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

MOTIF DETECTION IN PROKARYOTES BASED ON COMPARATIVE GENOMICS

Jury:
Prof. dr. ir. N.N., voorzitter
Prof. dr. ir. B. De Moor, promotor
Prof. dr. ir. K. Marchal, co-promotor
Prof. dr. ir. J. Van Impe
Prof. dr. ir. J. Vanderleyden
Prof. dr. Y. Van de Peer (U.Gent)
Prof. dr. M. McClelland (SKCC, San Diego)
Prof. dr. ir. J. Michiels

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Pieter MONSIEURS

UDC 681.3*J3

December 2006

© 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/2006/7515/99

ISBN 978-90-5682-766-3

Voorwoord

Om zeker niemand te vergeten in mijn voorwoord, heb ik mij steeds voorgenomen om op tijd hieraan te beginnen. Niet dus… Iedereen bedanken die bijgedragen heeft aan dit doctoraat, zal moeilijk zijn. Maar als ik het chronologisch op een rij zet, zou het zonder al te grote hiaten moeten lukken.

Wanneer het juist misgegaan is, valt natuurlijk moeilijk te zeggen, maar Meneer Stef Smet, leraar biologie en chemie aan het Sint-Jan Berchmanscollege van Mol, heeft daar zeker een cruciale rol in gespeeld. Het enthousiasme waarmee hij vertelde over de genetica van de fruitvlieg ("in de volksmond ook wel *Drosohphila melanogaster* genoemd") of *Escherichia coli* ("Eventjes oefenen op de uitspraak voor het mondelinge examen heren") waren de aanleiding om voor bio-ingenieur te gaan studeren. De eerste grote beslissing op studiegebied was een feit. Na ongeveer twee jaar moest nog de major ("specialisatie") gekozen worden. Deze keer waren het de lessen Biochemie van Jos (die ik in die tijd nog met Professor Vanderleyden aansprak) die de doorslag gaven om voor cel- en genbiotechnologie te kiezen. Vanuit de begeleidingscommissie en via verschillende samenwerkingen is Jos steeds nauw betrokken gebleven bij mijn onderzoek. Jos, bedankt voor alle hulp, je interesse voor mijn bio-informatica-onderzoek en vooral voor de kritische biologische vragen. Thesiskeuze, de derde belangrijke beslissing. Heimwee naar wiskunde, de interesse in computers en het avontuur van het onbekende deden mij beslissen om tijdens mijn laatste ingenieursjaar een thesis te gaan maken binnen de bio-informatica onderzoeksgroep van ESAT. Het was misschien wel de minst rationele beslissing van alle drie, maar ik heb er nog geen seconde spijt van gehad. Daarom moet ik mijn promotor Bart De Moor bedanken voor de kansen die ik gekregen hebben binnen zijn onderzoeksgroep, aanvankelijk als thesisstudent en later als doctoraatsstudent. Bart, bedankt voor alle steun tijdens mijn verblijf op ESAT.

De persoon die het meest bijgedragen heeft tot mijn thesis, en zonder wie er hier helemaal geen doctoraat zou liggen, is Kathleen (Professor Marchal voor de vrienden). Met je uitgebreide bio-informaticakennis, je enthousiasme, je gedrevenheid, je begrip (zelfs als de zoveelste deadline weer gevaarlijk dicht in de buurt kwam) en geduld (met een literair non-talent als mij) zorg je er voor dat elke computerleek binnen enkele maanden verkocht is aan bio-informatica. Er waren zelfs thesisstudenten die beweerden dat ze de aan/uit knop van een computer amper wisten staan maar nu al bijna een doctoraat in de bio-informatica op zak hebben (nietwaar Karen?). Ook al heb je nu 18 'kindjes' die je moet begeleiden, toch lukt het je om voor mij en de anderen tijd vrij te maken tussen alle andere professorbeslommeringen in. Merci Kathleen!

Voor elk hoofdstuk uit deze thesis heb ik mogen samenwerken met andere onderzoeksgroepen. Een speciale dankuwel gaat dan ook uit naar Sigrid De Keersmaecker die er mee verantwoordelijk voor is dat dit doctoraat niet vol biologische fouten staat. Haar expertise en *Salmonella* kennis heeft in sterke mate bijgedragen tot dit doctoraat, en is terug te vinden in bijna elk hoofdstuk. Ook Jan Michiels en Gunter Dirix zou ik willen bedanken voor hun inbreng. Al was het onderwerp afwijkend van de rest van mijn doctoraat, het was een leerrijke en aangename samenwerking. Het feit dat Gunter op het einde van ons project binnen CMPG al uitgelachen werd als bio-informaticus, zegt genoeg over onze samenwerking. Hoewel de samenwerking met Dirk Gevers en Yves Van de Peer niet tot officiële wetenschappelijke output heeft geleid, heb ik tijdens onze vergaderingen enorm veel ideeën opgestoken. En Dirk, laat maar weten wanneer de volgende conferentie in Canada doorgaat, ik zal er staan. I also want to thank Dr. William Navarre, Dr. Martin Bader and Prof. Ferric Fang from the University of Washington and Prof. Michael McClelland from the Sidney Kimmel Cancer Centre in San Diego for the work on the PhoPQ regulon. A special word of gratitude to Prof. McClelland for his cooperation and helpful discussions during his stay in Belgium. It’s a real honor for me that you agreed to be a member of my PhD jury. Tot slot wil ik ook nog Professor Jan Van Impe bedanken voor de tijd en moeite die hij geïnvesteerd heeft in mijn doctoraat als lid van de begeleidingscommissie en partner in het SQUAD project.

Natuurlijk kan ik de collega’s van de bio-informaticagroep niet vergeten te bedanken. De eilandvrienden uit de toren verdienen daarbij een speciale vermelding. Ruth (alleen koffiepauze houden is toch ook niet alles) en Kristof (binnen een maand zit ge terug opgescheept met mij aan eenzelfde eiland), merci he. Gert, de linux-goeroe voor alle computerhulp. Karen, Thomas, Tim, Valerie, Marleen, Wout, Abeer, Aminael voor de fijne samenwerking. De eerste werkmaanden op ESAT waren we met alle “nieuwkomers” een soort nomadenstam die doorheen Heverlee doolden, van de oude bibliotheek op ESAT tot 200F. Nochtans heeft dat er voor gezorgd dat je je direct op je gemak voelde in dat grote departement. Bert, Frizo, Ruth, Nathalie, Marcello, Steffen, Joke en Raf, thanx! En natuurlijk ook de rest van de onderzoeksgroep (maar mijn voorwoord zou een bladzijde langer zijn moest ik alle namen vermelden). Voor de administratieve rompslomp tijdens mijn doctoraat kon ik steeds reken op Ida, Ilse en Bart, merci!

Ontspanning is onontbeerlijk tijdens een doctoraat, en vooral tijdens het schrijven ervan. Daarom wil ik al de “bekwame” mensen van district Mol bedanken voor de nodige ontspannende momenten. Lore, merci om de gaten in mijn scoutsengagement op te vangen als ik het te druk had met mijn doctoraat. Drie van mijn vier doctoraatsjaren deelde ik een appartement in de Heverlee. Wim, merci voor alle hulp en steun tijdens de voorbije jaren. Je hebt meer bijgedragen tot dit doctoraat dan je waarschijnlijk zelf vermoed. De squashavonden (of waren het Duvel-avonden?) op donderdag zijn ondertussen legendarisch (en nee, dat squashballeke in je oog was geen revanche omdat ik achter stond). Ah ja, en voor die rode wijn op jouw nieuwe stoelen moeten we ook nog altijd een oplossing vinden. Merci makker, en tot in de Perel! Hoewel de avonden voor een doctoraatsstudent ook bedoeld zijn om te werken (ja toch?), was er tijdens de week soms toch nog tijd voor een bezoek aan de Leuvense horeca. Merci aan alle scoutsvrienden.

Natuurlijk moet ik ook nog de familie Daems-Vansweevelt bedanken: voor het uitlenen van een moutainbike om mij uit te kunnen leven tussen twee hoofdstukken door, de nodige terrasmomenten tijdens de warme zomermaanden (ook al moest dan al onze chips eraan geloven), en zeker en vast voor de zorgvuldig voorbereide receptie.

Zonder de onvoorwaardelijke steun van mijn ouders had dit boek er nooit gelegen. Van het brengen van eten tijdens de blokperiode in eerste kandidatuur, tot het zorgvuldig checken van de draft van mijn doctoraat op de avond vooraleer het naar de drukker moet: jullie hebben er altijd gestaan voor mij, en één dankuwel is veel te weinig om uit te drukken wat dat betekent. Janik, ik word waarschijnlijk nooit zo’n grote computer-nerd als jij, maar je blijft mijn grote voorbeeld (ahum). En Katleen, bedankt voor het delen van dezelfde doctoraatsbeslommeringen.

Lies, merci voor al je steun in moeilijke schrijfmomenten, je begrip voor het weekend- en avondwerk, om onze dromen uit te stellen tot na een doctoraat, en vooral om er altijd te zijn voor mij…

Pieter Monsieurs

December 2006

Abstract

Bacteria are dynamic organisms, able to survive in different environmental conditions. In order to adapt their cellular machinery to continuously changing conditions, bacteria are equipped with flexible regulatory networks.

As in bacteria the rate of transcriptional initiation is an important check point for control of gene expression, we focus in this thesis on unraveling the regulatory mechanism responsible for the transcriptional control. The basic functional element of a transcriptional regulatory network is the gene's promoter region which contains the regulatory binding sites for the transcription factors that regulate its expression. Over the past years considerable effort has been put in the in silico identification of these regulatory binding sites, which resulted in a diverse range of motif detection methods. With the availability of entire genomes new opportunities opened up for comparative genomics and motif detection. Motif detection methods based on comparative genomics (phylogenetic footprinting) exploit the conservation of motifs in orthologous promoter regions based on the idea that evolutionary forces tend to preferentially retain the biologically functional DNA sequences.

In this PhD we used the concept of phylogenetic footprinting to extend the information on two poorly characterized regulons involved in the pathogenicity of *Salmonella typhimurium*. For the PmrAB regulatory system, several novel targets were detected by our in silico analysis, a few of which were validated by experimental wet lab analysis. The PhoPQ systems, a sensor for magnesium ions and an important regulator of virulence genes in some pathogenic bacteria, were further characterised by combining microarray data with in silico motif prediction. By comparing to what extent this regulon overlapped between *Salmonella typhimurium* and its close relative *Escherichia coli* we could show that the PhoPQ two-component system seemingly quickly adopted novel targets during evolution, possibly giving rise to the difference in phenotypes between the two related species.

The fact that both regulons mentioned above were already partially characterized facilitated their analysis. However, if one wants to identify regulatory motifs without any prior information, one has to rely on the mere property of "statistical overrepresentation". In these cases, the existing motif detection tools will fail if the involved species are evolutionary too related or if the regulatory motifs are present only in a limited subset of genes. For this reason, we developed an adapted version of MotifSampler that allows detection of niche- or species-specific regulatory motifs or motifs that belong to sparsely connected hubs in the regulatory network.

The tools developed in this PhD study all apply to the identification of regulatory motifs. As the detection of regulatory motifs is complicated because they are short, degenerated and only present in a limited number of promoter regions, we can apply theses tools to biological questions facing the same limitations. We illustrate the wide application area of our tools by detecting potential targets of regulatory RNA and by detecting small signalling peptides.

Korte inhoud

Bacteriën zijn dynamische organismen die in staat zijn zich aan te passen aan uiteenlopende omgevingsomstandigheden. Om in staat te zijn hun cellulaire systeem voortdurend aan te passen aan de wisselende omgevingsomstandigheden, zijn deze organismen uitgerust met flexibele regulatorische netwerken.

In bacteriën wordt de expressie van een gen in sterke mate bepaald door transcriptiesnelheid. Daarom leggen we in deze thesis de nadruk op regulatorische systemen die deze transcriptie controleren. De belangrijkste bouwsteen van een transcriptioneel regulatorisch netwerk is de promoterregio van een gen. Dit gebied bevat immers de bindingsplaatsen voor regulatorproteïnen die de expressie van het overeenkomstige gen controleren. De voorbije jaren zijn reeds heel wat inspanningen geleverd m.b.t. de computationele identificatie van dergelijke regulatorische bindingsplaatsen, wat leidde tot een uiteenlopende aanbod van motiefdetectie-algoritmen. De beschikbaarheid van de volledige genoomsequenties van diverse species biedt echter nieuwe mogelijkheden voor motiefdetectie. Vertrekkend van de hypothese dat regulatorische motieven biologisch functionele sequenties zijn en dus in de evolutie bij voorkeur bewaard blijven, biedt dit de mogelijkheid om motieven te identificeren via vergelijking van orthologe promoterregio's uit verschillende species ("phylogenetic footprinting").

In deze thesis gebruiken we het idee van "phylogenetic footprinting" om een duidelijker beeld te krijgen van twee regulatiesystemen die van belang zijn voor het infectiemechanisme van *Salmonella typhimurium*. Via een bio-informatica analyse identificeerden we nieuwe biologisch relevante genen die betrokken zijn in het PmrAB regulatie systeem. Het PhoPQ regulatie systeem werd verder ontrafeld door gebruik te maken van een combinatie van expressie en motiefdata. Uit een vergelijking van de samenstelling van het PhoPQ regulatiesysteem tussen *S. typhimurium* and *Escherichia coli* concludeerden we dat de samenstelling van dit regulatiesysteem erg flexibel is. Deze waarneming geeft een mogelijke verklaring voor de uiteenlopende fenotypes die geobserveerd worden voor twee evolutionair nauw verwante species.

Voor beide bovenvermelde regulatorische systemen beschikken we over een beperkte hoeveelheid prior informatie. Indien men echter dergelijke motieven wil identificeren zonder prior informatie kan statistische overrepresentatie van motieven in promoterregio’s gebruikt worden voor hun identificatie. In bovenstaande gevallen zouden de beide motieven echter niet gedetecteerd worden met bestaande motief detectie algoritmen, enerzijds omdat de betrokken species evolutionair te nauw gerelateerd zijn, anderzijds omdat deze motieven slechts aanwezig zijn in een beperkt aantal promoterregio’s. Daarom ontwikkelde we een aangepaste versie van het MotifSampler algoritme dat in staat is om niche- of species specifieke regulatorische motieven te identificeren.

De methoden ontwikkeld tijdens dit doctoraat zijn allen toegespitst op de identificatie van regulatorische motieven. Vermits dergelijke motieven gedegenereerd zijn en aanwezig zijn in een beperkt aantal genen, kunnen de ontwikkelde methoden ook toegepast worden voor biologische vraagstukken die dezelfde beperkingen vertonen. We illustreren het ruime toepassingsgebied van onze methoden door de detectie van doelwitgenen van regulatorisch RNA enerzijds en de identificatie van kleine signaalpeptiden anderzijds.

Abbreviations

ABC	ATP-binding cassette
AMP	antimicrobial peptides
Ara4N	4-amino-4-deoxy-L-arabinose
BLAST	Basic Local Alignment Search Tool
ChIP	Chromatin ImmunoPrecipitation
CRP	cAMP receptor protein
DNA	deoxy-ribonucleic acid
EM	Expectation-Maximization
FACS	fluorescence-activated cell sorter
FDR	false discovery rate
FN	false negatives
FP	false positives
FRN	fumarate and nitrate reduction protein
HMM	Hidden Markov Model
HTH	Helix-Turn-Helix
LL	log-likelihood score
LPS	Lipopolysaccharide
MAST	Motif Alignment and Search Tool
MEME	Multiple EM for Motif Elicitation
mRNA	messenger RNA
Opp	oligopeptide permease
pEtN	phosphoethanolamine
PSSM	Position Specific Scoring Matrix
PWM	Position Weight Matrix
RNA	ribonucleic acid
RNAP	RNA polymerase
rRNA	ribosomal RNA
SCV	*Salmonella*-containing vacuole
SENS	sensitivity
SPEC	specificity
sRNA	regulatory RNA
TP	true positives

Nederlandse Samenvatting

1 Transcriptionele regulatie in prokaryoten

1.1 Context van de thesis

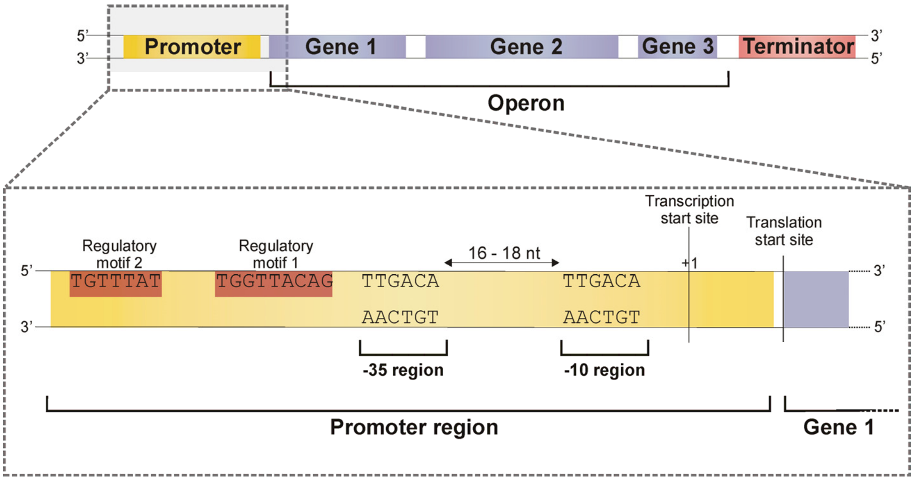
Ondanks het feit dat bio-informatica een recente gelanceerde term is, wordt het reeds tientallen jaren toegepast. Het gebruik van wiskundige modellen om biologische fenomenen te verklaren is inderdaad niet nieuw. Het gebruik van dergelijke modellen kende een enorme toename tijdens de jaren negentig door de ontwikkeling van nieuwe hoge doorvoer-data zoals microroosters, chromatine immunoprecipitatie technologie, proteoom en metaboloom analyse op genoomwijd niveau, etc. Waar een aantal jaren geleden elk gen of proteïne als aparte entiteit werd beschouwd, laten deze nieuwe technologieën toe om een groot aantal genen tegelijkertijd te analyseren. Een gen of proteïne wordt bestudeerd als deel van een complex netwerk. Het modelleren van dergelijke netwerken is het ultieme doel van systeembiologie.

Veel algoritmen voor systeembiologie zijn toegepast op eukaryote modelorganismen (bv. gist). Nochtans liggen er nog heel wat uitdagingen te wachten in het veld van de microbiologie. Systeemmicrobiologie legt de nadruk op het begrijpen van de verschillende bouwstenen en het dynamisch gedrag van bacteriële genetische netwerken. Een eerste uitdaging hierbij ligt in de reconstructie van de verschillende basisnetwerkstructuren (transcriptioneel, proteïne-interactie, proteoom, …). Vermits in bacteriën de expressie van een gen in sterke mate bepaald wordt door de transcriptiesnelheid, leggen we in deze thesis de nadruk op transcriptionele regulatorische netwerken. Regulatie op transcriptioneel niveau wordt gecontroleerd door regulatorproteïnen die binden met specifieke korte DNA-sequenties in de promoterregio's (i.e. regulatorische motieven) en op die manier de transcriptie van een of meerdere genen stimuleren of verhinderen. Aangezien regulatorische motieven de bouwstenen zijn van transcriptionele regulatorische netwerken, vormen ze een belangrijke databron voor netwerkinferentie.

1.2 Basiselementen voor transcriptie

Transcriptie is het biologisch proces waarbij DNA overgeschreven wordt tot RNA. Essentieel in dit transcriptieproces is de samenstelling van de promoterregio. In de meest basale vorm vereist transcriptie enkel het binden van het RNA polymerase met de -10 en -35 regio van de bacteriële promoter (σ^{70}-specifiek), waarna het DNA overgeschreven wordt tot RNA.

Naast de herkenningsplaatsen voor het RNA polymerase bevat de promoterregio ook korte geconserveerde DNA-sequenties die dienst doen als bindingsplaats voor regulatorproteïnen. Dergelijke regulatorproteïnen zijn essentieel in de controle van genexpressie vermits zij de binding van het RNA polymerase met de promoterregio verhinderen (repressor) of vergemakkelijken (activator). De bindingsplaatsen voor deze proteïnen worden regulatorische motieven genoemd.



Figuur N.1: **Overzicht van een prokaryoot operon.** In dit voorbeeld bestaat het operon uit drie genen die allemaal onder controle staan van de promoterregio van gen 1. In gedetailleerde beeld van de promoter van gen 1 zijn de -10 en -35 regio aangeduid (consensus sequenties voor de σ^{70}-factor). Posities zijn relatief berekend ten opzichte van de translatiestart. Stroomopwaarts van deze geconserveerde gebieden zijn nog twee hypothetische regulatorische motieven weergegeven.

1.3 Regulatorische motieven

Regulatorische motieven spelen een bepalende rol in transcriptionele regulatie. De exacte locatie en de bindingsaffiniteit van regulatorproteïnen voor deze motieven bepalen in belangrijke mate de expressie van een gen. Deze motieven kunnen op verschillende manieren voorgesteld worden. De consensussequentie is de meest eenvoudige weergave en geeft voor elke positie van het regulatorisch motief het meest voorkomende nucleotide weer (eventueel m.b.v. gedegenereerde symbolen). Een meer geavanceerde manier om een regulatorisch motief weer te geven is door middel van een matrix model. Voor elke positie in het motief kan de probabiliteit weergegeven worden waarmee een bepaald nucleotide op die plaats waargenomen wordt. Een derde representatiewijze van een regulatorisch motief is een motieflogo en is gebaseerd op deze matrixweergave. Hierbij wordt voor elke positie de frequentie van een specifiek nucleotide voorgesteld met zijn overeenkomstig symbool (A, C, G of T), waarbij de hoogte van het symbool evenredig is met de frequentie van het overeenkomstig nucleotide.

Er is reeds heel wat tijd geïnvesteerd in de computationele detectie van dergelijke regulatorische motieven. Over het algemeen kan men hierbij twee grote strategieën onderscheiden. De eerste benadering is gebaseerd op de hypothese dat genen die co-gereguleerd zijn, vermoedelijk ook eenzelfde transcriptioneel regulatiemechanisme (i.e. regulatorische motieven) vertonen. Hierbij kan men een onderscheid maken tussen deterministische en probabilistische algoritmen. Eventueel kunnen extra informatiebronnen (microroosterdata, ChIP-chip data, etc.) gebruikt worden om de regulatorische motieven op een meer betrouwbare manier te identificeren. De beschikbaarheid van volledige genoomsequenties opende echter nieuwe perspectieven voor de detectie van regulatorische motieven. Door het vergelijken van promoterregio’s van orthologen kunnen geconserveerde regulatorische motieven geïdentificeerd worden. De onderliggende hypothese hierbij is dat evolutie de biologische relevante sequenties bewaart, ook in promoterregio’s. Gezien regulatorische motieven fungeren als bindingsplaats voor regulatorproteïnen zullen zij doorheen de evolutie bewaard blijven (zie deel 2).

1.4 Transcriptionele regulatorische netwerken

Het transcriptioneel regulatorisch netwerk van een organisme geeft een overzicht van welk regulatorproteïne bindt op welke promoterregio, en wat het globale effect is van al deze interacties op de expressie van alle genen. Dergelijke regulatorische netwerken kunnen voorgesteld worden m.b.v. een “directed graph” waarbij transcriptiefactoren en doelwitgenen toegewezen worden aan bepaalde knooppunten in het netwerk, en de verbindingen tussen de verschillende knooppunten de mogelijke interacties weergeven. Door de vooruitgang in hoge-doorvoer technologieën en de beschikbaarheid van verschillende genoomsequenties werd het mogelijk om al deze verschillende databronnen te integreren en zodanig het transcriptioneel regulatorisch netwerk te reconstrueren. Een alternatief voor deze puur “datagedreven” reconstructie van regulatorische netwerken is de kennisgedreven reconstructie waarbij reeds goed gekarakteriseerde regulatorische netwerken als startpunt gebruikt worden in het algoritme.

1.5 Overzicht van de thesis

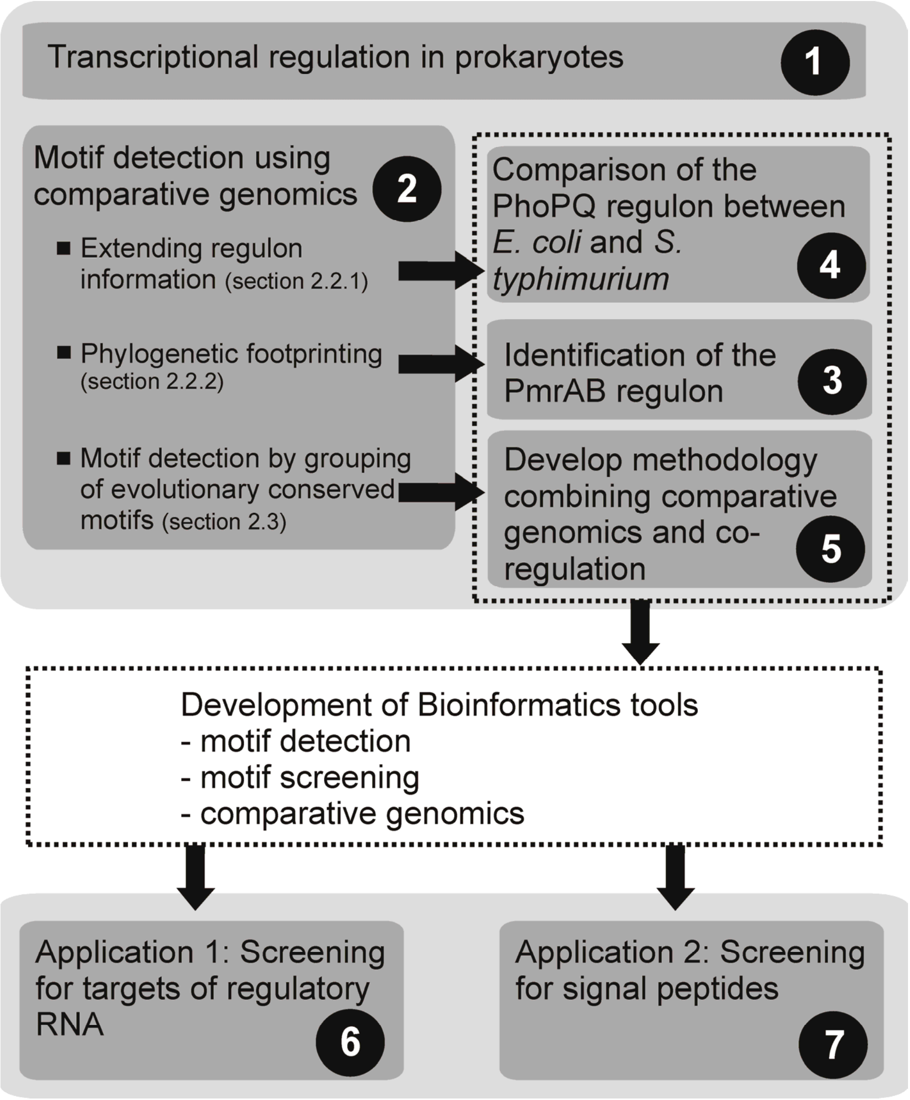
Bacteriën zijn dynamische organismen die in staan zijn zich aan te passen aan uiteenlopende omgevingsomstandigheden. Om in staat te zijn hun cellulaire systeem voortdurend aan te passen aan de wisselende omgevingsomstandigheden, zijn deze organismen uitgerust met flexibele regulatorische netwerken.

In bacteriën wordt de expressie van een gen in sterke mate bepaald door transcriptiesnelheid. Daarom leggen we in deze thesis de nadruk op regulatorische systemen die de transcriptie controleren. De belangrijkste bouwsteen van een transcriptioneel regulatorisch netwerk is de promoterregio van een gen. Dit gebied bevat immers de bindingsplaatsen voor regulatorproteïnen die de expressie van het overeenkomstige gen controleren. De voorbije jaren zijn reeds heel wat inspanningen geleverd met betrekking tot de computationele identificatie van dergelijke regulatorische bindingsplaatsen, wat leidde tot een uiteenlopende aanbod van motiefdetectie-algoritmen [**hoofdstuk 1**]. De beschikbaarheid van de volledige genoomsequenties van diverse species biedt echter nieuwe mogelijkheden voor motiefdetectie. Vertrekkend van de hypothese dat regulatorische motieven biologisch functionele sequenties zijn en dus in de evolutie bij voorkeur bewaard blijven, biedt dit de mogelijkheid om motieven te identificeren via vergelijking van orthologe promoterregio's uit verschillende species ("phylogenetic footprinting") [**hoofdstuk 2**].

In deze thesis gebruiken we het idee van "phylogenetic footprinting" om een duidelijker beeld te krijgen van twee regulatiesystemen die van belang zijn voor het infectiemechanisme van *Salmonella typhimurium*. Via een bio-informatica analyse identificeerden we nieuwe biologisch relevante genen die betrokken zijn in het PmrAB regulatie systeem [**hoofdstuk 3**]. Het PhoPQ regulatie systeem werd verder ontrafeld door gebruik te maken van een combinatie van expressie en motiefdata. Uit een vergelijking van de samenstelling van het PhoPQ regulatiesysteem tussen *S. typhimurium* and *Escherichia coli* concludeerden we dat de samenstelling van dit regulatiesysteem erg flexibel is. Deze waarneming geeft een mogelijke verklaring voor de uiteenlopende fenotypes die geobserveerd worden voor twee evolutionair nauw verwante species [**hoofdstuk 4**].

Voor beide bovenvermelde regulatorische systemen beschikken we over een beperkte hoeveelheid prior informatie. Indien men echter dergelijke motieven wil identificeren zonder prior informatie kan statistische overrepresentatie van motieven in promoterregio's gebruikt worden voor hun identificatie. In bovenstaande gevallen zouden de beide motieven echter niet gedetecteerd worden met bestaande motief detectie algoritmen, enerzijds omdat de betrokken species evolutionair te nauw gerelateerd zijn, anderzijds omdat deze motieven slechts aanwezig zijn in een beperkt aantal promoterregio's. Daarom ontwikkelde we een aangepaste versie van het MotifSampler algoritme dat in staat is om niche- of species specifieke regulatorische motieven te identificeren [**hoofdstuk 5**].

De methoden ontwikkeld tijdens dit doctoraat zijn allen toegespitst op de identificatie van regulatorische motieven. Vermits dergelijke motieven gedegenereerd zijn en aanwezig zijn in een beperkt aantal genen, kunnen de ontwikkelde methoden ook toegepast worden voor biologische vraagstukken die dezelfde beperkingen vertonen. We illustreren het ruime toepassingsgebied van onze methoden door de detectie van doelwitgenen van regulatorisch RNA enerzijds [**hoofdstuk 6**] en de identificatie van kleine signaalpeptiden anderzijds [**hoofdstuk 7**].


Transcriptional regulation in prokaryotes
1
Motif detection using comparative genomics
2
■ Extending regulon information (section 2.2.1)
■ Phylogenetic footprinting (section 2.2.2)
■ Motif detection by grouping of evolutionary conserved motifs (section 2.3)
Comparison of the PhoPQ regulon between E. coli and S. typhimurium
4
Identification of the PmrAB regulon
3
Develop methodology combining comparative genomics and co-regulation
5
Development of Bioinformatics tools
- motif detection
- motif screening
- comparative genomics
Application 1: Screening for targets of regulatory RNA
6
Application 2: Screening for signal peptides
7

Figuur N.2: **Overzicht van de thesis.** Nummers van de hoofdstukken zijn aangeduid in de zwarte cirkels. Hoofdstuk 1 en 2 geven een literatuuroverzicht van transcriptionele regulatie en motiefdetectie in bacteriën. Hoofdstuk 3 tot 5 beschrijven de resultaten van regulatorische motiefdetectie in bacteriën waarbij gebruik wordt gemaakt van comparatieve genoomanalyse. In hoofdstuk 6 en 7 worden de methodes ontwikkeld in de vorige drie hoofdstukken toegepast om biologische problemen op te lossen die gelijkenissen vertonen met regulatorische motiefdetectie.

2 Motiefdetectie met behulp van comparatieve genoomanalyse

2.1 Comparatieve genoomanalyse voor motiefdetectie: toepassingen

Een eerste toepassing van de comparatieve genoomanalyse spitst zich toe op de ontwikkeling van methoden die toelaten om de kennis van een bepaald regulatorisch systeem uit te breiden op basis van een beperkte hoeveelheid informatie m.b.t. de bindingsplaats van het betrokken regulatorproteïne. Deze benadering werd toegepast voor verschillende regulatiesystemen in een uitgebreide waaier van bacteriële species [82,176,194,195,259]. Op een gelijkaardige manier werd in dit doctoraat een methodologie ontwikkeld om het PhoPQ regulon te vergelijken tussen twee evolutionair gerelateerde bacteriën *E. coli* and *S. typhimurium* [178] (beschreven in hoofdstuk 4). In een later stadium werd comparatieve genoomanalyse ook gebruikt om *de novo* motieven te detecteren (i.e. “phylogenetic footprinting”). Hierbij worden de klassieke motiefdetectie-algoritmen gebruikt voor het identificeren van nieuwe motieven in sets van orthologe promoterregio’s [167]. Beide strategieën kunnen echter ook gecombineerd worden. Het resultaat van een “phylogenetic footprinting” stap kan gebruikt worden om op genoomwijde schaal de promoterregio’s te controleren op de aanwezigheid van dit regulatorisch motief, en aldus het regulon te identificeren [221]. Ook de omgekeerde benadering is mogelijk, waarbij men start met een genoomwijde detectiestap op basis van een gekend motiefmodel, gevolgd door een “phylogenetic footprinting” stap als *in silico* bewijs voor de biologische relevantie van de geïdentificeerde doelwitgenen [169]. Wij pasten een gelijkaardige strategie toe voor de identificatie van het PmrAB regulon in *S. typhimurium* (beschreven in hoofdstuk 3).

2.2 Motiefdetectie door groeperen van evolutionair geconserveerde motieven

Het initiële idee van “phylogenetic footprinting” werd slechts toegepast op een enkele set van orthologe promoterregio’s waarin het regulatorisch motief geïdentificeerd wordt. Gebaseerd op deze “phylogenetic footprinting” ontwikkelde men het idee dat motieven die gedetecteerd worden in één set van orthologe promoterregio’s ook teruggevonden zullen worden in andere sets van orthologe promoters. De uitdaging van deze geavanceerde methoden van “phylogenetic footprinting” is het groeperen van al deze regulatorische motieven – resulterend uit de verschillende “phylogenetic footprinting” stappen – die sterk op elkaar gelijken en dus de bindingsplaats vormen voor hetzelfde regulatorproteïne. Een eerste type van algoritmen vereist een set van cogereguleerde genen, terwijl een tweede type van algoritmen toegepast kan worden op genoomwijde schaal. De idee achter beide algoritmen is echter dezelfde: de eerste stap is steeds een "phylogenetic footprinting" stap die resulteert in een set van evolutionaire geconserveerde regio's in promoters, terwijl de tweede stap de geconserveerde promoters groepeert die een gelijkaardig regulatorisch motief bevatten. Een voorbeeld van het eerste type algoritme is PhyloCon [295]. Wij ontwikkelden een gelijkaardige methode die in staat is om regulatorische motieven te detecteren in een set van cogereguleerde genen waarbij de motieven slechts aanwezig zijn in een erg beperkte set van promoterregio's. Terwijl de twee hierboven vermelde algoritmen nog steeds een set van cogereguleerde genen vereisen (vaak afgeleid van experimentele data), werden ook algoritmen ontwikkeld die op een genoomwijde schaal kunnen toegepast worden, en waarvoor men enkel afhankelijk is van de beschikbaarheid van sequentiedata. Voor de clusteringstap (i.e. de groepering van geconserveerde promoterregio's die een gelijkaardig motief bevatten) ontwikkelde drie onderzoeksgroepen een algoritme gebaseerd op Gibbs Sampling [120,214,287]. Een alternatief voor deze Gibbs Sampling clustering is een "neighbor-joining" algoritme [3,296].

2.3 Integratie van fylogenetische afstanden

De algoritmen beschreven in 2.2 combineren coregulatie en fylogenetisch informatie op een sequentiële manier. Recent werden nieuwe algoritmen ontwikkeld die gelijktijdig gebruik maken van beide databronnen (sequentiedata en fylogenetische informatie). De meest eenvoudige implementatie kan enkel toegepast worden op twee verschillende species [209]. Daarnaast werden ook nog verschillende bestaande motiefdetectie-methoden – die vertrekken van sets van cogereguleerde genen – aangepast om ook fylogenetisch informatie in rekening te brengen: PhyME en EMnEM zijn gebaseerd op het klassieke "Expectation-Maximization" algoritme en PhyloGibbs bouwt verder op het Gibbs Sampling principe. Het meest geavanceerde algoritme werd ontwikkeld door Li en Wong [144]. Dit Gibbs Sampling algoritme maakt eveneens gebruik van fylogenetische informatie en coregulatie, maar brengt bovendien ook de evolutionaire afstand tussen de verschillende species in rekening.

3 Identificatie van het PmrAB regulon

3.1 Inleiding

De meest voor de hand liggende methode om een transcriptioneel netwerk te identificeren, is met behulp van experimentele analyses. Gezien het tijdrovende en arbeidsintensieve karakter van dergelijke analyses, werden methoden ontwikkeld die toelaten om regulatorische systemen te ontrafelen via een *in silico* benadering. Potentiële doelwitgenen van een regulatorisch systeem kunnen geïdentificeerd worden op basis van de aanwezigheid van het overeenkomstige regulatorisch motief in zijn promoterregio. Dergelijke genoomwijde zoektocht naar regulatorische motieven zal echter resulteren in een hoog aantal valse positieve doelwitgenen. De ontwikkeling van methoden voor comparatieve genoomanalyse gecombineerd met de toegenomen beschikbaarheid van bacteriële genomen laat toe om na te gaan of het regulatorisch motief ook bewaard is in promoterregio's van orthologe genen wat een extra evidentie zou vormen voor de biologische relevantie van het motief. Vertrekkend van een beperkte hoeveelheid experimentele informatie m.b.t. het PmrAB regulatorisch systeem, gebruikten we comparatieve genoomanalyse om potentiële nieuwe doelwitgenen te identificeren.

Het PmrAB regulatorisch systeem is vereist voor resistentie tegen kationische antimicrobiële peptiden en Fe^{3+} gecontroleerde celdoding [97,125,222,304,315], en daardoor essentieel voor de virulentie in *S. typhimurium*. Deze resistentie is voornamelijk het resultaat van PmrAB gecontroleerde wijzingen in het polyliposaccharide celmembraan. Bovendien is dit regulatorisch systeem ook onrechtstreeks afhankelijk van de concentratie aan Mg^{2+} gezien PmrAB onder controle staat van het PhoPQ regulatorisch systeem [77,88,246,247] via PmrD [125,134]. Deze Mg^{2+} afhankelijkheid is vooral van belang in intracellulaire omgeving (e.g. in macrofagen) [185] terwijl de Fe^{3+} regulatie van het PmrAB systeem vooral van belang zou zijn in extracellulaire condities [39].

In onze studie van het PmrAB regulatorisch systeem maken we gebruik van genoomwijde zoekmethoden voor regulatorische motieven gecombineerd met comparatieve genoomanalyse. In een eerste stap zoeken we op genoomwijde schaal naar potentiële regulatorisch motieven in de promoterregio's van *S. typhimurium*. Potentiële doelwitgenen worden in een tweede stap – indien mogelijk – gevalideerd m.b.v. "phylogenetic footprinting". Dit werk werd uitgevoerd in samenwerking met het Centrum voor Microbiële en Plant Gentica (Prof. J. Vanderleyden, Dr. S. De Keersmaecker) en is gepubliceerd in Genome Biology [160].

3.2 Twee-staps "phylogenetic footprinting" procedure

Een Gibbs sampling algoritme werd toegepast op een beperkt aantal experimenteel geverifieerde PmrAB target genen (*ugd*, *pmrC*, *pmrG*) om het PmrA motief te detecteren. Het resulterende PmrA motiefmodel vertoonde sterke overeenkomsten met de reeds experimenteel geverifieerde bindingsplaatsen [2,303]. Dit motiefmodel werd gebruikt om op genoomwijde schaal het PmrA motief te detecteren in de promoterregio's van *S. typhimurium*. Aangezien het PmrAB regulatorisch systeem goed bewaard is in evolutionair gerelateerde bacteriën, werd gebruik gemaakt van "phylogenetic footprinting" voor de validatie van de *in silico* geïdentificeerde regulatorische motieven. De ontwikkelde twee-staps "phylogenetic footprinting" procedure maakt in een eerste stap gebruik van een Gibbs Sampling algoritme voor de identificatie van potentiële "seeds" voor de locale alignering. In een tweede stap worden deze "seeds" gebruikt voor de aanmaak van een locale multipele alignering. Uit deze alignering kan afgeleid worden of het PmrA regulatorisch motief bewaard is in evolutionaire verwante species, wat een extra validatie vormt voor een potentieel doelwitgen.

Er zijn verschillende redenen waarom we terugvallen op de alignering van orthologe promoterregio's in plaats van een lijst met hoogst scorende motieven bekomen via MotifSampler. Eerst en vooral stelden we vast dat de intergenische sequenties tussen orthologe promoters vaak sterk op elkaar gelijken (omwille van evolutionair sterk verwante species). Hierdoor is niet enkel het biologische relevante motief bewaard, maar ook de flankerende sequenties. Indien een alignering duidelijk maakt dat niet enkel het regulatorische motief bewaard is gebleven, maar ook de omgeving van het motief, is het resulterende motief betrouwbaarder dan zonder conservering van de flankerende gebieden. Ten tweede is het algoritme dat we gebruiken voor de identificatie van de "seeds" (i.e. MotifSampler) een stochastisch algoritme ontwikkeld voor de identificatie van niet-gerelateerde sequenties (i.e. geen evolutionaire relatie). Zoals hierboven aangehaald is daardoor ook de omgeving van het motief bewaard gebleven. Er is met andere woorden geen enkele garantie dat de lijst met hoogst scorende motieven ook effectief het biologische relevante motief bevat. Indien we in onze methodologie echter de resultaten van het motiefdetectie algoritme enkel als "seed" gebruiken voor deze alignering, zullen al de topscorende motieven resulteren in eenzelfde alignering.

3.3 Biologische interpretatie

Potentiële regulatorische motieven werden geïdentificeerd in de intergenische gebieden van genen waarvan de functie gerelateerd is aan de werking van het PmrAB regulatorisch systeem (i.e. genen die coderen voor celmembraanproteïnen, flagellensynthese, wijziging van celmembraan, etc.). Indien orthologen in gerelateerde species teruggevonden werden, was het regulatorisch motief in vele gevallen bewaard in de orthologe promoters. Het ontbreken van een PmrA motief in orthologe promoterregio's kan wijzen op een unieke PmrAB regulatie van dit doelwitgen in *S. typhimurium*. Voor dergelijke motieven kan onze methodologie geen extra evidentie leveren voor de biologische relevantie.

Naast de gekende PmrAB gereguleerde genen (*pmrH*, *pmrC*, *ugd*) werden ook enkele nieuwe potentiële doelwitgenen gesuggereerd via onze *in* *silico* methode. Vier van deze nieuwe doelwitgenen werden geselecteerd voor biologische validatie i.e. *yibD*, *aroQ, mig-13* and *sseJ.* Expressie-analyses m.b.v. GFP-reporterfusies werden uitgevoerd voor wildtype en *pmrA* mutanten met variërende concentraties aan Mg^{2+} en Fe^{3+}. Met uitzondering van *sseJ* vertoonden alle nieuwe potentiële doelwitgenen een duidelijke afhankelijkheid van Mg^{2+} of Fe^{3+}. Voor *sseJ* was dit effect enkel zichtbaar voor één van de vijf condities.

Naast het testen van nieuwe PmrAB gereguleerde genen werd ook een set van mutante PmrA boxen aangemaakt met behulp van plaatsspecifieke mutagenese in het eerste deel van het PmrA motief, waarbij vooral de derde en vijfde positie van de eerste site van het PmrA motief essentieel zijn voor de activering door het PmrA regulatorproteïne. Voor de mutaties ter hoogte van de andere posities in de PmrA box was er nog steeds expressie van het PmrA gereguleerde gen, weliswaar in mindere mate.

3.4 Conclusie

We hebben aangetoond dat onze *in silico* methodologie in staat is om op een betrouwbare wijze nieuwe PmrAB doelwitgenen te identificeren. Hoewel het niet uitgesloten is dat onze methodologie vals positieve resultaten zal opleveren, schept het heel wat mogelijkheden om het genetisch netwerk in *S. typhimurium* te ontrafelen dat verantwoordelijk is voor het virulente karakter van *Salmonella* stammen. Onze methodologie kon vier nieuwe en biologisch relevante PmrAB doelwitgenen voorspellen.

4 Vergelijking van het PhoPQ regulon

4.1 Inleiding

Het vierde hoofdstuk behandelt de identificatie van het PhoPQ regulon in *E. coli* en *S. typhimurium.* Waar de identificatie van het PmrAB regulon enkel gebaseerd is op sequentiedata, gebruiken we voor het PhoPQ regulon een combinatie van sequentie- en microroosterdata. De voornaamste reden hiervoor is de onduidelijkheid met betrekking tot de karakteristieken van het PhoP motiefmodel. Het PhoPQ regulatorisch systeem is bewaard in zowel *E. coli* als *S. typhimurium*. Het PhoPQ systeem is in beide organismen verantwoordelijk voor het waarnemen van de extracellulaire Mg^{2+} en Ca^{2+} concentratie. De aan- of afwezigheid van deze ionen wordt waargenomen door het PhoQ proteïne, dat op zijn beurt de PhoP transcriptionele regulator activeert. In *S. typhimurium* en andere Gram-negatieve bacteriën reguleert dit twee-componentsysteem echter ook genen die betrokken zijn in virulentie [88]. Het waarnemen van de concentratie aan Mg^{2+} laat de pathogene bacterie immers toe om zijn subcellulaire locatie te bepalen (e.g. binnenin een macrofaag, darmflora). Indien de bacterie aanwezig is binnenin een macrofaag moeten immers een aantal virulentiefactoren geactiveerd worden. Naar aanleiding van de uiteenlopende eigenschappen van het regulatiesysteem vergeleken we de samenstelling van het PhoPQ regulon tussen beide organismen, wat resulteerde in een erg beperkte overlap tussen beide species. Deze analyse werd uitgevoerd in samenwerking met de Universiteit van Washington (Dr. W. Navarre, Prof. F. Fang) en het Sidney Kimmel Cancer centre (Prof. M. McClelland), die ons de nodige microroosterdata bezorgden. Dit hoofdstuk is gepubliceerd in Journal of Molecular Evolution.

4.2 Identificatie van het PhoPQ regulon

Voor de identificatie van PhoPQ gereguleerde genen in *E. coli* maakten we gebruik van de microroosterdata beschreven in Mingawa *et al.* [175]. Zij identificeerden 219 genen die positief gereguleerd worden door het PhoPQ regulatiesysteem. De 219 genen zijn gelegen in 193 operons. Voor de identificatie van dit regulon in *S. typhimurium* vergeleken we de genexpressie van een PhoP knock-out mutant met een PhoP constitutieve mutant. Op basis van zorgvuldig bepaalde selectiecriteria resulteerde dit in de identificatie van 189 operons die positief gereguleerd worden door het PhoPQ systeem. De kwaliteit van de gebruikte microroostergegevens werd bevestigd door de aanwezigheid van experimenteel geverifieerde PhoPQ-afhankelijke genen in de subset van positief gereguleerde genen.

Om het onderscheid te kunnen maken tussen direct en indirect PhoPQ gereguleerde operons werd in elke promoterregio van de subset van positief gereguleerde genen op zoek gegaan naar het PhoP motiefmodel. Initieel werd het PhoPQ motiefmodel beschreven als een directe herhaling van een hexanucleotide (TGTTTA) van elkaar gescheiden door 5 nucleotiden. Recent werd het motiefmodel verder verfijnd waarbij aangetoond werd dat het motiefmodel ook promoterregio's kan binden die meer variatie vertonen in de bindingssequentie [141,175,311]. Wij gebruikten een combinatie van de verschillende karakteristieken voor de identificatie van de direct PhoPQ-gereguleerde operons. Dit resulteerde in 42 en 34 direct gereguleerde operons in *S. typhimurium* en *E. coli* respectievelijk.

4.3 Overlap PhoPQ regulon tussen *E. coli* en *S. typhimurium*

Uit de vergelijking van beide datasets blijkt dat slechts 13 operons gemeenschappelijk gereguleerd worden door het PhoPQ regulatiesysteem in *E. coli* en *S. typhimurium.* Slechts 2 van de 13 operons zijn ook direct gereguleerd door PhoPQ (i.e. *phoPQ* en *slyB*). Voor 1 van de 13 operons werd een motief gevonden in *S. typhimurium*, maar evidentie voor deze directe regulatie werd niet gevonden in *E .coli*. Een gelijkaardige analyse werd ook uitgevoerd voor operons die enkel differentieel tot expressie kwamen in ofwel *E. coli* ofwel *S. typhimurium*. Het grote aantal speciesspecifieke operons die onder controle staan van het PhoPQ systeem suggereren dat het PhoPQ regulatiesysteem sterk gespecialiseerd is in deze organismen. De beperkte overlap tussen beide regulons kan verklaard worden door middel van de cruciale rol die het PhoPQ systeem speelt in het virulente fenotype van *S. typhiumurium*. Dit fenotype is volledig afwezig bij de niet-pathogene *E. coli* K12 [88] stam.

4.4 Biologische resultaten

Op basis van de functionele annotatie van de verschillende genen in *S. typhimurium* en *E. coli* achterhaalden we voor elk PhoPQ gereguleerd gen de functionele omschrijving uit de *S. typhi* Sanger [199] en EcoCyc [128] databank. Op basis van deze data werd nagegaan welke functionele klassen een significante oververtegenwoordiging van PhoP-gereguleerde genen bevatten. In *S. typhimurium* zijn de functioneel aangerijkte klassen voornamelijk betrokken in “het centraal intermediair metabolisme”, “aanmaak en wijziging van het celmembraan”, “transport van kationen” en “sensitiviteit voor geneesmiddelen”. Voor *E. coli* zijn de meest aangerijkte klassen betrokken in “algemeen metabolisme” en “cel- en membraanstructuur”. Ondanks de beperkte overlap in regulonsamenstelling blijkt het PhoPQ regulatiesysteem toch een gelijkaardige functie bewaard te hebben in beide organismen.

4.5 Conclusie

Onze analyse toont aan hoe een goed geconserveerd regulatiesysteem dat beantwoordt aan eenzelfde extracellulair signaal in twee organismen, geïntegreerd kan worden in verschillende cellulaire reactiewegen tijdens een relatief korte tijdspanne. Deze opname van nieuwe genen in het regulatiesysteem kan een verklaring vormen voor de enorme flexibiliteit van bacteriële genetische netwerken die de bacterie toelaten om zich aan te passen aan snel wijzigende omgevingscondities. In *S. typhimurium* zijn vermoedelijk extra genen die bijdragen tot het virulente fenotype onder controle komen te staan van het PhoPQ regulatiesysteem, terwijl dit niet gebeurd is in de niet-pathogene *E. coli* stam.

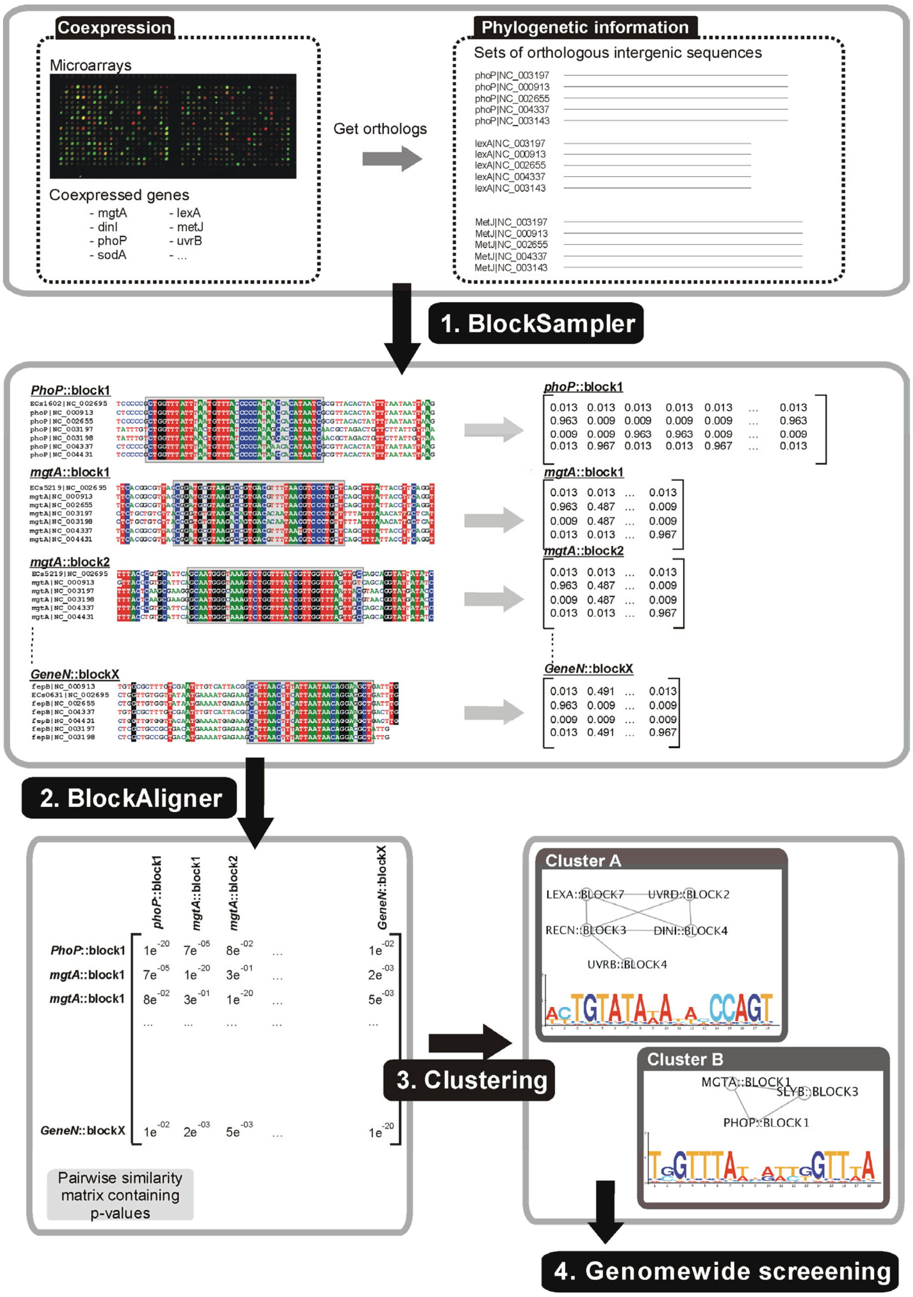
5 Robuuste detectie van regulatorisch motieven door gebruik van fylogenetisch informatie

5.1 Inleiding

In de vorige twee hoofdstukken werden twee regulatorische systemen geïdentificeerd op basis van een *in silico* methode. In beide gevallen was echter een beperkte hoeveelheid informatie beschikbaar met betrekking tot het respectievelijke motiefmodel. Identificatie van beide regulons zou immers heel wat moeizamer verlopen indien geen prior informatie beschikbaar is over het regulatorisch motief. In de studie naar het PmrAB regulon constateerden we immers dat intergenische regio's vaak sterk geconserveerd zijn tussen evolutionair gerelateerde species. Hierdoor is niet enkel het biologisch relevante motief geconserveerd, maar ook de flankerende sequenties, wat de *de novo* detectie van het motief erg bemoeilijkt. Voor het PhoPQ regulon bemerkten we dat een regulatorisch motief vaak evolutionair geconserveerd is in slechts een erg beperkte subset van genen.

Er zijn reeds een uitgebreid aantal algoritmen ontwikkeld voor de identificatie van regulatorische motieven (e.g. [14,110,113,136,146,268, 284,301]). Het basisidee hierbij is dat genen die cogereguleerd zijn vermoedelijk eenzelfde regulatorisch motief bevatten in hun promoterregio. Het rendement van deze algoritmen daalt echter snel wanneer de signaal-ruis verhouding kleiner wordt (i.e. het aantal promoterregio's dat het regulatorisch motief bevat is klein vergeleken met de regio's die het niet bevatten), zoals het geval voor het PhoPQ regulon. Dergelijke lage signaal-ruis verhoudingen komen echter frequent voor in biologische data (e.g. data afgeleid uit microroosterdata).

Om deze problemen op te vangen ontwikkelden we een methode die fylogenetische informatie combineert met co-expressie data voor *de novo* detectie van regulatorische motieven. Deze fylogenetische informatie is geïntroduceerd in de methodologie via een "phylogenetic footprinting" stap (zie hoofdstuk 2) gebaseerd op Gibbs Sampling. Onze methode is in staat om regulatorische motieven te identificeren die slechts aanwezig zijn in een erg beperkte set van cogereguleerde genen. Via een vergelijking van de resultaten van onze methode met andere gelijkaardige algoritmes op een testdataset, tonen we aan dat onze methode in staat is om regulatorische motieven te detecteren op basis van genoomwijde microroosterdata (i.e. een set van genen die tot co-expressie komt, maar waar het motief slechts aanwezig is in een erg beperkte subset van genen). Dit werk is gepubliceerd in BMC Bioinformatics [179].



Figuur N.3: **Overzicht van de methodologie.** Input data: Op basis van microroosterdata worden genen geïdentificeerd die tot co-expressie komen. Orthologe sequenties worden geïdentificeerd. Stap 1: BlockSampler: "phylogenetic footprinting". Stap 2: BlockAligner: aligneren van de geconserveerde promoterregio's resulterend in een p-waarde voor elke paarsgewijze alignering. Stap 3: clustering van de geconserveerde regio's op basis van de paarsgewijze p-waarden, en aflijnen van het motief. Stap 4: "screening" van beschikbare promoterregio's.

5.2 Overzicht van de ontwikkelde methode

Onze methode vertrekt van een set van potentieel co-gereguleerde genen. In een eerste stap zoeken we de orthologen van al deze genen, en voeren een “phylogenetic footprinting” stap uit op elke set van orthologe promoterregio’s. Hiervoor ontwikkelde we een nieuw algoritme, BlockSampler, dat een uitbreiding vormt van het motiefdetectiealgoritme MotifSampler [268]. Hieruit verkrijgen we een set van geconserveerde gebieden (“blocks”) voor elke set van orthologe genen, waarin alle potentiële motieven teruggevonden kunnen worden. In een tweede stap worden alle geconserveerde gebieden tussen de verschillende sets van orthologe promoters paarsgewijs met elkaar gealigneerd. Hiervoor ontwikkelden we het algoritme BlockAligner, dat een lokale alignering maakt van twee geconserveerde promotergebieden met de Kullback-Leibler afstand als scoringsfunctie. In een derde stap worden alle geconserveerde gebieden die mogelijk een gelijkaardig motief bevatten gegroepeerd m.b.v. een “graph-based” clusteringalgoritme, en wordt het gemeenschappelijk motief afgelijnd binnen eenzelfde cluster. In een laatste stap wordt in alle promoterregio’s van potentieel cogereguleerde genen gezocht naar een overeenkomst met elk van de motiefmodellen die resulteren uit onze methode. Dit laat ons toe om regulatorische motieven terug te vinden die initieel door onze methode gemist werden.

