KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

MICROARRAYS: ALGORITHMS FOR KNOWLEDGE DISCOVERY IN ONCOLOGY AND MOLECULAR BIOLOGY

Jury:
Prof. dr. ir. P. Verbaeten, voorzitter
Prof. dr. ir. B. De Moor, promotor
Prof. dr. ir. S. Van Huffel
Prof. dr. K. Kas (Harvard University; VIB)
Prof. dr. I. Vergote
Prof. dr. D. Timmerman

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de toegepaste wetenschappen

door

Frank DE SMET

UDC 681.3*J3:616-006

Mei 2004

Arenbergkasteel, Kasteelpark Arenberg 1, B-3001 Heverlee (Belgium)

D/2004/7515/37

ISBN 90-5682-503-8

Voorwoord

Toen ik in het voorjaar van 1999 op zoek was naar een manier om mijn studies geneeskunde en toegepaste wetenschappen op een evenwichtige wijze te combineren, kwam ik in contact met Prof. Bart De Moor. Hij bood me de gelegenheid om onderzoek te doen en een doctoraat te maken in een nieuwe en interdisciplinaire groep die bio-informatica ging bestuderen. Vermits ik hier ondermeer de kans zou krijgen om me te specialiseren in de klinische toepassingen van deze jonge wetenschap, heb ik geen moment geaarzeld en met veel enthousiasme deze opdracht aanvaard. Graag zou ik Prof. Bart De Moor willen bedanken voor de kansen die hij me gegeven heeft.

Voorts zou ik ook graag de leden van mijn begeleidingscommissie, Prof. Sabine Van Huffel en Prof. Koen Kas, willen bedanken voor de steun die ze me tijdens dit onderzoek hebben gegeven en voor het doornemen van dit proefschrift.

Bovendien zou ik ook Prof. P. Verbaeten, als voorzitter, Prof. Ignace Vergote en Prof. Dirk Timmerman willen bedanken dat zij deel willen uitmaken van de jury van dit doctoraatsproefschrift.

Natuurlijk zou de voorliggende tekst niet tot stand zijn gekomen zonder de interactie en hulp van de andere medewerkers van de bio-informaticagroep en SCD. Vooraleerst heb ik zeer veel waardering voor de post-docs die me altijd met raad en daad hebben bijgestaan: Dr. Yves Moreau voor de vele inspirerende ideeën en de gemeenschappelijke interesse in de klinische toepassingen, Dr. Kathleen Marchal en Dr. Janick Mathys om me in te leiden in de geheimen van de moleculaire biologie en voor de vele suggesties die mijn onderzoek op het goede pad hebben gebracht. Ook zou ik Prof. Johan Suykens willen bedanken voor de vele discussies en tips in verband met de meer wiskundige aspecten van dit onderzoek. Bovendien wil ik ook Prof. Joos Vandewalle, als hoofd van onze afdeling, bedanken voor alle steun.

Tevens wil ik de verschillende leden bedanken van de afdeling gynaecologie-verloskunde van het U.Z.Leuven waarmee er op regelmatige basis is samengewerkt. Een speciaal dankwoord voor Prof. Dirk Timmerman is zeker op zijn plaats. Hij heeft me geintroduceerd bij zijn collega's en me steeds met enthousiasme gestimuleerd om samen te werken. Bovendien heeft hij me op meerdere momenten de zo onmisbare data ter beschikking gesteld. Prof. Ignace Vergote wil ik ook bedanken voor het vertrouwen dat hij mij heeft gegeven om mee te werken aan het opstarten en aanvragen van meerdere projecten, die zonder zijn bijdrage nooit gerealiseerd zouden kunnen worden. In dit verband, zou ik hier ook Dr. Paul Van Hummelen (Microarray Facility van het V.I.B.) willen vermelden voor de aangename en professionele samenwerking tijdens deze projecten. Als laatste wil ik ook Prof. Thomas D'Hooghe bedanken om ons uit te kiezen als partner in verband met het onderzoek naar endometriose.

I would also like to thank Dr. Elisabeth Epstein and Prof. Lil Valentin for giving me the chance to collaborate in a joint paper.

Zeker mag ik mijn collega's binnen de bioinformaticagroep en SCD, niet vergeten die altijd klaar stonden als ik hulp nodig had en die er steeds voor zorgden dat iedereen zich gewaardeerd voelde. Een speciale vermelding verdienen zeker Bart, Bert (2x), Cynthia, Frizo, Geert, Gert, Joke, Jos, Kristof, Leentje, Maarten, Nathalie, Patrick, Pieter, Raf, Ruth, Steffen, Stein, Steven, Tijl, en Tom. Ook mag ik Bart, Ida, Ilse en Pela niet vergeten voor alle logistieke steun en de hartelijke gesprekken.

Dit onderzoek werd mogelijk gemaakt door de K.U.Leuven dat de nodige fondsen ter beschikking heeft gesteld om mij te financieren: eerst als wetenschappelijk medewerker (op het IWT-STWW-Genprom project) en vanaf september 2003 als doctoraatsbursaal (op het FWO-project G.0115.01).

Als laatste, maar zeker niet in het minste, zou ik mijn familie willen bedanken voor al de liefdevolle steun die ze me hebben gegeven. Een speciale vermelding voor mijn ouders is hier op zijn plaats omdat zij altijd in mij hebben geloofd en voor de kansen die ze mij hebben gegeven gedurende mijn lange (11 jaren) studies. Mijn lieve echtgenote, Ilse, en mijn twee schatten van kindjes, Lieselot en Stijn, zou ik willen bedanken om me een thuis te geven waar het mogelijk was om dit werk tot een goed einde te brengen.

Abstract

In this thesis we have studied a general data-mining framework (feature extraction, classification and clustering) that could be used to analyse clinical and microarray and, in the future, proteomic data. We have mainly applied this framework to oncology related problems.

For the prediction of the degree of myometrial invasion in endometrial cancer, we developed three models that aim to discriminate between patients with and without deep myometrial invasion using ultrasound and histopathological data.

For the analysis of microarray experiments, we evaluated the use of principal component analysis. In addition, we examined some elementary clustering techniques (K-means and hierarchical clustering). We applied and compared the performance of Fisher's linear discriminant analysis and Least Squares Support Vector Machines for the classification of expression patterns of malignancies. Based on these results, we concluded that regularization or dimensionality reduction is necessary. Subsequently, we gave a general overview of existing techniques to cluster gene expression profiles and noted that they do not have all the desired properties for this task. This observation was the basis for the development and validation of our own algorithm called adaptive quality-based clustering. Finally, we presented an in-depth study of univariate analysis in microarray data. We described a method to estimate the total number of genes whose expression is and is not affected by a difference in tumour type. We described how a Receiver Operating Characteristic (ROC) curve could be applied to define an optimal rejection level and showed that the area under the ROC curve could be used to assign a quality measure to microarray data.

In the description of our future research, we presented some concrete clinical projects in which we will use the data-mining framework for the analysis of microarray and proteomic data.

Samenvatting

In dit proefschrift hebben we een algemeen kader voor gegevensontginning (selectie van kenmerken, classificatie en clustering) bestudeerd dat kan gebruikt worden voor de analyse van klinische en microroosterdata en, in de toekomst, van proteoomdata. We hebben dit hoofdzakelijk toegepast voor problemen in de oncologie.

Betreffende de voorspelling van de diepte van myometriuminfiltratie bij endometriumcarcinomen, hebben we drie modellen ontwikkeld die gebruik maken van gegevens bekomen uit het echografisch en histopathologisch onderzoek en die een onderscheid trachten te maken tussen patiënten met en zonder diepe invasie.

Betreffende de analyse van microroosterexperimenten, hebben we het gebruik van Principale Component Analyse geëvalueerd. Bovendien hebben we in deze context enkele elementaire clusteringstechnieken bestudeerd (K-means-clustering en hiërarchische clustering). We hebben Lineaire Discriminant Analyse en kleinste kwadraten Support Vector Machines gebruikt en vergeleken met betrekking tot de classificatie van expressiepatronen van maligniteiten. Hieruit is gebleken dat regularisatie of een afname van de dimensionaliteit noodzakelijk is in combinatie met de classificatie van microroosterexperimenten. Vervolgens hebben we een overzicht gegeven van bestaande technieken voor het clusteren van genexpressieprofielen en opgemerkt dat deze methoden niet altijd optimaal zijn. Deze observatie heeft dan geleid tot de ontwikkeling en validatie van een nieuw algoritme dat we adaptief kwaliteitsgebaseerd clusteren hebben genoemd. Tot slot hebben we een grondige studie verricht van univariate analyse van microroostergegevens. We hebben een methode besproken die het mogelijk maakt om het aantal genen te schatten wiens expressie wel en niet wordt beïnvloed door een verschil in het type van de tumor. We hebben beschreven hoe een Receiver Operating Characteristic (ROC) curve kan gebruikt worden voor de bepaling van het optimaal niveau waarop de nulhypothese moet worden verworpen en hebben aangetoond dat de

oppervlakte onder de ROC-curve kan dienen om de kwaliteit van microroostergegevens te kwantificeren.

In de beschrijving van ons toekomstig onderzoek hebben we enkele concrete klinische projecten voorgesteld waarin de technieken beschreven in dit proefschrift kunnen gebruikt worden voor de analyse van zowel microrooster- als proteoomdata.

Nederlandse samenvatting

Microroosters: algoritmen voor kennisextractie in de oncologie en moleculaire biologie

Hoofdstuk 1: Inleiding

Motivatie

Het klinisch beleid bij kwaadaardige processen is in vele gevallen gedeeltelijk empirisch en wordt gestuurd door gegevens uit de literatuur (bekomen uit klinische studies) of de persoonlijke ervaring van de clinicus. De huidige diagnostische schema's vertonen nog dikwijls een significante variabiliteit tussen verschillende artsen en vereisen vaak een bijkomende en soms subjectieve beoordeling. Bovendien kan niet alle informatie die klinisch relevant is uit de gegevens worden gehaald die een clinicus op dit moment tot zijn beschikking heeft. Methoden die bijvoorbeeld toelaten om een meer objectieve en betere toewijzing aan de verschillende diagnostische klassen te bekomen, zouden dus nuttig kunnen zijn.

Moleculaire biologie

De fundamentele processen die aan de basis liggen van de carcinogenese worden in de meeste gevallen nog niet gebruikt om het klinisch beleid te helpen bepalen. Het ontstaan van kanker is immers een proces dat zich voor een groot deel afspeelt op het niveau van het genoom. Onder invloed van bepaalde factoren (bestraling, virale infecties, …) kunnen mutaties ontstaan in bepaalde genen (bijvoorbeeld proto-oncogenen en tumorsuppressorgenen) met eventueel ongecontroleerde celgroei en de mogelijkheid tot invasie en metastasering tot gevolg. Door deze mutaties kan echter ook de transcriptie of translatie van andere genen (waarin geen mutatie optreedt, maar waarvan de transcriptie of translatie direct of indirect wordt geregeld, bijvoorbeeld als het gemuteerd gen codeert voor een transcriptiefactor) ontregeld worden. Het is waarschijnlijk dat het betrekken van de effecten van deze mutaties in de klinische besluitvorming een verbetering zou betekenen in vergelijking met de meer empirische beslissingsschema's die nu gebruikt worden. Het behoort tot de verwachtingen dat de analyse van data (afkomstig van microroosters of de analyse van het proteoom - zie verder) die het moleculair biologisch gedrag van tumorcellen weerspiegelen, een belangrijke vooruitgang kan betekenen in het wetenschappelijk onderzoek naar het gedrag en ontstaan van tumoren.

In dit proefschrift bestuderen we een algemeen kader voor gegevensontginning dat kan gebruikt worden voor de analyse van klinische, microrooster- en proteoomdata. We passen dit voornamelijk toe voor problemen uit de of gerelateerd aan oncologie. Vooraleerst is het de bedoeling om diagnostische vraagstukken nauwkeuriger en objectiever te formuleren aan de hand van klinische data. Bovendien is het de bedoeling om microrooster- en proteoomdata, aan de hand van specifieke algoritmen, te integreren in de klinische besluitvorming en om ze te gebruiken om een meer fundamenteel inzicht te verkrijgen in de moleculaire biologie achter de carcinogenese.

In de volgende secties worden de verschillende datatypes en de verschillende elementen van het algemeen kader voor gegevensontginning verder toegelicht.

Datatypes

1. Klinische data: dit datatype bevat waarden voor klassieke klinische parameters (de variabelen; bijvoorbeeld gegevens uit de klinische biologie, uit de medische beeldvorming, uit het histopathologisch onderzoek, uit het klinisch onderzoek, uit de anamnese) die gewoonlijk worden vergaard in het kader van een zeker diagnostisch probleem voor een zekere groep van patiënten. In vergelijking met de volgende datatypes is het aantal variabelen meestal een aantal grootte-ordes kleiner.

2. Microroosterdata: microroosters bestaan uit een groot aantal sondes samengebracht op een klein oppervlak. Sterk vereenvoudigd kan gesteld worden dat ieder van deze sondes bestaat uit DNA dat complementair is aan één welbepaalde mRNA-streng (ze zijn dus specifiek voor één welbepaald gen). Iedere mRNA-streng (of het overeenkomstig cDNA) zal dus specifiek binden aan (of hybridiseren met) zijn complementaire sonde(s) wanneer het totaal mRNA, afkomstig uit cellen van een welbepaald celtype, in contact wordt gebracht met de sondes op het microrooster. De binding van iedere complementaire sonde met zijn overeenkomstig mRNA kan gemeten worden en is dus een maat voor de hoeveelheid mRNA

(expressieniveau) afkomstig van één welbepaald gen. De twee belangrijkste soorten microroosters zijn cDNA-microroosters (zie Duggan (1999) en Figuur 1.2) en oligonucleotideroosters (GeneChip®, Affymetrix Inc. - zie Lipshutz (1999)).

Zoals gezegd, kunnen mutaties die aan de basis liggen van het ontstaan van kwaadaardige processen, ook bij niet-gemuteerde genen verstoring van hun expressie veroorzaken. Het is nu de verzameling van deze ontregelde genexpressies die het fenotype van de tumorcel bepaalt (Sager, 1997). Het meten van een groot gedeelte van deze expressieniveaus met microroosters zou dus van grote waarde kunnen zijn om het werkelijk gedrag van de tumorcellen te kennen, te voorspellen en te begrijpen.

Vermits ieder experiment met een microrooster resulteert in een hoogdimensionale vector met duizenden waarden of componenten (één per sonde op het microrooster), moeten er aangepaste technieken worden toegepast voor de analyse van microroosterdata.

3. Proteoomdata: omwille van posttranscriptionele modificatie en regulatie van biologisch actieve moleculen is het mogelijk dat door de meting van de expressieniveaus met microroosters niet alle relevante fenomenen in een cel op het moleculair biologisch vlak worden waargenomen. Dat wil dus zeggen dat door de studie van het proteoom (verzameling van alle proteïnen in een cel) het eventueel mogelijk is om complementaire informatie te bekomen over de fundamentele processen die zich afspelen binnenin een bepaalde cel. Dit kan gebeuren door middel van recente technologieën die gebaseerd zijn op massaspectrometrie en die het mogelijk maken om de aanwezigheid van een brede subset proteïnen in een staal te kwantificeren (voor een voorbeeld zie Chapman (2002)). De gegevens die hieruit resulteren zullen niet expliciet worden geanalyseerd in dit proefschrift maar wel besproken worden in het kader van de voorstelling van enkele concrete toepassingen die gepland zijn tijdens ons toekomstig onderzoek (Hoofdstuk 7). Kwalitatief bestaat de uitvoer van deze technologieën uit spectra die bestaan uit duizenden discrete waarden of piekamplitudes elk geassocieerd aan een welbepaalde waarde voor massa/lading die op zijn beurt overeenkomt met een zeker (onbekend) proteïne. Deze spectra zijn dan karakteristiek voor de proteïnen of een subklasse van de proteïnen aanwezig in een staal. Dit resulteert dus eveneens in datavectoren die duizenden waarden bevatten en waarbij iedere component van deze vector representatief is voor de hoeveelheid van een niet nader bepaald proteïne in het bestudeerde staal. De uitvoer is dus kwalitatief gelijkaardig aan microroostergegevens en kan dus mogelijks geanalyseerd worden met gelijkaardige technieken.

Algemeen kader voor gegevensontginning

Het algemeen kader voor gegevensontginning bestaat uit de volgende drie elementen (zie ook Figuur 1.5):

1. Selectie van kenmerken: niet al de variabelen in een dataset zijn geschikt om in verdere analyses gebruikt te worden. Het is beter om een beperkte verzameling van kenmerken (bijvoorbeeld individuele variabelen, een groep van variabelen of een combinatie van variabelen) te selecteren die optimaal gebruikt kunnen worden bij classificatie en clustering (zie volgende twee punten). In deze tekst beschouwen we twee verschillende manieren om kenmerken te selecteren: univariaat en multivariaat.

   Bij univariate selectie van kenmerken veronderstelt men dat de datapunten tot een beperkt aantal klassen behoren en heeft men als doelstelling om de individuele variabelen te selecteren die maximaal gecorreleerd zijn met de verschillende klassen. In dit geval maakt men typisch gebruik van hypothesetesten (Dawson-Saunders en Trapp, 1994). Deze techniek wordt voor microroosterdata echter bemoeilijkt door het probleem van meervoudig testen.

   Een eerste techniek voor multivariate analyse betreft het selecteren van een groep van variabelen die, wanneer ze gecombineerd worden in een bepaald model, een statistisch significante bijdrage leveren tot de nauwkeurigheid van de voorspelling. Dit wordt modelselectie genoemd en gebeurt door een iteratief proces waarbij de variabelen achtereenvolgens worden toegevoegd aan of verwijderd uit het model. Deze techniek wordt veel gebruikt in combinatie met standaard logistieke regressie (zie Hosmer en Lemeshow (1989)). Een tweede techniek voor multivariate analyse betreft de identificatie van een (lineaire of niet-lineaire) functie of combinatie van variabelen die een gewenste eigenschap heeft. Bij Principale Component Analyse (Bishop, 1995), bijvoorbeeld, wordt er een lineaire combinatie gezocht van de variabelen die een maximale variantie vertoont over een verzameling datapunten. Dit is een techniek die we bij voorkeur zullen gebruiken bij de analyse van microroosterexperimenten.

2. Classificatie: hier worden wiskundige modellen geconstrueerd die kunnen voorspellen tot welke klasse een welbepaald datapunt behoort. Aan de hand van een modelstructuur, een verzameling van kenmerken en een trainingsset (d.i. een verzameling datapunten waarvan reeds geweten is tot welke klasse ze behoren, m.a.w. de kentekens of labels van de datapunten zijn gekend) worden de parameters of coëfficiënten van het model bepaald. Dit noemt men het trainen van het model. Dit model kan vervolgens worden getest op nieuwe datapunten waarvan wordt verondersteld dat de kentekens niet gekend zijn.

3. Clustering: met clusteranalyse is het mogelijk om automatisch verschillende klassen of clusters te ontdekken in een groep datapunten zonder voorafgaande kennis van de eigenschappen van die clusters (Kaufman en Rousseeuw, 1990). Een cluster zal in het algemeen een aantal datapunten bevatten die een zekere graad van overeenkomst vertonen volgens een bepaalde afstandsfunctie.

Hoofdstuk 2: Klinische data-analyse: voorspelling van de infiltratiediepte van endometriumcarcinomen

In dit hoofdstuk wordt het algemeen kader voor gegevensontginning toegepast voor klinische data afkomstig van patiënten met een endometriumcarcinoom (kwaadaardig proces van het slijmvlies van de baarmoeder of uterus). De graad van myometriale invasie (myometrium = spierlaag van de uterus) is een belangrijke prognostische factor met een belangrijke impact op het beleid. Hier wordt er een onderscheid gemaakt tussen patiënten met een invasiediepte die kleiner is dan 50% van de totale dikte van het myometrium (groep I - FIGO stadium Ia of Ib) of die groter is dan 50% van de totale dikte van het myometrium (groep II - FIGO stadium Ic of hoger). Een echografisch onderzoek (transvaginale echografie (TVS) met kleuren Doppler (CDI)) en een histopathologisch onderzoek van een endometriale biopsie horen meestal bij de initiële evaluatie van deze patiënten. Prof. Dr. D. Timmerman (afdeling gynaecologie-verloskunde, U.Z.Leuven) heeft gegevens die resulteren uit deze evaluatie verzameld voor 97 patiënten. Deze groep van patiënten noemen we verder ook de trainingsset en worden gebruikt voor de univariate analyse, voor de multivariate analyse of modelselectie en voor het trainen van drie modeltypes: standaard logistieke regressie en kleinste kwadraten Support Vector Machines (LS-SVM) met een lineaire en radiale basisfunctie (RBF) kernel.

Univariate analyse (zie ook Tabel 2.2) van de echografische parameters wees uit dat de ratio (EV/UV) van het endometriumvolume (EV) en het volume van de uterus (UV) de grootste oppervlakte (AUC) onder de Receiver Operating Characteristic (ROC) curve had (78%) en dat deze oppervlakte kleiner was dan deze van de subjectieve beoordeling door de expert (79%). Er was echter geen significant verschil tussen de AUC van EV/UV en de AUCs van de endometriumdikte (ET), de myometriumdikte (MT), EV, de ratio (ET/AP) van ET en de voorachterwaartse diameter van de uterus (AP) en MT/AP. De AUC van de CDI parameters (van de linker en rechter arteria uterina en intratumoraal gemeten) was klein.

Multivariate analyse met stapsgewijze logistieke regressie wees de differentiatiegraad, het aantal fibromen (leiomyomen), ET en EV aan als de variabelen die significant bijdragen in een standaard logistiek regressiemodel. CDI parameters droegen niet significant bij. Dit resulteerde dan in het volgende logistieke regressiemodel:

$$y = \frac{\exp(\beta_0 + \beta_1.\mathrm{DD1} + \beta_2.\mathrm{DD2} + \beta_3.\mathrm{NF} + \beta_4.\mathrm{ET} + \beta_5.\mathrm{EV})}{1 + \exp(\beta_0 + \beta_1.\mathrm{DD1} + \beta_2.\mathrm{DD2} + \beta_3.\mathrm{NF} + \beta_4.\mathrm{ET} + \beta_5.\mathrm{EV})} \quad (1)$$

waar DD1 and DD2 gelijk zijn aan 1 als, respectievelijk, de tumor matig en slecht gedifferentieerd is en gelijk zijn aan 0 in alle andere gevallen. De coëfficiënten zijn: β_0 = -3.70, β_1 = 2.36, β_2 = 2.42, β_3 = -2.45, β_4 = 0.20, en β_5 = -0.11. De AUC van dit logistieke regressiemodel geëvalueerd op de trainingsset is 89% (zie ook Tabel 2.2).

Aan de hand van de vier variabelen die werden geselecteerd door stapsgewijze logistieke regressie, hebben we ook een LS-SVM-model met een lineaire en een LS-SVM-model met een RBF-kernel getraind. Voor het LS-SVM-model met een lineaire kernel is het mogelijk om, na een herschikking van de termen, dit te schrijven als een eenvoudige lineaire functie van de variabelen:

$$y = \beta_0 + \beta_1.\mathrm{DD} + \beta_2.\mathrm{NF} + \beta_3.\mathrm{ET} + \beta_4.\mathrm{EV} \quad (2)$$

waar DD gelijk is aan 1, 2 en 3 als de tumor goed, matig en weinig gedifferentieerd is, respectievelijk. De coëfficiënten zijn: β_0 = -1.45, β_1 = 0.37, β_2 = -0.38, β_3 = 0.05, en β_4 = -0.03. Het LS-SVM-model met een RBF-kernel kan niet in een eenvoudige vorm worden neergeschreven en wordt hier daarom niet expliciet beschreven. De AUCs van de LS-SVM-modellen met een lineaire en RBF-kernel geëvalueerd op de trainingsset zijn 88% en 99%, respectievelijk (Tabel 2.2).

We hebben deze drie modellen eveneens prospectief gevalideerd op een nieuwe verzameling van 37 patiënten (zie Tabel 2.3). De AUCs van het standaard logistieke regressiemodel en de LS-SVM-modellen met een lineaire en RBF-kernel geëvalueerd op deze nieuwe dataset zijn respectievelijk: 81%, 90% en 92%. De drie modellen hebben allen een betere AUC dan de subjectieve beoordeling door de expert (74%) maar het verschil is enkel significant voor het LS-SVM-model met een RBF-kernel ($p = 0.0485$). Uit deze resultaten blijkt dus dat dit laatste model het beste presteert voor de onderzochte patiënten.

Als conclusie kunnen we zeggen dat CDI niet bijdraagt tot het voorspellen van de invasiediepte van endometriumcarcinomen en dat individuele morfologische parameters bepaald door TVS niet voldoende zijn om een nauwkeurige voorspelling te maken. Het combineren van de differentiatiegraad, de endometriumdikte, het endometriale volume en het aantal fibromen in een standaard logistiek regressiemodel, in een LS-SVM-model met een lineaire kernel en vooral in een LS-SVM-model met een RBF-kernel, zouden deze voorspelling kunnen verbeteren. Deze methodiek zou een eenvoudige en goedkope manier kunnen vertegenwoordigen die kan bijdragen tot een betere preoperatieve scheiding tussen patiënten met een laag en hoog risico. Er is echter nog veel werk nodig vooraleer de modellen die hier beschreven worden, echt bruikbaar worden in de klinische praktijk. Vooraleerst werden de modellen afgeleid met behulp van gegevens die afkomstig zijn van dezelfde expert. Omdat er verschillen mogelijk zijn tussen verschillende experts, is het nodig om deze modellen verder te valideren (en indien nodig aan te passen) met gegevens die afkomstig zijn van meerdere centra. Bovendien kunnen er wijzigingen optreden in de karakteristieken van de patiëntenpopulatie, wat het nodig maakt om deze modellen continu te evalueren.

Tenslotte merken we nog op dat we deelgenomen hebben aan een gelijkaardige studie (Epstein et al., 2002) waar we eveneens ROC-curven hebben gebruikt voor het vergelijken van verschillende modellen die de aanwezigheid van een endometriumcarcinoom trachten te voorspellen in vrouwen met postmenopausaal bloedverlies.

Hoofdstuk 3: Analyse van microroosterdata

In dit hoofdstuk wordt het algemeen kader voor gegevensontginning toegepast voor microroostergegevens afkomstig uit de oncologie, met de bedoeling om hieruit klinische en biologische informatie te halen (De Smet et al., 2001; Marchal et al., 2004).

Omdat ieder microroosterexperiment de expressie meet van duizenden genen, resulteert dit in enorme datavectoren met duizenden componenten. Voor de analyse hiervan zijn speciale technieken nodig die extreem hoogdimensionale datapunten aankunnen. Noteer dat de vectoren die worden gegenereerd door verschillende microroosterexperimenten kunnen geschikt worden in een expressiematrix (zie Figuur 3.1). In deze matrix bevatten de kolommen alle expressieniveaus van een specifiek experiment en de rijen de expressieniveaus van een zeker gen (gemeten in de verschillende experimenten). De rijen van de expressiematrix worden verder ook genexpressieprofielen genoemd. Afhankelijk van de toepassing kunnen zowel de kolommen als de rijen van deze matrix beschouwd worden als datapunten. In het eerste geval worden de expressieniveaus van de verschillende genen dan beschouwd als de variabelen en in het tweede geval is dit zo voor de experimenten. In dit hoofdstuk echter, beschouwen we in de meeste gevallen de microroosterexperimenten of de kolommen van de expressiematrix (elk geassocieerd aan een patiënt of tumorstaal) als de datapunten. Clusteranalyse van genexpressieprofielen is hierop de enige uitzondering. In dit hoofdstuk beschouwen we verder ook verzamelingen van microroosterexperimenten die tumorcellen bestuderen die afkomstig zijn van verschillende klassen (bijvoorbeeld experimenten afkomstig van patiënten met een verschillende histopathologische diagnose, een verschillende prognose, een verschillend antwoord op therapie).

In hetgeen volgt, bespreken we eerst enkele stappen die nodig zijn ter voorbereiding van de microroostergegevens voor verdere analyse. Hierna onderzoeken we de drie elementen van ons algemeen kader voor gegevensontginning toegepast op dit datatype: selectie van kenmerken, clustering en classificatie. Een grondige studie van twee delen van dit algemeen kader zal ondernomen worden in Hoofdstuk 4, 5 en 6 (clustering van genexpressieprofielen en univariate analyse). Om de hier beschreven methodologie te illustreren hebben we ondermeer gebruik gemaakt van twee verzamelingen van microroostergegevens die publiek beschikbaar zijn op het internet (data van Golub et al. (1999) die 72 patiënten (onderverdeeld in een trainingsset van 38 patiënten en een testset van 34 patiënten) bestudeerden met acute lymfatische (ALL) of myeloïde (AML) leukemie; data van Perou et al. (2000) die patiënten bestudeerden met mammacarcinomen - wij maken hier een onderscheid tussen matig en slecht gedifferentieerde tumoren).

Voorbereiding van de data

Voordat de microroostergegevens kunnen gebruikt worden met de methoden beschreven in de volgende paragrafen, is het mogelijk dat ze eerst nog enkele voorbereidende stappen moeten ondergaan. Hier bespreken we normalisatie, niet-lineaire transformatie en de verwerking van ontbrekende waarden. Twee andere stappen, standaardisatie en filteren, zullen worden besproken in het kader van het clusteren van genexpressieprofielen.

1. Normalisatie: In een experiment met een cDNA-microrooster bestaan er verschillende bronnen van ruis die systematische fouten kunnen veroorzaken (bijvoorbeeld veroorzaakt door verschillen in het groen en rood kanaal). Bij normalisatie is het de bedoeling om deze systematische fouten te berekenen en te verwijderen.

2. Niet-lineaire transformaties: In vele gevallen is het de gewoonte om een niet-lineaire functie, zoals het logaritme, toe te passen op de expressiewaarden. Bij het gebruik van expressieratios (afkomstig van een cDNA-microrooster, waar een test- en referentiestaal worden gebruikt en de uiteindelijke expressiewaarde wordt bekomen door de ratio van de overeenkomstige intensiteiten in het rode en groene kanaal te beschouwen) heeft dit een bijkomend voordeel, vermits deze niet symmetrisch rond 1 zijn verdeeld. Het gebruik van een logaritmische transformatie corrigeert dit.

3. Verwerking van ontbrekende waarden: Microroosterdata bevatten dikwijls ontbrekende waarden. Vele algoritmen die gebruikt worden om deze gegevens te analyseren hebben hier echter problemen mee. Daarom zijn er technieken nodig om deze ontbrekende waarden te vervangen of zijn er algoritmen nodig die hiermee op een meer directe manier kunnen omgaan. In deze context beschrijven we twee technieken: verwerking van ontbrekende waarden zonder vervanging en de methode van de meest nabije buren.

   In sommige gevallen maken algoritmen voor de analyse van microroostergegevens enkel gebruik van de berekening van (Euclidische) afstanden of gemiddelde expressievectoren. Door een kleine wijziging in de definitie van deze afstanden of gemiddelde expressievectoren, is het mogelijk om deze ontbrekende waarden te verwerken zonder ze te vervangen. Meer concreet berekenen we afstanden tussen twee expressievectoren door enkel de componenten te beschouwen die aanwezig zijn in beide vectoren. Bovendien berekenen we de componenten van de gemiddelde expressievector van een verzameling expressievectoren door enkel de overeenkomstige componenten in rekening te brengen in deze verzameling vectoren waarvoor er werkelijk waarden aanwezig zijn.

   In de methode van de meest nabije buren vervangen we de ontbrekende waarden in een genexpressieprofiel door deze te schatten aan de hand van de waarden in de meest gelijkende genexpressieprofielen.

Selectie van kenmerken

Een eerste doelstelling is het verminderen van het aantal gegevens (of waarden) per patiënt of per microroosterexperiment. Enkel de meest essentiële kenmerken die zo informatief mogelijk zijn over een zeker klassenverschil, moeten worden geselecteerd. Dit wordt ook het probleem van de afname van de dimensionaliteit genoemd. Deze afname is meestal noodzakelijk vooraleer gestart kan worden met classificatie of clustering. Bovendien is het mogelijk dat op deze manier de genen worden geïdentificeerd die verantwoordelijk zijn voor het verschil in eigenschappen tussen verschillende soorten tumoren. Wanneer bijvoorbeeld normale cellen en tumorcellen worden vergeleken, is het mogelijk dat er genen worden ontdekt die betrokken zijn in de carcinogenese.

Selectie van kenmerken kan met en zonder supervisie gebeuren. In selectie van kenmerken met supervisie worden de kentekens of klassenlabels van de verschillende patiënten expliciet gebruikt terwijl dit voor de selectie zonder supervisie niet het geval is.

We bespreken nu de twee verschillende manieren om kenmerken te selecteren: univariaat en multivariaat.

1. Univariate selectie: De meest eenvoudige manier is de selectie van individuele genen waarvan de expressie het best gecorreleerd is met een bepaald klassenverschil, waarin men op een bepaald moment geïnteresseerd is. Deze selectie is dus steeds gesuperviseerd. Dit is logisch vermits niet alle genen een expressiepatroon hebben dat informatie bevat over een bepaald klassenverschil zodat deze genen kunnen worden weggelaten. Verschillende technieken zijn mogelijk om de graad van correlatie van een gen met een zeker klassenverschil te kwantificeren. Zoals reeds vermeld kunnen hiervoor hypothesetesten worden gebruikt die echter bemoeilijkt worden door het probleem van meervoudig testen, dat verder zal besproken worden in Hoofdstuk 6. De AUC (oppervlakte onder de Receiver Operating Characteristic curve) is een maat die hiervoor ook kan gebruikt worden. In deze tekst zullen wij ook dikwijls gebruik maken van een score die werd geïntroduceerd door Golub et al., (1999) en die wordt gegeven door:

   $$G(g_i) = \frac{\mu_1(g_i) - \mu_2(g_i)}{\sigma_1(g_i) + \sigma_2(g_i)}, \tag{3}$$

   waar $\mu_1(g_i)$ and $\mu_2(g_i)$ de gemiddelde waarden zijn van het expressieprofiel g_i in respectievelijk klasse 1 en 2 en waarbij $\sigma_1(g_i)$ and $\sigma_2(g_i)$ de geassocieerde standaard deviaties zijn.

2. Multivariate selectie: Door de hoge dimensionaliteit van microroostergegevens is modelselectie niet onmiddellijk bruikbaar voor dit type data, althans niet zonder voorafgaande reductie van de dimensionaliteit met een andere methode.

   Zoals reeds vermeld is voor microroosters een andere methode voor multivariate selectie van de kenmerken echter meer gebruikelijk: Principale Component Analyse (PCA). Zo kunnen voor de trainingsset in de data van Golub et al., de twee principale componenten worden bepaald met de hoogste eigenwaarde en de microroosterexperimenten van de trainings- en testset kunnen hierop dan worden geprojecteerd. Dit resulteert dan in twee kenmerken voor iedere patiënt. Wanneer deze twee kenmerken worden uitgezet in een grafiek (Figuur 3.3), geeft dit een duidelijk zichtbare scheiding tussen patiënten met ALL en AML. Merk op dat in dit geval de selectie van de principale componenten op een niet-gesuperviseerde manier gebeurt aan de hand van de eigenwaarden (er wordt geen gebruik gemaakt van de klassenlabels). Dit kan echter ook op een gesuperviseerde manier gebeuren. Door gebruik te maken van de methodiek voor univariate analyse kan men de principale componenten uitkiezen die overeenkomen met kenmerken die een maximale correlatie vertonen met een zeker gekend klassenverschil. Voor de data van Perou et al. hebben we PCA toegepast met en zonder gesuperviseerde selectie van twee principale componenten (Figuur 3.4). PCA met niet-gesuperviseerde selectie van de principale componenten resulteerde echter in een slechte scheiding tussen patiënten met matig en slecht gedifferentieerde borsttumoren. Hieruit besluiten we dat in dit geval de richtingen met maximale spreiding niet gedomineerd worden door dit verschil in klassen. Gesuperviseerde selectie van de principale componenten (gebaseerd op de Golub-score van Vergelijking 3) resulteerde echter in een veel betere scheiding.

Clustering

Bij het clusteren van microroosterexperimenten beoogt men patiënten te groeperen die een zekere overeenkomst in expressie vertonen. De gevonden groepen kunnen overeenkomen met een bestaand diagnostisch schema (dat meestal gebaseerd is op klinische waarnemingen), maar het behoort tot de mogelijkheden dat door clustering van expressiepatronen nieuwe diagnostische categorieën kunnen gevonden worden die patiënten bevatten waarvan het klinisch gedrag minder variatie vertoont dan in de bestaande schema's. Met clustering is het dus niet de bedoeling om voorspellingen te gaan maken voor individuele patiënten, maar om te bepalen welke de verschillende tumorklassen en hun eigenschappen zijn. In deze tekst hebben we twee verschillende methoden toegepast om de 72 patiënten in de dataset van Golub et al. te clusteren: "K-means" en hiërarchische clustering (Figuren 3.5 en 3.6). Vermits K-means-clustering niet geschikt is om hoogdimensionale data te clusteren, hebben we eerst PCA toegepast met niet-gesuperviseerde selectie van de principale componenten (gesuperviseerde selectie is hier niet gepast vermits de klassenlabels worden verondersteld niet gekend te zijn bij clustering - ze zijn het resultaat van het algoritme zelf). K-means-clustering van de data van Golub et al. resulteerde in twee clusters die bijna perfect overeenkomen met het gekende verschil tussen ALL en AML en is er dus als het ware in geslaagd om de concepten ALL en AML te herontdekken. Hiërarchische clustering resulteerde in een boomstructuur waar de meeste patiënten met AML geconcentreerd zijn in één welbepaalde tak.

In verband met de clustering van microroosterexperimenten kan er echter een kritische opmerking worden gemaakt (Levenstien et al., 2003). In het algemeen is het mogelijk om zeer veel verschillende resultaten met clustering te bekomen (bijvoorbeeld door een verschillende instelling van de parameters van het algoritme of door verschillende algoritmen te gebruiken).

Meestal zal dan het resultaat worden gekozen dat het beste beantwoordt aan een hypothese die men vooraf wou bewijzen (men kiest bijvoorbeeld het clusterresultaat dat een maximaal verschil in overleving van de patiënten in de verschillende clusters vertoont). Het zou echter kunnen dat dit clusterresultaat per toeval werd gegenereerd (en die kans verhoogt indien meerdere clusterresultaten beschikbaar zijn) en niet resulteert in categorieën die een werkelijk biologisch of medisch proces weerspiegelen. In feite gaat het hier opnieuw over een probleem van meervoudig testen. Uit deze observatie concluderen we dat ieder clusterresultaat in de literatuur met de nodige reserve moet worden bekeken en dat de auteurs die dergelijke resultaten publiceren tenminste zouden moeten vermelden hoeveel verschillende verzamelingen van clusters ze in overweging hebben genomen.

Merk op dat ook de rijen van de expressiematrix (genexpressieprofielen) als basis kunnen dienen voor clustering. Deze problematiek zal verder worden besproken in Hoofdstuk 4 en 5.

Classificatie

In een klinische omgeving is het belangrijk dat, aan de hand van metingen met microroosters, voor individuele patiënten voorspellingen kunnen worden gedaan i.v.m. prognose, antwoord op therapie, stadiumbepaling, histopathologische diagnose, … Dit gebeurt aan de hand van wiskundige modellen. In deze tekst worden twee verschillende binaire classificatietechnieken voor microroosterexperimenten bestudeerd: Fisher’s Lineaire Discriminant Analyse (FDA) en LS-SVM. FDA is een lineaire classificatiemethode die geen regularisatie gebruikt en dus moet gecombineerd worden met voorafgaande selectie van kenmerken. LS-SVM-classificatie gebruikt wel regularisatie en kan in principe onmiddellijk worden toegepast op microroostergegevens. Deze technieken werden toegepast op de data van Golub et al. en Perou et al. Bovendien worden de conclusies van een studie besproken die, aan de hand van 9 datasets, deze technieken vergelijkt en die het belang van dimensionaliteitsreductie of regularisatie en het belang van niet-lineariteit bij de classificatie van microroosterexperimenten onderzoekt (Pochet et al., 2004).

Na toepassing van PCA met niet-gesuperviseerde selectie van twee principale componenten op de trainingsset van Golub et al., kunnen we een FDA-model trainen in twee dimensies. Dit model kunnen we vervolgens toepassen op de patiënten van de testset (Figuur 3.7). Dit resulteerde in 3 misclassificaties (91% nauwkeurigheid). De bekomen performantie van het model echter, is in dit geval afhankelijk van de specifieke onderverdeling tussen trainings- en testset en van het aantal gekozen principale componenten. Om een betere beoordeling van de modelperformantie te bekomen, hebben we het trainen en testen van het model herhaald voor 20 randomisaties van de originele trainings- en testset waarbij we bovendien het aantal geselecteerde principale componenten hebben geoptimiseerd met behulp van een "leave-one-out cross-validatie" (LOO-CV) op de trainingsset. Dit resulteerde in een gemiddelde nauwkeurigheid van het model (geëvalueerd op de testset) van 94.40% (met een standaard deviatie van 3.84%). Gesuperviseerde selectie van de principale componenten resulteerde hier niet in een betere performantie. Dezelfde randomisaties werden gebruikt om de performantie van LS-SVM-modellen met een lineaire en RBF-kernel te onderzoeken (zonder voorafgaande dimensionaliteitsreductie). Dit resulteerde in een nauwkeurigheid van 92.86% (σ = 4.12%) en 93.56% (σ = 4.12%), respectievelijk.

Het gebruik van FDA tesamen met de data van Perou et al. werd geëvalueerd met een LOO-CV in combinatie met een gesuperviseerde selectie van de principale componenten in iedere iteratie. Indien er telkens 5 principale componenten worden geselecteerd resulteerde dit in een nauwkeurigheid van 79%. Dit resultaat toont duidelijk aan dat de differentiatiegraad van borstcarcinomen kan worden voorspeld met expressiepatronen.

We sluiten deze paragraaf af met een opsomming van de 3 voornaamste conclusies van onze vergelijkende studie:

1. LS-SVM-modellen met een lineaire en RBF-kernel zonder voorafgaande dimensionaliteitsreductie en die regularisatie toepassen, geven goede resultaten wanneer ze geëvalueerd worden op een testset. Het gebruik van een RBF-kernel resulteert in een evenwaardige of in sommige gevallen een betere performantie in vergelijking met een lineaire kernel.

2. Onze studie bevestigt dat regularisatie belangrijk is wanneer lineaire classificatie wordt ondernomen zonder voorafgaande dimensionaliteitsreductie.

3. Het toepassen van kernel-PCA met RBF-kernel voor FDA geeft minderwaardige resultaten.

Hoofdstuk 4: Clusteranalyse van genexpressieprofielen

In dit hoofdstuk gaan we dieper in op een specifiek element van het algemeen kader voor gegevensontginning toegepast op microroostergegevens: clustering van genexpressieprofielen (rijen van de expressiematrix) (Moreau et al., 2002a; Thijs et al., 2004). In tegenstelling tot het vorige hoofdstuk, beschouwen we hier vooral microroosterdata die metingen bevatten van stalen die genomen zijn op verschillende tijdstippen van een biologisch proces. De genexpressieprofielen zijn in dit geval vectoren waarvan de componenten de expressieniveaus zijn van een specifiek gen genomen op verschillende ogenblikken in de tijd. Clusteranalyse van genexpressieprofielen zoekt groepen van genen waarvan de expressie zich op gelijkaardige wijze gedraagt. Met andere woorden, deze techniek zoekt genexpressieprofielen die voldoende dicht tegen elkaar liggen (volgens een zekere afstandsmaat). Dit is belangrijk omdat gelijkaardige expressie (ook wel co-expressie genoemd) van genen informatie kan opleveren over de biologische functie van die genen. Co-expressie van genen verhoogt bijvoorbeeld de kans dat de transcriptie van die genen op dezelfde manier wordt gereguleerd (co-regulatie), d.w.z. dat ze interageren met dezelfde transcriptiefactor. In hetgeen volgt bespreken we eerst twee stappen die meestal in combinatie met clusteranalyse van genexpressieprofielen worden gebruikt ter voorbereiding van de data. Daarna bespreken we enkele eigenschappen van algoritmen van de eerste en tweede generatie. Als laatste geven we een woordje uitleg over de validatie van de resultaten van clusteralgoritmen.

Voorbereiding van de data

Hier bespreken we twee technieken die, naast de drie stappen die in Hoofdstuk 3 werden besproken, meestal worden uitgevoerd vooraleer men overgaat tot clusteranalyse van genexpressieprofielen.

1. Filteren: Sommige genen waarvan de expressie wordt gemeten op een microrooster zijn niet betrokken in het biologisch proces dat wordt bestudeerd. Hun expressieniveaus vertonen dikwijls weinig variatie over de verschillende experimenten. Wanneer deze genen zouden betrokken worden in de clusteranalyse zouden ze de kwaliteit van het uiteindelijk resultaat in negatieve zin kunnen beïnvloeden. Het zou dus beter zijn om deze genen te verwijderen vooraleer over te gaan tot het clusteren. Dit noemt men filteren. Bij filteren is het de bedoeling om genen die niet beantwoorden aan een zeker criterium (bijvoorbeeld een minimum standaard deviatie) te verwijderen uit de dataset.

2. Standaardisatie: Biologen zijn over het algemeen geïnteresseerd in groepen van genen die hetzelfde relatief gedrag vertonen, d.w.z. op hetzelfde moment stijgende en dalende expressiewaarden vertonen. Het kan echter zijn dat genen met een gelijkaardig relatief gedrag toch een zeer verschillend absoluut gedrag vertonen en bijgevolg een grote Euclidische afstand hebben (bijvoorbeeld als ze een verschillende amplitude hebben of een verschillende basislijn). Om dit te vermijden kan men de genexpressieprofielen standaardiseren. Dit betekent dat ieder genexpressieniveau g^j in een genexpressieprofiel $g(g^1,g^2,...,g^j,...,g^e)$ moet worden vervangen door:

$$\frac{g^j - \mu}{\sigma}, \tag{4}$$

waarbij μ het gemiddelde expressieniveau is van g en σ de standaard deviatie.

Algoritmen van de eerste generatie

Alhoewel er met de clusteralgoritmen van de eerste generatie (zoals visuele inspectie, K-means, hiërarchische clustering en "Self-Organizing Maps" (SOM) die oorspronkelijk ontworpen werden voor andere doeleinden) biologisch relevante resultaten kunnen bekomen worden, bezitten deze technieken een aantal eigenschappen die ze minder geschikt maken voor het clusteren van genexpressiedata. Zo vereisen ze bijvoorbeeld dat de gebruiker een arbitraire waarde voor een zekere parameter definieert (bijvoorbeeld het aantal clusters in K-means) die een belangrijke impact kan hebben op het uiteindelijk resultaat. Deze algoritmen moeten dus gecombineerd worden met procedures die toelaten om de meest geschikte waarde voor deze parameter te vinden, wat allerminst triviaal is. Een ander probleem is dat deze technieken ieder expressieprofiel in een cluster dwingen. Dit geldt ook voor de genen die niet echt betrokken zijn in het biologisch proces dat wordt bestudeerd. Dit kan leiden tot vervuiling van de clusters en een verstoring van hun gemiddeld expressiegedrag. Als laatste kan men vermelden dat de eerste generatie clusteralgoritmen meestal een rekencomplexiteit bezitten die niet toelaat om grote verzamelingen van genexpressieprofielen te clusteren. Vermits de datasets die meestal bestudeerd worden een aanzienlijke aantal genen bevatten, is deze beperking dikwijls onaanvaardbaar.

Algoritmen van de tweede generatie

Recent zijn er een aantal clusteralgoritmen gepubliceerd die specifiek werden ontworpen voor het clusteren van genexpressieprofielen (bijvoorbeeld het "Self-organizing tree" algoritme (SOTA) (Herrero et al., 2001), modelgebaseerd clusteren (Ghosh en Chinnaiyan, 2002; Yeung et al., 2001a) en het kwaliteitsgebaseerd clusteren (Heyer et al., 1999)) en die een aantal van de problemen met de algoritmen van de eerste generatie trachten te verhelpen. De speciale vereisten voor het clusteren van genexpressieprofielen zijn ook de aanleiding geweest voor net ontwikkelen van een eigen clusteralgoritme dat adaptief kwaliteitsgebaseerd clusteren (AQBC) wordt genoemd en in het volgende hoofdstuk grondig wordt besproken. De techniek die werd geïntroduceerd door Heyer et al. (kwaliteitsgebaseerd clusteren) diende hiervoor als vertrekpunt. Hun aanpak resulteert in clusters die zoveel mogelijk genen bevatten die minstens een minimum aan co-expressie vertonen. Dit resulteert in clusters die beter geschikt zouden kunnen zijn voor verdere analyse. Vermits alleen clusters worden gegenereerd waarin het aantal genen boven een zeker minimum komt, worden niet alle genen in een dataset aan een cluster toegewezen. De minimale graad van co-expressie die de genen in een zekere cluster minstens moeten vertonen, wordt echter beschreven als een clusterdiameter (ook wel de kwaliteit van de cluster genoemd) dat door de gebruiker moet worden gespecificeerd en opnieuw redelijk arbitrair is en niet noodzakelijk aangepast aan de lokale structuur van de data. Bovendien is hun algoritme kwadratisch in het aantal expressieprofielen.

Clustervalidatie

Een bioloog is voornamelijk geïnteresseerd in de biologische relevantie van de clusters die gegenereerd worden door clusteralgoritmen en wil deze technieken gebruiken om nieuwe biologische processen te ontdekken. Dit wil zeggen dat er methoden nodig zijn om te testen of bestaande en nieuwe clusteralgoritmen betekenisvolle resultaten opleveren. Het zoeken naar verrijking in bepaalde functionele categorieën (Tavazoie et al., 1999), “Figure of merit” (FOM) (Yeung et al., 2001b), de Rand index (Yeung en Ruzzo, 2001c) en de silhouette (Kaufman en Rousseeuw, 1990) zijn enkele van de methoden die geschikt zijn om resultaten van een clusteringtechniek te valideren. Bovendien wordt de dataset van Cho et al. (1999) (die de celcyclus van gist bestudeert) dikwijls gebruikt om de performantie van clusteralgoritmen te vergelijken.

Een manier om een verzameling clusters te valideren is deze te vergelijken met bestaande schema’s die genen indelen volgens hun biologische functie. Als er clusters gevonden worden die een significant aantal genen bevatten uit eenzelfde functionele klasse kan dit bewijzen dat een clusterresultaat biologisch relevant is. De data van Cho et al. (celcyclus van gist), bijvoorbeeld, bevat genen die functioneel geclassificeerd zijn. Dit is een van de redenen dat deze dataset dikwijls gebruikt wordt voor clustervalidatie. Veronderstel dat een clusteralgoritme een zeker aantal clusters terugvindt in deze dataset. Veronderstel dat een welbepaalde cluster g genen bevat waarvan er k tot dezelfde functionele klasse behoren. Veronderstel bovendien dat deze functionele klasse op zijn beurt f genen en de volledige dataset n genen (in dit geval 6220) bevat. Door gebruik te maken van de cumulatieve hypergeometrische distributie kunnen we de kans of p-waarde berekenen dat dit niveau van verrijking per toeval is opgetreden, d.w.z, wat is de kans om minstens k genen te vinden in deze specifieke cluster van g genen uit een specifieke functionele klasse van f genen en uit een dataset met n genen:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{n-f}{g-i}}{\binom{n}{g}} = \sum_{i=k}^{\min(g,f)} \frac{\binom{f}{i}\binom{n-f}{g-i}}{\binom{n}{g}}. \tag{5}$$

Deze p-waarden kunnen worden berekend in iedere cluster voor iedere functionele categorie. Vermist er in dit specifiek voorbeeld ongeveer 200 functionele klassen bestaan, moet er rekening gehouden worden met het probleem van meervoudig testen wat in dit geval betekent dat alleen clusters weerhouden worden met een p-waarde voor een zekere functionele klasse die kleiner is dan 0.0003.

Hoofdstuk 5: Adaptief kwaliteitsgebaseerd clusteren van genexpressieprofielen

In het vorige hoofdstuk hebben we opgemerkt dat sommige van de klassieke algoritmen die gebruikt worden voor het clusteren van genexpressieprofielen, een aantal eigenschappen bezitten die hen minder geschikt maakt voor deze taak. In dit hoofdstuk stellen we een algoritme voor dat we zelf hebben ontworpen en dat tracht tegemoet te komen aan deze nadelen. We hebben deze aanpak adaptief kwaliteitsgebaseerd clusteren genoemd (AQBC (van “Adaptive quality-based clustering”)) (De Smet et al., 2002). Deze methode is, in essentie, een heuristisch algoritme dat in iedere iteratie twee stappen uitvoert. Een bijzondere eigenschap van dit algoritme is dat het enkel gestandaardiseerde genexpressieprofielen beschouwt. Daaruit volgt dat deze profielen op de doorsnede liggen van een hypervlak en een hypersfeer in de *e*-dimensionale ruimte (waarbij *e* het aantal componenten is van ieder genexpressieprofiel). Hieronder bespreken we de essentiële onderdelen van deze aanpak.

Algoritme

De gebruiker van AQBC moet twee parameters definiëren: *MIN_NR_GENES* en *S*. De eerste parameter geeft het minimum aantal genen in een cluster en de tweede parameter geeft het significantieniveau, d.w.z., de minimum kans dat een gen dat aan de cluster is toegewezen werkelijk tot de cluster behoort. Meestal wordt hiervoor 95% genomen. Merk op dat we in dit algoritme ervoor gekozen hebben om de ontbrekende waarden te verwerken zonder vervanging zoals besproken in Hoofdstuk 3.

Gedurende iedere iteratie voert het algoritme twee stappen uit die hieronder worden besproken:

Stap 1: lokalisatie van een clustercentrum

In de eerste stap wordt een clustercentrum gezocht waarrond een maximaal aantal genexpressieprofielen liggen binnen een zekere voorlopige straal (ook wel de kwaliteit van de cluster genoemd) waarvan de waarde gelijk is aan de straal die gevonden was in Stap 2 (zie verder) van de vorige iteratie. In de eerste iteratie wordt deze waarde geïnitialiseerd aan de hand van een formule die afhankelijk is van *e*. Dit clustercentrum wordt, samengevat, gevonden door het repetitief verplaatsen van het middelpunt van een hypersfeer naar zijn zwaartepunt (d.w.z. naar het gemiddelde van alle genexpressieprofielen die binnen de gegeven straal liggen - zie Figuur 5.1) totdat het middelpunt samenvalt met het zwaartepunt.

Stap 2: herberekening van de straal

In deze stap wordt de voorlopige waarde voor de straal die werd gebruikt voor het lokaliseren van het clustercentrum in Stap 1, herberekend zodanig dat alle genen van de cluster een significante co-expressie vertonen, d.w.z., dat ze een minimum kans (gegeven door *S*) moeten hebben om tot de cluster te behoren (het clustercentrum in deze stap blijft constant en wordt gegeven door het punt dat in Stap 1 werd gevonden). Om deze kans te berekenen hebben we de distributie van de Euclidische afstand *r* van de expressieprofielen tot het clustercentrum gemodelleerd. Dit model wordt gegeven door:

$$p(r) = P_C . p(r \mid C) + P_B . p(r \mid B) \tag{6}$$

waar

$$p(r \mid C) = \frac{S_d}{\left(2\pi\sigma^2\right)^{d/2}} r^{d-1} \exp\left(-\frac{r^2}{2\sigma^2}\right) \tag{7}$$

$$p(r \mid B) = \frac{S_d}{S_{d+1}\left(d+1\right)^{d/2}} r^{d-1} \tag{8}$$

$$P_C + P_B = 1 \tag{9}$$

en

$$d = e - 2 \tag{10}$$

$$S_d = \frac{2\pi^{d/2}}{\Gamma\left(d/2\right)} \tag{11}$$

$$\Gamma(x)=\int_0^\infty u^{x-1}e^{-u}\mathrm{d}u. \tag{12}$$

Het model in Vergelijking 6 bestaat uit twee termen. De eerste term beschrijft de distributie van de profielen die tot de huidige cluster behoren en de tweede term beschrijft de distributie van de profielen die niet tot de cluster behoren (dit worden ook de profielen genoemd die tot de achtergrond behoren). Ieder van de termen wordt ook vermenigvuldigd door zijn geassocieerde a-priori kans (P_C en P_B). De parameters van dit model (σ, P_C en P_B) worden door middel van een EM-algoritme geschat en aangepast aan de structuur van de data (zie Figuur 5.2). De straal van de cluster (R_k) wordt dan als volgt herberekend:

$$P(C \mid R_k)=\frac{P_C.p(R_k \mid C)}{P_C.p(R_k \mid C)+P_B.p(R_k \mid B)}=S. \tag{13}$$

Als deze herberekende straal meer dan 10% verschilt van de voorlopige waarde die werd gebruikt in Stap 1, dan wordt de hele procedure (Stap 1 en Stap 2) opnieuw opgestart maar waarbij de hier (her)berekende waarde voor de straal gebruikt wordt als voorlopige waarde in Stap 1. Als de hier herberekende straal niet meer dan 10% verschilt van de voorlopige straal die werd gebruikt in Stap 1, dan worden die genexpressieprofielen die gedefinieerd worden door deze herberekende straal en het clustercentrum (bepaald in Stap 1) uit de dataset verwijderd. Bovendien wordt deze verzameling van profielen als een geldige cluster beschouwd en getoond aan de gebruiker als het aantal profielen in deze verzameling groter is dan *MIN_NR_GENES*.

Het algoritme eindigt als aan het stopcriterium is voldaan. Dit is onder andere het geval als de verzameling genexpressieprofielen die uit de dataset wordt verwijderd een vast aantal maal en opeenvolgend minder elementen bevat dan *MIN_NR_GENES*. De rekencomplexiteit van het totale algoritme is lineair in *n* (*n* is het aantal genexpressieprofielen in de dataset). Deze methode is geïntegreerd en publiek beschikbaar in een pakket (INCLUSive) voor analyse van microroosterdata dat op het internet kan gevonden worden (Thijs et al., 2002; Coessens et al., 2003).

Resultaten

AQBC werd getest op een aantal datasets waaronder de data van Cho et al. (celcyclus in gist) die reeds werd vermeld in het vorige hoofdstuk. Na het filteren van de 3000 genen met de hoogste waarde voor σ / μ (voor standaardisatie) hebben we AQBC toegepast met S = 0.95 en *MIN_NR_GENES* = 10 (zie Figuur 5.3 voor het resultaat). We hebben het resultaat gevalideerd door te zoeken naar clusters die verrijkt waren in bepaalde functionele categorieën, zoals eveneens besproken in het vorige hoofdstuk (zie Tabel 5.3). We hebben de resulterende p-waarden vergeleken met de p-waarden van de functioneel overeenkomende clusters die gevonden waren door Tavazoie et al. (1999) door het K-means-algoritme toe te passen op dezelfde data set. De drie belangrijkste clusters gevonden door Tavazoie et al. werden ook door AQBC gevonden maar de verrijking lag gevoelig hoger bij AQBC.

