TECHNISCHEN UNIVERSITÄT MÜNCHEN

Professur für Bioinformatik

COMPUTATIONAL METHODS FOR THE PREDICTION OF BACTERIAL PATHOGEN-HOST PROTEIN PROTEIN INTERACTIONS

MARC-ANDRÉ JEHL

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. W. Liebl

Prüfer der Dissertation:

1. Univ.-Prof. Dr. D. Frischmann
2. Univ.-Prof. Dr. Th. Rattei (Universität Wien/Österreich)

Die Dissertation wurde am 08.10.2015 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 16.12.2015 angenommen.

Danksagung

An dieser Stelle möchte ich all jenen danken, die mich bei der Erstellung meiner Doktorarbeit unterstützt haben, insbesondere:

Herrn Prof. Hans-Werner Mewes für seine Zustimmung dazu, dass ich mit meiner Promotion im Fachbereich Bioinformatik in München beginnen konnte.

Meinem Doktorvater Prof. Dmitrij Frishman für seinen fachlichen Beistand und seine Mithilfe bei der Abwicklung formaler Angelegenheiten.

Meinem Betreuer Prof. Thomas Rattei, der für mich als ständiger Ansprechpartner mit seinem fachlichen Rat stets eine große Hilfe war.

Allen Kollegen am Lehrstuhl für Genomorientierte Bioinformatik in Weihenstephan, insbesondere Roland Arnold, Patrick Tischler und Stefka Tyanova für die vielen erkenntnisreichen Diskussionen.

Desweiteren den Kollegen am Lehrstuhl für Computational Systems Biology (CUBE) an der Universität Wien für das interessante wissenschaftliche Umfeld, speziell Raheleh Sheibani-Tezerji und Dmitrij Turaev für die gute Zusammenarbeit in unterschiedlichen Projekten.

Dr. Agathe Subtil und Lena Gehre vom Institut Pasteur in Paris für die experimentelle Überprüfung wesentlicher Ergebnisse meiner Arbeit.

Leonie Corry, Sekretärin am Lehrstuhl für Genomorientierte Bioinformatik in Weihenstephan und Gabriele Kaindl, Sekretärin am CUBE in Wien für deren unermüdlichen Einsatz beim meistern der Bürokratie deutscher und österreichischer Universitäten.

Meinem Freund und Kollegen Thomas Weinmaier, der zeitgleich mit mir seine Doktorarbeit begann und ein wichtiger Weggefährte war.

Meinen Eltern für die ständige emotionale Unterstützung und ihr großes Interesse an meiner Forschungsarbeit.

Nicht zuletzt meiner langjährigen Lebensgefährtin Anna Beifuß: Ich danke Dir, dass Du in dieser manchmal anstrengenden Zeit immer für mich da warst!

Abstract

Bacteria are part of almost any biological process on earth, often exhibiting symbiotic or pathogenic life-styles. Both pathogens and symbionts had to develop molecular strategies to manipulate host functions, cope with the hosts immune response and show parallel trends in genome evolution. Despite intense research, pathogenic infections are worldwide still one of the major health issues. For a molecular understanding of bacterial virulence and the development of effective diagnostics and therapy, e.g. based on novel drug targets, it is necessary to identify the key players in microbe-host interactions. Essential to the molecular cross talk between bacterium and host thereby is the transport of bacterial proteins, so called effectors, into the eukaryotic host cell. Within the host cytosol, effector proteins are capable of directly altering host cellular functions to the pathogens/symbionts advantage. Experimental identification of effector proteins remains a challenge and the number of characterized effectors is limited. Existing *in silico* prediction methods rely on the modeling of secretion signals in the amino acid sequence of effector proteins. It has been shown for several organisms that effector proteins not only have functions similar to eukaryotic proteins but that they possess similarity to eukaryotic proteins also on the sequence and structural level. Nevertheless, current effector prediction methods are still limited and unspecific. The aim of this work was to advance bioinformatics approaches for the investigation of pathogen-host protein interactions.

In this work, I present a taxonomically universal, signal independent effector prediction approach. By collaboration partners it was shown in an experimentally study that this method predicts the effectome of the intra-cellular human pathogen *Chlamydia trachomatis* with high quality. To make predicted effectomes easily accessible even to scientists without particular bioinformatic skills, I devised a web resource for the prediction of secreted bacterial proteins. The Effective web portal provides precalculations for all fully sequenced genomes of bacterial pathogens and symbionts. Users also have the possibility to predict effector candidates on their own in user defined sequence data by a state-of-the-art set of prediction methods. In this thesis I have been also working on the mainly unknown host sided part of pathogen-host interactions and thereby explored the potential of domain domain prediction methods to predict pathogen-host protein interactomes. By integrating gene expression data as well as information on the host protein interaction network, I was able to reduce the number of predicted interactors to an experimentally traceable set of candidates. Motivated by the ever increasing availability of completely sequenced bacterial genomes, I applied the newly developed methods beyond the prediction of pathogen-host interactomes and explored their potential to speed up recognition of potential bacterial threats, an important challenge in microbial diagnostics. I could show that to some extent the repertoire of effector candidates in the genome of a bacterial organism enables prediction of a host-interacting phenotype.

Zusammenfassung

Bakterien spielen eine Rolle in fast jedem biologischen Prozess auf der Erde. Dabei gehen sie oft symbiotische als auch pathogene Wechselbeziehungen zu ihrer Umwelt ein. Sowohl Krankheitserreger als auch Symbionten mussten molekulare Strategien entwickeln um die Immunantwort und Stoffwechselfunktionen der Wirtszelle zu manipulieren und zeigen parallele Trends in der Genomevolution. Trotz intensiver Forschung sind pathogenen Infektionen immer noch eines der wichtigsten gesundheitlichen Probleme weltweit. Für ein tieferes Verständnis der bakteriellen Virulenz und für die Entwicklung von wirksamen Therapiemethoden, zum Beispiel basierend auf neuartigen Wirkstoffzielen, ist es notwendig, die wichtigsten molekularen Akteure in der Interaktion zwischen Bakterium und Wirtszelle zu identifizieren. Wesentlich für diese molekulare Wechselwirkung ist der Transport von bakteriellen Proteinen, sogenannten Effektoren, in die eukaryontische Wirtszelle. Innerhalb der Wirtszelle sind Effektorproteine in der Lage, die Funktionen der Wirtszelle direkt zugunsten des Krankheitserregers bzw Symbionts zu verändern. Die experimentelle Identifizierung von Effektorproteinen ist nach wie vor eine Herausforderung und die Anzahl der bekannten Effektoren begrenzt. Existierende computergestützte Vorhersageverfahren beruhen auf der Modellierung von Sekretionssignalen in der Aminosäuresequenz der Effektorproteine. Für mehrere Organismen wurde gezeigt, dass Effektorproteine nicht nur ähnliche Funktionen wie eukaryontische Proteine übernehmen, sondern dass sie auch auf der Sequenz- und Strukturebene Ähnlichkeit zu eukaryotischen Proteinen besitzen. Dennoch sind die aktuellen Effektor-Vorhersageverfahren noch zu begrenzt und unspezifisch. Ziel dieser Doktorarbeit war es, die Entwicklung bioinformatischer Ansätze zur Untersuchung der Protein-Wechselwirkungen zwischen Bakterium und Wirtszelle voranzutreiben.

In dieser Arbeit stelle ich einen taxonomisch universellen, von Sekretionssignalen unabhängigen Ansatz zur Effektor-Vorhersage vor. Durch die experimentelle Arbeit unserer Kooperationspartner konnte gezeigt werden, dass diese Methode das Effektom des intrazellulären Humanpathogens *Chlamydia trachomatis* mit hoher Qualität vorhersagt. Um die vorhergesagten Effektome auch für Wissenschaftler ohne bioinformatische Kenntnisse zugänglich zu machen, entwickelte ich eine Web-Ressource für die Vorhersage von sekretierten bakteriellen Proteinen. Das ËffectiveẄeb-Portal bietet vorberechnete Effektomanalysen für alle vollständig sequenzierten Genome bakterieller Pathogene und Symbionten. Benutzer der Webseite haben ausserdem selbst die Möglichkeit, Effektor-Kandidaten in eigenen bereitgestellten Proteinsequenzen mit einem umfassenden Auswahl aktueller Verfahren vorherzusagen. Ich habe mich in dieser Doktorarbeit bei der Untersuchung der Wechselwirkung zwischen Bakterium und Wirt auch dem größtenteils unbekannten Bereich auf Seiten der Wirtszelle gewidmet. Dabei untersuchte ich das Potenzial von Verfahren zur Vorhersage von Interaktionen zwischen Proteindomänen zur Erforschung des Proteininteraktionsnetzwerks von Bakterium und Wirtszelle. Durch die

Integration von Genexpressionsdaten sowie struktureller Informationen über die Topologie des Host-Protein-Interaktionsnetzwerks war ich in der Lage, die Anzahl der vorhergesagten Interaktoren zu einem experimentell überprüfbaren Set von Kandidaten zu reduzieren. Durch die zunehmende Verfügbarkeit von vollständig sequenziert bakteriellen Genome motiviert, wandte ich die neu entwickelten Methoden über die Grenzen der Vorhersage von Host-Pathogen Interaktomen an und untersuchte ihr Potenzial zur Erkennung potenzieller bakterieller Krankheitserreger, einer wichtigen Herausforderung in der mikrobiellen Diagnostik. Ich konnten zeigen, dass in einem gewissen Ausmaß das Repertoire von Effektor-Kandidaten im Genom eines Bakteriums die Vorhersage eines Host-interagierenden Phänotyps ermöglicht.

