



NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

**SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS**

PROGRAM OF POSTGRADUATE STUDIES

PhD THESIS

“Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases”

Argiris D. Sakellariou

ATHENS

JUNE 2015



ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

**"Υπολογιστικές Μέθοδοι για τον Προσδιορισμό Στατιστικώς
Σημαντικών Γονιδίων: Εφαρμογές σε δεδομένα γονιδιακής
έκφρασης από διάφορες ανθρώπινες νόσους"**

Αργύρης Δ. Σακελλαρίου

ΑΘΗΝΑ

ΙΟΥΝΙΟΣ 2015

PhD THESIS

“Computational Methods for the Identification of Statistically Significant Genes:
Applications to Gene Expression Data of Various Human Diseases”

Argiris D. Sakellariou

SUPERVISOR: Sergios Theodoridis, Professor UoA

THREE-MEMBER ADVISORY COMMITTEE:

Sergios Theodoridis, Professor UoA

Dimitris Maroulis, Professor UoA

Emmanouil Sagkriotis, Associate professor UoA

SEVEN-MEMBER EXAMINATION COMMITTEE

(Signature)

(Signature)

**Sergios Theodoridis,
Professor UoA**

**Dimitris Maroulis,
Professor UoA**

(Signature)

(Signature)

**Emmanouil Sagkriotis,
Associate professor UoA**

**Elias Manolakos,
Associate professor UoA**

(Υπογραφή)

(Υπογραφή)

**Despina Sanoudou,
Assistant professor Medical School-
UoA**

**Aristidis Charonis,
Professor BRFAA**

(Signature)

**Kostas Vekrellis,
Associate professor BRFAA**

Examination Date 18/06/2015

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

"Υπολογιστικές Μέθοδοι για τον Προσδιορισμό Στατιστικώς Σημαντικών Γονιδίων:
Εφαρμογές σε δεδομένα γονιδιακής έκφρασης από διάφορες ανθρώπινες νόσους"

Αργύρης Δ. Σακελλαρίου

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Σέργιος Θεοδωρίδης, Καθηγητής ΕΚΠΑ

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ:

Σέργιος Θεοδωρίδης, Καθηγητής ΕΚΠΑ

Δημήτριος Μαρούλης, Καθηγητής ΕΚΠΑ

Εμμανουήλ Σαγκριώτης, Αν. Καθηγητής ΕΚΠΑ

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

(Υπογραφή)

**Σέργιος Θεοδωρίδης
Καθηγητής ΕΚΠΑ**

(Υπογραφή)

**Εμμανουήλ Σαγκριώτης
Αν. Καθηγητής ΕΚΠΑ**

(Υπογραφή)

**Δέσποινα Σανούδου
Επικ. Καθηγήτρια ΙΑΤΡΙΚΗΣ ΣΧΟΛΗΣ-
ΕΚΠΑ**

(Υπογραφή)

**Κωνσταντίνος Βεκρέλλης
Αν. Καθηγητής' ΙΙΒΕΑ**

(Υπογραφή)

**Δημήτριος Μαρούλης
Καθηγητής ΕΚΠΑ**

(Υπογραφή)

**Ηλίας Μανωλάκος
Αν. Καθηγητής ΕΚΠΑ**

(Υπογραφή)

**Αριστείδης Χαρώνης
Καθηγητής ΙΙΒΕΑ**

Ημερομηνία εξέτασης 18/06/2015

ABSTRACT

The advent of microarray technology has revolutionized our knowledge about the underlying mechanisms of human diseases, based on the simultaneous hybridization of thousands of genes. The analysis and mining in this immense amount of information necessitate the development of sophisticated algorithms and effective computational tools. The holy grail of those tools is to discern those maybe tens of genes among tens of thousands of genes that appear to differentiate their expression values systematically between two specific phenotypes. This endeavor is impeded by several factors including the inherent “noise” from the microarray technology and the poly-parametric nature of the diseases, which disguise the hunted patterns of differential gene expression.

Therefore, feature selection methods oriented to microarray gene expression data is a research topic that drawn scientific interest even from the late 90s. So far numerous algorithmic approaches have been proposed trying to identify the significant genes per dataset with the aspiration to be further characterized as marker genes relevant to the inspected disease. Despite many successful applications of such methods in a variety of datasets, no method considered as a gold standard yet regarding the discrimination accuracy, the robustness, the number of significant genes and their biological relevance.

In this dissertation we propose a new hybrid feature selection method (*mAP-KL*) based on the hypothesis that among the statistically significant ranked genes in a gene list, there should be clusters of genes that share similar biological functions related to the investigated disease. Thus, instead of keeping N top ranked genes, it would be more appropriate to define and keep a number of gene cluster exemplars. The *mAP-KL* combines successfully multiple hypothesis testing and affinity propagation clustering algorithm along with the Krzanowski & Lai cluster quality index, to select a small yet informative subset of genes.

We subjected our method across a variety of validation tests on simulated microarray data as well as on real microarray data. Regarding the real microarray data we employed datasets of six neuromuscular diseases and four cancer datasets covering a variety number of samples per phenotype. What is more, we engaged ten other feature selection approaches on the same real microarray data and compared the classification results according to several metrics, for example AUC. In addition to the classification analysis we exploited the produced gene lists from a biological perspective as a further assessment of our method in relation to the other approaches. The overall evaluation results (AUC= 0.86) suggest that *mAP-KL* generates concise yet biologically relevant and informative n -gene expression signatures, which can serve as a valuable discrimination tool for diagnostic and prognostic purposes, by identifying potential disease biomarkers in a broad range of diseases.

SUBJECT AREA: Feature Selection, Computational Intelligence on Genomic Data

KEYWORDS: microarrays, gene expression data, significance analysis, hybrid feature selection, biomarkers

ΠΕΡΙΛΗΨΗ

Η έλευση της τεχνολογίας των μικροσυστοιχιών, βασιζόμενη στην ταυτόχρονη υβριδοποίηση χιλιάδων γονιδίων έφερε επανάσταση στις μέχρι τότε γνώσεις μας σχετικά με τους μηχανισμούς που διέπουν τις ανθρώπινες ασθένειες. Η ανάλυση και η εξόρυξη γνώσης από ένα τέτοιο όγκο πληροφορίας απαιτεί την ανάπτυξη εξελιγμένων αλγορίθμων και αποτελεσματικών υπολογιστικών εργαλείων. Σκοπός αυτών των εργαλείων είναι να διακρίνουν μεταξύ δεκάδων χιλιάδων γονιδίων εκείνες τις δεκάδες ίσως γονιδίων που εμφανίζουν μια συστημική διαφοροποίηση στις τιμές της γονιδιακής τους έκφρασης μεταξύ δύο ή περισσότερων φαινοτύπων. Η προσπάθεια αυτή όμως παρεμποδίζεται από διάφορους παράγοντες όπως, τον εγγενή "θόρυβο" των μικροσυστοιχιών καθώς και από την πολυ-παραμετρική φύση των ασθενειών, τα οποία συγκαλύπτουν τα προς αναζήτηση μοτίβα αυτής της διαφορικής γονιδιακής έκφρασης.

Ως εκ τούτου, η ανάπτυξη μεθόδων επιλογής χαρακτηριστικών (γονιδίων) από δεδομένα γονιδιακής έκφρασης είναι ένα ερευνητικό θέμα που έχει κεντρίσει το επιστημονικό ενδιαφέρον από τα τέλη της δεκαετίας του '90 όταν και πρωτοεμφανίστηκαν. Μέχρι στιγμής πολλές αλγοριθμικές προσεγγίσεις έχουν προταθεί, οι οποίες προσπαθούν να εντοπίσουν τα σημαντικά εκείνα γονίδια, ανά σύνολο δειγμάτων, με τη φιλοδοξία κάποια από αυτά να χαρακτηριστούν ως γονίδια σήμανσης για την εξεταζόμενη νόσο. Παρά τις αρκετές επιτυχημένες εφαρμογές σε μια ποικιλία γονιδιακών δεδομένων έκφρασης, καμία μέθοδος δεν έχει καταφέρει να διακριθεί έναντι των υπολοίπων όσων αφορά την σταθερά υψηλή διαχωριστική ικανότητα, των αριθμό αλλά και την βιολογική σημαντικότητα των επιλεγμένων γονιδίων.

Σε αυτή την διατριβή προτείνουμε μια νέα υβριδική μέθοδο επιλογής γονιδίων (mAP-KL) η οποία βασίζεται στην υπόθεση ότι μεταξύ των στατιστικά σημαντικών γονιδίων σε μια ταξινομημένη λίστα, θα πρέπει να υπάρχουν ομάδες γονιδίων που μοιράζονται παρόμοιες βιολογικές λειτουργίες σε σχέση με την υπό διερεύνηση νόσο. Έτσι, αντί να επιλέγουμε τα N κορυφαία γονίδια μιας λίστας, θα ήταν σκόπιμο να επιλέγουμε τα χαρακτηριστικότερα γονίδια από κάθε ομάδα γονιδίων. Το mAP-KL συνδυάζει επιτυχώς μια μέθοδο πολλαπλού ελέγχου υποθέσεων με μια μέθοδο επιλογής ομάδων γονιδίων και με την χρήση ενός δείκτη ποιότητας συστάδων δεδομένων των Krzanowski & Lai για την τελική επιλογή ενός μικρού αλλά χαρακτηριστικού υποσυνόλου γονιδίων.

Υποβάλλαμε την μεθόδό μας σε μια σειρά δοκιμών αρχικά σε προσομοιωμένα δεδομένα μικροσυστοιχιών και στη συνέχεια σε πραγματικά δεδομένα χρησιμοποιώντας σύνολα δεδομένων από έξι νευρομυϊκές παθήσεις καθώς και από τέσσερις τύπους καρκίνου, καλύπτοντας έτσι ένα ευρύ φάσμα αριθμού δειγμάτων ανά φαινότυπο. Επιπλέον, εφαρμόσαμε δώδεκα άλλες μεθόδους επιλογής χαρακτηριστικών στα ίδια πραγματικά δεδομένα και συγκρίναμε τα αποτελέσματα ταξινόμησης με την χρήση διάφορων μετρικών αξιολόγησης. Επιπροσθέτως, θελήσαμε να ελέγξουμε και να συγκρίνουμε τις παραχθείσες γονιδιακές λίστες της μεθόδου μας αλλά και των άλλων μεθόδων σε σχέση με την βιολογική τους συνάφεια ως προς την εξεταζόμενη νόσο. Τα συνολικά αποτελέσματα των αξιολογήσεων ($AUC = 0.86$) δείχνουν ότι η mAP-KL επιλέγει ένα υποσύνολο από n -γονίδια τα οποία όχι μόνο διαχωρίζουν ικανοποιητικά άγνωστα δείγματα, αλλά είναι και βιολογικώς σχετιζόμενα. Συνεπώς, η mAP-KL μπορεί να αποτελέσει ένα πολύτιμο εργαλείο διαχωρισμού για διαγνωστικούς και θεραπευτικούς σκοπούς, με την ανάδειξη πιθανών βιοδεικτών σε ένα ευρύ φάσμα ασθενειών.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μέθοδοι Επιλογής Χαρακτηριστικών

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μικροσυστοιχίες, δεδομένα γονιδιακής έκφρασης, ανάλυση σημαντικότητας, υβριδική μέθοδος επιλογής χαρακτηριστικών, βιοδείκτες

...στους γονείς μου...

ΕΥΧΑΡΙΣΤΙΕΣ

Η εκπόνηση της παρούσας διδακτορικής διατριβής πραγματοποιήθηκε στο Ίδρυμα Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών (ΙΙΒΕΑΑ) υπό την εποπτεία και την καθοδήγηση του καθηγητή κυρίου Γεώργιου Σπύρου, τον οποίο θέλω να ευχαριστήσω ιδιαίτερα όχι μόνο για την εμπιστοσύνη που επέδειξε στο πρόσωπό μου αλλά και στη αμέριστη υποστήριξη του καθόλη την διάρκεια του ερευνητικού μου έργου.

Επιπλέον θα ήθελα να ευχαριστήσω την Δέσποινα Σανούδου, Επίκουρη καθηγήτρια στην Ιατρική σχολή Αθηνών η καθοδήγηση της οποίας ήταν απόλυτα κρίσιμη για την επιτυχή ολοκλήρωση της έρευνάς μου.

Τέλος, θα ήθελα να εκφράσω τις ιδιαίτερες ευχαριστίες μου στον κύριο Σέργιο Θεοδωρίδη, καθηγητή του τμήματος Πληροφορικής και Τηλεπικοινωνιών, που μου έδωσε την ευκαιρία να εμπλακώ και να γνωρίσω τον υπέροχο κόσμο της έρευνας, γιατί χωρίς την εμπιστοσύνη και αποδοχή του, τίποτα από όλα αυτά δεν θα είχαν γίνει.

ΣΥΝΟΠΤΙΚΗ ΠΑΡΟΥΣΙΑΣΗ ΤΗΣ ΔΙΔΙΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

ΕΙΣΑΓΩΓΗ

Η εμφάνιση της τεχνολογίας των μικροσυστοιχιών DNA έχει βελτιώσει τις δυνατότητές μας ως προς την καλύτερη κατανόηση των μηχανισμών που διέπουν τις ανθρώπινες ασθένειες και έχει βοηθήσει στην ακριβέστερη ταξινόμηση, διάγνωση και πρόγνωση. Λόγω της υψηλής διακίνησης δεδομένων που την χαρακτηρίζουν, είναι απαραίτητη η χρήση υπολογιστικών εργαλείων για την ανάλυση και εξόρυξη των δεδομένων, προκειμένου να βοηθηθούν οι ερευνητές στο να μεγιστοποιήσουν την εξαγόμενη γνώση από τα πειραματικά αποτελέσματα. Στον τομέα της διάγνωσης, οι προερχόμενοι από τις μικροσυστοιχίες βιοδείκτες έχουν εξελιχθεί σε ένα πολύτιμο εργαλείο. Παρόμοια με οποιαδήποτε άλλη κλινική δοκιμή, ο πρωταρχικός στόχος των μοριακών δοκιμών, συμπεριλαμβανομένων των δοκιμών με μικροσυστοιχίες, είναι η παροχή αξιόπιστων και έγκαιρων αποτελεσμάτων για τη βελτίωση της φροντίδας των ασθενών. Προκειμένου να μεγιστοποιηθεί η χρησιμότητα των μικροσυστοιχιών στην διάγνωση / πρόγνωση, είναι σημαντικό να ελαχιστοποιηθεί ο αριθμός των βιοδεικτών που πρέπει να ελεγχθούν ώστε να επιτευχθεί μια ακριβής διάγνωση.

Ωστόσο, η επιλογή αυτών των βιοδεικτών, αποτελεί μια πρόκληση κατά την οποία οι μέθοδοι επιλογής χαρακτηριστικών (FS) μπορούν να συμβάλουν σημαντικά. Πράγματι, από τα τέλη της δεκαετίας του '90 μια πληθώρα μεθόδων εμφανίστηκε και εφαρμόστηκε σε αρκετές μελέτες μικροσυστοιχιών. Παρά τις αλγοριθμικές διαφορές τους, όλες οι μέθοδοι έχουν τους ίδιους στόχους: 1) την αποφυγή της υπερπροσαρμογής και τη βελτίωση της απόδοσης των προβλέψεων 2) να παράγουν γρηγορότερα και αποδοτικότερα μοντέλα, και 3) να προσφέρουν μια βαθύτερη κατανόηση των υποκείμενων διεργασιών. Παρ' όλα αυτά, η επιλογή αυτών των «σημαντικών» γονιδίων που αποδίδουν το ίδιο υψηλό επίπεδο ταξινόμησης σε δεδομένα μιας συγκεκριμένης ασθένειας δεν είναι ακόμα εφικτό και αποτελεί ένα ανοιχτό ζήτημα.

ΣΥΝΑΦΕΙΣ ΕΡΕΥΝΗΤΙΚΕΣ ΠΡΟΣΠΑΘΕΙΕΣ

Στην πραγματικότητα, κάθε σύνολο από δεδομένα μικροσυστοιχιών μπορεί να οδηγήσει σε τόσες λίστες σημαντικών γονιδίων όσες και οι FS μέθοδοι που θα εφαρμοστούν. Ακόμη και στις περιπτώσεις όπου οι μέθοδοι μοιράζονται την ίδια φιλοσοφία οι παραγόμενες γονιδιακές λίστες είναι πιθανόν να αποκλίνουν. Σχετικά με τις FS μεθόδους που μοιράζονται κοινές αρχές, μπορούμε να ορίσουμε τις ακόλουθες ευρείες ομάδες: φιλτραρίσματος, περιτυλίγματος και ενσωματωμένες. Αυτές είναι οι 3 βασικές κατηγορίες, κάθε μία με τα αντίστοιχα πλεονεκτήματα και μειονεκτήματα. Επιπλέον από αυτές τις 3 κατηγορίες, έχει εμφανιστεί και μια νέα κατηγορία FS μεθόδων, οι υβριδικές μέθοδοι. Αυτές οι μέθοδοι συνδυάζουν μεθόδους διαφόρων κατηγοριών με σκοπό την αξιοποίηση των πλεονεκτημάτων τους και παράλληλα την άμβλυνση των μειονεκτημάτων τους προς όφελος της επιλογής «σημαντικών» γονιδίων.

Βεβαίως, ο συνδυασμός μεθόδων αποτελεί μια εποικοδομητική διαδικασία που βασίζεται πάντα σε επιστημονικές υποθέσεις, είτε βιολογικές είτε στατιστικές, και όχι σε κάποια τυχαία επιλογή. Επί παραδείγματι, στην μελέτη των Jaeger et al. ισχυρίστηκαν ότι οι αλγόριθμοι ταξινόμησης/φιλτραρίσματος παράγουν λίστες γονιδίων, όπου τα κορυφαία γονίδια έχουν υψηλή συσχέτιση μεταξύ τους, κυρίως επειδή ανήκουν στο ίδιο βιολογικό μονοπάτι. Επίσης, ο Hall στη διατριβή του, διερεύνησε την υπόθεση ότι "ένα καλό υποσύνολο χαρακτηριστικών (γονιδίων) είναι αυτό που περιέχει χαρακτηριστικά

υψηλής συσχέτισης με την κατηγορία, αλλά χαμηλής συσχέτισης μεταξύ τους». Αυτές οι πεπιοθήσεις έδωσαν το έναυσμα για πολλές υβριδικές μεθόδους, κάποιες από τις οποίες συνδύαζαν μια μέθοδο φιλτραρίσματος και μια μέθοδο ομαδοποίησης για να καταλήξουν σε μια λίστα «σημαντικών» γονιδίων.

Ειδικότερα, οι Jaeger et al. χρησιμοποίησαν έναν fuzzy αλγόριθμο ομαδοποίησης για να διαχωρίσουν αρχικά τα γονίδια, ομαδοποιώντας τα σύμφωνα με ένα μέτρο ομοιότητας. Στη συνέχεια, με τη βοήθεια ενός στατιστικού ελέγχου, όπως το t-test ή το Wilcoxon test, επέλεξαν ένα ή περισσότερα αντιπροσωπευτικά γονίδια από κάθε ομάδα για τον σχηματισμό λίστας «σημαντικών» αλλά και ασυσχέιστων μεταξύ τους γονιδίων. Σε αυτή τη μελέτη, ο ακριβής αριθμός των συστάδων που θα πρέπει να σχηματιστεί και ο αριθμός των αντιπροσωπευτικών γονιδίων ανά συστάδα παρέμειναν προβλήματα προς λύση.

Παρόμοια με τους Jaeger et al., στη μελέτη των Hanczar et al. προτάθηκε μια μέθοδος δύο βημάτων. Συγκεκριμένα, μια μη-επιτοπτευόμενη μέθοδος ομαδοποίησης, k-mean, συνδυάστηκε με μια μαθηματική έννοια, του «πρωτότυπου γονιδίου», προσπαθώντας να εντοπίσει τα αντιπροσωπευτικά γονίδια της κάθε ομάδας. Ανάλογα προβλήματα με αυτά της μελέτης των Jaeger et al. εμφανίστηκαν και σε αυτήν τη μελέτη, και τα οποία χαρακτηρίστηκαν ως μελλοντικοί στόχοι από τους ερευνητές. Μια εναλλακτική αλγοριθμική προσέγγιση, όπου η κατάταξη των γονιδίων προηγείται κάθε άλλης μεθόδου περιγράφεται στην mRMR μέθοδο. Συγκεκριμένα, η αρχική κατάταξη μέσω t-test ή F-test συνδυάζεται στη συνέχεια με μια διαδοχική και επαναλαμβανόμενη σύγκριση μεταξύ των ταξινομημένων ζευγών των γονιδίων, προκειμένου να καταλήξει σε ένα υποσύνολο «σημαντικών» γονιδίων, σύμφωνα με κάποια κριτήρια, όπως τη μέγιστη συνάφεια και τον ελάχιστο πλεονασμό. Ένα σημαντικό μειονέκτημα αυτής της προσέγγισης είναι ότι το κριτήριο του πλεονασμού μπορεί να αποκλείσει γονίδια που θεωρούνται σημαντικά από βιολογικής άποψης. Μια άλλη ενδιαφέρουσα προσέγγιση, η HykGene, είναι μια μέθοδος επιλογής γονιδίων τριών βημάτων, η οποία ενσωματώνει έναν αλγόριθμο φιλτραρίσματος, μια ιεραρχική μέθοδο ομαδοποίησης των κορυφαίων ταξινομημένων γονιδίων και, τέλος, έναν αλγόριθμο σάρωσης γραμμής. Αφού πρώτα προσδιοριστούν οι συστάδες από την ιεραρχική μέθοδο, ο αλγόριθμος σάρωσης γραμμής εφαρμόζεται στο δένδρογραμμα προκειμένου να επιλέξει ένα αντιπροσωπευτικό γονίδιο ανά συστάδα.

Λαμβάνοντας υπόψη από τη μια τα επιτυχή αποτελέσματα κατάταξης των παραπάνω μελετών, και από την άλλη τους περιορισμούς/προβλήματα των μεθόδων αυτών, αναπτύξαμε μια νέα υβριδική μέθοδο, την mAP-KL. Στην προτεινόμενη προσέγγιση, τα γονίδια πρώτα κατατάσσονται ανάλογα με την διαφορική τους έκφραση, χρησιμοποιώντας ένα t-test πολλαπλών υποθέσεων, και στη συνέχεια, τα κορυφαία N ταξινομημένα γονίδια ομαδοποιούνται με τη μέθοδο συσταδοποίησης Affinity Propagation (AP). Πριν από την AP εφαρμόζεται ένας αλγόριθμος αναγνώρισης του αριθμού των συστάδων μεταξύ των κορυφαίων- N -γονιδίων. Το αποτέλεσμα αυτής της μεθόδου είναι ένα υποσύνολο που περιλαμβάνει ένα αντιπροσωπευτικό γονίδιο ανά συστάδα.

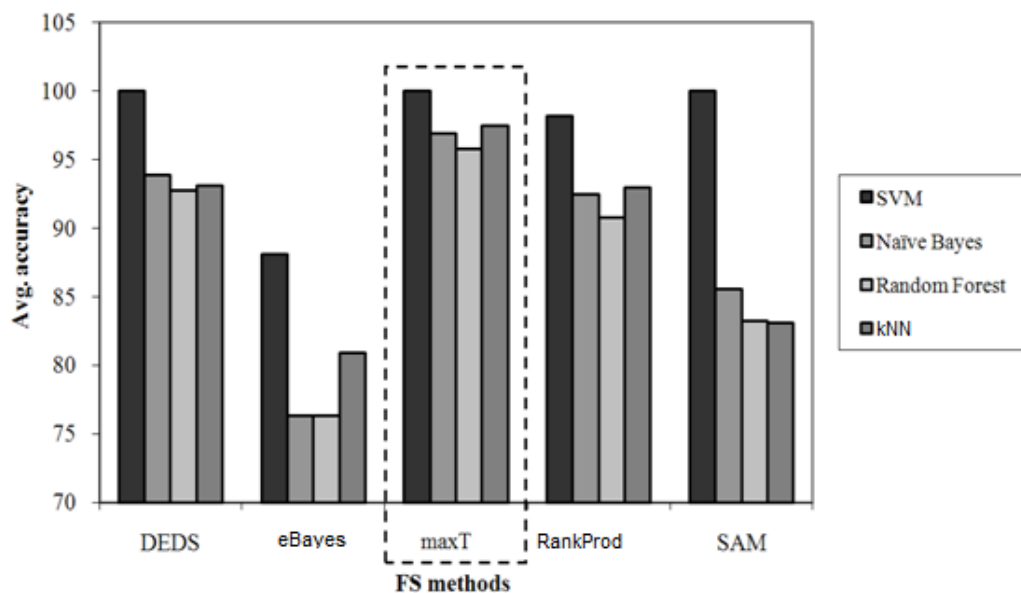
Η ΠΡΟΤΕΙΝΟΜΕΝΗ ΥΒΡΙΔΙΚΗ ΜΕΘΟΔΟΣ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ (mAP-KL)

Μια FS μέθοδος, για δεδομένα γονιδιακής έκφρασης από μικροσυστοιχίες, θα πρέπει να είναι ανεξάρτητη από τον τύπο της πλατφόρμας, από την νόσο και από το μέγεθος του συνόλου των δεδομένων. Η υπόθεσή μας είναι ότι μεταξύ των στατιστικά σημαντικών γονιδίων μιας ταξινομημένης λίστας, θα πρέπει να υπάρχουν ομάδες

γονιδίων που μοιράζονται παρόμοιες βιολογικές λειτουργίες σε σχέση με την υπό διερεύνηση ασθένεια. Έτσι, αντί να κρατάμε τα κορυφαία N γονίδια μιας ταξινομημένης λίστας, θα ήταν σκόπιμο να προσδιορίζουμε και να κρατάμε έναν αριθμό αντιπροσωπευτικών γονιδίων ανά συστάδα. Γι' αυτό το σκοπό προτείνουμε μια υβριδική μέθοδο FS (mAP-KL), η οποία συνδυάζει μια μέθοδο φιλτραρίσματος μέσω της εφαρμογής πολλαπλού ελέγχου υποθέσεων, έναν αλγόριθμο συσταδοποίησης και έναν δείκτη ποιότητας συστάδων, προκειμένου να επιλέξουμε ένα μικρό αλλά αντιπροσωπευτικό υποσύνολο γονιδίων.

Η μέθοδος φιλτραρίσματος

Η προτεινόμενη μεθοδολογία συνδυάζει κατάταξη/φιλτράρισμα και ομαδοποίηση για να επιλεγεί ένα μικρό σύνολο γονιδίων μη συσχετισμένων μεταξύ τους αλλά συσχετισμένο με την υπό διερεύνηση ασθένεια. Σε σχέση με το στάδιο του φιλτραρίσματος, αρχικά χρησιμοποιείται η συνάρτηση $\max T$ για να ταξινομήσει τα γονίδια του συνόλου εκπαίδευσης. Η απόφασή μας για το ποιά μέθοδο φιλτραρίσματος θα εφαρμόσουμε προέκυψε από τα συμπεράσματα μιας μελέτης που πραγματοποιήσαμε επάνω σε FS μεθόδους. Συγκεκριμένα, αξιολογήσαμε την απόδοση ταξινόμησης πέντε διαφορετικών FS μεθόδων σε δεδομένα από δέκα διαφορετικές νευρομυϊοπάθειες. Κάθε μέθοδος έδωσε μια διαφορετική λίστα γονιδίων, από την οποία στη συνέχεια χρησιμοποιήθηκαν από πάνω προς τα κάτω τα ταξινομημένα γονίδια από την τιμή κατάταξης 2 έως την θέση 400 με μοναδιαίο βήμα, προκειμένου κάθε φορά να συνθέτουμε ένα νέο σύνολο ταξινόμησης. Η αξιολόγηση των επιδόσεων κατάταξης όλων των συνόλων ταξινόμησης ανά FS μέθοδο απεικονίζεται στην Εικόνα 1, και δείχνει ότι η $\max T$ πέτυχε μέση ακρίβεια ταξινόμησης ίση με 95%, ανάμεσα σε υγιή δείγματα και δείγματα με την υποδιερεύνηση νόσο.



Εικόνα 1: Η συνολική ακρίβεια ταξινόμησης πέντε μεθόδων επιλογής χαρακτηριστικών σε δέκα σύνολα δεδομένων από νευρομυϊοπάθειες σύμφωνα με τέσσερις αλγόριθμους ταξινόμησης

Ο δείκτης ποιότητας συστάδων

Στη συνέχεια, και πριν το στάδιο της ομαδοποίησης, επιχειρούμε να προσδιορίσουμε τον αριθμό των συστάδων, που στην ουσία θα είναι και ο αριθμός των

αντιπροσωπευτικών γονιδίων που θα αποτελέσουν το υποσύνολο μας. Η απόφαση σχετικά με το ποιόν δείκτη ποιότητας συστάδων θα χρησιμοποιήσουμε, βασίστηκε τόσο στα αποτελέσματα μιας συγκριτικής μελέτης των Tibshirani et al. όσο και σε ένα πλήθος δοκιμών που εκτελέσαμε σε προσομοιωμένα δεδομένα ομαδοποίησης οι οποίες επίσης κατέδειξαν την αποτελεσματικότητα των δεικτών. Σύμφωνα με τα προηγούμενα, καταλήξαμε στην εφαρμογή του δείκτη των Krzanowski και Lai προκειμένου να προσδιορίσουμε τον αριθμό των συστάδων στα δείγματα της νόσου του συνόλου εκπαίδευσης. Η εφαρμογή του μόνο στα δείγματα της νόσου αποτελεί μια σημαντική λεπτομέρεια της μεθοδολογίας μας, δεδομένου ότι έχει άμεσο στο πλήθος των συστάδων που θα αναγνωριστούν και κατά συνέπεια στο πλήθος των επιλεγθέντων γονιδίων.

Ωστόσο, αρχικά αντιμετωπίσαμε το δίλημμα σε ποιό τμήμα των δεδομένων θα ήταν ορθότερο να εφαρμόσουμε τον δείκτη ποιότητας συστάδων. Η μία επιλογή ήταν να ψάξουμε για την δομή των συστάδων αποκλειστικά στα δείγματα που ανήκουν στο φαινότυπο ελέγχου ή υγιή δείγματα, ενώ η επόμενη εναλλακτική ήταν να διερευνήσουμε τη δομή των συστάδων στα δείγματα με την πάθηση. Εν τέλει καταλήξαμε ότι αυτό που πραγματικά έχει σημασία για τον προσδιορισμό των σημαντικών γονιδίων ως προς μια ασθένεια, είναι το τμήμα των δεδομένων που σχετίζονται με τη νόσο, διότι όλες οι πληροφορίες σχετικά με το «έναυσμα» των μοριακών διαδικασιών είναι σίγουρα παρούσες σε αυτό το υποσύνολο.

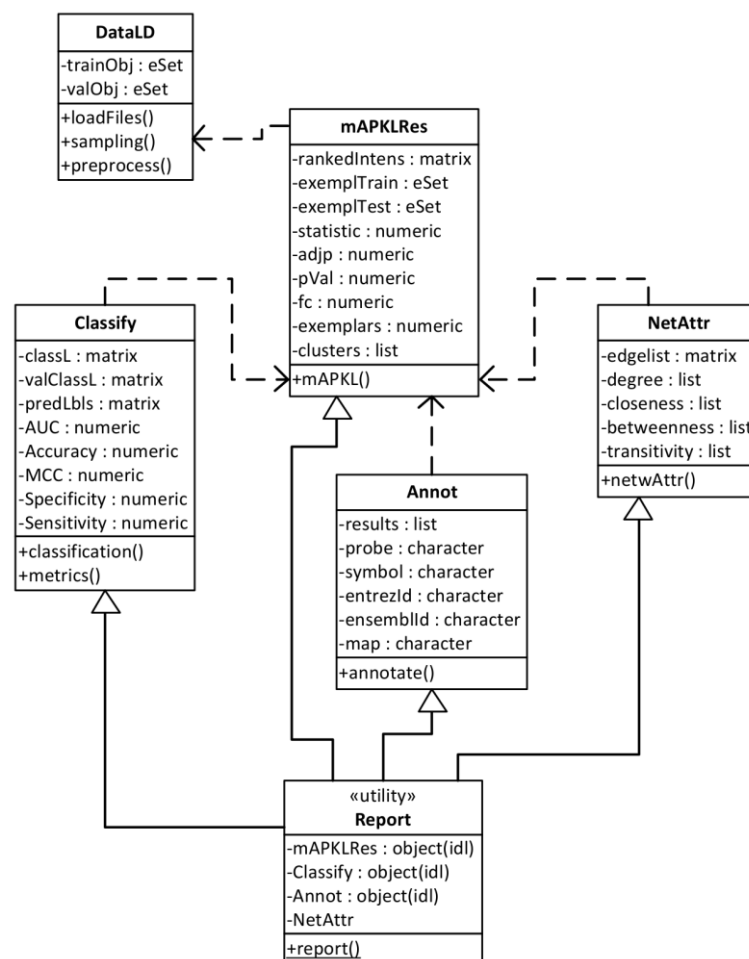
Ο αλγόριθμος συσταδοποίησης Affinity Propagation

Το τελικό βήμα της μεθοδολογίας μας, περιλαμβάνει την ανάλυση συστάδων με την AP μέθοδο. Ο αλγόριθμος της AP εμφανίστηκε στα τέλη της δεκαετίας του 2000 και σύμφωνα με τα αποτελέσματα μιας εκτεταμένης μελέτης με 15 άλλους αλγόριθμους ομαδοποίησης, π.χ. *k*-medians clustering, hierarchical agglomerative clustering κλπ, διακρίθηκε πετυχαίνοντας τα πιο ακριβεί αποτελέσματα ομαδοποίησης. Η εγγενής θεώρηση της μεθόδου, ότι αρχικά όλα τα σημεία δεδομένων (γονίδια) θεωρούνται ως πιθανά «exemplars», καθώς και η αποτελεσματική της ομαδοποίησης, μας προέτρεψαν στο να ενσωματώσουμε την AP ως ένα αναπόσπαστο μέρος της μεθοδολογίας μας. Συνεπώς, ορίζουμε στην AP τον αριθμό των k συστάδων σύμφωνα με το αποτέλεσμα του δείκτη των Krzanowski και Lai και στη συνέχεια αφήνουμε την AP να ανίχνευση αυτές τις k συστάδες μεταξύ των κορυφαίων N γονιδίων (όπου N ένας προκαθορισμένος αριθμός). Ο αλγόριθμος τις περισσότερες φορές συγκλίνει στον ζητούμενο αριθμό συστάδων και μας παρέχει μια λίστα με τα πιο αντιπροσωπευτικά γονίδια ανά συστάδα, τα επονομαζόμενα «exemplars». Αυτά τα n «exemplars» αναμένεται να σχηματίσουν ένα ταξινομητή ο οποίος θα διαχωρίζει επιτυχώς τα δείγματα μεταξύ δύο κλάσεων σε ένα σύνολο δοκιμής. Αφού έχουμε διαθέσιμα τα «exemplars» τα χρησιμοποιούμε για να διαμορφώσουμε τα αντίστοιχα σύνολα εκπαίδευσης και, και στη συνέχεια προχωράμε στη διαδικασία ταξινόμησης.

Η ανάπτυξη της mAP-KL σε ένα R-πακέτο

Προκειμένου να δώσουμε τη δυνατότητα στην επιστημονική κοινότητα να εφαρμόσει την μέθοδό μας σε οποιοδήποτε σύνολο δεδομένων γονιδιακής έκφρασης, αναπτύξαμε την μεθοδολογία μας σε ένα R πακέτο, ανοιχτού κώδικα, το mAPKL το οποίο φιλοξενείται στην διεθνή πλατφόρμα Bioconductor. Στο πακέτο αυτό συμπεριλάβαμε και επιπλέον λειτουργίες όπως της δειγματοληψίας (δημιουργία συνόλων εκπαίδευσης και ελέγχου) προεπεξεργασίας, ταξινόμησης, ανάλυσης δικτύων, γονιδιακών πληροφοριών, ανάλυση βιολογικών μονοπατιών καθώς και την παραγωγή έκθεσης με τα αποτελέσματα των προηγούμενων αναλύσεων. Όλες αυτές οι λειτουργίες

υποστηρίζονται από πέντε διακριτές κλάσεις, Εικόνα 2. Η κεντρική ιδέα κατά τη διάρκεια σχεδιασμού του πακέτου ήταν να ενσωματώσουμε λειτουργίες που να μπορούν είτε να οδηγήσουν σε μια εκτενή ανάλυση είτε να χρησιμοποιηθούν αυτόνομα. Επί παραδείγματι, ένας χρήστης μπορεί να εισάγει ένα οποιοδήποτε σύνολο δεδομένων γονιδιακής έκφρασης και να εκτελέσει με μία μόνο εντολή μέχρι και οκτώ διαφορετικές μεθόδους προεπεξεργασίας. Στη συνέχεια, μπορεί να αναλύσει τα προεπεξεργασμένα δεδομένα με τη μέθοδο *mAP-KL* και να παράξει λίστες σημαντικών γονιδίων. Ο χρήστης μπορεί επίσης να εκτελέσει ταξινόμηση δειγμάτων, εξόρυξη στοιχείων των γονιδίων, ανάλυση βιολογικών μονοπατιών και χαρακτηριστικών δικτύου. Από την άλλη πλευρά, ένας χρήστης μπορεί επίσης να χρησιμοποιήσει οποιαδήποτε από τις προηγούμενες λειτουργίες αυτόνομα όπως για παράδειγμα, τη συνάρτηση της δειγματοληψίας για να δημιουργήσει σύνολα εκπαίδευσης και αξιολόγησης.



Εικόνα 2: Η UML σχηματική αναπαράσταση των κλάσεων και των συναρτήσεων στο mAPKL

ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΧΟΛΙΑΣΜΟΣ

Υποβάλλαμε τη μέθοδό μας σε μια σειρά δοκιμών αρχικά σε δεδομένα προσομοίωσης και στη συνέχεια σε πραγματικά δεδομένα. Όσον αφορά τα πραγματικά δεδομένα, χρησιμοποιήσαμε σύνολα δεδομένων από έξι νευρομυοπάθειες ως εκπροσώπους τού μικρού πληθυσμού δειγμάτων ανά φαινότυπο και δεδομένα από τέσσερις τύπους καρκίνου ως εκπροσώπους τού μεγάλου πληθυσμού δειγμάτων ανά φαινότυπο. Σχεδιάσαμε και εκτελέσαμε ένα πλήρες σύνολο δοκιμών χρησιμοποιώντας 5 Cross-validation στο σύνολο εκπαίδευσης και στη συνέχεια Hold-out validation σε ανεξάρτητο σύνολο ελέγχου χρησιμοποιώντας τρεις διαφορετικούς ταξινομητές, RF - SVM - KNN. Σκοπός μας ήταν να αξιολογήσουμε τις επιδόσεις της μεθόδου μας τόσο σε μικρά όσο

και σε μεγάλα σύνολα δειγμάτων καθώς και την σταθερότητα της απόδοσης σε σχέση με τους ταξινομητές. Επιπλέον, στα ίδια σύνολα δεδομένων εφαρμόσαμε 12 άλλες FS μεθόδους και συγκρίναμε τα αποτελέσματα ταξινόμησης χρησιμοποιώντας 3 μετρικές απόδοσης, όπως τις AUC, TNR, TPR. Σε σχέση με τις FS μεθόδους, χρησιμοποιήσαμε έξι μονοπαραγοντικές μεθόδους φιλτραρίσματος (eBayes, ODP, maxT, η SAM, SNR και t-δοκιμής), μία πολυπαραγοντική μέθοδο φιλτραρίσματος (cat) , τρεις μεθόδους μείωσης των διαστάσεων (BGA-COA, η PCA, PLS-CV), μία ενσωματωμένη μέθοδο (Random Forest), και μία υβριδική μέθοδο (HykGene).

Η αποδοτικότητα μιας FS μεθόδου προκύπτει όχι μόνο από την απόδοση της κατά την ταξινόμηση, αλλά και από τη βιολογική συνάφεια της λίστας των γονιδίων της με τους αντίστοιχους φαινοτύπους. Γι'αυτό κι εμείς επιπλέον του ελέγχου απόδοσης των μεθόδων κατά την ταξινόμηση, ελέγξαμε επίσης και την βιολογική συνάφεια των λιστών ως προς την εξεταζόμενη νόσο. Συγκεκριμένα, αξιολογήσαμε τις παραγόμενες λίστες γονιδίων από την mAP-KL, τις λίστες των μεθόδων που διακρίθηκαν κατά την ταξινόμηση, (eBayes, PLS-CV, SAM, BGA-COA, RF-MDA), καθώς επίσης και τις λίστες από την μέθοδο maxT η οποία είναι η μέθοδος κατάταξης της mAP-KL. Κατά τη διάρκεια αυτών των αξιολογήσεων, προσπαθήσαμε να φωτίσουμε τη «σημασιολογία» πίσω από αυτές τις λίστες γονιδίων καθώς και τη σχέση τους με τις αντίστοιχες ασθένειες.

Τα αποτελέσματα ταξινόμησης σε πραγματικά δεδομένα

Τα συνολικά αποτελέσματα, βάση του ταξινομητή RF, όπως συνοψίζονται στην Εικόνα 3 τοποθετούν την mAP-KL στις κορυφαίες μεθόδους μεταξύ 12 άλλων FS αλγόριθμων. Ειδικότερα, η μέθοδος mAP-KL πέτυχε τη δεύτερη καλύτερη μέση AUC στις νευρομυοπάθειες, και συγκεκριμένα 0.91, ενώ συνολικά σε όλες τις δέκα ασθένειες πέτυχε μέση τιμή AUC ίση με 0.86, το οποίο αποτελεί την τρίτη καλύτερη επίδοση έχοντας μάλιστα και την μικρότερη τιμή τυπικής απόκλισης σε σχέση με μεθόδους που πέτυχαν καλύτερες επιδόσεις, π.χ. την eBayes, την PLS-CV. Σύμφωνα με τα αποτελέσματα ταξινόμησης, μπορούμε να υποστηρίξουμε ότι ο συνδυασμός μιας μονοπαραγοντικής μεθόδου φιλτραρίσματος και μίας μεθόδου συσταδοποίησης οδηγεί στην επιλογή υποσυνόλων γονιδίων υψηλής διαχωριστικής ικανότητας σε άγνωστα δείγματα ανεξαρτήτως ασθένειας και αριθμού δειγμάτων.