In het hierboven beschreven resultaat hebben we hetzelfde criterium gebruikt als Tavazoie et al. om te filteren (gebaseerd op σ / μ) omdat we de vergelijking tussen K-means en AQBC niet wilden beïnvloeden door een verschil in filtering. We hebben echter de data van Cho et al. opnieuw geanalyseerd met AQBC (met dezelfde waarden voor de parameters) maar waarbij we de 3000 genen hebben geselecteerd met de hoogste standaard deviatie σ. We hebben de resulterende clusters gevalideerd en kwamen tot het besluit dat verschillende onder hen waren verrijkt in functionele categorieën van het hoogste niveau (zie Tabel 5.4). Bovendien waren we in staat om de rol van iedere cluster in de celcyclus van gist te bepalen en deze rol te correleren met het gemiddelde expressieprofiel in iedere cluster. We hebben ook verschillende proteïnecomplexen gevonden waarvan bijna alle leden tot dezelfde cluster behoorden.

We hebben AQBC ook getest op een dataset die de ontwikkeling van het centraal zenuwstelsel in de rat bestudeert, op een dataset die bestaat uit expressiepatronen in verschillende weefsels bij muizen en op een kunstmatige dataset. De resultaten worden in deze samenvatting niet verder besproken.

Conclusie

In tegenstelling met de klassieke clusteralgoritmen, bezit AQBC enkele eigenschappen die het meer geschikt maken voor het clusteren van genexpressieprofielen:

1. Het kent niet alle expressieprofielen aan een cluster toe maar enkel diegenen die een significante co-expressie met de andere profielen van de cluster vertonen (significantieniveau wordt gegeven door S). Dit wil zeggen dat de clusters die resulteren uit deze methode mogelijks een beter vertrekpunt zijn voor verdere analyses.
2. De belangrijkste parameter die door de gebruiker moet worden gedefinieerd is S. De waarde die hiervoor moet gekozen worden heeft een specifieke statistische betekenis en is daardoor minder arbitrair en kan onafhankelijk van de dataset bepaald worden. Bovendien bestaat er een waarde (95%) voor deze parameter die in de meeste gevallen

betekenisvolle resultaten geeft. Het is dus meestal niet nodig om uitgebreid te zoeken naar een geschikte keuze voor deze parameter.

3. AQBC produceert clusters die geen vaste straal hebben en aangepast zijn aan de locale datastructuur.
4. AQBC is een snel algoritme dat lineair is in het aantal genexpressieprofielen.
5. Het algoritme is publiek beschikbaar voor data-analyse.
6. Deze aanpak werd uitgebreid biologisch gevalideerd.

Er zijn echter ook enkele nadelen:

1. Het is een heuristische aanpak waarvan het niet bewezen is dat ze convergeert in alle situaties.
2. Het model beschreven in Vergelijkingen 6-12 geldt enkel onder bepaalde voorwaarden. Dit omvat de noodzaak om gestandaardiseerde genexpressieprofielen te gebruiken. Bovendien veronderstelt dit model dat de Euclidische afstand wordt gebruikt wat wil zeggen dat AQBC niet onmiddellijk uitbreidbaar is voor andere afstandsmaten.

Hoofdstuk 6: Univariate analyse in microroosterdata

In dit hoofdstuk concentreren we ons op univariate analyse in microroosterdata en het probleem van meervoudig testen (De Smet et al., 2004). Om de genen in een dataset te ordenen volgens hun correlatie met een zeker klassenverschil (zie ook Hoofdstuk 3) - of anders gezegd, volgens hun graad van differentiële expressie - worden dikwijls hypothesetesten gebruikt die resulteren in een p-waarde voor ieder gen. Vervolgens wordt een arbitrair significantieniveau α gekozen. De genen met een kleinere p-waarde dan α worden dan verklaard differentiële expressie te hebben (of een positieve uitslag van de test te hebben) en de genen met een p-waarde kleiner dan α worden verklaard geen differentiële expressie te hebben (negatieve uitslag van de test). De genen waarvan de uitslag positief is worden dan geselecteerd om verder te worden geanalyseerd of gevalideerd (bij bijvoorbeeld het zoeken naar doelwitten voor geneesmiddelen).

De keuze van α heeft echter enkele gevolgen (zie Tabel 1). Ten eerste kunnen genen wiens expressie niet wordt beïnvloed door het klassenverschil en dus geen werkelijke differentiële expressie hebben, per toeval toch een p-waarde hebben die kleiner is dan α. Daardoor wordt de uitslag van de test voor deze genen verkeerdelijk positief verklaard (vals positieven). Dit noemt men een Type I fout. De vals positieve genen zullen dus geen resultaten opleveren in verdere analyses. Omdat het totaal aantal genen en het aantal genen zonder werkelijke differentiële expressie in microroosterdata extreem hoog kan zijn, kan het aantal vals positieve genen bij gebruikelijke waarden voor α (bijvoorbeeld 5%) behoorlijk hoog zijn. Dit noemt men ook het probleem van meervoudig testen.

Ten tweede kan de keuze van α ook resulteren in een aantal vals negatieve genen. Dit zijn de genen wiens expressie wordt beïnvloed door het klassenverschil (en dus werkelijk differentieel tot expressie komen) maar een p-waarde groter hebben dan α. Dit noemt men een Type II fout die ertoe kan leiden dat potentieel geldige doelwitten niet in overweging worden genomen voor verder onderzoek.

In de literatuur is er recent veel aandacht besteed aan het beheersen van de Type I fout in microroosterdata. Typisch beheerst of controleert men de "Family-Wise Error" (FWE) of de "False Discovery Rate" (FDR - dit is de ratio van het aantal vals positieven op het totaal aantal positieven). De controle van het aantal Type I fouten gaat echter dikwijls ten koste van het aantal Type II fouten dat niet gecontroleerd wordt en aanzienlijk kan zijn.

In dit hoofdstuk stellen we een op Receiver Operating Characteristic (ROC) curven gebaseerde procedure voor die niet tracht om de Type I of II fout te controleren maar die probeert om een optimale balans tussen deze twee fouten te bekomen. Bovendien stelt de oppervlakte onder deze ROC-curve (AUC (van "Area Under the Curve")) ons in staat om de graad van overlapping tussen de p-waarden van de genen met en zonder werkelijke differentiële expressie te kwantificeren. Deze graad van overlapping bepaalt op zijn beurt de relatie tussen de Type I en Type II fout en bepaalt daarom het niveau waarop de optimale balans tussen die twee bereikt wordt. De AUC kan daarom als kwaliteitskenmerk beschouwd worden die de mogelijkheid van microroosterdata beschrijft om te discrimineren tussen genen met en zonder differentiële expressie. Dit kwaliteitskenmerk kan bijvoorbeeld gebruikt worden om verschillende datasets te vergelijken die dezelfde condities bestuderen en om te beslissen welke data het best geschikt zijn voor verdere analyse.

Methodologie

Onze procedure start met het toekennen van een p-waarde aan ieder gen volgens een zekere hypothesetest. In deze tekst gebruiken we hiervoor de "Wilcoxon rank sum test". Vervolgens ordenen we de genen volgens hun p-waarde (in stijgende volgorde).

Hierna berekenen we het totaal aantal genen (verder n_I genoemd) dat werkelijk differentieel tot expressie komt door de grootheid V_i te berekenen voor ieder gen:

$$V_i = \frac{i - p_i.n}{1 - p_i}, \qquad (14)$$

waar i de rangorde (na ordening volgens de p-waarde) en p_i de p-waarde is van een gen ($i = 1,\ldots,n$) en waar n het totaal aantal genen is in de dataset. Wanneer men V_i tegenover i uitzet in een grafiek ziet men dat deze waarde een constant niveau bereikt voor hogere i (zie bijvoorbeeld Figuur 6.2). Men kan bewijzen dat dit constant niveau gelijk is aan n_1. Na de berekening van n_1 is het eenvoudig om n_0 (totaal aantal genen zonder werkelijke differentiële expressie) te berekenen, vermits $n_0 = n - n_1$.

Vervolgens kan men deze geschatte waarden voor n_1 en n_0 gebruiken om het aantal genen te schatten dat terecht positief (TP_i), terecht negatief (TN_i), vals positief (FP_i - van “False Positive”) en vals negatief (FN_i) is bij ieder mogelijk significantieniveau $\alpha = p_i$. Dit wordt gedaan door de formules van Tabel 1 toe te passen. Deze waarden weerspiegelen het verschil tussen werkelijke en verklaarde differentiële expressie.

Tabel 1: Definitie van de terecht en vals positieve genen (TP_i en FP_i) en van de terecht en vals negatieve genen (TN_i en FN_i) en hun aantallen bij een significantieniveau $\alpha = p_i$. Voor ieder van hen is de verwachte waarde gegeven

		Werkelijke differentiële expressie?		
		JA	NEE	
Verklaarde differentiële expressie?	JA ($p \le p_i$)	TP_i $\approx i - p_i.n_0$	FP_i $\approx p_i.n_0$ **Type I fout**	Pos_i $= i$
	NEE ($p > p_i$)	FN_i $\approx n_1 - i + p_i.n_0$ **Type II fout**	TN_i $\approx (1-p_i).n_0$	Neg_i $= n-i$
		n_1	n_0	

Deze waarden kan men gebruiken om de sensitiviteit ($SENS_i = TP_i/TP_i+FN_i$), specificiteit ($SPEC_i = TN_i/TN_i+FP_i$), en FDR ($FDR_i = FP_i/TP_i+FP_i$) te berekenen voor ieder mogelijk significantieniveau. Wanneer we vervolgens de sensitiviteit uitzetten versus 1 - specificiteit krijgen we een ROC-curve.

Deze ROC-curve kan men gebruiken om een significantieniveau α^{opt} te bepalen waarbij een optimale balans tussen het aantal Type I en Type II fouten wordt bereikt. Dit optimum kan op verschillende manieren worden gedefinieerd maar in deze tekst gebruiken we het punt op de ROC-curve dat een raaklijn met richtingscoëfficiënt 1 heeft. Hiervan kan het bewezen worden dat het de som van de kans op een Type I en Type II fout minimaal maakt (of anders gezegd, de som van de sensitiviteit en specificiteit maximaal maakt). De AUC heeft ook een speciale betekenis: ze is gelijk aan de kans dat de p-waarde van een willekeurig gen met werkelijke differentiële expressie kleiner is dan de p-waarde van een willekeurig gen zonder werkelijke differentiële expressie. Zoals reeds gezegd karakteriseert deze waarde dus de graad van overlapping tussen de p-waarden van de genen met en zonder werkelijke differentiële expressie en bepaalt ze dus de balans tussen de Type en Type II fout. Ze kan beschouwd worden als een kwaliteitskenmerk van een bepaalde dataset met betrekking tot de studie van differentiële expressie. Dit kwaliteitskenmerk is onafhankelijk van een significantieniveau.

Resultaten

We hebben de hierboven beschreven procedure toegepast op verschillende voorbeelden waarvan we de resultaten hier kort zullen samenvatten.

We hebben twee datasets vergeleken die expressiepatronen van patiënten bevatten met ALL en AML: de data van Golub et al. (1999) (zie ook Hoofdstuk 3) en de data van Armstrong et al. (2002). Zie ook Figuur 6.8 en Tabel 6.2. De AUC van de dataset van Armstrong et al. (95.13%) is significant ($p < 0.0001$) hoger dan de AUC van de data van Golub et al. (91.39%) wat weerspiegeld wordt in het niveau van de balans tussen de Type I en Type II fout in α^{opt}. De som van de sensitiviteit en specificiteit in α^{opt} ligt dus hoger bij de data van Armstrong et al. dan bij de data van Golub et al. (175.82% versus 166.09%). De relatieve waarde van n_1 (n_1/n) is bij Armstrong et al. (75.49%) ook beduidend groter dan bij Golub et al. (45.63%) Zowel de hogere AUC als de hogere relatieve waarde voor n_1 zijn de oorzaak van een veel gunstiger verloop van de FDR bij Armstrong et al. (ze stijgt veel minder vlug en bereikt een kleinere maximale waarde). Uit deze resultaten kunnen we dus besluiten dat de data van Armstrong et al. geschikter zijn om differentiële expressie tussen ALL en AML te bestuderen dan de data van Golub et al.

Armstrong et al. hebben ook nog een klasse acute leukemie bestudeerd die ze MLL noemden. Dit gaf ons de mogelijkheid om te onderzoeken wat de invloed op de differentiële expressie was van een wijziging in conditie. We hebben de expressiepatronen in de data van Armstrong et al. gebruikt om de differentiële expressie te vergelijken tussen

ALL en AML, tussen ALL en MLL en tussen AML en MLL (zie ook Tabel 6.2). Dit resulteerde in een significant lagere AUC voor de differentiële expressie tussen ALL en MLL (85.98%) in vergelijking met de differentiële expressie tussen ALL en AML (95.13%) en in vergelijking met de differentiële expressie tussen AML en MLL (94.83%). Hetzelfde geldt voor de relatieve waarde van n_1 (de waarden hiervoor waren, respectievelijk: 35.47%, 75.49% en 64.02%). Hieruit kunnen we dus besluiten dat de graad van differentiële expressie tussen ALL en MLL minder is dan de graad van differentiële expressie tussen ALL en AML of tussen AML en MLL. Dit klinkt aannemelijk vermits er reeds geweten was dat de blasten bij MLL een gelijkaardige morfologie hadden als bij ALL.

Als laatste voorbeeld hebben we onze procedure toegepast op twee datasets die expressiepatronen bevatten van patiënten met matig en slecht gedifferentieerde borsttumoren: Perou et al. (2000) (zie ook Hoofdstuk 3) en van ’t Veer et al. (2002) (zie Figuur 6.9). De resultaten gaven opnieuw een duidelijk verschil in kwaliteit met betrekking tot de studie van differentiële expressie tussen de twee beschouwde condities: de data van van ’t Veer et al. was hiervoor beter geschikt dan de data van Perou et al. De AUC en relatieve waarde voor n_1 bij de data van Perou et al. waren respectievelijk: 87.99% en 14%. Bij van ’t Veer et al. waren die: 90.54% en 42%. De AUC bij van ’t Veer was significant hoger ($p = 0.0001$) dan de AUC van Perou et al. Opnieuw hebben zowel het verschil in AUC als het verschil in relatieve waarde voor n_1 hun impact op het verloop van de FDR in beide datasets (gunstiger verloop bij van ’t Veer et al.).

Discussie

Volgens ons kan het verschil in geschiktheid van microroosterdata om differentiële expressie tussen welbepaalde condities te bestuderen (gedetecteerd door een verschil in AUC) te wijten zijn aan het gebruik van een andere of verbeterde microroostertechnologie en experimenteel protocol. Het behoort ook tot de mogelijkheden dat er een verschil bestaat in de specificiteit (voor een zekere pathologie of klassenverschil) van de genen die aanwezig zijn op het microrooster. Eventueel kunnen een verschil in de kwaliteit van de tumorbiopsies en een verschil in de beoordeling van de histopathologie, ook een wijziging in de AUC veroorzaken.

De methode beschreven in dit hoofdstuk zou ook kunnen gebruikt worden om de kwaliteit van verschillende platformen te vergelijken (bijvoorbeeld Affymetrix versus cDNA-microroosters), om het effect van een verschillende voorbereiding van de data te bestuderen op de detectie van differentiële expressie en om het effect van additionele experimenten te beoordelen. Bovendien kan het voor dezelfde dataset gebruikt worden om te beslissen welke hypothesetest het beste resultaat geeft.

Hoofdstuk 7: Conclusies en toekomstig onderzoek

In dit hoofdstuk vatten we de voornaamste conclusies en eigen bijdragen samen. Bovendien stellen we kort enkele specifieke onderzoeksprojecten voor waarin we in de toekomst willen bijdragen. Noteer dat twee van deze projecten het gebruik van proteoomdata inhouden. Als laatste deel van dit hoofdstuk bespreken we enkele algemene toekomstige onderzoeksvragen.

Specifiek toekomstig onderzoek

Ovariale tumoren: studie van het transcriptoom

In dit onderzoek willen we wiskundige modellen construeren die aan de hand van genexpressiedata van sereuze (meest voorkomende histopathologie) ovariumcarcinomen de volgende twee binaire classificatieproblemen trachten op te lossen:

1. Voorspelling of een patiënt met een stadium III (FIGO stadiumbepaling) ovariale tumor zal hervallen binnen 6 maanden na de laatste therapeutische interventie. Omdat de standaard chemotherapie voor ovariumcarcinomen meestal gebaseerd is op platinum, zullen deze modellen in staat zijn om platinumresistentie (chemosensitiviteit van de tumor) te voorspellen. Dit zal de geneesheer in de eerste plaats in staat stellen om de patiënt realistische informatie te geven in verband met zijn prognose, maar het kan ook toelaten om in de toekomst een alternatieve behandeling te ontwikkelen voor stadium III tumoren waarvan voorspeld wordt dat ze niet gevoelig zullen zijn aan de standaard chemotherapie.
2. Voorspelling of een patiënt met een stadium I ovariale tumor zal hervallen na de initiële chirurgie. De patiënten met een stadium I tumor die volgens onze modellen een hoge kans hebben op een recidief, zijn ideale kandidaten die maximaal voordeel zullen halen uit een adjuvante therapie (chemotherapie en/of lymfadenectomie) terwijl patiënten met een stadium I tumor en een lage kans op recidief gespaard zouden kunnen blijven van de bijwerkingen van een zinloze adjuvante behandeling en kunnen gerustgesteld worden dat ze een hoge kans op blijvende genezing hebben.

Endometriose: studie van het transcriptoom en proteoom

Hier plannen we om zowel het transcriptoom en proteoom te bestuderen van weefselstalen die bestaan uit normaal uterien endometrium van vrouwen met en zonder endometriose. Bovendien zullen de vrouwen met matige-ernstige endometriose nog onderverdeeld worden in diegenen met en zonder herval na chirurgie. Hierdoor hopen we wiskundige modellen te kunnen opstellen die de aanwezigheid van endometriose en de kans op herval na chirurgie kunnen voorspellen. In een eerste fase willen we modellen identificeren die gebaseerd zijn op één datatype (d.w.z. op microrooster- of proteoomdata alleen). In een volgende fase hopen we om de voorspellingen te optimaliseren door modellen te construeren die microrooster- en proteoomdata combineren (eventueel nog zelfs aangevuld met klinische gegevens). Bovendien is het de bedoeling om de patronen die bekomen werden uit de studie van het transcriptoom en proteoom met elkaar te vergelijken met een techniek die gebaseerd is op GSVD (Alter et al., 2003).

Cervix- en endometriumcarcinomen: studie van het proteoom

In dit onderzoek willen we serum- en weefselstalen van patiënten onderzoeken met cervix- of endometriumcarcinomen met de bedoeling om prognostische informatie te bekomen. De studie van het proteoom in serum kan eventueel leiden tot de identificatie van merkers die kunnen bepaald worden aan de hand van een eenvoudig te bekomen bloedstaal (in tegenstelling tot het nemen van een biopsie).

Algemene toekomstige onderzoeksvraagstukken

Terwijl er recent verschillende publicaties zijn verschenen die duidelijk het potentieel van microroosters bij het bepalen van het klinische beleid in de oncologie aantonen, zijn er echter nog veel hindernissen die het gebruik van deze technologie in de dagdagelijkse klinische praktijk verhinderen.

Vooraleerst zijn de meeste modellen die geconstrueerd zijn aan de hand van microroosterdata gebaseerd en getest op een beperkt aantal patiënten. Om betrouwbare modellen te bekomen moeten voldoende technische en biologische replica's voorhanden zijn. Bovendien moeten deze modellen prospectief worden gevalideerd in klinische studies met grotere groepen patiënten.

Bovendien is er ook nog het probleem van de standaardisatie. Omdat de experimentele procedure voor het bestuderen van het transcriptoom gevoelig kan variëren van plaats tot plaats, is het mogelijk dat klinische modellen die bekomen zijn in een bepaald centrum niet direct overdraagbaar zijn naar een ander centrum. Gedetailleerde experimentele richtlijnen zijn nodig vooraleer een implementatie in de klinische praktijk mogelijk is. Merk ook op dat het gebruik van een unieke referentie bij cDNA-microroosters een veralgemeende toepassing van de resulterende modellen onmogelijk maakt.

Zoals reeds vermeld, is het mogelijk dat men door de studie van het proteoom meer informatie kan bekomen over het fenotype van een tumorcel.

Bovendien zijn voor de studie van het transcriptoom steeds weefselstalen nodig, wat niet altijd het geval is voor de studie van het proteoom (dat bijvoorbeeld ook in serum kan bepaald worden). Het gebruik van proteoomdata kan hierdoor de volgende stap zijn om hoogdimensionale moleculair biologische data in het klinisch beslissingsproces te integreren.

Specifiek voor de mathematische analyse van hoogdimensionale moleculair biologische gegevens blijven er ook nog een aantal open onderzoeksvragen die het eventueel mogelijk maken om bijkomende informatie te bekomen. Dit omvat ondermeer het combineren van microroosterdata, proteoomdata en eventueel klinische data in hetzelfde model, het gebruik van Independent Component Analyse (ICA), de combinatie van modelselectietechnieken met andere methoden voor de selectie van kenmerken, het gebruik van andere afstandsmaten bij clustering en kernelversies van clusteralgoritmen, het gebruik van GSVD of CCA voor de vergelijking van microrooster- en/of proteoomdata en het gebruik van meta-analyse technieken voor de analyse van data die afkomstig zijn van verschillende bronnen of centra.

Notation

List of symbols

In the following table we alphabetically list and explain the symbols that are used in this text. Some of the symbols can have more than one meaning, which should be clear from the context.

Symbol	**Explanation**
#	Number of elements in a set
$\varnothing$	An empty set
$\{\}$	A set
$\{x \mid \text{Condition}(x)\}$	A set that contains x-values for which the Condition is true
$\cup$	Union of two sets
$\backslash$	Subtraction of two sets
$\in$	Is an element of
$\notin$	Is not an element of
$\neq$	Not equal to
$\approx$	Approximately equal to
*	Missing value
$n!$	$n.(n\text{-}1).(n\text{-}2)...3.2.1$
$\lVert\cdot\rVert_2$	2-norm of a vector
$\sum_{i=1}^{I} x_i$	Sum: $x_1+x_2+x_3+...+x_I$
$\prod_{i=1}^{I} x_i$	Product: $x_1 \,.\, x_2 \,.\, x_3 ... x_I$
$\lvert x\rvert$	Absolute value of x

Symbol	Explanation
$\begin{pmatrix} n \\ m \end{pmatrix}$	Number of combinations of *m* elements chosen from a set of *n* elements $\begin{pmatrix} n \\ m \end{pmatrix} = \frac{n!}{m!(n-m)!}$
$\int$	Integral
$\surd$	Square root
$\forall$	For all elements off
$[x_1\ x_2\ ...\ x_n]$	Row vector with components $x_1, x_2, ..., x_n$
1_N	Row vector of dimension *N* where the components equal 1
a	Number of gene pairs that are placed in the same cluster in two partitions
A	Expression matrix of a set of microarray experiments (*n* x *e* matrix)
A_k	Diagonal matrix whose elements are proportional to the eigenvalues of Σ_k
A_{ROC}	True AUC for an infinite sample
$\hat{A}_{ROC}$	Estimate of the AUC
A^{SORT}	Expression matrix where the gene expression profiles have been sorted (descending order) according to their correlation with g_{mv}
ACCUR_RAD	Internal tuning parameter of AQBC
AP	TVS parameter - Uterine anteroposterior diameter (mm)
α	Rejection level or confidence level
α^{opt}	Optimal rejection level
α^j	Lagrange multiplier (LS-SVM)
b	Model threshold
β_i	*i*th model parameter of a linear model
C	Cluster
C_i	Class *i* or cluster *i*
CC	Presence of a clear cell component in an endometrial tumour (0:not present, 1:present)
CEIL(*x*)	Smallest integer that is equal or larger than x
d	Number of gene pairs that are placed in different clusters in two partitions (in Chapter 5 (AQBC) this also refers to e - 2)
$d(g_i,C_l)$	Distance from a gene expression profile g_i to a cluster C_l.
$d(v_k,v_l)$	Euclidean distance between the expression vectors v_k and v_l
dr	Elementary thickness

Symbol	Explanation
dV	Elementary volume
D	Fixed threshold for the diameter of a cluster
$D(i,j)$	Evaluation of the status of the j^{th} component of the i^{th} expression vector. Equals 1 if this not a missing value and equals 0 otherwise
D_k	Orthogonal matrix of the eigenvectors of Σ_k (this notation is also used for a design variable)
DD	Degree of differentiation of an endometrial tumour (1:good, 2: moderate, 3: poor)
DIV	Internal tuning parameter of AQBC
λ_k	Constant of proportionality of Σ_k
e	Number of microarray experiments in a microarray data set – number of columns of the expression matrix A
e_i	Number of microarray experiments in a data set that belong to class i
e^i	Error variables (LS-SVM)
E(.)	Expected value
EE	TVS parameter - Endometrial Echogenicity (0:homogeneous, 1:heterogeneous)
EL	TVS parameter - Endometrial Lining (0:regular, 1:irregular)
ET	TVS parameter - Endometrial thickness (mm)
EV	TVS parameter - Endometrial volume (ml)
f	Number of genes in a functional category
F	Column vector (s x 1) that contains the s features of the microarray experiment with expression vector m after PCA
FDR_i	False discovery rate at rejection level $\alpha = p_i$
FN_i	Number of false negative genes at rejection level $\alpha = p_i$
FP_i	Number of false positive genes at rejection level $\alpha = p_i$
g	Gene expression profile (1 x e row vector) (g is also used for the number of genes in a cluster)
g_i	i^{th} gene expression profile (1 x e row vector)
g^j	j^{th} component of gene expression profile g
g_i^j	j^{th} component of the i^{th} gene expression profile g_i
g_{mv}	Gene expression profile with missing value
g_{si}	Gene expression profile with i^{th} largest correlation with g_{mv}
$g(x)$	Logit (logistic regression)
$G(g_i)$	Golub score of the i^{th} gene expression profile

Symbol	Explanation
G	Collection of gene expression profiles (this notation is also used for the intensity in the green channel of a cDNA-microarray and for the G-statistic in the likelihood ratio test (logistic regression))
γ	Regularization parameter of a LS-SVM model
$\Gamma(.)$	Gamma function
H	Intersection of a hypersphere and a hyperplane formed by standardizing gene expression profiles (AQBC)
H_L	Linearised version of H (AQBC)
H_0	Null hypothesis
H_1	Alternative hypothesis
I	Number of expression vectors in a data set
J	Number of measurements or components in an expression vector
$J(w)$	Fisher criterion
k	Number of genes in a cluster that belong to a certain functional category
K	Number of clusters
$K(r,s)$	Kernel function
$l(\beta)$	Likelihood function
ln	Logarithm with base e
$L(\beta)$	Log likelihood function
m	Expression vector of a microarray experiment (n x 1 column vector)
m^{Ci}	Mean expression vector of the expression vectors of the microarray experiments belonging to class i
m^j	Expression vector of the j^{th} microarray experiment (n x 1 column vector)
m_x	Sample mean of x
max X	Maximum value of set X
mean(X)	Mean value or average vector of a set X
min X	Minimum value of set X
M_i	Number of missing values in an expression vector v_i
MAXITER	Internal tuning parameter of AQBC
ME	Cluster mean (AQBC)
MI	Degree of myometrial invasion of an endometrial tumour (0:absence of deep invasion, 1:deep invasion)
MIN_NR_GENES	User-defined parameter of AQBC - minimum number of genes in a cluster
MT	TVS parameter - Myometrial thickness (mm)
μ	Mean

Symbol	**Explanation**
μ_i	Mean of the values belonging to class *i* or mean vector of cluster C_i
μ_x	True or population mean of *x*
$\mu(g_i)$	Mean expression level of g_i
n	Number of gene expression profiles in a microarray data set - Number of rows of the expression matrix *A*
n_0	Number of genes in a microarray data set without actual differential expression
n_1	Number of genes in a microarray data set with actual differential expression
n_{1calc}	Calculated value of n_1
n_s	Number of tests performed simultaneously
N	Number of data points in a data set
$\check{N}$	Set of *n* genes in a microarray data set
$\check{N}_0$	Set of n_0 genes without actual differential expression in a microarray data set
$\check{N}_1$	Set of n_1 genes with actual differential expression in a microarray data set
$N(j)$	Number of expression vectors in a set that do not have a missing value for their j^{th} component
N_A	Number of abnormal objects
N_N	Number of normal objects
NF	TVS parameter - Number of fibroids
O_k	Cluster center of cluster C_k (AQBC)
O(.)	Order of computational complexity
p_i	p-value of the i^{th} gene in a microarray data set after sorting the genes according to their p-values
p_E	Significance for entry into the model in model selection
p_R	Significance for removal out of the model in model selection
$p_k(g\|\mu_k,\Sigma_k)$	Multivariate Gaussian model for cluster C_k with mean μ_k and covariance matrix Σ_k
$p(g)$	Mixture model for gene expression profiles
$p(r)$	Probability density estimation for *r* in AQBC
$p(r\|C)$	Distribution of r in the cluster (AQBC)
$p(r\|B)$	Distribution of r in the background (AQBC)
$p(.\|C_i)$	Class conditional density function of class *i*
P	Matrix (*n* x *s*) that contains the *s* selected principal components of the expression matrix *A*
P_i	Set of measurement numbers of the missing values in an expression vector v_i or gene expression profile g_i
P_C	A priori probability of belonging to a cluster (AQBC)

Symbol	Explanation
P_B	A priori probability of belonging to the background (AQBC)
$P(C\|r)$	Posterior probability of belonging to the cluster given r (AQBC)
$P(C_i\|.)$	Posterior probability of class i
PI	CDI parameter - Pulsatility index
PSV	CDI parameter - Peak systolic velocity (cm/sec)
π_k	Prior probability of belonging to cluster C_k
Q_1	Probability that two randomly chosen abnormal objects will both be ranked with greater suspicion than a randomly chosen normal object (ROC curves)
Q_2	Probability that one randomly chosen abnormal object will be ranked with greater suspicion than two randomly chosen normal objects (ROC curves)
r	Euclidean distance of a gene expression profile to its cluster center O_k (AQBC)
$r_{i,j}$	Pearson correlation between g_i and g_j
R	Intensity in the red channel of a cDNA-microarray
RAD	Radius of a sphere (AQBC)
RI	CDI parameter - Resistance index
R_k	Radius or quality of cluster C_k (AQBC)
R_k_PRELIM	Preliminary estimate of the radius of a cluster (AQBC)
s	Number of selected principal components
$s(g_i)$	Silhouette of gene expression profile g_i
s_e	Standard error of $\hat{A}_{ROC}$
s_x	Sample standard deviation of x
sign(.)	Sign function
S	User-defined parameter of AQBC - significance level (in Appendix A this is also used for a finite sample)
S_c	Scoring function
S_d	Surface area of a unit sphere in d dimensions
S_A	Sample subset with abnormal subjects
S_N	Sample subset with normal subjects
S_W	Within-class covariance matrix
SA	Subjective assessment of the degree of invasion of an endometrial tumour (0:stage Ia, 1:Ib, 2:Ic, 3:II or higher)
$SENS_i$	Sensitivity at rejection level $\alpha = p_i$
$SENS_{opt}$	Sensitivity at optimal rejection level α^{opt}
$SPEC_i$	Specificity at rejection level $\alpha = p_i$
$SPEC_{opt}$	Specificity at optimal rejection level α^{opt}
SP	Presence of a serous papillary component in an endometrial tumour (0:not present, 1:present);

Symbol	Explanation
σ	Standard deviation
σ^2	Variance
σ_i	Standard deviation of the values belonging to class i
Σ	Covariance matrix
Σ_k	Covariance matrix of cluster C_k
t	Gene number of a gene that belongs to $\check{N}_0$ and with a p-value that is equal or larger than the p-values of all the genes from $\check{N}_1$
t_x	Test statistic (t-distribution)
T	Threshold
T^{opt}	Optimal threshold
TAMXV	CDI parameter - Time-averaged maximum mean velocity (cm/sec)
TN_i	Number of true negative genes at rejection level $\alpha = p_i$
TP_i	Number of true positive genes at rejection level $\alpha = p_i$
U^T	Transpose of the matrix U
UV	TVS parameter - Uterine volume (ml)
v_{av}	Mean expression vector of a set of expression vectors
v_{av}^j	j^{th} component of the mean expression vector
v	Expression vector
v_i	i^{th} expression vector
v_i^j	j^{th} component of the i^{th} expression vector
$v(g_i)$	Within dissimilarity of gene expression profile g_i
V_i	$\frac{i - p_i.n}{1 - p_i}$.
VC	Number of valid clusters (AQBC)
w	Vector with model parameters
$w(g_i)$	Between dissimilarity of gene expression profile g_i
W	Wilcoxon statistic
x^i	i^{th} data point
x_j^i	j^{th} component of the i^{th} data point
y	Output of a model
y^j	Output of a model for the j^{th} data point
y_A	Model output for an abnormal object
y_N	Model output for a normal object
Y	Outcome variable in logistic regression
Y^i	Outcome variable (0 or 1) for the i^{th} data point in logistic regression
z	z statistic
$\varphi(x)$	Mapping function (LS-SVM)

Acronyms

AFP	Alpha fetoprotein
ALL	Acute lymphoblastic leukemia
AML	Acute myeloid leukemia
ANOVA	Analysis of variance
AQBC	Adaptive quality-based clustering
AUC	Area under the ROC curve
BIC	Bayesian information criterion
CAST	Cluster affinity search technique
CCA	Canonical correlation analysis
CDI	Colour Doppler imaging
cDNA	Complementary DNA
CT	Computer tomography
DNA	Deoxyribonucleic acid
EM	Expectation-maximization
EST	Expressed sequence tag
FDA	Fisher’s linear discriminant analysis
FDR	False discovery rate
FIGO	International federation of gynaecology and obstetrics
FN	False negative
FOM	Figure of merit
FP	False positive
FWE	Family-wise error
GSVD	Generalized singular value decomposition
hCG	Human chorionic gonadotropin
IT	Intratumoral
LDH	Lactate dehydrogenase
LOO-CV	Leave-one-out cross-validation
LS-SVM	Least squares support vector machine
MALDI-TOF	Matrix-assisted laser desorption ionisation time-of-flight
MCLUST	Model-based cluster algorithm
MIPS	Munich information center for protein sequences
MLL	Acute leukemia involving the mixed-lineage leukemia gene
MR	Magnetic resonance
mRNA	Messenger RNA
tRNA	Transfer RNA
NaN	Not a number
NN	Nearest neighbour
OMIM	Online Mendelian inheritance in man
ORF	Open reading frame
PCA	Principal component analysis
PCR	Polymerase chain reaction
RBF	Radial basis function

RNA	Ribonucleic acid
ROC	Receiver Operating Characteristic
RT-PCR	Reverse transcription-coupled PCR
SELDI-TOF	Surface-enhanced laser desorption ionisation time-of-flight
SOM	Self-organizing map
SOTA	Self-organizing tree algorithm
TN	True negative
TNM	Tumour, node, metastases
TP	True positive
TVS	Transvaginal sonography
UA	Uterine artery

Contents

Publication list

Most of the work discussed in this dissertation has been published in one of the following articles, in which we contributed.

International journal

Thijs, G., Moreau, Y., De Smet, F., Mathys, J., Lescot, M., Rombauts, S., Rouze, P., De Moor, B. and Marchal, K. (2002) INCLUSive: INtegrated Clustering, Upstream sequence retrieval and motif Sampling. *Bioinformatics*, **18**, 331-332.

De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B. and Moreau Y. (2002) Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, **18**, 735-746.

Epstein, E., Skoog, L., Isberg, P.E., De Smet, F., De Moor, B., Olofsson, P.A., Gudmundsson, S. and Valentin, L. (2002) An algorithm including results of gray-scale and power Doppler ultrasound examination to predict endometrial malignancy in women with postmenopausal bleeding. *Ultrasound Obstet Gynecol*, **20**, 370-376.

Moreau, Y., De Smet, F., Thijs, G., Marchal, K. and De Moor, B. (2002) Functional bioinformatics of microarray data: from expression to regulation. *Proceedings of the IEEE*, **90**, 1722-1743.

Coessens, B., Thijs, G., Aerts, S., Mathys, J., Moreau, Y., Marchal, K., De Smet, F., Engelen, K., Glenisson P. and De Moor, B. (2003) INCLUSive - A Web Portal and Service Registry for Microarray and Regulatory Sequence Analysis. *Nucleic Acids Res*, **31**, 3468-3470.

Timmerman, D., De Smet, F., De Brabanter, J., Van Holsbeke, C., Jermy, K., Moreau, Y., Bourne, T. and Vergote I. (2003) OC118 : Mathematical models to evaluate ovarian masses - can they beat an expert operator ? *Ultrasound Obstet Gynecol*, **22(S1)**, 33.

De Smet, F., Moreau, Y., Tmmerman, D., Vergote, I. and De Moor, B. (2004) Balancing false positives and false negatives for the detection of differential expression in malignancies. *Br J Cancer*, submitted.

Pochet, N., De Smet, F., Suykens, J. and De Moor, B. (2004) Systematic benchmarking micorarray data classification: assessing the role of nonlinearity and dimensionality reduction. *Bioinformatics*, submitted.

Book chapter

Marchal, K., De Smet, F., Engelen, K. and De Moor, B. (2004) Computational biology and toxicogenomics. In Helma, C. (ed), *Predictive Toxicology*. Marcel Dekker, In press.

Thijs, G., De Smet, F., Moreau, Y., Marchal, K. and De Moor, B. (2004) Gene regulation bioinformatics of microarray data. In Akay, M. (ed), *Genomics and Proteomics Engineering*. Wiley/IEEE press, Accepted.

National journal

De Smet, F., Marchal, K., Timmerman, D., Vergote, I., De Moor, B. and Moreau, Y. (2001) Gebruik van microroosters in de klinische oncologie, *Tijdschr voor Geneeskunde*, **57**, 1225-1236.

International conference

Antal, P., Fannes, G., De Smet, F. and De Moor, B. (2001) Ovarian cancer classification with rejection by Bayesian Belief Networks. In *Proc of the Bayesian Models in Medicine workshop, the European Conference on Artificial Intelligence in Medicine (AIME'01)*, pp. 23-27.

Moreau, Y., Thijs, G., Marchal, K., De Smet, F., Mathys, J., Lescot, M., Rombauts, M., Rouzé, P., and De Moor B. (2002) Integrating quality-based clustering of microarray data with Gibbs smapling for the discovery of regulatory motifs In *Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)*, pp. 75-79.

Chapter 1

Introduction

1.1 Motivation

Cancer is the second leading cause of death after heart disease (Longo, 1998). The classical approach to cancer management has several aspects (Slapak and Kufe, 1998). Firstly, there is diagnosis and staging. An examination of a tumour sample under a microscope (histopathological diagnosis) allows verifying the malignancy, the origin, and degree of differentiation of the tumour. Subsequently, staging has to be performed or the extent of the malignant disease has to be determined. During staging, one aims for example to establish whether the tumour is still localized or has already invaded surrounding tissue, whether the lymph nodes are affected or whether distant metastases are present (e.g., in the lung, liver, bone or brain). Diagnosis and staging subsequently allow determining the most appropriate management strategy or therapy planning, which can be surgery, radiotherapy, chemotherapy or a combination. During therapy planning a trade-off has to be found between the goals of the treatment plan (e.g., curative intent, complete remission, gain in survival, palliation) and the possible side effects or morbidity of therapy (e.g., acute toxicity of chemotherapy, secondary tumours following radiotherapy and chemotherapy, mutilation after surgery). Diagnosis and staging also can give an indication of the prognosis (e.g., prediction of the therapy response, survival, disease-free survival, probability of disease eradication)

This classical approach to cancer management, however, is in many cases empirical and based on knowledge present in the literature (usually derived from clinical studies) and often on the personal experience of the clinician. The present diagnostic schemes often exhibit significant interobserver variability and thus still need a considerable amount of personal expertise and sometimes interpretation from the physician. Moreover, not all information that is clinically relevant (e.g., prognostic information) can be extracted using the data that physicians have access to at this moment. Better and more objective tools that, for example, allow assigning patients to a certain diagnostic category or provide prognostic information would be helpful, especially for non-experts.

1.2 Molecular biology

The fundamental mechanisms underlying carcinogenesis on a molecular biological level are in many cases still elusive and not taken into account to make the most optimal management decisions. In this context, we will discuss some elementary principles of molecular biology and describe the technologies that will be used to gather molecular biological data in this dissertation.

Genes are nucleic acid sequences (double-stranded DNA) that carry the information that represents a particular protein or polypeptide. This information is stored by a specific sequence of nucleotides (symbolized by A, G, C and T). The genes encode for proteins through the intermediate action of mRNA. Transcription generates a single-stranded mRNA identical in sequence with one of the DNA strands. The transcription process is initiated by the binding of several transcription factors (specific proteins) to regulatory binding sites in the promoter region upstream of the transcribed sequence. The transcription factor proteins bind to each other to form a complex that associates with an enzyme called RNA polymerase. This association enables the binding of RNA polymerase to a specific site in the promoter (see Figure 1.1). Subsequently, this complex catalyses RNA synthesis. It should be noted that the transcription rate can be positively or negatively affected or regulated by the action of the transcription factors. In a later stage, the mRNA is processed, transported out of the nucleus, and translated into a protein (Moreau et al., 2002a; Lewin, 1997).

Cancer is a genetic disease caused by mutations in the genes of a cell. Distinct processes such as contact with carcinogens, viral infections, radiation can induce mutations in the human genome. This can transform a normal cell into a tumour cell, induce its proliferation and finally lead to invasion and metastasis. Mutations leading to cancer can either occur in proto-oncogenes (genes involved in controlled cell proliferation and cell division), in tumour suppressor genes (encoding for inhibitors of cell proliferation), in genes linked with apoptosis (programmed cell death), genes linked with invasion and metastasis, DNA repair, and so on. These mutations can induce changes in or dysregulate the transcription or expression of other genes without mutations, but whose expression levels (amount of transcription or mRNA produced for a specific gene) are directly or indirectly controlled by the mutated genes. This is for example the case when the mutated gene codes for a transcription factor. It will be the

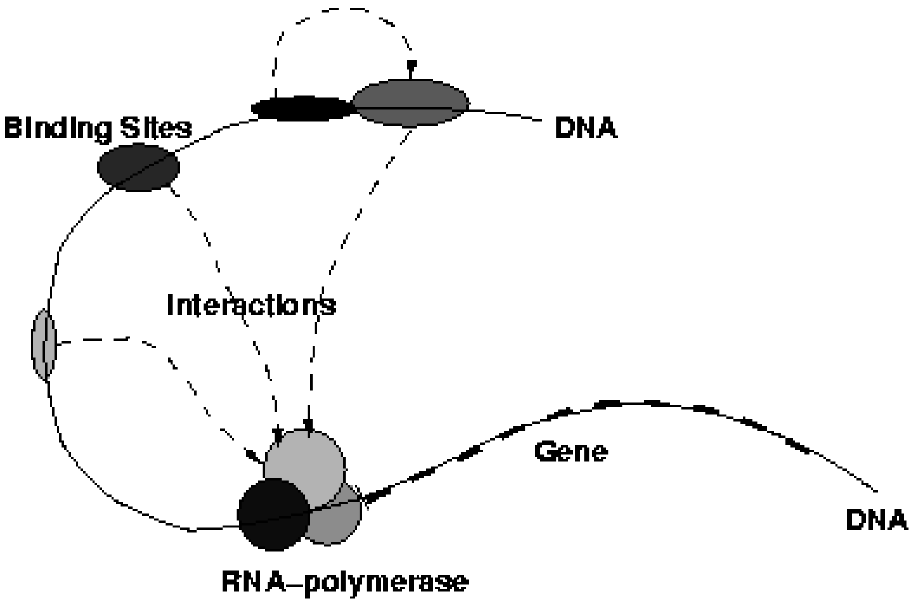


Figure 1.1: Initiation of the transcription process by the association of the complex of transcription factors (gene regulatory proteins), the RNA polymerase, and the promoter region of a gene.

collection of these disturbed expression levels that guide the phenotype of the tumour (Sager, 1997) and represent the fundamental mechanisms that cause malignant process. It can be expected that incorporation of the effects of these mutations on the global expression pattern of the tumour cells into the clinical decision making process could be of major importance. The measurement of these expression patterns will therefore be of great benefit to know, to determine and to understand the real clinical behavior of the tumour cells. Furthermore, studying such data will allow gaining a more profound insight into the processes that lead to and determine the phenotype of malignancies, which could open new perspectives for fundamental cancer research. This could, for example, ultimately lead to the discovery of new drug targets and the development of new drugs that might improve the prognosis of cancer patients.

One of the most promising technologies recently developed to measure expression patterns are microarrays. Microarrays allow to simultaneously measure the expression level of thousands or tens of thousands of genes (also called the transcriptome) in a biological sample. An array constitutes of a reproducible pattern of different DNAs (primarily PCR products or oligonucleotides - also called probes) attached to a solid support. Fluorescently labelled cDNA, prepared from mRNA, is hybridised to the complementary DNA present on the array. Hybridisation intensities are measured by a laser scanner and converted to a quantitative read out. Two basic types of arrays are available: cDNA-microarrays (Duggan, 1999 - see Figure 1.2) and oligonucleotide arrays (Lipshutz, 1999). These will be further discussed in Chapter 3. Since each microarray experiment measures the expression of thousands of genes, this results in a vector with thousands of components (one component for each probe present on the array). When entire microarray experiments need to be analysed, techniques have to be used that can cope with extremely high-dimensional data points. The data produced by microarrays have been the main focus of our research and we will devote the largest part of this dissertation to the methods that can analyse it.

It is possible although, that microarrays do not capture all relevant phenomena in a cell on a molecular level because of posttranscriptional modification and regulation of biologically active molecules. By studying the proteome (collections of all the proteins), it is therefore possible to obtain additional information about the molecular biology of a cell that is not captured by microarrays. The proteome can be examined using recently developed technology based on mass spectrometry that enables to quantify the presence of a large subset of proteins in a sample. We did not yet study this technology or the resulting data during our research, but some specific applications that could be investigated in the future are discussed in the last chapter of this thesis (Chapter 7, Section 7.2.1). In these applications we plan to use the ProteinChip technology developed by Ciphergen Biosystems (based on surface-enhanced laser desorption ionisation time-of-flight (SELDI-TOF) mass spectrometry - see http://www.ciphergen.com and Chapman (2002)) or the ClinProt system of Bruker Daltonics (based on matrix-assisted laser desorption ionisation time-of-flight (MALDI-TOF) mass spectrometry - see http://www.bdal.com/clinprot.html). ProteinChip technology has already been applied to some selected cases in oncology (Kozak et al., 2003; Petricoin et al., 2002a; Petricoin et al., 2002b). Qualitatively, these technologies result in spectra that contain thousands of discrete peak amplitude values each associated with a mass/charge value, which, in its turn, is associated to a (unknown) protein (see Figure 1.3). Therefore, these spectra are characteristic for the proteins or a subset of proteins present in a sample and the output consists of huge data vectors where every component is representative for the amount of an unspecified protein that is present in the sample at hand. The output is thus qualitatively similar to microarray data and can thus possibly be analysed using similar techniques.

In this thesis we will present and study a general data-mining framework, mainly applied to oncology, that intends to extract clinically and biologically meaningful information from microarray data (and proteomic

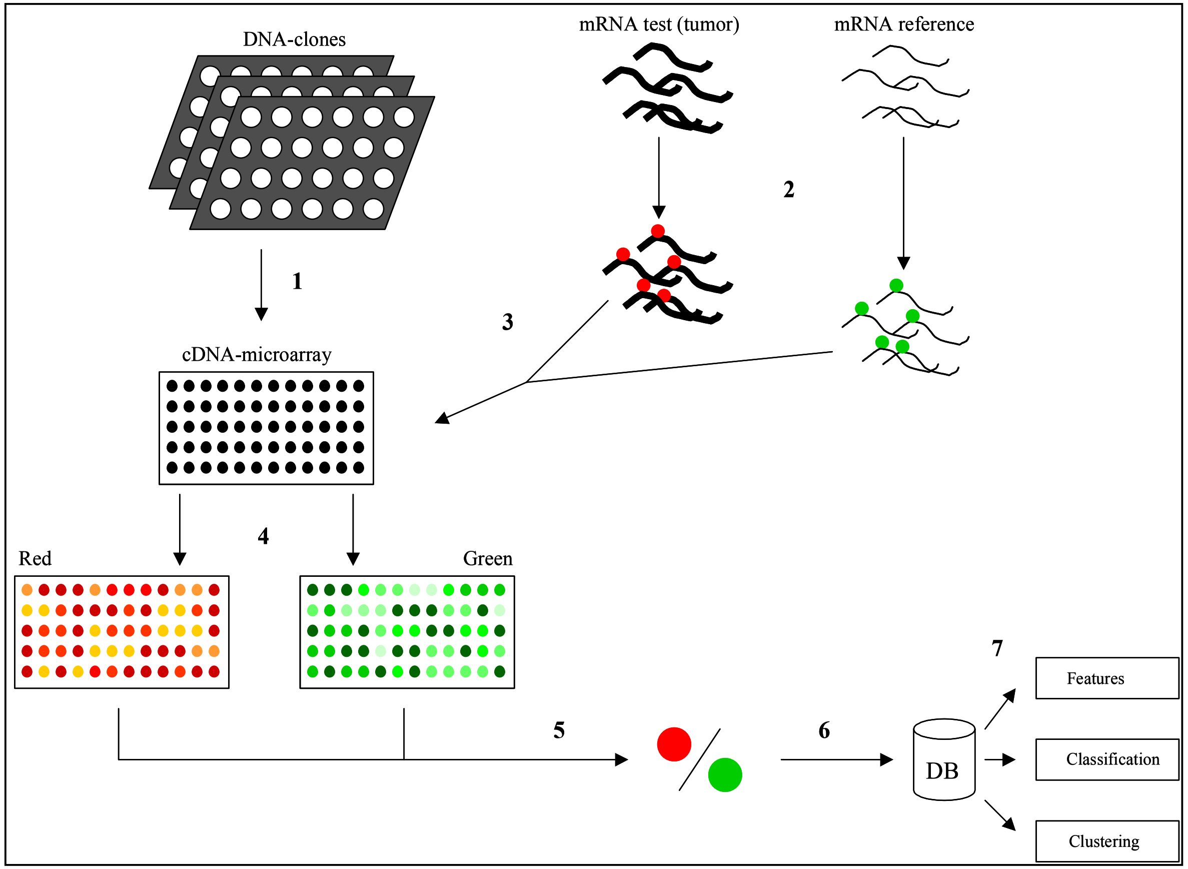


Figure 1.2: Schematic overview of an experiment with a cDNA-microarray. (1) Spotting of the pre-synthesized DNA-probes (derived from the genes to be studied) on the glass slide. These probes are the purified products from PCR-amplification of the associated DNA-clones. (2) Labelling (via reverse transcriptase) of the total mRNA of the test sample (tumour - red) and reference sample (green). (3) Pooling of the two samples and hybridisation (4) Read-out of the red and green intensities separately (measure for the hybridisation by the test and reference sample) in each probe. (5) Calculation of the relative expression levels (intensity in the red channel / intensity in the green channel). (6) Storage of results in a database. (7) Data-mining.

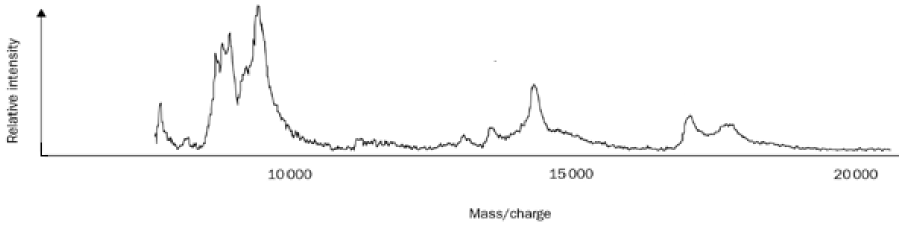


Figure 1.3: Typical mass spectrum obtained with ProteinChip technology (from Petricoin et al., 2002a). This specific spectrum consists of 15.200 peak amplitudes associated with a mass/charge value.

data - future research) and that aims to solve and formulate diagnostic problems more objectively and accurately using clinical data. This framework aims to apply specific algorithms to facilitate diagnosis, prognosis and therapy planning and to obtain a more fundamental insight into the molecular biology of carcinogenesis (see Figure 1.4 for the context of the framework in this thesis).

1.3 Data-mining framework

Although we will use or discuss several data sets in this thesis that contain patients with gynaecologic malignancies (e.g., endometrial, breast or ovarian cancer), we will illustrate our general data-mining framework with a hypothetical data set that contains patients with a malignancy that is exclusively male: testicular cancer. Testicular cancer is the most common type of cancer for men between the ages of 15 and 34 and the incidence rate is reported to be 4/100.000 (Güden et al., 2003). The etiology or cause is unknown but there is a strong association with cryptorchidism (non-descended testicle). In contrast with ovarian cancer (see Section 7.2.1) where most tumours have an epithelial origin, most testicular tumours arise from the primordial germ cells (95%). Testicular germ cell tumours are divided in two major subgroups: seminomas and non-seminomas. Approximately one-third of patients present with early or stage I disease (tumour limited to the testis - Motzer and Bosl, 1998). These patients are usually treated with inguinal orchidectomy (removal of the affected testis) followed by adjuvant therapy resulting in extremely high cure rates. Adjuvant therapy in most centers consists of radiotherapy for seminomas (irradiation of the para-aortic and sometimes ipsilateral iliac lymph nodes, resulting in a relapse rate of less than 5%) and adjuvant chemotherapy (bleomycin, etoposide and cisplatin combination) or retroperitoneal lymph node dissection (a major surgical procedure) for non-seminomas (Jones and Vasey, 2003). There is a problem with adjuvant therapy, however. Without adjuvant therapy after orchidectomy, only 20% of patients with early stage seminomas and only 30% of patients with early stage non-seminoma will have a relapse. Recent publications (Zagars et al., 2004; Huddart et al., 2003) indicate that, beside the acute side-effects, adjuvant therapy could have a profound impact on longevity in this population of predominantly young men, due to an increase in cardiovascular disease or cardiac death and second cancers. This means that for seminomas, for example, 80% of the patients that have received adjuvant radiotherapy will be exposed to this risk without any reason because they would not have had a relapse anyway. At this moment, however, there are no reliable clinical parameters that can distinguish between patients that will and will not have a recurrence without adjuvant therapy, which is the reason that in many centers adjuvant therapy is given to the majority of patients (although recently surveillance (wait-and-see) is also proposed as an option, which means no adjuvant therapy and a rigorous follow-up). We will use this example to illustrate how our data-mining framework could help to select the patients that would benefit from adjuvant therapy and to select the patients for which this would only mean an increase in morbidity and mortality.

Consider a set of patients with stage I seminoma that did not have adjuvant radiotherapy after orchidectomy (patients under surveillance) and consider two groups or classes of patients: without and with relapse (e.g., within five years). The latter group are the patients that would have benefited from adjuvant radiotherapy. Suppose that the class memberships or the class labels are already known for each patient in this data set. Also suppose that we have clinical data available (e.g., values for the tumour markers (β-)hCG (human chorionic gonadotropin), AFP (alpha fetoprotein) and LDH (lactate dehydrogenase), from histopathology (e.g., TNM stage, presence of vascular invasion), from ultrasound examination of the testis, from CT (computer tomography) of the chest and abdomen, from patient and family history, and so on) for each patient and that the primary tumours obtained after orchidectomy were analysed with microarrays and that the resulting expression patterns are available for analysis. It should be noted that the number of clinical parameters is some orders of magnitude lower then the number of gene expression levels available for each patient. We will now discuss how the different elements of our data-mining framework, classification, clustering and feature extraction, could be applied to this data set (see Figure 1.5 for a schematic overview of the application of the elements of this framework).

Figure 1.4: (see opposite page) Context and overview of the general data-mining framework in this thesis. This framework consists of three different components: feature extraction, classification and clustering, each associated with specific methods that are, in most cases, applied through custom MATLAB or SAS scripts. Adaptive quality-based clustering (AQBC - an algorithm specifically designed to cluster gene expression profiles - see Chapter 5) is integrated in an on-line tool for microarray data analysis called INCLUSive. In this thesis we will use, after appropriate preprocessing, the different components of the framework to study clinical and microarray data and discuss how this methodology could be applied for proteomic data in the description of our future research. We will illustrate the analysis of these different data types with concrete data sets that contain information about specific diseases or biological processes. For microarrays the analysis can be done in two different contexts, dependent on the definition of the objects that are studied. In the first context, the objects are entire microarray experiments (which are usually expression patterns associated with specific patients) and in the second context the objects are the expression measurements for a specific gene over the different experiments (called gene expression profiles). Feature selection results in the identification of individual or a set of variables (sometimes called biomarkers for microarray or proteomic data) or in combinations of variables that are as informative as possible about a certain class distinction. For univariate analysis in microarray data we will describe a methodology that can estimate the total number of genes that is and is not actually differentially expressed and introduce a quality label that reflects the appropriateness of a microarray data set to study differential expression (Chapter 6). Classification results in diagnostic models that can predict the diagnostic category of a patient using its expression or proteomic pattern or associated clinical parameters. Finally, clustering results in the identification of the diagnostic categories itself or in groups of genes with similar expression patterns (coexpressed genes) dependent on the context in which the analysis is done for microarray data. ROC = Receiver Operating characteristic curve; PCA = Principal Component Analysis; LS-SVM = Least-Squares Support Vector Machine; LDA = Linear Discriminant Analysis.