Publication record

[1] Raheleh Sheibani-Tezerji, Muhammad Naveed, **Marc-André Jehl**, Angela Sessitsch, Thomas Rattei, and Birgit Mitter. The genomes of closely related pantoea ananatis maize seed endophytes having different effects on the host plant differ in secretion system genes and mobile genetic elements. *Frontiers in Microbiology*, 6, 2015.

[2] Ilias Lagkouvardos, **Marc-André Jehl**, Thomas Rattei, and Matthias Horn. Signature protein of the PVC superphylum. *Applied and Environmental Microbiology*, 80(2):440–445, January 2014.

[3] **Marc-André Jehl**, Roland Arnold, and Thomas Rattei. Effective—a database of predicted secreted bacterial proteins. *Nucleic Acids Research*, 39(suppl 1):D591–D595, January 2011.

[4] Stefan Goetz, Roland Arnold, Patricia Sebastian-Leon, Samuel Martin-Rodriguez, Patrick Tischler, **Marc-André Jehl**, Joaquin Dopazo, Thomas Rattei, and Ana Conesa. B2g-FAR, a species-centered GO annotation repository. *Bioinformatics*, 27(7):919–924, January 2011.

[5] Monika R. Nuk, Andreas Reisner, Martina Neuwirth, Katrin Schilcher, Roland Arnold, **André Jehl**, Thomas Rattei, and Ellen L. Zechner. Functional analysis of the finO distal region of plasmid r1. *Plasmid*, 65(2):159–168, March 2011.

[6] Silvia Lang, Karl Gruber, Sanja Mihajlovic, Roland Arnold, Christian J. Gruber, Sonja Steinlechner, **Marc-André Jehl**, Thomas Rattei, Kai-Uwe Froehlich, and Ellen L. Zechner. Molecular recognition determinants for type IV secretion of diverse families of conjugative relaxases. *Molecular Microbiology*, 78(6):1539–1555, December 2010.

[7] Roland Arnold, **André Jehl**, and Thomas Rattei. Targeting effectors: the molecular recognition of type III secreted proteins. *Microbes and Infection*, 12(5):346–358, May 2010.

[8] Thomas Rattei, Patrick Tischler, Stefan Goetz, **Marc-André Jehl**, Jonathan Hoser, Roland Arnold, Ana Conesa, and Hans-Werner Mewes. SIMAP—a comprehensive

database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Research*, 38(suppl 1):D223–D226, January 2010.

[9] Valerie Eichinger, Thomas Nussbaumer, Alexander Platzer, **Marc-André Jehl**, Roland Arnold, and Thomas Rattei. Effective - updates and novel features for a better annotation of bacterial secreted proteins and type III, IV, VI secretion systems. *Nucleic Acids Research, in revision*, 2015.

Contents

Danksagung I

Abstract III

Zusammenfassung V

Publication record VII

Table of Contents IX

1 Introduction 1

1.1 Bacterial pathogens and host-associated bacteria 4

1.1.1 Chlamydiae 4

1.1.2 Legionella spp 6

1.2 Identification of bacterial pathogens 6

1.2.1 Experimental methods to identify bacterial pathogens 7

1.2.2 Computational methods for the predicition of bacterial phenotypes 8

1.2.3 Public resources collecting information on bacterial phenotypes . . 9

1.3 Bacterial protein secretion 10

1.3.1 General secretion (Sec) pathway and twin-arginine (Tat) pathway . 11

1.3.2 Type I secretion system (T1SS) 11

1.3.3 Type II secretion system (T2SS) 12

1.3.4 Type III secretion system (T3SS) 13

1.3.5 Type IV secretion system (T4SS) 15

1.3.6 Type V secretion system (T5SS) 17

1.3.7 Type VI secretion system (T6SS) 19

1.3.8 Type VII secretion system (T7SS) 19

1.4 Bacterial effector proteins 20

1.4.1 Effector similarity to eukaryotic proteins 20

1.4.2 Eukaryotic-like domains in effector proteins of *Legionella pneumophila* 21

1.5 Computational prediction of effector proteins 21
1.5.1 Signal based methods 22
1.5.2 Homology based approaches 25
1.5.3 Genomic and function based approaches 25
1.6 Prediction of pathogen host protein-protein interactions 25
1.6.1 Interolog based methods 26
1.6.2 Sequence based approaches 26
1.6.3 Domain-based prediction 27
1.7 Concepts used in this work 29
1.7.1 Machine learning 29
1.7.2 Statistical measures 31
1.7.3 Web portal/resource development 33

2 A secretion system independent approach to predict bacterial effector proteins 35
2.1 Motivation 35
2.2 Exemplary analysis of the eukaryotic F-Box protein domain in bacteria . . 36
2.3 Setup of a genome repository for comparative analysis 38
2.3.1 Phenotype annotations of completely sequenced bacterial genomes . 40
2.3.2 Bacterial contamination of eukaryotic genome sequences 47
2.3.3 Criteria for the prediction of a functional T3SS machinery 50
2.4 Phenotype specificity of eukaryotic-like protein domains in bacteria 56
2.4.1 Discussion 59
2.5 Identification of eukaryotic-like domains enriched in pathogenic/symbiotic bacteria 59
2.5.1 Methods 59
2.5.2 Results 60
2.5.3 Reported experimental validation of effector candidates predicted in the genomes of *Chlamydiae sp.* 62
2.5.4 Discussion 63

3 The Effective web portal - Development of a comprehensive resource for the prediction of secreted bacterial proteins 67
3.1 Motivation 67
3.2 Functionality and requirements 68

3.3 Concepts and implementation . . . 68

3.3.1 Software model of the Effective web portal . . . 69

3.3.2 Combination of complementary effector prediction approaches . . . 71

3.3.3 Automatic updates of the Effective database . . . 73

3.3.4 Implementation of basic technical aspects . . . 73

3.3.5 The Effective web portal provides precalculated effector candidates for all sequenced pathogens and symbionts . . . 74

3.3.6 Use case: Selection of effector candidates for pulldown analysis to investigate virulence strategies of *Legionella pneumophila* . . . 75

3.3.7 The Effective web portal allows on-the-fly prediction of effector proteins in user provided sequence data . . . 80

3.3.8 Use case: Comparative genomics analysis of three *Pantoea ananatis* strains causing diverse host phenotypes . . . 81

3.4 Discussion and outlook . . . 83

4 Application of a domain-based approach to predict pathogen-host interactomes 85

4.1 Motivation . . . 85

4.2 Domain-domain interaction candidates of the *C. trachomatis* effectome . . 87

4.2.1 Introduction . . . 87

4.2.2 Material and methods . . . 88

4.2.3 Results . . . 89

4.2.4 Discussion . . . 90

4.3 Prediction of the chlamydial pathogen-host interactome in human . . . 93

4.3.1 Introduction . . . 93

4.3.2 Material and methods . . . 94

4.3.3 Results . . . 95

4.3.4 Discussion . . . 98

5 Classification of bacterial phenotypes based on genomic features 101

5.1 Motivation . . . 101

5.2 Characteristics of eukaryotic-like protein domains in bacteria of diverse phenotype . . . 103

5.2.1 Diversity of eukaryotic-like domain profiles between closely related bacterial genomes . . . 103

5.2.2 The pathogenicity probability of a eukaryotic-like protein domains . 106

5.2.3 Predictive power of a genomic pathogenicity probability derived from eukaryotic-like domains 109
5.3 Pathofier - a machine learning approach to predict bacterial phenotypes based on effector protein candidates 112
5.3.1 Material and methods 112
5.3.2 Results 115
5.3.3 Discussion 118

6 Summary and conclusions **121**

7 Bibliography **125**

8 Appendix **159**

List of Figures **183**

List of Tables **187**

1 Introduction

Pathogenic infections are worldwide one of the major health issues [306]. Despite intense research and development of new anti-microbial treatments, the number of pathogenic infections is still severe and remains unchanged over recent years, compare graph 1.1.

