ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Σύντηξη Αλγορίθμων για την Επιλογή Χαρακτηριστικών Δεικτών από Μεγάλες Βάσεις Γονιδιακών Δεδομένων»

ΣΟΥΝΑΠΟΓΛΟΥ ΘΩΜΑΣ

2000030063

Εξεταστική Επιτροπή

Ζερβάκης Μιχάλης, *καθηγητής* (επιβλέπων)

Μπάλας Κωνσταντίνος, *αναπληρωτής καθηγητής*

Πετράκης Ευριπίδης, *αναπληρωτής καθηγητής*

Ευχαριστίες

Αρχικά, θέλω να ευχαριστήσω τον καθηγητή μου, κύριο Ζερβάκη για την πολύτιμη βοήθεια και αρωγή του στην εκπλήρωση αυτής της διπλωματικής εργασίας.

Επίσης, θέλω να ευχαριστήσω τον υποψήφιο διδάκτορα Μιχάλη Μπλαζαντωνάκη για τη διαρκή του υποστήριξη, καθόλη την περίοδο της εργασίας μέχρι και την εκπλήρωσή της.

Τέλος, ευχαριστώ τα μέλη της εξεταστικής επιτροπής για την παρουσία και την ενασχόλησή τους με το κείμενο της διπλωματικής εργασίας.

Περιεχόμενα

περιεχόμενα εικόνων- διαγραμμάτων- πινάκων

- **εικόνες**

- **διαγράμματα**

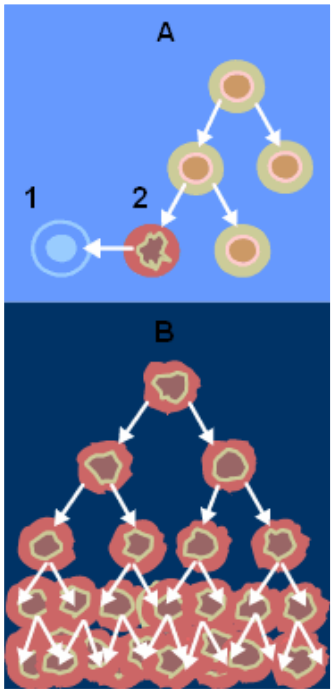
- πίνακες

Κεφάλαιο1: Εισαγωγή

Α. Εισαγωγή – Cancer Genomics

Γενετική έρευνα για τον καρκίνο

Ο καρκίνος είναι μια κατηγορία ασθενειών ή διαταραχών που χαρακτηρίζονται από ανεξέλεγκτη διαίρεση κυττάρων και την δυνατότητα αυτών των κυττάρων να εισβάλουν σε άλλους ιστούς, είτε από την άμεση αύξηση στον παρακείμενο ιστό μέσω εισβολής είτε από την εμφύτευση σε απόμακρες περιοχές μέσω μετάστασης. Αυτή η ανεξέλεγκτη αύξηση προκαλείται από ζημία στο DNA, με συνέπεια τις μεταλλάξεις στα γονίδια που ελέγχουν την κυτταροδιαίρεση. Διάφορες μεταλλάξεις μπορούν να απαιτηθούν για να μετασχηματίσουν ένα κανονικό κύτταρο σε ένα κακοήθες κύτταρο. Αυτή η διαδικασία, αν αφεθεί ανεξέλεγκτη, συνήθως έχει ως αποτέλεσμα το θάνατο. Κάθε χρόνο εμφανίζονται περίπου επτά εκατομμύρια νέα περιστατικά καρκίνου σε όλο τον κόσμο. Από το έτος 2000 ο καρκίνος είναι η υπ' αριθμόν ένα αιτία θανάτου στις βιομηχανικές χώρες, με τάσεις αύξησης τόσο της συχνότητας εμφάνισής του όσο και της διάδοσής του. Ενώ αναπτύσσονται νέες θεραπείες βασισμένες σε μεγάλο φάσμα νέων μηχανισμών, υπάρχουν ακόμα τεράστιες ανάγκες που δεν καλύπτονται ιατρικά.



Εικόνα1: σχηματική απεικόνιση δημιουργίας ενός καρκινικού κυττάρου:
Α- κανονική διαίρεση κυττάρων,
Β- καρκινική διαίρεση κυττάρων
1 - απόπτωση,
2 - κατεστραμμένο κύτταρο.

Οι όποιες προσπάθειες για να ορίσουμε υποκατηγορίες στη βάση της μορφολογίας έχουν σε πολύ μεγάλο βαθμό αποτύχει, εξαιτίας των διαγνωστικών ασυμφωνιών που προκύπτουν από την μεταξύ και ενδιάμεσα αδυναμία αναπαραγωγής των παρατηρήσεων

Έχει παρατηρηθεί ότι οι ασθενείς που υποφέρουν από καρκίνο του μαστού και βρίσκονται στην ίδια ακριβώς φάση της ασθένειας, μπορούν να έχουν αξιοσημείωτα διαφορετική απόκριση στην θεραπεία που υποβάλλονται. Οι πιο ισχυροί μέχρι τώρα παράγοντες που μπορούν να προβλέψουν μεταστάσεις (predictors), δεν μπορούν να ταξινομήσουν τους όγκους του μαστού σύμφωνα με την κλινική τους συμπεριφορά. Η χημειοθεραπεία ή η ορμονική θεραπεία μειώνουν τον κίνδυνο άμεσης μετάστασης κατά 30% περίπου, αλλά το 70%-80% των ασθενών που υπόκεινται σε αυτή τη θεραπεία θα μπορούσαν να είχαν επιβιώσει και χωρίς αυτήν.

Για αυτόν, ακριβώς το λόγο, υπάρχει αυξημένο επιστημονικό ενδιαφέρον στην προσπάθεια ταξινόμησης (classification) του όγκου με βάση όχι μορφολογικά στοιχεία, αλλά μοριακά. Η γονιδιακή έκφραση των όγκων μπορεί να προσφέρει πολύ περισσότερη πληροφορία από τη μορφολογική και να προωθήσει εναλλακτικά σε αυτή, πιο αξιόπιστα συστήματα ταξινόμησης των όγκων.

Περιγραφή της Εργασίας

Στη διαδικασία ταξινόμησης των φαινοτύπων, δημιουργούνται πολύ σημαντικά προβλήματα στην απόδοση της ταξινόμησης, όπως επίσης και στο υψηλότατο κόστος αυτής. Το παραπάνω πρόβλημα μπορεί να επιλυθεί με την επιλογή των πιο σημαντικών γονιδίων (informative genes), των οποίων ο συνδυασμός μπορεί να συγκεντρώσει τη συνολική πληροφορία του συνόλου του γονιδιώματος.

Κατά τη διαδικασία αυτή, υπάρχουν στη βιβλιογραφία πολλές μέθοδοι, χωρίς όμως κάποια από αυτές να οδηγεί σε αξιόπιστα αποτελέσματα για την επιλογή ενός ελάχιστου συνόλου γονιδίων στα οποία θα επικεντρωθεί η μελέτη της βιολογίας, στην κατεύθυνση της δυνατότητας πρόβλεψης τυχόν καρκινικής προδιάθεσης.

Σκοπός της εργασίας είναι η υλοποίηση της αλγοριθμικής σύντηξης ως μέθοδος συνδυασμού ενός συνόλου παραμετρικών και μη παραμετρικών μεθόδων υπολογισμού. Τελικός στόχος της μεθόδου είναι η εξαγωγή ενός ελάχιστου αριθμού γονιδίων που θα μπορεί να «προβλέψει» τα κλινικά αποτελέσματα καρκινικών παθήσεων.

Για την υλοποίηση της αλγοριθμικής σύντηξης, η εργασία επικεντρώθηκε αρχικά σε δεδομένα που αφορούν τον καρκίνο του μαστού, σε σύνολο 25.000 γονιδίων. Μετά τα πρώτα αποτελέσματα, έγινε εφαρμογή της μεθόδου και σε άλλα σύνολα δεδομένων, που αφορούσαν την ασθένεια της λευχαιμίας, όπως επίσης και το καρκίνο του μαστού, αφού πρώτα υπήρξε προεπεξεργασία των δεδομένων, που οδήγησε στη μελέτη 5000 γονιδίων. Επίσης, και για τα τρία παραπάνω σύνολα δεδομένων εφαρμόσαμε τη μέθοδο της συσχέτισης συντελεστών ώστε να προκύψουν συγκρίσιμα αποτελέσματα σε σχέση με την υπό μελέτη μέθοδο της αλγοριθμικής σύντηξης. Τέλος, βελτιστοποιήσαμε το κριτήριο επιλογής του αλγορίθμου που υλοποιεί την αλγοριθμική σύντηξη.

Ως γενικές παρατηρήσεις για τα αποτελέσματά μας, μπορούμε να παρατηρήσουμε ότι, η μέθοδος της αλγοριθμικής σύντηξης μπορεί να αποτελέσει μια αξιόπιστη μέθοδο για την επιλογή χαρακτηριστικών γονιδίων στη μελέτη πρόβλεψης των καρκινικών παθήσεων, ιδιαίτερα σε σχέση με πιο συμβατικές μεθόδους, όπως είναι αυτή της συσχέτισης συντελεστών. Σημαντικό ρόλο στη μελέτη της καταλληλότητας της μεθόδου έχει και το πεδίο στο οποίο εφαρμόζεται, δηλαδή το σύνολο δεδομένων, καθώς ανάλογα με την ποιότητά του (ύπαρξη θορύβου ή όχι), αλλά και την ασθένεια στην οποία αναφέρεται, μπορεί να καταστήσει τη μέθοδο περισσότερο ή λιγότερο κατάλληλη για κάποιο είδος καρκίνου.

Β. Microarray Τεχνολογία

Η microarray τεχνολογία είναι μια πολλά υποσχόμενη προσέγγιση για την ανάλυση μεγάλου μεγέθους δεδομένων και μας δίνει τη δυνατότητα να μελετήσουμε τα πρότυπα της γονιδιακής έκφρασης σε μια γενετική βαθμίδα. Με τη βοήθεια των microarrays μπορούμε να καθορίσουμε χιλιάδες τιμές έκφρασης σε εκατοντάδες διαφορετικές περιπτώσεις, επιτρέποντας τη συγκέντρωση των γενετικών διεργασιών σε μια συνολική γενετική βαθμίδα. Με αυτόν τον τρόπο μπορούμε να καθορίσουμε τη συνεισφορά του γενετικού υλικού σε πολύπλοκές

ανωμαλίες και να εξετάσουμε τις αλλαγές που μπορεί να υπάρχουν σε γονίδια (διαφορετικά επίπεδα έκφρασης) κατά την εκδήλωση ασθενειών.
Συνοπτικά, τα microarrays είναι:

- Ένα πείραμα της τάξης των 10000 στοιχείων
- Ένας τρόπος να εξερευνήσουμε τις λειτουργίες των γονιδίων
- Ένα στιγμιότυπο του επίπεδο έκφρασης ενός ολόκληρου φαινότυπου κάτω από συγκεκριμένες συνθήκες.

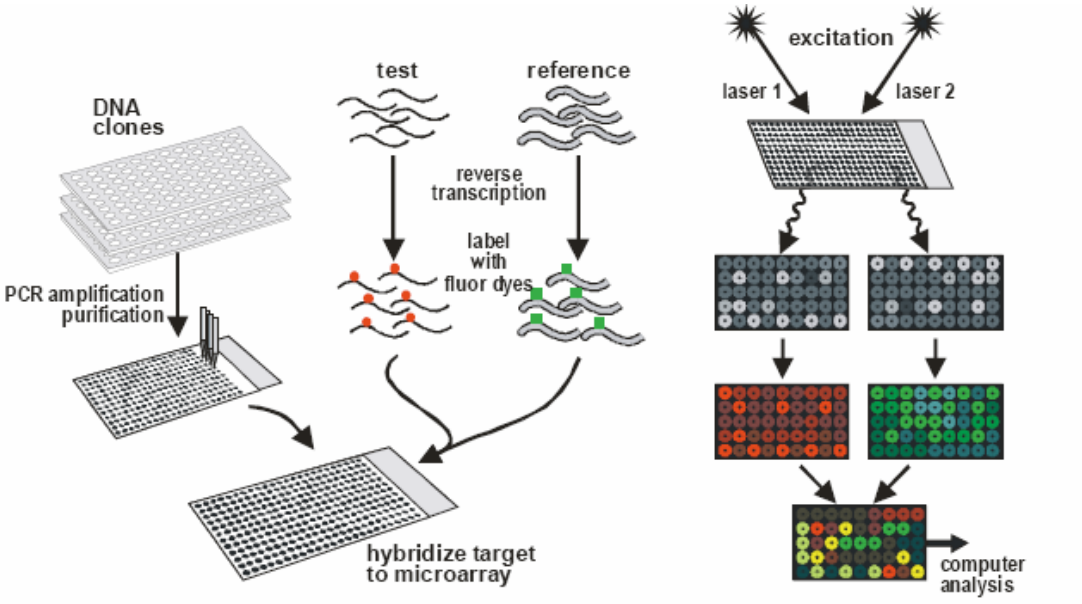
Η ουσία και οι λειτουργίες των κυττάρων χαρακτηρίζονται από την παραγωγή των πρωτεϊνών, οι οποίες αποτελούνται από αμινοξέα. Τα πρότυπα των αλληλουχιών των αμινοξέων κωδικοποιούνται σε γονίδια, τα οποία αποτελούν μέρος του γονιδιώματος. Για την παραγωγή της πρωτεΐνης, η οποία ορίζεται από το γονίδιο, ένα κύτταρο μεταγράφει τη γενετική του αλληλουχία σε αλληλουχία αγγελιοφόρου ριβονουκλεϊκού οξέος (mRNA). Έπειτα, η αλληλουχία mRNA μεταφράζεται σε τριπλέτες σε αλληλουχία αμινοξέων. Η **γενετική έκφραση (gene expression)** αναφέρεται στο επίπεδο παραγωγής των πρωτεϊνικών μορίων που ορίζονται από ένα γονίδιο. Η απεικόνιση της γενετικής έκφρασης είναι μια από τις πιο θεμελιακές προσεγγίσεις στη γενετική και μοριακή βιολογία. Η βασική τεχνική για τον υπολογισμό της γονιδιακής έκφρασης είναι να υπολογίσουμε το mRNA αντί των πρωτεϊνών, γιατί οι αλληλουχίες mRNA είναι υβριδικές με αυτές του συμπληρωματικού ριβονουκλεϊκού οξέος (cRNA) ή του δεοξυριβονουκλεϊκού οξέος (DNA), ιδιότητα η οποία απουσιάζει από τις πρωτεΐνες.

Οι πίνακες DNA (DNA arrays) είναι νέες τεχνολογίες οι οποίες έχουν σχεδιαστεί έτσι, ώστε να υπολογίζουν τη γενετική έκφραση δεκάδων χιλιάδων γονιδίων σε ένα μόνο πείραμα. Ένας πίνακας DNA αποτελείται από αλληλουχίες probe DNA, οι οποίες είναι εναποθετημένες σε μια επιφάνεια χρυσού ή γυαλιού. Για να αποτιμήσουμε τη γενετική έκφραση των ιστών, εξάγουμε αρχικά από αυτούς τις αλληλουχίες του mRNA. Οι αλληλουχίες mRNA, έπειτα, μεταγράφονται αντίστροφα σε αλληλουχίες DNA (cDNA). Ένα microarray είναι ένα ειδικά επικαλυμμένο γυαλί μικροσκοπίου στο οποίο τα μόρια cDNA προσκολλώνται σε σταθερά σημεία, τα οποία ονομάζονται σημεία (spots). Σύμφωνα με την πιο

σύγχρονη τεχνολογία 19.200 και περισσότερα σημεία μπορούν να τυπωθούν σε ένα γυαλί μικροσκοπίου, το καθένα από τα οποία αναπαριστά ένα μοναδικό γονίδιο. Οι αντίστροφα- μεταγραφόμενες αλληλουχίες DNA (cDNA), λαμβάνουν φθορίζουσα ετικέτα.

Οι ταμπέλες που λαμβάνουν τα cDNA probes είναι είτε Cy3 dye είτε Cy5 dye. Τα Fluorescent cDNA probes, που έχουν ετικέτα Cy5 προέρχονται από κάθε δείγμα mRNA του πειράματος, ενώ αυτά με ετικέτα Cy3 προέρχονται από μια δεξαμενή των mRNA. Κάθε πειραματικό cDNA probe με ετικέτα Cy5 συνδυάζεται με cDNA probes που έχουν ετικέτα Cy3. Όλα τα probes ταυτόχρονα «επωάζονται» στο microarray, καθιστώντας δυνατή στις αλληλουχίες των γονιδίων την συνένωσή τους (hybridization), κάτω από αυστηρές συνθήκες, στους συμπληρωματικούς τους κλώνους και να εναποτίθενται στην επιφάνεια του πίνακα.

Έπειτα, με τη διέγερση του λέιζερ των ενσωματωμένων στόχων, μπορούμε να πετύχουμε μια εκπομπή με ένα χαρακτηριστικό φάσμα, το οποίο μετριέται με τη βοήθεια ενός σαρωτικού, ομοεστιακού μικροσκοπίου. Τότε εισάγονται στο λογισμικό μας μονόχρωμες εικόνες, οι οποίες ψευδο- χρωματίζονται και συγχωνεύονται.



Εικόνα 2: cDNA microarray διάγραμμα. Αρχικά, οι κλώνοι DNA τοποθετούνται κατά διαστήματα πάνω σε ένα γυαλί μικροσκοπίου. Μετά τη συνένωση των δύο συμπληρωματικών αλυσίδων DNA (hybridization) το γυαλί σαρώνεται με τη διέγερση λέιζερ και μας δίνει δυο εικόνες για περαιτέρω επεξεργασία.

Ο λόγος φθορισμού (fluorescence ratio) ποσοτικοποιείται για κάθε γονίδιο και αντανακλά τη σχετική πληθώρα του γονιδίου σε κάθε πειραματικό δείγμα του mRNA σε σύγκριση με τη δεξαμενή αναφοράς mRNA. Η χρήση ενός ανιχνευτή κοινής αναφοράς μας επιτρέπει να θεωρήσουμε αυτούς τους λόγους φθορισμού ως μέτρα του σχετικού επιπέδου έκφρασης κάθε γονιδίου σε όλα τα πειραματικά μας δείγματα.

Κεφάλαιο2: Περιγραφή Προβλημάτων και Λύσεων

Α. Γονιδιακή επιλογή (Gene Selection)

Πρόσφατες μελέτες έχουν δείξει ότι η microarray γονιδιακή έκφραση των δεδομένων (microarray gene expression data) μπορεί να φανεί χρήσιμη στην ταξινόμηση (classification) φαινοτύπων σε πολλές περιπτώσεις ασθενειών. Σε αυτό το πρόβλημα ο αριθμός των χαρακτηριστικών (στη συγκεκριμένη περίπτωση των γονιδίων) ξεπερνά σε πολύ μεγάλο βαθμό το μέγεθος των στιγμιότυπων (δείγματα ιστών). Το γεγονός αυτό μπορεί σε πολύ μεγάλο βαθμό να βλάψει την απόδοση της ταξινόμησης καρκινικών ιστών , όπως επίσης και να αυξήσει το κόστος. Έχει επίσης αποδειχθεί, ότι επιλέγοντας ένα μικρό μέγεθος γονιδίων (informative genes) μπορεί να οδηγήσει σε βελτιωμένα αποτελέσματα ακρίβειας. Το παραπάνω πρόβλημα μπορεί να οριστεί ως **πρόβλημα γονιδιακής επιλογής (gene selection problem).** Η γονιδιακή επιλογή περιλαμβάνει την αναζήτηση των υποσυνόλων ομάδων γονιδίων που μπορούν να διακρίνουν τον καρκινικό ιστό από τον κανονικό και μπορούν να έχουν κάποια ευθεία ή έμμεση συμμετοχή στο μοριακό μηχανισμό της ογκογένεσης. Πιο συγκεκριμένα, η γονιδιακή επιλογή περιλαμβάνει τη μελέτη των κριτηρίων και την αναπαράσταση των τεχνικών που έχουν ως σκοπό τη μείωση του αριθμού των γονιδίων και την επιλογή ενός βέλτιστου (ή κοντά στο βέλτιστο) υποσυνόλου γονιδίων από κάποιο αρχικό σύνολο γονιδίων για την ταξινόμηση όγκων (tumor classification).

Τα πρακτικά πλεονεκτήματα της γονιδιακής επιλογής, τα οποία ισχύουν ακόμα και σε σχέση με κάποιες άλλες μεθόδους μείωσης των διαστάσεων ενός αρχικού συνόλου δεδομένων /dataset (principal component analysis κ.ά.) είναι:

- η μείωση της πολυπλοκότητας. Παρόλο που δύο χαρακτηριστικά (γονίδια) μπορούν να φέρουν σημαντική πληροφορία για την ταξινόμηση όταν αυτά τα μεταχειριζόμαστε ξεχωριστά, το κέρδος είναι πολύ μικρό όταν συνδυάζουμε τα δύο χαρακτηριστικά μαζί σε ένα διάνυσμα εξαιτίας της υψηλής αμοιβαίας

συσχέτισης. Έτσι με τη γονιδιακή επιλογή μπορούμε να μειώσουμε την πολυπλοκότητα με πολύ μικρή μείωση του κέρδους.

- η μείωση του κόστους υπολογισμού, αφού στη θέση ενός τεράστιου συνόλου δεδομένων (dataset) υπάρχει ένα σαφώς πολύ μικρότερο σύνολο για επεξεργασία, μειώνοντας έτσι, τις απαιτήσεις για τις μετρήσεις και την αποθήκευση των αποτελεσμάτων.

- η διατήρηση της πολύ βασικής ιδιότητας για την ταξινόμηση (classification) της γενίκευσης (generalization), αφού όσο μεγαλύτερη είναι η αναλογία των εκπαιδευόμενων δειγμάτων προς τις ελεύθερες παραμέτρους του ταξινομητή, τόσο καλύτερο είναι το αποτέλεσμα της ταξινόμησης που προκύπτει. Ένας μεγάλος αριθμός χαρακτηριστικών (γονιδίων) μεταφράζεται ευθέως σε μεγάλο αριθμό παραμέτρων ταξινόμησης (π.χ. τα βάρη σε έναν γραμμικό ταξινομητή ή τα συναπτικά βάρη σε ένα νευρωνικό δίκτυο). Έτσι, για έναν περιορισμένο αριθμό δειγμάτων ιστών, μειώνοντας τον αριθμό των γονιδίων όσο περισσότερο γίνεται, μπορούμε να επιτύχουμε καλύτερη γενίκευση στην σχεδιαζόμενη ταξινόμηση.