1.3.1 Classification

To predict the class membership and hence select the patients that need and do not need adjuvant therapy, one could develop mathematical models (e.g., logistic regression, Fisher’s linear discriminant analysis, Least Squares Support Vector Machines (LS-SVM)) that could anticipate whether patients will have a relapse without radiotherapy. This is called classification. The data set that is described above and for which the class labels are already known, could be used to determine the coefficients of a chosen model structure. This is called model training and the data that is used to train the model is called a training set. This model can subsequently be applied to classify a set of new patients (called the test set) that was not used for training and compare the model predictions with the true outcome

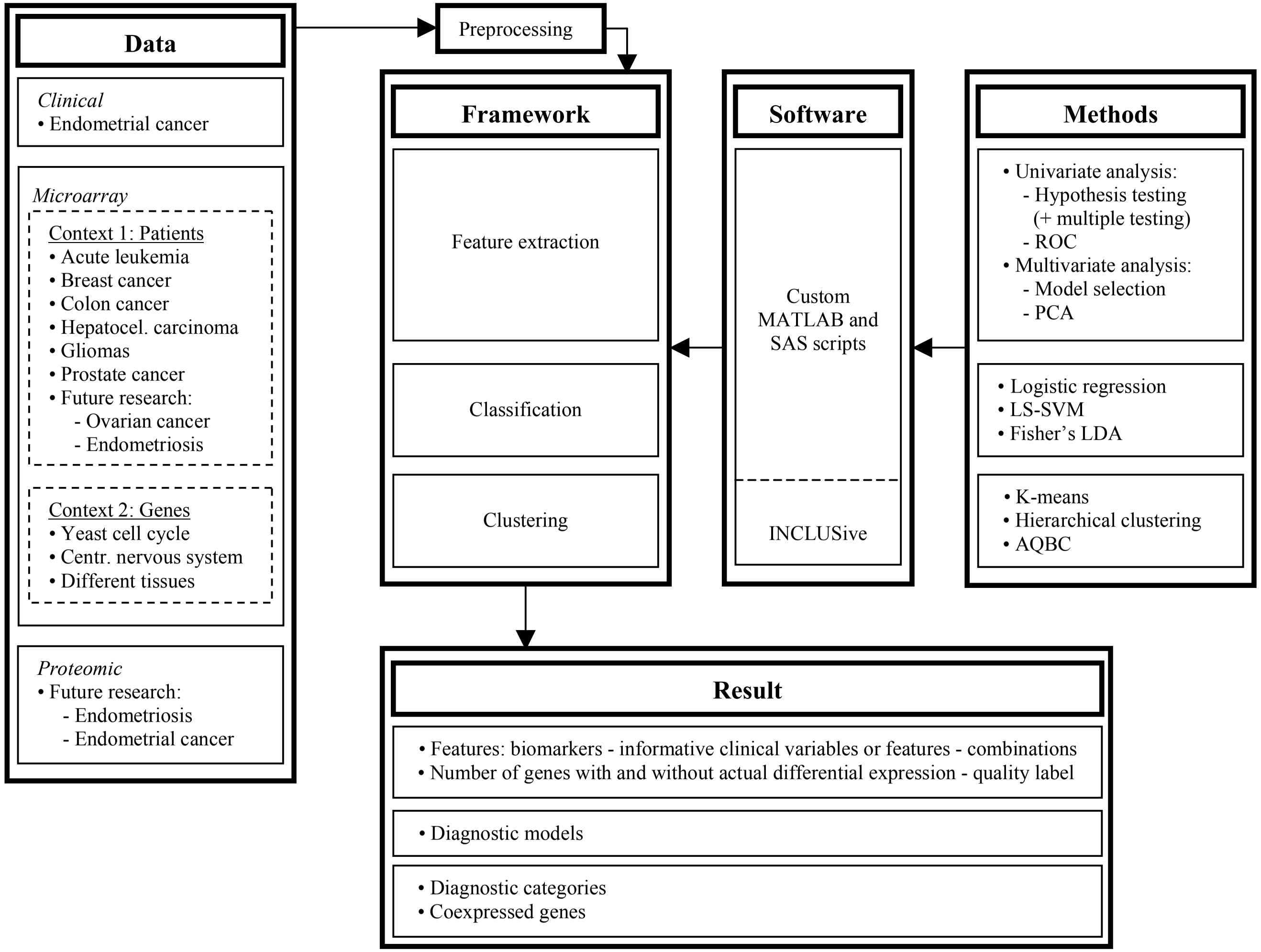

Data
Clinical
• Endometrial cancer
Microarray
Context 1: Patients
• Acute leukemia
• Breast cancer
• Colon cancer
• Hepatocel. carcinoma
• Gliomas
• Prostate cancer
• Future research:
- Ovarian cancer
- Endometriosis
Context 2: Genes
• Yeast cell cycle
• Centr. nervous system
• Different tissues
Proteomic
• Future research:
- Endometriosis
- Endometrial cancer
Preprocessing
Framework
Feature extraction
Classification
Clustering
Software
Custom MATLAB and SAS scripts
INCLUSive
Methods
• Univariate analysis:
- Hypothesis testing (+ multiple testing)
- ROC
• Multivariate analysis:
- Model selection
- PCA
• Logistic regression
• LS-SVM
• Fisher's LDA
• K-means
• Hierarchical clustering
• AQBC
Result
• Features: biomarkers - informative clinical variables or features - combinations
• Number of genes with and without actual differential expression - quality label
• Diagnostic models
• Diagnostic categories
• Coexpressed genes

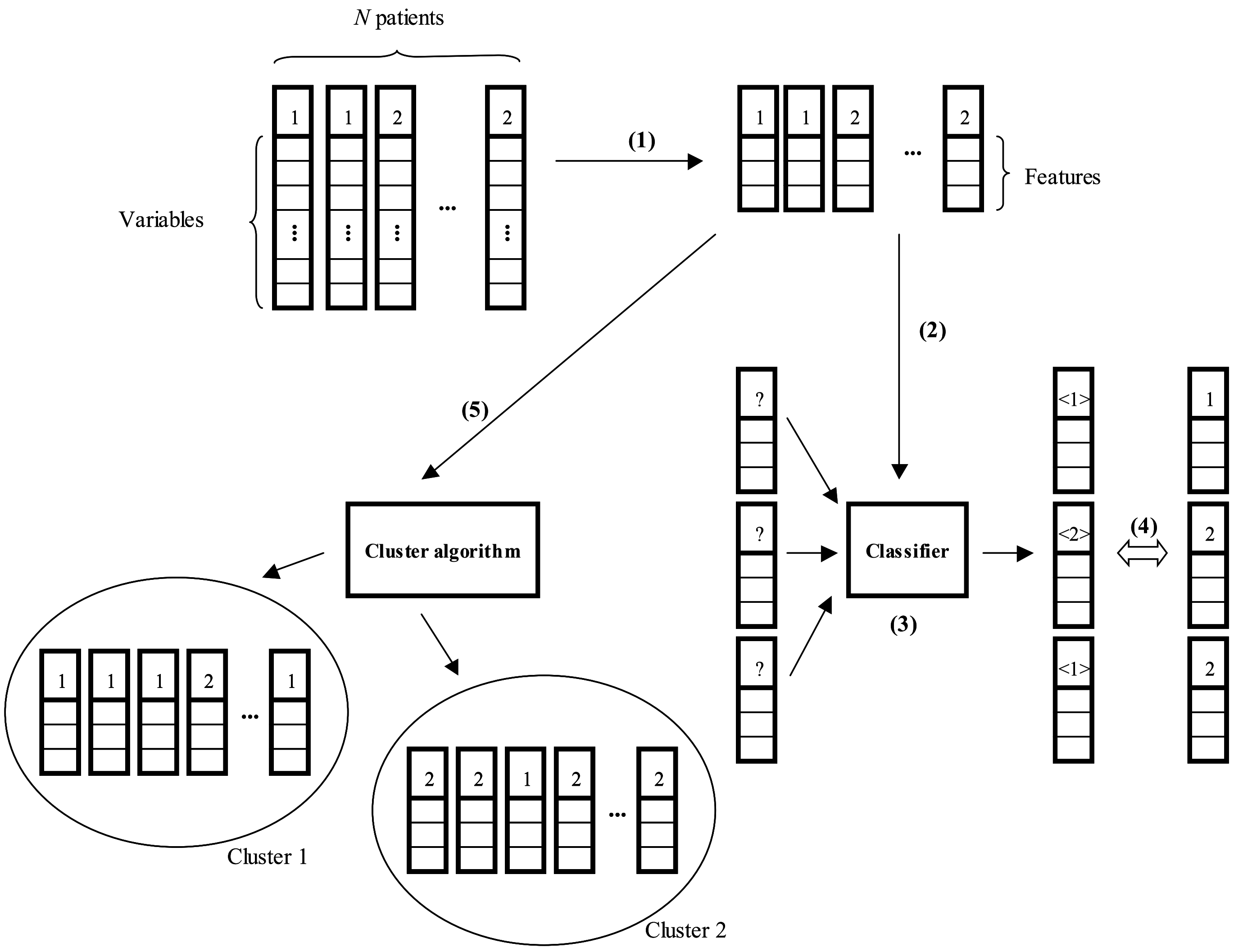
Figure 1.5: (see opposite page) Schematic overview of the different elements of our general data-mining framework applied to the data set containing patients with stage I seminoma under surveillance (i.e., that did not receive adjuvant radiotherapy). Several variables (which can be clinical parameters or gene expression levels measured with microarrays) are available for each patient. In this scheme, the variables are grouped in a column vector. Two classes are considered: patients with (class 1) and without (class 2) tumour recurrence after orchidectomy. The class membership or class label of each patient is indicated in the head of the column vector that represents it. (1) Feature extraction or dimensionality reduction: selection of a limited number of features (obtained after univariate or multivariate analysis) that allow optimal use in subsequent analyses. (2) Model training: the selected features and the class labels of our data or training set are used to determine the coefficients of a certain classifier. (3) Classification: the classifier is subsequently used to predict the class membership of new patients (test set) that were not used to train the model. Classification is based on the same features that were selected in the training set (predicted class labels are indicated by < >). (4) Model validation: comparison of the predicted class label with the true class label of the patients of the test set. (5) Cluster analysis: automatic discovery of groups or clusters of patients (based on the selected features and, since the algorithm has to define the classes by itself, not on the known class labels) with a certain similarity that might represent unknown diagnostic categories and that might contain a significantly different proportion of patients that will and will not have a relapse. Cluster analysis of gene expression profiles is not visualised in this figure.

of these patients. This allows estimating the accuracy or the predictive power of the model on independent data and is called model validation.

In first instance, we could try to construct models that only use clinical data, but since the available clinical parameters probably do not to contain sufficient information to distinguish between patients with and without relapse, the resulting model accuracy can be expected not to be adequate. Therefore it could be helpful to incorporate expression patterns measured with microarrays - that represent the fundamental mechanisms on a molecular biological level determining the phenotype of the tumour - into the mathematical model and the clinical decision making process. As said previously, special techniques need to be applied for the classification of expression patterns due to the high dimensionality of this data.

1.3.2 Clustering

It might be possible that stage I seminomas can be classified or divided in different but yet unknown fundamental entities on a molecular biological level. Tumours belonging to these different entities or diagnostic categories might exhibit different behaviour that is reflected by a different probability of relapse under surveillance. To discover these unknown entities


N patients
Variables
1
1
2
...
2
(1)
1
1
2
...
2
Features
(2)
(5)
?
?
?
Classifier
(3)
<1>
<2>
<1>
(4)
1
2
2
Cluster algorithm
1
1
1
2
...
1
Cluster 1
2
2
1
2
...
2
Cluster 2

one can apply clustering techniques to the expression patterns from the patients of the data set that is described above.

With cluster analysis or class discovery it is possible to automatically find different classes or clusters in a group of microarray experiments or data points without knowing the properties of these classes or the class labels in advance. A cluster, in general, will group data points with a certain degree of similarity, according to a certain distance measure. Ideally, after cluster analysis (with for example the K-means algorithm or hierarchical clustering) groups are formed in such a way that data points in the same cluster are as similar as possible, whereas objects in different clusters are as dissimilar as possible (Kaufman and Rousseeuw, 1990). These clusters might represent groups of patients that, in our example, might contain a significantly different proportion of patients that will and will not have a tumour recurrence under surveillance. These clusters could be the basis of a new diagnostic scheme in which the different categories contain patients with less clinical variability. In clustering, therefore, we do not make predictions for individual patients like in classification, but we try to discover the diagnostic entities or classes themselves.

Cluster analysis of microarray data can also be applied in a different context. Instead of clustering entire microarray experiments, one could try to cluster the expression measurements of the genes over the different experiments (which will be called gene expression profiles further on). For the seminoma data set, for example, one could aim to identify groups of gene that have similar behavior over the different patients (these genes are called coexpressed) and that might have similar roles in the pathway that determines the behaviour of these tumours (e.g., they might be regulated by the same transcription factor).

1.3.3 Feature extraction

Not all variables (clinical parameters or gene expression levels) in our seminoma data set data set are ideal candidates that can be used for further analysis in classification or clustering. In feature extraction we want to identify features (which can be individual variables, sets of variables or combinations of the variables – see further) that allow optimal use in subsequent analyses. Since microarray data consists of thousands of gene expression levels and many classification and clustering algorithms cannot deal with this directly, feature reduction also aims to diminish the dimensionality of the data vectors (dimensionality reduction) here. In this text we consider two different categories of feature extraction: univariate and multivariate.

In univariate feature extraction one aims to select the individual clinical variables or gene expression levels (also called biomarkers) whose value is maximally correlated with, for example, the difference between seminoma patients with and without relapse under surveillance or whose value shows, on the average, maximal difference between these two different classes. This can, for example, be achieved using classical hypothesis testing (Dawson-Saunders and Trapp, 1994). For microarray data the use of hypothesis testing is complicated by the problem of multiple testing to which we will devote Chapter 6 in this thesis.

However, a set of variables that, by themselves, are correlated with a certain class distinction can behave similarly and do not contain, as a whole, more information about the class distinction under consideration than one single variable (this set of variables could be called mutually dependent - e.g., this could be the case for coexpressed genes). A mathematical model that uses this set of variables to predict class membership of the patients could not be expected to perform significantly better than a model that uses only one variable or a fraction of the variables from this set. On the other hand, variables can exist that, on their own, do not contain sufficient information to construct a reliable model but can, when combined, result in a model that performs better. In model selection techniques, variables are selected that have a statistically significant contribution in a certain model. This is usually achieved by an iterative process where variables are sequentially added to or removed from a model (Hosmer and Lemeshow, 1989). Model selection techniques in the context of standard logistic regression are further discussed in Appendix A. The entity selected by this technique is a limited set of variables that in combination, results in an adequate model performance. Model selection is therefore considered to be a method for multivariate feature extraction. However, to prevent overfitting, one needs 6 to 10 patients for each variable that is considered for inclusion during model selection in for example logistic regression. This means that this method is not directly applicable to do feature extraction in microarray data (at least not without reducing the number of features first using some other feature selection technique) due to the extremely large number of genes expression levels that are available compared to the number of patients. In this thesis we will only use model selection in the context of clinical data analysis.

Another technique for multivariate feature extraction is the identification of a linear or non-linear function or combination of the different variables that has a desired property. In principal component analysis (Bishop, 1995), for example, one aims to find a linear combination of the variables that have maximal spread over a set of data points. This is the preferred technique for feature extraction in microarray data and will therefore result in linear combinations of gene expression levels.

1.4 Chapter-by-chapter overview of own contributions

Our own work can be divided in several topics each associated with a certain chapter. The relation between the different chapters is visualized in Figure 1.6. Throughout this text, the quantifiable results of our research will be accentuated by footnotes where appropriate. A list of our publications can also be found in the beginning of this text.

Chapter 2: Clinical data analysis: Prediction of the depth of invasion in endometrial cancer

In this chapter, we analyse a data set, that contains clinical parameters from patients with endometrial cancer (see Appendix B, Section B.1.1), according to some elements of the scheme set out by our general data-mining framework. By univariate and multivariate analysis (model selection - stepwise logistic regression analysis) we investigate which variables contribute in predicting the degree of myometrial invasion in endometrial cancer. Based on this, we construct, compare and validate a logistic regression model and LS-SVM models with a linear and RBF kernel that aim to help the physician in distinguishing between tumours with and without deep myometrial invasion. Although this is not discussed, we also applied some of the techniques presented in this chapter for another study (Epstein et al., 2002).

Chapter 3: Microarray data analysis

This chapter deals with our application of the general data-mining framework to microarray experiments in oncology as illustrated with the data from Golub et al. (1999) (De Smet et al., 2001; Marchal et al., 2004) and the data from Perou et al. (2000) (also see Appendix B). First, in the context of preprocessing we describe the strategies that we have used to manage missing values in microarray data. Further on, we demonstrate the use of univariate analysis and refer to Chapter 6 for an in-depth study of this topic and the problems associated to it. We show how principal component analysis can be applied to microarray data and suggest two methods (unsupervised and supervised) to select the principal components. Subsequently, we perform cluster analysis on the microarray experiments from the data from Golub et al., formulate some critical remarks about these techniques and refer to Chapter 4 and 5 for a detailed study of cluster analysis of gene expression profiles. Finally, we demonstrate how Fisher's linear discriminant analysis and LS-SVM models with linear and RBF kernels can be used to classify microarray experiments and compare these techniques in a benchmarking study that uses nine data sets (Pochet et al., 2004).

Chapter 4: Clustering of gene expression profiles

In this chapter we present a general review of cluster analysis of gene expression profiles (Moreau et al., 2002a; Thijs et al., 2004) and describe some algorithmic challenges. We discuss some specific preprocessing techniques and some of the existing first- and second-generation algorithms that are commonly used to cluster gene expression profiles. An inventory of the advantages and especially the disadvantages of these approaches will lead to the development of our own algorithm in Chapter 5. Finally, we will discuss some selected topics dealing with cluster validation.

Chapter 5: Adaptive quality-based clustering of gene expression profiles

In this chapter we develop our own method, called adaptive quality-based clustering (AQBC), that is specifically tailored to cluster gene expression profiles (De Smet et al., 2002). This method is validated on three existing (including the yeast cell cycle data from Cho et al. (1998) - see Appendix B) data sets and an artificial data set. The integration of our method in an on-line tool for automatic multistep analysis of microarray data, called INCLUSive, is also mentioned (Thijs et al., 2002a; Coessens et al., 2003). Finally, we compare our approach with some of the existing methods already mentioned in Chapter 4.

Chapter 6: Univariate analysis in microarray data

In this chapter we elaborate on the problems of univariate analysis and multiple testing in microarray data (De Smet et al., 2004). We present a method that enables to calculate the number of genes that is and is not affected by a certain class difference. Using this result we show how Receiver Operating Characteristic curves can be used to optimally balance the number of false positives (genes not affected by the difference between the classes but declared so) and false negatives (genes that are affected by the difference in classes but not declared so) and can be used to assign a quality measure to a certain microarray data set with respect to its ability to detect differential expression. Among others, we demonstrate how this quality measure can be used for microarray data by calculating this value for the data from Golub et al. and Perou et al. and by comparing this with the corresponding value for other data that study acute leukemia and degree of differentiation in breast cancer, respectively.

Chapter 7: Conclusions and future research

In this chapter we will describe our main accomplishments and devote a section to future research directions on the shorter term and future prospects on the longer term. Concerning future research on the short term, we will present some concrete projects that have already started or are planned, in which the techniques that are described in this thesis could be applied on microarray and proteomic data. They include a project for ovarian cancer management using microarrays, a project that plans to combine transcriptomic and proteomic patterns in the endometrium for the clinical management of endometriosis and a project that deals with the analysis of proteomic patterns for the study of patients with cervical and endometrial malignant tumours.

1.5 Other research

In the past few years, we have also investigated some other research topics that are not discussed in this dissertation. They include our work related to the development of a control system for the optimization of glycemia in critically ill patients[1] and our work related to the use of artificial intelligence methods for the preoperative assessment of ovarian tumours[2].

[1] This research has resulted in a patent appplication where we are co-inventor (see http://l2.espacenet.com/espacenet/viewer?PN=WO03080157&CY=gb &LG=en&DB=EPD).

[2] In this context, we co-authored two papers (Timmerman et al. (2003) and Antal et al. (2001)).

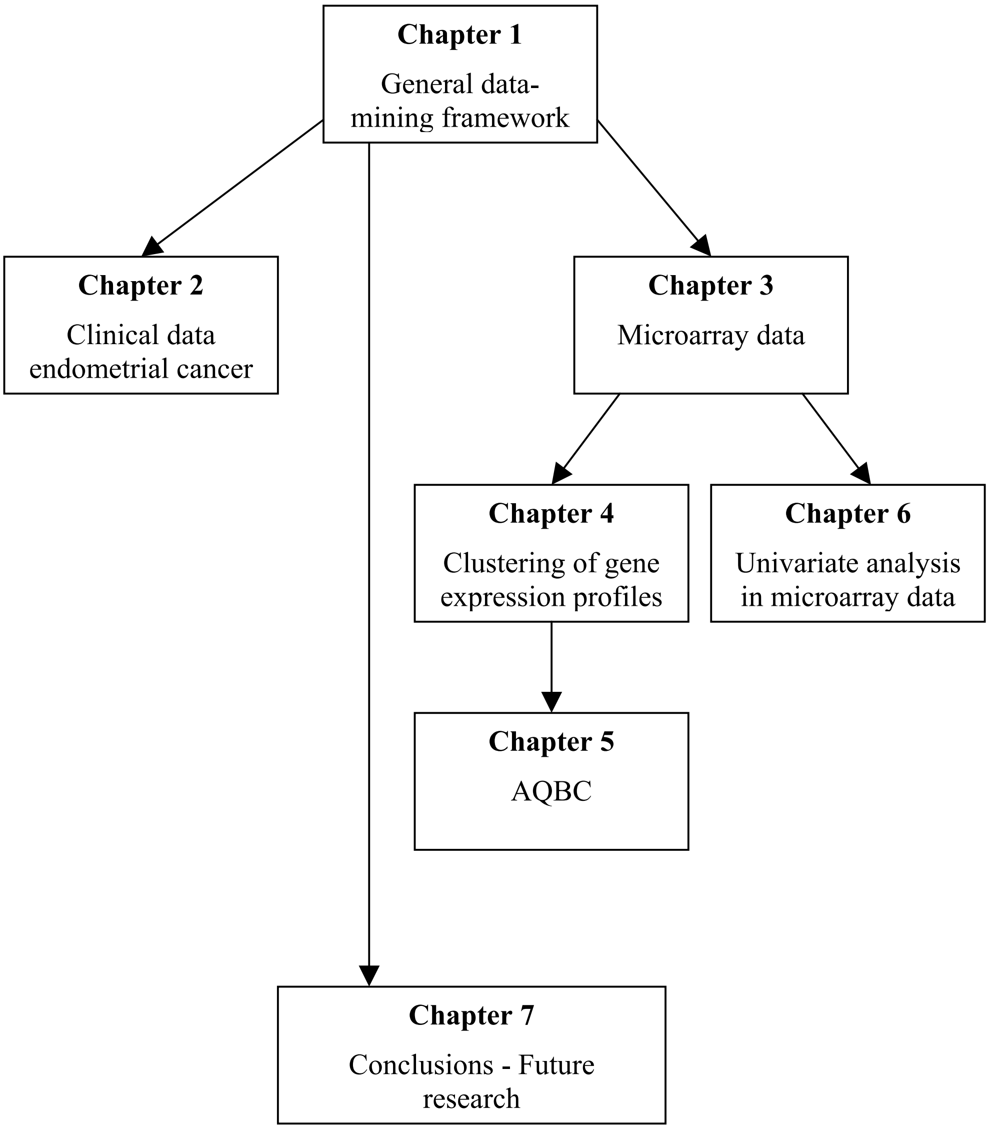

Chapter 1
General data-mining framework
Chapter 2
Clinical data endometrial cancer
Chapter 3
Microarray data
Chapter 4
Clustering of gene expression profiles
Chapter 6
Univariate analysis in microarray data
Chapter 5
AQBC
Chapter 7
Conclusions - Future research

Figure 1.6: Main relationships between the chapters of this thesis. For clarity, the arrows that connect every chapter with Chapter 7 (conclusions) are not indicated. After the description of the general data-mining framework in Chapter 1, we will apply this to clinical and microarray data in Chapter 2 and 3 and discuss its potential use for proteome data in Chapter 7 (future research). While Chapter 3 gives a general overview of the application of the framework to microarray data, a more thorough study of some specific items follow in Chapter 4, 5 and 6. In Chapter 4 we provide a general discussion of cluster analysis of gene expression profiles, which will result in the development of our own algorithm in Chapter 5. In Chapter 6 we present an in-depth study of univariate analysis in microarray data.

Chapter 2

Clinical data analysis: Prediction of the depth of invasion in endometrial cancer

2.1 Introduction

In this chapter we will apply the general data-mining framework to analyse clinical data obtained from patients with endometrial cancer. As already mentioned in Chapter 1, the number of available parameters or variables per patient present in clinical data sets is some orders of magnitude lower when compared to for example microarray data (where thousands of features per patient are available). This means that clinical data can be studied using classical biostatistical techniques, which are often not directly applicable to high dimensional (microarray) data. Note that in this context, we will not discuss methods that aim to cluster these data. Although it is possible that for example new diagnostic categories can be discovered using clustering methods, the clinical value of these techniques can be expected to be rather limited since, in general, the existing diagnostic categories have already been derived and fine-tuned based on clinical information (in fact, the existing diagnostic categories can be regarded as empirically derived clusters). The probability that new and relevant diagnostic schemes emerge by clustering clinical data alone is therefore smaller (in comparison with clustering microarray data that have not yet been incorporated in most of the existing diagnostic categories).

Carcinoma of the endometrium (inner lining of the uterus) is the most common female pelvic malignancy (Young, 1998). Most (75%) tumours are confined to the uterus at diagnosis and are usually curable. However, it is still the 7th leading cause of death from cancer in women. This malignancy occurs mostly in postmenopausal women and in the sixth and seventh decades of life. It is suggested that exposure to estrogens (endogenous or exogenous) may play an important etiologic role. Symptoms often include abnormal vaginal discharge or bleeding. Initial evaluation of these patients includes ultrasound examination (Transvaginal sonography (TVS - grey scale examination of the morphology) with Colour Doppler Imaging (CDI - measurement of the blood flow in the uterine arteries and in the tumour itself) - see Figure 2.1) and an endometrial biopsy.

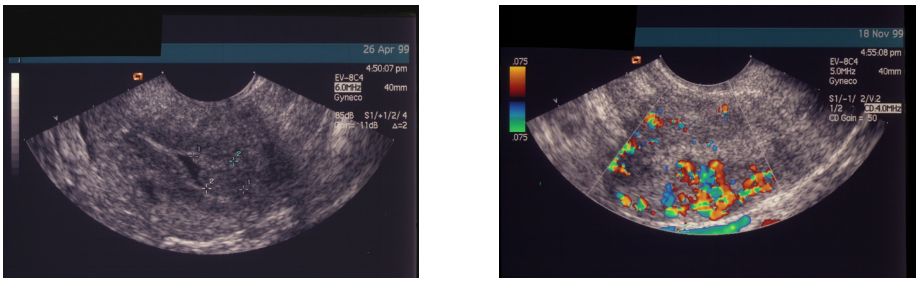


Figure 2.1: Transvaginal sonography (grey scale) and Colour Doppler Imaging of a stage IB endometrial tumour (images supplied by Prof. D. Timmerman).

The transition between FIGO surgical stage Ib and Ic endometrial carcinoma is determined by the degree of myometrial (muscle layer of the uterus) invasion (less or more than 50% (Levine and Hoskins, 2002)) and is an important prognostic factor (Ludwig, 1995) that determines the treatment schedule in many institutions. Accurate preoperative discrimination between patients with stage Ia or Ib disease (group I) and patients with stage Ic or higher (group II - patients with deep myometrial invasion) would allow to identify high-risk patients who might need pelvic and para-aortic lymphadenectomy. This might be important because in many countries patients who need lymphadenectomy are referred to a gynaecological oncologist while patients not needing lymphadenectomy are operated by the general gynaecologist or surgeon.

Several techniques are commonly used to estimate the final histopathological stage or degree of myometrial invasion, but all have specific limitations. Intraoperative gross visual inspection (Franchi et al. (2000) reported an accuracy of 85.3% in predicting the degree of myometrial invasion (403 patients)) or frozen section (Kucera et al. (2000) reported an accuracy of 88% in predicting the myometrial invasion (624 patients)) does not allow preoperative planning of the surgical procedure. MR Imaging (contrast-enhanced) is the most reliable method (in a meta-analysis, Kinkel et al. (1999) reported an area under the Receiver Operating Characteristic (ROC) curve (AUC - see Appendix A, Section A.2) of 91% with respect to the prediction of myometrial invasion) but is costly, has more limited availability, can induce contrast allergies, has a smaller resolution (some distance is present between the individual sections, which may allow small lesions to be missed) and is not appropriate for all patients (e.g., claustrophobia, obesity). TVS and CDI have been well studied but different groups report the use of different morphological or CDI parameters with a considerable variation in the results. Presently, the largest study that investigates the use of TVS and/or CDI to estimate the depth of myometrial invasion was published by Arko et al. (2000) and contains 120 patients. This study reported an accuracy of 73% in predicting myometrial invasion.

2.2 Aim and overview

In the study presented in this chapter, we assessed the value of several parameters in distinguishing between patients from group I or from group II by analysing a data set that contains ultrasound measurements obtained after TVS with CDI and histopathological data from patients with endometrial carcinoma. We constructed models that aim to predict the presence of deep myometrial invasion and that could help the clinician to identify patients that might need more extensive surgery.

In the Materials and Methods section we will describe the data set and its content and discuss the methods that we used to perform feature extraction and classification using clinical data. In the Results section the results of our analysis will be examined and their clinical value evaluated.

2.3 Materials and Methods

Prof. Dr. D. Timmerman from the department of Obstetrics and Gynaecology (University Hospitals Leuven) collected data from 97 consecutive patients (training set) with endometrial carcinoma between September 1994 and February 2000.

All patients underwent preoperative ultrasound examination with TVS and CDI by the same expert (Prof. Timmerman). Histopathology was assessed preoperatively using an endometrial biopsy. The mean age was 65.9 years (range 45-83) and 88 women were postmenopausal. The distribution of the different surgical FIGO stages was as follows: 24 stage Ia, 35 Ib, 12 Ic, 8 II, 13 III and 5 IV. The histopathological subtypes were: 76 endometrioid adenocarcinoma, 3 serous papillary and 18 mixed type (5 with a clear cell and 3 with a serous papillary component). Fifty-four tumours were highly, 18 moderately and 25 poorly differentiated. Tumours with a serous papillary or a clear cell component were considered to be poorly differentiated.

2.3.1 Feature extraction

In this section, we aim to identify the parameters that could be of value to a clinician in distinguishing between patients with and without deep myometrial invasion. More specifically, we want to examine which (if any) individual parameters obtained after TVS with CDI contribute in this distinction. Moreover, we want to identify which ultrasound and histopathological parameters significantly contribute in a standard logistic regression model that predicts the degree of myometrial invasion.

Univariate analysis of the ultrasound parameters and the subjective assessment

Several morphological parameters visualised by grey scale TVS are available for univariate analysis (endometrial (ET) and myometrial (MT) thickness; endometrial (EV) and uterine (UV) volume; ET/AP (uterine anteroposterior diameter); EV/UV; MT/AP; EE (endometrial echogenicity: homogeneous or heterogeneous); EL (endometrial lining: regular or irregular)). CDI parameters included intratumoral peak systolic velocity (PSV), time-averaged maximum mean velocity (TAMXV), resistance index (RI) and pulsatility index (PI) (for an exact definition of these terms, see Timmerman (1997)). Furthermore uterine artery PSV, TAMXV (maximum of the values measured at both the left and right uterine artery, i.e. the worst case), RI and PI (minimum of the values measured at both the left and right uterine artery) were measured. The subjective assessment by the gynaecologist of the depth of myometrial invasion (using a 4-value scoring system - 0: stage Ia; 1: Ib; 2: Ic; 3: II or higher) was also recorded. See Table 2.1 for an example of the possible values for these parameters and their units.

Univariate analysis was performed using the SAS software package (Release 8.01). We performed hypothesis testing (see Appendix A, Section A.1) and specifically used the Wilcoxon rank-sum test (for continuous data) or the Fisher’s exact test (for categorical data) to calculate p-values that reflect if there is a significant difference between patients from group I and group II for a certain variable (Dawson-Saunders and Trapp, 1994). Two-sided tests were used and $p < 0.05$ was used as the level of significance.

In first instance we did not apply a Bonferroni correction (also see Appendix A, Section A.1 for more details) to correct for multiple testing, but in the Results section we will discuss the effect if such a correction would have been applied for this data set. This correction controls the Type I or the family-wise error (FWE - probability of having one or more false positives). This method is, however, very conservative (Perneger, 1998) and can result in an inflation of the Type II error and a decrease in statistical power (which

Table 2.1: Example of the possible values for the different parameters in the training set (first nine patients are shown). Nr=patient number; ET=Endometrial Thickness (mm); MT=Myometrial Thickness (mm); EV=Endometrial Volume (ml); UV=Uterine Volume (mm); AP=uterine anteroposterior diameter (mm); EE=Endometrial Echogenicity (0:homogeneous, 1:heterogeneous); EL=Endometrial Lining (0:regular, 1:irregular); PSV=Peak Systolic Velocity (cm/sec); TAMXV=Time-Averaged Maximum Mean Velocity (cm/sec); RI=Resistance Index; PI=Pulsatility Index; IT=intratumoral; UA=Uterine Artery; SA=Subjective Assessment (0:stage Ia, 1:Ib, 2:Ic, 3:II or higher); DD=degree of differentiation (1:good, 2:moderate, 3:poor); NF=number of fibroids; CC=presence of a clear cell component (0:not present, 1:present); SP=presence of a serous papillary component (0:not present, 1:present); MI=degree of myometrial invasion (0:absence of deep invasion (group I), 1:deep invasion (group II)); NaN=missing value (Not a Number).

Nr	ET	MT	EV	UV	ET/AP	EV/UV	MT/AP	EE	EL	IT PSV	IT TAMXV	IT RI	IT PI	UA PSV	UA TAMXV	UA RI	UA PI	SA	DD	NF	CC	SP	MI
1	11	5	5.29	96.33	0.24	0.05	0.11	0	0	0.23	0.16	0.55	0.82	0.546	0.124	0.98	3.82	2	3	0	1	0	1
2	16	14	4.61	189.38	0.31	0.02	0.27	0	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0	1	0	0	0	0
3	40	2	33.83	118.98	0.80	0.28	0.04	1	1	0.23	0.19	0.27	0.32	0.265	0.186	0.5	0.71	2	2	0	0	0	1
4	20	5	16.47	93.97	0.43	0.18	0.11	1	1	0.23	0.17	0.41	0.56	0.629	0.312	0.41	0.56	2	3	1	0	0	1
5	4	7	0.46	16.05	0.22	0.03	0.39	0	0	0.07	0.04	0.62	1.04	0.227	0.106	0.75	1.44	0	1	0	0	0	0
6	29	4	16.71	77.98	0.69	0.21	0.10	1	1	0.10	0.07	0.44	0.62	0.31	0.201	0.49	0.68	3	3	0	0	1	1
7	13	5	4.04	40.86	0.42	0.10	0.16	1	1	0.10	0.08	0.5	0.66	0.472	0.136	0.84	2.24	1	3	1	1	0	0
8	17	9	4.27	66.19	0.44	0.06	0.23	1	1	0.08	0.027	0.95	2.84	0.584	0.227	0.89	2.29	0	1	0	0	0	1
9	21	6	9.71	95.95	0.44	0.10	0.13	1	1	0.20	0.11	0.69	1.29	0.627	0.174	0.93	2.99	1	1	0	0	0	1

can be extreme if this method is applied to microarray data - for further information on this, the problem of multiple testing and univariate analysis of microarray data see Chapter 6).

In addition, the ROC curves and the AUC were estimated and compared for the individual variables. On the ROC curves, the optimal cut-off point was defined as the point that obtained the best trade off between sensitivity and specificity (the point that maximalises the sum of the sensitivity and specificity). The resulting sensitivity, specificity and accuracy were also calculated. See Appendix A, Section A.2 for more information about the technical details of ROC curves and the choice of an optimal cut-off point.

Multivariate stepwise logistic regression

With multivariate stepwise logistic regression analysis (using stepwise selection in the LOGISTIC procedure from SAS) we aimed to develop a standard logistic regression model that included variables with a coefficient significantly different from zero (see Section A.3.4 on model selection techniques in Appendix A for more details - also see Hosmer and Lemeshow (1989)). We considered the following variables for inclusion in the model: the ultrasound parameters discussed in the previous section, degree of differentiation, number of fibroids detected during ultrasound examination (NF; range 0-2; this parameter was previously reported to be a potential factor disturbing sonographic prediction (overestimation of invasion) (Weber et al., 1995)), presence of a clear cell component and presence of a serous papillary component (based on Pipelle biopsies). In the model, obtained at the end of the stepwise logistic regression analysis, only variables having a coefficient significantly different from zero (p-value < 0.05 - Wald Chi-Square statistic) were allowed. Note that only 74 of the 97 patients from the training set were used for the stepwise logistic regression analysis because of missing values in some of the considered variables (SAS removes patients with one or more missing values).

To prevent overfitting, ideally, one needs 6 to 10 patients for each variable that is considered for inclusion during stepwise logistic regression. This means that in our case the number of patients is already on the low side.

2.3.2 Classification

The variables selected by now after the stepwise logistic regression analysis in the previous section, were used to fit three models, described below, to the training data. The single valued output of these models can also be analysed and compared using hypothesis tests and ROC[1] curves as described in Section 2.3.1. The ROC-analysis can also be used to construct an optimal cut-off point or threshold for these models. Patients with a model output larger than this cut-off are then predicted to belong to group II and thus have deep myometrial invasion.

Standard logistic regression

We fitted a standard logistic regression model with the LOGISTIC procedure from SAS (also see Appendix A, Section A.3) using the variables selected after multivariate analysis in Section 2.3.1. The class labels for patients from group I were 0 and 1 for patients from group II. The Wald Chi-Square statistic was used to assess the significance of the coefficient of a certain variable in the fitted model.

Unlike the two following model building techniques based on Least Squares Support Vector Machines, standard logistic regression does not use regularization, which makes this method prone to overfitting (i.e., the generalization or its performance on prospective or independent data can be sub optimal).

LS-SVM model with a linear kernel

Using LS-SVMlab version 1.5 (see http://www.esat.kuleuven.ac.be /sista/lssvmlab/ and Suykens et al. (2002)) we trained a Least Squares Support Vector Machine (LS-SVM) model using a linear kernel (see Appendix A, Section A.4 for a definition of these models). For all LS-SVM models, the class labels for patients from group I were -1 and 1 for patients from group II. We tuned the hyperparameter (only γ in this case) using a linesearch approach (in the tunelssvm function from LS-SVMlab) where the leave-one-out cross-validation performance (LOO-CV) on the training set was optimised. This hyperparameter setting was subsequently used when training the definitive model. Note that it is possible to write a LS-SVM model with a linear kernel, by rearranging the terms, as a simple linear equation in its variables. Also note that, as said above, we will use the optimal cut-off point following from the ROC analysis on the training set, which does not have to be equal to zero (like in the classical definition of a LS-SVM where a sign function is used).

[1] Together with the Department of Obstetrics and Gynaecology, Malmö University Hospital, Lund University, Sweden, we have contributed in a study and the associated publication in Ultrasound in Obstetrics and Gynecology (Epstein et al., 2002) where we also used ROC curves to compare the performance of different models that aim to predict the presence of an endometrial malignancy in women with postmenopausal bleeding using grey scale and power Doppler ultrasound.

Since regularization is performed (γ is finite), the generalization of this technique can be expected to be more optimal than standard logistic regression or other linear classifiers without regularization.

LS-SVM model with an RBF kernel

Using LS-SVMlab we trained a non-linear LS-SVM model using an RBF (Radial Basis Function) kernel. We tuned the hyperparameters (σ and γ in this case) using a gridsearch approach where, again, the LOO-CV performance was optimised.

If non-linear effects are important in the prediction of deep myometrial invasion, using an RBF kernel can be expected to yield better performance in comparison with the use of a linear kernel.

We fitted this LS-SVM model using the variables that significantly contributed in a linear logistic regression model (see Section 2.3.1), which is not necessarily the best selection for a non-linear LS-SVM model. This means that it is possible that this model is still not entirely optimal. Using model selection techniques in combination with LS-SVM models can thus possibly result in a more optimal selection of variables for LS-SVM models. We did not yet test this exhaustively, but using a (self developed) MATLAB script that implements a forward selection technique for LS-SVM models and that selects variables that improve the LOO-CV performance, did not result in models with an improved generalization.

Prospective validation

Due to the possibility of overfitting, applying the ROC analysis on the same collection of patients that was used to fit our models can result in optimistic estimates for the AUCs. Therefore, we prospectively validated our models using independent data from 37 consecutive and new patients that became available *after* the first 97 that were used to derive our models (this is also the main explanation for this specific subdivision between training and test set). The mean age of these patients was 67.1 years and 36 of them were postmenopausal. The distribution of the FIGO stages was: 7 stage Ia, 20 Ib, 7 Ic, 0 II, 2 III and 1 IV. The following histopathological subtypes were present: 30 endometrioid adenocarcinoma and 7 mixed type (5 with a serous papillary and 1 with a clear cell component). Twenty tumours were highly, 8 moderately and 9 poorly differentiated.

Using this data, we constructed the ROC curves and calculated the AUCs of the three models discussed above and compared them with the AUC of the subjective assessment of our expert (Prof. Timmerman). We also evaluated the performance of our models at the optimal cut-off points obtained after the ROC-analysis of the training set.

2.4 Results

The results of the univariate analysis of the ultrasound parameters and the subjective assessment can be inspected in Table 2.2. EV/UV had the largest AUC from all the ultrasound parameters but there was no significant difference with ET, MT, EV, ET/AP and MT/AP and it was still smaller (not significantly) than the AUC of the subjective assessment. The AUCs of the CDI parameters or the blood flow indices were low. Only the uterine artery RI and PI were (borderline) significant at the 5% level. After applying a Bonferroni correction ET, MT, EV, ET/AP, EV/UV and MT/AP would remain statistically significant. The variables that were borderline significant before the Bonferroni correction would no longer be considered as significant after this correction. In this case, we can state that the overall conclusions of the univariate analysis would remain the same and the decrease of statistical power due to the Bonferroni correction is limited here. Note that this would not be the case if the number of tests that was performed simultaneously, was much higher (which is the case for microarray data).

Multivariate stepwise logistic regression selected the degree of differentiation, the number of fibroids, ET and EV as variables that significantly contributed in a standard logistic regression model. None of the CDI parameters was included.

The resulting logistic regression model fitted to the training data is given by (note that due to missing values in the four selected variables, only 94 patients could be used to fit the models):

$$y = \frac{\exp(\beta_0 + \beta_1.\mathrm{DD1} + \beta_2.\mathrm{DD2} + \beta_3.\mathrm{NF} + \beta_4.\mathrm{ET} + \beta_5.\mathrm{EV})}{1 + \exp(\beta_0 + \beta_1.\mathrm{DD1} + \beta_2.\mathrm{DD2} + \beta_3.\mathrm{NF} + \beta_4.\mathrm{ET} + \beta_5.\mathrm{EV})} \qquad (2.1)$$

where DD1 and DD2 equal 1 if, respectively, the tumour is moderately and poorly differentiated and 0 in other cases. The coefficients are: β_0 = -3.70 (95% CI [-5.53, -1.86], p < 0.0001), β_1 = 2.36 ([0.82, 3.91], p = 0.0027), β_2 = 2.42 ([1.00, 3.84], p = 0.0008), β_3 = -2.45 ([-4.23, -0.67], p = 0.0070), β_4 = 0.20 ([0.07, 0.32], p = 0.0021), and β_5 = -0.11 ([-0.19, -0.03], p = 0.0054). The performance of the logistic regression model on the training data is also summarised in Table 2.2.

The resulting LS-SVM model (as previously said, without the sign function, since we used ROC analysis to define the optimal cut-off) with a linear kernel fitted to the training data is, after rearrangement of the terms, given by:

Table 2.2: Univariate analysis of the ultrasound parameters, the subjective assessment, the logistic regression model and the LS-SVM models with a linear and RBF kernel (training set, *N* = 97). ET=Endometrial Thickness; MT=Myometrial Thickness; EV=Endometrial Volume; UV=Uterine Volume; AP=uterine anteroposterior diameter; EE=Endometrial Echogenicity; EL=Endometrial Lining; PSV=Peak Systolic Velocity; TAMXV= Time-Averaged Maximum Mean Velocity; RI=Resistance Index; PI=Pulsatility Index. The optimal cut-off point was defined as the point that obtained the best trade-off between sensitivity and specificity. See the Results section for a discussion.

		Range	AUC [95% CI]	Optimal cut -off value	Sensitivity (%)	Specificity (%)	Accuracy (%)	Mean or proportion in group I	Mean or proportion in group II	p-value
ET (mm)		2-65	0.76 [0.66, 0.86]	14	80.56	64.41	70.53	14.64	24.88	<0.0001
MT (mm)		2-18	0.71 [0.59, 0.82]	8	74.29	61.02	65.95	8.80	6.41	0.001
EV (ml)		0.002-84.21	0.76 [0.66, 0.86]	4.93	71.43	69.49	70.21	8.17	17.77	<0.0001
UV (ml)		16.05-1074.99	0.61 [0.49, 0.72]	88.62	57.89	69.49	64.95	91.32	146.82	0.08
ET/AP		0.07-1.48	0.75 [0.65, 0.86]	0.429	72.22	71.19	71.58	0.37	0.54	<0.0001
EV/UV		<0.0001-0.75	0.78 [0.68, 0.87]	0.085	68.57	79.66	75.53	0.07	0.15	<0.0001
MT/AP		0.04-0.44	0.75 [0.64, 0.85]	0.174	74.29	74.58	74.47	0.24	0.15	<0.0001
EE (% heterogeneous)		-	0.60 [0.49, 0.72]	-	64.86	55.93	59.38	44.07 %	64.86 %	0.06
EL (% irregular)		-	0.61 [0.50, 0.73]	-	78.38	44.07	57.29	55.93 %	78.38 %	0.03
Intra-tumoral	PSV (cm/sec)	0-0.96	0.61 [0.49, 0.73]	0.13	58.82	64	61.9	0.14	0.21	0.09
	TAMXV (cm/sec)	0-0.77	0.61 [0.49, 0.73]	0.062	82.35	46	60.71	0.09	0.14	0.09
	RI	0.05-1	0.62 [0.48, 0.75]	0.5	50	77.55	66.27	0.62	0.54	0.08
	PI	0.23-5.98	0.61 [0.48, 0.74]	0.61	38.24	87.76	67.47	1.37	1.09	0.10
Uterine artery	PSV (cm/sec)	0.09-2.05	0.51 [0.39, 0.65]	0.621	31.43	84.31	62.79	0.49	0.53	0.81
	TAMXV (cm/sec)	0.04-0.75	0.57 [0.45, 0.70]	0.25	37.14	80.39	62.79	0.20	0.24	0.27
	RI	0.41-1.15	0.64 [0.52, 0.76]	0.71	48.57	78.43	66.28	0.78	0.71	0.03
	PI	0.16-5.96	0.64 [0.52, 0.76]	1.29	48.57	78.43	66.28	1.88	1.45	0.04
Subjective assessment		0-3	0.79 [0.69, 0.88]	1	60.53	86.44	76.29	0: 50.85 % 1: 35.59 % 2: 11.86 % 3: 1.69 %	0: 13.16 % 1: 26.32 % 2: 39.47 % 3: 21.05 %	<0.0001
Logistic regression		0-1	0.89 [0.83, 0.96]	0.45	77.14	86.44	82.98	0.21	0.65	<0.0001
LS-SVM with linear kernel		-1.52 − 1.42	0.88 [0.81, 0.95]	-0.31	91.43	72.88	79.79	-0.52	0.20	<0.0001
LS-SVM with RBF kernel		-1.15 − 0.93	0.99 [0.97, 1]	-0.30	97.14	100	98.94	-0.74	0.56	<0.0001

$$y = \beta_0 + \beta_1.\mathrm{DD} + \beta_2.\mathrm{NF} + \beta_3.\mathrm{ET} + \beta_4.\mathrm{EV} \quad (2.2)$$

where DD equals 1, 2 and 3 if the degree of differentiation is highly, moderately and poorly differentiated, respectively. The coefficients are: β_0 = -1.45, β_1 = 0.37, β_2 = -0.38, β_3 = 0.05, and β_4 = -0.03. Note that the LS-SVM model with an RBF kernel cannot be written in a simplified form and is therefore not explicitly stated here. The performance of the LS-SVM models with a linear and RBF kernel on the training data is also described in Table 2.2.

Evaluated on the training set, the logistic regression and the LS-SVM models with a linear and RBF kernel had a larger AUC than the subjective assessment ($p = 0.0595$, $p = 0.1412$ and $p < 0.0001$, respectively).

The results of the prospective validation can be inspected in Table 2.3 and Figure 2.2. From these we can conclude that prospective evaluation on the independent test set resulted in a better AUC for the standard logistic regression model and the LS-SVM model with a linear kernel, and in a significantly better AUC for the LS-SVM model with an RBF kernel when compared with AUC of the subjective assessment ($p = 0.4758$, $p = 0.0790$ and $p = 0.0485$, respectively). As could be expected (see Section 2.3.2), the performance on the test set or level of generalization of the LS-SVM model with a linear kernel was better than the performance of the standard logistic regression model. Evaluation on the training set (Table 2.2) gave the opposite order of performance, although the difference was small. This shows that the level of overfitting for the standard logistic regression model was higher than for the LS-SVM model with a linear kernel. Also note that the LS-SVM model with an RBF kernel had the best overall performance, both on the training as on the independent test set. This is an indication that non-linear effects might play a role in the distinction between patients with and without deep myometrial invasion.

2.5 Conclusions

In this chapter we used a data set containing 97 patients to assess the value of different ultrasound parameters, measured using TVS with CDI, in discriminating between endometrial cancer patients with and without deep myometrial invasion. Moreover, we used this data to construct a standard logistic regression model and LS-SVM models with a linear and RBF kernel that aim to predict the presence of deep myometrial invasion. Finally we validated these models using independent test data containing 37 patients and compared their performance with the subjective assessment of an expert ultrasonographer.

Table 2.3: Prospective validation: performance of the logistic regression model and the LS-SVM models with linear and RBF kernels for the patients of the independent test set (*N* = 37). Comparison with the ultrasound parameter (EV/UV) from Table 2.2 with the best discriminatory potential and the subjective assessment. The optimal cut-off values were taken from Table 2.2 as evaluated on the training set.

	AUC [95% CI]	p-value: comparison with AUC of subj. ass.	Optimal cut -off value from Table 2.2	Sensitivity (%)	Specificity (%)	Accuracy (%)
EV/UV	0.74 [0.55, 0.93]	-	0.085	70.00	75.00	73.53
Subjective assessment	0.74 [0.58, 0.90]	-	1	50.00	77.78	70.27
Logistic regression	0.81 [0.64, 0.97]	0.4758	0.45	60.00	84.00	77.14
LS-SVM with linear kernel	0.90 [0.80, 1]	0.0790	-0.31	90.00	80.00	82.86
LS-SVM with RBF kernel	0.92 [0.83, 1]	0.0485	-0.30	100.00	72.00	80.00

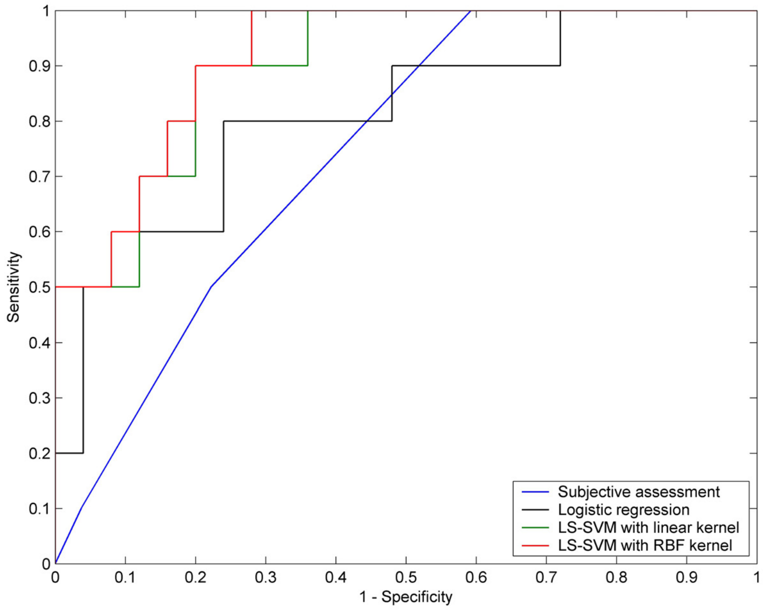


Figure 2.2: Comparison of the ROC curves for the subjective assessment, the logistic regression model, and the LS-SVM models with a linear and RBF kernel for the patients of the independent test set (*N* = 37).

In conclusion our study indicates that CDI does not contribute to the prediction of the degree of myometrial invasion in endometrial cancer. Single morphological parameters are not sufficient in making accurate predictions. Combining the degree of differentiation, the endometrial thickness and volume and the number of fibroids in a standard logistic regression model may deliver predictions more reliable than the subjective impression by an experienced ultrasonographer. Moreover, combining these variables in a LS-SVM model, preferably using an RBF kernel, might even improve these predictions. In our prospective study, which was of limited size though, only a LS-SVM with RBF kernel performed significantly better than the subjective assessment of the expert. These models could represent a simple and inexpensive method that might contribute to the preoperative discrimination between low- and high-risk patients allowing for better preoperative selection of patients with endometrial carcinoma.

However, the models, described in this study, although mathematically interesting and illustrative, are, in our opinion, still far away of being useful or reliable in real clinical practice. First of all, the measurements that were considered in our study all originated from the same expert ultrasonographer. Because differences might exist between different centers or even individual ultrasonographers (who use different ultrasound equipment for example), this means that the models discussed here should at least be tested and, if the performance proves to be unsatisfactory, derived again using multicenter prospective data. Moreover, even the techniques used by the same expert might undergo subtle changes throughout time, causing a drop in model performance when the model is applied on new patients. These comments also apply to the evaluation of the degree of differentiation, which is, at least partially, a subjective measure that can also differ between centers, between pathologists and in time. Secondly, the number of patients available in our training and test set is limited (although this is one of the largest studies available up till now), which contain patients that have been examined in a limited time frame. As already discussed, this might (have) cause(d) problems of overfitting when for example too many variables relative to the number of patients are considered for inclusion in the model during multivariate analysis. Moreover, the characteristics of the population of patients might evolve, causing new patients to be drawn from a different distribution than the one that was used to derive the models. Again, this might cause a drop in model performance when applied to new data. To be clinically useful, these models should, in our opinion, be continuously evaluated and updated (which is often easier said than done since the available data is usually sparse), which we have planned in the near future using new patient data.

Chapter 3

Microarray data analysis

3.1 Introduction

In this chapter we will use the general data-mining framework (feature extraction, clustering and classification), as described in Chapter 1, to analyse microarray data[1] and specifically apply this in oncology. We aim to show how specific methodology can be utilised in order to extract clinical and biological information out of the resulting data and to obtain a more fundamental insight in the molecular biology of carcinogenesis and to facilitate diagnosis, prognosis estimation, prediction of therapy response, and so on. While it is still possible to analyse clinical data manually (as is done daily by medical doctors), this is impossible for microarray data. The number of genes, for which the expression levels are measured in one single microarray experiment, can equal several thousands. This means that each microarray experiment results in a data vector that contains thousands of values. This also means that algorithms are needed that can deal with high dimensional data points and that the methods that were applied in Chapter 2 (methods to control the Type I error in multiple testing problems, model selection techniques, standard logistic regression - LS-SVMs are an exception) to analyse classical clinical data are not straightforward or indicated to be used here, at least not without appropriate dimensionality reduction, regularization or methods that can deal with the problem of multiple testing without severe loss of statistical power.

As already mentioned in Chapter 1, two basic types of microarrays exist and will both be encountered in this chapter:

[1] Some of the topics presented in this chapter have been published in ‘het Tijdschrift voor Geneeskunde’ (De Smet et al., 2001) and have also been included in a book chapter (Marchal et al., 2004).

1. Spotted arrays (Duggan, 1999) or cDNA-microarrays are small glass slides on which pre-synthesized single stranded DNA or double-stranded DNA is spotted. These DNA fragments are usually several hundred base pairs in length and are derived from ESTs (Expressed Sequence Tag) or known coding sequences from the organism studied. Usually each spot represents one single ORF (Open Reading Frame) or gene. A pair of cDNA samples is independently copied from the corresponding mRNA populations (usually derived from a reference and a test sample) with reverse transcriptase and labelled using distinct fluorochromes (green and red). These cDNA samples are subsequently pooled and hybridised to the array. Relative amounts of a particular gene transcript in the two samples are determined by measuring the signal intensities detected for both fluorochromes and calculating the ratios (here, only relative expression levels are usually obtained). A cDNA microarray is therefore a differential technique, which intrinsically normalizes for noise and background. Also see Figure 1.2 for a schematic overview of the procedure that can be followed with spotted arrays.

2. GeneChip® oligonucleotide arrays (Affymetrix, Inc., Santa Clara, CA) (Lipshutz, 1999) are high-density arrays of oligonucleotides synthesized in situ using light-directed chemistry consisting of thousands different oligomer probes (25-mers). Each gene is represented by 15-20 different oligonucleotides, serving as unique sequence-specific detectors. In addition mismatch control oligonucleotides (identical to the perfect match probes except for a single base-pair mismatch) are added. These control probes allow estimation of cross-hybridisation. With this technology, absolute expression levels are obtained (no ratios).

