

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

INTEGRATING SCIENTIFIC LITERATURE WITH LARGE SCALE GENE EXPRESSION ANALYSIS

Promotor:

Prof. dr. ir. B. De Moor

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

door

Patrick GLENISSON

June 2004



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

INTEGRATING SCIENTIFIC LITERATURE WITH LARGE SCALE GENE EXPRESSION ANALYSIS

Jury:

Prof. dr. ir. Y. Willems, voorzitter
Prof. dr. ir. B. De Moor, promotor
Prof. dr. ir. S. Van Huffel
Prof. dr. ir. P. Marynen
Prof. dr. ir. J. Suykens
Prof. dr. ir. Y. Moreau
Prof. dr. ir. H. Blockeel

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

door

Patrick GLENISSON

U.D.C. 575.117.2

June 2004

© Katholieke Universiteit Leuven – Faculteit Toegepaste Wetenschappen
Arenbergkasteel, B-3001 Heverlee (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

D/2004/7515/53

ISBN 90-5682-515-1

Voorwoord

In eerste instantie wil ik mijn promotor prof. Bart De Moor bedanken voor de talloze kansen en de financiering die hij me geboden heeft tijdens mijn doctoraat. Zijn bij mij al legendarische '*Go for it, Patje*' gaf altijd een opstoot van vertrouwen in de momenten dat de helling wat steiler werd.

Het is dankzij Prof. Yves Moreau dat ik de bioinformatica ben ingedoken; zijn stempel staat zeker gedrukt in dit doctoraat. Ik wil hem bedanken voor zijn ideeën en zijn steun, en ik ben dan ook blij dat hij deel uitmaakt van mijn jury.

Ik wens ook de overige leden van mijn begeleidingscomité en jury, Prof. Peter Marynen, Prof. Sabine Van Huffel, Prof. Johan Suykens, Prof. Hendrik Blockeel en Prof. Yves Willems te bedanken voor hun advies tijdens het doctoraat en de feedback op het proefschrift.

Zonder hulp van Dr. Janick Mathys hadden de ideeën in dit doctoraat minder impact gehad. Ik wens haar te bedanken om 'advocaat van de biologie' te spelen en voor het geduld en de moeite die ze heeft opgebracht voor de vele validaties. Ook Bert Coessens en Steven Van Vooren ben ik veel dank verschuldigd. Het was zéér fijn samenwerken.

I acknowledge Peter Antal for the fruitful collaborations. He laid the seeds of the presented textmining research and put me on the right track. Nagyon köszönöm !

De periode aan de MAF (de MicroArray Facility nvdr) zal ik ook niet gemakkelijk vergeten. Bedankt Paul Van Hummelen en de overige MAF'ers voor de leuke samenwerking.

Voorts wil ik ook Gert Thijs, Stein Aerts, Geert Fannes, Frank De Smet, Kathleen Marchal en Bart Hamers bedanken voor de interacties gedurende de laatste jaren. Ook de andere collega's van de 'Bio-groep' en SCD ben ik dank verschuldigd. Zij bepaalden immers mee de goede werksfeer die er op ESAT heerst.

Tot slot wens ik mijn ouders en zus te bedanken ; er is heel wat veranderd het laatste jaar, behalve hun onvoorwaardelijke steun en begrip. Merci ! And, of course, thanks Flori for your understanding, support, advice and much more. We have seen that *even for a PhD, one plus one can equal three...*

Abstract

The current tendency in the life sciences to produce ever growing amounts of high-throughput data has led to a situation where the interpretation of data and the formulation of hypotheses lag the pace at which information is produced.

Gene expression arrays or microarrays measure the activity of thousands of genes simultaneously, and are primary examples of post-genome technologies where the conversion of the bulk of gene-related data into information constitutes a major challenge. The research presented in this thesis addresses this problem and is made up of three components: (1) representation and storage of microarray information, (2) literature-based gene analysis, (3) integrated exploratory analysis of gene expression data with text.

Firstly, one of the early problems concerning gene expression experiments has been the absence of agreed, reproducible reporting schemes. To address this, we have contributed to the development of MIAME 1.1, an international standard for reporting about microarray experiments. In order to furnish this information from a laboratory environment, we discuss the development and implementation of a first-generation information management system (LIMS) at the VIB MicroArray Facility.

Secondly, we propose a vector-based text representation that compiles a term-based profile for each gene based on its associated literature. We show that such gene profiles are able to capture relevant functional information. To automatically assess the functional cohesion of a group of genes in the text world, we develop and test a text-based coherence score. We leverage this approach to a public text mining web service, called TXTGate. Through a use of multiple, complementary vocabularies, the application offers several *views* on gene literature and profiles groups of genes, typically stemming from gene expression cluster analysis.

Finally, we combine gene expression profiles and text profiles in a clustering framework based on statistical meta-analysis. As many parameter settings need to be evaluated in parallel, we propose a novel score that gives a one-shot estimate of the enrichment of known transcription factor binding sites in the meta-clustering solution. Through this independent validation, we demonstrate the overall benefit of using expression data and literature *jointly* in the clustering of gene expression data.

Nederlandstalige abstract

De huidige tendens in de biologische wetenschappen om steeds meer gegevens te produceren heeft tot een situatie geleid waar de interpretatie van gegevens sterk achterloopt op de productie ervan. Microroosters, waarbij het gedrag van soms duizenden genen simultaan wordt gemeten, vormen hier een goed voorbeeld van. Het onderzoek behandeld in dit eindwerk concentreert zich op drie punten: (1) representatie en opslag van microrooster gegevens, (2) tekst-gebaseerde gen analyze en (3) clustering van microrooster- and tekstuele gegevens.

Een van de eerste problemen die zich voordeden met microrooster gegevens was de afwezigheid van gestandaardiseerde rapportering. In dit kader hebben we bijgedragen tot de ontwikkeling van de internationale MIAME standaard. Aan praktische zijde hebben we een eerste versie van een Laboratorium Informatie Management Systeem (LIMS) ontworpen aan de VIB MicroArray Facility.

Ten tweede is een vector-gebaseerde tekst representatie opgesteld dat, voor elk gen, een profiel van termen opstelt en dit gebaseerd op de bijhorende wetenschappelijke literatuur. We hebben aangetoond dat deze voorstelling krachtig genoeg is om functionele informatie over genen te bevatten. Om de coherentie van groepen van genen automatisch te onderzoeken vanuit het standpunt van de literatuur, hebben we met succes een nieuwe tekst-gebaseerde score uitgewerkt en getest. Het voorgestelde raamwerk van tekstontginning is geïmplementeerd in TXTGate, een software component die toelaat om groepen van genen vanuit verschillende oogpunten te onderzoeken. Typische vragen die kunnen voorgelegd worden aan de applicatie zijn: 'Welke sleutelwoorden aangaande biologische functie of ziekte zijn karakteristiek voor een groep van genen ?' of 'Welke genen zijn in de literatuur mede gerapporteerd met een groep van genen ?'.

Tot slot, hebben we gen expressie data geïntegreerd met literatuur-gebaseerde informatie door gebruik te maken van een door ons ontworpen meta-clustering algoritme. Verschillende parameterisaties werden uitvoerig getest en door middel van een validatie in termen van transcriptionele regulatie hebben we aangetoond dat het gecombineerd gebruik van expressie- en literatuur-data een verbetering geeft ten opzichte van het geïsoleerd gebruik van de verschillende informatiebronnen.

Nederlandse samenvatting

Situering van het onderzoek

De vervolledigiging van het Menselijk Genoom[1] Project in 2001 door de tandem bestaande uit het International Human Genome Consortium en Celera Genomics [66][158] heeft een aanzienlijke impact gehad op de evolutie van de moleculaire biologie, genetica en geneeskunde. De manier waarop biologische vraagstukken worden geformuleerd is hierdoor fundamenteel veranderd. Nu ongeveer de meeste genen geïdentificeerd zijn in niet enkel mens, maar ook een aanzienlijk aantal andere organismen zoals muis, rat, gist,... [1] kan op zoek gegaan worden naar de functie van elk van die genen. Naast informatie omtrent het genoom zelf, laten recente biotechnologische ontwikkelingen toe om naar het gedrag, of samenspel, van genen en eiwitten[2] te kijken op een massief-parallelle manier. Waar er vroeger eerder reductionistisch werd tewerk gegaan (door te focussen op één enkel gen), kan nu het gedrag van duizenden genen simultaan geobserveerd worden. Dit zorgt echter voor een enorme hoeveelheid gegevens die tevens door een aantal intrinsieke moeilijkheden gekarakteriseerd worden:

1. Experimenteel biologische data is van hoge dimensie: soms wordt het simultaan gedrag van duizenden genen gemeten over meer dan een dozijn experimentele condities.

2. De data heeft meestal een lage signaal-ruis verhouding en vereist daarom robuuste analyse technieken.

3. Biologische data heeft meestal enkel betekenis in een bepaalde context en opereert op verschillende schalen: sequenties van DNA of eiwitten, transcriptionele processen, 3D eigenschappen van eiwitten en enzymes, chromosomale organisatie en fenotypes[3] zijn enerzijds onderling nauw verbonden, maar kunnen anderzijds, afhankelijk van de vraagstelling, apart benaderd worden door 'modulair' naar het probleem te kijken.

[1] Het genoom is de volledige genetische code van een organisme.

[2] Eiwitten worden geproduceerd met behulp van de informatie bevat in de genen, en spelen een centrale rol in de werking van een cel.

[3] Fenotypes zijn uitwendige, herkenbare kenmerken die zich manifesteren in het organisme.

4. Biologische kennis is verspreid over verschillende gegevensbanken over het Internet. Het inbrengen van zulke kennis tijdens de interpretatie van biologische data veronderstelt dat deze eenvoudig toegankelijk en hanteerbaar is. Een begrijpelijk gegevensformaat of een efficiente representatie is hierbij van belang.

Biologisch onderzoek wordt in groeiende mate gestuurd door reeds bestaande kennis die vervat zit in een massa aan databanken. De vervolledigingvan de sequentie van het menselijk genoom en het recente succes van grootschalige experimenten in de genomica en proteomica[4] dragen hier voor een groot stuk toe bij. Om alle gegevens en informatie die het resultaat zijn van veelal dure, grootschalige experimenten optimaler te kunnen benutten, is er een absolute nood aan intuïtieve en schaalbare technieken voor gegevensontginning en intelligent kennisbeheer. Enkel hierdoor kan de band tussen de kennis die vervat zit in

1. de expert,
2. de beschikbare kennis verspreid over honderden externe databanken, en
3. de bredere wetenschappelijke gemeenschap

verstevigd worden. Door het integreren van deze verschillende bronnen via (a) tekstontginning en (b) statistische gegevensverwerking is het mogelijk om het hoofd te bieden aan de explosieve groei van gegevens zoals we die kennen in domeinen zoals de functionele genomica – domeinen die veruit te complex zijn geworden om gevat te kunnen worden door één enkele menselijke geest.

Microroosters vormen een belangrijke technologie in de functionele genomica. Ze bepalen de simultane expressie van duizenden genen in welbepaalde omstandigheden.

In fundamenteel biologisch onderzoek enerzijds, kunnen deze omstandigheden bijvoorbeeld fazen zijn in de deling van de cel (cel cyclus) [141], of speciale condities waaraan de cel wordt blootgesteld (bv., hitteschok, hoge zuurtegraad,...) [77]. In dit soort experimenten zal men veelal geïnteresseerd zijn in het groeperen van genen naar hun biologische functie. Naar de organizatie van genen in groepen in zo'n experiment op basis van de data wordt in de literatuur veelal verwezen als clustering [73].

Anderzijds, in toepassingen die eerder verbonden zijn met het biomedische domein, worden genexpressies gemeten over een set van patiënten met een welbepaalde aandoening, en dit met het oog op een beter inzicht in de genetische mechanismen die er aan de basis liggen [37]. De achterliggende doelen situeren zich in het ontwikkelen van efficiëntere medicijnen, het tegengaan van nevenwerkingen en het toespitsen van de behandeling op individuele patienten.

In deze doctoraatstekst zal vooral toegespitst worden op het eerste soort van problemen (zijnde het functioneel karakteriseren van genen). Er zal onderzocht en aangetoond worden hoe het betrekken van informatie uit wetenschappelijke

[4]De wetenschap die zich bezighoudt met het karakteriseren van eiwitten.

literatuur in de analyse van microrooster data, de interpretatie ervan vergemakkelijkt en de uiteindelijke 'clustering'-resultaten (zie hierboven) verbetert.

Concepten uit de moleculaire biologie

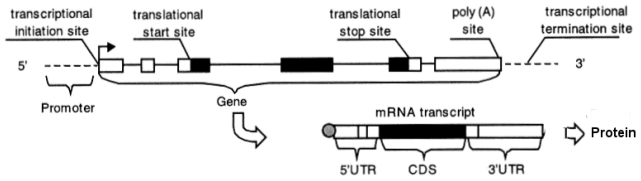
DNA

Een organisme zoals de mens bijvoorbeeld, is samengesteld uit miljarden cellen, waarvan de kern identieke paren chromosomen bevatten. Een chromosoom is een lang snoer van DNA (Desoxyribonucleinezuur). DNA heeft de structuur van een dubbele helix en is opgebouwd uit vier chemische componenten die basen genoemd worden: ze worden aangeduid met A, C, T en G. Een sterke bindingsaffiniteit tussen A en T enerzijds en C en G anderzijds waarborgt de structuur van het DNA. Deze eigenschap speelt ook een centrale rol in microroosters die verderop behandeld worden. Het DNA bevat de erfelijke informatie die identiek is in alle cellen van een organisme.

Genen en eiwitten

Eiwitten vervullen een centrale rol in de cellulaire machinerie en zowel hun scheikundige samenstelling als hun abundantie wordt bepaald en geregeld door de genen. Een gen is een stukje van het DNA dat de code bevat om een bepaald eiwit aan te maken. Het centrale dogma van de moleculaire biologie bestaat eruit dat de aanmaak van eiwitten bepaald wordt door het aflezen en vertalen van de informatie die in de genen vervat zit. Het proces kan onderverdeeld worden in twee grote stappen: Eerst vindt de transcriptie plaats van DNA naar boodschapper RNA or mRNA. Deze overschrijving of transcriptie gebeurt volgens een bepaalde chemische orientatie (5' to 3') tussen de transcriptie initiatie site (transcriptional initiation site) en de terminale transcriptie site (transcriptional termination site). De overgeschreven genetische informatie bevat cruciale informatie over de opeenvolging van aminozuren die het eiwit bepalen. Volgend op de transcriptie komt de vertaalslag van de coderende mRNA sequentie (CDS) naar het eiwit. Regio's van het mRNA die niet worden vertaald worden UTR's (untranslated regions) genoemd. De DNA regio voor de overgeschreven sequentie, wordt de promotor genoemd. Het bevat bindingsplaatsen voor eiwitten die de transcriptie regelen, de transcriptie factoren (transcription factors). Figuur 0.1 geeft een overzicht van dit gehele proces, ook wel gen expressie genoemd.

Transcriptie factoren (TF) binden op bepaalde plaatsen van het DNA die veelal stroomopwaarts ten opzichte van een gen liggen (waaronder de promoter regio). Figuur 0.2 toont hoe deze TF aan het DNA kunnen binden en hoe ze de transcriptie (en dus de expressie) van genen kunnen beinvloeden. Via veranderingen in scheikundige configuraties van eiwitten kunnen transcriptiefactoren zorgen voor het opstarten, het beëindigen het versterken of verminderen van de genexpressie. De bindingsplaats van een transcriptie factor wordt ook wel 'motief' genoemd.



Figuur 0.1: Overzicht van de gen uitdrukking of gen expressie: Eiwitten worden uit DNA geassembleerd via het mRNA. Transcriptie gebeurt volgens een bepaalde chemische orientatie (5' to 3') tussen de transcriptie initiatie site (transcriptional initiation site) en de terminale transcriptie site (transcriptional termination site). De overgeschreven genetische informatie bevat cruciale informatie over de opeenvolging van aminozuren die het eiwit bepalen. Volgend op de transcriptie komt de vertaalslag van de coderende mRNA sequentie (CDS) naar het eiwit. Regio's van het mRNA die niet worden vertaald worden UTR's (untranslated regions) genoemd. (Illustratie uit Alberts *et al.* [36])

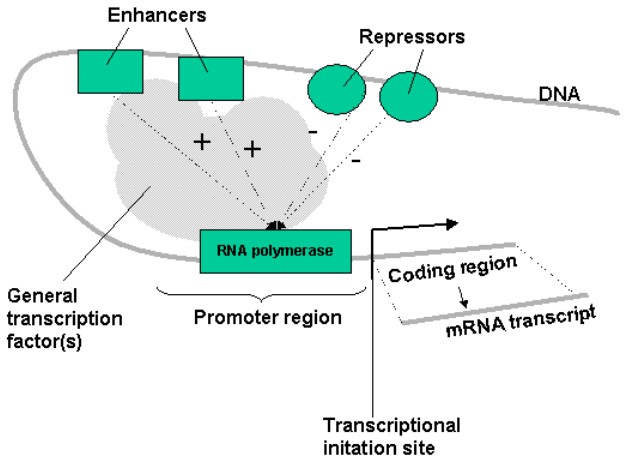
Microroosters

Recente technologische ontwikkelingen laten toe om biologische processen op genomische schaal te onderzoeken. De technologie die het meest in de kijker komen te staan de jongste jaren is die van de microroosters of gene chips. Door de miniaturisering van de biotechnologie kan hiermee nu de expressie van duizenden genen in één keer gemeten worden. Typische vraagstukken die hiermee aangepakt worden zijn

- Welke functies hebben bepaalde genen ?
- Hoe is de transcriptie van genen gereguleerd ?
- Hoe verandert de expressie van genen naar gelang bepaalde behandelingen, ziektes of andere omstandigheden ?

Een microrooster is een glasplaatje waarop enkel-strengig[5] DNA is vastgehecht op vaste plaatsen (ook 'spots' genoemd). Elke 'spot' op een rooster komt overeen met één gen en het totaal aantal spots op een rooster kan tot in de tienduizenden lopen. De sterke, unieke bindingsaffiniteit tussen twee complementaire (i.e., waarbij A met T en C met G overeenkomt) stukjes DNA maakt dat mRNA uit een onderzocht staal kan binden (of: hybridizeren) met zijn uniek complement op een spot op het rooster. De hoeveelheid mRNA van elk gen wordt veelal gemeten in een relatief kader, i.e., men meet het expressieniveau van een gen in een staal ten opzichte van een controle (bv. ziek t.o.v. gezond weefsel). Om het verschil te zien, merkt men het mRNA van controle en test staal met twee fluorescente merkers (rood en groen). Na beschijning van het microrooster met een laser krijgt met zo een beeld van de uitdrukking van elk gen op het rooster aanwezig. Elke spot op het rooster waar meer mRNA van het

[5]DNA is normaal een dubbele streng opgekruld als een helix.

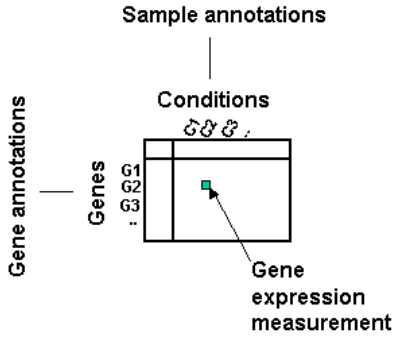


Figuur 0.2: Transcriptionele regulatie: Transcriptie factoren (voorgesteld als blokken) zijn eiwitten die op specifieke plaatsen van het DNA binden (binding sites). Bij modelorganismen zoals gist bevinden deze zich stroomopwaarts van het gen dat uitgedrukt wordt, i.e., vóór de de transcriptie startplaats (transcription inititiation site). Via wijzigingen in scheikundige configuraties kunnen transcriptiefactoren informatie doorgeven aan het RNA polymerase complex om gen expressie te activeren, te stoppen, te bevorderen of te verminderen.

test-staal is gehybridizeerd zal roder oplichten; het overeenkomstige gen is dan méér uitgedrukt, of over-uitgedrukt, in het test- ten opzichte van het controle-staal. Omgekeerd zal een onder-uitgedrukt gen groener oplichten. Wanneer het gen niet uitgedrukt is, licht de spot niet op (zwart). Na quantificatie van de fluorescentie-intensiteiten is het uiteindelijke resultaat van een microrooster experiment veelal[6] een matrix die de relatieve over- of onder-expressie van alle genen onder invloed van verschillende condities in een bepaald experiment bevat. In Figuur 0.3 visualizeren we een gen expressie matrix. Om bruikbaar te zijn dient deze matrix aangevuld te worden met gedetailleerde informatie aangaande (a) de genen die op het rooster aanwezig zijn en (b) de experimentele condities die bestudeerd worden. Enerzijds is er de informatie aangaande de laboratorium-specifieke gegevens die een microrooster faciliteit gebruikt om haar productieproces te ondersteunen, anderzijds is er de informatie aangaande de experimenten zelf (i.e., de hybridizaties). Het opslaan van het eerste type informatie wordt typisch behandeld door een labo informatie systeem of LIMS (Laboratory Information Management System). Informatie aangaande de experimenten vinden we eerder terug als bijlage bij wetenschappelijke publicaties of in publieke databanken die dit soort informatie groeperen.

In Figuur 0.4 tonen we hoe de twee types informatie zich verhouden. In de doorsnede bevindt zich de minimale informatie die benodigd is om een micro-

[6]Er zijn andere technologieën zoals oligonucleotide chips die absolute expressie meten. Deze worden hier niet besproken.



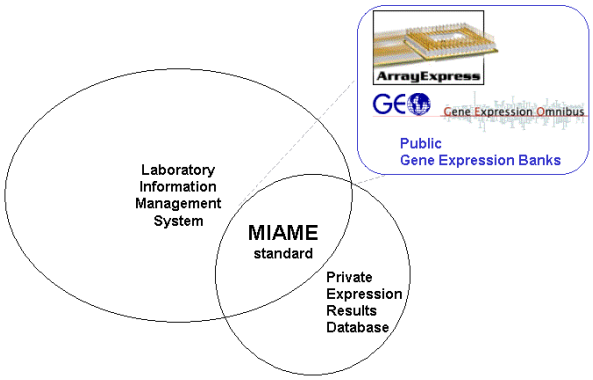
Figuur 0.3: Gen expressie matrix als resultaat van een microrooster experiment. Elke positie in de matrix bevat informatie dat het gen expressie niveau van een bepaald gen in een bepaald experimenteel staal, dat met een bepaalde conditie overeenkomt (gene expression measurement). Gedetailleerde informatie over zowel de genen (gene annotations) als de experimentele omstandigheden (sample annotations) die bestudeerd worden zijn cruciaal voor zowel de reproduceerbaarheid als de bruikbaarheid van een microrooster experiment.

rooster accuraat en reproduceerbaar te beschrijven. De precieze inhoud is vastgelegd in een internationale standaard, MIAME (Minimum Information About a Microarray Experiment). Eens geïmplementeerd in een gegevensformaat kan deze minimale informatie typisch gebruikt worden als leidraad om grote microrooster gegevensbanken aan te maken, of om informatie over microrooster experimenten uit te wisselen.

LIMS

Tijdens het doctoraatswerk hebben we een implementatie van een eerste laboratorium informatie systeem aan de VIB microarray facility[7] verwezelijkt. Het Vlaams Interuniversitair Instituut voor Biotechnologie (VIB[8]) is een non-profit, wetenschappelijk onderzoeksinstituut, waarin een zestigtal Vlaamse onderzoeksgroepen actief zijn in verschillende domeinen van de levenswetenschappen. De VIB microarray facility (VIB-MAF) is een centrale faciliteit die microroostertechnologie aanbiedt aan onderzoeksgroepen in het VIB. Ten tijde van het project werkte de VIB-MAF enkel met cDNA microroosters. cDNA's zijn gekloonde kopijen van boodschapper RNA, of mRNA. cDNA microroosters worden samengesteld door een selectie van fragmenten cDNA (EST's genoemd), gekozen uit gespecializeerde bibliotheken (bv. dBEST [5]), te plaatsen op gerichte posities op het rooster (i.e., een glasplaatje, zie hierboven). Dit proces wordt ook *spotten* genoemd. Aangezien elke spot bedoeld is om de activiteit van één enkel gen te meten, is het cruciaal dat een correcte link tussen spot en EST kloon, het productieproces dat deze doorlopen heeft en het gen dat deze

[7] www.microarrays.be
[8] www.vib.be



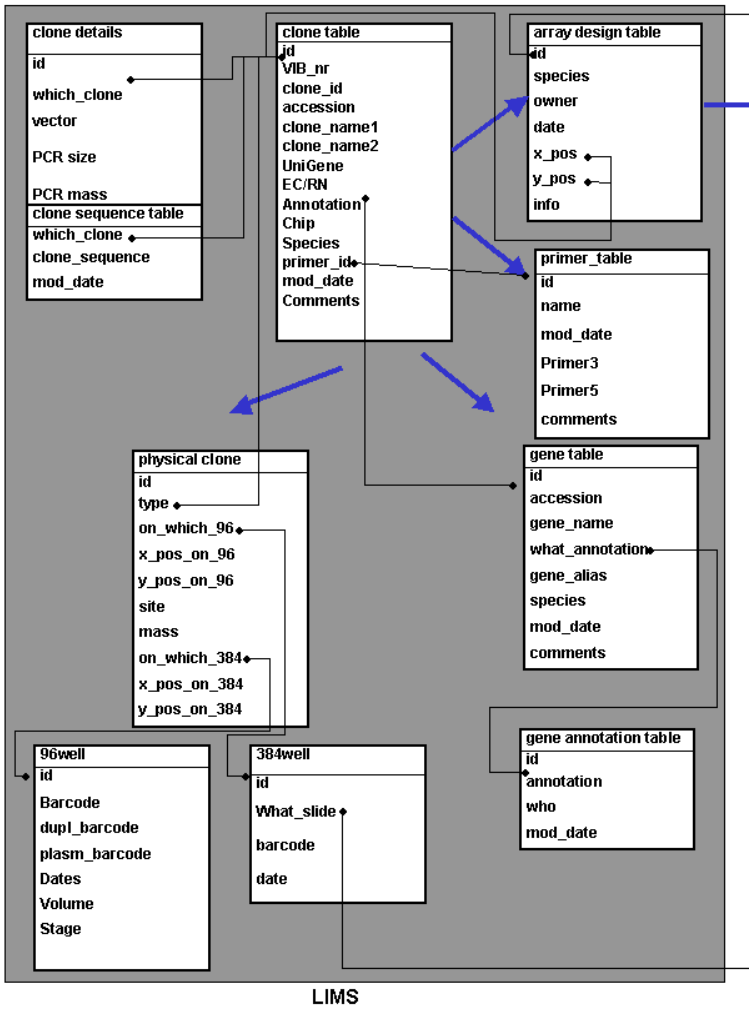
Figuur 0.4: Laboratorium informatie systemen (LIMS) ondersteunen het personeel bij microrooster experimenten door een leidraad te bieden bij het opslaan en opvragen van allerlei productiegegevens die het eigenlijke microrooster experiment voorafgaan. Alle informatie die cruciaal is voor de reproducibiliteit en klaarheid van een microrooster experiment is vastgelegd in de internationale standaard MIAME. Deze, alsook supplementaire, informatie wordt typisch opgeslagen in private of publieke gegevensbanken (public and private expression results database).

uiteindelijk voorstelt, beschikbaar is. Dit vormt deel 1 van het opgezet systeem. Het ondersteunend data model is weergegeven in Figuur 0.5.

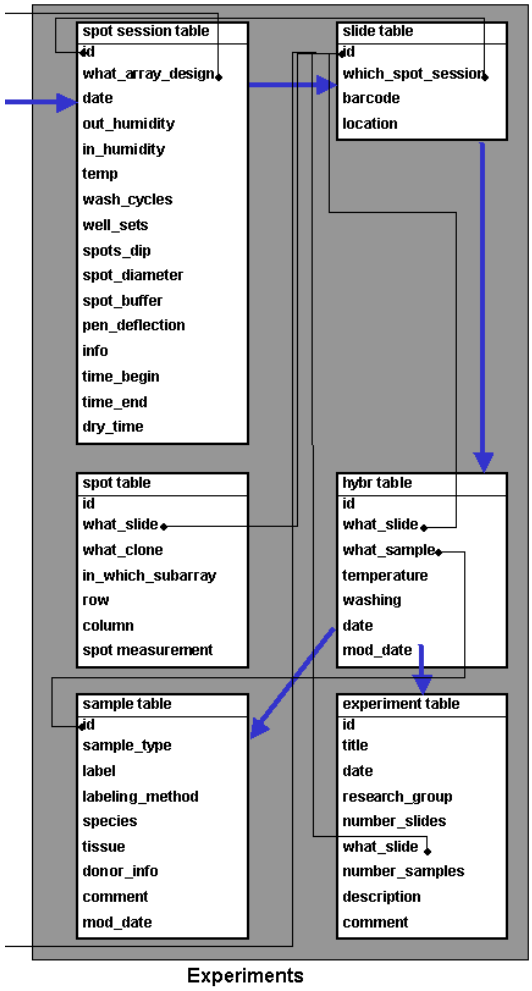
Informatie omtrent het microrooster experiment zelf (i.e., de hybridizaties) zit vervat in het tweede deel van het databank schema (Figuur 0.6). Hier vinden we informatie over de test en controle stalen gebruikt bij de verschillende hybridizaties op de verschillende roosters, alsook de opgemeten resultaten van het experiment en alle bijhorende protocols. Deze informatie is typisch benodigd door MIAME, beschreven in volgende sectie.

MIAME

De nood aan publieke databanken die microrooster gegevens bevatten wordt sinds enige tijd ervaren [71]. Het primaire doel is over een publiek platform te beschikken waar informatie die wetenschappelijke publicaties omtrent gen expressie ondersteunt, een gecentraliseerde plaats krijgt. Het is echter niet onmiddellijk duidelijk welke informatie dient opgeslagen te worden in zulke gegevensbank opdat zij achteraf bruikbaar zou zijn. Idealiter zou zulk een databank de mogelijkheid moeten bieden om analyses te doen over verschillende experimenten heen met behulp van bijvoorbeeld data mining. Het is duidelijk dat microrooster informatie gevoeliger complexer is dan bijvoorbeeld informatie omtrent DNA sequentie. Inderdaad, waar een DNA sequentie, eens bepaald, ondubbelzinnig is, heeft een gen expressie experiment enkel betekenis in een gegeven context. Zo dienen de condities waaronder de expressie gemeten werd duidelijk gestipuleerd te worden, en het aantal mogelijkheden hierbij is onmete-



Figuur 0.5: Ontwikkeld LIMS databank model – deel 1. Dit blok bevat informatie specifiek aan het productieproces met hoofdzakelijk informatie over de klonen. Er wordt bijgehouden naar welke microtiter platen de klonen verhuizen (tabellen *physical_clone*, *96well* and *384well*). Details over hoe de klonen vermeningvuldigd zijn, zit in tabel *primer*. Een goede beschrijving van een kloon is cruciaal als men de link terug wil maken naar de genen. Zie tabellen *clone*, *clone_details*, *clone_sequence* and *gene*.



Figuur 0.6: Ontwikkeld LIMS databank model – deel 2. Hier registreren we informatie over de experimentele waarden en de gebruikte protocols van de reeks hybridizaties uitgevoerd in een experiment. Het merendeel van de informatie benodigd door de MIAME standaard ligt hierin opgeslagen.

lijk. Ook de manier waarop expressie gemeten en geanalyseerd wordt, vertoont veel variaties gaande van verschillen in aangewende technologie, protocols, software en algoritmen.

De MIAME standaard [54] behandelt deze punten door middel van een breed gesteunde specificatie van de minimale informatie omtrent een microrooster experiment. Zulke standaard vergemakkelijkt het opzet en de bruikbaarheid van gen expressie databanken. MIAME richt zich tevens naar producenten van microroosters en software ontwikkelaars opdat zij de benodigde MIAME informatie op een gebruiksvriendelijke manier zouden leveren aan hun eindgebruikers teneinde dezen te ontlasten. MIAME bestaat uit zes grote stukken:

1. Experimenteel design: de set van hybridizatie-experimenten,
2. Rooster design: elk rooster met een beschrijving van elke spot op het rooster,
3. Stalen: gebruikte stalen, hoe ze geëxtraheerd zijn en hoe de fluorescerende merkers zijn aangebracht,
4. Hybridizaties: procedures en parameters ervan,
5. Metingen: de beelden en de quantificatie procedure,
6. Normalizatie: gebruikte controles, normalizatie procedures en parameters.

Voor een gedetailleerd overzicht verwijzen we naar de originele publicatie van Brazma *et al.* [54], tot dewelke we actief hebben bijgedragen. Het artikel is, twee en een half jaar na publicatie, ongeveer 200 maal geciteerd en via het data formaat MAGE-ML [140] wordt MIAME gebruikt als communicatie tussen databanken, software en gebruikers. Enkele belangrijke tijdschriften zoals Nature, Science, EMBO Journal en The Lancet eisen nu dat voor een microroostergerelateerde publicatie de bijhorende gegevens conform de MIAME standaard worden verstuurd naar een publieke databank [72].

De erkenning van MIAME is echter geen triviaal proces. Gezien de relatief hoge kost van microrooster experimenten heerst(e) er argwaan bij veel biologen om zomaar hun gegevens publiek te maken. Het volledig vrijgeven van data kan inderdaad hints geven aan concurrerende onderzoekers over genen die betrokken zijn bij een bepaald proces, maar waarnaar het onderzoek nog lopende is. MIAME vangt dit op door geen eis te stellen aan wat niet gerapporteerd wordt. Een ander probleem is dat sommige informatie over de microroosters confidentieel is, en dus niet kan geleverd worden door gebruikers. Dit is vooral het geval bij commerciele platformen zoals Affymetrix. De stuurgroep die MIAME opvolgt, MGED[9], streeft naar actieve betrokkenheid van deze industriesegmenten in MIAME.

Als besluit stellen we dat MIAME een belangrijke impact heeft op het gebruik en de bruikbaarheid van microrooster experimenten. Gegevensbanken

[9] www.mged.org

die MIAME ondersteunen zullen een waardevolle bron uitmaken om (a) nieuwe experimenten te plannen, (b) gedane experimenten te interpreteren of om (c) volledige *in silico* analyses uit te voeren.

Cluster analyse

Grootschalige studies zoals bijvoorbeeld microrooster experimenten bevatten de metingen van duizenden genen over tientallen condities. Om beter inzicht te krijgen in zulke gegevens wendt men technieken uit de exploratorische data analyse aan. Eén van die technieken is clustering, waarbij de objecten van een data set gegroepeerd worden in een aantal categorieën volgens bepaalde criteria. Door groepen van genen te maken die tesamen zijn uitgedrukt, hoopt men zo een idee te krijgen van welke (mogelijk ongekende) genen een biologische functie hebben die gelijkaardig zijn aan die van reeds gekende. Ook wanneer we genen analyseren op basis van de literatuur is het doel een groepering te maken naar gelijkaardige biologische functies. In ons werk maken we voornamelijk gebruik van hiërarchische clustering en K-means clustering. Voor deze uitgelegd worden, starten we met een overzicht van de algemene methodologie achter clustering.

Methodologie

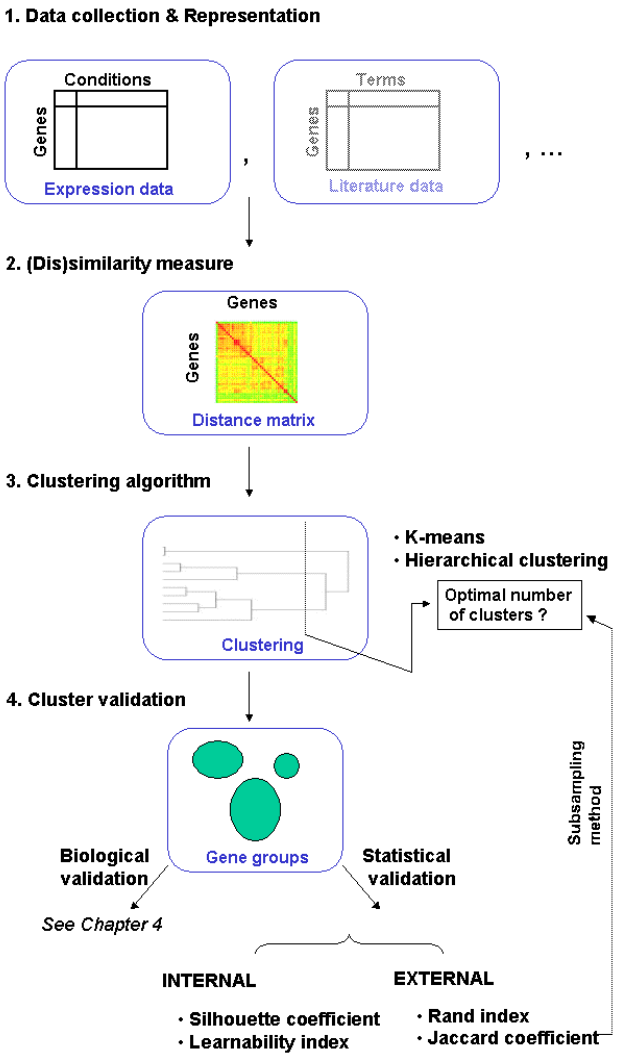
Een exploratorische analyse via clustering bestaat uit verschillende stappen, gevisualiseerd in Figuur 0.7:

1. Verzameling van data en representatie. Data dient verzameld te worden en bijgewerkt (preprocessing). Een goede representatie van het probleem en de gegevens kan de analyse beduidend efficiënter maken. We passen het clustering raamwerk toe op microrooster data en literatuur-gebaseerde gegevens.

2. Keuze van een goede metriek om similariteit tussen genen te meten. We merken op dat deze keuze soms inherent is aan de keuze van het cluster algoritme (volgende punt).

3. Keuze van een gepast cluster algoritme. In dit werk gebruiken we K-means and hiërarchische clustering. Beiden vereisen op een zeker ogenblik de keuze van het aantal clusters k. Dit is echter geen triviale keuze, en methodes om dit automatisch te bepalen worden voorgesteld in dit werk.

4. Statistische validatie van cluster resultaten. Aangezien clustering een exploratorische techniek is, zijn soms meerdere oplossingen mogelijk. Om de cluster oplossingen (statistisch) te valideren gebruiken we zowel interne als externe maten.



Figuur 0.7: Overzicht van de clustering methodologie: Data wordt voorgesteld door middel van een representatie (**1**) en na keuze van een gepaste metriek worden (dis)similariteiten berekend (**2**). Na toepassing van cluster algoritmes zoals K-means of hiërarchische clustering (k bepaald via de stabiliteitsmethode) (**3**), worden clusters gevalideerd met de M-score of via een statistische coëfficiënt (**4**).

Voorafgaande gegevensverwerking (preprocessing)

Gen expressie data is typisch georganiseerd in een $n \times p$ gen expressie matrix (zie ook Figuur 0.3), waarin elke rijvector $x_{i\cdot}$ de quantificatie voorstelt van het niveau van expressie van een gen i over verschillende condities $j = 1 \dots p$. Het gemiddelde en de variantie op deze matrix is gedefineerd als volgt:

$$\begin{cases} \overline{x_i} = \frac{1}{p} \sum_{l=1}^{p} x_{il} \\ \sigma_i^2 = \frac{1}{p} \sum_{l=1}^{p} (x_{il} - \overline{x_i})^2 \end{cases} \tag{0.1}$$

Bij cDNA microroosters wordt veelal met expressie ratio's gewerkt, i.e. de verhouding van expressie niveau van test versus controle staal voor elk gen. Om de schaal voor *over*- en *onder*-expressie gelijk te maken, neemt men meestal de logaritme van deze ratio's. In het geval er waarden ontbreken in de matrix, worden deze meestal geïnterpoleerd. Wij behandelen ze zoals beschreven in [24][67].

Het is nuttig om de cluster analyse te beperken tot de genen wier expressie niveau verandert doorheen de opgemeten condities. Hiermee vermijdt men achteraf overbodige interpretaties te doen, aangezien genen met een kleine variabiliteit soms een belangrijke component van ruis kunnen vormen. Een voorbeeld volgt hieronder.

Om het gedrag van genen onderling te vergelijken is hun relatief gedrag meer belangrijk dan de absolute expressiewaarden. Vandaar standardizeert men de gen expressie profielen $\{x_i\}_{l=1..p}$ als volgt:

$$x'_{i\cdot} = \frac{x_{i\cdot} - \overline{x_i}}{\sigma_i}. \tag{0.2}$$

Indien men genen met lage variantie over de condities niet zou filteren zoals hierboven, zou deze operatie willekeurige profielen opleveren.

Wanneer we echter literatuur-data gebruiken, gebeurt de preprocessing anders. Dit wordt echter behandeld in de sectie over tekstontginning.

Similariteitsmaten

De Pearson correlation coëfficiënt is een populaire maat in gen expressie analyse en meet de correlatie tussen gecentreerde gen profielen :

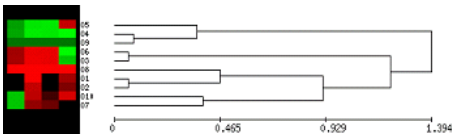
$$s_{\text{Pearson}}(i,j) = \frac{\sum_{l=1}^{p} (x_{il} - \overline{x_i})(x_{jl} - \overline{x_j})}{\sqrt{\sum_{l=1}^{p} (x_{il} - \overline{x_i})^2} \sqrt{\sum_{l=1}^{p} (x_{jl} - \overline{x_j})^2}}. \tag{0.3}$$

De cosinus meet ongecentreerde correlatie en is gedefinieerd als:

$$s_{\text{Cosine}}(i,j) = \frac{\sum_{l=1}^{p} x_{il} x_{jl}}{\sqrt{\sum_{l=1}^{p} x_{il}^2} \sqrt{\sum_{l=1}^{p} x_{jl}^2}}. \tag{0.4}$$

Om de similariteiten (0.3) en (0.4) in afstanden te vertalen gebruiken we

$$d(i,j) = d_{ij} = 1 - |s_{ij}|. \tag{0.5}$$



Figuur 0.8: Illustratie van een dendrogram op microrooster data van 10 genen over 4 condities. Het bepalen van het optimaal aantal clusters (i.e., waar het dendrogram wordt gesneden) is een niet-triviaal probleem. Vaak wordt de gen expressie matrix herordend en van een kleurcode voorzien (rood=overuitgedrukt; groen=onderuitgedrukt).

Tot slot, een laatste, belangrijke gebruikte afstandsmaat is de Euclidische afstand

$$d(i,j) = \sqrt{\sum_{l=1}^{p}(x_{il} - x_{jl})^2}. \tag{0.6}$$

Gebruikte cluster algoritmes

Hiërarchische methods

Hiërarchische clustering is deterministisch en organiseert de data in een binaire (soms ternaire, quaternaire,...) boom waarvan de laatste 'kinder'-knopen de objecten in de data set zijn. De verschillende 'ouder'-knopen in de boom zelf stellen kandidaat clusters voor. We onderscheiden agglomeratieve en divisieve hiërarchische clustering. Het eerste type wordt hier behandeld, het tweede type wordt via een toepassing geïllustreerd verderop.

Als startpunt geldt een dissimilariteitsmatrix D dat de afstanden tussen all n punten bevat volgens een gekozen afstandsmaat. Startend van singleton clusters, worden de dichtsbijzijnde clusters (voorgesteld als twee kinderknopen) paarsgewijs samengebracht in een nieuwe cluster die de ouderknoop vormt van deze twee clusters in de hiërarchische boom. Dit wordt iteratief herhaald tot er één cluster rest. Het resultaat is een volledige boom ook wel dendrogram genoemd. Een voorbeeld van een dendrogram voor microrooster data is gegeven in Figuur 0.8 en is voor het eerst toegepast door Eisen *et al.* [73]. Bij de visualizatie van de microrooster analyse wordt elk element typisch ingekleurd naargelang er over- (rood) of onder- (groen) expressie is.

Er bestaan verschillende manieren om kinderknopen in het dendrogram samen te voegen tot een nieuwe ouderknoop. Algemeen schrijven we het als volgt: Stel dat clusters r and s samengevoegd worden tot cluster rs. De afstand, $d(\cdot,\cdot)$, van rs tot de overige clusters k wordt berekend als volgt:

$$d(rs,k) = \alpha_r d(k,r) + \alpha_s d(k,s) + \beta d(r,s) + \gamma|d(k,r) - d(k,s)|. \tag{0.7}$$

De keuzes voor de parameters α_r, α_s, β en γ zijn weergegeven in Tabel 0.1.

Tabel 0.1: Overzicht van verschillende agglomeratieve hiërarchische cluster methoden. We tonen verschillende keuzes voor parameters α_r, α_s, β en γ met n_r, n_s en n_k de cluster grootte van clusters r, s, and k. 'Single link' komt overeen met de minimale afstand tussen de clusters, 'complete link' met een maximum afstand, en 'average link' met een gemiddelde cluster afstand. Ward's methode meet de verandering in de kwadratische afstand tot het cluster centrum als clusters r en s tot rs samengevoegd worden.

Method	α_r	α_s	β	γ
Single link	$1/2$	$1/2$	0	$-1/2$
Complete link	$1/2$	$1/2$	0	$1/2$
Average link	$\frac{n_r}{n_r+n_s}$	$\frac{n_s}{n_r+n_s}$	0	0
Ward's method	$\frac{n_r+n_k}{n_r+n_s+n_k}$	$\frac{n_s+n_k}{n_r+n_s+n_k}$	$\frac{-n_k}{n_r+n_s+n_k}$	0

Statistische validatie en optimale k

Aangezien een goed clustering resultaat niet éénduidig is gedefinieerd, zijn er verschillende manieren om het te valideren. Interne maten kijken naar statistische eigenschappen van de resulterende clusters (bv. de Silhouette coëfficiënt meet de afstanden binnen elke cluster versus de afstanden tussen de clusters onderling). Externe maten daarentegen quantificeren de overeenkomst met een uitwendig gedefinieerde partitionering van de data. Hier behandelen we enkel de Jaccard coëfficiënt. De overige maten zijn van minder belang in deze samenvatting en hiervoor refereren we naar de Engelstalige doctoraatstekst.

Jaccard coëfficiënt

Gegeven een set van n objecten, een externe partitionering $P = \{P_1, ..., P_u\}$, en een clustering $C = \{C_1, ..., C_v\}$, dan definiëren we voor elke clustering een matrix M met componenten

$$M_{ij} = \begin{cases} 1 & \text{als } x_i \text{ en } x_j \text{tot dezelfde cluster behoren} \\ 0 & \text{anders} \end{cases} \tag{0.8}$$

Een correlatie tussen partitioneringen P en C kan worden geschreven als

$$\langle P, C \rangle = \sum_{i,j} M_{ij}^P M_{ij}^C. \tag{0.9}$$

Stel nu dat $N_{11}, N_{10}, N_{01}, N_{00}$ het aantal keren telt dat M_{ij}^P and M_{ij}^C de elementsgewijze patronen $\{1,1\}$, $\{1,0\}$, $\{0,1\}$, $\{0,0\}$ bevat, dan schrijven we:

$$M_{ij} = \begin{cases} \langle P, C \rangle = N_{11} \\ \langle P, P \rangle = N_{11} + N_{10} \\ \langle C, C \rangle = N_{11} + N_{01}. \end{cases} \tag{0.10}$$

De Jaccard coëfficiënt wordt nu geschreven als

$$J(P,C) = \frac{\langle P,C\rangle}{\langle P,P\rangle + \langle C,C\rangle - \langle P,C\rangle} = \frac{N_{11}}{N_{01}+N_{10}+N_{11}}, \tag{0.11}$$

en beschrijft de overeenkomst tussen twee binaire matrices wanneer overeenkomsten tussen nullen wordt genegeerd. J neemt waarden aan tussen 0 en 1: wanneer het cluster algoritme een groupering voorstelt die gelijk is aan de externe partitionering van de objecten, is $J = 1$. Deze manier om J af te leiden laat zien dat J als een soort correlatie coëfficiënt kan gezien worden.

Stabiliteitsmethode om optimale k te bepalen

De stabiliteitsmethode voor clustering is eerst voorgesteld door Ben-Hur *et al.* [46]. De onderliggende veronderstelling is dat cluster resultaten 'stabiel' zijn naarmate ze bewaard blijven als de clustering wordt toegepast op een willekeurige fracties f van de data set (bv. $f \geq 0.5$). Op die manier kunnen cluster resultaten horende bij elke parameter setting zoals k (het aantal clusters) vergeleken worden.

In de voorgestelde methode wordt voor elke waarde van k de overeenkomst gemeten tussen clusterings uitgevoerd op s paren van willekeurige subsamples X_{sub1}, X_{sub2} van de data. Elk pair van clusterings C_1, C_2 toegepast op subsamples X_{sub1}, X_{sub2} respectievelijk worden vergeleken via de Jaccard coëfficiënt J. Dit wordt s maal herhaald zodat een distributie van J bekomen wordt, die onmiddellijk wordt omgezet naar een cumulatieve distributie functie (cdf). Dit wordt herhaald voor verschillende parameter keuzes (k in dit geval), zodat we verschillende cumulatieve distributie functies verkrijgen. Wanneer we deze uitzetten verwachten we dat plotse overgangen in de vorm van de cdf's aangeeft welke keuze voor k optimaal is. Samengevat ziet het stabiliteitsalgoritme er als volgt uit:

Gegeven: fractie f, aantal subsamples s, max aantal clusters K_{max}
Itereer van $1 \ldots K_{max}$
    Herhaal s keer
        $\mathcal{S}_1$=Willekeurige fractie f van de data
        $\mathcal{S}_2$=Willekeurig fractie f van de data
        Cluster $\mathcal{S}_1$
        Cluster $\mathcal{S}_2$
        Bereken Jaccard coëfficiënt J op de geclusterde $\mathcal{S}_1 \cap \mathcal{S}_2$
Visualiseer alle cdf's van J voor elke k om meest stabiele k te bepalen.

Een toepassing van deze methode op microrooster data over duizenden genen blijkt een uitdaging aangezien het aantal stabiele clusters in zulke complexe data set wordt onderschat zo blijkt uit onze studie. Daarom implementeren we deze methode als een divisief hiërarchisch cluster algoritme, dat samengevat wordt in Figuur 0.9. Dit iteratief scheme start met de gehele data set als één grote

cluster en creëert hieruit een initieel, optimaal aantal nieuwe clusters (berekend via de stabiliteitsmethode). Vervolgens wordt de procedure herhaald voor elk van de bekomen clusters. Divisieve hiërarchische methodes vinden we ook terug in ander werk [125][89][38]. Het combineren van het divisief schema met de stabiliteitsmethode van Ben-Hur *et al.* [46] vormt een eigen inbreng.

Evaluatie van clustering via transcriptionele regulatie

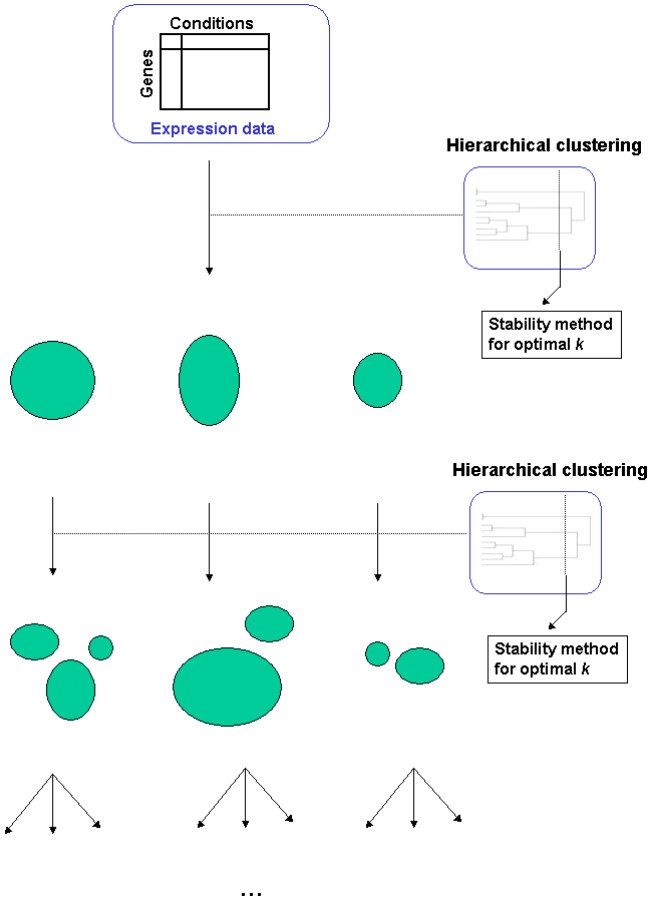
Zoals gezegd is clustering een 'slecht' gedefinieerd probleem aangezien de objectieven ervan vaag gesteld zijn. Zo heeft een cluster algoritme een aantal parameterizaties waar meerdere zinvolle keuzes mogelijk zijn (bv. het aantal clusters k). Er bestaan reeds een aantal quantitatieve methoden om cluster resultaten te evalueren. Methoden zoals de Silhouette-coëfficiënt [100][125], de stabiliteitsmethode uit vorige sectie [46], de 'Figure Of Merit' [165] en de 'Gap'-statistiek [153], maken een inschatting van de optimaliteit van een clustering resultaat zich baserende op bepaalde statistische eigenschappen van de geclusterde data set. Echter, deze methoden hebben als nadeel dat ze geen biologische kennis gebruiken, aangezien ze enkel rekening houden met de data set die geanalyseerd wordt. De complexiteit van microrooster gegevens maakt dat het soms nuttig is een ander standpunt in te nemen.

Zo is er reeds vanaf het begin dat microroosters hun intrede deden een interesse in transcriptionele regulatie. Het achterliggende idee is dat 'goed' geclusterde groepen van genen, genen zijn die samen tot uitdrukking komen en dus ook mogelijk regulatorische signalen in hun DNA delen [56][65][152][160]. Echter, veelvuldige cluster oplossingen (overeenkomstig een aantal algoritmen en bijhorende parameterisaties) evalueren in termen van transcriptionele regulatie is een tijdrovende onderneming aangezien de biologische validiteit van zulke resultaten door een biologisch expert moet nagegaan worden. Een voorbeeld kan men vinden in Gasch en Eisen [77]. We stellen dus vast dat:

1. Puur data-gebaseerde scores niet altijd een indicatie geven naar cluster oplossingen met een signifante aanrijking aan motieven.

2. De huidige manieren om grouperingen van genen te evalueren op aanwezigheid van interessante motieven/motief patronen zich beperken tot een tijdrovende, handmatige analyse.

Deze vaststellingen vormden de basis ter ontwikkeling van een heuristiek, op motieven gebaseerd, om snel de biologische significantie van een cluster oplossing in te schatten. In het geval meerdere cluster algoritmes aangewend worden, of veel verschillende parameter instellingen mogelijk zijn, blijkt deze heuristische score bijzonder nuttig. In wat volgt stellen we de score voor en schetsen we de voornaamste resultaten indien toegepast op gen expressie data van de gist cel cyclus.



Figuur 0.9: Hiërarchisch, divisief clustering schema. In een eerste ronde wordt de data gepartitioneerd in verschillende clusters via hiërarchische clustering, met optimale k bepaald via de stabiliteitsmethode. Elk van deze clusters nogmaals doorgegeven aan een hiërarchisch cluster algoritme volgens dezelfde procedure. Het proces wordt gestopt wanneer er geen stabiele oplossingen meer kunnen gevonden worden.

De M-score

Startend van een set van M DNA sequentie motieven en een clustering van een data set in k clusters, definiëren we een cluster-over-motief matrix A van dimensie $k \times M$. Elk element in de matrix bevat een p-waarde die de statistische overrepresentatie meet van het betrokken motief in die cluster. Voor een gegeven motief j in een gegeven cluster i, stel r het aantal genen (in i) dat j bevat. n is de grootte van de cluster en m het totaal aantal genen (in de data set) dat één of meer kopies van het motief j in hun stroomopwaarste regio bevat. N tenslotte is het totaal aantal genen in de data set. In de veronderstelling dat $N >> n$, kunnen we de waarschijnlijkheid dat we motief j *toevalligerwijs* bij r of meer genen waarnemen uitdrukken als:

$$p\text{-value} = 1 - \sum_{l=0}^{r-1} \mathcal{B}(l, n, \frac{m}{N}), \tag{0.12}$$

met $\mathcal{B}$ de klassieke binomiaalverdeling. We merken op dat de achtergrond probabiliteit, $\frac{m}{N}$, data-afhankelijk is, aangezien enkel de genen die overblijven na filtering (zie hierboven) in beschouwing genomen worden. Deze methode om significantie te berekenen wordt bijvoorbeeld ook toegepast in Gasch en Eisen [77]. Wanneer $r \leq 8$, gebruiken we de Fisher Exact test die meer robust is bij kleine samples [107]. Tot slot voeren we een correctie door voor multiple testing, d.w.z., aangezien elk motief over verschillende clusters tegelijkertijd getest wordt, corrigeren we, kolomsgewijs, alle p-waarden volgens Bonferroni:

$$p\text{-waarde}^* = \min(p\text{-waarde} \cdot k, 1).$$

Vooraleer we, vertrekkende van de matrix A met alle p-waarden, de M-score afleiden, sommen we de onderliggende veronderstellingen op:

- Een motief is minder interessant wanneer het (significant) voorkomt in veel clusters;
- Als de set van M motieven waarnaar gezocht wordt groot genoeg is, zal een cluster die een grote portie van deze motieven bevat, waarschijnlijk biologisch minder zinvol zijn;
- Een (erg) hoog aantal clusters overschat de biologische diversiteit die met een microrooster experiment bestudeerd wordt.

Onze M-score balanceert tussen deze drie criteria en is samengesteld als volgt:

$$\text{M-score} = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{M} \underbrace{\log(\frac{M}{f_{\{1..M\} \in i}})}_{\text{IMF}} \underbrace{\log(\frac{k}{f_{\{1..k\} \ni j}})}_{\text{ICF}} \underbrace{\cdot A(i,j)}_{p-\text{waarde}}, \tag{0.13}$$

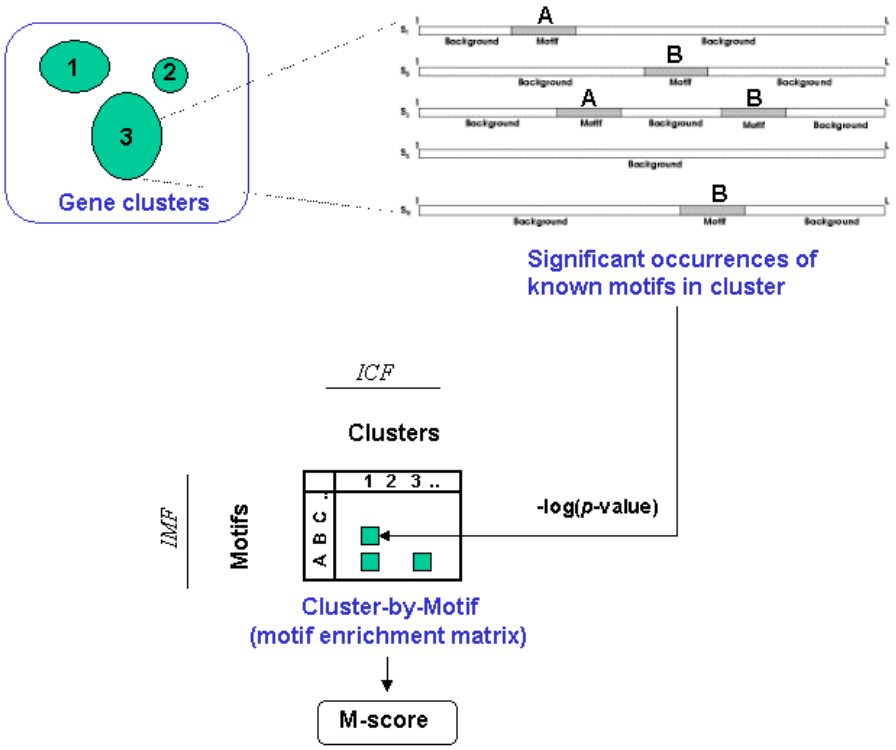
met $f_{\{1..M\}\in i}$ het aantal significante motieven in cluster i, $f_{\{1..k\}\ni j}$ het aantal keer dat motief j significant voorkomt in alle k clusters en $A(i,j)$ de p-waarde voor motief j in cluster i. De term IMF kan gezien worden als een *inverse motief frequentie*, ICF als een *inverse cluster frequentie*. Hiermee balanceren we tussen enerzijds clusterresultaten met clusters met 'te veel' significante motieven (typisch voorkomend als een cluster te groot is) en anderzijds resultaten waar motieven voorkomen doorheen veel clusters (typisch voorkomend als de clusters te sterk opgedeeld zijn).

Toepassing op cel cyclus microrooster data

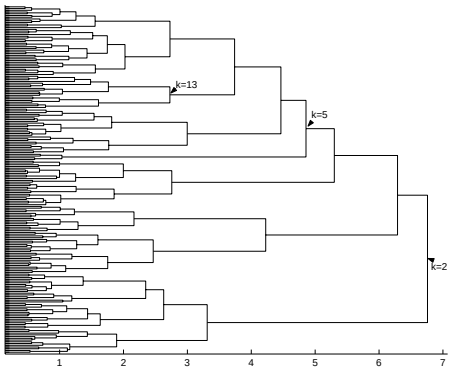
We passen de M-score toe op microrooster data van Spellman *et al.* [141] die de cel cyclus van gist beschrijft. De cel cyclus is een serie van activiteiten in de cel die uitmondt in cel groei en deling van de cel in twee dochter cellen. We analyseren de 1745 genen met grootste variabiteit (variantie) in hun profielen én met literatuur beschrijvingen in de SGD databank [29] gedateerd op January 2003. Om te zien of deze subselectie de resultaten nog vergelijkbaar maakt met vorige analyses op deze data zoals door Cho *et al.* [62] en Tavazoie *et al.* [148], berekenen we de lineaire correlatie tussen de a priori kans om een motief j te vinden in deze set van genen $\{P(j|N = 1745), j = 1..42\}$, versus de a priori kans om een motief j te vinden in het hele genoom $\{P(j|N = 6182), j = 1..42\}$. De correlatie is 0.998, wat maakt dat we geen ruis geïntroduceerd hebben door deze selectieprocedure. We clusteren deze data met Ward's agglomeratieve hiërarchische clustering en gebruiken de cosinus-gebaseerde afstand. Het dendrogram wordt getoond in Figuur 0.11.

Figure 0.12 toont hoe onze M-score varieert over verscheidene k en illustreert de Bonferonni en Fisher correcties, alsook het gedrag van de score in geval van gepermuteerde data.

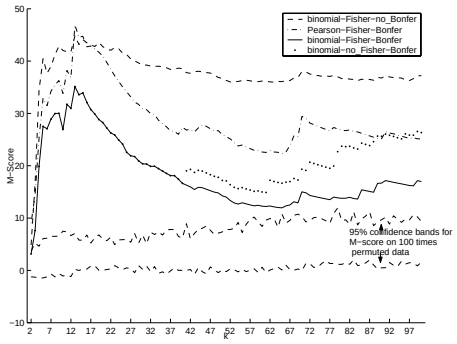
In Figuur 0.13 tonen we de cluster-over-motief matrix A voor $k = 5, 13, 30$. Het cluster resultaat met $k = 13$ geeft aanleiding tot de hoogste M-score en na evaluatie door een bioloog blijkt dit inderdaad de biologisch interessantste oplossing (beter dan $k = 5, 30$). Een partitionering in $k = 5$ clusters is biologisch minder plausibel wanneer we de expressieprofielen van de clusters bekijken (zie volledige doctoraatstekst). In Tavazoie *et al.* wordt een opdeling onderzocht in $k = 30$ clusters. Naar onze verwachting blijkt dit inderdaad te hoog, wat tevens wordt aangegeven door een lagere M-score en na vergelijkende studie door een bioloog. Inderdaad, bij $k = 30$ worden minder significante motief patronen teruggevonden. Deze resultaten geven aan dat de M-score een goed idee kan geven over de parameterregio die overeenstemt met biologisch plausibele oplossingen. In tegenstelling tot tijdsintensieve manuele validaties biedt de score een aantrekkelijk alternatief wanneer tussen veel clusteroplossingen moet gekozen worden.



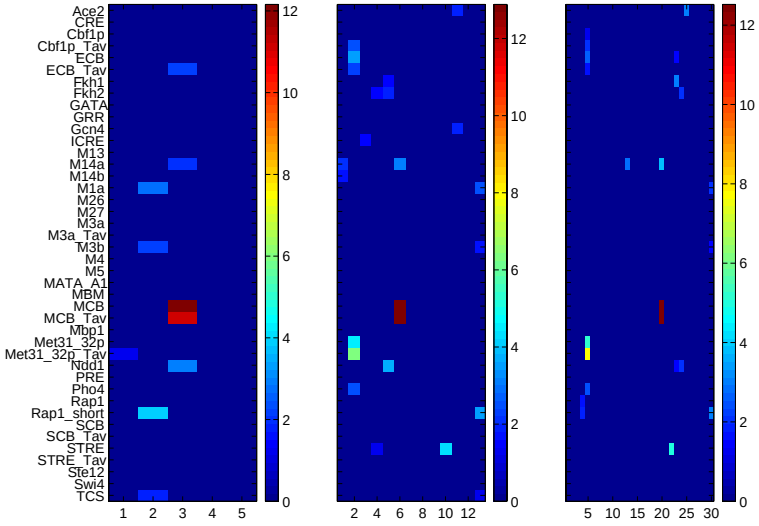
Figuur 0.10: Overzicht van de M-score. Om biologische relevantie van cluster resultaten in termen van transcriptionele regulatie te beoordelen, worden eerst, voor elke cluster, (gekende) significante motieven opgespoord in de stroomopwaartse regio's van genen in die cluster. We gebruiken hiervoor string-gebaseerde methodes. De significantie is uitgedrukt als een p-waarde en wordt opgeslagen in een motif-over-cluster matrix. Een visualizatie van die matrix kan bijvoorbeeld gebruikt worden om aanrijking van motieven te bekijken. Echter, als er veel cluster oplossingen te interpreteren zijn, zijn automatische methoden meer aangewezen. De M-score benadeelt cluster resultaten waarin clusters met (biologisch gezien) 'te veel' motieven voorkomen door de IMF factor die kolomsgewijs wordt berekend. Eveneens boeten motieven die overal voorkomen aan belang in door de ICF factor, die rijgewijs wordt berekend.



Figuur 0.11: Dendrogram van hiërarchisch geclusterde cel cyclus data. Waar het dendrogram wordt afgesneden, bepaalt het finale aantal clusters. Dit kan bepaald worden met de stabiliteitsmethode. Hier echter gebruiken we onze M-score die de cluster oplossing evalueert in termen van transcriptionele regulatie.