- η μεγαλύτερη πιθανότητα υιοθέτησης του μοντέλου σε κλινικό περιβάλλον. Τα επιλεχθέντα υποσύνολα γονιδίων με την μεγάλη ακρίβεια στην ταξινόμησης, είναι πιθανόν να εμπλέκονται σε έναν βαθμό στη διαδικασία ανάπτυξης του όγκου. Έτσι, τα υποσύνολα γονιδίων που επιλέγουμε είναι δυνατόν να έχουν σημαντική βιολογική αξία και μπορούν να αποτελέσουν πεδίο έρευνας για την επιστήμη της Βιολογίας στην προσπάθεια της θεραπείας, αλλά και της πρόληψης του καρκίνου.

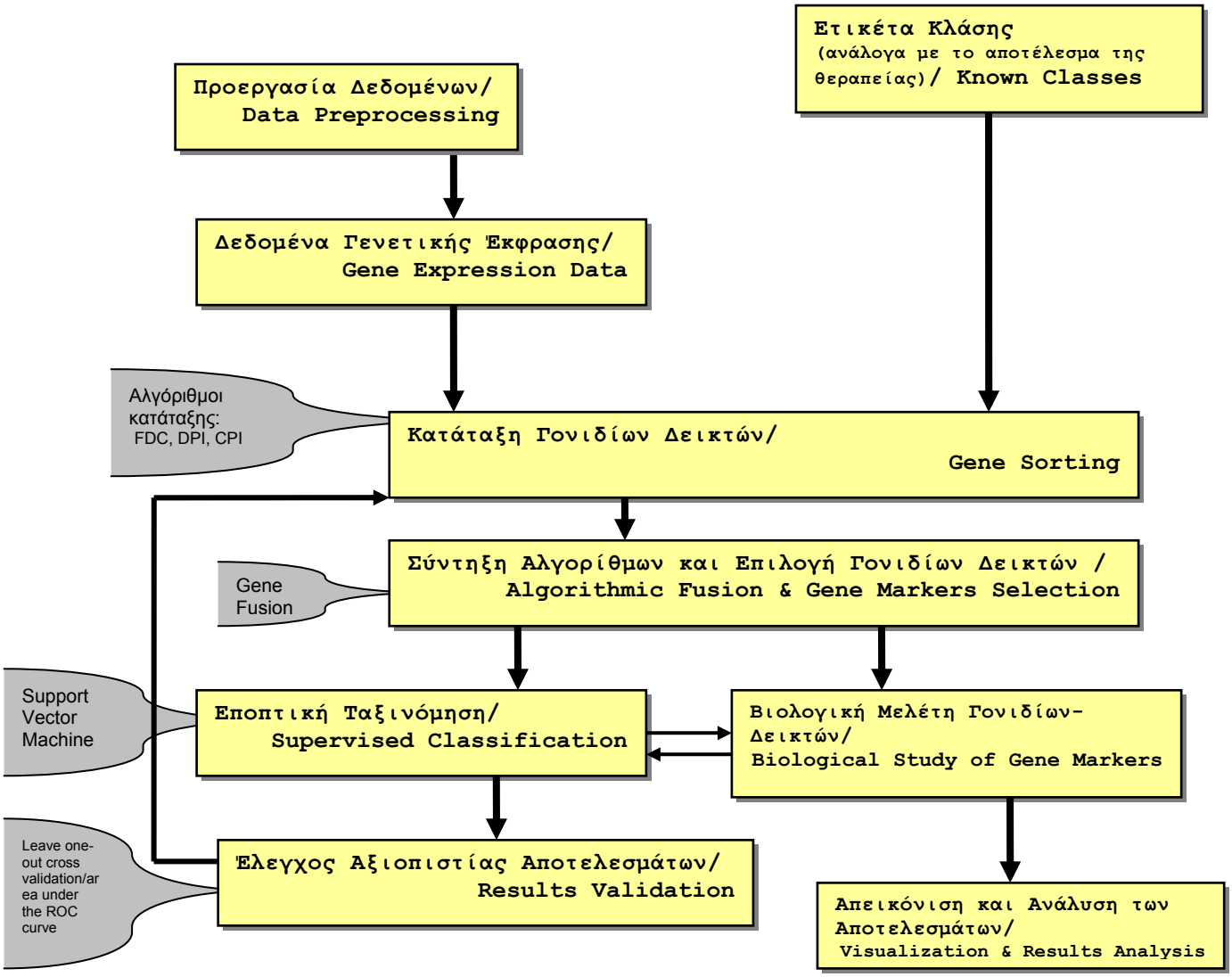
- Η βελτίωση του λάθους ταξινόμησης (classification error), αφού όπως φαίνεται και παραπάνω, ένα πολύ σημαντικό βήμα στο σχεδιασμό της ταξινόμησης είναι το στάδιο της εκτίμησης της απόδοσης, όπου υπολογίζεται η πιθανότητα σφάλματος του ταξινομητή.

Β. Μεθολογία της επιλογής γονιδίων- δεικτών (gene selection markers)

Πολλές διαφορετικές μέθοδοι και τεχνικές έχουν αναλυθεί για τη γενετική επιλογή από σετ δεδομένων γενετικής έκφρασης τεχνολογίας microarray. Η ακρίβεια και η αξιοπιστία των αποτελεσμάτων εξαρτώνται από τις μεθόδους που χρησιμοποιούμε. Από τα μέχρι τώρα αποτελέσματα των μετρήσεων, δεν μπορούμε με την εφαρμογή μιας και μοναδικής μεθόδου να εγγυηθούμε κάποιο ικανοποιητικό αποτέλεσμα.

Για τον παραπάνω λόγο, στην παρούσα διπλωματική εργασία πραγματοποιείται η υλοποίηση και η μελέτη της αλγοριθμικής σύντηξης για την εξαγωγή γονιδίων, τα οποία είναι πολύ πιθανόν να αποτελούν δείκτες (markers) πρόβλεψης των κλινικών αποτελεσμάτων (θανατηφόρα ή όχι) κατά την ανάπτυξη της ασθένειας του καρκίνου.

Η προσέγγισή μας περιλαμβάνει την ανάπτυξη μεθόδων μέτρησης και κατάταξης των γονιδίων βασιζόμενοι σε διαφορετικά κριτήρια σύμφωνα με τις ενδείξεις ασυμφωνίας των προτύπων γονιδιακής έκφρασης (gene expression patterns) και των κλινικών αποτελεσμάτων. Κατά τη διαδικασία της υλοποίησης, χρησιμοποιούμε έναν συνδυασμό στατιστικών και μη-στατιστικών κριτηρίων, συνεχών και διακριτών παραμετροποιήσεων, όπως επίσης και εκτιμήσεις, οι οποίες βασίζονται σε μοντέλα ή όχι. Οι περισσότερες από τις μεθόδους που έχουν μέχρι τώρα αναπτυχθεί επιλέγουν τα πενήντα με διακόσια πρώτα γονίδια της κατάταξης, όπως αυτή διαμορφώνεται από την κάθε μέθοδο. Τα γονίδια αυτά είναι σε πάρα πολλές περιπτώσεις συσχετιζόμενα μεταξύ τους. **Ο στόχος μας είναι η επιλογή ενός όσο το δυνατόν μικρότερου υποσυνόλου γονιδίων δεικτών, τα οποία μπορούν με ακρίβεια να χρησιμοποιηθούν για τη διαδικασία της ταξινόμησης.**



διάγραμμα 1: διάγραμμα της διαδικασίας επιλογής γονιδίων- δεικτών

Στην παραπάνω εικόνα απεικονίζεται σχηματικά η μεθοδολογία της επιλογής γονιδίων- δεικτών.

Γ. Προεργασία Δεδομένων

Τα ακατέργαστα δεδομένα έκφρασης (raw expression data) πρέπει να υποστούν την κατάλληλη επεξεργασία, ώστε να είναι σε θέση να χρησιμοποιηθούν για τη γονιδιακή επιλογή. Έτσι, λοιπόν, κάθε στήλη (δείγμα ιστού) στο σύνολο δεδομένων μας, πολλαπλασιάζεται με τον παράγοντα 1/κλίση, όπως αυτή προκύπτει από το λόγο της γραμμικής προσαρμογής ελάχιστων τετραγώνων

(least squares linear fit) του δείγματος προς ένα δείγμα αναφοράς (το πρώτο δείγμα του συνόλου δεδομένων). Αυτή η γραμμική προσαρμογή επιτυγχάνεται χρησιμοποιώντας μόνο γονίδια που έχουν «παρουσία» (P) σε όλα τα δείγματα που επεξεργαζόμαστε, όπως επίσης και στο δείγμα αναφοράς (κάθε «παρουσία» (P) αναπαριστά ένα γονίδιο με «παρουσία» RNA, όπως αυτή καθορίζεται από την ανάλυση των μέσων διαφορών των μετρήσεων έκφρασης από κάθε σύνολο γονιδίων του microarray). Το δείγμα που επιλέγουμε ως αναφορά είναι συνηθισμένο (π.χ. μπορεί να είναι αυτό που έχει τον αριθμό "P" πιο κοντά στη μέση τιμή όλων των δειγμάτων στο σύνολο δεδομένων). Για την πειραματική επεξεργασία συνήθως επιλέγεται ένα ανώτατο όριο μονάδων γύρω στις 16.000, λόγω του γεγονότος ότι σε αυτό το επίπεδο παρατηρούμε φθορίζουσα χρωματική καθαρότητα του σαρωτή και μπορούμε να έχουμε αξιόπιστες μετρήσεις. Επίσης, καθορίζεται ένα κατώτατο όριο (threshold) για το επίπεδο έκφρασης των 20 μονάδων, έτσι ώστε να μειώσουμε τις επιπτώσεις του θορύβου και να αποφύγουμε κάποια απώλεια πιθανών γονιδίων- δεικτών.

Μετά την προεργασία, οι τιμές της γονιδιακής έκφρασης υπόκεινται σε φιλτράρισμα απόκλισης, όπου τα γονίδια με ελάχιστη απόκλιση αποκλείονται, καθώς διαπερνώνται όλα τα δείγματα ιστού. Το φίλτρο απόκλισης ελέγχει τις αλλαγές τάξης μεγέθους και την απόλυτη απόκλιση σε όλα τα δείγματα (συγκρίνοντας τα μέγιστα/ ελάχιστα και τα ελάχιστα/ μέγιστα με προκαθορισμένες τιμές και αποκλείοντας τα γονίδια εκείνα που δεν εμπίπτουν σε καμία από τις δύο περιπτώσεις.

Δ. Πίνακας Έκφρασης

Τα σχετικά επίπεδα έκφρασης των n γονιδίων, που μπορούν να αποτελούν ολόκληρο το γονιδίωμα ενός οργανισμού, μπορεί να απεικονίζονται σε ένα μοναδικό microarray. Μια σειρά από *m πίνακες*, που είναι σχεδόν πανομοιότυπα φυσικά, απεικονίζουν τα επίπεδα έκφρασης του γονιδιώματος σε *m* διαφορετικά δείγματα. π.χ. κάτω από *m* διαφορετικές πειραματικές συνθήκες.

Έστω ότι ο πίνακας **X,** μεγέθους (*n*-γονίδια x *m*- δείγματα), συνοψίζει ολόκληρα τα δεδομένα έκφρασης, όπου το στοιχείο x_{ij} είναι το log2 του λόγου έκφρασης του $i^{ου}$ γονιδίου στο $j^{ο}$ δείγμα, όπως μετριέται από το array *j*. Το διάνυσμα στην *I* σειρά του πίνακα **X** αναπαριστά το λόγο φθορισμού του $i^{ου}$ γονιδίου έναντι προς τα διαφορετικά δείγματα, που αντιστοιχούν σε διαφορετικούς πίνακες. Το διάνυσμα στην $j^{ο}$ στήλη του πίνακα **X**, αναπαριστά τους λόγους φθορισμού του γονιδιώματος, όπως αυτοί υπολογίζονται στο $j^{ο}$ array.

$$x_{ij} = \log_2 \frac{C5_{ij}}{C3_{ij}}$$

όπου C5ij .....Cye-5 μέτρο φθορισμού του γονιδίου *I* στο microarray πείραμα *j*

C3ij .....Cye-3 μέτρο φθορισμού του γονιδίου *I* στο microarray πείραμα *j*

Το στοιχείο x_{ij} είναι αρνητικό, εάν C3>C5,

0, εάν C3=C5,

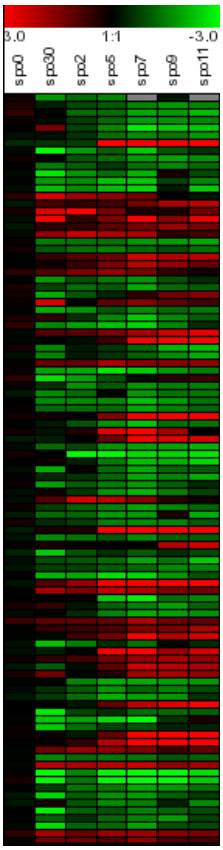
θετικό, εάν C3<C5.

Με άλλα λόγια, το στοιχείο x_{ij} είναι θετικό, εάν το γονίδιο *I* στο πείραμα *j είναι* over expressed, και αρνητικό, εάν το γονίδιο *I* στο πείραμα *j* είναι under expressed σε σχέση με το δείγμα ελέγχου.

Ε. Γραφική Αναπαράσταση

Εξαιτίας του υπερβολικά μεγάλου όγκου αριθμών, το οποίο είναι πολύ δύσκολο να μελετηθεί από μόνο του, τα βασικά δεδομένα συνδυάζονται με τη γραφική αναπαράσταση. Σε αυτήν το σημείο δεδομένου αναπαρίσταται με ένα χρώμα το οποίο ποσοτικά και ποιοτικά αντανακλά την αρχική πειραματική παρατήρηση. Συνήθως κάθε σημείο δεδομένου x_{ij} χρωματίζεται στη βάση του λόγου φθορισμού του. Αυτό μας βοηθά να αναπαραστήσουμε πολύπλοκης γενετικής έκφρασης δεδομένα και να επιτρέψουμε στους βιολόγους να αφομοιώσουν και να εξερευνήσουν τα δεδομένα με έναν πολύ πιο διαισθητικό τρόπο. Στη βιβλιογραφία τα χρησιμοποιούμενα χρώματα ποικίλουν από διαποτισμένο πράσινο (για μέγιστες αρνητικές τιμές) μέχρι διαποτισμένο κόκκινο (για μέγιστες θετικές τιμές. Κύτταρα με log ratio 0 (δηλαδή γονίδια που δεν διαφοροποιούνται) χρωματίζονται μαύρα, κύτταρα με αυξανόμενα θετικά log ratios με κόκκινο

ανάλογης αύξουσας έντασης, ενώ κύτταρα με αυξανόμενα αρνητικά log ratios, με πράσινο ανάλογης αύξουσας έντασης. Για τιμές οι οποίες απουσιάζουν το χρώμα το οποίο θα εμφανίζεται είναι το γκρι. (Sturn2001)



Εικόνα 3: γραφική αναπαράσταση ενός βασικού συνόλου δεδομένων (dataset)

ΣΤ. Κατάταξη Γονιδίων Δεικτών/ Gene Sorting

Στο στάδιο της κατάταξης γονίδιων, σύμφωνα με τη βιβλιογραφία υπάρχουν πολλές μέθοδοι, οι οποίες μπορούν να χρησιμοποιηθούν και για τον απευθείας προσδιορισμό των ενδεικτικών για παθολογικές ανωμαλίες γονιδίων, χωρίς όμως αξιόλογα αποτελέσματα.

Σε γενικές γραμμές οι προσεγγίσεις που υπάρχουν περιλαμβάνουν:

1. **παραμετρικές μεθόδους**,

όπως είναι η ανάλυση κύριων συνιστωσών (principal component analysis), η ανάλυση ανεξάρτητων μεταβλητών (independent component analysis), το μέτρο συσχέτισης διαχωρισμού (separation correlation metric), που είναι περισσότερο γνωστό ως κριτήριο διαχωρισμού το Fisher (Fisher's discrimination criterion-FDC).

Οι παραμετρικές μέθοδοι χρησιμοποιούν ένα σύνολο στατιστικών κριτηρίων, όπως αυτά προκύπτουν από την απεικόνιση της γονιδιακής έκφρασης στη βάση μοντέλων βέβαιης πιθανότητας κατανομής. Οι στατιστικές μέθοδοι είναι συνήθως αξιόπιστες και ακριβείς στην ανάλυση μεγάλων συνόλων δεδομένων. Παρόλα αυτά, η ασυμφωνία μεταξύ του σχετικά μικρού αριθμού δειγμάτων δεδομένων και του σχετικά μεγάλου αριθμού γονιδίων συχνά καθιστούν το όποιο στατιστικό μοντέλο, ανακριβές. Επιπλέον, τα στατιστικά μέτρα εύκολα αποκλίνουν ή διαταράσσονται από την αβεβαιότητα και την ανακρίβεια των γνωστών ετικετών κάθε δείγματος, από την ακριβή κατηγορία κάθε είδους και την διαταραχή των κατανομών των δειγμάτων.

Παράδειγμα παραμετρικής μεθόδου είναι η Ανάλυση Κύριων Συνιστωσών-Principal Component Analysis (PCA), και η Ανάλυση Ανεξαρτήτων Συνιστωσών-Independent Component Analysis (ICA). Η **Ανάλυση Κύριων Συνιστωσών** (PCA) είναι μια μέθοδος γραμμικού μετασχηματισμού του συνόλου δεδομένων, όπου οι άξονες αναπροσαρμόζονται ανάλογα με την διακύμανση των δεδομένων. Έτσι, η μέγιστη διακύμανση σε κάθε προβολή των δεδομένων, τοποθετείται στην πρώτη συντεταγμένη (Πρώτη Κύρια Συνιστώσα), η δεύτερη μεγαλύτερη διακύμανση στη δεύτερη συντεταγμένη κλπ. Με την παραπάνω μέθοδο δηλαδή, έχουμε την «περιστροφή» του συνόλου των δεδομένων σε τέτοια κατεύθυνση, ώστε να είναι εμφανής η μέγιστη διακύμανση. Η **Ανάλυση Ανεξαρτήτων Συνιστωσών** (ICA) είναι μια πολύ πιο δραστική μέθοδος σε σχέση με την ανάλυση κύριων συνιστωσών. Ορίζει ένα μοντέλο για τα μεγάλα σύνολα δεδομένων των δειγμάτων μας. Στο μοντέλο αυτό, οι μεταβλητές δεδομένων θεωρούνται γραμμικές συνθέσεις κάποιων υποβόσκων μεταβλητών, οι οποίες δεν ακολουθούν την κανονική

κατανομή, είναι ανεξάρτητες μεταξύ τους και ονομάζονται ανεξάρτητες συνιστώσες των δεδομένων. Τελικός σκοπός της μεθόδου είναι ο εντοπισμός αυτών των ανεξαρτήτων συνιστωσών, οι οποίες ονομάζονται και πηγές ή παράγοντες.

2. **μη παραμετρικές μεθόδους**,
όπως είναι νευρωνικά δίκτυα (neural networks), αλλά και άλλες μέθοδοι (projection pursuit regression, support vector machines, threshold number of misclassification TnoM). Οι περισσότερες από τις παραπάνω μεθόδους έχουν ως αποτέλεσμα στην απόδοση κάποιας τιμής σε κάθε γονίδιο για την ανίχνευση της σχετικότητας του καθενός.

Οι μη παραμετρικές μέθοδοι δε βασίζονται στην υπόθεση των στατιστικών μοντέλων και παραμέτρων. Αντίθετα, επεξεργάζονται απευθείας τα αντικείμενα, εφαρμόζοντας προκαθορισμένα μη στατιστικά κριτήρια στα δείγματα δεδομένων. Οι συγκεκριμένες μέθοδοι έχουν περισσότερα πλεονεκτήματα στον περιορισμό και την εξασθένιση των αποτελεσμάτων που προκύπτουν από μικρότερο αριθμό δειγμάτων δεδομένων. Παρόλα αυτά, η απόκλιση των μετρήσεων και η αβεβαιότητα, συμπεριλαμβανομένου και της ανατύπωσης και των ατελειών του συνόλου δεδομένων (dataset), του καθενός δείγματος δεδομένων, καθιστούν δύσκολη την απόκτηση κάποιου συνεπούς αποτελέσματος από όλα τα διαφορετικά πειράματα. Αυτό ισχύει, γιατί τα αποτελέσματα επηρεάζονται ιδιαίτερα από την κατάσταση του πειράματος, όπως ακόμα και από τις περιβαλλοντικές συνθήκες στις οποίες εκτελείται. Επίσης, η έμφυτη διακύμανση των περιπτώσεων από τις κλινικές/ παθολογικές διαγνώσεις επηρεάζουν τον εντοπισμό χαρακτηριστικών (γονιδίων). Για παράδειγμα, δεδομένα σε μια καρκινική ασθένεια δεν καθορίζονται από τις γενετικές μεταβολές στον όγκο, ή τη θεραπεία που μπορεί να εφαρμόζεται, αλλά και από διάφορα χαρακτηριστικά μεμονωμένων περιστάσεων, όπως είναι η ηλικία του ασθενούς, οι γενικές συνθήκες υγιεινής και ο μεταβολισμός λόγω της χρήσης φαρμάκων. Ο

συγκεκριμένος τύπος έμφυτης διακύμανσης επηρεάζει τόσο τις παραμετρικές όσο και τις μη παραμετρικές προσεγγίσεις.

Παράδειγμα μη- παραμετρικής μεθόδου είναι ο **αριθμός κατωφλίου λαθεμένης ταξινόμησης** – threshold number of misclassification (TnoM). Κατά τη μέθοδο αυτή, επιλέγουμε ως κατώφλι (threshold) μια τιμή της έκφρασης δεδομένων και υπολογίζουμε συνολικό αριθμό των λανθασμένων ταξινομήσεων (TnoM score). Από τον αριθμό αυτό, ένα γονίδιο θεωρείται τόσο πιο σημαντικό όσο αφορά την πληροφορία, όσο πιο μικρός είναι ο αριθμός λανθασμένων ταξινομήσεων.