The vectors generated by several microarray experiments can be arranged in an expression matrix where the columns contain the expression levels of a specific experiment and the rows contain the expression levels of a specific gene in the different experiments (see Figure 3.1). The number of rows of the expression matrix always is much higher than the number of columns. Further on in this text, the rows of the expression matrix will also be called gene expression profiles. Dependent on the objective or application, both the columns and the rows of the expression matrix can be considered as the data points or objects for data analysis. In the first case, the expression levels of the different genes are considered to be the variables while in the second case this is true for the experiments. In this chapter, however, we will, in most cases, consider the microarray experiments (each associated with a tumour or patient - column vectors of the expression

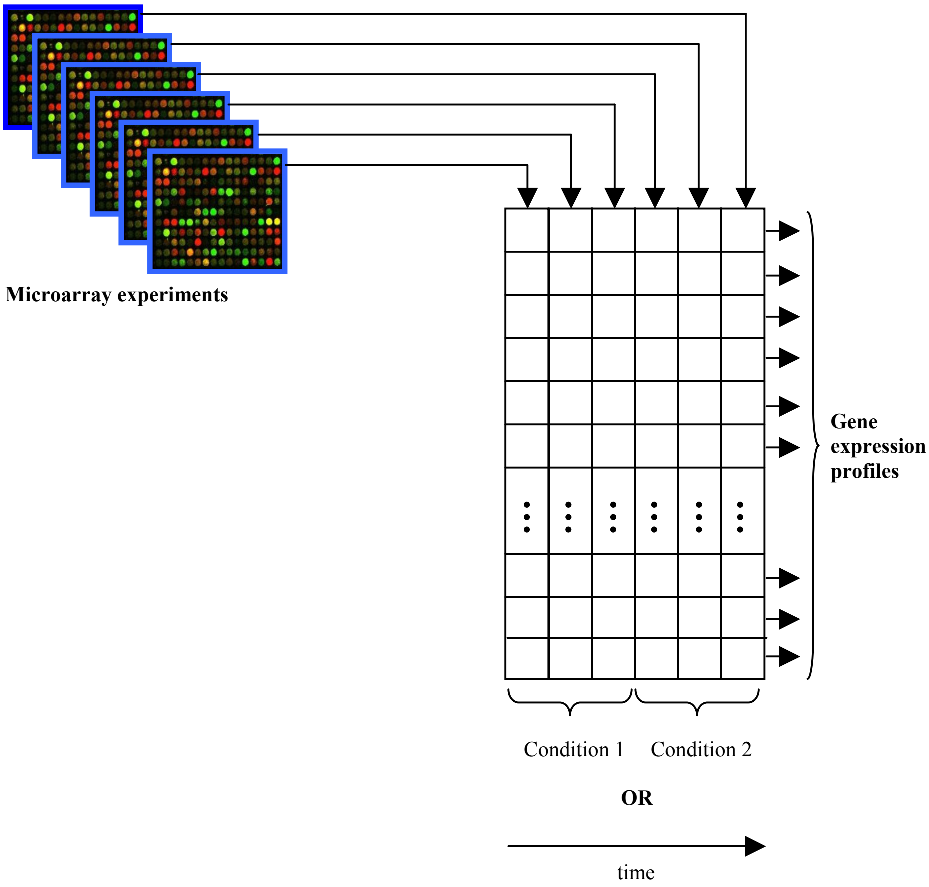


Figure 3.1: Construction of an expression matrix of 6 microarray experiments. The high dimensional vectors resulting from different microarray experiments (studying patients under different conditions or samples taken at different time points during a certain biological process) can be placed in the 6 columns of a matrix. One row of this matrix represents the different measurements of a specific gene over the different experiments and is called a gene expression profile.

matrix) as the data points or objects and the gene expression measurements as the variables. Cluster analysis of gene expression profiles forms an exception to this rule - see further in Section 3.4.2 and Chapter 4 and 5. In this case, the row vectors of the expression matrix are considered to be the data points.

In this Chapter, we will also consider data sets that contain microarray experiments that study tumour cells originating from different classes or conditions with different properties (while for the study of cluster analysis of gene expression profiles in Chapter 4 and 5 we will focus on data sets that contain samples taken at different time points during a certain biological process - also see Figure 3.1). These different classes could for example be:

- Tumours with a different histopathological diagnosis (Golub et al., 1999; Nielsen et al., 2002; Pomeroy et al., 2002).
- Tumours in a different stage of development (Shridhar et al., 2001, Tapper et al., 2001).
- Tumours with a different prognosis (Rosenwald et al., 2002; van de Vijver et al., 2002; van 't Veer et al., 2002; Huang et al., 2003; Iizuka et al., 2003; Nutt et al., 2003).
- Tumours with a different therapy response (Kihara et al., 2001; Chang et al., 2003).
- Benign versus malignant tumours (Alon et al., 1999).
- Primary tumour versus metastasis (Ramaswamy et al., 2003).
- Sporadic versus hereditary tumours (Hedenfalk et al., 2001).
- Tumours with different clinical behavior but using present clinical guidelines, assigned to the same diagnostic category (Alizadeh et al., 2000; Armstrong et al., 2002).

In the following sections, we will first discuss some issues related to preprocessing microarray data after which we will examine the different elements of our data-mining framework applied to this data type: feature extraction, clustering and classification. An in-depth study of two sub-items of our data-mining framework (clustering of gene expression profiles and univariate analysis) will be presented in Chapter 4, 5 and 6. To illustrate the methodology, we will apply the algorithms to the data from Golub et al. (1999) (acute leukemia - ALL versus AML) and Perou et al. (2000) (degree of differentiation in breast tumours - grade 2 versus grade 3) as they are described in Appendix B.

3.2 Preprocessing

Before submitting microarray data to the algorithms or methods described in the next sections, it often has to undergo some preparatory steps (preprocessing). In this section some of the most common preprocessing steps like normalization, non-linear transformation and missing value replacement will be examined. Two additional preprocessing steps - filtering and standardization - more often associated with clustering gene expression profiles, are described in Chapter 4. It is important to mention that these steps can have an important impact on the final result.

3.2.1 Normalization

The first preprocessing step that is customarily applied is the normalization of the hybridisation intensities within a single array experiment (Quackenbush, 2001; Engelen et al., 2003; Marchal et al., 2004). In a two-channel cDNA-microarray experiment several sources of noise (due to for example differences in dye, labelling, in detection efficiency, and in the quantity of initial RNA within the two channels) create systematic sources of bias. The bias can be computed and removed to correct the data. Since many sources can be considered and since they can be estimated and corrected in a variety of ways, many normalization procedures exist but will not be further discussed here. For an illustration, see Figure 3.2 where the dye related bias is removed using a Lowess fit.

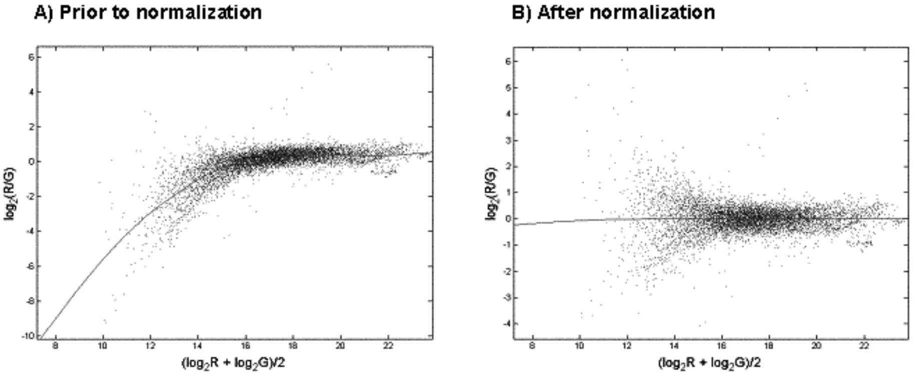


Figure 3.2: Illustration of the influence of an intensity-dependent normalization to remove the bias between the dyes in a cDNA-microarray experiment. Panel A: representation of the log-ratio $\log_2(R/G)$ versus the mean log intensity $(\log_2(R)+\log_2(G))/2$ of every spot on the array (R and G are the intensities in the red and green channel, respectively). At low average intensities the average ratios become negative indicating that the green dye is consistently more intense as compared to the intensity of the red dye. This phenomenon is referred to as the non-linear dye effect. Panel B: Representation of the ratio $\log_2(R/G)$ versus the mean log intensity $(\log_2(R)+\log_2(G))/2$ after performing a normalization based on the Lowess fit (Yang et al., 2002).

3.2.2 Non-linear transformations

It is common practice to pass expression values through a non-linear function (Quackenbush, 2001). Often the logarithm is used for this non-linear function. This is especially suited when dealing with expression ratios (coming from two-channel cDNA-microarray experiments, using a test and reference sample) since expression ratios are not symmetrical. Upregulated genes have expression ratios between 1 and infinity, while downregulated genes have expression ratios squashed between 1 and 0. Taking the logarithms of these expression ratios results in more symmetry between expression values of up- and downregulated genes.

3.2.3 Missing values management

Microarray experiments often contain missing values (measurements absent because of technical reasons) (Troyanskaya et al., 2001). The inability of many algorithms to handle such missing values necessitates their replacement or the development of methods that can deal with these missing values in a more direct way. Simple replacements, which are customarily, such as a replacement by zero or by the average of the expression profile often disrupt these profiles. Indeed replacement by average values relies on the unrealistic assumption that all expression values are similar across different experimental conditions.

In this paragraph we will describe two methods that we have used during our research.

Missing value management without replacement

In some cases, algorithms only need to calculate the (Euclidean) distance between expression vectors and/or calculate average expression vectors (like for example K-means, hierarchical clustering (see Section 3.4.1) or our algorithm AQBC for clustering gene expression profiles that is described in Chapter 5). By a slight change in the definition of how distances and average expression vectors have to be calculated, it is possible to handle these missing values without replacing them (Kaufman and Rousseeuw, 1990).

Suppose that $A = \{v_i(v_i^1, v_i^2, ..., v_i^j, ..., v_i^J)\}_{i = 1,...,I}$ is a set of I expression vectors v_i where J is the number of measurements for each expression vector. At this moment we do not specify whether the expression vectors of A are entire microarray experiments (columns of expression matrix) or gene expression profiles (rows of the expression matrix). This is dependent on the definition of the data points of the specific algorithm (e.g., cluster algorithm for microarray experiments versus cluster algorithm for gene expression profiles). Suppose that the measurement numbers of the missing values for expression vector v_i are given by the set $P_i = \{p_{i,m}\}_{m=1,...,Mi}$, where M_i is the number of missing values in v_i. For example, suppose that $v_1 = (1,3,-9,*,5,*,0)$ ('*' indicates a missing value), then $P_1 = \{4,6\}$ ($p_{1,1} = 4$; $p_{1,2} = 6$; $M_1 = 2$).

If we want to calculate the Euclidean distance $d(v_k,v_l)$ between v_k and v_l, we have to take their missing values into account. Suppose that $\#(P_k \text{ U } P_l) < J$, otherwise $d(v_k,v_l)$ is undefined. We define $d(v_k,v_l)$ as:

$$d(v_k,v_l) = \sqrt{\frac{J}{J - \#(P_k \cup P_l)} \sum_{j \notin (P_k \cup P_l)} (v_k{}^j - v_l{}^j)^2}. \tag{3.1}$$

This means that calculating distances is done by considering only those components for which there are values present in both expression vectors. Since this means that the number of terms in the sum in Equation 3.1 can vary, a weighing factor is applied to account for the different number of terms. For example if $v_1 = (1,*,*,-7,9,0,-1)$ and $v_2 = (*,2,*,5,1,*,*)$ then $P_1 = \{2,3\}$, $P_2 = \{1,3,6,7\}$, $P_1\text{U}P_2 = \{1,2,3,6,7\}$ and $\#(P_1\text{U}P_2) = 5$. The distance $d(v_1,v_2)$ is given by:

$$d(v_1,v_2) = \sqrt{\frac{7}{7-5}\left[(-7-5)^2 + (9-1)^2\right]}. \tag{3.2}$$

If we want to calculate the mean expression profile v_{av} of A, we also have to take the missing values into account. The j-th measurement of v_{av} ($v_{av}{}^j$) is defined as follows (Note that $* \,.\, 0 = 0$):

$$v_{av}{}^j = \begin{cases} \frac{1}{N(j)} \sum_{i=1}^{I} (v_i{}^j . D(i,j)) & \text{if} \quad N(j) \neq 0 \\ \quad * & \text{if} \quad N(j) = 0 \end{cases}, \tag{3.3}$$

where

$$D(i,j) = \begin{cases} 1 & \text{if} \quad j \notin P_i \\ 0 & \text{if} \quad j \in P_i \end{cases} \tag{3.4}$$

and

$$N(j) = \sum_{i=1}^{I} D(i,j). \tag{3.5}$$

This means that the components of v_{av} are the mean values of the corresponding components of the expression vectors in A for which there actually values present. For example if $A = (v_1,v_2,v_3)$ where $v_1 = (1,*,*,-7,9,0,-1)$, $v_2 = (*,2,*,5,1,*,*)$, and $v_3 = (2,3,*,-9,*,6,*)$ then

$$v_{av} = (\frac{1+2}{2}, \frac{2+3}{2}, *, \frac{-7+5-9}{3}, \frac{9+1}{2}, \frac{6+0}{2}, \frac{-1}{1}). \qquad (3.6)$$

Nearest neighbour approach

The second approach to deal with missing values is based on the hypothesis that in a microarray data set one can, for each gene with one or more missing values, find other genes with similar expression behavior (these genes are called coexpressed - also see Chapter 4) that can be used to estimate and replace the missing values.

We have implemented the method as follows (also see Van den Enden, 2001). Consider a gene expression profile g_{mv} with a missing value for the p^{th} component and that belongs to a set of n gene expression profiles $A = \{g_i(g_i^1, g_i^2, \ldots, g_i^j, \ldots, g_i^e)\}_{i = 1,\ldots,n}$ of dimension e. The algorithm to replace this missing value is given in Table 3.1. First we calculate the similarity (concretely, we used the absolute value of the Pearson correlation coefficient) between g_{mv} and every other gene expression profile in the microarray data set. Since the calculation of this similarity has to take the presence of missing values into account, this was done using an approach similar to the calculation of distances in the previous method for missing values management (only the values that are actually present in both expression profiles are used to calculate the correlation coefficient). Next, we select a fraction (default 5%) of the genes with the highest absolute value of the correlation with g_{mv} and from this fraction we again select the set of gene expression profiles without a missing value for the p^{th} measurement. Then we model the relationship between the components of g_{mv} and the components of every selected gene expression profile using linear regression. Each linear regression model (one for each selected gene expression profile) can be used to estimate the missing value in g_{mv} using the p^{th} measurement in the selected gene expression profile at hand. Finally, the missing value is replaced by an average of these estimates.

A slight variation on this method, in which linear regression models between *every* gene expression profile in the data and g_{mv} are considered and where the missing value is replaced by a weighted average (using the absolute value of the correlation coefficient as weights), could also be useful.

After the implementation of our method, Troyanskaya et al. (2001) published a similar method, which they called the K-nearest neighbours method.

3.2.4 Examples

As stated in the introduction, we will use the data from Golub et al.

Table 3.1: Nearest neighbour (NN) approach for replacing a missing value in the p^{th} component of $g_{mv} \in A = \{g_i(g_i^1, g_i^2, ..., g_i^j, ..., g_i^e)\}_{i = 1,...,n}$.

```
NN ($A = \{g_i(g_i^1, g_i^2, ..., g_i^j, ..., g_i^e)\}_{i = 1,...,n}, mv, p$)

FOR $i = 1,...,n$

    $r_{i,mv} = \frac{\sum_{j \notin P_i \cup P_{mv}} (g_i^j - \mu(g_i))(g_{mv}^j - \mu(g_{mv}))}{\sqrt{\sum_{j \notin P_i \cup P_{mv}} (g_i^j - \mu(g_i))^2} \sqrt{\sum_{j \notin P_i \cup P_{mv}} (g_{mv}^j - \mu(g_{mv}))^2}}$

/* Calculate correlation between $g_{mv}$ and every gene expression profile in $A$ */

END FOR

$A^{SORT} = \{g_{si}\}_{i = 1,...,n}$ where $|r_{s1,mv}| \geq |r_{s2,mv}| \geq ... \geq |r_{sn,mv}|$
/* Sort expression profiles according to correlation */

$EST = 0$

$COUNT = 0$

FOR $i = 1,...,C$ where $C$ = CEIL(0.05 x $n$)
/* Calculate estimates of $g_{mv}^p$ using profiles with 5% highest correlation*/

        IF $g_{si}^p \neq *$

                $COUNT = COUNT + 1$

                $EST = EST + w.\ g_{si}^p + b$
                /* Linear regression between $g_{si}$ and $g_{mv}$*/

                $w = \frac{\sum_{j \notin P_{si} \cup P_{mv}} (g_{si}^j - \mu(g_{si}))(g_{mv}^j - \mu(g_{mv}))}{\sum_{j \notin P_{si} \cup P_{mv}} (g_{si}^j - \mu(g_{si}))^2}$

                $b = \mu(g_{mv}) - w.\ \mu(g_{si})$

        END IF

END FOR

$EST = EST\ /\ COUNT$        /* Calculate average of all estimates */

$g_{mv}^p = EST$             /* Replace missing value */
```

and Perou et al. to illustrate the algorithms discussed in this chapter. Here we will briefly describe the steps that we performed to prepare the data for further analysis.

The data from Golub et al. (Affymetrix chips) had undergone a crude normalization step before downloading (such that the overall intensities for each chip were equivalent - the authors called this re-scaling). Next and according to the original publication, every expression value below 20 was replaced by 20 (application of a threshold), since, according to the authors, discrimination of expression below this level could not be performed with confidence. Finally, and also following the guidelines of the authors, a logarithmic transformation (base 10) was performed. No missing values were present.

For the data of Perou et al. (cDNA-microarray technology), we first selected the experiments associated with moderately or poorly differentiated tumours after downloading (resulting in 57 microarray experiments). Next, we calculated the ratio of the difference between the total and background intensity from the tumour and reference sample. Subsequently, a simple normalization was performed by multiplying each array with a single scaling factor so that the median ratio on each array was 1 (Alizadeh et al., 2000). Then a logarithmic transformation (in this case with base 2, but the actual value for this is not important) was performed. Finally the missing values (8% of the values were missing) were replaced using the nearest neighbour approach as described above.

3.3 Feature extraction

Not all genes are correlated with or contain information about the class distinction between samples. In this section we would like to determine a limited number of features that are as informative as possible about a certain class distinction. This is also called the problem of reduction of dimensionality (e.g., reduction of the number of dimension from 7129 genes to for example 5 features that are maximally correlated with the ALL-AML distinction in the data set from Golub et al.). This reduction will have several advantages. It allows identifying the set of features that could be responsible for the distinction between the different sample types. For instance when comparing expression patterns of tumour cells to normal cells, the genes responsible for carcinogenesis could be pinpointed, which could open perspectives for appropriate drug development (identification of therapeutic targets (Gerhold et al., 2002)). Furthermore, dimensionality reduction will facilitate or even enable the other components of the data-mining framework (clustering and classification). Often it is a mandatory preprocessing step before other algorithms can be applied.

Feature selection can be done in a supervised or unsupervised way. In supervised feature extraction, the distinction between the different classes is used to select the features while in unsupervised feature extraction the class labels of the different samples or microarray experiments do not have to be known. Supervised feature extraction is not an appropriate preparatory step before cluster analysis of microarray experiments since, by definition, class membership of the samples is not known in advance (but is a result of the cluster analysis itself - see further in Section 3.4.1). Both supervised and unsupervised feature extraction can be used in combination with classification of microarray experiments. This can, for example, be appropriate if classifiers are used that do not use regularization (e.g., Fisher's linear discriminant analysis - see Section 3.5.1). Without prior feature reduction the risk of overfitting would be extreme.

Below, the two types of feature extraction - univariate and multivariate - that were described in Chapter 1, will be discussed in the context of microarray data analysis.

3.3.1 Univariate feature extraction

In univariate feature selection, we want to rank the *individual* genes according to their correlation with a certain class distinction and select the genes with a maximum degree of correlation. This is the simplest method for feature extraction. This selection is logical because usually microarray data contains a considerable number of genes whose expression is not affected by the different conditions that are under consideration. In univariate analysis one aims to remove as much of those genes as possible and only retain the individual genes most closely related to the class distinction. Univariate feature extraction is always supervised.

Several strategies are possible to perform univariate analysis in microarray data and to quantify the degree of correlation with a certain class distinction. Golub et al. (1999) for example, have introduced a simple measure or score $G(g_i)$ that quantifies the correlation between a single gene expression profile g_i and two different classes:

$$G(g_i) = \frac{\mu_1(g_i) - \mu_2(g_i)}{\sigma_1(g_i) + \sigma_2(g_i)}, \tag{3.7}$$

where $\mu_1(g_i)$ and $\mu_2(g_i)$ are, respectively, the mean values of the expression levels of gene g_i belonging to samples from class 1 and 2 - $\sigma_1(g_i)$ and $\sigma_2(g_i)$ are the associated standard deviations. As an example, we calculated $G(g_i)$ corresponding to the class distinction ALL-AML for all the 7129 genes in the leukemia data set from Golub et al. We selected the 5 genes with the highest absolute value of $G(g_i)$. These are displayed in the Table 3.2.

Table 3.2: Univariate analysis using the score $G(g_i)$ introduced by Golub et al. Selection of the 5 genes with the highest absolute value for $G(g_i)$ in the leukemia data set from the same authors. These are therefore the genes that are most discriminative between ALL and AML, according to this score.

Gene description	$G(g_i)$
CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	-1.5956
CTPS CTP synthetase	1.5494
Leukotriene C4 synthase (LTC4S) gene	-1.4959
DF D component of complement (adipsin)	-1.3935
C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds	1.3719

A related method to do univariate feature extraction in microarray data is to employ classical hypothesis testing (Dawson-Saunders and Trapp, 1994 - also see Appendix A, Section A.1) where a test statistic and the associated p-value is assigned to each individual gene. But in the case of microarray data, a hypothesis test should be performed for thousands of genes simultaneously, which results in an extreme situation of multiple testing that cannot be corrected adequately using classical techniques (see Section A.1 (Appendix A) and Section 2.3.1 where a Bonferroni correction was discussed to correct for multiple testing in clinical data). This problem will be fully investigated in Chapter 6.

A final approach that could be used to quantify the relation between the individual gene expression profiles of a microarray data set and two classes is to construct a receiver operating characteristic curve for the gene expression levels of each gene and calculate the area under the curve (see Appendix A, Section A.2). This last value quantifies how well the expression levels of the gene at hand can discriminate between the two classes.

Univariate analysis is commonly used to select the genes that warrant further biological investigation or validation (e.g., for target discovery in drug development (Gerhold et al., 2002)). This could also be used as a preparatory step before classification or clustering, but for this task another feature extraction method, called principal component analysis, is more suitable and is discussed in the next section.

3.3.2 Multivariate feature extraction

In the previous chapter we noted that, ideally, 6 to 10 patients or data points are needed for each variable that is considered for inclusion during model selection techniques (like stepwise logistic regression analysis - see Appendix A, Section A.3.4). Therefore and due to the high dimensionality of microarray experiments, these methods cannot be used directly in combination with microarray data to select variables or genes (that significantly contribute in a logistic regression model aiming to discriminate between two classes of microarray experiments). Other feature reduction techniques that decrease the dimensionality of the data points drastically (univariate selection or principal component analysis - see further) are needed first.

We have not studied model selection techniques in combination with prior feature reduction for microarray data. For this type of data another multivariate feature selection technique that was already mentioned in Chapter 1 and called principal component analysis (PCA), is more common (Bishop, 1995; Quackenbush, 2001).

In univariate feature selection each feature corresponds to exactly one gene expression level. However, in general, the distinction between classes is not fully determined by the activity of a single gene, but rather by the interaction of several genes. It is therefore better to work with a (linear or non-linear) combination of genes. In PCA, linear combinations of the different gene expression values of a microarray experiment are selected. The coefficients of the linear combinations in PCA are determined in such a way that these linear combinations have maximal spread (or standard deviation) for a certain collection of microarray experiments. In fact, PCA searches for the combinations that are most informative. Each linear combination results in exactly one value for each microarray experiment and can thus be regarded as one feature. The coefficients of the linear combinations can also be arranged in (column) vectors, with the same dimensionality as the microarray experiments, called the principal components for the collection of experiments at hand. The principal components are orthogonal and can be found by calculating the eigenvectors of Σ:

$$\Sigma = \frac{1}{n-1} A.A', \tag{3.8}$$

where *A* is the expression matrix (*n* x *e* matrix - collection of *e* microarray experiments where *n* gene expression levels were measured). Σ is called the covariance matrix of the expression matrix *A* - *A* has to be centralized in Equation 3.8, i.e., the mean column vector of *A* has to lie in the origin. The eigenvectors or principal components with the largest eigenvalues also correspond to the linear combinations with the largest spread in the collection of microarray experiments represented by *A*. In general, if $n > e$ (which is always the case for microarray data), the rank of Σ cannot be higher than *e*-1 (because *A* is centralized and therefore the columns of *A* are linearly dependent) and one can find maximally *e*-1 principal components with an eigenvalue different from zero. Since, in practice, the microarray experiments of *A* are almost always linearly independent before centralization, exactly *e*-1 principal components can be identified corresponding to an eigenvalue that is different from zero. All these principal components span an *e*-1 dimensional subspace containing all the (centralized) microarray experiments in *A*.

The linear combinations or features themselves can be calculated by projecting the expression vector of a certain microarray experiment onto the principal components. If all the principal components are used, this would constitute a dimensionality reduction *n* to *e*-1. In this case, the centralized microarray experiments of *A* can be completely reconstructed after feature extraction (no information is lost). In practice, however, not all the *e*-1 principal components are used but a selection is made according to a certain criterion (see further), which means a further dimensionality reduction, but in this case with loss of information (the original microarray experiments of *A* cannot be fully reconstructed using this limited set of features). So if *m* (*n* x 1) is the centralized expression vector for a certain microarray experiment, the columns of *P* (*n* x *s*) contain *s* selected principal components of the expression matrix *A* and *F* (*s* x 1) is given by:

$$F = P^T.m, \tag{3.9}$$

then the *s* components of *F* contain the *s* features or linear combinations for the microarray experiment with expression vector *m* according to the *s* principal components of the collection of microarray experiments represented by *A*.

The selection of the principal components of *A* can be done in a unsupervised or supervised way. If the selection is unsupervised, the principal components that are associated with the largest eigenvalues of Σ (i.e., corresponding to the features with the largest spread in *A*), are chosen. The principal components associated with smaller eigenvalues are assumed to lie in the directions that are dominated by noise. In supervised selection of the principal components, the features of the microarray experiments of *A* (rows of $P^T.A$) are considered as univariate data that can be selected using the methods described in Section 3.3.1.

In the next sections we discuss the application of PCA to the data from Golub et al. and on the data from Perou et al.

PCA for the data from Golub et al.

We calculated the principal components of the training set (e = 38) from Golub et al. and selected the two principal components associated with the two largest eigenvalues (unsupervised selection). The associated features of the patients of the training *and* test can (e = 34) be inspected in Figure 3.3. The separation between patients with ALL and AML (also for patients from the test set that were not used to derive the principal components here) is clearly visible. This means that, in this case, the directions in which the distinction between ALL and AML is prominent are also the directions with the largest spread in the data. One could say that the distinction between ALL and AML is dominant here.

PCA for the data from Perou et al.

We derived the principal components of the 57 patients from Perou et al. The features of the patients associated with the principal components with the two largest eigenvalues (supervised selection) can be inspected in the upper plot of Figure 3.4. In this case there is no clear separation between patients with grade 2 or grade 3 breast tumours. One can conclude that in this example the directions with the largest spread are not dominated by distinction between moderately or poorly differentiated breast tumours (but could possibly be caused by other factors, but since we do not have additional clinical information about these patients, this cannot be investigated). For this data set it could therefore be meaningful to perform supervised selection of the principal components. Based on the absolute value of the score introduced by Golub et al. (Equation 3.7), we selected principal components 5 and 30 (Golub scores: 0.48 and 0.42 respectively). Using this score, the features associated with principal components 1 and 2 were ranked on the 33rd and 16th place, respectively (Golub scores: 0.07 and -0.13). The features associated with principal components 5 and 30 can be inspected in the lower plot of Figure 3.4. Although still not perfect (a large amount of overlap still exists), the separation between grade 2 and 3 breast tumours is clearly better when compared to the separation of these two classes if unsupervised selection of principal components was used. In this example, supervised selection of principal components would therefore be a better option if one aims to develop models that can discriminate between grade 2 and 3 tumours (see Section 3.5.3).

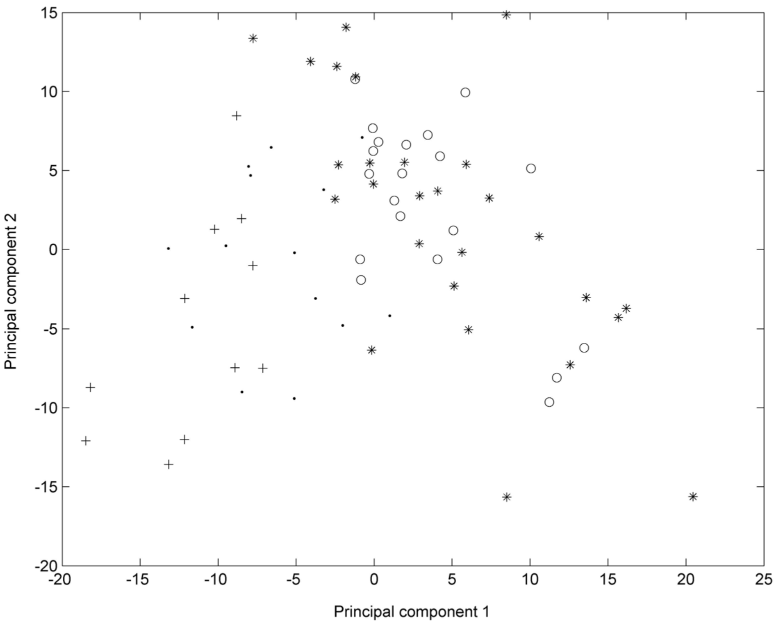
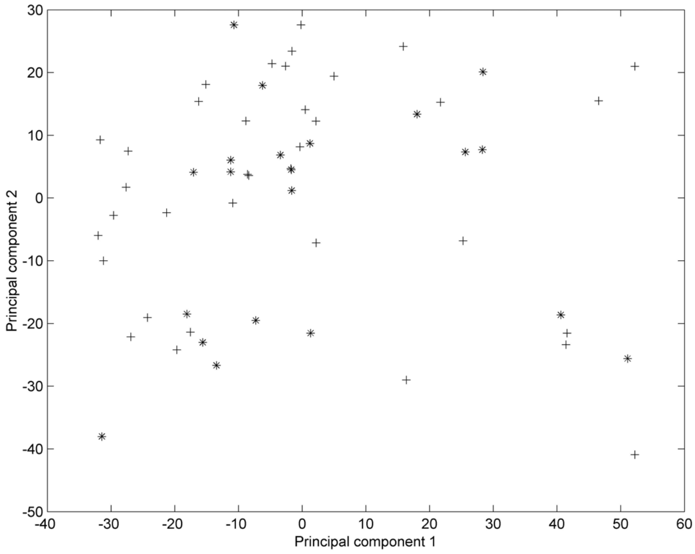


Figure 3.3: Principal component analysis for the data from Golub et al. The two principal components of the training set associated with the largest eigenvalues were selected (unsupervised). Every sample of the training and test set was projected onto these two components, resulting in two features or values (plotted in the X- and Y-axis here) for every microarray experiment. Training set: * = ALL, + = AML; Test set: O = ALL, • = AML.

3.4 Clustering

3.4.1 Cluster analysis of microarray experiments

As already stated in Chapter 1, with cluster analysis one aims to automatically find different classes in a group of data points without knowing the properties of these classes in advance. If these data points are microarray experiments, cluster analysis will group the tumour samples with a certain degree of similarity in expression behavior. The distinct classes or clusters generated by the clustering procedure will probably - at least partially - match with the existing diagnostic categories used for the current classification of tumours, which is predominantly based on clinical parameters. However since expression data are not customarily used for the present classification schemes, it is not excluded that novel, yet unknown



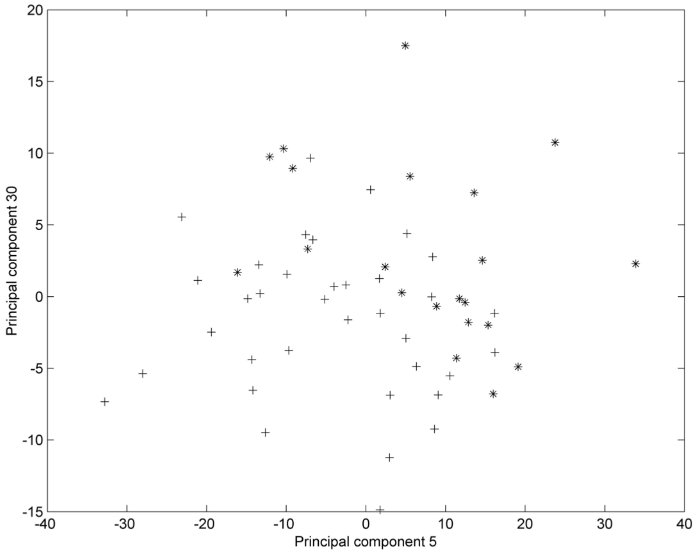


Figure 3.4: Principal component analysis for the data from Perou et al. The principal components of the complete data set (57 patients) were determined. Every sample was projected onto two selected principal components. * = grade 2 tumour, + = grade 3 tumour. Upper plot: unsupervised selection of principal components with the two largest eigenvalues. Lower plot: supervised selection of principal components based on the absolute value of the score introduced by Golub et al. (Equation 3.7). This resulted in the selection of the principal components with the 5th and 30th largest eigenvalue.

diagnostic entities might originate from these analyses, which could improve clinical management of cancer. Cluster analysis of microarray experiments could therefore be used to discover new diagnostic categories or subcategories that might group patients with less clinical variability.

For example, diffuse large B-cell lymphoma is a disease that is clinically heterogeneous. Some patients respond well to therapy and achieve a durable remission, while other patients have a less favourable prognosis. Although clinical parameters are available that can assess the risk profile of the patients, these prognostic variables are not ideal yet. Using hierarchical clustering of microarray data (see further) Alizadeh et al. (2000) claim to have found two clinically distinct forms of (or clusters in) patients with diffuse large B-cell lymphoma with a significantly different overall survival. The authors conclude that these two groups of patients might represent two distinct subentities that could be the basis of a new classification scheme.

Below we will apply two commonly used techniques to cluster microarray experiments: K-means and hierarchical clustering. We will illustrate these techniques using the data from Golub et al.

K-means

The K-means algorithm is described in Appendix A, Section A.5. This algorithm finds a prespecified number (*K*) of clusters in a set of data points or, in this case, microarray experiments. A form of (unsupervised) feature extraction has to be performed in advance if one wants to cluster high dimensional microarray experiments using this approach. We will use principal component analysis for dimensionality reduction here (see Figure 3.5).

We applied K-means clustering to cluster the complete data set from Golub et al. (72 patients - in this case, we do not consider the subdivision between training and test set). First, imagine that the difference between ALL and AML is not known. In this case we have simply a data set with 72 patients with acute leukemia. After principal component analysis (also based on the complete data set) with unsupervised selection of five principal components, we submitted the data to a K-means algorithm with $K = 2$. The result can be inspected in Figure 3.5. The algorithm has succeeded in finding two clusters. When looking at the first cluster, one can see that all the patients - except one - have ALL. When looking at the second cluster, one can see that all the patients have AML. This means that the procedure was in fact able to redefine the concepts ALL and AML. In this example, nothing new is learned (because ALL and AML were already known), but the result clearly shows the potential of this technique.

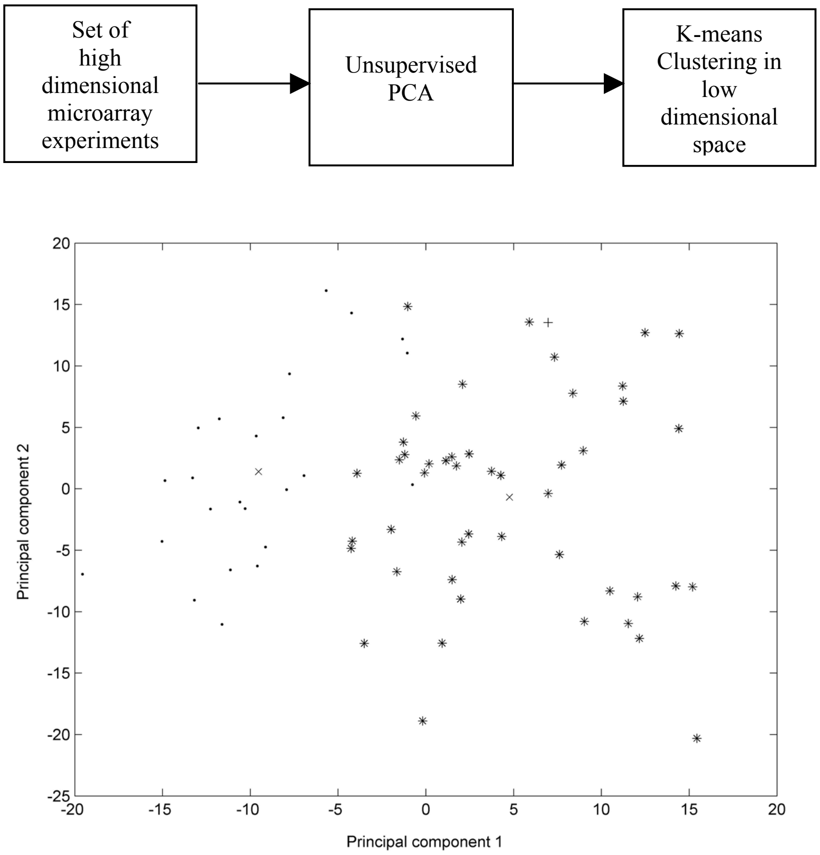


Figure 3.5: K-means clustering with *K*=2 for the complete data set from Golub et al. Upper plot: visualization of the different steps to cluster microarray experiments with K-means. Since K-means cannot be applied in combination with high dimensional data, unsupervised PCA has to be performed for dimensionality reduction. In this case we selected the five principal components associated with the largest eigenvalues. Lower plot: cluster result. Only the first two principal components are shown in this figure, although the clustering procedure was done in five dimensions. Note the almost perfect correlation between the clusters and the clinical classification (ALL-AML). Cluster 1: * = ALL, + = AML; Cluster 2: O = ALL, · = AML; × = cluster means.

Hierarchical clustering

Hierarchical clustering is the most commonly used method for cluster analysis of microarray data. This method places the data points in a tree structure and the clusters are formed by cutting the tree at a certain level. See Appendix A, Section A.6 for more information. Hierarchical clustering can be used in combination with high-dimensional data and therefore PCA or other feature reduction methods are not mandatory before analysis of microarray experiments with this method.

As an example, hierarchical clustering was also applied to group the samples of the complete data set from Golub et al. See Figure 3.6 for the resulting tree structure. We used average linkage clustering and chose the correlation coefficient as distance measure between the data points. Most patients with AML are concentrated in one single branch in Figure 3.6.

Critical remarks

In a recent article by Levenstien et al. (2003), the authors raised an important problem related to the results obtained with hierarchical cluster analysis of microarray experiments. Since hierarchical clustering results in several possible sets of clusters, the biologist or medical doctor has to choose an 'appropriate' set (i.e., choose a certain cut-off level). The most appropriate set, however, will often be the set that optimally supports a certain a-priori hypothesis, like a large difference in survival between the patients of the different clusters. Since there are multiple sets to choose between, it might well be that the most appropriate set (i.e., the largest difference in survival) was generated by accident (problem of multiple testing) and in fact does not represent a real biological or medical category. Levenstien et al. quantify this observation by assigning a global p-value to the result obtained by hierarchical clustering of a set of microarray experiments. This p-value represents, for example, the probability that the largest difference in survival between the patients of a set of clusters could be generated by accident.

In our opinion, the problem of multiple testing related to cluster analysis of microarray experiments may even be larger in some cases. Firstly, when hierarchical clustering is used, it is often customary to execute several runs of the algorithm with different parameter settings (e.g., choice between single, complete, average or centroid linkage clustering; choice of the distance measure between data points; different choice of preprocessing) each resulting in several possible sets of clusters. This can inflate the number of possible cluster results to choose the most appropriate result from. These different runs are, in our opinion, often not mentioned in publications (because they did not give meaningful results) while they can degrade the significance of the result that is finally published. Secondly, the definition of the most appropriate cluster result or the a-priori hypothesis that has to be supported can vary or is not fixed before the start of the cluster analysis. Not only cluster results with a large difference in survival can be useful, but also clusters with a large difference in other characteristics of the tumour cells (e.g., histopathology). This can increase the number of tests that has to be performed for each possible set of clusters and also augment the problem of multiple testing. Thirdly and finally, hierarchical clustering is not the only

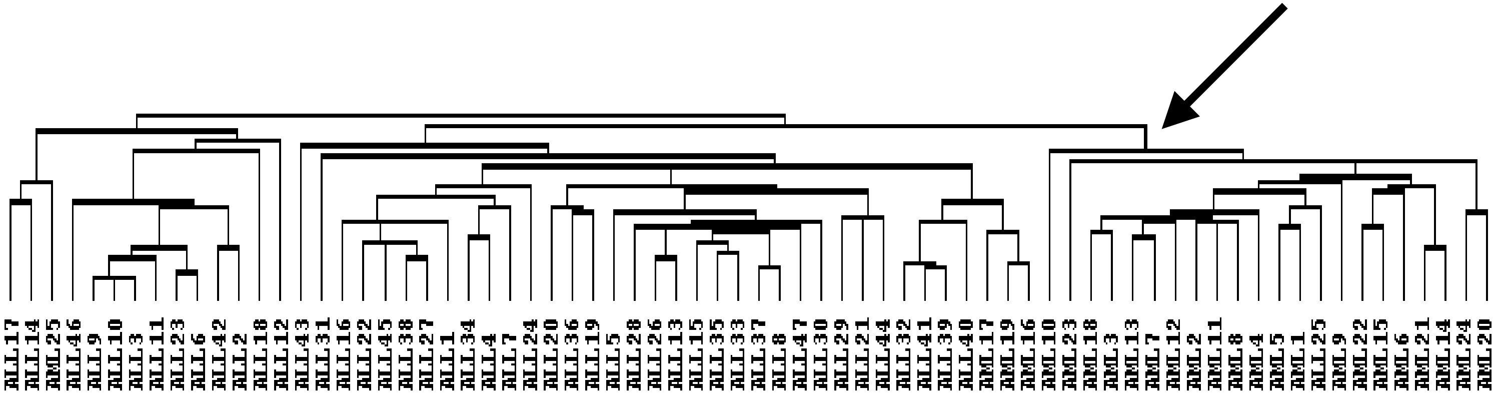


Figure 3.6: Result of average linkage hierarchical clustering (based on the correlation coefficients as distance measure between microarray experiments) of the patients of the complete data set from Golub et al. The terminal branches represent the individual patients (ALL/AML + number). The branch marked with an arrow contains only one patient with ALL and almost all patients (except 4) with AML.

cluster algorithm that is available to cluster microarray experiments. Before publication of a cluster result, the authors could have tried other algorithms (and possibly with different parameter settings again) before the selection of the result that will appear in the final report.

Conclusively, we can state that each cluster result of microarray experiments in literature should be scrutinized. Ideally, the authors should mention a measure like the global p-value introduced by Levenstien et al. Moreover, they should mention how many different cluster results they obtained and removed from consideration using other parameter settings, other algorithms, different preprocessing techniques, and so on. Also, if possible, they should mention if there was an a-priori hypothesis that they wanted to see supported by the cluster result.

3.4.2 Cluster analysis of gene expression profiles

Gene expression profiles can also be used as the basis for cluster analysis. Contrary to the other sections of this chapter, the rows of the expression matrix are considered as the data points or objects of our analysis here and the different measurements for a gene in the different microarray experiments as the variables. PCA could also be performed in this setting. In this case, PCA looks for linear combinations of the different measurements in a gene expression profile and the principal components lie in the row space of the expression matrix. But since the dimension of the gene expression vectors equals the number of experiments in the data set and since this number usually is several orders of magnitude lower than the dimension of a microarray experiment, feature reduction prior to cluster analysis is less important in this setting and usually not an issue. Moreover, performing PCA prior to cluster analysis of gene expression profiles often degrades the cluster quality (Yeung and Ruzzo, 2001c).

An in-depth study of the methodology and specific requirements associated with cluster analysis of gene expression profiles will be presented in Chapter 4. In Chapter 5 we will describe an algorithm that we have specifically designed to cluster this kind of data.

3.5 Classification

As discussed in the introduction of this dissertation, in a clinical environment it is important to be able to do predictions (with regard to diagnosis, prognosis, therapy response, and so on - see the different classes discussed in Section 3.1) for individual patients using microarray experiments. Here a prediction must be made for samples or patients for which class membership (e.g., good versus bad prognosis, benign versus malignant, and so on) is not known in advance. Based on a set of features and a training set, a model has to be trained. This model can then be used to classify new patients for whom the outcome is not known (or is supposed not to be known) - also see Figure 1.5. This approach could help to incorporate expression measurements that represent the fundamental mechanisms that guide the phenotype of the tumour, into the clinical decision making process for individual patients.

In this section we will consider and illustrate two different binary modelling techniques: Fisher's linear discriminant analysis (FDA) and Least Squares Support Vector Machines (LS-SVMs). We will apply these methods to the data from Golub et al. and Perou et al. Furthermore, we will briefly describe the conclusions of a systematic benchmarking study to compare several classification techniques using nine different microarray data sets. More specifically, we want to examine the importance of regularization or dimensionality reduction when classifying microarray experiments and we want to examine if non-linear classification can contribute in the accuracy of the predictions.

3.5.1 Fisher's linear discriminant analysis

Fisher's linear discriminant analysis (FDA) is a linear classification technique that can be used to assign data points or microarray experiments to one of two classes. In FDA (Bishop, 1995) one projects each microarray experiment m^j of the expression matrix $A = [m^1, m^2, ..., m^j, ..., m^e]$ that contains the training data (each microarray experiment of this training set has a known class label, i.e., it belongs to one of two classes: class 1 (C_1) or 2 (C_2)) onto a vector w resulting in a variable $y = [y^1, y^2, ..., y^j, ..., y^e]$:

$$y^j = w^T.m^j. \tag{3.10}$$

The vector w is chosen to maximize the following criterion $J(w)$:

$$J(w) = \frac{(\mu_2 - \mu_1)^2}{\sum_{m^j \in C_1} (y^j - \mu_1)^2 + \sum_{m^j \in C_2} (y^j - \mu_2)^2}, \tag{3.11}$$

where μ_1 and μ_2 are the mean values of y^j associated with the training samples from C_1 and C_2, respectively. Note the similarity between Equation 3.11 and 3.7 (Golub score). Maximizing Equation 3.11 gives the expression for w (only the direction of w is important not the magnitude of w, so scalar factors can be dropped, but we choose the sign of w so that $\mu_1 < \mu_2$):

$$w \propto S_W^{-1}(m^{C_2} - m^{C_1}), \qquad (3.12)$$

where m^{C_1} and m^{C_2} are the average expression vectors for the microarrays of the training set belonging to C_1 and C_2, respectively. They are given by:

$$m^{C_1} = \frac{1}{e_1} \sum_{m^j \in C_1} m^j, \qquad (3.13)$$

and

$$m^{C_2} = \frac{1}{e_2} \sum_{m^j \in C_2} m^j, \qquad (3.14)$$

where e_1 and e_2 are the number of microarray experiments from the training set that belong to class C_1 and C_2, respectively. This also means that $\mu_1 = w^T. m^{C_1}$ and $\mu_2 = w^T. m^{C_2}$. The within-class covariance matrix S_W is given by:

$$S_W = \sum_{m^j \in C_1} (m^j - m^{C_1})(m^j - m^{C_1})^T + \sum_{m^j \in C_2} (m^j - m^{C_2})(m^j - m^{C_2})^T. \qquad (3.15)$$

Now we have to choose a threshold b so that new microarray experiments m^t (from a test set), for which the outcome is not supposed to be known, can be classified. If $y^t < b$ then m^t is predicted to belong to C_1 and if $y^t > b$ then m^t is predicted to belong to C_2. Bishop suggests two methods to derive b from the training data. Firstly, by assuming that the variable y^j is the sum of a set of random variables (see Equation 3.10), we can invoke the central limit theorem and model the class-conditional density functions $p(y^j|C_1)$ and $p(y^j|C_2)$ using normal distributions and the training data. After using Bayes' theorem to calculate the posterior probabilities $P(C_1|y^j)$ and $P(C_2|y^j)$, the threshold b follows from solving:

$$P(C_1 \mid b) = P(C_2 \mid b). \qquad (3.16)$$

In practice we do not solve Equation 3.16 but we evaluate the posterior probabilities using the test sample and assign it to C_1 if $P(C_1|y^t) > P(C_2|y^t)$ and to C_2 if $P(C_1|y^t) < P(C_2|y^t)$. This is equivalent with the comparison of:

$$-\frac{(y^t - \mu_1)^2}{2\sigma_1^2} - \ln \sigma_1 + \ln \frac{e_1}{e} \qquad (3.17)$$

and

$$-\frac{\left(y^t-\mu_2\right)^2}{2\sigma_2^2}-\ln\sigma_2+\ln\frac{e_2}{e}, \tag{3.18}$$

where σ_1 and σ_2 are the standard deviations of y^j associated with the training samples from C_1 and C_2, respectively. If the value given by Equation 3.17 is larger than the value given by Equation 3.18, the test sample is assigned to C_1 and if the value given by Equation 3.17 is smaller than the value given by Equation 3.18, the test sample is assigned to C_2. From this discussion, it follows also that a new microarray experiments m^t with $y^t = w^T. m^t$ can be classified with greater confidence if the difference between y^t and b is greater (and therefore also the difference between $P(C_1|y^t)$ and $P(C_2|y^t)$). Also see Figure 3.7.

Secondly, Bishop proves that under certain assumptions the following is also a valid choice for b:

$$b = w^T.m, \tag{3.19}$$

where m is the average expression vector of the microarrays belonging to the training set. This is given by:

$$m=\frac{1}{e}\sum_{j=1}^{e} m^j. \tag{3.20}$$

Although we did not use this technique here, ROC analysis could, similarly to the method that was applied in Chapter 2, also be a valid approach to determine an optimal value of b in this context.

FDA is a linear classification technique where the number of model parameters that has to be estimated (components of w), is determined by the number of expression values (the variables) in each microarray experiment, which can be considerable. Moreover, this method does not use regularization and thus is prone to overfitting if the number of variables in the model is too high relative to the number of data points in the training set. Since this is certainly the case for microarray data, FDA cannot be applied without prior feature reduction in practice. The same remarks also apply to standard logistic regression (see Section 2.3.2), which is a method that is qualitatively similar to FDA. The need for prior feature reduction in linear classifiers without regularization will also be confirmed by our benchmarking study that we will discuss below.

3.5.2 Least Squares Support Vector Machines

We refer to Appendix A, Section A.4 for more details on this technique. Like in Chapter 2, Section 2.3.2, we used LS-SVMlab version 1.5 to derive our LS-SVM models and tuned the hyperparameters with a linesearch (linear kernel) or gridsearch (Radial Basis Function (RBF) kernel) approach on the training set.

LS-SVM models, unlike logistic regression or FDA, apply regularization (to prevent overfitting) and estimating the model parameters involves solving a dual problem where the number of equations is determined by the number of data points and not by the number of variables. This means that, in principle, LS-SVMs can be directly applied with good results for the classification of microarray experiments without prior feature reduction (which would be necessary if for example FDA is applied) and that the number of equations that has to be solved is equal to the number of microarray experiments (plus one for the constant term of the LS-SVM model) and not to the number of genes in the data set. The direct applicability of LS-SVM models to microarray data will also be confirmed by our benchmarking study.

3.5.3 Examples

Data from Golub et al.

In the original publication of Golub et al., the data was divided in a fixed training and test set. In first instance, we used this fixed training set to derive a model that can distinguish between ALL and AML. After PCA on the training set and unsupervised selection of the first two principal components (see Figure 3.3) we used FDA to construct a linear model in two dimensions. This model applied on the fixed test set can be seen in Figure 3.7 and resulted in three misclassifications (91% accuracy). The principal components of the training set were used to derive the two features for every patient of the test set.

The performance of the model visualized in Figure 3.7 is not necessarily completely representative of the general behaviour of this technique using similar data, since it uses a fixed training and test set. The specific partitioning between training and test that was used here could have accidentally resulted in an over- or underestimation of the model performance. In order to get a more optimal assessment of the model, the whole procedure of model training and testing should be repeated several times using data where the data points have been randomly reshuffled between training and test set. This procedure is called randomisation.

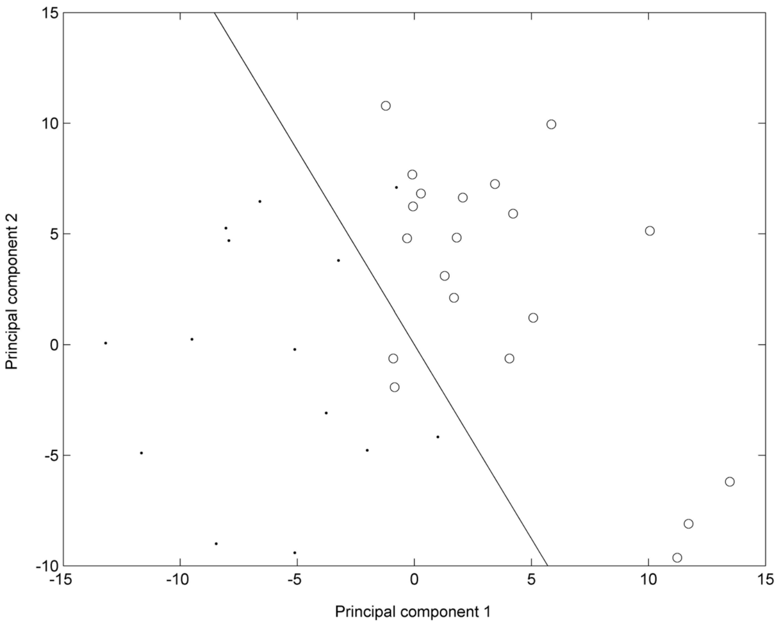


Figure 3.7: Model obtained using FDA for the classification of the patients of the Golub et al. data after PCA on the training set with unsupervised selection of the first two Principal Components. The parameters of the linear model (represented by the line here) were calculated using the patients of the training set (also see Figure 3.3). The patients (only the patients of the test set, for which the two features in the X and Y axis were calculated using the principal components of the training set, are shown in this figure) above the line are classified as ALL and below as AML. This results in 3 misclassifications and these occur in patients that have a relatively small distance to the classification line, confirming that the classification for patients that lie further from this line can be done with greater confidence. Test set: O = ALL, • = AML.

Moreover, the number of principal components that was selected was arbitrarily set to two here. It is possible that changing this number could increase the model performance. In the context of our systematic benchmarking study (see further) we evaluated the performance of FDA on the data from Golub et al. using 20 stratified (training and test set contain the same amount of samples from each class compared to the original training and test set) randomisations where the number of principal components was optimised for each randomisation using a leave-one-out cross-validation performance (LOO-CV) on the respective training sets. This resulted in an average (over the different randomisations) accuracy on the test sets of 94.40% (with a standard deviation σ of 3.84%). Supervised selection of the principal components did not result in a better performance in this case, which could have been expected since the separation between the two classes was already excellent for unsupervised selection (see Figure 3.3).

As a part of our benchmarking study, we also evaluated the performance of LS-SVM models with a linear and RBF kernel on the data from Golub et al. using the same randomisations (without prior feature reduction or PCA). This resulted in an average accuracy on the test sets of 92.86% (σ= 4.12%) for the LS-SVM model with a linear kernel and 93.56% (σ= 4.12%) for the LS-SVM model with an RBF kernel.

Data from Perou et al.

We evaluated the performance of FDA on the data of Perou et al. in distinguishing between grade 2 and 3 breast tumours by a LOO-CV approach. In each LOO-CV iteration a different sample is left out. Subsequently, PCA with supervised selection (based on the Golub score of Equation 3.7 - Figure 3.4 showed that for this data set, unsupervised selection could be expected not to be sufficient) of a fixed number of principal components is performed on the remaining data and the model is trained based on the resulting features. Finally, the left out data point is projected onto the selected principal components and evaluated using the trained model and the resulting prediction is compared with the real value.

When the number of selected principal components in each iteration was set to five, this approach resulted in an LOO-CV accuracy of 79%. This result clearly demonstrates that it is possible to predict the degree of differentiation in breast tumours with a certain degree of accuracy using expression data.

Benchmarking study[2]

As already announced, we performed a systematic benchmarking study to evaluate the role of regularization or dimensionality reduction and to evaluate the role of non-linear techniques in the context of the classification of microarray experiments. We will outline the main elements of this study here. We compared the following techniques for classifying microarray experiments:

1. LS-SVM models with a linear kernel (with (γ finite and tuned) and without regularization (γ infinite, which corresponds to FDA (Suykens et al., 2002))) and an RBF kernel without prior reduction of dimensionality.

[2] This study was submitted as a full paper to Bioinformatics (Pochet et al., 2004).

2. FDA combined with classical PCA and kernel PCA (we will not further discuss the details about the kernel version of classical PCA in this thesis - see Suykens et al. (2002)) with unsupervised and supervised selection of the principal components. The optimal number of selected principal components was determined by a LOO-CV approach on the training set.

For this comparison we examined 9 binary cancer classification problems using 7 data sets that were publicly available, including the data from Golub et al. The other data sets were: Alon et al. (1999) (colon cancer), Hedenfalk et al. (2001) (breast cancer - sporadic versus hereditary), Iizuka et al. (2003) (hepatocellular carcinoma), Nutt et al. (2003) (high-grade gliomas), Singh et al. (2002) (prostate cancer), and van ‘t Veer et al. (2002) (breast cancer - good versus bad prognosis). We refer to Appendix B for more details about these data sets. The performance of the different classification techniques was also evaluated using 20 stratified randomisations of the training and test set. As an illustration, the test set accuracies obtained using randomisations of the data set from Nutt et al. are given in Table 3.3.

Table 3.3: Test set accuracies of different classification techniques applied to 20 randomisations of the data set from Nutt et al. (2003). We tested LS-SVM models with a linear kernel (with and without regularization) and an RBF kernel and we tested FDA combined with classical and kernel PCA (also with a linear and RBF kernel) with supervised and unsupervised selection of the principal components. The average test set accuracies and their standard deviations σ (over the different randomisations) are given.

Classification technique	Accuracy test (%) (± σ)
LS-SVM linear kernel (with regularization)	61.25 ± 11.75
LS-SVM RBF kernel (with regularization)	69.95 ± 8.59
LS-SVM linear kernel (no regularization = FDA)	48.93 ± 10.88
PCA (unsupervised) + FDA	67.82 ± 7.24
PCA (supervised) + FDA	65.52 ± 11.01
kPCA linear kernel (unsupervised) + FDA	68.31 ± 6.78
kPCA linear kernel (supervised) + FDA	67.32 ± 11.04
kPCA RBF kernel (unsupervised) + FDA	64.20 ± 11.19
kPCA RBF kernel (supervised) + FDA	58.13 ± 12.24

The comparison of the different classification techniques applied to these nine classifications problems, resulted in the following three main conclusions:

1. Our study confirmed that LS-SVM models with linear and RBF kernels (γ finite and tuned) without prior dimensionality reduction never resulted in overfitting on all data sets that were examined. The results obtained with RBF kernels (non-linear classifiers) are never worse and sometimes even significantly better compared to results obtained with a linear kernel in terms of the test set performance.
2. Our study also confirmed that regularization appears to be very important when applying linear classification methods onto microarray data without dimensionality reduction. Linear classification techniques without dimensionality reduction and without regularization hardly perform better than random classifiers.
3. Performing kernel PCA with an RBF kernel before classification with FDA tends to result in overfitting.

