New statistical Methods of Genome-Scale Data Analysis in Life Science - Applications to enterobacterial Diagnostics, Meta-Analysis of *Arabidopsis thaliana* Gene Expression and functional Sequence Annotation

Neue statistische Methoden für genomweite Datenanalysen in den Biowissenschaften - Anwendungen in der Enterobakteriendiagnostik, Meta-Analyse von *Arabidopsis thaliana* Genexpression und funktionsbezogenen Sequenzenannotation

Doctoral thesis for a degree at the Graduate School of Life Sciences, Julius-Maximilians-Universität Würzburg, Section Infection and Immunity, ZINF/IMIB

submitted by

Torben Friedrich

from

Buchholz in der Nordheide

Würzburg 2009

Submitted on: . . . .

Office stamp

Members of the *Promotionskomitee*:

Chairperson: . . . . Prof. Dr. Manfred Gessler

Primary Supervisor: . . . . Prof. Dr. Drs. h. c. Jörg Hacker

Supervisor (Second): . . . . Prof. Dr. Thomas Dandekar

Supervisor (Third): . . . . Prof. Dr. Sven Rahmann

Supervisor (Fourth): . . . . PD Dr. Ulrich Dobrindt

Date of Public Defence: . . . .

Date of receipt of Certificates: . . . .

Danksagung

Ganz besonders möchte ich Tobias Müller und Ulrich Dobrindt für eine exzellente Betreuung, viele hilfreiche Diskussionen und gute Ideen sowie Ratschläge danken. Mein Dank gilt auch meinem Promotionskomitee. Thomas Dandekar hat nicht nur mit wertvollen Diskussionen zu allen Teilbereichen der Arbeit, sondern auch mit der Bereitstellung eines hervorragenden Arbeitsumfeldes sehr unterstützt. Jörg Hacker hat den Prozess der Doktorarbeit beratend begleitet und wertvolle Ideen einfließen lassen. Durch die vielen bestehenden Kontakte war eine ideale Ausgangsbasis bezüglich der Datenbeschaffung gegeben. Sven Rahmann danke ich für die Kooperation bei der Entwicklung des diagnostischen Microarrays und als Mitglied des Promotionskomitees. Er hat mit vielen Anregungen und seiner Erfahrung in den Bereichen Microarray-Design und -Auswertung zu dem Projekt beigetragen.

Finanzielle Unterstützung für die Arbeit kam von dem Forschungsverbund FORINGEN (TP-A1) der Bayerischen Forschungsstiftung. Für die Möglichkeit, die Testhybridisierungen bei der Scienion AG durchführen zu können, gilt mein besonderer Dank den dortigen direkten Ansprechpartnern Wilfried Weigel und Sabine Fischer. Für die Erlaubnis zur Verwendung der unveröffentlichten Genomsequenzen von *S. bongori* 12419 und *E. coli* O42 gilt mein Dank Julian Parkhill vom Sanger Institut, Cambridge, U.K..

Danken möchte ich auch allen Kollegen in der Bioinformatik und Infektionsbiologie, mit denen ich im Laufe der Doktorarbeit zusammengearbeitet habe. Sie haben wesentlich zu einem produktiven und kurzweiligen Arbeitsumfeld beigetragen, das in beiden Instituten gegeben war. Barbara und Markus möchte ich für das Korrekturlesen der Arbeit danken. Außerdem gilt mein Dank Julia Engelmann, Daniela Beißer, Juilee Thakar, Christian Koetschan, Philipp Seibel für eine tolle gemeinsame Zeit im Raum 117.

Aber vor allem möchte ich meinen Eltern danken, die mir mit ihrer uneingeschränkten Unterstützung das Studium und damit diese Doktorarbeit überhaupt erst möglich gemacht. Lydia und Peter danke ich für die Möglichkeit, ihr ‘Freiluftbüro’ nutzen zu dürfen und für ihre Unterstützung in allen Belangen. Barbara hat die Doktorandenzeit mit der richtigen Mischung aus Ablenkung, Führsorge und Geduld unvergesslich werden lassen.

Zusammenfassung

Die aktuellen Fortschritte und Entwicklungen in der Molekularbiologie stellen eine Fülle neuer, bisher kaum analysierter Daten bereit. Dieser Fundus umfasst unter Anderem biologische Daten zu genomischer DNA, zu Proteinsequenzen, zu dreidimensionalen Proteinstrukturen sowie zu Genexpressionsprofilen. In der vorliegenden Arbeit werden diese Informationen genutzt, um neue Methoden der Charakterisierung und Klassifizierung von Organismen bzw. Organismengruppen zu entwickeln und einen automatisierten Informationsgewinn sowie eine Informationsübertragung zu ermöglichen.

Die ersten beiden vorgestellten Ansätze (Kapitel 4 und 5) konzentrieren sich auf die medizinisch und wissenschaftlich bedeutsame Gruppe der Enterobakterien. Deren Bedeutung für Medizin und Mikrobiologie geht auf ihre Funktion als kommensale Bewohner des Darmtraktes, ihre Nutzung als leicht kultivierbare Modellorganismen und auf die vielseitigen Infektionsmechanismen zurück. Obwohl bereits viele Studien über einzelne Pathogruppen mit klinisch unterscheidbaren Symptomen existieren, sind die genotypischen Faktoren, die für diese Unterschiedlichkeit verantwortlich zeichnen, teilweise noch nicht bekannt. Der in **Kapitel 4** beschriebene umfassende Genomvergleich wurde anhand einer Vielzahl von Enterobakterien durchgeführt, die nahezu die gesamte Bandbreite klinisch relevanter Diversität darstellen. Dieser Genomvergleich bildet die Basis für eine Charakterisierung des enterobakteriellen Genpools, für eine Rekonstruktion evolutionärer Prozesse und Einflüsse und für eine umfassende Untersuchung spezifischer Proteinfamilien in enterobakteriellen Untergruppen. Die in diesem Kontext vorher noch nicht angewandte Korrespondenzanalyse liefert qualitative Aussagen zu bakteriellen Untergruppen und den ausschließlich in ihnen vorkommenden Proteinfamilien. In drei Hauptuntergruppen der Enterobakterien, die den Gattungen *Yersinia* und *Salmonella* sowie der Gruppe aus *Shigella* und *E. coli* entsprechen, wurden die jeweils spezifischen Proteinfamilien mit Hilfe statistischer Tests identifiziert. Zusammenfassend bilden die auf Genomvergleichen aufbauenden Methoden neue Ansatzpunkte, um aus der Übertragung der bekannten Funktionalität einzelner Proteine auf spezifische, genotypische Besonderheiten bakterieller Gruppen zu schließen.

Aufgrund ihrer hohen medizinischen Relevanz war die Typisierung enterobakterieller Isolate entsprechend ihrer Pathogenität Ziel zahlreicher Studien. Die Microarray-Technologie bietet ein schnelles, reproduzierbares und standardisierbares Hilfsmittel für bakterielle Typisierung und hat sich in der Bakteriendiagnostik, Risikobewertung und Überwachung bewährt. Das in **Kapitel 5** beschriebene Design eines diagnostischen Microarray beruht auf einer großen Anzahl verfügbarer Genomsequenzen von Enterobakterien. Ein hocheffizienter String-Matching-Algorithmus ist die Grundlage einer neuartigen Strategie der Sondenauswahl, die sowohl kodierende als auch nicht-kodierende Berei-

che genomischer DNA berücksichtigt. Im Vergleich zu Diagnostika, die ausschließlich auf Virulenz-assoziierten Sonden beruhen, verringert dieses Prinzip das Risiko einer inkorrekten Typisierung. Zusätzliche Sonden erweitern das Anwendungsspektrum auf eine simultane Diagnostik der Antibiotikaresistenz bzw. eine Überwachung der Resistenzausbreitung.

Umfangreiche Testhybridisierungen belegen eine überwiegende Zuverlässigkeit der Sonden und vor allem eine robuste Klassifizierung enterobakterieller Stämme entsprechend der Pathogruppen. Die Tests bilden zudem die Grundlage für das Training eines Regressionsmodells zur Klassifizierung der Pathogruppe und zur Vorhersage der Menge hybridisierter DNA. Das Regressionsmodell zeichnet sich durch kontinuierliche Lernfähigkeit und damit durch eine Verbesserung der Vorhersagequalität im Prozess der Anwendung aus. Ein Teil der Sonden repräsentiert intergenische DNA und bestätigt infolgedessen die Relevanz der zugrunde liegenden Strategie. Die Tatsache, dass ein großer Teil der von den Sonden repräsentierten Gene noch nicht annotiert ist, legt die Existenz bisher unentdeckter Faktoren mit Bedeutung für die Ausbildung entsprechender Virulenz-Phänotypen nahe.

Ein weiteres Haupteinsatzgebiet von Microarrays ist die Genexpressionsanalyse. Die Größe von Genexpressionsdatenbanken ist in den vergangenen Jahren stark gewachsen. Obwohl sie eine Fülle von Expressionsdaten bieten, sind Ergebnisse aus unterschiedlichen Studien weiterhin schwer in einen übergreifenden Zusammenhang zu bringen. In **Kapitel 6** wird die Methodik einer ausschließlich datenbasierten Meta-Analyse für genomweite *A. thaliana* Genexpressionsdatensätze dargestellt, die neue Erkenntnisse über Funktion und Regulation von Genen verspricht. Die Anwendung von Kernel-basierter Hauptkomponentenanalyse in Kombination mit hierarchischem Clustering identifizierte drei Hauptgruppen von Kontrastexperimenten mit jeweils überlappenden Expressionsmustern. In zwei Gruppen konnten deregulierte Gene wichtigen Funktionen bei Indol-3-Essigsäure (IAA) vermitteltem Pflanzenwachstum und -entwicklung sowie pflanzlicher Pathogenabwehr zugeordnet werden. Bisher funktionell nicht näher charakterisierte Serin-Threonin-Kinasen wurden über die Meta-Analyse mit der Pathogenabwehr assoziiert. Grundsätzlich kann dieser Ansatz versteckte Wechselbeziehungen zwischen Genen aufdecken, die unter verschiedenen Bedingungen reguliert werden.

Bei der funktionellen Charakterisierung von Proteinen oder der Vorhersage von Genen in Genomsequenzen werden Hidden-Markov-Modelle (HMMs) eingesetzt. HMMs sind technisch ausgereift und in der computergestützten Biologie vielfach eingesetzt worden. Trotzdem birgt die Methodik das Potential zur Optimierung bezüglich der Modellierung biologischer Daten, die hinsichtlich der Längenverteilung ihrer Sequenzen variieren.

Untereinheiten dieser Modelle, die Zustände, repräsentieren über ihre individuelle Verweildauer zugrunde liegende Verteilungen von Sequenzlängen. **Kapitel 7** stellt eine Methode zur Anpassung einfacher HMM-Topologien an biologische Daten, die glockenkurvenartige Längenverteilungen zeigen, vor. Die Modellierung solcher Verteilungen wird dabei durch eine serielle Verkettung vervielfältigter Zustände gewährleistet, ohne dass die Klasse herkömmlicher HMMs verlassen wird. Auswertungen der Modellierungsleistung bei unterschiedlich stark optimierten HMM-Topologien unterstreichen die Bedeutung der entwickelten Topologieoptimierung. Zusammenfassend wird hier eine generelle Methodik beschrieben, die die Modelleigenschaften von HMMs über Topologieoptimierungen verbessert. Die Parameter dieser Optimierung werden mit Hilfe von Maximum-Likelihood und einem leicht einzubindenden Momentschätzer bestimmt.

In **Kapitel 8** wird die Anwendung von HMMs zur Vorhersage von Interaktionsstellen in Proteindomänen beschrieben. Wie bereits gezeigt wurde, sind solche Stellen aufgrund einer variablen Konserviertheit ihrer Position und ihres Typs schwer zu bestimmen. Eine Vorhersage von Interaktionstellen in Proteindomänen wird über die Definition einer neuen HMM-Topologie erreicht, die sowohl Sequenz- als auch Strukturdaten einbindet. Interaktionsstellen werden mit einem Posterior-Decoding-Algorithmus vorhergesagt, der zusätzliche Informationen über die Wahrscheinlichkeit einer Interaktion für alle Sequenzpositionen bereitstellt. Die Implementierung der Interaktionsprofil-HMMs (ipHMMs) basiert auf den etablierten Profil-HMMs und erbt deren Effizienz und Sensitivität. Eine groß angelegte Vorhersage von Interaktionsstellen mit ipHMMs konnte mutationsbedingte Fehlfunktionen in Proteinen erklären, die mit vererbbaren Krankheiten wie unterschiedlichen Tumortypen oder Muskeldystrophie assoziiert sind. Wie Profile-HMMs sind auch ipHMMs für groß angelegte Anwendungen geeignet. Insgesamt verbessert die HMM-gestützte Methode sowohl die Vorhersagequalität für Interaktionsstellen als auch das Verständnis molekularer Hintergründe bei vererbbaren Krankheiten.

Im Hinblick auf aktuelle und zukünftige Anforderungen stelle ich in dieser Arbeit Lösungsansätze für eine umfassende Charakterisierung großer Mengen biologischer Daten vor. Alle beschriebenen Methoden zeichnen sich durch gute Übertragbarkeit auf verwandte Probleme aus. Besonderes Augenmerk wurde dabei auf den Wissenstransfer gelegt, der durch einen stetig wachsenden Fundus biologischer Information ermöglicht wird. Die angewandten und entwickelten statistischen Methoden sind lernfähig und profitieren von diesem Wissenszuwachs, Vorhersagequalität und Zuverlässigkeit der Ergebnisse verbessern sich.

Abstract

Recent progresses and developments in molecular biology provide a wealth of new but insufficiently characterised data. This fund comprises amongst others biological data of genomic DNA, protein sequences, 3-dimensional protein structures as well as profiles of gene expression. In the present work, this information is used to develop new methods for the characterisation and classification of organisms and whole groups of organisms as well as to enhance the automated gain and transfer of information.

The first two presented approaches (chapters 4 und 5) focus on the medically and scientifically important enterobacteria. Its impact in medicine and molecular biology is founded in versatile mechanisms of infection, their fundamental function as a commensal inhabitant of the intestinal tract and their use as model organisms as they are easy to cultivate. Despite many studies on single pathogroups with clinical distinguishable pathologies, the genotypic factors that contribute to their diversity are still partially unknown. The comprehensive genome comparison described in **Chapter 4** was conducted with numerous enterobacterial strains, which cover nearly the whole range of clinically relevant diversity. The genome comparison constitutes the basis of a characterisation of the enterobacterial gene pool, of a reconstruction of evolutionary processes and of comprehensive analysis of specific protein families in enterobacterial subgroups. Correspondence analysis, which is applied for the first time in this context, yields qualitative statements to bacterial subgroups and the respective, exclusively present protein families. Specific protein families were identified for the three major subgroups of enterobacteria namely the genera *Yersinia* and *Salmonella* as well as to the group of *Shigella* and *E. coli* by applying statistical tests. In conclusion, the genome comparison-based methods provide new starting points to infer specific genotypic traits of bacterial groups from the transfer of functional annotation.

Due to the high medical importance of enterobacterial isolates their classification according to pathogenicity has been in focus of many studies. The microarray technology offers a fast, reproducible and standardisable means of bacterial typing and has been proved in bacterial diagnostics, risk assessment and surveillance. The design of the diagnostic microarray of enterobacteria described **in chapter 5** is based on the availability of numerous enterobacterial genome sequences. A novel probe selection strategy based on the highly efficient algorithm of string search, which considers both coding and non-coding regions of genomic DNA, enhances pathogroup detection. This principle reduces the risk of incorrect typing due to restrictions to virulence-associated capture probes. Additional capture probes extend the spectrum of applications of the microarray to simultaneous diagnostic or

surveillance of antimicrobial resistance.

Comprehensive test hybridisations largely confirm the reliability of the selected capture probes and its ability to robustly classify enterobacterial strains according to pathogenicity. Moreover, the tests constitute the basis of the training of a regression model for the classification of pathogroups and hybridised amounts of DNA. The regression model features a continuous learning capacity leading to an enhancement of the prediction accuracy in the process of its application. A fraction of the capture probes represents intergenic DNA and hence confirms the relevance of the underlying strategy. Interestingly, a large part of the capture probes represents poorly annotated genes suggesting the existence of yet unconsidered factors with importance to the formation of respective virulence phenotypes.

Another major field of microarray applications is gene expression analysis. The size of gene expression databases rapidly increased in recent years. Although they provide a wealth of expression data, it remains challenging to integrate results from different studies. In **chapter 6** the methodology of an unsupervised meta-analysis of genome-wide *A. thaliana* gene expression data sets is presented, which yields novel insights in function and regulation of genes. The application of kernel-based principal component analysis in combination with hierarchical clustering identified three major groups of contrasts each sharing overlapping expression profiles. Genes associated with two groups are known to play important roles in Indol-3 acetic acid (IAA) mediated plant growth and development as well as in pathogen defence. Yet uncharacterised serine-threonine kinases could be assigned to novel functions in pathogen defence by meta-analysis. In general, hidden interrelation between genes regulated under different conditions could be unravelled by the described approach.

HMMs are applied to the functional characterisation of proteins or the detection of genes in genome sequences. Although HMMs are technically mature and widely applied in computational biology, I demonstrate the methodical optimisation with respect to the modelling accuracy on biological data with various distributions of sequence lengths.

The subunits of these models, the states, are associated with a certain holding time being the link to length distributions of represented sequences. An adaptation of simple HMM topologies to bell-shaped length distributions described in **chapter 7** was achieved by serial chain-linking of single states, while residing in the class of conventional HMMs. The impact of an optimisation of HMM topologies was underlined by performance evaluations with differently adjusted HMM topologies. In summary, a general methodology was introduced to improve the modelling behaviour of HMMs by topological optimisation with maximum likelihood and a fast and easily implementable moment estimator.

Chapter 8 describes the application of HMMs to the prediction of interaction sites in protein domains. As previously demonstrated, these sites are not trivial to predict because of varying degree in conservation of their location and type within the domain family. The prediction of interaction sites in protein domains is achieved by a newly defined HMM topology, which incorporates both sequence and structure information. Posterior decoding is applied to the prediction of interaction sites providing additional information of the probability of an interaction for all sequence positions. The implementation of interaction profile HMMs (ipHMMs) is based on the well established profile HMMs and inherits its known efficiency and sensitivity. The large-scale prediction of interaction sites by ipHMMs explained protein dysfunctions caused by mutations that are associated to inheritable diseases like different types of cancer or muscular dystrophy. As already demonstrated by profile HMMs, the ipHMMs are suitable for large-scale applications. Overall, the HMM-based method enhances the prediction quality of interaction sites and improves the understanding of the molecular background of inheritable diseases.

With respect to current and future requirements I provide large-scale solutions for the characterisation of biological data in this work. All described methods feature a highly portable character, which allows for the transfer to related topics or organisms, respectively. Special emphasis was put on the knowledge transfer facilitated by a steadily increasing wealth of biological information. The applied and developed statistical methods largely provide learning capacities and hence benefit from the gain of knowledge resulting in increased prediction accuracies and reliability.

Contents

List of Figures

List of Tables

Bioinformatical concepts of genomics, global evaluation of gene expression and sequence analysis

Modern biomedical research aims at a comprehensive understanding of entire systems. These systems correspond e. g. to metabolic networks, whole organisms, ecological communities or host pathogen interactions. Rapidly evolving high-throughput technologies like mass-spectrometry, DNA-sequencing or microarrays facilitate the realisation of such ambitious goals. These studies demand for time-efficient bioinformatical solution in planning, preparation and evaluation stages of experiments that mass-produced biological data. Initial milestones reached by the application of novel technologies comprise the generation of large collections of complete genome sequences, genes, proteins, protein domain representations, microarray experiments and structure information of proteins and protein complexes. The tremendous fund of information holds sources to raise new questions for the purpose of a deeper understanding of biological systems and their interdependencies.

In the following, different approaches will be described that aim at supporting the gain of knowledge from the variety of available sequence data. The application and combination of methods of statistics and sequence analysis yielded fundamental insights into characteristics of the important enterobacterial family harbouring many human and animal pathogens as well as model organisms in genetics. Meta-analysis of comparative genome hybridisation (CGH) of *Arabidopsis thaliana* unravelled genes involved in plant pathogen defence and plant growth. In basic studies, HMMs were methodically extended for the purpose of enhanced generality and the prediction of structural features in proteins. All these approaches imply a portable character, which enables its application to a wide range of organisms. The heterogeneity of these topics suggests a subdivision of the results part according to the following single projects:

Chapter 4 Enterobacterial strains were compared based on their genomic content by applying multivariate and statistical methods. They mediate the detection of characteristic genotypes that contributed to phenotypic divergence.

Chapter 5 In reference to the importance of enterobacteria as family of major human pathogens, genomic information was evaluated to develop a diagnostic microarray.

Chapter 6 A general meta-analysis concept was applied to the well studied plant model organism *A. thaliana*. The method profits from the existence of many publicly available CGH experiments. The experimental data was coherently compared to detect recurring patterns of differential expression that get lost in the analysis of single contrasts.

Chapter 7 Hidden Markov models are applied to model biological sequences. HMMs natively represent sequences of geometrically shaped length distributions. In order to overcome modelling deficiencies with respect to otherwise distributed sequences, the HMM architecture was optimised based on a moment estimator.

Chapter 8 Increasing availability of HMMs representing functional subunits of proteins as well as structural information of protein ligand complexes was fused and served as training data in the development of an HMM-based prediction method of protein interaction sites.

The first part covers the investigation of bacteria from the family of *Enterobacteriaceae*, a versatile bacterial taxon comprising many pathogens as well as commensals of eukaryotic hosts. Below, the bacterial family is introduced by focusing on biomedical aspects relevant in clinical therapeutics and diagnostics.

Enterobacteria and *E. coli* pathotypes

In reference to many outbreaks and large number of annual cases, enterobacteria constitute major problems of health care in developing countries. This branch of the gram-negative γ-proteobacteria have been in focus of numerous scientific research projects throughout many years. The widespread scientific interest in enterobacteria is attracted by a large variety of routes to colonise niches in a broad range of vertebrate hosts. Prominent scientific model organisms in genetics and molecular biology like the *E. coli* strain K-12 MG1655 allow simple *in vitro* cultivation and genetic manipulation. Among the *Enterobacteriaceae* Several pathogens of the genera *Salmonella*, *Yersinia*, *Klebsiella* and *Escherichia*, which differ in pathogenicity, origin and natural reservoir, are known and will be shortly characterised in the following.

Salmonellae The genus *Salmonella* consists of bacterial pathogens which are capable to infect a wide range of animal hosts. Common reservoirs of infectious agents like the isolate *S. bongori* 12419 are reptiles and amphibians. Many described serovars have generated a quite complex salmonellae nomenclature (Brenner *et al.*, 2000). Most serovars belong to the species *S. enterica*, which is subdivided into 6 subspecies with further subdivisions referring to antigenic formulae. Subspecies *enterica* is usually characterised by a habitat in warm-blooded hosts, while the other five subspecies are found in cold-blooded animals. In medical context, typhoid *S. enterica* serovars Typhi and Paratyphi are correlated to human-restricted symptoms like enteric fever. Non-typhoid serovars Typhimurium and Enteriditis are found in a broad range of hosts causing gastroenteritis (Haraga *et al.*, 2008). Typhoid fever predominantly occurs in Asian and African developing countries as a cause of contaminated water supply. The last outbreak in the republic of Congo (2004/2005) resulted in 42,564 cases of typhoid fever and 214 deaths (WHO, 2009).

Yersinia The genus *Yersinia* is known from three pandemic outbreaks of plaque caused by *Y. pestis*. It comprises two further species, *Y. pseudotuberculosis* and *Y. enterocolitica*. DNA-DNA hybridisation revealed a close relationship between *Y. pestis* and *Y. pseudotuberculosis*, though the former is transmitted by fleas and causes bubonic plaque, while the latter and *Y. enterocolitica* are enteropathogenic yersiniae transmitted by fecal or oral routes and normally do not lead to death. (Achtman *et al.*, 1999; Wren, 2003)

Klebsiella The genus *Klebsiella* mainly consists of commensal or soil bacteria. Merely the species *K. pneumoniae* and rarely *K. oxytoca* have been described as facultative pathogens. Both may cause infections of the urinary or respiratory tract in humans. Recently, cases of liver abscess were reported. Furthermore, hospitality acquired nosocomial infections frequently trace back to *Klebsiella* pathogens. (Brisse *et al.*, 2006)

Escherichia coli *E. coli* strains normally are commensal inhabitants of the human gastrointestinal tract. The colonisation of the gut, which begins a few hours after birth, plays an important role in human digestion. Other lineages act as pathogens in humans and warm-blooded animals. Three general clinical syndromes result from *E. coli* infections: diarrhoeal infections caused by intestinal pathogens (IPEC), urinary tract infections and sepsis or meningitis originating from colonisation with extraintestinal pathogens (ExPEC). Many research projects have focused on the versatile virulence mechanisms enabling the pathogens to establish an infection in their hosts. The differences in virulence mechanisms lead to the definition of numerous pathotypes. In the following the main characteristics associated with these pathotypes are shortly reviewed. They are described in more detail elsewhere (Nataro and Kaper, 1998; Kaper *et al.*, 2004).

The first described intestinal pathotype has been the enteropathogenic group of *E. coli* (EPEC). Common symptoms of EPEC infections are potentially fatal infant diarrhoea predominantly in developing countries. The mode of infection was termed 'attaching and effacing' (A/E) and is characterised by the formation of microcolonies, attachment to epithelial cells and induction of the reformation of microvilli to pedestal-like structures around attached bacteria. Host cell manipulation is probably mediated by secretion of effectors via the type-III-secretion system. Diarrhoea is caused by the injection of enterotoxins.

Enterohaemorrhagic *E. coli* (EHEC) represent another intestinal pathogroup originally inhabiting the bovine intestinal tract. The most common path of infection is contaminated food. Clinical symptoms upon human infection are diarrhoea and haemolytic uremic syndrome (HUS). Frequently occurring bloody diarrhoea is caused by the injection of a shiga-like toxin. Some EHEC additionally contain the locus of enterocyte effacement (LEE), which encodes for virulence genes inducing pedestal formation in host cells.

Enterotoxigenic *E. coli* (ETEC) cause mild as well as severe forms of watery diarrhoea with high rates of infant infections. ETEC infections are highly prevalent in developing countries and - like other types of enterobacterial infections - rarely occur in industrialised parts of the world. ETEC attach to cells of the small bowel mucosa and release heat-stable and/or heat-labile enterotoxins. Toxin β-subunits bind to receptors on the host cell surface, α-subunits induce an increased ion secretion.

A frequent cause of persistent diarrhoea in children and adults worldwide are enteroaggregative *E. coli* (EAEC). The pathotype is characterised by adherence to the intestinal mucosa in an autoaggregative stacked-brick fashioned biofilm. This aggregation seems to lead to mild mucosal damage. No constant equipment of virulence factors could be determined throughout EAEC isolates.

In contrast to previously described intestinal pathotypes, enteroinvasive *E. coli* (EIEC) are capable to enter host cells, lyse the endocytic vacuole, grow intracellularly and spread to neighbouring cells. The most frequent symptom of EIEC infections is watery diarrhoea. EIEC rarely cause dysentery, which is characterised by fever, abdominal cramps and diarrhoea. By injection of IpaABC and IpgD proteins via a type-III-secretion system EIEC induce epithelial signalling events, cytosceletal rearrangements, cellular uptake and lysis of the endocytic vacuole.

The most frequent cause of urinary tract infections are uropathogenic *E. coli* (UPEC). An infection is thought to begin with the colonisation of the bowel and the periurethral area of even immunocompetent hosts. The bacteria then ascend the urethra to the bladder and attach via F1-fimbriae to epithelial cells. Invasion of cells and formation of biofilms as source of recurrent infection are reported. UPEC strains are commonly equipped with virulence factors like haemolysin, cytotoxic necrotising factor or special adhesins that generally are not found in intestinal pathotypes.

Another extraintestinal pathotype comprises isolates of patients with new-born meningitis (MNEC). These *E. coli* spread haematogenously, translocate the blood-brain barrier without observable damage and adhere to the microvascular endothelium of the brain by S-fimbriae. An increasing incidence and mortality rates between 15% and 40% are reported for neonatal infection. Further pathotypes like sepsis-associated *E. coli* (SEPEC) and avian pathogenic *E. coli* (APEC) are part of the ExPEC group.

Shigella *Shigella* isolates are highly similar to EIEC regarding their patogenicity. The genus *Shigella* is now seen as a clonal lineage of *E. coli* (Lan and Reeves, 2002). Due to historical reason they maintained the position as a separate genus subdivided into the species *S. dysenteriae*, *S. flexneri*, *S. sonnei* and *S. boydii*. These species cause varying degrees of dysentery. Infections are acquired by the oral-fecal route and manifest in the colon or rectum where bacterial cells cross the epithelial barrier, enter macrophages, disrupt the membranes of phagosomes and reside in the cytosol. There the *Shigella* pathogens multiplicate and induce rapid cell death. (Ogawa *et al.*, 2008)

Comparative enterobacterial genomics

Through the years many different methods have been proposed to establish systems for bacterial phylogenies, strain typing and classification. But, difficulties in achieving general typing concepts arise due to a larger genetic variability in bacteria as compared to eukaryotes (Hacker and Carniel, 2001). Therefore, bacteriologists have established bacteria-by-bacteria solutions for subtyping based on various mechanisms, which changed with the rapid development of lab technologies. The *Enterobacteriaceae* are a good example for separately developed nomenclatures. The genus/species/strain concept can be found for the *E. coli* and *Klebsiella* clades, while *Salmonella* and *Yersinia* nomenclatures are extended by subspecies levels. An extra *Shigella* genus was introduced for a clonal *E. coli* lineage that was separately discovered and never fused with *E. coli* to one entire clade.

Former practice of microbial strain typing relied on the determination of phenotypic traits like the O-, H- and K-antigens. The development of nucleotide based technologies changed the common practice to more accurate genotypic reconstructions of bacterial strain typing and phylogeny. Initially, genotypic methods consist solely in the determinations of nucleotide polymorphisms within single commonly occurring genes. Among these the gene encoding the small ribosomal 16S rRNA subunit became the major determinant of macroscopic bacterial evolution. Even purposes like the reconstruction of the tree of life favour the use of 16S rRNA as it can be compared with eukaryotic 18S rRNA genes (Clarridge, 2004). Nevertheless, 16S rRNA phylogeny exhibited a lower resolution and robustness in species sublevels, because of the existence of multiple gene copies with different evolutionary background in bacterial genomes. Thus, Case *et al.* (2007) suggested the *rpoB* gene as phylogenetic determinant to compensate for these shortcomings in enterobacteria. Concurrently, a method based on multiple genes termed multi-locus sequence typing (MLST) was developed. The method was initially applied to *Neisseria* strain typing and comprised fragments of 11 housekeeping genes sized between 417 to 579 bp. The gene loci of the fragments were distributed across the whole genome to ensure that no co-inheritance contributed to single transformation events. So called sequence types (ST) were determined by the patterns of single nucleotide polymorphisms in the gene fragments. Later on the method was adapted to other bacteria. MLST transfer to *E. coli* comprised sequence fragments of the genes *arcA*, *aroE*, *dnaE*, *mdh*, *gnd*, *gapA*, *pgm*, *espA* and *ompA* (Maiden *et al.*, 1998). Recently, Wirth *et al.* (2006) redesigned the MLST determinants, and a large screening of sequence types was performed to set up a MLST database for *E. coli*. The rapidly changing techniques in bacterial strain typing again reflect the difficulties linked with the definition of an overall methodology.

Most recently, the availability of complete genomic sequences and of derived proteomes dramatically increased and enabled more detailed insights into bacterial evolution. The first comparisons of whole genomes or its proteomes were restricted to a basic set of already sequenced organisms like *H. sapiens*, *M. musculus*, *E. coli* or *B. subtilis*. Several approaches were developed to achieve comparability between these distantly related taxa as a basis for comparative analysis. Tatusov *et al.* (1997) established a method to cluster orthologuous groups of proteins (COG) across the three domains of life. They defined a COG as a cluster with at least three members of different phylogenetic domains. The members of a COG have to exhibit reciprocally highest similarity among all proteins of respective organisms in all-against-all sequence alignments. Another study considered the different genome sizes to determine organism specific similarity thresholds. The obtained orthology assignment was then subjected to factorial analysis in order to correlate the organisms according to ancestry (Tekaia *et al.*, 1999). Protein clustering has been in focus of several studies in recent years. The OrthoMCL programme refined previously introduced criteria in order to perform a clustering of orthologous proteins on multiple organisms. The programme employs the Markov cluster algorithm (Enright *et al.*, 2002) to determine orthology or recent paralogy, respectively (Li *et al.*, 2003).

In the methods described in chapter 4 the concept of multiple genome comparison was taken up. It was extended by multivariate analysis and statistical testing to unravel traits of specificity for subgroups of the taxonomic family of *Enterobacteriaceae*. The applied methods comprise unsupervised correlation of strain- and protein-wise differences based on an assignment of protein family presence across a variety of enterobacterial strains with diverse phenotypes. In a second step specific proteins were investigated in enterobacterial subgroups, which were inferred by whole proteome comparisons. Based on this prior knowledge, conserved protein families of evolutionary related groups were determined by applying statistical testing. Furthermore, the suitability of functional subunits of proteins as entities of evolutionary change and phenotypic determinants was appreciated by explorative comparisons. In summary, the described methods provide versatile, portable solutions to compare a steadily growing number of bacterial strains with available genome sequence information.

Diagnostics of Enterobacteria

Microarray technologies

Microarrays can roughly be described as platforms containing immobilised biomolecules that are capable to bind to stained target molecules. Preliminary developments of this technology were based on nylon membranes loaded with complementary DNA (cDNA). Analogous to applications of state-of-the-art microarrays, rRNA samples were hybridised to these known cDNAs. In modern microarrays the platform material changed to glass or translucent plastic with amino, aldehyde or epoxy derivatised surfaces (Venkatasubbarao, 2004). Simultaneously the capacity to immobilise reporter DNA tremendously increased. The cDNA determinants were replace with more sensitive oligonucleotides, which can reach densities up to several millions of probes. (Stoughton, 2005)

Current microarray technologies enable the design of whole genome or even multi-genome arrays (Willenbrock *et al.*, 2006, 2007) to perform CGH experiments. Beyond gene expression analysis and CGH, the range of microarray application comprises the detection of single nucleotide polymorphism and diagnostics (Cassone *et al.*, 2007). Diagnostic microarrays are characterised by a small number of specifically designed probes that map to genes specifically linked to certain target organisms. Microarray experiments are complex processes that involve many individual operations. Beginning with the determination of capture probes, over sample preparation to hybridisation, such studies as well demand for time-efficient but sensitive evaluation of hybridisation profiles. Hybridisation signal analysis is generally based on methods of hierarchical clustering, statistical testing and multivariate analysis. Hybridisation signal evaluation is an essential step to ensure the detection of important signals within a mass of data as well as background noise from unspecific hybridisation and the carrier material, respectively.

Existing diagnostic strategies

Enterobacteria formerly were classified by serological determination of O-, H- and K-antigens. The O-antigen refers to the polysaccharide side chain of lipopolysaccharide (LPS), the H-antigen to the flagellum and the K-antigen to an antigen resulting from capsular proteins. The serotype was defined by specific combinations of O- and H-antigens found on the surface of enterobacterial strains that cause similar pathological symptoms (Nataro and Kaper, 1998). Serotyping is an indirect determination of pathogenicity as LPS and flagella are not directly involved in pathogenicity and as the motility depends on the cellular and environmental state.

The introduction of PCR technology enabled the detection of genotypic virulence determinants. Classical PCR assays are only suitable for small-scale diagnostics using few markers and a narrow spectrum of target species. Real-time PCR, nested PCR, ligase chain reaction or PCR-ELISA enhanced the technology towards higher sensitivity and/or efficiency. The development of multiplex PCR enabled multiple target diagnostics with up to 100 markers within single reactions. Alternatively, non-amplification methods like fluorescence in situ hybridisation were applied to the detection of *Y. pestis* (Mothershed and Whitney, 2006). Microarrays are two-dimensional matrices allowing the incorporations of higher numbers of capture probes. The technology is principally not restricted in the number of capture probes and provides high reproducibility even across different platforms (Consortium *et al.*, 2006). Numerous Microarray-based diagnostics for a multitude of bacterial pathogens including single species, whole genera and even a broad spectrum of enterobacteria have been developed (Barl *et al.*, 2008; Kostić *et al.*, 2007; Pelludat *et al.*, 2005; Bekal *et al.*, 2003). In the majority of approaches either virulence genes or phylogenetic markers like 16S or 23S rRNA were selected as capture probes (Yoo *et al.*, 2009; Bruant *et al.*, 2006; Ikeda *et al.*, 2005; Lehner *et al.*, 2005). Several approach include or specifically focus on the detection of antimicrobial resistance (AMR) (Frye *et al.*, 2008; Bruant *et al.*, 2006).

Antimicrobial resistance

Since the last 60 to 70 years, antibiotics have become a common way to treat bacterial infections. The frequent use of antimicrobial agents in medical therapeutics generates an increased evolutionary selection pressure to develop AMR strategies in human-associated microbiotas (Cohen, 1992).

Studies on the progressing spread of antimicrobial resistance underline the need of its screening in clinical diagnostics (von Baum and Marre, 2005; Welch *et al.*, 2007). Naturally produced antibiotics are secondary metabolites of bacteria or other microbes to combat rival species.

Antibiotics interfere in three essential microbial processes: the cell-wall biosynthesis (β-lactams), the protein synthesis (aminoglycosides, macrolides, tetracyclines) or DNA replication and repair (fluoroquinolones). With the extensive use of antibiotics in medical care and agriculture, bacterial pathogens as well as commensal bacteria in animals and environment developed strategies of AMR. Rapid development and spread of antimicrobial resistance is related to a high mutation rate with short generation times, a high selective pressure in antimicrobial therapy and the collection of resistance mediating genes on mobile genetic elements.

One of different strategies to develop resistance is the expression of multi-drug efflux pumps. The protein complexes pump antimicrobial agents at high rates out of bacterial cells so that they could not act on intracellular target sites like the peptidyl transferases. The pumps are variants of transmembrane proteins occurring in all bacteria to transport lipophilic and amphipathic molecules. Secondly, hydrolytic enzymes were developed by bacteria that inactivate β-lactam antibiotics by destroying the β-lactam rings at high rates. Resistance is furthermore conferred by alteration of the target structure of the antimicrobial agents, as e. g. the methylation of residues in ribosome subunits or reprogramming of peptidoglycan composition. (Walsh, 2000)

Microarray-based enterobacterial diagnostics

Though the pool of acquired genomic, proteomic or interactomic data raises new questions and requires sophisticated techniques of automated analysis, it also provides a starting-point of yet unconsidered strategies in research and diagnostics.

The development of an enterobacterial diagnostic microarray, which is described in chapter 5, targets the identification of clinically relevant pathogroups from genus to even subspecies level. In contrast to previous work, we unravelled pathogroup-specific capture probes by probe selection across multiple genomes leading to yet unconsidered determinants of enterobacterial pathogenicity. Diagnostic classification as well as the quantification of pathogens in a sample is provided by the application of a regression model. The classifier features training in contrast to previous experiments which providing a constant learning ability in the process of application. An integrated approach is described to design a high level diagnostic microarray for a multitude of clinically relevant phenotypes among enterobacteria by unravelling and learning distinct genotypic traits.

Meta-analysis on gene expression experiments

In the last years, enormous amounts of data have been generated by microarray experiments from different organisms, tissues and platforms under various experimental conditions. Databases like the NCBI Gene Expression Omnibus (GEO) (Barrett *et al.*, 2007), ArrayExpress (Parkinson *et al.*, 2007) and NASCArrays (Craigon *et al.*, 2004) have been set up to archive these datasets and to make them available to the scientific community. The size of microarray databases is likely to increase exponentially in the future, as is typical for all molecular databases, increasing the need for sophisticated methods to analyse these large amounts of data appropriately.

Several factors impede a straight-forward analysis of microarray database content: standards for data submission vary between different databases, some microarray datasets do not provide raw data and on the experimental side, protocols and experimental conditions can differ between diverse laboratories conducting microarray hybridisations. However, microarray meta-analysis on a potentially large number of datasets can substantially advance the gain of additional insights into gene regulation. In single experiments, such new unravelled details could have been overseen or not detected. Unrecognised gene regulation could result from weak signals of a particular gene or group of genes in single experiments. Furthermore, the sensitivity to detect gene regulation can be increased by putting genes into a functional context and considering its regulation under other conditions or treatments.

Several methods for microarray meta-analysis have been proposed in recent years, most of them using models which compute an “effect size” and take care of inter-study variation (Choi *et al.*, 2003; Conlon *et al.*, 2006; Hu *et al.*, 2005; Moreau *et al.*, 2003). Thus, they often resemble procedures applied for the detection of differential expression but add the study as an extra explanatory variable. Several datasets from different microarray experiments are integrated in the meta-analysis to increase the number of replicates and thereby the power to detect differentially expressed genes. Because this design implies that datasets addressing the same topic such as the same cell type or treatment are used, microarray meta-analyses of this kind usually consist of only a small number of studies.

A second approach to supervised microarray meta-analysis is to integrate knowledge of biological functions into the analysis to predict global co-expression relationships and to infer functional relationships between co-regulated genes (Huttenhower *et al.*, 2006).

Nevertheless, all the above methods are based on parametric models, which have several biological and statistical assumptions. In classical microarray analysis, a first explorative analysis reveals possible signals in the data, which can then be verified or disproved by parametrical hypothesis testing. Similarly, the described approach of unsupervised meta-analysis yields insights into the biological structure of the data and may thus lead to precise biological hypotheses. These could then be tested by the parametric models described above. The aim of this study is to compare the results of a large number of microarray experiments on *Arabidopsis thaliana* using the well established Affymetrix ATH-1 Genome Array (`http://www.affymetrix.com/products/arrays/specific/arab.affx`) as a starting point. The analysis is restricted to this highly-standardised platform to reduce uninformative variability introduced by different technologies.

In this unsupervised meta-analysis, I show how to overcome the challenges posed by the heterogeneity of microarray data. This was achieved by applying exploratory data analysis methods. First, microarray datasets from public web sources were collected and pre-processed in order to remove noise from the data and build a common data basis for further analyses. Later, exploratory data analysis was applied to the processed datasets, namely kernel Principal Component Analysis (kPCA) and spectral and hierarchical clustering, to group contrasts from different microarray experiments and to find genes regulated in a specific cluster. These genes were identified in a specific cluster by unsupervised feature subset selection using the kernel principal component loadings. Although gene selection or feature subset selection is a challenging task for classification, many different approaches have been proposed for the same. According to my knowledge, gene selection or feature subset selection has not yet been performed using loadings of features on kernel PCA scores in the context of meta-analysis.

Genes selected to play a role in either plant growth and development (related to indole-3-acetic acid, a plant growth hormone) or pathogen defence were mapped onto physiological processes and functions and could be validated by previous studies. For genes which have not completely been characterised yet, the developed approach was able to propose a function and a possible regulatory mechanism as shown here for DUF26 (Domain of Unknown Function) kinase genes.

Optimisation of sequence length representation in hidden Markov models

HMMs are a widely applied class of probabilistic models. The fields of applications of this methodology range from speech recognition and spam deobfuscation to image processing. The theory of HMMs goes back to the 1960s and is established in several applications of computational biology since the early 1990s. HMMs were transferred to model biological motifs like DNA sequences (Churchill, 1989), protein families (Haussler *et al.*, 1993) and gene expression time course data (Schliep *et al.*, 2003). Prominent examples are programmes like GENSCAN (Burge and Karlin,

1997) for the detection of coding regions in DNA sequences (Krogh *et al.*, 1994b), TMHMM predicting transmembrane areas in protein sequences (Sonnhammer *et al.*, 1998), HMMer (Eddy, 1998) and the Sequence Alignment and Modelling System (SAM) (Hughey and Krogh, 1996) for the assignment of homology for a protein sequence to protein or domain families. The underlying profile hidden Markov model of the latter two solutions enables a probabilistic representation of a protein or domain family, respectively. The databases SMART (Letunic *et al.*, 2004; Schultz *et al.*, 1998), Pfam (Bateman *et al.*, 2004) and TIGRFAM (Haft *et al.*, 2003) are sources of these HMMs accessible via the Internet. SMART is a database of profile hidden Markov models (pHMM) of signalling, extracellular and chromatin-associated domains, while Pfam and TIGRFAM contain pHMMs of all types of domain families. The numerous applications implicate a large variety of source data, which is likely to exhibit large differences in the underlying length distribution of data types, especially in biological sequences.

HMMs consist of a network of states connected with certain transition probabilities. If states are self-transitive the duration of stay follows a geometric law (Durbin *et al.*, 1998). The distribution of retention time in self-transitive states often does not match the length distributions of biological signals. Several HMM-based approaches circumvent the restriction to geometrically distributed source data:

- The detection of protein domains with HMMer and SAM as well as the prediction of interaction sites in proteins (Friedrich *et al.*, 2006) is based on a profile-like topology. Each conserved amino acid therefore is represented by a single state. An exception is the insert state, which can model several amino acids. The applied model topology associates insertions with geometric length distributions. Detailed investigations rather found a power law to fit insertion lengths in protein sequence alignments (Qian and Goldstein, 2001).

- An extended model class, termed semi-HMM, connects the length distribution of sequence segments to the frequency of its observations. The trade-off for an explicit integration of the length of stay is the need to adapt training and decoding algorithms. Semi-HMMs were employed by several approaches to predict genes, especially in GENSCAN, Genemark.hmm (Lukashin and Borodovsky, 1998) and Genie (Kulp *et al.*, 1996).

- Two approaches altered conventional HMMs to model substructures of genetic information (Melodelima *et al.*, 2007; Munch and Krogh, 2006). They use differently adapted topologies, while in both cases the determination of adaptation parameters is based on computationally demanding, numerical optimisation methods.

The adaptation of geometrically distributed data modelling in HMMs to DNA and protein sequences with bell-shaped length distributions can be achieved by a sequential replication of states (Durbin *et al.*, 1998). The appropriate representation of the underlying source of data promises an enhancement of HMM-based predictive methods.

In chapter 7 I describe a general methodology to adjust HMM topologies with respect to length distributions of the modelled biological sequences. This methodology is based on an estimation of replication and transition parameters for the adjustment of state-associated retention time by maximum likelihood and the method of moments. Though important modelling characteristics are adjusted, the well established model class of HMMs is maintained.

Interaction site prediction using hidden Markov models

To date, sequence databases grow with a steadily increasing pace. Most of these sequences are generated within large scale sequencing projects. As the experimental characterisation of a protein is a time consuming process, the gap between uncharacterised and characterised protein sequences opens further and further. This for example is reflected in the size difference of TrEMBL (Wu *et al.*, 2006), a database of translated DNA-sequences, and Swiss-prot (Boeckmann *et al.*, 2003) containing manually curated entries. Whereas the first contains more than 8.5 M entries (release 40.4), the second holds only about 470 K sequences (release 57.4). This discrepancy underlines the importance of tools for the automated functional annotation of proteins.

Driven not only by different large scale projects, it became clear in the last years, that a major aspect of the function of a protein is its interaction with other proteins. Unravelling these partners allows placing a protein into its cellular context, giving insights into higher level function. Still, these data do not provide any details about the type of interaction or regions of the protein with substantial importance for the interaction. To address this problem, Aloy *et al.* (2004) performed three dimensional reconstructions of protein complexes. Indeed, this approach does reveal many details of the structural basis of an interaction, but it might be too time-consuming and too sophisticated for large scale applications. A trade-off will be the prediction of regions of a protein involved in interactions. Accordingly, different methods have been developed to analyse and predict residue patches involved in protein binding. For all of these tools, the Protein Data Bank (PDB) (Deshpande *et al.*, 2005) is the standard source of verified structural information on proteins and protein-ligand complexes.

Three main strategies were followed to approach the detailed analysis of binding interfaces and their interaction sites. For a large amount of proteins no data on binding interfaces is available. Features of binding sites like the accessible surface area, the hydrophobicity or the interface residue propensity were inferred from resolved protein-ligand complexes (Jones and Thornton, 1997) and transferred to predictions for new structures via SVM (Bradford and Westhead, 2005; Koike and Takagi, 2004; Chung *et al.*, 2006), neural networks (Zhou and Shan, 2001; Fariselli *et al.*, 2002) and via homology using FastA and further tools (Hendlich *et al.*, 2003; Milburn *et al.*, 1998). Although these approaches are useful in transferring knowledge of binding interfaces to protein structures, their application is restricted to only a small amount of proteins with known structure.

A combination of sequence and structure information provides an indication of evolutionary distance of functional sites. The evolutionary trace (ET) method searches for a structural cluster of conserved residues in a protein within a set of homologous sequences (Lichtarge *et al.*, 1996). All tools described above are restricted to work with protein structures as input.

As the amount of unidentified and uncharacterised protein sequences is growing very fast, the need of tools to automatically annotate them on the basis of existing knowledge is obvious. Ofran and Rost (2003) trained a neural network for the assignment of interaction sites in protein sequences where no structure information is available. This approach only performed quiet good in detecting interactions of strong evidence. Recently a profile based heuristic method for the localisation of binding patches for small molecules was published (Snyder *et al.*, 2006). It transfers annotated binding interfaces of small molecules from PDB entries to a query sequence and ranks them by calculating a ligand score for the binding patch. In a first step domains of the query were detected with RPS-Blast. Though this approach might give reasonable results for small ligands, difficulties in determining more variable interfaces like those targeting peptide or nucleotide ligands will probably occur.

A challenge for the prediction of interaction sites arises from the fact, that even within one protein or domain family, the position and the type of these sites can vary as highlighted for example by the sterile α motif (sam) domain. This domain is known to form homotypic and heterotypic oligomers (Thanos *et al.*, 1999; Schultz *et al.*, 1997). Other studies reported sam-mediated protein-protein interactions like the interaction between the ELK and the Grb10/2 proteins (Schultz *et al.*, 1997). In recent publications sam was described to bind RNA (Edwards *et al.*, 2005), and the domain is even thought to be involved in binding of p73 to lipid membranes (Barrera *et al.*, 2003). As described by Kim and Bowie (2003) for oligomerisation and RNA-binding, these interaction partners bind to different interfaces on the surface of the sam domains. It was shown in a recent large scale analysis of structurally characterised protein domains, that the variability exhibited by the sam domain is rather the rule than the exception. Within most of the analysed domain families, neither the position nor the type of amino acids involved in an interaction was conserved (Pils *et al.*, 2005). Obviously, this variability will hinder any straightforward prediction approaches simply transferring interaction sites of one family member to all other sequences members.

To address this challenge, I have adapted the statistical approach of HMMs to learn the patterns of functional sites in homologous sequences. In chapter 8 I describe a novel type of profile HMMs integrating information on sequence and function. One of its main features is the fully probabilistic detection of domains and interaction sites in proteins.

All together the approaches I propose here partially share methodical or topical intersections. The methods part (part I) refers to these overlaps by a pooling of methods related to studies concerning enterobacteria (see chapter 1) and those concerning HMMs (chapter 3).

Part I.

Applied methods

1. Enterobacterial genomics and diagnostics

1.1. Methods of genome comparison

1.1.1. Bacterial genomes

The publicly or elsewhere available enterobacterial genome sequences listed in Table 1.1 were subjected to genome comparison and probe selection. The genome sequences cover a broad range of pathotypes from *E. coli* as well as several subtypes or species in genera *Salmonella*, *Klebsiella* and *Yersinia*. Available plasmid sequences of the strains were incorporated in the studies.

1.1.2. Clustering of homologous proteins

The comparison of proteomes was based on all-against-all sequence alignments of enterobacterial proteins. Several billion pairwise protein sequence alignments were executed on a high-throughput Linux cluster. The sequences were aligned with an MPI-(Message Passing Interface) compiled version of the programme PARALIGN (Saebø *et al.*, 2005) in Smith-Waterman mode. The Smith-Waterman algorithm evaluates the full local alignment space and therefore produces more accurate alignments than the BLAST-heuristic, which is optimized for time-efficiency. The proteins were subsequently clustered according to sequence similarity with the programme OrthoMCL. The implemented Markov cluster algorithm interprets protein similarities as a graph. The graph is composed of nodes being proteins and edges representing protein similarities. Stochastic random walks through the graph are simulated by two operations on the similarity matrix, termed expansion and inflation, to separate clusters with respect to the local amount of flow between the nodes. The clusters match the definition of homologous proteins. The OrthoMCL algorithm was applied with lowest alignment coverage of 50% and an E-value-threshold of 10^{-6} for sequence similarity. Alternative thresholds for coverage and E-values did not substantially change the clustering result.

1.1.3. Core genome and dispensable genome

Common and variable parts of the gene-pool can be approximated by an assignment of presence and absence of coding sequences, here translated to amino acid sequences, in a set of related strains. The intersection of present proteins defines the core part across organisms, termed the core genome. The dispensable genome refers to the variable part of the gene-pool, which encodes for individual differ-

Table 1.1.: Table of bacterial genomes

Genus	Species	Isolate	Patho-/Serotype	Genbank-ID	Submission	Authors
Escherichia	*coli*	K-12 MG1655	non-pathogens	U00096.2	1997/09/26	Blattner *et al.* (1997)
		K-12 W3110	non-pathogens	AP009048.1	2005/08/22	Mori *et al.* (2005)
		Nissle 1917	commensal	–	–	–
		O9 HS	commensal	CP000802.1	2007/08/13	Rasko *et al.* (2007)
		DH10B	non-pathogens	CP000948.1	2003/08/14	Durfee *et al.* (2008)
		ATCC8739	commensal	CP000946.1	2008/02/14	Copeland *et al.* (2008)
		536	UPEC	CP000247.1	2006/01/20	Brzuszkiewicz *et al.* (2006)
		UTI89	UPEC	CP000243.1	2006/01/05	Chen *et al.* (2006)
		CFTO73	UPEC	AE014075.1	2002/06/20	Welch *et al.* (2002)
		APEC O1	APEC	CP000468.1	2006/09/14	Johnson *et al.* (2007)
		ACI 789	APEC	–	–	Eliora Ron, University of Tel Aviv
		O157:H7 EDL933	EHEC	AE005174.2	2000/10/22	Perna *et al.* (2001)
		O157:H7 Sakai	EHEC	BA000007.2	2000/06/26	Hayashi *et al.* (2001)
		O42	EAEC	–	–	Sanger Institute
		E2348-69	EPEC	FM180568	2008/07/16	Iguchi *et al.* (2009)
		E24377A	ETEC	CP000800.1	2009/07/11	Rasko *et al.* (2007)
		SMS-3-5	SECEC	CP000970.1	2003/08/20	Fricke *et al.* (2008)
Shigella	*flexneri*	2a 301	2a	AE005674.1	2004/12/06	Jin *et al.* (2004)
		5b 8401	5b	CP000266.1	2006/02/22	Nie *et al.* (2006)
		2a 2457T	2a	AE014073.1	2002/06/13	Wei *et al.* (2003)
	dysenteriae	Sd197	1	CP000034.1	2004/10/29	Yang *et al.* (2005)
	sonnei	Ss046	1	CP000038.1	2004/10/29	Yang *et al.* (2005)
Shigella	*boydii*	Sb227	4	CP000036.1	2004/10/29	Yang *et al.* (2005)
		CDC 3083-94	18	CP001063.1	2005/08/05	Rasko *et al.* (2008)
Klebsiella	*pneumoniae*	MGH78578		CP000647.1	2006/09/06	McClelland *et al.* (2006)
Salmonella	*enterica*	Paratyphi A ATCC9150	A	CP000026.1	2004/10/01	McClelland *et al.* (2004)
		Choleraesuis SC-B57	C1	AE017220.1	2004/09/02	Chiu *et al.* (2005)
		Typhi Ty2	D1	AE014613.1	2002/09/02	Deng *et al.* (2003)
		Typhi CT18	D1	AL513382.1	2001/10/25	Parkhill *et al.* (2001a)
		Arizonae	IIIa	CP000880.1	2007/11/21	McClelland *et al.* (2007)
	typhimurium	LT2	B	AE006468.1	2001/03/29	McClelland *et al.* (2001)
	bongori	12419	–	–	–	Sanger Institute
Yersinia	*pestis*	CO92	Orientalis	AL590842.1	2001/10/04	Parkhill *et al.* (2001b)
		KIM	Medievalis	AE009952.1	2002/02/21	Deng *et al.* (2002)
		91001	Microtus	AE017042.1	2003/04/24	Song *et al.* (2004)
		Antigua	Antiqua	CP000308.1	2006/04/06	Chain *et al.* (2006)
		Angola	Antiqua	CP000901.1	2007/12/12	Worsham *et al.* (2007)
		Nepal516	–	CP000305.1	2006/04/06	Chain *et al.* (2006)
		Pestoides F	–	CP000668.1	2007/04/13	Copeland *et al.* (2007)
	pseudotuberculosis	IP32953	–	BX936398.1	2004/02/08	Chain *et al.* (2004)
		IP31758	–	CP000720.1	2007/07/23	Eppinger *et al.* (2007)
		YPIII	–	CP000950.1	2008/11/03	Challacombe *et al.* (2008)
	enterocolitica	8081	–	AM286415.1	2006/06/30	Thomson *et al.* (2006)

ences and is not present throughout all organisms. The core genome was approximated by repetitive determination of the intersection between an iteratively increasing number of proteomes from different strains in permuted order. The imbalance in the number of reference strains per genus was compensated by appropriate sampling with equal sample sizes. The Extrapolation of dispensable and core genome size was performed according to the formalisation given in Algorithm 1, which followed

Algorithm 1 Calculation of dispensable and core genome

```
for 1 to max.iterate do                                                    ▷ max.iterate=10,000
    n.sample ← min group size                                     ▷ number of groupwise sampling
    genome sample +← sample from G1,...,$G_n$ n.sample times                ▷ vector of test genomes
    genome order ← permute ‘genome sample’ order
    for k = 2 to n genomes * min group size do
        core genome +← count present proteins in genomes[genome order]
        disp genome +← count absent in genomes [genome order[1:(k-1)]] & present in genome[k]
    end for
end for
```

the principle previously described by Tettelin *et al.* (2005) and Willenbrock *et al.* (2007). Values of dispensable and core genome sizes were fitted by non-linear regression. For the core genome approximation the best fit according to the error sum of squares was achieved by applying a triple exponential decay function

$$f_c(x) \quad = \quad \Omega_c + K_{C1}\, e^{(-T_{C1}\, x)} + K_{C2}\, e^{(-T_{C2}\, x)} + K_{C3}\, e^{(-T_{C3}\, x)} \tag{1.1}$$

while the dispensable genome was best fitted by a double exponential decay function

$$f_d(x) \quad = \quad \Omega_d + K_{D1}\, e^{(-T_{D1}\, x)} + K_{D2}\, e^{(-T_{D2}\, x)} \tag{1.2}$$

The variable x contains the amount of compared proteomes to determine the corresponding number of core protein clusters $f_c(x)$ or the number of specific protein clusters in the x-th proteome $f_d(x)$. $\Omega_{d/c}$ signify the approximated value of the dispensable or core genome for the whole group of strains. The T as well as K parameters represent the amplitude and exponential decay factors of the regression curve. Sampling outcomes for the dispensable and core genome were graphically displayed as bean plots. Bean plots follow the principle structure of box plots, but the boxes are replaced by so-called beans, which represent the density distribution of dispensable or core genome sizes for respective numbers of considered genomes. The density distributions are overlaid by horizontal lines providing values of dispensable and core genomes of respective single runs. The non-linear regression and corresponding confidence intervals were conducted with the statistical computing environment R (R Development Core Team, 2004) and the add-on packages `nlme` (Pineiro *et al.*, 2008) and `quantreg` (Koenker, 2008). Plotting was performed by applying the `beanplot` package (Kampstra, 2008).

1.1.4. Hierarchical clustering of proteome data

The hierarchical clustering was performed on the presence and absence, abundance or similarity values of protein clusters in related strains. As the results of strain comparisons based on E-value profiles did not substantially differ from those obtained by analysing profiles of abundance or binary assignments of protein clusters, the handy integer values were processed in subsequent explorations. The distances between binary and abundance profiles were calculated with the Jaccard index. The metric determines the fraction of entries (the protein clusters) that are simultaneously positive in two compared entities (the proteomes). Completely missing and therefore uninformative entries are not considered by Jaccard's method. The Jaccard index implementation of the R package `vegan` (Oksanen *et al.*, 2008) was applied to retrieve the distance matrix for the strains in focus.

The strains were hierarchically clustered with the method of Ward (1963) as it is implemented by the R function `hclust`. Ward's clustering intends to minimise the loss of information, which is determined by an approach derived from the error sum of squares principle. In every clustering step from single entities to one overall cluster the error sum of squares between clustering targets and the cluster means is minimised. Bootstrap resampling was conducted employing the `pvclust` R package (Suzuki and Shimodaira, 2006). Besides conventional bootstrapping, the algorithm provides the assessment of approximately unbiased bootstrap values. The latter bootstrapping corrects the influence of large data sets on bootstrap values by multi-scale bootstrap resampling.

1.1.5. Multivariate analysis

The multidimensional data resulting from proteome or virulence domain mappings were explored with respect to strain varieties by correspondence analysis (CA). Details about the methodology are reviewed elsewhere (Benzécri, 1992). Briefly, CA can be regarded as a way to simultaneously display and qualitatively correlate differences along multiple row and column entities of a multidimensional matrix by dimensional reduction. The method is based on the normalisation of matrix entries by division with respective row and column sums and on the calculation of χ^2-distances of row and column instances. The χ^2-distance indicates the cell-wise dependencies of row and column instances and serves as a measure of the divergence from expectation. Dimensional reduction is achieved by the determination of data planes in the projected data space, in which the first principal axis accounts for the largest fraction of the variance. Principal axis of higher CA-dimensions beginning with the second dimension are chosen according to orthogonality to the previous one, and account for the largest fraction of variance that is not covered by lower-dimensional principal axes. The orthogonal basis vectors can than be displayed in two- or three-dimensional coordinate systems. The origin of the coordinate system is termed centroid and characterises data independence of row and column instances. The distance to the centroid specifies the distinctness of dependence between row and column instances in the corresponding area of the plot. The superposition of row and column instances spatially discriminates positive from negative correlations by an opposite location of respective items in the plot.

The implementation of CA of the R packages `MASS` (Venables and Ripley, 2002) and `vegan` were applied to explore proteomic differences. Row and column instances of different main axis determined by CA were graphical displayed by using the `geneplotter` (Gentleman and Biocore, 2008) R package.

1.1.6. Phylogenetics and phylogenomics

Conventional enterobacterial phylogeny was inferred from 16S rRNA or *rpoB* marker genes, respectively. Multiple sequence alignments obtained from the alignment programme mafft using G-INSi parameter settings (Katoh *et al.*, 2005) were subjected to the programme PAUP, a commercial package of phylogenetic tools. The choice of an appropriate model of evolutionary processes was achieved by applying the programme modeltest (Posada and Crandall, 1998). Modeltest hierarchically determined the general time-reversible substitution model with invariant sites to be appropriate in 16S rRNA and *rpoB* phylogenies according to the Akaike information criterion. Phylogenetic trees were plotted using the visualisation tool splits tree (Huson, 1998). The programme provides, besides basic tree drawing, the alternative splits tree representation of phylogenetic distances. Splits trees reflect uncertainties in evolutionary divergence time of taxa by the insertion of parallelograms. The edge sizes of these parallelograms increase with the degree of uncertainty, which is calculated by the split decomposition algorithm (Bandelt and Dress, 1992).

Evolutionary analysis on genome rearrangements were based on multiple whole genome alignments applying the programme MAUVE (Darling *et al.*, 2004). This programme constructs alignments by a heuristic search of anchor sequences in locally homologous regions called local collinear blocks. The algorithm assigns local collinearity if such a block occurs in at least two genomes and fulfills a minimum weight criterion. The criterion refers to the number and length of anchor sequences in a block. The subsequent progressive alignment is guided by a phylogenetic tree constructed on the basis of the anchor sequences. The full alignment provides information about homologous blocks among genomic sequences, the order of these blocks and rearrangements with the corresponding position in the genomic landmark sequence. The sequence of homologous blocks and rearrangements was further subjected to the Spring server, which computes the breakpoint distances on reversals (also termed inversions) and block interchange (including transpositions) between chromosomes (Lin *et al.*, 2006). The distance is based on the operations needed to transform the sequence of reversals and block interchanges from one genome into the sequence of a second genome. Reversals are weighted down in distance measurements compared to block interchanges because of a higher observed occurrence of these evolutionary events.

1.1.7. Statistical tests

Significant presence and absence of protein clusters in subgroups of bacterial genomes were probed by different tests depending on the type of data. The significance of the presence ratio of protein clusters based on binary presence/absence in a subgroup of related enterobacteria compared to other subgroups was determined with Pearson's χ^2-statistical test.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - \bar{E})^2}{\bar{E}} \tag{1.3}$$

The χ^2-test statistic measures the deviance of the frequencies O_i in group i to the expected frequencies $\bar{E} = \frac{\sum_i O_i}{n}$ determined from all n groups. The test is based on the null hypothesis ($H_0 : O_1 = O_2 = \ldots = O_n = \bar{E}$) that all n assigned subgroups exhibit group-specific frequencies O_i of the presence of a protein cluster that are equal to an overall mean frequency $\bar{E}$. Protein clusters exclusively present in one group were selected by setting the alternative hypothesis to $H_1 : O_i > \bar{E}$. The test statistics for all protein clusters were calculated by the application of the R-specific implementation of Pearson's χ^2-test in the function `prop.test`.

Group-specific significant differences in the abundance of protein clusters were determined by applying the Kruskal-Wallis rank sum test. The non-parametric test statistic proposed by Kruskal and

Wallis (Kruskal and Wallis, 1952) is calculated as

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{n} s_i \bar{r}_{i\cdot}^2 - 3(N+1) \tag{1.4}$$

with the overall number of observations N, the size of the i-th group s_i and the average sum of ranks from observations $\bar{r}_{i\cdot}$ in group i. The Kruskal-Wallis test is based on a collective ranking of all abundances from all groups. Identical abundance values were assigned to averaged ranks. The test then evaluates the hypothesis by comparing group-wise sums over ranks assigned to the abundance values.

Finally, the p-values for test statistic T (either χ^2 or H) was approximated

$$\Pr(\chi_{n-1}^2 \geq T) \tag{1.5}$$

by selecting the χ^2-distribution with $n-1$ degrees of freedom. Obtained p-values of the significance of group-wise protein cluster presence were corrected for multiple testing errors with the method of Benjamini and Hochberg (1995), which is provided by the R function `p.adjust`.

1.1.8. Assignment of protein domains

HMMs of protein domains were obtained from the Pfam A database. Chapters 7 and 8 deal in more detail with the theory of HMMs including a general optimisation of its modelling behaviour in chapter 7 and the development of a prediction method for protein interaction sites based on profile HMMs in chapter 8. Protein domains were detected by the programme HMMer applied in `hmmpfam` mode with an E-value threshold of 10^{-2}. The use of `pfam_ls` versions of the selected HMMs assured the detection of complete domains.

Annotation of specific protein clusters and HMM database

Annotations to bacterial protein sequences were requested from NCBI's Entrez protein database (03/2008). Multiple sequence alignments were constructed of sequences from each protein cluster using the programme mafft. The G-INSi parameter setting of the alignment programme was applied to obtain global alignments based on the Needleman-Wunsch algorithm. The alignments are provided for download at the database interface. They additionally served as training data for a library of profile HMMs, which were trained with `hmmbuild` and `hmmcalibrate` from the HMMer2-package. Text file versions of the HMMs in HMMer format are provided in conjunction with the database.