Σύμφωνα με τη θεωρία υπολογισμού, είναι γνωστό ότι ο υπολογισμός κάποιων ιδιοτήτων ενός συνόλου αντικειμένων είναι η διαδικασία ανάθεσης αριθμών ή άλλων συμβόλων στα αντικείμενα με τέτοιο τρόπο ώστε να είναι δυνατός ο υπολογισμός της σχέσης μεταξύ των αντικειμένων που υπολογίζονται (Allen1979). Ένας συγκεκριμένος τρόπος ανάθεσης αριθμών ή άλλων συμβόλων σύμφωνα με της ιδιότητες των αντικειμένων ονομάζεται **κλίμακα υπολογισμού (scale of measurement)**. Η βασική ιδέα της θεωρίας υπολογισμού, λοιπόν, είναι ότι μια ποσοτική κλίμακα είναι η συσχέτιση (mapping) μεταξύ κάποιων αντικειμένων και ενός συνόλου σχετιζόμενων αριθμητικών τιμών. Η συσχέτιση αυτή βέβαια, δεν είναι αυθαίρετη, αλλά ικανοποιεί κάποιες απαιτήσεις. Γι' αυτό το λόγο θα πρέπει να είναι κανείς πολύ προσεκτικός στην υλοποίηση κάποιων μεθόδων υπολογισμού και να αναγνωρίζει τις δυνατότητες και τα μειονεκτήματά τους. Πιο συγκεκριμένα, στο πεδίο της γονιδιακής απεικόνισης διαφορετικοί αλγόριθμοι και τεχνικές εκτίμησης των αποτελεσμάτων μπορούν να έχουν ως αποτέλεσμα την δημιουργία διαφορετικών λιστών κατάταξης γονιδίων και την εξαγωγή διαφορετικών συνόλων γονιδίων που μπορούν να χρησιμοποιηθούν για την πρόβλεψη αποτελέσματος. Έτσι, είναι ζητούμενο η θέσπιση κριτηρίων στην ανάλυση, έτσι ώστε να μπορούμε να ελέγχουμε τη διαδικασία της κατάταξης και εξαγωγής γονιδίων. Υιοθετούμε, λοιπόν, ως κύρια κριτήρια την «διακριτότητα» κα την «αντιπροσωπευτικότητα». Με βάση αυτά, η υλοποίησή μας έχει ως βασικό στόχο τα υποψήφια για εξαγωγή γονίδια που κατατάσσονται πιο ψηλά στις λίστες που διαμορφώνονται τελικά, να είναι όσο το δυνατόν πιο «διακριτά», λαμβάνοντας υπόψη τις διαφορετικές κλάσεις των αποτελεσμάτων, έτσι ώστε οι

κατανομές των τιμών των γονιδίων να είναι εύκολα διαχωρισμένες, που σημαίνει ότι τα επίπεδα της γονιδιακής έκφρασης είναι και αριθμητικά διαχωρισμένα.

Ζ. Αλγόριθμοι Κατάταξης και Επιλογής Γονιδίων

Με βάση τα παραπάνω κριτήρια και τα διαφορετικά κλινικά αποτελέσματα, μπορούμε να παρουσιάσουμε τους βασικούς αλγορίθμους για την κατάταξη και επιλογή των γονιδίων που χρησιμοποιήσαμε κατά την υλοποίηση των μεθόδων υπολογισμού.

1) Μέτρο Συσχέτισης Διαχωρισμού/ κριτήριο διαχωρισμού το Fisher (Separation Correlation Metric SCM/ Fisher's discrimination criterion FDC)

Η FDC μέθοδος είναι μια από τις πιο δημοφιλής παραμετρικές μεθόδους εντοπισμού των ιδιοτήτων δεδομένων και των προβολών τους, που είναι πιο εύκολα διαχωρίσιμες ανάμεσα στις κλάσεις. Είναι μια μέθοδος η οποία χρησιμοποιείται κυρίως για διαφορετικά εκφραζόμενες ασθένειες ή για ασθένειες με διαφορετικά κλινικά αποτελέσματα. Παρόλα αυτά είναι επίσης γνωστό ότι η FDC μέθοδος δεν μπορεί να αποτελέσει ένα απόλυτο κριτήριο ανάπτυξης ακριβών ταξινομήσεων. Έτσι, θα πρέπει να χρησιμοποιείται συνδυαζόμενη με άλλες αναλύσεις συσχετισμού σε πρακτικές εφαρμογές, ώστε να ελαχιστοποιηθούν κάποιες από τις παράμετρες επιπτώσεις.

Περιγραφή αλγορίθμου

Έστω ω_1 και ω_2 οι ετικέτες των δύο διαφορετικών κλάσεων των δειγμάτων (των θανατηφόρων και των μη θανατηφόρων περιπτώσεων). Η FDC μέθοδος έχει ως στόχο τη μεγιστοποίηση του κριτηρίου:

$$J(\underline{W}) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} = \frac{\underline{W}^T S_B \underline{W}}{\underline{W}^T S_W \underline{W}}$$

με μ_i, i=1,2 να είναι η μέση τιμή της προβολής των δειγμάτων των δεδομένων της κλάσης ω_i στην κατεύθυνση έκφρασης του διανύσματος $\underline{W}$, αντίστοιχα. Έχουμε, $\mu_i = (1/n_i)\sum_{\underline{X}\in\omega_i} \underline{W}^T \underline{X}$ με $\underline{X}=[x_1,x_2, \ldots]$, να αναπαριστά το διάνυσμα της γονιδιακής έκφρασης, που αποτελείται από τις τιμές έκφρασης του κάθε γονιδίου σε όλα τα

$$\sigma_i^2 = (1/n_i)\sum_{\underline{X}\in\omega_i} (\underline{W}^T \underline{X} - \mu_i)^2,\ i = 1,2$$

κλινικά δείγματα. Ο n_i είναι ο αριθμός των δειγμάτων δεδομένων στην κλάση ω_i. Ο είναι η απόκλιση των προβαλλόμενων δειγμάτων της κλάσης ω_i στην κατεύθυνση έκφρασης του διανύσματος $\underline{W}$, αντίστοιχα. Το $\underline{W}=[w_1, w_2, \ldots, w_n]$ είναι ένα διάνυσμα στο χώρο $\underline{X}$ που λειτουργεί ως συνάρτηση μετασχηματισμού στην οποία τα δείγματα X προβάλλονται στο μονοδιάστατο χώρο $y = \underline{W}^T \underline{X}$. Ο $S_B = (m_1-m_2)(m_1-m_2)^T$ ονομάζεται καρτεσιανός πίνακας ενδιάμεσα στις κλάσεις, όπου m_i, i=1,2 είναι το διάνυσμα της μέσης τιμής των δειγμάτων δεδομένων της κλάσης ω_i στο αρχικό διάνυσμα χώρου $\underline{X}$. Ο $S_W = S_1+S_2$ ονομάζεται καρτεσιανός πίνακας μεταξύ των κλάσεων, όπου S_i, i=1,2 είναι ο πίνακας απόκλισης των δειγμάτων δεδομένων της κλάσης ω_i στο αρχικό διάνυσμα χώρου $\underline{X}$.

Όταν περιορίζουμε το διάνυσμα $\underline{W}$ στην μορφή [1 0 ...], [0 1 ...], τότε το πρόβλημα μεταφέρεται στις ευκλείδιες συντεταγμένες, ο πίνακας FDC αναπαριστά το μέτρο ενός γονιδίου σύμφωνα με τις παραμέτρους της μέσης τιμής και της τυπικής απόκλισης, σε συνάρτηση πάντα με τις τιμές των αρχικών κλάσεων. Έτσι, αν στο γονίδιο g_j αντιστοιχεί το μέτρο FDC (k), τότε αυτό ισούται με:

$$\mathrm{FDC}_k = \frac{(\mu_{k1}-\mu_{k2})^2}{\frac{\left(\sum_{j=1}^{n_1}\left|x_{kj}^{(1)}-\mu_{k1}\right|\right)}{n_1} + \frac{\left(\sum_{j=1}^{n_2}\left|x_{kj}^{(2)}-\mu_{k2}\right|\right)}{n_2}}$$

2) Cross Projection Index (CPI)

Πρόκειται για εφαρμογή της μεθόδου «projection pursuit» σε γονιδιακές εκφράσεις. Η μέθοδος projection pursuit αφορά τη συσχέτιση κάποιων τιμών σε κάθε μια προβολή μικρών διαστάσεων που αναδεικνύει τη δομή του συνόλου δεδομένων. Όταν δημιουργηθεί ένα συγκεκριμένο σύνολο προβολών, οι υπάρχουσες δομές (π.χ. ένα σύνολο γονιδίων, ομάδες δεδομένων κ.ά.) διαφορετικών προτύπων μπορούν να εξαχθούν και να μελετηθούν ξεχωριστά ανάλογα με τις τιμές ενός ευρετηρίου (index). Αυτή η τιμή του ευρετηρίου προκύπτει από μια συνάρτηση εκτίμησης που σχετίζεται με την προβολή.

Συνήθως, διεξάγεται έρευνα για την ανακάλυψη των συσχετιζόμενων προβολών με τη μεγιστοποίηση του ευρετηρίου στο χώρο προβολής.

Στην περίπτωσή μας, έχουμε προβολή των τιμών της γονιδιακής έκφρασης της κάθε μιας μεμονωμένης περίπτωσης δείγματος στην αντίστοιχη συνάρτηση πιθανότητας, λαμβάνοντας υπόψη και την κατανομή της συνολικής κλάσης. Ονομάζουμε την προβολή CP. Οι ποσοτικοί υπολογισμοί των προβολών σε όλες τις περιπτώσεις διαμορφώνουν το «ευρετήριο» («index»). Καθώς λαμβάνουν χώρα οι προβολές μεταξύ των διαφορετικών ζευγαριών των γονιδιακών προτύπων, τα γονίδια, ανάλογα με το βαθμό που μπορούν να αποτυπώνουν τα χαρακτηριστικά των γονιδιακών εκφράσεων ανάμεσα στις διαφορετικές κλάσεις κατατάσσονται σε έναν πίνακα- ευρετήριο.

Η διαδικασία υπολογισμού του Cross Projection Index (CPI) έχει ως εξής:

1. θεωρώντας δύο σύνολα δεδομένων D_1 και D_2, όπου

$D_1 = \{x_1^{(1)}, x_2^{(1)}, \ldots, x_{n1}^{(1)}\}$, και $D_2 = \{x_1^{(2)}, x_2^{(2)}, \ldots, x_{n2}^{(2)}\}$

τα σημεία δεδομένων $x_k^{(1)} \in D_1, k = 1, 2, \ldots, n_1$ αποτελούν τα δείγματα που ανήκουν στην κλάση ω_1

τα σημεία δεδομένων $x_k^{(2)} \in D_2, k = 1, 2, \ldots, n_2$ αποτελούν τα δείγματα που ανήκουν στην κλάση ω_2 . Κάθε σημείο $x_k^{(c)}, c \in \{1, 2\}$ είναι ένα διάνυσμα $x_k^{(c)} = [x_{k1}^{(c)}, x_{k2}^{(c)}, \ldots, x_{km}^{(c)}]$, όπου m είναι ο αριθμός των χαρακτηριστικών- γονιδίων σε κάθε περίπτωση δειγμάτων.

2. Εφόσον τα στοιχεία του συνόλου δεδομένων ακολουθούν την γκαουσιανή κατανομή έχουμε,

$$p^{(1)}(x) = p(x|x \in \omega_1) = p\left(x_{kj}^{(1)}\right) = \frac{1}{\sqrt{2\pi}\sigma_j^{(1)}} e^{-\frac{1}{2}\left(\frac{x_{kj}^{(1)} - \mu_j^{(1)}}{\sigma_j^{(1)}}\right)^2}$$

$$p^{(2)}(x) = p(x|x \in \omega_2) = p\left(x_{kj}^{(2)}\right) = \frac{1}{\sqrt{2\pi}\sigma_j^{(2)}} e^{-\frac{1}{2}\left(\frac{x_{kj}^{(2)} - \mu_j^{(2)}}{\sigma_j^{(2)}}\right)^2}$$

όπου, $\mu_j^{(c)} = (1/n_c)\sum_{k=1}^{n_c} x_{kj}^{(c)}$,

και $\sigma_j^{(c)} = (1/n_c)\sum_{k=1}^{n_c}(x_{kj}^{(c)} - \mu_j^{(c)})^2,\ c \in \{1,2\}$

με n_c τον αριθμών των σημείων δεδομένων στην κλάση ω_c

3. υπολογίζουμε το CP του $x_{kj}^{(1)}$ με το j στοιχείο του σημείου δεδομένου x_k της κλάσης ω_1 ως προς την κατανομή της κλάσης ω_2 ορίζεται ως:

$$p^{(2)}\left(x_{kj}^{(1)}\right) = \frac{1}{\sqrt{2\pi}\sigma_j^{(2)}} e^{-\frac{1}{2}\left(\frac{x_{kj}^{(1)} - \mu_j^{(2)}}{\sigma_j^{(2)}}\right)^2}$$

ομοίως το CP του $x_{kj}^{(2)}$ με το j στοιχείο του σημείου δεδομένου x_k της κλάσης ω_2 ως προς την κατανομή της κλάσης ω_1 ορίζεται ως:

$$p^{(1)}\left(x_{kj}^{(2)}\right) = \frac{1}{\sqrt{2\pi}\sigma_j^{(1)}} e^{-\frac{1}{2}\left(\frac{x_{kj}^{(2)} - \mu_j^{(1)}}{\sigma_j^{(1)}}\right)^2}$$

4. παίρνουμε τους λογαριθμικούς μετασχηματισμούς των παραπάνω εκφράσεων και έχουμε:

$$\begin{aligned}
P^{11} &= \log\left(p^{(1)}\left(x_{kj}^{(1)}\right)\right) \\
&= -\frac{1}{2}\left(\frac{x_{kj}^{(1)} - \mu_j^{(1)}}{\sigma_j^{(1)}}\right)^2 + \ln\frac{1}{\sqrt{2\pi}\sigma_j^{(1)}} \\
P^{12} &= \log\left(p^{(1)}\left(x_{kj}^{(2)}\right)\right) \\
&= -\frac{1}{2}\left(\frac{x_{kj}^{(2)} - \mu_j^{(1)}}{\sigma_j^{(1)}}\right)^2 + \ln\frac{1}{\sqrt{2\pi}\sigma_j^{(1)}} \\
P^{21} &= \log\left(p^{(2)}\left(x_{kj}^{(1)}\right)\right) \\
&= -\frac{1}{2}\left(\frac{x_{kj}^{(1)} - \mu_j^{(2)}}{\sigma_j^{(2)}}\right)^2 + \ln\frac{1}{\sqrt{2\pi}\sigma_j^{(2)}} \\
P^{22} &= \log\left(p^{(2)}\left(x_{kj}^{(2)}\right)\right) \\
&= -\frac{1}{2}\left(\frac{x_{kj}^{(2)} - \mu_j^{(2)}}{\sigma_j^{(2)}}\right)^2 + \ln\frac{1}{\sqrt{2\pi}\sigma_j^{(2)}}.
\end{aligned}$$

Οι μετρήσεις αυτές αποτελούν την πιθανότητα των διανυσμάτων των δειγμάτων σε σχέση με τις διαφορετικές κλάσεις των αποτελεσμάτων που εξετάζουμε.

5. τελικά το CPI υπολογίζεται ως:

εάν το $Q_j^{(1)}$ είναι ο αριθμός των σημείων δεδομένων $x_{kj}^{(1)}$ του γονιδίου j, έτσι ώστε $[P^{11} > P^{21}]$ και

$Q_j^{(2)}$ είναι ο αριθμός των σημείων δεδομένων $x_{kj}^{(2)}$ του γονιδίου j, έτσι ώστε $[P^{22} > P^{12}]$, τότε το CPI γονιδίου g_j ορίζεται ως:

$$\mathrm{CPI}(j) = Q_j^{(1)} + Q_j^{(2)}$$

Τα αποτελέσματα του CPI είναι διακριτοί αριθμοί. Είναι μια κατάταξη της μέγιστης πιθανότητας του σημείου δεδομένων x_k σε σχέση με την κλάση ω_i.

3) Discrete Partition Index (DPI)

Ενώ οι μέθοδοι FDC και CPI είναι παραμετρικές, η DPI μέθοδος είναι μη παραμετρική και διακριτή. Η DPI μέθοδος υπολογίζει τον ελάχιστο ρυθμό εσφαλμένης ταξινόμησης στο σύνολο δεδομένων και βασίζεται στην αρχή TnoM. Η αρχή TnoM προσπαθεί να βρει ένα σημείο δεδομένων V_k στο σύνολο της γονιδιακής έκφρασης, τέτοιο ώστε ο συνολικός αριθμός δειγμάτων που ταξινομήθηκαν λάθος θα μπορούσε να ελαχιστοποιηθεί, αν το σημείο V_k λαμβάνεται ως όριο διαχωρισμού που διαιρεί τα δεδομένα της γονιδιακής έκφρασης σε δύο διακριτές κλάσεις. Ο ελάχιστος αριθμός δειγμάτων που ταξινομούνται λάθος σε ένα βέλτιστο σημείο V_k ονομάζεται **Discrete Partition Index.** Ο αλγόριθμος υπολογισμού του DPI έχει ως εξής:

για κάθε γονίδιο g_j, οι τιμές έκφρασής του σε όλα τα δείγματα που βρίσκονται στη σειρά j του συνόλου δεδομένων

1. Κατατάσσουμε τις τιμές έκφρασης των δεδομένων της σειράς j σε αύξουσα σειρά
2. Αρχή βρόχου

   k=1;

V[k]: το k –οστό δείγμα στην καταταγμένη λίστα (από τα αριστερά προς τα δεξιά);

A[k]: ο αριθμός των δειγμάτων της κλάσης ω_1 που βρίσκονται αριστερά του V[k] + ο αριθμός των δειγμάτων της κλάσης ω_2 που βρίσκονται δεξιά του V[k]

B[k]: ο αριθμός των δειγμάτων της κλάσης ω_2 που βρίσκονται αριστερά του V[k] + ο αριθμός των δειγμάτων της κλάσης ω_1 που βρίσκονται δεξιά του V[k]

k++;

Τέλος βρόχου;

Μ. DPI(j) ισούται με τη μικρότερη τιμή ανάμεσα στα A[k] B[k];
Τέλος της διαδικασίας;

- Οι τιμές στα A[k] και B[k] είναι οι πιθανοί αριθμοί των εσφαλμένων ταξινομήσεων, όταν λαμβάνουμε τον αριθμό V[k] ως τιμή ορίου.
- Η τιμή DPI[j] είναι ο ελάχιστος αριθμός εσφαλμένων ταξινομήσεων για το γονίδιο g_j
- Τα γονίδια παίρνουν την τιμή των ελάχιστων δυνατόν εσφαλμένων ταξινομήσεων, όταν ο αλγόριθμος τελειώσει.
- Τα γονίδια δεν κατατάσσονται με μοναδικό τρόπο στο DPI, επειδή πολλαπλά γονίδια είναι δυνατόν να έχουν την ίδια τιμή DPI και άρα παίρνουν την ίδια τιμή και έχουν την ίδια κατάταξη.

Η. Σύντηξη Αλγορίθμων Γενετικής Έκφρασης για την Επιλογή Δεικτών Πρόβλεψης Κακοηθών Όγκων

Η θεωρία υπολογισμού αποδεικνύει ότι απαιτούνται πολύ ισχυρές υποθέσεις για στατιστικά δεδομένα, ώστε να έχουμε ουσιαστική πληροφόρηση για τα πραγματικά στοιχεία των αντικειμένων που μελετάμε. Σε πολλές, όμως, περιπτώσεις, είναι επιθυμητή μια ανάλυση, η οποία αποδίδει σταθερά αποτελέσματα. Στο συγκεκριμένο πρόβλημα, η θεωρία υπολογισμού δεν διαθέτει μια ολοκληρωμένη λύση. Πιο συγκεκριμένα, δε λαμβάνει υπόψη το τυχαίο σφάλμα υπολογισμού και έτσι, εάν τέτοια σφάλματα αποτελούν μια σημαντική πτυχή της διαδικασίας υπολογισμού, τότε απαιτούνται επιπλέον μέθοδοι ανάλυσης.

Από την υλοποίηση των τριών παραπάνω αλγορίθμων έχει γίνει γνωστό ότι δεν υπάρχει συσχέτιση του ενός με τον άλλο και ότι τα γονίδια από τις πρώτες θέσεις της κατάταξης από τις τρεις λίστες που προκύπτουν είναι δυνατόν να περιγράψουν ένα διακριτό πρότυπο. Έτσι, είναι δυνατή η επιλογή ενός συνόλου με καλύτερη απόδοση (Zhu.2004)

Με βάση, λοιπόν, όλα τα παραπάνω μπορούμε να αντιληφθούμε ότι θα μπορούσαμε να έχουμε καλύτερο αποτέλεσμα συνδυάζοντας έναν αριθμό διαφορετικών μεθόδων υπολογισμού. Παρόλα αυτά, η ποιότητα των αποτελεσμάτων εξαρτάται από τον τρόπο με τον οποίο οι διαφορετικές μέθοδοι υπολογισμού ενσωματώνονται, όπως επίσης, από το σύνολο των μεθόδων υπολογισμού που διαμορφώνει έναν καλό συνδυασμό τους για ένα συγκεκριμένο σύνολο δεδομένων. Ακόμα και με αυτόν τον τρόπο, είναι δυνατόν να υπάρχουν διάφορες δυσκολίες στην υλοποίηση κάποιων συνδυασμών. Μερικές από αυτές είναι η στατιστική φύση πολλών μεθόδων σε αντίθεση με κάποιες μη παραμετρικές, ή κάποιες βασίζονται σε κάποια μοντέλα, ενώ κάποιες άλλες όχι.

Το βέβαιο, όμως, είναι ότι δεν μπορούμε να επιτύχουμε έναν αποδοτικό συνδυασμό μεθόδων μέσω απλών εφαρμογών, ούτε επίσης, χρησιμοποιώντας την ιδιότητα της προσθετικότητας ανάμεσα στις διάφορες μεθόδους. Έτσι,

απαιτείται μια πιο επιδέξια προσέγγιση, η οποία θα λαμβάνει υπόψη τους διαφορετικούς τύπους υπολογισμού και κλιμάκων.