Figuur 0.12: M-score versus k voor gist cel cyclus data. De Fisher correctie corrigeert voor een overschatting van p-waarden in kleine clusters die veelal voorkomen bij hoge k. We tonen ook hoe de Bonferroni correctie de gepiektheid van de curve wijzigt. Tot slot zien we dat de M-score voor gepermuteerde data (i.e., waar de cluster toewijzingen aan alle genen willekeurig gepermuteerd worden) beduidend lager is en geen piekgedrag vertoont.



Figuur 0.13: Visualisatie van motief aanrijkingen voor geclusterde cel cyclus data met $k = 5, 13, 30$. Op de x-as staan de cluster nummers, op de y-as zijn de namen van de onderzochte motieven uitgezet.

Tekstontginning

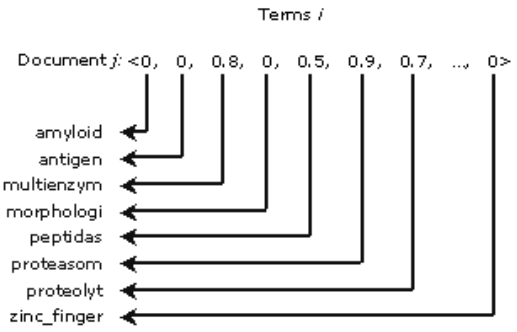
Representatie van literatuur in een vectorruimte

In het vectorruimte model [41], wordt een tekst document voorgesteld door middel van een vector, of tekstprofiel, waarvan elke component overeenkomt met één enkele term die mogelijk uit meerdere woorden bestaat. De verzameling van alle beschouwde termen wordt het vocabularium genoemd. Elke component van de vector bevat een gewicht dat het belang van die term aangeeft. Indexeren is de berekening van die gewichten:

$$w_{ij} = \log(\frac{N}{n_j}).$$

N is het totaal aantal documenten dat beschouwd wordt, en n_j het aantal documenten dat term j bevat in de document collectie. Het logaritme wordt vaak inverse document frequentie (IDF) genoemd. Elk gewicht w_{ij} in de vector van document i is een gewicht voor term j uit het vocabularium. Similariteit tussen paren documenten drukken we uit via de cosinus. De onderliggende hypothese is dat hoge similariteit tussen documenten ook betekent dat ze semantisch gerelateerd zijn.

Een schijnbaar logische keuze voor een vocabularium is het verzamelen van alle woorden die in een tekst voorkomen. In onze aanpak kiezen we voor gecustomizeerde vocabularia gebaseerd o.a. op de Gene Ontology (GO) [9]. Het gebruik van beperkende vocabularia wordt ook voorgesteld door Stephens *et al.*



Figuur 0.14: Illustratie van een index van een document van gen j met van nul verschillende gewichten voor de termen 'multienzyme', 'peptidase', 'proteasome' en 'proteolytic'. De verzameling van alle termen wordt een vocabularium genoemd en stop-woorden zoals 'from', 'the', 'often',... worden typisch verwijderd. Merk op dat termen ontdaan zijn van verbuigingen, vervoegingen en meervouden. Hiertoe werd de Porter stemmer voor Engelse taal gebruikt [126].

[142] Chiang *et al.* [61]. Aangezien GO een dynamisch gecontroleerde woordenschat is die een brede bedekking van terminologie in de functionele genomica aanbiedt, beschouwen wij het als een ideale bron voor ons vocabularium. We gebruiken GO tevens als bron voor begrippen die uit meer dan één woord bestaan (frasen), aangezien hun automatische afleiding uit teksten een niet-triviale taak is.

Nu we een index hebben geconstrueerd voor elk document, dienen we vervolgens een tekstueel gen profiel op te bouwen. Daartoe brengen we alle informatie die gerelateerd is aan elk gen samen. Hiertoe kan men gespecializeerde databanken consulteren zoals SGD voor gist [29] of LocusLink voor vele andere organismen (inclusief mens en muis) [17]. Een alternatief is voor elk gen de relevante abstracts op te vragen via PUBMED [28].

Met N_i relevante documenten voor elk gen i voorhanden, construeren we een gen index (i.e., weerom een vector van termen j) samen door een gemiddelde te nemen:

$$g_i = \{g_i\}_j = \{\frac{1}{N_i}\sum_{k=1}^{N_i} w_{kj}\}_j$$

Een voorbeeld van een gen profiel is gegeven in Figuur 0.14.

Tabel 0.2: Coherentie score op basis van tekst voor de onderzochte cel cylus groepen.

Group	p-value
Cell cycle control	1.01e-167
DNA repair	3.91e-61
Fatty acids, lipids	4.28e-08
Glycosylation	6.29e-06
Methionine	9.88e-28
Mitotic exit	1.50e-82
Nutrition	1.76e-18
Pseudohypae	2.79e-05
Secretion	1.11e-06
Sporulation	1.11e-01

Het scoren van functionele groepen berekend via tekst

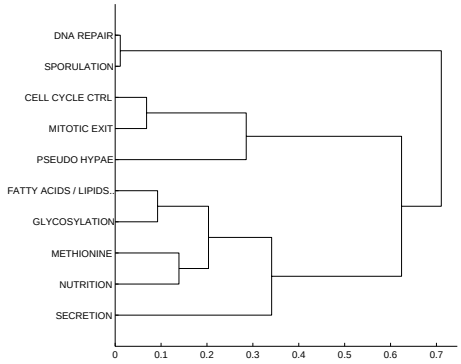
Om de geschiktheid van de vector-representatie voor functionele genomica aan te tonen, gebruiken we 10 cel cyclus specifieke groepen zoals voorgesteld in de gedetailleerde analyse van Spellman *et al.* [141](Figuur 7 aldaar). Hiertoe definiëren we een score die de coherentie van een gen groep **G** meet vanuit de tekstrepresentatie:

$$W = \text{median}(\{\cos(g_k, g_l)_{k,l}\})$$

met g_k, g_l genen uit **G**. Om de statistische significantie van deze score te bepalen (i.e., te bepalen hoe toevallig een waarde voor deze score is), genereren we een achtergrond verdeling door voor elk van de groepen een 100-voudige randomizatie te doen. De uiteindelijke functionele coherentie van een gen groep is dan uitgedrukt als een p-waarde die uitdrukt wat de kans is dat de geobserveerde W toevallig is. In Tabel 0.2 tonen we de resultaten voor de 10 groepen. Met een drempelwaarde van 0.05 zien we dat alle groepen, behalve de sporulatiegroep, gedetecteerd worden via onze vectorruimte voorstelling.

Genen clusteren op basis van literatuur

In een volgende stap gaan we na hoe via onze tekst representatie deze groepen relateert. Hiervoor gebruiken we weerom Ward's agglomeratieve hiërarchische clustering en het resultaat is afgebeeld in 0.15. Een biologische interpretatie leert ons dat de metabole processen (vetzuren (fatty acids), glycosylatie, methionine metabolisme, voeding (nutrition) and afscheiding (secretion)) dicht geclusterd zijn, wat geen verrassing blijkt aangezien metabolisme een sterk geïntegreerd proces is, waar verschillende 'pathways' elkaar doorkruisen. Tevens vinden een aantal van die processen (oxidative phosphorylatie, de TCA cyclus, aminozuur catabolisme en vetzuur oxidatie) in dezelfde cellulaire locatie plaats (namelijk de mitochondriën). Cel cyclus, de exit uit mitose (een sleutel gebeurtenis in de cel cyclus) en pseudohypae dat een reactie is op ontbering aan stikstof typisch in de G1/S faze van de cel cyclus, worden correct aan elkaar gerelateerd. Minder evident tot slot is de link tussen DNA reparatie en sporulatie, veroorzaakt door



Figuur 0.15: Dendrogram van de 10 onderzochte cel cyclus-specifieke groepen na clustering met behulp van de tekst representatie

de betrokkenheid van RAD proteinen in beide processen. We zien bijgevolg dat de tekst representatie niet enkel groepen herkent (zie vorige sectie), maar ook met elkaar kan relateren op een biologisch relevante manier. In wat volgt tonen we aan hoe sleutelwoorden voor groepen van genen geëxtraheerd worden en hoe deze kunnen worden gebruikt voor de interpretatie ervan.

TXTGate: een software platform om groepen van genen te profileren op basis van de literatuur

Motivatie

Naast een quantitieve analyse, waarvan de belangrijkste besluiten zijn samengevat in vorige sectie, hebben we onze methode ingebouwd in een praktisch raamwerk om de flexibiliteit en praktische bruikbaarheid van onze aanpak volledig tot haar recht te laten komen. Onze applicatie, TXTGate, is gestoeld op vijf belangrijke observaties aangaande biologische tekstontginning:

1. Goede literatuur referenties vinden voor genen is een onderzoeksonderwerp op zichzelf. Document classificatie strategieën zoals in Leonard *et al.* [104] of Raychaudhuri *et al.* [130] kunnen hiervoor gebruikt worden. Wij verkiezen echter manueel onderhouden gen-literatuur associaties zoals beschikbaar in een aantal belangrijke publieke databanken.

2. Informatie in wetenschappelijke (MEDLINE) abstracts geassocieerd aan genen blijft divers en kan gaan van informatie over DNA sequentie tot gerelateerde ziektes. Daarom achten we het nuttig probleem-specifieke vocabularia te hanteren.

3. Annotaties van genen voorzien in publieke databanken bestaan vaak uit sleutelwoorden, zinnen of paragrafen en komen daarom in aanmerking voor tekstontginning.

4. Interoperabiliteit tussen methoden voor tekstontginning en databanken kan verbeterd worden door het aanwenden van gestructureerde vocabularia en ontologieën, iets waarin sterk geïnvesteerd wordt in de huidige functionele genomica.

5. Het kennisvergaringsproces is typisch cyclisch, dit wil zeggen dat de gebruiker over en weer springt tussen zijn eigen verworven kennis (via expertise of gedane experimenten) enerzijds, en de state-of-the-art vastgelegd in externe bronnen zoals wetenschappelijke abstracts, artikels en gegevensbanken anderzijds.

TXTGate is een stuk web-gebaseerde software [33] dat verschillende literatuurgebaseerde gen indices aanbiedt, gebruikmakende van publieke gegevensbanken. De gebruiker geeft een groep van genen, die bijvoorbeeld voortkomt uit een microrooster analyse, aan het systeem en door middel van gecustomizeerde vocabularia kunnen betrokken MEDLINE abstracts en tekst-gebaseerde databankvelden onderzocht worden op veel voorkomende termen of gen-namen. Er zijn opties voorzien om de gen groep verder te clusteren of om nieuwe queries samen te stellen naar andere gegevensbanken.

Gevallenstudies: belangrijkste resultaten

Aan de hand van gevallenstudies hebben we geïllustreerd hoe TXTGate kan assisteren in het beantwoorden van vraagstukken zoals:

- Welke functionele, of ziekte-termen zijn gerelateerd aan een groep van genen ?

- Welke andere genen zijn in de literatuur gelinkt aan een bepaalde groep genen ?

- Kan een groep van genen verder onderverdeeld worden ?

Het laatste punt is reeds geïllustreerd in vorige sectie waar we tekst-gebaseerde clustering toepasten op cel-cyclus genen en voor meer voorbeelden hiervan verwijzen we naar de volledige doctoraatstekst. Het tweede punt is conceptueel analoog aan het eerste, dat hieronder gedemonstreerd wordt.

In Tabel 0.3 vinden we een vergelijking tussen termen voor cluster e in de tekst analyse van Chaussabel en Cher [60] en de TXTGate termen gevonden door gebruik te maken van het OMIM ziekte-vocabularium. Het microrooster experiment dat door Chaussabel en Cher geanalyseerd werd, heeft betrekking op menselijke immuunresponsen. Wanneer we de twee analyses vergelijken zien we dat de analyse met TXTGate díe ziekten aangeeft die door Chaussabel en Cher enkel gevonden werden door een uitgebreide, manuele analyse. Inderdaad, de termen komende uit hun tekst analyse zijn te algemeen en geven geen indicaties naar ziekten. De flexibiliteit van TXTGate bestaat erin verschillende vocabularia (GO [9], MeSH [22], OMIM [26], eVOC [6] en HUGO-gennamen) aan te bieden als venster naar verschillende bronnen van tekstuele informatie

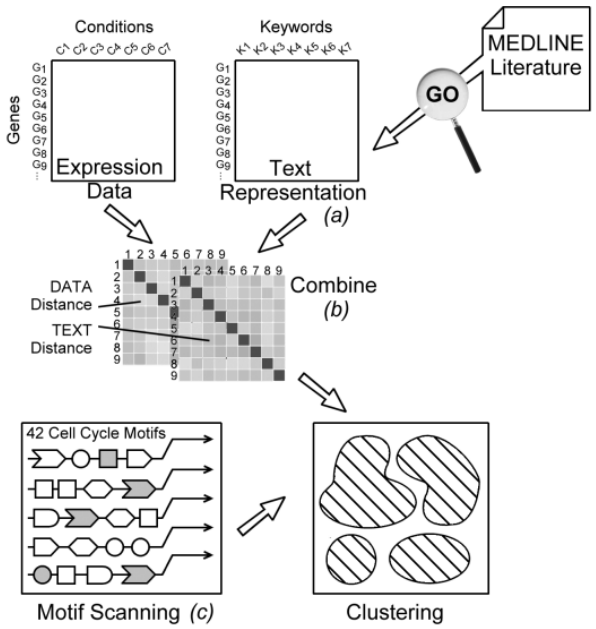
Tabel 0.3: Vergelijking van termen in cluster *e* gevonden door Chaussabel *et al.* en termen gevonden met TXTGate (gebruik makende van de OMIM setting). We vinden referenties naar ziektetermen die Chaussabel *et al.* niet recupereren in hun tekst analyse, maar wel bespreken in hun manuscript nadat ze manueel een aantal gegevensbanken hebben geraadpleegd. TXTGate biedt dus een flexibele manier om snel een idee te krijgen over wat reeds geweten is omtrent een groep van genen. Door verschillende vocabularia te voorzien in de applicatie bieden we daarenboven verschillende standpunten aan van waaruit literatuur informatie kan bekeken worden. In dit geval was men geïnteresseerd in de dimensie ziekte, waardoor we de OMIM-setting gebruikte in onze TXTGate analyse.

Gene Symbol	Cluster Terms in Chaussabel *et al.*	TXTGate terms
Cluster e		
CKB	Population	deaminas
AMPD3	Frequency	**lipoprotein_lipas**
ADA	Allele	creatin
ADORA2A	Unrelated	lipoprotein
ADORA3	Families	**krabb**
P2RX	Recessive	**epidermolysi_bullosa**
P2RX1	Autosomal	**alagil**
P2RX7	Disorder	bear
GEM	Severe	leukodystrophi
ARHH	Patient	receptor
LPL	Deficiency	down
CD36L1		**corneal_dystrophi**
LDLR		**deaf**
BF		hdl
GALC		nucleosid
LAMB3		retinoblastoma
GJB2		junction
TGFBI		adhesion
JAG1		congenit_heart_defect
DSCR1		**hear_loss**

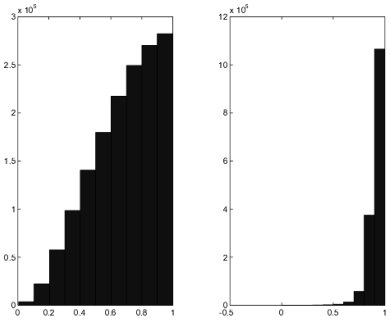
(MEDLINE en verscheidene selecties van velden uit publieke databanken). In de volledige doctoraatstekst worden nog verscheidene voorbeelden, die elk een ander aspect van TXTGate belichten, aangehaald.

Gecombineerde analyse van expressie gegevens en literatuur

De complexiteit van gen expressie data leidt tot een groot aantal mogelijke cluster oplossingen. Daarbij komt het feit dat de gegevens vaak aanzienlijke ruis bevatten. Om op een efficiënte manier genen te groeperen naar gelijkaardige functie kan literatuur een interessant hulpmiddel vormen. Tot hiertoe hebben we tekstuele informatie enkel gebruik *na* een cluster analyse, i.e., om de interpretatie te ondersteunen. In dit hoofdstuk wordt nagegaan hoe een *gelijktijdige* of *geïntegreerde* analyse van expressie en literatuur gegevens een gunstige invloed hebben op de resultaten. De achterliggende veronderstelling is dat een integratie van beide data types in één raamwerk resultaten geeft die niet naar voor komen bij het analyseren van de data types apart. Figuur 0.16 illustreert onze aanpak die we meta-clustering noemen.



Figuur 0.16: Overzicht over het meta-clustering raamwerk. Na tekstuele gen profielen te hebben opgesteld met het vectorruimte model, en gen expressie data te hebben verzameld (**a**), berekenen we alle paarsgewijze afstanden in de twee data ruimten. Expressie- en tekstgebaseerde afstanden worden gecombineerd (**b**), en vervolgens geclusterd. Via de door ons ontwikkelde M-score evalueren we de uitkomst in termen van transcriptionele regulatie (**c**).



Figuur 0.17: Histogrammen van paarsgewijze afstanden tussen genen op basis van microrooster data (links) en literatuur data (rechts). De tekst-gebaseerde afstandsverdeling piekt sneller naar 1 waardoor de expressie data een inherent voordeel krijgt wanneer de afstanden zonder meer bij elkaar worden opgeteld. Dit vormt het argument om de afstanden naar p-waarden om te zetten en dit gebruikmakende van deze empirische verdelingen.

Combinatie van heterogene data

Zoals aangegeven in Figuur 0.16 worden de datatypes gecombineerd via paarsgewijze afstanden. Dit laat toe om in elke data ruimte een gepaste metriek te zoeken. Voor zowel tekst als expressie gegevens hebben we de cosinus-gebaseerde afstand gebruikt. Figuur 0.17 toont de verdelingen van alle paarsgewijze tekst- en expressie-gebaseerde afstanden. We merken hierbij een veel scherpere verdeling naar 1 bij de tekstafstanden. Dit wil zeggen dat bij het samenvoegen van deze afstanden zonder meer, impliciet een voordeel wordt toegekend aan de expressie-data. De schaalfactor is echter niet evident te bepalen.

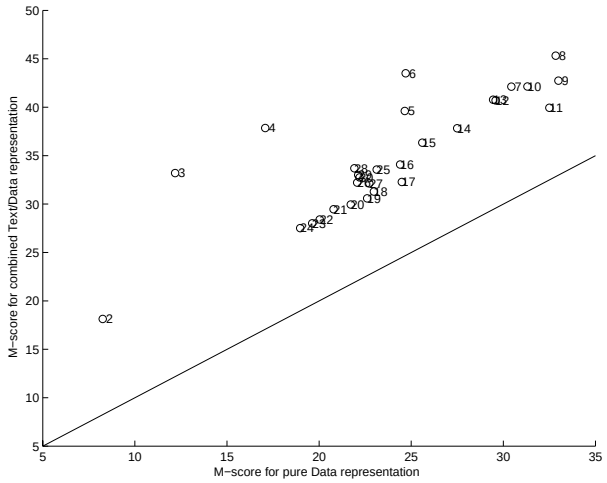
Daarom transformeren we elke afstand naar een p-waarde berekend via de empirische cumulatieve distributie functie (cdf). Vervolgens passen we het idee van de zogenaamde omnibus procedures uit de meta-analyse toe, die speciaal gericht zijn om gegevensanalyse over verschillende experimenten te doen [117]. Meer bepaald gebruiken we Fisher's omnibus methode om p-waarden voor elke paarsgewijze expressie- en tekst-gebaseerde afstand te combineren.

$$\mathcal{S} = -2\log p^{data} - 2\log p^{tekst},$$

volgt een χ^2 verdeling [88] en we gebruiken de cdf-waarden om de gecombineerde afstandsmatrix op te vullen. Zoals aangegeven in Figuur 0.16 gebruiken deze matrix om te de data te clusteren.

Efficiënte validatie

Om het effect van de meta-clustering te onderzoeken ten opzichte van pure expressie clustering, zetten berekenen we de M-score in beide gevallen voor alle waarden van $k = 1 \ldots 30$. In Figuur 0.18 zien we een significante ($\alpha \leq 10^{-3}$) verbetering op van de meta-clustering ten opzichte van de expressie clustering.



Figuur 0.18: M-scores voor hiërarchische clustering voor verschillende besnijdingen k, toegepast op expressie data (x-axis) versus expressie/tekst (y-axis). Beide data types zijn geïntegreerd door berekende p-waarden te combineren via Fisher's omnibus procedure.

Met de stabiliteitsmethode bepalen we de optimale k voor beide data types en vergelijken beide oplossingen in een meer diepgaand biologisch onderzoek, analoog aan de gevalstudie bij de M-score en gebruikmakende van de TXTGate web applicatie (zie hierboven). Hieruit destilleren we volgende conclusies:

- Integratie is in staat om sterk relevante patronen te behouden.
- Het toevoegen van literatuur informatie versterkt functionele verbanden tussen genen.
- Het toevoegen van literatuur leidt tot een reorganizatie van genen met gelijkaardige patronen en gelijkaardige functie in eenzelfde cluster.
- De integratie hergroepeert genen met een lage expressie correlatie, maar een hoge functionele correlatie.
- De geïntegreerde analyse ontdekt nieuwe motief patronen die bevestigd zijn door andere studies.

Hiermee hebben we aangetoond hoe ons meta-clustering raamwerk de premisse ondersteunt dat combinatie van heterogene data kan helpen bij het aanpakken van complexe problemen zoals exploratorische analyse van microrooster data.

Besluit

Als conclusie van dit eindwerk sommen we karakteristieken van data in de functionele genomica op, alsook welke bijdrage wij geleverd hebben.

1. Experimenten zoals microroosters produceren grootschalige data en dit aan een alsmaar groeiend tempo. Om de gelijkvormigheid, de reproducibiliteit en de bruikbaarheid van deze data te verhogen hebben we bijgedragen aan de ontwikkeling van MIAME [54], een standaard voor de rapportering van microrooster experimenten. Om zulke informatie te leveren vanuit het laboratorium is een doordacht informatie systeem nodig. Wij hebben de eerste versie van zulk een LIMS ontwikkeld aan de VIB Microarray Facility[10].

2. De complexiteit van functionele genomica, alsook experimentele imperfecties nopen het bijeenbrengen van verschillende informatiebronnen. In ons werk hebben we een strategie voor tekstontginning uitgewerkt toegespitst op dit soort problemen. TXTGate is een flexibel platform dat de voorgestelde ideeën bruikbaar maakt voor praktische vraagstukken zoals 'Welke sleuteltermen (functie of ziekte) zijn belangrijk bij een groep van genen ?', 'Hoe kan ik een groep van genen onderverdelen op basis van wat in de literatuur beschreven is ?', of 'Welke andere genen heb ik over het hoofd gezien ?'. Eveneens hebben we een meta-clustering aanpak ontwikkeld die het mogelijk maakt tekstuele informatie diep te integreren in de analyse van microrooster data.

3. Om het methode-gedreven onderzoek te ondersteunen en de validatie van verscheidene clustering uitkomsten te vereenvoudigen zijn bevattelijke scores nodig. Gedetailleerde biologische validaties zijn arbeidsintensief en zijn daarom ongeschikt als men een eerste indruk wil krijgen van een bepaald resultaat. Vooral in clustering, waar veel uitkomsten per se mogelijk zijn, zijn manieren om biologische relevantie te testen belangrijk. Onze voorgestelde M-score geeft een idee van significante aanrijking aan motief patronen in een geclusterde data set. Ze bouwt voort op een aantal eenvoudige veronderstellingen, werd uitvoerig getest en efficiënt aangewend bij het onderzoek naar de biologische relevantie van de meta-clustering.

Tot slot biedt dit werk verschillende perspectieven naar opvolgingen. Zo is het inbrengen van semantiek in ons 'blind' vectorruimte-model een aantrekkelijke piste, gegeven de investeringen die heden gepompt worden in projecten zoals bijvoorbeeld Gene Ontology [9][106]. Tot slot kan heterogene data ook op andere matrix-algebraïsche manieren gecombineerd worden zoals bijvoorbeeld in [102], [75] en [138].

[10] www.microarrays.be

Contents

Voorwoord **i**

Abstract **iii**

Nederlandstalige abstract **v**

Nederlandse samenvatting **vii**

Notation **xliii**

Publication list **xlvii**

1 Introduction **1**
1.1 Towards integrative data mining in bioinformatics 1
1.2 Large scale experiments in present-day molecular biology 2
1.3 Microarrays and databases 4
1.4 Microarray data analysis 6
1.5 Two visions on biological meaning 6
1.6 Cyclic nature of the knowledge discovery process 8
1.7 Text mining and functional genomics 8
1.8 Integrated data mining: expression, motifs and literature 11
1.9 Data sets 11
1.9.1 Yeast cell cycle gene expression data 11
1.9.2 External databases 13
1.9.3 Biological evaluations 15
1.10 Dissertation structure 15

2 Large scale gene expression experiments **19**
2.1 Concepts of molecular cell biology 21
2.1.1 DNA 21
2.1.2 Genes 21
2.1.3 Motifs 22
2.1.4 Cellular organization 22
2.2 Microarray technology 23
2.2.1 Technology 23
2.2.2 Gene expression matrix 26
2.2.3 Sources of noise 26
2.3 Microarrays and database systems 28
2.3.1 LIMS 28
2.3.2 MIAME 29
2.4 Discussion 34

3 Exploratory data analysis **37**
3.1 Introduction . . . 37
3.2 General clustering methodology . . . 39
3.3 Prevalent approaches to clustering microarray data . . . 41
3.3.1 Preprocessing issues . . . 41
3.3.2 Distance metrics . . . 42
3.3.3 Hierarchical methods . . . 43
3.3.4 Non-hierarchical methods . . . 46
3.4 Statistical cluster validity . . . 46
3.4.1 *Silhouette* coefficient . . . 47
3.4.2 k-NN learnability . . . 47
3.4.3 *Jaccard* coefficient . . . 48
3.4.4 *Rand* index . . . 49
3.5 The optimal number of clusters . . . 49
3.5.1 Stability based method to determine optimal k . . . 49
3.6 Iterative clustering . . . 51
3.7 Discussion . . . 54

4 Biological cluster validation **55**
4.1 Gene ontology . . . 55
4.1.1 Description . . . 55
4.1.2 Assessment of functional enrichment in gene groups . . . 57
4.1.3 Analysis of entire clustering solutions using GO . . . 58
4.2 Upstream regulatory analysis . . . 61
4.2.1 Motif-based score to evaluate various clustering runs . . . 61
4.2.2 M-score . . . 63
4.2.3 Results . . . 66
4.3 Discussion . . . 76

5 Text-based gene analysis **81**
5.1 Vector space model . . . 81
5.1.1 Document processing . . . 81
5.1.2 Vector representations . . . 83
5.1.3 Matrix formulations . . . 85
5.1.4 Document similarity . . . 85
5.1.5 Construction of a gene index . . . 86
5.2 Scoring groups of genes based on text . . . 86
5.2.1 Scoring method . . . 86
5.2.2 Text-based scoring to identify functional categories . . . 88
5.2.3 Text-based scoring to compare classical vector space representations . . . 90
5.2.4 Using text to score clusters from expression ana-lysis . . . 94
5.2.5 Conclusion . . . 96
5.3 Clustering genes based on text . . . 96
5.3.1 Performance of text clustering . . . 96
5.3.2 Reconstruction of functional segregation using text . . . 103
5.3.3 Conclusion . . . 104
5.4 Discussion . . . 106

6 TXTGate, a platform for profiling genes based on text **107**
6.1 Research context . . . 108
6.2 System design . . . 111
6.2.1 Features . . . 111
6.2.2 Literature indices in TXTGate . . . 111
6.3 Related software . . . 116
6.4 Experiments . . . 117
6.5 Study 1: *Saccharomyces cerevisiae* . . . 117
6.5.1 Text profiling of yeast expression clusters . . . 118
6.5.2 Text clustering and the detection of subclusters . . . 120

6.5.3 Random, corrupted and ambiguous clusters 123
6.5.4 Condition-specific yeast networks 130
6.5.5 Co-linkage analysis 133
6.6 Study 2: Human data 139
6.6.1 Subclustering human data 139
6.6.2 Textual information through the eyes of different vocabularies 140
6.7 Discussion 144

7 Integration of text and expression data **147**
7.1 Gene expression, keywords and motifs 149
7.1.1 Overview of the approach 149
7.1.2 Experimental setting 149
7.2 Mixing heterogeneous data via distance matrices 151
7.2.1 Linear combination of distance matrices 151
7.2.2 Fisher's omnibus to combine evidence 154
7.3 Motif scoring as independent evaluation 154
7.4 Analysis of expression clusters 156
7.5 Analysis of integrated clusters 160
7.5.1 Correspondence analysis of both clusterings 160
7.5.2 Loss of the SCB motif 165
7.5.3 Emergence of the Cb1fp/Met31_32p/Pho4/STRE motif pattern 168
7.6 Discussion 173

8 Conclusions **179**
8.1 Accomplishments 179
8.2 Discussion and Future directions 180
8.2.1 Semantically-oriented text representations 180
8.2.2 Advanced integration of information 181

A Overview of external databases **183**

B Supplementary text profiles **187**

Bibliography **183**

Notation

Ace Angiotensin-Converting Enzyme

BOOL Boolean

Cbf1p Centromere Binding Factor 1

CCA Canonical Correlation Analysis

cDNA Complementary DNA

CDS CoDing Sequence

DNA DeoxyriboNucleic Acid

ECB Early Cell cycle Box

EST Expressed Sequence Tag

eVOC [see Appendix A]

Fkh1 ForKHead transcription factor

G1 Gap 1

G2 Gap 2

GeneCards [see Appendix A]

GO Gene Ontology; [see Appendix A]

GUI Graphical User Inteface

HUGO HUman Genome Organization; [see Appendix A]

ICRE Inositol/Choline-Responsive Element

IDF Inverse Document Frequency

LIMS Laboratory Information Management System

ln natural logarithm

LocusLink [see Appendix A]

log base 10 logarithm

LSI Latent Semantic Indexing

M Mitosis

MAGE-ML MicroArray Gene Expression Markup Language

MCB Mlu I Cell cycle Box

MeSH Medical Subject Headings

Met31_32p METhionine-related factors

MGED Microarray Gene Expression Data society

MIAME Minimum Information About a Microarray Experiment

MIPS Munich Information center for Protein Sequences

ML MEDLINE; [see Appendix A]

mRNA messenger RiboNucleic Acid

OMIM Online Mendelian Inheritance in Man; [see Appendix A]

Ndd1 Nuclear Division Defective transcription factor

PCA Principal Component Analysis

Pho4 Phosphate responsive transcriptional activator

PMID PubMed identifier

PRE Pheromone Response Element

Rap1 mammalian telomere protein Repression and Activation Protein

TCA cycle TriCarboxylic Acid cycle

TCS TriChoSanthin

TF Transcription Factor

TF Term Frequency

TRANSFAC [see Appendix A]

S Synthesis

SCB Swi4-dependent Cell cycle Box

STRE Stress

SVD Singular Value Decomposition

SP SWISSPROT ; [see Appendix A]

UMLS Unified Medical Language System ; [see Appendix A]

UTR UnTranslated Region

VIB Vlaams Interuniversitair instituut voor Biotechnologie

YC YeastCard

Publication list

International journal papers

- A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C. Holstege, I.F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME): towards standards for microarray data. *Nat Genet*, 29:365–371, 2001.

- B. Coessens, G. Thijs, S. Aerts, K. Marchal, F. De Smet, K. Engelen, P. Glenisson, Y. Moreau, J. Mathys and B. De Moor. INCLUSive : a web portal and service registry for microarray and regulatory sequence analysis.*Nucleic Acids Res*, 31(13):3468-3470, 2003.

- P. Glenisson, J. Mathys, Y. Moreau, and B. De Moor. Meta-clustering of gene expression data and literature-extracted information. *SIGKDD Explorations, Special Issue on Microarray Data Mining*, 5(2):101–112, 2003.

- P. Glenisson, B. Coessens, S. Van Vooren, J. Mathys, Y. Moreau, and B. De Moor. TXTGate : Profiling gene groups with text-based information. *Genome Biol*, 2004. In press.

International conference papers

- P. Antal, P. Glenisson, G. Fannes, J. Mathys, B. De Moor B, and Y. Moreau. On the potential of domain literature for clustering and bayesian network learning. In *Proc of the 8th ACM-SIGKDD Int Conf on Knowl Discovery and Data mining (ACM-SIGKDD), Edmonton, Canada*, pages 405–414, 2002.

- P. Glenisson, P. Antal, J. Mathys, Y. Moreau, and B. De Moor. Evaluation of the vector space representation in text-based gene clustering. *Pac Symp Biocomput (PSB), Kauai, Hawai*, pages 391–402, 2003.

- P. Glenisson, B. Coessens, S. Van Vooren, Y. Moreau, and B. De Moor. Text-based gene profiling with domain-specific views. In *Proc. of the First Int Workshop on Semantic Web and Databases (SWDB 2003), Berlin, Germany*, pages 15–31, 2003.
- S. Aerts, B. Coessens, P. Glenisson, D. Lambrechts, B. De Moor. In silico validation of cis-regulatory modules and putative target genes using multiple genomic information sources, Submitted.

Chapter 1

Introduction

1.1 Towards integrative data mining in bioinformatics

The range of interest of molecular biologists has been shifting over the last ten years from a single gene to the genome[1] level. In a visionary article published on the dawn of the appearance of the first complete draft the Human Genome in 2001 [66][158], Marc Vidal advocates the necessity of a biological atlas of functional maps in this 'post-genome' era [159]. For a long time biologists were forced to adopt a reductionist approach, studying single genes at a time – perhaps best paraphrased by the ongoing proverb 'One gene, one PhD'. Public availability of the human genome and concurrent biotechnological advancements in high-throughput screening sparked a phase transition in the scientific approach to biology and biomedicine. The growing availability of data, not only on DNA or protein sequence, but also on protein structure, gene expression, genetic disease,... enables the formulation of more integrative questions aimed at the recognition of molecular functional modules as building blocks to cellular organization. Such questions demand the import of a new vocabulary into the biological sciences including concepts as 'robustness', 'significance', 'coincidence detection', 'error correction', 'adaptation' and 'conditional probability' – to name a few [87].

Integration of information from databases and large scale assays invoke data sets that are typically heterogeneous, exhibit a high dimensionality (both in number of measurements and in number of features) and are inextricably connected to expert knowledge (e.g., curator annotations[2] , MEDLINE abstracts,...). Integrative data mining thus enables a synthesis, rather than a reductionist analysis, of large scale surveys of genomic data and is geared at identifying patterns (i.e., a subset of variables, features, or both) of biological

[1]The entire DNA or 'book of life' of an organism.

[2]Annotations are concise descriptions attached to a biological object, and usually appear in biological database.

interest. Through an efficient merging of heterogeneous data into a common framework, integrated data mining will be one of the crucial components in constructing what Vidal refers to as an 'atlas' of biological function.

Key features of large scale biological data to be accounted for in intelligent data analysis include the following (see also Altman *et al.* [39], De Moor *et al.* [116]):

- High throughput data is usually collected with a low signal-to-noise ratio, requiring robust methods;
- Biology is still relatively knowledge rich and data poor, which demands complex representations;
- Biological data sources operate at multiple scales and are tightly linked;
- Biological knowledge is distributed, requiring interchangeable data formats, efficient cross linking and, in the end, incorporation of multiple knowledge sources.

1.2 Large scale experiments in present-day molecular biology

A human body for example is made up of trillions of cells, the nucleus of which contains identical complements of chromosomes. Each chromosome is a long strand of DNA (DesoxyriboNucleic Acid), which in its turn is a double helix structure composed of four molecules (bases) called A, C, T and G. A gene is a functional region within the DNA, whereas the genome refers to the full sequence of base pairs of an organism. Recent technological advances enable the collection of many different types of data at a genome-wide scale, including DNA sequences and so-called gene expression measurements.

Roughly speaking, a gene is said to be expressed in a certain cell when it is used to make proteins, which are part of the essential molecular machinery of the cell. Large scale gene expression experiments, often called microarrays, permit the measurement of the expression levels of thousands of genes simultaneously over various conditions. The result of a microarray experiment usually results in a gene expression matrix (shown in Figure 1.2), where each row contains measurements about the expression level of a gene under various conditions.

Microarrays are representative examples of experimental methods that are large in scope, open to multiple statistical approaches, but still hypothesis-driven. We distinguish two major application directions for microarrays:

1. Fundamental biology and biomedicine, where scientists are looking for gene regulation mechanisms in biological processes or trying to identify disease-related genes and their orchestration.
2. Applied biomedicine and pharmacology where microarrays open the path to genetically customized patient diagnoses and improved drug design through screening the effects of various drug treatments.

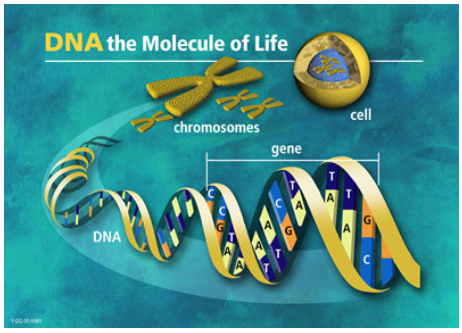


Figure 1.1: Each chromosome in a cell is a long strand of DNA (DesoxyriboNucleic Acid), which in its turn is a double helix structure composed of four molecules (bases) called A, C, T and G. The genome or 'book of life' of an organism refers to the full sequence of base pairs. (Figure courtesy Human Genome Project information, http://DOEgenomes.org)

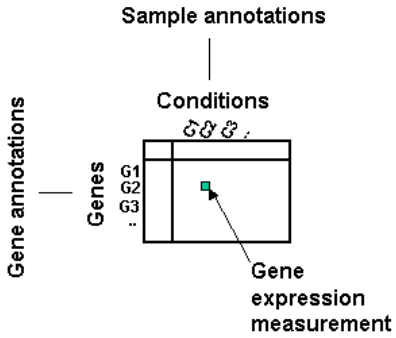


Figure 1.2: Gene expression matrix resulting from a microarray experiment. Each position in the matrix contains information characterizing the expression level of the particular gene in a particular sample (condition). An example of an microarray experiment is the simultaneous monitoring of the expression of thousands of genes in a cancer cell. Correct and detailed information about both genes and samples is crucial to both reproducibility and usefulness of a microarray experiment.

Addressing the first type of questions, 'functional genomics' is a discipline that aims at characterizing the function of genes along with the role they play in a given pathway. Our presented research will be mainly concerned with this kind of problems.

1.3 Microarrays and databases

As a single microarray experiment typically scans the behavior of thousands of genes over dozens of conditions, archiving and retrieving all supporting data is far from trivial. Indeed, although the gene expression matrix provides a conceptual view of a microarray experiment, it is the end product of a complex chain of biotechnological operations. One could ask what information needs to be reported about an experiment for optimal clarity.

Firstly, details about each studied condition (i.e., one column the expression matrix in Figure 1.2) needs to be accurately stored. When looking, for example, at the expression of genes in mice with one deficient gene, procedures on how mRNA (see Chapter 2) is extracted from the animal's tissue is important for interpretation and reproduction by third parties. Secondly, information about each element on the array and how it arrived there (i.e., one row in Figure 1.2) is crucial to know which genes are exactly under study. Finally, the measurements contained in each element of the gene expression matrix are a result of non-trivial image segmentation and intensity quantifications and many sources of noise can be identified (see Chapter 2).

Although many groups have organized their microarray data for a long time in some kind of file-based system (e.g., using spreadsheets), the large scale nature of the data made this practice quickly untenable for its limitations in terms of search and upkeep. Laboratory Information Management Systems (or LIMS) provide a solid alternative and are used to store, manage and update a wide range of protocols, procedures and settings that precede the actual gene expression experiment. On the other hand, all information surrounding the gene expression matrix in Figure 1.2), being gene annotation, sample annotation and expression measurements, as well as further analyses (e.g., clustering - see below) are typically recorded in an in-house results database. However, it is not immediately clear what the necessary level of detail is to reproduce a microarray experiment. Neither is it straightforward to see what reported information is needed to make experiments more valuable for various kinds of data mining. Such considerations are crucial in the establishment of expression compendia, where lots of heterogeneous microarray experiments are bundled into a public database. MIAME aims at solving these issues. In our PhD we have contributed to the development of MIAME [54], a proposed standard for reporting about gene expression experiments. As illustrated in Figure 1.3, MIAME lies at the intersection of LIMS and private result databases. Through data formats that implement the standard (MAGE-ML [140]), exchanges between such, and other, databases are made possible.

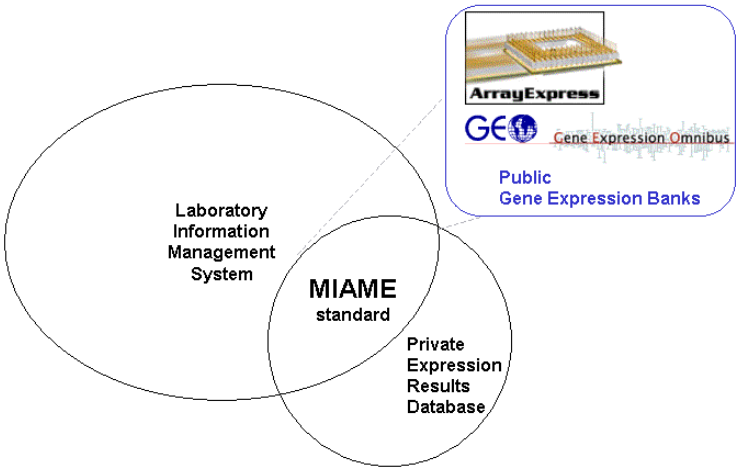


Figure 1.3: Laboratory Information Management Systems (LIMS) support the experimentalist by allowing a systematic registration of each step preliminary to the actual microarray gene expression experiment. All the information related *directly* the gene expression measurements is specified in a standard called MIAME. This information, jointly with supplementary data such as results from subsequent expression analyses, are typically stored in private 'result databases'. These issues are treated in Chapter 2 where we discuss the LIMS we developed at the VIB microarray Facility and where we provide more details on the MIAME standard to which we also contributed.

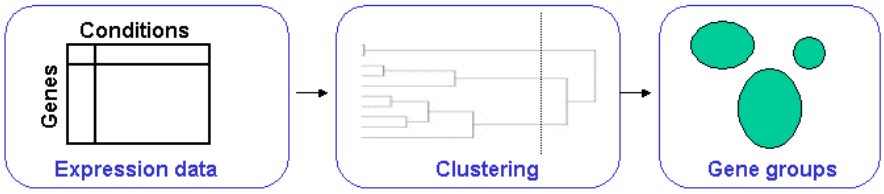


Figure 1.4: Clustering gene expression data. To organize expression profiles of thousands of genes over dozens of conditions, genes are placed into more tractable groups by clustering. The output of a so-called hierarchical clustering algorithm is typically a tree. The number of clusters for example is a typical parameter that needs to be supplied by the user beforehand and has a direct relation to the position where the tree has to be cut (dashed vertical line).

1.4 Microarray data analysis

As microarrays screen thousands of genes simultaneously, finding gene groups that are interesting is a tedious, if not impossible, task without the help of algorithms. A technique that is often relied upon to organize microarray gene expression data is clustering (Figure 1.4). With the aim of partitioning the data in a more tractable way, genes are divided in groups according to their level of mutual co-expression. The underlying assumption adopted in functional genomics is that similarly expressed genes also share a common function.

Popular clustering algorithms include K-means and hierarchical clustering, which are explained in Chapter 3. Clustering is an unsupervised technique meaning that there is no guidance to arrive to the best solution. As a result many plausible outcomes are possible. The number of clusters, for example, is a typical parameter that needs to be supplied by the user beforehand. We present an existing technique based on subsampling to tune this parameter in order to get the most mathematically stable solution. The large dimensionality of microarray data as well as its noisy nature made us modify the technique so to accommodate our research problems (Chapter 3).

1.5 Two visions on biological meaning

As gene expression clustering typically outputs lists of dozens or hundreds of genes that are co-expressed, typically one is faced with the assignment of relevance to those results. Throughout this dissertation we are interested in two aspects of 'biological meaning' (see also Figure 1.5); Firstly, we wish to *functionally* characterize genes, i.e, finding out whether a group of genes share a common molecular function or whether they are active within the same location of the cell. The Gene Ontology (GO) offers a taxonomy of important concepts (terms) in functional genomics and associates genes to nodes in this taxonomy. Hence, it has become a primary source for gene annotation. However, sources such as

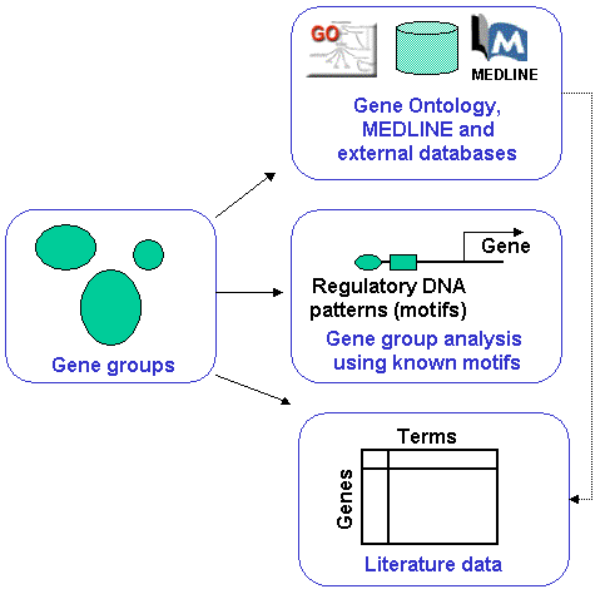


Figure 1.5: Experimentally defined groups of genes are approached in two ways in this work. Firstly, functional relatedness can be corroborated through an interactive use of external information sources including Gene Ontology, scientific MEDLINE abstracts or other specialized databases. Our application TXTGate (Chapter 6), for instance, provides an elegant, interactive framework that links a variety of external sources together through a text mining approach. The adopted representation places genes in a term vector space, thus giving rise to a gene-by-term matrix (dotted arrow). The algorithms and scores applied to this matrix are summarized in Figure 1.7. Secondly, we are interested in extracting evidence for *transcriptional regulation* in clusters of genes. We have developed a motif-based score (M-score) that assesses biological relevance of an entire clustering solution in terms of DNA patterns that regulate the expression (or more precisely the 'transcription') of the genes.

.

GO are hand-crafted and may not be comprehensive (for example, no disease information is present in GO) or up-to-date at any given time. Scientific literature on the other hand contains a wealth of information and is therefore also a crucial source of information. Text mining exploits this contention and aims at processing information that is encoded as unstructured, free text. The use of text mining in biology plays a central role in this dissertation.

Secondly, we are interested in extracting evidence for *transcriptional regulation* from clusters of co-expressed genes. We propose a novel motif-based score (M-score) that provides a quick estimate of the biological relevance of multiple clustering solutions in terms of DNA patterns that regulate the expression (or more precisely the 'transcription') of the genes. In Chapter 4 we demonstrate its behavior both on synthetic and real data and it is used as an independent control in Chapter 7 to assess the biological relevance of *fusing* expression data with literature information.

1.6 Cyclic nature of the knowledge discovery process

The process of successfully gaining insight into complex genetic mechanisms will increasingly depend on a complementary use of a variety of resources, including biological databases and specialized literature on the one hand, and the experimentalist's knowledge on the other hand. The knowledge discovery process could therefore be considered as cyclic, (i.e., requiring several iterations between heterogeneous information sources to extract a reliable hypothesis). For example, to date, linking up experimentally derived gene groups (from cluster analysis) to the existing databases and published literature still requires numerous queries and extensive user intervention. This process of drilling down into the entries of hundreds of genes is notably inefficient and requires higher-level views that can more easily be captured by a (non-)expert's mind. Figure 1.6 shows how a gene group derived from expression data ('Data World') is scanned for existing information ('Text World') to yield an improved characterization and understanding. The TXTGate application presented in Chapter 6 is an instantiation of such cyclic analysis. It is a text mining framework that involves scientific literature as well as various public databases in the functional interpretation of sets of genes (see also Figure 1.5).

1.7 Text mining and functional genomics

Rapidly recognizing whether a set of genes share a common function is useful for assessing its biological significance and for prioritizing on those sets that deserve follow-up. The interpretation of putative gene relations from expression data is a labor-intensive task due to the number of genes on an array and the complexity of how they might be related. Conversely, due to its increasing availability in computer-parsable format, published scientific literature that describe genes

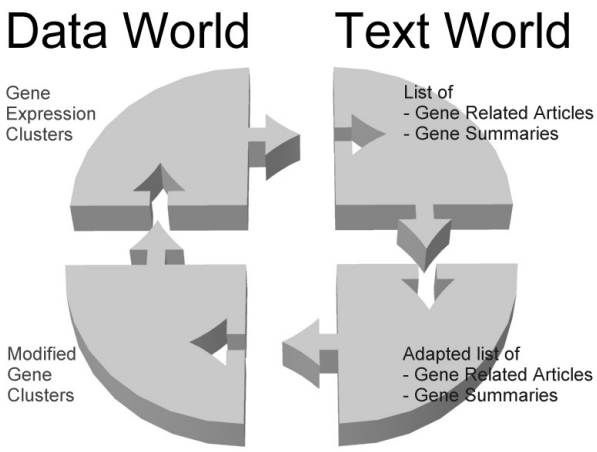


Figure 1.6: Cyclic nature of the knowledge discovery process. It shows a high-level view of how this is embodied in microarray cluster analysis: starting from a cluster of genes resulting from a gene expression analysis (the 'Data World'), the corresponding literature profiles are queried and analyzed (the 'Text World'), resulting in either the addition of extra genes of interest or the omission of irrelevant genes. This updated cluster can subsequently be reanalyzed in expression space, which concludes a first cycle (Figure courtesy: Steven Van Vooren).

and their functions constitutes a potentially important source of information to assist in the data mining process [110]. Particular motivations to include literature-based information in expression analysis include:

- Functionally related genes may not be grouped together due to experimental noise or clustering parameters used (e.g., distance metric, number of clusters);
- Simultaneously expressed genes do not necessarily all share the same biological function. One is typically interested in genes that are co-expressed and functionally related in the context of the microarray experiment. Conversely, genes that are expressed at different times, and therefore grouped differently, may still have some function in common;
- When expression and function correlate well, the underlying reason is not always apparent. Proper justification requires extensive additional database or literature searches.

Figure 1.7 depicts how we cluster genes according to their reported functions in literature and quantify the functional coherence of gene groups measured in term space. Various aspects, applications and original contributions of automated text analysis, overviewed more in detail below, are presented in Chapters 5, 6 and 7.

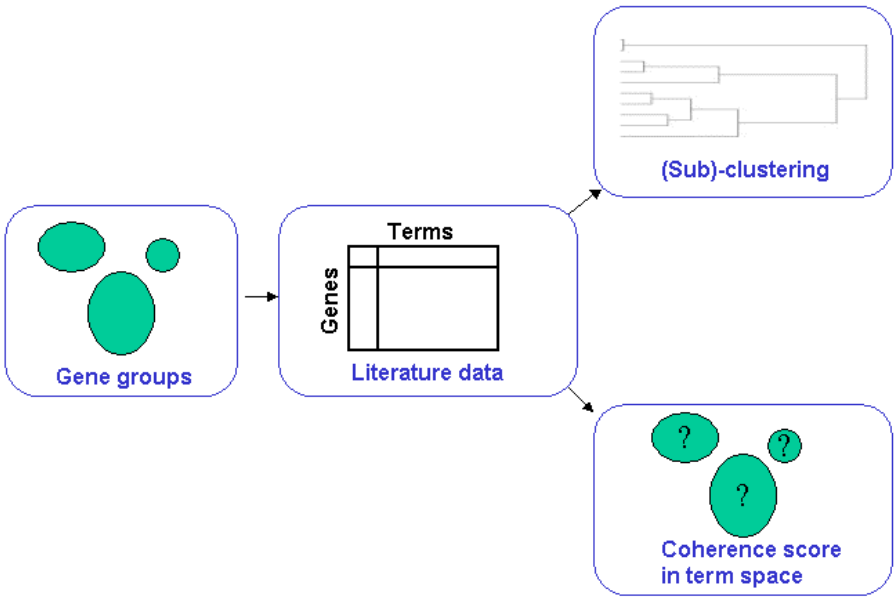


Figure 1.7: Analyzing experimentally established groups of genes from the perspective of prior work in the field is a challenging task. A text mining approach is used to represent a variety of literature information in a gene-by-term matrix. Genes can be analyzed in the 'Text world' (see also Figure 1.6) by (a) clustering genes according to their reported functions in literature, or by (b) quantifying the tightness -and hence functional relatedness- of gene groups in term space.

1.8 Integrated data mining: expression, motifs and literature

To date, clustering methods that include no background information remain the most popular ones. The scheme in Figure 1.7 where groups are processed in text space using a measure of coherence, or through, clustering are examples of how literature is implied 'late' in the analysis process, i.e., it is used as an aid to either (a) interpret cluster structures discovered in expression data, or to (b) discover novel substructures in an already defined (expression) cluster. However, this assumes the existence of a good clustering solution stemming from an appropriate algorithm that is applied to reliable expression data. Motivated by the noisy characteristic of post-genome data (summed up in Section 1.1) and the additional arguments presented in previous section, we have investigated how to *deeply integrate* expression data with literature information. The underlying idea is that as the sources of noise in the two data structures are unlikely to be similar, a proper *joint* analysis of text and data can improve the overall quality of the results. Therefore, as visualized in Figure 1.8, we developed a method to cluster genes based on both text and data at the same time and we validated this approach in terms of transcriptional regulation with our developed M-score.

1.9 Data sets

Various data sets are used throughout this thesis. The lack of benchmarks in functional genomics make 'method-driven' research[3] a challenging task. Hence, one needs to construct cases powerful enough to demonstrate the approach, yet simple enough to be understandable by a non-expert. Indeed, how to corroborate improvements on exploratory techniques such as clustering whose purposes are, by definition, loosely defined ? The investigation into new approaches to cluster expression data is a good example illustrating this problem. The lack of benchmarks is especially true for topics such as text mining, which still had an 'exotic' status around the start of the PhD (2000), but which came more to the forefront of bioinformatics over the past two years (2002-2004). Finally, the fast pace at which databases and their updates keep evolving pose additional problems and undermine the reproducibility of experiments. To accommodate for this we provide here a detailed overview of the data sets used. Explanation of some of the used terminology is postponed to the relevant positions in the thesis.

1.9.1 Yeast cell cycle gene expression data

The following description is taken from [4]. The cell cycle (Figure 1.9) is an ordered set of events, culminating in cell growth and division into two daughter

[3] In contrast to problem-driven research, which tries to find *a* solution to a given biological problem, method-driven research reflects the curiosity to see whether a certain algorithm, method or program can improve on certain biological problems.

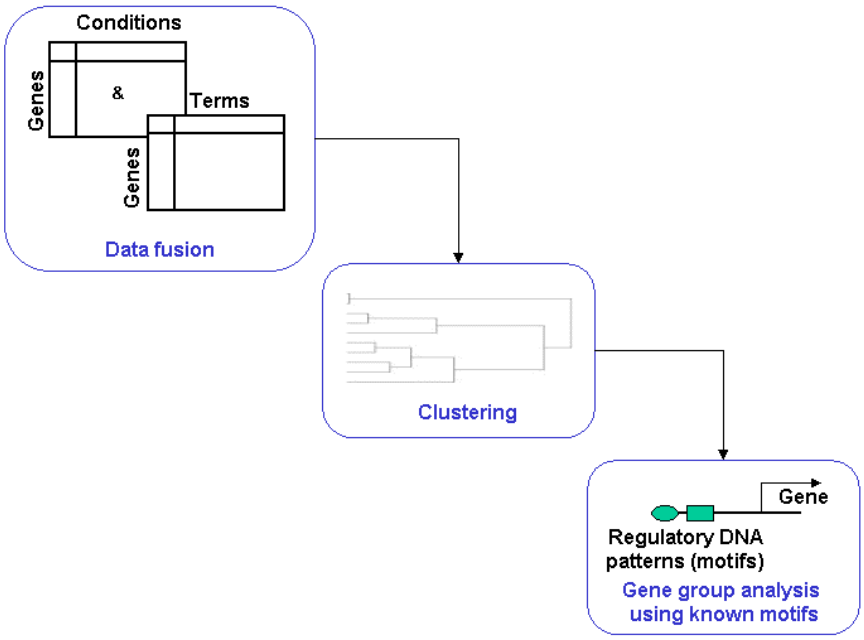


Figure 1.8: A combined analysis data sources can yield more information than using them separately. In Chapter 7 we show how an integrated analysis of expression and text improves the clustering outcome in terms of significant motifs found in the resulting clusters using our M-score.

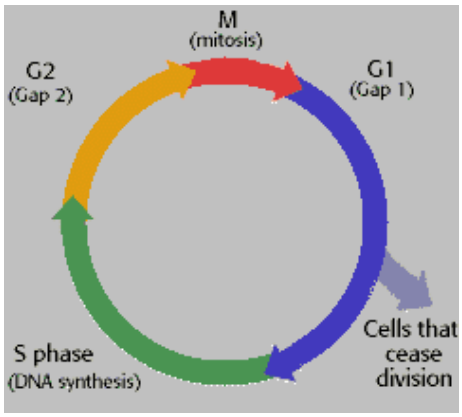


Figure 1.9: The cell cycle is an ordered set of events, culminating in cell growth and division into two daughter cells. The stages are G1, S, G2 and M. The G1 stage stands for 'GAP 1'. The S stage stands for 'Synthesis'. This is the stage when DNA replication occurs. The G2 stage stands for 'GAP 2'. The M stage stands for 'mitosis', and is when nuclear (chromosomes separate) and cytoplasmic (cytokinesis) division occur. (Figure taken from [4])

cells. Non-dividing cells are not considered to be in the cell cycle. The stages are G1, S, G2 and M. The G1 stage stands for 'GAP 1'. The S stage stands for 'Synthesis'. This is the stage when DNA replication occurs. The G2 stage stands for 'GAP 2'. The M stage stands for 'mitosis', and is when nuclear (chromosomes separate) and cytoplasmic (cytokinesis) division occur. How cell division (and thus tissue growth) is controlled is very complex. Cancer, for example, is a disease where regulation of the cell cycle goes awry and normal cell growth and behavior is lost.

Yeast (*Saccharomyces cerevisiae*) is a model organism in which the cell cycle is heavily studied [141]. As a result the first gene expression studies were conducted on the yeast cell cycle. However, the various screenings use different microarray platforms(cDNA versus Affymetrix), preprocessing and data analysis techniques (gene filtering, clustering,...). Moreover, data from early studies is often included in some way in the later work. Finally, various data sets are re-analyzed with our methods.

1.9.2 External databases

Databases play a crucial role in the storage, retrieval and dissemination of all kinds of biological knowledge. Throughout this thesis, information from a variety of databases will be used. Concise descriptions of all these databases (and their abbreviations) can be consulted in the Appendix.

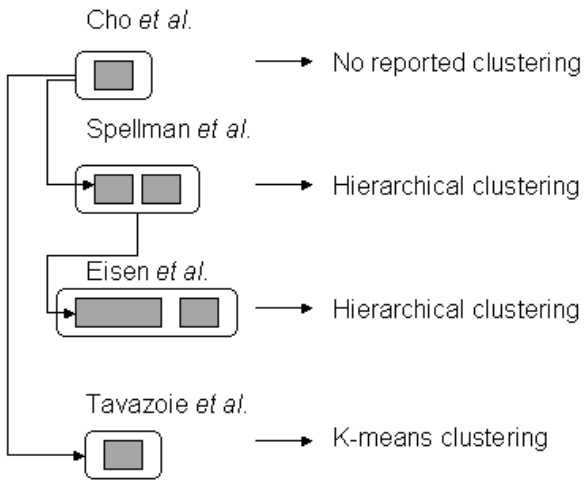


Figure 1.10: Overview of how various public yeast expression data results used in this thesis are interrelated. Cho *et al.* [62] was the first yeast cell cycle publication and was conducted on the Affymetrix platform. The published data accompanying the publication are rather unclear. Spellman *et al.* [141] processed the expression values of the Cho again and appended it to their expression analysis, conducted on a cDNA platform. Tavazoie *et al.* [148] did an *ab initio* re-analysis from the Cho data. Contrary to the other work, they used K-means clustering. Finally Eisen *et al.* [73] appended other yeast expression data (not necessarily cell cycle related) with data from Spellman *et al.* Throughout this thesis we will indicate accordingly (a) which data we use as well as (b) whether we used the reported clustering results or we did a new clustering analysis.

Table 1.1: Overview of the used data sets in this work. We use with various types of gene groups to test or demonstrate our methods.

Source	version	size	Description
Gene Ontology	Jan 2003	906	Genes from diverse GO catogories not size 200
Gene Ontology	Nov 2003	27	Autophagy- related genes
MIPS	Jul 2002	116	Three distinct functional groups: Translational control, autophagy and amino acid transport
OMIM	Sept 2003	25	All genes related to colon cancer
Spellman*et al.*[141]	n.a.	126	Diverse cell cycle induced genes
Spellman *et al.*[141]	n.a.	1745	All significantly expressed genes expression in yeast cell cycle microarray experiment
Ihmels *et al.*[95]	n.a.	20	Genes from two discussed modules in the yeast TCA pathway
Chaussabel *et al.* [60]	n.a.	60	Discussed gene modules related to human bacterial infection

1.9.3 Biological evaluations

The lack of benchmarks and gold standards at the time the presented research was carried out, made us take on a hybrid approach in which quantitative analyses are supported by qualitative evaluations. Therefore, the principal results in this thesis are always complemented by a biological discussion. The early work (Chapters 4 and 5) was done in close collaboration with a post-doctoral yeast biologist. In some of the later work on TXTGate (Chapter 6) and integrated clustering (Chapter 7), we performed the biological validation ourselves by combining accumulated experience with the consultation of biological databases.

1.10 Dissertation structure

Several aspects of microarray data, microarray data mining and biological text mining will be treated in this dissertation. Figure 1.11 brings all the previous parts together. Full arrows designate a directed propagation of information throughout the analysis process. For example: a matrix is submitted to a clustering algorithm, or a set of genes is passed as a query to a data structure. Dashed arrows on the other hand designate a more general connection. For example: a database has been used to generate the term matrix. Each of the components were introduced earlier in this chapter and will be developed more in detail throughout the remainder of the dissertation.

The first phase of our research was conducted at the MicroArray Facility of the VIB[4], an umbrella institute of (at that time) nine prominent biotech re-

[4]Vlaams Interuniversitair Instituut voor Biotechnologie.

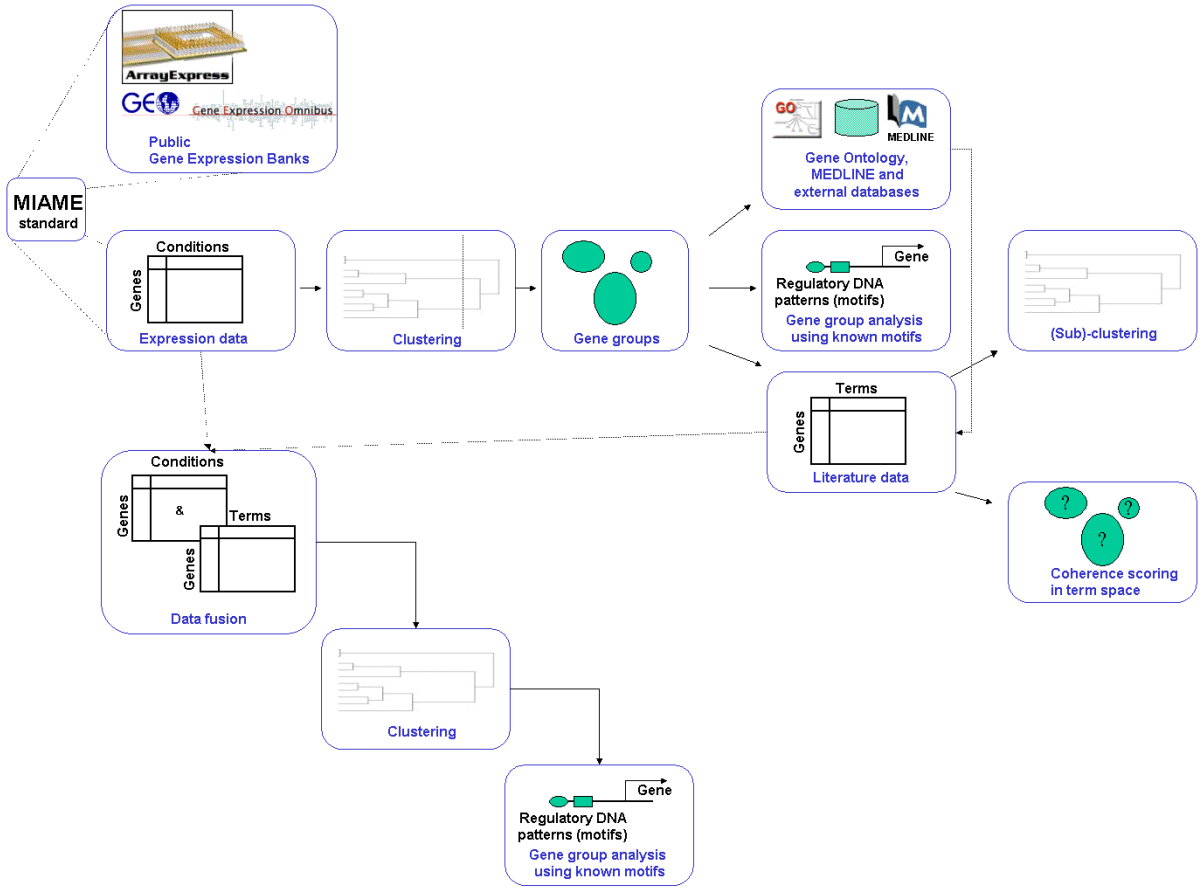


Figure 1.11: General overview of the biological data mining framework in this thesis.

search groups in Flanders. An operational analysis, design of a first generation database scheme, and implementation of a prototype Laboratory Information Management System (LIMS) was performed. This work was motivated by the fact that there is no better way to get to know the data than from the inside. This is discussed in Chapter 2. Joining several prominent international researchers during the first MGED(Microarray Gene Expression Databases) conferences[5] lead to a one-year membership (2000-2001) of the Steering Committee for the establishment of standards for microarray data, which ultimately resulted in a co-authorship on a seminal publication in Nature Genetics.

Chapter 3 overviews principles of microarray data analysis with a focus on clustering. Special attention is given to the assessment of cluster(ing) quality and automatic determination of the number of clusters, a common parameter in numerous clustering algorithms. These methods will be used later on to analyze both expression and literature data.

One step beyond statistical analysis of expression data is to interpret the results and link them to already existing knowledge such as public databases or scientific articles in order to render the results increasingly interpretable. Both existing and new approaches to biological cluster validation are provided in Chapter 4). In particular we present a motif-based heuristic to assess the biological relevance of a clustering results in terms of transcriptional regulation. Our M-score proves particularly useful when many clustering solutions need to be evaluated.