3.6 Conclusions

In this chapter we discussed and applied the three elements of our data-mining framework on expression patterns of entire microarray experiments and mentioned the application of clustering techniques for gene expression profiles, which will further be elaborated on in the next two chapters. We illustrated the technique using examples from oncology and explained how the results of the analysis of microarray data could help to improve the clinical management of cancer. Although several problems of a more technical nature still exist that can complicate the clinical use of microarrays (e.g., cost, heterogeneous composition of samples from solid tumours and existence of biological and technical variation), it can be expected that in the future this technology will find its way into clinical practice (Friend, 1999).

In the context of data analysis of microarray experiments, we described some frequently used preprocessing steps (normalization, non-linear transformations and missing value management). We noted that in univariate feature extraction and in cluster analysis of microarray experiments, multiple testing is a problem that has to be taken into account. For univariate analysis of microarray data, multiple testing will be studied in further detail in Chapter 6. In the context of cluster analysis of microarray experiments, we illustrated how clustering techniques can potentially discover diagnostic categories in a group of patients but also mentioned that cluster results of microarray experiments in literature should be approached with some caution. Furthermore, we showed that principal component analysis is an adequate multivariate feature selection technique that looks for linear combinations of the expression values of a microarray experiment. We also demonstrated that principal component analysis is an appropriate method that can be used in combination with classification and clustering techniques that cannot deal with the large number of dimensions characteristic for expression patterns of microarray experiments. We explained that the selection of the principal components could be done in an unsupervised and supervised way. We pointed out that supervised selection can only be used in the context of classification and that unsupervised selection is appropriate before cluster analysis and could, in some cases where the principal components with the largest eigenvalues sufficiently capture the class distinction under consideration, be appropriate before classification. Finally, we applied and compared Fisher's linear discriminant analysis and LS-SVM models with respect to the binary classification of microarray experiments and, in this context, presented our systematic benchmarking study. We concluded that regularization or dimensionality reduction is necessary when performing class prediction using microarray experiments and that the introduction of non-linear models can, in some cases, significantly increase model performance.

Chapter 4

Clustering of gene expression profiles

4.1 Introduction

In the previous chapter we have given a general overview of the data mining framework to analyse microarray data. In this chapter we will focus on a specific item of this framework: cluster analysis of gene expression profiles[1].

As previously said, with microarrays one can measure the expression levels of thousands of genes simultaneously. These expression levels can be determined for samples taken under different conditions (e.g., cells originating from tumour samples with different properties, as discussed in the previous chapter). But since clustering of gene expression profiles has been mainly used for microarray data containing samples taken at different time points during a certain biological process (e.g., different phases of the yeast cell cycle), we will focus on these types of data sets in this and the next chapter. The discussion will therefore not be limited or focus on data generated to study problems in oncology, but the methodology described here can of course also be used to analyse them.

For each individual gene, the arrangement of the expression measurements into a vector leads to what is generally called a gene expression profile. This is thus equivalent with a row of the expression matrix. These expression profiles or vectors are the objects that will be analysed in this chapter.

[1] The discussion presented in this chapter has appeared in a review paper in the Proceedings of the IEEE (Moreau et al., 2002a) and will appear in a book chapter (Thijs et al., 2004). We co-authored both publications. We were also actively involved in writing a survey paper that has appeared in the European Journal of Control (De Moor et al., 2003).

Because relatedness in biological function often implies similarity in expression behavior (and vice versa) and because several genes might be involved in the process being studied (e.g., they might be regulated by the same transcription factor - see Chapter 1, Section 1.2 for more details on the biology), it will, in general, be possible to identify subgroups or clusters of genes that will have similar expression profiles (i.e., according to a certain distance function, the associated expression vectors are sufficiently 'close' to one another). Genes with similar expression profiles are called coexpressed.

Conversely, coexpression of genes can thus be an important observation to infer the biological role of these genes. For example, coexpression of a gene with unknown biological function with a cluster containing genes with known (or partially known) function can give an indication of the role of the unknown gene. Also, coexpressed genes are more likely to be coregulated, i.e., they might interact with the same transcription factors.

Clustering algorithms are designed to detect unknown classes in the data (see Chapter 1, Section 1.3.2). This means that cluster analysis in a collection of gene expression profiles aims at identifying subgroups (= clusters) of such coexpressed genes, which thus have a higher probability of participating in the same pathway. An idealized example in two dimensions is shown in Figure 4.1.

Cluster analysis of gene expression profiles is only a first rudimentary step preceding further analysis, which includes motif finding, functional annotation, genetic network inference (Roth et al., 1998; Thijs et al., 2002a; van Helden et al., 2000). Moreover, clustering often is an interactive process where the biologist has to validate or further refine the results and combine the clusters with a priori biological knowledge. Claiming that the biologist can immediately obtain the desired results just by applying the clustering algorithm is, in our opinion, wishful thinking.

In the following sections we will discuss some of the specific problems related to cluster analysis of gene expression profiles, describe some of the solutions that are already available and show that these solutions are still not entirely optimal. This has motivated us to develop a clustering algorithm specifically tuned towards clustering gene expression profiles that aims to circumvent some of the disadvantages of the existing algorithms. This approach will be discussed in the next chapter.

4.2 Algorithmic challenges

The first generation of cluster algorithms (e.g., direct visual inspection (Cho et al., 1998), K-means (Tou and Gonzalez, 1979), self-

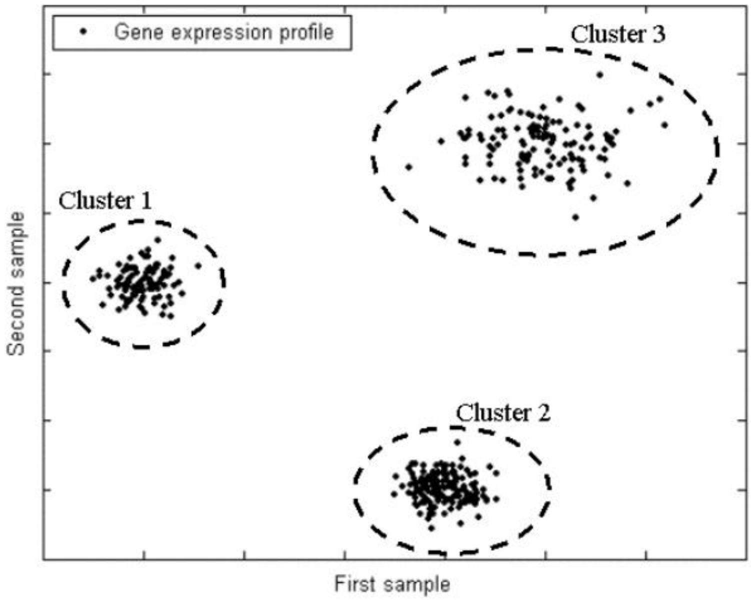


Figure 4.1: Visualization of 375 (simulated) gene expression profiles (each expression profile contains two expression levels measured in two different samples - data not standardized). It is clear that, in this case, cluster analysis will result in the identification of three well-separated clusters (representing three classes of genes, possibly associated with specific biological pathways).

organizing maps or SOMs (Tamayo et al., 1999), hierarchical clustering (Eisen et al., 1998)) applied to gene expression profiles were mostly developed outside biologically related research. Though possible to obtain biologically meaningful results with these algorithms, some of their characteristics often complicate their use for clustering expression data (these methods lack fine-tuning for biological problems) (Sherlock, 2000). They require, for example, the predefinition of one or more arbitrary user-defined parameters that are hard to estimate by a biologist (e.g., the predefinition of the number of clusters in K-means and SOM - this number is almost impossible to predict in advance). Moreover, changing these parameter settings will often have a profound impact on the final result. These methods therefore need extensive parameter fine-tuning, which means that a comparison of the results with different parameter settings is almost always necessary, which is not trivial. Another problem is that first-generation clustering algorithms often force every data point into a cluster. In general, a considerable amount of genes included in the microarray experiment do not really contribute to the biological process studied and these genes will therefore lack coexpression with other genes (they will have seemingly constant or even random expression profiles). Including these genes into one of the clusters will ‘contaminate’ their content (these genes represent noise) and make them less suitable for further analysis. Finally, the computational and memory complexity of some of these algorithms often limit the number of expression profiles that can be analysed at once. Considering the nature of our data sets (number of expression profiles often running up into thousands), this constraint is often unacceptable.

Recently, many new clustering algorithms have emerged claiming to solve some of the limitations of the earlier methods (e.g., self-organizing tree algorithm or SOTA (Herrero et al., 2001), quality-based clustering (Heyer et al., 1999), model-based clustering (Ghosh and Chinnaiyan, 2002; Yeung et al., 2001a), simulated annealing (Lukashin and Fuchs, 2001), gene shaving (Hastie et al., 2000), the cluster affinity search technique or CAST (Ben-Dor et al., 1999)). Also, some procedures were developed that could help the biologist to estimate some of the arbitrary parameters needed for the first generation of algorithms (e.g., like the number of clusters present in the data (Ghosh and Chinnaiyan, 2002; Lukashin and Fuchs, 2001; Yeung et al., 2001a)). We will discuss a selection of these clustering algorithms in more detail in the following sections. Many of these methods can be used with different distance measures, which can also have serious implications for the final result. One of the reasons that there are so many different clustering methods (sometimes giving very different results) is that, from a biological point of view, these different algorithms sometimes seem to expose different aspects present within the data and not always generate all the relevant clusters.

An important problem that arises when performing cluster analysis of gene expression profiles is the preprocessing of the data. Clustering implies more than just submitting the raw microarray data to the cluster algorithm of choice. A correct preprocessing strategy is almost as important as the cluster analysis itself. Normalization, non-linear transformations and management of missing values have been discussed in Chapter 3 and are equally important in this setting. Moreover, it is common to (crudely) filter the gene expression profiles (removing the profiles that do not satisfy a certain criterion - see further) before proceeding with the actual clustering (Eisen et al., 1998). A final customarily used pre-processing step is standardization or rescaling of the gene expression profiles (e.g., multiplying every expression vector with a scale factor so that their lengths are one - Quackenbush, 2001). This makes sense because the aim is to cluster gene expression profiles with the same relative behavior (expression levels go up and down at the same time) and not only the ones with the same absolute behavior. The two latter pre-processing steps will be discussed in more detail in the following sections.

Validation is another key issue when clustering gene expression profiles. When using existing algorithms or developing new ones it is not merely enough to submit the data to the algorithm and wait for the results. Cluster analysis is more than just producing clusters. The biologist using the algorithm is of course mainly interested in the biological relevance of these clusters and wants to use the results to discover new biological phenomena. This means that we need methods to (biologically and statistically) validate and objectively compare the results produced by new and existing clustering algorithms. Some standard methods for doing cluster validation have recently emerged (looking for enrichment of functional categories (Tavazoie et al., 1999), figure of merit or FOM (Yeung et al, 2001b), Rand index (Yeung and Ruzzo, 2001c), silhouette (Kaufman and Rousseeuw, 1990)) and will be discussed below. No real benchmark data set exists that can be used to unambiguously validate novel algorithms. However the yeast cell cycle data (Cho et al., 1998) as described in Appendix B is often used for this purpose.

4.3 Methods and algorithms

In this section some of the methods related to clustering gene expression will be discussed in more detail.

4.3.1 Specific preprocessing

Filtering

As stated in Section 4.2, a set of microarray experiments, generating gene expression profiles, frequently contains a considerable number of genes that do not really contribute to the biological process that is being studied. The expression values of these profiles often show little variation over the different experiments (they are called "constitutive" with respect to the biological process studied). Moreover, these constitutive genes will have seemingly random and meaningless profiles after standardization (division by a small standard deviation resulting in noise inflation), which is also a very common pre-processing step (see further). Another problem with microarray data sets is the fact that they regularly contain highly unreliable expression profiles with a considerable number of missing values. Due to their number, replacing these missing values in these expression profiles is not possible within the desired degree of accuracy.

The quality of the clusters would significantly degrade, if these data sets would be passed to the clustering algorithms as such. Most clustering algorithms assign every expression profile in the data to one of the clusters, even the ones of poor quality, corrupting the content and the average profile of these clusters making them less suitable for further analysis. A solution to this problem could be to use clustering algorithms that do not assign every profile to a cluster. The algorithm that is proposed in the next Chapter (AQBC) follows this approach. Another, more simple solution (that can also be used in combination with the previous solution), is to remove at least a fraction of the undesired genes from the data. This procedure is in general called filtering (Eisen et al., 1998). Filtering involves removing gene expression profiles from the data set that do not satisfy one or possibly more criteria. Commonly used criteria include a minimum threshold for the standard deviation of the expression values in a profile (removal of constitutive genes) and a threshold on the maximum percentage of missing values. Another similar method for filtering takes a fixed number or fraction of genes best satisfying one criterion (like the criteria stated above).

Standardization or rescaling

Biologists are mainly interested in grouping gene expression profiles that have the same relative behavior, i.e., genes that are up- and downregulated together. Genes showing the same relative behavior but with diverging absolute behavior (e.g., gene expression profiles with a different base line and/or a different amplitude but going up and down at the same time) will have a relatively high Euclidean distance. Cluster algorithms based on this distance measure will therefore wrongly assign these genes to different clusters.

Applying standardization or rescaling to the gene expression profiles can largely prevent this effect (Quackenbush, 2001). Gene expression profiles showing the same relative behavior will have a small(er) Euclidean distance after rescaling.

Consider a gene expression profile $g(g^1,g^2,...,g^j,...,g^e)$. Rescaling is commonly done by replacing every expression level g^j in g by

$$\frac{g^j - \mu}{\sigma}, \tag{4.1}$$

where μ is the average expression level of the gene expression profile and is given by:

$$\mu = \frac{\sum_{j=1}^{e} g^j}{e} \tag{4.2}$$

and σ is the standard deviation given by:

$$\sigma = \sqrt{\frac{1}{e-1}\sum_{j=1}^{e}\left(g^{j}-\mu\right)^{2}}. \tag{4.3}$$

This is repeated for every gene expression profile in the data set and results in a collection of expression profiles all having average zero and standard deviation one (i.e., the absolute differences in expression behavior have been largely removed). The division by the standard deviation is sometimes omitted (rescaling is then called mean centering).

4.3.2 Clustering algorithms

As already stated, several clustering methods (first and second generation algorithms) are available. We will discuss some of the important ones in more detail below.

First-generation algorithms

Not withstanding some of the disadvantages of these early methods, it has to be noted that many good implementations of these algorithms were already developed outside biologically related research and are ready to be used by biologists (which is not always the case with the newer methods) - see also Table 4.2.

a) Direct visual inspection:

This is of course the most simple and direct approach used by many biologists in the early days of gene expression analysis (Cho et al., 1998). This method is best suited where the patterns of interest are known in advance, but does not work for larger data sets (high number of dimensions or data points) or when one hopes to discover unexpected patterns.

b) Hierarchical clustering

Hierarchical clustering is the most widely used method for clustering gene expression data (Eisen et al., 1998; Quackenbush, 2001; Sherlock, 2000) and can be seen as the *de facto* standard. Hierarchical clustering has the advantage that the results can be nicely visualized (see Figure 4.2). This method can also be used to cluster entire microarray experiments (columns of the expression matrix - see Chapter 3, Section 3.4.1). For more information and a description of the possible algorithms, see Appendix A, Section A.6. Using this method, clusters are formed by cutting a tree structure at a certain level or height. This level corresponds to a certain pairwise distance, which in its turn is rather arbitrary (it is difficult to predict which level will give the most valid biological results). Finally, the computational complexity of hierarchical clustering is quadratic in the number of gene expression profiles, which can be a problem when considering the current size of the data sets.

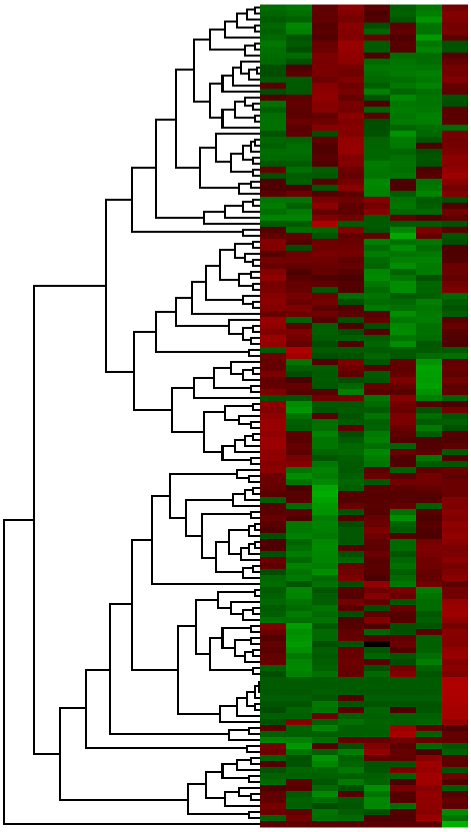


Figure 4.2: Typical result we obtained from an analysis using hierarchical clustering using 137 gene expression profiles of dimension 8. The left side of the figure represents the tree structure. The terminal branches of this tree are linked with the individual genes and the height of all the branches is proportional to the pairwise distance between the clusters. The right side of the figure (also called a heat map) corresponds to the expression matrix where each row represents a gene expression profile, each column a microarray experiment and the individual values are represented on a colour (green to red) or grey scale.

c) K-means clustering

K-means clustering of gene expression profiles (Tavazoie et al., 1999; Tou and Gonzalez, 1979) results in a partitioning of the data (every gene expression profile belongs to exactly one cluster) using a predefined number *K* of partitions or clusters (see Figure 4.3). K-means clustering was also applied to clustering microarray experiments and the algorithm is described in Appendix A, Section A.5. The predefinition of the number of clusters by the user is also rather arbitrary (it is very difficult to predict the number of clusters in advance). In practice, this makes it necessary to use a trial-and-error approach where a comparison and biological validation of several runs of the algorithm with different parameter settings is necessary.

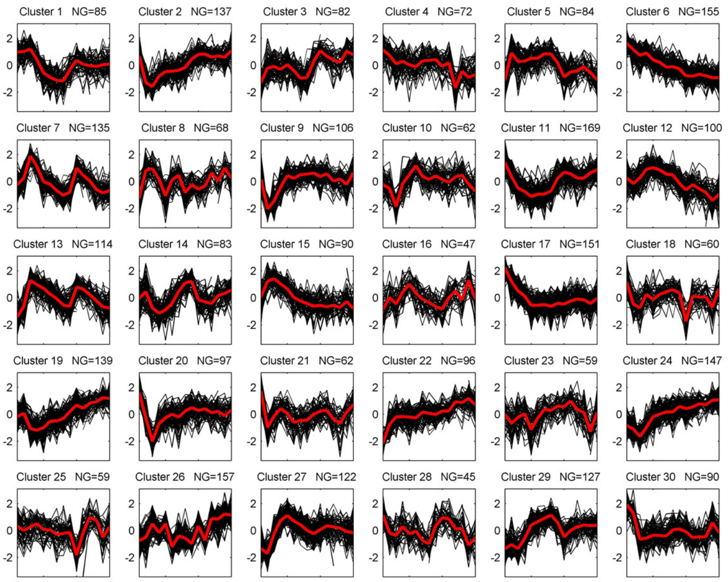


Figure 4.3: Typical result from an analysis using K-means clustering with 30 clusters using 3000 standardized expression profiles of dimension 15 (yeast cell cycle data - filtering: 3000 expression profiles with the highest standard deviation before standardization were chosen). The sum of the number of genes in each cluster equals the total number of genes submitted to the algorithm (=3000). NG = Number of Genes. Each plot shows the individual expression profiles and the mean expression profile of a cluster.

d) Self-organizing maps (SOM)

In SOM (Kohonen, 1997; Tamayo et al., 1999), the user has to predefine a topology or geometry of nodes (e.g., a two-dimensional grid - one node for each cluster), which again is not really straightforward. These nodes are then mapped into the gene expression space, initially at random and iteratively adjusted. In each iteration, a gene expression profile is randomly picked and the node that maps closest to it is selected. The mapping of this selected node is then moved into the direction of the selected expression profile. The mapping of the other nodes is also moved into the direction of the selected expression profile but to an extent proportional to the distance from the selected node in the initial two-dimensional node topology.

Second-generation algorithms

In this section we will describe several of the newer clustering methods that have specifically been designed to cluster gene expression profiles.

a) *Self-organizing tree algorithm:*

The SOTA (Herrero et al., 2001) combines both self-organizing maps and divisive hierarchical clustering. The topology or node geometry here takes the form of a dynamic binary tree. Similar to self-organizing maps, the gene expression profiles are sequentially and iteratively presented to the terminal nodes (located at the base of the tree - these nodes are also called cells). Subsequently, the gene expression profiles are associated with the cell that maps closest to it and the mapping of this cell plus its neighbouring nodes are updated (moved into the direction of the expression profile). The presentation of the gene expression profiles to the cells continues until convergence. After convergence the cell containing the most variable population of expression profiles (variation is defined here by the maximal distance between two profiles that are associated with the same cell) is split in two sister cells (causing the binary tree to grow) where after the entire process is restarted. The algorithm stops (the tree stops growing) when a threshold of variability is reached for each cell. To obtain a statistical definition for this threshold a randomised version of the entire data set is used (for each expression profile all its expression values are randomly and independently shuffled - this operation destroys the actual correlation between expression profiles) and the distances between all possible pairs of gene expression profiles in this version of the data are calculated. This results in the probability distribution of the distances that could occur by chance (i.e., the distribution that describes the probability that two unrelated expression profiles have a certain distance). The threshold of variability can now be defined by choosing a confidence level α (e.g., α=5%), so that only a fraction α of the randomised gene expression profiles have a distance smaller than this threshold. Using this threshold ensures that the fraction of misassignments (unrelated profiles assigned to the same cluster) in the actual cluster result is limited by the α-value.

The approach described by Herrero et al. (2001) has some properties that make it potentially useful for clustering gene expression profiles:

i. The clustering procedure itself is linear in the number of gene expression profiles (compare this with the quadratic complexity of standard hierarchical clustering).

ii. The number of clusters does not have to be known in advance. Moreover, Herrero et al. describe a statistical procedure to stop growing the tree. Therefore, the user is freed from choosing a (arbitrary) level where the tree has to be cut (like in standard hierarchical clustering).

iii. A server running the program is available (see Table 4.2).

In our opinion, this method, however, also has some disadvantages:

i. The procedure for finding the threshold of variability is time-consuming since it involves the actual construction of a randomised data set and the calculation of the distances between all possible pairs of randomised expression profiles (quadratic!). The entire process described in Herrero et al. (2001) is thus in fact quadratic in the number of gene expression profiles.

ii. No biological validation was provided showing that this algorithm indeed produces biologically relevant results.

b) Model-based clustering

Model-based clustering (Fraley and Raftery, 1999; Ghosh and Chinnaiyan, 2002; Yeung et al., 2001a) is an approach that is not really new and has already been used in the past for other applications outside bioinformatics. Its potential use for cluster analysis of gene expression profiles has been proposed only recently, however. In the context of clustering gene expression profiles we will thus treat it as a second-generation algorithm.

Model-based clustering assumes that a finite mixture of underlying probability distributions, where each distribution represents one cluster, generates the data. Usually, multivariate normal distributions are used for these probability distributions. In this case, each cluster C_k is represented by a multivariate Gaussian model p_k in e dimensions:

$$p_k(g \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{e}{2}} \mid \Sigma_k \mid^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(g-\mu_k)^T \Sigma_k^{-1}(g-\mu_k) \right), \qquad (4.4)$$

where g is a gene expression profile or vector and μ_k and Σ_k the mean and covariance matrix of the multivariate normal distribution respectively. The covariance matrix Σ_k can be represented by its eigenvalue decomposition, which in this case is written as follows:

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \qquad (4.5)$$

where D_k is the orthogonal matrix of the eigenvectors of Σ_k, A_k is a diagonal matrix whose elements are proportional to the eigenvalues of Σ_k and λ_k is the constant of proportionality. This decomposition implies a nice geometric interpretation of the clusters: D_k controls the orientation, A_k controls the shape and λ_k controls the volume of the cluster. Simpler forms for the covariance structure can be used (e.g., by having some of the parameters take the same values across clusters), decreasing the number of parameters that have to be estimated but also decreasing the model flexibility (capacity to model more complex data structures). The mixture model p itself takes then the following form:

$$p(g) = \sum_{k=1}^{K} \pi_k . p_k (g \mid \mu_k, \Sigma_k) \tag{4.6}$$

where K is the number of clusters and π_k is the prior probability that an expression profile belongs to cluster C_k so that:

$$\sum_{k=1}^{K} \pi_k = 1 \tag{4.7}$$

and

$$\pi_k \geq 0. \tag{4.8}$$

In practice we would like, given a collection of expression profiles $\{g_i\}_{i=1,\ldots,n}$ to estimate all the parameters (π_k, μ_k, Σ_k (k=1,…,K) and K itself) of this mixture model. In a first step these parameters are estimated with an EM-algorithm using a fixed value for K and a fixed covariance structure. This parameter estimation is then repeated for different values for K and different covariance structures. The result of the first step is thus a collection of different models fitted to the data and all having a specific value for K and a specific covariance structure. In a second step the best model in this group of models is selected (i.e., the most appropriate number of clusters and a covariance structure is chosen here). This model selection step involves the calculation of the Bayesian Information Criterion (BIC; Schwartz, 1978) for each model, which is not further discussed here.

A good implementation for model-based clustering (called MCLUST - Fraley and Raftery, 1999) is available (see Table 4.2). Yeung et al. (2001a) reported good results using this software on several synthetic data sets and real expression data sets. They claimed that the performance of MCLUST on real expression data was as least as good as could be achieved with a heuristic cluster algorithm (CAST - Ben-Dor et al. (1999) - not discussed here).

c) *Quality-based clustering*

In Heyer et al. (1999) a clustering algorithm (called QT_Clust - also see Table 4.1 for the basic steps of this approach) is described that produces clusters in a set of gene expression profiles $G = \{g_i\}_{i=1,\ldots,n}$ that have a quality guarantee that ensures that all members of a cluster should be coexpressed with all other members of this cluster. Heyer et al. define the quality of a cluster C as the maximum of the distance $d(g_k,g_l)$ between two gene expression vectors g_k and g_l of C (called the diameter of C). Heyer et al. use a specific distance measure (jackknife correlation - not further discussed here) but the method can be easily be extended to other distance measures. The quality guarantee itself is defined as a fixed and user-defined threshold D for the quality or diameter of each cluster.

Briefly said, the aim of QT_Clust is to find clusters, with a quality guarantee, containing a maximum number of expression profiles. It considers every expression profile in the data set as a cluster seed (one could call this a cluster center) and iteratively assigns the expression profiles to these clusters that cause a minimal increase in diameter until the diameter threshold D (=quality guarantee) is reached. At this stage every expression profile is made available to every candidate cluster and there are initially as many candidate clusters as there are expression profiles. At this point, the candidate cluster that contains the highest number of expression profiles is selected as a valid cluster and removed from the data set where after the whole process starts again. The algorithm stops when the number of points in the largest remaining cluster falls below a prespecified threshold (*MIN_NR_GENES*). This stop criterion implies that the algorithm will terminate before all expression profiles are assigned to a cluster.

This approach was designed with cluster analysis of expression data in mind and has some properties that could make it very useful for this task:

i. By using a stringent quality guarantee it is possible to find clusters with tightly related expression profiles (containing highly coexpressed genes). These clusters might therefore be good ‘seeds’ for further analysis.

ii. Genes not really coexpressed with other members of the data set are not included in any of the clusters.

There are, however, also some disadvantages:

i. The quality guarantee of the clusters is a user-defined parameter that is hard to estimate and too arbitrary. This

Table 4.1: Quality-based clustering algorithm (QT_Clust) proposed by Heyer et al. (1999)

```
QT_Clust ($G = \{g_i\}_{i=1,\dots,n}$, $MIN\_NR\_GENES$ , $D$)
FOR ALL $g_i \in G$          /* Consider every expression profile as a seed for
                              candidate cluster $C_i$ */
        $C_i = \{g_i\}$
        $FLAG$ = true
        WHILE $FLAG$ = true AND $C_i \neq G$
                FIND $g_a \in (G \setminus C_i)$ that minimizes
                        $\mathrm{Diam}(C_i \cup \{g_a\}) = \max\{d(g_k,g_l) | g_k, g_l \in (C_i \cup \{g_a\})\}$
                /* Find expression profile that causes minimal increase in
                diameter of $C_i$ */
                IF $\mathrm{Diam}(C_i \cup \{g_a\}) > D$
                        FLAG = false
                        /* Cluster $C_i$ stops growing if diameter threshold is
                        reached */
                ELSE
                        $C_i = C_i \cup \{g_a\}$
                END IF
        END WHILE
END FOR
FIND $C \in \{C_1, C_2, \dots, C_n\}$ such that $\#C$ is maximal
/* Select candidate cluster with maximum number of expression profiles */
IF $\#C < MIN\_NR\_GENES$
        STOP    /* Stop algorithm if number of elements of selected cluster
                falls below threshold */
ELSE
        OUTPUT $C$
        QT_Clust ($G \setminus C$, $MIN\_NR\_GENES$, $D$)        /* Find next cluster */
END IF
```

method is therefore, in practice, hard to use by biologists and extensive parameter fine-tuning is necessary.

ii. This algorithm produces clusters all having the same fixed diameter not optimally adapted to the local data structure.

iii. The computational complexity is quadratic in the number of expression profiles.

iv. No ready to use implementation is available.

Table 4.2: Availability of clustering algorithms

Package	URL
Cluster	http://rana.lbl.gov/EisenSoftware.htm
J-Express	http://www.molmine.com
Expression Profiler	http://ep.ebi.ac.uk
SOTA	http://bioinfo.cnio.es/sotarray
MCLUST	http://www.stat.washington.edu/fraley/mclust
AQBC (see Chapter 5)	http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html

4.3.3 Cluster validation

As mentioned before, depending on the pre-processing, the algorithms and the different distance measures, clustering will produce different results. Even random data often produces clusters. Therefore validation of the relevance of the cluster results is of utmost importance. Below, we will describe four methodologies that are often used for this task.

Looking for enrichment of functional categories: Biological validation

One way to validate results from clustering algorithms is to compare the gene clusters with existing functional classification schemes. In such schemes, the genes are allocated to one or more functional categories representing their biochemical properties and biological roles (Tavazoie et al., 1999). Finding clusters that have been significantly enriched for genes with similar function is proof that a specific clustering technique can produce biologically relevant results. Therefore the method discussed here is often called biological validation.

As stated in Section 4.2, the yeast cell cycle data (Cho et al., 1998) described in Appendix B is often used as a benchmark data set. One of the reasons is that the majority of the genes included in the data have been functionally classified (Mewes et al., 2000). A functional classification scheme is available (MIPS database - see http://mips.gsf.de/genre/proj/yeast/index.jsp), which makes it possible to biologically validate the results.

Assume that a certain clustering method finds a set of clusters in this data. We could objectively look for functionally enriched clusters as follows: Suppose that one of the clusters has *g* genes where *k* genes belong to a certain functional category in the MIPS database and suppose that this functional category in its turn contains *f* genes in total. Also suppose that the total data set contains *n* genes (in the yeast cell cycle data *n* would be 6220). Using the cumulative hypergeometric probability distribution, we could calculate the probability or p-value that this degree of enrichment could have occurred by chance, i.e., what is the probability of finding at least *k* genes in this specific cluster containing *g* genes from this specific functional category containing *f* genes (out of the whole *n* annotated genes) by chance:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{n-f}{g-i}}{\binom{n}{g}} = \sum_{i=k}^{\min(g,f)} \frac{\binom{f}{i}\binom{n-f}{g-i}}{\binom{n}{g}}. \tag{4.9}$$

These p-values can be calculated for each functional category in each cluster. Since there are about 200 functional categories in the MIPS database, only clusters where the p-value is smaller than 0.0003 for a certain functional category, are said to be significantly enriched (level of significance 0.05). These p-values can also be used to compare the results from functionally matching clusters identified by two different clustering algorithms on the same data. For an example see Chapter 5, where the results of our clustering algorithm have been validated using this method. Also see Appendix A, Section A.1 for more explanation about p-values.

Rand index: validation against an external criterion

The Rand index (Yeung, 2001b; Yeung and Ruzzo, 2001c) is a measure that reflects the level of agreement of a cluster result with an external criterion, i.e. an existing partition or a known cluster structure of the data. This external criterion could for example be an existing functional categorization (see previous method), a predefined cluster structure if one is clustering synthetic data where the clusters are known in advance, or another cluster result obtained using other parameter settings for a specific clustering algorithm or obtained using other clustering algorithms. The latter could be used to investigate how sensitive a cluster result is to the choice of the algorithm or parameter setting. If this result proves to be relatively stable, one can assume that pronounced signals are present in the data possibly reflecting biological processes.

Suppose we want to compare two partitions (the cluster result at hand and the external criterion) of a set of *n* genes. Suppose that *a* is the number of gene pairs that are placed in the same subset (or cluster) in both partitions. Suppose that *d* is the number of gene pairs that are placed in different subsets in both partitions. The Rand index is then defined as the fraction of agreement between both partitions:

$$\frac{a+d}{\binom{n}{2}}. \tag{4.10}$$

The Rand index lies between 0 and 1 (1 if both partitions are identical). The adjusted Rand index has similar properties but is not further discussed here.

Testing cluster coherence: Silhouette

A gene expression profile can be considered to be well clustered if its distance to the other expression profiles of the same cluster is small and the distance to the expression profiles of other clusters is larger. This criterion can be formalized by using silhouettes (Kaufman and Rousseeuw, 1990). This measure validates the cluster result on statistical grounds only (statistical validation). Biological information is not used here.

Suppose g_i is an expression profile that belongs to cluster C_k. Call $v(g_i)$ (also called the within dissimilarity) the average distance of g_i to all other expression profiles from C_k. Suppose C_l is a cluster different from C_k and define $d(g_i,C_l)$ as the average distance from g_i to all expression profiles of C_l. Now define $w(g_i)$ (also called the between dissimilarity) as follows:

$$w(g_i) = \min_{C_l \neq C_k} d(g_i, C_l). \tag{4.11}$$

The silhouette $s(g_i)$ of g_i is now defined as follows:

$$s(g_i) = \frac{w(g_i) - v(g_i)}{\max\{v(g_i), w(g_i)\}}. \tag{4.12}$$

Note that -1 ≤ $s(g_i)$ ≤ 1. Consider two extreme situations now. Firstly, suppose that the within dissimilarity $v(g_i)$ is significantly smaller than the between dissimilarity $w(g_i)$. This is the ideal case and $s(g_i)$ will be approximately 1. This occurs when g_i is 'well clustered' and there is little doubt that g_i is assigned to an appropriate cluster. Secondly, suppose that

$v(g_i)$ is significantly larger than *$w(g_i)$*. Now *$s(g_i)$* will be approximately -1 and g_i has in fact been assigned to the wrong cluster (worst case scenario).

We can now define two other measures: the average silhouette width of a cluster and the average silhouette width of the entire data set. The first is defined as the average of *$s(g_i)$* for all expression profiles of a cluster and the second is defined as the average of *$s(g_i)$* for all expression profiles in the data set. This last value can be used to mutually compare different cluster results and can be used as an inherent part of clustering algorithms, if its value is optimised during the clustering process.

Figure of merit

"Figure of merite" or FOM (Yeung et al., 2001b) is a simple quantitative data-driven methodology (statistical validation) that also allows comparisons to be made between outputs of different cluster algorithms. The methodology is related to jackknife-based or leave-one-out cross-validation. The method goes as follows. The clustering algorithm (for the gene expression profiles) to be tested is applied to all experimental conditions (in this case the data variables) except for one left-out condition. If the algorithm performs well, we expect that if we look at the genes from a given cluster, their values for the left-out condition will be highly coherent. Therefore, we compute the FOM, for the left-out condition, as the root mean square of the deviations of each gene relative to the mean of the genes in its cluster for this condition. The FOM measures the within-cluster similarity of the expression values of the removed experiment and therefore reflects the predictive power of the clustering. It is expected that removing one experiment from the data should not interfere with the cluster output if the output is robust. For cluster validation, each condition is subsequently used as a validation condition and the aggregate FOM (sum of the all the FOM) over all conditions is used to compare cluster algorithms.

4.4 Conclusion

In this chapter a general overview of clustering gene expression profiles and a discussion of the specific requirements related to this task was given. We described a selection of the first- (hierarchical clustering, K-means, SOM) and second-generation (SOTA, model-based and quality-based clustering) algorithms that are frequently used to solve this problem and discussed some of the preprocessing steps like filtering and standardization that are customarily associated with these methods. Since the aim of clustering expression profiles is to discover new biology, we discussed some of the methods (looking for enrichment of functional categories, Rand index, silhouette and FOM) that can be used to biologically or statistically validate the resulting clusters and to objectively show that the output of the algorithms at hand indeed produce relevant clusters. We noted that some of the algorithms have properties that make them less suited for clustering gene expression profiles. This includes the necessity to choose an arbitrary parameter setting or to perform extensive parameter fine-tuning, the inclusion of all the genes in the clusters, a high computational complexity and the lack of biological or other validation.

To solve some of the difficulties associated with clustering gene expression profiles, we developed an algorithm that is called adaptive quality-based clustering and that starts from the principles introduced in quality-based clustering by Heyer et al. This method is described and fully validated in the next chapter.

Chapter 5

Adaptive quality-based clustering of gene expression profiles

5.1 Introduction[1]

In the previous chapter we noted that algorithms for clustering gene expression profiles have special requirements and that the classical algorithms suffer from some drawbacks that make them less appropriate for this task. In this chapter we will present a specific solution to this challenge.

As said, much effort is currently being done to adapt clustering algorithms towards the specific needs of biological problems. In this context the idea of quality-based clustering (Heyer et al., 1999 - see Section 4.3.2) was developed. Heyer et al. proposed an algorithm (which they called QT_Clust) that tries to identify clusters that have a certain quality or diameter (representing the minimal degree of coexpression needed - see below for the exact definition used in this chapter) and where every cluster contains a maximal number of points. Genes not exhibiting this minimal degree of coexpression with any of the clusters are excluded from further analysis. A problem with the quality-based approach of Heyer et al., however, is that this quality is a user-defined parameter that is hard to estimate (it is hard to find a good trade-off or optimal value: setting the quality too strictly will exclude a considerable number of coexpressed genes, setting it too loosely will include too many genes that are not really coexpressed). Moreover, it should be noted that the optimal value for this quality is, in general, different for each cluster and data set dependent. The computational complexity of this approach is quadratic in the number of gene expression profiles.

[1] The discussion presented in this Chapter has been published as a full paper in Bioinformatics (De Smet et al., 2002).

In this chapter, we describe an adaptive quality-based clustering method starting from the principles described by Heyer et al. (quality-based approach; locating clusters, with a certain quality, in a volume where the density of points is maximal). The algorithm described below is in essence a heuristic, two-step approach that defines the clusters sequentially (the number of clusters is not known in advance, so it is not a parameter of the algorithm). The first step locates a cluster (quality-based approach) and the second step derives the quality of this cluster from the data (adaptive approach). We will make an assessment of the computational complexity of this approach and the performance of the algorithm is validated on real and artificial microarray data. We will make a theoretical comparison between our algorithm, the algorithm of Heyer et al., hierarchical clustering, K-means and self-organizing maps. Finally, we will refer to an on-line tool for integrated clustering, upstream sequence retrieval and motif sampling (INCLUSive) in which our algorithms has been integrated.

5.2 General methodology

5.2.1 Standardization

As previously mentioned, it is common practice to standardize gene expression vectors before cluster analysis so that their mean is zero and their variance is one before proceeding with the actual cluster algorithm. If $g_i(g_i^1,g_i^2,...,g_i^j,...,g_i^e)$ is a standardized expression vector, this means that:

$$\frac{1}{e}\sum_{j=1}^{e} g_i^{\,j} = 0, \tag{5.1}$$

$$\sqrt{\frac{1}{e-1}\sum_{j=1}^{e}\left(g_i^{\,j}\right)^2} = 1. \tag{5.2}$$

Standardized expression profiles or vectors therefore are located in an e-dimensional space on the intersection of a hyperplane (Equation 5.1) and a hypersphere with a radius equal to $\sqrt{(e\text{-}1)}$ (Equation 5.2).

5.2.2 Quality *R* of a cluster

The definition used in this chapter for the quality R of a cluster is slightly different from the definition proposed by Heyer et al. and is as follows: In a collection of gene expression profiles $G=\{g_i, i=1,\ldots,n\}$, a cluster C_k with center O_k (center not necessarily standardized) and quality R_k (also called radius of cluster C_k), will only contain the profiles satisfying the following property:

$$\left\|g_i - O_k\right\|_2 < R_k. \tag{5.3}$$

Equation 5.3 means that cluster C_k only contains gene expression profiles with a minimum degree of coexpression (represented by the quality guarantee R_k). The norm or distance measure we use here is the 2-norm or Euclidean distance.

5.3 Algorithm

5.3.1 Global algorithm

The global cluster algorithm (AQBC - see Table 5.1) is, as mentioned previously, a heuristic iterative two-step approach where the basic steps are as described below. In this implementation we use two user-defined parameters (*MIN_NR_GENES* and *S* - the values between brackets are default values), several internal tuning parameters that have a fixed value (the user is not allowed to change these values) and the data set itself (*G*).

During each iteration, this algorithm first finds a cluster center (O_k) using a preliminary estimate (R_k*_PRELIM*) of the radius or quality of the cluster (Step 1). When the cluster center has been located, the algorithm determines a new estimate for the radius (R_k) of the cluster (Step 2). Now there are two possibilities:

1. If this new estimate is approximately equal to the preliminary estimate (e.g., within 10% - *ACCUR_RAD*), the set of genes defined by the cluster center and the new estimate of the radius is removed from the data set *G*. Furthermore, if the number of genes in this set is equal or larger than a predefined value (*MIN_NR_GENES* - user-defined; default 2), this set is a valid cluster. The preliminary estimate of the radius to be used in Step 1 of the next iteration (for the next cluster) is updated with the new estimate of the radius calculated in Step 2 of the current iteration (in most cases, the best preliminary estimate for the radius of the next cluster seems to be the radius of the previous cluster).

2. If the new estimate of the radius is substantially different from the preliminary estimate, the preliminary estimate *R_k_PRELIM* is also updated with the new estimate R_k and a new iteration is started. This is repeated until the relative difference between R_k and *R_k_PRELIM* falls under *ACCUR_RAD*.

Table 5.1: Global cluster algorithm. The values between brackets are the default values for the user-defined parameters

```
AQBC ($G = \{g_i, i=1,\ldots,n\}$, MIN_NR_GENES <2>, S <0.95>)
ACCUR_RAD = 0.1                 /* Set internal tuning parameter */
Initialise $R_k$_PRELIM          /* Radius estimate initialisation */
WHILE Stop criterion NOT TRUE
        $O_k$ = locate_cluster_center (G, $R_k$_PRELIM)
                /* Localisation of a cluster center - Step 1*/
        $R_k$ = recalculate_radius (G, $O_k$, $R_k$_PRELIM, S)
                /* Re-estimation of radius - Step 2 */
        IF ( | $R_k$ - $R_k$_PRELIM | / $R_k$_PRELIM) < ACCUR_RAD
                /* Check accuracy of radius estimation */
                CLUSTER = $\{g_i \in G \mid \|g_i - O_k\| < R_k\}$
                G = G \ CLUSTER         /* Remove cluster from data set G */
                IF #CLUSTER >= MIN_NR_GENES         /* Valid cluster ? */
                        Output CLUSTER
                END IF
        END IF
        $R_k$_PRELIM = $R_k$                  /* Update radius estimate */
END WHILE
```

The iterations are terminated when the stop criterion is satisfied (see Section 5.3.5).

The algorithm was implemented in MATLAB. This implementation uses the method described in Chapter 3 (Section 3.2.3 - missing value management without replacement (Kaufman and Rousseeuw, 1990)) to deal with missing values often occurring in expression data.

Below we will discuss the initialisation of the preliminary estimate of the radius before the first iteration, the procedures used in Step 1 and 2, the stop criterion (WHILE loop) and the computational and memory complexity of the overall algorithm.

5.3.2 Radius estimate initialisation

In the global cluster algorithm, the preliminary estimate of the radius (*R_k_PRELIM*) has to be initialised before the first iteration (radius estimate for the first cluster - line 3 of AQBC). We use half of the radius of the hypersphere defined by standardization of the expression profiles (see above in Section 5.2.1). This is given by:

$$R_k\_PRELIM = \frac{\sqrt{e-1}}{2} \tag{5.4}$$

where *e* is the dimension of the gene expression vectors (number of expression vector components).

5.3.3 Localization of a cluster center - quality-based clustering (Step 1)

Given a collection *G* of gene expression profiles, the objective of Step 1 is to find a cluster center in an area of the data set where the ‘density’ (or number) of expression profiles, within a sphere with radius or quality equal to *R_k_PRELIM* (preliminary estimate of the radius), is locally maximal. The method described here is based on the principles used by Heyer et al. but is significantly faster (also see the discussion in Table 5.6). The disadvantage with this approach is that the quality or radius of the clusters is a parameter that is not very intuitive (it is often hard to find a ‘good’ value for this parameter; often a trial-and-error approach is used with manual validation of the clusters). Furthermore, all the clusters are forced to have the same radius.

The basic steps of the algorithm used for the first step are described in Table 5.2 (locate_cluster_center). After initialisation of the cluster center (with the mean profile of all the expression profiles in the data set *G*), all the expression profiles within a sphere with radius *RAD* are selected. Iteratively, the mean profile of these expression profiles is calculated and subsequently the cluster center is moved to this mean profile. This approach moves the cluster in the direction where the ‘density’ of profiles is higher (conceptually visualised in Figure 5.1).

Table 5.2: Algorithm for the localization of a cluster center.

```
$O_k$ = locate_cluster_center ($G$, $R_k$_PRELIM)
MAXITER = 50
/* Set internal tuning parameter - maximum number of iterations */
DIV = 1/30
/* Set internal tuning parameter - fraction needed to determine DELTARAD */
$O_k$ = mean ($G$)          /* Cluster center initialisation */
RAD = max $\{\|g_i - O_k\| \mid g_i \in G\}$          /* Start with maximal radius */
DELTARAD = (RAD - $R_k$_PRELIM) * DIV
/* Determine step for decreasing radius */
RAD = RAD - DELTARAD          /* Decrease radius */
GENES_IN_SPHERE = $\{g_i \in G \mid \|g_i - O_k\| < RAD\}$
/* Determine profiles within sphere */
ME = mean (GENES_IN_SPHERE)          /* Recalculate mean */
ITER = 1
WHILE ($ME \neq O_k$ AND ITER < MAXITER) OR RAD > $R_k$_PRELIM
/* Iterate until convergence or maximal number of iterations has been reached */
        ITER = ITER + 1
        $O_k$ = ME          /* Move cluster center to cluster mean */
        IF RAD > $R_k$_PRELIM
                RAD = RAD - DELTARAD                    /*
                Decrease radius if desired quality has not been reached */
        END IF
        GENES_IN_SPHERE = $\{g_i \in G \mid \|g_i - O_k\| < RAD\}$
        /* Determine profiles within sphere */
        ME = mean (GENES_IN_SPHERE)          /* Re-calculate mean */
END WHILE
IF $ME \neq O_k$
        $O_k$ = empty          /* Undefined cluster center if no convergence */
END IF
```

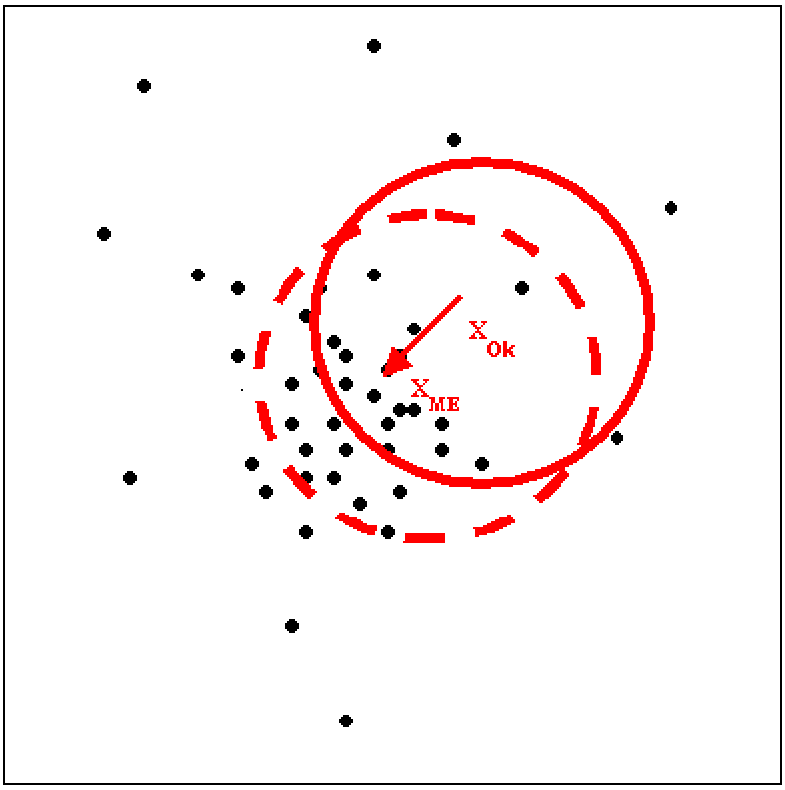


Figure 5.1: Conceptual visualisation of cluster center (X_{Ok}) relocation to the cluster mean (X_{ME}) in two dimensions (one iteration - cluster radius constant - data not standardized). The number of profiles (black dots) within the sphere after relocation is substantially higher than the number of profiles before relocation.

The radius *RAD* of the sphere is initialised so that all profiles in the data set are located within this sphere. Every iteration, this radius is decreased with a constant value (*DELTARAD*, a fraction (*DIV*) of the difference between the initial value of *RAD* and R_k*_PRELIM*) until the radius has reached the desired value (R_k*_PRELIM*) and then remains constant. In the first iterations (when *RAD* is still 'large') this technique will move the cluster center to regions of the data where the 'global' density is higher (these regions often contain the largest cluster(s)). After some iterations (when *RAD* is equal or close to R_k*_PRELIM*) the cluster center will move towards an actual cluster where the density is 'locally' higher.

Convergence is reached if the cluster center remains stationary after *RAD* has reached R_k*_PRELIM*. If this does not happen within a certain (*MAXITER*) number of iterations, O_k is emptied and the algorithm stops.

The number of distance calculations performed during each iteration of locate_cluster_center is equal to the number (= n) of all expression profiles in G (only the distances from the expression profiles to the current cluster center have to be calculated). Note also that the computational complexity of the calculation of one distance is O(e) (e is the dimensionality of the expression vectors). Because the number of iterations is limited (*MAXITER*), the computational complexity for the localization of one cluster center is thus O($n \times e$).

5.3.4 Re-estimation or adaptation of the cluster quality (Step 2)

In Step 1 of the algorithm we located a cluster center O_k in a collection G of gene expression profiles, using a preliminary estimate R_k*_PRELIM* of the radius of the cluster. The objective of the method described in this paragraph is, given the cluster center that remains fixed, to re-calculate the radius R_k of the current cluster as to assess that genes belonging to this cluster are significantly coexpressed.

To substantiate the method described here, we introduce a randomised version of the original data set where the components of each expression vector are randomly and independently permuted (Herrero et al., 2001). This randomised version of the data will only be used for conceptual reasons, it will not be used during the actual calculations. This process of randomisation destroys the correlation between the expression vectors that was introduced through non-accidental mechanisms (e.g., experimental setup). Any correlation still existing after this procedure can be attributed to chance.

First, we calculate the Euclidean distance r from every expression vector in the data set to the cluster center O_k. Imagine doing the same for every vector present in the randomised data. The distribution of these distances in the original data consists of two parts - see Figure 5.2:

1. Background: these are the expression profiles with a distance to the cluster center that is also significantly present in the distance distribution of the randomised data set. Genes belonging to the background of the current cluster center either do not belong to any cluster (noise; are not significantly coexpressed with other genes) or belong to another cluster. Genes belonging to other clusters (if not too dominant) will not significantly show up in the distance distribution for the current cluster center (they ‘drown’ in the noise or background).

2. Cluster: these are the expression profiles with a distance to the cluster center that is not significantly present in the distance distribution of the randomised data set (left-sided tail in the distribution of the original data set). Genes belonging to the cluster are significantly coexpressed.

To calculate the true radius of the cluster we need to construct a model (probability density estimation) describing the total distribution of the distance *r* in the original data. We propose the following model structure:

$$p(r) = P_C.p(r \mid C) + P_B.p(r \mid B) \tag{5.5}$$

where

$$P_C + P_B = 1. \tag{5.6}$$

The model structure assumes that the distance measure used for *r* is the Euclidean distance. This means that our method cannot be directly extrapolated to other distance measures.

The model for the total distribution described in Equation 5.5 is the sum of two terms (also see Figure 5.2). The first term represents the distribution of the cluster, the second term represents the distribution of the background, each multiplied by the associated a priori probability (P_C and P_B). As we will see further, this model is only valid for standardized gene expression vectors. Note also that this model is an approximation and only reliable in the neighbourhood of the cluster. Below we will discuss how *p(r|C)* and *p(r|B)* are constructed, how the parameters of the model are determined and how we will use this model to calculate the radius of the cluster.

Distribution of r in the cluster: *p(r|C)*

Assume that all the gene expression vectors g_i in *G* are standardized and therefore are located in an *e*-dimensional space on the intersection of a hypersphere (with a radius equal to √(*e*-1) (Equation 5.2)) and a hyperplane (Equation 5.1) going through the center of the hypersphere. The intersection itself (we will further refer to it as *H*) can therefore be seen as a curved space with an intrinsic dimensionality of *d* = *e*-2 (*H* itself is a hypersphere with radius √(*e*-1) located in the (*e*-1)-dimensional space defined by the hyperplane). We simplify the problem by neglecting the curved nature of *H* in the neighbourhood of the cluster (we assume the hypersphere to be locally flat - said otherwise, we linearise *H* in the neighbourhood of the cluster - we will refer to this linearised version of *H* as H_L). This approximation also implies that the cluster center O_k belongs to H_L and that the Euclidean distances to the cluster center measured in H_L are equal to the real Euclidean distances (= r) to the cluster center. The equations derived in this section are therefore an approximation and thus only reliable close to the current cluster center O_k ($r < \sqrt{}(e\text{-}1)$ = radius H), which is sufficient for our purpose, because we are only interested in modelling the area where the cluster is situated.

The cluster is assumed to be normally distributed around O_k within H_L (the variance is hypothesised to be equal in each direction (in H_L) and given by σ^2). This means that the probability of finding an expression vector g_i of the cluster in an elementary volume dV of H_L is given by (Bishop, 1995)

$$\frac{1}{\left(2\pi\sigma^2\right)^{d/2}}\exp\left(-\frac{\left\|g_i - O_k\right\|}{2\sigma^2}\right)\mathrm{d}V = \frac{1}{\left(2\pi\sigma^2\right)^{d/2}}\exp\left(-\frac{r^2}{2\sigma^2}\right)\mathrm{d}V, \tag{5.7}$$

where r is the Euclidean distance from the expression vector g_i to the cluster center O_k.

We know that the volume inside a shell with radius r around O_k in H_L (with elementary thickness dr) equals (Bishop, 1995)

$$S_d r^{d-1}\mathrm{d}r, \tag{5.8}$$

where S_d is the surface area of a unit sphere in d dimensions given by

$$S_d = \frac{2\pi^{d/2}}{\Gamma(d/2)} \tag{5.9}$$

and Γ is the gamma function given by

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}\mathrm{d}u. \tag{5.10}$$

Replacing dV in Equation 5.7 by Equation 5.8 gives us the probability of finding an expression vector of the cluster inside the elementary shell:

$$\frac{S_d}{\left(2\pi\sigma^2\right)^{d/2}} r^{d-1}\exp\left(-\frac{r^2}{2\sigma^2}\right)\mathrm{d}r = p(r \mid C)\mathrm{d}r. \tag{5.11}$$

Said otherwise, Equation 5.11 results in the probability density estimation *p(r|C)* describing the distribution of *r* in the current cluster.

Distribution of r in the background: *p(r|B)*

As previously mentioned, *H* can be described as a *d*-dimensional curved space (hypersphere with radius √(*e*-1)=√(*d*+1)). It has a finite volume given by (Bishop, 1995):

$$S_{d+1}(d+1)^{d/2}, \tag{5.12}$$

where S_{d+1} is the surface area of a unit sphere in *d*+1 dimensions. The background is assumed to be uniformly distributed in this finite volume. Dividing Equation 5.8 by Equation 5.12 gives us the probability of finding an expression vector of the background inside the elementary shell:

$$\frac{S_d}{S_{d+1}(d+1)^{d/2}} r^{d-1}\mathrm{d}r = p(r \mid B)\mathrm{d}r. \tag{5.13}$$

Said otherwise, Equation 5.13 results in the probability density estimation *p(r|B)* describing the distribution of *r* in the background.

