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

Reconstruction of regulatory modules based on heterogeneous data sources

Jury:
Prof. dr. ir. P. Van Houtte, voorzitter
Prof. dr. ir. B. De Moor, promotor
Prof. dr. ir. K. Marchal, promotor
Prof. dr. ir. Y. Moreau
Prof. dr. ir. J. Vanderleyden
Prof. dr. ir. T. De Bie (Bristol, UK)
Prof. dr. J. Collado-Vides (UNAM, Mexico)

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Karen LEMMENS

UDC 681.3*J3 September 2008

© Katholieke Universiteit Leuven – Faculteit Ingenieurswetenschappen
Arenbergkasteel, Kasteelpark Arenberg 1, 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/2008/7515/96

ISBN 978-90-5682-986-5

Voorwoord

Na vier jaren in de ESAT toren ben ik nu op het punt gekomen dat ik bijna een doctoraat op zak heb. Dit was nooit mogelijk geweest zonder de steun en hulp van vele mensen.

Ik zou de leden van mijn jury willen bedanken voor hun waardevolle suggesties en opmerkingen die deze tekst aanzienlijk verbeterd hebben.

Verder wil ik ook professor Bart De Moor bedanken voor de kansen die hij mij als thesisstudent en doctoraatsstudent geboden heeft, maar ook voor zijn steun en vertrouwen in mijn onderzoek.

De persoon die ik het meeste dank verschuldigd ben jij natuurlijk, Kathleen. Het is allemaal begonnen toen je mij tijdens mijn thesis leerde dat er meer aan een computer is dan de aan/uit knop. Tijdens dat jaar heb je mij met veel gedrevenheid en enthousiasme ingeleid in de bioinformatica en dat werkte aanstekelijk want nu, bijna 5 jaar later, doe ik nog steeds even graag bioinformatica. De voorbije jaren heb ik ontzettend veel van je geleerd, en stond je altijd klaar voor mij. Maar naast werk was er ook altijd tijd voor wat plezier en ontspanning, en dat ging van een spelletje weerwolven, bochtjes (heeeel veel) maken op de skipistes, salsa dansen met Indiërs tot hoe te overleven tussen die pinches alacranes tijdens de reisjes in Mexico (of was dat werk?). Kathleen, superhard bedankt voor alles!

Naast mijn promotoren wil ik ook professor Tijl De Bie speciaal bedanken. Tijl, als geen ander slaag jij erin inzicht te krijgen in onze verwarde biologische vragen en problemen en kan je die moeiteloos omzetten naar keiharde Java code. Ik ben heel blij dat de Berbers in Marokko je af en toe eens de weg getoond hebben, want zonder jou zou er nu geen doctoraat zijn. Heel erg bedankt voor de fijne samenwerking!

Tijdens dit doctoraat heb ik de kans gekregen om met heel wat uiteenlopende onderzoeksgroepen samen te werken. Veel dank aan Álvaro Cortés Calabuig en prof. Denecker van de afdeling Informatica (departement computerwetenschappen) om de confrontatie met biologische problemen aan te gaan en hun gedrevenheid om oplossingen te zoeken. Daarnaast heb ik ook kennis gemaakt met de expertise van Sigrid De Keersmaecker, Inge Thijs, Geert Schoofs en Amy De Weerdt van het CMPG. Bedankt om te geloven in onze predicties en voor de tijd en moeite die jullie hebben gestoken in de experimentele validatie van onze resultaten!

Muchas gracias al prof. Julio Collado-Vides y a todos los miembros de CCG por invitarnos a México y por ayudarnos a utilisar RegulonDB. Más allá de la colaboración científica, quisiera dar gracias a ustedes por introducirnos a la cultura Mexicana: el Día de Muertos, la comida rica, las fiestas, Mambo Cafe, la piscina cerrada y las expresiones Mexicanas malas pero importantes. Muchas gracias a Tomas, Sarath, Vladimir, Paty, Heladia, Romualdo, Socorro, Alberto, Victor, Martín y todos del CCG, y espero que nos veamos pronto! Prof. Collado-Vides, muchas gracias por venir a Bélgica y estar aquí hoy.

Veel dank gaat uit naar de bioinformaticagroepen van ESAT en CMPG. Eerst en vooral wil ik alle meisjes bedanken die samen met mij het door sommigen gedoopte "wijveneiland" hebben bevolkt: Cynthia, Joke, Abeer, Riet en onze laatste aanwinst Daniela, maar natuurlijk ook Shi en Roland de enige twee mannen die dapper genoeg waren om ons aan dit eiland te vergezellen. Raf, Olivier, Tim, Wout, Fabián, Kristof, los aseres cubanos Aminael & Roldán, Carolina, Hui, Valerie, en alle andere collega's van de bioinformatica groep, maar ook een aantal mensen van de 'oudere' generatie zoals Pieter en Bert P. wil ik bedanken voor de fijne samenwerking, maar ook voor alle plezante momenten in de toren en op conferenties, de koffiepauzes en de ViaVia spaghetti's! Ida en Ilse, heel erg bedankt voor de hulp bij alle administratieve rompslomp!

Zeker niet minder belangrijk voor de voltooiing van dit doctoraat waren vrienden op wie ik altijd kon rekenen voor de nodige afleiding en ontspanning. Bedankt Mieke, Anthony, Barbara, Cynthia, Jan en Tim voor de x-wekelijkse drank-en eetgelegenheden en voor jullie steun en begrip voor mijn afwezigheid de laatste maanden. Veel dank aan Tom, Mieke, Kevin, Britt, Kristiaan, Sabine, Bert, Martien, Kristof en Judith voor de vele spelavonden, weekendjes en de muzikale toets in mijn leven en aan Patrik, Yannick en Karolien voor de avondjes uit in Hasselt.

Papa, mama, zus & Dieter, bedankt dat jullie altijd in mij geloofd hebben en altijd voor mij klaar stonden. Dankzij jullie onvoorwaardelijke steun heb ik dit doctoraat tot een goed einde kunnen brengen en daar kan ik jullie niet genoeg voor bedanken!

Lieve Thomas, je hebt de laatste maanden heel wat geklaag moeten aanhoren ... bedankt voor je eindeloze geduld en steun tijdens deze stressvolle periode. Maar ik wil je vooral bedanken om er altijd te zijn voor mij ...

Karen

Abstract

Organisms are able to adapt their cellular machinery to changing conditions. This complex cellular behavior is mediated by the underlying genetic regulatory network. Traditionally, molecular biology focused on the study of individual genes or small sets of genes. However, the growing awareness of the importance of interactions between multiple genes to explain the behavior of organisms as an integrated system recently triggered the emergence of a novel, multidisciplinary research domain, referred to as systems biology. Systems biology exploits genome-wide data with the aim to obtain global insight in the cell's interaction networks. Unraveling these regulatory networks to gain insight in the ways different biological entities function and interact in various conditions is considered one of the foremost challenges in current bioinformatics research.

Since heterogeneous data provide complementary information on the same biological system, data integration holds the key to the successful reconstruction of gene regulatory networks. Several network inference methods that make use of data integration strategies have generated promising biological results. In this thesis, we developed and applied two methods for the identification of transcriptional regulatory modules. Both methods rely on association rule mining algorithms. These algorithms are able to integrate heterogeneous data sources in a natural way and do not suffer from problems with local optima.

In a first phase of this PhD, we developed ReMoDiscovery, a data integration framework for the detection of regulatory modules. It makes use of a well-known association rule mining algorithm to integrate heterogeneous data sources that provide complementary information on the transcriptional network. An application of ReMoDiscovery on some well-known yeast data sets identified regulatory modules covering many biological processes. The approach outperformed several existing module detection tools, proving that association rule mining based tools are well suited for module detection through data integration.

A second phase focused on issues related to data collection. Because creating and maintaining comprehensive data compendia is a difficult and time-consuming task, most network reconstruction algorithms are currently being applied on subsets of the data only. However, fully exploiting all available data providing information on the transcriptional network is

important to obtain a global view on the organism's genetic regulation mechanisms. For these reasons, an *E. coli* cross-platform microarray compendium containing data from different laboratories, from different public databases and even from different platforms was created. We also developed GeneReg, a data collection tool to automate the data collection process.

In a third and final phase of this PhD, we focused on the condition dependency of the regulatory network. Because in large expression compendia genes are less likely to be co-expressed in all conditions, a bicluster strategy becomes necessary. DISTILLER, our second data integration framework, is capable of such a condition selection. Application of DISTILLER on the *E. coli* expression compendium and regulatory motif data allowed us to study the condition-dependent transcriptional network of *E. coli* and to extend it considerably with novel regulator-gene interactions. Our analysis also revealed that compared to what can be expected from RegulonDB, a surprisingly small amount of sets of co-expressed genes seem to be regulated by two or more regulators in the same environmental conditions.

Korte inhoud

Door de werking van het regulatorische netwerk past een organisme zich aan aan wisselende omgevingscondities. In traditionele moleculaire biologie worden meestal een kleine geïsoleerde set van genen bestudeerd om inzicht te krijgen in het regulatorische netwerk. Een gen functioneert echter als onderdeel van een groot regulatorisch netwerk met veel interacties tussen alle onderdelen. Het doel van systeembiologie, een onderzoeksdomein dat recent veel aan belang gewonnen heeft, is om een globaal inzicht te verkrijgen in dit interactienetwerk door gebruik te maken van data op genoomwijd niveau. Het verwerven van inzicht in de manier waarop de verschillende biologische entiteiten van het netwerk functioneren en interageren in uiteenlopende omgevingstoestanden, is van groot belang om biologische processen beter te begrijpen.

Verschillende heterogene datatypes beschrijven hetzelfde cellulaire systeem vanuit een andere invalshoek. Door deze meetgegevens op de juiste manier te integreren kunnen we fundamenteel inzicht verwerven in de werking van dit systeem. In deze thesis werden twee methodes voor de identificatie van transcriptionele regulatorische modules ontwikkeld en toegepast. Beide methodes zijn gebaseerd op algoritmes voor het ontdekken van zogenaamde associatieregels of 'association rule mining' algoritmes. In tegenstelling tot andere methodes kunnen deze algoritmes heterogene data op natuurlijke wijze integreren en hebben zij geen problemen met lokale optima.

Tijdens de eerste fase van dit doctoraat werd ReMoDiscovery ontwikkeld, een data-integratie raamwerk voor de detectie van regulatorische modules. Het raamwerk maakt gebruik van een gekend "association rule mining" algoritme voor de integratie van verschillende soorten data die het transcriptionele netwerk op complementaire wijze beschrijven. ReMoDiscovery werd toegepast op enkele reeds uitvoerig bestudeerde gist datasets waarbij modules met uiteenlopende biologische functies werden geïdentificeerd. Onze methode presteerde beter dan voorgaande methodes en leverde daarom het bewijs dat methodes gebaseerd op 'association rule mining' algoritmes geschikt zijn voor detectie van modules via data-integratie.

Een tweede fase van het doctoraat spitste zich toe op het verzamelen van beschikbare data omtrent het transcriptionele netwerk. Omdat het uitermate moeilijk en tijdrovend is om data compendia te creëeren en onderhouden, gebruiken veel netwerk-inferentie methodes slechts een gedeelte van de

beschikbare data maar het bundelen van de informatie uit alle beschikbare data is cruciaal om een zo volledig mogelijk zicht te verkrijgen op het transcriptioneel netwerk. Om deze reden werd een compendium dat *E. coli* microroosters bevat van verschillende laboratoria, van verschillende publieke databanken en zelfs van verschillende platformen aangelegd. We ontwikkelden ook GeneReg, een software programma om het verzamelen van data te automatiseren.

In een derde en laatste fase van dit doctoraat concentreerden we ons op de conditie-afhankelijkheid van het regulatorische netwerk. In bovengenoemde grote expressiecompendia verwacht men niet dat genen samen tot expressie komen in alle experimentele condities. DISTILLER, een tweede data-integratie raamwerk maakt het mogelijk om regulatorische modules te identificeren waarbij enkel condities worden geselecteerd waarin de genen samen tot expressie komen. Toepassing van DISTILLER op het *E.coli* compendium van expressiedata en regulatorische motiefdata liet ons toe om het conditie-afhankelijke transcriptionele netwerk van *E.coli* te bestuderen en uit te breiden met nieuwe regulator-gen interacties. Onze analyse onthulde ook dat in vergelijking met wat verwacht kon worden van RegulonDB slechts een verrassend klein aantal sets van genen die samen tot expressie komen onder controle staan van twee of meerdere regulatoren in dezelfde experimentele condities.

Acronyms and terminology

Acronyms

AE	ArrayExpress
ARM	Association rule mining algorithms
BW	Band width
ChIP-chip	Chromatin immunoprecipitation (ChIP) on a microarray (chip)
CLR	Context of likelihood relatedness
COGRIM	Clustering of genes into regulons using integrated modeling
DISTILLER	Data Integration System To Identify Links in Expression Regulation
DNA	Deoxy ribonucleic acid
FDR	False discovery rate
FN	False negatives
FP	False positives
GEO	Gene expression omnibus
GRAM	Genetic Regulatory Modules
MAS 5.0	MicroArray Suite 5.0
mRNA	Messenger RNA
ODE	Ordinary differential equations
ReMoDiscovery	Regulatory module discovery
RMA	Robust Multi-array Average
RNA	Ribonucleic acid
SAMBA	Statistical-Algorithmic Method for Bicluster Analysis
s_c	Regulator support
s_m	Motif support
SMD	Stanford microarray database

t_c	Regulator threshold
t_e	Correlation threshold
t_m	Motif threshold
TN	True negatives
TP	True positives
UAS	Upstream activator sequence
Y2H	Yeast-two-hybrid

Terminology

(biological) network inference	Learning the structure of (biological) circuits from data
Bicluster	Set of genes that are co-expressed in a subset of conditions of a data set
Closed itemset	Frequent itemset that cannot be extended with an additional item without changing the support
Frequent itemset	Itemset of which the support exceeds the support threshold
High-throughput data	Data that were generated on a genome-wide scale
Item	A basic element in association rule mining algorithms. Items are grouped together to form itemsets
Itemset	A group of items
Maximal itemset	Frequent itemset that will not meet the support threshold anymore upon addition of an extra item
Module	Set of genes that are co-expressed and often co-regulated in (a subset of) conditions of a data set
Motif	The representation of a set of binding sites
Motif instance	A binding site in the promoter region of a gene
Regulatory program	Regulators and/or motifs
Support	The number of transactions in which the items of an itemset appear together
Support threshold	The minimum support
Transaction	A property shared by a goup of items

Nederlandse samenvatting

Hoofdstuk 1: Inleiding

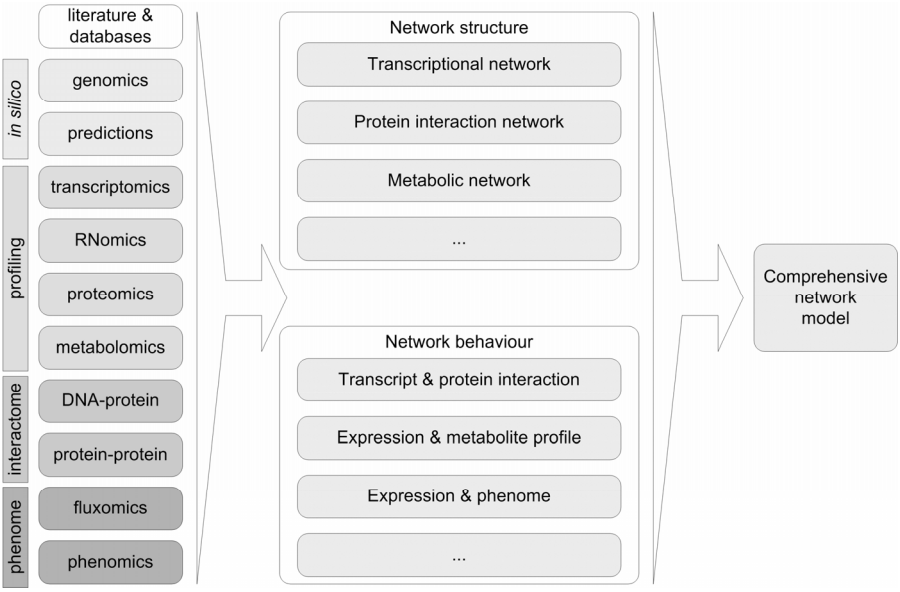
Situering en doel van de thesis

In het verleden werd reeds een groot aantal biologische principes ontrafeld door één gen of een kleine set van genen in detail te bestuderen. Dit soort studies geeft echter slechts een beperkt beeld van de cel, weefsels of organismen. Systeembiologie, daarentegen, stelt zich net tot doel om te begrijpen hoe de verschillende biologische componenten (zoals DNA, RNA, proteïnen en metabolieten) samenwerken om het organisme leefbaar te maken. Een gen wordt hierbij niet langer bestudeerd als een geïsoleerde entiteit maar eerder als onderdeel van een complex netwerk of systeem. Het ultieme doel is het modelleren van de dynamiek van deze netwerken om zo cellulair gedrag te kunnen voorspellen.

Tegenwoordig kunnen verschillende cellulaire entiteiten en hun interacties op grote schaal bestudeerd worden door de beschikbaarheid van hoge doorvoer methodes. Met deze methodes wordt zogenaamde 'omics' data verkregen die informatie geven over het onderliggende netwerk. In top-down systeembiologie worden deze heterogene data gecombineerd om het regulatorische netwerk te ontrafelen. Combinatie van verschillende soorten heterogene data biedt twee voordelen. Enerzijds belicht elke databron verschillende aspecten van het netwerk waardoor de integratie ervan leidt tot een meer volledig inzicht in het netwerk (De Keersmaecker *et al*, 2006). Anderzijds bevat de data vaak veel fouten of is de data niet volledig. Een goed voorbeeld hiervan zijn de fluctuaties in genexpressie tussen twee schijnbaar identieke experimentele condities (Baldi and Long, 2001). De experimentele en biologische ruis van omics data kunnen er voor zorgen dat deze data slechts beperkt bruikbaar zijn. Als we echter aannemen dat de fouten in verschillende datasets onafhankelijk zijn, dan kunnen de fouten in de resultaten die verkregen zijn door twee databronnen te combineren drastisch afnemen (Hartemink and Segal, 2005). Deze beperkingen van omics data kunnen dus overwonnen worden door data van twee of meerdere methodes te integreren. De integratie van meerdere datasets leidt daarom tot nieuwe inzichten die niet zouden kunnen verkregen worden door slechts één databron te bestuderen (Hartemink and Segal, 2005). In deze thesis

concentreren we ons daarom op de integratie van data voor de reconstructie van het regulatorische netwerk (zie Figuur N - 1).

In deze thesis onderzoeken we hoe genexpressie in cellen kan veranderen onder invloed van bepaalde omgevingstoestanden. Hoewel recent enkele methodes ontwikkeld werden voor de reconstructie van het transcriptionele netwerk, brengen deze methodes meestal niet alle aspecten van het netwerk in rekening. De meeste methodes concentreren zich ofwel op de conditie-afhankelijke expressie van de genen (zoals biclusteringsalgoritmes) ofwel op de integratie van verschillende types van data, maar niet op beide. Daarenboven kunnen niet alle methodes de regulatie van een gen door meerdere regulatoren in rekening brengen. In deze thesis werden daarom twee methodes, ReMoDiscovery en DISTILLER, ontwikkeld die gebruik



Figuur N - 1: Integratie van verschillende databronnen om een meer volledig beeld op het netwerk te verkrijgen. De *in silico* data verwijzen naar bijvoorbeeld voorspellingen van regulatorische motieven. De 'profiling' data meten welke cellulaire component aanwezig is en in welke hoeveelheid. De 'interactome' data beschrijven of er een interactie tussen twee componenten plaatsvindt in de cel. De 'phenome' data geven informatie over het algemene gedrag van het systeem. De kennis van literatuur en databanken kan gecombineerd worden met verschillende soorten omics data en *in silico* voorspellingen om de structuur van verschillende types van netwerken de bepalen en de relatie tussen de verschillende netwerken (netwerkgedrag of 'network behaviour'). Het ultieme doel van data integratie is om alle netwerkinformatie te integreren om zo één volledig netwerk model te bekomen. Figuur uit De Keersmaecker *et al* (2006).

maken van data integratie voor de studie van het (conditie-afhankelijke) modulaire transcriptionele netwerk. Deze methodes werden succesvol toegepast op twee modelorganismen om zo modulariteit, conditie-afhankelijke expressie, combinatoriële regulatie en de relatie tussen deze eigenschappen in detail te bestuderen.

Transcriptie en transcriptie regulatie

Het dubbelstrengige deoxyriboucleotidezuur (DNA) bevat de code van het leven, maar de proteïnen zijn de echte functionele elementen in een cel. Genen zijn segmenten in het DNA die coderen voor proteïnen via een intermediair produkt, het mRNA. Om een proteïne te vormen wordt een gen eerst afgeschreven naar een mRNA-sequentie via het transcriptieproces. Vervolgens kan dit mRNA verder vertaald worden naar een aminozuur-sequentie (of proteïne) tijdens het translatieproces.

Wanneer omgevingsfactoren veranderen, moet een cel zich kunnen aanpassen aan deze veranderingen: actieve genen worden uitgeschakeld en inactieve genen worden geactiveerd. Regulatiemechanismen zijn dus nodig om ervoor te zorgen dat genen enkel afgeschreven worden indien hun overeenkomstige proteïnen nodig zijn in de cel.

De transcriptie kan beïnvloed worden op verschillende manieren. Ten eerste zijn er in het promotergebied van zowel prokaryote als eukaryote organismen consensussequenties aanwezig. Deze consensussequenties worden herkend door het RNA-polymerase en beïnvloeden de binding van het RNA-polymerase. Een bijkomende transcriptionele regulatie is afkomstig van transcriptiefactoren. Deze regulatoren reageren op bepaalde omgevingsfactoren en kunnen de binding van een RNA-polymerase op de DNA-sequentie vergemakkelijken of bemoeilijken. Deze transcriptiefactoren herkennen korte geconserveerde DNA-sequenties, zogenaamde motieven, in de promoter regio van het gen.

Vermits transcriptionele regulatie een zeer complex proces is, worden genen vaak gereguleerd door meer dan één transcriptiefactor. Het effect van de binding van twee regulatoren kan zowel synergistisch als antagonistisch zijn. De kennis van de mechanismen van combinatoriële regulatie is tot op heden vrij beperkt. Vaak weet men bijvoorbeeld wel dat meerdere regulatoren hetzelfde gen reguleren maar kan men niet zeggen of dat deze regulatoren in dezelfde experimentele condities actief zijn of eerder onafhankelijk van elkaar werken in verschillende experimentele condities.

Reconstructie van het transcriptionele netwerk

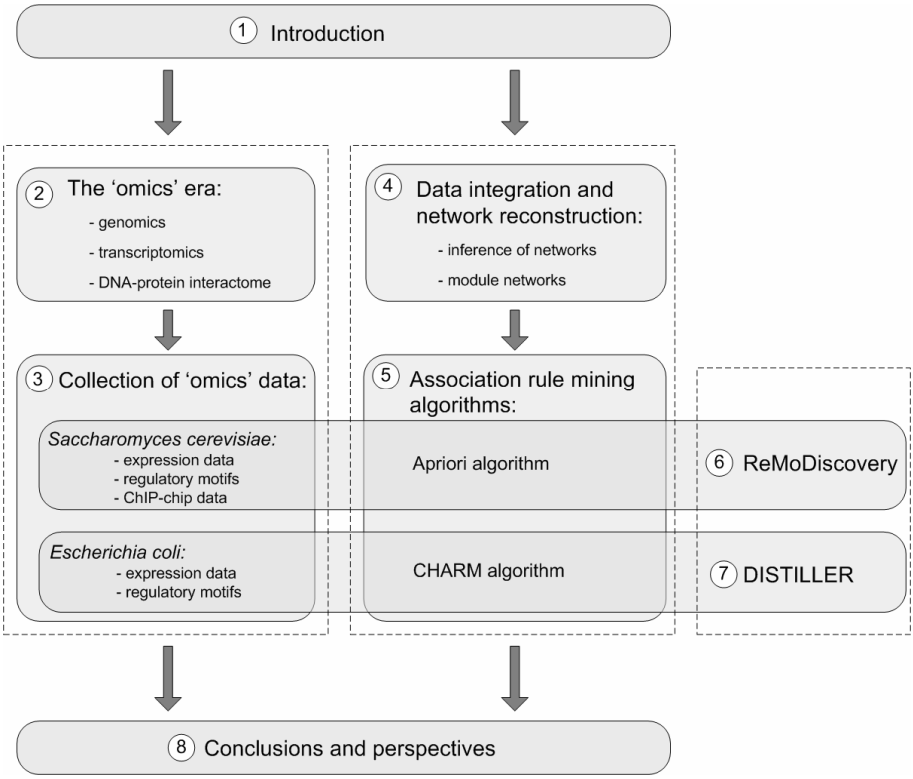
Transcriptionele regulatie kan voorgesteld worden in de vorm van een transcriptioneel netwerk. Het transcriptionele netwerk is gedefinieerd als een graaf waarin de nodes de genen voorstellen. Een pijl van een gen naar een ander gen wijst op het feit dat het eerste gen codeert voor een transcriptiefactor die het tweede gen reguleert. De architectuur van dit netwerk kan beschreven worden aan de hand van graaf-eigenschappen. In biologische netwerken interageren de meeste genen slechts met een klein aantal andere genen, maar enkele genen, ook wel hubs genoemd, zijn betrokken in veel interacties. Dit duidt erop dat het transcriptionele netwerk *scale-free* is. Het transcriptionele netwerk is ook modulair en hiërarchisch georganiseerd. Biologische processen zijn opgebouwd uit zogenaamde *pathways* die vooral zelfstandig functioneren maar waartussen wel communicatie aanwezig is. Biologische processen zijn bijgevolg georganiseerd in afgescheiden modules.

Het doel van de reconstructie van het transcriptionele netwerk is om voor alle transcriptiefactoren te kunnen bepalen welke regulatorische motieven ze herkennen, in welke condities ze actief zijn, met welke regulatoren ze samenwerken tijdens deze condities en welke genen ze reguleren. Twee technieken bestaan voor de reconstructie van het transcriptionele netwerk: een “bottom-up” en “top-down” benadering. De eerstvernoemde techniek maakt gebruik van modellen van gekende interacties tussen moleculaire entiteiten. Deze modellen worden vervolgens gebruikt om cellulair gedrag te simuleren en voorspellen. In deze thesis wordt de top-down techniek toegepast. Deze methodes starten van een globaal overzicht van het systeem door het gebruik van grote hoeveelheden omics data.

Tijdens netwerkreconstructie mag het dynamische aspect van het netwerk niet vergeten worden: sommige interacties tussen genen en regulatoren kunnen aanwezig zijn in bepaalde experimentele condities maar afwezig zijn in andere condities. Het regulatorische netwerk is dus verschillend in verschillende experimentele condities.

Overzicht van de thesis

Hoofdstuk 2 geeft een overzicht van de ‘omics’ data die rechtstreeks kunnen gerelateerd worden aan het transcriptieproces. Deze data zullen in latere hoofdstukken dan ook gebruikt worden voor de reconstructie van regulatorische netwerken (zie Figuur N - 2).



Figuur N - 2: Overzicht van de structuur van de thesis. In hoofdstuk 2 geven we een overzicht van de verschillende types van omics data die informatie geven over het transcriptionele netwerk en vervolgens beschrijven we in hoofdstuk 3 hoe we een data compendium creëerden voor twee modelorganismen. Hoofdstuk 4 geeft een overzicht van de bestaande data integratie en netwerk reconstructie methodes. Omdat onze ontwikkelde methodes voor module detectie gebaseerd zijn op association rule mining algoritmes, geeft hoofdstuk 5 een meer gedetailleerde beschrijving van deze algoritmes. Tenslotte worden onze twee data integratie raamwerken besproken. In hoofdstuk 6 worden de Apriori-gebaseerde methode ReMoDiscovery en de resultaten van de toepassing op gist data besproken. Vervolgens bespreken we in hoofdstuk 7 het DISTILLER raamwerk dat gebaseerd is op het CHARM algoritme. DISTILLER werd toegepast op *E. coli* data en laat toe om de conditie-afhankelijke co-expressie en combinatoriële regulatie te bestuderen.

De eerste stap in de reconstructie van regulatorische netwerken is het verzamelen van alle relevante beschikbare data. In **hoofdstuk 3** bespreken we hoe compendia van verschillende data gecreëerd werden voor twee modelorganismen, *Saccharomyces cerevisiae* en *Escherichia coli*. Vermits deze 'omics' data niet altijd even gemakkelijk kunnen verzameld worden, werd er in samenwerking met het departement Computerwetenschappen (KU Leuven, Belgium) software ontwikkeld voor automatisatie van het verzamelen van data:

Pastor D., Cortes-Calabuig A., Lemmens K., De Moor B., Marchal K., Denecker M. (2007) GeneReg: Integration of Experimental Data on the DNA Transcription Process. Proceedings of BNAIC 2007.

In **hoofdstuk 4** beschrijven we hoe de 'omics' data kunnen gebruikt worden voor de reconstructie van regulatorische netwerken. We maken een onderscheid tussen methodes die gebruik maken van data integratie en methodes die enkel expressiegegevens gebruiken. Een tweede indeling van de methodes is gebaseerd op het onderscheid tussen methodes die rechtstreeks netwerkinteracties zoeken en methodes die regulatorische modules identificeren. In dit hoofdstuk bespreken we tenslotte ook de nood aan vergelijkende studies van deze methodes. Een overzicht van de verschillende methodes is beschikbaar in volgende publicatie:

Van den Bulcke T., Lemmens K., Van de Peer Y., Marchal K. (2006) Inferring Transcriptional Networks by Mining Omics Data. Current Bioinformatics, vol. 1, no. 3, pp. 301-313.

Een methode voor de identificatie van biclusters op basis van enkel expressiedata is gepubliceerd:

Dhollander T., Sheng Q., Lemmens K., De Moor B., Marchal K., Moreau Y. (2007) Query-driven module discovery in microarray data. Bioinformatics, vol. 23, no. 19, pp. 2573-2580.

Hoewel verschillende algoritmes voor netwerk reconstructie bestaan, brengt geen van deze methodes alle aspecten van transcriptionele regulatie in rekening. In deze thesis werden twee methodes voor netwerk reconstructie ontwikkeld die gebaseerd zijn op *association rule mining* algoritmes. In **hoofdstuk 5** worden deze *association rule mining* algoritmes en hun gebruik in bioinformatica besproken.

In **hoofdstuk 6** beschrijven we ReMoDiscovery, een Apriori-gebaseerd data integratie algoritme. Onze methodologie maakt gebruik van microroostergegevens, regulatorische motieven en ChIP-chip data voor de identificatie van regulatorische modules. Een module bestaat uit een set van genen die samen tot expressie komen, die gereguleerd worden door dezelfde

regulators en die dezelfde motieven in hun promotergebied bezitten. ReMoDiscovery werd toegepast op enkele goed gekende gist datasets en modules die betrokken zijn in een reeks van biologische processen konden geïdentificeerd worden. De resultaten van ReMoDiscovery werden vergeleken met de resultaten van andere netwerk reconstructie methodes. Deze studie werd gepubliceerd in:

> Lemmens K., Dhollander T., De Bie T., Monsieurs P., Engelen K., Smets B., Winderickx J., De Moor B., Marchal K. (2006) Inferring transcriptional module networks from ChIP-chip-, motif- and microarray data. Genome Biology, vol. 7, no. 5, pp. R37.

Het gebruik van ReMoDiscovery toonde aan dat methodes gebaseerd op association rule mining algoritmes even geschikt zijn voor de reconstructie van netwerken als andere meer gebruikte graaf-gebaseerde of probabilistische algoritmes. Met het groeiend aantal beschikbare microroosterexperimenten werd een selectie van de condities of biclusterstrategie een noodzaak. In een tweede fase van dit doctoraat werd daarom DISTILLER ontwikkeld, een methode voor de identificatie van modules die ook een selectie van de condities toelaat. Vermits een dergelijke biclusterstrategie ook computationeel zwaarder is, werden we verplicht ons te baseren op een beter *association rule mining algoritme*, CHARM. Onze data integratie methode DISTILLER werd toegepast op een compendium van 870 *Escherichia coli* microroosters en regulatorische motief data voor 67 motieven. Dankzij onze bicluster strategie konden we de relatie tussen conditie-afhankelijke co-expressie en combinatoriële regulatie bestuderen. DISTILLER en zijn toepassing op *E. coli* zijn beschreven in **hoofdstuk 7**. De resultaten van deze studie zijn beschikbaar via:

> Lemmens K., De Bie T., Dhollander T., Monsieurs P., De Moor B., Collado-Vides J., Engelen K., Marchal K. (2008) The condition-dependent transcriptional network in *Escherichia coli*. Accepted for publication in Annals of the NYAS, proceedings of the DREAM2 conference.

en:

> Lemmens K., De Bie T., Dhollander T., De Keersmaecker S., Thijs I., Schoofs G., De Weerdt A., De Moor B., Vanderleyden J., Collado-Vides J., Engelen K., Marchal K. (2008) DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli*. submitted.

De belangrijkste resultaten van deze thesis en mogelijke uitbreidingen naar de toekomst toe zijn samengevat in **hoofdstuk 8**. Dit hoofdstuk beschrijft ook hoe ons eerder fundamenteel onderzoek past in het bredere kader van de systeembiologie.

Hoofdstuk 2: Het 'omics' tijdperk

Het doel van deze thesis is de reconstructie van het regulatorische netwerk door middel van integratie van verschillende soorten omics data. In dit hoofdstuk bespreken we daarom de data die rechtstreeks gerelateerd zijn aan de reconstructie van het transcriptionele netwerk: regulatorische motief data (genoom data), expressie data (transcriptoom data) en proteïne-DNA interactie data. Deze databronnen geven elk op zich informatie over telkens een ander aspect van het transcriptionele netwerk en de integratie ervan zal daarom helpen om een meer gedetailleerd inzicht te verwerven in het netwerk.

Genoom data

Volledige genoomsequenties zijn beschikbaar voor 628 prokaryote en 22 eukaryote organismen. Aan de hand van deze genoomsequenties konden regulatorische elementen ontdekt worden (Kellis *et al*, 2003; Monsieurs *et al*, 2005; Monsieurs *et al*, 2006) of kan evolutie bestudeerd worden (Delsuc *et al*, 2005). De bindingsplaatsen van transcriptiefactoren of regulatorische motieven vormen een zeer belangrijke bron van informatie voor de studie van het transcriptionele netwerk. Regulatorische motieven zijn immers korte geconserveerde DNA-sequenties in het promotergebied van een gen. Deze motieven worden herkend en gebonden door regulatoren die zo het gen reguleren. Hoewel de binding van een regulator op het regulatorische motief conditie-afhankelijk is, is het motief zelf inherent aanwezig in de DNA-sequentie.

Verscheidene computationele methodes werden reeds ontwikkeld voor de identificatie van regulatorische motieven. Een eerste groep van methodes heeft als doel om regulatorische motieven te identificeren zonder enige voorkennis over de motiefsequentie. Deze methodes zijn de zogenaamde *de novo* motief detectie methodes. Alternatieve motief-screening methodes maken gebruik van gekende of voorspelde motieven en zoeken in DNA-sequenties naar deze gekende motieven.

Transcriptoom data

Transcriptoom data meten de aan- en afwezigheid van elk type mRNA molecule in een cel in een bepaalde experimentele conditie. Tijdens de laatse jaren is de microroostertechnologie uitgegroeid tot een efficiënte techniek waarmee het mRNA- of expressieniveau van duizenden genen in één enkel experiment kan gemeten worden. Vermits mRNA belangrijke informatie

biedt over het regulatorische netwerk, zullen microroostergegevens gebruikt worden in deze thesis. Deze microrooster experimenten zijn in grote getallen aanwezig in publieke databanken zoals Gene Expression Omnibus (Barrett *et al*, 2007), Stanford microarray database (Demeter *et al*, 2007) of ArrayExpress (Parkinson *et al*, 2007). Tijdens dit doctoraat zullen we dan ook gebruik maken van deze databanken om alle beschikbare microroosterexperimenten te kunnen gebruiken.

Proteïne-DNA interacties

Het transcriptionele regulatorische netwerk bestaat uit interacties tussen transcriptiefactoren en de promotergebieden van hun doelwitgenen. Voor de reconstructie van het transcriptionele netwerk zullen we daarom ook data gebruiken die rechtstreeks informatie geven over de binding van regulatoren. Chromatine-immunoprecipitatie in combinatie met microroosters (ChIP-chip) laat toe om de fysische interactie tussen een regulator en de promotergebieden van zijn doelwitgenen te meten en is daarom ook een van de meest populaire technieken geworden om proteïne-DNA interacties te bepalen.

Hoofdstuk 3: Het verzamelen van 'omics' data

Een eerste en cruciale stap in de reconstructie van het transcriptionele netwerk is het samenstellen en *pre-processen* (normalizatie en eventueel herschaling) van omics data compendia. In het voorgaande hoofdstuk hebben we reeds de types van data besproken die in deze thesis gebruikt werden. In dit hoofdstuk beschrijven we hoe we de data compendia creëerden voor twee modelorganismen: de prokaryote bacterie *Escherichia coli* en de eukaryote gist *Saccharomyces cerevisiae*. Vermits deze organismen modelorganismen zijn, is er veel data beschikbaar en zijn deze organismen reeds goed bestudeerd.

De data die verzameld werden voor gist bestond uit twee goed gekende expressiedatasets (Gasch *et al*, 2000; Spellman *et al*, 1998), een ChIP-chip dataset (Harbison *et al*, 2004) en een compendium van regulatorische motieven (Kellis *et al*, 2003). Deze data werden als input gebruikt door ReMoDiscovery (hoofdstuk 6).

In een tweede fase van het doctoraat ontwikkelden we DISTILLER, een data integratie methode die bovendien ook condities kan selecteren waarin een set van genen samen tot expressie komen. DISTILLER werd toegepast op een grote heterogene expressiedataset van *E. coli* in combinatie met regulatorische motieven.

In eerste instantie creëerden we een compendium van regulatorische motieven. Alle gekende motieven van *E. coli* zijn beschikbaar in de RegulonDB databank (Salgado *et al*, 2006). Met deze motieven werden gezocht in alle intergenische regio's voor de aanwezigheid van deze motieven, waardoor ook nieuwe predicties mogelijk waren.

Naast deze regulatorische motieven werden ook alle beschikbare microroosters verzameld uit de drie grote microrooster databanken (Barrett *et al*, 2007; Demeter *et al*, 2007; Parkinson *et al*, 2007). Na verwijdering van alle redundante microroosters verkregen we een heterogeen compendium bestaande uit 870 microroosters. We creëerden zo het eerste heterogene compendium van *E. coli* dat microroosters bevat gegenereerd via verschillende microroostertechnieken en afkomstig van verschillende laboratoria. Pre-processing van de data was noodzakelijk om alle data vergelijkbaar te maken tussen verschillende experimentele condities en verschillende laboratoria.

Het creëren van een dergelijk compendium is een tijdsrovende en moeilijke taak. Ten eerste moeten alle experimenten gedownload worden (vaak één per één) en moet men weten welke nieuwe data er beschikbaar is. De data bevindt zich vaak verspreid over verschillende databanken met overlappende inhoud en is beschikbaar in verschillende formaten. Heterogene data zoals regulatorische motieven of ChIP-chip data zijn verspreid over verschillende websites en databanken zodat het nog moeilijker is om toegang te krijgen tot al deze gegevens. Om een zo gedetailleerd mogelijk inzicht te kunnen krijgen in het transcriptionele netwerk is het echter cruciaal om zoveel mogelijk informatie in rekening te brengen. Daarom werd in samenwerking met het departement Computerwetenschappen aan de K.U. Leuven (David Pastor Escuerdo, Álvaro Cortés Calabuig and Prof. Marc Denecker) software ontwikkeld, genaamd GeneReg, om dit proces van data collectie te automatiseren (Pastor *et al*, 2007). In zijn huidige implementatie kan GeneReg ChIP-chip data en regulatorische motief data opslaan en microrooster data uit de Stanford Microarray Database halen en opslaan. Het uiteindelijke doel is om de toepassing meer generisch to maken zodat voor elk organisme en voor verschillende types van databanken de data kan verzameld worden.

Hoofdstuk 4: Integratie van data en netwerk reconstructie

Tijdens dit doctoraat werden twee methodes voor de reconstructie van het transcriptionele netwerk ontwikkeld. In dit hoofdstuk bespreken we een aantal gerelateerde methodes voor de reconstructie van netwerken. Deze methodes kunnen ingedeeld worden in methodes die enkel

expressiegegevens gebruiken voor de netwerk reconstructie en methodes die verschillende soorten heterogene data integreren. Bovendien kunnen de methodes ook geclassificeerd worden al naargelang ze rechtstreeks het transcriptionele netwerk identificeren of dat er regulatorisch modules geïdentificeerd worden. Een regulatorische module bestaat uit een set van genen die samen tot expressie komen en onder invloed staan van hetzelfde regulatiemechanisme. Op het einde van dit hoofdstuk bespreken we ook enkele huidige uitdagingen in verband met het vergelijken van resultaten en performantie van de verschillende methodes.

Netwerk reconstructie op basis van expressiedata

Er bestaan vier grote klassen van methodes die het netwerk rechtstreeks kunnen identificeren (Bansal *et al*, 2007; Gardner and Faith, 2005): Booleaanse netwerken (Akutsu *et al*, 1999; Akutsu *et al*, 2000; Liang *et al*, 1998), 'ordinary differential equation' netwerken (Bonneau *et al*, 2006; di Bernardo *et al*, 2005; Gardner *et al*, 2003; Tegner *et al*, 2003; van Someren *et al*, 2006), Bayesiaanse netwerken (Friedman *et al*, 2000; Hartemink *et al*, 2001; Yu *et al*, 2004b) en associatie-netwerken (Faith *et al*, 2007; Margolin *et al*, 2006). Deze methodes gaan op grote schaal na welke RNA-transcripten een invloed hebben op een bepaald RNA-transcript.

In hoofdstuk 7 vergelijken we onze netwerk reconstructie methode met de associatie-netwerk methode van Faith *et al* (2007). Faith *et al* (2007) ontwikkelden de 'context of likelihood relatedness' (CLR). CLR start van een lijst van transcriptiefactoren en zoekt vervolgens naar doelwitgenen van deze transcriptiefactoren. Deze methode voorspelt een interactie tussen een transcriptiefactor en doelwitgen als hun beide expressieprofielen voldoende op elkaar gelijken.

In plaats van het regulatorische netwerk model rechtstreeks af te leiden, kan men ook regulatorische modules identificeren. Clustering (Eisen *et al*, 1998; Tamayo *et al*, 1999; Tavazoie *et al*, 1999) en biclustering algoritmes (Cheng and Church, 2000; Ihmels *et al*, 2002; Madeira and Oliveira, 2004; Murali and Kasif, 2003; Sheng *et al*, 2003; Tanay *et al*, 2002) zijn geschikt om de modulariteit van het transcriptionele netwerk te onthullen. Een cluster bestaat uit een set van genen met een gelijkaardig expressieprofiel in alle gemeten experimentele condities, terwijl een bicluster gedefinieerd is als een set van genen met een gelijkaardig profiel in slechts een gedeelte van de condities. In een derde benadering, query-driven biclustering genaamd, wordt er gezocht naar een bicluster of een set van biclusters rond een bepaald gen of een set van genen die betrokken zijn in hetzelfde biologische proces (Bergmann *et al*, 2003; Dhollander *et al*, 2007). Het doel van query-driven biclustering is om bijkomende genen te vinden die mogelijk

betrokken zijn in dezelfde functie als de oorspronkelijke genen en om de condities te identificeren waarin de genen samen tot expressie komen.

Hoewel deze clustering en biclustering algoritmes nuttig zijn om de modulariteit van het netwerk te bestuderen, identificeren deze methodes niet rechtstreeks het regulatorische programma, bestaande uit regulatoren en overeenkomstige motieven, dat verantwoordelijk is voor co-expressie van de genen. Segal *et al* (2003) ontwikkelden daarom een methode die de genen verdeelt in modules en tegelijkertijd zoekt naar de regulatoren die het expressiegedrag van de genen in de module kunnen verklaren. LeMoNe (Michoel *et al*, 2007) is gebaseerd op de methode van Segal *et al* (2003), maar gebruikt een alternatieve strategie om de het regulatorische programma te identificeren.

Netwerk reconstructie op basis van integratie van data

Ook hier kan een onderscheid gemaakt worden tussen methodes die het netwerk rechtstreeks afleiden en methodes die gebaseerd zijn op de modulaire eigenschappen van het netwerk. Bayesiaanse netwerken behoren tot de eerste klasse, maar vereisen een grote hoeveelheid data. Dit probleem kan verholpen worden door gebruik te maken van priorkennis. Zo kunnen bijvoorbeeld ChIP-chip of motif informatie gebruikt worden in combinatie met expressiedata (Bernard and Hartemink, 2005; Tamada *et al*, 2003).

Een andere recent gepubliceerde netwerk reconstructie methode, SEREND past een iteratief classificatieschema toe gebaseerd op logistieke regressie (Ernst *et al*, 2008). Deze methode gebruikt de kennis over gekende regulator-doelwitgen interacties tesamen met expressiedata en regulatorische motieven om zo nieuwe interacties te kunnen voorspellen. In hoofdstuk 7 vergelijken we onze methode met SEREND.

Een tweede type van netwerk reconstructie methodes tracht de modulariteit van het netwerk te onthullen op basis van data integratie. Verscheidene methodes bestaan reeds die gebruik maken van expressiedata, regulatorische motieven en/of ChIP-chip data (Bar-Joseph *et al*, 2003; Chen *et al*, 2007; Gao *et al*, 2004; Kato *et al*, 2004; Reiss *et al*, 2006; Tanay *et al*, 2004; Xu *et al*, 2004).

In hoofdstuk 6 vergelijken we onze methode voor de detectie van regulatorische modules op basis van data integratie met SAMBA (Tanay *et al*, 2004) en GRAM (Bar-Joseph *et al*, 2003). SAMBA bestaat uit een graaf-biclusteralgoritme om simultaan expressiedata, ChIP-chip, proteïne-proteïne interactie data en fenotype data te analyseren. GRAM zoekt naar modules op basis van microrooster data en ChIP-chip data en gebruikt hiervoor een iteratieve procedure. In de eerste stap worden de ChIP-chip data gebruikt om

een set van genen te vinden waarvan de promotergebieden gebonden worden door een gemeenschappelijke set van regulatoren. In een tweede stap worden de microrooster data gebruikt om in deze set van genen te zoeken naar genen die samen tot expressie komen. Tenslotte wordt deze basisset van genen verder uitgebreid met additionele genen.

Association rule mining

In dit doctoraat werden twee algoritmes ontwikkeld die beiden gebaseerd zijn op *association rule mining* algoritmes (ARM algoritmes). ARM algoritmes worden tegenwoordig meer gebruikt in de bioinformatica, maar tot nu toe werden ze nog niet vaak toegepast voor de identificatie van regulatorische modules op basis van data integratie. ARM algoritmes bieden echter vele voordelen:

- ARM algoritmes zoeken uitvoerig naar alle mogelijke oplossingen en ondervinden dus geen problemen met lokale optima. Omdat ze alle oplossingen of modules in één keer kunnen vinden, moeten deze algoritmes ook niet eerder gevonden modules maskeren om de volgende module te vinden.
- ARM algoritmes zijn ideaal voor data integratie: elke databron waarvan de relevante eigenschappen kunnen gedefinieerd worden op een deterministische en overerfbare manier kunnen in de analyse opgenomen worden.
- De meeste eerder vermelde data integratie methodes combineren de beschikbare data sequentieel terwijl ARM algoritmes toelaten om alle data tegelijkertijd te combineren.
- Terwijl sommige netwerk reconstructie methodes moeten steunen op de assumptie dat de expressieprofielen van transcriptiefactor en doelwitgen gelijkaardig zijn (Chen *et al*, 2007; Faith *et al*, 2007), is dit niet het geval voor ARM algoritmes.
- In tegenstelling tot sommige module detectie methodes waar een gen slecht tot één module kan horen (Cheng and Church, 2000; Segal *et al*, 2003), zijn ARM algoritmes in staat om rechtstreeks overlappende modules te detecteren.

Ondanks het feit dat ARM algoritmes een aantal voordelen bieden, werden ze tot nu toe zelden gebruikt voor de reconstructie van netwerken. Dit kan te wijten zijn aan twee problemen die ontstaan wanneer men ARM algoritmes gebruikt:

- Het eerste probleem is dat de standaard input voor ARM algoritmes binair of gediscretiseerd moet zijn. Echter, expressiedata

bijvoorbeeld bestaat uit continue waarden. Daarom wordt expressiedata vaak gediscretiseerd vooraleer ze gebruikt wordt door ARM algoritmes, wat tot een verlies aan informatie leidt.

- Een tweede probleem is dat de output van de meeste ARM algoritmes vaak een lange lijst van mogelijks interessante resultaten is zonder statistische significantie score. Dat wil zeggen dat meestal een filterstap noodzakelijk is om een selectie van de meest interessante resultaten te bekomen.

Omdat ARM algoritmes veel voordelen lijken te bieden ten opzichte van andere methodes zullen wij deze algoritmes gebruiken voor de detectie van regulatorische modules. Om deze te kunnen gebruiken zonder te veel informatie te verliezen onderzochten we mogelijke oplossingen voor het discretizatie probleem dat ontstaat bij het gebruik van microroostergegevens. Daarenboven bedachten we ook een efficiënte filterstrategie om alle biologische interessante modules te selecteren. Tenslotte ontwikkelden we ook een strategie om een selectie van condities toe te laten in een ARM raamwerk.

Welke methode kiezen als bioloog?

Er bestaan reeds een aanzienlijk aantal methodes voor de reconstructie van het transcriptionele netwerk. Op dit moment is er echter een gebrek aan informatie over de performantie van de verschillende methodes in verschillende situaties, zodat het moeilijk is voor een bioloog om te beslissen welke methode te gebruiken. De reden voor dit gebrek aan informatie is dat vele problemen zich voordoen wanneer men de methodes en hun resultaten probeert te vergelijken.

De meeste tijdschriften vragen tegenwoordig om een vergelijking met andere gerelateerde methodes bij publicatie van een nieuwe methode. Vaak worden dan de eigen resultaten vergeleken met de resultaten van de andere methodes zoals beschreven in het originele bijhorende artikel. Hoewel de verschillende methodes vaak gelijkaardige datasets als input gebruiken, is het meestal niet zo dat ze op exact dezelfde dataset werden toegepast. Vergelijkende studies zouden echter bij voorkeur op dezelfde biologische dataset moeten toegepast worden omdat alleen op deze manier de mogelijkheden en beperkingen van de methodes echt kunnen geanalyseerd worden. Bovendien is het zo dat de resultaten kunnen vergeleken worden met wat gekend is in de literatuur, maar geen enkel biologisch netwerk is tot op heden volledig gekend.

Nadat een dataset geselecteerd is kunnen de verschillende methodes toegepast en vergeleken worden op deze dataset. Een eerste uitdaging is dan het kiezen van de parameterwaardes. Meestal wordt een netwerk

reconstructie methode slechts toegepast op één of twee datasets en standaard parameterwaardes worden dan gekozen op basis van deze dataset. Toepassing op andere data betekent meestal dat andere parameterwaardes nodig zullen zijn. Informatie over de invloed van de parameters is dus cruciaal. Een tweede mogelijk probleem is dat de originele implementatie van het algoritme zoals beschreven in het bijhorende artikel niet altijd beschikbaar is. Als er toch een versie beschikbaar is, kan het zijn dat deze versie sterk verschilt van de versie in het originele artikel, of dat de implementatie fouten bevat.

Resultaten van de verschillende methodes moeten voorzichtig vergeleken worden. Alle methodes verschillen van elkaar in de manier waarop ze de data benaderen en de biologie modelleren. Sommige methodes gebruiken bijvoorbeeld enkel expressiedata terwijl andere methodes informatie van verschillende databronnen combineren. De resultaten zullen in dat geval niet enkel afhangen van de methode maar ook van de gebruikte datasets.

Vermits sommige methodes het regulatorische netwerk rechtstreeks reconstrueren terwijl andere regulatorische modules zoeken, is het ook niet vanzelfsprekend om resultaten van deze twee klasses van methodes te vergelijken. Zelfs de resultaten van twee module detectie methodes vergelijken is niet voor de hand liggend. Sommige methodes vinden modules die bestaan uit slechts enkele genen, terwijl andere methodes modules met veel genen vinden. Dit wil echter niet altijd zeggen dat één van beide methodes beter is dan de andere. Meestal duidt dit op het feit dat ze verschillende aspecten van de data belichten, of de modulariteit op een verschillende resolutie beschrijven.

Met al deze moeilijkheden en uitdagingen moet rekening gehouden worden opdat een eerlijke vergelijking van de methodes mogelijk zou zijn. Veel tijdschriften vereisen echter dat de nieuwe methode beter presteert dan reeds gepubliceerde methodes, wat soms leidt tot oneerlijke vergelijkingen.

Hoofdstuk 5: Association rule mining algoritmes in de bioinformatica

In het vorige hoofdstuk werden verscheidene methodes voor de inferentie van netwerken beschreven. *Association rule mining* (ARM) algoritmes worden zelden gebruikt voor netwerk reconstructie hoewel zij een aantal veelbelovende voordelen lijken te bieden (zie hoofdstuk 4). De tijdens dit doctoraat ontwikkelde methodes voor netwerk reconstructie maken gebruik van deze ARM algoritmes.

ARM algoritmes werden initieel ontwikkeld voor de analyse van *transacties* die worden opgeslaan in een databank. Dergelijke analyse van aankopen

helpt bijvoorbeeld het huidige management van een winkel om te beslissen welke produkten best in hetzelfde rek geplaatst worden. ARM algoritmes zijn dus erg nuttig om grote hoeveelheden van data te analyseren.

Vermits het doel van deze thesis erin bestaat om het transcriptionele netwerk te ontrafelen, zou het interessant kunnen zijn om bijvoorbeeld te weten welke genen één of meerdere motieven gemeenschappelijk hebben. Stel dat er een databank bestaat waarin alle genen en hun motieven zijn opgeslaan. ARM algoritmes kunnen op basis van deze databank op heel efficiënte manier alle sets van genen vinden die dezelfde motieven hebben. Volgens de terminologie van ARM algoritmes wordt hierbij een gen een *item* genoemd, terwijl het motief een *transactie* is. Een set van genen die dezelfde motieven bevatten in hun promotergebieden is een *itemset*. Het aantal motieven dat deze genen gemeenschappelijk hebben is de *support* van deze itemset. Een itemset wordt *frequent* genoemd als zijn support groter is dan een bepaalde minimum waarde: de *support constraint*. Een frequente itemset wordt *maximaal* genoemd als het geen subset is van een andere frequente itemset. Een frequente itemset is *closed* als de itemset geen subset is van een andere frequente itemset met dezelfde transacties.

Alle mogelijke combinaties van itemsets zouden kunnen getest worden om te kijken welke itemset frequent, maximaal of closed is. Voor middelgrote of grote databanken is deze benadering echter niet mogelijk en moeten we ons baseren op efficiënte algoritmes zoals de ARM algoritmes. In een eerste stap zoeken deze algoritmes naar de frequente itemsets. Vervolgens worden deze itemsets gebruikt om associatie-regels mee af te leiden. Voor het onderzoek in deze thesis is echter enkel de identificatie van de frequente itemsets van belang.

Het Apriori en CHARM algoritme

Twee ARM algoritmes zijn belangrijk voor de netwerk reconstructie methodes die werden ontwikkeld tijdens dit doctoraat: het Apriori algoritme (Agrawal and Imielenski, 1993) en het CHARM algoritme (Zaki and Hsiao, 2002).

Het Apriori algoritme is het eerste en meest bekende ARM algoritme. Dit algoritme zoekt naar frequente itemsets. Het begint hierbij met sets bestaande uit één item die voldoen aan een vooropgestelde *support constraint*. Vervolgens worden deze items gecombineerd tot frequente itemsets van twee items. Vervolgens worden deze itemsets weer gecombineerd tot frequente itemsets bestaande uit drie items, etc. Het Apriori algoritme maakt gebruik van het feit dat een itemset alleen maar frequent kan zijn als al zijn subsets ook frequent zijn. Dat wil zeggen dat enkel wanneer alle sub-itemsets frequent zijn, er voor een nieuwe itemset

moet nagegaan worden of deze ook frequent is. Indien daarentegen niet alle subsets frequent zijn, dan zal de nieuwe itemset ook nooit frequent kunnen zijn en moet deze bijgevolg nooit geanalyseerd worden. Door gebruik te maken van deze overerfbaarheid kan Apriori op zeer efficiënte wijze de frequente itemsets vinden.