Text mining refers to the computer-assisted retrieval and extraction of knowledge from electronically encoded information with little structure[6]. Due to lack of gold standards, porting text mining techniques to the biomedical domain requires some ad-hoc tuning, but above all carefully designed experiments and at times creative evaluations. Chapter 5 introduces the adopted vector space model and presents statistical evidence that our text mining methodology is suitable to represent information on biological function. A coherence measure in text space is introduced and used to score functionally related gene groups. Also, we quantitatively evaluate how well gene groups are reconstructed by clustering genes based only on literature information.

We have co-developed text mining software to link information extracted from specialized literature to the interpretation of microarray data, ultimately aiming a piece of software that is (a) capable of alleviating the problem of data interpretation (b) user-friendly, and (c) deployable within the existing framework of our research group. Chapter 6 describes our application, baptized TXTGate, and presents several biological case studies. With TXTGate we maximally exploit the flexibility of the vector space model by offering different vocabularies that act as perspectives to literature.

Combining different information sources in a statistically rigid way is getting increased attention within various scientific communities, most notably in bioinformatics. Deep integration of various information sources build upon the

[5] www.mged.org

[6] Database systems, taxonomies or ontologies can be considered highly structured representations.

idea that combining of information pays off more than using individual sources separately. Having shown that literature typically encodes functional information, our next goal was to integrate text with expression data. Using techniques from statistical meta-analysis we show in Chapter 7 how joining expression-based correlations and literature-based correlations benefit the overall outcome of gene expression cluster analysis.

Finally, Chapter 8 reviews and discusses the main issues in this work and provides indications for future work.

Chapter 2

Large scale gene expression experiments

The announcement of a near-complete draft of the human genome in 2001 by the International Human Genome Consortium [66] and the Celera Genomics [158] marked a milestone in the life sciences. In the same way mathematics and information technology had to reach biologists to accomplish this endeavor, molecular biology has to reach engineers and statisticians embarking upon bioinformatics. Therefore we start this chapter with a summary of the most important concepts of molecular biology.

As pointed out already in Chapter 1, functional genomics tries to identify the role of genes, gene products and the relationships between them. One of the technologies that have become prevalent to this end are microarrays, which simultaneously measure the level of expression of thousands of genes. We provide an overview of the technology, along with some characteristics that are important to data storage and analysis.

Microarray experiments are complex and information-rich, thus requiring data storage, data exchange and data reporting specifications. We will overview the Laboratory Information Management System (LIMS) developed at the MicroArray Facility [1] of the VIB. The VIB Interuniversity Institute for Biotechnology, is a non-profit scientific research institute with some 60 research teams across four universities in Flanders that are active in the frontline of the life sciences [2]. Finally, we will discuss MIAME[3], a proposed standard for reporting about gene expression experiments, to which we have also contributed. Figure 2.1 pictures the concepts covered in this chapter.

[1] http://www.microarrays.be

[2] http://www.vib.be

[3] We actively contributed to the development of MIAME 1.1, and are listed as a co-author in the original commentary published in *Nature Genetics* [54].

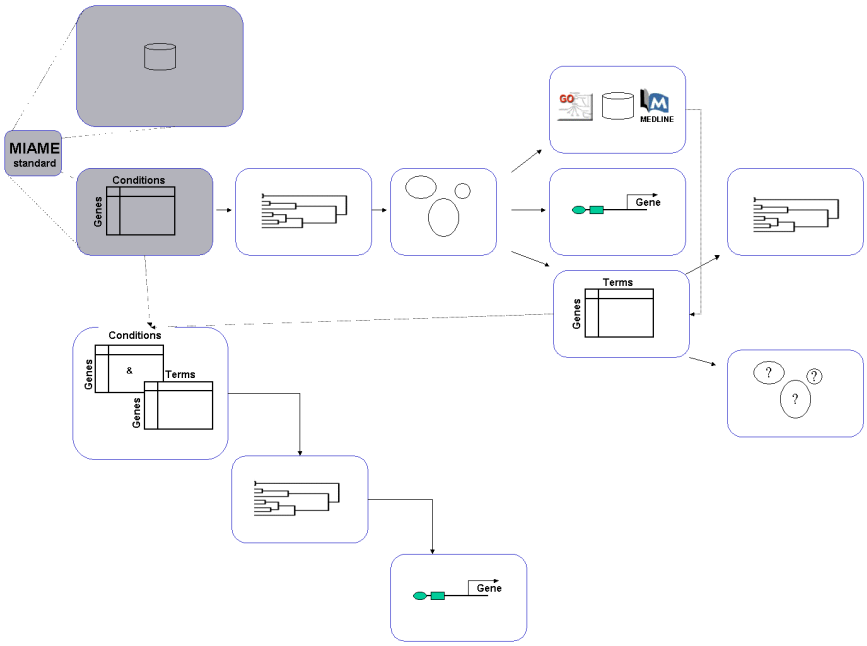


Figure 2.1: Microarrays can be considered the first mature genome-wide analysis technology. As other present-day biological data, microarray data is characterized by a large discrepancy between the pace with which biological data is produced and the rate at which information is extracted. One of the barriers has been the lack of transparency in publicly disclosed expression data. We have contributed to the development of MIAME 1.1 [54], a proposed standard for reporting about gene expression experiments. In order to furnish this kind of information from a laboratory environment we have developed and implemented a first generation information management system (LIMS) at the VIB MicroArray Facility. This chapter will introduce microarray technology along with its link to databases and data reporting schemes.

2.1 Concepts of molecular cell biology

2.1.1 DNA

DNA (deoxyribonucleic acid) holds the keys to how living organisms work. All organism's inherited characteristics are contained in this molecule via a code consisting of four nucleotides: A (adenine), C (cytosine), T (thymine) and G (guanine). The translation, however, of this code to the characteristics of a living being are complex. DNA is a double stranded molecule whose structure is driven by a highly selective binding affinity between A/T and C/G respectively. This chemical property is also central to microarray technology covered later in this chapter.

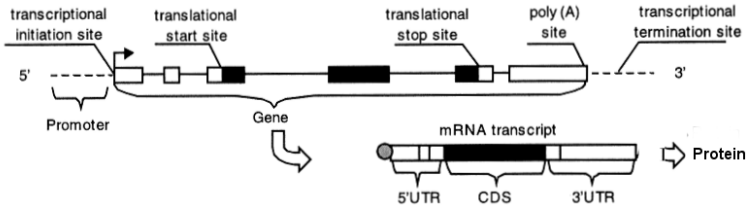


Figure 2.2: Overview of gene expression: Proteins are assembled from DNA through intermediate actions on mRNA. Transcription is directed (from 5' to 3') and takes place between the transcriptional initiation site and the the transcriptional termination site. Next, translation of the coding sequence (CDS) into the protein is carried out. Regions in the mRNA that are not translated (untranslated region or UTR) and untranscribed regions play an important role in the regulation of a gene's expression. In model eukaryotes such as *S. cerevisiae*, regions *upstream* the gene are crucial to how it comes to expression (Illustration from Alberts *et al.* [36]).

2.1.2 Genes

Proteins are central to a cell's machinery and their chemical composition as well as their number are determined by genes. The derivation of a chain of amino acids that form a protein is done by 'reading out' particular parts of the DNA – the genes. Roughly this process is divided in two steps: Firstly, transcription to so-called messenger RNA or mRNA takes place. Transcription is directed (from 5' to 3') and takes place between the transcriptional initiation site and the the transcriptional termination site (see Figure 2.2). The transcribed regions hold the crucial information about the ordered amino acids in the protein sequence to be constructed and this constitutes the central dogma of molecular biology. Next, translation of the coding sequence (CDS) into the protein is carried out. Regions in the mRNA that are not translated (untranslated region or UTR) and untranscribed regions play an important role in the regulation of a gene's expression (see Figure 2.2). Notably the so-called promoter region in front of the transcribed sequence contains binding sites for transcription factor proteins

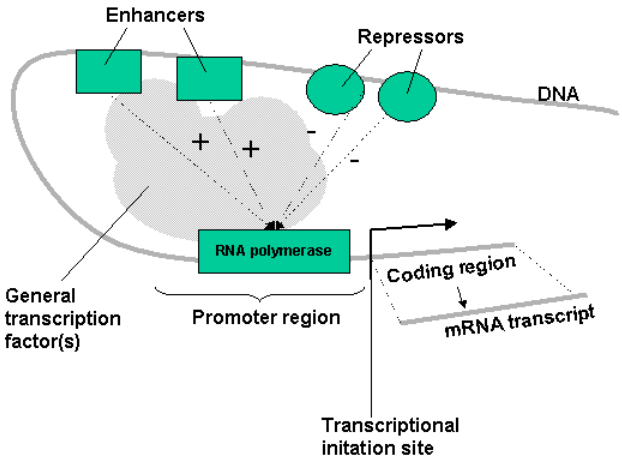


Figure 2.3: Transcriptional regulation in yeast: Proteins (visualized as various kinds of blocks) called transcription factors bind to specific regions of the DNA (called binding sites) that are *upstream* of the transcriptional initiation site. Through chemical conformational changes they can contribute to signaling (indicated as arrows) the RNA polymerase that binds to the region in the DNA called the promoter, to start, stop, enhance or repress the transcription of a gene. This constitutes a simplified view, valid for the model organism yeast used in this dissertation. For higher order organisms, however, regulation is much more complex.

that start up transcription. See Karp [99] and Alberts *et al.* [36] and for more extended and didactical expositions.

2.1.3 Motifs

Transcription factors (TF) are proteins that bind to regulatory sequences *upstream* of a gene in a DNA sequence. They can bind to the promoter region (i.e., immediately preceding the gene), or to other, more upstream regions. In the first case they initiate the transcription process, in the latter they can enhance or silence it. To see how, we have to look at some elements of the protein *structure* of a transcription factor. The affinity of a transcription factor with a particular gene is due to its specific DNA-binding domain. The TF can interact with polymerases that actually perform the transcription, or with other transcription factors. Such interaction can enhance or repress the activity of transcription. Figure 2.3 visualizes this. Often the binding site (or *motif*) is referred to by the name of the transcription factor it 'attracts'. See Wray *et al.* [161] for a extensive review on transcriptional regulation in eukaryotes.

2.1.4 Cellular organization

The function of a protein, and hence of a gene, is related to its location in the cell. Therefore we introduce the most important components and processes in an

eukaryotic cell (see Figure 2.4). We distinguish several internal components of a cell. The nucleus (core) contains the genetic information (DNA) of the organism, which is identical across all cells of the organism. The DNA is packed densely and organized in chromosomes[4]. The rest of the cell excluding the nucleus is termed the cytoplasm, which contains various organelles. The fluid portion of the cytoplasm is the cytosol. The mitochondria carry out the oxidation of food molecules and are crucial to the energy supply. The cytoskeleton are protein filaments throughout the cytosol that determine the cell's shape and movement. Lysosomes are responsible for intracellular digestion and break down unwanted molecules for recycling or excretion. The endoplastic reticulum (ER) is responsible for the assembly of export material as well as the assembly of the cell's membrane components. The Golgi apparatus acts as a kind of processing and distribution center. Finally, vesicles are formed from membranes and have a transport function. Some of the defined locations and functions will be further elaborated on throughout the dissertation.

Saccharomyces cerevisiae

As human cells are complex and difficult to work with, yeast researchers work with *Saccharomyces cerevisae*, or baker's yeast, to study eukaryotic cell biology. Its 'simple' structure, robustness and speed of reproduction has made it popular particularly in genetics and molecular biology. The entire genome of this species has been sequenced and it is used to do research on the basic cellular mechanics of replication, cell division and metabolism. Many data sets used in this dissertation will pertain to this model organism.

2.2 Microarray technology

2.2.1 Technology

The advent of microarrays testifies to the systematic shift of attention within the biomedical communities from the sequencing of DNA to systematic screen of gene function. Examples of typical questions include

- What are the functional roles of different genes ?
- How are genes regulated ?
- How does the expression of genes change under various diseases or treatments, conditions, mutants,... ?

Microarray technology enables a systematic screening of the behavior of thousands of genes simultaneously over time or under a variety of conditions.

A microarray is a glass slide onto which *single*-stranded DNA is attached at fixed locations (spots). Each spot on an array is meant to correspond to a

[4]In fact, DNA is packed into chromosomes only during the division of the cell. It is then that DNA becomes observable through a microscope.

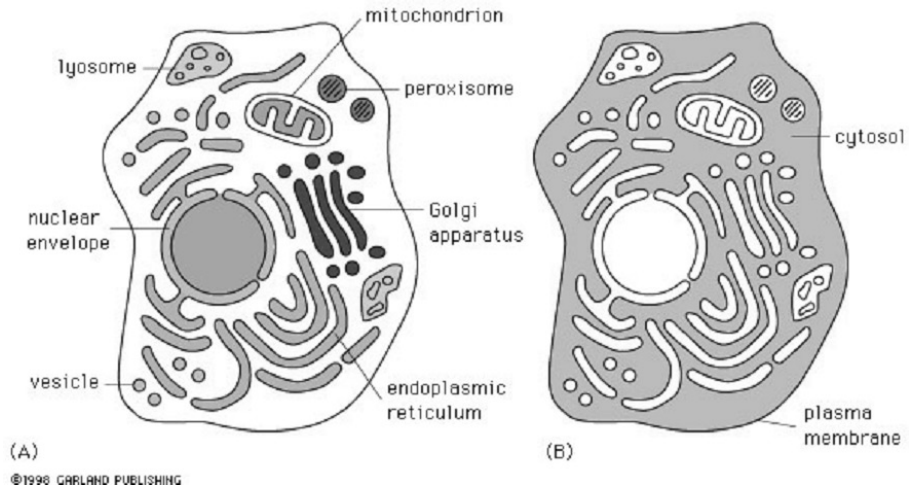


Figure 2.4: Structure and organization of a eukaryotic cell. We repeat the description in the text for clarity. The part cell excluding the nucleus is termed the cytoplasm and contains various organelles. The fluid portion of the cytoplasm is the cytosol. The mitochondria carry out the oxidation of food molecules and are crucial to the energy supply. The cytoskeleton are protein filaments throughout the cytosol that determine the cell's shape and movement. Lysosomes are responsible for intracellular digestion and break down unwanted molecules for recycling or excretion. The endoplastic reticulum (ER) is responsible for the assembly of export material as well as the assembly of the cell's membrane components. The Golgi apparatus acts as a kind of processing and distribution center. Finally, vesicles are formed from membranes and have a transport function. (Illustration from Alberts *et al.* [36])

single gene and their number might run into several thousands. The preferential binding that exists between two complementary strands of DNA (Section 2.1) enables a probing[5] of a sample cell or tissue for the abundance of mRNA from all the genes. This abundance is (almost) always measured in a relative setup, e.g., measuring expression level of a sample versus control. In practice, both sample and control RNA are labeled with a fluorescent marker. To measure the relative abundance of sample versus control RNA, the array is excited by a laser. Each spot where the sample RNA population is abundant appears red, or green if the control RNA population is abundant [55]. If sample and control bind equally the spot will be yellow; if neither bind it will appear black.

A popular brand of microarrays are so-called cDNA microarrays. cDNA's are cloned copies of otherwise fragile mRNA. An Expressed Sequence Tag or EST is a tiny portion of an entire gene that can be used to help identify unknown genes and to map their position within a genome. As they are derived from cDNA, they represent a region of the gene that is transcribed.

cDNA microarrays are made by mechanically depositing ('spotting') amplified ESTs from cloned cDNA libraries onto glass microscope slides. As ESTs are fragments of mRNA, linking ESTs back to the gene it is transcribed from, is an error-prone task. The annotation of the EST in its turn depends on the information available on the gene it matches. Good annotations of ESTs are therefore important to infer correctly interpret a cDNA microarray experiment. EST annotations are available in a multitude of databases (e.g., dbEST [5]).

Two different samples are hybridized to a single array and a ratio of the expression level between these two samples for each probe should be calculated. The reason for this is that the absolute amount of probe at each element varies due to the amplification step. In addition, there are variations from array to array due to the printing process such that inter-array experiments based on absolute intensity become difficult. The overall setup for cDNA microarrays is pictured in Figure 2.5.

For the sake of completeness, we also mention the existence of another important gene expression technology. The Affymetrix 'Genechip', or oligonucleotide platform is a commercial implementation. These arrays are manufactured through a patented process that is inspired by microchip technology. Each gene's expression is assayed by detecting specific hybridization between sample RNA and etched oligonucleotide probes of length 15-20. Various mismatch control probes are added and the system aims at a high specificity by discarding measurements where hybridization to differing probes is observed.

As with all large scale experiments, microarrays are costly in terms of time, equipment and labor. Hence setting up an experiment so that informativeness and efficiency are maximized is an art in its own. We refer to Yang and Speed [164] who review experimental design issues in microarrays.

[5]We define the elements that are fixed on a surface as *probe*, following the idea that a sample (e.g., from a treated tissue) of interest is probed for the presence of abundant mRNA.

2.2.2 Gene expression matrix

Conceptually, a microarray experiment can be viewed as a table with rows representing genes, columns representing various samples, and each position in the table describing the measurement for the particular gene in the particular sample (see Figure 1.2). Such a table is commonly referred to as a microarray gene expression matrix. In addition to this data matrix, a description of a microarray experiment should also contain information about the genes whose expression has been measured (gene annotation) as well as the experimental conditions under which the measurements were taken (sample annotation). Therefore the information required to describe a microarray experiment can be divided into three logical parts: gene annotation, sample annotation and the microarray data matrix.

2.2.3 Sources of noise

Following Schuchhardt *et al.* [134] we will give a list of the major sources of fluctuations to be expected in a microarray experiment:

- mRNA preparation (Target): depending on RNA extraction process from the tissue, targets may look different across samples.
- cDNA preparation (Probe): reverse transcription to cDNA will result in array elements of varying length. Amplification of clones with PCR is hard to quantify and cross-contamination of neighboring clones in well plates may occur. Finally, linking EST's to the respective gene may be non-trivial.
- Labeling efficiency: the fluorescent labeling may fluctuate randomly and depends on nucleotide composition.
- Variations across pin geometries: surface properties and therefore varying transport characteristics of spotting pins yield different amounts of cDNA on the array elements.
- Fluctuations within one print tip group: the amount of transported probes stochastically varies even for the same spotting pin.
- Fixing properties of probe: chemical interactions between array slide and probes are far from understood. Moreover, this can vary across the same slide (i.e., slide inhomogeneity).
- Hybridization efficiency: various parameters including temperature, time and amount of target molecules determine the hybridization process. Also hybridization can sometimes be non-specific.
- image analysis: overshining (signals from neighboring spots), non-linear transmission characteristics, saturation effects and variations in spot shape complicate the image analysis.

Many packages for image quantitation are available, and despite minor difference between them, most give high-quality, reproducible measures of hybridization intensities of each gene element [128]. Such intensities are always relative in the sense that they represent expression levels to be compared against a reference. Therefore, expression *ratios* are popular measures to quantify expression changes. Taking the logarithms of these ratios results in more symmetry in the range of values from up- and down-regulated genes. Normalization of microarray data[6] refers to transformations of and operations on all (ratio-based) data within a gene expression experiment to balance hybridization intensities with respect to the sources of variations outlined above. For a more detailed overview we refer to Quackenbush [128] and Marchal [108].

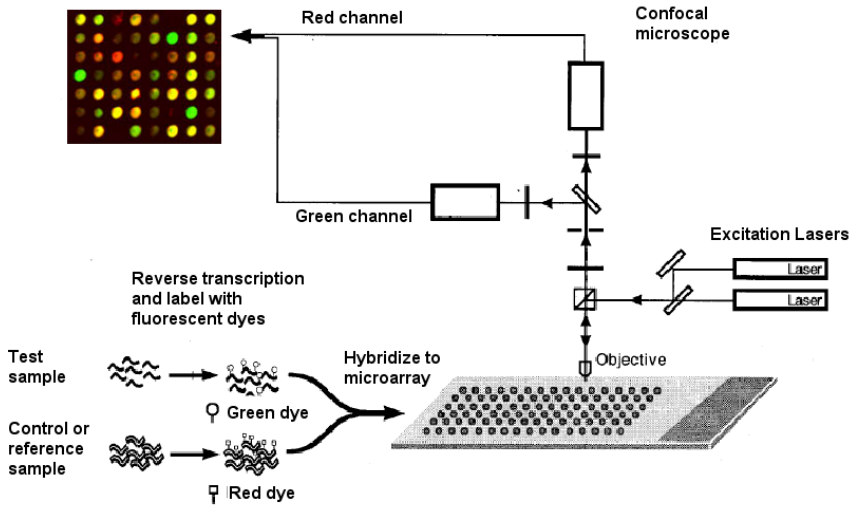


Figure 2.5: cDNA microarray procedure. RNA from the test sample and a control sample is extracted, subsequently labeled with fluorescent dyes and hybridized to a glass slide (microaray). Each microscopic slide is composed of a high number of deposited clones (spots), each representing one gene. Hybridization occurs if complementary DNA from either control or test sample is present. Relative abundance of test versus control RNA per spot is measured by imaging the slide with a confocal microscope imager.

[6]Note a clear distinction between normalization used in the context of a single gene expression *vector*, which corresponds to one mathematical operation, and normalization used more generally when talking about gene expression *data* in general, which refers to a whole set of preprocessing techniques.

2.3 Microarrays and database systems

The advent of high throughput screenings on the workbench of the biologist created a chronic need for information systems to manage and store vast amounts of data generated by these technologies. In previous section a flavor of the complexity of the technology and various sources of noise were given. For these reasons, it is not immediately clear what the necessary level of detail is to reproduce a microarray experiment. Neither is it straightforward to see what reported information is needed to make experiments more valuable for various kinds of data mining.

In this section we will discuss our developed laboratory information management system (LIMS) that includes a database scheme and a graphical user interface aimed to support the workflow at the VIB microarray facility. Next we will treat the MIAME minimum information requirements to which we contributed through the MGED community in the period 2000-2001. MIAME is meant to guide researchers in making (reports about) expression experiments increasingly clear and reproducible, and the specification can serve as a blueprint for the development expression databases.

2.3.1 LIMS

One of the difficulties in generating reliable cDNA expression data is the necessity to identify each spot with the respective gene. Currently, one glass slide at the VIB microarray facility can contain up to 26,000 spots. As explained before, typically, a proprietary 'clone' library is purchased, amplified, stored and ordered into so called well-plates, usually per 96 or 384. Contamination (leaking of DNA fragments to neighboring wells) and equipment constraints often require a reordering of the clones usually performed by pipetting robots. In the final preparation another robot spots the probes to the glass slide so that each spot represents a gene. The underlying biochemistry is complex and the possible configurations are numerous. Hence to deal with this error-prone process, mechanisms to backtrack efficiently on any errors that might have occurred during this production process are crucial. An automated administration of each action on the workbench is therefore necessary. Laboratory Information Management Systems (LIMS) support personnel by allowing a systematic registration of each step. We have developed a first generation LIMS system at the VIB Microarray Facility[7]. In Figures 2.8 and 2.9 we show screenshots of a Graphical User Interface (GUI) of the developed LIMS. The GUI graphically represents the different steps of the production process.

Figures 2.6 and 2.7 shows the database scheme that underlies the LIMS. We distinguish two main blocks: one registering all steps in the production process (stock management, clone identification and tracking, and quality control), the other containing information about the set of hybridizations within each experiment. We note that besides laboratory-specific information (such as stock

[7] http://www.microarrays.be

information and protocols), the first block also contains information crucial to reproduce an experiment (e.g., clones ids, clone amplification parameters, ...). Nevertheless the core of a single microarray experiment is encoded in the second block. Here we find the hybridizations of samples and controls on one or several arrays, along with the resulting measurements of relative abundance of the screened gene transcripts. Also each studied condition needs to be accurately reported. This involves not only the biological question addressed, but also details about for example mRNA extraction. Although tightly connected to the LIMS system, this 'results database' is not a part of it.

The presented system and underlying database scheme are typical for a facility that produces arrays. Publicly available expression databases that group microarray experiments from a broad user community serve a different goal and therefore advocate a different data model. This will be treated in the next section.

2.3.2 MIAME

It is widely acknowledged that there is a need for a public repository for microarray data [71], whose functions would include providing access to supporting data for publications based on microarray experiments. However, it has been much less clear exactly what information should be stored in such databases, in particular as far as providing supporting data for publications is concerned. As a single hybridization is a final step in a sensitive chain of biotechnological operations involving numerous materials, protocols, hardware and software, it has been not clear what level of detail is necessary to accompany a publication. Moreover, contrary to sequence data, expression data only makes sense in a certain context: while the sequence of a gene is uniquely determined, its transcriptome is highly dependent on the environmental conditions. Also, should databases store raw microarray scans (images), or is one final summary scalar per array element (e.g., one green/red ratio per spot for two channel platforms) sufficient ? Clearly, there are many possible intermediate levels between these two extremes such as storing complete outputs from image analysis software together with the information about the analysis methods and parameters.

The precise nature of the information to be stored will be dictated by the role of the particular database or repository. If the unique goal of the database is to archive supporting data for published experiments, then it can be assumed that the publications themselves will provide information explaining the database entries. It might in principle be left to the peer reviewing to determine whether a particular publication together with the respective database entry contains the information that is necessary to verify and potentially reproduce the experimental results. However, it is unlikely that such a system could be efficient, and moreover the value of such a non-standardized database would be rather limited. For instance, it would be difficult to use the database for high throughput automated data analysis or mining. The experience of the sequence databases over the past decade unequivocally demonstrates the strategic importance of structured, consistent annotation applied early in the process of data genera-

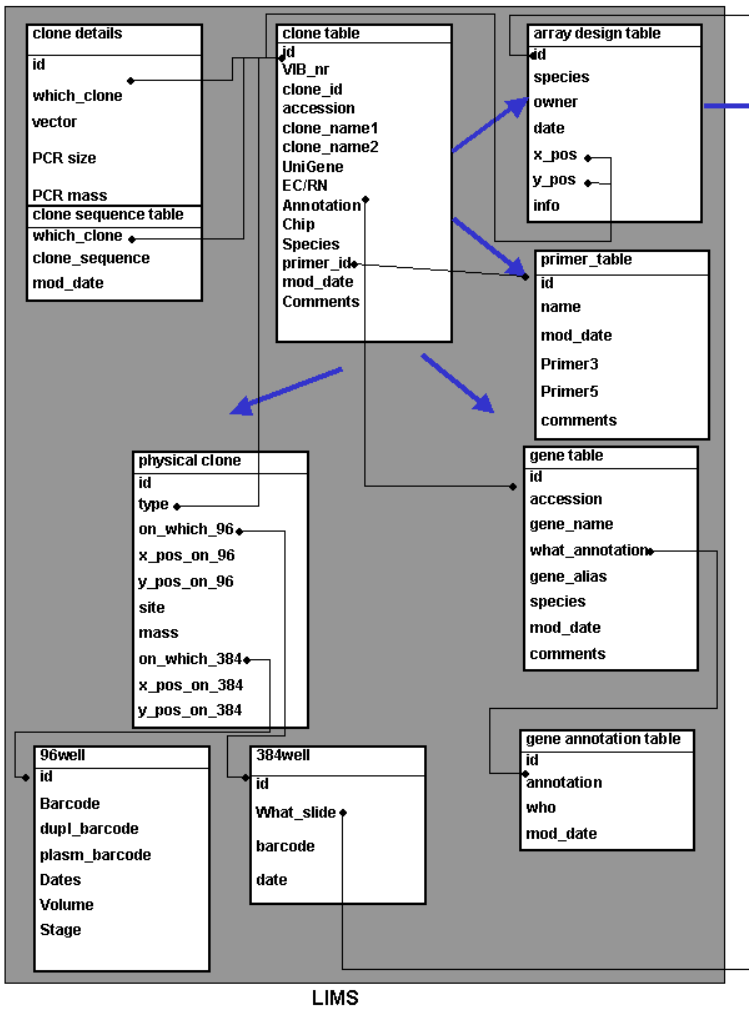


Figure 2.6: Developed LIMS database scheme – part 1. This block contains information specific to the production process involving the clones. We track localization of amplified clones (*physical_clone*) between different well plates (*96well* and *384well*). Details on the amplification process are encoded in table *primer*. As each clone is to represent one gene, proper annotation is crucial. Tables *clone*, *clone_details*, *clone_sequence* and *gene*. Tables are connected through foreign keys: for example, the primer information belonging to a given instance in table *clone* is encoded as a reference to a particular instance of table *primer*. Note some connections go out to Figure2.7.

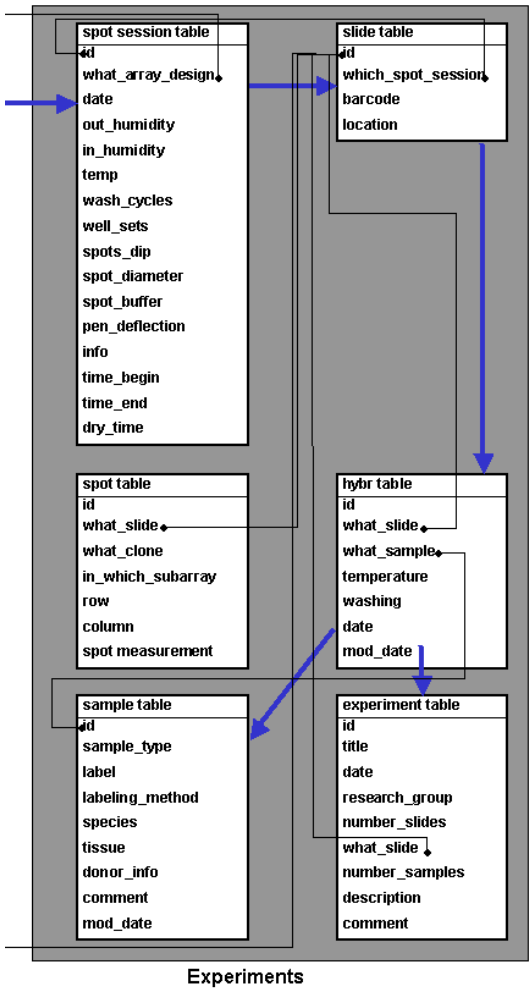


Figure 2.7: Developed LIMS database scheme – part 2. This block groups experimental values about the series of hybridizations for each experiment. Information on experiment, sample, slide, hybridization session and spots are encoded in the respective tables. Although this early design is not fully MIAME compliant, its logical organization corresponds to the major structure of the specification. Indeed, although not a goal in itself, MIAME is developed with the realization of gene expression data models in mind. The majority of the information for MIAME is found in this 'results database', rather than in the LIMS database (Figure 2.6), which is more geared to support the management of laboratory-specific data.



Figure 2.8: Screenshot of the developed LIMS: Labo personnel is guided through submission of new clones via an intuitive interface. We distinguish : (1) Arrival of new clones, (2) Custom array ordering and (3) Database querying

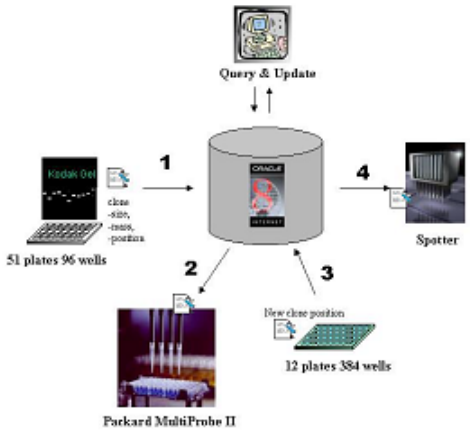


Figure 2.9: Screenshot of the developed LIMS: JAVA scripts shelled by a visual interface guide laboratory personnel through several steps of the recording of information. We recognize several ordered procedures that occur with the LIMS database in the back end: (1) quality control of cloned EST's, (2) Redirection of information to re-arraying robot, (3) Processing of reordering from 96 well plates to 384 well plates, (4) Data forwarding to the 'spotter', which deposits the clones on the array and (5) interaction with database

tion.

The MIAME (Minimum Information About a Microarray Experiment) standard addresses these issues in an open collaborative specification for the minimum information that should be reported about a microrarray experiment. Adopting such a standard by the microarray community would facilitate the establishment and usefulness of databases for microarray data. MIAME was also set up to guide microarray manufacturers and software producers so to easily capture MIAME information in LIMS systems. This enables MIAME compliant information generation without extra effort by the researchers. The MIAME specification is organized into six parts:

1. Experimental design: the set of hybridization experiments as a whole,
2. Array design: each array used and each element (spot) on the array,
3. Samples: samples used, extract preparation and labelling,
4. Hybridizations: procedures and parameters,
5. Measurements: images, quantification and specifications,
6. Normalization Controls: control types, normalization procedures and parameters.

The main components of MIAME are also retrieved in the second part of our LIMS database scheme (Figure 2.7). As illustrated in Figure 1.3, MIAME lies at the intersection of LIMS and private result databases. For a detailed specification of MIAME we refer to Brazma *et al.* [54]. The original article is cited over 200 times two and a half years after its date of publication. Through data formats that implement the standard (MAGE-ML [140]), exchanges (a) between public and private databases, (b) between databases and software and (c) between individual users, are made possible.

Over the past two years MIAME has been increasingly welcomed. Now it is even obligatory to publish results on microarray data in several high-impact international journals (Nature, Science, EMBO Journal and Lancet). Taken from the editorial in Nature, November 2002 [72]: '... more alarming [than the fact that web supplements are insufficient to disseminate full data and experimental conditions] is the concern that the [full] data are not being subjected to rigorous peer review ... Sparked by these concerns Nature research journals have agreed on a policy for submission and publication of microarray data'.

However, important problems remain. Although MIAME's data *reporting* guidelines increasingly enable cross-experimental queries, many data sets cannot be easily compared because of the absence of Quality Control of Standard Operating Procedures. Especially for cDNA arrays this is problematic due to considerable variations in array manufacturing. In a recent article [155] the Tumor Analysis Best Practices Working Group attempt to create guidelines, endorsed by various microarray core facilities, to develop platform-specific guidelines on data *generation* and data *interpretation*.

Table 2.1: Overview of important public gene expression databases (as of 02/2004). Entries marked with an asterisk adhere to the MIAME standard.

Name	♯	Web Address
ArrayExpress (*)	133	http://www.ebi.ac.uk/arrayexpress/
Gene Expression Omnibus (*) (GEO)	746	http://www.ncbi.nlm.nih.gov/geo/
Stanford MicroArray Database (*) (SMD,LAD)	158	http://genome-www.stanford.edu/microarray
HuGE Index	59	http://www.hugeindex.org
RAD (*)	13	http://www.cbil.upenn.edu/RAD/
yMGV	82	http://transcriptome.ens.fr/ymgv/
GXD (*)	n.a.	http://www.informatics.jax.org/

2.4 Discussion

Although cost factors initially formed a barrier to a wide dissemination of high throughput techniques, we can expect that more and more laboratories will have access to the technology. Moreover, as gene expression only makes sense in a particular cellular context the number of expression data is to quickly surmount the amount of sequence data. Next to IT infrastructure supporting the laboratory work flow (see Section 2.3.1), database systems that can record bulks of expression data in a scalable, organized, documented, easy-to-access, and cross-linked manner are thus of increasing importance. Such expression compendia will enable

- an easier comparison of analyses by different algorithms, methodologies or on different platforms,
- a faster identification of benchmark datasets,
- conduction of 'in silico' experiments over numerous microarray screenings.

In Table 2.1 we enlist some important publicly available gene expression repositories. Although not all databases or journals subscribe to MIAME (yet), the proposed standard is an important milestone in leveraging publicly available expression data. Gradual adherence to the standard by both academic and industrial players shows its ongoing dynamics. However, on the practical side, not all researchers are were initially keen to unconditionally disclose the data that souped up large chunks of their research budget. The MIAME check-list, however, never required authors to reveal every gene on the array, as this could allow to scoop experiments in progress. Data should not be published prematurely and MIAME adopts the logic that the identity of things you are not publishing can be left out. Some requirements in MIAME might demand proprietary information from commercial products used in the experiment. For example, information on the required hybridization protocols might be confidential and unknown to the researcher if the experiments have been carried out by commercial providers. On a higher level, the MGED forum that sustains the standard is successfully negotiating broad recognition of MIAME by funding agencies as well as commercial providers. Affymetrix, for example, have implemented MIAME-compliant data output in their new software releases.

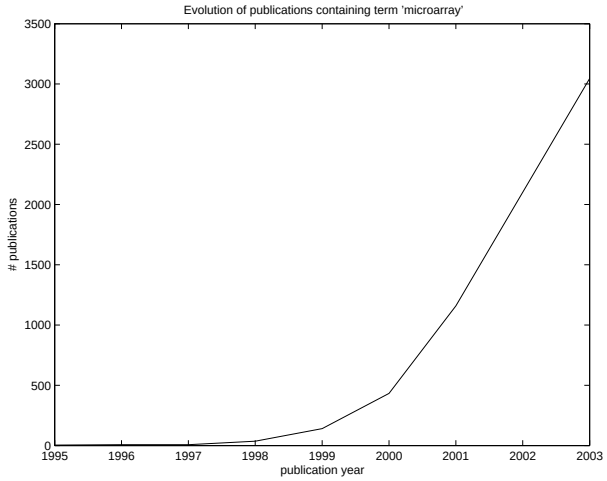


Figure 2.10: Evolution of number of publications between 1995 and 2003 containing term *microarray*. A total of 6791 publications were retrieved in PUBMED and we witness a continuing growth rate.

On the long run we expect expression databases adhering MIAME will gradually be considered as mines rather than as archives. Indeed, gene expression databases will increasingly impregnate the way microarray experiments are set up. Firstly, optimal design will be determined not only by statistical considerations, but also by a systematic screening of prior experimental work recorded in public repositories. Secondly, once a set of new hybridization experiments are planned and conducted, comparing their outcome to previously deposited assays can confirm, question or refute adopted hypotheses or methodologies. More ambitiously, previously accumulated expression data can be tied to the expression assay at hand, and evidence can be boosted by carrying out a so-called meta-analysis. Such endeavors are driven by the statistical principle that a combined analysis can reveal more that the sum of its parts (see Moreau *et al.* [117] for more details). Finally, after a careful validation of the candidate hypothesis (using independent information from other databases, other studies, literature or expert knowledge) the full data supporting the experiment is, in its turn, deposited to a repository –usually accompanying a scientific publication reporting the conclusions.

Figure 2.10 shows the number of publications carrying the term 'microarray' and indicates a steadily growing interest in the technology. Considering that over thousand experiments – each one often pertaining multiple array hybridizations – are already recorded in present-day public repositories, many opportunities arise to do *in silico* biology. Several recent publications provide convincing illustrations of the power of so-called 'dry' molecular biology in characterizing unknown genes and detecting novel regulatory gene networks (see e.g., Ihmels *et al.* [95], Wu *et al.* [162]), Bergmann *et al.* [47] and Stuart *et al.* [143]). Con-

trary to the somewhat controversial formulation, 'dry' or *in silico* approaches do not exclude experimental approaches. Rather than generating novel data to support pending hypotheses, *in silico* approaches generate novel hypotheses by processing vast amounts of already existing data. Promising outcomes of such exercises, however, still need corroboration in the lab.

Chapter 3

Exploratory data analysis

3.1 Introduction

Exploratory Data Analysis is an approach for data analysis that employs a variety of techniques to maximize insight into a data set. Its goal consists of (a) uncovering underlying structure, (b) extracting important variables, (c) detecting outliers and anomalies or (d) testing underlying assumptions [25]. One of the techniques include cluster analysis or clustering, where the data is organized into more tractable subgroups. We treat popular clustering algorithms such as K-means and hierarchical clustering, in Section 3.3. Clustering is an unsupervised technique meaning that there is no guarantee to arrive to an optimal partitioning of the data. As a result many plausible outcomes are possible. The number of clusters, for example, is a typical parameter that needs to be supplied by the user beforehand. We present an existing technique based on subsampling to tune this parameter in order to get the most mathematically stable solution. In this dissertation the presented techniques will be applied to divide genes into groups based on their expression profiles, as well as their literature-based information (See Figure 3.1 and the overview presented in Chapter 1). Other application of microarray data analysis are treated in detail in De Smet [139].

Genome expression data encompasses dozens of measurements about thousands of genes (a microarray chip at the VIB microarray facility, for example, can screen the activity of up to 26,000 genes) and the inherent noisy nature of expression data puts a challenge to the application of these clustering methods. To make them work with increased efficiency (in terms of computational time) and accuracy (in terms of the detection of biologically plausible patterns), a second generation of clustering algorithms have been developed. In Section 3.6, we describe how we had to modify the mentioned subsampling method to handle real-life gene expression data. We conclude the chapter with a short overview of other important microarray clustering methods and how they link to our work.

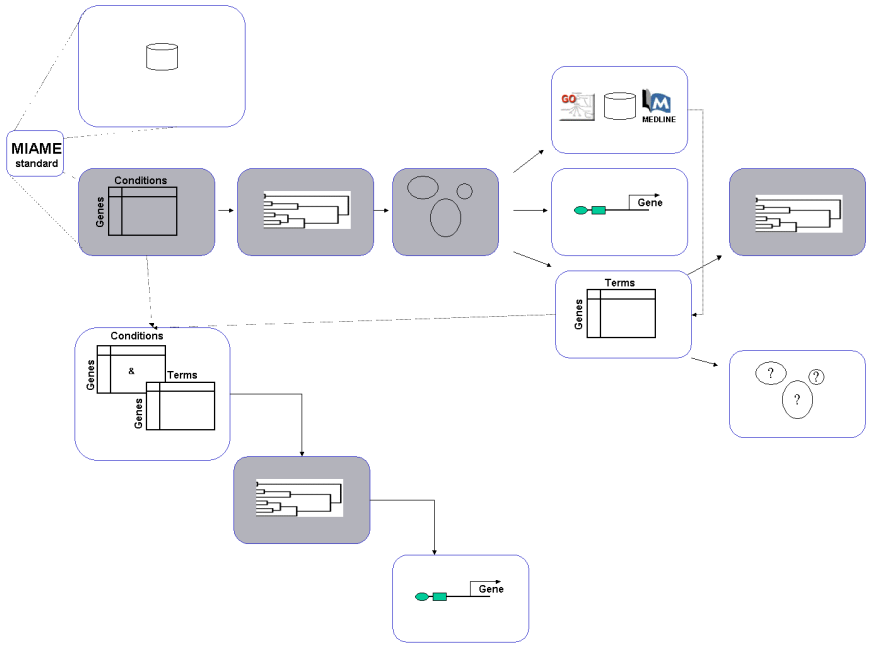


Figure 3.1: One important technique to construct tractable lists of genes is clustering. In this chapter we describe the 'classical' clustering techniques, often used in expression analysis (K-means and hierarchical clustering). However, these techniques require a predefinition of the number of clusters (k), which is rather arbitrary. We discuss how k can be determined statistically with our adaptation of an existing method based on subsampling. To enable a comparison of clustering results to other partitionings of data (e.g., from an expert or from another algorithm), comparative measures are presented.

3.2 General clustering methodology

Clustering is an exploratory tool and requires specific needs. Several and possibly recurring steps can be identified, which are depicted in Figure 3.2:

1. Data collection and representation: Careful recording of the data and appropriate preprocessing (missing value management, normalization,...) constitute first important steps in clustering analysis. A good mapping of the data in feature space, i.e, finding a good *representation*, constitutes an art in its own and can drastically determine the outcome of an exploratory analysis. Feature types can be continuous, discrete, binary, ordinal,... or a mixture of these. Additional transformation of the feature space to reduce dimensionality might be necessary. In the subsequent chapters we will apply the methodology on microarray gene expression data (Chapters 4 and 7) and gene literature information (Chapters 5, 6 and 7).

2. Similarity measure: Distance metrics directly influence the formation of clusters. Although many choices to quantify relations between objects may perform well, a careful understanding of the problem and chosen representation is required. Some algorithms have an inherent measure. In K-means clustering, for example, the Euclidean distance measure is intrinsically adopted.

3. Algorithm: The chosen clustering method is also problem-dependent. Important issues include transparency of parameterizations, computational and memory complexity, type of expected clusters (size, shape, number), level of desired interactivity, visualization capabilities and incorporation of prior knowledge. As clustering algorithms we use K-means clustering and hierarchical clustering. Both require a choice of the number of clusters k. Hierarchical clustering is computationally more expensive, but deterministic and offers the advantage of outputting a binary tree so that k can be determined visually.

4. Cluster validation: Assessing the relevance of a clustered data set can range from very easy to very hard. As clustering is an exploratory technique, it is an ill-defined problem because, although multiple good solutions may exist, objective criteria to define the optimal solutions are hard to determine. Assessment of cluster quality can be done based on the data that generated the result (i.e., 'statistically', which is covered in this chapter) or using external information (i.e., 'biologically', which is covered in Chapter 4). For the first kind we distinguish 'internal' and 'external' measures. Some of these measures can be integrated into the clustering algorithm itself (step 3), thus giving rise to a new clustering method.

   We will exemplify this with the subsampling method and our improvements hereto. The subsampling method determines optimal k by exploiting properties of the *Jaccard* coefficient that measures the correspondence between clusters from full data versus sampled data.

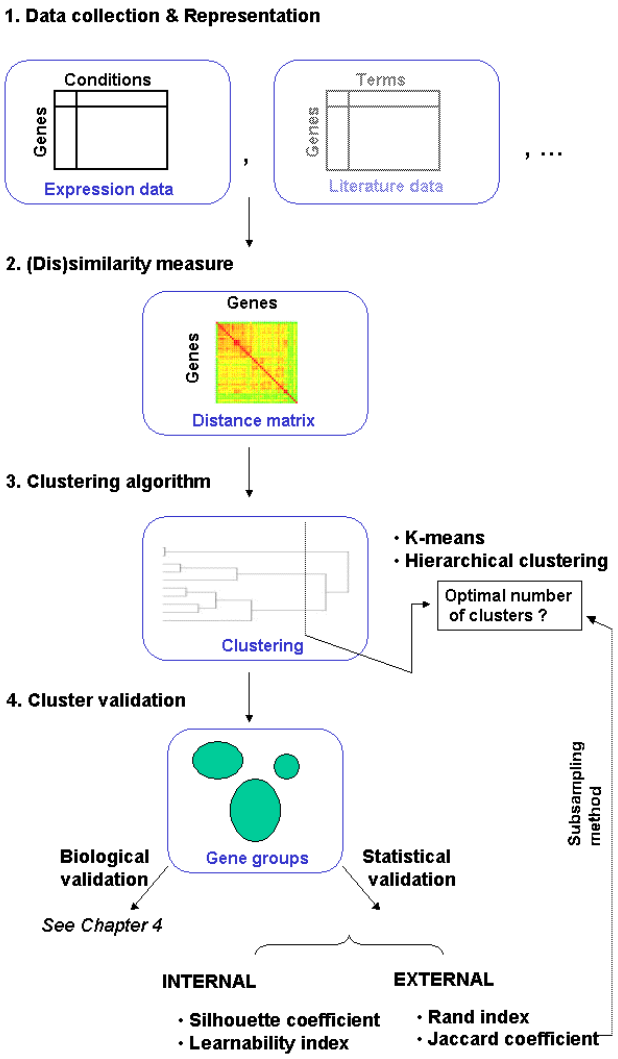


Figure 3.2: Overview of the methodology underlying clustering analysis. **1.** Data should be collected, appropriately preprocessed and poured into a representation. We will apply the methodology on microarray gene expression data and gene literature information. **2.** A distance measure should be chosen to quantify the relationship between genes. We note that in K-means clustering the Euclidean distance measure is intrinsically connected to the algorithm. **3.** As clustering algorithms we use K-means clustering and hierarchical clustering. Both require a choice of the number of clusters k. However, this is not always straightforward and at times misleading. **4.** Assessment of cluster quality can be done based on the data that generated the result (i.e., 'statistically') or using external information (i.e., 'biologically'). For the first type we distinguish internal and external measures. Some of these measures can be integrated in step **3**, thus giving rise to a new clustering method.

3.3 Prevalent approaches to clustering microarray data

Gene expression measurements are typically organized in a $n \times p$ matrix (Figure 1.2) where each rowvector $x_{i\cdot}$ represents the quantification of the level of expression of a single gene (out of n genes) over p conditions. This data structure is central to microarray analysis. The (row-wise) mean and (row-wise) variance vectors are defined as

$$\begin{cases} \overline{x_i} = \frac{1}{p} \sum_{l=1}^{p} x_{il} \\ \sigma_i^2 = \frac{1}{p} \sum_{l=1}^{p} (x_{il} - \overline{x_i})^2 \end{cases} \tag{3.1}$$

3.3.1 Preprocessing issues

Log transformation

Since we are looking at expression ratios (cfr. Section 2.2), we will get the profiles in an asymmetrical scale: over-expressions will have values between 1 and infinity while under-expressions will be between 0 and 1. In order to give the same weight to both over-expressions and over-repressions we need to transform the scale. The most common function used for this task is the log-transformation.

Missing value estimation

One of the characteristic of the gene expression profiles is the existence of missing values in the data set. The excess of missing values can be a problem for methods where calculations are based on a distance matrix. Measurements absent in the data matrix because of technical reasons are commonly replaced by zero, an average expression value or by using techniques as k-nearest neighbor [154] or singular value decomposition [52]. Additionally, profiles exceeding a threshold of missing values (e.g., expressed as percentage of measurements) in the gene expression matrix can be ignored. The way we treat missing values can be found on the supplementary website [24] of De Smet *et al.* [67].

Gene selection

Whatever data mining techniques performed on the data, it is very useful to limit the analysis to those genes showing most variability between two or more samples. Inclusion of many constitutive genes (i.e., genes not showing variation across the studies) in data analysis could not only put a needless burden on the final interpretation, their small standard deviations (3.1) interfere with standardization (3.2) as divisions by small, random values give rise to random profiles that disturb the clustering process. Also, we will get strange results when clustering if we use a correlation metric (see below) as flat patterns can correlate with nearly anything. Therefore, a filtering of profiles not attaining a minimum variance threshold has become common practice.

Standardization

Finally, to compare expression profiles, relative change of gene expression is more important then absolute values. Hence for an easier comparison profiles $\{x_i\}_{l=1..p}$ are often rescaled to have zero mean and unit variance (standardization):

$$x'_{i\cdot} = \frac{x_{i\cdot} - \overline{x_i}}{\sigma_i} \tag{3.2}$$

Normalization of an vector i refers to rescaling the values so $\sqrt{\sum_{l=1}^{p} x_{il}^2} = 1$ If the row-wise mean (3.1) equals zero, normalization is equivalent to standardization up to a factor p.

3.3.2 Distance metrics

In any clustering algorithm the notion of distance between objects is crucial to placing them into groups. A very common distance metric is the Euclidean distance, defined as

$$d(i,j) = d_{ij} = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{jl})^2}, \tag{3.3}$$

which fulfils the four requirements of a metric:

1. $d_{ij} \geq 0$,
2. $d_{ii} = 0$,
3. $d_{ij} = d_{ji}$,
4. $d_{ij} + d_{jk} \geq d_{ik}$ (triangular inequality),

for all i, j, k.

Distance measures that satisfy all but condition (4) are called semi-metric distances and are also often used in expression analysis. Cosine- and Pearson-based distances are examples of these and are based on the similarity that exist between two objects i and j, denoted $s(i,j)$ or s_{ij}. Pearson correlation is defined as:

$$s_{\text{Pearson}}(i,j) = \frac{\sum_{l=1}^{p} (x_{il} - \overline{x_i})(x_{jl} - \overline{x_j})}{\sqrt{\sum_{l=1}^{p} (x_{il} - \overline{x_i})^2} \sqrt{\sum_{l=1}^{p} (x_{jl} - \overline{x_j})^2}}, \tag{3.4}$$

whereas the cosine between two vectors is

$$s_{\text{Cosine}}(i,j) = \frac{\sum_{l=1}^{p} x_{il} x_{jl}}{\sqrt{\sum_{l=1}^{p} x_{il}^2} \sqrt{\sum_{l=1}^{p} x_{jl}^2}}. \tag{3.5}$$

When the data is mean-centered Pearson and cosine are equivalent. One way to transform similarities (3.4) and (3.5) to (semi-metric) distances is via

$$d_{ij} = 1 - |s_{ij}|. \tag{3.6}$$

Taking the absolute value means that correlation and anti-correlation are treated similarly. In gene expression analysis this mapping is supported by the following reasoning: if one is interested in genes belonging to the same biochemical pathway, genes that are over- or underexpressed in a similar fashion should be treated similarly. Replacing $|s_{ij}|$ by $\frac{s_{ij}+1}{2}$ in (3.6) will separate patterns of over- and underexpression.

Applying distance measures on expression data does not always yield intuitive results. In Chapter 4 methods that evaluate choices such as an optimal distance measures are presented. Euclidean distance measures focus on absolute expression values whereas the measures based on the angle between vectors tend to rather follow trends in the data. Also, the microarray platform used as well as the way the data is preprocessed exerts a significant influence on the behavior of the measures. In Figure 3.3 we show an example of how the Euclidean (3.3) and (absolute valued) cosine-based (3.6) measures quantify distances between two pairs of standardized expression profiles.

3.3.3 Hierarchical methods

Hierarchical clustering organizes the data into a binary tree in which the leaf nodes are the objects and the interior nodes, or branch points, are candidate clusters. We distinguish agglomerative and divisive strategies.

Agglomerative hierarchical clustering

The starting point is a pairwise dissimilarity matrix, D, containing distances between all n points using a given metric, and a disjoint clustering that places each object to a separate cluster. Gradually (starting from singleton clusters), nearest clusters are merged pairwise into a new cluster that is represented as one node in a binary tree called a dendrogram. The merging stops until one cluster is left and the resulting dendrogram contains $n-1$ nodes. The essential difference between various agglomerative, or bottom-up, methods lies in the way how the distances between a newly formed cluster and existing clusters are defined. Suppose that at a given stage clusters r and s merge into cluster rs. The distance, $d(\cdot,\cdot)$, of rs with respect to remaining clusters k is generally computed as:

$$d(rs,k) = \alpha_r d(k,r) + \alpha_s d(k,s) + \beta d(r,s) + \gamma |d(k,r) - d(k,s)|, \tag{3.7}$$

with choices for parameters α_r, α_s, β and γ shown in Table 3.1. Intuitively, single link corresponds to taking the minimum distance between clusters, complete link to a maximum distance and average link to a group mean distance. Ward's method measures the change in *within-cluster sum of squares* by merging r and s into rs.

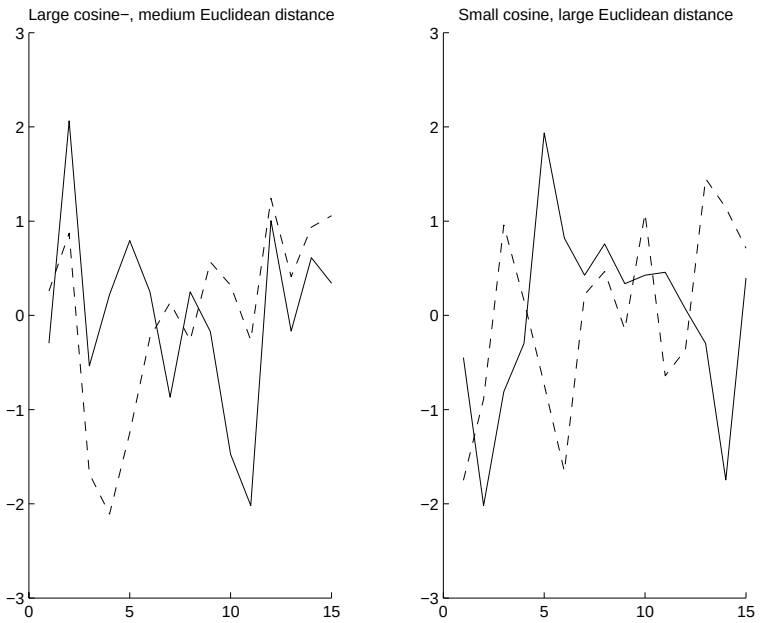


Figure 3.3: Differences between Euclidean and cosine-based $1 - |\cos|$ distance measures. We plot expression levels (y-axis) of a time course experiment in the cell cycle (x-axis). **A.** As it is a absolute quantification, the Euclidean distance is medium (5.37), whereas the cosine distance large (0.9702). **B.** Trends of pronounced anticorrelation are yielding a small cosine distance (0.0343), but a large Euclidean distance (7.226). Values were obtained from standardized data. Hence, using $1 - \cos$ instead of $1 - |\cos|$ yields equivalent results for Euclidean and cosine in this case. Hence, distance measures do not necessarily give intuitive quantifications of relations between expression profiles. Also, preprocessing steps as well as data type exert an important influence on distance measures.

Table 3.1: Overview of agglomerative hierarchical clustering methods. Suppose that at a given stage clusters r and s are merged into cluster rs. The distance, $d(\cdot, \cdot)$, of rs with respect to remaining clusters k is generally computed as in (3.7). Choices for parameters α_r, α_s, β and γ are shown, with n_r, n_s and n_k denoting cluster sizes of the clusters r, s, and k.

Method	α_r	α_s	β	γ
Single link	$1/2$	$1/2$	0	$-1/2$
Complete link	$1/2$	$1/2$	0	$1/2$
Average link	$\frac{n_r}{n_r+n_s}$	$\frac{n_s}{n_r+n_s}$	0	0
Ward's method	$\frac{n_r+n_k}{n_r+n_s+n_k}$	$\frac{n_s+n_k}{n_r+n_s+n_k}$	$\frac{-n_k}{n_r+n_s+n_k}$	0

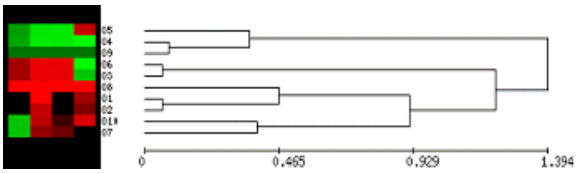


Figure 3.4: Illustration of dendrogram stemming from hierarchical clustering. The x-axis contains the *cophenetic* distances, i.e., the distances between subsequent clusters that are represented as nodes in the tree. Cutting the tree at a given level determines the final number of clusters k, and is far from straightforward. The reordered, color-coded (e.g., red=overexpression; green=underexpression) gene expression data matrix (10 objects (genes) in a four dimensional feature space) is shown on the left side.

As these methods require the full distance matrix (demanding $\mathcal{O}(n^2)$ computations), they exhibit quadratic memory complexity. An illustration of a dendrogram is shown in Figure 3.4. Often in gene expression analysis a reordered, color-coded data matrix is plotted jointly with the dendrogram to allow for visual identification of interesting clusters. For any dendrogram of n elements, there are $2^{(n-1)}$ possible linear reorderings consistent with the structure of the tree. Indeed, at each node of the binary tree either of the two joined elements can be ordered ahead of each other. Simple methods as distance to average expression level and time of maximal induction [73], or more advanced strategies [43] are available. Finally, a cutoff in the dendrogram which leads to final cluster assignments are commonly done by visually inspecting the clustered data matrix or the dendrogram. However, it is known that dendrograms can at times be misleading indicators for cluster structure [96]. Therefore, more principled methods for automatically determining the number of clusters as in Horimoto *et al.* [92] or in Section 3.5 are useful.

Divisive hierarchical clustering

Less common, top-down approaches to hierarchical clustering were developed in Pollard and van der Laan [125] and by Alon *et al.* [38]. To build up a binary tree one starts from a single cluster that is split up in k subclusters with $k = 2$ using a non-hierarchical method (such as K-medoids, see Section 3.3.4). More generally, generalization from binary to k-ary trees are also possible in which case an optimal k needs to be determined (e.g, using a *Silhouette* coefficient, see Section 3.4). Subsequently, a next level of the tree is constructed from the k clusters using the same procedure and the process is repeated until $k = n$ (singleton clusters). Refinements to merge clusters that are not near each other in the dendrogram, but yet very similar can be found in Pollard *et al.* [125].

3.3.4 Non-hierarchical methods

K-means

The K-means algorithm [146] adopts the Euclidean distance measure and minimizes the total of sum of squares within each cluster C_k in an iterative manner over proposed cluster assignments:

$$\min_{C} \sum_{k=1}^{K} \sum_{i \in C_k} |x_{i\cdot} - \overline{x_{k\cdot}}|^2 \tag{3.8}$$

with $\overline{x_{k\cdot}}$, for example, the average gene expression vector over all vectors $x_{i\cdot}$ in cluster k, and i running over all elements in cluster C_k. The parameter K denotes the number of clusters and is fixed in advance. Solutions are usually approximated using gradient descent optimization and the resulting computational complexity is $\mathcal{O}(Kn)$.

K-medoids

Contrary to K-means, the K-medoids algorithm [146] takes a full distance matrix (requiring $\mathcal{O}(n^2)$ computations) as input and attempts to solve following optimization problem:

$$\min_{C} \sum_{k=1}^{K} \sum_{i \in C_k} d(x_{i\cdot}, m_{k\cdot}), \tag{3.9}$$

with $\{m_1, ..., m_K\}$ the corresponding representative points (called medoids) of each cluster, and i running over all elements in cluster C_k. Also here, K denotes the number of clusters and is fixed in advance. One advantage of this algorithm over centroid-based methods, such as K-means, is that each medoid is a robust representative data point for each cluster. As with hierarchical clustering, various metrics can be tried out through the distance matrix. However, K-Medoids is computationally far more intensive than K-means as the determination of the medoid m_k is of order $\mathcal{O}(n_k)$ with n_k the number of elements in cluster k. Also, the heuristic to solve the optimization problem is not as straightforward as with K-means. For more details we refer to Kaufman and Rousseeuw [100].

3.4 Statistical cluster validity

To measure the performance and quality of the clustering one defines four scores: the *Silhouette* coefficient, the performance of the clustering as a classifier, and the *Jaccard* and the *Rand* index. The first two are termed 'internal' scores as they rely on statistical properties of the clustered data, the final two are called 'external' because they involve a comparison with known, external labelings.

3.4.1 *Silhouette* coefficient

As a first internal score for cluster quality we use the *Silhouette* coefficient. It is a metric-independent measure designed to describe the ratio between the concepts cluster coherence (*intra* cluster distance) and cluster separation (*inter* cluster distance) and to assist in choosing which clustering is preferable according to the data. The *Silhouette* value for each element i in cluster k is

$$sc_{ik} = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \tag{3.10}$$

where $a(i)$ is the average dissimilarity of member i to all other members of its cluster and $b(i)$ the dissimilarity of member i to the nearest member of the nearest cluster. With n_k the size of cluster k and n the total number of objects, the *Silhouette* coefficient per cluster is defined as

$$SC_k = \frac{1}{n_k} \sum_{i=1}^{n_k} sc_{ik}, \tag{3.11}$$

and the overall *Silhouette* coefficient becomes

$$SC = \frac{1}{n} \sum_{k} \sum_{i=1}^{n_k} sc_{ik}. \tag{3.12}$$

SC takes on values between -1 and 1.

3.4.2 *k*-NN learnability

To establish a second measure of internal cluster quality, we translate cluster validity into classification performance as in Pavlidis *et al.* [121]: suppose we use a clustering result that provides a label for all n points in some way, what is the performance of a classifier on a test set ? To this end, labelling, training and test set, performance measure and classification algorithm need to be explained;

Performance To assess the performance, we score each cluster consecutively by means of a misclassification probability, explained below. Performance for all clusters can then summarized by taking the average over the clusters.

Labelling For each cluster in turn, all its members are labelled as positive examples, whereas all members of the remaining clusters bear negative labels.

Training/test set construction For the construction of a training and test set, a leave-one-out cross-validation is used [144][52]. Each element of the cluster that is being scored is consecutively considered as a singleton test set, and the remaining $n - 1$ elements of the whole data set function as the training set.

Classifier A k-Nearest Neighbor[1] (k-NN) classifier [52] with the $(1 - |\cos|)$ distance measure is used. k-NN is particular among classification methods in that there is no explicit training step. Hence, leave-one-out cross-validation amounts to applying k-NN for each element of the cluster and computing misclassification probability for that cluster.

Misclassification probability The probability of misclassification is

$$P(i \in \mathcal{P}|i \in \mathcal{N})P(i \in \mathcal{N}) + P(i \in \mathcal{N}|i \in \mathcal{P})P(i \in \mathcal{P}), \tag{3.13}$$

where $\mathcal{P}$ is the positive class (i.e., members of the cluster under consideration) and $\mathcal{N}$ the negative class (i.e., members of other cluster). k-NN is computationally inexpensive as no training step is required and as $P(i \in \mathcal{P}|i \in \mathcal{N})$ and $P(i \in \mathcal{N}|i \in \mathcal{P})$ can be calculated analytically in this case. Indeed, for each element i that is consecutively considered as a singleton test set, assignment to the positive class $\mathcal{P}$ requires that more than $\frac{k}{2}$ neighbors of i are members of that cluster. Hence,

$$P(i \in \mathcal{N}|i \in \mathcal{P}) = \sum_{l=0}^{\lfloor \frac{k}{2} \rfloor} \frac{\binom{n_{\mathcal{P}}-1}{l}\binom{n-n_{\mathcal{P}}}{k-l}}{\binom{n-1}{k}}, \tag{3.14}$$

which is a hypergeometric distribution with $n_{\mathcal{P}}$ the size of the positive class and n the total number of objects. Substituting (3.14) into (3.13) yields a probability of misclassification that measures how learnable that cluster is in terms of the adopted k-NN classifier.

3.4.3 *Jaccard* coefficient

The first measure of external cluster validity is the *Jaccard* coefficient Given a set of n points, an external partition $P = \{P_1, ..., P_u\}$, and a clustering $C = \{C_1, ..., C_v\}$, we define for each partitioning a matrix M with components

$$M_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j\text{belong to the same cluster} \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

A correlation between partitionings P and C is written as

$$\langle P, C \rangle = \sum_{i,j} M_{ij}^P M_{ij}^C. \tag{3.16}$$

Assuming that $N_{11}, N_{10}, N_{01}, N_{00}$ count the number of times M_{ij}^P and M_{ij}^C exhibit the element-wise patterns $\{1,1\}$, $\{1,0\}$, $\{0,1\}$, $\{0,0\}$ respectively, we write following equalities:

$$M_{ij} = \begin{cases} \langle P, C \rangle = N_{11} \\ \langle P, P \rangle = N_{11} + N_{10} \\ \langle C, C \rangle = N_{11} + N_{01}. \end{cases} \tag{3.17}$$

[1] In this subsection, k denotes the number of neighbors.

The *Jaccard* coefficient is now written as

$$J(P,C) = \frac{\langle P,C\rangle}{\langle P,P\rangle + \langle C,C\rangle - \langle P,C\rangle} = \frac{N_{11}}{N_{01}+N_{10}+N_{11}}. \tag{3.18}$$

It quantifies the correspondence between two binary matrices when 'negative' (i.e., zero) matches are ignored. J takes on values between 0 and 1. This derivation presents the *Jaccard* coefficient as a form of correlation between matrices.