5.3 Performantie op testdataset

In een eerste fase voerden we onze methode uit op samengestelde testdatasets (“golden standard”) met gekende bindingsplaatsen voor vier verschillende regulatorproteïnen, waarbij telkens een verschillend aantal randomsequenties toegevoegd werd om variatie te verkrijgen in de signaal-ruisverhouding (variërend tussen 4% en 18%). De performantie van onze methode werd geëvalueerd voor vier verschillende karakteristieken: de mate waarin het correcte motiefmodel als resultaat teruggegeven wordt, het aantal vals positieve motiefmodellen, sensitiviteit en specificiteit. Eenzelfde analyse werd ook doorgevoerd voor twee andere motiefdetectie-algoritmen: AlignACE, een motiefdetectie-algoritme dat enkel gebruikt maakt van promoterregio’s van een set van genen die tot co-expressie komen [113], en PhyloCon, een algoritme dat net als onze methode gebruik maakt van een combinatie van co-expressie en fylogenetische informatie [295]. Wanneer onze methode werd toegepast op de testdatasets, resulteerde dit voor drie van de vier motieven in een correct motiefmodel voor meer dan 90% van de datasets, zelfs wanneer een hoog aantal randomsequenties toegevoegd is aan de gekende motiefinstanties. Enkel voor het Fur-motief lag dit percentage lager (ongeveer 50%) wat vermoedelijk te wijten is aan het meer gedegenereerde karakter van dit motiefmodel. Het aantal vals positieven was erg beperkt. Sensitiviteit en specificiteit waren algemeen bekeken erg hoog.

Deze waarden bleken echter wel iets lager te liggen als het motief bestond uit twee geconserveerde halve sites die van elkaar gescheiden zijn door een niet-geconserveerde tussensequentie. Een gelijkaardige analyse voor AlignACE toonde aan dat onze methode duidelijk een betere performantie vertoont voor alle karakteristieken. Bovendien bleek AlignACE ook sterk gevoelig aan ruis (i.e. toevoegen van random sequenties aan de gekende motiefinstanties). Het derde algoritme, PhyloCon, vertoonde een betere performantie voor het gedegenereerde Fur motief, maar presteerde minder goed voor de drie andere motieven. Bovendien bleek het PhyloCon algoritme ook sterk afhankelijk van ruis. Daarnaast heeft PhyloCon ook bepaalde algoritmische beperkingen waardoor de correcte identificatie van regulatorisch motieven bemoeilijkt wordt.

5.4 Performantie op microroosterdata

Naast de samengestelde datasets hebben we onze methode ook getest op biologische datasets afgeleid van microroosterdata. Een eerste dataset bestaat uit 47 differentieel tot expressie gekomen genen tussen een constitutieve *pmrA* mutant en een *pmrA* deletiemutant van *S. typhimurium* [256]. Ondanks het feit dat het PmrA motief slechts aanwezig was in 5 van de 47 genen was onze methode in staat om een motiefmodel als resultaat te geven waarvan de consensus-sequentie zeer sterk gelijkend is op het biologisch gevalideerde PmrA motief [2,160,303].

Een tweede dataset was samengesteld op basis van de publicatie van Salmon *et al.* [231] waarbij de expressieprofielen van genen vergeleken worden bij de omschakeling van aerobe naar anaerobe omstandigheden. Deze omschakeling wordt op transcriptioneel regulatorisch niveau gecontroleerd door de FNR regulator [98]. In de set van differentieel tot expressie gekomen genen is het FNR motief slechts aanwezig in erg beperkte subset van genen (4 van de 83 genen). Ook hier is onze methode in staat om het correcte FNR motiefmodel als resultaat te geven.

5.5 Conclusie

In dit hoofdstuk hebben we een methode ontwikkeld die in staat is om meerdere regulatorische motieven te detecteren die niet statistisch overgerepresenteerd zijn in een set van cogereguleerde genen (o.a. regulatorproteïnen in een genetisch netwerk die slechts een beperkt aantal genen controleren). Hiervoor maken we optimaal gebruik van de combinatie van co-expressiegegevens en fylogenetische informatie. Via een vergelijking met twee andere motiefdetectie-algoritmen tonen we de robuustheid van onze methode aan. Als proefstuk tonen we aan dat vertrekkend van genoomwijde expressiedata (i.e. veel ruis) nog steeds de correcte motiefmodellen gedetecteerd worden.

6 Detectie van regulatorisch RNA doelwitgenen

6.1 Inleiding

Recente ontdekkingen hebben geleid tot de identificatie van verschillende RNA's met een andere functie dan boodschapper RNA's, transport RNA's of ribosomaal RNA's. Verschillende onderzoeken wijzen uit dat niet-coderende RNA's vaak een cruciale rol spelen in bacteriële regulatorische netwerken. Deze niet-coderende RNA's vervullen hun regulatorische functie hetzij via baseparing met het mRNA van het doelwitgen, hetzij via rechtstreekse binding van het sRNA met een proteïne waardoor de activiteit van het proteïne gewijzigd wordt. Wij concentreren ons in dit hoofdstuk op de detectie van doelwitgenen voor de eerste groep van sRNA's. Op dit moment zijn meer dan 60 sRNA's in *E. coli* gekend waarvan minstens één derde gebruik maken van dit DNA-baseparing regulatiemechanisme [10,291]. De biologische functies van de meeste sRNA's zijn nog onbekend. Een van de recent geïdentificeerde sRNA moleculen is *sraD*, gelegen in de buurt van het *luxS* gen, maar op de complementaire DNA-streng. Aanvankelijk werd vermoed dat LuxS een essentiële rol zou spelen in de biofilmvorming van *S. typhimurium*. Recente experimentele resultaten tonen echter aan dat het vermoedelijk niet het *luxS* gen zelf is dat hiervoor verantwoordelijk is, maar wel het gebied 5' stroomopwaarts. In dit gebied is recent het regulatorisch RNA *sraD* geïdentificeerd.

Als we een duidelijk beeld willen krijgen op de exacte functie van het *sraD* molecule, moeten we de doelwitgenen van dit sRNA identificeren. Verschillende algoritmen zijn ontwikkeld voor de identificatie van sRNA moleculen, maar algoritmes voor de identificatie van doelwitgenen van dit sRNA laten voorlopig echter nog op zich wachten. Op dit moment is er slechts één algoritme beschikbaar dat potentiële doelwitgenen voorspelt van sRNA's [271]. In dit hoofdstuk stellen we een *in silico* methode voor – gebaseerd op comparatieve genoomanalyse – die toelaat om potentiële doelwitgenen te voorspellen van sRNA's. We gebruiken deze benadering om de biologische functie van *sraD* te achterhalen.

6.2 Identificatie van het *sraD* sRNA

Het *sraD* RNA was op het moment van onze analyse enkel gekend in *E. coli*. Via comparatieve genoomanalyse van het *sraD* sRNA in evolutionair gerelateerde bacteriën (*E. coli*, *S. typhimurium, Y. pestis, Erwinia carotovora, Serratia marcescens*) werden de orthologe sRNA moleculen in deze organismen geïdentificeerd. In eerste instantie werden hiervoor de promoterregio's geisoleerd van *luxS*. Hierin werden vervolgens de *sraD* orthologen geïdentificeerd d.m.v. een Waterman-Smith alignering met *sraD* in *E. coli*. Met behulp van BlockSampler [179] werden drie geconserveerde gebieden geïdentificeerd in de set van orthologe *sraD* sequenties. De goede bewaring van deze gebieden doet vermoeden dat deze sequenties essentieel zijn in de identificatie en binding van het mRNA van doelwitgenen.

6.3 Bepalen van potentiële doelwitgenen

Zoals hierboven vermeld concentreren we ons in dit hoofdstuk op de sRNA's die hun regulatorische functie uitvoeren via baseparing met het mRNA van de doelwitgenen. De doelwitgenen van deze sRNA's worden *in silico* geïdentificeerd door in translatie-initiatieregio's te zoeken naar sequenties die complementair zijn met de sRNA sequenties. Voor het *sraD* sRNA gebruikten we echter niet de volledige sRNA sequentie om sequentiecomplementariteit te detecteren maar maakten we enkel gebruik van de geconserveerde regio's in het sRNA. Deze geconserveerde regio's zijn beschreven als positiespecifieke scoringsmatrices. Dit laat ons toe om MotifLocator [160] te gebruiken om de flankerende gebieden in de buurt van de translatie-initiatiestart te doorzoeken naar complementaire sequenties. De beste hit voor de eerste geconserveerde regio in *sraD* was *yijC*, een transcriptionele regulator. Voor de ortholoog van *yijC* in *E. coli*, *fabR,* is aangetoond dat het een essentiële rol speelt in de concentratie aan onverzadigde vetzuren in het celmembraan. Voor de tweede geconserveerde regio werd *metC* teruggevonden als meest waarschijnlijke doelwitgen. Een insertie-inactivatie van het *metC* gen in *S. typhimurium* leidde in muizen tot een geattenueerd virulentiefenotype. Beide potentiële doelwitgenen werden ook experimenteel geverifieerd in samenwerking met het Centrum voor Microbiële en Plant Genetica (Dr. S. Dekeersmaecker, Prof. J. Vanderleyden). Mutaties in beide genen leidden tot een verminderde biofilmvorming, wat onze *in silico* analyse ondersteunt. De directe interactie tussen het sRNA en de doelwitgenen kon echter nog niet aangetoond worden.

6.4 Conclusie

Hoewel de methodes voor de identificatie van regulatorisch RNA reeds beschikbaar zijn en nog steeds verder ontwikkeld worden, is de volgende uitdaging doelwitgenen te identificeren die onder controle staan van een specifiek regulatorisch RNA. In dit hoofdstuk ontwikkelden we een erg rudimentaire methode om potentiële doelwitgenen te bepalen van sRNA's die gebruik maken van het basesparingsmechanisme. Naarmate meer details duidelijk worden over het exacte bindingsmechanisme van deze sRNA's met het mRNA van hun doelwitgenen, zal onze predictiemethode geoptimaliseerd kunnen worden door meer karakteristieken in rekening te brengen dan enkel sequentiesimilariteit.

7 Detectie van dubbel-glycine leidersequenties

7.1 Inleiding

In tegenstelling met de andere hoofdstukken in deze thesis is hoofdstuk 7 toegespitst op een biologisch probleem dat zich niet situeert op niveau van DNA-sequenties, maar op niveau van proteïnesequenties. Het doel van de studie was om de evolutionaire verspreiding van een proteïne transport systeem te onderzoeken in alle volledig gekende bacteriële genomen. Dit impliceert dat verschillende methoden voor comparatieve genoomanalyse hergebruikt kunnen worden. Voor proteïnespecifieke problemen moesten we echter terugvallen op publiek beschikbare algoritmen. Dit onderzoek is uitgevoerd in nauwe samenwerking met het Centrum voor Microbiële en Plant Genetica (Prof. J. Michiels, Dr. G. Dirix).

Het transportsysteem voor proteïnen wordt in alle organismen gecontroleerd door eenzelfde onderliggend mechanisme: elk polypeptide dat bestemd is voor extracellulair transport bevat een specifieke aminozuursequentie ook gekend als signaal- of leiderpeptide. Afhankelijk van het leiderpeptide wordt het overeenkomende transportsysteem geactiveerd, waarbij tijdens het transport de leidersequentie vaak afgesplitst wordt. Een interessant signaalpeptide is de dubbel-glycine (GG)-leidersequentie vermits het een sleutelrol speelt bij verschillende peptidesecretiesystemen en bovendien betrokken is in *quorum sensing* en bacteriocine productie. Proteïnen die het GG-motief bevatten worden geëxporteerd m.b.v. een corresponderende ATP bindings cassette (ABC) transporter, namelijk het peptidase C39. Uit voorgaande analyses bleek dat proteïnen die het GG-peptide bevatten en het peptidase C39 steeds in mekaars nabijheid gevonden worden op het bacteriële chromosoom.

In dit hoofdstuk gebruiken we bestaande kennis van goed gekarakteriseerde proteïnen met GG-motief samen met hun corresponderende ABC transporter om de aanwezigheid van het betreffende secretiesysteem na te gaan in alle volledig gekende bacteriële genomen.

7.2 Strategie

Op het moment van onze studie waren alle voorgaande onderzoeken naar GG-leidersequenties uitgevoerd op proteïneniveau. De korte lengte van peptiden die het GG-motief bevatten, heeft ervoor gezorgd dat deze peptiden slechts beperkt geannoteerd zijn in de databanken van bacteriële genomen. In deze studie negeren we daarom de bestaande annotatie en baseren onze analyse op de ruwe DNA-sequentie van de bacteriële genomen. Met behulp van de Wise2 software [24] worden de DNA-sequenties vertaald in de zes mogelijke leesramen waarbij in elk mogelijk leesraam gezocht wordt naar de GG-leidersequentie. Hiervoor werd een Hidden Markov Model (HMM) opgesteld van deze leidersequentie voor zowel Gram-positieve als Gram-negatieve bacteriën. Gezien de aanwezigheid van het GG-motief gerelateerd is aan de aanwezigheid van zijn overeenkomstige ABC transporter peptidase C39, voerden we een gelijkaardige motiefdetectie-analyse uit voor alle bacteriële genomen met een HMM dat het peptidase C39 domein beschrijft.

Het verband tussen het GG-motief en peptidase C39 weerspiegelt zich in de chromosomale locatie van beide genen: op basis van voorgaande analyses worden daarom enkel die GG-leidersequentie in rekening genomen die op minder dan 10kb van een peptidase C39 domein gelegen zijn. Andere criteria voor een GG-motief om als biologisch significant beschouwd te worden, zijn 1) de afwezigheid van inserties of deleties in het GG motief, 2) de afwezigheid van een stopcodon tussen de translatiestart en het einde van het GG motief en 3) de totale lengte van het proteïne moet minder zijn dan 150 aminozuren, en de regio voor het GG motief moet kleiner zijn 50 aminozuren.

7.3 Detectie van Peptidase C39

Het motiefmodel voor het peptidase C39 domein (aanwezig in de Pfam databank [19]) werd gebruikt om met behulp van de Wise2 software alle volledige gekende bacteriële genomen te doorzoeken naar het overeenkomstige proteïnedomein. Dit resulteert in 78 potentiële peptidase C39 domeinen, waarvan voor 3 hits geen correcte annotatie gevonden werd. Procentueel bleek het peptidase C39 domein ook meer teruggevonden te worden in Gram-positieve bacteriën (44%) dan in Gram-negatieve bacteriën (33%). De peptidase C39 domeinen werden gevalideerd door de aanwezigheid van twee geconserveerde proteïnemotieven, namelijk het cysteïne en histidine motief. Deze motieven zijn verantwoordelijk voor de binding en afsplitsing van het GG-motief [105,171,282]. Enkel voor 13 van de 78 potentiële hits kon de aanwezigheid van beide motieven niet bevestigd worden. Elk van deze 13 hits werd geïdentificeerd in Gram-negatieve bacteriën. Deze 13 ABC transporters zijn betrokken in de secretie van toxines uit de hemolysine-familie. Hemolysines bevatten geen leidersequentie, en de domeinen voor de herkenning en afsplitsing van de GG-leidersequentie zijn daardoor overbodig.

7.4 Detectie van dubbel-glycine motief peptides

In een eerste stap worden motiefmodellen opgesteld voor het GG-motief in Gram-positieve en Gram-negatieve bacteriën. Voor Gram-positieve bacteriën konden we ons model baseren op een training set van 31 gekende peptiden die het GG-motief bevatten. Gezien voor de Gram-negatieve species slechts een beperkt aantal GG-leidersequenties geïdentificeerd waren [171], bepaalden we extra GG-motief instanties via een iteratieve procedure van MEME en MAST [15]. Dit resulteerde voor de Gram-negatieve bacteriën in een trainingset van 38 GG-motieven. Met behulp van de HMMER2.2 software [61] werden HMMs opgesteld van beide motieven die gebruikt werden in de Wise2 software voor een genoomwijde zoektocht in de volledig gekende bacteriële genomen.

Op basis van de criteria vermeld hierboven (ligging van peptidase C39, aanwezigheid stopcodons, …), werden de biologische relevante peptiden met een GG-leidersequenties bepaald. Voor Gram-negatieve bacteriën leidde dit tot 58 potentiële hits, waarbij de lengte van het overeenkomstige peptide varieert tussen 23 en 142 aminozuren. Zoals hierboven vermeld werden 13 van de peptidase C39 domeinen die geen cysteïne en histidine motief bevatten teruggevonden in Gram-negatieve bacteriën. Als gevolg hiervan zouden geen GG-leidersequenties mogen teruggevonden worden in de buurt van deze peptidasen. Voor 12 van de 13 peptidase C39 domeinen is dit inderdaad het geval. De GG-leidersequentie die voorkomt in de buurt van een hemolysine-secreterende transporter, is vermoedelijk een vals positief resultaat.

De zoektocht in Gram-positieve bacteriën leidde tot een lijst van 48 kandidaat GG-leidersequenties. 92% van deze potentiële leidersequenties werd gevonden in melkzuurbacteriën. De lengte van de overeenkomstige peptiden varieerde tussen 29 en 126 aminozuren. Naast 17 hypothetische proteïnen, komen in de lijst van potentiële hits ook 15 bacteriocines en 10 bacteriocine-homologen voor. Wat betreft de peptidase C39 domeinen, werd voor 21 van de 29 proteïnen een GG-leidersequentie in de onmiddellijke omgeving teruggevonden.

7.5 Conclusie

Onze methodologie leidde tot nieuwe inzichten in de verspreiding van het GG-peptide verwerkings- en secretiesysteem in Gram-positieve en Gram-negatieve bacteriën. Hiervoor baseerden we ons niet op voorgaande annotatiegegevens van de verschillende genomen maar werd vertrokken van de ruwe DNA-sequentie.(vertaald in de zes potentiële leesramen). Omwille van de stringente criteria die we toepasten voor de detectie van de GG-peptiden, konden we niet alle gekende GG-leidersequenties terugvinden. De stringente criteria zorgden er ook voor dat voor de gedetecteerde peptidase C39 domeinen niet steeds een GG-leidersequentie gevonden werd in de directe omgeving. Wanneer echter meer GG-peptides experimenteel geverifieerd worden, kan ons algoritme verder verfijnd worden.

8 Conclusies en perspectieven

8.1 Conclusies

In deze thesis hebben we aangetoond dat comparatieve genoomanalyse een krachtige methode is voor de detectie van nieuwe transcriptiefactorbindingsplaatsen. Deze benaderingen zijn gebaseerd op de vergelijking van orthologe promotersequenties waarbij men verwacht dat ze bindingsplaatsen bevatten voor hetzelfde regulatorproteïne. Wij ontwikkelden in deze thesis methoden die gebruik maken van bestaande of nieuw ontwikkelde algoritmen om biologische problemen op te lossen waarvoor we gebruik konden maken van comparatieve genoomanalyse.

- Ontwikkeling van een “phylogenetic footprinting” methode die resulteert in een meer betrouwbare identificatie van regulons (e.g. PmrAB regulon).
- Ontwikkeling van een methode die toelaat om regulons te vergelijken tussen verschillende species (e.g. PhoPQ regulon).
- Een *de novo* motiefdetectiemethode gebaseerd op comparatieve genoomanalyse om niche-specifieke regulatorische motieven te identificeren in evolutionair nauw gerelateerde species.
- Toepassing van de ontwikkelde methodes en algoritmes op:
    - Detectie van potentiële doelwitgenen van regulatorisch RNA.
    - Detectie van kleine signaalpeptiden (GG-peptiden) in alle volledig gekende bacteriële genomen.

8.2 Perspectieven

Vandaag de dag zijn genetische netwerkinferentie en systeembiologie frequent weerkerende begrippen in bio-informatica. Identificatie van regulatorische motieven kan niet losgekoppeld worden van deze toepassing. Genetische netwerkreconstructie vereist vaak een combinatie van verschillende databronnen, waarbij regulatorische motieven een belangrijke plaats opeisen. Het opstellen van een motiefcompendium enkel en alleen op basis van sequentiedata (met behulp van comparatieve genoomanalyse) zou een meerwaarde betekenen voor genetische netwerkinferentie. Dergelijke motiefdata zijn immers onafhankelijk van experimentele data en bijgevolg conditieonafhankelijk. Algoritmische aanpassingen van de methodologie ontwikkeld in hoofdstuk 5 moet het mogelijk maken om dergelijk motiefcompendia op te stellen (Valerie Storms en Abeer Fadda). Deze motiefdata kunnen dan geïntegreerd worden in data-integratiealgoritmen zoals ReMoDiscovery [142].

De novo motiefdetectie met behulp van fylogenetische informatie kan echter nog verder geoptimaliseerd worden. Op dit moment worden de motieven geïdentificeerd door sequentieel co-expressie en fylogenetische informatie te gebruiken. Meer geavanceerde algoritmen kunnen ontwikkeld worden die simultaan gebruik maken van beide informatiebronnen. Hierbij kan bovendien de evolutionaire afstand tussen de betrokken organismen in rekening gebracht worden (Marleen Claeys en Sun Hong).

Een vaak onderschat probleem in de reconstructie van genetische netwerken is de invloed van regulatorisch RNA. De algoritmes voor genetische netwerkreconstructie zullen zodanig geïmplementeerd moeten worden dat ze rekening kunnen houden met de invloed van regulatorisch RNA. Hiervoor moet echter de identificatie van sRNA doelwitgenen nog verder op punt gesteld worden. Voor de optimalisatie van de huidige detectiemethode wordt samengewerkt met Dr. J. Vogel (Max Plank Institute).

De algoritmen en methoden ontwikkeld in deze thesis, gecombineerd met de toenemende beschikbaarheid van experimentele data, zullen bijdragen tot de ontrafeling van bacteriële regulatorische netwerken. Op lange termijn zal dit ook toelaten om inzicht te verwerven in de evolutie van volledige regulatorische netwerken.

Publications

1. Marchal K, Thijs G, De Keersmaecker S, Monsieurs P, De Moor B, Vanderleyden J: Genome-specific higher-order background models to improve motif detection. *Trends in Microbiology* 2003, 11:61-66.

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

3. Monsieurs P, De Keersmaecker S, Navarre WW, Bader MW, De Smet F, McClelland M, Fang FC, De Moor B, Vanderleyden J, Marchal K: Comparison of the PhoPQ regulon in *Escherichia coli* and *Salmonella typhimurium*. *Journal of Molecular Evolution* 2004, 60:462-474.

4. Dirix G, Monsieurs P, Dombrecht B, Daniels R, Marchal K, Vanderleyden J, Michiels J: Peptide signal molecules and bacteriocins in Gram-negative bacteria: a genome-wide in silico screening for peptides containing a double-glycine leader sequence and their cognate transporters. *Peptides* 2004, 25:1425-1440.

5. Dirix G, Monsieurs P, Marchal K, Vanderleyden J, Michiels J: Screening genomes of Gram-positive bacteria for double-glycine-motif-containing peptides. *Microbiology* 2004, 150:1121-1126.

6. De Bie T, Monsieurs P, Engelen K, De Moor B, Cristianini N, Marchal K: Discovering transcriptional modules from motif, chip-chip and microarray data. *Pacific Symposium on Biocomputing* 2005,483-494.

7. Van Hellemont R, Monsieurs P, Thijs G, De Moor B, Van de Peer Y, Marchal K: A novel approach to identify regulatory motifs in distantly related genomes. *Genome Biol* 2005, 6:R113.

8. Monsieurs P, Thijs G, Fadda A, De Keersmaecker S, Vanderleyden J, De Moor B, Marchal K: More robust detection of motifs in coexpressed genes by using phylogenetic information. *BMC Bioinformatics* 2006, 7:160.

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

Contents

VOORWOORD i

ABSTRACT v

KORTE INHOUD vii

ABBREVIATIONS ix

NEDERLANDSE SAMENVATTING xi

PUBLICATIONS xxxv

CONTENTS xxxvii

1 TRANSCRIPTIONAL REGULATION IN PROKARYOTES 1

1.1 Context of the thesis 1

1.2 Basal transcription process 4
- 1.2.1 Overview 4
- 1.2.2 Structure of a prokaryotic gene 5
- 1.2.3 Proteins controlling gene expression 6

1.3 Regulatory motifs 9
- 1.3.1 Representation of a motif 9
- 1.3.2 Computational detection of regulatory motif 12

1.4 Transcriptional Regulatory Networks 15
- 1.4.1 Structure and function of regulatory networks 15
- 1.4.2 Reconstruction of regulatory networks 15
- 1.4.3 Evolution of regulatory networks 16

1.5 Overview of the thesis 17

2 MOTIF DETECTION USING COMPARATIVE GENOMICS 21

2.1 Introduction 21

2.2 Comparative Genomics for motif detection: applications 22

2.2.1 Extending regulon information 22
2.2.2 Phylogenetic footprinting 24
2.2.3 Combined approaches 25

2.3 Motif detection by grouping of evolutionary conserved motifs 26

2.3.1 Motif detection using comparative genomics and coexpression 27
2.3.2 Genome-wide motif detection 28

2.4 Integration of phylogenetic distances 31

3 IDENTIFICATION OF THE PMRAB REGULON 33

3.1 Introduction 33

3.2 Results 35

3.2.1 Genome-wide screening using a PmrA motif model 35
3.2.2 Identification of close homologs 36
3.2.3 Phylogenetic footprinting using Gibbs sampling 36
3.2.4 Detailed analysis of the putative PmrAB targets 43
3.2.5 Experimental analysis 49
3.2.6 Other promising PmrAB targets 53

3.3 Discussion 54

3.4 Conclusion 59

3.5 Materials and methods 59

3.5.1 Selection of intergenic sequences 59
3.5.2 Construction of motif models 59
3.5.3 Genome-wide screening 60
3.5.4 Identification of datasets 61

3.5.5 Phylogenetic footprinting by Gibbs sampling 61
3.5.6 Functional annotation 63
3.5.7 Nomenclature 63

4 COMPARATIVE ANALYSIS OF THE PHOPQ REGULON 65

4.1 Introduction 65

4.2 Results 67
4.2.1 Identification of the PhoP dependent regulon based on microarray analysis 67
4.2.2 Motif screening 67
4.2.3 Combination of motif screening results and microarray data 70
4.2.4 Functional classes 74
4.2.5 Detailed analysis of novel direct PhoPQ targets 77

4.3 Discussion 79

4.4 Conclusion 81

4.5 Material and methods 81
4.5.1 Identification of PhoPQ dependent genes based on expression data 81
4.5.2 Sequence-based analysis 83
4.5.3 Calculating the overlap in regulon composition 85
4.5.4 Enrichment of functional classes 85

5 MORE ROBUST DETECTION OF MOTIFS IN COEXPRESSED GENES BY USING PHYLOGENETIC INFORMATION 87

5.1 Introduction 87

5.2 Results 89
5.2.1 General Strategy 89
5.2.2 Evaluation of the analysis flow on test data 93
5.2.3 Importance of the selected species 95
5.2.4 Comparison with other methodologies 95
5.2.5 Evaluation of the analysis flow on expression data 97

5.3 Discussion 99

5.4 Conclusion 101

5.5 Methods 102

5.5.1 Input data sets 102
5.5.2 Analysis flow 103
5.5.3 Benchmarking with other methodologies 106

6 DETECTION OF REGULATORY RNA TARGETS 109

6.1 Introduction 109

6.2 Results 112

6.2.1 Characterization of the regulatory RNA 112
6.2.2 Screening for sRNA targets 113

6.3 Discussion 115

6.4 Methods 117

6.4.1 Genomic context of the regulatory RNA 117
6.4.2 Detection of conserved regions in regulatory sRNA 118
6.4.3 Screening for putative targets of sRNA 118

7 DETECTION OF DOUBLE GLYCINE LEADER SEQUENCES 119

7.1 Introduction 119

7.2 Results 122

7.2.1 Overview 122
7.2.2 Peptidase C39 domain search 125
7.2.3 Search for genes encoding double-glycine motif containing peptides 129

7.3 Discussion 134

7.4 Methods 136

7.4.1 MEME/MAST system 136
7.4.2 Genome-wide screening using HMMs 137
7.4.3 BlastX 137

8 CONCLUSIONS AND PERSPECTIVES 139

8.1 Conclusions 139

8.2 Perspectives 141

REFERENCES 145

CURRICULUM VITAE 167

Chapter 1

Transcriptional regulation in prokaryotes

1.1 Context of the thesis

Bioinformatics, although a relatively recent term, already exists for several decades. Indeed, the use of mathematical models to explain biological phenomena and analyze data is certainly not new. So far, it was only common practice in certain fields of biology (phylogeny, molecular modeling, population genetics). However, the development of new high-throughput techniques during the 1990s opened up new opportunities in the field of computational biology. This technological progress dramatically changed the view on molecular biology. Whereas a few years ago each gene or protein was studied as a single entity, new high-throughput technologies allowed one to analyze a large number of genes or proteins simultaneously and put them in the context of an entire biological system [87]. Automated DNA sequencers enabled the sequencing of genomes, microarray analysis permitted global expression profiling, advances in mass spectrometry led to large-scale proteomic and metabolomic analysis, ChIP chip data allow detecting protein-DNA binding interactions, etc. As a result, a gene is no longer studied as an isolated entity, but as a part of a complex network that determines the phenotype of the cell. From this perspective the cell is considered as a system that interacts with its environment. Modeling the dynamic action of these networks to predict cellular behavior is the ultimate goal of systems biology [11,131].

Despite the fact that a high number of the algorithms for systems biology are tested and validated on known eukaryotic model organisms such as yeast [78,112] and *Caenorhabditis elegans* [78], there are still major challenges waiting in the field of microbiology. Systems microbiology has a great potential for both fundamental and industrial applications, ranging from adaptation and evolution to improved management of bacterial infections and designing better performing industrial strains.

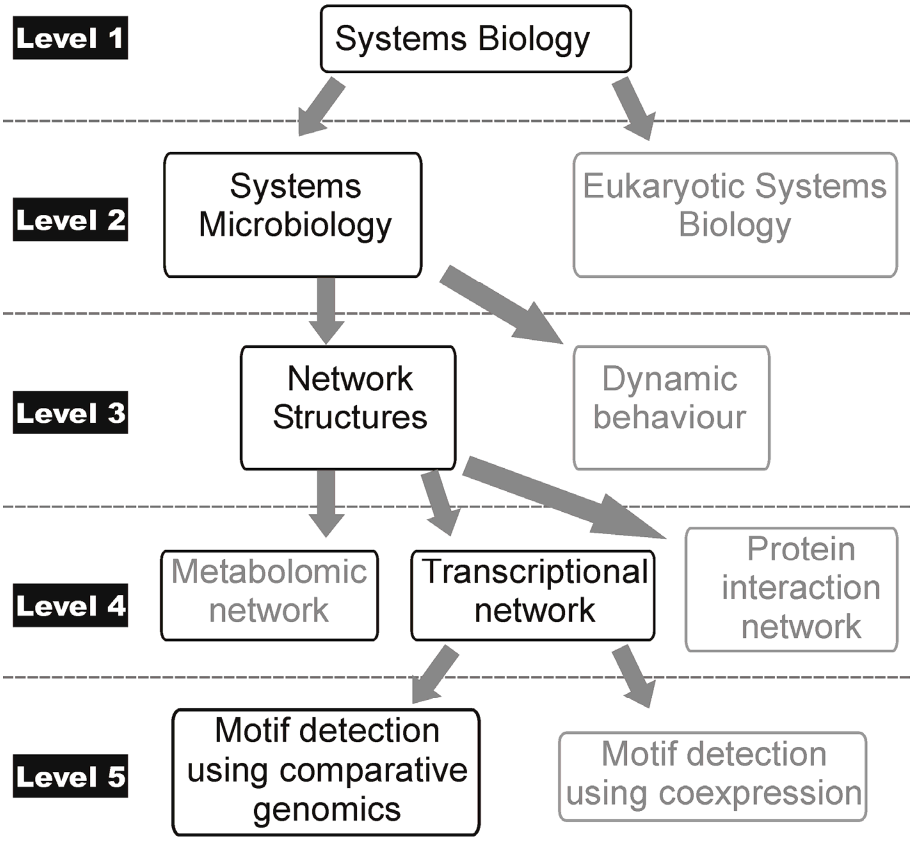


Figure 1.1: **Context of the thesis.** From the most general level (level 1: Systems Biology) to the most detailed level (level 5: motif detection using comparative genomics).

Systems (micro)biology focuses on the understanding of the component and dynamic behaviors of biological systems. In order to gain insight into the causal interactions between the molecular entities, a first challenge in systems biology is the reconstruction of the basic network structures. This can be achieved by combining different high-throughput data, such as transcriptomes and proteomes, to identify and reconstruct components and connectivity of biological networks (e.g. [18,124,142]). Based on these basic network structures, more detailed mechanistic models can be compiled taking into account dynamic behavior. Describing this dynamic behavior will have to give an explanation how organisms are responding to environmental stress and continuously changing growth conditions. Despite the fact that different studies have already been performed on biological system dynamics (e.g.[4,121]), major challenges are still waiting at the level of network reconstruction.

These network reconstructions can be applied at different levels, depending on the data sources used (transcriptional networks, protein interaction network, metabolomic networks, etc.) where the ultimate goal is to combine these different basic networks into one comprehensive network model. As the rate of transcriptional initiation is the predominant site for control of gene expression [109,252], studying the transcriptional regulatory network is essential in the understanding of complex behavior of bacteria. Despite the effort that has been put in the identification of transcriptional networks, several fundamental questions regarding the structure, evolution and different components of transcriptional regulatory networks remain unanswered. Regulation at transcriptional level is typically mediated by a DNA binding protein (transcription factor) that binds to target sites (i.e. a regulatory motif) in the genome and – either single or in combination with other factors – regulates the expression of one or more target genes. The total sum of all these transcriptional interactions can be conceptualized as a network, and is termed the transcriptional regulatory network [12,121,140,155,237]. As regulatory motifs form the linkage between the different components of transcriptional regulatory networks, reliable motif detection methods are indispensable for reconstruction of regulatory networks.

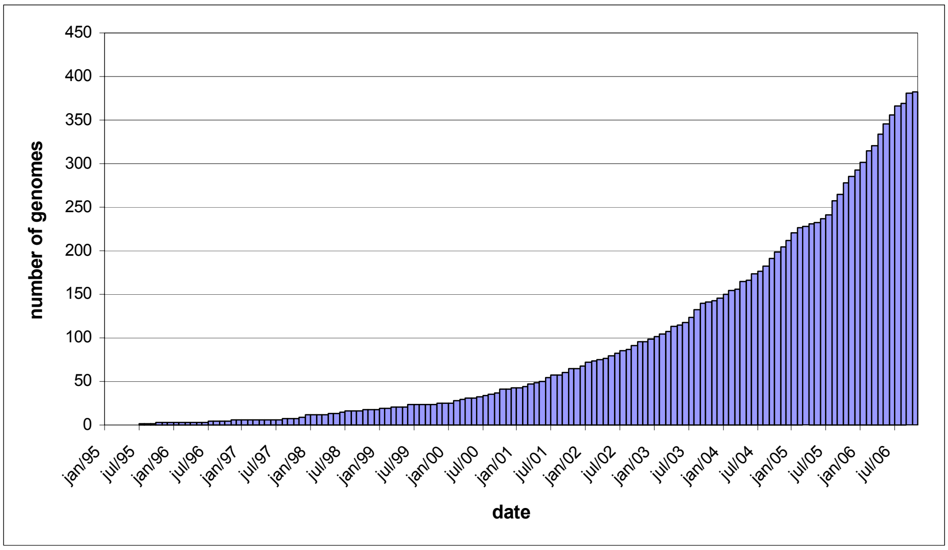


Figure 1.2: **Overview of the number of completely sequenced bacterial genomes over time.**

Initially motif detections methods were based on the idea that genes that are coregulated by the same transcription factor share a similar binding site. However, as the number of fully sequenced genomes is constantly increasing, this unprecedented resource of genomic data opens up new opportunities for computational biology, also for regulatory motif detection. At this moment, the NCBI Entrez database contains 383 fully sequenced microbial genomes, and the sequencing of an additional 644 microbial genomes is in progress (see figure 1.2). This overwhelming amount of sequence data can be used in a comparative genomics approach to evaluate the genomic characteristics between different organisms. Initially, this approach has mainly been applied to protein-coding sequences e.g. for revealing the function of a protein in a poorly characterized genome by comparison with proteins from well-studied microbial species. Similarly with protein-encoding sequences, evolutionary forces also tend to preferentially retain functional DNA sequences in promoter regions (i.e. regulatory motifs). By comparing sequences from different evolutionary related species – which cannot be a problem giving the number of fully sequenced genome – the conserved regulatory sequences can be detected.

As mentioned in above, bacteria have evolved complex (and often overlapping) transcriptional regulatory networks that respond to various changes in growth conditions, and this in response to the ever-present need to adapt to environmental stress. The regulation of prokaryotic gene expression can be controlled at different levels: at transcriptional, RNA processing or translation level. As the rate of transcriptional initiation is the important check point for the control of gene expression [109,252], we mainly focused on this process in the thesis. This first chapter contains an overview of transcriptional regulation in prokaryotes with main focus on the role and detection of regulatory motifs.

1.2 Basal transcription process

1.2.1 Overview

Transcription is the process during which the genetic information is transcribed from DNA into RNA. In its most basic form, transcription only needs the RNA polymerase to copy enzymatically the coding DNA sequence to RNA, which implies the transfer of genetic information from DNA to RNA. Transcription can be divided into three stages. The process starts when the RNA polymerase binds to the promoter region upstream of the gene that needs to be copied to RNA (template recognition). The strands of the DNA are separated to make the template strand available for base pairing with ribonucleotides. In the second stage, the initiation process synthesizes the first nucleotide bonds in RNA. The RNA polymerase remains at the promoter while it synthesizes the first approximately nine nucleotides. During elongation (i.e. the third stage) the RNA polymerase enzyme moves along the DNA and extends the growing RNA chain. The last stage is the termination step which involves the recognition of the point at which no further bases need to be added.

First we will describe the different compounds in a prokaryotic gene and the function of each of these compounds in this transcriptional process. Secondly, we describe the proteins responsible for the control of gene expression.

1.2.2 Structure of a prokaryotic gene

1.2.2.1 Bacterial promoter

As described above, initiation of transcription starts with the binding of the RNA polymerase (RNAP) to the promoter region of the transcriptional unit that needs to be transcribed to RNA. A key question in examining the interaction between the RNAP and a promoter region is how the protein recognizes a specific promoter sequence. This interaction is based on three conserved features of a bacterial promoter. 1) The -10 sequence or Pribnow-box is a hexamer located just upstream of the transcription start, centred around the -10 position before the transcription start site. 2) The -35 sequence is a hexamer centred around the -35 position. The high-resolution structure of the RNAP holoenzyme when bound to DNA, confirms that the sigma factor establishes specific contacts with both the -10 and -35 region [165,186]. 3) The distance separating the -10 and -35 site is between 16 and 18 bp in 90% of all promoters. Although the actual sequence in this intervening region is unimportant, the distance is critical in holding the two sites at appropriate distance for the geometry of the RNAP. Remark that some of these characteristics are only valid for the most prominent RNAP i.e. the σ^{70} containing RNAP (see 1.2.3.1).

The promoter region does not only contain signals necessary to recruit RNA polymerase. It also contains short, conserved DNA sequences that act as a binding site for regulatory proteins. These regulatory proteins are essential in the control of gene expression. The short DNA sequences are called regulatory motifs or transcription factor binding sites (see section 1.3).

1.2.2.2 Operon structure

Prokaryotic genes are often grouped together into operons which contain genes that are co-regulated at the transcriptional level as all genes in one operon share the same promoter region (i.e. located upstream of the first gene of the operon) and are thus transcribed together. So identifying operon structures is essential for a further understanding of gene regulation and function (also see chapter 4). In addition, genes grouped in operons are often coding for proteins whose functions are related. Indeed, analysis by Overbeek *et al.* [192] of several complete bacterial genomes has led to the conclusions about the tendencies of genes with related function to remain together trough evolution. Different operon prediction methods have been developed by several groups. The first approach is based on simultaneous modelling of transcriptional and translational signals present in operons (ribosome binding sites, promoter, terminator regions, …) [310]. The second approach takes advantage of the differences in distances of adjacent genes in the same operon [181,229]. A third approach is based on a comparative genomics approach where gene order and orientation is conserved between two or more genomes [63,68]. A last type of algorithms adopts a machine-learning approach and utilizes a variety of data types including gene expression, functional annotation, intergenic distances, transcription signal features, … [28,119,191,299,316].

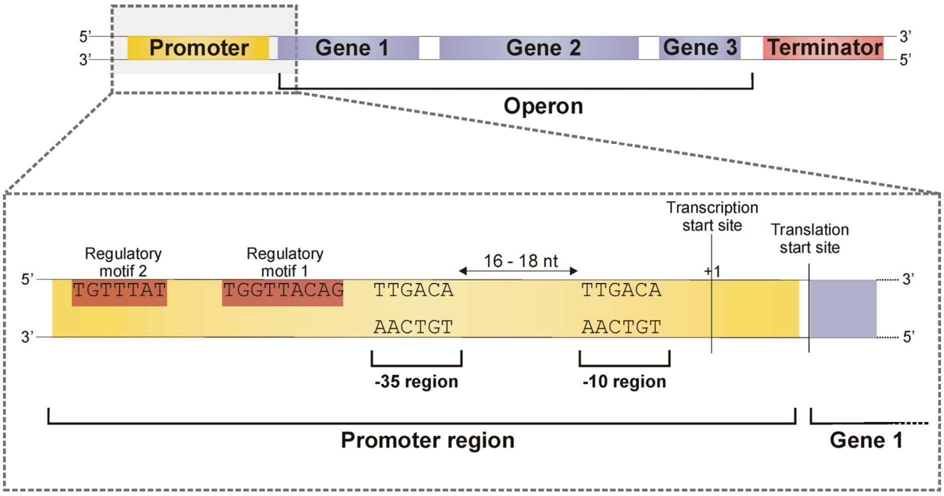


Figure 1.3: **Overview of a prokaryotic operon.** In this example, the operon consists of three genes all under control of the promoter region upstream of gene 1. In a detailed view of the promoter region, the -10 region and -35 region are indicated (consensus sequences for the σ^{70}-factor). Positions are relative to the transcription start site. Upstream of these conserved regions, two hypothetical examples of regulatory motifs are given.

1.2.3 Proteins controlling gene expression

1.2.3.1 RNA polymerase

RNA polymerase is the key enzyme of transcription and gene expression in all living organisms. In bacteria, the catalytically competent core consist of five subunits of which the structure and function are evolutionary conserved [60,254]. This core enzyme – containing four out of the five subunits – has the ability to synthesize RNA on a DNA template, but cannot initiate transcription at proper sites. Together with the σ-factor (i.e. the fifth subunit), the RNAP core forms the holoenzyme, where this σ-factor is capable of specific promoter recognition and efficient initiation of transcription [34]. During the initiation stage of the transcription process, the σ-factor of the holoenzyme specifically binds to two conserved hexamers in the promoter, located at nucleotide positions -10 and -35, relative to the transcription start site. After transition to the elongation stage, the σ-factor dissociates from the core enzyme and is free to bind a new RNAP core enzyme starting a new cycle of transcription. This key feature of RNA polymerase σ-factors allows cells to adjust to transcription patterns rapidly to optimize cellular metabolism in response to changing external conditions and cellular signals by using different σ-factors with different promoter specificities to regulate different sets of genes [180].

Among many different σ-factors expressed in bacteria [148,302], σ^{70} is the major factor responsible for the bulk of transcription activity in the cell. Bacterial cells also use alternative σ-factors, specific to different subsets of promoters, to adapt to environmental changes [107,117,302]. The number of σ-factors varies from one in *Mycoplasma genitalia* [74], to seven in *Escherichia coli* [117] to more than 60 in *Streptomyces coelicolour* [22]. An overview of the different σ-factors in *E. coli* is given in table 1.1.

Sigma factor	Gene	Consensus sequence			Function
		-35 region	Separation	-10 region	
σ^{70}	rpoD	`TTGACA`	16-18 bp	`TATAAT`	housekeeping genes
σ^{32}	rpoH	`CCCTTGAA`	13-15 bp	`CCCGATAT`	heat-shock genes
σ^{54}	rpoN	`CTGGCAC`	5bp	`TTGCA`	nitrogen-regulated
σ^{E}	unknown	`GAACTT`		`TCTGA`	heat-shock genes
σ^{F}	fliA	`CTAAA`	15 bp	`GCCGATAA`	flagellar, chemotaxis
σ^{K}	rpoK	unknown		unknown	katE

***Table 1.1*: Summary of *E. coli* sigma factors.**

Despite the fact that a large number of biochemical and genetic data on RNA polymerase accumulated in the past 20 years, our understanding of the basis transcriptional mechanisms is far from complete e.g. the high-affinity binding of the σ-factor to the core enzyme, the precise mechanism of promoter recognition and DNA melting and the mechanism of promoter escape when going from initiation to elongation stage. Extra high-resolution structures of RNA polymerase complexes (with and without extra transcription factors) and biochemical and genetic data are required to get a full understanding of the complexity of every step at the transcription cycle.

1.2.3.2 Transcription factors

Changes in the synthesis of σ-factors and competition of these different σ-factors to bind to the core RNA polymerase, permits the cell to induce and repress different programs of gene expression. However, this global regulation mechanism only permits to respond to general conditions, while often a more specific reaction is required (e.g. in pathogenic bacteria). Therefore, in addition to the RNA polymerase, transcription factors that respond to specific conditions in the environment, can bind the regulatory motifs in the promoter region and facilitate or inhibit the binding of the RNA polymerase to the promoter region and thus influence the transcription rate of the corresponding operon. Estimations on the total number of transcription factors in *E. coli* vary between 270 and 400 [153,204,265]. Mainly in *E. coli,* different attempts have been made to classify these transcription factors into different groups. Every classification system is based on one of the typical features of a regulatory protein:

- **Domain architecture**: More than 90% of the transcription factors are multiple-domain proteins, and more than three-quarters are two-domain proteins [8,153,182]. Transcription factors have been classified based on the domain architecture (i.e. the combination of domains) [153]. One of the domains needs to be a DNA-binding domain, while the other domains most of the time have a different function, such as small molecule-binding, protein interaction or enzymatic domain.

- **DNA-binding domain**: A second way of classifying the transcription factors is based on their DNA-binding domain. The helix-turn-helix (HTH) domain is detected in most of the *E. coli* regulatory proteins (more than 75%) [103]. Additional DNA-binding domains such as zinc fingers [219], antiparallel β-sheets [245], RNA-binding like motifs [86] and helix-loop-helix [224] have been described in regulatory proteins of *E. coli*, although they contribute a small fraction compared with the regulatory proteins with an HTH motif.

- **Functional categories**: The most straightforward way of classifying regulatory proteins, is based on their function. Regulatory proteins have been grouped into evolutionary families based on their sequence similarity [204,205] as this similarity most of the time reflects a common function.

- **Activators and repressors**: Transcription factors can be activators, repressors or both. Whether a transcription factor acts as an activator or a repressor, can be computationally be predicted based on the relative position of the regulatory motif in the promoter region [43,154] and the relative position of the

HTH motif within the protein sequence [206]. Following these results, the hypothesis is made that prokaryotic activators function by stabilizing the polymerase from the promoter region and repressors act by steric hindrance blocking RNA polymerase binding or processing [154].

1.3 Regulatory motifs

Gene expression is regulated by transcription factor binding to specific transcription factor binding sites (also called regulatory motifs). Therefore, the basic functional element of a regulatory network and thus transcriptional regulation is the promoter region of a gene or operon which contains regulatory motifs for the relevant transcription factors that regulate the expression of that particular gene. The location and affinity of the transcription factors for these binding sites determine the expression levels of a gene in response to changes in active transcription factor concentration inside the cell. Therefore, identification of regulatory motifs is a prerequisite for understanding gene regulation.

1.3.1 Representation of a motif

Since variability in regulatory binding sites exists for one particular regulatory protein, the representation of the regulatory motif is not straightforward. Three different ways of representing a motif can be considered: a string-based approach, a matrix representation and a motif logo (figure 1.4).

1.3.1.1 String-based representation

A string-based approach that has been widely used to represent the binding specificity of a regulatory motif, is the consensus sequence. However, the definition of this consensus sequence is rather arbitrary [48]. In general it refers to a sequence that matches all of the binding sites closely, but not necessarily exactly. At each position in this consensus sequence, the most prominent nucleotide or combination of nucleotides is displayed. The alphabet used in the representation of this consensus sequences is based on the IUPAC (International Union of Pure and Applied Chemistry) code symbols. This code permits to designate more than one nucleotide at a given position in the consensus sequence. In this work one nucleotide is displayed in the consensus sequence if it is present in more than 60% of the binding sites. If there are two dominant nucleotides, the degenerate alphabet for two nucleotides is used if there joint presence in the binding sites is higher than

60% (e.g. if at a certain position the nucleotide A appears in 30% of the binding sites, and nucleotide T in 40 %, the degenerate symbol ‘w’ is assigned to this position).

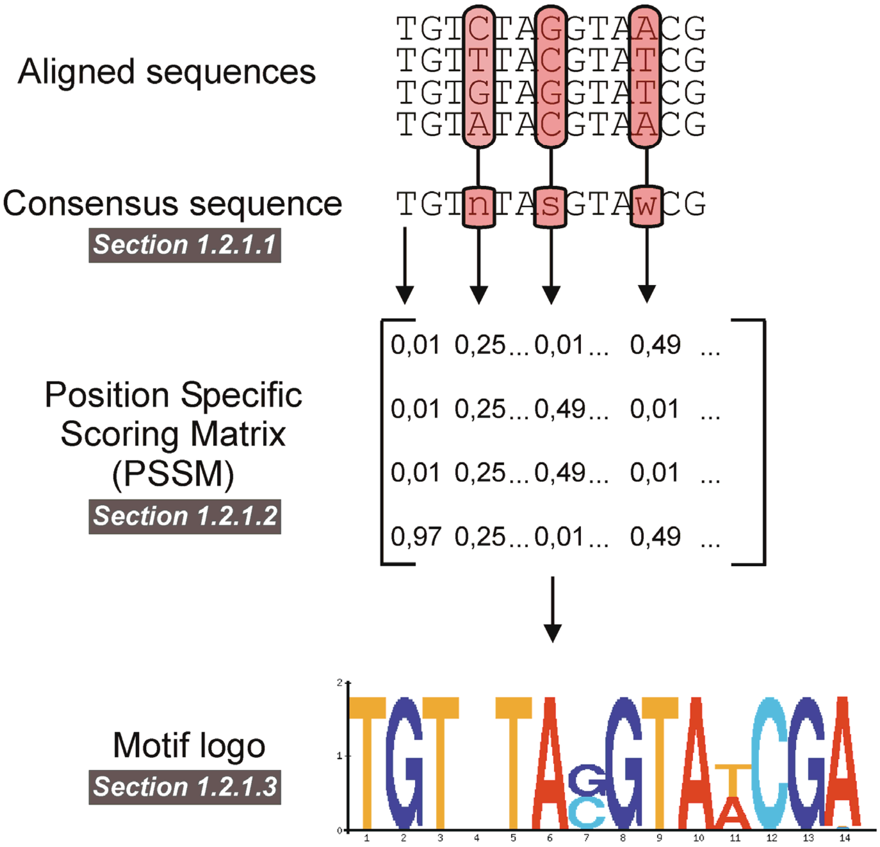


Figure 1.4: **Different ways of representing a motif**. Aligned subsequences are displayed at the top. Based on this alignment a consensus sequences can be built. The highlighted positions indicate degenerate nucleotides where “n” indicates a random nucleotide, “s” indicates a C or a G and “w” indicates an A or a T. Another representation is the Position Specific Scoring Matrix (PSSM) that contains the frequency of a particular nucleotide at a specific position. Only the frequencies for the first position and the highlighted positions are displayed. Based on this matrix, a motif logo can be constructed, giving a visual representation of the binding site.

1.3.1.2 Matrix representation

Despite the fact that the consensus sequence gives an easy to interpretable impression of the binding site, it merely serves as an average binding site and is rather limited in the true specificities of the binding site. A more advanced way of representation of a motif model is a matrix model. In its simplest application, this matrix is an alignment or count matrix which lists the number of occurrences of each letter at each position of an alignment. From the count matrix, a position specific frequency matrix can be constructed by calculating the frequencies of each letter at each position and by introducing pseudocounts. The pseudocounts compensate for the zero occurrences in the count matrix and make it possible to interpret the values in this frequency matrix as probabilities. These pseudocounts are mostly chosen proportional to the prior probability of finding the nucleotide in the genome (i.e. the single nucleotide frequency in the genome). This type of matrix is used in the MotifLocator, MotifSampler, BlockSampler and BlockAligner algorithm (see chapter 3, 4 and 5). A third type of matrix is a position weight matrix (PWM) from which the elements can be calculated using the following formula:

$$w_{ij} = \ln \frac{(n_{ij} + p_i)/(N+1)}{b_i} \approx \ln \frac{f_{ij}}{b_i} \qquad (1.1)$$

where N is the total number of sequences, b_i is the single nucleotide frequency in the genome for nucleotide i, n_{ij} is the total number of occurrences of nucleotide i at position j, f_{ij} is the frequency of nucleotide i at position j, p_i is the pseudocount for nucleotide i.

1.3.1.3 Motif logo

A third way of representing a regulatory binding site is a motif logo. Information theory is used in this representation to visualize the aligned binding sites. A motif logo is made in such a way that the sum of the heights of each of the possible nucleotides corresponds to the amount of information that is stored at that position. In its most straightforward way, the information content at a particular position can easily be computed based on the frequency matrix.

$$I_i = 2 + \sum_{b=A}^{T} f_{b,i} \log_2 f_{b,i} \text{ bits} \qquad (1.2)$$

where i is the position within the site, b refers to each of the possible nucleotides and f_{ij} is the observed frequency of each nucleotide at that position. I_i varies between 0 for positions where the frequency of all nucleotides is 0.25, and 2 bits for positions completely conserved for one position. This information content is used as total height of the stack in the motif logo representation. In addition, Berg and Von Hippel showed that the logarithms of nucleotide frequencies should be proportional to the binding energy contribution of the nucleotides [292]. This idea fits nicely with the information content analysis and suggests that the information content is related to the average binding energy for the collection of binding sites. However, equation (1.2) and the analysis on bind energy contribution are only appropriate for genomes with 25% for each nucleotide. Therefore, a more general form of equation (1.2) is:

$$I_i = \sum_{b=A}^{T} f_{b,i} \log_2 \frac{f_{b,i}}{p_b} \qquad (1.3)$$

where p_b is the frequency of nucleotide b in the whole genome. This equation is also known as the relative entropy or the Kullback-Leibler distance.

1.3.2 Computational detection of regulatory motif

As experimental identification and verification of regulatory motifs is challenging and time consuming, much effort has been put into the development of computational approaches [272]. Good computational approaches can potentially provide high-quality prediction of regulatory motifs and reduce the time needed for experimental verification. However, the computational approach has turned out to be at least challenging as the experimental one, and a very large number of different methods have been developed [233]. The first implementations only used sequence data (i.e. promoter regions) to detect regulatory motifs. At a later stage, different researchers introduced other (experimental) data sources in addition to upstream DNA sequences for the identification of regulatory motifs.

In this chapter, we only describe the motif detection strategies that start from sets of coregulated genes (Figure 1.5). A second approach is the detection of regulatory motifs using phylogenetic information i.e. searching for motifs in orthologous promoter regions. This latter approach and its different variants are described in chapter 2.

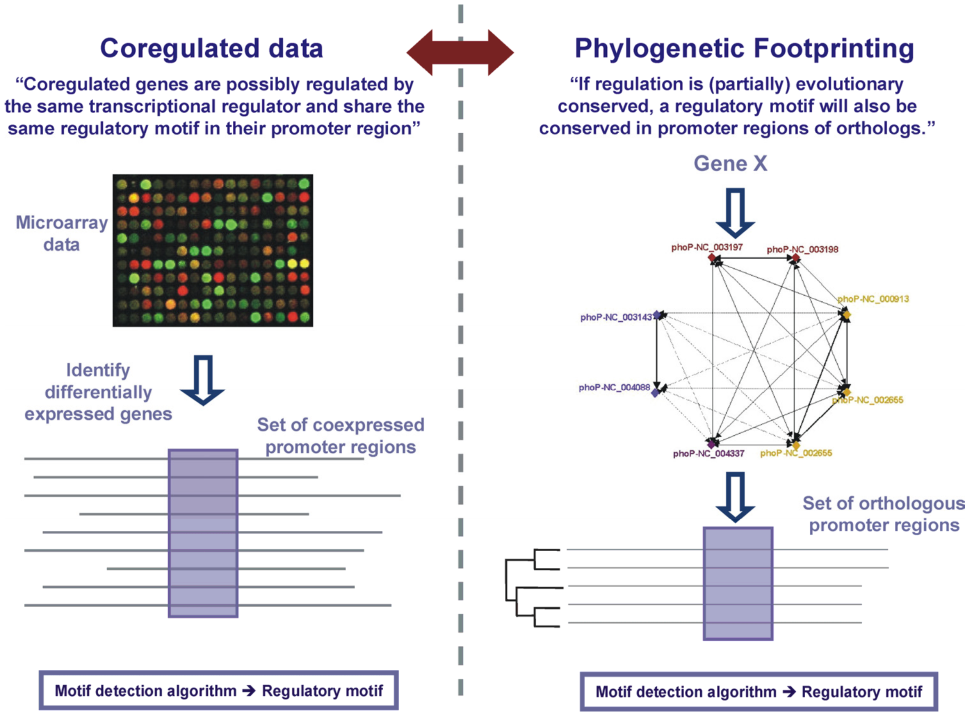

Coregulated data
"Coregulated genes are possibly regulated by the same transcriptional regulator and share the same regulatory motif in their promoter region"
Microarray data
Identify differentially expressed genes
Set of coexpressed promoter regions
Motif detection algorithm ➔ Regulatory motif
Phylogenetic Footprinting
"If regulation is (partially) evolutionary conserved, a regulatory motif will also be conserved in promoter regions of orthologs."
Gene X
phoP-NC_003197
phoP-NC_003198
phoP-NC_000913
phoP-NC_003143
phoP-NC_004088
phoP-NC_002655
phoP-NC_004337
phoP-NC_002655
Set of orthologous promoter regions
Motif detection algorithm ➔ Regulatory motif

Figure 1.5: **Two types of motif detection tools**. The first strategy starts from a set of coregulated genes within one species, and searches for a motif that is statistically overrepresented in this set of genes. The second approach (i.e. phylogenetic footprinting) looks for motifs that are evolutionary conserved in promoter regions of different orthologs.