Η βιολογική συνάφεια των επιλεγθέντων γονιδίων

Συνήθως, το αρχικό προϊόν μιας FS μεθόδου είναι ένας κατάλογος από probe ids, αντί συμβόλων γονιδίων, μιας και τα δεδομένα γονιδιακής έκφρασης προέρχονται από microarray chips. Ως εκ τούτου, μια απαραίτητη ενέργεια που εκτελούμε συνήθως είναι να ταιριάξουμε τα probe ids με τα αντίστοιχα σύμβολα γονιδίων. Ένα ενδιαφέρον χαρακτηριστικό της τεχνολογίας των chips είναι ότι ένα γονίδιο (σύμβολο) είναι πολύ πιθανό να αντιπροσωπεύεται από περισσότερα του ενός probe ids. Έτσι, ένα σημαντικό υπερεκφρασμένο ή υποεκφρασμένο γονίδιο μπορεί να είναι παρόν σε μια ταξινομημένη λίστα περισσότερες από μία φορές. Ως αποτέλεσμα, αυτές οι πολλαπλές εμφανίσεις ενός γονιδίου πρέπει να αφαιρούνται από τις λίστες με τα κορυφαία γονίδια προκειμένου να καταλήξουμε σε λίστες με μοναδικά γονίδια. Αυτό είναι ένα σημαντικό βήμα όσον αφορά τον επικείμενο έλεγχο εμπλουτισμού, μιας και σε μια ταξινομημένη λίστα 20 ή 50 probe ids μπορεί για παράδειγμα τα μοναδικά αντιπροσωπευτικά γονίδια να είναι αντίστοιχα 14 ή 35. Επιπλέον, τα chips περιλαμβάνουν και κάποια probe ids υπεύθυνα για τον έλεγχο της ποιότητας του υβριδισμού που δεν θα πρέπει να συμπεριλαμβάνονται στην κορυφή ταξινομημένων λιστών οποιασδήποτε διαφορικής ανάλυσης. Για όλους αυτούς τους λόγους, ο βαθμός μοναδικότητας 'degree of

uniqueness' (DoU) μιας ταξινομημένης λίστας αποτελεί ένα πρώτο μέτρο αξιολόγησης από βιολογικής πλευράς του ενδυνάμει εμπλουτισμού μιας λίστας.

	eBayes	PLS-CV	SAM	BGA-COA	RF-MDA	mAP-KL	cat	Hyk Gene	maxT	ODP	SNR	t-test	PCA	MEAN
Diseases with Small Sample Size available	ALS	1.00	1.00	1.00	1.00	1.00	1.00	0.64	1.00	1.00	1.00	1.00	1.00	1.00
	DMD	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.61	0.97
	JDM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LGMD2A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.94	1.00	0.94	1.00	0.58
	LGMD2B	0.48	1.00	0.52	0.98	1.00	0.70	0.36	0.82	0.91	0.73	0.88	0.82	0.21
	NM	-	0.42	0.65	0.47	0.22	0.74	0.78	0.88	0.37	0.25	0.90	0.89	0.55
	MEAN	0.90	0.90	0.86	0.91	0.87	0.91	0.86	0.88	0.87	0.83	0.95	0.95	0.66
Diseases with Large Sample Size available	BREAST	-	0.82	0.77	0.76	0.82	0.87	0.75	0.76	0.77	0.74	0.77	0.73	0.75
	COLON	0.80	0.79	0.80	0.87	0.81	0.89	0.80	0.81	0.79	0.82	0.79	0.79	0.83
	LEUKEMIA	1.00	0.99	0.99	1.00	0.99	0.71	0.99	0.97	0.96	-	0.50	0.50	0.64
	PROSTATE	0.86	0.87	0.92	0.73	0.83	0.80	-	0.69	0.50	-	0.50	0.50	0.50
	MEAN	0.89	0.87	0.87	0.84	0.86	0.82	0.85	0.81	0.76	0.78	0.64	0.63	0.68
TOTAL MEAN	0.89	0.89	0.87	0.87	0.87	0.86	0.85	0.84	0.81	0.81	0.80	0.79	0.67	
TOTAL STD	0.184	0.185	0.173	0.179	0.242	0.127	0.214	0.129	0.223	0.259	0.191	0.197	0.240	

< 0.50	0.50-0.69	0.70-0.79	0.80-0.89	0.90-0.95	0.96-0.99	1.00
--------	-----------	-----------	-----------	-----------	-----------	------

Εικόνα 3: Τα συνολικά αποτελέσματα ταξινόμησης (AUC) για τον RF ταξινομητή

Στους ακόλουθους πίνακες, έχουμε παραθέσει τον αριθμό των probe ids και τον αντίστοιχο αριθμό των γονιδιακών συμβόλων τους ανά μέθοδο και σύνολο δεδομένων. Στην τελευταία στήλη, έχουμε υπολογίσει την τιμή DoU ως τον μέσο όρο του πηλίκου των συμβόλων των γονιδίων προς τα probe ids. Όσο πιο κοντά στην μονάδα είναι η τιμή DoU τόσο πιο μοναδική είναι η λίστα κατάταξης. Όσον αφορά τα δεδομένα από νευρομυοπάθειες, Πίνακας 1, η mAP-KL πέτυχε την υψηλότερη βαθμολογία με την μέθοδο maxT να είναι αρκετά κοντά. Σε σχέση με τα δεδομένα από καρκίνους, Πίνακας 2, η μέθοδος eBayes ξεπέρασε τις άλλες μεθόδους, αν και η τιμή της βασίζεται σε τρία και όχι σε τέσσερα σύνολα δεδομένων. Η mAP-KL κατέλαβε τη δεύτερη θέση δείχνοντας την υψηλή «μοναδικότητα» που χαρακτηρίζει τις παραγόμενες λίστες της.

Ένα δεύτερο μέτρο αξιολόγησης είναι η συνάφεια των μοναδικών γονιδίων με τα σχετικά προς την εξεταζόμενη πάθηση βιολογικά μονοπάτια. Σε αυτό το σημείο είναι σημαντικό να αναφερθούμε σε μια άλλη παράμετρο προτού περιγράψουμε τα αποτελέσματα αυτού του ελέγχου, η οποία αφορά τα γονίδια που κωδικοποιούν πρωτεΐνες (P-C-Gns) από μια λίστα κατάταξης. Στην ουσία, όλα τα γνωστά γονίδια δεν κωδικοποιούν πρωτεΐνες και συνεπώς δεν εμπλέκονται στην μοριακή λειτουργία. Η ανάλυση βιολογικών μονοπατιών προσπαθεί να απλοποιήσει την πολυπλοκότητα στο κυτταρικό επίπεδο μέσω της αντιπροσώπευσης μιας σειράς βημάτων όπου «το κάθε βήμα είναι ένα γεγονός που μετατρέπει εισερχόμενες φυσιολογικές οντότητες σε

εξερχόμενες οντότητες". Τέτοιες οντότητες μεταξύ άλλων παραγόμενων μικρών μορίων ή σωματιδίων είναι σίγουρα οι παραγόμενες πρωτεΐνες, , και ως εκ τούτου μόνο τα γονίδια που κωδικοποιούν πρωτεΐνες είναι απαραίτητα για την ανάλυση μονοπατιών.

Πίνακας 1: Η τιμή DoU επτά FS μεθόδων σε δεδομένα από νευρομυοπάθειες

FS	ALS		DMD		JDM		LGMD2A		LGMD2B		NM		DoU
	Prbs	Gns	Prbs	Gns	Prbs	Gns	Prbs	Gns	Prbs	Gns	Prbs	Gns	
mAP-KL	21	20	14	14	21	20	6	6	15	15	18	18	0.984
maxT	20	20	20	20	20	20	20	20	20	20	20	18	0.983
RF-MDA	20	20	20	20	20	20	20	19	20	20	20	18	0.975
SAM	20	14	20	20	20	18	20	16	20	16	20	20	0.867
eBayes ¹	20	17	20	20	20	18	20	16	20	15	-	-	0.860
PLS-CV	20	13	20	20	20	19	20	18	20	16	20	17	0.858
BGA-COA	20	15	20	17	20	18	20	14	20	17	20	17	0.817

¹ Η eBayes αξιολογήθηκε σε πέντε σύνολα

Πίνακας 2: Η τιμή DoU επτά FS μεθόδων σε δεδομένα από καρκίνους

FS	Breast		Colon		Leukemia		Prostate		DoU
	Prbs	Gns	Prbs	Gns	Prbs	Gns	Prbs	Gns	
eBayes ¹	-	-	20	18	20	18	20	19	0.917
mAP-KL	6	4	20	16	5	5	12	12	0.867
PLS-CV	20	14	20	18	20	19	20	17	0.850
BGA-COA	20	12	20	18	20	19	20	18	0.838
SAM	20	11	20	18	20	18	20	19	0.825
maxT	20	11	20	16	20	17	20	20	0.800
RF-MDA	20	9	20	14	20	18	20	19	0.750

¹ Η eBayes αξιολογήθηκε σε τρία σύνολα

Μέσα από μια πληθώρα εργαλείων ανάλυσης βιολογικών μονοπατιών, επιλέξαμε τη βάση δεδομένων «Reactome», η οποία περιλαμβάνει επιμελημένη και αξιολογημένη πληροφορία για τα βιολογικά μονοπάτια και τις επιδράσεις στην ανθρώπινη βιολογία. Ελέγξαμε τις κορυφαίες λίστες επιλεγμένων FS μεθόδων για όλες τις ασθένειες και αξιολογήσαμε τον εμπλουτισμό τους. Κατά την αξιολόγηση, λάβαμε υπόψη την τιμή DoU, τον αριθμό των γονιδίων που κωδικοποιούν πρωτεΐνες καθώς και το πλήθος των βιολογικών μονοπατιών σύμφωνα με τα αποτελέσματα της βάσης δεδομένων της «Reactome». Το τελικό σκορ εμπλουτισμού μονοπατιών (PE) για κάθε FS (m) είναι ο μέσος όρος των αθροισμάτων των πηλίκων των P-C-Gns προς το πλήθος των μονοπατιών, πολλαπλασιαζόμενο με την τιμή DoU για όλες τις ασθένειες (d)

$$PE_m = \sum_{d=1}^{10} \frac{Protein-coding-genes_d}{Pathways_d} \times DoU.$$

Συνοψίσαμε τα αποτελέσματα, Πίνακας 3, κατατάσσοντας τις FS μεθόδους σε φθίνουσα σειρά με βάση τον μέσο όρο της PE τιμής τους και σύμφωνα με την ανάλυση μονοπατιών, η μέθοδος maxT φαίνεται να επιτύγγάνει την υψηλότερη βαθμολογία PE σε όλες τις ασθένειες. Άλλωστε είναι και η μέθοδος με την δεύτερη υψηλότερη τιμή DoU οριακά πίσω από την mAP-KL. Ωστόσο, αυτό το σημαντικό αποτέλεσμα της maxT έναντι των mAP-KL και RF-MDA που ακολουθούν, φαίνεται να οφείλεται κυρίως στο αξιοσημείο PE σκορ που πέτυχε η maxT στον καρκίνο του προστάτη (4.33), όπου προσδιόρισε τρία (3) μονοπάτια με 13 μοναδικά γονίδια. Όπως και να έχει, αυτές οι τρεις μέθοδοι φαίνεται να συγκροτούν μια ομάδα μεθόδων με PE σκορ κοντά στην μονάδα, το οποίο όχι μόνο είναι ικανοποιητικό, αλλά και ενδιαφέρον για περαιτέρω ανάλυση από βιολόγους.

Πίνακας 3: Τα συνολικά αποτελέσματα ανάλυσης βιολογικών μονοπατιών

FS	Pathway Analysis											Mean	Stdev
	ALS	DMD	JDM	LGMD2A	LGMD2B	NM	Breast	Colon	Leukemia	Prostate			
maxT	1.00	1.08	1.08	0.43	1.36	1.01	0.47	0.80	0.79	4.33	1.24	1.12	
mAP-KL	1.43	0.78	1.38	0.43	0.88	1.40	0.67	0.63	0.80	1.17	0.95	0.36	
RF-MDA	0.75	1.10	1.40	0.74	0.63	1.80	0.54	0.63	0.80	1.03	0.94	0.40	
eBayes ¹	0.37	1.50	0.90	0.64	0.67	-	-	1.08	1.26	0.86	0.91	0.36	
PLS-CV	0.37	0.89	1.21	0.66	0.90	0.85	0.98	0.90	1.07	1.04	0.89	0.23	
SAM	0.29	1.13	1.00	0.64	0.80	1.08	0.46	1.15	0.98	1.27	0.88	0.32	
BGA-COA	0.68	1.06	0.63	0.70	1.19	0.85	0.60	0.90	1.14	1.00	0.87	0.22	

¹ Η eBayes αξιολογήθηκε σε οκτώ σύνολα

ΣΥΜΠΕΡΑΣΜΑΤΑ

Προτείνουμε μια υβριδική FS μέθοδο (mAP-KL), η οποία καταδεικνύει με σαφήνεια πόσο αποτελεσματικός είναι ο συνδυασμός μιας μεθόδου ελέγχου πολλαπλών υποθέσεων με έναν αλγόριθμο ομαδοποίησης για την επιλογή ενός μικρού αλλά αντιπροσωπευτικού συνόλου γονιδίων, σε δυαδικά προβλήματα ταξινόμησης. Συγκεκριμένα, σε μια πληθώρα ασθενειών και συνόλων δεδομένων, η mAP-KL πέτυχε ανταγωνιστικά αποτελέσματα κατάταξης σε σύγκριση με άλλες 12 FS μεθόδους και ειδικότερα σε σχέση με τη μέθοδο HykGene, η οποία ακολουθεί παρόμοια φιλοσοφία, δηλαδή αρχικά κατάταξη και στη συνέχεια ομαδοποίηση. Τα πλεονεκτήματα της mAP-KL έναντι της HykGene αλλά και άλλων παρόμοιων προσεγγίσεων οφείλονται σε τρία βασικά χαρακτηριστικά: στην καθοδηγούμενη από τα δεδομένα φύση της, στην χρήση της μεθόδου συσταδοποίησης AP, και στην ανεξαρτησία της από ταξινομητές. Πράγματι, η χρήση ενός δείκτη ποιότητας συστάδων, του Krzanowski και Lai, μειώνει οποιαδήποτε ασάφεια και παρέχει στον αλγόριθμο συσταδοποίησης έναν αντιπροσωπευτικό αριθμό πιθανών συστάδων. Επιπλέον, στην mAP-KL τα δεδομένα είναι αυτά που καθορίζουν το μήκος της λίστας, συγκεκριμένα η δομή των δεδομένων υπαγορεύει τον αριθμό των συστάδων και ο αλγόριθμος ομαδοποίησης αποφασίζει για τους εκπροσώπους από την κάθε συστάδα. Σε αντίθεση με άλλες μεθόδους, όπως για παράδειγμα στη HykGene, όπου χρησιμοποιείται ένας ταξινομητής περιτυλίγματος, στην περίπτωση μας κατά την επιλογή του συνόλου δεν εμπλέκεται κανένας ταξινομητής. Αυτό το μεθοδολογικό

χαρακτηριστικό έχει μεγάλη σημασία, μιας και τα υποσύνολά μας δεν εμφανίζουν το φαινόμενο της υπερπροσαρμογής το οποίο σχετίζεται με την εμπλοκή των ταξινομητών κατά τη διαδικασία επιλογής.

Σχετικά με τον προσδιορισμό των συστάδων, η εφαρμογή του AP αλγόριθμου, αντιμετωπίζει επιτυχώς το θέμα των αντιπροσωπευτικών γονιδίων ανά συστάδα. Άλλες παρόμοιες προσεγγίσεις με την mAP-KL παραδέχθηκαν την δυσκολία τους ως προς την αποτελεσματική επιλογή ενός ή περισσότερων αντιπροσωπευτικών γονιδίων ανά συστάδα. Εκτός αυτού, η AP ακολουθεί ένα μηχανισμό «δικτύου-γονιδίων» με την θεώρηση ότι αρχικά όλα τα γονίδια αποτελούν πιθανούς κόμβους ενός δικτύου. Τα επιλεγθέντα «exemplars» είναι τα κεντρικά γονίδια μιας συστάδας γονιδίων και πιθανώς οι βασικοί κόμβοι σε ένα δίκτυο γονιδίων. Ως εκ τούτου, η εξόρυξη των «exemplars» μπορεί να θεωρηθεί ως το πρώτο βήμα της διαδικασίας επαγωγής δικτύου και όχι μόνο το αποτέλεσμα μιας FS προσέγγισης. Στα μελλοντικά μας σχέδια, προτιθέμεθα να κατασκευάσουμε δίκτυα με βάση τα κορυφαία N γονίδια της μεθοδολογία μας και στη συνέχεια να διερευνήσουμε τα χαρακτηριστικά δικτύου των «exemplars». Μια πρώτη προσπάθεια προς αυτή την κατεύθυνση είναι ήδη διαθέσιμη στο πακέτο mAPKL, αν και περισσότερες επαγωγικές μέθοδοι δικτύου για την επανακατασκευή ρυθμιστικών δικτύων καθώς και μέθοδοι ελέγχου του εμπλουτισμού θα υιοθετηθούν στο επόμενο διάστημα.

TABLE OF CONTENTS

PREFACE.....	35
1. INTRODUCTION.....	37
2. MEASURING GENE EXPRESSION.....	41
2.1. Microarrays.....	41
2.1.1 Hybridization and gene expression.....	41
2.2. Data analysis.....	43
2.2.1 Low level analysis.....	44
2.2.2 High level analysis.....	46
2.3. Interpretation of genomic results.....	47
2.4. Next Generation Sequencing.....	48
3. COMPUTATIONAL INTELLIGENCE METHODS TO ANALYZE AND EXPLOIT GENE EXPRESSION MEASUREMENTS.....	51
3.1. Feature selection.....	51
3.1.1 Filter methods.....	52
3.1.2 Wrapper methods.....	59
3.1.3 Embedded methods.....	60
3.1.4 Dimension reduction methods.....	60
3.1.5 Hybrid methods.....	64
3.2. Clustering.....	65
3.2.1 Element selection.....	66
3.2.2 Variable selection.....	66
3.2.3 Variable standardization.....	66
3.2.4 Selecting a measure of association (similarity/dissimilarity).....	66
3.2.5 Selection of clustering method.....	67
3.2.6 Determining the number of clusters.....	67
3.2.7 Interpretation, validation and replication.....	68
3.3 Clustering algorithms.....	68
3.3.1 Hierarchical clustering.....	69

3.3.2	Self-Organizing Maps (SOMs)	70
3.3.3	Affinity Propagation	71
3.4	Classification	73
3.4.1	Support Vector Machines	73
3.4.2	Instance-based learning	77
3.4.3	Random Forests	78
3.5	Measuring the classification performance	80
3.5.1	Performance measures	80
4.	MAP-KL: A NEW HYBRID METHOD FOR FEATURE (GENE) SELECTION	83
4.1.	Introduction.....	83
4.2.	The general framework and implementation of our Methodology	83
4.2.1	The filtering method	83
4.2.2	The clustering quality index	85
4.2.3	The clustering algorithm	85
4.3.	Simulated data	87
4.3.1	The clusters setup	89
4.3.2	The 'choedata' setup	90
5.	FEATURE SELECTION WITH MAP-KL.....	93
5.1.	Introduction	93
5.2.	Microarray data	93
5.3.	Neuromuscular disease data	94
5.4.	Cancer data	104
5.5.	Analysis of previous experiments	110
5.6.	Summary	114
6.	BIOLOGICAL RELEVANCE OF DISCRIMINATORY GENE LISTS	117
6.1.	Introduction	117
6.2.	The gene lists from a Systems Biology perspective	117

6.3.	The gene lists from a disease point of view	122
7.	R-PACKAGE IMPLEMENTATION	125
7.1.	Introduction	125
7.2.	Classes and functions of the mAPKL package.....	125
7.3.	An analysis scenario with mAPKL package.....	127
7.4.	Availability and Future Directions.....	135
8.	DISCUSSION, CONCLUSIONS AND FUTURE WORK	137
8.1.	Discussion.....	137
8.2.	Conclusions and future directions	137
	REFERENCES	141

LIST OF FIGURES

Figure 2.1: Hybridization of an unknown sample to a 25-mer oligonucleotides array. . .	42
Figure 2.2: The PM and MM probe pairs	42
Figure 2.3: The arRP-I sequencing array	43
Figure 2.4: Boxplots of 3 normalization algorithms showing the different effect when applying on the same raw data.....	45
Figure 2.5: Boxplot of raw probe intensity values	46
Figure 2.6: Boxplot of \log_2 transformed probe intensity values.....	46
Figure 2.7: Identification of novel potential disease mutations against (A) MYH7 and (B) <i>ILK</i> genes [30]	49
Figure 3.1: The categorization of feature selection techniques [5].....	52
Figure 3.2: The relationship between fold change, t-score and cat score [50].....	59
Figure 3.3: The impact of feature selection in cluster analysis	66
Figure 3.4: A single-link agglomerative clustering dendrogram [60]	69
Figure 3.5: Principle of SOMs [62].....	71
Figure 3.4: The functionality under affinity propagation	72
Figure 3.5: The overfitting problem [20].....	73
Figure 3.6: The separating hyperplane and the relevant margins [64].....	75
Figure 3.7: A feature mapping from a two-dimensional input space to a two-dimensional feature space [66].....	77
Figure 3.8: The kNN algorithm for $k=5$	78
Figure 3.9: The top-down majority voting procedure in RF	79
Figure 3.10: The performance metrics in binary classification [69]	81
Figure 4.1: The overall classification accuracy of five feature selection methods on ten datasets of neuromuscular disease data according to four classification algorithms.....	84
Figure 4.2: The influence on the accuracy when differentiating the length of the training set.....	84
Figure 4.3: The mAP-KL methodology flowchart	86

Figure 4.4: A scatter plot of five clusters with DEGs and one cluster, the 6th, with non-DEGs.....	88
Figure 4.5: The boxblots of the average log ₂ expression summary intensity as a function of spiked-in fold change.	88
Figure 4.6: The relationship between DEGs and top-ranked genes	90
Figure 5.1: The overall classification results (AUC metric) with RF classifier	115
5.2: The classification performance (AUC) of mAP-KL across diseases for three classifiers.....	116
Figure 6.1: The overview of the PE scores.....	122
Figure 7.1: A UML schematic representation of the classes and functions of the mAPKL. The solid rectangles with the three compartments represent classes. In the first compartment is the name of the class, in the second compartment is the attributes of the class, and in the third is the methods/functions relevant to the class. The 'Report' rectangular is a special type of class called utility that has static attributes and methods and no instances. The dotted lines represent 'dependencies' between classes. The lines with the arrowhead represent 'generalizations' and show the parts (static attributes) of the 'Report' class.	127
Figure 7.2: Loading the packages and sampling the data	128
Figure 7.3: The density plots per normalization method	128
Figure 7.4: The density plots per normalization method	129
Figure 7.5: The density plots per normalization method	130
Figure 7.6: Pathway analysis results	131
Figure 7.7: The exemplars' network characteristics.....	132
Figure 7.8: The exemplars that are also network hubs	133
Figure 7.9: A network graph of the weighted local degree of centrality	133
Figure 7.10: The summarized mAP-KL analysis report	134

LIST OF TABLES

Table 4.1: The statistical parameters under the simulated data	87
Table 4.2: The number of clusters identified by mAP-KL for several top N ranked genes compared to three other FS methods (the number of genes per subset is in parenthesis).....	89
Table 4.3: The subsets of genes selected from the 'choedata' according to mAP-KL. We have marked bold the DEGs.....	91
Table 5.1: The real microarray data divided in train and test sets	94
Table 5.2: The classification results in ALS and DMD neuromuscular diseases according to RF classifier	95
Table 5.3: The classification results in ALS and DMD neuromuscular diseases according to SVM classifier	96
Table 5.4: The classification results in ALS and DMD neuromuscular diseases according to KNN classifier.....	97
Table 5.5: The classification results in JDM and LGMD2A neuromuscular diseases according to RF classifier	98
Table 5.6: The classification results in JDM and LGMD2A neuromuscular diseases according to SVM classifier	99
Table 5.7: The classification results in JDM and LGMD2A neuromuscular diseases according to KNN classifier.....	100
Table 5.8: The classification results in LGMD2B and NM neuromuscular diseases according to RF classifier	101
Table 5.9: The classification results in LGMD2B and NM neuromuscular diseases according to SVM classifier	102
Table 5.10: The classification results in LGMD2B and NM neuromuscular diseases according to KNN classifier.....	103
Table 5.11: The classification results in breast and colon cancers according to RF classifier.....	104

Table 5.12: The classification results in breast and colon cancers according to SVM classifier.....	105
Table 5.13: The classification results in breast and colon cancers according to KNN classifier.....	106
Table 5.14: The classification results in leukemia and prostate cancers according to RF classifier.....	108
Table 5.15: The classification results in leukemia and prostate cancers according to SVM classifier.....	109
Table 5.16: The classification results in leukemia and prostate cancers according to KNN classifier.....	110
Table 5.17: An overview of the published classification results in van 't Veer et al. breast cancer data.....	112
Table 5.18: An overview of the published classification results in Singh et al. prostate cancer data.....	112
Table 5.19: An overview of the published classification results in Golub et al. ALL/AML leukemia data.....	113
Table 5.20: An overview of the published classification results in Alon et al. colon cancer data.....	114
Table 6.1: The DoU of seven FS methods across neuromuscular data.....	118
Table 6.2: The DoU of seven FS methods across cancer data.....	118
Table 6.3: The pathway analysis results on neuromuscular data.....	120
Table 6.4: The pathway analysis results on cancer data.....	121
Table 6.5: The pathway lists in the LGMD2B disease.....	122
Table 6.6: The disease enrichment per gene list.....	123
Table 7.1: Classification performance of gene exemplars per preprocessing method.....	130

PREFACE

I carried out my research at the premises of the The Biomedical Research Foundation of the Academy of Athens (BRFAA) under the supervision of Dr. George Spyrou.

1. INTRODUCTION

The dawn of DNA microarray technology has improved our potential to comprehend the underlying mechanisms of human diseases and to aid in more accurate classification, diagnosis, and/or prognosis [1]. Because of its high throughput nature, computational tools are essential in data analysis and mining in order to help biomedical researchers to maximize the extracted knowledge from the experimental results [2].

In the area of diagnostics, microarray-derived markers are emerging as a valuable tool. Similar to any other clinical test, the primary goal of molecular tests, including microarray tests, is to provide reliable and timely results for improving patient care. In order to maximize the usefulness of microarrays in the diagnostic/prognostic arena it is important to minimize the number of biomarkers that need to be tested for an accurate diagnosis to be reached. Two prime examples of successful identification of such biomarkers and their effective transition to the clinic are the MammaPrint [3] and the Oncotype [4] molecular tests for breast cancer with a 70-gene and a 21-gene molecular signatures, respectively.

The selection of those biomarkers, however, is a challenging process in which feature selection methods could make a significant contribution. Indeed, from the late 90s a plethora of methods emerged and applied on several microarray studies. Despite differences in their fundamental algorithms, they all share the same objectives: 1) to avoid overfitting and improve prediction performance; 2) to make faster and cost effective models; and 3) to offer a deeper insight into the underlying processes [5]. Nevertheless, selecting those “significant” genes that perform the same level of classification in relation to a specific disease is far from feasible at the moment and still an open issue.

In reality, every microarray dataset may result to as many significant gene lists to as many feature selection methods we apply. Even in cases where methods share the same principles the produced gene lists are bound to diverge. Speaking of methods that share common principles, we may define the following broad groups of feature selection methods. Filtering, wrapper and embedded feature selection methods are the key categories in the field, each one with the respective advantages and disadvantages. In addition to this classification, a new class of feature selection methods, hybrid methods, has emerged. Hybrid methods’ combine methods of different categories aiming at taking advantage of their pros while alleviating their cons of benefit to the “significant” gene list selection.

Though, combining methods is a constructive decision making process based always on scientific assumptions, either biological or statistical, rather than on pot luck. For instance, Jaeger et al. [6] claimed that ranking algorithms produce lists of genes, where the top ranked genes are highly correlated with each other, mainly because they belong to the same pathway. Additionally, Hall in his thesis [7] investigated the hypothesis that “A good feature subset is one that contains features highly correlated with the class, yet uncorrelated with each other”. Those beliefs were the springboard for several hybrid methods which combined a ranking (filtering) method and a clustering method to conclude to a list of significant genes.

In particular, Jaeger et al. employed a fuzzy clustering algorithm to prefilter the genes by grouping them according to their similarity. Then, with the aid of a statistical test like t-test or Wilcoxon, selected one or more representative genes from each cluster to form a list of “significant” yet uncorrelated genes. In this study, the number of clusters to be formed and the number of representative genes remained unaddressed. Similar to

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

Jaeger et al., Hanczar et al. [8] proposed a two step method where an unsupervised clustering method, K-mean, combined with a mathematical notion, *prototype gene*, that tries to identify the representative genes of each cluster. Analogous issues to Jaeger et al. appeared in this study, and characterized as objectives for future work by the researchers. An alternative algorithmic approach, where ranking of genes precedes any other method is described in the mRMR [9] method. Particularly, the initial ranking through t-test or F-test is then combined with a sequential iteration between pairs of the ranked genes, to conclude to a subset of “significant” genes according to some criteria, maximum relevance and minimum redundancy. One considerable drawback of this approach is that the redundancy criterion may exclude genes that considered important from a biological point of view. Another interesting approach, HykGene [10], proposed a three step gene selection, which incorporates a filtering algorithm, a hierarchical clustering on the top-ranked genes and finally a sweep-line algorithm that first identifies the clusters from the dendrogram and then selects one representative gene per cluster (for more details see section 3.1.5).

Taking into account the promising classification results of those combined methods as well as their intrinsic limitations, we considered a new hybrid method, mAP-KL. In the proposed approach, the genes are first ranked according to their differential expression using a multiple hypothesis t-test, which controls successfully the Type I error. Then the top N ranked genes are held and grouped to clusters with the Affinity Propagation (AP) clustering algorithm [11]. Prior to AP a clustering index algorithm determines the number of clusters among the top- N -genes. The output of this method is a subset of genes, one exemplar per cluster that best describes the phenotypes’ characteristics.

We subjected our method to a series of evaluation tests on simulated microarray data in the first part and real microarray data in the second. Regarding the real microarray data we employed datasets of six neuromuscular diseases as representatives of small cohorts and four cancer datasets with numerous samples per phenotype. Moreover, we applied twelve other feature selection approaches on the same real microarray data along with mAP-KL and compared the classification results using several metrics, for example AUC, TNR, TPR. Apart from the classification analysis we investigated the produced gene lists from a biological point of view to have a further assessment of our method towards the other competitors. The overall evaluation results suggest that mAP-KL is a feature selection method that delivers robust gene lists of biological relevance that may assist biologists to gain valuable insights.

This dissertation is organized in eight chapters. Chapter 2 provides a basic yet necessary introduction to microarray technology and particularly to Affymetrix gene chips and how we should treat and analyze microarray data. In the last part we also present the Next Generation Sequencing technology as a promising alternative in the forthcoming years.

Chapter 3 provides a thorough representation on computational methods ranging from feature selection to machine learning and clustering. In particular, we first introduce the field of feature selection and representative methods per category in the context of microarray data. The presented methods are also those applied and compared with our method in the real microarray data. Moreover, we discuss the fundamentals of clustering and introduce the Affinity Propagation method as a promising contemporary method, which is also part of our hybrid method. In the final part of this chapter we shed light on machine learning techniques and specifically on Support Vector Machines, KNN and Random Forests since those algorithms utilized through the Weka environment to classify the real microarray datasets according to the “significant” gene lists per method.

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

In chapter 4 we elaborate on the proposed methodology by explaining the rationale behind the method, the distinct parts as well as the relevant pipeline. Moreover, we present the evaluation tests applied on simulated data that support the potential virtue of the method.

Chapter 5 incorporates the evaluation tests and results applied on real microarray data. In particular, we have included the comparative results among our method and twelve other approaches presenting the achieved classification results in small and large cohorts. Additionally, we review the classification results on the same large cohorts that achieved during other studies.

In the following chapter we provide the biological analysis of the gene lists produced by *mAP-KL*. The aim of this analysis is to inspect from a biological perspective the strength of each method and particularly for the proposed methodology.

The software implementation of *mAP-KL* method is described in chapter 7. Specifically, we discuss the classes and the functions included in the r-package named *mAPKL* that is archived in the Bioconductor software project. Besides, we present a case study scenario to demonstrate the functionality of the package.

Finally, in chapter 8 we summarize the evaluation results and discuss in more depth the findings of this study trying to emerge the advantages of *mAP-KL* against other feature selection methods in the context of microarray gene expression analysis. Not to mention that we also pinpoint the shortcomings of our method and propose future directions that may evolve the proposed algorithm.

2. MEASURING GENE EXPRESSION

2.1. Microarrays

It was since 1995, when the genome of the bacterium *Haemophilus influenzae* was completely sequenced. So far, the genomes of more than 4,100 organisms have been sequenced [12], deluging us with billions of bases. However, this huge amount of information is inadequate by itself to enlighten us about the genes' functionality and collaboration, the genes' malfunctions that induce diseases, the development of efficient drugs, or even the basic cell functions. Therefore, genomic studies intend to understand biology instead of just providing us with a list of genes and maybe their functionality. Towards this goal, there are several tools and technologies among which are high-density arrays of oligonucleotides or complementary DNAs (cDNAs) [13].

A variety of DNA microarray chip devices, fabricated on glass, silicon, or plastic substrates, is commercially available. The underlying principle of microarrays is the hybridization of an unknown sample to DNA molecules of known sequence, attached at specific location on a surface. Generally, in each array there are thousands of different DNA probe sequences arranged in a defined matrix. Unlike conventional nucleic-acid hybridization methods, microarrays can identify thousands of genes simultaneously thus, revolutionizing the gene expression analysis in cells and tissues [14, 15].

2.1.1 Hybridization and gene expression

Oligonucleotide arrays take advantage of the nucleic acid strands capacity to recognize or hybridize complementary sequences through base pairing. Oligonucleotide probes are designed and synthesized based solely on sequence information to serve as sensitive, unique, and sequence-specific detectors. A given gene is represented by 15–20 different 25-mer oligonucleotides Figure 2.1, which overlaps slightly only if necessary or inevitable [16]. In relation to eukaryotic organisms, probes are chosen typically from the 3' end of the gene or transcript (nearer to the poly(A) tail) to control the effect of a partially degraded messenger RNA (mRNA). A further control element is the use of mismatch (MM) and perfect match (PM) oligonucleotides, which are identical except for a single base in a central position, Figure 2.2. In particular, the MM probes allow the discrimination between 'real' and cross-hybridization signals. Non-specific or semi-specific hybridization produces higher signal for the PM probes than for the MM probes leading to consistent patterns that are highly unlikely to occur by chance. The PM/MM pairs hybridization produces recognizable and quantitative fluorescent patterns even for low RNA concentrations [17].

Gene expression (mRNA abundance) monitoring may produce quantitative results for as many as 40,000 genes in a single hybridization. A central benefit of representing the whole genome or a large chunk of different genes on an array is a broader and unbiased analysis over the genes related to the inspected condition. The collection of the expressed or transcribed genes, referred as the expression profile or transcriptome, is the first step towards protein synthesis and is responsible for both morphological and phenotypic differences. Besides, the transcriptome is characterized by its rapid response to either environmental perturbations or normal cellular events. As a consequence, it provides us with a

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

valuable knowledge that propels us into understanding regulatory mechanisms, cellular functions and biochemical pathways as well as into determining diseases' causes, and efficient drug development [13].

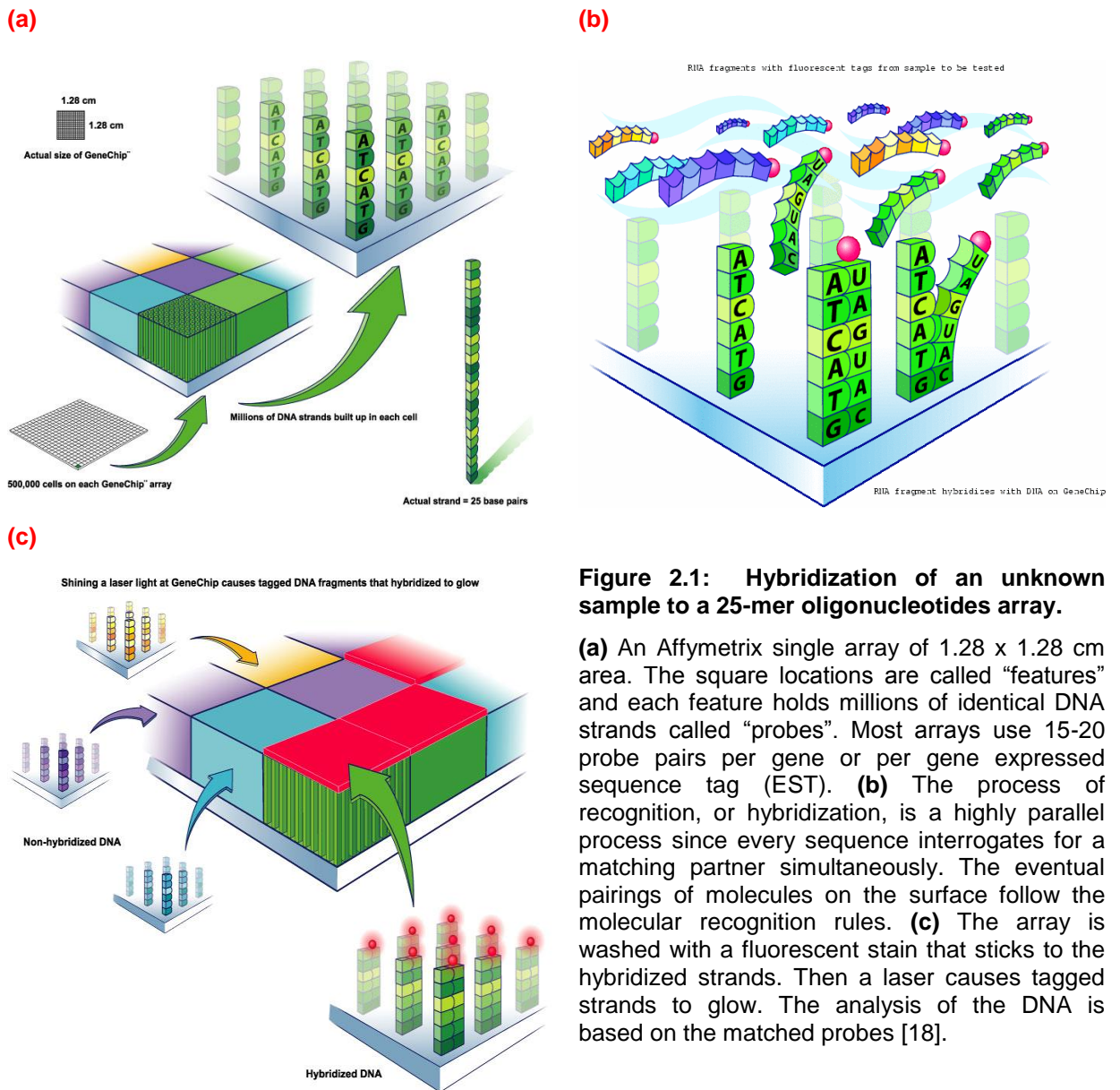


Figure 2.1: Hybridization of an unknown sample to a 25-mer oligonucleotides array.

(a) An Affymetrix single array of 1.28 x 1.28 cm area. The square locations are called “features” and each feature holds millions of identical DNA strands called “probes”. Most arrays use 15-20 probe pairs per gene or per gene expressed sequence tag (EST). **(b)** The process of recognition, or hybridization, is a highly parallel process since every sequence interrogates for a matching partner simultaneously. The eventual pairings of molecules on the surface follow the molecular recognition rules. **(c)** The array is washed with a fluorescent stain that sticks to the hybridized strands. Then a laser causes tagged strands to glow. The analysis of the DNA is based on the matched probes [18].

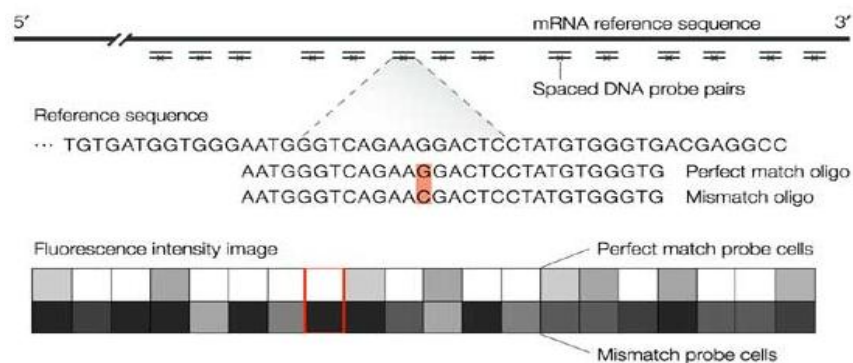


Figure 2.2: The PM and MM probe pairs

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

A successful example of how gene chips may shed light on the causes of a disease and eventually become the platform of a clinical test is the arRP-I sequencing array, Figure 2.3. In particular, scientists at the Kellogg Eye Center (KEC) developed a rapid genetic test for the retinitis pigmentosa (RP) blinding disease, to assist the diagnosis of physicians. This test screens simultaneously for mutations, both previously known and novel, in multiple genes and up to now more than 30 genes related to RP have been identified thus, allowing physicians to develop and apply genetic therapies. From a disease classification viewpoint, Golub et al. used a dataset of 34 samples, and monitored more than 6,000 genes per array to conclude to a set of 50 genes that discriminate highly accurate samples between individuals with and without acute leukaemia. This study indicated that microarray experiments require a sufficient number of samples per condition e.g. healthy vs tumor to account for possible tumor markers and also that a set of significant genes rather than single genes is necessary for reliable predictions. However, before integrate multiple samples into a single analysis, the hybridization intensities have to undergone a preprocessing step to maintain high standards of data quality [19].

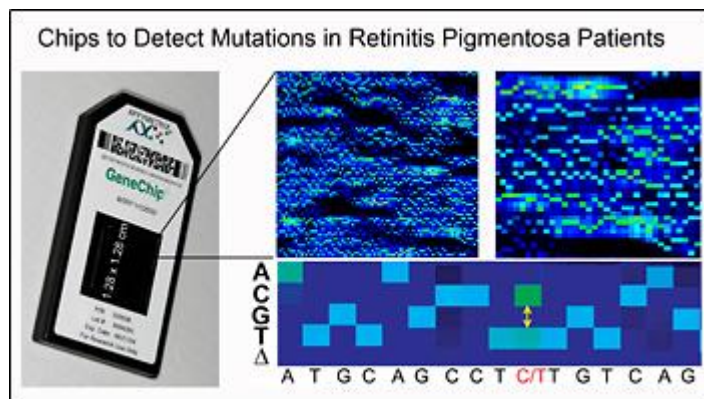


Figure 2.3: The arRP-I sequencing array

2.2. Data analysis

The analysis of the microarray data is mainly divided into two levels: a low level analysis and a high level analysis. The low level also called data preprocessing includes image analysis, data transformation and normalization, whereas the high level analysis incorporates inference and/or classification. Image analysis deals with appropriate ways to quantify spots on microarrays. So far, many image processing algorithms have been developed particularly for Affymetrix arrays, where mainly all of them try to estimate the amount of RNA while minimizing the extraneous sources of variation owing to array-specific physical defects. During the normalization step we strive to control any technical variation included in the data whilst maintaining the prospective biological variation [20]. These two types of variation coexist within the intensity values, though the technical variation is believed to predominate over biological. Non-biological sources of variation can be introduced during sample preparation (e.g., dye effects), array manufacture (e.g., probe concentration), hybridization (e.g., amount of sample) and in the measurement process (e.g., scanner inaccuracies). Normalization methods can be applied either within arrays, two-color case, or between arrays for single-channel arrays (Affymetrix chips). In relation to data transformation, we usually

imply any logarithmic transformations necessary to make our data more normal like [21].

2.2.1 Low level analysis

The Affymetrix GeneChip Operating Software incorporates the MAS 5.0 (MAS5) method. MAS5 is a single array method that can be applied on individual arrays and carries out Global background correction, local background correction, summarization, and normalization, Figure 2.4(a). At first an intensity quintile of 2% is defined as Global background and is subtracted from all probe intensities. Then, during local background, an Ideal Mismatch intensity (IM) value is used to restrict the negative values issue, since approximately the 30% of MM is greater than PM. The IM is either equal to the MM when $PM > MM$ or equal to a fraction of the PM otherwise. The background corrected PM intensities of each probe set are employed during the summarization step, and an expression index is computed through the one-step Tukey biweight M-estimator. Finally, in the normalization step each expression index is multiplied by a scaling factor (sf) specific for the array. This factor arises as follows: a trimmed mean of the indexes is computed excluding the 2% of the highest and the lowest values. Then this trimmed mean value divides a target intensity (S_c) value, which by default in MAS5 algorithm is 500, and the outcome is the array's sf value [22].

Along with MAS5, the Robust Multichip Average (RMA) methods are the most commonly employed preprocessing approaches for Affymetrix image analysis chips, Figure 2.4(b). Contrary to MAS5, RMA is an academic alternative proposed by Irizarry et al. in 2003, which takes into account information between arrays after an initial background correction step. Apart from that, RMA utilizes only the PM values and the normalization step precedes the summarization. In brief, the RMA initially corrects arrays for background using a convolution model, where the PM values are considered as the sum of background intensity and real signal intensity. Then, the background corrected PM values are normalized through the quantile normalization algorithm based on normal distribution, and finally the expression indexes for each probe-set are computed separately through a linear model on a log₂ scale [22].

In 2004 Wu et al. proposed the Gene Chip RMA (GCRMA), which is a modification of the RMA applying a different background correction, Figure 2.4(c). In GCRMA the background signal of both PM and MM probe pairs is divided into optical noise and non-specific binding defined as $PM = O + NPM + S$ and $MM = O + NMM$, where O represents the optical noise and is a constant specific to array, N stands for the non-specific binding and S is the actual -biological signal. The background corrected PM values are computed either as a maximum likelihood estimator or as a random through an empirical Bayes approach. Overall, the GCRMA produces more accurate results for differential expression analysis at the expense of lower precision in clustering because of the introduction of artifacts [11]. Although plasmide data sets – real data with known structure – are used to test and evaluate proposed analytical methods, it is still unclear which method performs best in all cases [20].