1.2. Microarray-related methods

1.2.1. Longest common factor statistics

Rahmann (2003) proposed an algorithm based on enhanced suffix arrays to identify all common, contiguous subsequences, termed factors, in a subset of reference genomes. The method is based on the definition of appropriate matching and cross-hybridisation thresholds to ensure a save matching to all target sequences and to prevent for undesired matches. Briefly summarised, the algorithm

Table 1.2.: Table of recently published bacterial genomes

Genus	Species	Isolate	Pathogroup	Genbank-ID	Submission	Authors
Escherichia	*coli*	DH10B	non-pathogens	CP000948.1	03/14/08	Durfee *et al.* (2008)
		Ed1a	non-pathogens	CU928162.2	12/18/08	Genoscope -,C.E.A.
		SE11	non-pathogens	AP009240.1	10/22/08	Oshima *et al.* (2008)
		ATCC8739	non-pathogens	CP000946.1	14-FEB-2008	Copeland et al. 2008
		IAI1	non-pathogens	CU928160.2	12/18/08	Genoscope -,C.E.A.
		IAI39	UPEC	CU928164.2	12/18/08	Genoscope -,C.E.A.
		UMN026	UPEC	CU928163.2	12/18/08	Genoscope -,C.E.A.
		SMS-3-5	SECEC[1]	CP000970.1	03/20/08	Fricke *et al.* (2008)
		O157:H7 EC4115	EHEC	CP001164.1	10/08/08	Eppinger et al. 2008
		55989	EAEC	CU928145.2	12/18/08	Genoscope -,C.E.A.
		E24377A	ETEC	CP000800.1	09/11/07	Rasko et al. 2007
		S88	MNEC	CU928161.2	12/18/08	Genoscope -,C.E.A.
Shigella	*boydii*	CDC 3083-94	18	CP001063.1	05/05/08	Rasko et al. 2008
Salmonella	*enterica*	Enteritidis P125109	PT4	AM933172.1	09/25/08	Thomson *et al.* (2008)
Klebsiella	*pneumoniae*	342		CP000964.1	09/24/08	Fouts *et al.* (2008)

[1] The SECEC pathotype is assigned to an environmental isolate from an industrial area. The strain causes diarrhoea and exhibits multiple antimicrobial resistances.

decomposes the target genomes into all possible factors and stores them in a lexicographical order together with information about sequence positions and longest common prefixes as enhanced suffix arrays. If a collection of several sequences is investigated, a generalized suffix array of all sequences is constructed. Matching statistics and longest common factors can further on be obtained according to the algorithm described by Rahmann (2002).

1.2.2. Sequence alignments and annotation

The occurrence of oligonucleotide probes in genomes of recently published enterobacteria was determined by Smith-Waterman sequence alignments over the whole 70 bp of probe length. The alignments were performed using an mpi-compiled version of the PARALIGN programme (Saebø *et al.*, 2005) for Linux clusters. In accordance to match/mismatch criteria applied to initial probe selection, the longest consecutive matching between probe and target genome sequences was determined. PARALIGN was also applied to align candidate probes with the human genome to prevent cross-hybridisation in clinical samples. Similarly, the performance of the probe set was assessed on recently published enterobacterial genomes. Table 1.2 lists detail of the genomes subjected to this screening.

All oligonucleotides related to pathogroup typing were functionally annotated by performing NCBI-BLAST searching against the enterobacterial sequence database. Annotations were obtained manually from the most abundant function assigned to respective genomic regions.

New AMR-specific capture probes were designed by the programme OligoPicker (Wang and Seed, 2003). Probe uniqueness was validated against the genomic DNA of reference strains with BLAST. Table S1 of supplementary material provides the whole set of markers for AMR.

1.2.3. Genomic DNA preparation

Cultures were grown overnight at 37°C with aeration by constant shaking in 4ml LB (Luria Bertani) medium. Two different DNA-extraction methods were applied to prepare genomic DNA (gDNA) from the collection of enterobacterial strains. The TNE DNA extraction was used as standard preparation method. In cases where the standard method failed, Phenol extraction was applied.

TNE DNA extraction Cells were washed once in 1ml TNE-buffer (10mM Tris [pH 7.5], 10 mM NaCl, 10mM EDTA) and resuspended in 600 µl TNEX (TNE, 1 Vol. Triton X 100) with 3 µl lysozyme (50 mg/ml) for 10 min at 37°C. After the addition of 30 µl proteinase K (20 mg/ml), the lysate was incubated for 1-2h at 65°C until clearance. DNA was precipitated by adding 30 µl 5 M NaCl and 1.3ml 100% ethanol, followed by two washing steps with 100% and 70% ethanol. The DNA was dried for 10 minutes and resuspended in 200 µl H_2O.

Phenol extraction After centrifugation cell pellets were resuspended in 500 µl lysis buffer (50 mM Tris-HCl [pH 8.0], 50 mM EDTA [pH 8.0]) for 1h at -20°C. Lysozyme (10mg/ml in 0.25M Tris-HCl [pH 8.0]) was added to the frozen cells. The cell lysates were thawed by reversing and incubated for 45 min on ice. Cellular proteins were enzymatically degraded by proteinase K solution (0.5% SDS, 50mM Tris-HCl [pH 7.5], 0.4M EDTA [pH 7.5], 1mg/ml Proteinase K) at 50°C. The DNA was extracted by addition of 500 µl Phenol (Tris-HCl) and precipitation from the aqueous phase with 0.1 vol. 3M Sodium-Acetate [pH 5.2]. The DNA was purified by the addition of 2.5 vol. of 100% ethanol, followed by washing, drying and resuspension as described before.

1.2.4. Microarray technology and hybridisation

The HTA™Slide12 from Greiner Bio-One provide 12 separate wells for independent parallel hybridisation. They are composted of polymer and coated with a 3D-Epoxy surface. Each well provides a printable area of 12 x 36 mm^2 bordered by a rim of 0.5 mm in height. The 70-mer oligonucleotides were synthesised by Metabion and spotting of microarrays was conducted by Scienion AG with a sciFLEXARRAYER S100 spotting machine.

Genomic DNA of enterobacterial isolates in Table 1.3 were hybridised to the microarray in order to test its hybridisation reliability on reference and new strains. Test hybridisations with different combinations and ratios of mixed culture samples were set up in addition to pure culture test in order to evaluate the performance of the microarray on community samples. Table 1.4 summarizes the ratios and isolates involved in mixed culture tests. The experiments comprise a dilution series of a mixture of the commensal *E. coli* strain K-12 MG1655 and the EHEC isolate *E. coli* O157:H7 EDL933 (M01-M05). Table 1.4 specifies the composition these and further equally balanced mixed culture samples ranging over the whole diversity if the pathogroup tree (M06-M12). The spike-in experiments were intended to evaluate the accuracy to predict simultaneously the DNA content and therefore the amount of two or more bacterial groups in a test sample. For the spike-in mixtures of EHEC and commensals, the pathogroup-specific rates varied in a range between 0.8 and 0.2 of overall hybridised DNA in a counterrotated mode starting with an amount of 1.6 µg commensal DNA in plot M01. The applied linear regression model was trained with all hybridisation patterns of isolates belonging to one of the groups indicated as annotation of the x-axis in the plots. To calibrate the coefficient matrix for the prediction of mixed cultures, the training was extended by the mixed-culture patterns. No cross-validation was performed because of the lack of biological repeats. All experiments were merely conducted with a technical replicate, an identical composition of the mixed sample.

Table 1.3.: Table of isolates used in microarray tests.

Species	Patho-/Serotype	Isolate
E. coli	MNEC	IHE3034
		A21
	SEPEC	4405/1
		B10363
	UPEC	536
		AD110
		EcoR55
	EHEC	ED142
		EDL933
		SF493/89
		5720/96
		2907/97
	EAEC	5777/94
		O42
		17-2
		DPT065
	EIEC	76-5
		EDL-1284
		HN280
		O164
	non-pathogens	EcoR28
		K-12 MG1655
		Nissle 1917
		M3/6
		EcoR7
		EcoR23
	APEC	BEN79
		Ben2908
		AC/I
	EPEC	179/2
		E2348/69
		37-4
		Z412-94
		TB156A
	ETEC	F18
		IMI590
		H10407
		E1392-75
		E34420A
		B34212c
S. dysenteriae	4	2095
	9	2088
S. sonnei	LT06	2084
		2083
	LT50	2098
S. flexneri	1a	2092
	2a	2089
	2b	2090
	3a	2093
		2082
	3b	2091
	4a	2081
	5	2097
S. boydii	4	2087
		2085
	11	2086
	14	2094
S. typhimurium	B	DT104
		DT17
		DT12
		DT170
		PTU302
		LT2
		ATCC14028
S. Bareilly	C1	
S. infantis	C1	
S. Virchow		
S. Livingstone		
S. Bovismorbificans	C2	
S. Manhatten		
S. Hadar		
S. Give	E1	
S. Derby	B	
K. pneumoniae		MGH78578
		U983
		375
		E492
		625
		Bk098/2
		3091
		Kp52145
		U047
		SB3464
		110
K. ozeanae		SB3431
K. edwardsii		S15
Y. enterocolitica		WA314
		1208-79
Y. pseudotuberculosis		25201A
		H260/91

The table lists all isolates applied for test hybridisations of the developed diagnostic chip. An abbreviated nomenclature was used in Salmonella listings providing genus and serovars.

Sample preparation and labelling The concentration of genomic DNA samples was determined both before labelling and after purification. Therefore, the absorption of 30 µl sample DNA (Abs_i) was measured at 260 nm wavelength with an Axon photometer. The DNA concentration of sample i results from $[gDNA]_i^{260} = Abs_i^{260}\ [gDNA]_0^{260}\ D_i$ with D_i as factor of dillution and $[gDNA]_0^{260} = 50\mu g\, ml^{-1}$ as gDNA concentration, when absorbance at 260nm is $Abs_0 = 1$.

Genomic DNA was labelled with the DecaLabel DNA Labeling Kit from Fermentas (30 rxns., No.

Table 1.4.: Table of mixed culture tests of microarray hybridisations.

gDNA 1	ratio	gDNA 2	ratio
E. coli K-12 MG1655	0.8	*E. coli* O157:H7 EDL933	0.2
E. coli K-12 MG1655	0.6	*E. coli* O157:H7 EDL933	0.4
E. coli K-12 MG1655	0.5	*E. coli* O157:H7 EDL933	0.5
E. coli K-12 MG1655	0.4	*E. coli* O157:H7 EDL933	0.6
E. coli K-12 MG1655	0.2	*E. coli* O157:H7 EDL933	0.8
E. coli M3/6	0.5	*S. flexneri* 1a	0.5
E. coli ED142	0.5	*S. boydii* 2094	0.5
E. coli M3/6	0.5	*S. typhimurium* LT2	0.5
E. coli ED142	0.5	*S. infantis*	0.5
E. coli M3/6	0.5	*Y. pestis* KUMA	0.5
E. coli ED142	0.5	*Y. pseudotuberculosis* H260/91	0.5
E. coli M3/6	0.5	*E. coli* 536	0.5

The table lists the compositions of test samples prepared for hybridisation experiments with mixed cultures. The first 5 spike-in experiments refer to the evaluation of detection accuracy in samples of varying gDNA amounts of a commensal against an EHEC strain.

K0622). Initially, 4 µg of genomic DNA were resuspended in 35 µl nuclease-free H_2O and after addition of 10 µl decanucleotide denatured for 5 min at 95°C. After 2 min on ice 5 µl labelling mix (3 µl mix T (0.33 mM dATP, 0.33 mM dCTP, 0.33 mM dGTP), 1 µl Cy5-dUTP (Enzo Life Science), 1 µl Klenow fragment) was mixed with each denatured sample. The labelling process is completed by incubation for 20 min at 37°C, the addition of 4 µl 0.25 mM dNTP, another incubation period of 15 min and finally the arrest of reaction.

Purification of labelled gDNA The MinElute PCR purification kit from Quiagen was used for purification. The kit is designed to recover DNA fragments of sizes in the range of 70 bp to 4 kb while small oligonucleotides shorter than 40 bp are removed. The labelled samples were mixed with 300 µl PB buffer and transferred to purification columns. After centrifugating at 12,000 g for 1 min and dropping of the filtrate 400 µl of 35% Guanidinium-HCl were added. Another centrifugation was followed by 2 washing steps with 700 µl PE mix and two times centrifugation to get rid of the resting PE in the filter. The sample was then eluted with two times 20 µl EB buffer for 2 min each and subsequent centrifugation.

Processing of slides All solution applied in processing and washing procedures of the slides were demineralised and filtrated with 0.22 µm pore filters. Spotted slides can exhibit clumping effects of oligonucleotides. To reduce these effects of the spotting procedure the slides were treated with water vapour and subsequent drying to “straighten” the probes. At first, the slides were treated for 5 min under agitation with 0.1% Triton X-100. Afterwards, they were transferred twice to a processing chamber filled with 6 mM HCl and agitated for 2 min. The following step comprised a bath in 100 mM KCl solution for 10 min and in water for 2 min, both at agitation. Then slides were transferred to a chamber filled with pre-warmed (50°C) 50 mM Ethanolamine, 0.1% SDS in 0.1 M Tris [pH 9.0] for 15 min. The processing was completed by two washing steps with H_2O for 2 min again under agitation, bathing in cold Ethanol and drying for 3 min under centrifugation at 1,000 g.

Hybridisation and washing In preparation for hybridisation, 2 µg of labelled and purified samples were dried in a SpeedVac and resuspended in 15 µl hybridisation buffer (Scienion SciHyb, prewarmed for 10 min to 42°C). The cavities of the hybridisation chamber were loaded with 20 µl H_2O, samples were dropped contactless on the spotted areas of the slides and the slides were hybridised overnight (about 15 h) in a 42°C water basin.

After removal of hybridisation fluid the arrays were washed three times with 30 µl washing solution 1 (5% 20x SSC, 0.033% SDS). In all successive steps the slides were kept in darkness if possible. The slides were consecutively transferred to chambers with washing solution 1, 2 (1% 20x SSC) and 3 (0.25% 20x SSC) and agitated for 5 min each. Finally, the slides were dried by centrifugation at 1,000 g for 3 min.

Scanning and image processing The slides were scanned in 5 µm resolution with an Axon GenePix® 4000B microarray scanner. Scan images were processed by applying the GenePix 6.0 software to obtain raw intensities.

1.2.5. Disc diffusion test

Strains were cultivated overnight at 37°C with aeration by constant shaking in 4 ml Mueller-Hinton (MH) medium (23 g/l Mueller-Hinton Broth, Composition: 2 g/l beef infusion solids, 1.5 g/l starch, 17.5 g/l casein hydrolysate, pH 7.4±0.2 (37°C)). 100 µl of the overnight culture were transferred to 4 ml MH medium and cultivated for 4 hours under constant shaking at 37°C. These cultures were diluted to a final culture containing between $5 \times 10^6 - 1 \times 10^6$ CFU/ml. 100 µl of each dilution were plated on a MH-agar plate (23 g/l MH-medium, 20 g/l agar). Susceptibility discs containing the antibiotic substances that are listed in Table 1.5 were placed on the agar plate in a sufficiently large distance regarding the zones of inhibition. The cells on the MH-agar plates were cultivated overnight at 37°C. The assignment of susceptibility, intermediate behaviour or resistance was subsequently determined by the measurement of the diameter of the zone of inhibition around the susceptibility discs. Table 1.5 lists the corresponding reference thresholds of this assignment, which were defined by the Clinical and Laboratory Standards Institute (USA).

1.2.6. Evaluation of hybridisation Patterns

Subsequent microarray analyses were performed using the statistical programming software R.

Between array normalisation

Raw intensities were normalised by the algorithm for variance stabilisation between arrays (Huber *et al.*, 2002). The method homogenises the variance of hybridisation intensities from a set of samples by transformation of the data with the model $h\left(x\right) = arsinh\left(a + bx\right)$. This transformation corrects for an underweighting of differences in lower intensities.

Separation of hybridisation intensities in signal and noise

Microarray experiments yield two kinds of outcomes: the signal intensities upon binding of complementary DNA and an unspecific fluorescence of the microarray surface or dye remnants. For log

Table 1.5.: Standards of antimicrobial resistance

Antibiotic	Class	Concentration [µg/ml]	Resistent [mm]	Intermediate [mm]	Susceptible [mm]
Amocillin	β-Lactam (Aminopenicillin)	2	≤15	16-22	≥23
Oxacillin	β-Lactam (Isoxazolylpenicilline)	5	≤15	–	≥16
Imipenem	β-Lactam (Carbapeneme)	10	≤13	14-15	≥16
Ceftriaxone	β-Lactam (Cephalosporine)	5	≤15	–	>15
Gentamicin	Aminoglycoside	10	≤14	15-20	≥21
Erythromycin	Macrolide	15	≤16	17-20	≥21
Tetracycline	Tetracycline	30	≤16	17-21	≥22
Chloramphenicol	Amphenicol	10	≤20	–	≥21
Sulphometoxazole/Trimethoprii	Sulfonamide/Dr inhibitor	25	<15	15-17	>17

Standard values to assign susceptibility to antibiotics. The thresholds were defined by the Clinical and Laboratory Standards Institute (USA). The listed antimicrobial agents cover all classes for which equivalent resistance probes were designed.

normalized hybridisation patterns each type of intensity values follows a normal distribution. The classification accuracy of microarray intensities in either one of these classes is strongly dependent on the degree of overlap of the two distributions. In experimentally generated hybridisation patterns the bimodal Gaussian mixture model is able to fit the two intrinsic normal distributions. Figure 1.1 graphically illustrates the distributional fitting of the two normal distributions to intensities of AMR-associated probes as green coloured background noise fraction and a red coloured signal fraction. Parameter estimation of the Gaussian mixture and calculation of posterior probabilities of the classification was achieved by using the R-package `Mclust` (Fraley and Raftery, 2002).

Analysis of variance and simultaneous inference of multiple comparisons

The analysis of variance (ANOVA) is a statistical test to determine if the mean value of groups within a test set significantly differs to a higher extent than single values of a group from its mean. In other words, the test determines if a significant difference in the average value between at least two groups is observed. Formally the test statistic is calculated as

$$F = \frac{\text{variance of group means}}{\text{mean of within-group variances}} \tag{1.6}$$

The uni-factorial and multi-factorial ANOVA are distinguished. A factor in the context of ANOVA is an independent variable consisting of two or more internal groups, the factor levels. In case of two groups, the ANOVA yields the same result as a two-sided t-test. In microarray applications the factors are vectors of signal intensities corresponding to probes and the groups are biological repeats of certain experimental condition or as in this case bacterial virulence phenotypes. ANOVA was calculated by applying the R-function `aov` from the `stats`-package.

An ANOVA is often complemented with additional tests to determine the nature of differences between groups. The Tukey honestly significant difference (HSD) test is often applied in such context. This test is replaced in the described analysis by the simultaneous inference of one-sided multiple comparisons (Hothorn *et al.*, 2008). The algorithm evaluates individual test hypothesis derived from

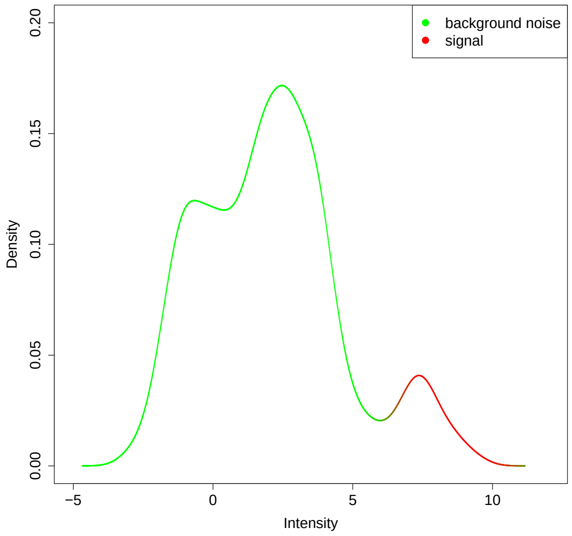


Figure 1.1.: Distribution of signal intensities of AMR probes. The parameters of the bimodal distribution were fitted with a Gaussian mixture model. The intensities were then classified into a background noise (green) and a signal fraction (red) based on the fitted distribution.

ANOVA to calculate adjusted p-values. The method is implemented in the R-package `multcomp`.

2. Meta-analysis on gene expression datasets

2.1. Data pre-processing

Microarray data were collected from the Gene Expression Omnibus (GEO) database (Barrett *et al.*, 2007). For our analysis, we defined a *dataset* as a GEO entry with a unique GSE series accession number. Each dataset consisted of several Affymetrix CEL-files, each one representing the raw data from one microarray hybridization. The raw data of one microarray is termed a *sample* in the following section. Instead of comparing whole GEO datasets with each other, we broke down each dataset into *contrasts* and used these as 'entities' for our analysis (Fig. 1, Everitt 2005). A *contrast* is the difference in gene expression between any two sample groups of the same dataset. A sample group contains all replicate samples from one condition (e.g. treatment, mutant, see Table 2). Therefore, for most GEO datasets, several contrasts were set up. For example, a contrast could be a comparison of an *Arabidopsis thaliana* mutant with a wild type plant.

A contrast was then represented by a vector of the logarithmic (base 2) fold changes of all 22810 probe sets on the ATH1 chip. The majority of probe sets on the ATH1 chip interrogates the expression level of one gene, some match to two or more genes. Before computing the fold changes, raw intensity values of all samples of a contrast were normalized using the gcRMA algorithm implemented in the *gcrma* package (Wu *et al.*, 2005) which is part of Bioconductor (Gentleman *et al.*, 2004) and runs under the statistical software R. Logarithmic fold changes and p-values adjusted for multiple testing using the false discovery rate method (Benjamini and Hochberg, 2000) were computed using the *limma* package (Smyth, 2004) which is also integrated into Bioconductor.

We imposed the following selection criteria on the datasets: a) Availability of the Affymetrix raw data (CEL-files) for download, b) at least two replicates of each condition are available c) time-course experiments were excluded. 20 GEO datasets fulfilled these criteria as of November 2006. From these datasets, 76 contrasts could be set up on the basis of 424 CEL-files. The final data matrix used for the unsupervised meta-analysis was a 76×22810 matrix, 76 contrasts with 22810 log fold changes.

2.2. Outlier removal and transformation

To remove experimental outliers from the data which could negatively influence any further analysis, a filtering criterion was set up as follows. Across all experiments, 15% and 85% quantiles of the distributions of medians and variances of the log fold changes were calculated. Experiments whose medians laid outside the inter-quantile-range or whose variances were below the 15% quantile threshold were excluded from further analysis. This resulted in a reduced data matrix X with 41 remaining contrasts. We randomly inspected the 35 removed contrasts for detectable problems and found several contrasts having a low-variant distribution of multiple-testing corrected p-values with almost all p-values close to one.

When dealing with heterogenous experimental datasets from different laboratories and experimental settings, efficient data transformation methods are necessary to produce a reasonable level of comparability. Log fold changes from microarray experiments deserve special attention in that they implicitly define a "direction" of differential expression by their algebraic sign which is semantically not sustainable when comparing contrasts from divergent settings. We therefore only evaluated the absolute value of the log fold changes and brought all remaining 41 contrasts approximately to a standard normal distribution by applying the *Box-Cox-Transformation* (Eq.2.1, Box and Cox 1964) using Maximum-Likelihood estimated power coefficients.

For a power coefficient p and data x the box-cox-transformed data x' is defined as follows:

$$x' = \begin{cases} (x^p - 1)/p & \text{if } p \neq 0 \\ log(x) & \text{if } p = 0 \end{cases} \tag{2.1}$$

The average p values were about 0.13, resulting in an approximately logarithmic transformation of the log fold changes. Subsequently, all datasets were standardized to zero mean and unit variance to analyze datasets without regard to their scale and location.

2.3. Kernel PCA

Principal Component Analysis (PCA) aims to provide a lower dimensional view of high dimensional data by projecting the data points from a data matrix X onto a new coordinate system retrieved by eigen-decomposition of the associated covariance matrix. The axes of the new coordinate system are thereby chosen in a way that each axis or principal component explains as much of the (remaining) variance of the data as possible and that all axes after the first are orthogonal to the ones before.

Kernel PCA (Schölkopf *et al.*, 1998) is a non-linear extension of the regular PCA, performing the same projection in a possibly even higher dimensional feature space. The data points are implicitly projected from the input space I into the feature space F by replacing the standard Euclidean dot product with a positive-semidefinite symmetric bilinear form, the kernel function κ (Eq. 2.2). The algorithm is represented in a dual form such that all computation takes place using only the matrix of pairwise dot products XX' (Shawe-Taylor and Cristianini, 2004), the Gram or Kernel matrix K (Eq. 2.3), instead of using the data points or its variances directly.

More precisely, for a row-indexed data matrix X and a mapping $\phi : I \to F$, $x \mapsto \phi(x)$ the kernel function κ and its associated kernel matrix K is defined as

$$\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \tag{2.2}$$

$$K_{ij} = \kappa(x_i, x_j). \tag{2.3}$$

Kernel PCA has the advantage of being able to detect non-linear patterns in the data which might be overlooked or not covered appropriately when using conventional PCA.

For our analysis we used the Kernel PCA algorithm implemented in the "kernlab" package (Karatzoglou *et al.*, 2004), for the kernel function κ we chose a polynomial kernel

$$\kappa(x_i, x_j) = (s \langle x_i, x_j \rangle + k)^d$$

of degree $d = 2$, scale $s = 1$ and offset $k = 0$.

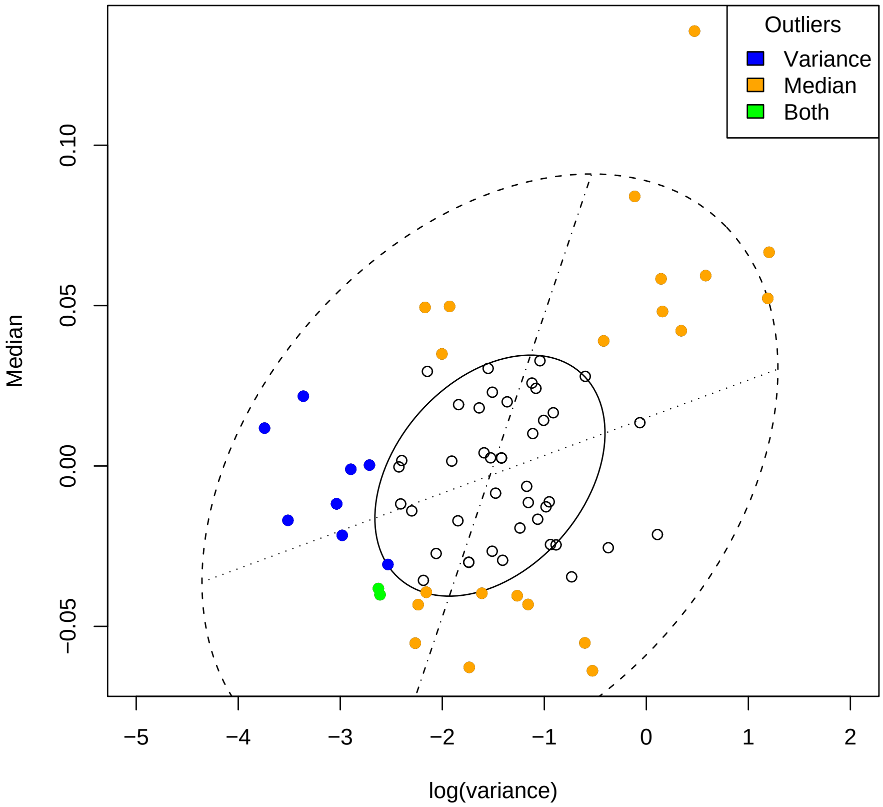


Figure 2.1.: Outlier removal. Median vs. log(variance) plot of all 76 contrasts and the associated bivariate box plot, colors indicate the type of outlier (see legend). The bivariate box plot is the two-dimensional analog of the familiar box plot of univariate data and consists of a pair of concentric ellipses, the hinge and the fence (Everitt, 2005). This box plot is based upon a robust estimator for location, scale and correlation. Uncolored contrasts were kept for further analysis.

2.4. Clustering

Clustering was performed on all remaining contrasts after removal of outliers. For an initial identification of the three main clusters of contrasts, we applied a spectral clustering algorithm from the "kernlab" package (Karatzoglou *et al.*, 2004). Spectral clustering algorithms cluster points using eigenvectors of matrices derived from the data, the kernel matrix K in this case. Similar to k-means clustering for data in the input space, the initial number of clusters has to be specified.

To gain structured clustering results, we applied hierarchical clustering using Ward's minimum variance method, which aims to find compact and spherical clusters based on Euclidean distance.

Decomposition of the symmetric kernel matrix K

$$K = S\Lambda S' \tag{2.4}$$

leads to a product of the orthogonal matrix S of its eigenvectors, a diagonal matrix Λ consisting of its eigenvalues and the transpose of S, S'. As the eigenvalues of K are directly linked to the proportion of explained variance of the principal component axes, the axes were scaled by the square roots of their respective eigenvalues, i. e.

$$\widetilde{X} = S\Lambda^{1/2}. \tag{2.5}$$

The result is a Euclidean distance

$$d(x_i, x_j) = \sqrt{\langle \widetilde{x}_i, \widetilde{x}_j \rangle} \tag{2.6}$$

weighted by the information content of each of the vector coefficients, thus scaling down axes that were given a low information content in the previous kPCA analysis.

Uncertainty of the predicted clusters was estimated by a 1000-fold multi-scale bootstrap resampling using the `pvclust` algorithm.

3. Sequence analysis with HMMs

3.1. Hidden Markov models

HMMs are applied in this study according to the well described theory by Rabiner (1989) and Durbin *et al.* (1998). Briefly, a HMM is a probabilistic network of nodes $\mathcal{Q} = \{q_1, \dots, q_m\}$, so called states. Each state q_i except for terminal states is connected to other states q_j by a transition probability τ_{ij}. Non-silent states are able to emit an alphabet of symbols $\mathcal{O} = \{\omega_1, \dots, \omega_n\}$.The transition and emission parameters of HMMs can be estimated if the state path of training samples is known. If no state path is available, Baum-Welch training provides an iterative refinement of the parameter space according to the likelihood of the data given the current model. Algorithms like Viterbi (Viterbi, 1967) and posterior decoding (Durbin *et al.*, 1998) determine a best path through the model and state-specific posterior probabilities for a sequence of observations (though not necessarily a valid path), respectively.

Profile hidden Markov models This type of linear HMM for a family of protein or DNA sequences can be considered as a probabilistic description of homologous protein sequences. A match state represents amino acids in conserved positions of a multiple sequence alignment. If the sequence lacks an amino acid to match in such a column, this position is defined as a deletion and it is assigned with a delete state. In unconserved alignment positions, amino acids are assigned with insert states. Frequencies of amino acid occurrences are separately modelled in each state as discrete emission probabilities and states are connected by transition probabilities in a left-to-right architecture. Various flanking states as well as begin and end states confer the adaptation to overhanging sequences and repeated protein domains. Several databases like SMART, Pfam and TIGRFAM allow for an assignment of homology in different categories of protein sequences. The SMART database provides profile hidden Markov models of signalling, extracellular and chromatin-associated domains based on expert-curated seed alignments. Domains of this HMM library were used as a source of domain family alignments.

3.2. Maximum likelihood

The maximum likelihood method provides an optimisation of free parameters of a mathematical model by the determinations of the set of parameters maximising the likelihood of the data to the model (Johnson *et al.*, 2005).

3.3. Method of moments

The method of moments goes back to Pearson (1902). In general, the moments of a function can be easily determined and directly yield distributional parameters. The moments of any random variable can be obtained from the equation

$$E\left[X^k\right] = \begin{cases} \sum\limits_x x^k p(x) & \text{if X is discrete,} \\ \int\limits_{-\infty}^{+\infty} x^k f(x)\,dx & \text{if X is continuous} \end{cases} \tag{3.1}$$

with the probability mass function $p(x)$ and the probability density function (PDF) $f(x)$ of the random variable X. The corresponding moment generating function of X equates to

$$M(t) = E(e^{tX}). \tag{3.2}$$

The kth differentiation of $M(t)$ at time $t = 0$ is equal to the expectation of the kth moment of the function

$$\frac{d^k}{dt} M(t)|_{t=0} = E(X^k). \tag{3.3}$$

Parameters are estimated by expressing them as functions of moments and replacing the moments by its sample moments

$$E[X^k] = \frac{1}{n} \sum_{i=1}^{n} x_i^k. \tag{3.4}$$

3.4. Model choice and goodness of fit

The Bayesian information criterion (BIC, Schwarz 1978) evaluates the likelihood of a model in relation to its complexity and therefore serves as a criterion of model choice. The BIC is calculated according to $BIC = -2\mathcal{L}(\theta, x) + k\,ln(n)$ with the log likelihood $\mathcal{L}(\theta, x)$ of the model, the number k of estimated parameters and the sample size n. The MM estimator does not yield likelihood values for the calculation of the BIC. Therefore, the L$_1$-distance was applied here as criterion of model selection. The L$_1$-distance of two discrete distributions x and y is given by the equation

$$d_1(x, y) = ||x - y||_1 := \sum_{i=1}^{n} |x_i - y_i|. \tag{3.5}$$

3.5. Receiver operating characteristic

The receiver operating characteristic (ROC) is a performance measurement of a binary classifier in dependence of a varying discrimination threshold. The conduction of ROC curves requires a classification of outcomes of a predictor in a positive and a negative class as well as the knowledge of the true nature of these outcomes. With this information the true positive and negative rates equate to

$$tpr = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \text{ and } tnr = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}.$$

ROC curves are then visualised in a scatter plot of the tpr versus the false positive rate $(1 - tnr)$. Predictors can be directly compared by the area under corresponding ROC curves.

3.6. Protein interaction data

As described in more detail by Pils *et al.* (2005) a HMMer search of all PDB sequences (October 2004 version, 27,969 structures) against the SMART database was performed to get all SMART sequences with a structure representation in PDB. All structures without ligands and all homodimer complexes were excluded, because homodimers are often an artefact of the crystallisation process. The remaining sequences were scanned for atom-atom distances smaller 4 Å between protein and ligand atoms. This length is consistent with distance between two oxygen atoms in a hydrogen bond. After filtering, the training set contained 5,590 sequences each associated to one of 248 domains. Every sequence was linked to its ligand-specific interaction profiles. Interaction site information was grouped according to the three considered ligand categories: peptides, nucleotides and ions. Other ligand types were not incorporated in our analyses because of low amount of data or unclassifiable ligands.

3.7. Validation with generated sequences

The recognition of self-emitted interaction profiles by an ipHMM is a necessary condition for the prediction of binding sites in new sequences. In a first validation step we used the feature of trained ipHMMs to emit domain-specific sequences according to their model parameters. Interaction sites of generated sequences were predicted and these predictions were compared with generated state paths in the same way as described below. This evaluation considered the same ligand-specific ipHMMs as in cross-validation tests with a limit of at least 20 sequences in the ipHMM-alignment. The process of generation was repeated 10 times for every domain.

3.8. Cross-validation and ROC curves for ipHMMs

We tested the prediction accuracy of the interaction profile HMM with 5-fold cross-validation. This was done for all domains with at least 20 sequences in the training set. All domain specific sets of sequences were partitioned into 5 equally dimensioned parts. We isolated 5 times a unique part as test set and estimated ipHMMs with the remaining 4 parts. Testing was done by applying the developed posterior decoding algorithm on all test sequences of a domain and finding matches between the predicted binding sites and the extracted interaction profile of the sequences (true positives, TP). The initial threshold for the assignment of an interaction site was set to a posterior probability of 0.5. True negatives (TN) are defined as correctly predicted non-interacting sites, false positives (FP) are predicted interacting positions that are characterised as non-interacting in the preceding structure scan. Finally, the false negative (FN) definition is the reverse case. We then calculated sensitivity, specificity as stated above and false positive rates as $false\ positive\ rate = 1 - specificity$. To get a closer look on the quality of our interaction site predictions we determined ROC for ipHMMs from all ligand categories in the peptide binding category (see figure 8.2). Therefore, the false positive rate of equation was plotted against the sensitivity. These values were calculated for increasing discrimination thresholds (steps of 0.02) in the range from 0 to 1.

Part II.

Achieved methods, results and conclusions

4. Comparative enterobacterial genomics

The developed approaches comprise new ways to study characteristics and differences of enterobacteria. The concept benefits from the availability of many completely sequenced genomes. Concurrently, bacterial genome sequences harbour the information to reinforce the evolutionary reconstructions and strain characterisations obtained in the past.

4.1. Multi-level phylogeny of enterobacteria

On phylogenetic markers for enterobacteria The established concepts for phylogenetic analysis in bacteria are based on the definition of phylogenetic markers. These marker genes are found in all organisms of the taxonomic groups under investigation and exhibit a high degree of conservation as well as areas of variation in their sequences (Doolittle, 1999). The overall conservation is required to build robust multiple sequence alignments, while variability allow for the differentiation of even closely related bacteria. In previous analyses several candidate marker genes were proposed in order to trace back the evolutionary paths of bacterial clades. The most common marker among these is the 16S rRNA gene encoding the small ribosomal subunit. Numerous studies on the reconstruction of the bacterial tree of life are based on 16S rRNA phylogenies (Fox *et al.*, 1980; Ibrahim *et al.*, 1993) and recently the gene became a barcoding unit in metagenomic analysis (Tringe *et al.*, 2005; Ley *et al.*, 2008b,a). But, contradictories resulting from the use of 16S rRNA as a phylogenetic determinant were reported as a consequence of multiple 16S rRNA gene copies in *E. coli* genomes (Case *et al.*, 2007). Inconsistencies in subtree topologies were obtained from phylogenetic reconstructions of all intragenomic 16S rRNA gene copies of *E. coli* and *Shigella* strains. Even single gene copies of recently diverged enterohaemorrhagic O157:H7 strains revealed more distant relationship to one another than to respective copies of other *E. coli* or *Shigella* isolates.

In a separate analysis we could confirm the stated scepticism against the application of 16S rRNA as a molecular marker to reconstruct subspecies level phylogenies of enterobacteria. In general, the splits tree in Figure 4.1 reveals consistency on the genus level between the *Yersinia*, *Salmonella* and *E. coli* clades. But, intragenomic copies of *E. coli* K-12 MG1655, *E. coli* HSO9 and *Shigella flexneri* 2a 2457T lead to confusions in the reconstruction of recent evolutionary events in the *E. coli* clade. Although phylogeny on distantly related taxa profits from the high degree of structural conservation of the 16S rRNA gene, its employment in evolutionary reconstruction on subspecies level of enterobacteria is inadvisable. Case *et al.* (2007) instead suggested the RNA polymerase β-subunit gene (*rpoB*) as a marker for phylogenetic analysis. Enterobacterial genomes largely contain only a single variant of the conserved gene (exception: one recent paralog in the genome of *S. enterica* Arizonae 62z4z23) with essential functionality. The *rpoB* marker was described to perform as good as 16S rRNA on distantly related taxa and to provide a better resolution on closely related organisms. Therefore, the gene

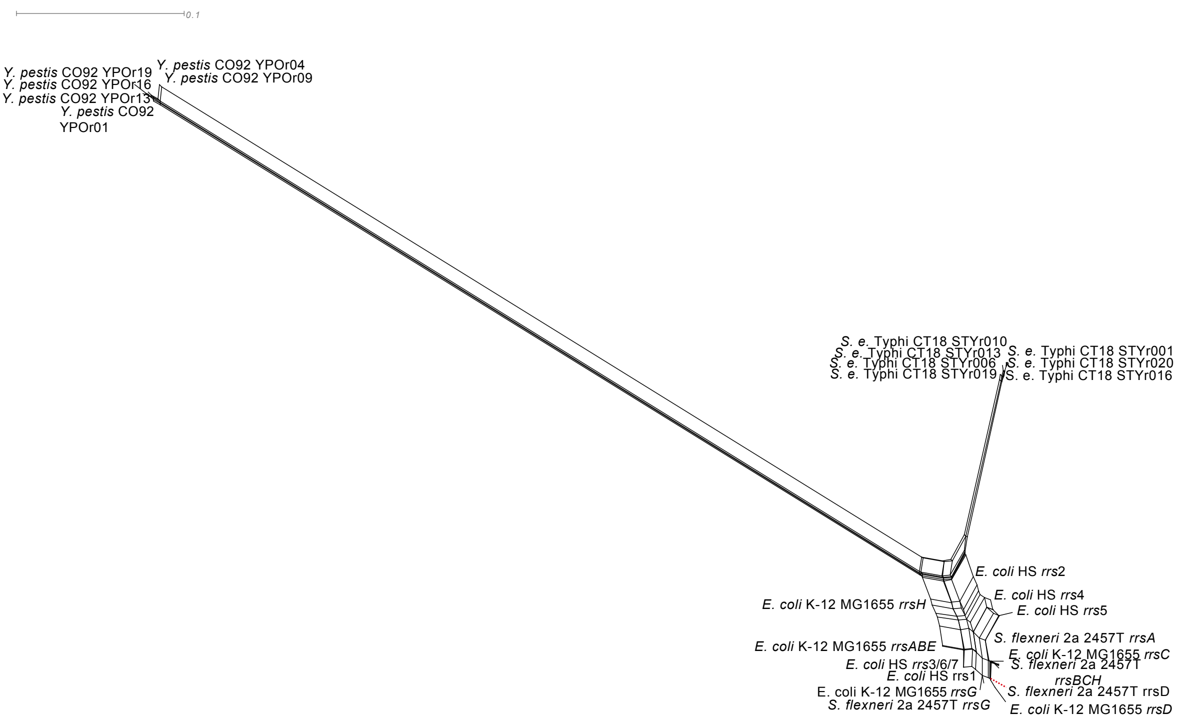


Figure 4.1.: Splits tree of enterobacterial 16S rRNA genes from different cistrons per genome. The splits tree exhibits ambiguities in the evolutionary traits of 16S rRNA sequences from different genomic loci in selected enterobacteria. The analysis clearly disqualifies 16S rRNA genes as marker for phylogenetic reconstructions in closely related enterobacterial subgroups.

was applied to the construction of a comprehensive phylogenetic tree of enterobacterial isolates. The *rpoB*-based unrooted phylogenetic tree in Figure 4.2 reveals the existence of three main subgroups consisting of the most clearly separated genus *Yersina*, as well as the genera *Salmonella* and *E. coli* (including *Shigella*). According to the reconstructions *E. coli* diverged early after the appearance of the taxon into intestinal, enterohaemorrhagic and extra-intestinal strains. The lineage attributed with an intestinal habitat further subdivided into *Shigella* as well as pathogenic and non-pathogenic *E. coli* strains.

As a consequence of the sparse representation of the genus *Klebsiella*, its exact evolutionary path within the family of *Enterobacteriaceae* remains to be determined in detail. But, the genus most likely separates early in the evolutionary history of *Enterobacteriaceae* in parallel to the other main clades. The evolution of the genus *Yersinia* is characterised by an early split with the appearance of the species *Y. enterocolitica*. Though *Y. enterocolitica* is more distantly related to *Y. pseudotuberculosis* than *Y. pestis*, the latter species developed a distinct pathogenicity in a short time concerning the evolutionary context (Wren, 2003). The phylogenetic tree underlines the close relationship between *Y. pseudotuberculosis* and *Y. pestis*. The low phylogenetic distance between *Y. pestis* isolates furthermore highlights the recent emergence of the species.

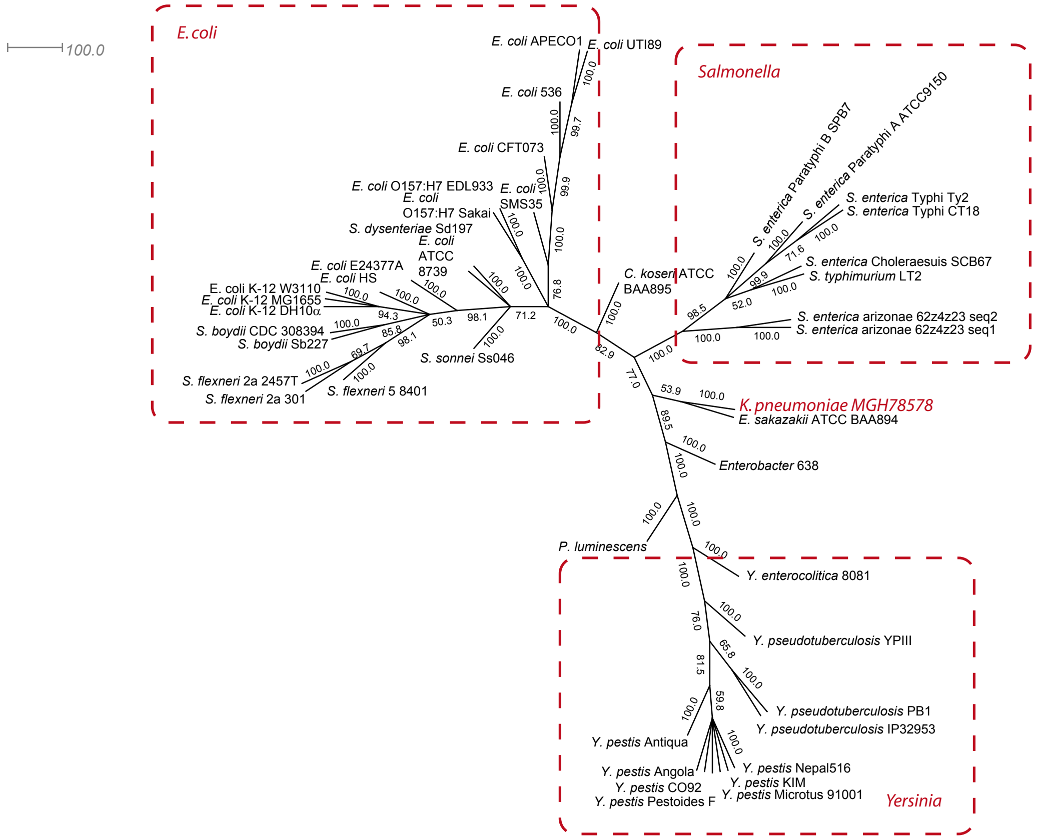


Figure 4.2.: Phylogenetic tree of enterobacterial strains with representation in public genome databases. The phylogenetic reconstruction is based on the *rpoB* gene, which has been applied to phylogenetic analysis for enterobacteria previously. By application of model test the GTK+I+G phylogenetic model was chosen according to hierarchical likelihood-based decisions. Values at the nodes of trees indicate the bootstrap values in percentages of split occurrence. Dashed rectangles enclose the three major groups considered in these analyses. Furthermore the only member of the genus *Klebsiella* is highlighted in red.

4.1.1. Global evolutionary aspects

The gene pools The completion of many genome projects of enterobacterial strains enables the comparison of genomic content. Differences in the genome sizes of enterobacterial strains imply genomic variability. This first picture is confirmed by the presence of numerous mobile elements in enterobacterial genomes. Multiple whole genome alignments of *E. coli* and *Shigella* isolates revealed large blocks of collinearity between these closely related bacteria as well as concise areas of higher variability (Mau *et al.*, 2006). These observations indicate the existence of a core genome common to all enterobacteria and a complemental part, which varies among the different strains. The core genome approximates the number of genes that are probably essential to establish an enterobacterial lifestyle. Previously, the core and complemental genome sizes only across *E. coli* strains were determined to 1,573 and 79, respectively (Willenbrock *et al.*, 2007).

As an important characteristic of bacterial clades, the size of complemental and core genome fraction were approximated for the larger group of enterobacteria. Figure 4.3 visualises the procedure of repeated and stepwise evaluation of the complemental (left) and core genome (right) in randomly sampled orders of evaluated proteomes. The amount of contribution to the overall enterobacterial complemental genome per newly discovered strain was estimated by non-linear regression to 278 protein families ($CI_{95} = 77 - 1,076$). Similarly, the level of core genome size was predicted for the whole group of enterobacteria to 1,209 ($CI_{95} = 1,273 - 1,557$) essential genes, which undershoots the value determined for *E. coli* alone. The increase in the complemental part and the decrease in the core genome size is owed to the higher degree of divergence across enterobacteria compared to the *E. coli*. Additionally, enterobacteria colonise a broader range of host and have developed new virulence mechanisms. Both facts imply larger genotypic diversity.

Genome rearrangements The complemental genome has its seeds in different evolutionary mechanisms. Genomic variation could arise from mutations, transduction or transformation. A sometimes neglected, but at least as important source of variation constitutes the shuffling of genomic content by mobile genetic elements. Recombinations at sites of direct repeats can lead to several types of genome rearrangements like duplication, translocation and inversion. Different types of mobile elements occur in bacterial genomes: genomic islands, bacteriophages and insertion elements (IS elements). They can give rise to gain, loss or shuffling of genomic content by insertion and excision. The reorganisation of genes or whole genomic regions could lead to considerable changes in cellular processes. Thus, these mechanisms represent important factors in bacterial evolution. The availability of whole genome sequences largely enables to investigate the genomic differences introduced by mobile elements and recombination.

In order to study the genome plasticity, conserved collinear blocks were determined based on multiple genome alignments of the enterobacterial strains under investigation. Evolutionary distances were calculated on the order and strand orientation of sequences of these genomic landmarks. The splits tree in Figure 4.4 reflects the phylogeny of enterobacterial genome rearrangements. The rearrangement metric clearly separates the *Yersinia* clade as in the previous phylogenetic analysis based on a single housekeeping gene. Likewise, the split of *Y. enterocolitica* appears, but the recent split evoking the appearance of *Y. pestis* and *Y. pseudotuberculosis* could not be detected on the rearrangement level. A second pole within the inferred tree consists of the other strains under investigation. In terms of genome rearrangements, the clades *E. coli*, *Klebsiella* and *Salmonella* largely exhibit collinearity, whereas *Shigella* strains were separated from their close *E. coli* relatives. The substantial difference in the order of genomic regions in *Shigella* isolates results from a large number of IS elements, which increase the genome plasticity. Previous genome analyses revealed 10-20 fold larger number of IS elements in *Shigella* compared to *E. coli* K-12 (Yang *et al.*, 2005). The reported amount of IS elements thereby correlates with the distances between *Shigella* and *E. coli* strains in the split decomposition. The largest distance in this context was obtained for *S. dysenteriae* Sd197, the strain with the largest number of IS elements among the investigated *Shigella* isolates.

4.2. Comparative proteome analysis

4.2.1. Identification of enterobacterial protein families

Ambiguous results from evolutionary considerations on different levels of genomic information in enterobacteria reflect the complexity of phylogenetic processes. At the same time these results highlight the need to investigate all units of genetic information in genome comparisons. An important element

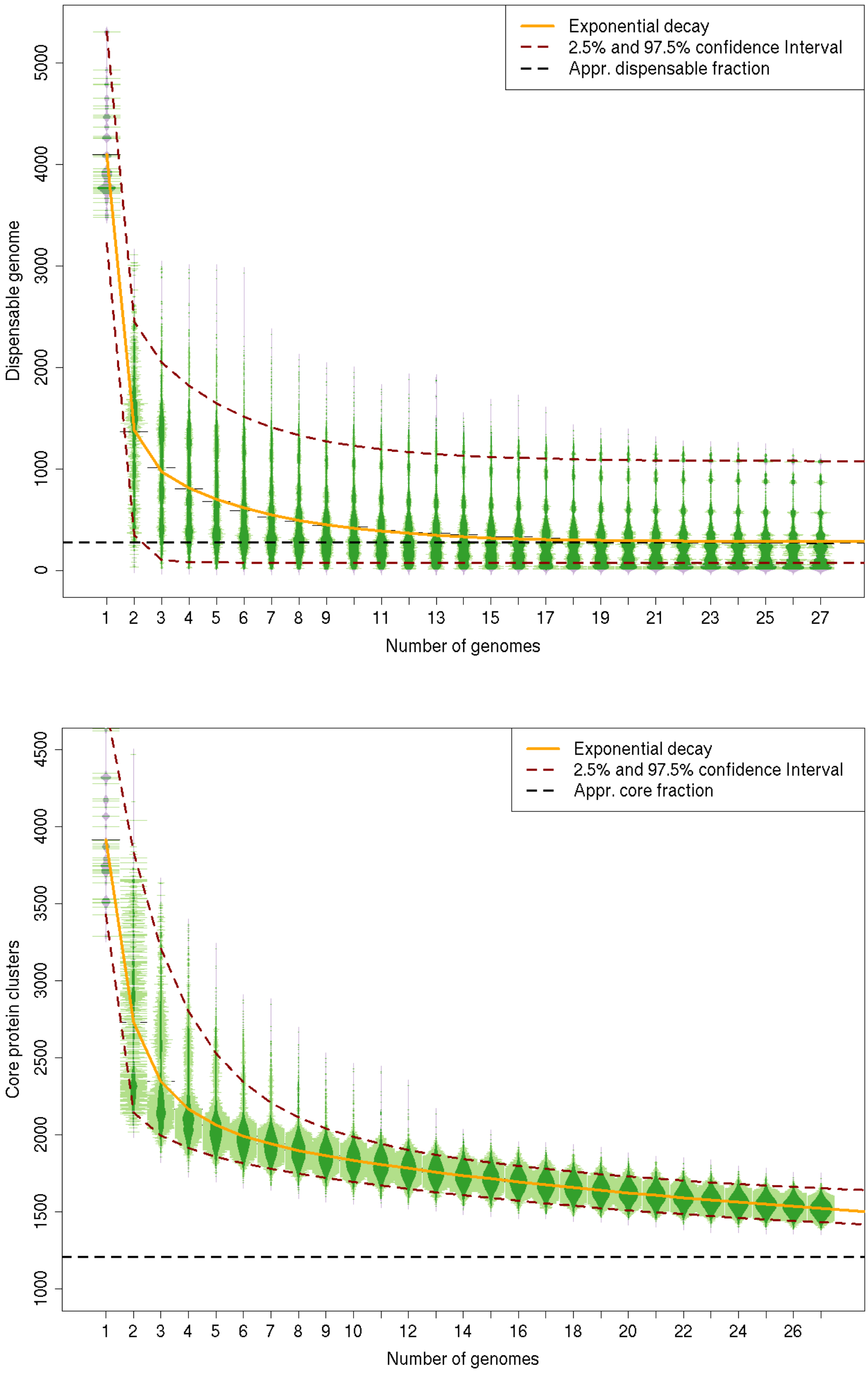


Figure 4.3.: Approximation of the complemental and core genome derived from previously conducted protein clustering. The size of complemental (left plot) and core genome (right) was determined by 10,000 random permutations of the order, in which the respective intersection or introduction of protein clusters in a growing group of targeted genomes was simulated. The values obtained from the iterative approximations were summarised as so-called bean plots. The vertical bean represents the density of complemental or core genome sizes for a respective group size of strains (x-axis). The beans are extended by a so-called ‘rug’, which indicate exact values of approximated sizes. The fitted curves refer to the estimated regression model (solid, orange line), the 95% confidence interval (dashed, dark red) and the asymptotic threshold of the overall sizes of complemental ω_d and core genomes ω_c (dashed, black).

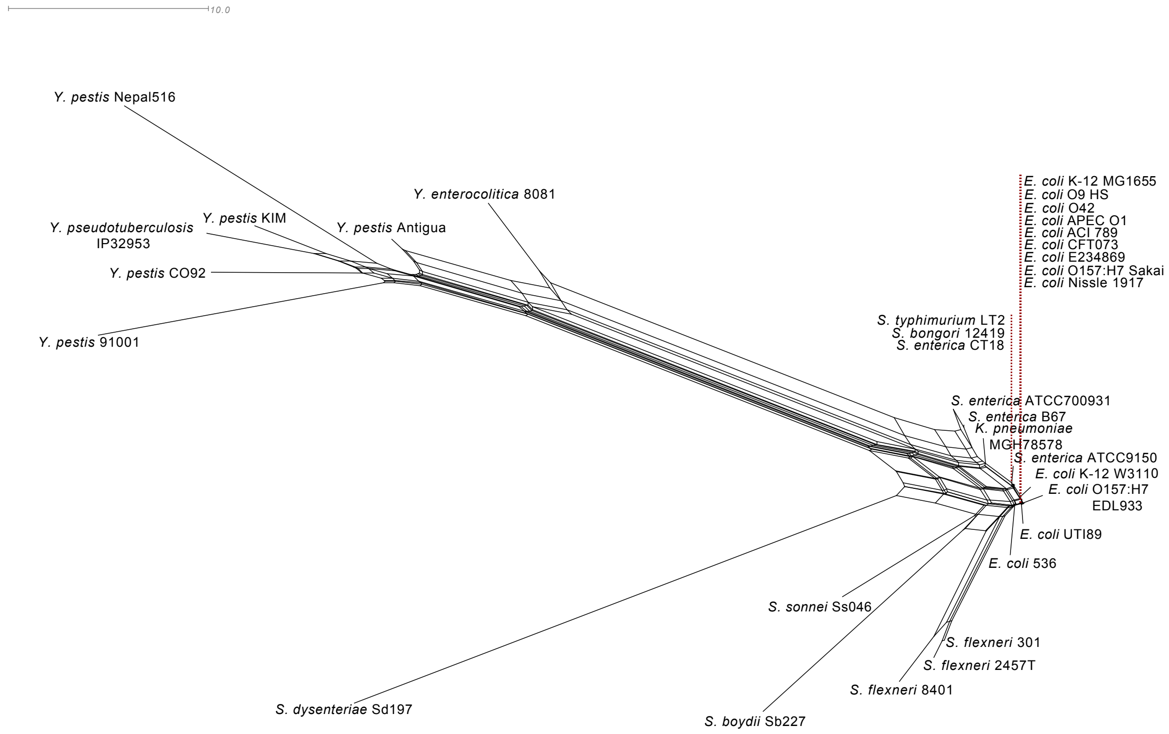
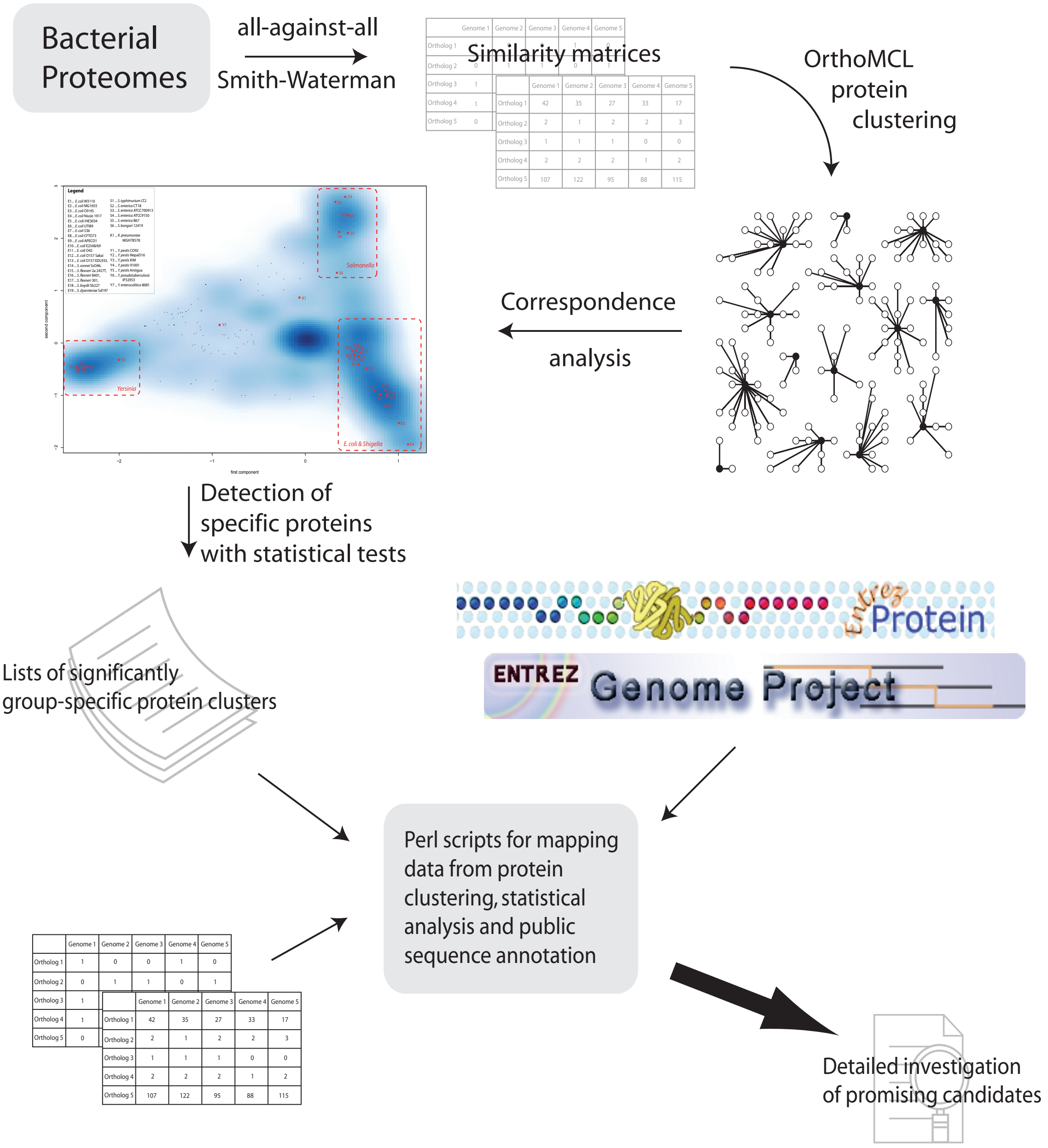


Figure 4.4.: Splits tree of the order and orientation of large genomic blocks in enterobacteria. A multiple whole genome alignment resulted in the identification of large genomically conserved regions, which order and strand location may differ between the strains. Based on differences in the order of genomic blocks, a distance matrix was calculated by applying the rearrangement metric implemented in the web application SPRING. Groups of strains exhibiting rarely any difference according to the applied metric are separately listed in order to increase readability. A dashed red line points to the position of these strains in the splits tree. The block The tree shows a clear separation of *Yersinia* strains. *Salmonella*, *Klebsiella*, *Shigella* and *E. coli* share a similar genome organisation. Nevertheless, *Shigella* strains exhibit a higher rate of genome rearrangements due to its large number of IS elements present in their genomes.

in this context is the compendium of the building blocks of cellular life, the proteins. The bacterial proteome is encoded by chromosomal and plasmid DNA, and protein expression is controlled by various regulatory mechanisms. The comparison of whole proteomes provides a more comprehensive picture of changes in lifestyle and pathogenicity than conventional phylogeny can afford.

The investigation of proteomic differences presupposes the mapping of homologs between bacterial strains. According to the flow diagram in Figure 4.5, an assignment of protein homology was achieved by performing large-scale protein clustering based on an MCL algorithm. In advance, an all-against-all similarity matrix was constructed by vast pairwise sequence alignments using the Smith-Waterman algorithm. Existing mapping approaches like the COG database were tested for suitability in protein mapping, but the coverage of COGs comprising enterobacterial proteins was too low. In addition, the OrthoMCL concept of protein clustering is applicable to closely related taxa while the COG approach was optimised for distantly related organisms. The OrthoMCL programme assigned 10,040 protein clusters consisting of two or more enterobacterial homologs or recent paralogs, respectively. A fraction of 1,321 protein clusters was detected in all strains, while 2,128 protein clusters only occurred in two genomes. Protein cluster sizes range from specific clusters comprising only two proteins to frequently occurring clusters with up to 751 members of the respective strains.


Bacterial Proteomes
all-against-all
Smith-Waterman
Similarity matrices
OrthoMCL protein clustering
Correspondence analysis
Legend
Salmonella
Yersinia
E.coli & Shigella
first component
second component
Detection of specific proteins with statistical tests
Entrez Protein
ENTREZ Genome Project
Lists of significantly group-specific protein clusters
Perl scripts for mapping data from protein clustering, statistical analysis and public sequence annotation
Detailed investigation of promising candidates

	Genome 1	Genome 2	Genome 3	Genome 4	Genome 5
Ortholog 1	42	35	27	33	17
Ortholog 2	2	1	2	2	3
Ortholog 3	1	1	1	0	0
Ortholog 4	2	2	2	1	2
Ortholog 5	107	122	95	88	115

Figure 4.5.: Flow diagram of the proteome comparison among enterobacteria. The comparison is based on all-against-all protein sequence alignments. The resulting similarity matrix was subjected to the OrthoMCL programme to determine clusters of homologs and recent paralogs. Patterns of presence/absence and abundance of protein clusters were globally explored by hierarchical clustering and correspondence analysis. Bacterial groups sharing similar patterns could be identified. The features of similarity - specific protein clusters - were determined by applying statistical tests in order to determine significantly different occurrence frequencies of respective proteins. Finally, these protein clusters were manually investigated and merged to functional units.

4.2.2. Unravelling proteomic differences

Whole proteome comparison Proteome mapping data represents a new kind of strain characterisation to unravel specificities and commonalities of enterobacterial subgroups. Different types of outcomes were obtained from protein mapping, a binary presence-absence notation, the abundance of cluster members existent per strain and similarity indicating E-values. Analytics based on these different data types resulted in similar outcomes. Therefore, subsequent analyses were conducted using the simpler binary or abundance data sets.

Evolutionary processes tackle with different strength at many sites of bacterial proteomes. The comparison of enterobacteria on the basis of the whole proteomes therefore promises comprehensive insights into the intrinsic evolution within the enterobacterial family. Overall proteome similarities were investigated by hierarchical clustering on protein abundance data. The resulting dendrogram visualised by Figure 4.6 again reveals a tripartition into distinct *Yersinia* (green colour), *Salmonella* (blue) and *E. coli* (red) groups. *Shigella* proteomes (orange) form a distinct subgroup compared to proteomes of other *E. coli* (dark red). Another separate group within *E. coli* refers to ExPEC isolates. All other intestinal strains show high similarities and comprise the closely related subgroups of EHEC and K-12 proteomes. The robustness of the hierarchical clustering was confirmed by adjusted bootstrap resampling as all splits with relevance for considered subgroups yielded high confidence (adjusted p-values > 90).

The *Salmonella* cluster in the dendrogram is composed of *S. enterica*ssp. *enterica* isolates as well as one *S. bongori* and one *S. enterica*ssp. *arizonae* strain. The latter two strains form a distinct group, which refers to a shared habitat in reptile hosts. Typhoid (serovars Typhi and Paratyphi) and non-typhoid (serovar Typhimurium) salmonellae did not reveal substantial differences on the proteome level, though they cause distinct disease patterns in humans. *K. pneumoniae* MGH78578 and *Y. enterocolitica* 8081 (dark blue) are assigned as outgroup of the *Salmonella* cluster. Such a classification does not surprise regarding the intermediate position of these strains in phylogenetic reconstructions with *rpoB*. A bipartite structure with a separation of *Y. pseudotuberculosis* and *Y. pestis* dominates the *Yersinia* genus. Though the two species recently evolved from the same lineage, they exhibit strong proteomic differences. The observations could be explained at least in parts by the different clinical pathologies and ways of infection of these two *Yersinia* species.

Correlation between strains and protein clusters Hierarchical clustering reveals the basic overall distances between enterobacterial proteomes, but does not allow for the investigation of the proteins that underlie these differences. The CA is an intuitive method to correlate strain-wise profiles of present and absent protein clusters with the occurrence patterns of protein clusters in respective enterobacterial strains. The matrix of presence and absence of protein clusters is far too large to extract important differences either concerning protein clusters or strains 'by eye'. Likewise, it is not possible to display the raw data in conventional two-dimensional plots without losing a substantial amount of information. CA provides a statistical framework to transform the data with the objective of the reduction of dimensionality. The method is suited for contingency tables of the type like the presence/absence or abundance data that was acquired for enterobacteria. After data transformation according to an ortho-normal projection of the proteome data into a data space, where the axes are orientated according to directions of maximum variance in the data cloud, two-dimensional visualisation preserves a large part of intrinsic information. The fraction of preserved inertia decreases with dimensionality.