1.3.2.1 Sequence-based algorithms

The earliest motif detection algorithms are based on the hypothesis that genes showing synchronized changes in expression level or having a similar functional annotation, are co-regulated and share the same transcriptional regulation mechanism i.e. regulatory motifs. These algorithms can be divided into two groups based on the match model they use to represent the regulatory motif: deterministic models representing the motif as a string or regular expression versus probabilistic models representing the motif as position specific scoring matrix (PSSM).

Deterministic models are mostly used in enumerative methods, also called word-counting methods. The enumerative methods typically involve exhaustive enumeration of words up to some maximum size in a data set, and are thus best suited to consensus sequence motif models. Once the words are catalogued, they can be scored using an appropriate measure of statistical significance [202,284]. Recently, dictionary-based motif discovery methods have been proposed that are related to word counting methods, but which incorporate a probabilistic model of how sequences are generated from a dictionary of possible words [35,240,285].

The motif detection algorithms using a probabilistic model, aim at constructing a multiple alignment by locally aligning small conserved regions in a set of unaligned sequences. Once again, these algorithms can largely be divided into two different approaches. The oldest motif detection tools use the Expectation-Maximization (EM) method [15,137]. EM is a local optimization procedure that is guaranteed to monotonically improve the expected likelihood, but is sensitive to its initialization point and is therefore not guaranteed to converge to the global maximum. However, heuristics for reasonable initialization points may solve this problem partially [27,152].

Gibbs Sampling is the stochastic variant of the EM algorithm [136,189]. Because of the stochastic nature of the Gibbs Sampling methods, this approach does not suffer from getting stuck in a local optimum. However, due to the stochastic nature of the approach, the algorithm may have to be run for many iterations to obtain adequate results. Different implementations of this approach exist, where different extra features have been added (automatic detection of length, higher order background model, …) [113,146,268,301].

1.3.2.2 Algorithms using additional data

Recent advances in experimental technology like microarrays and Chromatin ImmunoPrecipitation (ChIP-chip data) have influenced the field of motif detection. Where traditional motif detection algorithms only use these experimental data to define their input data sets, recent algorithms have integrated these data in their search for regulatory motifs.

- **microarray data**: Initially, microarray data were only used to create clusters of coexpressed genes, from which the promoter regions were used as input for motif detection algorithms. However, different algorithms have been developed that simultaneously use microarray data and sequence data, often applying regression-based techniques [36,44,127].
- **ChIP-chip data**: ChIP-chip experiments measure the binding of a particular protein to DNA using microarray technology. Analyzing the data with motif discovery programs can reveal motifs bound by the studied regulator. This new technique has been used as source for identification of co-regulated genes and has recently been applied in a number of studies, mainly to eukaryotes [102,129,147,217,218]. However, this sort of data can also be used in a more advanced way – using regression techniques similar to microarray data – simultaneously with sequence data to detect regulatory motifs [241].

- **Transcription factor structure**: Information about transcription factor structure can improve motif discovery results and reveal connections between specific transcription factors and motifs. The structure of a DNA-binding protein is closely linked to the motifs it binds and specialized algorithms have been developed that take advantage of this knowledge [23,70,156,258,309].

1.4 Transcriptional Regulatory Networks

1.4.1 Structure and function of regulatory networks

Advances in sequencing and generating high-throughput expression data have created a situation in which it is possible to integrate these different data sources for the deciphering of transcriptional regulatory networks. For prokaryotic organisms, the primary role of transcriptional regulation is the response to changes in environmental conditions such as nutritional status and environmental stresses. Owing to the central role that transcriptional regulation plays in cellular function and the availability of different data sources, reconstruction of these networks has emerged as a key task in biology [252,306].

The basic functional element of a regulatory network is the promoter region of a gene or operon, which contains the cis-regulatory binding sites for the relevant transcription factors that regulate the expression of a particular gene. The location and orientation of these binding sites, as well as the affinity of the transcription factor to particular variants of the binding site, determine the expression level of a gene in response to changes in the active concentration of the transcription factor inside the cell. The transcriptional regulatory network is then defined by which transcription factors bind to which promoters and what the integrated effect of all these transcription factors is on the expression of all genes [109]. These regulatory networks can be visualized using directed graphs with transcription factors and target genes as nodes, and regulatory interactions as edges [92,237].

1.4.2 Reconstruction of regulatory networks

Deriving full regulatory network structure solely based on experimental data appears to be challenging. However, large quantities of high-throughput data permit this kind of analysis. Different combinations of expression data, ChIP-chip data and motif analysis data have allowed the generation of hypothetical regulatory network structures using a variety of data integration methods in well studied organism like yeast and *E. coli* [18,46,104,140,151,236]. One alternative to this purely data-driven approach would be to utilize well-curated regulatory network structures (derived from databases and literature) as a starting point for expanding the network on the basis of high-throughput data [100,108,313]. In the future, such combinations of knowledge-driven and data-driven regulatory network construction strategies may allow the acceleration of network reconstruction in well-studied organisms [109].

1.4.3 Evolution of regulatory networks

Evolution of transcriptional regulatory networks can occur at different levels. At the level of the binding sites for regulatory proteins, these regulatory motifs seem to evolve in a step-wise manner accumulating different mutations, with loss and gain of individual interaction probably playing a greater role than loss and gain of whole motifs [12]. This is supported by the results of Wray *et al.* [305] showing that small changes in transcription factor binding sites can lead to major changes in transcription factor binding affinity. Another nice example of the evolution of regulatory motifs, is the work on the regulation of ribosomal protein genes in various yeasts. It was shown that there was a gradual sequence divergence in the regulatory motif of the involved transcription factor where the degree of divergence was proportional to the distances in the phylogenetic tree [260].

Evolutionary studies have also been performed at the global level, i.e. the evolution of transcriptional regulatory networks. Several of these studies have shown that both transcription factors and target genes can evolve by duplication, with inheritance of the regulatory interactions [162,198]. About 45% of the interactions in the known *E. coli* and yeast networks can be attributed to this mechanism [263]. However, although there are numerous examples of recently duplicated transcription factors and their targets, another explanation might be the massive horizontal transfer of fragments containing both regulated genes and their regulators. This hypothesis is supported by the observation that recent horizontal gene transfers of bacterial enzymes outnumber duplication events by about tenfold. When using comparative genomics in reconstructing transcriptional regulatory networks, it is evident that genes and regulatory interactions are conserved to varying degrees in closely related organisms, and this can be exploited to reconstruct the regulatory networks of poorly characterised organisms[81,312]. This approach also lead to the observation that transcription factors are less conserved than target genes which suggests that regulation of genes evolves faster than the genes themselves [12,162].

1.5 Overview of the thesis

Bacteria are dynamic organisms, able to survive in different environmental conditions. In order to adapt their cellular machinery to continuously changing conditions, bacteria are equipped with flexible regulatory networks.

As in bacteria the rate of transcriptional initiation is an important check point for control of gene expression, we focus in this thesis on unraveling the regulatory mechanism responsible for the transcriptional control. The basic functional element of a transcriptional regulatory network is the gene's promoter region which contains the regulatory binding sites for the transcription factors that regulate its expression. Over the past years, considerable effort has been put in the in silico identification of these regulatory binding sites, which resulted in a diverse range of motif detection methods [**chapter 1**]. With the availability of entire genomes new opportunities opened up for comparative genomics and motif detection. Motif detection methods based on comparative genomics (phylogenetic footprinting) exploit the conservation of motifs in orthologous promoter regions based on the idea that evolutionary forces tend to preferentially retain the biologically functional DNA sequences [**chapter 2**].

In this PhD we used the concept of phylogenetic footprinting to extend the information on two poorly characterized regulons involved in the pathogenicity of *Salmonella typhimurium*. For the PmrAB regulatory system, several novel targets were detected by our in silico analysis, a few of which were validated by experimental wet lab analysis [**chapter 3**]. The PhoPQ systems, a sensor for magnesium ions and an important regulator of virulence genes in some pathogenic bacteria, were further characterised by combining microarray data with in silico motif prediction. By comparing to what extent this regulon overlapped between *Salmonella typhimurium* and its close relative *Escherichia coli* we could show that the PhoPQ two-component system seemingly quickly adopted novel targets during evolution, possibly giving rise to the difference in phenotypes between the two related species [**chapter 4**].

The fact that both regulons mentioned above were already partially characterized facilitated their analysis. However, if one wants to identify regulatory motifs without any prior information, one has to rely on the mere property of "statistical overrepresentation". In these cases, the existing motif detection tools will fail if the involved species are evolutionary too related or if the regulatory motifs are present only in a limited subset of genes. For this reason, we developed an adapted version of MotifSampler that allows detection of niche- or species-specific regulatory motifs or motifs that belong to sparsely connected hubs in the regulatory network [**chapter 5**].

The tools developed in this PhD study all apply to the identification of regulatory motifs. As the detection of regulatory motifs is complicated because they are short, degenerated and only present in a limited number of promoter regions, we can apply theses tools to biological questions facing the same limitations. We illustrate the wide application area of our tools by detecting potential targets of regulatory RNA [**chapter 6**] and by detecting small signalling peptides [**chapter 7**].

Type	Achievement	Pub.
Chapter 2: Motif detection using comparative genomics		
literature	Description of motif detection methodologies using comparative genomics	
Chapter 3: Identification of the PmrAB regulon		
biological	detect new targets of the PmrAB regulon and create a more accurate motif model of the PmrAB regulatory motif	[160]
software	two-steps method for phylogenetic printing based on 1) motif detection procedure based on Gibbs Sampling 2) generate multiple alignments with the output motifs of the Gibbs Sampling step as seeds	
Chapter 4: Comparative analysis of the PhoPQ regulon		
biological	comparison of the composition of the PhoPQ regulon between *E. coli* and *S. typhimurium,* well conserved regulatory system can evolve very quickly resulting in novel phenotypes	[178]
software	methodology to compare the composition of regulons between two species	
Chapter 5: More robust detection of motifs in coexpressed genes by using phylogenetic information		
software	*de novo* motif detection method based on comparative genomics to detect regulatory motifs that 1) belong to sparsely connected hubs in the regulatory network 2) are derived from orthologous promoter regions from evolutionary closely related bacteria	[179]
Chapter 6: Detection of regulatory RNA targets		
biological	identify potential targets of the *sraD* regulatory RNA	
software	screening method to identify potential target genes of regulatory RNA	
Chapter 7: Detection of double glycine leader sequences		
biological	detect GG-peptides and their corresponding secretion system in Gram-positive and Gram-negative bacteria	[55,56]

Table 1.2: **Overview of the different chapters with the achievements.** Level indicates the level of the achievement: literature overview, developing software or solving biological problems. Pub. Refers to the corresponding article where the study is published.

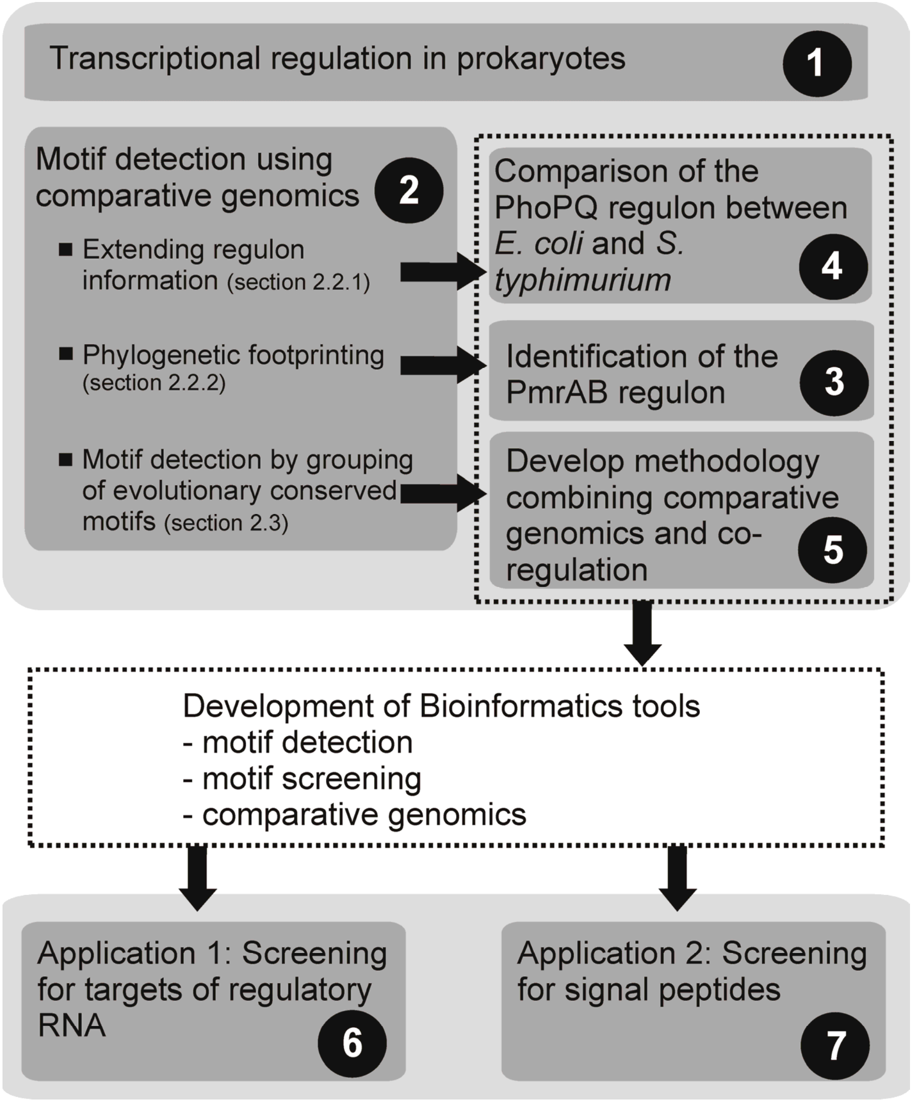

Transcriptional regulation in prokaryotes
1
Motif detection using comparative genomics
2
Extending regulon information (section 2.2.1)
Phylogenetic footprinting (section 2.2.2)
Motif detection by grouping of evolutionary conserved motifs (section 2.3)
Comparison of the PhoPQ regulon between E. coli and S. typhimurium
4
Identification of the PmrAB regulon
3
Develop methodology combining comparative genomics and co-regulation
5
Development of Bioinformatics tools
- motif detection
- motif screening
- comparative genomics
Application 1: Screening for targets of regulatory RNA
6
Application 2: Screening for signal peptides
7

Figure 1.6: **Overview of the thesis.** Chapter numbers are mentioned in the circles. Chapter 1 and 2 give a literature overview on transcriptional regulation and motif detection in bacteria. Chapter 3 until 5 describe the results of regulatory motif detection in bacteria using comparative genomics approaches and motif detection tools. In chapter 6 and 7, tools developed in the previous three chapters are used to solve biological problems similar to regulatory motif detection.

Chapter 2

Motif Detection using Comparative genomics

2.1 Introduction

With the sequencing of multiple complete bacterial genomes, computational biology entered a new era. The availability of the sequences of all genes in a wide range of prokaryotic species created the opportunity to compare at large scale different prokaryotic genomes. Initially, the main efforts have been directed at comparison of genomes at protein-encoding level with the aim of identifying all genes and their corresponding function in well-studied bacteria as well as poorly characterized ones [114,133,192]. However, an adequate understanding of bacterial cell functioning is impossible without a knowledge of the transcription regulatory networks. Therefore, an important further step in the functional annotation of genomes is the identification of regulatory signals, i.e. binding sites for transcription factors. Although the problem of prediction of regulatory sites has been addressed over more than 15 years, it is still far from being solved [272]. In the first chapter, we briefly overviewed motif detection algorithms that search for statistically overrepresented motifs in a search space reduced by coexpression. In this chapter we focus on the motif detection methods using comparative genomics. These methods are all based on the idea that evolutionary forces tend to preferentially retain regulatory motifs. If a regulatory motif is conserved in the promoter regions of orthologs derived from evolutionary related organisms, it is more likely to be biologically relevant. The search space is thus reduced by orthologous relationships.

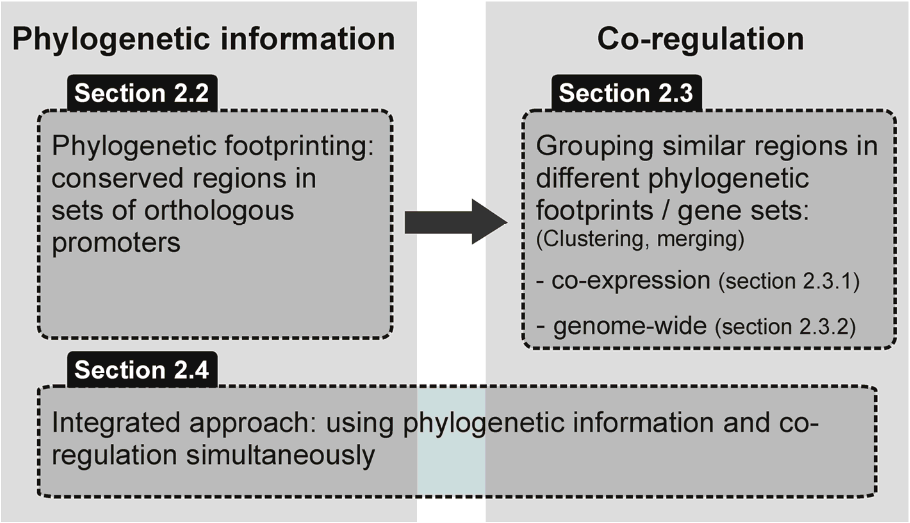

Phylogenetic information
Co-regulation
Section 2.2
Phylogenetic footprinting: conserved regions in sets of orthologous promoters
Section 2.3
Grouping similar regions in different phylogenetic footprints / gene sets:
(Clustering, merging)
- co-expression (section 2.3.1)
- genome-wide (section 2.3.2)
Section 2.4
Integrated approach: using phylogenetic information and co-regulation simultaneously

Figure 2.1: **Overview of chapter 2.**

2.2 Comparative Genomics for motif detection: applications

Comparative genomics has been used in the past in different ways for motif detection. In the beginning, comparative genomics was only used to extend the limited information we have about known regulatory motifs and corresponding regulons in order to get a more detailed view on the studied regulon. At a later stage, comparative genomics was also used to detect new regulatory motifs. Anyhow, in both cases the underlying assumption in all these approaches is that regulation tends to be – at least partially – conserved in evolutionary related bacteria.

2.2.1 Extending regulon information

The first time comparative genomics was used for motif detection, had as goal to augment the knowledge about partially characterized regulons in *E. coli*. A regulon is defined as a set of genes that is controlled by the same transcription factor. Mironov *et al.* [176] combined genome comparisons with motif screening methods to analyze the evolutionary conservation of the purine (PurR), arginine (ArgR) and aromatic amino acid regulons (TrpR and TyrR). In this study, the information on the transcription factor binding sites was derived from *E. coli* and was used to predict binding sites and additional target genes in *E. coli* and to find their corresponding regulons in *Haemophilus influenzae* which is closely related to *E. coli*. A target gene was considered relevant if a motif instance was found in both orthologs. Based on this comparative analysis, the authors make a suggestion on the different types of evolutionary mechanisms than can explain the variation in regulon composition between both species 1) presence and absence of individual genes in otherwise conserved operons. 2) breaking operons into two parts, both of which retain regulation. 3) loss or switch of regulation in operons. An approach based on the same two species was also used to predict the regulons controlled by the fumarate and nitrate reduction protein (FNR) and the cAMP receptor protein (CRP) [259]. However in this study, evidence of comparative analysis was only required for putative target genes where initial motif screening returned a low score. Putative target genes with a high motif screening score did not need this in silico evidence.

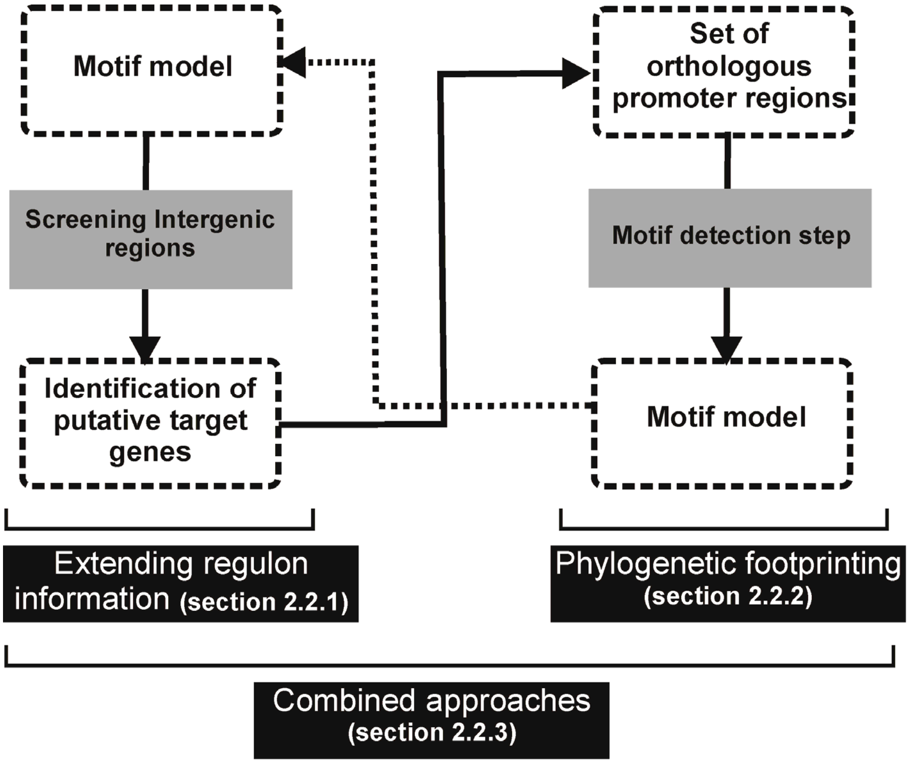

Motif model
Screening Intergenic regions
Identification of putative target genes
Set of orthologous promoter regions
Motif detection step
Motif model
Extending regulon information (section 2.2.1)
Phylogenetic footprinting (section 2.2.2)
Combined approaches (section 2.2.3)

Figure 2.2: **Schematic overview of section 2.2.** The left panel of the figure displays the flow when comparative genomics methods are used to extend information on regulons (section 2.2.1). The right panel is a schematic representation of phylogenetic footprinting (section 2.2.2). Both approaches can also be combined (section 2.2.3) where the motif model obtained from the phylogenetic footprinting step is used to screen intergenic regions to identify putative target genes (dotted line), or orthologous promoter regions can be selected from the putative target genes and validated using a phylogenetic footprinting step (full line).

A similar study using comparative genomics in archaea was performed by Gelfand *et al.* [82]. Starting from the limited experimental data on archaeal transcriptional regulation or using analogy with known bacterial regulons, putative motif models were constructed representing the binding sites of different regulatory proteins. Using a comparative genomics approach exploiting six fully sequenced archaeal genomes, the authors were able to predict four different regulons (nitrogen fixation, heat shock, purine metabolism, aromatic amino acid synthesis) in the different archaeal species. The same strategy was applied to reveal the complete Fur regulons in gamma-proteobacteria [196] and zinc regulons in proteobacteria and Gram-positive bacteria from the *Bacillus*-group [195]. These results demonstrate the feasibility of prediction of at least some transcriptional regulatory sites by comparing poorly characterized prokaryotic species, particularly when closely related genome sequences are available.

Similarly we compared the PhoPQ regulon between the evolutionary related organisms *E. coli* and *S. typhimurium* using a similar approach to study the evolution of this regulon [178]. Details about this observation are described in chapter 4.

2.2.2 Phylogenetic footprinting

The studies discussed in 2.2.1 mainly focussed on methods to augment known regulons in well-studied species and to identify their counterparts in lesser-studied species. However, comparative genomics can also be used to detect new regulatory motifs using a phylogenetic footprinting approach. In this phylogenetic footprinting approach, traditional motif detection algorithms are used to detect *de novo* regulatory motifs in sets of orthologous promoter regions.

McCue *et al.* [167] were the first to describe a phylogenetic footprinting algorithm, eliminating the needs for the identification of co-regulated genes or knowledge of any prior information on the regulon. Given the promoter sequence of an individual gene and the promoter regions from a number of corresponding orthologs in evolutionary related species, their method permits the detection of regulatory motifs. Applying an extended Gibbs Sampling algorithm [136] to a study set of 184 *E. coli* genes with documented regulatory motifs, revealed that when orthologous data were available from at least two other gamma- proteobacterial species, 81% of the predictions corresponded with documented sites. If only one other species was available, 67% corresponded with documented sites. Next, this phylogenetic footprinting procedure was applied genome-wide to 2097 data sets of orthologous promoter regions consisting of *E. coli* genes and their respective orthologs in nine other gamma-proteobacteria.

2.2.3 Combined approaches

Where in previous section comparative genomics was used to extend the information about known regulons (2.2.1) or to detect new regulatory motifs using phylogenetic footprinting (2.2.2), many studies use a combination of both approaches. A first type of these combined approaches uses the output of a phylogenetic footprinting step to execute a genome-wide screening (dotted line in figure 2.2). The second type of methodologies first performs a genome-wide screening and uses phylogenetic footprinting as an *in silico* evidence for the putative targets (full line in figure 2.2).

An example of a combined approach was the comparative study of the biotin regulons in all available prokaryotic genomes (46 at the moment of the study) [221]. Based on a phylogenetic analysis of the DNA binding domains of all BirA regulatory proteins (i.e. the transcription factor controlling the biotin regulons), the genomes were divided into two major groups, proteobacterial and non-proteobacterial. Consistent with this, two different sets of promoter regions were subjected to a motif detection step resulting in two different regulatory motif models. The newly detected motif models were used to detect new candidate members of the BirA regulons by screening intergenic regions of all genomes and identifying those genes for which the BirA motif is conserved in its orthologs. Evolutionary comparison of these computationally identified BirA regulons showed that the BirA regulon is widely distributed in prokaryotes and suggests that BirA regulatory systems for biotin biosynthesis are ancient.

McGuire *et al.* [169] used this phylogenetic footprinting strategy simultaneously with a traditional motif detection step performed on a set of coregulated genes. They conducted a systematic search for novel transcription factor binding sites using 17 complete microbial genomes. First, potentially coregulated genes were grouped based on three methods: 1) genes that make up functional pathways; 2) genes homologous to regulons from a well-studied species (e.g. *E. coli*); 3) groups of genes derived from conserved operons. Next, the AlignACE motif detection algorithm [113] was applied to the sets of coregulated genes in each of the 17 species. If a motif was detected in a set of coregulated genes of one of the 17 species, for each of these genes the upstream sequences from the corresponding orthologs were selected and on their turn subjected to a motif detection step with AlignACE. If the same motif was found with this phylogenetic footprinting step as with the first coregulated strategy, the authors put higher confidence in these predictions as representing true biological binding sites.

We used a similar approach in studying the PmrAB regulon in *S. typhimurium* [160]. In a first step, the binding site of the regulatory protein PmrA was detected based on a traditional motif detection step with MotifSampler [268] on a set of known PmrAB-regulated genes. Genome-wide screening of promoter regions with the PmrA motif model returned all possible direct targets of the PmrAB system. All putative targets were subjected to a phylogenetic footprinting step to increase the confidence in our predictions. Detailed information about this study can be found in chapter 3.

2.3 Motif detection by grouping of evolutionary conserved motifs

Computational strategies for the discovery of regulatory sites began with algorithms that identified regulatory motifs in sets of coexpressed or functionally related genes (see chapter 1) or phylogenetic footprinting approaches (see 2.2). Building on top of the concept of phylogenetic footprinting is the idea that motifs discovered in one set of orthologous promoter regions often reoccur in other sets of orthologous promoters. The challenge of advanced motif detection methods is to group those motifs coming from different phylogenetic footprinting outputs that are similar enough and thus represent the same transcription factor binding site. Methods following this approach can largely be divided into two groups. The first group requires a relatively small dataset of genes as input from which is known that a subset of these genes is coregulated. The second group works at a genome-wide level and as such does not need any information on co-regulation. However, the steps that need to be performed are the same in both applications: the first step is always the phylogenetic footprinting step returning local alignments of the orthologous promoter regions, where the second step groups those phylogenetic footprints that represent the same regulatory binding site.

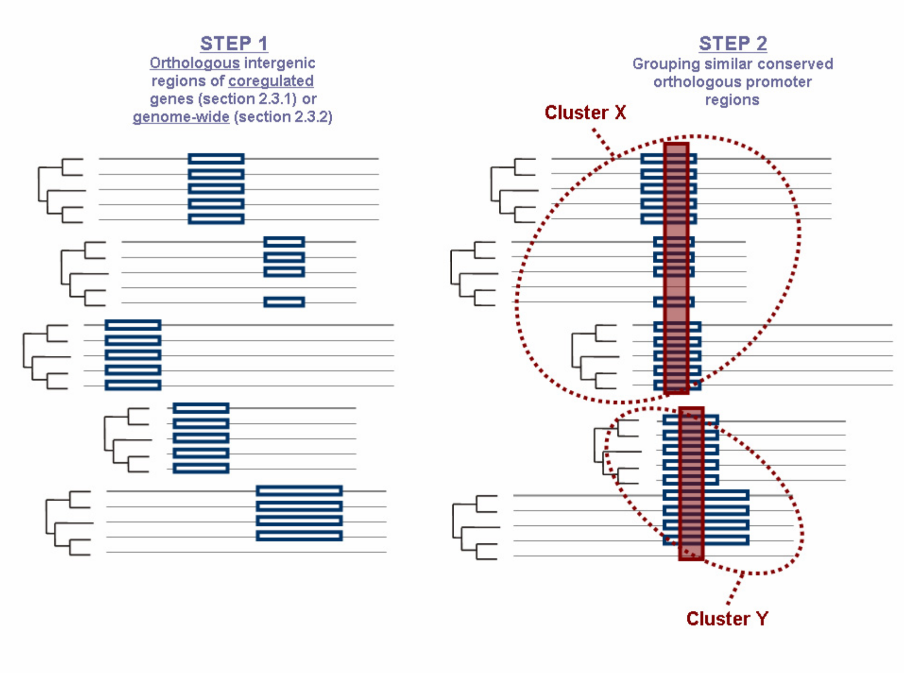

STEP 1
Orthologous intergenic regions of coregulated genes (section 2.3.1) or genome-wide (section 2.3.2)
STEP 2
Grouping similar conserved orthologous promoter regions
Cluster X
Cluster Y

Figure 2.3: **Two steps when combining phylogenetic footprinting and coexpression.** The left panel of the figure displays the phylogenetic footprinting step. The blue boxes indicate the conserved sequences in the different sets of orthologous promoter regions. The right panel displays the clustering of those conserved regions resulting from different sets of orthologous promoter regions that share a similar regulatory motif. The red box contains the part of the conserved promoter regions that is shared between the different members of the cluster.

2.3.1 Motif detection using comparative genomics and coexpression

The rapid accumulation of genomic sequences from multiple organisms and the mounting evidence from microarray gene profiles make it possible to combine knowledge of co-regulation among different genes with conservation among orthologous genes to improve motif finding. Wang and Stormo [295] developed PhyloCon which is an algorithm that is capable of combining both data sources This two-step procedure starts with aligning orthologous intergenic sequences and creating position specific scoring matrices (called profiles) based on the Wconsensus program [110]. Then PhyloCon compares these profiles generated from different genes and identifies the common regions in these profiles using a greedy approach. Comparison of these profiles is scored with a new score called Average Log Likelihood Ratio that measures the likelihood of the observed nucleotide frequencies in column *i* from one matrix being generated by the distribution of nucleotide frequencies of column *j* in a second matrix. Based on a Waterman-Smith approach the highest local alignment is obtained by setting all negative scoring column pairs to zero and tracking back from the highest scoring column pair to the first positive pair. Profile comparison results are used in a merging step that combines all those profiles and returns a common region between all these profiles that potentially represents a regulatory motif. Their method was tested on simulated data and biological data.

We developed a methodology that is based on the same idea of combining phylogenetic information and co-regulation, however using different algorithmic components [179]. The first step of our algorithm is a phylogenetic footprinting step where we use a Gibbs Sampling based algorithm called BlockSampler. The second step consists of aligning all conserved regions of different sets of orthologs using BlockAligner, which uses a Waterman-Smith approach with the Kullback-Leibler distance as scoring function. The third step consists of a clustering of the conserved regions based on the Kullback-Leibler distance obtained with BlockAligner. Conserved regions grouped together in the same cluster are merged to one recurring matrix representing a putative regulatory motif. The last step consists of screening all input intergenic regions with the retrieved motif models to recover those instances of the motif that were missed.

For a detailed description of our approach, we refer to chapter 5 where all components of our methodology are described. In this chapter, we also perform an extensive comparison of our methodology with PhyloCon and traditional motif detection algorithms based on compiled and true biological data sets.

2.3.2 Genome-wide motif detection

The strategy applied in section 2.3.1 still requires assembling a set of coexpressed or functionally related genes before motif detection can be applied. However, this assembly of coexpressed or functionally related genes depends highly on experimental data, so limitations in experiments are propagated to the computational methods that infer motifs from data [296]. This limitation can be overcome by systematically analyzing sequence from multiple species at the whole-genome level i.e. without presumption of gene coregulation.

In general, and similar to 2.3.1, all these algorithms consist of two steps (see figure 2.3) where the first step is a traditional phylogenetic footprinting approach that result in position specific frequency matrices describing the conserved regions in sets of orthologous promoter regions. For this phylogenetic footprinting step, results are derived from previous studies [167,215] for the implementation of Qin *et al.* [214] and Van Nimwegen *et al.* [287]. Jensen *et al.* [120] and Alkema *et al.* [3] applied a Gibbs Sampling approach to sets of orthologous promoter regions, while Stormo and Wang [296] apply the Wconsensus algorithm [110] to detect phylogenetic footprints.

The second step consists of a clustering of all the conserved promoter regions that share the same regulatory motif [3,110,120,214,287]. Three of these algorithms implemented a Gibbs Sampling based clustering algorithm to group those conserved promoter regions – represented as matrices – based on the similarity between all these matrices [120,214,287]. Similar with a Gibbs Sampling approach in motif detection, the algorithm begins by randomly assigning all matrices (i.e. conserved regions in one set of orthologous promoters) into an arbitrary number of clusters. On each iteration the algorithm considers the reassignment of one of the matrices to another cluster, given the current assignment of all other regions to different clusters. Therefore, the probability that this matrix belongs to one of the existing clusters, or begins a new cluster itself, is calculated. In the sampling step, the matrix is assigned to one of these clusters by sampling from this probability distribution, and the clusters are updated according to this sampling step. This process of calculating probabilities and sampling from this distribution is repeated several times, each time for a different matrix, until convergence is reached. In contrast with the Gibbs Sampling clustering algorithm, Wang and Stormo [296] and Alkema *et al.* [3] both use a variant of the Neighbor-joining method in their clustering step. This implies that in each step of the clustering, all matrices are compared with each other, and the most similar ones are merged into a new matrix. This merging step is repeated but the previous two matrices are now substituted by the merged matrix. Merging of these matrices ends when no similar matrices are detected anymore. The resulting merged matrices are the summarized representation of the different clusters.

Despite the fact that at first sight all implementations look similar, every algorithm has it own drawbacks. A reoccurring limitation in the clustering algorithm is the presumption that all potential regulatory motifs have a fixed width [3,120,214,287]. Requiring a fixed length for the phylogenetic footprints puts limits on the evolutionary distance that is allowed between different organisms. This inhibits the application of the algorithm to phylogenetic footprints derived from evolutionary closely related species as these footprints may vary in size from small conserved motif to long stretches of evolutionary conserved regions what makes it difficult to assigned a fixed length (see chapter 3 and 5). The only algorithm that does not require a fixed width of the motif, is the PhyloNet algorithm of Stormo and Wang.[296]. Two of these genome-wide applications have extra algorithmic limitations. The implementation of Qin *et al.* [214] requires to choose a priori between a palindromic or non-palindromic motif model. A similar limitation also applies to the algorithm of Jensen *et al.* [120] where the user has to choose between a one-block or two-block (i.e. gapped or dyad) motif model.

All of these genome-wide algorithms are tested by comparing their output with a set of known regulatory motifs in well studied model organisms. Two of the studies used *E. coli* as model organism [214,287] where genome-wide output was benchmarked with the DPInteract database [220]. Alkema *et al.* [3] and Jensen *et al.* [120] used *Bacillus subtilis* as model organism and genome-wide output was compared with the DBTBS database containing all known regulatory motif in *B. subtilis* [159]. The last model organism that is used, is S*accharomyces cerevisiae* by Wang and Stormo for benchmarking their PhyloNet algorithm [296].

An important issue in all these algorithms is the selection of the species involved in the phylogenetic footprinting step [168]. For all test cases, the species are selected out of an evolutionary diverse set of species. In Van Nimwegen *et al.* [287] and Qin *et al.* [214], representative species are selected over almost the complete evolutionary range of proteobacteria. In Jensen *et al.* [120], for more than 50% of the *B. subtilis* genes no ortholog is detected in any of the selected species used in the phylogenetic footprinting step. As all the algorithms described above are applied to species that are evolutionary distant, detecting regulons that are present in only a small set of closely related organisms is a difficult task, and consequently, regulatory motifs that are missed in the genome-wide output are typically these motifs that belong to such a specific regulatory system. This problem can be solved by selecting species that are evolutionary more related, which improves the detection of such subtle signals. However, this complicates the correct delineation of a true biological motif as not only the regulatory motif itself is conserved, but also the flanking regions because of evolutionary relatedness. We developed an algorithm that successfully tackles these problems, and is discussed in chapter 5. A second way to handle this problem is the use of phylogenetic information. In the algorithms mentioned above, the phylogenetic footprinting step does not take into account the phylogenetic information, i.e. each sequence within a particular set of orthologous promoter sequences is weighted equally despite the fact that these sequences come from different species with unequal phylogenetic distances between them. This kind of information is exploited when using the algorithms mentioned in paragraph 2.4, where the evolutionary distances are incorporated in the motif detection procedures. This will permit to use a mixture of evolutionary related and evolutionary distant organisms as input for the motif detection algorithm without a decrease in performance. These adaptations will allow detecting weaker motif signals, as the phylogenetic footprinting step starts detecting motifs in the most evolutionary related organisms and the motif detection step can be terminated when the motif is not conserved in evolutionary more distant organism, retaining the species- or lineage-specific motif. This should permit to detect regulatory motifs only present in a subset of related organisms, as well as motifs conserved in a wide range of organisms.

The second step applied in all five algorithms, i.e. clustering of similar conserved regions between different sets of orthologous promoter regions, also causes the failure of detection of some regulatory motifs, namely those regulons that only contain a limited number of genes (e.g. a transcriptional regulatory system that controls less than five genes). All five algorithms require the conservation of the regulatory motif in at sufficiently high number of different sets of orthologous promoter regions before detection is possible. As mentioned in chapter 5, the algorithm we developed is capable of detecting regulatory motifs conserved in a limited number of genes. However, we still need to prove that our algorithm still can detect such motifs lacking a high degree of overrepresentation when extended to a genome-wide scale.

2.4 Integration of phylogenetic distances

In previous section we described motif detection algorithms that combine co-regulation and phylogenetic information in a sequential way. Recently, novel algorithms are developed that combine simultaneously the search for overrepresented motifs with incorporation of phylogenetic information.

The most basic implementation that displays this feature is orthoMEME [209] which is based on the Expectation-Maximization (EM) algorithm MEME (see chapter 1). During the motif detection procedure, a motif instance is only considered as a true biological motif if it has an orthologous copy in one other species. Indeed, orthoMEME can only handle two species, i.e. one reference species in which one needs to find regulatory motifs, and a second species used as phylogenetic validation of the detected motif instances.

Previous algorithms – including orthoMEME – threat orthologous sequences as statistically independent observations ignoring the fact that all orthologous promoter sequences descend from a common ancestor. Therefore, algorithms are developed that take into account the varying phylogenetic distances among the species (i.e. the evolution of DNA sequences) by implementing a probabilistic model of evolution [183,238,239]. By using this probabilistic model, the phylogenetic information is integrated in the scoring function of the advanced motif detection algorithms. If during the motif detection step a motif model is detected for which the variation in motif instances in a set of orthologous promoters corresponds with the phylogenetic tree, a higher confidence, and thus a higher score, is assigned to this motif model. The general strategy for the *de* novo detection of regulatory motifs is similar to the original algorithms: PhyME [239] and EMnEM [183] use an EM-algorithm and PhyloGibbs uses a Gibbs Sampling strategy [238]. The major disadvantage of all these advanced algorithms is the need for creating multiple alignments of orthologous promoters (i.e. the algorithms still require two steps). These alignments are used to check the conformity of the detected motif instances with the phylogenetic tree. However, creating these multiple alignments in this first step implies setting an arbitrary cut-off alignment score to identify candidate functional sites.

The most advanced implementation that integrates phylogenetic information and co-regulation simultaneously is the algorithm of Li and Wong [144]. Their method also does take into account the phylogenetic information and in addition does not depend on the alignments of orthologous sequences to obtain candidate motif instances as this step is fully integrated in their algorithm. Instead of using aligned regions they implemented a Gibbs Sampling approach where all orthologous and co-regulated sequences are searched simultaneously for conserved motif instances. In an initialization step, motif instances are assigned at random in every sequence present in the input (i.e. at the leaves of the phylogenetic tree). Based on a substitution matrix, starting from these instances, an instance can be derived at the different nodes and ultimately at the root of the phylogenetic tree that resembles the ancestor binding site for that set of orthologous sequences. Using a Gibbs Sampling approach, all promoter regions in one random chosen set of orthologous promoter regions are screened, the instances in these promoter sequences are updated by a sampling step based on the screening score, a modified root instance is calculated based on these updated instances, and the motif model at the root level (calculated based on the root instances of all orthologous promoter sequences) is updated by adding this new root instance. This process is iterated several times and will ultimately lead to convergence resulting in one motif model describing a potential regulatory motif. Because the updating and sampling step are similar to the steps used in traditional Gibbs Sampling motif detection, the authors refer to their method as sampling motifs on a phylogenetic tree [144].

Chapter 3

Identification of the PmrAB regulon

3.1 Introduction

The most straightforward approach for the elucidation of a transcriptional regulatory system is the use of experimental techniques (e.g. site-directed mutagenesis). However, such research is laborious and time-consuming. As the methods for comparative genomics have developed and bacterial genomic data have been accumulated, an in silico methodology can be applied to identify transcriptional regulatory systems. Such a bioinformatics approach can point out interesting putative targets of a regulatory protein based on the detection of its corresponding binding site in promoter regions. The use of a cross-species comparison can increase the confidence in these putative target genes. Starting from limited experimental information available in literature on the binding site of the PmrAB regulatory system, we developed a new bioinformatics approach based on comparative genomics to suggest new putative PmrAB target genes. In collaboration with the Centre for Microbial and Plant Genetics (Prof. J. Vanderleyden, Dr. S. De Keersmaecker), four of these putative targets were experimentally validated demonstrating the power of our bioinformatics approach. This work – together with the details on the experimental validation – has been published in Genome Biology [160].

The PmrAB two-component regulatory system is part of a multicomponent feedback loop that acts as one of the key regulatory mechanisms of *Salmonella typhimurium* virulence [97,125,315]. The PmrAB regulatory system is itself responsive to Fe^{3+} and mild acid [304] and senses Mg^{2+} indirectly by communicating with the Mg^{2+}-sensitive PhoPQ system [77,88,246,247] via PmrD [125,134]. PmrD is hypothesized to transduce the signal from the PhoPQ system to the PmrAB system via a posttranslational modification. *pmrD* itself is transcriptionally activated by the PhoPQ system but repressed by the PmrAB system [31,125,134]. The PmrAB system is required for resistance to the cationic antibiotic polymyxin B [222] and to

Fe^{3+}-mediated killing [304]. The Mg^{2+}-dependent regulation of PmrAB was shown to be important for gene expression in an intracellular environment [185]. Fe^{3+}-dependent PmrAB regulation, on the other hand, has been hypothesized to be essential for survival in extracellular environments [39]. A DNA region to which the PmrA protein binds, has been identified by DNA footprinting analysis [2,303].

In contrast to *pmrD*, other known targets in *S. typhimurium* are transcriptionally activated by PmrAB. A first class of targets is involved in LPS modification. PmrAB-induced modifications include the addition of 4-amino-4-deoxy-L-arabinose (Ara4N) and phosphoethanolamine (pEtN) to lipid A [317]. Involved in the Ara4N modification of lipid A are *ugd* [247] and the *pmrHFIJKLM* loci, both responsible for Ara4N biosynthesis [30,95,97,317] and incorporation of Ara4N into lipid A [275,276]. LPS modifications are hypothesized to allow bacterial survival within macrophages by lowering the affinity for amphiphatic cationic peptides with antimicrobial activity, that are produced as a consequence of the innate immune response.

A second class consists of targets, directly dependent of PmrAB, but with yet undefined functions. *pmrC* (cotranscribed with pmrAB [96]) and *pmrG* (located upstream of the *pmrHFIJKLM* operon) are both transcriptionally activated by PmrAB. Mutations in *pmrG* did not affect the resistance to polymyxin B [97]. Recently, Tamayo *et al.* identified two additional targets of PmrAB, *yibD* and *dgoA*. However, none of these were involved in resistance to polymyxin B or to high Fe^{3+}-concentrations [257]. These genes might therefore represent a group of yet unidentified functionalities regulated by the PmrAB system. Also PmrAB-regulated genes involved in resistance to Fe^{3+} and pEtN addition to LPS remain to be identified [257]. This, taken together with the recent indications for new PmrAB-dependent functionalities, raises the possibility that not all PmrAB targets have yet been identified. Therefore, in this study we used an in silico approach to predict targets of the PmrAB regulatory system.

Several methodologies exist that allow genome-wide screenings using a motif model (or mathematical representation) of experimentally verified regulatory sites [1,110,196,249,284]. These assign to each possible motif position in the genome a score (of which the specificities depend on the methodology) that indicates how well the subsequence located at that motif position matches the motif model. Genome-wide screenings of this type have proven to be successful in detecting additional targets of the studied regulator. However, more reliable predictions on motifs for specific pathways have been obtained by incorporating cross species comparisons (phylogenetic footprinting) [57,82,135,167,168,196,221,259]. Because evolutionary forces tend to preferentially retain functional DNA sequences, motifs that are conserved in the intergenic regions of orthologs derived from several related species are more likely to be biologically relevant [75,80].

In this study, we combine both approaches. Putative targets identified by a genome-wide screening were, whenever possible, analyzed by phylogenetic footprinting based on Gibbs sampling [167-169]. Four interesting targets were validated by wet lab experiments and the PmrA box of a representative target was subjected to site-directed mutagenesis.

3.2 Results

3.2.1 Genome-wide screening using a PmrA motif model

Gibbs sampling was used to detect motifs in the intergenic regions of three experimentally verified PmrAB targets (*ugd*, *pmrC*, *pmrG*). The logo of the detected statistically overrepresented motif is represented in Figure 3.1. This retrieved motif corresponded to the PmrA binding site that was experimentally identified by Aguirre *et al.* [2] and partially overlapped the PmrA binding site delineated by Wosten *et al.* [303]. They detected this site upstream of the transcription start of *pmrC*, in the intergenic region between *pmrG* and *pmrH* and upstream of *ugd* (on the plus strand) [2,303].

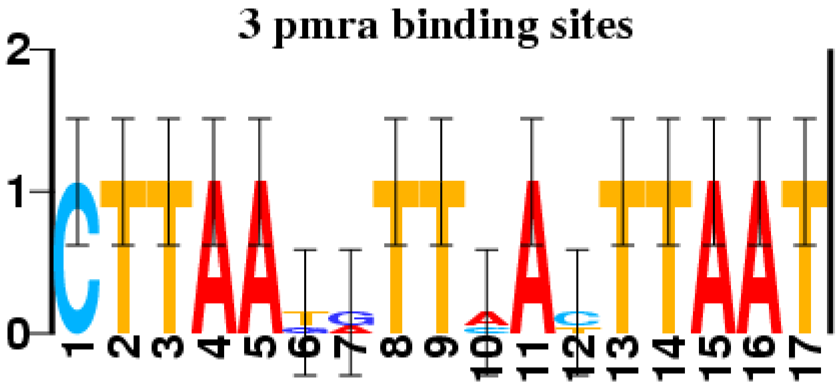


Figure 3.1: **Consensus sequence of the PmrA box.** Motif logo representing the initial motif model used to screen the S. typhimurium intergenic sequences.

We used the obtained motif model in a genome-wide screening of the *S. typhimurium* intergenic sequences [166]. Table 3.1 (column 1) summarizes the results of our screening, using a threshold as described in Materials and Methods. From experimentally verified examples, it appears that the PmrA motif can be biologically functional when present on the plus strand (e.g. in case of *pmrH*), but also when it is located on the minus strand (example *pmrG*) [303]. Therefore, both strands of the genome sequence were screened.

3.2.2 Identification of close homologs

Only in species that have a functional counterpart of the *S. typhimurium* PmrAB system, we can expect to detect conserved biologically active PmrA motifs. From all completely sequenced species, only the species (serovars) *S. typhimurium*, *Salmonella typhi*, *Escherichia coli*, *Shigella flexneri* and *Yersinia pestis*, contain the amino acids that determine the specificity of the sensor protein PmrB (amino acids suggested to be involved in binding of ferric iron [304]). Also the protein domains involved in DNA binding of PmrA, were almost perfectly conserved in the PmrA orthologs of the selected species (PF00486 domain on the Sanger website: http://www.sanger.ac.uk/Software/Pfam/). Therefore, these γ-proteobacterial species were used to perform phylogenetic footprinting analysis. For each gene containing a potential hit of the PmrA motif in the *S. typhimurium* genome sequence, close homologs were selected as described in Materials and Methods.

3.2.3 Phylogenetic footprinting using Gibbs sampling

For each dataset we aimed at constructing a local multiple alignment. We used Gibbs sampling to generate motifs that can be used as alignment seeds. Alignments were subsequently constructed based on the positions of these motif seeds. Potential seeds were selected using a heuristic described in the supplementary information. Such multiple alignments summarize the motifs that are conserved in the intergenic sequences among species. We used the alignments to verify whether the putative PmrA motifs retrieved by the genome-wide screening were conserved in other species. Table 3.1 gives an overview of the results of the genome-wide screening and the phylogenetic footprinting approach.

***Table 3.1*: List of the putative PmrAB targets in *S. typhimurium*:** Name: name of the gene in the *S. typhimurium* genome (NC_003197). For genes that are divergently transcribed and have a shared intergenic region, the gene for which the motif is detected on the plus strand is indicated first and the gene for which the motif is on the minus strand is indicated after the "/".; Description: annotation of the encoded proteins and genome location of the genes (derived from GenBank and Sanger annotation). Score: normalized score assigned to the respective motifs by MotifLocator; Site: instance of the motif as detected in the respective intergenic sequence. Distribution (COG): distribution of the protein as determined by our analysis. The distribution is indicated by a binary profile that indicates the presence 1 versus absence 0 of the protein in species (serovars) of respectively *Salmonella*, *E. coli*, *Shigella* and *Yersinia* (e.g. 1111 protein present in all 4 species; 1000: protein present in Salmonella species only). Distribution (McClelland *et al.*): distribution of the protein encoded by the corresponding gene in 9 bacterial genomes as determined by McClelland *et al.* [166]. Proteins, having close homologs in at least one *Salmonella* strain but not in *E. coli* or *K. pneumoniae*, are indicated by "some *Salmonella* only". Genes that contain close homologs in all genomes are indicated by "all nine genomes". Other combinations are indicated by "other distributions". "?" indicates that the authors were not certain about the statement. Differences between the distribution as determined by McClelland *et al.* [166] versus the one determined by our analysis is due to the difference in selection criteria used to identify close homologs (see Material and Methods). Alignment: indicates whether the intergenic regions in the dataset could be locally aligned (nd: no local alignment detected that contained the original sequence of *S. typhimurium*, m: local alignment detected; If the dataset only contained homologs from *Salmonella* species, local alignments were considered non-informative (indicated by /)). Footprint: denotes whether the PmrA motif is conserved in the close homologs. +: the retrieved putative PmrA motif is conserved. -: the intergenic sequences of the orthologs could be locally aligned but the PmrA motif was not part of the conserved regions. Most promising PmrAB targets that contained a PmrA motif matching the PmrA consensus (Fig. 3.4) are in bold face. PmrA motifs that are experimentally validated in this study are indicated by an "*".

Name	Description	Score	Instance	Alignment	Footprint	Distribution COG	Distrubution ref
minus strand							
STM1273	putative nitric oxide reductase	0.848436	CTTAATGTTTTCTTAAT	/	/	1000	All *Salmonella* only
STM2132	pseudogene; frameshift; Putative RBS for *STM2133*;	0.814252	TTTTAGATTCACTTAAT	/	/	1000	Some or all *Salmonella* only
STM4596	Paralog of *E. coli* orf, hypothetical protein (AAC73478.1); Blast hit to putative inner membrane protein	0.806962	TTTAATATTCACTTAAA	/	/	1000	Some *Salmonella* only
STM3131	putative cytoplasmic protein; Putative RBS for *STM3130*; putative first gene of operon with *STM3130* (putative hypothetical protein)	0.801641	CTTAATTTTTACTTATT	/	/	1000	All *Salmonella* only
STM1020	Gifsy-2 prophage	0.791616	CTTATTGTTAAGTCAAT	/	/	1000	Other distributions
stdA	*STM3029;* Paralog of *E. coli* putative fimbrial-like protein (AAC73813.1); Blast hit to putative fimbrial-like protein	0.788548	CAAAACATTAACTTAAT	/	/	1000	Subspecies 1 only?
Ugd	*STM2080; S. typhimurium* UDP-glucose 6-dehydrogenase	0.781719	CTCAGAATTAACTTAAT	m	+	1100	All nine genomes
sinR	*STM0304; S. typhimurium* SINR protein. (SW:SINR_SALTY) transcriptional regulator	0.780204	CTTGATATCATCTTAAT	/	/		Subspecies 1 only
STM3131	putative cytoplasmic protein; Putative RBS for *STM3130*; putative first gene of operon with *STM3130* ; (putative hypothetical protein	0.772846	CTTAATACTCACATTAT	/	/	1000	Other distributions
STM4413	putative imidazolonepropionase and related amidohydrolases; Putative RBS for *STM4412*; first gene of operon with *STM4412* (D-galactonate transport)	0.771153	GTGAATGTTAAATTAAT	/	/	1000	Some or all *Salmonella* only
ybdO	*STM0606;* Ortholog of *E. coli* putative transcriptional regulator LYSR-type (AAC73704.1); Blast hit to putative transcriptional regulator, LysR family	0.769839	CTTAATGTAGAGTTTAT	m	+	1110	All *Salmonella* only
oraA	*STM2828;* Ortholog of *E. coli* regulator, OraA protein (AAC75740.1); Blast hit to regulator	0.766748	CTTGATGGTAATTTAAC	m	-	1110	All nine genomes
sdhC	*STM0732;* Ortholog of *E. coli* succinate dehydrogenase, cytochrome b556 (AAC73815.1); Putative RBS for *sdhD*; first gene of putative operon encoding succinate dehydrogenase	0.765950	CTTATTATTCCCTTAAG	/	/	1000	All nine genomes

ycaR	*STM0987;* Ortholog of *E. coli orf*, hypothetical protein (AAC74003.1); Blast hit to putative inner membrane protein; Putative RBS for *kdsB*; first gene of a putative operon with *ksdB* (CMP-3-deoxy-D-manno-	0.765889	`TTCAATATTAACATAAT`	/	/	1000	All nine genomes
last	*STM4600;* Ortholog of *E. coli orf*, hypothetical protein (AAC77356.1); Blast hit to putative tRNA*tRNA methyltransferase	0.765754	`ATTTAGGATAATTTAAT`	nd	/	1110	All nine genomes
STM2137	putative cytoplasmic protein	0.764036	`TTTAACCTTAATTTAAT`	nd	/	1100	Some *Salmonella* only
STM1672	putative cytoplasmic protein	0.762904	`ATTAATAGTCACTTATT`	/	/	1000	Subspecies 1 only?
gcvA	*STM2982;* Ortholog of *E. coli* positive regulator of *gcv* operon (AAC75850.1); first gene of putative operon (*gcvA, ygdD, ygdE* containing a SAM dependent methyltransferase)	0.761166	`CTTAATGTCGAATGAAT`	m	+	1111	All nine genomes
ycgO	*STM1801;* Ortholog of *E. coli* orf, hypothetical protein (AAC74275.1); Blast hit to putative CPA1 family, Na:H transport protein	0.760685	`TTTAACATTAACATAAT`	m	+	1110	All nine genomes?
STM2287	Paralog of *E. coli* putative sulfatase * phosphatase (AAC75329.1); Blast hit to putative cytoplasmic protein	0.759519	`CTTATTATTCACATAAC`	/	/	1000	Some or all *Salmonella* only?
yebW	*STM1852;* Ortholog of *E. coli* orf, hypothetical protein (AAC74907.1); Blast hit to putative inner membrane lipoprotein	0.754895	`CTCAATGTTAACTACTT`	/	/	1000	All nine genomes?
STM0897	Hypothetical protein Fels-1 prophage	0.754468	`CGTAAGGCTCTTTTAAT`	/	/	1000	Some *Salmonella* only
lpfA	*STM3640; S. typhimurium* long polar fimbria protein A precursor; first gene of a putative fimbriae synthesis operon	0.753228	`ATTAAGAATAAATTAAT`	/	/	1000	Other distributions
Plus strand							
yjdB*	*STM4293; S. typhimurium* hypothetical 61.6 Kda protein in *basS*pmrA-adiY* intergenic region. (SW:YJDB_SALTY) putative integral membrane protein; Putative RBS for *basR*; first gene of the putative operon (*yjdB basR basS*)	0.930146	`CTTAAGGTTCACTTAAT`	m	+	1111	All nine genomes
Ugd	*STM2080; S. typhimurium* UDP-glucose 6-dehydrogenase	0.913666	`CTTAATATTAACTTAAT`	M	+	1100	All nine genomes
yfbE / ais	*STM2297;* Ortholog of *E. coli* putative enzyme (AAC75313.1); first gene of the *yfbE* operon; shared intergenic with *ais*	0.912660	`CTTAATGTTAATTTAAT`	M	+	1111	All nine genomes?