In contrast to RMA and GCRMA, the MAS5 provides us with expression indexes in exponential form. As a result, prior to any further analysis it is crucial to apply a logarithmic transformation, usually logarithm with base two, Figures 2.5 and 2.6, just like RMA and GCRMA. The first and most obvious reason is to make them

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

more symmetric hence, obliged to the normality assumptions of many parametric statistical tests e.g t-test, SAM, ANOVA etc. In addition, logarithmic transformations cope successfully with random error minimization. Random error describes inevitable uncertainties in all scientific measurements rather than mistakes. For example, although the log ratio of a non differential expressed gene across

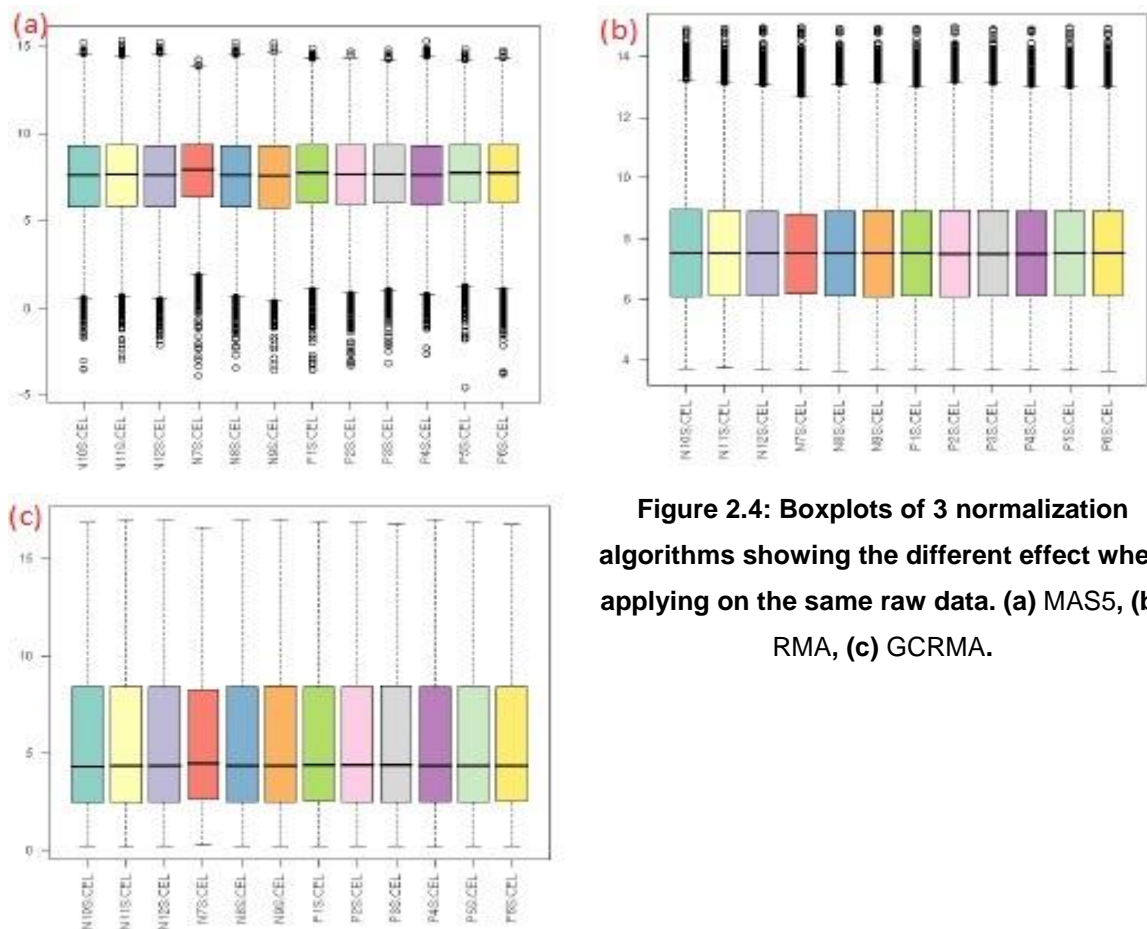


Figure 2.4: Boxplots of 3 normalization algorithms showing the different effect when applying on the same raw data. (a) MAS5, (b) RMA, (c) GCRMA.

all arrays should be 1:1, the existence of random error as a stochastic model, imposes deviations from this ratio. In raw intensity values, the random error is roughly proportional to signal intensity, therefore even equivalent fold changes are not equally reproducible [23].

In case of ratios between two conditions (normal:disease), the logarithmic transformation is considered mandatory. As a paradigm consider the following binary case (normal:disease), firstly with raw data values and secondly with log₁₀ transformed values. Suppose we have three samples per condition N₁=1.1, N₂=1.4, N₃=5 and D₁=2, D₂=5, D₃=15. The disease:normal ratios of those samples have a mean=0.39, standard deviation=0.14 and coefficient of variation=0.37. Now if we invert the ratios, normal:disease, the relevant values are: mean=2.8, sd=0.9 and cv=0.32. In the second case, where the ratio is logarithmically transformed (log₁₀(D₁/N₁)), the relevant values for the disease:normal case are mean=0.43, sd=0.15 and cv=0.35. Inverting the ratios to normal:disease, the absolute values of the metrics are the same but with a negative sign in front of the mean and the coefficient of variation, which reflects that the numerator is smaller (-) than the denominator [23].

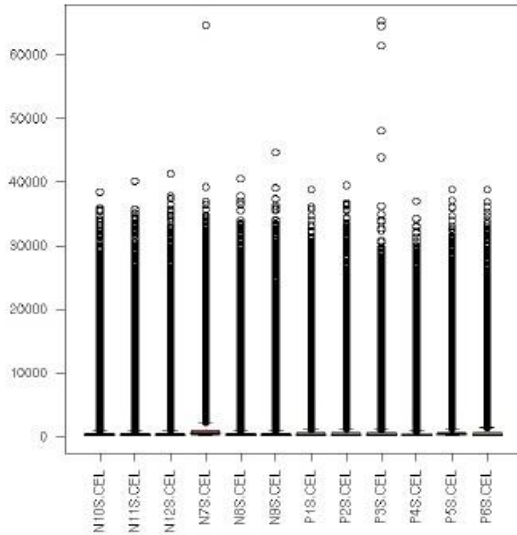


Figure 2.5: Boxplot of raw probe intensity values

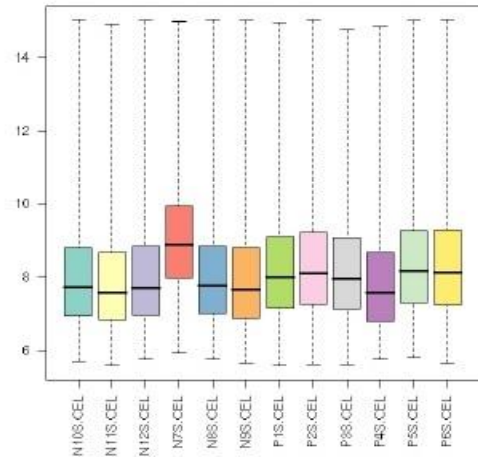


Figure 2.6: Boxplot of \log_2 transformed probe intensity values

2.2.2 High level analysis

During this state of gene expression analysis we turn our focus on statistical tests aiming at detecting differentially expressed genes in samples from two distinct conditions (e.g. normal and disease). Although in the preprocessing state we tried to eliminate the non-biological variation among genes, such variation still exists and it is the scope of statistical tests to detect those genes related only to biological differentiation factors. Usually an alternative (research) hypothesis is stated in positive terms (e.g. whether a particular gene or group of genes is related to the inspected disease) contrary to the statistical null hypothesis, which is stated in negative terms (e.g. whether a particular gene or group of genes is unrelated to the inspected disease). A threshold value (α) is necessary to determine over the significance of the genes, in other words whether a gene deviates from the null hypothesis simply by chance or because the alternative hypothesis stands. The null and the alternative hypothesis of the mean values of a population between two conditions can be stated as follows:

$$\begin{aligned}
 H_0 : \mu_{DISEASE} - \mu_{NORMAL} &= 0 \\
 H_1 : \mu_{DISEASE} - \mu_{NORMAL} &\neq 0
 \end{aligned}
 \tag{2.1}$$

Although a zero difference is usually stated, a non-zero value can be stated too. Besides, the hypotheses can be directional and as a result the = sign can be replaced by \geq or by \leq and additionally the \neq sign by $<$ or $>$, respectively [23].

Since a typical experiment involves n number of samples, between ten and a few hundreds, and m number of genes, usually more than 10,000, the biological question can be restated as a problem of multiple hypothesis testing. Indeed, thousands of null hypotheses H_j are tested simultaneously for each gene j trying to reject the potential association between the expression level X_j and the inspected condition Y . However, in any hypothesis testing, there is always a

probability of incorrect inferences, committing two types of errors. A false positive, or Type I error, when declaring that a gene is differentially expressed when it isn't, and a false negative, or Type II error, when failing to identify a truly differentially expressed gene. Behind the multiple comparisons problem, there are two aspects to consider: firstly, a test statistic T_j for each gene j have to be computed through a statistical method (like t-test, ANOVA, e.t.c.), and secondly, a multiple testing procedure should be applied such that on the one hand to determine which hypothesis to reject while on the other to control the Type I error rate [24].

Taking into account the large number of genes in microarray experiments, it is of great importance to successfully control the Type I error, since a false positive rate of 1% results in 100 false calls when monitoring 10.000 genes. Usually for a single statistical test we set beforehand an acceptable threshold value, for example the P -value <0.05 , to control the false-positive rate. "The P -value is the smallest level of significance that results in rejection of the null hypothesis. The smaller the P -value the stronger the evidence against the null hypothesis." Though in differential expression analysis, where numerous of tests are conducted a P -value of 0.05 leads to 5% of differentially expressed genes even if none of them are actually differentially expressed [23].

Hence, there is a need to adjust the P -value produced in simultaneous testing. The Bonferroni adjustment procedure is widely applicable to simultaneous testing situations, but lacks of power since its product, a new false positive rate, is more stringent thus increases drastically the false-negative rate. In particular, given n hypotheses, H_1, H_2, \dots, H_n , and a nominal error rate of α , we test each individual hypothesis H_i at a reduced significance level α_i such that $\sum \alpha_i = \alpha$ and typically $\alpha_i = \alpha/n$. Let p_i be the unadjusted P -value for H_i hypothesis, then the H_i is rejected when $np_i \leq \alpha$ and that is the Bonferroni adjusted P -value, p_{Bonf} [25]. Thus, in the previous example the Bonferroni procedure yields a new false positive threshold of $0.000005 = 0.05/10000$. This new threshold reduces the probability to 0.05, which means that in the entire dataset the probability of making at least one false positive error is 0.05. As a consequence, other methods addressing this highly stringent threshold appeared, to provide a more balanced control between sensitivity and specificity like the false discovery rate (FDR) [23] and will be discussed in the feature selection chapter.

2.3. Interpretation of genomic results

Measurements of tens of arrays and thousands of genes found to conclude to robust expression markers necessary to produce reliable and highly accurate predictions relevant to phenotype discrimination. What's more, such broad experiments are also equally important in understanding basic biological processes or even understanding and treating complex human diseases like cancer. Indeed, by identifying those genes that are upregulated in a tumor type we may conclude to causative effects that transform cells from normal to cancerous state and more interestingly to deduce potential therapeutic targets. However, making biological meaningful assertions requires sophisticated systems of knowledge representation, knowledge bases, which organize the data, facts, observations, relationships and even hypotheses that outline the ground of our current scientific insight. Furthermore, such knowledge bases no need to just store the information but also to provide it to scientists in a structural and

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

meaningful way to assist their understanding and interpretation of even complex observations [13].

Taking full advantage of the abilities of such knowledge bases, entails a restricted vocabulary and a well-defined semantic and grammar, which essentially incorporates the facts, ideas, connections, and observations existing in the scientific literature as well as in the scientists' minds. Unluckily, the scientific literature did not evolve this way and therefore a great deal of dedicated, systematic human effort is required to convert all the stored info into a systematic, organized, linked, visualized and searchable form. In accordance to these prerequisites the Gene Ontology Consortium (GO) produced a comprehensive controlled set of terms describing genes across organisms. The GO project functions in conjunction with organism databases such as FlyBase, Mouse Genome Informatics Database, the Saccharomyces Genome Database (SGD), and utilizes terms that describe molecular function, cellular location and biological processes [26]. Additionally, there are other knowledge databases like the Munich Information Center for Protein Sequences (MIPS), WormBase, the Kyoto Encyclopedia of Genes and Genomes (KEGG), the Encyclopedia of E. coli Genes and Metabolism (EcoCyc) which also incorporate sequence, genetics, gene expression, homology, regulation, function and phenotype information in a structured and functional form. A step forward of these databases would be biological 'expert systems' in which concepts and facts will be more fully integrated and related, thus allowing connections between initially unrelated observations and information to be made, as well as across organisms. As a consequence, scientists would be eligible to state any insightful question and receive the most meaningful interpretations from a biological perspective [13].

2.4. Next Generation Sequencing

A potential alternative to microarrays for high throughput studies is the RNA_Seq and more recently the next generation sequencing (NGS) technologies. Whereas microarrays are based on the simultaneous hybridization of thousands of genes, the sequencing techniques allow for the complete sequencing of the whole transcriptome of an organism without prior knowledge for any particular gene. The need for such an holistic approach emerged from the intrinsic limitations of microarray technology to control successfully the background levels of hybridization, particularly in cases of transcripts with low abundance [27], as well as to shed light on exon-level expression and alternative splicing. Alternative splicing, i.e., the process where individual exons of a gene are spliced and produce different isoforms of mRNA, is responsible for proteins variability within an organism and almost 50% of disease mutations in exons may originate to mRNA defects. Hence, it was of vital importance the development of a technology, which on the one hand measures the exon expression while on the other identify isoforms of mRNA [28].

It was since 1975 when the first automated sequence method developed by Edward Sanger (the chain-termination method) and for almost two and a half decades was regarded as the gold standard for nucleic acid sequencing. Indeed, the Human Genome Project accomplished in 2003 based solely on Sanger sequencing. However, the growing demand for faster and cheaper sequencing forced the development of second-generation or next-generation sequencing methods (NGS). Those new methods carry out massively parallel sequencing, Figure 2.7, through which a complete genome may be sequenced within a single

3. COMPUTATIONAL INTELLIGENCE METHODS TO ANALYZE AND EXPLOIT GENE EXPRESSION MEASUREMENTS

3.1. Feature selection

Microarray data analysis is widely used for the identification of ‘informative’ genes. However, due to the ‘curse’ of dimensionality, where the number of gene probes represented on microarrays far exceeds the available number of cases (samples) as well as the inherent noise in microarray data, feature selection (FS) approaches strive to achieve this goal. Typically, informative genes are selected according to a two-sample statistical test combined with multiple testing procedures to guard against Type 1 errors [31]. This methodology generates gene lists, which then can be either ranked or filtered according to certain statistical criteria, e.g. p-value, q-value etc. The selected subset of genes is assumed to construct better classifiers, both in terms of accuracy and efficiency. In particular, we expect improved classification performance and generalization by avoiding over-fitting. Furthermore, the classifiers will be more efficient in time and space because of the fewer features, and biologists’ insights will be augmented [5].

An FS algorithm should perform efficiently and independently of the sample size and yield its subset within a reasonable period, to enable numerous experiments. Moreover, the subset’s length should be small, for instance, less than 50 genes, and the selected genes should present biological relevance to the inspected disease so as to facilitate further analysis. Despite the plethora of available FS methods, none of them has managed to successfully deal with all the aforementioned issues playing the role of a milestone. For instance, some methods are effective with small cohorts while others with large ones [32]. Aside from this, there are methods that are developed and tested for specific diseases, leaving their suitability for broader use unexplored [3]. Furthermore, some FS algorithms are so sophisticated that they either need specialized and expensive hardware to operate or an impractically long run time [33].

A wide variety of FS algorithms has been proposed [34-36] and depending on how they combine the feature selection search with the construction of the classification model, they can be classified into 3 categories: filter, wrapper, and embedded [5], Figure 3.1. Complementary to this categorization, hybrid approaches have drawn researchers’ interest. Specifically the benefits of usually two different techniques are combined towards the identification of an improved gene subset selection, for example, a univariate filter with a wrapper or an embedded method [6, 37-41]. Apart from FS methods, there are also data reduction techniques such as principal component analysis and partial least squares, which search for linear combinations of all genes to provide us with a small subset of ‘metagenes’ [42].

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases


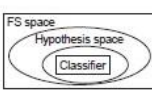
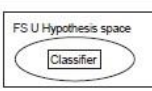
	Model search	Advantages		Disadvantages	Examples
Filter		Univariate	Fast Scalable Independent of the classifier	Ignores feature dependencies Ignores interaction with the classifier	Chi-square Euclidean distance t-test Information gain, Gain ratio
		Multivariate	Models feature dependencies Independent of the classifier Better computational complexity than wrapper methods	Slower than univariate techniques Less scalable than univariate techniques Ignores interaction with the classifier	Correlation based feature selection (CFS) Markov blanket filter (MBF) Fast correlation based feature selection (FCBF)
Wrapper		Deterministic	Simple Interacts with the classifier Models feature dependencies Less computationally intensive than randomized methods	Risk of over fitting More prone than randomized algorithms to getting stuck in a local optimum (greedy search) Classifier dependent selection	Sequential forward selection (SFS) Sequential backward elimination (SBE) Plus q take-away r Beam search
		Randomized	Less prone to local optima Interacts with the classifier Models feature dependencies	Computationally intensive Classifier dependent selection Higher risk of overfitting than deterministic algorithms	Simulated annealing Randomized hill climbing Genetic algorithms Estimation of distribution algorithms
Embedded		Interacts with the classifier Better computational complexity than wrapper methods Models feature dependencies		Classifier dependent selection	Decision trees Weighted naive Bayes Feature selection using the weight vector of SVM

Figure 3.1: The categorization of feature selection techniques [5]

3.1.1 Filter methods

Filter techniques are fast and efficient considering the high dimensionality of most microarray experiments, and that's the main reason for attracting most of the researchers' attention. Those techniques calculate a feature (gene) relevance score taking into account only the intrinsic properties of the data. Afterwards, the genes are ranked according to this score and only the top genes, either through a numeric threshold e.g. p-value or an arbitrary number of genes e.g. top20 genes, are kept and form the input of a classification algorithm. The calculation of the relevance score is usually based on univariate methods, which are computationally simple and fast, and the output is easy to understand and independent of the classification algorithm. So far, there is a plethora of univariate gene ranking techniques ranging from the simple fold change to parametric and model-free methods [5].

Regarding the parametric methods, they are based on the Gaussian distribution and assumptions of the samples and the most widely used representatives are the two-sample t-test and the ANOVA. Furthermore, modifications primarily in the variance's estimation has resulted in a number of t-test like statistics and Bayesian frameworks that better address the small sample size of the microarrays experiments as well as the inherent noise of gene expression data. However, our inability to validate the true underlying distributional assumptions due to small sample sizes, has given room to nonparametric or model-free methods as an alternative to Gaussian stringent distributional assumptions. The Wilcoxon rank-sum test, the rank products and the between-within classes sum of squares (BSS/WSS) are among the non-parametric methods engaged in gene expression studies. Those methods employ random permutations of the data to estimate the reference distribution of the statistic, which alleviates the small sample sizes problem and enhances the robustness against outliers [5].

The main disadvantage of univariate methods is that during the assessment of gene's significance the potential gene dependencies are ignored and that may affect the classification performance as well as the following biological analysis. Therefore, a number of multivariate filter techniques have emerged in order to incorporate genes dependencies to some degree. Those methods ranges from simple bivariate interactions to more sophisticated algorithms that try to explore higher order

interactions, such as correlation based feature selection (CFS), several variants of the Markov blanket filter method, The Minimum Redundancy - Maximum Relevance (MRMR), and Uncorrelated Shrunken Centroid (USC) [5]. During our study we employed seven univariate filter methods (maxT, ODP, eBayes, SAM, SNR and t-test), and one multivariate filter algorithm (cat).

3.1.1.1 The single-step maxT adjusted p-values

The biological problem of identifying the differential expressed genes in a number of mRNA samples can be restated as a multiple hypothesis testing problem. In particular, let X_j to be the expression value for gene j and Y the covariate of interest (e.g. treatment/control). Multiple hypothesis testing entails the simultaneous null hypothesis testing H_j for each gene j of no association between X_j and Y . Typically, this approach involves two phases: (1) computing a test statistic T_j for each gene j , and (2) applying a multiple testing procedure to decide upon the rejected hypotheses in relation to a properly defined *Type I* error rate. Regarding the first aspect, there is a plethora of univariate statistical methods and the decision depends on the experimental design and the type of covariate. For instance, in the case of binary covariates either a t-statistic or a Mann-Whitney statistic are acceptable choices [24].

As regards the multiple testing procedures, there are three types: single-step, step-down, and step-up procedures. In single-step procedures the evaluation of each hypothesis is based on a critical value equal for all hypotheses regardless of the results of the other tests. This type of procedures lacks of power i.e. minimize a suitably defined *Type I* error rate, which stepwise procedures try to encounter by taking into account not only the total number of hypotheses but also their outcome. The step-down and step-up procedures fall into this category of multiple testing procedures. Specifically, in step-down procedures those hypotheses with either the smallest unadjusted p-values or the largest absolute test score are examined sequentially. As long as one hypothesis is accepted, the rest of the hypotheses are considered accepted too. On the contrary, in step-up procedures the hypotheses with the least significant scores are considered successively, and as soon as one hypothesis is rejected, the rest of the hypotheses are rejected too [24].

The purpose of each of the above multiple testing procedures, is mainly to control the *Type I* error rate i.e rejecting the null hypothesis when it actually is true. There are several approaches dealing with *Type I* error control, though the following four are the most standard [24].

- i. *Per-comparison error rate (PCER)*. The PCER is defined as the expected value of (number of *Type I* errors/number of hypotheses), $PCER = E(V) / m$.
- ii. *Per-family error rate (PFER)*. The PFER is defined as the expected number of *Type I* errors, $PFER = E(V)$.
- iii. *Family-wise error rate (FWER)*. The FWER is defined as the probability of at least one *Type I* error, $FWER = pr(V \geq 1)$.
- iv. *False discovery rate (FDR)*. The FDR of Benjamini & Hochberg is defined as the expected proportion of *Type I* errors among the rejected hypotheses, $FDR = E(Q)$ where by definition $Q = \begin{cases} V/R, & \text{if } R > 0, \\ 0, & \text{if } R = 0 \end{cases}$ and R the number of rejected hypotheses.

In order to have a strong control, of the FWER at a significance level α , the Bonferroni procedure is the most widely engaged in multiple testing. At this point, we may notice that the term strong control refers to control of the *Type I* error rate under any

combination of true and false hypotheses, i.e., any value of m_0 . On the other hand, weak control refers to control of the *Type I* error rate only when all the null hypotheses are true $m_0 = m$. In the context of microarray experiments, it is very unlikely that no genes are differentially expressed therefore, it is meaningful to have strong control of the *Type I* error rate. Besides, due to the co-regulation among group of genes, the test statistics and the resultant p-values of those genes are correlated too. Towards this phenomenon, Westfall & Young proposed the use of adjusted p-values for less conservative multiple testing procedures that take into consideration the dependence structure among test statistics like the *single-step minP adjusted p-values* and the *single-step maxT adjusted p-values*. During our experiment, we employed the *single-step maxT adjusted p-values* since its p-values require fewer computations than those in *single-step minP* and are defined as follows

$$\tilde{p}_j = \text{pr}(\max_{1 \leq l \leq m} |T_l| \geq |t_j| \mid H_0^C) \quad (3.1)$$

where H_0^C denotes the complete null hypothesis and the T_l the test statistics of the l^{th} hypothesis and t_j the test statistic of gene j [24]. Regarding the permutation algorithm that implements the *single-step maxT adjusted p-values* is included in the multtest r-package and has as follows:

For the original data, order the observed test statistics such that $|t_{s1}| \geq |t_{s2}| \geq \dots \geq |t_{sm}|$.

For the b th permutation, $b = 1, \dots, B$:

1. Permute the n columns of the data matrix X .
2. Compute test statistics $t_{1,b}, \dots, t_{m,b}$ for each hypothesis (i.e. gene).
3. Next, compute $u_{i,b} = \max_{l=i, \dots, m} |t_{sl,b}|$ (see equation (3.1)), the successive maxima of test statistics by

$$u_{m,b} = |t_{sm,b}|$$

$$u_{i,b} = \max(|u_{i+1,b}|, |t_{si,b}|) \text{ for } i = m-1, \dots, 1.$$

The above steps are repeated B times and the adjusted p-values are estimated by $\tilde{p}_{si}^* = \frac{\#\{b : u_{i,b} \geq |t_{si}| \}}{B}$ for $i = 1, \dots, m$ with the monotonicity constraints enforced by setting [43].

3.1.1.2 The ‘optimal discovery procedure’ (ODP)

In relation to single significance tests, the Neyman-Pearson proposed a procedure for optimal testing given the null and alternatives distributions. This procedure is based on the likelihood ratio

$$\frac{\text{probability of data under alternative distribution}}{\text{probability of data under null distribution}}$$

that rejects the null hypothesis if exceeds a predefined threshold. The strength of this procedure stems from the comparison of the exact likelihoods between alternative versus null hypothesis. In general, a single-hypothesis test involves a number of steps, which in the case of multiple hypothesis tests abstractly can be fall into two major steps.

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

- i. determining the order in which the tests should be called significant and
- ii. choosing an appropriate significance cut-off somewhere along this ordering [44].

The Neyman and Pearson notion is most relevant to the first step, but for a single hypothesis perspective, whereas the domain of "multiple hypothesis testing" deals with the second step. Typically, the main goal of a multiple-hypothesis testing approach is to estimate a cut-off error rate based either on the familywise error rate or on the false discovery rate in order to sort the tests according to their significance. However, the ordering of the tests is accomplished solely on the p-values obtained from each significance test rather than on information across tests, which ultimately affects the quality of the entire procedure [44].

The ODP method copes with the first step, trying to incorporate an optimal testing, in the area of multiple hypotheses, to provide a significance framework for the second step. In particular, the objective is to maximize the expected number of true positive for each fixed expected number of false positive results. This criterion directly relates to optimality in terms of false discovery rates. The ODP approach implements the creation of a statistic for each hypothesis test that engages the relevant information from all other tests, similar to shrinkage estimators employed in simultaneous point estimation. As a result, providing an improved way to order tests that should be called significant also improves the performance of the forthcoming multiple-testing procedures [44].

The ODP procedure involves three components:

- i. defining the optimality goal;
- ii. properly constraining the set of procedures over which the optimality is to be found;
- iii. deriving the procedure that achieves this optimality.

Concerning the first component, the optimality goal is to maximize the expected number of true positive results, ETP, for each fixed expected number of false positive results, EFP. Regarding false discovery rates (FDR) the key observation is that it may be interpreted and characterized in terms of EFP and ETP,

$$FDR \approx \frac{EFP}{EFP + ETP} \quad (3.2)$$

where the approximate equality it may turn into exact equality for large number of tests with certain convergence properties [44].

As regards to the second component, defining and employing significance thresholding functions has the advantage over critical functions that a positive score for each test is produced which can be used to sort the tests from most significant to least. This function is defined to be $S: \mathbb{R}^n \rightarrow [0, \infty)$ such that the null hypothesis is rejected if and only if $S(x) \geq \lambda$ for some λ chosen to satisfy an acceptable level of significance. In the case of multiple tests a 'single thresholding procedure' (STP) is defined to be a multiple-testing procedure equivalent to applying a single significance thresholding function S and cut-off λ to every test, where each test i is significant if and only if $S(x_i) \geq \lambda$ for a given S and λ . For example suppose that a standard two-sided t-test is applied to each x_i the statistic is $S(x_i) = \left| \frac{\bar{x}_i}{s_i/\sqrt{n}} \right|$, where \bar{x}_i is the sample mean and s_i is the sample standard deviation of x_i [44].

Finally, based on the previous components the ODP is defined to be the multiple-testing procedure that maximizes ETP for each fixed EFP among all STPs

$$S_{(ODP)}(x) = \frac{g_{m0+1}(x) + g_{m0+2}(x) + \dots + g_m(x)}{f_1(x) + f_2(x) + \dots + f_{m0}(x)} \quad (3.3)$$

where m stands for significance tests performed on observed data sets x_1, x_2, \dots, x_m , and f_i is the null density and g_i is the alternative density of significance test i . Null hypothesis i is rejected if and only if $S(x_i) \geq \lambda$ for some $0 \leq \lambda < \infty$. For each fixed λ this procedure yields the maximum number of expected true positive results ETP among all simultaneous thresholding procedures that have an equal or greater number of expected false positive results EFP. Although it seems that the ODP requires the knowledge of the true distribution of each significance test and that is not feasible in practice, it can be estimated from the observed data for each test since. The data reflect their true distribution either null or alternative thus, the ODP can be estimated effectively regardless of any prior knowledge of the tests' distributions [44].

3.1.1.3 The empirical Bayes moderated t-statistic (eBayes)

The eBayes ranks genes by testing whether all pairwise contrasts between different outcome-classes are zero. It is applied to extract information across genes thus making the final analyses more stable even for experiments with limited number of arrays. Moderated t-statistics lead to p-values with increased degrees of freedom for the individual variances hence, reflecting the greater reliability associated with the smoothed standard errors [45].

This approach requires a design matrix and a contrast matrix to be specified. The design matrix illustrates the different RNA targets that have been hybridized to the arrays. The contrast matrix facilitates the combination of the coefficients defined by the design matrix into contrasts of interest where each contrast corresponds to a comparison of interest between the RNA targets. However, during simple experiments the contrast matrix may not be explicitly specified [45].

The whole algorithm is implemented in three steps: During the first step, a linear model is fitted to the data to estimate their variability. Each row of the resultant design matrix corresponds to an array and each column corresponds to a coefficient. For one-channel data, the number of coefficients equals to the number of distinct RNA sources. In the second step, the contrast matrix allows the comparison of the fitted coefficients in as many ways as the questions to be answered, regardless of the number of the coefficients. Finally, in the third step the posterior odds are reformulated in terms of a moderated t-statistic, where posterior residual standard deviations are utilized instead of ordinary standard deviations. The moderated t-statistic as opposed to posterior odds reduces the number of hyperparameters necessary for the hierarchical model. Moreover, it follows a t-distribution with increased degrees of freedom, and may accommodate tests for more than two contrasts with the aid of moderated F-statistics. Linear Models for Microarray Data (Limma) is an r-package, which implements this statistic [46].

From a mathematical point of view, we may describe the prior description as follows: suppose that we have a set of n microarrays with a response vector of log-intensities $y_g^T = (y_{g1}, \dots, y_{gn})$ for the g th gene. The probes should be suitably normalized to produce an expression summary, represented here as y_{gi} , for each gene on each array. We assume a linear model $E(y_g) = X\alpha_g$ where X is a design matrix and α_g is a coefficient vector, and estimated covariance matrices $\text{var}(y_g) = W_g \sigma_g^2$ where W_g is a known non-negative definite weight matrix that may contain diagonal weights with zero value. Then,

the contrasts of interest are given by $\hat{\beta}_g = C^T \hat{\alpha}_g$ where C is the contrast matrix and $\text{var}(\hat{\beta}_g) = C^T V_g C s_g^2$ are the estimated covariance matrices. The posterior values for the residual variances are given by $\tilde{s}_g^2 = \frac{f_0 s_0^2 + f_j s_j^2}{f_0 + f_j}$ where f_j is the residual degree of freedom for the j th gene. The moderated t -statistic is

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{s_g \sqrt{u_{gj}}} \quad (3.4)$$

that follows an approximate t -distribution on $f_0 + f_j$, and u_{gj} is the j th diagonal element of $C^T V_g C$ [46].

3.1.1.4 The Significance Analysis of Microarrays (SAM)

The SAM approach based on the analysis of random fluctuations in gene expression data. Specifically, the fluctuations of gene expression appear to be gene specific even for a given expression level. To elucidate this observation SAM was developed to take into account the ratio of change in gene expression to standard deviation in the data for a specific gene. The “relative difference” of the gene expressions for a given gene is

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0} \quad (3.5)$$

where $\bar{x}_I(i)$ and $\bar{x}_U(i)$ are the average expression levels of gene (i) in conditions I and U. In addition, the “gene-specific scatter” $s(i)$ is the standard deviation of repeated measurements

$$s(i) = \sqrt{a \left\{ \sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_n [x_n(i) - \bar{x}_U(i)]^2 \right\}} \quad (3.6)$$

where \sum_m and \sum_n are summations of the expression measurements in conditions I and U, and $a = (1/n_1 + 1/n_2) / (n_1 + n_2 - 2)$, where n_1 and n_2 are the number of measurements in the conditions I and U respectively. Moreover, to compare the “relative difference” across all genes the distribution of $d(i)$ should be independent of gene expression values. Therefore, a small positive constant value s_0 was added to the denominator of $d(i)$ such that to minimize the coefficient of variation [47].

Furthermore, the engagement of balanced permutations among the available samples was employed to alleviate potential confounding effects between the two conditions so as to conclude to robust “relative difference” scores per gene. Then, the genes were ranked according to their score $d_p(i)$ in descending order. The expected relative difference, $d_E(i)$ is defined as the average over the N balanced permutations as $d_E(i) = \sum_p d_p(i) / N$. Comparing the “relative difference” and the “expected relative difference” per gene we may characterize a gene as significant if the subtraction of those two values exceeds an adjustable threshold Δ (delta) [47].

In order to control the number of falsely detected significant genes, SAM employs horizontal cutoffs representing the smallest $d(i)$ among the significantly induced genes and the least negative $d(i)$ among the significantly repressed genes. The number of genes that exceeds the horizontal cutoffs for induced and repressed genes per

permutation corresponds to the falsely significant genes. Hence, the estimated number of falsely significant genes is the average of the number of genes called falsely significant from all permutations. Respectively, the FDR is the ratio of the falsely significant genes to the significant genes for a given. As Δ decreases, the number of genes called significant increases but at the cost of an increasing FDR. SAM was first applied to analyze the transcriptional response of lymphoblastoid cells to ionizing radiation (IR) [47].

3.1.1.5 Student *t*-test and Signal-to-Noise Ratio

In our univariate analysis we engaged two common statistical approaches, the student *t*-test and the Signal-to-Noise Ratio. Both of these methods are included in the Comparative Marker Selection suite [48], which in turn is part of the GenePattern software [49]. This module is freely available and allows users to apply and compare different methods of computing significance for each marker gene, for example p-value, q-value, FDR, FWER, e.t.c, a viewer to assess the results, and a tool to create derivative datasets and marker lists based on user-defined significance criteria. In our analysis we engaged the “rank” estimate which derives from the value of the test statistic.

Concerning the *t*-test, the default method is the two-paired test which assumes that differentially expressed genes can be up-regulated in either class. It is calculated by the

$$\text{formula } t\text{-test} = \frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \quad (3.7)$$

where μ is the mean value, σ is the standard deviation, n is the number of samples, and A is the one class and B is the second class in a binary case. The SNR statistic is given as the subtraction of the class means divided by the sum of their standard

$$\text{deviations } SNR = \frac{\mu_A - \mu_B}{\sigma_A + \sigma_B} \quad (3.8).$$

3.1.1.6 Correlation-adjusted *t*'-scores (cat)

The aforementioned approaches disregard the potential correlations among genes, which may have a negative influence on the gene ranking and the subsequent classification results. In general, there are three possible strategies to follow when dealing with the correlation among genes. During the first approach, the conventional *t*-scores are computed and then the correlation structure is taken into account. In the second strategy, the correlation structure model is generated, and the inferences about genes' significance are based on that. Finally, the third approach tries to combine *t*-scores with the estimated correlations to form a new gene-wise statistic. The proposed “correlation-adjusted *t*'-scores”, or for short “cat” scores are influenced by the third strategy [50].

The cat scores bear in mind the close relationship between gene ranking and feature selection for class discrimination. Therefore, it is exploited a close link between gene ranking and two-class linear discriminant analysis (LDA). If there are two distinct class labels, $K = 2$, the difference $\Delta^{LDA}(x) = d_1^{LDA}(x) - d_2^{LDA}(x)$ between the discriminant scores of the classes results in the following prediction rule:

if $\Delta^{LDA} \geq 0$ then the assigned class label is 1, otherwise is 2. We may rewrite the $\Delta^{LDA}(x)$ as $\Delta^{LDA}(x) = \omega^T \delta(x) + \log\left(\frac{\pi_1}{\pi_2}\right)$ (3.9)

where ω is a weight vector $\omega = P^{-1/2}V^{-1/2}(\mu_1 - \mu_2)$ (3.10)

and $\delta(x)$ is a vector-valued distance function $\delta(x) = P^{-1/2}V^{-1/2}\left(x - \frac{\mu_1 + \mu_2}{2}\right)$ (3.11), with P are the correlations and V the variances of the diagonal matrix. The “cat” scores can be defined as a vector proportional to the feature weight vector ω as follows:

$$\begin{aligned} \tau^{adj} &\equiv \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2} \omega \\ &= P^{-1/2} \times \left\{ \left(\frac{1}{n_1} + \frac{1}{n_2}\right) V \right\}^{-1/2} (\mu_1 - \mu_2) \\ &= P^{-1/2} \tau \end{aligned} \quad (3.12)$$

The vector τ includes the gene-wise t -scores, and n_k stands for the number of samples in class k . To sum up, it would be quite accurate to state that the “cat” score is the natural and intuitive extension of fold change and t -score, as illustrated in Figure 3.2. Indeed, whilst the t -score represents the standardized mean difference $\mu_1 - \mu_2$ with constant $c = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2}$, the “cat” score is the standardized as well as the decorrelated mean difference, with factor $P^{-1/2}$ responsible for the decorrelation [50].

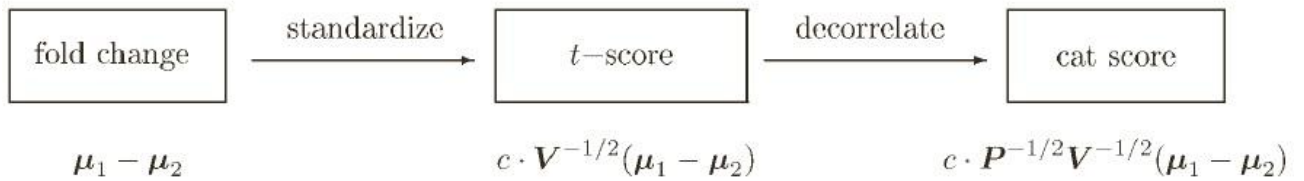


Figure 3.2: The relationship between fold change, t-score and cat score [50]

3.1.2 Wrapper methods

As opposed to filter methods, the wrapper methods select a good subset of genes in conjunction with the classification model. In particular, there is a search procedure that generates possible feature subsets, which then are evaluated by training and testing on a specific classification model. In other words, we have a search algorithm “wrapped” around a specific classification algorithm to determine a “good” subset of genes. Since the number of inspected genes varies usually from tens to hundreds heuristic search methods are employed to facilitate the search for an optimal subset. Those methods fall into two categories: deterministic and randomized search algorithms. The main

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

advantage of wrapper methods is their intrinsic ability to take into account gene dependencies through the tailored relationship of a search algorithm to a specific classification model. However, this notion has a higher risk of overfitting than filter techniques and depending on the size of the feature space is computationally intensive [5].

3.1.3 Embedded methods

Similar to wrapper techniques the embedded approaches are tailored to a specific learning algorithm with the exception that the features algorithm is enclosed into the classifier construction. Hence, this category of methods employs a combined space of feature subsets and hypotheses, which incorporates the interaction with the classification model. This common characteristic between embedded and wrapper methods can be seen as an alternative way to carry out multivariate gene selection analysis. Indeed, incorporating the classifier's bias into the gene selection process enhance the construction of accurate classifiers. Compared to wrapper methods the embedded algorithms are far less computationally intensive [5].

In relation to the available applications, the random forests algorithm is a distinctive example of an embedded algorithm where a classifier combines several single decision trees to calculate the importance of each gene. Other examples use the weights of each feature in linear SVMs or logistic regression classifiers. These weights have been computed through a multivariate analysis and reflect the relevance of each gene. Those genes with small weights are filtered from the rest of the analysis. However, mainly the wrapper and to a lesser degree the embedded methods have failed to draw the attention in the domain because of being highly sophisticated compared to filter methods. Nevertheless, a sensible tactic is to employ a univariate filter method to reduce the features subspace prior to any wrapper or embedded analysis procedure, something that restricts the computational time to reasonable levels [5].

3.1.4 Dimension reduction methods

Dimension reduction or feature extraction is an alternative to feature selection in relation to the "curse" of dimensionality problem. Contrary to feature selection, those methods utilize all the available genes and project them onto a low-dimensional space which also facilitates visual representation. Besides, the resultant components of the projection usually provide us with information of the intrinsic structure of the data. Though, a point of criticism against those methods is the questionable interpretability particularly when scientists are interested in specific genes [51].

Those methods are classified into linear and non-linear as well as supervised and unsupervised. Employing a supervised method rather than an unsupervised is a preferable choice since the construction of the projection components takes into account the class information. On the other hand, non-linear methods are more computationally expensive than linear methods and additionally lack of robustness, therefore they are not adopted microarray classification studies. So far, the Partial Least Squares (PLS) method is a supervised linear dimension reduction method, which performs even if the number of genes far exceeds the number of available samples. Another method that is in close relation to PLS is the between-group-analysis (BGA). Finally, a well known dimension reduction method is the Principal Components Analysis (PCA), which is an unsupervised linear method, thus not recommended for classification problems [51].

3.1.4.1 Partial Least Squares

Partial Least Squares is a multivariate regression method that originally developed for the chemometrics field and it is particularly suitable to predict a univariate or multivariate continuous response from a large number of continuous predictors. The principal notion behind PLS is to find uncorrelated linear transformations of the original predictor variables such that have high covariance with the response variables. In the sequel, those linear transformations can be engaged as predictors in conventional linear regression models to predict the response variables. The profound benefit of this idea is the efficient performance of linear regression since the produced components are always much smaller in number than the original variables p and irrespective of the available observations [51].

In particular, suppose we have a train set L with known class labels and a test set T with labels that have to be predicted. The corresponding data matrices are X_L and X_T and the class labels vector is Y_L . We may formulate a classification method as a function δ_{PLS} of X_L and Y_L and the vector of predictors $x_{new,i}$ corresponding to the i th observation of the test set:

$$\begin{aligned} \delta_{PLS}(., X_L, Y_L): \mathbb{R}^p &\rightarrow 1, \dots, K \\ x_{new,i} &\rightarrow \delta_{PLS}(x_{new,i}, X_L, Y_L) \end{aligned} \quad (3.13)$$

where K is the dimension, and $K = 2$ for binary problems [51].

This function involves two steps; dimension reduction and linear discriminant analysis (LDA).

1. In the dimension reduction step, we find m appropriate linear transformations Z_1, \dots, Z_m of the vector of predictors x . However, the appropriate number of m is user defined and there is no widely accepted method. A simple yet effective approach based on cross validation is proposed by Boulesteix where only the train set is used to determine the m PLS components. The classifier δ_{PLS} is build using a percentage $\alpha\%$ of the available observations and applied to the remaining observations trying several values for m . The procedure is repeated for N_{run} runs and an error rate is computed for each one. After the completion of the N_{run} runs, the mean error rate for each m value. The m value that minimizes the error rate is denoted as m_{opt} and is the one used to predict the class labels of the test set. Then the SIMPLS algorithm is employed to determine the $p \times 1$ vectors $\alpha_1, \dots, \alpha_m$ which are used to construct the linear transformations Z_1, \dots, Z_m :

$$\begin{aligned} Z_1 &= \alpha_1^T x, \\ \dots &= \dots, \\ Z_m &= \alpha_m^T x. \end{aligned} \quad (3.14)$$

Thus, if A denotes the $p \times m$ matrix containing the vectors $\alpha_1, \dots, \alpha_m$ in its columns, the matrix with the new components of the train set L is obtained as $Z_L = X_L A$.