3.4.4 *Rand* index

As a second external measure for cluster validity we use the adjusted *Rand* index. Using the same notation from previous section, the *Rand* index is defined by

$$R = \frac{N_{11}+N_{00}}{N_{11}+N_{01}+N_{10}+N_{00}}. \tag{3.19}$$

The correction for random partitioning is $R_{\text{adj}} = \frac{R-E(R)}{\max(R)-E(R)}$, where a hypergeometric baseline distribution is used to compute the expected values. In a comparative study, the adjusted *Rand* index is recommended as the external measure of choice [96]. When the overlap is completely random $R_{adj} \to 0$.

3.5 The optimal number of clusters

In order to apply K-means, K-medoids and hierarchical clustering successfully the number of clusters need to be determined. As mentioned already, in case of hierarchical methods proper visualization can be helpful, but at times also misleading. Therefore statistical methods capable of determining K automatically are usually applicable to a broad class of algorithms (including the ones that parameterize the number of clusters indirectly, e.g., by requiring an optimal cluster radius). Once statistical 'fitness' of a clustering result is defined, one can examine the effect on it exerted by the parameter settings. The methods outlined in Section 3.4 are primary candidates for methodologies addressing this problem. Jain and Dubes [96] and Pollard *et al.* [124] describe strategies based on within- versus between-cluster distances, using the *Silhouette* coefficient or similar scores. Yeung *et al.* [165] use a learnability score termed Figure of Merit (FOM) that is inspired on the principle of leave-one-out cross validation in supervised learning. The data is clustered using expression information over $p-1$ conditions and validated by measuring tightness of the resulting clusters on the left-out condition. In what follows we will present more in-depth the stability-based method from Ben-Hur *et al.* [46] to determine the number of clusters quantitatively and visually.

3.5.1 Stability based method to determine optimal k

The general assumption of the stability method is that cluster structure within a data set remains preserved even if only a fraction (say $f \geq 0.5$) of the data is

used. In other words, inherently present cluster structure is to be stable with respect to subsampling.

Therefore, for each parameter setting (K in this case) correspondences are measured between clusterings conducted on s pairs of random subsamples X_{sub1}, X_{sub2} of the data. Each pair of clusterings C_1, C_2 applied to X_{sub1}, X_{sub2} (with a fixed parameter setting) is compared via the *Jaccard* coefficient (3.4.3). Repeating this s times yields a distribution of coefficients that is converted to a cumulative distribution function (cdf). Iteration over several parameter choices gives several cdf's that can be mutually compared. The expectation is that 'kinks' in the spectrum of cdf's that range over various settings of K reveal where the optimal value is reached. Summarized, the algorithm goes as follows:

Given: a fraction f, nr of subsamples s, maximum nr of clusters K_{max}
Iterate over k=1...K_{max}
    Repeat s times
        $\mathcal{S}_1$=Random subsample representing fraction f of the data
        $\mathcal{S}_2$=Random subsample representing fraction f of the data
        Cluster $\mathcal{S}_1$
        Cluster $\mathcal{S}_2$
        Compute *Jaccard* coefficient J (3.4.3) on clustered $\mathcal{S}_1 \cap \mathcal{S}_2$
Plot cumulative distribution functions of J to determine most stable K.

It is particularly efficient when hierarchical clustering is used as a clustering algorithm. Indeed, once a hierarchy is constructed, computing solutions for various K boils down to cutting the tree at a certain level. Note also that parameterizations can be more general than K and include e.g., stepwise feature selection.

We illustrate the stability method using the Iris data set, a benchmark in the machine learning community. It presents 150 random samples of flowers from three *iris* species. From each species there are 50 observations of four features: sepal length, sepal width, petal length, and petal width. The data is clustered using the Euclidean metric and Ward's method. In Figure 3.5 we show the corresponding dendrogram and stability plot. We computed for each $K = 2 \dots 20$, 150 pairs of subsamples taken from a fraction $f = 0.8$ of the data. Consistent with the reports of Ben-Hur [46] we found these settings sufficient for this type of problem. We see that $K = 2$ is the most stable setting as we observe a gap in the transition to the cdf of $K = 3$. Nevertheless, $K = 3, 4, 5$ remain interesting settings, whereas $K \geq 9$ corresponds to totally unstable settings. The cdf for $K = 3$ - the real number of clusters - is particularly interesting as it shows that some subsamples maintain cluster membership in some cases (i.e., considerable probability mass for *Jaccard* correlations close to one), but notably don't in other cases (i.e., significant probability mass in the lower regions of *Jaccard* correlation). Exploring scatter plots of the Iris data in Figure 3.6 indeed confirm the existence of close, partly overlapping clusters. Hence, visualizing the proposed cdf's and comparing them over various settings is useful to determine optimal K.

3.6 Iterative clustering

As already mentioned in previous sections, combining established clustering algorithms (Section 3.3) with methods for automated parameter estimations (Section 3.5) into a divisive clustering strategy is a powerful approach to tackle data that holsters more complex patterns. Figure 3.7 explains how we adopt a top-down clustering approach when applying the stability method to microarray data. Other examples of this approach can be found in Pollard *et al.* [125], Herrero *et al.* [89] and Alon *et al.*[38].

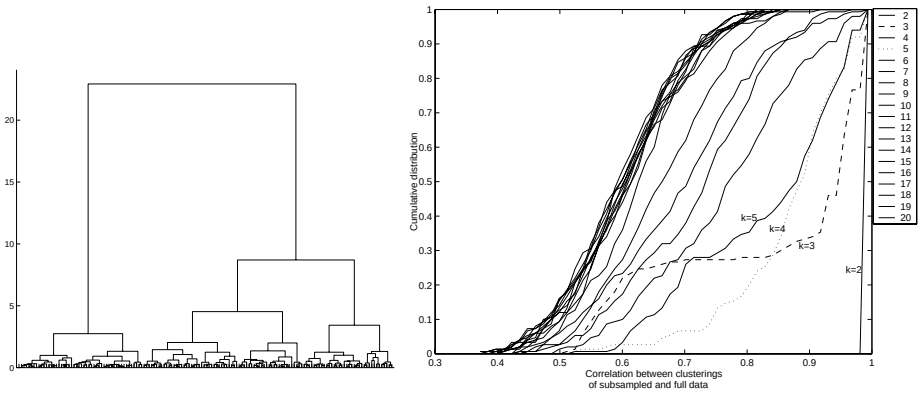


Figure 3.5: Stability-based method to determine optimal number of clusters illustrated on Iris data set. On the left side the dendrogram is shown. Visual inspection reveals the presence of either two or three clusters. The stability diagram plots values of cumulative distribution functions (y-axis), computed for each $K = 2 \ldots 20$, of the *Jaccard* coefficients over pairs of clustered subsamples (x-axis). K=2 seems the most stable solution. For K=3 we see that the cdf has a particular shape indicating that subsampling maintains cluster membership in some cases, but notably doesn't in others. Exploring scatter plots of the Iris data indeed confirm that two clusters are ill-separated.

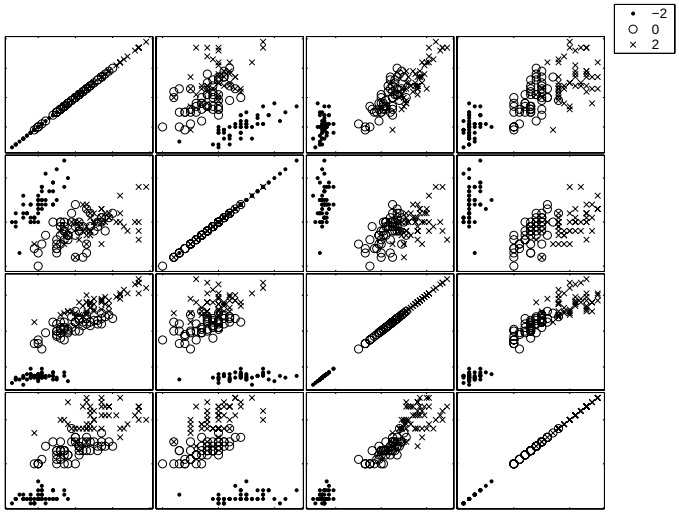


Figure 3.6: Scatter plot of the Iris data for each pair of dimensions (four in total). Although three groups are hidden in the data, only two groups are unequivocally detectable. Determining optimal K on this type of data requires therefore exploratory methods that can reveal such ambiguities.

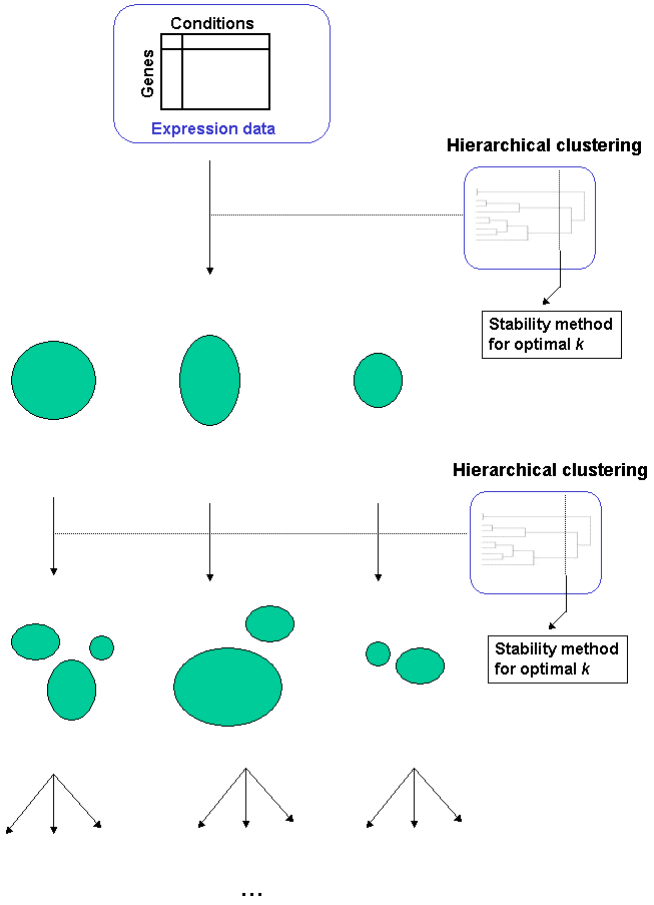


Figure 3.7: Iterative clustering scheme. Starting with a first clustering round, the data is partitioned into several clusters using hierarchical clustering. The optimal number of clusters is determined with the stability method. Each of the clusters are re-clustered with the same algorithm and this process is repeated for each of the clusters until no more 'trustable' subdivisions can be made. The process thus yields a divisive, stability-based hierarchy of clusters.

3.7 Discussion

Underlying a microarray experiment is always a biological question one likes to address. The expression levels of thousands of genes originating from samples taken under a variety of conditions (e.g., various stages of a tumor) or in different stages in a biological process (e.g., phases in the cell cycle) are organized in a gene expression matrix. Although debatable and subject to nuances, the central assumption of gene expression analysis is that genes that share similar expression patterns also have similar functions. A comprehensive functional mapping of genes requires the organization of genes into tractable subgroups containing both characterized and uncharacterized genes. Formulating hypotheses on the function of such unknown genes by transferring knowledge about the known genes in the same group is often referred to as 'guilt by association'. Classical methods such as K-means and hierarchical clustering provide a first way to cluster genes based on their expression profiles.

Different algorithms or even the same algorithm with different parameterizations can give very different clustering results. Measures describing statistical cluster quality can assist in eliminating spurious solutions or help in finding good clustering parameters (Section 3.4). However, the scale of gene expression data and the complexity of the biological problems addressed mandate further improvements on the exploratory analysis. For example, the application of the stability method to real life expression data underestimates the true number of clusters present in yeast cell cycle data (see Chapter 7). Therefore, we use an iterative scheme to drill deeper into clusters and find stable subclusters. On the one hand this can be considered as a modification to the subsampling method, on the other hand it is the implementation of a divisive clustering scheme also adopted by other authors [125][89][38].

Finally, we not only use clustering algorithms to analyze expression data, but also to cluster genes based on literature information. Many of the algorithms presented in this section will therefore consistently recur throughout this dissertation.

Chapter 4

Biological cluster validation

As gene expression studies typically output lists of dozens or hundreds of genes that are co-expressed, typically one is faced with the assignment of biological meaning to those results. Throughout this dissertation we are interested in two aspects of 'biological meaning' (See Figure 4.1). Firstly, we wish to *functionally* characterize genes or gene clusters and, as treated in the following chapters, we will use scientific literature as main information source to this end. However, for its impact on both the database development and data analysis communities, we will also discuss the use of the Gene Ontology (GO) [149] for functional annotation in Section 4.1. The rest of databases used throughout this work are described in the Appendix. Secondly, we are interested in extracting evidence for *transcriptional regulation* from clusters of co-expressed genes. In Section 4.2 we propose a novel motif-based score that provides a quick estimate of the biological relevance of multiple clustering solutions[1]. We demonstrate its behavior both on synthetic and real data and it will be used later on in Chapter 7 to assess our integrated clustering framework.

4.1 Gene ontology

4.1.1 Description

From the multitude of publicly available biological resources [44], the Gene Ontology (GO) is one of the most intriguing, not in the least because of its influence on the development of biological databases. Historically, bioinformatics has largely grown from efforts to deal with increasingly large amounts of data produced by molecular biology in the form of protein or DNA sequences [116]. This sort of data can be modelled straightforwardly as a list of characters, and subsequently searched, stored, and manipulated by a computer. During the development of the many repositories that store these data, a large

[1]This score was presented in a broader context (see Chapter 7) in the Special Issue on Microarray Data Mining from the journal *SIGKDD Explorations* [83].

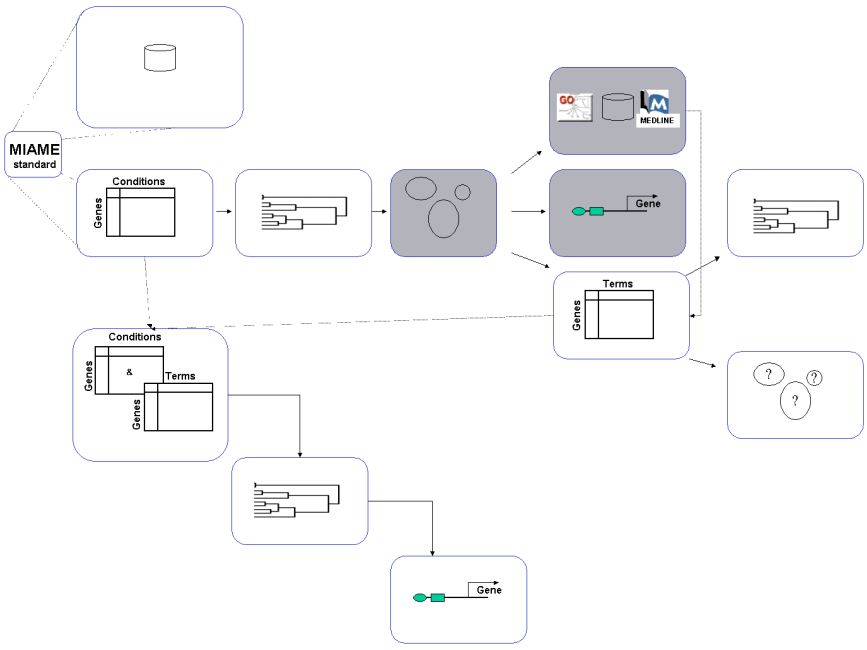


Figure 4.1: Experimentally defined groups of genes are approached in two ways in this work. Firstly, functional relatedness can be corroborated through an interactive use of external information sources including Gene Ontology (GO). Because of its central role in functional genomics and its use in later chapters, we will discuss GO here. Secondly, we are interested in extracting evidence for *transcriptional regulation* in clusters of genes. We have developed a motif-based score (M-score) that assesses biological relevance of an entire clustering solution in terms of DNA patterns that regulate the expression (or more precisely the 'transcription') of the genes.

amount of annotation has been associated with these sequences. This ranges from semi-structured data, such as sequence or species information, to unstructured descriptions in free text format.

SWISS-PROT, for example is often described as a protein sequence database, but due its descriptive nature, it could also be considered to be a protein annotation database. This has served the community well in the past, when the annotation was meant for humans to read. However it causes difficulties when trying to analyze annotations computationally as in high-throughput screening. While such free text is accessible by computer applications, it is not easy to interpret it automatically. In Chapter 5 we will discuss text mining methods to cope with free text information and the forthcoming chapters will deal largely with integrating such information in various ways in large scale gene analysis. In part for similar reasons, there has also been a growing interest in ontologies within bioinformatics. Ontologies provide a mechanism for capturing a community's view of a domain in a shareable form, that is both accessible by humans and computationally manageable. It provides a set of vocabulary terms that label domain concepts. These terms should have definitions and be placed within a structure of relationships, the most important being the 'is-a' relationship between parent and child and the 'part-of' relationship between part and whole [106].

GO is a collaborative effort to address the need for consistent descriptions of gene products in different databases. However, it is not a way to unify biological databases. Rather, GO describes by means of controlled, structured vocabularies how gene products behave in a cellular context and facilitates the process of extracting biological insight from enormous sets of data. At the highest level GO consists of three branches that describe biological processes, cellular components and molecular functions in a species-independent manner. *Molecular function* describes activities, such as catalytic or binding activities, at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. A *biological process* is accomplished by one or more ordered assemblies of molecular functions. A *cellular component* is a component of a cell that it is part of some larger object which may be an anatomical structure (e.g., rough endoplasmic reticulum or nucleus) or a gene product group (e.g., ribosome or proteasome). In Figure 4.2 we show an example.

4.1.2 Assessment of functional enrichment in gene groups

Although the range of bioinformatic applications in which GO takes an important position is growing rapidly, we focus here on the particular use of GO to understand groups of genes as they are generated by microarray data analysis. As GO provides an association of terms to gene products, one basic feature is to use this information to look for correspondences between microarray gene clustering results and GO terms. In Table 4.1 we exemplify the GO annotations of three co-expressed genes from a microarray experiment in Spellman *et al.*

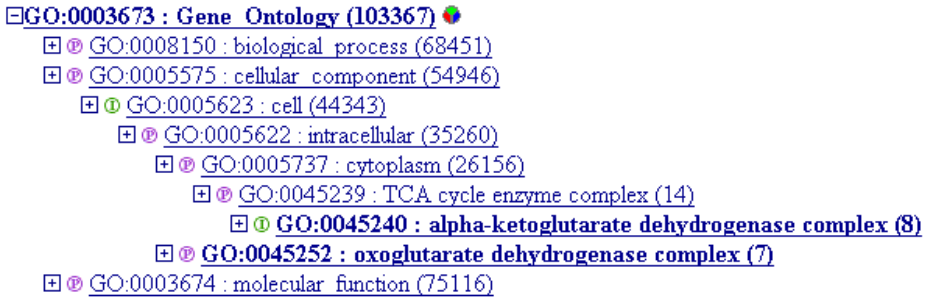


Figure 4.2: Example of Gene Ontology Tree. Location of the alpha ketoglutarate dehydrogenase complex in cellular component branch. At the highest level GO consists of three branches that describe biological processes, cellular components and molecular functions in a species-independent manner. *Molecular function* describes activities, such as catalytic or binding activities, at the molecular level. A *biological process* is accomplished by one or more ordered assemblies of molecular functions. A *cellular component* is a component of a cell that it is part of some larger object which may be an anatomical structure (e.g., cytoplasm) or a gene product group (e.g., TCA cycle enzyme complex).

[141]. The GO annotations suggest that this experimental expression cluster groups gene products involved in the biological process of protein folding. In contrast, the molecular function and cellular component annotations of these gene products correlate less well with the clustered expression patterns of these gene products. This illustrates the rationale to split up GO according to three components, each describing another aspect of biological function.

Given a list of over- or underexpressed genes one can calculate the significance of enrichment (expressed as a p-value) of each observed GO category with the binomial distribution $\mathcal{B}$,

$$p\text{-value} = 1 - \sum_{l=1}^{r} \mathcal{B}(l, n, p_{backgr}), \tag{4.1}$$

with n the number of co-expressed genes, r the number of co-expressed genes having the GO category associated to them, and p_{backgr} the prior chance to draw genes from this category or its descendants in the GO tree. For a more detailed discussion we refer to Zeeberg *et al.* [168]. See also GO4G in Aerts [35].

4.1.3 Analysis of entire clustering solutions using GO

The number of clusters in expression clustering constitutes a parameter that heavily influences the biological relevance of the outcome. To overcome the limitations of data-centric measures (Section 3.4) for determining the optimal number of clusters, Gibbons *et al.* [79] developed a method to determine the

Table 4.1: Example of GO annotations according to its three main taxonomies - biological process, molecular function and subcellular location- of three co-expressed genes from microarray data by Spellman *et al.* [141]. Note that multiple annotations per gene are possible.

Gene name	Biol process	Mol function	Cell location
HSP78	protein folding	chaperone	mitochondrial matrix
HSP104	protein folding	cochaperone	cytoplasm
CPR6	protein folding	peptidylprolyl isomerase	cytoplasm

optimal number of clusters in a gene expression data set by using a 'figure of merit' that uses GO annotations. Using a database of gene annotations that describes for each gene i, the presence of an attribute (a GO term) j, they basically define biological significance of a cluster result C as the total mutual information (MI) between C and all attributes A_j:

$$MI(C, A_1, A_2, \ldots, A_{N_A}) = N_A H_C + \sum_i H_{A_i} - \sum_i H_{A_i C}, \tag{4.2}$$

with $H_{A_i C}$ the entropy for each cluster-attribute pair, H_{A_i} the entropy for each of the N_A attributes independent of the cluster and H_C the entropy of the clustering result. Their final figure of merit is expressed as a z-score

$$\frac{MI - \text{Ave}(MI)}{\text{Std}(MI)}, \tag{4.3}$$

measuring statistical significance of Equation 4.2 with respect to random cluster assignment. Ave(MI) and Std(MI) are the mean and standard deviation of the quantity MI computed on random samples. Attributes that occurred exceedingly often or rarely - irrespective of the clustering outcome - were filtered. Also, attributes were compared pairwise using mutual information to eliminate the occurrence of highly dependent attributes.

Studying four public datasets (two ratio-based, two Affymetrix-based) using their score, they recommend to use the Euclidean distance for ratio-based or Pearson distance for non-ratio-based measurements. Their figure of merit versus the number of clusters, k, applied to data from Cho *et al.* and Spellman *et al.* are reprinted in Figure 4.3. For both Cho and Spellman data sets, they report the performance of single link hierarchical clustering (with Pearson distance measure 3.4) to be worse than random, with average linkage faring only slightly better. The 'best' (i.e., most significant, and highest z-score) distance measures (including K-means with Pearson distance measure) shows optimal value for k between 7 and 10 for the Cho data. For the Spellman data, Euclidean and Pearson distance on K-means perform comparably and show optimal k around 20. The peak-to-average ratio, however, is poor in both cases.

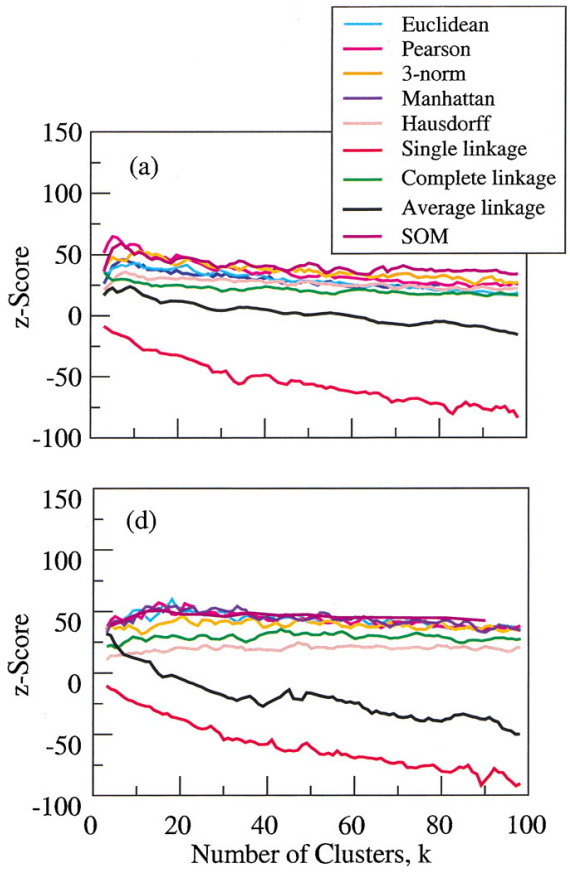


Figure 4.3: Figure of Merit from Gibbons *et al.* applied to expression data of Cho *et al.*(indicated as (a)) and Spellman *et al.* (indicated as (d)) to determine optimal number of biologically relevant clusters using various similarity measures. The figure of merit is expressed as a basic z-score. Performance of single link hierarchical clustering (with Pearson distance measure (3.4)) is reported to be worse than random, with average linkage faring only slightly better. For the Cho data the optimal value for k between 7 and 10 . For the Spellman data, Euclidean and Pearson distance on K-means perform comparably and show optimal k around 20. The peak-to-average ratio, however, is poor in both cases. (Illustration adapted from Gibbons *et al.* [79]).

The work of Gibbons *et al.* is particularly interesting for three reasons. Firstly, they share our rationale to develop a score that evaluates an entire clustering solution in a more biological fashion. In Section 4.2.2 we will present our M-score that originated from similar motivations. Their analysis on Cho and Spellman provide useful reference points for our analyses presented in Section 4.2.2 and Chapter 7. Secondly, the use of gene annotations to complement expression data lies also at the heart of the research presented in this thesis (Figure 1.11). Whereas the work of Gibbons and colleagues uses a vector model with GO attributes and defines biological coherence using mutual information, we will encode unstructured information about genes in the literature using a resembling vector representation. Finally, their work illustrates well the potential of GO for quantitative approaches in high throughput gene analysis. In Chapter 6 we will demonstrate how we use GO as an approach to text-based knowledge.

We conclude with mentioning the existence of many more applications of GO in biology and data mining, and a complete listing can be found on the GO website [3].

4.2 Upstream regulatory analysis

Microarray studies have driven the search for common features in the regulatory mechanism of genes that behave similarly under certain conditions. As it seems reasonable to hypothesize that genes with similar expression profiles are also coregulated, clustering together genes based on expression can therefore reveal putative regulatory signals (or 'motifs'). Figure 4.4 illustrates the idea. Although many approaches exist to search for patterns in upstream sequences of groups of genes, they can be divided in two main categories: string-based and matrix-based methods. The former include oligonucleotide frequency analysis [156][157], suffix trees [109] and graph-based methods [145]. The latter encompass inherent probabilistic methods such as Expectation Maximization [58][42] and Gibbs sampling [151][64]. Both string- and matrix-based methods can be used to locate instances of known motifs or to detect novel motifs.

Throughout this dissertation we will restrict ourselves to string-based searches for previously established motifs within upstream regions of co-expressed genes. We refer to Thijs [150] for advances in motif finding.

4.2.1 Motif-based score to evaluate various clustering runs

As clustering gene expression data constitutes an 'ill-posed' problem in the sense that definite objectives are often hard to define, labor-intensive biological interpretations usually start from educated guesses on good clustering parameterizations (e.g., initial points, number of clusters,...). Especially when experimenting *in silico* with various methodologies or parameterizations, quantitative methods for cluster validation can be of great help in choosing 'good' solutions. A wide range of techniques to validate genome-wide clusterings al-

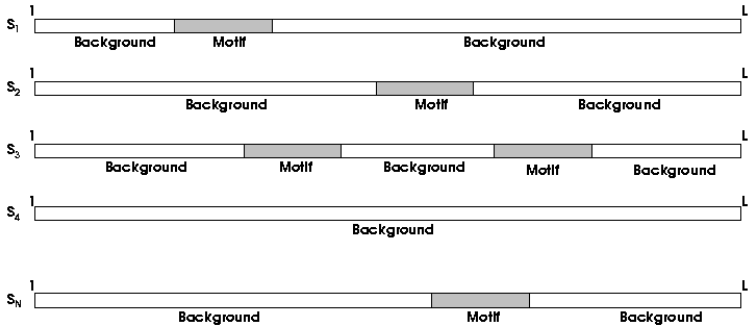


Figure 4.4: Schematic representation of the upstream region of a set of co-expressed genes. Upstream sequences of length L from N co-expressed genes are scanned for the presence of recurring patterns or motifs (adapted from Thijs *et al.*).

ready exist and each formulates the optimality principle differently. On the one hand, *data*-based scores such as the *Rand* index [166], the *Silhouette* coefficient [100][125] or the local stability-based method [46] (see Section 3.4), the Figure Of Merit [165] and the *Gap* statistic [153], estimate good solutions based on the statistical properties of the clustered data. However, these classes of methods suffer the drawback that validation is performed on the same data that produced the clusters, without taking into account biological constraints. On the other hand, an increasing amount of tools are becoming available that link cluster results to information from other sources and compare them (GOMINER [13] [168],GeneXpress [11] [135],INCLUSIV [15] [64], TXTGate [33](see also Chapter 6), ...).

Over the last years there has been a great interest in detecting regulatory signals (or motifs) in upstream regions of coexpressed genes [56][65][152][160]. However, exploring multiple clustering results over various parameterizations involves a lot of overhead in assessing biological relevance to each of the results (for example, see [77], where authors went to great lengths to validate their clustering method in terms of regulation patterns). We observe that

1. Purely data-based scores do not necessarily correlate well with clustering solutions that yield significant and good motif patterns and,

2. Most current ways to evaluate gene groupings in terms of motifs are limited to a manual investigation of statistically enriched clusters.

To the best of our knowledge, none of them provide a one-shot estimate of the relevance of all patterns found in the upstream promotor regions in a clustering solution. This supported our motivation for the development of a *motif*-based heuristic to economize on biological validations when the parameter-space is prohibitively large. We present our motif-inspired cluster validation score, abbreviated as M-score, for quickly estimating the validity of multiple clustering runs in terms of significant upstream sequence patterns. For similar reasons of

Table 4.2: Contingency Table C. Elements of this table in the text are denoted by c_{ij} $(i, j = 1, 2)$.

	Motif	¬Motif
Cluster	r	$n - r$
¬Cluster	$m - r$	$N - n - m + r$

validation overhead, we decided to biologically check our score by focusing on a well-understood data set, for which we have accumulated support from both published results and the expertise of a yeast biologist.

In Section 4.2.2 we treat several ways to obtain p-values that describe the statistical overrepresentation of motifs in a clustered data set. We state our biological assumptions and propose the derived M-score in Section 4.2.2. We test the score, and some variations, on synthetic data. Subsequently, we demonstrate how it can be used to determine the optimal number of clusters (in terms of motif enrichment of the solution) using two cluster algorithms on cell cycle expression data. We complement the discussion of the score's performance with a biological evaluation.

4.2.2 M-score

Our proposed M-score is used to assess the quality of an overall clustering result. It starts from values describing the significance of a set of defined (or found) motifs in each of the clusters. In this section, we therefore first describe how these significance values are computed, followed by a description of the M-score itself.

Statistical overrepresentation of known motifs

Starting from a predefined set of M Motifs and a partitioning of a data set into k clusters, we define a cluster-by-motif matrix A of size $k \times M$. Each entry contains a p-value that expresses the statistical enrichment of genes within a cluster that contain the given motif within 800bp upstream of the ORF. We define a 2×2 contingency table $\{C(i, j)\}_{i=1,2;j=1,2}$, as in Table 4.2, where r is the number of genes containing the motif, n the cluster size, m the total number of genes that have one or more copies of the given motif upstream, and N the total number of genes in the data set. We use the information captured in the contingency table to derive three options for one-sided p-values that test the hypothesis whether the motif enrichment in the given cluster is higher than what is expected by chance:

a. Binomial approximation to the hypergeometric distribution When sampling without replacement, the significance of observing at least r genes whose upstream region contain one or more copies of a given motif

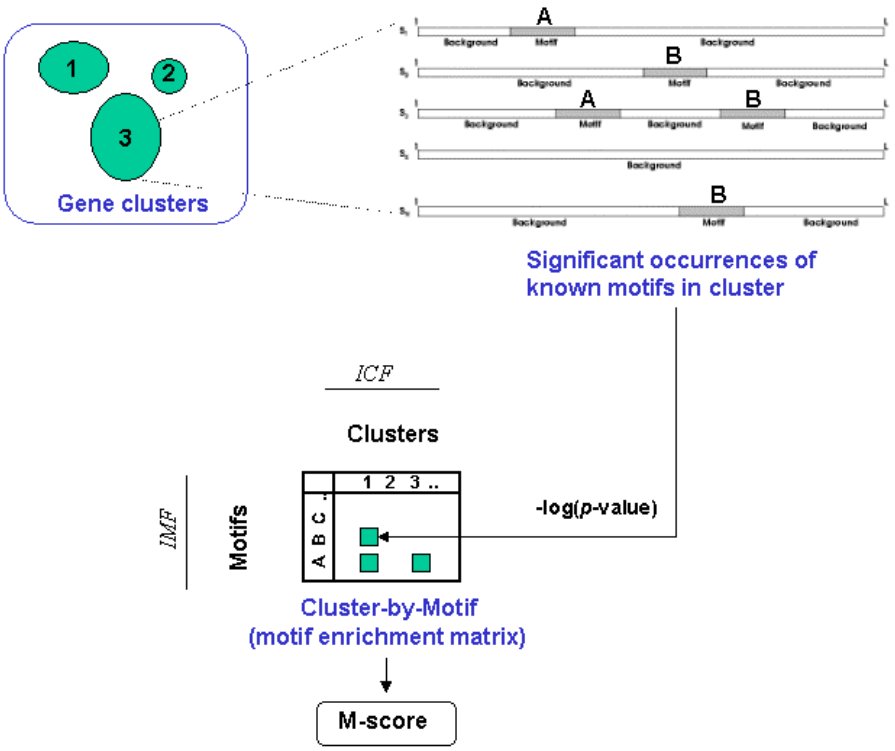


Figure 4.5: Overview of the M-score. To assess biological relevance of clusters in terms of transcriptional regulation (see Figure 2.3), significant occurrences of known motifs are computed using string-based methods. These are poured into a motif-by-cluster matrix as p-values, which be visually inspected for enrichment of motifs in clusters. Based on simple biological assumptions, a summarizing 'M-score' is developed to assess biological relevance of the entire clustering in one shot. The score tries to balance between groupings with clusters containing 'too much' significant motifs (by downweighing the corresponding column with IMF) on the one hand, and groupings in which motifs are too much distributed over all clusters (by downweighing the corresponding row with ICF) on the other hand.

sequence within a cluster of n genes is given by

$$1 - \sum_{l=0}^{r-1} \frac{\binom{m}{l}\binom{N-m}{n-l}}{\binom{N}{n}}, \tag{4.4}$$

the hypergeometric distribution As this quantity is hard to compute for $r \approx 50$, we use the binomial approximation of the hypergeometric which holds when $N >> n$. The significance then becomes

$$p\text{-value} = 1 - \sum_{l=0}^{r-1} \mathcal{B}(l, n, \frac{m}{N}), \tag{4.5}$$

with $\mathcal{B}$ the binomial distribution. We note that the background probability, $\frac{m}{N}$, is data-dependent since it involves only those genes that remain, for example, after filtering genes with low variance from expression data. This method is used for example in Gasch and Eisen [77].

b. Pearson χ^2 test for independence The Pearson χ^2 statistic for contigency table C (Table 4.2) is defined as

$$G^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(c_{ij} - \frac{c_{i\cdot}c_{\cdot j}}{N})^2}{\frac{c_{i\cdot}c_{\cdot j}}{N}}, \tag{4.6}$$

where $c_{\cdot j}$ and $c_{i\cdot}$ denote row- and columnsums respectively. This statistic measures the independence of the random variables *motif occurrence* and *cluster membership*. G^2 is χ^2-distributed with $(2-1)(2-1)$ degrees of freedom and we assume that large values indicate an overrepresentation of the given motif in a certain cluster with at the one-sided significance level

$$p\text{-value} = 1 - \chi^2(1). \tag{4.7}$$

The distributional properties of (4.6) are based on the assumption that the underlying sampling model is multinomial [98]. Accordingly, in the numerical simulations we conducted, we found the Pearson χ^2 test-statistic less stringent (i.e., smaller p-values) than when modelling explicitly the binomial (4.5).

c. Fisher Exact test for small sample sizes When any of the entries in C becomes smaller than 8, we use the Fisher Exact test, which is more robust to small sample sizes [107]. The hypergeometric probability for C (Table 4.2) is written as

$$p\text{-value} = \frac{\prod_{i=1}^{2} c_{i\cdot}! \prod_{j=1}^{2} c_{\cdot j}!}{N! \prod_{i,j=1}^{2} c_{ij}!}. \tag{4.8}$$

The one-sided version of Fisher's method requires the construction of the set of all possible 2×2 contingency tables, $\{D^l\}_l$, each table having elements d^l_{ij} $(i, j = 1, 2)$, that are mutually different and satisfy the condition

$$|\frac{d^l_{11}}{d^l_{11} + d^l_{12}} - \frac{d^l_{21}}{d^l_{21} + d^l_{22}}| \geq |\frac{c_{11}}{c_{11} + c_{12}} - \frac{c_{21}}{c_{21} + c_{22}}|.$$

For all contingency tables $\{D^l\}$, we derive p-values as in (4.8), and sum them to obtain a final p-value.

Multiple testing

As the significance of each motif is tested along multiple clusters simultaneously, we correct each of the p-value entries in A for multiple testing with the Bonferroni method,

$$p\text{-value}^* = \min(p\text{-value} \cdot k, 1). \tag{4.9}$$

M-score

All information about the p-values contained in a cluster-by-motif matrix A is compressed into an M-score, which measures the amount of biological evidence (based on motif analysis) present in one single clustering run.

Before we proceed, we formulate the three biological assumptions it is built on. Given a cluster-by-motif matrix A as defined in Section 4.2.2, we assume that

- a motif is less interesting when it (significantly) occurs in many clusters;
- provided the set of M motifs is large enough, a cluster that contains a large proportion of the motifs is less likely to be biologically relevant;
- a 'too large' number of clusters is less likely to reflect the true biological diversity underlying the experiment.

The proposed heuristic balances between these three criteria and is composed as follows:

$$\text{M-score} = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{M} \underbrace{\log(\frac{M}{f_{\{1..M\}\in i}})}_{\text{IMF}} \underbrace{\log(\frac{k}{f_{\{1..k\}\ni j}})}_{\text{ICF}} \cdot \underbrace{A(i,j)}_{p-\text{value}}, \tag{4.10}$$

where $f_{\{1..M\}\in i}$ is the number of (significantly) found motifs contained in cluster i, $f_{\{1..k\}\ni j}$ the (significant) occurrences of motif j over all k clusters, and $A(i,j)$ the p-value for the motif j in cluster i.

The term IMF can be seen as an *inverse motif frequency*, ICF as an *inverse cluster frequency*. They balance between groupings with clusters containing too much significant motifs (typically if that cluster is too big) on the one hand, and groupings in which motifs are too much distributed over all clusters (typically if clusters are too small) on the other hand.

4.2.3 Results

We test the M-score (4.10) on both synthetic and real data of *Saccharomyces cerevisiae*. Rather than exhaustively examining results over multiple data sets or cluster methodologies, our purpose is to demonstrate and validate the score's properties. To this end we select a set of 42 cell cycle motifs which are listed in Table 4.4. Consensus sequences for these motifs were extracted from the TRANSFAC Database [111], and string searches for each motif sequence over all 800bp upstream regions results in a gene-by-motif matrix containing raw counts. Although much more advanced methods than string-based searches for motif detection are available (e.g., [94][64]), this choice still constitutes a reasonable approach to validate our setup (e.g., [56][77]).

Synthetic data

As a proof of principle we select the following set of six motifs : Ace2, ECB, SCB, ICRE, Met31_32p, Rap1 and SCB. We collect all 358 genes carrying either of these motifs in their upstream region and divide them into six groups as in Table 4.3. Each motif j lies at the basis of a separate group $(1, \ldots, 6)$, with exception of the SCB-motif whose genes are divided over two groups. We simulate the simple case that the simple motif patterns in Table 4.3 regulate the expression profiles of all genes in particular way: each gene g (out of 358) is assigned a simple expression profile according to the groups as constructed in Table 4.3:

$$\{e_{g,j}\}_{j=1,\ldots,6} = \delta_{(j,\text{group\_nr})} + \epsilon,$$

with

$$\delta_{(j,\text{group\_nr})} = \begin{cases} 1 & \text{if } j = \text{group\_nr of gene } g \\ 0 & \text{otherwise,} \end{cases}$$

and $\epsilon \sim \mathcal{N}(0, \sigma)$ gaussian noise. We call this data synthetic (instead of artificial) as it involves real motifs present in real genes, which are nevertheless assigned to simplified clusters and have simplistic expression profiles. We cluster these 358

Table 4.3: Partitioning of 358 genes bearing the six selected motifs. Ninety-four genes carrying the SCB motif are merged with the ECB genes in group 2.

Group	Motifs(nr_genes)
1	Ace2(19)
2	ECB(15)+SCB(94)
3	ICRE(75)
4	Met31_32p(52)
5	Rap1(2)
6	SCB(101)

genes hierarchically and consider all possible cutoffs for k in the range $2, \ldots, 40$. Figure 4.6 plots the influence of the various motif p-values from Section 4.2.2 on the M-score in case of $\sigma = 0.2$. We see that the binomial method results in a

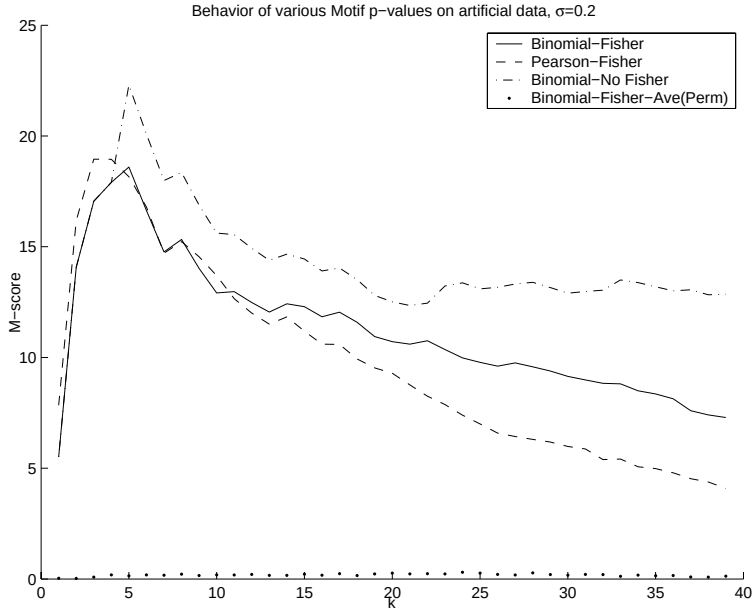


Figure 4.6: M-score versus k on synthetic data. We show the influence of various p-values (derived from the binomial or Pearson χ^2 distribution, with an optional Fisher exact correction for low cluster sizes) on the M-score. The real number of clusters, $k_{real} = 6$, is correctly identified by the binomial-related scores. A Fisher correction for clusters of small size (i.e., less than 8) is necessary to avoid bias in the range of high k. The M-score averaged over 100 permutations is denoted as Ave($Perm$) and we see all M-scores are significant with respect to permuted data

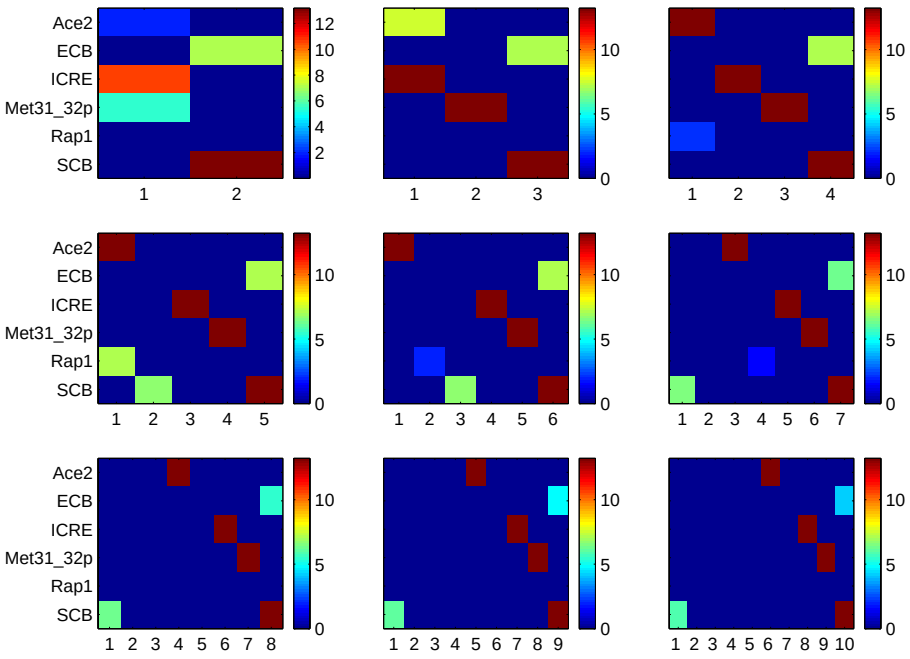


Figure 4.7: Visualization of motif enrichment matrix A for various k on synthetic data ($\sigma = 0.2$). The colorbar indicates the negative logarithm of p-values that stem from the binomial-based M-score with a Fisher correction. The Rap motif, which occurs only in two genes in this synthetic setup, is detected only for $k = 4, \ldots, 7$. The entire pattern of motifs is found back for $k = 6$ and $k = 7$.

peak at the encoded number of clusters, $k_{real} = 6$. According to expectations, computations based on Pearson χ^2-based values behave similarly, but give a slightly different peak. On real data these differences are more pronounced as will be shown further on. Additionally we observe that a Fisher correction for clusters of small size (i.e., less than 8) is necessary to avoid bias in the range of high k, where small clusters would be overweighted (see also Figure 4.10). The significance of the M-score with respect to permuted data is clear by inspecting profile of the averaged score over 100 permutations, denoted as Ave($Perm$) in Figure 4.6.

To illustrate the M-score we visualize in Figure 4.7 how motif enrichment matrices A change over $k = 2, \ldots, 10$ on synthetic data (binomial method with Fisher correction). We see that, except for the Rap motif, all motifs are detected, though with varying significance (expressed as $-\log(p)$). The Rap motif, which occurs only in two genes in this synthetic setup, is detected only for $k = 4, \ldots, 7$. The overall hidden pattern of motifs is found back for $k = 6$ and $k = 7$. Although displaying the same overall motif pattern, $k = 7$ yields a lower M-score as it contains a cluster void of significant motifs. For $k = 5$ we observe a stronger significance of the Rap motif than for $k = 6$, but a weaker correspondence to the original pattern. This illustrates that our method is of an exploratory nature: it is meant to point to a region of interest rather than to a single 'true' value.

Finally, Figure 4.8 shows the behavior of the M-score on synthetic data for various σ. We see that up to $\sigma = 0.4$, the correct region around k_{real} is identified. Microarray data is found to typically exhibit fluctuations around $\sigma = 0.25$ (see Bussemaker *et al.* [56]).

Yeast cell cycle data

To evaluate the performance of the M-score on real data, we use the yeast expression data from Spellman *et al.* [141] (See also Table 1.1 and Figure 1.10). For computational reasons, we select from the 3000 genes with highest variance, those 1745 genes that had literature annotations in the SGD database [29] as of January 2003. To check whether the choice of this subsample would bias our findings we calculate the linear correlation between the a priori chances to find a motifs j in this set of genes, $\{P(j|N = 1745), j = 1..42\}$, versus the a priori chances to find motifs j over the entire genome, $\{P(j|N = 6182), j = 1..42\}$. A correlation of 0.998 is observed, ensuring that this gene selection procedure does not drastically change the motif composition in the gene set with respect to analyses in Cho *et al.* [62] and Tavazoie *et al.* [148]. Hence, the results from the following analyses cannot not be attributed to the gene selection procedure.

Optimal k in terms of the M-score

To demonstrate the ability of the M-score to evaluate various cluster batches, we cluster the data hierarchically using the cosine distance metric and Ward's method, and consider all cutoffs in the range $k = 2, \ldots, 100$. The dendrogram is shown in Figure 4.9. The data was preprocessed as in De Smet *et al.* [67]

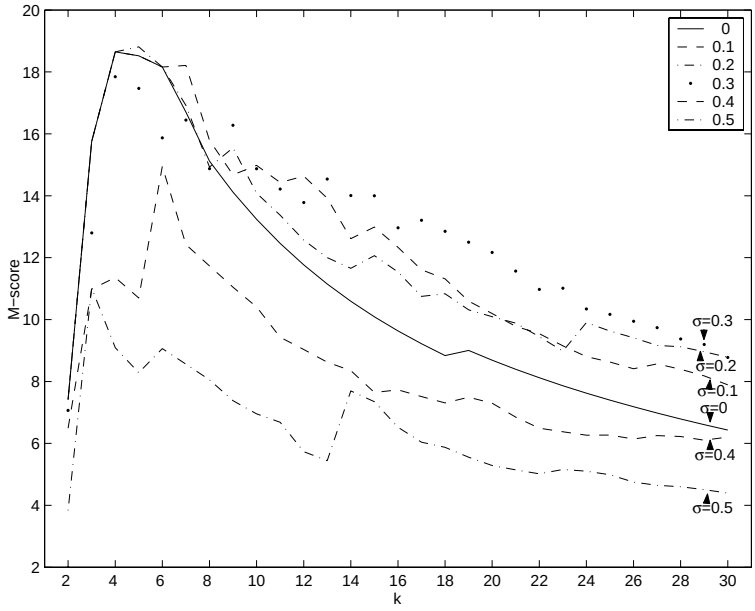


Figure 4.8: M-score versus k on synthetic data with various σ. We illustrate how robust the M-score is with respect to noise distributed as $\mathcal{N}(0, \sigma)$. Up to $\sigma = 0.4$, the score identifies the region around the hidden number of clusters, k_{real}, as the one of interest.

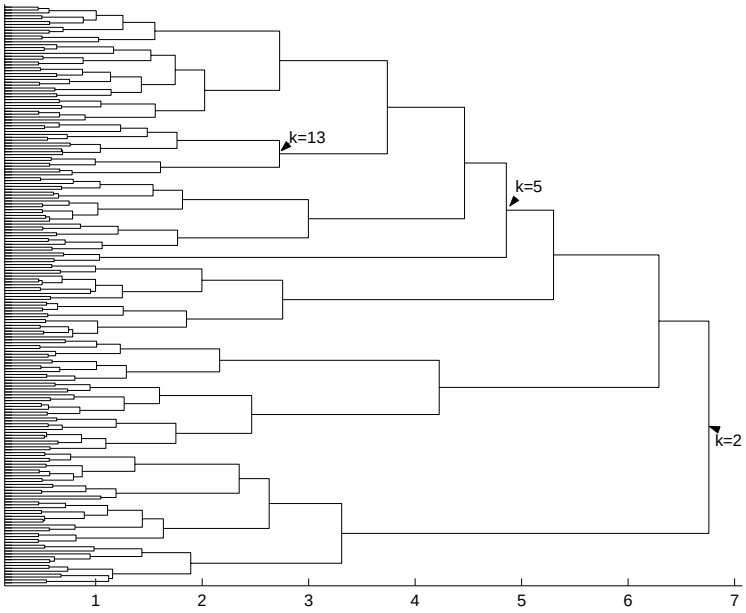


Figure 4.9: Dendrogram of hierarchically clustered cell cycle expression data (1705 genes). Decision on a good cutoff depends on the optimality criterium used. Contrary to Chapter 3, we attempt to estimate k based on the biological criteria that underly our M-score.

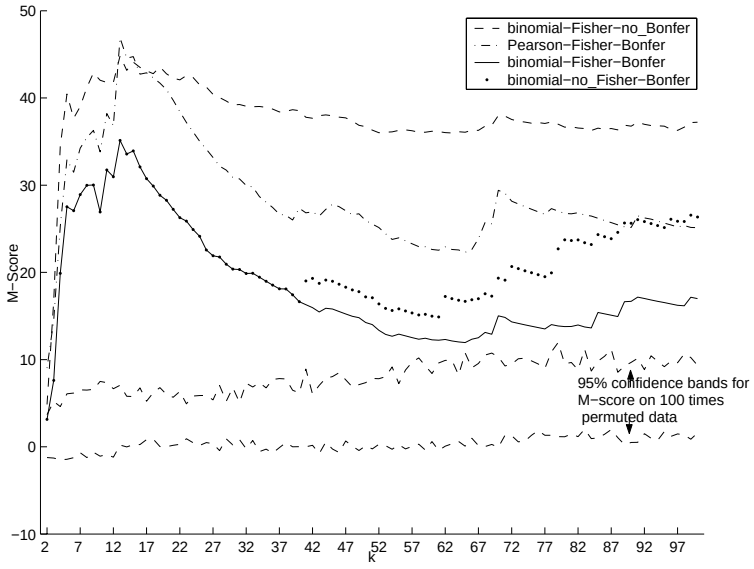


Figure 4.10: M-score versus k on yeast expression data. The Fisher correction adjusts overestimation of p-values in small clusters that are typical for high k. The Bonferroni correction accommodates for the overestimation of motif significance and thereby improves average-to-peak ratio of the scoring curve. The M-score is significant with respect to permuted data.

(see also Section 3.3.1).

Figure 4.10 plots the M-score over k. Looking at the solid curve, we see that the M-score is peaked around $k = 13$, and dropping towards a saturation level (around 10) for lower k. Apart from the fact that Pearson is less conservative (i.e., lower p-values), the overall behavior of the Pearson-based score does not differ from using binomial p-values (provided a correction for small clusters and multiple testing in both cases). This was already observed in Section 4.2.3. The Bonferroni correction accommodates for the overestimation of motif significance and thereby improves average-to-peak ratio of the scoring curve. After permuting the cluster assignments 100 times for each k, we see in that the permuted data occasionally results in significant motifs (i.e., yielding M-scores > 10), but we find the complete M-score profile over k to be significant at the p=0.05 level (see confidence band in Figure 4.10).

Having established the score's significance, we proceed with evaluating its biological properties by visualizing the cluster-by-motif matrix A for the cases $k = 5, 13, 30$ in Figure 4.11.

We see that 5 clusters (k=5) is not sufficient for clustering the data in a biologically meaningful way as all genes with cell cycle specific profiles are grouped in the same cluster (cluster 3) regardless of their phase specificity. This explains why in this cluster both the ECB (M/G1 phase specific) and the MCB motif

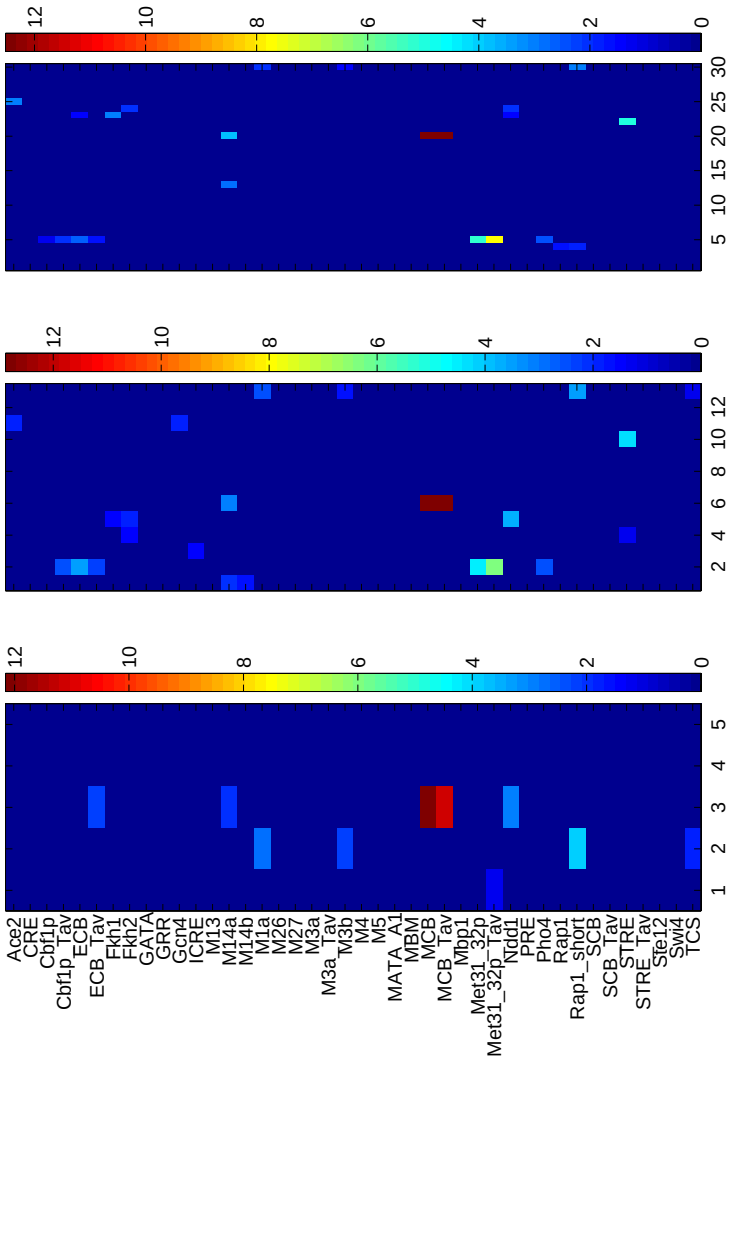


Figure 4.11: Visualization of motif enrichment matrices A for Ward's hierarchical clustering on cell cycle data with $k = 5, 13, 30$.

(G1/S phase specific) are enriched. Cluster 2 is characterized by an enrichment of the motifs M1a, M3b, Rap1 and TCS as well as a significant change in expression in the G1 phase. This cluster contains genes that encode ribosomal proteins, characterized by the Rap1 and M1a motif [148] and other proteins involved in translation and RNA metabolism, found with genes carrying the M3b motif [148]. These genes are tightly coregulated since this cluster is also obtained with $k = 13$ (cluster 13) and $k = 30$ (cluster 30), each time containing approximately the same genes. The other three clusters obtained with $k = 5$ are too diverse to yield any enrichment of motifs. These findings were confirmed by visual inspection of the expression profiles of the clusters (results not shown).

The clustering with $k = 13$ yields the highest number of significant motifs and generates the most consistent biological results. It leads to five periodic clusters, each one specific to a single phase of the cell cycle. In cluster 1 we find an enrichment of M14a and M14b motifs in genes whose expression peaks during the S and the S/G2 transition phase, which is in accordance with the results of Tavazoie *et al.* [148]. Cluster 6 shows an enrichment of the G1/S specific MCB motif [148][141] and the M14a motif, which was thought to be S/G2 specific [148]. The expression profiles of these genes clearly show their specificity for the S phase (Figure 4.12). This phase-specificity is shared by cluster 5 (Figure 4.13), which shows an enrichment for binding motifs for the Fkh1, Fkh2 and Ndd1 transcription factors, which are known to cooperate during the S/G2 transition [103]. Cluster 2 is characterized by an enrichment in ECB, Pho4, Cbf1p and Met31-32p motifs. The latter two factors are known to cooperatively regulate gene expression according to the demand for the products of the amino acid biosynthesis pathway, while Pho4 is implied in the biosynthesis of the amino acid histidine [50]. The expression of the genes in this cluster peaks from the S/G2 transition to the G2/M transition (Figure 4.14), which is consistent with the reported phase-specificity of the Cbf1p and Met31-32p motifs [148]. However, up to now the cell cycle specific ECB motif was thought to be specific to the M phase (Tavazoie *et al.* [148]) or M/G1 boundary (Spellman *et al.* [141]), and not to co-occur with the Met31-32p and Cbf1p motifs. Finally, the STRE motif is characteristic for cluster 10 and this cluster was also found by Tavazoie *et al.* [148]. Additionally, five non-periodic clusters were found that showed a significant enrichment of one or several motifs.

On the other hand, when the data are clustered into a large number of groups ($k = 30$), the cluster with the combined M14a and M14b motifs disappears and Fkh2 is no longer found in combination with its known interaction partners Fkh1 and Ndd1 [103]. It should be stressed that we only clustered a subset of the genes that were analyzed by Tavazoie *et al.*, therefore it is not surprising that our motif results are different. However, taking into account the fact that we clustered only half of the genes, our results seem promising considering Tavazoie *et al.* only found 3 periodic and 5 relevant non-periodic clusters.

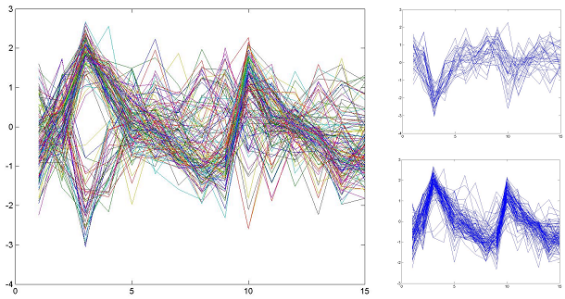


Figure 4.12: Expression profiles of cluster 6, enriched with M14a and MCB motifs. We append a discoupled view of the up- and down-regulated profiles for clarity.

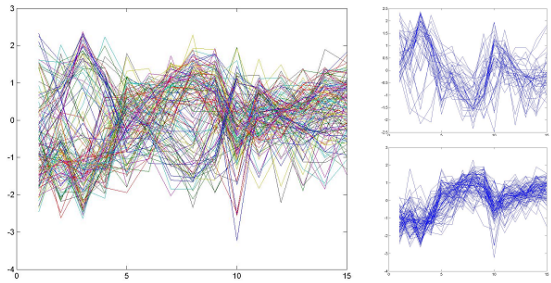


Figure 4.13: Expression profiles of cluster 5, enriched with Fkh1, Fkh2 and Ndd1 motifs. We append a discoupled view of the up- and down-regulated profiles for clarity.

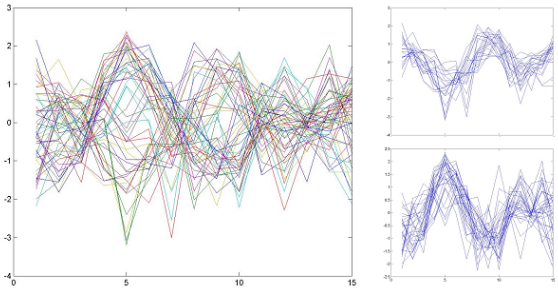


Figure 4.14: Expression profiles of cluster 2, enriched with ECB, Pho4, Cbf1p and Met31-32p motifs. We append a discoupled view of the up- and down-regulated profiles for clarity.

M-score to compare cluster methods

We illustrate how the proposed heuristic behaves for some other standard clustering methods. Figure 4.15 compares complete linkage and Ward's method in hierarchical clustering (cosine distance) with K-means.

Single linkage and average linkage are omitted since they did not produce significant clusters in terms of motifs. Indeed, this is consistent with the analysis of Gibbons *et al.* summarized earlier in Figure 4.3.

As K-means is not deterministic, an average profile over 30 runs with different initializations is plotted. We see that K-means reaches optima of the M-score in the range $k = 5, \ldots, 15$, which is in partial correspondence with the results from preceding Section 4.2.3. Nevertheless K-means scores overall better, and stops yielding interesting clusters above $k = 70$. Additionally, a peak is reached for $k = 5$ which is rather unexpected.

To interpret this result we visualize the motif enrichment for K-means in two cases in Figure 4.16. Comparing the already discussed hierarchical result for $k = 5$ (see Figure 4.11) with the corresponding K-means solution for $k = 5$, we clearly observe that K-means produces a richer outcome in terms of motifs. As the score balances between a moderate presence of biologically relevant motifs on the one hand, and a moderate number of clusters on the other, occasionally biases arise. We find an example in the case of K-means with $k = 5$: although the result might have been reasonable, previous analyses show that further exploration in the range of slightly higher k is necessary.

Juxtaposing Ward's solution for $k = 12$ en K-means for $k = 17$ (i.e., the best parameterizations in terms of the M-score for both algorithms), we see that an increased significance associated with stronger groupings of M14a-, M3a-, SCB- and ECB-related genes respectively, build up the difference in M-score. Finally, we observe that complete linkage clustering does not appear to be an interesting choice in this setup.

4.3 Discussion

Microarray clustering solutions require biological validation. In Section 4.1 we have exemplified how this can be achieved in a quantitative way in terms of GO. The approach of Gibbons *et al.* in particular is interesting for the way it uses external knowledge to enhance the analysis of expression data. On the other hand, we have proposed a new motif-based score (M-score) to quickly estimate the validity of multiple clustering runs in terms of significant sequence patterns or motifs. We have tested it on synthetic and real data about the yeast cell cycle. Our purpose was to treat, verify and demonstrate the M-score, rather than to exhaustively explore its behavior on various data sets or clustering methodologies. We have shown that starting from binomial p-values is a reasonable approach provided a Fisher correction for small cluster sizes and an overall Bonferroni correction for multiple testing. Our background aim was to economize on biological validations when the parameter-space is prohibitively

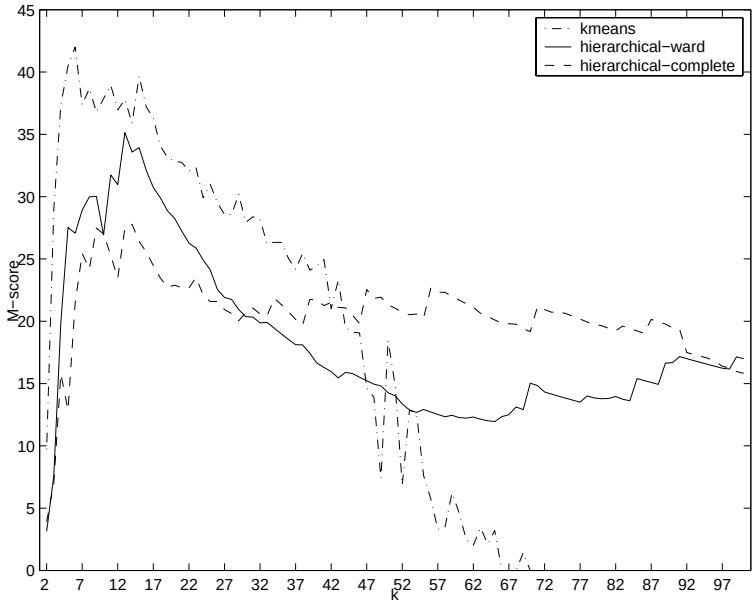


Figure 4.15: M-score versus k for various basic cluster algorithms. All M-score profiles are generated with binomial p-values, with the correction for small clusters (Fisher method) and multiple testing

large. Therefore, we do not consider the M-score an absolute quantification of biological relevance and we illustrated possible biases in an example where k=5. We also tested the M-score on two mainstream cluster algorithms using the number of clusters k as variable parameter. However, also other clustering algorithms with parameterizations different from the number of clusters k can be explored. In this respect we have also experimented with the parameter S in Adaptive Quality-based Clustering [67], which yielded comparable results that did not add extra arguments (results not shown).