STM1269* / ST1268	putative chorismate mutase; intergenic shared with *STM1268*	0.888478	`CTTAATGTTATCTTAAT`	/	/	1000	All *Salmonella* only
STM0692	Paralog of *E. coli* nitrogen assimilation control protein (AAC75050.1); putative transcriptional regulator, LysR family	0.814773	`CTTGATGTTGATTTAAT`	/	/	1000	All *Salmonella* only
ybjG / mdfA*	*STM0865;* Ortholog of *E. coli orf*, hypothetical protein (AAC73928.1); putative permease; intergenic shared with *mdfA* (multidrug translocase)	0.810981	`CTTTAAGGTTAATTTAA`	m	+	1111	All nine genomes
STM2901	Hypothetical protein putative cytoplasmic protein; located downstream of pathogenicity island 1	0.803712	`CTTAATATCAATATAAT`	/	/	1000	Other distributions
yhjC / yhjB	*STM3607;* Ortholog of *E. coli* putative transcriptional regulator LysR-type (AAC76546.1); intergenic shared with *yhjB* (putative transcriptional regulator)	0.796967	`TTGAATATTAATTTAAT`	nd	/	1110	All nine genomes?
yjbE / pgi	*STM4222;* Ortholog of *E. coli orf*, hypothetical protein (AAC76996.1); Blast hit to putative outer membrane protein; first gene of the putative operon (*yjbE, yjbF, yjbG, yjbH*) consisting of putative outer membrane (lipo)proteins; intergenic shared with *pgi* (glucosephosphate isomerase)	0.791181	`TTTAATTTTAACTTATT`	/	/	1000	All nine genomes?
yibD*	*STM3707;* Ortholog of *E. coli* putative regulator (AAC76639.1); Blast hit to putative glycosyltransferase	0.790879	`CTTAATAGTTTCTTAAT`	m	+	1100	Other distributions
STM1926 / flhC	Putative cytoplasmic protein Putative RBS for *STM1926*; first gene of a putative operon with *yecG* (putative universal stres protein); shared intergenic with *flhC* en *flhD* (flagellar transcriptional activator)	0.790699	`CCTAATGTTCACTTTTT`	/	/	1000	Some or all *Salmonella* only
STM0334 / STM0335	putative cytoplasmic protein; shared intergenic with *STM0335*	0.789514	`TTTCATATTCATTTAAT`	/	/	1000	Some *Salmonella* only
ybdN	STM0605; Ortholog of *E. coli orf*, hypothetical protein (AAC73703.1); Blast hit to putative 3-phosphoadenosine 5-phosphosulfate sulfotransferase (PAPS reductase)*FAD synthetase Putative RBS for *ybdM*; first gene of a putative operon with *ybdM* (hypothetical transcriptional regulator)	0.788778	`ATTAATATAAATTTAAT`	nd	/	1100	All nine genomes?
glgB	STM3538; Ortholog of *E. coli* 1,4-alpha-glucan branching enzyme (AAC76457.1); Blast hit to 1,4-alpha-glucan branching enzyme; Putative RBS for *glgX*; putative first gene of operon involved in glycogen synthesis	0.779808	`TTTAAGGGTAGCTTAAT`	m	-	1111	All nine genomes

leuO	*STM0115; S. typhimurium* probable activator protein in leuabcd operon. (SW:LEUO_SALTY) putative transcriptional regulator (LysR family)	0.776490	ATTAATGTTAACTTTTT	m	-	1111	All nine genomes
STM0343	Paralog of *E. coli* orf, hypothetical protein (AAC75237.1); Blast hit to AAC75237.1 identity in aa 10 - 512 putative Diguanylate cyclase*phosphodiesterase domain	0.77427	ATTAATGTTACTTTAGT	nd	/	1100	Subspecies 1 only
orf242	*STM1390 S. typhimurium orf242* (gi|4456866) putative regulatory proteins, *merR* family	0.773644	CTTAGTCTTCATTTGAT	/	/	1000	Other distributions
STM1868A / mig-3	lytic enzyme; intergenic shared with *mig-3* (phage assembly protein)	0.773462	CTTAATGATTATTTATT	/	/	1000	?
STM2763 / STM2726	Paralog of *E. coli* prophage CP4-57 integrase (AAC75670.1); Blast hit to putative integrase; intergenic shared with *STM2726* (putative inner membrane)	0.772053	ATTAATGTCCATTTAGT	/	/	1000	*S. typhimurium* only
pntA	*STM1479;* Ortholog of *E. coli* pyridine nucleotide transhydrogenase, alpha subunit (AAC74675.1); Blast hit to AAC74675.1 pyridine nucleotide transhydrogenase (proton pump), alpha subunit; Putative RBS for *pntB*; first gene of the putative operon (*pntA,pntB*)	0.770547	TTTAATGTTAATTTCTT	m	-	1111	All nine genomes
STM0057 / cit2	putative citrate-sodium symport; intergenic shared with *citC2* (citrate lyase synthetase)	0.767968	CTCATGGTTCATTGAAT	nd	/	1110	Other distributions
yrbF	*STM3313;* Ortholog of *E. coli* putative ATP-binding component of a transport system (AAC76227.1); Blast hit to AAC76227.1 putative ABC superfamily (atp_bind) transport protein; Putative RBS for *yrbE*; RegulonDB:STMS1H003330; first gene of putative *yrb* operon (ABC transporter)	0.766758	CCTAATTTTGACTTTAT	m	+	1111	All nine genomes
yejG	*STM2220;* Paralog of *E. coli orf*, hypothetical protein (AAC75242.1); Blast hit to putative cytoplasmic protein	0.767099	CTTTATGTTTATTTTAT	m	+	1111	All nine genomes
slsA	*STM3761;* putative inner membrane protein	0.765418	CTTTATGTTATTTAAAT	nd	/	1110	Other distributions
yhcN	*STM3361;* Ortholog of *E. coli orf*, hypothetical protein (AAC76270.1); Blast hit to putative outer membrane protein	0.764452	ATTAGTGTATACTTAAT	m	+	1111	All nine genomes?
yceP	*STM1161;* Ortholog of *E. coli orf*, hypothetical protein (AAC74144.1); Blast hit to putative cytoplasmic protein	0.764191	TTTATTGTTCATATAAT	m	+	1100	All nine genomes

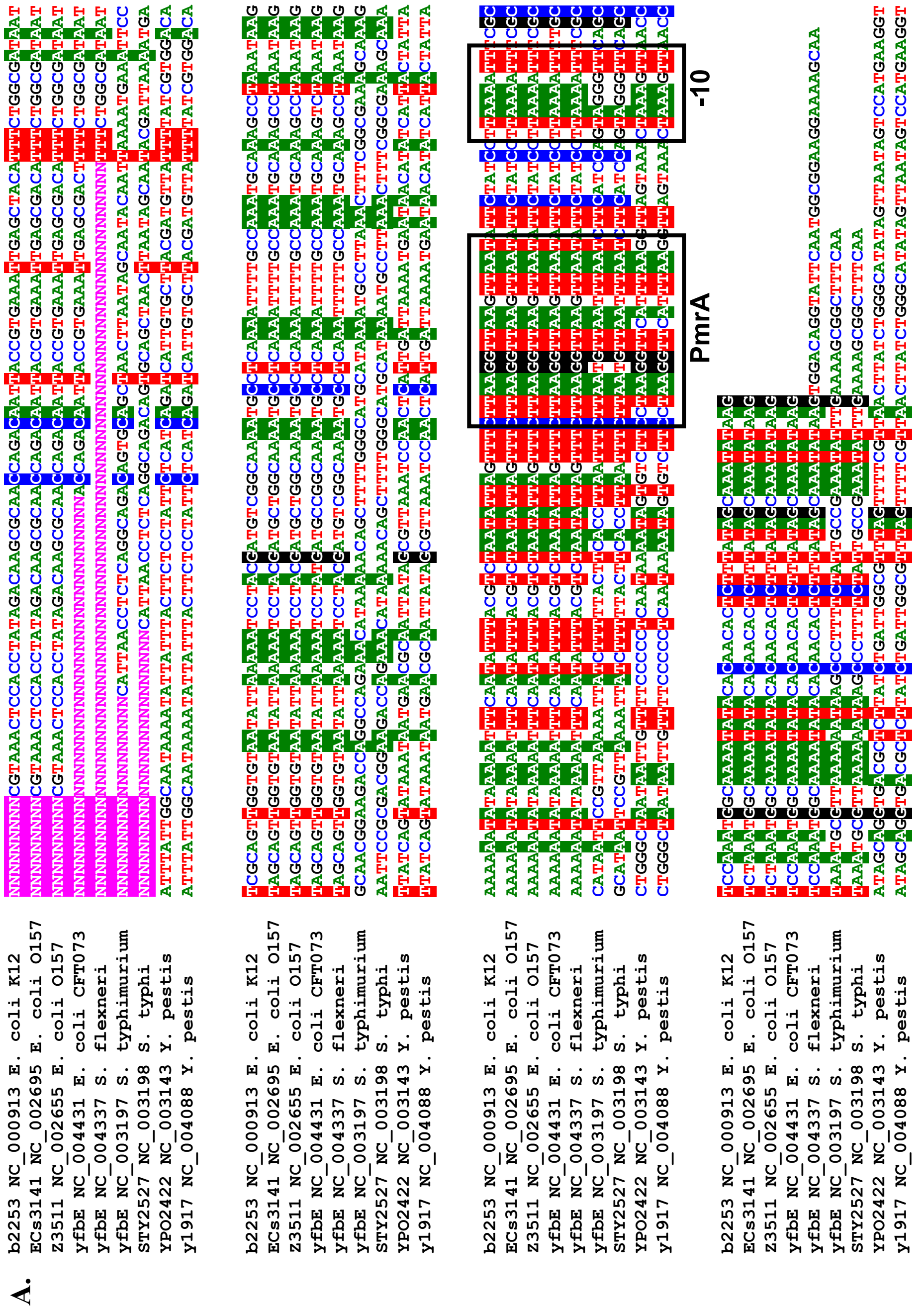
STM4098	putative arylsulfate sulfotransferase	0.763003	`TCTAATATTTATTTAAT`	nd	/	1100	Subspecies 1 only?
stfA	*STM0195; S. typhimurium* major fimbrial subunit StfA	0.762241	`ATCAATTTTAATTTAAT`	/	/	1000	Some *Salmonella* only
atpF	*STM3869;* Ortholog of *E. coli* membrane-bound ATP synthase, F0 sector, subunit b (AAC76759.1); Blast hit to imembrane-bound ATP synthase, F0 sector, subunit b; Putative RBS for *atpH*; first gene of a putative operon encoding putative ATP synthase	0.760841	`CAGAAGGTTAACTAGAT`	m	+	1111	All nine genomes
yegH / wza	*STM2119;* Ortholog of *E. coli* putative transport protein (AAC75124.1); Blast hit to putative inner membrane protein; intergenic shared with *wza* (putative polysaccharide export protein)	0.760004	`ATTAATATTAAATGAAT`	m	-	1111	All nine genomes
yjgD / argI	*STM4470; S. typhimurium* hypothetical protein in *argI-miaE* intergenic region (ORF15.6) putative cytoplasmic protein; Putative binding site for ArgR; shared intergenic regions with *argI* (arginine ornithine transferase); first gene of a putative operon with *miaE*	0.759514	`ATTAAAATTCACTTTAT`	m	+	1111	All nine genomes
sseJ / STM1630*	*STM1631; S. typhimurium* secreted effector; regulated by SPI-2; shared intergenic with *STM1630* (putative inner membrane protein)	0.758303	`CTTAAGAAATATTTAAT`	/	/	1000	Some *Salmonella* only
csrA	*STM2826; S. typhimurium* carbon storage regulator	0.756990	`CTTAGGTTTAACAGAAT`	m	+	1111	All nine genomes
dinP / yafK	*STM0313;* Ortholog of *E. coli* damage-inducible protein P; putative tRNA synthetase (AAC73335.1); Blast hit to AAC73335.1 DNA polymerase IV, devoid of proofreading, damage-inducible protein P; intergenic shared with *yafKJ* (periplasmic putative amido transferase)	0.756938	`CATACTGTACACTTAAA`	m	+	1111	All nine genomes
STM0346	putative outer membrane protein; Homolog of *ail* and *ompX*	0.756369	`CATTAGGTGCTCTTAAT`	/	/	1000	Some *Salmonella* only
ybfA / STM0707	*STM0708;* Ortholog of *E. coli orf*, hypothetical protein (AAC73793.1); Blast hit to putative periplasmic protein; intergenic shared with *STM0707* (hypothetical protein)	0.754265	`ATTAGTATTAATTTAAC`	m	+	1111	All nine genomes?
yncD / STM1587	*STM1587;* Ortholog of *E. coli* putative outer membrane receptor for iron transport (AAC74533.1); Blast hit to paral putative outer membrane receptor; intergenic shared with *STM1586* (putative receptor)	0.754063	`CATTTTCTTAACTTAAT`	m	-	1100	All nine genomes
yafC / STM0275	*STM0256;* Ortholog of *E. coli* putative transcriptional regulator LysR-type (AAC73313.1); Blast hit to putative transcriptional regulator, LysR family; intergenic shared *STM0275* (drug efflux protein)	0.753257	`CAAAATATCAATTTAAT`	m	-	1111	Other distributions

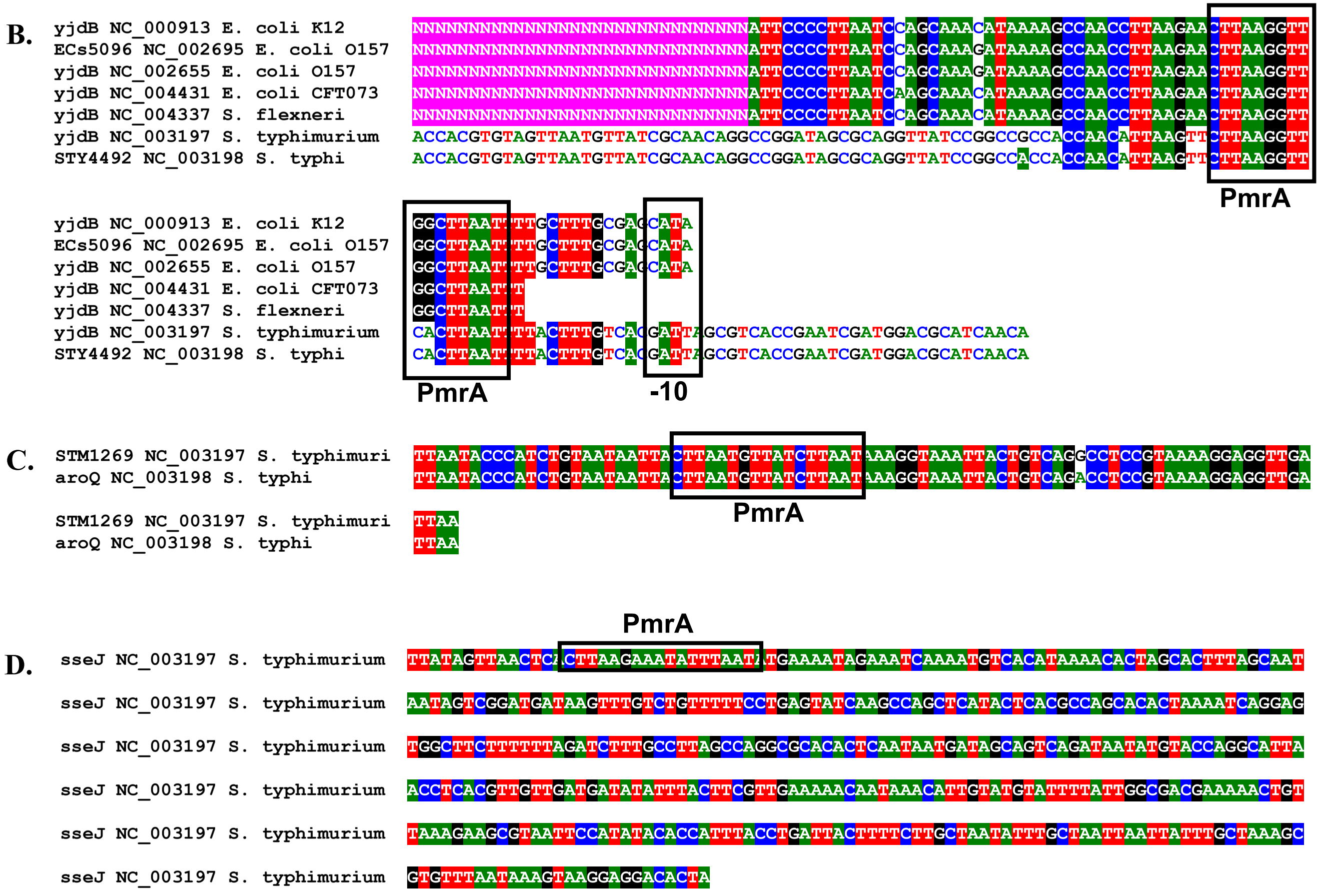
3.2.4 Detailed analysis of the putative PmrAB targets

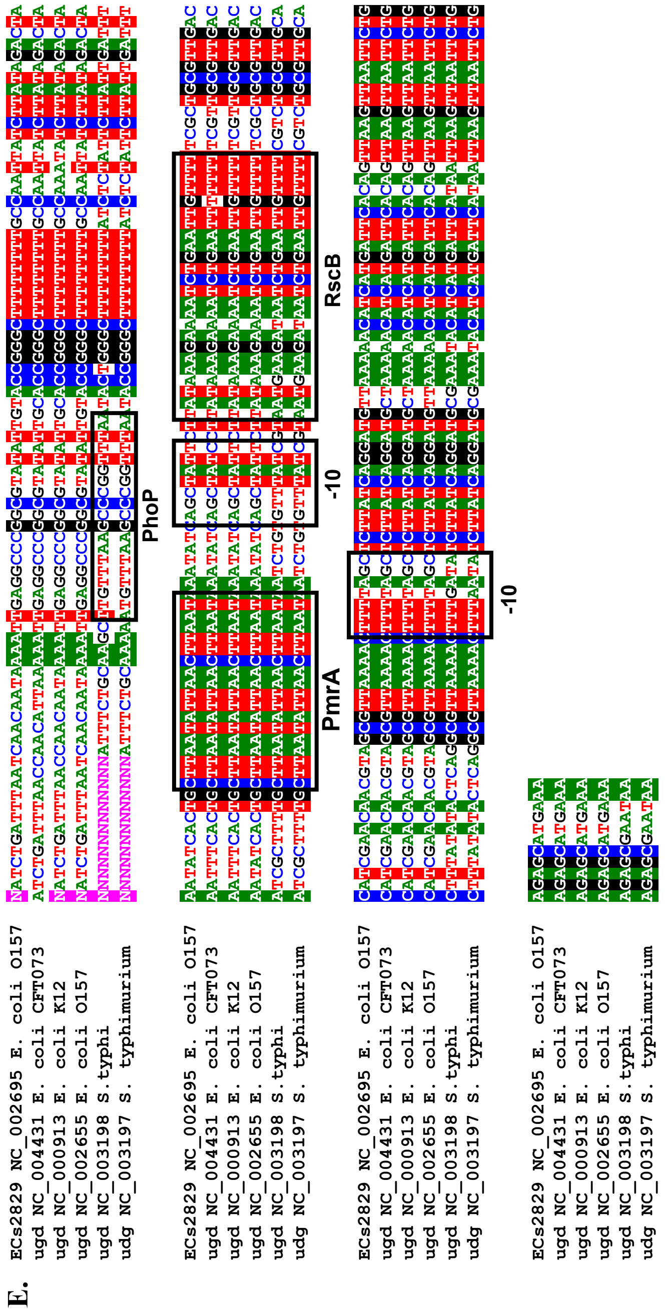
Putative PmrA motifs were detected in the intergenic regions of genes encoding transcriptional regulators, outer membrane and secreted proteins, proteins with functions involved in flagellae and fimbrae synthesis, proteins with a function related to the modification of cellular components, putative transport proteins, proteins involved in amino acid synthesis and also in phage remnants. As mentioned before, if the putative PmrAB-regulated genes contained close homologs in other species, the intergenic sequences of these close homologs were locally aligned to check whether putative PmrA motifs were conserved in these other species as well. For some of the datasets, however, no local alignment could be identified (no motif detected). Closer inspection showed that most of these datasets contained highly homologous paralogs of the original sequence. The intergenic sequences of these paralogs showed an overall low degree of conservation (e.g. *STM0057*) with the original intergenic sequence in *S. typhimurium* (data not shown). In some of the datasets, a local alignment of the respective intergenic regions could be detected, but the putative PmrA motif was not present within the conserved parts of the alignment (e.g. *leuO*). For these putative PmrAB targets, phylogenetic footprinting could not strengthen the confidence in the prediction of the PmrA motif. If such putative motifs are biologically active, their activity will be restricted to Salmonella serovars or *S. typhimurium*.

Our analysis revealed that PmrA motifs, present in the intergenic sequences of known PmrAB-dependent *S. typhimurium* genes, were also conserved in the intergenic sequences of the orthologs of these genes in related species (Fig. 3.2). An overview of the alignments of these known targets is given below.

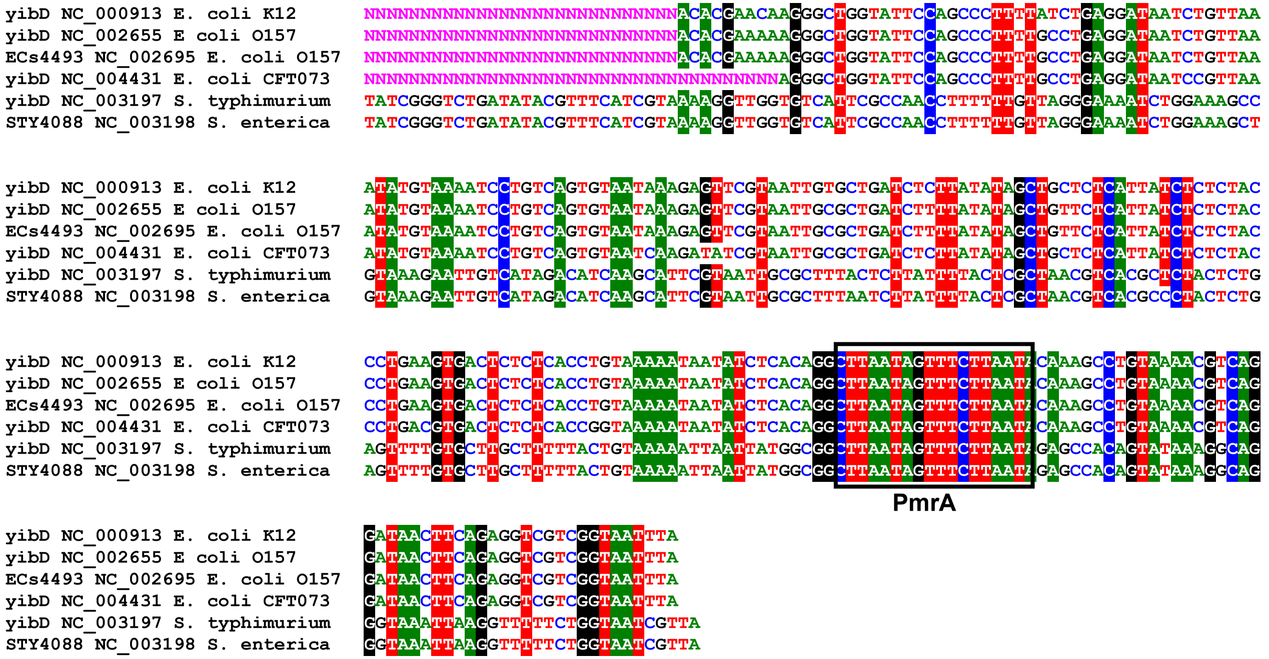
Figure 3.2: **Local alignments of the most promising targets.** Examples of local alignments obtained by phylogenetic footprinting of known PmrAB targets and of some promising potential targets. Known motifs or (putative) PmrA motifs are indicated by a box. A: *yfbE (pmrH),* B: *yjdB (pmrC)*, C: *STM1269 (aroQ)* D: *sseJ*, E: *ugd*, F: *yibD*, G: *ybjG (mig-13).*


A.
b2253 NC_000913 E. coli K12
ECs3141 NC_002695 E. coli O157
Z3511 NC_002655 E. coli O157
yfbE NC_004431 E. coli CFT073
yfbE NC_004337 S. flexneri
yfbE NC_003197 S. typhimurium
STY2527 NC_003198 S. typhi
YPO2422 NC_003143 Y. pestis
y1917 NC_004088 Y. pestis
PmrA
-10


B.
yjdB NC_000913 E. coli K12 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATTCCCCTTAATCCAGCAAACATAAAAGCCAACCTTAAGAACTTAAGGTT
ECs5096 NC_002695 E. coli O157 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATTCCCCTTAATCCAGCAAAGATAAAAGCCAACCTTAAGAACTTAAGGTT
yjdB NC_002655 E. coli O157 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATTCCCCTTAATCCAGCAAAGATAAAAGCCAACCTTAAGAACTTAAGGTT
yjdB NC_004431 E. coli CFT073 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATTCCCCTTAATCAAGCAAACATAAAAGCCAACCTTAAGAACTTAAGGTT
yjdB NC_004337 S. flexneri NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATTCCCCTTAATCCAGCAAACATAAAAGCCAACCTTAAGAACTTAAGGTT
yjdB NC_003197 S. typhimurium ACCACGTGTAGTTAATGTTATCGCAACAGGCCGGATAGCGCAGGTTATCCGGCCGCCACCAACATTAAGTTCTTAAGGTT
STY4492 NC_003198 S. typhi ACCACGTGTAGTTAATGTTATCGCAACAGGCCGGATAGCGCAGGTTATCCGGCCACCACCAACATTAAGTTCTTAAGGTT
PmrA
yjdB NC_000913 E. coli K12 GGCTTAATTTTGCTTTGCGAGCATA
ECs5096 NC_002695 E. coli O157 GGCTTAATTTTGCTTTGCGAGCATA
yjdB NC_002655 E. coli O157 GGCTTAATTTTGCTTTGCGAGCATA
yjdB NC_004431 E. coli CFT073 GGCTTAATTT
yjdB NC_004337 S. flexneri GGCTTAATTT
yjdB NC_003197 S. typhimurium CACTTAATTTTACTTTGTCACGATTAGCGTCACCGAATCGATGGACGCATCAACA
STY4492 NC_003198 S. typhi CACTTAATTTTACTTTGTCACGATTAGCGTCACCGAATCGATGGACGCATCAACA
PmrA
-10
C.
STM1269 NC_003197 S. typhimuri TTAATACCCATCTGTAATAATTACTTAATGTTATCTTAATAAAGGTAAATTACTGTCAGGCCTCCGTAAAAGGAGGTTGA
aroQ NC_003198 S. typhi TTAATACCCATCTGTAATAATTACTTAATGTTATCTTAATAAAGGTAAATTACTGTCAGACCTCCGTAAAAGGAGGTTGA
PmrA
STM1269 NC_003197 S. typhimuri TTAA
aroQ NC_003198 S. typhi TTAA
D.
PmrA
sseJ NC_003197 S. typhimurium TTATAGTTAACTCACTTAAGAAATATTTAATATGAAAATAGAAATCAAAATGTCACATAAAACACTAGCACTTTAGCAAT
sseJ NC_003197 S. typhimurium AATAGTCGGATGATAAGTTTGTCTGTTTTTCCTGAGTATCAAGCCAGCTCATACTCACGCCAGCACACTAAAATCAGGAG
sseJ NC_003197 S. typhimurium TGGCTTCTTTTTTAGATCTTTGCCTTAGCCAGGCGCACACTCAATAATGATAGCAGTCAGATAATATGTACCAGGCATTA
sseJ NC_003197 S. typhimurium ACCTCACGTTGTTGATGATATATTTACTTCGTTGAAAAACAATAAACATTGTATGTATTTTATTGGCGACGAAAAACTGT
sseJ NC_003197 S. typhimurium TAAAGAAGCGTAATTCCATATACACCATTTACCTGATTACTTTTCTTGCTAATATTTGCTAATTAATTATTTGCTAAAGC
sseJ NC_003197 S. typhimurium GTGTTTAATAAAGTAAGGAGGACACTA


E.
ECs2829 NC_002695 E. coli O157
ugd NC_004431 E. coli CFT073
ugd NC_000913 E. coli K12
ugd NC_002655 E. coli O157
ugd NC_003198 S.typhi
udg NC_003197 S. typhimurium
PhoP
PmrA
-10
RscB
-10

F.

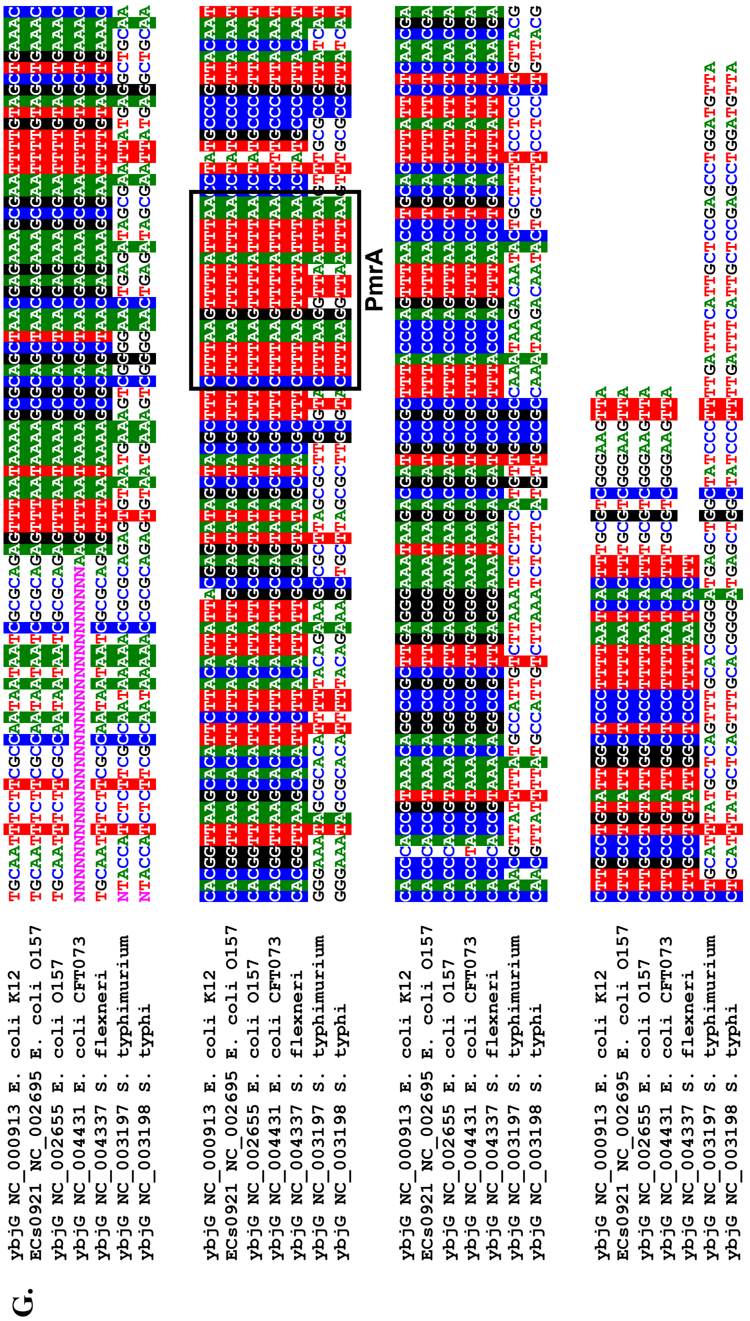


```
yibD NC_000913 E. coli K12          NNNNNNNNNNNNNNNNNNNNNNNNNNNNACACGAACAAGGGCTGGTATTCCAGCCCTTTTATCTGAGGATAATCTGTTAA
yibD NC_002655 E coli O157          NNNNNNNNNNNNNNNNNNNNNNNNNNNNACACGAAAAAGGGCTGGTATTCCAGCCCTTTTGCCTGAGGATAATCTGTTAA
ECs4493 NC_002695 E. coli O157      NNNNNNNNNNNNNNNNNNNNNNNNNNNNACACGAAAAAGGGCTGGTATTCCAGCCCTTTTGCCTGAGGATAATCTGTTAA
yibD NC_004431 E. coli CFT073       NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGGGCTGGTATTCCAGCCCTTTTGCCTGAGGATAATCCGTTAA
yibD NC_003197 S. typhimurium       TATCGGGTCTGATATACGTTTCATCGTAAAAGGTTGGTGTCATTCGCCAACCTTTTTTGTTAGGGAAAATCTGGAAAGCC
STY4088 NC_003198 S. enterica      TATCGGGTCTGATATACGTTTCATCGTAAAAGGTTGGTGTCATTCGCCAACCTTTTTTGTTAGGGAAAATCTGGAAAGCT

yibD NC_000913 E. coli K12          ATATGTAAAATCCTGTCAGTGTAATAAAGAGTTCGTAATTGTGCTGATCTCTTATATAGCTGCTCTCATTATCTCTCTAC
yibD NC_002655 E coli O157          ATATGTAAAATCCTGTCAGTGTAATAAAGAGTTCGTAATTGCGCTGATCTTTTATATAGCTGTTCTCATTATCTCTCTAC
ECs4493 NC_002695 E. coli O157      ATATGTAAAATCCTGTCAGTGTAATAAAGAGTTCGTAATTGCGCTGATCTTTTATATAGCTGTTCTCATTATCTCTCTAC
yibD NC_004431 E. coli CFT073       ATATGTAAAATCCTGTCAGTGTAATCAAGATATCGTAATTGCGCTGATCTCTTATATAGCTGCTCTCATTATCTCTCTAC
yibD NC_003197 S. typhimurium       GTAAAGAATTGTCATAGACATCAAGCATTCGTAATTGCGCTTTACTCTTATTTTACTCGCTAACGTCACGCTCTACTCTG
STY4088 NC_003198 S. enterica      GTAAAGAATTGTCATAGACATCAAGCATTCGTAATTGCGCTTTAATCTTATTTTACTCGCTAACGTCACGCCCTACTCTG

yibD NC_000913 E. coli K12          CCTGAAGTGACTCTCTCACCTGTAAAAATAATATCTCACAGGCTTAATAGTTTCTTAATACAAAGCCTGTAAAACGTCAG
yibD NC_002655 E coli O157          CCTGAAGTGACTCTCTCACCTGTAAAAATAATATCTCACAGGCTTAATAGTTTCTTAATACAAAGCCTGTAAAACGTCAG
ECs4493 NC_002695 E. coli O157      CCTGAAGTGACTCTCTCACCTGTAAAAATAATATCTCACAGGCTTAATAGTTTCTTAATACAAAGCCTGTAAAACGTCAG
yibD NC_004431 E. coli CFT073       CCTGACGTGACTCTCTCACCGGTAAAAATAATATCTCACAGGCTTAATAGTTTCTTAATACAAAGCCTGTAAAACGTCAG
yibD NC_003197 S. typhimurium       AGTTTTGTGCTTGCTTTTTACTGTAAAAATTAATTATGGCGGCTTAATAGTTTCTTAATAGAGCCACAGTATAAAGGCAG
STY4088 NC_003198 S. enterica      AGTTTTGTGCTTGCTTTTTACTGTAAAAATTAATTATGGCGGCTTAATAGTTTCTTAATAGAGCCACAGTATAAAGGCAG

yibD NC_000913 E. coli K12          GATAACTTCAGAGGTCGTCGGTAATTTA
yibD NC_002655 E coli O157          GATAACTTCAGAGGTCGTCGGTAATTTA
ECs4493 NC_002695 E. coli O157      GATAACTTCAGAGGTCGTCGGTAATTTA
yibD NC_004431 E. coli CFT073       GATAACTTCAGAGGTCGTCGGTAATTTA
yibD NC_003197 S. typhimurium       GGTAAATTAAGGTTTTTCTGGTAATCGTTA
STY4088 NC_003198 S. enterica      GGTAAATTAAGGTTTTTCTGGTAATCGTTA
```


G.
ybjG NC_000913 E. coli K12
ECs0921 NC_002695 E. coli O157
ybjG NC_002655 E. coli O157
ybjG NC_004431 E. coli CFT073
ybjG NC_004337 S. flexneri
ybjG NC_003197 S. typhimurium
ybjG NC_003198 S. typhi
PmrA

pmrH (first gene of an operon that contains the genes *pmrHFIJKLM*) (Table 3.1)) is the only known PmrAB-regulated gene for which the PmrA motif is conserved in all genome sequences analyzed (including *Y. pestis*). In *pmrC*, encoding a gene with unknown function [2,257], the PmrA motif is conserved in the intergenic regions of its orthologs in *E. coli* strains, *Salmonella* species and *Shigella*. *ugd* encodes a UDP-D-glucose dehydrogenase required for the synthesis of Ara4N. Three two-component systems are involved in its regulation (PmrAB, PhoPQ and RcsCB) [2,185] and this is reflected in the presence of the corresponding motifs: *ugd* contains a PmrA, PhoP and RcsB motif. The experimentally confirmed PmrA motif on the plus strand and part of the -10 sequence as determined by Aguirre *et al*., have been conserved in *S. typhimurium*, *S. typhi* and *E. coli* [2]. Remark that the promoter of *ugd* also has a hit of the PmrA motif on the minus strand. This was, however, not confirmed by DNA footprint analysis [2] and might represent a false positive. The PhoP motif on the plus strand in *ugd* of *Salmonella*, though occurring as a dyad, is not conserved in close orthologs and was recently demonstrated to be non functional [185]. The recognition site for the RcsB protein [185] is also conserved in *E. coli*. Lastly, *yibD* encodes a putative glycosyltransferase. The PmrA motif is conserved in *E. coli*. *yibD* has recently been identified as a PmrAB target by a genome-wide mutagenesis study. Its actual function is still unknown [257].

3.2.5 Experimental analysis

As our in silico predictions suggested some interesting potential targets of the PmrAB regulatory system, some of these targets were biologically validated in collaboration with the Centre of Microbial and Plant Genetics (CMPG, K.U. Leuven) to confirm the predictive power of our methodology (section 3.2.5.1). In addition, site-directed mutagenesis of the PmrA-binding site revealed specific sequence requirements (section 3.2.5.2). Details on the materials and methods used in this experimental analysis are not extensively described in this work. For this we refer to the article published in Genome Biology [160].

3.2.5.1 Experimental validation by expression analysis

Our in silico predictions pointed towards putative targets of the PmrAB regulatory system. Some of these have functionalities that were previously not associated with the PmrAB system. To further prove the strength of our in silico approach, four potential targets were selected for biological validation: *yibD* (novel at the time of our analysis), *aroQ* (*STM1269*), *mig-13* and *sseJ*. *aroQ* and *yibD* were selected because a perfect repeat of the previously described PmrA half site (CTTAAT [2]) was detected in their respective intergenic regions. *mig-13* (Fig. 3.2) was chosen because it has previously been reported as a gene selectively induced in macrophages, but with further unknown regulation. *sseJ* (Fig. 3.2) was further analyzed because although PmrAB regulated genes have been implicated in animal virulence [97], no direct link between SPI-2 (*Salmonella* pathogenicity island 2) gene regulation and PmrAB has been demonstrated yet.

Table 3.2: **Expression analysis of the GFP reporter fusions.** All experiments were performed twice. Values indicate the average mean peak fluorescence measurements of at least three samples for the populations grown under the conditions indicated for one representative experiment. Values between brackets represent the standard deviations. All values are expressed in arbitrary units. Strains used: WT = wildtype (ATCC14028s) and *pmrA*$^-$ = *pmrA* mutant.

Fusion	Strain	10 mM $MgCl_2$	10 µM $MgCl_2$	100 µM $FeCl_3$	10 mM $MgCl_2$ 100 µM $FeCl_3$	10 µM $MgCl_2$ 100 µM $FeCl_3$
pmrC::GFP	**WT**	6.06 (0.18)	16.8 (1.42)	70.53 (3.84)	27.39 (4.41)	83.2 (3.21)
	***pmrA*$^-$**	1.00 (0.01)	1.02 (0.02)	1.08 (0.03)	1.03 (0.03)	1.16 (0.12)
mig-13::GFP	**WT**	6.17 (1.55)	13.50 (2.02)	35.81 (4.67)	17.86 (5.04)	49.23 (5.43)
	***pmrA*$^-$**	2.69 (0.11)	4.32 (0.48)	5.2 (0.09)	2.67 (0.16)	9.64 (1.19)
aroQ::GFP	**WT**	2.32 (0.22)	20.39 (1.54)	19.39 (0.53)	4.38 (0.19)	19.48 (2.07)
	***pmrA*$^-$**	1.06 (0.02)	1.09 (0.02)	1.71 (0.09)	1.02 (0.01)	1.09 (0.03)
yibD::GFP	**WT**	1.25 (0.02)	1.67 (0.26)	33.35 (7.01)	27.52 (5.64)	52.46 (8.98)
	***pmrA*$^-$**	1.26 (0.02)	1.21 (0.06)	1.30 (0.02)	1.14 (0.02)	1.81 (0.44)
sseJ::GFP	**WT**	7.68 (1.55)	11.25(1.46)	22.58 (1.01)	3.80 (1.13)	8.03 (1.27)
	***pmrA*$^-$**	5.64 (0.72)	8.72 (1.05)	7.35 (1.55)	2.99 (0.43)	6.47 (1.36)

For each of these targets, GFP reporter fusions were constructed and expression was determined by fluorescence-activated cell sorter (FACS) analysis in a wild type and a *pmrA* mutant. Because the PmrAB system is sensitive to Mg^{2+} and Fe^{3+} concentrations, we tested the effect of these signals on the expression of the fusions [257] (Table 3.2). In all experiments, *pmrC* was used as a positive control. The *pmrC* fusion showed a clear induction by either Mg^{2+} deprivation or Fe^{3+} excess. The observed level of induction was higher for the Fe^{3+}-dependent signal than for the Mg^{2+}-dependent signal and the combination of both signals seemed to act synergistically. For both signals, induction was abrogated in a *pmrA* mutant, indicating that induction by Mg^{2+} and Fe^{3+} is solely PmrAB-dependent. For the *mig-13* fusion similar observations were made, although induction by low Mg^{2+} and the synergistic effect of both signals were less pronounced.

mig-13 also exhibited a considerable background expression level both in a *pmrA* mutant and in the uninduced state in a wild type background. *aroQ* was strongly induced by low Mg^{2+} and induction was abrogated in a *pmrA* mutant. The influence of Fe^{3+} was less pronounced. In case of *yibD*, the opposite was found: the *yibD* gene was barely induced by low Mg^{2+} but the Fe^{3+} excess resulted in a large induction. Although for the *yibD* fusion Fe^{3+} excess but not Mg^{2+} deprivation seemed to be a sufficient signal to trigger expression, both signals acted synergistically. Also induction of *yibD* was abrogated in a *pmrA* mutant. Compared to the other fusions, the observed expression levels of the *sseJ* fusion were rather low under the conditions tested. Results show an upregulation of *sseJ* expression in elevated Fe^{3+} concentrations that was absent in the *pmrA* mutant. As observed for *mig-13*, *sseJ* was expressed at a background level in the *pmrA* mutant. Interestingly, even at low concentrations, Mg^{2+} seemed to counteract the Fe^{3+}- dependent induction.

3.2.5.2 Site-directed mutagenesis of the PmrA box

We constructed a set of mutant PmrA box sequences by site-directed mutagenesis of the PmrA box of yibD. AT->GC and GC->AT substitutions were introduced in the first half site of the PmrA box (Figure 3.3A). We focused on the first half site since in the experimentally verified target *pmrC*, the second half site overlaps with the -35 promoter site [303]. Expression was compared between different mutagenized fusions and the non-mutated fusion in the wild type and in the *pmrA* mutant strain in all conditions mentioned before. For simplicity only the expression values for two inducing conditions are displayed in Figure 3.3B 1) the induction by the combined action of high Fe^{3+} and low Mg^{2+} concentrations and 2) the induction by elevated Fe^{3+} levels in the presence of a high Mg^{2+} concentration. Observations under all other conditions allowed to draw similar conclusions. Substitutions of the third and fifth positions of the motif box completely abrogated PmrAB-dependent expression. Mutations of the first, second, fourth or sixth position reduced PmrAB-dependent induction. Remark that for the mutation of the second position, the expression was very low but not completely abrogated. Results from this site-directed mutagenesis experiment of one representative PmrAB target allowed to unequivocally demonstrate that the PmrA box we identified was responsible for PmrAB-dependent transcriptional activation. It also allowed to further delineate the sequence requirements of the PmrA consensus.

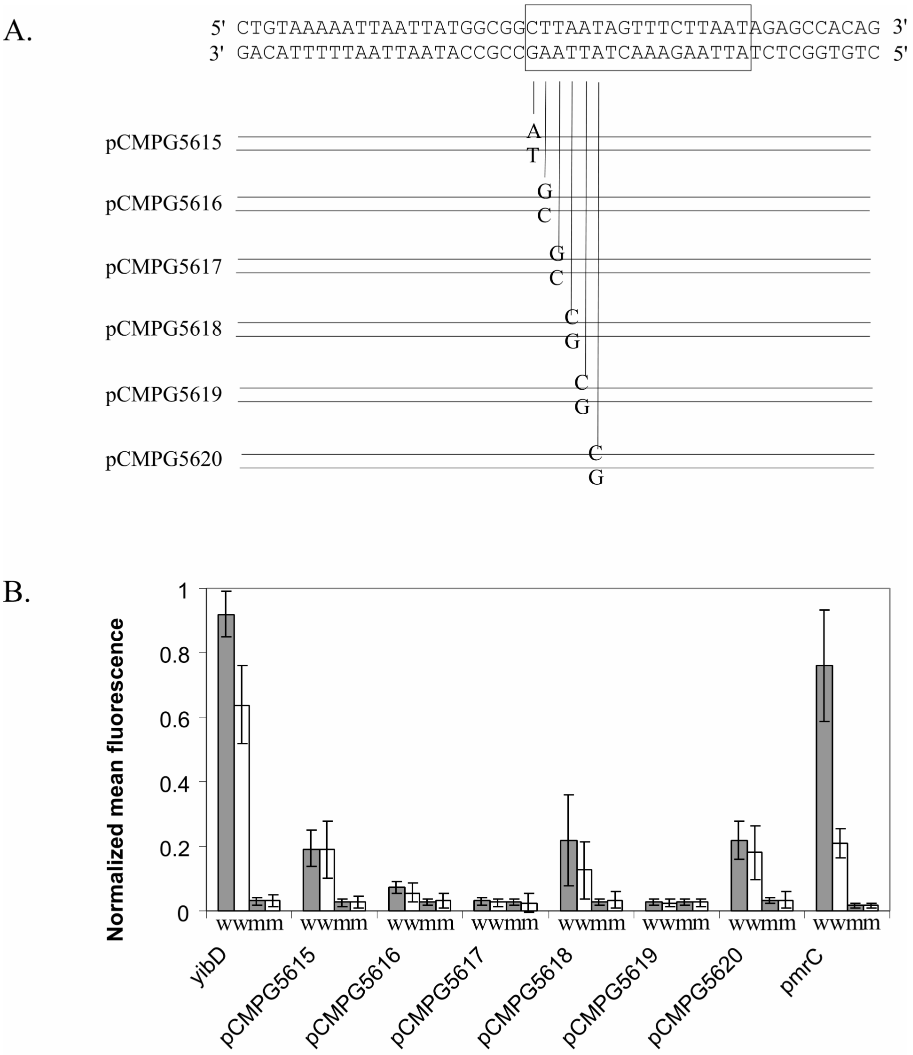


Figure 3.3:* Site-directed mutagenesis of the PmrA box in *yibD. Six species of the *yibD* promoter mutant designated pCMPG5615-pCMPG5620 each with a single base substitution (T->G or A->C) in the PmrA box were constructed as outlined in Panel A. Promoters were fused to GFP and promoter activity was assessed by FACS analysis. Panel B. The normalized expression values of the six mutant fusions and the wild type fusion measured in two distinct conditions in the wild type and *pmrA*::Tn*10*d mutant background are plotted. Grey bars represent condition 1 (100 μM FeCl$_3$ + 10 μM MgCl$_2$), white bars correspond to the expression values observed in condition 2 (100 μM FeCl$_3$ + 10 mM MgCl$_2$). w: wild type background; m: *pmrA* mutant background. The *pmrC::GFP* fusion was included as a positive control. Bars represent the standard deviations of 3 independent measurements.

3.2.6 Other promising PmrAB targets

Based on the instances of the PmrA motif in experimentally verified PmrAB targets of *Salmonella* (verified previously or validated in this study), a PmrA consensus was built (Fig. 3.4). The motif consensus of PmrA was converted into a regular expression (A/C)(C/T)T(A/T)A(T/G/A)N_5 NTT(A/T)A(T/A/G). To construct this regular expression we only considered the two conserved half sites because the PmrA motif is believed to be a dyad [2]. We preferred the intermediate part between the conserved half sites of the regular expression to be represented degenerated (i.e. N_5). Indeed, the observed degree of conservation in the intermediate part of the motif model (Fig. 3.4.B) is probably rather related to the restricted sample size of the training set than being an intrinsic property of the motif. Promising motifs (indicated in bold in Table 3.1), therefore, are motifs that match this regular expression and thus contain nucleotides that occur in the conserved half sites of one of the experimentally verified examples. Promising targets for which the putative PmrA motif was also conserved in species other than *Salmonella*, were *mig-13*, *yrbF*, *yjgD*, *ybdO*, *yejG*, *lasT* and *ybdN*. Promising targets, only present in *S. typhimurium* and/or *S. typhi* were *STM1269* (*aroQ*), *STM1273*, *sseJ* and *lpfA*. Remark that this listing is just based on an arbitrary selection criterium i.e. a preliminary PmrA motif consensus that will ameliorate as more PmrAB targets will become experimentally validated. Besides the targets mentioned above, Table 3.1 contains other targets that are of interest because their annotation relates to the PmrAB system (such as *yncD*).

A. Experimentally verified PmrA targets of *S. typhimurium*

ugd	(*S. typhimurium*)	CTTAAT	ATTAA	CTTAAT
pmrC	(*S. typhimurium*)	CTTAAG	GTTCA	CTTAAT
pmrC	(*E. coli*)	CTTAAG	GTTGG	CTTAAT
pmrH	(*S. typhimurium*)	CTTAAT	GTTAA	TTTAAT
pmrH	(*E. coli*)	CTTAAG	GTTAA	GTTAAT
pmrH	(*Y. pestis*)	CCTAAG	GTTCA	TTTAAG
pmrD	(*S. typhimurium*)	ATTAAT	GTTAG	GTTAAT
mig-13	(*S. typhimurium*)	CTTTAA	GGTTA	ATTTAA
mig-13	(*E. coli*)	CTTTAA	GTTTT	ATTTAA
STM1269	(*S. typhimurium*)	CTTAAT	GTTAT	CTTAAT
yibD	(*S. typhimurium*)	CTTAAT	AGTTT	CTTAAT
sseJ	(*S. typhimurium*)	CTTAAG	AAATA	TTTAAT

B. Adapted motif logo

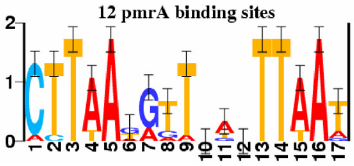


Figure 3.4: **Refined consensus of the PmrA box.** A. Alignment of all experimentally verified PmrA sites (previous [2,125] or this work) in *S. typhimurium*. PmrA sites in the orthologs of these respective experimentally verified genes are also displayed if these PmrA motif instances deviated from the PmrA motif in *S. typhimurium*. B. An adapted motif model of the PmrA site was built (represented by its logo) based on the sequences represented in A.

3.3 Discussion

Putative PmrAB targets were detected by genome-wide screening of *S. typhimurium* intergenic sequences using a PmrA motif model. If possible, the confidence in the predicted motifs was strengthened by a cross species comparison: we tested whether the PmrA motif was conserved in the intergenic regions of close homologs in related species. To this end, we developed a two-step procedure for phylogenetic footprinting. In the first step, a motif detection procedure based on Gibbs sampling was performed to generate a list of motifs. In the second step, these motifs were used as seeds to generate local multiple alignments. Eventually, the biological relevance of the obtained alignments was assessed. Several reasons urged us to use the alignments rather than a listing of the high scoring motifs obtained by Gibbs sampling. At first, we observed as also reported by McCue *et al.*, a high overall similarity in intergenic regions of the selected species [168]. In general the overall degree in conservation between the intergenic sequences of close homologs is about 93.56% for the sequenced representatives of *Escherichia* and *Shigella* species, 69.21% for *Shigella* and *Salmonella* and 53.31% for *Salmonella* and *Yersinia*. As a result of this property (high correlation in the data), not only the motif itself turns out to be conserved, but also its local neighbourhood. Moreover, the degree of conservation between the aligned sequences in a biologically relevant alignment will reflect, in most cases, the phylogenetic relatedness of the species from which the sequences are derived (see Fig. 3.2 for examples). By selecting the most promising alignment seeds (based on the appropriate heuristics for the scores) and constructing a local alignment with these seeds, we could also evaluate the local neighbourhood of the seed. If this one seemed to be conserved as well, we could be more confident in the obtained alignment and in the motifs contained within the conserved parts. Therefore, the use of local alignments allows a better judgment on the reliability of the motifs.

Secondly, Gibbs sampling is a stochastic procedure. The algorithm has to be run repeatedly on the same dataset, each time generating potentially different motifs. As a consequence, the output of a motif detection approach can be simultaneously redundant and non exhaustive: some statistically strong motifs are detected repeatedly in different runs. On the other hand, some motifs might never be detected. Indeed, because Gibbs sampling was originally designed for unrelated sequences and because of the high correlation in the data, the number of possible equally scoring motifs (local optima) might be so high that many runs have to be performed before all motifs have been covered. All these local optima coincide with motifs that, when used as seeds, will result in a similar alignment. The same alignment can thus be obtained by several motifs, but there is no guarantee that all possible motifs that result in the same alignment will be detected by Gibbs sampling. Therefore an alignment is a better summary of the degree of conservation between the intergenic regions than a listing of the highest scoring motifs.

Moreover, regulatory systems such as PmrAB might have acquired some very species-specific targets. For such highly specialized regulatory systems, motifs are likely to be present in the intergenic sequences of a selected subset of orthologs only. However, such motifs, because they occur in a restricted number of sequences of the dataset only, will not necessarily correspond to the highest scoring motifs. As such, they might be overlooked when selecting on high scoring motifs, for instance, by setting a threshold on the score. Once a reliable local alignment of a set of intergenic sequences is obtained, one can judge on the degree of confidence put in the prediction of the motif of interest by checking in which subset of species the motif is conserved, but also by taking into consideration other factors, such as the functional annotation of the putative target. The motifs that we select based on our heuristic will result in a biological relevant alignment that includes the maximal number of species. As such, our heuristic tries to overcome the fact that intrinsically Gibbs sampling is unable to cope with correlated data. Remark also that the motif of interest (PmrA motif) does not necessarily have to correspond to the motif used to produce the alignment.

We demonstrated that our in silico phylogenetic footprinting approach can be used to confirm targets detected by genome-wide screening. So far, it can only be used for species that show a high degree of conservation in their intergenic regions, similar to the one observed in this study. As more complete genomic data will become available, the approach might become extendable to other species.

As was also suggested previously [168], the high observed similarity in intergenic sequences might be due to the small phylogenetic distance between the species we analyzed. However, it can not be excluded that because of the small size of the intergenic regions in bacteria and the very similar habitat and mechanism of regulation among the γ-proteobacterial species used in this study, a large part of the complete intergenic region is functional and therefore conserved. This hypothesis was also posed by Rajewsky *et al.* [215]. The alignment of the intergenic region of the well-characterized *ugd* indeed points into that direction. Large parts of the conserved regions of the alignments correspond to experimentally verified motifs.

Remarkably, most potential PmrAB-regulated genes exhibited a footprint of the PmrA motif in *E. coli* only and several target genes had no counterpart at all in organisms other than *Salmonella* species. This indicates a high degree of specialization of the PmrAB two-component system in *Salmonella* species. Such high specialization could also explain the considerable differences between PmrAB-dependent regulons in related species. For instance, both in *Y. pestis* and *S. typhimurium* the attenuated virulence of the respective *phoP* mutants of both species is ascribed to a defect in LPS modification, a process shown to be PmrAB-dependent [215]. So far, two *S. typhimurium* loci have been postulated to be involved in this LPS remodelling: *pmrHFIJKLM* and *ugd*. Only for *pmrH* we detected an ortholog in *Y. pestis* and a conserved footprint of the PmrA motif in the promoter region of this ortholog. Ugd does not even have a functional counterpart in *Y. pestis*. This low similarity in PmrAB regulon composition indicates that a different network of genes must be responsible for a similar phenotype in distinct species. The latter is not completely unexpected in view of the very different LPS composition between *Salmonella* and *Y. pestis* [111].

For most of the known experimentally verified targets, clear phylogenetic footprints of the PmrA motif could be detected in the intergenic regions of close homologs. In the intergenic region of *pmrD*, we could recover the consensus sequence only partially (i.e. one half site) because the second half site overlaps with the coding region (data not shown) and this was not included in the current analysis. Another PhoPQ-dependent gene that contributes to resistance against antimicrobial peptides is *mig-14* [31]. However, we could not find the presence of a clear PmrA consensus in the promoter of *mig-14*. Also in *dgoA*, previously shown to be PmrAB-regulated [257], we could not detect a PmrA motif. This would indicate that both targets are only indirectly dependent on PmrAB. It can, however, not be excluded that they represent false negatives of our screening.

Besides the known targets, several putative new predictions could be made. Some of these predictions are consistent with previously published observations. Indeed, the PmrAB system is part of a complex regulatory cascade, acting downstream of the pleiotrophic PhoPQ system. The PhoPQ regulon is responsible for intracellular bacterial survival and genes dependent on PhoPQ are induced in macrophages. Part of the PhoPQ regulon has been discovered to be only indirectly PhoPQ-dependent via PmrAB. This PmrAB-dependent subset is known to confer resistance to cationic peptides by encoding genes involved in LPS modification and genes contributing to the resistance against elevated Fe^{3+}. Genes encoding proteins involved in modification of membrane components and outer membrane proteins therefore are sensible additional putative PmrAB targets. Another target worth mentioning in view of the Fe^{3+}-sensitivity of the PmrAB system is *yncD*, encoding a putative outer membrane receptor for iron transport. Phage remnants, such as *mig-3*, have been described as macrophage inducible, PhoPQ-dependent genes [280]and thus can be PmrAB-dependent. This might explain the PmrA motif in the intergenic region between *STM1868A* and *mig-3*. Detweiler *et al*. showed that 2 genes, *virK* and *somA*, both co-expressed with the SPI-2 system, confer resistance to cationic peptides and that their expression is PhoPQ-dependent. Also four fimbrial operons had genes that were co-expressed with SPI-2 [54]. Predictions on the PmrAB-dependency of *sseJ*, encoding a secreted effector of the SPI-2 system, or of genes, that are involved in fimbriae synthesis could therefore be in agreement with these findings (such as in *lpfA*, encoding the *S. typhimurium* long polar fimbria A precursor). Recently, the study of Kim *et al.* also related PmrAB-dependent regulation to swarming motility functions in *S. typhimurium* [130]. This could explain why we detected a putative PmrA motif in the intergenic region of *flhC* encoding a master flagellar transcriptional activator.