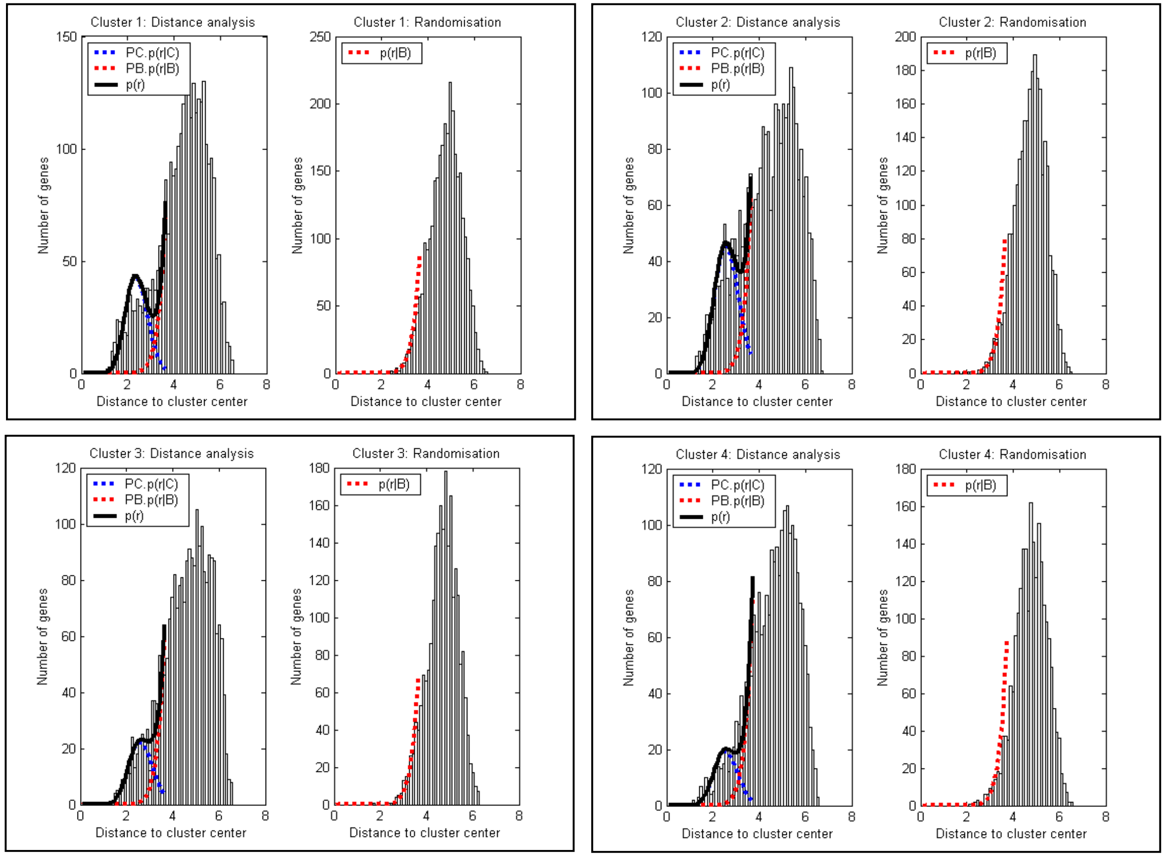
Estimation of the parameters

Two parameters (σ and P_C (or P_B)) still have to be determined by fitting the model to the distance distribution of the original data (the randomised data is *not used* for the actual calculations). This is done by an EM-algorithm (Bishop, 1995). We use the preliminary estimate of the radius *R_k_PRELIM* (see localisation of a cluster center) to initialise the two parameters to be determined by the EM-algorithm. Because the model only has to fit the distribution of *r* (distance to the cluster center - one dimension), the computational complexity of the EM-algorithm is low as compared to the computational complexity of the cluster center localisation in Step 1 and therefore can be neglected if *e* is sufficiently large. The accuracy of the fit (which represents the validity of the assumptions we made to construct our model) for the clusters found in the yeast cell cycle data (see Figure 5.3 and Section 5.4.1) can be inspected in Figure 5.2 for the first four clusters of Figure 5.3.

Calculation of R_k

After the estimation of σ and P_C, we determine the radius R_k of the current cluster so that points that are assigned to the cluster have a

Figure 5.2: Distribution of r (distance from expression vectors to a certain cluster center) for the first 4 clusters found in the yeast cell cycle data (see Figure 5.3). In each box, the histogram on the left represents the distribution of r in the actual data and the histogram on the right represents the distribution of r in the randomised data (the cluster center for the randomised distribution is identical to the cluster center for the actual distribution - the randomisation is not applied to the cluster center itself). For each cluster, the model (see Equation 5.5 - 5.13) fitted by the EM-algorithm is superimposed on the distribution of the actual data (after multiplication with an appropriate factor to fit the scale (this accounts for the bin size and the number of expression profiles in the data set)). The model for the background (see Equation 5.13) is also superimposed on the distribution of the randomised data.



probability *S* or more (significance level - user-defined; default setting: *S* = 95%) to belong to the cluster:

$$P(C \mid R_k) = \frac{P_C.p(R_k \mid C)}{P_C.p(R_k \mid C) + P_B.p(R_k \mid B)} = S. \tag{5.14}$$

To summarise, the complete input-output relation of the method explained in this section is given by: R_k = recalculate_radius (*G*, O_k, R_k*_PRELIM*, *S*). R_k will be empty if O_k is empty (cluster center localisation did not converge) or if the EM-algorithm to determine the model parameters did not converge.

5.3.5 Stop criterion

The iteration (WHILE loop) in the global algorithm ends when the stop criterion is satisfied. This is the case when one of the three following conditions holds true:

1. Step 1 or 2 stops converging.
2. If, for a specific cluster, the number of iterations necessary to decrease the relative difference between R_k and R_k*_PRELIM* (under *ACCUR_RAD*), is larger than a predefined number.
3. If the clusters removed from the data are not valid (number of genes below *MIN_NR_GENES*) for a predefined and consecutive number of times.

5.3.6 Computational and memory complexity of the global algorithm

It is difficult to give an exact measure for the computational complexity of this heuristic approach. However, we can give an indication of the role of the most important variables. As previously said, the computational complexity of one cluster center localisation is approximately O($n \times e$) (*n* is the number of gene expression profiles in the data set, *e* is the dimensionality of the expression vectors) and the computational complexity of the re-estimation of the cluster quality is negligible. So, the computational complexity of one iteration in the global algorithm (WHILE loop) is also approximately O($n \times e$). Notice also that Condition 2 of the stop criterion sets a limit for the maximum number of iterations in the global algorithm needed to define one cluster (which is only valid if the number of genes in this cluster equals or exceeds *MIN_NR_GENES*). Moreover, the number of invalid clusters (number of genes less than *MIN_NR_GENES*) found before one of the conditions of the stop criterion is true, is in practice also more or less proportional to the number of valid clusters found (e.g., for each invalid cluster found, two valid clusters will be found). This number of valid clusters is no classical attribute of the data (like *n* or *e*) used to express computational complexity but it is rather a measure for the complexity of the structure of the data. Taken together, this means that the number of iterations in the global algorithm is also more or less proportional to this number of valid clusters in the data set and since the computational complexity of one iteration is approximately $O(n \times e)$, the computational complexity of the global algorithm is thus approximately $O(n \times e \times VC)$ (VC = number of valid clusters). Notice also that, after finding a certain number of clusters, the number of genes left in the data is smaller than *n* (clusters are discarded from the data). The computational complexity, as described above, is thus an upper limit.

Since only the distances from the expression profiles to the current cluster center have to be kept in memory (this is true at any stage of the algorithm), the memory complexity of the global algorithm is $O(n)$.

5.4 Results

5.4.1 Mitotic cell cycle of Saccharomyces cerevisiae

The algorithm was tested on the yeast cell cycle data as it is described in Appendix B. As previously said, this data set can be considered as a benchmark and contains expression profiles for 6220 genes over 17 time points taken at 10-min intervals, covering nearly two full cell cycles.

Our preprocessing included the following steps: data corresponding to the 90 and 100-min measurements were removed (Tavazoie et al., 1999). Also, we selected the 3000 most variable genes using σ / μ as a metric of variation (see Tavazoie et al.) (filtering). Finally, we standardized the gene expression profiles as described in the standardization section. The final data still contained 2779 missing values. The results of the cluster analysis with our algorithm (MIN_NR_GENES = 10, S = 0.95) are shown in Figure 5.3. Table 5.3 summarises the biological validation of this result by looking for enrichment of functional categories in individual clusters as described in Chapter 4 (Section 4.3.3). We mapped the genes in each cluster to the functional categories in the Munich Information center for Protein Sequences (MIPS) Comprehensive Yeast Genome Database. For each cluster we calculated p-values for observing the frequencies of genes in particular functional categories using the cumulative hypergeometric probability distribution. In the same table we also show, as a comparison and

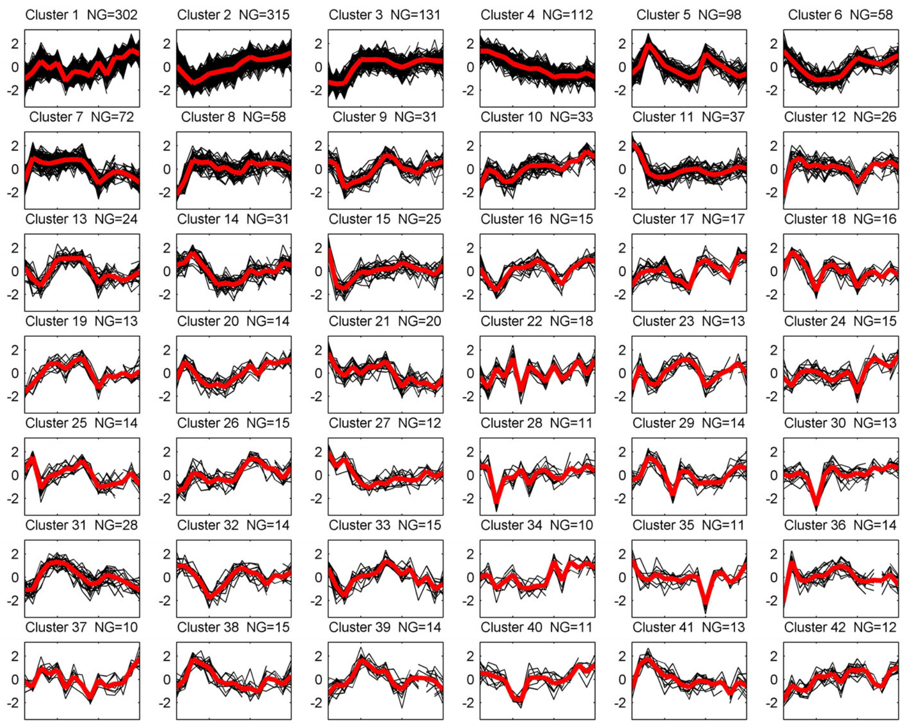


Figure 5.3: Mitotic cell cycle of Saccharomyces cerevisiae: Cluster analysis with AQBC (*MIN_NR_GENES* = 10, *S* = 0.95). Preprocessing included: removal of 90 and 100-min measurements, selection of the 3000 most variable genes using σ / μ as a metric of variation and standardization. Each box corresponds to one cluster and shows the standardized expression profiles of the genes assigned to it. For each cluster, the mean expression profile is also shown. NG = Number of Genes assigned to each cluster.

Table 5.3: Biological validation of the cluster result in Figure 5.3 and comparison with the result of Tavazoie et al. The genes in each cluster have been mapped to the functional categories in the MIPS database and (-log10 transformed) p-values (representing the degree of enrichment - also see main text) have been calculated for each functional category in each cluster. Only significantly enriched functional categories are shown (-log10 transformed p-values ≥ 4) and clusters without enrichment are not listed. As a comparison and in parallel (functionally matching clusters are shown in the same row), the results obtained by Tavazoie et al. (K-means) are also included. NR = Not Reported.

Cluster number AQBC	Cluster number K-means (Tavazoie et al.)	Number of ORFs AQBC	Number of ORFs K-means (Tavazoie et al.)	MIPS functional category	ORFs within functional category AQBC	ORFs within functional category K-means (Tavazoie et al.)	P-value ($-\log_{10}$) AQBC	P-value ($-\log_{10}$) K-means (Tavazoie et al.)
1	1	302	164	ribosomal proteins	101	64	80	54
				organisation of cytoplasm	146	79	77	39
				protein synthesis	119	NR	74	NR
				cellular organisation	211	NR	34	NR
				translation	17	NR	9	NR
				organisation of chromosome structure	4	7	1	4
2	4	315	170	mitochondrial organization	62	32	18	10
				energy	35	NR	8	NR
				proteolysis	25	NR	7	NR
				respiration	16	10	6	5
				ribosomal proteins	24	NR	4	NR
				protein synthesis	33	NR	4	NR
				protein destination	49	NR	4	NR
5	2	98	186	DNA synthesis and replication	20	23	18	16
				cell growth, cell division and DNA synthesis	48	NR	17	NR
				recombination and DNA repair	12	11	8	5
				nuclear organization	32	40	8	4
				cell-cycle control and mitosis	20	30	7	8

Table 5.3 - Continued

Cluster number AQBC	Cluster number K-means (Tavazoie et al.)	Number of ORFs AQBC	Number of ORFs K-means (Tavazoie et al.)	MIPS functional category	ORFs within functional category AQBC	ORFs within functional category K-means (Tavazoie et al.)	P-value ($-\log_{10}$) AQBC	P-value ($-\log_{10}$) K-means (Tavazoie et al.)
6		58		mitochondrial organisation peroxisomal organization energy	15 4 9		7 4 4	
8		58		tRNA-synthetases organisation of cytoplasm	5 14		5 4	
16		15		cellular transport and transport mechanisms	6		4	
21	17	20	83	transcription	9	21	4	4
31	14	28	74	organisation of centrosome nuclear biogenesis organisation of cytoskeleton	3 1 2	6 3 7	4 2 2	6 5 4
36		14		tRNA transcription	3		4	
37	18	10	101	organisation of cytoplasm ribosomal proteins protein synthesis cellular organisation	7 4 4 7	30 16 20 55	6 4 3 2	9 7 7 5
40		11		organization of endoplasmatic reticulum	4		4	
42		12		cellular transport and transport mechanisms	6		4	

Table 5.3 - Continued

Cluster number AQBC	Cluster number K-means (Tavazoie et al.)	Number of ORFs AQBC	Number of ORFs K-means (Tavazoie et al.)	MIPS functional category	ORFs within functional category AQBC	ORFs within functional category K-means (Tavazoie et al.)	P-value ($-\log_{10}$) AQBC	P-value ($-\log_{10}$) K-means (Tavazoie et al.)
	5		152	cell rescue, defense, cell death		22		5
				carbohydrate metabolism		24		4
				stress response		12		4
				energy		16		4
				metabolism of energy reserves		6		4
	7		101	cell-cycle control and mitosis		17		5
				budding, cell polarity, filament formation		10		4
				DNA synthesis and replication		7		4
	8		148	TCA pathway		5		4
				Carbohydrate metabolism		22		4
	21		70	protein synthesis		14		5
				organisation of cytoplasm		18		5
				ribosomal proteins		10		4
	30		60	nitrogen and sulphur metabolism		9		8
				amino acid metabolism		12		7

in parallel (where possible, we compare p-values of functionally matching clusters), the results obtained by Tavazoie et al. on the same data using the K-means algorithm. The three most important clusters found by Tavazoie et al. (cluster 1, 4 and 2 in Tavazoie et al.) could be matched with three clusters discovered by AQBC (cluster 1, 2 and 5). The degree of enrichment in the clusters identified by AQBC, however, was considerably higher and biologically more consistent.

In the biological validation and comparison discussed above, we filtered the data using the same metric of variance (σ / μ) as proposed by Tavazoie et al. because different filtering strategies could produce different clusters independent of the clustering technique (we did not want different filtering to interfere with our comparison). However, in general, if filtering is performed, we recommend using simple measures of variation, like the standard deviation σ (not σ / μ) or the difference between the minimum and maximum value, together with AQBC. Using AQBC with the yeast cell cycle data indeed resulted in biologically more relevant results when using the standard deviation (σ) as the metric of variance to select the 3000 most variable genes (resulting in data with 2563 missing values). This analysis, with the same parameter settings as previously, produced several clusters enriched in top-level functional categories (see Table 5.4).

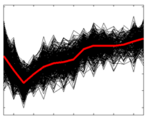
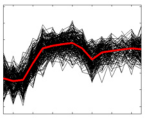
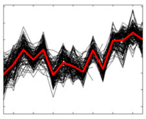
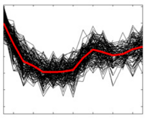
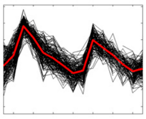
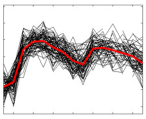
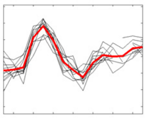
We were able to determine the role of every cluster presented in Table 5.4 within the yeast cell cycle context and correlate this role with the behaviour of the average profiles of the clusters. We have also found several protein complexes where nearly all members belong to the same cluster. Since this is beyond the scope of this text, we will not further discuss this but more information on this can be found on the supplementary website of De Smet et al. (2002) (http://www.esat.kuleuven.ac.be/~fdesmet/paper/adaptpaper.html).

The results of AQBC in this section have been obtained without additional fine-tuning (we used the default value for S) of one or more parameters (unlike, for example, K-means (used by Tavazoie et al.) where the number of clusters has to be estimated in advance, which is certainly not trivial) and that these results can be obtained very easily and almost instantaneously (maximum 1.5 minutes for the examples above on a typical PC).

5.4.2 Central nervous system development

Wen et al. (1998) generated gene expression levels of 112 genes on 9 time points during central nervous system development of the rat (also see Appendix B). In the original reference, clustering of gene expression profiles was performed by using a form of hierarchical clustering. Each gene was

Table 5.4: Biological validation of the results of AQBC on the yeast cell cycle data (*MIN_NR_GENES* = 10, *S* = 0.95) using σ as the metric of variance for filtering. The algorithm retrieved 38 clusters. We looked for enrichment of top-level functional categories in individual clusters. Notice the periodic behaviour of the clusters enriched with cell-cycle specific genes (cluster 3, 6 and 9).

Cluster number	Graphical representation of cluster	Number of ORFs	MIPS functional category (top-level)	ORFs within functional category	P-value (-log$_{10}$)
1		426	energy transport facilitation	47 40	10 5
3		196	cell growth, cell division and DNA synthesis	48	5
4		149	protein synthesis cellular organisation	71 107	50 19
5		159	cell rescue, defense, cell death and ageing	20	4
6		171	cell growth, cell division and DNA synthesis	76	24
9		78	cell growth, cell division and DNA synthesis	23	4
37		11	metabolism	9	6

represented by a 17 dimensional vector (consisting of the 9 expression values and 8 slopes based on a reduced time interval of 1). The hierarchical clustering was based on the 112x112 Euclidean distance matrix calculated using these 17 dimensional vectors. The hierarchical clustering resulted in four basic clusters (or 'major waves') identifying distinct phases of development and a group with largely invariant gene expression profiles (which we could call the constitutively expressed genes).

We applied AQBC (*MIN_NR_GENES* = 10 and *S* = 0.95) to this data set after standardization. We only used the 9 dimensional vectors consisting of the 9 expression values. No missing values were present and no filtering was performed. The algorithm discovered 4 distinct gene groups, each highly correlated with one of the four major waves found by Wen et al. (cluster 1 corresponds to wave 2; cluster 2 corresponds to wave 3; cluster 3 corresponds to wave 1; cluster 4 corresponds to wave 4). The invariant wave was not found, as could be expected after standardization (the division by a small standard deviation inflates the noise, resulting in quasi-random profiles not assigned to any of the clusters).

Figure 5.4 shows the standardized expression profiles (use this figure to compare the average profile in each cluster with the average expression patterns in the major waves 1-4 found by Wen et al. - the similarity is striking).

5.4.3 Measurement of expression levels in different tissues

Seven two-channel cDNA microarray-experiment to characterise 4595 expression patterns in 7 mouse tissues (brain, heart, kidney, liver, lung, skeletal muscle and testis) were performed in the lab of Dr. P. Van Hummelen of the Microarray Facility of the V.I.B. The intention of this experiment was to identify groups of tissue-specific genes. See Appendix B for more details on the data.

We used AQBC to cluster these gene expression vectors after standardization. We used the following parameter settings: *S* = 0.8 and *MIN_NR_GENES* = 5. We did not use the default value 0.95 for *S* (In this data set there are almost no clusters containing expression vectors that reached the default significance level. This could be caused by the rather low dimensionality (*e* = 7) of the data (too few experiments), giving a larger overlap between a cluster and the background). Except standardization, no other preprocessing steps were performed.

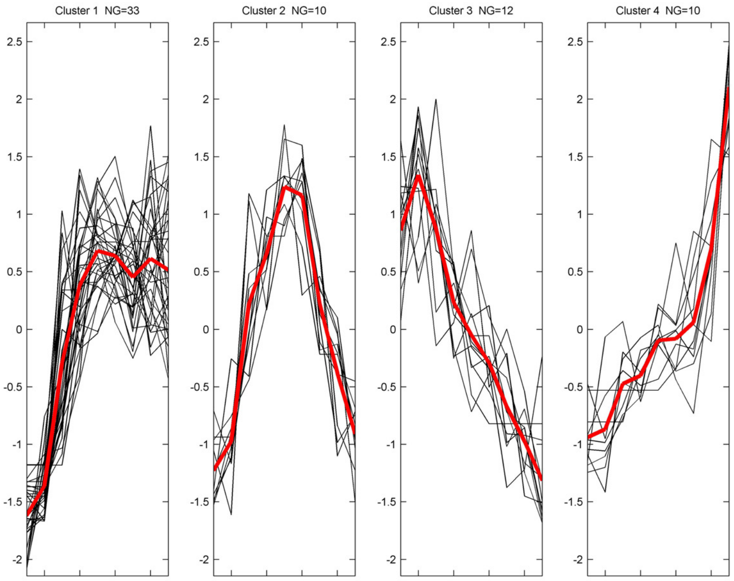


Figure 5.4: Central nervous system development data from Wen et al.: Cluster analysis with AQBC (*MIN_NR_GENES* = 10, *S* = 0.95). Except standardization, no further preprocessing was performed after downloading the raw data. Each box corresponds to one cluster and shows the standardized expression profiles of the genes assigned to it and the mean expression profile. Note the similarity of these mean expression profiles with the four major waves found by Wen et al. NG = Number of Genes assigned to each cluster.

AQBC identified 33 clusters, which can be inspected in Figure 5.5. A considerable number of clusters are tissue-specific (i.e., they contain genes differentially expressed in one or two tissues), reflecting the aims of the experimental setup. An overview of the most important tissue-specific clusters is given in Table 5.5.

5.4.4 Artificial data

We constructed data containing artificially created gene expression profiles of dimension 51. The largest part (1500 profiles) of the data contained totally random profiles (before standardization, these expression profiles were normally distributed around the origin - after standardization, these expression profiles were uniformly distributed on the hypersphere defined by standardization - see Section 5.2.1). In this set of random profiles we introduced 7 small clusters, containing profiles exhibiting significant

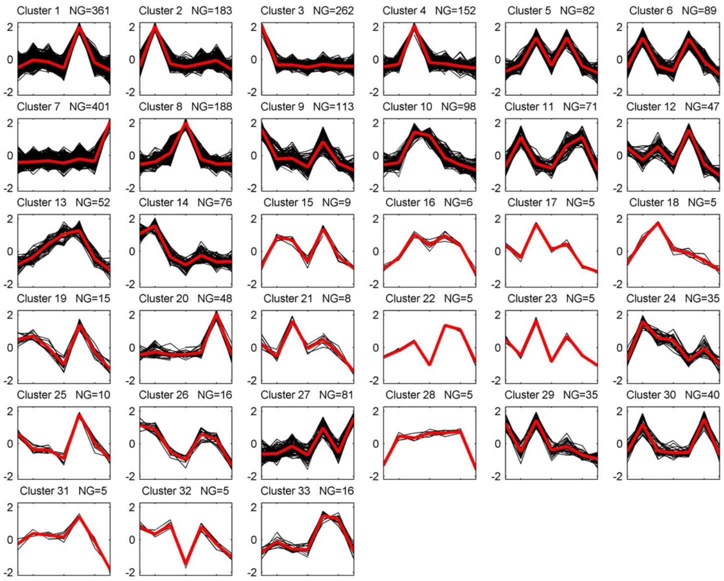


Figure 5.5: Measurement of expression levels in different mouse tissues: Cluster analysis with AQBC (*MIN_NR_GENES* = 5, *S* = 0.8). No filtering was performed. Each box corresponds to one cluster and shows the standardized expression profiles of the genes assigned to it and the mean expression profile. Note the presence of several tissue-specific clusters - see Table 5.5. NG = Number of Genes assigned to each cluster.

Table 5.5: Overview of the most important tissue specific clusters from Figure 5.5.

Cluster number	Tissue specificity
1	Lung
2	Heart (Skeletal muscle)
3	Brain
4	Kidney
5	Kidney Lung
6	Heart Lung
7	Testis
8	Liver
10	Kidney Liver
20	Skeletal muscle (Heart)
30	Heart Skeletal muscle

coexpression. These clusters were created by superposing normally distributed noise (the variance of this noise was different for each cluster) on copies of 7 template profiles (the number of copies was also different for each cluster) - see Figure 5.6. We used 5 cosine-like template profiles (3 with phase shifts and 2 with frequency shifts) and 2 template profiles that were random. A data set created by this procedure was used for cluster analysis with our algorithm. After standardization, we used the default settings, except for *MIN_NR_GENES* (which we set equal to 15, which is the number of profiles in the smallest cluster, to avoid finding small clusters accidentally present in the 1500 random profiles). The algorithm was able to identify the clusters introduced in the data set and separate the 1500 random profiles from the profiles in the clusters (these random profiles were not assigned to any of the clusters). This result is shown in Figure 5.7.

We also created a second artificial data set by introducing 8377 missing values (about 10% of the entries) in the first set. The introduction of these missing values did not change the results obtained by our algorithm (using the same settings for *S* and *MIN_NR_GENES* as before).

The results above demonstrate the ability of the algorithm to separate small subsets of significantly coexpressed gene expression profiles from a large collection of unrelated profiles and the ability to discriminate between individual subsets or clusters (even between the 3 clusters on the left side of Figure 5.6 - clusters with cosine-like template profiles with phase shifts).

5.5 INCLUSive[2]

Our algorithm AQBC is publicly available for data analysis and can be found on http://www.esat.kuleuven.ac.be/~thijs/Work/Clustering.html. This method has also been integrated in an on-line tool, called INCLUSive (see http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html), which is a suite of web-based tools and is aimed at the automatic multistep analysis of microarray data. The goal is to provide an integrated platform where several sources of information can be linked together to facilitate the analysis of microarray data. Currently, preprocessing of microarray data (Engelen et al., 2003), AQBC, information retrieval of genes in clusters (cluster validation), retrieval of upstream sequences and motif finding algorithms (Thijs et al., 2001; Thijs et al., 2002b) are accessible from this website.

[2] INCLUSive has appeared as an Application Note in Bioinformatics (Thijs et al., 2002a) and an updated version has been published in the web software issue of Nucleic Acids Research (Coessens et al., 2003).

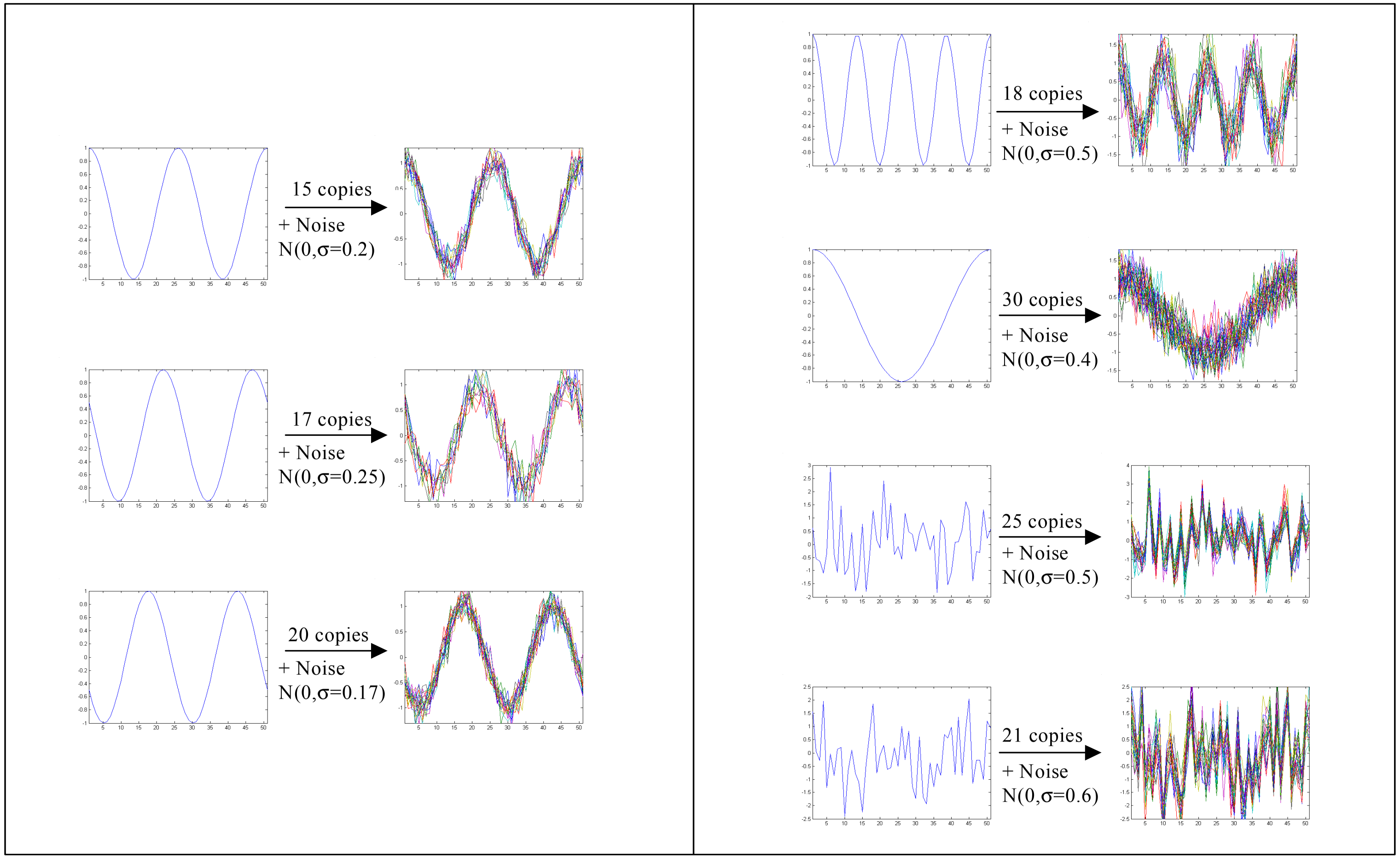


Figure 5.6: Construction of the 7 clusters in the artificial data set that were introduced into the 1500 random profiles.

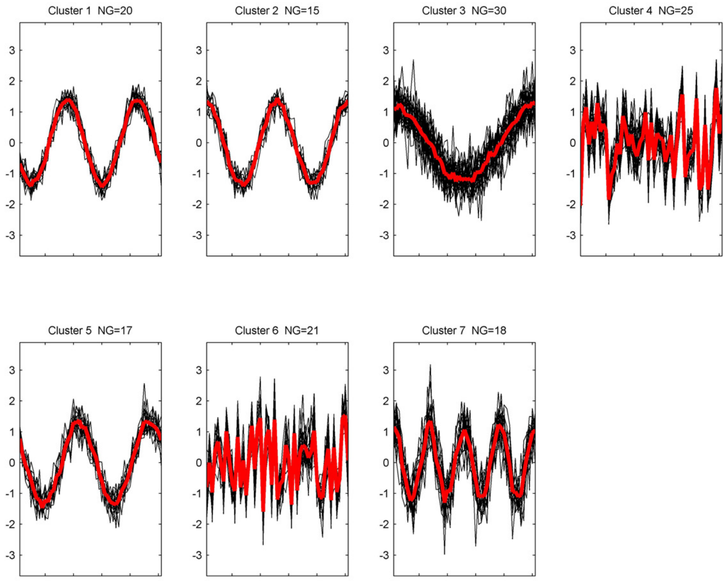


Figure 5.7: Cluster analysis with AQBC (*MIN_NR_GENES* = 15, *S* = 0.95) of the standardized artificial data set. Each box corresponds to one cluster and shows the standardized expression profiles of the genes assigned to it and the mean expression profile. The clusters introduced in Figure 5.6 have all been completely recovered and the random profiles have been excluded from the result. NG = Number of Genes assigned to each cluster.

5.6 Discussion and conclusion

The algorithm proposed in this chapter is designed to find clusters of significantly coexpressed genes (higher degree of coexpression than could be expected by chance) in high-density areas of the data (high-density areas were assumed by Heyer et al. to be, biologically seen, the most interesting regions in the data). Genes not exhibiting an expression profile significantly similar to the expression profile of other genes in the data are not assigned to any of the clusters. The same applies to genes lying in low-density areas of the data. The size or radius for each cluster separately is determined - through the significance level *S* - by limiting the probability of a false positive result (a gene assigned to the cluster that is not really coexpressed with the other members of the cluster). The default value for the significance level *S* guarantees that a gene, which has been assigned to the cluster, has a probability of 95% or more to belong to the cluster (this means that the probability of being a false positive is 5% or less). In other words, the genes in the cluster are significantly coexpressed with a certain confidence.

Therefore, clusters formed by our algorithm might be good ‘seeds’ for further analysis of expression data (see INCLUSive and Thijs et al. (2002a)) since they only contain a limited number of false positives. When the presence of false positives in a cluster is undesirable, a more stringent value for the significance level *S* might be applied (e.g., 99%; for noise-sensitive analyses such as motif finding) which will result in smaller clusters exhibiting a more tightly related expression. The presence of a lower number of false positives was confirmed by the comparison in Table 5.3 of the cluster result of AQBC and K-means applied to the data from Cho et al. This comparison showed that the degree of enrichment in the clusters retrieved by AQBC was substantially higher. From a biological point of view, the control of the number of false positives is the main advantage of our algorithm.

The significance level *S*, in turn, can be seen as a constant quality criterion for the clusters (while the quality criterion *R* as defined in Equation 5.3 differs among the clusters defined by our algorithm). Our algorithm can thus be regarded as being a pure quality-based clustering method where all the clusters have a constant quality represented by the significance level *S* (the term adaptive quality-based clustering is thus only valid when using Equation 5.3 as quality criterion). When compared to the previous definition (quality measure *R*), this new quality measure *S* has the advantage that it has a strict statistical meaning (it is therefore much less arbitrary) and that, in most cases, it can be chosen independently of a specific data set or cluster. In addition, it allows for the setting of a meaningful default value (95%).

In Table 5.6 a detailed comparison between our global algorithm (AQBC) and the algorithm proposed by Heyer et al. (QT_Clust) is made. Because we focus on algorithmic aspects, the QT_Clust algorithm in our comparison uses the same distance and quality measure as we did (Euclidean distance and quality defined as in Equation 5.3 - In Heyer et al. the jackknife correlation was used together with a quality measure defined as a diameter). This change of distance and quality measure does not significantly change the structure of QT_Clust and in essence, there is no fundamental difference between a quality defined as a radius and a quality defined as a diameter.

To complete the picture, Table 5.7 gives a summary of the differences between our method, hierarchical clustering, SOM and K-means.

In summary, some of the properties of the AQBC approach make it very suited for cluster analysis of gene expression profiles:

i. AQBC can be considered as an intuitively appealing and user-friendly algorithm where the principal user-defined parameter is a significance level *S*, which has a strict

Table 5.6: Comparison between AQBC and QT_Clust (Heyer et al., 1999).

	AQBC	**QT_Clust**
User-defined parameters	1. Data set *G* 2. Significance level *S* 3. Minimum number of genes *MIN_NR_GENES*	1. Data set 2. Quality (radius *R* or diameter *d*) 3. Minimum number of genes (termination criterion)
Computational Complexity	$\sim \mathrm{O}(n \times e \times VC)$	$\sim \mathrm{O}(n^2 \times e \times VC)$
Cluster radius R	Automatically calculated for each cluster separately - not constant	Constant and user-defined
Quality measure	Significance level *S*: statistical parameter that can be chosen independently of data set (default value (0.95) almost always gives meaningful results)	Arbitrary parameter that has to be set by the user in function of a specific data set, after visual inspection of clusters formed at different quality-levels (optimal value is not straightforward) - no meaningful default value
Number of clusters	Not predefined	Not predefined
Inclusion of all genes in clusters	No	No
Result	Always gives the same results for the same parameter settings - results are reproducible	Always gives the same results for the same parameter settings - results are reproducible

Table 5.7: Comparison between AQBC, hierarchical clustering, Self-Organizing Maps and K-means.

	AQBC	**Hierarchical clustering** Eisen et al. (1998)	**SOM** Tamayo et al. (1999)	**Standard K-means** Tou and Gonzalez (1979)
Format of result	Set of clusters	Tree structure difficult to interpret for large data sets	Set of predefined number of clusters	Set of predefined number of clusters
Principal user-defined parameter	Significance level *S*	-	Number of clusters / Node topology	Number of clusters
Additional requirements from the user	Limited (fine-tuning of *S* is rarely necessary)	Definition of (an arbitrary) level where the tree structure has to be cut	Extensive parameter fine-tuning (comparison of several runs with different parameter settings) is almost always necessary	Extensive parameter fine-tuning (comparison of several runs with different parameter settings) is almost always necessary
Statistical definition of clusters	Yes	No	No	No
Inclusion of all genes in clusters	No	Yes	Yes	Yes
Missing values management	Yes	Yes	Not discussed	Not standard
Computational complexity of one run of the algorithm	Linear in *n*	Quadratic in *n*	Linear in *n*	Linear in *n*

statistical meaning and is therefore much less arbitrary than for example the predefinition of the number of clusters or the quality guarantee used in standard quality-based clustering. It can be chosen independently of a specific data set or cluster and it allows for a meaningful default value that often gives meaningful results. There is no need for extensive parameter fine-tuning.

ii. AQBC produces clusters adapted to the local data structure (the clusters do not have the same radius).

iii. Only genes that are significantly coexpressed are assigned to a cluster.

iv. AQBC is a fast algorithm with a computational complexity that is linear in the number of expression profiles.

v. A server running the program is publicly available for data analysis.

vi. Our implementation has an integrated approach for missing values without the necessity to replace them.

vii. AQBC was extensively biologically validated.

AQBC, however, also has some limitations:

i. It is a heuristic approach not proven to converge in every situation.

ii. Due to the model structure used in Step two (Section 5.3.4) some additional constraints have to be imposed. They include:

    a. Only standardized expression profiles are allowed.

    b. AQBC has to be used in combination with the Euclidean distance and cannot directly be extended to other distance measures.

Chapter 6

Univariate analysis in microarray data

6.1 Introduction

As already announced, in this chapter we will focus on the problem of univariate analysis and multiple testing in microarray data[1].

Microarrays allow for the simultaneous measurement of expression levels of thousands of genes in a certain tissue (e.g., in a tumour). These measurements can be repeated under different conditions (e.g., originating from tumours or tissues with different properties such as normal and malignant tissues (Alon et al., 1999); tumours that are and are not sensitive to chemotherapy (Kihara et al., 2001); tumours with good and poor prognosis (van 't Veer et al, 2002); tumours with and without metastatic potential (Ramaswamy et al., 2003); and so on). Also see Figure 3.1.

Usually a test statistic or a hypothesis test (resulting in a p-value for *each* gene - univariate analysis) is used to rank the genes with respect to their differential expression between the different tumour types or experimental conditions. See Appendix A, Section A.1 for more details about hypothesis testing. Subsequently, an arbitrary threshold or rejection level α (genes with a p-value smaller than α are *declared* to be positive or differentially expressed) is chosen to select the genes that warrant further investigation or validation (e.g., for target discovery in drug development (Gerhold et al., 2002)).

However, due to the overlap of the p-values of the genes that are and are not *actually* differentially expressed (i.e., the genes whose expression is and is not affected by the difference between the experimental conditions), the choice of this rejection level has some consequences (also see Table 6.1).

[1] A summary of the discussion in this chapter has been submitted to the British Journal of Cancer as a full paper (De Smet et al., 2004).

Firstly, genes whose expression is not affected by the difference between the different tumour types and therefore have no actual differential expression, can accidentally have a p-value that is lower than the rejection level and are therefore wrongly declared to be differentially expressed. In statistics this is also called a Type I error. This results in a number of false positive genes that will not yield any results in further investigations. Since the number of genes in a microarray, that is not actually differentially expressed, usually is high in microarray data, the number of false positive genes at commonly used rejection levels (e.g., 5%), can be considerable (problem of multiple testing). See the Results section for some examples.

Secondly, the choice of the rejection level can also result in a certain number of false negative genes (Type II error). These are the genes that are actually differentially expressed but that have a p-value that is larger than the rejection level, resulting in discarding potentially valid targets.

Recently, much attention has been paid in literature to the control of the number of false positives or Type I error (Keselman et al., 2002; Reiner et al., 2003; Storey and Tibshirani, 2003). Classically and as already discussed in Chapter 2 for the analysis of clinical data, by applying a Bonferroni correction (also see Appendix A, Section A.1), one can control the family-wise error (FWE - probability of having one or more false positives) at a given level, fixed beforehand. However, for microarray data, where usually a considerable number of genes is actually differentially expressed, controlling the FWE is too stringent and results in an unacceptable Type II error (leading to an unacceptable loss of statistical power). Controlling the False Discovery Rate (FDR; expected fraction of genes falsely declared positive among all the genes declared differentially expressed) (Benjamini and Hochberg, 1995; Reiner et al., 2003; Storey and Tibshirani, 2003) is less stringent and seems a more sensible approach for microarray data but still does not control the Type II error, which could still be large and lead to the loss of a considerable number of missed targets. Control of the Type I error in microarray data often goes at the expense of the Type II error that remains uncontrolled and (too) large. While the study of multiple testing finds its roots in genetic studies where the number of positives is usually small and control of false positives is paramount, the number of positives in studies of differential expression between patient biopsies is large and false negatives become an equally important issue. Because of this historical reason, we believe that the control of false negatives in multiple testing methods has been somewhat overlooked.

In this chapter we will first describe a method to estimate the total number of genes that is actually differentially expressed starting from the p-values assigned by a certain hypothesis test to every gene and independent of a certain rejection level defined in advance. Using this result, we present a method based on Receiver Operating Characteristic (ROC) curves that does not control the Type I or Type II errors but that obtains an *optimal balance* between them. We aim to obtain a sensible or optimal - according to a certain criterion - trade-off between false positives and negatives.

Moreover, the use of ROC curves enables us to estimate the degree of overlap between the p-values of genes that are and are not actually differentially expressed. This amount of overlap in its turn determines the relationship between the false positives and negatives and the level of the (optimal) trade-off or balance between them (i.e., the lower the amount of overlap, the better the balance). The assessment of the amount of overlap between the p-values by ROC curves can therefore be used to assign a quality measure to a specific microarray data set. This quality measure can help to compare different data sets that study the same experimental conditions with respect to their ability to discriminate between genes whose expression is and is not affected by the different conditions. This can help the biologist to decide which data set is best suited for further analysis, without first having to choose an arbitrary rejection level.

Below, we will first describe the methodology in detail and apply this, among others, using two pairs of data sets that are publicly available (one pair dealing with acute leukemia and one pair dealing with degree of differentiation in breast cancer).

6.2 Methodology

Consider microarray data containing several sets of experiments, each analysing tissues originating from a specific group of malignancies or a specific condition, and containing expression levels for *n* genes g_i (we call this set of genes $\check{N}$ - so $n = \# \check{N}$). Assume that we have already used a certain hypothesis test to calculate the p-values p_i of the respective genes (also see Appendix A, Section A.1 for more details about hypothesis testing). These p-values reflect the probability that an equally good or better test statistic, quantifying the difference between the gene expression levels of the different conditions, is generated if a certain null hypothesis is true. In general, the null hypothesis states that there is no actual differential expression. Also assume that the genes are ordered according to this p-value, so that $p_1 < p_2 < \ldots < p_n$. Note, that in this chapter, we chose the Wilcoxon rank sum test (a nonparametric test that examines the null hypothesis that the medians of the expression levels from *two* conditions for a certain gene are identical) to generate the p-values (Pagano and Gauvreau, 2000; Troyanskaya et al., 2002). This test uses a test statistic that is based on the ranks of the expression levels of one gene rather than on the values themselves. Note, that in principle, every procedure (e.g., through random column permutations of the data to simulate the distribution of the test statistic under the null hypothesis (Tusher et al., 2001)) or hypothesis test (e.g., Kruskal-Wallis test if there are more than two conditions), that generates p-values for every individual gene, is suitable as long as its underlying assumptions are checked or assumed.

Now assume that the number of genes that is actually differentially expressed is n_1 (for these genes the null hypothesis is false - we call this set of genes $\check{N}_1$, so $n_1 = \# \check{N}_1$). Assume further that the number of genes that is not actually differentially expressed is n_0 (for these genes the null hypothesis is true - we call this set of genes $\check{N}_0$, so $n_0 = \# \check{N}_0$). Of course, these numbers are not known in advance and have to be estimated from the data.

Starting with the estimation of n_1 and n_0, we proceed by calculating the number of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*) at each rejection level. Using these estimates, the sensitivities and specificities at each rejection level can be calculated. Finally, we use these quantities to construct a ROC curve. All methods described in this chapter were implemented in MATLAB but are straightforward to implement using other packages.

6.2.1 Estimation of n_1 and n_0

Assume that a gene g_t with associated p-value p_t can be found with t defined as follows:

$$t = \min\left\{ j \mid g_j \in \breve{N}_0 \text{ and } \forall g_i \in \breve{N}_1 : p_i \leq p_j \right\}. \tag{6.1}$$

The assumption of the existence of such a gene g_t comes down to the fact that one supposes that the largest p-value in the data set belongs to $\check{N}_0$ (which is logical since genes belonging $\check{N}_1$ will, in general, have relatively small p-values because they are not generated under the null hypothesis).

Now choose any gene g_k with $p_k \geq p_t$ (by definition g_k belongs to $\check{N}_0$, since all genes belonging to $\check{N}_1$ have p-values smaller than p_t). Since the genes were ordered according to their p-value, k is the number of genes belonging to $\check{N}$ with a p-value equal to or smaller than p_k. Since $\check{N} = \check{N}_1 \cup \check{N}_0$, we can write the following set of equations:

$$\begin{cases} k = \#\left\{g_i \in \breve{N}_1 \mid p_i \leq p_k\right\} + \#\left\{g_i \in \breve{N}_0 \mid p_i \leq p_k\right\} & (6.2) \\ n = n_1 + n_0. & (6.3) \end{cases}$$

Since, by definition, all genes belonging to $\check{N}_1$ have p-values smaller than p_k, the first term in Equation 6.2 equals n_1. To calculate the second term, we assume that the test statistics of the gene expression profiles of $\check{N}_0$ (that are generated under the null hypothesis) are independent (all genes, that exhibit coexpression that can change the test statistic, are assumed to belong to $\check{N}_1$). Under this condition and by definition, the probability, that a gene from $\check{N}_0$ has an equally good or better test statistic than g_k (i.e., has a p-value equal to or smaller than p_k), equals p_k. This means that the expected value (mean of the binomial distribution) of the second term in Equation 6.2 equals $p_k.n_0$ and that we can approximate Equation 6.2 as follows:

$$k \approx n_1 + p_k.n_0. \tag{6.4}$$

Deriving n_1 from the set of Equations 6.3 and 6.4 gives:

$$n_1 \approx \frac{k - p_k.n}{1 - p_k}. \tag{6.5}$$

For a given data set, n_1 is constant. Now define V_i, for *every* gene g_i, as follows:

$$V_i = \frac{i - p_i.n}{1 - p_i}. \tag{6.6}$$

According to Equation 6.5 and for $p_i \geq p_t$, V_i is constant and equals n_1. Moreover, it is easy to prove that $V_i < n_1$ when $p_i < p_t$ and that V_i goes to zero when p_i gets smaller.

Using this information, we can present an easy method to derive n_1 (and n_0 through Equation 6.3): Calculate V_i for every gene g_i and plot these values in a graph (e.g., i on the X-axis and V_i on the Y-axis). If this graph reaches a constant level at a certain gene, this gives us respectively n_1 and g_t. In practice, after reaching the constant level, the graph will slightly fluctuate around a mean value (because of the approximation we used to derive Equation 6.4). So for the calculation of n_1, it is better to take the mean of V_i in a certain interval $[r,s]$ where $r > t$ and $s << n$, if possible (if $i \approx n$, $p_i \approx 1$ and the denominator in Equation 6.6 gets very small and the formula for V_i becomes ill conditioned). See the Results section for some examples of this method.

Alternative derivation

Storey and Tibshirani (2003) recently reported a method (in PNAS), using a somewhat different reasoning than given above, to estimate n_0 (this was published *after* the development of our method). We will discuss the main ideas of their approach below (using a notation that is consistent with the one that is used here) and will show that their result is completely equivalent with the method described above.

First consider a data set that does not contain any genes with actual differential expression (this can be approximated by randomising an existing microarray data set containing genes with actual differential expression through an independent and random permutation of the elements of every row in the expression matrix). The test statistic of the genes of such data follows the null distribution and the p-values are uniformly distributed between 0 and 1, which can be seen in Figure 6.1 where a histogram of the p-values of a randomised data set is shown.

Now consider a real microarray data set containing genes with and without actual differential expression. A histogram of the p-values of a representative data set can also be inspected in Figure 6.1. The distribution of the p-values in this case is a superposition of a uniform distribution assumed to be generated by the genes that are not actually differentially expressed (like in the randomised data set) and a distribution assumed to be generated by the genes with actual differential expression (whose test statistic does not follow the null distribution). The genes in this last distribution have p-values that are concentrated in the lower range and that are almost absent in the higher range (close to one).

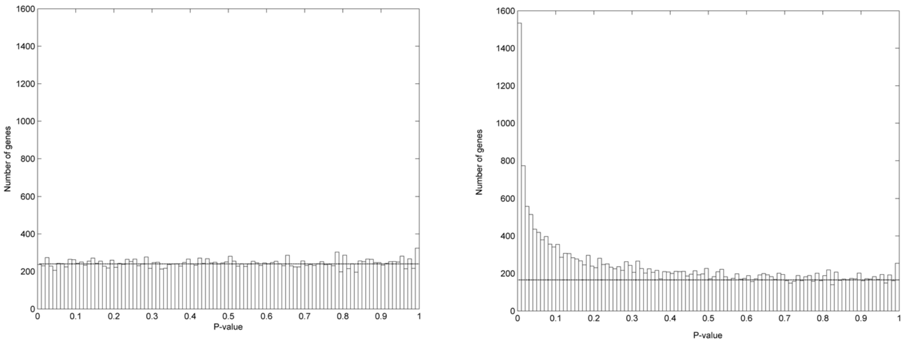


Figure 6.1: Left: Histogram of the p-values from a data set only containing genes without actual differential expression (randomisation of the data set that was used for the histogram on the right). This represents a uniform distribution; Right: Histogram of the p-values of a data set that contains genes with and without actual differential expression. The distribution of the p-values is a superposition of two distributions separated by the horizontal line. The distribution under the horizontal line is a uniform distribution representing the genes without actual differential expression and the distribution above the horizontal line represents the genes with actual differential expression where the p-values are concentrated in the lower range.

The number of genes that generate the uniform distribution under the horizontal line in the histogram on the right in Figure 6.1 is given by n_0.

Storey and Tibshirani propose to estimate this number as follows: consider a gene from $\breve{N}$ with a p-value p_i. Since the distribution of the p-values of the genes without actual differential expression is assumed to be uniform, the number of genes from $\breve{N}_0$ with a p-value larger than p_i can be estimated by:

$$\#\{g_j \in \breve{N}_0 \mid p_j > p_i\} = n_0(1 - p_i). \tag{6.7}$$

If $p_i \to 1$, the following applies:

$$\#\{g_j \in \breve{N}_0 \mid p_j > p_i\} = \#\{g_j \in \breve{N} \mid p_j > p_i\} = n - i, \tag{6.8}$$

since most of the genes from $\breve{N}$ with p-value close to 1 have no actual differential expression and therefore belong to $\breve{N}_0$. From Equation 6.7 and 6.8 follows that

$$n_0 = \lim_{p_i \to 1} \frac{n - i}{1 - p_i}. \tag{6.9}$$

Equation 6.9 is the main formula that has to be evaluated in Storey and Tibshirani. Since $n_1 = n - n_0$, the following applies:

$$n_1 = \lim_{p_i \to 1}\left(n - \frac{n - i}{1 - p_i} \right) = \lim_{p_i \to 1} \frac{i - p_i n}{1 - p_i}. \tag{6.10}$$

Equation 6.8 also applies for p_i in the neighbourhood of 1 since the distribution (as represented in the right histogram in Figure 6.1) becomes reasonably flat for p-values relatively close to 1. This also means that, in Equation 6.10, the limit value will already be reached and the expression behind the limit will become more or less constant for p_i in the neighbourhood of 1. Since this expression is equal to the one given in Equation 6.6, this shows that the method of Storey and Tibshirani is equivalent to the method described at the start of this section. Storey and Tibshirani assume that the p-values of the genes without actual differential expression follow a uniform distribution. For the derivation of Equation 6.6 we have assumed that that the test statistics of the gene expression profiles without actual differential expression are independent and that all genes, that exhibit coexpression that can influence the test statistic, are assumed to belong to $\breve{N}_1$. The assumptions concerning the uniform distribution and the independence of the test statistics of the genes that belong to $\breve{N}_0$ are in fact equivalent and follow from each other.

6.2.2 Estimation of the number of true positive, true negative, false positive and false negative genes

Suppose that we declare the genes with a p-value smaller than or equal to a certain rejection level $\alpha = p_i$ as differentially expressed (i.e., the null hypotheses for these genes are rejected - one could say that the expression of these genes is predicted not to be affected by the difference in conditions) and the genes with a p-value larger than this rejection level as not differentially expressed (i.e., the null hypotheses for these genes are not rejected - one could say that the expression of these genes is predicted not to be affected by the difference in conditions). When the declared status of differential expression is compared with the actual status of differential expression (or with the actual status of the null hypothesis - false or true), four categories of genes (true positive (TP_i), true negative (TN_i), false positive (FP_i) and false negative (FN_i) genes) emerge that are defined in Table 6.1. Using the value of n_1 and n_0, derived in the previous section, we can calculate the number of genes in each category using the formulas from Table 6.1.

Table 6.1: Definition of True and False Positive genes (TP_i and FP_i) and of True and False Negative genes (TN_i and FN_i) at a certain level of rejection $\alpha = p_i$ (p-value of the ith gene after ranking them in ascending order by p-value). For each of them, the formula of the expected value is given.

		Actually differentially expressed?		
		YES	NO	
Declared differentially expressed?	YES ($p \leq p_i$)	TP_i $\approx i - p_i.n_0$	FP_i $\approx p_i.n_0$ **Type I error**	Pos_i $= i$
	NO ($p > p_i$)	FN_i $\approx n_1 - i + p_i.n_0$ **Type II error**	TN_i $\approx (1\text{-}p_i).n_0$	Neg_i $= n\text{-}i$
		n_1	n_0	

6.2.3 Sensitivity and specificity

Using the values calculated in Table 6.1, the sensitivity ($SENS_i$) at a certain rejection level $\alpha = p_i$ is defined as (Pagano and Gauvreau, 2000)

$$SENS_i = \frac{TP_i}{TP_i + FN_i} = \frac{TP_i}{n_1}, \tag{6.11}$$

which is the fraction of actually differentially expressed genes that are declared differentially expressed. Note that 1 - sensitivity equals the probability that a gene with actual differential expression is not declared differentially expressed, which is exactly the probability of a Type II error.

The specificity ($SPEC_i$) at a certain rejection level $\alpha = p_i$ is defined as (Pagano and Gauvreau, 2000)

$$SPEC_i = \frac{TN_i}{TN_i + FP_i} = \frac{TN_i}{n_0}, \tag{6.12}$$

which is the fraction of genes without actual differential expression that are not declared differentially expressed. Note that 1 - specificity equals the probability that a gene without actual differential expression is declared differentially expressed, which is exactly the probability of a Type I error.

6.2.4 Construction and interpretation of ROC curves

Suppose that we calculate the sensitivities and specificities at *all* possible rejection levels $\alpha = p_i$ ($i = 1,\ldots,n$) and that we construct a Receiver Operating Characteristic (ROC) curve (sensitivity plotted versus 1-specificity - also see Appendix A, Section A.2.1). ROC curves are a popular method to compare and characterise the performance of diagnostic tests in medicine (e.g., Epstein et al. (2002)). We will discuss and use them here to quantify our ability to discriminate between genes with and without actual differential expression.

First of all, a ROC curve shows the trade-off or balance between specificity and sensitivity (and hence between the Type I and Type II errors) for every possible rejection level and therefore allows for the selection of a rejection level α^{opt} with an optimal balance between specificity and sensitivity or between the Type I and Type II errors. Optimal can be defined in several ways and depends on the context or the requirements of the application. Often, the point on the ROC curve (and associated rejection level) with a tangent line with slope 1 is chosen, for which it can be proven that it maximizes the sum of the sensitivity and specificity (and hence minimises the sum of the probability of a Type I and Type II error) - this is also the definition of optimal that will be used in this chapter. Alternatively, one can also try to optimise a more custom defined cost function of the Type I and Type II errors that meets some specific requirements. One could, for example, use a cost function that puts more weight on either the Type I or Type II error, dependent on which is most important for a specific situation. In fact, by minimising the sum of the probability of the Type I and Type II errors (as said, this is done in this chapter), the number of false positives and negatives are weighed by the inverse of the number of genes without and with actual differential expression, respectively. This means, for example, that the 'cost' of a false negative will be higher if the number of genes that are actually differentially expressed (or that are actually positive) is lower and vice versa, which is logical since the impact of missing a rare target is higher than the impact of missing one of many targets.

Secondly, the Area under the ROC curve (AUC) has a special meaning (see Appendix A, Section A.2.2 for a method to calculate the AUC and its standard deviation). Suppose we randomly select a gene g_i with actual differential expression with p-value p_i and a gene g_j without actual differential expression with p-value p_j, then it can be proven that

$$AUC = P(p_i < p_j), \tag{6.13}$$

i.e., the AUC equals the probability that the p-value of the gene with actual differential expression is lower than the p-value of the gene without actual differential expression and therefore it is the probability that p_i and p_j are ranked correctly. The AUC quantifies how well the genes whose expression is and is not affected by the difference between the tumour types can be discriminated using the p-values of these genes independently of the choice of an arbitrary rejection level and independently of the relative values for n_1 and n_0. The AUC increases if the overlap between the p-values of the genes with and without actual differential expression decreases. This means that the level of the (optimal) balance between Type I and Type II errors (e.g., reflected by the sum of the specificity and sensitivity) improves if the AUC increases. Therefore, the AUC can be seen as a quality measure with respect to the detection of differential expression for a specific set of microarray experiments, given a certain hypothesis test. Provided the same hypothesis test is consistently applied, the AUC can be used to compare (see Appendix A, Section A.2.3 for a method to compare AUCs) the ability of different gene expression data sets to discriminate between genes whose expression is and is not affected by the difference in conditions. For example, one could calculate this quality measure for several data sets, which study gene expression levels under the same conditions, from different sources or institutions. As another example, one could try to study the effect on the differential expression and on this quality measure by a change in one or both conditions (see Results section).