Experiments on cell cycle data also pointed out that single and average linkage hierarchical clustering perform poor. This was also found by the GO-based analysis from Gibbons *et al.* Also, the range in which the optimal number of clusters is situated for the Spellman data corresponds well to the outcome of Gibbons and colleagues.

Further on in Chapter 7 we will heavily rely on the desirable properties of the score to check how integrating text-based information with microarray data can reveal gene groupings with higher biological relevance that are not detectable by expression data alone. We have demonstrated the score using sets of known motifs, although this is not part of the stated assumptions and the method is also applicable for sets of newly discovered conserved sequences [160]. A general limitation of the method however is that it attempts to quantify solutions in terms of *direct* regulation and makes no claim to score genetic

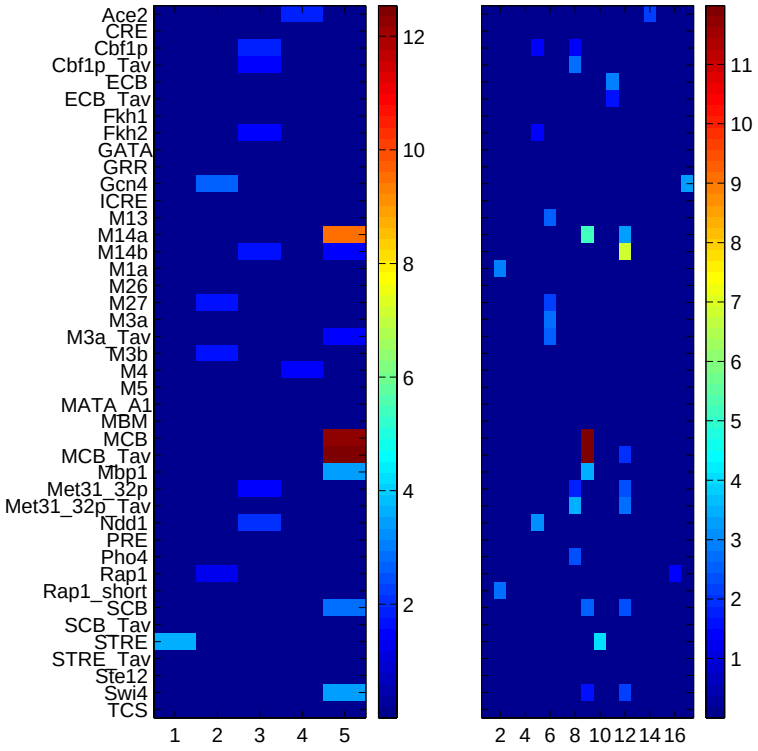


Figure 4.16: Visualization of motif enrichment matrices A for K-means on cell cycle data with $k = 5$ and $k = 17$

networks. Finally, the evaluation of clustering solutions in 'motif space' provides a useful alternative to other validation procedures as described in Sections 3.5 and 4.1.3.

As a conclusion we state that although the biological assumptions that underpin the heuristic constitute a simplification of the biological reality and advanced cluster algorithms that integrate sequence information already exist, the proposed M-score is a useful tool to explore the effect of multiple clustering parameterizations and methods in terms of detecting regulatory patterns over an entire data set.

Table 4.4: Additional Table: Motifs considered in cell cycle experiment. The consensus sequence for each motif was extracted from the TRANSFAC Database. Besides the regular DNA nucleotides A,C,T and G, degenerate consensus symbols are used, which are conform the standard of the International Union of Pure and Applied Chemistry (IUPAC) [16].

Motif	sequence
Ace2	GTGTGTGTGTGTGTG
Cbf1p_Tav	TCACGTG
Cbf1p	RTCACRTG
CRE	TKACGTCA
ECB_Tav	TTWCCNNNWNNGGAAA
ECB	TTWCCCNNWNAGGAAA
Fkh1	AAWNRTAAAYARGN
Fkh2	GGNAAWANGTAAACAA
GATA	GATAA
Gcn4	TGACTCA
GRR	CTTCTT
ICRE	WYTTCAYRNTGS
M13	WWWTTTTTCA
M14a	TTTSGCKTY
M14b	TGTTTRKT
M1a	TTCYCWYYG
M26	GTKTNNGTGNG
M27	AWATTTTTC
M3a_Tav	MGATGAG
M3a	CGATGAG
M3b	TGAAAAA
M4	GAARKYTTT
M5	YWKNTCCYNTTT
MATA_A1	TGATGTANNT
MBM	TATTTGCTC
Mbp1	GNRRNANACGCGTNR
MCB_Tav	ACGCGKNR
MCB	ACGCGTNA
Met31_32p_Tav	AAANTGTGG
Met31_32p	AAACTGTGG
Ndd1	GGNAAWWNGTMAACA
Pho4	CACGTG
PRE	TGAAACA
Rap1_short	ACCCAT
Rap1	ACACCCATACATTT
SCB_Tav	MAMGAAA
SCB	CACGAAAA
Ste12	TGAAACA
STRE_Tav	AGGRG
STRE	AGGGG
Swi4	CACGAAA
TCS	CATTCT

Chapter 5

Text-based gene analysis

The evaluation of the functional significance of heterogeneous gene groups constitutes a major challenge for microarray users. One particular problem is that the biological scope of an expression experiment often exceeds the capacity of the researcher to have an overview of all related in-depth knowledge. As a result, assigning meaning to a set of hundreds of genes involves intense querying and managing of information from various expert databases. One particular strategy to narrow down on a set of genes of potential relevance is text mining. Figure 5.1 pictures the concepts covered.

Using primarily MEDLINE abstract information, we present in Section 5.1 an intuitive literature mining framework in terms of the *vector space model*. In Section 5.2 we demonstrate the performance of a new text-based score for the detection of a priori defined functional groups. We will demonstrate that our text-based score behaves favorably, and hence constitutes a quantitative validation of our text mining framework[1]. In Section 5.3 we will quantify the added value of literature-based information over database annotations and the feasibility of text-based gene clustering. The results in this chapter should be considered as a 'proof of principle' and they are the foundations upon which we build the rest of the work presented in this dissertation.

5.1 Vector space model

5.1.1 Document processing

The construction of a literature database starts with the collection of a set of documents in ASCII format. These documents might be MEDLINE abstracts, internal reports, database annotations, emails,... The document corpus is processed by removing punctuation, case and document structure. For this reason the vector space representation is often referred to as *bag-of-words*.

[1]The text mining framework was first presented to a knowledge discovery and data mining audience at the SIGKDD conference in 2002 [40]. The biological implications of the approach were expounded and discussed at the Pacific Symposium on Biocomputing in 2003 [80].

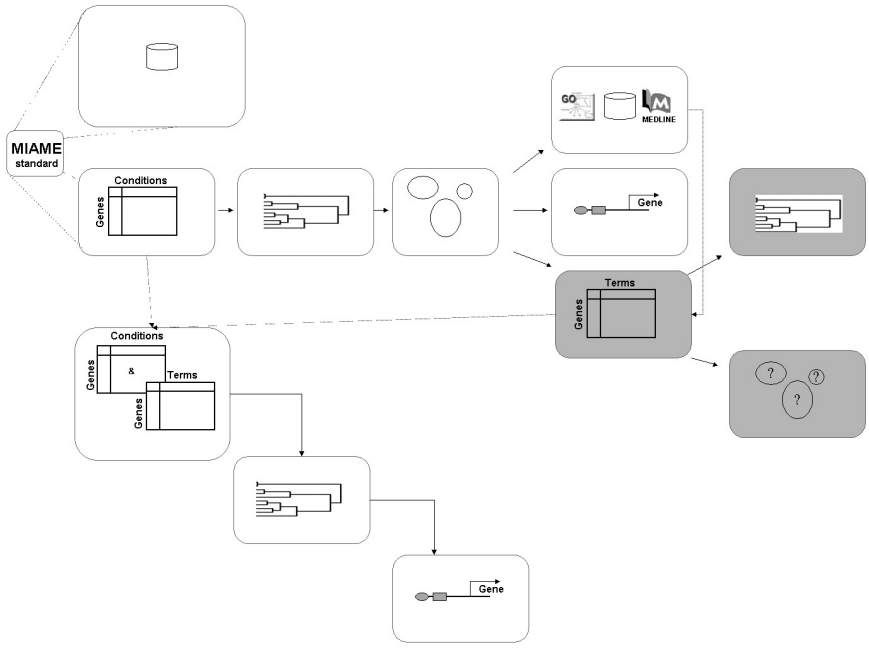


Figure 5.1: Analyzing experimentally established groups of genes from the perspective of prior work as reported in literature. A text mining approach is used to represent a variety of literature information in a gene-by-term matrix. Genes or gene groups can then be analyzed in the term space by (a) clustering genes according to their reported functions in literature, or by (b) quantifying the tightness -and hence functional relatedness- of gene groups in term space. See also Figure 1.7.

Standard English *stemming* (Porter's method [126]) canonizes the words according to morphological and inflexional endings (e.g., plurals, tenses,...) and helps to reduce, to a certain extent, the dimensionality as well as the dependency between the terms. A removal of words including articles, prepositions and conjunctions is desirable to reduce noise and is done using term statistics (see below), and/or using a predefined *stopword* list. *Phrases* are terms consisting of several tokens (e.g., 'amino acid transport') and although little is known on how they affect the performance of learning text (e.g., in document classification or document retrieval), they are important when generating comprehensible summaries (document summarization). *Synonyms* are different terms conveying the same meaning or referring to the same object (e.g., 'tumour' and 'tumor' ;'P53' and 'TP53'). *Polysemy* refers to terms conveying different meanings according to the context it appears in (e.g., 'CD' as compact disk, Crohn's disease, cytosine deaminase, ...).

The idiosyncrasies of biomedical text however pose particular challenges to the implementation of the concepts introduced above. Biological nomenclatures for example, can contain comma's, dots, hyphens, primes, parentheses, brackets,... and can be case-sensitive (e.g., '1-chloro-2,2-bis(4'-chlorophenyl)ethane dehydrogenase'). Moreover, inspiring gene names such as 'CELL', 'AND' or 'ALPHA' renders accurate extraction of crucial biological entities such as genes and proteins, a daunting task. Conversely, applying procedures such as stemming can modify harmless words such as 'may' (if not appearing in the stopword list) into gene names ('MAI') or vice versa. Such issues can seriously trouble the outcome when looking for co-cited genes in literature. We deal with these issues in a pragmatic way by hand-crafted preprocessing. Nevertheless, the detection of gene names [85] and relationships between terms [45][122] remain important research topics, albeit not further covered in this thesis.

5.1.2 Vector representations

The representation called the vector space model encodes a document in a k-dimensional term space where each component w_{ij} represents the weight of term t_j in document d_i. Hence, each point i in this vector space is a row vector containing weights for each term j. We call this a *text-* or *term-profile*.

As explained before any structure in the text is thus obliterated. When all terms are assigned equal importance, binary weights are adopted:

$$w_{ij}^{\text{BOOL}} = \begin{cases} 1 & \text{if } t_j \in d_i \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

This is referred to as Boolean weighing. The IDF, TF-IDF and ln(TF)-IDF[2] weighing schemes on the other hand allow for a partial matching of correspond-

[2] ln is the natural logarithm.

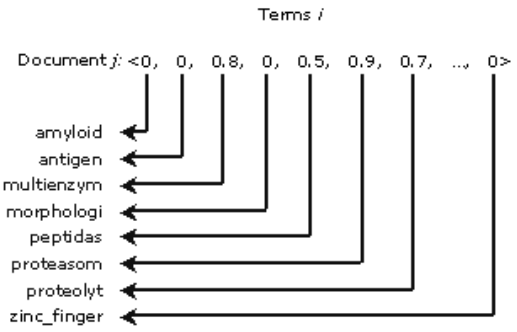


Figure 5.2: Illustration of term index of a given document j containing terms 'multienzyme', 'peptidase', 'proteasome' and 'proteolytic' (i.e., non-zero weights). The set of all terms is named a vocabulary. Typically stopwords such as 'from', 'the', 'often',... are removed. Note that keywords are matched according to their stemmed from, this is, stripped from inflections or plurals.

ing terms and are defined as follows:

$$\begin{array}{lcl} w_{ij}^{\text{IDF}} & = & \log(\frac{N}{n_i}), \\ w_{ij}^{\text{TF-IDF}} & = & f_{ij}\log(\frac{N}{n_i}), \\ w_{ij}^{\ln(\text{TF})\text{-IDF}} & = & \ln(f_{ij})\log(\frac{N}{n_i}), \end{array} \tag{5.2}$$

where f_{ij} is the number of occurrences of t_j in d_i and is called term frequency (TF). It accounts for the assumption that terms occurring often within a document are more important. The logarithmic term $\log(\frac{N}{n_i})$ is called inverse document frequency (IDF), where N represents the total number of documents and n_i is the number of documents containing term i in the collection. It proportionally downweighs terms that often recur throughout the document set. For the many other term weighting schemes that exist, we refer to the book of M. Berry [48]. In Figure 5.2 we illustrate the term index for a given document.

Additionally, we define another type of index called the reference representation (see Shatkay *et al.* [136]). When a concept (e.g., a gene or a document) contains references to other documents in the same or another repository, we can encode this as follows:

$$w_{ij}^{\text{ref}} = \begin{cases} 1 & \text{if concept } i \text{ contains a reference to document} j, \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

The reference representation hence encodes objects in document space instead of term space. We include this representation in some of our studies for its importance at the time of its first appearance [137].

5.1.3 Matrix formulations

We can rewrite the defined weighting schemes as a transformation of the $n \times m$ document-term matrix A containing raw term frequency counts (or their logarithms):

$$\begin{aligned} A &= A_{IDF}Q, , \\ A &= PA_{TF-IDF}Q, \\ A &= ln(P)A_{TF-IDF}Q, , \end{aligned} \tag{5.4}$$

where P is a $n \times n$ diagonal matrix with term normalization constants for each document and Q a $m \times m$ diagonal matrix holding the inverse document frequencies for each term. We haven explicitly written A on the left side of the equation to show analogy with the LSI formulation (5.6) described below. The reference representation (5.3) has no direct analogy with LSI as features other than terms are used.

Latent Semantic Indexing (LSI) extends this vector space model by modeling the term-document relationship using a reduced-dimension representation computed by the singular value decomposition (SVD) of the document-term matrix [49][68]. More specifically the SVD of the document-term matrix A is written as

$$A = U\Sigma V^T, \tag{5.5}$$

with

$$\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{\min(n,m)}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(n,m)}$ sorted singular values and U, V orthogonal $n \times n$ and $m \times m$ matrices respectively. The closest rank k approximation of A is proven to be

$$\text{SVD}(k) = A_k = U_k\Sigma_k V_k^T, \tag{5.6}$$

with U_k and V_k the first k columns of U and V, and Σ_k the $k \times k$ diagonal matrix containing the k largest singular values of A [84]. Choosing a rank k that models best the semantic structure of a collection remains an open question and is mostly done by empirical testing [49].

As the number of terms often run into thousands, of which a fair amount are relevant ones, we also define a representation, IDF-SVD(k), which describes the application of (5.6) to the IDF representation A_{IDF}. The underlying hypothesis is that using IDF as a preprocessing step to SVD can enhance the performance.

5.1.4 Document similarity

Similarity between pairs of documents d_1 and d_2, or between a text document d_1 and a query document d_2, is expressed by the cosine of the angle between the corresponding vector representations:

$$\text{Sim}(d_1, d_2) = \cos(d_1, d_2) = \frac{\sum_j w_{1j}w_{2j}}{\sqrt{\sum_j w_{1j}^2}\sqrt{\sum_j w_{2j}^2}}. \tag{5.7}$$

Since $w_{ij} \geq 0$, for all i, the degree of similarity $\mathrm{Sim}(\cdot,\cdot) \in [0,1]$. The underlying hypothesis states that high similarity between documents testifies to a strong semantic connection between them.

5.1.5 Construction of a gene index

As genes constitute elementary concepts in biological reasoning, there is a need for mappings between genes and textual sources. In practice, gene-literature associations are available in several forms:

- Hits from PUBMED when entering terms related to a gene,
- Gene-to-literature mappings curated repositories such as SGD, GO or LocusLink,
- Gene-to-literature links by using document classification techniques.

After computing a vector-based literature index for each document in a given collection, we summarize for each gene the text indices of all documents associated to it (see also Figure 5.3). The text index of a gene i is then a vector of terms j obtained by taking the average over the N_i indexed documents to which it is linked:

$$g_i = \{g_i\}_j = \{\frac{1}{N_i}\sum_{k=1}^{N_i} w_{kj}\}_j. \tag{5.8}$$

This operation pools the keyword information contained in all documents related to a gene into a single term vector. As a result, documents containing different, but related terms and describing the same gene are joined.

5.2 Scoring groups of genes based on text

We will show how the vector-based framework can be successfully used to couple text-based information and experimental data. In what follows, we present a method to score groups of genes using distances in text space, and apply these scores to

- test to which extent the TF-IDF and LSI text representations can model known relationships between genes;
- explore the significance of clustered gene expression expression data in terms of the proposed score.

5.2.1 Scoring method

We define the average mutual relatedness, or within-group coherence of a set $\mathbf{G}$ of genes in text space (cfr. Figure 5.3) as

$$W_{\mathbf{G}} = \mathrm{median}(\cos(g_k, g_l))_{kl}, \tag{5.9}$$

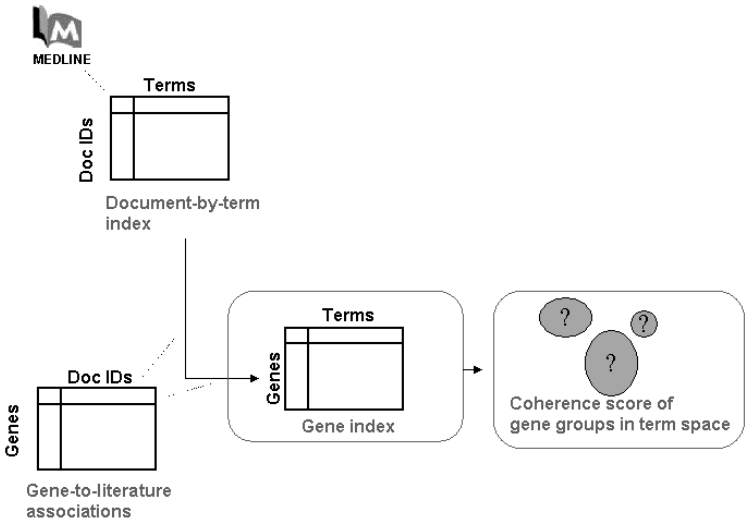


Figure 5.3: After computing a vector-based literature index for each document in a given collection, we summarize for each gene the text indices of all documents associated to it by taking the average text profile. Gene-document associations stem from a retrieval engine such as PUBMED, a document classification algorithms or a database containing gene literature references.

with g_k, g_l gene members of $\mathbf{G}$. To assess the significance of $W_{\mathbf{G}}$ we compute p-values with respect to two different background distributions.

Firstly, we construct a randomization distribution, denoted D_{global} by sampling 100 random gene groups of the same size as $\mathbf{G}$ from all indexed genes. Secondly, we use the fact that multiple groups are simultaneously evaluated and sample a permutation distribution, denoted D_{local}, which is generated from a 100-fold permutation of the group labels.

Each p-value assessing the significance of observed score $W_{\mathbf{G}}$, is computed by modelling either of these distributions, D_{global} or D_{local}, with a parametric Gaussian normal distribution (Figure 5.4). Normality assumptions on this parametric fit were checked using a conservative Kolmogorov-Smirnov goodness-of-fit test [59] at a rejection level of $\alpha = 0.05$. The null hypotheses H_0 for the Kolmogorov-Smirnov test in this case are that the tested D_{local} or D_{global} are normal distributions. For each potential value x, the Kolmogorov-Smirnov test compares the proportion of values less than x with the expected number predicted by the standard normal distribution. It uses the maximum difference over all x values is its test statistic:

$$\max(|F(x) - \Phi(x)|)$$

where F is the proportion of values less than or equal to x and Φ is the standard normal cumulative distribution function evaluated at x. We reject H_0 if

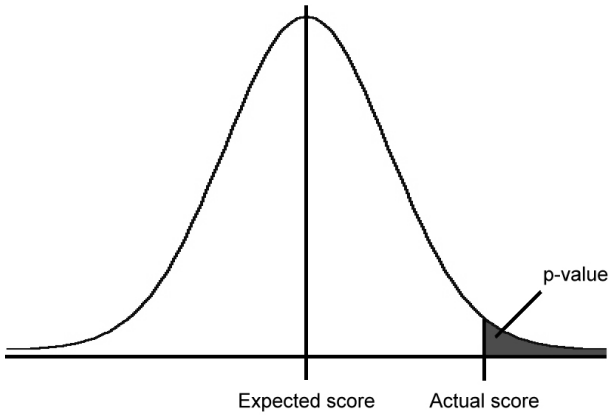


Figure 5.4: p-Value significance of the text coherence score W_G (5.9) is measured with respect to a background distribution, which is computed by either randomization (D_{global}) or permutation D_{local}. The background distribution is modelled by fitting a parametric Gaussian. (Figure courtesy: Steven Van Vooren)

the resulting p-value is 0.05 or lower. Critical values for this test statistic are obtained from tables that are embedded in software such as MATLAB.

Hence, functional relatedness of a group of genes is measured as a p-value denoting the chance that the observed within-group coherence $W_{\mathbf{G}}$ is generated by either of these background distributions. We mention that Herrero *et al.* [89] applied a similar technique when clustering expression data.

5.2.2 Text-based scoring to identify functional categories

As a basic test we score 13 functional groups from GO with various representations (see also Table 1.1 for an overview of the data sets), the parameterizations of which are governed by empirically established guidelines that will be treated in the next section. All the textual information related to the genes in this study was extracted from MEDLINE identified using the curated literature references in the Saccharomyces Genome Database as of 11 Jan 2003. For the association of genes to a predetermined set of functional groups we used the 11 Dec 2002 GO release. As a term space for each document in the collection, we construct a vocabulary consisting of 15,057 (possibly multi-word) terms extracted from GO. The vocabulary was further pruned by considering those terms that occur more than once and less than five thousand times.

In Table 5.1 we list the textual coherence score of IDF and IDF-SVD(40) with respect to a permutation distribution. A detailed comparison between various representations is the topic of the next Section. We see that the IDF representation is capable of correctly identifying 85% of the groups, whereas IDF-SVD(40) attains 92%. We see that the very small membrane fusion group (3 genes) falls is not significant, while the lipid metabolism and cell adhesion

Table 5.1: Text-based coherence of functional groups from GO (906 unique yeast genes). For each functional group we list the number of genes that were linked to it in GO *and* for which we were able to extract curated literature abstracts. We list the textual coherence score (as p-value) of IDF and IDF-SVD(40) in the third and fourth column. All scores herein are computed with respect to D_{local} and scores *above* the one-sided significance 0.025 threshold are highlighted. The last column contains the p-values of the Kolmogorov-Smirnov statistic to check the normality assumptions on the background distribution (here, IDF-SVD(40)). With all values being above the $\alpha = 0.05$ level, we see that these assumptions are valid.

Function	♯ Genes	IDF	IDF-SVD(40)	Support for H_0
Signal transduction	128	9.49E-57	4.55E-17	9.62E-01
Cell adhesion	8	**3.51E-02**	1.41E-03	9.60E-01
Authophagy	27	2.73E-11	7.21E-11	9.72E-01
Budding	98	4.85E-24	5.04E-15	9.93E-01
Shape size control	73	1.83E-06	3.60E-05	6.20E-01
Cell fusion	11	3.72E-55	8.17E-10	8.07E-01
Ion homeostasis	82	1.73E-04	1.20E-04	6.31E-01
Membrane fusion	3	6.29E-04	**2.91E-02**	9.36E-01
Sporulation	68	**8.26E-02**	**9.48E-01**	8.96E-01
Amino acid metabolism	145	9.16E-06	5.49E-21	5.87E-01
Carbohydrate metabolism	168	2.47E-04	3.28E-15	7.00E-01
Electron transport	20	1.51E-11	8.12E-10	4.87E-01
Lipid metabolism	149	**6.30E-01**	3.04E-18	6.52E-01
Nitrogen metabolism	33	7.36E-27	1.28E-05	9.02E-01

group are only recognized in the best representation. All representations failed to detect the sporulation group. The poor results of the sporulation group may be explained by the fact that this is a very diverse group of genes that overlap with many of the other groups under consideration. For instance, the processes of sporulation and budding share a number of components (e.g., ACT1, CDC10, SUR7). Furthermore, the sporulation group contains genes involved in autophagy (AUT10, PRE1, PUP2, UBC1), signal transduction (BMH1, BMH2, GPA2, IRA1, RAS2), metabolism (GTS1, SHP1) and membrane fusion (MSO1, NEM1).

As explained before, we fit a normal distribution through the sampled background distributions. The validity of this model assumption for IDF-SVD(40) is illustrated in the last column of Table 5.1, where we see that the permutation distributions for each of the 13 groups follow a normal distribution. We tested the null hypothesis of normality at the $\alpha = 0.05$ level with a Kolmogorov-Smirnov test. No exceptions were encountered for all other tested representations and background distributions.

Hence we see that the functional coherence measure (5.9) as explained in Section 5.2.1 is successful at identifying functional groups. A fortiori, we have an indication that the adopted text mining representation is useful for our purposes as stated in the introduction. In the next section we treat more in detail several issues about the representation and methodology.

5.2.3 Text-based scoring to compare classical vector space representations

In what follows a quantitative analysis of various representations is carried out on 10 cell cycle specific functional groups encompassing 126 genes selected from Spellman *et al.* All genes are listed in Table 5.2.

Benchmark

The IDF, ln(TF)-IDF, and TF-IDF model increasingly favor term occurrence within abstracts. In Table 5.3 we see that, with a one-sided p-value threshold set at 0.025, TF-IDF is capable of detecting the glycosylation group at the expense of the fatty acid group that is detected by ln(TF)-IDF.

In terms of overall precision-recall, Figure 5.5 points out TF-IDF as the method of choice among these three, followed by IDF and ln(TF)-IDF respectively. The LSI model, SVD(40), that maps all genes in a 40-dimensional reduced vector space, is of comparable performance to TF-IDF but misclassifies another group (secretion). Finally, when using the IDF as preprocessed input to LSI (denoted as IDF-SVD(40)), we achieve best overall performance with one misclassification (sporulation group). IDF-SVD(40) achieves the best learning curve corresponding to 90% recall versus a precision of 1-1.38E-03=99.998% (see Figure 5.5).

Randomization versus permutation

The second part of Table 5.3 shows results when significance is measured with respect to the randomization distribution D_{global}. In this case all but the sporulation group are detected with the TF-IDF and IDF-SVD(40) representations. This is to be expected as samples drawn from D_{local} maintain some of the structure present within the gene-term matrix of the examined cell cycle genes, while samples from D_{global} exhibit much less structure as they are drawn from a much larger set of genes. Although computing p-values with respect to D_{global} appears the method of choice in this setup, we argue that in certain contexts, and gene expression analysis in particular, considering the more conservative D_{local} might be preferable. We will illustrate this in section 5.2.4.

Parameterizations of LSI

One problem in the LSI model is that the choice of optimal rank remains an open question and is normally decided via empirical testing [48]. In Figure 5.6 we show the scree plot of sorted singular values of the document-term matrix of the 10 cell cycle groups.

One plausible cutoff would lie between rank 7 and 10, around the jump in the curve's first derivative. In practice however, this proves to be too stringent and we choose rank 40. Applying the LSI model on a IDF-processed document-term matrix increases expected precision over various ranks. In Figure 5.7 we plot, for the hardest detectable group (Secretion), the effect of rank reduction

Table 5.2: Cell cycle specific genes used to compare vector representations.

FATTY ACIDS/LIPIDS/..	NUTRITION	SECRETION	DNA REPAIR	CELL CYCLE CONTROL	METHIONINE
EPT1	BAT2	EMP24	DUN1	CLB5	MUP1
LPP1	PHO8	SEC28	MSH2	CLB6	MET1
PSD1	AGP1	SLY41	MSH6	CLN1	MET6
SUR1	BAT1	UFE1	OGG1	CLN2	MET10
SUR2	GAP1	ERV25	PMS1	HSL1	MET13
SUR4	DIP5	SSO2	RAD27	PCL1	MET14
AUR1	FET3	GYP6	RAD5	PCL2	MET28
ERG3	FTR1		RAD51	RME1	MET17
LCB3	PFK1	**SPORULATION**	RAD53	SWE1	MET3
ERG2	PHO3	SPO16	RAD54	CLB4	SAM1
ERG5	PHO5	HOP1	RDH54	HSL7	
PMA1	PHO11	HDR1	RHC18	WHI3	**GLYCOSYLATION**
PMA2	PHO12	SPS4	UNG1	ACE2	MNN1
PMP1	PHO84	SSP2	HPR5	CLB1	OCH1
ELO1	RGT2		MEC3	CLB2	PMT1
FAA1	SUC2	**MITOTIC EXIT**	ALK1	CLN3	PMT3
FAA3	SUT1	DBF20		SWI5	PMT5
FAA4	VCX1	APC1		PCL9	PSA1
FAS1	ZRT1	CDC5		SWI4	QRI1
	GLK1	CDC20			SVS1
PSEUDOHYPAE	HXT1	DBF2			SSO1
ELM1	HXT2	MOB1			GDA1
PHD1	HXT4	SPO12			PMI40
TEC1	HXT7	TEM1			ALG7
	TAT1	SIC1			VRG4

Table 5.3: Tested functional groups as reported in Spellman *et al.* (126 unique yeast cell cycle related genes). For each functional group, we list the representation and the type of background distribution used. All permutations or randomizations were taken identical across the representations. Scores above the one-sided significance 0.025 detection threshold (i.e., groups that are not detected as functionally coherent) are highlighted in bold. The IDF, ln(TF)-IDF, and TF-IDF model increasingly favor term occurrence within abstracts. We see that, with a one-sided p-value threshold set at 0.025, TF-IDF is capable of detecting the glycosylation group at the expense of the fatty acid group which is detected by ln(TF)-IDF. Finally, preprocessing LSI with IDF favors the detection of the secretion group with respect to regular LSI.

Groups	IDF	ln(TF)-IDF	TF-IDF	SVD(40)	IDF_SVD(40)	IDF	IDF-SVD(40)
	D_local					D_global	
CELL CYCLE CTRL	1.14E-148	6.04E-69	0.00E+00	5.11E-57	9.54E-34	1.01E-167	5.06E-33
DNA_REPAIR	1.53E-33	3.36E-19	1.52E-102	8.45E-11	7.72E-13	3.91E-61	1.52E-16
FATTY_ACIDS/LIPIDS	**2.76E-02**	7.84E-03	**2.00E-01**	3.79E-09	2.63E-13	4.28E-08	1.99E-21
GLYCOSYLATION	**1.55E-01**	**5.30E-01**	1.25E-03	7.77E-05	2.09E-06	6.29E-06	7.05E-08
METHIONINE	1.82E-15	1.37E-09	2.54E-27	2.52E-05	2.26E-07	9.88E-28	5.40E-06
MITOTIC EXIT	4.85E-76	1.08E-55	9.07E-117	7.44E-14	2.04E-09	1.50E-82	1.02E-07
NUTRITION	2.12E-11	8.50E-06	1.99E-07	4.24E-06	7.32E-23	1.76E-18	9.87E-17
PSEUDOHYPAE	6.74E-07	2.74E-03	1.45E-21	1.70E-03	1.90E-04	2.79E-05	4.25E-03
SECRETION	6.39E-03	**1.48E-01**	1.48E-05	**4.47E-02**	1.38E-03	1.11E-06	2.29E-03
SPORULATION	**9.13E-02**	**4.47E-01**	**1.25E-01**	**6.35E-01**	**9.59E-01**	**1.11E-01**	8.81E-01

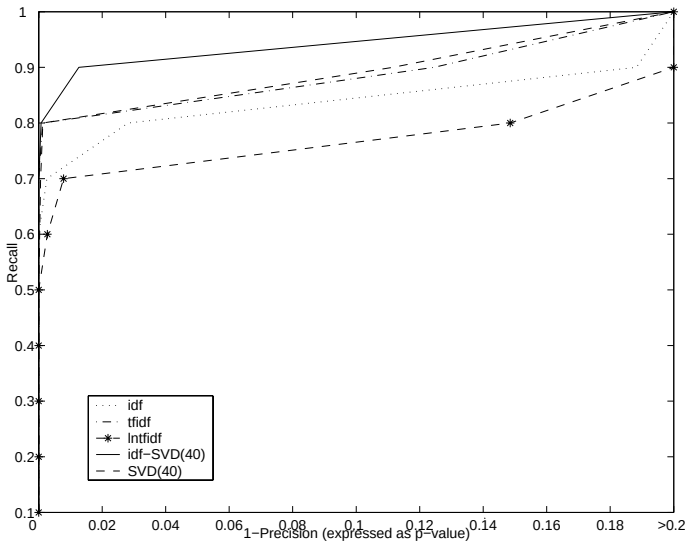


Figure 5.5: Precision-recall plot for various text representations of the cell cycle groups. The x axis denotes $1 - \text{Precision}$, i.e., the larger the p-value, the higher the chance to recover random groups, and consequently the less precise the classification method is. This figure is an extension of the results in Table 5.1. We see that LN(TF)-IDF performs worst, SVD(40), and TF-IDF both score better than IDF, while a combined representation IDF-SVD(40) produces the best curve, corresponding to 90% recall versus a precision of (1-1.38E-03 = 99.9986%).

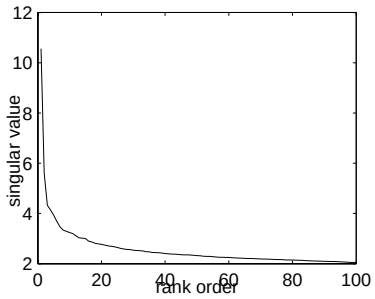


Figure 5.6: Scree plot of singular values of the document-term matrix of the MEDLINE collection associated to the 10 cell cycle groups. A classical cutoff value would be in the range $[7, \ldots, 10]$, but this generally proves to be overly stringent in LSI.

by computing the permutation scores (Section 5.2.1) for all SVD and IDF-SVD representations of ranks in the interval $[3, \ldots, 80]$. The figure shows that when establishing significance threshold at 0.025, IDF-SVD detects this group for 71% of the ranks, in contrast to the 'pure' SVD case where only 13% of the ranks give rise to recognition. This suggests that processing (i.e., weighting) with IDF robustifies LSI in these type of problems, which is usually applied to the raw frequency matrices.

Finally, we report that postponing the LSI indexing to after the summarization process (i.e., computing the SVD of the gene-term matrix instead of the doc-term matrix, see Figure 1), and choosing the rank on similar considerations, did not essentially change our results.

5.2.4 Using text to score clusters from expression analysis

We test how coherence in expression space corresponds to the presented functional coherence score, which is based on text. To this end we collect the cluster memberships of the set of 126 cell cycle genes (see Section 5.2.3) as assessed by the expression analysis by Tavazoie *et al.* [148] who performed a K-means clustering (k=30) on a genome-wide set. The first part of Table 5.4 plots the text-based score and the expression-based score, both computed in the same fashion and measured with respect to D_{local}. The second block of Table 5.4 on the other hand, shows significance of text and data determined with respect to D_{global}. The results show that the text coherence corresponds well to the data-based coherence for groups that are highly functionally related. Tavazoie *et al.* performed a biological validation of their clustering based on functions described in the MIPS catalogue. We see that for very diverse groups, for which Tavazoie and colleagues did not find any significant functional enrichment, the text representation also fails to group the genes together. For clusters that are functionally enriched but diverse (3 or more functional classes grouped together as in cluster 1, 2, 7, and 14), the data-based clusters are textually coherent but

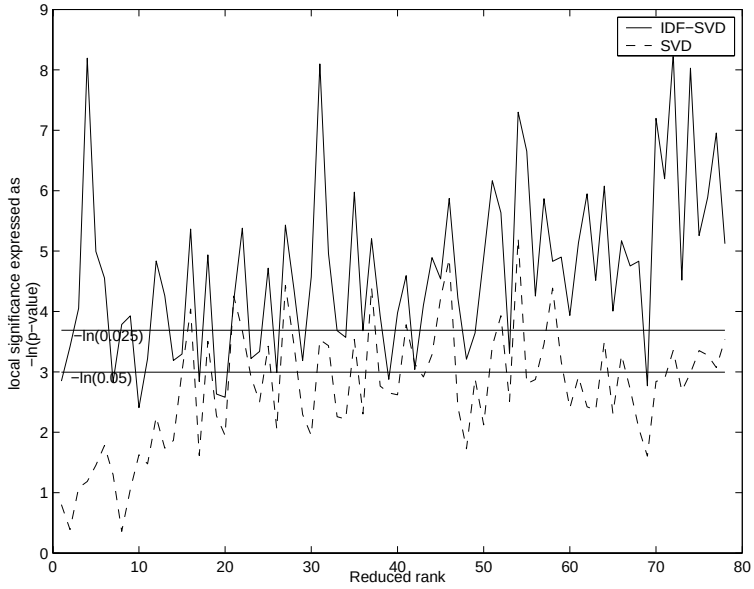


Figure 5.7: Application of LSI model to an IDF processed document-term matrix increases expected precision over various ranks. For the hardest detectable group (Secretion), we examined the effect of rank reduction with respect to the permutation score by computing permutation scores for all SVD and IDF-SVD representations of ranks in the interval [3..80]. The figure shows that when establishing significance threshold at 0.025, IDF-SVD detects this group in 71% of the cases, in contrast to the 'pure' SVD case where only 13% of the ranks gives rise to recognition. This suggests that preprocessing with IDF robustifies LSI for these type of problems.

only when the significance is measured globally. For tightly functionally related clusters (1 or 2 functional classes as in cluster 4, 8, and 30) the coherence is significant with respect to both distributions. Hence the choice for either strict or diverse functional enrichment can be governed by the choice of the background distribution the score is evaluated against. We argue that depending on the type of questions formulated about a clustered expression data set, both approaches are relevant.

5.2.5 Conclusion

Various ways to derive a quantitative score that screens for literature-based functional enrichment in a group of genes were treated and tested on both general and specific problem settings. We report IDF-SVD(40) to display best overall performance when scoring gene groups. The interpretation, though, of SVD-transformed data is more challenging and makes it less suitable for the extraction of terms (as in Chapter 6). Additionally, we have investigated how 'local' permutation or 'global' randomization can be used to compute the significance of a group. Both of them are relevant given the characteristics of biological data (see Chapter 1).

5.3 Clustering genes based on text

In previous sections we established a correspondence between gene groups represented in text space and externally defined gene grouping (from GO or expression data). Our approach to *recognize* functionally related groups from text by means of a score was meant to prove the general viability of our framework. In what follows, we explore how to *reconstruct* (instead of recognize) functionally separated groups by clustering textual gene representations. First, we conduct a controlled experiment involving three groups (see Table 1.1 for an overview of the data sets). We continue to explore various aspects of the vector space model - this time from a clustering perspective. We quantify (a) the effect of textual information used, (b) indexing schemes and (c) cluster quality. Next, we attempt to segregate the groups of genes studied earlier in this chapter and approach the outcome from the biological side.

5.3.1 Performance of text clustering

For the following experiment, we have collected and compiled (as of September 2001) several sources for textual annotations of the genes:

- Gene descriptions from the Saccharomyces Genome Database (SGD) [29],
- functional information from SWISS-PROT (SP), a curated protein sequence database,
- SGD and SP information that is pooled into a local database denoted YeastCard (YC),

Table 5.4: Text Coherence of expression cluster from Tavazoie *et al.* with respect to **(a)** Local permutation (Columns 3-4). The gene groups nitrogen and sulfur metabolism (Cluster 0), amino acid metabolism (Cluster 30), mitochondrial organisation and respiration (Cluster 4) and TCA pathway (Cluster 8) were identified as significant groupings based on clustered data. **(b)** Global permutation (Columns 5-6). Global significance is computed by taking a single, random group of 126 genes and generating 100 randomized versions for each of the 13 clusters. In this table the *significant* scores (i.e., *below* one-sided 0.025 detection threshold) are highlighted in bold.

Cluster nr	Cluster size	p-value Text D_{local}	p-value Data D_{local}	p-value Text D_{global}	p-value Data D_{global}
0	35	**0.01822**	0.49151	**3.43E-06**	**3.30E-08**
1	4	0.32109	0.23747	**0.013266**	0.04566
2	25	0.71614	**1.20E-36**	0.033733	**2.40E-133**
4	8	**0.0002287**	**0.0019**	**1.33E-07**	**2.00E-12**
7	16	0.15159	**2.90E-05**	**0.0080045**	**1.50E-25**
8	5	**0.0052197**	0.03028	**1.51E-06**	**0.0018**
11	3	0.63933	0.03897	0.21308	**2.20E-06**
12	4	0.87697	**0.00854**	0.60303	**2.70E-06**
14	6	0.061157	**6.70E-08**	**0.0094759**	**6.80E-09**
16	2	0.70254	0.96235	0.57749	0.96574
23	3	0.34164	**0.00097**	0.10734	**6.10E-08**
28	5	0.16944	0.27417	0.031223	**0.00098**
30	5	**0.0040335**	**0.00017**	**0.00044264**	**3.60E-40**

- Expansions of the above fields with retrieved PUBMED abstracts.

We construct a set of genes for which the functional associations are well-established. From the MIPS catalogue [44], we select three biologically distinct functional groups consisting of 116 genes in total. For all genes we select their corresponding SGD and YC annotations and proceed with the 105 genes that have entries in both databases.

The first group holds 63 genes that encode lysosomal proteins. The second group consists of 30 genes involved in translational control and the third contains 23 genes related to amino acid transport. Following the text mining approach explained in Section 5.2, all gene annotations are represented by various indices and subsequently *expanded* with the 20 best matching (in cosine sense) MEDLINE abstracts. More specifically, we perform the expansion by re-indexing the enriched annotations, again following various indexing schemes. Table 5.5 summarizes the impact of these settings on cluster performance, expressed by means of the *Rand* index R_{adj} (see Chapter 3). Firstly we discuss the effect of information source, afterwards follow the results on the indexing schemes.

Table 5.5: R_{adj} scores for clustering the three groups using various representations. Note that some results are duplicated along the blocks to facilitate discussion. We observe that, over all experiments, using IDF in the representations outperforms the TF, BOOL and *ref* weighting schemes. Database annotations outperform keyword-based information and, in turn, abstracts improve the clustering over database annotations.

	Representation	Weight	R_{adj}	$\sharp(t_i)$
Source	SP Keywords	BOOL	0.1767	3
	SGD	TF	0.4050	
	YeastCard	TF-IDF	0.4617	
Index	SGD	BOOL	0.3386	8
	SGD	TF	0.4050	
	YeastCard	BOOL	0.3323	
	YeastCard	TF	0.4028	
	YeastCard	TF-IDF	0.4617	26
	$YC\text{-}ML_{20}$	BOOL	0.3726	
	$YC\text{-}ML_{20}$	TF	0.2953	
	$YC\text{-}ML_{20}$	TF-IDF	0.7344	396
	$YC\text{-}ML_{20}$	ref	0.2354	20
Expansion	$SGD\text{-}ML_{20}$	TF-IDF	0.5920	
	$YC\text{-}ML_{20}$	TF-IDF	0.7344	

Effect of indexing scheme

In the second block of Table 5.5 we write the performance of the boolean (BOOL), frequency (TF) and TF-IDF index on typical free-text entries in annotation databases, and on a set of our top 20 retrieved MEDLINE abstracts. For very brief keyword-based descriptions (less than 8 words) the boolean representation is found to be the best one. If all fields from SGD are used, TF

(0.4050) improves on BOOL (0.338). For the YC database, typically containing 30 to 50 terms per entry, TF-IDF (0.4617) outperforms BOOL (0.3323) and TF (0.4028) slightly.

In the expansion step, we collect and re-index the 20 best matching MEDLINE abstracts for each gene. This operation provides a profile for each gene with the number of terms ranging typically between 200 and 400. Among the indexing options for this set of abstracts, TF-IDF (0.7344) scores considerably higher than BOOL (0.3726) and TF (0.2953), even after stopword removal.

Basing ourselves on the same 20 top-scoring abstracts we also evaluate the performance of the reference representation ref, which characterizes a gene in document space instead of term space. It has an R_{adj} value of 0.2354, indicating that it is a less descriptive representation. This can be explained by the fact that ref is probably more dependent on the retrieval of shared, relevant abstracts (see also Shatkay *et al.* [136]).

Effect of information source

In the first block of Table 5.5, we see that for the gene groups considered, the *keywords* field in SWISS-PROT does not provide sufficient information for an acceptable clustering result (0.1767). For instance, the SWISS-PROT keyword list only provides an average of 2 to 3 meaningful keywords for 86 out of 105 genes. The remaining genes are described with no or irrelevant keywords such as *hypothetical protein*, which will not allow for correct classification. Using the GO entries and especially the description line of SGD improves the results and raises the *Rand* score to 0.4050. Only two genes have no biologically relevant terms. When resorting to our pooled information source YC, we obtain a score of 0.4617, misclassifying 21 out of 105 genes. Although the clustering itself is not dramatically influenced by the expansion with YC, for most of the genes, the textual representation is greatly improved (e.g., the weights of specific terms are increased and additional specific terms are incorporated). We refer to Chapter **??** for a detailed illustration of text profiles.

Cluster quality

Because of the absence of a gold standard or prior knowledge in regular clustering problems, internal measures of quality are used to evaluate a cluster result. They are based on various statistical properties of the grouped data and provide clues to choose between different parameterizations of one algorithm (such as the number of clusters) or between various cluster algorithms. Here we use two measures, the *Silhouette* coefficient and a k Nearest Neighbour (k-NN) learnability index (see Chapter 3) to study the influence of the text representation on a standard clustering procedure. Our concern is that, although high *Rand* scores may be encouraging, they do not provide any level of confidence in the result: it might be that clusters or groups lie very close to each other or that clusters exhibit a high spread. Therefore, we compute for each clustering their score over the major text representations.

Table 5.6: Various cluster quality scores for the three major text representations. High *Silhouette* coefficients (3.12), low p-values (k-NN misclassification rates) and high *Rand* indices correspond to desirable clusterings. We observe a discrepancy between *Silhouette* coefficient on the one hand and k-NN misclassification rate and R_{adj} on the other hand.

Representation	*Silhouette*	10-NN p-value (miscl. rate)	R_{adj}
SGD	0.220	4.634^{-4}(0.2286)	0.4050
YC	0.1576	1.2^{-3}(0.2095)	0.4617
$YC\text{-}ML_{20}$	0.2192	10^{-9}(0.1143)	0.7344

From Table 5.6 we see that the *Silhouette* score does not contain any indications towards the optimal representation (i.e., the one with the highest R_{adj}). The more local 10-NN misclassification rate correlates better to R_{adj}. Therefore we examine individual cluster structures more carefully. We expect that groups that are easy (and therefore do not need an elaborate representation to be learned) will end up in clusters having low misclassification rates over all the representations. Harder groups will behave opposite.

In Figure 5.8 we plot the misclassification rate against the *Silhouette* coefficient. We observe there exists a discrepancy between the *Silhouette* and learnability score. We estimate the group of translational control as the hardest to learn from text, since it has the highest misclassification rate, even with the MEDLINE expanded representation. For the amino acid group there exists significant variation in the *Silhouette* value while the misclassification rate stays below 0.1. This indicates that the shape of the cluster changes over the sources used, but not its relative position with respect to the other clusters. The group involved in lysosomal organization behaves more or less as expected, as the *Silhouette* increases and the misclassification drops with the gradual addition of information (SGD; YC; YC-ML_{20}).

The quality of the cluster is highly affected by the presence of distant genes: genes that have a poor or biased description (and hence representation) will end up far away from the cluster center (i.e., the medoid). We illustrate this in Figure 5.9, where we plot the growing of the *Silhouette* score (from right to left on the x-axis) for the YC-ML_{20} representation while increasingly dropping members beyond a given distance. Flat regions indicate the absence of members in that distance region and sudden changes in *Silhouette* scores show the detrimental effect of those, more distant, genes on the score. As before we deem the translational control group the hardest to learn as it drops most sharply. Conversely, the group describing amino acid transport seems the most coherent. Similar as before, the lysosomal group takes up an intermediate position. Since a biologist is not always interested in clustering *all* the genes *per se*, this information can be utilized to prune genes from the clustering process or to check the information given by that gene.

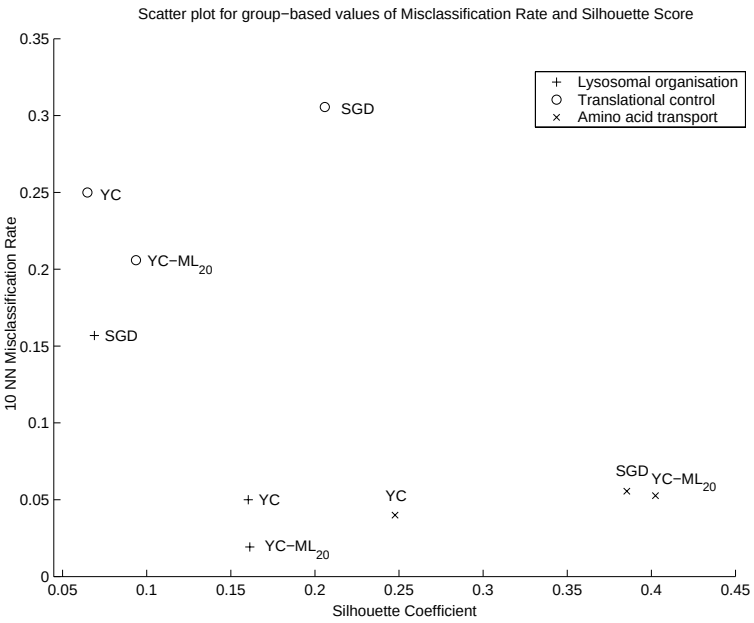


Figure 5.8: Correspondence between spatial cluster information as captured by the *Silhouette* coefficient and learnability in Nearest Neighbour sense. We estimate the group of translational control as the hardest to learn from text, as it has the highest misclassification rate, even with the MEDLINE expanded representation. For the amino acid group there exists significant variation in the *Silhouette* value while the misclassification rate stays below 0.1. The group involved in lysosomal organization behaves more or less as expected, as the *Silhouette* increases and the misclassification drops with the gradual addition of information.

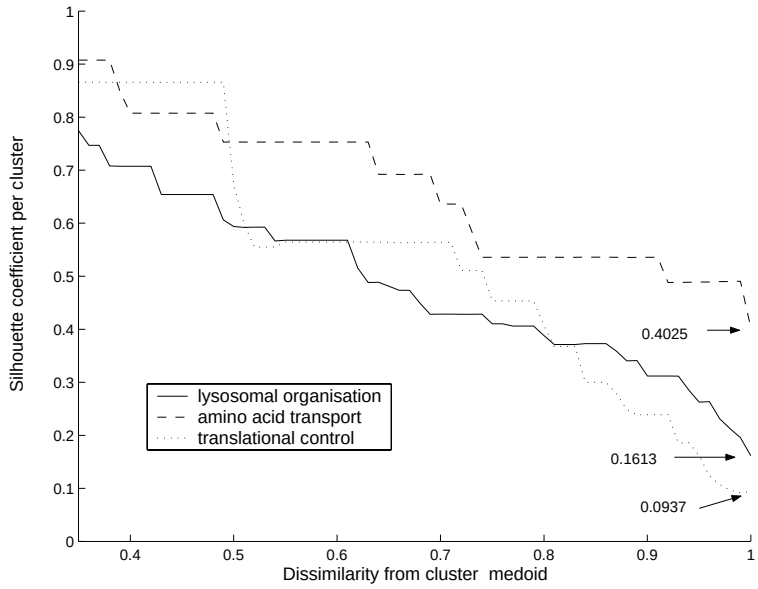


Figure 5.9: Effect of distant members in each cluster on its *Silhouette* score in text-based clustering. We plot the growing of the *Silhouette* score (i.e., from right to left on the x-axis) for the YC-ML$_{20}$ representation while increasingly dropping members beyond a given distance. The translational control group seems the hardest to learn as it drops most sharply. Conversely, the group describing amino acid transport seems the most coherent. Similar as before, the lysosomal group takes up an intermediate position.

5.3.2 Reconstruction of functional segregation using text

To illustrate how the representation interrelates gene groups, we cluster the text profiles of the 906 genes from Section 5.2.2 and the 126 cell cycle induced genes from Section 5.2.3 using hierarchical clustering.

Selected GO groups

To see how the 13 general GO groups (906 genes) are interrelated or discriminated among each other, we study the dendrogram in Figure 5.10. We see that the various metabolic processes are clustered closely together. This is no surprise since metabolism is a highly integrated process. Individual metabolic pathways are linked into complex networks through common, shared substrates. Additionally, the majority of these processes, oxidative phosphorylation, the citric acid cycle, amino acid catabolism, and fatty acid oxidation share the same subcellular location, the mitochondrion.

For the autophagy-related genes we see that they are grouped together with the genes involved in ion homeostasis. Autophagy is a bulk protein degradation process that takes place in the lysosomes, membrane-bound organelles that serve as the major degradative compartment within the central vacuolar system of eukaryotic cells. These lysosomes also play crucial roles in metal ion homeostasis and plasma membrane repair. This explains why we see a relation between the autophagy and the ion homeostasis group.

Saccharomyces cerevisiae is unusual in that it has two methods of reproduction, vegetative growth and sexual reproduction. Cell budding is the most common mode of vegetative growth in yeasts. Yeast buds are initiated when mother cells attain a critical cell size at a time coinciding with the onset of DNA synthesis. The subsequent localized weakening of the cell wall, together with tension exerted by osmotic pressure, allows extrusion of cytoplasm into an area bounded by new cell wall material. During this process, the mitotic cell cycle ensures the formation of a duplicated genome for the daughter cell through a number of consecutive steps such as DNA synthesis, nuclear division, spindle formation, bud emergence, nuclear migration, and cytokinesis.

The cellular machine, responsible for distributing each set of the duplicated genome into daughter cells during mitosis is called the mitotic spindle and consists of the genes from the 'cell fusion' group, mainly microtubular components. The dendrogram affirms the close relatedness between budding and cell fusion. Though vegetative growth is the major way of yeast reproduction, sexual reproduction is an alternative when nutrient supplies fall short. The latter process involves the conjugation of cells of opposite mating type. Under starvation conditions, meiosis is induced, which leads to sporulation and finally to the propagation of four haploid spores that segregate.

The nucleus moves during mating towards the tips of the mating cells and some of the molecular mechanisms are shared with the movements of the nucleus and spindle in budding cells. When mating partners make contact, the cell walls knit together to form a continuous outer layer. To complete formation

of a zygote, the cell wall separating the partners must be degraded, plasma membranes must come into contact and fuse, and finally the haploid nuclei must merge into a single diploid nucleus. The described fusion of the plasma membranes is carried out by the genes from the 'membrane fusion' group. Based on this definition of sporulation, we would expect to observe a relation between the sporulation, the cell fusion and the membrane fusion group. This is, however, not the case in the dendrogram, which once again reflects the poor results of the textual representations for the sporulation group as described earlier. The fact that signal transduction, cell shape and size control, cell adhesion and sporulation are related according to the dendrogram is no surprise either. For instance, the activation of the MAPK pathway, the main signal transduction pathway in yeast, results in a complex series of cellular events leading to mating, sporulation, filamentous growth and so on. These events include changes in cell shape ('shmooing', elongation), cell cycle, budding pattern, and cell-cell connections and in increased transcription of specific genes.

Selected cell cycle groups

The interrelation of the cell cycle groups is visualized in the dendrogram printed in Figure 5.11. We see that the various metabolic processes (fatty acids, glycosylation, methionine metabolism, nutrition and secretion) are clustered closely together. This is no surprise since metabolism is a highly integrated process. Individual metabolic pathways are linked into complex networks through common, shared substrates. Additionally, the majority of these processes (oxidative phosphorylation, the citric acid cycle, amino acid catabolism and fatty acid oxidation) share the same subcellular location, namely the mitochondrion. Cell cycle control, mitotic exit (one of the key events in the cell cycle) and the formation of pseudohyphae (a response to nitrogen starvation that is tightly controlled at the G1/S transition of the cell cycle) are closely related as expected. The processes of DNA repair and sporulation are also linked together, most probably because a number of proteins (RAD proteins), which are implicated in post-replication repair and damage-induced mutagenesis, are also required for sporulation by modulating the chromatin structure via histone ubiquitination.

5.3.3 Conclusion

Through a quantitative analysis on several test cases we have shown that clustering genes based on their text representation is a viable way to divide genes into functional groups. As with any clustering approach (see Section 3.2), success depends highly on a good representation. For example the reference, representation proved particularly unsuccessful here. A main reason for this is that genes sharing no referenced articles are completely dissimilar in this representation. The term representations on the other hand show a strong dependence on the used weighting schemes, as well as on the type of information source (database annotations, MEDLINE abstracts, ...) used. Contrary to Section 5.2 where we used instead of curated literature references, we have worked here with MED-

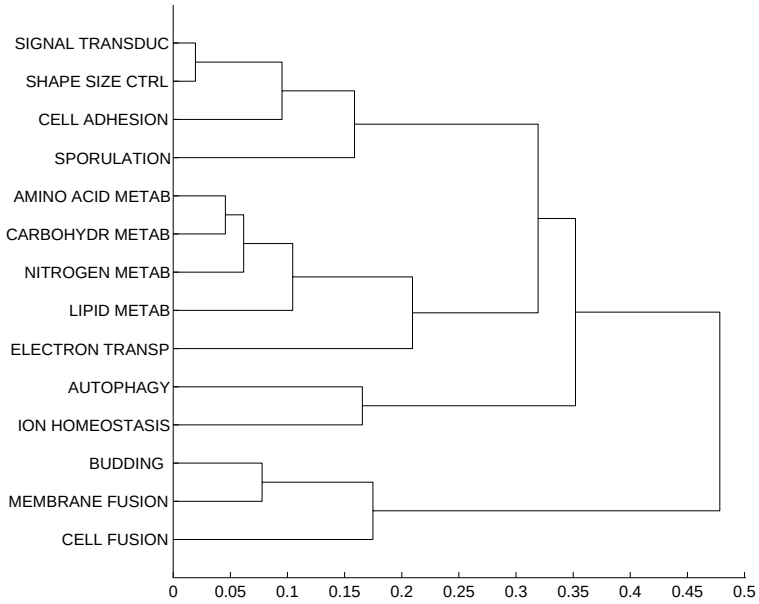


Figure 5.10: Dendrogram of the 13 functional groups from GO (906 genes). We refer to the text for an extended discussion.

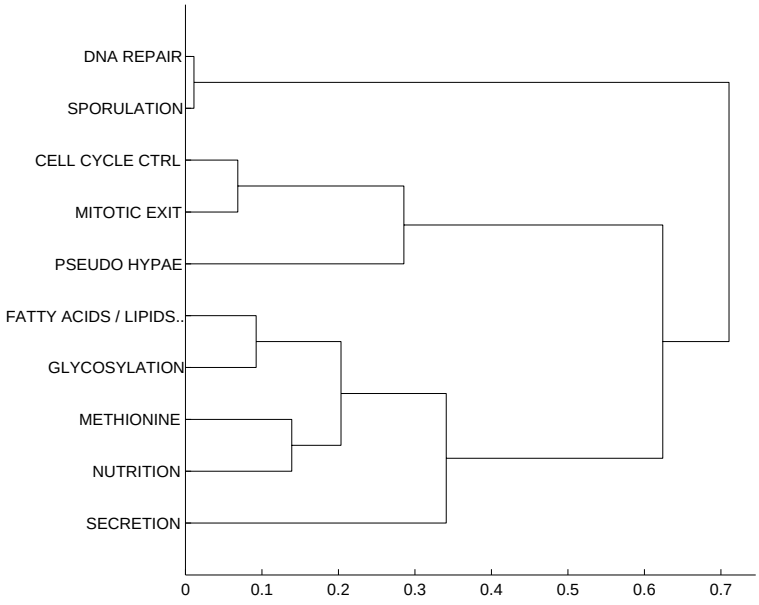


Figure 5.11: Dendrogram of the 10 cell cycle specific groups (126 genes). We refer to the text for an extended discussion.

LINE abstracts that were retrieved using the cosine in the vector space model. Nevertheless, we do not expect the outcomes to change in the reverse case. Finally, some limitations of the text representation were pointed out but they will be discussed more in depth in the following chapters.

5.4 Discussion

In this chapter we investigated the potential of the vector-based representation to score and cluster groups of genes based on text. Assuming the existence of relevant MEDLINE assigments to each of the genes under study, we presented a framework to represent genes in term space by pooling linked textual information. We presented an analysis on how these textual representations can *recognize* functional gene groups based on text. Various ways to derive a quantitative text-based score that screens for functional enrichment in a group of genes were treated.

In a following step we continued to investigate various aspects of the text representation, including how it *interrelates* functional gene groups through clustering. We provided a biological discussion as well as a set of quantitative indications supporting the added value of MEDLINE abstract information with respect to database annotations.

These results point out that textual data should be treated on equal footing with other types of data, including high-throughput data, sequence data, or ontological data. Moreover, a developed representation of text opens the gate to application of the same variety of statistical methods as in the field of microarray data analysis (as illustrated with the text clustering). In Chapter 7 we will further develop this idea and process textual information *jointly* with expression data, whereas the *interpretation* of gene groups through high-valued keywords is the subject of the following Chapter 6.

Chapter 6

TXTGate, a platform for profiling genes based on text

Recent advances in high-throughput methods like microarrays enable a more systematic testing of the function of multiple genes, their interrelatedness and the controlled circumstances in which ensuing observations hold. As a result, scientific discoveries and hypotheses are stacking up, all primarily reported in the form of free text. However, since the data and information lack usability when offered in a raw state, various specialized database systems have been implemented to provide a complementary resource in designing, performing, or analyzing large scale experiments. Linking up experimentally derived gene groups (e.g., from microarray cluster analysis) to the existing databases and published literature still requires numerous queries and extensive user intervention. This process of drilling down into the entries of hundreds of genes is notably inefficient and requires higher-level views that can more easily be captured by a (non-)expert's mind. Figure 6.1 demonstrates the role of the text mining application TXTGate in this problem. We also recall Figures 1.6 and 1.7 in the introductory chapter.

In Chapter 5 we provided a *quantitative* analysis of the text representation using gold standards. With TXTGate, the methodology presented in this chapter, we attempt to leverage the practical usability of text mining. TXTGate [33] combines literature indices of selected public biological resources in a flexible text-mining system designed towards the analysis of groups of genes. By means of tailored vocabularies, term- as well as gene-centric views are offered on selected textual fields and MEDLINE abstracts. Subclustering and links to external resources allow for in-depth analysis of the resulting term profiles[1].

[1]The main ideas underlying the TXTGate web application were first presented at the First International Workshop on Semantic Web and Databases [82]. A description of the mature version of the software, along with elaborated case studies demonstrating its full potential is

We present cases studies that illustrate how TXTGate assists in answering questions such as:

- Which functional terms are related to a group of genes ?
- Which disease terms are related to a group of genes ?
- Which other genes are linked in literature to a group of genes[2] ?
- Can this group of genes be further divided according to thematic information ?

By qualitatively comparing our results to other work and extending the analysis on some provocative examples, we transform the previously presented text mining approach from a 'proof of principle' to a genuine knowledge discovery tool that is useful to a biologist. We adopted two validation mechanisms: Firstly, comparisons were made to a variety of previous studies listed in Table 6.3. Secondly, opinions on the text profiles were provided by a biological expert in cases where no supporting information was present in the accompanying publications.

As a concluding remark we mention that many references to external databases occur in this chapter; therefore, we remind the reader that concise descriptions can be consulted in Appendix.

6.1 Research context

Until now it has been largely overlooked that there is little difference between retrieving an abstract from MEDLINE and downloading an entry from a biological database [78]. Fading boundaries between text from a scientific article and a curated[3] annotation of a gene entry in a database is readily illustrated by the GeneRIF feature in LocusLink [127], where snippets of a relevant article pertaining to a gene's function are manually extracted and directly pasted as an attribute in the database. The broadening of the biologist's scope, along with the swelling amount of information, results in a growing need to move from single-gene or keyword-based queries to more refined schemes that allow comprehensive views of text-oriented databases.

As gene expression studies typically output a list of dozens or hundreds of genes that are co-expressed, a researcher is typically faced with the assignment of biological meaning to such lists. Several text mining approaches have been developed to this end. Masys et al. [110] link groups of genes with relevant MEDLINE abstracts through the PUBMED engine. Each cluster is characterized by a pool of the relevant keywords derived from both the MeSH headings and UMLS ontology. Their interface reports the quantitative significance of

currently in press in *Genome Biology* [81].

[2]We will call this co-linkage.

[3]Curation refers to the hand-crafted maintenance of quality and consistency of biological information including sequences, annotations and cross-references.

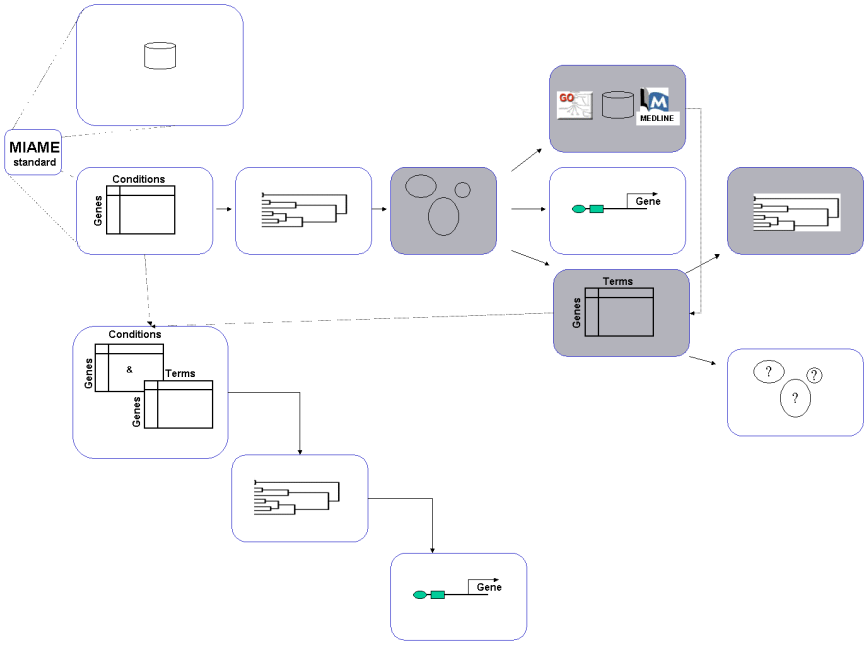


Figure 6.1: Starting from a cluster of genes resulting from a gene expression analysis in the 'Data World', corresponding literature profiles are queried and analyzed in the 'Text World'. This results in either the addition of extra genes of interest or the omission of irrelevant genes. Various information sources such as GO, LocusLink,... are implicated in the text analysis.

each result and provides links to different databases to allow further browsing. Jenssen et al. [97] constructed a pioneering online system to link co-expression information from a microarray experiment with their constructed co-citation network. This literature network covers co-occurrence information of gene identifiers in over ten million MEDLINE abstracts. Their system characterizes co-expressed genes using the MeSH keywords attached to the abstracts about those genes. Shatkay et al. [136] link abstracts to genes in a probabilistic scheme that uses the EM algorithm to estimate the parameters of the word distributions underlying a 'theme'. Genes are identified as similar when their corresponding gene-by-documents representations are close. Chaussabel et al. [60] and Glenisson et al. [80] provide a proof of principle on how clustering genes encoded in a keyword-based representation can further discern relevant subpatterns. Finally, Raychaudhuri et al. [131] developed a method called neighborhoud divergence, to quantify the functional coherence of a group of genes using a database that links genes to documents. The score is successfully applied to both gold standard and expression data, but has the slight drawback that it does not tell the function itself. Their method is indeed geared at the identification of biologically coherent groups, rather than at their interpretation. Although all these approaches represent nice text mining work, we have witnessed, as of February 2003, the lack of a flexible and user-friendly approach that maximally combines the advantages of term-based text mining with free-text information stored in important public resources. As a result, we have created a pilot text mining system named TXTGate that is built using several prevalent biological resources and that aims at implementing the interactive (or cyclic) nature of the knowledge discovery process (see Figure 1.6).