Het CHARM algoritme zoekt naar closed itemsets. Een van de grote verschillen met het Apriori algorime is dat CHARM de itemset- en transactieruimte tegelijkertijd doorzoekt in een zogenaamde IT of itemset-transactie boom. In deze boomstructuur bestaat een node uit een itemset-transactie paar. CHARM doorzoekt deze boomstructuur aan de hand van een *depth-first* strategie (Apriori daarentegen gebruikt een breadth-first strategie) en maakt gebruik van de notie van equivalentie klassen. In deze IT boom is elke node eigenlijk een prefix-gebaseerde klasse. Twee itemsets behoren tot dezelfde klasse als ze eenzelfde prefix van lengte *k* hebben. Deze prefix is in feite een geordende lijst van *k* gennamen. Een klasse bestaat dus uit de items waarmee de prefix kan uitgebreid worden om een nieuwe frequente node te bekomen. Geen enkele boom die begint bij een infrequente node moet verder onderzocht worden. De kracht van deze methode is dat het de originele zoekruimte opsplitst in kleinere onafhankelijke problemen. Elke boom beginnende bij een bepaalde node kan als een nieuw probleem gezien worden en enkel dit stukje moet in het geheugen passen.

Toepassingen in de bioinformatica

ARM algoritmes worden in toenemende mate gebruikt in de bioinformatica. Een aantal toepassingen maakt gebruik van het Apriori algoritme (Chiu *et al*, 2006; Ivan *et al*, 2007; Morgan *et al*, 2007; Oyama *et al*, 2002) of Apriori-gebaseerde algoritmes (Artamonova *et al*, 2005; Artamonova *et al*, 2007; Becquet *et al*, 2002; Brazma *et al*, 1997; Carmona-Saez *et al*, 2006; Creighton and Hanash, 2003; Rodriguez *et al*, 2005). Andere toepassingen zoeken naar closed itemsets (Huang *et al*, 2007; Okada *et al*, 2007; Pham *et al*, 2004). Hoewel de meeste toepassingen in de bioinformatica gebruik maken van reeds bestaande ARM algoritmes, worden er soms ook nieuwe algoritmes ontwikkeld om de specifieke eigenschappen van een bepaald biologisch probleem in rekening te kunnen brengen (Georgii *et al*, 2005; Lopez *et al*, 2008; Tamura and D'haeseleer, 2008).

ARM algoritmes werden voor de meest uiteenlopende bioinformatica toepassingen gebruikt: om relaties te zoeken tussen HIV-behandelingen en HIV mutaties (lin *et al*, 2006; Tamura and D'haeseleer, 2008), voor de annotatie van proteïnen (Chiu *et al*, 2006; Rodriguez *et al*, 2005), voor de studie van proteïne-proteïne interacties (Ivan *et al*, 2007; Oyama *et al*, 2002), om combinaties van regulatorische motieven in de promotergebieden

van genen te identificeren (Brazma *et al*, 1997; Doi *et al*, 2008; Morgan *et al*, 2007) of voor de analyse van expressiedata (Becquet *et al*, 2002; Carmona-Saez *et al*, 2006; Creighton and Hanash, 2003). Ondanks deze toepassingen worden ARM algoritmes niet vaak gebruikt voor netwerk reconstructie.

Hoofdstuk 6: Het transcriptionele netwerk van gist

Het doel van deze thesis is het bestuderen van het transcriptionele netwerk. In dit hoofdstuk bespreken we de module detectie methode ReMoDiscovery die ontwikkeld werd in samenwerking met Professor Tijl De Bie (University of Bristol, UK). Op basis van expressiedata, regulatorische motieven en proteïne-DNA interacties kan ReMoDiscovery sets van genen identificeren die samen tot expressie komen en die onder controle staan van hetzelfde regulatorische programma. Om de identificatie van dergelijke interessante genen mogelijk te maken baseert ReMoDiscovery zich op het Apriori algoritme (Lemmens *et al*, 2006).

ReMoDiscovery

Onze methode voor netwerk reconstructie, ReMoDiscovery, bestaat uit twee stappen: een *seed discovery* stap en een *seed extension* stap.

Tijdens de *seed discovery* stap vindt ReMoDiscovery zogenaamde *seed* modules. Deze modules bestaan uit genen met een gelijkaardig expressieprofiel (hoge correlatiecoëfficiënt) en die onder controle staan van eenzelfde regulatorisch programme met een minimum aantal regulatoren en een minimum aantal regulatorische motieven. Sets van genen die aan deze strenge eisen voldoen en die niet meer uitgebreid kunnen worden met andere genen zonder deze eisen te overtreden worden *seed* modules genoemd. Deze sets van genen worden gezocht aan de hand van het Apriori algoritme.

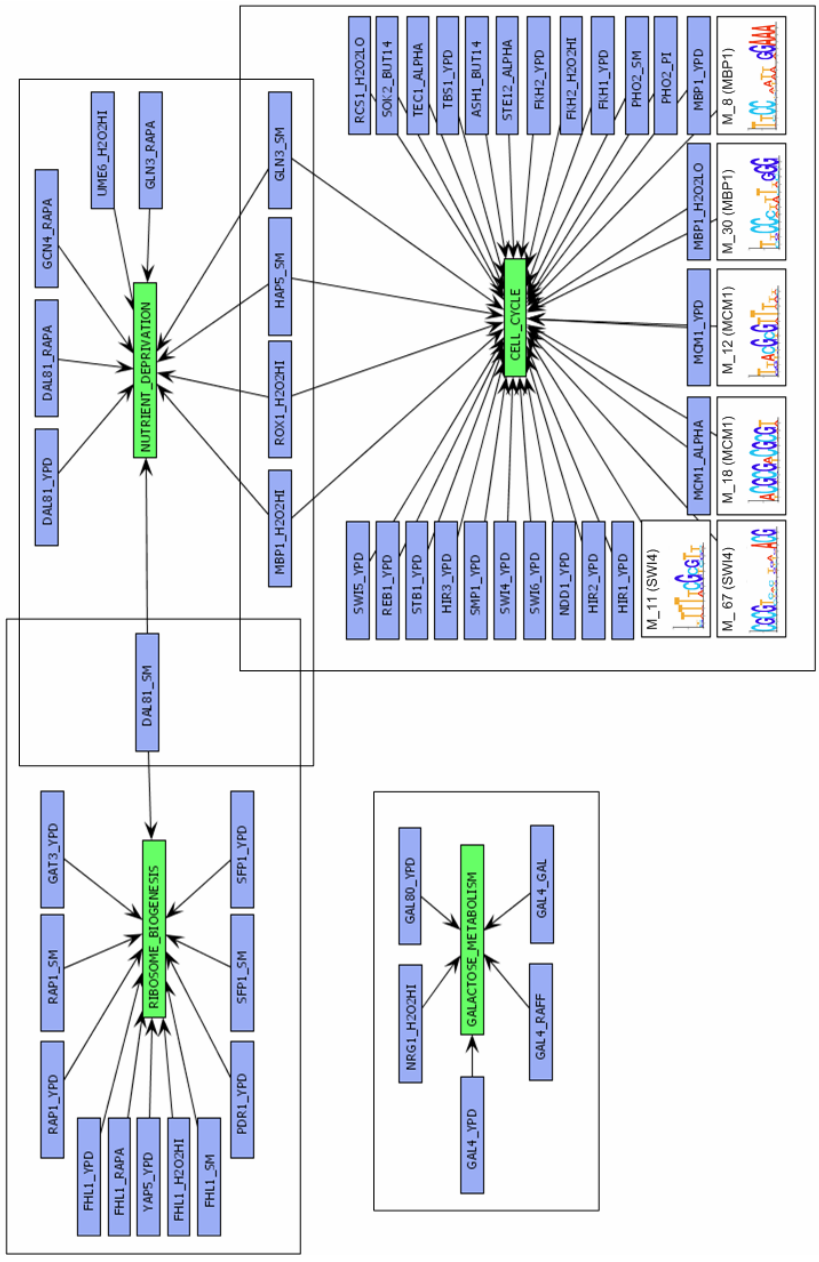
De strenge criteria om de seed modules te bepalen blijken voldoende om regulatoren en motieven op een betrouwbare manier te vinden. De maximale geninhoud van deze modules is echter waarschijnlijk onderschat door deze strikte criteria. Daarom zal in een tweede *seed extension* stap, de geninhoud van de modules uitgebreid worden.

Identificatie van regulatorische modules

ReMoDiscovery werd toegepast op een dataset die genexpressie meet tijdens de cel cyclus (Spellman *et al*, 1998) en op een expressiedataset die genexpressie meet tijdens verschillende stress-gerelateerde experimentele condities (Gasch *et al*, 2000), telkens in combinatie met ChIP-chip data (Harbison *et al*, 2004) en regulatorische motieven (Kellis *et al*, 2003).

In de eerste stap van het algoritme werden 20 regulatorische modules geïdentificeerd in de cel cyclus expressiedata (Spellman *et al*, 1998) (zie ook Figuur N - 3) en 104 modules in de stress-gerelateerde expressiedata (Gasch *et al*, 2000). Om de biologische relevantie van onze resultaten te verifiëren, vergeleken we ze met wat reeds gekend is in de literatuur. We beschouwden een module als 'bevestigd' door de literatuur indien alle regulatoren betrokken waren bij hetzelfde biologische proces volgens de literatuur. In totaal werden 15 van de 20 en 53 van de 104 modules bevestigd.

Een groot aantal van de *seed* modules (18 van de 20 en 63 van de 104) konden uitgebreid worden in de tweede stap van het algoritme. Voor al deze uitgebreide modules geldt dat de regulatoren in de module betrokken zijn bij hetzelfde biologische proces als de genen in de module. In sommige gevallen konden de *seed* modules echter niet uitgebreid worden: geen additionele gecorreleerde genen waren blijkbaar aanwezig in de dataset. Dit wijst erop dat ofwel de modulegrootte zeer klein is ofwel het regulatorische programma van de module niet actief was in de condities die getest werden in de expressiedata. De identificatie van het regulatorische programma in de seed discovery stap is grotendeels afhankelijk van de ChIP-chip and motif data. Vermits motiefdata conditie-onafhankelijk zijn betekent de aanwezigheid van een gemeenschappelijk motief niet noodzakelijk dat de genen samen tot expressie moeten komen in de condities die getest werden in een bepaalde dataset. Omwille van het verschil in experimentele condities waarin de ChIP-chip en expressiedata getest werden, wordt informatie van ChIP-chip data niet noodzakelijk bevestigd door alle expressiedata. Daarom kan een module enkel uitgebreid worden met additionele genen als zijn regulatorische programma actief is in de condities van het microroosterexperiment. Gebaseerd op deze observatie werden de modules opgedeeld in modules die actief zijn in de twee datasets of in slechts één van de twee datasets. De eerste groep van modules zijn betrokken bij meer algemene functies zoals ribosoom synthese, terwijl de tweede groep van modules betrokken zijn in processen zoals cel cyclus, stress, respiratie, meiose, aminozuurmetabolisme of nutriënt tekort.



Figuur N - 3: legende zie pagina xxix

Vergelijking met andere methodes

Om de verschillen te bepalen tussen ReMoDiscovery en eerder gepubliceerde algoritmes voor module detectie pasten we enkele van de goed gekende methodes (i.e. SAMBA (Tanay *et al*, 2004) and GRAM (Bar-Joseph *et al*, 2003)) toe op de cel cyclus expressie data (Spellman *et al*, 1998) in combinatie met ChIP-chip data (Harbison *et al*, 2004).

We analyseerden de tijd die de methodes nodig hebben om resultaten te genereren op basis van bovenvermelde datasets. SAMBA en ReMoDiscovery zijn eerder snelle methodes in vergelijking met GRAM. Indien er genen in de dataset waren met meer dan 10 significante ChIP-chip interacties, kon GRAM niet toegepast worden. Na deze genen uit de dataset gefilterd te hebben, bleek GRAM nog altijd trager dan ReMoDiscovery of SAMBA.

Om de gen- en regulatorinhoud van de modules verkregen door GRAM, SAMBA en ReMoDiscovery te vergelijken, gebruikten we een Jaccard coëfficiënt score voor geninhoud en regulatorinhoud apart. Voor een bepaalde parameterwaarde keken we voor elk paar van genen (regulatoren) of deze samen voorkwamen in minstens één module. Door dit te doen voor elk paar van genen (regulatoren) en voor elke methode konden het aantal genparen (regulatorparen) dat samen voorkomt in beide methodes bepaald worden. Vermits parameterwaardes de samenstelling van de modules beïnvloeden, hebben we de Jaccard coëfficiënt voor verschillende waardes bepaald.

Figuur N- 3: Overzicht van de seed modules die werden geïdentificeerd in de cel cyclus dataset (Spellman *et al*, 1998). Regulatoren en motieven die horen bij een regulatorisch programma van een module die uitgebreid kon worden zijn in de figuur aangeduid aan de hand van een vetgedrukte pijl van die regulator of motief naar de module. In de cel cyclus dataset werden modules geïdentificeerd waarvan de genen voornamlijk in vier biologische processen betrokken zijn: ribosome biogenese, reactie bij tekort aan nutriënten, cel cyclus en galactose metabolisme. De meeste modules waren echter betrokken in de cel cyclus, wat logisch is aangezien de genexpressie in de data gemeten was tijdens verschillende stadia van de cel cyclus. Sommige regulatoren reguleren genen die betrokken zijn bij meerdere processen. Dal81 bijvoorbeeld, reguleert genen betrokken in ribosoom biogenese maar ook genen betrokken in stress omwille van een tekort aan nutriënten.

Voor alle parameterwaardes bleek de overlap in zowel gen- als regulatorinhoud tussen de modules van GRAM en ReMoDiscovery groot te zijn. In het algemeen vindt GRAM regulatorische modules met een kleiner aantal regulatoren maar groter aantal genen. De overlap met ReMoDiscovery was dan ook het grootst voor meer stringente ChIP-chip parameterwaardes en lagere minimum correlatiewaarde in ReMoDiscovery. GRAM en ReMoDiscovery kunnen dus gelijkaardige patronen vinden in de datasets. De regulatorische programma's die geïdentificeerd werden door SAMBA leken niet op die van ReMoDiscovery of GRAM, maar de geninhoud wel. Dezelfde analyse kan uitgevoerd worden op de uitgebreide modules in plaats van de *seed* modules. De uitgebreide modules resulteren in een kleiner aantal regulatoren per module maar een groter aantal genen. De uitgebreide modules van ReMoDiscovery lijken dan nog sterker op de modules van GRAM.

Alle module detectie methodes in deze studie beogen hetzelfde tweeledige doel: de identificatie van een set van genen die samen tot expressie komen en het overeenkomstige regulatorische programma. Om het eerste aspect te evalueren hebben we de gemiddelde functionele aanrijking berekend van de modules die geïdentificeerd werden door elk van de methodes bij standaard parameterwaardes op onze dataset. Uit deze resultaten blijkt dat de modules inderdaad aangerijkt zijn voor gekende functionele klassen. Voor ReMoDiscovery verbeterde de aanrijking bij uitbreiding van de seed modules.

Vervolgens gingen we na hoe goed de verschillende methodes het regulatorische programma van de modules kunnen bepalen. Vermits onze dataset bestond uit cel cyclus expressiedata, stelden we eerst een lijst op van alle gekende cel cyclus regulatoren. Voor elk van deze regulatoren werd geteld hoe vaak ze voorkwamen in de regulatorische programma's van modules die functioneel aangerijkt waren voor cel cyclus. Er werd ook gekeken naar de verhouding van het aantal gekende cel cyclus regulatoren ten opzichte van het totaal aantal regulatoren in een module, uitgemiddeld over alle modules. De resultaten tonen aan dat zowel ReMoDiscovery als GRAM beter zijn in het detecteren van cel cyclus gerelateerde regulatoren dan SAMBA. Dit besluit moet met enige voorzichtigheid geïnterpreteerd worden omdat het transcriptionele netwerk van gist nog niet volledig gekarakteriseerd is en het dus niet mogelijk is om de echte positieve interacties van de vals positieve te onderscheiden zonder verdere experimentele validatie.

Conclusie

ReMoDiscovery is een twee-staps-methodologie om actieve modules te detecteren op basis van drie onafhankelijke databronnen. De methode is gebaseerd op het association rule mining algoritme Apriori. Aan de hand van dit algoritme kunnen de regulatorische modules op zeer snelle en efficiënte wijze geïdentificeerd worden.

ReMoDiscovery werd toegepast op cel cyclus en op stress-gerelateerde expressiedata. Biologisch relevante regulatorische modules werden geïdentificeerd. De modules waren betrokken in zeer uiteenlopende processen zoals cel cyclus, ribosoom biogenese, galactose metabolisme, aminozuur metabolisme of respiratie. In vergelijking met andere methodes voor de detectie van modules presteerde ReMoDiscovery beter op de bestudeerde dataset.

ReMoDiscovery is dus in staat om biologisch relevante resultaten te genereren en presteert tegelijkertijd beter dan gerelateerde methodes voor module detectie. Hiermee konden we aantonen dat methodes gebaseerd op association rule mining algoritmes geschikt zijn voor de identificatie van regulatorische modules.

Hoofdstuk 7: Studie van de conditie-afhankelijke combinatoriële regulatie in *E. coli*

Hoewel we in hoofdstuk 6 reeds aantoonden dat ReMoDiscovery biologisch relevante resultaten kan genereren op kleinere expressie datasets, wordt een selectie van de condities of biclusterstrategie noodzakelijk als we alle beschikbare microrooster data willen gebruiken. In samenwerking met professor De Bie (University of Bristol, UK) werd daarom een tweede data integratie methode ontwikkeld onder de naam DISTILLER. DISTILLER is gebaseerd op het association rule mining algoritme CHARM (Zaki and Hsiao, 2002) en laat een selectie van de condities toe. Hoewel ReMoDiscovery en DISTILLER beide zoeken naar sets van genen die samen tot expressie komen en hun regulatorische programma, zoekt DISTILLER dus ook naar de experimentele condities waarin de genen samen tot expressie komen. Dankzij deze biclusterstrategie moeten genen in een module niet langer samen tot expressie komen in alle condities, maar slechts in een gedeelte van de condities.

De onderzoeksgroep van professor Collado-Vides (UNAM, Cuernavaca, Mexico) bood ondersteuning bij het gebruik van de databank RegulonDB (Salgado *et al*, 2006). RegulonDB bevat alle gekende informatie over het transcriptionele netwerk van *Escherichia coli*. Uit deze databank blijkt dat

meer dan 50% van de genen combinatorieel gereguleerd zijn. Desalniettemin komt het bewijs voor elk van de interacties vaak van afzonderlijke studies. Deze studies kunnen in verschillende experimentele condities uitgevoerd zijn. De huidige manieren om netwerken voor te stellen brengen deze conditie-afhankelijkheid niet in rekening. Uit deze statische netwerken is het bijgevolg niet duidelijk of twee regulatoren die hetzelfde gen reguleren samenwerken in dezelfde condities of onafhankelijk van elkaar functioneren in verschillende experimentele condities (Balaji *et al*, 2006; Balaji *et al*, 2007). De relatie tussen modulariteit, combinatoriële regulatie en conditie-afhankelijkheid van bacteriele netwerken werd bestudeerd met 'DISTILLER' (Data Integration System To Identify Links in Expression Regulation).

DISTILLER

De kern van ons raamwerk is een data integratie strategie die gebaseerd is op een ARM algoritme. De methode bestaat uit 3 stappen: (1) de identificatie van seed modules; (2) filteren van de seed modules tot een set van niet-redundante modules; en (3) de uitbreiding van de seed modules met additionele genen.

In een eerste stap worden seed modules geïdentificeerd die bestaan uit een set van genen die samen tot expressie komen in een set van experimentele condities en die minstens één regulatorisch motief gemeenschappelijk hebben. Het uitbreiden van ARM algoritmes om een selectie van condities toe te laten is niet triviaal, onder meer omdat voor de hand liggende afstandsmaten om co-expressie te bepalen zoals correlatie niet zomaar toepasbaar zijn in de context van 'association rule mining'. Een nieuwe afstandsmaat moest daarom bepaald worden. Vermits deze strategie voor conditie selectie het probleem van module detectie combinatorieel nog moeilijker maakt, baseert DISTILLER zich op het CHARM algoritme (Zaki and Hsiao, 2002), dat toelaat om de modules op nog efficiëntere wijze te identificeren en om het aantal parameters te beperken.

In een tweede stap moeten de geïdentificeerde modules gefilterd worden. Een voordeel van de ARM algoritmes is dat ze alle mogelijke oplossingen onderzoeken. Het resultaat is dan een lange lijst van mogelijk interessante modules. Vermits deze modules geen expliciete score hebben, is het niet duidelijk welke modules het meest interessant zijn om eerst te onderzoeken. Daarenboven kunnen de modules ook een sterke overlap vertonen, bijvoorbeeld als twee modules slechts verschillen in enkele condities. Daarom introduceerden we tijdens dit doctoraat een iteratieve procedure die de meest interessante modules één voor één selecteert. Deze methode brengt de significantie van de individuele module in rekening maar straft

tegelijkertijd overlap met andere eerder geselecteerde modules af. Het resultaat is een lijst van significante en biologisch interessante modules.

Tijdens de derde stap van de methode worden additionele genen geselecteerd die niet voldeden aan de strenge criteria van de eerste seed discovery stap.

Identificatie van regulatorische modules

In deze studie werd DISTILLER toegepast op twee *E. coli* datasets: een compendium van expressiedata en een compendium van regulatorische motieven (zie chapter 3). Door motief en expressie data te integreren kon DISTILLER 150 conditie-afhankelijke regulatorische modules detecteren. Uit elke module kunnen de regulator-doelwitgen interacties afgeleid worden door elk gen te verbinden met de regulator die bindt op het motief dat alle genen van de module gemeenschappelijk hebben.

In de modules zijn zo in totaal 732 afzonderlijke interacties aanwezig voor 67 regulatoren. Van deze 732 interacties waren 454 interacties reeds gekend in RegulonDB (Salgado *et al*, 2006), wat overeenkomt met 62% van de totaal aantal gekende interacties in RegulonDB. Voor 37 van de 67 regulatoren kon op zijn minst een gedeelte van hun regulon bevestigd worden door DISTILLER. Voor de andere 30 regulatoren werden geen interacties gevonden. Waarschijnlijk was het aantal genen in de modules kleiner dan het vereiste aantal genen of waren de condities waarin de interacties voorkomen niet aanwezig in het expressiecompendium. Naast de 454 eerder beschreven interacties vond DISTILLER ook 278 nieuwe interacties. Voor vele gekende regulatoren kon het gekende gedeelte van hun regulon nog uitgebreid worden.

Om aan te tonen dat onze methode goede interacties kan voorspellen werden in samenwerking met Dr. I. Thijs en Dr. S. De Keersmaecker (CMPG, KU Leuven) nieuwe interacties voor de regulator FNR experimenteel gevalideerd via ChIP-qPCR. Er werden 11 nieuwe doelwitgenen (*ydhY*, *yfgG*, *hscC*, *treF*, *yjhB*, *ydjX*, *yjtD*, *ydaT*, *yehD*, *yhjA* en *ftnB*) getest. Al deze genen werden inderdaad gebonden door FNR.

Conditie-afhankelijke transcriptie

De microroosters in het compendium werden gegroepeerd in conditionele categorieën gebaseerd op de experimentele opzet. De categorie aeroob-anaeroob, bijvoorbeeld, bevat alle microroosters waarin het effect van een verandering in zuurstofniveau gemeten wordt. Vervolgens berekenden we in hoeverre de condities van de modules van een bepaalde regulator aangerijkt zijn voor een specifieke conditionele categorie. Aanrijking van een

conditionele categorie duidt erop dat de doelwitgenen van een bepaalde regulator vooral samen tot expressie komen in condities die tot de aangerijkte categorie behoren. Dit geeft indirect informatie over de condities waarin een regulator actief is. De meeste regulatoren in onze resultaten bleken inderdaad actief te zijn in condities die overeenkomen met hun annotatie, wat erop wijst dat onze conditie selectie relevant is.

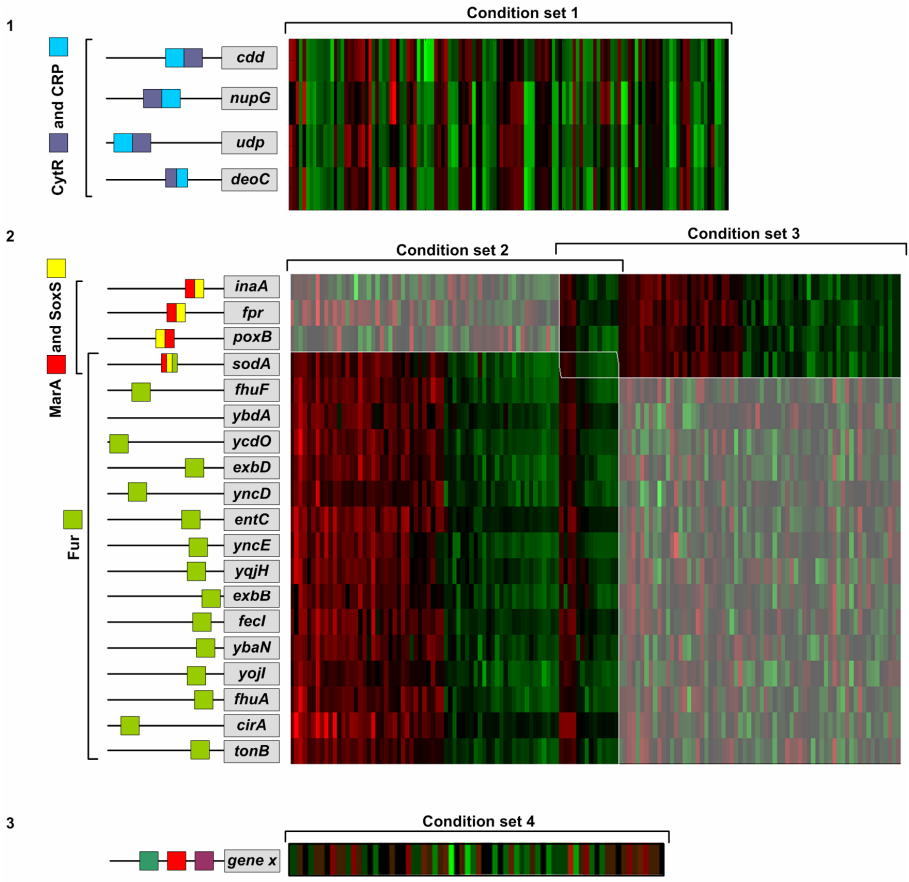
Experimentele condities die grote veranderingen in de energiestatus van de cel veroorzaken, zoals veranderingen in zuurstofconcentratie of pH, hebben het meest uitgesproken effect op transcriptionele regulatie. Andere condities blijken enkel een effect te hebben op zeer specifieke biologische processen zoals de Fe^{2+} concentratie of het toedienen van schade aan DNA.

Globale regulatoren zoals ArcA, CRP, Fis, FNR, IHF of Lrp zijn duidelijk betrokken in meerdere conditionele categorieën en zijn verantwoordelijk voor de algemene cellulaire reactie. Meer specifieke regulatoren daarentegen zijn belangrijk voor een meer verfijnde reactie.

Combinatoriële regulatie

DISTILLER identificeert ook het niveau van combinatoriële regulatie van de doelwitgenen in elke module. Met combinatoriële regulatie wordt hier een set van genen bedoeld die gereguleerd worden door meerdere regulatoren, onafhankelijk van het feit of dat deze regulatoren complexe interacties hebben of onafhankelijk van elkaar functioneren. In RegulonDB (Salgado *et al*, 2006) zijn 42 operons beschreven die door één regulator gereguleerd worden, 66 operons door twee en 70 operons door drie of meerdere regulatoren. Onze modules vertonen niet dezelfde regulatorische complexiteit: slechts 25 van de 150 modules staan onder controle van twee of meer regulatoren. Voor 24 modules bestaat het regulatorische programma uit minstens één globale regulator (Figuur N - 4).

Hoewel op niveau van de modules er inderdaad een laag niveau van combinatoriële regulatie is, kan de expressie van individuele genen beïnvloed worden door meer dan één regulator. Wij identificeerden 85 zogenaamde *connector* genen in onze modules. Dit zijn individuele genen die voorkomen in meerdere modules en waarvan de modules onder controle staan van telkens andere regulatoren. Modules die een connector gen gemeenschappelijk hebben vertonen meestal weinig overlap in de condities, hetgeen erop wijst dat slechts één regulator voldoende is om de expressie van een connector gen te veranderen bij een specifieke experimentele verandering. Voor nog andere genen is het expressiegedrag zodanig specifiek dat het nooit gelijkaardig is aan dat van een voldoende aantal andere genen, zodat deze genen nooit in modules gevonden worden door DISTILLER.



Figuur N - 4: Drie types van combinatoriële regulatie. Type 1 toont combinatoriële regulatie op het niveau van modules. De genen *cdd*, *nupG*, *udp* en *deoC* hebben telkens twee motieven gemeenschappelijk en komen tot expressie in conditieset 1. Type 2 toont combinatoriële regulatie op het niveau van een 'connector' gen. Alle genen van module 1 hebben twee motieven, MarA en SoxS, gemeenschappelijk en komen samen tot expressie in conditieset 2. Alle genen van module 2 zijn gereguleerd door Fur. Het gen *sodA*, een connector gen, is aanwezig in beide modules en wordt gereguleerd door de regulatoren van zowel module 1 als module 2 maar in verschillende condities. Dit wijst erop dat de regulatoren van beide modules onafhankelijk van elkaar functioneren. Zowel type 1 als type 2 van combinatoriële regulatie kunnen door DISTILLER teruggevonden worden. In het geval dat een conditie-specifieke complexe interacties tussen regulatoren ervoor zorgen dat een zeer specifiek expressiepatroon verkregen wordt en dat er geen andere genen ditzelfde patroon vertonen (type 3), kan dergelijke interactie niet door DISTILLER gevonden worden.

Vergelijking met andere methodes

Tenslotte hebben we de resultaten van DISTILLER vergeleken met resultaten van twee recent gepubliceerde netwerk reconstructie methodes. Op deze manier konden we de betrouwbaarheid van onze resultaten en de complementariteit van de verschillende methodes evalueren. We kozen voor de CLR methode (Faith *et al*, 2007) en SEREND (Ernst *et al*, 2008) omdat beide methodes initieel ook werden toegepast op *E. coli* (zie ook hoofdstuk 4). CLR werd toegepast op een compendium van 445 Affymetrix microroosters, terwijl SEREND werd toegepast op ditzelfde compendium maar in combinatie met motiefinformatie van RegulonDB.

De interacties geïdentificeerd door CLR en DISTILLER vertonen een beperkte overlap: slechts 21 gekende en 6 nieuwe interacties werden gevonden door beide methodes. Terwijl CLR voor een groter aantal regulatoren interacties kan detecteren dan DISTILLER, is CLR minder goed in het terugvinden van alle interacties voor een gegeven regulator. DISTILLER kan zo bijvoorbeeld gemakkelijker interacties identificeren voor globale regulatoren, terwijl CLR ook interacties voor genen met een heel specifiek expressiepatroon kan vinden zolang het expressieprofiel van regulator en doelwitgen gerelateerd zijn. De complimentariteit van beide methodes was te verwachten vermits beide methodes gebaseerd zijn op fundamenteel verschillende principes.

Van 981 nieuwe interacties geïdentificeerd door SEREND en de 278 door DISTILLER werden 78 interacties gevonden door beide methodes. Het feit dat beide algoritmes dezelfde nieuwe interacties voorspellen hoewel ze gebaseerd zijn op volledig verschillende methodes en datasets wijst erop dat beide geschikt zijn om significante nieuwe signalen in de dataset te detecteren.

Conclusies

DISTILLER is een data integratie raamwerk voor de detectie van regulatorische modules. Het is gebaseerd op het CHARM algoritme. DISTILLER zoekt naar sets van genen die samen tot expressie komen in een set van condities en die minstens één motief gemeenschappelijk hebben in hun promotergebieden. Door de biclusterstrategie van onze netwerk reconstructie methode kon alle beschikbare microrooster data optimaal gebruikt worden.

Door DISTILLER toe te passen op publiek beschikbare regulatorische motief data en microrooster data kon de relatie tussen conditie-afhankelijke co-expressie en combinatoriële regulatie in het transcriptionele netwerk van *E. coli* geanalyseerd worden. Er werden 62% van 736 gekende interacties in

RegulonDB bevestigd en 278 nieuwe interacties voorspeld. Door nieuwe doelwitgenen van de zeer goed bestudeerde regulator FNR experimenteel te valideren toonden we aan dat onze methode betrouwbare voorspellingen doet. We konden ook de conditie-afhankelijkheid van vele gekende regulatoren bevestigen zoals ArcA of FNR. Voor minder gekende regulatoren zoals CpxR konden we de condities waarin ze mogelijk actief zijn voorspellen. Daarnaast ontdekten we dat het niveau van regulatorische complexiteit van de modules eerder laag leek in vergelijking met wat verwacht werd van RegulonDB. Dit duidt erop dat complexe regulatorische programma's de graad van modulariteit verlagen en kunnen resulteren in meer specifieke expressieprofielen.

Hoofdstuk 8: Conclusies en vooruitzichten

In deze thesis lag de nadruk op de detectie van transcriptionele modules, i.e. de basiseenheden van het regulatorische netwerk. Deze regulatorische modules bestaan uit sets van genen die samen tot expressie komen en die hetzelfde regulatorische programma hebben. Het regulatorische programma bestaat uit een set van regulatoren en de overeenkomstige regulatorische motieven. Tijdens dit doctoraat werd een methodologie ontwikkeld voor de detectie van modules gebaseerd op association rule mining algoritmes. Dit resulteerde in een set van state-of-the-art module detectie en data integratie methodes. Door deze methodes toe te passen op biologische data konden we aantonen dat association rule mining algoritmes even geschikt zijn voor netwerk reconstructie als de meer gebruikte graaf-gebaseerde of probabilistische methodes. De voornaamste bijdragen van deze thesis zijn:

- De ontwikkeling van een methode voor de geautomatiseerde verzameling van data wat resulteerde in een microrooster compendium voor *E. coli*.
- De toepassing van ReMoDiscovery op gist data waarbij aangetoond werd dat ARM algoritmes geschikt zijn voor de reconstructie van netwerken.
- De toepassing van DISTILLER op *E. coli* data waarbij de conditie-afhankelijke combinatoriële regulatie kon bestudeerd worden.

Vooruitzichten

In deze thesis werden reeds gekende regulatorische motieven gebruikt. Naar de toekomst toe zouden ook nieuwe motieven kunnen geïntegreerd worden in de analyse. Aan de hand van *in silico* motief detectie werden in het

CMPG/BioI (KU Leuven) voor zowel *E. coli* (Dr. P. Monsieurs) als *B. subtilis* (A. Fadda) nieuwe motieven voorspeld. Bovendien kunnen na de toepassing van DISTILLER op beide organismen, de regulatorische netwerken vergeleken worden. Dergelijke analyses zouden toelaten nieuwe interessante biologische ontdekkingen en bevindingen te doen.

Naast het gebruik van motief, ChIP-chip en microrooster data kunnen ook andere databronnen in de analyse opgenomen worden. Small RNA's bijvoorbeeld spelen ook een belangrijke rol in transcriptionele regulatie. Tot nu toe werden regulatie-mechanismen via deze small RNA's echter nog niet op genoomwijde schaal bestudeerd.

Naast deze conceptuele aanpassingen zijn er natuurlijk ook nog meer algoritmische verbeteringen mogelijk. Zo is bijvoorbeeld een gebruiksvriendelijke interface van DISTILLER noodzakelijk. DISTILLER zou ook rechtstreeks aan het microroostercompendium kunnen gekoppeld worden. Op deze manier wordt telkens met de meest up-to-date versie van het compendium gewerkt. De filterstap voor het samenvatten van de resultaten zou ook nog verbeterd kunnen worden.

Van fundamenteel onderzoek naar toepassingen van de systeembiologie

De systeembiologie heeft als doel het biologische systeem als geheel te karakteriseren. Het is dus essentieel om te begrijpen hoe de verschillende biologische componenten (zoals DNA, RNA, proteïnen en metabolieten) samenwerken om het organisme leefbaar te maken, m.a.w. om het onderliggende regulatorische netwerk te ontrafelen. De ontwikkeling van methodes voor de detectie van transcriptionele modules door de integratie van verschillende databronnen is dus zeer belangrijk in dit systeembiologie tijdperk vermits het helpt om het volledige regulatorische netwerk beter te begrijpen.

De methodes ontwikkeld in deze thesis werden toegepast op *E. coli* en gist. Het zijn echter generische methodes die ook op hogere eukaryoten kunnen toegepast worden. De verkregen kennis zou dan kunnen leiden tot toepassingen in medische domeinen (Chuang *et al*, 2007; Ibrahim and Gold, 2005; Joyce and Palsson, 2006; Khalil and Hill, 2005) en in de ontwikkeling van medicijnen (Kitano, 2002).

Het begrijpen van de onderliggende principes van het biologische systeem laat ook toe om deze kennis te gebruiken om organismen te veranderen zodat het organisme andere nieuwe taken kan uitvoeren. De produktie van terpenoiden, die gebruikt kunnen worden als anti-malaria medicijn, door *E.coli* (Martin *et al*, 2003) of *S. cerevisiae* (Aebersold *et al*, 2000; Lindahl *et*

al, 2006) is een voorbeeld van hoe organismen kunnen aangepast worden om nieuwe taken te vervullen. Vanzelfsprekend profiteert synthetische biologie van de kennis en inzichten die verkregen werden uit de systeembiologie (Andrianantoandro *et al*, 2006; Endy, 2005; Heinemann and Panke, 2006; Serrano, 2007).

Hoewel het aantal succesvolle toepassingen in systeembiologie en synthetische biologie stijgt, blijven er toch ook nog een groot aantal uitdagingen over die opgelost moeten worden. Zo worden er bijvoorbeeld steeds nieuwe genoomwijde meettechnieken ontwikkeld. Hoe deze nieuwe databronnen in de analyse kunnen verwerkt worden en hoe zelfs nog meer nieuwe hoge doorvoer technieken te ontwikkelen, blijft één van de belangrijkste uitdagingen voor het huidige bioinformatica tijdperk. Een tweede grote uitdaging heeft te maken met de dynamica van het netwerk. De conditie-afhankelijkheid van interacties kan nu in rekening gebracht worden, maar een volledige kwantitatieve dynamische modellering van het genoomwijde biologische systeem is tot op heden nog niet mogelijk.

Publications

Marchal K., De Keersmaecker S., Monsieurs P., van Boxel N., **Lemmens K.**, Thijs G., Vanderleyden J., De Moor B. (2004) "In silico identification and experimental validation of PmrAB targets in *Salmonella typhimurium* by regulatory motif detection", Genome Biology, Vol. 5, no. 2, 20 p.

Van den Bulcke T., **Lemmens K.**, Van de Peer Y., Marchal K. (2006) "Inferring Transcriptional Networks by Mining Omics Data", Current Bioinformatics, vol. 1, no. 3, pp. 301-313.

Lemmens K., Dhollander T., De Bie T., Monsieurs P., Engelen K., Smets B., Winderickx J., De Moor B., Marchal K. (2006) "Inferring transcriptional module networks from ChIP-chip-, motif- and microarray data", Genome Biology, vol. 7, no. 5, pp. R37.

Dhollander T., Sheng Q., **Lemmens K.**, De Moor B., Marchal K., Moreau Y. (2007) "Query-driven module discovery in microarray data", Bioinformatics, vol. 23, no. 19, pp. 2573-2580.

Pastor D., Cortes-Calabuig A., **Lemmens K.**, De Moor B., Marchal K., Denecker M. (2007) "GeneReg: Integration of Experimental Data on the DNA Transcription Process". Proceedings of BNAIC 2007.

Lemmens K., De Bie T., Dhollander T., Monsieurs P., Bart De Moor B., Collado-Vides J., Engelen K., Marchal K. (2008) "The condition-dependent transcriptional network in *Escherichia coli*". Accepted for publication in Annals of NYAS, DREAM2.

Lemmens K., De Bie T., Dhollander T., De Keersmaecker S., Thijs I., Schoofs G., De Weerdt A., Bart De Moor B., Vanderleyden J.,Collado-Vides J., Engelen K., Marchal K. (2008) "Condition-dependent combinatorial regulation in *Escherichia coli*". Submitted.

Contents

Voorwoord ..... **i**
Abstract ..... **iii**
Korte inhoud ..... **v**
Acronyms and terminology ..... **vii**
Nederlandse samenvatting ..... **ix**
Publications ..... **xli**
Contents ..... **xliii**

Chapter 1: Introduction **1**
1.1 Context of the thesis 1
1.2 Transcription 2
1.3 Transcriptional regulation of gene expression 3
1.4 The transcriptional network 5
1.5 Reconstruction of the transcriptional network 8
1.6 Objectives of the thesis 9
1.7 Overview of the thesis 10

Chapter 2: The 'omics' era **17**
2.1 Genomics data 18
2.2 Transcriptomics 21
2.2.1 Spotted microarrays 21
2.2.2 GeneChip arrays 22
2.2.3 Filter-based DNA arrays 24
2.3 Protein-DNA interactome 24
2.4 Complementarity of the data 26

Chapter 3: Collection of omics data 29

3.1 *Saccharomyces cerevisiae* data 31

3.1.1 Expression data 31

3.1.2 Transcription factor binding site data 31

3.1.3 ChIP-chip data 32

3.2 *Escherichia coli* data 32

3.2.1 Expression data 32

3.2.2 Transcription factor binding site data 36

3.3 GeneReg 37

Chapter 4: Data integration and network reconstruction 39

4.1 Inference of regulatory transcriptional networks 40

4.1.1 Network inference using expression data 41

4.1.1.1 Network inference 41

4.1.1.2 Module network inference 43

4.1.2 Network inference by data integration 46

4.1.2.1 Data integration 46

4.1.2.2 Network inference by data integration 47

4.1.2.3 Module network inference by data integration 48

4.2 Association rule mining algorithms for module network inference by data integration 51

4.3 Which tool to choose as a biologist? 53

Chapter 5: Association rule mining in bioinformatics 57

5.1 The Apriori algorithm 60

5.2 Charm algorithm 61

5.3 Application of association rule mining algorithms in bioinformatics 64

Chapter 6: Unraveling the yeast transcriptional network 71

6.1 ReMoDiscovery 73

6.1.1 Seed discovery step 73
6.1.2 Parameter sensitivity 75
6.1.3 Seed extension step 78
6.2 Application of ReMoDiscovery on *Saccharomyces cerevisiae* data 79
6.2.1 Modules involved in ribosome biogenesis 83
6.2.2 Modules involved in galactose metabolism 83
6.2.3 Cell cycle 84
6.2.4 Stress related conditions 84
6.2.5 Respiration 86
6.2.6 Amino acid metabolism 86
6.2.7 Modules involved in filamentous growth 86
6.2.8 Modules involved in meiosis 87
6.3 Comparison with other module inference tools 87
6.4 Discussion 93

Chapter 7: Condition-dependent combinatorial regulation in *E. coli* 95
7.1 Motivation 96
7.2 DISTILLER 97
7.2.1 Identification of seed modules 98
7.2.2 Selection of interesting non-redundant modules 103
7.2.3 Seed module extension 104
7.2.4 Running parameters 104
7.3 Application of DISTILLER on *Escherichia coli* data 109
7.3.1 Identifying regulatory modules 110
7.3.2 Experimental validation of predicted interaction for FNR 115
7.3.3 Conditional dependency of the regulatory network 117
7.3.3.1 Oxygen responsive regulators 118
7.3.3.2 Regulators of the carbon metabolism 118

7.3.3.3 Regulators of nucleoside and amino acid biosynthesis 120
7.3.3.4 Regulators sensitive towards Iron/Cu 120
7.3.3.5 Nucleoid associated regulators 120
7.3.3.6 Stress responsive regulators 121
7.3.3.7 Biofilm related conditions 121
7.3.4 Combinatorial regulation 121
7.3.5 Comparison with the related methods 124
7.4 Discussion 125
7.5 Materials and methods 127
7.5.1 Benchmarking with RegulonDB and novel interactions 127
7.5.2 Static versus condition-dependent combinatorial regulation 127
7.5.3 Conditional dependency of the network 127
7.5.4 Comparison with other methods 128

Chapter 8: Conclusions and perspectives 129
8.1 Conclusions 129
8.2 Perspectives 131
8.3 From fundamental research to applications of systems biology 133
8.4 From applied systems biology to synthetic biology 134
8.5 Systems biology challenges 136

Bibliography 139
Curriculum vitae 163

Chapter 1

Introduction

1.1 Context of the thesis

Traditional molecular biology research usually focuses on the study of one or a few genes or proteins. Although an impressive number of biological principles have been revealed this way, these approaches are limited in giving us a comprehensive picture of cells, tissues and organisms. A systems biology approach aims at improving the understanding of how the different biological components (i.e., DNA, RNA, proteins, and metabolites) combine to make the organism viable (Greenbaum *et al*, 2001). A gene is therefore not longer studied as an isolated entity but as part of a complex network or system. The network of cellular components and their interactions mediates signal transduction by translating dynamically changing environmental cues into the observed behavior. Modeling the dynamic action of these networks to predict cellular behavior is the ultimate goal of systems biology research (Arita *et al*, 2005; Kitano, 2002).

The advent of high-throughput techniques has allowed large-scale identification of cellular entities, their activity patterns, and their biochemical and genetic interactions. These techniques have been used to generate so-called 'omics' data. Top-down systems biology is increasingly relying on the integration of heterogeneous omics data. Indeed, using distinct data sources instead of a single high-throughput data set to reconstruct networks yields several advantages. First, different omics data (e.g. genomics, transcriptomics, proteomics, interactomics, and metabolomics) unveil distinct aspects of networks and integrating those data leads to a more complete insight into these networks. Second, experimental and biological noise in the individual data measurements can be so prohibitive that each data type alone has limited predictive power. The integration of data from different sources provides an effective means to deal with the high noise level in the individual data sources by lowering the false discovery rate (FDR) and increasing the sensitivity (De Keersmaecker *et al*, 2006). Thus,

the extension of molecular biology to genome-wide analyses has caused a 'paradigm shift': research has evolved from studying a few genes into a systems thinking in which high-throughput data are being analysed in an integrative way.

Cells are able to alter the expression of genes according to their needs in a specific environment. Although the transcriptional network is only a portion of the total regulatory system, it is the major mechanism of regulation of gene expression, especially in prokaryotes, and therefore the subject of this thesis. In the following sections we will therefore briefly discuss transcription, regulation of transcription and the transcriptional network. The final sections of this introductionary chapter put forward the objectives and structure of the thesis.

1.2 Transcription

While deoxyribonucleic acid, or DNA, contains the code of life, the proteins are the real functional elements within a cell. It is the code of these building blocks of a cell that is contained in the genes, which are embedded in the DNA sequence. Genes are segments of DNA that encode for proteins through the intermediate action of mRNA. To form a protein, a gene is first transcribed into an mRNA sequence, a single stranded molecule similar to DNA (see Figure 1- 1). The transcription of a gene involves the creation of a mRNA sequence that is complementary to one of the DNA strands. In prokaryotes, mRNA is not modified, while eukaryotic mRNA is modified through RNA splicing, 5' end capping, and the addition of a polyA tail to obtain a template of the protein sequence. mRNA is further translated into an amino acid sequence (or protein) using a specific genetic code. Each three nucleotides of the mRNA sequence are translated into one particular amino acid during translation. Such a triplet of nucleotides is called a codon. These processes are illustrated in Figure 1- 1.

The transcription process consists of 3 steps: initiation, elongation and termination. During the initiation step the RNA polymerase binds to a specific site in the DNA, called promoter. At the promoter region it unwinds and unpairs a small region of the DNA double strand. In the elongation phase, ribonucleotides form base pairs with the DNA strand and bind to the previous ribonucleotide. A short DNA-RNA hybrid is formed that will not exceed a length of 10-20 bp because of rewinding of the DNA and displacement of the newly formed RNA. During termination the RNA polymerase recognizes a stop signal in the DNA sequence that triggers the separation of the complex of DNA, RNA and RNA polymerase.

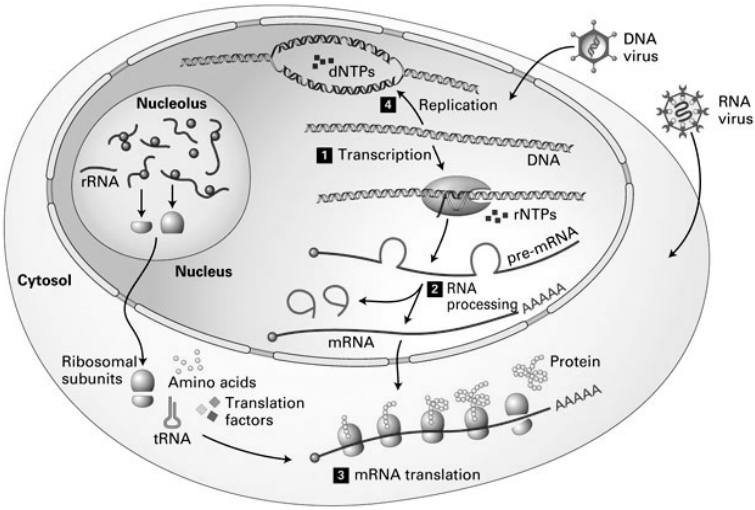


Figure 1- 1: Biological processes in a eukaryotic cell. The transcription of a gene involves the creation of a mRNA sequence that is complementary to one of the DNA strands. In eukaryotes, this mRNA needs to processed further: there can be a RNA splicing, 5' end capping, and the addition of a polyA tail. During translation proteins are produced using the mRNA as a template. Taken from Scott *et al* (2003).

In prokaryotes, DNA, RNA and protein synthesis take place in the same cellular compartment. In contrast, DNA and RNA synthesis of eukaryotes takes place in the nucleus, while protein synthesis takes place in the cytoplasm. In eukaryotes the RNA and protein synthesis are thus completely separated while in prokaryotes the translation process can already start while the transcription process has not yet completed.

1.3 Transcriptional regulation of gene expression

When environmental conditions change, cells need to adapt to these changes: active genes must be switched off and other inactive genes must be turned on. Regulation mechanisms are needed such that genes will only be transcribed when the corresponding proteins are needed in the cell.

The transcription rate can be influenced in several ways, including[*]:

- The promoter region of prokaryotes contains two consensus sequences: TTGACA and TATAAT, respectively at 35 base pairs

[*] Alternative transcriptional regulation mechanisms exist, for instance small RNA mediated regulation

and 10 base pairs before the transcription start position. In eukaryotes a TATA box or TATAAA consensus sequence is present at 25-30 base pairs before the transcription start position. These consensus sequences are recognized by the RNA polymerase and the more similar the DNA sequence is to these consensus sequences, the stronger the binding of RNA polymerase is and the higher the promoter strength is.

- Additional transcription regulation is mediated by transcription factors. These regulators respond to specific conditions in the environment and can facilitate or inhibit the binding of the RNA polymerase to the DNA sequence. The transcription rate can be positively or negatively affected by the action of transcription factors, called respectively activators or repressors.

The transcription factors bind to specific binding sites in the promoter region of the gene. These binding sites or regulatory motifs are short conserved DNA sequences. The location of these sites and the affinity of the transcription factor for these binding sites determine the expression level of a gene in a specific environmental condition. Motifs are usually located close to the transcription start site. In yeast, upstream activator sequences (UAS) are usually located at 40-200 base pairs upstream from the transcription start site. In higher eukaryotes, however, promoter regions can also be located as far as 10kb from the gene they are controlling. Enhancers can be found upstream, downstream and even in the genes themselves. Both the UAS and the enhancers serve as binding sites for transcriptional regulators and usually more specific regulators will bind to these binding sites.

Since gene regulation is a very complex process, both prokaryotic and eukaryotic genes are often regulated by more than one transcription factor. The effects of two regulators binding to the promoter region of a gene in the same experimental condition can be cooperative (or synergistic) or antagonistic (Kel *et al*, 1995). If the binding effect of the two regulators on the gene expression is larger than the sum of the binding effects of each of the regulators separately, we refer to it as a cooperative binding. In an antagonistic binding, the repressor interferes with the activator while binding simultaneously to DNA. It could also be that the activator and repressor are in competition for binding to the same or overlapping binding sites.

The action of multiple regulators on the same gene is called combinatorial regulation or control. Combinatorial control is believed to form an important aspect of transcription regulation. Nevertheless, while a number of specific molecular details of the mechanisms of combinatorial regulation have emerged, our understanding of the general principles of combinatorial regulation on a genomic scale is still limited (Balaji *et al*, 2006; Balaji *et al*, 2007). Often it is only known whether a gene is regulated by two or more

transcription factors irrespective of whether these regulators actually work together in similar conditions or act independently from each other under different conditions.

1.4 The transcriptional network

Transcriptional regulation can be represented in the form of a transcriptional network defined as a directed graph where the nodes represent the genes and a directed edge from one node to another indicates that the first gene codes for a transcription factor that regulates the second gene (Figure 1- 2). The architecture of this network can then be described by means of graph features such as distance, diameter or degree. We briefly introduce the latter concepts and their relevance for the study of regulatory networks below.

The shortest path between two nodes is called the distance. The graph diameter is the maximum distance between any two nodes in the network. If this diameter is small, the network is called a small world network. Many biological networks, for instance the metabolic network, are known to exhibit this small world property (Wagner and Fell, 2001).

The degree of a node is defined as the number of edges that connect to this node. The number of incoming edges is called the in-degree, while the number of outgoing edges is called the out-degree. If a node has a high degree, this means that this node is connected to many other nodes in the network. Biological networks are not randomly organized but seem to have a scale-free architecture with the typical power law degree distribution: a limited number of nodes have a high degree while most nodes have a small degree. This means that there are some hubs present in the network: these are nodes connected to many other nodes. The protein-protein interaction network, transcriptional network and metabolic network are all scale-free networks (Barabasi and Albert, 1999; Guelzim *et al*, 2002; Jeong *et al*, 2000). The advantage of this kind of organization is that the loss of one non-hub link is not as disruptive in scale-free networks as in random networks. In other words, scale-free networks are generally more robust. The hubs are of course extremely important and usually they play essential roles in biological systems (Yu *et al*, 2004a).