To further confirm the predictive power of our methodology, four putative PmrAB targets were subjected to biological validation. Expression analysis clearly demonstrated PmrAB-dependency of *yibD*, which confirms the recent observations of Tamayo *et al.* [257]. The observed PmrA dependency of *mig-13* is in accordance with it being upregulated in macrophages in in vivo conditions [280]. Striking is the clear PmrAB dependency of *aroQ*, encoding a periplasmic chorismate mutase. In general, chorismate mutases are involved in the synthesis of tyrosine and phenylalanine and they are the key to the synthesis of a plethora of secondary metabolites [58]. The function of periplasmic chorismate mutases which differ from the cytoplasmic chorismate mutases in their long C-terminal extension [37], is still unclear. Periplasmic AroQ proteins have also been detected in *Y. pestis*, *Pseudomonas aeruginosa*, *Mycobacterium* species, *Erwinia herbicola* and in the phytoparasitic nematode *Meloidogyne javanica* [37]. Strikingly, all these organisms containing AroQ interact with a eukaryotic host. This observation, taken together with the fact that AroQ is dependent on the key virulence regulator PmrAB in *S. typhimurium* suggests that the so far unknown function of AroQ might be involved in bacterial-host interactions.

Despite its low expression level in the in vitro conditions we used, the *sseJ* fusion showed a clear PmrAB-dependent induction by Fe^{3+}-excess. SseJ is a secreted effector protein that is translocated across the membrane of the Salmonella-containing vacuole (SCV) by SPI-2. From recent evidence it was speculated that the putative acyltransferase activity of SseJ would be involved in modifying the lipid composition of the SCV [76,225], thereby interfering with the trafficking and maturation properties of the SCV in the infected cells. The PmrA dependency of *sseJ* would therefore link expression of genes involved in bacterial LPS modification with those involved in regulation of the lipid composition of the SCV membrane.

Further experimental analysis will shed light on how these previously undescribed PmrAB-dependent proteins, with so far unknown function, relate to the known part of the PmrAB-dependent regulon.

Interestingly, the extent to which each of the tested strains reacted to the signal Fe^{3+} or/and Mg^{2+} varied considerably. This is not surprising in view of the complex regulatory system that integrates both signals. Indeed, both signals are transduced via the PhoPQ, PmrD, PmrAB multicomponent system that includes a posttranslational signal transduction and a transcriptional feedback loop [125]. Depending on the affinities between the interacting components of such dynamic systems, small changes in initial concentrations of the components might result in large differences of the observed expression levels [237]. A more detailed study of the dynamics of this system might unveil how such system can integrate signals so differently.

Site-directed mutagenesis of the PmrA box in the *yibD* promoter indicated a crucial role for the T at position 3 and the A at position 5 of the first half site of the motif. As can be derived from the consensus site in Fig. 3.4, indeed no degeneracy is allowed at positions 3 and 5. This observation allows us to extrapolate to a certain extent the sequence requirements of the PmrA box in *yibD* to other PmrAB targets. Some positions seem essential while, seemingly the specific choice of the nucleotide at the other positions affects the level of induction. By altering the nucleotides, binding affinities of the regulatory protein to the box can be modified allowing specific fine-tuning of gene expression in a cell.

3.4 Conclusion

Conclusively, we could demonstrate that our in silico approach allowed the reliable inference of additional PmrAB-dependent targets. Although false positives will still be present among the predicted PmrAB targets, it offers an interesting guideline for further elucidation of genetic networks involved in *S. typhimurium* virulence gene expression. As such, we could predict PmrAB-dependent regulation of four additional targets, that is *yibD*, *aroQ*, *mig-13* and *sseJ*. Our approach might become extendable to other species when more genome sequences will become available.

3.5 Materials and methods

3.5.1 Selection of intergenic sequences

Genome sequences were obtained from GenBank (ftp://ncbi/nlm/nih/gov/genbank/genomes/Bacteria/). All intergenic regions used in this study were extracted using the modules of INCLUSive [269] to automatically parse GenBank entries [21]. Here, we define an intergenic sequence as a region that contains the non-coding sequence between two coding regions. No overlap with coding regions is allowed. Intergenic regions with lengths lower than 10 bp were discarded because of computational reasons.

3.5.2 Construction of motif models

A motif model (probabilistic representation of the consensus DNA pattern that is recognized by the respective regulatory protein) for PmrA was constructed using MotifSampler (release 3.0) [268]. The PmrA training set consisted of the promoter regions of three known PmrAB-regulated genes (*ugd* [2,95,246], *pmrH* [95,246,275,303] and *pmrC* [97,303]) for which the binding of the PmrA protein to the promoter regions was verified by DNA footprinting analysis [2,303].

3.5.3 Genome-wide screening

The intergenic regions of the complete genome of *S. typhimurium* LT2 (NC_003197 [166]) were screened using MotifLocator (version 3.0) [160]. The scoring scheme of MotifLocator uses an extension of the classical position-weight matrix scoring scheme [234]. Given the motif model θ and the background model B_m a score $W(x)$ is computed for each window x of length l in the sequence S. $W(x)$ compares the score of the subsequence within the window being generated by the motif model to the score of the subsequence within the window being generated by the background model. b_j: nucleotide at position j in the segment.

$$W(x) = \log(\frac{P(x|\theta)}{P(x|S,B_m)} = \sum_{j=1}^{l}[\log(\theta_j^{b_j}) - \log(P(b_j|S,B_m))]$$

Both the plus and minus strand were screened using a background model of order 3. The higher order background allows implicit compensation for motifs that are located in a context, highly resembling the global nucleotide composition of the genome. In order to apply a threshold on the scores, scores of different motifs were normalized as such that their values range between 0 and 1. The normalized scores $\overline{W}(x)$ are displayed in Table 3.1.

$$\overline{W}(x) = \frac{W(x) - W_{\min}}{W_{\max} - W_{\min}} \text{ with } W_{\min} = \min_x W(x) \text{ and } W_{\max} = \max_x W(x).$$

Hits with a score above 0.75 were retained (corresponding to a selection of the 0.003 % top scoring hits of the total number of possible motif positions in the genome. Possible positions are identified as overlapping windows of length l). To give a rough assessment on the number of hits with a score similar to the chosen threshold that could be expected based on the specific nucleotide composition of the genome, we generated 100 random sets of intergenic sequences using a 3rd order background model. These random sets were scored with the same PmrA motif model. From these results it appeared that the true set contained 3 times more hits with a score above the threshold than an average random genome.

3.5.4 Identification of datasets

Highly similar homologs of the putative PmrAB-regulated genes were identified in the genome sequences of *S. typhimurium* (NC_003197), *S. typhi* (NC_003198), *S. flexneri* (NC_004337), *E. coli* O157:H7 (NC_002695), *E. coli* O157:H7 EDL933 (NC_002655), *E. coli* K12 (NC_000913), *Y. pestis* CO92 (NC_003143) and *Y. pestis* KIM (NC_004088). In general, only true orthologs are likely to have retained a similar function and therefore a similar mechanism of regulation [75] between true orthologs and paralogs is not always straightforward. Because our motif detection algorithm is, to some extent, robust against the presence of noise and allows for the presence of sequences that do not contain the motif [161], we did not make an a priori distinction between true orthologs and paralogs, if both appeared highly similar to the original protein. This motivated us to use the principle of clusters of orthologs for dataset construction [261]. The pairwise blast scores obtained by mutually aligning the whole genome sequences using BlastP [5] (release 2.1.2) were used as input of the cluster program TRIBE-MCL [67] (version 02.277). Stringent criteria were applied to only retain closely related orthologs and paralogs (cut-off of the blast hit (E-value of $1e^{-80}$)). For those proteins which, when blasted against themselves, gave rise to an E-value higher than $1e^{-80}$ (*yjbE*, *STM1926*, *STM0344*, *yhcN*, *STM1868A*, *yceP*, *atpF*, *yjgD*, *csrA*, *ybfA*) the threshold was relaxed (E-value $1e^{-20}$). The choice of the stringent threshold was essential to maximally reduce the noise in the datasets.

3.5.5 Phylogenetic footprinting by Gibbs sampling

We used a two-step procedure for phylogenetic footprinting. In the first step Gibbs sampling is performed to generate a list of motifs. Subsequently, local alignments are generated by 1) selecting motifs that can be used as alignment seeds, 2) assessing the relevance of the alignments by a test statistic.

3.5.5.1 Motif detection by Gibbs sampling

An advanced Gibbs sampling procedure for motif detection was utilized (MotifSampler). MotifSampler 3.1 allows searching for overrepresented motifs in each dataset. The motif length, the maximal number of different motifs and the background model are user-defined parameter settings of the algorithm described in Thijs *et al.* [268]. A recent extension of the algorithm allows to automatically determine the number of instances of a certain motif per sequence and requires a predefined indication on the prior probability of expecting at least one motif per sequence. For each dataset, 100 runs of the MotifSampler were performed under the following conditions: motif lengths varying from 6, 8, 10, 12, background order 0, prior probability default value 0.7. Because Gibbs sampling is a stochastic procedure, each run can give rise to different motifs. To summarize the results from the 100 runs, all detected motifs were mutually compared and similar motifs were grouped. The information content was used as similarity measure to compare motifs. Therefore, for each dataset a list of different potential motifs was obtained. Motifs in this list were ranked according to their log-likelihood score (LL-score) [268].

3.5.5.2 Generating reliable local alignments using the detected motifs

From the obtained list, motifs that could be used as seeds to generate a biologically relevant alignment were selected using a heuristic procedure. Motifs with a high LL-score that occurred preferentially once in each sequence were chosen (starting with those motifs that had the highest consensus score). Moreover, we preferentially selected motifs that occurred in the maximal number of species. For each dataset, multiple alignments were constructed using the position of these retrieved motifs as alignment initializations (seeds) until a reliable alignment was obtained. The alignment was considered biologically relevant if within a window of 100 bp around the motif, it exhibited a degree of conservation that reflected the overall observed homology between intergenic sequences of the selected species (interspecies homology). To assign a more quantitative criterium to the alignment, a p-value was assigned to each alignment that was calculated as follows: for each window of 100 bp around the motif, the largest conserved block not overlapping with the core motif was identified (using the consensus score of minimal 0.7 as minimal similarity measure). This p-value expresses the probability of observing a conserved block of the same length in a randomly aligned dataset of similar composition. Distributions of conserved blocks in randomly aligned sequences were constructed. These random datasets take into account the observed high pairwise sequence homology between intergenic sequences derived from similar species (serovars) (homology between *E. coli* sequences, homology between *Salmonella* sequences, homology between *Y. pestis* sequences), but not the interspecies homology between intergenic regions (e.g. between *E. coli* and *Salmonella*). All alignments with a p-value < 0.15 (p-value of *ugd*) were considered as relevant (indicated in Table 3.1 with “m”). Because the obtained alignments are local, they are gapless. In some cases, more than one alignment might be essential to cover the complete intergenic region.

3.5.6 Functional annotation

Functional annotation was derived from NCBI [21,166] and from the Sanger annotation of *S. typhi*. Specific genomic context was derived from NCBI [21,166]. The distribution of the putative targets (unique for *Salmonella* species versus more widely distributed), as derived from our COG's, was verified by comparison to the analysis of McClelland *et al.* [166], who included in addition to the species we used, several subspecies of *Salmonella* (6 genomes), and species more distantly related to *Salmonella* (*K. pneumoniae*).

3.5.7 Nomenclature

As the gene names used in the annotation of the *S. typhimurium* genome sequence do not always match the "common" names used in the PmrAB literature, we give a summary of the synonyms below. *STM1269* (*aroQ*); *ybjG* (*mig-13*); *pmrAB* (*basSR*); *ugd* (*udg*, *pagA*, *pmrE*); *pmrHFIJKLM* (*yfbE*, *pmrF*, *yfbG*, *STM2300*, *pqa*, *STM2302*, *STM2303*); *pmrC* (*yjdB*); *pmrG* (*ais*); *pbgP* (*yfbE*, *pmrH*).

Chapter 4

Comparative analysis of the PhoPQ regulon

4.1 Introduction

In this chapter we will develop a bioinformatics approach to reconstruct the PhoPQ regulon in *Salmonella typhimurium* and *Escherichia coli*. Where the identification of the PmrAB regulatory system was purely based on sequence data, the uncertainty about the motif requirements of the PhoPQ motif model forced us to use a combination of sequence and microarray data for the elucidation of the PhoPQ regulon. As this regulatory system is conserved in both *E. coli* and *S. typhimurium*, but regulates some important virulence factors in *S. typhimurium*, we compared the composition of the PhoPQ regulons between the pathogenic *S. typhimurium* and the non-pathogenic *E. coli* strain. This analysis was done in collaboration with the University of Washington (Dr. W. Navarre, Prof. F. Fang) and the Sidney Kimmel Cancer centre (San Diego; Prof. M. McClelland) who provided us with the *S. typhimurium* microarray data. This chapter has been published in the Journal of Molecular Evolution.

The PhoPQ regulatory system governs the adaptation to Mg^{2+} limiting conditions in *Salmonella typhimurium* and *Escherichia coli*. In this two-component system, the PhoQ protein acts as a sensor for extra-cytoplasmatic Mg^{2+} and Ca^{2+} [290] that controls the activity of the response regulator PhoP. This two-component system has also been shown to determine the virulence characteristics of *S. typhimurium* and other Gram-negative species [88]. In *S. typhimurium*, the sensing of extracellular Mg^{2+} permits the pathogen to determine its subcellular location i.e. inside macrophages and to activate the virulence factors essential for survival. Mutations in the *phoPQ* operon of *S. typhimurium* result in an attenuated virulence phenotype i.e. an increased sensitivity to cationic antimicrobial peptides [13,71,91,93,95,99], a decreased resistance to bile salts [211,288]. The amino acid sequences of PhoP and PhoQ of *E. coli* are 93% and 86% identical respectively to those of *S. typhimurium* [123]. This implies that the

PhoPQ systems in *E. coli* and *S. typhimurium* are most probably functional counterparts of each other. A short consensus site (T/G)GTTTA, occurring as an interrupted dyad, has been suggested to be the binding site for the PhoP regulatory protein in *Salmonella* spp. [248] and *Escherichia coli* [126]. The biological relevance of this motif was experimentally verified by Yamamoto *et al.* [311] and Lejona *et al.* [141] in *E. coli* and *S. typhimurium* respectively.

In this study, we estimated the size of the direct PhoPQ dependent regulons in both species by combining the evidence gained from microarray data and motif information. Comparing the gene composition of the PhoPQ regulon revealed a very small overlap in the genes that were PhoPQ regulated in both species, indicating that the difference in virulence phenotype between the pathogenic and the non-pathogenic bacteria might be attributed to a group of target genes that are specifically PhoP regulated in the pathogenic *S. typhimurium* strain but not in *E. coli* K12.

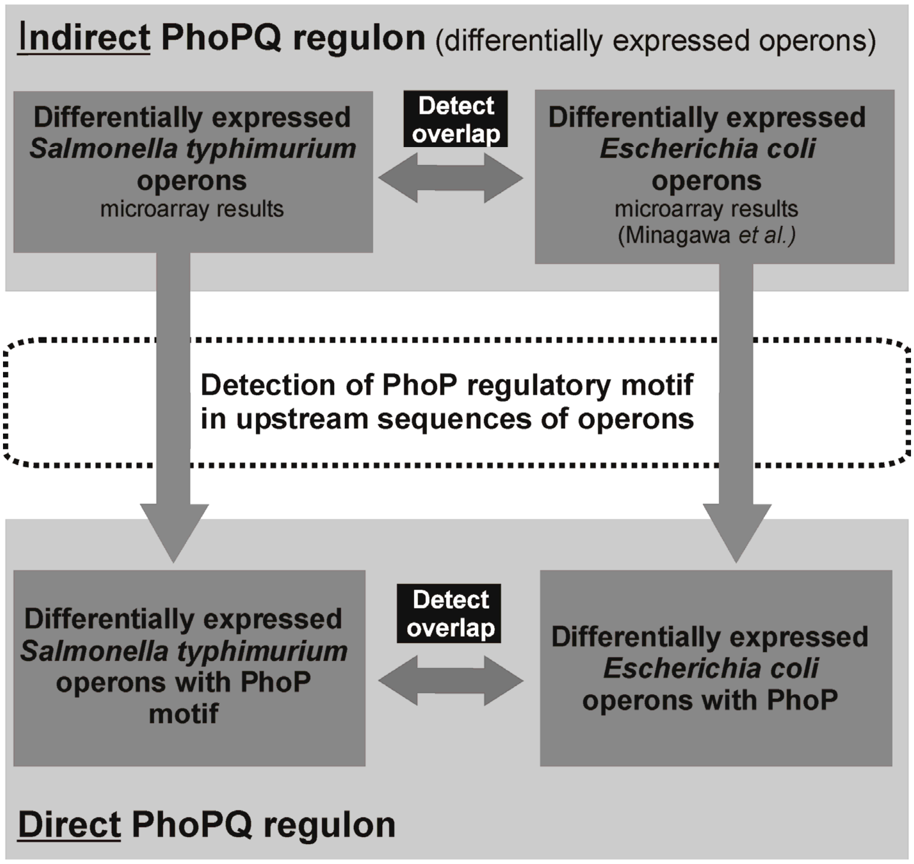

Indirect PhoPQ regulon (differentially expressed operons)
Differentially expressed Salmonella typhimurium operons microarray results
Detect overlap
Differentially expressed Escherichia coli operons microarray results (Minagawa et al.)
Detection of PhoP regulatory motif in upstream sequences of operons
Differentially expressed Salmonella typhimurium operons with PhoP motif
Detect overlap
Differentially expressed Escherichia coli operons with PhoP
Direct PhoPQ regulon

Figure 4.1: **Overview of the methodology.**

4.2 Results

4.2.1 Identification of the PhoP dependent regulon based on microarray analysis

To identify genes dependent on PhoPQ in *E. coli*, we used the results described by Minagawa *et al.* [175]. They identified 219 genes as being upregulated by PhoP.

To identify the regulon regulated by PhoPQ in *S. typhimurium* we compared expression in a knock out and constitutive mutant of the genes encoding the PhoPQ system. Based on the microarray analysis, the expression of approximately 2855 genes was found affected to some extent by the constitutive mutation. Based on selection criteria outlined in the Materials and Methods section, we selected from the 324 most significantly differentially expressed genes, the 214 most upregulated genes. This number approximated the number of *E. coli* PhoPQ upregulated genes selected by Minagawa *et al.* [175]. Conclusively, for both datasets we obtained a subset of genes most severely upregulated by PhoPQ. The quality of the microarray results and the biological relevance of the used cut-off values were confirmed by the presence of experimentally verified PhoPQ regulated genes in the selected subset of genes.

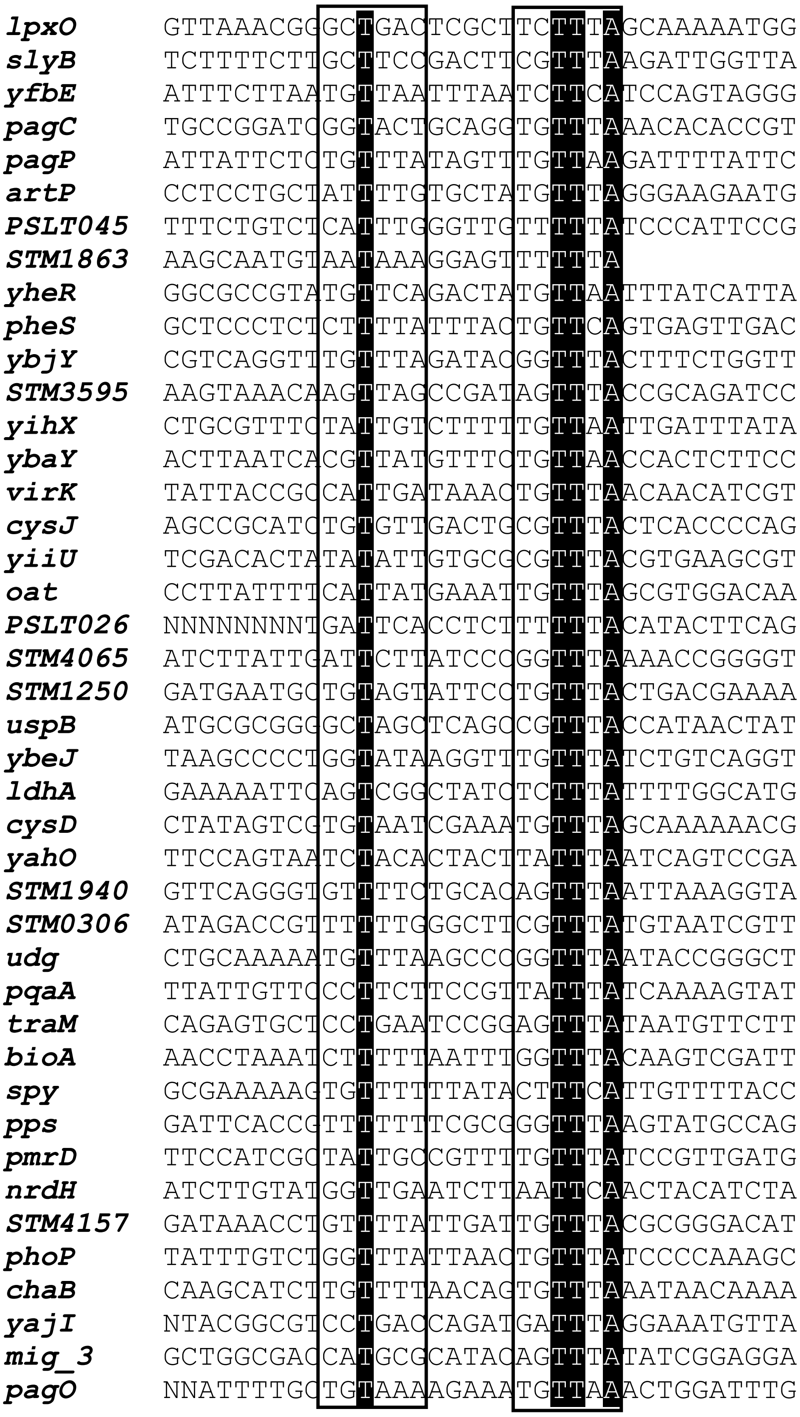
Based on our analysis, the 219 genes differentially PhoP regulated in *E. coli* were organised in 193 operons (on a total of 2848 predicted operons of which 2027 are singletons). Similarly, the 214 most affected *S. typhimurium* genes were located in 189 operons (on a total of 3026 predicted operons of which 2195 are singletons).

4.2.2 Motif screening

To identify target operons directly regulated by the PhoPQ regulatory system, intergenic regions of the first genes of the operons that were identified as differentially expressed in either one or both organisms, were screened with the PhoP motif model. Initially the PhoP box was postulated to be a direct repeat of the hexanucleotide (T/G)GTTTA separated by a spacer of 5 nucleotides. Recent experimental evidence allowed refinement of the specific sequence requirements of the PhoP box [141,175,311]. These studies confirmed that the PhoP regulatory protein is able to bind to a promoter region that does not display an intact dyad motif. A conserved thymine in the first half site (at position 3) and two conserved thymines together with one conserved adenine in the second half site (at positions 3, 4 and 6 respectively) were sufficient for detection of the DNA binding of the PhoP protein in *in vitro* DNA footprinting analysis.

slyB	AATAATCATCATGAATGTTTTGTTTATAATTGGTTG
acrA	GCAGCAATGGGTTTATTAACTTTTGACCATTGACCA
ybaD	TGCGCCTTTGTTGTATCGTCAGTTCAGGGTAAAATA
malS	AAATCTGAAACTATGTCACGTGTTAACGATTCAGAT
yeaD	AGCGACTTCGGTCGCTCTTTTTTTTACCTGATAAAA
napF	AAATGGCTTATTAATTATGCGGTTTATTTGGTCGCT
yhcL	CATCGTTGTTTTCAATCTGCCGTTTATGGGATTGAC
cchB	GTCACGCTCTCTCCTTTTTCATTTTACCTTCTGCGG
nlp	TTCACACTCTTTACAGGAACTTTTTAGAGCAATAGG
ygfF	TCTAAAGGCGCTTCGGCGCCTTTTTAGTCAGATGAC
ybdK	AAGGGCATATTTCGTCAGCATGTTTATATTGCCTCC
b1012	TTATGTGCAACTGTTTTGACCGTTTAGTCCACTTTT
ycgK	AGCTGTAGATTTCTACGGTTTGTTGAGTCCATGCCC
yrbL	AAGAGGCATTGTTTAGGTTTTGTTTAAGTTAATCGA
yjcQ	GCTGGCGTTCTTTCATGAAGAGTTAAGCGAGCGGCG
yjaD	AGTGAGAAATGTAAAAACCATGTTAAACATGCCAGT
purH	AAACTTCGTAATGAATTACGTGTTCACTCTTGAGAC
ppdA	GCGTCGGTTTTTTTACCCTCCGTTAAATTCTTCGAG
argD	TGTGGTTATAATTTCACATTTGTTTATGCGTAACAG
sucC	ATACGAAATATTCGGTCTACGGTTTAAAAGATAACG
fadL	CCGGAAAGTGCTGCTCCAGTTGTTAATTCTGCAAAA
yaaF	GCCGGTCTTGTTACCGGCATTTTTTATGGAGAAAAC
ybcU	AATATGAAATTTCAACTCATTGTTTAGGGTTTGTTT
metB	ATTGACGTCCATTAACACAATGTTTACTCTGGTGCC
rstA	ATGAAAACTTGTTTAGAAACGATTGATAGTAAGTAA
ybeQ	ATTAGCCGCATTGCGAGGCTCGTTTATTTATACTTT
ais	TTGCGCTTGTCTATAGGTGGAGTTTACG
yijF	CTCCCACGTTGTTCAGGAATTTTTTATCCGCTTCTG
lar	GCACCGCCAACTTGAAATATTTTTTATGAGAAAAAT
yiiE	CTACATCACTGTTCAGGAGCATTTAAATATTATTTG
phoP	CTCCCCGCTGGTTTATTTAATGTTTACCCCCATAAC
ytfR	ATCATGCCTACTGGGAGCACGCTTTACACCGGGGGA
uvrB	TTGCTCATGATTGACAGCGGAGTTTACGCTGTATCA
malM	GAAAGCCCCTCTGATTATCGGGTTTAGCGCGCTATT

Figure 4.2.A.: **Alignment of the detected PhoP motifs in *E. coli*.** Conserved positions are highlighted.



```
lpxO     GTTAAACGGGCTGACTCGCTTCTTTAGCAAAAATGG
slyB     TCTTTTCTTGCTTCCGACTTCGTTTAAGATTGGTTA
yfbE     ATTTCTTAATGTTAATTTAATCTTCATCCAGTAGGG
pagC     TGCCGGATCGGTACTGCAGGTGTTTAAACACACCGT
pagP     ATTATTCTCTGTTTATAGTTTGTTAAGATTTTATTC
artP     CCTCCTGCTATTTTGTGCTATGTTTAGGGAAGAATG
PSLT045  TTTCTGTCTCATTTGGGTTGTTTTTATCCCATTCCG
STM1863  AAGCAATGTAATAAAGGAGTTTTTTA
yheR     GGCGCCGTATGTTCAGACTATGTTAATTTATCATTA
pheS     GCTCCCTCTCTTTTATTTACTGTTCAGTGAGTTGAC
ybjY     CGTCAGGTTTGTTTAGATACGGTTTACTTTCTGGTT
STM3595  AAGTAAACAAGTTAGCCGATAGTTTACCGCAGATCC
yihX     CTGCGTTTCTATTGTCTTTTTGTTAATTGATTTATA
ybaY     ACTTAATCACGTTATGTTTCTGTTAACCACTCTTCC
virK     TATTACCGCCATTGATAAACTGTTTAACAACATCGT
cysJ     AGCCGCATCTGTGTTGACTGCGTTTACTCACCCCAG
yiiU     TCGACACTATATATTGTGCGCGTTTACGTGAAGCGT
oat      CCTTATTTTCATTATGAAATTGTTTAGCGTGGACAA
PSLT026  NNNNNNNNTGATTCACCTCTTTTTTACATACTTCAG
STM4065  ATCTTATTGATTCTTATCCCGGTTTAAAACCGGGGT
STM1250  GATGAATGCTGTAGTATTCCTGTTTACTGACGAAAA
uspB     ATGCGCGGGGCTAGCTCAGCCGTTTACCATAACTAT
ybeJ     TAAGCCCCTGGTATAAGGTTTGTTTATCTGTCAGGT
ldhA     GAAAAATTCAGTCGGCTATCTCTTTATTTTGGCATG
cysD     CTATAGTCGTGTAATCGAAATGTTTAGCAAAAAACG
yahO     TTCCAGTAATCTACACTACTTATTTAATCAGTCCGA
STM1940  GTTCAGGGTGTTTTCTGCACAGTTTAATTAAAGGTA
STM0306  ATAGACCGTTTTTTGGGCTTCGTTTATGTAATCGTT
udg      CTGCAAAAATGTTTAAGCCCGGTTTAATACCGGGCT
pqaA     TTATTGTTCCCTTCTTCCGTTATTTATCAAAAGTAT
traM     CAGAGTGCTCCTGAATCCGGAGTTTATAATGTTCTT
bioA     AACCTAAATCTTTTTAATTTGGTTTACAAGTCGATT
spy      GCGAAAAAGTGTTTTTTATACTTTCATTGTTTTACC
pps      GATTCACCGTTTTTTTCGCGGGTTTAAGTATGCCAG
pmrD     TTCCATCGCTATTGCCGTTTTGTTTATCCGTTGATG
nrdH     ATCTTGTATGGTTGAATCTTAATTCAACTACATCTA
STM4157  GATAAACCTGTTTTATTGATTGTTTACGCGGGACAT
phoP     TATTTGTCTGGTTTATTAACTGTTTATCCCCAAAGC
chaB     CAAGCATCTTGTTTTAACAGTGTTTAAATAACAAAA
yajI     NTACGGCGTCCTGACCAGATGATTTAGGAAATGTTA
mig_3    GCTGGCGACCATGCGCATACAGTTTATATCGGAGGA
pagO     NNATTTTGCTGTAAAAGAAATGTTAAACTGGATTTG
```

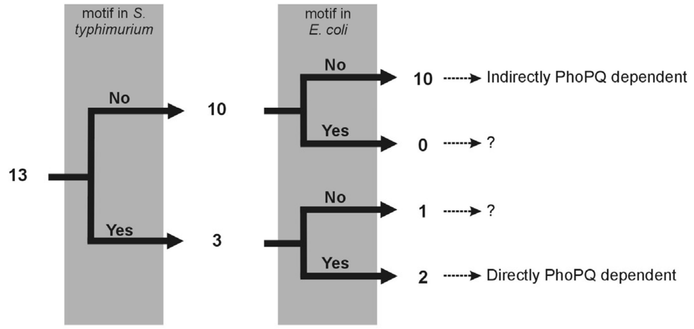
Figure 4.2.B: **Alignment of the detected PhoP motifs in *S. typhimurium*.** Conserved positions are highlighted.

However, as was also mentioned in these experimental studies, it cannot be excluded that alterations at some of these positions just reduce the PhoPQ dependent regulation to a level that cannot be observed by *in vitro* footprinting analysis [141]. Therefore, in order not to bias our initial genome wide screening towards previously made observations and to guarantee a maximal sensitivity, we first screened the selected promoter regions with a motif model that describes only one half site of the PhoP box (high sensitive, low specific screening). In addition, we also performed a second more stringent screening taking also into account the specific sequence requirements described above (low sensitive, high specific screening). The first screening resulted in 130 and 147 putative target operons in *E. coli* and *S. typhimurium* respectively. When taking into account the more stringent criterion, 42 and 34 operons potentially directly dependent on PhoP were retained in *S. typhimurium* and *E. coli* respectively (see Fig. 4.2.A and Fig. 4.2.B).

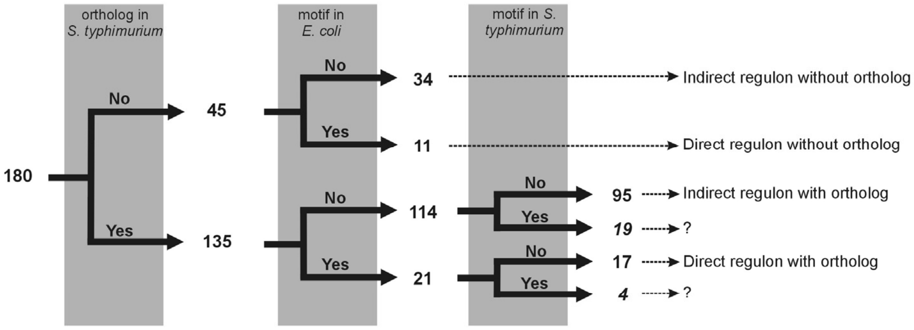
4.2.3 Combination of motif screening results and microarray data

In this section we compare the overlap of the PhoPQ regulon between *E. coli* and *S. typhimurium* i.e. between the 193 operons of *E. coli* and the 189 operons of *S. typhimurium* that were identified as upregulated by PhoPQ in each of the respective organisms. The directly regulated operons were distinguished from the indirectly regulated ones by the presence of a PhoP motif in the region upstream of a gene. Results are based on the stringent screening criteria. A schematic representation can be found in Fig. 4.3.

A. Part of the PhoPQ regulon present in *S. typhimurium* and *E. coli*



B. Regulon exclusively PhoPQ dependent in *E. coli*


ortholog in S. typhimurium
motif in E. coli
motif in S. typhimurium
180
No
Yes
45
135
No
Yes
No
Yes
34 → Indirect regulon without ortholog
11 → Direct regulon without ortholog
114
21
No
Yes
No
Yes
95 → Indirect regulon with ortholog
19 → ?
17 → Direct regulon with ortholog
4 → ?

C. Regulon exclusively PhoPQ dependent in *S. typhimurium*

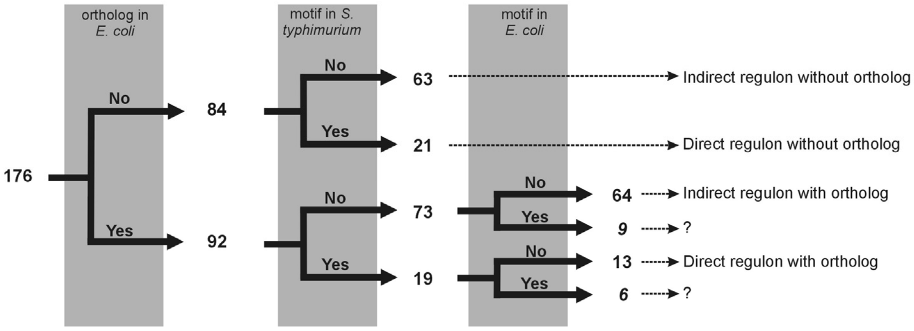

ortholog in E. coli
motif in S. typhimurium
motif in E. coli
176
No
Yes
84
92
No
Yes
No
Yes
63 → Indirect regulon without ortholog
21 → Direct regulon without ortholog
73
19
No
Yes
No
Yes
64 → Indirect regulon with ortholog
9 → ?
13 → Direct regulon with ortholog
6 → ?

Figure 4.3*: Overview of the PhoPQ dependent genes. A.**: Overview of the PhoPQ dependent genes, differentially expressed both in *S. typhimurium* and *E. coli.* Each branch of the tree corresponds to a specific subcategory. The number of operons for the different subcategories is displayed. ***B: Overview of the PhoPQ dependent genes, differentially expressed in *E. coli* but not in *S. typhimurium.* Legend as in Fig. 4.3.A. Subcategories for which the results need to be interpreted with caution are indicated by a “?”. ***C*:** Overview of the PhoPQ dependent genes, differentially expressed in *S. typhimurium* but not in *E. coli*. Legend as in Fig. 4.3.A.

4.2.3.1 Corresponding genes in both data sets

When comparing both datasets, 13 operons were differentially expressed both in *S. typhimurium* and *E. coli* (Fig. 4.3.A). Only 2 of these 13 orthologous operons contained a PhoP box in their upstream region both in *S. typhimurium* gene and *E. coli* (i.e. *phoPQ* and *slyB*). In both these genes, a PhoP motif was also conserved in the evolutionary related bacteria *Shigella flexneri* and *Yersinia pestis* offering an extra validation for the biological relevance of the identified regulatory motif. In one gene (*pagP*), a PhoP box was found in *S. typhimurium* but was missing in the promoter region of the corresponding ortholog of *E. coli* (*crcA*). These results suggest that there is a direct regulation of *pagP* in *S. typhimurium* while this regulation is indirect in *crcA* in *E. coli*. Alternatively, this motif might be a false positive result. In the remaining 10 operons, no PhoP box was present neither in *E. coli* nor in *S. typhimurium* indicating that these genes are probably indirectly regulated by the PhoPQ two-component system. A detailed list of the operons that are differentially expressed in both organisms can be found in Table 4.1.

Table 4.1: List of PhoPQ dependent operons that are differentially expressed both in *E. coli* and in *S. typhimurium*.

E. coli		*S. typhimurium*		
Operon	Motif	Operon	Motif	Function
galETK	no	*galETK*	no	Galactose metabolism
tktAB	no	*talA-tktB*	no	Transaldolase/transketolase
gcvTH	no	*gcvTH*	no	Glycine metabolism
ompT	no	*pgtE*	no	Outer membrane protein
atpFHA	no	*atpFHA*	no	Membrane bound ATP synthase
malFG	no	*malFG*	no	Maltose transport
dacC	no	*dacC*	no	D-alanyl-D-alanine carboxypeptidase
yraNOP	no	*yraNOP*	no	Endonuclease-isomerase-periplasmatic protein
bioBFCD	no	*bioBFCD*	no	Biotine synthesis
cmtBA	no	*STM2344*	no	Transport of small molecules
crcA	no	*pagP*	yes	Lipid A modification
slyB	yes	*slyB*	yes	Outer membrane protein
phoPQ	yes	*phoPQ*	yes	Regulatory protein – Magnesium sensitive

4.2.3.2 *E. coli* microarray data

Besides the 13 operons that were differentially expressed in both organisms according to the corresponding microarray data, remarkable differences were found for other operons (Fig. 4.3.B). 45 operons that were differentially expressed in *E. coli*, do not have an orthologous operon in *S. typhimurium* nor in evolutionary related bacteria like *Yersinia pestis* and

Shigella flexneri. 11 of these operons displayed a PhoP box in their promoter region while the remaining 34 *E. coli* operons were most likely indirectly regulated by the PhoPQ system (no PhoP box found).

On the other hand, 135 *E. coli* operons that were differentially expressed in the microarray of Minagawa *et al.* [175], and that had an orthologous operon in *S. typhimurium* were not differentially expressed according to our data in *S. typhimurium*. For 17 of these 135 orthologous operons, differential expression and direct regulation (i.e. a PhoP box found) was confined to *E. coli* only. 95 of these 135 Mg^{2+} repressed *E. coli* operons appeared to be indirectly regulated by the PhoPQ two-component system (no PhoP box was detected upstream these operons). Besides the previous results, some inconsistent observations were made. 19 of the PhoPQ dependent differentially expressed *E. coli* operons did not contain a motif while in the corresponding non-differentially expressed *S. typhimurium* orthologs, a PhoP box was present. For a total of 4 operons, in both organisms a PhoP box was found but differential expression was observed in *E. coli* only. These latter two classes might be due to false positive motif hits, measurement errors in the microarray data or incomplete sampling of relevant PhoPQ dependent conditions. Therefore, these results should be interpreted with caution.

4.2.3.3 *S. typhimurium* microarray data

Performing a similar analysis for *S. typhimurium* (Fig. 4.3.C), 84 PhoPQ dependent *S. typhimurium* operons do not have an ortholog in *E. coli* nor in evolutionary related species. In 21 of these 84 unique operons, we found evidence for direct PhoPQ dependency (i.e. a PhoP box is found in the promoter region of the first gene of the operon) while the remaining 63 operons were indirectly PhoPQ dependent (i.e. no PhoP box present in the promoter region of the gene). 92 of the significantly differentially expressed *S. typhimurium* operons contained an orthologous operon in *E. coli*. However, these operons in *E. coli* seemed not to be differentially expressed in a PhoPQ dependent way. In 13 of these 92 *S. typhimurium* operons, evidence for direct PhoPQ dependent regulation was found. On the other hand, 64 of the 92 operons did not display a PhoP motif in their promoter region pointing towards indirect regulation by the PhoPQ system in *S. typhimurium* operons. Also in this dataset, a limited number of inconsistencies was observed for which no clear conclusions could be drawn. For 9 of the 92 PhoPQ dependent differentially expressed *S. typhimurium* operons that did not contain a PhoP motif, a PhoP box was present in the corresponding non-differentially regulated *E. coli* orthologs. Another 6 operons contained a PhoP box in both organisms although differential expression only was observed in *S. typhimurium*.

4.2.4 Functional classes

Based on the functional classification schemes of the EcoCyc and the *S. typhi* Sanger database, we retrieved for each PhoPQ regulated gene in *E. coli* and *S. typhimurium* its corresponding functional class. We subsequently determined which functional classes were significantly enriched in the sets of directly/indirectly PhoPQ dependent genes in both *E. coli* and *S. typhimurium*. It should be mentioned that many of the detected PhoPQ dependent targets (especially the unique genes) are not yet annotated in these databases, resulting most probably in an underestimation of the degree of overrepresentation of certain functional classes.

For both the stringent and the non-stringent estimation of the direct PhoPQ dependent regulon, we calculated the enrichment of functional categories in the selected gene sets. The enrichment was most pronounced using the gene sets resulting from the stringent criteria (lower p-values). Moreover, the functional classes that were enriched in these gene sets better corresponded to the known functionalities of PhoPQ dependent genes and the average size of the stringent direct PhoPQ dependent regulon approximated the previously predicted direct PhoPQ regulon size of 40 loci [246,247]. Therefore, the most stringent criteria seemed to be the best approximation of the true regulon composition and these were used for further characterization of this regulon.

In Table 4.2, the overrepresented functional classes in both the directly and indirectly PhoPQ dependent regulons are shown for *S. typhimurium*. The categories of “small molecule metabolism” (subcategories “degradation” (1.A), “energy metabolism” (1.B)), “protein translation and modification” (3.A.8), “transport of anions and carbohydrates, organic acids and alcohols” (4.A.3 and 4.A.5) and “adaptations and atypical conditions (5.F)” are mainly overrepresented in the indirect regulon. Overrepresentation of the class “aminoacyl tRNA synthetases” (3.A.5) (containing *pheS, pheT, rbn*), “transport of amino acids and amines” (4.A.1 containing *artP, artI, nrdF, STM4157, artM, PSLT043, ybj*) and the class “drug/analogue sensitivity” (5.D to which belong *pmrD*, *pqaA* and *pagP*) is confined to the direct regulon.

Significantly enriched both within the subset of direct and indirect PhoPQ target genes is the category for “central intermediary metabolism (1.C)”, e.g. the directly regulated genes of the sulphur metabolism (*cysC, cysI, cysJ, cysN, cysD*)), the functional class involved in the “synthesis and modification of the cell envelope (3.C.1)”, to which belong the directly PhoPQ regulated genes *slyB, pmrF, yfbG, STM2303, STM1940, spy, STM1864, lpxO, STM2302, ybjY, ybaY, STM4065, yahO* and *pagO*, and the functional category responsible for “cation transport (4.A.2)”, to which belong the directly regulated genes *chaB, artQ* and *kefB*.

Table 4.2: **Overview of enriched functional classes for PhoPQ dependent *S. typhimurium* genes**. Column 1 and 2: Number and description of the functional categories according to the *S. typhi* Sanger database. Column 3: p-values for the indirect PhoPQ regulon, Column 4: p-values for the direct PhoPQ regulon genes. p-values describe the statistical significance of the enrichment of a functional class within the subset of PhoPQ regulated genes. The most significant fields (p-value below 0.2) are indicated in bold.

Number	Description	Indirect	Direct
1	**Small molecule metabolism**	**0.10**	0.60
1.A	*Degradation*	**0.16**	1.00
1.A.1	Carbon compounds	**0.18**	1.00
1.B	*Energy metabolism*	**0.10**	0.97
1.B.1	Glycolysis	0.23	1.00
1.B.5	Pentose phosphate pathway	**0.01**	1.00
1.B.5.b	Non-oxidative branch	**0.01**	1.00
1.B.7	Respiration	0.99	1.00
1.B.7.a	Aerobic	0.86	1.00
1.B.7.b	Anaerobic	0.98	1.00
1.B.8	Fermentation	**0.01**	0.22
1.C	*Central intermediary metabolism*	**0.01**	**0.01**
1.C.2	Gluconeogenesis	1.00	**0.06**
1.C.3	Sugar nucleotides	**0.02**	**0.17**
1.C.5	Sulphur metabolism	1.00	**0.01**
1.D	*Amino acid biosynthesis*	0.78	1.00
1.D.2	Aspartate family	0.73	1.00
1.D.3	Serine family	0.36	1.00
1.D.4	Aromatic amino acid family	0.70	1.00
1.D.6	Pyruvate family	0.65	1.00
1.F	Purines, pyrimidines, nucleosides and nucleotides	1.00	0.31
1.F.5	Miscellaneous nucleoside/nucleotide reactions	1.00	**0.01**
1.G	*Biosynthesis of cofactors, prosthetic groups and carriers*	0.61	0.60
1.G.1	Biotin	0.32	0.33
1.G.10	Thioredoxin	0.42	**0.16**
1.G.9	Riboflavin	0.22	1.00
2	**Broad regulatory functions**	0.23	0.75
3	**Macromolecule metabolism**	0.74	0.30
3.A	*Synthesis and modification of macromolecules*	0.89	0.51
3.A.2	Ribosomal protein synthesis and modification	0.80	1.00
3.A.3	Ribosome maturation and modification	0.36	1.00
3.A.5	Aminoacyl tRNA synthetases and their modification	1.00	**0.04**
3.A.7	DNA replication, modification, recombination and repair	0.95	0.77
3.A.8	Protein translation and modification	**0.10**	1.00
3.A.9	RNA synthesis, RNA modification and DNA transcription	1.00	0.32
3.B	*Degradation of macromolecules*	0.88	1.00
3.B.3	Proteins, peptides and glycopeptides	0.46	1.00
3.C	*Cell envelope*	0.39	**0.19**
3.C.1	Membranes, lipoproteins and porins	**0.20**	**0.08**
3.C.2	Surface polysaccharides, lipopolysaccharides and antigens	0.39	0.38
3.C.4	Murein sacculus and peptidoglycan	0.33	1.00
4	**Cell processes**	0.63	0.69
4.A	*Transport/binding proteins*	0.31	**0.14**
4.A.1	Amino acids and amines	1.00	**0.00**

4.A.2	Cations	**0.16**	**0.01**
4.A.3	Carbohydrates, organic acids and alcohols	**0.04**	1.00
4.A.5	Anions	**0.01**	1.00
4.A.6	Other	0.81	1.00
4.C	*Cell division*	0.75	1.00
4.D	*Chemotaxis and mobility*	0.53	1.00
4.G	*Detoxification*	0.29	1.00
4.H	*Cell killing*	0.22	1.00
4.I	*Pathogenicity*	0.93	0.97
5	**Other**	0.28	0.28
5.A	*IS elements, Phage-related functions and prophage*	1.00	0.54
5.D	*Drug/analogue sensitivity*	1.00	**0.01**
5.F	*Adaptions and atypical conditions*	**0.03**	1.00
5.H	*Conserved hypothetical protein*	0.21	0.92
5.H.a	Hypothetical protein (conserved in *E. coli*)	0.72	0.97
5.H.b	Hypothetical protein (conserved in org. other than *E. coli*)	**0.10**	0.74
5.I	*Unknown*	0.43	**0.10**

In *E. coli,* a significant part of, both direct and indirect dependent PhoPQ genes is also enriched in functional categories involved in general metabolism (7 e.g. “central intermediary metabolism and carbon utilization”), in “transport” (9) and in functions related to “cell wall and membrane structure” (2). Besides in “these previous classes”, the indirect regulon is also enriched in the categories “cell processes” (1), and information transfer (5). Some genes mainly those of the direct regulon are involved in “regulation” (8). Note, however, that the functional analysis of these *E. coli* genes is only partial because 30 of the potentially directly regulated PhoPQ genes are not assigned to any functional class in the EcoCyc database.

***Table 4.3*: Overview of enriched functional classes for PhoPQ dependent *E. coli* genes.** Column 1 and 2: Number and description of the functional categories according to the EcoCyc database. Categories of ontology level 1 are indicated in bold, categories of level 2 in italic, and categories of level 3 in normal. Column 3: p-values for the indirectly PhoPQ regulated genes, Column 4: p-values for the directly PhoPQ regulated genes. p-values describe the statistical significance of the enrichment of a functional class within the subset of PhoPQ regulated genes. The most significant fields (p-value below 0.2) are indicated in bold.

Number	Description	Indirect	Direct
1	**Cell Processes**	**0.00**	**0.13**
	adaptations	0.92	1
	Cell division	0.24	1
	protection	0.38	**0.18**
2	**Cell Structure**	**0.01**	**0.04**
	flagella	**0.01**	1
	membrane	**0.05**	**0.03**
	murein	0.50	1
	pilus	0.25	**0.01**
	ribosomes	**0.12**	1
4	**Extrachromosomal**	0.21	0.68

	transposon related	1.00	1
5	**Information transfer**	**0.08**	0.71
	DNA related	**0.11**	**0.18**
	protein related	**0.00**	0.5
	RNA related	1.00	0.98
6	**Location of Products**	1.00	1
	cytoplasm	0.54	0.74
	extracellular	**0.15**	1
	inner membrane	0.31	**0.09**
	outer membrane	0.65	0.53
	periplasm	0.67	**0.01**
7	**Metabolism**	**0.00**	**0**
	Biosynthesis building blocks	1.00	1
	flagellum	**0.01**	1
	large molecule carriers	0.86	**0**
	lipopolysaccharide	**0.00**	1
	Biosynthesis macromolecules	1.00	1
	Carbon Utilization	0.64	0.26
	amines	**0.19**	**0.01**
	amino acids	**0.15**	**0.01**
	Carbon compounds	0.43	0.8
	fatty acids	0.47	0.3
	Central intermediary metabolism	**0.02**	0.46
	energy metabolism	1.00	1
	aerobic respiration	**0.01**	1
	anaerobic respiration	**0.01**	**0.01**
	fermentation	**0.10**	1
	glycolysis	0.77	1
	TCA cycle	**0.19**	**0.02**
	energy productions/transport	1.00	1
	electron acceptors	**0.03**	**0.04**
	electron carriers	**0.04**	**0.01**
	electron donors	**0.01**	1
8	**regulation**	0.24	**0.07**
	genetic unit regulated	0.99	0.85
	type of regulation	0.62	**0.14**
	DNA structure level	0.60	1
	posttranscriptional	0.63	0.28
	unknown	**0.13**	1
9	**transport**	**0.00**	**0.11**
	group translocators	**0.10**	1

4.2.5 Detailed analysis of novel direct PhoPQ targets

Most of the known PhoPQ regulated genes with an experimentally verified PhoP box could be retrieved by our analysis (*phoP*, *slyB, pdgL/pcgL*, *pmrD* in *S. typhimurium* [141] and (*phoP, slyB, yrbL, ybcU/vboR* and *rstA*) in *E. coli* [175]). The *mgtA* gene was missing in our analysis because it was inaccurately measured (low significance) on our arrays and was not spotted on the microarray of Minagawa *et al*. [175]. Retrieving the known PhoPQ targets illustrates the predictive power of our analysis and allows suggesting the presence of new promising directly PhoPQ regulated targets in *S. typhimurium*.

To the functional category related to the "cell envelope" belong several directly PhoPQ regulated genes with as most promising novel targets *pagO, lpxO* and *STM1940*. *pagO* encodes an integral membrane protein, previously described as PhoPQ dependent [95] that is similar to a product of the *Yersinia* virulence plasmid. According to our analysis, *pagO* would be directly PhoPQ dependent. *lpxO* codes for a dioxygenase that plays a role in lipid synthesis [83]. Belonging to the same functional class (subcategory "membranes and lipoproteins") is the novel potential direct PhoPQ target STM1940. This gene codes for a cell wall-associated hydrolase in *S. typhimurium*.

The set of directly PhoPQ dependent genes we identified was also statistically enriched for the functional class related to "drug sensitivity". Three of the genes we identified from this category (*pmrD*, *pqaB*, *pagP*) are important for resistance to antimicrobial peptides. The *pmrD* gene is shown to be directly PhoPQ regulated [125]. The *pagP* gene was shown to be PhoPQ dependent and important for the modification of lipid A [99], but in addition, our results point towards its direct regulation. A third gene, *pqaB* is shown to be involved in antimicrobial peptide resistance in *S. typhi* but is suggested to be indirectly PhoPQ regulated via the PmrAB system [16]. This means that the detected PhoP box is either false positive or points towards a complex dual regulation of this gene.

Although not significantly enriched within the set of directly PhoPQ regulated genes, the functional class related to pathogenicity and virulence also contained potential PhoPQ dependent targets (*pagC, mgtC, virK* and *STM0306*). Although the three former genes have previously been related to PhoPQ dependency, we find here in addition evidence for a potential direct dependency on PhoPQ. PagC is a membrane protein [94] that was shown to be essential for survival within macrophages and for virulence in *S. typhimurium* [212]. MgtC was also shown to be required for intramacrophage survival and growth in low Mg^{2+} conditions [25]. The exact function of MgtC, however, is still unknown. VirK contributes to the resistance of *S. typhimurium* against polymyxine B and is important for the systemic infection of the bacteria [54]. Moreover, both *pagC* and *mgtC* are located on the SPI-3 pathogenicity island and were previously suggested to be acquired via horizontal gene transfer [26,94]. This would imply that – after acquisition – both genes were integrated into the PhoPQ dependent regulatory cascade of *S. typhimurium*. *STM0306,* a fourth PhoPQ regulated gene that is involved in pathogenicity, is a paralog of the *S. typhimurium*

SapA protein which was shown to play a role in virulence [201]. An ortholog of SapA is present in *E. coli* and *Erwinia chrysanthemi.* In the latter, SapA was shown to play a role in the resistance of this organism to antimicrobial peptides [149].

4.3 Discussion

Based on the combination of microarray- and motif data, the PhoPQ dependent regulon in both *E. coli* and *S. typhimurium* was reconstructed. This reconstruction is the best estimate of the true regulon composition that can be made at this stage, due to the presence of inherent variation commonly observed in microarray experiments, the uncertainty about the motif requirements of the PhoP motif model and the restricted availability of experimental data. Remark that, for instance the *E. coli* and the *Salmonella* microarray experiments were not performed in exactly the same conditions. However, we can expect that the conditions that were used, trigger a large part of the PhoPQ regulon in both organisms. Indeed, the *E. coli* dataset tests the influence of Mg^{2+}, the most important PhoPQ signal. The *Salmonella* experiment on the other hand makes use of a constitutive mutant and therefore is relatively independent on the conditions applied. Conclusively, the size of the *E. coli* regulon might be underestimated as compared to the size of the *S. typhimurium* regulon. This, however, does not prevent us from studying the overlap in regulon composition: a large overlap in regulon composition would imply that at least the targets induced by the major PhoPQ trigger i.e. Mg^{2+} detected in *E. coli* would be contained within the PhoPQ regulon in *Salmonella*. This not being the case as we observed in this study, indicates that the true overlap indeed will be low.

Statistical analysis of the PhoPQ related expression data of *S. typhimurium* clearly pointed out the pleiotropic nature of the PhoPQ regulatory system. We found evidence for at least 2855 genes being affected by the PhoP mutation. From these, only a limited subset of approximately 42 operons was directly regulated by PhoP, which is in accordance with previous predictions [246,247]. The seemingly contradictory observations of the pleiotropic nature of the PhoPQ system on the one hand and the small size of the direct PhoPQ dependent regulon on the other hand can be explained by the high number of regulatory proteins that are part of the direct regulon (e.g. *mig-14, slyA, pmrD*, *traM, STM0859*). Each of these regulators can activate other regulatory cascades [141], allowing a combinatorial increase of PhoPQ affected genes e.g. *pmrD* encodes a protein that posttranscriptionally activates the two-component PmrAB system [88,125].

In an initial attempt to compare the overlap in the PhoPQ regulon composition between *E. coli* and *S. typhimurium* we used a very strict definition of overlap: an overlap was defined when two orthologs were in both organisms differentially expressed (data not shown). Although such stringent analysis will result in a highly reliable set of potential PhoPQ targets, overlapping targets might escape detection because of potential type II error in the expression data (presence of false negatives in either one of the organisms). Therefore, we repeated the analysis using the low stringent criteria described in the materials and methods section. Using this higher sensitivity analysis, however, did not drastically increase the detected overlap in regulon composition proving the biological relevance of the detected low overlap. Indeed, when comparing the average regulon composition using non-stringent criteria, an overlap of only 13 PhoPQ regulated operons was observed (i.e. 26 genes).

For the estimation of the direct regulon, we made approximations based on both stringent and non-stringent sequence requirements of the PhoP motif, assuming that the true sequence requirements must be somewhere in between. Besides the *mgtA* gene that was missing in our analysis for reasons explained before, only two of these 13 overlapping operons (*phoPQ* and *slyB*) were directly regulated by the PhoPQ two-component system in both organisms.

In addition to these overlapping operons, both organisms had a considerable set of operons for which the PhoPQ dependency was only confined to either one of the two organisms. These results point towards a high specialisation of the PhoPQ regulon in either one of the species. This might not be so unexpected in view of the major role the PhoPQ system plays in determining the virulence phenotype of *S. typhimurium*, a phenotype that is absent in the non-pathogenic related strain *E. coli* K12 [88]. This also explains why a large part of the direct regulon comprises genes that are unique for both species. These general conclusions seemed relatively independent of the stringency of the definition of "regulon overlap" and of the motif requirements used, indicating that the true overlap between the PhoPQ regulon compositions in both species is indeed low.

In *S. typhimurium*, significantly enriched functional classes within the set of directly PhoPQ regulated genes are involved in the "central intermediary metabolism", "synthesis and modification of cell envelope", "cation transport" and "drug sensitivity". The most significantly enriched classes (i.e. lowest p-values) in the subset of *E. coli* PhoPQ directly dependent genes are also related to "general metabolism" and "cell and membrane structure". Despite the limited overlap in regulon composition between both species, the PhoPQ regulatory system seems to have conserved common functions in both species.