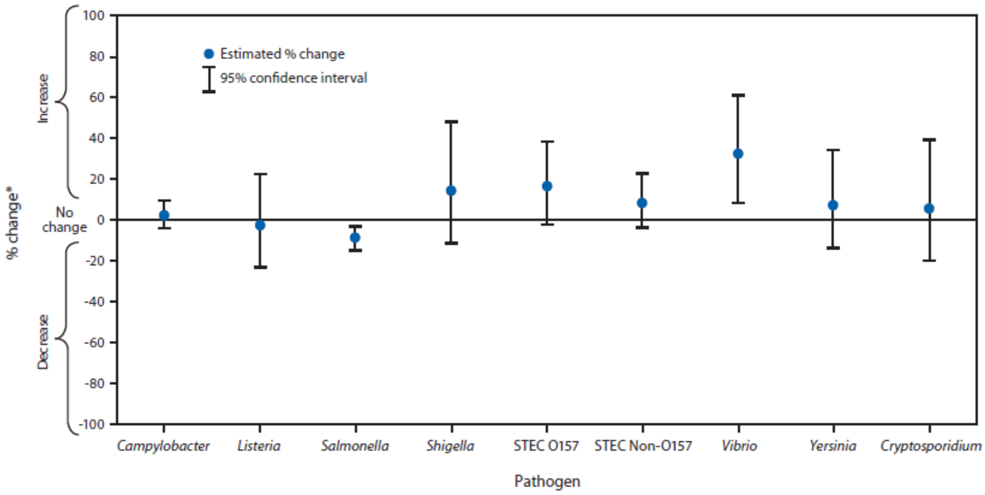


Figure 1.1: ***Estimated change of incidences of pathogenic infections in the United States for 2012.*** *Shown is the estimated change in incidence of laboratory-confirmed pathogenic infections in the United States in 2012, compared with average annual incidence during 2006-2008 as reported by the Centers for Disease Control and Prevention (CDC) [51].*

For example, chronic lower respiratory disease surpassed stroke as third leading cause of death in the United States, according to reports of the American Lung Association. Beside persistent problems, public health faces various new threats. Acquisition of antibiotic resistance is observed for some strains of the most notorious agents [128, 206]. High rates in the spread of resistance genes throughout the genomes of diverse pathogens make experts rise the alarm of an upcoming post-antibiotic area [150]. Further challenges are the revival of supposedly contained pathogens [192, 270], observed in the emergence of Mycobacterium tuberculosis in eastern Europe and other parts of the developing world

[30]. Novel agents put public health at risk in situations like the epidemic spread of Enterohaemorrhagic Escherichia coli (EHEC) in northern parts of Germany in 2011 [242]. In figure 1.2, the diversity of infecting agents is illustrated in an overview of bacterial threats to human health.

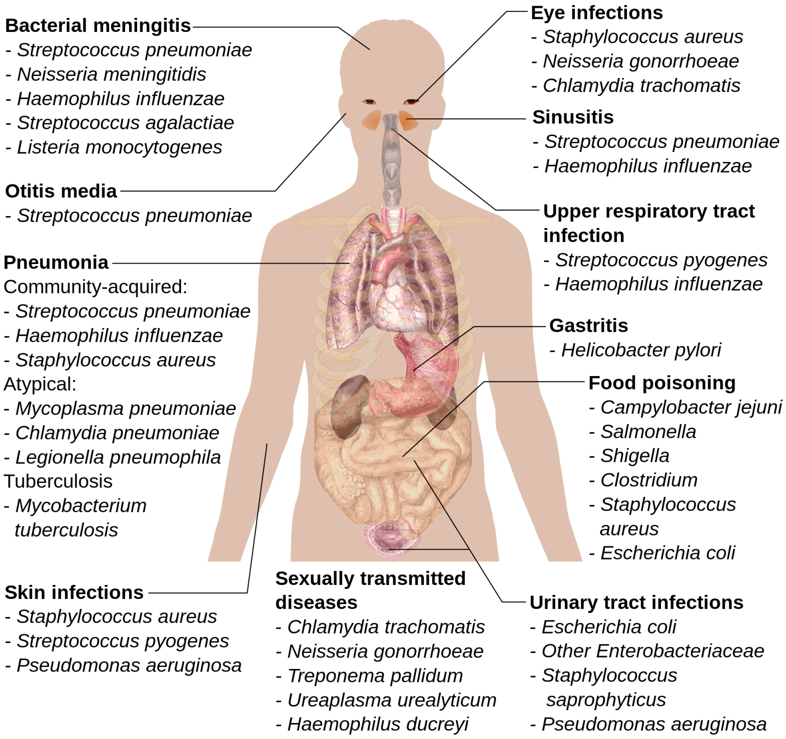


Figure 1.2: ***Overview of bacterial infections in human.*** *Shown are known bacterial agents and tissue tropism upon infection in human [129].*

Increasing efforts on the development of pathogen specific vaccines are necessary to be able to keep up with the challenges of a future, in which current antibiotic treatments will have lost much of their power [150]. For the design of novel efficient methods, detailed knowledge about bacterial virulence is essential. Pathogens are able to initiate and maintain infection by circumventing the hosts immune system while at the same time limiting the impact on the host cell. Essential for triggering these molecular processes is the secretion of bacterial proteins. Bacteria secrete proteins into the host cell, so called effectors, to alter host pathways in their favor [108]. Once secreted into the host cell, these effector proteins can influence a broad range of functions. Effectors enable disruption of the host immune defense e.g. by hijacking apoptotic pathways [110]. The effector NleB inhibits NF-kappaB-dependent host innate immune responses upon infection by several attaching/

effacing (A/E) human pathogens [113]. Effector SipA of *Salmonella typhimurium* permits membrane ruffling and rearrangement of the actin cytoskeleton [323]. Transcription activator–like (TAL) effectors of plant pathogenic bacteria contain a modular DNA binding domain to manipulate transcription of host proteins [37]. Also symbionts interact with the protein network of the host cell by secreting effector proteins. Pathogenic and symbiotic bacteria therefore depend on similar genomic features to implement their host-interacting lifestyle [275]. Many mechanisms that shape host interactions of symbionts are factors that also contribute to the processes involved in pathogenic infections [217]. Altering host functions is a major virulence strategy determining bacteria-host interactions in a variety of pathogens and symbionts [184]. Studies of multihost bacterial pathogens like *Pseudomonas* spp suggest the existence of universal bacterial virulence mechanisms that are highly conserved across phylogeny [234]. Functional similarity to host proteins is a commonly observed theme in effector proteins [97]. Mimicking of host cell proteins is observed by effector proteins of several host-associated bacteria. E.g. the human pathogen *Legionella pneumophila* is able to modulate host functions by the secretion of eukaryotic-like proteins. It has been shown for several organisms that effector proteins not only have functions similar to eukaryotic proteins but that they possess similarity to eukaryotic proteins also on the sequence and structural level [118, 7, 4]. These effectors contain domain signatures that are characteristic for and mainly observed in eukaryotic proteins. The domains/proteins are called eukaryotic-like domains (ELDs)/proteins (ELPs). Revealing the arsenal of bacterial effector proteins is key to the understanding of bacterial virulence and bacterial interaction with the host [309]. Experimental identification and characterization of effector proteins remains a challenge and the number of characterized effectors still limited. *In silico* prediction of effector proteins offers the possibility to select for high-quality experimental candidates. Existing methods rely on the prediction of secretion signals in the amino acid sequence of effector proteins. Momentarily, out of seven bacterial secretion systems, only for the Type III secretion system a signal-peptide can be predicted with good accuracy. For a comprehensive prediction of bacterial effector proteins, alternative approaches are needed. In this work, a function-based prediction approach based on the identification of eukaryotic-like protein domains is developed. Furthermore, it is evaluated, to which extent this extended repertoire of effector candidates in the genome of a bacterial organism enables recognition of its host-associated phenotype.

1.1 Bacterial pathogens and host-associated bacteria

Bacterial organisms are part of almost every important biological process and shape the ecological systems on earth. Many bacteria have developed symbiotic or parasitic lifestyles and live in close relationship with diverse hosts. For example, several bacterial species can be found as endosymbionts in amoeba, including *Chlamydiae* and *Bacteroidetes* [253]. Some bacteria are important agents of human and animal infections causing a wide range of diseases as *Yersinia* [310], *Salmonella* [125], *Pseudomonas* [83], *Listeria* [91] and *Helicobacter* [66]. Many of these pathogenic and symbiotic bacteria exhibit a facultative or obligate intra-cellular life-style. The compositions of bacterial genomes can change rapidly through a variety of processes including genome rearrangements and horizontal gene transfer [127].

The lifestyles of both pathogens and symbionts are similar as they both rely on interaction with a host system. Strong similarities can also be observed on the genomic level [217]. Pathogens and symbionts had to develop molecular strategies to alter host functions, cope with the hosts immune response and show parallel trends in genome evolution [275]. Pathogens and symbionts are described as host-associated or of host-interacting phenotype, in contrast to the non-pathogenic phenotype that includes all non-pathogenic bacterial organisms without any observed host-interaction in nature.

Evolutionary and ecological principles shaping pathogenic as well as mutualistic microbe-host relationships are important to the understanding of disease [77]. Two human pathogens are presented in detail as they serve as model organism in this study.

1.1.1 Chlamydiae

Chlamydiae are Gram-negative obligate intracellular bacteria. The phylum of Chlamydiae is very old and comprises several different species [63, 203]. Its evolutionary history makes it a model for the development of bacterial pathogenicity. While individual species and strains usually are infecting a limited set of hosts, the overall host coverage of Chamydiae is wide, ranging from amoebae [133] to human [106].