In addition to being scale-free, Shen-orr *et al* (2002) discovered the presence of network motifs in the transcriptional network of *E. coli*. Network motifs are topologically distinct regulatory interaction patterns that are present more frequently in true biological networks than in random networks. Therefore these motifs must have a specific biological function: they are postulated to be the basic signal transduction elements, each with their own characteristic properties. Shen-orr *et al* (2002) were the first to identify these network

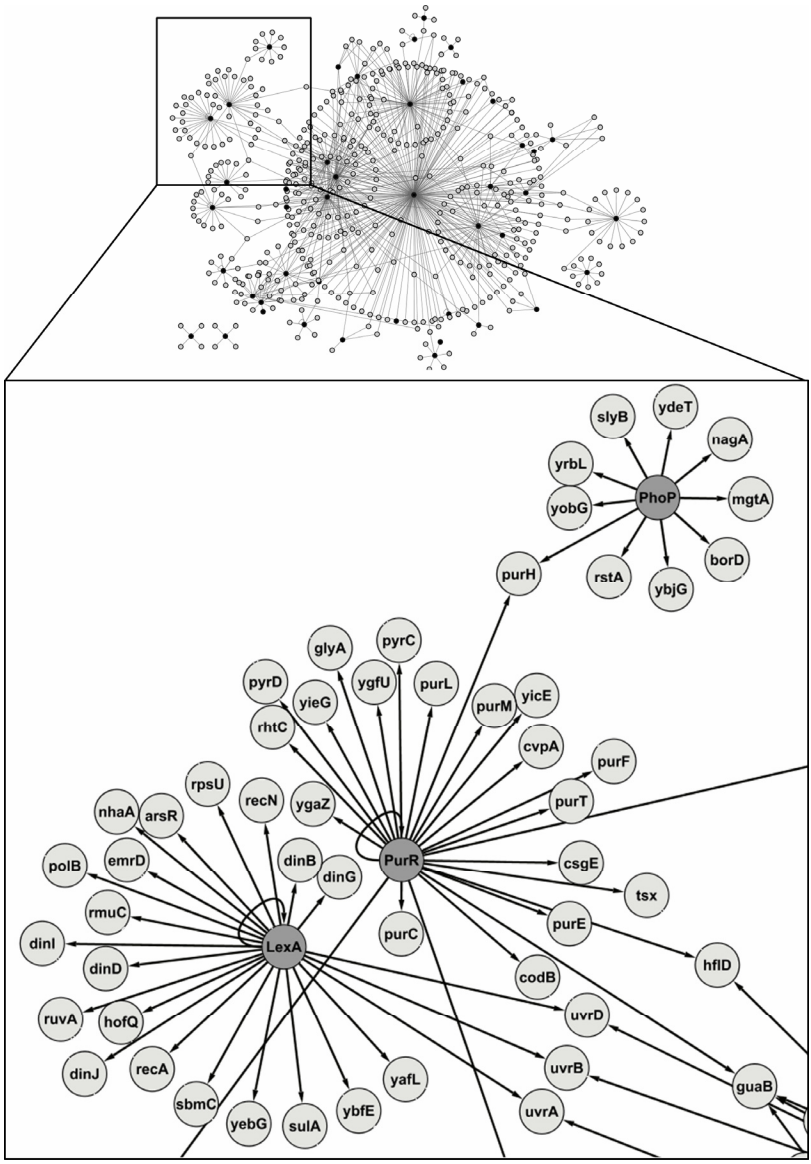


Figure 1- 2: Representation of the *Escherichia coli* regulatory network. The black nodes represent regulators; the grey nodes represent target genes. An edge between a regulator and gene indicates that the gene is regulated by the regulator. Regulator PhoP, for instance, regulates 10 genes. One of the genes, *purH*, is regulated by two transcription factors, PhoP and PurR. This type of network representation does not show the condition dependency of the interactions, i.e. all interactions are shown although it may happen that two interactions never 'appear' in the same experimental condition. This transcriptional network was obtained by applying our data integration framework DISTILLER on expression and regulatory motif data (see chapter 7).

motifs. Examples of network motifs are the single input motif, multiple input and feed-forward loop motifs (Figure 1- 3).

Previous studies have unveiled the modular and hierarchic organization of biological networks (Guelzim *et al*, 2002; Hartwell *et al*, 1999; Ravasz *et al*, 2002; Rives and Galitski, 2003; Tanay *et al*, 2004). Indeed, biological processes consist of pathways that mainly act on their own although communication exists between these pathways. Therefore one might expect that the distinct biological processes are organized in discrete and separable modules (Figure 1- 3). A module in a network can be defined in various ways (Wolf and Arkin, 2003):

- A topological module can be defined by means of graph-based clustering techniques. The clustering coefficient of a node in the network can be calculated as the number of edges between the nodes in the neighbourhood divided by the number of possible edges between them.
- Another popular definition of a module involves co-expressed genes, with or without environmental context dependence and assigning a regulatory motif or regulator to these genes. This latter definition of a module will also be used in the remainder of this text.

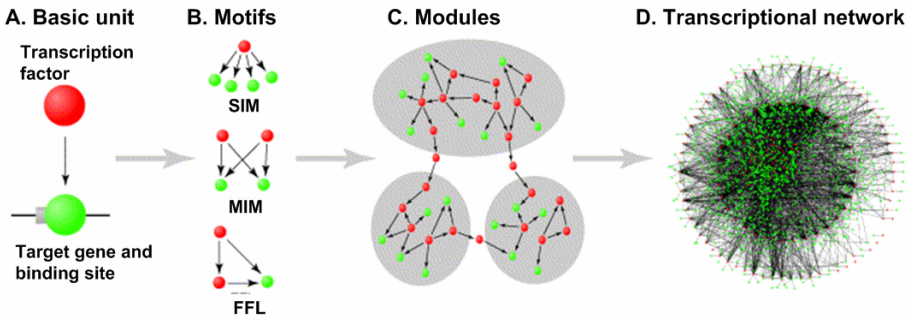


Figure 1- 3: The different layers in the transcriptional regulatory network. The basic unit of the regulatory network consists of a transcription factor that regulates a target gene (A). These basic units are organized into network motifs (B). Three examples are shown: a single input motif (SIM), a multiple input motif (MIM) and a feedforward loop. The transcription factors are indicated in red, the target genes in green. The network motifs can be organised in modules (C). The modules themselves are also linked by hubs. Finally the complete transcriptional regulatory network is shown (D). Figure taken from Babu *et al* (2004).

1.5 Reconstruction of the transcriptional network

Revealing the complete regulatory network underlying the system's behavior is the ultimate goal of systems biology. In this thesis, we focus on the transcriptional network, one aspect of the regulatory network. The purpose of transcriptional network inference is to determine for all transcription factors the regulatory motifs they recognize, the conditions in which they are active, the regulators they cooperate with in these conditions, and their target genes in these conditions. By using a top-down data-driven approach we aim at reconstructing the transcriptional network on a large scale.

Top-down inference can be contrasted with bottum-up inference (Bruggeman and Westerhoff, 2007; De Keersmaecker *et al*, 2006):

- *Bottom-up* inference starts from a comprehensive expert model of known interactions between molecular entities as described in literature and curated databases. Such models can be used to simulate cellular behavior or to predict the outcome of a perturbation experiment. Inconsistencies between observed data and simulations point at deficiencies in the current network structure and outline hypotheses of novel interactions that can better explain the observations.
- *Top-down* network inference methodologies start from a global view of the behavior of the system by using high-throughput data. This kind of inference methods do not rely on expert knowledge on the relationships between the molecular components. Top-down inference is a data-driven and thus data-demanding approach. Given the current data availability, top-down network inference problems are often underdetermined (i.e. the network that is reconstructed from the data is not unique, many equally likely solutions can explain the observations). However, the top-down inference can be made increasingly tractable by integrating data from different sources, and holds great promise for future bioinformatics research.

It is important to keep in mind that a network is a dynamic entity: some regulator-gene interactions might be present in one experimental condition while not in other conditions. Therefore the regulatory network is different in different environmental or experimental conditions (Luscombe *et al*, 2004; Van den Bulcke *et al*, 2006). In current representations of the regulatory network, this condition-dependency is usually not taken into account, i.e. all possible regulator-gene interactions are shown in a static way (Figure 1- 2). The detailed study of condition dependency is one of the innovative aspects of this thesis.

1.6 Objectives of the thesis

From the overview above, it should be clear that transcriptional regulation is a very complex process. During transcription, a transcription factor binds to a specific binding site in the promoter region of a gene to regulate this gene. Often, multiple regulators are in control of the same gene. These regulators might cooperate or they may work independently from each other. In addition, the transcription of a gene is condition-dependent, so the interactions in the transcriptional network change depending on the environmental condition.

Several data sources are available that provide genome-wide information on for instance gene expression or transcription factor binding sites. One data source, however, does not provide sufficient information to obtain a holistic insight in the transcriptional network. By integrating multiple data sources providing complementary information on the expression of genes, transcription factor binding and regulatory motifs, we can obtain a more detailed and complete insight in the transcriptional network. In this thesis we therefore focus on the integration of data that all provide information on the transcriptional network (Figure 1- 4).

Although recently several data integration methods for the inference of regulatory networks have been developed (chapter 4), these methods usually do not take all previously mentioned aspects of the transcriptional network into account. Most methods either focus on condition dependent expression (for, instance, biclustering algorithms described in chapter 4) or data integration. In addition, not all methods are capable of identifying regulation of a gene by multiple regulators. Therefore, we developed two network inference methods in this thesis that exploit data integration and allow to study the (condition-dependent) modular transcriptional network. Both our methods rely on "association rule mining" algorithms (chapter 5) that, although currently not yet used very often in the bioinformatics field, are very efficient and allow the integration of several data sources. Our two approaches, ReMoDiscovery (chapter 6) and DISTILLER (chapter 7) have been successfully applied on two model organisms, yeast and *Escherichia coli* respectively, indicating that association rule mining algorithms are indeed well suited for data integration. By using multiple data sets we were able to reconstruct the transcriptional network for both organisms. Moreover, important aspects of transcriptional networks such as modularity, condition dependency and combinatorial regulation, as well as the relationship between these properties were studied in detail.

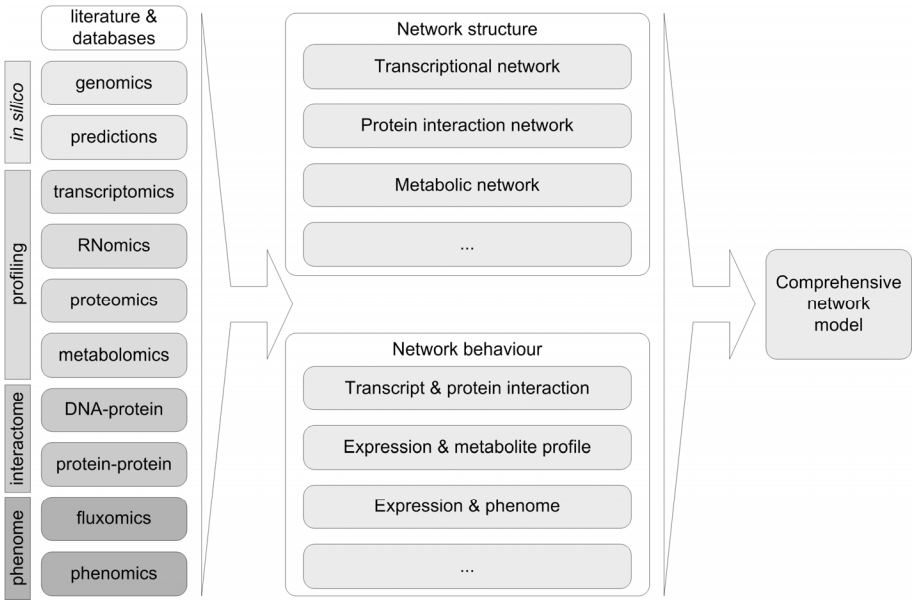


Figure 1- 4: Integration of different data sources to obtain a comprehensive network model. The in silico data refer to for instance predictions of transcription factor binding sites. The 'profiling' omics data (or component data) measure which molecule is available and in which amount it is available in a particular experimental condition. The 'interactome' data describe whether there is an interaction between two molecules in the cell. 'Phenome' data provide information on the overall behaviour of the system. The knowledge present in literature and databases can be integrated with different types of omics data (profiling, interactome or phenome) and *in silico* predictions to obtain the network structure for different types of networks and the relation between the different networks (network behaviour). The ultimate goal of data integration is to integrate all network information into one comprehensive network model. Figure taken from De Keersmaecker *et al* (2006).

1.7 Overview of the thesis

This section gives a chapter-by-chapter overview of the thesis (see Figure 1-5). Table 1- I indicates for each chapter whether that chapter is a literature study and whether it is associated with our own methodological or biological contributions to the research topic.

This thesis focuses on the study of the transcriptional regulatory network via data-driven top-down network inference methods. The advent of high-throughput DNA sequencing techniques triggered the development of other high-throughput techniques to measure cellular entities and their

Table 1- I: Overview of the different chapters of the thesis. For each chapter a brief summary of its content is shown. The type of contribution is also indicated (Literature: literature overview; Method: methodological contribution; Biology: biological contribution). Publications and external collaborations (with research groups other than ESAT and CMPG) related to the chapters are also specified.

Ch.	Content	Contr.	Publication	Collaboration
2	Description of the omics data that will be used in the thesis	Literature		
3	Collection of the omics data	Biology		Prof. Collado-Vides (UNAM, Mexico)
	Development "GeneReg" to automatically collect the data	Method	Pastor *et al*, 2007	Álvaro Cortés Calabuig (KU Leuven, Belgium)
4	Overview on the different network reconstruction methods	Literature Method	Van den Bulcke *et al*, 2006; Dhollander *et al*, 2007	
5	Introduction to association rule mining algorithms and their applications in bioinformatics	Literature		
6	Development of a module detection tool ReMoDiscovery	Method	Lemmens *et al*, 2006	Prof. De Bie (University of Bristol, UK)
	Application of ReMoDiscovery on yeast data: identification of modules covering many biological processes	Biology		
7	Development of DISTILLER for the inference of regulatory modules	Method	Lemmens *et al*, 2008b	Prof. De Bie (University of Bristol, UK)
	Application of DISTILLER on *E.coli* data: insight in condition dependency, modularity and combinatorial regulation	Biology	Lemmens *et al*, 2008a	Prof. Collado-Vides (UNAM, Mexico)

interactions. Many types of omics data are currently available and although they all are important to better understand the cell as a system, not all of these data provide information on the transcriptional regulatory network. **Chapter 2** gives an overview of 'omics' data that are directly related to the transcription process, such as the transcriptomics, protein-DNA interactome and genomics (regulatory motifs). These data sources are used in later chapters.

The first step in the reconstruction of the transcriptional network is the collection of relevant omics data. In **chapter 3** we describe how a compendium of omics data was created for two model organisms: the prokaryote *Escherichia coli* and the eukaryote *Saccharomyces cerevisiae*. These omics data are not always easily accessible. This makes collecting them in general a hard and tedious task. In collaboration with the computer science department of the KU Leuven, we have developed a software tool called GeneReg, to make the collection and storage of heterogeneous data sources feasible. This software has been published:

> Pastor D., Cortes-Calabuig A., Lemmens K., De Moor B., Marchal K., Denecker M. (2007) GeneReg: Integration of Experimental Data on the DNA Transcription Process. Proceedings of BNAIC 2007.

In **chapter 4** we describe how the collected data can be used for network inference. We distinguish data integration methods from methods that deal with a single data source only (usually expression data). A second classification is based on the distinction between direct network inference or inference of regulatory modules (i.e. the basic building blocks of the network). Although it is encouraging to see that many network inference tools have been developed and that promising biological results have been generated, this wealth of available algorithms can be overwhelming for the biological user. The need for benchmarking is therefore also discussed in this chapter. A review on the different module detection tools has been published:

> Van den Bulcke T., Lemmens K., Van de Peer Y., Marchal K. (2006) Inferring Transcriptional Networks by Mining Omics Data. Current Bioinformatics, vol. 1, no. 3, pp. 301-313.

A query-driven biclustering method based on expression data only has been published:

> Dhollander T., Sheng Q., Lemmens K., De Moor B., Marchal K., Moreau Y. (2007) Query-driven module discovery in microarray data. Bioinformatics, vol. 23, no. 19, pp. 2573-2580.

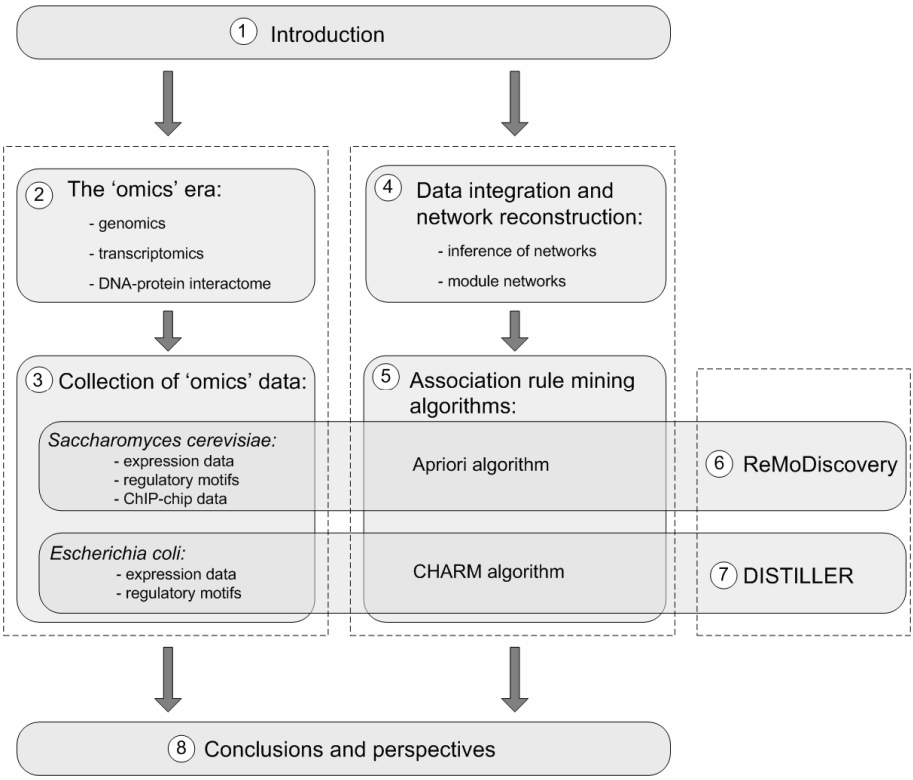


Figure 1- 5: Chapter-by-chapter description of the thesis structure. In chapter 2 we give an overview of the different types of omics data that are relevant for reconstruction of the transcriptional network, while chapter 3 describes how a data compendium was created for two model organisms, *Saccharomyces cerevisiae* and *Escherichia coli*. In chapter 4 of the thesis we start with an overview of existing data integration and network inference methods, followed by a more detailed discussion of the association rule mining algorithm class, since our developed module detection tools are based on these algorithms (Chapter 5). In chapter 6 and 7 of the thesis we discuss our two data integration frameworks. The Apriori-based ReMoDiscovery was developed and applied on yeast data. This study is described in chapter 6. In chapter 7 we describe DISTILLER, which is based on the more advanced CHARM algorithm. DISTILLER has been applied on *E. coli* data and allowed to study condition-dependent co-expression and combinatorial regulation.

Although many network inference tools are available, we feel that these tools do not take into account all aspects of transcriptional regulation and are therefore not perfectly suited to obtain a complete insight in the transcriptional network. We therefore developed our own network inference tools based on highly efficient association rule mining algorithms (data mining). In **chapter 5** these algorithms are first described in general terms. Since our tools are based on two specific algorithms, the Apriori and the CHARM algorithm, chapter 5 focuses on these two algorithms. We also discuss prior art regarding the application of association rule mining algorithms in the bioinformatics field.

In **chapter 6** we present our Apriori-based module detection tool ReMoDiscovery. The presented methodology uses microarray information, regulatory motif data and ChIP-chip data to identify modules. These modules consist of sets of co-expressed genes that have a motif in common in their promoter regions and that are regulated by a common set of regulators. ReMoDiscovery proceeds in two steps. In a first step (seed discovery) seed modules are identified, while in the second step (seed extension) the gene content of the seed modules is further optimized based on statistical criteria. ReMoDiscovery was applied on yeast data and modules covering a whole range of biological processes were identified: cell cycle, nutrient deprivation, stress, respiration, amino acid metabolism, filamentous growth and meiosis. We compared our results with the results of other module detection tools. Our proposed approach seemed to outperform alternative methods. This work was published in Lemmens *et al* (2006).

Lemmens K., Dhollander T., De Bie T., Monsieurs P., Engelen K., Smets B., Winderickx J., De Moor B., Marchal K. (2006) Inferring transcriptional module networks from ChIP-chip-, motif- and microarray data. Genome Biology, vol. 7, no. 5, pp. R37.

ReMoDiscovery proved that tools based on association rule mining algorithms are equally suitable for module detection as other more frequently used graph-based or probabilistic algorithms. However, with the growing number of available experiments, a condition selection or bicluster strategy became necessary. We therefore developed DISTILLER, a module detection tool that allows condition selection. Since biclustering poses additional challenges for modeling and computation, we relied on a more advanced association rule mining algorithm called CHARM. The data integration framework DISTILLER was applied on a compendium of 870 *Escherichia coli* microarrays and regulatory motif data (67 motifs). Because of the condition selection strategy we were able to study the relation between condition-dependent co-expression and combinatorial regulation. DISTILLER and its application on the *E. coli* data are described in **chapter**

7. The methodological results of this study have been accepted for publication:

Lemmens K., De Bie T., Dhollander T., Monsieurs P., Bart De Moor B., Collado-Vides J., Engelen K., Marchal K. (2008) "The condition-dependent transcriptional network in *Escherichia coli*". Accepted for publication in Annals of NYAS, DREAM2.

The biological results of this study are submitted for publication

Lemmens K., De Bie T., Dhollander T., De Keersmaecker S., Thijs I., Schoofs G., De Weerdt A., De Moor B., Vanderleyden J., Collado-Vides J., Engelen K., Marchal K. (2008) DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli*. Submitted.

The results of this thesis are summarized in **chapter 8**, together with possible future extensions of this work. The chapter also describes how our rather fundamental research fits in the broader context of systems biology.

Chapter 2

The 'omics' era

The completion and publication of the *Haemophilus influenzae* genome sequence in 1995 (Fleischmann *et al*, 1995) constitutes a milestone in the history of biological research: it was the first genome to be published and many more followed, including the human genome in 2001 (Lander *et al*, 2001; Venter *et al*, 2001). The advent of high-throughput DNA sequencing techniques triggered the development of other novel high-throughput techniques to measure the different components and interactions between the components in the cell. Nowadays, thousands of genes or proteins can be studied in one single experiment, giving rise to genome-wide data sets. These so-called 'omics' data can be classified into three categories: data on the components, on the interactions between the components or on the functional states (Greenbaum *et al*, 2001; Joyce and Palsson, 2006). The components data provide an inventory of the molecules in the cell. Genomics, transcriptomics, proteomics and metabolomics are examples of such components data. High-throughput techniques for the measurement of lipids (lipidomics) or carbohydrates (glycomics) are also available but relatively few such data sets have been generated so far. The interactions data yield information on the connectivity between the different cell components. Protein-DNA interactome or protein-protein interactome data belong to this class. The functional-states data specifies the overall behavior of the system, for instance in the form of phenomics data (Greenbaum *et al*, 2001; Joyce and Palsson, 2006).

This thesis focuses on methods that reconstruct the regulatory network by integrating different omics data (see Figure 1- 4). Since our main goal is to reveal the transcriptional network in two model organisms (yeast (chapter 6) and *E. coli* (chapter 7)), we only discuss those omics data available for yeast and bacteria that are directly relevant for our research: regulatory motif data (genomics), transcriptomics data and protein-DNA interactomics data (see Figure 2- 1). In chapter 3 we explain how these omics data were collected and subsequently used as input for network reconstruction methods.

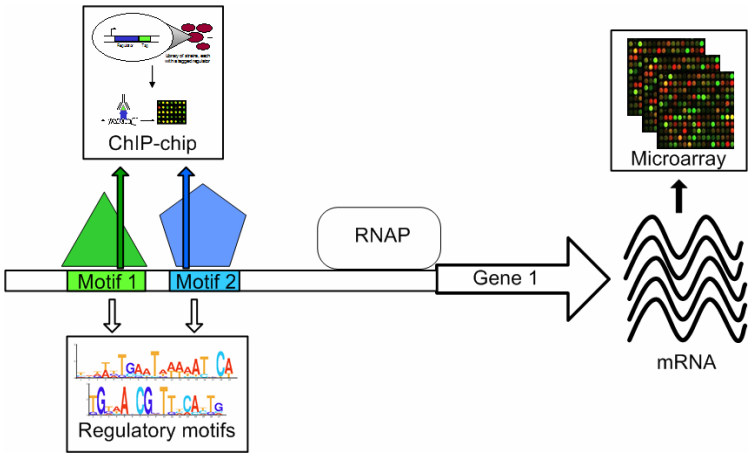


Figure 2- 1: Simplified representation of transcriptional regulation. A regulator (indicated with green triangle or blue pentagon) binds to the transcription factor binding site (indicated with 'Motif 1' and 'Motif 2'). The regulators can as such influence the transcription level of gene 1. The omics data that give information on this process and that were used in this thesis are shown. ChIP-chip data provide on a genome-wide scale information on the direct physical interaction between a regulator and DNA, while regulatory motifs can be detected *in silico*. Microarray data measure the level of mRNA present in the cell in a given experimental condition.

2.1 Genomics data

Genomics is defined as the study of the whole genome sequence, where both the coding and non-coding parts are being investigated. Complete genome sequences are available for about 628 prokaryotic and 22 eukaryotic organisms[†]. These genome sequences have been utilized to discover regulatory elements (Kellis *et al*, 2003; Monsieurs *et al*, 2005; Monsieurs *et al*, 2006) or to study evolution (Delsuc *et al*, 2005).

This thesis aims at unraveling the transcriptional network. Transcription factor binding sites or regulatory motifs are therefore a very informative data source to include in our research.

Regulatory motifs are short, conserved DNA-sequences present in the promoter region of a gene. They are the tags that are recognized by the regulators and hence play a very important role in the transcriptional regulatory network. Although the binding of a regulator to a regulatory motif is condition dependent, the motif itself is inherently present in the DNA

[†] From http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html

sequence. Therefore, motifs of an organism can in principle be identified independent of the experimental conditions.

Although the transcription factor recognizes a specific transcription factor binding site, some variation is usually allowed in the binding site. For the same regulator the motif instance in the upstream region of one gene should thus not be exactly the same as the motif instance in the upstream region of another gene. By aligning all motif instances of a regulator, the regulatory motif for that regulator can be represented (see Figure 2- 2). A first way of representing the motif is by a consensus sequence. The consensus sequence consists of the most frequently occurring nucleotide at each position. A motif matrix provides a more detailed way of representing a regulatory motif: the frequency (or probability) of each nucleotide at each position is given. The motif logo is a third way of representing the regulatory motif. It is a graphical representation of the amount of information at each position.

Because empirical validation of binding sites is laborious, several computational methods have been developed for identification of regulatory motifs. A first group of methods aims at identifying regulatory motifs without any prior knowledge about which specific binding site is searched for. These methods are the so called *de novo* motif identification methods. Alternative motif-screening approaches use known or predicted transcription factor binding sites to scan sequences for potential binding sites.

De novo motif identification algorithms fall apart in two main categories: methods starting from co-expressed genes and methods that make use of comparative genomics, i.e., phylogenetic footprinting. The rationale of the former category is that genes showing a similar expression pattern are likely to share the same transcriptional regulation mechanism, and hence the same regulatory motifs. Phylogenetic footprinting methods rely on the fact that parts of the promoter region of a gene that are non-functional can undergo mutations during evolution. In contrast, the regulatory motifs embedded in these evolving promoter regions are functional and therefore tend to be more conserved during evolution. By comparing promoter regions of orthologous genes of evolutionary closely related species, motifs can be detected.

Although computational methods have generated interesting results, the resulting motif compendia usually contain many false positives. These are motifs that were predicted by the *in silico* techniques but that are biologically not relevant. In theory motif compendia can be constructed for any organism for which a sufficient number of sequence data is available, but applying the motif detection procedures is far from trivial in organisms with long intergenic sequences such as vertebrates. Specialized databases exist (such as TRANSFAC (Matys *et al*, 2006), RegulonDB (Gama-Castro *et al*, 2008), …) that contain information on regulatory motifs of diverse organisms.

Once a reliable list of motifs is compiled, the corresponding motif models can be used to screen all intergenic regions of the organism to identify additional genes containing a motif instance in their promoter region. In this thesis we use such motif screening techniques (Hertzberg *et al*, 2005; Marchal *et al*, 2004) to obtain genome-wide information about the presence of a motif instance in the promoter region of a gene (chapter 3, 6 and 7).

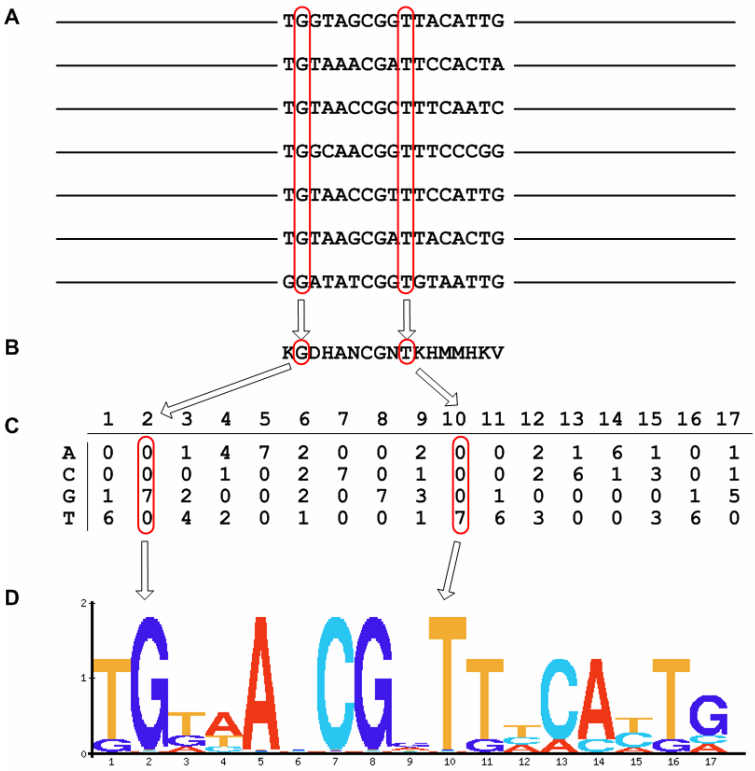


Figure 2- 2: FNR binding sites and sequence motif. A) Seven motif instances (or known binding sites) of the regulator FNR in *E. coli*. B) The consensus sequence of the motif. Note that 'K', for instance, stands for G or T, as defined by IUPAC‡. C) Frequencies of the nucleotides at each position. D) Sequence logo showing the frequencies scaled relative to the information content (measure of conservation) at each position. In chapter 7 we predict and experimentally validate novel targets for the FNR regulator.

‡ www.chem.qmul.ac.uk/iubmb/misc/naseq.html#tab1

2.2 Transcriptomics

Transcriptomics data describe snapshots of the activity of a cell by measuring the presence and abundance of each type of mRNA molecule. In the past few years, microarray technology has emerged as an effective technique to measure the level of expression of thousands of genes in a single experiment. Because of their capacity to monitor many genes, microarrays have become one of the main workhorses of molecular biology gene regulation research. mRNA provide important information for reverse engineering of regulatory networks (Wyrick and Young, 2002) and will also be used in this thesis.

In the past, microarray experiments did not have to be made publicly available upon publication of analysis results, but journals recently started to require that authors make their microarray results available in public databases. Gene Expression Omnibus (Barrett *et al*, 2007), Stanford microarray database (Demeter *et al*, 2007) or ArrayExpress (Parkinson *et al*, 2007) contain transcriptome data for many organisms including yeast and bacteria. To ensure interpretability and to allow independent verification of the results, the "Minimum Information About a Microarray Experiment" standard has been introduced (Brazma *et al*, 2001).

In this thesis we will make use of these public databases to fully exploit all available microarray data. Three important types of microarrays, namely spotted microarrays, GeneChips and filter-based arrays, are available from these databases and will be discussed below.

2.2.1 Spotted microarrays

Spotted microarrays are a relatively cheap and popular technology platform that uses small glass slides on which pre-synthesized DNA is spotted (see Figure 2- 4). Each spot represents one gene. Spotted arrays use two samples: a reference and a test sample, for instance normal versus malignant tissue. These samples are labeled using distinct fluorescent molecules, Cy3 (green) and Cy5 (red), and simultaneously hybridized to the microarray. Relative amounts of a particular gene transcript in the two samples are determined by measuring the signal intensities detected at both fluorescence wavelengths and calculating the ratios of the two signal intensities.

Normalization of the raw, extracted intensities is a necessary and important step that precedes any high-level analysis. Normalization aims to remove consistent and systematic sources of variation to ensure comparability of the measurements, both within and across slides.

Dye-related discrepancies, for instance, stem from a variety of factors, including physical properties of the dyes and efficiency of dye incorporation. These dye biases can be removed based on the "global normalization principle" that states that most genes will not change their expression level during an experiment and those that do tend to have equal probabilities of being upregulated and downregulated. This principle implies that a so-called MA-plot, in which one plots the log ratios of the intensities of the two channels (or M-value = $\log_2(Cy5/Cy3)$) versus the mean intensities (or A-value = $\log_2(\sqrt{Cy5 \cdot Cy3})$) for each gene, should be centered on M=0, but due to dye-related discrepancies it often is not (see Figure 2- 3). The most popular, and widely applied intensity-dependent normalization procedure to remove these artefacts was first described by (Yang *et al*, 2002) and will also be used in this thesis for the normalization of spotted arrays (chapter 3).

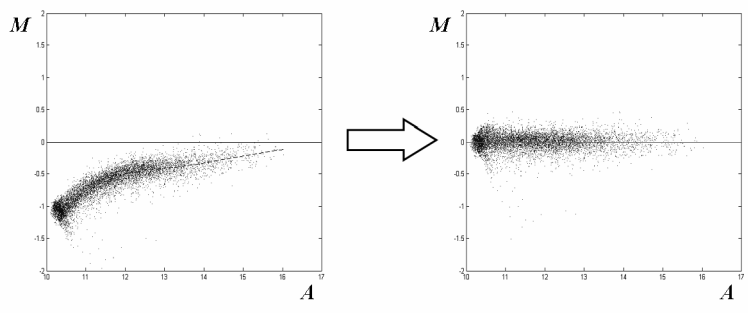


Figure 2- 3: MA-plot before (left) and after (right) a Loess intensity-dependent normalization (Yang *et al*, 2002). Before normalization, the measurement cloud is not centered on M=0 because of dye effects, while after normalization the centering is clearly on M=0. Figure taken from Engelen (2005).

2.2.2 GeneChip arrays

GeneChip arrays (Affymetrix Inc., Santa Clara) are high-density arrays of oligonucleotides or probes that are synthesized *in situ*. Each gene is represented by several short sequences of 15-20 different complementary oligonucleotides; control oligonucleotides are also added. In contrast to spotted arrays, each GeneChip measures one biological sample or condition (see Figure 2- 4).

Because GeneChip arrays have a different probe design and are single-channel arrays, they require a different preprocessing than the spotted arrays. At the moment MicroArray Suite 5.0 (MAS 5.0) and Robust Multi-array Average (RMA) are the most popular normalization methods. The main

differences between MAS 5.0 and RMA are the fact that RMA does not use the mismatch values and that RMA applies normalization between the slides. None of these methods is currently set as the best standard method. Lim *et al* (2007) claim that MAS 5.0 would be better for the inference of networks, but other papers indicate that RMA is the better normalization technique (Faith *et al*, 2007; Millenaar *et al*, 2006). In this thesis we used RMA normalization if raw GeneChip intensities were available (chapter 3).

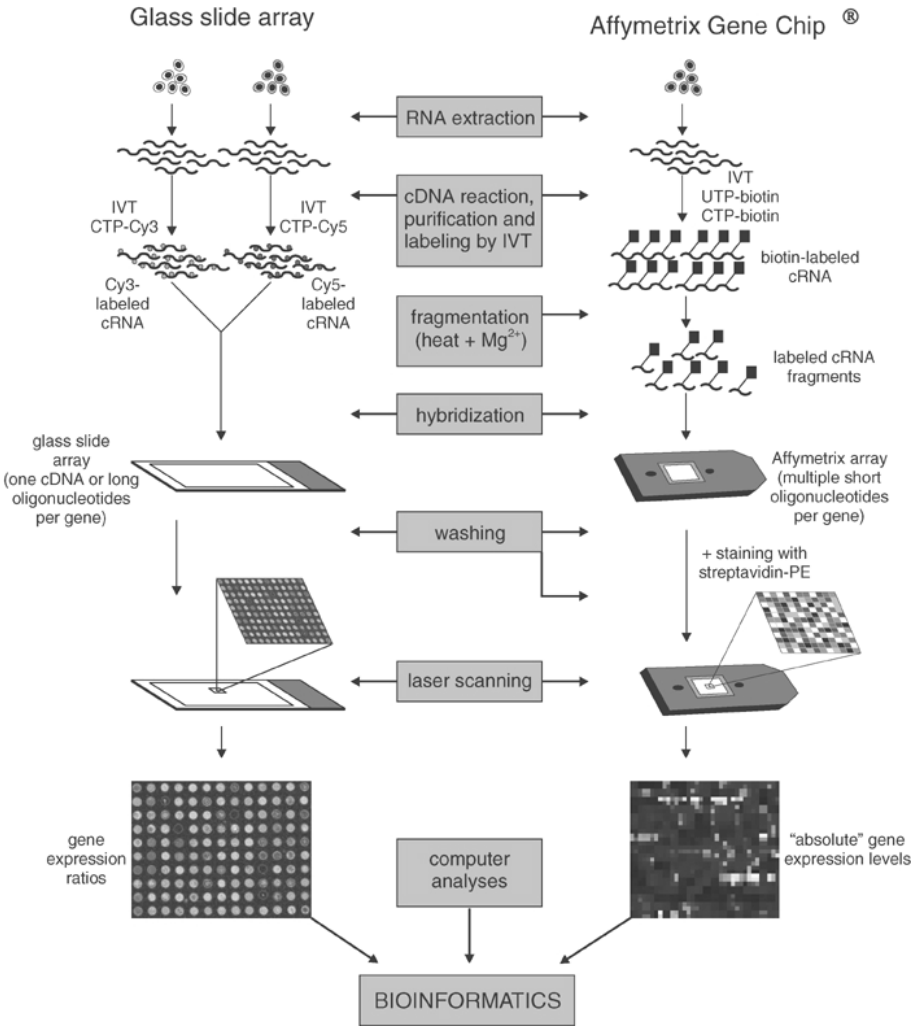


Figure 2- 4: Spotted arrays versus GeneChip arrays. The main difference between the two types of microarrays, is that two biological samples can be hybridized on the spotted arrays, while one sample is hybridized on GeneChip arrays. Taken from Staal *et al* (2003).

2.2.3 Filter-based DNA arrays

In addition to the solid support matrices arrays (spotted arrays or GeneChip arrays), another category of microarrays exists that use a nylon filter as a support matrix. Nylon filter-based arrays are single channel arrays like the GeneChip, meaning that only one sample is hybridized to the nylon. The sample is labeled with radioactive P^{33}. The P^{33} labeling is more sensitive than the fluorescent labeling, but gives rise to health risks (Baldi and Hatfield, 2002).

Filter-based arrays thank their popularity to the fact that profiling is based on standard southern blotting protocols, techniques that are quite well-known to most biologists. The necessary equipment to perform this kind of analysis (for instance, a phosphorimager) is present in most research institutes. Nylon filters can also be reused several times, which offers a cost advantage. Although these nylon filters are not suited for large genomes, bacterial genomes pose no real problems. For example, Sigma-Genosys produces a nylon filter DNA array for transcription profiling in *E. coli* (Baldi and Hatfield, 2002). This type of array is also present in the microarray deposits and will be used as a data source in this thesis.

2.3 Protein-DNA interactome

The transcriptional regulatory network is composed of interactions between transcription factors and the promoter regions of target genes. For the reconstruction of the transcriptional network we will therefore also use data providing direct information on the binding of regulators. Chromatin immunoprecipitation on chip (or ChIP-chip) provides information on the direct physical interaction between a regulator and the promoter regions of its target genes and has become the most popular technique to determine the protein-DNA interactome.

Chromatin immunoprecipitation (or ChIP) has previously been used to study protein-DNA interactions for a small number of possible target genes of a specific regulator. The method consists of several steps. In the first step formaldehyde is used to fixate the regulator in the DNA *in vivo* since formaldehyde can produce reversible protein-DNA crosslinks. In a second step the cells get disrupted and the DNA will be sonicated into fragments. In the third step, the DNA fragments that are crosslinked to the regulator will be enriched via immunoprecipitation with a specific antibody directed against the regulator. Then the crosslinks will be reversed and the enriched DNA will be identified with for instance a quantitative PCR.

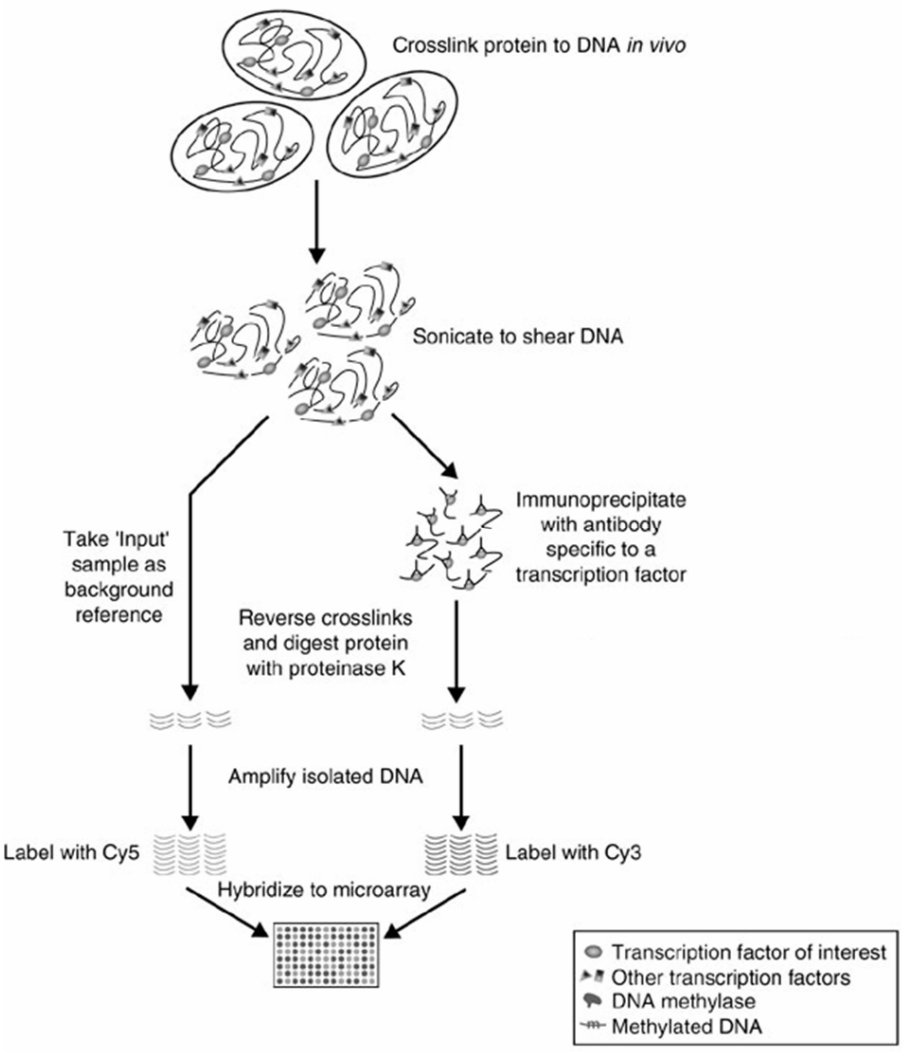


Figure 2- 5: ChIP-chip methodology. First, the transcriptional regulator is tagged with an epitope that can be recognized by antibodies. The proteins that are bound to DNA are fixed and the DNA is isolated and sheared by sonication. An antibody specific to the epitope is used for immunoprecipitation. Then the regulator will be removed from the DNA. As a result, promoter regions of genes that were bound by this regulator are enriched. The enriched DNA is labelled with Cy3, the background reference DNA with Cy5. Both enriched and reference DNA sets are hybridized to a microarray to identify the promoters that were bound by the regulator *in vivo*. Taken from Taverner *et al* (2004).

Instead of identifying the enriched DNA via a quantitative PCR, the enriched DNA can also be amplified and labeled with for instance Cy5. A pool of non-enriched DNA is labeled with Cy3. Both the enriched and non-enriched set of DNA can subsequently be hybridized to a microarray (see Figure 2-5). With this ChIP on chip or ChIP-chip method the binding of a specific regulator in a specific environmental condition to the promoter region of a gene can be investigated on a genome-scale in one single experiment (Lee *et al*, 2002; Orlando, 2000; Ren *et al*, 2000). ChIP-chip data are always measured in a specific condition since a regulator will only bind to the promoter region of a gene when the regulator is available and active and the promoter region of the target gene is accessible.

Lee *et al*. (2002) used the ChIP-chip technique to identify protein-DNA interactions for 106 regulators in yeast in rich medium conditions (Lee *et al*, 2002). In a subsequent study (Harbison *et al*, 2004), this set was extended to 203 regulators measured in rich medium conditions; 84 of these 203 regulators were also investigated in other conditions. In *E. coli* the regulatory interactions of specific regulators were investigated, such as CRP, FNR, MelR and the nucleoid associated transcription factors IHF, Fis, and H-NS (Grainger *et al*, 2005; Grainger *et al*, 2006; Grainger *et al*, 2007).

2.4 Complementarity of the data

Genome-wide expression data sets have typically been used in cluster analysis to detect sets of genes with a similar expression profile. Although these co-expressed genes are potentially co-regulated, co-expressed genes can also have a different regulatory program.

ChIP-chip analysis provide another way of studying regulation: physical interactions between a transcription factor and DNA regions are identified. However, one should keep in mind that the binding of a regulator in a specific condition does not necessarily imply that this regulator will regulate the gene in this condition. In the case of combinatorial control, for instance, an additional regulator might be required. Alternatively the presence of a particular ligand or metabolite may be needed for the activation of the regulator. For this reason, ChIP-chip data can contain false-positives. On the other hand, some interactions of a regulator with its target genes only occur in very specific conditions, that may not (yet) be present in the ChIP-chip compendium.

Regulatory motif data also provide information on transcription regulation. The motifs are sites in the DNA that are recognized by a particular transcription factor. They indicate that a regulator might regulate a particular gene by binding to the motif in the promoter region of that gene. Regulatory

motifs are however often derived from *in silico* techniques and many false positives and negatives may be present.

All these approaches are useful for studying the transcriptional network but their individual power is limited because they only provide partial information on the network: expression data only provide indirect evidence for regulation, ChIP-chip data indicate regulator binding but the binding might not be functional, and regulatory motifs only correspond to potential binding sites that might never be bound by a regulator. In addition, high-throughput data inherently contain noise.

In spite of these shortcomings, these three data types offer complementary information on the transcriptional network. The goal of this thesis is therefore to integrate them and thus exploit as much available information as possible to obtain a more comprehensive view on the transcriptional network.

Chapter 3

Collection of omics data

Access to publicly available high-throughput data is of utmost importance to many bioinformaticians. Several network inference methods and module detection tools have been developed that use microarray data to reconstruct the regulatory network. Some of these methods combine microarray data with other available high-throughput data like ChIP-chip or motif data. Therefore, a first and crucial step in the reconstruction of the transcriptional network is the compilation and pre-processing of omics data compendia.

In chapter 2 we already discussed the different types of omics data that we will use for the inference of the transcriptional network (chapter 6 and 7). In this chapter we describe how we collected our data and created our data compendia for two model organisms: *Escherichia coli*, a prokaryote bacterium species and *Saccharomyces cerevisiae*, a eukaryote yeast species (see Figure 3- 1). Since both organisms are model organisms, they have been extensively studied and a substantial amount of high-throughput data is available for both organisms. In chapter 6 and 7 we describe how this collected data can be used for the inference of regulatory networks.

The collected yeast data consisted of two well-studied expression data sets in addition to a ChIP-chip data set and a regulatory motif compendium. Several network inference tools have been applied on the same data sets. By using this data we could therefore more easily compare the performance of our algorithm ReMoDiscovery with other available tools. In addition, because the data are so well-known, we were able to asses the biological relevance of our network reconstruction tool. This enabled us to prove that our "association rule mining" based method was well suited for data integration and detection of regulatory modules (chapter 6).

One of the disadvantages of our tool is that it assumes that a set of co-regulated genes is co-expressed in all experimental conditions. This limits the application of ReMoDiscovery to smaller and less heterogeneous expression data sets. To overcome this shortcoming, we developed a second network inference tool called DISTILLER that includes a condition selection

strategy: DISTILLER fully exploits all available microarray data (chapter 7). We applied DISTILLER on a collection of available microarray data for *E. coli*, together with motif data present in RegulonDB (Salgado *et al*, 2006). To our knowledge, the *E. coli* expression compendium of 870 arrays that we constructed was the first cross-platform compendium of *E. coli*.

Collecting and pre-processing of microarray compendia is a time-consuming and tedious task. Therefore we developed a software tool called GeneReg, in collaboration with the Computer Science Department at the K.U.Leuven (David Pastor Escuredo, Álvaro Cortés Calabuig and Prof. Marc Denecker):

Pastor D., Cortes-Calabuig A., Lemmens K., De Moor B., Marchal K., Denecker M. (2007) GeneReg: Integration of Experimental Data on the DNA Transcription Process. Proceedings of BNAIC 2007.

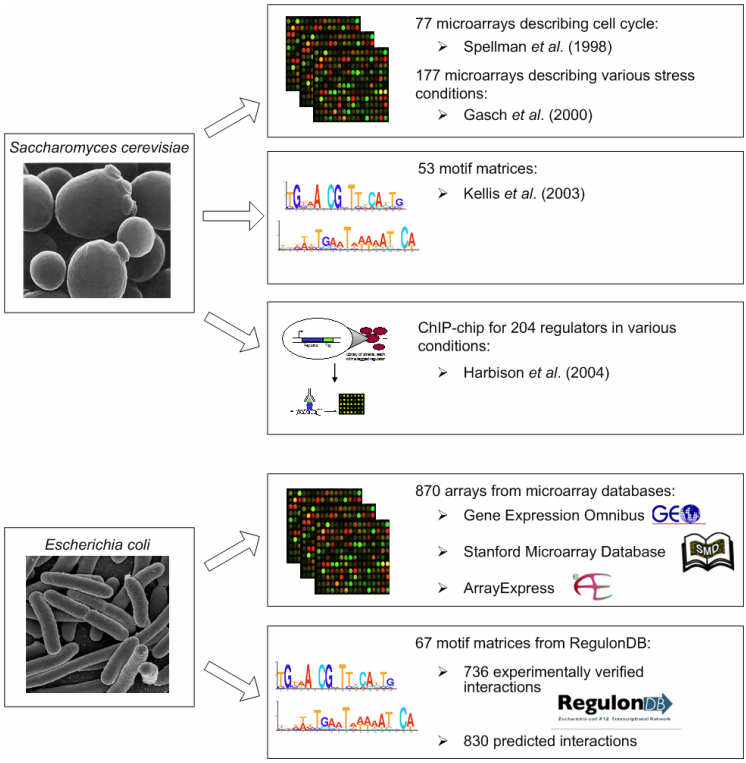


Figure 3- 1: The omics data that were collected for two model organisms, *S. cerevisiae* and *E. coli*. Microarray data, ChIP-chip data and motif data were available for yeast. The collected data are well-studied data sets that serve as benchmark for our module detection tool (chapter 6). For *E. coli*, motif data and microarray data are available. Data were collected from large databases, such that a maximal exploitation of the available data became possible. The *E. coli* data are used as input for DISTILLER (chapter 7).

3.1 *Saccharomyces cerevisiae* data

3.1.1 Expression data

The Gasch and Spellman data sets (Gasch *et al*, 2000; Spellman *et al*, 1998) are two well-known and well-studied yeast expression data sets. These data were used as microarray benchmark sets. We downloaded the data[§] and rescaled the expression profiles by subtracting the mean of each profile and dividing by the standard deviation across the different conditions or time points.

The Spellman dataset (Spellman *et al*, 1998) contains 77 spotted arrays describing the dynamic changes of yeast genes during the cell cycle. Three independent methods for synchronization of the cell cultures were used to generate the data: 1) α pheromone to arrest MATa cells in G1; 2) centrifugal elutriation to obtain small G1 cells; 3) a temperature-sensitive mutation, *cdc15-2*, which, at the restrictive temperature, arrests cells late in mitosis was also used. Each of these methods introduces characteristic artifacts. For instance, use of pheromone has regulatory consequences characteristic of mating, whereas use of temperature-sensitive mutants can cause heat shock.

The Gasch dataset (Gasch *et al*, 2000) consists of 173 spotted arrays, examining gene expression behavior during various stress conditions: heat shock, hydrogen peroxide, superoxide generated by menadione, a sulfhydryl oxidizing agent (diamide), and a disulfide reducing agent (dithiothreitol), hyper-osmotic shock, amino acid.

3.1.2 Transcription factor binding site data

The motif data were obtained from a comparative genome analysis between distinct yeast species (Kellis *et al*, 2003). Kellis *et al* (2003) only provide a list of motif predictions in promoter regions of genes; no score is given to indicate the likelihood of a all motif-gene combinations. For our analysis, however, we prefer a score for each motif-gene combination since this allows to vary the threshold on the score and in addition, it allows us to predict new motif-gene interactions. For each motif-gene combination, the motifs that are available online as regular expressions were therefore downloaded and transformed into their corresponding weight matrices: for each motif, the 20 *Saccharomyces cerevisiae* genes in which the motif was

[§] http://www-genome.stanford.edu/yeast_stress and http://cellcycle-www.stanford.edu

most reliably detected according to the scoring scheme of Kellis *et al* (2003) were selected. The intergenic sequences of these genes were subjected to motif detection based on Gibbs sampling (Thijs *et al*, 2002). If the statistically overrepresented motif in this set of putatively co-expressed genes corresponded to the motif that was detected by the comparative motif search of Kellis *et al* (2003) the motif model was retained. Out of the 71 putative motifs described by Kellis *et al* (2003), 53 motifs were retained. The weight matrices corresponding to these motifs were subsequently used to screen all intergenic sequences of yeast using MotifLocator (Marchal *et al*, 2004) . The higher the score of a motif hit in a gene, the more likely it will be a true instance.

3.1.3 ChIP-chip data

Lee *et al* (2002) performed a genome-wide ChIP-chip analysis for 106 regulators in rich medium. Harbison *et al* (2004) extended this data set to 204 regulators profiled in rich medium. In addition, they profiled 84 regulators in at least one environmental condition other than rich medium. Harbison *et al* (2004) combined these ChIP-chip results with phylogenetically conserved sequences and prior knowledge to search for regulatory motifs. Highly significant motifs could be identified for 116 regulators.

Genome-wide ChIP-chip data performed by Harbison *et al* (2004) were downloaded[**] from their website and used as input for our module detection algorithm.

3.2 *Escherichia coli* data

3.2.1 Expression data

Many *E. coli* microarrays are publicly available in three major microarray databases: Stanford Microarray Database or SMD (Demeter *et al*, 2007), Gene Expression Omnibus or GEO (Barrett *et al*, 2007), and ArrayExpress or AE (Parkinson *et al*, 2007). To obtain a large and heterogeneous compendium of microarrays, we collected all available *E. coli* data from these three databases††. Additionally, we added four microarray

[**] http://www.broad.mit.edu/annotation/fungi/comp_yeasts/motiflist.html

†† http://www.ncbi.nlm.nih.gov/geo/; http://genome-www5.stanford.edu/; http://www.ebi.ac.uk/microarray-as/ae/

experiments described in literature that were available as supplementary information. After removing redundant information in the microarray databases, we obtained a total of 870 microarrays, i.e. the first cross-platform heterogeneous microarray compendium for *E. coli*. Table 3- 1 gives an overview of the contribution of the different platforms to the compendium. A full description of the microarray compendium can be found in Table 3- 2.

To make the data comparable between different experimental conditions and platforms the following normalization procedures were conducted:

- If possible, raw intensities were preferred as data source over normalized data provided by the public repository.
- Dual-channel data were Loess fitted to remove nonlinear, dye-related discrepancies (Yang *et al*, 2002). No background correction procedures were performed to avoid an increase in expression log ratio variance for lower, less reliable intensity levels.
- Whenever raw data were available, single-channel data were first normalized per experiment with RMA (Irizarry *et al*, 2003). Log ratios were then created for the single-channel data in order to combine them with the dual channel measurements. For each single-channel array, expression log ratios were computed by comparing the normalized values against an artificial reference array. This

Table 3- 1: Overview of the contribution of the different platforms and data sources to the compendium. The number of arrays that were collected from each data source and platform are indicated. Both single channel and dual channel arrays were collected from literature and the different databases, Gene Expression Omnibus (GEO), Stanford microarray database (SMD) and ArrayExpress (AE).

		data source			
		GEO	**SMD**	**AE**	**Literature**
single channel	**Affymetrix**	183	-	-	115
	P33	46	-	-	63
dual channel	**spotted DNA/cDNA**	278	44	12	3
	spotted oligonucleotide	110	-	16	-

Table 3- 2: Overview of the different experiments that we collected in our expression compendium. **Exp. ID:** a unique identifier for each experiment included in the compendium; **Experiment:** the name of each experiment as obtained from the corresponding database or according to the corresponding paper; **No. of arrays:** the number of arrays in each experiment; **Article:** the PubMed ID of the paper in which the experiment is described; **Data type:** the platform of the experiment, i.e. Affymetrix (affy), filter-based arrays (P33), spotted arrays (cDNA or oligo); **Data source:** the source from which the data was obtained: literature, the Gene Expression Omnibus (GEO), the Stanford Microarray database (SMD) or ArrayExpress (AE).