To ensure the quality and interpretability of term profiles we put forward the three major considerations on which our system is built. Firstly, the construction of high-quality associations between genes and relevant MEDLINE abstracts is often problem-dependent and constitutes a research track on its own that requires advanced document classification strategies as, for example, proposed by Leonard et al. [104] or Raychaudhuri et al. [130]. Despite some shortcomings, curated gene-literature references are therefore grateful resources to exploit. Secondly, the information contained within curated gene references is sometimes diverse and can range from sequence to disease. Additionally, the research questions that scientists are addressing when they scrutinize gene groups from high-throughput assays, display a similar diversity. Therefore, considering all terms occurring in a large set of documents (i.e., a general vocabulary) might be detrimental to the extraction of terms that are relevant to the question at hand. Hence, the construction of separate vocabularies according to gene name, disease and function seems a logical choice to provide increased insight. Thirdly, as mentioned previously, annotations offered by curated gene databases are often in semi-structured form and encompass keywords, sentences or paragraphs. To facilitate integration of such annotations with existing knowledge, controlled vocabularies that describe conceptual properties are of great value when constructing interoperable and computer-parsable systems. A number of structured vocabularies have already arisen (GO [9], MeSH [22], eVOC [101])

and, slowly but surely, certain standards are systematically being adopted to store and represent biological information [12].

Armed with these three insights, we developed TXTGate, a platform that offers multiple 'views' on vast amounts of gene-based free-text information available in selected curated database entries and linked scientific publications. TXTGate enables detailed functional analysis of interesting gene groups by displaying key terms extracted from associated literature and by offering options to link out to other resources or to subcluster the genes based on text. We address the need for easy validation of gene clusters arising from e.g., microarray experiments on the one hand, and the problem of using scientific literature (in the form of free text) as a source of functional information about genes on the other hand. The strength of TXTGate is its use of tailored vocabularies in order to visualize only the information most relevant to the query at hand. The application is publicly available at http://www.esat.kuleuven.ac.be/txtgate.

6.2 System design

6.2.1 Features

TXTGate is implemented as a web application designed to process queries of up to 200 genes, although this is not a strict limitation of the system. We believe the application scales well enough to be of use in, for example, microarray cluster validation. A conceptual overview of the system is shown in Figure 6.2. Various literature indices were created based on selected annotation fields and linked MEDLINE information both present in curated repositories (LocusLink and SGD). Several tailored vocabularies derived from public resources (GO, MeSH, OMIM, eVOC and HUGO) act as a perspective to the textual information. A user-defined query on either of these indices regarding a gene group (currently TXTGate smoothly accomodates queries of ~ 200 genes) results in a summary profile that can be used for further query building for a variety of other databases (Figure 6.3). Alternatively, a gene group of interest can be subclustered based on the selected textual information to discern substructures not apparent in the original summary profile.

6.2.2 Literature indices in TXTGate

All textual information is indexed with a normalized IDF weighting scheme (see Chapter 5), as it turned out to be a reasonable choice for modeling pieces of text comprising up to ~ 200 terms, as observed in database annotations and MEDLINE abstracts. We offer various operations on the resulting indices:

Combining multiple, linked documents into a single gene profile
As explained in Section 5.2, when a given gene has more curated MEDLINE references associated to it, we combine these abstracts into an indexed gene entry by taking the *mean* profile.

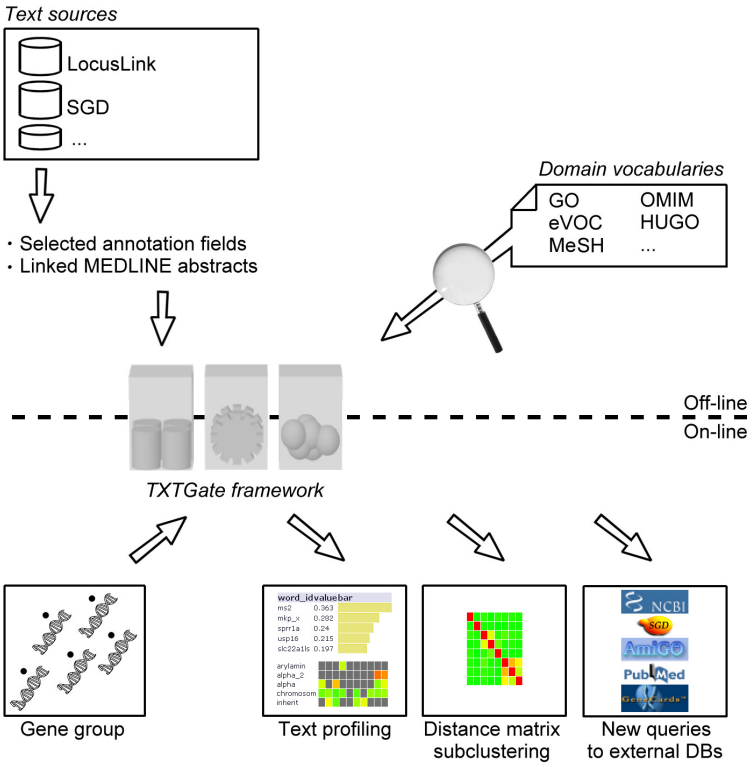


Figure 6.2: Conceptual overview of TXTGate. **Offline:** We indexed two different sources of textual information about genes (LocusLink and SGD) using different domain vocabularies.**Online:** These indices are used for (a) term-based gene profiling and (b) clustering of interesting gene groups. TXTGate's link-out feature (c) makes it possible to use interesting terms for further querying external databases (Figure courtesy: Steven Van Vooren).

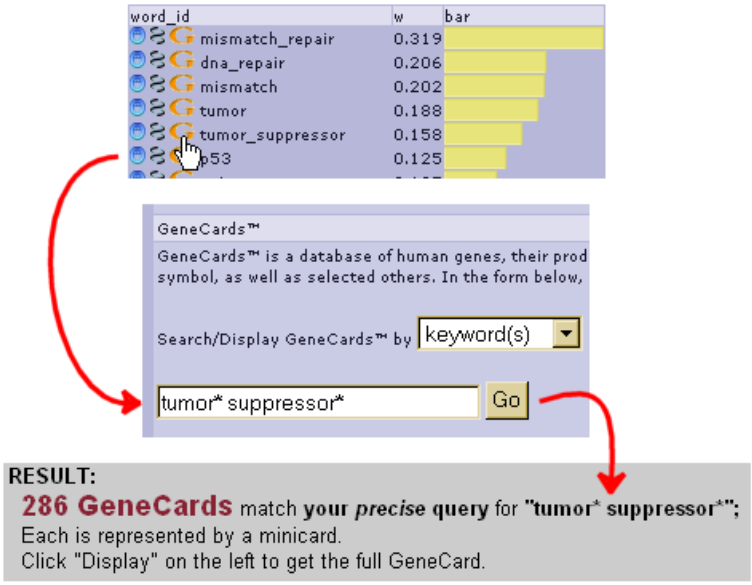


Figure 6.3: TXTGate profile and link-out feature. High scoring terms (e.g., 'tumor suppressor') of a gene group are presented and can be directly used to compose a query to external databases (GeneCards in this case) by clicking the corresponding icons that precede the terms.

Combining multiple gene profiles into a group profile

To summarize a cluster of genes and explore the most interesting terms they share, we compute the mean and variance of the terms over the group. Although simple, these statistics already reveal information on interesting terms characterizing the gene group. This is performed on-line.

Subclustering gene profiles

We offer the possibility to subcluster, on-line, a gene group of maximum 100 genes by means of hierarchical clustering (Ward's method), which was chosen for its deterministic nature and the computational advantage of using the same solution when consecutively considering different numbers of clusters k.

We construct five different *term*-centric domain vocabularies that provide different views on the gene-specific information we index. A first vocabulary is derived from the Gene Ontology (GO) [9] and comprises 17,965 terms. GO is a dynamic controlled hierarchy of (multi-word) terms with a wide coverage in life science literature and genetics in particular. We consider it as an ideal source to extract a highly relevant and relatively noise-free domain vocabulary. For the yeast indices, we prune this vocabulary by retaining those terms occurring at least twice and in less than 20% of all MEDLINE abstracts referred to in Saccharomyces Genome Database (SGD) [29], hereby obtaining a new vocabulary of 3,867 terms. Two other domain vocabularies are rather similar in scope but differ in size. One is based upon the Medical Subject Headings (MeSH) [22], the National Library of Medicine's controlled vocabulary thesaurus and counts 27,930 terms. The other is based on the Online Mendelian Inheritance in Man's (OMIM) [26] Morbid Map. The latter is a cytogenetic[4] location of all disease genes present in OMIM and their associated diseases. We extract all disease terms to construct a 2,969-term vocabulary. A fifth domain vocabulary is drawn from eVOC [101], a thesaurus consisting of four orthogonal controlled vocabularies encompassing the domain of human gene expression data. It includes anatomical system-, cell type-, pathology- and developmental stage-related terms.

Beside these term-centric domain vocabularies we construct two *gene (symbol)*-centric vocabularies with the screening of co-occurring and co-linked genes in mind. *Co-occurrence* denotes simultaneous presence of gene names within a *single* abstract, as in Jenssen *et al.* [97]. We define *co-linkage* here as a weaker form of co-occurrence that screens for simultaneous presence in the *pool* of abstracts that are linked to a given group of genes. We derive from the HUGO database [113] a vocabulary of all uniquely defined human gene symbols and their synonyms. In total this vocabulary consists of 26,511 gene symbols. The second vocabulary consists of all uniquely defined yeast gene symbols found in SGD and contains 11,319 terms. Since these official gene symbols are frequently requested and used by scientists, journals and databases, we assume

[4] a cytogenetic marker on a (cytogenetic) map denotes a part of a chromosome with a particular structure or composition that is interesting for biological or medical reasons (often: disease).

Table 6.1: Overview of the domain vocabularies in TXTGate. The vocabularies are named after the resource they stem from. Copious examples will be provided further on.

Domain Vocabulary	**Nr. of terms**
Term-centric	
GO	17965
GO-pruned yeast	3867
MESH	27930
OMIM	2969
eVOC	1553
Gene-centric	
HUGO_symbols	26511
SGD_gene_names	11319

Table 6.2: Overview of the indexed resources of textual information in TXTGate. In the second column we specify which fields of the resource were used. The third column lists the domain vocabularies with which the information was indexed.

Resource	**Information fields**	**Used domain vocabularies**
Locuslink	linked MEDLINE	GO, MeSH, eVOC,OMIM, HUGO_symbols
Locuslink	grif_txt	GO
Locuslink	summary sum_func	GO
Locuslink	go_term ext_annotation_term	GO
SGD	linked MEDLINE	GO-pruned, SGD_gene_names
SGD	GO_term	GO-pruned

they constitute a good first approximation to detect gene occurrence in MEDLINE abstracts. In Table 6.1 we list the adopted domain vocabularies.

We downloaded the entire LocusLink (as of April 8 2003) and SGD (as of January 15 2003) databases and identify and index subsets of fields (such as GO annotations and functional summaries) that were most sensible in the presented context. Although indexing these database entries could have been done on all fields at once, we deem a preservation of selected parts of LocusLink's and SGD's logical field structure more appropriate for functional gene profiling. Besides the textual annotations, we index the 73,152 MEDLINE abstracts referred to in all entries of LocusLink, as well as the 24,909 abstracts linked to in SGD. All information sources are processed with different domain vocabularies and the resulting indices are used in TXTGate as a basis for literature profiling on and further query building of genes of interest. Table 6.2 overviews the indexed resources of textual information and connects them to the domain vocabularies used.

6.3 Related software

The work presented hereafter extends the general ideas of textual profiling and clustering presented in Blaschke et al. [53] and Chaussabel et al. [60], where the usability of literature indices for profiling gene groups in yeast and human is proven. TXTGate implements the vector space model for gene profiling and provides indices, not only for MEDLINE abstracts, but also for selected functional annotations from two public databases. Various engineered domain-specific vocabularies (term- as well as gene-centric) act as perspectives to the literature and the tool provides direct links to external resources. In what follows, we compare TXTGate to other reported biological text mining software. MedMiner [147][19] retrieves relevant abstracts by formulating expanded queries to PUBMED. It uses entries from the GeneCards database [132][10] to fish for additional relevant keywords to expand a query. The resulting filtered abstracts are comprehensibly summarized in keywords and sentences, and feedback loops are provided. Nevertheless, the system is directed at querying terms and specific gene-drug or gene-gene relationships, rather than at scrutinizing gene clusters. MedMOLE [57][20] is also a system to query MEDLINE more intelligently and detects HUGO names in abstracts via a natural language processing (NLP)-based gene name extractor. The retrieved abstracts can be clustered, and top keywords are presented. However, the application scales less well (20 genes were experienced problematic as of November 2003), is not effective at profiling groups of genes, and the presented summaries provide much less detail than MedMiner and TXTGate. GEISHA [53][8] is a tool for profiling gene clusters with an emphasis on comprehensible summarization within a shallow parsing framework. This system was implemented for *E. Coli* but is no longer available. PubGene [97][27] is a database containing gene co-occurrence and co-citation networks of human genes derived from the full MEDLINE database. For a given set of genes it reports the literature network they reside in, together with their

high-scoring MeSH terms. As not all relevant information can be captured by gene symbols or MeSH terms, the functionalities offered by TXTGate provide complementary views to interpret groups of genes. Although our co-linkage feature (being a weaker form of co-occurrence that spans only the set of 73,152 MEDLINE abstracts used in LocusLink) is less elaborated than the possibilities offered by PubGene, we will show its usability and added value through the integration in the broader TXTGate framework (see further). MedGene [123][18] and G2D [93][7] are specialized databases that, contrarily to TXTGate, are geared at ranking genes by disease. They accept user-defined queries scrutinizing gene-disease, disease-disease, or gene-gene relationships extracted from literature. Finally, MeKE [61][21] is an application listing gene functions extracted by an ontology-based NLP system. Their current setup is more directed towards a functional knowledge base, rather than comprehensibly profiling information coming from groups of genes as offered by our software.

6.4 Experiments

To demonstrate the proper functioning of TXTGate we test various features of the application on a variety of data sets in the following two sections. We overview the data sets in Table 6.3, which supports Table 1.1 in the introductory chapter. We distinguish gene groups derived from expression data and those matching a given functional query (against a database or ontology). The presented studies are organized according to the two organisms involved. *Saccharomyces cerevisiae* (baker's yeast) is a heavily researched model organism for eukaryotes and many publicly available (expression) data sets exist. On the other hand, its great complexity, extended uncharted territory and the rewarding prospect to better understand diseases makes the human genome a still greater challenge to explore. In both cases comparisons to results of previous text mining assays as well as novel investigations are presented.

The overall setup is qualitative (i.e., demonstrating biological relevance) and is meant to address

- differences in information captured by various database fields and domain vocabularies;
- the usability of text profiling, subclustering and link-out features offered by TXTGate;
- how text-mining human data opens the path to exploring diseases through automated literature processing.

6.5 Study 1: *Saccharomyces cerevisiae*

Various data sets on this model organism will be used

Table 6.3: Overview of the data sets used in our experiments. We use data sets stemming explicitly from gene expression analysis on the one hand ('Expression-based') and data resulting from queries to GO or OMIM on the other hand ('Function-based'). In case of expression data the first column shows the original publication(s) accompanying the data set. The second column shows subsequent analyses from which the results are used to compare against.

Data	Analysed in	Organism
Expression-based		
DeRisi[70]; Spellman[141]; Chu[63]*et al.*	Eisen *et al.*[73]	yeast
DeRisi[70]; Spellman[141]; Chu[63]*et al.*	Blaschke *et al.*[53]	yeast
All (30) yeast expression datasets as of 06/2001[14]	Ihmels *et al.*[95]	yeast
Nau *et al.*[118]	Chaussabel *et al.*[60]	human
Function-based		
Autophagy group from GO		yeast
Colon cancer gene group from OMIM		human

- to compare and benchmark our (GO-based) term profiling to previous work and demonstrate assets of the TXTGate software (Section 6.5.1 and 6.5.4),
- to demonstrate practical situations where clustering of text profiles can be prove useful (Section 6.5.2),
- to corroborate our setup with respect to noisy information (Section 6.5.3), and
- to show how an occurrence analysis of gene names in abstracts can help to find new, potentially relevant, genes (Section 6.5.5).

6.5.1 Text profiling of yeast expression clusters

We profile the clusters derived from the yeast gene expression data set in Eisen *et al.* [73]. The experiments describe variation in gene expression under various experimental conditions, including mitotic cell cycle, sporulation, diauxic shift and temperature. In Tables B.1, B.2 and B.3 (Appendix) we show detailed text profiles for clusters $\{B, C, D, E, F, G, H, J$ and $K\}$(corresponding to 136 genes in total) in Figure 2 in Eisen *et al.*. Terms from MEDLINE abstracts as well as from GO annotations in SGD are shown. In clusters B and C we demonstrate the complementary nature of the MEDLINE information: Terms extracted from MEDLINE such as 'spindle pole body','cytokinesis' and 'anaphase promotor complex' are more reminiscent of the overall function of cluster B than the most of the less-specific GO-extracted terms such as 'cytoskeleton', 'protein binding' and 'bud neck'. For cluster C, the difference is even more pronounced as the

GO term 'degrad' is the only one referring (vaguely) to the asserted proteasome function of this cluster. One exception is the GO-term 'microtubuli nucleation', which refers to a process that contributes to the biogenesis of the microtubule cytoskeleton. For the remaining clusters all terms extracted from MEDLINE and GO annotations clearly refer to the same biological function. This result strengthens the findings in Section 5.3 that term-based analysis of literature can be valuable to complement annotations in public databases. We compare the quality of our analysis which is based on the vector space model with the analysis in Blaschke *et al.* [53]). Their approach called GEISHA extract group-specific keywords by means of a z-score analogous to (4.3). Mutually corresponding keywords are highlighted in bold.

For clusters B, C, G, H, J and K we were able retrieve most keywords labelled relevant by Blaschke and colleagues. For cluster D we do not find back 'zinc finger' and 'cytochrome b', but they constitute more general terms. We found the TXTGate term 'glyceraldehyde-3-phosphate-dehydrogenase' in cluster E more closely related to glycolysis than 'glycerol-3-phosphate dehydrogenase', the latter being specifically related to alcoholic fermentation. Examination of the annotations in Eisen *et al.* do not suggest an anaerobic breakdown of energy and hence confirm the TXTGate result. 'Aminoacyl tRNA synthetase' returned by GEISHA for cluster F is an enzyme playing a crucial role in the coupling of tRNA's to the correct amino acid in the translation process. However, this enzyme is not specific to the *mitochondrial* synthesis of proteins and the occurrence of this term does not add information to the function of this cluster. 'Cytochrome oxidase' on the other hand, a term shared by TXTGate and GEISHA, is a mitochondrial-specific term that refers to one of the three respiratory enzyme complexes in the inner mitochondrial membrane. In our MEDLINE profiles, we find 'cytochrom' as isolated term and due to the limitations of our phrase detection method we retrieve the more specific 'cytochrome-c-oxidase' only in the GO profiles. This again is not surprising as our domain vocabulary is built on GO and we can not always expect they are consistently used in literature. Incorporation of a more curated phrase list, however, can easily accommodate for this problem. Some of the terms produced in the GEISHA analysis are irrelevant and labeled '(-)'. Their incidence is due to the retrieval of spurious abstracts (e.g., retrieval of abstracts pertaining the organism *Saccharomyces pombe*, or fission yeast, or problems with polysemy of gene nomenclature) . TXTGate did not suffer from the particular difficulties they report, mainly due the adoption of the gene-literature resource in SGD and the separation of term- and gene-centric vocabularies in the TXTGate conception.

We conclude that our methodology is competitive with the GEISHA approach to organize functional gene information. In what follows we demonstrate how the flexibility of the underlying vector space representation is exploited in TXTGate.

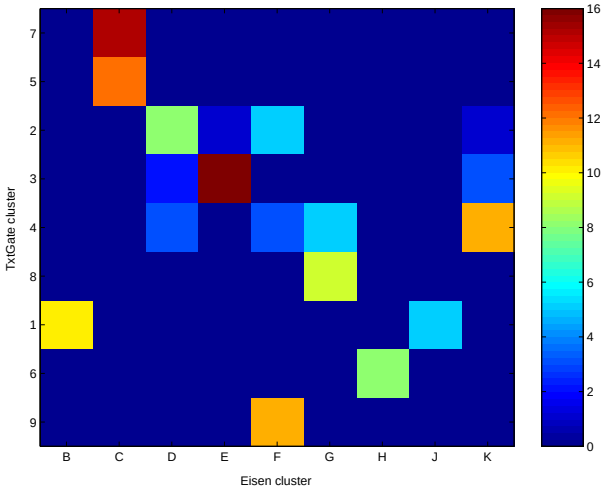


Figure 6.4: A comparison of TXTGate clusters with clusters computed in Eisen *et al.* The color bar indicates how many genes are shared by corresponding TXTGate and Eisen clusters. R_{adj} is 0.5351.

6.5.2 Text clustering and the detection of subclusters

Clustering a set of expressed genes

To understand in which way the text representation differs from expression data, we divide the set of genes from Eisen *et al.* into $k = 9$ clusters based on text and compare this with the composition of the expression clusters. The partitioning of genes in text clusters versus the clusters computed by Eisen *et al.* is depicted in Figure 6.4. The adjusted *Rand* index, R_{adj}, is 0.5351 which indicates a significant, moderate correspondence. We proceed with examining how the two clusterings differ.

Cluster B en J are grouped in Cluster 1, hereby aggregating genes involved in the chromosome replication (J) and progression to cytokinesis (B) in the M-phase of the cell cycle. Cluster 3 corresponds entirely to cluster E augmented with two genes involved in glucose repression (STD1) and acetyl-COA metabolism (ACH1) respectively. Genes related to oxidative phosphorylation and the cytochrome c complex that occur over clusters K, G, D and F are joined in TXTGate cluster 4. In cluster 5 we observe a subdivision of cluster C containing all components related to the 26s proteasome complex, whereas cluster 7 contains all components related to the 20s proteasome complex in cluster C. Cluster 6 corresponds to H and we find all ATP-related components G of cluster G in TXTGate cluster 8. Cluster 9 maps to F for the majority of the genes, whereas the other genes are contained in cluster 2 (cfr. supra). Finally, cluster 2 groups genes related to mRNA splicing (9/13 genes from cluster D), protein synthesis in the mitochondria (cluster F) and genes involved in the cell's oxida-

tive reactions, including succinate dehydrogenase activity in the mitochondria (cluster K). This indicates cluster 2 is spurious, and should be further divided. In the following section we will resolve this by further subclustering cluster 2.

Subclustering a group of interest

The nature of the expression data or the particular behavior of a clustering algorithm can yield the presence of heterogeneous gene functions within a gene cluster, hereby complicating at times its interpretation. Conducting a deeper analysis by further subclustering can give more insight in substructures not apparent when analyzing the whole cluster.

Firstly, we perform a subclustering of cluster E from Eisen *et al.* as done in Blaschke *et al.*, and we compare our results to their similar analysis. In Table 6.4 we show the text profiles of cluster E after subclustering with $k = 2$. High scoring terms are shown for both subclusters *E1* and *E2*. We also print the terms (excluding those denoting gene symbols) resulting from a similar analysis conducted by Blaschke *et al.* Terms that were labeled specific to a subcluster by Blaschke are marked with an asterisk (*). Although several of the text mining settings in Blaschke *et al.* are different from ours (due to the differences in MEDLINE corpus, textual analysis methodology and used clustering algorithm), a comparison of the term profiles in both analyses shows that TXTGate also identifies *E1* as being related to glycerol, whereas *E2* is more related to the pyruvate metabolism and ethanol fermentation. We thus distinguished two viable sub-patterns not immediately apparent from the expression data alone.

Secondly, we sub-divide the (spurious) TXTGate cluster 2 from Section 6.5.2, with $k = 3$. In Figure 6.5, we visualize how the top scoring terms for cluster 2 are distributed over the three subclusters 2.1, 2.2 and 2.3. Subcluster 2.3 groups genes related to succinate dehydrogenase (part of TCA pathway), which is supported by strong weighing (red) of this term in Figure 6.5. Genes involved in RNA splicing aggregate in subcluster 2.2. Nevertheless, closer inspection learns that subcluster 2.1 contains genes POR1, NDI1, MDH1 and PPA2 that are functionally more related to genes that are involved in the aerobic metabolism and TCA cycle and that belong to subcluster 2.3. This is also apparent in the text profiles where terms as 'respiratory', 'malate dehydrogenase' and 'succine' score high for these genes. Failure to classify these genes correctly is in part due to failure to encode correlations between these terms (and hence the genes) in the vector space model. Also a word such as 'mitochondria', which refers to subcellular localization of the gene product, causes erroneous functional associations between genes by their contributions to the similarity measure. Elimination of such words from the domain vocabulary is not evident as many functional descriptions include terms not only from the biological process the gene products participate in but also from their targeted locations.

From the comparison between text- and expression-based partitioning we see that literature information tends to divide genes into functional categories that can span multiple phases. Hence cohesion in expression space will not always be equivalent to cohesion in text space. This observation will be further

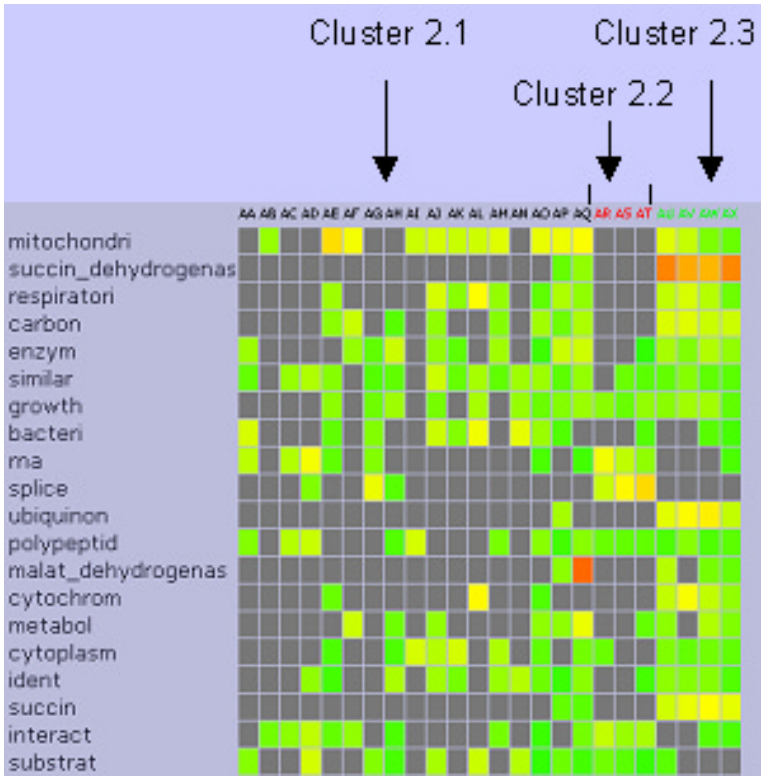


Figure 6.5: Term-by-gene matrix for TXTGate cluster *2* of the investigated Eisen data. To understand the poor interpretability of this cluster in terms of function, we subclustered and reorganized it into subgroups *2.1*, *2.2* and *2.3*, marked on top. The color in each square marks the importance (grey= not relevant, green=moderately relavant, yellow=highly relevant) of term (row) to a particular gene (column). Subcluster 2.3 groups genes related to succinate dehydrogenase (part of TCA pathway), which is supported by strong weighing (red) of this term. Genes involved in RNA splicing aggregate in subcluster 2.2. Subcluster 2.1, however contains some genes that are functionally more related to genes involved in the aerobic metabolism and the TCA cycle and were rather expected in subcluster 2.3. Failure to classify these genes correctly is in part due to failure to encode correlations between these terms (and hence the genes) in the vector space model.

elaborated on in Chapter 7. Further analyzing given gene clusters (Eisen cluster E and TXTGate cluster 2) by subclustering can reveal interesting subtopics. Nevertheless, the text representation should ideally also incorporate associations between terms.

6.5.3 Random, corrupted and ambiguous clusters

In this section we conduct a (qualitative) robustness analysis of the term profiles. We discuss a number of factors that can influence coherence in the themes identified by a literature-based analysis with TXTGate. To this end we extract from GO a group of 27 genes related to autophagy. Some of these genes have other functional classifications as well. In total 316 references to 165 distinct MEDLINE abstracts are retrieved for these genes.

Corrupted clusters

A quantitative assessment of the robustness of the vector space model was already treated in Chapter 5. Here we are interested in the behavior of the term profiles given by TXTGate: Using the autophagy cluster we examine how the term profiles are affected by a random cluster spanning 30 genes, and a 50% diluted autophagy cluster where half of the autophagy genes have been replaced by random ones. In Table 6.5 we see that the random cluster contains mostly unspecific terms with small weights assignments. The corrupted cluster on the other hand still identifies the major consistent theme in the cluster (autophagy, degradation, vacuolar), though with smaller weights than in the full autophagy cluster. The occurrence of the unrelated term 'vesicle' is explained below.

Ambiguity through multiple gene function and quality of associated abstracts

It has been argued [131] that as most genes have multiple functions, the associated literature will not be coherent either and as a result, text mining methods should address this problem. To understand how our gene representation (which collapses all related articles into a single vector) deals with this issue, we focus on the following two particular autophagy genes, AUT10 and SEC23, whose GO categorizations are provided in Table 6.6: AUT10 is clearly related to autophagy which is supported by five references, whereas SEC23 is only marginally related to autophagy as only one (!) out of fifty-eight references covers this topic. The bulk literature on SEC23 focuses on the gene's primary role in secretion.

We observe two related difficulties here. Firstly, in contrast to what is suggested from GO, the weak association between SEC23 and autophagy may have detrimental effects on the term profiles of the autophagy cluster. However, experiments on replacements with random genes conducted in previous section, already suggested the robustness of the profiles. Secondly, as some genes are better described than others, an unbalanced number of literature references can

Table 6.4: TXTGate profiling of cluster *E* from Eisen *et al.* by subclustering (k=2). High scoring terms are shown for each subcluster *E1* and *E2*. We also print the terms (excluding gene names) resulting from a similar analysis conducted by Blaschke *et al.* Terms that were labeled specific to a subcluster by Blaschke are marked with an asterisk (*). Although several of their settings are different from ours (due to the differences in MEDLINE corpus, textual analysis and used cluster algorithm), a comparison of the term profiles in both analysis shows that TXTGate also identifies *E1* being related to glycerol, whereas *E2* is more related to the pyruvate metabolism and ethanol fermentation.

Gene Symbol		**Cluster Terms in Blaschke *et al.***	**Terms from TXTGate**
subcluster E1			
TPT1	FBA1	glyceraldehyde-3-phosphate*	glyceraldehyd_3_phosphat_dehydrogenas
GPM1	TKL1	glyceraldehyde-3-phosphate dehydrogenase*	glycolyt
PGK1	CDC19	phosphoglycerate kinase*	glucos
TDH3	HXK2	phosphoglycerate*	enzym
TDH2	TYE7	mutase*	glycolysi
ENO2	PFK1	dehydrogenase	carbon
TDH1	ACS2	enolase	pyruv_kinas
		glycerol-3-phosphate dehydrogenase	ethanol
		osmotic stress	phosphoglycer_kinas
		phospoglycerate	growth
subcluster E2			
PDC5	PDC1	alcohol*	pyruv_decarboxylas
PDC6		transketolase*	pyruv
		catabolite repression	glucos
		decarboxylase	enzym
		ethanol	alcohol
		glucose	decarboxyl
		glucose repression	ethanol
		hexokinases	ferment
		pyruvate	thiamin
		pyruvate decarboxylase	decarboxylas

Table 6.5: Term profiles from a random cluster of 30 genes, an autophagy-related cluster encompassing 27 genes and a 50% mixture of these two. The random cluster contains mostly unspecific terms with a small weights. The corrupted cluster still identifies the major consistent theme in the cluster, though with smaller weights than in the full autophagy cluster.

Random cluster		**50% corrupted autophagy cluster**		**Full autophagy cluster**	
term	mean term	mean	term	mean	
cytoplasm	0,0613	autophagi	0,160739882	autophagi	0,364347316
growth	0,0567	vacuol	0,095910451	vacuol	0,157904001
similar	0,0548	autophag	0,074861691	autophag	0,148262374
chang	0,0491	vacuolar	0,071953443	starvat	0,13367178
codon	0,0481	cytoplasm	0,068951504	aminopeptidas	0,113242263
duplic	0,0476	starvat	0,063492904	vacuolar	0,108454675
characterist	0,0446	degrad	0,058827786	lysosom	0,093778269
chromosom	0,0442	vesicl	0,054584484	cytoplasm	0,090473303
sensit	0,0408	aminopeptidas	0,053926445	vesicl	0,089509491
synthesi	0,0388	characterist	0,052494531	degrad	0,089506498

Table 6.6: GO annotation of two particular autophagy related genes. AUT10 is clearly related to autophagy, which is supported by five references. SEC23 is only marginally related to autophagy as only one out of sixty-three references cover this. The bulk literature focuses on the gene's primary role in secretion.

Gene	Subcellular location	Molecular function	Biological process
AUT10	cytosol	unknown	autophagy sporulation vacuolar protein processing protein-vacuolar targeting
SEC23	COPII vesicle coat cytoplasm	GTPase activator	ER to Golgi transport autophagy

bias the profiles in favor of excessively referenced functions. With the above two genes being excellent instantiations of these issues, we address following points:

1. How informative is the text profile of the document pool (i.e., $58 + 5 = 63$ documents) ?

2. Can we efficiently detect the sub-topics (i.e., the two major functions) by clustering this document pool ?

3. How does it relate to our gene index where document vectors are merged per gene and subsequently normalized ?

Not completely surprising, when profiling the document pool, i.e., computing highest mean and variance terms, a first reference to autophagy is only found at the 26th position in the term ranked by variance. Nearly all other top scoring terms refer to elements of the secretory pathway SEC23 participates in. We proceeded with two subclustering iterations on the documents to see whether the 'autophagy' theme could be ultimately extracted from this document set, as well as to understand how the document clustering works.

In Table 6.7 we see that the overall profile of the 63 documents clearly refers to the secretory pathway. A first clustering iteration divides the documents in two groups. The one is more related to the biological processes *ER to Golgi transport* and *autophagy*, whereas the other groups articles refering to the cellular component *COPII vesicle coat*. By further subclustering the first group we finally retrieve autophagy-related terms in the corresponding article group. However, the persistent occurrence of terms as 'golgi transport' and 'secretory pathway' needs further clarification. We therefore examine some documents belonging to subcluster 1.1. by printing their title and key sentences extracted from the abstract in Table 6.8. Besides articles typically treating autophagy (such as PMID:11707261), two articles covering the connection between autophagy and secretion were found (PMID:11694599 and PMID:10491387). This explains the difficulty to isolate autophagy-related documents in this collection.

Comparing the overall profile of these 63 documents (Table 6.7.A) to the average profile of AUT10 and SEC23 extracted from the gene index (Table 6.7.B), we conclude that indexing textual information in a gene-centric way by using literature references constitutes an efficient way to extract the underlying themes. By normalizing each gene profile we correct for an unbalanced number of references that resulted in an overrepresentation of terms related to the secretory pathway.

Quality of associated abstracts

Our method is dependent on the functional information contained in the linked abstracts. In Table 6.9 we give some examples of spurious abstracts that occasionally troubles the analysis. We mainly distinguish articles that report on system dynamics, the performance of biotechniques, genetic characteristics of a sequence and homologues or mutations of a gene. As these articles sometimes contain function-related terms they constitute a source of noise.

Table 6.7: A. Term profiles resulting from subsequent clustering of the document profiles (instead of gene profiles) related to SEC23 and AUT10. The overall profile of the 63 documents clearly refers to the secretory pathway. A first clustering iteration divides the documents in two groups. The one is more related to the biological processes *ER to Golgi transport* and *autophagy*, whereas the other groups articles refering to the cellular component *COPII vesicle coat*. By further subclustering the first group we finally retrieve autophagy-related terms in the corresponding article group (in **bold**). **B.** Average gene profile from SEC23 and AUT10 derived from the gene index, from which the two major themes can be immediately derived. Organizing gene information in a gene-centric way aids in uncovering relevant themes despite unbalanced and biased literature references.

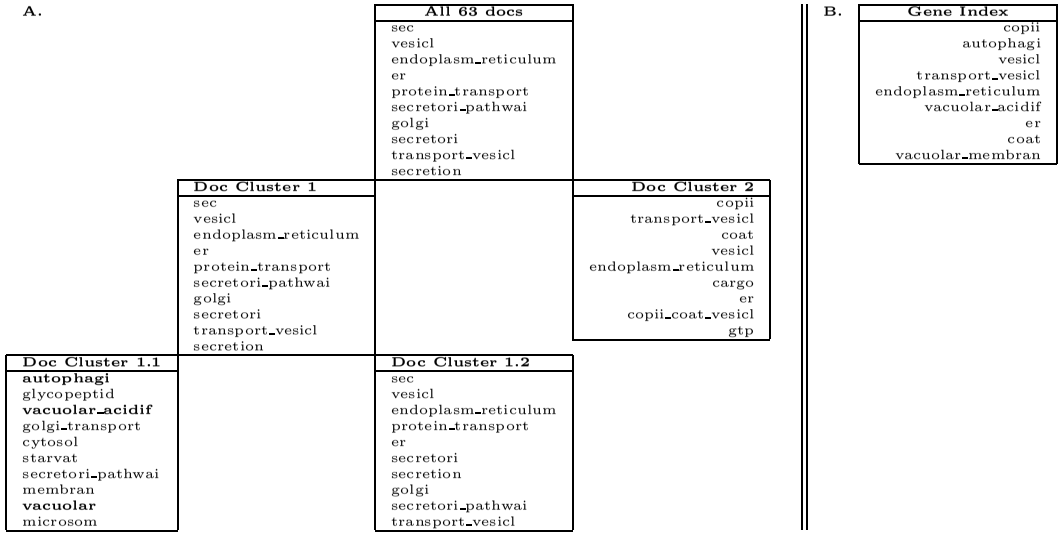


Table 6.8: Title and informative sentences from MEDLINE abstracts in subcluster 1.1 that connect autophagy with the secretory pathway.

PMID	Title	Informative sentence
11694599	Autophagosome requires specific early Sec proteins for its formation and NSF/SNARE for vacuolar fusion	To further understand autophagy, we analyzed the involvement of factors engaged in the secretory pathway
10491387	Starvation induces vacuolar targeting and degradation of the tryptophan permease in yeast	internal TAT2 is diverted from the late secretory pathway to the vacuolar pathway
11707261	Autophagy and the cytoplasm to vacuole targeting pathway both require Aut10p	

Table 6.9: Titles of typical MEDLINE abstracts that troubles term-based functional text analysis. Sometimes database gene annotations contain references that don't pertain to the function of a gene, but to another aspect. Examples are articles that treat homologues of a gene in another organism, sequence characteristics, biotechniques or other aspects. As such abstracts are not directly related to the function of a gene, they constitute a noise factor.

PMID	Title
11208060	Morphogenesis and dynamics of the yeast Golgi apparatus
10051619	Yeast and human genes that affect the Escherichia coli SOS response
9046099	The sequence of a nearly unclonable 22.8 kb segment on the left arm chromosome VII from *Saccharomyces cerevisiae* reveals ARO2, RPL9A, TIP1, MRF1 genes and six new open reading frames
10383756	Genetic redundancy and gene fusion in the genome of the Baker's yeast *Saccharomyces cerevisiae*: functional characterization of a three-member gene family involved in the thiamine biosynthetic pathway
9348540	Nup93, a vertebrate homologue of yeast Nic96p, forms a complex with a novel 205-kDa protein and is required for correct nuclear pore assembly

6.5.4 Condition-specific yeast networks

With TXTGate we primarily aimed at developing a methodology that could support gene expression analysis. As gene expression makes only sense if the corresponding cellular conditions are well-characterized, we test whether our method is capable of discriminating between different states of a given pathway. To this end, we use the analysis performed in Ihmels et al. [95], where the authors use their signature method to discover expression 'modules'[5]. They identified two subparts of the yeast TCA cycle that are autonomously regulated in different cellular contexts (see Figure 6.6).

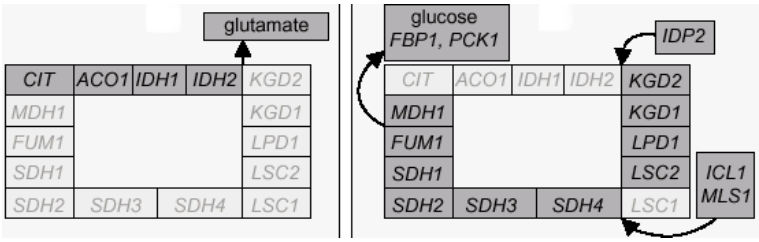


Figure 6.6: Two modules from the TCA pathway found by Ihmels *et al.*

[5]A module is defined here as a set of genes that are jointly active under very specific conditions.

We profiled these two subsets and examined how TXTGate differentiates between the two subnetworks by examining both term- and gene-centric summaries. The first module constitutes four genes that Ihmels and collegues reported to be 'upstream of alfa-ketoglutarate, a primary precursor of glutamate and coregulated under experimental conditions of deletion of RTG1 and genes that affect mitochondrial function'. In Table 6.10 we print the term-centric TXTGate profile of the first transcriptional module. 'Glutamate' is a module-specific term whereas 'isocitrate dehydrogenase' and 'TCA cycle' are also ranked highly when profiling the entire pathway (see Figure 6.7). Using these three terms to query PUBMED we are able to retrieve three highly specific abstracts (see Table 6.10) related to reduction in respiration capacity (or hypoxia). We note that the reference PMID:10490611 is exactly the one used by Ihmels *et al.* to support their results.

Under a different set of conditions the authors found a module containing 'the genes upstream of alfa-ketoglutarate coexpressed with genes whose expression is dependent on the transscriptional activator CAT8, which suggests they are involved in gluconeogenesis'. Profiles for the second transcriptional module are displayed in Table 6.11. Again, although terms as 'carbon' and 'succinate dehydrogenase' are shared with profiles for the more general TCA pathway, they rank particularly high in this cluster and, used as building blocks for a PUBMED query in combination with 'glyoxylate cycle' or 'gluconeogenesis' , we are again able to retrieve very specific MEDLINE abstracts (Table 6.11). Interestingly, the abstract PMID:10328823 contains references to gene ACN9, which also scores high in our gene-centric profile (cfr. infra) and is new with respect to the analysis by Ihmels (see also below). The abstract PMID:11514507 refers to the role of CAT8 in this module and confirms the manual analysis of Ihmels and colleagues. This means our system is a powerful alternative to extensive manual analyses in PUBMED as through set of keywords it can quickly lead to relevant information.

For the sake of completeness Figure 6.7 shows a screenshot of TXTGate's term profile of *all* genes belonging to the two modules (i.e., when ignoring the modular structure of the TCA pathway inferred from expression analysis), succeeded by a text-clustering (k=3). Not surprisingly, coherent groups are formed based on strong functional similarity (succinyl-CoA metabolism, gluconeogenesis, succinate dehydrogenase), rather than on membership to a pathway module. Extraction of full pathways solely from text requires NLP-oriented methods and is outside the scope of this application.

In this section we have shown that our method does not merely recover general functional terms shared by a given group genes (as treated in previous sections), but can also assist in the interpretation gene groupings stemming from advanced clustering methods. Secondly, we demonstrated how the TXTGate methodology supports the cyclic nature of the analysis process where literature-extracted terms are not only used to comprehensibly summarize gene groups, but also for the construction of subsequent queries to the PUBMED interface. This approach can considerable speed up the analysis by facilitating retrieval of highly relevant information. Moreover, we already report the discovery of a

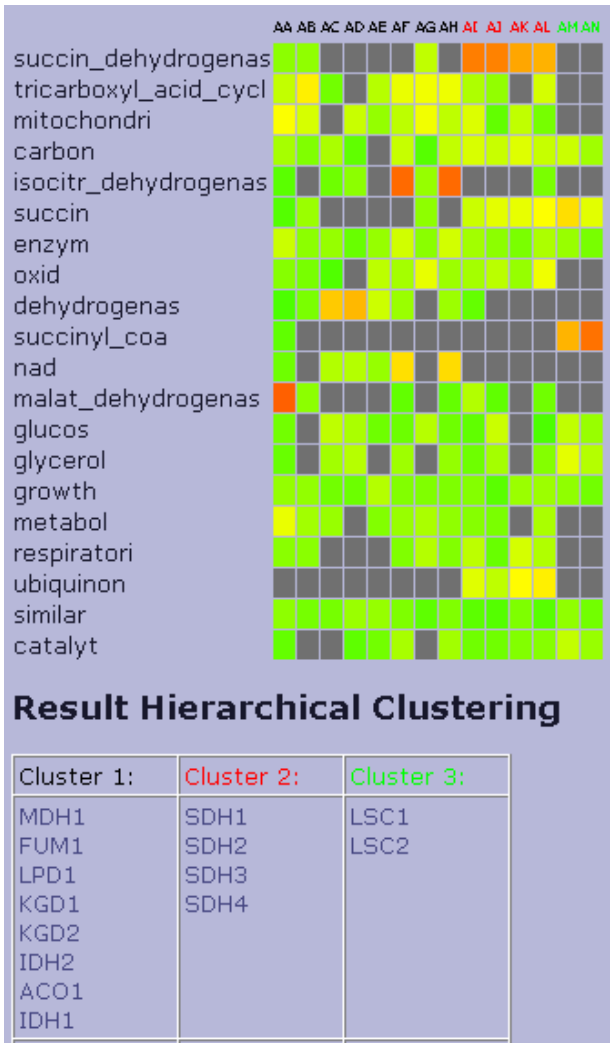


Figure 6.7: Term profiles and subclustering of all genes from the TCA pathway in Ihmels *et al.*

new putative gene through text analysis (ACN9), which will be further covered in next section.

6.5.5 Co-linkage analysis

As genes take a central position in molecular biology they can be regarded as one of the building blocks in biological reasoning. Contrary to regular terms, which, in our context, confer either a general or specific functional connotation, genes (or, more specifically, gene names) convey information in a more 'compact' way. Indeed, they often function as primary keys between many public genome databases. We start from the hypothesis that genes are often cited by their official symbol and look for simultaneous presence in the *pool* of abstracts that are linked to the group of genes under consideration. As gene interactions networks are known to be scale-free[6] [91] and gene citation is sensitive to scientific *trends* [90], we can expect a handful of genes to dominantly occur in these kinds of analyses. We interpret such hub-genes as less informative and therefore correct partly for their invasiveness by applying, as before, the IDF weighting scheme to the gene-centric index.

In Figure 6.9 we show the co-linkage network for module 2 from Ihmels *et al.*. The query is composed of the module's 15 genes and results in 212 nodes and 457 edges. The idiosyncracy of the gene network (i.e., lots of genes with few connections and few genes with lots of connections) troubles a transparent analysis. Our method proposes candidate genes of interest by assigning a high score to genes that are co-linked multiple times with the genes from module 2. In Table 6.12 we list the top 30 genes ranked by our system. Three genes from the original list are not recovered because of their small weights.

Contrary to a 'whole network' analysis, our approach comprehensibly summarizes, in a first approximation, the most prevalent gene links from literature. For example, CAT8 is a transscriptional regulator suggesting the implication of the member genes of module 2 in gluconeogenesis. This is also indicated by Ihmels and colleagues as well as our term-based MEDLINE analysis. ACN9 is a weakly understood gene (3 curated abstracts) involved in gluconeogenesis that also appeared in the course of our term-based analysis and constitutes a good candidate gene for further analysis. Indeed, when submitting ACN9 jointly with the genes from module 2 to SGD's *Expression Connection*[7], we find expression correlation (Pearson correlation ≥ 0.8) of this gene with KGD1 and PCK1 in an experiment describing the evolution of expression during glucose limitation (see Figure 6.8 and [76]). We note that the accompanying paper was not present as a literature reference for ACN9 nor was this gene addressed in this paper according to SGD curators.

Finally, we see some genes from module 1 (marked with an asterisk) present

[6] A scale-free network is a specific kind of network in which the distribution of connectivity is extremely uneven. Instead, in scale-free networks, some nodes act as intensively connected hubs using a power-law distribution, $\text{Prob}(\text{nr\_connections} > x) = x^{-\gamma}$, with $\gamma \geq 0$.

[7] Expression Connection allows for searching gene expression data from multiple yeast microarray studies.

Table 6.10: Term-centric TXTGate profile of the first transcriptional module from Ihmels *et al.*. 'Glutamate' is a module-specific term whereas 'isocitrate dehydrogenase' and 'TCA cycle' are also ranked highly when profiling the entire pathway. However, using these three terms (plus 'Saccharomyces cerevisiae') to query PUBMED we are able to retrieve highly specific abstracts related to reduction in respiration capacity (or hypoxia). We note that the first reference we retrieve is the one used by Ihmels *et al.* to explain their results. This supports the cyclic nature of the analysis process where literature-extracted terms are used to comprehensibly summarize gene groups and to subsequently conduct a further exploration.

Gene symbol	Term-centric profile	Weights	Retrieved PMID	MEDLINE title
IDH2	**isocitr_dehydrogenas**	0,4503	10490611	A transcriptional switch in the expression of yeast tricarboxylic acid cycle genes in response to a reduction or loss of respiratory function.
CIT	**tricarboxyl_acid_cycl**	0,2201		
ACO1	nad	0,2140		
IDH1	isocitr	0,1857		
	glutam	0,1786	8672488	Expression and gene disruption analysis of the isocitrate dehydrogenase family in yeast.
	mitochondri	0,1687		
	enzym	0,1384		Isocitrate dehydrogenases and oxoglutarate dehydrogenase activities of baker's yeast grown in a variety of hypoxic conditions.
	oxid	0,1355	1091851	
	carbon	0,1088		
	retrograd	0,0986		
	metabol	0,0949		
	amp	0,0905		
	acet	0,0889		
	respiratori	0,0861		
	increas	0,0768		
	oxid_phosphoryl	0,0761		
	glucos	0,0734		
	citrat	0,0730		
	catalyt	0,0710		
	elev	0,0696		

Table 6.11: Term-centric TXTGate profile of the second transcriptional module from Ihmels *et al.*. Again, although the terms 'carbon' and 'succinate dehydrogenase' are shared with profiles from more the general TCA pathway, they rank particularly high in this cluster and when used in combination with 'glyoxylate cycle' or 'gluconeogenesis' to query PUBMED, we are able to retrieve very specific MEDLINE abstracts. Interestingly, the first abstract contains references to a gene whose symbol scores high in our gene-centric profile.

Gene symbol	Term-centric profile	Weights	Retrieved PMID	MEDLINE title
MDH1	**carbon**	0,1499		
FUM1	**succin_dehydrogenas**	0,1458	10328823	Yeast mutants of glucose metabolism with defects in the coordinate regulation of carbon assimilation
SDH1	glucos	0,1074		
SDH2	enzym	0,0943		Pyruvate carboxylase deficiency in yeast: a mutant
SDH3	mitochondri	0,0902	3882052	affecting the interaction between the glyoxylate and
SDH4	succin	0,0873		Krebs cycles
LSC2	growth	0,0866		
LPD1	metabol	0,0830		Three target genes for the transcriptional activator
KGD1	**glyoxyl_cycl**	0,0814	11514507	Cat8p of Kluyveromyces lactis: acetyl coenzyme A
KGD2	gluconeogenesi	0,0805		synthetase genes KlACS1 and KlACS2 and lactate
FBP1	dehydrogenas	0,0801		permease gene KlJEN1.
PCK1	tricarboxyl_acid_cycl	0,0796		
IDP2	oxid	0,0749		
ICL1	malat_dehydrogenas	0,0707		
MLS1	catabolit	0,0675		
	cytosol	0,0661		
	ethanol	0,0659		
	isocitr_lyase	0,0654		
	similar	0,0625		
	isocitr_dehydrogenas	0,0618		

in profiles in Figure 6.12, indicating a difficulty to retrieve context-specific genes from a module. There are two possible reasons for this: either our term-inspired method to detect subnetworks is simple and leaves room for improvement [131][97] or the analyzed conditioned gene relations are weakly, or not, reported in literature.

As a conclusion we state that our ranking of co-linked genes via a gene-centric vocabulary can offer useful clues to interpret and further unravel sub-networks from expression clusters. Moreover, the gene-centric approach is found to be complementary to the term-based analysis as terms refer more to candidate functional characteristics, while co-linkage analysis infer hypothetical relations to the investigated gene group. The putative relation of ACN9 to module 2 is also retrieved in a particular expression experiment. This supports our cyclic views on the analysis process (see also Figure 1.6), where modules or clusters from the data world are analyzed in the text world and subsequently yield candidates that are can be confirmed again in the data world.

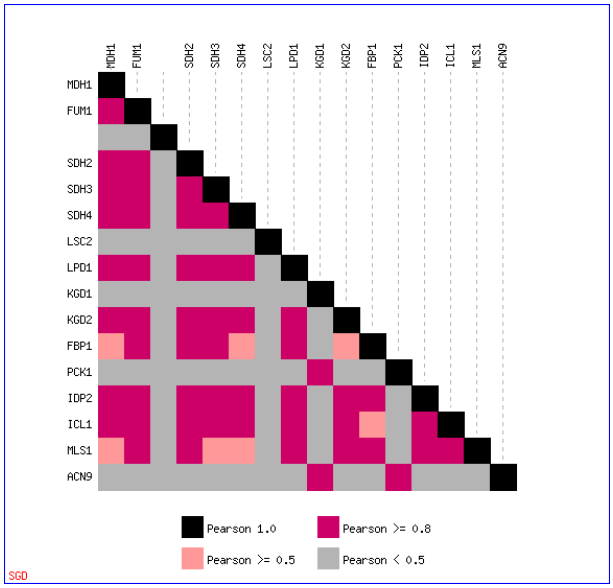


Figure 6.8: Visualization from SGD's *Expression Connection* of expression correlation between member genes of module 2 and ACN9 in a microarray experiment studying the effect of glucose limitation. ACN9 was independently found by term- and gene-centric analyses with TXTGate and shows here a high expression correlation with KGD1 and PCK1, making it an additional candidate to this module (see also Figure 6.6). This supports our cyclic views on the analysis process as implemented in TXTGate, where modules or clusters from the data world are analyzed in the text world and subsequently yield candidate genes that can be further scrutinized in the data world.

Table 6.12: Gene-centric profile of module 2 from Ihmels *et al.*. We show the top 30 'terms' and append their biological function according to GO. Genes in **bold** belong to the original query. Gene names marked with an asterisk (*) are member genes of module 1 of the TCA pathway (see Figure 6.6). We marked ACN9 and CAT8 in *italic*, as they also appear in our term-based analysis presented earlier. We demonstrate here that our ranking of co-linked genes via a gene-centric vocabulary can offer primary clues to interpret and further unravel sub-networks from expression clusters. Contrary to a whole network analysis, our approach comprehensibly summarizes, in a first approximation, the most important gene links from literature.

Gene 'term'	Weight	GO annotations (biological process)
SDH4	0,1235	
SDH2	0,1140	
MDH1	0,0947	
IKI1	0,0924	regulation of transcription from Pol II promoter
ACO1(*)	0,0921	
ACN9	0,0905	carbon utilization by utilization of organic compounds
		gluconeogenesis
ICL1	0,0884	
MDH2	0,0867	gluconeogenesis
		glyoxylate cycle
		malate metabolism
MLS1	0,0849	
TCM62	0,0847	protein complex assembly
IDH2(*)	0,0834	
SOR1	0,0822	fructose metabolism
		mannose metabolism
IDH1(*)	0,0775	
FBP1	0,0774	
SDH3	0,0739	
CIT1(*)	0,0725	
PCK1	0,0712	
RTG2	0,0705	intracellular signaling cascade
ELP6	0,0606	regulation of transcription from Pol II promoter
SIP4	0,0597	positive regulation of gluconeogenesis
		regulation of transcription from Pol II promoter
CAT8	0,0594	gluconeogenesis
		positive regulation of transcription from Pol II promoter
KGD2	0,0593	
HSP60	0,0592	mitochondrial matrix protein import
		protein folding
SKO1	0,0589	negative regulation of transcription from Pol II promoter
SFC1	0,0584	fumarate transport
		succinate transport
ACS1	0,0584	acetate fermentation
		acetyl-CoA biosynthesis
KGD1	0,0569	
IDP2	0,0537	
LPD1	0,0514	

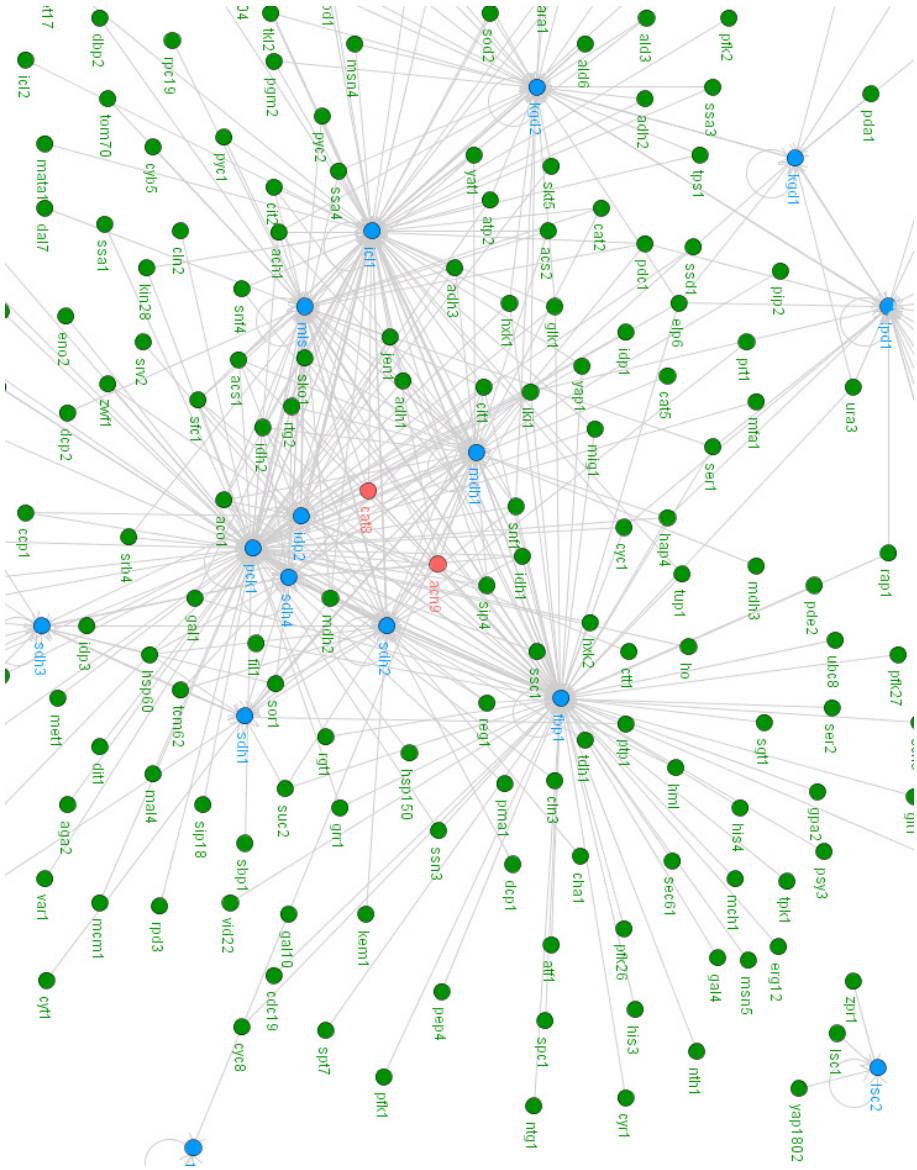

kgd2
kgd1
icl1
mls1
mdh1
pck1
idp2
sdh4
sdh2
sdh1
fbp1
lpd1
lsc2
cat8
acn9
sod2
ald3
ald6
pfk2
ssa3
adh2
tps1
pda1
yat1
atp2
slt5
cat2
pdc1
ssd1
pip2
adh3
hxk1
glk1
elp6
idp1
prt1
ura3
jen1
adh1
cit1
iki1
yap1
cat5
mfa1
mig1
ser1
snf1
idh1
cyc1
hap4
tup1
mdh3
rap1
pde2
sip4
hxk2
ho
ubc8
pfk27
ssc1
ctt1
ser2
sgt1
reg1
ptp1
tdh1
hsp150
hml
his4
gpa2
psy3
pma1
cln3
cha1
sec61
mch1
tpk1
ssn3
dcp1
erg12
gal4
msn5
pfk26
his3
nth1
zpr1
lsc1
yap1802
cyr1
ntg1
spc1
att1
pep4
pfk1
spt7
kem1
gal10
cdc19
cyc8
vid22
rpd3
mcm1
var1
cyt1
aga2
mal4
sip18
sbp1
suc2
grr1
rgt1
sor1
fil1
tcm62
hsp60
met1
dit1
idp3
gal1
sdh3
srb4
ccp1
aco1
dcp2
zwf1
sfc1
eno2
snf4
acs1
idh2
rtg2
sko1
kin28
srv2
ssa1
dal7
mata1
cyb5
tom70
cln2
cit2
ach1
pyc1
pyc2
ssa4
msn4
tkl2
pgm2
rpc19
dbp2
icl2
tf17
ara1

Figure 6.9: Our co-linkage network for module 2 discovered by Ihmels *et al.*. The query includes the 15 genes of the module and are indicated in blue to show their place in the network. In total there were 212 nodes and 457 edges. As the connectivity tends to grow with the number of genes of interest, such networks quickly become intractable for manual analysis. Our text mining method is a first realization to prioritize on and to identify potentially interesting genes in a large network. Relevant genes derived from literature analysis are indicated in red. We also observe that few genes have many links, whereas many genes are sparsely connected. Visualization was done with the BioLayout software [74].

6.6 Study 2: Human data

Advances in genomics have pushed approaches to medicine increasingly to the molecular and genetic level. In this section we discuss cases where human genomic data is (or should be) related with disease information. Along the line of the yeast studies in previous sections we show in a concise way the consistency and added value of TXTGate with respect to previous work in text profiling (Section 6.6.1) and how the other vocabularies that are based on OMIM, eVOC and HUGO make up powerful tools to discover disease-based information (Section 6.6.2).

6.6.1 Subclustering human data

We compare the output from TXTGate with results presented in Chaussabel *et al.* [60], where the authors describe, among other experiments, text profiling and clustering of nearly 200 genes involved in the common transcriptional program induced in human macrophages upon bacterial infection (original data describe din Nau *et al.* [118]).

We re-interpret the results by retrieving the MEDLINE textual profiles of all genes in the clusters and compared TXTGate's best scoring terms to the cluster terms in Chaussabel *et al.* The results of the first four (non-overlapping) biclusters (cluster *a*, *b*, *c* and *d*) can be found in Table B.4. The terms 'Adipose', 'Metastasis' and 'NM' did not show up in the profiles from TXTGate because they are not contained in the GO domain vocabulary. For cluster *e* there are no common terms found. All other cluster terms from Chaussabel and colleagues have a comparable equivalent in the TXTGate profiles.

When running TXTGate using the OMIM vocabulary, however, we are able to uncover exactly those terms associated to diseases that are retrieved by Chaussabel *et al.* by manually investigating genes from this cluster in the OMIM database. In Table 6.13 we highlight these terms in bold. As the set of diseases related to these genes is heterogeneous, the relevant terms display a high variance rather than a high mean, which supports our choice to include also a variance profile. Moreover, the fact that we retrieve those disease terms

only by means of the OMIM vocabulary points out that the use of various vocabularies in TXTGate leads to improved insights, a point treated further in the next section.

6.6.2 Textual information through the eyes of different vocabularies

One major feature of TXTGate is its ability to present textual information (most importantly MEDLINE abstracts) from different perspectives. This is implemented by offering indices built on GO-, OMIM-, MeSH-, eVOC-, and gene nomenclature-based domain vocabularies respectively. Each configuration is meant to expose a different aspect of the literature. TXTGate mirrors the dual approach adopted by external databases to separate keyword and gene symbol queries. This subscribes to our strategy to construct both term- and gene-centric vocabularies.

In order to compare the various term-based vocabularies we profile a group of genes involved in colon and colorectal cancer extracted from the OMIM Morbid Map database. Table 6.14 shows the top ten terms for each of the retrieved profiles. As can be seen, there is little difference between the MeSH and OMIM profiles whose terms are mainly medical- and disease-related ('colorectal cancer', 'colon cancer', 'colorectal neoplasm', 'hereditary'), while the scope of the GO profile is more focused upon metabolic functions of genes ('mismatch_repair', 'DNA repair', 'tumor suppressor', 'kinase') and the eVOC profile contains terms more related to cell type and development ('growth', 'cell', 'carcinoma', 'metabolic', 'fibroblast'). TXTGate's link-out feature allows a more profound analysis of the retrieved terms. Top-ranking terms can be sent to PUBMED to retrieve relevant publications as demonstrated in the yeast studies earlier. Since all MEDLINE entries are tagged with MeSH keywords, using terms from the MeSH vocabulary assures a successful query. When using the GO-derived vocabulary, terms can be mapped back directly to the GO tree with AmiGO [2] to investigate the term's neighborhood. Other databases made accessible in the application for querying include LocusLink and OMIM.