4.4 Conclusion

Our analysis shows how a regulatory system that is very well conserved [123] and that corresponds to the same extracellular signal can become integrated in a relatively short time period (120-160 million years [45]) in seemingly completely different pathways. This acquisition of novel target genes might explain the high ability of prokaryotic organisms to evolve novel phenotypes and adapt to specific niches. The PhoPQ regulon might have recruited genes that contribute to a virulence phenotype in *S. typhimurium* but not in *E. coli* K12.

4.5 Material and methods

Firstly, PhoP dependent operons were identified in *S. typhimurium* based on the microarray dataset of this study and in *E. coli* based on a previously published dataset [175]. Subsequently, promoter regions of these PhoPQ dependent operons were screened for the presence of the PhoP regulatory motif. This motif information was used to distinguish in each of the respective organisms the direct from the indirect regulon. Eventually, overlap between the regulons in both organisms was identified.

4.5.1 Identification of PhoPQ dependent genes based on expression data

To identify genes dependent on PhoPQ in E. coli, we used the results described by Minagawa *et al.* [175]. They used cDNA microarray experiments to identify target genes of the Mg^{2+} stimulon that responded to the availability of external Mg^{2+} in a PhoPQ dependent manner. In their study, wild type, *phoP* and *phoQ* defective *E. coli* strains were grown in the presence and absence of Mg^{2+} respectively.

For the identification of the PhoPQ regulon in *S. typhimurium,* we set up a microarray experiment (data available at the supplementary information) in which RNA was isolated from strains of *Salmonella enterica* serovar Typhimurium ATCC 14028s harbouring either a null mutation (*phoP::Tn10dCm*) or constitutive mutation (*phoQ24*) in the genes encoding the PhoPQ two-component system [174]. Strains were grown to mid-log phase in M9 minimal media prior to harvesting RNA and the array experiments were performed as described in Bader *et al*. [13].

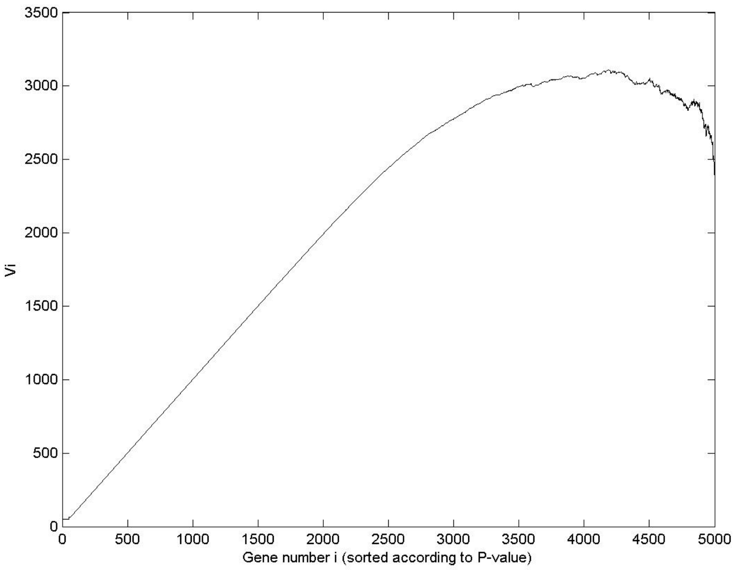


Figure 4.4.A: **Plot of Vi versus the gene number (sorted according to their p-value).** $V_i = (i\ \ p_i.n) / (1 - p_i)$ where *n* is the total number of genes in the dataset, *p* is the pvalue of gene *i*, *i* is the rank order of a gene after sorting all genes according to their p-value, and V_i reaches a constant level at 2855 genes, which is an estimate for the number of actually differentially expressed genes.

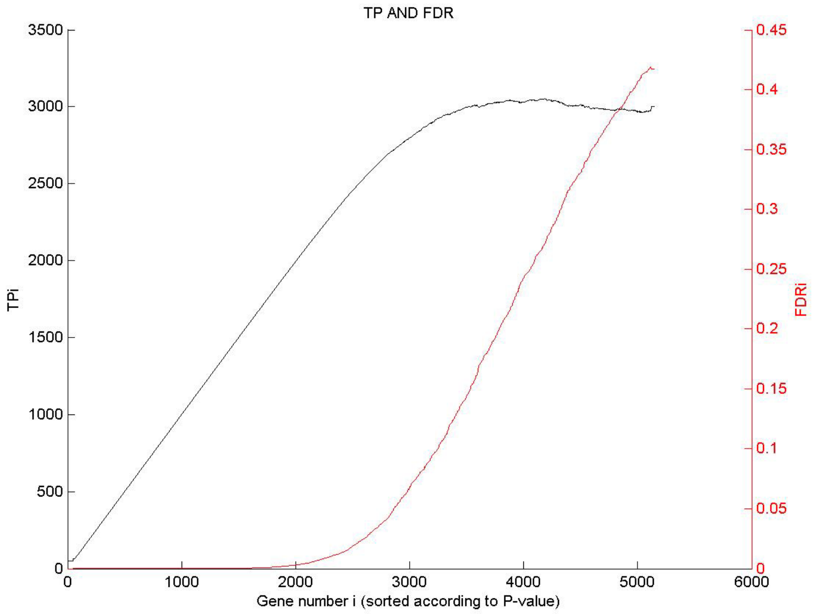


Figure 4.4.B: **Number of true positives (TP, dashed line) / FDR (solid line) versus the gene number (genes are sorted according to their p-value).**

Data were statistically analysed by combining six replicates for each experiment using the maximum-likelihood analysis of Ideker *et al.* [115,116]. This method calculates for each gene a generalized likelihood ratio test statistic (λ value) and the ratio of the mean intensities for the two conditions (μ-ratio). The λ values were converted into p-values using a χ^2 cumulative probability distribution with 1 degree of freedom [116]. These p-values were used to estimate the number of actually differentially expressed genes (Fig. 4.4.A) and subsequently to plot the false discovery rate (FDR) and the number of true positives for each critical value of λ (Fig. 4.4.A and 4.4.B). Detailed description on how to interpret these plots can be obtained from Storey and Tibshirani [251] and De Smet *et al.* [53]. In order to minimize the type-1 error (number of false positives), we chose for a high critical threshold of λ (λ cut-off of 40) resulting in 324 significantly differentially expressed genes. Under the assumption that genes, directly regulated by PhoPQ will be most severely affected by the PhoP mutation, we selected from the 324 most significantly differentially expressed genes (using a λ cut-off of 40) those for which the μ-ratio was less than –0.30 (genes upregulated by PhoP). This additional selection step reduced this number of selected genes to a subset of the 214 most influenced genes. This number approximated the number of *E. coli* PhoPQ upregulated genes selected by Minagawa *et al.* [175].

4.5.2 Sequence-based analysis

4.5.2.1 Intergenic sequences

For sequence analysis, genome sequences were obtained from GenBank (NC_003197 for *S. typhimurium* LT2, NC_003277 for *S. typhimurium* LT2 plasmid and NC_000913 for *E. coli* K12) [21]. Intergenic regions used in this study were extracted from the genome sequences using the modules implemented in INCLUSive 3.0 [269] to automatically parse GenBank files. Here, we define intergenic regions as the non-coding sequence between two coding sequences. No overlap was allowed with the coding sequences.

4.5.2.2 Identification of orthologs

Cluster of orthologs were identified using TribeMCL (version 02.277) [67]. The obtained clusters contain proteins that are each others close homologs (either orthologs or paralogs). The pairwise BLAST scores obtained by mutually aligning the whole-genome sequences using blastp (version 2.1.2) [5] were used as input of TribeMCL. Stringent criteria were applied to only retain closely related orthologs and paralogs (cut-off of the BLAST hit was an e-value of $1e^{-80}$). When the BLAST hit of a protein against itself resulted in an expectation value higher than $1e^{-80}$, we used less stringent criteria (a cut-off of $1e^{-20}$). Because the analysis of the PhoPQ regulon is performed at an operon level, we defined in the context of this study an operon as being unique if for all genes of an operon in one organism (e.g. *E. coli*) no orthologs are found in the other organism (e.g. *S. typhimurium*).

4.5.2.3 Operon prediction

Note that for our analysis it was of importance not to detect highly reliable operons (high selectivity, low sensitivity), but to unveil anything that could possibly be an operon (high sensitivity, low selectivity). Therefore, as compared to studies focusing on operon prediction, we used deliberately non-stringent criteria for our operon prediction: the intergenic distance was used as the main criterion for defining an operon [181,229]. When two genes were located on the same strand and the distance of the intergenic sequence between these two genes was smaller than a predefined cut-off value, these two genes were considered residing in the same operon. For both *E. coli* and *S. typhimurium* we used a cut-off value of 40 nucleotides to predict operons.

These cut-off values were determined as follows: we first predicted operons in *E. coli* using different cut-off values. Each of these predicted sets was compared to the operon prediction of RegulonDB version 3.2 that was used as a benchmark set [230]. An operon in RegulonDB was considered as identical with our own predicted operon if both operons contained exactly the same genes. A maximal match of the predicted operons in *E. coli* with the RegulonDB database was obtained using the cut-off value of 40 nucleotides, mentioned above. Since for *S. typhimurium* no such benchmark dataset existed, we used the cut-off value on the *S. typhimurium* intergenic distance that maximized the match between the predicted set in *E. coli* and *S. typhimurium*. A match in this context means that an operon in *E. coli* and *S. typhimurium* contains orthologous genes only. For *E. coli* the average number of genes in an operon was predicted to be 2.88 while for *S. typhimurium* an average of 2.90 genes per operon was found.

4.5.2.4 Motif detection

We used MotifSampler 3.0 [268] to construct the PhoP motif model (i.e. a probabilistic representation of the DNA pattern). The MotifSampler is a motif detection algorithm based on Gibbs sampling that allows retrieving statistically overrepresented patterns in the promoter regions of coregulated genes [267]. Based on a training set containing promoter regions of known PhoPQ regulated genes (*phoP, mgtA, pmrD, pdgL* and *slyB* in *S. typhimurium* [141,311] and *yrbL, ybcU/vboR* and *rstA* in *E. coli* [175]), the motif detected with the highest information content had a length of six nucleotides and a consensus site (T/G)GTTTA. This corresponds to one half site of the previously suggested dyad PhoP motif [248,311]. The motif model corresponding to this motif was used for the genome wide motif screening of all intergenic sequences of respectively *E. coli* and *S. typhimurium* using MotifLocator 3.0 [160]. This algorithm uses the motif model to calculate a score for each window (with length similar to the length of the motif model) in the intergenic sequences. The threshold score was set at 0.75 (non-stringent screening criterion). This threshold corresponds to the selection of approximately the 2% best scoring motif hits of all possible motif positions in the genome if the intergenic regions of the whole genome of respectively *S. typhimurium* and *E. coli* would be screened (*S. typhimurium* 1.91% and *E. coli* 1.96%).

For the stringent screening criterion, we included extra restrictions on the PhoP motif model that were derived from previous publications [175,175,311], in which the PhoP motif was described as a direct repeat of (T/G)GTTTA. A conserved thymine in the first half site (at position 3), two conserved thymines and one conserved adenine in the second half site (at positions 3, 4 and 6 respectively) were shown to be essential for the binding of the PhoP regulatory protein. Therefore, as additional criterion we explicitly required that these four positions were conserved inside and around the motif instances retrieved from the first screening.

4.5.3 Calculating the overlap in regulon composition

To identify the direct regulon, we checked whether the intergenic regions of the first genes of operons that were selected as differentially expressed in either one (or both) of the organisms contained a PhoP motif. To prove univocally the presence of a restricted overlap in regulon composition of the PhoPQ system between the two organisms compared (see discussion), we used extremely non-stringent criteria to calculate this overlap: for each gene differentially expressed by PhoP in one organism, we first identified the operon to which it belonged (based on the operon prediction outlined above). If at least one of the orthologs of the genes belonging to that operon turned out to be differentially expressed in the other organism as well, the operon was considered differentially expressed in both organisms (i.e. identifying an operon as PhoPQ dependent in both organisms does not necessarily imply that all genes of the operon were identified as differentially expressed in both organisms).

4.5.4 Enrichment of functional classes

To identify which functional classes were enriched in the directly PhoPQ dependent *E. coli* and *S. typhimurium* genes, the functional classifications of respectively EcoCyc release 7.6 [128] and the *S. typhi* Sanger database [199] were used. To use the annotation of the Sanger database of *S. typhi* strain C18 for *S. typhimurium*, a mapping between the *S. typhi* and *S. typhimurium* gene names based on ortholog sequence information was made. Functional enrichment of PhoPQ dependent targets was calculated using the hypergeometric distribution [262], which assigns to each functional class a p-value. This p-value describes for each functional class the probability that in a random set of genes, the same number of genes of that specific functional class would be observed. In detail, the probability of observing at least *k* predicted PhoPQ regulated genes from a functional category within the total number of predicted PhoPQ target genes (*n*) is given by:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}}$$

where *f* is the total number of genes in the functional category in the EcoCyc database or the *S. typhi* Sanger database and *g* is the total number of genes within the genome of *E. coli* and *S. typhimurium* respectively.

Chapter 5

More robust detection of motifs in coexpressed genes by using phylogenetic information

5.1 Introduction

In previous two chapters we identified two regulatory systems based on an in silico approach. However, in both cases information was already available on the motif model that describes the binding sites for these regulatory systems. Identification of these regulons would be more complicated if no prior information on the regulatory motif was available in literature.

In the study of the PmrAB regulatory system, we observed a high overall similarity in the intergenic species of the studies species. As a result, not only the motif itself turns out to be conserved, but also its local neighbourhood. This characteristic will complicate the detection of regulatory motifs using a phylogenetic footprinting approach as traditional motif detection algorithms cannot delineate the true biological motif in these long stretches of conserved intergenic sequences.

For the PhoPQ regulation system we concluded that a regulatory system might be conserved in only a small subset of genes even between evolutionary related bacteria. Several motif detection algorithms have already been developed to discover overrepresented motifs in sets of coexpressed genes (for instance [15,110,113,136,146,268,284,301]). The rationale behind these methodologies is that a set of genes regulated by the same transcription factor should contain in their intergenic regions motifs that are statistically overrepresented as compared to their occurrences in unrelated sequences. These methodologies have proven to be useful in many applications (for instance [50,122,145,165]). However, usually their performance rapidly drops as the signal to noise ratio (defined as the number of sequences that actually contain the motif versus the number of sequences lacking the motif) in the data set decreases [208,255], a situation we also observed in the study of the PhoPQ regulon. This drop in performance is also evident in an assessment study of 13 different computational *de novo* motif detection tools, where the number of correct motifs retrieved was very low in sets of noisy data [272]. However, noisy data is mostly the case encountered in sequence sets derived from genome wide expression profiles, such as microarrays. The high noise level in such sequence sets comes as a result of the processing of the microarray data (for instance, the filtering and clustering procedures used), as well as the nature of the biological processes in the cell itself. For instance, when comparing mRNA expression between a wild type and a regulator knock out strain, besides the direct targets of the regulator, indirect targets are also affected in their mRNA expression in the knock out. In such mutants the complete regulatory network acting downstream of the mutated regulator is disturbed. Lists of genes derived from these experiments contain targets of more than one regulatory protein lowering the relative overrepresentation of a particular motif

In order to get over these problems, we describe in this chapter a methodology we developed that exploits orthology information besides coexpression, to discover *de novo* motifs. Orthology information is introduced by phylogenetic footprinting which is based on the assumption that among phylogenetically related species, the regulating sequences in the upstream regions of orthologous genes are selectively conserved by evolution. Phylogenetic footprinting has been used for the detection of motifs with relative success [75,80,160,167,169,194,195,215]. In these approaches, however, the orthology information is only used within a set of orthologous intergenic regions to create motif models. Recently, a number of algorithms were developed that permit for the mutual comparison of the motif models derived from different sets of orthologous intergenic regions, followed by the clustering of the conserved regions that share the same regulatory motif. In one approach, Jensen *et al.* [120] applied phylogenetic footprinting using Gibbs Sampling and then grouped the conserved regions using a Bayesian clustering algorithm. A similar way of clustering motif models derived from phylogenetic footprinting was developed by Qin *et al.* [214] and Van Nimwegen *et al.* [287]. Another approach is developed by Wang and Stormo [295], where conserved regulatory regions were detected using Wconsensus [110] and where the regions sharing the same regulatory motif were clustered by gradually merging motif models of different orthologous sets (PhyloCon). In their recent paper, Wang and Stormo [296] used a different clustering approach; they would first align the models and consequently cluster them according to their alignment scores (PhyloNet). Our methodology is a combination of Gibbs Sampling-based phylogenetic footprinting, with two-step clustering (first aligning motif models, then clustering based on the alignment score). For phylogenetic footprinting, we developed a new algorithm called BlockSampler, which is an extension of the Gibbs Sampling-based MotifSampler [268], optimized towards phylogenetic footprinting. For the alignment of the different motif models returned from BlockSampler, we developed a second algorithm called BlockAligner, which aligns matrices describing conserved regions using a Smith-Waterman approach [242].

The methodology based on these two new algorithms is capable of detecting motifs with weak overrepresentation in a set of coregulated genes. By applying our procedure on gene lists derived from real genome wide expression studies, we show its ability to function effectively in noisy data. In this context, the use of orthology information compensates for a lower degree of motif overrepresentation We also compared its performance on a test data set, with that of two other motif detection tools: one that is exclusively based on coexpression information, AlignACE [113], and another that is most similar to our method, PhyloCon [295]. With this comparison, we demonstrated the robustness of our method over the two mentioned. Using our method on two real biological data sets proved its biological applicability.

5.2 Results

5.2.1 General Strategy

Several motif detection algorithms have been developed to discover overrepresented motifs in sets of coexpressed genes. However, in a noisy data set the motif might not be sufficiently overrepresented to allow detection by classical tools. To be able to recover motifs in such context, we developed a procedure in which we first use phylogenetic footprinting to delineate all potential motifs in each gene. Then, we mutually compare all detected motifs and identify the ones that are shared by at least a few genes in the data set as potential candidates.

The complete analysis flow is represented in figure 5.1 and consists of the following steps: starting from a list of differentially expressed or coexpressed genes, we find the orthologs for each gene in a number of closely related species. The obtained data sets, consisting of the intergenic sequences of orthologous genes, are subjected to phylogenetic footprinting using a Gibbs Sampling tool called BlockSampler (Step 1). This yields a list of conserved regions (blocks) for each of the orthologous data sets, corresponding to all potential motifs. Searching for the motifs shared by at least some of the original list of coexpressed genes, we mutually align our blocks with BlockAligner (step 2), and then we construct a multiple alignment using the pairwise alignment scores to delineate further the potential regulatory motifs (step 3). As our methodology only detects motifs in genes for which orthologous information is available, some motif hits in the initial gene list might escape detection. To recover these motifs, we use the motif models we obtained from the previous step to screen the intergenic sequences of the remaining genes (step 4).

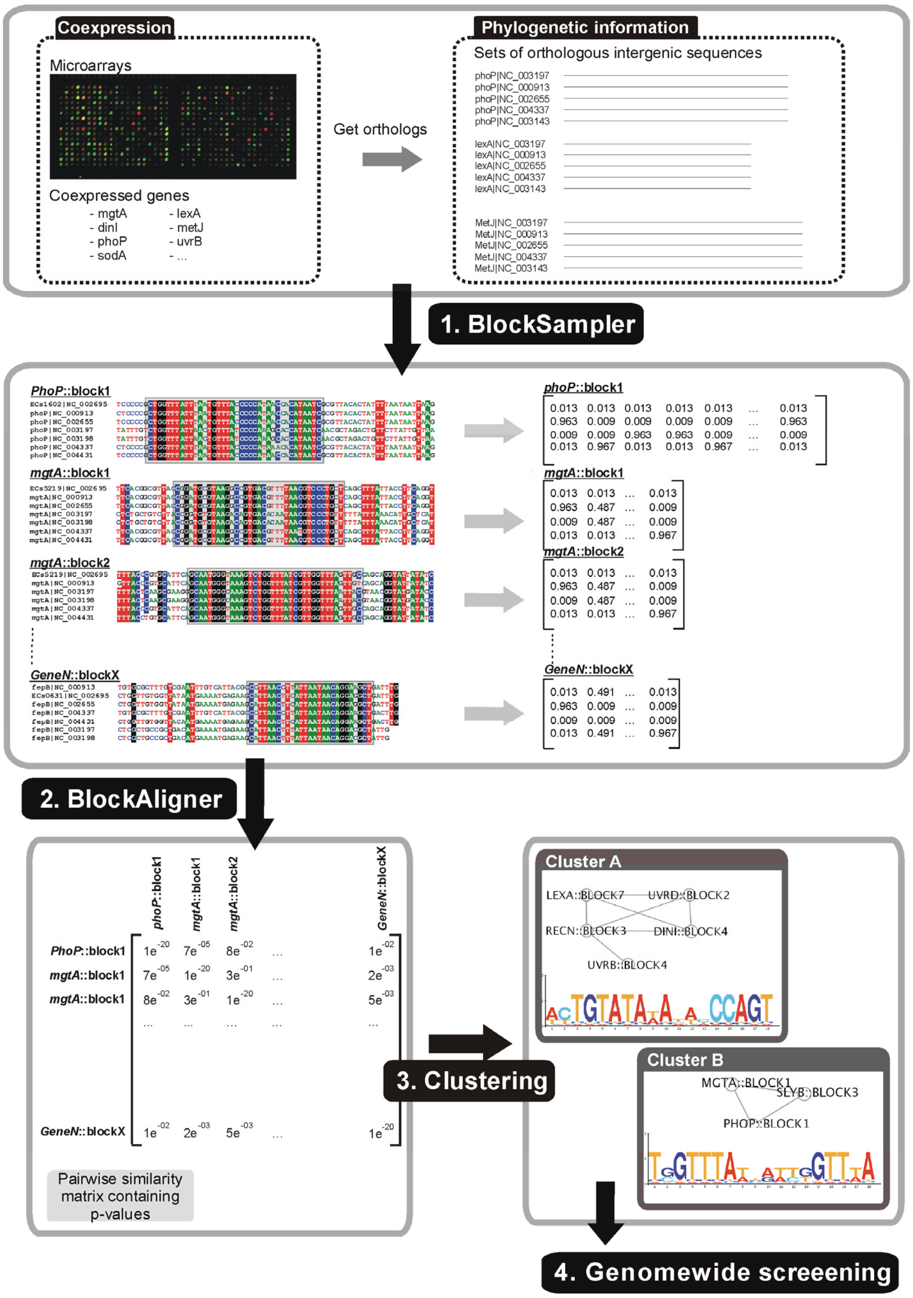


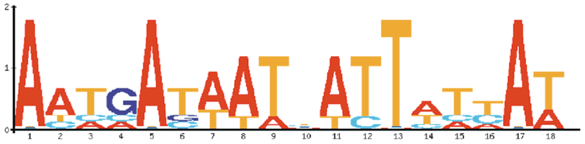
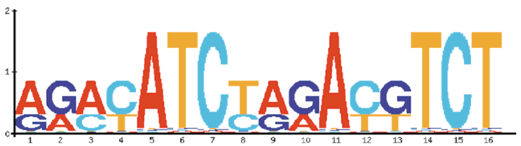
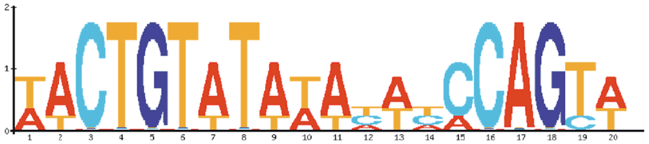
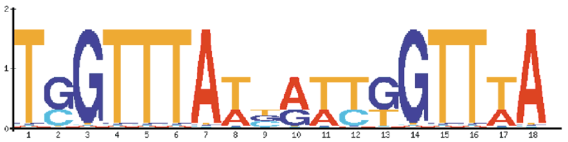
Figure 5.1: **Flowchart of our algorithm.** Input data: Based on microarray data, coexpressed genes are identified. For each of these coexpressed genes, the orthologous intergenic sequences are identified. Step 1: BlockSampler: within a set of orthologous intergenic sequences, conserved blocks are detected and stored as a Position Specific Scoring Matrix (PSSM). Step 2: BlockAligner: The resulting PSSM's corresponding to conserved blocks, are aligned to each other with BlockAligner, which is a local alignment tool that assigns a p-value to each alignment as a quality measure. Step 3: based on all these pairwise comparisons and their corresponding p-values, a clustering is performed to group intergenic regions with a similar regulatory motif. Based on the BlockAligner results, a regulatory motif model is constructed for each cluster. Step 4: the detected motif models are used to screen all intergenic regions that were the input for the methodology.

Here we give an overview of the main tools we used in our flow. Algorithmic details can be found in the methods section.

BlockSampler is used to detect conserved motifs in intergenic sequences of orthologous genes. The algorithm is an adapted version of MotifSampler, a motif detection algorithm based on Gibbs sampling [268]. Due to the specificities of the scoring scheme (see methods) the algorithm does not require the motif to be present in each single orthologous sequence. Although regions conserved in the entire set of sequences will receive a better score, the regions conserved in less than the entire set could still be retained. In contrast to the original implementation [267], BlockSampler allows for scoring each orthologous intergenic sequence in the input data set with its corresponding species specific background model. Since it is often the case with closely related species that big stretches around the motifs are conserved [160,168], BlockSampler was adapted to look for long conserved sequence blocks. BlockSampler also allows for the choice of a reference species, usually corresponding to the species of interest for which the experimental data are available. The consequence of using a reference species is that only the conserved regions that include the reference species will be retained.

BlockAligner performs a local ungapped pairwise alignment, based on dynamic programming for the mutual comparison of sequence blocks, represented as frequency matrices. The scoring scheme of BlockAligner is based on the Kullback-Leibler distance. We used a graph clustering algorithm to group the blocks shared by multiple genes [67]. This algorithm is optimized to cluster sequences based on pairwise alignment scores.

Table 5.1: **Composition of test data sets.** Table containing details about the test data set. For each motif, the genes are shown of which the intergenic region is used in the test data set. L gives the length of the motif model. Cs indicates the consensus score of the motif model created based on the instances in these genes; # indicates the number of genes that contain the corresponding motif; the motif logo gives a visual representation of the information content of the motif model.

Motif	L	Description	Cs	Genes	#	Motif logo
Fur	18	Less conserved motif present in 5 genes	1.02	*entC, fepB, fepD, sodA, yhhX*	5	
MetJ	16	Well conserved motif present in 3 genes	1.21	*metA, metE, metJ*	3	
LexA	20	Well conserved motif present in 5 genes	1.26	*uvrB, uvrD, lexA, dinI, recN*	5	
PhoP	18	Well conserved motif present in 5 genes (dyad motif present in 3 genes, half site present in 2 genes)	1.27	*phoP, mgtA, slyB, bioA, pmrD*	5	

5.2.2 Evaluation of the analysis flow on test data

We applied our analysis flow on test data sets with known motif instances. As shown in table 5.1, the data sets are made up of sequences with instances for the motifs PhoP, LexA, Fur and MetJ, imbedded in a varying number of random sequences. Five collections of data sets were created with an increasing number of random sequences, decreasing the signal to noise ratio from 18- 7% for Fur and LexA and 11- 4% for MetJ and PhoP (dyad motif). For the purpose of evaluating the performance of the methodology, we adopted four measurements: recovery rate, number of false positives, specificity and sensitivity. The former two are a reflection of how well a true motif model is extracted, while the latter two are measurements of the extracted motif's quality. Specifically, the recovery rate measures the number of times a motif model corresponding to a true motif is recovered by the algorithm; the number of false positives measures the average number of motif models not corresponding to the true motifs recovered by the algorithm; the sensitivity measures the ratio of true motif instances that contributed to the recovered motif model; the specificity measures the ratio of false motif instances that contributed to the recovered motif model (for details see methods). Note that these values were calculated before the last screening step of the algorithm, so that all instances recovered by screening are not included in the calculations.

As shown in figure 5.2, our methodology performed well in retrieving the MetJ, LexA and PhoP motif models, with relatively low effect of the increasing noise in the data set. Only for the Fur motif the recovery rate was low, but with high resistance to noise. Analysis of the results showed that during the run of the methodology, the Fur motif instances were clustered in a single cluster (step 3), however, together with degenerated blocks lacking Fur instances. The presence of such blocks in the subsequent motif delineation step causes the consensus score of the Fur motif model to drop below the selected threshold, explaining why it was not frequently recovered using our stringent criteria.

The number of false positives picked up by the algorithm was low (table 5.2), ranging between an average of 0.5 in the low noise data sets to 0.8 in the high noise sets. This shows the robustness of the algorithm against noise in the data.

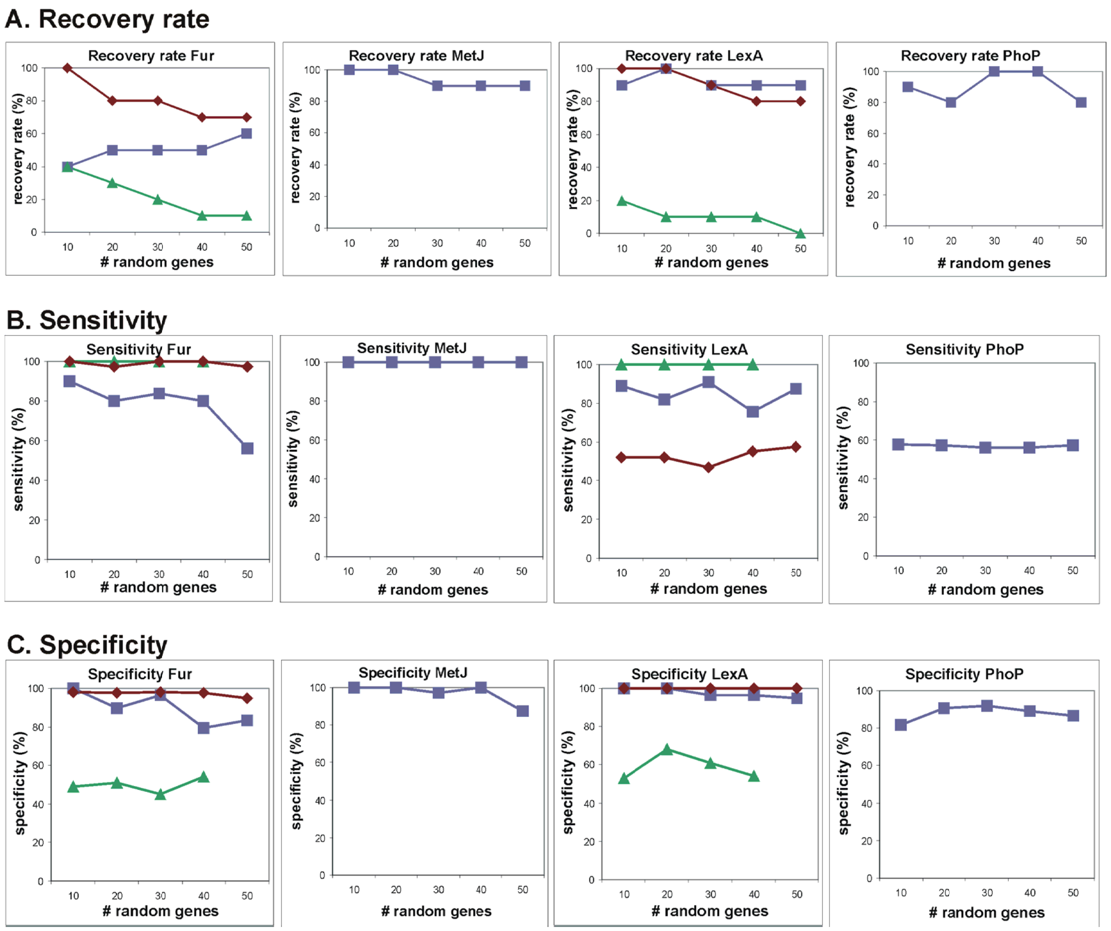


***Figure 5.2:* Plot of the recovery rate, sensitivity and specificity.** For each of the four motifs (Fur, MetJ, LexA and PhoP), the recovery rate (A), the sensitivity (B) and specificity (C) is plotted for each of the three algorithms (BlockSampler: ■, AlignACE: ▲, PhyloCon: ♦). At the X-axis of each plot, the number of added random genes is indicated. No sensitivity or specificity data for a specific algorithm means that this characteristic could not be calculated due to a recovery rate of 0.

***Table 5.2*: Number of false positives.** Number of false positives on average per run obtained by applying the different algorithms on the respective test sets. Number of false positive motif models is given for each number of added random genes (# Random) to the initial test set.

# Random	New methodology	AlignACE	PhyloCon
10	0.50	17.1	5.4
20	0.60	18.1	5.4
30	0.80	17.4	5.1
40	0.50	18.3	5.7
50	0.80	16.9	5.3

Regarding the quality of the motif models recovered, figure 5.2 shows that the algorithm achieved relatively high sensitivity and specificity for Fur, MetJ and LexA, indicating that most true instances contributed to the model, and that few false ones did. In the case of PhoP, the sensitivity was low due to the fact that the model recovered by our algorithm was the dyad variant, while the two instances of the half site variant (*bioA* and *pmrD*) did not contribute to the model.

5.2.3 Importance of the selected species

To assess the importance of adding a distantly related species to the analysis (in our case *Y. pestis*), we tested to what extent motif instances in this species contributed to the motif models returned by our algorithm. Instances of *Y. pestis* seemed to have contributed to the resulting motif models in more than 90% of the cases for MetJ and LexA, more than 80% for PhoP, and more than 75% for the Fur motif. On the other hand, adding *Y. pestis* did not decrease the sensitivity of our motif models: motif instances that were only conserved in *S. typhimurium* and *E. coli,* but not in *Y. pestis* were also recovered. These results indicate that aligned blocks are not dominated by the sequences of the most related species and that adding a distantly related species helps delineating the true motifs without decreasing the sensitivity, even when motif instances are no longer conserved across all species.

5.2.4 Comparison with other methodologies

We compared the performance of our methodology to that of two other motif detection algorithms: AlignACE [113], a Gibbs sampling algorithm designed to detect overrepresented motifs in a set of coregulated genes and PhyloCon, an algorithm that uses information on coexpression and orthology for motif detection [295].

AlignACE was applied to our test data sets without taking orthology information into account. Execution parameters are described in Methods. As displayed in Figure 5.2, only two out of the four motifs could be detected by AlignACE with very low recovery rates. In addition, the performance for those two motifs drops significantly when increasing the number of random genes (Figure 5.2 panel A). While the number of false positives seems to be independent of the noise, it is clearly high (varying between 16.9 and 18.3 compared to 0.5-0.8 for our algorithm (Table 5.2)). Regarding the quality of the motifs recovered, Fur and LexA, the sensitivity was always high, however, at the expense of a lower specificity. No values for sensitivity and specificity were calculated for PhoP and MetJ as these motifs were not detected. Our methodology clearly outperformed AlignACE in three aspects: the number of true motifs recovered, the quality of the motifs and the robustness against noise. This result is expected as AlignACE uses one source of information i.e. genes from one species, while our methodology depends on two sources i.e. single species genes and their orthologs.

In the case of PhyloCon, it also failed to recover MetJ and PhoP (figure 5.2). Although it was previously reported that it is possible to detect the MetJ motif in a dataset containing only MetJ intergenic regions [295] (confirmed in our lab), PhyloCon clearly failed to do so in the noisy context of our study. For LexA, PhyloCon recovered the motif at a rate comparable to that of our algorithm, and only marginally affected by the increasing noise in the data. PhyloCon performed well in retrieving the Fur motif, but with a recovery rate slightly sensitive to noise. Calculating the false positive discovery rate for PhyloCon illustrates that this algorithm is not sensitive to low signal to noise ratios (false positives ranging between 5.1 and 5.7) (Table 5.2). However, the absolute number of false positives is higher than that in the case of our methodology.

Regarding the quality of the motif models, the Fur motif model retrieved by PhyloCon seemingly outperforms our retrieved model both in sensitivity and specificity with hardly any effect for the noise in the data set. For the LexA motif, an optimal specificity was obtained but with mediocre sensitivity, indicating that the motif model, although frequently recovered, is often based on just two or three LexA instances.

It is important to emphasize that all the values obtained for the performance assessment of PhyloCon are biased to its favour in this case. One needs to note that in all the occasions when PhyloCon did not converge to a known motif, we retrieved the motif model from a cycle premature of convergence (for details see Methods: Benchmarking with other methodologies). Thus, slightly better sensitivities and recovery rates of

PhyloCon over our methodology need to be interpreted in the light of this fact.

Again, our methodology proved to be more robust to noise than similarly based algorithms as seen in its ability to recover all the true motif models in the data set with very few false positives.

5.2.5 Evaluation of the analysis flow on expression data

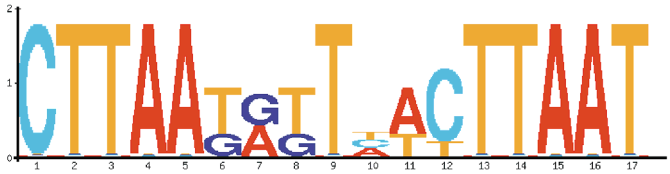
Besides the test data sets, we applied our methodology on two data sets derived from genome wide expression studies. A first data set consisted of a list of 47 differentially expressed genes between a constitutive and a null *pmrA* mutant of *S. typhimurium* [256]. The PmrAB two-component regulatory system is part of a multi-component feedback loop that acts as a major virulence regulator in *S. typhimurium*. The system itself is responsive to Fe^{3+} and mild acid and senses Mg^{2+} indirectly by communicating with the Mg^{2+} sensitive PhoPQ system via PmrD. Applying our analysis flow on the intergenic sequences of these coexpressed genes (BlockSampler, followed by BlockAligner and clustering) resulted in the detection of two potential motifs. The first motif which was derived from the motif instances of the known PmrA regulated genes *udg*, *yfbE*, *yibD* and *yjdB,* corresponded to the consensus of the biologically validated PmrA motif (consensus CTTAA-N_5-CTTAA) [2,160,303]. The motif logo is shown in figure 5.3.A. Screening the intergenic regions of the remaining differentially expressed genes from the input set with this motif model, resulted in the detection of six additional potential PmrA targets. Two of these six genes (*S. typhimurium* (*STM1269* and *ais*) did not contain an ortholog in *E. coli* or *Y. pestis,* and in the four remaining genes (*ybjG, ygiW, yijP* and *yegH*) the PmrA motif was not conserved in the orthologs of *E. coli* or *Y. pestis,* explaining the reason these motif instances could not be recovered initially unless by screening.

The (direct or indirect) regulation of eight of these reported ten genes by PmrAB is supported by evidence from previous studies. Seven of these genes – *ais (pmrG), yjdB (pmrC), yfbE (pmrH), ugd* (*udg, pmrE, pagA), yibD*, *ybjG* (*mig-13*) and *STM1269* (*aroQ*) – are known members of the PmrAB regulon that were confirmed in different experiments [2,97,160,256,257,303]. The last gene, *yijP (STM4118)*, was also discovered in an *in silico* screening of this same microarray data by Tamayo *et al.* [256].

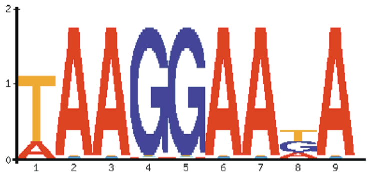
We also report two new potential PmrA targets, *ygiW* and *yegH*, that were not recovered by Tamayo *et al* [256]. This could be due to the more stringent threshold they used when defining differentially expressed genes, than ours (see above).

In addition to the well known PmrA motif, we predicted yet an uncharacterized motif (consensus (T/A)AAGGAAnA) (figure 5.3.B) which was based on conserved instances present in *tufA*, *yihT* and *STM2186*. Screening the remaining differentially expressed genes with the motif model corresponding to this motif resulted in the discovery of motif instances in two additional genes (*sopD* and *STM1472*). No similarity between this motif and any of the motifs in the existing databases could be found (RegulonDB 4.0 [227]).

A. PmrA motif model (PmrA testcase)



B. Unknown motif model (PmrA testcase)



C. FNR motif model (FNR testcase)

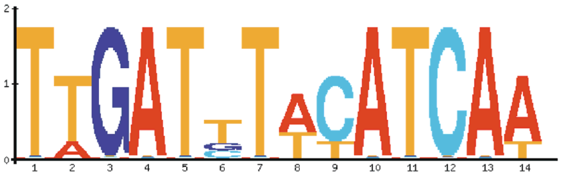


Figure 5.3: **Motif logos of the motif models resulting from the analysis of expression data.** A. PmrA motif model (PmrA testcase). B. Motif model of unknown regulatory motif (PmrA testcase). C. FNR motif model (FNR testcase).

The second data set was derived from the study of Salmon *et al.* [231]. *E. coli* and related species respond to oxygen depletion by switching to anaerobic respiration. This transition is mainly controlled by the global regulator FNR [98]. Salmon *et al.* [231] performed genome-wide expression profiling experiments to identify genes differentially expressed in response to oxygen and controlled by FNR. Applying our methodology to 83 differentially expressed genes derived from their study resulted in the detection of a motif model corresponding to the well known FNR motif (motif logo: figure 5.3.C). The model was based on conserved instances found in 4 known FNR regulated genes *narX*, *nirB, ndh* and *cydA* (RegulonDB 4.0). The instances present in three differentially expressed genes, marked as FNR regulated in RegulonDB (*cyoA, dcuC* and *frdA*) did not contribute to our FNR motif model. It seems like the degree of conservation of the motif among the respective orthologs of these genes was too low to be detected by our analysis flow. However, these genes could be retrieved in the screening step.

5.3 Discussion

If we want to be able to get a complete view on the transcriptional network of an organism, revealing all regulatory motifs remains one of the main challenges [272]. Despite the fact that several motif detection algorithms have been developed and optimized, detection of regulatory motifs with low overrepresentation is still difficult using the common motif detection tools. Therefore we developed a tool that is able to retrieve regulatory motifs planted in a noisy environment i.e. only a very small subset of submitted intergenic regions contains the motif. In a first step of our method, large conserved regulatory regions (i.e. blocks) are identified in a set of orthologous intergenic regions (BlockSampler). After aligning all these conserved blocks with each other, blocks containing a shared regulatory motif are clustered. In a final step, screening of all input intergenic regions permits for the detection of those target genes where the motif was not conserved in its orthologs (if existing).

To test its reliability, we applied our methodology on a “golden standard” consisting of a set of randomly selected, non-related genes among which known targets of the Fur, MetJ, LexA and PhoP regulons were hidden. Although these known targets contained previously described motif instances for each of these regulators, their statistical overrepresentation in this test set was low (18% going down to 7% for Fur and LexA, 11% down to 4% for MetJ and PhoP (dyad motif)). When applying our methodology we found that for 3 out of the 4 hidden motifs the recovery rate was above 90% even in the presence of a high amount of noisy genes. For the Fur motif it was slightly lower (50% on average) because of the lower degree of conservation of the motif model (consensus score of 1.02 for the Fur motif versus 1.21, 1.27 and 1.26 for the MetJ, PhoP and LexA motif respectively). Notwithstanding this high recovery rate, the number of false positive motifs (i.e. motifs not corresponding to any of the motifs hidden in the data set) remained low (an average of 0.50 false positives per run when 10 random genes are added versus 0.80 when 50 random genes are added). The achieved quality of the retrieved motif models, reflected by the motif model sensitivity (the number of true instances contributing to the motif model) and specificity (the number of false positive instances degrading the motif model) depends on the characteristics of the motifs; well conserved, non-dyad motifs such as the LexA and MetJ motifs are seemingly easy to retrieve and have a high quality; dyad motifs (PhoP and Fur), although having high specificity (above 80%), their sensitivity was lower than in the case of non-dyad motifs, especially for the PhoP dyad. In the case of the PhoP motif, the low sensitivity is due to the fact that it also occurs as a single half site. The score of the alignment of the single half site with the dyad falls below the thresholds we used and therefore these single sites could not be recovered by our methodology.

To assess the influence of taking into account orthology information in addition to information from coexpression, we applied our methodology to a combined data set of orthologous and coexpressed genes and compared its results to those obtained by applying AlignACE [113], which is a motif detection tool based on Gibbs sampling, to the coexpression data only. Our methodology clearly outperformed AlignACE in recovering the true motifs in the data set. In addition, the performance of AlignACE was seriously influenced by increasing noise in the data, in contrast to what is observed with our methodology. These results indicate that by the incorporation of orthology information, the retrieval of motifs with weak overrepresentation (ranging between 11% and 4%) becomes possible.

In addition to the above mentioned advantage of using the combined sources of information of coexpression and orthology, an extra value can be deduced. When using phylogenetic footprinting alone on orthologous gene sets of closely related species sharing a large part of the regulatory mechanism, long conserved blocks are detected rather than small distinct motifs [160]. Because this complicates the delineation of specific motifs, the sequential use of BlockSampler (utilizing orthology information) and BlockAligner (utilizing coexpression information) allows for an improved delineation of individual regulatory motifs from those long conserved regions across orthologs.

We also compared our methodology to PhyloCon [295], as it has a similar strategy as ours. PhyloCon identifies conserved regulatory regions (profiles, called blocks in our study) in sets of orthologs based on the Wconsensus program [110]. Subsequently, these profiles are merged between different sets of orthologs using a greedy algorithm, in order to detect a common regulatory motif.

PhyloCon was originally developed to detect motifs in a set of coexpressed genes, containing only one motif. The test sets used in the original paper usually contain a single motif per data set. As was noted by the authors, PhyloCon would perform better on less conserved motifs. Indeed, our results show that PhyloCon had a better performance in retrieving the more degenerated Fur motif than the well conserved LexA, MetJ and PhoP motifs. Despite this, the overall recovery rate was lower than that of our methodology in noisy data sets containing more than one motif. The MetJ and PhoP motifs were not recovered at all. Although the authors stated that their methodology can retrieve regulatory regions present in only a small set of genes (e.g. MetJ), in our experience PhyloCon fails to retrieve such motifs in a noisy context. In addition, as PhyloCon proceeds cycle per cycle, where in each cycle a new motif instance is added to the motif model, the authors suggest that the motif model building process be stopped before convergence in order to detect more motifs in a single dataset. However, they do not provide a clear stop criteria or heuristics to retrieve the optimal model. In order not to bias our results towards a sub-optimal stop criterion, we checked at each cycle whether a motif model was found that matched our test data set. If found during at least one cycle, the motif was considered "recovered". Note that this approach is feasible when one knows which motif to look for, but not in the case when a novel motif is searched for. This approach led to the bias of the assessment results to the favor of PhyloCon, as seen in the higher recovery rate, sensitivity and specificity acquired for PhyloCon over our method.

As a final proof of concept, we also applied our methodology to two gene sets derived from genome wide expression profiling experiments. Firstly, out of 47 differentially expressed genes between constitutive and null *pmrA* mutants of *S. typhimurium* [256], we could retrieve the expected PmrA motif model [2,160,303], and predict the presence of two new putative PmrA targets, *ygiW* and *yegH*. We also predicted a yet uncharacterized motif of the consensus (T/A)AAGGAAnA that was based on conserved instances in the genes *tufA, yihT* and *STM2186*. In a second test, using a list of 83 FNR regulated genes [231], we could retrieve the FNR motif model based on 4 genes containing a clear FNR regulatory motif.

5.4 Conclusion

Conclusively, we have developed an approach that can reliably identify multiple regulatory motifs lacking a high degree of overrepresentation in a set of coexpressed genes (motifs belonging to sparsely connected hubs in the regulatory network) by exploiting the advantages of using both coexpression and phylogenetic information. Through comparing our methodology to two other motif detection programs, we show the robustness of our implementation. As a proof of concept, analysis of genome wide expression data with our methodology successfully retrieves the present regulatory motifs.

5.5 Methods

5.5.1 Input data sets

The selection of intergenic regions and the construction of species specific background models, relied on modules implemented in INCLUSive [267]. Intergenic regions are defined as the non-coding parts between two coding sequences. The regions used in this study were derived from the following genomes: *Escherichia coli* K12 [GenBank: NC_000913], *Escherichia coli* plasmid R721 [GenBank: NC_002525], *Escherichia coli* O157:H7 EDL933 [GenBank: NC_002655], *Escherichia coli* O157:H7 [GenBank: NC_002695], *Escherichia coli* CFT073 [GenBank: NC_004431], *Escherichia coli* O157:H7 plasmid pO157 [GenBank: NC_002128], *Escherichia coli* plasmid pB171 [GenBank: NC_002142], *Shigella flexneri* 2a str. 301 [GenBank: NC_004337], *Shigella flexneri* virulence plasmid pWR501 [GenBank: NC_002698], *Salmonella typhimurium* LT2 [GenBank: NC_003197], *Salmonella typhimurium* LT2 plasmid pSLT [GenBank: NC_003277], *Salmonella Typhi* CT18 [GenBank: NC_003198], *Salmonella Typhi* CT18 plasmid pHCM1 [GenBank: NC_003384], *Salmonella Typhi* CT18 plasmid pHCM2 [GenBank: NC_003385], *Yersinia pestis* CO92 plasmid pPCP1 [GenBank: NC_003132], *Yersinia pestis* CO92 [GenBank: NC_003143], *Yersinia pestis* CO92 plasmid pCD1 [GenBank: NC_003131], *Yersinia pestis* KIM [GenBank: NC_004088], *Yersinia pestis* CO92 plasmid pMT1 [GenBank: NC_003134].

Close homologs (either orthologs or paralogs) were identified as described in Marchal *et al.* [160].

For benchmarking, different test data were compiled consisting of a core set of genes with known motifs (PhoP, LexA, Fur and MetJ) supplemented with sets of random genes varying in number and composition. A core set of genes with known binding sites was selected based on the RegulonDB database 4.0 [227] for respectively LexA, Fur, and MetJ. Genes containing known PhoP motifs were selected from Monsieurs *et al.* [178]. The composition of the core gene set is heterogeneous in terms of the number of instances and conservation of each motif. An overview of this composition is given in table 5.1. For MetJ (3 genes) and LexA (5 genes), the motif instances where conserved in all the species used in this study. For 3 out of the 5 Fur regulated genes and 1 out of the PhoP regulated genes, a motif instance was not present in *Y. pestis*. Starting from this core set, different test data were generated by gradually adding an increasing number of random genes. To this end, genes having an intergenic sequence larger than 100 nucleotides and a sufficient number of orthologs in the organisms of interest were selected randomly from the *Salmonella* genome. By sampling ten times respectively 10, 20, 30, 40 and 50 random genes and adding these to the core gene set, a total of 50 test sets was created.

For the construction of the PmrA data set, the data corresponding to an experiment described by Tamayo *et al.* [256] were downloaded from the Stanford Microarray Database [17]. In their analysis, Tamayo *et al.* compared the mRNA expression between PmrA-constitutive and PmrA-null strains at two different time points (early- and mid-logarithmic phase of growth). As input data set, we selected two times 40 genes out of this microarray results that were most up- or down-regulated respectively in both conditions. Notice that we used a less stringent threshold than Tamayo *et al.* [256] who only selected 41 genes that exhibited a minimal fold change of 2 at both time points. After elimination of those genes with in an intergenic region smaller than 50 nucleotides, only 47 out of these 80 genes were retained. For the application of our methodology on data sets derived from genome wide expression studies, we relied on the data previously published in Salmon *et al.* [231] to build the FNR data set. All 125 genes assigned by Salmon *et al.* to a cluster affected by a mutation of *fnr* (i.e. 6 out of the 8 different clusters), were combined and used as input data for our methodology. From these 125 genes, 83 genes with an intergenic region longer than 50 nucleotides were retained.

5.5.2 Analysis flow

5.5.2.1 Step 1: Detecting conserved blocks with BlockSampler.

BlockSampler 3.1 is based on the original Gibbs sampling algorithm of MotifSampler. Briefly, the Gibbs sampling procedure starts by searching for a motif shared by at least 2 sequences and having one occurrence in the reference sequence (the sequence of interest, see below). After convergence, short motif seeds are identified. The identification of these seeds is predicated on the log likelihood score [268]. This score depends on the degree of conservation of the motif and the number of instances detected in the species. Thus, the more species in which the motif is conserved and the higher its degree of conservation, the higher is its corresponding score. High scoring seeds are subsequently extended using a simple protocol: if the consensus score over a 5-nt region adjacent (upstream or downstream) to the current motif seed exceeds a given threshold the motif is extended with one nucleotide (in that direction). Detected conserved intergenic regions (i.e. blocks) of variable length are eventually reported. To select the most promising hits from the output of BlockSampler, we designed a score that is independent of block sequence length, but increases with the degree of conservation of the motifs. This normalized consensus score is appropriate because short motifs have a higher chance of resulting in a high consensus score. Normalization was done by recalculating the consensus scoring according to the following formula: Cs_{ad} = (L/L+E) Cs, where Cs_{ad} is the normalized consensus score, L is the length of the conserved block, E is an empirical factor (set to 6) and Cs the consensus score. Different empirical factors were tested on different data sets, and 6 appeared to give the best balance between block sequence length and conservation. Depending on the interest of a particular study, the empirical factor can be enlarged to favor larger blocks. Blocks are then ranked according to this normalized consensus score

BlockSampler requires six user-defined parameters: 1) the definition of a reference sequence: the reference sequence is the sequence in which the presence of the conserved block is required (in our case, it was set to be *Salmonella typhimurium*); 2) the number of runs: as BlockSampler is based on Gibbs sampling the algorithm should be repeatedly applied on the same input set (set at 100 runs); 3) strand: only the plus strand is searched; 4) prior: set at default value of 0.2; 5) threshold of the consensus score: only blocks exceeding a consensus score of 1.3 are retained; 6) minimal motif length: minimal width of the block is set to 8.

As is the case with other Gibbs sampling based motif detection procedures, the same block can be detected several times over the different runs of the algorithm. To compile a list of non redundant blocks, blocks overlapping for more than 75% were grouped. From each set of overlapping blocks, the one displaying the highest (normalized) consensus score was chosen as representative and retained for further analysis. Each block is represented by a motif model, in the form of a frequency matrix.

5.5.2.2 Step 2: Aligning conserved blocks using BlockAligner

The algorithm (version 3.1) uses a local ungapped alignment strategy based on dynamic programming to mutually compare conserved blocks represented by their respective motif models (frequency matrices M1 and M2). The following additive scoring scheme is used: the total alignment score of two motif models is the sum of the individual column scores. A column score (S) is defined as the distance between two aligned columns of the frequency matrices. As a measure, the Kullback-Leibler distance between two probability distributions is used, since the columns of a motif model can be considered to be the parameters of multinomial distributions. To make the scoring scheme compatible with dynamic programming, matching columns should score positively and non-matching columns negatively. Therefore the minimal match value of the Kullback-Leibler distance T was introduced. T is a user defined parameter that determines the stringency of the alignment. Columns with a score below T receive a negative score, while columns with a score above T receive a positive score. As a result, the following score of the alignment of two columns, *i* and *j* of two conserved blocks (represented by the frequency matrices *M1* and *M2*) can be calculated:

$$S(i,j) = T - \sum_{b=A}^{T} M1_{b,i} * \log \frac{M1_{b,i}}{M2_{b,j}} + M2_{b,j} * \log \frac{M2_{b,j}}{M1_{b,i}}$$

S(i,j) will be equal to *T* if column *i* and *j* of motif models *M1* and *M2* respectively, are exactly the same.

As a biological motif is often "gapped" i.e. consisting of conserved nucleotides intersected by some non-conserved nucleotides, we introduced a small non-match penalty. Remark that this is different from a "gap score" in alignment algorithms [242,270], as insertions and deletions are not explicitly modelled (we use a local ungapped alignment).

This leads to the following scheme for the alignment matrix:

A(i,j) = max(0,A(i-1,j-1) + S(i,j),A(i-1,j-1) - NonMatchScore) for i>1 && j>1

A(i,1) = max(S(i,1),0)

A(1,j) = max(S(i,1),0)

In our setup, BlockAligner was used with the following parameter set: T value = 0.40; minimal length of the reported common motif = 6 nucleotides.

To assess the significance of the results, the alignment procedure was repeated 100 times on the same motif models (*M1* and *M2*) after randomly shuffling their columns. The distribution of the scores obtained by aligning these randomly shuffled motif models was used to estimate the parameters of an extreme value distribution. This background distribution allowed obtaining a p-value for the genuine alignment (i.e. assessing the probability of obtaining by coincidence the score observed when aligning the unshuffled blocks). Blocks with a p-value below 0.001 were considered significant.

5.5.2.3 Step 3: Clustering conserved blocks and delineating regulatory motifs.

Conserved blocks shared by different orthologous gene sets were grouped using a graph based clustering algorithm TribeMCL version 02.277 [67]. Nodes of the graph represent conserved blocks and edges represent the quality of the alignment between these conserved blocks. We used the $-\log_{10}$

of the p-value of the pairwise alignments (see previous step) as weight measure for the edges. To prevent inflating spurious relations between blocks based on low scoring alignment scores, only alignments with a p-value lower than 0.001 were taken into account.

Based on the pairwise alignment scores between the conserved blocks grouped within the same cluster, a multiple alignment is created, which is subsequently converted into a frequency matrix. Such matrix representing the multiple alignment of conserved blocks in a cluster can be seen as a model of the average motif that is conserved in the intergenic region of several sets of orthologous genes and to which all orthologous genes from the original orthologous sets of the cluster contribute. This multiple alignment was converted in a frequency matrix. From this frequency matrix, the minimal regulatory motif was defined as the 1) the region conserved in at least three reference genes, 2) with a minimal length of 6 nucleotides and 3) with a consensus score higher than or equal to 1.10. This minimal motif is extended with additional motif positions in both directions until the consensus score drops below the threshold of 1.10.

To construct a motif model specific for the reference species, the multiple alignment is based only on the motif instances contributing to the blocks that originate from the reference species.

5.5.2.4 Step 4: Genome wide screening for additional targets

Screening of promoter regions with the obtained motif models is performed using MotifLocator 3.0 [42,160]. The cut-off value for a screening was derived based on the lowest MotifLocator score of known target genes of the corresponding regulatory protein.

5.5.3 Benchmarking with other methodologies

5.5.3.1 Running AlignACE.

AlignACE can be obtained at the AlignACE website. AlignACE is a Gibbs Sampling algorithm for detecting regulatory motifs that are overrepresented in the promoter regions of a set of potentially coregulated genes. Therefore, the test sets used for AlignACE only contained information from coexpressed genes i.e. the intergenic regions of the reference genes from *S. typhimurium* sharing a similar motif. Orthologous information was not explicitly used. We run AlignACE release 4.0 with default parameter setting except that we give the GC content of species from which the intergenic regions are used as input (0.52 for *S. typhimurium*). AlignACE returns series of motif models that are overrepresented in the input promoter regions.