Exp. ID	Experiment	No. of arrays	Article	Data type	Data source
1	Palsson_2005_Genome_Res	99	16204189	affy	Literature
2	Gossett_2005_J_of_Bact	16	15659676	affy	Literature
3	GSE1121	43	15129285	affy	GEO
4	GSE2037	15	15705577	affy	GEO
5	GSE2928	12	16199573	affy	GEO
6	GSE2999	24	NO	affy	GEO
7	GSE3105	12	16199566	affy	GEO
8	GSE3937	4	16597943	affy	GEO
9	GSE4511	15	15601715	affy	GEO
10	Chang_2002_Mol_Biol	63	12123445	P33	Literature
11	GSE2827	6	16549663	P33	GEO
12	GSE533	12	12672900	P33	GEO
13	E-MEXP-222	6	15659689	oligo	AE
14	E-MEXP-244	4	NO	oligo	AE
15	E-MEXP-245	6	NO	oligo	AE
18	E-MEXP-267	12	NO	cDNA	AE
19	Blattner_2005_Genome_Res	3	14597655	cDNA	Literature
20	GSE33	16	11967071	P33	GEO
21	GSE1421	3	15466047	oligo	GEO
22	GSE1780	9	15817786	oligo	GEO
23	GSE1981	3	15821913	oligo	GEO
24	GSE2095	4	15647275	oligo	GEO
25	GSE3250	81	16141204	cDNA	GEO
26	GSE3437	21	16336047	cDNA	GEO
27	GSE3591	55	16377617	oligo	GEO
28	GSE4112	4	14526013	cDNA	GEO
29	GSE4321	16	16818608	cDNA	GEO
30	GSE4357	6	16626502	cDNA	GEO
31	GSE4358	6	16626502	cDNA	GEO
32	GSE4359	7	16626502	cDNA	GEO
33	GSE4360	4	16626502	cDNA	GEO
34	GSE4361	4	16626502	cDNA	GEO

35	GSE4362	6	16626502	cDNA	GEO
36	GSE4363	11	16626502	cDNA	GEO
37	GSE4364	7	16626502	cDNA	GEO
38	GSE4365	5	16626502	cDNA	GEO
39	GSE4366	5	16626502	cDNA	GEO
40	GSE4367	5	16626502	cDNA	GEO
41	GSE4368	5	16626502	cDNA	GEO
42	GSE4369	5	16626502	cDNA	GEO
43	GSE4370	7	16626502	cDNA	GEO
44	GSE4371	7	16626502	cDNA	GEO
45	GSE4372	4	16626502	cDNA	GEO
46	GSE4373	6	16626502	cDNA	GEO
47	GSE4374	6	16626502	cDNA	GEO
48	GSE4375	6	16626502	cDNA	GEO
49	GSE4376	6	16626502	cDNA	GEO
50	GSE4377	5	16626502	cDNA	GEO
51	GSE4378	4	16626502	cDNA	GEO
52	GSE4379	3	16626502	cDNA	GEO
53	GSE4380	6	16626502	cDNA	GEO
54	GSE4408	16	17009874	cDNA	GEO
55	GSE4417	4	16626502	cDNA	GEO
56	GSE2129	4	15647275	oligo	GEO
57	Courcelle_2001_Genetics	15	11333217	cDNA	SMD
58	Khodursky_2000_24_okt_PNAS	26	11027315	cDNA	SMD
59	Tani_2002_15_okt_PNAS	3	12374860	cDNA	SMD
61	GSE5356	6	16952965	oligo	GEO
62	GSE5139	4	17189370	oligo	GEO
63	GSE5137	4	17189370	oligo	GEO
64	GSE5076	4	17189370	oligo	GEO
65	GSE5075	4	17189370	oligo	GEO
66	GSE4706	6	16804185	oligo	GEO
67	GSE5084	4	17189370	oligo	GEO
68	GSE4941	12	16849805	P33	GEO
70	GSE4556	15	17026754	affy	GEO
71	GSE3905	8	17222132	affy	GEO
72	GSE4724	9	17074904	affy	GEO
73	GSE5904	2	NO	affy	GEO
74	GSE5239	24	NO	affy	GEO

artificial reference array was constructed on a per experiment basis by taking the median expression of each gene across all arrays in the corresponding experiment. When deemed necessary (e.g. experiments normalized by MAS5.0 for which the raw data was not available), a Loess fit was performed on these log ratios. To ensure that the artificial reference was not altered by this intensity dependent non-linear rescaling, the artificial reference expression levels were chosen for the average log intensity (instead of the mean expression levels of the respective array and the artificial reference).

- To ensure comparability between arrays, gene expression profiles were median centered across arrays that share the same reference. An additional variance rescaling of the gene expression profiles was performed to render genes with differing magnitudes of expression changes more comparable.

Analysis of the obtained modules (see chapter 7) indicated that the array composition of the retrieved modules is not biased towards arrays from a specific platform: the average percentage of arrays from a particular platform in a module approximates the overall percentage of arrays of that platform in the total compendium. Only 6 arrays (from 5 different experiments) are not included in any module. This suggests that the pre-processing of the microarray compendium is indeed appropiate.

3.2.2 Transcription factor binding site data

RegulonDB (Gama-Castro *et al*, 2008) is currently one of the largest databases with curated knowledge of the transcriptional regulatory network of *E. coli*. It contains information on experimentally verified binding sites. In chapter 7 we explain how the motif data that we collected from this database as described below are used in combination with the previously described microarray compendium for the inference of the transcriptional network of *E. coli* (chapter 7). In collaboration with Prof. Collado-Vides (UNAM, Cuernavaca, Mexico) we were able to fully exploit the information present in the RegulonDB database. Our goal is to obtain a score for all motif-gene combinations indicating the probability that a particular motif instance is present in the promoter region of a gene. The motif-gene interaction predictions that follow from this scoring schema can be used to extend the transcriptional network of *E. coli*.

To predict novel binding site instances, we first downloaded motif weight matrices corresponding to the binding sites of 67 regulators from the

RegulonDB website (version 5.6)[‡‡] (Salgado *et al*, 2006). Upstream regions on the direct strand of all annotated *Escherichia coli* K12 (NC_000913) genes were screened with these motif matrices in order to find novel motif instances. These upstream regions include the intergenic region between the gene of interest and its upstream gene together with the first 50 nucleotides of the genes' coding region. If an upstream region was smaller than 150 nucleotides, it was extended with the region overlapping the coding region of the previous gene until a maximum of 150 nucleotides was reached. For motif screening and p-value calculations of the identified motif instances we used the method of Hertzberg *et al* (2005). The p-values were used to construct the 'motif matrix', a binary matrix which assigns a motif instance to a gene whenever the gene's upstream sequence contains at least one instance of the motif, with a p-value below a threshold of 0.001.

In addition to these predictions, we also derived the known binding sites from RegulonDB. Whenever a motif instance in the promoter region of a gene was experimentally confirmed according to RegulonDB, its corresponding regulator-target interaction was set to '1' in the motif matrix, irrespective of its motif screening p-value. The 34 motif instances present in the upstream sequences of non-coding RNAs (tRNA or miscellaneous RNA) were omitted. The resulting motif matrix contains a total of 736 experimentally verified and 830 predicted motif instances. This motif matrix together with the expression compendium is used as input for our network reconstruction tool DISTILLER in chapter 7.

Note that since only the first operon gene will contain the motif in its promoter region, the interactions presented in this motif matrix will not involve downstream operon genes.

3.3 GeneReg

As mentioned previously, we collected microarray data from publicly available databases to create our *E. coli* microarray compendium (section 2.2.1). Although these public microarray databases are very useful, it is hard for a biologist to download each single microarray (one by one) from these databases or to keep track of new experiments. Moreover, the data is distributed over several databases with overlapping content and microarray data are available in different formats. Heterogeneous biological data, like regulatory motifs or ChIP-chip data, are distributed over different web sites and databases, making it even harder for the biologist to get access to all these publicly available data simultaneously. In addition, the information is

[‡‡] http://regulondb.ccg.unam.mx/

not only stored in different formats and data models, but also contains redundant, incomplete and/or inconsistent data. Because creating and maintaining compendia of all available data, like we did for *E. coli*, is such a difficult and time-consuming task, most network reconstruction algorithms are currently being benchmarked (and applied) on subsets of the data only. However, it is crucial to incorporate as much data as possible to obtain more insight into the properties of different network inference methods and, more importantly, into the nature of the transcriptional network itself.

Therefore we developed GeneReg, a tool to automate the data collection process (Pastor *et al*, 2007). GeneReg includes functionality for data retrieval from the data source, pre-processing of the information (by means of wrappers), global database population and finally the querying process. The application runs on a Windows platform and is based on a MySQL 5.1 database management system. In its current version, GeneReg is able to store ChIP-chip data and regulatory motif data and to extract and store information from the Stanford Microarray Database. The final goal is to make the application more generic, enabling data extraction for any organism and from other databases like the Gene Expression Omnibus or ArrayExpress. GeneReg was developed in cooperation with the department of Computer Science at the K.U.Leuven (David Pastor Escuredo, Álvaro Cortés Calabuig and Marc Denecker) and published in Pastor *et al* (2007).

With the growing number of microarrays in particular, stored in different databases and in different formats, software like GeneReg becomes increasingly useful and important to fully exploit all available data.

Chapter 4

Data integration and network reconstruction

In the previous chapters we already discussed the different types of omics data and described how we collected them into a compendium (chapter 2 and 3). After data collection, omics data can be used for network reconstruction. Different types of biological networks exist such as the transcriptional network, the protein-protein interaction network or the metabolic network. In this thesis two methods for the inference of the transcriptional network have been developed. Before introducing these methods, we discuss related algorithms for transcriptional network inference (section 4.1). In chapter 6 and 7, our two methods for network reconstruction will be compared with these state-of-the-art algorithms. A review on the different network inference methods was also published in:

> Van den Bulcke T., Lemmens K., Van de Peer Y., Marchal K. (2006) Inferring Transcriptional Networks by Mining Omics Data. Current Bioinformatics, vol. 1, no. 3, pp. 301-313.

One class of methods are the biclustering algorithms (section 4.1.1.2). During this PhD, a query-driven biclustering method was developed and published in Bioinformatics:

> Dhollander T., Sheng Q., Lemmens K., De Moor B., Marchal K., Moreau Y. (2007) Query-driven module discovery in microarray data. Bioinformatics, vol. 23, no. 19, pp. 2573-2580.

We also explain in this chapter why we explore yet another approach for network reconstruction, based on association rule mining algorithms (section 4.2). Finally, many difficulties arise when trying to compare network reconstruction tools. These problems make it hard for a biologist to know which tool to choose. We discuss these problems and possible solutions in the final section of this chapter (section 4.3).

4.1 Inference of regulatory transcriptional networks

With the advent of high-throughput data, molecular biology has evolved from a rather data-poor area to an extremely data-rich research area. Because of this huge amount of available data, top-down or data-driven network inference methods flourish (see also Chapter 1, section 1.4). Deciphering a transcriptional network is often looked at as a reverse-engineering problem, where the goal is to understand how a biological network functions by examining data sets produced by the network.

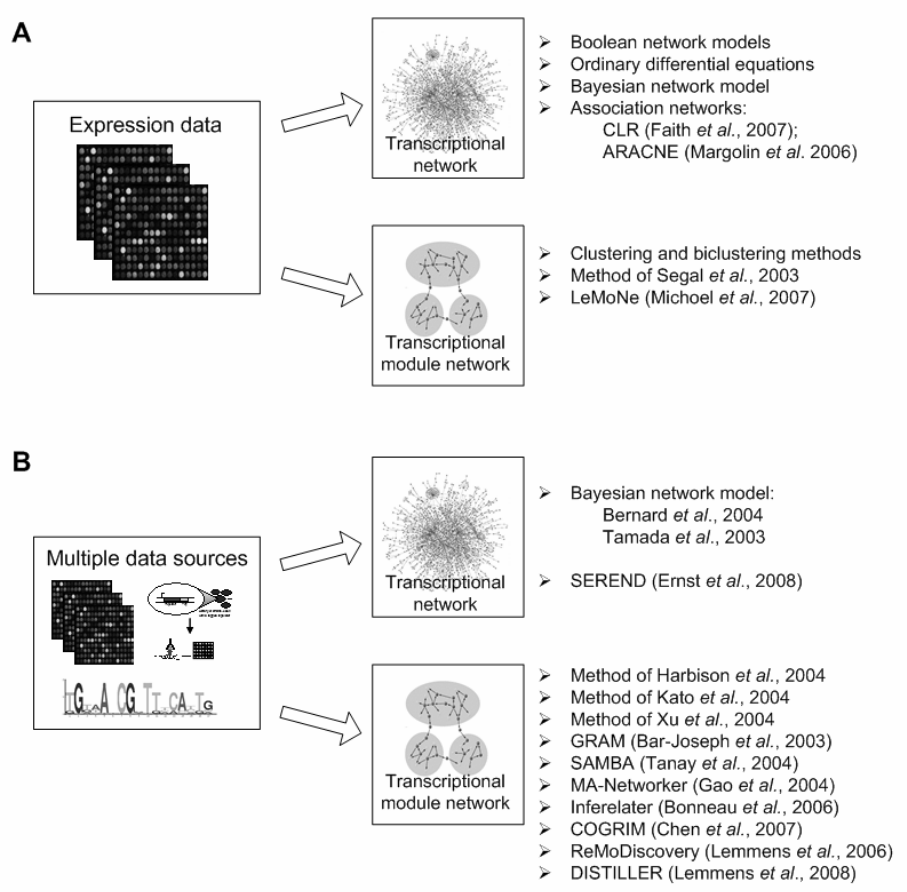


Figure 4- 1: Different network reconstruction methods. The transcriptional network can be inferred based on only microarray data (Panel A) or by using multiple heterogeneous data sources, for instance, microarray data, ChIP-chip data or regulatory motif data (Panel B). Methods will either search for the transcriptional network directly or they can identify transcriptional modules. The methods that we discuss in this section are mentioned.

Two main approaches exist to unravel the regulatory network (Bernard and Hartemink, 2005). In the first approach, the regulatory network models are learned directly. Usually this kind of inference algorithms requires more high-throughput data than currently available. In other words, the network inference problem is underdetermined. A second approach exploits the notion of modules or the basic building blocks of the network and generally requires less data. These modules are sets of co-expressed genes that are regulated in concert. Usually the genes of a module are only co-expressed in a limited set of conditions, i.e. modules are condition-dependent. Modules can exhibit overlap in gene content, regulator content or experimental condition content. Both the network inference and module network inference approaches can either use expression data only or combine multiple data sources. Figure 4- 1 illustrates this dual classification. Various network and module network inference methods are described below.

4.1.1 Network inference using expression data

4.1.1.1 Network inference

In contrast to the transcriptional network, a gene network usually does not describe physical interactions between a transcription factor and a gene but rather the regulatory influence of a particular RNA transcript on another RNA transcript. Since only a limited set of transcripts influences a given transcript, the networks are usually sparse. These networks can give a global view on the regulatory network since they are not restricted to regulator-promoter interactions and the regulator transcripts may exert their regulatory effect indirectly through proteins, metabolites, etc. Four major classes of reverse engineering algorithms exist that learn the network directly: Boolean networks, ordinary differential equation networks, Bayesian networks and association networks (Bansal *et al*, 2007; Gardner and Faith, 2005).

Boolean network models describe the transcript levels as binary values. A gene is either in the “ON” state or “OFF” state. Each gene is regulated by other genes through general logical or Boolean functions. Solving Boolean networks requires a large amount of experimental data. For a fully connected Boolean network, the algorithm must receive samples of all possible combinations of input expression states (D'haeseleer *et al*, 2000). Of course, biological networks are never fully connected, but even for more sparse networks the data requirements are still considerable. Examples of Boolean regulatory network models are (Akutsu *et al*, 1999; Akutsu *et al*, 2000; Liang *et al*, 1998).

Reverse-engineering methods based on ordinary differential equations (ODE) form a second category. ODEs relate the change in expression level of a gene to the expression levels of the other genes. The goal is to detect

which RNAs regulate particular transcripts and to determine the parameters of the interaction function - for instance a sigmoidal or linear function – that relates them. Examples of ODE-based reverse-engineering algorithms include NIR (Gardner *et al*, 2003), MNI (di Bernardo *et al*, 2005) and other methods (Bonneau *et al*, 2006; Tegner *et al*, 2003; van Someren *et al*, 2006).

A third type of network inference methods uses the Bayesian network model. A Bayesian network is a graphical model that infers probabilistic relationships between random variables, X_i. These relationships are encoded in a directed acyclic graph of which the nodes represent the random variables that correspond to genes. Inferring the network comes down to determining the correct graph structure and modeling the relationships between the random variables by using the expression data. In fact, such a relationship is represented by the probability distribution of the expression level of a child node, given the expression levels of its parent nodes in the graph (conditional probability distributions). From a computational point of view it is intractable to optimise over all possible graph structures. Instead, different heuristic search algorithms have been used although they may lead to a local optimum. In addition, often several equally plausible graph structures exist and hence the problem of network inference is underdetermined. To bias the inference towards the real underlying biological graph structure, prior knowledge can be incorporated (see section 3.1.2.2). Several Bayesian networks have been developed (Friedman *et al*, 2000; Hartemink *et al*, 2001; Yu *et al*, 2004b).

A last category of networks are the so-called association networks. They assign interactions to pairs of transcripts that have a high statistical similarity in their expression profile. To measure this similarity, the Pearson correlation coefficient which assumes a linear relationship between the two expression levels can be used. Mutual information is another similarity measure that has the extra advantage that it does not make any assumptions on the form of dependence of the variables. Because only similarity is measured, there is no information about the direction of the connection.

One example of an association network method is ARACNE (Margolin *et al*, 2006). In a first step, ARACNE calculates the mutual information between all genes and creates a graph displaying an edge between two genes if the mutual information was higher than a threshold. Subsequently some indirect interactions are pruned by considering all triplets. A triplet is a set of three genes that are fully connected. The algorithm uses the data processing inequality to decide upon removal of the edge with the lowest mutual information. ARACNE has been applied on mammalian data.

In chapter 7 (section 7.3.5), we compare our network reconstruction method with the context of likelihood relatedness (CLR) association network method of Faith *et al* (2007). This method has been applied on an *E. coli* Affymetrix compendium (Faith *et al*, 2007). The underlying algorithm starts from a list of known transcription factors and an interaction between a regulator and a target gene is predicted whenever the mutual information between the expression profiles of the target and the regulator exceeds a certain threshold. The choice of this threshold determines a tradeoff between the true positive and false positive rate. Therefore, CLR compares the mutual information of the transcription factor-gene pair to a background distribution of mutual information scores. This distribution is determined by all possible transcription factor-gene pairs that include either the transcription factor or the target gene. If the mutual information of a transcription factor-gene pair is significantly higher than the background distribution, this interaction is considered a true interaction. By using this clever thresholding, many false positive interactions could be removed. CLR identified 1079 regulatory interactions, of which 338 were in the previously known network and 741 were novel predictions. By using ChIP, the authors were able to confirm 21 novel interactions.

4.1.1.2 Module network inference

Instead of inferring the regulatory network model directly, alternative methods identify transcriptional modules based on expression data.

In particular, clustering and biclustering algorithms can reveal the modularity of the transcriptional network. A cluster is defined as a set of genes with a similar expression profile in all experimental conditions and can be detected by clustering algorithms (Eisen *et al*, 1998; Tamayo *et al*, 1999; Tavazoie *et al*, 1999). Genes are however not expected to be co-expressed in all conditions of large expression compendia. Therefore several algorithms are available that detect biclusters or sets of genes that are co-expressed in only a set of experimental conditions (Cheng and Church, 2000; Ihmels *et al*, 2002; Madeira and Oliveira, 2004; Murali and Kasif, 2003; Sheng *et al*, 2003; Tanay *et al*, 2002). In query-driven biclustering, a bicluster or a set of biclusters is detected around one specific gene or a set of genes that is known to be involved in the same pathway (Bergmann *et al*, 2003; Dhollander *et al*, 2007). The goal of query-driven biclustering is to recruit additional genes that are potentially involved in the same function as the query gene(s) and to simultaneously identify the conditions in which the genes are co-expressed.

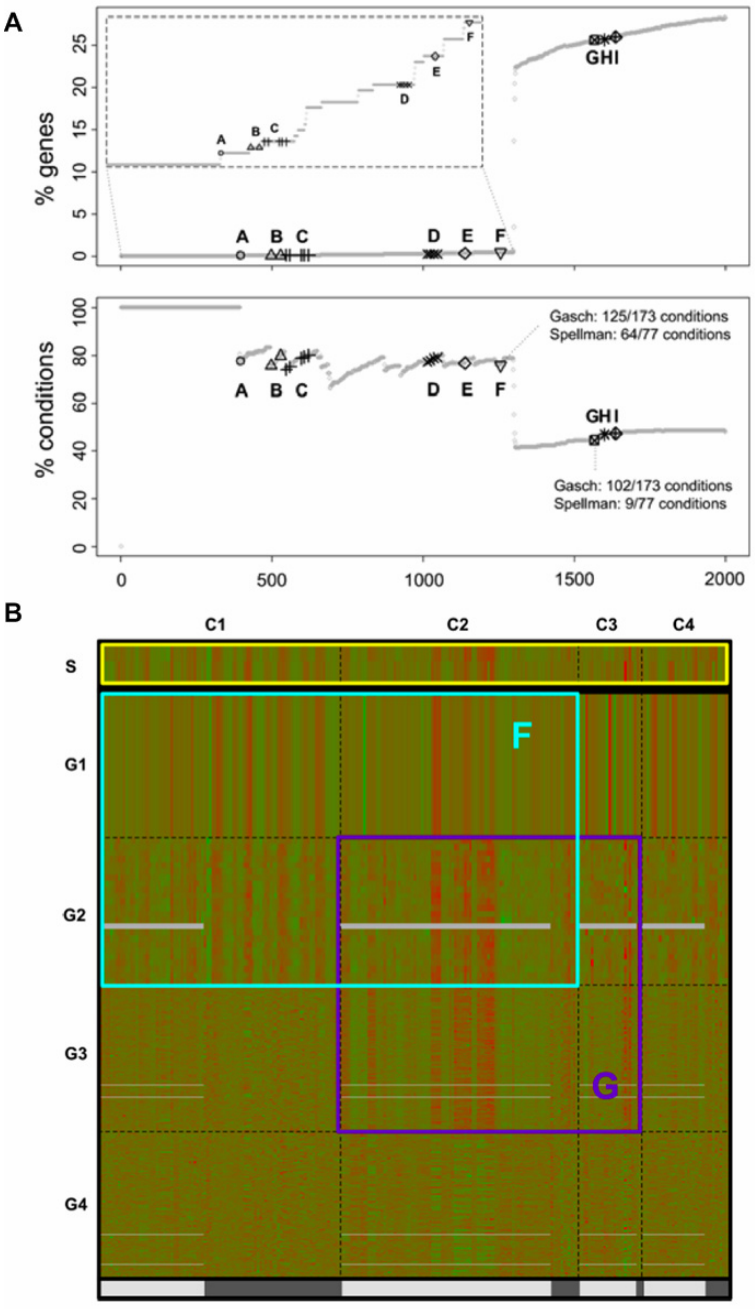


Figure 4- 2: Caption on page 45

During this PhD, such a query-driven biclustering method was developed (Dhollander *et al*, 2007). The sets of query genes on which this approach was applied, were obtained from the results of ReMoDiscovery (Lemmens *et al*, 2006), a transcriptional network inference method that we developed and benchmarked during this PhD (see chapter 6). An example of the biclusters that were detected around a set of query genes is shown in Figure 4- 2.

Although clustering and biclustering algorithms are useful to study modularity, they do not explicitly detect the regulators controlling the set of co-expressed genes. These associated regulators have to be determined indirectly, for instance by detection of regulatory binding sites in the upstream regions of the genes in a cluster or bicluster.

Figure 4- 2: Resulting biclusters around one query of genes when using the query-driven biclustering approach of Dhollander *et al* (2007). The goal of the query-driven approach is to start from a set of genes of interest, i.e. the query genes, and to recruit additional genes of which the expression profile is similar to the expression profile of the query gene(s). Simultaneously, the condition content of the bicluster is adjusted. The evolution of the bicluster (gene and condition content) for decreasing resolution parameters is shown in panel A. Initially, the expression of the recruited genes must be very similar to the expression values of the query genes in the bicluster conditions. For less stringent resolution settings, the co-expression criterion is less strict. Letters A-I indicate automatically selected statistically interesting modules at various resolutions. Modules with the same gene content were grouped in one letter symbol. Functionally enriched GO categories for the selected modules range from DNA-dependent DNA replication (bicluster A-C) over mitotic cell cycle (D) over cell division and cytokinesis (E–F) to ribosome biogenesis and assembly (G–H–I) as the parameters became less stringent. Panel B shows the expression profiles of bicluster F (cyan) and G (purple), illustrating the transition from a cell cycle bicluster into a ribosomal bicluster. The profiles of the seed genes (S) are framed in yellow. As expected, the seed genes can be traced back to the intersection (G2) of the gene sets belonging to both biclusters. The bar on the bottom indicates which conditions are from the Gasch (light gray) and Spellman (dark gray) data sets. For clarity, all gene sets are stretched to similar sizes (S: 2 genes, G1: 1 gene, G2: 24 genes, G3: 1550 genes; G4: 4569 genes; C1: 100 conditions, C2: 90 conditions, C3: 21 conditions, C4: 39 conditions). Missing values in the expression array are indicated in gray.

In contrast to clustering and biclustering algorithms that cannot reveal the regulatory program of the sets of co-expressed genes, Segal *et al* (2003) have developed the *module networks method* that searches simultaneously for a partition of the genes into modules and for a regulatory program for each module that can explain the expression behaviour of the genes in the modules. The regulatory program is defined as the set of regulators that regulate the module. It specifies the expression profile of the genes in the module as a function of the expression of the regulators. Thus, the approach relies on the assumption that the regulators are transcriptionally regulated themselves and that their expression level gives information about their activity level. This assumption is not always true, for instance, constitutively expressed regulators that are activated post-transcriptionally. In addition, it was shown in yeast that over 80% of the tested pairs of expression profiles between regulators and targets were not significantly correlated (Herrgard *et al*, 2003). Another disadvantage of the method of Segal *et al* (2003) is that for algorithmic reasons the modules are not allowed to overlap. Segal *et al* (2003) applied their method on a yeast expression data sets (Gasch *et al*, 2000), revealing 50 modules and their associated regulatory program. The biological potential of their method was proved by experimentally validating three of the hypotheses that followed from their predictions.

LeMoNe (Michoel *et al*, 2007) adapted the method of Segal *et al* (2003) by incorporating an alternative strategy for inferring regulatory programs. The regulatory program is identified based on the expression profile of the genes in the module and does not depend anymore on the expression profiles of the potential regulators. This improvement did not only allow for a faster learning of the regulatory programs, but also missing or hidden regulators (i.e. regulators that are missing from the list of potential regulators) can be handled in this way.

4.1.2 Network inference by data integration

4.1.2.1 Data integration

The advent of high-throughput techniques has allowed large-scale identification of the different cellular entities, their expression patterns, and their biochemical and genetic interactions. In recent years, many researchers have put a lot of effort in extracting insights from these omics data. Although omics data can provide researchers with unprecedented, large-scale views of biological systems, the data from such studies are often incomplete and contain errors. These kind of errors in genomic data can mislead the investigator. A good example are gene expression measurements, which are known to fluctuate between experiments even if the conditions are apparently identical (Baldi and Long, 2001) due to for

instance measurement errors. Another example is the determination of the protein-protein interactome by yeast-two-hybrid techniques that are notorious for false positives and negatives (Ito *et al*, 2000; Ito *et al*, 2001; Legrain *et al*, 2001). In addition, Ito *et al* (2002) (Ito *et al*, 2002) compared two yeast-2-hybrid protein-protein networks (Ito *et al*, 2001; Uetz *et al*, 2000) with each other and remarked that there was little overlap between both studies, most probably due to noise present in both data sets.

These examples illustrate that experimental and biological noise in the individual data measurements can be so prohibitive that each data type alone has limited utility. However, if we assume that the errors in different data sets are independent, then the errors in results that are obtained by combining two data sources can be drastically decreased (Hartemink and Segal, 2005). It has therefore been proposed that these limitations of omics data may be overcome by integrating data obtained from two or more distinct approaches. In addition, different omics data (e.g. genome sequence, transcriptome, proteome, interactome, metabolome) unveil distinct aspects of networks and their integration leads to a more complete insight into these networks (De Keersmaecker *et al*, 2006). Results that are obtained from multiple data sets can therefore lead to novel insights that cannot be found when analyzing only one data source in isolation (Hartemink and Segal, 2005).

4.1.2.2 Network inference by data integration

A first approach for network inference by data integration aims at revealing the complete transcriptional network directly. As mentioned in section 4.1.1.1, Bayesian networks are capable of inferring transcriptional networks. Unfortunately, they require large amounts of expression data to infer statistically reliable networks. Although a huge amount of high-throughput expression data are available at the moment, it is still not sufficient for the network inference when considering hundreds or thousands of genes. As a consequence this shortage of data may lead to false positive relationships between genes. One solution is incorporating prior knowledge, which can be done in a very natural way in Bayesian network models. For instance, Bernard and Hartemink (2004) have incorporated ChIP-chip data, while Tamada *et al* (2003) used regulatory motif information.

A recently published network inference tool called SEREND applies an iterative classification scheme based on logistic regression (Ernst *et al*, 2008). The approach exploits existing knowledge on regulator-target interactions in a semi-supervised way in order to predict novel interactions for these regulators together with the sign of the predicted interactions (positive or negative). SEREND uses a co-expression score and a motif score to classify unknown interactions: a prediction between a regulator and a target gene will be ranked as highly reliable if the predicted target gene

contains a motif instance similar to the motif instances in the known target genes of that regulator and if the target gene is co-expressed with the previously described targets of that regulator. Ernst *et al* (2008) use confirmed *E. coli* regulator-target interactions for 123 regulators and extended the known regulatory network with 981 novel predictions by integrating a microarray compendium of 445 arrays (Faith *et al*, 2007) and regulatory motif data for 71 transcription factors (Gama-Castro *et al*, 2008).

4.1.2.3 Module network inference by data integration

In a second approach for network inference by data integration, transcriptional modules are identified. During this PhD, two methods have been developed for the inference of modules through data integration. One of our approaches was compared with two related methods, GRAM (Bar-Joseph *et al*, 2003) and SAMBA (Tanay *et al*, 2004) (chapter 6). We will therefore discuss these two algorithms and other related work in more detail (see below and Table 4- 1).

Kato *et al* (2004) describe a pragmatic approach to analyze heterogeneous data. They used a sequential procedure to combine ChIP-chip data, microarray expression data and regulatory motifs to find correct combinations of transcription factors and motifs.

The approach of Segal *et al* (2003) was extended to incorporate ChIP-chip data (Xu *et al*, 2004). To assign a regulatory program to a module, the authors first search for a candidate set of regulators. To find this candidate set, both the mutual information between the expression profile of the regulator and the expression profiles of the target genes and the binding p-value of ChIP-chip data are taken into account. This joint scoring strategy allows different weaker indications from separate data sources to be merged to generate evidence for the hypothesis that a gene is part of a module. Using a ChIP-chip compendium (Lee *et al*, 2002) and microarray experiments (Gasch *et al*, 2000; Spellman *et al*, 1998; Xu *et al*, 2004). Xu *et al* (2004) identified 50 modules involving 86 regulators that cover a wide range of cellular processes in yeast.

GRAM (Genetic Regulatory Modules) learns modules from microarray and ChIP-chip data using a sequential data analysis strategy (Bar-Joseph *et al*, 2003). In a first step, ChIP-chip data is used to find a set of genes whose upstream regions are likely to bind a common set of transcriptional regulators. In a second step, the microarray data is used to find a subset of this gene set, containing only those genes whose expression profiles are similar to each other. Finally, the resulting core set is expanded with additional genes that have a small combined p-value for the same set of regulators in the ChIP-chip data. GRAM was applied on the yeast ChIP-chip

Table 4- 1: Overview of methods and types of data used for module inference.

Study	Algorithm	Condition selection	Data sources			
			expression	ChIP-chip	regulatory motifs	other
Cheng and Church, 2000; Ben-Dor *et al*. 2001; Sheng *et al*, 2003; …	Biclustering	x	x			
Bergmann *et al*, 2003; Dhollander *et al*, 2007	Query-driven Biclustering	x	x			
Kato *et al*, 2004			x	x	x	
Segal *et al*, 2003			x			
Michoel *et al*, 2007	LeMoNe		x			
Xu *et al*, 2004			x	x		
Bar-joseph *et al*, 2003	GRAM		x	x		
Gao *et al*, 2004	MA-Networker		x	x		
Tanay *et al*, 2004	SAMBA	x	x	x		growth profiles, protein-protein interactions, literature
Chen *et al*, 2007	COGRIM		x	x	x	
Reiss *et al*, 2006; Bonneau *et al*, 2006	cMonkey and Inferelator	x	x		x	protein expression
Lemmens *et al*,2006	ReMoDiscovery		x	x	x	
Lemmens *et al*,2008	DISTILLER	x	x	x	x	

data (Lee *et al*, 2002) and a compilation of 500 expression arrays. The authors identified 106 modules for which the genes in a module function in a similar pathway.

Tanay *et al* (2004) developed an advanced graph bicluster algorithm, SAMBA (Statistical Algorithmic Method for Bicluster Analysis), to simultaneously integrate expression data, ChIP-chip, protein interaction and phenotypic data. Different data items such as a condition in expression data, a phenotype, a regulator, etc. are treated as gene properties. A bipartite graph is constructed with gene nodes and property nodes: genes are associated with properties if they exhibit the property. In terms of expression data, genes are associated to conditions in which they are over- or under-expressed. The authors then proceed by applying a heuristic search strategy to find heavy subgraphs in the bipartite graph. The identified modules are allowed to overlap in both gene content and property content. SAMBA was applied on yeast data and identified 665 modules covering a broad range of biological processes. Small modules could be clustered into supermodules that characterize common behaviour of the smaller modules.

MA-Networker is an algorithm that combines ChIP-chip data with expression data (Gao *et al*, 2004). In a first step, multivariate regression analysis is used to calculate the transcription factor activity profiles by using ChIP-chip and expression data. Per experimental condition, the transcription factor activity profile indicates whether the regulator is responsible for the expression changes. If a gene is regulated by a regulator, one expects the expression profile of the gene to be similar to the activity profile of the regulator. Therefore, in a second step the correlation is calculated between the regulator's activity profile and the mRNA expression profile of the target gene. Although the method searches for the condition dependent activation of a gene by one regulator it does not yet use this information to determine the concerted action of multiple regulators, i.e. to compile complete regulatory programs. Gao *et al* (2004) applied their approach on a yeast compendium of 750 arrays and ChIP-chip data (Lee *et al*, 2002). They discovered that only 58% of the genes whose promoter region is bound by a regulator are true regulatory targets.

cMonkey detects sets of genes that are co-regulated in a subset of conditions by combining biclustering of expression data and *de novo* motif detection (MEME) in the promoter regions of the set of genes (Reiss *et al*, 2006). Reiss *et al* (2006) applied their method on data of *Halobacterium NRC-1*, *Helicobacter pylori*, *Saccharomyces cerevisiae* and *Escherichia coli*. The identified biclusters cover both known interactions and novel predictions. These biclusters can subsequently be used as input for the Inferelator (Bonneau *et al*, 2006), a framework that detects regulatory influences for genes and/or gene clusters from mRNA and/or protein expression levels via standard regression and model shrinkage techniques.

COGRIM integrates expression, ChIP-chip and binding site information (Chen *et al*, 2007). Via a Bayesian network, the authors try to find a matrix of which the genes are represented in the rows and the regulators in the columns. A one in this matrix indicates that a gene is regulated by a particular regulator, based on a combination of the ChIP-chip data, the motif data and the degree to which the expression profile of the regulator corresponds to the expression profile of the gene. From this matrix COGRIM derives which genes are regulated by the same regulator, i.e. regulons are detected. A disadvantage is that the authors cannot directly identify combinatorial regulation but they have to do a post-processing step to identify significant transcription factor interactions. In addition, the available implementation of the software cannot be applied on large data sets (see also section 4.3). COGRIM was applied on yeast ChIP-chip (Lee *et al*, 2002) and 500 expression arrays and also on human data.

4.2 Association rule mining algorithms for module network inference by data integration

A particular interesting, but yet unexplored class of algorithms for the inference of transcriptional modules by data integration is the class of association rule mining algorithms (ARM algorithms). In this PhD two module detection tools that rely on ARM algorithms, ReMoDiscovery (chapter 6) and DISTILLER (chapter 7), have been developed. ARM algorithms were developed in the database community, but are currently also being applied in the bioinformatics field. Although these algorithms offer many advantages, they have not been used frequently for the inference of networks or module networks. In this section, we discuss some of their advantages in comparison with the previously mentioned network inference methods (section 4.1). A more detailed description is given in chapter 5.

1. Exhaustive search: One of the main advantages of ARM algorithms is that they perform an exhaustive search for all feasible solutions and thus are not subject to problems associated with local optima. Simultaneously clustering both the rows and the columns of an expression matrix, for instance, is a combinatorial problem and many biclustering methods are therefore based on deterministic or stochastic greedy algorithms, often resulting in local optima. The method of Cheng and Church (2000) is based on a greedy algorithm that tries to find biclusters that have a coherence score below a user-defined threshold. Another biclustering algorithm (Ihmels *et al*, 2002; Ihmels *et al*, 2004) uses a deterministic iterative procedure that converges to a local stable point, depending on the initialisation. To obtain a better coverage of the total solution space, the method is repeated with many

randomly chosen initialisations. Sheng *et al* (2003) fit a probabilistic model using Gibbs sampling search strategies. Although Gibbs sampling garantees global convergence in the long run, in practice this method only converges locally due to the huge dimensions and difficult structure of the search space.

Many biclustering approaches need to mask the previously identified biclusters to be able to detect other biclusters (Cheng and Church, 2000; Sheng *et al*, 2003) or need multiple initialisations to obtain multiple biclusters (Ihmels *et al*, 2002; Ihmels *et al*, 2004). In contrast, ARM algorithms identify all possible solutions in one run.

2. Elegant data integration: Standard association rule mining algorithms allow including any data source in the analysis, as long the relevant properties derived from the data sets can be defined in a deterministic and hereditary manner. ARM algorithms can thus also be used for data integration. Carmona-Saez *et al* (2006) combined microarray data with additional information like functional annotation. Pham *et al* (2004) first search for sets of genes that are regulated by a common set of regulators, based on ChIP-chip data. Subsequently, the co-expressed genes in each set are listed. Clearly, the described approach uses different data sources in a sequential and hence suboptimal manner.

3. Concurrent data integration: Most data integration methods mentioned in section 4.1.2 combine the available data sources in a sequential manner. Bar-Joseph *et al* (2003), for instance, start with determining all genes that are regulated by the same set of regulators by using ChIP-chip data. Subsequently, expression data is used to check whether those genes are also co-expressed. In contrast, ARM algorithms are able to integrate the data in a concurrent way.

4. No co-expression assumption between regulator and target: Some network inference methods (Chen *et al*, 2007; Faith *et al*, 2007) assume that the expression profiles of the transcription factor and its target gene are similar. ARM algorithms do not have to rely on such assumptions.

5. Overlapping modules: ARM algorithms identify overlapping modules in a straight-forward way. Some module detection tools do not allow for overlap between modules, meaning that a gene can only belong to one module (Cheng and Church, 2000; Segal *et al*, 2003). Other module detection methods do allow for overlap but the overlap has to be modeled explicitly (Lazzeroni and Owen, 2002).

Despite the fact that ARM algorithms offer a substantial number of advantages, they have not been used very often for transcriptional network inference. This is possibly due to two problems that arise when relying on ARM algorithms.

The first problem is that the standard input data for ARM algorithms needs to be binary or discretized. Expression data, however, contain continuous values. Therefore, in most ARM applications that use expression data, this expression data first needs to be discretized. Okada *et al* (2007) for instance propose a biclustering method based on closed itemset minining. Although they can deal with multi-valued discretization levels discretization still causes a significant loss of information. Huang *et al* (2007) try to circumvent the discretization problem by calculating the expression correlation between each pair of genes for each of 65 microarray experiments. If a set of pairs of genes was co-expressed in a sufficient set of microarray experiments this set was used as input for a biclustering algorithm based on simulated annealing.

In addition, the output of most ARM algorithms is often a long list of possibly interesting results without rigorous statistical significance scores. This means that a filtering step is necessary to obtain a selection of the most interesting results. Several filtering steps are possible such as rule filtering or rule grouping (Tuzhilin and Adomavicius, 2002). Obviously, the better filtering procedure for the problem at hand is the procedure that guarantees the biologically most interesting results.

Because association rule mining algorithms have many advantages in comparison to other methods, we will rely on these algorithms for the detection of module networks by means of data integration. To be able to use them without loosing to much information, we examined possible solutions for the discretization problem that arises when using microarray data. Moreover, we came up with an efficient filtering strategy that allows us to select all biologically interesting modules. Finally, we developed a strategy to enable condition selection in an ARM framework for transcriptional module detection. In chapter 6 and 7 we will explain how these problems were tackled.

4.3 Which tool to choose as a biologist?

Many tools exist for the reconstruction of regulatory networks and module networks. Currently, biologists are overwhelmed by the increasing number of tools due to the lack of information and guidance on the relative performance of different tools in different situations. The reason is that many problems arise when trying to compare the tools and their results.

When a new network or module network inference method is submitted for publication, most journals ask for a comparison with other ‘similar’ tools (see also chapter 6 and 7 for a comparison of our module detection tools with related tools). Authors then often use the results of the different tools as described in the corresponding papers. This is possible since most network inference and module detection tools have been applied on yeast data.

However, the algorithms have usually not been applied on the exact same data sets, complicating the comparison. Preferably, comparative studies should be applied on a single biological benchmark data set. Only in this manner possibilities and limitations of the tools can be truly understood. Determining true positive results is possible by comparing the results with databases based on literature knowledge. However, no single biological network has been completely characterized. Therefore, determining the number of false positive results is impossible since no databases contain evidence for the absence of interactions between regulators and genes. The inclusion of negative evidence may become important to allow further progress and comparisons.

Due to these limitations of real experimental data, it may be interesting to benchmark the different tools on simulated or artificial data that mimic biological data, for instance generating by network generators that generate synthetic networks and simulated genome-wide data derived from those networks (for instance, (Van den Bulcke *et al*, 2006)). Of course artificial data can never cover all complexities of biological data, but it can at least be used to study some qualitative properties of the tested algorithm (Van Leemput *et al*, 2008). A consensus on sensible artificial benchmark data sets may become important in the near future.

Once a benchmark data set has been selected, the different tools can be run on this data set and the results can be compared. There are however more problems that need to be tackled:

- Very often network inference or module detection tools have been applied on one data set and parameters were tuned by the developers of the algorithm such that good results on this particular data set could be obtained. Applying the tool on other data, however, might require different parameter settings. Very often, however, users of the algorithm will adopt default parameters because there usually is insufficient information available on the meaning of model parameters or on parameter sensitivity. Although it is a hard and tedious task, more attention should be paid to this issue to get a clear picture on the influence of algorithmic settings on the results.
- For many network or module network inference methods the original algorithm implementation that was used in the article is not available. Instead, more user-friendly versions of the software are presented to the users. Although this may seem desirable at first sight, it can result in many problems when trying to run the software on actual data sets. Sometimes the online version of a tool differs considerably from the algorithm used in the corresponding article. The authors put for instance a version online that can only run on small data sets, although the original algorithm has been used on

real-life biological data sets. In other cases, the implementation of the user-friendly online version of the tool contains errors. Comparing network inference and module network inference tools can therefore be very hard or even impossible. During the research in chapters 6 and 7, we had to deal with this problem. In some cases, even the authors of the algorithm were unable to get it work on our data sets, so we were unable to include the corresponding algorithm in the comparative analysis.

Results of the different tools on the benchmark data set need to be compared with caution:

- Each of the tools differs in the way they approach the data and each tool has an underlying model that represents the biology. This model should be able to model sufficient complexity to capture different aspects of the biological processes in a comprehensive and intuitive way, but at the same time the model must be simple enough such that computational limitations do not pose a problem. Some tools include only expression data in their model, while other tools are capable of integrating multiple data sets. Including different information in the analysis allows to obtain more accurate results, such that it may be difficult to compare two tools that use different data sets.
- Comparing tools for network inference with tools for module network inference is non-trivial. Modules consist of sets of genes that are co-expressed in (a set of) conditions. A regulator can be assigned to these modules based on expression, ChIP-chip and/or regulatory motif data. When trying to compare a network with modules, single regulator-target gene interactions need to be extracted from the module network to be able to compare both. As such information about co-expression and condition dependency is lost.
- When comparing two module detection tools, the interactions extracted from the resultating modules can be compared, but also modules as a whole could be compared. Some module detection tools detect modules consisting of many genes and less experimental conditions, while others detect modules with a small number of genes coexpressed in many conditions. However, this does not necessarily imply incompatibility of the results. In some cases, it merely reflects the fact that different methods reveal different aspects of the data, such as a different level of modularity (resolution). D'haeseleer *et al* (2000) said about clustering algorithms "there is no one-size fits-all solution to clustering, or

> even a consensus of what a 'good' clustering should look like" (D'haeseleer *et al*, 2000) and the same can be said about module detection tools. Nevertheless, comparing the different tools remains crucial to get an idea on the type of interactions that can be found.

As mentioned above, comparing different tools is a very difficult task and in addition, most journals demand that the new method outperforms the previous tools, which sometimes results in unfair comparison strategies. Chen *et al* (2007), for example, compared the performance of their method COGRIM on a yeast benchmark data set with the performance of, among others, ReMoDiscovery (Lemmens *et al*, 2006) and GRAM (Bar-Joseph *et al*, 2003). Default parameter settings were used for ReMoDiscovery (Lemmens *et al*, 2006) and GRAM (Bar-Joseph *et al*, 2003). The within-regulon correlation as well as the within-regulon MIPS category enrichment (Mewes *et al*, 2008) were calculated. Within-regulon expression correlation might be a useful measure to judge the results of COGRIM (Chen *et al*, 2007) since they detect regulons, i.e. all genes regulated by the same regulator, based on ChIP-chip, regulatory motif and expression data. COGRIM starts by identifying single regulator-target gene interactions: the regulator has to bind to the promoter region of the gene and the expression profiles of the regulator and the target gene have to be similar. Afterwards all target genes for one regulator are combined into a regulon. Of course since COGRIM requires similarity of the expression profiles of regulator and target gene, the correlation coefficient for all genes in a regulon will be good. Although GRAM and ReMoDiscovery use the same data sets, they handle the different data sets completely different. Instead of searching for regulons, they detect modules, i.e. sets of co-expressed genes that are regulated by the same regulator. No similarity is required for the expression profiles of the regulator and the target genes. Therefore, even if the genes in two modules are regulated by the same regulator, the genes of one module do not have to be co-expressed with the genes of the other module. Since Chen *et al* (2007) wanted to compare regulons identified by different methods, they most probable took the union of all genes in modules that were regulated by the same regulator (although it is not mentioned how they exactly combine modules to regulons). Not surprisingly, the within-regulon correlation for module detection tools like ReMoDiscovery or GRAM is low, although the within-module correlation would be high. A comparison based on the within-regulon correlation is therefore unfair for methods like ReMoDiscovery or GRAM, that even get penalized for adopting less limiting assumptions. In conclusion, comparison of different algorithms and the results of different module detection tools is thus a non-trivial task and should be done in a fair way.

Chapter 5

Association rule mining in bioinformatics

In chapter 4 several network inference methods were described. Association rule mining (ARM) algorithms have not been used frequently for network reconstruction but do seem to offer many advantages, although some challenges remain to be solved. We therefore discuss ARM algorithms in more detail in this chapter. In chapter 6 and 7 we introduce our module detection tools based on ARM algorithms and show how some of the problems concerning the application of ARM algorithms to real-life biological data sets have been approached.

ARM algorithms were initially developed in the database community to analyse market basket data. Basket data consists of information on transactions or sets of items that have been purchased together. Transactions are stored in a database. Analysis of these past purchases helps the management of a store to decide on products to put on sale, the design of coupons, the way to place merchandise on the shells to maximize profit, etc. ARM algorithms are thus useful for mining large collections of data. Since our module detection tools, ReMoDiscovery (chapter 6) and DISTILLER (chapter 7) make use of this algorithm class a description of ARM algorithms is given below. We focus in particular on the Apriori and CHARM algorithm because ReMoDiscovery and DISTILLER, respectively, are based on these two algorithms. In the last section of this chapter we discuss applications of ARM algorithms in the bioinformatics domain.

Since our goal is to unravel the transcriptional network, it may be interesting to know, for instance, which genes have (multiple) motifs in common. Assume one has a database containing all genes together with the motifs that are present in the promoter regions of these genes. Given these data, ARM algorithms are able to find gene sets that share motifs in a very efficient way. In the usual terminology of ARM algorithms, a gene is called an *item*, while the motif corresponds to a *transaction*. A set of genes that share a number of motifs is an *itemset*. The number of common motifs is the *support* of that itemset. An itemset is called *frequent* if its support exceeds a prespecified

threshold: the support constraint. A frequent itemset is called *maximal* if it is not a subset of any other frequent itemset (in terms of items). A frequent itemset is called *closed* if it is not a subset of any other frequent itemset with the same transactions (see Figure 5- 1).

All possible itemsets can be represented in a lattice structure, i.e. all possible combinations of items in different itemsets of various sizes. In a naive way, all these combinations could be tested one by one to check whether they are frequent, maximal or closed. However for large databases this approach is computationally not feasible and we need to rely on efficient algorithms such as ARM algorithms. These algorithms start from a database of items and transactions and in a first step they search for frequent itemsets. In the second step they learn association rules from the frequent itemsets. An example of such an association rule is: if gene A and gene B are included in the itemset, there is a 90% probability to have gene C in the itemset as well.

Since the association rules themselves are not very important for our research, our focus will be on the efficient identification of frequent, maximal and closed itemsets. These itemsets or sets of genes that satisfy particular constraints or supports can be interpreted as transcriptional modules. We will thus make use of the association rules mining algorithms to find transcriptional modules.

Figure 5- 1: Example of a database of gene-motif combinations (A) and frequent (B), closed and maximal itemsets (C). Panel A shows a transaction base in which the genes are the items and the motifs are the transactions: motif 1, for instance, is present in the promoter regions of gene A, gene B and gene C. In the lower part of panel A, the support of the itemsets, or the number of motifs the genes have in common, is shown. Itemset (gene A, gene B, gene C), for instance, has a support of two, meaning these three genes have two motifs in common. Panel B shows the lattice of all possible itemsets (grey). The black nodes indicate the frequent itemsets if a minimum support of two is required. The red dashed line indicates that itemsets have the same support and will result in the same closed itemset, indicated with the red arrow. Panel C shows the closed and maximal itemsets.

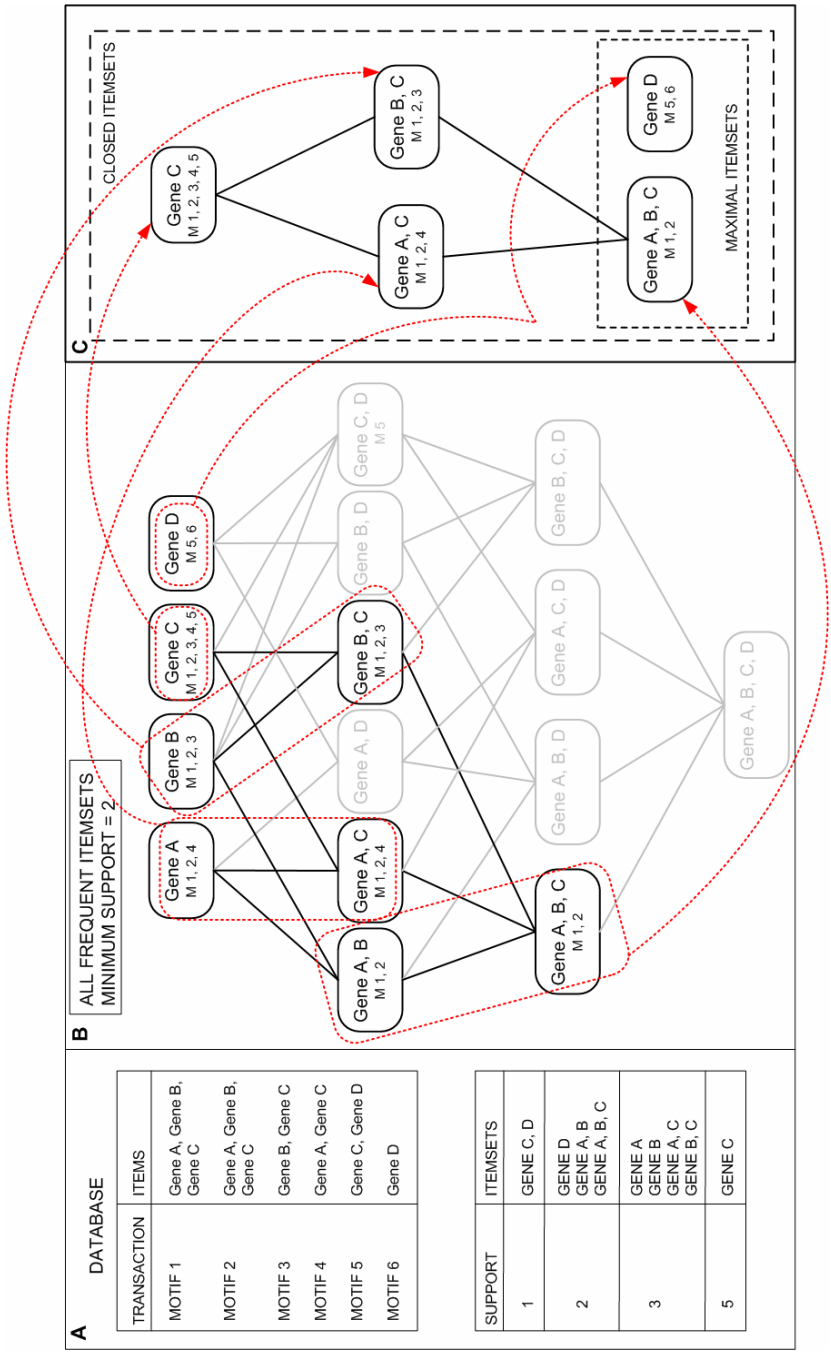


Figure 5- 1 caption on page 58

5.1 The Apriori algorithm

The Apriori algorithm (Agrawal and Imielenski, 1993) is the classical ARM algorithm to detect frequent itemsets. It scans in the lattice structure of all possible itemsets in a breadth-first way (Figure 5- 1). The algorithm builds up the frequent itemsets incrementally, starting from a list of frequent itemsets of size 1 and gradually expanding them. The list L^k, consisting of itemsets of size k, is composed based on the list L^{k-1} of frequent itemsets of size (k-1): each frequent itemset of size (k-1) will be expanded with each item that is not yet included in the itemset. Of course, by using this strategy the same itemset can be composed in multiple ways by adding the items in a different order. For instance, the two itemsets (gene A, gene B, gene C) and (gene A, gene C, gene B) are the same itemsets. Therefore the list of itemsets should be ordered and an itemset will only be expanded with an item of higher rank than the highest ranked item of the itemset. For instance, if the genes are ordered alphabetically then itemset (gene A, gene C, gene B) will not be formed since gene B has a lower than gene C. Using this non-redundant incremental strategy, candidate itemsets are created only once and the algorithm then needs to check whether these are frequent.

To accomplish this task efficiently, Apriori relies on the observation that whenever an itemset is frequent, all of its subsets must necessarily be frequent as well. This so-called hereditary property allows us to conclude *a priori* whether a particular candidate itemset is possibly frequent and can be exploited to reduce the number of times the constraints need to be evaluated. Indeed, only itemsets for which all subsets have been found to be frequent itemsets need to be checked. In other words, sets can be discarded *a priori* if one of their subsets turns out to be not frequent, even before checking whether the itemset under study is frequent: we simply have to check if all of its subsets of size (k-1) exist in the list L^{k-1}. Applying this approach thus helps to reduce the number of itemsets for which checks have to be performed.