We use the same colon cancer case to test the potential of our gene-based HUGO vocabulary in screening for co-linkage of genes. We construct two different index tables: one with and one without taking alternative gene symbols into account, by mapping all synonymous symbols to the primary gene symbol. The former has the disadvantage of not being able to disambiguate alternate gene symbols that mapped to different primary gene symbols. The latter does not take into account synonyms since only true occurrences of a symbol are counted. As a consequence, often-used symbols are ranked highly, while not being the official gene symbols. Examples of this are 'p21' and 'DRA' whose primary symbols are respectively 'CDKN1A' and 'SLC26A3'. The top 25 gene symbols using the latter index table are given in Table 6.15. Most of the retrieved gene names are also in the query list.

We used TXTGate's link-out feature to investigate the role of the genes that were not in the query list by sending them as a query to LocusLink

Table 6.13: Comparison of the terms in cluster *e* found by Chaussabel *et al.* with those found by TXTGate (using OMIM vocabulary setting). The diversity of the diseases the member genes are related to, makes the relevant terms displaying high variance, rather than high mean. When running TXTGate using the OMIM vocabulary we are able to uncover exactly those terms associated to diseases that are retrieved by Chaussabel *et al.* by manually investigating genes from this cluster in the OMIM database. Our system hence provides automatic support for time-consuming expert investigations into the prior art of a field. The flexibility of doing so with various domain vocabularies helps in achieving this efficiently.

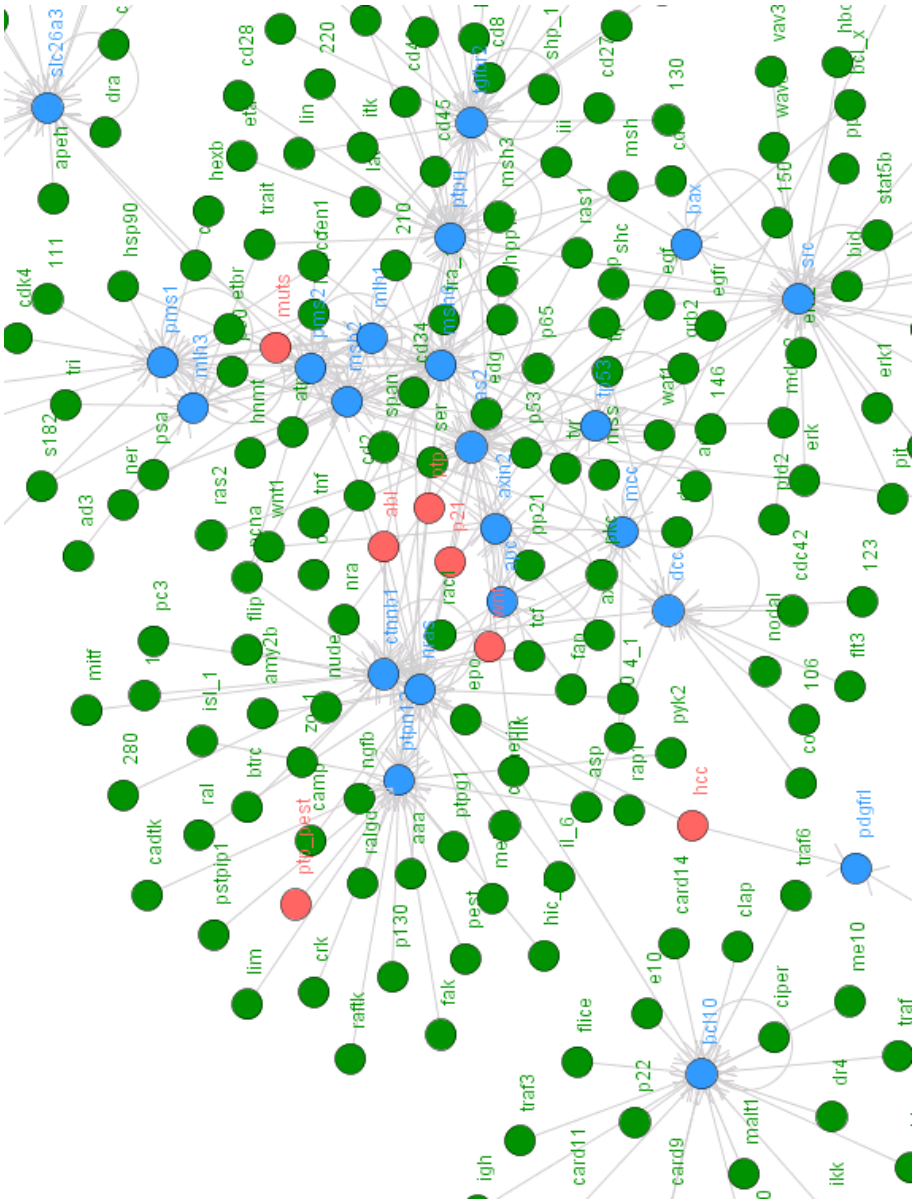
Gene Symbol	**Cluster Terms in Chaussabel *et al.***	**TXTGate terms**
Cluster e		
CKB	Population	deaminas
AMPD3	Frequency	**lipoprotein_lipas**
ADA	Allele	creatin
ADORA2A	Unrelated	lipoprotein
ADORA3	Families	**krabb**
P2RX	Recessive	**epidermolysi_bullosa**
P2RX1	Autosomal	**alagil**
P2RX7	Disorder	bear
GEM	Severe	leukodystrophi
ARHH	Patient	receptor
LPL	Deficiency	down
CD36L1		**corneal_dystrophi**
LDLR		**deaf**
BF		hdl
GALC		nucleosid
LAMB3		retinoblastoma
GJB2		junction
TGFBI		adhesion
JAG1		congenit_heart_defect
DSCR1		**hear_loss**

Table 6.14: Various perspectives on textual information in TXTGate. Here we show how term-centric GO-, OMIM-, MeSH- and eVOC-based vocabularies profile a group of genes involved in colon and colorectal cancer.

GO	OMIM	MeSH	eVOC
mismatch_repair	colorect	colorect_neoplasm	colorect
tumor	colorect_cancer	mismatch	tumour
dna_repair	tumor	cancer	malign_tumour
mismatch	kinas	colorect	colon
pair	colon	mutat	growth
tumor_suppressor	hereditari	repair	cell
apc	cancer	dna_repair	carcinoma
kinas	colon_cancer	colon	metabol
somat	associ	neoplasm_protein	fibroblast
ra	on	tumor	chain

and GeneCards. We found 'mutS' to be an *Escherichia coli* mismatch repair gene that has several homologues in human ('MSH2', 'MSH3', 'MSH4', 'MSH5', 'MSH6'). 'p21' is a cyclin-dependent kinase inhibitor that plays a role in cell cycle regulation and whose expression is governed by p53. The term 'wnt' denotes an entire family of genes active in the WNT pathway which encode secreted signaling proteins involved in oncogenesis and in several developmental processes, including regulation of cell fate. We did not find an associated locus for gene symbol 'hcc' in LocusLink, but a plain text search in the abstracts revealed that HCC is used as abbreviation of 'hepatocellular carcinoma'. Gene 'abl' is a known proto-oncogene encoding a cytoplasmic and nuclear protein tyrosine kinase implicated in processes of cell differentiation, cell division, cell adhesion and stress response. 'ptp' is the abbreviation of 'protein tyrosine phosphatase'. PTPs are signaling molecules regulating cell growth and differentiation, the mitotic cell cycle and oncogenic transformation. 'ptp_pest' denotes those PTPs that contain a C-terminal PEST-motif. The protein encoded by 'prlts' has a high sequence similarity with the ligand binding domain of PDGF receptor beta. Mutations in this gene can cause colorectal cancer. This concludes the colon and colorectal cancer case where we illustrated the use of term- and gene-centric domain vocabularies in combination with the link-out feature.

Figure 6.10 (next page) : Constructed co-linkage gene network for the colon cancer genes we extracted via OMIM. Analogously to Figure 6.9, genes present in the query are indicated in blue, whereas relevant genes derived from literature analysis are indicated in red in the network. Table 6.15 shows the description of these co-linked genes. Visualization was done with the BioLayout software [74].


slc26a3
dra
apeh
cdk4
111
hsp90
tri
s182
ad3
ner
psa
pms1
mlh3
hnmt
ras2
pcna
wnt1
tnf
abl
p21
pp21
nra
rac1
pc3
mitf
flip
amy2b
nude
etnnb1
nck
epo
tcf
fap
isl_1
btrc
zo_1
camp
ngfb
ptpn12
280
ral
caditk
pstpip1
lim
ptk_dest
crk
raftk
ralgds
aaa
p130
fak
ptpg1
pest
hic_
il_6
asp
rap1
hcc
pyk2
d4_1
dcc
nodal
cdc42
123
fft3
106
traf6
pdgfrl
clap
card14
e10
ciper
me10
bcl10
p22
flice
traf3
igh
card11
card9
malt1
ikk
dr4
traf
p65
p53
ras1
shc
egf
egfr
bax
msh3
msh2
mlh1
cd34
cd45
ptpu
cd28
cd27
cd8
cd4
shp_1
iii
msh
130
150
vav3
src
bid
stat5b
erk1
pit
pld2
erk
146
wart
mcc
apc
axin2
lin
itk
trait
hexb
cdfen1
210
220
muts
atr
span
ser
edg
tp53
tyr
a
dcc
ptpu
hcc

6.7 Discussion

In this chapter we have presented a framework for advanced textual profiling of groups of genes. Our primary aim was to create a text mining platform to manage and exploit information about groups of genes. Having provided a proof of principle on the vector space model in Chapter 5, our focus with TXTGate in this chapter has been more on the practical usability, consistency and, especially, the added value of our approach with respect to earlier systems: TXTGate is implemented as a web application designed to process queries of up to 200 genes, although this is not a strict limitation of the system. We believe the application scales well enough to be of use in, for example, microarray cluster validation. Supported by work of Stephens et al. [142] and more recently Chiang et al. [61] we aimed to complement the limitations of a single, more general text index by offering different views. Nevertheless, some vocabularies could still be optimized to improve the information content of the profiles. For example, some general or non-informative terms are still scoring high because of our stemming and phrase detection methods (e.g., 'ii', 'protein', 'alpha').

Although the citations in LocusLink and SGD constitute good sources for retrieving relevant gene-related MEDLINE abstracts, a weighting of information according to the context and the elimination of poorly informative or contaminating annotations (such as sequence-related articles) still need to be taken into account in future incarnations of the software. Document classification strategies as in Leonard et al. [104] or Raychaudhuri et al. [130] can be adopted to this end.

Much in the same way as with GO annotations, transfer of literature references according to homology can be used to characterize poorly annotated genes [129][149]. At this stage, the application allows for the study of homologues within all organisms contained in LocusLink, provided the user inputs the corresponding LocusLink identifiers. This type of operations will be increasingly supported with future additions of literature indices from other organisms and databases. As a conclusion, its approach to summarize database annotations and literature via specific vocabularies along with its options to perform further analysis via clustering or query building, make TXTGate a flexible gateway to comprehensibly explore text-based information.

Table 6.15: TXTGate co-linkage analysis of genes with gene-centric vocabularies. This table shows the top-25 co-linked gene symbols in the pool of abstracts of the colon and colorectal cancer case. Genes that were not in the query list are indicated in **bold**.

Gene name	Description
HNPCC	Hereditary nonpolyposis colon cancer
APC	Adenomatous polyposis coli protein
P53	Cellular tumor antigen P53 (tumor suppressor P53)
MLH1	DNA mismatch repair protein MLH1 (mutL protein homolog 1)
MUTS	E. coli mismatch repair gene mutS
P21	Cyclin-dependent kinase inhibitor 1A
MSH2	DNA mismatch repair protein MSH2 (mutS protein homolog 2)
BAX	BAX protein, cytoplasmic isoform delta
WNT	Wingless-type MMTV integration site family members
PMS2	DNA mismatch repair protein PMS2
SRC	Proto-oncogene tyrosine protein kinase SRC
DCC	Tumor suppressor protein DCC precursor
MCC	Colorectal mutant cancer protein MCC
BRAF	Proto-oncogene serine/threonine protein kinase B-RAF
FGFR3	Fibroblast growth factor receptor 3 precursor
HCC	Hepatocellular carcinoma
DRA	Chloride anion exchanger DRA
AXIN2	AXIS inhibition protein 2
PMS1	DNA mismatch repair protein PMS1
ABL	Abelson murine leukemia viral oncogene homolog 1
BUB1	Mitotic checkpoint serine/threonine protein kinase BUB1
PTP	Protein tyrosine phosphatase family
BCL10	B cell lymphoma/leukemia 10
PTP_PEST	Protein tyrosine phosphatase family with C-terminal PEST-motif
PRLTS	PDGF-receptor beta-like tumor suppressor

Chapter 7

Integration of text and expression data

To date, clustering methods that include no background information as the ones presented in Chapter 3 remain the most popular. We have argued in detail in Chapter 5 why literature could leverage experimental data sets. Both the the coherence score from Chapter 5 and the TXTGate application presented in Chapter 6 approach literature as an aid to (a) interpret cluster structures discovered in expression data, or to (b) discover novel genes functionally related to a particular cluster. Yet this assumes the existence of a good clustering solution stemming from an appropriate algorithm applied to reliable data.

The complexity of gene expression data, however, leads to a whole range of potentially relevant clustering solutions. Also, fluctuations in the reliability of gene expression data undermine a fast extraction of co-regulated or functionally coherent gene groups. Although the use of literature representations *after* clustering can validate or partially improve cluster solutions, a *joint* analysis of expression and literature data can exert a further-reaching influence on the clustering outcome as both background and experimental information are processed in the same framework. In this chapter we explore various aspects of data integration (see also Figure 7.1 for its role in the dissertation) , including (a) ways to combine heterogeneous information, (b) the conundrum of independent and time-efficient validation, and (c) the effect of an integrated analysis on biological relevance. More specifically we investigate the combination and joint analysis of yeast expression data and literature-extracted information. We evaluate our setup independently in motif space by conducting a cis-regulatory motif analysis on the results supported by the M-score presented in Chapter 4[1].

[1]Our integrated approach to clustering genes through a combination of information from text and expression space and a validation of these results in motif space has been recently published in the Special Issue on Microarray Data Mining from the journal *SIGKDD Explorations* [83].

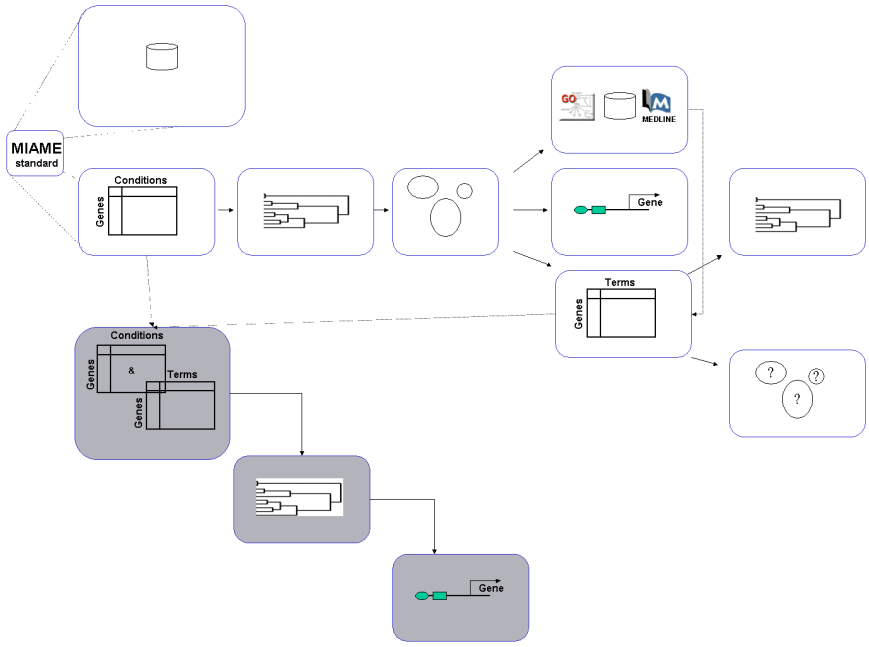


Figure 7.1: A combined analysis of data sources can yield more information than using them separately. We show how an integrated analysis of expression and text improves the clustering outcome in terms of significant motifs found in the resulting clusters using our M-score.

7.1 Gene expression, keywords and motifs

7.1.1 Overview of the approach

Typically, the current expert's environment is composed of a *data* world, which encompasses high-throughput data and statistical methods on the one hand, and a *knowledge* world, which contains existing domain information dominantly present in free-text form on the other hand. Within this terminology, data analysis is increasingly shifting towards a deep interaction of human expertise with those two worlds. To increase the efficiency of such interaction, we aim at overcoming the artificial separation of the two worlds (i.e., separation between tools for data analysis and those for information retrieval) by using domain literature in the same way as expression data, after a transformation of textual domain knowledge into a suitable numerical format.

In Figure 7.2 we give an overview of our approach: starting from a literature repository we compute a document index based on the vector space model which results in a document-term matrix. For each gene we summarize all documents that are linked to it (e.g., as query results from PUBMED, or entries in curated gene-literature repository) by merging the information as explained in Chapter 5. Having all genes represented in term vector space (indicated as (**a**) in Figure 7.2), we mathematically combine the text profiles with values in the gene expression matrix of a microarray experiment (indicated as (**b**) in Figure 7.2). This combination is performed either by pooling the feature vectors from expression and text space, or by combining the corresponding (possibly transformed) distance matrices. Further on, we will show that in some cases there exists an equivalence between these ways of combining data. Subsequently, we cluster the augmented data structure and validate our approach in motif space (indicated as (**c**) in Figure 7.2) by scoring and comparing various resulting solutions to the ones where solely expression data is used. In what follows we will introduce the case study and specify the information sources used.

7.1.2 Experimental setting

Our text-based information source consists of a literature index for yeast genes we constructed from a corpus of 24,909 yeast-related MEDLINE abstracts. These abstracts and their associations with the genes were extracted from the curated literature references in the Saccharomyces Genome Database as of 11 Jan 2003. We adopt the IDF text representation as before. Central to the theme of our work is the gene expression experiment. We use the yeast expression data from Cho *et al.* [62][141]. From the 3000 variance-normalized expression profiles, we withhold those 1745 that have literature references and therefore text profiles (See also Table 1.1 and Figure 1.10). Time points 10 and 11 of the data (a repeat of the G1 measurements) are omitted because of their unreliable nature. We further refer to Section 4.2.3 for how the data was processed. Finally, for the regulatory sequence analysis, we used the set of 42 cell cycle motifs listed in Table 4.4. Consensus sequences for these motifs were extracted from

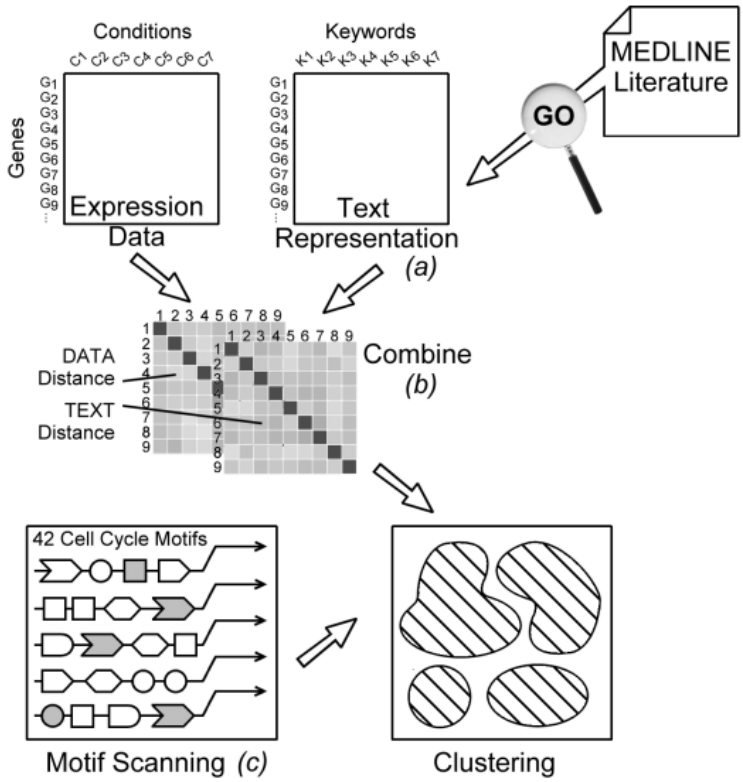


Figure 7.2: Overview of the Meta-clustering framework of expression data and textual information. After constructing textual gene profiles in term vector space (indicated as (**a**)), we mathematically combine them (indicated as (**b**)) with values in the gene expression matrix of a microarray experiment. This combination is performed either by pooling the feature vectors from expression and text space, or by combining the corresponding (possibly transformed) distance matrices. The augmented data is subsequently clustered and an evaluation in terms of cis-regulatory elements is conducted (indicated as (**c**)) (Figure courtesy: Steven Van Vooren).

the TRANSFAC Database [111], and string searches for each motif sequence over all 800bp upstream regions results in a gene-by-motif matrix containing raw counts (indicated (**c**) in Figure 7.2). We note that although much more advanced methods than string-based searches for motif detection are available (e.g., AlignACE [94], INCLUSive [64]), this choice still constitutes a reasonable approach to validate our setup (e.g., Bussemaker *et al.* [56], Gasch *et al.* [77]).

7.2 Mixing heterogeneous data via distance matrices

With multiple information sources simultaneously available, it is a challenging question how to conduct *integrated* exploratory analyses of microarray data with the aim of extracting more information than one would from the expression measurements alone. To prove this we will show how combining text-based information (essentially capturing functional relatedness) with expression data (registering co-expression) can add biological significance to the overall clustering analysis.

In this section we propose two ways of data combination (see (**b**) in Figure 7.2). Both belong to a class of integration methods sometimes referred to as 'intermediate' integration [121]. Figure 7.3 provides an illustration: whereas 'early' integration appends two (or more) types of data and passes it to a learning algorithm, 'intermediate' integration creates variable-to-variable dissimilarity matrices for each data type, combines them, and finally passes the result on to a learning algorithm. 'Late' integration involves a separate, rather sequential, analysis of the various data types [121]. Here we present and discuss two integration methodologies, whereas in Section 7.5 we elaborate on they contribute to an improved clustering of expression data.

7.2.1 Linear combination of distance matrices

With distance matrices D for both expression and literature data at hand (see previous Sections and 7.2), the most obvious solution to merge information, is to add the distance information for each pair of genes together in a new matrix

$$D^{Integr} = (1 - \lambda)D^{data} + \lambda D^{txt}, \tag{7.1}$$

with λ a parameter controlling the importance of the textual information. The rationale behind this procedure is that dissimilarity between two genes can be expressed after appropriate preprocessing steps and a choice of distance measures in each information space. When using the $1 - \cos$ distance measure in both spaces (or any type of covariation measure), the intermediate integration can be shown to be equivalent to combining the two original data matrices early on. Suppose that we have two data matrices A and B such that $\|B(i,:)\| = c \cdot \|A(i,:)\|$ for all i, with c a rescaling factor between the norms of

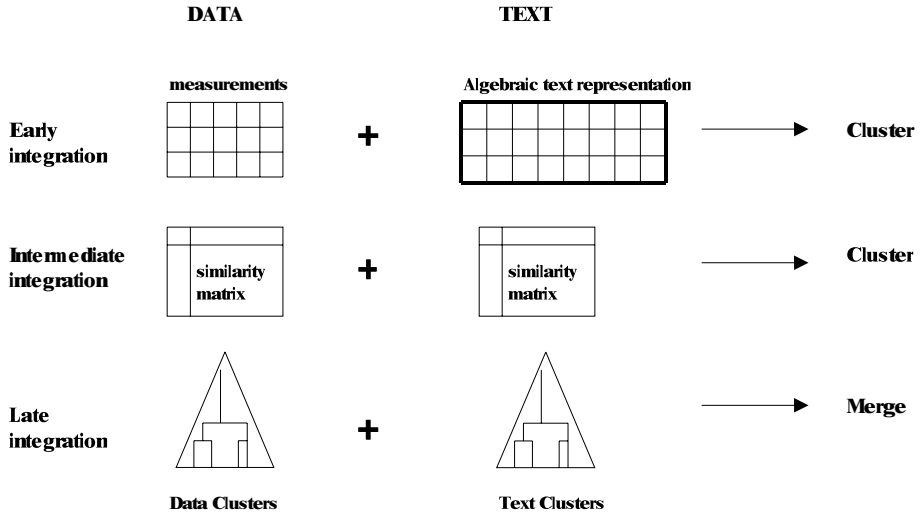


Figure 7.3: Various ways to integrate heterogeneous data. Whereas 'early' integration appends two (or more) types of data and passes it on to a learning algorithm, 'intermediate' integration creates variable-to-variable dissimilarity matrices for each data type, combines them, and finally passes the result on to a learning algorithm. 'Late' integration involves a separate, rather sequential, analysis of the various data types.

the two data types. For every two genes i, j we can then compute that

$$\cos([A(i,:)B(i,:)], [A(j,:)B(j,:)]) = \frac{1}{1+c^2}\{\cos(A(i,:), A(j,:)) + \cos(B(i,:), B(j,:))\}$$

where the square brackets denote a horizontal concatenation of the rowvectors and $c = \sqrt{\frac{\lambda}{1-\lambda}}$. This means that in case of $1 - \cos$ distance, early and intermediate integration are equivalent and we can see $\lambda \sim c$ as a scaling constant between the norms of the data- and text representations. The choice for λ is typically governed by the confidence attributed to either of the two data types. However, even when distance measures span the same range, some caution is required.

Indeed, when plotting the histograms of all mutual distances for the data and text representations in Figure 7.4, we observe a much sharper distribution towards one for the text-based distances. This means that, for example, $\lambda = 0.5$ would not correspond to an equal contribution of both sources in the mixed representation. Rather, this setting will favor the expression data over the text representation. Although not necessarily detrimental, this scaling issue invokes problems on the transparency or interpretability of λ as most choices will appear *ad hoc* given this observation.

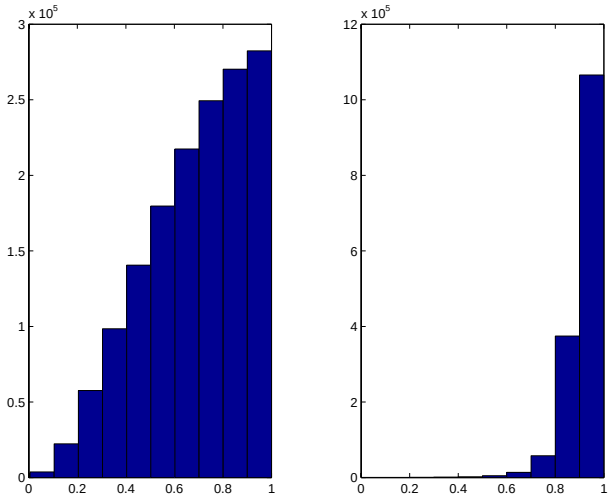


Figure 7.4: Histograms of mutual gene distances in expression space (left) and text space (right). Even when distance measures span the same range, we observe a much sharper distribution towards 1 for the text-based distances. The underlying reason is the inherent difference between the ways information is encoded in expression and text space. As a result $\lambda = 0.5$, for example, does not correspond to an equal contribution of both sources in the mixed representation. This anomaly forms our key argument to use p-valued distances for data integration.

7.2.2 Fisher's omnibus to combine evidence

To overcome the scaling problems introduced in the previous section, we transform all entries in distance matrices D to p-values by computing one-sided cumulative distribution function (cdf) values for each distance value in both representations. Once p-values are available, we are freed from the distribution of the data they are generated from and we can apply tools (called omnibus procedures) from meta-analysis, which encompasses a set of classical statistical techniques to combine evidence from multiple sources [117]. We use Fisher's omnibus method to combine the p-values derived for each expression-based and text-based distance. In this method the combined statistic

$$\mathcal{S} = -2 \log p^{data} - 2 \log p^{txt},$$

follows a χ^2 distribution [88] and we use the resulting cdf-values to make up the combined distance matrix. We note that this method can be generalized by including weights analogous to λ, but we will not further explore this option.

7.3 Motif scoring as independent evaluation

We use the M-score introduced in Chapter 4 to check how integrating text-based information with microarray data can reveal gene groupings with overall motif enrichments that are not detectable by expression data alone. As we investigate numerous ways and parameterizations to combine and cluster the data, the proposed M-score proves a useful tool in detecting promising directions. To isolate the effect of adding text, we construct a controlled experiment eliminating as much as possible variation due to differences in initializations, parameter settings or methodological choices. As partitioning method we have therefore chosen hierarchical clustering using Ward's method (see Section 3.3.3) (a) because it takes dissimilarity matrices as input, (b) for its deterministic nature and (c) for the computational advantage to use the same solution when considering multiple number of clusters through the cut-off value k. In both text- and data spaces the $1 - \cos$ distance measure is used. As the Euclidean distance metric is a frequently used alternative in case of expression data, we will also provide results for this case at the end of Section 7.5.

In Figure 7.5, we depict a scatter plot describing the difference in M-scores when the p-value-transformed distance data are combined via Fisher's method. We notice a significant improvement of the M-score when fusing data. The null hypothesis H_0 that the two settings yield the same result is rejected at level $\alpha \leq 10^{-3}$. However, it is not sure that the underlying structure of both simple and merged data are exactly the same. This means we should determine an optimal value for k for each data type and elaborate further on these, instead of comparing structures with identical k in both representations.

We use the stability-based method (see Section 3.5.1) to determine the optimal number of clusters for expression and text-augmented data. We recall that

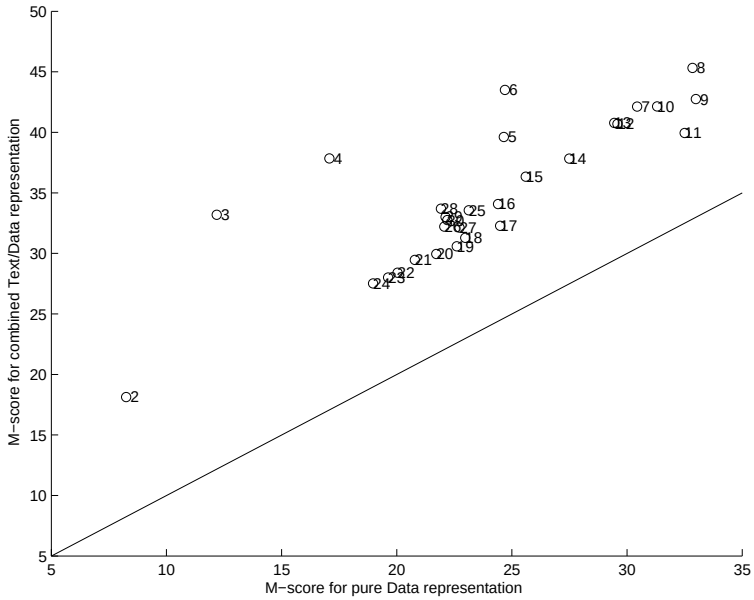


Figure 7.5: Motif Scores of hierarchical clusterings for various cutoffs k, applied on expression matrices (x-axis) versus combined expression/text distance matrices (y-axis). Both data types were integrated by combining p-values derived from the corresponding distance matrices. We notice a significant improvement of the M-score when fusing expression and text data.

this method uses the distribution of pairwise similarity between clusterings of subsamples of a data set as a measure of cluster stability. We have chosen this method for its exploratory nature and its ability to exploit the computational advantages of hierarchical clustering. As explained in Section 3.6, we modified the stability method so to apply it in an iterated manner in order to accommodate for the complexity of the data. We recall Figure 3.7 for the exact procedure. After choosing an initial partitioning into five clusters, the stability method on each of these gene clusters yield the stability diagrams plotted in Figure 7.6 and 7.7. As it happens, we find, for both expression and combined data, an optimal number of $k = 14$ stable clusters at the end of this second iteration. Although our iterated stability procedure can be applied in different ways (e.g., in more than two iterations and starting from smaller k), we converged to similar results in all cases.

The adjusted *Rand* index, R_{adj}, which measures the difference between pure data and text-augmented clustering, attains a value of 0.0997 still indicating an significant, but weak correspondence[2]. This is our starting point to proceed with a detailed discussion of the correspondences and differences between the two solutions.

[2]for random overlap $R_{adj} \simeq 10^{-4}$.

7.4 Analysis of expression clusters

In what follows we provide a 'classical' (i.e., along the line of seminal functional analyses as done by Tavazoie [148] and Spellman [141]) biological interpretation of a (purely) expression-based clustering result in terms of found motifs and extracted literature terms. This exercise is an instantiation of the pipeline depicted in Figures 1.4 and 1.5 in the introductory chapter. The following discussion enables a deeper understanding of cluster analysis and, above all, it acts as a reference point to compare the results from our integrated framework as treated in the following Section 7.5.

For the clustering on pure data, 14 clusters are obtained of which three show both a periodic profile and an enrichment of relevant motifs (see Figure 7.8).

Cluster 3 (201 genes) shows an enrichment of the G1/S-specific MCB (*Mlul* cell Cycle Box) and Mbp1 motifs. The latter is related to DNA replication and repair. The phase specificity is confirmed by both expression (Figure 7.9) and text profiles (Table B.5) of the cluster, and by the results of Tavazoie *et al.* [148] and Lee *et al.* [103]. High-scoring terms such as 'cyclin', 'cell cycle', 'bud', 'G1', 'mismatch repair' and 'DNA replication' are related to the presence of MCB and Mbp1 motifs.

Cluster 6 (78 genes) is characterized by the occurrence of ECB motifs (Early cell Cycle Box), specific for the M/G1 phase [148][141]. The expression profile (Figure 7.9) confirms the observed phase specificity, peaking in the late S and M phase. Additionally, the cell cycle specificity, the observed phase-specificity and the significant motifs for this cluster are supported by the text profile of the cluster with high-scoring terms such as 'cell cycle', 'bud', 'spindle pole body', 'DNA replication', 'MCM' and 'kinetochore' (ECB). In the S phase, DNA replication takes place, small buds develop and spindle formation starts. In the M phase spindle assembly takes place and the buds reach full size.

Cluster 9 (117 genes) shows enrichment for binding motifs for the Fkh2 and Ndd1 transcription factors, which are known to cooperate during the G2 phase to activate mitosis [103]. The text profile of the cluster in Table B.5 reflects the presence of observed motifs only through the high-scoring term 'mitosis'. Both MEDLINE and GO profiles indicate that this cluster is indeed functionally heterogeneous.

Additionally, the clustering on pure data yields four non-periodic clusters with relevant motifs: cluster 7 (Met31_32p) groups genes involved in biosynthesis of methionine in the S phase. Cluster 8 and 14 (M1a, Rap1) group ribosomal proteins, whereas cluster 2 contains genes with the stress motif STRE.

Hence our analysis is consistent with the important previous studies on the yeast cell cycle. It also proves that the transformation of distances into p-values, which constitutes the core of the proposed statistical integration, gives valid biological results.

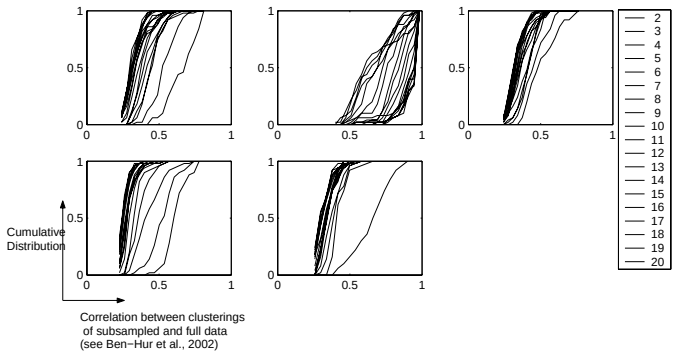


Figure 7.6: Stability diagram for expression data to determine underlying cluster structure (here: number of clusters) from data. Using the last cdf profile that is separated from the continuum as a reference, we estimate from this figure the optimal number of clusters to be $\{3, 1, 3, 5, 2\}$, totalling to 14.

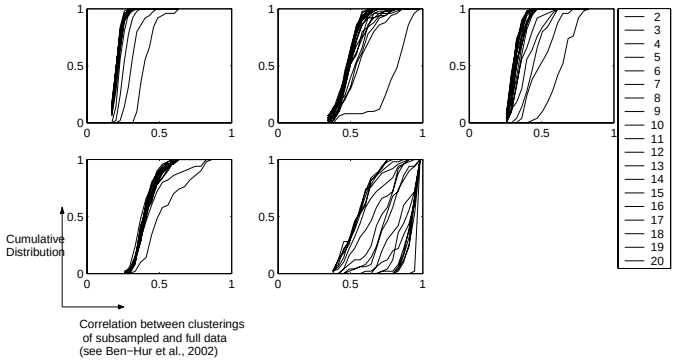


Figure 7.7: Stability diagram for combined data for determining underlying cluster structure (here: number of clusters) from data. Taking the last cdf profile separated from the continuum, we estimate from this figure the optimal number of clusters to be $\{4, 2, 4, 3, 1\}$, a total of 14.

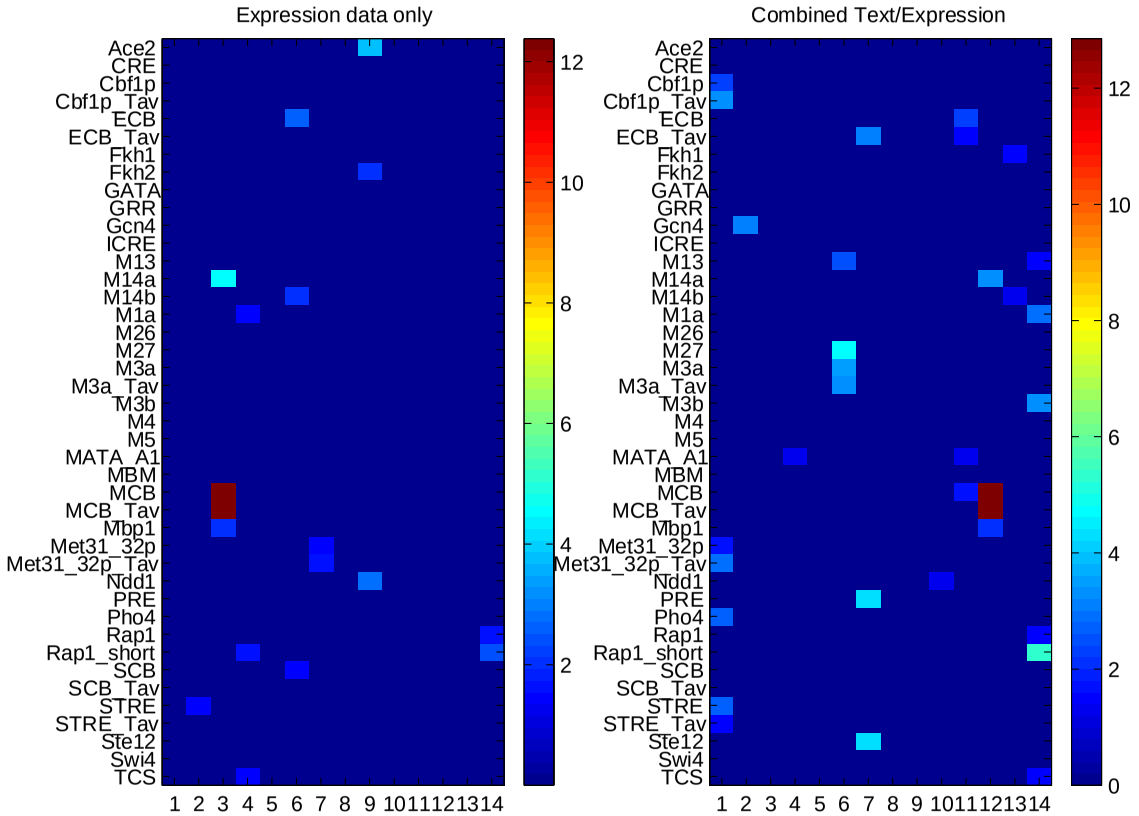


Figure 7.8: Visualization of motif enrichment for expression data (left) and combined text-expression data (right). For both representations we obtained k=14 from a stability analysis .

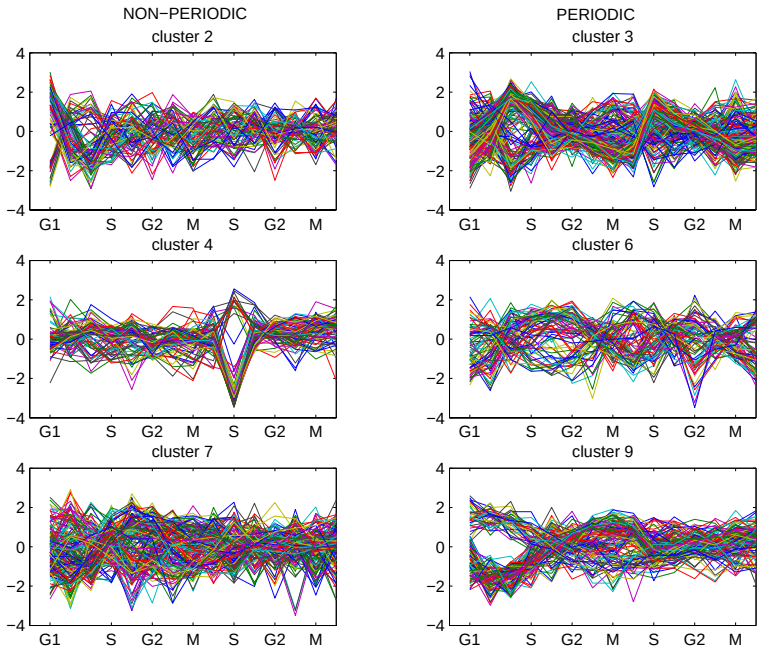


Figure 7.9: Expression profiles of selected gene clusters generated from microarray gene expression data. We show three periodic and three non-periodic clusters.

7.5 Analysis of integrated clusters

7.5.1 Correspondence analysis of both clusterings

As tracking cluster membership of each gene individually over the two representations is hardly attainable, we try to map clusters onto each other by looking at major overlaps and shared significant motifs between the expression clusters and integrated clusters. Consistent with this approach we will contrast expression and text profiles of entire clusters (and not of subsets of clusters for example).

The integrated clustering yields 14 clusters, with an overview of the relevant motifs found in Figure 7.8. In Figure 7.10 we visualize how the cluster structure has changed by adding text information to expression data. We use these two plots as a starting point for the following discussion.

A strong overlap is seen between data cluster 3 and integrated cluster 12, as 40% of the genes in the former are regrouped and make up 64% of the latter. This is supported by the overrepresentation of MCB, Mbp1 and M14a motifs in both cases (Figure 7.8), and by the corresponding expression profiles (Figure 7.11) and text profiles (Table 7.1). Hence, we see that **the integration can conserve strongly relevant cluster structures.**

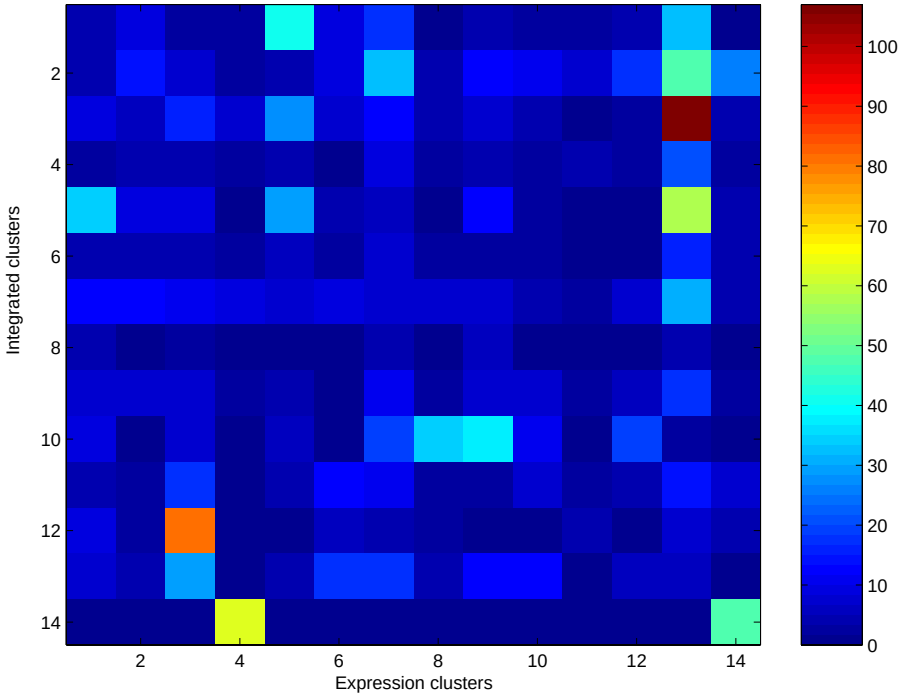


Figure 7.10: A comparison of expression clusters (x-axis) with clusters computed in the combined text/data representation (y-axis). The colors indicate how many genes are shared by the corresponding clusters.

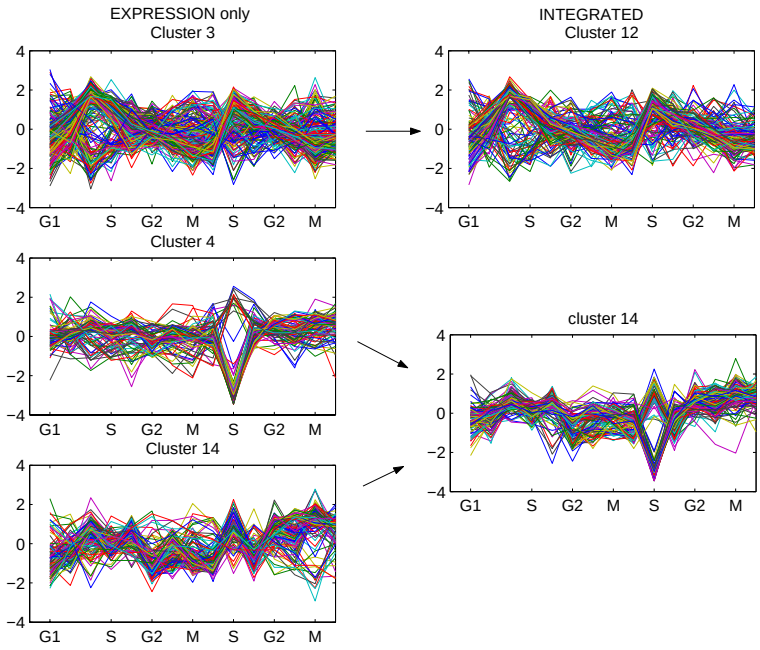


Figure 7.11: Expression profiles of cluster 3 (expression-only representation displayed on the left side) whose cluster structure and phase specificity remain preserved in cluster 12 (integrated representation displayed on the right side), followed by Expression profiles of cluster 4 and 14 (expression-only representation) and cluster 14 (integrated representation). In this case the integrated representation results in a strengthened connection between the Rap1 and TCS motif, although the genes in expression clusters 4 and 14 already share very similar profiles.

Next, we observe a *merging* of data clusters 4 and 14 into integrated cluster 14, resulting in a strengthened connection between the Rap1 and TCS motifs. Although the genes in data clusters 4 and 14 share similar profiles that indicate strong regulation in the S-phase of the cell cycle, they are only clustered together in the integrated representation (Figure 7.11). They represent ribosomal genes corroborated by the term term 'ribosomal subunit' in the text profile (Table 7.1). As we will elaborate further on, this constitutes an example of how **text reenforces functional relationships between genes.**

Based on figure 7.10 large chunks of data clusters 8 (50%) and 9 (37%) join and make up 48% of integrated cluster 10. However, since only Ndd1 is detected (Figure 7.8), a loss of the motif pattern Fkh2/Ndd1 is observed with respect to data cluster 9. Scanning the latter cluster for genes that generate this pattern, we see that ALK1, SWI5 and BUD4 turn out to have Ndd1 as well as Fkh1 in their upstream regulatory region (no genes that solely carry Fkh1 are present). These three genes, however, appear to be present in the corresponding integrated cluster. Failure to detect Fkh2 as a significant motif (and, hence,

Table 7.1: Comparison of text profiles between corresponding data clusters and integrated clusters. We show the text profiles of entire clusters and not of possible subsets (see text for more details). A selection of terms reminiscent of gene function is marked in bold for clarity. **A.** Text profiles support the conservation of data cluster 3 in the integrated clustering (cluster 12). It contains genes peaking in G1/S phase of the cell cycle where DNA replication, synthesis and repair takes place, as well as genes active in the formation of spindles in the early mitosis phase. **B.** Through the addition of text to expression data a strengthened connection between ribosomal genes merge them from expression clusters 4 and 14 into integrated cluster 14. This is an example on how text can reenforce functional relationships between genes.

A		**B**		
Data cluster 3	Integrated cluster 12	Data cluster 4	Data cluster 14	Integrated cluster 14
cyclin	**cyclin**	**duplic**	**duplic**	**duplic**
mismatch_repair	**mismatch_repair**	**ribosom**	histon	**ribosom**
spindl_pole_bodi	**dna_replic**	codon	cytoplasm	pp
dna_replic	g1	auxiliari	codon	**small_ribosom_sub**
replic	left	cytoplasm	**ribosom**	auxiliari
g1	replic	xanthin	**small_ribosom_subunit**	cytoplasm
checkpoint	cell_cycl	characterist	pp	codon
cell_cycl	topoisomeras	acetyltransferas	characterist	acetyltransferas
spindl	rna_helicas	pp	auxiliari	characterist
sister_chromatid_cohesion	checkpoint	**small_ribosom_subunit**	adenyl_kinas	acetyl
repair	**repair**	acetyl	ubiquitin	g
bud	**bud**	chitin_deacetylas	h3	xanthin
cohesin	phase	pyruv_kinas	h4	w
microtubul	**spindl_pole_bodi**	phosphoglycer_mutas	sterol	amp
mitosi	kinas	cell_wall	cell_wall	chitin_deacetylas
cohesion	dna_primas	g	acetyltransferas	pyruv_kinas
telomer	trna	sorbitol	ergosterol	translat
cell_wall	chromosom	chitinas	g	r
histon	septin	elong	chang	chang
chitin_synthas	protein_kinas	cell_cycl_regul	elong	left

Table 7.2: Comparison of text profiles between corresponding data clusters and integrated clusters. **A.** Genes involved in cytokinesis (The division of a cell into two daughter cells following mitosis) and budding from data clusters 8 and 9 are regrouped in integrated cluster 10, indicated by recurrence of the terms 'phospholipase', 'cell wall', 'bud' and 'cytokinesis'. The text profile additionally reveals the biological processes 'thiamine biosynthesis', 'heat shock response' to be present. As a control sub-clustering the latter cluster with TXTGate indeed confirmed the presence of genes involved in biosynthesis of amino acids, stress, regulation of mitosis and budding. **B.** The term profile of expression cluster 1 versus integrated cluster 5 indicates an improved grouping of genes involved in the respiratory chain with reminiscent terms like 'cytochrome c', 'respiratory chain', and 'mitochondrial'.These profiles demonstrate a plausible reorganization of the genes by the integrated representation.

A			**B**	
Data cluster 8	Data cluster 9	Integrated cluster 10	Data cluster 1	Integrated cluster 5
checkpoint	mitosi	hsp90	trehalos	**cytochrom_c_oxidas**
phospholipas_d	succin_dehydrogenas	hydroxymethylpyrimidin	**mitochondri_ribosom**	**mitochondri_ribosom**
transcript_factor	arp2	thiamin	fatti_acid	**mitochondri**
protein_kinas	**cytokinesi**	chaperon	**mitochondri**	ubiquitin
phosphatas	hsp90	swi	synthas	proteasom
mannosyltransferas	peroxisom	transketolas	peroxisom	protein_biosynthesi
cell_wall	**bud**	**cell_wall**	allantoin	**cytochrom**
tfiiic	cyclin	transcript_factor	rna_helicas	mrp
prospor_membran	hydroxymethylpyrimidin	thiamin_biosynthesi	copper	translat_activ
kinetochor	**cell_wall**	kynurenin	autophagi	deubiquitin
histon	inositol	quinolin	proteasom	assembli
chaperon	myosin	heat_shock_protein	coenzym	**atp**
mate	transcript_factor	snf	reductas	translat
formaldehyd	photolyas	floccul	signalosom	oxidas
cell_cycl	cell_cycl	pyruv_decarboxylas	water_channel	synthas
spindl	singlet_oxygen	**bud**	translat	commun
zinc	pyrophosphorylas	**cytokinesi**	origin_recognit_complex	rub1
sporul	mitochondri	nicotin	**cytochrom_c_oxidas**	**respiratori**
chromatin_assembli	6_phosphofructo_2_kinas	**phospholipas_d**	zinc	**mitochondri_small_ ribosom_subunit**
bud	mitot	phosphoribosyl	cell_wall	**succin_dehydrogenas**

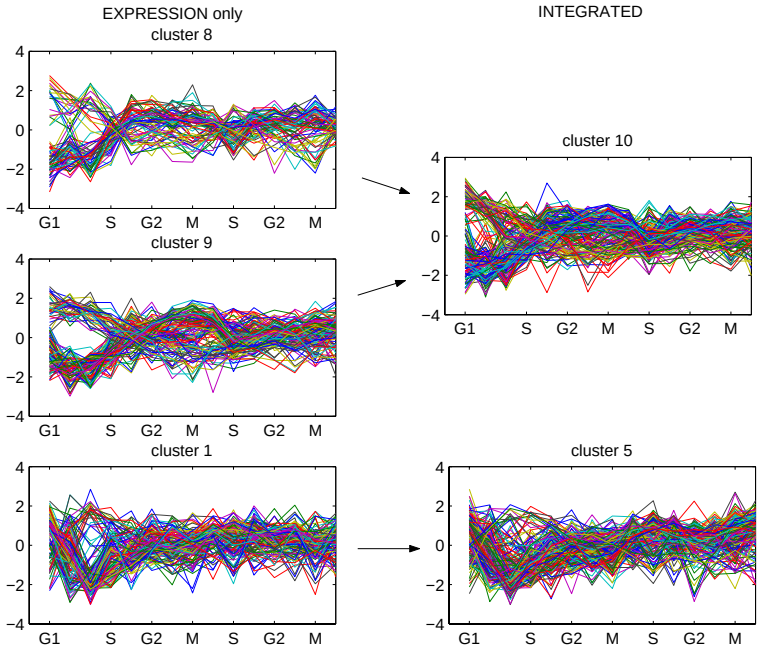


Figure 7.12: Expression profiles of cluster 8 and 9 (expression-only representation) versus cluster 10 (integrated representation). Conservation of phase specificity is less apparent and genes are grouped increasingly according to their function. Similar observations hold for comparison between cluster 1 (expression-only representation) and 5 (integrated representation).

the pattern Fkh2/Ndd1) is therefore a consequence of the difference in cluster size between expression cluster 9 (117) and integrated cluster 10 (151) pushing the motif below the detection threshold (see Equation 4.5). In this case, **adding text influences the outcome indirectly through the structure of the new cluster.**

Examining the expression profiles of the respective data and integrated cluster (Figure 7.12), we see that although the integrated profile is more diffuse, it still displays a strong regulation signal during mitosis and transition to G1. The text profile in Table 7.2 reveals multiple active biological processes ('thiamine biosynthesis', 'heat shock', 'cytokinesis', 'bud', 'transcription factor') in the integrated cluster. Sub-clustering with TXTGate indeed confirmed the presence of genes involved in the biosynthesis of amino acids, heat shock response and stress, regulation of mitosis and cytokinesis.

For data cluster 1 and integrated cluster 5 the expression profiles display a trend of regulation in the G1 phase (Figure 7.12). The term profile indicates an improved grouping of genes involved in the respiratory chain. A part of these mitochondrial genes were already clustered together, the other part stem from the data cluster 13. So, **the addition of text leads to a reorganization of genes with similar profiles and similar functions into a single cluster.**

From Figure 7.10 we see that data cluster 13 is mainly scattered over integrated clusters 2, 3 and 5. Their expression profiles are shown in Figure 7.13. We have discussed integrated cluster 5 above. Integrated cluster 2 is a larger, more heterogeneous cluster grouping 200 genes involved in the synthesis of various proteins (arginine, histidine) and in energy production (gluconeogenesis, galactose metabolism). Finally, integrated cluster 3 holds 217 genes mainly involved in both the synthesis and the metabolism of sterols and fatty acids. These observations suggest that **integration regroups genes with weakly similar expression profiles according to reported functions.**

Integrated cluster 6 groups genes predominantly involved in the transscription process, RNA metabolism and mRNA splicing indicated by terms such as 'ATP dependent RNA helicase', 'pre mRNA splice factor' and 'mRNA process' (results not shown in detail). Significant occurrence of motifs M3a were independently found by Tavazoie *et al.* Co-occurrence of motifs PRE and Ste12 as found in integrated cluster 7 (Figure 7.8) is accordingly retrieved in literature, where STE12 is reported to bind to the pheromone response element (PRE) present in the upstream region of genes whose transcription is induced by pheromone [167]. This supports the contention that **the addition of text can unveil new relevant motif patterns.**

7.5.2 Loss of the SCB motif

We observe that data cluster 6, which exhibits the cell cycle specific motifs ECB, M14b and SCB and contains genes involved in the organization of the centrosome (spindle pole body and kinetochore formation), is shattered over various integrated clusters. In the motif plot, Figure 7.8, we see that for integrated cluster 11 motifs ECB and M14b are retrieved. However, the SCB motif is no

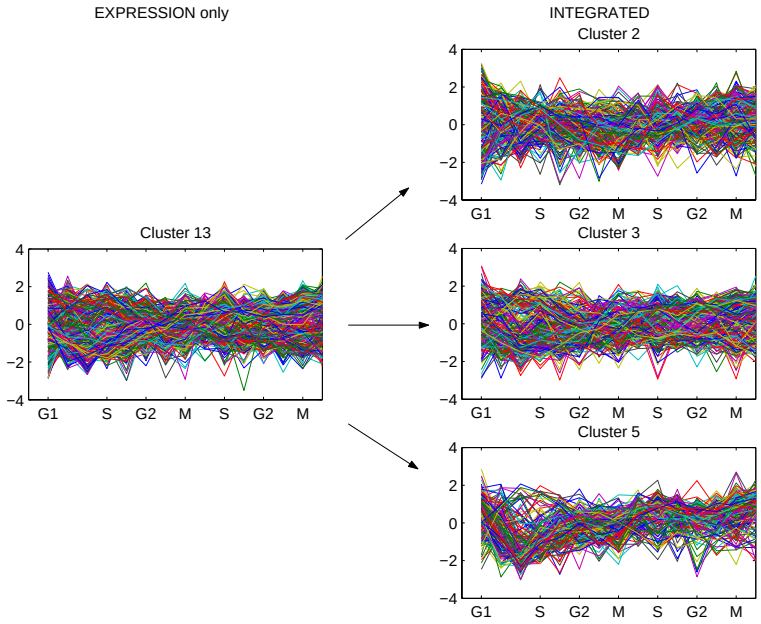


Figure 7.13: Expression profiles of the cluster 13 (expression-only representation) versus clusters 2,3 and 5 (integrated representation). Genes with poor expression similarities are redistributed over other clusters according to known functions.

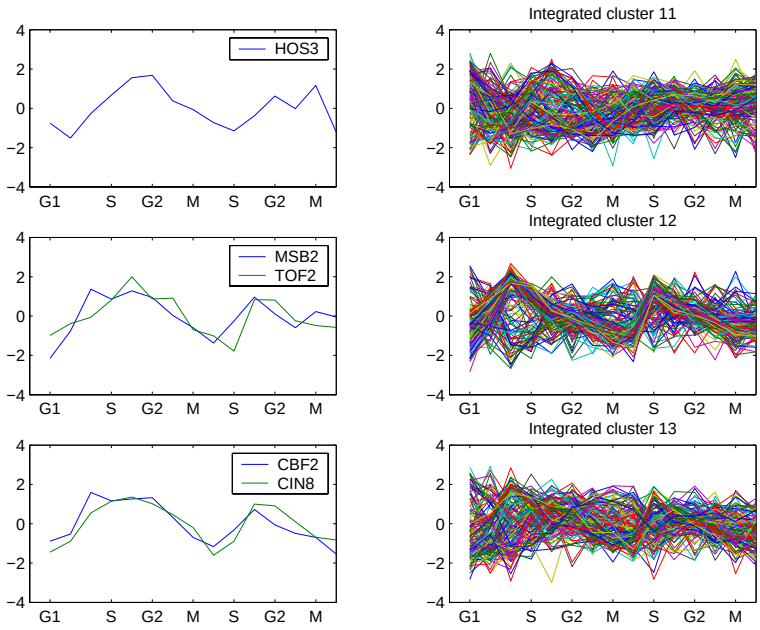


Figure 7.14: Expression profiles of those genes from data cluster 6 that carry the SCB motif versus profiles of integrated clusters 11, 12 and 13 they appear in.

longer significantly present in any of the integrated clusters. To gain insight in the underlying reasons we trace the seven genes that contain SCB in the original data cluster 6.

One gene, HOS3, affects the histones (proteins complexed to DNA in chromatin and chromosomes). It is assigned to aforementioned integrated cluster 11, which indeed groups genes related to chromatin structure with terms as 'telomere', 'histone deacetylase' and 'chromatin assembly'. MSB2 and TOF2 move to integrated cluster 12, a pronounced periodic cluster discussed previously. CBF2 and CIN8, both related to spindle microtubules reside in integrated cluster 13 characterized in its turn by keywords like 'kinetochore', 'spindle pole body' and 'sister chromatide cohesion'. Figure 7.14 shows the expression profiles of these genes along with those of the integrated clusters. It teaches us that, although they are periodically similar with peaks in the S and G2 phase, these genes are clustered more according to function due to the addition of literature information.

ABF1 is a transcriptional regulator that is not explicitly linked to the cell cycle [133] and moves to the rather heterogeneous cluster 2 in the integrated representation for its the high scoring terms 'DNA binding' and 'transcriptional activation', and due to its weak periodicity, as shown in Figure 7.15.

One particular case that shows a limitation of our current approach is YOP1, a currently unannotated gene in SGD, assigned to integrated cluster 6 for the

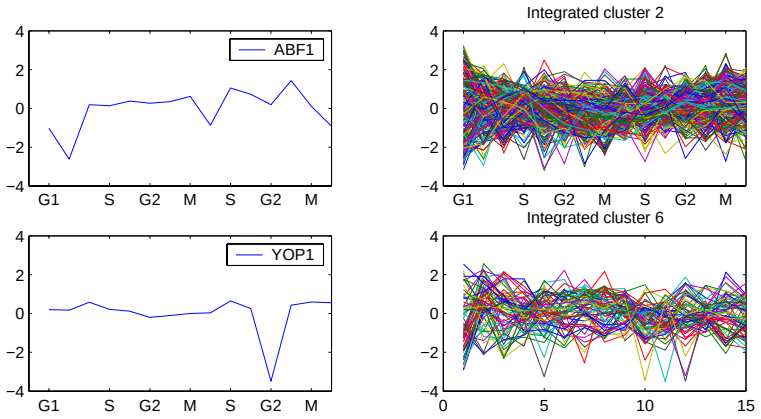


Figure 7.15: Expression profiles of those genes from data cluster 6 that carry the SCB motif versus profiles of integrated clusters 2 and 6 they appear in.

keyword 'splicesome' in its MEDLINE text profile. However, this keyword stems from an article that doesn't pertain to function (PMID 10734188: 'Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast'). We conclude that integrated clusters provide more functionally coherent grouping at the expense of periodicity and significance of some motifs (SCB in this example).

7.5.3 Emergence of the Cb1fp/Met31_32p/Pho4/STRE motif pattern

A particularly interesting motif pattern is Cb1fp/Met31_32p/Pho4/STRE, found in integrated cluster 1. All genes are implicated in amino acid biosynthesis and this cluster was also found in Tavazoie *et al.* [148] and Spellman *et al.* [141], but not in the data clustering. Seven genes (Table 7.3) were found to exhibit both the Cb1fp and Met31_32 motif. Two (marked in bold) were assigned to other clusters in the data clustering.

Examining the relation between MET30, MET2 and CYS3 in Table 7.4, we first observe an (expected) strengthening (0.3497 to 0.0431) of the relationship between MET2 and MET30 due to their involvement in the synthesis of methionine. Indeed, MET2 is involved in the first step of the methionine biosynthetic pathway, whereas MET30 is known to control cell cycle function, sulfur metabolism, and methionine biosynthesis as part of the ubiquitin ligase complex. Less trivial is the reenforcement between MET30 and MET2 on the one hand and CYS3 on the other hand ($D_{ij}^{data} = 0.5942$ to $D_{ij}^{integr} = 0.1899$ and $D_{ij}^{data} = 0.5447$ to $D_{ij}^{integr} = 0.0228$ respectively). A search in literature confirms cysteine to be essential for transcriptional regulation of sulfur assimilation and methionine biosynthesis [86] [119]. These observations support that functional relations from literature underly the detection of new motif patterns

Table 7.3: Member genes of integrated cluster 1 carrying motifs Cb1fp and Met31_32p in their upstream regulatory region. The addition of text to expression data changed cluster membership of the genes indicated in bold, rendering the motif pattern Cb1fp/Met31_32p/Pho4/STRE significant.

Gene	Data Cluster	Integrated Cluster
MET30	13	1
CYS3	7	1
MET32	7	1
MET2	13	1
MET1	7	1
MET14	7	1
MET28	7	1
MET3	7	1

in expression data.

Table 7.4: p-Value-based distances between MET30, CYS3 and MET2 in both representations. Adding text strengthens the connection between these three genes and critically contributes to rendering the motif pattern Cb1fp/Met31_32p significant in the corresponding integrated cluster.

	Mutual expression distances			Mutual integrated distances		
	MET30	CYS3	MET2	MET30	CYS3	MET2
MET30	0			0		
CYS3	0.5942	0		0.1899	0	
MET2	0.3497	0.5447	0	0.0431	0.0228	0

We have elaborated on how adding text affects cluster structures or motif patterns. In the following examples we continue to demonstrate more in detail how pairwise gene relations are influenced by the integrated representation. Figure 7.17 visualizes the overall behavior of how the pairwise distances of genes in integrated cluster 1 change by including literature information with our method.