5.5.3.2 Running PhyloCon

PhyloCon uses the same information sources as our methodology [295]. This algorithm (release 3a) also starts from a set of genes that are potentially co-regulated (e.g. derived from microarray data) and uses orthologous information to detect novel regulatory motifs. This two-step procedure starts with aligning orthologous intergenic sequences and creating position specific scoring matrices (called profiles) based on the Wconsensus program [110]. Then PhyloCon compares these profiles generated from different genes and identifies the common regions in these profiles using a greedy approach. Because PhyloCon is optimised to use the same sources of information (both coexpression and orthology) as our methodology, the same test data sets could be used. When running PhyloCon (downloaded from the PhyloCon website), the number of standard deviations was set between 0.5 and 2, but this only marginally affected the results. The way PhyloCon works is that the motif model grows cycle per cycle. In each cycle a new motif instance is added to the motif model. In the original article no clear stop criteria or heuristics were provided, so it is difficult to decide at which cycle an optimal motif model is detected. By screening all different cycles of the PhyloCon runs, we looked for runs of which at least one cycle shows a match to one of the known motif models. In a single run, more than one cycle can contain a motif of interest. For each particular motif the cycle that shows the highest combined sensitivity specificity score for the corresponding motif model was used for the calculation of the sensitivity, specificity and false positives.

5.5.3.3 Test Run

Each test run consisted of applying one of the specified algorithms to 10 test sets of similar composition (i.e. same number of random genes is added). The recovery rate of detecting a particular motif was defined as the percentage of test sets in which this motif could be recovered. For instance, if the LexA motif model was found in 6 of the 10 test sets each of which contained LexA regulated genes, the performance was defined to be 60%.

The exact content of a test run depends on the specificities of the algorithm applied. One test run of the methodology presented in this study was defined as applying the complete procedure on the test set (100 runs of BlockSampler, BlockAligner, Clustering, motif delineation). A single test run of PhyloCon or AlignACE was defined as running once the algorithm on the test set (see running PhyloCon and Running AlignACE).

5.5.3.4 Performance evaluation

To evaluate the performance of the different motif detection algorithms, we reported 1) the average recovery rate and 2) the average number of false positives. The average recovery rate reports how many times a motif model corresponding to a motif, known to be present in the test sets, was found on average in the test runs on different test sets of the same composition. 2) the average number of false positives reports how many times a motif model not corresponding to any motif known to be present in the test sets, was found on average in the test runs on different data sets of the same composition. Motifs known to be present in the data sets were represented by a benchmark of curated motif models extracted from the RegulonDB database 4.0 [227]. A detected motif model was considered identical to a benchmark motif model when the Kullback-Leiber distance (as implemented in MotifComparison) between the two motif models was lower than 0.65 (default parameter); otherwise it was considered as a false positive. The calculation of the number of false positives for AlignACE and PhyloCon was done as follows: all motif models that were returned from the AlignACE algorithm, were aligned with the benchmark motif models. If such a motif model did not show similarity to any of the four motifs known to be present in the test sets, this motif model was regarded a false positive. For PhyloCon, no clear stop criterion is described. For that reason, we only took into account those cycles for which a hit with a benchmark motif model was detected. All motifs in these cycles that did not show a match with a benchmark motif model were treated as a false positive.

5.5.3.5 Motif Quality

To represent the quality of the obtained motif models, we calculated motif model sensitivity and specificity using the following definitions:

$$SENS = \frac{TP}{TP + FN} \text{ x } 100\%$$

where TP is the number of true positives motif instances (i.e. motif instances known to be present in the data set that contributed to the detected motif model) and FN is the number of false negatives (i.e. motif instances known to be present in the data set that did not contribute to the detected motif model).

For definition of the specificity (SPEC), we used:

$$SPEC = \frac{TP}{TP + FP} \text{ x } 100\%$$

where TP is defined as stated above and FP is the number of false positives (i.e. motif instances not corresponding to any of the known motifs present in the data set contributing to the detected motif model). Remark that the definition of FP is dependent on the accuracy and completeness of the existing annotation in the motif databases.

Chapter 7

Detection of double glycine leader sequences

7.1 Introduction

Different from all previous research described in the thesis, this chapter focuses on a biological problem situated at the protein level instead of at the nucleotide level. In addition, the methodology proposed here to solve a specific biological question is based on algorithms developed in previous chapters, or algorithms developed by other research groups. As the aim of the research was to study the evolutionary distribution of a protein transport system among all fully sequenced bacteria, different comparative genomics methodologies developed in previous chapters could be recycled. For the protein-specific problems, we needed to integrate new algorithms in our methodology. This work has been done in close collaboration with the Centre of Microbial and Plant Genetics (Prof. J. Michiels, Dr. G. Dirix) and resulted in two publications [55,56].

Protein transport in all systems is accomplished by a single underlying mechanism: each polypeptide destined for transport to the extracellular environment contains an amino acid sequence known as signal or leader peptide that identifies the polypeptide to the appropriate transporting system. Frequently, the signal sequence is cleaved from the parent polypeptide during the transport process. An interesting leader peptide is the double glycine leader sequences as it plays a key role in many peptide secretion systems involved in quorum sensing and bacteriocin production in Gram-positive bacteria.

Quorum sensing, which involves the production, release, and subsequent detection of chemical signalling molecules called autoinducers, allows bacteria to regulate gene expression in response to changes in cell-population density. As a population of bacteria grows, the extracellular concentration of autoinducer increases. When a threshold is reached, the group responds with a population-wide alteration in gene expression.

Processes controlled by quorum sensing are usually ones that are unproductive when undertaken by an individual bacterium but become effective when undertaken by the group. However, the signalling molecules differ among Gram-negative and Gram-positive bacteria. A broad variety of these signalling molecules has been identified in the past 20 years which can roughly be divided in (1) acylhomoserinelactones in Gram-negatives [300]; (2) processed oligopeptides in Gram-positives [253]; and (3) AI-2 in both Gram-negatives and Gram-positives [307,308].

The quorum sensing system in Gram-positive bacteria using oligopeptide signal molecules, induces processes like virulence in *Staphylococcus aureus* [190] and *Enterococcus faecalis* [213], genetic competence in *Streptococcus pneumoniae* [41] and *Bacillus subtilis* [273], and the production of antimicrobial peptides (AMP) in many lactic acid bacteria [132]. These oligopeptide signal molecules consist of small peptides that are processed and transported by an ATP-binding cassette (ABC) transporter. The precursor peptides of the autoinducers involved in the competence development of *S. pneumoniae* (called the competence stimulating peptides), and in the production of AMP (i.e. bacteriocins) in lactic acid bacteria, contain a GG-motif leader sequence that is removed by their dedicated ABC-transporter concomitant with export [106,207]. Comparable to these autoinducer peptides, a similar mechanism exists to export bacteriocins, of which the prepeptide also contains a GG-motif leader sequence that is recognized and removed by an ABC transporter [188,226].

The consensus sequence of this GG-motif was proposed as LSX_2ELX_2IXGG where X can be any amino acid [106]. It was shown, as suggested above, that the presence of a GG-motif containing peptide is clearly related to the presence of a dedicated ABC-transporter, which contains a specific N-terminal extension of about 150 amino acids that is not found in other ABC-transporters [105] and that is characterized as the Peptidase C39 domain [19]. In this N-terminal proteolytic domain, two conserved motifs can be distinguished, called the cysteine and the histidine motifs (C/H motifs). These motifs are responsible for recognition and cleavage of the GG-motif containing peptides [106,282].

Most of these studies on the GG-motif containing peptides and their cognate ABC-transporter is performed in Gram-positive bacteria. However, Michiels *et al.* [171] already identified a number ABC-transporters in Gram-negative species. The presence of the specialized exporters suggests that peptides containing the double-glycine leader sequence should also be found in Gram-negative bacteria. Indeed, several peptides containing the GG-motif were already detected in the study of Michiels et al. [171]. To screen for additional putative GG-motif containing peptides at a genome-wide level in Gram-positive and Gram-negative bacteria, an in silico strategy was designed. Using a curated training set, a motif model of the leader peptide was built and used to screen 120 fully sequenced bacterial genomes. The screening methodology was applied at the nucleotide level as probably many small peptide genes have not been annotated and may be absent in the existing databases.

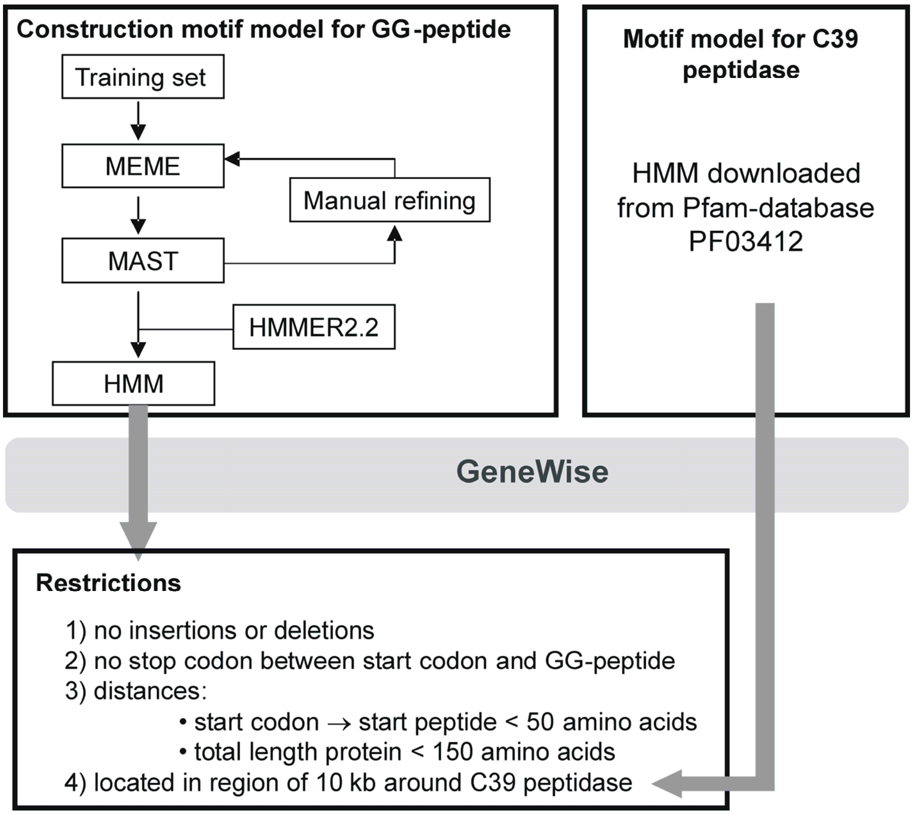

Construction motif model for GG-peptide
Training set
MEME
Manual refining
MAST
HMMER2.2
HMM
Motif model for C39 peptidase
HMM downloaded from Pfam-database PF03412
GeneWise
Restrictions
1) no insertions or deletions
2) no stop codon between start codon and GG-peptide
3) distances:
• start codon → start peptide < 50 amino acids
• total length protein < 150 amino acids
4) located in region of 10 kb around C39 peptidase

Figure 7.1: **Schematic overview of the screening strategy used.**

7.2 Results

7.2.1 Overview

As mentioned above, this research aims at performing a genome-wide in silico screening for peptides containing the double glycine leader sequence and their cognate transporters. So far, searches for GG-leader sequences were performed at the amino acid level. In case of a search for small peptides, the major drawback of using protein databases is the dependence on the correct annotation of DNA sequences. Small open reading frames are not always annotated and consequently, the corresponding peptides will not be present in the protein databases. Therefore, a DNA based search strategy is needed.

The Wise2 program (http://www.ebi.ac.uk/Wise2) translates DNA sequences in the six reading frames and compares the translations with a Hidden Markov Model (HMM) [24]. Wise2 was used to screen all fully sequenced bacterial genomes present in the NCBI genome database (March 2003) for GG-motifs and for Peptidase C39 domains. Beside the chromosomes, also plasmids were taken into the search, but only those plasmids that belong to a fully sequenced strain. A list of all screened genomes and plasmids is given in table 7.1. For the Peptidase C39 domain search, the HMM describing this protein family domain was obtained from the Pfam database (www.sanger.ac.uk/Software/Pfam; accession number PF03412) [19]. For the peptide search, two HMMs were built, describing the GG-motif based on peptides from either Gram-negative or Gram-positive bacteria. A schematic overview of the screening strategy is shown in Fig. 7.1.

***Table 7.1:* List of genomes and plasmids screened for the presence of Peptidase C39 domains and GG-motifs.** Legend: Accnr. Chr.: Accession number of the chromosome.

	Organism	Accnr. Chr.	Accnr. Plasmids
1	*Agrobacterium tumefaciens str. C58, (Cereon)*	NC_003062, NC_003063	NC_003064, NC_003065
2	*Agrobacterium tumefaciens str. C58, (U. Washington)*	NC_003304, NC_003305	NC_003306, NC_003308
3	*Aquifex aeolicus*	NC_000918	NC_001880
4	*Bacillus anthracis str. A2012*	NC_003995	NC_003980, NC_003981
5	*Bacillus anthracis str. Ames*	NC_003997	
6	*Bacillus cereus*	NC_004722	NC_004721
7	*Bacillus halodurans*	NC_002570	
8	*Bacillus subtilis*	NC_000964	

9	*Bacteroides thetaiotaomicron*	NC_004663	
10	*Bifidobacterium longum*	NC_004307	NC_004943
11	*Bordetella bronchiseptica*	NC_002927	
12	*Bordetella parapertussis*	NC_002928	
13	*Bordetella pertussis*	NC_002929	
14	*Borrelia burgdorferi*	NC_001318	NC_000948, NC_000950, NC_000952, NC_000954, NC_000956, NC_001849, NC_001851, NC_001853, NC_001855, NC_001857, NC_001904, NC_000949, NC_000951, NC_000953, NC_000955, NC_000957, NC_001850, NC_001852, NC_001854, NC_001856, NC_001903
15	*Bradyrhizobium japonicum*	NC_004463	
16	*Brucella melitensis*	NC_003317, NC_003318	
17	*Brucella suis*	NC_004310, NC_004311	
18	*Buchnera aphidicola str. APS*	NC_002528	NC_002252, NC_002253
19	*Buchnera aphidicola str. Bp*	NC_004545	
20	*Buchnera aphidicola str. Sg*	NC_004061	
21	*Campylobacter jejuni*	NC_002163	
22	*Candidatus Blochmannia floridanus*	NC_005061	
23	*Caulobacter crescentus*	NC_002696	
24	*Chlamydia muridarum*	NC_002620	NC_002182
25	*Chlamydia trachomatis*	NC_000117	
26	*Chlamydophila caviae*	NC_003361	NC_004720
27	*Chlamydophila pneumoniae AR39*	NC_002179	
28	*Chlamydophila pneumoniae CWL029*	NC_000922	
29	*Chlamydophila pneumoniae J138*	NC_002491	
30	*Chlamydophila pneumoniae TW-183*	NC_005043	
31	*Chlorobium tepidum*	NC_002932	
32	*Chromobacterium violaceum*	NC_005085	
33	*Clostridium acetobutylicum*	NC_003030	NC_001988
34	*Clostridium perfringens*	NC_003366	
35	*Clostridium tetani*	NC_004557	
36	*Corynebacterium efficiens*	NC_004369	
37	*Corynebacterium glutamicum*	NC_003450	
38	*Coxiella burnetii*	NC_002971	NC_004704
39	*Deinococcus radiodurans*	NC_001263, NC_001264	NC_000958
40	*Enterococcus faecalis*	NC_004668, NC_004670	NC_004671
41	*Escherichia coli CFT073*	NC_004431	
42	*Escherichia coli K12*	NC_000913	
43	*Escherichia coli O157:H7*	NC_002695	NC_002127, NC_002128
44	*Escherichia coli O157:H7 EDL933*	NC_002655	
45	*Fusobacterium nucleatum*	NC_003454	
46	*Haemophilus ducreyi*	NC_002940	
47	*Haemophilus influenzae*	NC_000907	
48	*Helicobacter hepaticus*	NC_004917	
49	*Helicobacter pylori 26695*	NC_000915	
50	*Helicobacter pylori J99*	NC_000921	
51	*Lactobacillus plantarum*	NC_004567	
52	*Lactococcus lactis*	NC_002662	NC_002502, NC_004955, NC_004922, NC_004966

53	*Leptospira interrogans*	NC_004342, NC_004343	
54	*Listeria innocua*	NC_003212	NC_003383
55	*Listeria monocytogenes*	NC_003210	
56	*Mesorhizobium loti*	NC_002678	NC_002679, NC_002682
57	*Mycobacterium bovis*	NC_002945	
58	*Mycobacterium leprae*	NC_002677	
59	*Mycobacterium tuberculosis CDC1551*	NC_002755	
60	*Mycobacterium tuberculosis H37Rv*	NC_000962	
61	*Mycoplasma gallisepticum*	NC_004829	
62	*Mycoplasma genitalium*	NC_000908	
63	*Mycoplasma penetrans*	NC_004432	
64	*Mycoplasma pneumoniae*	NC_000912	
65	*Mycoplasma pulmonis*	NC_002771	
66	*Neisseria meningitidis MC58*	NC_003112	
67	*Neisseria meningitidis Z2491*	NC_003116	
68	*Nitrosomonas europaea*	NC_004757	
69	*Nostoc sp.*	NC_003272	NC_003240, NC_003267, NC_003273, NC_003241, NC_003270, NC_003276
70	*Oceanobacillus iheyensis*	NC_004193	
71	*Pasteurella multocida*	NC_002663	NC_001774, NC_004772, NC_004771
72	*Pirellula sp.*	NC_005027	
73	*Porphyromonas gingivalis*	NC_002950	
74	*Prochlorococcus marinus str. MIT9313*	NC_005071	
75	*Prochlorococcus marinus str. CCMP1375*	NC_005042	
76	*Prochlorococcus marinus str. CCMP1378*	NC_005072	
77	*Pseudomonas aeruginosa*	NC_002516	
78	*Pseudomonas putida*	NC_002947	
79	*Pseudomonas syringae*	NC_004578	NC_004632, NC_004633
80	*Ralstonia solanacearum*	NC_003295	NC_003296, NC_001399
81	*Rickettsia conorii*	NC_003103	
82	*Rickettsia prowazekii*	NC_000963	
83	*Salmonella enterica sv. Typhi*	NC_003198	NC_003384, NC_003385
84	*Salmonella enterica sv. Typhi Ty2*	NC_004631	
85	*Salmonella typhimurium*	NC_003197	NC_003277
86	*Shewanella oneidensis*	NC_004347	NC_004349
87	*Shigella flexneri str. 2457T*	NC_004741	
88	*Shigella flexneri str. 301*	NC_004337	
89	*Sinorhizobium meliloti*	NC_003047	NC_003037, NC_003078, NC_004965
90	*Staphylococcus aureus Mu50*	NC_002758	NC_002774
91	*Staphylococcus aureus MW2*	NC_003923	
92	*Staphylococcus aureus N315*	NC_002745	NC_003140
93	*Staphylococcus epidermidis*	NC_004461	NC_005003, NC_005005, NC_005007,NC_00504, NC_00506, NC_00508
94	*Streptococcus agalactiae 2603V/R*	NC_004116	
95	*Streptococcus agalactiae NEM316*	NC_004368	
96	*Streptococcus mutans*	NC_004350	
97	*Streptococcus pneumoniae R6*	NC_003098	

98	*Streptococcus pneumoniae TIGR4*	NC_003028	
99	*Streptococcus pyogenes M1 GAS*	NC_002737	
100	*Streptococcus pyogenes MGAS315*	NC_004070	
101	*Streptococcus pyogenes MGAS8232*	NC_003485	
102	*Streptococcus pyogenes SSI-1*	NC_004606	
103	*Streptomyces avermitilis*	NC_003155	NC_004719
104	*Streptomyces coelicolor*	NC_003888	NC_003903, NC_003904
105	*Synechococcus sp.*	NC_005070	
106	*Synechocystis sp.*	NC_000911	NC_004967
107	*Thermoanaerobacter tengcongensis*	NC_003869	
108	*Thermosynechococcus elongates*	NC_004113	
109	*Thermotoga maritima*	NC_000853	
110	*Treponema pallidum*	NC_000919	
111	*Tropheryma whipplei TW08/27*	NC_004551	
112	*Tropheryma whipplei str.Twist*	NC_004572	
113	*Ureaplasma urealyticum*	NC_002162	
114	*Vibrio cholerae*	NC_002505, NC_002506	NC_004982
115	*Vibrio parahaemolyticus*	NC_004603, NC_004605	
116	*Vibrio vulnificus*	NC_004459, NC_004460	
117	*Wigglesworthia glossinidia*	NC_004344	
118	*Wolinella succinogenes*	NC_005090	
119	*Xanthomonas axonopodis*	NC_003919	NC_003921, NC_003922
120	*Xanthomonas campestris*	NC_003902	
121	*Xylella fastidiosa 9a5c*	NC_002488	NC_002489, NC_002490
122	*Xylella fastidiosa Temecula1*	NC_004556	NC_004554
123	*Yersinia pestis CO92*	NC_003143	NC_003131, NC_003134, NC_003132
124	*Yersinia pestis KIM*	NC_004088	NC_004835, NC_004837, NC_004839

7.2.2 Peptidase C39 domain search

By using the Wise2-software in combination with the HMM describing the Peptidase C39 family (PF03412), all fully sequenced bacterial genomes (Table 7.1) were screened. The search resulted in 78 candidate Peptidase C39 domain encoding loci, on which a blastx analysis was performed to retrieve the complete protein sequence (www.ncbi.nlm.nih.gov/Blast) [5]. For three hits no correct annotation was found. Therefore the putative protein sequences were derived from the corresponding genome sequences. Information about the distribution of the Peptidase C39 hits among the different screened genomes is given in Table 7.2.

Table 7.2: **Distribution of the Peptidase C39 hits amongst Gram-negative and Gram-positive bacteria.** Percentages are calculated with respect to the Gram-negative and Gram-positive strains or hits.

	Gram-	Gram+	Total
Number of strains screened	79 (64%)	45 (36%)	124
Number of hits	49 (63%)	29 (37%)	78
Number of strains with one or more hits	26 (57%)	20 (43%)	46
Number of strains with more than 1 hit	10 (67%)	5 (33%)	15

Although originally discovered in Gram-positive bacteria, the domain is clearly present in many Gram-negative bacteria. 26 out of 79 Gram-negative bacterial strains (33%) contained one or more Peptidase C39 domain, compared to 20 out of 45 (44%) for Gram-positive bacteria. The genomes of 10 out of the 26 Gram-negative bacteria containing a Peptidase C39 domain (38%), have more than one Peptidase C39 domain protein encoding gene, whereas for Gram-positive bacteria this is 5 out of 20 (25%). Interestingly, when we compare the genomic context by usage of the STRING database [293], most genes coding for a Peptidase C39 domain have one or more genes in close proximity that code for a membrane fusion protein or other proteins related to membrane transport, indicating that they are likely involved in protein/peptide secretion. A list of all Peptidase C39 domain hits that resulted from our search is given in table 7.3.

Table 7.3: **List of the Peptidase C39 domain containing proteins.** # Not annotated. [a] Consists of a three-letter abbreviation of the species followed by the class to which the protein belongs and a number to differentiate between multiple transporters within the same strain. [b] Accession number of the genome; / protein containing a truncated Peptidase C39 domain. [c] Protein containing a truncated ABC_membrane domain (PF:000664).

ID	Name[a]	Organism	Acc.Nr.[b]	Gene	L	Description
1	Bha_1	*Bacillus halodurans*	NC_002570	BH0451	724	Lantibiotic mersacidin transporter
2	Bsu_1	*Bacillus subtilis*	NC_000964	sunT	705	Lantibiotic sublancin 168 transporter
3	Bth_1	*Bacteroides thetaiotaomicron*	NC_004663	BT4288	741	ABC transporter
4	Bbr_1	*Bordetella bronchiseptica*	NC_002927	cyaB	735	Cyclolysin secretion ATP-binding protein
5	Bpa_1	*Bordetella parapertussis*	NC_002928	cyaB	735	Cyclolysin secretion ATP-binding protein
6	Bpe_1	*Bordetella pertussis*	NC_002929	cyaB	712	Cyclolysin secretion ATP-binding protein
7	Bja_1	*Bradyrhizobium japonicum*	NC_004463	bll6293	835	HlyB/MsbA family ABC transporter
8	Cje_1	*Campylobacter jejuni*	NC_002163	Cj0058	199	Putative periplasmic protein
9	Ccr_1	*Caulobacter crescentus*	NC_002696	CC0684	726	HlyB/MsbA family ABC transporter
10	Cvi_1	*Chromobacterium violaceum*	NC_005085	CV0068	727	Colicin V secretion ATP-binding

						protein
11	Cvi_2	*Chromobacterium violaceum*	NC_005085	CV4400	706	Colicin V secretion ABC transporter
12	Cvi_3	*Chromobacterium violaceum*	NC_005085	hlyB	709	Hemolysin B
13	Cvi_1	*Chromobacterium violaceum*	NC_005085	CV0686	234	Conserved hypothetical protein
14	Cac_1	*Clostridium acetobutylicum*	NC_001988	CAP0073	731	ABC transporter
15	Efa_1	*Enterococcus faecalis*	NC_004671	EFB0050	700	Toxin ABC transporter
16	Efa_1	*Enterococcus faecalis*	NC_004668	EF0528	270	Cytolysin B transporter, truncated
17	Eco_1	*Escherichia coli CFT073*	NC_004431	hlyB	707	Hemolysin B
18	Eco_2	*Escherichia coli CFT073*	NC_004431	mchF	704	Probable microcin H47 transporter
19	Ecp_1	*Escherichia coli O157:H7*	NC_002128	hlyB	706	Hemolysin B
20	Hhe_1	*Helicobacter hepaticus*	NC_004917	HH0187	228	Conserved hypothetical protein
21	Lpl_1	*Lactobacillus plantarum*	NC_004567	plnG	716	Bacteriocin ABC transporter
22	Lla_1	*Lactococcus lactis*	NC_002662	lcnC	715	Lactococcin A ABC transporter
23	Mga_1	*Mycoplasma gallisepticum*	NC_004829	sunT	667	Bacteriocin/lantibiotic ABC exporter
24	Mge_1	*Mycoplasma genitalium*	NC_000908	MG390	660	ABC transporter
25	Mpe_1	*Mycoplasma penetrans*	NC_004432	MYPE7470	691	ABC transporter
26	Mpn_1	*Mycoplasma pneumoniae*	NC_000912	lcnDR3	660	Hemolysin ABC exporter
27	Mpu_1	*Mycoplasma pulmonis*	NC_002771	MYPU_3760	680	ABC transporter
28	Nme_1	*Neisseria meningitidis MC58*	NC_003112	NMB1400	742	ABC transporter
29	Nme_1	*Neisseria meningitidis MC58*	NC_003112	NMB0098	165	Putative ABC transporter, truncated
30	Nme_2	*Neisseria meningitidis MC58*	NC_003112	NMB0103	164	Putative bacteriocin resistance protein
31	Nme_3	*Neisseria meningitidis MC58*	NC_003112	NMB0582	180	Putative bacteriocin resistance protein
32	Nme_4	*Neisseria meningitidis MC58*	NC_003112	NMB0855	218	Putative bacteriocin resistance protein
33	Nmf_1	*Neisseria meningitidis Z2491*	NC_003116	NMA1620	742	Putative cytolysin transporter
34	Nmf_1	*Neisseria meningitidis Z2491*	NC_003116	NMA0173	164	Hypothetical protein
35	Nmf_2	*Neisseria meningitidis Z2491*	NC_003116	NMA1066	218	Putative periplasmic protein
36	/	*Neisseria meningitidis Z2491*	NC_003116	#	63	Truncated protein
37	Neu_1	*Nitrosomonas europaea*	NC_004757	NE0789	231	Conserved hypothetical protein
38	Nsp_1	*Nostoc sp.*	NC_003272	all2021	712	ABC transporter
39	Nsp_2	*Nostoc sp.*	NC_003272	alr1201	722	ABC transporter
40	Nsp_3	*Nostoc sp.*	NC_003272	alr5147	716	ABC transporter
41	Nsp_4	*Nostoc sp.*	NC_003276	alr7014	714	ABC transporter
42	Nsp_5	*Nostoc sp.*	NC_003276	alr7295	736	ABC transporter
43	Nsp_1	*Nostoc sp.*	NC_003272	alr1927	1011	ABC transporter
44	Nsp_2	*Nostoc sp.*	NC_003272	hetC	1044	Heterocyst differentiation protein
45	Pma_1	*Prochlorococcus marinus*	NC_005071	PMT0978	738	Multidrug efflux family ABC transporter
46	Pae_1	*Pseudomonas aeruginosa*	NC_002516	PA4143	719	Probable toxin transporter
47	Pae_1	*Pseudomonas aeruginosa*	NC_002516	PA1762	253	Hypothetical protein
48	Pae_2	*Pseudomonas aeruginosa*	NC_002516	PA1953	226	Hypothetical protein
49	Ppu_1	*Pseudomonas putida*	NC_002947	PP4927	731	HlyB family ABC transporter
50	Ppu_1	*Pseudomonas putida*	NC_002947	PP2855	226	Conserved hypothetical protein
51	Rso_1	*Ralstonia solanacearum*	NC_003296	cyaB	725	Probable cyclolysin-type ABC transporter
52	Rso_1	*Ralstonia solanacearum*	NC_003295	RSc0445	234	Hypothetical signal peptide protein
53	Son_1	*Shewanella oneidensis*	NC_004349	SOA0049	725	Putative toxin secretion ABC transporter

54	Smu_1	*Streptococcus mutans*	NC_004350	SMU.1881c	763	Putative ABC transporter
55	Smu_2	*Streptococcus mutans*	NC_004350	SMU.286	760	Putative ABC transporter
56	Smu_1	*Streptococcus mutans*	NC_004350	SMU.1897	160	Putative ABC transporter
57	Spn_1	*Streptococcus pneumoniae R6*	NC_003098	clyB	702	Hemolysin ABC transporter
58	Spn_2	*Streptococcus pneumoniae R6*	NC_003098	comA	717	ABC transporter ComA
59	Spn_1	*Streptococcus pneumoniae R6*	NC_003098	spr0469	197	Conserved hypothetical protein, truncation
60	/	*Streptococcus pneumoniae R6*	NC_003098	spr0105	77	Truncated protein
61	Spo_1	*Streptococcus pneumoniae TIGR4*	NC_003028	SP0042	717	Competence factor transporter ComA
62	Spo_2	*Streptococcus pneumoniae TIGR4*	NC_003028	SP1953	698	Toxin secretion ABC transporter
63	Spo_1	*Streptococcus pneumoniae TIGR4*	NC_003028	SP0530	195	Putative ABC transporter, truncated
64	Spy_1	*Streptococcus pyogenes MGAS315*	NC_004070	SpyM3_1650	214	Putative salivaricin A modification enzyme
65	Spz_1	*Streptococcus pyogenes MGAS8232*	NC_003485	spyM18_0543	717	Putative ABC transporter
66	/	*Streptococcus pyogenes MGAS8232*	NC_003485	spyM18_0524	78	Truncated protein
67	Spa_1	*Streptococcus pyogenes SSI-1*	NC_004606	#	222	Similar to SalT of S. Salivarius, truncated
68	Sav_1	*Streptomyces avermitilis*	NC_003155	SAV7493	749	Putative ABC transporter
69	Sco_1	*Streptomyces coelicolor*	NC_003888	SCO0755	740	Putative ABC transporter
70	Ssp_1	*Synechococcus sp.*	NC_005070	SYNW2111	740	Putative ABC transporter
71	Ssq_1	*Synechocystis sp.*	NC_000911	hlyB	733	Hemolysin ABC transporter
72	Uur_1	*Ureaplasma urealyticum*	NC_002162	ABC-2	658	ABC transporter
73	Xfa_1	*Xylella fastidiosa 9a5c*	NC_002488	XF1220	707	Colicin V secretion ABC transporter
74	Xfa_2	*Xylella fastidiosa 9a5c*	NC_002488	XF2397	720	Toxin secretion ABC transporter
75	Xfb_1	*Xylella fastidiosa Temecula1*	NC_004556	cvaB	707	Colicin V secretion ABC transporter
76	Xfb_2	*Xylella fastidiosa Temecula1*	NC_004556	hlyB	720	Toxin secretion ABC transporter
77	Ype_1[c]	*Yersinia pestis CO92*	NC_003143	#	301	Putative secretion ATPase, truncated
78	Ypf_1	*Yersinia pestis KIM*	NC_004088	y0286	704	Putative ABC transporter

As mentioned above, two conserved motifs, called the cysteine and histidine motifs (C/H motifs), can be distinguished in the peptidase C39 domain, which are responsible for the recognition and cleavage of the GG-motif containing peptides. However, the exact consensus sequence of these motifs differs for Gram-positive and Gram-negative bacteria. For the Gram-positive bacteria, the consensus sequences are $QX_4(D/E)CX_2AX_3MX_4(Y/F)GX_4(I/L)$ and $H(Y/F)(Y/V)VX_{10}(I/L)XDP$ for the cysteine and histidine motifs respectively [105,282]. Based on seven ABC-transporters from Gram-negative species, Michiels *et al.* adapted the consensus sequence of the cysteine and histidine motifs for Gram-negative bacteria to $QX_4(D/E)C(G/A)XAXLX_2(I/V)X_4GX_4(I/L)X_2LR$ and $H(Y/F)(Y/V)V(L/V)X_9(I/L/V)XDP$, respectively [171]. For every Peptidase C39 domain, the presence of the C/H motifs was manually checked based on these consensus sequences. With the exception of 13 instances of the 78 candidate Peptidase C39 domains (indicated in table 7.3) which lack both conserved motifs, and three hits truncated in the Peptidase C39 domain, which only contain the cysteine motif, all Peptidase C39 domain hits contain the C/H motifs, suggesting that those domains have peptidase activity. For two hits (Nme_2 and Nmf_1), which have an identical protein sequence, the cysteine motif was only partially present, but after translation of the region flanking the 5' end of both genes, the full motif was found, suggesting that both genes are not correctly annotated. The 13 hits lacking the C/H motifs all belong to Gram-negative bacteria and constitute ABC-transporters that secrete toxins of the hemolysin family. It was already shown that hemolysins do not contain leader peptides and are not proteolytically processed during export and that their cognate transporters lack the C/H motifs in their N-terminal domain [105,294]. However, the hits Ccr_1, Efa_1, Mpn_1, Spn_1 and Ssq_1, also annotated as proteins transporting hemolysin-like toxins, contain the conserved C/H motifs.

7.2.3 Search for genes encoding double-glycine motif containing peptides

For the GG-motif search, two HMMs were built, based on peptides from either Gram-negative or Gram-positive bacteria. A Gram-positive training set was obtained from 31 already known GG-motif containing peptides. Since only a limited number of GG-motif sequences from Gram-negative bacteria was available [171], additional possible GG-motif containing peptides were first obtained via an iterative MEME/MAST approach.

Based on the twelve possible GG-motif containing peptides from Gram-negative bacteria obtained from a previous search [171], a probabilistic model of the GG-motif was constructed using MEME (see methods) [14]. Only the N-terminal leader sequences of those peptides were taken as input for MEME. The MEME output (i.e. a motif model describing the GG-motif) was used to screen all peptides from the non-redundant database from NCBI using MAST [15]. Hits with an E-value less than 10 were retained for further analysis. Subsequently, hits were manually refined by retaining only those peptides from Gram-negative bacteria that did not have any gap or insertion in their GG-motif and that ended with the amino acid pairs Gly-Gly or Gly-Ala (since all GG-motifs described so far end with such a pair). The obtained curated set of peptides was used as training set for a new MEME/MAST iteration using the same parameter settings as described above and also in the methods. The resulting output was manipulated in a way similar to the first iteration. Based on the second MEME/MAST cycle output, a curated training set of 38 possible GG-motifs from Gram-negative bacteria was obtained. Since most members of the training set had a length of 15 amino acids, only the 15 C-terminal amino acids from members containing longer sequences were taken into the training set (Fig. 7.4).

The Gram-positive and Gram-negative training sets were used to build HMMs by using the HMMER2.2 software (http://hmmer.wustl.edu) [61]. Using the Wise2 program, the GG-motif search was performed with both HMMs on the same set of DNA sequences given in table 7.1.

In a first step, since all GG-motif containing peptides described so far are genetically linked to their cognate transporters, a limit was set on the distance between both genes. Only those peptides were retained from which the coding region was located less than 10 kb from the coding region of a Peptidase C39 domain. Secondly, the length of the leader sequence and the total peptide length was set to a maximum of 30 and 150 amino acids respectively. Therefore, fore every GG-motif hit sequence an open reading frame was built, starting with one of the three possible start codons ATG, GTG or TTG. Finally the remaining hits were blasted using blastx and if a perfect match with a non-hypothetical protein (other than an already known GG-motif containing peptide) was found, the hit was removed. The results of the screening step are described separately for Gram-negative and Gram-positive bacteria.

7.2.3.1 Gram-negative bacteria

Initially, the search was performed with the HMM derived from the Gram-negative training set. Taking into account the restrictions described above, 38 possible GG-motif containing peptides were retrieved. The search with the HMM derived from the Gram-positive training set, resulted in 20 additional candidate peptides. So in total, 58 possible GG-motif containing peptides from Gram-negative bacteria were found. The size of the peptides was 23 to 142 amino acids, or in the mature form (i.e. without the leader peptide) from 5 to 124 amino acids. A list of the possible GG-motif containing peptides and their cognate transport proteins is given in Table 7.4.

As mentioned above, 13 from the 49 Peptidase C39 domain containing proteins from Gram-negative bacteria do not have the conserved C/H motifs, and as a consequence, no GG-motif coding genes should be found next to their coding region. For all transporters this is the case, except for Xfa_2, for which one GG-motif hit was retrieved. This could be a false positive hit. Concerning the 36 remaining transporters, 27 of them have at least one possible GG-motif containing peptide. For the nine other transporters, no possible GG-motif hits were observed. This could be due to the restrictions that were applied in this search. To conclude, many possible

GG-motif containing peptides that resulted from our search, show structural similarity with bacteriocins and peptide hormones.

Table 7.4: **List of the possible GG-motif containing peptides from Gram-negative bacteria.** Legend: #: not annotated. [a] The position of the GG-motif coding region on the corresponding replicon is given. [b] The name of the transport protein to which the possible peptide is genetically linked (<10 kb). [c] Length of the GG-leader sequence (GG) and the total peptide (Total) in amino acids.

ID	Organism	Acc.Nr.	Start[a]	End[a]	Gene	Transporter[b]	Length[c]	
							GG	Total
1	*Caulobacter crescentus*	NC_002696	747957	747913	#	Ccr_1	15	84
2	*Caulobacter crescentus*	NC_002696	748227	748183	#	Ccr_1	19	83
3	*Caulobacter crescentus*	NC_002696	754303	754347	CC0688	Ccr_1	19	142
4	*Caulobacter crescentus*	NC_002696	757113	757066	#	Ccr_1	17	28
5	*Caulobacter crescentus*	NC_002696	758792	758839	CC0692	Ccr_1	20	50
6	*Chromobacterium violaceum*	NC_005085	81944	81991	CV0071	Cvi_1	27	50
7	*Chromobacterium violaceum*	NC_005085	4746267	4746220	CV4401	Cvi_2	17	68
8	*Chromobacterium violaceum*	NC_005085	709116	709163	CV0683	Cvi_1	30	109
9	*Chromobacterium violaceum*	NC_005085	709170	709217	CV0683	Cvi_1	23	84
10	*Escherichia coli CFT073*	NC_004431	1176822	1176866	mchB	Eco_2	15	75
11	*Escherichia coli CFT073*	NC_004431	1183119	1183163	#	Eco_2	15	92
12	*Neisseria meningitidis MC58*	NC_003112	98911	98958	NMB0086	Nme_1	21	83
13	*Neisseria meningitidis MC58*	NC_003112	104406	104450	NMB0091	Nme_1	15	78
14	*Neisseria meningitidis MC58*	NC_003112	610925	610972	#	Nme_3	15	52
15	*Neisseria meningitidis MC58*	NC_003112	883878	883834	NMB0861	Nme_4	17	141
16	*Neisseria meningitidis MC58*	NC_003112	885348	885304	NMB0865	Nme_4	17	117
17	*Neisseria meningitidis Z2491*	NC_003116	154396	154349	NMA0169	Nmf_1	23	101
18	*Neisseria meningitidis Z2491*	NC_003116	1027673	1027629	NMA1073	Nmf_2	17	141
19	*Neisseria meningitidis Z2491*	NC_003116	1029700	1029656	NMA1078	Nmf_2	17	100
20	*Nitrosomonas europaea*	NC_004757	859268	859321	NE0790	Neu_1	19	34
21	*Nitrosomonas europaea*	NC_004757	851653	851606	NE0784	Neu_1	24	37
22	*Nostoc sp.*	NC_003272	2422705	2422661	asl2024	Nsp_1	23	37
23	*Nostoc sp.*	NC_003272	1418171	1418215	asr1202	Nsp_2	21	58
24	*Nostoc sp.*	NC_003272	1418704	1418751	asr1203	Nsp_2	25	95
25	*Nostoc sp.*	NC_003272	1419515	1419559	asr1204	Nsp_2	25	68
26	*Nostoc sp.*	NC_003272	6136580	6136624	asr5139	Nsp_3	20	80
27	*Nostoc sp.*	NC_003272	6136990	6137034	alr5140	Nsp_3	20	78
28	*Nostoc sp.*	NC_003272	6137394	6137438	alr5141	Nsp_3	20	78
29	*Nostoc sp.*	NC_003272	6137793	6137837	asr5142	Nsp_3	20	80
30	*Nostoc sp.*	NC_003272	6138151	6138195	alr5143	Nsp_3	27	91
31	*Nostoc sp.*	NC_003272	6139056	6139100	asr5144	Nsp_3	20	78
32	*Nostoc sp.*	NC_003272	6139788	6139832	asr5145	Nsp_3	24	85
33	*Nostoc sp.*	NC_003272	6140355	6140399	asr5146	Nsp_3	23	84
34	*Nostoc sp.*	NC_003276	13799	13846	#	Nsp_4	15	31
35	*Prochlorococcus marinus*	NC_005071	1048057	1048104	#	Pma_1	15	23
36	*Prochlorococcus marinus*	NC_005071	1059298	1059348	#	Pma_1	22	27
37	*Pseudomonas aeruginosa*	NC_002516	4629958	4630002	PA4139	Pae_1	20	99

38	*Pseudomonas aeruginosa*	NC_002516	4632476	4632520	PA4141	Pae_1	15	99
39	*Pseudomonas aeruginosa*	NC_002516	4632632	4632679	PA4141	Pae_1	16	48
40	*Pseudomonas aeruginosa*	NC_002516	1912484	1912528	PA1768	Pae_1	24	88
41	*Pseudomonas aeruginosa*	NC_002516	2137508	2137555	PA1951	Pae_2	15	49
42	*Pseudomonas aeruginosa*	NC_002516	2141561	2141608	#	Pae_2	21	77
43	*Ralstonia solanacearum*	NC_003295	472941	472988	RSc0444	Rso_1	19	87
44	*Shewanella oneidensis*	NC_004349	38530	38486	#	Son_1	15	67
45	*Shewanella oneidensis*	NC_004349	46897	46941	SOA0057	Son_1	25	60
46	*Synechocystis sp.*	NC_000911	3449398	3449354	#	Ssq_1	29	39
47	*Xylella fastidiosa 9a5c*	NC_002488	1166571	1166615	XF1217	Xfa_1	15	79
48	*Xylella fastidiosa 9a5c*	NC_002488	1166975	1167019	XF1218	Xfa_1	15	96
49	*Xylella fastidiosa 9a5c*	NC_002488	1167500	1167544	XF1219	Xfa_1	15	79
50	*Xylella fastidiosa 9a5c*	NC_002488	1167885	1167929	#	Xfa_1	15	66
51	*Xylella fastidiosa 9a5c*	NC_002488	2276882	2276838	XF2402	Xfa_2	15	27
52	*Xylella fastidiosa Temecula1*	NC_004556	593159	593203	PD0497	Xfb_1	15	79
53	*Xylella fastidiosa Temecula1*	NC_004556	593561	593605	PD0498	Xfb_1	15	96
54	*Xylella fastidiosa Temecula1*	NC_004556	594077	594121	#	Xfb_1	15	77
55	*Xylella fastidiosa Temecula1*	NC_004556	594460	594504	#	Xfb_1	15	66
56	*Yersinia pestis CO92*	NC_003143	105395	105442	#	Ype_1d	22	45
57	*Yersinia pestis CO92*	NC_003143	109863	109907	YPO0100	Ype_1d	15	79
58	*Yersinia pestis KIM*	NC_004088	309522	309566	y0288	Ypf_1	15	79

7.2.3.2 Gram-positive bacteria

The screening of Gram-positive bacteria for peptides containing a GG-motif resulted in a total of 48 candidate peptides. Although out of the 45 screened bacterial genomes only 12 genomes were from lactic acid bacteria, 92% of all GG-motif-containing hits were found in lactic acid bacteria (of which 80% belong to streptococcal strains). The size of the peptides ranges from 29 to 126 amino acids, or in mature form (i.e. without the leader sequence) from 11 to 103 amino acids. A list of possible GG-motif-containing peptides, their cognate transporter protein and their length is given in table 7.5. If available, the gene name of the GG-peptide encoding sequence was taken from the genome annotation data and included in the table.

Among the 48 hits, three were not annotated in the corresponding genome sequence project. Seventeen hits, annotated as hypothetical proteins, did not display similarity to any known protein or peptide. The remaining hits are bacteriocins (n=15) or bacteriocin homologs (n=10), a conserved domain protein (n=1), a plantaricin biosynthesis protein (n=1) and a phage-related proteins (n=1). Physical linkage to one or more possible GG-peptides was obtained in 21 out of the 29 Peptidase C39 domains retrieved.

Table 7.5: **List of the possible GG-motif containing peptides from Gram-negative bacteria.** Legend: #: not annotated. ¶: protein containing a truncated Peptidase C39 domain. [a] The position of the GG-motif coding region on the corresponding replicon is given. [b] The name of the transport protein to which the possible peptide is genetically linked (<10 kb). [c] Length of the GG-leader sequence (GG) and the total peptide (Tot.) in amino acids.

							Length[c]	
ID	**Organism**	**Acc.Nr.**	**Start[a]**	**Stop[a]**	**Gene**	**Transporter[b]**	**GG**	**Tot.**
1	*Bacillus subtilis*	NC_000964	2271945	2271901	yolB	sunT	16	57
2	*Clostridium acetobutylicum*	NC_001988	76046	76093	CAP0072	CAP0073	15	58
3	*Enterococcus faecalis*	NC_004671	49438	49485	EFB0056	EFB0050	26	74
4	*Lactococcus lactis*	NC_002662	86271	86315	#	lcnC	21	41
5	*Lactococcus lactis*	NC_002662	88983	89027	#	lcnC	17	41
6	*Lactobacillus plantarum*	NC_004567	366994	366947	plnJ	plnG	21	46
7	*Lactobacillus plantarum*	NC_004567	368259	368306	plnN	plnG	25	55
8	*Lactobacillus plantarum*	NC_004567	371389	371436	plnA	plnG	15	41
9	*Lactobacillus plantarum*	NC_004567	375965	375918	plnF	plnG	18	52
10	*Lactobacillus plantarum*	NC_004567	376145	376098	plnE	plnG	23	56
11	*Lactobacillus plantarum*	NC_004567	383668	383715	plnY	plnG	18	29
12	*Streptococcus mutans*	NC_004350	269347	269391	SMU.283	SMU.286	20	69
13	*Streptococcus mutans*	NC_004350	1776619	1776575	SMU.1882c	SMU.1881c/1897	23	117
14	*Streptococcus mutans*	NC_004350	1781366	1781319	SMU.1889c	SMU.1881c/1897	23	87
15	*Streptococcus mutans*	NC_004350	1781849	1781805	SMU.1892c	SMU.1881c/1897	25	61
16	*Streptococcus mutans*	NC_004350	1783593	1783549	SMU.1895c	SMU.1881c/1897	23	53
17	*Streptococcus mutans*	NC_004350	1783889	1783845	SMU.1896c	SMU.1881c/1897	18	78
18	*Streptococcus mutans*	NC_004350	1787899	1787855	SMU.1902c	SMU.1897	22	47
19	*Streptococcus mutans*	NC_004350	1789887	1789843	SMU.1905c	SMU.1897	22	62
20	*Streptococcus mutans*	NC_004350	1790260	1790213	SMU.1906c	SMU.1897	18	65
21	*Streptococcus mutans*	NC_004350	1794717	1794670	SMU.1914c	SMU.1897	23	76
22	*Streptococcus pneumoniae R6*	NC_003098	39538	39585	thmA	comA	18	71
23	*Streptococcus pneumoniae R6*	NC_003098	117752	117796	Spr0109	Spr0105¶	23	126
24	*Streptococcus pneumoniae R6*	NC_003098	119712	119756	Spr0111	Spr0105¶	23	123
25	*Streptococcus pneumoniae R6*	NC_003098	123834	123878	Spr0115	Spr0105¶	23	124
26	*Streptococcus pneumoniae R6*	NC_003098	472023	471979	Spr0465	Spr0469	24	51
27	*Streptococcus pneumoniae R6*	NC_003098	1732074	1732118	Spr1765	clyB	29	74
28	*Streptococcus pneumoniae R6*	NC_003098	1732293	1732337	Spr1766	clyB	25	62
29	*Streptococcus pneumoniae TIGR4*	NC_003028	39895	39942	SP0041	SP0042	18	71
30	*Streptococcus pneumoniae TIGR4*	NC_003028	507823	507779	SP0528	SP0530	24	42
31	*Streptococcus pneumoniae TIGR4*	NC_003028	511742	511786	SP0531	SP0530	18	60
32	*Streptococcus pneumoniae TIGR4*	NC_003028	512406	512453	SP0532	SP0530	18	84
33	*Streptococcus pneumoniae TIGR4*	NC_003028	512744	512791	SP0533	SP0530	18	71
34	*Streptococcus pneumoniae TIGR4*	NC_003028	515115	515159	SP0539	SP0530	18	79
35	*Streptococcus pneumoniae TIGR4*	NC_003028	515370	515414	SP0540	SP0530	18	67
36	*Streptococcus pneumoniae TIGR4*	NC_003028	515832	515879	SP0541	SP0530	18	44
37	*Streptococcus pneumoniae TIGR4*	NC_003028	1850404	1850448	SP1948	SP1953	29	74
38	*Streptococcus pneumoniae TIGR4*	NC_003028	1850623	1850667	SP1949	SP1953	25	62

39	*Streptococcus pyogenes MGAS315*	NC_004070	1664879	1664835	salA	SpyM3_1650	19	41
40	*Streptococcus pyogenes MGAS8232*	NC_003485	423034	423081	SpyM18_0525	SpyM18_0524¶	22	74
41	*Streptococcus pyogenes*	NC_003485	423161	423207	#	SpyM18_0524¶	21	82
42	*Streptococcus pyogenes*	NC_003485	424632	424679	SpyM18_0528	SpyM18_0524¶	18	70
43	*Streptococcus pyogenes*	NC_003485	430596	430552	SpyM18_0540	SpyM18_0524¶	24	41
44	*Streptococcus pyogenes*	NC_003485	434538	434582	SpyM18_0544	SpyM18_0543	23	75
45	*Streptococcus pyogenes*	NC_003485	434763	434807	SpyM18_0545	SpyM18_0543	15	63
46	*Streptococcus pyogenes*	NC_004606	1658665	1658618	SPs1650	/	19	41
47	*Streptomyces avermitilis*	NC_003155	8931651	8931604	SAV7495	SAV7493	16	64
48	*Streptomyces coelicolor*	NC_003888	796608	796652	SCO0753	SCO0755	23	71

7.3 Discussion

As opposed to Gram-negative bacteria, a lot of research has been done on peptide signal molecules and bacteriocins containing a double-glycine leader sequence in Gram-positive bacteria. In these bacteria it was shown that the GG-motif leader sequence plays a key role in many peptide secretion systems involved in quorum sensing and bacteriocin production [41,132]. This leader sequence, typically between 15 and 30 amino acids in length, is recognized and proteolytically removed during secretion by its cognate ABC-transporter carrying a Peptidase C39 domain. Processed peptides vary in length from 17 to over 80 amino acids. Peptide pheromones (which are quorum sensing related auto-inducers) and class II bacteriocins are generally synthesized as inactive prepeptides containing this conserved GG-type leader sequence. These bacteriocins are functionally related as these peptide pheromones act as auto-inducers for the start of the production of bacteriocins. Besides this functional relation, these peptides also show structural similarities (similar length, net positive charge between 3 and 6 at pH 7, length between 15 and 60 amino acids, amino acid composition).

In our screening of the Gram-negative bacterial genomes, the peptides identified constitute good candidates for bacteriocins and / or peptide pheromones. All peptides possess a GG-motif leader sequence and their corresponding genes are found in close proximity to a dedicated ABC transporter containing the Peptidase C39 domain. In addition, also at the structural level the peptides have similar physical characteristics (length, amino acid composition, net charge) which correspond to the peptide pheromones and bacteriocins found in Gram-positive bacteria. Both of these peptides have been shown to play a role in quorum sensing in Gram-positive bacteria. Therefore, a signalling function for these peptides can also be hypothesized in Gram-negative bacteria. However, the question can be raised as how these molecules are sensed by the cells. In general, being surrounded by two concentric lipid bilayer membranes, these peptides should either 1) be sensed at the outer membrane, periplasm or inner membrane, or 2) they should be active intracellularly.

In Gram positive bacteria, sensing at the cytoplasmatic membrane generally occurs trough a two-component signal-transduction system consisting of a membrane-located sensor and a cytoplasmatic response regulator [59,172]. Two-component regulatory systems are ubiquitous in most Gram-negative bacterial species [250]. It is therefore not unlikely that also Gram-negative bacteria could possess two-component regulatory systems able to detect peptides. For example, in several pathogenic bacteria, such as *Salmonella typhimurium* and *Pseudomonas aeruginosa*, cationic antimicrobial peptides can activate the expression of resistance genes [13,170]. In *S. typhimurium*, it was suggested that the two-component regulatory system PhoP-PhoQ could directly sense the presence of peptides [170].

Peptides that function intracellularly, need to pass the cytoplasmatic membrane to interact with their targets. In Gram-positive bacteria, the peptides are transported into the cell by oligopeptide permeases (Opp) [138,244,273]. As Opp systems are also ubiquitous in Gram-negative bacteria, it is not unlikely that short or even relatively long peptides could be imported in the cytoplasm by this system. An alternative way for signalling peptides to enter the cell is by diffusion across the cytoplasmatic membrane, comparable to the mechanism used by α-helical antimicrobial peptides [274]. In contrast to Gram-positive bacteria however, the signalling peptides working intracellularly in Gram-negative bacteria need to pass an extra barrier i.e. the relatively impermeable outer membrane. A major class of outer membrane proteins are the porins, i.e. integral membrane proteins which form β-barrels [57] that can catalyse the uptake of these oligopeptides. However, the question of whether Gram-negative bacteria could sense peptide signals remains an intriguing problem that still requires experimental validation.

Comparable to the screening of the Gram-negative bacteria, the genome-wide screening of 45 bacterial genomes resulted in a total of 48 candidate peptides containing a GG-motif. More than half of the GG-hits retrieved are bacteriocins, and some of them also function as pheromones. More than 40% of the identified peptide genes were either not annotated or had not yet been characterized as secreted peptides in the genome sequencing projects. These peptide genes were detected in the genomes of lactic acid bacteria, but also in the genera *Bacillus, Clostridium* and *Streptomyces.*

To conclude, our screening strategy led to new insights in the distribution of GG-peptide processing and secretion systems among Gram-positive and Gram-negative bacteria. The results are not dependent on previous annotations as the screening was performed at the nucleotide level. However, not all known GG-motif containing peptides in Gram-positive bacteria were found in the current analysis. Also, in the case of several Peptidase C39 domains no corresponding peptides were present. This may be due to the 10 kb restriction, but this is clearly not always the case. On the other hand, the motif of the GG-leader sequence used may be too specific. Therefore, as more GG-containing peptides will be characterized biochemically in the future, the algorithm could be further refined.

7.4 Methods

7.4.1 MEME/MAST system

MEME (Multiple EM for Motif Elicitation; http://meme.sdsc.edu/meme/website/meme.html version 3.0) is a tool for discovering motifs in a group of related DNA or protein sequences that uses the expectation-maximization algorithm [14,15]. The algorithm estimates how many times each motif occurs in each sequence in the dataset together with the width of the putative motifs, and outputs an alignment of the occurrences of the motif. The algorithm is capable of discovering several different motifs with differing numbers of occurrences in a single dataset. We run the MEME algorithm on a training set of peptides containing (putative) GG-motif containing peptides using the following parameters: minimum motif width: 6; maximum motif width: 30; zero or one instance of the motif per sequence; maximum number of motifs: 1. For the other parameters, default setting were used.

MAST (Motif Alignment and Search Tool; http://meme.sdsc.edu/meme/website/mast.html version 3.0) is a tool for searching biological sequence databases for sequences that contain one or more instances of a group of known motifs [15]. MAST takes as input a file containing the descriptions of one or more motifs and searches a selected sequence database for sequences that match the motifs. As a motif file, the output of the MEME motif discovery tool can be used directly. For the retrieval of extra instances of the GG-motif leader sequence, we screened the non-redundant peptide database of NCBI. For each sequence, the match scores are converted into an E-value. For a database hit with a given score, an E-value gives the number of hits that one would have expected to see with this score or higher by chance in a sequence database of this size. Because of the small width of the GG-motif leader peptide, the cut-off E-value was set to the MAST default E-value of 10. As this is a high E-value cutoff, manual refinement was necessary to remove false positive hits (see results).

7.4.2 Genome-wide screening using HMMs

For the screening of the different bacterial genomes with the GG-motif leader peptide and Peptidase C39 domain, those amino acid sequences are described as Hidden Markov Models (HMMs). Profile HMMs are statistical models of multiple sequence alignments. They use position-specific scores for amino acids (or nucleotides) and position specific penalties for opening and extending an insertion or deletion.

7.4.2.1 Building HMM model for GG-motif leader peptides

The HMM for the Peptidase C39 domain is downloaded from the Pfam database release 10.0 [19]. The profile HMM for the Gram-positive and Gram-negative leader peptide was constructed using the HMMER software package (release 2.3.2). The *hmmbuild* program is used to create a profile HMM based on the alignment of a set of putative GG-motif leader sequences. In a next step, *hmmcalibrate* is run on the HMM which produces some extra parameters by fitting a distribution to the scores obtained from a random (Monte Carlo) simulation of a small sequence database. These extra parameters will greatly increase the sensitivity of a database search.

7.4.2.2 Genome-wide screening with HMMs using Genewise

Wise2 is a package focused on comparisons of biopolymers, commonly DNA sequences and protein sequences. Wise2's particular strength is that it permits the comparison of DNA sequence at the level of its protein translation. We used the *genewise* algorithm of the Wise2 package version 2.2.0 which is able to translate DNA sequences into the six reading frames and compares the translations with a HMM. As a HMM model of a protein sequence, we use either the Gram-positive or Gram-negative HMM describing the GG-motif leader sequence, or the HMM describing the Peptidase C39 domain. We screened both strands (resulting in six reading frames, three on each strand), and selected the *alg333*-option that forces *genewise* to use the implementation optimised for non-spliced bacterial genomes. For the other parameters, default settings were used.