Such a transformation was applied to the presence/absence profiles of protein clusters in enterobacteria to unravel those candidates that contribute to the group's versatility in host colonisation, pathogenicity or metabolism. The biplot in Figure 4.7 superimposes differences in the composition

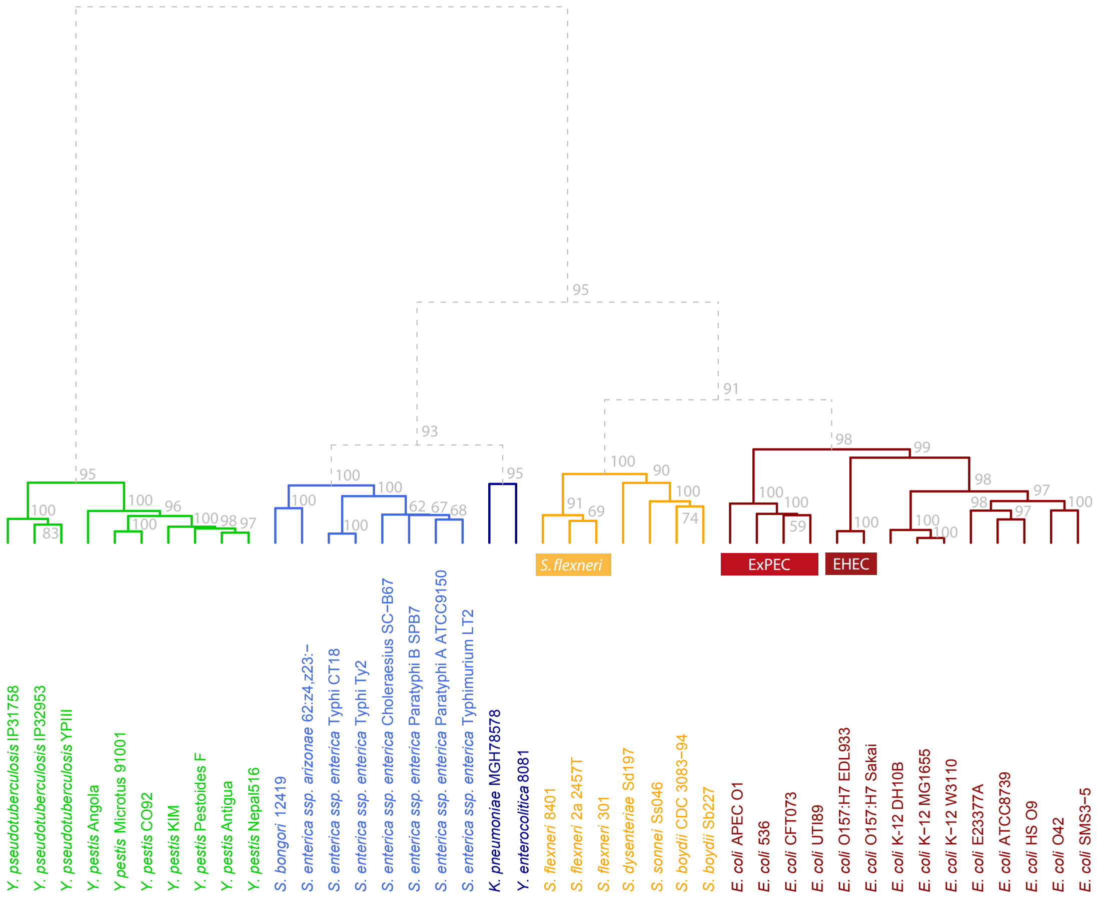


Figure 4.6.: Hierarchical clustering on the abundance of protein clusters in enterobacterial strains. The colors refer to the globally recognised groups of *Yersinia* (green), *Salmonella* (blue), *E. coli* (dark red) and *Shigella* (orange). *K. pneumoniae* MGH78578 and *Y. enteocolitica* 8081 are assigned to an extra group (dark blue), which coincides with previous phylogenetic results. Furthermore the *Shigella* subgroup composed of *S. flexneri* isolates (yellow bar) and the *E. coli* subgroups of EHEC (dark red bar) and ExPEC strains (red bar) exhibit high degree of proteomic similarity.

of enterobacterial proteomes with the separation of enterobacterial protein clusters according to its occurrences in enterobacterial strains. The separation of protein clusters is visualised as a smoothed scatter-plot, in which colours refer to the density of proteins. Single black dots are locations of protein clusters in low-density areas. Axis annotations contain arbitrary scales with no direct impact on the displayed entities. The origin of the coordinate system marks the area that has no influence to the criteria of separation underlying the chosen dimensions of the CA.

The first two principal axes of the CA on enterobacterial proteome differences separate the three main groups *E. coli* with *Shigella*, *Salmonella* and *Yersinia*. The composition and the sub-groupings in *Salmonella* and *Yersinia* regions are consistent with groupings in the hierarchical clustering. The highest protein density is located around the origin that corresponds to the commonly present core proteome, and therefore does not contribute to the separation of strains. Other areas of high density of protein clusters coincide with the locations of strains or groups of strains and represent protein clusters with specificity for corresponding single strains or whole groups.

The third principal axis in Figure 4.8 focuses on differences between intestinal and extra-intestinal

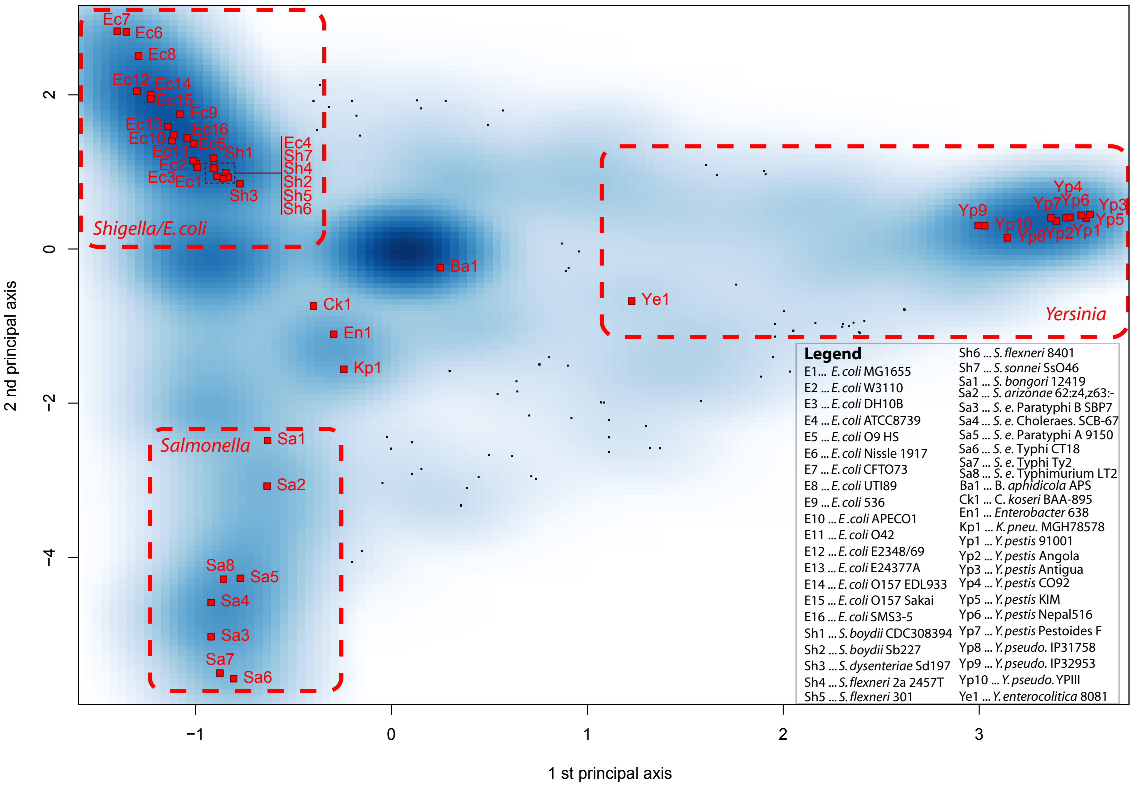


Figure 4.7.: Association graph of enterobacterial protein clusters and the strains they are occurring in. The correspondence analysis divides the set of investigated enterobacteria into three main groups: *Shigella*/*E. coli*, *Salmonella* and *Yersinia*. The distribution of protein clusters is indicated by the intensity of blue colour (dark blue = high density of protein clusters). The cloud of protein clusters around zero determines the core genome that influences to the characterisation of all strains equally. Single black dots mark the position of single protein clusters in areas of low protein cluster density.

E. coli proteomes. Principal axis 4 contrasts *K. pneumoniae* MGH78578 against other enterobacteria. Appendix A provides plots of further principal axes, which explain less obvious differences like those between closely related strains. The interpretation of the multivariate exploration of enterobacterial strains provides links to specific features of numerous grouping constellations among enterobacteria. CA applied in such context yields rapid overview of the interrelations between groups of strains and concurrently annotates the uncovered groupings with respective important features. Hence, CA is suggested as first stage analysis in genome comparisons across a multitude of bacterial strains.

Detection of characteristic proteins for bacterial groups The application of CA on proteome mapping data of enterobacterial strains provided indications of similarities in proteomes and the features contributing to the similarities. These protein clusters are specific or specifically absent for strains of a certain group. The general suitability of CA to assign characteristic proteins to bacterial groups was analysed by a mapping of the presence ratio of protein clusters in predefined enterobacterial subgroups. The groups were chosen according to previous results from proteome comparisons and common knowledge. In purpose of simplicity, the study was initially focused on the main groups *Yersinia*, *Salmonella* and *E. coli*. In order to correlate the location of protein clusters in

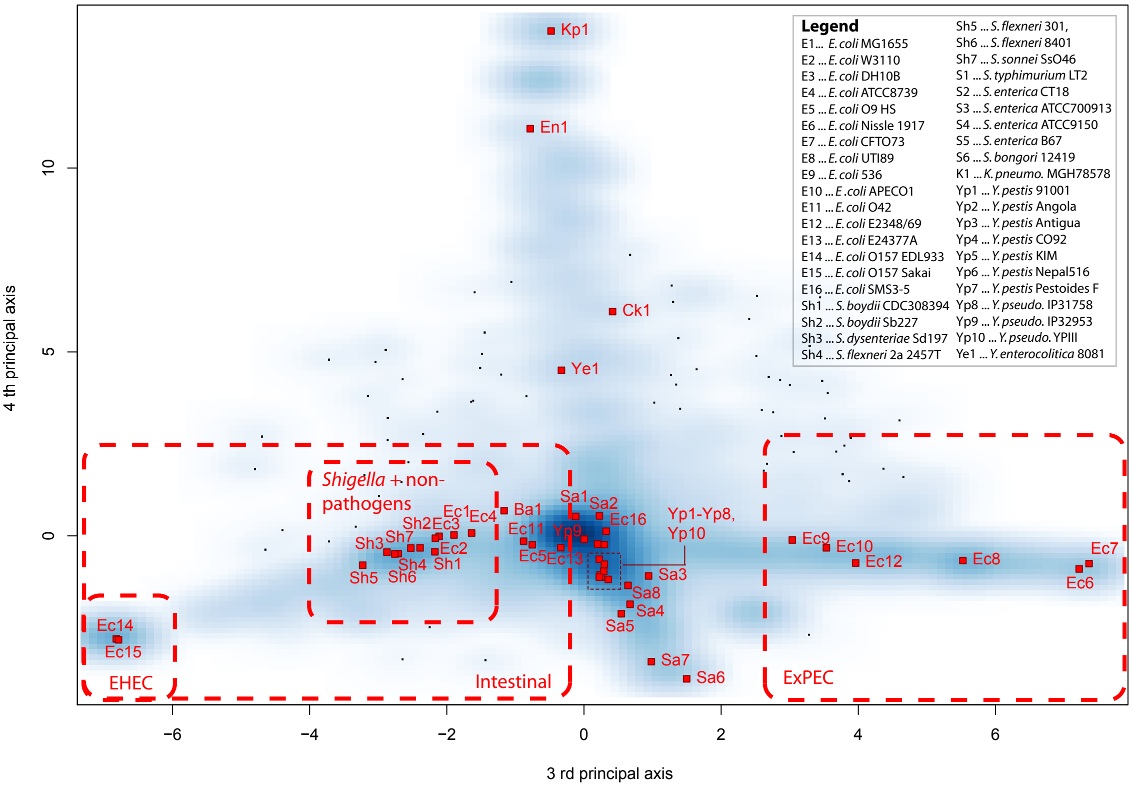


Figure 4.8.: CA plot of the third and fourth principal axes, which discriminate intestinal *E. coli* from extra-intestinal pathogenic *E. coli* and *Klebsiella* from other enterobacteria. Several subgroups can be distinguished within the intestinal group of *E. coli* strains comprising *Shigella*+commensal and enteohaemorrhagic isolates, respectively.

a CA plot with the coverage of their occurrences in defined groups, we assigned protein cluster items in conventional biplots with a colour scheme. The colours in RGB vectors were derived from group-wise ratios of protein cluster presence as $Col[RGB] = \{R_{PC}^{G1}, R_{PC}^{G2}, R_{PC}^{G3}\}$ with $R_{PC}^{G} = \frac{n_{pres}}{N}$. The colour $Col[RGB]$, which is assigned to a protein cluster (small diamond) in Figure 4.9, was calculated by the ratio R_{PC}^{G} of present protein clusters n_{pres} in respective groups $G1 \ldots G3$ divided by the group size N. The red colour channel was assigned to *E. coli*, the green channel to *Salmonella* and the blue one to *Yersinia*. Fully specific protein clusters assigned with the lightest pure red, green and blue colours are located around the local centroids of bacterial groups rather than at the edges. White colour indicates a presence in all proteomes and therefore characterises housekeeping proteins of the core proteome. These protein clusters are located at the origin. An extraction of the group-specific clusters by its position in the CA is difficult as these candidates assigned with light RGB colours are surrounded by protein clusters of partial specificity. The plot of protein cluster specificity clarifies the difficulty to directly obtain characteristic protein cluster for enterobacterial subgroups from CA.

Statistical tests for protein cluster specificity In order to overcome the limitations of the CA with respect to the assignment of specific proteome features, we developed a method based on statistical tests to reliably determine protein cluster specificity in enterobacteria. Specificity in a

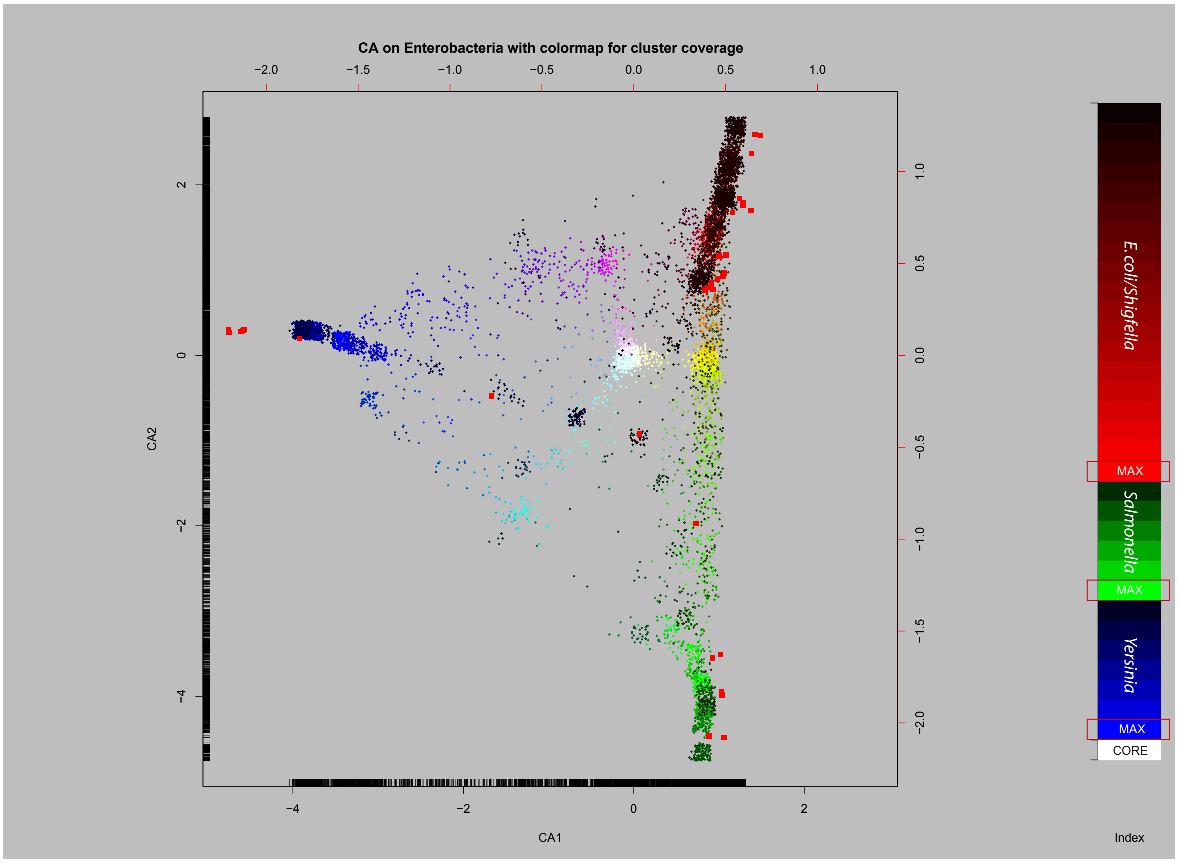


Figure 4.9.: CA plot with a colour-coded mapping of protein cluster presence in the predefined main groups *E. coli*, *Salmonella* and *Yersinia*. The CA plot maximally discriminates protein clusters according to the occurrence patterns in selected enterobacteria. Enterobacterial strains are categorised into the groups *E. coli* (red gradient), *Salmonella* (green) and *Yersinia* (blue). Small points refer to relative positions of protein clusters and large red squares to relative positions of strains. The protein clusters with highest specificity for one group are assigned with the lightest colour of the gradient. The colour gradient refers to the coverage of protein cluster presence in a group. Protein clusters with highest group specificity are located within the CA in the centre of each group. Thus, the CA does not provide an easy access to the assignment of characteristic protein clusters.

binary presence/absence assignment and abundance data was tested with Pearson's χ^2-test and the Kruskal-Wallis test, respectively. A protein cluster was determined as specific for a group of strains if the coverage or the ranks of abundance values significantly differ in one group. The p-values derived from statistical test were multiple testing corrected. They communicate the confidence of the assignment. The group-wise determination of characteristic protein clusters with p-values of group specificity below 5×10^{-2} resulted in large libraries of special features in enterobacterial genera. The protein clusters were made available as a database with computationally generated annotation derived from NCBI, multiple sequence alignments and profile HMMs for the assignment of these protein families in newly sequenced genomes.

Characteristics of enterobacterial subgroups The described approach yielded proteome characteristics based on statistical tests of significantly different occurrence patterns of protein clusters in enterobacterial subgroups. Group-specific protein clusters were manually selected with respect to annotations quality for detailed analysis and merged to operons or regulons if possible. Table 4.1

lists the curated library of specific traits from *E. coli*, *Salmonella* and *Yersinia*. In the following, some interesting protein clusters and whole operons will be described in more detail.

The performed analysis assigned specificity for *E. coli* strains to the protein clusters encoded in the *uidA* and *uidR* genes (first entry in Table 4.1; clusters CL3469 and CL3261). The β-glucuronidase enzyme UidA hydrolyses mucopolysaccharides at β-D-glucuronic-acid residues. The expression of UidA is repressed by the DNA-transcriptional repressor UidR. The *uid*-gene-locus was previously described as a specific region for *E. coli* isolates and was proposed as marker for the detection of *E. coli* and *Shigella* species (Cleuziat and Robert-Baudouy, 1990). The KEGG database of metabolic pathways (Kanehisa *et al.*, 2008) associates β-glucuronidase activity [EC:3.2.1.31] amongst others with the pathway for starch and sucrose metabolism (KEGG-Pathway: ecj00500, ecd00500, ece00500, ecc00500) for *E. coli* isolates, but not for *Salmonella* or *Yersinia* strains. *In vitro* cultivation of *E. coli* isolates on starch as sole carbon source lead to an induction of β-glucuronidase expression (Cenci *et al.*, 1998). Starch is available in gut environments of herbivorous vertebrate host, and can be a factor of increased growth for intestinal pathogens like the EHEC strain O157:H7 (Callaway *et al.*, 2003). The exclusive presence of the enzyme is a facet of *E. coli* lifestyle with implications in successful niche colonisation of vertebrate hosts.

Another enzyme, the oxalacetate decarboxylase [EC:4.1.1.3], was found to be specific for *Salmonella* isolates (clusters CL3591, CL3869 and CL6138). In salmonellae the enzyme is part of the arginine and proline metabolism and catalyses the transformation between glyoxylate, D-4-hydroxy-2-oxoglutarate and pyruvate (KEGG-Pathway: sty00330, stt00330, spt00330, sec00330, . . .). In *E. coli* and *Yersinia* strains the transformation is mediated by the 2-dehydro-3-deoxyphosphogluconate aldolase [EC:4.1.2.14], which complements the oxaloacetate decarboxylase in salmonellae. The oxaloacetate decarboxylase functions as a Na^+-ion pump that mediates anaerobic fermentation in *Salmonella* by establishing a proton gradient across the inner membrane (Woehlke and Dimroth, 1994). The Oad proteins in *Salmonella* are another example of specific metabolic functions acquired to enable vertebrate host colonisation.

All investigated *Salmonella* isolates contain the large anaerobic vitamin-B_{12}-synthesis pathway (*cbi-* and *ttr*-genes; clusters CL5345-53, CL5604-05, CL5770-71, CL4915 and CL4936-38). Vitamin B_{12} has a complex structure, is able to chelate an iron-ion and functions as a co-factor of important enzymes throughout all bacteria. The anaerobic pathway to synthesise vitamin B_{12} is absent in *E. coli* and *Yersinia*. Though the large operon of vitamin B_{12} synthesis was evolutionary maintained by selection to stay functionally conserved in *Salmonella*, it was only found to be essential under anaerobic conditions with tetrathionate as sole carbon source (Price-Carter *et al.*, 2001). Even though the benefit of the existence of the complementary pathway in *Salmonella* is not yet clear, elements of pathway for vitamin B_{12} synthesis were described to be involved in the regulation of B_{12}-independent enzymes (Rodionov *et al.*, 2003).

Salmonella pathogenicity incorporates the activity of a type-III-secretion system, which mediates the injection of effector proteins like SifA, SifB and SptP into the infected cell. SifA and SifB secretion normally induces the formation of *Salmonella*-induced filaments (Sifs) in infected epithelial cells. Furthermore, SifA was reported to mediate intracellular survival of *Salmonella* in murine macrophages (Brumell *et al.*, 2001). SptP is a secreted tyrosine phosphatase that disrupts the actin cytosceleton in host cells (Fu and Galán, 1998b). The chaperon SicP is required for the virulence of SptP (Fu and Galán, 1998a). These effector proteins partially shape *Salmonella* virulence and its assignment as specific characteristics contribute to uncover the *Salmonella* pathogenicity.

Yersinia strains harbour several siderophore systems for the binding and uptake of the limited iron sources available in host organisms. Beside the *fec-feb*-system, *Yersinia* isolates are equipped with the *yfuABC* siderophore system. The *yfu*-siderophore operon is specific for *Yersinia* strains and reveals highest similarity to a *Serratia sfu*-system, while the *feb-*/*fec*-system exhibits homology to the corresponding siderophore system in *E. coli* (Schubert *et al.*, 1999). The simultaneous expression of different iron uptake systems surely is an advantage in the highly competitive environments of vertebrate microbiotas.

Table 4.1.: Protein families, operons and regulons specific for one of the three major enterobacterial groups

Group	Name	Gene Tag(s)	Function	Cluster ID	References
E. coli/*Shigella*	β-glucuronidase operon	*uidAR*	enzymes truncating glucuronic acid residues, proposed as determinants for *E. coli* and *Shigella*	CL3469, CL3261	Cleuziat and Robert-Baudouy (1990)
E. coli/*Shigella*	β-galactosidase operon	*ebgCR*	β-subunit and regulator, used to study evol. adaptation in δLacZ mutants	CL3385-86	Hall (2003); Hazkani-Covo and Graur (2005)
E. coli/*Shigella*		*ygcE(G)OPQRUW*	cluster of proteins with predicted similarity to flavoproteins	CL3746, CL3349-51, CL3423, CL3565, CL3747	
E. coli/*Shigella*		*ykgEFGHK*	transcriptional unit (Ecocyc) of predicted oxidoreductases, predicted metabolic enzymes	CL3419-21, CL3697, CL3792	
E. coli/*Shigella*		*yhfSTUWXY*	put. alanine racemase (yhfX), put. mutase (yhfW)	CL3413-16, CL3485, CL3565	
E. coli/*Shigella*	*E. coli* multidrug resistance operon	*(emrKY), yibH*	multi-drug resistance	CL2372	Lomovskaya and Lewis (1992)
E. coli/*Shigella*		*yhdWYZ*	hyp. amino acid ABC transporter ATP-binding proteins	CL3201-02, CL3142	Hazkani-Covo and Graur (2005)
E. coli/*Shigella*		*yfaL*	autotransporter domain, adhesin	CL3818	
E. coli/*Shigella*	*tdcABC*-operon transcription activator	*tdcRF*	positive regulation by binding to operon promotor, operon implicated in anaerobic threonine metabolism	CL3293	Goss *et al.* (1988); Ganduri *et al.* (1993); Hazkani-Covo and Graur (2005)
E. coli/*Shigella*		*ypdABCFEGHI*	put. PTS system II	CL3167-69, CL3223, CL3291-92, CL3440, CL3892	Tchieu *et al.* (2001)
E. coli/*Shigella*	cation efflux system operon	*cusABF*	PTS system II B/C/D components	CL2015,CL2743,CL3:	Tchieu *et al.* (2001)
E. coli/*Shigella*		*agaABC*	PTS system II components	CL3612, CL3707, CL3801	Tchieu *et al.* (2001)
E. coli/*Shigella*	dihydroxyacetone kinase subunit M and regulator	*dhaHR*	glycerol metabolism operon, put. relation to PTS system	CL3143, CL3285	Tchieu *et al.* (2001)
E. coli/*Shigella*		*yddABHMU*	cluster of metabolic proteins	CL3626, CL3335, CL3577, CL3262, CL3511	
E. coli/*Shigella*	xanthine dehydrogenase subunits A and B	*yagYZ,xdhAB*	iron-sulfur and molybden binding subunits	CL3836-37, CL3460-61	Xi *et al.* (2000)
E. coli/*Shigella*		*ygeVWYZ*	protein cluster for ornithine metabolism	CL3813-15, CL3714	Xi *et al.* (2000)
E. coli/*Shigella*	carbamate kinase-like prot. (yqeA), xanthine and CO dehydrogenase maturation factor (yqeB)	*yqeABC*		CL3712-13	
E. coli/*Shigella*		*yahABDEIJ*	metabolic protein cluster with put. cytosine deaminase (yahJ) and carbamate kinase-like protein (yahI)	CL3830-35	
E. coli/*Shigella*	glycolate oxidase operon	*glcBC*	glc operon for malate synthase	CL3807-08	Pellicer *et al.* (1996)
E. coli/*Shigella*	RNA 3'-terminal phosphate cyclase	*rtcABR*		CL3179, CL3313, CL3063	
E. coli/*Shigella*	hyp. proteins	*yfdEVW*	metabolic function, formyl-CoA transferase (yfdW)	CL3182, CL3199, CL3321	
E. coli/*Shigella*		*yggCD*	Pantothenate kinase and transcriptional regulator	CL3205-06	
E. coli/*Shigella*		*ydiLNQT*	cluster of proteins involved in electron transport	CL3257-59, CL3401	Campbell *et al.* (2003)
E. coli/*Shigella*		*yjdAFIJ*	unknown function	CL2510, CL3381, CL3486-87	
E. coli/*Shigella*	DNA damage inducible protein	*dinD*		CL3245	Khil and Camerini-Otero (2002)
E. coli/*Shigella*	chemosensory pili system protein	*chpA*	put. involvement in regulation of cell growth, toxin of ChpA/R toxin-antitoxin system	CL3617	Masuda *et al.* (1993)
E. coli/*Shigella*		*ydcAHSU*	cluster for extracellular proteins	CL3404, CL3367, CL3551, CL3720	
E. coli/*Shigella*	type-1 fimbriae operon	*fimABCEFGHI*		CL3751-55, CL3490, CL3504, CL3569	Boyd and Hartl (1999)
Salmonella	oxalacetate decarboxylase	*oadABG*	contains sodium ion pump	CL3591, CL3869, CL6138	Woehlke *et al.* (1992)
Salmonella	suppression of copper sensitivity	*scsACD*		(CL4023-24), CL4611	Gupta *et al.* (1997)
Salmonella	pathogenicity island encoded proteins	*pipBB2D*	PAI-encoded proteins A, SPI5 and a protein similar to pipB	CL6145, CL6175, 5459	Wood *et al.* (1998)

Continued on next page

Group	Name	Gene Tag(s)	Function	Cluster ID	References
Salmonella	tetrathionate reductase genes, thisulfate reductase electron transport, anaerobic sulfide reductase	*ttrBS, phsB, asrABC*	sulfur reduction	CL5770-71, CL4915, CL4936-38	Price-Carter *et al.* (2001)
Salmonella	vitamin B12 synthesis pathway proteins	*cbiADEFGHLNOPQT, cobD*		CL5345-53,CL5604-05	Rodionov *et al.* (2003)
Salmonella	invasion protein antigen, surface presentation antigen	*sipACD, sigDE*		CL5744, CL4170, CL4303, CL4458-59	Hermant *et al.* (1995); Wallis and Galyov (2000)
Salmonella		*sicP, sptP, sifAB*	virulence related chaperon and effector proteins	CL5746-47, CL6147, CL6587	Marcus *et al.* (2000); Brumell *et al.* (2001); Lin *et al.* (2003)
Salmonella	secretion aparatus needle assembly	*invIJH, orgABC*	cell adhesion/ invasion	CL4865/5743/6184, CL5749-51	Collazo *et al.* (1995)
Salmonella	type III secretion system	*ssaBEIKMOP,sscAB, sseABCDEFG, STM1410*	adhesion and toxin injection	CL6150-65, CL6558	Marcus *et al.* (2000)
Salmonella		*pagD*	put. outermembrane virulence protein, PhoP activated genes, activation inside of host cells, antimicrobial peptide resistance	CL6148	Ernst *et al.* (2001); Navarre *et al.* (2005); Gunn *et al.* (1995)
Salmonella	type III secreted protein effector	*sopDE2*		CL5741, CL5455	Marcus *et al.* (2000)
Salmonella	fimbrial proteins	*sthADE, steDF, bcfABDH, fimIWY, safAD, STM4595*		CL5329/ 41/ 5477, CL7012-13, CL5788-89/ 4847/ 4160, CL5991-92/ 6436, CL7061/ 6546, CL5799	Boyd and Hartl (1999); Townsend *et al.* (2001)
Salmonella	tricarboxylic transport	*STM2786/87, tctDE*	two-component system with catabolite repression	CL5577-78, CL5579-80	
Salmonella		STM4258-62, *sfbABC*, STM0509	put. ABC transport systems	CL5727-30, CL4965/4982/5591, CL5590	Pattery *et al.* (1999)
Salmonella	cytochrome BD2 proteins	STM0360-61	subunits I and II	CL3851-52	
Salmonella	Methyl viologen resistance	*smvA*	multi-drug efflux pump	CL4619	Santiviago *et al.* (2002)
Salmonella	DNA-damage inducible protein	*dinI*		CL4979	
Salmonella		*hilD*	invasion protein regulator	CL5748	Olekhnovich and Kadner (2002)
Salmonella	aminoethylphosphonate operon	*phnSUVWX*		CL6214-17	Huang *et al.* (2005)
Salmonella	tetrathionate metabolism operon	*ttrBS*		CL5770-71	
Salmonella	put. envelope protein	*envE*		CL5777	Gunn *et al.* (1995)
Salmonella		STM2244	virulence protein, homolog of MsgA	CL6567	
Salmonella		SPA1609	toxin subunit	CL7056	
Salmonella	citrate transport	*citB*		CL6198	
Yersinia		*tccC1, tcaA1C1A*	put. toxin subunit, insecticidal toxins	CL2320, CL4632/ 4784/ 5198	Pinheiro and Ellar (2007)
Yersinia		*senA*	enterotoxin-like protein	CL5658	
Yersinia	heme aquisition system	YPO2999, *hasADE*	HlyD-family secretion protein, hemophore	CL4056, CL4769-71	Rossi *et al.* (2001)
Yersinia		*shlB, hcp6*	hemolysin activation/secretion/coregulation	CL3949, CL5414	
Yersinia	alcaligin biosynthesis	*ysuA, alcAB, entF3*	put. iron-siderophore transport system, siderophore biosynthesis	CL5074, CL5073/ 5385	
Yersinia		*fecB2B3B4E, febC, btuC3C4*	put. solute-binding iron ABC transport system	CL5382/ 4653/ 56/ 69, CL5054, CL4654-55	Koster (2005)
Yersinia		*yfuABC*	iron-(III)-binding system	CL4725-27	Schubert *et al.* (1999); Gong *et al.* (2001)
Yersinia		*livF1F2G1K1M1M2*	put. substrate binding/ABC-transporter, periplasmic transport protein	CL4283-87/4499	
Yersinia		*potB1B2C1C2D*	put. binding-protein-dependent transport system	CL4799-4800/ 5119-21	Shah and Swiatlo (2008)
Yersinia		*artI2M1M2*	put. arginine ABC-transporter	CL4812-14	Saitoh *et al.* (2005); Wissenbach *et al.* (1995); Lu (2006)
Yersinia		*mdlB2B3B7*	ABC-type multidrug-protein-lipid transport	CL4301/ 4306/ 6040	
Yersinia		*togBM*	solute-binding periplasmic protein of oligogalacturonide ABC transporter, lower part of pectin degradation pathway	CL4313-14	Abbott and Boraston (2008); Hugouvieux-Cotte-Pattat *et al.* (2001)
Yersinia	ribose and arabinose operons	*rbsB1B3B4B6B10B11DK. araC8H1H2H4H7H8H13*	sugar transport	CL4928/ 5030/ 5050/ 5125-26/ 5132/ 5388/ 5425-26, CL4939/ 5029/ 5053/ 5131/ 5374/ 5387/ 89/ 5427	Laikova *et al.* (2001)
Yersinia		*malF2F4F5F7E2G1G4G ugpB1B3B4*	sugar transport system	CL4483-84/ 4752-53/ 5101-02/ 5144-45, CL4754/ 5146/ 5145	
Yersinia		*proP3P16P17P29P34*	metabolite transport system	CL5155/ 5160/ 5177/ 5621/ 5652	
Yersinia		*pstA1B1*	put. phosphate transport system	CL5117-18	Lamarche *et al.* (2005)
Yersinia		YP_2473	urease protein cluster	CL4507	
Yersinia	urease operon	*ureABCDEFG*	can be inactivated because of point mutation (premature stop codon) in *ureD*	CL3919-21/ 4017/ 4744-45/ 6115	Sebbane *et al.* (2001)
Yersinia		*acrA9*	multi-drug efflux pump	CL4831	
Yersinia		*mgtE*	put. divalent cation transport protein	CL4295	

Continued on next page

Group	Name	Gene Tag(s)	Function	Cluster ID	References
Yersinia	general secretion pathway	YP_2839(*gspC*), *gspKL, hofG2G4G7*		CL4661-62/ 5168, CL5167/ 5644/ 4665	Sandkvist (2001)
Yersinia		YP_0412/16/19/22	put. type-III-secretion aparatus (beside the Yop-aparatus?)	CL5001-02/ 04/ 06	Cornelis (2002)
Yersinia		YP_4094/3665/67/3674-77	hyp. type-VI-secretion system	CL5016-20/ 23-24	Angot *et al.* (2007)
Yersinia		*nqrA1B1D1E1F*	$Na^{(+)}$-translocation, NADH-quinone reductase	CL4487-91	Ravcheev *et al.* (2007)
Yersinia		*ypeIR, yspR*	quorum sensing	CL4714-15, CL5135	
Yersinia		YPO3923/18, YPO3886	Colicin Js-sensitivity, Colicin S-type Pyocin	CL4039-40, CL6049	Foultier *et al.* (2002)
Yersinia		*flaA3, fliEFGH-NMQPR, flgABCDEGHJK, flhB, fleR*, YPO0720/43	flaggelar system	CL3355, CL3929/ 4250-58/ 4420, CL3930/ 4028/ 4261-63/ 4421-23, CL5420/ 5988	Soutourina and Bertin (2003)
Yersinia		*smfA, fimC2C3C4C5D3*, YPO0302-03/ 0700/ 1707/ 10/ 1922/ 2881/ 2940/ 45/ 50/ 3798-3801	fimbrial proteins	CL3517, CL4315/4495/4761/4871/5175, CL4026-27/ 4149/ 4238/ 4685-87/ 4760/ 62-63/ 5176/ 5309/ 5420	
Yersinia	put. tellurite resistance proteins	*treABDEXYZ*	induced by streptomycin in *Y. pestis*, mediates resistance to tellurite, bacteriophages and microcins in *E. coli*	CL4283-87/4499	Orth *et al.* (2007); Whelan *et al.* (1995)
Yersinia		*pilNM*	put. Tfp-pilus assembly	CL4638-39	Nudleman and Kaiser (2004)
Yersinia		YP_3000-08, YPA_3099	Flp-pilus	CL4645-50, CL4651	
Yersinia		*hmwA*	put. adhesin	CL?	Nelson *et al.* (2001)
Yersinia	carnithine metabolism, polyketide synthesis	*caiD1D2, pksG*	genetic neighbourhood	CL5066-67, CL5065	Shen (2003)
Yersinia		*uvrA2*		CL6041	
Yersinia	fatty acid biosynthesis	*fabD1F3G5G6G7G11*	short chain dehydrogenase	CL5061/ 5063-64/ 5068/ 5070/ 5655	DiRusso *et al.* (1999)
Yersinia	tight adherence	*tadG*		CL5165	Tomich *et al.* (2007)
Yersinia	transposase	*tra5D*	integrase core subunit	CL2996	
Yersinia	PH6 antigen precursor	*psaAFE*	forms individual distinct fimbrial strands on bacterial surface for adhesion to host cells	CL4717-19	Liu *et al.* (2006)
Yersinia	cold shock-like proteins	*cspC3E2*	stress adaptation	CL4716/ 5666	Yamanaka *et al.* (1998)
Yersinia	haemin storage system	*hmsS*		CL5179	Lillard *et al.* (1997)

The table lists specific protein families merged to operons or regulons wherever possible. This excerpt of the library of specific protein clusters provides information about the bacterial group in which the proteins occur, the name of the whole units if existent, the corresponding gene names, functional descriptions, a reference to the constructed database (cluster ID) and literature references.

4.3. Abundance of virulence-associated protein domains

Virulence factors were deeply studied in many strains that exhibit diverse clinical pathology (Johnson, 1991; Law, 2000; Reid *et al.*, 2000). CA was applied to profiles of virulence-associated protein domains both as a proof of concept and as a new method to explore the repertoire of virulence in whole bacterial clades. Virulence-structures were detected by HMMs of protein domains, so called profile HMMs. Protein domains are subunits of proteins with distinct functionality. The publicly available profile HMM library Pfam contains numerous profile HMMs that model virulence-associated protein domains or single domain proteins. Virulence determinants in the set of *E. coli* strains were detected with a collection of all Pfam domains annotated with a virulence-function in bacteria. If necessary, profile HMM assignments were finally merged to translate protein domain annotations to the presence of virulence structures in the categories toxins, secretion systems, lipopolysaccharides (LPS)/capsules, siderophores, microcins and fimbriae. Detailed results of the assignment were listed in Table A.1 of the appendix. Differences in virulence between *E. coli* strains were summarised by CA (Figure 4.10).

The investigated strains can be categorised into non-pathogenic (K-12 MG1655 and W3110), extra-intestinal (536, CFT073, UTI89), intestinal (O157:H7 EDL933 and Sakai, O42, E2348/69) and com-

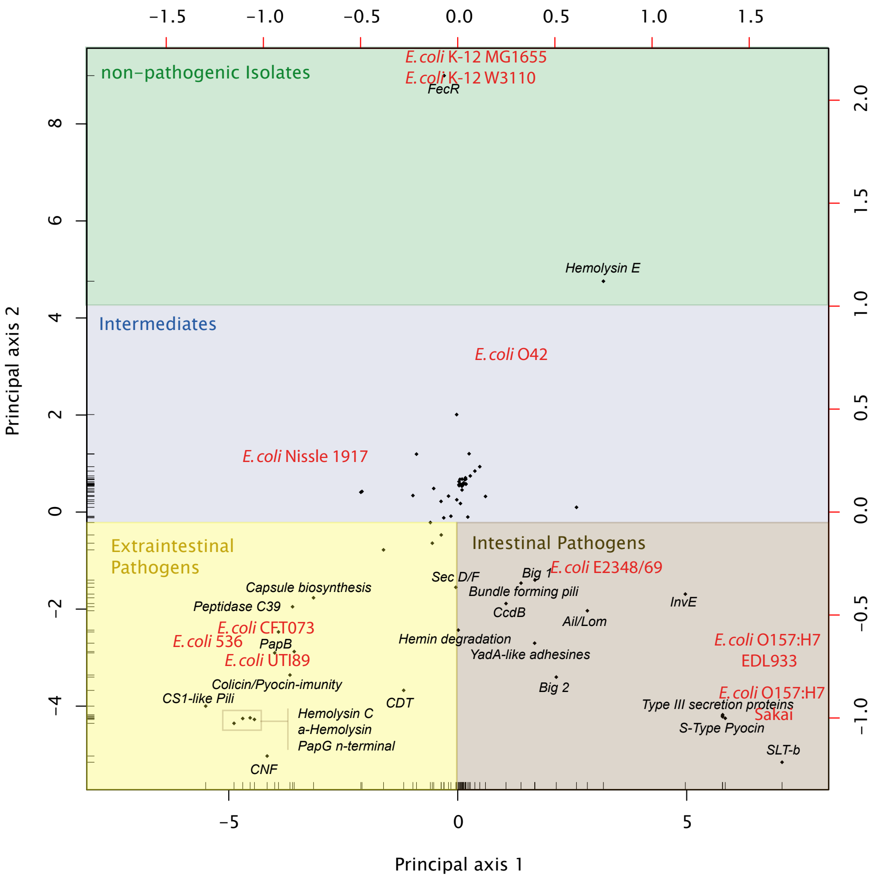


Figure 4.10.: Association of the occurrence of virulence-associated protein domains and differences in virulence between selected *E. coli* strains. The strains are selected from EHEC, EPEC, EAEC and UPEC pathogroups as well as from the group of commensal isolates. The separation according to virulence exhibits a tripartition to non-pathogenic, intestinal pathogenic and extra-intestinal pathogenic strains. The associated virulence domains correlate well with knowledge from previous reports. Abbreviations represent: Ail/Lom-like proteins (Ail/Lom), bacterial Ig-like domains groups 1 and 2 (Big1/2), cytolethal distending toxin (CDT), cytotoxic necrotising factor (CNF), shiga-like toxin β-subunit (SLT-b).

mensal (Nissle 1917) *E. coli* isolates. As described previously, the CA correlates strain specific patterns of virulence factors with occurrence patterns of virulence structures among strains. In other words, the method overlays and correlates column-wise differences (the strains) with row-wise differences (virulence determinants). Strong correlation between virulence structure occurrences and strain-specific virulence patterns is visualised in a two-dimensional plot as spatial proximity.

The subsequently described interpretation derived from the CA-plot in Figure 4.10 nicely coincides with knowledge about virulence determinants in the strains. The two non-pathogenic strains are separated from other strains by the second principal axis. They are rarely associated with any virulence factors except for FecR, a putative sensor in siderophore systems. Iron uptake is generally required for vertebrate gut colonisation and therefore siderophore systems also exist in non-pathogenic *E. coli*. The strain Nissle 1917 is positioned in the intermediate zone between non-pathogenic and pathogenic strains. The intermediate *E. coli* strain is considered as a commensal but shares many traits with UPEC isolates. The strain's location near the origin in the plot signifies the absence of specific virulence determinants. The O42 strain is an enteroaggregative *E. coli* pathogen. The classification in the intermediate zone near the non-pathogenic isolates probably arises from an absence of profile HMMs for EAEC-specific virulence factors.

The lower third of the plot is divided into a right and left part, which characterises intestinal and extra-intestinal *E. coli* isolates, respectively. The support of the first principal axis of the CA designates this separation as the most obvious difference in the whole data set. Extra-intestinal isolates were associated with toxins and adhesion factors, which were frequently described as typical virulence factors mediating UPEC infections (Oelschlaeger *et al.*, 2002). The same holds true for the intestinal pathogens and their association with virulence factors. Here, the EHEC and EPEC isolates were correlated with the presence of shiga-like toxins and the type-III-secretion system. A specific strategy of intestinal *E. coli* pathogens to combat concurrent bacteria in the gut habitat is the secretion of S-type pyocin. Interestingly, the extraintestinal group of strains was characterised amongst others by defence mechanisms against colicins and pyocins.

4.4. Conclusions

We introduced a novel methodology to draw comprehensive pictures of molecular divergence within bacterial clades. The prerequisite for a detailed investigation of specific bacterial characteristics is a fundamental knowledge about the relationships between the strains of a clade. Therefore, the phylogeny of the bacterial group in focus, the *Enterobacteriaceae*, was reconstructed regarding different levels of genomic organisation. Due to the existence of multiple intragenomic copies with different evolutionary background, the conventional 16S rRNA marker is prone to reveal ambiguous or even wrong phylogenetic reconstructions on enterobacterial subspecies levels. Thus, the marker was replaced with the single-copy gene for the RNA polymerase β-subunit, which has proved before to yield a better phylogenetic resolution. The relationships derived from the *rpoB*-based phylogenetic inference coincide with common knowledge and separate the enterobacteria involved in this study into the three macro-groups *E. coli* (with *Shigella* as a clonal lineage), *Yersinia* and *Salmonella*. The obtained macroscopic structure was taken over as a prior in subsequent analysis that required a supervised grouping.

Genome rearrangements were also considered as an important process in bacterial evolution. This aspect of enterobacterial evolution was set in relation to molecular phylogeny. On the macroscopic level both measures clearly separate the investigated enterobacteria into three macro-groups. But, genome rearrangement occur with different rates between otherwise closely related strains and rather reflect a rapid mechanism of genome reorganisation. The effect of genome rearrangements was previously stated as a rewiring on the transcriptional level with a potential impact on adaptation and altering lifestyle (Pérez *et al.*, 2008). Obviously, rearrangements play an important role in *Shigella* divergence from *E. coli* and may contribute to the *Shigella*-specific ability to invade host cells upon infections. The ability to acquire considerable frequencies of genome rearrangements was only found in *Yersinia* and *Shigella* strains. *Shigella* strains are equipped with a large amount of IS elements, which mediate genomic reorganisations (Touchon *et al.*, 2009). The number of IS elements varies in *Yersinia* isolates, but is not as large as in *Shigella*. Though recent studies located rearrangements among *Yersinia* mainly at sites of IS elements (Chain *et al.*, 2006), other processes like recombination events seem to be the major cause for the divergence in genome organisation from *E. coli*, *Klebsiella* and *Salmonella*.

Recombination events and mobile genetic elements are mainly responsible for the flexible part of the enterobacterial gene pool. In accordance to large genome plasticity, an open complemental genome was determined with an introduction of more then 250 new genes per discovered strain. The complemental genome fraction far exceeds the complemental genome that was recently approximated to 79 genes for the *E. coli* clade alone (Willenbrock *et al.*, 2007). The large increase can surely be attributed to the introduction of genomic variability by *Shigella* (Touchon *et al.*, 2009) and *Yersinia* strains. The core part of the gene pool in enterobacteria is smaller than the core genome of *E. coli*, a fact which could either indicate non-orthologous replacement or a far smaller essential genome size.

With the knowledge of basic differences between enterobacterial genomes at hand, I approached the detailed determination of patterns in the flexible part of the proteomes that are common to members of enterobacterial subgroups. In order to determine sets of proteins that support a divergence of whole groups of strains and thereby a divergence of lifestyle, pathogenicity and/or metabolism, respectively, we developed independent methodologies based on CA and statistical testing. The CA provides an unsupervised view on striking differences between features of strains as well as characterisations of whole bacterial groups. Unfortunately, the CA was unsuitable in filtering certain group-specific characteristics, as it spatially locates characteristics of whole groups and those of its single members at similar coordinates of the principal axes. These shortcomings were circumvented by the application of statistical tests for the assignment of a significantly specific occurrence of a protein family in the bacterial group in focus. The application of the methodology on the three major groups of our dataset resulted in numerous assignments of group-specificity to protein families and even whole operons or regulons. The specific character could be confirmed throughout all groups by detailed expert investigations of selected protein families, operons or metabolic pathways. The suggested methods substantially improve the concurrent detection specific traits of members of a bacterial group. Many yet uncharacterised specificities constitute starting points of future investigation and concurrently indicate a source of uncertainties concerning previously derived characterisations of enterobacterial subgroups.

Functional characteristics of enterobacterial subgroups could also be broken down to functional subunits of proteins, the protein domains. The potential to discriminate the functional repertoire of bacterial groups by CA on the basis of protein domains therefore was evaluated by a mapping of virulence-related domains in a selection of representative *E. coli* isolates. This analysis revealed a strong conformity with prior expectation concerning the association of virulence factors to *E. coli* isolates. Various virulence factors have been described as determinants for a uropathogenic phenotype in human infection like type-1 and P-fimbriae, cytotoxic necrotising factor and α-haemolysin (Lloyd *et al.*, 2007). In the presented CA of virulence factors in *E. coli* these occurrences could be confirmed as specific determinants for UPEC virulence. The same holds true for intestinal pathogenic *E. coli*. EHEC isolates are generally characterised to harbour a type-III-secretion system, shiga-like toxin, while bundle forming pili (BFP) are considered as a major virulence factor in EPEC (Kaper *et al.*, 2004).

Comparisons based on protein domain representations, especially HMMs, provide the advantage of a well established framework for their detection even across large phylogenetic distances. However, the generality of the protein domain concept might lead to misinterpretations, if certain functionality depends on a co-occurrence of domains in a protein. An example would be the group of proteins harbouring a fimbrial domain (PF00419) that is involved in fimbriae construction. The domain is found in enterobacterial proteins exhibiting three different domain architectures. Proteins with a single fimbrial domain constitute structural subunits in fimbriae, while proteins composed of a fimbrial and a FimH-domain (PF09160) mediate fimbrial adhesion to mannose. The combination of a fimbrial and a lectin domain (PF09222) is found in ETEC-specific fimbriae that mediate adherence by binding to lectin (Buts *et al.*, 2003). Although in general the functional repertoire of enterobacterial strains can be deduced from the nature of protein domains, domain modularity needs to be considered in comprehensive functional analysis.

The completeness of the picture that can be drawn from the developed approaches to unravel enterobacterial properties strongly depends on the considered genomic information. At the time of these studies genome sequences of enterobacterial subgroups like the genus *Klebsiella*, the species *Y. enterocolitica* or several *E. coli* pathotypes were underrepresented. Most likely, this will change in near future and the conclusions that can be drawn from comparative genome or proteome analysis will gain in reliability. A crucial factor in all genomic analyses is the available amount of annotation for protein families. The assigned specific protein families largely lack a reliable functional annotation of its family members. Nevertheless, our approach provides starting points for detailed experimental investigations of potentially elementary proteins for the development of bacterial phenotypes. The method is flexible concerning the nature of groups to be compared in a certain bacterial clade and as well can be applied to investigate differences in other bacterial taxa.

5. Development of a diagnostic microarray for clinically relevant enterobacteria

5.1. Design of a microbial diagnostic microarray

5.1.1. Concept of microarray design

Our developed strategy to design a diagnostic microarray based on a new set of pathogroup-specific determinants is structured according to clinically distinct pathogroups of enterobacteria. Figure 5.1 depicts these subdivision assigned to the *Enterobacteriaceae* and illustrates the nested relations associated with the large group of clinically relevant and versatile *Shigella* and *E. coli* strains. The hierarchical dendrogram is further denoted as the pathogroup tree. The subdivisions and comparisons applied in this case were guided by clinical relevance. Due to co-evolution and horizontal gene transfer of virulence-associated traits the taken divisions do not coincide with phylogenetic reconstructions (Wirth *et al.*, 2006). The groups with main impact in the described analysis generally occupy terminal positions in the diagnostic structure. But, superior relationships between the strains were, however, simultaneously considered as complementary indications of the pathogroups' nature. The comparisons were therefore split into the intrinsic three main levels of organisation within the pathogroup tree: (I) the genus level, (II) the distinction between *Shigella*, pathogenic and non-pathogenic *E. coli* as well as (III) the diversity among intestinal and extraintestinal *E. coli* pathotypes.

5.1.2. Probe selection

The sequences of genomes and available plasmids of reference strains (see Table 1.1) were subjected to a probe selection procedure in order to find capture probes that provide a high discrimination between the entities of the pathogroup tree. The strategy of probe selection was based on a global extraction of group-specific 70-mer oligonucleotides by the application of longest common factor statistics. The string matching algorithm yielded sets of commonly found oligonucleotides without prior restrictions in strain composition of the resulting groups. Oligonucleotides had to meet the criteria of unique full length matches to all reference genomes of the respective group and a maximum of 14 consecutive matches in alignments to any other sequence in the set of other enterobacterial genomes. The provisional pool of probes ($\sim$18,000 oligonucleotides) was expert curated to select only those probes characterising enterobacterial subgroups with clinical importance. An even larger provisional pool of probes ($\sim$360,000 oligonucleotides) was obtained by applying less stringent cross-matching criteria, but the large size of the stringent provisional pool put the need for an additional oligonucleotide source aside. A set of candidate probes was carefully selected from the pool of provisionals according to cross-matching behaviour to human DNA and conventional hybridisation parameters like compositional complexity, GC-content, change in Gibb's free energy and melting

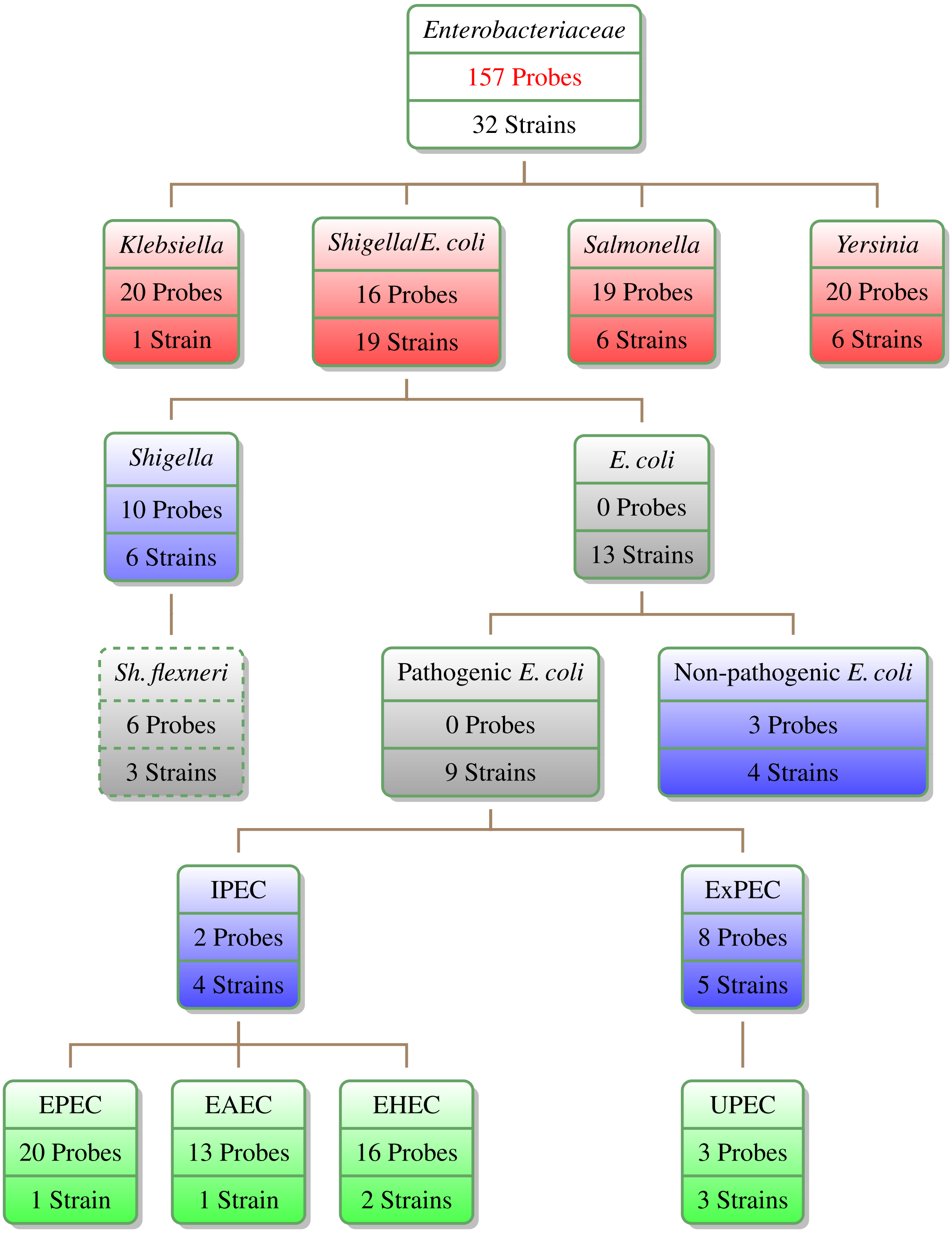


Figure 5.1.: Overview of assigned clinically relevant *Enterobacteriaceae*. Each node corresponds to a pathogroup entity and the respective box comprises information about the number of probes designed for the respective group as well as the number strains assigned to the group according to prior knowledge. The colours refer to the genus level (red), the intermediate *E. coli* level (blue) and the *E. coli* pathotype level (green). Gray colour refers to pathogroups for which no probes could be found and the white box titled '*Enterobacteriaceae*' summarises the assignment.

temperature. Reverse complementary oligonucleotides were considered as autonomous candidate probes even if they fully overlap, as the difference in base composition may have an influence in hybridisation properties.

The choice of probe length implicates a trade-off between the signal sensitivity and the size of the pool of capture probes. While longer oligonucleotides yield higher sensitivity, they implicate in parallel a smaller source of oligonucleotides from which capture probes can be selected. By reducing the probe length, the resolution of subspecies level diagnostics would be enlarged while accepting a higher risk of cross hybridisations. Previously, 70-mer oligonucleotides were determined as optimally sized microarray probes, which even exceed signal intensities yielded by 100-mer capture probes (Letowski *et al.*, 2004).

The objective to construct a slim and cost efficient diagnostic tool as well as technical specifications of the slide format raised the need for a strict limitation of the probe set size. Thus, no more than 20 capture probes were selected per pathogroup. Interestingly, it was impossible to define pathogroup determinants for the generic entities '*E. coli*' and 'pathogenic *E. coli*' implicating the absence of concise genotypes across the respective strains. Pathogroups like UPEC or *Shigella* yielded less than 20 capture probes. But, the discrimination power rather depends on the uniqueness and quality of probes than on the number of determinants. Figure 5.1 provides a detailed overview of the pathogroups and the number of capture probes assigned to them. The topmost node titled '*Enterobacteriaceae*' does not characterise a pathogroup but provides a summary of probe selection, which resulted in a probe set of 157 capture probes derived from 32 reference genomes.

The number of reference strains per enterobacterial subgroup depended on the availability of completed or nearly finished sequencing projects. Thus certain assigned subgroups were underrepresented at the time of chip design (06/2006). Furthermore, genome sequences were not yet available for the *E. coli* pathotypes ETEC, EIEC, SEPEC and MNEC. To compensate for limitations in the availability of genomic sequences in certain pathogroups, comprehensive test hybridisations were conducted to verify the discriminative power of the chosen capture probes.

By means of initially unrestricted group-wise probe selection we could specify probes separating *S. flexneri* as a *Shigella* subgroup, though no special emphasis was put on such a subdivision. As *S. flexneri* does not cause substantially different clinical symptoms upon infection, the subgroup was not separately analysed. But, the ability to distinguish between *Shigella* subgroups underlines the high sensitivity in strain typing mediated by the applied probe selection strategy.

5.1.3. Characterisation of capture probes

Selected oligonucleotide probes were mapped to the genomes of the respective groups by a BLAST search to find general annotations of corresponding group-specific, genomic regions. The annotations were summarised to the categories listed in Table 5.1 as column labels. In accordance to the applied, generalised probe selection strategy, nearly 13% of probes originated from intergenic regions. The high number of intergenic probes reflects the importance of such regions in pathotype diversity and putatively as well in virulence. Some oligonucleotides represent regions with direct connection to virulence-associated genes, like toxins (IPEC probe 2638_3: *espF*) or genes located on pathogenicity islands (PAIs, UPEC probe 1540_10). Others were assigned to genes with a regulatory function or with genes of extracellular structures (genes encoding proteins residing in periplasma, the outermembrane or secreted proteins). Interestingly, the majority of selected probes refers to genes with poor or missing annotation. Clearly, these probes were not yet considered as characteristic markers for certain enterobacterial subgroups.

Table 5.1.: Overview of oligonucleotide markers and their categorisation.

Group	Intergenic	Virulence	Uncharacterised	Transcription	DNA mobility	Adhesion	Extracellular	Metabolism	Transport	Others	Probe set
Yersinia	4	0	11	0	0	0	0	4	0	1	20
Klebsiella	7	0	5	1	0	1	0	4	2	0	20
Salmonella	1	0	6	0	0	2	0	4	3	3	19
Shigella/E. coli	1	0	3	4	0	0	1	7	0	0	16
Shigella	1	4	0	0	0	1	0	0	0	4	10
Non-pathogenic	1	0	0	0	0	0	0	1	1	0	3
ExPEC	0	0	4	0	0	0	2	2	0	0	8
IPEC	0	1	1	0	0	0	0	0	0	0	2
UPEC	0	0	2	0	0	0	1	0	0	0	3
EHEC	4	0	10	0	0	1	1	0	0	0	16
EPEC	0	0	16	1	0	1	0	1	0	1	20
EAEC	1	0	17	0	0	1	0	0	0	1	20
In total	20	5	75	6	0	7	5	23	6	10	157

The term “Probe set” refers to the contribution of genomic groups to the final set of probes. Beside several probes in categories like virulence, extracellular (secreted proteins) or transcription, the probe set comprises a relatively large fraction of probes originating from intergenic regions.

5.1.4. Test hybridisations

Samples of genomic DNA extracted from representative strains of various enterobacterial pathogroups were hybridised to the microarray in order to determine its classification performance. As inferable from Table 1.3, both, genomic DNA of reference strains and of a large number of clinical isolates representing the whole range of pathogroups was prepared as test samples. Moreover, I conducted test hybridisations with genomic DNA from *E. coli* pathotypes ETEC, EIEC and SEPEC without representation on the microarray. As no special probes were selected for the classification of these pathotypes, they could be considered as a kind of negative test with respect to the pathotypes in focus.

Faecal samples as well as many clinical specimens are composed of mixed bacterial communities comprising pathogens and non-pathogens. The evaluation of the microarray accounts for these types of clinical diagnostics by specifically designed spike-in experiments. Mixed culture gDNA was prepared according to Table 1.4. The experiments target evaluations with respect to the contrasting ability of the microarray in the background of multiple bacteria and the sensitivity in determining proportions of their occurrence in clinical samples.

5.2. Assessment of single probe performance

Comprehensive test hybridisations delivered insights into the reliability of single group-specific capture probes in the classification of respective pathogroups. The probe-specific contribution in group separation was estimated by an analysis of variance (ANOVA) on signal intensities. The direction of

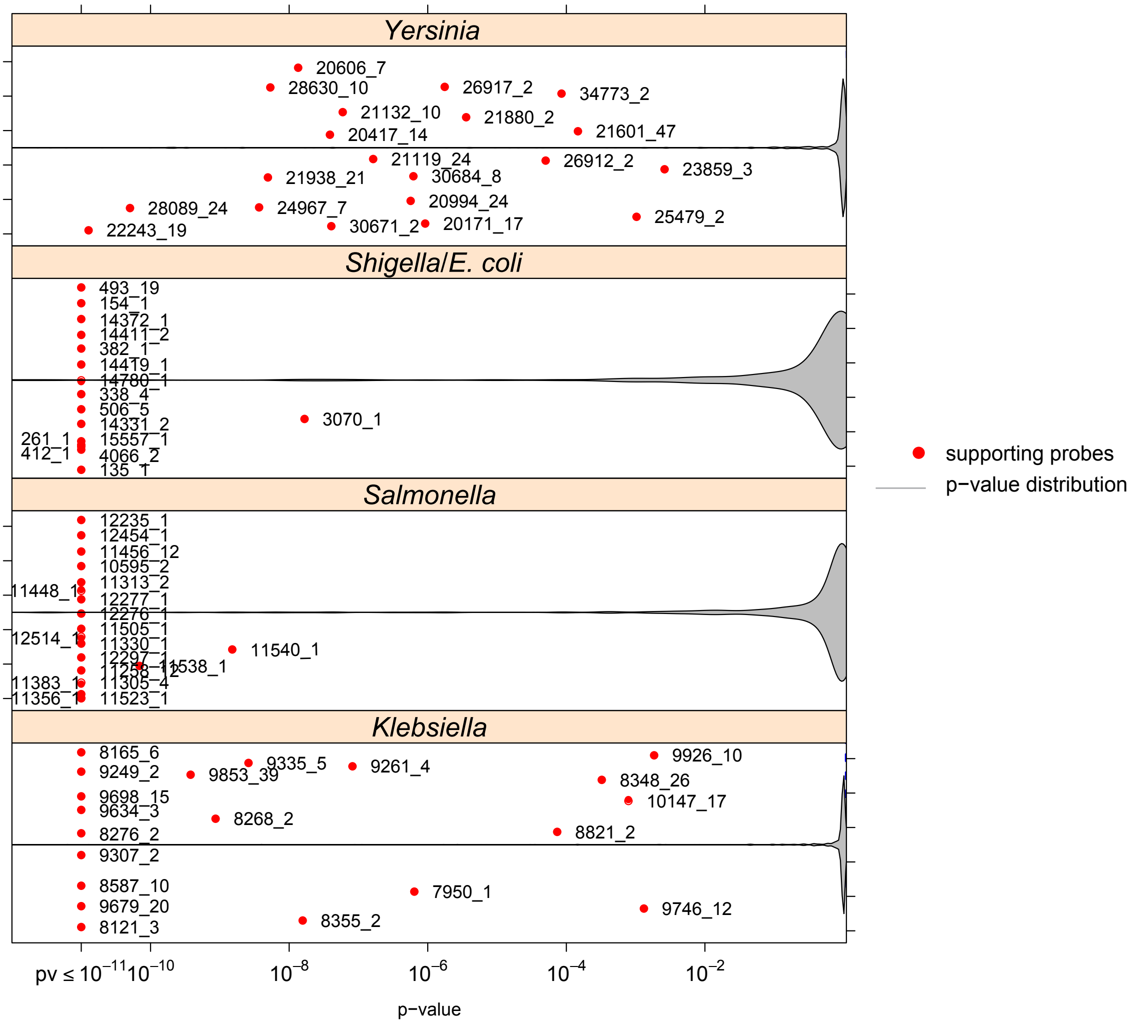


Figure 5.2.: Probe-specific contribution to the detection of diagnostic groups in the genus level of the pathogroup tree. The performance of single group-specific probes in the detection of pathogroups was determined by a combination of an ANOVA and simultaneous inference of one-sided multiple comparisons. The resulting adjusted p-values communicate the robustness of intensity difference between pairs of pathogroups. The p-values from single comparisons were averaged for each pathogroup in log space. Violin plots indicated the overall distribution of non-averaged p-values on a log-scaled x-axis as relative densities (density values are not reflected by the y-axis). The y-axis follows arbitrary units in order to improve readability of single points. Probes of the genus level largely exhibit significant single discrimination, though *Yersinia* and *Klebsiella* probe support is inferior putatively because of higher intra-group diversity and lower coverage, respectively.

the probe support was determined by the method of simultaneous inference of multiple comparisons. Adjusted p-values of one-sided pairwise inference against all pathogroups were averaged in log space in order to obtain group-specific indications of support of single capture probes. The averaged p-values of probes in respective groups are contrasted in Figures 5.2 to 5.4 against p-value distributions of all probes in the corresponding group. The p-value distributions are visualised as arbitrary densities of so-called violin plots on a log-scaled p-value axis (x-axis), which is cut at a p-value of 10^{-11}. Small densities at low p-values highlight the success of probe selection as the group-specific probes form the body of lowest p-values.