6.2.5 False discovery rate

The False discovery rate (FDR_i) (Benjamini and Hochberg, 1995; Tusher et al., 2001; Rhodes et al., 2002; Keselman et al., 2002; Reiner et al., 2003; Storey and Tibshirani (2003)) at a certain rejection level $\alpha = p_i$ can be defined as

$$FDR_i = E\left(\frac{FP_i}{TP_i + FP_i} \right) \approx \frac{p_i . n_0}{i}, \tag{6.14}$$

which is the expected value of the fraction of genes falsely declared differentially expressed from all the genes that are declared differentially expressed. The false discovery rate is a measure that is often used to quantify and control the Type I error. Using the formulas from Table 6.1, we estimate this quantity by the expression in the right hand side of Equation 6.14. If one would, for example, try to validate all the genes that are declared differentially expressed, the false discovery rate reflects the fraction of genes where the validation procedure is expected to be unsuccessful. Selecting a rejection level with a low FDR_i limits the Type I error and yields higher efficacy for the target validation. The estimated number of false positive genes (in the nominator for FDR_i) is based on n_0 and not on n (like for example in Tusher et al. (2001) or Rhodes et al. (2002) - the false discovery rate is overestimated there because the number of false positives is based on the number of null hypotheses that would be rejected if the null hypotheses were true for *all* the genes in the data set). This is important if n_1 is large, which is often the case.

Two main factors independently determine the behaviour of the false discovery rate: the AUC and the relative value of n_1 (reflected by the fraction n_1/n). An increased value for the AUC (reflecting less overlap between the distributions of the p-values of $\check{N}_1$ and $\check{N}_0$) causes FDR_i to start approaching its maximum value at higher values of i. An increased value for n_1/n (or a decreased value for n_0/n = 1- n_1/n, which is the maximum value for the false discovery rate in Equation 6.14, since $p_n \approx 1$) causes an overall decrease of the false discovery rate.

6.3 Results

6.3.1 Acute leukemia

In this paragraph we will apply the methodology described above on microarray data from two sources that contain measurements for two or three classes of patients with acute leukemia.

The first set contains the data from Golub et al. (1999) as it is described in Appendix B and already used for data analysis in Chapter 3. In summary, it contains expression profiles of 72 patients with acute lymphoblastic (ALL - Condition 1) or myeloid (AML - Condition 2) leukemia. In the original publication the patients are divided into a training (38 patients; 27 ALL and 11 AML) and a test set (34 patients; 20 ALL and 14 AML). The data contains n = 7129 genes. No additional preprocessing was performed after downloading.

The second data set (Armstrong et al., 2002) also contains several microarray experiments obtained from patients with acute leukemia but contains patients from a third condition (called MLL leukemia) besides ALL and AML. Also see Appendix B for more information. The data contains expression profiles for 12582 genes measured using Affymetrix technology. In total, 24 ALL patients, 28 AML patients and 20 MLL patients are available.

We will first illustrate our procedure for univariate analysis of microarray data using only the patients of the training set from Golub et al. We will also analyse a randomised version of this training set and use this as a basis to construct an artificial data set. Next, we will use our methodology to compare the complete data from Golub et al. (training + test set) with the data set from Armstrong et al. with respect to the detection of differential expression between ALL and AML. Finally, we will investigate the effect of a change in condition (replacement of ALL or AML patients with MLL patients in the data from Armstrong et al.).

Training set from Golub et al.

The results of our analysis using the 38 patients from the training set from Golub et al. can be inspected in Figure 6.2-6.4. In Figure 6.2, V_i reaches a fairly constant level of about 2821 (n_1 = mean of V_i for i between 5000 and 6000; n_1/n = 40%) at about g_{5000} (t = 5000). This means that if we would use p_{5000} = 0.509 as rejection level α, we can expect to have retained all the genes that are actually differentially expressed. Moreover, increasing the rejection level will only include genes whose expression is not affected

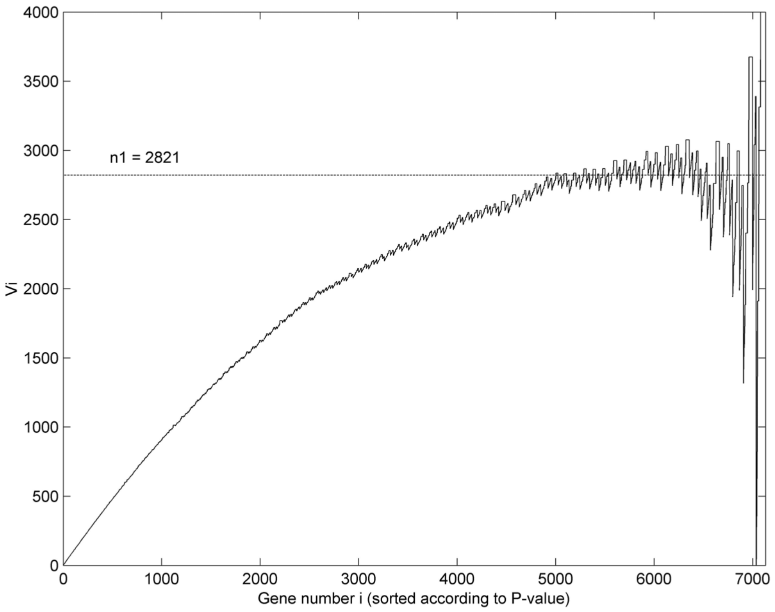


Figure 6.2: Analysis of the training set of Golub et al. Plot of V_i versus the gene number i (sorted according to their p-value). V_i reaches a constant level of about 2821 at g_{5000}, which is the estimate for n_1.

by the difference between ALL and AML and for which biological validation is not expected to yield any positive results.

The behaviour of the number of true positives TP_i ($= i - p_i.(n-n_1)$) in Figure 6.3 confirms these findings and gives additional proof that the calculated value of n_1 (in Figure 6.3 called $n_{1calc} = 2821$) is indeed the correct one. The correct curve for TP_i (curve in the middle in Figure 6.3) rises until g_{5000} and then reaches a constant level of 2821. If we evaluate the formula for TP_i with a value for n_1 that is smaller than 2821, this would result in a curve like the two lowest ones in Figure 6.3 (curve reaches a maximum level and then starts declining again). If we evaluate the formula for TP_i with a value for n_1 that is larger than 2821, this would result in a curve like the two upper ones in Figure 6.3 (curve keeps rising without reaching a constant level).

In the original paper of Golub et al. and also based on the patients of the training set alone, it was stated that roughly 1100 genes were more highly correlated with the AML-ALL class distinction than would be expected by chance (this number was derived using a method called neighbourhood analysis at an *arbitrary* level of significance - moreover, the number of false positives at this level of significance was derived (and

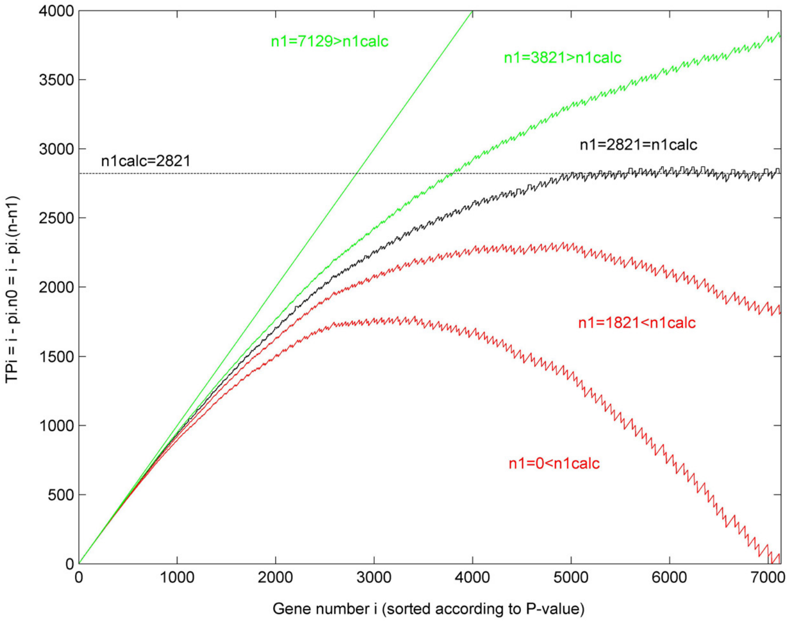


Figure 6.3: Analysis of the training set of Golub et al. Plot of the estimated number of true positives ($TP_i = i - p_i.(n-n_1)$) versus the gene number i for different values of n_1. If n_1 is set to the correct value $n_{1calc} = 2821$, the curve in the middle is obtained, which reaches a constant level of 2821, as expected. If n_1 is smaller than n_{1calc}, curves like the two lowest ones are obtained. If n_1 is set to a value larger than n_{1calc}, the result is like the two upper curves.

overestimated) by calculating the median number of genes that would accidentally reach this level of significance assuming that *none* of the genes were correlated with the class distinction). Our result suggests that this number should be more than doubled.

In Figure 6.4 one can inspect the sensitivity, specificity and false discovery rate plotted versus *i* and the ROC curve. The AUC equals 90.13% with a standard deviation of 0.41%. The point that maximizes the sum of the sensitivity and specificity (optimal sensitivity-specificity trade-off) has an associated rejection level α = 0.227 with a sensitivity of 86.29% and a specificity of 77.26%.

Randomised training set from Golub et al.

We randomly and independently permuted the components of each gene expression vector, resulting in a data set expected not to contain genes with actual differential expression between ALL and AML (the conditions or class labels remained constant) - also see Figure 6.1. After analysis, one can

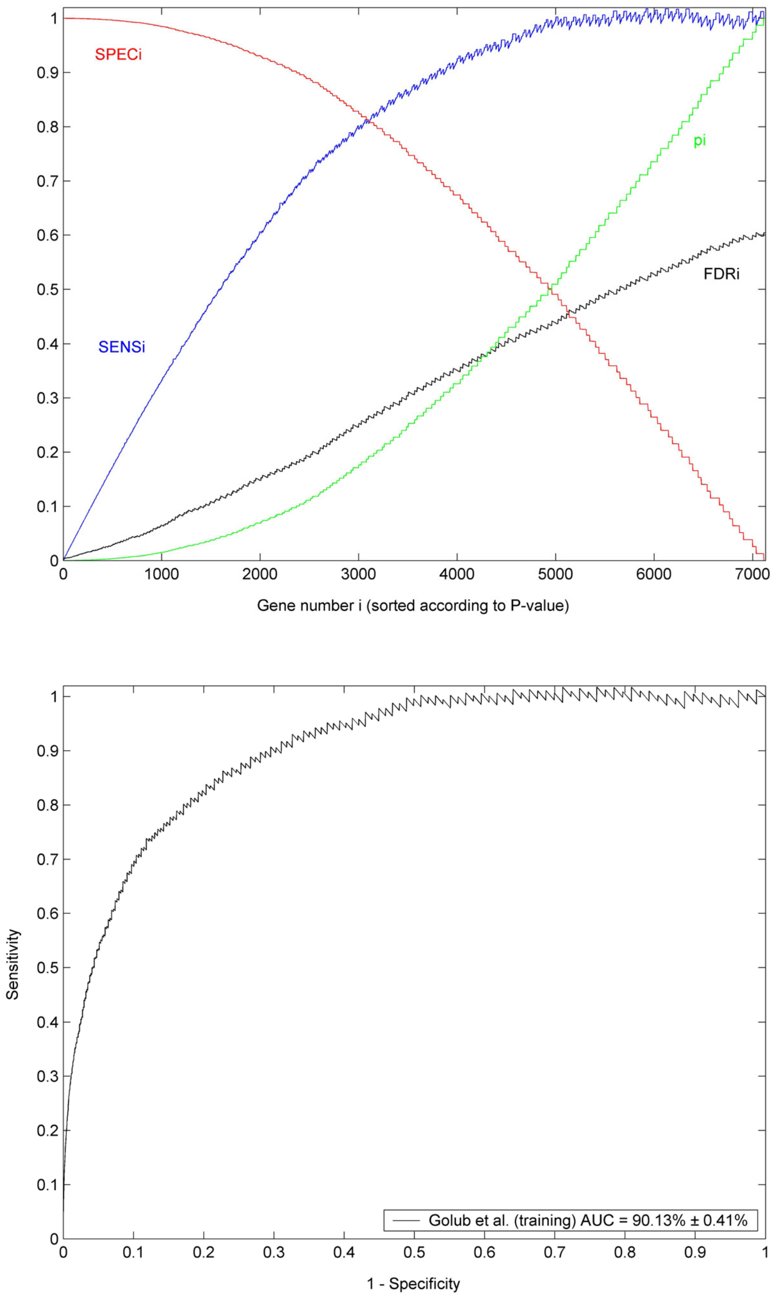


Figure 6.4: Analysis of the training set of Golub et al. Upper curves: sensitivity ($SENS_i$), specificity ($SPEC_i$), false discovery rate (FDR_i) and the p-values (p_i) versus the gene number i. Lower curve: ROC curve.

see (Figure 6.5) - as expected - that V_i reaches its constant level of approximately zero (so $n_1 \approx 0$, the null hypothesis is true for all the genes) starting from the first gene ($t = 1$), confirming that this data does not contain genes that, individually, contain real information about the difference between ALL and AML.

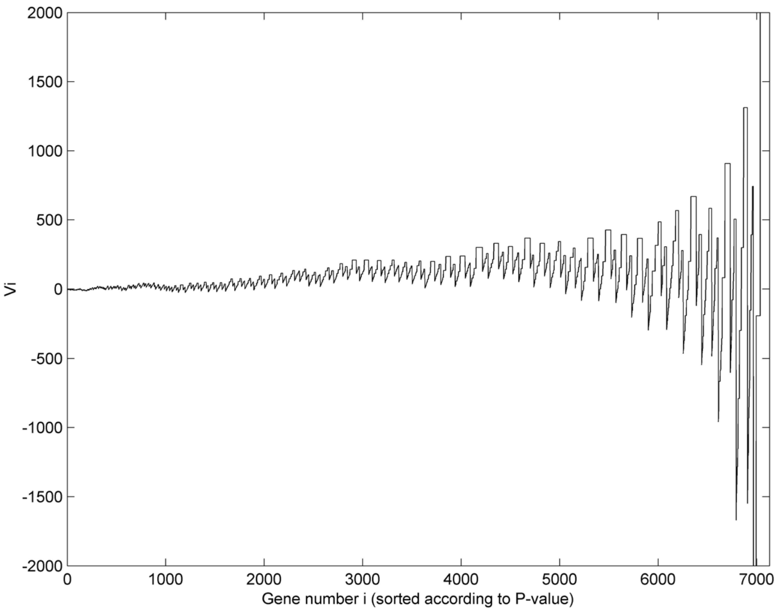


Figure 6.5: Plot of V_i versus the gene number i for the randomised training set of Golub et al. The constant level of V_i is approximately zero.

Simulated data

To construct an artificial data set we arbitrarily selected the gene expression profile from the non-randomised training set from Golub et al. that, after sorting according to the p-value, was on the 1000th place (= g_{1000}). This gene had a p-value of 0.015 and therefore can, *on its own*, be considered as differentially expressed between ALL and AML. Consequently, we superimposed noise to the components of this expression profile drawn from a uniform distribution in the range of [$-\sigma/4,\sigma/4$], where σ was the standard deviation of the components of g_{1000} ($\sigma = 396$). This was repeated 1000 times and resulted in 1000 expression profiles (with p-value ranging from 0.00079 to 0.38), which are, by design, not accidentally correlated with the class distinction ALL-AML and therefore can be considered actually differentially expressed. Finally, we added these 1000 expression profiles to the 7129 profiles without actual differential expression from the randomised training set described above, resulting in a data set with known values of n = 8129, n_1 = 1000 and n_0 = 7129. The distribution of p-values in this data set was similar to the distribution of p-values in all the real data sets we studied (see Figure 6.1).

A plot of V_i can be inspected in Figure 6.6. It reaches a constant level of about 1009 (mean of V_i for i between 1800 and 3000, which is our estimate for n_1) at the 1800th gene. Since, by design, we know the actual status for each individual gene expression vector in this data, we can calculate the real value of the false discovery rate and sensitivity at each level of rejection and compare this with the estimated false discovery and sensitivity by our method. This is done in Figure 6.7. The difference is minimal.

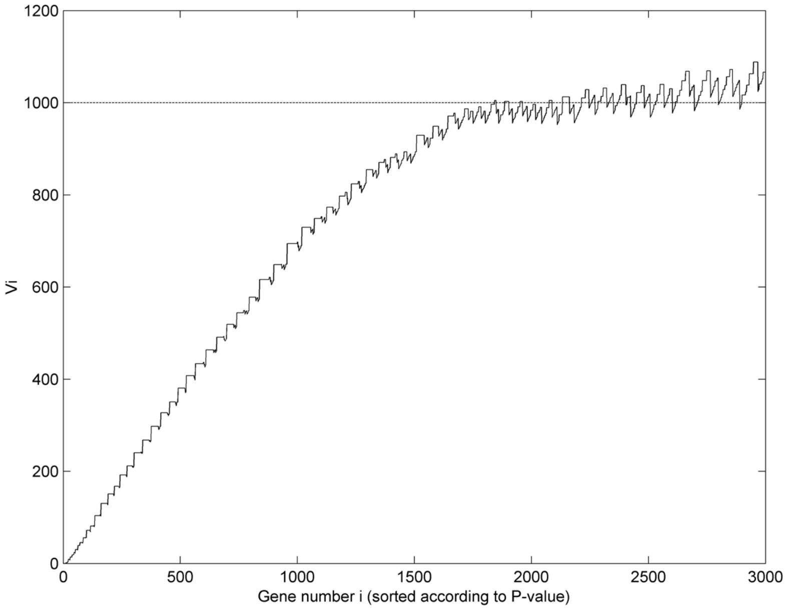


Figure 6.6: Analysis of the simulated data with known values for n_1 = 1000 and n_0 = 7129. Plot of V_i versus the gene number i. V_i reaches a constant level of about 1009, which is the estimated value for n_1.

Complete data set from Golub et al.

In order to compare the (complete) results of Golub et al. with the results of Armstrong et al. with respect to the detection of differential expression between ALL and AML, we first performed univariate analysis using all 72 patients from Golub et al. (training + test set). The results can be inspected in Table 6.2. A graph of the ROC and the false discovery rate can be inspected in Figure 6.8.

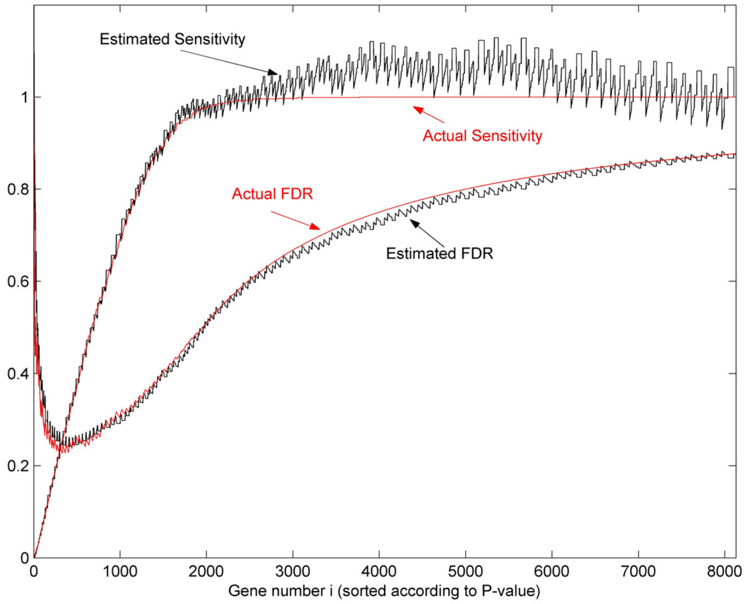


Figure 6.7: Analysis of the simulated data with known values for n_1 = 1000 and n_0 = 7129. Comparison between the actual values for the false discovery rate (FDR) and sensitivity and the estimated values for the FDR and sensitivity (derived using the estimated value of n_1 (Figure 6.6), Equations 6.11 and 6.14. and the formulas in Table 6.1). In this case, the estimated value for the sensitivity does not always stay below is theoretical limit of one. Calculating the value for the sensitivity as max $((i-p_i.n_0)/n_1$, 1) would reduce the difference between the actual and estimated sensitivity even more.

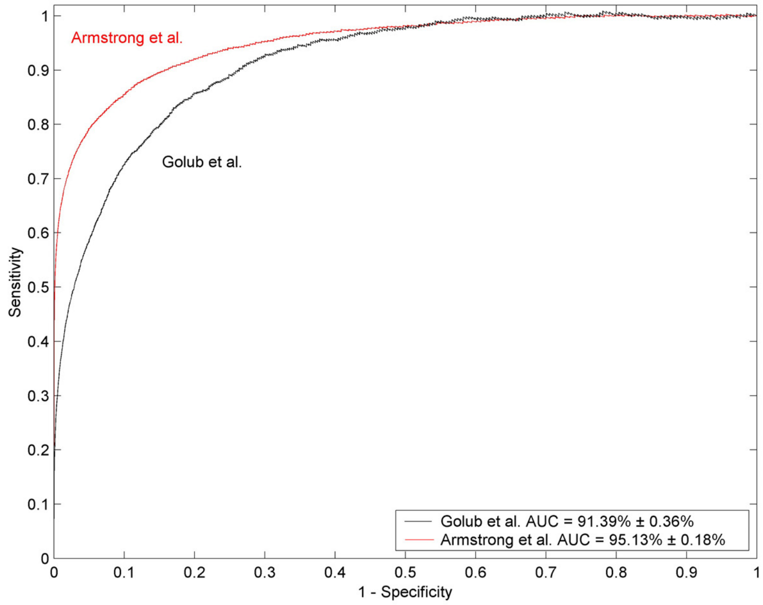
Comparison with the data from Armstrong et al. (ALL versus AML)

We removed the 20 MLL patients from the study of Armstrong et al. and analysed the resulting data (24 ALL patients and 28 AML patients) with respect to the detection of differential expression between ALL and AML. The results can also be inspected in Table 6.2 and Figure 6.8.

The AUC of the data from Armstrong et al. (95.13%) is significantly ($p < 0.0001$; two-sided, unpaired test (see Appendix A, Section A.2.3)) different from the AUC derived from the complete data set from Golub et al. (91.39%), which is reflected in the fact that the level of the optimal balance between (or, in our case, the maximum sum of) sensitivity and specificity is higher in the data from Armstrong et al. when compared to the data from Golub et al. (175.82% versus 166.09%).

Table 6.2: Results of the univariate analysis for the complete data from Golub et al. (detection of differential expression between ALL and AML) and from Armstrong et al. (detection of differential expression between ALL and AML, between ALL and MLL and between MLL and AML). n = total number of genes; n_0 = number of genes without actual differential expression; n_1 = number of genes with actual differential expression; AUC = area under the ROC curve; α^{opt} = rejection level where the optimal balance between specificity and sensitivity is reached (i.e., the rejection level that maximizes the sum of sensitivity and specificity - for the first two columns, these are also the rejection levels associated with the points on the ROC curves in Figure 6.8 with tangent lines with slope 1); $SENS^{opt}$ = sensitivity at α^{opt}; $SPEC^{opt}$ = specificity at α^{opt}.

	Golub et al. ALL-AML	Armstrong et al. ALL-AML	Armstrong et al. ALL-MLL	Armstrong et al. MLL-AML
n	7129	12582	12582	12582
n_0	3876	3084	8119	4527
n_1	3253	9498	4463	8055
AUC (%) [95% CI]	91.39 [90.68, 92.10]	95.13 [94.78, 95.48]	85.98 [85.24, 86.72]	94.83 [94.46, 95.20]
α^{opt}	0.18 (= p_{3429})	0.11 (= p_{8633})	0.22 (= p_{5193})	0.13 (= p_{7589})
$SENS^{opt}$ (%)	84.03	87.26	76.75	86.97
$SPEC^{opt}$ (%)	82.06	88.56	77.97	86.78
$SENS^{opt}$ + $SPEC^{opt}$ (%)	166.09	175.82	154.71	173.76



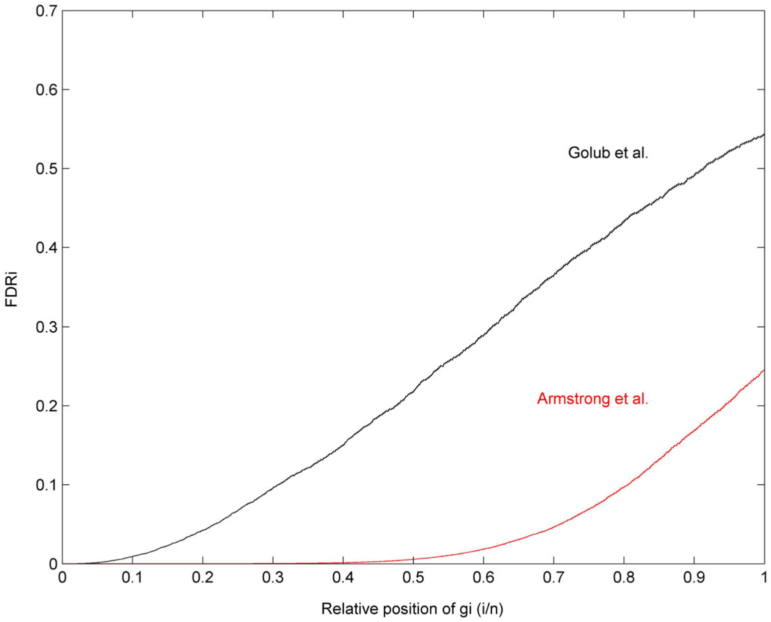


Figure 6.8: Comparison of the results of the univariate analysis for the complete data set from Golub et al. and for the data set from Armstrong et al. with respect to the difference between ALL and AML. Upper curves: ROC curves. Lower curves: false discovery rates versus the relative position of the genes (= *i*/*n*).

The fraction n_1/n is considerably higher in the data from Armstrong et al. (75%) than in the complete data set from Golub et al. (46%). As previously said, both the difference in AUC and in the relative value of n_1 have an independent impact on the relative behaviour of the false discovery rate of both studies (the false discovery rate for the data of Armstrong et al. starts increasing later and its maximum value is lower).

Effect of a change in condition

We analysed the data from Armstrong et al. with respect to the detection of differential expression between ALL and MLL (after removal of the 28 AML patients) and with respect to the detection of differential expression between MLL and AML (after removal of the 24 ALL patients) and compared this with the previous results with respect to the detection of differential expression between ALL and AML on the same data set. The results can also be inspected in Table 6.2.

The difference between MLL and AML did not result in any statistically significant change in AUC when compared with the difference between ALL and AML. However, the difference between ALL and MLL did result in a significant decrease in AUC when compared with the difference between ALL and AML (85.98% versus 95.13%, p<0.0001), which also resulted in a considerable decrease of the level of the optimal balance between sensitivity and specificity (maximum of sensitivity + specificity = 154.71% versus 175.82%), as could be expected.

6.3.2 Breast cancer: degree of differentiation

In this section we will compare two microarray data sets that study human breast tumours that are moderately or poorly differentiated (grade 2 or 3 - the degree of differentiation reflects the degree of anaplasia or the degree of malignancy of the tumour and is an important prognostic factor).

The first data set was published by Perou et al. (2000) (see Appendix B) and was already used in Chapter 3. In short, this data contains 21 microarray experiments with grade 2 and 36 with grade 3 breast tumours. In each experiment, the expression levels for 9216 genes were measured. A similar preprocessing strategy as was used in Chapter 3, Section 3.2.4 was followed, except the missing values replacement, which was omitted since missing values do not interfere with our analysis here (p-values can be calculated using only the values that are really present).

The second data set was produced by van ‘t Veer et al. (2002) and is also described in Appendix B. These authors studied primary breast tumours using a cDNA-microarray (24481 genes). In total 27 patients had a tumour with grade 2 and 78 patients had a tumour with grade 3. We did no further preprocessing, since this was already appropriately done.

For the study of Perou et al. and with respect to the detection of differential expression between grade 2 and grade 3 breast tumours, n_1 was calculated to be about 1306 (n_1/n = 14%) and the AUC was 87.99% ± 0.63%. For the study of van ‘t Veer et al. both n_1/n and the AUC were higher (n_1 was 10208 (n_1/n = 42%) and the AUC was 90.54% ± 0.21%, which was significantly different (p = 0.0001) from the AUC from Perou et al.) and explain the more optimal behavior of the associated false discovery rate. See Figure 6.9 for a comparison. Although the balance between Type I and Type II errors was better for rejection levels in the lower range, in this specific case the higher AUC for the study of van ‘t Veer et al. did not result in a dramatic improvement of the balance between Type I or Type II errors at α^{opt} in comparison with the study of Perou et al. This is caused by the fact that the two ROC curves almost coincide for rejection levels in the higher range and at α^{opt}.

6.3.3 Breast cancer: prognosis of sporadic lymph node negative patients

van ‘t Veer et al. also studied the expression signature of breast cancer patients with negative lymph nodes with a good prognosis (i.e., who did not develop distant metastases within 5 years - 51 patients) and a bad prognosis (i.e., who did develop metastases within 5 years - 46 patients). van ‘t Veer et al. developed a classifier based on the expression levels of 70 genes to distinguish between these two groups and proved it to be a powerful predictor (van de Vijver et al. (2002)). Clinically this is extremely important because this enables us to give adjuvant systemic therapy specifically to the patients who might benefit from it while withholding it from patients for which this might only mean unnecessary toxicity (presently, the available prognostic factors are not ideal to predict the clinical behaviour of this disease; on a clinical level, the phenotype of the two tumours is not that different). We used the procedure for univariate analysis to determine the total number of genes that are actually differentially expressed between good and bad prognosis breast tumours to see whether the differences on the molecular level between these two phenotypes are only subtle or whether we are dealing with tumour cells that are profoundly different. n_1 was calculated to be about 6449 (n_1/n = 27% - see Figure 6.10) and the AUC was 88.54% ± 0.28%.

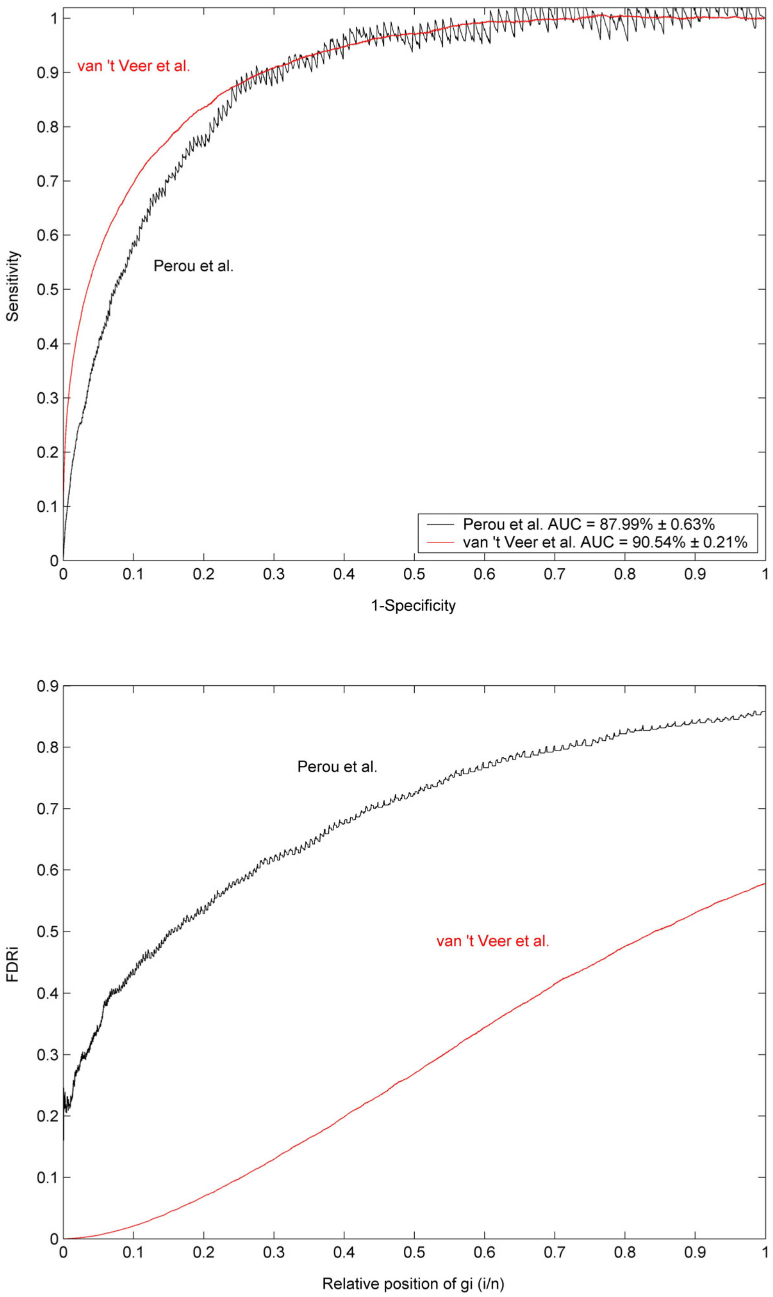


Figure 6.9: Comparison of the results of the univariate analysis for the data set from Perou et al. and for the data set from van ‘t Veer et al. with respect to the difference between grade 2 and 3 breast tumours. Upper curves: ROC curves. Lower curves: false discovery rates versus the relative position of the genes (= *i*/*n*).

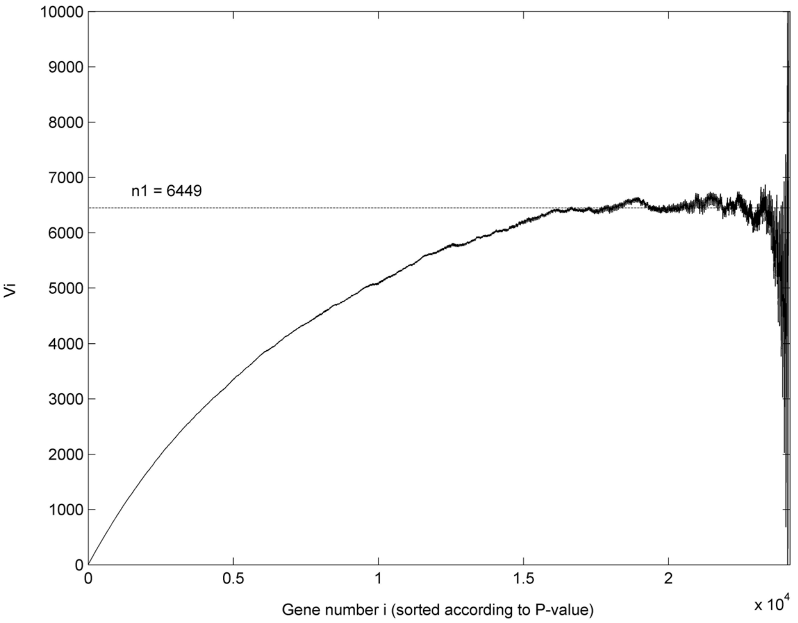


Figure 6.10: Analysis of the data from van 't Veer et al. with respect to the difference between good and bad prognosis patients. V_i reaches a constant level of about 6449 (our estimate for n_1) at about g_{16000}.

6.4 Discussion and conclusion

In this chapter we described a procedure for univariate analysis of microarray data that accounts for multiple testing starting from the p-values assigned by a certain hypothesis test to every gene. Here we have used p-values that were generated using the Wilcoxon rank sum test that does not assume a specific distribution of the gene expression levels. Troyanskaya et al. (2002) showed that this test was a robust and valid choice for studying differential expression and concluded that it was more conservative than two other nonparametric approaches. Moreover, this test does not require calculating test statistics in a great number of randomly permuted data sets (like in, for example, the nonparametric *t*-test described in Troyanskaya et al. or in SAM (Tusher et al. (2001)), which can be computationally intensive. Using this test in combination with our procedure therefore results in a method with low computational complexity, which takes at most a few minutes for the largest data sets.

In theory, using other procedures or tests to derive the p-values, can have an effect on the final result of our analysis. Therefore, as an example, we repeated all the analyses in the Results section using a two-sample (parametric) *t*-test (although we could be far from certain that its distributional assumptions were satisfied) instead of the Wilcoxon test. In general, this resulted in about the same values for n_1. In some cases the AUCs differed somewhat, but their ranking did not change, resulting in the exactly the same conclusions for our comparisons. Therefore, it is important to stress that a comparison of microarray data sets with respect to the detection of differential expression should only be done when the p-values in both data sets were derived using the same method or hypothesis test.

Using this set of p-values and independent of a certain rejection level, we described a method (determination of the constant level of V_i) for the estimation of the total number of genes that are and are not actually differentially expressed (n_1 and n_0) and therefore can be expected to be affected by the difference in conditions. We applied this method, among others, on a randomised and simulated data set and accurately estimated n_1 and n_0. We also used this method to see if, on a molecular level, profound differences exist between breast tumours with good and bad prognosis (Section 6.3.3). The results of this analysis seem to indicate that we are dealing with more than subtle changes between these two categories of tumours and that it should indeed be possible to accurately distinguish these two types of breast tumours using gene expression levels.

Subsequently, the estimates for n_1 and n_0 are used to assess the difference between actual and declared differential expression at each rejection level (i.e., to estimate the number of true and false positives and the number of true and false negatives). Using these estimates, the sensitivities, specificities, and false discovery rates can be calculated, which, in one way or another, reflect the quality of the prediction of actual differential expression at a certain rejection level. Finally, the knowledge of the sensitivities and specificities allow us to construct an ROC curve, which shows the trade-off or balance of the Type I and II errors at different rejection levels.

In contrast with current practice only to control the Type I error, ROC curves enable to optimally balance the Type I and Type II errors according to a certain criterion or cost function and enable, through the use of the AUC, to quantify our ability to discriminate between genes with and without actual differential expression in a specific data set using a certain hypothesis test. The AUC also reflects how well the Type I and Type II errors can be (optimally) balanced. We therefore propose to use the AUC as a quality measure to compare microarray data sets for their appropriateness to detect differential expression provided the same hypothesis test is used consistently. This quality measure could be used for different types of comparisons. We illustrated two of these comparisons.

As a first type of comparison, we investigated how this quality measure could be used to compare data sets that study the same conditions (ALL versus AML and grade 2 versus grade 3 breast tumours) but that originate from different sources or institutions. Firstly, after comparing the AUCs, we concluded that the data from Armstrong et al. is more appropriate to discriminate between genes that are and are not differentially expressed between ALL and AML than the data from Golub et al., although the last data set contained more experiments than the first (72 versus 52). In our opinion, the optimisation of the Affymetrix technology and protocol (year 2002 versus 1999) and perhaps a more optimal selection of the genes arrayed on the chip for Armstrong et al. could have contributed to this difference in quality, which was accurately detected by the rise in AUC. The methodology described here could be suited to compare the performance of different microarray platforms (e.g., cDNA-microarrays versus Affymetrix). Secondly, both the values for n_1/n and for the AUC indicate that the study of van ‘t Veer et al., when compared to the study of Perou et al., is of substantially higher quality to study differential gene expression in breast tumours with grade 2 or 3. Again, possible causes that could have attributed to this gap in quality are differences in technology, differences in experimental protocol and experimental setup, differences in surgical procedure and quality of the resected tumour biopsy, the choice of the genes on the array (more specifically chosen to study breast cancer in van ‘t Veer et al.) and so on. Since the determination of the degree of differentiation can vary between pathologists, this is also a factor that could have contributed. Both the absolute and relative value for n_1 are considerable (especially in the study of van ‘t Veer), suggesting that tumour cells with a different grade have a profoundly different phenotype.

As a second type of comparison, we examined what the effect on the AUC could be of a change in condition (replacement of ALL or AML patients by MLL patients). The difference between MLL and AML did not result in a significant decrease in AUC when compared to the difference between ALL and AML, while the difference between ALL and MLL did. The lower number of experiments that was available for the analysis of the difference between ALL and MLL (44 versus 52 for the analysis of the difference between ALL and AML) could have partially caused the significant drop in AUC, but this was, in a lesser extent, also true for the analysis of the difference between MLL and AML (48 patients), which did not show a drop in AUC. The behavior of the AUC and the results in Table 6.2 suggest that the degree of differential expression between ALL and MLL is less pronounced than the degree of differential expression between ALL and AML or between MLL and AML. This seems plausible, because the leukemic cells in MLL patients have a lymphoblastic morphology and have previously been classified as ALL. Again, this has been accurately detected by our analysis of the AUCs.

In our opinion, several other situations can be conceived where a comparison of AUCs could be informative, although we did not study them in detail here. For example, one could compare the AUCs of a data set for which the raw experimental data have been preprocessed or normalized - in order to remove different systematic sources of experimental variation from microarray data (also see Chapter 3, Section 3.2.1) - using different strategies (e.g., Lowess fit (Yang et al., 2002), ANOVA based methods (Kerr et al., 2000), …) and select a preprocessing strategy that results in a maximal AUC or maximal discrimination between the genes that are and are not differentially expressed.

Evaluation of the usefulness of additional experiments with respect to the detection of differential expression is another example where a ROC analysis could be valuable. Suppose one has done a basic set of microarray experiments (under two or more conditions) and suppose one performs a set of additional experiments in order to obtain a more optimal identification of the genes that are actually differentially expressed. Comparison of the AUCs of the basic set and of the basic + additional set could quantify if this has succeeded and could even help us to decide if more additional experiments would be beneficial (e.g., if the set of additional experiments has not resulted in a satisfactory rise in AUC, it could be expected that more additional experiments also will fail to do this).

Another situation where ROC analysis of microarray data could be useful is to select a test statistic, hypothesis test or method to calculate the p-values that gives a maximal AUC for *one specific* microarray data set. This is another setting than described in this chapter where we emphasised the comparison between *different* microarray data sets (evaluated using the same hypothesis test). In a recent publication, Broberg (2003) suggests such an approach, although the author uses a less refined method to estimate n_0 and another measure than the AUC to quantify the balance between Type I and II errors.

Finally, we have shown in this chapter that the relative value for n_1 (n_1/n) and the AUC can accurately summarise the behavior of the false discovery rate, which is a quantity that is often used to describe and control the Type I error. A higher value for n_1/n results in generally lower values for the false discovery rate and a lower maximum value for this quantity. For equal values of n_1/n, a higher value for the AUC results in lower values for the false discovery rate when the p-values are in the lower range (but the maximum value of the false discovery rate remains unchanged).

Chapter 7

Conclusions and future research

7.1 General conclusions and accomplishments

The application of the general data-mining framework to clinical and microarray data in this thesis has lead to several concrete results and observations, which we will summarize in this section. We will conclude this dissertation with a short description of some concrete clinical problems that will be studied in the future.

In the context of the prediction of deep myometrial invasion in endometrial cancer with ultrasound measurements and histopathological data, univariate and multivariate analysis have showed that Colour Doppler Imaging does not contribute to this prediction. Stepwise logistic regression analysis selects the degree of differentiation, the endometrial thickness and volume and the number of fibroids as significantly contributing in a logistic regression model. In a prospective study of limited size, we showed that a logistic regression model and LS-SVM models with linear and RBF kernel - based on the selected variables and in ascending level of performance - performed better than the subjective assessment of an expert ultrasonographer. This difference was only statistically significant for the LS-SVM model with an RBF kernel. In a concluding remark, we added a word of caution with respect to the clinical use of these models and noted that they should be evaluated using multicenter studies and regularly updated.

We applied the three elements of our data-mining framework to microarray data containing expression patterns from patients with acute leukemia (Golub et al., 1999) and from patients with breast tumours (Perou et al., 2000). In this context we implemented and used two methods to deal with missing values: missing values management without replacement and a nearest neighbour approach. We performed principal component analysis on these data and noted that for the data from Golub et al. and Perou et al., unsupervised selection of the principal components did and did not, respectively, capture the class difference under consideration (ALL versus AML for Golub et al. and grade 2 and 3 breast tumours for Perou et al.) and concluded that, for the data from Perou et al., supervised selection of principal components before classification would be a better option. Furthermore, cluster analysis of the data from Golub et al. succeeded in redefining the concepts ALL and AML when using a K-means algorithm based on the features after unsupervised PCA (which could have been expected since, as showed, the directions with the largest spread are also the directions in which the distinction between ALL and AML were prominent). In the context of cluster analysis of microarray experiments, we noted that due to the large number of possible cluster results and/or the presence of several a-priori hypotheses, multiple testing is a problem that has to be accounted for when interpreting a cluster result. Finally, in a systematic benchmarking study we evaluated the performance of several approaches to perform linear and non-linear binary classification with and without regularization and dimensionality reduction. We concluded that regularization or dimensionality reduction is necessary for the classification of microarray experiments. Moreover, we noted that, in general and within the bounds of our study, a non-linear LS-SVM model with an RBF kernel could be the model of choice to do class prediction with microarray experiments.

In a general overview of techniques related to the cluster analysis of gene expression profiles we noted that the properties of existing clustering algorithms complicate their use for this task. This includes the choice of user-defined and arbitrary parameter settings or the need for extensive parameter fine-tuning, inclusion of all the genes - even the ones that do not participate in the biological process under study - in a cluster, a high computational complexity and the lack of biological validation or ready to use implementation. These observations were the basis of the development of our own algorithm called adaptive quality-based clustering that was specifically designed to cluster gene expression profiles and to tackle some of the problems of the other algorithms. In summary, this algorithm, which was integrated in an on-line tool for microarray data analysis (INCLUSive), is a heuristic two-step approach in which the radius of a cluster is adapted to the local data structure after localisation of a cluster center. Among others, we applied the algorithm to a data set that studies the yeast cell cycle and biologically validated it by looking for clusters that have been significantly enriched with genes that belong to a certain functional category. We noted that the degree of enrichment in our result was significantly higher when this was compared to the most prominent and functionally matching clusters obtained by another group using K-means on the same data set.

The sixth chapter of this thesis was devoted to univariate analysis and the related problem of multiple testing in microarray data. We noted that the p-values for genes that are and are not affected by a certain difference between tumour classes overlap and that using a certain rejection level results in a number of false positive and negative results. After calculation of these p-values using a certain hypothesis for every gene, we showed, based on a plot of a simple quantity and independent from a certain rejection level, how to estimate the number of genes that are and are not differentially expressed in different tumour types. Moreover, we showed that this approach is completely equivalent with a method recently published in PNAS. These estimates can subsequently be used to derive the number of true positives and negatives, the number of false positives and negatives, the sensitivity, the specificity and the false discovery rate for every possible rejection level and to construct an ROC curve. In contrast with current practice only to control the Type I error, we described how this ROC curve could be used to define a rejection level that results in an optimal balance between the Type I and II error according to a certain criterion or cost function that describes the relative importance of a false positive versus a false negative result. Moreover, we proved that the area under the ROC curve could be used as a quality measure for microarray data with respect to its ability to detect differential expression that quantifies the amount of overlap between the p-values of the genes that are and are not actually differentially expressed. Using this quality measure, we demonstrated, among others, that the data from Armstrong et al. (2002) is more suited to discriminate between genes that are and are not differentially expressed between ALL and AML than the data from Golub et al. Moreover, we showed that the degree of differential expression between MLL (a third class of acute leukemias) and ALL is less pronounced than the degree of differential expression between ALL and AML or between MLL and AML. In a second test case, we concluded that the study of van ‘t Veer et al. is of substantially higher quality to study differential expression between grade 2 and 3 breast tumours than the study of Perou et al.

In this thesis we have used ROC curves in two contexts with a subtle difference between them. Firstly, ROC curves were applied to test or compare the ability of univariate data or single valued output of a model to discriminate between patients belonging to two classes. In this case the class membership is known for each patient or data point individually. Using the class labels and the value for the univariate variable or model output, the number of true positives and negatives and the number of false positives and negatives (and hence the sensitivity and specificity) can be derived exactly (by simple counting) for every possible cut-off level and set of data points at hand. Secondly, ROC curves were applied to test and compare the ability of p-values - assigned using a certain hypothesis test to every gene in a microarray data set - to discriminate between genes that are and are not actually differentially expressed. In this context, however, the class labels (i.e., the actual status of differential expression) for the individual data points or genes are not known or taken into account. In this case the ROC curve is constructed through an estimate of the sensitivity and specificity for every possible rejection level. Although it is possible to estimate the number of true positives and negatives and the number of false positives and negatives for every rejection level using microarray data alone, it is impossible to predict which individual genes exactly are true positive or negative or are false positive or negative. This means that the algorithm (and associated MATLAB script) used to construct the ROC curves in the first context (where the input of the algorithm consists of the class labels for every data point and the associate model output or value for the univariate variable) needed to be adapted to be useful for constructing ROC curves in the second context (where the input consists of the estimate of the number of true positives and negatives and the number of false positives and negatives for every possible rejection level).

7.2 Future research

In this section we will first discuss some specific ongoing or submitted project proposals in which we are involved. In this research we aim to apply some of the techniques described in this thesis for concrete clinical problems. Two of these projects involve the use of proteomic data that have not been explicitly analysed in this dissertation and that are, as stated in Chapter 1 (Section 1.2), qualitatively similar to microarray data with respect to the use of our methodology.

At the end of this section, we will briefly examine some general research prospects.

7.2.1 Specific future research

Ovarian cancer: transcriptomics

Ovarian cancer accounts for 4% of new cases of cancer and for 6% of cancer deaths in women. The prognosis of the disease is generally poor with an overall five year survival of approximately 30%. Approximately 85-90% of ovarian neoplasms are of epithelial origin (derived from tissues that come from the mesothelium). These tumors may be benign (50%), malignant (33%), or borderline malignant (16%). The serous histologic type is the most common epithelial tumour of the ovary (46-75%) and will be the focus of our attention here. About 30% of ovarian cancer patients are diagnosed with early-stage disease and about 10%-50% of them will have a recurrence after initial surgery. Most women with advanced disease will respond to initial (chemo)therapy but most of them will eventually relapse.

Presently, no clinical parameters are available that can reliably predict chemosensitivity in FIGO stage III ovarian cancer (tumour with abdominal extension or extension to regional nodes) or the probability of recurrence after initial surgery in FIGO stage I ovarian cancer (tumour limited to one or both ovaries). Therefore we (in cooperation with Prof. I. Vergote and Prof. D. Timmerman, department of Obstetrics and Gynaecology of the University Hospitals Leuven, and Dr. P. Van Hummelen of the Microarray Facility of the Flanders Interuniversity Institute for Biotechnology (V.I.B.)) aim to develop and test models that use cDNA-microarray data and that:

1. Predict if a stage III ovarian tumour will relapse within 6 months after the last therapeutic intervention. Since standard chemotherapy for advanced ovarian cancer is usually platinum based (e.g., carboplatinum + paclitaxel), this model will be able to predict platinum resistance (or chemosensitivity of the tumour). This has mainly prognostic significance but might allow to develop new therapeutic strategies in the future for tumours that are predicted not to respond adequately to the standard chemotherapeutic regimen.
2. Predict if a stage I ovarian tumour will have a recurrence after initial surgery. The subset of women with early-stage disease and, according to our model, with a high probability of recurrence are ideal candidates that might maximally benefit from adjuvant treatment (chemotherapy and/or lymphadenectomy) while the women with early-stage disease and a low probability of recurrence might be spared the side-effects of adjuvant therapy.

In the first phase of the study, the models will be trained using appropriate training sets of microarray experiments (100 are planned). This will include expression patterns from tumour samples obtained after initial surgery from patients with stage III disease that have relapsed within 6 months after the last therapeutic intervention, from patients with stage III disease that have had a therapy-free interval of minimum 12 months and from stage I patients that have and have not had a recurrence. In a second phase, the models will be validated using data from additional microarray experiments with new tumour samples (100 additional experiments are planned). The resulting model predictions will be compared with the true outcome of the patients in order to evaluate what the usefulness of the models in real clinical practice would be. If, during this validation phase, the predictive power of some of the models trained in the first phase would seem inadequate, the additional microarray experiments could be used to refine the first version of the models.

Another aim of this project is to identify differentially expressed genes between the different diagnostic classes considered (univariate analysis) and that might represent clinically useful biomarkers. All 200 microarray experiments that are planned will be available for this analysis.

At this moment we are preparing to perform the first 21 experiments. We have carried out an extensive search in literature (more than 80 papers were screened), several databases that are publicly available (e.g., LocusLink, OMIM) and other on-line sources to discover known genes that are involved in the distinction between several classes of ovarian tumours. This search resulted in a list of about 5000 UnigeneID's of which about 85% was finally spotted onto the microarray. In our opinion, this effort was necessary to ensure that the microarray will be sufficiently enriched in ovarian cancer related genes. RNA extraction and amplification was already performed for 14 stage III tumours (7 with and 7 without relapse) and 7 stage I tumours. In first instance, we have chosen to use a common pool of reference RNA for all the experiments (classical reference design – with colour flip). Sufficient reference RNA (obtained from the first 21 test samples and from a limited number of ovarian tumours for which sufficient tissue was available) was extracted to provide for about 200 experiments. At this moment the hybridisation and labelling process is being refined and the expression patterns from the first 21 experiments should be available soon.

Endometriosis: proteomics and transcriptomics

Endometriosis is an important and benign gynaecological disorder associated with pain and infertility and is defined as a benign proliferation of endometrial tissue outside the uterine cavity. This condition can be found in 80% of women with dysmenorrhea (discomfort or pain during menstrual bleeding), dyspareunia (pain during sexual intercourse) and/or chronical pain in the lower abdomen and in about 50% of women with subfertility. This disease can be diagnosed through laparoscopic surgery, which is an invasive procedure that can visualise the involvement of the internal genitalia. The lesions can be minimal but can also consist of large endometriosis cysts and extensive adhesions that can distort the organs involved and deform the anatomy of the small pelvis. Therefore endometriosis is classified in four stages: minimal, mild, moderate and severe. This disease cannot be cured completely. Surgery can improve the symptoms like pain and infertility but relapse is frequent (50%), certainly in severe forms. Hormonal treatment can inhibit the lesions but has important side effects and the disease recurs when the treatment is interrupted.

In this research we (in cooperation with Prof. T. D'Hooghe, coordinator Leuven University Fertility Center) aim to analyse transcriptomic patterns (measured with microarrays), proteomic patterns and possibly clinical data related to the study of endometriosis. The transcriptomic and proteomic patterns will be obtained from normal eutopic (from the uterus itself) endometrium from women with and without endometriosis. The tissue samples were or will be acquired through an endometrial biopsy taken during general anaesthesia for surgery (with hysteroscopy and laparascopy, planned for pain or subfertility) or taken during consolation (Pipelle de Cornier) on an outpatient basis. Due to the possible effect of the menstrual cycle on the state of the endometrium, we will only analyse samples obtained during the luteal phase and preferably samples histologically dated on day 19-21 of the menstrual cycle. These samples will be specifically selected from a tissue bank constructed for this study.

We plan to perform 100 microarray experiments: 25 using endometrium from women with a normal pelvis, 25 using endometrium from women with minimal-mild endometriosis (of which minimally 10 are treated for pain and minimally 10 are treated for subfertility), 25 using endometrium from women with moderate-severe endometriosis without relapse within 2 years after surgery, and 25 from women with moderate-severe endometriosis with relapse within 2 years after surgery. Moreover, endometrial biopsies originating from the same patients will be analysed by the technology described in Section 1.2. to measure proteomic patterns.

Since we will study eutopic endometrium of patients with and without endometriosis and since women with (moderate-severe) endometriosis will be subdivided in a group with and a group without relapse after surgery, two binary classification problems can be defined using these transcriptomic and/or proteomic patterns: prediction of absence or presence of endometriosis and prediction of absence or presence of relapse after surgery. These models might help the clinician in detecting endometriosis and in assessing its prognosis using only eutopic endometrium. In a first phase, we aim to construct models (and compare their performances) that are based on microarray data or on proteomic data alone. In a next phase, we will investigate if it is possible to further optimise the predictions by combining microarray and proteomic data, potentially complemented with clinical data. The results of this combined approach will be compared to the results of the analysis of proteomic, transcriptomic or clinical data alone. This comparison will possibly allow assessing the complementarity of the different data sources with respect to clinical predictions.

Moreover, it might be possible to compare the microarray data set with the corresponding proteomic data set using an approach introduced by

Alter et al. (2003). They describe a method based on Generalized Singular Value Decomposition (GSVD) to compare two microarray datasets of different origin. Simplified, the goal is to identify fundamental gene expression profiles that are present in one or in both datasets, and that represent biological processes exclusive for one dataset or common between both. The main condition for this method to be useful is that experiments in both datasets need to be paired to each other (for any experiment in one of the data sets, there is a corresponding experiment in the other). This makes the setup in the work of Alter et al. methodologically equivalent with the setup as proposed in our project (here microarray and proteomic data are paired, since they originate from the same patients). In our setting, this method makes it possible to detect fundamental patterns that are present in the microarray data and not in the proteomic data or vice versa, or to detect patterns that appear in both data sets. This again can provide information concerning the complementarity of microarray and proteomic data and can provide information concerning the correspondence and differences between processes that take place at the level of the transcriptome or proteome.

Finally, another important goal of this project again concerns the identification of genes that show a different RNA expression between the classes under study and, on the other hand, the identification of mass/charge values corresponding to peak amplitudes (and the corresponding proteins) that differ between the classes (univariate analysis – identification of biomarkers). Moreover, multivariate feature extraction methods (e.g., PCA) might be able to identify combinations of gene expression levels (microarray data) and peak amplitudes (proteomic data) that might result in more optimal separation between the classes.

Cervical and endometrial cancer: proteomics

In this research project (submitted by Prof. Vergote - we were asked to collaborate for data analysis) proteomic patterns in serum and tissue samples of patients with endometrial and cervical cancer will be obtained and analysed. Again, in this study we wish to develop mathematical models that can provide prognostic information (e.g., prediction of the presence of subclinical metastases, prediction of response to chemo- or radiotherapy) and we aim to discover new biomarkers with different behaviour between patients with a different prognosis. The study of proteomic patterns in serum might lead to markers that can be easily determined by a simple blood sample (while this might not be the case for biomarkers identified through tissue sampling since these proteins might not be secreted and therefore could only be determined through a more invasive procedure or biopsy).

7.2.2 General research prospects

When we wrote our article in 'Tijdschrift voor Geneeskunde' (De Smet et al., 2001) we predicted how genome-wide analysis technologies like microarrays could be used in guiding clinical management in oncology. We described, for example, how microarrays might be used to distinguish between tumours with and without metastatic phenotype, to predict therapy response or to provide prognostic information that is impossible to derive from clinical parameters only. At the moment of writing, these examples were of a rather hypothetical nature and not yet supported by concrete cases in literature. At this moment however, several publications have appeared that confirm the potential clinical applicability of microarrays we hypothesized earlier (e.g., van 't Veer et al.,2002; Ramaswamy et al., 2003; Chang et al., 2003).

While these publications clearly prove that microarrays could be an invaluable clinical tool, a considerable amount of work and research needs to be done before widespread use of expression patterns in real clinical practice is feasible. Several issues or problems need to be addressed in this context. First of all, most of the models have been developed and tested using a limited number of patients. Before reliable statistical conclusions can be drawn, microarray data sets need to contain a sufficient amount of technical and biological replicates (e.g., in order to account for technical variation, inter-individual variation (which can be considerable in humans), variation in the composition of the tissues analysed (tissue heterogeneity)). Moreover, mathematical models need to be validated in prospective clinical trials where larger patient groups are studied. Furthermore, there is the issue of standardization (Tumor Analysis Best Practices Working Group, 2004). Since the experimental procedure (e.g., surgical procedures, tissue processing, RNA extraction, labelling, data preprocessing, and so on) can vary extensively from place to place and can have a significant impact on the data, clinical models reported by one group are not directly applicable in other centers. Moreover, the use of a uniquely constructed reference pool in cDNA-microarrays makes extended use of the derived models impossible. Before widespread implementation into clinical practice of algorithms based on expression patterns is possible, detailed experimental guidelines and standards have to be agreed upon.