Strong enhancement of data-based pairwise gene relations

In a first query we look for pairs of genes from integrated cluster 1 with a big discrepancy between their mutual distances resulting from expression data on the one hand, and the integrated data on the other hand. Put differently, we look for cases where literature radically enhanced the pairwise relationships. We therefore selected two gene pairs, IDH2/ZWF1 and MDH2/GUT2, that displayed a difference

$$(D_{ij}^{data} - D_{ij}^{integr}) \geq 0.9,$$

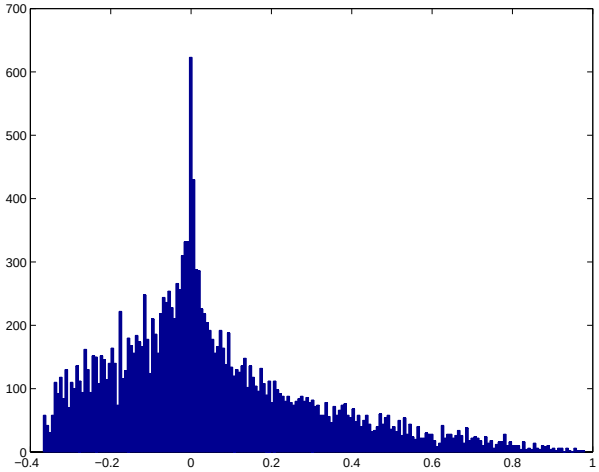


Figure 7.16: Histogram of $(D_{ij}^{data} - D_{ij}^{integr})$, the difference in gene relationships from expression-only and text-augmented data, for genes in integrated cluster 1. There is a tendency to re-enforce as well as to weaken gene relationships, although strengthening can be more radical: Differences close to 1 indicate that the addition of text turns *unrelated* gene pairs into *very related* gene pairs.

as shown in Table 7.5. The corresponding text profiles can be found in Table 7.6. All genes bear a strong connection to the TCA cycle and glucose metabolism with terms as 'tricarboxylic acid cycle', 'NADP', 'isocitrate dehydrogenase', 'glucose' and 'mitochondrial'. Their corresponding expression profiles, plotted in Figure 7.17, show that the cosine distance indeed classify the relations between these gene pairs as overly distant. As can be seen in the histogram (Figure 7.16), this kind of radical overruling of text occurs rarely.

Enhancement of moderate data-based pairwise gene relations

In a second query we look for gene pairs where

$$\begin{cases} (D_{ij}^{data} - D_{ij}^{integr}) \geq 0.2 \\ 0.2 \leq D_{ij}^{data} \leq 0.25, \end{cases}$$

meaning a strengthening of existing data-derived relationships. NDE1/GUT2 are in antiphase (Figure 7.17 and $D_{ij}^{data} = 0.2391$, whereas $D_{ij}^{integr} = 0.0015$ (see Table 7.5). A query in PUBMED elucidates that both are part of systems primarily responsible for mitochondrial oxidation of cytosolic NADH [120]. The pair MDH2/ACO1 is described to be involved in acetate metabolism [112], which reenforces the similarity shared in their expression profiles ($D_{ij}^{data} = 0.2226$). Another interesting example is the pair AAD4/AAD15 which share identical MEDLINE references. The resulting identical similarity between the text profiles upgrades the distance $D_{ij}^{data} = 0.2447$ to $D_{ij}^{integr} = 0.0003$. The underlying

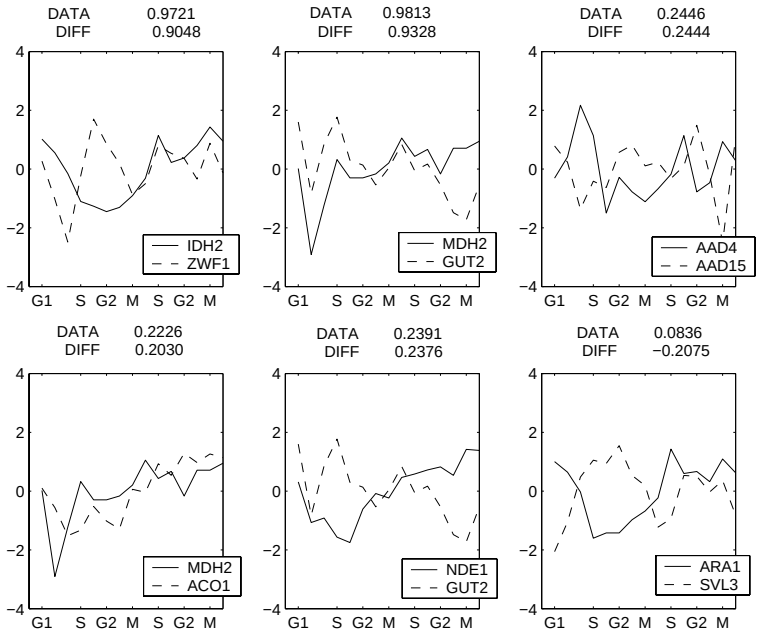


Figure 7.17: Expression profiles of the selected genes from integrated cluster 1 showing strong and medium enhancements, and moderate downgrading of expression-based pairwise relationships. 'DATA' shows the expression-based p-valued distance, 'DIFF' the difference ($D_{ij}^{data} - D_{ij}^{integr}$) for the respective genes. For gene pairs IDH2/ZWF1 and MDH2/GUT2, the cosine-based expression distance quantifies these genes as very distant, which is overruled by the addition of text. Gene pairs AAD4/AAD15, MDH2/ACO1 and NDE1/GUT2 bear a moderate similarity to each other according to the cosine-based expression distance. Also here adding text re-enforces their relationship. Finally, the distance between ARA1/SVL3 is *increased* by the addition of text, due to the absence of matching known annotations.

Table 7.5: Modification of pairwise gene relations by adding text. All genes belong to integrated cluster 1. The data cluster they reside in is indicated between brackets.

Genes		Expression distance	Integrated distance
IDH2 (5)	ZWF1 (2)	0.9721	0.0673
MDH2 (1)	GUT2 (13)	0.9813	0.0485
NDE1 (13)	GUT2 (13)	0.2391	0.0015
MDH2 (1)	ACO1 (13)	0.2226	0.0196
AAD4 (7)	AAD15 (1)	0.2447	0.0003
ARA1 (5)	SVL3 (6)	0.083612	0.2911

Table 7.6: Text profiles of gene pairs IDH2/ZWF1 and MDH2/GUT2. Although these pairs are weakly linked by expression data, they have a strong connection in literature, which is taken into account in the integrated representation.

IDH2	ZWF1	MDH2	GUT2
isocitr_dehydrogenas	dehydrogenas	malat_dehydrogenas	glycerol_phosphat_dehydrogenas
nad	nadph	metabol	nadh_dehydrogenas
isocitr	nadp	acet	glycerol
tricarboxyl_acid_cycl	oxid	carbon	extern
enzym	enzym	enzym	mitochondri
carbon	oxid_stress	cytosol	shuttl
mitochondri	pathwai	glucos	nadh
amp	loss	mitochondri	glycerol_kinas
citrat	transketolas	inactiv	lactat
catalyt	reduc	amino	oxid
acet	metabol	ethanol	ethanol
oxid	sensit	peroxisom	enzym
glycerol	growth	mutagenesi	cytosol
glutam	increas	degrad	glucos
cycl	reactiv	cellular	flavoprotein
dehydrogenas	gener	common	dehydrogenas
metabol	decreas	catabolit	respiratori
nadp	glutathion_reductas	compart	metabol
oxoglutar_dehydrogenas	isocitr_dehydrogenas	isocitr_lyase	reduct
regulatori	glutathion	glyoxyl	redox

reason for the apparent 'transfer' of annotation is explained by the very high level of nucleotide sequence similarity between the AAD genes that makes it difficult to determine if they are all involved in the oxidative-stress response [69]. This case illustrates that the risk of our approach lies in cases where the confidence in the MEDLINE annotations is limited, rather than in case where the number of annotations[3] is low.

Repression of data-based pairwise gene relations

A final illustration is provided by ARA1/SVL3, that resulted from querying instances where

$$\begin{cases} (D_{ij}^{data} - D_{ij}^{integr}) \leq -0.2 \\ D_{ij}^{data} \leq 0.09. \end{cases}$$

Here a strong expression-based relation ($D_{ij}^{data} = 0.0836$) is downgraded ($D_{ij}^{integr} = 0.0836$) by a absolute text-based dissimilarity ($D_{ij}^{txt} = 1$). As our setup automatically avoids that distances between genes without annotations are 1, this means the text profiles of the gene pair exist but share no common terms. ARA1 is involved in carbohydrate metabolism and has two MEDLINE references, whereas SVL3 is related to endocytosis and showed one reference. Upon examining their expression profile we can indeed interpret the cosine metric to have been overly optimistic about the relationship between these two genes.

Notice that from the profiles in Figure 7.17 a strong expression similarity between ARA1 and SVL3 is slightly counterintuitive, as is the dissimilarity of IDH2/ZWF1. A shallow interpretation would be not to trust the adopted cosine distance measure on expression data. However, as already discussed in Chapter 3 we argue that both Euclidean and cosine metric suffer from drawbacks and yield occasional counterintuitive outcomes. Also, the favorable motif analysis of the expression data (see Section 7.4) confirms the validity of our setup. As a final argument we show in Figure 7.18 the scatter plot describing the difference in M-scores when the Euclidean metric (instead of $1 - \cos$) is used for computing the expression-based dissimilarities. Also here the integrated representation performs better providing an independent indication of the favorable effect of our method.

7.6 Discussion

One important question that arises when using expression data and textual information interchangeably, is in which aspects the two data types differ. While expression data tends to favor clusters of co-expression (e.g., phases in the cell cycle), textual data on the other hand enlightens a more functional dimension of a gene group – a point also made in [79]. We have demonstrated in our

[3]The number of annotations can be considered to be (e.g., logarithmically) related to the level of understanding of a gene.

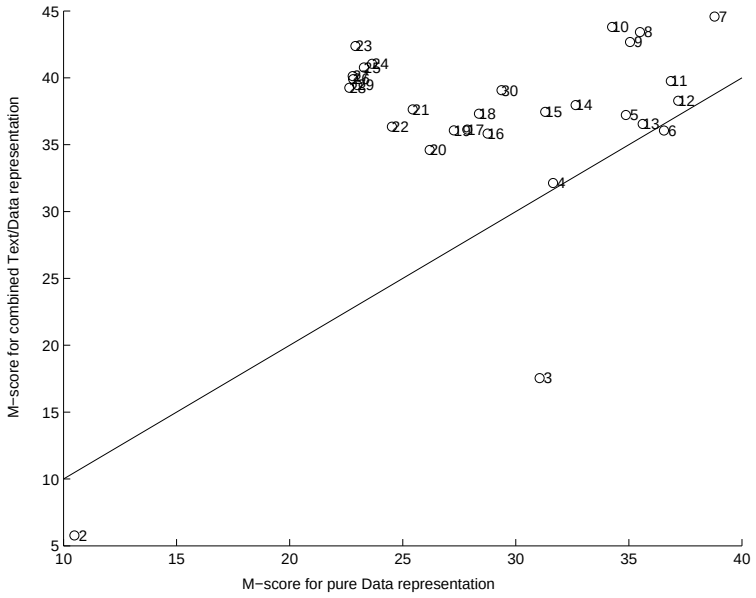


Figure 7.18: Motif Scores of hierarchical clusterings for various cutoffs k, applied on expression matrices (x-axis) versus combined expression/text distance matrices (y-axis). Both data types were integrated by combining p-values derived from the corresponding distance matrices. Here Euclidean distance lies at the basis for the expression data, whereas cosine was used for text. This provides an independent indication of the favorable effect of our method.

controlled experiment how fusing data changes the clustering results in such a way that all expression, text and motif profiles remain biologically relevant. Depending on the questions asked however the emphasis in the results can shift to either of the three directions. For example, the formation, duplication and assembly of spindle pole bodies is not limited to a single phase of the cell cycle, and therefore adding text to expression profiles will diminish the phase specificity of the clusters. However, it is not necessarily biologically more relevant to focus on phase specificity (and thus purely on the data). If one is really interested in obtaining phase-specific clusters rather than functionally coherent clusters, the results would only be improved by extending the vocabulary with terms and especially phrases that are very specific to the different phases of the cell cycle by making a clear distinction between e.g., spindle pole body formation and spindle duplication. In GO these concepts are registered as 'spindle assembly' and 'spindle pole body duplication', and unless they are reported as such in literature, only the constituting keywords are recognized in the text. A more complete phrase list would certainly accommodate for this problem, but variations in natural language such as 'the formation of the spindle pole' would remain a problem.

We have demonstrated in our controlled experiment how fusing data changes the clustering results such that expression, text and motif profiles remain biologically relevant. The choice of validating results in motif space was driven by the motivation to use a framework independent enough to draw justified conclusions. We explicit that the motif framework addresses *direct* regulation of gene expression by the given transcription factors that bind on the motifs, and does not aim the reconstruction of genetic networks for example. In response to the overhead involved in evaluating clustering solutions on the presence of significant motifs, we used our developed M-score. We additionally mention that, although perhaps a more straightforward choice (see e.g., Section 4.1.3 for a GO-based clustering score), we did not use GO in our validation framework as it lies at the basis of the domain vocabulary that acts as a view on literature. Moreover, as information from GO is partly built on information embedded in MEDLINE abstracts, we might have ended up in circular confirmations of truth. As multiple sources will be increasingly considered simultaneously, exploring correlations between 'summarizing' scores based on expression, GO [79], literature (Chapter 5, [131]), pathways [169], sequence [160],... will be of increasing importance.

With both data types treated *on equal footing* in the p-value framework – typically used to combine evidence coming from repeated experiments – we faced the question to which extent pairwise relations from the data were enhanced, or obfuscated, by adding text. The emergence of significant motifs that did not appear previously in the controlled setup confirmed the net beneficial effect of our approach. However, the present keyword-based representation has its limitations and pushes forward several important issues for future improvements:

Phrases To balance between complexity and efficiency, we chose GO as a canvas to literature-encoded information. However, many open questions remain on what to choose as an atomic entity for the text index (be it

a stemmed word, a phrase or a concept), an issue already illustrated in [105]. We acknowledge that better phrases, for one thing, would improve *interpretability* of the text profiles significantly.

GO structure It is still unclear whether a separation of the domain vocabulary according to the top-level GO branches - molecular function, biological process and subcellular location - would eliminate erroneous associations between genes. Especially terms referring to molecular functions such as kinases, enzymes, transcription factors,... showed a tendency to dominate the text profiles.

Genes with multiple function As a gene with multiple functions will have a more diverse text profile, its relation to an gene having a unique, shared function will be lower than the relation between two genes sharing their unique functions. Proper function disambiguation requires that some form of contextual information is provided (e.g., using terms describing the experimental question or using information from neighboring genes in a given space).

Spurious abstracts Curated literature annotations are not perfect, and abstracts describing genetic properties, sequencing efforts, or irrelevant mutational analysis regularly occur. Document classification strategies as in Leonard *et al.* [104] and Raychaudhuri *et al.* [131][130] already accommodate well for this problem.

Negations Although negative assertions are an important source of information, they require more parsing-related techniques.

Hence, improvements on how the text model represents biological function will directly affect the quality of the integrated representation. Likewise, advances on how to generate more reliable expression data, how to calculate more accurate similarities between expression profiles, or how to generate better cluster patterns, will exert their influence on the integration. For example, we mainly worked with the Euclidean and cosine-based distances in expression space and witnessed cases where text overcame the limitations of these choices. However, measures that combine the advantages from Euclidean and correlation-based distances exist (e.g., see Bickel [51]). Including such modifications, we expect similar overall trends when applying our framework, but is yet unclear in which aspect the results will differ.

Finally, we observed that fusing heterogeneous distance matrices requires caution in terms of scaling problems. We circumvented this issue by transforming distances to p-values allowing a statistically more principled integration of text and data. Also within the proposed p-value framework possibilities to include weights attributing *confidence* to the adopted information sources exist. The weighted Fisher method combines p-values for r sources via the statistic

$$\mathcal{S} = p_1^{\lambda_1} p_2^{\lambda_2} \cdot \ldots \cdot p_r^{\lambda_r},$$

whose distribution is analytically determined only when $\lambda_1, \lambda_2, \ldots \lambda_r$ are mutually distinct [88]. One way to determine weighting factors as λ automatically would be to rank the the data sources according to classification performance on a validated, functionally coherent gene set.

We conclude that although a joint analysis of heterogeneous data – exemplified with expression data and text-based information – poses several barriers to successful experimentation, it has rewarding effects when trying to overcome the erroneous nature of large scale genomic data.

Chapter 8

Conclusions

To conclude this dissertation we give a general discussion of the main accomplishments of this work along with some future research directions.

8.1 Accomplishments

The main objective of our work was to develop strategies to incorporate text-based information in the analysis of microarray gene expression data. It is a contribution to the more general need currently experienced in biology and biomedicine for more intelligent computational data analysis. Present-day biological data is characterized by several, intertwined, features. We reformulate these features (also presented in Chapter 1) and integrate our contributions into their description.

1. There is a large discrepancy between the pace with which biological data is produced and the rate at which information is extracted. One of the barriers responsible for this has been the lack of transparency in publicly disclosed expression data. To address this, we have contributed to the development of MIAME 1.1 [54], a proposed standard for reporting about gene expression experiments. In order to furnish such kind of information from a laboratory environment, we have developed and implemented a first generation information management system (LIMS) at the VIB MicroArray Facility.

2. The complexity of functional genomics requires approaches that include multiple, possibly heterogeneous, perspectives. This urges methods to integrate data from different information sources. In Chapter 5 we did preparatory work and developed a literature-based score to assess the cohesiveness of a gene group in the text world. We have shown that gene profiles based on text tends to favor functional relationships between genes. Having established a good representation, we continued to deeply integrate literature-based information with microarray analysis. We combined gene

expression profiles and text profiles in a meta-analysis framework in Chapter 7 and demonstrated the benefit of fusing both sources by implicating a third perspective. Indeed, through validation in terms of known transcription factor binding sites, we have demonstrated the overall benefit of using two information sources jointly in an exploratory analysis.

3. Interpreting and managing the steady streams of results (as produced by gene expression analysis), and efficiently connecting them to existing knowledge makes up a considerable bottleneck in the scientific analysis of biological data. After having quantitatively validated the ability of our text mining approach to detect functional cohesion of gene groups in Chapter 5, we leveraged our framework to a useful platform for text-based gene analysis. TXTGate, the text mining web service presented in Chapter 6 offers several *views* on gene literature. The application profiles groups of genes (that typically stem from gene expression cluster analysis) by returning (a) overrepresented function- or disease-related keywords characterizing the gene group, or (b) related gene names that are co-cited with these genes in literature. Sub-clustering and links to external resources allow for an in-depth analysis of the resulting term profiles. We illustrated through various case studies how TXTGate speeds up browsing through relevant literature and how new, potentially relevant genes (e.g., ACN9) can be identified.

4. Exploratory undertakings as they occur in this thesis often involve a wide range of possible outcomes to explore. This is especially true for gene expression clustering which, in itself, is an ill-defined problem in the sense that often many solutions are biologically advisable. To keep these cumbersome degrees of freedom under control, and to spare the biological expert, we have developed a motif-based heuristic that estimates the relevance of a clustering solution in terms of transcriptional regulation. This M-score, along with an elegant method to determine an optimal number of clusters based on subsampling, was presented in Chapter 4.

Evidently, there are many open ends to the research presented. In the following section we will provide some future directions with respect to the main topics in our work.

8.2 Discussion and Future directions

8.2.1 Semantically-oriented text representations

The bag of words representation which encodes terms in a vector space is a standard setup in text mining research. Although in essence a simple model, we have spent considerable effort to make it capable of efficiently encoding information about gene function. Indeed, the idiosyncrasies of biological information make this a far from trivial task. We have shown the usability of this framework in several ways. TXTGate exploits the availability of multiple publicly available

resources and implements our idea to explore large collections of text through multiple views offered by several domain-specific vocabularies. The integrated clustering method shows how text can be used as prior knowledge in a learning algorithm.

Stand-alone multi-word terms however, are not always sufficient to capture subtle biological facts. To import semantics into the vector space model one could consider exploiting the structure present in some of the ontologies used. For example, in our work we use GO merely as a term list. Although this is a first step towards more interoperability between various biological resources, increasingly exploiting the structure of the ontology can bring in important nuances about the extracted information. One example is to construct an ontology-projected, literature-based similarity measure. The idea is to parse gene-related documents for terms that occur in the ontology (hence, literature-based), but instead of computing gene similarities in term vector space, term-based similarity can be measured within the ontology, which has a richer structure (hence, ontology-projected; see e.g., Lord *et al.* [106]). To this end, a prior mapping of terms to the nodes in the ontology needs to be constructed.

Instead of using external sources such as ontologies, related terms can also be organized into term clusters using co-occurrence networks, term clustering or Latent Semantic Indexing. Implicitly, such approaches boil down to automatic ontology construction[1]. Porting these approaches to genomic text mining still remains an open challenge.

Finally, documents often not only contain terms, but also link to a variety of other documents, concepts,.... The evident example is references to genes or diseases within a document, which is already exploited by some applications (see Chapter 6). However, one could envision a text mining approach incorporating disease links, functional links, gene links, document links, database links and terms in one mathematical framework (see e.g., Modja [115]). Text mining and the emerging discipline called 'link analysis' indeed appear similar in their requirements for effective techniques to extract and utilize information hidden in the data resource or a data network [31].

8.2.2 Advanced integration of information

Recent technological advances enable us to collect many different types of data at a genome-wide scale, including DNA sequences, gene and protein expression measurements, protein interactions, database annotations and literature references. Results jointly learned from more than one type of data are likely to lead to new insights that might not be apparent from analyzing one type of data in isolation. We focused on how fusing text and data can yield the expected improvements. In Chapter 7, we fused data by transforming distances in both spaces to p-values. It allowed for a more principled integration of both data types and circumvented the observed scaling issues. Our meta-clustering has nice similarities to statistical meta-analysis and allowed for an improvement of

[1]The term Latent Semantic Indexing, for example, is indeed reminiscent of this goal.

the overall clustering result.

Rather than combining kernels as proposed, other extensions such as Canonical Correlations Analysis (CCA) as in Yamanishi *et al.* [163] could be envisioned. In CCA, one searches for directions in two different spaces (e.g., one in expression space, and one in text space), such that the projections of corresponding data points in both spaces show the largest possible correlation among all such directions. Whereas in principal component analysis (PCA) one projects data in a strictly unsupervised way onto directions of maximal variance, in CCA, projections are made so the different information sources supervise each other.

Alternatively, distance matrices can be regarded as linear kernels [144] and hence generalized to nonlinear cases. Kernel-based data fusion as in (7.1) is currently a active research topic and include issues such as automatic determination of importance weights similar to λ [102], or the investigation into other kernel combinations [75]. Also, many ways remain open to integrate heterogeneous data 'early on' (i.e., by appending data types instead of joining kernels). Proper discretization, for example, can allow to merge feature vectors without the encountered scaling problems and to apply, for example, Latent Semantic Indexing on the merged data matrix to discover low rank approximations that take into account both expression and text [138].

Appendix A

Overview of external databases

eVOC

eVOC [101][6] is a thesaurus consisting of four orthogonal controlled vocabularies encompassing the domain of human gene expression data. It includes anatomical system-, cell type-, pathology-, and developmental stage-related terms.

GeneCards

The GeneCards database [132][10] offers concise information about human genes (and their mouse homologs), with a focus on their cellular functions and medical aspects.

GO

GO is a dynamic controlled hierarchy of (multi-word) terms with a wide coverage in life science literature and genetics in particular [149][9]. GO is more extensively discussed in Section 4.1.

HUGO

The HUGO nomenclature database contain approved symbols for over one half of the genes in the human genome [113].

LocusLink

LocusLink [127][17] provides a single query interface to curated sequence and descriptive information about genetic loci. It presents information on official nomenclature, aliases, sequence accessions, phenotypes and related web sites. Next to literature references also brief, literature-extracted summaries are available through a feature called GeneRIF. An example is shown in Figure A.2.



Display Abstract Show: 20 Sort Send to Text

1: Clin Chim Acta. 2002 Mar;317(1-2):77-84. Related Articles, Links

ELSEVIER FULL-TEXT ARTICLE

Residual galactosylsphingosine (psychosine) beta-galactosidase activities and associated GALC mutations in late and very late onset Krabbe disease.

Harzer K, Knoblich R, Rolfs A, Bauer P, Eggers J.

Institut fur Hirnforschung, Universitat Tubingen, Calwer Strasse 3, D-72076, Tubingen, Germany. Harzer-Rottenburg@t-online.de

BACKGROUND: Krabbe disease (globoid-cell leukodystrophy; GLD) is caused by mutations in the GALC gene. Beta-galactocerebrosidase (GALC) is a specific beta-galactosidase which is defective in GLD. About 90% of GLD patients have an infantile course by fatal cerebral demyelination, but 10% have a later onset (LOGLD) of symptoms and survive for one or several decades. METHODS: Activities of GALC towards galactosylceramide (GC) and galactosylsphingosine (psychosine; PS) were determined in white blood cells and cultured fibroblasts derived from GLD patients and controls using tritium-labelled natural substrates. In the galactosylsphingosine (psychosine) beta-galactosidase (GALC-PS) assay, a thin layer chromatographic technique was used to separate enzymatically released radioactive galactose. RESULTS: Both galactosylceramide beta-galactosidase (GALC-GC) and GALC-PS activities were reduced by at least 85% of the normal in all but 2 of the 10 GLD patients studied. In particular, one 23-year-old severely demyelinated LOGLD patient was strongly deficient (11% of the normal) in GALC-GC but apparently normal for GALC-PS activity. This patient's GALC genotype was the 30-kb-deleted/502T allele combined with a wild-type allele in the 1637C background known to slightly reduce GALC-GC activity. Further, of six LOGLD patients, both of 62- and 63-year-old brothers had the deleted allele combined with an 809G>A mutated 1637C allele. The sibs had strongly reduced GALC-GC and GALC-PS activities but became clinically remarkable only in their 50s with a severe mental downhill course in one of them. CONCLUSIONS: A GALC genotype with one deleted and one polymorphic GALC activity-reducing allele can lead to enzymatic and clinical signs of LOGLD in the absence of marked GALC-PS deficiency. If an active PS hydrolysis in the fibroblasts of a LOGLD patient also reflected such hydrolysis in the brain, the psychosine hypothesis for GLD may need to be revised.

PMID: 11814461 [PubMed - indexed for MEDLINE]

Figure A.1: Example MEDLINE abstract retrieved through the PUBMED interface. It discusses the link between a mutations in the gene GALC and Krabb's disease.

MEDLINE

The world's largest medical library, MEDLINE, contains over ten million references to scientific publications in the biomedical domain. In many cases the abstract is available. PUBMED [28] is the web interface to the MEDLINE database. An example is given in Figure A.1.

MeSH

The Medical Subject Headings (MeSH) is a controlled vocabulary produced by the National Library of Medicine and used for indexing, cataloging, and searching for biomedical and health-related information and documents [22].

MIPS

The Mips Functional Catalogue Database [114] provides a search tool to browse the functional categories from various organisms, including yeast[23].

OMIM

This database is a catalog of human genes and genetic disorders. OMIM [26] contains textual information and references. It also contains copious links to MEDLINE and sequence records in the Entrez system (http://www.ncbi.nlm.nih.gov/entrez/), and links to additional related resources at NCBI and elsewhere.

TRANSFAC

TRANSFAC is the database on eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles [111][32].

SGD

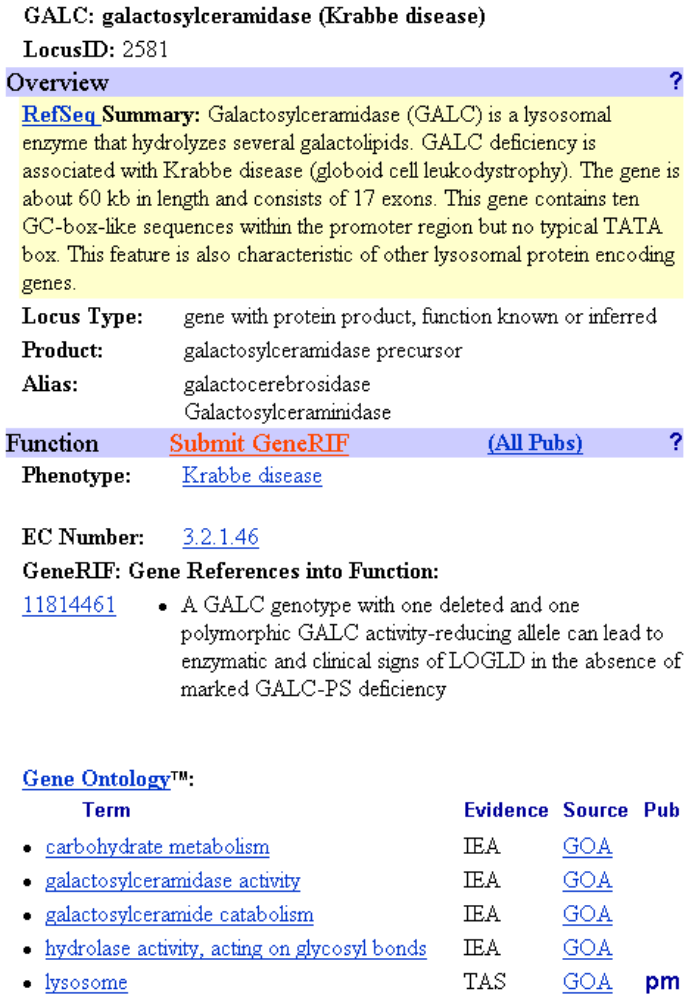
The Saccharomyces Genome Database is a scientific database of the molecular biology and genetics of yeast. It includes a variety of gene annotations (including GO), gene-literature references and gene expression information [29].

SWISS-PROT

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases[30].

UMLS

The Unified Medical Language System is large database designed to link biomedical vocabularies together from disparate sources such as clinical terminologies, drug sources, vocabularies in different languages, and clinical terminologies. Contrary to MeSH, it is an ontology in the sense that relationships between terms are encoded in the database [34].



GALC: galactosylceramidase (Krabbe disease)

LocusID: 2581

Overview ?

RefSeq Summary: Galactosylceramidase (GALC) is a lysosomal enzyme that hydrolyzes several galactolipids. GALC deficiency is associated with Krabbe disease (globoid cell leukodystrophy). The gene is about 60 kb in length and consists of 17 exons. This gene contains ten GC-box-like sequences within the promoter region but no typical TATA box. This feature is also characteristic of other lysosomal protein encoding genes.

Locus Type: gene with protein product, function known or inferred

Product: galactosylceramidase precursor

Alias: galactocerebrosidase
Galactosylceraminidase

Function Submit GeneRIF (All Pubs) ?

Phenotype: Krabbe disease

EC Number: 3.2.1.46

GeneRIF: Gene References into Function:

11814461 • A GALC genotype with one deleted and one polymorphic GALC activity-reducing allele can lead to enzymatic and clinical signs of LOGLD in the absence of marked GALC-PS deficiency

Gene Ontology™:

Term	Evidence	Source	Pub
• carbohydrate metabolism	IEA	GOA	
• galactosylceramidase activity	IEA	GOA	
• galactosylceramide catabolism	IEA	GOA	
• hydrolase activity, acting on glycosyl bonds	IEA	GOA	
• lysosome	TAS	GOA	pm

Figure A.2: Example of LocusLink database for a gene related to Krabb's disease. A summary for the gene product, along with aliases and GO annotations is provided. GeneRIF provides not only links to seminal MEDLINE papers (e.g., 11814461), but also (manually) extracted text snippets.

Appendix B

Supplementary text profiles

Table B.1: TXTGate profiling of yeast expression clusters *B,C,D* in Eisen *et al.* For each gene cluster we display the top 10 terms extracted from associated MEDLINE abstracts, as well as terms from the GO annotation field in the SGD database. We compare our MEDLINE profile with the analysis in Blaschke *et al.* and indicate corresponding terms in **bold**. Terms marked with '(-)' are considered noisy by Blaschke and collegues and should therefore be ignored.

Cluster ID			MEDLINE-based profiles	GO-based profiles	Terms in Blaschke *et al.*	Biological description
CLUSTER *B* Spindle pole assembly and function						
stu2	cdc16		**spindl_pole**_bodi	cytoskeleton	**spindles**	Control of mitosis from S phase to cytokinesis
dhs1			**cyclin**	cyclin	**cytokinesis**	
bnr1			bud	axial_bud	temperature sensitive(-)	
spc42			mitosi	microtubul_nucleat	**anaphase promoting complex**	
cnm67			**cytokinesi**	protein_bind	**b-type cyclins**	
clb4			septin	g2		
cdc10			**spindl**	bud_neck		
cdc3			cell_cycl	division		
clb3			**anaphas_promot_complex**	degrad		
apc4			anaphas	g1_transition		
CLUSTER *C* Proteasome						
rpn11	pup2	pre7	proteasom	degrad	**26s proteasome**	Proteasome, protein degradation, peroxisome and transport across membrane
ufd1	scl1	rpn10	**26s_proteasom**	endopeptidas	ubiquitinated	
rpn9	pre5	rpt3	regulatori	cytoplasm	**degradation**	
rpt1	pre9	prt5	**degrad**	nuclear	**proteolytic**	
rpt6	pre1	rpn12	**ubiquitin**		peroxisomal targeting(-)	
pre4	pre2	rpn5	particl			
rpn6	pre3	rpn8	19s			
rpt4	pre10		**proteolysi**			
rpn7	pup1		core			
rpn3	pre6		similar			
CLUSTER *D* mRNA splicing						
pop6	std1		mitochondri	mrna_splice	transcription factor	mRNA maturation, transcription regulation and the cytochrome b
caf16	mgm1		rna	nuclear	zinc finger	
mrs1	prp19		splice	transport	cytochrome b	
smd3	sls1		mrna	mrna_process	**pre-mRNA splicing**	
taf40			interact	rrna_process	RNA polymerase	
pet309			nuclear	snrnp_recycl		
pzf1			transcript	gtpase		
ysh1			u6	pre_mrna_splice_factor		
cbp1			u4	mitochondri_genom_mainten		
prp24			*pre_mrna_splice*	integr_membran_protein		

Table B.2: TXTGate profiling of yeast expression clusters *E,F,G* in Eisen *et al.*. For each gene cluster we display the top 10 terms extracted from associated MEDLINE abstracts, as well as terms from the GO annotation field in the SGD database. We compare our MEDLINE profile with the analysis in Blaschke *et al.* and indicate corresponding terms in **bold**. Terms marked with '(-)' are considered noisy by Blaschke and collegues and should therefore be ignored. Terms marked in italic stem from top 15 (instead of top 10) results.

Cluster ID			MEDLINE-based profiles	GO-based profiles	Terms in Blaschke *et al.*	Biological description
CLUSTER *E* Glycolysis						
tpt1	pdc6		glyceraldehyd_3_phosph...	gluconeogenesi	metabolism	Glycolysis
gpm1	pdc1		**glycolyt**	glycolysi	**glycolytic**	
pgk1	cdc19		**pyruv_decarboxylas**	cytoplasm	hexokinases	
tdh3	hxk2		glucos	glyceraldehyd_3_phosph...	**pyruvate decarboxylase**	
tdh2	tye7		enzym	phosphoryl	glycerol-3-phospate dehydrogenase	
eno2	pfk1		ethanol	transketolas		
tdh1	acs2		carbon	hexokinas		
fba1			glycolysi	phosphoglycer_kinas		
tkl1			ferment	phosphoglycer_mutas		
pdc5			pyruv_kinas	pyruv_kinas		
CLUSTER *F* Mitochondrial ribosome						
dbi56	mrp7	ppa2	**mitochondri_ribosom**	mitochondri	**mitochondrial ribosomal**	Cytochromes and mitochondrial synthesis of proteins
mss51	none	img2	mitochondri	biosynthesi	cytochrome oxidase	
msf1	cox17		mrp	mitochondri_matrix	**cytochrome c**	
mrps5	mrp17		protein_biosynthesi	protein_biosynthesi	aminoacyl-tRNA synthetases	
msm1	mrpl6		commun	peptidyltransferas	**translational activator**	
cox10	mrpl8		cytoplasm	ribosom		
msk1	rml2		**translat**	respir		
mef1	mrpl13		individu	mrna_process		
none	mrpl9		ribosom	mitochondri_intermembran_space		
mrpl35	pet123		*cytochrom*	cytochrom_c_oxidas_biogenesi		
CLUSTER *G* ATP synthesis						
sds22	atp14		**synthas**	atpas	**ATP synthase**	ATP syntesis in mitochondria
cox4	atp3		**atp**	transport	mitochondrial f1-atpase	
cox13	atp16		mitochondri	cytochrom_c_oxidas	**oxidative phosphorylation**	
cox8	atp7		assembli	mitochondri	**cytochrome oxidase**	
atp1	qcr8		**cytochrom_c_oxidas**	mitochondri_inner_membran		
atp5			**oxid_phosphoryl**	nuclear		
atp2			atpas	ubiquinol_cytochrom_c_reductas		
cox12			f			
atp4			growth			
atp17			aerob			

Table B.3: TXTGate profiling of yeast expression clusters *H,I,K* in Eisen *et al.*. For each gene cluster we display the top 10 terms extracted from associated MEDLINE abstracts, as well as terms from the GO annotation field in the SGD database. We compare our MEDLINE profile with the analysis in Blaschke *et al.* and indicate corresponding terms in **bold**. Terms marked with '(-)' are considered noisy by Blaschke and collegues and should therefore be ignored. Terms marked in italic stem from top 15 (instead of top 10) results.

Cluster ID	MEDLINE-based profiles	GO-based profiles	Terms in Blaschke *et al.*	Biological description
CLUSTER *H* Chromatin structure				
hht1 hht2 hhf2 hhf1 htb1 hta2 hta1 htb2	**histon** h4 h2a h3 h2b **nucleosom** **chromatin** core tail transcript	nuclear disassembli chromatin_assembli dna_bind	**histone proteins** **chromatin structure** **nucleosome assembly**	Chromatine structure, histones
CLUSTER *J* DNA replication				
cdc54 mcm3 mcm2 cdc47 dbf2	mcm **dna_replic** **replic** initi **origin** cell_cycl phase mainten *licens* *origin_recognit_complex*	atp_depend_dna_helicas dna_replic_initi cell_cycl nuclear	**licensing factor** **DNA replication** S phase **replication origins** minichromosome maintenance	Control of DNA replication
CLUSTER *K* Tricarboxylic acid cycle and respiration				
por1 sdh4 none sdh2 sdh1 mdh1 ndi1 qcr6 qcr7 cyt1 cox6 ach1 rip1 cox15 cor1 sdh3	cytochrom succin_dehydrogenas mitochondri carbon **respiratori** oxid reductas respir cytochrom_bc assembli	mitochondri_inner_membran ubiquinol_cytochrom_c_reductas ubiquinon succin_dehydrogenas_complex succin cytochrom_c_oxidas malic_enzym mitochondri_outer_membran cytochrom_c_oxidas_biogenesi acetyl_coa_hydrolas	voltage dependent(-) outer membrane(-) pores(-) channel(-) succinate malate dehydrogenase **respiratory chain**	Respiration and membrane channels (voltage dependent)

Table B.4: TXTGate profiling of biclusters *a*, *b*, *c* and *d* from Chaussabel *et al.* (GO vocabulary). We compare the terms of Chaussabel *et al.* with those from TXTGate and mark correspondances in bold. Compared with the annotations of Chaussabel *et al.*, we observe that TXTGate's profiles are comparably informative.

Gene Symbol		Bicluster Terms in Chaussabel *et al.*		Terms from TXTGate
Cluster a				
LPL		Lipoprotein		**lipoprotein**
CD36L1		Density		lipas
LDLR		Cholesterol		ldl
		Lipid		ldl_receptor
		Adipose		**cholesterol**
				hdl
				scaveng_receptor
				high_densiti_lipoprotein
				low_densiti_lipoprotein_receptor
				low_densiti_lipoprotein
Cluster b				
UPA		Invasive	Collagenase	**metalloproteinas**
PLAUR		Invasion	Collagen	**matrix**
SERPIN		Metastasis	Matrix	metalloendopeptidas
MMP1		UPAR	MMP	**collagenas**
MMP10		UPA	Metalloproteinase	extracellular_matrix
MMP14		Plasminogen	Molecule-1	alpha
SPARC		Urokinase-type	Adhesion	**upar**
		Urokinase	Vascular	**plasminogen_activ**
		Plasmin	Endothelial	interstiti
		Activator		**invasion**
Cluster c				
AMPD3		Adenosine		purinerg
ADA		A2A		**adenosin**
ADORA2A		A1		deaminas
ADORA3		Antagonist		p2
P2RX		Agonist		p2x
P2RX1		NM		p1
P2RX7				**agonist**
				receptor
				adenosin_receptor
				ada
Cluster d				
IP10	FGL2	Interferon		tumor_necrosi_factor
MIP1A	TNFRSF1A	IFN-alpha		cytokin
MIP1B	TNF	IFN		induc
IL8	CD14	Interferon-gamma		**interferon**
STAT4	PTX3	IFN-gamma		inflammatori
IL12B	IL6	Inducible		antigen
TNFRSF9	MX1			lymphocyt_activ
TNFSF9	GBP1			stimul
SLAM	ISG15			chemokin
TNFRSF5	IRF1			monocyt
CD83	INDO			

Table B.5: Text Profiles generated with TXTGate for clusters 3, 6 and 9 resulting from clustering yeast cell-cycle data. We show both MEDLINE and GO profiles of terms with highest variance. **A.** Cluster 3 (201 genes) contains genes peaking in G1/S phase of the cell cycle where DNA replication, synthesis and repair takes place, indicated by terms such as 'mitmatch repair' and 'DNA replication'. Cyclins activate kinases that control transition to, for example, the M(itosis) phase. Some genes active in the early M phase are also included in this cluster, yielding terms as 'spindle pole body' and 'microtubuli'. **B.** Cluster 6 (78 genes) groups genes active in the M phase and the transition to the G1 phase. Indeed, microtubules pull each sister chromatide via the kinetochores to the spindle poles. This mainly happens in a subphase of the M phase termed the anaphase. **C.** Cluster 9 (117 genes) contains a diversity of genes involved in energy production (succinate dehydrogenase, mitochondrial), protein degradation (peroxisome) and cytokinesis (i.e., final stage of mitosis).

A		B		C	
Cluster 3	201 genes	Cluster 6	78 genes	Cluster 9	117 genes
MEDLINE	GO	MEDLINE	GO	MEDLINE	GO
cyclin	nuclear	spindl	nuclear	mitosi	integr_membran_protein
mismatch_repair	integr_membran_protein	spindl_pole_bodi	microtubul_nucleat	succin_dehydrogenas	axial_bud
spindl_pole_bodi	dna_repair	kinetochor	integr_membran_protein	arp2	biosynthesi
dna_replic	cell_cycl	mcm	cell_polar	cytokinesi	protein_kinas
replic	transcript	motor	cytoplasm	hsp90	cytoplasm
g1	axial_bud	microtubul	protein_kinas	peroxisom	nuclear
checkpoint	dna_replic_initi	anaphas	ubiquitin_protein_ligas	bud	biogenesi
cell_cycl	protein_kinas	telomer	cytoskeleton	cyclin	protein_fold
spindl	cytoskeleton	dna_replic	mate	hydroxymethylpyrimidin	heat_shock_protein
sister_chromatid_cohesion	dna_recombin	h1	microtubul_motor	cell_wall	cell_cycl
repair	transport	isomeras	interact_protein	inositol	cytokinesi
bud	microtubul_nucleat	septin	secret	myosin	cytoskeleton
cohesin	meiosi	bud	endocytosi	transcript_factor	mitochondri
microtubul	mrna_splice	kinesin	glutathion_peroxidas	photolyas	meiosi
mitosi	g1_transition	replic	histon_deacetylas	cell_cycl	cyclin
cohesion	cyclin	rub1	meiotic_recombin	singlet_oxygen	interact_protein
telomer	mitot_cell_cycl	polar	meiosi	pyrophosphorylas	polar_bud
cell_wall	microtubul_base_process	mitosi	respir	mitochondri	v_snare
histon	protein_fold	hml	mitot_spindl_checkpoint	6_phosphofructo_2_kinas	glycosyl
chitin_synthas	type_ii	cell_polar	ubiquitin_conjug_enzym	mitot	uridylyltransferas

Bibliography

[1] [http://www.genomenewsnetwork.org/sequenced_genomes/genome_guide_index.shtml].

[2] Amigo Gene Ontology browser. [http://www.godatabase.org].

[3] Applications of Gene Ontology. [http://www.geneontology.org/GO.biblio.html].

[4] Cell cycle tutorial. [http://www.biology.arizona.edu/cell_bio/tutorials/cell_cycle/].

[5] dbEST. [http://www.ncbi.nlm.nih.gov/dbEST/].

[6] eVoc. [http://www.sanbi.ac.za/evoc/ontologies.html].

[7] G2D. [http://www.bork.embl-heidelberg.de/g2d/].

[8] GEISHA. [http://www.pdg.cnb.uam.es/blaschke/cgi-bin/geisha].

[9] The Gene Ontology. [http://www.geneontology.org].

[10] GeneCards. [http://bioinformatics.weizmann.ac.il/cards/].

[11] Genexpress. [http://genexpress.stanford.edu/].

[12] The GOA project. [http://www.ebi.ac.uk/GOA/].

[13] GOMINER. [http://discover.nci.nih.gov/gominer/].

[14] Ihmels *et al.* webpage. [http://www.weizmann.ac.il/home/jan/NG/MainFrames.html].

[15] INCLUSive. [http://www.esat.kuleuven.ac.be/inclusive/index.jsp].

[16] IUPAC. [http://www.chem.qmul.ac.uk/iupac/misc/naseq.html].

[17] LocusLink. [http://www.ncbi.nlm.nih.gov/LocusLink/].

[18] MedGene. [http://hipseq.med.harvard.edu/MEDGENE/].

[19] MedMiner. [http://discover.nci.nih.gov/textmining/].

[20] MedMOLE. [http://www.cineca.it/HPSystems/Chimica/medmole/].

[21] MeKE. [http://ismp.csie.ncku.edu.tw/ yuhc/meke/].

[22] MeSH. [http://www.nlm.nih.gov/mesh/meshhome.html].

[23] MIPS. [http://mips.gsf.de/funcat].

[24] Missing value estimation. [ftp://ftp.esat.kuleuven.ac.be/pub/sista/desmet/missing.pdf].

[25] Nist e-handbook of statistical methods. [http://www.itl.nist.gov/div898/handbook/].

[26] Online Mendelian Inheritance in Man. [http://www.ncbi.nlm.nih.gov/omim/].

[27] PubGene. [http://www.pubgene.org].

[28] PUBMED. [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed].

[29] Saccharomyces Genome Database. [http://www.yeastgenome.org/].

[30] SWISS PROT. [http://us.expasy.org/sprot/].

[31] Text mining and link analyis. [http://www-2.cs.cmu.edu/ dunja/TextLink2003/].

[32] TRANSFAC. [http://www.gene-regulation.com/].

[33] TXTGate. [http://www.esat.kuleuven.ac.be/txtgate].

[34] UMLS. [http://www.nlm.nih.gov/research/umls/].

[35] S. Aerts. *Computational discovery of* cis-regulatory *modules in animal genomes,* PhD thesis. Faculty of Engineering, K.U.Leuven (Leuven, Belgium), 2004.

[36] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science Publishing, fourth edition, 2002.

[37] A. Alizadeh, M. Eisen, R.E. Davis, C. Ma, H. Sabet, T. Tran, J.I. Powell, L. Yang, G.E. Marti, D.T. Moore, J.r. Hudson JR, W.C. Chan, T. Greiner, D. Weisenburger, J.O. Armitage, I. Lossos, R. Levy, D. Botstein, P.O. Brown, and L.M. Staudt. The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. *Cold Spring Harb Symp Quant Biol*, 64:71–8, 1999.

[38] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. 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, 1999.

[39] R. Altman. Challenges for intelligent systems in biology. *IEEE Intelligent Syst, Special Issue on Intelligent Syst in Biol*, April, 2002.

[40] P. Antal, P. Glenisson, G. Fannes, J. Mathys, B. De Moor B, and Y. Moreau. On the potential of domain literature for clustering and bayesian network learning. In *Proc of the 8th ACM-SIGKDD Int Conf on Knowl Discovery and Data mining (ACM-SIGKDD), Edmonton, Canada*, pages 405–414, 2002.

[41] R.A. Baeza-Yates and B.A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[42] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51–80, 1995.

[43] Z. Bar-Joseph, D.K. Gifford, and T.S. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17:S22–9, 2001.

[44] A.D. Baxevanis. The molecular biology database collection: 2003 update. *Nucleic Acids Res*, 31:1–12, 2003.

[45] K.G. Becker, D.A. Hosack, G. Dennis, R.A. Lempicki, T.J. Bright, C. Cheadle, and J. Engel. PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*, 4:61, 2003.

[46] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. *Pac Symp Biocomput*, pages 6–17, 2002.

[47] S. Bergmann, J. Ihmels, and N. Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*, 2:E9, 2004.

[48] M. Berry and M. Browne. *Understanding Search Engines: mathematical modeling and text retrieval*. SIAM Press, 1999.

[49] M. Berry, S.T. Dumais, and G.W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37:573–595, 1995.

[50] L.T. Bhoite, J.M Allen, E. Garcia, L.R. Thomas, I.D. Gregory ID, W.P. Voth WP, K. Whelihan, R.J. Rolfes, and D.J. Stillman. Mutations in the PHO2 (bas2) transcription factor that differentially affect activation with its partner proteins BAS1, PHO4, and SWI5. *J Biol Chem.*, 277:37612–37618, 2002.

[51] D.R. Bickel. Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically. *Bioinformatics*, 19:818–824, 2003.

[52] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.

[53] C. Blaschke, J.C. Oliveros, and A. Valencia. Mining functional information associated with expression arrays. *Funct Integr Genomics*, 1:256–268, 2001.

[54] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C. Holstege, I.F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME): towards standards for microarray data. *Nat Genet*, 29:365–371, 2001.

[55] A. Brazma and J. Vilo. Review: Gene expression data analysis. *FEBS Letters*, 23893:1–8, 2000.

[56] H.J. Bussemaker, H. Li, and E.D. Siggia. Regulatory element detection using correlation with expression. *Nat Genet*, 27:167–171, 2001.

[57] R.A. Calogero, G. Iazzetti, S. Motta, G. Pedrazzi, S. Rago, E. Rossi, and R. Turra. MedMOLE: Mining literature to extract biological knowledge by microarray data. In *Proc Virt Conf Genom and Bioinf*, North Dakota, USA, 2001.

[58] L.R. Cardon and G.D. Stormo. Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J Mol Biol*, 223:159–170, 1992.

[59] I.M. Chakravart, R.G. Laha, and J. Roy. *Handbook of Methods of Applied Statistics, volume I.* John Wiley and Sons, 1967.

[60] D. Chaussabel and A. Sher. Mining microarray expression data by literature profiling. *Genome Biol*, 3(10):1–16, 2002.

[61] J.H Chiang and H.C. Yu. MeKE: discovering the functions of gene products from biomedical literature viasentence alignment. *Bioinformatics*, 19:1417–1422, 2003.

[62] R.J. Cho, M.J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, 2:65–73, 1998.

[63] S . Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705, 1998.

[64] B. Coessens, G. Thijs, S. Aerts, K. Marchal, F. De Smet, K. Engelen, P. Glenisson, Y. Moreau, J. Mathys, and B. De Moor. INCLUSive - a web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Res*, 31:3468–3470, 2003.

[65] E.M. Conlon, X.S Liu, J.D. Lieb, and J.S. Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci USA*, 100:3339–3344, 2003.

[66] The International Human Genome Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

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

[68] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the Am Soc for Information Science*, 41:391–407, 1990.

[69] D. Delneri, D.C. Gardner, and S.G. Oliver. Analysis of the seven-member AAD gene set demonstrates that genetic redundancy in yeast may be more apparent than real. *Genetics*, 153:1591–600, 1999.

[70] J.L. DeRisi, V.R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

[71] Editorial. Free and public expression. *Nature*, 410:851, 2001.

[72] Editorial. Coming to terms with microarrays. *Nat Genet*, 32:333–334, 2002.

[73] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95:14863–14868, 1998.

[74] A.J. Enright and C.A. Ouzounis. BioLayout–an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, 17(9):853–4, 2001.

[75] E. Eskin and E. Agichtein. Combining text mining and sequence analysis to discover protein functional regions. *Pac Symp Biocomput*, pages 288–99, 2004.

[76] T.L. Ferea, D. Botstein, P.O. Brown, and R.F. Rosenzweig. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc Natl Acad Sci USA*, 96:9721–9726, 1999.

[77] A.P. Gasch and M.B. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol*, 3:1–22, 2002.

[78] M. Gerstein and J. Junker. Blurring the boundaries between scientific papers and biological databases. *Nature Online, http://www.nature.com/nature/debates/e-access/Articles/gernstein.html*, web debate, on-line 7 May 2001.

[79] F.D. Gibbons and F.P. Roth. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res*, 12:1574–1581, 2002.

[80] P. Glenisson, P. Antal, J. Mathys, Y. Moreau, and B. De Moor. Evaluation of the vector space representation in text-based gene clustering. *Pac Symp Biocomput*, pages 391–402, 2003.

[81] P. Glenisson, B. Coessens, S. Van Vooren, J. Mathys, Y. Moreau, and B. De Moor. TXTGate : Profiling gene groups with text-based information. *Genome Biol*, 2004. In press.

[82] P. Glenisson, B. Coessens, S. Van Vooren, Y. Moreau, and B. De Moor. Text-based gene profiling with domain-specific views. In *Proc. of the First Int Workshop on Semantic Web and Databases (SWDB 2003)*, pages 15–31, 2003.

[83] P. Glenisson, J. Mathys, Y. Moreau, and B. De Moor. Meta-clustering of gene expression data and literature-extracted information. *SIGKDD Explorations, Special Issue on Microarray Data Mining*, 5(2):101–112, 2003.

[84] G. Golub and C. Van Loan. *Matrix Computations*. John Hopkins University Press, third edition, 1996.

[85] D. Hanisch, J. Fluck, H.T. Mevissen, and R. Zimmer. Playing biology's name game: identifying protein names in scientific text. *Pac Symp Biocomput*, pages 403–414, 2003.

[86] J. Hansen and P.F. Johannesen. Cysteine is essential for transcriptional regulation of the sulfur assimilation genes in *Saccharomyces cerevisiae*. *Mol Gen Genet*, 263:535–542, 2000.

[87] L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray. From molecular to modular cell biology. *Nature*, 402:C47–52, 1999.

[88] L.V. Hedges and I. Olkin. *Statistical methods for Meta-Analysis*. Academic Press, 1985.

[89] J. Herrero, A. Valencia, and J. Dopazo. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17:126–136, 2001.

[90] R. Hoffmann and A. Valencia. Life cycles of successful genes. *Trends Genet*, 19:79–81, 2003.

[91] R. Hoffmann and A. Valencia. Protein interaction: same network, different hubs. *Trends Genet*, 19:681–3, 2003.

[92] K. Horimoto and H. Toh. Statistical estimation of cluster boundaries in gene expression profile data. *Bioinformatics*, 17:1143–51, 2001.

[93] Y. Hu, L.M. Hines, H. Weng, D. Zuo, M. Rivera, A. Richardson, and J. LaBaer. Analysis of genomic and proteomic data using advanced literature mining. *J Proteome Res*, 2:405–412, 2003.

[94] J.D. Hughes, P.W. Estep, S. Tavazoie, and G.M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *saccharomyces cerevisiae*. *J Mol Biol*, 296:1205–1214, 2000.

[95] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nat Genet*, 31:370–377, 2002.

[96] A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice Hall, 1988.

[97] T.K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28:21–28, 2001.

[98] J.D. Jobson. *Applied Multivariate Data Analysis, v.2. Categorical and multivariate methods*. Springer-Verlag, 1992.

[99] G. Karp. *Cell and molecular biology*. John Wiley and Sons, 2002.

[100] L. Kaufman and P. Rousseeuw. *Finding groups in data*. Wiley-Interscience, 1990.

[101] J. Kelso, J. Visagie, G. Theiler, A. Christoffels, S. Bardien, D. Smedley, D. Otgaar, G. Greyling, C.V. Jongeneel, M.I. McCarthy, T. Hide, and W. Hide. eVOC: A controlled vocabulary for unifying gene expression data. *Genome Res*, 13:1222–1230, 2003.

[102] G.R. Lanckriet, M. Deng, N. Cristianini, M.I. Jordan MI, and W.S. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. *Pac Symp Biocomput*, 18:300–11, 2004.

[103] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J.-B. Tagne, T.L. Volkert, E. Fraenkel, D.K. Gifford, and R.A. Young. Transcriptional regulatory networks in *saccharomyces cerevisiae*. *Science*, 298:799–804, 2002.

[104] J.E. Leonard, J.B. Colombe, and J.L. Levy. Finding relevant references to genes and proteins in MEDLINE using a bayesian approach. *Bioinformatics*, 18:1515–1522, 2002.

[105] D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *Proc of the 15th Int ACM SIGIR Conf on Research and Development in Information Retrieval*, pages 37–50, 1992.

[106] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Semantic similarity measures as tools for exploring the Gene Ontology. *Pac Symp Biocomput*, pages 601–612, 2003.

[107] R. Lowry. Concepts and applications of inferential statistics. 2003. [http://faculty.vassar.edu/lowry/VassarStats.html].

[108] K. Marchal, K. Engelen, J. De Brabanter, S. Aerts, and B. De Moor. Comparison of different methodologies to identify differentially expressed genes in two-sample cDNA arrays. *J Biol Sys*, 10:409–430, 2002.

[109] L. Marsan and M.F. Sagot. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J Comput Biol*, 7:345–362, 2000.

[110] D.R. Masys. Linking microarray data to the literature. *Nat Genet*, 28:9–10, 2001.

[111] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A.E. Kel, O.V. Kel-Margoulis, D.U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31:374–378, 2003.

[112] M.T. McCammon. Mutants of *Saccharomyces cerevisiae* with defects in acetate metabolism: isolation and characterization of Acn-mutants. *Genetics*, 144:57–69, 1996.

[113] V.A. McKusick. *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*. Johns Hopkins University Press, twelfth edition, 1998.

[114] H.W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen, J. Warfsmann, and A. Ruepp. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*, 32 Database issue:D41–4, 2004.

[115] D.S. Modja. Clustering hypertext with applications to web searching. In *Proc. ACM Hypertext Conf*, 2000.

[116] B. De Moor, K. Marchal, J. Mathys, and Y. Moreau. Bioinformatics : Organisms from Venus, technology from Jupiter, algorithms from Mars. *European Journal of Control*, 9:237–278, 2003.

[117] Y. Moreau, S. Aerts, B. De Moor, B. De Strooper, and M. Dabrowski. Comparison and meta-analysis of microarray data : from the bench to the computer desk. *Trends in Genetics*, 19:570–577, 2003.

[118] G.J. Nau, J.F. Richmond, A. Schlesinger, E.G. Jennings, E.S. Lander, and R.A. Young. Human macrophage activation programs induced by bacterial pathogens. *Proc Natl Acad Sci USA*, 99:1503–1508, 2002.

[119] B.I. Ono, T. Hazu, S. Yoshida, T. Kawato, S. Shinoda, J. Brzvwczy, and A. Paszewski. Cysteine biosynthesis in *Saccharomyces cerevisiae*: a new outlook on pathway and regulation. *Yeast*, 15:1365–1375, 1999.

[120] K.M. Overkamp, B.M. Bakker, P. Kotter, M.A. Luttik, J.P. Van Dijken, and J.T. Pronk. Metabolic engineering of glycerol production in *Saccharomyces cerevisiae*. *Appl Environ Microbiol*, 68:2814–2821, 2002.

[121] P. Pavlidis, D.P. Lewis, and W.S. Noble. Exploring gene expression data with class scores. *Pac Symp Biocomput*, pages 474–485, 2002.

[122] D. Pearce. A comparative evaluation of collocation extraction techniques. In *Third Int Conf on Language Resources and Evaluation*, 2002.

[123] C. Perez-Iratxeta, P. Bork, and M.A. Andrade. Association of genes to genetically inherited diseases using data mining. *Nat Genet*, 31:316–319, 2002.

[124] K.S. Pollard and M.J. van der Laan. A method to identify significant clusters in gene expression data. In *In Proc of Systemics, Cybernetics and Informatics 2002 (SCI 2002)*, 2002.

[125] K.S. Pollard and M.J. van der Laan. Statistical inference for simultaneous clustering of gene expression data. *Math Biosci*, 176:99–121, 2002.

[126] M.F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.

[127] K.D. Pruitt and D.R Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*, 29:137–140, 2001.

[128] J. Quackenbush. Microarray data normalization and transformation. *Nat Genet*, 32:496–501, 2002.

[129] S. Raychaudhuri, J.T. Chang, F. Imam, and R.B. Altman. The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res*, 31:4553–4560, 2003.

[130] S. Raychaudhuri, J.T. Chang, P.D. Sutphin, and R.B. Altman. Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res*, 12:203–214, 2002.

[131] S. Raychaudhuri, H. Schutze, and R.B. Altman. Using text analysis to identify functionally coherent gene groups. *Genome Res*, 12:1582–1590, 2002.

[132] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. GeneCards: A novel functional genomics compendium withautomated data mining and query reformulation support. *Bioinformatics*, 14:656–664, 1998.

[133] J.R. Rohde and M.E. Cardenas. The tor pathway regulates gene expression by linking nutrient sensing to histone acetylation. *Mol Cell Biol*, 23:629–635, 2003.

[134] J. Schuchhardt, D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach, and H. Herzel. Normalization strategies for cDNA microarrays. *Nucleic Acids Res*, 28(10):E47, 2000.

[135] E. Segal, M. Shapira, A. Rev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34:166–176, 2003.

[136] H. Shatkay, S. Edwards, M., and Boguski. Information retrieval meets gene analysis. *IEEE Intelligent Syst, Special Issue on Intelligent Syst in Biol*, April, 2002.

[137] H. Shatkay, S. Edwards, W.J. Wilbur, and M. Boguski. Genes, themes and microarrays: Using information retrieval for large-scale gene analysis. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, Menlo Park, CA, USA*, pages 317–328. AAAI, 2000.

[138] S.B. Singh, R.D. Hull, and E.M. Fluder. Text influenced molecular indexing (timi): a literature database mining approach that handles text and chemistry. *J Chem Inf Comput Sci*, 43:743–52, 2003.

[139] F. De Smet. *Microarrays: algorithms for knowledge discovery in oncology and molecular biology,* PhD thesis. Faculty of Engineering, K.U.Leuven (Leuven, Belgium), 2004.

[140] P.T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W.L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B.J. Aronow, A. Robinson, D. Bassett, J.r. Stoeckert CJ, and A. Brazma. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*, 3(9):R0046, 2002.

[141] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*, 9:3273–3297, 1998.

[142] M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, and J. Mostafa. Detecting gene relations from Medline abstracts. *Pac Symp Biocomput*, pages 483–495, 2001.

[143] J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302:249–255, 2003.

[144] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific Publishing Co., Pte, Ltd. (Singapore), 2002.

[145] S.H. Sze, M.S. Gelfand, and P.A. Pevzner. Finding weak motifs in DNA sequences. *Pac Symp Biocomput*, pages 235–246, 2002.

[146] J. H. Friedman T. Hastie, R. Tibshirani. *The Elements of Statistical Learning*. Springer Verlag, 2001.

[147] L. Tanabe, U. Scherf, L.H. Smith, J.K. Lee, L. Hunter, and J.N. Weinstein. MedMiner: An internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques*, 27:1210–1217, 1999.

[148] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nat Genet*, 22:281–285, 1999.

[149] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat Genet*, 25:25–29, 2000.

[150] G. Thijs. *Probabilistic methods to search for regulatory elements in sets of coregulated genes,* PhD thesis. Faculty of Engineering, K.U.Leuven (Leuven, Belgium), 2003.

[151] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouzé, and Y. Moreau. A higher order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17:1113–1122, 2001.

[152] G. Thijs, K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouzé, and Y. Moreau. A Gibbs sampling meyhod to detect over-represented motifs in the upstream regions of co-expressed genes. In T. Lengauer, D. Sankoff, S. Istrail, P. Pevzner, and M.S. Waterman, editors, *Fifth Annual International Conference on Computational Biology*, volume 5, pages 305–312. ACM press, 2001.

[153] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistics. Technical Report 208, Stanford University, USA, 2000.

[154] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–5, 2001.

[155] Tumor Analysis Best Practices Working Group. Expression profiling–best practices for data generation and interpretation in clinical trials. *Nat Rev Genet*, 5(3):229–37, 2004.

[156] J. van Helden, B. André, and L. Collado-Vides. Extracting regulatory sites from upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, 281:827–842, 1998.

[157] J. van Helden, A.F. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, 28(8):1808–1818, 2000.

[158] J.C. Venter and colleagues. The sequence of the human genome. *Science*, 291:1304–51, 2001.

[159] M. Vidal. A biological atlas of functional maps. *Cell*, 104:333–339, 2001.

[160] J. Vilo, A. Brazma, I. Jonassen, A. Robinson, and E. Ukkonen. Mining for putative regulatory elements in the yeast genome using gene expression data. *Proc Int Conf Intell Syst Mol Biol*, 8:384–94, 2000.

[161] G.A. Wray, M.W. Hahn, E. Abouheif, J.P. Balhoff, M. Pizer, M.V. Rockman, and L.A. Romano. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, 20(9):1377–419, 2003.

[162] L.F. Wu, T.R. Hughes, A.P. Davierwala, M.D. Robinson, R. Stoughton, and S.J. Altschuler. Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat Genet*, 31:255–265, 2002.

[163] Y. Yamanishi, J.P. Vert, A. Nakaya, and M. Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19:I323–I330, 2003.

[164] Y.H. Yang and T. Speed. Design issues for cDNA microarray experiments. *Nat Rev Genet*, 3:579–588, 2002.

[165] K.Y. Yeung, D.R. Haynor, and W.L. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17:309–318, 2001.

[166] K.Y. Yeung and W.L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17:763–774, 2001.

[167] Y.L. Yuan and S. Fields. Properties of the DNA-binding domain of the *Saccharomyces cerevisiae* STE12 protein. *Mol Cell Biol*, 11:5910–5918, 1991.

[168] B.R. Zeeberg, W. Feng, G. Wang, M.D. Wang, A.T. Fojo, M. Sunshine, S. Narasimhan, D.W. Kane, W.C. Reinhold, S. Lababidi, K.J. Bussey, J. Riss, J.C. Barrett, and J.N. Weinstein. GOMINER: A resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28, 2003.

[169] A. Zien, R. Kuffner, R. Zimmer, and T. Lengauer. Analysis of gene expression data with pathway scores. *Proc Int Conf Intell Syst Mol Biol*, 8:407–17, 2000.

Curriculum vitae

Patrick Glenisson was born in Antwerp on November 26th, 1975. He obtained a Master of Science in Physics (Licentiaat Natuurkunde) in 1997 and a Master in Artificial Intelligence in 1999 at the Katholieke Universiteit Leuven, Belgium. In August 1999 he joined the ESAT-SCD bioinformatics group at the Department of Electrical Engineering at the same university. He was a research assistant at the KULeuven from 1999-2000 and from 2001-2004. In 2000-2001 he was affiliated with the VIB (Vlaams Interuniversitair Instituut voor Biotechnologie) through their core Microarray Facility (VIB-MAF). He was a member of the steering committee of the international MGED (Microarray Gene Expression Data) Society in 2000-2001.