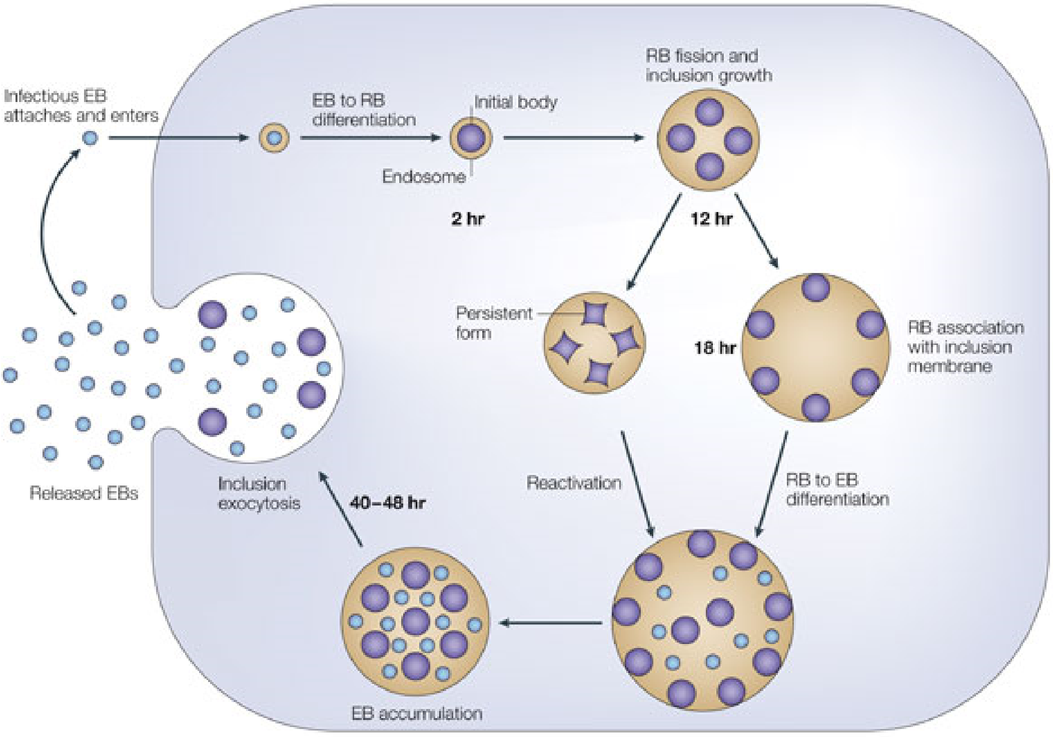


Figure 1.3: ***Illustration of biphasic developmental lifecycle of Chlamydia trachomatis.*** *Uptake of infectious, metabolically inactive elementary body (EB) by mucosal epithelial cells. An endosomal membrane is created to form an inclusion vacuole in which the EB transforms into a metabolically active reticulate body (RB). After replication, the RBs transform back into EBs and are released from the vacuole to infect surrounding host cells [43].*

Clinical and economic relevance of chlamydia infections are high, several diseases are caused by chlamydial agents in human, e.g. pelvic inflammatory disease [243], infection of the urinary tract [211], the respiratory tract [123] and the eye [245]. Chlamydia infection is the most frequent sexually transmitted diseases in developed countries [13]. Chlamydiae have a complex bi-phasic developmental cycle and form separate inclusions in the host cells, compare figure 1.3. Due to the intra-cellular lifestyle, *Chlamydia* rely on the host to survive and lack several biosynthetic pathways leading to genome reduction [22]. The obligate intracellular lifestyle of *Chlamydiae* makes a cultivation in the lab challenging and hinders experimental identification and characterization of novel effector proteins. This strengthens the role of computational biology for chlamydia research.

1.1.2 Legionella spp

Legionella spp. are aerobic, Gram-negative bacteria infecting environmental protozoa and also causing severe infections in human. Inhaled *Legionella pneumophila* replicates in macrophages of the innate immune system, leading to potentially fatal pneumonia (Legionnaires disease). *L. pneumophila* replicates in a privileged, membrane-bound compartment, the 'Legionella-containing vacuole' (LCV). LCVs acquire secretory (Arf1, Rab1 and Rab8) as well as endosomal (Rab7 and Rab14) small GTPases, which allow the pathogen vacuole to communicate with the secretory and endocytic vesicle trafficking pathway [294]. Key virulence factor of *L. pneumophila* are multiple copies and classes of Type IV secretion systems (T4SS) that secrete more than 250 effector proteins into the eukaryotic host cell [137], compare genome map 1.4. This allows for a robust and redundant modulation of host pathways. When lacking single effector proteins, most *L. pneumophila* mutant strains grow at wild-type rate in host cells [174]. Adaptation to the host environment and exploitation of host cell functions are enabled by an extended set of eukaryotic like proteins (ELP). Many of these ELP were shown to modulate host cell functions to the pathogen's advantage. For example, the eukaryotic-like methyltransferase RomA is secreted by *Legionella pneumophila* to target the host cell nucleus and methylate histone H3 to alter gene expression [241]. The motifs in these proteins that were found predominantly in eukaryotes are ankyrin repeats, SEL1 (TPR), Set domain, Sec7, serine threonine kinase domains (STPK), U-box, and F-box protein domains. *Legionella pneumophila* was extensively analyzed for eukaryotic-like proteins and serves as a model organism for this pathogenic strategy [119].

1.2 Identification of bacterial pathogens

Undiscovered pathogens are supposed to be the causitive agents of many human diseases [237, 173]. Recognition of the causative agents of pathogenic threads is crucial to an efficient medical response. Rapid methods to classify novel bacteria into harmful pathogens, potentially harmful symbionts and harmless non-pathogens could improve medical treatment as well as containment of infections [238].

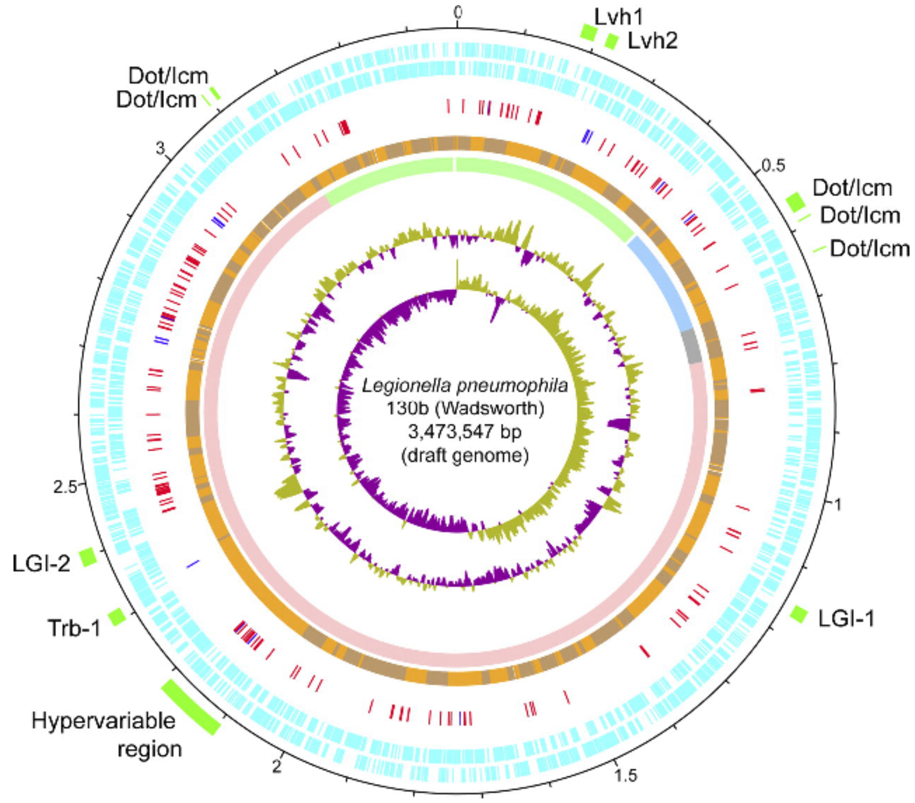


Figure 1.4: ***Circular map of the L. pneumophila 130b draft genome.*** *From the outside in, the first circle shows the positions of T4SSs and a hypervariable region. The second circle is the scale (in Mbp). The third and fourth circles show the predicted CDSs transcribed clockwise and anti clockwise, respectively. The fifth circle shows known Legionella T4SS effector genes and their paralogues (colored red) and putative effectors (dark blue). The sixth circle shows the 159 contigs of the draft genome, circle 7 shows the 4 scaffolds that link 145 of these contigs. Circle 8 shows the GC content (percent), the innermost circle is a plot of GC deviation* $((G-C)/(G+C))$ *[254].*

1.2.1 Experimental methods to identify bacterial pathogens

Currently several experimental procedures for the detection of pathogens exist: Isolation and bacteriological culturing of a pathogenic agent is the traditional, rather insensitive approach to determining the cause of an infection [215]. The enzyme-linked immunosorbent assay (ELISA) is a method which relies on the recognition of antibodies generated in a pathogen specific host response to infection [169]. Furthermore, biosensors involving bacterial phages or the respective bacteriophage receptor binding proteins as recognition elements allow for very specific detection of a growing number of well-known pathogens [264]. Polymerase chain reaction (PCR) techniques provide the possibility to detect bac-

terial agents based on their DNA sequence upon generating pathogen specific primers [172, 167].
Despite great progress and ongoing development of new experimental strategies in the field, the established methods in practice are still expensive, time consuming and require trained laboratory personnel to successfully execute these tests. Costs increase rapidly in the effort to adjust these techniques to the identification of novel bacterial agents. In practice, specific PCR based tests are often missing and the development of suitable primers is laborious [40, 182].