As revealed by Figure 5.2, p-values of probes specific to genus-level pathogroups generally com-

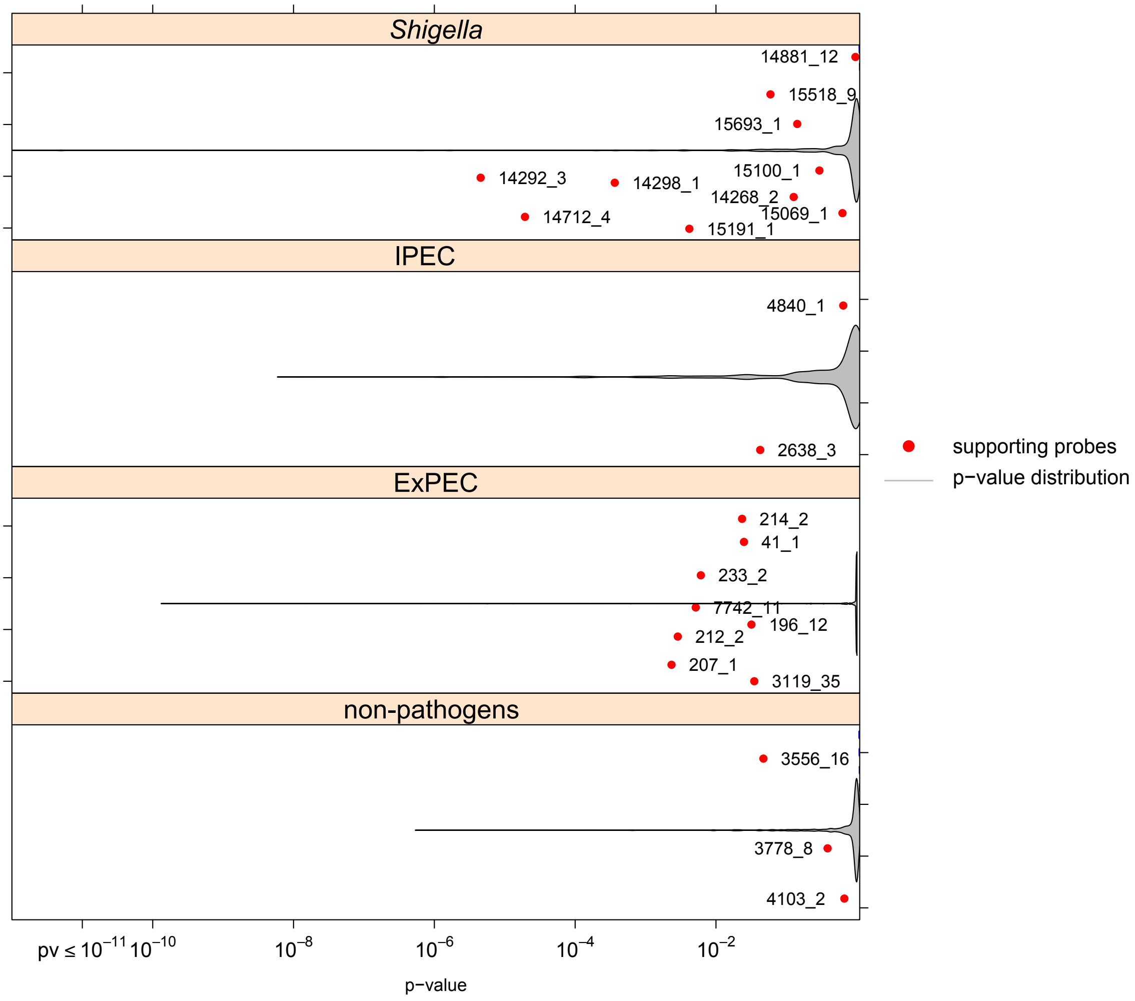


Figure 5.3.: Evaluation of single probe support within the intermediate level of the pathogroup tree. The significance of discrimination power of single group-specific probes was determined as described above. The minor support in six *Shigella* probes arises from its specificity to the *S. flexneri* subgroup of *Shigella*. The generally lower p-values in comparison to the genus level indicate putatively results from closer relationships of the groups and therefore a smaller genetic variability.

municate high confidence in the ability to classify respective strains. In comparison, the genera exhibit differences in the overall performance of their specifically designed probes. Best support was obtained for *Salmonella* and *E. coli* pathogroups while lower but still significant p-values were assigned to probes selected from *Klebsiella* and *Yersinia* genomes. These results seem to arise from quite different influences. The probes of the *E. coli* group were selected against the background of large amount of genomic data. *Salmonella* constitutes a pathogroup with a largely homogenous genotype (Porwollik *et al.*, 2004). The observed larger variability in *Klebsiella* probe performance reflects the sparse genomic data available in this group. *Yersinia* probe variability seems to mirror the genotypic diversity among *Yersinia* ssp. strains (Hinchliffe *et al.*, 2003; Zhou *et al.*, 2004).

In general, lower p-values of capture probe performance were obtained in the intermediate level downstream of the *E. coli* group mainly because of closer evolutionary distances between these groups compared to the distances in the genus level. Figure 5.3 depicts probe performance evaluations among the pathogroups denoted ‘*Shigella*’, ‘non-pathogens’, ‘IPEC’ and ‘ExPEC’. The evaluation of

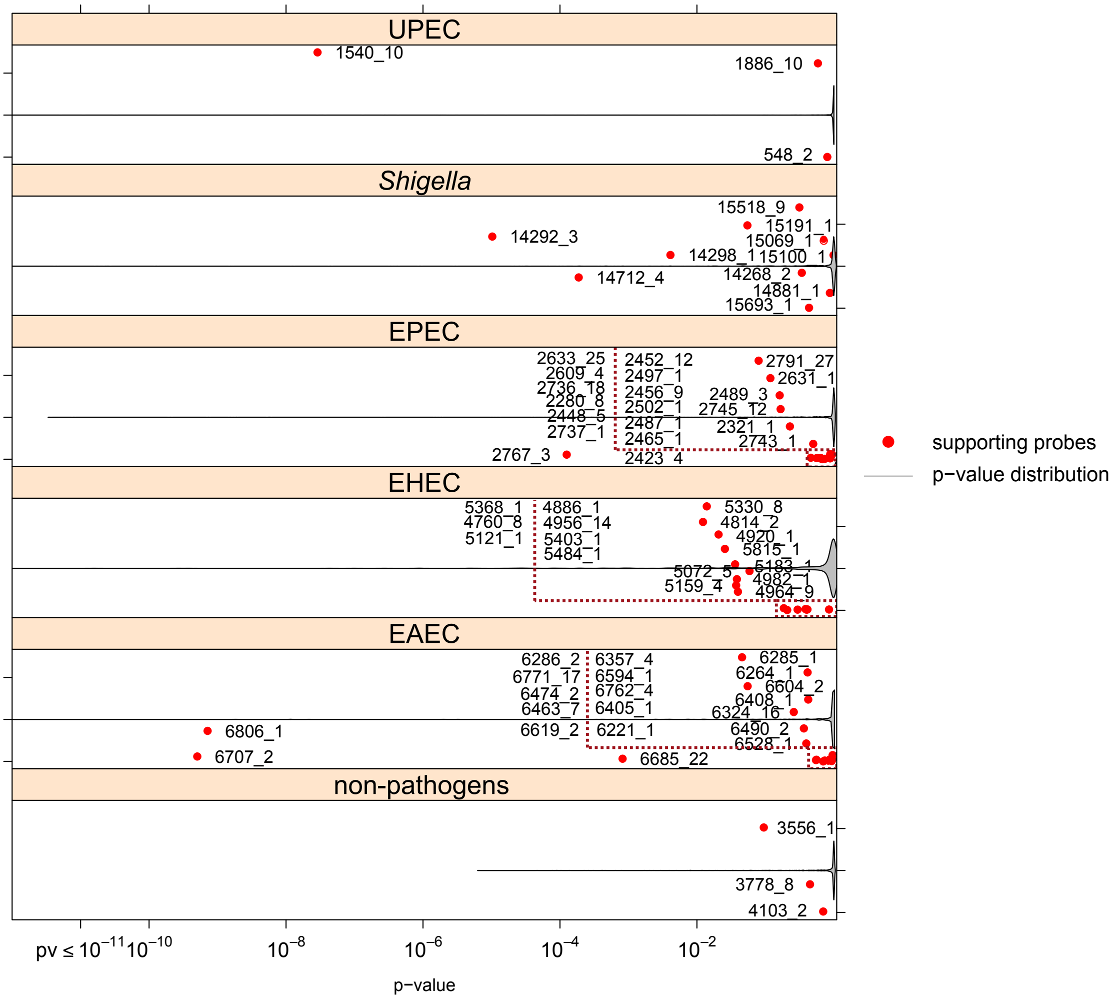


Figure 5.4.: Discriminative power of group-specific probes among *E. coli* pathotypes. The ability of probe-wise discrimination derived from ANOVA-based simultaneous inference of one-sided multiple comparisons further decreases in the *E. coli* pathotype level of the pathogroup tree. Single red dots mark averaged adjusted p-values of group-specific probes. The overall distribution of p-values as indicated by violin plots exhibits a general increase of p-values. The increase is influenced by a larger number of comparisons and by the close relation of the target groups.

Shigella-specific determinants comprises four capture probes classifying all *Shigella* strains as well as those specific only for *S. flexneri*. The corresponding plot reveals significant support by the capture probes representing the whole group. The three top-performing *Shigella*-specific probes originate from the locus of the invasion plasmid antigen H gene (*ipaH*). Venkatesan *et al.* (1989) described motifs of the gene locus to be effective predictors of *Shigella* and EIEC virulence.

Figure 5.4 reflects averaged single capture probe performance in terminal pathogroups. Increased p-value level resulted first of all from a higher number of pairwise comparisons. But, the group-specific probes of *E. coli* pathotypes still constitute the lowest fraction of the overall p-value distribution in each group. The separation of the enteroaggregative pathotype is strongly supported by two probes. One of these high-performing probes with the ID 6806_1 is located in the plasmid encoded *aatD* gene locus.

5.3. Evaluation of hybridisation

The global aim of any diagnostic means is the detection of and the distinction between targets, here antimicrobial resistance and clinically relevant enterobacterial pathogroups, respectively. In the following, the ability of the developed microarray to come to such decisions is described. Comprehensive test hybridisations provide the basis for these investigations.

5.3.1. Regression analysis

In order to predict the membership to a diagnostically relevant group (see Figure 5.1) a regression model was trained with the results from test hybridisations analogous to the method described by Engelmann *et al.* (2009). The regression model treats intensities of single probes independently from one another because of probe specific hybridisation behaviour. The target affinity to perfect match probes is dependent on the probe-specific sequence composition and does not allow for direct comparison of intensities from hybridisations to different probes. Given the intensity matrix of hybridisations Y with probes $i = 1 \ldots n$ as rows and samples $k = 1 \ldots m$ as columns and a master table containing hybridised amounts of DNA X of the same size, the linear regression model equates to

$$Y = AX. \tag{5.1}$$

The affinity matrix A is trained by solving the equation

$$\hat{A} = YX^T(XX^T)^{-1} \tag{5.2}$$

The prediction performance of the regression model was determined by leave-one-out cross-validation. In a recurrent sampling procedure the regression model was trained in each run by all but a single hybridisation pattern, which further on served as test pattern. Based on the test pattern the amount of corresponding gDNA $\hat{x}_k$ was predicted to

$$\hat{x}_k = y_k\, A^T \tag{5.3}$$

with y_k being the intensity vector of test sample k. According to the specifications for the microarray technology 2 µg of bacterial gDNA were hybridised in each single experiment. Based on prior knowledge on the true nature of test strains a master table X was generated, which refers to the hybridised amount of DNA in each pathogroup. All capture probes characterising a certain pathogroup or its parent group of a test strain were set to an appropriate factor of hybridised DNA (for pure cultures 1.0 $\cong$ 2 µg), while 0.0 was assigned to all other probes. The factor corresponds to the proportion of the sample DNA coming from a certain pathogroup and drops only below one in mixed culture hybridisations. Predicted amounts of hybridised DNA for single probes are mapped back to the pathogroup by taking the median of all pathogroup-specific probes. Each pathogroup was evaluated by samples from different strains. Groups with no explicit representations in the probe set were treated separately. In these cases the amount of hybridised DNA was determined by a regression model trained on all core pathogroups.

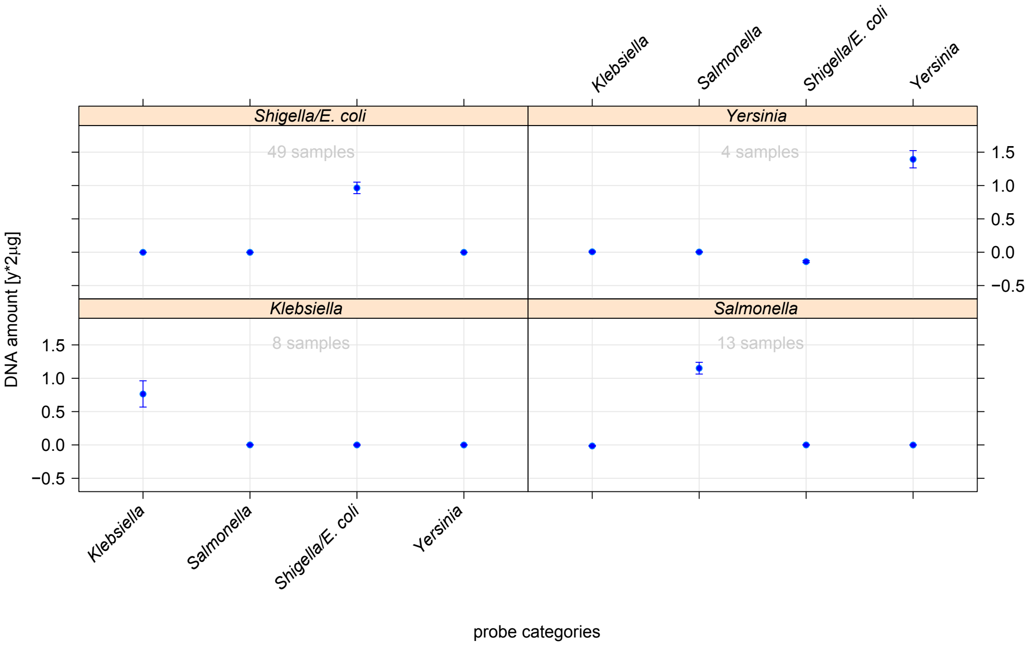


Figure 5.5.: Evaluation of the prediction performance on the genus level of the decision tree. In this case the linear regression model was trained on signal intensities of probes representing the main genera of considered Enterobacteria (*Salmonella*, *Shigella*/*E. coli*, *Klebsiella* or *Yersinia*, the x-axis). The model was trained with all hybridisation patterns. The medians with standard error of predicted DNA amounts were obtained by leaf-one-out cross-validation.

Pure cultures The regression model-based cross-validation has been determined in the context of the previously denoted intrinsic levels of the pathogroup tree. Figure 5.5 summarises the classifications on the genus level of enterobacteria subdivided to prediction outcomes of the pathogroups '*Shigella*/*E. coli*' (top left), '*Yersinia*' (top right), '*Klebsiella*' (bottom left) and '*Salmonella*' (bottom right). The headline of each plot refers to the true nature of the test samples and the x-axis represents the pathogroups, which are contrasted in regression model analysis. No misclassifications were obtained from cross-validations on the genus level. In contrast, the regression model exhibited the ability to accurately predict DNA amounts used for hybridisation. The tests furthermore suggest an influence of sample coverage in the accuracy of quantitative predictions.

Next, the classification accuracy among the branch of *Shigella* and *E. coli* strains was evaluated in more detail, starting with the intermediate level of the pathogroup tree (pathogroups shaded in blue in Figure 5.1). Figure 5.6 depicts the prediction results subdivided according to classes on this level. The training of the regression model was restricted to intensity data from probes designed for respective pathogroups of the intermediate level. Even in the narrow evolutionary spectrum of *E. coli* isolates the regression model was able to safely separate hybridisation patterns of *Shigella*, ExPEC and intestinal strains. Again, the level of prediction noise in non-target pathogroups was basically absent except for a reciprocal interference between non-pathogenic and IPEC. As described below, this interference could be resolved in predictions contrasting the non-pathogenic pathogroup against *E. coli* pathotypes (see Figure 5.7).

Certainly, the classification of *E. coli* pathotypes depicted in Figure 5.7 constitutes the most difficult

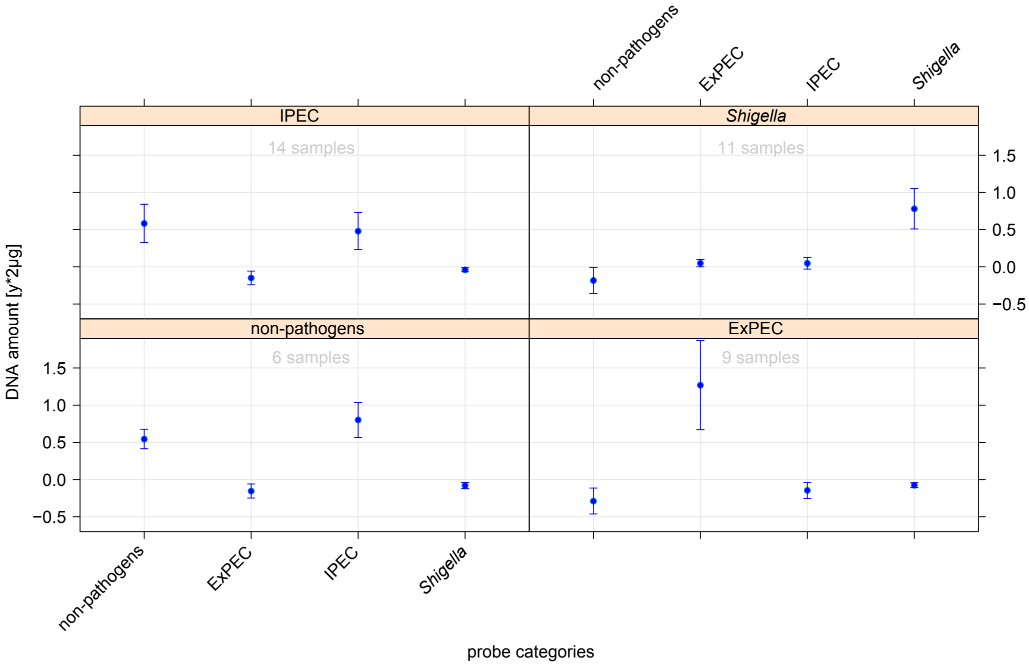


Figure 5.6.: The prediction of hybridised DNA of the groups beneath the node of *E. coli* and *Shigella* isolates. For the distinction of hybridisation of these groups the regression model was trained only with signal intensities of probes associated to contrasted groups (x-axis). For the commensals and intestinal pathogens we obtained a reciprocal cross-hybridisation, which probably has arisen from the low number of probes in these groups. In contrast, extraintestinal pathogens and *Shigella* isolates were predicted to an expected type and amount.

classification scenario because of a generally low number of reference genomes in these pathogroups, close phylogenetic relation with largely similar genotypes and high frequency of genetic interchange. Figure 5.7 refers to cross-validations of hybridisation patterns of the *E. coli* pathotype level, which was extended with clinically relevant contrasts to *Shigella* and non-pathogenic *E. coli*. The contrasting of non-pathogenic *E. coli* patterns against concrete intestinal and extraintestinal pathotypes yielded clear predictions in each pathogroup. In all classification, the prediction level of the true class can be robustly separated from prediction levels of respective negative classes.

The robustness of model predictions was further evaluated by predictions on hybridisation patterns from isolates of new pathotypes in terms of the developed microarray. These groups comprise EIEC, ETEC and SEPEC pathotypes. With respect to equivalence, patterns of these pathogroups were set in contrast to other *E. coli* pathotypes. The predictions are graphically displayed in Figure 5.8. They do not reveal a clear tendency to any pathotypes. Only the hybridisation patterns of EIEC isolates show a certain hybridisation to probes of intestinal pathotypes and *Shigella* isolates. The observed interrelation between *Shigella* and EIEC classes coincide with the high similarity of enteroinvasive *E. coli* and *Shigella* isolates concerning pathogenicity and genotype. The absence of any positive prediction for ETEC and SEPEC pathotypes as well correlate with prior expectation.

Classification of bacterial communities Furthermore, the linear regression model was trained with specifically designed spike-in experiments to detect different pathotypes within mixed bacterial cultures. Though the regression model was not specifically trained with hybridisation patterns of

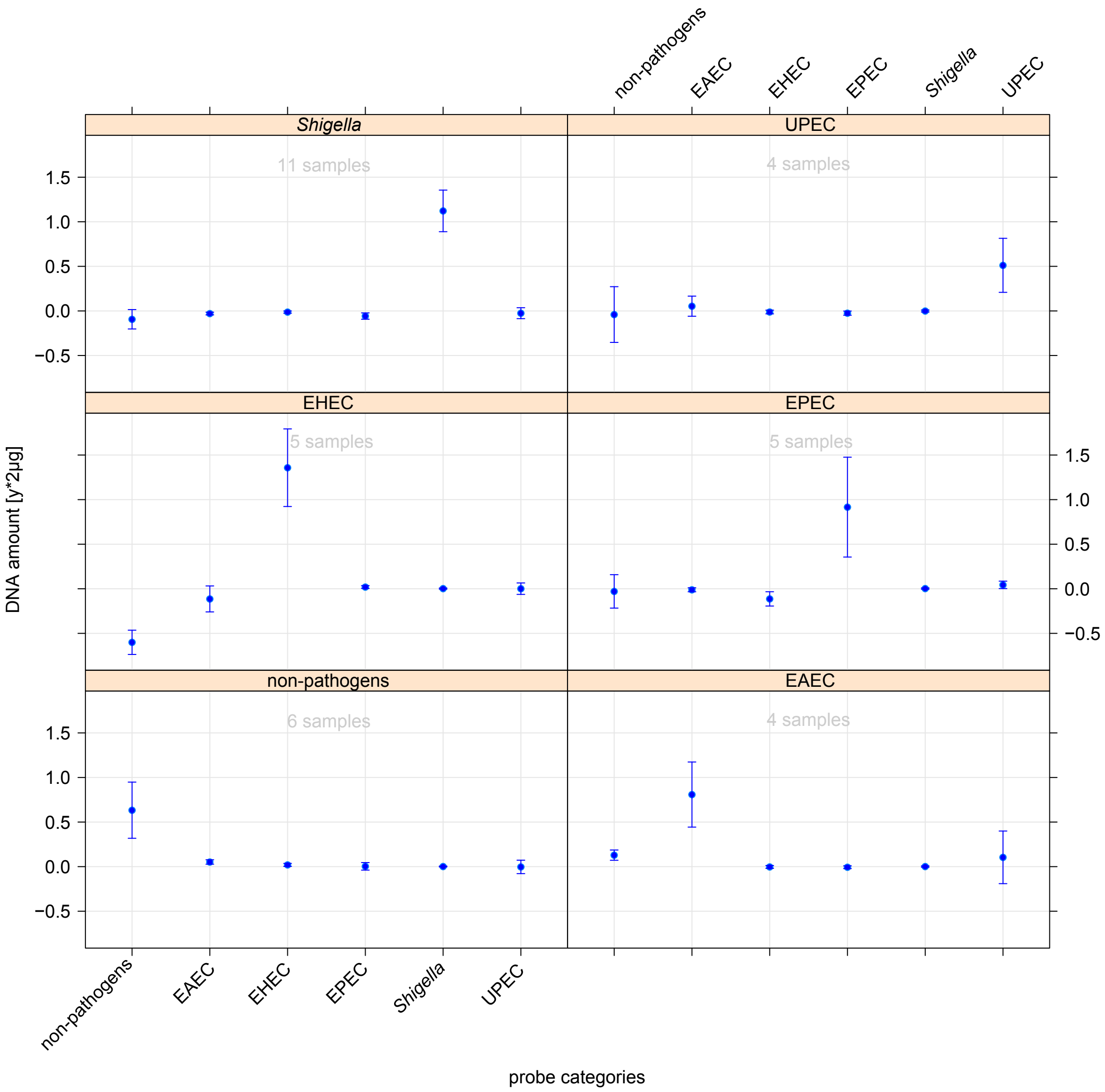


Figure 5.7.: Prediction performance of the regression model applied to the classification of *E. coli* pathotypes. The model was trained with hybridisation patterns from all contrasted pathotypes, which are indicated on the x-axis. The blue dots correspond to median values of predicted amount of hybridised gDNA from respective bacterial pathogroups. The error bars signify standard errors determined from the test samples in each class.

mixed culture sample, the predictions shown in Figure 5.9 did not only correlate with the true nature of test strains but also correctly quantify the underlying proportions. Especially the spike-in series with counterrotated proportions of a non-pathogenic *E. coli* and an EHEC strain (Plots M01-M05) demonstrate the sensitivity of the regression model in estimations of quantities of bacterial DNA and its mixtures. Though the tests for the prediction quality on mixed cultures did not fully characterise the model performance, they indicate the potential to establish an accurate predictor. By conducting comprehensive tests with biological repeats, the prediction performance of the regression model can certainly even be improved as shown for pure culture predictions. Mixed culture test hybridisations did not reveal any limit of detectable rates of pathogens though it most likely exists. If such a limit is under-run - a possible scenario for faecal-samples diagnostics - appropriate measures have to be

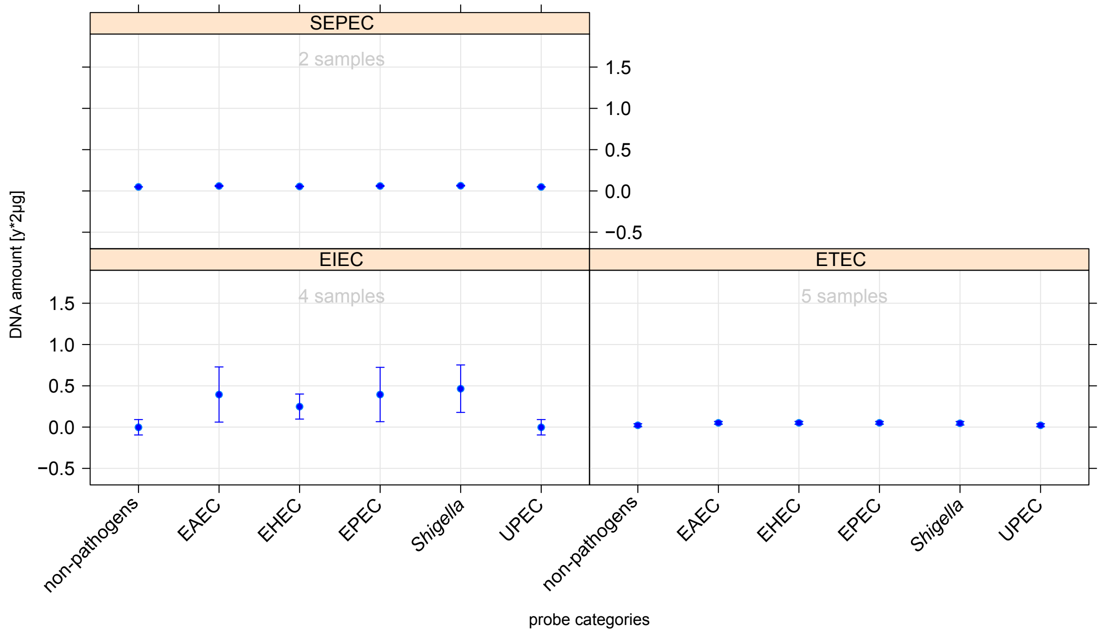


Figure 5.8.: Regression model behavior on the categorical prediction of hybridisation patterns from new pathotypes that are not represented by specifically designed oligonucleotides. The model training was based on the core pathotypes. The unspecific representation resulted in diffuse prediction outcome, where only the group of enteroinvasive *E. coli* shows cross-reactions to probes of *Shigella* and intestinal pathogens.

taken to scale up group-specific DNA ratios in question.

5.3.2. Antimicrobial resistance screening

The developed diagnostic microarray comprises features to screen for basic antimicrobial resistance patterns in enterobacterial samples and communities. A set of 30 previously published AMR markers was extended with 12 newly designed probes. The AMR probe set comprises resistance mediating enzymes and efflux pumps against aminoglycosides, β-lactams, sulfonamides, tetracyclines, dihydrofolate reductase (Dhfr) inhibitors, amphenicols and macrolides. Due to the parallel architecture of the microarray platform, the capacity of the spotting area was limited. Therefore the AMR diagnostics could only afford to provide a detection of selected AMR markers.

AMR relevant, log normalised signal intensities of hybridisation patterns from all test strains were classified into a signal and a noise fraction by fitting a Gaussian mixture model composed of two normal distributions on all data points. The left plot of Figure 5.10 summarises single posterior signal probabilities of AMR probe intensities obtained from numerous test hybridisations as a heatmap (red colour gradient). For about one third of hybridisation profiles, mainly originating from *E. coli* and *Shigella* isolates, no resistance could be detected. All but one tested *Salmonella* strains exhibited resistance to trimethoprim (genes *dhfrXIII* and *dhfrXV*), whereas neither isolate revealed any resistance to sulfonamides. These two therapeutics are frequently applied in combination. A second large right-most cluster, mainly consisting of pathogenic *E. coli* and *Shigella* strains, hold multiple resis-

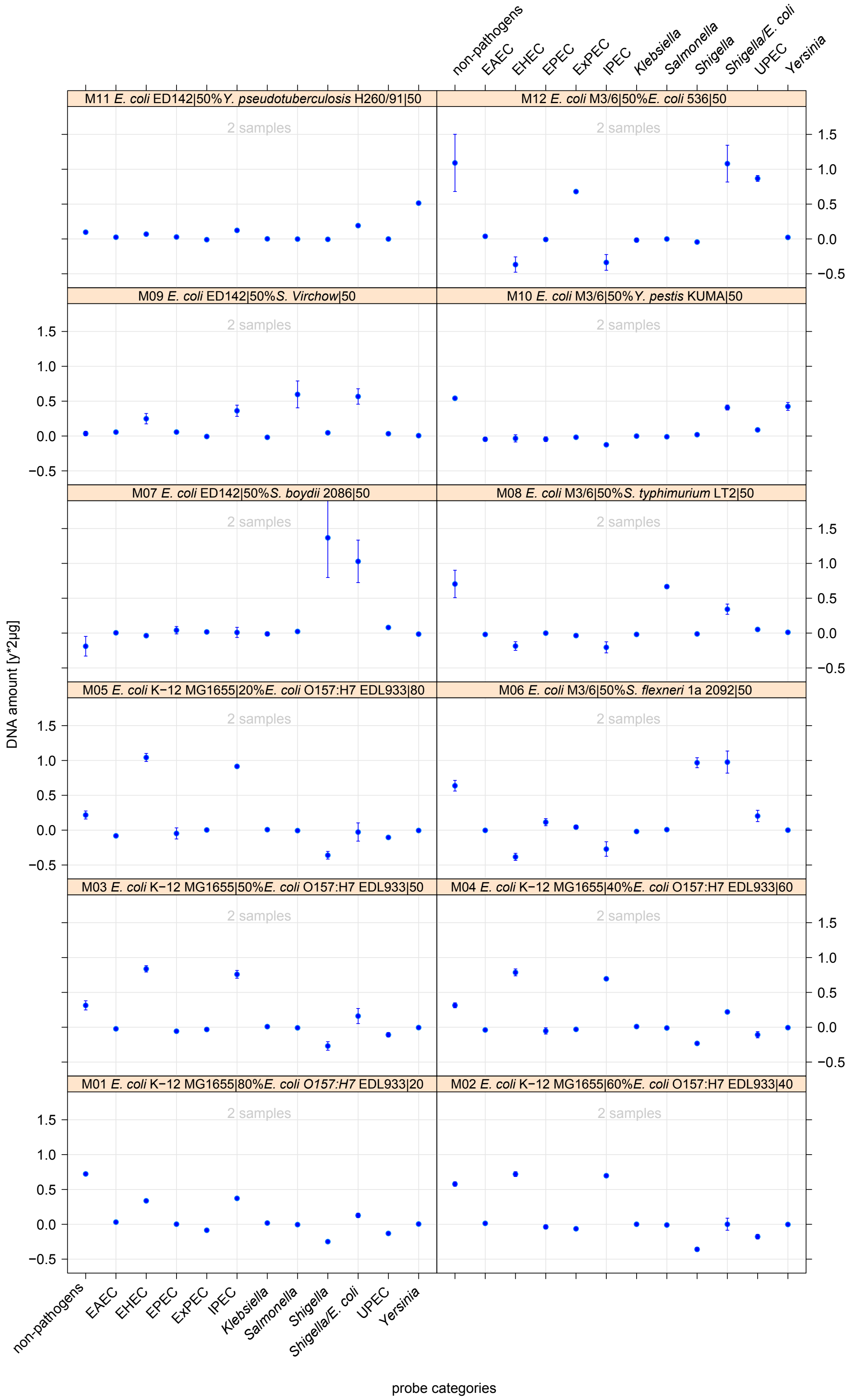


Figure 5.9.: Regression model behavior on the categorical prediction of various mixed hybridisations. The regression model was trained with pure sample and the mixed-culture hybridisation patterns (excluding new pathotypes like ETEC, EIEC and SEPEC).

tances. SHV-type (sulfhydryl variable) β-lactamases were in correspondence with a previous report only detected in *K. pneumoniae* isolates (Paterson *et al.*, 2003).

Microarray results were validated by susceptibility tests with the disc diffusion method. Tests were conducted for those strains and only for antimicrobial agents, for which resistance was detected by the microarray. Figure 5.11 summarises the disc experiments and comprises a susceptibility screening for the antimicrobial agents listed in table 1.5. The values in the plot signify the zone of inhibition around the susceptibility discs, values in brackets refer to zones of moderate inhibition and dashes indicate skipped tests. The class of β-lactamases was covered by 3 test antibiotics in accordance to the frequency of observed resistance and to the complexity of resistance mechanisms and proteins in this class (Giamarellou, 2005).

In far most tests the disc diffusion method confirms the resistances detected by the microarray analysis. Disc diffusion also revealed susceptibilities for single tests. The laboratory *E. coli* strain K-12 MG1655 served as a control in disc experiments. The K-12 genome contains the AMR genes *ampC*, *macAB*, *emrAB* and *acrAB*. AmpC functions as a penicillinase which especially affects ampicillin and other penicillins and therefore mediates resistance to oxacillin and amoxicillin. MacAB, EmrAB and AcrAB form efflux proteins in the extracellular matrix, which are specialised transporters of macrolides and provides erythromycin resistance (Sánchez *et al.*, 1997; Kobayashi *et al.*, 2001). As these protein complexes constitute frequently occurring chromosomally encoded AMR structures, the respective genes were not considered in the described design of an AMR diagnostic. The K-12 strain was susceptible for all other tested antimicrobial agents (with the tetracycline value at the threshold between intermediate status and susceptibility). The disc experiments further revealed widespread susceptibilities to ceftriaxone. Resistance to third-generation cephalosporines mainly arises from the CTX-class (cefotaxime) of β-lactamases, and the hybridisation experiments did not exhibit any positive signals for the corresponding probes. Sporadic ceftriaxone resistances can be traced back to oxacillinases (*bla*OXA) or to PER-type (Pseudomonas extended resistant) extended-spectrum β-lactamases (ESBLs) (Giamarellou, 2005). The capture probes selected by Bruant *et al.* (2006) (representing transport proteins FloR and CmlA) did not well predict the establishment of chloramphenicol resistance. In contrast, chloramphenicol resistance was correctly detected by the microarray, when conferred by type I chloramphenicol acyltransferases.

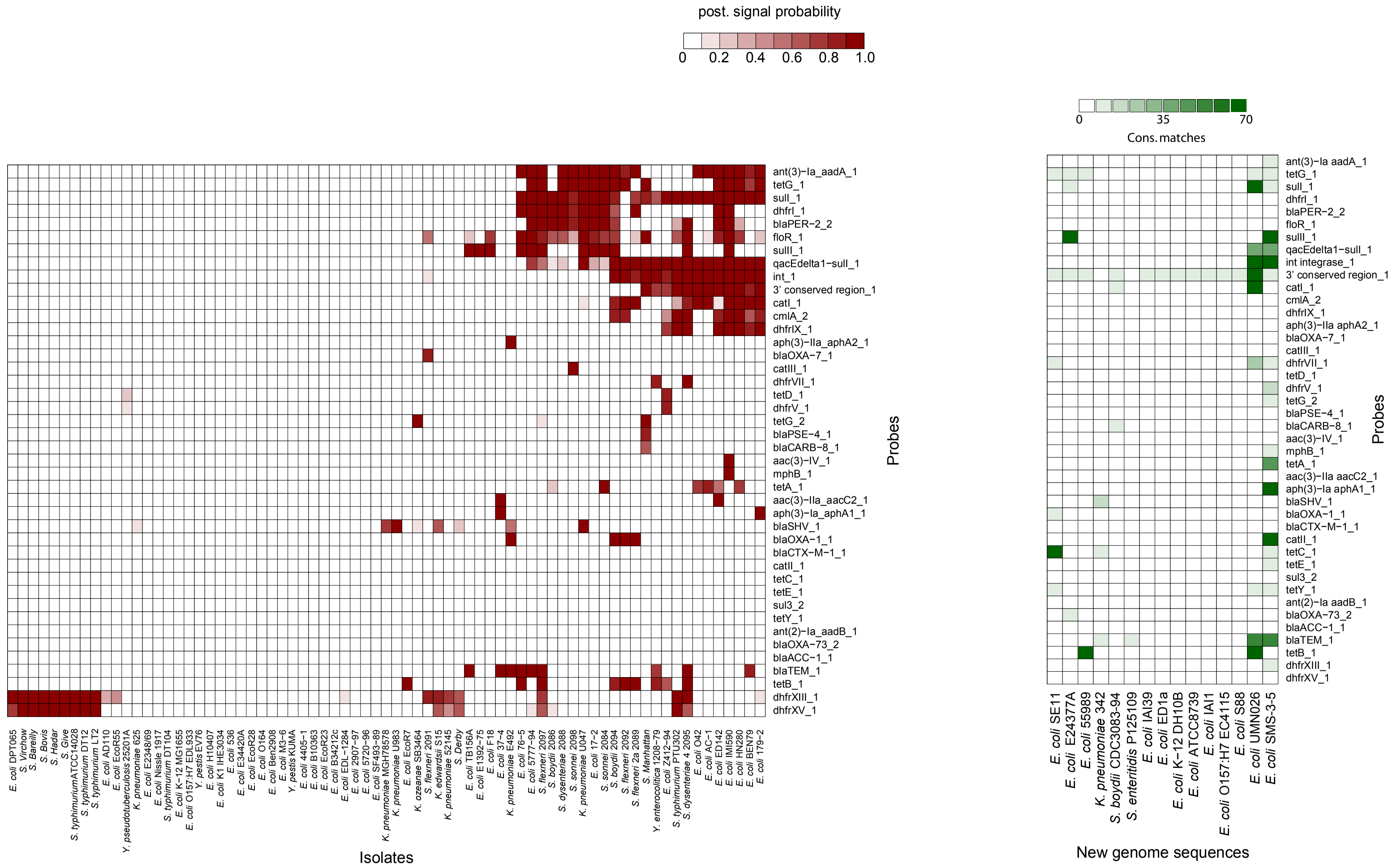


Figure 5.10.: Experimental antimicrobial resistance (AMR) screening of test isolates (left plot) and theoretical assessment of AMR on isolates with recently published genome (right). Colours in the left plot indicate posterior probabilities of signal membership to the signal fraction of the overall distribution of signal intensities. Test hybridisations of pure culture DNA are plotted against AMR probes while the type of AMR can be deduced from the first part of the probe names. Colours on the right side indicate the length of maximum consecutive matches in Smith-Waterman alignments of AMR probes against genome sequences.

The microarray-based detections of aminoglycoside resistances mainly refer to signals in the probe for the streptomycin 3'-adenyltransferase which generally does not confer gentamycin resistance.

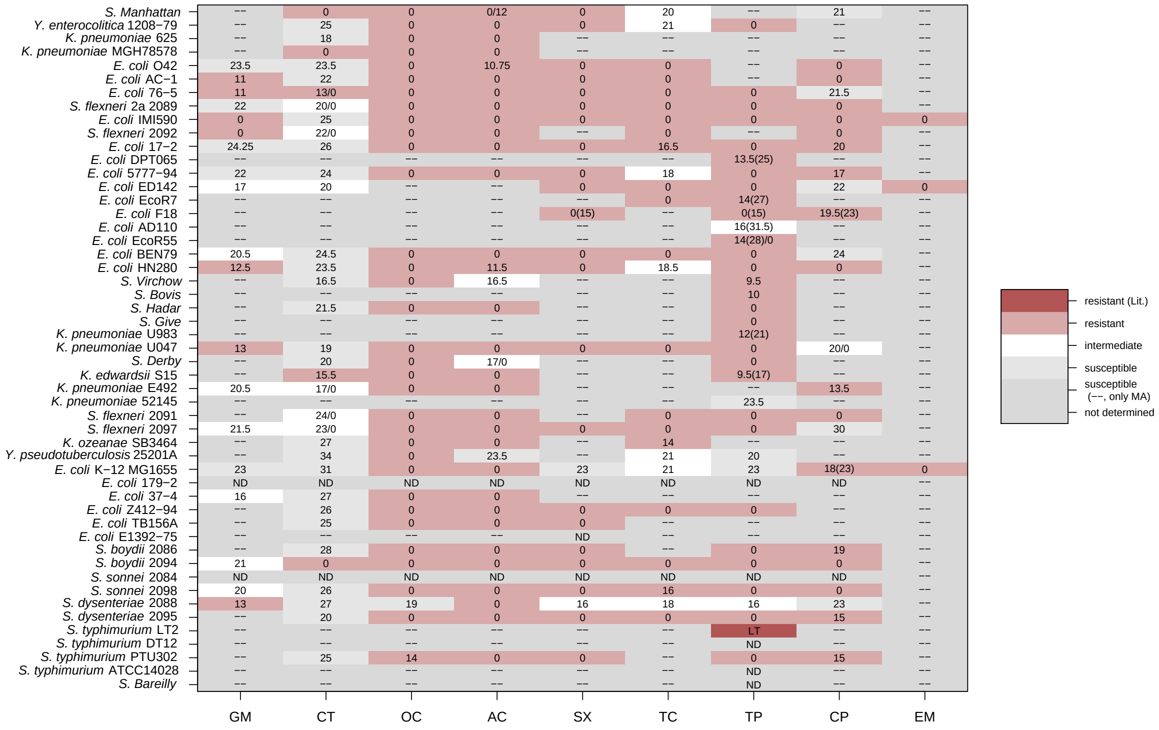
5.3.3. Designed probes and recently published enterobacterial genomes

As a consequence of the increase in sequencing efficiency and the decrease of its costs at the same time, several new enterobacterial genomes were published recently. They contain novel sequence information, a knowledge that impacts strain typing and diagnostics in general. This knowledge especially of strains from new pathotypes could, however, not be integrated in the developed microarray. Nevertheless, the microarray's diagnostic behaviour on these strains was appreciated by Smith-Waterman alignments of all probe sequences against the genome sequences specified in Table 1.2.

Typing of pathogroups The updated data regarding recently published genome sequences of enterobacteria mainly comprised *E. coli* strains belonging to non-pathogens, UPEC, MNEC, EHEC, ETEC, SECEC, EAEC and *Shigella* pathogroups as well as strains *S. Enteritidis* and *K. pneumoniae*. The set of strains covers mostly the full range of pathogroups represented by the microarray and even comprises yet unconsidered *E. coli* pathotypes (SECEC SMS-3-5 and ETEC E24377A). The alignment results were summarised in Figure 5.12 as an image plot of strains against pathogroups. The plot indicates a correspondence of matching category and true pathogroup (green scale), no matching though it was expected (grey colour) or cross-matching (red scale). Colour intensities refer to the length of the respective longest consecutive stretch of matches.

The ability of genus level capture probes to discriminate between *Shigella/E. coli* and *Salmonella* isolates is confirmed by the alignments. The *K. pneumoniae* 342 genome shows sequence similarity to almost all *Klebsiella*-specific capture probes. Moreover, the strains safe detection will be assured by the absence of similarities to probes from other pathogroups. *Salmonella* and non-pathogenic strains as well promise good detectability based on the alignment results although they do not match the full set of capture probes designed for this pathogroup. Representatives of so called new pathogroups (ETEC: E24377A, MNEC: S88) can be correctly classified as *E. coli* isolates and do not reveal substantial cross-hybridisation risk. Additionally, the UPEC strains exhibit detectable patterns of the ExPEC pathogroup. Globally, cross-matches occurred only to the same few probes, a finding which indicates a good performance of the majority of selected probes for the classification of new isolates. The single occurrences of cross-matching would be balanced by the linear regression model with dominating full matches. Therefore, the theoretical assessment verifies the appropriateness of the selected capture probes in pathogroup classification of enterobacteria.

Theoretical AMR detection In addition, the part of the probe set characterising AMR properties was evaluated by determining sequence similarities to strains with recently published genomes. The right image plot in Figure 5.10 provides lengths of the longest consecutive matches encoded in a green colour gradient. The SECEC strain SMS-3-5 was reported to harbour multiple antimicrobial resistances Fricke *et al.* (2008). This finding could be confirmed by our sequence alignments which uncover resistance loci coding for a TEM (Temoneira, name of patient) β-lactamase (*bla*TEM), a chloramphenicol acetyltransferase II (catII), an aminoglycoside 3'-phosphotransferase (aph(3)-Ia aphA1), a tetracycline efflux protein (tetA) and a type II sulfonamide resistant dihydropteroate synthase (sulII).



	GM	CT	OC	AC	SX	TC	TP	CP	EM
S. Manhattan	−−	0	0	0/12	0	20	−−	21	−−
Y. enterocolitica 1208−79	−−	25	0	0	0	21	0	−−	−−
K. pneumoniae 625	−−	18	0	0	−−	−−	−−	−−	−−
K. pneumoniae MGH78578	−−	0	0	0	−−	−−	−−	−−	−−
E. coli O42	23.5	23.5	0	10.75	0	0	−−	0	−−
E. coli AC−1	11	22	0	0	0	0	−−	0	−−
E. coli 76−5	11	13/0	0	0	0	0	0	21.5	−−
S. flexneri 2a 2089	22	20/0	0	0	0	0	0	0	−−
E. coli IMI590	0	25	0	0	0	0	0	0	0
S. flexneri 2092	0	22/0	0	0	−−	0	−−	0	−−
E. coli 17−2	24.25	26	0	0	0	16.5	0	20	−−
E. coli DPT065	−−	−−	−−	−−	−−	−−	13.5(25)	−−	−−
E. coli 5777−94	22	24	0	0	0	18	0	17	−−
E. coli ED142	17	20	−−	−−	0	0	0	22	0
E. coli EcoR7	−−	−−	−−	−−	−−	0	14(27)	−−	−−
E. coli F18	−−	−−	−−	−−	0(15)	−−	0(15)	19.5(23)	−−
E. coli AD110	−−	−−	−−	−−	−−	−−	16(31.5)	−−	−−
E. coli EcoR55	−−	−−	−−	−−	−−	−−	14(28)/0	−−	−−
E. coli BEN79	20.5	24.5	0	0	0	0	0	24	−−
E. coli HN280	12.5	23.5	0	11.5	0	18.5	0	0	−−
S. Virchow	−−	16.5	0	16.5	−−	−−	9.5	−−	−−
S. Bovis	−−	−−	−−	−−	−−	−−	10	−−	−−
S. Hadar	−−	21.5	0	0	−−	−−	0	−−	−−
S. Give	−−	−−	−−	−−	−−	−−	0	−−	−−
K. pneumoniae U983	−−	−−	−−	−−	−−	−−	12(21)	−−	−−
K. pneumoniae U047	13	19	0	0	0	0	0	20/0	−−
S. Derby	−−	20	0	17/0	−−	−−	0	−−	−−
K. edwardsii S15	−−	15.5	0	0	−−	−−	9.5(17)	−−	−−
K. pneumoniae E492	20.5	17/0	0	0	−−	−−	−−	13.5	−−
K. pneumoniae 52145	−−	−−	−−	−−	−−	−−	23.5	−−	−−
S. flexneri 2091	−−	24/0	0	0	−−	0	0	0	−−
S. flexneri 2097	21.5	23/0	0	0	0	0	0	30	−−
K. ozeanae SB3464	−−	27	0	0	−−	14	−−	−−	−−
Y. pseudotuberculosis 25201A	−−	34	0	23.5	−−	21	20	−−	−−
E. coli K−12 MG1655	23	31	0	0	23	21	23	18(23)	0
E. coli 179−2	ND	ND	ND	ND	ND	ND	ND	ND	−−
E. coli 37−4	16	27	0	0	−−	−−	−−	−−	−−
E. coli Z412−94	−−	26	0	0	0	0	0	−−	−−
E. coli TB156A	−−	25	0	0	0	−−	−−	−−	−−
E. coli E1392−75	−−	−−	−−	−−	ND	−−	−−	−−	−−
S. boydii 2086	−−	28	0	0	0	−−	0	19	−−
S. boydii 2094	21	0	0	0	0	0	0	0	−−
S. sonnei 2084	ND	ND	ND	ND	ND	ND	ND	ND	−−
S. sonnei 2098	20	26	0	0	0	16	0	0	−−
S. dysenteriae 2088	13	27	19	0	16	18	16	23	−−
S. dysenteriae 2095	−−	20	0	0	0	0	0	15	−−
S. typhimurium LT2	−−	−−	−−	−−	−−	−−	LT	−−	−−
S. typhimurium DT12	−−	−−	−−	−−	−−	−−	ND	−−	−−
S. typhimurium PTU302	−−	25	14	0	0	−−	0	15	−−
S. typhimurium ATCC14028	−−	−−	−−	−−	−−	−−	ND	−−	−−
S. Bareilly	−−	−−	−−	−−	−−	−−	ND	−−	−−

Figure 5.11.: Validation of antimicrobial resistance with the disc diffusion test. Resistances found by microarray hybridisations were tested in correspondence to the hybridisation results by the exposure of resistant strains to the following antimicrobial substances: Gentamicin (GM, Aminoglycoside), Ceftriaxone (CT, β-lactam), Oxacillin (OC, β-lactam), Amoxicillin (AC, β-lactam), Sulphometoxazole (SX, Sulfonamide), Tetracycline (TC), Trimethoprim (TP, DR inhibitor), Chloramphenicol (CP, Amphenicol) and Erythromycin (EM, Macrolide). Values specify the size of the zone of full inhibition, those in brackets the zone of partial inhibition. Dashes mark cases where no resistance was found in hybridisations and therefore no experimental validation was conducted. The colours are mapped according to the resulting categories from susceptibility to resistance. Most microarray-based resistance predictions could be confirmed by the experiments, though we also obtain susceptibilities in single strains.

The corresponding genes were found in the published genomic sequence. Sequence analysis determined a second multiple resistant strain, the UPEC isolate UMN026. The strain's genome encodes in correspondence to probe alignments for the TEM-type β-lactamase, the aminoglycoside/multi-drug efflux protein AcrD, the dihydropteroate synthase type-1 and several efflux pumps. Single resistances were also obtained by sequence alignments to *E. coli* genomes of strains 55989 and SE11 for tetracycline and of strain E24377A for sulfonamides. Although our AMR probes did only reveal moderate similarity to three different regions in the *K. pneumoniae* isolate 342, the strain was described to be highly resistant. The resistance mechanisms in *K. pneumoniae* 342 rely on β-lactamases and on the existence of many efflux pumps Fouts *et al.* (2008). The β-lactam resistance could be detected with the performed alignments to the designed probes.

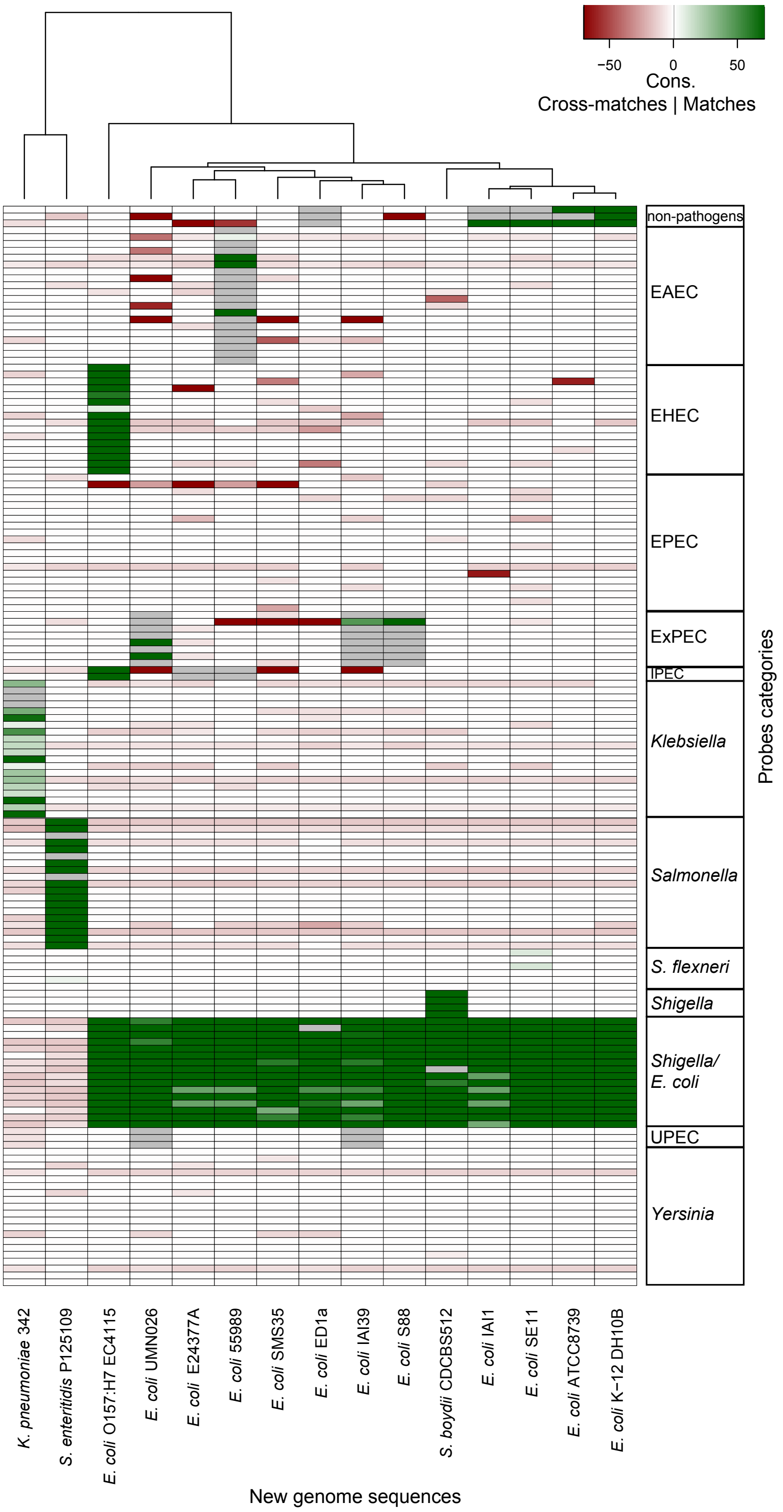


Figure 5.12.: Theoretical assessment of diagnostic probe performance based on alignments of probe sequences to recently published enterobacterial genomes. The imageplot summarises lengths of longest stretches of consecutive matching (green scale) and cross-matches (red scale) of probes to new genomes. Grey colour indicates an expectation of matching without the observation of matches. Fields coloured in light red represent weak similarities that will not lead to cross-hybridisation. The genus level categories show high similarity to corresponding probes, in downstream levels few cross-matching was observed between *E. coli* pathotypes. The cross-matching goes back to only few probes.

5.4. Overview of prediction results

In summary, the predictions of DNA hybridisation on signal intensities of specifically designed markers of enterobacterial pathogroups yields accurate results throughout all levels of hierarchical diagnostic decisions. The prediction outcome is stable regarding different compositions in training sets of the regression model and regarding contrasts between groups from different pathogroup levels. Overall, the regression model exhibits low levels of prediction noise in non-target classes. Accuracies in predictions of the amount of hybridised DNA depend on the number of biological repeats, the distinction power and amount of group-specific probes and the homogeneity of the pathogroup in focus. Spike-in experiments of mixed cultures underline the ability of the diagnostic microarray in conjunction with regression analysis to decode the proportions of bacteria in clinical specimens. The microarray proofed to detect major AMR conferred by degrading enzymes or efflux proteins and the established signal analysis provides information on the reliability of resistance prediction as posterior probabilities.

5.5. Conclusions

Here, I present a novel strategy in the design and analysis of a diagnostic microarray for the distinction of subgroups within the versatile family of *Enterobacteriaceae*. The branch of the γ-proteobacteria comprises frequently studied model organisms in molecular biology, genetics and computational biology. Members of this family are known as versatile pathogens causing gastrointestinal and urinary tract infection, new-born meningitis, plague, diarrhoea or pneumonia, to name but a few (Nataro *et al.*, 1995; Butler, 1994; Brisse *et al.*, 2006; Ogawa *et al.*, 2008). The multiplicity in clinical symptoms implies a large gene pool, genetic exchange and the requirement of complex diagnostic tests. New technologies like DNA microarrays provide suitable high-throughput environments to determine a large number of traits within a single diagnostic test.

The diagnostic strategy applied here is based on an initial categorisation of the target group of bacteria. The subsequent probe selection is geared to the prior categorisation and its quality and discrimination power certainly depends on a proper choice of meaningful sub-entities in the reference set of target genomes. The initial search algorithm of probe selection, longest common factor statistics, explicitly scans the whole genomes with coding and non-coding regions. The consideration of non-coding areas as robust markers with respect to specify a group of bacteria is not straight-forward. Intuitively, non-coding regions are expected to be less conserved and normally do not have a direct impact on infection of a host and survival within the host environment. But, highly conserved intergenic motifs like repetitive sequences termed ERIC (enterobacterial repetitive intergenic consensus) (Wilson and Sharp, 2006) or conserved transcriptional regulatory elements (Pritsker *et al.*, 2004) were described for enterobacteria previously. The selection of intergenic probes distributed on nearly all levels of considered clinically relevant subgroups confirms the existence of characteristic traits outside of coding regions. The high number of capture probes characterising the pathogroups, which refer to poorly or not annotated genes, indicates the existence of unrecognised genotypic traits with an impact in pathogenicity.

Microarray-based diagnostics in comparison The microarray technology is well suited for diagnostic applications due to its highly parallel architecture. In the past few years many workgroups studied the applicability of microarrays to microbial ecology and phylogenetics (Gentry *et al.*, 2006; Wagner *et al.*, 2007), comparative genomics (Dorrell *et al.*, 2005; Willenbrock *et al.*, 2006, 2007) and clinical diagnostics (Loy and Bodrossy, 2006). Microbial diagnostic microarrays (MDM) are generally characterised by a low number of probes, which either target sequence differences in single diagnostic markers or represent a library of virulence-associated genes. MDM from the first category rely on probes designed from sequence differences in single markers like 16S rRNA (Lehner *et al.*, 2005) and *gyrB* (Kakinuma *et al.*, 2003; Kostić *et al.*, 2007). Though these single marker diagnostics perform well in the distinction between distantly related organisms, its distinction performance on subspecies level was found to be limited (Case *et al.*, 2007). Further MDM were based on libraries of determinants for virulence-associated genes (Dobrindt *et al.*, 2003; Bekal *et al.*, 2003; Bruant *et al.*, 2006; Korczak *et al.*, 2005). However, high rates of horizontal gene transfer, which have been reported to occur especially among *E. coli* strains (Wirth *et al.*, 2006), frequently affect virulence-associated genes due to selection pressure in the host. Furthermore, the virulence factors are mainly organised in pathogenicity islands (Blum *et al.*, 1994; Hacker *et al.*, 1997), which can be transferred, deleted and re-inserted (Dobrindt *et al.*, 2004). Other studies reveal overlaps in virulence genotypes of *E. coli* pathotypes (Kariyawasam *et al.*, 2007; Rodriguez-Siek *et al.*, 2005). Finally, the overall genome content of many non-pathogenic *E. coli* isolates resembles that of extraintestinal pathogenic isolates and thus does not allow proper strain typing and risk assessment (Grozdanov *et al.*, 2004; Hejnova *et al.*, 2005; Zdziarski *et al.*, 2008). Thus, the proposed strategy in the development of a MDM rather relies on the determination of the genome-wide most stable subgroup-specific traits among available non-redundant genomic information of the target group of bacteria.

AMR screening An important part in clinical treatment of bacterial infections is the choice of an appropriate drug therapy. Many publications described an increase of antimicrobial resistances in clinical isolates over the last years (Diekema *et al.*, 2004; Lautenbach *et al.*, 2001; Hyle *et al.*, 2005). In this context, the integration of a screening for important determinants of antimicrobial resistances was mandatory in the development of a diagnostic tool. The AMR screening feature does not only provide an assessment of the applicability of antimicrobial agents, but also enables the tracking of AMR progression. Such a screening based on probes for the major classes of AMR mediated by enzymes or efflux proteins was also found in a previous study, which targets the use of microarrays for bacterial diagnostics (Bruant *et al.*, 2006). Therefore, the AMR specific part of the probe set extends previous work by selected new markers of AMR. Hybridisation with a large number of test strains and *in vitro* verification of resistances by the disc diffusion method largely correlate. The choice of the array format does not allow for an establishment of a fully detailed AMR tracking. Such a feature would require many additional probes to target further resistance mediating genes as well as known single nucleotide polymorphisms in AMR target structures. The challenge to establish microarray-based diagnostics of AMR with differences between microarray detection and conventional testing was already stated in previous studies (Frye *et al.*, 2006). As *E. coli* strains possess a high number of drug efflux systems and an even higher number of other membrane transporters (Paulsen *et al.*, 2001), a functional shift mediated by mutations could be the cause for such observed differences.

Nevertheless, microarray based detection of AMR has been described previously as an enhancement to conventional susceptibility testing (Bruant *et al.*, 2006; Frye *et al.*, 2006). Here, it was shown to robustly detect AMR in a wide range of enterobacterial isolates.

Diagnostics of enterobacteria The microarray design strategy was optimised for the detection and classification of enterobacteria. Probe selection was based on a previously approved longest common factor approach and on subsequent filtering of candidate capture probes according to strict match and mismatch limits, which conferred robust signalling with low cross-hybridisation. Extensive test hybridisations were conducted in order to assess the quality of the selected probe set and to obtain training data for the calibration of the linear regression model. Probe-wise performance evaluations based on these tests legitimate the separation of sense and anti-sense capture probes, which exhibited divergence of support quality e. g. in classifications of *Yersinia* test isolates. Detailed investigation concerning the nature of the selected probes reveals single markers, which were previously described because of their group specificity. As an example two capture probes of the EAEC pathogroup indicating strong group-specific support are derived from the *aat* gene locus. The whole *aat* and *aap* loci were previously reported to be specific for EAEC strains and suggested for diagnostic purposes (Nataro 2008, EP 1 917 975 A1; Jenkins *et al.* 2006). The function of nearly half of the capture probes is still uncharacterised and to my knowledge these markers were not applied in enterobacterial diagnostics before. The finding underlines the importance of an unsupervised probe selection mechanism considering both coding and non-coding genomic regions.

Test strains were classified to enterobacterial subgroups by a regression model. The model was able to provide clear separation of the considered subgroups while the prediction accuracy of nature and amount of hybridised DNA increased with the size of the training set and the distance between the groups. Spike-in experiments with mixed culture hybridisations containing isolates from two groups in various proportions were intended to evaluate the power of classification for bacterial communities. The tendencies of predictions based on these mixed culture hybridisations were mainly correct. The regression model is generally able to determine the composition of bacterial communities. The accuracy in determining the proportions in community samples can certainly be enhanced by using a larger set of training data. In extremely unbalanced mixtures, especially if single strains are highly underrepresented, the implementation of an amplification technology can circumvent the existence of detection limits (Park *et al.*, 2006).

In a separate *in silico* analysis we matched the probe set to recently published enterobacterial genomes. The assessment of probe validity on yet unconsidered sequence information confirmed the appropriateness of selected probes. Major AMR patterns reported for these strains could be recognised by the corresponding capture probes of the developed microarray thus recommending it for AMR diagnostics.

Regarding the numerous existing approaches to construct a diagnostic for the versatile group of enterobacteria or its subgroup *E. coli*, our introduced design strategy differs because of its genome-wide probe selection, the broad range of targets and an intuitive but powerful regression model for the analysis of hybridisation patterns. The regression model features training on previous hybridisations and thus is approved by application, which leads to constant simultaneous learning. The probe selection was based on genomic data of published strains that represent clinically relevant phenotypes. With an increase in genomic data the method of probe selection even gains in accuracy of detecting stable traits of the bacterial groups in focus. The chosen microarray platform with 12 separate spotting areas provides a tool for highly parallel diagnostics to reduce analysis time and costs. The trade-off is a limited number of probes. But, the obtained test results proof the suitability of the probe set size for the distinction of the assigned clinical phenotypes. Further efforts should be focused on the reduction of costs for a single hybridisation. A recently developed label-free system might be a step in the right direction as it reduces the preparation and hybridisation time of sample and in parallel increases the sensitivity (Wang *et al.*, 2006).

A manuscript for publication is in preparation

6. Meta-Analysis on diverse gene expression data sets

6.1. Dimension reduction by kernel principal component analysis (kPCA)

The ATH-1 whole genome chip consists of 22810 probe sets, this led to a 41×22810 data matrix (contrasts $\times$ log fold changes of probe sets) after outlier removal. To reduce the dimension of the data matrix, a kernel PCA algorithm was applied which was able to cover virtually the complete information content by defining an orthonormal system of 38 principal component axes. The 22810 log fold changes could therefore be represented by a 41×38 data matrix without any measurable loss of information. Using only the first 25 principal components, 80.585% of the variance could be described. If we state that the remaining 20% of the variance in the data describe noise, an estimation which is certainly not too strict in the context of large-scale gene expression measurements, an effective de-noising can be reached by considering only the first 25 principal components in further steps of the analysis. For a detailed overview of the variance distribution on the first 15 principal components, see Table 6.1.

6.2. Unsupervised analysis reveals three clear clusters of contrasts

The principal component plot (Fig. 6.1) revealed three major clusters of contrasts and several minor ones. In contrast to typical meta-analyses these clusters were not a priori defined, but detected by the proposed unsupervised meta-analysis. Based on this clustering we used an implementation (Karatzoglou *et al.*, 2004) of the spectral clustering algorithm proposed by Ng *et al.* (2001), a variant of the k-means clustering algorithm in a kernel defined feature space, to support the clusters shown in Fig. 2. According to the annotation of the datasets retrieved from GEO, the three clusters were related to indole-3-acetic acid (IAA) addition or inhibition (cluster 1, triangles), pathogen defence activation (cluster 2, solid circles) and “others” (cluster 3, outlined circles). For a detailed biological interpretation, see section “Biological interpretation of clusters”. Additionally, inspection of the pairwise plots of the other principal components contributing to a lower extent to the variance of the data revealed more contrast clusters.

To get further structural insights into the relationships between contrasts and the experimental settings, we performed hierarchical clustering assessed by multi-scale bootstrapping (Fig. 6.2). In agreement with the spectral clustering performed earlier and the graphical inspection of the pairwise scatter plots of contrasts on the kPCA axes, the three main clusters of contrasts could also be found

Table 6.1.: Variance of kernel principal components.

	PC1	PC2	PC3	PC4	PC5
PV	0.10035	0.05383	0.05003	0.04640	0.03887
CP	0.10035	0.15418	0.20422	0.25062	0.28949

	PC6	PC7	PC8	PC9	PC10
PV	0.03725	0.03250	0.03226	0.03142	0.02973
CP	0.32674	0.35925	0.39151	0.42293	0.45267

	PC11	PC12	PC13	PC14	PC15
PV	0.02793	0.02699	0.02647	0.02606	0.02470
CP	0.48061	0.50761	0.53409	0.56016	0.58486

Variance of the first 15 principal components on the 41×22810 data matrix of *Arabidopsis thaliana* microarray data, explaining close to 60% of the variance of the data. Abbreviations: PV = Proportion of Variance, CP = Cumulative Proportion of variance.

as the first two splits in the resulting dendrogram with high bootstrap support.