As an example, consider the database of Figure 5- 1. The items, i.e. gene A, gene B, gene C and gene D, are ordered. If a minimum support of two motifs is required, all itemsets of size one are frequent. In a second step, the algorithm combines every itemset of size one with an item not yet in the itemset and according to the ordering of the items. So gene A will be combined with gene B, gene C and gene D. Gene B will be combined with gene C and gene D, but not with gene A since gene A has a lower rank, meaning that the combination gene A–gene B has already been formed. For each of these itemsets the algorithm has to check in the database whether they are frequent. There are only three combinations that are frequent: (gene A, gene B), (gene A, gene C) and (gene B, gene C). Gene C and gene D only share one motif and therefore this itemset is not frequent. In the next

iteration, itemset (gene A, gene B and gene C) has to be checked and it happens to be frequent. Itemset (gene A, gene C), for instance, is a frequent itemset that could be extended with gene D to form (gene A, gene C, gene D). However, since itemset (gene A, gene D) and (gene C, gene D) are not frequent, we already know *a priori* that itemset (gene A, gene C, gene D) cannot be frequent. For similar reasons, other size three itemsets cannot be frequent and do not have to be checked.

Many frequent pattern mining algorithms are variations and extensions of the well-known Apriori algorithm (Agrawal and Imielenski, 1993). The Apriori algorithm searches for all frequent itemsets by using a bottom-up breadth first strategy. In each iteration, the algorithm has to check the database to find the frequent itemsets among possible candidate itemsets. The Apriori algorithm performs well on sparse databases but if the itemsets grow larger, the performance of Apriori-based algorithms decreases (Zaki and Hsiao, 2002). A first solution to this problem is to search for maximal itemsets instead of all frequent itemsets, since the number of maximal itemsets is much smaller than the number of frequent itemsets. However, maximal itemsets do not contain all information i.e. they do not uniquely determine all frequent itemsets and their exact support. In contrast, closed itemsets are lossless in the sense that knowing all closed itemsets allows you to infer the support of any given itemset. The number of closed itemsets is also much smaller than the number of frequent itemsets. The difference between frequent itemsets, maximal itemsets and closed itemsets is indicated in Figure 5- 1.

5.2 Charm algorithm

The CHARM algorithm (Closed Association Rule Mining) is an efficient algorithm for identifying closed itemsets (Zaki and Hsiao, 2002). In contrast to the Apriori algorithm (section 5.1), CHARM explores the itemset space and transaction space simultaneously over an IT or itemset-transaction tree search space. In this tree, a node consists of an (itemset x transaction) pair (see Figure 5- 2). CHARM searches this tree using a depth-first search strategy (Apriori, in contrast, uses a breadth-first strategy) exploiting the notion of equivalence classes. In the IT-tree, each node is in fact a prefix-based class. Two itemsets belong to the same class if they share a common *k*-length prefix, determined by an ordered list of *k* gene names. By construction, the children of a node all belong to the same equivalence class *X* since they all share the same prefix *X*. In Figure 5- 2, gene B and gene C, for instance, belong to the same equivalence class [A]. Note that gene D does not belong to this class since itemset (gene A, gene D) is not frequent. So a class represents items with which the prefix can be extended to obtain a new frequent node. No subtree of an infrequent prefix has to be examined.

The frequent itemsets can readily be determined in the IT-tree framework: for a given node or prefix class the intersections of the transactions of all pairs of elements is determined and it is checked whether they meet the minimum support. A pass over the database to check the support of an itemset is not necessary anymore. Each resulting frequent itemset is a class on its own that can be expanded recursively. The power of this class approach is that it breaks the original search space into independent sub-problems. Any subtree rooted at node X can be treated as a new problem and only this subproblem has to fit in the memory.

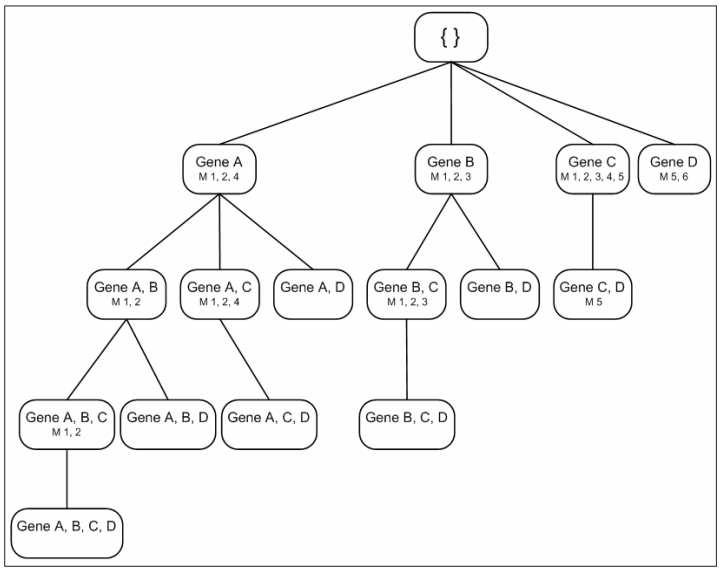


Figure 5- 2: The itemset-transaction (IT) tree used by CHARM to search both the itemset and transaction space: each node in the gaph indicates both the itemset and the support of the itemset.

CHARM makes use of four properties of the equivalence classes to skip levels in the IT-tree structure. Assuming two members, X_i and X_j, of the same equivalence class that are ordered such that $X_i < X_j$, the following properties apply:

- Rule 1: if X_i and X_j have the same support, then X_i can be replaced by $X_i \cup X_j$ and element X_j does not need to be considered anymore.
- Rule 2: if the support of X_i is a subset of the support of X_j, then X_i can be replaced by $X_i \cup X_j$ since they have identical closures (i.e. they result in the same closed itemset), but element X_j cannot be removed.

- Rule 3: if the support of X_j is a subset of the support of X_i, then X_j can be replaced by X_i U X_j since they have identical closures, but element X_i cannot be removed.
- Rule 4: if the support of X_i and X_j is different, then no element of the class can be eliminated since both X_i and X_j will lead to different closures.

By making use of these rules, parts of the IT-tree can be passed over and the IT-tree can be searched in very efficient way. Because rule 1 and 2 favor this skipping of levels, CHARM orders the items in increasing order of support such that there will be more occurrences of rule 1 and 2.

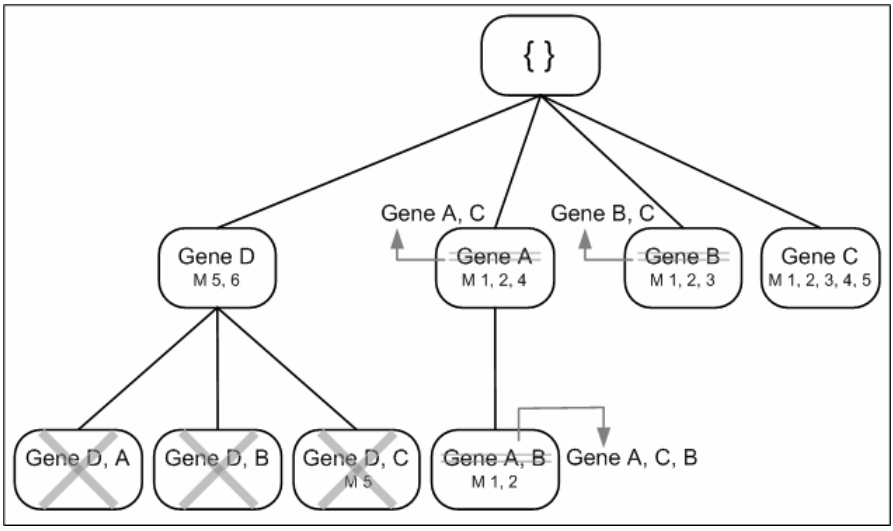


Figure 5- 3: The search strategy adopted by CHARM (Zaki and Hsiao, 2002). CHARM starts by ordering the items (genes) by increasing support. In the next step, the item with the lowest support, gene D, will combined with the other items, but none of the combinations is frequent. Subsequently, gene A will be extended with gene B. Since the support of gene A and gene B is different, no elements can be discarded. The resulting frequent itemset is (gene A, B) with a support of two. Then this itemset (gene A, B) will be extended with gene C. Since the support of gene A is part of the support of gene C, CHARM will replace gene A in the tree by (gene A, C), resulting in the itemsets (gene A, C) and (gene A, C, B). When extending gene B with gene C, the same is true and itemset (gene B) will be replaced by itemset (gene B, C). This search strategy drastically reduces the number of combinations that need to be tested.

Figure 5- 3 shows how CHARM works on our previous example. First the genes are ordered according to their support. Gene D will be combined with the remaining elements in the class: gene A, gene B and gene C. The support of gene D is different from the support of gene A, therefore rule 4 applies and all combinations need to be considered. Since the support of the combinations (gene D, A), (gene D, B) and (gene D, C) does not meet the support constraint of two motifs, all these itemsets were pruned. None of their extensions needs to be checked anymore. Therefore, the first identified closed itemset is gene D. In the next step gene A will be extended with gene B. Again rule 4 applies and no elements can be discarded. The itemset (gene A, B) is a frequent set. This set can be extended with gene C. Since the support of gene A is part of the support of gene C, rule 2 applies and we can simply replace gene A by (gene A, C). By doing this, we already completely checked the subtree starting at node A, i.e. we do not have to check whether the branch (gene A, C) is frequent anymore because we already know it is. That part of the tree can thus be skipped. Again we found two closed itemsets: (gene A,C) and (gene A, C, B). When we extend gene B with gene C, again rule 2 applies and we can substitute gene B with (gene B, C) and identified a fourth closed itemset. Gene C is the gene with highest order and therefore can not be extended with any gene and determines the last closed itemset. Using this strategy, the IT-tree can be searched in a very efficient way.

5.3 Application of association rule mining algorithms in bioinformatics

Recently, ARM algorithms found their way to the field of bioinformatics. Several applications make use of the Apriori algorithm (Chiu *et al*, 2006; Ivan *et al*, 2007; Morgan *et al*, 2007; Oyama *et al*, 2002) or Apriori-based algorithms like (Artamonova *et al*, 2005; Artamonova *et al*, 2007; Becquet *et al*, 2002; Brazma *et al*, 1997; Carmona-Saez *et al*, 2006; Creighton and Hanash, 2003; Rodriguez *et al*, 2005). Some of those applications search for closed itemsets (Huang *et al*, 2007; Okada *et al*, 2007; Pham *et al*, 2004). Although most of these bioinformatics applications rely on existing itemset and rule mining algorithms, sometimes new algorithms are being developed to take into account specific properties of the biological problems at hand (Georgii *et al*, 2005; Lopez *et al*, 2008; Tamura and D'haeseleer, 2008). In almost all cases, the approach consists of four steps. In a first step, data is gathered and transformed into a matrix format. Usually, the entries in the matrix are binary. In a second step, the matrix is processed to obtain frequent itemsets. In a third step association rules can be derived from the frequent itemsets. Some of the approaches skip the third step. In the latter case, the frequent itemsets themselves form the result. Because ARM methods tend to

generate large amounts of itemsets or association rules, a final filtering or post-processing step is usually introduced to obtain biologically interesting itemsets or rules.

Diverse applications of ARM algorithms in the field of bioinformatics have been reported (see Table 5- 1). Lin *et al* (2006), for example, make use of the database of the HIV Drug Resistance database and ARM algorithms to find relationships between mutations in the HIV protease gene and antiretroviral drug treatment. In another example, ARM algorithms were used to obtain sets of COGs (Clusters of Orthologous Groups of Proteins) associated with a phenotype (Tamura and D'haeseleer, 2008).

ARM algorithms have also been used to study protein interactions or to annotate proteins. Oyama *et al* (2002) reveal rules related to protein-protein interactions, while Ivan *et al* (2007) study ligand-protein interactions. Both studies collect information on different properties of the proteins like functional category, protein domain information or residue composition. Subsequently, rules are derived that provide information on which of these properties occur together very frequently in interactions. These rules could provide novel insight in the characteristics of the interactions. ARM algorithms have also been used for the annotation of proteins based on protein domain composition (Chiu *et al*, 2006) or protein sequence similarities (Rodriguez *et al*, 2005).

One of the earliest bioinformatics applications of ARM approaches concerned the search for combinations of transcription factor binding sites in the upstream regions of yeast genes that occur more frequently than expected by chance (Brazma *et al*, 1997). Brazma *et al* (1997) screened the promoter regions of yeast for the presence of motif instances. Subsequently they searched for the frequent sets of regulatory motifs and used these frequent itemsets to derive association rules, such as " if motif 1 and motif 2 are present, then motif 3 is also present". The order of occurrence of the motifs could not be taken into account by the approach of Brazma *et al* (1997) Despite this shortcoming, this kind of research is still very useful to study combinatorial regulation. More recently, ARM was used to search for frequent combinations of regulatory motifs that are located close to each other in the DNA sequence of the human genome (Morgan *et al*, 2007). Another approach, developed by Doi *et al* (2008), uses as input a set of user-defined genes and searches for significant combinations of regulatory motifs in their upstream regions. The search for motifs is performed via both *de novo* motif detection and screening with known motif matrices.

Table 5- 1: Overview of ARM algorithm applications in the bioinformatics field. The ARM algorithm is indicated, but also whether the application uses the derived association rules or only the frequent itemsets. The table also mentions which data sources were used in the particular study, and in the case microarrays were used, it also indicates whether the microarray data had to be discretized and if condition selection was possible.

Study	Research domain	Organism	ARM algorithm	Itemsets or association rules	Data source	Discretization of expression data	Condition selection
Brazma *et al*, 1997	regulatory motifs	yeast	Toivonen *et al*, 1996 - Apriori based	association rules	motif screening results	/	/
Morgan *et al*, 2007	regulatory motifs	human	Agrawal *et al*, 1993, 1994 - Apriori	association rules	motif screening results	/	/
Doi *et al*, 2008	regulatory motifs	rice	not mentioned	association rules	motif screening and *de novo* motif detection results	/	/
Oyama *et al*, 2002	protein-protein interactions	yeast	Agrawal *et al*, 1993, 1994 - Apriori	association rules	structural and functional information on proteins from Swiss-prot	/	/
Iván *et al*, 2007	protein-ligand interactions	several species?	Agrawal *et al*, 1993, 1994 - Apriori	association rules	ligand-binding residue sets	/	/
Chiu *et al*, 2006	protein annotation	several species	Agrawal *et al*, 1993, 1994 - Apriori	association rules	protein domain and protein functional information	/	/
Rodriguez *et al*, 2005	protein annotation	five prokaryote species	Rodriguez *et al*, 2005 - Apriori based	association rules	protein sequence similarity and protein functional information	/	/
Artamonova *et al*, 2005	protein annotation	several species	Webb *et al*, 2000 - implementation of Apriori	association rules	structural and functional information on proteins from Swiss-prot	/	/

Artamonova *et al*, 2007	protein annotation	several species	Borgelt *et al*, 2002 - Apriori based	association rules	structural and functional information on proteins from Pedant	/	/
Lin *et al*, 2006	HIV mutants	human	Agrawal *et al*, 1993, 1994 - Apriori	association rules	information on HIV mutations and drug resistance	/	/
Tamura *et al*, 2008	genotype-phenotype mapping	prokaryotic species	Own method: class association rule mining	association rules	phenotypes and clusters of orthologous groups of proteins	/	/
Becquet *et al*, 2002	analysis of gene expression	human	Boulicaut *et al*, 2003 - Apriori based	association rules	SAGE data	yes	yes
Creighton *et al*, 2003	analysis of gene expression	yeast	Own method: Apriori based	association rules	microarray data	yes	yes
Carmona-Saez *et al*, 2006	analysis of gene expression	yeast and human	Rodriguez *et al*., 2005 - Apriori based	association rules	microarray data and gene annotations	yes	yes
Lopez *et al*, 2008	analysis of gene expression	yeast	Own method: fuzzy version of Wang *et al*, 2002	association rules	structural and functional information and bicluster membership data	/	/
Okada *et al*, 2007	analysis of gene expression	yeast	Uno *et al*, 2002 - Closed itemset mining	closed itemsets	microarray data	yes	yes
Georgii *et al*, 2005	analysis of gene expression	yeast	Ruckert *et al*, 2004 - Quantitative association rules	association rules	microarray data	no	yes
Huang *et al*, 2007	analysis of gene expression	human	Grahne and Zhu 2003 - Closed itemset mining	closed itemsets	microarray data	no	condition group
Pham *et al*, 2004	analysis of gene expression	yeast	Grahne and Zhu 2003 - Closed itemset mining	closed itemsets	ChIP-chip data *	/	/

In spite of the previously mentioned applications, ARM methods are not used very frequently for network inference although they offer many advantages:

- ARM algorithms search for all solutions in an exhaustive manner so in contrast to optimization-based methods, ARM algorithms do not suffer from problems with local optima.
- Because ARM methods find all solutions at once, they do not need to mask previous results to find the next result or to use multiple initializations to find all possible results.
- In contrast to other methods, overlapping solutions are possible and generated in a natural and straight-forward way. This is obviously an interesting property, for instance while searching for biclusters. Since a gene can be involved in more than one biological function, a gene can also belong to more than one bicluster.
- In principle, each data set from which hereditary properties can be derived, can be included in the analysis in a straight-forward way. These hereditary constraints allow the ARM algorithms to search efficiently by employing pruning steps. ARM methods are therefore extremely useful for data integration since many data sources can be included.

The reason why these methods have not been used more often can be traced back to two major challenges associated with the use of ARM algorithms for network inference.

A first major challenge when applying ARM algorithms is the large amount of generated frequent itemsets or rules. The use of maximal or closed itemsets (instead of all frequent itemsets) can partly solve this problem but still too many itemsets and rules remain. Therefore, the resulting itemsets and rules need to be filtered and analyzed in a computationally efficient and biologically meaningful way to remove redundancy (for instance, in the form of heavily overlapping sets) and separate interesting results from less interesting results. Tuzhilin and Adomavicius (2002) investigated strategies to tackle this problem. They suggest rule filtering in a context where the user is only interested in the rules concerning for instance a particular group of genes. Rule grouping can be applied when many similar rules are generated. In this case, it would be useful for the biologist to analyze similar rules together to obtain a high-level overview of the inferred rule classes.

The definition of biological relevance may differ from application to application. The most appropriate filtering approach can therefore be very different in each ARM application. For instance, Artamonova *et al* (2005, 2007) for instance employed ARM techniques for the discovery of

incorrectly annotated proteins in protein databases. They calculated the confidence of all derived association rules. The confidence of a rule X => Y was defined as the ratio of the support of all itemsets containing X and Y to the support of all itemsets containing X. Artamonova *et al* (2005, 2007) were especially interested in those association rules with a confidence close to one, since the exceptions on this rule might indicate that the corresponding proteins are wrongly annotated. Oyama *et al* (2002) investigated properties of interacting proteins. Many association rules were derived, including rules of the form "if the protein has domain A then it also contains domain B". These kinds of rules only say something about one protein, not about an interaction. Since Oyama *et al* (2002) were only interested in those rules that describe a relationship between two interacting proteins, they had to apply a filtering procedure to obtain only those rules that contain information on two interacting proteins. In summary, it is a non-trivial challenge to define the biologically relevant information and a corresponding measure of interestingness for each specific application. In chapter 7, we suggest the use of a module selection filtering step that is based on statistical significance and overlap of the inferred modules.

A second challenge is the requirement for the discretization of the input data that is often imposed. In bioinformatics, and especially in the field of network inference, expression data are used very frequently. These data, however, naturally contain continuous values. Therefore, in most ARM applications that use expression data, discretization becomes necessary. The continuous expression values are, for instance, transformed to "underexpressed", "overexpressed" or "not differentially expressed". For example, association rules have been utilized for the analysis of discretized SAGE data (Becquet *et al*, 2002) and discretized microarray data (Creighton and Hanash, 2003). Rules of the form "if gene 1 and gene 2 are overexpressed, then gene 3 will be overexpressed" can then be derived from these data. Carmona-Saez *et al* (2006) combined discretized microarray data with additional information like functional annotation. They identify rules like "if a set of genes are cell cycle genes, then they will be upregulated in condition 1, 4, 5 and downregulated in condition 8". Okada *et al* (2007) propose a biclustering method based on closed itemset mining and discretized microarray data. They did not explicitly derive association rules but employed closed itemset mining as an efficient strategy to find the biclusters.

Discretization of the data leads to a significant loss of information. Moreover, the choice of discretization thresholds may introduce artefacts. Georgii *et al* (2005) use quantitative association rules to avoid the need for discretization. Lopez *et al* (2008) make use of bicluster membership in addition to other different data sources such as GO terms, length of the gene, GC content of the gene. By using bicluster membership they avoided the

need to discretize the microarray data. Huang *et al* (2007) found another solution to deal with the continuous values in expression data. For each of 65 microarray experiments (consisting of minimum 8 microarrays), they calculated the expression correlation between each pair of genes, although this a computationally costly task. Subsequently this information was used to check whether sets of pairs of genes were co-expressed in a sufficient set of microarray experiments by using frequent itemset mining. Next these frequent itemsets were used as seeds in a biclustering algorithm based on simulated annealing. This additional step serves to allow small amounts of noise in the resulting bicluster patterns. Although Huang *et al* (2007) can select microarray experiments in which genes are co-expressed, they cannot select single conditions or microarrays. Yet another way of dealing with continuous expression data directly has been proposed by Pham *et al* (2004). They combined ChIP-chip data with expression data to obtain sets of co-expressed genes that are regulated by the same set of transcription factors. In a first phase, all sets of genes that are regulated by a common set of regulators are detected by means of closed itemset mining. Subsequently, for each closed itemset the co-expressed genes are listed. The approach of Pham *et al* (2004) uses different data sources in a sequential and hence suboptimal manner. In addition, their method does not allow for condition selection or biclustering.

In this thesis we deal with the above challenges and employ ARM algorithms for the inference of module networks by data integration. Modules consisting of sets of co-expressed genes that are regulated by the same set of transcription regulators and that have the same regulatory motifs in their upstream regions are identified in a very efficient way (see chapter 6 and 7). To obtain these sets of genes, our approaches, ReMoDiscovery and DISTILLER, make use of expression data, ChIP-chip data and motif data. In contrast to sequential data integration strategies, such as the method of Pham *et al* (2004), we combine all data sources in a concurrent way. In addition, DISTILLER is able to bicluster microarray expression data without discretizing the microarray data (Becquet *et al*, 2002; Carmona-Saez *et al*, 2006; Creighton and Hanash, 2003; Okada *et al*, 2007) or grouping the arrays into microarray experiments (Huang *et al*, 2007). In our methods, ARM algorithms are employed to obtain frequent itemsets. We introduce a statistical approach for the filtering of the obtained frequent itemsets. Since the goal is the reconstruction of the module regulatory network, no association rules are derived.

Chapter 6

Unraveling the yeast transcriptional network

The goal of this thesis is to obtain a detailed and holistic view on the transcriptional network. Although several network inference methods already exist (as described in chapter 4), we will explore the ability of association rule mining algorithms for this task since they seem to offer many advantages (chapter 5). In this chapter, we discuss our module detection tool ReMoDiscovery, developed in collaboration with Prof. De Bie (University of Bristol, UK). Based on expression, regulatory motif and ChIP-chip data (see chapter 2), sets of co-expressed genes are identified that have the same regulatory program (section 6.1). The regulatory program consists of transcription factors that are in control of those genes and the corresponding regulatory motifs. To make the discovery of such interesting gene sets feasible, ReMoDiscovery relies on the Apriori algorithm that belongs to the class of association rule mining algorithms (chapter 5).

Most module detection tools make use of probabilistic or graph-based strategies, while association rule mining algorithms are hardly ever used (chapter 4). In a comparative study, ReMoDiscovery outperforms other published module discovery tools (section 6.3). The GRAM algorithm (Bar-Joseph *et al*, 2003) is slower than ReMoDiscovery and running times of GRAM were prohibitive if the data contained genes with more than ten significant ChIP-chip interactions. In comparison to SAMBA (Tanay *et al*, 2004), ReMoDiscovery is equally fast. SAMBA, however, seems not to be able to detect biologically interesting regulatory programs. Next we discuss the data integration functionality of the ReMoDiscovery platform on yeast data (chapter 3), where it was shown to generate biological meaningful findings and allowed predicting potential novel roles of transcriptional regulators (section 6.2).

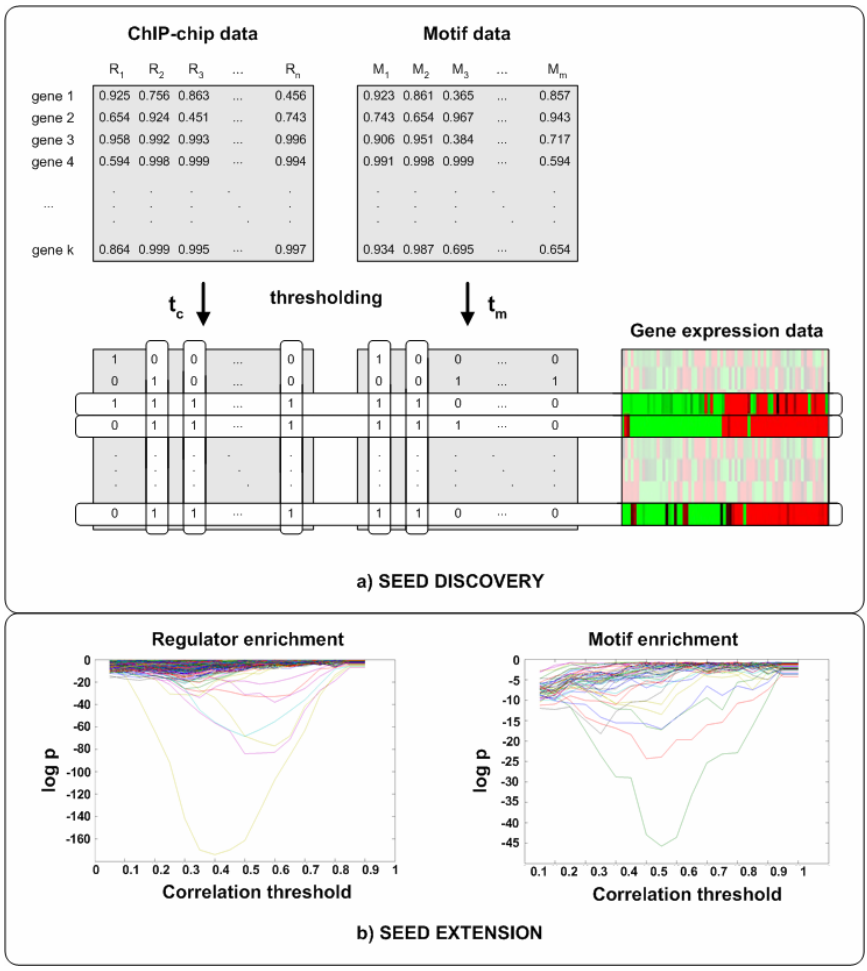


Figure 6- 1: ReMoDiscovery analysis flow. ReMoDiscovery consists of a seed discovery step followed by a seed extension step. ChIP-chip data, motif data, and expression data are used as input for the algorithm. These three datasets can be represented as matrices in which the rows represent the genes. For the ChIP-chip data (matrix R) the columns represent the regulators, for the motif data (matrix M) they represent the motifs and for the expression data (matrix A) the different experiments. **a)** The seed discovery step identifies sets of genes that are co-expressed, bind the same regulators, and have the same motifs in their intergenic region. **b)** The gene content of the seed modules can be extended during the seed extension step using less stringent criteria. The logarithm of the module enrichment p-values (Y-axis) are plotted for all regulators (motifs) as a function of the correlation threshold (X-axis). Each line in the sample plot shows the module enrichment p-values for the enrichment of its corresponding regulator (motif) as a function of the gene expression correlation threshold used.

By outperforming related methods and by generating good biological results, we proved that ARM based algorithms are well suited for the inference of the transcriptional network through data integration. This study was published in Genome Biology (2006):

Lemmens K., Dhollander T., De Bie T., Monsieurs P., Engelen K., Smets B., Winderickx J., De Moor B., Marchal K. (2006) Inferring transcriptional module networks from ChIP-chip-, motif- and microarray data. Genome Biology, vol. 7, no. 5, pp. R37.

6.1 ReMoDiscovery

ReMoDiscovery (Lemmens *et al*, 2006) aims at identifying transcriptional modules by examining microarray data for target genes with a common expression profile that also share the same regulatory program, based on evidence from ChIP-chip and motif data. Our approach is different in spirit from previously suggested methods in that this algorithm takes distinct data sources related to transcriptional regulation, that is, microarray, ChIP-chip and motif data, into account in a concurrent (non-iterative or sequential) way. In contrast to previous methods, where motifs are usually defined in a downstream analysis step, we exploited motif data as an independent information source.

The ReMoDiscovery workflow consists of two steps. In a first seed discovery step, seed modules are identified (see Figure 6- 1). This seed discovery problem translates into finding gene sets (row dimension in Figure 6- 1) that are co-expressed in microarray data (matrix M), that bind the same regulators (share the same columns in the ChIP-chip matrix) and that have the same motifs in their intergenic region (same columns in the motif matrix (Figure 6- 1)). In a second seed extension step, the gene content of the module is extended using less stringent criteria. In the following, we discuss the specifics of this two-step procedure.

6.1.1 Seed discovery step

In the seed discovery step, we detect large modules with tightly co-expressed genes, directed by a common regulatory program with a minimum number of regulators and a minimum number of conserved motifs in the upstream region of the genes included in the module. Modules that meet these user-defined stringent criteria are defined as valid seed modules. We solely report 'maximal modules', defined as valid seed modules that become invalid upon extending them with any gene they do not yet contain.

An exhaustive search for all valid gene sets is not feasible, as the number of possible sets drastically increases with the number of genes in the sets (in a

data sets with 6000 genes there are for instance about 10E+10 possible sets of three genes and 10E+13 possible sets of four genes). However, by defining the constraints in such a way that extensions of an invalid module are never valid (that is, as hereditary constraints), we can adopt a fast Apriori-like algorithm to solve the problem. In ReMoDiscovery, the constraints specify the minimum number of regulators (or regulator support constraint s_c), the minimum number of motifs (or motif support constraint s_m), and a minimal pairwise correlation between genes in a module t_e. We apply these constraints to find regulatory modules that contain as many genes as possible, i.e. we search for the maximal itemsets.

Since the regulator binding and motif data consist of non-binary score values, the support values are estimated by using thresholded regulator and motif scores, equal to 1 if the score is larger than a threshold (t_c or t_m, respectively) and 0 otherwise. After thresholding, the regulator and motif data are binary data sets that can be represented as two matrices R and M (Figure 6- 1). The current implementation uses correlation as a measure for co-expression, but other similarity measures could be applied in the Apriori framework, as long as the co-expression constraint remains hereditary.

In summary, a high level description of the steps in the algorithm can be given as follows:

1) choose parameter values s_c, s_m, t_e, t_m and t_c;

2) threshold the regulator and motif data using thresholds t_m and t_c, yielding the binary matrices R and M;

3) find all maximal modules for which the support constraints specified by s_c and s_m are satisfied, and for which the correlation between the gene expression profiles of any pair of genes in the module exceeds the required threshold t_e;

4) report maximal modules along with the motifs and regulators that support them.

To determine the statistical significance of the obtained modules, we assigned a 'seed module' p value to each seed module. To this end, we randomly permuted the gene labels for each dataset (ChIP-chip, motif data, gene expression) separately. This randomization procedure was repeated 100 times. The results of ReMoDiscovery seed discovery on these random datasets were used to construct an empirical joint distribution on the number of regulators and genes from which we calculated a *seed module p value* for each of the seeds found in the real data sets. As expected, seeds with a high number of genes were highly significant (see Figure 6- 2). Modules with one gene were only significant if they contained many regulators.

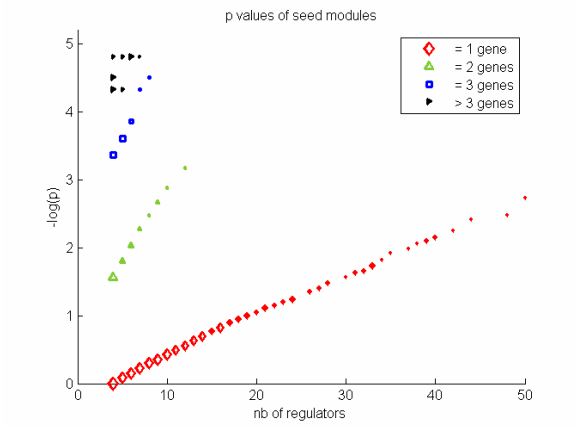


Figure 6- 2: Significance scores (i.e. minus the logarithm of the seed module p-values) for seed modules found on the Gasch dataset (Gasch *et al*, 2000) with a ChIP-chip support of four regulators. These significance scores reflect the probability score in a joint distribution of the number of regulators and genes for seed modules obtained on randomized data. The size of the markers is proportional to the (logarithm of the) number of modules of a specific size, mined on the original non-randomized data. All modules with more than one gene are highly significant.

6.1.2 Parameter sensitivity

To test the sensitivity of the seed discovery step with respect to the parameters, we compared results obtained at different parameter settings using a normalized Jaccard similarity score.

The Jaccard similarity measure is defined as:

$$J = \frac{TP}{TP + FN + FP} \tag{1}$$

In this equation, *TP* stands for True Positives, *FP* for False Positives, and *FN* for False Negatives, all defined with respect to two binary matrices that represent pairwise occurrences for two module partitions. When focussing on genes for example, a True Positive would indicate a pair of genes (one entry in the gene x gene co-occurrence matrix) that occurs together in at least one module, in both partitions.

The original implementation was based on the empirical distribution of Jaccard values, obtained by repeatedly randomizing the module composition. This distribution can then be used to attach a significance score to the Jaccard similarity between results obtained with two parameter settings, either from the same algorithm, or across algorithms.

An obvious disadvantage of this strategy is the need for many randomizations in order to get reasonable estimates for the mean and standard deviation on the distribution of the Jaccard coefficients. In the case of sparse matrices, it is unlikely that any *TP* occur upon randomization, making (almost) all values equal to zero. This makes it practically impossible to work with empirical p-values for such sparse matrices. Therefore, an analytical derivation for the distribution of Jaccard similarities between randomized compositions is useful to reduce computational cost. It also allows to produce smoother sensitivity analysis plots.

Define n as the total number of entries in the co-occurrence matrix For instance, n could be the number of genes squared in a gene x gene matrix. Assume that p_1 and p_2 represent the densities of ones in the binary co-occurrence matrices for partition one and two respectively. Obviously, *TP*, *FP* and *FN* are related, because

$$TP + FP = n \cdot p_1 \tag{2}$$

$$TP + FN = n \cdot p_2 \ . \tag{3}$$

Using this to rewrite (1) yields

$$y = \frac{TP}{TP + FN + FP} = \frac{TP}{\left(TP + FN\right) + \left(TP + FP\right) - TP} = \frac{TP}{n \cdot \left(p_1 + p_2\right) - TP} \tag{4}$$

The latter equation is of the form x/(a-x) with x a random variable indicating the probability of observing a specific number of *TP*.

Let us assume we are looking at two matrices with given densities that are uncorrelated. In this case, the number of TP is a random binomial variable. The probability of observing exactly k TP is given by $Bin(k, n, p_1 p_2)$, the probability of having k successes in n trials with success rate $(p_1 p_2)$. If the number of TP x is much smaller than a, $ay \approx x$ is binomial too:

$$p_{ay}\left(s\right) = Bin\left(s, n, p_1 p_2\right) \tag{6}$$

In the plots of parameter sensitivity and module comparison, we therefore expressed the significance of the Jaccard coefficient as the number of standard deviations from the mean of this distribution.

The overall similarity in gene and regulator content of the resulting modules from different parameter settings was examined separately. We varied the correlation threshold on the expression profiles, the threshold on the ChIP-chip data t_c and the minimum number of regulators s_c (see Figure 6- 3). Parameter settings that are more similar generally resulted in more similar gene and regulator module content. This consistency eases parameter tuning.

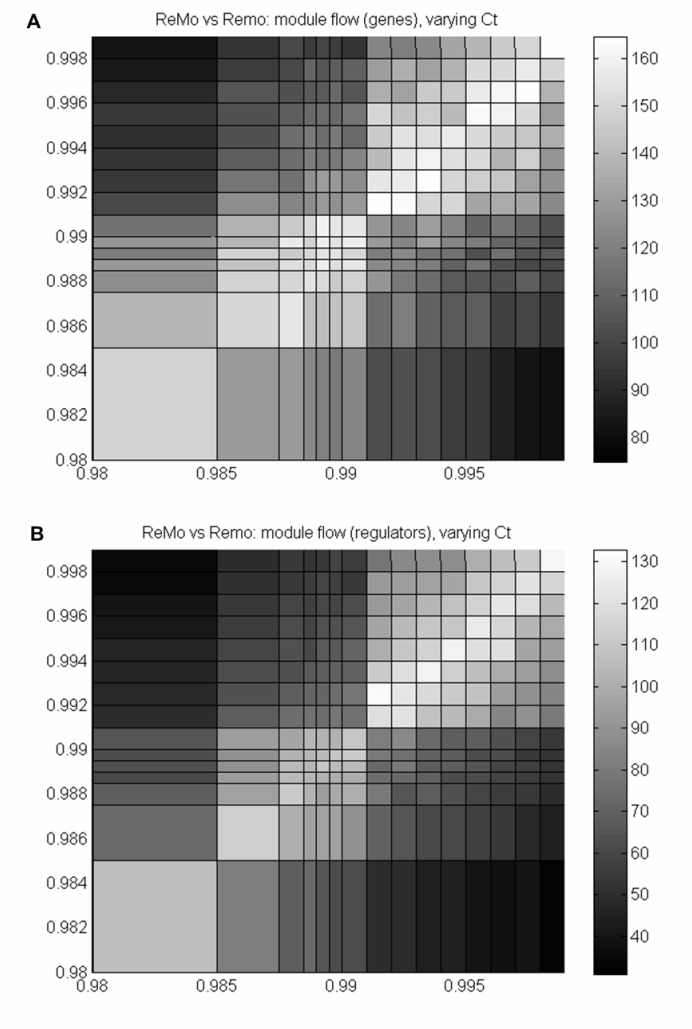


Figure 6- 3: The significance of the similarity in module content between ReMoDiscovery seed modules at different parameter settings. Significance was calculated using a normalized Jaccard similarity score: the color bar on the right indicates the number of standard deviations from the mean of the distribution. For the ChIP-chip threshold c_t the tested settings were not selected equidistantly: close to the default value (0.99) more ChIP-chip threshold values have been examined. These plots indicate the parameter sensitivity of the ReMoDiscovery algorithm. A. Influence of variations in ChIP-chip threshold on gene content; B. Influence of variations in ChIP-chip threshold on regulator content. From these plots it is clear that parameter settings that resemble each other result in modules with similar gene and regulator content. Modules obtained with regulator threshold t_c of 0.999, for instance, resemble modules obtained with a regulator threshold of 0.995 while they resemble modules obtained with regulator threshold 0.985 less. This consistency makes it easy to tune the parameter settings of ReMoDiscovery.

6.1.3 Seed extension step

The stringent criteria for the valid modules in the seed discovery step appear to be sufficient for a reliable detection of regulators and motifs. The reported maximal gene content of such modules, however, is likely to be underestimated in size because of these stringent criteria. For this reason, ReMoDiscovery uses a second step, in which the gene content of statistically significant seed modules is extended.

This extension is performed by computing the module's mean expression profile, and ranking the remainder of the genes in the dataset according to their correlation with this seed profile. The genes at the top of the ranking are the best candidates for inclusion into the module. However, it is not obvious where to choose the cutoff on the correlation with the seed profile that is minimally required for additional genes in order to belong to the module. To solve this problem, module *enrichment p values* are computed according to the enrichment of all regulators (motifs) in the extended modules as a function of the correlation cutoff. To determine the module *enrichment p value* for the enrichment of a particular motif (regulator) in an extended module with n genes, we first calculate the mean score of that motif in the module by averaging out the entries in the original motif (regulator) data matrix in the column corresponding to the motif (regulator) and the rows corresponding to genes in the module. We then compare this mean score to the distribution of scores obtained on a random selection of *n* genes, for the same motif (regulator). The mean score of a module by random gene selection is approximately Gaussianly distributed (central limit theorem), with mean equal to the mean over all genes, and variance equal to the overall variance divided by the size of the module.

This Gaussian approximation of the H_0-hypothesis is used to calculate the module *enrichment p value* for each motif or regulator. This enrichment p value is calculated at different correlation thresholds. Since the number of genes that are recruited in the extended module depends on the correlation threshold, the central limit theorem (valid from 20 genes in the extended module) will be satisfied at almost all correlation thresholds. If motifs and regulators identified in the seed discovery step appear to be over-represented in the extended sets, the correlation resulting in the largest enrichment is considered optimal (see Figure 6- 1).

6.2 Application of ReMoDiscovery on *Saccharomyces cerevisiae* data

We applied ReMoDiscovery on a dataset assessing gene expression during cell cycle (Spellman *et al*, 1998) and a dataset assessing gene expression in stress related conditions (Gasch *et al*, 2000), each in combination with the ChIP-chip (Harbison *et al*, 2004) and the regulatory motifs (Kellis *et al*, 2003)(for a description of all data sources see chapter 3).

In the seed discovery step, we detected 20 seed modules for the Spellman dataset (Spellman *et al*, 1998) and 104 seed modules for the Gasch dataset (Gasch *et al*, 2000) that were all statistically significant at a cutoff of 0.05 for the *seed module p-value*[§§]. To assess the biological relevance of the seed discovery step we compared our results with literature knowledge. We consider a seed module 'confirmed' if all of its regulators could be linked to the same biological process by the available literature. For the Spellman dataset (Spellman *et al*, 1998) 15 out of 20, and for the Gasch dataset (Gasch *et al*, 2000) 53 out of 104 seed modules were supported by literature. For the Spellman dataset (Spellman *et al*, 1998) the seed modules are displayed in Figure 6- 4, while for the Gasch dataset (Gasch *et al*, 2000) they are represented in Figure 6- 5. A large part of the seed modules (18 out of 20 for the Spellman dataset (Spellman *et al*, 1998); 63 out of 104 for the Gasch dataset (Gasch *et al*, 2000)) could be extended by the second step of the algorithm. All regulatory programs of the extended modules were supported by literature meaning that the regulators of the genes in a module are indeed involved in a similar biological function as the genes in the module.

In some cases, seed modules could not be extended, i.e. no additional correlated genes appeared to be present in the dataset under study. This implies that either the true module size was extremely small (only a few genes belong to the module) or the module's regulatory program, although being biologically relevant, was not active in the conditions tested in the expression data. The identification of the regulatory program in the seed discovery step is to a large extent determined by the ChIP-chip and motif data. Since motif data are condition independent, a common motif does not necessarily imply co-expression of the corresponding genes in the tested microarray conditions. Similarly, because of the discrepancies in experimental conditions between available ChIP-chip and expression data, evidence from the ChIP-chip data does not automatically imply support by all microarray data. As a result, a module can only be extended with additional genes if its regulatory program appears to be active in the

[§§] Significant seed modules that only contained one gene were omitted.

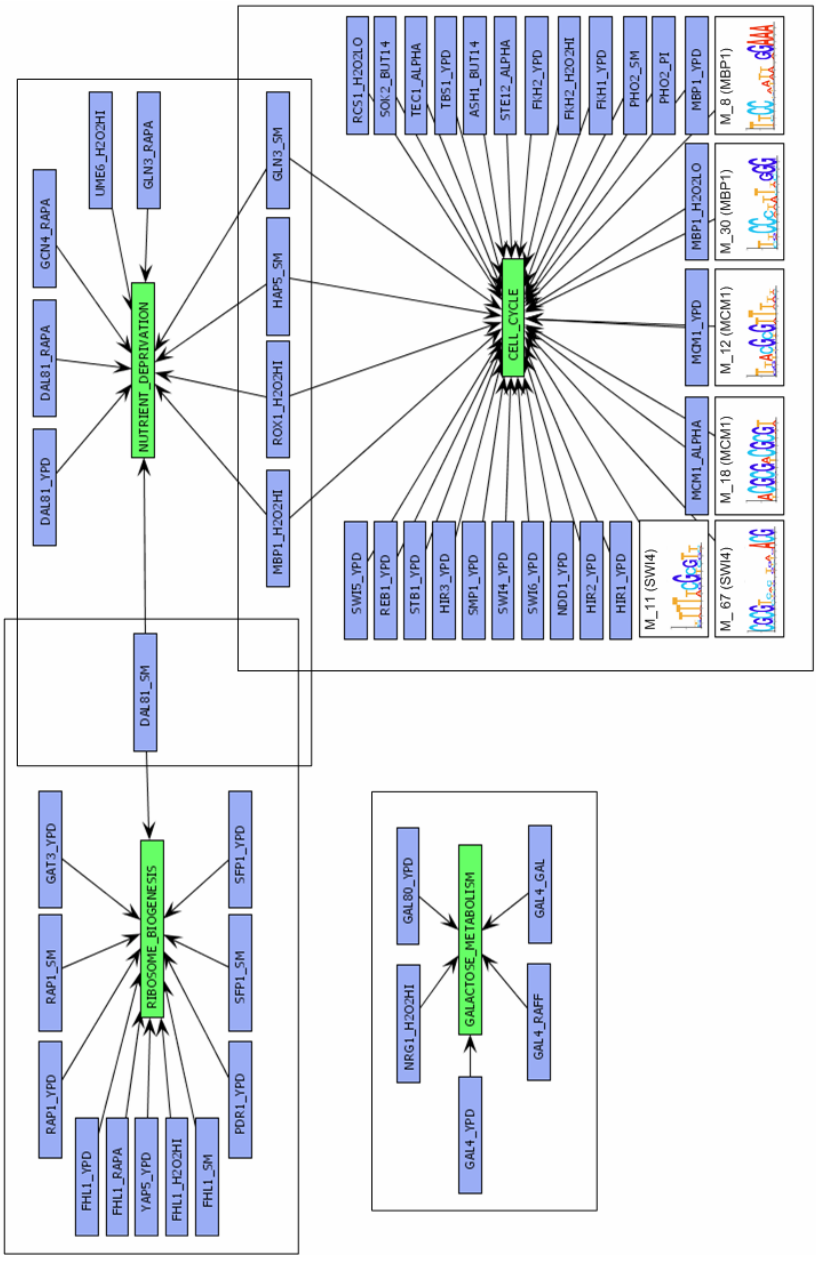


Figure 6- 4: caption on page 82

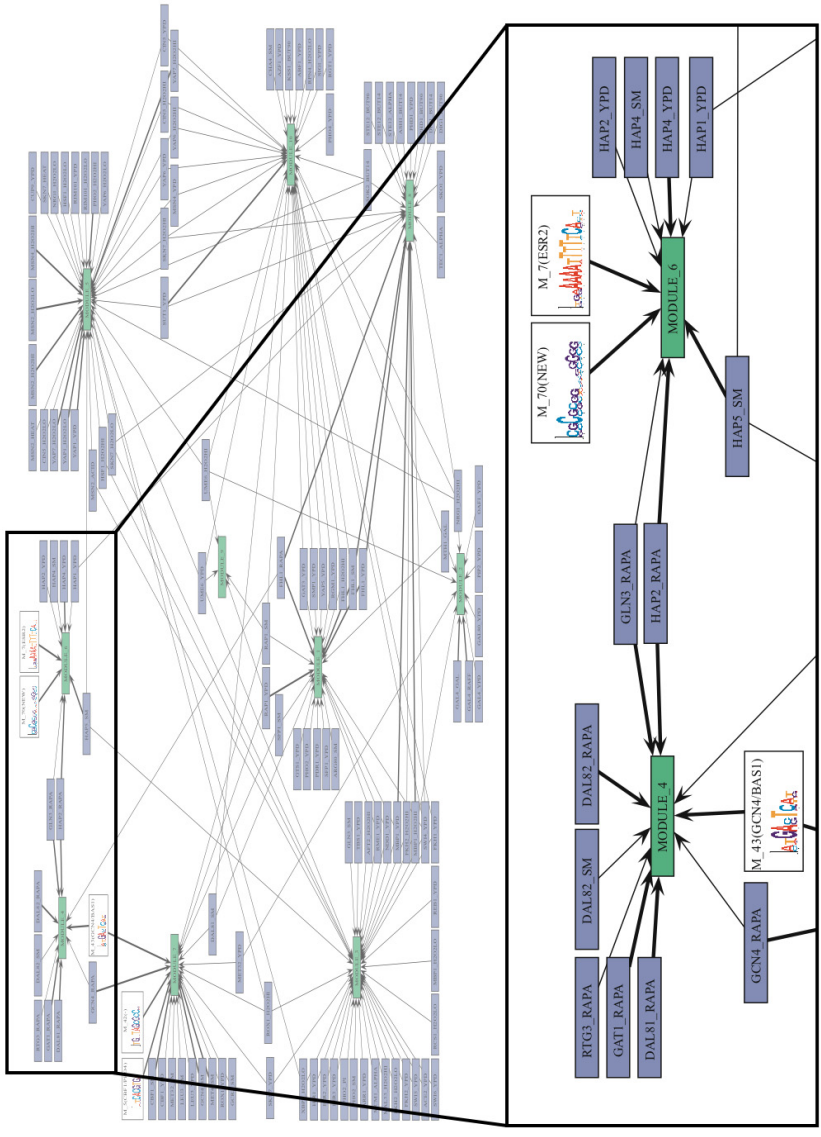


Figure 6- 5: caption on page 82

conditions underlying the used microarray study. Based on this observation, we subdivided modules in a category involved in general metabolism (such as ribosome synthesis, galactose metabolism) that was found active in both datasets and a category related to processes of which the activity was restricted to either one of the datasets. The latter group includes modules involved in cell cycle, which were extended in the Spellman dataset (Spellman *et al*, 1998), and modules related to nutrient deprivation, stress, respiration, amino acid metabolism, filamentous growth and meiosis extended in the Gasch dataset (Gasch *et al*, 2000).

A more detailed description of the modules is given below. To summarize results, modules with respective regulatory programs involved in the same biological process are discussed together.

Figure 6- 4: Overview of the seed modules identified in the Spellman data set (Spellman *et al*, 1998). For visualization purposes, seed modules with similar function are combined (indicated in green). A regulator or motif that is part of a regulatory program of an extended module is indicated in the figure by a bold edge from the regulator or motif to its module. In the Spellman data set modules were identified of which the genes involved in mainly four biological processes: ribosome biogenesis, nutrient deprivation, cell cycle and galactose metabolism. Most modules were involved in cell cycle, which makes sense since the Spellman data set measures gene expression levels during cell cycle. Some regulators regulate genes that are involved in multiple processes. Dal81, for instance, regulates genes involved in ribosome biogenesis but also genes that function during nutrient deprivation. The regulatory motifs of some extended modules are indicated in the figure.

Figure 6- 5: Overview of the seed modules identified in the Gasch data set (Gasch *et al*, 2000). For visualization purposes, seed modules with similar function are combined (indicated in green). A regulator or motif that is part of a regulatory program of an extended module is indicated in the figure by a bold edge from the regulator or motif to its module. In the Gasch data set that measures gene expression levels in various stress conditions, ReMoDiscovery identified modules involved in ribosome biogenesis, galactose metabolism, cell cycle, nutrient deprivation, stress, respiration, amino acid metabolism, filamentous growth and meiosis. Many regulators regulate genes that are involved in different biological processes. The figure focusses on modules involved in nutrient deprivation (module_4) and respiration (module_6). The modules involved in these processes consist of regulators and motifs. Some regulators, such as GLN3 or HAP2, regulate genes that are involved in both biological processes.

6.2.1 Modules involved in ribosome biogenesis

Modules involved in ribosome biogenesis are active in both the Spellman (Spellman *et al*, 1998) and the Gasch (Gasch *et al*, 2000) dataset. This could be expected as ribosome biogenesis is known to be tightly coupled to cell cycle progression as well as to environmental changes that affect growth rate. Different regulators were found to be associated with these ribosome related modules, including Arg80, Dal81, Fhl1, Gat3, Gts1, Mbp1, Mth1, Ndd1, Pdr1, Pho2/Bas2, Rap1, Rgm1, Rme1, Sfp1, Smp1, Swi4, and Yap5. Of these, Fhl1 and Rap1 were found in most modules. Consistently, both factors have been reported as main transcriptional regulators of ribosomal gene expression (Jorgensen *et al*, 2004; Martin *et al*, 2004). Also Sfp1 and Rgm1 have been implicated in ribosome biogenesis and most recent data indicate that the former could act as receiver of nutritional and stress derived signals (Cipollina *et al*, 2005; Jorgensen *et al*, 2004; Marion *et al*, 2004).

To our knowledge, no data are available that may confirm a direct involvement of the other transcription factors in ribosomal gene expression. Nevertheless, the processes in which these factors are known to be involved can be linked to ribosome biogenesis. For instance, Arg80, Dal81 and Pho2/Bas2 all function in the sensing and metabolic control of essential nutrients such as amino acid and phosphate, and it is well established that ribosomal protein genes expression is directly related to availability of essential nutrients (Jorgensen *et al*, 2004; Klein and Struhl, 1994; Powers and Walter, 1999; Thevelein *et al*, 2000; Winderickx *et al*, 2003). Another example is the transcriptional regulator Gat3, an uncharacterized member of the GATA family of transcription factors that controls the expression of nitrogen catabolic genes. The GATA factors are regulated by the Tor-pathway, a pathway that also regulates the expression of genes involved in ribosome biogenesis (Powers and Walter, 1999).

6.2.2 Modules involved in galactose metabolism

Both the Spellman (Spellman *et al*, 1998) and the Gasch (Gasch *et al*, 2000) dataset revealed active modules controlling so-called GAL genes (e.g., GAL3, GAL1, GAL7, GAL10), which encode proteins involved in galactose metabolism. These modules comprise the transcriptional regulators Gal4 and Gal80, which are key regulators of the galactose metabolism (Diep *et al*, 2005; Pilauri *et al*, 2005; Timson *et al*, 2002) and the transcriptional repressor Nrg1 that is known to mediate glucose repression of the GAL genes (Zhou and Winston, 2001).

Some transcriptional regulators that were retained only from the Gasch (Gasch *et al*, 2000) dataset point towards interactions between this module

for galactose metabolism and modules for other processes such as for cell cycle control via Mbp1 (see also cell cycle module) (Koch *et al*, 1993) and for amino acid metabolism via Met32 (see also amino acid module) (Blaiseau *et al*, 1997). In addition, the module for galactose metabolism contains regulators Oaf1, Pip2 and Ume6, involved in the induction of peroxisomal genes participating in β-oxidation (Schuller, 2003), potentially linking galactose metabolism to this process.

6.2.3 Cell cycle

Nine modules involved in cell cycle control were found active in the Spellman dataset (Spellman *et al*, 1998). The transcriptional regulators connected to these cell cycle related modules include components, such as Swi4, Swi6, Mbp1 and Stb1 constituting the transcriptional complexes SBF and MBF that operate during progression from G1 to S phase (Costanzo *et al*, 2003; Ho *et al*, 1999), as well as components involved in G2/M-specific transcription such as Fkh1, Fkh2 and Ndd1 (Kumar *et al*, 2000; Loy *et al*, 1999; Wittenberg and Reed, 2005). Further support for our analysis comes from the observation that other factors with a role in cell cycle regulation were also retrieved. These include the transcriptional repressor Xbp1, the corepressors Hir1, Hir2 and Hir3, and the transcription factors Pho2, Reb1 and Rcs1. Xbp1 is a repressor sharing homology with Swi4, Mbp1 (Mai and Breeden, 2005). Pho2 is involved with the early G1 transcription factor Swi5 in the control of the HO gene (Bhoite *et al*, 2001). Hir1, Hir2 and Hir3 are involved in cell cycle regulated transcription of histone genes (Prochasson *et al*, 2005; Spector *et al*, 1997). The transcription factor Reb1 is known to bind with high affinity to a sequence upstream of CLB2 (Van Slyke and Grayhack, 2003), a gene of which the regulation is important for completion of the normal vegetative cell cycle. The regulator Rcs1 is involved in timing the budding event of the cell cycle (Gil *et al*, 1991). Additional factors identified are Ash1, Sok2, Ste12 and Tec1. Their presence in our modules might link cell cycle to processes discussed below like mating type switching (Cosma, 2004) and the filamentous growth pathway (see also filamentous growth module) (Gagiano *et al*, 2002; Gancedo, 2001; Pan and Heitman, 2000).