1.2.2 Computational methods for the predicition of bacterial phenotypes

In the face of these challenges, the advances in next-generation sequencing (NGS) have a huge impact on microbial diagnostics [78]. The sequencing time for complete bacterial genomes declines rapidly to be soon in the range of hours [59]. This almost instant availability of genomic information about a pathogenic threat is expected to transform current approaches in clinical microbiology [101]. Pathogen detection strategies which include sequence data of bacterial genomes and *in silico* analysis have many advantages compared to the traditional solely experimental methods. Sequence-based bioinformatics approaches in microbial research have the potential to enhance or replace many complex experimental methods. Beside others, essential tasks like the identification of bacterial species in an isolate as well as testing of the properties of a novel agent, such as antibiotic resistance and virulence, profit the most [79].
Existing computational approaches to pathogen detection, e.g. from metagenomic data samples, focus on effective methods to identify known pathogens: taxonomic mapping, whole genome assembly, sequence composition analysis and statistical frameworks. The assignment of sequence reads to known microorganisms via taxonomic mapping was first used in the work of Huson et al. [139]. These approaches try to determine the last common ancestor or taxonomic group for each sequence read. Bhaduri et al. identifies the origin of microbial sequences in sequence samples by whole genome assembly [32]. Compositional analysis searches for patterns in the sequence data, taking into account GC-content, k-mer frequency and taxonomic markers to assign input sequences to known populations or clades [216]. In their Pathoscope approach, Francis et al. identify the source genome by using a Bayesian statistics to map sequence reads against a database of known agents [102].

These approaches work on genomes or metagenomic read sequences. In medical applications, patient samples are commonly metagenomic samples, containing sequence reads of different sources. Identification of bacterial organisms in the probe is performed by identifying a low number of species specific reads and mapping them to a database of known agents. Closely related species/strains are hard to distinguish. E.g. the pathogenic E.coli EHEC is nearly identical to several non-pathogenic E.coli. While, in the recognition of characterized pathogens, these methods achieve good performance, they are not able to classify novel genotypes. Due to the implicit limitations of the reference database mapping, novel agents with weak or no similarities to genotypes part of the reference databases stay undetected. The large majority of bacterial microorganisms have not been characterized yet [75]. To develop a methods that overcome the boundaries of mapping-based approaches is critical. One if not the major issue is to clarify whether a particular sample contains harmless non-pathogenic bacteria or comprises harmful novel pathogens and symbionts putting a potential host at risk. With the current approaches, this basic question stays mostly unsolved. Complementary methods using genomic data for a direct classification of bacterial phenotypes are needed.

De novo assembly of prokaryotic genomes is limited and usually results in tens of contigs which can not be joined further for a completely closed genome sequence [153]. Especially in the case of unknown bacterial agents, approaches that rely on the identification of genomic phenotype specific features have an important advantage. They work for completely assembled genomes and contigs as well.

Pathogens encode a large repertoire of virulence-associated genomic features that enable infection. They possess an arsenal of secreted effector proteins to alter host functions in the cell, as well as other virulence factors, e.g. encoding means of transportation to transfer effectors into the host cytoplasm.

1.2.3 Public resources collecting information on bacterial phenotypes

Reliable evidence for the phenotype of bacterial organisms comes from experiments. Several public resources collecting genome projects also collect specific metadata regarding observed characteristics of the organism at hand. Resources providing information on phenotypic annotations of bacterial organisms are BacMap [68], GOLD [212] and NCBI phenotypes list [303].

1.3 Bacterial protein secretion

Bacteria manipulate their environment and interact with host cells by the secretion of proteins, so called “effectors”. Effectors fulfill their function outside the prokaryotic cell and interact with proteins within the host cell cytoplasm, serving manifold functions. Transport of effector proteins from the bacterial cell into the host cytoplasm is regulated by secretion systems [52].

Besides the Sec/Tat general secretion pathway, seven different secretion systems are known in bacteria, compare figure 1.5. These systems have only little homology between each other. They show considerable differences and are associated with diverse functions. Six secretion pathways were identified in Gram-negative bacteria, named Type I to Type VI secretion systems (T1SS to T6SS). Gram-negative bacteria contain an inner and outer membrane with a periplasmic space in between. Secretion systems III, IV and VI are able to transport proteins through both membranes. Furthermore they establish direct transport of effector proteins into the host cell and are especially associated with bacterial virulence. In Gram-positive bacteria that contain one single membrane, protein secretion by the two-arginine (Tat) or Sec pathway leads by default to release into the extra-cellular medium. Direct secretion into eukaryotic cells is enabled by the Type VII secretion system which is exclusive to Gram-positive bacteria and specific for virulence [2]. Many bacteria do not secrete effectors by one system alone but have multiple secretion systems encoded in their genome [36]. Even for some of the few characterized effector proteins, details about the secretion mechanism remain unknown, compare [202].There is still limited knowledge for most secretion systems about the respective core set of molecular components that are necessary for a fully functional secretion machinery.

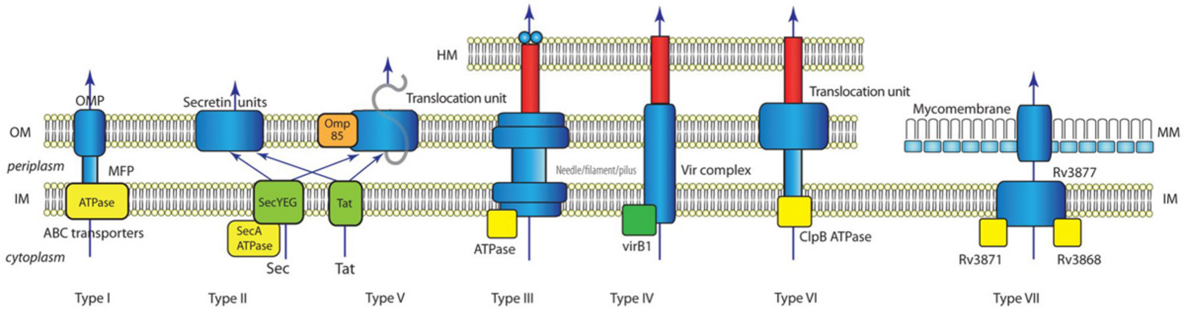


Figure 1.5: ***Overview of the 7 secretion systems identified in bacteria.*** *Shown are the basic outlines of the macro-molecular secretion apparatus in bacteria. HM: Host membrane; OM: outer membrane, IM: inner membrane; MM: mycomembrane; OMP: outer membrane protein; MFP: membrane fusion protein. ATPases and chaperones are shown in yellow [292].*

1.3.1 General secretion (Sec) pathway and twin-arginine (Tat) pathway

The general secretion pathway (Sec) and the two-arginine (Tat) translocation pathway are present in all domains of life. In bacteria, the Sec pathway is involved in both the secretion of unfolded proteins across the cytoplasmic inner membrane and the insertion of membrane proteins into the cytoplasmic membrane. The Tat pathway is able to transfer folded proteins. The majority of inner membrane proteins use the co-translational, the majority of secreted proteins the post-translational pathway. Although the mechanisms of transport are fundamentally different, both pathways also share common elements [201], compare figure 1.6. Substrates of the Sec pathway are recognized by a hydrophobic N-terminal sequence. The signal for Tat secretion is also located in N-terminal region of large co-factor containing proteins, recognized as a motif rich in basic amino acid residues (S-R-R-x-F-L-K) [292]. The secretion signals can be predicted by computational methods with high accuracy [224].

1.3.2 Type I secretion system (T1SS)

The T1SS is widespread in Gram-negative bacteria. Three membrane proteins are required for transport: A specific outer membrane protein (OMP), an ATP-binding cassette (ABC) and the so-called membrane fusion or adaptor protein (MFP). Assembled, they span the cell envelope and allow for a single-step secretion independent of the Sec pathway. The ABC component guaranties specific recognition of substrates. Necessary for secretion

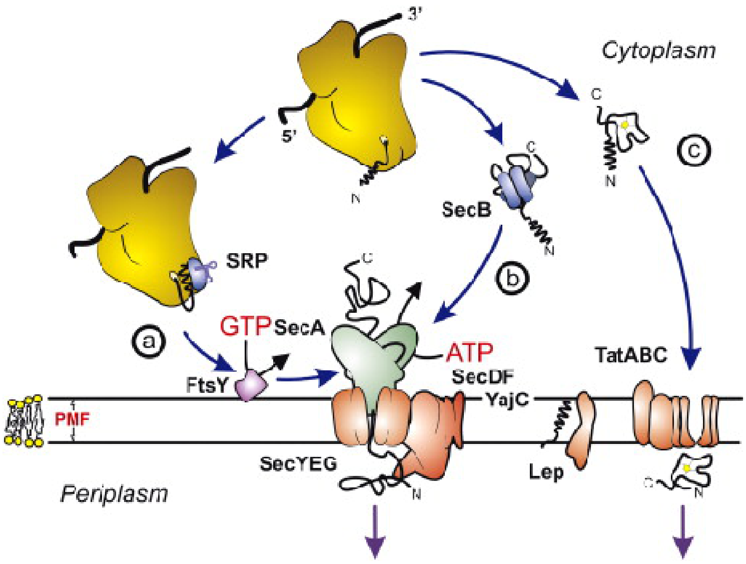


Figure 1.6: ***Schematic overview of Sec/Tat translocase secretion pathways in Escherichia coli.*** *(a) Co-translational and (b) post-translational secretion of unfolded proteins by the general secretion pathway. (c) Secretion of folded precursor proteins by the Twin arginine pathway [201].*

of a substrate is an uncleaved secretion signal in the C-terminal amino acid sequence. It it recognized by the specific ABC transporter protein, initial to the sequential assembly of the membrane spanning complex. Secreted proteins usually contain glycine rich repeats (GGXGXDXXX) in various numbers (up to 50) that form distinctive beta-sandwiches/ beta-roll structures. There are also substrates that contain no repeats or other repeats with other patterns, e.g. being homologous to regions in adhesion molecules. [74].