As the three clusters were mainly separable through the x-axis on the kPCA scatter plot using the first two axes (Fig. 6.1), we postulated that the first principal component alone might be enough to select genes whose co-regulation patterns could clearly distinguish between IAA related, pathogen-defence related and other contrasts.

6.3. Gene selection with kPCA loadings

To accomplish an efficient feature subset selection, i.e. to identify genes that are responsible for the clustering, a variety of methods have been described, e.g. Self-Organizing Maps (SOMs) (Tamayo *et al.*, 1999), Maximal Margin Linear Programming (MAMA) (Antonov *et al.*, 2004), Correlation Based Feature Selection (CFS) (Hall, 1999) or Recursive Feature Elimination (RFE) using Support Vector Machines (SVM) (Guyon *et al.*, 2002; Zhang *et al.*, 2006). In consequent continuation of our approach of exploratory meta-analysis, we looked for genes that have a strong association with the first kPCA axis, i.e. we calculated the loadings of each of the genes onto the principal components. To achieve this with respect to the kernel defined feature space we projected single artificial contrasts containing only one deregulated gene onto the new coordinate system. Each of the 22810 artificial contrasts was set up in a way that it showed a high absolute fold change value in one of the genes and all others being set to zero. From the resulting 22810×38 matrix of loadings of each of the genes onto the 38 principal components, we selected the 500 top genes for both positive (IAA related) and negative (pathogen related) extrema. To assess the accuracy of the gene selection process exploratively, we repeated the previous kernel PCA analysis using only the selected genes, i.e. on the remaining 41×500 data matrices, and inspected pairwise scatter plots of the first 20 principal components for each dataset of either IAA-related or pathogen-associated genes. All kPCA plots of the IAA-related gene set, even the one of the first two axes which contribute most to the overall variance of the data, showed a wide spread of IAA contrasts along the principal component axes.

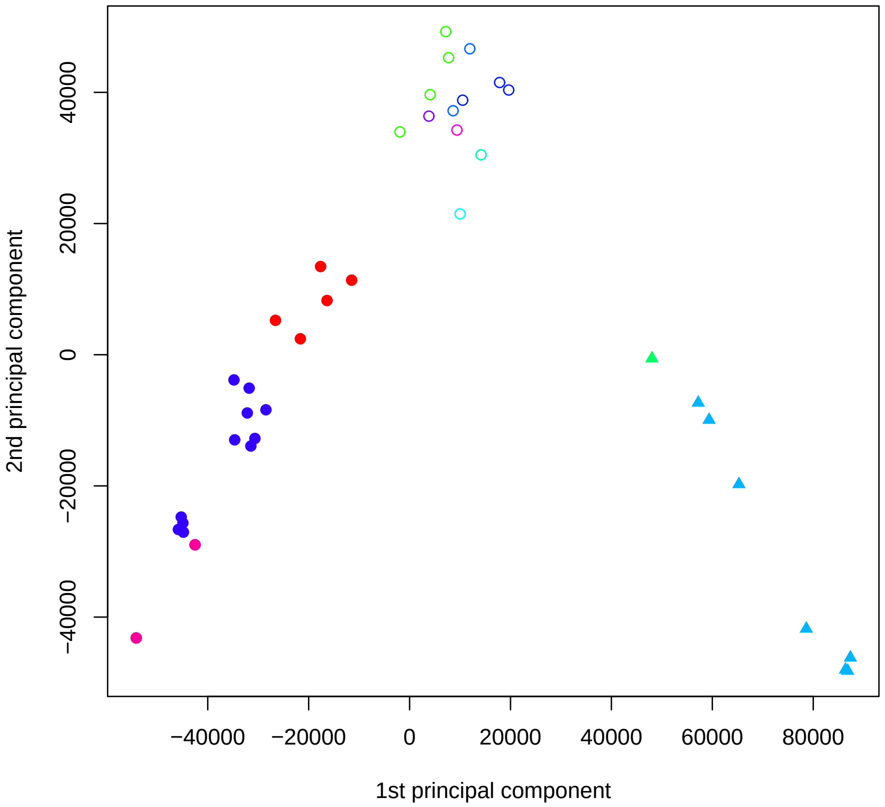


Figure 6.1.: Kernel PCA on 41 *Arabidopsis thaliana* contrasts. Plot of all 41 contrasts using the first two principal component axes. Comparisons are colored according to the experiment they originated from and correspond to the colors used in Figure 3, different shapes indicate the three different clusters obtained from spectral clustering: Indole-3-acetic acid (IAA) related contrasts (solid circle), pathogen related contrasts (triangles) and others (outlined circle).

This indicated a high variance of the selected genes in IAA-related contrasts. All other contrasts were projected onto a compact local cluster by kPCA, demonstrating that the selected genes do not vary in these contrasts. The same was found in the kPCA plots of the matrix with pathogen-associated genes (data not shown). These findings indicate that expression patterns related neither to IAA nor pathogen treatment were efficiently stripped off by the gene selection process.

6.4. Biological interpretation of clusters

The hierarchical clustering on all kPCA scores in Figure 6.2 revealed three main clusters of contrasts: contrasts studying pathogen defence (blue), contrasts analyzing indole-3-acetic acid (IAA) effects (violet) and other contrasts studying various effects (grey). These three clusters were well-supported by high bootstrap values. The labels at the edges include the GEO accession number followed by an index indicating the contrast number. For a detailed description of contrasts see Table 6.2. For each contrast, two groups of samples were compared and for each group, the genetic background and treatment is listed. The last column of Table 6.2 indicates the cluster this contrast was assigned to in kernel PCA clustering.

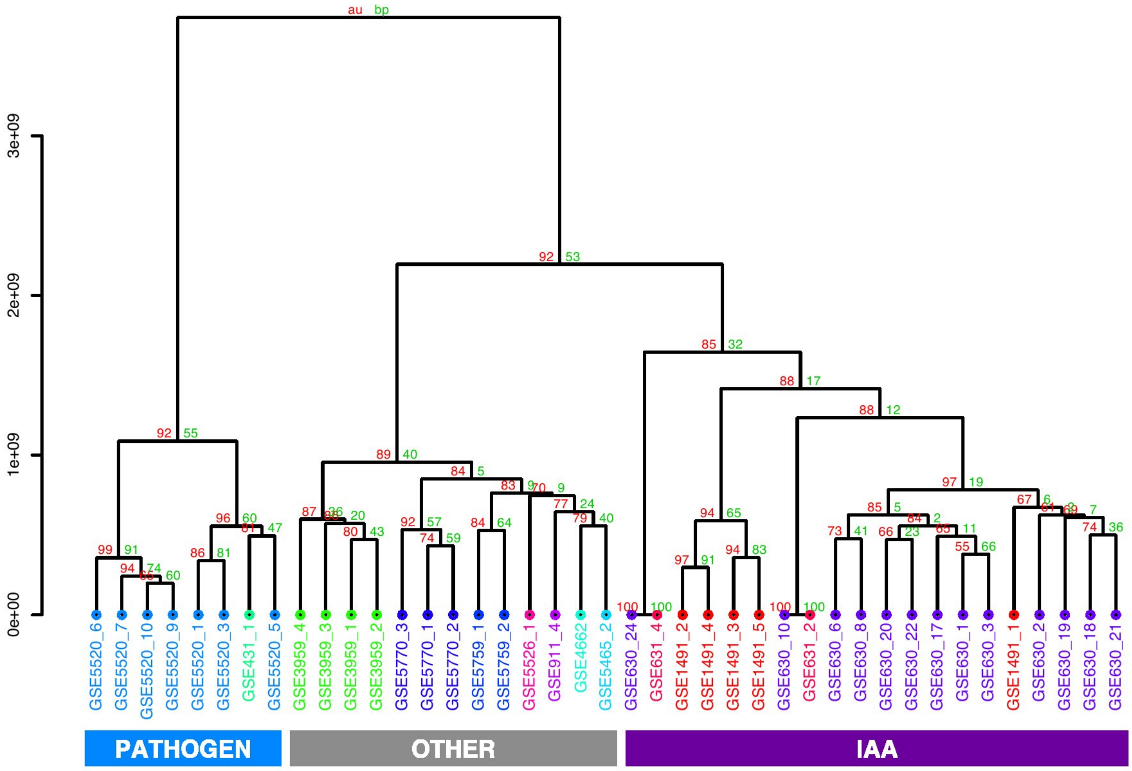


Figure 6.2.: Hierarchical clustering on 41 *Arabidopsis* contrasts. Cluster dendrogram using hierarchical ward clustering on all 38 principal component vectors resulting from kernel PCA. Contrasts are colored according to their experimental affiliation. Approximately unbiased (au, (Suzuki and Shimodaira, 2006)) and standard bootstrap (bp) values are given for all splits and support the results from the previous spectral clustering (Fig. 2).

Zooming into the IAA cluster, a cluster containing only contrasts with IAA inhibition (GSE1491_2, GSE1491_3, GSE1491_4 and GSE1491_5) was well-separated from the remaining contrasts, including GSE1491_1, a contrast from the same dataset, but where IAA instead of an IAA inhibitor was added to one sample group. The remaining contrasts in the IAA cluster mainly studied the effect of IAA on different mutants with defects in IAA biosynthesis or signalling. Indole-3-acetic acid (IAA) belongs to a group of plant growth hormones called auxins. The “others”- cluster consisted of contrasts studying various effects like the effect of lincomycin which is an inhibitor of plastid protein translation, regulation changes of an embryogenesis transcription factor mutant or of stress tolerant mutants. Naturally, in this cluster of divergent contrasts, contrasts from the same dataset clustered closely together. The architecture of the hierarchical cluster tree shows that data preprocessing followed by kernel PCA adjusted the data in such a way that contrasts stemming from biologically similar experiments are indeed more similar to each other than to other contrasts. Thus, with our analysis, we were able to achieve comparability of microarray datasets from different laboratories addressing different biological questions. This is nontrivial and important considering the numerous sources of variation that affect the nature of the datasets underlying this analysis.

Table 6.2.: Overview of all contrasts included in the explorative meta-analysis.

	Sample Group 1		Sample Group 2		
Contrast	Genetic background	Treatment	Genetic background	Treatment	Cluster
GSE1491_1	WT Col-0	IAA	WT Col-0	non	IAA
GSE1491_2	WT Col-0	IAA inhibitor A	WT Col-0	non	IAA
GSE1491_3	WT Col-0	IAA inhibitor B	WT Col-0	non	IAA
GSE1491_4	WT Col-0	IAA/IAA inhibitor A	WT Col-0	non	IAA
GSE1491_5	WT Col-0	IAA/IAA inhibitor B	WT Col-0	non	IAA
GSE3959_1	MU LEC2GR	1h LEC2 induction	MU LEC2GR	no LEC2 induction	other
GSE3959_2	MU LEC2GR	4h LEC2 induction	MU LEC2GR	no LEC2 induction	other
GSE3959_3	MU LEC2GR	1h LEC2 induction	WT WS-0	4h LEC2 induction	other
GSE3959_4	MU LEC2GR	4h LEC2 induction	WT WS-0	NA	other
GSE431_1	pmr4-1 MU	non	pmr4-1 MU	powdery mildew	pathogen
GSE4662_1	MU STA1	non	WT	NA	other
GSE5465_2	MU OETOP6B	non	WT	NA	other
GSE5520_1	WT Col-0	DC1318 Cor 10e6	MU STA1	non	pathogen
GSE5520_10	WT Col-0	EcTUV86-2 fliC 10e8	WT Col-0	non	pathogen
GSE5520_3	WT Col-0	DC3000 10e6	WT Col-0	non	pathogen
GSE5520_5	WT Col-0	DC1318 Cor 5x10e7	WT Col-0	non	pathogen
GSE5520_6	WT Col-0	DC3000 hrpA-fliC 10e8	WT Col-0	non	pathogen
GSE5520_7	WT Col-0	DC3000 hrpA 10e8	WT Col-0	non	pathogen
GSE5520_9	WT Col-0	EcO157H7 10e8	WT Col-0	non	pathogen
GSE5526_1	WT?	non	WT?	non	other
GSE5759_1	WT Col-0	dark plus lincomycin	WT Col-0	dark	other
GSE5759_2	WT Col-0	red light plus lincomycin	WT Col-0	red light	other
GSE5770_1	WT Col-0	lincomycin	WT Col-0	non	other
GSE5770_2	abi4-102 MU	lincomycin	abi4-102 MU	non	other
GSE5770_3	gun1-1 MU	lincomycin	gun1-1 MU	non	other
GSE630_1	WT Col-0	IAA (2h 5μM)	WT Col-0	EtOH (2h)	IAA
GSE630_10	MU arf2-6	IAA (2h 5μM)	MU arf2-6	EtOH (2h)	IAA
GSE630_17	MU IAA17-6	EtOH (2h)	WT Col-0 I	EtOH (2h)	IAA
GSE630_18	MU arx3-1	EtOH (2h)	WT Col-0 I	EtOH (2h)	IAA
GSE630_19	MU i5i6i19	EtOH (2h)	WT Col-0 I	EtOH (2h)	IAA
GSE630_2	MU nph4-1	IAA (2h 5μM)	MU nph4-1	EtOH (2h)	IAA
GSE630_20	MU IAA17-6	IAA (2h 5μM)	WT Col-0 I	IAA (2h 5μM)	IAA
GSE630_21	MU arx3-1	IAA (2h 5μM)	WT Col-0 I	IAA (2h 5μM)	IAA
GSE630_22	MU i5i6i19	IAA (2h 5μM)	WT Col-0 I	IAA (2h 5μM)	IAA
GSE630_24	MU arf2-6	IAA (2h 5μM)	WT Col-0 A2	IAA (2h 5μM)	IAA
GSE630_3	MU arf19-1	IAA (2h 5μM)	MU arf19-1	EtOH (2h)	IAA
GSE630_6	MU IAA17-6	IAA (2h 5μM)	MU IAA17-6	EtOH (2h)	IAA
GSE630_8	MU i5i6i19	IAA (2h 5μM)	MU i5i6i19	EtOH (2h)	IAA
GSE631_2	MU arf2-6	IAA (2h 5μM)	MU arf2-6	non	IAA
GSE631_4	MU arf2-6	IAA (2h 5μM)	WT Col-0	IAA (2h 5μM)	IAA
GSE911_4	35S::LFY	non	WT ler	35S::LFY	other

Each contrast consists of two groups which are described by their genetic background (genotype) and treatment. The last column “Cluster” derives from the clustering of the kernel PCA scores. Contrasts are labelled with the GEO series number followed by contrast index.

6.4.1. *Arabidopsis thaliana* genes regulated by indole-3-acetic acid (IAA)

To get an overview of the functions of the selected genes representative for the contrast clusters “IAA” or “pathogen”, the *Arabidopsis thaliana* pathway analysis program MapMan (Usadel *et al.*, 2005) was used. With MapMan, gene expression values can be displayed onto diagrams of functional categories and metabolic and regulatory pathways. In this study, MapMan was used to visualise the representative genes for the two clusters “IAA” and “pathogen”.

Among the genes representative for IAA contrasts, the functional category “hormones” with the subgroup “IAA” defined by MapMan showed the highest proportion of regulated genes (diagram not shown). The subgroup “IAA” consists of 215 genes in MapMan. We selected 500 genes representative for IAA with our approach and out of these, 43 genes are cataloged in the MapMan subgroup “IAA”.

Thus, by selecting 500 genes from the ATH1 microarray which comprises roughly 2% of the array, we were able to capture 20% of the genes annotated as IAA-related in MapMan.

In the "hormones" subgroup "ethylene", and in the category "transcription factor" many genes are regulated under IAA treatment, while a smaller number of genes is regulated in the categories "Cytochrome P450" and "cell wall" (data not shown).

Regulated genes in the subgroup "ethylene" are either involved in ethylene synthesis or signal transduction. Ethylene plays a role in the regulation of a number of developmental processes, often in interaction with other plant hormone signals. For example, auxins can induce ethylene formation and in turn ethylene can trigger an auxin increase. Some processes such as root elongation, differential growth in the hypocotyl and root hair formation and elongation are regulated by both auxin and ethylene in *Arabidopsis thaliana* (Stepanova *et al.*, 2005). All the GEO datasets we annotated as IAA-related originate from seedling RNA extracts. Since IAA belongs to the group of auxins, the aforementioned processes are likely to be regulated under IAA treatment.

Cytochrome P450 monooxygenases are involved in various biosynthetic reactions which synthesise for example plant hormones or defence compounds. Regulation of cell wall genes is also expected as auxins mediate cell elongation by stretching of the cell wall which requires restructuring processes.

In conclusion, the gene selection of our unsupervised meta-analysis approach chose many genes which are annotated and independently validated as being IAA regulated.

6.4.2. *Arabidopsis thaliana* genes regulated by pathogen exposure

Gene selection for contrasts studying plant response to pathogens revealed a high number of regulated genes in the following functional categories of MapMan (Usadel *et al.*, 2005): "biotic stress", "receptor kinases", "photosynthesis" (light reactions), "alkaloid-like proteins" from "secondary metabolism", "nitrilases", "cell wall" genes and "WRKY transcription factors". For all of the functional categories mentioned above, it has been reported that genes in these categories are regulated after pathogen attack and play a role in plant defence. Figures 6.3 and 6.4 show details of the MapMan maps which harbour these categories. In the figures, grey areas inside the diagrams represent all the individual genes present on the ATH1 chip and annotated in MapMan. The selected genes representative for contrasts studying the effects of pathogen exposure are highlighted by small dark blue squares. For example, Fig. 6.3 C shows that there are 41 DUF26 receptor kinases present on the ATH1 chip, of which 9 are regulated after pathogen exposure. In the following, we give a short description of the functions of the genes regulated after pathogen exposure.

A change in carbohydrate metabolism after pathogen attack as observed here (Fig. 6.3 A, upper right: "light reactions") has also been reported by Berger *et al.* (2004) for the pathogens *Pseudomonas syringae* or *Botrytis cinerea*. The authors have shown a co-regulation of defence, sink and photosynthetic gene expression in response to the pathogens under study.

As the cell wall is a natural barrier for plant pathogens, plant defence includes cell wall modifications and biosynthesis to thicken cell walls and impede further pathogen attack (Cheong *et al.*, 2002). Figure 6.3 A shows that several genes of the cell wall metabolism are regulated after pathogen exposure.

The regulation of WRKY transcription factors (Fig. 6.3 B, upper left) is also described in the publication accompanying the GEO dataset GSE5520 (Thilmony *et al.*, 2006). Our findings confirm their suggestion that these transcription factors regulate plant response to bacteria.

Alkaloids (Fig. 6.3 A, lower left) are secondary metabolites listed in the “N-misc.” category of MapMan. They are generally not essential for the basic metabolic processes of the plant but play an important role in plant defence (Dixon, 2001). They are produced by the plant to restrict pathogen feeding. The accumulation of antimicrobial substances is often regulated by signal-transduction pathways which require the perception of the pathogen by a plant receptor encoded by host resistance genes (Dangl and Jones, 2001; Piroux *et al.*, 2007). Thus, the regulation of DUF26 containing genes postulated by our analysis of the *Arabidopsis thaliana* transcriptome (Fig. 6.3 C) might reflect their function in pathogen recognition. Receptor kinases are discussed in more detail in the next section.

The functional category “biotic stress” (Fig. 6.4 A) comprises a number of different genes which are annotated to be pathogen related.

Nitrilases (Fig. 6.4 B, upper right) are involved in IAA biosynthesis and catalyze the conversion of indole-3-acetonitrile to IAA. The induction of four *Arabidopsis thaliana* nitrilases by the pathogen *Pseudomonas syringae* has been shown by Bartel and Fink (1994).

Thus, gene selection by unsupervised meta-analysis was able to pinpoint biologically important genes of which many are experimentally validated to be regulated by pathogen attack. Clearly, one could postulate that the remaining genes of unknown function are also associated with responses to pathogen attack.

6.4.3. Serine-threonine kinases involved in plant response to pathogens

As presented in Figure 6.3 C, the extracted set of genes deregulated in response to pathogens includes a number of receptor kinases. Many kinases belong to the group of serine/threonine kinases of the DUF26 subfamily. They all share the same domain composition and order consisting of a signal peptide, an extracellular region containing two domains of unknown function (DUF26, PF01657) and a cytosolic serine/threonine kinase domain (pkinase, PF00069). According to the SMART database (Letunic *et al.*, 2006), proteins of this family are exclusively found in Streptophyta. The 9 putative receptor kinases exhibit high similarity in domain composition and nucleotide sequence with the receptor-like kinase 4 of *Arabidopsis thaliana* (Swiss-Prot-ID Q9C5T0). This enzyme is reported to be a member of the systemic acquired resistance pathway in higher plants. Its expression can be activated by a regulatory protein induced via pathogen and salicylic acid interaction (Du and Chen, 2000). Salicylic acid is a signalling molecule which induces systemic acquired resistance in the host plant (Ryals *et al.*, 1996). These findings suggest a function for the putative receptor-like kinases in host defence processes.

Two of the DUF26 kinase genes (At4g21400, At4g21410) were also regulated in the contrasts from dataset GSE3959 and in one contrast from the dataset GSE5770. In the former dataset, the function of B3 domain protein LEAFY COTYLEDON2 (LEC2) was studied. This transcription factor is required for several aspects of embryogenesis including the maturation phase. In the latter contrast, *abi4* mutant plants were treated with lincomycin and compared to untreated mutants. ABI4 is a transcription factor, lincomycin inhibits plastid protein translation. From this finding it may be concluded that these two DUF26 kinase genes either play a role in more than one signalling pathway or that the same pathway is used to regulate several functions. This might be an interesting starting point to study these pathways in more detail.

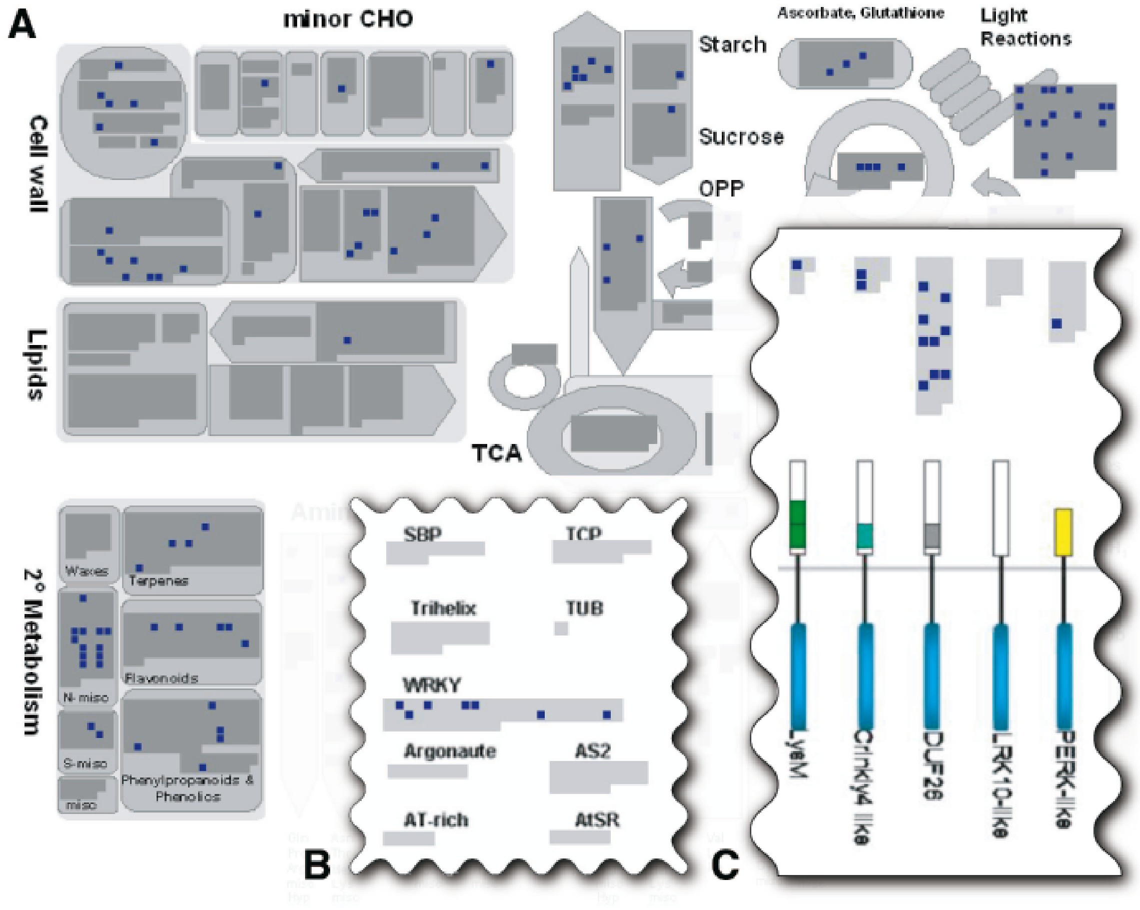


Figure 6.3.: Overview of genes regulated in pathogen associated contrasts. The grey areas inside the individual diagrams of the functional categories represent all genes present on the ATH1 chip. Dark blue squares highlight genes regulated in contrasts of the "pathogen" cluster. Regulation of cell wall genes (upper left), alkaloids which fall into the category "N-misc." of "secondary metabolism" and "Light Reactions" of photosynthesis (upper right) is apparent. B) Part of the "transcription" map indicating regulation of WRKY transcription factors. C) Section of the "receptor like kinases" map indicating regulation of DUF26 kinases. Figure reading example: In subfigure C, a total of 41 DUF26 kinases are represented on the ATH1 chip of which 9 are regulated after pathogen exposure. The figure is based on maps from the pathway analysis program MapMan (Usadel *et al.*, 2005).

As can be seen from Figure 6.5, the DUF26 kinase genes were not regulated in all of the contrasts involving pathogen exposure. This could be due to several reasons. For example either the variance in the single microarray intensities was so high that differential expression could not be detected in the contrast or the difference in expression levels (i.e. the logarithmic fold change) was too low to be significant because of biological reasons. Again, this finding might be an interesting starting point to analyze the function and regulation of the DUF26 kinase genes.

6.5. Conclusions

Public microarray data repositories accumulate large amounts of data which have so far rarely been used for large-scale analyses. Using this wealth of information, additional implications for the function and regulation of genes can be made which could not be derived from single microarray datasets. This stresses the importance of meta-analyses and their benefit over classical microarray experiments.

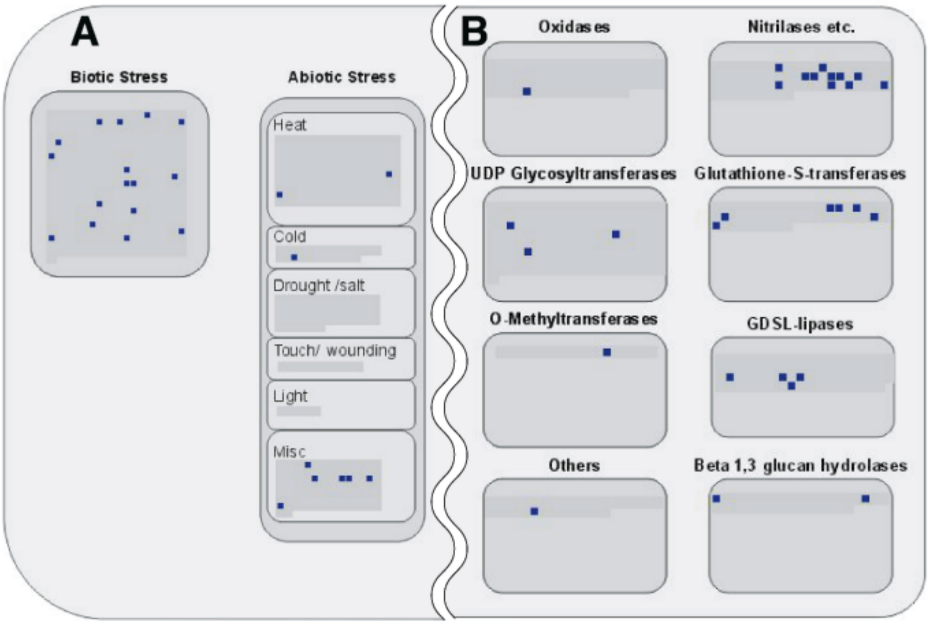


Figure 6.4.: Overview of (A) stress genes and (B) genes of large enzyme families regulated in pathogen-associated contrasts. The grey areas inside the individual diagrams of the functional categories represent all genes present on the ATH1 chip. Dark blue squares indicate regulated genes. Subcategories "Biotic Stress" (A) and "Nitrilases etc." (B) contain a high number of genes regulated after pathogen exposure. The figure is based on maps from the pathway analysis program MapMan (Usadel *et al.*, 2005).

In this study, we apply a novel approach of an unsupervised meta-analysis on a large number of gene expression microarrays. Before conducting the analysis, we performed a pre-processing which included a conservative outlier removal. Kernel PCA, followed by hierarchical clustering, revealed robust and significant clusters of contrasts which reflect similar experimental conditions. Thus we were able to detect biologically important known and unknown factors (e.g. IAA- or pathogen-associated) through an unsupervised analysis.

To find genes specifically regulated in these clusters, a novel approach of gene selection was conceived. Gene selection was performed using loadings of features on kernel PCA scores, which has to our knowledge not been performed in the context of meta-analysis before. Gene selection based on loadings of features on kernel PCA scores circumvents a major drawback of most proposed methods of feature selection: They tend to find linear combinations of features, i.e. genes, that separate the given experimental classes best (e.g. different cancer types, etc.). This is challenging as the search space for all possible linear combinations is too large to be searched exhaustively and sophisticated heuristics and optimization methods have to be chosen which likely yield differing results, see e.g. Zhang *et al.* (2006). An unsupervised analysis as proposed here circumvents this problem efficiently by working directly on the loadings from the kPCA analysis. Eigen-decomposition of the kernel matrix is deterministic and so are the results from our gene selection process, provided the projection is capable of clustering the contrasts appropriately. The genes selected by our feature extraction were found to be representative of a group of contrasts and could in part be experimentally validated. Furthermore, adding random noise to the data did not change the set of selected genes, proving the robustness of the proposed gene selection method.

It is the gene-selection in the first place that benefits most from an analysis across several datasets. Weak regulation signals can easily be overlooked in a single dataset, i.e. the genes will likely receive

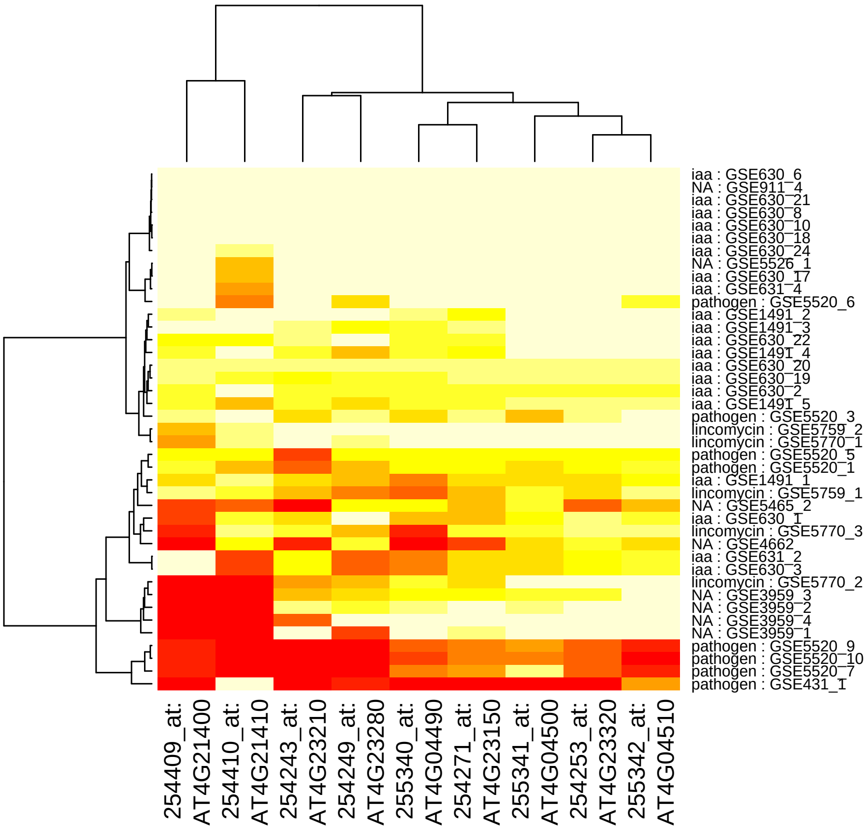


Figure 6.5.: Regulation of DUF26 kinase genes. Red cells indicate low p-values for a gene in a particular contrast, light yellow cells represent high p-values. The DUF26 kinase genes are strongly regulated in four pathogen-associated contrasts.

an insignificant p-value due to their low fold changes compared to a relatively high variance. The situation becomes even worse after a correction for multiple testing has raised the overall p-value level, efficiently removing those subtle signals. In a meta-analysis approach which integrates many datasets, even a small signal that is consistent across several contrasts can be detected. To ensure this surplus and to prevent early losses of information, we used fold changes and not p-values for our analysis. We performed the unsupervised meta-analysis on absolute fold changes to reduce variation introduced by different experimental settings. For example, when there are contrasts in the dataset which compare a surplus of a factor with a control and other contrasts comparing a lack of a factor with another control, we might expect fold changes with opposite signs but still want the contrasts to cluster closely together because the same factor was studied in both. In some cases the direction of the experimental setup was not even apparent from the description of the dataset.

To ensure that results of similar quality could not be obtained by a simpler model and thus to prevent overfitting of the data we compared the results to the ones obtained from traditional linear PCA. Even though linear PCA was also able to detect some of the major clusters in principle, its accuracy as assessed by hierarchical clustering as well as by the gene selection process fell far short of the results from the kernelised version. Additionally, it should be noted that kernel PCA outperforms the traditional approach significantly, considering that the dimension of the kernel matrix as a matrix of pairwise scalar products between the data points is independent of the dimension of the data, which is 22810 (the number of probe sets) in the case of the ATH-1 arrays.

For a large *Arabidopsis thaliana* microarray dataset, we demonstrate here that gene selection, based on the study of principal components, proposed genes typical for either IAA- or pathogen-associated contrasts. These genes were proved to be related to either IAA effects or plant reactions in response to pathogen exposure by previous studies. Furthermore, starting from our finding that DUF26 kinases are regulated in pathogen-associated contrasts, we applied homology modeling to propose that DUF26 kinases have a function in plant pathogen defence. Further experiments are needed to confirm this hypothesis. Nonetheless, this example demonstrates how unsupervised analysis can aid and guide the next steps of such an analysis.

In general, unsupervised meta-analysis embracing several highly divergent experimental settings can suggest novel gene functions by revealing the regulation of a gene under different conditions. It is noteworthy that these analyses are not restricted to datasets addressing the same topic, but that they profit from the divergence of the experimental settings.

However, it has to be mentioned that an unsupervised meta-analysis is suggestive rather than definitive. But since it is common in classical statistics to precede a supervised, parametric analysis with an explorative approach to check the integrity and quality of the data, we recommend the same here for microarray meta-analyses. Hypotheses from unsupervised analyses can then be tested with supervised methods and biological experiments.

We have shown here that it is feasible to integrate various datasets spanning a large range of experimental questions and originating from various laboratories into a coherent unsupervised analysis. This analysis can be applied to find genes representative of a cluster of related contrasts. Based on expression changes between clusters, the function and regulation of genes can be predicted. Our study is based on the Affymetrix ATH1 Genome Array platform here, but our approach can be transferred to any platform, organisms and experimental design which allows one to compute a logarithmic fold change, e.g. human or mouse microarray datasets. To achieve easy access to our unsupervised meta-analysis results, we intend to set up a database web server where new datasets can easily be added and compared to our curated database of *Arabidopsis thaliana* ATH-1 microarrays.

This project is published in Bioinformatics and Biology Insights (Engelmann et al.*, 2008).*

7. Optimal adjustment of HMM toplogies

7.1. Adaptation of HMM states to length distributions

HMMs consist of states, which emit a sequence of observables with a certain length. A two-state HMM with a symbol alphabet $\mathcal{O}_{S1} = 1$ of state S_1 and $\mathcal{O}_{S2} = 2$ of state S_2 emits a sequence of symbols $O_{hmm} = 111122221112211111$. The partial sequences of ones and twos arise from single emission events and refer to the holding time of the two states. The holding time is related to a self-transition of the state occurring with a certain probability. Concurrently, the probability of self-transition determines the length distribution of observation sequences associated to a self-transitive state in conventional HMM topologies. Single self-transitive states are characterised by the emission of geometrically distributed observation sequences.

Previous studies revealed in contrast that the length distributions of genetic elements in *E. coli* genes like the spacer after the ribosome binding site or the 3'-UTR spacer (Yada *et al.*, 1999) as well as exons (Melodelima *et al.*, 2007) and M-isochores (Melodelima *et al.*, 2006) in eukaryotic DNA are bell-shaped. Thus, we propose a statistical framework in order to optimise HMM topologies with respect to an appropriate representation of sequence length. An adequate modelling of length properties of emitted sequences is achieved by serially chain-linking self-transitive states (Durbin *et al.*, 1998), which results in sequence length of observables distributed according to the negative binomial law.

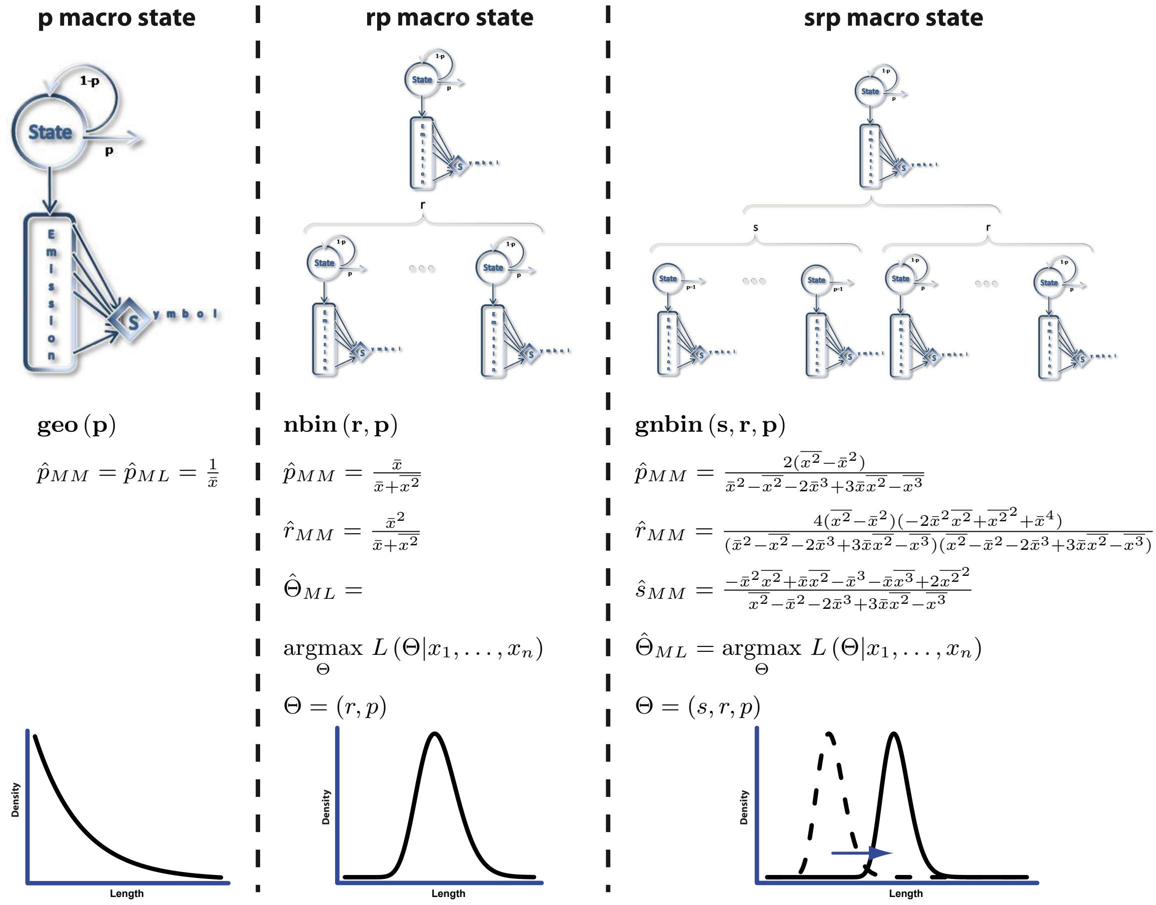
Description of optimisation structures

The types of state optimisation can be divided into three nested parts, which will be described in an order of increasing complexity. The nesting of types of state optimisation is derived from the nesting in the family of negative binomial distributions. In other words, our approach benefits from the fact that the sum of geometric distributions results in a negative binomial distribution. A chain of serially linked self-transitive states therefore emits sequences distributed according to the negative binomial law. In the following we will denote a chain of identical, sequentially linked states mediating the adaptation of a single self-transitive state to a certain length distribution as a macro state (MS).

7.1.1. The geometrically distributed p macro state

While modelling the duration of stay in a certain hidden state in the HMM by a geometric distribution (geo(p)), we get the likelihood:

$$L(p|x_1, \ldots, x_n) = \prod_{i=1}^{n} p(1-p)^{x_i - 1}, \tag{7.1}$$



geo (p)

$$\hat{p}_{MM} = \hat{p}_{ML} = \frac{1}{\bar{x}}$$

nbin (r, p)

$$\hat{p}_{MM} = \frac{\bar{x}}{\bar{x}+\overline{x^2}}$$

$$\hat{r}_{MM} = \frac{\bar{x}^2}{\bar{x}+\overline{x^2}}$$

$$\hat{\Theta}_{ML} =$$

$$\underset{\Theta}{\operatorname{argmax}}\ L\left(\Theta|x_1,\ldots,x_n\right)$$

$$\Theta = (r, p)$$

gnbin (s, r, p)

$$\hat{p}_{MM} = \frac{2(\overline{x^2}-\bar{x}^2)}{\bar{x}^2-\overline{x^2}-2\bar{x}^3+3\bar{x}\overline{x^2}-\overline{x^3}}$$

$$\hat{r}_{MM} = \frac{4(\overline{x^2}-\bar{x}^2)(-2\bar{x}^2\overline{x^2}+\overline{x^2}^2+\bar{x}^4)}{(\bar{x}^2-\overline{x^2}-2\bar{x}^3+3\bar{x}\overline{x^2}-\overline{x^3})(\overline{x^2}-\bar{x}^2-2\bar{x}^3+3\bar{x}\overline{x^2}-\overline{x^3})}$$

$$\hat{s}_{MM} = \frac{-\bar{x}^2\overline{x^2}+\bar{x}\overline{x^2}-\bar{x}^3-\bar{x}\overline{x^3}+2\overline{x^2}^2}{\overline{x^2}-\bar{x}^2-2\bar{x}^3+3\bar{x}\overline{x^2}-\overline{x^3}}$$

$$\hat{\Theta}_{ML} = \underset{\Theta}{\operatorname{argmax}}\ L\left(\Theta|x_1,\ldots,x_n\right)$$

$$\Theta = (s, r, p)$$

Figure 7.1.: Overview of the developed types of state adjustment and the estimation of the parameters s, r and p in terms of method of moment and maximum likelihood. The scheme illustrates the nested modular composition of macro states with the simple p macro state, the rp macro state consisting of a chain of self-transitive hidden states and the srp macro state with additional shifting states.

for given length data $x_1, \ldots, x_n$. Thus we obtain the maximum likelihood estimator for p by maximizing the log-likelihood:

$$\hat{p} = \underset{p}{\operatorname{argmax}}\, \mathcal{L}(p|x_1,\ldots,x_n) \tag{7.2}$$

$$= \underset{p}{\operatorname{argmax}} \log(\prod_{i=1}^{n} p(1-p)^{x_i-1}) \tag{7.3}$$

$$= \underset{p}{\operatorname{argmax}} \log p + \sum_{i=1}^{n}(n_i-1)\log(1-p) \tag{7.4}$$

Therefore solving the equation

$$\frac{\delta\mathcal{L}(p|x_1,\ldots,x_n)}{dp} = \tfrac{1}{p} - \textstyle\sum_{i=1}^{n}(n_i-1)\tfrac{1}{1-p} = 0 \tag{7.5}$$

yields the maximum-likelihood estimator $\hat{p} = \frac{1}{\bar{x}}$. The simplest macro state is represented by a single self-transitive state with geometrically shaped sequence emissions. The left column of Figure 7.1 graphically illustrates the macro state together with an exemplified distributional shape of sequence lengths and the concurrent estimator of ML and MM to obtain the p-parameter of state duration.

7.1.2. The rp macro state – a negative binomial law

The probability density function in Table 7.1 indicates that the geometric distribution is a special case of the negative binomial law with the parameter $r = 1$. As visualised in the middle column of Figure 7.1, sequence lengths following a negative binomial distribution can be modelled by simply chain-linking copies of self-transitive states in an appropriate r-times repetition. In order to obtain the MM estimator we initially express the first two moments μ^1_{nbin} and μ^2_{nbin} in terms of the unknown parameters r and p

$$\mu^1_{nbin} = \frac{r}{p}, \qquad \mu^2_{nbin} = r\frac{1-p}{p^2}. \tag{7.6}$$

In a second step we replace theoretical by empirical moments

$$\bar{x} = \frac{r}{p}, \qquad \bar{x^2} = r\frac{1-p}{p^2} \tag{7.7}$$

and finally solve both equations for r and p

$$\hat{p} = \frac{\bar{x}}{\overline{x^2} + \bar{x}}, \qquad \hat{r} = \frac{\bar{x}^2}{\overline{x^2} + \bar{x}}. \tag{7.8}$$

In respect of model simplicity and time efficiency, the ML estimate for the r-parameter was restricted to integer values. Hence, the determination of macro state parameters relied on maximising the likelihood $L(r, p|x_1, \ldots, x_n)$ given a discretised r-parameter space.

	geometric	negative binomial	generalised negative binomial
macro state	p	rp	srp
PDF	$p(1-p)^{n-1}$	$\binom{n-1}{r-1}p^r(1-p)^{n-r}$	$\binom{n-s-1}{r-1}p^r(1-p)^{n-s-r}$
$E[X]$	$\frac{1}{p}$	$\frac{r}{p}$	$\frac{r}{p} + s$
$Var[X]$	$\frac{1-p}{p^2}$	$r\frac{1-p}{p^2}$	$r\frac{1-p}{p^2}$

Table 7.1.: Summary of properties characterising the different macro states. The table contains the type of distribution, the probability density function (PDF), the expectation value ($E[X]$) and the variance ($Var[X]$).

7.1.3. The srp macro state – a shifted negative binomial law

While screening the length distribution of diverse biological sequences and motifs (see section biological examples), the negative binomial law did not always provide an optimal fit. The incorporation of a shifting parameter s aims at increasing the distributional flexibility especially for distributions of higher location parameter. The s-parameter represents a fraction of single-visit states without self-transition, resulting in a shift of the distribution along the x-axis. We termed this extension to the negative binomial distribution the generalised negative binomial law ($gnbin(s, r, p)$). MM Estimators given in the right panel of Figure 7.1 were derived from the moment generating function

$$\begin{aligned} M_{nbin}(t) &= \sum_{x_i=s+r}^{\infty} e^{tx} f(x) && (7.9) \\ &= e^{t(r+s)p^r(1-e^t+e^tp)^{-r}}. && (7.10) \end{aligned}$$

The estimators of s, r and p were obtained as described for the rp macro state.

7.1.4. Evaluation

With estimators for an optimisation of HMM topologies at hand a crucial aim should be the verification of their beneficial performance. For the purpose of a simple test framework, artificially constructed data sets served as sources of test sequences. These sequences were subjected to a re-estimation procedure in order to evaluate the estimation accuracy of MM against ML in a plain setup. Similarly, the prediction error of HMMs, which topologies were gradually adjusted to length distributions of artificial test sequences, was determined. Several sources of biological sequences were investigated with regard to their length distributions and corresponding parameters were fitted by ML and MM estimation of the various distribution subtypes.

7.2. Artificial test scenario

7.2.1. Construction of artificial test sets

Differently distributed test sets were generated by self-constructed HMMs consisting either of 3 macro states (ROC analysis) or of a single macro state (comparison between MM and ML). The tripartite HMM architecture consists of two flanking macro states, which simulate various length distributions (either by means of p or rp macro states) and a central macro state, which was always designed to generate sequences with negative binomial length distribution ($nbin(32, 0.03)$).

7.2.2. Estimation performance of maximum likelihood vs. method of moments

The performance of MM and ML in determining distributional parameters of empirical sequence length may differ in dependence of its location parameter. This hypothesis was tested by setting up a series of sequence generation processes with differently designed HMMs. HMM topologies in this kind of analysis ranged from a simple p macro state to rp macro states with up to 100 times chain-linked single states. For each of these data sets the underlying length distribution was further on back estimated by ML and MM and the entropy between the reference and the back-estimated distributions was determined. Surprisingly, the analysis revealed an increasing error of the ML estimates and the reference distributions with increasing r-parameter. The MM estimates remain in contrast at a low entropy level throughout the whole range of reference distributions. Only in the first section of the r-parameter ML exhibits slightly better re-estimates than MM (data not shown).

Both, ML and MM are point estimators and their ability to accurately re-estimate parameters of lengths distributed according to the negative binomial family based on a sample of previously generated sequences was investigated. In the setup of the corresponding experiment, reference HMM topologies with different, r-times chain-linked states generated 1000 sequences each.

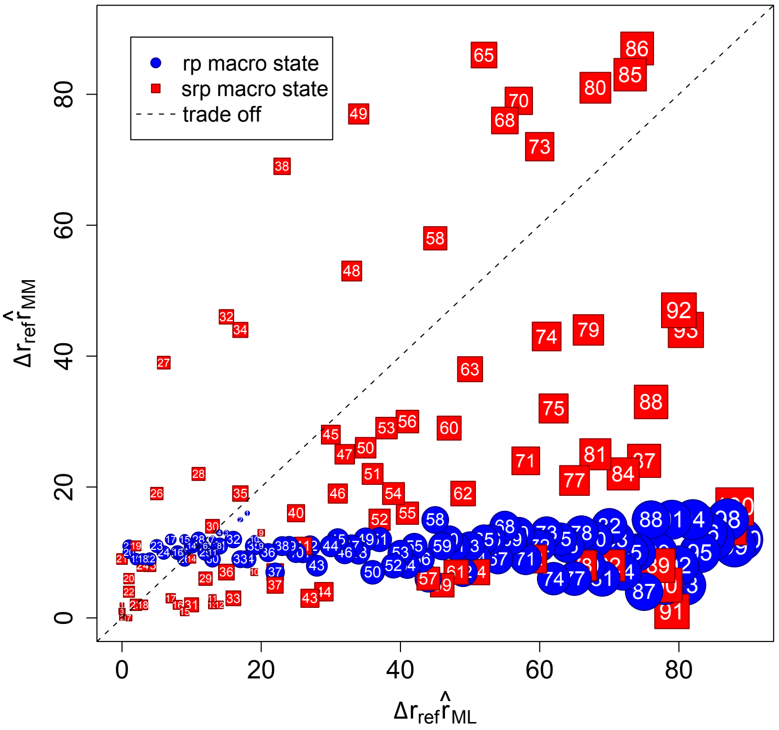


Figure 7.2.: The plot contrasts the point estimation accuracy of maximum likelihood and method of moments in the light of an increasing location parameter of the generating distribution. The axes x and y refer to the distance between the r-parameter of the generating HMM and estimates $\hat{r}_{MM/ML}$ of a re-estimation with MM and ML, respectively. The dashed diagonal marks the balance line between ML and MM. Surprisingly, re-estimations with MM are in contrast to back-estimation results with ML not influenced by the location parameter of the distribution and significantly more accurate regarding r-parameter estimation above $r = 10$.

Figure 7.2 contrasts the distances between r-parameters from re-estimations and original distributions of ML against MM for re-estimations with rp (blue circles) and srp (red squares) macro states. Likewise in the previous comparison of differences in the shapes of distributions, MM estimates reveal more accurate point estimation properties. ML is again able to compete with MM only for samples from reference distributions with small r-parameters. The picture changes when comparing point estimation properties of srp macro states. Here the difference between both methods is small, especially because the point estimates of MM were less stable. The macro state-dependant behaviour of MM could reflect an increasing variance of higher sample moments (Johnson *et al.*, 2005), and therefore the estimator also shows a higher variability.

ROC curves on artificial test data

The following paragraph is guided by the question if and to which extent the prediction quality changes with the suggested topology adjustment of HMMs. Test sequences were sampled 100 times each with HMMs composed of the following macro states:

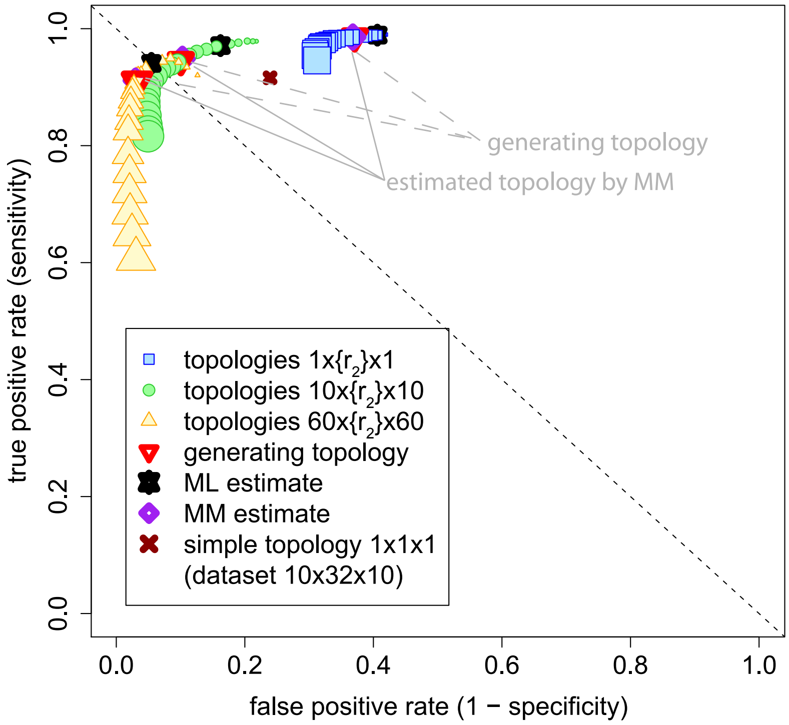


Figure 7.3.: The ROC curve reflects the prediction accuracy of different HMM topologies on datasets generated by 3 HMMs (the central macro state was set to $r = 32$) and the flanking states were varied with repetitions $r = \{1, 10, 60\}$ (squares, circles and triangles). Symbols reflect re-estimations on these datasets with HMMs adjusted by a running value in the range of $r_2 = 1$ to the minimum length of generated test sequences of each HMM. The rank of symbol size indicates the value of r_2 beginning with the smallest one (the p macro state). Sensitivity and specificity are calculated with a focus on the central macro state (positive class), while the flanking macro states were regarded as negative class. The dashed line marks the trade-off between sensitivity and specificity. Further symbols indicate the original (upside-down triangle, red), the simplest topology (dark red cross) as well as the favoured topologies for the data set estimated by ML (black star) and MM (violet diamond).

$MS_1 = MS_3 = \{geo(0.03), nbin(10, 0.03), nbin(60, 0.03)\}$, $MS_2 = nbin(32, 0.03)$. Discrete emissions with values 0.4, 0.3 and 0.3 were assigned in permutated order to a single alphabet of 3 symbols for all macro states. Subsequent analysis exclusively focused on MS_2, assigned as the positive class, while MS_1 and MS_3, denoted as negative class, represent the influence of preceding and successional parts of a HMM topology. In each run, the set of sample sequences was subjected to posterior decoding with a set of re-estimation HMMs which differ in the r-parameter of $MS_2 = \{geo(0.03), \ldots, nbin(r_2^{max}, 0.03)\}$ with r_2^{max} set to the minimum length of all test sequences generated by MS_2. The different HMMs were indicated in Figure 7.3 with symbols of varying size in correlation to the running r_2-parameter. To further highlight the importance of the topology adjustment, a prediction was performed with a simple 3-state HMM consisting exclusively of p macro states (dark red cross).

In summary, the ROC curves underline our expectation of improvements in model prediction through the adaptation of HMM topologies to length distribution of target sequence motifs. Both, the generating HMM topology (up-side-down triangle, red) and the topology suggested by MM (violet diamond) represent a compromise between sensitivity and specificity, while the ML estimate (black star) seems to put increased weight on topologies leading to an optimal sensitivity. Similar analy-

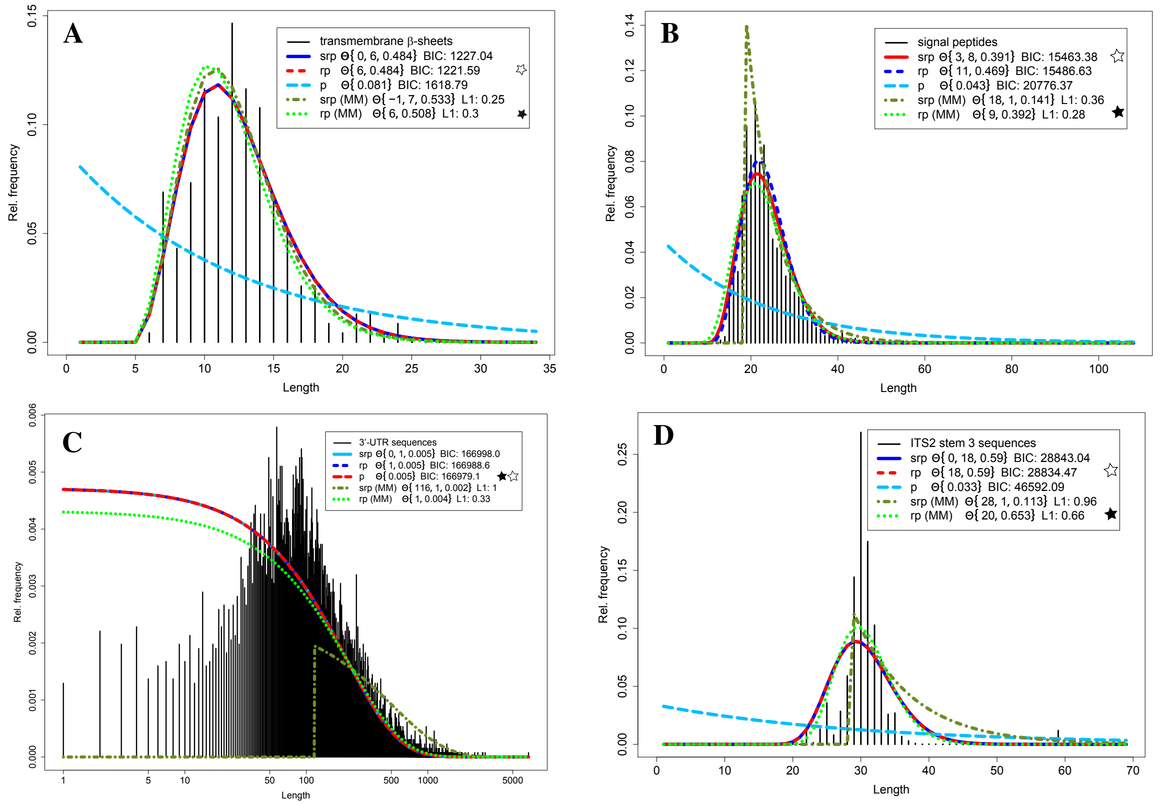


Figure 7.4.: Histograms of empirical length of motifs in protein and DNA sequences and estimations of corresponding distributions of srp, rp and p macro states with maximum likelihood and method of moments. Macro states with identical parameters like the ML estimates for MS_p, MS_{rp} and MS_{srp} in graph C result in the same distributional shape. The BIC selection is plotted on top in these cases. Best models according to either the BIC for ML estimates or the L$_1$ for MM estimates are highlighted by an open or filled star, respectively. The sources of these examples are (A) the lengths of transmembrane β-sheets, (B) the lengths of signal peptides, (C) the lengths of 3'-UTR sequences from *C. elegans* and (D) the lengths of the opening stem 3 of internal transcript spacer 2 sequences from *Asteraceae*.

sis with varying flanking conditions revealed a strong dependency between a decreasing variance in length distributions associated with MS_1 and MS_3 and an increasing specificity. Concurrently, the posterior decoding score of predictions increases with the variance in distributions of flanking macro states.

7.3. Biological examples

Pattern search with HMMs is applied to a wide variety of data types with diverse properties concerning amino acid/nucleotide composition and length distribution. Here, we exemplarily investigated biological sequences and motifs for cases following a negative binomial law. The empirical distributions of the sequences in Figure 7.4 are overlaid with fitted distributions from geometric (ML and MM: long dashes), neg. binomial (ML: short dashes, MM: dots) and generalised neg. binomial (ML: solid line, MM: dashes and dots) macro states. Line colours indicate an order of fitting quality according to the BIC in case of ML estimation. The best fitting estimators are additionally annotated in the legend with a filled (MM) and open star (ML).

A prominent application to predict transmembrane regions in proteins, TMHMM, is based on HMMs, which model length varieties by jumps in their topology (Sonnhammer *et al.*, 1998). Graph

A of Figure 7.4 reflects the length distribution of 232 β-sheet core regions of transmembrane proteins obtained from TMPDB (Ikeda *et al.*, 2003). The ML fitting of motif lengths resulted in identical models for srp and rp macro states, which exhibit a beneficial modelling compared to the geometric macro state. MM provides similar estimations for the srp and rp state types. The estimations suggest the core membrane structure as a target for state optimisation, which would provide a slim alternative to state-of-the-art modelling.

A closely related subject in automated protein annotation is the search for signal peptides. In previous publications signal peptides were modelled with a topology consisting of 3 main structural parts, the n-, h- and c-regions (Käll *et al.*, 2004). Considering signal peptides as an entire structure, the competing model fitting proposed a bell-shaped distribution (see plot B, Figure 7.4) with a repetition factor between $r = 8$ and $r = 11$. In reference to the r-parameters of fitted distribution at least some sub-elements are likely to follow a negative binomial law and its modelling can be improved by adjusted states.

The rapidly growing number of sequenced genomes requires efficient and accurate methods to determine genomic structures like exons and introns, intergenic regions, splice sites or untranslated flanking regions. In Plot C of Figure 7.4, variants of the negative binomial distribution were fitted to 3'-UTR sequence length of *C. elegans* from UTRome database (Mangone *et al.*, 2008). The length distribution of 3'-UTR sequences was consistently predicted to follow a geometric distribution suggesting conventional HMM modelling.

Internal transcript spacer 2 (ITS2) sequences are frequently used as markers in phylogenetic analysis. They separate the 5.8S and 28S rRNA sequences within ribosomal cistrons and are applied to phylogenetic reconstructions because of its sequence variability while maintaining its highly conserved secondary structure with four stem-loops (Coleman, 2007). Example D represents the length distribution and fitted models of the third opening stem of *Asteraceae* sequences from the ITS2 database (Schultz *et al.*, 2006). Recently, a profile HMM based approach for the detection of ITS2 sequences was developed (Keller *et al.*, 2009), providing potential for further improvement with appropriate length models.

7.4. Conclusion

In the previous sections we introduced a methodology to optimise the modelling characteristics of HMM topologies with respect to signal length distributions. We implemented two different ways to estimate distributional parameters, based on maximum likelihood and on an efficient alternative, the method of moments. The presented results of parameter re-estimation on artificial test samples suggest the application of MM as a valuable complement to ML, especially when the location of the expectation value of the target data is large.

Within the framework to optimise HMM topologies we integrated a generalised negative binomial distribution to gain flexibility in distribution modelling. Though tests revealed a good generalisation even with the conventional negative binomial law, some length distributions of biological data sources could be modelled more accurate with the extended negative binomial law (see Figure 7.4 plot B). This finding may as well be due to the restriction of the parameter space for ML estimations to integer values. In principle, the adjustment method can be suited to deal with real-values as chaining parameters r by an additional serially linked state with a decreased probability of stay appropriate to the real fraction of the r-parameter. This adaptation may increase the accuracy of ML estimations in general as well as the flexibility of the conventional negative binomial distribution to fit empirical data. But, the parameter space and in parallel the computational demand would be dramatically increased.

The performance tests for adjusted HMM topologies comprised different scenarios like a central macro state with and without adjustment in flanking macro states. We also varied the extent of adjustment in the central as well as in the flanking macro states. It turned out from these test cases that the error rate in state prediction is dependent on the variance in length distributions of the macro states.

Differences between estimation methods The method of moments provides in general a simple and efficient structure to derive estimators for distributional parameters. As an advantage compared to ML in the presented scenario, MM provides direct parameter estimation without the need of numerical approximation. Though ML is known, at least for large data sets, to provide desirable estimation accuracy, we found in comparative tests higher parameter accuracy with MM. The trade-off was less control of the desired parameter space. Formal restriction to a valid parameter interval did not lead to MM estimators, which implicates the possibility to obtain estimates outside of the parameter domain. Nevertheless, MM outperformed ML in comparative tests. Similarly, previous publications reported lower bias and mean squared error of MM compared to other estimation methods, especially ML (Yamamoto and Yanagimoto, 1992; Allison *et al.*, 2002).

Decoding algorithms Our results indicate a superior prediction accuracy of posterior over Viterbi decoding. The recent implementation of posterior decoding in new HMMer version, HMMer3, might support this observation (Eddy, 2008).

Comparison issues The existence of a number of approaches which adapt HMMs to cope with bell-shaped length varieties indicates the high impact of the topic. Most of the studies focus on the assignment of genes. But, our screening of the length distributions in a broad range of sequence motifs underlines the universal interest associated with the ability to incorporate length information. The sequence length will have an explicit impact when motif detection is based on length-flexible, compositional features rather then on position specific conservation. State-of-the-art gene predictors like GENSCAN rely on semi-HMMs to model typical codon usage as well as length distribution of coding regions. But, explicit connection of states with duration of stay is paid with an increased algorithmic complexity and requires special implementation of training and decoding algorithms.

In contrast, "in-house"-adjustment of conventional topologies as proposed here preserves the applicability of efficient algorithms like posterior decoding. Previously proposed approaches implicate either an accelerating heuristic (Bobbio *et al.*, 2002) or a discretisation of parameter space, while both is avoided by the use of the proposed simple and efficient MM-based method.

In principal, HMMs bear no limitations in distributions which can be modelled. Bilmes (2004) described HMM topologies, which are able to model bimodal distributions via parallelisation of unimodal architectures by a distributing state. The generalisation of the approach is paid by an increase in estimation complexity and may require numerical algorithms for parameter estimation.

With the proposed method at hand, the adjustment of HMM topologies to length distributions of various source data can be easily achieved as a tool either in rapid prototyping or to optimise existing approaches.

A manuscript of the project is in review for publication in Statistical Applications of Genetics and molecular Biology.

8. Modelling interaction sites in protein domains

8.1. Interaction profile hidden Markov model

We applied the probabilistic approach of hidden Markov models to the problem of predicting protein-ligand interaction sites. Our approach is based on the assumption that sequence patterns encoding protein function are shared between members of a domain family. These patterns are often weak and variable. To describe the above mentioned features of domain families, a novel HMM topology was designed by the adaptation of the pHMM (Eddy, 1998; Krogh *et al.*, 1994a) architecture, which is the method of choice for homology detection (Madera and Gough, 2002). The state repertoire was extended by one further match state, namely an interacting match state ($\mathrm{M_i}$). It inherits all features of a match state in the pHMM architecture. The resulting hidden Markov model topology is shown in Figure 8.1.

