης. Επιπλέον, αξίζει να σημειωθεί ότι τα αποτελέσματα της ταξινόμησης ήταν όλα τα δείγματα να κατηγοριοποιηθούν τελικά σε μία εκ των δύο κατηγοριών.

Βιβλιογραφία

[A.02] Butte A. The use and analysis of microarray data. In *Nat Rev Drug Discov., 1(12):951-60*, 2002.

[AJZ02] Sturn A, Quackenbush J, and Trajanoski Z. Genesis: cluster analysis of microarray data. In *Bioinformatics., 18(1):207-8.*, 2002.

[CCJP06] Heichinger C, Penkett CJ, Bahler J, and Nurse P. Genome-wide characterization of fission yeast dna replication origins. In *EMBO J., 25(21):5171-9*, 2006.

[D.03] Stekel D. Microarray bioinformatics. In *Cambridge University Press, UK*, 2003.

[DXGM06] Allison DB, Cui X, Page GP, and Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. In *Nat Rev Genet., 7(1):55-65*, 2006.

[G.A02] Churchill G.A. Fundamentals of experimental design for cdna microarrays. In *Nature Genet., 32, 490⊠495*, 2002.

[GT03] Smyth GK and Speed T. Normalization of cdna microarray data. In *Methods. 31(4):265-73*, 2003.

[IEJ+02] Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, and Quackenbush J. Within the fold: assessing differential expression measures and reproducibility in microarray assays. In *Genome Biol. 3(11)*, 2002.

[J.01] Quackenbush J. Computational analysis of microarray data. In *Nat Rev Genet., 2(6):418-27. Review*, 2001.

[JD06] Hoheisel JD. Microarray technology: beyond transcript profiling and genotype analysis. In *Nat Rev Genet. 7(3):200-10. Review*, 2006.

[KL.07] Kroll KL. Geminin in embryonic development: coordinating transcription and the cell cycle during differentiation. In *Front Biosci., 12:1395?409. Review*, 2007.

[Koh92] T. Kohonen. Self-organized formation of topologically correct feature maps. In *Biol Cybernetics, 43:59-69.*, 1992.

[LM04] Luo L. and Kessel M. Geminin coordinates cell cycle and developmental control. In *Cell Cycle. 3(6):711-4*, 2004.

[LW.01] C. Li and W. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. In *Proc. Natl Acad. Sci., USA 98, 31⊠36*, 2001.

[M.96] Schena M. Genome analysis with gene expression microarrays. In *Bioessays., 18(5):427-31. Review*, 1996.

[M.03] Schena M. Microarray analysis. In *John Wiley & Sons, Inc., Hoboken, New Jersey*, 2003.

[M.08] Ringner M. What is principal component analysis nat biotechnol. In *26(3):303-4.*, 2008.

[MPPD98] Eisen MB, Spellman PT, Brown PO, and Botstein D. Cluster analysis and display of genome?wide expression patterns. In *Proc Natl Acad Sci, 95(25):14863-8.*, 1998.

[MWD+00] Brown M.P., Grundy W.N., Lin D., Cristianini N., Sugnet C.W., Furey T.S., Ares M. Jr., and Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. In *Proc. Natl. Acad. Sci., 97:262-267.*, 2000.

[PD99] Brown PO and Botstein D. Exploring the new world of the genome with dna microarrays. In *Nat Genet.,21(1):33-7. Review*, 1999.

[PDJ+99] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, and Golub TR. Interpreting patterns of gene expression with self?organizing maps: methods and application to hematopoietic differentiation. In *Proc Natl Acad Sci, 96(6):2907-12*, 1999.

[R.03] Movellan J. R. Tutorial on principal component analysis. 2003.

[SCSF00] A. Soukas, P. Cohen, N.D. Socci, and J.M. Friedman. Leptin-specific patterns of gene expression in white adipose tissue. In *Genes Dev., 14:963-980.*, 2000.

[SK06] Seo S and Kroll KL. Geminin's double life: chromatin connections that regulate transcription at the transition from proliferation to differentiation. In *Cell Cycle., 5(4):374-9*, 2006.

[U.05] Nuber U. Dna microarrays. In *Taylor & Francis Group, UK*, 2005.

[VBH07] H. Van Bakel and F.C.P. Holstege. A tutorial for dna microarray expression profiling. In *Cambridge, MA: Cell Press*, 2007.

[VRG01] Tusher VG, Tibshirani R, and Chu G. Significance analysis of microarrays applied to the ionizing radiation response. In *Proc Natl Acad Sci, 98(18):10515*, 2001.

[WS79] Cleveland W.S. Robust locally weighted regression and smoothing scatterplots. In *J. Amer. Stat. Assoc., 74, 829-836*, 1979.

[WS.06] Noble WS. What is a support vector machine? nat biotechnol. In *4(12):1565-7. Review*, 2006.

[Yan02] Y.H. et al. Yang. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. In *Nucleic Acids Res., 30, e15*, 2002.

[YD03] Leung YF and Cavalieri D. Fundamentals of cdna microarray data analysis. In *Trends Genet. 19(11):649-59.*, 2003.