6.2.4 Stress related conditions

Twentysix modules directing general and specific stress responses were identified and extended in the Gasch dataset (Spellman *et al*, 1998). These modules contain several transcriptional regulators and subsets of them are known to help fine-tuning stress responses to particular conditions.

Six modules with transcriptional regulators that mediate control of target genes under nutrient deprived conditions were active in the Gasch dataset (Gasch *et al*, 2000). The regulators include Gat1, Dal81, Dal82, and Gln3 which are all involved in nitrogen catabolite repression (Daugherty *et al*, 1993; Hofman-Bang, 1999; Winderickx *et al*, 2003), Gcn4, which is the main regulator in the general amino acid control (Albrecht *et al*, 1998; Hinnebusch, 2005; Natarajan *et al*, 2001), Rtg3 which is a transcription factor involved in regulation of genes required for de novo biosynthesis of glutamine and glutamate (Dilova *et al*, 2004), Fhl1, the forkhead factor that regulates ribosome biosynthesis in response to nutrient availability (Martin *et al*, 2004), and Hap2, a transcription factor of the TCA cycle (Cutler *et al*, 2001). Especially the ChIP-chip data obtained after treatment with rapamycin were informative to identify the different modules comprised in this nutrient deprivation module. Rapamycin is known to inhibit Tor (i.e. target of rapamycin) protein kinases, which function in a nutrient-sensing signal transduction pathway. Consistently, the processes and regulators for this module all show connections to the Tor-mediated nutrient-sensing signal transduction pathway (Cooper, 2002; Cutler *et al*, 2001; Dilova *et al*, 2004; Valenzuela *et al*, 2001).

The regulators Msn2 and Msn4 present in our modules are known key regulators of stress-responsive gene expression in yeast (Gorner *et al*, 1998; Martinez-Pastor *et al*, 1996; Schmitt and McEntee, 1996). Several regulators identified by our analysis can be related to triggering responses upon oxidative stress, like Skn7, Yap1, Hap5, Rox1, Hsf1, Nrg1, Pho2/Bas2 and Yap4/Cin5 (Nevitt *et al*, 2004; Pinson *et al*, 2000; Raitt *et al*, 2000; Vyas *et al*, 2005; Wong *et al*, 2003). A connection with oxidative stress may as well exist for Sut1, a factor that relieves according to literature hypoxic genes from repression by the Cyc8–Tup1 (Regnacq *et al*, 2001) co-repressor complex that is recruited to many promoters via regulatory proteins such as Rox1 (Deckert *et al*, 1995). Besides with oxidative stress, links with other stress responses could be derived. Indeed, Cup9 mediates copper resistance (Knight *et al*, 1994) while Yap6 confers resistance to cisplatin (Furuchi *et al*, 2001).

Some regulators present in the stress related module have been reported to be operative in aspects indirectly related to stress response, for instance, Ngr1, Rim101, Sok2 and Ume6 are linked by their roles in meiosis and sporulation (see also module for filamentation and meiosis) (Bogengruber *et al*, 1998; Rothfels *et al*, 2005; Shenhar and Kassir, 2001; Steber and Esposito, 1995) and Xbp1 is a stress-induced transcriptional repressor of the cell cycle (see also cell cycle module) (Mai and Breeden, 1997).

6.2.5 Respiration

The Gasch dataset (Gasch *et al*, 2000) allowed to identify an extendable module dedicated to respiration that includes the heme-responsive factor Hap1 and the subunits Hap2, Hap4, Hap5 of the heme-activated CCAAT-binding complex (Schneider and Guarente, 1991; Zitomer and Lowry, 1992). Two motifs, motif 7 (Esr2: GRRAAAWTTTTCACT) and 70 (CGCGnnnnnGGGS), of which the latter is defined as a "new" motif by Kellis *et al* (2003), could be associated with this module.

6.2.6 Amino acid metabolism

The modules for amino acid metabolism were recovered upon analysis of the Gasch dataset (Gasch *et al*, 2000). Support for the validity of this module came from the presence of Dal81, a positive regulator of multiple nitrogen catabolite repression genes (Daugherty *et al*, 1993; Hofman-Bang, 1999; Winderickx *et al*, 2003) and from the presence of Gcn4, the main regulator in the general amino acid control (Albrecht *et al*, 1998; Hinnebusch, 2005; Natarajan *et al*, 2001). Also present was Leu3, a transcriptional regulator of genes involved in nitrogen assimilation and in biosynthetic pathways of branched-chain amino acids (Hu *et al*, 1995; Kohlhaw, 2003). The regulators Cbf1, Met4 and Met 32 of our module have previously been shown to be required for the coordinated expression of the structural genes from the sulfur amino acid biosynthesis pathway (Blaiseau and Thomas, 1998; Thomas and Surdin-Kerjan, 1997).

Additional regulators present in this module may provide links to other regulatory programs. The presence of Rox1 and Skn7 can couple this network to the program for oxidative stress response (see also stress related module) (Raitt *et al*, 2000; Wong *et al*, 2003), while Sfp1, Rap1, and Gcr2, a coactivator of Rap1 (Menon *et al*, 2005), reflect links with ribosome biosynthesis (see also module for ribosome biogenesis) (Devlin *et al*, 1991; Jorgensen *et al*, 2004; Winderickx *et al*, 2003).

6.2.7 Modules involved in filamentous growth

Five modules related to filamentous growth could be retrieved from the Gasch dataset (Gasch *et al*, 2000). The filamentous growth pathway induces a morphogenetic switch under adverse growth, such as nutrient deprivation resulting. This switch induces the formation of pseudohyphae which are believed to facilitate foraging for scarce nutrients (Gagiano *et al*, 2002; Gancedo, 2001; Pan and Heitman, 2000). Consistent with literature, these

modules included the regulators Ste12 and its interacting partners Dig1 and Tec1, as well as Sok2 and its downstream regulators Ash1 and Phd1 (Gagiano *et al*, 2002; Gancedo, 2001; Pan and Heitman, 2000). The regulator Nrg1 also present in our module is known to function as a negative regulator of filamentous growth and as a repressor of FLO11, which encodes a cell surface glycoprotein required for filamentous growth (Kuchin *et al*, 2002) (see also galactose metabolism module).

Filamentous growth is known to be intimately linked to growth and cell cycle control. As such, it is not surprising that our analysis also retrieved for this module factors involved in cell cycle control such as components of SBF and MBF, i.e. Swi4 and Mbp1 (Costanzo *et al*, 2003; Ho *et al*, 1999) or Fkh2 (Zhu *et al*, 2000) (see also cell cycle module). This close link to growth also explains the presence in our modules of Fhl1 (Jorgensen *et al*, 2004; Martin *et al*, 2004), Hap1 (Schneider and Guarente, 1991; Zitomer and Lowry, 1992) and Sut1 (Regnacq *et al*, 2001) as they all have important functions in determining the growth potential of yeast cells. Our analysis additionally retrieved Sko1, an important regulator allowing cells to cope with osmotic stress. Osmotic stress can under some conditions induce filamentous and as such the presence of Sko1 in our modules makes sense (Pascual-Ahuir *et al*, 2001; Zaragoza and Gancedo, 2000).

6.2.8 Modules involved in meiosis

Finally, we identified one extendable module in the Gasch dataset (Gasch *et al*, 2000) that is regulated by Ume6 and Rap1. We refer to this module as being important for meiosis because literature confirmed that Ume6 has a key regulator function in this process (Mallory and Strich, 2003; Mitchell, 1994) while Rap1 is believed to control Ume1, a regulator that is required for the repression of early meiotic genes (Mallory and Strich, 2003).

6.3 Comparison with other module inference tools

In order to assess the differences between ReMoDiscovery and previously described algorithms for module detection, we applied some of the well known module detection tools for which we had access to a workable implementation (i.e. SAMBA (Tanay *et al*, 2004) and GRAM (Bar-Joseph *et al*, 2003)) along with ReMoDiscovery on the combined Spellman (microarray) (Spellman *et al*, 1998) and Harbison (ChIP-chip) (Harbison *et al*, 2004) dataset. In chapter 3 there is a description of the data sets, while chapter 4 describes the two related module discovery tools SAMBA (Tanay

et al, 2004) and GRAM (Bar-Joseph *et al*, 2003). In this section we will first compare gene and regulator content of our modules with the modules identified by GRAM and SAMBA at different parameter settings in order to check whether similar modules can be identified by the different module discovery methods. Subsequently, the biological results obtained by the three approaches will be analysed and compared.

We analyzed running times on these datasets at distinct parameter settings for each of the tested algorithms on an Intel Pentium 2GHz laptop with 512Mb RAM. Independent of the setting for the *overlap prior factor*, the SAMBA algorithm (Tanay *et al*, 2004) was rather quick, with running times around three minutes. For parameter settings close to its defaults, ReMoDiscovery performed slightly better. Running times were on the order of one minute. In general, the speed of the Apriori algorithm is roughly proportional to the number of modules that satisfy the constraints. In contrast, running times of the GRAM algorithm (Bar-Joseph *et al*, 2003) were infeasible if the data contained genes with more than 10 significant ChIP-chip interactions, due to the exponential increase in the number of candidate core modules (see (Bar-Joseph *et al*, 2003) for details). After filtering out those genes, running times decreased to about 20 minutes at the default parameter setting.

To compare the gene and regulator content between modules obtained by GRAM (Bar-Joseph *et al*, 2003), SAMBA (Tanay *et al*, 2004) and ReMoDiscovery, we used an unsupervised scoring scheme that considers gene content and regulator content separately, i.e. the normalized Jaccard similarity score (Shakhnovich *et al*, 2004). For a specific parameter setting, we verify for each pair of genes (regulators) whether these two genes (regulators) occur together in at least one module. Doing so for all gene (regulator) pairs and for both methods, one can define the number of true positives TP as the number of gene pairs occurring together at least once in both methods. Analogously, the number of false positives FP, true negatives TN and false negatives FN can be defined. As in Shakhnovich *et al* (2004), we used the Jaccard similarity score TP / (TP + FP + FN) to score the overlap between two module compositions. In addition, randomizations were used to determine the significance of a specific score. This leads to the notion of normalized similarity scores, expressed as the number of standard deviations from the mean of the distribution of Jaccard similarity scores for randomized module compositions.

Since parameter settings influence the module composition, we calculated normalized Jaccard similarity scores on the results for a number of parameter settings. We used ReMoDiscovery with a ChIP-chip threshold (one minus the *ChIP-chip p-value*) equal to 0.99, a correlation threshold of 0.75 and a minimum of one motif and four regulators respectively. When comparing regulator content from ReMoDiscovery to SAMBA (Tanay *et al*, 2004) and

GRAM (Bar-Joseph *et al*, 2003), we looked at the ReMoDiscovery seed modules for a minimum number of regulators equal to 3, 4, 6, 8 and 10. We also examined the influence of a variation in ChIP-chip threshold, in the range [0.98 – 0.999] (values below 0.98 were not tested as the quality of the biological outcome started to decrease). The SAMBA (Tanay *et al*, 2004) *overlap prior factor* was varied between 0 and 1, in steps of 0.05. The latter parameter describes the extent of overlap that is permitted between different modules in the same solution. For the GRAM algorithm (Bar-Joseph *et al*, 2003), we varied all user defined parameters in a wide range: the base 10 logarithm of the *core profile p-value cutoff* between minus 12 and minus 3, the *num in core cutoff* between 5 and 97 and *module p-value cutoff* between 0.001 and 0.05. When comparing gene content from ReMoDiscovery to SAMBA (Tanay *et al*, 2004) and GRAM (Bar-Joseph *et al*, 2003), we considered ReMoDiscovery output for varying correlation threshold, in the range [0.65 – 0.9].

For all settings, we observed that both the overlap in gene and regulator content between the GRAM (Bar-Joseph *et al*, 2003) modules and the seed modules of ReMoDiscovery was highly significant (normalized Jaccard similarity scores around 15 and 25 standard deviations respectively). Since GRAM (Bar-Joseph *et al*, 2003) generally returns modules with less regulator content, the similarity in regulatory programs was best for the most stringent ReMoDiscovery ChIP-chip threshold (see Figure 6- 6). Accordingly, gene content was most similar if the ReMoDiscovery correlation threshold was lowered. From these results, we conclude that the ReMoDiscovery seed modules and the GRAM (Bar-Joseph *et al*, 2003) modules represent similar patterns in the data, the former focusing on modules with fewer genes and more regulators, the latter on modules with more genes and less regulators. The regulatory programs discovered by SAMBA (Tanay *et al*, 2004), using the discretization method suggested by the authors, did not significantly resemble those of ReMoDiscovery or GRAM (Bar-Joseph *et al*, 2003). The gene content on the other hand did show some overlap (see Figure 6- 6).

We performed a similar analysis, this time with the extended seed modules of ReMoDiscovery. Extending the seeds generally results in a smaller number of statistically overrepresented regulators in the modules (Table 6-1), but an increase in gene content. Accordingly, scores for overlap in gene content with GRAM (Bar-Joseph *et al*, 2003) and SAMBA (Tanay *et al*, 2004) now improved. The normalized Jaccard similarity score increased from 15 standard deviations to about 100 standard deviations for GRAM (Bar-Joseph *et al*, 2003) and from 6 to about 21 standard deviations for SAMBA (Tanay *et al*, 2004). At the same time, the regulator overlap with GRAM (Bar-Joseph *et al*, 2003) increased to about 50 standard deviations.

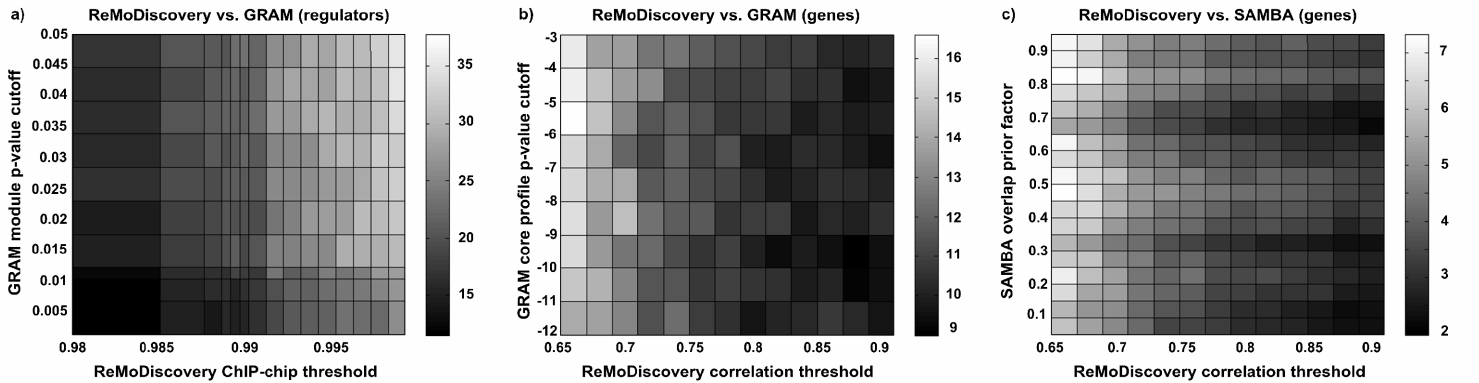


Figure 6- 6: Representative examples from the module content similarity analysis. The significance of the similarity in module content between ReMoDiscovery seed modules and GRAM (Bar-Joseph *et al*, 2003) and SAMBA (Tanay *et al*, 2004) output is shown at different parameter settings. The color bar on the right indicates the normalized Jaccard similarity score, that is, the number of standard deviations from the mean of the distribution of Jaccard similarity scores on randomized module partitioning. For the ChIP-chip threshold c_t the tested settings were not selected equidistantly: close to the default value (0.99) more ChIP-chip threshold values have been examined. **(a)** Regulator content similarity between ReMoDiscovery and GRAM, with varying GRAM module p value cutoff and ReMoDiscovery Chip-chip threshold. Lowering the ChIP-chip threshold of ReMoDiscovery results in a decreasing overlap in regulator content, indicating that GRAM in general finds sets of genes regulated by a smaller number of regulators. **(b)** Gene content similarity between ReMoDiscovery and GRAM, with varying GRAM core profile p value cutoff and ReMoDiscovery correlation threshold. Lowering the ReMoDiscovery correlation threshold results in an increasing overlap in gene content. This points to the fact that GRAM in general detects modules consisting of a larger number of genes. **(c)** Gene content similarity between ReMoDiscovery and SAMBA, with varying SAMBA overlap prior factor and ReMoDiscovery correlation threshold. Lowering the ReMoDiscovery correlation threshold again results in an increasing overlap in gene content, indicating that SAMBA identifies modules consisting of a larger number of genes.

In other words, increasing the number of genes in a module (higher gene content) corresponded to decreasing the number of regulators (lower regulator content), making results even more similar to those of GRAM (Bar-Joseph *et al*, 2003).

All tools discussed in this study serve two purposes: they simultaneously identify clusters of co-expressed genes and the corresponding regulatory programs. To evaluate the first aspect, we calculated the average functional overrepresentation of the modules detected by each of the tools on our benchmark dataset (see Table 6- 1). We used default parameter settings for GRAM (Bar-Joseph *et al*, 2003), SAMBA (Tanay *et al*, 2004) and ReMoDiscovery. From these results it appears that, for all tools, regulatory modules are well enriched for known functional classes. For ReMoDiscovery, the enrichment score improved upon extension of the seeds.

In order to test the sensitivity of these tools in retrieving regulators known to be involved in cell cycle , we compiled a list containing every regulator that was present in the regulatory program of at least one cell cycle enriched module identified by ReMoDiscovery, GRAM (Bar-Joseph *et al*, 2003) or SAMBA (Tanay *et al*, 2004). The regulators in this list that are involved in cell cycle according to the Saccharomyces Genome Database (Balakrishnan *et al*, 2005) were considered "cell cycle regulators": ACE2_YPD, FKH1_YPD, FKH2_H2O2Hi, FKH2_H2O2Lo, FKH2_YPD, MBF1_YPD, MBP1_H2O2Hi, MBP1_H2O2Lo, MBP1_YPD, MCM1_Alpha, MCM1_YPD, NDD1_YPD, RFX1_YPD, RPN4_YPD, STB1_YPD, SWI4_YPD, SWI5_YPD SWI6_YPD, YOX1_YPD (nomenclature adopted

Table 6- 1: Summary of the results of the GRAM, SAMBA and ReMoDiscovery module discovery methods when applied on our benchmark data set (Harbison *et al*, 2004; Spellman *et al*, 1998). The number of modules, the mean, minimum and maximum number of genes in the modules, the mean fuctional enrichment of the genes in the modules and the mean, minimum and maximum number of regulators is indicated for each method. All methods are able to identify modules with a good functional enrichment. For ReModiscovery the mean functional enrichment increases upon extension of the seed modules.

Method	No	Genes				Regulatory program		
		Mean	Min	Max	Mean functional enrichment	Mean	Min	Max
ReMoDiscovery (seed modules)	20	2.05	2	3	0.05	6.15	3	12
ReMoDiscovery (extended modules)	18	67.72	6	200	2.00E-03	3.50	2	6
GRAM	274	6.80	5	33	0.02	2.35	1	8
SAMBA	205	57.53	5	265	1.10E-02	4.16	0	31

from Harbison *et al* (2004)). Subsequently we tested how many of these regulators occurred in the regulatory programs of any of the cell cycle related modules (see Table 6- 2). We also displayed the ratios of the number of known cell cycle regulators over the total number of regulators detected in a module's program, averaged over all modules. These results show that both ReMoDiscovery and GRAM (Bar-Joseph *et al*, 2003) had a considerably higher sensitivity than SAMBA (Tanay *et al*, 2004) in retrieving cell cycle related regulators.

Conclusions about specificity should be treated with care because in the absence of a golden standard (i.e. a completely characterized network of interactions) the number of false positive predictions can never be quantified. Although the regulatory programs of GRAM (Bar-Joseph *et al*, 2003) and ReMoDiscovery seem to be more enriched in cell cycle related regulators (larger ratio of known cell cycle related regulators over the total number of regulators than SAMBA (Tanay *et al*, 2004)), it is not possible to distinguish between true and false positives without further experimental validation.

Table 6- 2: Summary of the significantly cell cycle enriched modules, identified by the GRAM (Bar-Joseph *et al*, 2003), SAMBA (Tanay *et al*, 2004) and ReMoDiscovery (Lemmens *et al*, 2006) module discovery methods. The number of modules, the mean, minimum and maximum number of genes in the modules and the mean, minimum and maximum number of regulators is indicated for each method. In addition, we mention the ratios of the number of known and unknown cell cycle regulators over the total number of regulators detected in a module's program, averaged over all modules and the total number of cell cycle regulators that were identified. Although ReMoDiscovery identifies less cell cycle enriched modules (8 modules by ReMoDiscovery versus 33 by GRAM or 14 by SAMBA), the regulatory program of these modules consists mostly of known cell cycle regulators. In contrast, SAMBA seems not to be able to assign cell cycle regulators to cell cycle modules.

Method	No	Genes			Regulatory program					
								No cell cycle R/ all R	No non cell cycle R/ all R	No cell cycle R
		Mean	Min	Max	Mean	Min	Max			
ReMoDiscovery (seed modules)	2	2.00	2	2	4.00	3	5	0,8	0.20	6
ReMoDiscovery (extended modules)	8	97.38	12	200	3.50	2	6	0.92	0.08	10
GRAM	33	6.47	5	11	2.66	1	6	0.74	0.26	17
SAMBA	14	58.29	17	155	2.57	0	12	0.29	0.71	5

6.4 Discussion

ReMoDiscovery is a two-step methodology to unravel active modules based on the concurrent analysis of three independently acquired data sources. The seed discovery step predicts putative seed modules (consisting of genes, regulators and corresponding motifs). The seed extension step optimizes the gene content of the modules and indicates whether the seed modules' regulatory program is active in the microarray data.

The data integration problem is tackled in a very direct way: using the Apriori algorithm, no iteration over the different data sources is required. As regards the algorithmic properties, a comparison of ReMoDiscovery with other module detection tools revealed that speed is one of the major advantages of the Apriori strategy. ReMoDiscovery's running times and memory requirements are drastically smaller than those of certain other module detection algorithms such as the GRAM (Bar-Joseph *et al*, 2003) algorithm. This is important as most module discovery algorithms require repeated testing to find the optimal parameter settings. Together with the straightforward biological interpretation of the parameters, its speed turns ReMoDiscovery into a user-friendly, readily tunable tool.

The biological relevance of our method was assessed by applying it on the extensively studied Spellman (Spellman *et al*, 1998) and Gasch (Gasch *et al*, 2000) datasets. Comparison of our results with literature showed that for many of our statistically significant seed modules experimental evidence existed. For modules of which no direct evidence existed so far, very often a plausible explanation for their composition could be inferred from literature and potential new links between the detected pathways and modules could be derived. A seed module that can be extended with more genes in the seed extension step gives a clue on the regulatory program being active in the prevailing conditions of the tested microarray experiment. Based on this observation, a distinction could be made between modules involved in general metabolism found active in both datasets (for instance, ribosome synthesis, galactose metabolism) and the more specialized modules (for instance, cell cycle, nutrient deprived conditions, stress related conditions, amino acid metabolism, respiration, filamentous growth or meiosis) of which the activity was restricted to either one of the datasets.

In contrast to previous approaches (Bar-Joseph *et al*, 2003; Xu *et al*, 2004), our method uses motif information as an additional independent input source. To avoid circular reasoning, we ensured that this information was derived from sequence information only and did not rely on any other experimental data source (for instance, as available in the motif compendium of Kellis *et al* (2003)). This implies that the compendium of motifs we used as input dataset is far from complete, explaining why we detected less motifs

for each module as compared to, for instance, Kato *et al* (2004) or Harbison *et al* (2004).

In order to assess to what extent ReMoDiscovery discovers modules similar to those detected by other module identification tools, we compared it with previously described tools on the same benchmark set. Compared to GRAM (Bar-Joseph *et al*, 2003), we found a significant overlap in both gene and regulator content of the detected modules over a sweep of different parameters. The similarity between both algorithms was larger when comparing the results of GRAM (Bar-Joseph *et al*, 2003) with those of the extended seed modules than with the original seed modules. This difference reflects the trade-off between the number of regulators and the number of genes in biological modules: modules comprising a regulatory program with many regulators (such as our seed modules) can be expected to contain few genes with a potentially highly related function. In a module, the number of genes will usually increase with a decreasing number of regulators. Obviously, there will be more genes that only share part of their regulatory program, i.e. the part that is active under the tested set of conditions. While our seed modules give a view on the complete regulatory program, our extended modules highlight the program active in the microarray dataset. They contain more genes and are more similar to the GRAM (Bar-Joseph *et al*, 2003) output. Hence, with ReMoDiscovery we offer an algorithm that can be used to focus on very specific regulatory programs (seed modules) as well as on less specific modules with more genes (extended seed modules). The most appropriate choice depends on the specific research question under study, so usually there is no single best solution for the outcome of a module detection algorithm.

In addition, we compared the output of SAMBA and ReMoDiscovery on the same benchmark data set. The regulatory programs that we obtained with the SAMBA algorithm (Tanay *et al*, 2004) did not significantly resemble those of ReMoDiscovery or GRAM (Bar-Joseph *et al*, 2003). The possibility might exist that SAMBA (Tanay *et al*, 2004) focuses on other aspects of the data and therefore detects fundamentally different modules. However, based on our analysis we believe that the regulatory programs recovered by SAMBA (Tanay *et al*, 2004) are unlikely to be biologically meaningful as the sensitivity in detecting cell cycle related regulators was low. Most likely the available download of the SAMBA-Expander application (Tanay *et al*, 2004) is not yet fully adjusted to the use of heterogeneous data sources.

Chapter 7

Condition-dependent combinatorial regulation in *E. coli*

In chapter 6 we already showed that tools based on association rule mining algorithms are equally suitable for module detection as other more frequently used graph-based or probabilistic algorithms. However, if we want to use all available microarray data, a condition selection or bicluster strategy is necessary.

In collaboration with Prof. De Bie (University of Bristol, UK) we have therefore developed DISTILLER a module detection tool that allows for condition selection. Since this bicluster strategy makes it harder to compute, we relied on the more advanced association rule mining algorithm, CHARM. CHARM uses a depth-first strategy for closed itemset, which guarantees a better scalability, especially on non-sparse data sets such as an expression compendium. By relying on CHARM, a bicluster component (condition selection) could be added and reduces the number of parameters. In section 7.1 we will describe our approach.

In collaboration with the research group of Prof. Collado-Vides (UNAM, Cuernavaca, Mexico), we obtained access and support with regard to the well-known database RegulonDB which contains curated information on the transcriptional network of the model organism *Escherichia coli*. By applying DISTILLER on publicly available regulatory motif and expression data (see chapter 3 for a detailed description of the data), the relation between condition-specific co-expression and combinatorial regulation in the transcriptional network of *E. coli* was examined, confirming 62% of 736 interactions described in RegulonDB and predicting 278 novel interactions (section 7.2.1). Experimental validation of predicted targets for the well-studied oxygen sensor "FNR" showed the effectiveness of our approach. We also confirm the condition dependency of many well known regulators, like ArcA or FNR, and predict for less studied regulators, like CpxR, the conditions under which they might be active (section 7.2.2). Surprisingly the level of regulatory complexity of the inferred modules seemed lower than

what one would expect from RegulonDB, indicating how complex regulatory programs tend to decrease the degree of modularity and may result in more specific expression profiles (section 7.2.3).

To conclude, DISTILLER was used to extend the transcriptional interaction network of RegulonDB by using curated information and to describe the condition dependency and modularity of the inferred transcriptional network.

The methodological results of this study have been accepted for publication in:

Lemmens K., De Bie T., Dhollander T., Monsieurs P., Bart De Moor B., Collado-Vides J., Engelen K., Marchal K. (2008) "The condition-dependent transcriptional network in *Escherichia coli*". Accepted for publication in Annals of NYAS, DREAM2.

The biological study is currently submitted for publication:

Lemmens K., De Bie T., Dhollander T., De Keersmaecker S., Thijs I., Schoofs G., De Weerdt A., De Moor B., Vanderleyden J., Collado-Vides J., Engelen K., Marchal K. (2008) DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli*. Submitted.

7.1 Motivation

The transcriptional network of *Escherichia coli* is among the best characterized transcriptional networks. Based on our current knowledge of this network it is clear that complex regulons are prevalent: more than 50 % of the genes are regulated by more than one transcriptional regulator (Gutierrez-Rios *et al*, 2003; Salgado *et al*, 2006). However, most of these complex regulons were inferred by curating experimental evidence for a regulator-target interaction from independent studies, each of which focuses on an individual interaction (Salgado *et al*, 2006). Evidence from these independent studies is obtained from measurements in different environmental conditions. Current network representations do not take into account this condition dependency of the regulatory interactions (Luscombe *et al*, 2004; Shen-Orr *et al*, 2002). As a consequence, it is not clear from these static networks whether regulators controlling the same gene are indeed needed together in the same conditions or act independently from each other in different conditions (Balaji *et al*, 2006; Balaji *et al*, 2007).

Bicluster strategies are well suited to map both the condition dependency and the modularity of the transcriptional network from microarray compendia (Cheng and Church, 2000; Ihmels *et al*, 2002; Murali and Kasif, 2003; Van den Bulcke *et al*, 2006), but do not give any information on the transcriptional program of the modules. Methods have been developed to infer transcriptional interactions from microarrays only, by assuming that the

transcription profile of the regulator is related to that of its targets genes (Faith *et al*, 2007; Michoel *et al*, 2007; Segal *et al*, 2003). Integrative approaches can avoid this assumption by exploiting data sources that are complementary to microarrays. These methods have been successfully used to infer simple regulons (Chen *et al*, 2007; Ernst *et al*, 2008; Gao *et al*, 2004) or to directly infer complex regulons, i.e. the set of genes regulated by several regulators (Bar-Joseph *et al*, 2003; Lemmens *et al*, 2006; Tanay *et al*, 2004). Most of the previously mentioned integrative approaches use the level to which the target genes of a particular regulator share a similar expression pattern as a feature for inferring regulator-target interactions, but do not include an explicit condition selection strategy as is the case with bicluster strategies (Van den Bulcke *et al*, 2006). A few exceptions exist, including the graph-based data integration tool SAMBA (Tanay *et al*, 2004) and the sequential approach described by Bonneau *et al* (2006). The latter approach searches simultaneously for biclusters and *de novo* motifs in the promoter region of the bicluster genes by using cMonkey (Reiss *et al*, 2006), and subsequently apply a regression strategy (Bonneau *et al*, 2006), to associate a regulatory program with the inferred biclusters.

To study the yet unknown relation between modularity, combinatorial regulation and condition dependency of bacterial networks, we developed the data integration framework 'DISTILLER' (Data Integration System To Identify Links in Expression Regulation). DISTILLER simultaneously identifies condition-dependent modularity and complex regulatory programs by integrating expression data and interaction data.

7.2 DISTILLER

The core of our framework is a data integration strategy that relies on itemset mining. In chapter 6 we already showed that approaches based on itemset mining are equally suitable for reconstructing networks than the more frequently used graph based (Tanay *et al*, 2004) or probabilistic methodologies (Segal *et al*, 2003). Although both our previous and our current approach are based on item set mining, the set up of DISTILLER is completely different from the one used in ReMoDiscovery (Lemmens *et al*, 2006) (see also chapter 6). In contrast to ReMoDiscovery, DISTILLER not only searches for sets of highly co-expressed genes that share controlling regulators, but also selects the experimental conditions for which the selected genes are co-expressed. By including this 'bicluster strategy' genes are no longer required to be co-expressed over all conditions. This allows applying the algorithm to heterogeneous expression compendia in order to assess the condition dependency of the interaction network. Extending itemset mining approaches towards biclustering is a non-trivial task since commonly used distance measures for assessing co-expression such as

correlation do not longer meet the basic subset relation constraints of an itemset mining framework. We therefore designed a novel distance measure (see section 7.1.1). Since the condition selection increases the combinatorial nature of the problem, DISTILLER relies on the closed itemset mining strategy CHARM (Zaki and Hsiao, 2002) instead of Apriori (Agrawal and Imielenski, 1993). Not only made this change in itemset mining algorithm the search for modules more efficient, but it also drastically reduced the number of user defined parameters, thereby enhancing the interpretability of the results.

One of the main advantages of itemset mining approaches in comparison to "optimization based" methods is that they investigate all possibly interesting solutions (in this case, modules) and thus are not subject to problems associated with local optima. However, this also implies that the output of virtually all itemset mining algorithms is a long list of possibly interesting results without rigorous statistical significance scores. In order to make interpretation of such lists feasible, we introduced in this work an intelligent filtering step that is based on a statistically inspired interest score (see section 7.1.2). The result is a concise list of statistically significant and biologically interesting modules. Although in this study our method was only applied on an expression compendium and motif data, other data sources related to transcriptional interactions such as additional microarrays or ChIP-chip can be integrated as well with our approach.

Our methodology consists of three steps: (1) the identification of seed modules (section 7.1.1); (2) the reduction of the set of all seed modules to a manageable set of non-redundant and significant seed modules (section 7.1.2); and (3) the extension of the thus obtained seed modules with additional genes (section 7.1.3).

7.2.1 Identification of seed modules

Valid seed modules are seed modules that contain a minimal number of genes (i.e. gene content threshold) while the genes in the module must be co-expressed in a sufficiently large number of conditions and share motif instances for the same regulator(s). In comparison to the definition of a valid module in the simpler ReMoDiscovery framework (chapter 6), a condition selection is now possible.

Although our approach can be extended to any number of input matrices (data sources), we used two input matrices to generate the results in this paper:

- An appropriately normalized expression data compendium **A** with dimensions $N_G \times N_C$, where N_G indicates the total number of genes in the compendium and N_C the total number of conditions.
- A binary 'regulatory motif matrix' **R** (input interaction matrix) with dimensions $N_G \times N_R$, where N_R is the total number of regulators for which motif data is available. Each element r_{ip} of this matrix indicates whether the upstream region of a specific gene *i* contains a statistically significant motif instance of the known regulatory motif model of that specific regulator *p*.

Let $\mathbf{g}^{(m)}$ be the set of gene indices that correspond to genes in module m. DISTILLER initially identifies seed modules that satisfy three constraints:

- (CG) The module should contain a minimum number *G* of genes (i.e. gene content threshold), or $\left|\mathbf{g}^{(m)}\right| \geq G$.
- (CR) All genes $\mathbf{g}^{(m)}$ in the module should contain motif instances for a sufficient number *R* of common, a priori unspecified regulators. Let **1** represent an all-ones vector of appropriate dimensions. Then, this constraint can be formulated mathematically as $\left|\left\{p \mid \mathbf{1}^{\mathrm{T}}\mathbf{R}_{\mathbf{g}^{(m)},p} = \left|\mathbf{g}^{(m)}\right|\right\}\right| \geq R$. Recall that *p* indicates a regulator. Curly brackets specify the set of regulators that regulate all genes in the module, while vertical bars indicate the size of a set.
- All genes $\mathbf{g}^{(m)}$ in the module should be significantly co-expressed in a sufficiently large, a priori unspecified set of experimental conditions. We indicate this required number of conditions by *C*. In order to calculate the maximal valid condition set $\mathbf{c}^{(m)}$ given a certain gene set $\mathbf{g}^{(m)}$, we first compute the difference between the largest and smallest expression levels in the gene set for each of the conditions *j*: $BW_j = \max(\mathbf{A}_{\mathbf{g}^{(m)},j}) - \min(\mathbf{A}_{\mathbf{g}^{(m)},j})$. We call BW_j the bandwidth for condition *j*. Subsequently, these bandwidths are sorted in increasing order to obtain a sorted bandwidth sequence $BW_j^{(s)}$ that satisfies $BW_{j-1}^{(s)} \leq BW_j^{(s)} \leq BW_{j+1}^{(s)}$ (see Figure 7- 1). The sequence $BW_j^{(s)}$ is then compared with a pre-specified threshold bandwidth sequence $BW_j^{(th)}$ that is sorted as well: $BW_{j-1}^{(th)} \leq BW_j^{(th)} \leq BW_{j+1}^{(th)}$. Gene set $\mathbf{g}^{(m)}$ is said to be co-expressed in C^{max} conditions if the sorted bandwidth sequence

$BW_j^{(s)}$ is completely within the threshold bandwidth sequence $BW_j^{(th)}$ for $1 \le j \le C_{\max}$, i.e. if $BW_j^{(s)} \le BW_j^{(th)}$ for $1 \le j \le C_{\max}$. Constraint CC is satisfied if this property holds for $C_{\max} \ge C$. There are various ways to choose the pre-specified bandwidth threshold sequence. One data-driven approach is described below.

A naive exhaustive search for valid seed modules would require checking all possible combinations of genes (4000), motif instances (67), and experimental conditions (870). This is unfeasible for data sets of any reasonable size. To find valid modules more efficiently we relied on an item set mining approach called CHARM (Zaki and Hsiao, 2002), which drastically restricts the search space without running the risk to skip valid modules.

CHARM can be used to efficiently limit the number of combinations to be tested if different item sets (or gene sets) are related to each other by a valid "subset" relation, i.e. an item set can only satisfy all constraints if all of its subsets do. A consequence is that we can search for modules by starting with very small gene sets (containing just one gene), gradually expanding them, and stopping (or pruning) the search once a gene set is reached for which one of the module properties is violated. This pruning step results in a massive speed-up, making the method applicable to large data sets.

Figure 7- 1: Strategy of DISTILLER for the introduction of condition selection. Extending association rule mining approaches towards biclustering is a non-trivial task. Commonly used distance measures for assessing co-expression such as correlation do not longer meet the basic subset relation constraints of an itemset mining framework. We therefore designed a novel distance measure, the bandwidth.
A) Definition of the bandwidth and its use for condition selection. The top half shows hypothetical expression profiles for four different genes. For each of the four conditions, the bandwidth for this set of genes is indicated with a vertical bidirectional arrow. The lower half shows the bandwidth sequence for these expression data, obtained by sorting the bandwidths in increasing order. Note that the solid lines in the lower part represent bandwidth sequences whereas the dashed lines in the upper panel denote expression profiles. Two threshold bandwidth sequences (BW(th,1) and BW(th,2)) are shown as well (bold lines). If we were to use BW(th,1), this set of three genes would qualify as a module for the expression data, since the bandwidth sequence lies entirely below BW(th,1). However, if we were using BW(th,2), this set of genes would not qualify as a module.
B) Selection of the band width threshold based on random data. To illustrate how a threshold bandwidth sequence is obtained from the data, we show five bandwidth sequences corresponding to randomly sampled sets of genes. The dashed line is a candidate threshold bandwidth sequence, obtained by connecting all second smallest values for all four sorted conditions. The threshold bandwidth sequence that is actually used is the one that qualifies a fixed fraction of randomly sampled gene sets as a module (box p-value).

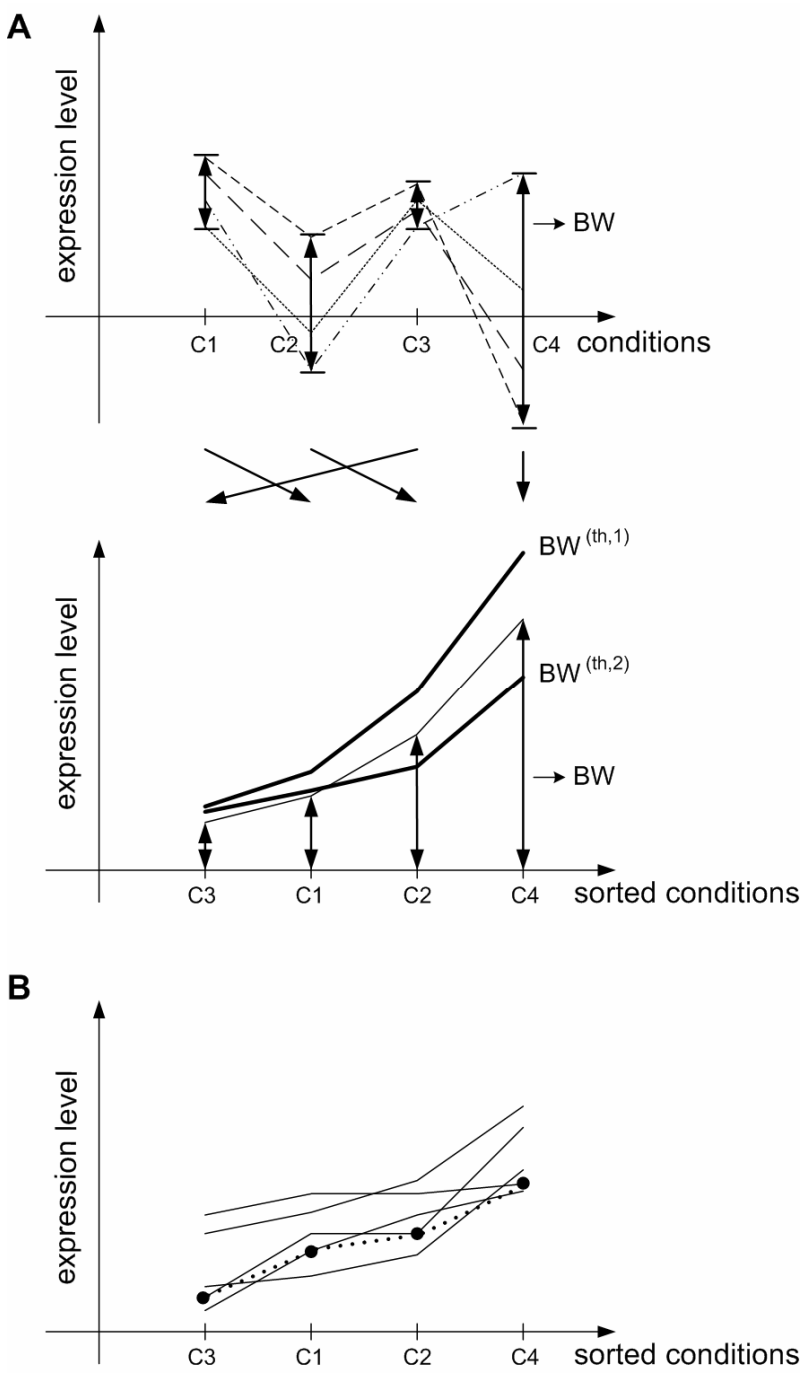


Figure 7- 1: caption on page 100

Implementing this subset relation for the integration of the motif data is straightforward as the motif matrix is a binary matrix (see chapter 3): a target gene has a motif instance for a regulator if the corresponding gene-regulator entry in the motif matrix is equal to one. However, defining "sufficient co-expression" in a way that the subset relation also allows for condition selection in the expression matrix is less trivial. To this end we used the concept of the bandwidth, which is defined as the difference between the largest and smallest expression levels in the gene set (Figure 7- 1). The bandwidth sequence, the set of bandwidths for all conditions sorted in increasing order, is subsequently compared with a pre-specified threshold bandwidth sequence and the genes are said to be co-expressed in a set of conditions if their bandwidth sequence is completely within the threshold bandwidth sequence.

To determine the threshold bandwidth sequence, we use the pre-specified probability that a randomly sampled gene set of size G_k passes the module constraint on the expression data ('box p-value', see Running parameters). We randomly select a large number of random gene sets of size G_k, and compute the sorted bandwidth sequence for each set. Subsequently, we determine the j'th position of the b'th candidate threshold bandwidth sequence as b'th largest elements among all j'th positions in those sorted bandwidth sequences. The consequence is that all candidate threshold bandwidth sequences are monotonically increasing, and different candidate threshold bandwidth sequences never intersect. Finally, the threshold bandwidth sequence b that achieves our pre-specified probability (or a smaller probability) is selected and used as threshold bandwidth sequence in the above procedure. Figure 7- 1 illustrates the threshold bandwidth selection process and the verification of constraint C_C on a hypothetical example. As the threshold bandwidth sequence is determined by randomization it has a statistical interpretation (p value).

Since DISTILLER is based on the CHARM algorithm, it reports only closed modules. These are modules that cannot be further extended by any other gene without reducing the number of motifs that all of its genes share, or the number of conditions in which its genes are expressed similarly. DISTILLER applies techniques used in CHARM to efficiently ignore non-closed modules while selecting all the closed ones. We report only closed modules with a number of genes larger than or equal to the 'gene content threshold' because only those modules containing a minimum number of genes are interesting.

7.2.2 Selection of interesting non-redundant modules

Despite the massive reduction in the number of modules achieved by using CHARM, the output may still be too large to explore. As no explicit score is assigned to the modules, it is not clear which modules are "most interesting" to analyze first. Also the output might contain partially redundant modules: noise in the data may cause modules to appear as a number of separate, partially overlapping modules, for instance differing from each other in a few conditions only. To address this problem we apply an iterative procedure that selects the most interesting modules one by one. It takes into account the significance of individual modules but penalizes at the same time overlap with modules that have already been reported. Thus, interesting modules are selected one by one depending on their statistical significance and on how well they help covering the complete solution space and thus do not overlap with modules that had already been selected.

We initialize the procedure by attaching an 'interest score' to every module. This interest score is obtained by multiplying the p-value of the module expression pattern with the p-value of the module motif content. For the motif data, the latter p-value is the probability that a module with at least the same number of genes and containing these motifs is found by chance. The corresponding null distribution assumes that each motif is present in all genes with the same probability. We estimate this probability from the data by taking the ratio of the number of genes with the motif at hand to the total number of genes. The p-value of the module motif content can then be computed using the cumulative binomial distribution. For the expression data, the p-value of the pattern is the probability that a module with at least the same number of genes and containing the same conditions is found by chance. The corresponding null distribution is specified as being invariant with respect to permutations of the gene expression values within the same condition. The corresponding p-value of the module expression pattern can then be computed by Monte Carlo sampling. The interest score is obtained by multiplying the p-values for both separate data sources.

The interest score allows ranking all modules according to decreasing significance. However, it is still possible that several mutually redundant (heavily overlapping) modules rank high. In the first step of the iterative procedure the most interesting (top ranking) module is selected. Then, in each of the following steps another module is selected in such a way that it adds as much information as possible to the modules that have already been already selected. To do this, we update the null distribution for the motifs each time a new module has been selected. This is done by changing the probability that a gene has a motif instance to one in the null model, for those entries in the motif matrix that are already covered by previously selected modules. The result is that the interest scores of modules that

heavily overlap with previously selected modules increase, such that they move down in the ranking. In every iteration, we select the most interesting module with respect to the corresponding adapted null distribution.

DISTILLER selects conditions in which the genes of a module are significantly co-expressed. However, in some of these conditions the expression ratios of the modules genes will be close to zero, indicating that their expression is not really varying (i.e. spurious conditions). Therefore as an additional post-processing step, conditions in a module for which the coefficient of variation (ratio of the gene expression standard deviation over the gene expression mean) exceeded a specified threshold were excluded from the modules.

7.2.3 Seed module extension

In a subsequent extension step we recruit additional candidate module genes that did not pass the stringent seed discovery step but should be considered part of the module (e.g. downstream operon genes that do not contain a motif instance in their promoter regions but are subject to its regulatory influence).

The relaxed criteria for adding additional genes to the module are the following: 1) the gene's expression profile should have a correlation with the module's mean expression profile of at least a fraction α of the module correlation (defined as the lowest correlation value between a seed gene's expression profile and the average expression profile for the modules conditions), and 2) the genes should have a motif instance with p-value below a threshold β. Both requirements have to be fulfilled unless a gene is part of an operon for which the first gene is present in the seed module. In this case only the first criterium has to be satisfied.

7.2.4 Running parameters

To test the influence of different parameters, we compared results obtained with different settings with known interactions from the RegulonDB network by calculating the recall and precision:

- Recall = $\frac{TP}{TP + FN}$
- Precision = $\frac{TP}{TP + FP}$

where TP = true positives, FN = false negatives, FP = false positives . In this case, $TP+FN$ equals the total number of interactions in RegulonDB, i.e. 736

interactions, while *TP+FP* is the total number of interactions reported by DISTILLER. The TP are confirmed interactions present in RegulonDB that were also identified by DISTILLER. Hence, recall and precision indicate the trade-off between the number of predicted and the percentage of reported interactions that corresponds to known links in the regulatory network. Our goal is to obtain high confident predictions, corresponding to parameter settings that yield a large product of recall and precision.

For each parameter setting the 200 best-scoring non-redundant modules were selected (see section 7.1.2). Subsequently the recall and precision were calculated for the best-scoring module, the two best-scoring modules, the three best-scoring modules, etc. All tested parameter settings are shown in Table 7- 1. We varied the minimum number of genes in a module (i.e. the gene content threshold), the minimum number of conditions in which the genes should be co-expressed, the p-value of the bandwidth threshold and the motif score threshold. The recall, precision and the product of the precision and recall for the different parameter settings are shown in Figure 7- 2.

In general, modules obtained with a motif threshold of 0.01 clearly yield a lower recall and precision than modules obtained with motif threshold 0.001. This means that with a lower motif threshold we recover less known RegulonDB interactions (at least, in the top 200 modules) and obtain many

Table 7- 1 Different parameter settings that were used to assess the parameter sensitivity. No. conditions: the minimum number of conditions in which a set of genes should be co-expressed; Bandwidth p-value: p-value to determine the threshold on the band width sequence; Motif threshold: the minimum p-value a motif instance should have; No. genes: the minimum number of genes in a module.

Test	no. conds	bandwidth p-value	motif threshold	no. genes
Parameter setting 1	50	0.00001	0,01	3
Parameter setting 2	50	0.00001	0,01	4
Parameter setting 3	50	0.0001	0,01	3
Parameter setting 4	870	0.00001	0,01	3
Parameter setting 5	870	0.00001	0,01	4
Parameter setting 6	870	0.0001	0,01	3
Parameter setting 7	870	0.00001	0,001	3
Parameter setting 8	870	0.00001	0,001	4
Parameter setting 9	870	0.0001	0,001	3
Parameter setting 10	870	0.0001	0,001	4
Parameter setting 11	50	0.00001	0,001	3
Parameter setting 12	50	0.00001	0,001	4
Parameter setting 13	50	0.0001	0,001	3
Parameter setting 14	50	0.0001	0,001	4

more predicted interactions. Since we want to both study the condition dependency of interactions and obtain reliable novel interactions, a motif threshold of 0.001 clearly is the better choice. Modules obtained with a bandwidth p-value of 0.0001 correspond to a larger recall and a slightly smaller precision than those obtained with a bandwidth p-value that equals 0.00001. From Figure 7- 2 it is clear that the product of recall and precision is clearly better for modules obtained with a lower p-value. These p-values are not as stringent as they may seem at first sight: in a set of 4000 genes, for instance, there are on the order of 1e10 possible sets of three.

Changing the gene content threshold from three to four genes has a positive influence on the recall, while the precision decreases. The product of precision and recall is however better for a gene content threshold of four genes. To verify that condition selection indeed has a beneficial effect, we also tested parameter settings where the minimum number of conditions in a module was set equal to all conditions (870). From Figure 7- 2 it is clear that although a rather high precision can be obtained with all conditions, the recall was low in those runs.

From the above analysis, it was clear that parameter setting 14 was most likely to yield biologically relevant predictions. To choose the number of modules for further biological analysis, we checked how the product of precision and recall changed over the number of modules (see Figure 7- 2). No large changes were observed in the last 50 modules, indicating increasing redundancy of the additionally selected modules. We therefore choose to keep 150 modules in our further analysis.

In summary, parameter settings were selected such that the seed module consists of at least four genes (i.e. gene content threshold) that share at least one motif and 50 conditions. The threshold for the p-value of the motif instances was set to 0.001. In order to choose the box threshold, 100000 randomizations (random sets of four genes) were carried out and the box p-value threshold was set to 0.0001. In the post-processing steps, a subset of 150 minimally redundant modules was selected, and the threshold for the coefficient of variation was set to 0.6. For the seed module extension step, correlation parameter α was set to 90% and the relaxed p-value threshold β to 0.05.

Figure 7- 2: Recall, precision and product of precision and recall for different parameter settings. For each parameter setting, recall, precision and product of precision and recall were calculated for a selection of up to 200 best-scoring non-redundant modules.Parameter setting 14 was selected for further use in the paper. Clearly, recall was best for this parameter setting (A). Precision of parameter setting 14 is better than those of parameter setting 1-6, but it is not optimal (B). The plot of the product of precision and recall allows to check for the trade-off between good precision and good recall (C). Parameter setting 14 outperforms the other settings and was therefore selected for futher use in the remainder of the paper.

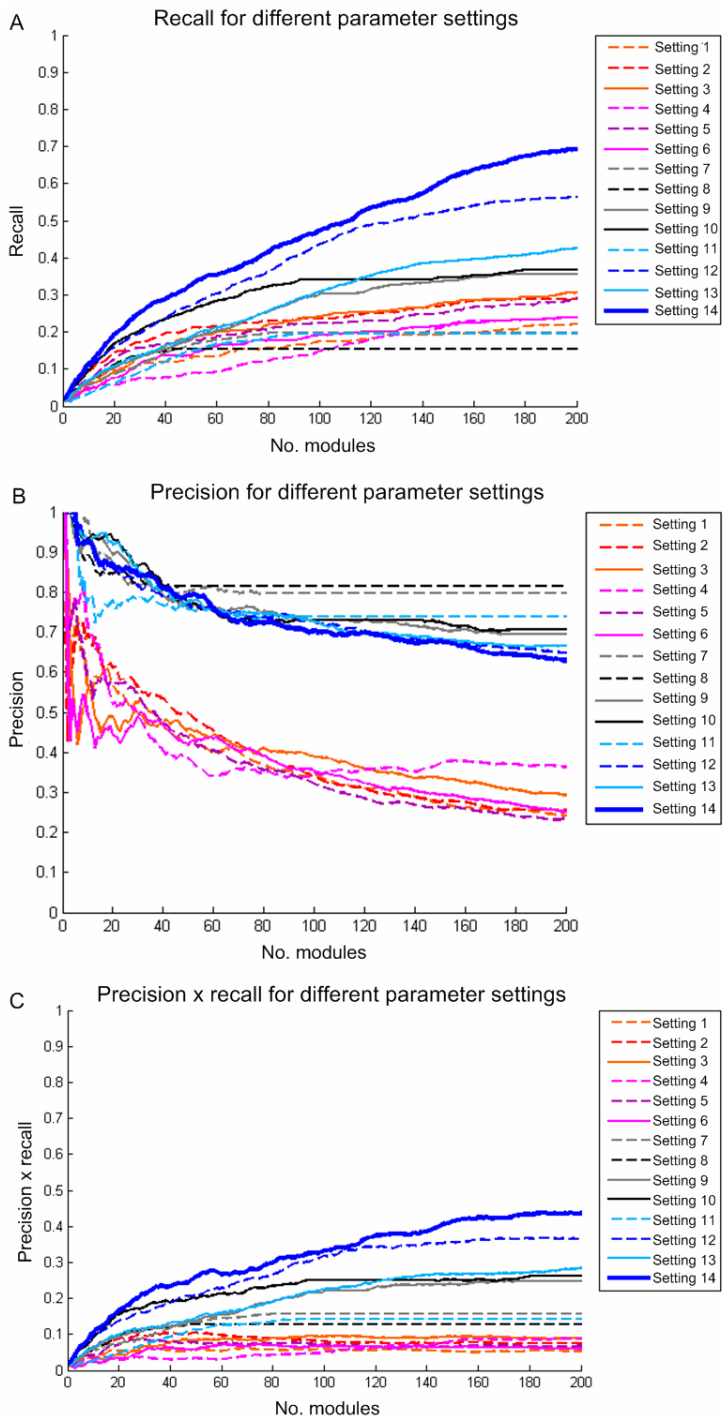


Figure 7- 2: caption on page 106

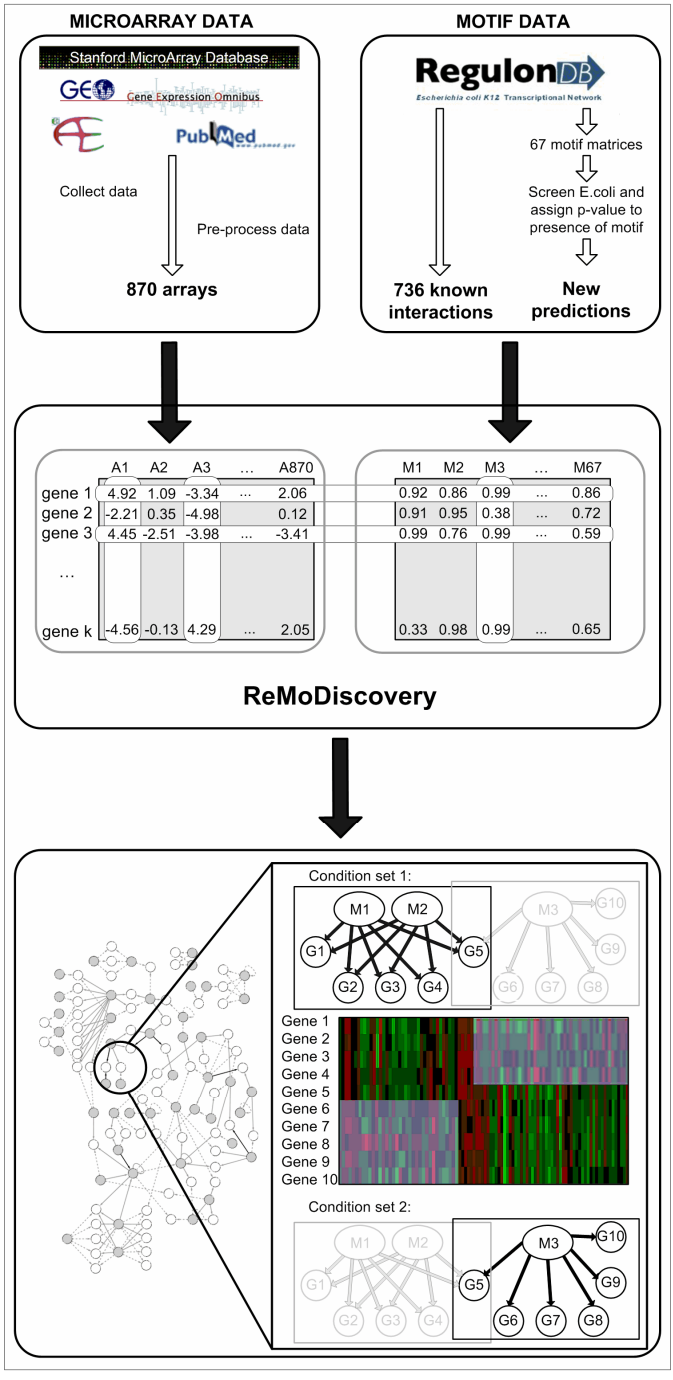


Figure 7- 3: caption on page 109

7.3 Application of DISTILLER on *Escherichia coli* data

DISTILLER (Data Integration System To Identify Links in Expression Regulation) is our data integration framework that searches for transcriptional modules by combining expression data with information on the direct interaction between a regulator and its corresponding target genes. The inferred modules contain a minimal number of genes (i.e. gene content threshold) that must be co-expressed in a sufficiently large number of conditions and share motif instances for the same regulator(s). In this study, we applied DISTILLER to simultaneously analyze two distinct *E. coli* data sources: a cross-platform expression compendium consisting of 870 microarrays and a regulatory motif compendium consisting of both predicted and experimentally verified motif instances (see chapter 3 for a description of the data sources). As such we gain insight into the modularity, combinatorial regulation and condition dependency of bacterial networks. In addition we considerably extended the current knowledge of the *E. coli* transcriptional network.