Every ipHMM is like a pHMM a probabilistic representation of a protein or domain family. The parameters of an ipHMM are estimated from a multiple sequence alignment of domain family members incorporating data on their binding sites and ligands from Pils *et al.* (2005). The same classification of alignment columns as for pHMMs is used here except that an additional occurrence of $\mathrm{M_i}$ states is allowed in matching columns. The new kind of states is provided with the same properties as a match state in the classic profile hidden Markov model architecture. These interacting match states are able to emit all amino acid symbols with probabilities according to their fitted parameters. In Figure 8.1, all bold arrows indicate new transition possibilities. Transition events are restricted to delete, insert and the two match states (main states). The last non-interacting match state demands a transition to the end state.

The information content of domain specific training data influences the accuracy of model parameters. Therefore, ipHMMs were only built for protein domains with more than 20 domain family members in heterocomplexes with resolved structure information in PDB. All sequence positions were labelled with the corresponding interaction status (0 for not interacting and 1 for interacting). The model estimation of the ipHMMs is achieved by maximum likelihood. Transition events were counted together with state emission. A position based weighting scheme (Henikoff and Henikoff, 1994) was applied to compensate for sequence redundancies that occur because of PDB-entries of one protein with different ligands. The weighting calculates sequence weights by associating column-specific weights with the degree of redundancy within one alignment column. The fact that there may be small amount of data in some domains requires the integration of a regularisation method. It prevents zero probabilities in the HMM especially in case of small training sets. Weighted pseudocounts (Durbin *et al.*, 1998) with a total value of 20 for emissions and 5 for a transition set of each type of states are used to solve these problems. The models were estimated for all ligand groups separately

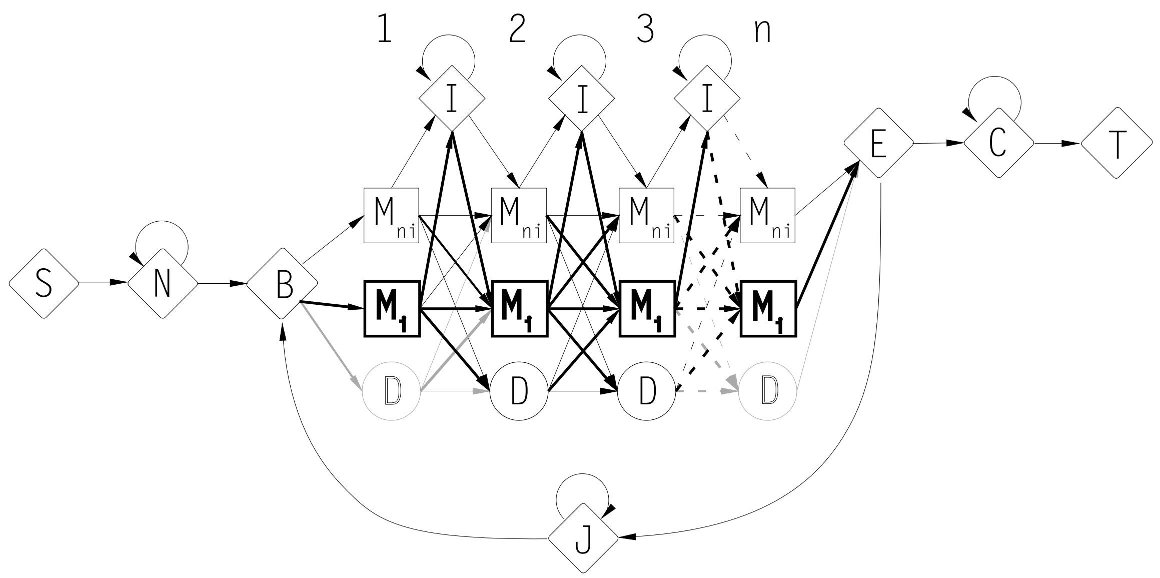


Figure 8.1.: Topology of the interaction profile hidden Markov model following the restrictions and connectivity of the HMMer architecture. The match states of the classical pHMM are split into a non-interacting (M_{ni}) and an interacting match state (M_i). Bold arrows indicate inserted transitions to or from new match states.

with the intention to increase the power of prediction.

Now the problem of applying ipHMMs to the prediction of binding sites in proteins of unknown function has to be faced. A major advantage of the approach is the adaptation of the posterior decoding to the new topology. The algorithm calculates probabilities for all emitting states at each sequence site as shown in Figure 8.3 for the EF-hand domain displaying probabilities of non-interacting and interacting match state. Additionally, delete state probabilities can be displayed to get alignment information corresponding to the domain family. It was necessary to adapt the recursion of forward and backward algorithm to the extended architecture of the interaction profile hidden Markov models.

$$\begin{aligned} f_{M_i^k}(j) &= \frac{e_{M_i^k}(x_j)}{e_{Null}(x_j)}\Big(f_{M_{ni}^{k-1}}(j-1)\tau_{M_{ni}M_i}(k-1) \\ &\quad +f_{M_i^{k-1}}(j-1)\tau_{M_iM_i}(k-1) \\ &\quad +f_{I^{k-1}}(j-1)\tau_{IM_i}(k-1) \\ &\quad +f_{D^{k-1}}(j-1)\tau_{DM_i}(k-1)\Big) \end{aligned} \tag{8.1}$$

$$\begin{aligned} b_{M_i^k}(j) &= \frac{e_{M_i^k}(x_j)}{e_{Null}(x_j)}\Big(b_{M_{ni}^{k+1}}(j+1)\tau_{M_iM_{ni}}(k+1) \\ &\quad +b_{M_i^{k+1}}(j+1)\tau_{M_iM_i}(k+1) \\ &\quad +b_{I^k}(j+1)\tau_{M_iI}(k) \\ &\quad +b_{D^k}(j+1)\tau_{M_iD}(k)\Big) \end{aligned} \tag{8.2}$$

In equations (8.1) and (8.2) the adaptation in the case of the forward and backward values of the interacting match state M_i at sequence site j and profile position k is presented. The emission probability is denoted by e for the indicated state corresponding to a certain sequence and profile position. The transition probability τ is subscripted with indices of the present and the following state as well as the profile position. Forward and backward probabilities for other states are achieved analogously (Rabiner, 1989; Durbin *et al.*, 1998).

The posterior probabilities could be calculated from the knowledge of forward and backward values. The final step is the search of the state path with maximum posterior probabilities via backtracking.

8.2. Validation tests with generated sequences

Estimated ipHMMs are able to emit typical sequences for the corresponding domain family. This feature was used to generate a large test set for a first validation of the prediction power of ipHMMs. The sequences were derived from ipHMM-specific emission and transition probabilities. The dependence of a sequence to model parameters of a certain domain family was the prerequisite of predicting its binding sites. The predicted state path was aligned to the one that was generated afterwards. Sensitivity, specificity and accuracy were calculated in the analysis of the alignment of state paths as mentioned below. Detailed results of all considered domains are listed in Tables C.1 to C.3 of supplementary material. The average sensitivity and specificity values of 0.64 and 0.70 reveal a good quality of predictions for domain-related sequences in contrast to alternative methods (see below).

8.3. Receiver operator characteristics

The evaluation of the prediction method was performed by receiver operator characteristics for several SMART domains. As shown in Figure 8.2 A, ROC curves were calculated for peptide-ligand ipHMMs of the EF-Hand domain, the pancreatic RNAse domain, the alkaline phosphatase domain and the extension to Ser-/Thr-type protein kinase. Figure 8.2 B presents ROC curves of the prediction of ion binding sites. The ipHMMs correspond to the alkaline phosphatase, EF-Hand, PBPe and Villin headpiece domain. In part C of Figure 8.2 ROC of ipHMMs focused on nucleotide-ligands comprising Pumilio-like repeats, pancreatic RNAse domain, HTH lactose operon repressor and C4 zinc finger domain were plotted. Table C.7 of supplementary material summarises test values of all considered ipHMMs. The evaluation consists of cross-validation for a varying discrimination threshold. In this case ROC curves allow to estimate the expected prediction quality of a predictor on new data. Accurate predictors exhibit areas under ROC curves near one.

The examined ipHMMs trained on nucleotide-ligand data showed on average the largest areas under their ROC curves (AUC) and consecutively the highest prediction power. The diversity of prediction quality is higher in the other two ligand-categories. The EF-hand ipHMM is an example of a non-optimal predictor in the cases of peptide- and ion-binding. This might be caused by too few or too similar training data. In contrast the alkaline phosphatase-ipHMM turned out to be a good predictor of ion-ligand interaction sites, while the prediction of peptide-interactions is not perfect. Though the prediction quality of ipHMM varied in some cases, AUC values were overall at a high level.

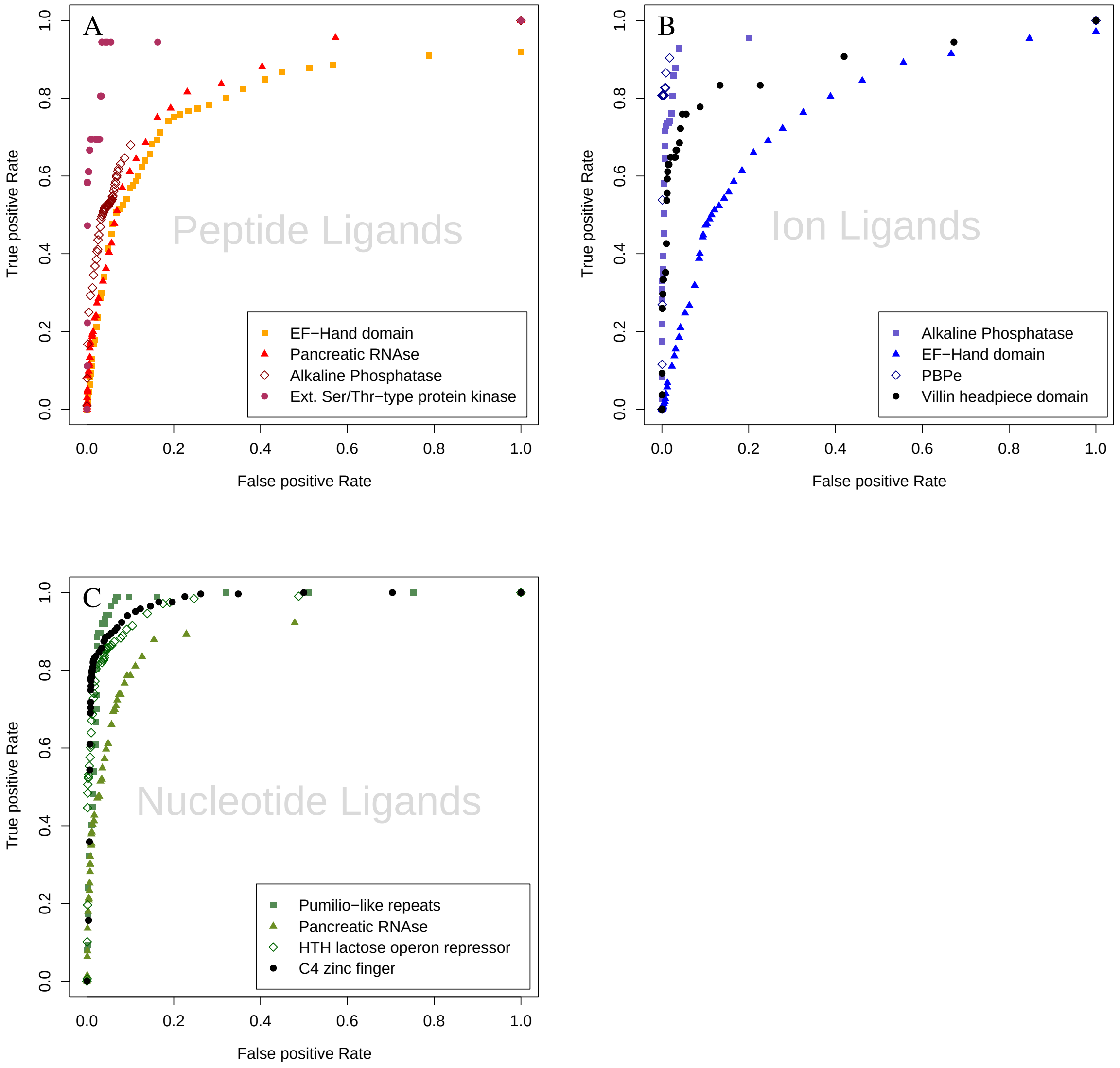


Figure 8.2.: ROC curves indicating the prediction power at various thresholds for the prediction of peptide, ion and nucleotide interaction sites. These calculations were performed for ipHMMs concerning peptide (A), ion (B) and nucleotide ligands (C).

8.4. Validation of predictions on SMART domains

Further testing was enlarged to the whole set of estimated ipHMMs with at least 20 sequences of known structure. We performed a 5-fold cross-validation to calculate the expected prediction accuracy on new data. Referring to results of the described ROCs, a discrimination threshold of 0.2 for posterior match probabilities was chosen as a switching point between an interaction and no interaction to balance average sensitivity and specificity.

With the evaluation of the new approach in mind, different prediction quality indicators were calculated including sensitivity, specificity and accuracy. Their values of all ipHMMs are given in the supplementary material. Results of the validation methods and of ROC indicate the best prediction performance for sites, which interact with nucleotide ligands. These findings are supported by the higher sensitivity values for predictions of nucleotide binding sites.

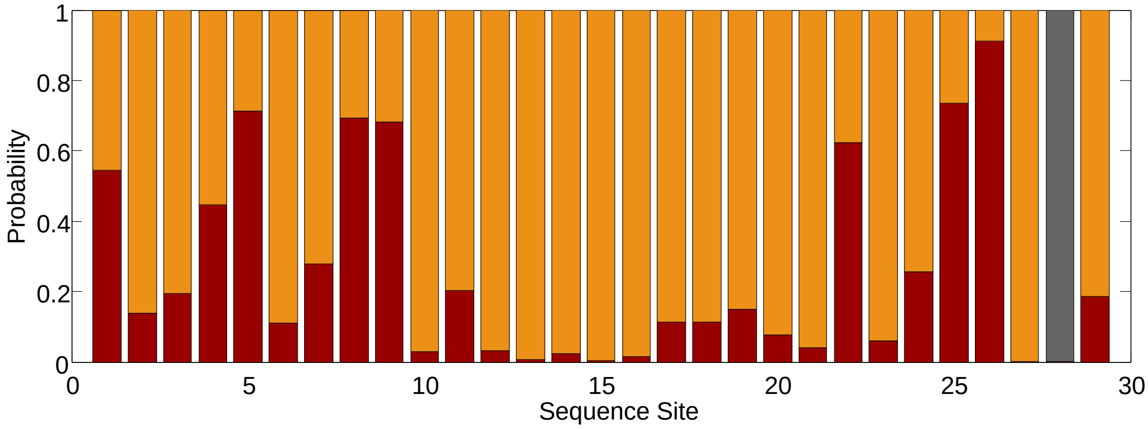


Figure 8.3.: The stacked bar graph represents the prediction result of the posterior decoding for the C-terminal EF-hand motif of *Xenopus laevis*. It contains posterior probabilities of interacting (dark red) and non-interacting match states (orange) and delete states depending on the sequence position. The probabilities for all other states are not displayed because of their low level. All sites with a posterior probability higher than 0.5 for the interacting match were predicted to interact with a peptide ligand.

An investigation of the prediction performance in case of the EF-Hand domain reveals accuracies between 0.73 and 0.86 depending on test and ligand type.

8.5. Interaction site prediction for the EF-hand domain

As an example of use the method was applied to EF-Hand domains of calmodulin from *Xenopus leavis*, whose structure has already been resolved in complex with a peptide of *Caenorhabditis elegans* CaM-kinase kinase (Kurokawa *et al.*, 2001). Predictions of peptide binding sites were performed for all sequences of the EF-Hand family separately using the trained ipHMM for the EF-Hand domain. The output of posterior probabilities for the C-terminal domain is displayed in Figure 8.3. The initial threshold for the prediction of interacting sites was set to a posterior probability of 0.5. The upper graph contains probabilities of non-interacting match states while the graph below shows posterior values of possible interacting sites. Overall, we find tendencies for higher interacting probabilities for match states at the edges of the domain. These areas correspond to its α-helices. The observed interacting positions are localised at sites 1, 5, 9, 22, 25 and 26 while sites 4 and 8 are incorrectly predicted as interactions (false positives, FP). The alignment of the EF-Hand sequence to the ipHMM resulted in 6 correct out of 8 predicted interacting sites.

Figure 8.4 visualises the mapping of correct and false predictions focused on peptide-binding for all four EF-Hand domains of calmodulin from *Xenopus laevis*. All proposed interaction sites were located on the α-helices and their residues were orientated towards the ligand.

8.6. Alternative approaches

An alternative approach to predict protein-protein interactions from sequence information only is based on a neural network with back-propagation (Ofran and Rost, 2003). The underlying data set was derived from PDB by defining interactions as atom-atom distances smaller 6 Å. The method

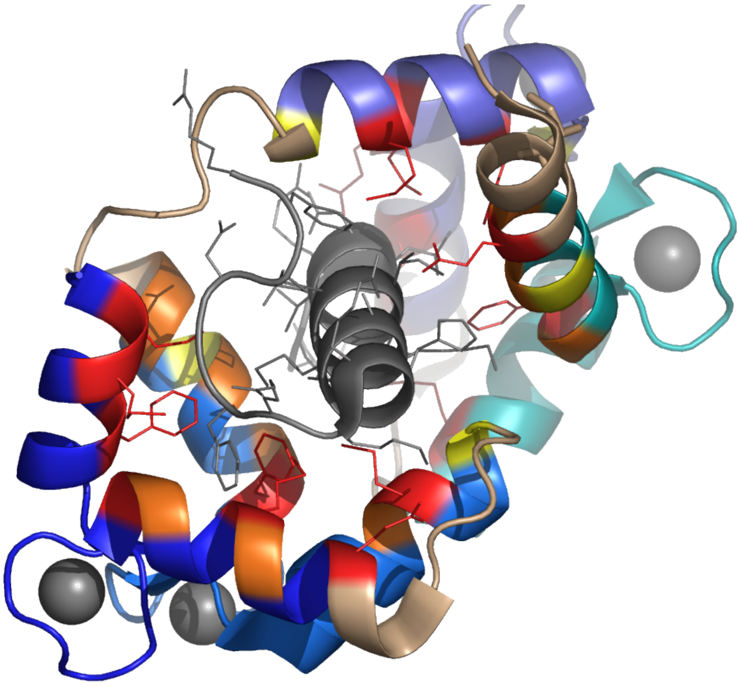


Figure 8.4.: The 3-dimensional protein structure of *Xenopus laevis* calmodulin in a calcium induced ligand binding conformation. The α-helix in the centre of the molecule is a ligand group from a CaM-kinase kinase. The four EF-hand domains are indicated in different blue colours. Residues marked in red are correctly predicted as interacting sites, orange residues are not detected as interactions and yellow residues are erroneous labelled as interacting.

showed a high rate of contact site detection, when only trying to predict sites with highest interacting evidence. But when the algorithm was trimmed to predict less evident contact sites, the prediction quality dropped significantly. The application of the neural network to an unfiltered prediction of binding sites revealed a low sensitivity of approximately 30%. The results of this neural network based approach suggest that ipHMMs will exhibit a better performance for predictions of whole binding interfaces.

Many studies dealt with the binding characteristics of protein sequences. Most of them focused, in contrast to the method presented here, on the investigation of binding patches in proteins of known structure. For this reason we will only concentrate on a comparison to the most recently published method for predicting small molecule binding interfaces of proteins (Snyder *et al.*, 2006). In order to exemplify differences between the approaches, the peroxisome proliferator-activated receptor-gamma (PPAR-γ) was choosen as query protein. A large scale comparison of both approaches was not reasonable because of the restriction of SMID-BLAST to small molecule interactions. The transcription factor in complex with coactivating ligands influences important cellular processes like adipogenesis, anti-inflammatory effects and antiproliferating function in many types of cancer (Lehrke and Lazar, 2005). Experimentally determined interaction sites to a protein were derived from the homodimeric crystal structure of PPAR-γ with a fragment of the steroid receptor coactivator 1 (SRC-1) and rosiglitazone, a high affinity ligand for PPAR-γ (PDB-Identifiyer: 2PRG, Figure 8.5, Nolte *et al.* (1998)). The PPAR-γ sequence was excluded from the training set (consisting of 32 sequences) of a new HOLI-ipHMM, which was built for comparison purpose.

The binding interface of PPAR-γ is located on the hormone receptor binding domain (SMART

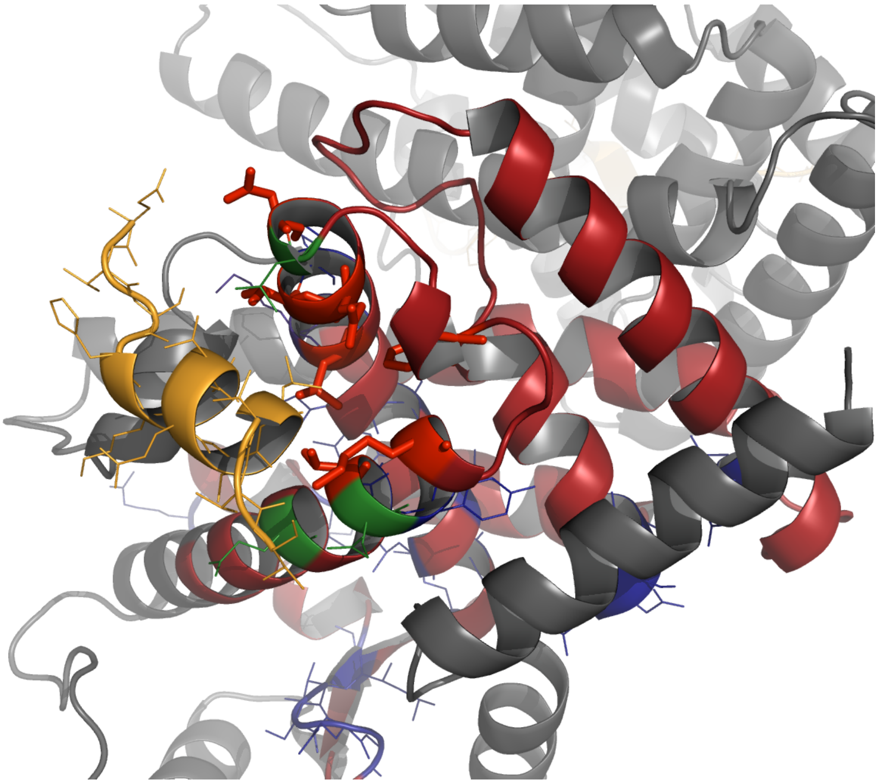


Figure 8.5.: A comparison of SMID-BLAST and ipHMMs mapped to the crystal structure of the homodimeric peroxisome proliferator-activated receptor γ (PPAR-γ) in complex with an LXXLL helix of the SRC-1 co-activator (yellow). PPAR-γ contains a ligand binding domain (SMART: HOLI, brick red) at amino acids 81 to 220. All verified interacting residues were highlighted by red sticks. The colors blue and green represent false positives of SMID-BLAST and ipHMM predictions respectively.

domain Holi). Then the interaction sites were determined as described above. The predictions are mapped on the structure of the PPAR-γ complex. The holi domain is coloured in brick red, correctly predicted peptide-binding sites with the ipHMM are displayed in red colour. The SMID-BLAST prediction overlaps only at position 314Q with the experimentally derived binding interface. Further incorrectly assigned interactions by SMID-BLAST are shown in blue. SMID-BLAST provides a list of binding patches to the user, each binding one small molecule. While evaluating the results of SMID-BLAST, all top ten binding patches were scanned to get an entire set of different predicted interaction sites.

The example prediction underlines the advantages of the HMM-based method in predicting the more variable protein interaction sites. The knowledge of interaction sites is in contrast to SMID-BLAST not directly transferred from single members of the domain family, but probabilistically assigned taking all known interactions in the domain at a given position into account. The ipHMM was able to find all verified peptide interactions except those at sites 20 and 26, while SMID-BLAST only found one contact site which overlaps with the binding patch derived from the structure of the complex. In contrast SMID-BLAST seems to reveal a high rate of false positives in this case. The ligands proposed by SMID-BLAST were mainly ions or organic compounds. With the purpose of a general comparison of Blast-based approaches to ipHMMs, a simple predictor was created, which searches for homologous sequences in the dataset of proteins with verified interactions. The best hit at a given identity threshold was taken to transfer its interactions to the query. The Sensitivity of the prediction was significantly lower compared to the average values observed for ipHMMs. Data on the results of this predictions can be found in supplementary material.

8.7. Large-scale analysis of point mutations

The described method was furthermore applied to the detection of interactions involved in diseases. The aim of this investigation was to understand the mechanisms of molecular dysfunctions arising from point mutations. The OMIM database contains informations of known mutations that cause diseases. In cases where these mutations were located within SMART domains, an interaction site prediction outlined consequences of mutated protein binding sites. We found 38 cases, where a disease-triggering mutation was associated with an interaction site. The results of this investigation are presented in Table 8.1. An interaction site to an ion-ligand at position 317 concerns a severe mutation in the human alkaline phosphatase (PPBT_HUMAN). The structure of this protein is not yet available. Mutations in this domain cause different forms of hypophosphatasia, a defect in bone mineralisation. Glycine at position 317 is located in the highly conserved region of the active site (Greenberg *et al.*, 1993). The residue is described to form hydrogen bonds to residues 315 and 320 (Zurutuza *et al.*, 1999). The first is involved in Mg^{2+}-coordination and the last in Zn^{2+}-binding. This information explains a severe defect in the enzymatic activity of the alkaline phosphatase because of a non-conservative mutation in the ion-binding site. A detailed look at mutated interaction sites of tissue-non-specific human alkaline phosphatase corroborates the predictions. All listed mutations were found in patients with hypophosphatasia (Taillandier *et al.*, 1999, 2000, 2001; Zurutuza *et al.*, 1999; Greenberg *et al.*, 1993). Structure prediction via homology modelling located the active site of the enzyme at the following residues: 43, 92, 93, 94, 154, 156, 167, 170, 315, 320, 324, 361, 362, 364 and 437 (Zurutuza *et al.*, 1999). A severe mutation was described at position 317 of the active enzyme (Zurutuza *et al.*, 1999; Greenberg *et al.*, 1993). This amino acid builds hydrogen bonds to the residues 315 and 320 of the active site. Residue 315 is involved in Mg^{2+}-coordination and site 320 interacts with Zn^{2+}. The prediction by the ipHMM method determined the residue as an ion- and peptide-ligand binding site.

The human cyclic nucleotide-gated cation channel α 3 (CNGA3_HUMAN) protein belongs to a family of ion-channels that share a common structure containing six transmembrane domains and a carboxy-terminal cGMP-binding site. A mutation at site 529 from valine to methionine destroys a conserved VVA motif, which is associated with cGMP-binding. Miss-sense mutations provoke achromatopsia, the total colour blindness (Kohl *et al.*, 1998). In the ipHMM-based analysis this residue was identified as interacting to nucleotide ligands. As in the case above no structure information to this protein is currently available.

These two cases demonstrate the importance of knowledge about interacting positions in understanding molecular reasons of hereditary diseases. The application of ipHMMs will provide the profit of prior information about interaction sites in mutational analysis.

Table 8.1.: Interaction site prediction in sequences with disease-related mutations.

Sequence	Domain	Diseases	OMIM-ID	Mutation[1]
BTK_HUMAN	SH2	Hypogammaglobulinemia and isolated growth hormone deficiency, X-Linked; X-linked agammaglobulinemia (XLA) and isolated growth hormone deficiency	307200, 300300	288P, 308P, 334P
PPBT_HUMAN	alkPPc	Hypophosphatasia	171760, 241500, 241510, 146300	71P, 211I, 220PI, 223PI, 235P, 249PI, 334PI, 426I, 436I, 450PI, 456PI
RB_HUMAN	CYCLIN	Retinoblastoma; osteosarcoma; bladder cancer; pinealoma with bilateral retinoblastoma	180200, 109800, 259500	661P, 712P
CAN3_HUMAN	EFh	Muscular dystrophy, limb-girdle, type 2A	253600, 114240	705I, 744P
DAX1_HUMAN	HOLI	Congenital adrenal hypoplasia with hypogonadotropic hypogonadism; dosage-sensitive sex reversal	300200, 300018, 300473	267P
INS_HUMAN	IlGF	Diabetes mellitus, rare form; MODY, one form; familial hyperproinsulinemia	176730	89PI, 92PI
ANDR_HUMAN	ZnF_C4	Androgen insensitivity, several forms; spinal and bulbar muscular atrophy of Kennedy; prostate cancer; perineal hypospadias; male breast cancer with Reifenstein syndrome	300068, 312300, 313200, 313700	568I, 571IN, 580IN, 581IN, 582N, 585N, 608N, 615N
CNGA3_HUMAN	cNMP	Achromatopsia 2	216900, 600053	529N, 547N
ABCD1_HUMAN	AAA	Adrenoleukodystrophy; adrenomyeloneuropathy	300100, 300371	514P, 518P, 515N
NKX25_HUMAN	HOX	Atrial septal defect with atrioventricular conduction defects	108900, 600584	188N, 191N
GELS_HUMAN	GEL	Amyloidosis, finnish type	137350, 105120	214I
GLI3_HUMAN	ZnF_C2H2	Greig cephalopolysyndactyly syndrome; Pallister-Hall syndrome; preaxial polydactyly, type IV; postaxial polydactyly, types A1 and B	165240, 175700, 174700, 174200, 146510	515I

[1] In this column the mutation site is displayed together with a shortcut indicating the corresponding type. The following abbreviations occur: “P” for Peptide ligands, “I” for Ion ligands and “N” for Nucleotide ligands. Combinations of shortcuts denote binding sites interacting with different types of ligands

8.8. Conclusion

In this article, a new method for the prediction of protein binding sites to different types of protein ligands was introduced. It is the first in incorporating information about homology as well as binding sites in a hidden Markov model topology. Those HMMs have already been applied to various analytical tasks as a result of their efficiency and comparatively high accuracy. The main novelty in the architecture of the interaction profile HMM is a second match state that represents interacting sequence positions. It was demonstrated in validation tests and on the example of calmodulin binding sites that the algorithm is able to detect the majority of existing interactions in a protein sequence. Interacting positions were determined from structures of protein-ligand complexes according to the length of a hydrogen bond (4 Å).

The detection of a wide range of ligand binding sites is enabled with the introduced approach. In contrast to most alternative solutions, the ipHMMs predict interaction sites in the context of domain families, which leads to a higher prediction quality. The increase of predictive power is indicated by a significantly higher sensitivity. IpHMMs provide in comparison to alternative methods like SMID-BLAST a larger spectrum of predictable types of interaction sites. Furthermore, interfaces consisting of a novel combination of known interaction sites in a domain family could be detected by ipHMMs.

For all existing predictors including ipHMMs, initial structure information is necessary for the training process. Once the corresponding ipHMMs have been trained, binding sites could be determined in sequences of unknown structure. The sensitivity for contact site detection of ion ipHMMs is slightly lower than for peptide and nucleotide ipHMMs, because of lower sequence coverage. Increasing amounts of identified protein structures will improve the prediction power of ipHMMs in general and especially in cases where still little sequence information is available.

The proposed method provides further information on the quality of single interaction site predictions. The state-associated posterior probabilities of sequence positions indicate how well the used ipHMM can distinguish between state alternatives. This is a valuable assistance in interpreting prediction results. The new ipHMMs inherited all features of profile hidden Markov models. Once the amount of interaction site data reaches a certain level, existing HMMs in frequently used databases like SMART and Pfam could be replaced by ipHMMs.

The developed approach supplements existing experimental tools for the investigation of changes in molecular mechanisms caused by miss-sense mutations. The ability of ipHMMs to predict interaction sites and ligand types assists the analysis of mutated sites in proteins. Due to the homology-based approach, these studies can be performed for proteins which structure is still unknown. This technique highlighted the consequence of mutations e. g. in the ion-binding region of the human tissue non-specific alkaline phosphatase and in the cGMP-binding motif of the human cyclic nucleotide gated cation channel 3. The large-scale screening of mutated interaction sites in human protein highlight the impact of interaction site prediction in elucidating causes of severe inheritable diseases like prostate cancer, breast cancer, diabetes mellitus or muscular dystrophy.

8.8.1. Future perspective

Increasing data of protein sequences and structures will lead to a good sequence coverage for the majority of domain families and consecutively to improved interaction profile hidden Markov models. Furthermore, these new protein sequences and structures open up the possibility to build ipHMMs of new domain families or known families that are not yet included in the ipHMM library because of a small basis of data. Other types of binding interfaces like those for carbohydrates or lipids could easily be modelled with the same HMM-topology.

This project is published in Bioinformatics (Friedrich et al.*, 2006).*

Part III.

General discussion

The rapid development of high-throughput technologies in many fields like ecology, systems biology, microbiology or molecular biology faces new challenges in the analysis of massively generated data. This work describes many facets of modern biomedical research beginning with inference of knowledge by large-scale comparisons over a holistic approach of a data driven development of high-throughput diagnostics to methodical improvements in sequence analysis. The first project comprises a concept to gain insights into taxonomic diversity via a highly parallel comparison of enterobacterial genomes. Novel methods for the comparison and processing of genomic sequences, including correspondence analysis and statistical tests, have been established. These methods contribute to the construction of an overall picture of enterobacteria and to the specification of factors of genus diversity. A second genomic approach concerns the development of a diagnostic microarray for enterobacteria based on the determination of maximally discriminating oligonucleotide probes. According to clinical pathology, this medically important bacterial family was diagnostically addressed by a completely new concept of microarray design and analysis. Enhanced evaluation of microarray experiments concerning *A. thaliana* gene expression was in focus of another project. As in the first approach, I demonstrate the advancing application of multivariate analysis to multiple genomic datasets, in this case genome-wide expression profiles under different conditions. Algorithmic improvements have been in focus of two HMM-related projects. The first approach provides a general enhancement of modelling properties of HMMs in sequence analysis. This goal is achieved by topological optimisation based on a moment estimator or maximum likelihood. The second approach extends conventional profile HMMs to cope with structural data in order to predict interaction sites in protein domains.

Chapter 4 describes **the comparison of a multitude of enterobacterial genomes** on different levels of genomic organisation. Preliminary phylogenetic and phylogenomic reconstructions highlighted the fundamental need for the consideration of the different levels as they represent independent evolutionary processes. Novel methods for the simultaneous comparison in unsupervised and supervised manner are introduced. Unsupervised investigations do not require prior knowledge about the nature of compared strains, though subsequent interpretations require the integration of the results into a broader context. Here, CA was applied on protein-related and domain-related mapping data to cluster the strains into functionally distinct groups. Guided by this first-order exploration, follow up supervised analysis assure the determination of specific protein families among groups of genotypically related strains. The obtained candidates were functionally investigated and grouped according to metabolic context. A considerable fraction of the specific protein family exhibited only puristic or totally lacking functional annotation. Although homology based transfer of knowledge is invaluable to get global insights into the diversity within taxonomic groups, the applied methods cannot fully replace experimentally determined functionality and rather require a baseline annotation. Therefore, it remains an important future challenge to increase the efficiency of functional annotation in genomic projects. Starting points for such experimental analysis are supplied by the described methods. Within a target taxon the applied methods allow nearly free scaling of subgroups excepting a minimal group size (6) due to limitations of statistical tests. Up-coming genome sequences will further strengthen the drawn conclusions on characteristic traits not only in enterobacterial groups, as the presented approaches are in principally portable to any bacterial taxon.

Enterobacteria do not only represent an interesting bacterial family with prominent model organisms in life-science and biotech, they are predominantly known as versatile pathogens causing several distinct clinical symptoms with high annual incidences worldwide. The distinction of strains to pathogroups responsible for these different symptoms has been in focus of several studies (Cassone *et al.*, 2007; Loy and Bodrossy, 2006). In chapter 5 the complete process including microarray layout, probe selection, sample preparation, testing, development of analytical methods as well as evaluation of analytical results is presented. The designed **diagnostic microarray** is based on novel principles concerning diagnostic target, probe selection and data analysis. As the main novelty, the oligonucleotide probes were recruited from both coding and non-coding areas of reference genomes using a sophisticated string matching algorithm. Some pathogroups exhibited small sets of candidate probes suggesting a liberalisation of probe selection criteria or a reduction of probe length to increase the pool of backup probes. The selected probes exhibit interesting links e. g. to intergenic regions or to virulence associated genes. Further investigation should be performed regarding the many uncharacterised traits underlying the probes, as these genes likely contribute to specific features of respective pathogroups. A main objective in clinical diagnostic constitutes cost reduction of single tests in conjunction with efficiency in testing time. Consequently, the chosen slide format, the HTA$^{\text{TM}}$ Slide12 from Greiner Bio-One, allows for parallel hybridisations. Certainly, further measures of cost-reduction have to be taken with respect to broad clinical applications. Cost-reduction is achievable by a reduction of the number of probes and especially by removing suboptimal performing reverse complementary probes. Future technical developments will also contribute to progress in this direction, though it is not yet clear if microarrays will replace conventional diagnostics.

Another strength of the developed microarray constitutes the applied regression model to evaluate hybridisation profiles. The model is able to predict hybridised amounts of DNA and serves in parallel as a classificator for enterobacterial pathogroups. Classifications based on comprehensive test hybridisations generally coincide with expected pathogroups of test samples. Ambiguities arise from classifications in the group of avian pathogenic *E. coli* (APEC, data not shown). The pathogroup has not been included in the core study as it is only indirectly relevant in clinical diagnostics. Merely a single and maybe atypical APEC genome (Johnson *et al.*, 2007) has been available as reference in probe selection. Furthermore, a large study aiming at a characterisation of the APEC pathotype failed to identify common patterns of known virulence genes amongst isolates from Ireland (McPeake *et al.*, 2005). Non-pathogenic *E. coli* strains form a second inhomogeneous and rarely characterised subgroup. Beside commensal intestinal isolates, the subgroup is compost of a mixture of laboratory strains like the K-12 isolates and the Nissle strain, which genotypically resembles UPEC strains without expressing UPEC-specific virulence factors (Grozdanov *et al.*, 2004). *E. coli* K-12 is in use as a laboratory strains for nearly 90 years and was frequently passaged and genetically manipulated (Bachmann, 1972). Therefore the K-12 lineage does probably not represent ‘typical’ commensals. Though this heterogeneous subgroup could be characterised by oligonucleotide determinants, it would be advantageous in a diagnostic context to focus on ‘true’ commensals that were isolated from the intestinal tract, given the respective genomic data. These examples underline the importance of well defined bacterial subgroups in order to enhance the performance of any microbial diagnostic device.

The basic concept and analytical elements of the described microarray development can be easily transferred to other bacterial clusters and even beyond. Although the microarray design was focused on clinical diagnostics, its application to further fields like quality control of food or water as well as veterinary medicine is imaginable. In summary, a novel, complete developmental process of a diagnostic microarray, which enhances the diagnostic reliability especially on subspecies levels, is demonstrated. The specifically adapted regression model further improves the diagnostic performance via continuous learning abilities in the process of its application

Gene expression analysis is an important and well established method to unravel and connect metabolic functions of genes. Despite the availability of microarray platforms for many organisms and a wealth of expression profiles generated under various conditions, only few approaches that are capable to integrate all these results exist. In chapter 6 I describe a **meta-analysis methodology** based on kPCA and hierarchical clustering to integrate results of different experiments. The meta-analysis requires a certain data quality which is not always provided by data sets in public databases. Especially processed expression data sets largely vary in quality. Thus, the restriction to unprocessed data in conjunction with rigorous outlier removal is advisable and was implemented to assure reasonable results. KPCA provides an unsupervised clustering of similar contrasts and in parallel allows for deriving most influencing features. Principal component analysis is methodically related to CA. Therefore, gene selection might even be optimised by appropriate statistical tests as described in the first project on enterobacterial genomics.

The integration of expression results obtained by using platforms with differences in spotted gene libraries remains a challenge. Appropriate solutions are the restriction of the analysis to the intersection of gene libraries or the restriction to a single platform, as preferred here for reason of simplicity. Although whole genome microarrays of different developers may differ in probe length, slide chemistry, target labelling or printing variables (Hardiman, 2004), the general comparability and reproducibility of gene expression results across platforms has been reported (Consortium *et al.*, 2006). Once setup, the meta-analysis steadily gains in robustness concerning the detection of commonly deregulated genes. In contrast, growing data sources shift the focus of the analysis towards more distinct differences in expression profiles. Minor aspects of differential gene expression might become more difficult to detect. However, as a strength of the described meta-analysis, the ‘resolution’ can be modulated by reducing the data set to relevant contrasts. The approach is portable to other platforms like whole genome oligonucleotide microarrays designed for *H. sapiens*, *M. musculus* or *E. coli*. The most relevant objective in the choice of appropriate platform types certainly is acceptance and establishment of these platforms in the microarray community, as a seed amount of expression profiles is required to obtain the maximum performance with the meta-analysis approach.

HMMs found wide-spread use in many applications of computational biology. The probabilistic models provide flexible frameworks, which can be adapted to various modelling problems by changes in its topology. Despite its wide-spread usage, I demonstrated the suboptimal modelling behaviour due to distributional characteristics of single self-transitive states. Therefore, a **methodology to optimise the modelling behaviour of HMMs** by serial chain-linking of self-transitive states was developed (see chapter 7), while the number of chain-linked states and the respective holding time was estimated by maximum likelihood and the straight forward moment estimator. Optimised HMMs revealed better modelling properties in artificially constructed test scenarios and especially in modelling of biological sequence data. In contrast to existing solutions implemented in algorithms for the prediction of genes the proposed optimisation maintains the class of conventional HMMs. The efficient and highly sensitive decoding algorithms remain unchanged. The moment estimator outperforms maximum likelihood concerning distributional re-estimation on artificially constructed test scenarios. Nevertheless, no moment estimator could be defined, which includes restrictions of parameter space to positive r-values. But, both tests on artificially constructed and real life data sets confirm the could confirm the validity of estimates obtained from method of moments In the current implementation, optimised models are restricted to representations of unimodal negative binomial distributions. Principally, the distributional repertoire, which can be modelled by simple topological changes, enables extensions to multi-modal distributions and beyond (Bilmes, 2006).

Although protein sequence databases have dramatically increased in size over the last years, structural information especially on protein complexes is still rare. Several approaches exist to determine binding interfaces, accessible surface area and other parameters of existing protein structures (Zhou and Qin, 2007). A more challenging task constitutes the determination of structural properties on sequence information alone. The development of an HMM-based method as the first fully probabilistic **approach to predict interaction sites in protein domains** (see chapter 8) enables large-scale structural annotation by homology-based knowledge transfer.

Limitations have arisen from rarely available structure information of protein complexes in a majority of domains. Although a recently generated update resulted in substantial increase in the amount of training data (data not shown), the overall coverage of interaction data to protein domains is still not optimal. Even the incorporation of Pfam domains does not fill this gap. The applied SMART database of profile HMMs obviously contains models of protein domain families showing high connectivity. Therefore, SMART domains exhibit higher abundance in resolved structures of protein complexes than Pfam. Protein structures deposited in PDB are generated by individual research projects. The nature of resolved proteins and protein complexes reveals a certain bias towards candidates of high interest e. g. from certain model organisms or simply towards candidates showing appropriate properties to facilitate structure determination (Peng *et al.*, 2004). The influence of this bias is higher in the small PDB database than in large protein repositories like UniProt. Potential overfitting arising from such bias was overcome by the implementation of sequence weighting in HMM training. Recent developments further prevent overfitting by an alternative training approach, which incorporates distributional parameters of profile HMMs. Thus, ipHMMs inherit the properties of profile HMMs trained on large, manually curated seed alignments as background distributions of transitions and emissions events. Currently, these new features are incorporated into a web application to provide access to ipHMMs for a broader community. In contrast to several other approaches (Bradford and Westhead, 2005; Fariselli *et al.*, 2002; Jones and Thornton, 1997) ipHMMs are capable to detect and classify different categories of interactions. Initially the ipHMMs were separately trained according to the categories peptide ligands, ion ligands and nucleotide ligands. The spectrum is currently extended to the detection of carbohydrate and miscellaneous ligands.

In applications to proteins associated with inheritable diseases, ipHMMs provided prediction-based explanations for protein dysfunctions. Severe diseases like certain types of cancer or muscular dystrophy are associated with mutated interaction sites indicating the method's medical impact. In combination with threading or molecular modelling, ipHMMs could be applied to large-scale assessments of effects arising from missense mutations. Previously, the application of automatic design algorithms has successfully modulated binding properties in an apoptosis-related ligand (van der Sloot *et al.*, 2006). A large study of interaction sites in single nucleotide polymorphisms (SNPs) of the human genome revealed an unexpectedly large number of 1710 SNPs at interacting positions (data not shown). Exemplarily, the investigation of SNPs at interacting positions revealed substantial disruptions in binding capacity. Although binding interfaces of proteins mostly rely on more than one interaction site, at least decreased ability of binding could be stated. Mutational events could principally also lead to the formation of new interaction sites in vicinity to the binding interface, but such effects have not yet been investigated. Overall, these applications underline the importance of ipHMMs in studying functional sites of protein domains and they give an impression of the methodical potential in the light of increasing amounts of structural information.

The presented work describes analytical methods for different levels of biological information including genomes, proteomes, transcriptomes, single protein sequences and protein domain structures. The simultaneous consideration of several sources of information about organisms and whole environments will become more and more important to understand the systems, in which these organisms participate. In all fields I observed the further requirement of additional high quality source data to increase the reliability of analytical results. Due to recent technical progress, it seems to be just a matter of time that these objectives will be met.

References

Abbott, D. W. and Boraston, A. B. (2008) Structural biology of pectin degradation by *Enterobacteriaceae*. *Microbiol Mol Biol Rev* **72**(2): 301–16, table of contents.

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A. *et al.* (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci U S A* **96**(24): 14043–14048.

Allison, D. B., Fernandez, J. R., Heo, M., Zhu, S., Etzel, C. *et al.* (2002) Bias in estimates of quantitative-trait-locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. *Am J Hum Genet* **70**(3): 575–585.

Aloy, P., Böttcher, B., Ceulemans, H., Leutwein, C., Mellwig, C. *et al.* (2004) Structure-based assembly of protein complexes in yeast. *Science* **303**(5666): 2026–2029.

Angot, A., Vergunst, A., Genin, S. and Peeters, N. (2007) Exploitation of eukaryotic ubiquitin signaling pathways by effectors translocated by bacterial type III and type IV secretion systems. *PLoS Pathog* **3**(1): e3.

Antonov, A. V., Tetko, I. V., Mader, M. T., Budczies, J. and Mewes, H. W. (2004) Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics* **20**(5): 644–652.

Bachmann, B. J. (1972) Pedigrees of some mutant strains of *Escherichia coli* K-12. *Bacteriol Rev* **36**(4): 525–557.

Bandelt, H. J. and Dress, A. W. (1992) Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol Phylogenet Evol* **1**(3): 242–252.

Barl, T., Dobrindt, U., Yu, X., Katcoff, D. J., Sompolinsky, D. *et al.* (2008) Genotyping DNA chip for the simultaneous assessment of antibiotic resistance and pathogenic potential of extraintestinal pathogenic *Escherichia coli*. *Int J Antimicrob Agents* **32**(3): 272–277.

Barrera, F. N., Poveda, J. A., González-Ros, J. M. and Neira, J. L. (2003) Binding of the C-terminal sterile alpha motif (SAM) domain of human p73 to lipid membranes. *J Biol Chem* **278**(47): 46878–46885.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D. *et al.* (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res* **35**(Database issue): D760–D765.

Bartel, B. and Fink, G. R. (1994) Differential regulation of an auxin-producing nitrilase gene family in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A* **91**(14): 6649–6653.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res* **32**(Database issue): D138–D141.

von Baum, H. and Marre, R. (2005) Antimicrobial resistance of *Escherichia coli* and therapeutic implications. *Int J Med Microbiol* **295**(6-7): 503–511.

Bekal, S., Brousseau, R., Masson, L., Prefontaine, G., Fairbrother, J. *et al.* (2003) Rapid identification of *Escherichia coli* pathotypes by virulence gene detection with DNA microarrays. *J Clin Microbiol* **41**(5): 2113–2125.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 289–300.

Benjamini, Y. and Hochberg, Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25**: 60–83.

Benzécri, J. (1992) Correspondence Analysis Handbook. CRC Press.

Berger, S., Papadopoulos, M., Schreiber, U., Kaiser, W. and Roitsch, T. (2004) Complex regulation of gene expression, photosynthesis and sugar levels by pathogen infection in tomato. *Physiologia Plantarum* **122**: 419–428.

Bilmes, J. A. (2004) What HMMs can't do. *IEIC Technical Report* **104**(541): 25–30.

Bilmes, J. A. (2006) What HMMs can do. *IEICE Transactions in Information and Systems* **E89-D**(3): 869–891.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**(5331): 1453–1474.

Blum, G., Ott, M., Lischewski, A., Ritter, A., Imrich, H. *et al.* (1994) Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen. *Infect Immun* **62**(2): 606–614.

Bobbio, A., Horváth, A. and Telek, M. (2002) PhFit: A General Phase-type Fitting Tool. In Proceedings of the International Conference on Dependable System and Networks (DNS'02).

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**(1): 365–370.

Box, G. E. P. and Cox, D. R. (1964) An Analysis Of Transformations. *Journal Of The Royal Statistical Society Series B-Statistical Methodology* **26**(2): 211–252.

Boyd, E. F. and Hartl, D. L. (1999) Analysis of the type 1 pilin gene cluster fim in *Salmonella*: its distinct evolutionary histories in the 5' and 3' regions. *J Bacteriol* **181**(4): 1301–1308.

Bradford, J. R. and Westhead, D. R. (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics* **21**(8): 1487–1494.

Brenner, F. W., Villar, R. G., Angulo, F. J., Tauxe, R. and Swaminathan, B. (2000) *Salmonella* nomenclature. *J Clin Microbiol* **38**(7): 2465–2467.

Brisse, S., Grimont, F. and Grimont, P. (2006) The Prokaryotes - The Geuns *Klebsiella*, vol. 6.

Bruant, G., Maynard, C., Bekal, S., Gaucher, I., Masson, L. *et al.* (2006) Development and validation of an oligonucleotide microarray for detection of multiple virulence and antimicrobial resistance genes in *Escherichia coli*. *Appl Environ Microbiol* **72**(5): 3780–3784.

Brumell, J. H., Rosenberger, C. M., Gotto, G. T., Marcus, S. L. and Finlay, B. B. (2001) SifA permits survival and replication of *Salmonella typhimurium* in murine macrophages. *Cell Microbiol* **3**(2): 75–84.

Brzuszkiewicz, E., Brüggemann, H., Liesegang, H., Emmerth, M., Olschläger, T. *et al.* (2006) How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains. *Proc Natl Acad Sci U S A* **103**(34): 12879–12884.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**(1): 78–94.

Butler, T. (1994) *Yersinia* infections: centennial of the discovery of the plague bacillus. *Clin Infect Dis* **19**(4): 655–61; quiz 662–3.

Buts, L., Bouckaert, J., Genst, E. D., Loris, R., Oscarson, S. *et al.* (2003) The fimbrial adhesin F17-G of enterotoxigenic *Escherichia coli* has an immunoglobulin-like lectin domain that binds N-acetylglucosamine. *Mol Microbiol* **49**(3): 705–715.

Callaway, T. R., Elder, R. O., Keen, J. E., Anderson, R. C. and Nisbet, D. J. (2003) Forage feeding to reduce preharvest *Escherichia coli* populations in cattle, a review. *J Dairy Sci* **86**(3): 852–860.

Campbell, J. W., Morgan-Kiss, R. M. and Cronan, J. E. (2003) A new *Escherichia coli* metabolic competency: growth on fatty acids by a novel anaerobic beta-oxidation pathway. *Mol Microbiol* **47**(3): 793–805.

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F. *et al.* (2007) Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol* **73**(1): 278–288.

Cassone, M., Giordano, A. and Pozzi, G. (2007) Bacterial DNA microarrays for clinical microbiology: the early logarithmic phase. *Front Biosci* **12**: 2658–2669.

Cenci, G., Caldini, G. and Strappini, C. (1998) Effect of different starches on *Escherichia coli* (S1) beta-glucuronidase expression. *J Basic Microbiol* **38**(2): 95–100.

Chain, P. S. G., Carniel, E., Larimer, F. W., Lamerdin, J., Stoutland, P. O. *et al.* (2004) Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci U S A* **101**(38): 13826–13831.

Chain, P. S. G., Hu, P., Malfatti, S. A., Radnedge, L., Larimer, F. *et al.* (2006) Complete genome sequence of *Yersinia pestis* strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen. *J Bacteriol* **188**(12): 4453–4463.

Chen, S. L., Hung, C. S., Xu, J., Reigstad, C. S., Magrini, V. *et al.* (2006) Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc Natl Acad Sci U S A* **103**(15): 5977–5982.

Cheong, Y. H., Chang, H.-S., Gupta, R., Wang, X., Zhu, T. *et al.* (2002) Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in *Arabidopsis*. *Plant Physiol* **129**(2): 661–677.

Chiu, C. H., Tang, P., Chu, C., Hu, S., Bao, Q. *et al.* (2005) The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. *Nucleic Acids Res* **33**(5): 1690–1698.

Choi, J. K., Yu, U., Kim, S. and Yoo, O. J. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19**: 84–90.

Chung, J.-L., Wang, W. and Bourne, P. E. (2006) Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins* **62**(3): 630–640.

Churchill, G. A. (1989) Stochastic models for heterogeneous DNA sequences. *Bull Math Biol* **51**(1): 79–94.

Clarridge, J. E. (2004) Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev* **17**(4): 840–62, table of contents.

Cleuziat, P. and Robert-Baudouy, J. (1990) Specific detection of *Escherichia coli* and *Shigella* species using fragments of genes coding for β-glucuronidase. *FEMS Microbiol Lett* **60**(3): 315–322.

Cohen, M. L. (1992) Epidemiology of drug resistance: implications for a post-antimicrobial era. *Science* **257**(5073): 1050–1055.

Coleman, A. W. (2007) Pan-eukaryote ITS2 homologies revealed by RNA secondary structure. *Nucleic Acids Res* **35**(10): 3322–3329.

Collazo, C. M., Zierler, M. K. and Galán, J. E. (1995) Functional analysis of the *Salmonella typhimurium* invasion genes *invl* and *invJ* and identification of a target of the protein secretion apparatus encoded in the *inv* locus. *Mol Microbiol* **15**(1): 25–38.

Conlon, E. M., Song, J. J. and Liu, J. S. (2006) Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics* **7**: 247.

Consortium, M. A. Q. C., Shi, L., Reid, L. H., Jones, W. D., Shippy, R. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**(9): 1151–1161.

Cornelis, G. R. (2002) *Yersinia* type III secretion: send in the effectors. *J Cell Biol* **158**(3): 401–408.

Craigon, D. J., James, N., Okyere, J., Higgins, J., Jotham, J. *et al.* (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res* **32**(Database issue): D575–D577.

Dangl, J. L. and Jones, J. D. (2001) Plant pathogens and integrated defence responses to infection. *Nature* **411**(6839): 826–833.

Darling, A. C. E., Mau, B., Blattner, F. R. and Perna, N. T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**(7): 1394–1403.

Deng, W., Burland, V., Plunkett, G., Boutin, A., Mayhew, G. F. *et al.* (2002) Genome sequence of *Yersinia pestis* KIM. *J Bacteriol* **184**(16): 4601–4611.

Deng, W., Liou, S.-R., Plunkett, G., Mayhew, G. F., Rose, D. J. *et al.* (2003) Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. *J Bacteriol* **185**(7): 2330–2337.

Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res* **33**(Database issue): D233–D237.

Diekema, D., BootsMiller, B., Vaughn, T., Woolson, R., Yankey, J. *et al.* (2004) Antimicrobial Resistance Trends and Outbreak Frequency in United States Hospitals. *Clinical Infectious Diseases* **38**(1): 78–85. PMID: 14679451.

DiRusso, C. C., Black, P. N. and Weimar, J. D. (1999) Molecular inroads into the regulation and metabolism of fatty acids, lessons from bacteria. *Prog Lipid Res* **38**(2): 129–197.

Dixon, R. A. (2001) Natural products and plant disease resistance. *Nature* **411**(6839): 843–847.

Dobrindt, U., Agerer, F., Michaelis, K., Janka, A., Buchrieser, C. *et al.* (2003) Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. *J Bacteriol* **185**(6): 1831–1840.

Dobrindt, U., Hochhut, B., Hentschel, U. and Hacker, J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* **2**(5): 414–424.

Doolittle, W. F. (1999) Phylogenetic classification and the universal tree. *Science* **284**(5423): 2124–2129.

Dorrell, N., Hinchliffe, S. J. and Wren, B. W. (2005) Comparative phylogenomics of pathogenic bacteria by microarray analysis. *Curr Opin Microbiol* **8**(5): 620–626.

Du, L. and Chen, Z. (2000) Identification of genes encoding receptor-like protein kinases as possible targets of pathogen- and salicylic acid-induced WRKY DNA-binding proteins in *Arabidopsis*. *Plant J* **24**(6): 837–847.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, UK.

Durfee, T., Nelson, R., Baldwin, S., Plunkett, G., Burland, V. *et al.* (2008) The complete genome sequence of *Escherichia coli* DH10B: insights into the biology of a laboratory workhorse. *J Bacteriol* **190**(7): 2597–2606.

Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics* **14**(9): 755–763.

Eddy, S. R. (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol* **4**(5): e1000069.

Edwards, T. A., Butterwick, J. A., Zeng, L., Gupta, Y. K., Wang, X. *et al.* (2005) Solution Structure of the Vts1 SAM Domain in the Presence of RNA. *J Mol Biol* **356**(5): 1065–1072.

Engelmann, J. C., Schwarz, R., Blenk, S., Friedrich, T., Seibel, P. N. *et al.* (2008) Unsupervised Meta-Analysis on Diverse Gene Expression Datasets Allows Insight into Gene Function and Regulation. *Bioinformatics and Biology Insights* **2**: 271–286.

Engelmann, J. C., Sven Rahmann, M. W., Schultz, J., Fritzilas, E., Kneitz, S. *et al.* (2009) Modelling cross-hybridization on phylogenetic DNA microarrays increases the detection power of closely related species. *Molecular Ecology Resources* **9**(1): 83–93.

Enright, A. J., Dongen, S. V. and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**(7): 1575–1584.

Eppinger, M., Rosovitz, M. J., Fricke, W. F., Rasko, D. A., Kokorina, G. *et al.* (2007) The complete genome sequence of *Yersinia pseudotuberculosis* IP31758, the causative agent of Far East scarlet-like fever. *PLoS Genet* **3**(8): e142.

Ernst, R. K., Guina, T. and Miller, S. I. (2001) *Salmonella typhimurium* outer membrane remodeling: role in resistance to host innate immunity. *Microbes Infect* **3**(14-15): 1327–1334.

Everitt, B. (2005) An R and S-PLUS Companion to Multivariate Analysis. Springer-Verlag London Limited.

Fariselli, P., Pazos, F., Valencia, A. and Casadio, R. (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem* **269**(5): 1356–1361.

Foultier, B., Troisfontaines, P., Müller, S., Opperdoes, F. R. and Cornelis, G. R. (2002) Characterization of the *ysa* pathogenicity locus in the chromosome of *Yersinia enterocolitica* and phylogeny analysis of type III secretion systems. *J Mol Evol* **55**(1): 37–51.

Fouts, D. E., Tyler, H. L., DeBoy, R. T., Daugherty, S., Ren, Q. *et al.* (2008) Complete genome sequence of the N_2-fixing broad host range endophyte *Klebsiella pneumoniae* 342 and virulence predictions verified in mice. *PLoS Genet* **4**(7): e1000141.

Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J. *et al.* (1980) The phylogeny of prokaryotes. *Science* **209**(4455): 457–463.

Fraley, C. and Raftery, A. E. (2002) Model-Based Clustering, Discriminant Analysis and Density Estimation. *Journal of the American Statistical Association* **97**: 611–631.

Fricke, W. F., Wright, M. S., Lindell, A. H., Harkins, D. M., Baker-Austin, C. *et al.* (2008) Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3-5. *J Bacteriol* **190**(20): 6779–6794.

Friedrich, T., Pils, B., Dandekar, T., Schultz, J. and Müller, T. (2006) Modelling interaction sites in protein domains with interaction profile hidden Markov models. *Bioinformatics* **22**(23): 2851–2857.

Frye, J. G., Fedorka-Cray, P. J., Jackson, C. R. and Rose, M. (2008) Analysis of *Salmonella enterica* with reduced susceptibility to the third-generation cephalosporin ceftriaxone isolated from U.S. cattle during 2000-2004. *Microb Drug Resist* **14**(4): 251–258.

Frye, J. G., Jesse, T., Long, F., Rondeau, G., Porwollik, S. *et al.* (2006) DNA microarray detection of antimicrobial resistance genes in diverse bacteria. *Int J Antimicrob Agents* **27**(2): 138–151.

Fu, Y. and Galán, J. E. (1998a) Identification of a specific chaperone for SptP, a substrate of the centisome 63 type III secretion system of *Salmonella typhimurium*. *J Bacteriol* **180**(13): 3393–3399.

Fu, Y. and Galán, J. E. (1998b) The *Salmonella typhimurium* tyrosine phosphatase SptP is translocated into host cells and disrupts the actin cytoskeleton. *Mol Microbiol* **27**(2): 359–368.

Ganduri, Y. L., Sadda, S. R., Datta, M. W., Jambukeswaran, R. K. and Datta, P. (1993) TdcA, a transcriptional activator of the *tdcABC* operon of *Escherichia coli*, is a member of the LysR family of proteins. *Mol Gen Genet* **240**(3): 395–402.

Gentleman, R. and Biocore (2008) geneplotter: Grapics related functions for Bioconductor. R package version 1.18.0.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**(10): R80.

Gentry, T. J., Wickham, G. S., Schadt, C. W., He, Z. and Zhou, J. (2006) Microarray applications in microbial ecology research. *Microb Ecol* **52**(2): 159–175.

Giamarellou, H. (2005) Multidrug resistance in Gram-negative bacteria that produce extended-spectrum beta-lactamases (ESBLs). *Clin Microbiol Infect* **11 Suppl 4**: 1–16.

Gong, S., Bearden, S. W., Geoffroy, V. A., Fetherston, J. D. and Perry, R. D. (2001) Characterization of the *Yersinia pestis* Yfu ABC inorganic iron transport system. *Infect Immun* **69**(5): 2829–2837.

Goss, T. J., Schweizer, H. P. and Datta, P. (1988) Molecular characterization of the *tdc* operon of *Escherichia coli* K-12. *J Bacteriol* **170**(11): 5352–5359.

Greenberg, C. R., Taylor, C. L., Haworth, J. C., Seargeant, L. E., Philipps, S. *et al.* (1993) A homoallelic Gly317-->Asp mutation in ALPL causes the perinatal (lethal) form of hypophosphatasia in Canadian mennonites. *Genomics* **17**(1): 215–217.

Grozdanov, L., Raasch, C., Schulze, J., Sonnenborn, U., Gottschalk, G. *et al.* (2004) Analysis of the genome structure of the nonpathogenic probiotic *Escherichia coli* strain Nissle 1917. *J Bacteriol* **186**(16): 5432–5441.

Gunn, J. S., Alpuche-Aranda, C. M., Loomis, W. P., Belden, W. J. and Miller, S. I. (1995) Characterization of the *Salmonella typhimurium pagC/pagD* chromosomal region. *J Bacteriol* **177**(17): 5040–5047.

Gupta, S. D., Wu, H. C. and Rick, P. D. (1997) A *Salmonella typhimurium* genetic locus which confers copper tolerance on copper-sensitive mutants of *Escherichia coli*. *J Bacteriol* **179**(16): 4977–4984.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning* **46**(1-3): 389–422.

Hacker, J., Blum-Oehler, G., Mühldorfer, I. and Tschäpe, H. (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol* **23**(6): 1089–1097.

Hacker, J. and Carniel, E. (2001) Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep* **2**(5): 376–381.

Haft, D. H., Selengut, J. D. and White, O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res* **31**(1): 371–373.

Hall, B. G. (2003) The EBG system of *E. coli*: origin and evolution of a novel β-galactosidase for the metabolism of lactose. *Genetica* **118**(2-3): 143–156.

Hall, M. (1999) Correlation-based Feature Selection for Machine Learning. Ph.D. thesis, Hamilton NZ: Waikato University, Department of Computer Science.

Haraga, A., Ohlson, M. B. and Miller, S. I. (2008) Salmonellae interplay with host cells. *Nat Rev Microbiol* **6**(1): 53–66.

Hardiman, G. (2004) Microarray platforms–comparisons and contrasts. *Pharmacogenomics* **5**(5): 487–502.

Haussler, D., Krogh, A., Mian, I. and Sjolander, K. (1993) Protein modeling using hidden Markov models: analysis of globins. In Proc. Proceeding of the Twenty-Sixth Hawaii International Conference on System Sciences, pages 792–802.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K. *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* **8**(1): 11–22.

Hazkani-Covo, E. and Graur, D. (2005) Evolutionary conservation of bacterial operons: does transcriptional connectivity matter? *Genetica* **124**(2-3): 145–166.

Hejnova, J., Dobrindt, U., Nemcova, R., Rusniok, C., Bomba, A. *et al.* (2005) Characterization of the flexible genome complement of the commensal *Escherichia coli* strain A0 34/86 (O83 : K24 : H31). *Microbiology* **151**(Pt 2): 385–398.

Hendlich, M., Bergner, A., Günther, J. and Klebe, G. (2003) Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol* **326**(2): 607–620.

Henikoff, S. and Henikoff, J. G. (1994) Position-based sequence weights. *J Mol Biol* **243**(4): 574–578.

Hermant, D., Ménard, R., Arricau, N., Parsot, C. and Popoff, M. Y. (1995) Functional conservation of the *Salmonella* and *Shigella* effectors of entry into epithelial cells. *Mol Microbiol* **17**(4): 781–789.

Hinchliffe, S. J., Isherwood, K. E., Stabler, R. A., Prentice, M. B., Rakin, A. *et al.* (2003) Application of DNA microarrays to study the evolutionary genomics of *Yersinia pestis* and *Yersinia pseudotuberculosis*. *Genome Res* **13**(9): 2018–2029.

Hothorn, T., Bretz, F. and Westfall, P. (2008) Simultaneous Inference in General Parametric Models. *Biometrical Journal* **50**(3): 346–363.

Hu, P., Greenwood, C. M. T. and Beyene, J. (2005) Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinformatics* **6**: 128.

Huang, J., Su, Z. and Xu, Y. (2005) The evolution of microbial phosphonate degradative pathways. *J Mol Evol* **61**(5): 682–690.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**: S96–104.

Hughey, R. and Krogh, A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput Appl Biosci* **12**(2): 95–107.

Hugouvieux-Cotte-Pattat, N., Blot, N. and Reverchon, S. (2001) Identification of TogMNAB, an ABC transporter which mediates the uptake of pectic oligomers in *Erwinia chrysanthemi* 3937. *Mol Microbiol* **41**(5): 1113–1123.

Huson, D. H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**(1): 68–73.

Huttenhower, C., Hibbs, M., Myers, C. and Troyanskaya, O. G. (2006) A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics* **22**(23): 2890–2897.

Hyle, E. P., Lipworth, A. D., Zaoutis, T. E., Nachamkin, I., Fishman, N. O. *et al.* (2005) Risk factors for increasing multidrug resistance among extended-spectrum beta-lactamase-producing *Escherichia coli* and *Klebsiella* species. *Clin Infect Dis* **40**(9): 1317–1324.

Ibrahim, A., Goebel, B. M., Liesack, W., Griffiths, M. and Stackebrandt, E. (1993) The phylogeny of the genus *Yersinia* based on 16S rDNA sequences. *FEMS Microbiol Lett* **114**(2): 173–177.

Iguchi, A., Thomson, N. R., Ogura, Y., Saunders, D., Ooka, T. *et al.* (2009) Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127:H6 strain E2348/69. *J Bacteriol* **191**(1): 347–354.

Ikeda, M., Arai, M., Okuno, T. and Shimizu, T. (2003) TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucleic Acids Res* **31**(1): 406–409.

Ikeda, M., Yamaguchi, N., Tani, K. and Nasu, M. (2005) Development of Phylogenetic Oligonucleotide Probes for Screening Foodborne Bacteria. *Journal of Health Science* **51**(4): 469–476.