1.3.3 Type II secretion system (T2SS)

Pathogenic as well as non-pathogenic Gram-negative bacteria use the type II secretion system (T2SS) to translocate folded proteins from the periplasm through the outer membrane and into the extracellular milieu. In an initial step, Type II secretion therefore depends on the Sec/Tat pathway for prior transport over the inner membrane into the periplasmic space and is the main terminal branch of the general secretory pathway. An example for an effector that contains a signal peptide for translocation by the Sec pathway and has evidence to be further secreted by T2SS is the chlamydial protease Cpaf [267]. Several plant pathogens use the system predominantly for the secretion of extracellular toxins, surface-associated virulence factors and hydrolytic enzymes. The multi-protein secretion machinery consists of 12–15 different proteins that are generally encoded in a single operon. There are still major gaps in the understanding of the mechanism and architecture of the T2SS system [158].

1.3.4 Type III secretion system (T3SS)

The T3SS - also called injectisome - allows to secrete effectors directly into the cytoplasm of eukaryotic host cells, compare schematic view 1.7. This feature makes it a key factor of virulence regarding some of the most severe human and animal pathogens. Species of *Chlamydia, Xanthomonas, Pseudomonas, Ralstonia, Shigella, Salmonella, Escherichia* and *Yersinia* all possess the ability to translocate effectors via the Type 3 system. Besides pathogenic bacteria, the T3SS is also used by several symbionts to interact with the host protein network [25]. The T3SS machinery is composed of approximately 25 different protein subunits. It resembles the overall shape of a molecular syringe and incorporates

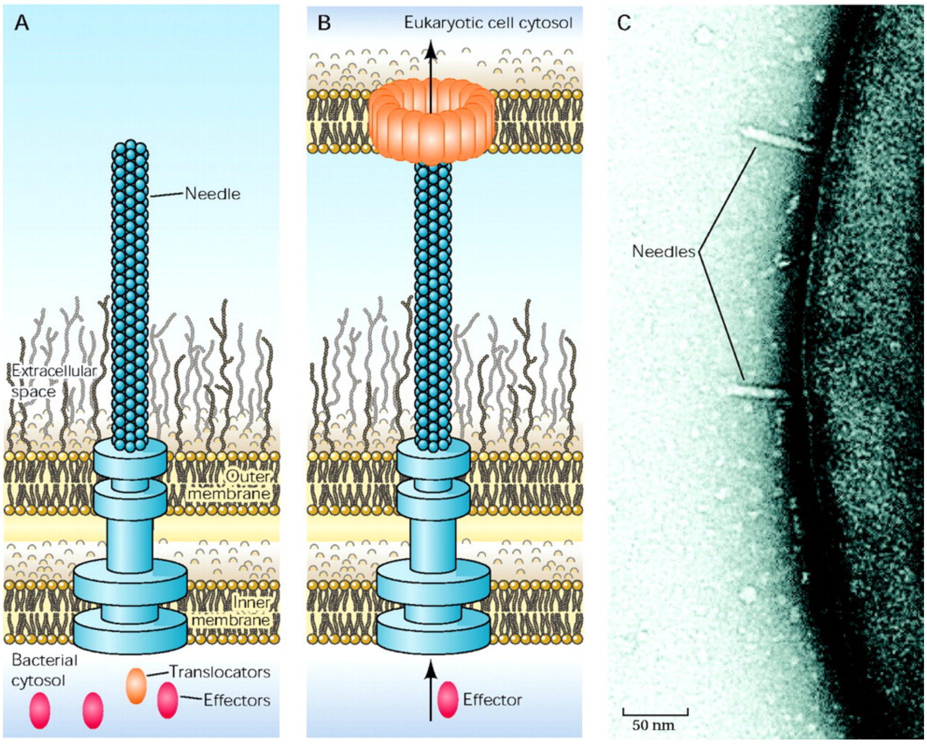


Figure 1.7: ***Schematic figures and electrone microscope image of the T3SS injectisome.*** *(A): a resting T3 injectisome spanning both bacterial membranes and the needle protruding (B): an active T3 injectisome with translocators forming a pore in the membrane of the eukaryotic target cell (C): electron micrograph of the surface of Yersinia enterocolitica with protruding needles [290].*

one to more than hundred copies into one of the most complex known bacterial nanomachines. The structural core of the system is the so-called needle complex that spans the bacterial cell envelope as a tripartite ring system and culminates in a needle protruding from the bacterial cell surface. Substrate targeting and translocation are accomplished by an export machinery consisting of various inner membrane embedded and cytoplasmic components. Formation of this multimembrane-spanning machinery is possible by precise orchestration in the assembly of all components [81]. The T3SS has a common evolutionary origin with the bacterial flagellum. Structures of the basal body are similar and many proteins share significant homology in both macro-molecular systems [290].
Regarding the nature of the T3 secretion signal, two different theories exist, both supported by experimental evidence. Either an N-terminal signal peptide [268] or a signal hidden in the mRNA sequence [16] could be responsible for substrate recognition, compare figure 1.8. It has become very probable by increasing evidence that a N-terminal signal in the effector sequence carries sufficient information to initiate translocation. This is further supported by experiments of Subtil et al. that showed positive secretion of chimeric proteins including the N-terminal part of chlamydial effectors by a heterologous T3SS in *Shigella* [279]. Findings are expected to be general due to the strong conservation of the T3SS and additional experiments by Anderson et al. [15]. Chaperone seem to have

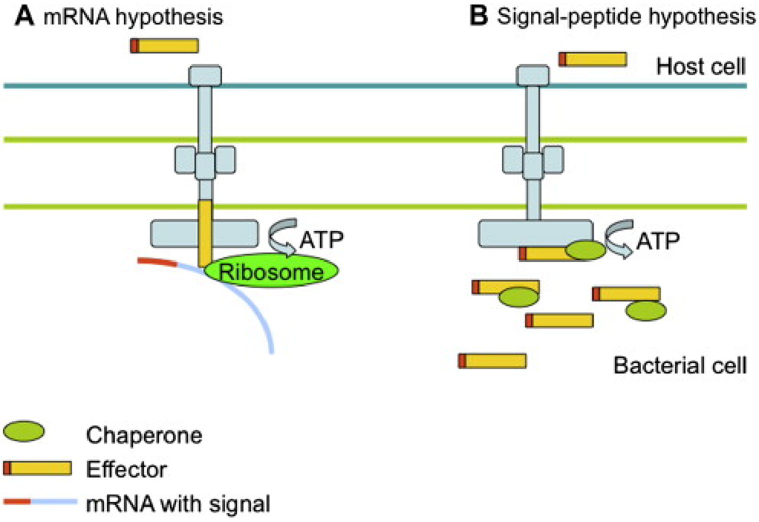


Figure 1.8: ***Schema illustrating the two hypotheses regarding T3 secretion signal location:*** *(A) mRNA based and (B) peptide based. (A) The effector mRNA contains the signal and the effector protein is synthesized during T3SS transport. (B) An N-terminal signal peptide is recognized by T3SS in the translated effector protein. Depending on the mechanism, chaperones either enhance signals or hold the protein in an unfolded, transportable state [21].*

multiple roles and play an important part in the translocation process. Yet no general mechanism based on the interaction of effectors with chaperones can be observed.

Experimental methods to verify protein secretion by T3SS Subtil et al. developed an in vivo assay to investigate the secretion of chlamydial effector proteins by a heterologous Type III secretion system. The N-terminal part of effector protein candidates that contains the signal peptide is fused to the Cya protein of Bordetella pertussis which serves as a reporter. After expression in various strains of *Shigella flexneri*, it was demonstrated that these hybrid proteins are secreted by the Type III secretion system of *Shigella flexneri.* The recognition of chlamydial secretion signals by the secretion machinery of another pathogen opens new possibilities for the study of chlamydial effector proteins [279].

1.3.5 Type IV secretion system (T4SS)

T4SS complexes are observed in Gram-positive as well as in Gram-negative bacteria. Pathogens with T4SS are Agrobacterium tumefaciens C58 (VirB), Helicobacter pylori (CAG,ComB), Pseudomonas aeruginosa (TraS/TraB), Bordetella pertussis (Ptl), E. coli (Tra), Legionella pneumophila (Dot) and the nitrogen-fixing plant mutualist Mesorhizobium loti [292]. Identification of the key components of eight different classes of type IV secretion systems were reported up to now [126]. In addition to effector proteins, the T4SS can transport nucleic acids into eukaryotic host cells, as well as into yeast and other bacteria. The T4SSs can be separated into three subtypes on the basis of their primary functions: DNA conjugation, DNA uptake/release, and cargo translocation, compare figure 1.9. Conjugative transfer of plasmid DNA or transposons is mediated into a wide range of bacterial species and eukaryotic host cells. The exchange of genetic material promotes genomic plasticity and adaptive responses of bacteria to changes of environmental conditions. The second subgroup mediates DNA uptake from and release into the extracellular milieu observed for some Gram-negative bacteria, including *Helicobacter pylori* and *Neisseria gonorrhoeae.* The third group delivers virulence proteins into eukaryotic (and mammalian) host cells. Gram-negative pathogens like *H.pylori*, *Brucella suis* and *Legionella pneumophila* make use of this Type IV secretion system to modulate the host system [104].