2. In the second step, LDA, the new components Z_1, \dots, Z_m are employed as predictor variables, and the test set matrix Z_T of new components is computed as $Z_T = X_T A$. The description of LDA in brief has as follows: for p predictor variables the random vector $x = (X_1, \dots, X_p)^T$ is assumed to follow a multivariate normal distribution within class $k (k = 1, \dots, K)$ with mean μ_k and covariance matrix Σ_k . In LDA it is assumed that Σ_k is the same for all k classes i.e. $\Sigma_k = \Sigma$. Therefore, using estimates $\hat{\mu}_k$ and $\hat{\Sigma}$ instead of μ_k and Σ , the maximum-likelihood discriminant rule assigns the i th new observation $x_{new,i}$ to the corresponding class through the following formula:

$$\delta(x_{new,i}) = \arg \min_k (x_{new,i} - \hat{\mu}_k) \hat{\Sigma}^{-1} (x_{new,i} - \hat{\mu}_k)^T \quad (3.15).$$

The PLS method is performed by gene selection by ranking genes according to the BSS/WSS -statistic, where BSS stands for the between-group sum squares and WSS the within-group sum of squares. Hence, for gene j the BSS/WSS -statistic is computed as

$$BSS_j / WSS_j = \frac{\sum_{k=1}^K \sum_{i:Y_i=k} (\hat{\mu}_{jk} - \hat{\mu}_j)^2}{\sum_{k=1}^K \sum_{i:Y_i=k} (x_{ij} - \hat{\mu}_{jk})^2} \quad (3.16)$$

where $\hat{\mu}_j$ is the sample mean of X_j and $\hat{\mu}_{jk}$ is the sample mean of X_j within class $k (k = 1, \dots, K)$. The genes with the highest BSS/WSS -statistic value are ranked first and considered as significant. Though, there is no well-established criterion to define the number of significant genes to be chosen [51].

3.1.4.2 Between-Group Analysis

BGA is a multiple discriminant approach that can be applied irrespective of the number of genes and available samples. The basic idea behind BGA is to ordinate groups of samples rather than individual samples with the intention of separating them maximally in some space. For N groups we find $N-1$ eigenvectors or axes such that to maximize the between group variances. Then with the aid of conventional ordination techniques such as Correspondence Analysis (COA) or PCA the individual samples are projected and plotted along those axes. Similarly, new samples are placed on those axes and classified on an axis-by-axis basis or by proximity to the group centroids. The combination of BGA with COA is quite effective since it allows us to inspect the association between the genes that discriminate group of samples with the grouped samples and it is the approach employed in our analysis [52].

Suppose we have a raw data table (N) of gene expression data with rows I representing genes and columns J representing microarray samples, and elements n_{ij} . The COA method requires non-negative elements (usually integers), therefore it is obligatory to add a constant to all values if necessary. The row sums and column sums of N are denoted as n_{i+} and n_{+j} respectively. The sum of all elements is denoted as n_{++} . With r_i it is denoted the weight or the relative contribution of each gene i to the total variation of the data set and is calculated as $r_i = n_{i+}/n_{++}$, whereas the relative contribution of sample j is declared as c_j and is calculated as $c_j = n_{+j}/n_{++}$. Likewise, the relative contribution of each element of N to the total variation in the data set is denoted as p_{ij} and is calculated as $p_{ij} = n_{ij}/n_{++}$. Those definitions produce two vectors

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

R and C of length I and J and one matrix $I \times J$. All these are converted into a table X of $I \times J$ with x^2 values using the formula $x_{ij} = (p_{ij} - r_i c_j) / \sqrt{r_i c_j}$. This table X incorporates the associations between genes and samples and is the one used to produce the correspondence analysis. The overall association between all genes and samples is given by the overall x^2 value for the data set (x_{++}) which is also the total value of all elements of X . This overall x^2 value is then decomposed by COA into components for each gene and sample along each of K eigenvectors, where K is $\min(I-1, J-1)$. Those eigenvectors are ranked by their eigenvalues where their sum equals the overall x^2 value of the data set. In relation to the eigenvectors, the method used to derive them is general singular value decomposition in which a matrix B is calculated with the formula $B = D_c^{1/2} X D_r X^T D_c^{1/2}$. In this formula the $D_c^{1/2}$ is a $J \times J$ matrix with the square roots of the elements of the C vector along the diagonal and zeros elsewhere, the D_r is a $I \times I$ matrix with the elements of the R vector along the diagonal and zeros elsewhere, and B is a $J \times J$ matrix which is diagonalized to produce J eigenvalues (where at least one of which will be zero) and eigenvectors [52].

In the case of two groups, the results of the analysis will be a single vector with the positions of all samples and genes. The most significant genes are those that separate the groups and are located at the end of the axes i.e. have the most extreme co-ordinates.

3.1.4.3 Principal Components Analysis

PCA is one of the oldest dimension reduction approaches where its main initiative is to reduce the dimensions of a data set with a large number of interrelated variables, whilst preserving the intrinsic variation in the data set to the best possible degree. This idea is accomplished with the transformation of the original variables to a new set of variables, the principal components, which are uncorrelated and ordered by the degree of variation. In other words, the first few of the new components incorporate most of the information related to the original variables [53].

Assume there is a vector x with p random variables, and that the variances of the p variables as well as the structure of the covariances or correlations between those variables are of interest. Looking at the p variances and all of the $\frac{1}{2} p(p-1)$ correlations or covariances it will not be very helpful, except the number of p variables is not that large or the structure is straightforward. An intriguing approach is instead of including all the p variables to an analysis to produce a few new variables ($\ll p$) such that preserve most of the intrinsic information given by these variances and correlations or covariances. PCA's focus is on variances rather than on covariances and correlations [53].

The initial step is to search for a linear function $\alpha_1^T x$ of those elements of x with maximum variance, where α_1 is a vector of p constants $\alpha_{11}, \alpha_{12}, \dots, \alpha_{1p}$, so that

$$\alpha_1^T x = \alpha_{11}x_1 + \alpha_{12}x_2 + \dots + \alpha_{1p}x_p = \sum_{j=1}^p a_{1j}x_j \quad (3.17).$$

Similarly, a linear function $\alpha_2^T x$ having maximum variance and been uncorrelated with $\alpha_1^T x$ has to be found. This procedure continuous until a linear function $\alpha_k^T x$ with

maximal variance and uncorrelated with the previous linear functions, $\alpha_1^T x, \alpha_2^T x, \dots, \alpha_{k-1}^T x$, has been found, too. That final variable, $\alpha_k^T x$, represents the k th PC. Although k can be equal to p , the Holy Grail is to conclude to m PCs, where $m \ll p$ [53].

The second step involves the procedure used to find those PCs. Suppose that the random vector x has a known covariance matrix Σ whose (i, j) th element reflects the known covariance between the i th and j th elements of x when $i \neq j$ and the variance of the j th element of x when $i = j$. It is proved that for $k = 1, 2, \dots, p$, the k th PC is given by $z_k = \alpha_k^T x$ where α_k is an eigenvector of Σ corresponding to its k th largest eigenvalue λ_k . Moreover, if α_k has chosen to be of unit length ($\alpha_k^T \alpha_k = 1$), then the variance of z_k is $\text{var}(z_k) = \lambda_k$ [53].

Finally, the form of the PCs has to be derived. Consider the $\alpha_1^T x$ linear function. The vector α_1 maximizes $\text{var}[\alpha_1^T x] = \alpha_1^T \Sigma \alpha_1$. In order to maximize $\alpha_1^T \Sigma \alpha_1$ subject to $\alpha_1^T \alpha_1 = 1$, the usual approach is to engage the technique of Lagrange multipliers such that $\text{var}[\alpha_1^T x] = \alpha_1^T \Sigma \alpha_1 - \lambda(\alpha_1^T \alpha_1 - 1)$, where λ is the Lagrange multiplier. Thus, the quantity to be maximized is $\alpha_1^T \Sigma \alpha_1 = \alpha_1^T \lambda \alpha_1 = \lambda \alpha_1^T \alpha_1 = \lambda$ and therefore λ must be the possibly larger to determine which of the p eigenvectors maximize the variance of the $\alpha_1^T x$. In general, the k th PC of vector x is $\alpha_k^T x$ and $\text{var}(\alpha_k^T x) = \lambda_k$, where λ_k is the k th largest eigenvalue of Σ , and α_k is the corresponding eigenvector [53].

PCA has been extensively used in bioinformatics studies, because of its computational simplicity and satisfactory statistical properties. Particularly in gene expression studies, not only manages to reduce the dimensionality of high-throughput measurements but also create subsets of genes, which in turn achieve satisfactory classification performance [37].

3.1.5 Hybrid methods

3.1.5.1 Hybrid system for marker Gene selection

The Hybrid system for marker Gene selection (HykGene) is a hybrid approach that combines gene ranking and cluster analysis. HykGene aims at selecting a limited number of non-redundant though highly discriminative genes. For that scope, it follows a three-step procedure: in the first step, a feature filtering algorithm is applied on a training set, in the second step, hierarchical clustering is performed on the top-ranked genes and a dendrogram is produced, whereas in the third step, a sweep-line algorithms is utilized to discern clusters from the dendrogram from which marker genes, one per cluster, are selected through clusters' collision. The method also includes a final step where only the selected marker genes are used to classify a test set [10].

During the first step, the method employs a filtering technique that ranks genes according to a calculated score. For this purpose, HykGene engaged Relief-F, Information Gain, and χ^2 statistic:

Relief-F: this algorithm first draws random instances, then computes their nearest neighbors and finally adjusts a feature-weighting vector that augments those features that best discriminate instances from neighbors of different classes. Particularly, the weight assigned per gene f is calculated through the formula $w_f = P(\text{different value of } f \mid \text{different class}) - P(\text{different value of } f \mid \text{same class})$ (3.18).

Information Gain: it measures the amount of information in bits relevant to class prediction based on the value (discretized) of a feature. Suppose that $\{c_i\}_{i=1}^m$ is the set of classes, and V the set of possible values for feature f . Then, the Information Gain of a feature f is given by the formula

$$G_{(f)} = -\sum_{i=1}^m P(c_i) \log P(c_i) + \sum_{v \in V} \sum_{i=1}^m P(f=v) P(c_i | f=v) \log P(c_i | f=v) \quad (3.19).$$

χ^2 -**statistic:** it tests the independence between two paired variables, in that case between feature f and class c . The χ^2 -statistic is estimated through the following type:

$$\chi^2(f) = \sum_{v \in V} \sum_{i=1}^m \frac{[A_i(f=v) - E_i(f=v)]^2}{E_i(f=v)} \quad (3.20), \text{ where } A_i(f=v) \text{ is the number of instances}$$

in class c_i with $f=v$, $E_i(f=v)$ is the expected value of $A_i(f=v)$ and is calculated with $E_i(f=v) = P(f=v)P(c_i)N$, and N is the total number of instances [10].

After the completion of the first step, the genes are ranked accordingly and a clustering approach, hierarchical clustering (HC) is engaged in the second step. Specifically, the top 50 or top 100 genes from the whole dataset are kept for clustering analysis through HC in order to conclude to homogeneous clusters. Since the dendrogram representation of HC clustering is far from definite regarding the exact number of clusters, the engagement of a sweep-line algorithm on the training samples deals with this issue in the third step. This method tries to discern representative genes from each cluster and then collapse the clusters onto those genes. With the aid of a cross-validation technique (LOOCV), all possible ways to extract clusters from the dendrogram are evaluated before concluding to the best possible number of clusters and consequently to the minimum set of representative genes. The gene with the minimum sum of squares of distances to all other genes within the cluster is characterized as representative. In the final step, the discrimination power of those marker genes is evaluated on the test set data. For this purpose, HykGene employs one of four possible classifiers: k-nearest neighbor (k-NN), linear support vector machine (SVM), C4.5 decision tree, and Naive Bayes (NB) [10].

3.2. Clustering

Clustering, or else unsupervised learning due to lack of prior knowledge, is one of those machine learning techniques that engaged widely to microarray analysis. The main objective of cluster analysis is to classify objects in a data set into meaningful classes according to a pre-specified similarity measure. This data summarization is solely based on the internal structure of the data, a data driven approach, where the researcher is free of making any data assumptions about sample size, data quality or experimental design. Clustering has several applications in gene expression data analysis including data reduction and visualization, inferring functions from clusters of genes, detecting classes or sub-classes of diseases or even predicting the categorization of new samples [20, 54, 55]. According to Milligan, cluster analysis entails seven critical steps: 1) Clustering element selection; 2) Clustering variable selection; 3) Variable standardization; 4) Choosing a measure of association (dissimilarity/similarity measure; 5) Selection of clustering method; 6) Determining the number of clusters; 7) Interpretation, validation and replication [55].

3.2.1 Element selection

Clustering is a data-driven approach where the data set elements should represent the underlying cluster structure. In order to conclude to distinct and reliable clusters, outliers (data points ranging outside the general region) are better to be excluded unless they form a single cluster [55].

3.2.2 Variable selection

Selecting an appropriate set of variables, enough and representative number of variables to the underlying data structure, imposes a significant impact on the subsequent cluster analysis. Indeed, suppose we generate four two-variate normal distribution data with mean vectors $\mu_1 = (0,0)'$, $\mu_2 = (0,2)'$, $\mu_3 = (2,0)'$, $\mu_4 = (2,2)'$ respectively, and the same covariance matrix $\Sigma = I$. Consequently, there are four distinct clusters, Figure 3.3(a). In case we deliberately omit variable 2, we notice that only two clusters rather than four can be distinguished, Figure 3.3(b). On the other hand, including unnecessary variables, noise or masking variables, might also dramatically deteriorate cluster discovery [55].

3.2.3 Variable standardization

The variable standardization usually involves the transformation of raw data into a more normal-like distribution form. However, such an intervention will definitely alter the relative distances between pairs of objects, hence modifying the underlying cluster structure of the data. As a consequence, it is of great importance for the outcome of the cluster analysis to retain the original structure by selecting an appropriate standardization measure [55].

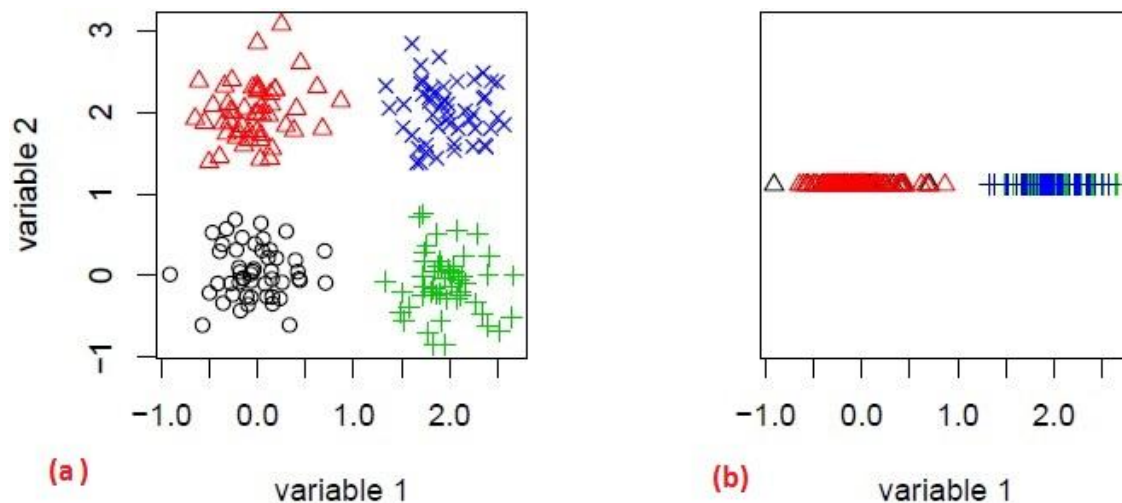


Figure 3.3: The impact of feature selection in cluster analysis

3.2.4 Selecting a measure of association (similarity/dissimilarity)

In cluster analysis, objects are assigned to the same cluster according to a similarity/dissimilarity measure. Such a metric should reflect the data characteristics necessary to differentiate the present clusters. There are numerous of measures relevant to the types of variables (e.g. interval-scaled variable, nominal variable, ordinal

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

variable or mixed data). Gene expression data belong to the continuous data type, and a commonly used dissimilarity measure is the Euclidean distance. During this measure each gene is considered as a point in multidimensional space, where each axis represents a separate biological sample and the coordinate on each axis is the amount of gene expression in that sample. However, Euclidean distance is vulnerable to unnormalized data and to negative gene associations, leading to missed correlation measurements [55].

A further dissimilarity measure is the Pearson correlation coefficient which, measures the linear dependence between two variables (genes) treated as vectors of measurements. This metric is based on two assumptions that relatively stand in microarray gene expression values. In particular, it assumes that genes' intensities follow the normal distribution, which is not the case even after applying normalization methods. Microarray data is best characterized as following a normal-like distribution. The other assumption states that genes interaction follows the underlying linear model. Though in reality, a specific gene may regulate other genes even if it is not at its peak expression values. An additional shortcoming for Pearson correlation is its sensitivity to outliers. On the other hand, mutual information is another dissimilarity measure calculated on discrete expression values (for example, gene values are discretized into 'low' and 'high' or 'low', 'medium', and 'high' states). This metric considers each expression-level measurement equally, regardless of the actual intensity values, and is therefore robust to outliers. By setting a threshold we may keep only the high mutual information scores of gene-pairs, reflecting the non-random association. However, there is always a possibility of noise high mutual information scores as well as of novel hypotheses for lower mutual information scores [55].

3.2.5 Selection of clustering method

One of the most critical steps prior to cluster analysis is the selection of an appropriate clustering method. Ideally, the successful candidate should fulfill four prerequisites. First, it should be designed to recover the suspected clusters in the data. Second, the method should effectively recover the structures for which it was designed. Third, it should be robust against errors in data and finally, there should be the available software package implementing the method [55].

3.2.6 Determining the number of clusters

Usually most of the clustering algorithms expect a predefined number of possible clusters lying in the data structure. Thus, selecting the number of clusters or partitions present in a dataset under analysis is an ordinary yet trivial task faced by any researcher. Clustering algorithms provide either little information about the potential number of clusters in the data, hierarchical methods, or no information, nonhierarchical algorithms. So far, several approaches have been developed to address this necessitate by determining an accurate as possible estimate of the number of clusters in a data set. The closer the number to the true number of clusters the more efficient the clustering result.

In relation to the hierarchical methods, where a number of possible solutions is provided somehow ranging from n clusters to one cluster, those cluster determination techniques when applied to hierarchical clustering results are referred to as stopping criteria. Apart from the hierarchical methods, such criteria can be also applied to nonhierarchical methods. However, engaging such a stopping criterion in a cluster analysis entails the

possibility of incorrect assumptions about the actual number of clusters in the underlying data set structure. Those incorrect assumptions can be either positive, meaning more clusters than the actual number or negative when fewer clusters are indicated. According to the area of the clustering problem under analysis the severity of those two errors might be considered differently. However, determining fewer clusters than what actually exists in a data set is generally considered a more serious error, since valuable information is ignored through merging of clusters [56].

Therefore, numerous of strategies have been proposed towards the more accurate estimation of the underlying structure. According to Tibshirani et al. indices comparison survey, the index of Krzanowski and Lai achieved excellent discriminatory results, and is the one selected to be part of our methodology [55]. In particular, we apply the index of Krzanowski and Lai [57] as included in the 'clusterSim' package [58] to determine the number of clusters. Krzanowski and Lai is defined by

$$DIFF(k) = (k-1)^{2/p} W_{k-1} - k^{2/p} W_k \quad (3.21)$$

when choosing the number of clusters (k) to maximize the quantity

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right| \quad (3.22)$$

where W_k denotes the within-cluster sum of squared errors.

3.2.7 Interpretation, validation and replication

The holy grail of cluster analysis is to produce interpretable classification results with respect to a specific case study. Graphical representations may play a useful role in interpreting the resultant cluster structure in conjunction with special knowledge and expertise in the inspected area of study. Ideally, a clustering method is considered as good if objects grouped in the same cluster are similar to each other and different from objects in other clusters. There are several approaches, external or internal, that intend to validate the performance of a cluster analysis. External criteria include the Rand index, the adjusted Rand index, the Fowlkes and Mallows index and the Jaccard index. In general, these approaches evaluate the clustering results based on external classification information (as it might be with simulated data) independent of the clustering procedure, by computing the goodness-of-fit between the data and the partitioning result. Replication entails a further validation test where the clustering results should be also found in replicated samples [55].

3.3 Clustering algorithms

Gene expression data can be clustered either based on genes or samples. The genes-based clustering considers the genes as the objects and the samples as the features contrary to the sample-based approach where the samples are treated as the objects and the genes as the features. In the first approach we look for coexpressed genes that imply coregulation and cofunction whereas in the second we anticipate to identify particular phenotypes. In the current dissertation we are interested in identifying clusters of genes therefore in the rest of this study we will discuss clustering from this perspective. Regarding the field of gene expression analysis, two well known clustering

algorithms have been widely used: the hierarchical clustering and the self-organizing maps. Though, the advent of new algorithms with improved characteristics that promise advance cluster discrimination has drawn our attention and as such we will present the Affinity Propagation algorithm.

3.3.1 Hierarchical clustering

Hierarchical clustering is one of the most commonly used unsupervised methods that iteratively groups genes with similar expression patterns to clusters. It is identified by the graphical representation of a tree, called dendrogram, where leaves represent all genes, Figure 3.4. Each branch of the tree may link two leaves (genes), two other branches, or one leaf with another branch. Any new gene is added through a connection to the branch that most resembles. Not to mention that branches are of different lengths representing the degree of similarity. In particular, shorter branch lengths represent increased similarity between genes or branches whereas longer branch lengths indicate increased dissimilarity. As a result, a dendrogram is not always symmetric. In relation to the number of nested clusters as described in a dendrogram, it is specified by cutting the dendrogram at some level [59].

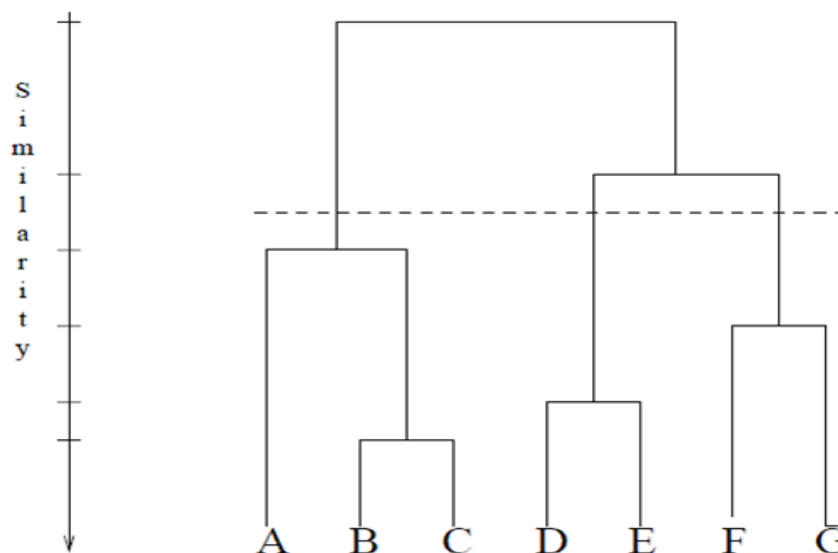


Figure 3.4: A single-link agglomerative clustering dendrogram [60]

The way that a dendrogram is formed specifies two subcategories in hierarchical clustering; the agglomerative and the divisive. Regarding the agglomerative method, it is considered as a bottom-up approach where initially each object is treated as an individual cluster and iteratively the closest pairs of clusters are merged until one final cluster. On the other hand, the divisive approach is a top-down technique meaning that initially the method starts with all objects included in one cluster and then iteratively concludes to one cluster per object [61].

Between those two approaches, the most preferred among many biologists during gene expression analysis is the agglomerative approach and specifically the one as applied by Eisen et al. called UPGMA (Unweighted Pair Group Method with Arithmetic Mean). According to this method, each cell of a gene expression matrix is represented by a color in accordance with the measured fluorescence ratio. Then the rows of the matrix are reordered following the hierarchical dendrogram structure and a defined node-ordering rule. As a result, the gene expression matrix has been converted into a colored

table where large adjacent patches of color represent clusters of genes that share similar expression patterns across multiple conditions [61].

The key advantage of hierarchical clustering is the graphical representation of the overall similarities in expression patterns for a whole data set, hence allowing the user to have quickly a preliminary impression about the data distribution [59, 61]. Albeit hierarchical clustering suffers from several drawbacks including, lack of robustness, nonuniqueness and inversion problems that hinder interpretation of the hierarchy. Moreover, due to the deterministic nature of hierarchical clustering some genes may be erroneously grouped based on local decisions, without the ability to reevaluate the clustering [62].

3.3.2 Self-Organizing Maps (SOMs)

Self-organizing maps is a clustering algorithm that likewise hierarchical clustering produces a visual representation (typically two-dimensional) of gene expression patterns of a data set. Though, SOMs not only are significantly different to hierarchical clustering algorithm but also more suitable for clustering analysis of gene expression data. Indeed, SOMs have been successfully applied on gene expression data studies in comparison to hierarchical clustering and achieved notably clustering results for both accuracy and robustness. Some of the key characteristics of SOMs include the ability of someone to impose partial structure on the clusters contrary to the rigid structure of hierarchical clustering that entails visualization simplicity and straightforward interpretation. Moreover, their computational burden is lower than that of dendrograms since SOMs do not require complete pairwise comparisons [59, 62].

Regarding the underlying algorithm itself, first the data samples define the multidimensional space, each sample is considered a separate dimension, and then genes are represented as points (nodes) in that multidimensional space using their expression levels as coordinates, where the i th coordinate represents the expression level in the i th sample. Therefore a SOM has a set of nodes with a two-dimensional lattice topology, for example 3×2 in Figure 3.5, and a distance function $d(N1, N2)$ on the nodes. The initial lattice mapping f_0 of the nodes is arbitrary and then iteratively adjusted. Then, a data point P is selected and the node closest to P is identified (N_p) with the aid of dissimilarity measures, typically Euclidean distance. The (N_p) is moved closest to P , whereas the other nodes are also moved closer to P but with different amounts of distance following the rule $f_{i+1}(N) = f_i(N) + \tau(d(N, N_p), i)(P - f_i(N))$, where τ is the learning rate that decreases depending on the distance between node N and N_p and the iteration number i until it becomes zero. The number of iterations range between 20,000 and 50,000. Regarding the benchmark point P , it is determined once by random ordering of the n nodes and recycled as needed [62].

Nevertheless, SOMs have also certain shortcomings that affect their overall performance when cluster gene expression data. First and foremost, the arbitrary initial mapping of the nodes (genes) makes the final mapping non-reproducible. Besides, there is a difficulty in identifying negative associations between nodes. Even if we reach to an accurate final mapping configuration, the centroids of each cluster are in the centre, delineating the boundaries among clusters is hard without the engagement of other techniques [59].

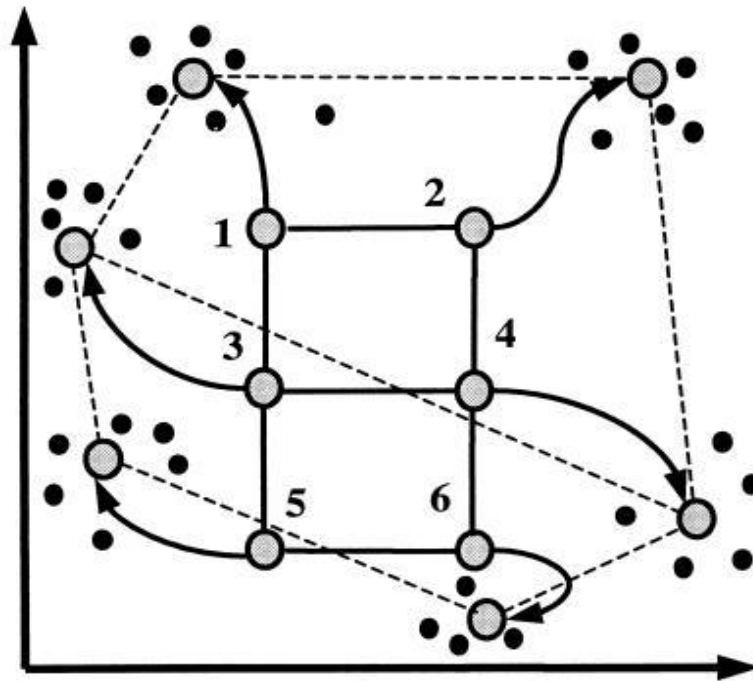


Figure 3.5: Principle of SOMs [62]

3.3.3 Affinity Propagation

Partition or centroid algorithms, composes a family of clustering methods widely used in expression-data analysis. These clustering techniques, for instance the k-means method, start with a predefined number of k data points utilized as multidimensional center points, centroids, setting an initial group of clusters. After that randomly or deliberately cluster initialization the algorithm iteratively assign samples to the nearest centroid's cluster and try to refine the centroids based on the optimization of the sum of squared errors metric. However, these techniques are quite susceptible to the initial cluster selection and the final cluster solution is directly related to the initial set up [11, 54].

On the contrary, Frey and Dueck proposed the Affinity Propagation (AP) clustering method, which simultaneously considers all data points as potential centroids. In particular, AP regards each data point as a network's node and through the recursive transmission of real-valued messages along the network's edges, tries to minimize the sum of squared errors between data points and their nearest centers, called "exemplars" when they reflect actual data points. Those messages represent the magnitude of the current affinity between one data point and a potential exemplar and are updated according to simple formulas that search for minima of an energy function [11].

The input of AP is a similarity matrix $s(i, k)$ which depicts whether data point with index k is a potential exemplar for data point i . The dissimilarity measure used is the negative squared error of Euclidean distance and the purpose is to minimize that error such as $s(i, k) = -\|x_i - x_k\|^2$ for two points x_i and x_k . This dissimilarity measure can be loose enough and applied quite successfully to pairs of images, of microarray measurements, of English sentences, or even pairs of cities. Rather than providing a prespecified number of clusters, we may input a real number $s(k, k)$ for each data

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

point k . The magnitude of the values influence whether a data point will be chosen as exemplar and the list of those values referred to as “preferences.” However, during our methodology pipeline we preferred to employ the Krzanowski and Lai index to determine the number of potential clusters and then provide it to the AP algorithm. Overall, the final number of the identified clusters is also influenced from the message-passing procedure [11].

In relation to messages exchanged between data points, there are two different types which can be viewed as log-probability ratios, the “responsibility” and the “availability” messages, Figure 3.4. The responsibility $r(i,k)$ messages sent from data point i to data point k , try to assess the suitability of a data point k being an exemplar to data point i while considering other potential exemplars, too. On the other hand, the availability $a(i,k)$ messages sent from candidate exemplar k to data point i , aim at appraising the degree of availability for each candidate exemplar to be a cluster center for the data point i . The responsibilities are computed through the formula:

$$r(i,k) \leftarrow s(i,k) - \max_{k',s.t.k' \neq k} \{a(i,k') + s(i,k')\} \quad (3.23)$$

During the first iteration the availabilities are set to zero, $a(i,k) = 0$, so the $r(i,k)$ uses the similarity value between point i and point k as its exemplar, minus the maximum value of the similarities between point i and other candidate exemplars. Whereas the responsibility update enables all candidate exemplars compete for ownership of a data point, the availability update gathers evidence from data points as to whether each candidate exemplar would make a good exemplar. The availability $a(i,k)$ is set to the “self-responsibility” $r(k,k)$, which depicts a strong indication that point k is an exemplar, plus the sum of the positive responsibilities candidate exemplar k receives from other points [11].

$$a(i,k) \leftarrow \min \left\{ 0, r(k,k) + \sum_{i',s.t.i' \notin \{i,k\}} \max\{0, r(i',k)\} \right\} \quad (3.24)$$

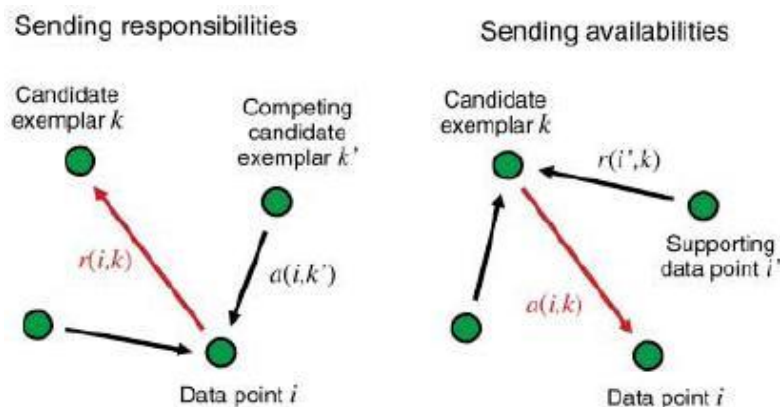


Figure 3.4: The functionality under affinity propagation

3.4 Classification

Contrary to clustering, the classification or else class prediction includes algorithms designed to classify objects into prior defined groups. Usually, such methods are applied onto a “training” data set and validated on an independent “test” data set, both of which are already labeled to specific categories. The main idea behind those techniques is to generalize from the training data to the testing data by identifying correctly the labels of the test set samples. However, classification algorithms are susceptible to the phenomenon of overfitting. The overfitting occurs when we have a small training set (a limited number of samples) and many features (genes) to model. As a consequence, we often achieve to minimize the training error while increasing the validation error as shown in Figure 3.5. The holy grail in classification is to balance between model complexity and prediction accuracy. In microarray experiments, we aim at discriminate samples of patients related to either a disease or a disease subtype or even the response to a treatment. The ultimate goal of classification applications may be a better way to distinguish among similar-looking diseases or disease subtypes, diagnostic, or it may be used to predict a clinical outcome in relation to a treatment, drug discovery. So far no method is widely accepted as optimal among the plethora of available algorithms. Therefore, we inclined to employ three well accepted in microarray studies classification algorithms to validate the feature selection algorithms; support vector machines (SVMs), k-nearest neighbor (KNN) and random forests (RF) [20, 54].

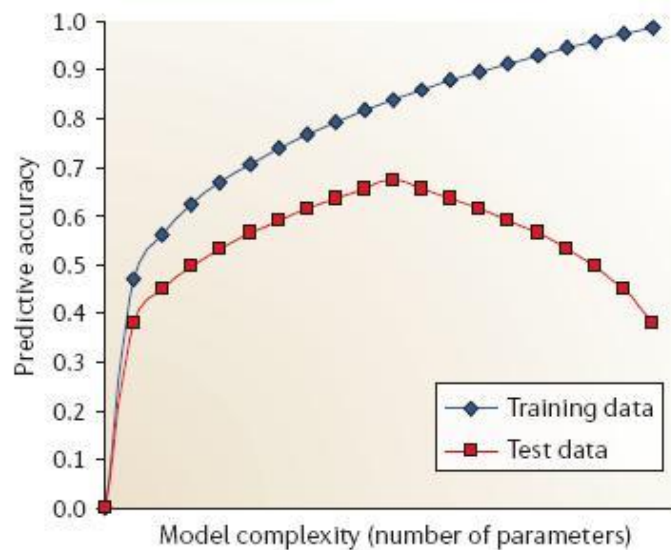


Figure 3.5: The overfitting problem [20]

3.4.1 Support Vector Machines

Support vector machines try to separate training samples describing a binary classification problem by drawing a hyperplane H in n dimensional gene-expression space. This is the main notion behind many classification algorithms in that we employ non-linear methods to transformation the input features into a high dimensional space, the *feature space*, where linear methods may be applied to separate the data points. Each sample is a point in multidimensional space, with n dimensions representing the number of genes and coordinates the expression levels of the genes. In case there is no a separating hyperplane during the initial map of the samples, the SVMs re-map the training samples into a higher-dimensional space where such a hyperplane exists. This hyperplane is best if its margin is largest. As *margin* we mean the largest distance

between the separating hyperplane H and two hyperplanes H_1 and H_2 parallel to H on both sides, on which are lying the closest sample points the so called *support vectors*. This final plane minimizes the overfitting problem because it has the largest possible margins from the training samples thus, is more robust to minor errors in the hyperplane's direction. The overfitting problem in microarray classification problems, has its origins in the fact that the number of genes far more exceeds the number of samples, called the "curse of dimensionality" in statistics. Because of this characteristic, a classification algorithm can discern training samples very accurately while being very inaccurate with new samples [63, 64].

3.4.1.1 Linear separation

In particular, suppose we have a linear binary microarray classification problem of l samples $\{(x_i, y_i), \dots, (x_l, y_l)\}$ called the training set, where x_i is a vector of n components (genes) corresponding to the expression measurements of the i^{th} sample, and y is a vector with the binary class labels for each sample, for instance 0 and 1, or -1 and +1. The main goal is to estimate a linear function $g(x)$ consisting of a weight vector w and a threshold vector w_0 , which assigns unknown samples into the correct classes based on the training samples:

$$\text{sign}(g(x_i)) \geq 0 \rightarrow y_i = +1 \quad (3.25)$$

$$\text{sign}(g(x_i)) < 0 \rightarrow y_i = -1, \forall x_i \quad (3.26)$$

A sample x_i is classified accurately if $g(x_i) \cdot y_i > 0 \quad (3.27)$

or else $g(x_i) \cdot y_i = (w^T x_i + w_0) \cdot y_i > 0, \forall x_i \quad (3.28).$

This condition controls the misclassification error on the training set, which is inversely proportional to the number of samples fulfilling this criterion. Now taking into account the margin, we change it to $(w^T x_i + w_0) \cdot y_i \geq b$ which gives for all samples x_i a solution with a distance greater than $\frac{b}{|w|}$ from the separating hyperplane H . Without affected the generality, we may scale the values of b , w , w_0 and still have the distance unchanged. By setting $b=1$, on the one hand we define the canonical hyperplanes

$$H_1 : w^T x + w_0 = +1 \quad (3.29)$$

$$H_2 : w^T x + w_0 = -1 \quad (3.30)$$

while on the other all training samples x_i satisfy the criterion

$$w^T x_i + w_0 \geq 1 \text{ for } y_i = +1 \quad (3.31)$$

$$w^T x_i + w_0 \leq -1 \text{ for } y_i = -1 \quad (3.32)$$

As a result, the separating hyperplane is defined as $g(x) = w^T x + w_0 = 0 \quad (3.33)$ and the margin from the canonical hyperplanes as $\frac{1}{|w|}$, Figure 3.6.

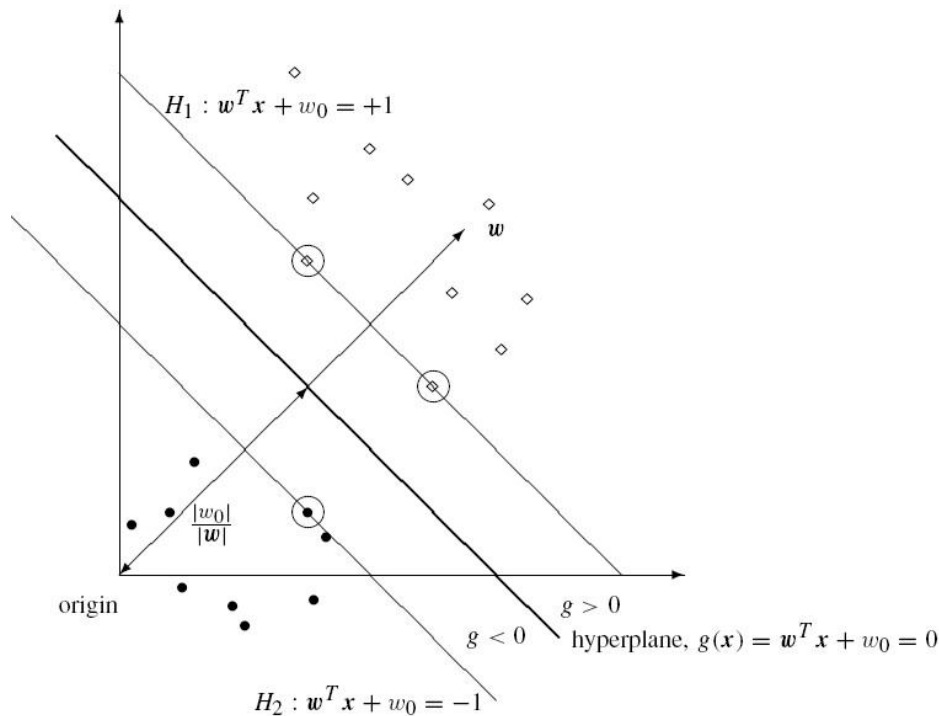


Figure 3.6: The separating hyperplane and the relevant margins [64]

The learning problem of SVM is formulated as follows:

$$\max \left(\frac{1}{|w|} \right) \text{ s.t. } (w^T x_i + w_0) \cdot y_i \geq 1 \quad i = 1, \dots, n \quad (3.34)$$

which can be re-written as

$$\min \left(\frac{1}{2} w^T w \right) \text{ s.t. } (w^T x_i + w_0) \cdot y_i \geq 1 \quad i = 1, \dots, n \quad (3.35)$$

to enable the Lagrange formalism, where the non-negativity constraints are multiplied by positive multipliers a_i where $\{a_i : i = 1, \dots, n; a_i \geq 0\}$ and subtracted from the *primal form* of the objective function L_p .

$$L_p = \frac{1}{2} w^T w - \sum_{i=1}^n a_i ((w^T x_i + w_0) \cdot y_i - 1) \quad (3.36)$$

Finding the values of w, w_0 and $a_i \geq 0$ that minimize the L_p give us the solution to the minimization problem. Therefore, we first differentiate L_p with regard to w, w_0 and then substitute the derivatives

$$w = \sum_{i=1}^n a_i x_i y_i \quad (\text{with respect to } w) \quad (3.37)$$

$$\sum_{i=1}^n a_i y_i = 0 \quad (\text{with respect to } w_0) \quad (3.38)$$

into the primal objective function, which yields the *dual form* of the Lagrangian

$$L_d = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_i a_j y_i y_j x_i^T x_j \quad (3.39)$$

$$\text{subject to } a_i \geq 0 \quad \sum_{i=1}^n a_i y_i = 0 \quad (3.40)$$

After the computation of w, w_0 we may classify a query pattern x_q by simply finding the sign of $g(x_q) = w^T x_q + w_0$ (3.41) [64].

3.4.1.2 Non-linear separation

However, most of the times, the data are not linearly separable and the minimization problem as stated previously is infeasible. Thus, we have to incorporate into the objective function an additional cost as a penalty

$$\frac{1}{2} w^T w + C \sum_{i=1}^n \xi_i \quad (3.42)$$

and the minimization problem is restated as:

$$\min \left(\frac{1}{2} w^T w + C \sum_{i=1}^n \xi_i \right) \text{ s.t. } (w^T x_i + w_0) \cdot y_i \geq 1 - \xi_i \quad i = 1, \dots, n \quad (3.43)$$

The regularization parameter C influences the penalty for ‘outliers’ and ‘softer’ margin. There is no a gold rule for setting the C value, and usually we either use several values randomly starting from 1 or employ a leave-one-out procedure on the training samples to find the value with the lowest error.

In this case, the only difference in the *dual form* of the Lagrangian is the upper bound of the α_i , $0 \leq a_i \leq C \quad \sum_{i=1}^n a_i y_i = 0$ [64].

3.4.1.3 Kernel functions

Instead of using every time the dot product of the input space $x_i^T x_j$ in the L_d we may use the dot product of the feature space supposing there is a kernel function that satisfies the equation $k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$. Indeed, using such a kernel function we just have to calculate the dot product of two vectors in the feature space without having to compute explicitly the $\varphi(x)$ transformation. As a result, the L_d takes the following form:

$$L_d = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_i a_j y_i y_j \varphi(x_i)^T \varphi(x_j) \quad (3.44)$$

subject to $0 \leq a_i \leq C \quad \sum_{i=1}^n a_i y_i = 0$. Feature mapping can simplify the classification task by separating data that cannot be separated by a linear function in the input space, but can be in the feature space [64], Figure 3.7.

The libsvm package that we engaged during the classification process of our experiment includes four different kernels [65]:

- linear: $K(x_i, x_j) = x_i^T x_j$
- polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

- radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, $\gamma > 0$
- sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$.

After trying several classification tests with the linear and the RBF kernel, we proceed with the linear since the results were better, the execution time was less and we had to engage only one parameter, the ‘regularization’ factor C , hence lessening the overfitting problem.