Η μέθοδος που χρησιμοποιούμε είναι η αλγοριθμική σύντηξη (gene fusion) που βασίζεται στην ενοποίηση ενός συνόλου παραμετρικών και μη παραμετρικών μεθόδων υπολογισμού σε διαφορετικά μέτρα κατάταξης. Η μέθοδος έχει ως σκοπό την εξαγωγή ενός τελικού αριθμού γονιδίων που θα μπορεί να «προβλέψει» τα κλινικά αποτελέσματα καρκινικών παθήσεων.

Η μέθοδος της αλγοριθμικής σύντηξης που χρησιμοποιούμε διαφέρει από τη συνηθισμένη χρήση της μεθόδου, αφού στις συντριπτικά περισσότερες περιπτώσεις αυτή εφαρμόζεται στο στάδιο της ταξινόμησης. Πιο συγκεκριμένα, η σύντηξη αφορά τον συνδυασμό διαφόρων ταξινομητών, που έχει ως στόχο τη διατήρηση της ακριβούς απεικόνισης χαρακτηριστικών, διατηρώντας την απλότητα και την ποικιλότητα. Η βασική διάκριση ανάμεσα στις μεθόδους σύντηξης αυτού του τύπου έχουν να κάνουν με τα δεδομένα εισόδου που χρησιμοποιούμε. Οι αρχικές μέθοδοι σύντηξης χρησιμοποιούσαν τις διακριτές τιμές ετικέτας (crisp labels) του κάθε ταξινομητή. Βασίζονταν σε σχήματα, όπως είναι πλειοψηφικής επιλογής (majority voting), η οποία εξάγει τα στατιστικά δεδομένα της υστερογενούς πιθανότητας κλάσης (posterior class probability statistics), συγκεντρώνοντας τις πραγματικές και ετικέτες για κάθε κλάση. Κάποιες άλλες μέθοδοι σύντηξης χρησιμοποιούν τις συνεχείς τιμές (soft labels) των ταξινομητών. Αυτές οι μέθοδοι είναι θεωρητικά πιο αποτελεσματικές και συνδυάζουν διάφορες συναρτήσεις από πολύ απλές (ελάχιστο, μέγιστο, γινόμενο, μέσος όρος) μέχρι πιο εξελιγμένες (γινόμενο πιθανότητας, γραμμικοί συνδυαστές, γραμμικές, τετραγωνικές συναρτήσεις, συναρτήσεις Fisher).

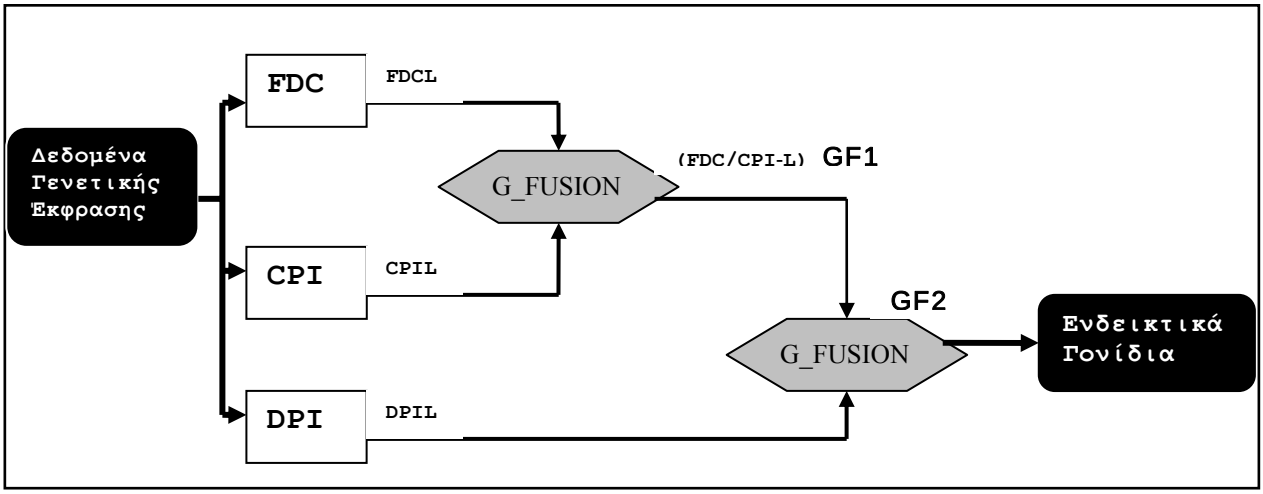
Η εφαρμογή της σύντηξης στην περίπτωσή μας διαφοροποιείται από τις κλασσικές εφαρμογές της μεθόδου, αφού η υλοποίησή μας προβλέπει τη σύντηξη αλγορίθμων για την επιλογή χαρακτηριστικών δεικτών από βάσεις γονιδιακών δεδομένων **πριν** από την ταξινόμηση. Ο ρόλος της ταξινόμησης έπειτα είναι η επιβεβαίωση ή όχι της επιλογής των γονιδίων –δεικτών του προηγούμενου σταδίου.

Η αλγοριθμική σύντηξη (gene fusion) εφαρμόζεται στις τρεις βασικές μεθόδους υπολογισμού που περιγράφηκαν παραπάνω. Για την υλοποίηση της μεθόδου χρησιμοποιούμε τον αλγόριθμο G_FUSION. Βασικό χαρακτηριστικό του είναι η διεκπεραίωση του προβλήματος του συνδυασμού πολλαπλών μεθόδων υπολογισμού – κατάταξης γονιδίων οι οποίες βρίσκονται σε διαφορετικές κλίμακες και διαστήματα αριθμητικών τιμών. Για παράδειγμα, στη μέθοδο FDC κάθε γονίδιο έχει διαφορετική τιμή εξαιτίας της απόδοσης συνεχών τιμών. Έτσι, τα γονίδια κατατάσσονται με μοναδικό τρόπο (ένα γονίδιο ανά επίπεδο γονιδιακής έκφρασης) . Αντίθετα, ένας αριθμός διαφορετικών γονιδίων στις μεθόδους CPI και DPI μπορεί να έχει την ίδια τιμή, άρα κατατάσσονται στο ίδιο επίπεδο- βαθμίδα γονιδιακής έκφρασης. Το παραπάνω χαρακτηριστικό, όπως και άλλα, είναι πολύ σημαντικό να ληφθεί υπόψη από τον G_FUSION αλγόριθμο που εφαρμόζουμε.

Μέθοδος	ιδιότητες					
	παραμετρική	**μη-παραμετρική**	**Συνεχής**	**Διακριτή**	**Βασισμένη σε μοντέλο**	**Όχι βασισμένη σε μοντέλο**
FDC	X		X			X
CPI	X			X	X	
DPI		X		X		X

Πίνακας 1: χαρακτηριστικές ιδιότητες των τριών μεθόδων υπολογισμού

Ο G_FUSION αλγόριθμος υλοποιεί μια παράλληλη «διασταύρωση» τεχνική συνένωσης στα αποτελέσματα των δυο παραμετρικών μεθόδων, των FDC και CPI, και μια μη παραμετρικής, της DPI. Ο τρόπος με τον οποίο υλοποιείται η αλγοριθμική σύντηξη φαίνεται στο διάγραμμα 2



διάγραμμα 2: υλοποίηση αλγοριθμικής σύντηξης τριών μεθόδων υπολογισμού

Συμβολίζουμε τις λίστες με τις κατατάξεις των γονιδίων που προκύπτουν από τις μεθόδους FDC, CPI και DPI, ως FDCL, CPIL και DPIL, αντίστοιχα. Όπως τονίσαμε και παραπάνω στην FDCL τα γονίδια κατατάσσονται με μοναδικό τρόπο, ενώ στις CPIL και DPIL μπορεί σε κάθε κόμβο της λίστας να υπάρχουν περισσότερα του ενός γονίδια με την ίδια τιμή CPI ή DPI. Έτσι, ο αριθμός των κόμβων στην FDCL είναι ίσος με τον αριθμό των συνολικών γονιδίων, ενώ στις CPIL και DPIL ο αριθμός των κόμβων ισούται με τα διαφορετικά επίπεδα τιμών των μετρήσεων.

Επίσης, ορίζουμε σε κάθε κόμβο κάθε λίστας μια αριθμητική τιμή μετρητής(k) για κάθε κόμβο k των λιστών για να καταγράφουμε τον αριθμό των γονιδίων που συμπίπτουν στη θέση κατάταξης.

$$\text{Μετρητής}(k) = \sum_{i=0}^{k} \text{αριθμός γονιδίων στον κόμβο } i \text{ του FDCL, CPIL, DPIL}$$

Ο G_FUSION αλγόριθμος κατά τη διάρκεια της σύντηξης εξάγει τα γονίδια που εκτιμώνται ως καλύτερα σε όλες τις μεθόδους υπολογισμού που συμμετέχουν

στην σύντηξη. Τα γονίδια στην τελική λίστα που προκύπτει συνεχίζουν να έχουν τις θέσεις κατάταξης που είχαν και στις αρχικές λίστες.

Η διαδικασία της σύντηξης, όπως φαίνεται και στο διάγραμμα 2 γίνεται σε δύο βήματα:

a. Γίνεται η σύντηξη των FDCL και CPIL, για να προκύψει η λίστα FDC/ CPI-L
b. Γίνεται η σύντηξη των FDC/ CPI-L , DPI-L, για να προκύψει η λίστα με τα τελικά γονίδια ως τελικό μας αποτέλεσμα

Η διαδικασία του G_FUSION αλγορίθμου έχει ως εξής:

Διαδικασία G_FUSION: Διασταύρωση Λιστών (έστω ότι παίρνουμε ως παράδειγμα τις λίστες FDCL-CPIL)

Είσοδος:

1. Διαμορφώνουμε τη λίστα FDCL με τα γονίδια να κατατάσσονται ανάλογα με τις τιμές των FDC σε φθίνουσα σειρά.
2. Διαμορφώνουμε τη λίστα CPIL με τα γονίδια να κατατάσσονται ανάλογα με τις τιμές των CPI σε φθίνουσα σειρά. (στην περίπτωση του DPIL η κατάταξη είναι σε αύξουσα σειρά.

Έξοδος:

Μια νέα λίστα FDC/CPI-L, όπου τα γονίδια κατατάσσονται ανάλογα με τις τιμές FDC μετρά τη διαδικασία διασταύρωσης στις λίστες FDCL-CPIL

Βήματα Υπολογισμού:

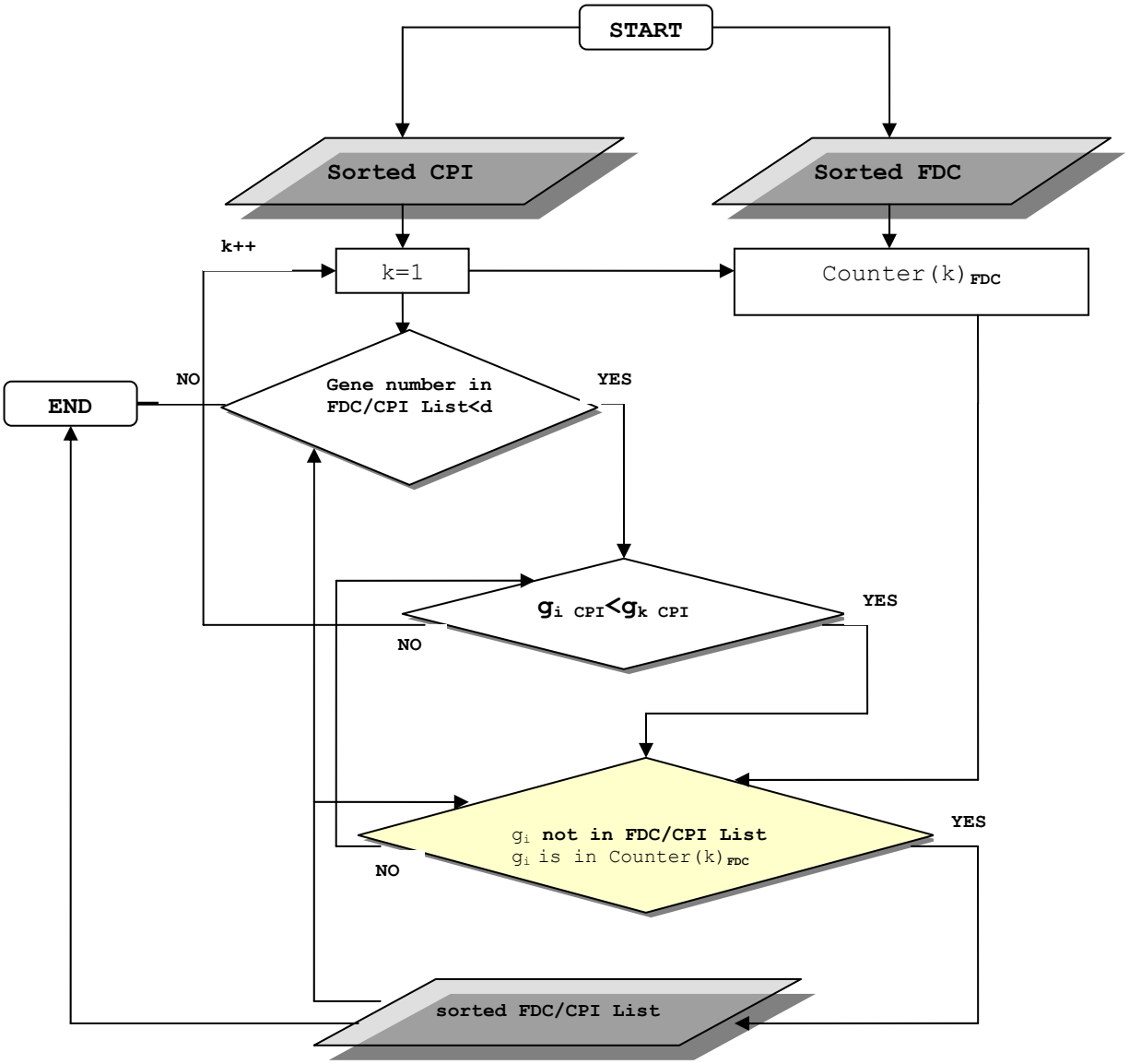
1. Έστω d η προκαθορισμένη σταθερά, που ορίζει τον αριθμό των γονιδίων που πρόκειται να επιλέξουμε
2. Αρχικοποιούμε τη λίστα FDC/CPI-L να είναι κενή
3. Έστω k να είναι η σταθερά που αρχικοποιείται στο πρώτο επίπεδο τιμών του CPIL
4. (***κριτήριο επιλογής***) Όσο ο αριθμός των γονιδίων στη λίστα FDC/CPI-L είναι μικρότερος του d:
   α. Για κάθε γονίδιο g_i,τέτοια ώστε το g_i να έχει τιμή CPI μικρότερη ή ίση με k στη CPIL και να μην βρίσκεται ήδη στη λίστα FDC/CPI-L.
   Εάν το γονίδιο g_i είναι παρόν στις τιμές του Μετρητή(k) των γονιδίων του FDCL, **τοποθετούμε το γονίδιο g_i** στη λίστα FDC/CPI-L
   β. Αυξάνουμε την τιμή του k στο επόμενο επίπεδο της λίστας CPIL

5. Κατατάσσουμε τα γονίδια στη λίστα FDC/CPI-L, σύμφωνα με τις FDC τιμές

- Ο αριθμός των γονιδίων που τελικά επιλέγονται στην λίστα FDC/ CPI-L ενδέχεται να είναι μεγαλύτερος από τον αριθμό d που εμείς ορίζουμε στο $1^{ο}$ βήμα του αλγορίθμου, εξαιτίας της πολλαπλότητας των γονιδίων στους κόμβους της λίστας CPIL. Μόλις φτάσουμε στη σειρά k της CPIL, επιλέγονται όλα τα γονίδια της ίδιας θέσης κατάταξης στη λίστα που πληρούν το κριτήριο του βήματος 4(α).

- Η ίδια διαδικασία ακολουθείται και για τη σύντηξη των λιστών FDC/ CPI-L , DPIL. Για να διασφαλίσουμε ότι η δεύτερη εφαρμογή του αλγορίθμου μας δίνει επαρκή αριθμό γονιδίων (>=d), θα πρέπει στην πρώτη εφαρμογή του αλγορίθμου να ορίσουμε ένα σχετικά μεγάλο αριθμό d.

Το διάγραμμα ροής του αλγορίθμου είναι:



διάγραμμα 3: Το διάγραμμα ροής του αλγορίθμου G_FUSION

Θ. Εποπτική Ταξινόμηση

Για να εκτιμήσουμε την αποτελεσματικότητα της μεθόδου της αλγοριθμικής σύντηξης, όπως φαίνεται και από την διάγραμμα1 πραγματοποιούμε εποπτική ταξινόμηση στα γονίδια κάθε πειραματικής εφαρμογής (η οποία όπως φαίνεται και από την περιγραφή του G_FUSION αλγορίθμου καθορίζεται από όλες τις επιλογές και τους συνδυασμούς στις τιμές του d σε κάθε εφαρμογή του αλγορίθμου).

Για την εποπτική ταξινόμηση χρησιμοποιούμε τη μέθοδο **support vector machines (SVM)**. Θεωρητικά, για να δημιουργήσουμε έναν δυαδικό ταξινομητή δεν έχουμε παρά να οικοδομήσουμε ένα υπερεπίπεδο (hyper plane) που θα διαχωρίζει τα δεδομένα σε δύο κλάσεις (θετικά και αρνητικά παραδείγματα) στο χώρο. Δυστυχώς, όμως στην πραγματικότητα τα περισσότερα ρεαλιστικά προβλήματα περιλαμβάνουν μη διαχωρίσιμα δεδομένα για τα οποία δεν μπορεί να υπάρξει κάποιο υπερεπίπεδο που θα διαχωρίζει τα θετικά από τα αρνητικά παραδείγματα. Η χρήση των support vector machines μας βοηθά να υπερκεράσουμε όλες τις δυσκολίες και τα μειονεκτήματα των τυπικών ταξινομητών. Πρόκειται για μη γραμμικούς αλγόριθμους ταξινόμησης, οι οποίοι βασίζονται σε μεθόδους πυρήνα (kernel methods). Σε αντίθεση με τις μεθόδους γραμμικής ταξινόμησης, οι μέθοδοι πυρήνα προχωρούν σε απεικόνιση των δεδομένων σε έναν χώρο περισσοτέρων διαστάσεων και η οικοδόμηση υπερεπιπέδου διαχωρισμού εκεί. Ο μεγάλων διαστάσεων χώρος αυτός ονομάζεται **χώρος δεδομένων**, σε αντιπαράθεση με τον χώρο εισόδων, όπου βρίσκονται τα παραδείγματα προς εκπαίδευση. Με την κατάλληλη επιλογή ενός χώρου δεδομένων με επαρκείς διαστάσεις, κάθε σύνολο εκπαιδευόμενων παραδειγμάτων μπορεί να εξελιχθεί σε διαχωρίσιμο. Στο χώρο δεδομένων μπορούμε να υπολογίσουμε τα εσωτερικά γινόμενα των διανυσμάτων επαρκώς, χωρίς να απαιτείται να υπολογίσουμε τη μη γραμμική απεικόνιση. Παρόλα αυτά, «μεταφράζοντας» το εκπαιδευόμενο σύνολο σε έναν χώρο υψηλών διαστάσεων, είναι σίγουρο ότι θα έχουμε μεγάλο κόστος υπολογισμού. Επιπλέον, στην προσπάθεια διαχωρισμού των δεδομένων με τον παραπάνω τρόπο, υπάρχει ο κίνδυνος να επιτύχουμε επουσιώδεις λύσεις, όπου η απόδοση των εκπαιδευόμενων παραδειγμάτων αυξάνει, ενώ η απόδοση των αθέατων

δεδομένων χειροτερεύει (overfitting). Η μέθοδος των support vector machines, όμως, δίνει λύσεις και στο συγκεκριμένο πρόβλημα (του overfitting), καθώς επιλέγοντας εκείνο το υπερεπίπεδο διαχωρισμού με το μέγιστο περιθώριο, ανάμεσα στα πολλά υπερεπίπεδα που μπορούν να διαχωρίσουν τα θετικά από τα αρνητικά παραδείγματα του χώρου δεδομένων. Επίσης, η συνάρτηση απόφασης για την ταξινόμηση των σημείων περιλαμβάνει μόνο το εσωτερικό γινόμενο των σημείων στο χώρο δεδομένων. Επειδή, ο αλγόριθμος που καθορίζει το υπερεπίπεδο διαχωρισμού στο χώρο δεδομένων μπορεί να διατυπωθεί εξολοκλήρου με όρους των διανυσμάτων εισόδου και των εσωτερικών γινομένων του χώρου δεδομένων, μία support vector machine μπορεί να ορίσει το υπερεπίπεδο με τον ορισμό της συνάρτησης πυρήνα (kernel function), χωρίς να χρειάζεται η ακριβής αναπαράσταση του χώρου. Με τον παραπάνω τρόπο μειώνουμε σημαντικά επίσης, και το κόστος υπολογισμού.