Jenkins, C., Chart, H., Willshaw, G. A., Cheasty, T. and Smith, H. R. (2006) Genotyping of enteroaggregative *Escherichia coli* and identification of target genes for the detection of both typical and atypical strains. *Diagnostic Microbiology and Infectious Disease* **55**(1): 13 – 19.

Johnson, J. (1991) Virulence factors in *Escherichia coli* urinary tract infection. *Clinical Microbiology Reviews* **4**(1): 80–128.

Johnson, N. L., Kemp, A. W. and Kotz, S. (2005) Univariate Discrete Distributions. Wiley Series in Probability and Statistics. Wiley, third edn.

Johnson, T. J., Kariyawasam, S., Wannemuehler, Y., Mangiamele, P., Johnson, S. J. *et al.* (2007) The genome sequence of avian pathogenic *Escherichia coli* strain O1:K1:H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes. *J Bacteriol* **189**(8): 3228–3236.

Jones, S. and Thornton, J. M. (1997) Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol* **272**(1): 133–143.

Kakinuma, K., Fukushima, M. and Kawaguchi, R. (2003) Detection and identification of *Escherichia coli*, *Shigella*, and *Salmonella* by microarrays using the *gyrB* gene. *Biotechnol Bioeng* **83**(6): 721–728.

Käll, L., Krogh, A. and Sonnhammer, E. L. L. (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**(5): 1027–1036.

Kampstra, P. (2008) Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets* **28**(1): 1–9.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**(Database issue): D480–D484.

Kaper, J. B., Nataro, J. P. and Mobley, H. L. (2004) Pathogenic *Escherichia coli*. *Nat Rev Microbiol* **2**(2): 123–140.

Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A. (2004) kernlab - An S4 package for kernel methods in R. *Research Report Series / Department of Statistics and Mathematics* .

Kariyawasam, S., Scaccianoce, J. and Nolan, L. (2007) Common and specific genomic sequences of avian and human extraintestinal pathogenic *Escherichia coli* as determined by genomic subtractive hybridization. *BMC Microbiology* **7**(1): 81.

Katoh, K., ichi Kuma, K., Toh, H. and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**(2): 511–518.

Keller, A., Schleicher, T., Schultz, J., Müller, T., Dandekar, T. *et al.* (2009) 5.8S-28S rRNA interaction and HMM-based ITS2 annotation. *Gene* **430**(1-2): 50–57.

Khil, P. P. and Camerini-Otero, R. D. (2002) Over 1000 genes are involved in the DNA damage response of *Escherichia coli*. *Mol Microbiol* **44**(1): 89–105.

Kim, C. A. and Bowie, J. U. (2003) SAM domains: uniform structure, diversity of function. *Trends Biochem Sci* **28**(12): 625–628.

Kobayashi, N., Nishino, K. and Yamaguchi, A. (2001) Novel macrolide-specific ABC-type efflux transporter in *Escherichia coli*. *J Bacteriol* **183**(19): 5639–5644.

Koenker, R. (2008) quantreg: Quantile Regression. R package version 4.20.

Kohl, S., Marx, T., Giddings, I., Jägle, H., Jacobson, S. G. *et al.* (1998) Total colourblindness is caused by mutations in the gene encoding the alpha-subunit of the cone photoreceptor cGMP-gated cation channel. *Nat Genet* **19**(3): 257–259.

Koike, A. and Takagi, T. (2004) Prediction of protein-protein interaction sites using support vector machines. *Protein Eng Des Sel* **17**(2): 165–173.

Korczak, B., Frey, J., Schrenzel, J., Pluschke, G., Pfister, R. *et al.* (2005) Use of diagnostic microarrays for determination of virulence gene patterns of *Escherichia coli* K1, a major cause of neonatal meningitis. *J Clin Microbiol* **43**(3): 1024–1031.

Koster, W. (2005) Cytoplasmic membrane iron permease systems in the bacterial cell envelope. *Front Biosci* **10**: 462–477.

Kostić, T., Weilharter, A., Rubino, S., Delogu, G., Uzzau, S. *et al.* (2007) A microbial diagnostic microarray technique for the sensitive detection and identification of pathogenic bacteria in a background of nonpathogens. *Anal Biochem* **360**(2): 244–254.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994a) Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235**(5): 1501–1531.

Krogh, A., Mian, I. S. and Haussler, D. (1994b) A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res* **22**(22): 4768–4778.

Kruskal, W. H. and Wallis, W. A. (1952) Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association* **47**(260): 583–621.

Kulp, D., Haussler, D., Reese, M. G. and Eeckman, F. H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol* **4**: 134–142.

Kurokawa, H., Osawa, M., Kurihara, H., Katayama, N., Tokumitsu, H. *et al.* (2001) Target-induced conformational adaptation of calmodulin revealed by the crystal structure of a complex with nematode Ca^{2+}/calmodulin-dependent kinase kinase peptide. *J Mol Biol* **312**(1): 59–68.

Laikova, O. N., Mironov, A. A. and Gelfand, M. S. (2001) Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of Proteobacteria. *FEMS Microbiol Lett* **205**(2): 315–322.

Lamarche, M. G., Dozois, C. M., Daigle, F., Caza, M., Curtiss, R. *et al.* (2005) Inactivation of the *pst* system reduces the virulence of an avian pathogenic *Escherichia coli* O78 strain. *Infect Immun* **73**(7): 4138–4145.

Lan, R. and Reeves, P. R. (2002) *Escherichia coli* in disguise: molecular origins of *Shigella*. *Microbes Infect* **4**(11): 1125–1132.

Lautenbach, E., Strom, B. L., Bilker, W. B., Patel, J. B., Edelstein, P. H. *et al.* (2001) Epidemiological investigation of fluoroquinolone resistance in infections due to extended-spectrum beta-lactamase-producing *Escherichia coli* and *Klebsiella pneumoniae*. *Clin Infect Dis* **33**(8): 1288–1294.

Law, D. (2000) Virulence factors of *Escherichia coli* O157 and other Shiga toxin-producing *E. coli*. *Journal of Applied Microbiology* **88**(5): 729–745.

Lehner, A., Loy, A., Behr, T., Gaenge, H., Ludwig, W. *et al.* (2005) Oligonucleotide microarray for identification of *Enterococcus* species. *FEMS Microbiol Lett* **246**(1): 133–142.

Lehrke, M. and Lazar, M. A. (2005) The Many Faces of PPARγ. *Cell* **123**(6): 993–999.

Letowski, J., Brousseau, R. and Masson, L. (2004) Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays. *J Microbiol Methods* **57**(2): 269–278.

Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J. *et al.* (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* **34**(Database issue): D257–D260.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T. *et al.* (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res* **32**(Database issue): D142–D144.

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R. *et al.* (2008a) Evolution of mammals and their gut microbes. *Science* **320**(5883): 1647–1651.

Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. and Gordon, J. I. (2008b) Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* **6**(10): 776–788.

Li, L., Stoeckert, C. J. and Roos, D. S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**(9): 2178–2189.

Lichtarge, O., Bourne, H. R. and Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* **257**(2): 342–358.

Lillard, J. W., Fetherston, J. D., Pedersen, L., Pendrak, M. L. and Perry, R. D. (1997) Sequence and genetic analysis of the hemin storage (*hms*) system of *Yersinia pestis*. *Gene* **193**(1): 13–21.

Lin, S. L., Le, T. X. and Cowen, D. S. (2003) SptP, a *Salmonella typhimurium* type III-secreted protein, inhibits the mitogen-activated protein kinase pathway by inhibiting Raf activation. *Cell Microbiol* **5**(4): 267–275.

Lin, Y. C., Lu, C. L., Liu, Y.-C. and Tang, C. Y. (2006) SPRING: a tool for the analysis of genome rearrangement using reversals and block-interchanges. *Nucleic Acids Res* **34**(Web Server issue): W696–W699.

Liu, F., Chen, H., Galván, E. M., Lasaro, M. A. and Schifferli, D. M. (2006) Effects of Psa and F1 on the adhesive and invasive interactions of *Yersinia pestis* with human respiratory tract epithelial cells. *Infect Immun* **74**(10): 5636–5644.

Lloyd, A. L., Rasko, D. A. and Mobley, H. L. T. (2007) Defining genomic islands and uropathogen-specific genes in uropathogenic *Escherichia coli*. *J Bacteriol* **189**(9): 3532–3546.

Lomovskaya, O. and Lewis, K. (1992) Emr, an *Escherichia coli* locus for multidrug resistance. *Proc Natl Acad Sci U S A* **89**(19): 8938–8942.

Loy, A. and Bodrossy, L. (2006) Highly parallel microbial diagnostics using oligonucleotide microarrays. *Clin Chim Acta* **363**(1-2): 106–119.

Lu, C.-D. (2006) Pathways and regulation of bacterial arginine metabolism and perspectives for obtaining arginine overproducing strains. *Appl Microbiol Biotechnol* **70**(3): 261–272.

Lukashin, A. V. and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**(4): 1107–1115.

Madera, M. and Gough, J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res* **30**(19): 4321–4328.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E. *et al.* (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**(6): 3140–3145.

Mangone, M., MacMenamin, P., Zegar, C., Piano, F. and Gunsalus, K. (2008) UTRome.org: a platform for 3'UTR biology in *C. elegans*. *Nucleic Acids Research* **36**(Database issue): D57.

Marcus, S. L., Brumell, J. H., Pfeifer, C. G. and Finlay, B. B. (2000) *Salmonella* pathogenicity islands: big virulence in small packages. *Microbes Infect* **2**(2): 145–156.

Masuda, Y., Miyakawa, K., Nishimura, Y. and Ohtsubo, E. (1993) *chpA* and *chpB*, *Escherichia coli* chromosomal homologs of the *pem* locus responsible for stable maintenance of plasmid R100. *J Bacteriol* **175**(21): 6850–6856.

Mau, B., Glasner, J. D., Darling, A. E. and Perna, N. T. (2006) Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*. *Genome Biol* **7**(5): R44.

McClelland, M., Sanderson, K. E., Clifton, S. W., Latreille, P., Porwollik, S. *et al.* (2004) Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nat Genet* **36**(12): 1268–1274.

McClelland, M., Sanderson, K. E., Spieth, J., Clifton, S. W., Latreille, P. *et al.* (2001) Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature* **413**(6858): 852–856.

McPeake, S. J. W., Smyth, J. A. and Ball, H. J. (2005) Characterisation of avian pathogenic *Escherichia coli* (APEC) associated with colisepticaemia compared to faecal isolates from healthy birds. *Vet Microbiol* **110**(3-4): 245–253.

Melodelima, C., Gautier, C. and Piau, D. (2007) A markovian approach for the prediction of mouse isochores. *J Math Biol* **55**(3): 353–364.

Melodelima, C., Guéguen, L., Piau, D. and Gautier, C. (2006) A computational prediction of isochores based on hidden Markov models. *Gene* **385**: 41–49.

Milburn, D., Laskowski, R. A. and Thornton, J. M. (1998) Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Eng* **11**(10): 855–859.

Moreau, Y., Aerts, S., Moor, B. D., Strooper, B. D. and Dabrowski, M. (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet* **19**(10): 570–577.

Mothershed, E. A. and Whitney, A. M. (2006) Nucleic acid-based methods for the detection of bacterial pathogens: present and future considerations for the clinical laboratory. *Clin Chim Acta* **363**(1-2): 206–220.

Munch, K. and Krogh, A. (2006) Automatic generation of gene finders for eukaryotic species. *BMC Bioinformatics* **7**: 263.

Nataro, J. P. and Kaper, J. B. (1998) Diarrheagenic *Escherichia coli*. *Clin Microbiol Rev* **11**(1): 142–201.

Nataro, J. P., Seriwatana, J., Fasano, A., Maneval, D. R., Guers, L. D. *et al.* (1995) Identification and cloning of a novel plasmid-encoded enterotoxin of enteroinvasive *Escherichia coli* and *Shigella strains*. *Infect Immun* **63**(12): 4721–4728.

Navarre, W. W., Halsey, T. A., Walthers, D., Frye, J., McClelland, M. *et al.* (2005) Co-regulation of *Salmonella enterica* genes required for virulence and resistance to antimicrobial peptides by SlyA and PhoP/PhoQ. *Mol Microbiol* **56**(2): 492–508.

Nelson, K. M., Young, G. M. and Miller, V. L. (2001) Identification of a locus involved in systemic dissemination of *Yersinia enterocolitica*. *Infect Immun* **69**(10): 6201–6208.

Ng, A., Jordan, M. and Weiss, Y. (2001) On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems* **14**.

Nie, H., Yang, F., Zhang, X., Yang, J., Chen, L. *et al.* (2006) Complete genome sequence of *Shigella flexneri* 5b and comparison with *Shigella flexneri* 2a. *BMC Genomics* **7**: 173.

Nolte, R. T., Wisely, G. B., Westin, S., Cobb, J. E., Lambert, M. H. *et al.* (1998) Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor-gamma. *Nature* **395**(6698): 137–143.

Nudleman, E. and Kaiser, D. (2004) Pulling together with type IV pili. *J Mol Microbiol Biotechnol* **7**(1-2): 52–62.

Oelschlaeger, T. A., Dobrindt, U. and Hacker, J. (2002) Pathogenicity islands of uropathogenic *E. coli* and the evolution of virulence. *Int J Antimicrob Agents* **19**(6): 517–521.

Ofran, Y. and Rost, B. (2003) Predicted protein-protein interaction sites from local sequence information. *FEBS Lett* **544**(1-3): 236–239.

Ogawa, M., Handa, Y., Ashida, H., Suzuki, M. and Sasakawa, C. (2008) The versatility of *Shigella* effectors. *Nat Rev Microbiol* **6**(1): 11–16.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Simpson, G. L. *et al.* (2008) vegan: Community Ecology Package. R package version 1.13-2.

Olekhnovich, I. N. and Kadner, R. J. (2002) DNA-binding activities of the HilC and HilD virulence regulatory proteins of *Salmonella enterica* serovar Typhimurium. *J Bacteriol* **184**(15): 4148–4160.

Orth, D., Grif, K., Dierich, M. P. and Würzner, R. (2007) Variability in tellurite resistance and the *ter* gene cluster among Shiga toxin-producing *Escherichia coli* isolated from humans, animals and food. *Res Microbiol* **158**(2): 105–111.

Oshima, K., Toh, H., Ogura, Y., Sasamoto, H., Morita, H. *et al.* (2008) Complete genome sequence and comparative analysis of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult. *DNA Res* **15**(6): 375–386.

Park, H. G., Song, J. Y., Park, K. H. and Kim, M. H. (2006) Fluorescence-based assay formats and signal amplification strategies for DNA microarray analysis. *Chemical Engineering Science* **61**(3): 954 – 965. Biomolecular Engineering.

Parkhill, J., Dougan, G., James, K. D., Thomson, N. R., Pickard, D. *et al.* (2001a) Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413**(6858): 848–852.

Parkhill, J., Wren, B. W., Thomson, N. R., Titball, R. W., Holden, M. T. *et al.* (2001b) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**(6855): 523–527.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R. *et al.* (2007) ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* **35**(Database issue): D747–D750.

Paterson, D. L., Hujer, K. M., Hujer, A. M., Yeiser, B., Bonomo, M. D. *et al.* (2003) Extended-spectrum beta-lactamases in *Klebsiella pneumoniae* bloodstream isolates from seven countries: dominance and widespread prevalence of SHV- and CTX-M-type beta-lactamases. *Antimicrob Agents Chemother* **47**(11): 3554–3560.

Pattery, T., Hernalsteens, J. P. and Greve, H. D. (1999) Identification and molecular characterization of a novel *Salmonella enteritidis* pathogenicity islet encoding an ABC transporter. *Mol Microbiol* **33**(4): 791–805.

Paulsen, I. T., Chen, J., Nelson, K. E. and Saier, M. H. (2001) Comparative genomics of microbial drug efflux systems. *J Mol Microbiol Biotechnol* **3**(2): 145–150.

Pearson, K. (1902) On the systematic fitting of curves to observations and measurements. *Biometrika* **1**(3): 265–303.

Pellicer, M. T., Badía, J., Aguilar, J. and Baldomà, L. (1996) *glc* locus of *Escherichia coli*: characterization of genes encoding the subunits of glycolate oxidase and the *glc* regulator protein. *J Bacteriol* **178**(7): 2051–2059.

Pelludat, C., Prager, R., Tschäpe, H., Rabsch, W., Schuchhardt, J. *et al.* (2005) Pilot study to evaluate microarray hybridization as a tool for *Salmonella enterica* serovar Typhimurium strain differentiation. *J Clin Microbiol* **43**(8): 4092–4106.

Peng, K., Obradovic, Z. and Vucetic, S. (2004) Exploring bias in the Protein Data Bank using contrast classifiers. *Pac Symp Biocomput* pages 435–446.

Perna, N. T., Plunkett, G., Burland, V., Mau, B., Glasner, J. D. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**(6819): 529–533.

Pils, B., Copley, R. and Schultz, J. (2005) Variation in structural location and amino acid conservation of functional sites in protein domain families. *BMC Bioinformatics* **6**(1): 210.

Pineiro, J., Bates, D., DebRoy, S., Sarkar, D. and the R Core team (2008) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-88.

Pinheiro, V. B. and Ellar, D. J. (2007) Expression and insecticidal activity of *Yersinia pseudotuberculosis* and *Photorhabdus luminescens* toxin complex proteins. *Cell Microbiol* **9**(10): 2372–2380.

Piroux, N., Saunders, K., Page, A. and Stanley, J. (2007) Geminivirus pathogenicity protein C4 interacts with *Arabidopsis thaliana* shaggy-related protein kinase AtSKeta, a component of the brassinosteroid signalling pathway. *Virology* **362**(2): 428–440.

Porwollik, S., Boyd, E. F., Choy, C., Cheng, P., Florea, L. *et al.* (2004) Characterization of *Salmonella enterica* subspecies I genovars by use of microarrays. *J Bacteriol* **186**(17): 5883–5898.

Posada, D. and Crandall, K. A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**(9): 817–818.

Pérez, A. D. G., González, E. G., Angarica, V. E., Vasconcelos, A. T. R. and Collado-Vides, J. (2008) Impact of Transcription Units rearrangement on the evolution of the regulatory network of gamma-proteobacteria. *BMC Genomics* **9**: 128.

Price-Carter, M., Tingey, J., Bobik, T. A. and Roth, J. R. (2001) The alternative electron acceptor tetrathionate supports B_{12}-dependent anaerobic growth of *Salmonella enterica* serovar Typhimurium on ethanolamine or 1,2-propanediol. *J Bacteriol* **183**(8): 2463–2475.

Pritsker, M., Liu, Y.-C., Beer, M. A. and Tavazoie, S. (2004) Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res* **14**(1): 99–108.

Qian, B. and Goldstein, R. A. (2001) Distribution of Indel lengths. *Proteins* **45**(1): 102–104.

R Development Core Team (2004) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc of the IEEE* **77**(2): 257–286.

Rahmann, S. (2002) Rapid large-scale oligonucleotide selection for microarrays. *Proc IEEE Comput Soc Bioinform Conf* **1**: 54–63.

Rahmann, S. (2003) Fast large scale oligonucleotide selection using the longest common factor approach. *J Bioinform Comput Biol* **1**(2): 343–361.

Ravcheev, D. A., Gerasimova, A. V., Mironov, A. A. and Gelfand, M. S. (2007) Comparative genomic analysis of regulation of anaerobic respiration in ten genomes from three families of γ-proteobacteria (*Enterobacteriaceae*, *Pasteurellaceae*, *Vibrionaceae*). *BMC Genomics* **8**: 54.

Reid, S., Herbelin, C., Bumbaugh, A., Selander, R. and Whittam, T. (2000) Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406**(6791): 64–67.

Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. and Gelfand, M. S. (2003) Comparative genomics of the vitamin B_{12} metabolism and regulation in prokaryotes. *J Biol Chem* **278**(42): 41148–41159.

Rodriguez-Siek, K. E., Giddings, C. W., Doetkott, C., Johnson, T. J., Fakhr, M. K. *et al.* (2005) Comparison of *Escherichia coli* isolates implicated in human urinary tract infection and avian colibacillosis. *Microbiology* **151**(Pt 6): 2097–2110.

Rossi, M. S., Fetherston, J. D., Létoffé, S., Carniel, E., Perry, R. D. *et al.* (2001) Identification and characterization of the hemophore-dependent heme acquisition system of *Yersinia pestis*. *Infect Immun* **69**(11): 6707–6717.

Ryals, J. A., Neuenschwander, U. H., Willits, M. G., Molina, A., Steiner, H. Y. *et al.* (1996) Systemic Acquired Resistance. *Plant Cell* **8**(10): 1809–1819.

Saebø, P. E., Andersen, S. M., Myrseth, J., Laerdahl, J. K. and Rognes, T. (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res* **33**(Web Server issue): W535–W539.

Saitoh, M., Tanaka, K., Nishimori, K., ichi Makino, S., Kanno, T. *et al.* (2005) The *artAB* genes encode a putative ADP-ribosyltransferase toxin homologue associated with *Salmonella enterica* serovar Typhimurium DT104. *Microbiology* **151**(Pt 9): 3089–3096.

Sandkvist, M. (2001) Type II secretion and pathogenesis. *Infect Immun* **69**(6): 3523–3535.

Santiviago, C. A., Fuentes, J. A., Bueno, S. M., Trombert, A. N., Hildago, A. A. *et al.* (2002) The *Salmonella enterica* sv. Typhimurium *smvA*, *yddG* and *ompD* (porin) genes are required for the efficient efflux of methyl viologen. *Mol Microbiol* **46**(3): 687–698.

Schliep, A., Schönhuth, A. and Steinhoff, C. (2003) Using hidden Markov models to analyze gene expression time course data. *Bioinformatics* **19 Suppl 1**: i255–i263.

Schölkopf, B., Smola, A. and Müller, K.-R. (1998) Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* **10**(5): 1299–1319.

Schubert, S., Fischer, D. and Heesemann, J. (1999) Ferric enterochelin transport in *Yersinia enterocolitica*: molecular and evolutionary aspects. *J Bacteriol* **181**(20): 6387–6395.

Schultz, J., Milpetz, F., Bork, P. and Ponting, C. P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* **95**(11): 5857–5864.

Schultz, J., Müller, T., Achtziger, M., Seibel, P. N., Dandekar, T. *et al.* (2006) The internal transcribed spacer 2 database–a web server for (not only) low level phylogenetic analyses. *Nucleic Acids Res* **34**(Web Server issue): W704–W707.

Schultz, J., Ponting, C. P., Hofmann, K. and Bork, P. (1997) SAM as a protein interaction domain involved in developmental regulation. *Protein Sci* **6**(1): 249–253.

Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistic* **6**(2): 461–464.

Sebbane, F., Devalckenaere, A., Foulon, J., Carniel, E. and Simonet, M. (2001) Silencing and reactivation of urease in *Yersinia pestis* is determined by one G residue at a specific position in the *ureD* gene. *Infect Immun* **69**(1): 170–176.

Shah, P. and Swiatlo, E. (2008) A multifaceted role for polyamines in bacterial pathogens. *Mol Microbiol* **68**(1): 4–16.

Shawe-Taylor, J. and Cristianini, N. (2004) Kernel Methods for Pattern Analysis. Cambridge University Press.

Shen, B. (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Curr Opin Chem Biol* **7**(2): 285–295.

van der Sloot, A. M., Tur, V., Szegezdi, E., Mullally, M. M., Cool, R. H. *et al.* (2006) Designed tumor necrosis factor-related apoptosis-inducing ligand variants initiating apoptosis exclusively via the DR5 receptor. *Proc Natl Acad Sci U S A* **103**(23): 8634–8639.

Smyth, G. K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**: Article 3.

Sánchez, L., Pan, W., Viñas, M. and Nikaido, H. (1997) The *acrAB* homolog of *Haemophilus influenzae* codes for a functional multidrug efflux pump. *J Bacteriol* **179**(21): 6855–6857.

Snyder, K. A., Feldman, H. J., Dumontier, M., Salama, J. J. and Hogue, C. W. V. (2006) Domain-based small molecule binding site annotation. *BMC Bioinformatics* **7**: 152.

Song, Y., Tong, Z., Wang, J., Wang, L., Guo, Z. *et al.* (2004) Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. *DNA Res* **11**(3): 179–197.

Sonnhammer, E. L., von Heijne, G. and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6**: 175–182.

Soutourina, O. A. and Bertin, P. N. (2003) Regulation cascade of flagellar expression in Gram-negative bacteria. *FEMS Microbiol Rev* **27**(4): 505–523.

Stepanova, A. N., Hoyt, J. M., Hamilton, A. A. and Alonso, J. M. (2005) A Link between ethylene and auxin uncovered by the characterization of two root-specific ethylene-insensitive mutants in *Arabidopsis*. *Plant Cell* **17**(8): 2230–2242.

Stoughton, R. B. (2005) Applications of DNA microarrays in biology. *Annu Rev Biochem* **74**: 53–82.

Suzuki, R. and Shimodaira, H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**(12): 1540–1542.

Taillandier, A., Cozien, E., Muller, F., Merrien, Y., Bonnin, E. *et al.* (2000) Fifteen new mutations (-195C>T, L-12X, 298-2A>G, T117N, A159T, R229S, 997+2T>A, E274X, A331T, H364R, D389G, 1256delC, R433H, N461I, C472S) in the tissue-nonspecific alkaline phosphatase (TNSALP) gene in patients with hypophosphatasia. *Hum Mutat* **15**(3): 293.

Taillandier, A., Lia-Baldini, A. S., Mouchard, M., Robin, B., Muller, F. *et al.* (2001) Twelve novel mutations in the tissue-nonspecific alkaline phosphatase gene (ALPL) in patients with various forms of hypophosphatasia. *Hum Mutat* **18**(1): 83–84.

Taillandier, A., Zurutuza, L., Muller, F., Simon-Bouy, B., Serre, J. L. *et al.* (1999) Characterization of eleven novel mutations (M45L, R119H, 544delG, G145V, H154Y, C184Y, D289V, 862+5A, 1172delC, R411X, E459K) in the tissue-nonspecific alkaline phosphatase (TNSALP) gene in patients with severe hypophosphatasia. Mutations in brief no. 217. Online. *Hum Mutat* **13**(2): 171–172.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* **96**(6): 2907–2912.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997) A genomic perspective on protein families. *Science* **278**(5338): 631–637.

Tchieu, J. H., Norris, V., Edwards, J. S. and Saier, M. H. (2001) The complete phosphotranferase system in *Escherichia coli*. *J Mol Microbiol Biotechnol* **3**(3): 329–346.

Tekaia, F., Lazcano, A. and Dujon, B. (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Res* **9**(6): 550–557.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**(39): 13950–13955.

Thanos, C. D., Goodwill, K. E. and Bowie, J. U. (1999) Oligomeric structure of the human EphB2 receptor SAM domain. *Science* **283**(5403): 833–836.

Thilmony, R., Underwood, W. and He, S. Y. (2006) Genome-wide transcriptional analysis of the *Arabidopsis thaliana* interaction with the plant pathogen *Pseudomonas syringae* pv. *tomato* DC3000 and the human pathogen *Escherichia coli* O157:H7. *Plant J* **46**(1): 34–53.

Thomson, N. R., Clayton, D. J., Windhorst, D., Vernikos, G., Davidson, S. *et al.* (2008) Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Res* **18**(10): 1624–1637.

Thomson, N. R., Howard, S., Wren, B. W., Holden, M. T. G., Crossman, L. *et al.* (2006) The complete genome sequence and comparative genome analysis of the high pathogenicity *Yersinia enterocolitica* strain 8081. *PLoS Genet* **2**(12): e206.

Tomich, M., Planet, P. J. and Figurski, D. H. (2007) The *tad* locus: postcards from the widespread colonization island. *Nat Rev Microbiol* **5**(5): 363–375.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S. *et al.* (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* **5**(1): e1000344.

Townsend, S. M., Kramer, N. E., Edwards, R., Baker, S., Hamlin, N. *et al.* (2001) *Salmonella enterica* serovar Typhi possesses a unique repertoire of fimbrial gene sequences. *Infect Immun* **69**(5): 2894–2901.

Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K. *et al.* (2005) Comparative metagenomics of microbial communities. *Science* **308**(5721): 554–557.

Usadel, B., Nagel, A., Thimm, O., Redestig, H., Blaesing, O. E. *et al.* (2005) Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. *Plant Physiol* **138**(3): 1195–1204.

Venables, W. N. and Ripley, B. D. (2002) MASS: Modern Applied Statistics with S. Springer, New York, fourth edn. ISBN 0-387-95457-0.

Venkatasubbarao, S. (2004) Microarrays–status and prospects. *Trends Biotechnol* **22**(12): 630–637.

Venkatesan, M. M., Buysse, J. M. and Kopecko, D. J. (1989) Use of *Shigella flexneri ipaC* and *ipaH* gene sequences for the general identification of *Shigella* spp. and enteroinvasive *Escherichia coli*. *J Clin Microbiol* **27**(12): 2687–2691.

Viterbi, A. J. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* **13**(2): 260 – 269.

Wagner, M., Smidt, H., Loy, A. and Zhou, J. (2007) Unravelling microbial communities with DNA-microarrays: challenges and future directions. *Microb Ecol* **53**(3): 498–506.

Wallis, T. S. and Galyov, E. E. (2000) Molecular basis of *Salmonella*-induced enteritis. *Mol Microbiol* **36**(5): 997–1005.

Walsh, C. (2000) Molecular mechanisms that confer antibacterial drug resistance. *Nature* **406**(6797): 775–781.

Wang, X., L.Cooper, K., Wang, A., Xu, J., Wang, Z. *et al.* (2006) Label-free DNA sequence detection using oligonucleotide functionalized optical fiber. *Applied Physics Letters* **89**(16): pp.163901.

Wang, X. and Seed, B. (2003) Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics* **19**(7): 796–802.

Ward, J. H. (1963) Hierarchical Grouping To Optimize An Objective Function. *Journal Of The American Statistical Association* **58**(301): 236–244.

Wei, J., Goldberg, M. B., Burland, V., Venkatesan, M. M., Deng, W. *et al.* (2003) Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect Immun* **71**(5): 2775–2786.

Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P. *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A* **99**(26): 17020–17024.

Welch, T. J., Fricke, W. F., McDermott, P. F., White, D. G., Rosso, M.-L. *et al.* (2007) Multiple antimicrobial resistance in plague: an emerging public health risk. *PLoS ONE* **2**(3): e309.

Whelan, K. F., Colleran, E. and Taylor, D. E. (1995) Phage inhibition, colicin resistance, and tellurite resistance are encoded by a single cluster of genes on the IncHI2 plasmid R478. *J Bacteriol* **177**(17): 5016–5027.

Willenbrock, H., Hallin, P., Wassenaar, T. and Ussery, D. (2007) Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. *Genome Biol* **8**(12): R267.

Willenbrock, H., Petersen, A., Sekse, C., Kiil, K., Wasteson, Y. *et al.* (2006) Design of a seven-genome *Escherichia coli* microarray for comparative genomic profiling. *J Bacteriol* **188**(22): 7713–7721.

Wilson, L. A. and Sharp, P. M. (2006) Enterobacterial repetitive intergenic consensus (ERIC) sequences in *Escherichia coli*: Evolution and implications for ERIC-PCR. *Mol Biol Evol* **23**(6): 1156–1168.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P. *et al.* (2006) Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* **60**(5): 1136–1151.

Wissenbach, U., Six, S., Bongaerts, J., Ternes, D., Steinwachs, S. *et al.* (1995) A third periplasmic transport system for L-arginine in *Escherichia coli*: molecular characterization of the *artPIQMJ* genes, arginine binding and transport. *Mol Microbiol* **17**(4): 675–686.

Woehlke, G. and Dimroth, P. (1994) Anaerobic growth of *Salmonella typhimurium* on L(+)- and D(-)-tartrate involves an oxaloacetate decarboxylase Na^+ pump. *Arch Microbiol* **162**(4): 233–237.

Woehlke, G., Wifling, K. and Dimroth, P. (1992) Sequence of the sodium ion pump oxaloacetate decarboxylase from *Salmonella typhimurium*. *J Biol Chem* **267**(32): 22798–22803.

Wood, M. W., Jones, M. A., Watson, P. R., Hedges, S., Wallis, T. S. *et al.* (1998) Identification of a pathogenicity island required for *Salmonella* enteropathogenicity. *Mol Microbiol* **29**(3): 883–891.

Wren, B. W. (2003) The yersiniae–a model genus to study the rapid evolution of bacterial pathogens. *Nat Rev Microbiol* **1**(1): 55–64.

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* **34**(Database issue): D187–D191.

Wu, J., Irizarry, R. and with contributions from James MacDonald and Jeff Gentry (2005) gcrma: Background Adjustment Using Sequence Information. R package version 2.2.1.

Xi, H., Schneider, B. L. and Reitzer, L. (2000) Purine catabolism in *Escherichia coli* and function of xanthine dehydrogenase in purine salvage. *J Bacteriol* **182**(19): 5332–5341.

Yada, T., Nakao, M., Totoki, Y. and Nakai, K. (1999) Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics* **15**(12): 987–993.

Yamamoto, E. and Yanagimoto, T. (1992) Moment estimators for the beta-binomial distribution. *Journal of Applied Statistics* **19**(2): 273–283.

Yamanaka, K., Fang, L. and Inouye, M. (1998) The CspA family in *Escherichia coli*: multiple gene duplication for stress adaptation. *Mol Microbiol* **27**(2): 247–255.

Yang, F., Yang, J., Zhang, X., Chen, L., Jiang, Y. *et al.* (2005) Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res* **33**(19): 6445–6458.

Yoo, S. M., Lee, S. Y., Chang, K. H., Yoo, S. Y., Yoo, N. C. *et al.* (2009) High-throughput identification of clinically important bacterial pathogens using DNA microarray. *Molecular and Cellular Probes* **23**(3-4): 171 – 177.

Zdziarski, J., Svanborg, C., Wullt, B., Hacker, J. and Dobrindt, U. (2008) Molecular basis of commensalism in the urinary tract: low virulence or virulence attenuation? *Infect Immun* **76**(2): 695–703.

Zhang, X., Lu, X., Shi, Q., Xu, X.-Q., Leung, H.-C. E. *et al.* (2006) Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics* **7**: 197.

Zhou, D., Han, Y., Song, Y., Tong, Z., Wang, J. *et al.* (2004) DNA microarray analysis of genome dynamics in *Yersinia pestis*: insights into bacterial genome microevolution and niche adaptation. *J Bacteriol* **186**(15): 5138–5146.

Zhou, H.-X. and Qin, S. (2007) Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics* **23**(17): 2203–2209.

Zhou, H. X. and Shan, Y. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* **44**(3): 336–343.

Zurutuza, L., Muller, F., Gibrat, J. F., Taillandier, A., Simon-Bouy, B. *et al.* (1999) Correlations of genotype and phenotype in hypophosphatasia. *Hum Mol Genet* **8**(6): 1039–1046.

Part IV.

Appendix

A. Supplementary data on comparative enterobacterial genomics

A.1. Proteome comparison

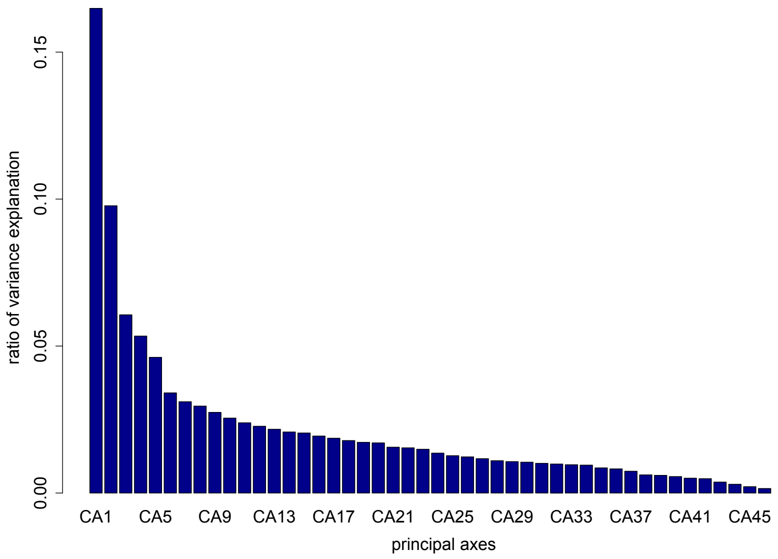


Figure A.1.: Variance explanation of principal axes in the correspondence analysis of proteome mapping data from enterobacteria. The values refer to the ratio of variance in the common data space of enterobacterial strains and protein families that is covered by respective principal axes.

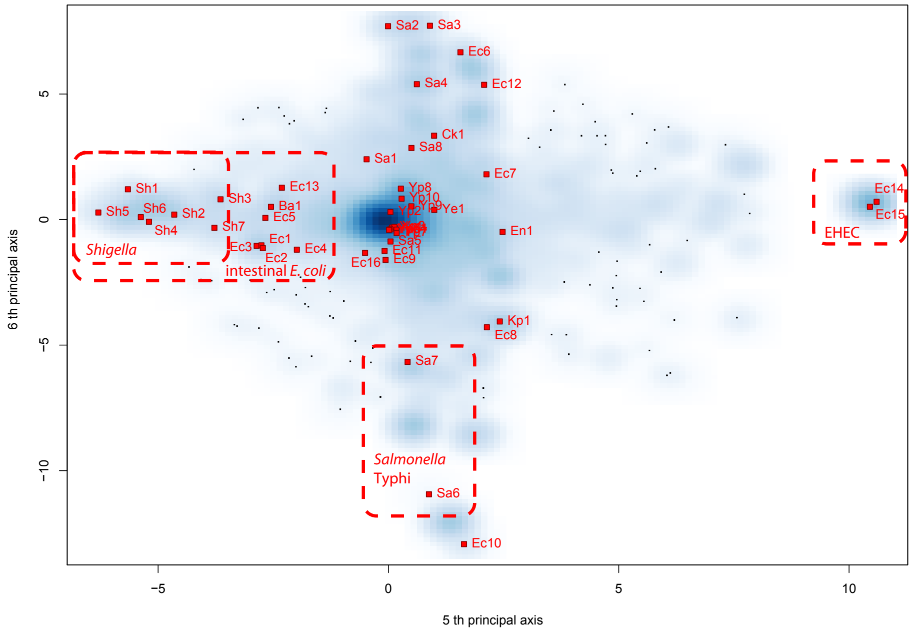


Figure A.2.: Association graph of enterobacterial protein clusters and the strains, in which they occur. The distribution of protein clusters is indicated by the intensity of blue colour (dark blue = high density of protein clusters).

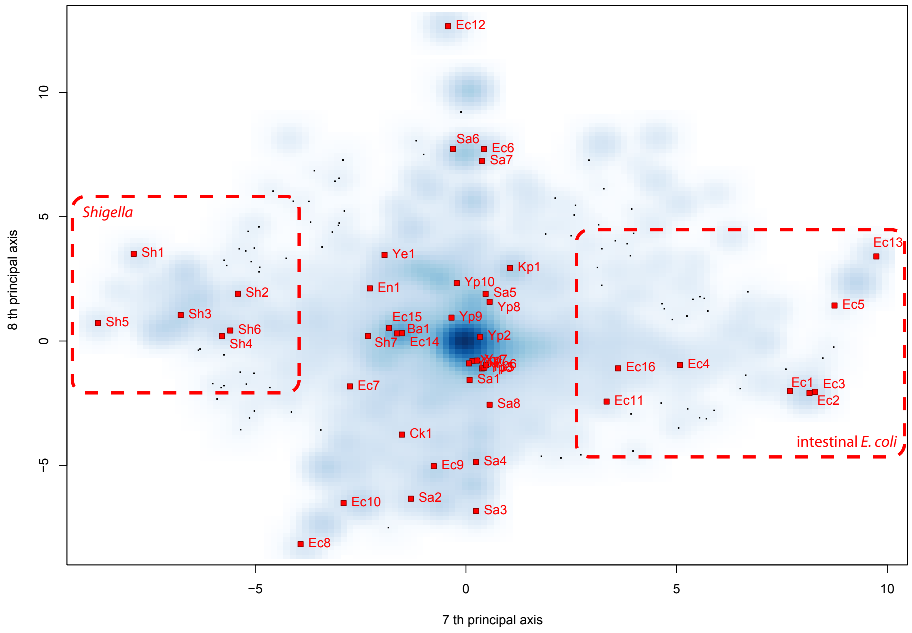


Figure A.3.: Association graph of enterobacterial protein clusters and the strains, in which they occur. The distribution of protein clusters is indicated by the intensity of blue colour (dark blue = high density of protein clusters).

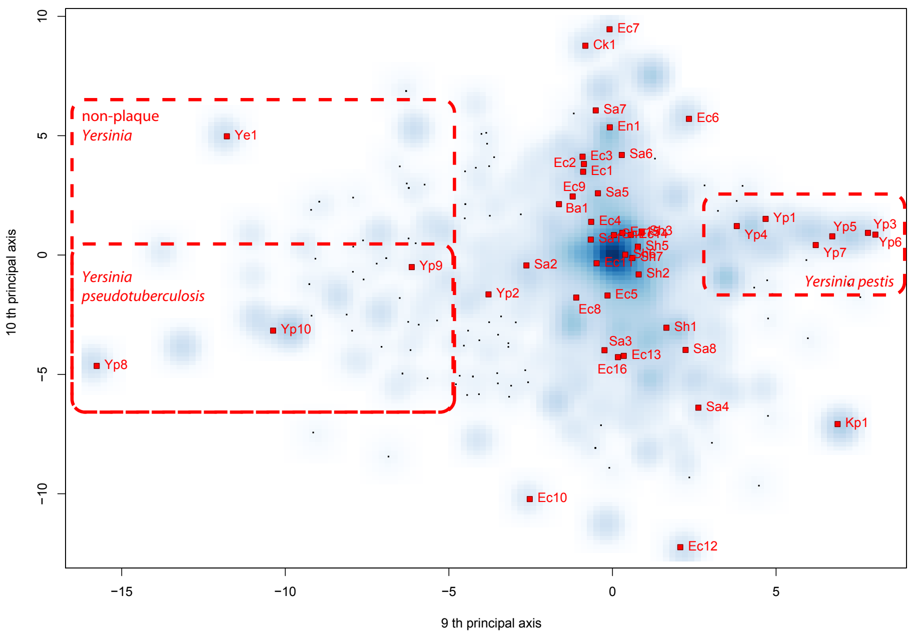


Figure A.4.: Association graph of enterobacterial protein clusters and the strains, in which they occur. The distribution of protein clusters is indicated by the intensity of blue colour (dark blue = high density of protein clusters).

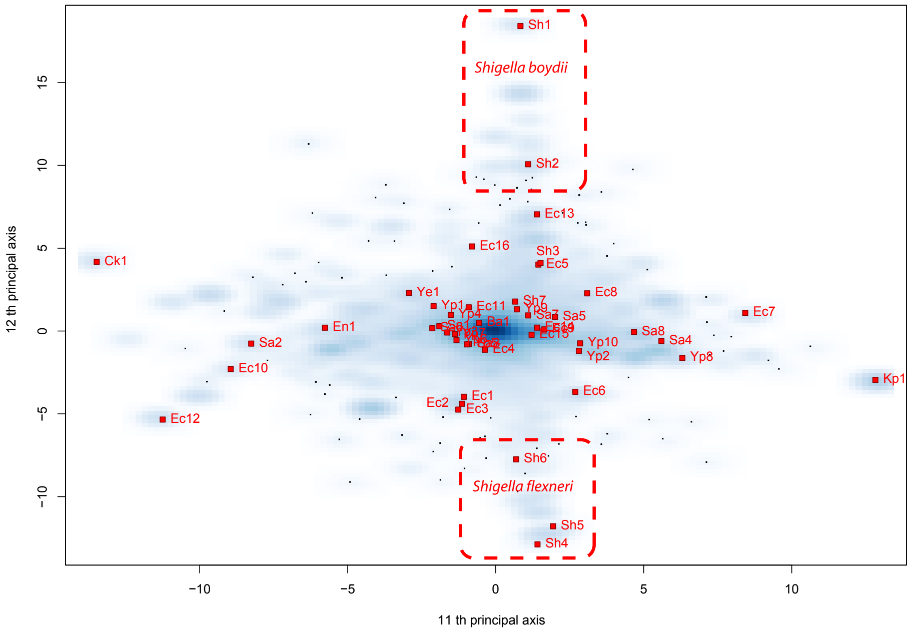


Figure A.5.: Association graph of enterobacterial protein clusters and the strains, in which they occur. The distribution of protein clusters is indicated by the intensity of blue colour (dark blue = high density of protein clusters).

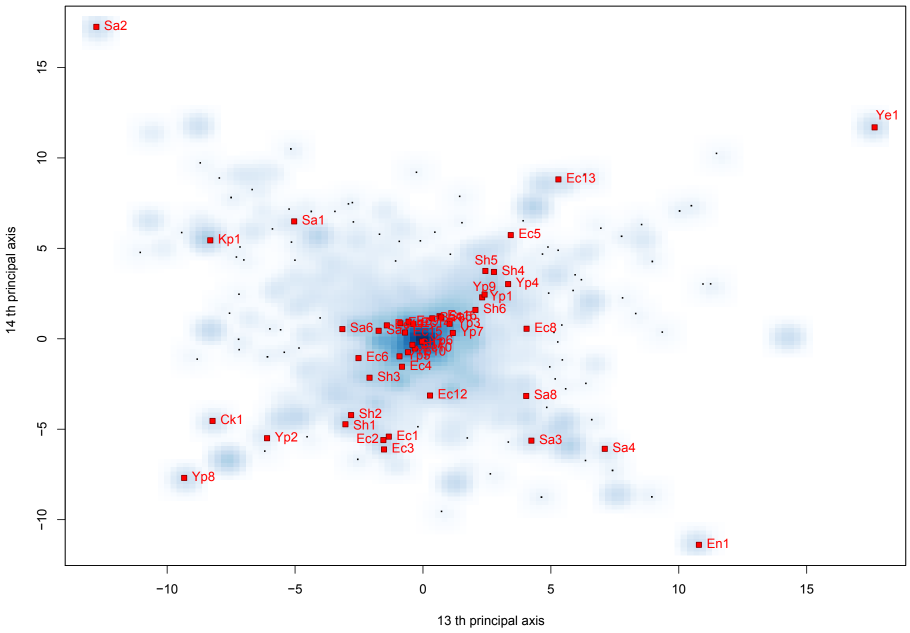


Figure A.6.: Association graph of enterobacterial protein clusters and the strains, in which they occur. The distribution of protein clusters is indicated by the intensity of blue colour (dark blue = high density of protein clusters).

A.2. Virulence structures in *E. coli*

Table A.1.: Abundance of virulence structures in selected *E. coli* strains

Category	Vir. structure	*Escherichia coli* strains										
		non-pathogens			UPEC			EHEC		EPEC	EAEC	MNEC
		MG1655	W3110	Nissle	CFT073	UTI89	536	Sakai	EDL933	E2348-69	O42	IHE3034
Toxins												
	α-Hemolysin	0	0	0	1	1	2	0	0	0	0	0
	Hemolysin-C	0	0	0	2	1	2	0	0	0	0	0
	Hemolysin-E	1	1	0	0	0	0	1	1	0	1	0
	CNF	0	0	0	0	1	0	0	0	0	0	0
	CDT	0	0	0	0	0	0	0	0	0	0	2
	hs-enterotoxin	0	0	0	0	0	0	0	0	0	0	0
	hl-enterotoxin	0	0	0	0	0	0	0	0	0	0	0
	Ecotin	1	1	1	1	1	1	1	1	1	1	1
	Sm. Mul. Drug Res.	4	4	4	4	6	6	4	4	3	3	4
	Big_1	1	1	2	2	2	2	5	4	2	3	2
	Big_2	0	0	2	2	3	1	7	6	3	2	5
	Enterotoxin ShET2	1	2	1	1	1	1	4	4	2	3	0
	CcdB	0	0	1	1	0	1	1	1	2	1	0
	Ail_Lom	2	2	1	3	3	1	12	12	5	5	7
	SLT β	0	0	0	0	0	0	2	2	0	0	0
	InvE	0	0	0	0	0	0	1	1	0	1	0
	Type III sec. proteins	0	0	0	0	0	0	1	1	1	0	0
Secretion Systems												
	ABC-Transporter (fam 1)	64	78	90	90	88	88	84	84	79	82	87
	ABC-Transporter (fam 2)	4	4	5	5	5	5	4	4	4	5	5
	ABC-Transporter (fam 3)	1	1	3	3	3	4	1	1	1	3	3
	Sec-System	1	1	1	1	1	1	1	1	1	1	1
	Type-II	1	1	1	1	1	1	1	1	1	1	1
	Type-III	0	0	0	0	0	0	1	1	1	0	0
	Type-IV	0	0	0	0	0	0	0	0	0	0	0
	HlyD	13	13	15	16	15	16	15	15	14	14	14
	Autotransporter (Type-V)	7	9	11	9	5	7	9	9	4	12	6
Fimbriae												
	CS1-like pili (CS2,CFA)	0	0	0	0	0	1	0	0	0	0	0
	Chaperone/Usher	11	11	8	10	10	10	11	13	7	10	9
	K88 pili	0	0	0	0	0	0	0	0	0	0	0
	P-pili	0	0	0	1	1	1	1	0	0	0	0
	Type-IV-pili	0	0	0	0	0	0	0	0	0	0	0
	Afimbrial adhesive sheath	0	0	0	0	0	0	0	0	0	0	0
	Dr-family adhesin	0	0	0	0	0	0	0	0	0	0	0
	Bundle-forming pili	0	0	1	0	0	0	0	1	0	0	0
	Fimbrial proteins	15	15	13	21	19	19	18	19	10	10	14
	Flp/Fap pili	0	0	0	0	0	0	0	0	0	0	0
Siderophors												
	Ferric dicitrate uptake	1	1	1	0	0	0	0	0	0	1	0
	TonB-dependent receptor	9	9	19	19	16	16	13	14	10	12	16
	siderophore interaction protein	1	1	1	1	1	1	1	1	1	1	1
	Heme-binding protein A	0	0	0	0	0	0	0	0	0	0	0
	aerobactin biosynthesis	0	0	2	2	2	0	0	0	0	0	0
	Hemin degradation	0	0	1	1	1	1	1	1	1	1	1
Microcins												
	Colicin E1 immunity	0	0	0	0	0	0	0	0	0	0	0
	Colicin/Pyocin immunity	0	0	2	2	3	3	0	0	0	0	3
	Cloacin immunity	0	0	0	0	0	0	0	0	0	0	0
	Cloacin	0	0	0	0	0	0	0	0	0	0	0
	S-type Pyocin	0	0	1	0	1	1	0	0	0	0	1
	Colicin	0	0	0	0	0	0	0	0	0	0	0
	Colicin V production	1	1	1	1	1	1	1	1	1	1	1
	HNH endonucleases	1	1	0	1	1	0	0	1	0	2	2
	Bacteriocins	0	0	0	0	0	0	0	0	0	0	0
	α/β enterocin/ lactococcin G	0	0	0	0	0	0	0	0	0	0	0
	Colicin release/ translocation	1	1	1	1	1	1	1	1	1	1	1
LPS/Capsule												
	Assembly core region	1	1	1	1	1	1	1	1	1	1	1
	O-Antigen biosynthesis	1	1	1	1	1	1	1	1	1	1	1
	Polysaccharide biosynthesis	2	2	2	2	2	2	1	1	1	1	2
	LPS kinase (Kdo/WaaP)	2	2	2	2	2	2	2	2	2	2	2
	LPS Saccharide biosynthesis	0	0	0	0	0	0	0	0	0	0	0
	capsule polysaccharide biosynthesis	0	0	2	2	2	3	0	0	0	2	2

The occurrences of virulence structures was detected for all components with a representation in the Pfam database.

B. Supplementary information on the diagnostic microarray for enterobacteria

B.1. Probe specification

Table B.1.: A list of probe features with information about sequence, ID, melting temperature (T_m), binding strength to its complement (ΔG), complexity of base composition, annotation of corresponding genes and funcional category.

Group	Probe-ID	T_m	ΔG	Complexity	Description	Category
Shigella_Ecoli	261_1	74.8931	-120.7399	0.62782	galactitol utilization operon repressor	metabolism
Sequence:	CCATGAAAGCGCGTAACCACGCCTTTTTGTTCGAGAAAGCGCAAATCGGCACGGATTGTCGCTTCCGAGG					
Shigella_Ecoli	14372_1	72.0792	-114.4514	0.63964	citrate reductase, cytochrome c-type, periplasmic	extracellulare
Sequence:	CGTAACTGTCCAACTTCGAGTATATGGATACAACCGCCCAGAAATCGGTTGCCGCGAAGATGCATGACC					
Shigella_Ecoli	14331_2	69.1654	-109.7477	0.61913	intergenic, between threonyl-tRNA synthetase and hypothetical protein	intergenic
Sequence:	AGCGTTTTGCTGGTGTACTCACTACAAACGAATTGCGAATCGAATCAATGTGAAACGGAAAGGTACAATC					
Shigella_Ecoli	506_5	76.463	-120.921	0.62518	lysine-sensitive aspartokinase III	metabolism
Sequence:	CGGTGGTGTAGATGCCCGGGACGTCGGTCCAGATATCAACACGAGATGCGTGTAAAGCCTCCGCCAGCAA					
Shigella_Ecoli	14780_1	75.9133	-120.1123	0.62576	lysine-sensitive aspartokinase III (reverse complement of Seq. 506)	metabolism
Sequence:	CTTGCTGGCGGAGGCTTTACACGCATCTCGTGTTGATATCTGGACCGACGTCCCGGGCATCTACACCACC					
Shigella_Ecoli	154_1	72.0325	-116.3752	0.60261	conserved hypothetical protein / predicted phosphodiesterase	uncharacterised
Sequence:	CGTAATCGAATGATTGTTTGTACGATTTGCGCACTCGTTACGCTCGCTTCGACCCTGAGCGTGCGATTTA					
Shigella_Ecoli	493_19	75.068	-118.3099	0.63074	transcriptional repressor cytR / regulator for deo operon / dual DNA-binding transcriptional regulator	transcription
Sequence:	CGTGTTGATCCCCGGATGATAAGTTCGCAGTCCATTAAACGAGAGCCACTGCCAACGTGTTGCCCCTGCA					
Shigella_Ecoli	4066_2	73.9542	-116.0461	0.632	transcriptional repressor cytR / regulator for deo operon / dual DNA-binding transcriptional regulator (reverse complement of Seq. 493)	transcription
Sequence:	AAATGCAGGGGCAACACGTTGGCAGTGGCTCTCGTTTAATGGACTGCGAACTTATCATCCGGGGATCAAC					
Shigella_Ecoli	338_4	73.8079	-116.0727	0.63519	D-erythrose 4-phosphate dehydrogenase	metabolism
Sequence:	ACCTGTTGATCGTGCATGGCGGAGTGAATTGTGGTCACAGTGCCGGACTCAATACCGTACGCATCATCTA					
Shigella_Ecoli	14411_2	72.1177	-113.2268	0.63548	D-erythrose 4-phosphate dehydrogenase (reverse complement of Seq. 338)	metabolism
Sequence:	TCATCAAATTGTTAGATGATGCGTACGGTATTGAGTCCGGCACTGTGACCACAATTCACTCCGCCATGCA					
Shigella_Ecoli	382_1	72.7648	-113.6822	0.61627	conserved hypothetical protein	uncharacterised
Sequence:	ATATGGCCGAAAAAGACGATACCGCCCAGGTGGTCGTCAACAACAATGATCCCACCTATAAAAGCGAGCA					
Shigella_Ecoli	14419_1	72.7648	-113.6822	0.61627	conserved hypothetical protein (reverse complement of Seq. 382)	uncharacterised
Sequence:	TGCTCGCTTTTATAGGTGGGATCATTGTTGTTGACGACCACCTGGGCGGTATCGTCTTTTTCGGCCATAT					
Shigella_Ecoli	412_1	70.6651	-111.5733	0.60764	DNA repair protein radC	transcription
Sequence:	AAAAGAAACATACTCTCCACGCCCAATCACGATATGGTCGAGCACGCGTAAATCCATGAACTGACAACTC					
Shigella_Ecoli	15557_1	72.2442	-114.249	0.60639	Sensor protein dcuS	transcription
Sequence:	TTTTGACAAAGGTGTCTCGACAAAAGGAAGCGAGCGAGGCGTCGGTTTAGCACTTGTCAAACAACAGGTA					
Shigella_Ecoli	135_1	74.3653	-117.7334	0.63519	NAD-linked malate dehydrogenase	metabolism
Sequence:	GGCATCACGATCGGACGCGGACAGTGTTTATGCATCTCACGGATGATCTCTTCCGTAAACAGCCCGGTCT					
Shigella_Ecoli	3070_1	74.3538	-117.9495	0.64178	NAD-linked malate dehydrogenase (reverse complement of Seq. 135)	metabolism
Sequence:	GGCGTCTCAGGACAGACCGGGCTGTTTACGGAAGAGATCATCCGTGAGATGCATAAACACTGTCCGCGTC					
Shigella	14292_3	75.3455	-119.3968	0.6194	invasion plasmid antigen	virulence effector
Sequence:	TTTTCCGCGTTCCTTGACCGCCTTTCCGATACCGTCTCTGCACGCAATACCTCCGGATTCCGTGAACAGG					
Shigella	14298_1	75.3455	-119.3968	0.6194	invasion plasmid antigen (reverse complement of Seq. 14292)	virulence effector
Sequence:	CCTGTTCACGGAATCCGGAGGTATTGCGTGCAGAGACGGTATCGGAAAGGCGGTCAAGGAACGCGGAAAA					
Shigella	14268_2	75.1278	-119.0055	0.61702	invasion plasmid antigen	virulence effector
Sequence:	CTGTTCACGGAATCCGGAGGTATTGCGTGCAGAGACGGTATCGGAAAGGCGGTCAAGGAACGCGGAGAAT					

Continued on next page…

Group	Probe-ID	T_m	ΔG	Complexity	Description	Category
Shigella	14712_4	75.1732	-119.1986	0.60932	invasion plasmid antigen (reverse complement of Seq. 14268)	virulence effector
Sequence:	CATTCTCCGCGTTCCTTGACCGCCTTTCCGATACCGTCTCTGCACGCAATACCTCCGGATTCCGTGAACA					
Shigella	15191_1	70.3741	-110.6709	0.63737	putative fimbriae usher in S. flex. 8401, partial conserved hypothetical protein in S. flex. 301, intergenic between IS2 orfA and hypothetical protein in S. flex. 2457T	others
Sequence:	ACGATAGGTAAGAATTACGACTTACCACAAACTAGTCAATTGGTTGTACCGACAGCTGGAGCCGTTGTGC					
Shigella	15518_9	71.9692	-113.264	0.63969	putative fimbriae usher in S. flex. 8401, partial conserved hypothetical protein in S. flex. 301, intergenic between IS2 orfA and hypothetical protein in S. flex. 2457T (reverse complement of Seq. 15191)	others
Sequence:	CGCAGGCACAACGGCTCCAGCTGTCGGTACAACCAATTGACTAGTTTGTGGTAAGTCGTAATTCTTACCT					
Shigella	15069_1	71.1627	-111.9833	0.64062	intergenic, between two hypothetical proteins	intergenic
Sequence:	ATCAGCGCCAAGATTTTTCGTCTTTTCGGCGGAATCATGAGACTAGTGCTTACGTCCTGATAACCTCCAG					
Shigella	15100_1	74.7833	-117.9522	0.60425	putative phage transposase	DNA mobility
Sequence:	AACAACAAACCGTGTTCCACCATCGATCACAAACGTGACTTCCGGTGAGAATGGACGCCCGTGGATCGGA					
Shigella	14881_12	75.4554	-118.7764	0.6331	putative Rhs-family protein	repetitive elements, others
Sequence:	CAATCTCCATACCGACCGGGGAGAACGAAGCCGGACAACGTTTGTGTGGGAAGGGTTCCGGCTGTTACAG					
Shigella	15693_1	76.4698	-119.6896	0.62459	putative Rhs-family protein, low similarity with Rhs-family protein from Salmonella strains (reverse complement of Seq. 14881)	repetitive elements, others
Sequence:	CTCCTGTAACAGCCGGAACCCTTCCCACACAAACGTTGTCCGGCTTCGTTCTCCCCGGTCGGTATGGAGA					
commensal	4103_2	65.7577	-101.9755	0.58919	intergenic, between rhsB element core protein RshB and DNA-binding transcriptional regulator, Ni-binding	intergenic
Sequence:	GGAATATATTTAGAACGTTACAAAAGAGAAGGTTCGGATAGGGTATTTAGAGAGGAATCAGGTGTGCTAG					
commensal	3556_16	77.2654	-123.2171	0.6377	acyl-CoA hydratase	metabolism
Sequence:	AGCGACTTGCCGGACGGAGCGCAGATTATCGTGAAGGTGTCAGTGCGTTCCTGGCTAAACGCTCACCGCA					
commensal	3778_8	75.0489	-120.1071	0.6254	predicted enzyme IIB component of PTS	transport
Sequence:	GTTGTTTTAACGCGTCGTCGTTGATCCCTGTGAGCAGGGCGGCGCCATTAAGCGTAGGGATGCCGTAATC					
ExPEC	233_2	70.5657	-110.3437	0.62531	hypothetical protein, putative membrane protein	uncharacterised
Sequence:	TTTATCTGGTCACGATCATCGGAACACTGTTTGCACTGGTCGGTGCAATTGTTACCACTATTCCGGAATA					
ExPEC	196_12	74.5908	-117.6522	0.61916	probable amidase	uncharacterised
Sequence:	CATCAACCCCAAGCTACAACGATGGGCATAGACACCTGTGTAGTGTGCGGGAATGCGCAGTGAGCCAAAC					
ExPEC	3119_35	74.5908	-117.6522	0.61916	probable amidase (reverse complement of Seq. 196)	uncharacterised
Sequence:	GTTTGGCTCACTGCGCATTCCCGCACACTACACAGGTGTCTATGCCCATCGTTGTAGCTTGGGGTTGATG					
ExPEC	41_1	74.9505	-117.1753	0.6393	Outer membrane heme/hemoglobin receptor	extracellulare
Sequence:	ACCCGGTTAACATTATCGTCGGTTCCCGCTATGACCGGTACAAGAGCTTCAATCCCCGTGCCGGAGAACT					
ExPEC	7742_11	74.9505	-117.1753	0.6393	Outer membrane heme/hemoglobin receptor (reverse complement of Seq. 41)	extracellulare
Sequence:	AGTTCTCCGGCACGGGGATTGAAGCTCTTGTACCGGTCATAGCGGGAACCGACGATAATGTTAACCGGGT					
ExPEC	212_2	76.2737	-121.8419	0.62251	putative polyketide synthase	metabolism
Sequence:	GCAATCAAACACGGGGCGCGCCACTTGTTGTGTGTAGGCTAGCGTTGACGCATCAAGGGTAGCCAGCAAC					
ExPEC	214_2	78.8766	-125.9929	0.61718	putative polyketide synthase	metabolism
Sequence:	TCGTTAGCGGCAGACGGTACACCGCGATCCACCCGGTTGCCCCGCGCTAATCTATCGATCGCCCGAGGTG					
ExPEC	207_1	75.9631	-119.392	0.61552	hypothetical protein, probable peptide synthetase protein	uncharacterised
Sequence:	CCTGCTGTACCCCCGGGAAACGCATCGCATGGGCTTCGATTTCTCCCAATTCAACGCGCAACCCTCTAAT					
IPEC	2638_3	76.1072	-119.3759	0.61203	EspF protein	virulence
Sequence:	AGTTCCCCCCGCAGAGCTCACTCGACTTGCGATACCTACAAGCTGCCGCCCTAGTGTAGAAGCAGCGTTA					

Continued on next page…

Group	Probe-ID	T_m	ΔG	Complexity	Description	Category
IPEC	4840_1	77.4384	-124.7486	0.60656	putative transport protein	uncharacterised
	Sequence: CCGAAATGTTCGCCTTCCCACCGACTCACCATCGCAGCGCCCATTGCGTTGCTCATGACGTTTGTCGCGG					
UPEC	548_2	73.8747	-114.8376	0.63151	putative histidine kinase-like ATPases, hypothetical protein of PAI II from E. coli 536	uncharacterised
	Sequence: CATCAGTGTTATGTCATGAACCGTCAGAAGAGCCCCGGGGAGGACTGGGATTGCTGATTGTACGCAGGAT					
UPEC	1886_10	74.8013	-116.8735	0.6272	putative histidine kinase-like ATPases, hypothetical protein of PAI II from E. coli 536, (full match with hemolysin operon from serotype O83:K24:H31)	uncharacterised
	Sequence: CCAGCATCCTGCGTACAATCAGCAATCCCAGTCCTCCCCGGGGCTCTTCTGACGGTTCATGACATAACAC					
UPEC	1540_10	74.1917	-115.7487	0.61432	probable hemagglutinin-related protein, putative member of ShlA/HecA/FhaA exoprotein family, located on PAI II of strain 536 and genomic island I of strain Nissle	extracellulare
	Sequence: AATGGATACCTGGTTCCGTCCACGGACCCGGACAGTCCGTATCTGATTACGGTGAACCCGAAACTGGATG					
EPEC	2609_4	72.2549	-113.5498	0.61491	putative fimbrial out membrane protein	adhesion
	Sequence: CGCCATGCCCCAAGGTTAAGGCCGCTTCGCATACTTAAGTAGTAGCTATCGTTAGTCTGGCTAGTTTCGT					
EPEC	2633_25	75.5655	-120.0445	0.63359	no annotaion available	uncharacterised
	Sequence: GCTGGTGTACCAGACGGCCAGCTATACGACGGATGAACCTCGTCTGCGTATTGTGTGCGAGCTTTCCCGA					
EPEC	2736_18	72.3039	-113.7429	0.63418	no annotaion available	uncharacterised
	Sequence: ATGAACCAAATGGAAGCTATGCGCAACGCAGATTTCAAACGCTACGGAGACTCAGGGGAGTTTCGTCAAT					
EPEC	2280_8	71.8879	-113.9458	0.62639	no annotaion available	uncharacterised
	Sequence: TTGTTACGTTTGCGAGCTTCTAGTAGCTCGGTGAATAGGCGATTAAAATTCTCAACGCGGGCACGGAGTT					
EPEC	2767_3	70.9519	-111.3653	0.64497	Lipoprotein (bfbP) located on plasmid pMAR7	metabolism
	Sequence: GGTGGCGCACTGAATGGAAGTGCTTCCCGAACAGTCACTGTCATTTGTATAACCCCGAGAATTATTGATC					
EPEC	2745_12	71.3827	-113.4129	0.6183	no annotaion available	uncharacterised
	Sequence: CTTTCGCTTCTACATAGCAGCTCACCGACAGGTTATAGTCATTAGCGACAACGGTCTCAAACGCGACAGA					
EPEC	2448_5	78.127	-123.5825	0.61639	no annotaion available	uncharacterised
	Sequence: ACCTGCACAGCACTAAGGGACTGCGCGACGGGGATATGCGCACAGCTAAAGCGTGTCGGGGTGCAGAGTT					
EPEC	2737_1	75.1709	-120.3791	0.62812	no annotaion available	uncharacterised
	Sequence: GACGGATAATGTTGTCTCGCTCGACCTGACAGCGGGGAGCGAAGAGGACGCCGCAAACGAATTGTCTCAC					
EPEC	2791_27	71.319	-112.5522	0.62512	hypothetical protein	uncharacterised
	Sequence: TTTTCACCTTCATACCCGTGCAGACAAAGAATTTCGCTATGCCGGCGAATCCGACAGATATATTCGCGAT					
EPEC	2452_12	73.3554	-115.8453	0.61776	putative major head subunit protein of bacteriophage	uncharacterised
	Sequence: AACTACACCATCCGTAACAAGGACTGGGAAGCCACGGTTGAAGTCGATCGTAACGACATCGAGGACGACC					
EPEC	2497_1	71.0888	-111.4827	0.632	no annotaion available	uncharacterised
	Sequence: ACGGTAAAAGTGACATCTTGCGTCAAGGGAGTACTCATGCCGGTAGCAGTAATGTACAGCGGTGTTTTTA					
EPEC	2321_1	73.2847	-113.8761	0.63485	no annotaion available	uncharacterised
	Sequence: TGCCACTGGTGGGGGGTCTGACCGAGGAACGAATCCGAATTATAACGCTGATGTTCACATTCCAATGTCT					
EPEC	2489_3	75.0142	-117.0803	0.63737	putative terminase large subunit of enterobacterial phage	uncharacterised
	Sequence: TCGCCAACTGATTTAAGCCCGGTCCCGTACCCGTAATCGATAAACACAGCATCGGCCTGGTACTGGTCCT					
EPEC	2743_1	68.6944	-108.4416	0.60114	type I restriction and modification enzyme on plasmid R124/3	transcription
	Sequence: TTTTCGTCTTATATTGCGGGTTTAACGGACTGAGTTTTGGCAATGTCTCGTTTAATTCCGTGCCATTTTC					
EPEC	2631_1	75.8002	-121.1227	0.59587	no annotaion available	uncharacterised
	Sequence: CCCTACATTCATGCCGTACGAGCCGAACGGAACCTACGCTACAAACCGCCATACCCGCCGTATGCTACGC					
EPEC	2456_9	76.1932	-122.348	0.62971	no annotaion available	uncharacterised
	Sequence: ATCAGAACGCTGGACTACACCCGCAAGGCGTGCCGCGAACGTATCAGTCTTCGCTTCCCGCGTGAAAAAC					
EPEC	2502_1	70.8437	-110.712	0.63747	no annotaion available	uncharacterised

Continued on next page...