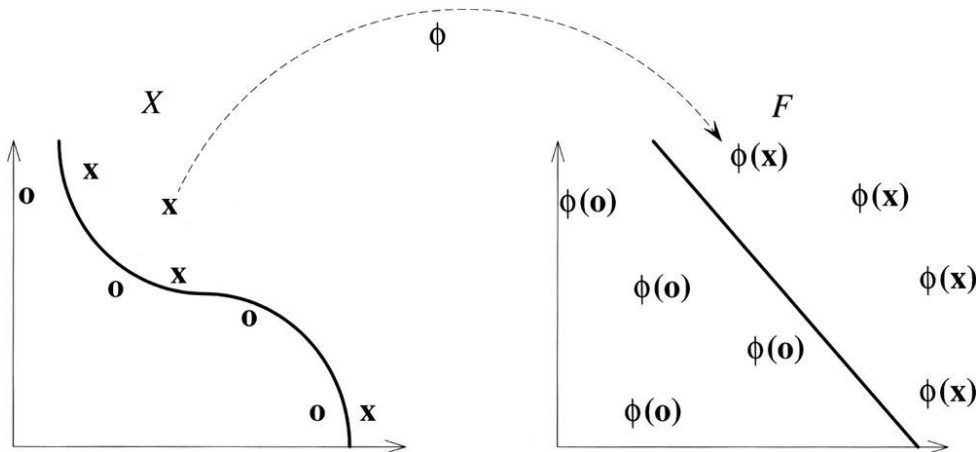


Figure 3.7: A feature mapping from a two-dimensional input space to a two-dimensional feature space [66]

3.4.2 Instance-based learning

Instance –based learning (IBL) algorithms originate from the nearest neighbor pattern classifier, which produce classification predictions based only on keeping consistency with an initial set of training instances without taking into account novel instances to maximize classification performance. The IBL algorithms do not construct and store explicit abstractions and generalizations based on the initial instances rather than computing similarities at presentation time between a novel instance and their saved instances. Due to this latency in model construction, the IBL algorithms are sometimes mentioned as “lazy” learning methods. This “laziness” entails several advantages and disadvantages. Regarding the advantages, the IBL methods are able to construct different approximations per query instance, which can be also applied locally in the neighborhood of the new query instance rather than over the entire instance space. On the other hand, practically all computation occurs during the classification time hence, classifying novel instances can be a time consuming process. Besides, IBL approaches consider all features from the new instances and accordingly trying to retrieve similar instances from memory. In case the truly most “similar” instances in memory depend on less attributes than the new instances, may appear erroneously as relatively distant in that n -attributes dimensional space, thus affecting the prediction accuracy [64, 67].

3.4.2.1 The k-Nearest Neighbor algorithm

The k-Nearest Neighbor (kNN) algorithm is the simplest instance-based algorithm. It assumes that all instances are represented by a set of attributes and correspond to points in the n -dimensional space R^n . It employs distance metrics, such as the Euclidean measure to compute the distance between two instances. For example,

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

suppose two instances x_i and x_j where $x_i = \langle x_i^1, x_i^2, \dots, x_i^n \rangle$, and x_i^r denotes the value of the r -th feature, then the distance between them is

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (x_i^r - x_j^r)^2} \quad (3.45)$$

Suppose that we have a query sample x_q ; then the KNN algorithm assigns it to the class where the maximum number of the k training samples are closer to it. If $k=1$ the algorithm assigns the new sample to the class of the nearest training sample, whereas if $k=5$ the class is determined by the majority of the k training samples closer to the new sample, Figure 3.8. In this example, the training samples are represented as points in a 2-dimensional space and the target function has a boolean value “-“ or “+” respectively. Hence, for $k=1$ the new sample x_q is classified as negative whereas for $k=5$ is classified as positive [64].

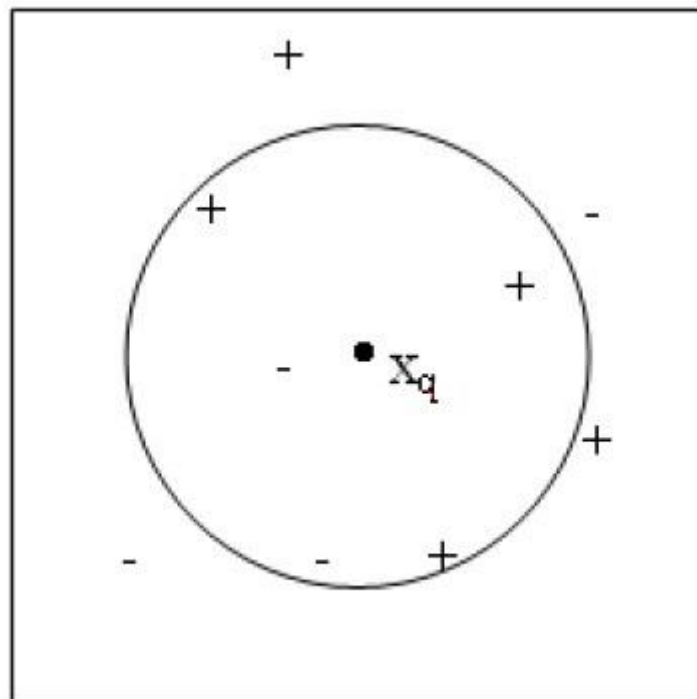


Figure 3.8: The kNN algorithm for $k=5$

During this example it is apparent the problem of selecting an appropriate k , and although there is no a particular rule for that, one common approach is to select several possible k values and through cross validation on the training samples to keep the k with the lowest error estimation.

3.4.3 Random Forests

Random Forests (RF) are a combination of tree-structured predictors where each of the trees grows using a random process. In relation to decision trees' semantics, the inner nodes define a test of some features of the sample, the branches from that nodes corresponds to the possible range of values for these features, whereas each leaf corresponds to a class label. Finally, following a top-down approach from the root to

some leaf nodes according to the branch conditions, the samples are classified through a majority voting of the decision trees, Figure 3.9. On the one hand, each path from the root to a particular leaf represents a conjunction of feature values, while on the other hand, each tree comprises a disjunction of these conjunctions [64].

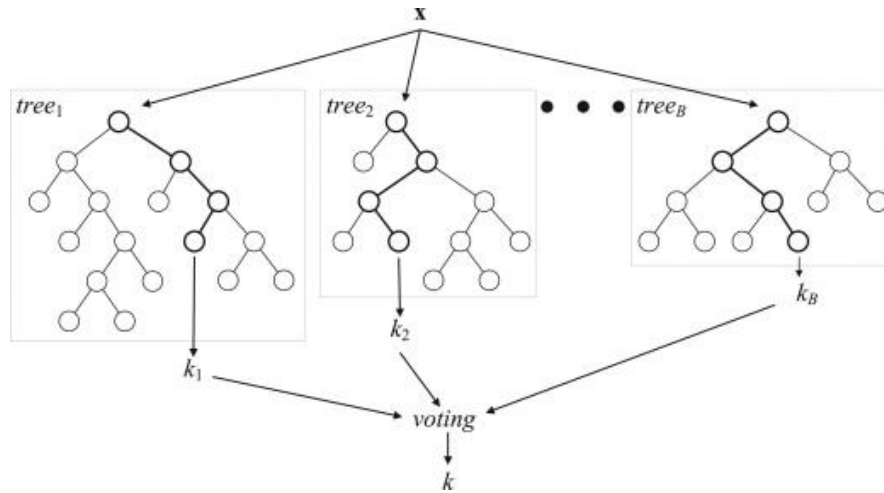


Figure 3.9: The top-down majority voting procedure in RF

Most of the algorithms engaged in growing a decision tree employ a top-down greedy construction, in which all features undergone through a statistical test to evaluate their classification performance on the training samples, and then keep the best features as tree nodes. Given a training set with N samples and M features, the N instances are sampled at random (with replacement), so as to generate a random vector Θ for each tree. For the K_{th} tree, there is a random vector Θ_K which is independent of the previous random vectors, $\Theta_1, \dots, \Theta_{K-1}$, but with the same distribution for all trees in the forest. Hence, every tree is grown using the training set and its random vector, resulting in a classifier, which votes for the most popular class. Not to mention, that each tree grows to the largest extent possible, and that this greedy construction never backtracks to reconsider earlier choices [64].

Regarding the statistical measure engaged in selecting the best features, either initially as root nodes or later as subtrees' roots, the information gain (Gain) measure from information theory is a potential candidate. This measure computes how well a given feature (F) separates the training samples (S) according to their class labels, and it is defined as follows:

$$Gain(S, F) \equiv Entropy(S) - \sum_{v \in Values(F)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (3.46)$$

where $Values(F)$ is the set of all possible values for feature (F), and S_v is the subset of S consisting of samples for which F has the value v . Hence, $Gain(S, F)$ is the information provided (the reduction in entropy) about the target function value (the class label), given the values of a particular F . Concerning the entropy measure, is defined as

$$Entropy(S) \equiv \sum_{i=1}^c -p_i \log_2 p_i \quad (3.47)$$

where c is the number of different classes (labels), and p_i is the proportion of S (the group of samples) belonging to class i . In case where all members of S belong to the same class, the entropy is 0 [64].

Alike to SVM learning algorithm, over-fitting is a significant problem to encounter in decision tree learning, too. In general, there are several approaches that try to deal with this phenomenon, and can be grouped into two classes. The first perception restricts the trees' growth before reaching their full potential, while the other let them fully grow and then prune it. By the term "prune", we mean the removal of a sub-tree rooted at a node, thus making it a leaf node. Although each individual tree may severely overfit the data, the final RF analysis is quite resistant to over-fitting because of averaging over numerous different trees. In particular, when RF draws the training set for the current tree by sampling with replacement, about one-third of the cases are left out of the sample, and called out-of-bag data (OOB). This OOB data is used to get estimates of variable importance. To measure the importance of variable x_j , values of x_j are permuted in the OOB sample, and the class membership of the OOB samples are predicted again from the tree. The number of correctly classified samples after permutation is subtracted from the original count of correctly classified samples and divided by the number of OOB samples for that tree, thus giving the decrease in classification accuracy as a proportion of samples. This permutation procedure is repeated for each tree in the forest, and the mean decrease in accuracy (MDA) is defined as the average of these values over all trees in the forest (multiplied by 100 and presented as a mean percentage decrease in accuracy) [64, 68]. In this experiment, a random forest classifier with 1,000 trees is applied.

3.5 Measuring the classification performance

The main objective of a good classifier is to balance between over-fitting and generalization error. In particular, a good classifier should achieve highly accurate scores on training samples as well as on independent test samples. However, this objective is far from being easily accomplished. In many cases, we conclude to complex models, incorporating many parameters that achieve a highly accurate classification on the training samples, but fail to distinguish validation samples. Therefore, it is welcomed to accept a modest classification error during the training process on the benefit of the validation phase. In order to evaluate the quality of the models produced from the different feature selection methods that we employed during our experiment, we first conducted a 5-fold cross-validation (5-CV) on training sets to assess the potential classification strength of the models' and then estimated its prediction power on separate test sets.

3.5.1 Performance measures

Assessing the prediction power entails the use of several performance measures where each of them provides different insights. In relation to binary classification problems, we may characterize as *positive* and *negative* the two classes respectively. Based on this notion we draw a table, which can be the basis for numerous metrics, Figure 3.10. Specifically, we use four different abbreviations to represent the four different prediction outcomes - TP (True Positive) for the correctly classified positive samples, FP (False Positive) for the negative samples that classified as positive, TN (True Negative) for the correctly classified negative samples and FN (False Negative) for the positive samples that classified as negative. One of the widely used measures in classification problems is accuracy (ACC), which is defined as the ratio of the number of correct predictions to

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

the total number of samples. However, ACC ignores the difference between *false alarms* (*Type I errors*), measured by FP, and *missed detections* (*Type II errors*), measured by FN. Moreover, ACC is influenced by the class distribution, thus can be quite misleading in unbalanced datasets. For instance, in a 90-10 class ratio, a model can achieve 90% accuracy by just identifying correctly the dominant class samples [69].

Other metrics include the *True Positive Rate* (TPR), the *False Positive Rate* (FPR), the *specificity* and the *precision*. The TPR which is also called as the *hit rate*, *recall*, or *sensitivity*, is defined as the ratio of positives correctly classified to the actual number of positives, whereas the FPR, also referred to as the *false alarm rate*, is the ratio of negatives incorrectly classified to the actual number of negatives. Likewise, the ratio of correctly classified negatives to the actual number of negatives defines the *specificity* measure. We should take note that the *sensitivity* and *specificity* metrics are independent since knowing one tell us nothing about the value of the other, contrary to FPR and *specificity*, where their sum is equal to one thus , knowing one allows us to calculate the other. Finally, *precision* is a further valuable measure that deals with the positive samples and is defined as the ratio of correctly classified positives sample over the total number of samples classified as positive [69, 70].

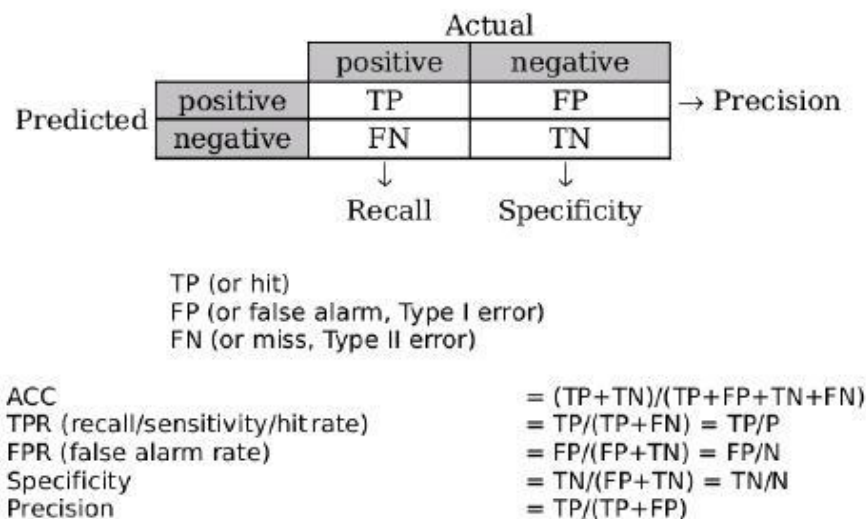


Figure 3.10: The performance metrics in binary classification [69]

An additional performance measure, which has been introduced as a better measure for evaluating the predictive ability of machine learners than accuracy is the Area Under the ROC Curve (*AUC*). The ROC curve is a two-dimensional plot between the TPR (Y-axis) against the FPR (X-axis) of the predictions with their values ranging from zero to one. Connecting the set of points (FPR, TPR) gives the ROC curve or space that facilitates the comparison among different learners on a dataset. The closer the curve is to the Y-axis (high true positives) and the further away it is from the X-axis (low false positives), the more accurate the predictions are. An ROC curve similar to a 45 degrees straight line stands for predictions made by random guessing referred to as the no-discrimination line. During our experiment, we decided to employ the AUC metric as a general performance measure and the TPR and TNR as two independent but informative metrics related specifically to the negative and positive classes [70].

4. mAP-KL: A NEW HYBRID METHOD FOR FEATURE (GENE) SELECTION

4.1. Introduction

A feature selection method in microarray gene expression data should be independent of platform, disease and dataset size. Our hypothesis is that among the statistically significant ranked genes in a gene list, there should be clusters of genes that share similar biological functions related to the investigated disease. Thus, instead of keeping N top ranked genes, it would be more appropriate to define and keep a number of gene cluster exemplars. We propose a hybrid FS method (mAP-KL), which combines multiple hypothesis testing and affinity propagation clustering algorithm along with the Krzanowski & Lai cluster quality index, to select a small yet informative subset of genes.

4.2. The general framework and implementation of our Methodology

4.2.1 The filtering method

The proposed methodology combines ranking-filtering and cluster analysis to select a small set of non-redundant but still highly discriminative genes. In relation to the filtering step, we first employ the maxT function (see 3.1.1.1) from the 'multtest' [71] r-package to rank the genes of the training set and then we reserve the top N genes ($N = 200$) for further exploitation. Our decision on which feature selection method to employ follows the findings of an analysis that we carried on feature selection methods [72]. Specifically, we assessed the classification performance of five different feature selection methods on data from ten different neuromuscular diseases. Each method yielded a different ranked list of genes, which was then used iteratively from top to bottom, in the range of 2 to 400 genes, to compose a new classification scheme in each iteration. The evaluation of the classification performance of all the produced schemes per feature selection method is depicted in Figure 4.1, and shows that the maxT achieved an average discrimination accuracy of 95%, between normal and disease samples.

In the same experiment, we also inspected the robustness of the classification performance when differentiating the number of genes in the training set. According to the accuracy graphs as depicted in Figure 4.2, we notice extreme top and bottom values of the accuracy score in DEDS, and LIMMA methods and in a lesser degree in RankProd method. This observation implies that biologically significant genes are merged with statistically significant genes and the level of enrichment from each group may affect the robustness of the classification procedure. Hence, the top ranked genes are not necessarily biologically relevant in the context of the specific diseases. Besides, some of the statistically significant genes may act as "noise" in the classification procedure. The more unstable the performance of the classifier the more "noisy" is considered to be the selected gene subset. Moreover, each classifier has its own inertia regarding the influence of the corresponding parameters and as a result, the "noise" in the classification is also affected from the characteristics of the classifier [72].

On the contrary, maxT and SAM appear to identify better those genes that play a significant role toward the samples discrimination. Indeed, the respective classification graphs from these two methods seem considerably stable when differentiating the number of genes and additionally, no particular fluctuation appears between 200 and 400 genes. However, setting the N parameter arbitrarily does not guarantee robust and

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

efficient classification results, so we decided to exploit several cases of top ranked genes in a simulation scenario (see section 4.3.1) before coming to a decision.

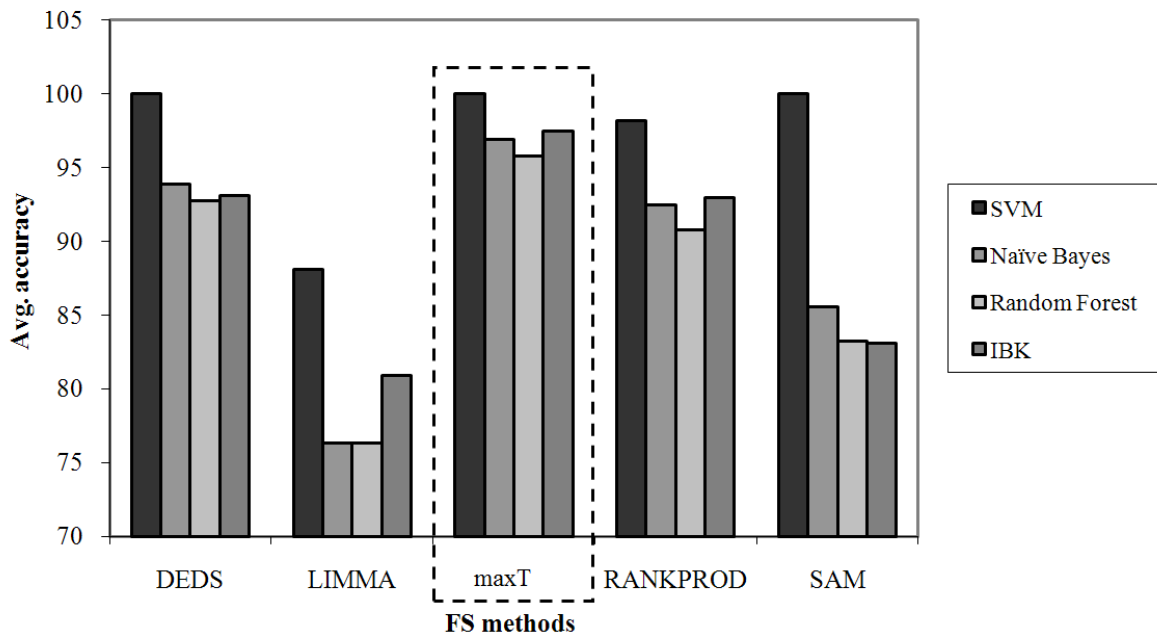


Figure 4.1: The overall classification accuracy of five feature selection methods on ten datasets of neuromuscular disease data according to four classification algorithms

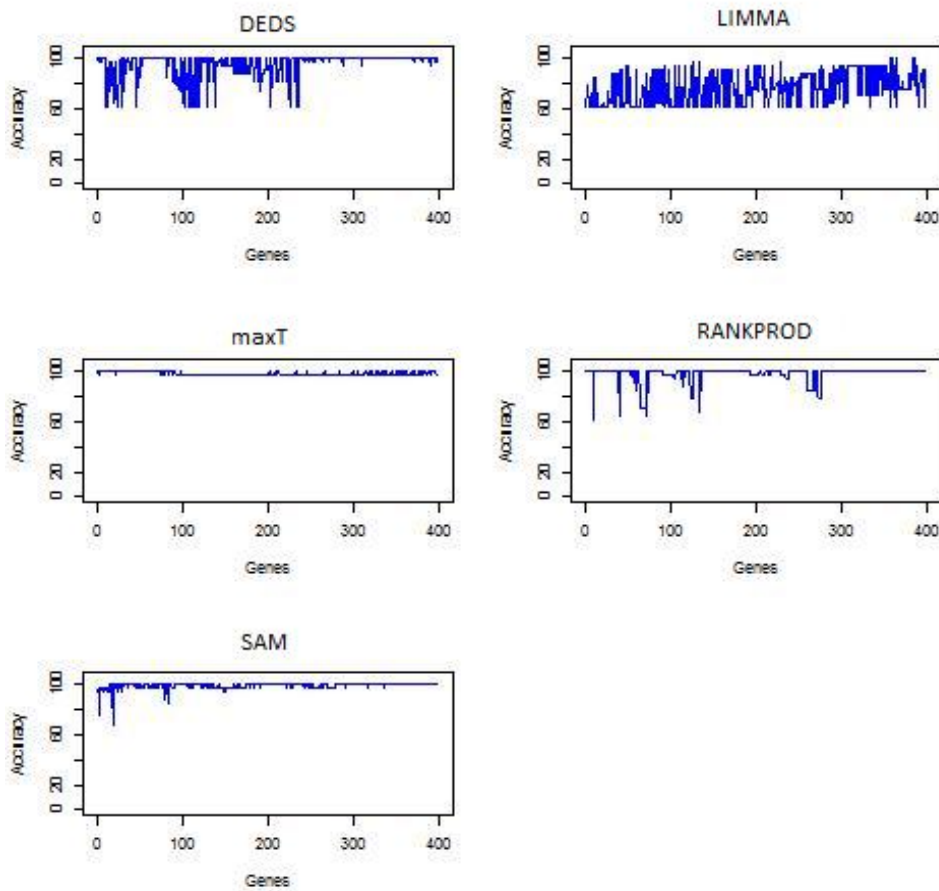


Figure 4.2: The influence on the accuracy when differentiating the length of the training set

4.2.2 The clustering quality index

In the sequel, prior to clustering analysis with AP we define the number of clusters, which in essence will be the number of representative genes that finally will compose our subset. The decision about which quality index to use, was based first on the results of the Tibshirani et al. indices comparison survey, where the index of Krzanowski and Lai excelled as well as on several trials on simulated clustering data that also proved the efficiency of the index. Hence, we employed the index of Krzanowski and Lai as included in the 'ClusterSim' package [58] to determine the number of clusters solely on the disease samples of the training test set.

This is actually a very fine detail in our methodology, since it has a direct impact on the clusters identification and consequently on the selected genes. There were two options in which part of the data it would be the most proper and advantageous for the rest of the analysis. The first option was to search for the clustering structure solely in the samples belonging to the normal phenotype or in the control phenotype in generally. The second alternative was to investigate the samples in the disease phenotype. We finally reckoned that what actually is of interest for the identification of significant genes relevant to a disease, is the disease part of the data because all the information about the "triggered" molecular processes is definitely present in it.

4.2.3 The clustering algorithm

The final step of our methodology involves the cluster analysis through the AP clustering method. AP algorithm appeared in the late 20s and according to a benchmark analysis [73] across 15 other clustering algorithms, including k-means and k-medians clustering, hierarchical agglomerative clustering e.t.c., excelled at finding the more accurate clustering solution. Besides its intrinsic belief that initially all data points (genes) are considered as potential exemplars and its efficient convergence to the final clustering, urged us to adopt AP through the APCluster r-package [74] as indispensable part of our methodology. Thus, we pass into AP the number of k clusters according to the Krzanowski and Lai index and then let AP to detect those n clusters where $(n = k)$ clusters among the top N genes (a pre-defined number). The algorithm converges to the requested number of clusters (most of the times) and provides us with a list of the most representative genes of each cluster, the so called exemplars. These n exemplars are expected to form a classifier that shall discriminate between the normal and disease classes in a test set.

Finally, we formulate the updated train and test sets by keeping only those n genes, and proceed with the classification process. The general flowchart of our methodology appears in Figure 4.3. The mAP-KL pipeline is currently integrated into a software package developed under the R environment (see chapter 7).

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

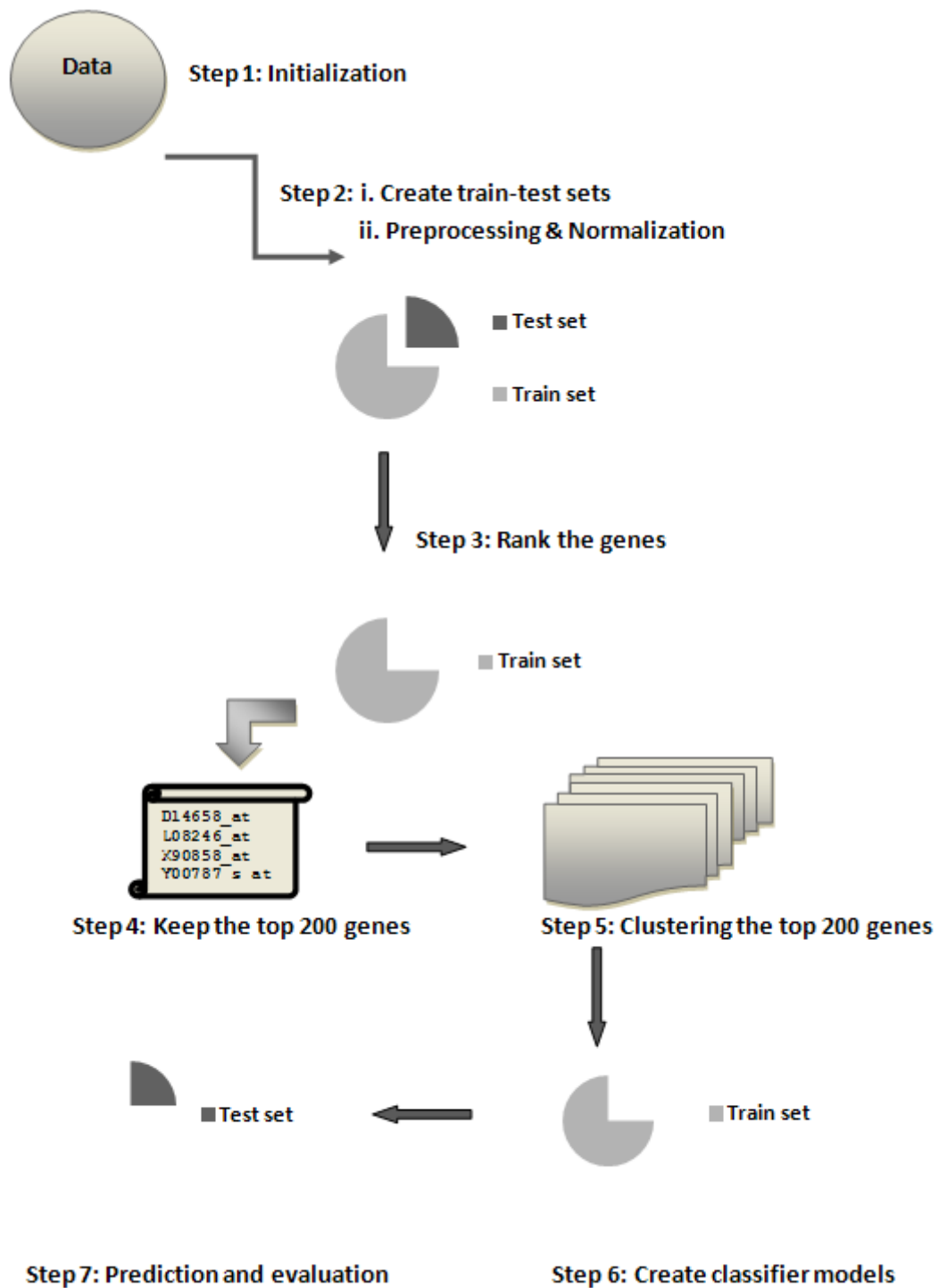


Figure 4.3: The mAP-KL methodology flowchart

4.3. Simulated data

Initially, we investigated mAP-KL’s performance on two synthetic datasets prior to any real microarray data. We intentionally utilized two different simulation setups to examine two different hypotheses. In the first hypothesis, we wanted to verify that mAP-KL provides us with a small subset of representative features, at least one gene per cluster, adequate for accurate classification. Therefore, we considered a binary classification problem simulating a normal-disease case with six different scenarios, Table 4.1, in relation to the number of differentially expressed genes (DEGs) that are included in the disease class samples. All simulated gene expression values follow the normal distribution for the respective mean and variance values as presented in Table 4.1.

In particular, we started with 50 DEGs, Figure 4.4, belonging to five clusters of 10 ‘genes’ and reached to 500 DEGs spreading in 25 clusters of 20 ‘genes’ per cluster, trying to imitate pathways. It is obvious in the figure that there is a considerable overlap among the data points per cluster, hence making the accurate discrimination a harsh task. The normal and the disease classes have 1,200 samples of 10,000 ‘genes’ per sample, where the first 200 samples from each class compose the train set and the rest form the test set. The non-differentially expressed genes are independently drawn from normal distribution with mean = 0 and variance = 0.5.

In the second hypothesis, we employed a subset of the publicly available ‘Golden Spike’ [75] Affymetrix case–control experiment, incorporated in the ‘st’ package [50, 76] under the name ‘choedata’. In this scenario, it was intriguing to explore the number of the known DEGs included in mAP-KL’s subset and whether they are capable of providing us with accurate models. The ‘choedata’ describes a binary classification problem with three replicates per class and 1,331 DEGs scattered randomly among 11,475 genes. The number of DEGs is considered as adequate towards the accurate estimation of the false-negative and false-positive rates at each fold-change level, Figure 4.5. Besides, there are intensity values with low fold-changes of 1.2- fold trying to imitate subtle biologically relevant differences that are frequently ignored or excluded during microarray analysis [75].

Table 4.1: The statistical parameters under the simulated data

		Simulated data parameters																								
DEGs (Genes/ Cluster)	var	mean																								
		50 (10)	0.2	2	3	4	5	6																		
100 (20)	0.2	2	3	4	5	6																				
200 (20)	0.2	2	3	4	5	6	7	8	9	10	11															
300 (20)	0.2	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9										
400 (20)	0.2	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10	10.5	11	11.5					
500 (20)	0.2	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12	12.5	13	13.5	14
Non-DEGs	0.5	0																								

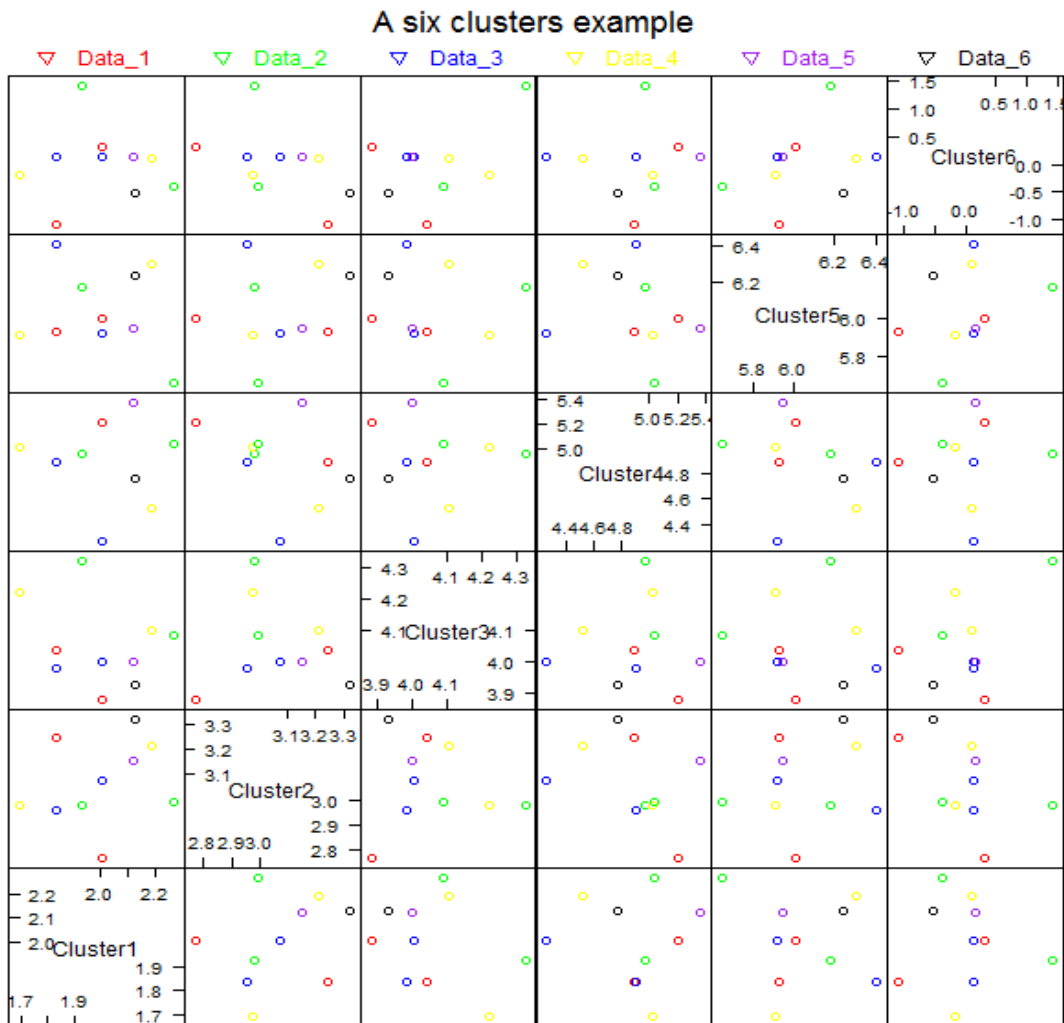


Figure 4.4: A scatter plot of five clusters with DEGs and one cluster, the 6th, with non-DEGs.

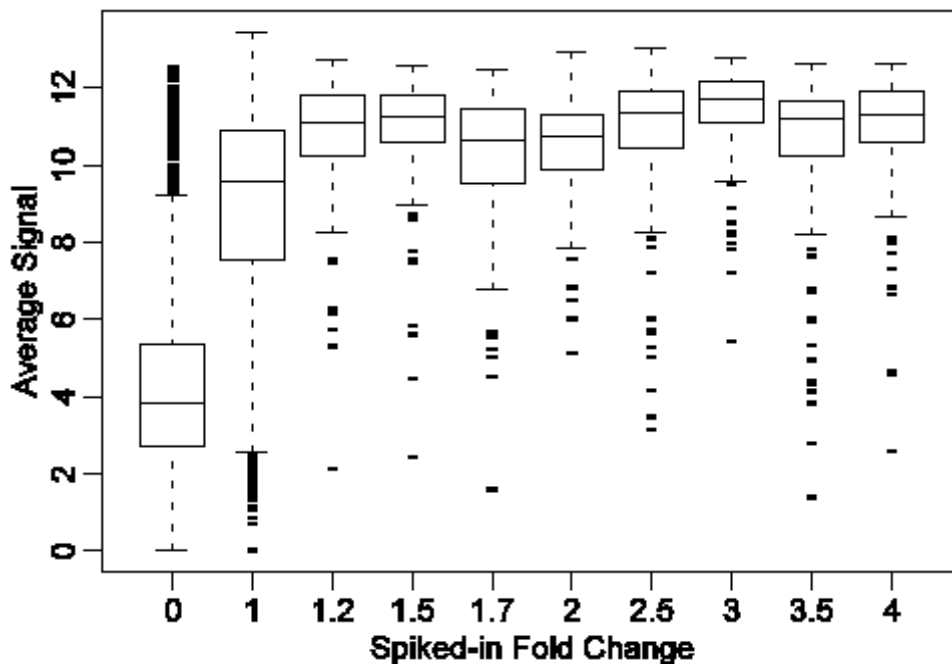


Figure 4.5: The boxblots of the average log₂ expression summary intensity as a function of spiked-in fold change. The probe sets that were not spiked i.e. fold change = 1, are placed at zero fold change to separate them from the probe sets which were spiked in at fold change = 1.

4.3.1 The clusters setup

We applied the mAP-KL on training sets of 200 samples with 10.000 ‘genes’ and diverse number of DEGs. Moreover, for each training set we differentiated the number of the top ranked genes kept for clustering, Table 4.2. The purpose of this case study was twofold. On the one hand, we wanted to investigate how many DEGs are included in our final subset along with their cluster origin. Furthermore, we explored the influence on the DEGs’ selection when differentiating the number of the top ranked genes. We also employed three other FS methods, (eBayes, maxT and RF-MDA), keeping either the top 20 ranked ‘genes’ (cases of 50 DEGs, 100 DEGs, 200 DEGs, 300 DEGs) or the top 30 ranked ‘genes’ (cases of 400 DEGs and 500 DEGs) trying to keep their length comparable with the subset’s length of mAP-KL.

As far as the identification of DEGs belonging to different clusters is concerned, the mAP-KL managed to compose subsets with at least one representative ‘gene’ from each cluster. Besides, as shown in Table 4.2, in almost all cases the maximum subsets’ length does not exceed the actual number of clusters in the training set. In relation to the other FS methods, only the RF-MDA method composed subsets of ‘genes’ with satisfactory representation of the actual clusters and comparable to mAP-KL. The eBayes and maxT methods demonstrated poor enrichment.

With respect to the effect of the number of top ranked ‘genes’ kept for clustering, it is evident that the closer to the real number of DEGs, the better the identification and selection of representative genes, Figure 4.6. Specifically, in cases where the number of DEGs is considerably lower than the number of N top ranked genes (e.g. 50 DEGs with 200 top ranked genes) the identified clusters are less than the actual. Similarly, when the number of DEGs far exceeds the number of N top ranked genes the identified clusters are fewer, for instance 500 DEGs with 200 top ranked genes parameter. Nonetheless, during the real gene expression data experiment, we employed a moderate value for the parameter $N=200$ top ranked genes.

As a final point, we formed the respective train-test sets for all methods and evaluated their performance with the aid of three classifiers (SVM-linear, KNN, RF). All methods performed accurately (ACC=100%) for all three classifiers.

Table 4.2: The number of clusters identified by mAP-KL for several top N ranked genes compared to three other FS methods (the number of genes per subset is in parenthesis).

DEGs	Identified Clusters										eBayes	maxT	RF-MDA
	Top ranked genes (mAP-KL)												
	50	100	150	200	250	300	350	400	450	500			
50	5 (5)	6 (6)	4 (4)	3 (3)	3 (3)	3 (3)	2 (2)	2 (2)	2 (2)	2 (2)	2 (20)	2 (20)	5 (20)
100	3 (3)	5 (5)	6 (6)	6 (14)	5 (5)	4 (4)	4 (4)	4 (4)	3 (3)	3 (3)	1 (20)	2 (20)	5 (20)
200	3 (3)	6 (6)	8 (8)	10 (10)	11 (11)	11 (11)	8 (8)	5 (5)	5 (5)	5 (5)	1 (20)	2 (20)	10 (20)
300	3 (3)	6 (6)	8 (8)	10 (10)	13 (13)	15 (15)	11 (11)	7 (7)	7 (7)	6 (6)	2 (20)	4 (20)	10 (20)
400	4 (4)	6 (6)	8 (8)	11 (11)	13 (13)	15 (15)	18 (18)	20 (20)	21 (23)	10 (10)	3 (30)	4 (30)	16 (30)
500	4 (4)	7 (7)	9 (9)	11 (11)	13 (13)	16 (16)	18 (18)	20 (20)	23 (23)	25 (25)	3 (30)	4 (30)	19 (30)

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

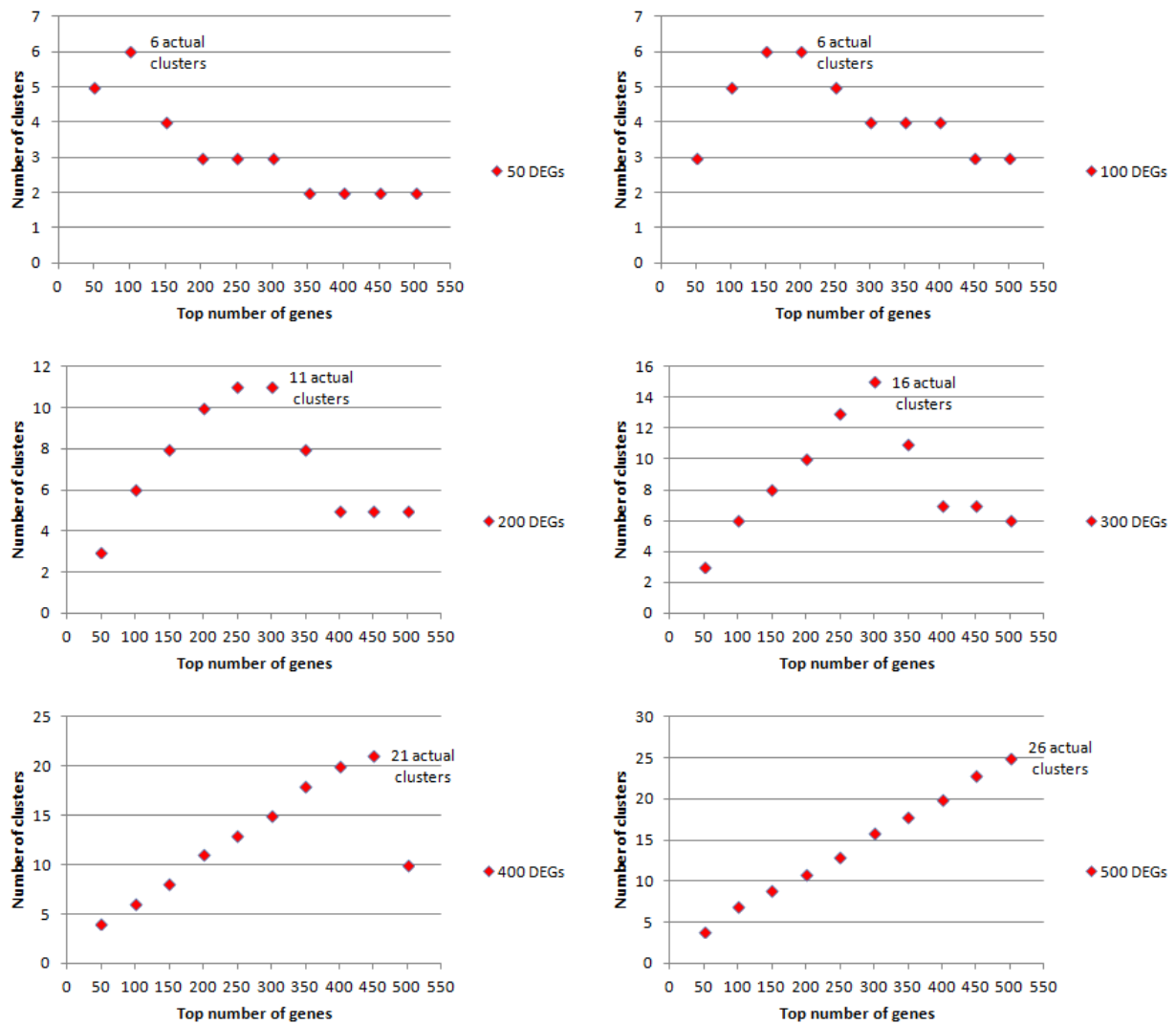


Figure 4.6: The relationship between DEGs and top-ranked genes

4.3.2 The 'choedata' setup

In this setup, we were interested in exploring, the length of the *mAP*-KL's subset in relation to the known DEGs included in it. Therefore, we applied on the 'choedata' the *mAP*-KL, which produced a subset of 15 genes with eight DEGs in it, Table 4.3. We then formed classification models with the three classifiers and concluded to accurate classification results.

However, we were intriguing to examine the impact in the *mAP*-KL's subset quality i.e. the number of DEGs, when using different statistical methods to rank the genes. Indeed, engaging the parametric Welch-t test statistical method, led us to a subset of 16 genes with 13 DEGs included. On the contrary, the Wilcoxon's subset includes 8 out of 15 DEGs. Despite this remarkable difference in the number of DEGs included in the two subsets, the classification results were accurate in both cases. Nonetheless, including more DEGs in a classifier is of benefit to the biological analysis if not to the classification process itself.