I. Έλεγχος Αξιοπιστίας Αποτελεσμάτων

i. leave one-out cross validation

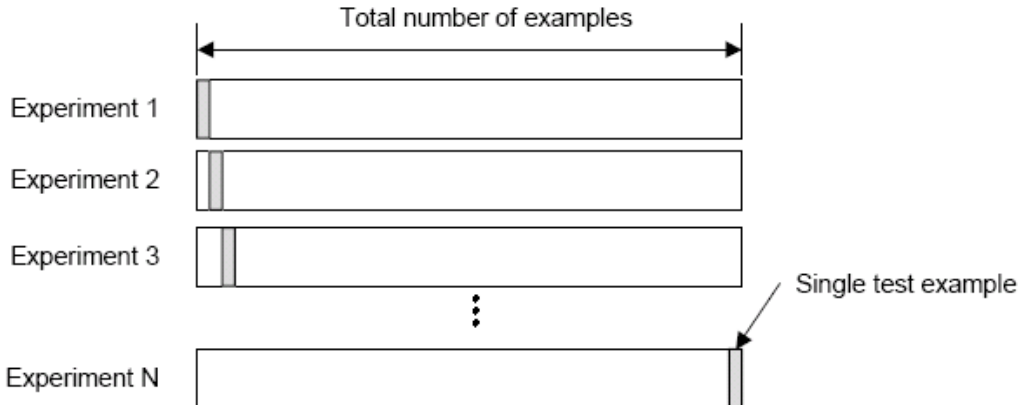
Για την εκτίμηση των αποτελεσμάτων, χρησιμοποιούμε leave one-out cross validation (loocv), που αποτελεί την πιο αξιόπιστη κατηγορία υλοποίησης της μεθόδου cross validation. Η μέθοδος cross validation χρησιμοποιείται για την εκτίμηση της ακρίβειας του μοντέλου ταξινόμησης. Η κατηγορία leave one-out της μεθόδου cross validation είναι η πιο δημοφιλής μέθοδος στην εκτίμηση αυτή. Συνήθως, μετά τη διαδικασία της γονιδιακής επιλογής, το σύνολο δεδομένων που αποκτούμε είναι μικρών διαστάσεων, αφού αυτός είναι και ο τελικός στόχος, ένα όσο το δυνατό μικρότερο σύνολο γονιδίων που θα μπορεί να χαρακτηρίζει τη συμπεριφορά του αρχικού συνολο δεδομένων. Για αυτά τα μικρά διαστάσεων σύνολα δεδομένων, η κατηγορία leave one-out cross validation μέθοδος εκτίμησης είναι μια άμεση τεχνική της οποίας η απόδοση δεν χαρακτηρίζεται από μεγάλη στατιστική απόκλιση (Goutte1997), αν και παρουσιάζει υψηλές τιμές διακύμανσης (Molinaro 2005, Braga- Neto and Dougherty 2003). Αντίθετα, εξαιτίας του μεγάλου υπολογιστικού κόστους, η μέθοδος δεν χρησιμοποιείται για μεγάλων διαστάσεων σύνολα δεδομένων και η συμπεριφορά της

στην εκτίμηση του σφάλματος γενίκευσης δεν έχει μελετηθεί ενδελεχώς.

Η μέθοδος leave one-out cross validation χρησιμοποιείται γενικά για την εκτίμηση του σφάλματος γενίκευσης (generalization error) ενός δεδομένου μοντέλου ή μπορεί να χρησιμοποιηθεί για την επιλογή ενός μοντέλου ανάμεσα σε πολλά. Για παράδειγμα, μπορούμε να χρησιμοποιήσουμε leave one-out cross validation για την επιλογή του αριθμού των κρυφών μονάδων ενός νευρωνικού δικτύου ή μπορούμε να επιλέξουμε ένα καλύτερο υποσύνολο εισόδων κατά τη γραμμική παλινδρόμηση (linear regression), με μικρότερο σφάλμα γενίκευσης. Επίσης, η μέθοδος λειτουργεί σωστά την εκτίμηση του σφάλματος γενίκευσης για συνεχείς συναρτήσεις σφάλματος, όπως είναι το σφάλμα των τετραγώνων μέσης τιμής (mean squared error), ενώ δεν προκύπτουν αξιόπιστα αποτελέσματα για ασυνεχείς συναρτήσεις σφάλματος, όπου συνήθως χρησιμοποιείται η μέθοδος k-fold.

Σύμφωνα με την leave one-out cross validation, χωρίζουμε το σύνολο των δεδομένων σε τόσα μέρη όσα είναι τα δείγματα- παραδείγματα. Έπειτα, αν N είναι ο συνολικός αριθμός των δειγμάτων, τότε για κάθε πείραμα χρησιμοποιούμε N-1 δείγματα για εκπαίδευση (training) και το εναπομένον για δοκιμή (testing). Το συνολικό λάθος ταξινόμησης υπολογίζεται από το μέσο ρυθμό λάθους σε κάθε δείγμα που δοκιμάζουμε:

$$E = \frac{1}{N}\sum_{i=1}^{N} E_i$$



Διάγραμμα 4: υλοποίηση leave one-out cross validation

ii. Μέθοδος της περιοχής κάτω από την καμπύλη ROC- Area under the ROC Curve

Για να αξιολόγηση της απόδοσης του δυαδικού ταξινομητή (binary classifier) χρησιμοποιούμε τη μέθοδο της περιοχής κάτω από την καμπύλη ROC- Area under the ROC Curve (AUC). Η καμπύλη ROC αποτελεί την έκφραση της σχέσης της ευαισθησίας (sensitivity) με την ιδιαιτερότητα (specificity) του ταξινομητή. Η λειτουργία του ταξινομητή καθορίζεται από την ύπαρξη ενός κατωφλίου. Η επιλογή της τιμής του κατωφλίου μπορεί να επιτύχει τη βέλτιστη λειτουργία του ταξινομητή. Η περιοχή που βρίσκεται κάτω από την καμπύλη μπορεί να αποτελέσει ένα αξιόπιστο μέτρο της απόδοσης του ταξινομητή. Η τιμή του εμβαδού αυτής της περιοχής μπορεί να κυμαίνεται από 0.5 μέχρι 1. όσο πιο κοντά στο 1 είναι η τιμή του εμβαδού, τόσο πιο καλή η λειτουργία του ταξινομητή. Η ακραία περίπτωση της τιμής 0.5 δείχνει ταξινόμηση των δεδομένων με βάση και μόνο την τύχη, ενώ η τιμή 1 σημαίνει άριστη λειτουργία του ταξινομητή.

Κεφάλαιο 3: Παρουσίαση Αποτελεσμάτων

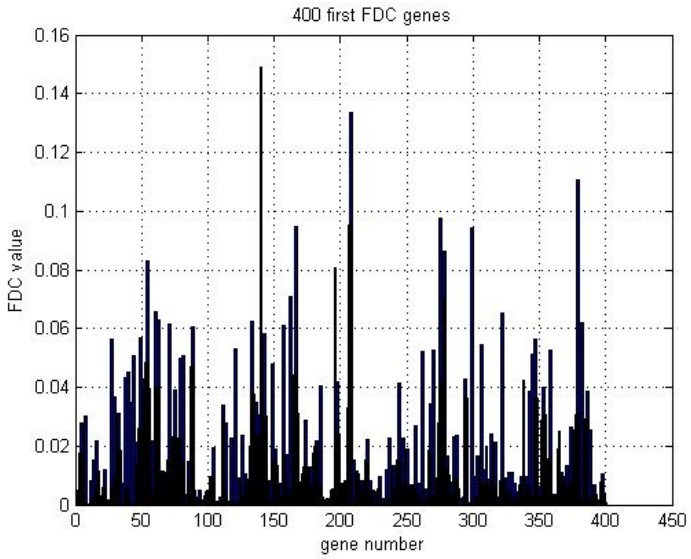
Για την πρώτη υλοποίηση της μεθόδου χρησιμοποιήσαμε τα δεδομένα που περιγράφονται στη δημοσίευση των Van't Veer et al. (2002) από το «Ολλανδικό Ινστιτούτο για τον καρκίνο (The Netherlands Cancer Institute)». Τα δεδομένα αφορούν την ασθένεια του καρκίνου του μαστού και χωρίζονται σε δύο κατηγορίες- κλάσεις:

1. 34 δείγματα από ασθενείς, οι οποίοι ανέπτυξαν άμεση μετάσταση μέσα σε διάστημα πέντε ετών
2. 44 δείγματα από ασθενείς, οι οποίοι δεν ανέπτυξαν κάποια σχετική μετάσταση μετά από διάστημα πέντε ετών

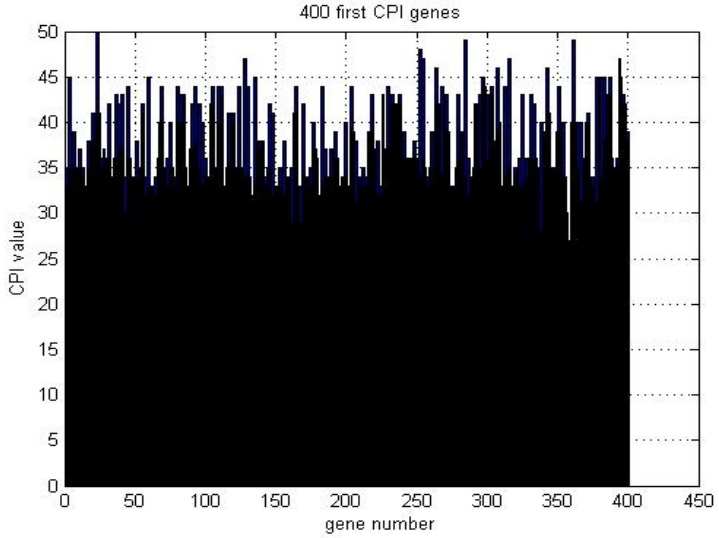
Σύμφωνα με τη διαδικασία που περιγράφηκε στο κεφάλαιο 1, 1.2.Microarray Technology, επιλέξαμε περίπου 25000 (24481) γονίδια για μελέτη. (Για την μέγιστη/ ελάχιστη fold απόκλιση, αποκλείστηκαν τα γονίδια με λιγότερη από τρεις τάξεις μεγέθους απόκλιση και για την μέγιστη/ ελάχιστη απόλυτη απόκλιση αποκλίσαμε τα γονίδια με απόλυτη απόκλιση μικρότερη των 100 μονάδων, βλέπε κεφάλαιο 2, 2.3 Προεργασία Δεδομένων).

Α. Αποτελέσματα των Αλγορίθμων Κατάταξης

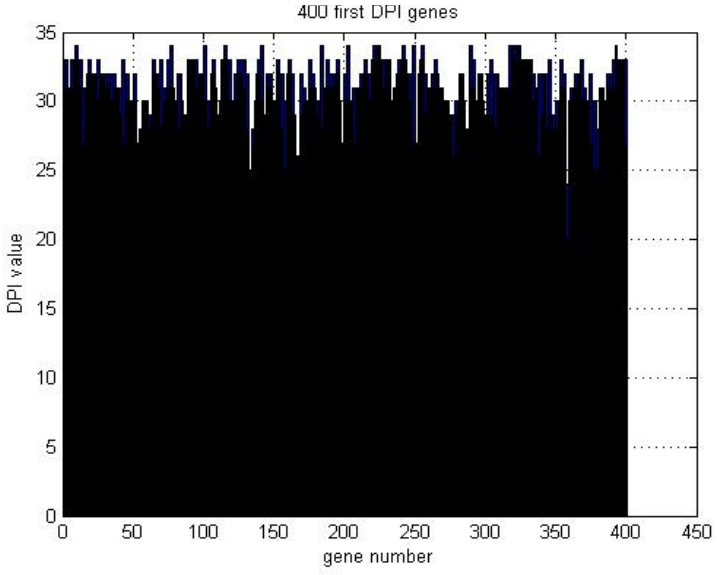
Αρχικά δημιουργήσαμε τις λίστες κατάταξης των γονιδίων FDCL, CPIL, DPIL. Ενδεικτικά, για τις τιμές των τετρακοσίων πρώτων γονιδίων της κάθε λίστας έχουμε τα αντίστοιχα διαγράμματα:



διάγραμμα 5: ενδεικτική απεικόνιση των FDC τιμών των τετρακοσίων πρώτων γονιδίων



διάγραμμα 6: ενδεικτική απεικόνιση των CPI τιμών των τετρακοσίων πρώτων γονιδίων



διάγραμμα 7: ενδεικτική απεικόνιση των DPI τιμών των τετρακοσίων πρώτων γονιδίων

Για κάθε λίστα κατάταξης των γονιδίων, εφαρμόσαμε εποπτική ταξινόμηση. Η εκτίμηση των αποτελεσμάτων, ήταν η εξής:

Δεδομένα:Van't Veer (unpreprocessed)	
#GENES*	**accuracy(%)****
1000	**85.90**
600	**82.05**
400	**82.05**
200	**85.90**
150	**83.33**
120	**82.05**
100	**80.77**
80	**79.49**
50	**69.23**
20	**67.95**
*the first top-ranked	

Πίνακας 2: ακρίβεια αποτελεσμάτων των δεδομένων από τη λίστα FDCL

Δεδομένα:Van't Veer (unpreprocessed)	
#GENES*	**accuracy(%)****
1000	**66.67**
600	**67.95**
400	**78.21**
200	**76.92**
150	**70.51**
120	**67.95**
100	**73.08**
80	**70.51**
50	**66.67**
20	**66.67**
*the first top-ranked	

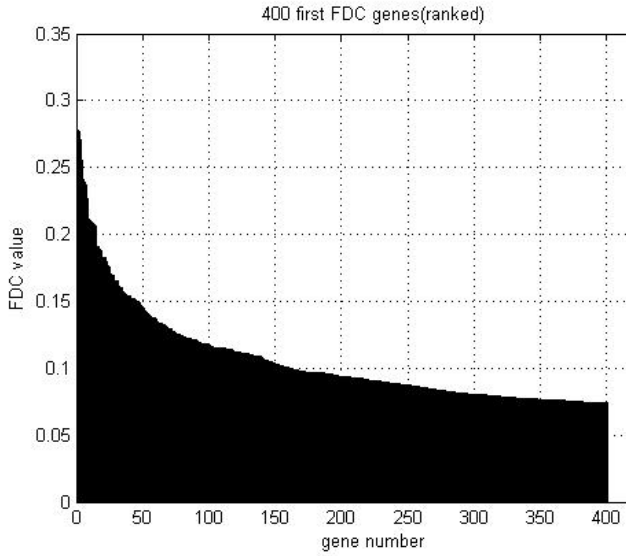
Πίνακας 3: ακρίβεια αποτελεσμάτων των δεδομένων από τη λίστα CPIL

Δεδομένα:Van't Veer (unpreprocessed)	
#GENES*	**accuracy(%)****
1000	**76.92**
600	**78.21**
400	**82.05**
200	**80.77**
150	**82.05**
120	**82.05**
100	**82.05**
80	**78.21**
50	**70.51**
20	**70.51**
*the first top-ranked	

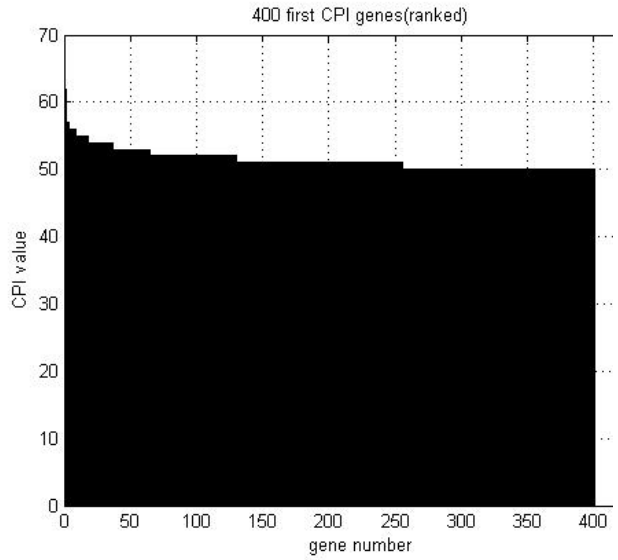
Πίνακας 4: ακρίβεια αποτελεσμάτων των δεδομένων από τη λίστα DPIL

Β. Αποτελέσματα της Αλγοριθμικής Σύντηξης

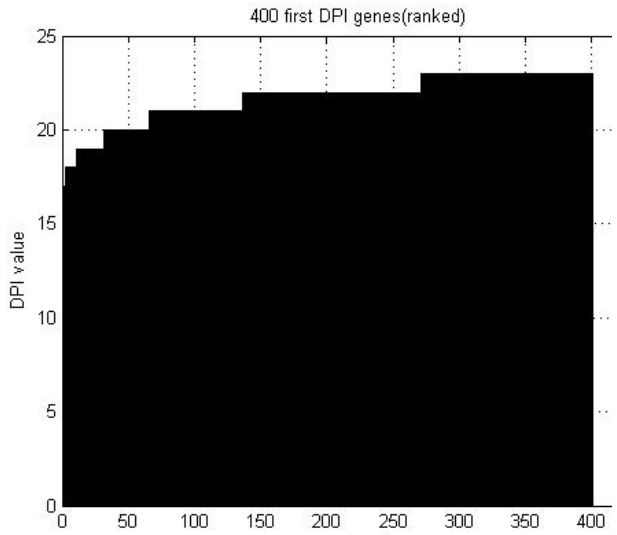
Με τη διαμόρφωση της εισόδου στον G_FUSION αλγόριθμο (βλέπε στον ορισμό του αλγορίθμου: είσοδοι), για τις τιμές των τετρακοσίων πρώτων γονιδίων της κάθε λίστας έχουμε τα αντίστοιχα διαγράμματα:



διάγραμμα 8: ενδεικτική απεικόνιση των FDC τιμών των τετρακοσίων πρώτων γονιδίων σε φθίνουσα κατάταξη



διάγραμμα 9: ενδεικτική απεικόνιση των CPI τιμών των τετρακοσίων πρώτων γονιδίων σε φθίνουσα κατάταξη



διάγραμμα 10: ενδεικτική απεικόνιση των DPI τιμών των τετρακοσίων πρώτων γονιδίων σε αύξουσα κατάταξη

Έπειτα, εφαρμόσαμε σε πολλούς συνδυασμούς (μεταβάλλοντας τις τιμές του d) τον G_FUSION αλγόριθμο.

Τα αποτελέσματα, μετά την εποπτική ταξινόμηση και την εκτίμηση των αποτελεσμάτων, ήταν τα εξής (η επιλογή των γονιδίων έγινε τυχαία σε πολλές επαναλήψης, ώστε το αποτέλεσμα που προκύπτει να είναι στατιστικά όσο το δυνατό πιο ακριβές):

Δεδομένα:Van't Veer (unpreprocessed)	
#GENES*	**accuracy(%)****
24000	**61.54**
15000	**61.54**
10000	**63.37**
6000	**64.1**
2000	**64.47**
1500	**65.02**
1000	**61.17**
600	**62.72**
400	**63.49**
200	**58.54**
150	**60.32**
120	**64.45**
100	**60.22**
80	**61.3**
50	**62.43**
20	**53.46**
*random selection	
**10-100 repetitions	
Πίνακας 5: ακρίβεια αποτελεσμάτων του συνόλου δεδομένων, **χωρίς** την εφαρμογή της αλγοριθμικής σύντηξης (τυχαία επιλογή πολλών επαναλήψεων)	

GENE SELECTION METHOD of algorithmic fusion for 25,000 genes *(Veer dataset)*			
GF1		**GF2**	
#GENES	**accuracy(%)****	**#GENES**	**accuracy(%)**
1000	**82.05**	**200**	**82.05**
		150	**85.90**
		120	**84.62**
		100	**84.62**
		80	**84.62**
		50	**83.33**
		20	**80.77**

GENE SELECTION METHOD of algorithmic fusion for 25,000 genes *(Veer dataset)*			
GF1		**GF2**	
#GENES	**accuracy(%)****	**#GENES**	**accuracy(%)**
600	**83.33**	200	82.05
		150	85.90
		120	83.33
		100	84.62
		80	84.62
		50	83.33
		20	80.77

GENE SELECTION METHOD of algorithmic fusion for 25,000 genes *(Veer dataset)*			
GF1		**GF2**	
#GENES	**accuracy(%)****	**#GENES**	**accuracy(%)**
400	**85.90**	200	80.77
		150	82.05
		120	82.05
		100	82.05
		80	82.05
		50	83.33
		20	80.77

GENE SELECTION METHOD of algorithmic fusion for 25,000 genes *(Veer dataset)*			
GF1		**GF2**	
#GENES	**accuracy(%)****	**#GENES**	**accuracy(%)**
200	**82.05**	150	78.21
		120	78.21
		100	80.77
		80	79.49
		50	82.05
		20	78.21

GENE SELECTION METHOD of algorithmic fusion for 25,000 genes *(Veer dataset)*			
GF1		**GF2**	
#GENES	**accuracy(%)****	**#GENES**	**accuracy(%)**
150	**78.21**	120	74.36
		100	75.64
		80	76.92
		50	78.21
		20	71.79

GENE SELECTION METHOD of algorithmic fusion for 25,000 genes *(Veer dataset)*			
GF1		**GF2**	
#GENES	**accuracy(%)****	**#GENES**	**accuracy(%)**
120	**79.49**	100	76.92
		80	76.92
		50	76.92
		20	71.79

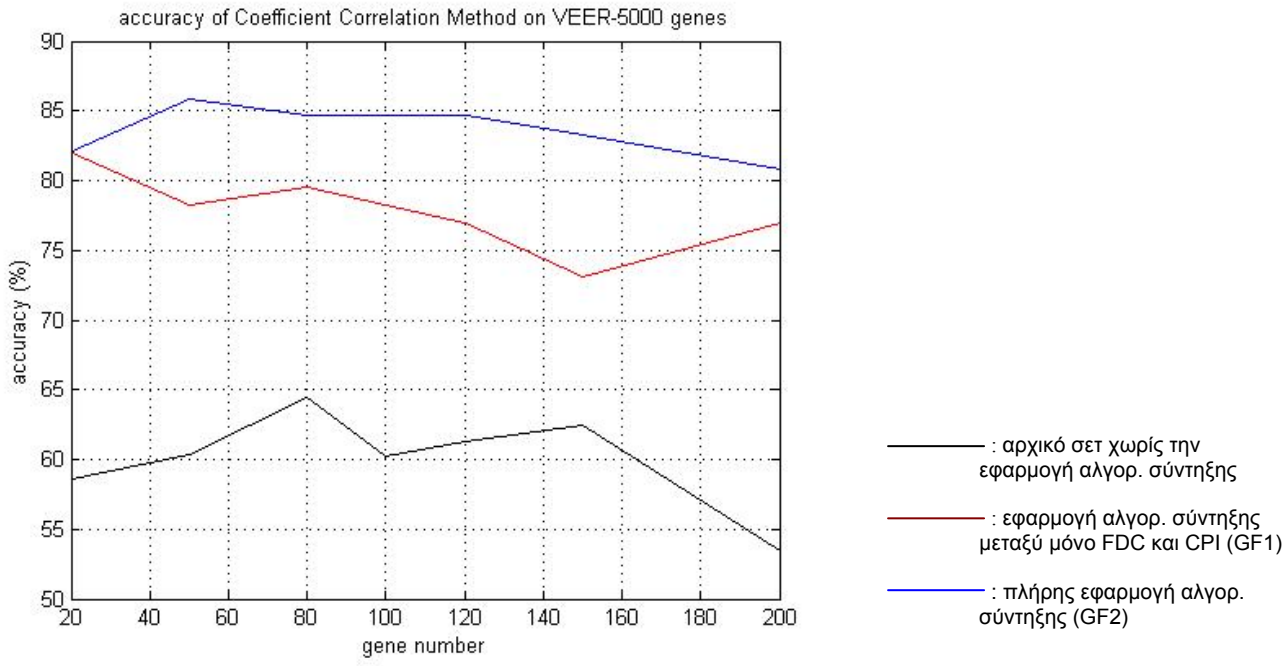
GENE SELECTION METHOD of algorithmic fusion for 25,000 genes (Veer dataset)			
GF1		**GF2**	
#GENES	**accuracy(%)****	**#GENES**	**accuracy(%)**
		80	78.21
		50	78.77
100	78.21	20	70.51