As previously mentioned, microarrays do not capture all relevant phenomena in a cell on a molecular level and by studying the proteome it is possible to obtain more information about the phenotype of a (tumour) cell. Moreover, since microarrays measure intracellular RNA levels, tissue samples are always needed, which can be difficult or impossible (e.g., if macroscopic tumour residues are not longer present in a patient) in some situations. Since tumour cells can exhibit aberrant secretion of several proteins, the study of proteomic patterns in serum (see the study of cervical and endometrial cancer in the previous section) could be helpful in these cases. The use of proteomic patterns could therefore be the next step in the integration of high throughput technologies into the clinical decision making process. Moreover, microarray, proteomic and possibly clinical data might be, at least partially, complementary and a combined analysis might improve the clinical performance of the resulting methods (also see the study of endometriosis in the previous section).

From a mathematical point of view, some techniques applied to high dimensional biological data might merit further investigation in the future. These include: classifiers that combine different data types (microarray, proteome and clinical data - e.g., committee networks), independent component analysis (ICA), the combination of model selection techniques with other methods for feature extraction, the use of different distance measures and kernel-based algorithms in clustering, the use of GSVD or canonical correlation analysis (CCA) to compare microarray and/or proteome data sets and the use of meta-analysis techniques to analyse data from different sources.

In conclusion, the use and development of the techniques mentioned in this thesis for the analysis of patient specific transcriptomic and proteomic patterns and the implementation of the results into clinical practice will be and remain the main focus of our research.

Appendix A

Methods

In this appendix we will give some technical details about some of the methods that have been applied or referred to on multiple occasions throughout this thesis. The following methods will be discussed: hypothesis testing and Bonferroni correction, receiver operating characteristic curves, logistic regression and model selection, least squares support vector machines, K-means clustering, and hierarchical clustering.

A.1 Hypothesis testing and Bonferroni correction

Hypothesis testing examines the belief in a certain property of a population parameter (or populations parameters) based on the data in a statistical sample (Dawson-Saunders and Trapp, 1994). Suppose, for example, we want to examine if the true or population mean μ_x of a certain variable *x* (e.g., the mean cholesterol level in all patients with cardiovascular disease) is equal to a given value (e.g., 190 mg/dl) based on the measurements of this variable in a certain sample (e.g., the measurement of the total cholesterol levels in N=100 patients). The sample mean and standard deviation are noted as m_x and s_x, respectively (e.g., the mean and standard deviation of the total cholesterol levels in our sample of 100 patients). Hypothesis testing involves the following steps:

1. Definition of the null and alternative hypothesis: the null hypothesis H_0 states that there is no difference between the population parameter and its hypothesized value. In our example the null hypothesis states that μ_x=190 mg/dl. The alternative hypothesis H_1 states the opposite: $\mu_x \neq$190 mg/dl.

2. Definition of a test statistic that reflects in one way or another how the sample at hand deviates from the null hypothesis and for which the distribution is assumed to be know if the null

hypothesis is true. In our example we can define the following test statistic and calculate its value for our sample of 100 patients:

$$t_x = \frac{m_x - 190}{s_x / \sqrt{N}}. \tag{A.1}$$

Under the null hypothesis and if x is normally distributed, this test statistic follows a t-distribution with N-1 degrees of freedom.

3. Calculation of the p-value: the p-value equals the probability that, under the null hypothesis, the test statistic will have a value that is as extreme as or more extreme than the test statistic for the sample at hand. This can be easily calculated since the distribution of the test statistic under the null hypothesis is known. In our example, the p-value is given by the probability that the test statistic is larger or equal than $|t_x|$ plus the probability that the test statistic is smaller or equal than $-|t_x|$, which in this case can be computed by calculating the appropriate areas under the t-distribution.

4. Drawing the final conclusion: if the calculated p-value is smaller than a predefined rejection level α (usually set at 5%), it is unlikely that the sample at hand was generated under the null hypothesis. In this case the null hypothesis is rejected in favor of the alternative hypothesis. One can state that the test result is significant, i.e., according to the evidence presented by the sample, one can conclude that the population parameter is different from its hypothesized value. In our example this would mean that we conclude that the mean cholesterol level in patients with cardiovascular disease is different from 190 mg/dl.

   If, on the other hand, the p-value is larger than the rejection level α, the null hypothesis is not rejected and there is not sufficient evidence to accept a real difference between the population parameter and its hypothesized value.

It should be noted that if there is no real difference between the population parameter and its hypothesized value, it is still possible that the null hypothesis will be (erroneously) rejected. The probability of falsely rejecting the null hypothesis is called a Type I error and its probability is given by the rejection level that is applied. If multiple tests are performed simultaneously, the probability that at least one test is declared significant due to chance increases. A common method to protect against this is to apply a Bonferroni correction (Keselman et al, 2002). In this procedure the rejection level that is applied for every individual test is set equal to the original rejection level α divided by the number of tests performed simultaneously (e.g., $0.05/n_s$, where n_s is the number of tests performed at the same time). It can be proven that a Bonferroni correction guarantees that the probability of committing at least on Type I error (also called the family-wise error (FWE)) will not be larger than α.

A.2 Receiver Operating Characteristic curves

A.2.1 Definition, use and interpretation

Suppose we have a set of objects (e.g., patients, genes, microarray experiments) that belong to one of two classes. Suppose that objects that do not and do exhibit a certain property belong to class 1 and class 2, respectively (e.g., patients without and with a certain disease, genes without and with differential expression - in some situations it is appropriate to call the objects of class 1 normal and the objects of class 2 abnormal). Also suppose that each object *i* is associated with a single value or variable y^i (e.g., the output of a model, a p-value, a measurement - i=1,...,N) that is generated to predict the class membership of this object.

Now consider a certain threshold or cut-off level *T*. If $y^i > T$, the test result for object *i* is declared positive (i.e., object *i* is predicted to belong to class 2). If $y^i \leq T$, the test result for object *i* is declared to be negative (i.e., object *i* is predicted to belong to class 1). If we compare the test results with the actual class memberships of the objects, four categories emerge: true and false positive and true and false negative objects. These categories are defined in Table A.1. Subsequently, the number of objects in each of these categories can be used to define the sensitivity ($TP/TP+FN = TP/N_A$) and specificity ($TN/TN+FP = TN/N_N$), which summarize the correlation between the test results and the actual class memberships for the set of objects under consideration.

The sensitivity and specificity are dependent on the choice for the threshold *T* and can be recalculated for other values of *T*. In this context, there is a trade-off between sensitivity and specificity because each change in the threshold that results in a higher sensitivity will also result in a lower specificity and vice versa. The plot of the sensitivity versus 1 - specificity (1 - specificity is also called the false positive rate) for varying values of *T* is called a Receiver Operating Characteristic (ROC) curve (Dawson-Saunders and Trapp, 1994; Swets, 1996) for the set or sample of objects under consideration. For an example, see Figure A.1. An ROC curve therefore, summarizes the trade-off between sensitivity and specificity for all possible values of the threshold *T* in one single plot and can therefore be used to

Table A.1: Definition of True and False Positive objects (*TP* and *FP*) and of True and False Negative objects (*TN* and *FN*) for a certain choice of the threshold *T*. N_N = number of objects belonging to class 1; N_A = number of objects belonging to class 2.

		Class membership	
		Class 1 (Normal)	Class 2 (Abnormal)
Test result	Positive ($y^i > T$)	*FP*	*TP*
	Negative ($y^i \le T$)	*TN*	*FN*
		N_N	N_A

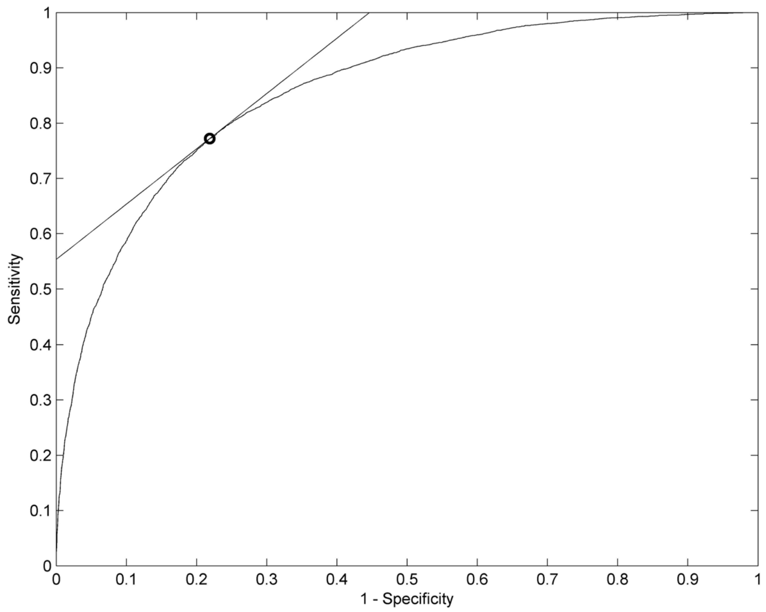


Figure A.1: Example of an ROC curve. A small circle indicates the point on the ROC curve that maximizes the sum of the sensitivity and specificity and that has a tangent line with slope 1.

select an optimal (according to a certain criterion) threshold or cut-off point T^{opt}. In this thesis we consistently use the point that maximizes the sum of the sensitivity and specificity and for which it can be proven that the associated point on the ROC curve has a tangent line with slope 1 (see Figure A.1).

The area under the ROC curve (AUC) has a special meaning (Hanley and McNeil, 1982). It is a measure for the ability of the variable under consideration to discriminate between objects from class 1 and class 2. If this variable represents the output of a certain model for example, the AUC quantifies the discriminatory power or accuracy of the associated model. Suppose we randomly select an object from class 1 with associated value y_N and that we randomly select an object from class 2 with associated value y_A. Then, it can be proven that the AUC equals the probability that $y_A > y_N$. Said otherwise, the AUC equals the probability that an object randomly selected from class 1 and an object randomly selected from class 2 are ranked correctly. It reflects the degree of overlap of y_i for objects from class 1 and class 2. The AUC does not depend on the choice of the threshold T. Two extreme situations are possible. If the variable under consideration has no discriminatory power whatsoever, the AUC will equal 0.5. If, on the other hand, the variable under consideration can result in a perfect classification of the objects, the AUC will equal 1.

In the next sections we will discuss how the AUC can be derived from a finite sample and how different AUCs can be compared. All the methods are available in the form of own MATLAB scripts. Worth mentioning is that some of these scripts were integrated in LS-SVMlab (see http://www.esat.kuleuven.ac.be/sista/lssvmlab/ and Suykens et al. (2002)).

A.2.2 Estimation of the AUC from a finite sample

Suppose we have a finite sample S consisting of a subset S_N with N_N normal objects and a subset S_A with N_A abnormal objects. Suppose we want to estimate the true AUC (i.e., the AUC for an infinite sample), denoted as A_{ROC}, using this finite sample. This estimate of the AUC, denoted as $\hat{A}_{ROC}$, can be obtained by calculating the Wilcoxon statistic W given by (Hanley and McNeil, 1982):

$$W = \frac{1}{N_A . N_N} \sum_{i \in S_A} \sum_{j \in S_N} S_c(y^i, y^j) = \hat{A}_{ROC}, \tag{A.2}$$

where S_c is a scoring function given by:

$$S_c(y^i, y^j) = \begin{cases} 1 & \text{if} \quad y^i > y^j \\ 1/2 & \text{if} \quad y^i = y^j. \\ 0 & \text{if} \quad y^i < y^j \end{cases} \tag{A.3}$$

The estimate $\hat{A}_{ROC}$ is a stochastical variable (it depends on the specific finite sample *S*) with mean A_{ROC} (the true are AUC) and standard error *s* given by:

$$s_e = \sqrt{\frac{A_{ROC}(1 - A_{ROC}) + (N_A - 1)(Q_1 - A_{ROC}{}^2) + (N_N - 1)(Q_2 - A_{ROC}{}^2)}{N_A N_N}}, \tag{A.4}$$

where Q_1 is the probability that two randomly chosen abnormal objects will both be ranked with greater suspicion than a randomly chosen normal object and where Q_2 is the probability that one randomly chosen abnormal object will be ranked with greater suspicion than two randomly chosen normal objects. Q_1 and Q_2 can be estimated from the finite sample but can also be approximated by the following equations:

$$\begin{aligned} \hat{Q}_1 &= \frac{\hat{A}_{ROC}}{2 - \hat{A}_{ROC}} \\ \hat{Q}_2 &= \frac{2\hat{A}_{ROC}^2}{1 + \hat{A}_{ROC}}. \end{aligned} \tag{A.5}$$

Replacing Q_1, Q_2 and A_{ROC} by their estimates (obtained in Equation A.5 and A.2) in Equation A.4 results in an estimate $\hat{s}_e$ for the standard error of $\hat{A}_{ROC}$.

A.2.3 Comparison of the AUC

Suppose that we have two variables y_1 and y_2 (e.g., given by the output of two different models) that have been generated to distinguish between objects from class 1 and 2. In this section we want to examine whether the discriminatory potential of these two variables is different, i.e., whether there is a difference in the respective true AUCs: A_{ROC1} and A_{ROC2}. Since, in practice, we can only estimate these AUCs ($\hat{A}_{ROC1}$ and $\hat{A}_{ROC2}$) and their standard errors ($\hat{s}_{e1}$ and $\hat{s}_{e2}$) from a finite sample, we have to investigate whether there is a significant difference between these estimates. In this context, two designs are possible: unpaired and paired (Hanley and McNeil, 1983).

Unpaired design

In an unpaired design, the AUCs for the two variables y_1 and y_2 are estimated from two different finite samples S_1 and S_2, respectively, that do not contain the same objects. In this case it is assumed that under the null hypothesis (that states that the true AUCs are equal), the following statistic follows a standard normal distribution:

$$z = \frac{\hat{A}_{ROC1} - \hat{A}_{ROC2}}{\sqrt{s_{e1}^2 + s_{e2}^2}}, \tag{A.6}$$

which can be used to calculate the probability or p-value that an equally large or larger value for z (or $|z|$ if a two-sided test is used) will be obtained if the null hypothesis is true. If this p-value is smaller than a certain rejection level (e.g., 5%), the null hypothesis is rejected and the estimates of the AUCs are declared significantly different.

Paired design

In a paired design, the AUCs for the two variables y_1 and y_2 are estimated from the same finite sample *S, i.e.*, the values for variables y_1 and y_2 are both available for the objects belonging to *S*. This situation, for example, is often encountered when the discriminatory performance of different mathematical models is being compared because usually the models can be evaluated using all available objects. In general and if possible, a paired design is preferred in comparison with an unpaired design, since a paired design results in an increase in statistical power (i.e., a true difference between the AUCs will be detected with a higher probability - i.e., the Type II error is lower).

In a paired design the estimates $\hat{A}_{ROC1}$ and $\hat{A}_{ROC2}$ are no longer independent but are positively correlated and an adapted z-statistic can be applied:

$$z = \frac{\hat{A}_{ROC1} - \hat{A}_{ROC2}}{\sqrt{\hat{s}_{e1}^2 + \hat{s}_{e2}^2 - 2.r.\hat{s}_{e1}\hat{s}_{e2}}}, \tag{A.7}$$

where r is a quantity that represents the correlation between $\hat{A}_{ROC1}$ and $\hat{A}_{ROC2}$, caused by using the same sample of objects to estimate both AUCs. This quantity can be found in tabular form (Hanley and McNeil, 1983).

A.3 Logistic regression and model selection

A.3.1 Definition

Logistic regression (Hosmer and Lemeshow, 1989) describes the relationship between one or several independent or explanatory variables (or data point $x = (x_1,x_2,...,x_p)$) and a binary (i.e., can only take on two possible values: 0 or 1) outcome variable Y. This outcome variable Y has a binomial distribution where the probability $P(Y = 1|x)$ (conditional probability of $Y = 1$ given the explanatory variables) is represented by $y(x)$. In a standard logistic regression model, $y(x)$ is written or modelled in a specific form:

$$y(x) = \frac{e^{g(x)}}{1+e^{g(x)}}, \tag{A.8}$$

where $g(x)$ is called the logit and is given by:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \tag{A.9}$$

where $\beta = (\beta_0,\beta_1,...,\beta_p)$ are the parameters or coefficients of the logistic regression model.

If some of the independent variables are discrete, nominal scaled variables (e.g., colour), these variables cannot be included in the logistic regression model as such. This situation requires the definition of design variables. In general, if the nominal scaled variable has k possible values, then k-1 design variables are needed. For example, suppose one considers a variable that represents a colour and that can take on three values: 'black', 'white' and 'grey'. In this case two design variables have to be introduced in the logistic regression model: D_1 and D_2. One possible coding strategy is then as follows: if the value for the nominal scaled variable is 'white', D_1 and D_2 are set equal to zero. If this value is 'black', D_1 is set to one and D_2 is set to zero. If this value is 'grey', D_1 is set to zero and D_2 is set to one.

A.3.2 Model fitting: maximum likelihood

Suppose we have a sample of N independent observation (training set) of the pair $\{x^i(x_1^i,x_2^i,...,x_p^i),Y^i\}_{i=1,...,N}$ and we want to estimate the parameters or coefficients $\beta = (\beta_0,\beta_1,...,\beta_p)$ of a logistic regression model that agrees most closely with the data. This is done by maximizing the likelihood function $l(\beta)$, which equals the probability of finding the observed data given the model parameters. This is given by:

$$l(\beta) = \prod_{i=1}^{N} y(x^i)^{Y^i} \left[1 - y(x^i)\right]^{1-Y^i}. \tag{A.10}$$

Mathematically it is easier (but equivalent) to maximize the logarithm of the likelihood function $\ln[l(\beta)]$. This is called the log likelihood $L(\beta)$:

$$L(\beta) = \sum_{i=1}^{N} \left\{Y^i \ln\left[y(x^i)\right] + (1 - Y^i) \ln\left[1 - y(x^i)\right]\right\}. \tag{A.11}$$

The values for the parameters that maximize this (log) likelihood function are called the maximum likelihood estimates of these parameters. To find these maximum likelihood estimates we have to differentiate $L(\beta)$ with respect to $\beta = (\beta_0, \beta_1, ..., \beta_p)$. However the resulting equations are non-linear in the parameters and therefore numerical and iterative methods built into logistic regression software have to be used.

The coefficients in a logistic regression model can be interpreted as the log of the odds ratio of the outcome for a unit increase of the associated variable.

A.3.3 Significance of an individual coefficient

Several hypothesis tests are available to test whether the maximum likelihood estimate of an individual coefficient differs significantly from zero (i.e., to test whether the true value of this coefficient is zero or whether the associated variable is significantly related to the outcome). We will mention two of these tests here: the likelihood ratio test and the Wald test.

Likelihood ratio test

Suppose we want to test whether the true value of a coefficient β_j is zero. Under the null hypothesis that β_j is zero, the statistic G given by

$$G = -2 \ln\left[\frac{\text{likelihood of fitted model without } x_j}{\text{likelihood of fitted model with } x_j}\right], \tag{A.12}$$

will follow a chi-square distribution with one degree of freedom. The likelihood of a fitted model can be found by evaluating Equation A.10 using the maximum likelihood estimates of the coefficients of the model at hand.

A specific sample of observations will result in a specific value for *G* and using the chi-square distribution, the probability can be calculated that an equally large or larger value for *G* will be obtained under the null hypothesis. If this probability or p-value is smaller than a certain rejection level (e.g., 5%), the null hypothesis is rejected and the true value of β_j is declared to be different from zero.

Wald test

In the Wald test it is assumed that under the null hypothesis, the ratio of the maximum estimate of the coefficient of a certain variable and its standard error will follow a standard normal distribution (or equivalently, it is assumed that the square of this ratio follows a chi-square distribution with one degree of freedom). Again, this can be used to calculate a p-value for a specific value for the maximum likelihood estimate of a certain coefficient.

A.3.4 Model selection

As described in Chapter 1, Section 1.3.3, in model selection we aim to select the most parsimonious set of variables from a group of considered variables that, when combined in a model, adequately explains the data. In this context, the model in which the variables are combined is a logistic regression model. In this section we will explain three possible strategies to perform model selection in logistic regression: forward, backward and stepwise selection.

Forward selection

In forward selection, the following procedure is applied based on a sample of observations and a group of variables that is considered for inclusion in the model:

5. Begin with a logistic regression model with only the intercept (constant term) and that does not include any variables.
6. Choose a significance level p_E for entry into the model (e.g., p_E = 0.15 - not too stringent).
7. For each variable that has not been included into the model: analyse a separate logistic regression model using this variable and the variables that already have been included in the model. Calculate the significance level or p-value (e.g., with the likelihood ratio or Wald test) for the coefficient of the variable in this model.

8. Select the variable associated with the smallest p-value. If this p-value is smaller than p_E: include the variable in the model and, if there remain variables that have not been included, return to step 3. Stop if this p-value is equal or larger than p_E or if all variables have been included.

Backward selection

In backward selection, the following procedure is applied:

1. Begin with a fitted logistic regression model where all the variables that are considered for inclusion are effectively included and calculate the significance level of the coefficient of each variable.
2. Choose a significance level p_R for removal out of the model (e.g., $p_R = 0.20$ - again, not too stringent).
3. Select the variable associated with the largest p-value. If this p-value is larger than p_R: remove the variable from the model. Stop if this p-value is equal or smaller than p_R.
4. Fit a logistic regression model with the variables that remain included in the model and calculate the significance level of the coefficient of each variable. Return to step 3. Stop if no variables remain included.

Stepwise selection

Stepwise selection is a combination of forward and backward selection. The basic scheme is the same as for forward selection with the following modification: after each inclusion of a variable, a logistic regression model is fitted using all the variables presently included, the significance level of the coefficient of each variable is calculated and the variable associated with the largest p-value is removed if this p-value is larger than p_R. This means that each forward selection step can be followed by a backward selection step. The algorithm stops if no variables can be included or removed.

From the previous it follows that a stepwise logistic regression procedure requires the specification of a value for p_R and p_E. As mentioned, 0.15 and 0.20, respectively, are reasonable choices, which are not too stringent and prevent that important variables would not be included in the model. The stepwise logistic regression therefore results in a logistic regression model that includes or contains variables that are important relative to the criteria p_R and p_E. If p_R and p_E do not correspond to our belief for statistical significance (usually fixed at a lower level of 5%), these may not be the variables reported in the final model and further selection is necessary. The methodology that may be used to achieve this will not be discussed further here.

A.4 Least Squares Support Vector Machines

Least Squares Support Vector Machine (LS-SVM) classifiers are a modified version of Support Vector Machines that can be used for binary classification (Suykens et al., 2002). Given a training set $\{x^i(x_1^i,x_2^i,...,x_p^i),y^i\}_{i=1,...,N}$ with input data x^i and corresponding class labels $y^i \in \{-1,+1\}$. The LS-SVM classifier takes the following (primal) form:

$$y(x) = \text{sign}\left[w^T \varphi(x) + b\right], \tag{A.13}$$

where the input data (that is said to belong to the input space) is mapped to a high dimensional feature space (which can be infinite dimensional) by a mapping function $\varphi(x)$ (that does not have to be specified – see further). This means that, conceptually, the classification is done in a high dimensional feature space where w is an element.

To determine the parameters of this model, the following optimisation problem has to be solved:

$$\min_{w,b,e}\left[\frac{1}{2}w^T w + \gamma\frac{1}{2}\sum_{i=1}^{N}\left(e^i\right)^2\right], \tag{A.14}$$

subject to:

$$y^i\left[w^T \varphi(x^i) + b\right] = 1 - e^i, i = 1,..., N. \tag{A.15}$$

The first term in Equation A.14 is called the regularization term and is representative of the model complexity and the second term in Equation A.14 is representative for the training set error. The optimisation problem therefore seeks a balance between minimizing model complexity and minimizing training set error. The process of limiting model complexity is called regularization and is necessary to prevent overfitting and enhance the generalization performance of this model. Note that γ is a hyperparameter of the model, which is called the regularization parameter. Also note that γ determines the degree of balance that has to be reached between model complexity and training set error. If γ is chosen to be infinite, no regularization is performed and only the training set error is minimized.

The optimisation problem is dealt with by solving the Lagrangian, which results in:

$$w = \sum_{i=1}^{N} \alpha^i y^i \varphi(x^i) \tag{A.16}$$

and the following dual problem to be solved in α and b:

$$\begin{bmatrix} 0 & y^T \\ y & \Omega + I/\gamma \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1_N \end{bmatrix}, \tag{A.17}$$

where

$$\begin{aligned} y &= \left[y^1 y^2 ... y^N\right]^T, \\ 1_N &= [1 1 ... 1]^T, \\ \alpha &= \left[\alpha^1 \alpha^2 ... \alpha^N\right]^T, \end{aligned} \tag{A.18}$$

and where the elements of the matrix Ω are given by:

$$\Omega_{kl} = y^k y^l \varphi(x^k)^T \varphi(x^l). \tag{A.19}$$

As said previously, the mapping function $\varphi(x)$ does not have to be constructed explicitly. We only have to specify the inner product in the feature space. This is represented by a (symmetric) kernel function that satisfies the Mercer condition:

$$K(r,s) = \varphi(r)^T \varphi(s), \tag{A.20}$$

where r and s belong to the input space. In this thesis we consider two possible kernel functions: a linear kernel given by:

$$K(r,s) = r^T s, \tag{A.21}$$

and a Radial Basis Function (RBF) kernel given by:

$$K(r,s) = \exp\left(-\frac{\|r-s\|_2^2}{\sigma^2} \right), \tag{A.22}$$

where σ again is a hyperparameter. Applying this kernel trick to Equation A.19 results in

$$\Omega_{kl} = y^k y^l K(x^k, x^l). \tag{A.23}$$

After applying Equation A.21 or A.22 to Equation A.23, all the coefficients of the set of equations given in Equation A.17 are known.

Finally, substitution of Equation A.16 in Equation A.13 results in the LS-SVM classifier in the dual form:

$$\begin{aligned} y(x) &= \operatorname{sign}\left[\sum_{i=1}^{N} \alpha^i y^i \varphi(x^i)\varphi(x) + b\right] \\ &= \operatorname{sign}\left[\sum_{i=1}^{N} \alpha^i y^i K(x, x^i) + b\right]. \end{aligned} \tag{A.24}$$

In the dual form, the model parameters are α and b that have been determined in Equation A.17. Equation A.17 is a set of N+1 linear equations in N+1 unknowns (α^1, α^2,...,α^N and b). Its size is therefore not determined by the number of dimensions in the input space but by the number of objects in the training set. Also not that using a linear kernel results in a linear classifier and using an RBF kernel results in a non-linear classifier.

Finally and in order to optimise model performance, the hyperparameter(s) γ (and possibly σ if an RBF kernel is chosen) have to be determined by a procedure that optimises the leave-one-out cross-validation performance on the training set.

A.5 K-means clustering

K-means (Tou and Gonzalez, 1979 - also see Table A.2 for the basic steps of this algorithm) is a cluster algorithm in which the user has to define in advance the number of clusters K that the algorithm will retrieve. The algorithm starts by assigning all the data points $\{x^i\}_{i=1,\ldots,N}$ to one of the K clusters at random (or pseudo-random like in Table A.2). Iteratively, the center (which corresponds to the average vector) of each cluster is calculated, followed by a re-assignment of the data points to the cluster with the closest (according to the Euclidean distance - the use of other distance measures is also possible but not discussed here) cluster center. Convergence is reached when the clusters do not further change. The result of the algorithm is dependent on K and the initial assignment of the data points to the K clusters. A form of (unsupervised) feature extraction has to be performed in advance if one wants to cluster high dimensional data (i.e., microarray experiments) using the K-means algorithm.

Table A.2: K-means algorithm.

```
K-means ($\{x^i\}_{i=1,...,N}$, $K$)
$C_1 = C_2 = ... = C_K = \varnothing$
FOR $i = 1,...,N$
        $C_{i - K.(\text{CEIL}(i/K) - 1)} = C_{i - K.(\text{CEIL}(i/K) - 1)} \cup \{x^i\}$
                /* assign $x^1$ to $C_1$, $x^2$ to $C_2$, ... , $x^K$ to $C_K$,
                   $x^{K+1}$ to $C_1$, $x^{K+2}$ to $C_2$, ... , $x^{2K}$ to $C_K$, ... */
END FOR
REPEAT
        FOR $i = 1,...,K$
                $\mu_i = \text{mean}(C_i)$      /* (Re)calculate average cluster vector */
        END FOR
        $C_1^{bef} = C_1$; $C_2^{bef} = C_2$; ... ; $C_K^{bef} = C_K$
        $C_1 = C_2 = ... = C_K = \varnothing$
        FOR $i = 1,...,N$
                $C_k = C_k \cup \{x^i\}$ if $\|x^i - \mu_k\|_2 < \|x^i - \mu_j\|_2 \quad \forall j \neq k$
                /* Re-assign each data point to the cluster with the nearest
                average vector */
        END FOR
UNTIL $C_1 = C_1^{bef}$ AND $C_2 = C_2^{bef}$ AND ... AND $C_K = C_K^{bef}$
/* Convergence if clusters have not changed */
OUTPUT $C_1$, $C_2$,..., $C_K$
```

A.6 Hierarchical clustering

Another method to cluster data points is hierarchical clustering. The results of this method can be visualized in a tree structure. Two approaches are possible: a top down approach (divisive clustering - see Alon et al. (1999) for an example) and a bottom-up approach (agglomerative clustering - see Eisen et al. (1998)). The latter is the most commonly used method and is discussed and used in this thesis. In the agglomerative approach each data point is initially assigned to a single cluster. Iteratively, the distance between every couple of clusters is determined and the two clusters that are closest are merged. This approach gives rise to the tree structure where the height of the branches is proportional to the pairwise distance between the clusters. Merging stops if only one cluster is left. Finally, clusters are formed by cutting the tree at a certain level or height. Different types of agglomerative clustering are possible, dependent on the definition of the distance between clusters:

1. Single linkage clustering: in this case the distance between two clusters is defined as the minimum of all pairwise distances between two data points (again according to a certain distance measure; e.g., correlation coefficient, Euclidean distance) belonging to the different clusters.
2. Complete linkage clustering: here the distance between two clusters is defined as the maximum of all pairwise distances between members of the different clusters.
3. Average linkage clustering: In this type of hierarchical clustering the distance between two clusters is defined as the mean of all pairwise distances between two vectors of the different clusters.
4. Centroid linkage clustering: in this case the distance between two clusters is defined as the distance between their centroids (average of the data points).

Feature reduction methods are not mandatory prior to the analysis of high-dimensional data with hierarchical clustering.

Appendix B
Data sets

In this appendix we will list and give an overview of the characteristics of most data sets that were used in this dissertation.

B.1 Clinical data

B.1.1 Endometrial cancer

This data set, kindly provided to us by Prof. Dr. D. Timmerman from the department of Obstetrics and Gynaecology (University Hospitals Leuven), contains patients diagnosed with endometrial cancer and typifies clinical data. Each patient is associated with a set of variables obtained after ultrasound and histopathological examination. The patients are divided into two classes dependent on the degree of invasion (with or without deep invasion) into the surrounding myometrium, which is an important prognostic parameter that has to be determined during staging. The training set contains 97 and the test set 37 patients. For each patient the subjective assessment of the degree of invasion by our expert ultrasonographer is available, which can be used as a reference. This data is analysed and discussed in further detail in Chapter 2 and serves as a typical example of clinical data analysis there.

B.2 Microarray or expression data

See Table B.1 for an overview of the URLs where the different data sets can be downloaded, if available.

B.2.1 Acute leukemia (1)

Golub et al. (1999) studied microarray data obtained from bone marrow or peripheral blood of 72 patients with acute lymphoblastic (ALL) or myeloid leukemia (AML) using an Affymetrix chip. Although the structure of this data set is simple and the separation between the two conditions is more pronounced than in most other cases, it can still be considered as a benchmark (paper cited over 1203 times) and serves as an illustration on several occasions in this text. In the original publication, the patients are divided into two sets: a fixed training set with 38 patients (27 ALL and 11 AML) and a fixed test set with 34 patients (20 ALL and 14 AML). The expression matrix contains 7129 genes or rows.

B.2.2 Acute leukemia (2)

Armstrong et al. (2002) also produced several microarray experiments obtained from patients with ALL or AML and from a third class or condition (called MLL leukemia) containing acute lymphoblastic leukemias with a chromosomal translocation involving the mixed-lineage leukemia gene. Armstrong et al. discovered that MLL leukemias have a distinct expression pattern and can be considered as a separate disease distinguishable from ALL and AML. It contains expression profiles for 12582 genes measured using Affymetrix technology. In total, 24 ALL patients, 28 AML patients and 20 MLL patients are available. This resulted in a data set containing 72 patients.

B.2.3 Breast cancer: degree of differentiation (1)

Perou et al. (2000) analysed surgical specimens of human breast tumours using cDNA-microarray technology with a common reference sample. Their study contained, among others, 37 tumours that were moderately or poorly differentiated (grade 2 or 3 - the degree of differentiation is assessed by the pathologist and reflects the degree of anaplasia or the degree of malignancy of the tumour and is an important prognostic factor). Twenty of these tumours were sampled twice (before and after a 16-week course of doxorubicin chemotherapy or paired with a lymph node metastasis) resulting in 57 microarray experiments (21 with grade 2 and 36 with grade 3). The raw data for each experiment (9216 genes) and the associated grade are available for downloading.

B.2.4 Breast cancer: degree of differentiation (2) and prognosis

van ‘t Veer et al. (2002) studied primary breast tumours with cDNA-microarrays (from sporadic lymph node negative patients and from patients with *BRCA1* or *BRCA2* germline mutations). In total 117 patients were analysed (24481 gene expression profiles are present in the data). The data included 51 patients with sporadic breast cancer (we also included the patients from the independent set) that did and 46 patients that did not develop distant metastases within five years. With respect to the degree of differentiation, 27 patients had a tumour with grade 2 and 78 had a tumour with grade 3. The data that is available can be downloaded under the form of log-ratios (and was already appropriately preprocessed, which was not the case for the data from Perou et al.).

B.2.5 Breast cancer: sporadic versus hereditary

Hedenfalk et al. (2001) studied sporadic breast tumours (7 patients), breast tumours carrying a *BRCA1* mutation (7 patients) and breast tumours carrying a *BRCA2* mutation (7 patients) using a cDNA-microarray. Three binary classification problems follow from this study (one class versus the rest). The authors selected 3226 genes for there analyses according to a set of prespecified criteria.

B.2.6 Colon cancer

Alon et al. (1999) studied 40 tumour and 22 normal colon tissue samples using an Affymetrix chip. The array contained probes for more than 6500 genes but the data that can be downloaded includes only the 2000 genes with highest minimal intensity across the 62 tissues.

B.2.7 Hepatocellular carcinoma

Using an Affymetrix chip, Iizuka et al. (2003) studied hepatocellular carcinomas with and without early intrahepatic recurrence after surgery for hepatic resection. Their data contained 60 patients originally divided in a training set of 33 and a test set of 27. In total, 20 patients had an early recurrence of their disease. The data contained 7129 gene expression profiles.

B.2.8 High-grade gliomas

Nutt et al. (2003) analysed the expression patterns of 50 patients with high-grade gliomas with an Affymetrix chip (12625 genes). They compared two histopathological subclasses: glioblastomas (poor prognosis - 28 patients) and anaplastic oligodendrogliomas (more favourable prognosis - 22 patients). The training set contained 21 patients (14 glioblastomas and 7 anaplastic oligodendrogliomas).

B.2.9 Prostate cancer

Singh et al. (2002) studied, among others, expression patterns of normal and malignant prostate samples using oligonucleotide arrays with probes for 12600 genes. A training set with 102 patients (52 prostate tumours and 52 normal samples) and a test set with 34 patients (25 prostate tumours and 9 normal samples) are available for downloading.

B.2.10 Yeast cell cycle

Cho et al. (1998) studied the yeast cell cycle in a synchronised culture on an Affymetrix chip (also see Spellman et al. (1998)). This data set contains expression profiles for 6220 genes over 17 time points taken at 10-min intervals, covering nearly two full cell cycles. Although, this data does not originate directly from oncology, it is related because dysregulation of the cell cycle plays an important role in carcinogenesis. Moreover, we chose to include this data because it studies microarray experiments taken at different time points of a biological process rather than microarray experiments belonging to different classes and is especially suited to examine cluster analysis of gene expression profiles, a topic that is investigated thoroughly in this thesis. Moreover in this context it can be considered as a benchmark (De Smet et al., 2002; Jakt et al., 2001; Yeung et al., 2001b; Heyer et al., 1999; Tamayo et al., 1999; Tavazoie et al., 1999).

B.2.11 Central nervous system development

Wen et al. (1998) studied gene expression levels of 112 genes on 9 time points during central nervous system development of the rat, using tissue of the cervical spinal cord using reverse transcription-coupled PCR (RT-PCR). Unfortunately, the website where the data was downloaded from, is no longer available.

B.2.12 Measurement of expression levels in different tissues

Seven two-channel cDNA microarray-experiments (obtained from Dr. P. Van Hummelen of the Microarray Facility of the V.I.B. - data not publicly available) were performed to characterise 4595 expression patterns in 7 mouse tissues. A common reference pool was used for each of the experiments (green channel). The red channel corresponds to a RNA pool obtained from one of 7 tissues:

- Experiment 1: Brain
- Experiment 2: Heart
- Experiment 3: Kidney
- Experiment 4: Liver
- Experiment 5: Lung
- Experiment 6: Skeletal muscle
- Experiment 7: Testis

The intention of this experimental setup was to detect groups of tissue-specific genes (mostly upregulated genes in one or two tissues). The data set itself contains 4595 seven-dimensional expression vectors. A fraction of the expression ratios is missing (approximately 3.5%).

Table B.1: Overview of the URLs of the different microarray data sets

Authors	URL
Golub et al.	http://www-genome.wi.mit.edu/cancer/
Armstrong et al.	http://research.dfci.harvard.edu/korsmeyer/MLL.htm
Perou et al.	http://genome-www.stanford.edu/molecularportraits/
van ’t Veer et al.	http://www.rii.com/publications/default.htm (log ratios) http://www.nature.com (suppl. inform. - degree of diff.)
Alon et al.	http://microarray.princeton.edu/oncology/affydata/index.html
Iizuka et al.	http://surgery2.med.yamaguchi-u.ac.jp/research/DNAchip/hcc-recurrence/index.html
Nutt et al.	http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
Singh et al.	http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
Cho et al.	http://cellcycle-www.stanford.edu
Wen et al. (1998)	http://rsb.info.nih.gov/mol-physiol/PNAS/GEMtable.html (no longer available)

Bibliography

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and Staudt, L.M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.

Alon, U., Barkai. N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA*, **96**, 6745-6750.

Alter, O., Brown, P.O. and Botstein, D. (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci USA*, **100**, 3351-3356.

Antal, P., Fannes, G., De Smet, F. and De Moor, B. (2001) Ovarian cancer classification with rejection by Bayesian Belief Networks. In *Proc of the Bayesian Models in Medicine workshop, the European Conference on Artificial Intelligence in Medicine (AIME'01)*, pp. 23-27.

Arko, D. and Takac, I. (2000) High frequency transvaginal ultrasonography in preoperative assessment of myometrial invasion in endometrial cancer. *J Ultrasound Med*, **19**, 639-643.

Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R. and Korsmeyer, S.J. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, **30**, 41-47.

Ben-Dor, A., Shamir, R. and Yakhini, Z. (1999) Clustering gene expression patterns. *J Comput Biol*, **6**, 281-297.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B*, **57**, 289-300.

Bishop, C.M. (1995) *Neural Networks for pattern recognition.* Oxford University Press Inc., New York.

Broberg, P. (2003) Statistical methods for ranking differentially expressed genes. *Genome Biol*, **4**, R41.

Chang, J.C., Wooten, E.C., Tsimelzon, A., Hilsenbeck, S.G., Gutierrez, M.C., Elledge, R., Mohsin, S., Osborne, C.K., Chamness, G.C., Allred, D.C. and O'Connell, P. (2003) Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet*, **362**, 362-369.

Chapman, K. (2002) The ProteinChip Biomarker System from Ciphergen Biosystems: a novel proteomics platform for rapid biomarker discovery and validation. *Biochem Soc Trans*, **30**, 82-87.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. and Davis, R.W. (1998) A genome wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, **2**, 65-73.

Coessens, B., Thijs, G., Aerts, S., Mathys, J., Moreau, Y., Marchal, K., De Smet, F., Engelen, K., Glenisson P. and De Moor, B. (2003) INCLUSive—A Web Portal and Service Registry for Microarray and Regulatory Sequence Analysis. *Nucleic Acids Res*, **31**, 3468-3470.

Dawson-Saunders, B. and Trapp, R.G. (1994) *Basic & Clinical Biostatistics, 2nd edition*. Appleton & Lange, Connecticut.

De Moor, B., Marchal, K., Mathys, J., and Moreau, Y. (2003) Bioinformatics: Organisms from Venus, Technology from Jupiter, Algorithms from Mars. *European Journal of Control*, **9**, 237-278.

De Smet, F., Marchal, K., Timmerman, D., Vergote, I., De Moor, B. and Moreau, Y. (2001) Gebruik van microroosters in de klinische oncologie, *Tijdschr voor Geneeskunde*, **57**, 1225-1236.

De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B. and Moreau Y. (2002) Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, **18**, 735-746.

De Smet, F., Moreau, Y., Tmmerman, D., Vergote, I. and De Moor, B. (2004) Balancing false positives and false negatives for the detection of differential expression in malignancies. *Br J Cancer*, submitted.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J.M. (1999) Expression profiling using cDNA microarrays. *Nat Genet* , **21**, 10-14.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, **95**, 14863-14868.

Engelen, K., Coessens, B., Marchal, K. And De Moor, B. (2003) MARAN: a web-based application for normalizing micro-array data. *Bioinformatics*, **19**, 893-894.

Epstein, E., Skoog, L., Isberg, P.E., De Smet, F., De Moor, B., Olofsson, P.A., Gudmundsson, S. and Valentin, L. (2002) An algorithm including results of gray-scale and power Doppler ultrasound examination to predict endometrial malignancy in women with postmenopausal bleeding. *Ultrasound Obstet Gynecol*, **20**, 370-376.

Fraley, C. and Raftery, E. (1999) MCLUST: Software for Model-Based Cluster Analysis. *Journal of Classification*, **16**, 297-306.

Franchi, M., Ghezzi, F., Melpignano, M., Cherchi, P.L., Scarabelli, C., Apolloni, C. and Zanaboni, F. (2000) Clinical value of intraoperative gross examination in endometrial cancer. *Gynecol Oncol*, **76**, 357-361.

Friend, S.H. (1999) How DNA microarrays and expression profiling will affect clinical practice. *BMJ*, **319**, 1306-1307.

Gerhold, D.L., Jensen, R.V. and Gullans, S.R. (2002) Better therapeutics through microarrays. *Nat Genet*, **32 Suppl**, 547-551.

Ghosh, D. and Chinnaiyan, A.M. (2002) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18**, 275-286.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

Guden, M., Goktas, S., Sumer, F., Ulutin, C. and Pak, Y. (2003) Retrospective analysis of 74 cases of seminoma treated with radiotherapy. *Int J Urol*, **10**, 435-438.

Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29-36.

Hanley, J.A. and McNeil, B.J. (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, **148**, 839-843.

Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D. and Brown, P. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol*, **1**, research0003.1-0003.21.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P., Wilfond, B., Borg, A. and Trent, J. (2001) Gene-expression profiles in hereditary breast cancer. *N Engl J Med*, **344**, 539-548.

Herrero, J., Valencia, A. and Dopazo, J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126-136.

Heyer, L.J., Kruglyak, S. and Yooseph, S. (1999) Exploring expression data: Identification and analysis of coexpressed genes. *Genome Res*, **9**, 1106-1115.

Hosmer, D.W. and Lemeshow, S. (1989) *Applied Logistic regression*. John Wiley & Sons, New York.

Huang, E., Cheng, S.H., Dressman, H., Pittman, J., Tsou, M.H., Horng, C.F., Bild, A., Iversen, E.S., Liao, M., Chen, C.M., West, M., Nevins, J.R. and Huang, A.T. (2003) Gene expression predictors of breast cancer outcomes. *Lancet*, **361**, 1590-1596.

Huddart, R.A., Norman, A., Shahidi, M., Horwich, A., Coward, D., Nicholls, J. and Dearnaley, D.P. (2003) Cardiovascular disease as a long-term complication of treatment for testicular cancer. *J Clin Oncol*, **21**, 1513-1523.

Iizuka, N., Oka, M., Yamada-Okabe, H., Nishida, M., Maeda, Y., Mori, N., Takao, T., Tamesa, T., Tangoku, A., Tabuchi, H., Hamada, K., Nakayama, H., Ishitsuka, H., Miyamoto, T., Hirabayashi, A., Uchimura, S. and Hamamoto, Y. (2003) Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet*, **361**, 923-929.

Jakt, L.M., Cao, L., Cheah, K.S. and Smith, D.K. (2001) Assessing clusters and motifs from gene expression data. *Genome Res*, **11**, 112-123.

Jones, R.H. and Vasey, P.A. (2003) Part I: testicular cancer-management of early disease. *Lancet Oncol*, **4**, 730-737.

Kaufman, L. and Rousseeuw, P.J. (1990) *Finding groups in data: an introduction to cluster analysis*. Wiley, New York.

Kerr, M.K., Martin, M. and Churchill, G.A. (2000) Analysis of variance for gene expression microarray data. *J Comput Biol*, **7**, 819-837.

Keselman, H.J., Cribbie, R. and Holland, B. (2002) Controlling the rate of Type I error over a large set of statistical tests. *Br J Math Stat Psychol*, **55**, 27-39.

Kihara, C., Tsunoda, T., Tanaka, T., Yamana, H., Furukawa, Y., Ono, K., Kitahara, O., Zembutsu, H., Yanagawa, R., Hirata, K., Takagi, T. and Nakamura, Y. (2001) Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by cDNA microarray analysis of gene-expression profiles. *Cancer Res*, **61**, 6474-6479.

Kinkel, K., Kaji, Y., Yu, K.K., Segal, M.R., Lu, Y., Powell, C.B. and Hricak, H. (1999) Radiologic staging in patients with endometrial cancer: a meta-analysis. *Radiology*, **212,** 711-718.

Kohonen, T. (1997) *Self-organizing maps.* Springer-Verlag, Berlin.

Kozak, K.R., Amneus, M.W., Pusey, S.M., Su, F., Luong, M.N., Luong, S.A., Reddy, S.T. and Farias-Eisner, R. (2003) Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: potential use in diagnosis and prognosis. *Proc Natl Acad Sci USA*, **100**, 12343-12348.

Kucera, E., Kainz, C., Reinthaller, A., Sliutz, G., Leodolter, S., Kucera, H. and Breitenecker G. (2000) Accuracy of intraoperative frozen-section diagnosis in stage I endometrial adenocarcinoma. *Gynecol Obstet Invest*, **49**, 62-66.

Levenstien, M.A., Yang, Y. and Ott, J. (2003) Statistical significance for hierarchical clustering in genetic association and microarray expression studies. *BMC Bioinformatics*, **4**, 62.

Levine, D.A. and Hoskins, W.J. (2002) Update in the management of endometrial cancer. *Cancer J*, **8 Suppl 1**, S31-40.

Lewin, B. (1997) *Genes – 6th edition*. Oxford University Press Inc., New York.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. and Lockhart, D.J. (1999) High density synthetic oligonucleotide arrays. *Nat Genet* , **21**, 20-24.

Longo, D.L. (1998) Approach to the patient with cancer. In Fauci, A.C., Braunwald, E., Isselbacher, K.J., Wilson, J.D., Martin, J.B., Kasper, D.L., Hauser, S.L. and Longo, D.L. (eds), *Harrison's principles of internal medicine - 14th edition*. McGraw-Hill, New York, pp. 493-499.

Ludwig, H. (1995) Prognostic factors in endometrial cancer. *Int J Gynaecol Obstet*, **49 Suppl**, S1-7.

Lukashin, A.V. and Fuchs, R. (2001) Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, **17**, 405-414.

Marchal, K., De Smet, F., Engelen, K. and De Moor, B. (2004) Computational biology and toxicogenomics. In Helma, C. (ed), *Predictive Toxicology*. Marcel Dekker, In press.

Mewes, H.W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., Stocker, S. and Weil,

B. (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, **28**, 37-40.

Moreau, Y., De Smet, F., Thijs, G., Marchal, K. and De Moor, B. (2002a) Functional bioinformatics of microarray data: from expression to regulation. *Proceedings of the IEEE*, **90**, 1722-1743.

Moreau, Y., Marchal, K. and Mathys, J. (2002b) *Computational biomedicine: a multidisciplinary crossroads*. Siemens Prize, FWO (Flanders, Belgium).

Motzer, R.J. and Bosl, G.J. (1998) Testicular cancer. In Fauci, A.C., Braunwald, E., Isselbacher, K.J., Wilson, J.D., Martin, J.B., Kasper, D.L., Hauser, S.L. and Longo, D.L. (eds), *Harrison's principles of internal medicine - 14th edition*. McGraw-Hill, New York, pp. 602-605.

Nielsen, T.O., West, R.B., Linn, S.C., Alter, O., Knowling, M.A., O'Connell, J.X., Zhu, S., Fero, M., Sherlock, G., Pollack, J.R., Brown, P.O., Botstein, D and, van de Rijn, M. (2002) Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet*, **359**, 1301-1307.

Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Cairncross, J.G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M.E., Batchelor, T.T., Black, P.M., von Deimling, A., Pomeroy, S.L., Golub, T.R. and Louis, D.N. (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*, **63**, 1602-1607.

Pagano, M. and Gauvreau, K. (2000) *Principles of Biostatistics, 2nd edition*. Duxbury Press.

Perneger, T.V. (1998) What's wrong with Bonferroni adjustments. *BMJ*, **316**, 1236-1238.

Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lonning, P.E., Borresen-Dale, A.L., Brown, P.O. and Botstein, D. (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747-752.

Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C. and Liotta, L.A. (2002a) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, **359**, 572-577.

Petricoin, E.F., Ornstein, D.K., Paweletz, C.P., Ardekani, A., Hackett, P.S., Hitt, B.A., Velassco, A., Trucco, C., Wiegand, L., Wood, K., Simone, C.B., Levine, P.J., Linehan, W.M., Emmert-Buck, M.R., Steinberg, S.M., Kohn, E.C. and Liotta, L.A. (2002b) Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst*, **94**, 1576-1578.

Pochet, N., De Smet, F., Suykens, J. and De Moor, B. (2004) Systematic benchmarking micorarray data classification: assessing the role of nonlinearity and dimensionality reduction. *Bioinformatics*, submitted.

Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S. and Golub, T.R. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436-442.

Quackenbush J. (2001) Computational analysis of microarray data. *Nat Rev Genet*, **2**, 418-427.

Ramaswamy, S., Ross, K.N., Lander, E.S. and Golub, T.R. (2003) A molecular signature of metastasis in primary solid tumors. *Nat Genet*, **33**, 49-54.

Reiner, A., Yekutieli, D. and Benjamini, Y. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368-375.

Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D. and Chinnaiyan, A.M. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res*, **62**, 4427-4433.

Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Muller-Hermelink, H.K., Smeland, E.B., Giltnane, J.M., Hurt, E.M., Zhao, H., Averett, L., Yang, L., Wilson, W.H., Jaffe, E.S., Simon, R., Klausner, R.D., Powell, J., Duffey, P.L., Longo, D.L., Greiner, T.C., Weisenburger, D.D., Sanger, W.G., Dave, B.J., Lynch, J.C., Vose, J., Armitage, J.O., Montserrat, E., Lopez-Guillermo, A., Grogan, T.M., Miller, T.P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T. and Staudt, L.M. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*, **346**, 1937-1947.

Roth, F.P., Hughes, J.D., Estep, P.W. and Church, G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole genome mRNA quantitation. *Nat Biotechnol*, **16**, 939-945.

Sager, R. (1997) Expression genetics in cancer: shifting the focus from DNA to RNA. *Proc Natl Acad Sci USA*, **94**, 952-955.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann Stat*, **6**, 461-464.

Sherlock, G. (2000) Analysis of large-scale gene expression data. *Curr Opin Immunol*, **12,** 201-205.

Shridhar, V., Lee, J., Pandita, A., Iturria, S., Avula, R., Staub, J., Morrissey, M., Calhoun, E., Sen, A., Kalli, K., Keeney, G., Roche, P., Cliby, W., Lu, K., Schmandt, R., Mills, GB., Bast, R.C. Jr, James, C.D., Couch, F.J., Hartmann, L.C., Lillie, J. and Smith, D.I. (2001) Genetic analysis of early-versus late-stage ovarian tumors. *Cancer Res*, **61**, 5895-5904.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R. and Sellers, W.R. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203-209.

Slapak, C.A. and Kufe, D.W. (1998) Principles of cancer therapy. In Fauci, A.C., Braunwald, E., Isselbacher, K.J., Wilson, J.D., Martin, J.B., Kasper, D.L., Hauser, S.L. and Longo, D.L. (eds), *Harrison's principles of internal medicine - 14th edition*. McGraw-Hill, New York, pp. 523-537.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, **9**, 3273-3297.

Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*, **100**, 9440 - 9445.

Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B. and Vandewalle, J. (2002) *Least Squares Support Vector Machines*. World Scientific, Singapore.

Swets, J.A. (1996) *Signal detection theory and ROC analysis in psychology and diagnostics, collected papers.* Lawrence Erlbaum Associates, New Jersey.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T. R. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA*, **96**, 2907-2912.

Tapper, J., Kettunen, E., El-Rifai, W., Seppala, M., Andersson, L.C. and Knuutila, S. (2001) Changes in gene expression during progression of ovarian carcinoma. *Cancer Genet Cytogenet*, **128**, 1-6.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) Systematic determination of genetic network architecture. *Nat Genet*, **22**, 281-285.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouzé, P. and Moreau, Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113-1122.

Thijs, G., Moreau, Y., De Smet, F., Mathys, J., Lescot, M., Rombauts, S., Rouze, P., De Moor, B. and Marchal, K. (2002a) INCLUSive: INtegrated Clustering, Upstream sequence retrieval and motif Sampling. *Bioinformatics*, **18**, 331-332.

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P. and Moreau Y. (2002b) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol*, **9**, 447-464.

Thijs, G., De Smet, F., Moreau, Y., Marchal, K. and De Moor, B. (2004) Gene regulation bioinformatics of microarray data. In Akay, M. (ed), *Genomics and Proteomics Engineering*. Wiley/IEEE press, Accepted.

Timmerman, D. (1997) *Ultrasonography in the assessment of ovarian and tamoxifen-associated endometrial pathology*. PhD thesis, Leuven University Press, Leuven.

Timmerman, D., De Smet, F., De Brabanter, J., Van Holsbeke, C., Jermy, K., Moreau, Y., Bourne, T. and Vergote I. (2003) OC118 : Mathematical models to evaluate ovarian masses - can they beat an expert operator ? *Ultrasound Obstet Gynecol*, **22(S1)**, 33.

Tou, J.T. and Gonzalez, R.C. (1979) Pattern classification by distance functions. In Tou, J.T. and Gonzalez, R.C. (eds), *Pattern recognition principles*. Addison-Wesley, Reading (Mass.), pp. 75-109.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, RB. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520-525.

Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D. and Altman, R.B. (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454-1461.

Tumor Analysis Best Practices Working Group (2004) Expression profiling - best practices for data generation and interpretation in clinical trials. *Nat Rev Genet*, **5**, 229-237.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, **98**, 5116-5121.

Van den Enden, E. (2001) *Clustering van Microrooster Gegevens: Evaluatie van de 'Gene Shaving' methode.* Master thesis, K.U.Leuven., Faculty of Applied Sciences, Department of Electrical Engineering.

van de Vijver, M.J., He, Y.D., van 't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H. and Bernards, R. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, **347**, 1999-2009.

van Helden, J., Rios, A.F. and Collado-Vides, J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucl. Acids Res*, **28**, 1808-1818.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530-536.

Weber, G., Merz, E., Bahlmann, F., Mitze, M., Weikel, W. and Knapstein, P.G. (1995) Assessment of myometrial infiltration and preoperative staging by transvaginal ultrasound in patients with endometrial carcinoma. *Ultrasound Obstet Gynecol*, **6,** 362-367.

Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L. and Somogyi, R. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci USA*, **95**, 334-339.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. and Speed, T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, **30**, e15.

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. and Ruzzo, W.L. (2001a) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977-987.

Yeung, K.Y., Haynor, D.R. and Ruzzo, W.L. (2001b) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309-318.

Yeung, K.Y. and Ruzzo, W.L. (2001c) Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**, 763-774.

Youngh, R.C. (1998) Gynecologic malignancies. In Fauci, A.C., Braunwald, E., Isselbacher, K.J., Wilson, J.D., Martin, J.B., Kasper, D.L., Hauser, S.L. and Longo, D.L. (eds), *Harrison's principles of internal medicine - 14th edition*. McGraw-Hill, New York, pp. 605-611.

Zagars, G.K., Ballo, M.T., Lee, A.K. and Strom, S.S. (2004) Mortality after cure of testicular seminoma. *J Clin Oncol*, **22**, 640-647.

Curriculum Vitae

Personal Data

Name:	Frank De Smet
Born:	03/08/1969, Bonheiden, Belgium
Married to:	Ilse Lamberts
Father of:	Lieselot (°21/12/2001) and Stijn (°25/09/2003)
Address:	Zemstbaan 156 2800 Mechelen
E-mail:	frank.desmet@esat.kuleuven.ac.be frank_desmet@pandora.be

Education

1992-1998: Medical Doctor - K.U.Leuven
Magna cum laude
Additional degree in electrocardiography

1987-1992: Master of Science in Electrical Engineering - K.U.Leuven
Specialization in microelectronics
Cum Laude
Master thesis: Implementation of the Fermi-Dirac statistic in the device simulator PRISM.

Professional Experience

12/10/1998-15/08/1999: Software Engineer at COMPEX N.V. (Belgian company specialized in the development of

Laboratory Information Management Systems) - see http://www.compex.be

Research Experience

15/08/1999-present: Research assistant of the K.U.Leuven at the bioinformatics group of ESAT-SCD(SISTA), (see http://www.esat.kuleuven.ac.be/~dna/BioI) under the supervision of Prof. dr. ir. Bart De Moor.