Table 4.3: The subsets of genes selected from the 'choedata' according to mAP-KL. We have marked bold the DEGs.

Wilcoxon	Welch-t
tun	Rim
CG6904	CG14254
SH3PX1	Cyp4p2
CG10283	CG10483
Tgt	CG8193
CG17930	Gdh
CG8300	CG17600
b	Gprk2
CG12213	kek3
RhoGEF2	CG5880
Imp	CG3544
Dip2	CG4785
Spred	CG32043
NA	CG18125
NA	CG7069
	orb

5. Feature selection with *mAP-KL*

5.1. Introduction

Following the development and successful testing in simulated data of *mAP-KL*, we designed and executed an elaborate set of analytical experiments with 5-CV on the training set and hold-out validation on a separate set using three different classifiers, RF – SVM – KNN, to assess its performance across whole genome expression datasets from both small and large patient cohorts. In relation to small cohorts, we employed data from 6 neuromuscular diseases, while for large cohorts we utilized data from four different types of cancer. On those microarray datasets, we also applied 12 other feature selection/elimination approaches and compared the classification results. In particular, we employed six univariate filter methods (eBayes, ODP, maxT, SAM, SNR and t-test), one multivariate filter algorithm (cat), three dimension reduction approaches (BGA-COA, PCA, PLS), one embedded method (Random Forest), and one hybrid method (Hyk-Gene). We further assessed the *mAP-KL*'s performance towards other feature selection and/or classification studies, conducted on the same cancer datasets.

5.2. Microarray data

Apart from the synthetic data, we utilized real data, including neuromuscular and cancer diseases data, to assess *mAP-KL*'s performance. Neuromuscular diseases are rare among the general population, thus the available tissue samples and whole transcriptome data are very limited. This characteristic is crucial since we intended to develop a FS method that produces robust models even in studies with limited number of samples. We therefore included data from Bakay et al. [77] related to 'amyotrophic lateral sclerosis' (ALS), 'Duchenne muscular dystrophy' (DMD), 'juvenile dermatomyositis' (JDM), 'limb-girdle muscular dystrophy type 2A' (LGMD2A), and 'limb-girdle muscular dystrophy type 2B' (LGMD2B), as well as 'nemaline myopathy' (NM) data from Sanoudou and Beggs [78] and Sanoudou et al. [79]. The gene expression data for the first five diseases originate from Affymetrix HG_U133A gene chips and share a set of 18 normal samples, whereas the NM data originate from Affymetrix HG_U95A gene chips and have been compared to 21 normal samples. We divided the data approximately in half, and kept the first half to build a balanced train sets and the second half to validate the classification models (Table 6.1). Concerning the preprocessing approach, all neuromuscular data underwent log₂ transformation and quantile normalization across samples.

Regarding the cancers datasets, we utilized microarray data from breast cancer, colon cancer, leukemia, and prostate cancer, all of which are considered benchmark datasets and have been widely used in gene expression classification studies. Van't Veer [3] explored breast cancer patients' clinical outcome following modified radical mastectomy or breast-conserving treatment combined with radiotherapy. Patients with good and poor 5-year prognosis following initial diagnosis were included. The breast cancer data was already normalized so we omitted the preprocessing step. The colon datasets [80] consisted of 62 samples of colon epithelial tissue taken from colon cancer patients. Sample were obtained both from tumor tissue as well as adjacent, unaffected parts of the colon of the same patients, and measured using high density oligonucleotide arrays. For the analysis of the colon microarray data we followed the same pre-processing approach as we did for the neuromuscular data i.e. we performed log₂ transformation and quantile normalization across samples. Datasets from acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) [81], two distinct acute leukemias,

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

were used for cancer subtype classification. The train set consisted of 27 ALL samples and 11 AML samples. Finally, prostate cancer [82] training data consisted of 52 prostate tumour tissue and 50 normal prostate tissue datasets, while the testing data consisted of 25 tumour and 9 normal datasets [83]. In relation to the preprocessing of the leukemia and the prostate data, we first set the Golub's floor and ceiling values (floor=100 and ceiling =16.000), though without filtering the genes, and then applied log10 transformation and quantile normalization across samples. For all cancers datasets we kept the train and test sets as provided, see Table 5.1.

Table 5.1: The real microarray data divided in train and test sets

Datasets		Attributes (nr of genes)	Train set samples (class1:class2)	Test set samples (class1:class2)
Amyotrophic lateral sclerosis	(ALS)	22,283	6:6	12:3
Duchenne muscular dystrophy	(DMD)	22,283	7:7	11:3
Juvenile dermatomyositis	(JDM)	22,283	10:10	8:11
Limb-girdle muscular dystrophy type 2A (LGMD2A)		22,283	7:7	11:3
Limb-girdle muscular dystrophy type 2B (LGMD2B)		22,283	7:7	11:3
Nemaline myopathy	(NM)	12,600	8:8	13:5
BREAST CANCER		(4348)24,481	44:34	7:12
COLON CANCER		7,129	15:15	7:25
ALL/AML LEUKEMIA		7,129	27:11	20:14
PROSTATE CANCER		12,600	52:50	25:9

5.3. Neuromuscular disease data

The use of small cohorts in biomedical research is common in some types of studies such as those of rare diseases. These small cohorts make feature selection algorithms prone to overfitting and thus less reliable [59] compared to larger cohorts. It was therefore intriguing to explore the robustness and generalization of *mAP-KL* on train sets with length ranging from 12 to 20 samples and test sets with 15 to 19 samples respectively. The majority of the methods in ALS and DMD validation achieved the highest classification score (AUC =1.00) in RF and SVM classifiers (Tables 5.2, 5.3) except for the HykGene in ALS and the PCA in DMD. In KNN classifier though, half of the methods achieved scores lower than AUC=1.00, Table 5.4.

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

Table 5.2: The classification results in ALS and DMD neuromuscular diseases according to RF classifier

	FS methods	5-CV			Hold-out Validation		
		AUC	TNR	TPR	AUC	TNR	TPR
ALS	mAP-KL	1.00 (0.00)	1.00 (0.00)	0.98 (0.14)	1.00	1.00	1.00
	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	PCA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	RF-MDA	1.00 (0.00)	1.00 (0.00)	0.93 (0.25)	1.00	1.00	1.00
	Rnd	1.00 (0.00)	1.00 (0.01)	1.00 (0.00)	0.99	0.92	0.97
HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.64	0.42	0.67	
DMD	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.91	1.00
	BGA-COA	0.98 (0.14)	0.85 (0.32)	1.00 (0.00)	1.00	1.00	1.00
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.91	1.00
	HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	RF-MDA	1.00 (0.00)	1.00 (0.00)	0.99 (0.07)	1.00	1.00	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	Rnd	1.00 (0.00)	0.99 (0.04)	1.00 (0.00)	0.99	0.96	0.93
PCA	0.48 (0.42)	0.48 (0.46)	0.41 (0.45)	0.61	0.55	0.67	

Table 5.3: The classification results in ALS and DMD neuromuscular diseases according to SVM classifier

	FS methods	5-CV			Hold-out Validation		
		AUC	TNR	TPR	AUC	TNR	TPR
ALS	mAP-KL	0.93 (0.16)	1.00 (0.00)	0.86 (0.32)	1.00	1.00	1.00
	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	RF-MDA	0.96 (0.04)	1.00 (0.00)	0.99 (0.07)	1.00	1.00	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	Rnd	0.97 (0.03)	1.00 (0.00)	0.97 (0.06)	0.98	0.97	1.00
	HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.79	0.58	1.00
PCA	0.85 (0.24)	0.86 (0.30)	0.83 (0.36)	0.75	0.50	1.00	
DMD	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	RF-MDA	0.94 (0.14)	1.00 (0.00)	0.87 (0.28)	1.00	1.00	1.00
	Rnd	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.95	0.91	1.00
PCA	0.49 (0.28)	0.51 (0.45)	0.46 (0.48)	0.18	0.36	0.00	

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

Table 5.4: The classification results in ALS and DMD neuromuscular diseases according to KNN classifier

	FS methods	5-CV			Hold-out Validation		
		AUC	TNR	TPR	AUC	TNR	TPR
ALS	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.98	0.97	1.00
	Rnd	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.98	0.95	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.96	0.92	1.00
	RF-MDA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.96	0.92	1.00
	PCA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.79	0.58	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.79	0.58	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.75	0.50	1.00
HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.67	0.67	0.67	
DMD	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	RF-MDA	0.99 (0.05)	1.00 (0.00)	0.98 (0.10)	1.00	1.00	1.00
	Rnd	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.95	0.91	1.00
PCA	0.73 (0.27)	0.79 (0.37)	0.67 (0.44)	0.48	0.64	0.33	

In JDM almost all of the methods achieved the highest AUC score (1.00) during hold-out validation irrespective of the classifier, Tables 5.5, 5.6, 5.7, though with RF the respective TNR score was 0.88 for the BGA-COA, eBayes, ODP, SNR and cat methods. The mAP-KL had a marginal performance deterioration with the SVM classifier (AUC=0.94). In 5-CV the PCA was the only method that failed to distinguish correctly all samples in all three classification schemes.

In relation to the LGMD2A, the RF classifier benefits the majority of the methods to discriminate accurately all the samples during hold-out validation, Table 5.5. In particular, ten methods achieved the highest AUC value, though only BGA-COA, mAP-KL and maxT (200) achieved the

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

highest TNR and TPR, too. The TNR score for PLS-CV was 0.91, for RF-MDA, ODP and SNR was 0.73, while for HykGene was 0.45 and for eBayes 0.36. It is worth noticing that the TNR score of the maxT with the 20 genes subset was considerably lower to that of maxT (200). Unlike RF, only two of the methods managed to excel with the SVM (BGA-COA and maxT(200)) and KNN (BGA-COA and maxT(200)) classifiers, Tables 5.6, 5.7. mAP-KL achieved the same high classification score in those two classifiers (AUC=0.95). The rest of the methods had AUC score above 0.70 with the exception of HykGene and PCA methods.

Table 5.5: The classification results in JDM and LGMD2A neuromuscular diseases according to RF classifier

FS methods	5-CV			Hold-out Validation			
	AUC	TNR	TPR	AUC	TNR	TPR	
JDM	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	HykGene	1.00 (0.00)	1.00 (0.00)	0.95 (0.15)	1.00	1.00	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	PCA	0.90 (0.19)	0.77 (0.31)	0.73 (0.32)	1.00	1.00	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	RF-MDA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.88	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.88	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.88	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.88	1.00
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.88	1.00
	Rnd	1.00 (0.00)	1.00 (0.00)	0.99 (0.03)	1.00	0.99	0.98
LGMD2A	mAP-KL	1.00 (0.00)	0.87 (0.30)	1.00 (0.00)	1.00	1.00	1.00
	BGA-COA	1.00 (0.00)	0.96 (0.17)	1.00 (0.00)	1.00	1.00	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.91	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.73	1.00
	RF-MDA	1.00 (0.00)	0.98 (0.10)	1.00 (0.00)	1.00	0.73	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.64	1.00
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.64	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.64	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.36	1.00
	HykGene	1.00 (0.00)	0.97 (0.12)	0.98 (0.10)	0.94	0.45	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.94	0.45	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.94	0.73	1.00
	Rnd	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.89	0.70	0.93
	PCA	0.83 (0.30)	0.61 (0.43)	0.99 (0.03)	0.58	0.27	1.00

Table 5.6: The classification results in JDM and LGMD2A neuromuscular diseases according to SVM classifier

	FS methods	5-CV			Hold-out Validation		
		AUC	TNR	TPR	AUC	TNR	TPR
JDM	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	PCA	0.62 (0.22)	0.73 (0.29)	0.51 (0.34)	1.00	1.00	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	RF-MDA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	Rnd	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.94	0.88	1.00
	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.94	0.88	1.00
LGMD2A	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.95	0.91	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.95	0.91	1.00
	RF-MDA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.91	0.82	1.00
	Rnd	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.90	0.83	0.97
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.86	0.73	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.86	0.73	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.86	0.73	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.86	0.73	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.82	0.64	1.00
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.77	0.55	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.77	0.55	1.00
	HykGene	0.92 (0.16)	0.85 (0.32)	0.98 (0.10)	0.68	0.36	1.00
	PCA	0.71 (0.26)	0.81 (0.35)	0.61 (0.44)	0.44	0.55	0.33

Table 5.7: The classification results in JDM and LGMD2A neuromuscular diseases according to KNN classifier

	FS methods	5-CV			Hold-out Validation		
		AUC	TNR	TPR	AUC	TNR	TPR
JDM	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	PCA	0.53 (0.23)	0.57 (0.32)	0.48 (0.32)	1.00	1.00	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	RF-MDA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	Rnd	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	0.99
	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.94	0.88	1.00
LGMD2A	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	maxT (200)	0.92 (0.16)	0.97 (0.12)	0.87 (0.32)	1.00	1.00	1.00
	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.95	0.91	1.00
	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.95	0.91	1.00
	RF-MDA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.91	0.82	1.00
	Rnd	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.89	0.82	0.97
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.86	0.73	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.82	0.64	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.82	0.64	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.77	0.55	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.77	0.55	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.77	0.55	1.00
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.73	0.45	1.00
	HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.65	0.64	0.67
	PCA	0.80 (0.22)	0.95 (0.18)	0.64 (0.42)	0.64	0.27	1.00

Unlike the previous datasets, in LGMD2B validation, the majority of the methods failed to discriminate accurately the test samples. In particular, with the RF classifier only three of the methods (RF-MDA, maxT (200) and PLS-CV) achieved the highest AUC (1.00) but their TNR scores were 0.73, 0.64 and 0.55 respectively, Table 5.8. Although many methods distinguish all disease samples correctly i.e. TPR = 1.00, all of them failed to discern all normal samples i.e. TNR < 1.00. Approximately half of the methods had a TNR below 0.50 (included, eBayes, SAM and mAP-KL) and no method had TNR greater than 0.80. The RF-MDA was the only of the three previous methods that achieved the highest AUC score with the SVM classifier with TNR and TPR equally high

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

scores, Table 5.9. In the KNN though, it performed poorly with an AUC score of 0.77, Table 5.10. The *mAP*-KL achieved its best score with the SVM classifier (AUC=0.91) and underperformed with the RF classifier. Regarding the 5-CV classification, the results were very promising since all methods but PCA achieved the highest score i.e. 1.00 for all metrics and classifiers.

Table 5.8: The classification results in LGMD2B and NM neuromuscular diseases according to RF classifier

FS methods	5-CV			Hold-out Validation		
	AUC	TNR	TPR	AUC	TNR	TPR
LGMD2B						
RF-MDA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.73	1.00
maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.64	1.00
PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	0.55	1.00
BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.98	0.73	1.00
maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.91	0.64	1.00
Rnd	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.90	0.56	1.00
SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.88	0.73	1.00
HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.82	0.64	1.00
t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.82	0.73	0.67
ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.73	0.45	1.00
<i>mAP</i> -KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.70	0.36	0.67
SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.52	0.27	1.00
eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.48	0.27	0.67
cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.36	0.09	1.00
PCA	0.89 (0.25)	0.74 (0.38)	0.61 (0.44)	0.21	0.09	1.00
NM						
SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.90	0.77	1.00
t-test	1.00 (0.00)	0.98 (0.10)	1.00 (0.00)	0.89	0.77	0.80
HykGene	1.00 (0.00)	1.00 (0.00)	0.99 (0.07)	0.88	0.69	0.80
maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.80	0.69	0.80
cat	1.00 (0.00)	1.00 (0.00)	0.99 (0.07)	0.78	0.46	1.00
<i>mAP</i> -KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.74	0.69	0.60
Rnd	0.98 (0.03)	0.87 (0.09)	0.96 (0.06)	0.67	0.49	0.76
SAM	1.00 (0.00)	0.87 (0.28)	0.98 (0.10)	0.65	0.15	1.00
PCA	0.82 (0.30)	0.77 (0.35)	0.73 (0.39)	0.55	0.92	0.40
BGA-COA	0.96 (0.14)	0.87 (0.28)	0.91 (0.19)	0.47	0.23	0.60
PLS-CV	0.97 (0.12)	0.87 (0.28)	0.99 (0.07)	0.42	0.08	1.00
maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.37	0.38	0.40
ODP	1.00 (0.00)	0.92 (0.23)	1.00 (0.00)	0.25	0.38	0.20
RF-MDA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.22	0.15	0.60
eBayes	-	-	-	-	-	-

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

Table 5.9: The classification results in LGMD2B and NM neuromuscular diseases according to SVM classifier

	FS methods	5-CV			Hold-out Validation		
		AUC	TNR	TPR	AUC	TNR	TPR
LGMD2B	RF-MDA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.95	0.91	1.00
	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.91	0.82	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.91	0.82	1.00
	HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.82	0.64	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.82	0.64	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.77	0.55	1.00
	Rnd	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.75	0.53	0.97
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.73	0.45	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.68	0.36	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.68	0.36	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.68	0.36	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.55	0.09	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.55	0.09	1.00
	PCA	0.48 (0.30)	0.58 (0.44)	0.37 (0.43)	0.26	0.18	0.33
NM	SAM	0.94 (0.14)	0.88 (0.28)	1.00 (0.00)	0.90	1.00	0.80
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.88	0.77	1.00
	ODP	0.94 (0.14)	0.87 (0.28)	1.00 (0.00)	0.86	0.92	0.80
	PLS-CV	0.87 (0.16)	0.84 (0.29)	0.90 (0.23)	0.85	0.69	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.81	0.62	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.77	0.54	1.00
	Rnd	0.95 (0.04)	0.94 (0.08)	0.96 (0.05)	0.75	0.75	0.76
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.75	0.69	0.80
	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.70	1.00	0.40
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.68	0.77	0.60
	HykGene	0.95 (0.14)	0.90 (0.27)	0.99 (0.07)	0.61	1.00	0.60
	BGA-COA	0.94 (0.14)	0.88 (0.28)	1.00 (0.00)	0.55	0.31	0.80
	RF-MDA	0.99 (0.05)	1.00 (0.00)	0.98 (0.10)	0.47	0.54	0.40
	PCA	0.51 (0.28)	0.60 (0.40)	0.42 (0.44)	0.43	0.46	0.40
	eBayes	-	-	-	-	-	-

Table 5.10: The classification results in LGMD2B and NM neuromuscular diseases according to KNN classifier

	FS methods	5-CV			Hold-out Validation		
		AUC	TNR	TPR	AUC	TNR	TPR
LGMD2B	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.91	0.82	1.00
	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.86	0.73	1.00
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.86	0.73	1.00
	PLS-CV	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.86	0.73	1.00
	HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.82	0.64	1.00
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.77	0.55	1.00
	RF-MDA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.77	0.55	1.00
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.73	0.45	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.73	0.45	1.00
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.73	0.45	1.00
	Rnd	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.70	0.47	0.93
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.55	0.09	1.00
	eBayes	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.55	0.09	1.00
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.55	0.09	1.00
	PCA	0.76 (0.26)	0.84 (0.34)	0.68 (0.43)	0.23	0.45	0.00
NM	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.90	1.00	0.80
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.88	0.77	1.00
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.82	0.85	0.80
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.82	0.85	0.80
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.76	0.92	0.60
	HykGene	0.93 (0.14)	0.85 (0.27)	1.00 (0.00)	0.75	0.69	0.80
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.75	0.69	0.80
	Rnd	0.96 (0.04)	0.96 (0.05)	0.95 (0.05)	0.75	0.66	0.84
	PLS-CV	0.88 (0.17)	0.88 (0.28)	0.88 (0.26)	0.71	0.62	0.80
	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.63	0.46	0.80
	mAP-KL	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.62	0.85	0.40
	BGA-COA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.62	0.23	1.00
	PCA	0.63 (0.28)	0.66 (0.36)	0.59 (0.44)	0.57	0.54	0.60
	RF-MDA	0.96 (0.11)	1.00 (0.00)	0.91 (0.22)	0.52	0.23	0.80
	eBayes	-	-	-	-	-	-

Likewise in NM validation, all of the methods faced considerable difficulties in distinguishing disease and normal samples. In RF classifier only the SNR, the t-test and the HykGene methods managed to reach an AUC score close to 0.90, Table 5.8. On the other hand, the SAM method achieved an AUC score of 0.90 in SVM and KNN classifiers with the same TNR and TPR scores of 1.00 and 0.80 respectively, Tables 5.9, 5.10. In this dataset the mAP-KL failed to achieve comparable results to the top methods mainly due to the difficulty to discern the disease samples in the validation test (TPR << 1.00). In contrast, during the 5-CV with the RF classifier ten methods achieved AUC score of 1.00, but only mAP-KL, maxT, maxT (200), RF-MDA, and SNR achieved the optimum score in TNR and TPR metrics. Though, with the SVM and KNN classifiers

the majority of the methods excelled in all three metrics. The PLS-CV and BGA-COA had the same TNR score (0.87) but different TPR (0.99 and 0.91) and AUC (0.97 and 0.96). The PCA method had the worst overall performance whereas the eBayes method failed to produce a list of significant genes.

5.4. Cancer data

As far as the large patient cohorts is concerned, we utilized microarray data from four different types of cancer (breast cancer, colon cancer, leukemia, and prostate cancer), with train sets length ranging from 30 to 102 samples and test sets from 19 to 34. In breast cancer hold-out validation with RF classifier, *mAP-KL* attained the optimum score (1.00) in TNR metric and the best AUC score (0.87). Two methods, PLS-CV and RF-MDA, achieved competitive TNR and AUC scores of 0.86 and 0.82 respectively. However, all methods faced difficulties to distinguish the non-responsive samples, and except the *maxT* (200) with a TPR score of 0.83, followed by the RF-MDA, the HykGene and the SAM methods (0.75), Table 5.11. Though, taking into account the classification results in the SVM and KNN classifiers, the methods with the most robust performance were the *maxT*(200) and the SNR methods, Tables 5.12, 5.13. The rest of the methods had a fluctuated performance, including the *mAP-KL* method. During the 5-CV validation, PLSCV, RF-MDA, HykGene and *cat* had an AUC score of 0.91, which was also the highest score attained with the RF classifier. The *cat* and HykGene methods also achieved the highest AUC score (0.82) with SVM classifier, whereas *cat* outperformed all other methods with the KNN classifier having an AUC score of 0.85. Regarding the *mAP-KL*, it had average performance with RF and SVM classifiers (AUC scores of 0.80 and 0.71 respectively) but failed during the KNN classification scheme. The eBayes method similarly to NM dataset failed to fulfill the analysis task.

Table 5.11: The classification results in breast and colon cancers according to RF classifier

FS method	5-CV			Hold-out Validation		
	AUC	TNR	TPR	AUC	TNR	TPR
<i>mAP-KL</i>	0.80 (0.11)	0.79 (0.16)	0.73 (0.18)	0.87	1.00	0.50
<i>maxT</i> (200)	0.85 (0.11)	0.83 (0.13)	0.69 (0.17)	0.83	0.71	0.83
PLS-CV	0.91 (0.08)	0.85 (0.13)	0.77 (0.15)	0.82	0.86	0.42
RF-MDA	0.91 (0.07)	0.91 (0.11)	0.70 (0.16)	0.82	0.86	0.75
<i>maxT</i>	0.87 (0.10)	0.84 (0.13)	0.74 (0.18)	0.77	0.71	0.58
SAM	0.82 (0.11)	0.79 (0.15)	0.69 (0.19)	0.77	0.71	0.75
SNR	0.86 (0.10)	0.85 (0.14)	0.72 (0.20)	0.77	0.71	0.67
BGA-COA	0.83 (0.10)	0.79 (0.15)	0.67 (0.15)	0.76	0.57	0.58
HykGene	0.91 (0.06)	0.86 (0.12)	0.76 (0.17)	0.76	0.71	0.75
Rnd	0.79 (0.01)	0.76 (0.03)	0.65 (0.03)	0.76	0.70	0.78
<i>cat</i>	0.91 (0.07)	0.86 (0.12)	0.78 (0.16)	0.75	0.71	0.50
PCA	0.72 (0.14)	0.66 (0.18)	0.56 (0.19)	0.75	0.43	0.67
ODP	0.83 (0.10)	0.80 (0.14)	0.69 (0.18)	0.74	0.71	0.58
t-test	0.82 (0.10)	0.81 (0.14)	0.69 (0.19)	0.73	0.71	0.58
eBayes	-	-	-	-	-	-

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

	mAP-KL	0.99 (0.03)	0.95 (0.12)	0.97 (0.09)	0.89	0.71	0.84
	BGA-COA	0.98 (0.06)	0.89 (0.22)	0.87 (0.19)	0.87	0.71	0.80
	Rnd	0.98 (0.02)	0.90 (0.06)	0.90 (0.03)	0.84	0.73	0.82
	maxT(200)	1.00 (0.00)	0.94 (0.13)	0.94 (0.13)	0.83	0.71	0.88
	PCA	0.79 (0.19)	0.80 (0.23)	0.72 (0.26)	0.83	0.43	0.84
	ODP	0.99 (0.03)	0.97 (0.13)	0.93 (0.13)	0.82	0.71	0.80
	HykGene	0.98 (0.06)	0.93 (0.14)	0.95 (0.12)	0.81	0.71	0.88
COLON	RF-MDA	0.99 (0.03)	0.96 (0.11)	0.93 (0.13)	0.81	0.71	0.80
	eBayes	0.99 (0.03)	0.97 (0.11)	0.93 (0.13)	0.80	0.71	0.80
	SAM	1.00 (0.02)	0.99 (0.09)	0.93 (0.13)	0.80	0.71	0.80
	cat	0.99 (0.04)	0.97 (0.14)	0.93 (0.13)	0.80	0.57	0.80
	maxT	1.00 (0.02)	0.97 (0.10)	0.94 (0.13)	0.79	0.71	0.80
	PLS-CV	1.00 (0.02)	0.94 (0.16)	0.94 (0.13)	0.79	0.71	0.80
	SNR	0.99 (0.03)	1.00 (0.00)	0.93 (0.13)	0.79	0.71	0.80
	t-test	0.99 (0.03)	0.99 (0.05)	0.93 (0.13)	0.79	0.71	0.80

In relation to colon cancer, the mAP-KL method excelled over the other methods in RF and SVM classifiers with AUC scores of 0.89 and 0.87 respectively, Tables 5.11, 5.12. Particularly, in SVM classifier the second best performance achieved by a bunch of methods with AUC score of 0.80. During the classification with the KNN classifier, the performance of almost all of the methods range from 0.78 to 0.80. Contrary to breast cancer, the TPR scores were higher than the TNR scores and range from 0.80 to 0.92 for all methods but PCA with KNN classifier. The classification results in 5-CV were very promising with AUC values above 0.90 for the majority of the methods with the exception of PCA, which attained much lower AUC scores below 0.79.

Table 5.12: The classification results in breast and colon cancers according to SVM classifier

FS methods	5-CV			Hold-out Validation			
	AUC	TNR	TPR	AUC	TNR	TPR	
	maxT(200)	0.75 (0.10)	0.79 (0.15)	0.72 (0.17)	0.79	1.00	0.58
	SNR	0.76 (0.11)	0.80 (0.15)	0.72 (0.20)	0.76	0.86	0.67
	Rnd	0.69 (0.04)	0.74 (0.04)	0.63 (0.05)	0.74	0.73	0.75
	maxT	0.77 (0.08)	0.77 (0.14)	0.78 (0.14)	0.73	0.71	0.75
	ODP	0.74 (0.11)	0.76 (0.14)	0.72 (0.21)	0.73	0.71	0.75
BREAST	SAM	0.70 (0.12)	0.73 (0.16)	0.67 (0.19)	0.73	0.71	0.75
	cat	0.82 (0.10)	0.87 (0.13)	0.77 (0.15)	0.72	0.86	0.58
	t-test	0.76 (0.10)	0.76 (0.15)	0.76 (0.18)	0.69	0.71	0.67
	RF-MDA	0.79 (0.09)	0.87 (0.13)	0.70 (0.16)	0.66	0.57	0.75
	mAP-KL	0.71 (0.10)	0.75 (0.14)	0.67 (0.15)	0.64	0.86	0.42
	BGA-COA	0.64 (0.11)	0.71 (0.16)	0.56 (0.20)	0.61	0.71	0.50
	HykGene	0.82 (0.08)	0.84 (0.13)	0.80 (0.13)	0.57	0.71	0.42
	PCA	0.51 (0.12)	0.60 (0.20)	0.42 (0.20)	0.55	0.43	0.67
	PLS-CV	0.77 (0.07)	0.77 (0.13)	0.77 (0.13)	0.55	0.86	0.25
	eBayes	-	-	-	-	-	-

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

COLON	mAP-KL	0.94 (0.09)	0.95 (0.12)	0.93 (0.13)	0.87	0.86	0.88
	cat	0.98 (0.05)	0.96 (0.11)	1.00 (0.00)	0.80	0.71	0.88
	eBayes	0.99 (0.04)	0.98 (0.08)	1.00 (0.00)	0.80	0.71	0.88
	HykGene	0.84 (0.14)	0.75 (0.25)	0.93 (0.14)	0.80	0.71	0.88
	maxT	0.96 (0.07)	0.93 (0.14)	0.99 (0.07)	0.80	0.71	0.88
	maxT (200)	0.95 (0.08)	0.94 (0.13)	0.96 (0.11)	0.80	0.71	0.88
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.80	0.71	0.88
	PLS-CV	0.92 (0.11)	0.89 (0.20)	0.95 (0.12)	0.80	0.71	0.88
	RF-MDA	0.93 (0.10)	0.87 (0.16)	0.99 (0.07)	0.80	0.71	0.88
	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.80	0.71	0.88
	SNR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.80	0.71	0.88
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.80	0.71	0.88
	BGA-COA	0.99 (0.05)	1.00 (0.00)	0.97 (0.09)	0.78	0.71	0.84
	Rnd	0.92 (0.04)	0.91 (0.05)	0.93 (0.05)	0.77	0.67	0.85
	PCA	0.70 (0.18)	0.71 (0.27)	0.69 (0.24)	0.53	0.14	0.92

Table 5.13: The classification results in breast and colon cancers according to KNN classifier

	FS methods	5-CV			Hold-out Validation		
		AUC	TNR	TPR	AUC	TNR	TPR
BREAST	maxT (200)	0.74 (0.11)	0.79 (0.16)	0.69 (0.13)	0.76	0.86	0.67
	RF-MDA	0.71 (0.10)	0.74 (0.13)	0.69 (0.15)	0.73	0.71	0.75
	SNR	0.69 (0.11)	0.80 (0.15)	0.58 (0.19)	0.73	0.71	0.75
	BGA-COA	0.62 (0.11)	0.76 (0.13)	0.48 (0.17)	0.72	0.86	0.58
	ODP	0.73 (0.12)	0.77 (0.15)	0.70 (0.17)	0.72	0.86	0.58
	Rnd	0.66 (0.04)	0.71 (0.04)	0.62 (0.07)	0.70	0.77	0.62
	SAM	0.67 (0.12)	0.79 (0.14)	0.54 (0.21)	0.69	0.71	0.67
	cat	0.85 (0.10)	0.88 (0.10)	0.82 (0.16)	0.68	0.86	0.50
	t-test	0.70 (0.11)	0.78 (0.15)	0.63 (0.17)	0.68	0.86	0.50
	PLS-CV	0.68 (0.10)	0.74 (0.15)	0.62 (0.16)	0.64	0.86	0.42
	maxT	0.78 (0.10)	0.80 (0.13)	0.76 (0.17)	0.61	0.71	0.50
	PCA	0.60 (0.10)	0.56 (0.18)	0.63 (0.14)	0.54	0.57	0.50
	HykGene	0.73 (0.08)	0.83 (0.12)	0.63 (0.16)	0.52	0.71	0.33
	mAP-KL	0.57 (0.11)	0.53 (0.14)	0.60 (0.19)	0.30	0.43	0.17
	eBayes	-	-	-	-	-	-

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.80	0.71	0.88
	eBayes	0.98 (0.05)	0.96 (0.11)	1.00 (0.00)	0.80	0.71	0.88
	HykGene	0.92 (0.10)	0.87 (0.16)	0.96 (0.11)	0.80	0.71	0.88
	maxT	0.97 (0.07)	0.93 (0.13)	1.00 (0.00)	0.80	0.71	0.88
	ODP	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.80	0.71	0.88
	RF-MDA	0.96 (0.07)	0.93 (0.13)	0.99 (0.05)	0.80	0.71	0.88
COLON	SAM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.80	0.71	0.88
	SNR	1.00 (0.00)	1.00 (0.00)	0.99 (0.05)	0.80	0.71	0.88
	t-test	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.80	0.71	0.88
	Rnd	0.93 (0.04)	0.92 (0.06)	0.93 (0.03)	0.79	0.74	0.84
	BGA-COA	0.99 (0.03)	0.99 (0.05)	0.99 (0.05)	0.78	0.71	0.84
	mAP-KL	0.96 (0.07)	0.99 (0.07)	0.93 (0.13)	0.78	0.71	0.84
	maxT (200)	0.95 (0.08)	0.96 (0.11)	0.94 (0.13)	0.78	0.71	0.84
	PLS-CV	0.99 (0.03)	0.99 (0.07)	1.00 (0.00)	0.78	0.71	0.84
	PCA	0.65 (0.18)	0.59 (0.24)	0.71 (0.25)	0.59	0.43	0.76

Concerning the leukemia dataset, three methods, BGA-COA, maxT (200) and eBayes, excelled in RF classifier while other four methods achieved an AUC score of 0.99. The mAP-KL, although achieving high classification scores in 5-CV, failed to predict correctly all AML samples (TPR = 0.43), and as a result its overall performance was 0.71 during the hold-out validation, Table 5.14. The cat method, was the method that overall achieved the best performance across all classifiers, with an average AUC score close to 1.00. Specifically in SVM and KNN classifiers had scores of 1.00 in AUC, TNR and TPR metrics, Tables 5.15, 5.16. On the other hand, the mAP-KL method appeared the same behavior with the RF classifier and had considerably low TPR scores that led to low AUC scores. Interestingly, the PCA, SNR and t-test methods failed to predict any or almost any of the 14 AML samples, although they identified all or almost all of the ALL samples. Similarly, those three methods achieved poorly results during the 5-CV compared to the other methods. Finally, the ODP algorithm failed to analyze the leukemia dataset.

Finally, in prostate cancer, no method succeeded in discriminating the samples in both types of validation, alike to NM in neuromuscular diseases section. Even more importantly, during the hold-out validation, three of the methods (SNR, t-test, maxT(200)) failed to identify even a single sample from the normal class across classifiers, Tables 5.14, 5.15, 5.16, whereas others, like eBayes, SAM and maxT, failed in two out of three classifiers i.e TNR=0.00. However, because of the normal/disease ratio (9 normal and 25 disease samples), the AUC values of some methods e.g. eBayes (0.86) and SAM (0.92) are deceptive. Conversely, the mAP-KL method achieved a notable performance across all classifiers, with high AUC scores (0.80, 0.94, 0.90) and non-zero TNR and TPR scores. The zero TNR score was also present in 5-CV by the SNR and t-test methods. The rest of the classification results were either close to the hold-out classification results or fairly optimistic. Besides, the ODP and cat algorithms, failed to deal with the prostate data.

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

Table 5.14: The classification results in leukemia and prostate cancers according to RF classifier

	FS methods	5-CV			Hold-out Validation			
		AUC	TNR	TPR	AUC	TNR	TPR	
LEUKEMIA	BGA-COA	0.99 (0.04)	1.00 (0.00)	0.81 (0.27)	1.00	1.00	0.86	
	maxT(200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	0.86	
	eBayes	1.00 (0.00)	1.00 (0.00)	0.91 (0.19)	1.00	0.95	0.93	
	RF-MDA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.99	1.00	0.86	
	PLS-CV	1.00 (0.00)	1.00 (0.00)	0.89 (0.25)	0.99	0.95	0.93	
	SAM	1.00 (0.00)	1.00 (0.00)	0.91 (0.19)	0.99	0.95	0.93	
	cat	1.00 (0.00)	1.00 (0.00)	0.95 (0.14)	0.99	0.95	0.93	
	HykGene	1.00 (0.00)	1.00 (0.00)	0.90 (0.20)	0.97	0.90	0.93	
	Rnd	0.99 (0.01)	0.98 (0.02)	0.86 (0.06)	0.97	0.99	0.75	
	maxT	1.00 (0.02)	0.98 (0.07)	0.85 (0.27)	0.96	1.00	0.64	
	mAP-KL	1.00 (0.00)	1.00 (0.00)	0.97 (0.17)	0.71	0.90	0.43	
	PCA	0.56 (0.16)	1.00 (0.00)	0.00 (0.00)	0.64	0.95	0.14	
	SNR	0.50 (0.00)	1.00 (0.00)	0.00 (0.00)	0.50	1.00	0.00	
	t-test	0.50 (0.00)	1.00 (0.00)	0.00 (0.00)	0.50	1.00	0.00	
	ODP	-	-	-	-	-	-	
	PROSTATE	SAM	0.96 (0.04)	0.97 (0.05)	0.88 (0.10)	0.92	0.00	1.00
		maxT(200)	0.95 (0.10)	0.95 (0.10)	0.89 (0.10)	0.88	0.00	1.00
PLS-CV		0.97 (0.03)	0.95 (0.08)	0.92 (0.07)	0.87	0.33	1.00	
eBayes		0.96 (0.04)	0.98 (0.04)	0.89 (0.10)	0.86	0.00	1.00	
RF-MDA		0.97 (0.04)	0.97 (0.06)	0.90 (0.09)	0.83	0.11	1.00	
mAP-KL		0.93 (0.06)	0.90 (0.09)	0.85 (0.11)	0.80	1.00	0.36	
BGA-COA		0.95 (0.05)	0.91 (0.09)	0.89 (0.10)	0.73	0.22	0.88	
Rnd		0.93 (0.02)	0.89 (0.04)	0.86 (0.03)	0.70	0.18	0.94	
HykGene		1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.69	0.89	0.24	
maxT		0.89 (0.07)	0.88 (0.09)	0.79 (0.13)	0.50	0.00	1.00	
PCA		0.84 (0.09)	0.77 (0.15)	0.75 (0.15)	0.50	0.00	1.00	
SNR		0.50 (0.00)	0.08 (0.27)	0.92 (0.27)	0.50	0.00	1.00	
t-test		0.50 (0.00)	0.08 (0.27)	0.92 (0.27)	0.50	0.00	1.00	
ODP		-	-	-	-	-	-	
cat	-	-	-	-	-	-		

Table 5.15: The classification results in leukemia and prostate cancers according to SVM classifier

FS methods	5-CV			Hold-out Validation			
	AUC	TNR	TPR	AUC	TNR	TPR	
LEYKEMIA	cat	1.00 (0.04)	1.00 (0.00)	0.99 (0.07)	1.00	1.00	1.00
	eBayes	0.99 (0.05)	1.00 (0.00)	0.98 (0.10)	0.96	1.00	0.93
	HykGene	1.00 (0.04)	1.00 (0.00)	0.99 (0.07)	0.96	1.00	0.93
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.96	1.00	0.93
	SAM	1.00 (0.04)	1.00 (0.00)	0.99 (0.07)	0.96	1.00	0.93
	BGA-COA	0.92 (0.11)	1.00 (0.00)	0.85 (0.23)	0.94	0.95	0.93
	PLS-CV	0.92 (0.14)	1.00 (0.00)	0.84 (0.27)	0.94	0.95	0.93
	RF-MDA	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.94	0.95	0.93
	Rnd	0.95 (0.05)	0.97 (0.03)	0.92 (0.08)	0.90	0.97	0.84
	maxT	0.97 (0.07)	0.99 (0.04)	0.96 (0.13)	0.82	1.00	0.64
	mAP-KL	0.98 (0.04)	0.96 (0.07)	1.00 (0.00)	0.75	1.00	0.50
	PCA	0.50 (0.00)	1.00 (0.00)	0.00 (0.00)	0.50	1.00	0.00
	SNR	0.50 (0.00)	1.00 (0.00)	0.00 (0.00)	0.50	1.00	0.00
	t-test	0.50 (0.00)	1.00 (0.00)	0.00 (0.00)	0.50	1.00	0.00
	ODP	-	-	-	-	-	-
PROSTATE	BGA-COA	0.92 (0.06)	0.92 (0.09)	0.92 (0.08)	0.98	1.00	0.96
	mAP-KL	0.88 (0.07)	0.89 (0.09)	0.86 (0.10)	0.94	0.89	1.00
	PCA	0.77 (0.08)	0.82 (0.14)	0.73 (0.12)	0.90	1.00	0.80
	maxT	0.82 (0.07)	0.82 (0.11)	0.82 (0.11)	0.83	0.78	0.88
	HykGene	0.86 (0.07)	0.84 (0.12)	0.87 (0.09)	0.78	0.56	1.00
	PLS-CV	0.95 (0.05)	0.94 (0.08)	0.95 (0.05)	0.56	0.11	1.00
	Rnd	0.88 (0.02)	0.89 (0.04)	0.88 (0.03)	0.51	0.01	1.00
	eBayes	0.94 (0.05)	0.98 (0.04)	0.90 (0.09)	0.50	0.00	1.00
	maxT (200)	0.95 (0.04)	0.97 (0.05)	0.92 (0.07)	0.50	0.00	1.00
	RF-MDA	0.95 (0.05)	0.95 (0.07)	0.94 (0.07)	0.50	0.00	1.00
	SAM	0.92 (0.05)	0.95 (0.07)	0.88 (0.09)	0.50	0.00	1.00
	SNR	0.50 (0.00)	0.00 (0.00)	1.00 (0.00)	0.50	0.00	1.00
	t-test	0.50 (0.00)	0.00 (0.00)	1.00 (0.00)	0.50	0.00	1.00
	cat	-	-	-	-	-	-
	ODP	-	-	-	-	-	-

Table 5.16: The classification results in leukemia and prostate cancers according to KNN classifier

FS methods	5-CV			Hold-out Validation			
	AUC	TNR	TPR	AUC	TNR	TPR	
LEYKEMIA	cat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00	1.00	1.00
	BGA-COA	0.96 (0.09)	1.00 (0.00)	0.92 (0.18)	0.96	1.00	0.93
	eBayes	0.99 (0.08)	1.00 (0.00)	0.97 (0.16)	0.96	1.00	0.93
	maxT (200)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.96	1.00	0.93
	PLS-CV	0.90 (0.14)	0.96 (0.07)	0.83 (0.27)	0.94	0.95	0.93
	SAM	0.99 (0.08)	1.00 (0.00)	0.97 (0.16)	0.93	1.00	0.86
	RF-MDA	0.96 (0.09)	1.00 (0.00)	0.91 (0.19)	0.90	0.95	0.86
	HykGene	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.89	1.00	0.79
	maxT	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.89	1.00	0.79
	Rnd	0.96 (0.03)	0.96 (0.03)	0.96 (0.04)	0.87	0.93	0.82
	mAP-KL	0.94 (0.09)	0.94 (0.09)	0.93 (0.17)	0.66	0.90	0.43
	PCA	0.57 (0.12)	1.00 (0.00)	0.00 (0.00)	0.53	0.95	0.14
	SNR	0.50 (0.00)	1.00 (0.00)	0.00 (0.00)	0.50	1.00	0.00
	t-test	0.50 (0.00)	1.00 (0.00)	0.00 (0.00)	0.50	1.00	0.00
	ODP	-	-	-			
PROSTATE	mAP-KL	0.82 (0.09)	0.82 (0.13)	0.83 (0.11)	0.90	1.00	0.80
	SAM	0.86 (0.08)	0.84 (0.12)	0.89 (0.10)	0.89	0.78	1.00
	eBayes	0.91 (0.07)	0.91 (0.11)	0.91 (0.08)	0.87	0.78	0.96
	PLS-CV	0.91 (0.06)	0.93 (0.07)	0.90 (0.09)	0.85	0.78	0.92
	RF-MDA	0.93 (0.05)	0.94 (0.07)	0.91 (0.08)	0.84	0.89	0.80
	BGA-COA	0.87 (0.07)	0.87 (0.10)	0.87 (0.11)	0.67	0.33	1.00
	Rnd	0.80 (0.05)	0.82 (0.07)	0.77 (0.04)	0.63	0.29	0.97
	HykGene	0.88 (0.06)	0.90 (0.09)	0.87 (0.11)	0.50	1.00	0.00
	maxT	0.80 (0.09)	0.85 (0.13)	0.75 (0.14)	0.50	0.00	1.00
	maxT (200)	0.82 (0.09)	0.89 (0.09)	0.76 (0.13)	0.50	0.00	1.00
	PCA	0.67 (0.11)	0.66 (0.15)	0.68 (0.15)	0.50	1.00	0.00
	SNR	0.50 (0.00)	0.00 (0.00)	1.00 (0.00)	0.50	0.00	1.00
	t-test	0.50 (0.00)	0.00 (0.00)	1.00 (0.00)	0.50	0.00	1.00
	cat	-	-	-			
	ODP	-	-	-			

5.5. Analysis of previous experiments

At a different level of assessment, we compared the mAP-KL's classification results of the specific cancer datasets, against those published in previous classification studies of the same data. For the purposes of this comparison, we have cited the author's name, the classification type, the number of the features used, and finally the achieved accuracy (ACC). Since we utilized three different classifiers to build and test mAP-KL's models, in this comparison we present all three results achieved.