GENE SELECTION METHOD of algorithmic fusion for 25,000 genes (Veer dataset)			
GF1		**GF2**	
#GENES	**accuracy(%)****	**#GENES**	**accuracy(%)**
80	76.92	50	76.92
		20	69.21

GENE SELECTION METHOD of algorithmic fusion for 25,000 genes (Veer dataset)			
GF1		**GF2**	
#GENES	**accuracy(%)****	**#GENES**	**accuracy(%)**
50	73.08	20	76.92

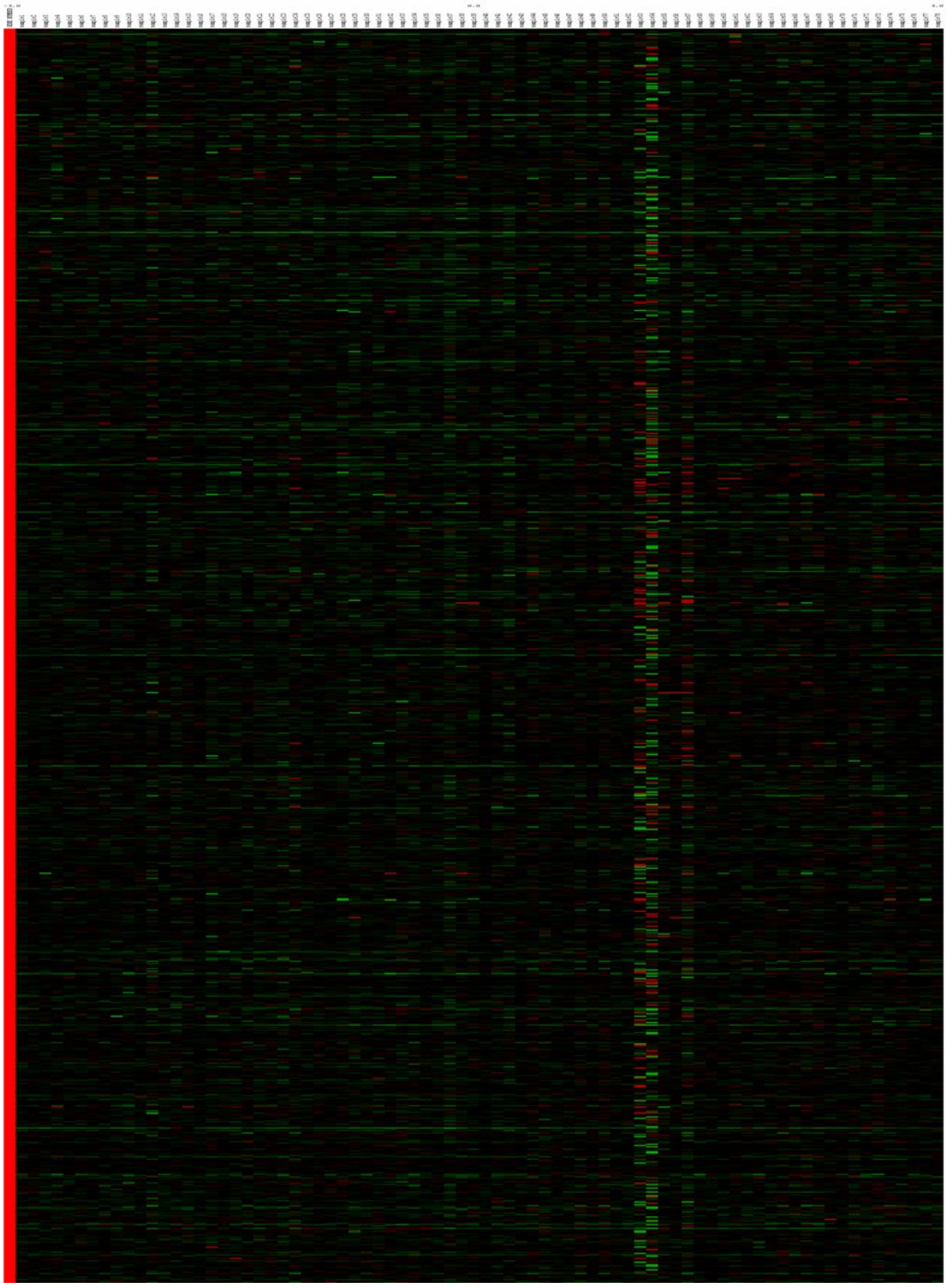
GENE SELECTION METHOD of algorithmic fusion for 25,000 genes (Veer dataset)	
GF1	
#GENES	**accuracy(%)****
20	76.92

Πίνακας 6: αποτελέσματα της αλγοριθμικής σύντηξης μετά από τη διαδοχική εφαρμογή του αλγορίθμου G_FUSION για διάφορες τιμές επιλογής γονιδίων

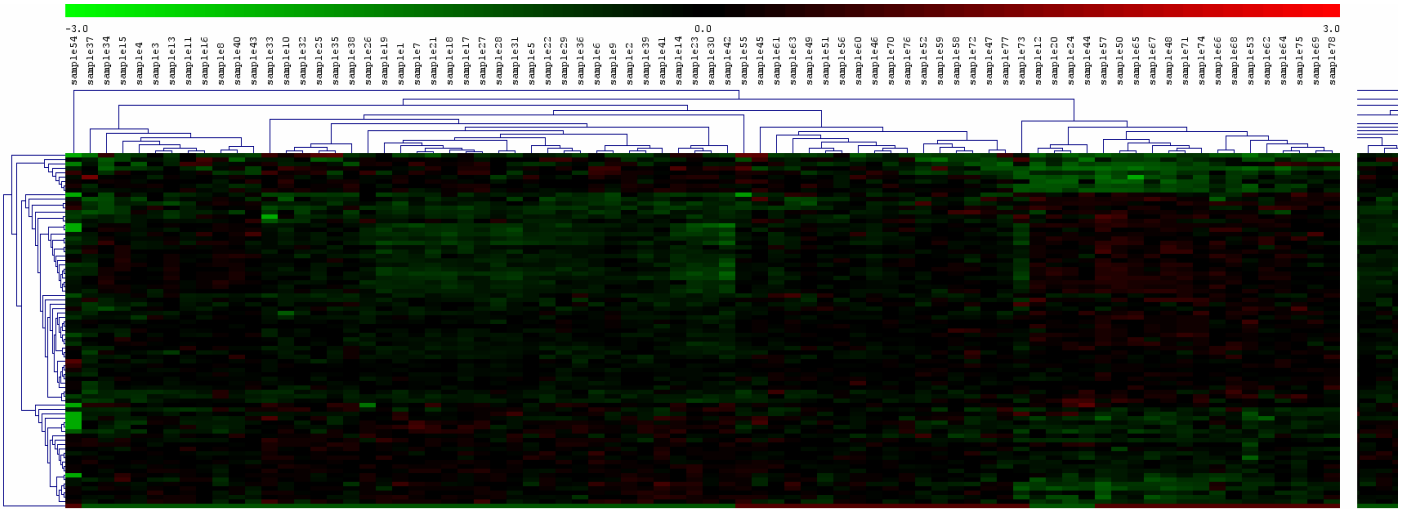


Διάγραμμα 11: γραφική αναπαράσταση των αποτελεσμάτων της αλγοριθμικής σύντηξης για 25.000 γονίδια

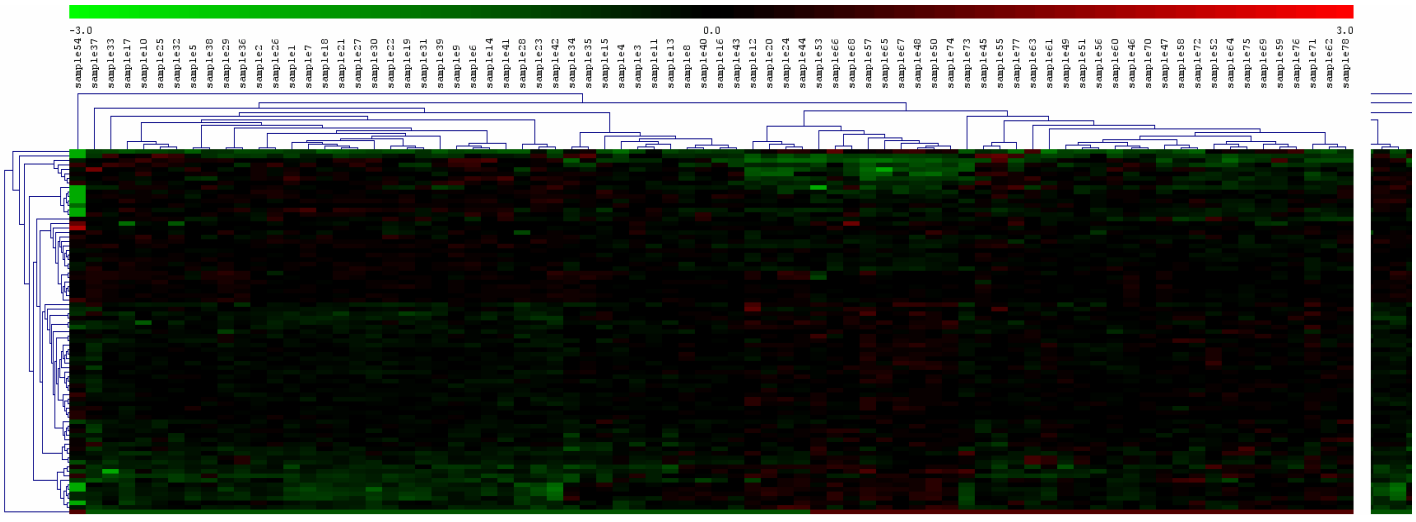
Παραθέτουμε κάποιες εικόνες όπου απεικονίζονται τα γονίδια, σε κάποιες ενδεικτικές εφαρμογές της μεθόδου της αλγοριθμικής σύντηξης:



Εικόνα 4: απεικόνιση των 1000 γονιδίων που προκύπτουν από την 1η αλγοριθμική σύντηξη



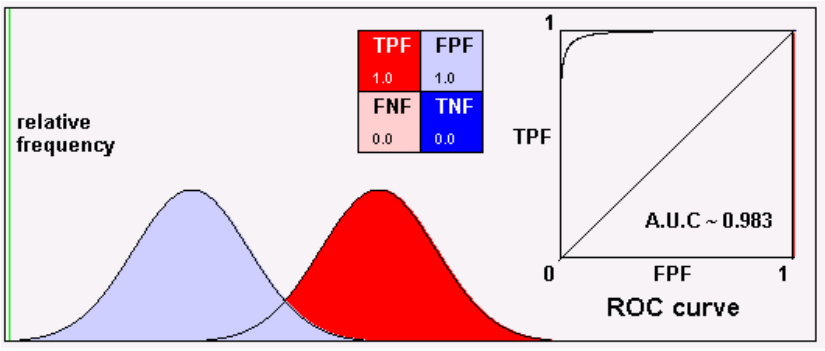
Εικόνα 5: απεικόνιση των 50 γονιδίων που προκύπτουν απευθείας από την 1η εφαρμογή του G_FUSION στο σύνολο των δεδομένων της Van't Veer



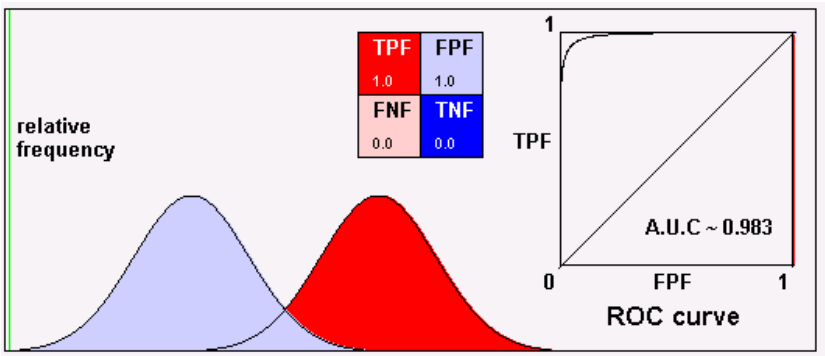
Εικόνα 6: απεικόνιση των 50 γονιδίων που προκύπτουν από την 2η αλγοριθμική σύντηξη των 1000 γονιδίων της FDC/CPI-L και της DPI-L

Γ. Μέθοδος της περιοχής κάτω από την καμπύλη ROC- Area under the ROC Curve

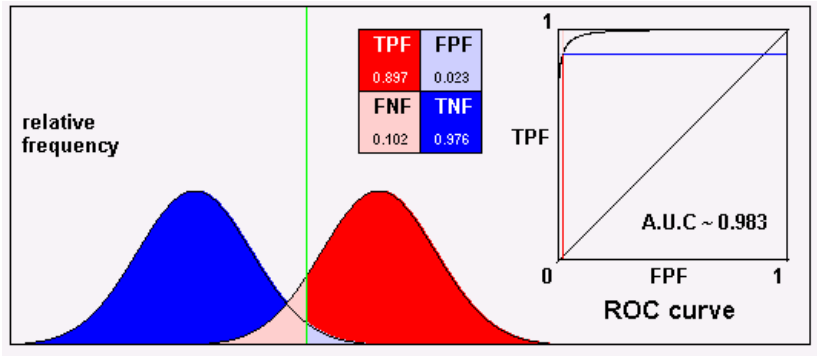
Για τα δεδομένα μας, όπως αυτά προκύπτουν από την εφαρμογή της αλγοριθμικής σύντηξης στην περίπτωση επιλογής 1000 γονιδίων στην πρώτη εφαρμογή του αλγορίθμου GENE_FUSION και στη διαδοχική επιλογή 80, 50 και 20 γονιδίων κατά την δεύτερη εφαρμογή του, είναι τα εξής:



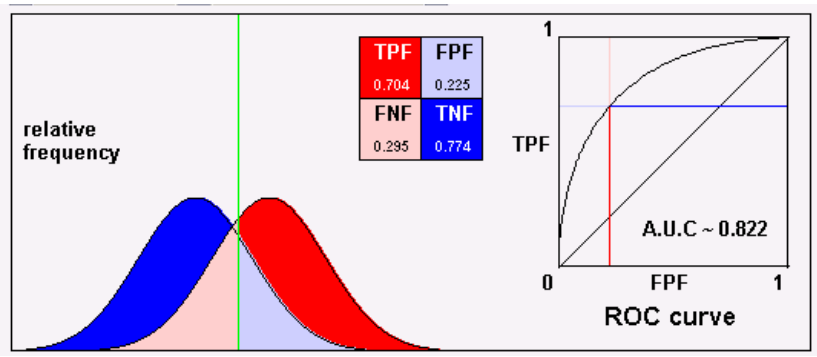
Διάγραμμα 12. Αποτέλεσμα της αξιολόγησης του ταξινομητή, όπως προκύπτει από την επιλογή 1000 γονιδίων κατά τη πρώτη εφαρμογή του αλγορίθμου GENE_FUSION στη διαδικασία της αλγοριθμικής σύντηξης



Διάγραμμα 13. Αποτέλεσμα της αξιολόγησης του ταξινομητή, όπως προκύπτει από την επιλογή 80 γονιδίων κατά τη δεύτερη εφαρμογή του αλγορίθμου GENE_FUSION στη διαδικασία της αλγοριθμικής σύντηξης



Διάγραμμα 14. Αποτέλεσμα της αξιολόγησης του ταξινομητή, όπως προκύπτει από την επιλογή 50 γονιδίων κατά τη δεύτερη εφαρμογή του αλγορίθμου GENE_FUSION στη διαδικασία της αλγοριθμικής σύντηξης



Διάγραμμα 15. Αποτέλεσμα της αξιολόγησης του ταξινομητή, όπως προκύπτει από την επιλογή 20 γονιδίων κατά τη δεύτερη εφαρμογή του αλγορίθμου GENE_FUSION στη διαδικασία της αλγοριθμικής σύντηξης

Δ. Παρατηρήσεις

- Η ακρίβεια των αποτελεσμάτων της αλγοριθμικής σύντηξης είναι εμφανώς μεγαλύτερη από το αρχικό σύνολο δεδομένων.

- Για τον ίδιο αριθμό επιλογής γονιδίων (ίδιο d), κατά τη δεύτερη εφαρμογή του G_FUSION αλγορίθμου (GF2), η ακρίβεια των αποτελεσμάτων είναι πολύ πιο μεγάλη σε σχέση με την πρώτη εφαρμογή του G_FUSION αλγορίθμου (GF1), κάτι που μπορούμε να διαπιστώσουμε και από τις εικόνες 5 και 6, όπου φαίνεται ξεκάθαρα η διαφορετική έκφραση των επιλεγόμενων γονιδίων ανάμεσα στα δείγματα που είναι προσβεβλημένα από καρκίνο και σε εκείνα που δεν είναι.

- Με μια πιο συνολική ματιά, η ακρίβεια των αποτελεσμάτων μας επιτρέπει να επιλέξουμε έναν ελάχιστο αριθμό γονιδίων, αλλά όχι με ασφάλεια, γι' αυτό και απαιτείται περαιτέρω μελέτη επάνω στη μέθοδο.

Κεφάλαιο 4: Βελτίωση των Αποτελεσμάτων

Στην προσπάθεια βελτίωσης των αποτελεσμάτων, προσπαθήσαμε αρχικά να βελτιώσουμε την ίδια τη διαδικασία της αλγοριθμικής σύντηξης, συνδυάζοντας με διαφορετικό τρόπο, τις τρεις μεθόδους αλγοριθμικής κατάταξης. Τα αποτελέσματά μας όμως, ήταν σε όλες τις περιπτώσεις πολύ χειρότερα από τα πρώτα, αφού δεν κατάφεραν σε καμία περίπτωση να ξεπεράσουν το 67%.

Στη συνέχεια, επιβεβαιώσαμε τα αποτελέσματα του Qiuming Zhu (Algorithmic Fusion of Gene Expression Profiling for Diffuse Large B-Cell Lymphoma Outcome Prediction, IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 8, NO. 2, JUNE 2004). Έτσι, επιλέγοντας στην πρώτη εφαρμογή του αλγορίθμου G_FUSION 50 γονίδια και στη δεύτερη 13, βρίσκουμε τα εξής γονίδια:

Γονίδιο (κατάταξη κατά σειρά εμφάνισης στην έξοδο του αλγορίθμου)

αριθμός γονιδίου	γονίδιο
6641	Bytyrophilin (BTF1) mRNA
5177	mRNA (non-coding)
6991	Dystrobrevin-alpha mRNA
4721	Splicing factor SF3a120
6206	PRKACB
6873	Transducin-like enhancer protein 1
3673	SM15
4097	PRKACB1
5849	IGF2 insulin like growth factor 2
7045	Growth hormone
5244	HADHA Hyndroxyacyl- Coenzyme
1394	PDE4B
3861	Nuclear antigen H731 mRNA

Πίνακας 7: τα γονίδια- δείκτες που προκύπτουν από την εφαρμογή της αλγοριθμικής σύντηξης στα δεδομένα του Zhu

Πρόκειται για τα **ίδια ακριβώς γονίδια** στα οποία καταλήγει ο καθηγητής Zhu. Το ποσοστό ακρίβειας του παραπάνω συνόλου γονιδίων ανέρχεται στο 82.05% παραπάνω επιβεβαίωση των αποτελεσμάτων, μπορεί να αποδείξει ότι η υλοποίηση της μεθόδου είναι ακριβής, σύμφωνα με τα πρότυπα των συγγραφέων.

Τέλος, στη συνέχεια της προσπάθειας για βελτίωση των αποτελεσμάτων, κινηθήκαμε σε τρεις κατευθύνσεις:

- εφαρμόσαμε στα δεδομένα μας μια πιο αυστηρή διαδικασία προεργασίας. Επίσης, εφαρμόσαμε τη μέθοδο της αλγοριθμικής σύντηξης σε άλλα δεδομένα, που αφορούσαν την ασθένεια της λευχαιμίας (golub).

- υλοποιήσαμε την μέθοδο κατάταξης γονιδίων της συσχέτισης συντελεστών (coefficient correlation) στα τρία σύνολα δεδομένων, όπου εφαρμόσαμε την αλγοριθμική σύντηξη (veer, veer- preprocessed, golub).