Figure 7- 3: Overview of analysis strategy for *E. coli*. *E. coli* microarray experiments were collected and combined in a cross-platform compendium of 870 arrays. We collected the binding site models (weight matrices) of 67 regulators from RegulonDB. These were used to screen *E. coli* upstream sequences in order to predict potential motif instances. Together with the known motif instance data from RegulonDB, these predictions resulted in the motif matrix. Based on this motif matrix and the microarray compendium, DISTILLER identifies modules consisting of gene sets that interact with the same regulators and that are co-expressed in a subset of conditions of the compendium. Modules can overlap in regulator, gene and/or condition content. This allows identifying condition-dependent combinatorial regulation. In the example on the figure, module 1 consists of gene 1, 2, 3, 4 and 5. These genes have two motifs (M1 and M2, corresponding to two regulators) in common and they are co-expressed in condition set 1. Module 2 consists of gene 5, 6, 7, 8, 9 and 10, co-expressed in condition set 2 and sharing one motif (M3). Note that gene 5 is present in both modules, regulated by the regulators binding to M1 and M2 and M3, but under a different set of conditions.

7.3.1 Identifying regulatory modules

By integrating motif data and a large scale expression compendium, two complementary information sources, DISTILLER detects condition-dependent regulatory modules (Figure 7- 3). From each module, regulator-target interactions that are functionally active under the experimental conditions included in the module, can be extracted by linking each gene in the module with the regulator(s) corresponding to the shared motif instance(s). The 150 statistically most significant modules recovered by DISTILLER represent a total of 732 interactions. Of these, 454 interactions correspond to 62% of 736 interactions for 67 regulators with known binding sites described in RegulonDB (Salgado *et al*, 2006). Most modules are enriched for functions in which the regulator was known to be involved. For 37 of the 67 regulators at least part of their regulon could be confirmed. For the remaining 30 regulators no interaction was found; most likely either the number of genes in the corresponding modules falls below the gene content threshold, or the conditions needed to trigger these interactions are not present in our compendium (e.g. MelR, triggered by melibiose).

In addition to identifying 454 previously described interactions, we predict 278 novel interactions that have not yet been documented in RegulonDB (Table 7- 2 and Table 7- 3). For many well studied regulators, the known part of their regulon could be considerably extended. As for most of the newly predicted interactions no additional confirmation existed in literature, we assigned them a level of confidence based on the gene composition of the module the target was retrieved in. If the module contained many previously confirmed targets, tightly co-expressed with the novel target, we attached larger confidence to its prediction(s). Many of the new predictions correspond to interactions of a known regulator with hypothetical genes, expanding the knowledge on regulons of well studied regulators. Novel interactions from modules that contain at least 50% known interactions can be assumed very reliable (Category I: 157 interactions). Predicted interactions from modules that contain at least one previously known interaction form the second most reliable group (Category II: 118 interactions). Finally, novel interactions from modules that contain no previously known interaction (Category III: 20 interactions) are more interesting but can no further be verified based on existing knowledge.

Table 7- 2 gives an overview of the number of known interactions per regulator that were identified, as well as the number of new, predicted interactions. The large number of predictions for FNR, CRP, ArcA, Fis, and IHF confirms their role as global hubs in the network. For some regulators only the previously described interactions, and/or a few additional ones could be retrieved (e.g. CueR, GlpR).

Table 7- 2: Interactions recovered by DISTILLER. The number of known and novel interactions for each regulator that were identified by applying DISTILLER is shown. The large number of predictions for FNR, ArcA, CRP … confirms their role as global hubs in the network. For some regulators, only few additional interactions were identified. Regulators for which no interaction could be found by DISTILLER may point to conditions that have not yet been tested. Regulator: a list of all regulators in RegulonDB for which the binding site is known; Modules: the number of modules that were found to be dependent on a specific regulator; Overlap RegulonDB: the number of experimentally confirmed interactions described in RegulonDB that could be detected by DISTILLER; Total new: the number of new interactions not previously described in RegulonDB that were detected by DISTILLER; New cat I: new interactions originating from modules that contain at least 50% known interactions; New cat II: new interactions originating from modules that contain at least 1 known interaction; New cat III: new interactions originating from modules that contain no known interactions. For 37 of the 67 regulators at least part of their regulon could be confirmed. For the remaining 30 regulators no interaction was found; most likely either the number of genes in the corresponding modules falls below the gene content threshold, or the conditions needed to trigger these interactions are not present in our compendium (e.g. MelR, triggered by melibiose).

Regulator	Modules	Overlap RegulonDB	Total new	New cat I	New cat II	New cat III
Ada	0	0 out of 3	0	0	0	0
AgaR	0	0 out of 3	0	0	0	0
AraC	0	0 out of 5	0	0	0	0
ArcA	13	28 out of 35	23	13	12	0
ArgP	0	0 out of 2	0	0	0	0
ArgR	2	7 out of 9	5	2	3	0
CRP	38	111 out of 153	63	49	24	0
CaiF	0	0 out of 2	0	0	0	0
CpxR	3	8 out of 17	5	5	0	0
CueR	1	3 out of 3	1	1	0	0
CysB	2	4 out of 6	6	2	0	4
CytR	1	3 out of 7	1	1	0	0
DcuR	0	0 out of 3	0	0	0	0
DeoR	1	1 out of 2	3	0	3	0
DnaA	0	0 out of 6	0	0	0	0
FNR	19	48 out of 57	25	7	18	0
FadR	1	4 out of 7	0	0	0	0
FhlA	0	0 out of 4	0	0	0	0
Fis	10	21 out of 25	20	12	8	0
FlhDC	2	9 out of 10	5	5	0	0
FruR	4	8 out of 12	7	2	6	0
Fur	7	23 out of 29	15	13	3	0
GadE	1	2 out of 3	2	2	0	0
GalR	0	0 out of 4	0	0	0	0

GalS	0	0 out of 4	0	0	0	0
GcvA	0	0 out of 2	0	0	0	0
GlpR	1	4 out of 4	0	0	0	0
GntR	0	0 out of 6	0	0	0	0
H-NS	0	0 out of 6	0	0	0	0
IHF	15	39 out of 51	18	10	8	0
IclR	0	0 out of 2	0	0	0	0
IdnR	0	0 out of 3	0	0	0	0
IscR	1	4 out of 7	0	0	0	0
LacI	0	0 out of 1	0	0	0	0
LexA	6	11 out of 14	13	3	6	4
Lrp	5	13 out of 18	5	3	3	0
MalT	1	4 out of 5	0	0	0	0
MarA	2	7 out of 15	1	1	0	0
MelR	0	0 out of 2	0	0	0	0
MetJ	2	7 out of 8	0	0	0	0
MetR	0	0 out of 4	0	0	0	0
Mlc	5	5 out of 5	11	3	9	0
ModE	0	0 out of 6	0	0	0	0
Nac	0	0 out of 5	0	0	0	0
NagC	5	5 out of 6	15	0	3	12
NanR	0	0 out of 2	0	0	0	0
NarL	7	19 out of 20	0	0	0	0
NarP	1	4 out of 7	0	0	0	0
NorR	0	0 out of 1	0	0	0	0
NsrR	1	3 out of 4	1	1	0	0
NtrC	5	7 out of 7	8	3	5	0
OmpR	1	2 out of 8	2	2	0	0
OxyR	1	2 out of 8	2	2	0	0
PhoB	1	1 out of 10	4	0	4	0
PhoP	3	9 out of 19	1	1	0	0
PspF	0	0 out of 2	0	0	0	0
PurR	6	14 out of 15	11	8	3	0
RcsB	1	2 out of 7	2	2	0	0
RhaS	0	0 out of 2	0	0	0	0
Rob	0	0 out of 5	0	0	0	0
SoxS	2	8 out of 15	0	0	0	0
TorR	0	0 out of 3	0	0	0	0
TrpR	0	0 out of 5	0	0	0	0
TyrR	2	4 out of 8	3	3	0	0
UlaR	0	0 out of 2	0	0	0	0
UxuR	0	0 out of 3	0	0	0	0
XylR	0	0 out of 2	0	0	0	0
			278	157	118	20

Table 7- 3: The complete list of novel interactions for each regulator. As mentioned in Table 7- 2, many of these novel interactions originate from modules that contain at least 50% known interactions. We can have most confidence in these interactions. For FNR, for instance, four of our predictions were already confirmed in a recent ChIP-chip analysis (Grainger *et al*, 2007), indicated in italic. We experimentally validated 11 additional novel targets for FNR (see section 7.3.2). These interactions are indicated in bold.

Regulator	Regulated genes
ArcA	b0007 (yaaJ), b0155 (clcA), b0280 (yagN), b0318 (yahD), b0598 (cstA), b0854 (potF), b1132 (hflD), b1388 (paaA), b1853 (yebK), b2070 (yegI), b2092 (gatC), b2093 (gatB), b2096 (gatY), b2193 (narP), b2380 (ypdA), b2518 (ndk), b2844 (yqeF), b3025 (qseB), b3128 (garD), b3237 (argR), b3739 (atpI), b4328 (iadA), b4376 (osmY)
ArgR	b0295 (ykgL), b0674 (asnB), b0860 (artJ), b2669 (stpA), b2818 (argA),
CRP	b0162 (cdaR), b0268 (yagE), b0459 (maa), b0486 (ybaT), b0592 (fepB), b0822 (ybiV), b0880 (cspD), b0898 (ycaD), b0953 (rmf), b1002 (agp), b1138 (ymfE), b1139 (lit), b1188 (ycgB), b1241 (adhE), b1255 (yciC), b1397 (paaJ), b1440 (ydcS), b1492 (gadC), b1498 (ydeN), b1513 (lsrA), b1685 (ydiH), b1704 (aroH), b1778 (yeaA), b1814 (sdaA), b1976 (yeeI), b2097 (fbaB), b2146 (yeiT), b2155 (cirA), b2536 (hcaT), b2597 (yfiA), b2702 (srlA), b2715 (ascF), b2775 (yqcE), b2800 (fucA), b2844 (yqeF), b2869 (ygeV), b2870 (ygeW), b3072 (aer), b3073 (ygjG), b3093 (exuT), b3238 (yhcN), b3239 (yhcO), b3253 (yhdH), b3403 (pck), b3453 (ugpB), b3510 (hdeA), b3683 (glvC), b3830 (ysgA), b3831 (udp), b3882 (yihU), b3909 (kdgT), b3947 (ptsA), b4142 (groS), b4189 (yjfO), b4208 (cycA), b4216 (ytfJ), b4227 (ytfQ), b4238 (nrdD), b4299 (yjhI), b4354 (yjiY), b4355 (tsr), b4380 (yjjI), b4383 (deoB)
CpxR	b0970 (yccA), b3913 (), b3914 (), b4143 (groL), b4219 (msrA)
CueR	b0209 (yafD)
CysB	b0058 (rluA), b0427 (yajR), b1287 (yciW), b1627 (rsxA), b2375 (yfdX), b4311 (yjhA)
CytR	b3831 (udp)
DeoR	b0683 (fur), b4380 (yjjI), b4383 (deoB)
FNR	b0643 (ybeL), **b0650 (hscC)**, b1002 (agp), b1003 (yccJ), b1004 (wrbA), *b1256 (ompW),* **b1358 (ydaT)**, b1376 (uspF), *b1541 (ydfZ)*, *b1593 (ynfK)*, **b1674 (ydhY)**, **b1750 (ydjX)**, b1901 (araF), **b1902 (yecI)**, **b2111 (yehD)**, **b2504 (yfgG)**, *b2580 (ung),* b2729 (hypD), **b3518 (yhjA)**, **b3519 (treF)**, b4278 (insG), **b4279 (yjhB)**, b4326 (yjiD), b4380 (yjjI), **b4403 (yjtD)**
Fis	b0469 (apt), b0779 (uvrB), b0835 (yliG), b1185 (dsbB), b1344 (ydaO), b1779 (gapA), b1876 (argS), b2296 (ackA), b2398 (yfeC), b2549 (yphG), b2550 (yphH), b2821 (ptr), b2927 (epd), b3000 (yghY), b3072 (aer), b3260 (dusB), b3813 (uvrD), b4108 (phnA), b4142 (groS), b4355 (tsr)
FlhDC	b1078 (flgG), b1926 (fliT), b1943 (fliK), b1945 (fliM), b2625 (yfjI)

FruR	b2096 (gatY), b2169 (fruB), b2340 (sixA), b2597 (yfiA), b3299 (rpmJ), b3357 (crp), b3956 (ppc)
Fur	b0150 (fhuA), b0468 (ybaN), b1018 (ycdO), b1374 (pinR), b1451 (yncD), b1452 (yncE), b1705 (ydiE), b2211 (yojI), b2392 (mntH), b2393 (nupC), b2675 (nrdE), b3008 (metC), b3070 (yqjH), b3337 (bfd), b3440 (yhhX)
GadE	b3490 (yhiL), b4326 (yjiD)
IHF	b0098 (secA), b0280 (yagN), b0432 (cyoA), b0565 (ompT), b0754 (aroG), b0992 (yccM), b1338 (abgA), b1344 (ydaO), b1908 (yecA), b2508 (guaB), b3320 (rplC), b3554 (yiaF), b3650 (spoT), b3734 (atpA), b3736 (atpF), b3784 (rfe), b4138 (dcuA), b4332 (yjiJ)
LexA	b0019 (nhaA), b0226 (dinJ), b0227 (yafL), b0231 (dinB), b0685 (ybfE), b1061 (dinI), b1848 (yebG), b2009 (sbmC), b3391 (hofQ), b3501 (arsR), b3645 (dinD), b3673 (emrD), b3832 (rmuC)
Lrp	b0891 (lolA), b0922 (mukF), b0966 (yccV), b1660 (ydhC), b3617 (kbl)
MarA	b4107 (phnB)
Mlc	b0426 (yajQ), b0678 (nagB), b0796 (ybiH), b1387 (maoC), b1498 (ydeN), b1617 (uidA), b3147 (yraM), b3417 (malP), b3862 (yihG), b3884 (yihW), b3900 (frvA)
NagC	b0191 (yaeJ), b0394 (mak), b0681 (ybfM), b1266 (yciV), b1654 (ydhD), b2112 (yehE), b2443 (yffL), b2774 (ygcW), b2846 (yqeH), b2987 (pitB), b3427 (yzgL), b3688 (yidQ), b3717 (cbrC), b3746 (yieN), b4311 (yjhA)
NsrR	b1426 (ydcH)
NtrC	b0470 (dnaX), b1012 (ycdM), b1823 (cspC), b1824 (yobF), b2309 (hisJ), b3871 (bipA), b3973 (birA), b4146 (yjeK)
OmpR	b0598 (cstA), b0811 (glnH)
OxyR	b1688 (ydiK), b4387 (ytjB)
PhoB	b2720 (hycF), b2884 (), b3211 (yhcC), b4464 (ygfQ)
PhoP	b1505 (ydeT)
PurR	b0336 (codB), b0411 (tsx), b1039 (csgE), b1849 (purT), b2313 (cvpA), b2498 (upp), b2682 (ygaZ), b2888 (ygfU), b3654 (yicE), b3714 (yieG), b3823 (rhtC)
RcsB	b0419 (yajO), b4377 (yjjU)
TyrR	b0406 (tgt), b1842 (holE), b1894 (insA-5)

7.3.2 Experimental validation of predicted interaction for FNR

To demonstrate the reliability of our approach, we used ChIP-qPCR (Thijs *et al*, 2007) for the *in vivo* validation of predicted interactions for FNR, one of the most extensively studied regulators in *E. coli*. DISTILLER recovered 48 of the 57 FNR targets described in RegulonDB, and predicted 25 novel FNR targets, four of which (*ung*, *ompW*, *ydfZ* and *ynfK*) were confirmed by a recent ChIP-chip analysis (Grainger *et al*, 2007). We tested 11 additional targets that were selected based on their difference in prediction confidence.

Therefore, *E. coli* MG1655 (Blattner *et al*, 1997) and JCB1101, a MG1655 encoding *fnr*3XFLAG (Constantinidou *et al*, 2006) cells were grown under aerobic (250 ml conical flasks containing 50 ml Luria-Bertani (LB) medium supplemented with 0.4% glucose, shaken at 200 r.p.m.) or microaerobic (250 ml conical flasks containing 150 ml LB medium shaken at 100 r.p.m. conditions at 37°C until an OD_{595} of 0.4 was reached. Chromatin immunoprecipitation was performed as described in Thijs *et al* (2007), except that the tagged FNR was immunoprecipitated with anti-FLAG M2 antibodies (Sigma, F1804). Immunoprecipitation was performed for six biological replicates of both MG1655 wild-type (referred to as 'mock') and the *fnr*3XFLAG strain (referred to as 'IP').

Physical interaction between FNR and the selected targets genes was determined by quantitative real time PCR (qRT-PCR). qRT-PCR was performed on a StepOnePlus instrument (Applied Biosystems) in a 20 µl reaction mixture containing 1 µl of the IP- or mock-DNA preparation, 900 nM of each primer and 10 µl Power SYBR Green kit (Applied Biosystems). A standard qRT-PCR run consisted of a holding stage to activate enzyme for 10 minutes at 95°C, followed by 40 cycles of 15 seconds at 95°C and 1 minute at 60°C.

Primers for qRT-PCR, designed with PrimerExpress (Applied Biosystems) are displayed in Table 7- 4. The well known FNR targets *ndh* and *pdhR* (Grainger *et al*, 2007; Green and Guest, 1994; Quail *et al*, 1994) were chosen as positive controls. Negative controls consisted of genes for which no evidence of FNR-mediated transcription regulation exists. Primers were targeted against the coding regions of these genes in order to avoid overlap with potential FNR binding sites in the intergenic regions. For the predicted targets, the intergenic sequence surrounding the predicted FNR motif was amplified.

To evaluate the enrichment of DNA sequences bound to FNR, the initial concentrations (N_0) were compared between IP and mock for both aerobic and microaerobic growth conditions. For each amplification reaction N_0 was

Table 7- 4: The set of genes that were tested in the ChIP experiment and their forward and reverse primers are shown. Among the genes are predicted FNR targets, negative controls and positive controls. The number of known FNR targets in each module is indicated (Module FNR targets).

Genes	Description	Module FNR targets	forward primer	reverse primer
ydhY (b1674)	Module 80	2	GCAATGAAAGTCACAAATAATTGTCG	TCTGAGCCCTTAACATTGATCGTT
yfgG (b2504)	Module 80	2	ATTCAGACAGGCATCCCACCT	AAAGCGTCAGTGGATACATATTTTAATG
hscC (b0650)	Module 83	3	CATTGATTAAGATCACCACGCG	CCATTGCATTTATATCTTGAAAGAAAA
yjhB (b4279)	Module 84	1	TGGAGGAAAAGATTGTACTGATTAGGT	GGAGCCTCAACAATATCCAGAAA
ydjX (b1750)	Module 84	1	GCTCAAACTGGGTCAGGAGAAT	GGCAGGCAACAAATCATTACTG
yjtD (b4403)	Module 84	1	GGCCGGACACCCAATAAAA	TTTCATCAAACTGTTAACGTGCTACA
ydaT (b1358)	Module 84	1	TCCTCATTTGATCATACCTGAAACA	GATGTGCTCATGCTTGATTTTCAT
treF (b3519)	Module 86	3	CAGGGAGATACGCTTTTTGATAGG	TGCCATTCCTGATCTGGGTAA
yehD (b2111)	Module 87	1	GCATTTAGAAAGCCGAAATCATTTATA	TCTCACATCAAAAAAGTTGCCG
yhjA (b3518)	Module 87	1	TGCCATTCCTGATCTGGGTAA	CAGGGAGATACGCTTTTTGATAGG
ftnB (b1902)	Module 87	1	ACGTATAATGAGAGCCATCTCGC	CGGGTATAAGAAAATGCTTAAATCATC
metC (b3008)	Neg. control		GGCTATTTTCCTTTGTGCTTA	CCCACGAGTAGGCCATGCTGAA
rfaH (b3842)	Neg. control		TCAGCCACTTCGTGCGCTT	TGCCGGATCGACAATGTCTT
yafZ (b0252)	Neg. control		CCCTGCTCGACAGCCTACA	CGGGTCTGACAGGCAAAGAA
kefB (b3350)	Neg. control		TCAGGTGATTGGTCGTTTGC	CCGCTCCAGCACGGTAAT
yfcI (b2305)	Neg. control		CGGCAATGCAAAACCATCTT	TCGGGAGCACCAATGGAA
secD (b0408)	Neg. control		GCGCGTGCGAAAGAGATT	GGTGTTTACCAGACGGAATTCC
endA (b2945)	Neg. control		AAGGCCAGTACGGTCAATGC	TGGTTCGGCAGCTTTTTCTT
php (b3379)	Neg. control		TGGTGGCCTGTACCGGTTAT	GCGGGTCGCCACATGT
yeiM (b2164)	Neg. control		TTGGCGGCTGGTTTGGT	CAGCACATAGCCAAAAATACTTTCC
ybhC (b0772)	Neg. control		ACGATAGCTGCCAGAGCAAAC	AGAAGACCGCAGAGCAGAGAA
tolA (b0739)	Neg. control		TCGACGCTGTCATGGTTGAT	GCTTTGCATGCGTTTGTACTG
ndh (b1109)	Pos. control		CCCGGTAAAGTCGCCTATCTT	GAATGGTGCATTGGTCATACCA
pdhR (b0113)	Pos. control		ATGTGCACAGTTTCATGATTTCAA	TGTGACGCTAAAGTAACAAAGTATTCAC

estimated by performing a linear regression on the logarithm of the fluorescence intensities per cycle number (Ramakers *et al*, 2003). This linear regression was fitted on 4 subsequent measurements, which were selected as the ones showing the highest slope (i.e. 4 measurements in the 'window-of-linearity'). To account for sample variation, N_0 values were rescaled by a sample-specific factor that renders the mean N_0 of 11 negative controls equal for all biological replicates of both aerobic and microaerobic growth conditions. To asses the validity of FNR binding to a specific target, *p*-values were calculated by performing a *t*-test on the mean logratios of IP over mock for both growth conditions. An interaction was deemed significant if its *p*-value was below 0.05.

All 11 predictions, consisting of four high confidence predictions (*ydhY*, *yfgG*, *hscC* and *treF*) and seven medium confidence predictions (*yjhB*, *ydjX*, *yjtD*, *ydaT*, *yehD*, *yhjA* and *ftnB*), were all shown to bind FNR in vivo (see

Table 7- 5). In the course of this study, two of these validated FNR targets *yhjA* and *ydh* have also been confirmed by two independent experimental studies (Partridge *et al*, 2007; Partridge *et al*, 2008)

Table 7- 5: The p-values and logratios (IP versus mock) that were obtained for each gene in both aerobic and micro-aerobic conditions are shown. An interaction was deemed significant if its p-value was below 0.05.

gene	description	p-value		logratio	
		Aer.	LB-Micro-aer.	Aer.	LB-Micro-aer.
ydhY (b1674)	Module 80	0,0212	0,0079	0,821	0,7399
yfgG (b2504)	Module 80	0,0137	0,0102	0,8896	0,6498
hscC (b0650)	Module 83	0,0611	0,0185	0,5999	0,6292
ydaT (b1358)	Module 84	0,0359	0,0254	0,7421	0,6873
ydjX (b1750)	Module 84	0,0127	0,2657	0,7355	0,204
yjhB (b4279)	Module 84	0,0012	0,0115	0,9176	0,5977
yjtD (b4403)	Module 84	0,1095	0,0168	0,2186	0,4854
treF (b3519)	Module 86	0,01	0,0484	0,4816	0,4141
ftnB (b1902)	Module 87	0,0261	0,1399	0,3379	0,1721
yehD (b2111)	Module 87	0,0044	0,4933	0,7689	0,0055
yhjA (b3518)	Module 87	0,0092	0,0408	0,7195	0,5711
endA (b2945)	Neg. control	0,3171	0,1548	0,068	0,0993
kefB (b3350)	Neg. control	0,7461	0,6144	-0,0417	-0,0092
metC (b3008)	Neg. control	0,0695	0,4546	0,1949	0,0171
php (b3379)	Neg. control	0,9328	0,3735	-0,2913	0,0162
rfaH (b3842)	Neg. control	0,7671	0,6362	-0,2136	-0,067
secD (b0408)	Neg. control	0,7727	0,6867	-0,1616	-0,0884
tolA (b0739)	Neg. control	0,1174	0,0765	0,107	0,0739
yafZ (b0252)	Neg. control	0,8678	0,7786	-0,0445	-0,0214
ybhC (b0772)	Neg. control	0,223	0,6324	0,0581	-0,008
yeiM (b2164)	Neg. control	0,7793	0,5028	-0,0437	-0,0002
yfcI (b2305)	Neg. control	0,2266	0,4794	0,0771	0,0038
ndh (b1109)	Pos. control	0,0824	0,0001	0,5162	0,6667
pdhR (b0113)	Pos. control	0,0062	0,0616	0,34	0,3841

7.3.3 Conditional dependency of the regulatory network

Arrays were grouped into conditional categories depending on the major cue that was changed in the experiments. For instance, the category aerobic-anaerobic groups all arrays in which the effect of changing the oxygen level on gene expression was measured. Figure 7- 4 shows to what extent the conditions of the modules of a particular regulator are enriched for a specific category. Enrichment of a conditional category implies that the target genes of a particular regulator are mainly co-expressed in conditions belonging to the enriched category; this indirectly gives information on the conditions where a particular regulator is active. Most regulators were found to be active in conditions that are in agreement with their annotation, illustrating the effectiveness of our novel condition selection (bicluster) strategy.

Environmental conditions that trigger major changes in the energy status of a cell seem to have the most pronounced effect on transcriptional regulation: changes in oxygen concentration, diauxic shift, pH and carbon source trigger a whole range of transcriptional regulators that mediate the transition to a novel metabolic state. Other conditions seem to trigger very specific pathways. Changes in the Fe^{2+} concentration or application of DNA damage, for instance, respectively induce the Fur and LexA pathways.

The role of global regulators such as ArcA, CRP, Fis, FNR, IHF, or Lrp that tune the overall cellular response towards the simultaneous interplay of energy, carbon source and amino acid availability is clearly visible. The more conditional categories a regulator is involved in, the more global its role. For instance, CRP and the nucleoid associated proteins Fis and IHF are the most pleiotropic regulators (Hengge-Aronis, 1999), but also FruR, Fur, and LexA seem to have a considerable impact on gene expression in a variety of conditions. In contrast to the global regulators, more specific regulators are important for fine-tuning the response. A more detailed description follows next.

7.3.3.1 Oxygen responsive regulators

The well-known regulators responsive to oxygen, redox levels and alternative electron acceptors (FNR, ArcA, NarL and NarP) (Unden and Bongaerts, 1997) are among the most overrepresented in the conditional category aerobic-anaerobic. To some extent their activity is also visible during diauxic shift. This is to be expected as changes in the expression levels of genes of the respiratory chain should be tuned towards the reducing power generated during carbon catabolism switches.

7.3.3.2 Regulators of the carbon metabolism

The regulators of carbon metabolism (CRP, MalT, Mlc, NagC, FruR, GlpR) are active in different conditional categories. The global regulator CRP has the most pronounced pleiotropic behavior and is active in the categories aerobic-anaerobic, diauxic shift, carbon source and amino acids. FruR and Mlc (and its paralog NagC (Plumbridge, 2001)), regulators of respectively glyconeogenesis and glycolysis, seem mainly active in diauxic shift conditions. FruR activity also depends on changes in the oxygen concentration and to a lower extent on amino acid availability, indicating that FruR together with CRP tunes the level of carbon source utilization towards oxygen and amino acid availability. Although they are paralogs, Mlc is mainly active during diauxic shift, while NagC modules are also linked to amino acids conditions. This is to be expected as Mlc is mainly

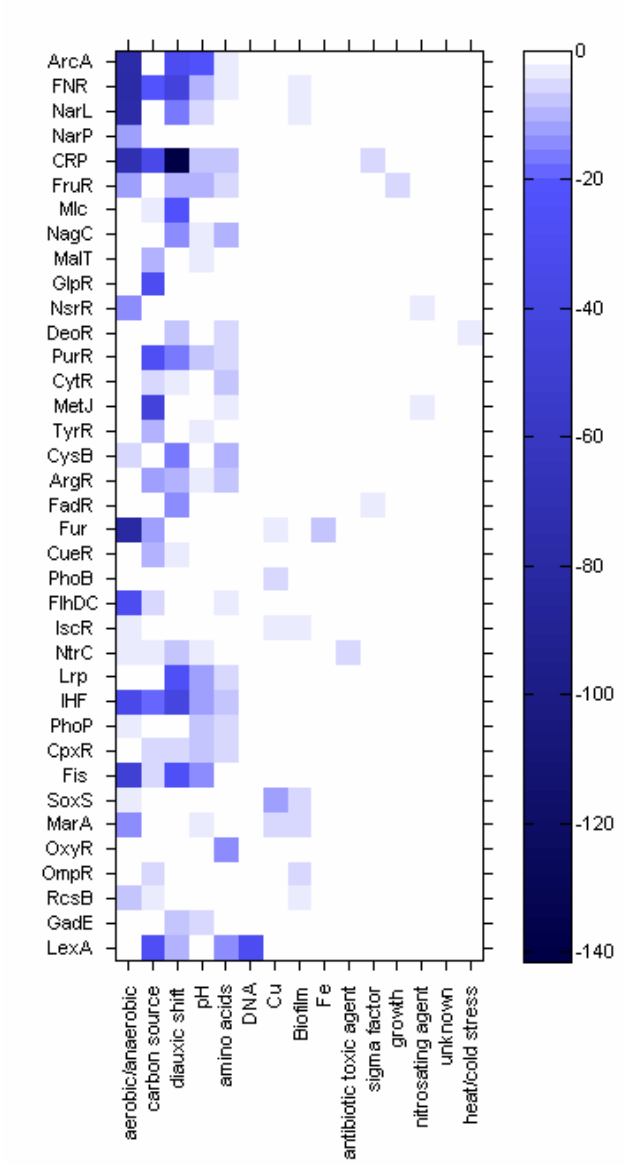


Figure 7- 4: Condition dependency of the regulatory modules. Columns: Conditional categories. Rows: regulators for which modules were detected by DISTILLER. Each entry indicates to what extent the conditions of the modules of a particular regulator are enriched (log p-value) for a specific category: dark blue entries correspond to the most significant enrichments. Most regulators were found to be active in conditions that are in agreement with their annotation, illustrating the effectiveness of our condition selection (bicluster) strategy. LexA, for instance, is a regulator involved in the process for repairing damaged microbial DNA and is the only regulator with a clear activity during conditions that induce DNA damage.

sensitive to glucose levels while NagC repressed genes are derepressed in the presence of amino sugars (N-acetylglucosamine (GlcNAc) (Plumbridge, 2001). Modules of the more specific regulators MalT and GlpR, involved respectively in the regulation of maltose and glycerol catabolism, are mainly expressed in the conditional category carbon source, upon addition of their respective carbon sources to the medium.

7.3.3.3 Regulators of nucleoside and amino acid biosynthesis

The conditions of modules involved in synthesis of novel molecular building blocks (nucleotides (DeoR, PurR, CytR), amino acids (CysB, Lrp, ArgR) and fatty acids (FadR)) are all enriched for the diauxic shift category. This is not surprising as most diauxic experiments assess growth recovery after starvation: the growth renewal requires the synthesis of novel building blocks. The conditional categories where regulators of purine and pyrimidine metabolism (DeoR, PurR, and CytR) are active suggest partial complementary: DeoR is mainly active during diauxic shift and amino acid starvation, CytR during amino acid starvation and PurR during diauxic shift and alterations of the carbon source. The condition sets of the modules involved in the biosynthesis of methionine, tyrosine and arginine (MetJ, TyrR and ArgR modules) are overrepresented in the carbon source category. Modules regulated by the nitrogen sensor NtrBC contain mainly conditions related to diauxic shift, and amino acid and carbon source alterations, confirming the role of the two-component system NtrBC in tuning the nitrogen availability to the carbon source and energy availability (Mao *et al*, 2007).

7.3.3.4 Regulators sensitive towards Iron/Cu

The role of Fur is very pronounced during iron related conditions but also during conditions which alter the oxygen level (aerobic-anaerobic). Indeed, Fe^{2+} is known to be an essential component of the prosthetic groups in enzymes involved in the electron transport chain and in protection against oxidative stress. The synthesis of these proteins not only depends on the oxygen concentration but also needs to be tuned towards the iron availability (McHugh *et al*, 2003). CueR, the regulator of the copper efflux system copA, is induced upon elevated copper levels (Outten *et al*, 2000), but does not seem active during comparable conditions in our compendium.

7.3.3.5 Nucleoid associated regulators

By acting both as a specific transcriptional regulator and nucleoid-associated protein, Lrp functions as a global regulator that is responsive to medium quality changes (starvation, osmolarity, …) and involved in the biosynthesis and transport of molecules (amino acids, rRNA) (Hung *et al*, 2002; Peterson *et al*, 2007; Pul *et al*, 2007). According to our results its modules are mainly

enriched for diauxic shift and pH related stress conditions. Similarly, modules of the nucleoid associated proteins Fis and IHF are active in a whole range of (stress) related conditions (carbon source starvation, pH dependent and O_2 related stress conditions) (Hengge-Aronis, 1999).

7.3.3.6 Stress responsive regulators

The other stress related regulators seem to act in more specific conditions. LexA, for instance, is apparently the only regulator with a clear activity during conditions that induce DNA damage (Brent and Ptashne, 1981; Little *et al*, 1981). Starvation conditions (carbon source and amino acid starvation) also induce the LexA dependent pathways.

The modules of both the PhoPQ two-component system, a pleiotropic regulator known to be induced by low pH (Groisman, 2001; Monsieurs *et al*, 2005), and CpxR, activated by envelope stress which results in protein misfolding (such as alkaline conditions) (Raivio, 2005), are mainly enriched in pH induced stress conditions.

The SoxS (regulator of the superoxide response regulon) and MarA regulators (multiple antibiotic resistance) seem to act in concert in the detoxification of externally added copper, during growth on biofilm and upon changes in the cellular oxygen level. The OxyR regulator, mainly active during oxygen changes (Liu *et al*, 2005), may activate the Fur regulon and protect the cell by scavenging iron (expression of dps, a stationary phase nucleoid protein that sequesters iron and protects DNA from damage).

7.3.3.7 Biofilm related conditions

OmpR, a known major regulator of membrane remodeling during growth on biofilms (Prigent-Combaret *et al*, 2001), and RscB, also a known regulator of growth during biofilm formation (Huang *et al*, 2006), are indeed overrepresented in biofilm conditions but may also play a role in alterations of the carbon source (both OmpR and RscB) or in oxygen changes i.e. aerobic-anaerobic (only RscB). GadE, known to be involved in the resistance to low pH (Mates *et al*, 2007), is mainly activated upon diauxic shift but also upon changes in the pH of the medium. Although CpxR has recently been described as a biofilm related regulator (Dorel *et al*, 2006), it does not seem to be overrepresented in the biofilm related conditions present in our compendium.

7.3.4 Combinatorial regulation

DISTILLER also identifies the level of combinatorial regulation of the target genes within each module. With combinatorial control we refer here to the

fact that a set of genes is regulated by at least two different regulators, irrespective of whether these regulators effectively undergo complex interactions or act independently from each other. According to RegulonDB, 42 transcription units (operons) are regulated by one regulator, 66 by two regulators and 70 by three or more regulators (with a maximum in-degree of eight regulators for a single transcription unit). Our inferred modules do not seem to exhibit the same amount of regulatory complexity: in our dataset, only 25 modules out of 150 were found to be regulated by at least two regulators and the maximum level of multiple regulation at the module level was restricted to three regulators. Out of 25 modules regulated by at least two regulators, 24 modules involve at least one global regulator such as CRP, FNR or ArcA, confirming the role of global regulators as hubs in the "co-regulatory network" (Balaji *et al*, 2007). To test whether this low number of modules that are regulated by multiple regulators is not only due to the fact that the number of complex regulons annotated in RegulonDB is lower than the number of simple regulons, we calculated the number of complex regulons containing at least four genes (four operons) in RegulonDB: 283 interactions belong to complex regulons of minimal two regulators. Only 83 of these 283 interactions (29%) were actually found co-expressed in our modules. In contrast, of the total of 663 interactions in RegulonDB that belong to simple regulons of at least four genes, 398 interactions (60%) were present in our transcriptional modules. Thus a significantly larger fraction of genes that share a single transcription factor could be found co-expressed than the fraction of genes that share at least two transcription factors.

However, this low level of control by more regulators at the module level does not exclude that the expression of individual genes is often influenced by more than one regulator. We identified 85 'connector genes' in our modules (see Figure 7- 5). These are individual genes that are shared by distinct modules, each of which is controlled by different regulators. Modules sharing the same connector gene often show little overlap in their conditions, suggesting that one regulator may in many cases be sufficient to alter the expression of a connector gene upon a specific environmental cue. One example of such a connector gene is the SodA gene product, manganese superoxide dismutase (Fawcett and Wolf, 1995; Jair *et al*, 1995; Tardat and Touati, 1993). The gene *sodA* is present in a module regulated by MarA and SoxS and in a module regulated by Fur, coupling its expression to the multiple antibiotic resistance (MarA), superoxide (SoxS) resistance and the intracellular iron pool (Fur). For other genes, the expression behavior may be highly specific and is therefore never shared with enough other genes to meet our gene content threshold (see Figure 7- 5). Those genes cannot be found in transcriptional modules.

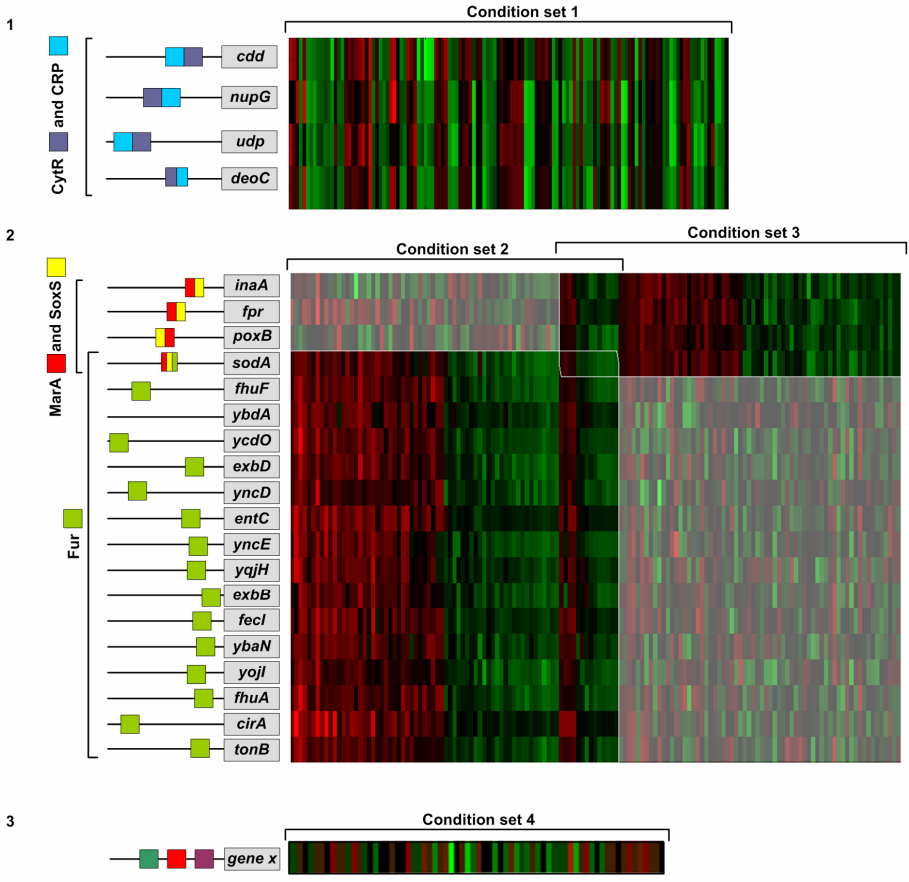


Figure 7- 5: Types of combinatorial regulation. Type 1 shows combinatorial regulation at the module level. The genes cdd, nupG, udp and deoC have two motifs in common (corresponding to the regulators CytR and CRP) and are co-expressed in condition set 1. This kind of control often seems to occur as a combination of a global regulator and a more specific one. Type 2 shows combinatorial regulation at the level of a connector gene. All genes of module 1 share two motifs, MarA and SoxS, and are co-expressed in a subset of conditions. For module 2 all genes are regulated by Fur. SodA, a connector gene, is shared by both modules and thus is regulated by the regulators of module 1 and module 2, but under a different set of conditions (as shown by the heatmap image), indicating that the corresponding regulators of both modules act independently from each other. Both types of interactions mentioned above can be identified by DISTILLER. Cases where condition-specific complex interactions between regulators result in such highly gene-specific expression patterns that genes are no longer found co-expressed in modules (type 3) cannot be detected by DISTILLER.

7.3.5 Comparison with the related methods

We compared our results with those of two recently published network reconstruction methods in order to assess the reliability of our predictions and complementarity between approaches. We selected the method of Faith *et al* (2007) that relies only on microarray data to infer interactions between regulators and target genes and the recently developed data integration framework named SEREND (Ernst *et al*, 2008) because both methods have initially been applied on *E. coli* data. Moreover, for a number of other tools either the software was not available or it was not applicable on large datasets. The goal of SEREND (Ernst *et al*, 2008) resembles our aim best: a (semi-)supervised way of using complementary available data sources to extend the known regulatory network in a reliable way (see chapter 4 for a detailed description of the methods).

Faith *et al* (2007) developed the Context Likelihood of Relatedness (CLR) to infer regulator-target interactions from an *E. coli* Affymetrix compendium (445 arrays) (Faith *et al*, 2007). This method is an extension of the "relevance networks approach" where an interaction between a regulator and a target gene is predicted if the mutual information between the expression profiles of the target and the regulator exceeds a certain threshold. CLR detected 159 interactions (see Materials and Methods). The interactions reported by CLR and DISTILLER cover a very different set of regulators, resulting in a low overlap: only 21 known and 6 novel interactions were identified by both methods. While CLR is able to predict interactions for a larger variety of regulators, i.e. 51 regulators, it is less exhaustive than DISTILLER in recovering all interactions of a given regulator. In particular, DISTILLER can more easily detect interactions involving global regulators where the expression of a regulator and its target are not necessarily directly related, while CLR can identify interactions for those regulators that trigger the expression of a single gene or a very limited set of co-expressed genes, as long as the expression profiles of the regulator and its targets are similar. The low overlap between DISTILLER and Faith *et al* (2007) is to be expected considering the fundamental differences in the underlying assumptions and working principle of both methods: while DISTILLER focuses on data integration, condition dependency, modularity and regulation by combined sets of transcription factors, CLR was designed to deal with gene-specific expression profiles.

In contrast to CLR and similar to DISTILLER, SEREND (Ernst *et al*, 2008) does not rely on the assumption that the transcription factor should have an expression profile related to the profile of its target genes. SEREND (Ernst *et al*, 2008) applies an iterative classification scheme that exploits existing knowledge on regulator-target interactions in a semi-supervised way in order to predict novel interactions for these regulators together with the sign of the

predicted interactions (positive or negative). Ernst *et al* (2008) use confirmed regulator-target interactions for 123 regulators and extended the known regulatory network with 981 novel predictions by integrating a microarray compendium of 445 arrays (Faith *et al*, 2007) and regulatory motif data for 71 transcription factors.

SEREND uses a co-expression score and a motif score to classify unknown interactions: a prediction between a regulator and a target gene will be ranked as highly reliable if the predicted target gene contains a motif instance similar to the motif instances in the known target genes of that regulator and if the target gene is co-expressed with the previously described targets of that regulator. Note that SEREND can also predict interactions for regulators with experimentally confirmed targets but without known binding sites: in that case it is sufficient that the novel target is co-expressed with the known targets, while for DISTILLER only the highly reliable targets for which both the co-expression and the motif constraint are fulfilled will be withheld.

Of the 981 novel interactions identified by SEREND and the 278 by DISTILLER, 78 interactions were found in common by both methods, 903 were uniquely identified by SEREND and 206 were identified by DISTILLER only. Of the 903 unique interactions of SEREND, only 177 predictions corresponded to interactions for regulators without known motif model. Although the latter interactions make up only 18% of the total interactions, they cannot be detected by DISTILLER. Note that these 903 interactions uniquely predicted by SEREND most probably still contain a significant proportion of false positives (see Materials and Methods for the choice of the threshold). The fact that both algorithms consistently retrieve the same predictions, although they are based on completely different methods and datasets, indicates that they are both able to recover the most significant signals from the data.

7.4 Discussion

Data integration frameworks like DISTILLER can enhance gene annotation by exploiting publicly available data in combination with curated information. The main difference of our approach over most previously developed algorithms is its ability to explicitly derive both the conditions under which the interactions take place and the combination of regulators which are responsible for the observed expression. This more detailed level of annotation will become increasingly important with the inclusion of a growing number of experiments and conditions in available expression compendia. Although the transcriptional network of *E. coli* is among the best characterized networks (Resendis-Antonio *et al*, 2005), current

representations are mainly static (Luscombe *et al*, 2004; Shen-Orr *et al*, 2002), because they do not take into account the condition dependency of the interactions.

By applying DISTILLER on the binding site information and microarray compendium, we confirmed 62% of the known transcriptional interactions in *E. coli* and extended the regulons of 29 regulators with 278 putative novel targets. For 30 regulators no interaction was found; most likely either the number of genes in the corresponding modules falls below the gene content threshold, or the conditions needed to trigger these interactions are not present in our compendium (e.g. MelR, triggered by melibiose).

To demonstrate the effectiveness of our approach we chose to validate predicted interactions for FNR. Because FNR is one of the best studied regulators in *E. coli* and genomewide ChIP-chip experiments are available (Grainger *et al*, 2007), finding new targets for this regulator is particularly challenging. In spite of this fact, we selected 11 predictions that have not been reported in previous studies and experimentally demonstrated a physical interaction with FNR for all of them using a ChIP-qPCR analysis.

Considering the condition dependency of transcriptional regulation enables to take a novel perspective on the transcriptional network. Although our results are preliminary and only based on a fraction of well characterized regulators, they reveal a first glimpse on real condition-dependent modularity in the *E. coli* transcriptional network: it seems that modularity in co-expression exists at the level of a single regulator, but that combinatorial regulatory programs seem to decrease the level of modularity and contribute to the network's evolvability (Hartwell *et al*, 1999): the fraction of genes sharing a single transcription factor for which significant co-expression was detected, was significantly larger than the fraction of genes sharing at least two transcription factors for which the co-expression constraint is satisfied. Combinatorial regulation inserts connections between different modules (through so-called connector genes) or generates novel gene specific expression behaviour that is not shared with other genes. The apparently large tolerance of prokaryotes against disruption of modularity may at least partially be explained by the existence of polycistronic transcription: a minimal degree of modularity in expression is always guaranteed by the operon structure (Balaji *et al*, 2007).

7.5 Materials and methods

7.5.1 Benchmarking with RegulonDB and novel interactions

For genes that are organized into operons, usually only the promoter region of the first operon gene contains a motif instance. Because in RegulonDB the direct interaction between a regulator and a target gene is derived from the presence of an experimentally verified motif instance, only the interaction between a regulator and the first operon gene is reported. RegulonDB contains information on 736 such interactions (Salgado *et al*, 2006). Therefore, when comparing the interactions inferred by DISTILLER with the direct interactions in RegulonDB, we only consider those genes inferred by DISTILLER that have the motif instance in their promoter region. All direct interactions inferred by DISTILLER that are not direct interactions according to RegulonDB are considered novel. Some of these interactions might have been reported in recent literature, but not yet covered by RegulonDB.

7.5.2 Static versus condition-dependent combinatorial regulation

To compare the level of static combinatorial regulation present in RegulonDB with the level of combinatorial regulation obtained by additionally taking into account expression data, we applied DISTILLER on one data set only i.e. an interaction matrix containing the known motif-gene interactions from RegulonDB. From this analysis, which does not take into account expression constraints, we counted the number of genes found in modules that were regulated by at least two regulators (a gene was counted more than once if it appeared in multiple modules). This number was compared with a similar figure obtained from the co-expression-constrained modules (see main results). The same procedure was followed for the analysis of non-combinatorial modules. The default gene content threshold was used for all analyses mentioned above.

7.5.3 Conditional dependency of the network

All arrays were grouped into 15 conditional categories assigned by manual curation. For each module, the enrichment of its conditions for each of the functional categories was calculated by means of the hypergeometric

distribution. Subsequently, for each regulator we selected the corresponding modules, and their enrichments for the conditional categories were combined using Fisher's method (Fisher, 1932). This results in a p-value for each combination of a regulator and conditional category. Strong enrichment of one module for a particular category or enrichment of multiple modules belonging to one regulator for the same conditional category can yield significant p-values.

7.5.4 Comparison with other methods

We compared our results on regulator-target interactions for *E. coli* with those initially published by Ernst *et al* (2008) and Faith *et al* (2007). Although SAMBA (Tanay *et al*, 2004) could theoretically be used in a set up similar as the one in this paper we did not include it in our current work as we already exhaustively tested it in a previous study (Lemmens *et al*, 2006).

For the comparison of the DISTILLER network with, on the one hand the networks inferred by CLR and SEREND and, on the other hand the network of RegulonDB, only interactions between a regulator and the first operon gene are taken into account. Interactions between a regulator and the other, non-first operon genes are omitted. For comparison with CLR, we only compared the interactions inferred for those 67 regulators for which a binding site is described in RegulonDB. Some regulators consist of two subunits, like for example FlhDC (Liu and Matsumura, 1994). As Faith *et al*. (Faith *et al*, 2007) have results for subunit 1 (FlhC) and for subunit2 (FlhD), we counted their results for these kind of regulators only once. An alternative comparison with the results obtained by applying CLR on our microarray compendium data set can be found in Additional file 4.

SEREND (Ernst *et al*, 2008) assigns a prioritization for all the predictions per regulator (highest scoring prediction is the best), but does not describe a statistical way to select the number of reliable interactions per regulator. For the comparison with DISTILLER we choose an arbitrary threshold to end up with a defined number of predictions per regulator. We used the threshold defined in the original work of Ernst *et al* (2008), i.e. for each transcription factor we selected the same number of best-scoring predicted targets as there were targets already described for that regulator. The choice of this arbitrary threshold might explain why for SEREND so many uniquely identified interactions were found not confirmed by our approach. These might correspond to false positive predictions for some regulators. Note that this observation does not imply any judgment on the quality of the SEREND output, as it is just an artifact of comparing algorithms. Choosing this threshold upon the usage of SEREND is not an issue as a biologist will focus on the highly ranked predictions first.

Chapter 8

Conclusions and perspectives

8.1 Conclusions

In this thesis we focused on the discovery of transcriptional modules, the basic building blocks of the regulatory network. These regulatory modules consist of sets of genes that are co-expressed and share common regulators and corresponding regulatory motifs. A methodology for module detection based on frequent itemset mining algorithms has been developed. This resulted in a set of state-of-the-art module detection and data integration tools. By applying these module detection tools on biological data, we showed that association rule mining algorithms are equally suitable for reconstructing networks than the more frequently used graph based or probabilistic methodologies. The main contributions of this thesis are:

- A method for the automated collection of data was developed, resulting in the creation of a microarray compendium for *E. coli*.
- The application of ReMoDiscovery on yeast data as a proof-of-concept of itemset mining algorithms.
- The application of DISTILLER on *E. coli* data to study condition-dependent combinatorial regulation in *E. coli*.

In a first phase, a tool named ReMoDiscovery has been developed (Lemmens *et al*, 2006). ReMoDiscovery is based on the Apriori algorithm for discovery of maximal itemsets. A two-step methodology is proposed, in which information from three independently acquired data sources is exploited: microarray data, ChIP-chip data and regulatory motif data. In a first seed discovery step, modules that satisfy three stringent requirements are identified, while coherence criteria during the subsequent seed module extension step are more relaxed. In order to assess to what extent ReMoDiscovery discovers modules similar to those detected by other module identification tools, we compared it with previously described tools on the same benchmark set. Compared to the GRAM algorithm (Bar-Joseph

et al, 2003), we found a significant overlap in both gene and regulator content of the detected modules, but GRAM seems to be much slower and only allows sparse ChIP-chip data. The SAMBA algorithm (Tanay *et al*, 2004) is equally fast as ReMoDiscovery and the identified modules are similar in gene content. The regulator content, however, differs significantly. Based on our analysis, we concluded that the regulatory programs recovered by SAMBA are unlikely to be biologically meaningful.

The biological relevance of our method was assessed by applying it on the extensively studied expression data sets (Gasch *et al*, 2000; Spellman *et al*, 1998) in combination with motif data (Kellis *et al*, 2003) and ChIP-chip data (Harbison *et al*, 2004). Experimental evidence existed for most of the statistically significant modules identified by ReMoDiscovery. In addition, several promising new links between the detected pathways and modules were derived. Some modules, mainly those involved in general metabolism, were detected in both datasets (for instance, ribosome synthesis, galactose metabolism). More specialized modules (for instance, cell cycle, nutrient deprived conditions, stress related conditions, amino acid metabolism, respiration, filamentous growth or meiosis) were restricted to either one of the datasets only. The application of ReMoDiscovery on these yeast data sets can be considered a proof-of-concept that demonstrates that itemset mining algorithms are well suited for the detection of biologically meaningful transcriptional modules.