Group	Probe-ID	T_m	ΔG	Complexity	Description	Category
	Sequence: TTTCATCTAGCTGCCCAATACCACTGGTCGTGCGCTCTGTTGAATACTTCGTGTAACCTGTAAGAAGAGT					
EPEC	2487_1	76.2216	-119.597	0.62231	high similarity with region of phage epsilon15 from Enterobacteria	uncharacterised
	Sequence: AGGTGTGTAGCAGTACGGCATATGGCACATGTGCCGCAGCGGTCCGGATGGGTTCCCTTGATGCTACTTC					
EPEC	2465_1	72.0931	-112.4468	0.63525	high similarity with a region of phage epsilon15 from enterobacteria	others
	Sequence: TTCAGACCATTACCGCAATCCCGGTAGTTGTTCAGACCAGTGATCTGGATAAGTCCACGACCGCGATAAT					
EPEC	2423_4	66.2986	-104.3305	0.61698	putative bacteriophage protein	uncharacterised
	Sequence: TTAATCGCAATCTTCGCTTTAAGAATGACACGATAAGTTTCATCGCTGAGATACATGTATCCTGAATCAG					
EHEC	5121_1	74.8069	-116.9964	0.6331	unknown protein possibly encoded by cryptic prophage CP-933P	uncharacterised
	Sequence: CAGGCCAAATAAGCTCCCAATCATGGGGTCGTAGCTCCGCCCTACTTACTTGGCCTTCCGTCGCAGATTC					
EHEC	4964_9	74.6604	-115.637	0.64328	putative outer membrane protein	Extracellulare
	Sequence: CTGGAGTATCCGGGGTGAGTCGCACCTTCTAATGTTGCCCCTTCAAATACAGGATACCACCCCGCTGCAT					
EHEC	5330_8	77.1694	-120.0413	0.58412	translocated intimin receptor Tir	Adhesion
	Sequence: GCACCGCTATTTGACTCCCTCAACTCCCCAACGCCTTTTGACTCCCCAGCACCTTTGGACTCCCCGGTCC					
EHEC	4956_14	74.9616	-119.4538	0.61684	putative anti-repressor protein encoded by prophage BP-933W	uncharacterised
	Sequence: GCCCGCTGAACTACGAAACTCGACTGGTCAAGCCGTGCCGCCTCTCTCAGTAGCACATCTCTGTTCTTCG					
EHEC	4814_2	70.8437	-111.3392	0.64606	intergenic, between two hypothetical proteins	intergenic
	Sequence: GATCAGTAGCGAGGACCTTATCACCTCTGACTTTCATCTTATGACGGCAGATGAGCGCAACGTGGTTAAA					
EHEC	5403_1	70.8639	-110.6083	0.63602	putative membrane protein	uncharacterised
	Sequence: TTTCAGTAATTGTTTCACCTAACGTGCTGGTTAGGGTTTCGGAGGAATTGCTGACATCATGTGCCATCGA					
EHEC	4982_1	75.0453	-119.8506	0.63432	intergenic, between two hypothetical proteins	intergenic
	Sequence: CTTGAGAATTTAACGGTTGTGGGCTCCTGCGGACACAGCGTTGGTGAATACAGTGCGCTCTACGCCGCTG					
EHEC	5159_4	75.9878	-122.8491	0.61243	unknown protein encoded within prophage	uncharacterised
	Sequence: GGTATTCAGGTTCGGCTGTCTGCAACGACGCGTCATTCAACTCACCGGTCGTAGCACTCTCGCGCATGCG					
EHEC	5183_1	70.3013	-109.8822	0.62546	intergenic, between a putative outer membrane protein and a hypothetical protein	intergenic
	Sequence: AAACTATAACCTCGAATGGCGTGTGTCCTCACGCCATTTACTCTAATTCAGCTTAACGTCACTTGTCCCT					
EHEC	4760_8	72.4182	-114.7142	0.61727	putative fimbrial protein	uncharacterised
	Sequence: CGGTCTTACATCGGTTACCAATCTACCTTCCGCAACGGTCATCCGTGCACTTAACGGGTAATCGTAAGGC					
EHEC	4886_1	75.1622	-118.2342	0.63292	hypothetical protein	uncharacterised
	Sequence: TTCTTGCAGTGTAAACCGTAAGAAACCGCAGCGGGGACAACAGTCCCCGCTATCCGGTTGCCAAAGGTTC					
EHEC	5072_5	76.984	-122.3464	0.62158	hypothetical protein	uncharacterised
	Sequence: AGACCGTGCGGCACTCAAACTGGCTCGACTCGGAAAGCCGGTTAAAATCATGCTCGGCGGGATAACCGGC					
EHEC	5815_1	77.1607	-122.9855	0.6103	hypothetical protein	uncharacterised
	Sequence: CCCAGCCGGTTATCCCGCCGAGCATGATTTTAACCGGCTTTCCGAGTCGAGCCAGTTTGAGTGCCGCACG					
EHEC	4920_1	71.3783	-110.8637	0.64312	hypothetical protein	uncharacterised
	Sequence: AGCGAGTGTCCCCATATACCGATTCTGGAACTGTTCGGAGCAATAGTACTAACCCATGGCATGATCTGAG					
EHEC	5484_1	76.4181	-118.799	0.46668	intergenic	intergenic
	Sequence: ACCTCACACCTCACACCTCACACCTCACACCTCACACCTCACACCTCACACCTCACAACAGCGGGTCTGG					
EHEC	5368_1	75.0402	-117.5001	0.6193	putative permease	uncharacterised
	Sequence: ACCGGGAATGACGCGTACAAAATGAAAGGGCGGAACTTACCTTTCGGGCCAATCTTACGCCGGGAATCCA					
EAEC	6463_7	75.6966	-117.9661	0.62806	no annotation available	uncharacterised
	Sequence: ACCGCCACTGGGGTAGCCCGACAACTTAGCAGTCCTATACTGCTTGACATGACCTCCGTTCCATCACCCG					
EAEC	6619_2	70.9652	-110.717	0.62922	no annotation available	uncharacterised
	Sequence: CCCCTTCCACTACTAAAACTTCTCGTATACGAAGCTGTGAACCAAGCGGACTCACTTCACCTGGTTTGAT					

Continued on next page...

Group	Probe-ID	T_m	ΔG	Complexity	Description	Category
EAEC	6324_16	69.721	-109.3318	0.62022	no annotation available	uncharacterised
Sequence:	GTACTTGAGCAGCTGTACTATCATAAGCACGCATTAGACAGCAACTGGTCTGAAGGTGTAAGGGCTTATG					
EAEC	6771_17	71.5053	-112.3796	0.6379	no annotation available	uncharacterised
Sequence:	AAGTCTTCAGGAGTGGTTCCACTCGACCGCGGTGCATAACTACGCAAACATGTCTCAATATCTCGAGAAA					
EAEC	6604_2	68.493	-105.8365	0.63271	no annotation available	uncharacterised
Sequence:	ATTGCAAGGTTTAATACGGCTCTAAACTTAAACACTCCCAGCACTGAAGTCTTGGTACCCTCATTAATGA					
EAEC	6357_4	67.1797	-105.1753	0.62239	no annotation available	uncharacterised
Sequence:	TCATTGTTAATTCGAGACGCTTGATTGTCCACTATATCAAAACAGTCATACCTCGGAATTTTCTCGCATT					
EAEC	6594_1	73.5035	-115.8801	0.6206	no annotation available	uncharacterised
Sequence:	GGGGATAGAGCTTCTTGATCGTTGCGTCACTAAGGACGGATTGACTGACGTCGGGCGTGAGATATGGCAA					
EAEC	6806_1	60.0299	-92.1298	0.51111	AatD protein located on plasmid pAA2 and on plasmid pO86A1	uncharacterised
Sequence:	TTTTTTTACTATTCTTTTATACTCTATCATTTTCTCTGGGATGTATTTTTCTAGTGGTTTTTTAGAATAT					
EAEC	6762_4	73.3401	-116.339	0.63214	putative phage-related protein	uncharacterised
Sequence:	TCACCAGCATTAAGCATCGGACGAACCCACTGTTCGAAGGGCGCGAGTTTTAACTGACGAGTACACCAAC					
EAEC	6490_2	72.5284	-111.9557	0.62835	no annotation available	uncharacterised
Sequence:	ATCGGCATCAATCTCTGATTGGGGAGCATAGGCCTTATAACCATTAGGCTGTGAAGTGGCTGTGCGGATA					
EAEC	6707_2	71.0353	-111.4337	0.63444	no annotation available	uncharacterised
Sequence:	GCGGGATCACTCCTCAGACAGTTAGTGTGACAGGTCTGGCATACGTTGCATTCTATTCACAACCAGTAAG					
EAEC	6264_1	65.1872	-101.2705	0.61843	no annotation available	uncharacterised
Sequence:	ATGAAGAGTTATCATATAGCAGGGTTATACTAAACGAGTCAATTCCGTCAAGCTTTTGCATAAATTTGTT					
EAEC	6408_1	69.7476	-109.0014	0.61906	no annotation available	uncharacterised
Sequence:	CAACCGTTATCAATTATCCCCTTTTTTTTTCGGGTAGTTCCTGAACATCTCACCGCCAGGTTTCTGTAAGC					
EAEC	6474_2	73.3466	-111.7678	0.62434	putative prophage integrase	uncharacterised
Sequence:	ATGGGACCCTTATATGGACCCACCGACTGGGTCCAATAATTTGAGGGTCCAATACATGGCAAGGCAGACT					
EAEC	6285_1	70.7773	-112.6377	0.61859	hypothetical protein of phages from enterobacteria	uncharacterised
Sequence:	ACTCGTTTTTACCACGCTCTCCAAATGCGTCTTTAGAGTCGTTGTATCCGCAATCGCAGCACACATAATC					
EAEC	6286_2	70.3271	-110.6112	0.63688	putative phage-related DNA recombination protein	uncharacterised
Sequence:	TCTTCTTCGTTTAGTACGTGAATAGCACTATCAAGACGTGATGCCTTAGGCCAATACTTGCTTGCACGCT					
EAEC	6405_1	71.0202	-112.154	0.61465	putative major head protein or intergenic	intergenic
Sequence:	CAGGATTTCCTGCGTGGTTTTCGGGTAGGCGTAGATCGGATTCAGCTTGATGGTTACACGTTCAATTTTC					
EAEC	6528_1	70.0972	-109.3872	0.63854	MchS1 protein	others, microcins
Sequence:	CCAGTTATACAGAGACTACCCGAATATCGTATGGGCTGGTCGTTCTAATACTATACTGGCCGGTCAGAGC					
EAEC	6221_1	66.2677	-103.7859	0.61117	predicted fimbrial-like adhesin protein	adhesion
Sequence:	TTTTAATACTGCTTATACGAACACAATTTTGCAGCGAGATCTGAAAATTAACAGGTGTTGCAGAATCTGA					
EAEC	6685_22	76.8679	-120.9149	0.62309	hypothetical protein	uncharacterised
Sequence:	CACCGCCCGGGTACGTGGGATACGGTCAGGGGGGTATTCTGACGGAAGCTGTACGTAAGCGCCCTTACAG					
Salmonella	10595_2	70.2742	-109.2252	0.62637	negative regulator of flagellin synthesis (anti-*fliA*, anti-sigma factor)	others
Sequence:	ATTTATTTATCCTCATCGAGGGTTACGTTGTAGCGGCCAGCTACCATCATGGTTGAATATCTCATCGGCA					
Salmonella	11456_12	70.3147	-109.789	0.62942	negative regulator of flagellin synthesis (anti-*fliA*, anti-sigma factor), (reverse complement of Seq. 10595)	others
Sequence:	CTGCCGATGAGATATTCAACCATGATGGTAGCTGGCCGCTACAACGTAACCCTCGATGAGGATAAATAAA					
Salmonella	11258_12	69.0496	-109.3825	0.6256	regulatory protein, similarities to E. coli strains	others
Sequence:	TCTTCTTCCGAATCGCGATTGTAATTGTAATTCGCAATGAGAGTTTGGTAGTTATGACCCGACGTTACGG					

Continued on next page...

Group	Probe-ID	T_m	ΔG	Complexity	Description	Category
Salmonella	11523_1	69.3503	-110.2023	0.6179	ferrioxamine B receptor precursor	transport
Sequence:	GTCGAGATAGACGTTATCCAGACTACCGTCCGAAAAACCACGCAATACGATGTAATCAAAGCGGTTAGAG					
Salmonella	12276_1	71.0937	-112.9321	0.63761	ferrioxamine B receptor precursor (shifted reverse complement of Seq. 11523)	transport
Sequence:	CAGATAGGCGCCTCTAACCGCTTTGATTACATCGTATTGCGTGGTTTTTCGGACGGTAGTCTGGATAACG					
Salmonella	11305_4	76.3122	-120.9477	0.61438	putative inner membrane protein or intergenic	intergenic
Sequence:	CCGGGGGCGGTTAAGCGCGCTATCGCAAACAGGAGGACTCACGGGACAATTATTGGCGACAGGCGTATTT					
Salmonella	11330_1	69.5052	-109.7118	0.63203	putative lipoprotein	uncharacterised
Sequence:	CAGTTTCTGCGAGCGAATCATTGACAGATAGTACGCGGAACAGTTGTCAATTGATGATCCTGGCAATTTA					
Salmonella	11540_1	72.8965	-117.4573	0.63015	probable lipoprotein	uncharacterised
Sequence:	CGCAAAATAAAGCTGTCGCGACGCACTAAATCGTCGGTAAGGACGGATTTGTCGTTGTCATGGAGCGACC					
Salmonella	12235_1	72.3149	-115.4277	0.63513	probable lipoprotein (reverse complement of Seq. 11540)	uncharacterised
Sequence:	GATTACTGGTCGCTCCATGACAACGACAAATCCGTCCTTACCGACGATTTAGTGCGTCGCGACAGCTTTA					
Salmonella	11538_1	81.1997	-128.8638	0.57264	hypothetical protein	uncharacterised
Sequence:	ACCAGCGGCCACGGGCGTCCCACGCTATCCCGTGACGCCATCAGCACTCCGGTCAGCCAGTCAGTCGTCC					
Salmonella	11448_1	72.0449	-112.7983	0.63704	hypothetical protein	uncharacterised
Sequence:	AACCGAATGATGCCGATCAAGCTCTCTATGCGACGGATTATCTTAATACCCGGCGTGTAAGTCTGACCCA					
Salmonella	12277_1	73.6758	-116.4788	0.63611	ferrioxamine B receptor precursor	transport
Sequence:	CGTCTATCTCGACGGGTTAAAAATGATGGGCGACACCAATTCGCACAGTTCGTTGGTGGTTGACCCCTGG					
Salmonella	11383_1	79.3909	-127.961	0.61253	6-phosphogluconate dehydratase	metabolism
Sequence:	TTGCCGCTCGTCAGCCTCATATTCCGGACCTGAGCGCGTCGCGCGTCGGAACGGGGCGTGAGTTGTTTGG					
Salmonella	12454_1	79.98	-129.6285	0.60627	6-phosphogluconate dehydratase (reverse complement of Seq. 11383)	metabolism
Sequence:	CAGCGCGCCAAACAACTCACGCCCCGTTCCGACGCGCGACGCGCTCAGGTCCGGAATATGAGGCTGACGA					
Salmonella	11356_1	73.9415	-117.0633	0.62463	putative inner membrane protein	uncharacterised
Sequence:	TGTTCAACGGTGGCAAGCTACTGACGCAAATGTTTGTCGCCGGAATGGGCACTAACGTTATTGCGGGTAA					
Salmonella	12297_1	73.3507	-116.5234	0.60357	outer membrane usher protein FimD precursor	adhesion
Sequence:	GCCCTACGCCACGGTATATCGCTATAACCGCGTCGCGTTAGATACCAACACGATGGACAACCATACCGAT					
Salmonella	11505_1	71.9705	-115.0419	0.61223	outer membrane usher protein FimD precursor (reverse complement of Seq. 12297)	adhesion
Sequence:	GACATCGGTATGGTTGTCCATCGTGTTGGTATCTAACGCGACGCGGTTATAGCGATATACCGTGGCGTAG					
Salmonella	11313_2	76.7723	-123.1654	0.60334	NADP-dependent malate dehydrogenase (decarboxylating)	metabolism
Sequence:	GAAACCCTTGAGCGGGTACGCGAACGAGCGCCCGATCTGATGATTGACGGTGAGATGCACGGTGATGCGG					
Salmonella	12514_1	78.529	-125.4736	0.59744	NADP-dependent malate dehydrogenase (decarboxylating),(reverse complement of Seq. 11313)	metabolism
Sequence:	ACGCCGCATCACCGTGCATCTCACCGTCAATCATCAGATCGGGCGCTCGTTCGCGTACCCGCTCAAGGGT					
Klebsiella	9746_12	74.6613	-116.294	0.63237	gi|43937|emb|X66059.1|K.pneumoniae genes sorC, sorD, sorF, sorB, sorA, sorM and sorE	metabolism
Sequence:	GGTTAATCATGCGTAATATCCTCCGCGTGGGCCAGCAGATACGCTTCAGCATCCTGGGACCAGATCCCAT					
Klebsiella	9335_5	74.9044	-118.4912	0.63566	gi|349480|gb|L23111.1|Klebsiella pneumoniae fimbrial adhesin (fimH) and fimbrial adhesin (fimK) genes, complete cds	adhesion
Sequence:	GGTGTGCTGTCGAGTTTTTCAGGCACCGTGAAATATAACGGCACCTCTTACCCGTTCCCGACCACCACGG					
Klebsiella	8348_26	77.4276	-121.3337	0.60851	intergenic, between peptide transport periplasmic protein and putative transcriptional regulator	intergenic
Sequence:	ACCCGACGGGCGACTAGCCGTACCCGTCGGGGGTGATACAGGGGGAGCGTGTTTCGATTATTGTGACAGT					
Klebsiella	8165_6	71.979	-111.1259	0.63665	intergenic, between putative diaminopropionate ammonia lyase and putative transmembrane amino acid transporter protein	intergenic
Sequence:	ATCGACAATAACAAGGGGGCCATCTTACCTACCCTACAAGCCCTCGGTATGACAAGCGTTCTGACTCATA					
Klebsiella	8276_2	76.2289	-119.135	0.62526	hypothetical protein	uncharacterised

Continued on next page. . .

Group	Probe-ID	T_m	ΔG	Complexity	Description	Category
Sequence:	GGGCCAGGGTGTCCGCGATGTCTCCCAGTTCCCCAGGACGTTCCTGAGGTAACTTTCGTATCAGAGGACG					
Klebsiella	8355_2	80.9502	-128.9451	0.60914	putative transcriptional regulator (LysR family)	uncharacterised
Sequence:	TTCTGCCGTCGGCCGTACGTGCCGGGTGGTACGATGTAGCAGCCGCGCCCCGAGACTGACCTCCAACGCT					
Klebsiella	9926_10	71.1664	-111.0064	0.63581	intergenic, between two hypothetical proteins	intergenic
Sequence:	TCTGGGACGTGCACTTCCAAGGCCGGAAATCTATCTCGATTATGATATCGCAAGGGATATTTGCCACTTC					
Klebsiella	8121_3	76.0332	-119.8714	0.61089	anaerobic C4-dicarboxylate transporter	Transport
Sequence:	GCGCGTTAAGATAATGGGTAAGCGGGCTAAGCTTGTCCGGAGTACGGGCTACCTCTCCGGGGTCGCGATA					
Klebsiella	8587_10	75.3155	-120.486	0.62501	methionine aminopeptidase	metabolism
Sequence:	GAAAGCGGCCCACCGTCATAGAGAAAGGGAAGCAACTCAGCGGATCCGCGACTGTTCATCGTCGCGTAAC					
Klebsiella	8268_2	78.3941	-124.2179	0.60048	putative PTS family enzyme IIBC, glucitol/sorbitol-specific	transport
Sequence:	TGATTCGGTAGCCCTGTCGCGGACCGATGTGGCTCGCGATGTACGTACCCCGGTCACCGAGTCGTGGGTA					
Klebsiella	8821_2	72.4911	-114.7316	0.63448	intergenic, between transposase, IS4 and L, D-carboxypeptidase A (in murein recycling)	intergenic
Sequence:	AAATTAATCCGGACCATTGTCCGCGCCTTGCACGAATCGGAAAAGTTAACTGCACTGAAGATCCGCTCTG					
Klebsiella	9261_4	81.6321	-130.7319	0.59391	formate hydrogen-lyase transcriptional activator	transcription
Sequence:	CTGGCGGCGGACGGGCGGATCTTTGGCGGTTGCGAATTCCTCCGCCGTGACAACCGGCCGTGGAGCGAGA					
Klebsiella	9307_2	78.4223	-125.5599	0.6223	propionate kinase	metabolism
Sequence:	CGGCATAAACGACGCGCCTCCTCTTGCGGAGAATCATGGTAACCGGCGGGTCGCCATGCCACTATGCGCC					
Klebsiella	9634_3	77.9707	-123.9271	0.61678	putative regulator	uncharacterised
Sequence:	TGGCCGCTGCTGTCGTCGCAGGTCCCGCATCACTCGCATATAGGGCCCGCTAAAATCGTAGAACTGCCGG					
Klebsiella	9698_15	72.3813	-113.8218	0.64818	intergenic, between periplasmic repressor of cpx regulon by interaction with CpxA and ferrous iron efflux protein F	intergenic
Sequence:	GGTAAATCAACTCCTTCGCCTATCCCGTGAGCGTCACAAGTTCGGGTTATACTAAGCGCATTGCAGGAGA					
Klebsiella	10147_17	77.2086	-122.4707	0.63145	intergenic on plasmid pKPN5, between putative anti-restriction protein and hypothetical protein	intergenic
Sequence:	CAGCGGCCGGGCTAAAGTGAAGGCACGGCACAACCCGTCGCGATGCAGGAGGTTGAATACAGGGTTTCTC					
Klebsiella	7950_1	69.8566	-107.608	0.62186	intergenic, between two hypothetical proteins	intergenic
Sequence:	CTATTTCCCCTACCTGATAGACTTAGTGTCACCGTATCCTGTTACTAAGAGCACGGGGCTACCTACTCAT					
Klebsiella	9249_2	77.75	-123.6693	0.62922	hypothetical protein	uncharacterised
Sequence:	ACCATTCGCACGCCCTGCACTCGGCGTGTGACTTCTATGGCTCGTTTAGCGCTCTCCGCGGAGGAGACAA					
Klebsiella	9679_20	73.9034	-115.9249	0.62716	ribonuclease BN	metabolism
Sequence:	GCTGATCCTCTCCTGGGCAGCGTTCTGGCTGCTCTATAGTATTGTACCGACTACCCAGGTACGTAACCGC					
Klebsiella	9853_39	76.0811	-119.2846	0.63473	putative carbohydrate kinase	uncharacterised
Sequence:	ACCAATCTGCTGGGCATAGAGGTCGACGGCCACCCGACCTATACAGATGACGTCGAGCCGCTTTACTGCT					
Yersinia	20994_24	75.4227	-119.0857	0.62443	putative oxidoreductase	uncharacterised
Sequence:	CCCTCCGTGTAAGACAAGCTTGGGGTTCGACCCTAACAAGCGTGCAGAGTAAGTGCGAAGCCCCAACTCG					
Yersinia	21601_47	75.0344	-116.8537	0.64109	NADH dehydrogenase I chain A	metabolism
Sequence:	ATTCATAAGGCACGTTTTTGGCACGAGCCTGGGCTCTCCCCCGAGGAAGTACGCACCTAAAAGCATCAA					
Yersinia	26912_2	75.0344	-116.8537	0.64109	NADH dehydrogenase I chain A (reverse complement of Seq. 21601)	metabolism
Sequence:	TTGATGCTTTTAGGTGCGTACTTCCTCGGGGGAGAGCCCAGGCTCGTGCCAAAAACGTGCCTTATGAAT					
Yersinia	22243_19	68.4494	-106.7133	0.61257	intergenic, between aspartate semialdehyde dehydrogenase and putative membrane protein	intergenic
Sequence:	CTAACACCACTACAACGTCAAGTACGAGGGGTATATCTAGCACGGCAGACTGACTAAGATAGACTTGATA					
Yersinia	28630_10	68.4494	-106.7133	0.61257	intergenic, between aspartate semialdehyde dehydrogenase and putative membrane protein (reverse complement of Seq. 22243)	intergenic
Sequence:	TATCAAGTCTATCTTAGTCAGTCTGCCGTGCTAGATATACCCCTCGTACTTGACGTTGTAGTGGTGTTAG					

Continued on next page…

Group	Probe-ID	T_m	ΔG	Complexity	Description	Category
Yersinia	21132_10	75.0225	-116.4128	0.63379	putative fimbrial protein	uncharacterised
Sequence:	TAGCAGCGATTGGACCCAACTAGAGGGTTTTGGGCTCGTAGACGGCACGGTGAGTGCCCCATTAATAGGA					
Yersinia	26917_2	76.7351	-121.6179	0.63456	putative lipoprotein	uncharacterised
Sequence:	GAGAGCGTGCGCTCAAGTGATGCACTCGTTTCCGGCCGGGTATCCATTGAGACACGGTAGGGGTTCAGCC					
Yersinia	30671_2	72.1165	-111.5209	0.64109	putative fimbrial protein (shifted reverse complement of Seq. 21132)	uncharacterised
Sequence:	TCATTATAAGGAATCGGTCCTATTAATGGGGCACTCACCGTGCCGTCTACGAGCCCAAAACCCTCTAGTT					
Yersinia	20171_17	70.7941	-110.4468	0.64109	putative outer membrane fimbrial usher protein	uncharacterised
Sequence:	ATGAATATAAGGACGGCCTATTCACAATTAACCGGTGGGGTTTCTCAAGCGACAAACAATAGCCGGTCTG					
Yersinia	20606_7	72.326	-114.1838	0.64594	putative thiamine pyrophosphate-dependent protein	uncharacterised
Sequence:	GTTATCAAACAGTAGGACGGTAATTTTTAGCCCCTCTTGTACCGCCGTGTGCAGCTCGGAATGCAGCATC					
Yersinia	21880_2	70.4252	-109.6711	0.63805	putative exported protein, putative chondroitin lyase	uncharacterised
Sequence:	TTTCGTTATAGATCCAAAAAGAGAGCACCTGTGTGGCTTTACGTCCCCAGGTTTTAGAGGCGACCTTATC					
Yersinia	21938_21	65.5431	-102.4803	0.6199	exodeoxyribonuclease V beta chain	Metabolism
Sequence:	CAACATACTTTATTTAGTGCGATAGACTCGATATTTGAACAGCCCCTAACGTTGAGGGATCTTATTTTAG					
Yersinia	23859_3	71.6727	-110.7278	0.63184	intergenic between putative chorismate mutase and putative lipoprotein, conserved hypothetical protein	intergenic
Sequence:	TTGACAATGACGGCTCGTGGGGTTGATAGGGAATGGTCTTGGTATCATCTCTTCCTATCAAAGAGGCCAG					
Yersinia	25479_2	73.5744	-115.9696	0.63184	intergenic between DNA-binding protein Fis and hypothetical protein	intergenic
Sequence:	TTCAGGTGTTCTTTCACGGGCGGGCACCCATACGTCAACAGAGAGTAAGGTTTACTCGCTAGAGCGTGAC					
Yersinia	28089_24	66.9191	-104.4046	0.61873	exodeoxyribonuclease V beta chain	metabolism
Sequence:	GGGCTAAAATAAGATCCCTCAACGTTAGGGGCTGTTCAAATATCGAGTCTATCGCACTAAATAAAGTATG					
Yersinia	34773_2	70.7941	-110.4468	0.64109	putative outer membrane fimbrial usher protein	uncharacterised
Sequence:	CAGACCGGCTATTGTTTGTCGCTTGAGAAACCCCACCGGTTAATTGTGAATAGGCCGTCCTTATATTCAT					
Yersinia	20417_14	74.93	-116.55	0.64072	putative acyltransferase	uncharacterised
Sequence:	ATCTGGCATTTAAGTCAGCGGTGCCCGGAAGTCCCCATTATTCCTGTCTACATGCACGGCCTTGACCGTT					
Yersinia	21119_24	77.9589	-122.492	0.61399	integral membrane protein PqiA family, low similarity with Y. enterocolitica	uncharacterised
Sequence:	GCAGCGTGGGCAATAGGCTGTTTGGTTGCCCCCCAATGGTGGCAAGGTGAATAGGGCATCGCACTCGCAG					
Yersinia	24967_7	75.3959	-117.4959	0.64478	putative acyltransferase (shifted reverse complement to Seq. 20417)	uncharacterised
Sequence:	TTTACCCATCGAACGGTCAAGGCCGTGCATGTAGACAGGAATAATGGGGACTTCCGGGCACCGCTGACTT					
Yersinia	30684_8	77.213	-121.1268	0.6182	integral membrane protein PqiA family (reverse complement of Seq. 21119)	uncharacterised
Sequence:	GATGCTGCGAGTGCGATGCCCTATTCACCTTGCCACCATTGGGGGGCAACCAAACAGCCTATTGCCCACG					
Beta-lactams	blaTEM_1				ampicilin resistance gene (AF307748, 8674-8605)	AMR
Sequence:	AAAGTTCTGCTATGTGGCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATAC					
Beta-lactams	blaSHV_1				ampicilin resistance gene (AF148850, 86-17)	AMR
Sequence:	CTCAAGCGGCTGCGGGCTGGCGTGTACCGCCAGCGGCAGGGTGGCTAACAGGGAGATAATACACAGGCGA					
Beta-lactams	blaOXA-1_1				ampicilin resistance gene (AJ238349, 256-187)	AMR
Sequence:	AAACAACCTTCAGTTCCTTCAAATAATGGAGATGCGACAGTAGAGATATCTGTTGATGCACTGGCGCTGC					
Beta-lactams	blaOXA-7_1				ampicilin resistance gene (X75562, 295226)	AMR
Sequence:	GTAGCGCAGGCTAATTTACTGCTACTTTTACAAAGCACGAAAACACCATTGACGGCTTCGGCAGAGAACTA					
Beta-lactams	blaPSE-4_1				ampicilin resistance gene (J05162, 348-279)	AMR
Sequence:	CGCTGATTGCCATTGTAATCCCAATATTCTCCATTTTGAGTATCAAGAACGGAAACACCTATACGAGCAG					
Beta-lactams	blaCTX-M-1_1				ampicilin resistance gene (X92506, 143-74)	AMR
Sequence:	ATACAGCGGCACACTTCCTAACAACAGCGTGACGGTTGCCGTCGCCATCAGCGTGAACTGACGCAGTGA					

Continued on next page. . .

Group	Probe-ID	T_m	ΔG	Complexity	Description	Category
Aminoglycosides	ant(3)-Ia aadA_1				gene for streptomycin and spectinomycin resistance (X12870, 1588-1519)	AMR
Sequence:	ATGATGTCGTCGTGCACAACAATGGTGACTTCTACAGCGCGGAGAATCTCGCTCTCTCCAGGGGAAGCCG					
Aminoglycosides	ant(2)-Ia aadB_1				gene for kanamycin, neomycin and gentamicin resistance (M86913, 1778-1709)	AMR
Sequence:	CCCGAGTGAGGTGCATGCGAGCCTGTAGGACTCTATGTGCTTTGTAGGCCAGTCCACTGGTGGTACTTCA					
Aminoglycosides	aac(3)-IIa aacC2_1				gene for gentamicin resistance (S68058, 200-131)	AMR
Sequence:	CACCGGTTTGGACTCCGAGTTTTCGAATTGCCTCCGTTATTGCCTTCCGCGTATGCATCGCGATATCTCC					
Aminoglycosides	aac(3)-IV_1				gene for entamicin resistance (X01385, 380-311)	AMR
Sequence:	TCGATCAGTCCAAGTGGCCCATCTTCGAGGGGCCGGACGCTACGGAAGGAGCTGTGGACCAGCAGCACAC					
Aminoglycosides	aph(3)-Ia aphA1_1				gene for kanamycin and neomycin resistance (V00359, 1310-1241)	AMR
Sequence:	GGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATT					
Aminoglycosides	aph(3)-IIa aphA2_1				gene for kanamycin and neomycin resistance (V00618, 220-151)	AMR
Sequence:	AGTCATAGCCGAATAGCCTCTCCACCCAAGCGGCCGGAGAACCTGCGTGCAATCCATCTTGTTCAATCAT					
Tetracycline	tet(A)_1				gene for tetracycline resistance (X00006, 1390-1321)	AMR
Sequence:	GATGCCGACAGCGTCGAGCGCGACAGTGCTCAGAATTACGATCAGGGGTATGTTGGGTTTCACGTCTGGC					
Tetracycline	tet(B)_1				gene for tetracycline resistance (V00611, 190-121)	AMR
Sequence:	CAAAGTGGTTAGCGATATCTTCCGAAGCAATAAATTCACGTAATAACGTTGGCAAGACTGGCATGATAAG					
Tetracycline	tet(C)_1				gene for tetracycline resistance (J01749, 130-61)	AMR
Sequence:	GACTGGCGATGCTGTCGGAATGGACGATATCCCGCAAGAGGCCCGGCAGTACCGGCATAACCAAGCCTAT					
Tetracycline	tet(D)_1				gene for tetracycline resistance (X65876, 1770-1701)	AMR
Sequence:	CAAACGCGGCACCCGCCAGGGATAACAGCAGCACCGGTCTGCGCCCCAGCTTATCTGACCATCTGCCCAG					
Tetracycline	tet(E)_1				gene for tetracycline resistance (L06940, 370-301)	AMR
Sequence:	GTTGAGGCTGCAACAGCTCCAGTCGCACCGGTAATACCAGCAATTAAGCGTCCCAAATACAACACCCACA					
Tetracycline	tet(Y)_1				gene for tetracycline resistance (AF070999, 1770-1701)	AMR
Sequence:	TTAATAAAGCCGGAACCACCGGCATGATTAATCCCAAACCAATCGCATCAAGCGCGACAACAATGAGTGC					
Phenicols	catI_1				gene for chloramphenicol resistance (M62822, 550-481)	AMR
Sequence:	TTTACGGTCTTTAAAAAGGCCGTAATATCCAGCTGAACGGTCTGGTTATAGGTACATTGAGCAACTGACT					
Phenicols	catII_1				gene for chloramphenicol resistance (X53796, 300-231)	AMR
Sequence:	AGCGGTAATATCGAGTTTGGTGGTCAGGCTGAATCCGCATTTAATCTGCTGACGATAAAGGGCAAAGTGT					
Phenicols	catIII_1				gene for chloramphenicol resistance (X07848, 370-301)	AMR
Sequence:	TTTGCTTGTTAAGCTAAAACCACATGGTAAACGATGCCGATAAAACTCAAAATGCTCACGGCGAACCCAA					
Phenicols	floR_1				gene for chloramphenicol and florfenicol resistance (AF252855, 384-315)	AMR
Sequence:	GACAAAGGCCGGTGCAGTTGAAGACCAAGCTGCTCCCAGAGACGCAATGACGAAAGCCGTTGCGCCCGCA					
Trimethoprim	dhfrI_1				gene for trimethoprim resistance (X00926, 490-421)	AMR
Sequence:	GGTTAAAGCATCTTTAATTGATGGAAAGATCAATACGTTCTCATTGTCAGATGTAAAACTTGAACGTGTT					
Trimethoprim	dhfrV_1				gene for trimethoprim resistance (X12868, 1560-1491)	AMR
Sequence:	GTACATGGCCTCTTCGATCGACGGGAATACTATTACGTTGTCATTATCGGCCGTCCAGGCTGAGCGAGTA					
Trimethoprim	dhfrVII_1				gene for trimethoprim resistance (X58425, 753-684)	AMR
Sequence:	GAACACCCATAGAGTCAAATGTTTTCCTTCCAACAAGGAGCCACTGATTATATGTGAGCGCTTTAAAGAG					
Trimethoprim	dhfrIX_1				gene for trimethoprim resistance (X57730, 830-761)	AMR
Sequence:	AGCTTTGAAGTGTTTTAAATCTTCTGGTTCATGCCACGGAATCTGATTTTCAAATCCGATACCTCCTGTC					
Trimethoprim	dhfrXIII_1				gene for trimethoprim resistance (Z50802, 929-860)	AMR
Sequence:	TGGCGCGAGAGCACCACTGTGTGGCGGTTTGGTAAGGGCTTGCCTATGGACTCAAATGTCTTGCGGCCCA					
Trimethoprim	dhfrXV_1				gene for trimethoprim resistance (Z83311, 620-551)	AMR

Continued on next page…

Group	Probe-ID	T_m	ΔG	Complexity	Description	Category
Sequence:	CTTCAGATGATTTAGCGCTTCATCGATAGATGGAAATACCAATACATTCTCATCACTGGAAGTGAAGCTT					
Sulfonamids	sulI_1				gene for sulphonamid resistance (X12869, 960-891)	AMR
Sequence:	AGCGCCGGCGGGGTCTAGCCGCCGGCTCTCATCGAAGAAGGAGTCCTCGGTGAGATTCAGAATGCCGAAC					
Sulfonamids	sulII_1				gene for sulphonamid resistance (M36657, 420-351)	AMR
Sequence:	TACGCGCCTGCGCAATGGCTGCGTCTGGCGCCAGATACCGGCCTCCATCGGAGAAACTGTCCGAGGTTAT					
Class 1 integron	qacEdelta1-sulI_1				site mediating resistance to quaternary ammonium compounds and sulfonamids (M33633, 1200-1131)	AMR
Sequence:	TTGGATGCCCGAGGCATAGACTGTACCCCAAAAAACAGTCATAACAAGCCATGAAAACCGCCACTGCGCC					
Class 1 integron	3’-conserved region_1				conserved region of integron (AY152821, 2368-2299)	AMR
Sequence:	GGCTGTAATTATGACGACGCCGAGTCCCGACCAGACTGCATAAGCAACACCGACAGGGATGGATTTCAGA					
Class 1 integron	int integrase_1				gene coding for the integrase enzyme in conjunction with blaVIM-2 gene (AY781413, 284-215)	AMR
Sequence:	CGTTCGGTCAAGGTTCTGGACCAGTTGCGTGAGCGCATACGCTACTTGCATTACAGTTTACGAACCGAAC					
Class 1 integron	cmlA_2	95.8			class 1 integron gene for chloramphenicol resistance (gi121647059, 278-347)	AMR
Sequence:	AATGGGCCTCGCTCTTACGTCATCGGCTGAAGTCTTTCTGGGGCTTCGGATTCTTCAGGCTTGTGGTGCC					
Class 1 integron	sul3_2	88.2			class 1 integron sulphonamide resistance protein (sul3) gene (gi121647059, 202-271)	AMR
Sequence:	TGAATTTCAGGGAAACGCTTCAAAACAAGAATAGATGTTTCTGGATTAGAGCCTAAAAGAAGCCCATAC					
Beta-lactams	blaOXA-73_2	94.0			resistance against extended spectrum beta-lactamases (AY762325.1, 472-541)	AMR
Sequence:	GGAGAAGCCATGAAGCTTTCTGCAGTCCCAGTCTATCAGGAACTTGCGCGACGTATCGGTCTTGATCTCA					
Beta-lactams	blaPER-2_2	92.3			extended-spectrum class A beta-lactamase (gi1524367, 649-718)	AMR
Sequence:	TTGAAACCACCACAGGACCACAGCGGTTAAAAGGCTTGTTACCTGCTGGTACTATAGTGGCGCATAAAAC					
Beta-lactams	blaCARB-8_1	100.5			Beta-lactamase (gi22203985, 195-264)	AMR
Sequence:	AATTATTGTGAGCATCTATCTAGCTCAAACACAGGCTTCAATGGCAGAGCGAAATGATGCGATTGTTAAA					
Beta-lactams	blaACC-1_1	90.5			class C Beta-lactamase (gi58333811, 414-483)	AMR
Sequence:	GGTGGTTTTGATACCATAAGCTTCGTTACCCAAAATCTCCATATTCACGTGCACTGGCTCATCTTTCTTG					
Tetracycline	tet(G)_1	96.4			tetracycline resistance protein tet(G) gene (gi12719011, 65-134)	AMR
Sequence:	GGTCGCTGGACACTATGGTGCCTTGCTGTCGCTCTATGCATTGATGCAGGTCGTCTTCGCGCCCATGCTT					
Tetracycline	tet(G)_2	99.3			tetracycline resistance protein tet(G) gene (gi12719011, 224-293)	AMR
Sequence:	GCCGGTCTTATGGGTGCTCTATATCGGCCGACTCGTGTCCGGCGTCACGGGCGCAACCGGAGCTGTAGCA					
Macrolides	mphB_1	88.2			macrolide-2’-phophotransferase II (gi4218049, 354-423)	AMR
Sequence:	AAACTGGATCATTTACACAGAGGAACTAATAGCTTATAAAAAGTTAGATGGTGTGCCAGCAGGTACGATA					
Negative control	gfp1_1	88.2			green fluorescence protein 1 (gfp1) gene (X83959.1, 236-305)	others
Sequence:	ACAGCATGACTTTTTCAAGAGTGCCATGCCCGAAGGTTATGTACAGGAAAGAACTATATTTTACAAAGAT					
Negative control	At3g51820_1				Arabidopsis thaliana chlorophyll synthetase (AY081481, 788-719)	others
Sequence:	GTAAGAGTGCCAAACAATGCTTGGCCAGCCCACCATGGCAAACTAATATAGCTTGCTCCAAGTGCAAAAT					
Positive control	nadB_5	93.4			quinolinate synthetase, B protein; L-aspartate oxidase (AF403416, 23-92)	metabolism
Sequence:	ATCCAGAAAATGGTACGCGACGCAGAAGCTAACGCCGAAGCTGACCGTAAGTTTGAAGAGCTGGTACAGA					
Positive control	23SrDNAa1_18	73.9		0.61	23S ribosomal rRNA gene a1	translation
Sequence:	CTGAAACATCTAAGTACCCCGAGGAAAAGAAATCAACCGAGATTCCCCCAGTAGCGGCGAGCGAACGGGG					
Positive control	23SrDNAa2_6	72.0		0.62	23S ribosomal rRNA gene a2	translation
Sequence:	GTCTGAATATGGGGGGACCATCCTCCAAGGCTAAATACTCCTGACTGACCGATAGTGAACCAGTACCGTG					
Positive control	23SrDNAa6_8	72.2		0.64	23S ribosomal rRNA gene a6	translation
Sequence:	CAGGATGTTGGCTTAGAAGCAGCCATCATTTAAAGAAAGCGTAATAGCTCACTGGTCGAGTCGGCCTGCG					
Positive control	23SrDNAa7_34	74.2		0.65	23S ribosomal rRNA gene a7	translation

Continued on next page. . .

Group	Probe-ID	T_m	ΔG	Complexity	Description	Category
Sequence:	GTAGCGAAATTCCTTGTCGGGTAAGTTCCGACCTGCACGAATGGCGTAATGATGGCCAGGCTGTCTCCAC					
Positive control	16SrDNAa1_46	76.3		0.62	16S ribosomal rRNA gene a1	translation
Sequence:	GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTG					
Positive control	16SrDNAa2_8	75.6		0.63	16S ribosomal rRNA gene a2	translation
Sequence:	GAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTG					
Positive control	16SrDNAa4_38	74.9		0.63	16S ribosomal rRNA gene a4	translation
Sequence:	ATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTAC					
Positive control	16SrDNAa5_12	74.7		0.63	16S ribosomal rRNA gene a5	translation
Sequence:	GCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTATC					

C. Supplementary data on ipHMMs

C.1. Validation of ipHMMs based on generated sequences

Table C.1.: Validation Results of Peptide-ligand ipHMMs using generated Sequences

Domain	SMART Abr.	Sensitivity	Specificity	Accuracy
Domain present in cyclins, TFIIB and Retinoblastoma	CYCLIN	0.57	0.80	0.77
Histone H3	H3	0.97	0.49	0.69
WD40 repeats	WD40	0.50	0.83	0.79
Alkaline phosphatase homologues	alkPPc	0.76	0.69	0.71
Actin	ACTIN	0.74	0.59	0.63
Insulin / insulin-like growth factor / relaxin family	IlGF	0.88	0.70	0.75
Trypsin-like serine protease	TrypSPc	0.35	0.77	0.72
Beta-propeller repeat	PQQ	0.59	0.60	0.60
BPTI/Kunitz family of serine protease inhibitors	KU	0.61	0.71	0.70
Src homology 2 domains	SH2	0.74	0.80	0.79
Kazal type serine protease inhibitors	KAZAL	0.82	0.68	0.71
Epidermal growth factor-like domain	EGF	0.22	0.79	0.66
Alpha-lactalbumin / lysozyme C	LYZ1	0.70	0.77	0.76
Histone H2B	H2B	0.97	0.35	0.66
Caspase, interleukin-1 beta converting enzyme (ICE) homologues	CASc	0.68	0.47	0.53
Pancreatic ribonuclease	RNAsePc	0.74	0.58	0.62
EF-hand, calcium binding motif	EFh	0.67	0.72	0.73
Immunoglobulin	IG	0.36	0.83	0.75
Ligand binding domain of hormone receptors	HOLI	0.49	0.79	0.74
Serine/Threonine protein kinases, catalytic domain	STKc	0.26	0.86	0.82
Ricin-type beta-trefoil	RICIN	0.58	0.76	0.73
Extension to Ser/Thr-type protein kinases	STKX	0.70	0.65	0.66
Immunoglobulin C-Type	IGc1	0.62	0.73	0.72
Zinc-dependent metalloprotease	ZnMc	0.45	0.69	0.65
ATPases associated with a variety of cellular activities	AAA	0.24	0.87	0.77
Immunoglobulin V-Type	IGv	0.29	0.86	0.82
ankyrin repeats	ANK	0.78	0.64	0.67
Serine Proteinase Inhibitors	SERPIN	0.46	0.72	0.68
Armadillo/beta-catenin-like repeats	ARM	0.28	0.90	0.84
Immunoglobulin C-2 Type	IGc2	0.61	0.74	0.72
Histone H4	H4	0.90	0.36	0.61
Histone 2A	H2A	0.88	0.43	0.63
Alpha-amylase domain	Aamy	0.34	0.77	0.70
Gelsolin homology domain	GEL	0.61	0.65	0.64
Rho (Ras homology) subfamily of Ras-like small GTPases	RHO	0.90	0.61	0.68
Src homology 3 domains	SH3	0.74	0.71	0.72
Summary		0.61	0.69	0.70

The table lists constructed HMMs for SMART domains with known interactions to peptide ligands and with a minimum set of 20 seed sequences. Sensitivity, specificity and accuracy are calculated based on the re-estimation of generated sequences as parameters of prediction quality.

Table C.2.: Validation results of nucleotide ligand ipHMMs using generated sequences

Domain	SMART Abr.	Sensitivity	Specificity	Accuracy
Zinc finger	ZnF_C2H2	0.62	0.55	0.56
RNA recognition motif	RRM	0.76	0.66	0.68
Actin	ACTIN	0.88	0.73	0.76
Cyclic nucleotide-monophosphate binding domain	cNMP	0.63	0.68	0.67
DNA polymerase X family	POLXc	0.80	0.91	0.90
3’-5’ exonuclease	35EXOc	0.68	0.73	0.72
Histidine kinase-like ATPases	HATPasec	0.46	0.68	0.63
Pancreatic ribonuclease	RNAsePc	0.80	0.62	0.66
Pumilio-like repeats	Pumilio	0.81	0.66	0.70
Serine/Threonine protein kinases, catalytic domain	STKc	0.37	0.78	0.72
DNA polymerase A domain	POLAc	0.79	0.77	0.77
ATPases associated with a variety of cellular activities	AAA	0.51	0.95	0.93
Homeodomain	HOX	0.79	0.60	0.66
Helix-hairpin-helix DNA-binding motif class 1	HhH1	0.85	0.80	0.80
Basic region leucin zipper	BRLZ	0.78	0.58	0.65
C4 zinc finger in nuclear hormone receptors	ZnF_C4	0.83	0.61	0.67
Rab subfamily of small GTPases	RAB	0.78	0.68	0.70
Helix-turn-helix lactose operon repressor	HTHLACI	0.88	0.62	0.70
Summary		0.72	0.70	0.72

Table of HMM prediction quality for SMART domains with interactions to nucleotide ligands. Domains can exhibit binding interfaces of different ligand types (like the actin and zinc finger domains).

Table C.3.: Validation results of peptide-ligand ipHMMs using generated sequences

Domain	SMART Abr.	Sensitivity	Specificity	Accuracy
Alkaline phosphatase homologues	alkPPc	0.67	0.55	0.58
Zinc finger	ZnFC2H2	0.55	0.56	0.55
Insulin / insulin-like growth factor / relaxin family.	IlGF	0.83	0.60	0.65
Trypsin-like serine protease	TrypSPc	0.38	0.75	0.69
Alpha-lactalbumin / lysozyme C	LYZ1	0.73	0.77	0.76
EF-hand, calcium binding motif	EFh	0.75	0.74	0.74
DNA polymerase A domain	POLAc	0.85	0.76	0.77
Zinc-dependent metalloprotease	ZnMc	0.50	0.78	0.74
ATPases associated with a variety of cellular activities	AAA	0.33	0.89	0.83
Immunoglobulin V-Type	IGv	0.52	0.79	0.76
Eukaryotic homologues of bacterial periplasmic substrate binding proteins	PBPe	0.88	0.65	0.70
Alpha-amylase domain	Aamy	0.30	0.84	0.78
C-type lectin (CTL) or carbohydrate-recognition domain (CRD)	CLECT	0.23	0.83	0.77
Gelsolin homology domain	GEL	0.66	0.67	0.68
Bacterial periplasmic substrate-binding proteins	PBPb	0.52	0.65	0.63
Summary		0.58	0.72	0.71

C.2. Cross-validation results of ipHMMs

Table C.4.: Cross validation results of peptide ligand ipHMMs

Domain	SMART Abr.	Sensitivity	Specificity	Accuracy
Domain present in cyclins, TFIIB and Retinoblastoma	CYCLIN	0.00	0.70	0.68
Histone H3	H3	1.00	0.65	0.79
WD40 repeats	WD40	0.00	0.91	0.87
Alkaline phosphatase homologues	alkPPc	0.90	0.59	0.62
Actin	ACTIN	0.47	0.48	0.48
Insulin / insulin-like growth factor / relaxin family	IlGF	0.77	0.62	0.65
Trypsin-like serine protease	TrypSPc	0.20	0.82	0.72
Beta-propeller repeat	PQQ	1.00	0.47	0.50
BPTI/Kunitz family of serine protease inhibitors	KU	0.33	0.60	0.57
Src homology 2 domains	SH2	1.00	0.68	0.69
Kazal type serine protease inhibitors	KAZAL	0.46	0.80	0.72
Epidermal growth factor-like domain	EGF	0.10	0.82	0.66
Alpha-lactalbumin / lysozyme C	LYZ1	0.80	0.66	0.67
Histone H2B	H2B	1.00	0.31	0.63
Caspase, interleukin-1 beta converting enzyme (ICE) homologues	CASc	0.00	0.55	0.37
Pancreatic ribonuclease	RNAsePc	0.82	0.64	0.67
EF-hand, calcium binding motif	EFh	0.57	0.86	0.79
Immunoglobulin	IG	0.00	1.00	0.96
Ligand binding domain of hormone receptors	HOLI	1.00	0.78	0.79
Serine/Threonine protein kinases, catalytic domain	STKc	0.41	0.91	0.88
Ricin-type beta-trefoil	RICIN	0.56	0.73	0.72
Extension to Ser/Thr-type protein kinases	STKX	0.00	0.60	0.59
Immunoglobulin C-Type	IGc1	0.45	0.65	0.62
Zinc-dependent metalloprotease	ZnMc	1.00	0.66	0.66
ATPases associated with a variety of cellular activities	AAA	0.00	1.00	0.97
Immunoglobulin V-Type	IGv	0.00	0.95	0.93
Ankyrin repeats	ANK	1.00	0.71	0.76
Serine Proteinase Inhibitors	SERPIN	0.29	0.73	0.72
Armadillo/beta-catenin-like repeats	ARM	0.20	0.92	0.83
Immunoglobulin C-2 Type	IGc2	0.56	0.73	0.71
Histone H4	H4	0.96	0.33	0.62
Histone 2A	H2A	0.88	0.43	0.58
Alpha-amylase domain	Aamy	0.19	0.77	0.75
Gelsolin homology domain	GEL	0.38	0.72	0.68
Rho (Ras homology) subfamily of Ras-like small GTPases	RHO	1.00	0.51	0.55
Src homology 3 domains	SH3	0.44	0.60	0.55
Summary		0.52	0.69	0.69

The same parameters as above were determined in cross-validation experiments to assess the prediction quality in yet unseen protein sequences.

Table C.5.: Cross validation results of nucleotide ligand ipHMMs

Domain	SMART Abr.	Sensitivity	Specificity	Accuracy
Zinc finger	ZnFC2H2	0.67	0.48	0.50
RNA recognition motif	RRM	0.89	0.62	0.64
Actin	ACTIN	0.56	0.61	0.61
Cyclic nucleotide-monophosphate binding domain	cNMP	0.67	0.75	0.75
DNA polymerase X family	POLXc	1.00	0.85	0.85
3’-5’ exonuclease	35EXOc	1.00	0.55	0.55
Histidine kinase-like ATPases	HATPasec	0.56	0.64	0.63
Pancreatic ribonuclease	RNAsePc	0.67	0.51	0.52
Pumilio-like repeats	Pumilio	1.00	0.85	0.87
Serine/Threonine protein kinases, catalytic domain	STKc	0.55	0.82	0.81
DNA polymerase A domain	POLAc	1.00	0.63	0.64
ATPases associated with a variety of cellular activities	AAA	0.50	0.98	0.98
Homeodomain	HOX	0.71	0.58	0.61
Helix-hairpin-helix DNA-binding motif class 1	HhH1	1.00	0.82	0.84
Basic region leucin zipper	BRLZ	0.88	0.71	0.73
C4 zinc finger in nuclear hormone receptors	ZnFC4	1.00	0.58	0.65
Rab subfamily of small GTPases	RAB	1.00	0.63	0.65
Helix-turn-helix lactose operon repressor	HTHLACI	1.00	0.41	0.41
Summary		0.81	0.67	0.68

Table C.6.: Cross validation results of ion ligand ipHMMs

Domain	SMART Abr.	Sensitivity	Specificity	Accuracy
Alkaline phosphatase homologues	alkPPc	0.63	0.49	0.49
Zinc finger	ZnFC2H2	0.33	0.52	0.50
Insulin / insulin-like growth factor / relaxin family	IlGF	0.00	0.67	0.66
Trypsin-like serine protease	TrypSPc	0.00	0.76	0.75
Alpha-lactalbumin / lysozyme C	LYZ1	0.33	0.68	0.67
EF-hand, calcium binding motif	EFh	1.00	0.83	0.86
DNA polymerase A domain	POLAc	1.00	0.64	0.64
Zinc-dependent metalloprotease	ZnMc	0.00	0.71	0.68
ATPases associated with a variety of cellular activities	AAA	1.00	0.95	0.95
Immunoglobulin V-Type	IGv	0.00	0.85	0.83
Eukaryotic homologues of bacterial periplasmic substrate binding proteins	PBPe	1.00	0.57	0.57
Alpha-amylase domain	Aamy	0.00	0.81	0.80
C-type lectin (CTL) or carbohydrate-recognition domain (CRD)	CLECT	0.25	0.82	0.81
Gelsolin homology domain	GEL	0.50	0.70	0.69
Bacterial periplasmic substrate-binding proteins	PBPb	0.17	0.44	0.43
Summary		0.41	0.70	0.69

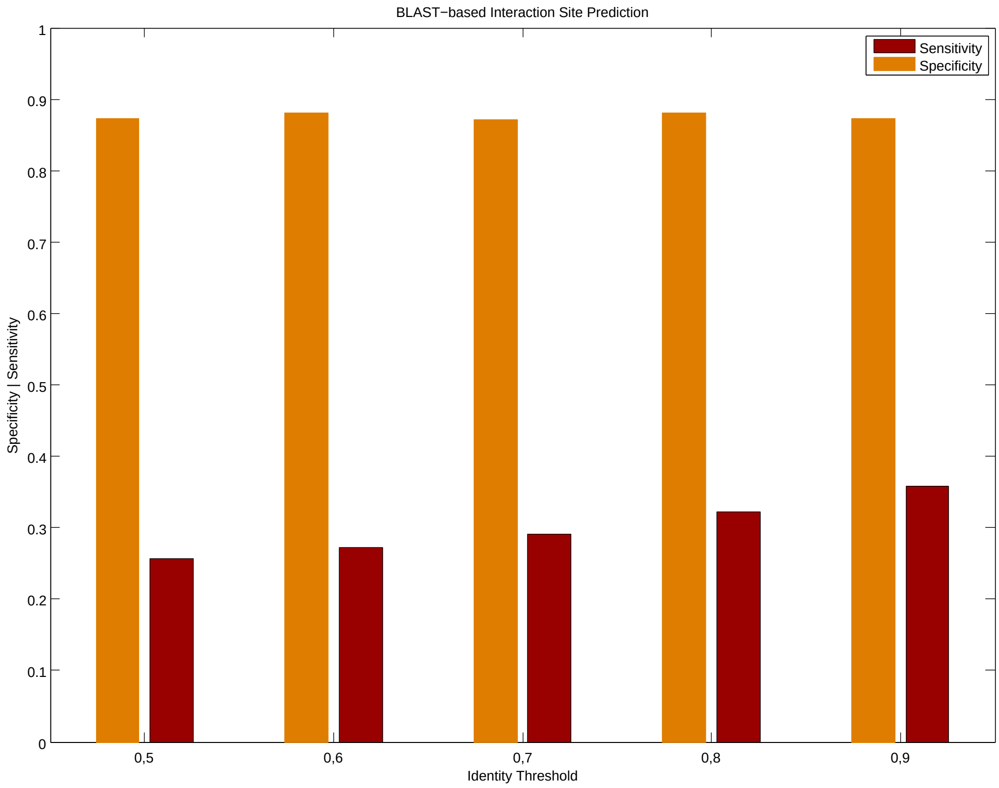
C.3. Receiver operating characteristic

Table C.7.: ROC and Cross-validation Results of ipHMMs

Ligand Group	Domain Name	AUC	Specificity	Sensitivity
Peptides	EF-Hand	0.80	0.75	0.77
	Pancreatic RNAse	0.85	0.90	0.61
	Alkaline Phosphatase	0.81	0.94	0.58
	Ext. Ser-/Thr-type protein kinase	0.96	0.97	0.81
Nucleotides	Pumilio-like repeats	0.98	0.94	0.97
	Pancreatic RNAse	0.90	0.92	0.74
	HTH lactose operon repressor	0.97	0.92	0.88
	C4 zinc finger	0.98	0.88	0.96
Ions	Alkaline Phosphatase	0.96	0.98	0.74
	EF-Hand	0.77	0.79	0.66
	PBPe	0.95	1.00	0.81
	Villin headpiece	0.89	0.96	0.69

Performance of interaction site prediction in selected domains was measured by the area under the ROC curves (AUC) as well as by the trade-off between sensitivity and specificity. An optimal prediction performance would yield a value of 1.0 for the area under the curve.

C.4. BLAST-based interaction site prediction



A simple method for the transfer of knowledge concerning the location of functional sites relies on a BLAST search. This comparative approach was applied to the training data of the ipHMM method. All homologous sequences were detected for every sequence from this set using BLAST (version 2.2.13). Clustering effects of sequences could be observed in the data set due to sequence redundancies within the PDB. In order to establish a better comparability to the HMM-based method, only hits up to an identity threshold ranging from 90 down to 50% and more than 60% coverage were considered as a source of prediction. Gradual thresholding allows for the evaluation of the decrease of predictive power from closely to remotely related sequences. The coverage is calculated as the percentage of the query sequence which is covered by the BLAST alignment to the considered hit. The interaction site profile of each first hit sequence that fits the thresholds was subsequently transferred to the query sequence according to the alignment of the BLAST program. Predictions were restricted to the area which was covered by the BLAST alignment. The prediction quality was evaluated in the same way as described for the ipHMM methodology. As expected, the prediction quality drops with lower sequence identity. For an identity threshold of 90% the method predicted 35.88% of all observed interaction sites correctly while 87.23% of all non-binding positions were precisely detected. In case of 50% maximum identity the sensitivity decreases to 25.65% whereas the specificity remains at the level of 87%. Regarding the more important prediction quality of interaction sites the BLAST-predictor performed substantially worse than ipHMMs.

Contributions

The work presented in this thesis has partially been conducted in collaboration. In the following, contributions to the scientific work are specified in detail:

Chapter 4: “Comparative enterobacterial genomics”

Ulrich Dobrindt and Torben Friedrich acquired enterobacterial genome sequences. Torben Friedrich and Chunguang Liang conducted all-against-all protein sequence alignments of enterobacterial strains. Torben Friedrich clustered the protein similarity data, calculated dispensable and core genome, performed multivariate analysis and the detection of specific protein families in enterobacterial groups. Torben Friedrich further annoted and functionally correlated the obtained specific protein families. Tobias Müller and Ulrich Dobrindt supervised the project and Thomas Dandekar, Jörg Hacker and Sven Rahmann cosupervised it. Torben Friedrich is preparing a manuscript.

Chapter 5: “Development of a diagnostic microarray for clinically relevant enterobacteria”

Torben Friedrich collected genomic sequences. Ulrich Dobrindt and Jörg Hacker set up the project. Ulrich Dobrindt, Tobias Müller, Sven Rahmann and Torben Friedrich designed research. Sven Rahmann performed initial probe selection by longest common factor statistics. Torben Friedrich evaluated the data resulting from longest common factor statistics, selected the final probe set by the application of filtering criteria in MATLAB, performed sequence similarity analysis to human DNA and designed and selected probes for the detection of antimicrobial resistance. Ulrich Dobrindt provided enterobacterial strain for test hybridisations. Torben Friedrich prepared genomic DNA of the bacteria. Test hybridisations were performed by Torben Friedrich at Scienion AG, Berlin. Torben Friedrich experimentally validated indications of antimicrobial resistances by disc diffusion experiments. Torben Friedrich developed the regression model for the diagnostic microarray and conducted cross-validation and ANOVA analysis. Ulrich Dobrindt and Tobias Müller supervised the project and Thomas Dandekar, Jörg Hacker and Sven Rahmann cosupervised it. Torben Friedrich drafted a manuscript, which was revised by Ulrich Dobrindt.

Chapter 6: “Meta-Analysis on diverse gene expression data sets”

Julia C Engelmann and Torben Friedrich selected microarray datasets from GEO database, annotated and normalised them. Julia C Engelmann set up the contrasts and calculated the fold changes and p-values for each contrast. Roland Schwarz performed the clustering, kernel PCA and gene selection in R. Julia C Engelmann interpreted the results biologically, discussed the functions of genes representative for IAA and pathogen related contrasts and performed the analysis in MapMan. Torben Friedrich performed homology modelling of the DUF26 kinases and wrote the respective part of the manuscript. Steffen Blenk performed literature research. Tobias Müller supervised the project. Julia C Engelmann and Roland Schwarz drafted the manuscript. Thomas Dandekar revised the manuscript.

Chapter 7: “Optimal adjustment of HMM topologies”

Christian Koetschan developed the java code to train and decode HMMs, estimate optimal topologies and generate test sequences. Tobias Müller, Christian Koetschan and Torben Friedrich established the maximum likelihood and moment estimators for topology optimisations. Christian Koetschan and Torben Friedrich calculated the moment estimator for the generalised negative binomial distribution. Torben Friedrich analysed and visualised HMM predictions. Tobias Müller and Torben Friedrich supervised the project. Torben Friedrich wrote the paper.

Chapter 8: “Modelling interaction sites in protein domains”

Birgit Pils generated the interaction annotation to SMART domains. Torben Friedrich developed the MATLAB code of ipHMM for training and posterior decoding and sequence generation. Torben Friedrich set up the training data, trained ipHMMs in different ligand categories and performed evaluations. Jörg Schultz provided data on SMART domains in proteins with representation in the OMIM database. Torben Friedrich performed predictions on OMIM protein sequences and analysed the links between interactions and mutations. Tobias Müller supervised the project. Torben Friedrich drafted the manuscript in cooperation with Tobias Müller and Jörg Schultz.

List of Publications associated with this thesis

Torben Friedrich, Birgit Pils, Thomas Dandekar, Jörg Schultz and Tobias Müller (2006) Modelling interaction sites in protein domains with interaction profile hidden Markov models. *Bioinformatics* **22**(23): 2851–2857.

Torben Friedrich, Christian Koetschan and Tobias Müller (in revision) Optimisation of HMM Topologies enhances DNA and Protein Sequence Modelling. *Statistical Applications in Genetics and Molecular Biology*

Julia C. Engelmann, Roland Schwarz, Steffen Blenk, **Torben Friedrich**, Philipp Seibel, Thomas Dandekar and Tobias Müller (2008) Unsupervised Meta-Analysis on Diverse Gene Expression Datasets Allows Insight into Gene Function and Regulation. *Bioinformatics and Biology Insights* **2**: 271–286.

Torben Friedrich, Sven Rahmann, Thomas Dandekar, Jörg Hacker, Tobias Müller and Ulrich Dobrindt (in preparation) Design of a diagnostic microarray for clinically relevant Enterobacteria by a global probe selection strategy

Conference Contributions

Torben Friedrich, Birgit Pils, Thomas Dandekar, Jörg Schultz and Tobias Müller (2005) Interaction Profile Hidden Markov Model - A Method for Interaction Site Prediction. GCB, Hamburg, Germany, October 5-7 (*Poster*)

Torben Friedrich, Sven Rahmann, Thomas Dandekar, Jörg Hacker, Ulrich Dobrindt and Tobias Müller (2007) Explorative analysis of genomic differences in enterobacteria. Ecoli 2007 EMBO-FEMS-LEOPOLDINA Symposium, Kloster Banz, Staffelstein, Germany, October 9-12 (*Poster*)

Torben Friedrich, Chunguang Liang, Jörg Hacker, Thomas Dandekar, Ulrich Dobrindt, Tobias Müller (2008) Comparative Proteomics of Enterobacteria using multivariate Analysis. Genomes 2008, Institute Pasteur, Paris, France, April 8-11 (*Poster*) and FORINGEN Symposium, Herrsching am Ammersee, Germany, April 3-4 (*Poster*)

Torben Friedrich, Birgit Pils, Thomas Dandekar, Jörg Schultz and Tobias Müller (2006) Modelling interaction sites in protein domains with interaction profile hidden Markov models. Bioinformatics Symposium, University of Würzburg, Germany, July 27 (*Oral Presentation*)