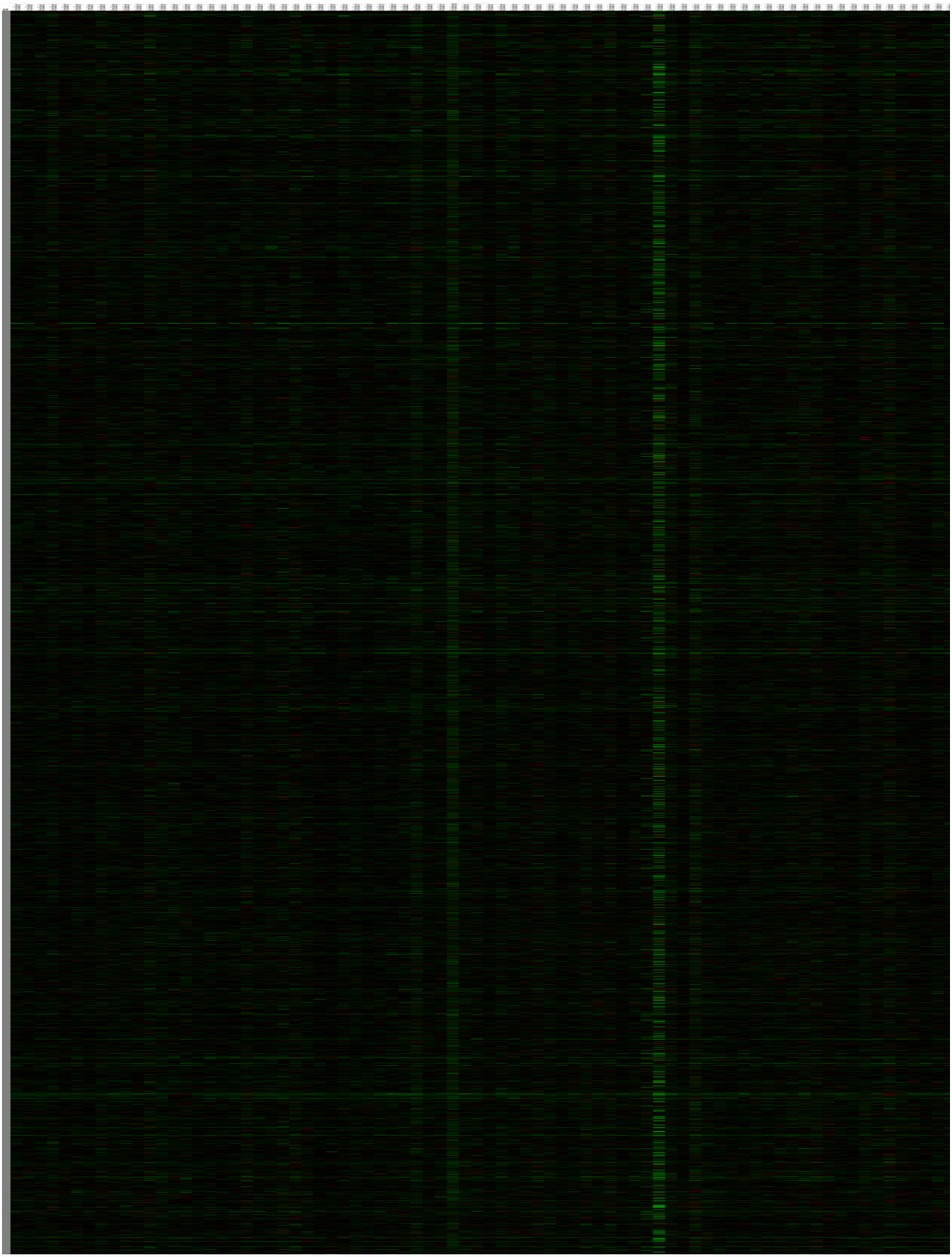
- Επεξεργαστήκαμε τα κριτήρια επιλογής- εξαγωγής του αλγορίθμου G_FUSION (βήμα 4^{α}), με σκοπό τη βελτιστοποίηση του αλγορίθμου.

Α. Εφαρμογή της γονιδιακής σύντηξης σε διαφορετικά σύνολα δεδομένων

a. Προεπεξεργασία συνόλου δεδομένων Van't Veer:

Υλοποιώντας πιο αυστηρά το στάδιο της προεργασίας μπορούμε να έχουμε λιγότερα γονίδια με καλύτερες όμως προοπτικές ως προς τα αποτελέσματα της επεξεργασίας τους. Έτσι, επιλέξαμε περίπου 5000 (5188) γονίδια. Ο τρόπος επιλογής αυτών των γονιδίων είναι: Για την μέγιστη/ ελάχιστη fold απόκλιση, αποκλείστηκαν τα γονίδια με λιγότερη από δύο τάξεις μεγέθους απόκλιση και τιμή P μικρότερη από 0.01 για περισσότερα από πέντε δείγματα όγκου, (βλέπε Κεφάλαιο2 , 2.3. Προεργασία Δεδομένων).

Στην παρακάτω εικόνα φαίνονται τα γονίδια που επιλέξαμε, αν και η εικόνα δεν είναι αρκετά ευκρινής λόγω του ακόμα μεγάλου αριθμού γονιδίων. Παρόλα αυτά με μια πιο προσεκτική παρατήρηση, ακόμα και με γυμνό μάτι, μπορεί να φανεί η αντιπροσωπευτικότητα αυτού του συνόλου δεδομένων σε σχέση με το αρχικό.



Εικόνα 7: απεικόνιση των 5000 γονιδίων που προκύπτουν από την προεργασία στα δεδομένα της Veer

Τα αποτελέσματά, μετά την εποπτική ταξινόμηση και την εκτίμηση των αποτελεσμάτων, ήταν τα εξής:

Δεδομένα: Van't Veer (preprocessed)	
#GENES*	**accuracy(%)****
5000	**61.54**
4000	**65.38**
3000	**65.38**
2000	**65.38**
1000	**65.38**
500	**67.95**
200	**62.82**
150	**65.38**
120	**69.23**
100	**66.46**
80	**59.87**
50	**65.56**
20	**62.04**
*random selection	
**10-100 repetitions	

Πίνακας 8: ακρίβεια αποτελεσμάτων του συνόλου δεδομένων, **χωρίς** την εφαρμογή της αλγοριθμικής σύντηξης (τυχαία επιλογή πολλών επαναλήψεων)

GENE SELECTION METHOD of algorithmic fusion for 5,000 genes (Veer dataset)			
GF1		**GF2**	
#GENES	**accuracy(%)**	**#GENES**	**accuracy(%)**
1000	**69.23**	**200**	**73.08**
		150	**74.36**
		120	**71.79**
		100	**74.36**
		80	**73.08**
		50	**78.21**
		20	**78.21**

GENE SELECTION METHOD of algorithmic fusion for 5,000 genes (Veer dataset)			
GF1		**GF2**	
#GENES	**accuracy(%)**	**#GENES**	**accuracy(%)**
200	**76.92**	**150**	**73.08**
		120	**71.79**
		100	**71.79**
		80	**73.08**
		50	**73.08**
		20	**73.08**

GENE SELECTION METHOD of algorithmic fusion for 5,000 genes (Veer dataset)			
GF1		**GF2**	
#GENES	**accuracy(%)**	**#GENES**	**accuracy(%)**
150	**71.79**	**120**	**71.79**
		100	**71.79**
		80	**73.08**
		50	**73.08**
		20	**71.79**

GENE SELECTION METHOD of algorithmic fusion for 5,000 genes (Veer dataset)			
GF1		**GF2**	
#GENES	**accuracy(%)**	**#GENES**	**accuracy(%)**
120	**76.92**	**100**	**71.79**
		80	**73.08**
		50	**73.08**
		20	**71.79**

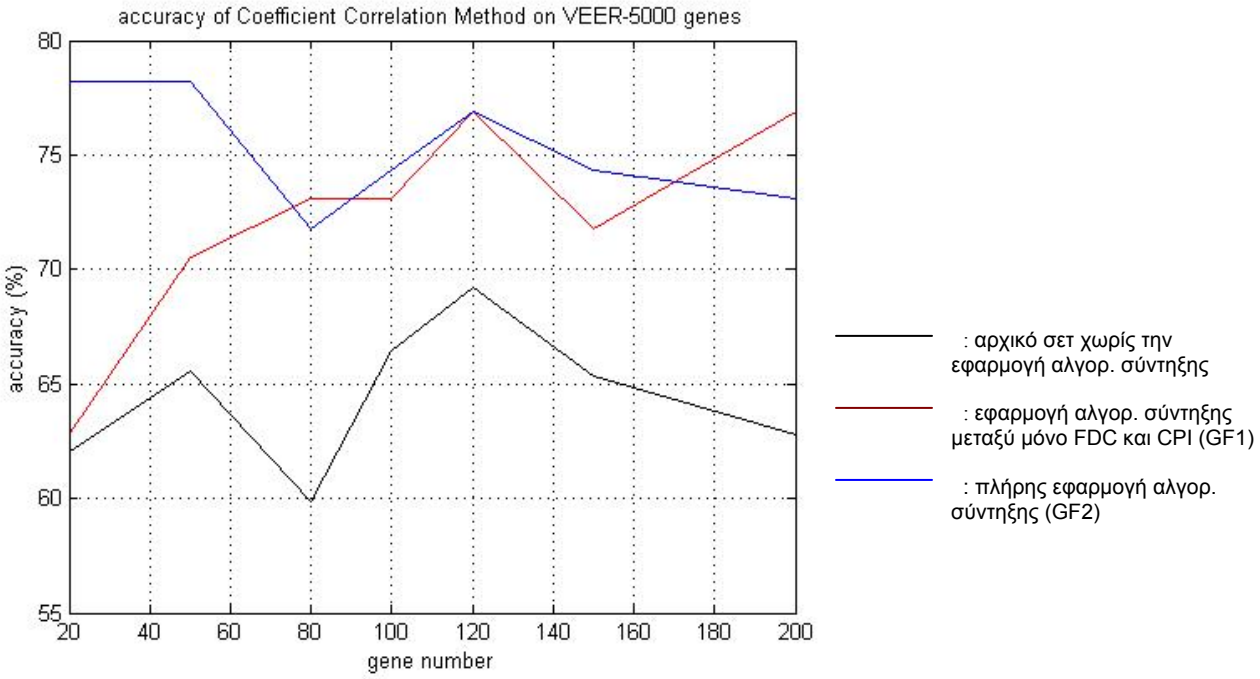
GENE SELECTION METHOD of algorithmic fusion for 5,000 genes (Veer dataset)			
GF1		**GF2**	
#GENES	**accuracy(%)**	**#GENES**	**accuracy(%)**
100	**76.92**	**80**	**73.08**
		50	**71.79**
		20	**71.79**

GENE SELECTION METHOD of algorithmic fusion for 5,000 genes (Veer dataset)			
GF1		**GF2**	
#GENES	**accuracy(%)**	**#GENES**	**accuracy(%)**
80	**73.08**	**50**	**66.67**
		20	**66.67**

GENE SELECTION METHOD of algorithmic fusion for 5,000 genes (Veer dataset)			
GF1		**GF2**	
#GENES	**accuracy(%)**	**#GENES**	**accuracy(%)**
50	**70.51**	**20**	**66.67**

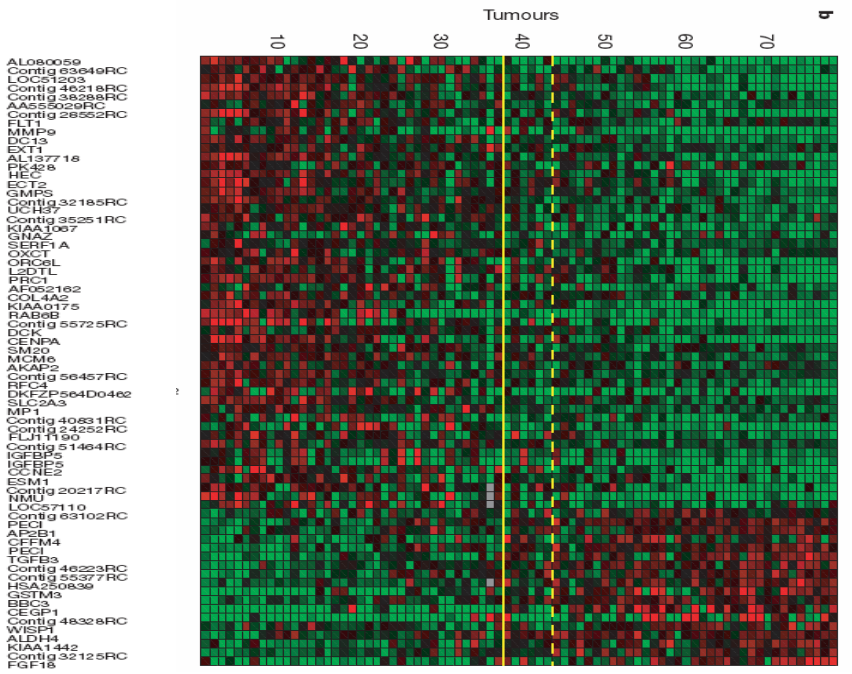
GENE SELECTION METHOD of algorithmic fusion for 5,000 genes (Veer dataset)	
GF1	
#GENES	**accuracy(%)**
20	**62.82**

Πίνακας 9: αποτελέσματα της αλγοριθμικής σύντηξης μετά από τη διαδοχική εφαρμογή του αλγορίθμου G_FUSION για διάφορες τιμές επιλογής γονιδίων (Veer, preprocessed)



Διάγραμμα 16: γραφική αναπαράσταση των αποτελεσμάτων της αλγοριθμικής σύντηξης για 5.000 γονίδια

Από την ταξινόμηση της Van't Veer, στα ίδια δεδομένα (Van't Veer et al.2002), όπου τα δείγματα των όγκων κατατάσσονται ανάλογα με την τιμή της συσχέτισης συντελεστών, προκύπτει η παρακάτω εικόνα:



Εικόνα 8: απεικόνιση των δεδομένων της Veer, όπως αυτά προκύπτουν από την ταξινόμηση με τη μέθοδο συσχέτισης συντελεστών από τη μελέτη της Van't Veer (2002)

Ο ταξινομητής προέβλεψε (loocv) σωστά 65 από τα 78 συνολικά δείγματα, ποσοστό επιτυχίας 83%.

b. **Υλοποίηση με τα δεδομένα του Golub** (Harvard –MIT Division of Health Sciences and Technology, Division of Bioengineering and Environment)

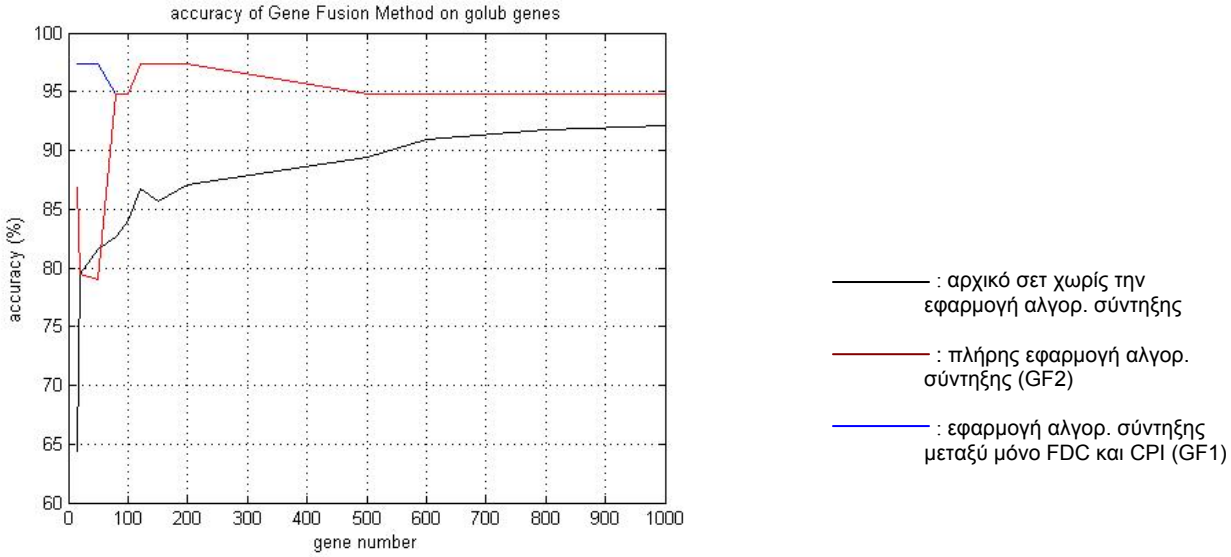
Τα αποτελέσματά, μετά την εποπτική ταξινόμηση και την εκτίμηση των αποτελεσμάτων, ήταν τα εξής:

Δεδομένα: Golub	
# GENES*	**accuracy(%)****
7000	**94.74**
6000	**94.74**
5000	**94.74**
4000	**94.74**
3000	**94.74**
2000	**94.74**
1000	**92.11**
500	**89.47**
100	**82.63**
50	**79.47**
10	**64.44**
5	**63.16**
*random selection	
**10-100 repetitions	

Πίνακας 10: ακρίβεια αποτελεσμάτων του συνόλου δεδομένων, **χωρίς** την εφαρμογή της αλγοριθμικής σύντηξης (τυχαία επιλογή πολλών επαναλήψεων)

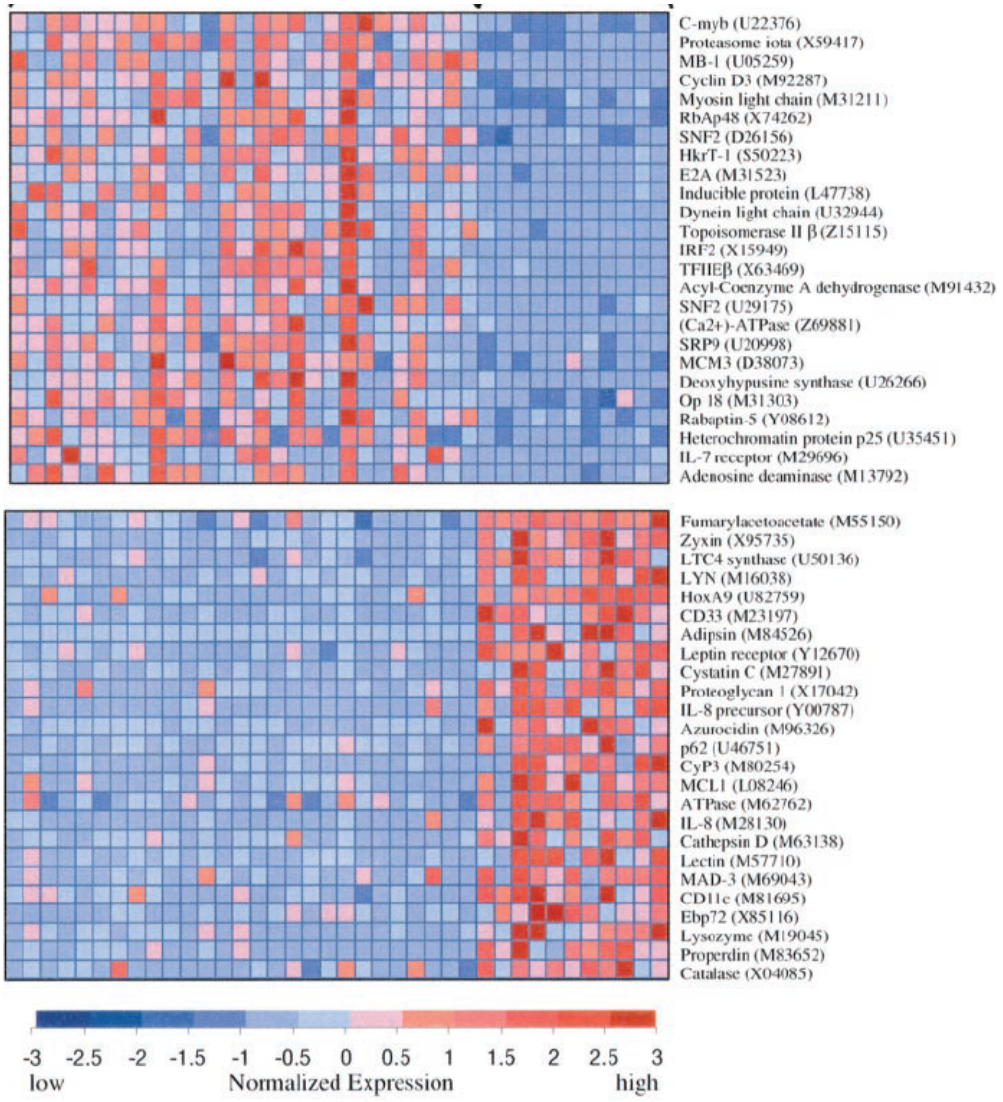
GENE SELECTION METHOD of algorithmic fusion for Golub Dataset			
GF1		**GF2**	
#GENES	**accuracy(%)**	**#GENES**	**accuracy(%)**
1000	**97.37**	**200**	**97.37**
		150	**97.37**
		120	**97.37**
		100	**97.37**
		80	**97.37**
		50	**97.37**
		20	**94.40**

Πίνακας 11: αποτελέσματα της αλγοριθμικής σύντηξης από τη διαδοχική εφαρμογή του αλγορίθμου G_FUSION (Golub)



Διάγραμμα 17: γραφική αναπαράσταση των αποτελεσμάτων της αλγοριθμικής σύντηξης για τα γονίδια από το σύνολο δεδομένων του Golub

Από την ταξινόμηση του Golub (Golub et al.1999), όπου τα δείγματα των όγκων κατατάσσονται ανάλογα με την τιμή της συσχέτισης συντελεστών, προκύπτει η παρακάτω εικόνα:



Εικόνα 9: απεικόνιση των δεδομένων του golub, όπως αυτά προκύπτουν από την ταξινόμηση με τη μέθοδο συσχέτισης συντελεστών (Golub et al.1999)

Με βάση τα παραπάνω μπορούμε να κάνουμε τις εξής παρατηρήσεις:

- Τα αποτελέσματα από την εφαρμογή αλγοριθμικής σύντηξης στο επεξεργασμένο σύνολο δεδομένων του Van't Veer έχει σίγουρα χειρότερα αποτελέσματα από αυτή του πρωτότυπου συνόλου

- Για τον ίδιο αριθμό επιλεχθέντων γονιδίων και στην δεύτερη φάση της αλγοριθμικής σύντηξης (στην GF2), όσο πιο μεγάλος είναι ο αριθμός που επιλέγουμε στην πρώτη φάση (GF1), τόσο πιο μεγάλη είναι η ακρίβεια των αποτελεσμάτων.

- Από το σύνολο δεδομένων του Golub αντιλαμβανόμαστε ότι η μέθοδος μπορεί να είναι αποδοτική και ότι η βασική υπευθυνότητα των χαμηλών ποσοστών ακρίβειας στις προηγούμενες περιπτώσεις οφείλεται στην ποιότητα των δεδομένων. (ύπαρξη θορύβου, κ.ά.)

Β. Υλοποίηση της μεθόδου κατάταξης γονιδίων της συσχέτισης συντελεστών (coefficient correlation) στα τρία σύνολα δεδομένων, όπου εφαρμόσαμε την αλγοριθμική σύντηξη (veer, veer- preprocessed, golub).