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

In relation to the van 't Veer et al. [23] breast cancer datasets, we present the classification results from 9 different approaches stemming from 7 studies, see Table 5.17. Regarding the CV test, Hassan et.al [21] and Hu et al. [60] achieved ACC above 90.00%, higher than van 't Veer et al. and with less features. However, they utilized all of the samples contrary to van 't Veer et al. Our method achieved moderate results (ACC = 75.93%) as absolute numbers for the 78 samples but with only 6 features and 5-CV contrary to LOO-CV that engaged by the others. In the hold-out test, although the ACC of mAP-KL is the lowest score, we did manage to identify correctly all responsive samples. However, we should consider why we discern only half of the non-responsive samples (type II error).

Singh et al. [40] first employed the specific prostate cancer datasets and we have included the results from three studies, Table 5.18. mAP-KL with the aid of SVM-linear classifier, misclassified one sample in hold-out validation just like Liu et al. [67]. However, in CV we misclassified approximately eight samples more than Liu et al., but with only 12 genes.

The ALL/AML discrimination in the leukemia datasets, Table 5.19, as first presented by Golub et al. [6], is the one most often analyzed among the datasets considered. More than 16 studies and 29 methods have based their evaluation on this set of data. Comparing mAP-KL to Golub classification results, we notice that in CV we identify one more sample, whereas in hold-out we misclassify two samples from Golub, though we did that with only 5 genes. There are many methods that distinguish correctly all samples in CV although only Hewett and Kijisanayothin [38] achieved an ACC of 98.61% with only two genes, but using all of the 72 samples. Regarding the hold-out validation, several methods achieved high classification scores with ACC above 95.00%, though only Mukherjee et al. [70] reached the 100%, with only 40 genes. Liu et al. [67] predict correctly all samples in both validation assessments, but we are unaware of the subset's length.

Finally, fourteen methods employed the Alon et al. [43] colon cancer datasets to assess their classification performance, see Table 5.20. During the CV assessment we achieved ACC = 96.00% with RF and KNN classifiers higher than the one achieved by Tan and Gilbert [63] (95.16%). Regarding the hold-out validation, Li et al. [64], Nguyen and Rocke [65] and Furey et al. [66] achieved ACC of 94.1%, 93.5% and 90.30% respectively. We reached to 81.25% and 87.50% ACC with 20 genes contrary to Nguyen and Rocke with 50 genes.

Table 5.17: An overview of the published classification results in van 't Veer et al. breast cancer data

Authors	Cross Validation		Train-Test		Features
	Samples	Accuracy (%)	Samples	Accuracy (%)	
van 't Veer et al., 2002 [3]	65/78	83.3	17/19	89.5	70
Hassan et al., 2009 [42]	-	92.13	-	91.67	3
Shen et al., 2006 [84]	60/78	76.90	15/19	78.9	231
Shen et al., 2006	-	76.20	15/19	78.9	231
Shen et al., 2006	62/78	81.40	17/19	89.5	44
Hu et al., 2006 [85]	88/97	90.7	-	-	50
Moon et al., 2006 [86]	49/78	62.9	-	-	-
Tan and Gilbert, 2003 [87]	-	-	17/19	89.47	834
Hewett and Kijisanayothin, 2008 [69]	66/97	68.04	-	-	8
<hr/>					
mAP-KL (RF)	-	75.93	13/19	68.42	6
mAP-KL (KNN)	-	56.35	5/19	26.32	6
mAP-KL (SVM-linear)	-	71.47	11/19	57.89	6

Table 5.18: An overview of the published classification results in Singh et al. prostate cancer data

Authors	Cross Validation		Train-Test		Features
	Samples	Accuracy (%)	Samples	Accuracy (%)	
Liu et al., 2004 [88]	98/102	96.08	33/34	97.06	-
Tan and Gilbert, 2003 [87]	-	-	25/34	73.53	3071
Hewett and Kijisanayothin, 2008 [69]	124/136	91.18	-	-	6
<hr/>					
mAP-KL (RF)	-	87.33	18/34	52.94	12
mAP-KL (KNN)	-	82.22	29/34	85.29	12
mAP-KL (SVM-linear)	-	87.82	33/34	97.06	12

Table 5.19: An overview of the published classification results in Golub et al. ALL/AML leukemia data

Authors	Cross Validation		Train-Test		Features
	Samples	Accuracy (%)	Samples	Accuracy (%)	
Golub et al., 1999 [81]	36/38	94.73	29/34	85.29	50
Liu et al., 2004 [88]	38/38	100.00	34/34	100	-
Liu et al., 2004	-	-	33/34	97.06	-
Li et al., 2001 [89]	-	-	-	94.1	-
Furey et al., 2000 [90]	-	-	-	94.1	-
Ben-Dor et al., 2000 [91]	-	-	-	91.6	-
Ben-Dor et al., 2000	-	-	-	94.4	-
Ben-Dor et al., 2000	-	-	-	95.8	-
Nguyen and Rocke, 2002 [92]	-	-	-	94.17	50
Nguyen and Rocke, 2002	-	-	-	95.44	50
Nguyen and Rocke, 2002	-	-	-	95.94	50
Nguyen and Rocke, 2002	-	-	-	96.44	50
Mukherjee et al., 1999 [93]	38/38	100	31/34	91.17	7129
Mukherjee et al., 1999	38/38	100	34/34	100	999
Mukherjee et al., 1999	38/38	100	32/34	94.11	99
Mukherjee et al., 1999	38/38	100	30/34	88.23	49
Mukherjee et al., 1999	-	-	34/34	100	40
Mukherjee et al., 1999	-	-	32/34	94.11	5
Dudoit et al., 2002 [94]	-	-	-	95.0~	-
Dudoit et al., 2002	-	-	-	95.0~	-
Dudoit et al., 2002	-	-	-	95.0~	-
Antonov et al., 2004 [95]	37/38	98	34/34	100	185
Liu and Chen, 2004 [96]	38/38	100	34/34	100	3800
Tibshirani et al., 2002 [97]	37/38	98	32/34	94.11	21
Moon et al., 2006 [86]	71/72	98.6	-	-	-
Hewett and Kijsanayothin, 2008 [69]	71/72	98.61	-	-	2
Antoniadis et al., 2003 [98]	38/38 (DLDA)	100	33/34 (DLDA)	97.06	50
Hu et al., 2006 [85]	38/38	100	-	-	50
Tan and Gilbert, 2003 [87]	-	-	31/34	91.18	1038
<hr/>					
mAP-KL (RF)	-	98.93	24/34	70.59	5
mAP-KL (KNN)	-	93.61	24/34	70.59	5
mAP-KL (SVM-linear)	-	97.36	27/34	79.41	5

Table 5.20: An overview of the published classification results in Alon et al. colon cancer data

Authors	Cross Validation		Train-Test		Features
	Samples	Accuracy (%)	Samples	Accuracy (%)	
Liu et al., 2004 [88]	57/62	91.94	-	-	-
Liu et al., 2004	53/62	85.48	-	-	-
Furey et al., 2000 [90]	-	-	-	90.3	-
Li et al., 2001 [89]	-	-	-	94.1~	-
Ben-Dor et al., 2000 [91]	-	-	-	80.6	-
Ben-Dor et al., 2000	-	-	-	74.2	-
Ben-Dor et al., 2000	-	-	-	72.6	-
Nguyen and Rocke, 2002 [92]	-	-	-	87.1	-
Nguyen and Rocke, 2002	-	-	-	87.1	-
Nguyen and Rocke, 2002	-	-	-	93.5	50
Nguyen and Rocke, 2002	-	-	-	91.9	1000
Antoniadis et al., 2003 [98]	52/62 (MAVE-LD)	83.87	-	-	50
Hu et al., 2006 [85]	56/62	90.3	-	-	50
Tan and Gilbert, 2003 [87]	59/62	95.16	-	-	135
<hr/>					
mAP-KL (RF)	-	96.00	26/32	81.25	20
mAP-KL (KNN)	-	96.00	26/32	81.25	20
mAP-KL (SVM-linear)	-	94.00	28/32	87.50	20

5.6. Summary

The overall results, based on the RF classifier, as summarized in Figure 5.1 places mAP-KL at the top among twelve (12) other feature selection algorithms developed for the mining of gene expression data. In particular, the mAP-KL method achieved the second best mean AUC in neuromuscular diseases i.e. 0.91 and the sixth best in cancer data. Eventually, the classification performance of mAP-KL across all ten diseases reached the AUC score of 0.86, which is the third best AUC score with the minimum standard deviation value compared to the methods with better classification performance e.g. eBayes, PLS-CV. Hence, we may firmly state that the combination of a univariate and a clustering method isolates subsets of genes that may discriminate unknown samples from a variety of diseases and number of samples quite accurately.

Furthermore, the mAP-KL methodology selects the significant genes without any classifier involvement, thus our method is considered as classifier independent. Indeed, the classification results across three classification algorithms, Figure 5.2, shows a similar classification performance i.e. standard deviation < 0.1 in most of the cases, and certainly no preference towards a particular classifier.

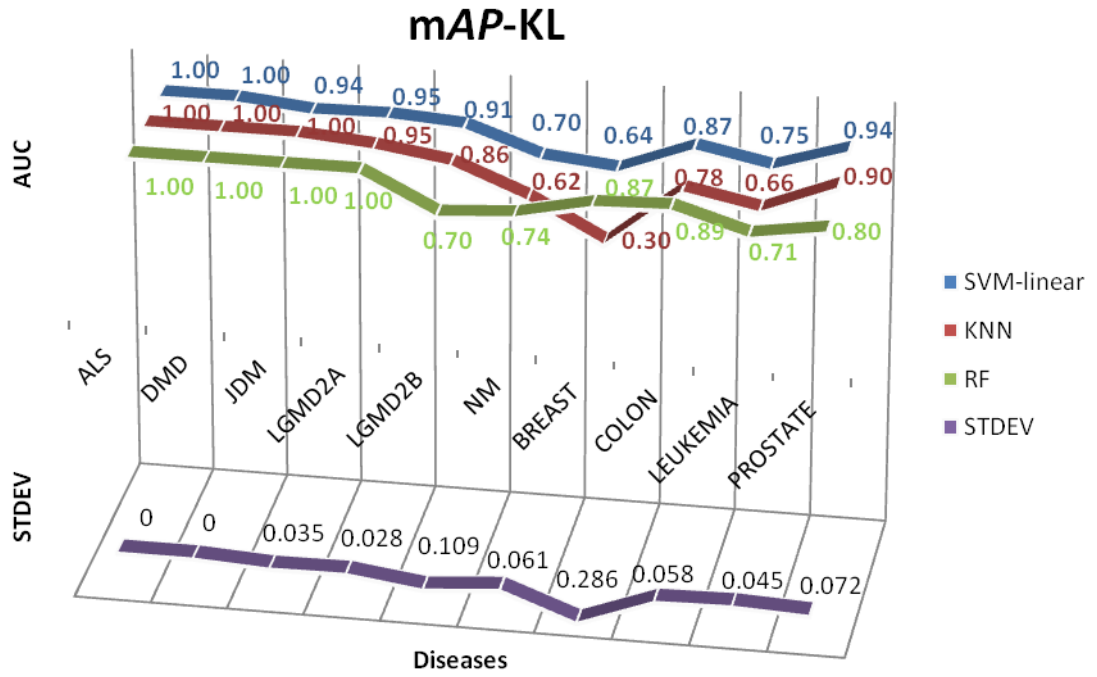
Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

	eBayes	PLS-CV	SAM	BGA-COA	RF-MDA	mAP-KL	cat	Hyk Gene	maxT	ODP	SNR	t-test	PCA	MEAN	
Diseases with Small Sample Size available	ALS	1.00	1.00	1.00	1.00	1.00	1.00	0.64	1.00	1.00	1.00	1.00	1.00	1.00	
	DMD	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.61	0.97	
	JDM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
	LGMD2A	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.94	1.00	0.94	1.00	0.58	0.96	
	LGMD2B	0.48	1.00	0.52	0.98	1.00	0.70	0.36	0.82	0.91	0.73	0.88	0.82	0.21	0.72
	NM	-	0.42	0.65	0.47	0.22	0.74	0.78	0.88	0.37	0.25	0.90	0.89	0.55	0.57
MEAN	0.90	0.90	0.86	0.91	0.87	0.91	0.86	0.88	0.87	0.83	0.95	0.95	0.66		
Diseases with Large Sample Size available	BREAST	-	0.82	0.77	0.76	0.82	0.87	0.75	0.76	0.77	0.74	0.77	0.73	0.75	0.78
	COLON	0.80	0.79	0.80	0.87	0.81	0.89	0.80	0.81	0.79	0.82	0.79	0.79	0.83	0.82
	LEUKEMIA	1.00	0.99	0.99	1.00	0.99	0.71	0.99	0.97	0.96	-	0.50	0.50	0.64	0.84
	PROSTATE	0.86	0.87	0.92	0.73	0.83	0.80	-	0.69	0.50	-	0.50	0.50	0.50	0.70
	MEAN	0.89	0.87	0.87	0.84	0.86	0.82	0.85	0.81	0.76	0.78	0.64	0.63	0.68	
TOTAL MEAN	0.89	0.89	0.87	0.87	0.87	0.86	0.85	0.84	0.81	0.81	0.80	0.79	0.67		
TOTAL STD	0.184	0.185	0.173	0.179	0.242	0.127	0.214	0.129	0.223	0.259	0.191	0.197	0.240		

< 0.50	0.50-0.69	0.70-0.79	0.80-0.89	0.90-0.95	0.96-0.99	1.00

Figure 5.1: The overall classification results (AUC metric) with RF classifier

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases



5.2: The classification performance (AUC) of mAP-KL across diseases for three classifiers

6. Biological relevance of discriminatory gene lists

6.1. Introduction

The power of the proposed FS approach is evident not only from its performance in the statistical metrics, but also from the biological relevance of the selected genes either to a broad range of different molecular pathways and biological processes or more importantly to the respective pathological phenotypes. Therefore we engaged the produced gene lists from our method and the methods that excelled in the classification processes, (eBayes, PLS-CV, SAM, BGA-COA, RF-MDA), as well as the maxT method which is the ranking method of *mAP-KL*, into a series of validations. During those validations we tried to unravel the “semantics” behind those gene lists and its association with the respective diseases.

6.2. The gene lists from a Systems Biology perspective

Usually the initial product of an FS method is a list of ids rather than gene symbols, since the expression data stem from microarray chips technology. Therefore a necessary action that we typically take is to match those probe ids with the relevant gene symbols. Another interesting thing from chip technology is that one gene symbol is regularly represented by more than one probe ids. Thus, an over or under expressed gene may be present in a top ranked list more than one times according to the chip specifications. As a result, those multiple instances of a gene shall be removed from any top ranked list to conclude to a list of unique top genes. This is an essential step regarding the anticipated gene enrichment since a top list of 20 or 50 probe ids may for example represent 14 or 35 unique gene symbols. Furthermore, gene chips include internal and external spiked in controls responsible for the hybridization quality that should be not included in the top ranking of any differential analysis. For all those reasons, the “degree of uniqueness” (DoU) of a top ranked list is a first validation measure directly connected to the list’s potential from a biological standpoint.

In the following tables, Table 6.1 and 6.2, we have cited the number of probe ids and the respective number of gene symbols per method and per dataset. In the last column of the tables we have calculated the DoU value as the average of the division between gene symbols and probe ids. The closest to the unit the more unique is the ranked list. Regarding the neuromuscular data, our method achieved the highest score with the maxT being quite close. In contrast, the BGA-COA had the most discrepancies between ranked probe ids and their respective gene symbols. In relation to cancer data, the eBayes method surpassed the other methods although its average quantity is based on three rather than four datasets. The *mAP-KL* is placed second setting a direct inference about the high “uniqueness” of the produced lists. On the contrary, the RF-MDA failed to identify enough unique gene symbols particularly in the breast cancer dataset and that was the cause for taking the final place.

Table 6.1: The DoU of seven FS methods across neuromuscular data

FS	ALS		DMD		JDM		LGMD2A		LGMD2B		NM		DoU
	Prbs	Gns	Prbs	Gns	Prbs	Gns	Prbs	Gns	Prbs	Gns	Prbs	Gns	
mAP-KL	21	20	14	14	21	20	6	6	15	15	18	18	0.984
maxT	20	20	20	20	20	20	20	20	20	20	20	18	0.983
RF-MDA	20	20	20	20	20	20	20	19	20	20	20	18	0.975
SAM	20	14	20	20	20	18	20	16	20	16	20	20	0.867
eBayes *	20	17	20	20	20	18	20	16	20	15	-	-	0.860
PLS-CV	20	13	20	20	20	19	20	18	20	16	20	17	0.858
BGA-COA	20	15	20	17	20	18	20	14	20	17	20	17	0.817

* The eBayes method evaluated in five datasets

Table 6.2: The DoU of seven FS methods across cancer data

FS	Breast		Colon		Leukemia		Prostate		DoU
	Prbs	Gns	Prbs	Gns	Prbs	Gns	Prbs	Gns	
eBayes *	-	-	20	18	20	18	20	19	0.917
mAP-KL	6	4	20	16	5	5	12	12	0.867
PLS-CV	20	14	20	18	20	19	20	17	0.850
BGA-COA	20	12	20	18	20	19	20	18	0.838
SAM	20	11	20	18	20	18	20	19	0.825
maxT	20	11	20	16	20	17	20	20	0.800
RF-MDA	20	9	20	14	20	18	20	19	0.750

* The eBayes method evaluated in three datasets

A second validation criterion is the enrichment of the unique gene symbols in relation to the associated pathways. Ideally a one-to-one relationship between genes and pathways could embrace all the necessary information for further biological insights. However, this relation is not only hard to achieve since most of the times we have either one-to-many relationship or many-to-one relationships but also can be misleading because one gene is usually involved in more than one pathways that are not necessarily involved in the inspected disease. Hence, taking into consideration the other two types of relationship the most desirable is the many-to-one. By having more than one gene related to the same pathway it is far more certain that this pathway is indeed active and related to the disease. Therefore a gene per pathway ratio between one and two i.e. one or at least two genes per pathway satisfies adequately the concept of genes' enrichment.

At this point is crucial to refer to another parameter before mentioning the results of this validation measure, which are the protein-coding-genes (P-C-Gns) in the ranked list. In essence, not all of the known genes are protein coding and thus involved in molecular functions. Pathway analysis tries to simplify the complexity at the cellular level through the representation of a series of steps where "each step is an event that transforms

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

input physical entities into output entities” [99]. Such entities are definitely the produced proteins, among other small molecules or particles, and as a consequence only the protein coding genes are requisite for a pathway analysis.

Through a plethora of pathway analysis tools, we utilized the “Reactome” pathway database [99], which is a curated and peer reviewed database of pathways and reactions in human biology. We uploaded the top lists of the selected FS methods for all diseases and evaluated their pathway enrichment. During the pathway evaluation we took into consideration the DoU and the number of protein-coding genes parameters as well as the number of pathways according to the “Reactome” database. The final pathway enrichment (PE) score for each FS (m) is the average of the summation of pathways per protein-coding genes multiplied by the DoU for all diseases (d)

$$PE_m = \sum_{d=1}^{10} \frac{Protein-coding-genes_d}{Pathways_d} \times DoU . \quad (6.1)$$

We formed detailed tables per method and disease, Table 6.3 and Table 6.4, to present the outcome of “Reactome” analysis and finally we summarized the results into a graph where the FS methods are in descending order based on their average PE score, Figure 6.1. In accordance with the pathway analysis the maxT method appears to achieve the highest PE score across all diseases. Besides is the method with the second highest DoU score marginally behind mAP-KL. However, this significant advantage over mAP-KL and RF-MDA that follow is mainly due to the PE score in prostate cancer (4.33), where the maxT achieved to identify 3 pathways with 13 unique genes. Otherwise those three methods appear to constitute a group with PE scores close to unit, which is a satisfactory if not intriguing case for biologists. The rest of the methods performed better in cancer data, four datasets, contrary to neuromuscular, six datasets, and that explain their minor deviation from the first three methods.

An additional remark about this pathway analysis has to do with the commonality of the pathways itself among the FS methods. In general, there is a small to moderate overlap among the pathways per method, Table 6.5. However, we cannot state the point that there are good and bad pathway lists having in mind the classification performance because there are no strong evidences that this diversity is directly connected to the classification process. Indeed, the PLS-CV and RF-MDA methods that achieved the highest AUC score in LGMDA2B disease have four out of eight common pathways. On the contrary, the eBayes pathway list owns three out of four of those common pathways though its classification performance is the worst achieved, 0.48 AUC, among the seven FS methods.

Table 6.3: The pathway analysis results on neuromuscular data

FS	ALS						
	Prbs	Gns	P-C-Gns	Pathways	Gns/Pathway	DoU	PE
mAP-KL	21	20	12	8	1.50	0.952	1.428
maxT	20	20	18	18	1.00	1.000	1.000
RF-MDA	20	20	6	8	0.75	1.000	0.750
BGA-COA	20	15	9	10	0.90	0.750	0.675
eBayes	20	17	7	16	0.44	0.850	0.372
PLS-CV	20	13	9	16	0.56	0.650	0.366
SAM	20	14	7	17	0.41	0.700	0.288
DMD							
eBayes	20	20	9	6	1.50	1.000	1.500
SAM	20	20	9	8	1.13	1.000	1.125
RF-MDA	20	20	11	10	1.10	1.000	1.100
maxT	20	20	13	12	1.08	1.000	1.083
BGA-COA	20	17	5	4	1.25	0.850	1.063
PLS-CV	20	20	8	9	0.89	1.000	0.889
mAP-KL	14	14	7	9	0.78	1.000	0.778
JDM							
RF-MDA	20	20	14	10	1.40	1.000	1.400
mAP-KL	21	20	13	9	1.44	0.952	1.376
PLS-CV	20	19	14	11	1.27	0.950	1.209
maxT	20	20	13	12	1.08	1.000	1.083
SAM	20	18	10	9	1.11	0.900	1.000
eBayes	20	18	9	9	1.00	0.900	0.900
BGA-COA	20	18	7	10	0.70	0.900	0.630
LGMD2A							
maxT	20	20	13	12	1.08	1.000	1.083
RF-MDA	20	19	7	9	0.78	0.950	0.739
BGA-COA	20	14	5	5	1.00	0.700	0.700
PLS-CV	20	18	8	11	0.73	0.900	0.655
eBayes	20	16	8	10	0.80	0.800	0.640
SAM	20	16	8	10	0.80	0.800	0.640
mAP-KL	6	6	3	7	0.43	1.000	0.429
LGMD2B							
maxT	20	20	15	11	1.36	1.000	1.364
BGA-COA	20	17	7	5	1.40	0.850	1.190
PLS-CV	20	16	9	8	1.13	0.800	0.900
mAP-KL	15	15	7	8	0.88	1.000	0.875
SAM	20	16	8	8	1.00	0.800	0.800
eBayes	20	15	8	9	0.89	0.750	0.667
RF-MDA	20	20	5	8	0.63	1.000	0.625

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

	NM						
RF-MDA	20	18	8	4	2.00	0.900	1.800
mAP-KL	18	18	7	5	1.40	1.000	1.400
SAM	20	20	13	12	1.08	1.000	1.083
maxT	20	18	9	8	1.13	0.900	1.013
BGA-COA	20	17	11	11	1.00	0.850	0.850
PLS-CV	20	17	9	9	1.00	0.850	0.850

Table 6.4: The pathway analysis results on cancer data

FS	Breast						
	Prbs	Gns	P-C-Gns	Pathways	Gns/Pathway	DoU	PE
PLS-CV	20	14	7	5	1.40	0.700	0.980
mAP-KL	6	4	1	1	1.00	0.667	0.667
BGA-COA	20	12	4	4	1.00	0.600	0.600
RF-MDA	20	9	6	5	1.20	0.450	0.540
maxT	20	11	6	7	0.86	0.550	0.471
SAM	20	11	5	6	0.83	0.550	0.458
Colon							
SAM	20	18	14	11	1.27	0.900	1.145
eBayes	20	18	12	10	1.20	0.900	1.080
BGA-COA	20	18	14	14	1.00	0.900	0.900
PLS-CV	20	18	11	11	1.00	0.900	0.900
maxT	20	16	9	9	1.00	0.800	0.800
RF-MDA	20	14	9	10	0.90	0.700	0.630
mAP-KL	20	16	11	14	0.79	0.800	0.629
Leukemia							
eBayes	20	18	14	10	1.40	0.900	1.260
BGA-COA	20	19	12	10	1.20	0.950	1.140
PLS-CV	20	19	9	8	1.13	0.950	1.069
RF-MDA	20	18	10	9	1.11	0.900	1.000
SAM	20	18	13	12	1.08	0.900	0.975
mAP-KL	5	5	4	5	0.80	1.000	0.800
maxT	20	17	12	13	0.92	0.850	0.785
Prostate							
maxT	20	20	13	3	4.33	1.000	4.333
SAM	20	19	8	6	1.33	0.950	1.267
mAP-KL	12	12	7	6	1.17	1.000	1.167
PLS-CV	20	17	11	9	1.22	0.850	1.039
RF-MDA	20	19	13	12	1.08	0.950	1.029
BGA-COA	20	18	10	9	1.11	0.900	1.000
eBayes	20	19	9	10	0.90	0.950	0.855

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

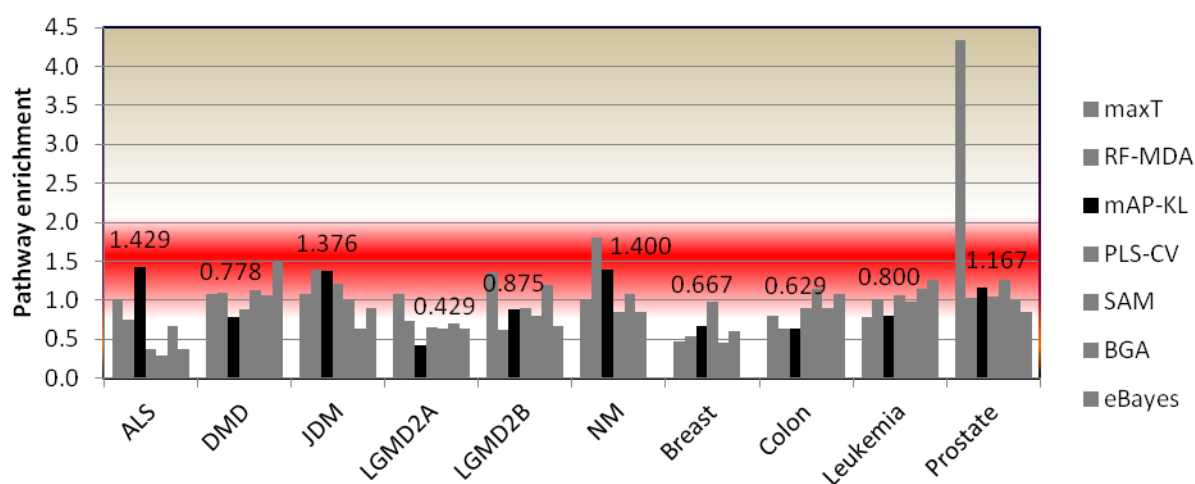


Figure 6.1: The overview of the PE scores

Table 6.5: The pathway lists in the LGMD2B disease

Pathways	mAP-KL	eBayes	PLS-CV	SAM	BGA-COA	RF-MDA	maxT
Apoptosis	x						
Binding and Uptake of Ligands by Scaveng			x				
Cell Cycle	x						x
Cell-Cell communication							x
Circadian Clock						x	x
Developmental Biology						x	x
Disease	x	x	x	x		x	x
Extracellular matrix organization			x			x	
Gene Expression	x	x		x		x	x
Hemostasis		x		x		x	x
Immune System		x	x	x	x	x	x
Metabolism	x	x	x	x	x		x
Metabolism of proteins	x						x
Muscle contraction	x	x	x	x	x		
Neuronal System		x		x			
Signal Transduction	x	x	x		x	x	x
Transmembrane transport of small molcul		x	x	x	x		

6.3. The gene lists from a disease point of view

During this final validation we explored the potential association of the gene lists with the respective pathological phenotypes. For this purpose we utilized a “WEB-based Gene SeT AnaLysis Toolkit” (WebGestalt) [100], to identify those genes from each list that are either directly or closely related to the diseases under analysis. This tool provides an integrated data mining analysis in several areas including “Disease association analysis” with the aid of “Gene List Automatically Derived For You”

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

(GLAD4U) [101] retrieval and prioritization tool, which exploits the PubMed literature. The overall findings of this disease enrichment analysis are presented in Table 6.6, thought we comment and refer only to mAP-KL's disease related genes.

Table 6.6: The disease enrichment per gene list

FS	Relevant Genes	
	ALS	DMD
mAP-KL	FHL1 ALDOA	COL3A1 SPARC
eBayes	TTN MYH7 FHL1 ACTA1 ALDOA	-
PLS-CV	TTN ATP2A1 NEB CKM ALDOA TNNC2	AQP4 MYH8 MYH3 FRZB COL1A1
SAM	TTN MYH7 FHL1 ACTA1 ALDOA	COL1A2 ASPN DMD MYH3
BGA-COA	NEB CKM	AQP4 MYH3
RF-MDA	-	-
maxT	-	-
	JDM	LGMD2A
mAP-KL	MX1 CCL5 RGS1 COL6A3 ISG20 HIF1A GBP1	ANXA5 PRKCQ
eBayes	-	SERPINE1 ANXA2
PLS-CV	PSMB8 ISG20 MYH3	SERPINE1 BMP7 CHI3L1 MYH3
SAM	-	SERPINE1 BMP7 CHI3L1 MYH3
BGA-COA	PSMB8 ISG20	CHI3L1 AQP4
RF-MDA	TRIM21 TYMS IL1R1 MAP2K6	-
maxT	TARDBP CCL5 S100A11	-
	LGMD2B	NM
mAP-KL	RAX MYH3	FGG PTAFR GNB2L1 CEACAM3 PTPRB MYH7
eBayes	IGHMBP2 FUS	-
PLS-CV	AQP4 MYH3 BMP7	TNNI2 ACTN3 ATP2A1 SLPI TGM2 CHI3L1
SAM	IGHMBP2 FUS	-
BGA-COA	MTM1 AQP4 MYH3 PTPN2	-
RF-MDA	-	TGM2 GRIN2A
maxT	GRN FUS	TGM2 GRIN2A
	Breast	Colon
mAP-KL	AGTR1 S100A8	MUC2 IL8 CD46 MAP2K2
eBayes	-	VIP IL8
PLS-CV	SCGB2A2 PTHLH PIP IGFBP5 CA9	MUC2 TSPAN1 ALDH1A1 CEACAM1 IL8
SAM	MMP9 CA9	MUC2 VIP IL8
BGA-COA	SCGB2A2 PRAME CA9	CDH3 ALDH1A1 CDK4 S100P
RF-MDA	GLS ESM1 AGTR1 ESM1	CDH3 TSPAN1 S100A11 IL8
maxT	MMP9 CA9	NPM1 HMGA1 TSPAN1 CEACAM1
	Leukemia	Prostate
mAP-KL	-	CLU GSTM1
eBayes	ELANE TCF3 MYB CD33 CCND3	TARP GSTM1 HPN TSPAN1
PLS-CV	IGH@ TCL1A CD79A ELANE IGK@ MPO	PAGE4 GDF15 TARP HPN
SAM	ELANE LYN CD33 CCND3	GDF15 TARP GSTM1 HPN
BGA-COA	TCL1A CD79A ELANE MPO CD79B CD79A	GDF15 CLU TARP HPN KLK2
RF-MDA	ELANE STMN1	TARP GSTM1
maxT	-	-

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

In ALS, representative examples include the FHL1 [102] gene that regulates skeletal muscle mass and ALDOA [103] gene that found to discern successfully systemic sclerosis patients through its increased concentration in plasma. Moreover, COL3A1 and SPARC genes are related to extracellular matrix formation and fibroblast growth, biological processes consistent with the increased fibrosis that is observed in skeletal muscles affected by DMD [104]. In NM and LGMD2B, the structure associated MYH7, MYH3 genes were depicted, in agreement with the reports of cytoskeletal disorganization in the affected muscle fibers of these patients [79, 105], whereas in LGMD2A the PRKCQ gene is considered as a valuable pharmacological target for both immune cells and skeletal muscles [106]. As opposed to the other skeletal muscle diseases included in this study, JDM is an inflammatory myopathy of presumed autoimmune dysfunction. Consistently with the disease pathology, the short-listed genes CCL5 and ISG20 are related to interferon or to chemokine and cytokine production, all key molecules of the immune system [107].

In relation to breast cancer, the AGTR1 have been found to be over-expressed across multiple independent breast cancer cohorts [108], similarly to the S100A8 gene which is also considered as a molecular marker [109]. With respect to colon cancer the IL8 products has been ascribed to angiogenesis promotion [110], the MAP2K2 appears to suppresses the proliferation of colon carcinoma cell lines when silenced [111] and the MUC2 in conjunction with Galectin-3 play a significant role in colon cancer metastasis [112]. Finally, in prostate cancer the CLU is considered as a valid therapeutic target when combined with androgen ablation [113] and the GSTM1 polymorphisms are closely related to mortality and are potential prognosis markers [114].

These findings jointly, demonstrate that despite their small size, the discriminatory 'lists of selected genes' depicted by the proposed FS approach contain biologically relevant genes, representative of the respective disease related molecular pathways.

7. R-package implementation

7.1. Introduction

To provide the research community with the capability to apply *mAP-KL* in any given gene expression dataset, we have implemented this methodology to an R package accompanied with extra functionalities including data sampling preprocessing, classification, network analysis, gene annotation analysis and reporting [115]. Concerning the data sampling functionality, a dataset of samples may be split into train and test sets following a user-defined proportion. In relation to data preprocessing we provide several normalization and transformation alternatives along with density plots that provide the user with the necessary hints about the effect of the methods on the input data. Regarding the classification performance of the selected genes, the user may perform any cross-validation on the training data or even a hold out validation on a separate test set with the aid of SVM and provides estimates of their discrimination ability. As regards the network analysis, the user may compute several network characteristics of the “exemplars” including degree of centrality, closeness, betweenness and clustering coefficient as well as to construct the edge list table (Node1 – Node2 – weight) based on the N top ranked genes. Finally, an *html* report summarizing the results of all types of analysis is produced to assist user towards a structured and archived analysis logbook.

7.2. Classes and functions of the *mAPKL* package

The *mAPKL* implemented in R as an S4 package that takes advantage of the rich functionality of the “ExpressionSet” (*eSet*) class [116]. This type of class is designed to accommodate a variety of information including expression data from microarray experiments (*assayData*), “meta-data” describing samples in the experiment (*phenoData*), annotations and meta-data about the features on the chip (*featureData*, *annotation*), information about the protocol used for processing the samples (*protocolData*), and a flexible structure to describe the experiment (*experimentData*). All those different sources of information are handled by class-methods thus the proper manipulation is guaranteed. Besides, using this class objects throughout this package we make feasible any collaboration with other bioconductor packages hence, extending the meta-analysis options.

The *mAPKL* includes four distinct functional modules and five classes, Figure 7.1. The core function of this package is the *mAPKL* that implements the hybrid feature selection methodology. It takes as input an *eSet* class object with the training data and several predefined parameters necessary for the intrinsic statistical analysis and clustering methods. It may also accept a validation *eSet* object to directly apply on it the results of the *mAP-KL* analysis. This function returns an object of ‘*mAPKLRes*’ S4 class where its slots embody the matrix of the top N ranked genes, the clusters and their respective exemplars, the training and validation *eSet* objects of the exemplars, along with statistical information such as *p*-value, adjusted *p*-value and fold-change for all genes.

The following functional module provides classification estimates for the selected genes. In particular, it utilizes the exemplars’ *eSet* objects from the ‘*mAPKLRes*’ class to run an SVM based cross-validation classification test to quantify the discrimination power of the gene exemplars. The necessary parameters for running the SVM classifier are computed dynamically with the *tune.svm* function of the ‘*e1071*’ R-package [117]. The classification measures are calculated through a *mAPKL*’s function called *metrics* and include the Area Under the Curve (AUC), the Matthews correlation coefficient (MCC),

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

the accuracy (ACC), the true negative rate (TNR) or specificity and the true positive rate (TPR) or sensitivity.

The next functional module exploits the microarray chip annotation file, if available, to collect necessary genome oriented information so as to facilitate other types of genome analysis such as pathway analysis. The 'Annot' S4 class provides slots for gene 'symbol', 'entrezId', 'ensemblId' and chromosomal location info of the exemplars. Thus, the user not only has at hand a valid conversion mechanism between probes and genes but also several additional meta-data for other types of analysis like pathway or Gene Ontology.

The *netwAttr* function deals with the network characteristics of the top N ranked genes but more importantly with the gene exemplars. Three different types of centralities (degree, closeness, betweenness) and a measure for clustering coefficient called transitivity are estimated with this function. The degree centrality of a node refers to the number of connections or edges of that node to other nodes. The closeness centrality describes the reciprocal accumulated shortest length distance from a node s to all other connected nodes. The betweenness centrality depicts the number of times a node intervenes along the shortest path of two other nodes. Transitivity measures the degree of nodes to create clusters within a network. For all four network measures we provide both global and local values. Moreover, the *netwAttr* provides a weighted edge list (Node1-Node2-weight) based on the top N ranked genes, as a front end to network and graph packages for advance analysis and visualization.

Finally, the package incorporates functions that assist data importing from 'txt' files to eSet class objects, preprocessing of the gene expression values and reporting. Concerning the preprocessing functional unit, it supports log 2 transformation and four different normalization methods including mean-centering, z-score, quantile and cyclic loess. In particular, this function produces an S3 class object, a list, with maximum nine available options, see section 7.3. Moreover, an efficient sampling method is available that assist user to split any dataset into a train and a test sets to a user defined percentage, while keeping a stratified analogy between the two classes of the samples. With regard to the *report* function, the user may produce a summarized report in html format that presents the results in all different stages of analysis. In the following section we will present a case study to display thoroughly the functionality of our package using the 'mAPKldata' experiment data package.

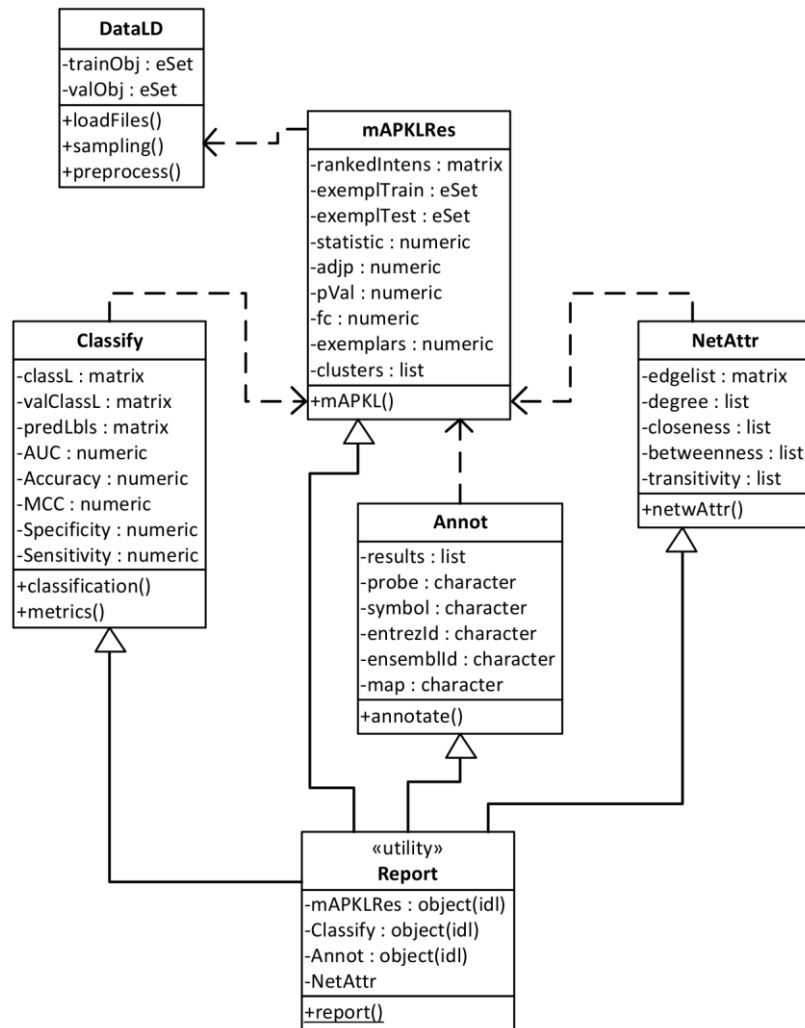


Figure 7.1: A UML schematic representation of the classes and functions of the mAPKL. The solid rectangles with the three compartments represent classes. In the first compartment is the name of the class, in the second compartment is the attributes of the class, and in the third is the methods/functions relevant to the class. The 'Report' rectangular is a special type of class called utility that has static attributes and methods and no instances. The dotted lines represent 'dependencies' between classes. The lines with the arrowhead represent 'generalizations' and show the parts (static attributes) of the 'Report' class.

7.3. An analysis scenario with mAPKL package

For the purposes of the following case study we engaged the “mAPKLData” bioconductor experiment data package that we built as a supplement to the “mAPKL” package. It provides the GSE5764 dataset, which is available at the NCBI Gene Expression Omnibus and includes gene expression data from a breast cancer study published by Turashvili et al.[118] that contains 30 samples related to breast cancer (20 normal and 10 tumor samples), based on Affymetrix HG-U133_Plus_2 microarray platform.

Initially, we load the two packages and then the breast cancer data. Then with the aid of the “sampling” function we create a separate training and validation sets where 60% of the samples will be used for training and the rest 40% of the samples will be used for evaluation purposes. The selection of samples follows a random selection based on the defined seed number hence, no bias is inserted. The resulted train set has twelve

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

normal and six tumor samples, and the validation set eight normal and four disease samples.

```
library(mAPKL)
library(mAPKLData)
data(mAPKLData)
varLabels(mAPKLData)
breast <- sampling(Data=mAPKLData, valPercent=40, classLabels="tvpe", seed=135)
```

Figure 7.2: Loading the packages and sampling the data

Then we employ the “*preprocess*” function that produces an S3 class object, a list, with maximum nine available options, Figure 7.3. In particular, the attributes of that list may contain the following values:

1. the initial gene expression values (*rawdata*),
2. the values after “*mean-centering*” normalization (*mc.normdata*),
3. the values after “*z-score*” normalization (*z.normdata*),
4. the values after “*quantile*” normalization (*q.normdata*),
5. the values after “*cyclic loess*” normalization (*cl.normdata*),
6. the values after log2 transformation and “*mean-centering*” normalization (*mcL2.normdata*),
7. the values after log2 transformation and “*z-score*” normalization (*zL2.normdata*),
8. the values after log2 transformation and “*quantile*” normalization (*qL2.normdata*),
9. the values after log2 transformation and “*cyclic loess*” normalization (*clL2.normdata*).