7.4.3 BlastX

To retrieve the full protein sequence of the Peptidase C39 domain, or to check whether a GG-motif leader sequence is located in an annotated, non-hypothetical protein, we used blastx to compare the DNA sequence coding for a putative peptidase or GG-motif leader sequence with the non-redundant protein database of NCBI. Translated BLAST services (i.e. blastx, release 2.1.2) are useful when trying to find homologous proteins to a nucleotide coding region. Blastx compares translational products of the nucleotide query sequence to a protein database. Because blastx translates the query sequence in all six reading frames and provides combined significance statistics for hits to different frames, it is particularly useful when the reading frame of the query sequence is unknown or it contains errors that may lead to frame shifts or other coding errors. As we are not interested in the identification of homologs, but rather want to check whether the retrieved DNA sequence is already annotated as a non-hypothetical protein, we require a perfect hit (determined on the number of identical amino acids).

Chapter 8

Conclusions and perspectives

8.1 Conclusions

In this thesis we have shown that cross-species comparison is a powerful tool for discovering new transcription factor binding sites and their corresponding regulons. These comparative genomic approaches are based on the analysis of orthologous intergenic sequences that are expected to contain binding sites (i.e. regulatory motifs) of a common regulatory protein. We developed methodologies that combine existing or newly developed algorithms to solve biological problems for which we could rely on comparative genomics. The main contributions of this thesis are:

- Development of two-step phylogenetic footprinting procedure resulting in a more reliable identification of regulons (e.g. PmrAB regulon).
- Development of a methodology to compare regulons between different species (e.g. PhoPQ regulon)
- A *de novo* motif detection tool based on comparative genomics to detect niche-specific regulatory motifs in closely related species.
- Application of comparative genomics tools on:
    - detection of sRNA target genes (e.g. for the *sraD* sRNA)
    - detection of small signalling peptides (GG-peptides) in all fully sequenced bacteria

Based on existing motif detection and motif screening tools, we further characterized two regulons that play an essential role in the pathogenicity of *Salmonella typhimurium*, i.e., the PmrAB and the PhoPQ regulons. The former regulon was only poorly characterized, and no *in silico* analysis had yet been performed. We screened the genome for potential PmrA targets and validated potential new targets by a newly developed two-step procedure for phylogenetic footprinting. Based on the developed approach, we could further delineate the PmrAB regulon and derive a more accurate motif model describing the corresponding PmrA regulatory motif.

For the PhoPQ regulon, we did not only rely on comparative genomics to extend the knowledge about this regulatory system, but also on expression data. Indeed, the small width of the PhoP motif resulted in a high number of false positive hits during motif detection, which could be filtered by combining motif predictions with the expression data. Using both data sources allowed us to reconstruct the PhoPQ regulon in *E. coli* and *S. typhimurium*. Surprisingly, only a very small part of the PhoPQ regulon of each respective species overlapped between both species (i.e. 9 orthologous operons seemed indirectly regulated by PhoPQ in both species and 3 orthologous operons were found to be directly regulated in both species). This analysis shows that a very well conserved regulatory system (such as the PhoPQ two-component systems can evolve fast with novel phenotypes as a result). Indeed, the acquisition of new target genes of the PhoPQ regulatory system might explain why *S. typhimurium* is a pathogen and *E. coli* is not.

For the identification of the PmrAB and the PhoPQ regulon, some information on the motif model and experimentally validated target genes was already available from literature. This allowed us to search in the lists of potential motifs generated by MotifSampler for the motif of interest. However, if one wants to identify regulatory motifs without any prior information, one has to rely on the mere property of “statistical overrepresentation”. This is not trivial as we could experience during the analysis of the PhoPQ and PmrAB motifs: at first it appeared that – when comparing orthologous intergenic sequences from evolutionary related organisms – not only the regulatory motif is conserved, but also long stretches of DNA flanking the biologically true regulatory motif. A traditional motif detection algorithm based on Gibbs sampling (e.g. MotifSampler) is hampered by the many “local optima”: and will rarely result in detecting the true biological motif. Secondly, it appeared that although two regulatory systems are evolutionary related (orthologs conserved in closely related species), their target genes might be very different. This of course, lowers the statistical overrepresentation and the motif detection algorithm needs to be able to detect a regulatory motif that is conserved in only a very small subset of genes.

In order to tackle both problems we developed an approach where phylogenetic footprinting is combined with information on co-regulation. We first use phylogenetic footprinting to detect conserved sequences within sets of orthologous promoter regions. Next, orthologous conserved regions are mutually compared to search for those conserved intergenic sequences that share a common regulatory motif. From this common region, a motif model is compiled that is used to screen the whole genome for missing targets. The combination of blocks resulting from different sets of orthologs permits a more correct delineation of the true biological motif in long stretches of evolutionary conserved promoter regions (first bottleneck). Secondly, by using a motif model based on conserved orthologous targets to screen the remainder of the genome for yet other targets, we can also detect niche- or species-specific regulatory motifs (second bottleneck).

Tools and methods on comparative genomics developed in chapter 3 to 5 were used in the subsequent two chapters to solve some real biological problems related to regulatory motif detection. Detection of regulatory motifs is a difficult task as these motifs are short (6 – 20 nucleotides), degenerated and hidden in a large amount of sequence data. The same problems also apply to the problem of detecting regulatory RNA targets (chapter 6) and signalling peptides (chapter 7). For the former, we performed a genome-wide screening to predict putative targets of a regulatory RNA potentially involved in biofilm formation. This work is still in progress.

For the latter problem, the full genomic sequence of all complete sequenced bacteria was screened for the signalling peptide located in the neighbourhood of a transporter protein. This screening strategy led to new insights in the distribution of GG-peptide processing and secretion systems among Gram-positive and Gram-negative bacteria.

8.2 Perspectives

In these days where genetic network inference and systems biology are reoccurring themes in bioinformatics research, detection of regulatory motifs cannot longer be considered as a discipline on its own, but needs to be viewed within the larger context of regulatory network reconstruction. Regulatory network reconstruction often requires a combination of different data sources (mostly experimental data) [18,92,140,236]. Motif data are an essential component in network reconstruction as these motifs are the basic functional elements of a regulatory network. Phylogenetic footprinting has an added value for network reconstruction because regulatory motifs are derived from sequence data exclusively. As a result, they can be used as an additional data source for network reconstruction algorithms, totally independent from the other (experimental) data. Exploiting the methodology we developed in Chapter 5 would allow us to construct for each organism a genome-wide motif compendium based on sequence data only. In the framework of this PhD, the methodology has only been applied to test data sets with a limited size of less than 100 genes [179]. Some steps in the current implementation are quite time-consuming, so algorithmic changes are required in order to leverage the motif detection to a genome-wide scale. This is planned as future work. Subsequently, integrating this genome-wide compendium with the in house collected expression compendia of prokaryotic species (*E. coli*, *B. subtilis*, *S. typhimurium*) will provide us with a completely new source of information to study bacterial networks and their evolution [this work is in progress together with Valerie Storms and Abeer Fadda (CMPG)]. Preliminary results of the genome-wide approach in *E. coli* have already been used as input data for a data integration tool: ReMoDiscovery [142] uses a combination of microarray data, ChIP chip data and motif information to retrieve regulatory modules. First results are promising, however further refinement of the motif detection step is required to get biologically more reliable results.

Despite the successful application of the methodology developed in chapter 5, the phylogenetic footprinting step had some essential flaws. The major issue with BlockSampler is the fact that this algorithm considers the orthologous sequences as statistically independent. Therefore we are now working on an optimized algorithm that will take into account the phylogenetic relationships between orthologous sequences, which can be obtained by introducing a motif substitution weight matrix. In addition motif detection method developed in chapter 5 uses the information from coexpression and orthologs in a sequential way (first phylogenetic footprinting, secondly aligning the conserved regions), while in the new algorithm both orthologous and coexpressed spaces will be searched simultaneously. However, this optimization is not straightforward as major adaptations are required to the traditional Gibbs Sampling strategy to make the algorithm work in two directions (coexpression and orthology). This tool will have some important advantages as it will allow studying motif evolution across species. This algorithm is currently implemented and tested in collaboration with Marleen Claeys and Sun Hong.

Another often underestimated issue in inferring regulatory networks is the role of sRNAs. If one wants to interpret the networks inferred by using microarray data correctly, integrating the influence of regulatory RNAs in the network reconstruction is indispensable. Therefore, in the near future, identification of the different regulatory RNAs and their corresponding targets will become an important issue. Methods for detecting the small RNAs themselves are already available. The next challenge will be the development of equally effective methods for finding the targets of these RNAs. In the framework of this PhD work, we developed an initial screening procedure for the detection of regulatory RNA targets. However, as more detailed working mechanisms of regulatory RNA will become available, this screening step will be further fine-tuned. As this moment we are setting up collaboration with Dr. J. Vogel (Max Plank Institute) on this topic. Concomitantly with this, it will be important to study the mechanism by which the sRNAs influence the expression of their targets. To this end, models underlying network inference tools need to be adapted to also take into account the influence of sRNA on the mRNA level.

Ultimately, by applying the methods generated in this PhD (such as the generation of genome wide motif compendia and sRNA) on other genomes, and combining the results with the increasing the amount of publicly available experimental data, will lead to the elucidation of the regulatory networks in a wide range of microorganisms. At this stage the identification of the static regulatory network is already a major challenge. At the long term however, the ultimate goal is to also model the dynamic behaviour of bacteria. This will not only result in a better understanding of prokaryotic transcriptional regulation but also will obtain insight into the evolution of the complete transcriptional regulatory networks.

References

1. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B: Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res* 2003, 31:1753-1764.
2. Aguirre A, Lejona S, Vescovi EG, Soncini FC: Phosphorylated PmrA interacts with the promoter region of *ugd* in *Salmonella enterica* serovar typhimurium. *J Bacteriol* 2000, 182:3874-3876.
3. Alkema WB, Lenhard B, Wasserman WW: Regulog analysis: detection of conserved regulatory networks across bacteria: application to *Staphylococcus aureus*. *Genome Res* 2004, 14:1362-1373.
4. Alon U, Surette MG, Barkai N, Leibler S: Robustness in bacterial chemotaxis. *Nature* 1999, 397:168-171.
5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25:3389-3402.
6. Altuvia S, Zhang A, Argaman L, Tiwari A, Storz G: The *Escherichia coli* OxyS regulatory RNA represses fhlA translation by blocking ribosome binding. *EMBO J* 1998, 17:6069-6075.
7. Andersen J, Forst SA, Zhao K, Inouye M, Delihas N: The function of *micF* RNA. *micF* RNA is a major factor in the thermal regulation of OmpF protein in *Escherichia coli*. *J Biol Chem* 1989, 264:17961-17970.
8. Aravind L, Koonin EV: DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res* 1999, 27:4658-4670.
9. Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EG, Margalit H, Altuvia S: Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr Biol* 2001, 11:941-950.
10. Argaman L, Altuvia S: *fhlA* repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *J Mol Biol* 2000, 300:1101-1112.
11. Arita M, Robert M, Tomita M: All systems go: launching cell simulation fueled by integrated experimental biology data. *Curr Opin Biotechnol* 2005, 16:344-349.
12. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA: Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* 2004, 14:283-291.
13. Bader MW, Navarre WW, Shiau W, Nikaido H, Frye JG, McClelland M, Fang FC, Miller SI: Regulation of *Salmonella typhimurium* virulence gene expression by cationic antimicrobial peptides. *Mol Microbiol* 2003, 50:219-230.

14. Bailey TL, Elkan C: The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* 1995, 3:21-29.
15. Bailey TL, Elkan C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994, 2:28-36.
16. Baker SJ, Gunn JS, Morona R: The *Salmonella typhi* melittin resistance gene *pqaB* affects intracellular growth in PMA-differentiated U937 cells, polymyxin B resistance and lipopolysaccharide. *Microbiology* 1999, 145:367-378.
17. Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernandez-Boussard T, Jin H, Matese JC, Nitzberg M, Wymore F et al.: The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res* 2005, 33:D580-D582.
18. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA et al.: Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* 2003, 21:1337-1342.
19. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: The Pfam protein families database. *Nucleic Acids Res* 2002, 30:276-280.
20. Bearson BL, Wilson L, Foster JW: A low pH-inducible, PhoPQ-dependent acid tolerance response protects *Salmonella typhimurium* against inorganic acid stress. *J Bacteriol* 1998, 180:2409-2417.
21. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL: GenBank. *Nucleic Acids Res* 2002, 30:17-20.
22. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D et al.: Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 2002, 417:141-147.
23. Bi C, Rogan PK: Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucleic Acids Res* 2004, 32:4979-4991.
24. Birney E, Durbin R: Using GeneWise in the *Drosophila* annotation experiment. *Genome Res* 2000, 10:547-548.
25. Blanc-Potard AB, Lafay B: MgtC as a horizontally-acquired virulence factor of intracellular bacterial pathogens: evidence from molecular phylogeny and comparative genomics. *J Mol Evol* 2003, 57:479-486.
26. Blanc-Potard AB, Groisman EA: The *Salmonella selC* locus contains a pathogenicity island mediating intramacrophage survival. *EMBO J* 1997, 16:5376-5385.
27. Blekas K, Fotiadis DI, Likas A: Greedy mixture learning for multiple motif discovery in biological sequences. *Bioinformatics* 2003, 19:607-617.
28. Bockhorst J, Craven M, Page D, Shavlik J, Glasner J: A Bayesian network approach to operon prediction. *Bioinformatics* 2003, 19:1227-1235.
29. Brantl S: Antisense-RNA regulation and RNA interference. *Biochim Biophys Acta* 2002, 1575:15-25.

30. Breazeale SD, Ribeiro AA, Raetz CR: Oxidative decarboxylation of UDP-glucuronic acid in extracts of polymyxin-resistant *Escherichia coli*. Origin of lipid A species modified with 4-amino-4-deoxy-L-arabinose. *J Biol Chem* 2002, 277:2886-2896.
31. Brodsky IE, Ernst RK, Miller SI, Falkow S: *mig-14* is a *Salmonella* gene that plays a role in bacterial resistance to antimicrobial peptides. *J Bacteriol* 2002, 184:3203-3213.
32. Buhler J, Tompa M: Finding motifs using random projections. *J Comput Biol* 2002, 9:225-242.
33. Bullas LR, Ryu JI: *Salmonella typhimurium* LT2 strains which are r- m+ for all three chromosomally located systems of DNA restriction and modification. *J Bacteriol* 1983, 156:471-474.
34. Burgess RR, Travers AA, Dunn JJ, Bautz EK: Factor stimulating transcription by RNA polymerase. *Nature* 1969, 221:43-46.
35. Bussemaker HJ, Li H, Siggia ED: Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A* 2000, 97:10096-10100.
36. Bussemaker HJ, Li H, Siggia ED: Regulatory element detection using correlation with expression. *Nat Genet* 2001, 27:167-171.
37. Calhoun DH, Bonner CA, Gu W, Xie G, Jensen RA: The emerging periplasm-localized subclass of AroQ chorismate mutases, exemplified by those from *Salmonella typhimurium* and *Pseudomonas aeruginosa*. *Genome Biol* 2001, 2:research0030.
38. Campbell EA, Muzzin O, Chlenov M, Sun JL, Olson CA, Weinman O, Trester-Zedlitz ML, Darst SA: Structure of the bacterial RNA polymerase promoter specificity sigma subunit. *Mol Cell* 2002, 9:527-539.
39. Chamnongpol S, Dodson W, Cromie MJ, Harris ZL, Groisman EA: Fe(III)-mediated cellular toxicity. *Mol Microbiol* 2002, 45:711-719.
40. Chen S, Zhang A, Blyn LB, Storz G: MicC, a second small-RNA regulator of Omp protein expression in *Escherichia coli*. *J Bacteriol* 2004, 186:6689-6697.
41. Cheng Q, Campbell EA, Naughton AM, Johnson S, Masure HR: The com locus controls genetic transformation in *Streptococcus pneumoniae*. *Mol Microbiol* 1997, 23:683-692.
42. Coessens B, Thijs G, Aerts S, Marchal K, De Smet F, Engelen K, Glenisson P, Moreau Y, Mathys J, De Moor B: INCLUSive: A web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Res* 2003, 31:3468-3470.
43. Collado-Vides J, Magasanik B, Gralla JD: Control site location and transcriptional regulation in *Escherichia coli*. *Microbiol Rev* 1991, 55:371-394.
44. Conlon EM, Liu XS, Lieb JD, Liu JS: Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci U S A* 2003, 100:3339-3344.

45. Cotter PA, DiRita VJ: Bacterial virulence gene regulation: an evolutionary perspective. *Annu Rev Microbiol* 2000, 54:519-565.
46. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO: Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 2004, 429:92-96.
47. Darwin KH, Miller VL: Molecular basis of the interaction of *Salmonella* with the intestinal mucosa. *Clin Microbiol Rev* 1999, 12:405-428.
48. Day WH, McMorris FR: Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Res* 1992, 20:1093-1099.
49. De Bie T, Monsieurs P, Engelen K, De Moor B, Cristianini N, Marchal K: Discovering transcriptional modules from motif, chip-chip and microarray data. *Pac Symp Biocomput* 2005,483-494.
50. De Keersmaecker SC, Marchal K, Verhoeven TL, Engelen K, Vanderleyden J, Detweiler CS: Microarray analysis and motif detection reveal new targets of the *Salmonella enterica* serovar Typhimurium HilA regulatory protein, including *hilA* itself. *J Bacteriol* 2005, 187:4381-4391.
51. De Keersmaecker SC, Sonck K, Vanderleyden J: Let LuxS speak up in AI-2 signaling. *Trends Microbiol* 2006, 14:114-119.
52. De Keersmaecker SC, Varszegi C, van Boxel N, Habel LW, Metzger K, Daniels R, Marchal K, De Vos D, Vanderleyden J: Chemical synthesis of (S)-4,5-dihydroxy-2,3-pentanedione, a bacterial signal molecule precursor, and validation of its activity in *Salmonella typhimurium*. *J Biol Chem* 2005, 280:19563-19568.
53. De Smet F, Moreau Y, Engelen K, Timmerman D, Vergote I, De Moor B: Balancing false positives and false negatives for the detection of differential expression in malignancies. *Br J Cancer* 2004, 91:1160-1165.
54. Detweiler CS, Monack DM, Brodsky IE, Mathew H, Falkow S: *virK*, *somA* and *rcsC* are important for systemic *Salmonella enterica* serovar Typhimurium infection and cationic peptide resistance. *Mol Microbiol* 2003, 48:385-400.
55. Dirix G, Monsieurs P, Dombrecht B, Daniels R, Marchal K, Vanderleyden J, Michiels J: Peptide signal molecules and bacteriocins in Gram-negative bacteria: a genome-wide in silico screening for peptides containing a double-glycine leader sequence and their cognate transporters. *Peptides* 2004, 25:1425-1440.
56. Dirix G, Monsieurs P, Marchal K, Vanderleyden J, Michiels J: Screening genomes of Gram-positive bacteria for double-glycine-motif-containing peptides. *Microbiology* 2004, 150:1121-1126.
57. Dombrecht B, Marchal K, Vanderleyden J, Michiels J: Prediction and overview of the RpoN-regulon in closely related species of the Rhizobiales. *Genome Biol* 2002, 3:R0076.
58. Dosselaere F, Vanderleyden J: A metabolic node in action: chorismate-utilizing enzymes in microorganisms. *Crit Rev Microbiol* 2001, 27:75-131.
59. Dunny GM, Leonard BA: Cell-cell communication in gram-positive bacteria. *Annu Rev Microbiol* 1997, 51:527-564.

60. Ebright RH: RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J Mol Biol* 2000, 304:687-698.
61. Eddy SR: Profile hidden Markov models. *Bioinformatics* 1998, 14:755-763.
62. Edgar R, Domrachev M, Lash AE: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002, 30:207-210.
63. Edwards MT, Rison SC, Stoker NG, Wernisch L: A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context. *Nucleic Acids Res* 2005, 33:3253-3262.
64. Ejim LJ, D'Costa VM, Elowe NH, Loredo-Osti JC, Malo D, Wright GD: Cystathionine beta-lyase is important for virulence of *Salmonella enterica* serovar Typhimurium. *Infect Immun* 2004, 72:3310-3314.
65. Engelen K, Coessens B, Marchal K, De Moor B: MARAN: normalizing micro-array data. *Bioinformatics* 2003, 19:893-894.
66. Engelen K, Naudts B, De Moor B, Marchal K: A calibration method for estimating absolute expression levels from microarray data. *Bioinformatics* 2006, 22:1258.
67. Enright AJ, Van Dongen S, Ouzounis CA: An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002, 30:1575-1584.
68. Ermolaeva MD, White O, Salzberg SL: Prediction of operons in microbial genomes. *Nucleic Acids Res* 2001, 29:1216-1221.
69. Eskin E, Pevzner PA: Finding composite regulatory patterns in DNA sequences. *Bioinformatics* 2002, 18:S354-S363.
70. Favorov AV, Gelfand MS, Gerasimova AV, Ravcheev DA, Mironov AA, Makeev VJ: A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics* 2005, 21:2240-2245.
71. Fields PI, Groisman EA, Heffron F: A *Salmonella* locus that controls resistance to microbicidal proteins from phagocytic cells. *Science* 1989, 243:1059-1062.
72. Fields PI, Swanson RV, Haidaris CG, Heffron F: Mutants of *Salmonella typhimurium* that cannot survive within the macrophage are avirulent. *Proc Natl Acad Sci U S A* 1986, 83:5189-5193.
73. Foster JW, Hall HK: Adaptive acidification tolerance response of *Salmonella typhimurium*. *J Bacteriol* 1990, 172:771-778.
74. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM et al.: The minimal gene complement of *Mycoplasma genitalium*. *Science* 1995, 270:397-403.
75. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC: Cross-species sequence comparisons: a review of methods and available resources. *Genome Res* 2003, 13:1-12.
76. Freeman JA, Ohl ME, Miller SI: The *Salmonella enterica* serovar typhimurium translocated effectors SseJ and SifB are targeted to the *Salmonella*-containing vacuole. *Infect Immun* 2003, 71:418-427.

77. Garcia Vescovi E, Soncini FC, Groisman EA: Mg^{2+} as an extracellular signal: environmental regulation of *Salmonella* virulence. *Cell* 1996, 84:165-174.
78. Ge H, Walhout AJ, Vidal M: Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* 2003, 19:551-560.
79. Geissmann TA, Touati D: Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J* 2004, 23:396-405.
80. Gelfand MS, Koonin EV, Mironov AA: Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res* 2000, 28:695-705.
81. Gelfand MS: Recognition of regulatory sites by genomic comparison. *Res Microbiol* 1999, 150:755-771.
82. Gelfand MS: Evolution of transcriptional regulatory networks in microbial genomes. *Curr Opin Struct Biol* 2006, 16:420-429.
83. Gibbons HS, Lin S, Cotter RJ, Raetz CR: Oxygen requirement for the biosynthesis of the S-2-hydroxymyristate moiety in *Salmonella typhimurium* lipid A. Function of LpxO, A new Fe2+/alpha-ketoglutarate-dependent dioxygenase homologue. *J Biol Chem* 2000, 275:32940-32949.
84. Gottesman S: Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet* 2005, 21:399-404.
85. Gottesman S: Small RNAs shed some light. *Cell* 2004, 118:1-2.
86. Graumann P, Marahiel MA: A case of convergent evolution of nucleic acid binding modules. *Bioessays* 1996, 18:309-315.
87. Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M: Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res* 2001, 11:1463-1468.
88. Groisman EA: The pleiotropic two-component regulatory system PhoP-PhoQ. *J Bacteriol* 2001, 183:1835-1842.
89. Groisman EA, Chiao E, Lipps CJ, Heffron F: *Salmonella typhimurium phoP* virulence gene is a transcriptional regulator. *Proc Natl Acad Sci U S A* 1989, 86:7077-7081.
90. Groisman EA, Heffron F, Solomon F: Molecular genetic analysis of the *Escherichia coli phoP* locus. *J Bacteriol* 1992, 174:486-491.
91. Groisman EA, Parra-Lopez C, Salcedo M, Lipps CJ, Heffron F: Resistance to host antimicrobial peptides is necessary for *Salmonella* virulence. *Proc Natl Acad Sci U S A* 1992, 89:11939-11943.
92. Guelzim N, Bottani S, Bourgine P, Kepes F: Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* 2002, 31:60-63.
93. Guina T, Yi EC, Wang H, Hackett M, Miller SI: A PhoP-regulated outer membrane protease of *Salmonella enterica* serovar typhimurium promotes resistance to alpha-helical antimicrobial peptides. *J Bacteriol* 2000, 182:4077-4086.

94. Gunn JS, Miller SI: PhoP-PhoQ activates transcription of *pmrAB*, encoding a two-component regulatory system involved in *Salmonella typhimurium* antimicrobial peptide resistance. *J Bacteriol* 1996, 178:6857-6864.

95. Gunn JS, Lim KB, Krueger J, Kim K, Guo L, Hackett M, Miller SI: PmrA-PmrB-regulated genes necessary for 4-aminoarabinose lipid A modification and polymyxin resistance. *Mol Microbiol* 1998, 27:1171-1182.

96. Gunn JS, Ryan SS, Van Velkinburgh JC, Ernst RK, Miller SI: Genetic and functional analysis of a PmrA-PmrB-regulated locus necessary for lipopolysaccharide modification, antimicrobial peptide resistance, and oral virulence of *Salmonella enterica* serovar typhimurium. *Infect Immun* 2000, 68:6139-6146.

97. Gunn JS, Alpuche-Aranda CM, Loomis WP, Belden WJ, Miller SI: Characterization of the *Salmonella typhimurium pagC/pagD* chromosomal region. *J Bacteriol* 1995, 177:5040-5047.

98. Gunsalus RP, Park SJ: Aerobic-anaerobic gene regulation in *Escherichia coli*: control by the ArcAB and Fnr regulons. *Res Microbiol* 1994, 145:437-450.

99. Guo L, Lim KB, Gunn JS, Bainbridge B, Darveau RP, Hackett M, Miller SI: Regulation of lipid A modifications by *Salmonella typhimurium* virulence genes *phoP-phoQ*. *Science* 1997, 276:250-253.

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

101. Hall TA: BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nuceicl Acids Symp* 1999, 41:95-98.

102. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J et al.: Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004, 431:99-104.

103. Harrison SC: A structural taxonomy of DNA-binding domains. *Nature* 1991, 353:715-719.

104. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA: Combining location and expression data for principled discovery of genetic regulatory network models. *Pac Symp Biocomput* 2002,437-449.

105. Havarstein LS, Diep DB, Nes IF: A family of bacteriocin ABC transporters carry out proteolytic processing of their substrates concomitant with export. *Mol Microbiol* 1995, 16:229-240.

106. Havarstein LS, Holo H, Nes IF: The leader peptide of colicin V shares consensus sequences with leader peptides that are common among peptide bacteriocins produced by gram-positive bacteria. *Microbiology* 1994, 140:2383-2389.

107. Hengge-Aronis R: Recent insights into the general stress response regulatory network in *Escherichia coli*. *J Mol Microbiol Biotechnol* 2002, 4:341-346.

108. Herrgard MJ, Covert MW, Palsson BO: Reconstruction of microbial transcriptional regulatory networks. *Curr Opin Biotechnol* 2004, 15:70-77.

109. Herrgard MJ, Covert MW, Palsson BO: Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res* 2003, 13:2423-2434.

110. Hertz GZ, Stormo GD: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 1999, 15:563-577.

111. Hitchen PG, Prior JL, Oyston PC, Panico M, Wren BW, Titball RW, Morris HR, Dell A: Structural characterization of lipo-oligosaccharide (LOS) from *Yersinia pestis*: regulation of LOS structure by the PhoPQ system. *Mol Microbiol* 2002, 44:1637-1650.

112. Hohmann S: The Yeast Systems Biology Network: mating communities. *Curr Opin Biotechnol* 2005, 16:356-360.

113. Hughes JD, Estep PW, Tavazoie S, Church GM: Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* 2000, 296:1205-1214.

114. Huynen MA, Bork P: Measuring genome evolution. *Proc Natl Acad Sci U S A* 1998, 95:5849-5856.

115. Ideker T, Galitski T, Hood L: A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2001, 2:343-372.

116. Ideker T, Thorsson V, Siegel AF, Hood LE: Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J Comput Biol* 2000, 7:805-817.

117. Ishihama A: Functional modulation of *Escherichia coli* RNA polymerase. *Annu Rev Microbiol* 2000, 54:499-518.

118. Ishii T, Yoshida K, Terai G, Fujita Y, Nakai K: DBTBS: a database of *Bacillus subtilis* promoters and transcription factors. *Nucleic Acids Res* 2001, 29:278-280.

119. Jacob E, Sasikumar R, Nair KN: A fuzzy guided genetic algorithm for operon prediction. *Bioinformatics* 2005, 21:1403-1407.

120. Jensen ST, Shen L, Liu JS: Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics* 2005, 21:3832-3839.

121. Kalir S, Alon U: Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. *Cell* 2004, 117:713-720.

122. Kang Y, Weber KD, Qiu Y, Kiley PJ, Blattner FR: Genome-wide expression analysis indicates that FNR of *Escherichia coli* K-12 regulates a large number of genes of unknown function. *J Bacteriol* 2005, 187:1135-1160.

123. Kasahara M, Nakata A, Shinagawa H: Molecular analysis of the *Escherichia coli phoP-phoQ* operon. *J Bacteriol* 1992, 174:492-498.

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

125. Kato A, Tanabe H, Utsumi R: Molecular characterization of the PhoP-PhoQ two-component system in *Escherichia coli K-12*: identification of extracellular Mg^{2+}-responsive promoters. *J Bacteriol* 1999, 181:5516-5520.

126. Kato A, Latifi T, Groisman EA: Closing the loop: The PmrA/PmrB two-component system negatively controls expression of its posttranscriptional activator PmrD. *Proc Natl Acad Sci U S A* 2003, 100:4706-4711.

127. Keles S, van der LM, Eisen MB: Identification of regulatory elements using a feature selection method. *Bioinformatics* 2002, 18:1167-1175.

128. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD: EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res* 2005, 33:D334-D337.

129. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B: A high-resolution map of active promoters in the human genome. *Nature* 2005, 436:876-880.

130. Kim W, Killam T, Sood V, Surette MG: Swarm-cell differentiation in *Salmonella enterica* serovar Typhimurium results in elevated resistance to multiple antibiotics. *J Bacteriol* 2003, 185:3111-3117.

131. Kitano H: Computational systems biology. *Nature* 2002, 420:206-210.

132. Kleerebezem M, Quadri LE, Kuipers OP, de Vos WM: Quorum sensing by peptide pheromones and two-component signal-transduction systems in Gram-positive bacteria. *Mol Microbiol* 1997, 24:895-904.

133. Koonin EV, Galperin MY: Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr Opin Genet Dev* 1997, 7:757-763.

134. Kox LF, Wosten MM, Groisman EA: A small protein that mediates the activation of a two-component system by another two-component system. *EMBO J* 2000, 19:1861-1872.

135. Laikova ON, Mironov AA, Gelfand MS: Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of Proteobacteria. *FEMS Microbiol Lett* 2001, 205:315-322.

136. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993, 262:208-214.

137. Lawrence CE, Reilly AA: An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* 1990, 7:41-51.

138. Lazazzera BA: The intracellular function of extracellular signaling peptides. *Peptides* 2001, 22:1519-1527.

139. Lease RA, Cusick ME, Belfort M: Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proc Natl Acad Sci U S A* 1998, 95:12456-12461.

140. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I et al.: Transcriptional

regulatory networks in *Saccharomyces cerevisiae*. *Science* 2002, 298:799-804.

141. Lejona S, Aguirre A, Cabeza ML, Garcia VE, Soncini FC: Molecular characterization of the Mg^{2+}-responsive PhoP-PhoQ regulon in *Salmonella enterica*. *J Bacteriol* 2003, 185:6287-6294.

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

143. Lenz DH, Mok KC, Lilley BN, Kulkarni RV, Wingreen NS, Bassler BL: The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholerae*. *Cell* 2004, 118:69-82.

144. Li X, Wong WH: Sampling motifs on phylogenetic trees. *Proc Natl Acad Sci U S A* 2005, 102:9481-9486.

145. Liu J, Tan K, Stormo GD: Computational identification of the Spo0A-phosphate regulon that is essential for the cellular differentiation and development in Gram-positive spore-forming bacteria. *Nucleic Acids Res* 2003, 31:6891-6903.

146. Liu X, Brutlag DL, Liu JS: BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* 2001,127-138.

147. Liu XS, Brutlag DL, Liu JS: An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol* 2002, 20:835-839.

148. Lonetto M, Gribskov M, Gross CA: The sigma 70 family: sequence conservation and evolutionary relationships. *J Bacteriol* 1992, 174:3843-3849.

149. Lopez-Solanilla E, Garcia-Olmedo F, Rodriguez-Palenzuela P: Inactivation of the *sapA* to *sapF* locus of *Erwinia chrysanthemi* reveals common features in plant and animal bacterial pathogenesis. *Plant Cell* 1998, 10:917-924.

150. Luscombe NM, Austin SE, Berman HM, Thornton JM: An overview of the structures of protein-DNA complexes. *Genome Biol* 2000, 1:REVIEWS001.

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

152. Macisaac KD, Gordon DB, Nekludova L, Odom DT, Schreiber J, Gifford DK, Young RA, Fraenkel E: A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. *Bioinformatics* 2006, 22:423-429.

153. Madan BM, Teichmann SA: Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res* 2003, 31:1234-1244.

154. Madan BM, Teichmann SA: Functional determinants of transcription factors in *Escherichia coli*: protein families and binding sites. *Trends Genet* 2003, 19:75-79.

155. Madan BM, Teichmann SA, Aravind L: Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* 2006, 358:614-633.

156. Mahony S, Golden A, Smith TJ, Benos PV: Improved detection of DNA motifs using a self-organized clustering of familial binding profiles. *Bioinformatics* 2005, 21:S283-S291.

157. Majdalani N, Cunning C, Sledjeski D, Elliott T, Gottesman S: DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proc Natl Acad Sci U S A* 1998, 95:12462-12467.

158. Majdalani N, Hernandez D, Gottesman S: Regulation and mode of action of the second small RNA activator of RpoS translation, RprA. *Mol Microbiol* 2002, 46:813-826.

159. Makita Y, Nakao M, Ogasawara N, Nakai K: DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res* 2004, 32:D75-D77.

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

161. Marchal K, Thijs G, De Keersmaecker S, Monsieurs P, De Moor B, Vanderleyden J: Genome-specific higher-order background models to improve motif detection. *Trends Microbiol* 2003, 11:61-66.

162. Maslov S, Sneppen K, Eriksen KA, Yan KK: Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evol Biol* 2004, 4:9.

163. Masse E, Majdalani N, Gottesman S: Regulatory roles for small RNAs in bacteria. *Curr Opin Microbiol* 2003, 6:120-124.

164. Masse E, Gottesman S: A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proc Natl Acad Sci U S A* 2002, 99:4620-4625.

165. Masuda N, Church GM: Regulatory network of acid resistance genes in *Escherichia coli*. *Mol Microbiol* 2003, 48:699-712.

166. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F: Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature* 2001, 413:852-856.

167. McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE: Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res* 2001, 29:774-782.

168. McCue LA, Thompson W, Carmack CS, Lawrence CE: Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res* 2002, 12:1523-1532.

169. McGuire AM, Hughes JD, Church GM: Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res* 2000, 10:744-757.

170. McPhee JB, Lewenza S, Hancock RE: Cationic antimicrobial peptides activate a two-component regulatory system, PmrA-PmrB, that regulates resistance to polymyxin B and cationic antimicrobial peptides in *Pseudomonas aeruginosa*. *Mol Microbiol* 2003, 50:205-217.

171. Michiels J, Dirix G, Vanderleyden J, Xi C: Processing and export of peptide pheromones and bacteriocins in Gram-negative bacteria. *Trends Microbiol* 2001, 9:164-168.

172. Miller MB, Bassler BL: Quorum sensing in bacteria. *Annu Rev Microbiol* 2001, 55:165-199.

173. Miller SI, Mekalanos JJ: Constitutive expression of the *phoP* regulon attenuates *Salmonella* virulence and survival within macrophages. *J Bacteriol* 1990, 172:2485-2490.

174. Miller SI, Kukral AM, Mekalanos JJ: A two-component regulatory system (*phoP phoQ*) controls *Salmonella typhimurium* virulence. *Proc Natl Acad Sci U S A* 1989, 86:5054-5058.

175. Minagawa S, Ogasawara H, Kato A, Yamamoto K, Eguchi Y, Oshima T, Mori H, Ishihama A, Utsumi R: Identification and molecular characterization of the Mg^{2+} stimulon of *Escherichia coli*. *J Bacteriol* 2003, 185:3696-3702.

176. Mironov AA, Koonin EV, Roytberg MA, Gelfand MS: Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res* 1999, 27:2981-2989.

177. Moller T, Franch T, Udesen C, Gerdes K, Valentin-Hansen P: Spot 42 RNA mediates discoordinate expression of the *E. coli* galactose operon. *Genes Dev* 2002, 16:1696-1706.

178. Monsieurs P, De Keersmaecker S, Navarre WW, Bader MW, De Smet F, McClelland M, Fang FC, De Moor B, Vanderleyden J, Marchal K: Comparison of the PhoPQ regulon in *Escherichia coli* and *Salmonella typhimurium*. *J Mol Evol* 2004, 60:462-474.

179. Monsieurs P, Thijs G, Fadda A, De Keersmaecker S, Vanderleyden J, De Moor B, Marchal K: More robust detection of motifs in coexpressed genes by using phylogenetic information. *BMC Bioinformatics* 2006, 7:160.

180. Mooney RA, Darst SA, Landick R: Sigma and RNA polymerase: an on-again, off-again relationship? *Mol Cell* 2005, 20:335-345.

181. Moreno-Hagelsieb G, Collado-Vides J: A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* 2002, 18:S329-S336.

182. Morett E, Segovia L: The sigma 54 bacterial enhancer-binding protein family: mechanism of action and phylogenetic relationship of their functional domains. *J Bacteriol* 1993, 175:6067-6074.

183. Moses AM, Chiang DY, Eisen MB: Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac Symp Biocomput* 2004,324-335.

184. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB: MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol* 2004, 5:R98.

185. Mouslim C, Groisman EA: Control of the *Salmonella ugd* gene by three two-component regulatory systems. *Mol Microbiol* 2003, 47:335-344.

186. Murakami KS, Masuda S, Campbell EA, Muzzin O, Darst SA: Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science* 2002, 296:1285-1290.

187. Nelson DL, Kennedy EP: Magnesium transport in *Escherichia coli*. Inhibition by cobaltous ion. *J Biol Chem* 1971, 246:3042-3049.

188. Nes IF, Diep DB, Havarstein LS, Brurberg MB, Eijsink V, Holo H: Biosynthesis of bacteriocins in lactic acid bacteria. *Antonie Van Leeuwenhoek* 1996, 70:113-128.

189. Neuwald AF, Liu JS, Lawrence CE: Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci* 1995, 4:1618-1632.

190. Novick RP, Muir TW: Virulence gene regulation by peptides in *staphylococci* and other Gram-positive bacteria. *Curr Opin Microbiol* 1999, 2:40-45.

191. Okuda S, Katayama T, Kawashima S, Goto S, Kanehisa M: ODB: a database of operons accumulating known operons across multiple genomes. *Nucleic Acids Res* 2006, 34:D358-D362.

192. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* 1999, 96:2896-2901.

193. Pal C, Papp B, Lercher MJ: Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 2005, 37:1372-1375.

194. Panina EM, Mironov AA, Gelfand MS: Comparative genomics of bacterial zinc regulons: enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. *Proc Natl Acad Sci U S A* 2003, 100:9912-9917.

195. Panina EM, Vitreschak AG, Mironov AA, Gelfand MS: Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria. *FEMS Microbiol Lett* 2003, 222:211-220.

196. Panina EM, Vitreschak AG, Mironov AA, Gelfand MS: Regulation of aromatic amino acid biosynthesis in gamma-proteobacteria. *J Mol Microbiol Biotechnol* 2001, 3:529-543.

197. Panina EM, Mironov AA, Gelfand MS: Comparative analysis of FUR regulons in gamma-proteobacteria. *Nucleic Acids Res* 2001, 29:5195-5206.

198. Papp B, Pal C, Hurst LD: Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet* 2003, 19:417-422.

199. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT et al.: Complete

genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* 2001, 413:848-852.

200. Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M et al.: ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2005, 33:D553-D555.

201. Parra-Lopez C, Baer MT, Groisman EA: Molecular genetic analysis of a locus required for resistance to antimicrobial peptides in *Salmonella typhimurium*. *EMBO J* 1993, 12:4053-4062.

202. Pavesi G, Mauri G, Pesole G: An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* 2001, 17:S207-S214.

203. Pavesi G, Mereghetti P, Mauri G, Pesole G: Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res* 2004, 32:W199-W203.

204. Perez-Rueda E, Collado-Vides J, Segovia L: Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. *Comput Biol Chem* 2004, 28:341-350.

205. Perez-Rueda E, Collado-Vides J: The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res* 2000, 28:1838-1847.

206. Perez-Rueda E, Gralla JD, Collado-Vides J: Genomic position analyses and the transcription machinery. *J Mol Biol* 1998, 275:165-170.

207. Pestova EV, Havarstein LS, Morrison DA: Regulation of competence for genetic transformation in *Streptococcus pneumoniae* by an auto-induced peptide pheromone and a two-component regulatory system. *Mol Microbiol* 1996, 21:853-862.

208. Pevzner PA, Sze SH: Combinatorial approaches to finding subtle signals in DNA sequences. *Proc Int Conf Intell Syst Mol Biol* 2000, 8:269-278.

209. Prakash A, Blanchette M, Sinha S, Tompa M: Motif discovery in heterogeneous sequence data. *Pac Symp Biocomput* 2004,348-359.

210. Prouty AM, Gunn JS: *Salmonella enterica* serovar typhimurium invasion is repressed in the presence of bile. *Infect Immun* 2000, 68:6763-6769.

211. Prouty AM, Brodsky IE, Falkow S, Gunn JS: Bile-salt-mediated induction of antimicrobial and bile resistance in *Salmonella typhimurium*. *Microbiology* 2004, 150:775-783.

212. Pulkkinen WS, Miller SI: A *Salmonella typhimurium* virulence protein is similar to a *Yersinia enterocolitica* invasion protein and a bacteriophage lambda outer membrane protein. *J Bacteriol* 1991, 173:86-93.

213. Qin X, Singh KV, Weinstock GM, Murray BE: Characterization of *fsr*, a regulator controlling expression of gelatinase and serine protease in *Enterococcus faecalis* OG1RF. *J Bacteriol* 2001, 183:3372-3382.

214. Qin ZS, McCue LA, Thompson W, Mayerhofer L, Lawrence CE, Liu JS: Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol* 2003, 21:435-439.

215. Rajewsky N, Socci ND, Zapotocky M, Siggia ED: The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res* 2002, 12:298-308.

216. Rasmussen AA, Eriksen M, Gilany K, Udesen C, Franch T, Petersen C, Valentin-Hansen P: Regulation of *ompA* mRNA stability: the role of a small regulatory RNA in growth phase-dependent control. *Mol Microbiol* 2005, 58:1421-1429.

217. Ren B, Dynlacht BD: Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors. *Methods Enzymol* 2004, 376:304-315.

218. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E et al.: Genome-wide location and function of DNA binding proteins. *Science* 2000, 290:2306-2309.

219. Rey L, Murillo J, Hernando Y, Hidalgo E, Cabrera E, Imperial J, Ruiz-Argueso T: Molecular analysis of a microaerobically induced operon required for hydrogenase synthesis in *Rhizobium leguminosarum* biovar viciae. *Mol Microbiol* 1993, 8:471-481.

220. Robison K, McGuire AM, Church GM: A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J Mol Biol* 1998, 284:241-254.

221. Rodionov DA, Mironov AA, Gelfand MS: Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. *Genome Res* 2002, 12:1507-1516.

222. Roland KL, Martin LE, Esther CR, Spitznagel JK: Spontaneous *pmrA* mutants of *Salmonella typhimurium* LT2 define a new two-component regulatory system with a possible role in virulence. *J Bacteriol* 1993, 175:4154-4164.

223. Rombauts S, Florquin K, Lescot M, Marchal K, Rouze P, Van de PY: Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol* 2003, 132:1162-1176.

224. Roth A, Messer W: The DNA binding domain of the initiator protein DnaA. *EMBO J* 1995, 14:2106-2111.

225. Ruiz-Albert J, Yu XJ, Beuzon CR, Blakey AN, Galyov EE, Holden DW: Complementary activities of SseJ and SifA regulate dynamics of the *Salmonella typhimurium* vacuolar membrane. *Mol Microbiol* 2002, 44:645-661.

226. Sahl HG, Bierbaum G: Lantibiotics: biosynthesis and biological activities of uniquely modified peptides from gram-positive bacteria. *Annu Rev Microbiol* 1998, 52:41-79.

227. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J: Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci U S A* 2000, 97:6652-6657.

228. 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 et al.: RegulonDB (version 5.0): *Escherichia*

coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 2006, 34:D394-D397.

229. Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V, Santos-Zavaleta A, Bonavides-Martinez C et al.: RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res* 2004, 32 Database issue:D303-D306.

230. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C, Collado-Vides J: RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res* 2001, 29:72-74.

231. Salmon K, Hung SP, Mekjian K, Baldi P, Hatfield GW, Gunsalus RP: Global gene expression profiling in *Escherichia coli* K12. The effects of oxygen availability and FNR. *J Biol Chem* 2003, 278:29837-29855.

232. Salmon KA, Hung SP, Steffen NR, Krupp R, Baldi P, Hatfield GW, Gunsalus RP: Global gene expression profiling in *Escherichia coli* K12: effects of oxygen availability and ArcA. *J Biol Chem* 2005, 280:15084-15096.

233. Sandve GK, Drablos F: A survey of motif discovery methods in an integrated framework. *Biol Direct* 2006, 1:11.

234. Schneider T, Stormo G, Gold L, Ehrenfeucht J: Information content of binding sites on nucleotide sequences. *J Mol Biol* 1986, 188:415-431.

235. Schneider TD, Stephens RM: Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 1990, 18:6097-6100.

236. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 2003, 34:166-176.

237. Shen-Orr SS, Milo R, Mangan S, Alon U: Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* 2002, 31:64-68.

238. Siddharthan R, Siggia ED, van Nimwegen E: PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* 2005, 1:e67.

239. Sinha S, Tompa M: Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* 2002, 30:5549-5560.

240. Sinha S, Blanchette M, Tompa M: PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* 2004, 5:170.

241. Smith AD, Sumazin P, Das D, Zhang MQ: Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics* 2005, 21:S403-S412.

242. Smith TF, Waterman MS: Identification of common molecular subsequences. *J Mol Biol* 1981, 147:195-197.

243. Snavely MD, Miller CG, Maguire ME: The *mgtB* Mg^{2+} transport locus of *Salmonella typhimurium* encodes a P-type ATPase. *J Biol Chem* 1991, 266:815-823.

244. Solomon JM, Magnuson R, Srivastava A, Grossman AD: Convergent sensing pathways mediate response to two extracellular competence factors in *Bacillus subtilis*. *Genes Dev* 1995, 9:547-558.

245. Somers WS, Phillips SE: Crystal structure of the met repressor-operator complex at 2.8 A resolution reveals DNA recognition by beta-strands. *Nature* 1992, 359:387-393.

246. Soncini FC, Groisman EA: Two-component regulatory systems can interact to process multiple environmental signals. *J Bacteriol* 1996, 178:6796-6801.

247. Soncini FC, Garcia Vescovi E, Solomon F, Groisman EA: Molecular basis of the magnesium deprivation response in *Salmonella typhimurium*: identification of PhoP-regulated genes. *J Bacteriol* 1996, 178:5092-5099.

248. Soncini FC, Vescovi EG, Groisman EA: Transcriptional autoregulation of the *Salmonella typhimurium phoPQ* operon. *J Bacteriol* 1995, 177:4364-4371.

249. Staden R: Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci* 1989, 5:89-96.

250. Stock AM, Robinson VL, Goudreau PN: Two-component signal transduction. *Annu Rev Biochem* 2000, 69:183-215.

251. Storey JD, Tibshirani R: Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 2003, 100:9440-9445.

252. Stormo GD, Tan K: Mining genome databases to identify and understand new gene regulatory systems. *Curr Opin Microbiol* 2002, 5:149-153.

253. Sturme MH, Kleerebezem M, Nakayama J, Akkermans AD, Vaugha EE, de Vos WM: Cell to cell communication by autoinducing peptides in gram-positive bacteria. *Antonie Van Leeuwenhoek* 2002, 81:233-243.

254. Sweetser D, Nonet M, Young RA: Prokaryotic and eukaryotic RNA polymerases have homologous core subunits. *Proc Natl Acad Sci U S A* 1987, 84:1192-1196.

255. Sze SH, Gelfand MS, Pevzner PA: Finding weak motifs in DNA sequences. *Pac Symp Biocomput* 2002,235-246.

256. Tamayo R, Ryan SS, McCoy AJ, Gunn JS: Identification and genetic characterization of PmrA-regulated genes and genes involved in polymyxin B resistance in *Salmonella enterica* serovar Typhimurium. *Infect Immun* 2002, 70:6770-6778.

257. Tamayo R, Prouty AM, Gunn JS: Identification and functional analysis of *Salmonella enterica* serovar Typhimurium PmrA-regulated genes. *FEMS Immunol Med Microbiol* 2005, 43:249-258.

258. Tan K, Moreno-Hagelsieb G, Collado-Vides J, Stormo GD: A comparative genomics approach to prediction of new members of regulons. *Genome Res* 2001, 11:566-584.

259. Tan K, McCue LA, Stormo GD: Making connections between novel transcription factors and their DNA motifs. *Genome Res* 2005, 15:312-320.

260. Tanay A, Regev A, Shamir R: Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A* 2005, 102:7203-7208.

261. Tatusov RL, Galperin MY, Natale DA, Koonin EV: The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000, 28:33-36.

262. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: Systematic determination of genetic network architecture. *Nat Genet* 1999, 22:281-285.

263. Teichmann SA, Babu MM: Gene regulatory network growth by duplication. *Nat Genet* 2004, 36:492-496.

264. Tetart F, Bouche JP: Regulation of the expression of the cell-cycle gene *ftsZ* by DicF antisense RNA. Division does not require a fixed number of FtsZ molecules. *Mol Microbiol* 1992, 6:615-620.

265. Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J: From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays* 1998, 20:433-440.

266. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y: A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol* 2002, 9:447-464.

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

268. Thijs G: Probabilistic methods to search for regulatory elements in sets of coregulated genes. *PhD thesis* 2003.

269. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y: A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 2001, 17:1113-1122.

270. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994, 22:4673-4680.

271. Tjaden B, Goodwin SS, Opdyke JA, Guillier M, Fu DX, Gottesman S, Storz G: Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res* 2006, 34:2791-2802.

272. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ et al.: Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005, 23:137-144.

273. Tortosa P, Dubnau D: Competence for transformation: a matter of taste. *Curr Opin Microbiol* 1999, 2:588-592.

274. Tossi A, Sandri L, Giangaspero A: Amphipathic, alpha-helical antimicrobial peptides. *Biopolymers* 2000, 55:4-30.

275. Trent MS, Ribeiro AA, Lin S, Cotter RJ, Raetz CR: An inner membrane enzyme in *Salmonella* and *Escherichia coli* that transfers 4-amino-4-deoxy-L-arabinose to lipid A: induction on polymyxin-resistant mutants and role of a novel lipid-linked donor. *J Biol Chem* 2001, 276:43122-43131.

276. Trent MS, Ribeiro AA, Doerrler WT, Lin S, Cotter RJ, Raetz CR: Accumulation of a polyisoprene-linked amino sugar in polymyxin-resistant *Salmonella typhimurium* and *Escherichia coli*: structural characterization and transfer to lipid A in the periplasm. *J Biol Chem* 2001, 276:43132-43144.

277. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* 2001, 98:5116-5121.

278. Udekwu KI, Darfeuille F, Vogel J, Reimegard J, Holmqvist E, Wagner EG: Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes Dev* 2005, 19:2355-2366.

279. Urbanowski ML, Stauffer LT, Stauffer GV: The *gcvB* gene encodes a small untranslated RNA involved in expression of the dipeptide and oligopeptide transport systems in *Escherichia coli*. *Mol Microbiol* 2000, 37:856-868.

280. Valdivia RH, Falkow S: Fluorescence-based isolation of bacterial genes expressed within host cells. *Science* 1997, 277:2007-2011.

281. Valentin-Hansen P, Eriksen M, Udesen C: The bacterial Sm-like protein Hfq: a key player in RNA transactions. *Mol Microbiol* 2004, 51:1525-1533.

282. van Belkum MJ, Worobo RW, Stiles ME: Double-glycine-type leader peptides direct secretion of bacteriocins by ABC transporters: colicin V secretion in *Lactococcus lactis*. *Mol Microbiol* 1997, 23:1293-1301.

283. Van den BT, Van Leemput K, Naudts B, van Remortel P, Ma H, Verschoren A, De Moor B, Marchal K: SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics* 2006, 7:43.

284. van Helden J, Andre B, Collado-Vides J: Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 1998, 281:827-842.

285. van Helden J, Rios AF, Collado-Vides J: Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* 2000, 28:1808-1818.

286. Van Hellemont R, Monsieurs P, Thijs G, De Moor B, Van de Peer Y, Marchal K: A novel approach to identify regulatory motifs in distantly related genomes. *Genome Biol* 2005, 6:R113.

287. van Nimwegen E, Zavolan M, Rajewsky N, Siggia ED: Probabilistic clustering of sequences: inferring new bacterial regulons by comparative genomics. *Proc Natl Acad Sci U S A* 2002, 99:7323-7328.

288. Van Velkinburgh JC, Gunn JS: PhoP-PhoQ-regulated loci are required for enhanced bile resistance in *Salmonella* spp. *Infect Immun* 1999, 67:1614-1622.

289. Vandepoele K, Vlieghe K, Florquin K, Hennig L, Beemster GT, Gruissem W, Van de PY, Inze D, De Veylder L: Genome-wide identification of potential plant E2F target genes. *Plant Physiol* 2005, 139:316-328.

290. Vescovi EG, Ayala YM, Di Cera E, Groisman EA: Characterization of the bacterial sensor protein PhoQ. Evidence for distinct binding sites for Mg^{2+} and Ca^{2+}. *J Biol Chem* 1997, 272:1440-1443.

291. Vogel J, Bartels V, Tang TH, Churakov G, Slagter-Jager JG, Huttenhofer A, Wagner EG: RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res* 2003, 31:6435-6443.

292. von Hippel PH, Berg OG: On the specificity of DNA-protein interactions. *Proc Natl Acad Sci U S A* 1986, 83:1608-1612.

293. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* 2005, 33:D433-D437.

294. Wandersman C: Protein and peptide secretion by ABC exporters. *Res Microbiol* 1998, 149:163-170.

295. Wang T, Stormo GD: Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 2003, 19:2369-2380.

296. Wang T, Stormo GD: Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proc Natl Acad Sci U S A* 2005, 102:17400-17405.

297. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S: Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev* 2001, 15:1637-1651.

298. Wassarman KM: Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes. *Cell* 2002, 109:141-144.

299. Westover BP, Buhler JD, Sonnenburg JL, Gordon JI: Operon prediction without a training set. *Bioinformatics* 2005, 21:880-888.

300. Whitehead NA, Barnard AM, Slater H, Simpson NJ, Salmond GP: Quorum-sensing in Gram-negative bacteria. *FEMS Microbiol Rev* 2001, 25:365-404.

301. Workman CT, Stormo GD: ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput* 2000,467-478.

302. Wosten MM: Eubacterial sigma-factors. *FEMS Microbiol Rev* 1998, 22:127-150.

303. Wosten MM, Groisman EA: Molecular characterization of the PmrA regulon. *J Biol Chem* 1999, 274:27185-27190.

304. Wosten MM, Kox LF, Chamnongpol S, Soncini FC, Groisman EA: A signal transduction system that responds to extracellular iron. *Cell* 2000, 103:113-125.

305. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* 2003, 20:1377-1419.

306. Wyrick JJ, Young RA: Deciphering gene expression regulatory networks. *Curr Opin Genet Dev* 2002, 12:130-136.

307. Xavier KB, Bassler BL: Interference with AI-2-mediated bacterial cell-cell communication. *Nature* 2005, 437:750-753.

308. Xavier KB, Bassler BL: LuxS quorum sensing: more than just a numbers game. *Curr Opin Microbiol* 2003, 6:191-197.

309. Xing EP, Karp RM: MotifPrototyper: a Bayesian profile model for motif families. *Proc Natl Acad Sci U S A* 2004, 101:10523-10528.

310. Yada T, Nakao M, Totoki Y, Nakai K: Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics* 1999, 15:987-993.

311. Yamamoto K, Ogasawara H, Fujita N, Utsumi R, Ishihama A: Novel mode of transcription regulation of divergently overlapping promoters by PhoP, the regulator of two-component system sensing external magnesium availability. *Mol Microbiol* 2002, 45:423-438.

312. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M: Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res* 2004, 14:1107-1118.

313. Yu H, Luscombe NM, Qian J, Gerstein M: Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet* 2003, 19:422-427.

314. Zhang YM, Marrakchi H, Rock CO: The FabR (YijC) transcription factor regulates unsaturated fatty acid biosynthesis in *Escherichia coli*. *J Biol Chem* 2002, 277:15558-15565.

315. Zhao Y, Jansen R, Gaastra W, Arkesteijn G, van der Zeijst BA, van Putten JP: Identification of genes affecting *Salmonella enterica* serovar enteritidis infection of chicken macrophages. *Infect Immun* 2002, 70:5319-5321.

316. Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S: Computational identification of operons in microbial genomes. *Genome Res* 2002, 12:1221-1230.

317. Zhou Z, Ribeiro AA, Lin S, Cotter RJ, Miller SI, Raetz CR: Lipid A modifications in polymyxin-resistant *Salmonella typhimurium*: PmrA-dependent 4-amino-4-deoxy-L-arabinose, and phosphoethanolamine incorporation. *J Biol Chem* 2001, 276:43111-43121.

Curriculum vitae

Pieter Monsieurs was born in Mol, Belgium, on April 18th, 1979. In 1997, he started his education in applied biological sciences at the K.U. Leuven, where he received the Candidacy diploma in Bioscience Engineering in 1999, and the Masters diploma in Cellular and Biotechnological Engineering in 2002. Since October 2002 he has been pursuing his PhD as a Research Assistant in the research group ESAT-SCD, under the supervision of Prof. Bart De Moor and Prof. Kathleen Marchal.