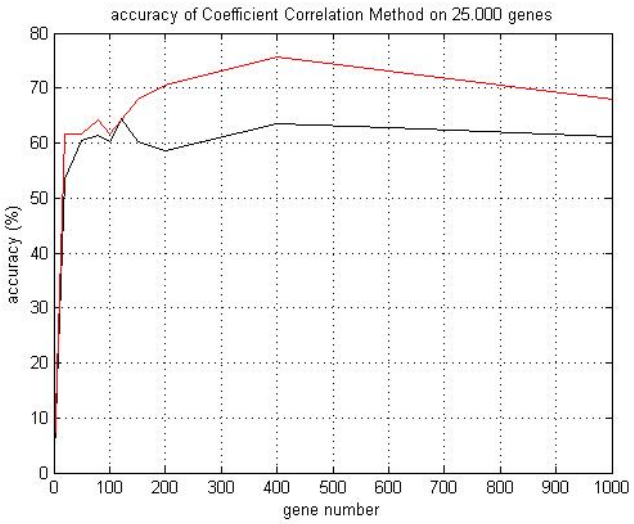
Η μέθοδος της συσχέτισης συντελεστών (golub et al.1999) αποτελεί μια αρκετά απλή μέθοδο κατάταξης και επιλογής γονιδίων. Κατά την υλοποίηση της μεθόδου, κάθε γονίδιο αναπαρίσταται από ένα διάνυσμα έκφρασης $v(g)= (e_1,e_2,\ldots,e_n)$, όπου e_i είναι το επίπεδο έκφρασης του γονιδίου g στο i-οστό δείγμα του αρχικού συνόλου των δειγμάτων. Επίσης, με τη βοήθεια του πίνακα $c= (c_1,c_2,\ldots,c_n)$, όπου $c_i = 1,0$ ανάλογα με το εάν το i-οστό δείγμα ανήκει στην κλάση 1 ή 2, αντίστοιχα, αναπαριστούμε την αντιστοίχηση του κάθε δείγματος στις κλάσεις. Για τη «συσχέτιση» ανάμεσα σε ένα γονίδιο και σε μια τιμή κλάσης χρησιμοποιούμε ένα μέτρο της «συσχέτισης», P(g,c) που χρησιμοποιεί τα γονίδια ως παράγοντες πρόβλεψης. Αν τα $\mu_1(g)$, $\sigma_1(g)$ και $\mu_2(g)$, $\sigma_2(g)$ συμβολίζουν τη μέση τιμή και την τυπική απόκλιση των λογαρίθμων των επιπέδων έκφρασης του γονιδίου g για τις κλάσεις 1 και 2, αντίστοιχα, τότε η συσχέτιση συντελεστών ισούται με $P(g,c)= [\mu_1(g)- \mu_2(g)]/[\sigma_1(g)+ \sigma_2(g)]$, που αναδεικνύει τη διαφορά ανάμεσα στις κλάσεις σε σχέση με την τυπική απόκλιση μεταξύ των κλάσεων.

Τα αποτελέσματά, μετά την εποπτική ταξινόμηση και την εκτίμηση των αποτελεσμάτων για κάθε σύνολο δεδομένων, ήταν τα εξής:

a. Van't Veer (unpreprocessed)

Δεδομένα: Van't Veer (unpreprocessed)	
#GENES*	**accuracy(%)****
24000	**61.54**
15000	**61.54**
10000	**63.37**
6000	**64.1**
2000	**64.47**
1500	**65.02**
1000	**61.17**
600	**62.72**
400	**63.49**
200	**58.54**
150	**60.32**
120	**64.45**
100	**60.22**
80	**61.3**
50	**60.43**
20	**53.46**
*random selection	
**10-100 repetitions	
Πίνακας 12: ακρίβεια αποτελεσμάτων του συνόλου δεδομένων, **χωρίς** την εφαρμογή της συσχέτισης συντελεστών (τυχαία επιλογή πολλών επαναλήψεων)	

Δεδομένα:Van't Veer (unpreprocessed) Coefficient Correlation Method	
GENES	**accuracy(%)**
1000	**67.95**
600	**73.08**
400	**75.64**
200	**70.51**
150	**67.95**
120	**64.1**
100	**61.54**
80	**64.1**
50	**61.54**
20	**61.54**
Πίνακας 13: ακρίβεια αποτελεσμάτων του συνόλου δεδομένων, **μετά** την εφαρμογή της συσχέτισης συντελεστών	



: αρχικό σετ χωρίς την εφαρμογή της συσχέτισης συντελεστών:

: εφαρμογή της συσχέτισης συντελεστών

Διάγραμμα 18: γραφική αναπαράσταση των αποτελεσμάτων της συσχέτισης συντελεστών για τα δεδομένα της Veer (unprerocessed)

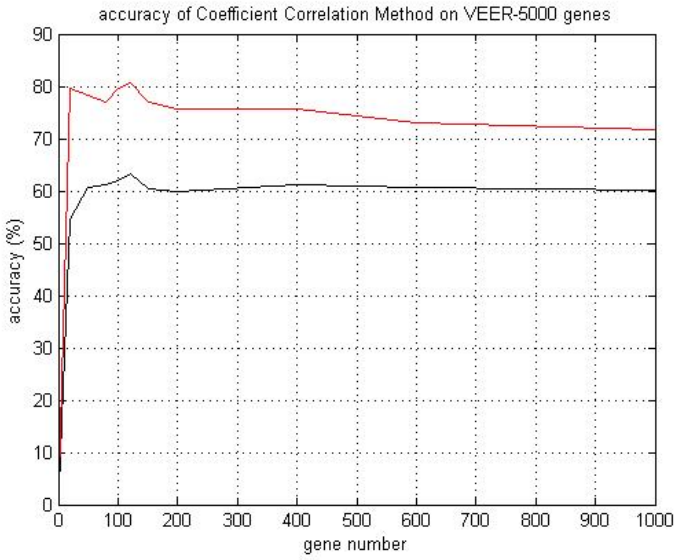
b. Van´t Veer (preprocessed)

Δεδομένα: Van't Veer (preprocessed)	
#GENES*	**accuracy(%)****
5000	**61.43**
3000	**61.68**
1500	**62.42**
1000	**60.22**
600	**60.72**
400	**61.27**
200	**59.94**
150	**60.32**
120	**63.38**
100	**61.87**
80	**61.3**
50	**60.59**
20	**54.69**
*random selection	
**10-100 repetitions	

Πίνακας 14: ακρίβεια αποτελεσμάτων του συνόλου δεδομένων, **χωρίς** την εφαρμογή της συσχέτισης συντελεστών (τυχαία επιλογή πολλών επαναλήψεων)

Δεδομένα: Van't Veer (preprocessed) Coefficient Correlation Method	
GENES	**accuracy(%)**
1000	**71.79**
600	**73.08**
400	**75.64**
200	**75.64**
150	**76.92**
120	**80.77**
100	**79.49**
80	**76.92**
50	**78.21**
20	**79.49**

Πίνακας 15: ακρίβεια αποτελεσμάτων του συνόλου δεδομένων, **μετά** την εφαρμογή της συσχέτισης συντελεστών



: αρχικό σετ χωρίς την εφαρμογή της συσχέτισης συντελεστών:

: εφαρμογή της συσχέτισης συντελεστών

Διάγραμμα 19: γραφική αναπαράσταση των αποτελεσμάτων της συσχέτισης συντελεστών για τα δεδομένα της Veer (prerocessed)

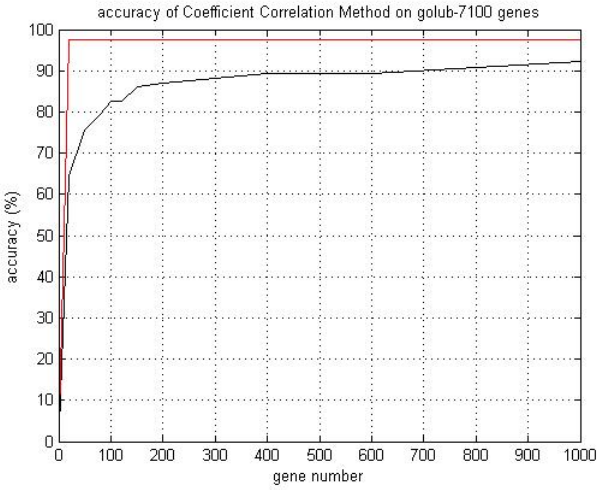
c. Golub dataset

Δεδομένα: Golub	
#GENES*	accuracy(%)**
7000	**94.74**
6000	**94.74**
5000	**94.74**
4000	**94.74**
3000	**94.74**
2000	**94.74**
1000	**92.11**
500	**89.47**
100	**82.63**
50	**79.47**
10	**64.44**
5	**63.16**
*random selection	
**10-100 repetitions	

Πίνακας 16: ακρίβεια αποτελεσμάτων του συνόλου δεδομένων, **χωρίς** την εφαρμογή της συσχέτισης συντελεστών (τυχαία επιλογή πολλών επαναλήψεων)

Δεδομένα: Golub Coefficient Correlation Method	
GENES	accuracy(%)
1000	**97.37**
600	**97.37**
400	**97.37**
200	**97.37**
150	**97.37**
120	**97.37**
100	**97.37**
80	**97.37**
50	**97.37**
20	**97.37**

Πίνακας 17: ακρίβεια αποτελεσμάτων του συνόλου δεδομένων, **μετά** την εφαρμογή της συσχέτισης συντελεστών



: αρχικό σετ χωρίς την εφαρμογή της συσχέτισης συντελεστών:

: εφαρμογή της συσχέτισης συντελεστών

Διάγραμμα 20 : γραφική αναπαράσταση των αποτελεσμάτων της συσχέτισης συντελεστών του συνόλου δεδομένων του Golub

Γ. Επεξεργασία των κριτηρίων επιλογής – εξαγωγής γονιδίων του αλγορίθμου G_FUSION (βήμα 4^{α}), με σκοπό τη βελτιστοποίησή του.

Το κριτήριο του G_FUSION (4α) ορίζει την επιλογή των γονιδίων, ανάλογα με τον αριθμό d, που εμείς ορίζουμε στο 1^{ο} βήμα και τον αριθμό του επιπέδου τιμών της μιας λίστας γονιδίων k, να γίνεται ως εξής:

4 α. Για κάθε γονίδιο g_i,τέτοια ώστε το g_i να έχει τιμή CPI μικρότερη ή ίση με k στη CPIL και να μην βρίσκεται ήδη στη λίστα FDC/CPI-L.
Εάν το γονίδιο g_i είναι παρόν στις τιμές του Μετρητή(k) των γονιδίων του FDCL, **τοποθετούμε το γονίδιο g_i** στη λίστα FDC/CPI-L

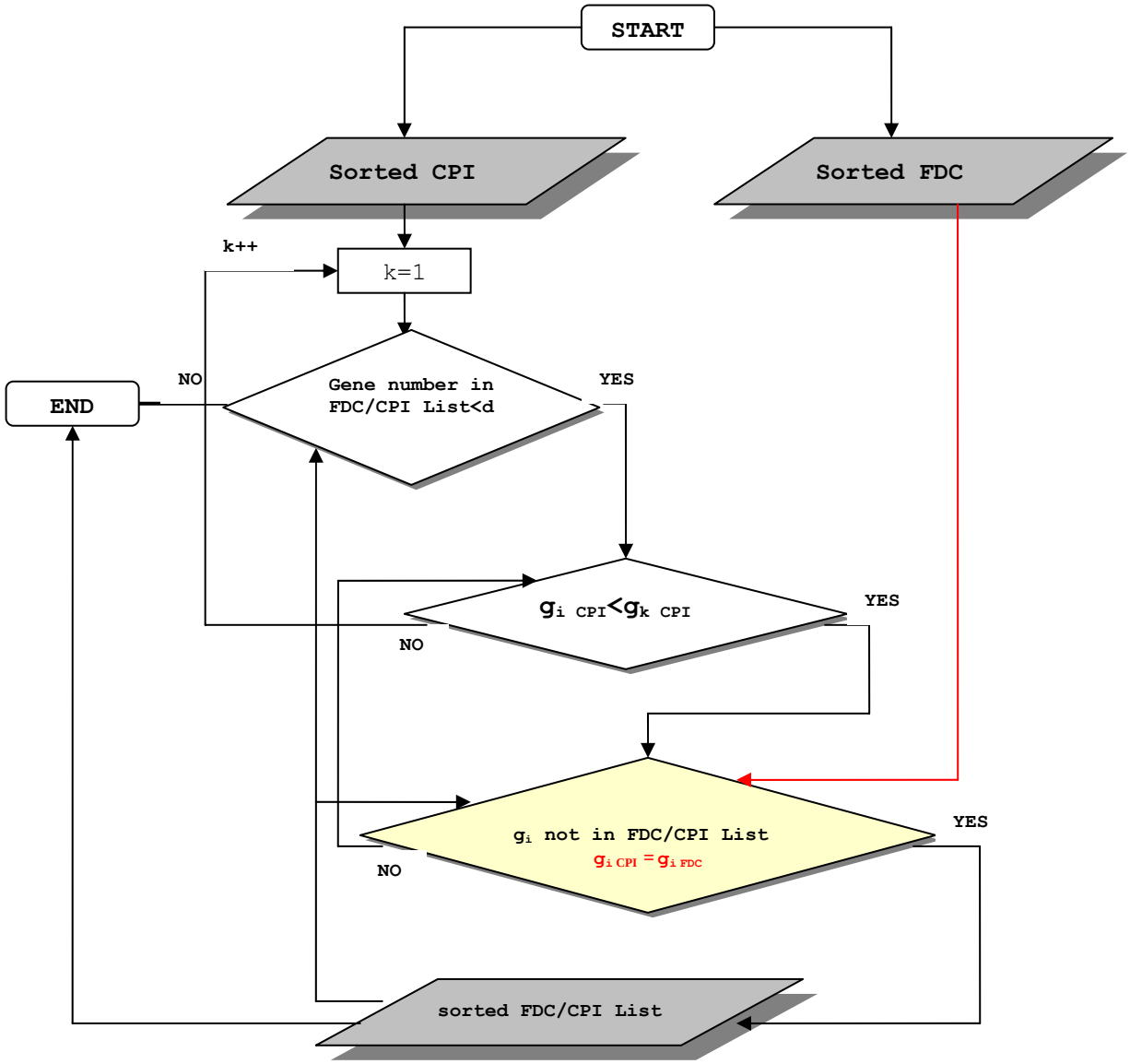
Αυτό σημαίνει ότι ξεκινώντας από το πρώτο επίπεδο τιμών (κόμβος) της λίστας CPIL και συνεχίζοντας μέχρι το τελευταίο επίπεδο τιμών, «συγκρίνουμε» τα γονίδια από τον πρώτο κόμβο της λίστας CPIL μέχρι τον κόμβο στον οποίο βρισκόμαστε, με τα αντίστοιχα της αντίστοιχης λίστας FDCL. Επειδή οι τιμές της λίστας FDCL είναι μοναδικές, αυτό σημαίνει ότι η σύγκρισή για την ύπαρξη του ίδιου γονιδίου στις δυο λίστες γίνεται από το 1^{ο} γονίδιο της λίστας CPIL μέχρι το τελευταίο του υπό εξέταση κόμβου κάθε φορά, και από το 1^{ο} γονίδιο της λίστας FDCL μέχρι το τελευταίο σε σειρά γονίδιο, που βρίσκεται στην αντίστοιχη θέση με το τελευταίο γονίδιο του υπό εξέταση κόμβο της λίστας CPIL

Μια βελτιστοποίηση που μπορούμε να υλοποιήσουμε έχει να κάνει με τον τρόπο των συγκρίσεων. Έτσι, μπορούμε να συγκρίνουμε τα γονίδια από την αρχή της λίστας CPIL μέχρι μόνο το αντίστοιχο κάθε φορά γονίδιο της λίστας FDCL. Με τον τρόπο αυτό μπορούμε να κάνουμε πρώτα τις συγκρίσεις ανάμεσα στα γονίδια με τις μεγαλύτερες τιμές FDC και CPI και κατόπιν τις υπόλοιπες.

Το κριτήριο επιλογής 4(α) του G_FUSION, θα μπορούσε να επαναδιατυπωθεί ως εξής:

4 α. Για κάθε γονίδιο g_i,τέτοια ώστε το g_i να έχει τιμή CPI μικρότερη ή ίση με k στη CPIL και να μην βρίσκεται ήδη στη λίστα FDC/CPI-L.
Εάν το γονίδιο g_i **είναι το ίδιο με το γονίδιο στην αντίστοιχη θέση της FDCL, τοποθετούμε το γονίδιο g_i** στη λίστα FDC/CPI-L

Σχηματικά, στο παρακάτω διάγραμμα φαίνεται το διάγραμμα ροής του αλγορίθμου μετά την τροποποίησή του, όσον αφορά το κριτήριο επιλογής:



Διάγραμμα 21: διάγραμμα της διαδικασίας επιλογής γονιδίων- δεικτών μετά τη βελτιστοποίησή της

Παρόλα αυτά, τα αποτελέσματα της υλοποίησης κυμάνθηκαν στα ίδια περίπου επίπεδα με αυτά της πρώτης υλοποίησης, με μικρές βελτιώσεις, ιδιαίτερα όταν επιλέγουμε μικρό αριθμό γονιδίων προς εξαγωγή.

Κεφάλαιο 5: Τελικές Παρατηρήσεις- Περαιτέρω Εργασία

Από τα παραπάνω μπορούμε να εξάγουμε τα εξής συμπεράσματα:

- Η μέθοδος της αλγοριθμικής σύντηξης μπορεί να αποφέρει πολύ πιο αξιόπιστα αποτελέσματα στην επιλογή χαρακτηριστικών γονιδίων στη μελέτη πρόβλεψης των καρκινικών παθήσεων, από μια συμβατική μέθοδο, όπως είναι αυτή της συσχέτισης συντελεστών. Βασικό ζήτημα για την εξαγωγή συμπερασμάτων αποτελεί το πεδίο εφαρμογής της μεθόδου (σύνολο δεδομένων).

- Η περαιτέρω έρευνα στον αλγόριθμο G_FUSION που υλοποιεί την αλγοριθμική σύντηξη και πιο συγκεκριμένα στα κριτήρια επιλογής των γονιδίων, όπως επίσης και η εφαρμογή της μεθόδου σε περισσότερα σύνολα δεδομενών, μπορεί να αποφέρει βελτιώσεις στη μέθοδο, αλλά και να καταδείξει τυχόν καταλληλότητα της μεθόδου για την πρόβλεψη καρκίνου συγκεκριμένων ειδών καρκίνου.

Βιβλιογραφία

[1] Watson JD, Crick FH. *Molecular Structure of Nucleic Acids. A Structure for Deoxyribose Nucleic Acid*. Nature 1953 April 25; Vol.171;737

[2] Lockhart DJ, Winzeler EA. *Genomics, gene expression and DNA arrays*. Nature. 2000 Jun 15;405(6788):827- 36. Review.

[3] M.J.Allenand, W.M.Yen, *Introduction to Measurement Bioinformatics and software development in Theory*. Belmont, CA: Wadsworth, 1979

[4] R.A.Fisher, *The use of multiple measurements in taxonomic problems* pt. II, vol. 7, pp. 179–188,1936.

[5] J.Yang, and D.Zhang, *What's wrong with Fisher criterion?* Pattern Recognit., vol. 35, no. 11, pp. 2665–2668, 2002.

[6] J.H.Friedman and J.W.Tukey, *A projection pursuit algorithm for exploratory data analysis*, IEEETrans.Comput.,vol.C-23,pp.881–890. 1979

[7] R.O.Duda, P.E.Hart and D.G.Stork, "*Pattern Classification*, 2nd ed.NewYork: Wiley, 2001".

[8] A.Rosenwald, G.Wright and W.Chan et al., *The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma*, NewEng.J.Med, vol. 346, no.25, pp1937–1947, 2000.

[9] S. N. Mukherjee¤ and S. J. Roberts: *A Theoretical Analysis of Gene Selection*

[10] Isabelle Guyon, Andre Elisseeff: "*An Introduction to Variable and Feature Selection*"

[11] Yuhang Wang, Fillia Makedon, James Ford, Justin Pearlman, "*HykGene: A Hybrid Approach for Selecting Marker Genes for Phenotype Classification using Microarray Gene Expression Data"* Bioinformatics Advance Access published December 7, 2004

[12] Qiuming Zhu, Hongmei Cui, Kajia Cao, and Wing C. Chan "*Algorithmic Fusion of Gene Expression Profiling for Diffuse Large B-Cell Lymphoma Outcome Prediction*"
IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 8, NO. 2, JUNE 2004

[13] M.A.Shipp, K.N.Ross and P.Tamayo et al., *Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning*,
Nature Medicine, vol. 8, no. 1, pp. 68–74, 2002.

[14] Michael P. S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S. Furey, Manuel Ares, Jr., and David Haussler: *Knowledge-based analysis of microarray gene expression data by using support vector machines*

[15] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH.Q: "*Gene expression profiling predicts clinical outcome of breast cancer.*" Nature. 2002 Jan 31, 415(6871):484-5.

[16] Momiao Xiong, Wuju Li, Jinying Zhao, Li Jin and Eric Boerwinkle "*Feature (Gene) Selection in Gene Expression-Based Tumor Classification"* Molecular Genetics and Metabolism **73,** 239–247 (2001)

[17] Alizadeh AA**,** Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM, et al. *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*. Nature. 2000 Feb 3;403(6769):503-11

[18] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*. Science. 1999 Oct 15;286 (5439):531-7

[19] Qiuming Zhu, Senior Member, IEEE, Hongmei Cui, Kajia Cao, and Wing C. Chan *Algorithmic Fusion of Gene Expression Profiling for Diffuse Large B- Cell Lymphoma Outcome Prediction*

[20] M.Brownetal., “Knowledge-based analysis of microarray gene expression data by using support vector machines,” NationalAcademy Sciences, vol. 97, 2000, pp. 262–267

[21] Isabelle Guyon, Jason Weston, Stephen Barnhill, M.D.and Vladimir Vapnik *“Gene Selection for Cancer Classification using Support Vector Machines”*

[22] Sudhir Varma and Richard Simon:
“*Bias in error estimation when using cross-validation for model selection”*
Published: 23-2-2006 BMC Bioinformatics 2006, **7**:91 doi:10.1186/1471-2105-7-91

[23] I.N.Dimou, G.C. Manikis, Michalis E.Zervakis, member IEEE:
“*Classifier Fusion Approaches for Diagnostic Cancer Models*”