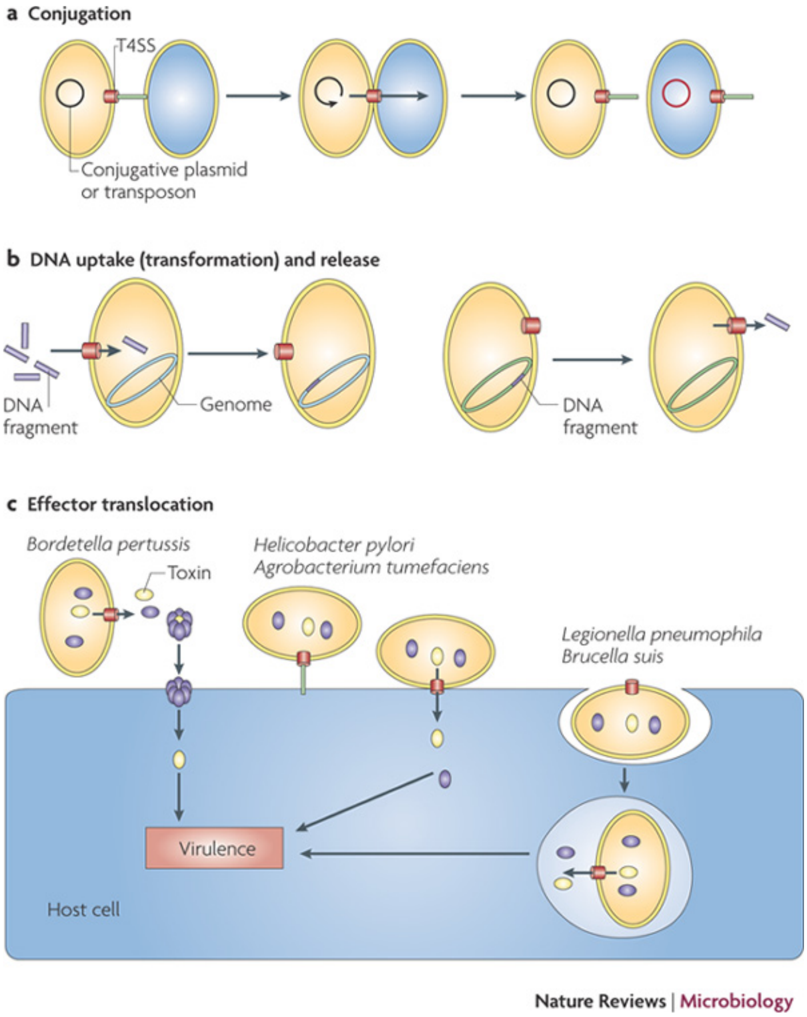


Figure 1.9: ***Schematic illustration of secretion mechanisms of Type IV secretion system subgroups.*** *(A): Conjugative T4SS - transport of plasmids and transposons (B): Transformative T4SS - mediated DNA uptake and release (C): Effector translocation - secretion of proteins (and/or) DNA into recipient cells [104].*

1.3.6 Type V secretion system (T5SS)

The T5SS is the simplest protein secretion system, consisting of several subtypes: the classic autotransporter system (type Va or AT-1), the two-partner secretion system (Vb), and the Vc system (AT-2), compare overview 1.10. Autotransporter proteins (AT) are present in all pathogenic Gram-negative bacteria. They constitute the largest number of secreted virulence factors of all seven secretion systems [143]. Secretion by T5SS is dependent on the Sec machinery for transport into the periplasm. Once in the periplasm, the autotransporter domain of the substrate inserts into the outer membrane to form a pore. The folded passenger domain is passed through the center of the autotransporter domain to be presented on the outside of the cell. In some cases, the protein is anchored to the outer membrane (like for adhesins). When the passenger domain is cleaved off, it forms a soluble enzyme or toxin [29]. Secreted proteins share several characteristics: a N-terminal signal sequence for Sec secretion into the periplasm, a functional passenger domain, a linker region necessary for translocation and a C-terminal region associated with the formation of the transmembrane pore [25].

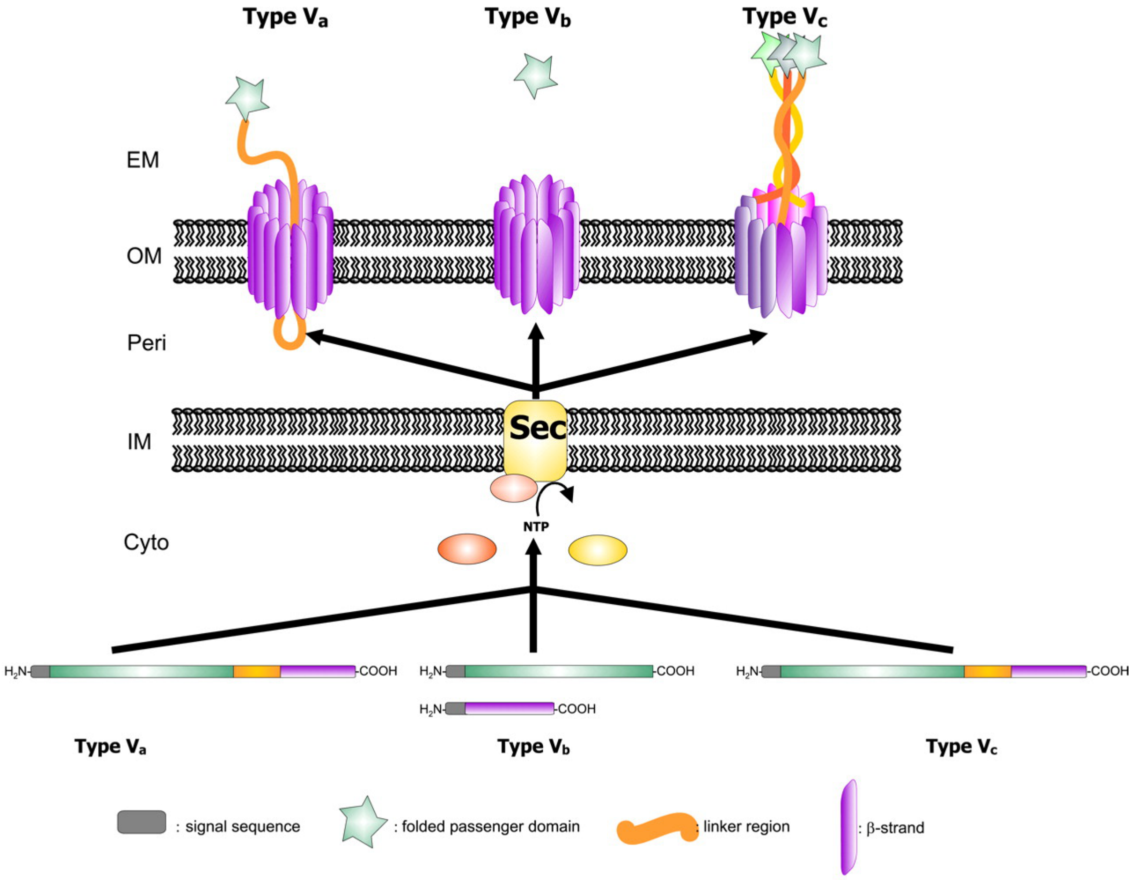


Figure 1.10: ***Schematic overview of Type V secretion system subtypes.*** *The secretion pathway of the autotransporter proteins (type Va) at the bottom left of the diagram, the two-partner system (type Vb) in the center and the type Vc or AT-2 family on the right. The signal sequence, the passenger domain, the linker region and the β-domain comprise the four functional parts. Once through the inner membrane, the signal sequence is cleaved and the β-domain inserts into the outer membrane and forms a pore. The passenger domain inserts into the pore and is translocated to the bacterial cell surface, where it may or may not undergo further processing [130].*

1.3.7 Type VI secretion system (T6SS)

The type VI secretion system (T6SS) is widespread in Gram-negative bacteria and transports effectors across the cell envelope. Secretion is Sec independent and happens in one single step [49]. The T6SS is typically encoded within a single gene cluster consisting of 13 conserved core components and a number of accessory ones. Conserved among all T6SSs include the T4SS IcmF- and IcmH-like proteins, a putative lipoproteins, the ClvP AAA+ ATPase (a potential energy source) and the Hcp (Haemolysing co-regulated protein) and VgrG (valine-glycine repeats) proteins. Together, those proteins form a membrane-embedded, syringe-like system. The T6SS components VipA/VipB (TssB/TssC) form a contractile sheath around a hollow, inner tube composed of Hcp (hemolysin-coregulated protein). Attached on top of the Hcp tube is a trimeric, spike-like cap consisting of VgrG (valine-glycine repeat protein G), similar to the tail spike complex of bacteriophages. Upon contraction of the VipA/VipB (TssB/TssC) sheath, the Hcp tube together with VgrG is pushed outward and penetrates the target cell. The penetration of target cells is accompanied by the delivery of specific toxins. Effectors can be attached to the VgrG spike or might be delivered directly through the hollow channel of the Hcp tube [147]. T6SS effectors can play important roles in virulence. They modify the eukaryotic host cytoskeleton through actin crosslinking, promoting host cell toxicity. Other T6SS effector molecules include the VasX protein secreted by V. cholerae that binds membrane lipids [190]. These toxic proteins target specifically prokaryotes to provide a competitive advantage against other microorganisms occupying the same niche. They are efficient weapons in interbacterial warfare and secretion of these effectors provides a fitness advantage by hydrolyzing cell walls of opponent bacteria. Up to now, no signal peptide could be identified that enables transport of effector proteins by the T6SS machinery. Bioinformatics analyses revealed a N-terminal motif, named MIX (marker for type six effectors) that is found in the sequence of T6SS effector proteins [247].