In a second phase DISTILLER, was developed. DISTILLER is a novel framework for data integration that simultaneously analyzes microarray and motif information to find modules of genes that are co-expressed in a subset of conditions, as well as their corresponding regulators. By including a condition selection strategy we were able to successfully apply DISTILLER on large expression compendia where interesting gene sets are not necessarily co-expressed in all measured conditions.

A microarray compendium for *E. coli* was created as input for DISTILLER. All microarray experiments present in the three major microarray databases (Gene Expression Omnibus, Stanford Microarray Database and ArrayExpress) were collected. In cooperation with the department of Computer Science, GeneReg was developed to automatically download, parse and store yeast microarray data from the Stanford microarray database (Demeter *et al*, 2007). GeneReg also collects and stores regulatory motif data and ChIP-chip data. At the moment we are extending this application to also automate the extraction of microarrays from Gene Expression Omnibus (Barrett *et al*, 2007) and ArrayExpress (Parkinson *et al*, 2007).

DISTILLER consists of three steps: a seed module identification step and a seed module extension step. In the first step, DISTILLER searches for sets of genes that are co-expressed in a subset of conditions and that have a set of

regulators in common. To be able to find these sets, DISTILLER relies on the CHARM algorithm for mining closed itemsets (Zaki and Hsiao, 2002). Afterwards, a clever filtering step was necessary to identify the biologically most interesting modules with limited overlap. This filtering steps makes it possible to use association rule mining based methods. During the last module extension step, the parameters settings are relaxed.

By applying DISTILLER to publicly available regulatory motif and expression data, we evaluated the relation between condition-specific co-expression and combinatorial regulation in the transcriptional network of *Escherichia coli*. We confirm 62% of 736 interactions described in RegulonDB and predict 278 novel interactions. Experimental validation of predicted targets for the well studied oxygen sensor "FNR" showed the effectiveness of our approach. An important observation was that, despite the prevalent number of genes in *E. coli* that are known to be regulated by multiple regulators, the number of modules subject to combinatorial control seems surprisingly low and involves mainly global regulators. However, this low level of combinatorial control at the module level does not exclude that the expression of individual genes is often influenced by more than one regulator, like is the case for 'connector genes' in our modules. These are individual genes that are shared by distinct modules, each of which is controlled by different regulators. Our results provide important clues on the nature of the transcriptional network in *E. coli*: while modularity in co-expression clearly exists at the level of a single regulator, combinatorial regulation inserts connections between different modules (through so-called connector genes) or generates novel gene specific expression behaviour.

8.2 Perspectives

We were able to gain insight into the modularity, combinatorial regulation and condition dependency of bacterial networks. In this thesis, only known regulatory motifs from RegulonDB have been used to screen the intergenic regions of *E. coli*. In future work, new motif instances could be identified and even confirmed with experimental validation. Based on phylogenetic footprinting and in silico motif detection, new motifs were predicted in the CMPG/Bioi for both *E. coli* (Pieter Monsieurs) and *B. subtilis* (Abeer Fadda). These motif predictions can be integrated with the known part of the regulatory network and microarrays. In addition, the application of DISTILLER on two phylogenetic related model organisms should enable comparing prokaryote networks. These kinds of analyses may lead to novel interesting biological discoveries and findings.

DISTILLER makes it straightforward to include any number of data sources related to transcriptional interactions. For instance, ChIP-chip data provide

information about the direct interaction between a regulator and the promoter region of a gene. At the moment, ChIP-chip data are available for a limited number of regulators in *E. coli* only. If more ChIP-chip become publicly available, these data can be included in the analysis. Small RNAs form another data source for inferring the regulatory network as they also play an important role in transcriptional regulation. Methods for detecting small RNAs are available and finding their targets has become an important challenge. Until now, regulation mechanisms via small RNAs have not been studied yet on a genome-wide scale, but the availability and use of high-throughput small RNA data would help in obtaining an improved view and novel insight in the regulatory network.

In addition to the above mentioned conceptual adaptations and extensions of the work presented in this thesis, several more algorithmic and implementational improvements are possible.

Potential implementation related improvements include:

- Implementation of a user-friendly interface of DISTILLER. The current version is a command line version that uses intermediately generated text files as input. Although this command line version is very useful for bioinformaticians, it is not very accessible for biologists.
- Adaptation of DISTILLER to run on the microarray compendium database instead of text files. This adaptation would guarantee that DISTILLER always uses up-to-date microarray information. To achieve this, organism-specific compendia that are updated on a regular base and that collect and parse data from the different microarray databases are needed. This is work in progress together with Riet De Smet, Kristof Engelen and Qiang Fu at the CMPG. Currently compendia for *Escherichia coli*, *Bacillus subtilis* and *Salmonella typhimurium* are already generated in the CMPG. The *Escherichia coli* compendium might also be coupled to the well-known RegulonDB database (in collaboration with prof. J. Collado-Vides, UNAM, Mexico).

Potential algorithmic improvements of DISTILLER:

- Improvement of the filtering step to summarize the results for the user. Association rule mining algorithms usually generate a long list of (overlapping) modules, making their usage not so popular. By determining an interest score for each module, the search method and presentation of the results can focus more on biological interesting or surprising results.

- Incorporation of motif detection algorithms in the itemset mining procedure, for instance by using suffix trees (this is work in progress together with Carolina Fierro and Tijl De Bie).

8.3 From fundamental research to applications of systems biology

The goal of systems biology is to understand how different biological entities function and interact in various conditions. This complex cellular behavior is mediated by the underlying regulatory network. Since systems biology aims at obtaining global insight into the biological system, integration of different high-throughput data sources that show different aspects of the systems' behavior is important in the top-down systems biology approach. Heterogeneous data sources can describe the same biological system from a different point of view. Combining these data can as such help obtaining a holistic view on the system. Moreover, genome-wide scale data have a noisy character. If independently acquired data sources confirm the same observation, we can have more confidence in this observation.

Data integration is thus a key component of systems biology research. Several approaches exist for network inference based on data integration (see also chapter 4). In this thesis, two successful methodologies to identify regulatory modules have been developed (chapter 6 and 7). Our approaches are based on association rule mining algorithms that allow data integration in a very straight-forward way and do not suffer from problems with local optima. Therefore, association rule mining methods are well suited for network reconstruction by integrating multiple data sources. Nevertheless, these algorithms have not been used very frequently for network inference. A reason for this might be that microarray data are not easy to include in the analysis. Most approaches would require discretization or ignore condition dependency of these microarray measurements. In addition, because association rule mining algorithms search for all possibly interesting solutions in an exhaustive manner, a filtering procedure of the results is necessary. During the development of our two data integration frameworks we have tackled these two major problems. By outperforming related module detection tools and by generating biological meaningful results, ReMoDiscovery and DISTILLER proved that methods based on association rule mining are very well suited for the detection of regulatory module networks. In the current systems biology era, the detection of transcriptional modules by the integration of several data sources is important, as it contributes to the unraveling of the regulatory network.

The network inference methods presented in this thesis were used to reconstruct the transcriptional network of *E. coli* and yeast. Since ReMoDiscovery and DISTILLER are generic methods, they could also be applied on higher eukaryotes. Revealing the regulatory network of higher eukaryotes is of utmost importance for medical applications. Many researchers predict that knowledge on the regulatory network may lead to more precise and effective disease treatments and eventually a predictive and preventive mode and even personalized medicine (Hood *et al*, 2004). A growing number of clinically relevant applications that rely the on integration of heterogeneous data sources are being developed (Joyce and Palsson, 2006):

- Studies of cancer or complex chronic inflammatory neurological diseases like multiple sclerosis benefit from the integrated analysis of omics data (Chuang *et al*, 2007; Ibrahim and Gold, 2005; Khalil and Hill, 2005).
- Systems biology provides a powerful tool for drug discovery by identifying specific proteins that are good drug targets and by revealing possible side effects of the new drug (Kitano, 2002). Estimates of the cost of developing a new drug range between 800 and 1200 million dollars (Adams and Brantner, 2006; Yao and Rzhetsky, 2008). Reducing the probability that a drug will fail or reducing the length of the drug development is therefore of utmost economic importance. Yao and Rzhetsky (2008) have analyzed successful drug targets and revealed some common properties about the network, sequence and tissue-expression levels that distinguish them from other human genes.
- In the field of toxicogenomics, the interpretation of toxicological studies takes advantage of genomics, proteomics and metabolomics (Stierum *et al*, 2005).

8.4 From applied systems biology to synthetic biology

During this PhD, two methods for the inference of transcriptional networks were developed. Understanding the underlying principles of the biological system enables the use of this knowledge to re-engineer organisms, a discipline referred to as called synthetic biology. The goal of synthetic biology is to extend and modify the behavior of organisms and engineer them to perform new tasks (recent reviews on this topic (Andrianantoandro *et al*, 2006; Endy, 2005; Heinemann and Panke, 2006; Serrano, 2007). Like systems biology, synthetic biology adopts a system-level approach in which

interactions between network units are crucial, in addition to individual genes and gene products. Because systems biology focuses on a quantitative understanding of the natural biological system it is obvious that synthetic biology relies on and benefits from the insights and knowledge obtained from it: in order to be able to adopt or reengineer a system, it is necessary to understand it. Regulatory module detection approaches, such as ReMoDiscovery or DISTILLER, can help revealing this underlying regulatory network. Good reverse engineering methods are therefore crucial for all current and future successes in synthetic biology (Andrianantoandro *et al*, 2006; Serrano, 2007).

The basic parts in synthetic biology are called devices. These devices are sets of one or more biochemical reactions, like transcription, translation, protein phosphorylation, allosteric regulation, ligand/receptor binding, and enzymatic reactions. A transcriptional device may for instance exist of a regulated gene, a transcription factor, a promoter and RNA polymerase. A growing number of these devices are stored in a central repository called 'Registry of Standard Biological Parts' at MIT (http://parts.mit.edu). Subsequently these basic parts are combined and connected. Integration of these parts in the host cell allows modifying the behavior of the host cell. Usually populations of communicating cells are needed to perform more sophisticated tasks (Andrianantoandro *et al*, 2006).

Small gene circuits have already been developed, including feedforward transcription networks (Mangan and Alon, 2003), regulatory feedback systems such as a 'toggle' switch (Gardner *et al*, 2000) and a cell-cell communication system through acetate metabolism (Bulter *et al*, 2004). Although these circuits usually lack a direct application, they can offer fundamental insight into biological processes and ways to adopt them. In addition to these examples of more fundamental research, there are already a number of exciting examples of synthetic biology as an applied discipline:

- Anderson *et al* (2005) have engineered bacteria that can localize and invade various mammalian tumor cells. The *inv* element of *Yersinia pseudotuberculis*, coding for invasin, is responsible for invasion of mammalian cells when expressed in *E.coli*. Several bacterial species were shown to localize tumor cells after intravenous injection and remain in the neighborhood of the tumor cells because of lower immune surveillance, hypoxic environment and increased nutrients. The *inv* gene was therefore put under control of the *Vibrio fischeri lux* quorum sensing circuit. The manipulated bacterial cells only invade mammalian cells when the density of bacteria is sufficiently large (Anderson *et al*, 2006).
- The study of Weiss (2007) describes preliminary results regarding spatiotemporal control of stem cells. In diabetes, beta cells are under

attack of the autoimmune system. Weiss (2007) have genetically engineered stem cells to maintain a certain level of beta cells even under attack of the autoimmune response. This system relies on bacterial cell-cell communication elements to detect the density of the cells, various regulatory network motifs, and differentiation of the stem cells into beta cells. This kind of application hints at potential future developments in diabetes treatment strategies (Weiss, 2007).

- Another interesting example is the production of terpenoid compounds in *E.coli* (Martin *et al*, 2003) and *Saccharomyces cerevisiae* (Aebersold *et al*, 2000; Lindahl *et al*, 2006). These compounds can be used as an anti-malaria drug.

From these examples it is clear that very often parts or devices from bacteria and yeast are used in synthetic biology. To be able to use these parts, it is necessary to have sufficient insight in the underlying regulatory networks. The developed methods and the corresponding results in this thesis can thus substantially contribute to further developments and applications in synthetic biology.

8.5 Systems biology challenges

Network inference may enable predicting effects of internal perturbations (like genetic mutations) and external perturbations (like drugs) on a sub-cellular, cellular or multi-cellular scale. It is encouraging to see how the number of successful applications of systems biology and synthetic biology gradually increase (see section 8.3 and 8.4). Nevertheless, a number of important challenges remain to be solved:

- At the level of transcriptional network inference, data sources describing gene expression and regulator binding are the most straightforward data sources to include. There are however other factors that carry out regulatory roles, such as small RNAs or histones. Only recently these types of regulation are being investigated (Ioshikhes *et al*, 2006; Segal *et al*, 2006). In addition, new high-throughput techniques, such as deep sequencing, provide new means for studying and investigating the regulatory network in more detail. The incorporations of these novel data sources, and the development of yet other high-throughput measurement techniques remains an important challenge.
- To obtain insight in the overall regulatory network, additional regulation mechanisms need to be taken into account. A number of these mechanisms give rise to extra network layers. Layers at

different levels (different time scales, different levels of details, different mechanisms) need to be combined. In Chuang *et al* (2007), for instance, protein-protein interaction data and expression data was combined for breast cancer metastasis classification. Nevertheless, one of the foremost challenges in systems biology remains the combination of data sources and networks at various scales in a biological meaningful way. Covert *et al* (2004) investigated possibilities for combining transcriptional and metabolic networks. This is however one of the rare examples in which regulatory and metabolic information has been integrated.

- The understanding of the dynamics of the regulatory network, i.e. the change of the different components and their interactions in time, is a very important aspect that is currently missing in most genome scale methods. The condition dependency of transcriptional interactions has recently attracted significant interest with the advent of biclustering algorithms for expression data and increasing computational power. This thesis contains significant contributions in this respect. Although these methods allow capturing some aspects of dynamics in regulation, full quantitative dynamic modeling of genome scale biological systems is yet unfeasible.

Bibliography

Adams CP, Brantner VV (2006). Estimating the cost of new drug development: is it really 802 million dollars? *Health Aff* **25:** 420-428.

Aebersold R, Rist B, Gygi SP (2000). Quantitative proteome analysis: methods and applications. *Ann N Y Acad Sci* **919:** 33-47.

Agrawal R, Imielenski T (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data: May 26-28*, Buneman P, Jajodia S (Eds.),pp. 207-216. ACM Press,New York

Akutsu T, Miyano S, Kuhara S (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac Symp Biocomput* 17-28.

Akutsu T, Miyano S, Kuhara S (2000). Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics* **16:** 727-734.

Albrecht G, Mosch HU, Hoffmann B, Reusser U, Braus GH (1998). Monitoring the Gcn4 protein-mediated response in the yeast *Saccharomyces cerevisiae*. *J Biol Chem* **273:** 12696-12702.

Anderson JC, Clarke EJ, Arkin AP, Voigt CA (2006). Environmentally controlled invasion of cancer cells by engineered bacteria. *J Mol Biol* **355:** 619-627.

Andrianantoandro E, Basu S, Karig DK, Weiss R (2006). Synthetic biology: new engineering rules for an emerging discipline. *Mol Syst Biol* **2:** 2006.

Arita M, Robert M, Tomita M (2005). All systems go: launching cell simulation fueled by integrated experimental biology data. *Curr Opin Biotechnol* **16:** 344-349.

Artamonova II, Frishman G, Frishman D (2007). Applying negative rule mining to improve genome annotation. *BMC Bioinformatics* **8:** 261.

Artamonova II, Frishman G, Gelfand MS, Frishman D (2005). Mining sequence annotation databanks for association patterns. *Bioinformatics* **21 Suppl 3:** iii49-iii57.

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004). Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* **14:** 283-291.

Balaji S, Babu MM, Aravind L (2007). Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of E. coli. *J Mol Biol* **372:** 1108-1122.

Balaji S, Babu MM, Iyer LM, Luscombe NM, Aravind L (2006). Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J Mol Biol* **360:** 213-227.

Balakrishnan R, Christie KR, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hong EL, Nash R, Oughtred R, Skrzypek M, Theesfeld CL, Binkley G, Dong Q, Lane C, Sethuraman A, Weng S, Botstein D, Cherry JM (2005). Fungal BLAST and Model Organism BLASTP Best Hits: new comparison resources at the Saccharomyces Genome Database (SGD). *Nucleic Acids Res* **33:** D374-D377.

Baldi P and Hatfield GW (2002). *DNA microarrays and gene expression.* Cambridge University Press, Cambridge.

Baldi P, Long AD (2001). A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17:** 509-519.

Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D (2007). How to infer gene networks from expression profiles. *Mol Syst Biol* **3:** 78.

Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK (2003). Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* **21:** 1337-1342.

Barabasi AL, Albert R (1999). Emergence of scaling in random networks. *Science* **286:** 509-512.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R (2007). NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res* **35:** D760-D765.

Becquet C, Blachon S, Jeudy B, Boulicaut JF, Gandrillon O (2002). Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. *Genome Biol* **3:** RESEARCH0067.

Bergmann S, Ihmels J, Barkai N (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys Rev E Stat Nonlin Soft Matter Phys* **67:** 031902.

Bernard A, Hartemink AJ (2005). Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pac Symp Biocomput* 459-470.

Bhoite LT, Yu Y, Stillman DJ (2001). The Swi5 activator recruits the Mediator complex to the HO promoter without RNA polymerase II. *Genes Dev* **15:** 2457-2469.

Blaiseau PL, Isnard AD, Surdin-Kerjan Y, Thomas D (1997). Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Mol Cell Biol* **17:** 3640-3648.

Blaiseau PL, Thomas D (1998). Multiple transcriptional activation complexes tether the yeast activator Met4 to DNA. *EMBO J* **17:** 6327-6336.

Blattner FR, Plunkett G, III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997). The complete genome sequence of Escherichia coli K-12. *Science* **277:** 1453-1474.

Bogengruber E, Eichberger T, Briza P, Dawes IW, Breitenbach M, Schricker R (1998). Sporulation-specific expression of the yeast DIT1/DIT2 promoter is controlled by a newly identified repressor element and the short form of Rim101p. *Eur J Biochem* **258:** 430-436.

Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V (2006). The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol* **7:** R36.

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

Brazma A, Vilo J, Ukkonen E, Valtonen K (1997). Data mining for regulatory elements in yeast genome. *Proc Int Conf Intell Syst Mol Biol* **5:** 65-74.

Brent R, Ptashne M (1981). Mechanism of action of the lexA gene product. *Proc Natl Acad Sci U S A* **78:** 4204-4208.

Bruggeman FJ, Westerhoff HV (2007). The nature of systems biology. *Trends Microbiol* **15:** 45-50.

Bulter T, Lee SG, Wong WW, Fung E, Connor MR, Liao JC (2004). Design of artificial cell-cell communication using gene and metabolic networks. *Proc Natl Acad Sci U S A* **101:** 2299-2304.

Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM, Pascual-Montano A (2006). Integrated analysis of gene expression by Association Rules Discovery. *BMC Bioinformatics* **7:** 54.

Chen G, Jensen ST, Stoeckert CJ, Jr. (2007). Clustering of genes into regulons using integrated modeling-COGRIM. *Genome Biol* **8:** R4.

Cheng Y, Church GM (2000). Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* **8:** 93-103.

Chiu SH, Chen CC, Yuan GF, Lin TH (2006). Association algorithm to mine the rules that govern enzyme definition and to classify protein sequences. *BMC Bioinformatics* **7:** 304.

Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol 2007* **3:** 140.

Cipollina C, Alberghina L, Porro D, Vai M (2005). SFP1 is involved in cell size modulation in respiro-fermentative growth conditions. *Yeast* **22:** 385-399.

Constantinidou C, Hobman JL, Griffiths L, Patel MD, Penn CW, Cole JA, Overton TW (2006). A reassessment of the FNR regulon and transcriptomic analysis of the effects of nitrate, nitrite, NarXL, and NarQP as *Escherichia coli* K12 adapts from aerobic to anaerobic growth. *J Biol Chem* **281:** 4802-4815.

Cooper TG (2002). Transmitting the signal of excess nitrogen in *Saccharomyces cerevisiae* from the Tor proteins to the GATA factors: connecting the dots. *FEMS* **26:** 223-238.

Cosma MP (2004). Daughter-specific repression of *Saccharomyces cerevisiae* HO: Ash1 is the commander. *EMBO* **5:** 953-957.

Costanzo M, Schub O, Andrews B (2003). G1 transcription factors are differentially regulated in *Saccharomyces cerevisiae* by the Swi6-binding protein Stb1. *Mol Cell Biol* **23:** 5064-5077.

Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429:** 92-96.

Creighton C, Hanash S (2003). Mining gene expression databases for association rules. *Bioinformatics* **19:** 79-86.

Cutler NS, Pan X, Heitman J, Cardenas ME (2001). The TOR signal transduction cascade controls cellular differentiation in response to nutrients. *Mol Biol Cell* **12:** 4103-4113.

D'haeseleer P, Liang S, Somogyi R (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* **16:** 707-726.

Daugherty JR, Rai R, el Berry HM, Cooper TG (1993). Regulatory circuit for responses of nitrogen catabolic gene expression to the GLN3 and DAL80 proteins

and nitrogen catabolite repression in *Saccharomyces cerevisiae*. *J Bacteriol* **175:** 64-73.

De Keersmaecker SC, Thijs IM, Vanderleyden J, Marchal K (2006). Integration of omics data: how well does it work for bacteria? *Mol Microbiol* **62:** 1239-1250.

Deckert J, Perini R, Balasubramanian B, Zitomer RS (1995). Multiple elements and auto-repression regulate Rox1, a repressor of hypoxic genes in *Saccharomyces cerevisiae*. *Genetics* **139:** 1149-1158.

Delsuc F, Brinkmann H, Philippe H (2005). Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* **6:** 361-375.

Demeter J, Beauheim C, Gollub J, Hernandez-Boussard T, Jin H, Maier D, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G, Ball CA (2007). The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res* **35:** D766-D770.

Devlin C, Tice-Baldwin K, Shore D, Arndt KT (1991). RAP1 is required for BAS1/BAS2- and GCN4-dependent transcription of the yeast HIS4 gene. *Mol Cell Biol* **11:** 3642-3651.

Dhollander T, Sheng Q, Lemmens K, De Moor B, Marchal K, Moreau Y (2007). Query-driven module discovery in microarray data. *Bioinformatics* **23:** 2573-2580.

di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, Elliott SJ, Schaus SE, Collins JJ (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotechnol* **23:** 377-383.

Diep CQ, Peng G, Bewley M, Pilauri V, Ropson I, Hopper JE (2005). Intragenic suppression of Gal3C interaction with Gal80 in the *Saccharomyces cerevisiae* GAL gene switch. *Genetics 2005*.

Dilova I, Aronova S, Chen JC, Powers T (2004). Tor signaling and nutrient-based signals converge on Mks1p phosphorylation to regulate expression of Rtg1.Rtg3p-dependent target genes. *J Biol Chem* **279:** 46527-46535.

Doi K, Hosaka A, Nagata T, Satoh K, Suzuki K, Mauleon R, Mendoza MJ, Bruskiewich R, Kikuchi S (2008). Development of a novel data mining tool to find cis-elements in rice gene promoter regions. *BMC Plant Biol* **8:** 20.

Dorel C, Lejeune P, Rodrigue A (2006). The Cpx system of *Escherichia coli*, a strategic signaling pathway for confronting adverse conditions and for settling biofilm communities? *Res Microbiol* **157:** 306-314.

Eisen MB, Spellman PT, Brown PO, Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95:** 14863-14868.

Endy D (2005). Foundations for engineering biology. *Nature* **438:** 449-453.

Engelen K. Normalizing microarray data: estimating absolute expression levels. 2005. Leuven, Katholieke Universiteit Leuven.

Ernst J, Beg QK, Kay KA, Balazsi G, Oltvai ZN, Bar-Joseph Z (2008). A semi-supervised method for predicting transcription factor-gene interactions in *Escherichia coli*. *PLoS Comput Biol* **4:** e1000044.

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5:** e8.

Fawcett WP, Wolf REJr (1995). Genetic definition of the *Escherichia coli* zwf "soxbox," the DNA binding site for SoxS-mediated induction of glucose 6-phosphate dehydrogenase in response to superoxide. *J Bacteriol* **177:** 1742-1750.

Fisher RA (1932). *Statistical Methods for Research Workers*. Oliver and Boyd, London.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496-512.

Friedman N, Linial M, Nachman I, Pe'er D (2000). Using Bayesian networks to analyze expression data. *J Comput Biol* **7:** 601-620.

Furuchi T, Ishikawa H, Miura N, Ishizuka M, Kajiya K, Kuge S, Naganuma A (2001). Two nuclear proteins, Cin5 and Ydr259c, confer resistance to cisplatin in *Saccharomyces cerevisiae*. *Mol Pharmacol* **59:** 470-474.

Gagiano M, Bauer FF, Pretorius IS (2002). The sensing of nutritional status and the relationship to filamentous growth in *Saccharomyces cerevisiae*. *FEMS Yeast Res* **2:** 433-470.

Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muniz-Rascado L, Martinez-Flores I, Salgado H, Bonavides-Martinez C, Abreu-Goodger C, Rodriguez-Penagos C, Miranda-Rios J, Morett E, Merino E, Huerta AM, Trevino-Quintanilla L, Collado-Vides J (2008). RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* **36:** D120-D124.

Gancedo JM (2001). Control of pseudohyphae formation in *Saccharomyces cerevisiae*. *FEMS Microbiol Rev* **25:** 107-123.

Gao F, Foat BC, Bussemaker HJ (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* **5:** 31.

Gardner TS, Cantor CR, Collins JJ (2000). Construction of a genetic toggle switch in Escherichia coli. *Nature* **403:** 339-342.

Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301:** 102-105.

Gardner TS, Faith J (2005). Reverse-engineering transcription control networks. *Physics of Life Reviews* **2:** 65-88.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11:** 4241-4257.

Georgii E, Richter L, Ruckert U, Kramer S (2005). Analyzing microarray data using quantitative association rules. *Bioinformatics* **21 Suppl 2:** ii123-ii129.

Gil R, Zueco J, Sentandreu R, Herrero E (1991). RCS1, a gene involved in controlling cell size in *Saccharomyces cerevisiae*. *Yeast* **7:** 1-14.

Gorner W, Durchschlag E, Martinez-Pastor MT, Estruch F, Ammerer G, Hamilton B, Ruis H, Schuller C (1998). Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. *Genes Dev* **12:** 586-597.

Grainger DC, Aiba H, Hurd D, Browning DF, Busby SJ (2007). Transcription factor distribution in *Escherichia coli*: studies with FNR protein. *Nucleic Acids Res* **35:** 269-278.

Grainger DC, Hurd D, Goldberg MD, Busby SJ (2006). Association of nucleoid proteins with coding and non-coding segments of the Escherichia coli genome. *Nucleic Acids Res* **34:** 4642-4652.

Grainger DC, Hurd D, Harrison M, Holdstock J, Busby SJ (2005). Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proc Natl Acad Sci U S A* **102:** 17693-17698.

Green J, Guest JR (1994). Regulation of transcription at the ndh promoter of Escherichia coli by FNR and novel factors. *Mol Microbiol* **12:** 433-444.

Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M (2001). Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res* **11:** 1463-1468.

Groisman EA (2001). The pleiotropic two-component regulatory system PhoP-PhoQ. *J Bacteriol* **183:** 1835-1842.

Guelzim N, Bottani S, Bourgine P, Kepes F (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* **31:** 60-63.

Gutierrez-Rios RM, Rosenblueth DA, Loza JA, Huerta AM, Glasner JD, Blattner FR, Collado-Vides J (2003). Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res* **13:** 2435-2443.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* **431:** 99-104.

Hartemink AJ, Gifford DK, Jaakkola TS, Young RA (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac Symp Biocomput* 422-433.

Hartemink AJ, Segal E (2005). Joint learning from multiple types of genomic data. *Pac Symp Biocomput* 445-446.

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999). From molecular to modular cell biology. *Nature* **402:** C47-C52.

Heinemann M, Panke S (2006). Synthetic biology--putting engineering into biology. *Bioinformatics* **22:** 2790-2799.

Hengge-Aronis R (1999). Interplay of global regulators and cell physiology in the general stress response of *Escherichia coli*. *Curr Opin Microbiol* **2:** 148-152.

Herrgard MJ, Covert MW, Palsson BO (2003). Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res 2003 Nov ;13(11):2423 -34 Epub 2003 Oct 14* **13:** 2423-2434.

Hertzberg L, Zuk O, Getz G, Domany E (2005). Finding motifs in promoter regions. *J Comput Biol* **12:** 314-330.

Hinnebusch AG (2005). Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol* **59:** 407-450.

Ho Y, Costanzo M, Moore L, Kobayashi R, Andrews BJ (1999). Regulation of transcription at the *Saccharomyces cerevisiae* start transition by Stb1, a Swi6-binding protein. *Mol Cell Biol* **19:** 5267-5278.

Hofman-Bang J (1999). Nitrogen catabolite repression in *Saccharomyces cerevisiae*. *Mol Biotechnol* **12:** 35-73.

Hood L, Heath JR, Phelps ME, Lin B (2004). Systems biology and new technologies enable predictive and preventative medicine. *Science* **306:** 640-643.

Hu Y, Cooper TG, Kohlhaw GB (1995). The *Saccharomyces cerevisiae* Leu3 protein activates expression of GDH1, a key gene in nitrogen assimilation. *Mol Cell Biol* **15:** 52-57.

Huang Y, Li H, Hu H, Yan X, Waterman MS, Huang H, Zhou XJ (2007). Systematic discovery of functional modules and context-specific functional annotation of human genome. *Bioinformatics* **23:** i222-i229.

Huang YH, Ferrieres L, Clarke DJ (2006). The role of the Rcs phosphorelay in Enterobacteriaceae. *Res Microbiol* **157:** 206-212.

Hung SP, Baldi P, Hatfield GW (2002). Global gene expression profiling in *Escherichia coli K12*. The effects of leucine-responsive regulatory protein. *J Biol Chem* **277:** 40309-40323.

Ibrahim SM, Gold R (2005). Genomics, proteomics, metabolomics: what is in a word for multiple sclerosis? *Curr Opin Neurol* **18:** 231-235.

Ihmels J, Bergmann S, Barkai N (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics* **20:** 1993-2003.

Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N (2002). Revealing modular organization in the yeast transcriptional network. *Nat Genet* **31:** 370-377.

Ioshikhes IP, Albert I, Zanton SJ, Pugh BF (2006). Nucleosome positions predicted through comparative genomics. *Nat Genet* **38:** 1210-1215.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4:** 249-264.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98:** 4569-4574.

Ito T, Ota K, Kubota H, Yamaguchi Y, Chiba T, Sakuraba K, Yoshida M (2002). Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol Cell Proteomics* **1:** 561-566.

Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y (2000). Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A* **97:** 1143-1147.

Ivan G, Szabadka Z, Grolmusz V (2007). Being a binding site: Characterizing residue composition of binding sites on proteins. *Bioinformation* **2:** 216-221.

Jair KW, Martin RG, Rosner JL, Fujita N, Ishihama A, Wolf RE, Jr. (1995). Purification and regulatory properties of MarA protein, a transcriptional activator of *Escherichia coli* multiple antibiotic and superoxide resistance promoters. *J Bacteriol* **177:** 7100-7104.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000). The large-scale organization of metabolic networks. *Nature* **407:** 651-654.

Jorgensen P, Rupes I, Sharom JR, Schneper L, Broach JR, Tyers M (2004). A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size. *Genes Dev* **18:** 2491-2505.

Joyce AR, Palsson BO (2006). The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* **7:** 198-210.

Kato M, Hata N, Banerjee N, Futcher B, Zhang MQ (2004). Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol* **5:** R56.

Kel OV, Romaschenko AG, Kel AE, Wingender E, Kolchanov NA (1995). A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res* **23:** 4097-4103.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241-254.

Khalil IG, Hill C (2005). Systems biology for cancer. *Curr Opin Oncol* **17:** 44-48.

Kitano H (2002). Computational systems biology. *Nature* **420:** 206-210.

Klein C, Struhl K (1994). Protein kinase A mediates growth-regulated expression of yeast ribosomal protein genes by modulating RAP1 transcriptional activity. *Mol Cell Biol* **14:** 1920-1928.

Knight SA, Tamai KT, Kosman DJ, Thiele DJ (1994). Identification and analysis of a *Saccharomyces cerevisiae* copper homeostasis gene encoding a homeodomain protein. *Mol Cell Biol* **14:** 7792-7804.

Koch C, Moll T, Neuberg M, Ahorn H, Nasmyth K (1993). A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science* **261:** 1551-1557.

Kohlhaw GB (2003). Leucine biosynthesis in fungi: entering metabolism through the back door. *Microbiol Mol Biol Rev* **67:** 1-15.

Kuchin S, Vyas VK, Carlson M (2002). Snf1 protein kinase and the repressors Nrg1 and Nrg2 regulate FLO11, haploid invasive growth, and diploid pseudohyphal differentiation. *Mol Cell Biol* **22:** 3994-4000.

Kumar R, Reynolds DM, Shevchenko A, Shevchenko A, Goldstone SD, Dalton S (2000). Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase. *Curr Biol* **10:** 896-906.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la BM, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lazzeroni L, Owen A (2002). Plaid models for gene expression data. *Statistica Sinica* **2:** 61-86.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799-804.

Legrain P, Wojcik J, Gauthier JM (2001). Protein--protein interaction maps: a lead towards cellular functions. *Trends Genet* **17:** 346-352.

Lemmens K, De Bie T, Dhollander T, De Keersmaecker S, Thijs I, Schoofs G, De Weerdt A, De Moor B, Vanderleyden J, Collado-Vides J, Engelen K, Marchal K (2008a). DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli*. *submitted*.

Lemmens K, De Bie T, Dhollander T, Monsieurs P, De Moor B, Collado-Vides J, Engelen K, Marchal K (2008b). The condition-dependent transcriptional network in *Escherichia coli*. *Ann N Y Acad Sci , accepted for publication*.

Lemmens K, Dhollander T, De Bie T, Monsieurs P, Engelen K, Smets B, Winderickx J, De Moor B, Marchal K (2006). Inferring transcriptional modules from ChIP-chip, motif and microarray data. *Genome Biol* **7:** R37.

Liang S, Fuhrman S, Somogyi R (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput* 18-29.

Lim WK, Wang K, Lefebvre C, Califano A (2007). Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics* **23:** i282-i288.

lin RS, Rhee SY, Shafer RW, Das AK (2006). A combined data mining approach for infrequent events: analyzing HIV mutation changes based on treatment history. *Comput Syst Bioinformatics Conf* 385-388.

Lindahl AL, Olsson ME, Mercke P, Tollbom O, Schelin J, Brodelius M, Brodelius PE (2006). Production of the artemisinin precursor amorpha-4,11-diene by engineered Saccharomyces cerevisiae. *Biotechnol Lett* **28:** 571-580.

Little JW, Mount DW, Yanisch-Perron CR (1981). Purified lexA protein is a repressor of the recA and lexA genes. *Proc Natl Acad Sci U S A* **78:** 4199-4203.

Liu H, Colavitti R, Rovira II, Finkel T (2005). Redox-dependent transcriptional regulation. *Circ Res* **97:** 967-974.

Liu X, Matsumura P (1994). The FlhD/FlhC complex, a transcriptional activator of the *Escherichia coli* flagellar class II operons. *J Bacteriol* **176:** 7345-7351.

Lopez FJ, Blanco A, Garcia F, Cano C, Marin A (2008). Fuzzy association rules for biological data analysis: a case study on yeast. *BMC Bioinformatics* **9:** 107.

Loy CJ, Lydall D, Surana U (1999). NDD1, a high-dosage suppressor of cdc28-1N, is essential for expression of a subset of late-S-phase-specific genes in *Saccharomyces cerevisiae*. *Mol Cell Biol* **19:** 3312-3327.

Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431:** 308-312.

Madeira SC, Oliveira AL (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform* **1:** 24-45.

Mai B, Breeden L (1997). Xbp1, a stress-induced transcriptional repressor of the *Saccharomyces cerevisiae* Swi4/Mbp1 family. *Mol Cell Biol* **17:** 6491-6501.

Mai B, Breeden LL (2005). Identification of target genes of a yeast transcriptional repressor. *Methods Mol Biol* **317:** 267-278.

Mallory MJ, Strich R (2003). Ume1p represses meiotic gene transcription in *Saccharomyces cerevisiae* through interaction with the histone deacetylase Rpd3p. *J Biol Chem* **278:** 44727-44734.

Mangan S, Alon U (2003). Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A* **100:** 11980-11985.

Mao XJ, Huo YX, Buck M, Kolb A, Wang YP (2007). Interplay between CRP-cAMP and PII-Ntr systems forms novel regulatory network between carbon metabolism and nitrogen assimilation in *Escherichia coli*. *Nucleic Acids Res* **35:** 1432-1440.

Marchal K, De Keersmaecker S, Monsieurs P, van Boxel N, Lemmens K, Thijs G, Vanderleyden J, De Moor B (2004). In silico identification and experimental validation of PmrAB targets in *Salmonella typhimurium* by regulatory motif detection. *Genome Biol* **5:** R9.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla FR, Califano A (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1:** S7.

Marion RM, Regev A, Segal E, Barash Y, Koller D, Friedman N, O'Shea EK (2004). Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proc Natl Acad Sci U S A* **101:** 14315-14322.

Martin DE, Soulard A, Hall MN (2004). TOR regulates ribosomal protein gene expression via PKA and the Forkhead transcription factor FHL1. *Cell 2004* **119:** 969-979.

Martin VJ, Pitera DJ, Withers ST, Newman JD, Keasling JD (2003). Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat Biotechnol* **21:** 796-802.

Martinez-Pastor MT, Marchler G, Schuller C, Marchler-Bauer A, Ruis H, Estruch F (1996). The *Saccharomyces cerevisiae* zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). *EMBO J* **15:** 2227-2235.

Mates AK, Sayed AK, Foster JW (2007). Products of the *Escherichia coli* acid fitness island attenuate metabolite stress at extremely low pH and mediate a cell density-dependent acid resistance. *J Bacteriol* **189:** 2759-2768.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34:** D108-D110.

McHugh JP, Rodriguez-Quinones F, Abdul-Tehrani H, Svistunenko DA, Poole RK, Cooper CE, Andrews SC (2003). Global iron-dependent gene regulation in *Escherichia coli.* A new mechanism for iron homeostasis. *J Biol Chem* **278:** 29478-29486.

Menon BB, Sarma NJ, Pasula S, Deminoff SJ, Willis KA, Barbara KE, Andrews B, Santangelo GM (2005). Reverse recruitment: the Nup84 nuclear pore subcomplex mediates Rap1/Gcr1/Gcr2 transcriptional activation. *Proc Natl Acad Sci U S A* **102:** 5749-5754.

Mewes HW, Dietmann S, Frishman D, Gregory R, Mannhaupt G, Mayer KF, Munsterkotter M, Ruepp A, Spannagl M, Stumpflen V, Rattei T (2008). MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res* **36:** D196-D201.

Michoel T, Maere S, Bonnet E, Joshi A, Saeys Y, Van den BT, Van Leemput K, van Remortel P, Kuiper M, Marchal K, van de PY (2007). Validating module network learning algorithms using simulated data. *BMC Bioinformatics* **8 Suppl 2:** S5.

Millenaar FF, Okyere J, May ST, van Zanten M, Voesenek LA, Peeters AJ (2006). How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics* **7:** 137.

Mitchell AP (1994). Control of meiotic gene expression in *Saccharomyces cerevisiae*. *Microbiol Rev* **58:** 56-70.

Monsieurs P, De Keersmaecker S, Navarre WW, Bader MW, De Smet F, McClelland M, Fang FC, De Moor B, Vanderleyden J, Marchal K (2005).

Comparison of the PhoPQ regulon in *Escherichia coli* and *Salmonella typhimurium*. *J Mol Evol* **60:** 462-474.

Monsieurs P, Thijs G, Fadda AA, De Keersmaecker SC, Vanderleyden J, De Moor B, Marchal K (2006). More robust detection of motifs in coexpressed genes by using phylogenetic information. *BMC Bioinformatics* **7:** 160.

Morgan XC, Ni S, Miranker DP, Iyer VR (2007). Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining. *BMC Bioinformatics* **8:** 445.

Murali TM, Kasif S (2003). Extracting conserved gene expression motifs from gene expression data. *Pac Symp Biocomput* 77-88.

Natarajan K, Meyer MR, Jackson BM, Slade D, Roberts C, Hinnebusch AG, Marton MJ (2001). Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Mol Cell Biol* **21:** 4347-4368.

Nevitt T, Pereira J, Rodrigues-Pousada C (2004). YAP4 gene expression is induced in response to several forms of stress in *Saccharomyces cerevisiae*. *Yeast* **21:** 1365-1374.

Okada Y, Fujibuchi W, Horton P (2007). A biclustering method for gene expression module discovery using a closed itemset enumeration algorithm. *IPSJ Transactions on Bioinformatics* **48:** 39-48.

Orlando V (2000). Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* **25:** 99-104.

Outten FW, Outten CE, Hale J, O'Halloran TV (2000). Transcriptional activation of an *Escherichia coli* copper efflux regulon by the chromosomal MerR homologue, cueR. *J Biol Chem* **275:** 31024-31029.

Oyama T, Kitano K, Satou K, Ito T (2002). Extraction of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics* **18:** 705-714.

Pan X, Heitman J (2000). Sok2 regulates yeast pseudohyphal differentiation via a transcription factor cascade that regulates cell-cell adhesion. *Mol Cell Biol* **20:** 8364-8372.

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A (2007). ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* **35:** D747-D750.

Partridge JD, Browning DF, Xu M, Newnham LJ, Scott C, Roberts RE, Poole RK, Green J (2008). Characterization of the Escherichia coli K-12 ydhYVWXUT operon: regulation by FNR, NarL and NarP. *Microbiology* **154:** 608-618.

Partridge JD, Poole RK, Green J (2007). The Escherichia coli yhjA gene, encoding a predicted cytochrome c peroxidase, is regulated by FNR and OxyR. *Microbiology* **153:** 1499-1507.

Pascual-Ahuir A, Posas F, Serrano R, Proft M (2001). Multiple levels of control regulate the yeast cAMP-response element-binding protein repressor Sko1p in response to stress. *J Biol Chem* **276:** 37373-37378.

Pastor D, Cortes-Calabuig A, Lemmens K, De Moor B, Marchal K, Denecker M (2007). GeneReg: Integration of Experimental Data on the DNA Transcription Process. In *BNAIC 2007*

Peterson SN, Dahlquist FW, Reich NO (2007). The role of high affinity non-specific DNA binding by Lrp in transcriptional regulation and DNA organization. *J Mol Biol* **369:** 1307-1317.

Pham TH, Satou K, Ho TB (2004). Mining yeast transcriptional regulatory modules from factor DNA-binding sites and gene expression data. *Genome Inform* **15:** 287-295.

Pilauri V, Bewley M, Diep C, Hopper J (2005). Gal80 dimerization and the yeast GAL gene switch. *Genetics* **169:** 1903-1914.

Pinson B, Gabrielsen OS, Daignan-Fornier B (2000). Redox regulation of AMP synthesis in yeast: a role of the Bas1p and Bas2p transcription factors. *Mol Microbiol* **36:** 1460-1469.

Plumbridge J (2001). DNA binding sites for the Mlc and NagC proteins: regulation of *nagE*, encoding the N-acetylglucosamine-specific transporter in *Escherichia coli*. *Nucleic Acids Res* **29:** 506-514.

Powers T, Walter P (1999). Regulation of ribosome biogenesis by the rapamycin-sensitive TOR-signaling pathway in *Saccharomyces cerevisiae*. *Mol Biol Cell* **10:** 987-1000.

Prigent-Combaret C, Brombacher E, Vidal O, Ambert A, Lejeune P, Landini P, Dorel C (2001). Complex regulatory network controls initial adhesion and biofilm formation in *Escherichia coli* via regulation of the *csgD* gene. *J Bacteriol* **183:** 7213-7223.

Prochasson P, Florens L, Swanson SK, Washburn MP, Workman JL (2005). The HIR corepressor complex binds to nucleosomes generating a distinct protein/DNA complex resistant to remodeling by SWI/SNF. *Genes Dev* **19:** 2534-2539.

Pul U, Wurm R, Wagner R (2007). The role of LRP and H-NS in transcription regulation: involvement of synergism, allostery and macromolecular crowding. *J Mol Biol* **366:** 900-915.

Quail MA, Haydon DJ, Guest JR (1994). The pdhR-aceEF-lpd operon of Escherichia coli expresses the pyruvate dehydrogenase complex. *Mol Microbiol* **12:** 95-104.

Raitt DC, Johnson AL, Erkine AM, Makino K, Morgan B, Gross DS, Johnston LH (2000). The Skn7 response regulator of *Saccharomyces cerevisiae* interacts with Hsf1 in vivo and is required for the induction of heat shock genes by oxidative stress. *Mol Biol Cell* **11:** 2335-2347.

Raivio TL (2005). Envelope stress responses and Gram-negative bacterial pathogenesis. *Mol Microbiol* **56:** 1119-1128.

Ramakers C, Ruijter JM, Deprez RH, Moorman AF (2003). Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci Lett* **339:** 62-66.

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297:** 1551-1555.

Regnacq M, Alimardani P, El Moudni B, Berges T (2001). SUT1p interaction with Cyc8p(Ssn6p) relieves hypoxic genes from Cyc8p-Tup1p repression in *Saccharomyces cerevisiae*. *Mol Microbiol* **40:** 1085-1096.

Reiss DJ, Baliga NS, Bonneau R (2006). Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* **7:** 280.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000). Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306-2309.

Resendis-Antonio O, Freyre-Gonzalez JA, Menchaca-Mendez R, Gutierrez-Rios RM, Martinez-Antonio A, Avila-Sanchez C, Collado-Vides J (2005). Modular analysis of the transcriptional regulatory network of *E. coli*. *Trends Genet* **21:** 16-20.

Rives AW, Galitski T (2003). Modular organization of cellular networks. *Proc Natl Acad Sci U S A* **100:** 1128-1133.

Rodriguez A, Carazo JM, Trelles O (2005). Mining association rules from biological databases. *Journal of the American society for information science and technology* **56:** 493-504.

Rothfels K, Tanny JC, Molnar E, Friesen H, Commisso C, Segall J (2005). Components of the ESCRT pathway, DFG16, and YGR122w are required for Rim101 to act as a corepressor with Nrg1 at the negative regulatory element of the DIT1 gene of *Saccharomyces cerevisiae*. *Mol Cell Biol* **25:** 6772-6788.

Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J (2006). RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* **34:** D394-D397.

Schmitt AP, McEntee K (1996). Msn2p, a zinc finger DNA-binding protein, is the transcriptional activator of the multistress response in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* **93:** 5777-5782.

Schneider JC, Guarente L (1991). Regulation of the yeast CYT1 gene encoding cytochrome c1 by HAP1 and HAP2/3/4. *Mol Cell Biol* **11:** 4934-4942.

Schuller HJ (2003). Transcriptional control of nonfermentative metabolism in the yeast *Saccharomyces cerevisiae*. *Curr Genet* **43:** 139-160.

Scott MP, Matsudaira P, Lodish H, Darnell J, Zipusky L, Kaiser CA, Berk A, and Krieger M (2003). *Molecular cell biology.* W. H. Freeman.

Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J (2006). A genomic code for nucleosome positioning. *Nature* **442:** 772-778.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34:** 166-176.

Serrano L (2007). Synthetic biology: promises and challenges. *Mol Syst Biol* **3:** 158.

Shakhnovich BE, Reddy TE, Galinsky K, Mellor J, Delisi C (2004). Comparisons of predicted genetic modules: identification of co-expressed genes through module gene flow. *Genome Inform Ser Workshop Genome Inform* **15:** 221-228.

Shen-Orr SS, Milo R, Mangan S, Alon U (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet* **31:** 64-68.

Sheng Q, Moreau Y, De Moor B (2003). Biclustering microarray data by Gibbs sampling. *Bioinformatics* **19 Suppl 2:** ii196-ii205.

Shenhar G, Kassir Y (2001). A positive regulator of mitosis, Sok2, functions as a negative regulator of meiosis in *Saccharomyces cerevisiae*. *Mol Cell Biol* **21:** 1603-1612.

Spector MS, Raff A, DeSilva H, Lee K, Osley MA (1997). Hir1p and Hir2p function as transcriptional corepressors to regulate histone gene transcription in the *Saccharomyces cerevisiae* cell cycle. *Mol Cell Biol* **17:** 545-552.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* **9:** 3273-3297.

Staal FJ, van der BM, Wessels LF, Barendregt BH, Baert MR, van den Burg CM, van Huffel C, Langerak AW, van dV, V, Reinders MJ, van Dongen JJ (2003). DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. *Leukemia* **17:** 1324-1332.

Steber CM, Esposito RE (1995). UME6 is a central component of a developmental regulatory switch controlling meiosis-specific gene expression. *Proc Natl Acad Sci U S A* **92:** 12490-12494.

Stierum R, Heijne W, Kienhuis A, van Ommen B, Groten J (2005). Toxicogenomics concepts and applications to study hepatic effects of food additives and chemicals. *Toxicol Appl Pharmacol* **207:** 179-188.

Tamada Y, Kim S, Bannai H, Imoto S, Tashiro K, Kuhara S, Miyano S (2003). Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics* **19 Suppl 2:** ii227-ii236.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* **96:** 2907-2912.

Tamura M, D'haeseleer P (2008). Microbial genotype-phenotype mapping by class association rule mining. *Bioinformatics* **24:** 1523-1529.

Tanay A, Sharan R, Kupiec M, Shamir R (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A* **101:** 2981-2986.

Tanay A, Sharan R, Shamir R (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics* **18 Suppl 1:** S136-S144.

Tardat B, Touati D (1993). Iron and oxygen regulation of *Escherichia coli* MnSOD expression: competition between the global regulators Fur and ArcA for binding to DNA. *Mol Microbiol* **9:** 53-63.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999). Systematic determination of genetic network architecture. *Nat Genet* **22:** 281-285.

Taverner NV, Smith JC, Wardle FC (2004). Identifying transcriptional targets. *Genome Biol* **5:** 210.

Tegner J, Yeung MK, Hasty J, Collins JJ (2003). Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc Natl Acad Sci U S A* **100:** 5944-5949.

Thevelein JM, Cauwenberg L, Colombo S, De Winde JH, Donation M, Dumortier F, Kraakman L, Lemaire K, Ma P, Nauwelaers D, Rolland F, Teunissen A, Van Dijck P, Versele M, Wera S, Winderickx J (2000). Nutrient-induced signal transduction through the protein kinase A pathway and its role in the control of metabolism, stress resistance, and growth in yeast. *Enzyme Microb Technol* **26:** 819-825.

Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y (2002). A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol* **9:** 447-464.

Thijs IM, De Keersmaecker SC, Fadda A, Engelen K, Zhao H, McClelland M, Marchal K, Vanderleyden J (2007). Delineation of the Salmonella enterica serovar Typhimurium HilA regulon through genome-wide location and transcript analysis. *J Bacteriol* **189:** 4587-4596.

Thomas D, Surdin-Kerjan Y (1997). Metabolism of sulfur amino acids in *Saccharomyces cerevisiae*. *Microbiol Mol Biol Rev* **61:** 503-532.

Timson DJ, Ross HC, Reece RJ (2002). Gal3p and Gal1p interact with the transcriptional repressor Gal80p to form a complex of 1:1 stoichiometry. *Biochem J* **363:** 515-520.

Tuzhilin A, Adomavicius G (2002). Handling very large numbers of association rules in the analysis of microarray data. ACM,

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403:** 623-627.

Unden G, Bongaerts J (1997). Alternative respiratory pathways of *Escherichia coli*: energetics and transcriptional regulation in response to electron acceptors. *Biochim Biophys Acta* **1320:** 217-234.

Valenzuela L, Aranda C, Gonzalez A (2001). TOR modulates GCN4-dependent expression of genes turned on by nitrogen limitation. *J Bacteriol* **183:** 2331-2334.

Van den Bulcke T, Lemmens K, Van de Peer Y, Marchal K (2006). Inferring Transcriptional Networks by Mining Omics Data. *Current Bioinformatics* **1:** 301-313.

Van Leemput K, Van den BT, Dhollander T, De Moor B, Marchal K, van Remortel P (2008). Exploring the operational characteristics of inference algorithms for transcriptional networks by means of synthetic data. *Artif Life 2008 Winter ;14(1):49 -63* **14:** 49-63.

Van Slyke C, Grayhack EJ (2003). The essential transcription factor Reb1p interacts with the CLB2 UAS outside of the G2/M control region. *Nucleic Acids Res* **31:** 4597-4607.

van Someren EP, Vaes BL, Steegenga WT, Sijbers AM, Dechering KJ, Reinders MJ (2006). Least absolute regression network analysis of the murine osteoblast differentiation network. *Bioinformatics* **22:** 477-484.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di F, V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D,

Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M (2001). The sequence of the human genome. *Science* **291:** 1304-1351.

Vyas VK, Berkey CD, Miyao T, Carlson M (2005). Repressors Nrg1 and Nrg2 regulate a set of stress-responsive genes in *Saccharomyces cerevisiae*. *Eukaryot Cell* **4:** 1882-1891.

Wagner A, Fell DA (2001). The small world inside large metabolic networks. *Proc Biol Sci* **268:** 1803-1810.

Weiss R (2007). Synthetic biology: from bacteria to stem cells.,pp. 634-635. ACM,

Winderickx J, Holsbeeks I, Lagatie O, Giots F, Thevelein J, de Winde H (2003). From Feast to Famine: Adaptation to nutrient availability in yeast. In *Yeast stress responses*, Hohmann S, Mager WH (Eds.),pp. 306-386. Springer publishers,Berlin

Wittenberg C, Reed SI (2005). Cell cycle-dependent transcription in yeast: promoters, transcription factors, and transcriptomes. *Oncogene* **24:** 2746-2755.

Wolf DM, Arkin AP (2003). Motifs, modules and games in bacteria. *Curr Opin Microbiol* **6:** 125-134.

Wong CM, Ching YP, Zhou Y, Kung HF, Jin DY (2003). Transcriptional regulation of yeast peroxiredoxin gene TSA2 through Hap1p, Rox1p, and Hap2/3/5p. *Free Radic Biol Med* **34:** 585-597.

Wyrick JJ, Young RA (2002). Deciphering gene expression regulatory networks. *Curr Opin Genet Dev* **12:** 130-136.

Xu X, Wang L, Ding D (2004). Learning module networks from genome-wide location and expression data. *FEBS Lett* **578:** 297-304.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30:** e15.

Yao L, Rzhetsky A (2008). Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res* **18:** 206-213.

Yu H, Greenbaum D, Xin LH, Zhu X, Gerstein M (2004a). Genomic analysis of essentiality within protein networks. *Trends Genet* **20:** 227-231.

Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED (2004b). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* **20:** 3594-3603.

Zaki MJ, Hsiao C (2002). CHARM: An efficient algorithm for Closed Itemset Mining. In *Proceedings of the Second SIAM International Conference on Data Mining (SDM '02)*, Grossman R, Han J, Kumar V, Mannila H, Motwani R (Eds.),pp. 457-473.

Zaragoza O, Gancedo JM (2000). Pseudohyphal growth is induced in *Saccharomyces cerevisiae* by a combination of stress and cAMP signalling. *Antonie Van Leeuwenhoek* **78:** 187-194.

Zhou H, Winston F (2001). NRG1 is required for glucose repression of the SUC2 and GAL genes of *Saccharomyces cerevisiae*. *BMC Genet* **2:** 5.

Zhu G, Spellman PT, Volpe T, Brown PO, Botstein D, Davis TN, Futcher B (2000). Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature 2000* **406:** 90-94.

Zitomer RS, Lowry CV (1992). Regulation of gene expression by oxygen in *Saccharomyces cerevisiae*. *Microbiol Rev* **56:** 1-11.

Curriculum vitae

Karen Lemmens was born in Hasselt, Belgium, on January 17th, 1980. In 1998 she started her education in applied biological sciences at the K.U. Leuven, where she received the Candidacy diploma in Bioscience Engineering in 2000, and the Masters diploma in Cellular and Biotechnological Engineering in 2003. In 2004 she received the Masters diploma of Sciences in Bioinformatics. Since October 2004 she has been pursuing her PhD as a Research Assistant in the research group ESAT-SCD, under the supervision of Prof. Bart De Moor and Prof. Kathleen Marchal.