Besides density plots per method are produced and saved, Figure 7.4, to assist the user upon which normalization approach to employ for the following *mAP-KL* analysis. Though, this decision is not exclusive and the user may run a *mAP-KL* analysis multiple times trying any of the available approaches and concluding to possible different subsets of exemplars. Those subsets will form different classifiers and will be assessed for their discrimination power with the aid of the “*classification*” function, Figure 7.5. This function performs classification through the SVM algorithm and produces a classification result either on the training set or on a validation set. The default SVM settings are: “*linear*” kernel and 5-folds cross-validation although other options are feasible.

```
normTrainData <- preprocess(breast$trainData)
normTestData <- preprocess(breast$testData)
attributes(normTrainData)

## $names
## [1] "rawdata"      "mc.normdata"  "z.normdata"  "q.normdata"
## [5] "cl.normdata"  "mcL2.normdata" "zL2.normdata" "qL2.normdata"
## [9] "clL2.normdata"
```

Figure 7.3: The density plots per normalization method

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

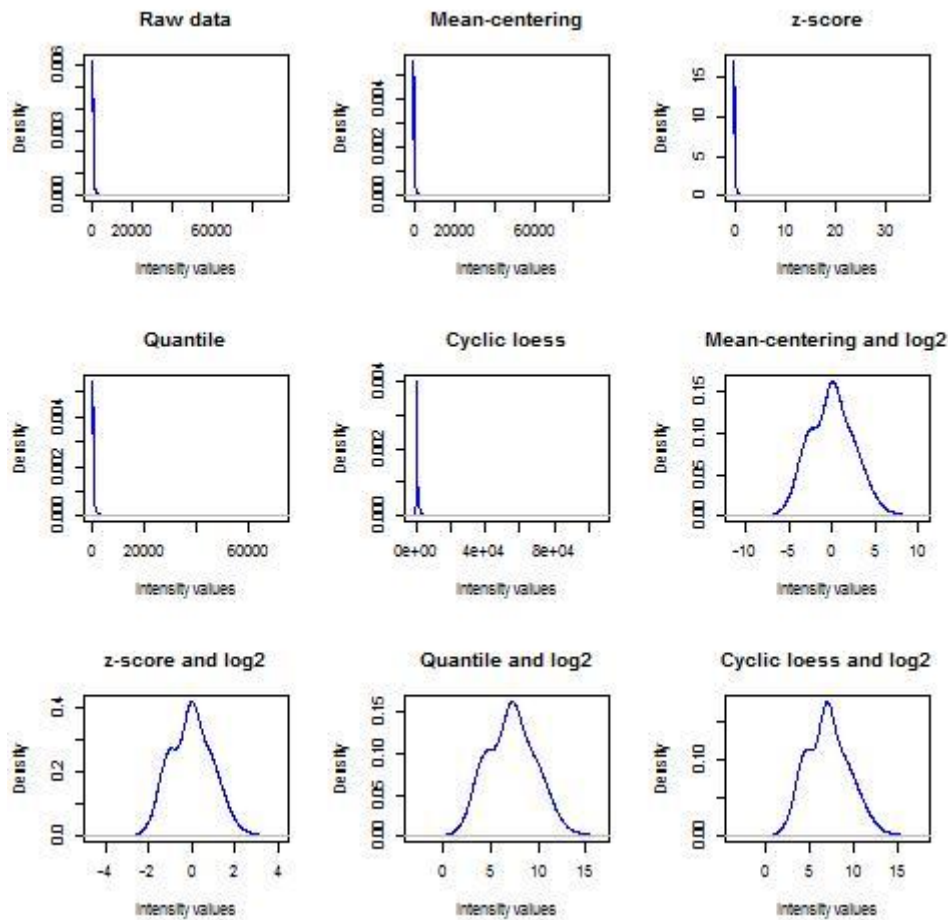


Figure 7.4: The density plots per normalization method

Indeed, we carried out eight different mAP-KL analyses and concluded to eight different subsets of exemplars. Those subsets are bound to form different classifiers where all of them will be assessed for their discrimination power with the aid of the classification function, Table 7.1. This function performs classification through the SVM algorithm and produces a classification result either on the training set or on a validation set. During this analysis we assessed the performance on the validation set using the following SVM parameters: 'linear' kernel and 5-folds cross-validation (although other options are feasible). According to the classification results, the exemplars' list produced after log2 transformation and cyclic loess normalization achieved the best discrimination results and consequently will be further explored from a pathway and a network-topology perspectives.

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

```

# log2 transformation and cyclic loess normalization
exprs(breast$trainData) <- normTrainData$cL2.normdata
exprs(breast$testData) <- normTestData$cL2.normdata
out.cL2 <- mAPKL(trObj = breast$trainData, classLabels = "type",
  valObj = breast$testData, dataType = 7)
# Hold-out classification
clasPred <- classification(out.cL2@exemplTrain, "type", out.cL2@exemplTest)
# log2 transformation and mean-centering normalization
exprs(breast$trainData) <- normTrainData$mcL2.normdata
exprs(breast$testData) <- normTestData$mcL2.normdata
out.mcL2 <- mAPKL(trObj = breast$trainData, classLabels = "type",
  valObj = breast$testData, dataType = 7)
# Hold-out classification
clasPred <- classification(out.mcL2@exemplTrain, "type", out.mcL2@exemplTest)
# log2 transformation and quantile normalization
exprs(breast$trainData) <- normTrainData$qL2.normdata
exprs(breast$testData) <- normTestData$qL2.normdata
out.qL2 <- mAPKL(trObj = breast$trainData, classLabels = "type",
  valObj = breast$testData, dataType = 7)
# Hold-out classification
clasPred <- classification(out.qL2@exemplTrain, "type", out.qL2@exemplTest)
# log2 transformation and z-score normalization
exprs(breast$trainData) <- normTrainData$zL2.normdata
exprs(breast$testData) <- normTestData$zL2.normdata
out.zL2 <- mAPKL(trObj = breast$trainData, classLabels = "type",
  valObj = breast$testData, dataType = 7)
# Hold-out classification
clasPred <- classification(out.zL2@exemplTrain, "type", out.zL2@exemplTest)

```

Figure 7.5: The density plots per normalization method

Table 7.1: Classification performance of gene exemplars per preprocessing method

Method	Exemplars	AUC	MCC	ACC	TNR	TPR
cL2	15	0.94	0.84	92.00	0.88	1.00
mcL2	40	0.88	0.82	92.00	1.00	0.75
qL2	40	0.88	0.82	92.00	1.00	0.75
z	17	0.81	0.62	83.0	0.88	0.75
mc	28	0.81	0.62	83.0	0.88	0.75
cl	17	0.75	0.63	83.0	1.00	0.50
q	14	0.69	0.41	75.0	0.88	0.50
zL2	39	0.62	0.43	75.00	1.00	0.25

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

Prior to pathway analysis we have to obtain relevant annotation info to the exemplars. For this purpose we first run the “*annotate*” function with the argument “*chip*” equal to “*hgu133plus2.db*” since this is the relevant microarray chip platform for that dataset. In the sequel, we exploit the “ENTREZID” property to perform a pathway analysis utilizing the “Reactome” pathway database [99].

```
gene.info <- annotate(out.cll2@exemplars, "hgu133plus2.db")
gene.info@results

##          PROBEID      SYMBOL ENTREZID      ENSEMBL      MAP
## 1  215717_s_at      FBN2      2201  ENSG00000138829  5q23-q31
## 2  1561358_at      TXLNA     200081  ENSG00000084652  1p35.1
## 3  222752_s_at     TMM206     55248  ENSG00000065600  1q32.3
## 4   233922_at      <NA>      <NA>      <NA>      <NA>
## 5  218871_x_at  CSGALNACT2    55454  ENSG00000169826 10q11.21
## 6   33323_r_at      SFN      2810  ENSG00000175793  1p36.11
## 7   244311_at      <NA>      <NA>      <NA>      <NA>
## 8   220932_at      <NA>      <NA>      <NA>      <NA>
## 9   205508_at     SCN1B     6324  ENSG00000105711 19q13.1
## 10  209596_at     MXRA5    25878  ENSG00000101825  Xp22.33
## 11  215180_at      <NA>      <NA>      <NA>      <NA>
## 12 1560638_a_at      <NA>      <NA>      <NA>      <NA>
## 13  201852_x_at    COL3A1     1281  ENSG00000168542  2q31
## 14   229947_at     PI15    51050  ENSG00000137558  8q21.11
## 15  221731_x_at     VCAN     1462  ENSG00000038427  5q14.3

library("reactome.db")
# We first remove the 'NA' entries
genes <- gene.info@entrezId[!is.na(gene.info@entrezId)]
# Then we map the Entrez ID to Reactome pathway identifiers
qExtID2PathID <- mget(genes, reactomeEXTID2PATHID, ifnotfound = NA)
notNA.idx <- unlist(lapply(qExtID2PathID, function(i) !all(is.na(i))))
qExtID2PathID <- qExtID2PathID[notNA.idx]
pathID <- as.character(qExtID2PathID[[1]])
# Finally we map Reactome pathway identifiers to pathway
# names
pathName <- unlist(mget(pathID, reactomePATHID2NAME))
pathName

##          1474244
## "Homo sapiens: Extracellular matrix organization"
##          1566948
## "Homo sapiens: Elastic fibre formation"
##          2129379
## "Homo sapiens: Molecules associated with elastic fibres"
##          1474228
## "Homo sapiens: Degradation of the extracellular matrix"
```

Figure 7.6: Pathway analysis results

A further functionality of this package is the computation of the exemplars' network characteristics, Figure 7.7. Particularly, the “*netwAttr*” function computes three different types of centralities (degree, closeness, and betweenness) and a measure for clustering coefficient called transitivity. The degree centrality of a node refers to the number of connections or edges of that node to other nodes. The closeness centrality describes the reciprocal accumulated shortest length distance from a node to all other connected nodes. The betweenness centrality depicts the number of times a node intervenes along the shortest path of two other nodes. The transitivity measures the degree of nodes to create clusters within a network. For all four network attributes we provide both global and local values. Based on the exemplars' network characteristics we may also identify potential hubs, Figure 7.8.

```
net.attr <- netwAttr(out.c1L2)
wDegreeL <- net.attr@degree$WdegreeL[out.c1L2@exemplars]
wClosenessL <- net.attr@closeness$WclosenessL[out.c1L2@exemplars]
wBetweennessL <- net.attr@betweenness$WbetweennessL[out.c1L2@exemplars]
wTransitivityL <- net.attr@transitivity$WtransitivityL[out.c1L2@exemplars]
Global.val <- c(net.attr@degree$WdegreeG, net.attr@closeness$WclosenessG,
  net.attr@betweenness$WbetweennessG, net.attr@transitivity$WtransitivityG)
Global.val <- round(Global.val, 2)
exempl.netattr <- rbind(wDegreeL, wClosenessL, wBetweennessL,
  wTransitivityL)
netAttr <- cbind(Global.val, exempl.netattr)
netAttr <- t(netAttr)
netAttr
```

##	wDegreeL	wClosenessL	wBetweennessL	wTransitivityL
## Global.val	330.18	0.93	741.81	0.57
## 215717_s_at	308.35	1.25	886.00	0.14
## 1561358_at	346.92	1.34	1141.00	0.14
## 222752_s_at	327.89	0.65	0.00	0.14
## 233922_at	317.58	0.79	2.00	0.15
## 218871_x_at	293.73	0.53	768.00	0.14
## 33323_r_at	338.19	0.27	0.00	0.13
## 244311_at	294.80	0.63	0.00	0.15
## 220932_at	359.10	0.66	0.00	0.14
## 205508_at	309.07	0.89	4.00	0.14
## 209596_at	345.13	1.34	278.00	0.14
## 215180_at	333.37	1.37	1440.00	0.14
## 1560638_a_at	368.23	1.38	4615.00	0.14
## 201852_x_at	353.34	0.93	24.67	0.15
## 229947_at	317.11	1.19	496.00	0.15
## 221731_x_at	331.01	0.61	14.00	0.15

Figure 7.7: The exemplars' network characteristics

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

Finally, the overall analysis is summarized in an html report produced by the “report” function, Figure 7.10. It covers the dataset representation depicting the samples' names and their respective class labels, the exemplars section where statistical results and network characteristics are included. The classification performance section illustrates the performance metrics achieved in either cross-validation or hold-out validation. The last section of this report presents annotation info relevant to the chip technology.

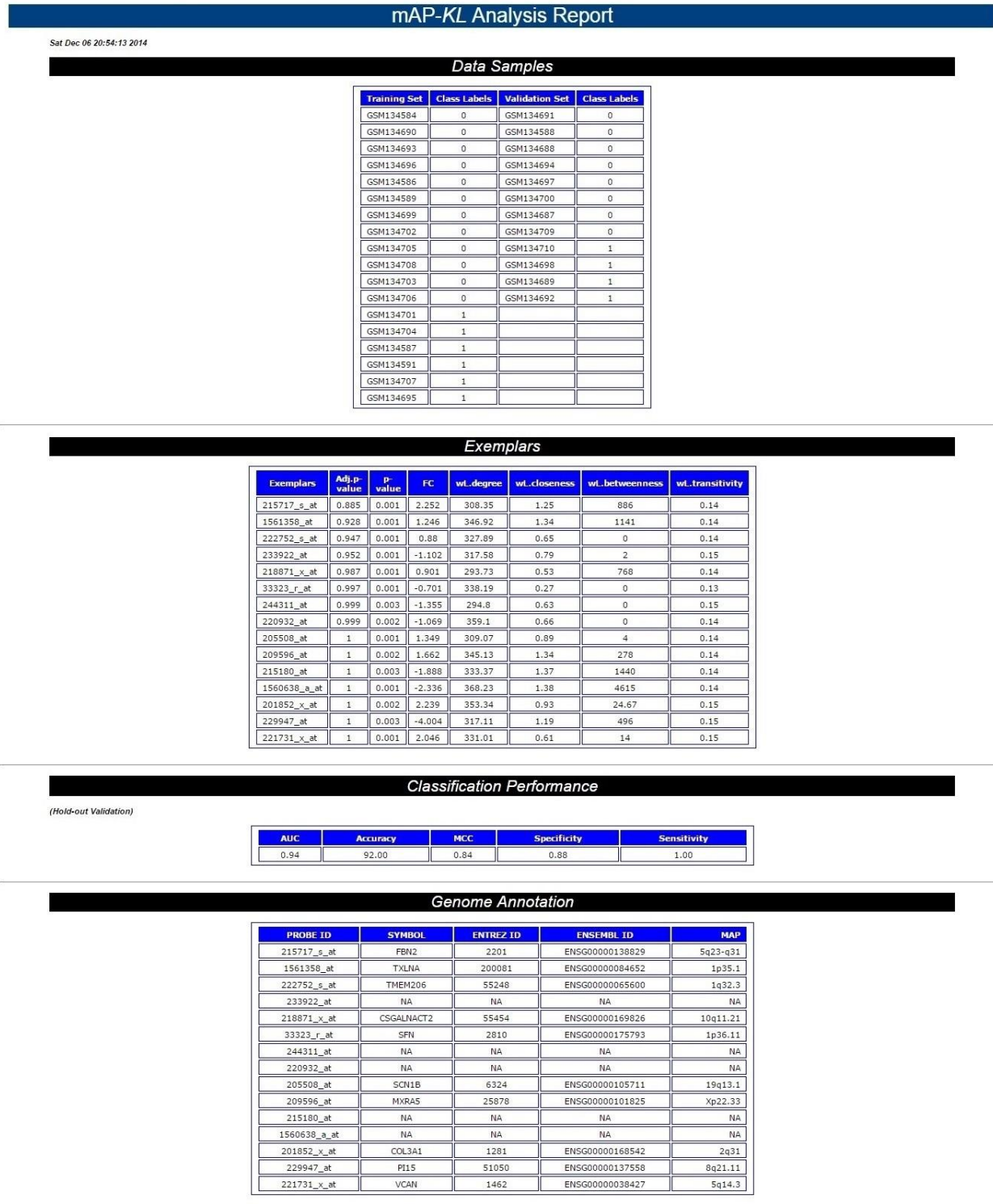


Figure 7.10: The summarized mAP-KL analysis report

7.4. Availability and Future Directions

As part of the Bioconductor project, the mAPK package is freely available under the GPL-2 or later license accompanied with detailed help pages per class and function. Besides, an elaborate vignette introduces all available functionalities through a case study scenario that is based on the 'mAPKData' bioconductor experiment data package. Thus, the user can see both illustrated codes and executed outputs and get easily accustomed to the package. Moreover, the Bioconductor project guarantees the easy implementation and platform independence, the versioning of the forthcoming package releases and the obliged that the package will be maintained by the author, which includes response to bug reports or queries from other users as well as checking periodically the functionality of the package. The potential expansions of the mAPKL package will include the adoption of methods for network reconstruction other than the 'clr' method [119] which is employed in the current version of the netWAttr function. Besides, methods related to functional enrichment and advanced graphics designate our subsequent directions.

8. DISCUSSION, CONCLUSIONS AND FUTURE WORK

8.1. Discussion

Feature selection in microarray data based on the differential expression of genes between two phenotypes, is a research topic which has drawn scientific interest from the late 90s. Numerous algorithmic approaches have been proposed so far trying to identify those significant genes that can be characterized as marker genes. Marker genes are supposed to encompass both, discrimination ability and biological relevance. The discrimination ability characteristic, envisages the accurate discrimination of samples between two phenotypes (e.g. normal vs. disease) with a limited number of genes. Although this criterion appears to have been accomplished, according to the published classification results, in essence, this is not the case. The reason is a complementary characteristic called generalization. Certainly, discriminating samples of a specific disease for a particular dataset is not adequate. The ultimate goal is to conclude to a set of genes that achieve accurate classification at any dataset relevant to the disease and phenotypes.

As far as the biological relevance is concerned, the marker genes shall be related to the disease. In reality the selected genes are not a priori associated with the specific disease. The inherent noise in gene expression data, the diversity of microarray platforms and normalization methods that at the end influence the measured intensities and their variability across the dataset, are some of the reasons that some of the differential expressed genes are from other causes rather than biological. On the other hand, biologists ask for gene lists of a reasonable number of genes, approximately less than 50 genes depending on the disease, which also include all or the majority of the in vivo identified relevant genes.

So far no method has addressed those two goals to a widespread number of cases and diseases to be considered as the gold standard. As a consequence a plethora of methods have been developed trying to achieve the best possible compensation between classification accuracy and biological relevance. Despite any differences among those methods there are also some common key characteristics that enable their classification into distinctive categories. The filter, wrapper and embedded are three principal classes of feature selection methods with the respective advantages and disadvantages. Methods belonging to one of those categories may be combined with other computational intelligence methods, e.g. clustering, to provide new methods with improved characteristics.

Those new offspring methods are generally called as “hybrid” methods and aspire to capitalize the benefits of the parent methods to achieve significantly improved gene selection. Therefore, the development of new approaches is actually an ongoing process in the advent of either new biological notions, for instance about genes' correlation, or computational hardware advances, for example multi-threading or parallel computing, where all of which may bring a new era in the process of gene selection.

8.2. Conclusions and future directions

In this dissertation we proposed a hybrid method (*mAP-KL*), which clearly demonstrates how effective the combination of a multiple hypothesis testing approach with a clustering algorithm can be to select small yet informative subsets of genes in binary classification problems. Particularly, across a variety of diseases and datasets, *mAP-KL* achieved competitive classification results (Figure 5.1), compared to other FS methods and specifically to HykGene method, which follows a similar philosophy i.e. first ranking and then clustering. The advances of *mAP-KL* over HykGene or other similar

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

approaches discussed during chapter 1 stem from three key characteristics; the data-driven nature, the affinity propagation clustering, and the classifier independence. Indeed, the engagement of a cluster quality index, the Krzanowski and Lai, diminishes any fuzziness and provides the clustering algorithm with a representative number of potential clusters, as clearly presented in the first simulation data setup. According to the clustering results in the simulated data of six different datasets with variable number of clusters, the mAP-KL managed to identify successfully the underlying cluster distribution. Though, we have to emphasize that accurate cluster quality indexing is in close relation to the size of data applied. Especially, by differentiating the number of differential genes and the number of top N ranked genes, we concluded that the closer to the DEGs is the number of the top N ranked genes the more accurate the identification of the clusters is.

Following the identification of the number of clusters, the employment of AP clustering algorithm, deals effectively with the issue of representative genes per cluster. Other comparable approaches to mAP-KL admitted considerably difficulties on selecting effectively one or more representative genes per cluster. For instance, in the Hanczar et al. study a mathematical notion, the *prototype gene*, was formulated towards the representative genes issue but considered as an attempt that needed further improvement. On the other hand, according to AP the exemplars are the central genes within a cluster of genes presenting a network oriented approach. This network driven perspective of the association of genes during molecular processes has gain ground lately through the systems biology field and it is a springboard for further improvements in the mAP-KL's methodology.

In relation to the exemplars, we assessed them from a classification as well from a biological point of view. The main reason is that representative genes may considered as marker genes if and only if are also related to the disease. Therefore, the classification results are inadequate to characterize a set of genes as marker genes unless they discriminate unknown samples of the relevant disease with a similar accuracy, generalization property, and contain genes that are associated with the disease. Actually, it is believed that the association is the reason for generalization and not contrariwise. In other words, if a set generalizes during several datasets it is bound that some of its genes are closely related to the disease.

Hence, in chapter 6 we conducted a biological relevance analysis on the selected gene sets among the best FS methods, according to their classification results in the study, including the mAP-KL. The disease association analysis, clearly demonstrates that the existence of relevant genes influence the classification process. Indeed, the mAP-KL achieved a 0.71 AUC score in leukemia, where none of the selected genes found to be related to the disease. In LGMD2B the AUC score was 0.70 with two relevant genes whereas in LGMD2A was 1.00 with two genes, too. This observation might be important for the improvement of the method and further analysis is necessary to unveil the intrinsic reasons for this outcome. However, in many other diseases the mAP-KL had enough representatives including in its subset, considering that mAP-KL concluded usually to a shorter list than the other FS with the fixed 20-genes length, reflecting the positive classification performance.

So far, in mAP-KL the data determine the size of the subset i.e. the number of clusters/features and the clustering algorithm decides on which informative genes are to be included. Contrary to other methods, for example HykGene, where a classifier is wrapped around its method, in our case no classifier takes part during the subset construction. This methodological characteristic is of great importance since our subsets lack of any overfitting phenomenon pertinent to classifiers. We verified this belief by

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

applying a diverse of classification algorithms on our subsets. Particularly, we employed the SVM with linear kernel classifier, the KNN and the Random Forest classifiers, all of which follow a different algorithmic perspective. In most of the diseases the classification performance is almost identical, Figure 5.2, although we used the same parameters set up in each classifiers across all diseases.

Taking into account all the aforementioned issues, we may claim that the novelty and strength of *mAP-KL* is the efficient sampling of the ranked gene list, selecting those genes that are necessary for improved classification, rather than keeping just a predefined number of top N ranked genes. A further advantage of the employment of *mAP-KL* is that the clustering correlation on the gene expression values may reflect biological relevance of the selected genes with the respective disease, thus providing a reasonable basis for discovering prognostic biomarkers.

In addition, the clustering nature within the *mAP-KL* methodology raises expectations for potential expansions to gene-network-inference. Indeed, on the one hand the initial ranking and on the other hand the subsequent clustering, confront to the general view of functional units i.e. groups of genes with similar functions based on their expression values [120]. Therefore, mining the “exemplars” it can be considered as the forefront of a network inference process rather than just the outcome of a feature selection approach. As such, we intent to construct networks based on the top N genes of our methodology and then to exploit the network characteristics of the “exemplars” to produce graphical representations of the cellular network topology, where genes are represented as vertices that are connected by edges representing potential direct regulatory interactions. An initial attempt towards this expansion has already been applied in the *mAPKL* r-package (see section 7.3). Though, more network inference methods for the reconstruction of gene regulatory networks and tests for functional enrichment designate our subsequent directions.

REFERENCES

- [1] M. S. Chua and M. M. Sarwal, "Microarrays: new tools for transplantation research," *Pediatr Nephrol*, vol. 18, pp. 319-27, Apr 2003.
- [2] S. Saviozzi, G. Iazzetti, E. Caserta, A. Guffanti, and R. A. Calogero, "Microarray data analysis and mining," *Methods Mol Med*, vol. 94, pp. 67-90, 2004.
- [3] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530-6, Jan 31 2002.
- [4] J. A. Sparano and S. Paik, "Development of the 21-gene assay and its application in clinical practice and clinical trials," *J Clin Oncol*, vol. 26, pp. 721-8, Feb 10 2008.
- [5] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507-17, Oct 1 2007.
- [6] J. Jaeger, R. Sengupta, and W. L. Ruzzo, "Improved gene selection for classification of microarrays," *Pac Symp Biocomput*, pp. 53-64, 2003.
- [7] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. Thesis, Computer Science, The University of Waikato, Hamilton, New Zealand, 1999.
- [8] B. Hanczar, M. Courtine, A. Benis, C. Hennegar, K. Clement, and J.-D. Zucker, "Improving classification of microarray data using prototype-based feature selection," *ACM SIGKDD Explorations Newsletter*, vol. 5, p. 7, December 2003 2003.
- [9] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J Bioinform Comput Biol*, vol. 3, pp. 185-205, Apr 2005.
- [10] Y. Wang, F. S. Makedon, J. C. Ford, and J. Pearlman, "HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data," *Bioinformatics*, vol. 21, pp. 1530-7, Apr 15 2005.
- [11] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972-6, Feb 16 2007.
- [12] [Online]. Available: <http://www.genomesonline.org/>
- [13] D. J. Lockhart and E. A. Winzeler, "Genomics, gene expression and DNA arrays," *Nature*, vol. 405, pp. 827-36, Jun 15 2000.
- [14] M. J. Heller, "DNA microarray technology: devices, systems, and applications," *Annu Rev Biomed Eng*, vol. 4, pp. 129-53, 2002.
- [15] D. Gershon, "Microarray technology: an array of opportunities," *Nature*, vol. 416, pp. 885-91, Apr 25 2002.
- [16] C. A. Harrington, C. Rosenow, and J. Retief, "Monitoring gene expression using DNA microarrays," *Curr Opin Microbiol*, vol. 3, pp. 285-91, Jun 2000.
- [17] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart, "High density synthetic oligonucleotide arrays," *Nat Genet*, vol. 21, pp. 20-4, Jan 1999.
- [18] Available: http://www.autismspeaks.org/docs/Affy_gene_chip.pdf
- [19] M. N. Mandal, J. R. Heckenlively, T. Burch, L. Chen, V. Vasireddy, R. K. Koenekoop, *et al.*, "Sequencing arrays for screening multiple genes associated with early-onset human retinal degenerations on a high-throughput platform," *Invest Ophthalmol Vis Sci*, vol. 46, pp. 3355-62, Sep 2005.
- [20] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nat Rev Genet*, vol. 7, pp. 55-65, Jan 2006.
- [21] J. S. Verducci, V. F. Melfi, S. Lin, Z. Wang, S. Roy, and C. K. Sen, "Microarray analysis of gene expression: considerations in data mining and statistical treatment," *Physiol Genomics*, vol. 25, pp. 355-63, May 16 2006.
- [22] Å. Magnus, "Normalization and differential gene expression analysis of microarray data," Ph.D. Thesis, Department of Mathematics, Chalmers University of Technology and Göteborg University, 2008.

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

- [23] R. Nadon and J. Shoemaker, "Statistical issues with microarrays: processing and analysis," *Trends Genet*, vol. 18, pp. 265-71, May 2002.
- [24] S. Dudoit, J. Popper, and J. Boldrick, "Multiple hypothesis testing in microarray experiments," *Statistical Science*, vol. 18, pp. 71-103, 2003.
- [25] S. P. Wright, "Adjusted P-Values for Simultaneous Inference," *Biometrics*, vol. 48, pp. 1005-1013, 1992.
- [26] A. J. Holloway, R. K. van Laar, R. W. Tothill, and D. D. Bowtell, "Options available--from start to finish--for obtaining data from DNA microarrays II," *Nat Genet*, vol. 32 Suppl, pp. 481-9, Dec 2002.
- [27] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Res*, vol. 18, pp. 1509-17, Sep 2008.
- [28] W. Xu, J. Seok, M. N. Mindrinos, A. C. Schweitzer, H. Jiang, J. Wilhelmy, *et al.*, "Human transcriptome array for high-throughput clinical studies," *Proc Natl Acad Sci U S A*, vol. 108, pp. 3707-12, Mar 1 2011.
- [29] A. Grada and K. Weinbrecht, "Next-generation sequencing: methodology and application," *J Invest Dermatol*, vol. 133, p. e11, Aug 2013.
- [30] B. Meder, J. Haas, A. Keller, C. Heid, S. Just, A. Borries, *et al.*, "Targeted next-generation sequencing for the molecular genetic diagnostics of cardiomyopathies," *Circ Cardiovasc Genet*, vol. 4, pp. 110-22, Apr 2011.
- [31] R. Hu, X. Qiu, G. Glazko, L. Klebanov, and A. Yakovlev, "Detecting intergene correlation changes in microarray analysis: a new approach to gene selection," *BMC Bioinformatics*, vol. 10, p. 20, 2009.
- [32] I. B. Jeffery, D. G. Higgins, and A. C. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," *BMC Bioinformatics*, vol. 7, p. 359, 2006.
- [33] V. Trevino and F. Falciani, "GALGO: an R package for multivariate variable selection using genetic algorithms," *Bioinformatics*, vol. 22, pp. 1154-6, May 1 2006.
- [34] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.
- [35] I. Inza, P. Larranaga, R. Blanco, and A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains," *Artif Intell Med*, vol. 31, pp. 91-103, Jun 2004.
- [36] M. Hauskrecht, R. Pelikan, D. E. Malehorn, W. L. Bigbee, M. T. Lotze, H. J. Zeh, *et al.*, "Feature Selection for Classification of SELDI-TOF-MS Proteomic Profiles," *Appl Bioinformatics*, vol. 4, pp. 227-46, 2005.
- [37] S. Ma and Y. Dai, "Principal component analysis based methods in bioinformatics studies," *Brief Bioinform*, vol. 12, pp. 714-22, Nov 2011.
- [38] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, pp. 265-286, 2006.
- [39] R. K. Agrawal and R. Bala, "A hybrid approach for selection of relevant features for microarray datasets," *International Journal of Computer and Information Engineering*, vol. 1, pp. 196-202, 2007.
- [40] L. Chuang, C. Ke, and C. Yang, "A Hybrid Both Filter and Wrapper Feature Selection Method for Microarray Classification," in *In Proc. of the International MultiConference of Engineers and Computer Scientists (IMECS)*, Hong Kong, 2008, pp. 19-21.
- [41] P. Yang and Z. Zhang, "An embedded two-layer feature selection approach for microarray data analysis," *IEEE Intelligent Informatics Bulletin*, vol. 10, pp. 24-32, 2009.
- [42] M. R. Hassan, M. M. Hossain, J. Bailey, G. Macintyre, J. W. Ho, and K. Ramamohanarao, "A voting approach to identify a small number of highly predictive genes using multiple classifiers," *BMC Bioinformatics*, vol. 10 Suppl 1, p. S19, 2009.
- [43] Y. Ge, S. Dudoit, and T. P. Speed, "Resampling-based multiple testing for microarray data analysis," *Test*, vol. 12, pp. 1-77, 2003.

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

- [44] J. D. Storey, "The optimal discovery procedure: a new approach to simultaneous significance testing," *Journal of the Royal Statistical Society: Series B* vol. 69, pp. 347-368, 2007.
- [45] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Stat Appl Genet Mol Biol*, vol. 3, p. Article3, 2004.
- [46] G. K. Smyth, *Limma: linear models for microarray data*. New York: Springer, 2005.
- [47] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci U S A*, vol. 98, pp. 5116-21, Apr 24 2001.
- [48] J. Gould, G. Getz, S. Monti, M. Reich, and J. P. Mesirov, "Comparative gene marker selection suite," *Bioinformatics*, vol. 22, pp. 1924-5, Aug 1 2006.
- [49] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and J. P. Mesirov, "GenePattern 2.0," *Nat Genet*, vol. 38, pp. 500-1, May 2006.
- [50] V. Zuber and K. Strimmer, "Gene ranking and biomarker discovery under correlation," *Bioinformatics*, vol. 25, pp. 2700-7, Oct 15 2009.
- [51] A. L. Boulesteix, "PLS dimension reduction for classification with microarray data," *Stat Appl Genet Mol Biol*, vol. 3, p. Article33, 2004.
- [52] A. C. Culhane, G. Perriere, E. C. Considine, T. G. Cotter, and D. G. Higgins, "Between-group analysis of microarray data," *Bioinformatics*, vol. 18, pp. 1600-8, Dec 2002.
- [53] I. T. Jolliffe, *Principal component analysis*, 2nd ed. New York: Springer, 2002.
- [54] D. K. Slonim, "From patterns to pathways: gene expression data analysis comes of age," *Nat Genet*, vol. 32 Suppl, pp. 502-8, Dec 2002.
- [55] M. Yan, "Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion," Ph.D. Thesis, Statistics, Virginia Polytechnic Institute and State University, 2005.
- [56] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *PSYCHOMETRIKA*, vol. 50, pp. 159-179, 1985.
- [57] W. J. Krzanowski and Y. T. Lai, "A criterion for determining the number of groups in a data set using sum of squares clustering," *Biometrics*, vol. 44, pp. 23-34, 1988.
- [58] M. Walesiak, "Cluster analysis with ClusterSim computer program and R environment," *Acta Universitatis Lodziniensis Folia Oeconomica*, vol. 216, pp. 303-311, 2008.
- [59] A. Butte, "The use and analysis of microarray data," *Nat Rev Drug Discov*, vol. 1, pp. 951-60, Dec 2002.
- [60] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys (CSUR)*, vol. 31, pp. 264-323 September 1999.
- [61] D. Jiang, C. Tang, and a. A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 16, pp. 1370-1386, 2004.
- [62] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, *et al.*, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proc Natl Acad Sci U S A*, vol. 96, pp. 2907-12, Mar 16 1999.
- [63] S. Mukherjee, "Classifying Microarray Data Using Support Vector Machines," in *Understanding And Using Microarray Analysis Techniques: A Practical Guide*, ed Boston: MA Kluwer Academic Publishers, 2003.
- [64] M. Gutkin, "Feature selection methods for classification of gene expression profiles," MSc Thesis, School of Computer Science, Tel-Aviv University, 2008.
- [65] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification," ed, 2003.
- [66] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*: CAMBRIDGE UNIVERSITY PRESS 2000.
- [67] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, pp. 37-66, 1991.
- [68] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

- [69] R. Hewett and P. Kijsanayothin, "Tumor classification ranking from microarray data," *BMC Genomics*, vol. 9 Suppl 2, p. S21, 2008.
- [70] A. Sakellariou, D. Sanoudou, and G. Spyrou, "Combining multiple hypothesis testing and affinity propagation clustering leads to accurate, robust and sample size independent classification on gene expression data," *BMC Bioinformatics*, vol. 13, p. 270, 2012.
- [71] K. S. Pollard, S. Dudoit, and M. J. v. d. Laan, "Multiple Testing Procedures: the multtest Package and Applications to Genomics," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, ed, 2005, pp. 249-271.
- [72] A. Sakellariou, D. Sanoudou, and G. Spyrou, "Investigating the minimum required number of genes for the classification of neuromuscular disease microarray data," *IEEE Trans Inf Technol Biomed*, vol. 15, pp. 349-55, May 2011.
- [73] D. Delbert, "Affinity Propagation: Clustering Data by Passing Messages," Doctor of Philosophy, Graduate Department of Electrical & Computer Engineering, University of Toronto, 2009.
- [74] U. Bodenhofer, A. Kothmeier, and S. Hochreiter, "APCluster: an R package for affinity propagation clustering," *Bioinformatics*, vol. 27, pp. 2463-4, Sep 1 2011.
- [75] S. E. Choe, M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon, "Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset," *Genome Biol*, vol. 6, p. R16, 2005.
- [76] R. Opgen-Rhein and K. Strimmer, "Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach," *Stat Appl Genet Mol Biol*, vol. 6, p. Article9, 2007.
- [77] M. Bakay, Z. Wang, G. Melcon, L. Schiltz, J. Xuan, P. Zhao, *et al.*, "Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration," *Brain*, vol. 129, pp. 996-1013, Apr 2006.
- [78] D. Sanoudou and A. H. Beggs, "Clinical and genetic heterogeneity in nemaline myopathy--a disease of skeletal muscle thin filaments," *Trends Mol Med*, vol. 7, pp. 362-8, Aug 2001.
- [79] D. Sanoudou, J. N. Haslett, A. T. Kho, S. Guo, H. T. Gazda, S. A. Greenberg, *et al.*, "Expression profiling reveals altered satellite cell numbers and glycolytic enzyme transcription in nemaline myopathy muscle," *Proc Natl Acad Sci U S A*, vol. 100, pp. 4666-71, Apr 15 2003.
- [80] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc Natl Acad Sci U S A*, vol. 96, pp. 6745-50, Jun 8 1999.
- [81] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-7, Oct 15 1999.
- [82] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, *et al.*, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, pp. 203-9, Mar 2002.
- [83] J. B. Welsh, L. M. Sapinoso, A. I. Su, S. G. Kern, J. Wang-Rodriguez, C. A. Moskaluk, *et al.*, "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer," *Cancer Res*, vol. 61, pp. 5974-8, Aug 15 2001.
- [84] R. Shen, D. Ghosh, A. Chinnaiyan, and Z. Meng, "Eigengene-based linear discriminant model for tumor classification using gene expression microarray data," *Bioinformatics*, vol. 22, pp. 2635-42, Nov 1 2006.
- [85] H. Hu, J. Li, A. Plank, H. Wang, and G. Daggard, "A Comparative Study of Classification Methods For Microarray Data Analysis," in *In Proceedings of the Fifth Australasian Conference on Data Mining and Analytics*, Sydney, Australia, 2006, pp. 33-37.
- [86] H. Moon, H. Ahn, R. L. Kodell, C. J. Lin, S. Baek, and J. J. Chen, "Classification methods for the development of genomic signatures from high-dimensional data," *Genome Biol*, vol. 7, p. R121, 2006.
- [87] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," *Appl Bioinformatics*, vol. 2, pp. S75-83, 2003.

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

- [88] B. Liu, Q. Cui, T. Jiang, and S. Ma, "A combinational feature selection and ensemble neural network method for classification of gene expression data," *BMC Bioinformatics*, vol. 5, p. 136, Sep 27 2004.
- [89] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, pp. 1131-42, Dec 2001.
- [90] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, pp. 906-14, Oct 2000.
- [91] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue classification with gene expression profiles," *J Comput Biol*, vol. 7, pp. 559-83, 2000.
- [92] D. V. Nguyen and D. M. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, pp. 39-50, Jan 2002.
- [93] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, P. J. Messirov, *et al.*, "Support vector machine classification of microarray data," MIT AI memo 182, 2000.
- [94] S. Dudoit, J. Fridlyand, and P. T. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, pp. 77-87, 2002.
- [95] A. V. Antonov, I. V. Tetko, M. T. Mader, J. Budczies, and H. W. Mewes, "Optimization models for cancer classification: extracting gene interaction information from microarray expression data," *Bioinformatics*, vol. 20, pp. 644-52, Mar 22 2004.
- [96] Z. Liu and D. Chen, "Gene expression data classification with revised kernel partial least squares algorithm," in *Proceedings of the 17th International FLAIRS Conference*, South Beach, Florida, USA, 2004, pp. 104-108.
- [97] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proc Natl Acad Sci U S A*, vol. 99, pp. 6567-72, May 14 2002.
- [98] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc, "Effective dimension reduction methods for tumor classification using gene expression data," *Bioinformatics*, vol. 19, pp. 563-70, Mar 22 2003.
- [99] I. Vastrik, P. D'Eustachio, E. Schmidt, G. Gopinath, D. Croft, B. de Bono, *et al.*, "Reactome: a knowledge base of biologic pathways and processes," *Genome Biol*, vol. 8, p. R39, 2007.
- [100] B. Zhang, S. Kirov, and J. Snoddy, "WebGestalt: an integrated system for exploring gene sets in various biological contexts," *Nucleic Acids Res*, vol. 33, pp. W741-8, Jul 1 2005.
- [101] J. Jourquin, D. Duncan, Z. Shi, and B. Zhang, "GLAD4U: deriving and prioritizing gene lists from PubMed literature," *BMC Genomics*, vol. 13 Suppl 8, p. S20, 2012.
- [102] B. S. Cowling, M. J. McGrath, M. A. Nguyen, D. L. Cottle, A. J. Kee, S. Brown, *et al.*, "Identification of FHL1 as a regulator of skeletal muscle mass: implications for human myopathy," *J Cell Biol*, vol. 183, pp. 1033-48, Dec 15 2008.
- [103] C. Toledano, M. Gain, A. Kettaneh, B. Baudin, C. Johanet, P. Cherin, *et al.*, "Aldolase predicts subsequent myopathy occurrence in systemic sclerosis," *Arthritis Res Ther*, vol. 14, p. R152, 2012.
- [104] J. N. Haslett, D. Sanoudou, A. T. Kho, R. R. Bennett, S. A. Greenberg, I. S. Kohane, *et al.*, "Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle," *Proc Natl Acad Sci U S A*, vol. 99, pp. 15000-5, Nov 12 2002.
- [105] A. Oldfors, "Hereditary myosin myopathies," *Neuromuscul Disord*, vol. 17, pp. 355-67, May 2007.
- [106] L. Madaro, A. Pelle, C. Nicoletti, A. Crupi, V. Marrocco, G. Bossi, *et al.*, "PKC theta ablation improves healing in a mouse model of muscular dystrophy," *PLoS One*, vol. 7, p. e31515, 2012.
- [107] S. A. Greenberg, J. L. Pinkus, G. S. Pinkus, T. Bursleson, D. Sanoudou, R. Tawil, *et al.*, "Interferon-alpha/beta-mediated innate immune mechanisms in dermatomyositis," *Ann Neurol*, vol. 57, pp. 664-78, May 2005.

Computational Methods for the Identification of Statistically Significant Genes: Applications to Gene Expression Data of Various Human Diseases

- [108] D. R. Rhodes, B. Ateeq, Q. Cao, S. A. Tomlins, R. Mehra, B. Laxman, *et al.*, "AGTR1 overexpression defines a subset of breast cancer and confers sensitivity to losartan, an AGTR1 antagonist," *Proc Natl Acad Sci U S A*, vol. 106, pp. 10284-9, Jun 23 2009.
- [109] G. Bode, A. Luken, C. Kerkhoff, J. Roth, S. Ludwig, and W. Nacken, "Interaction between S100A8/A9 and annexin A6 is involved in the calcium-induced cell surface exposition of S100A8/A9," *J Biol Chem*, vol. 283, pp. 31776-84, Nov 14 2008.
- [110] C. Alfaro, N. Suarez, I. Martinez-Forero, A. Palazon, A. Rouzaut, S. Solano, *et al.*, "Carcinoma-derived interleukin-8 disorients dendritic cell migration without impairing T-cell stimulation," *PLoS One*, vol. 6, p. e17922, 2011.
- [111] L. Voisin, C. Julien, S. Duhamel, K. Gopalbhai, I. Claveau, M. K. Saba-El-Leil, *et al.*, "Activation of MEK1 or MEK2 isoform is sufficient to fully transform intestinal epithelial cells and induce the formation of metastatic tumors," *BMC Cancer*, vol. 8, p. 337, 2008.
- [112] S. Song, J. C. Byrd, N. Mazurek, K. Liu, J. S. Koo, and R. S. Bresalier, "Galectin-3 modulates MUC2 mucin expression in human colon cancer cells at the level of transcription via AP-1 activation," *Gastroenterology*, vol. 129, pp. 1581-91, Nov 2005.
- [113] L. V. July, M. Akbari, T. Zellweger, E. C. Jones, S. L. Goldenberg, and M. E. Gleave, "Clusterin expression is significantly enhanced in prostate cancer cells following androgen withdrawal therapy," *Prostate*, vol. 50, pp. 179-88, Feb 15 2002.
- [114] C. A. Acevedo, L. A. Quinones, J. Catalan, D. D. Caceres, J. A. Fulla, and A. M. Roco, "Impact of CYP1A1, GSTM1, and GSTT1 polymorphisms in overall and specific prostate cancer survival," *Urol Oncol*, vol. 32, pp. 280-90, Apr 2014.
- [115] A. Sakellariou, "mAPKL," ed. Bioconductor, 2014.
- [116] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biol*, vol. 5, p. R80, 2004.
- [117] Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer, and A. Weingessel, "e1071: Misc Functions of the Department of Statistics (e1071)," ed. TU Wien, 2010.
- [118] G. Turashvili, J. Bouchal, K. Baumforth, W. Wei, M. Dziechciarkova, J. Ehrmann, *et al.*, "Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis," *BMC Cancer*, vol. 7, p. 55, 2007.
- [119] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, *et al.*, "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS Biol*, vol. 5, p. e8, Jan 2007.
- [120] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc Natl Acad Sci U S A*, vol. 95, pp. 14863-8, Dec 8 1998.