1.3.8 Type VII secretion system (T7SS)

The type VII secretion system is exclusive to Gram-positive bacteria. It was studied in Mycobacterium tuberculosis and found to transport immunogenic proteins that are characterized by a Trp-Xaa-Gly (WXG) motif and a size of about 100 amino acids (WXG-100 proteins) [262].

1.4 Bacterial effector proteins

A single bacterium may deliver up to 100 different, often multifunctional effector proteins into individual host cells to produce specific responses in the host [72]. The host-interacting lifestyle of pathogenic and symbiotic bacteria results in a host–pathogen co-evolutionary arms race that imposes intense selective pressures on these bacterial virulence factors [273]. Due to this pressure, different hosts and different survival strategies, known effectors can vary widely between different species and even between different strains of the same bacterial organism [250]. They do not show a typical folds or uniform domain composition, which could be used to identify them with certainty. Among effector proteins are e.g. Phosphatases, PiPases, Kinases, Lipases, Proteases, Cyclases and Lyases. Characterization of effector proteins provides insights into the strategies bacteria use to manipulate host cells. The YopJ effector which is secreted by *Yersinia* is a well-studied example of a protease that binds the mitogen-activated protein kinase (MAPK), preventing its phosphorylation and inhibiting inflammation [210].
Public resources for characterized bacterial effector proteins are VFDB [54] and MvirDB. The MvirDB database integrates DNA and protein sequence information on virulence factors from a comprehensive selection of public resources: Tox-Prot, SCORPION, the PRINTS database of virulence factors, TVFac, Islander, ARGO, CONUS, KNOTTIN, a subset of VIDA, the complete VFDB database and sequences derived by literature search [322]. Depending on the applied exact definition, virulence factors include also non-secreted endotoxins and components of the secretion system apparatus [163]. Effector proteins are a subset of bacterial virulence factors. Up to now, no completely characterized bacterial effectome was reported.

1.4.1 Effector similarity to eukaryotic proteins

Many effector proteins contain similar distinct functional and structural features, for example secretion signal peptides and specific domains that enable interaction with the host system. A common pathogen strategy is to mimic domains and binding motifs of eukaryotic proteins to alter parts of the host protein interaction network [88]. Besides analogous, merely structural mimicry [9], molecular mimicries are often the result of horizontal gene transfer (HGT) combined with high pathogenic genome flexibility due to host-pathogen co-evolution [180, 38]. E.g. Al-Khodor et al. suggest a common ancient origin for a common eukaryotic-like domain signature in bacterial effectors, the eukaryotic-like Ankyrin repeats [7]. A well characterized example of a protein protein interaction between the host and bacterial effectors is the Yersinia YopJ effector. Its eukaryotic-like SH2 domain inhibits NF-kappaB activation [252]. To mimic parts of the ubiquitin proteasome by effector proteins is a typical pathogenic attack used by plant and animal pathogens [17].

1.4.2 Eukaryotic-like domains in effector proteins of *Legionella pneumophila*

When analyzing effector proteins in *Legionella pneumophila*, Buchrieser et al. discovered signatures of protein domains otherwise only found in eukaryots [50]. Upon translocation, these proteins are able to directly modulate host response [175] and alter complex host-specific processes, e.g. mediate post-transcriptional modification of host proteins [241]. Phylogenetic analysis reveals that these proteins originate often in lateral gene transfer from the eukaryotic hosts and from other bacterial genes [179]. The phylogenetic patterns of selected eukaryotic-like signaling domains was described by Ponting et al., compare the representation of phylogenetic tree 1.11. Molecular mimicry of eukaryotic proteins was found to be an efficient strategy in the pathogenicity of *Legionella* species [205] and the dominant way to alter host cell functions [118].
Besides from effector proteins characterized in *Legionella* sp, eukaryotic-like domains are observed in effector proteins of taxonomically diverse pathogens [88], [258]. Pathogens comprising proteins with detectable eukaryotic-like domains, are for example the amoeba symbiont Candidatus Amoebophilus asiaticus or intra-cellular and environmental Chlamydia [220, 276].

1.5 Computational prediction of effector proteins

Identification and characterization of secreted bacterial effector proteins is laborious and not free of caveats, increasing the need for *in silico* prediction methods [124]. Computational prediction of effector proteins can be specific for the transport system identified in a particular pathogen. These secretion system dependent methods still have to overcome several challenges to enhance accuracy of the prediction. Alternative approaches that detect effector proteins independent of the secretion pathway are needed since not all ways of effector transport are known. Beside homology and signal-based approaches, prediction methods rely on the genomic and functional properties such as specific domain signatures

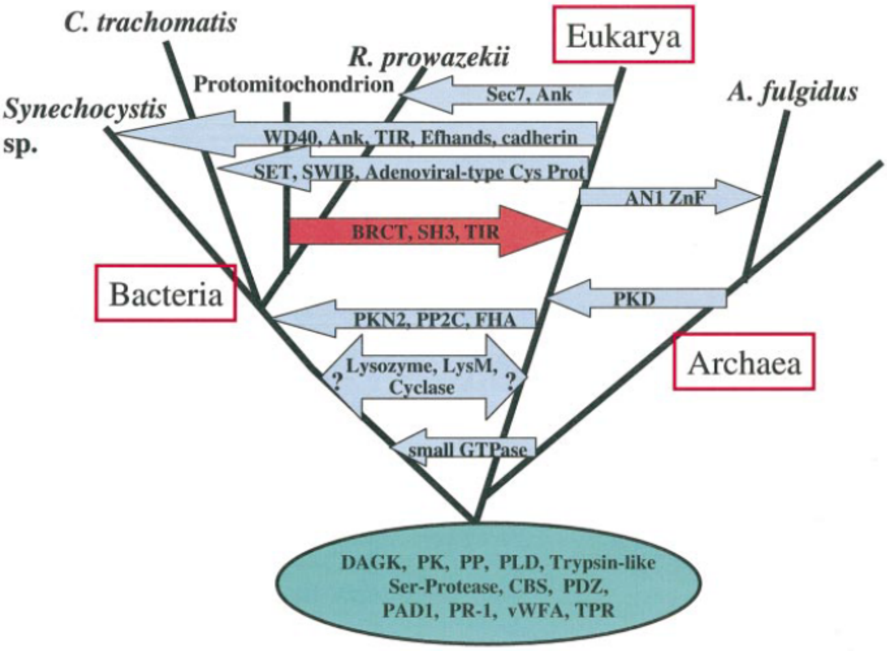

C. trachomatis
R. prowazekii
Eukarya
Protomitochondrion
Sec7, Ank
A. fulgidus
Synechocystis sp.
WD40, Ank, TIR, Efhands, cadherin
SET, SWIB, Adenoviral-type Cys Prot
AN1 ZnF
BRCT, SH3, TIR
Bacteria
PKD
PKN2, PP2C, FHA
Archaea
? Lysozyme, LysM, Cyclase ?
small GTPase
DAGK, PK, PP, PLD, Trypsin-like Ser-Protease, CBS, PDZ, PAD1, PR-1, vWFA, TPR

Figure 1.11: ***Schematic phylogenetic tree of selected eukaryotic-like domains.*** *Blue arrows indicate proposed horizontal gene transfer events. Red arrow show gene acquisition from mitochondrial endosymbiosis. Domains represented within the green oval are suggested to have been present in the last common ancestor of archaea, eukaryota and bacteria. The directions of HGTs regarding lysozyme, LysM, LRR and cyclase domains could not be stated [227].*

to identify effector proteins, compare the overview in figure 1.12. In the following, general effector prediction approaches are discussed into more detail.

1.5.1 Signal based methods

Existing methods for the in silico prediction of effectors focus on the recognition of secretion signal sequences within the proteins amino acid sequence. These signal peptides guide effectors through the specific secretion machinery. These methods can only be applied to bacteria for which the specified secretion system was identified. Up to now, modeling of the secretion signal is feasible for the Sec/Tat pathway and T3SS [186]. For all other secretion systems, signal peptides are not well defined [19].

The Type IV secretion signal is very hard to detect. Several bioinformatics screening approaches attempt to identify a secretion signal that is assumed in the C-terminus of T4 effectors, but the approaches are not applicable in general [171, 298]. A SVM based approach was able to discriminate T4 effectors and non-effector proteins of eight different

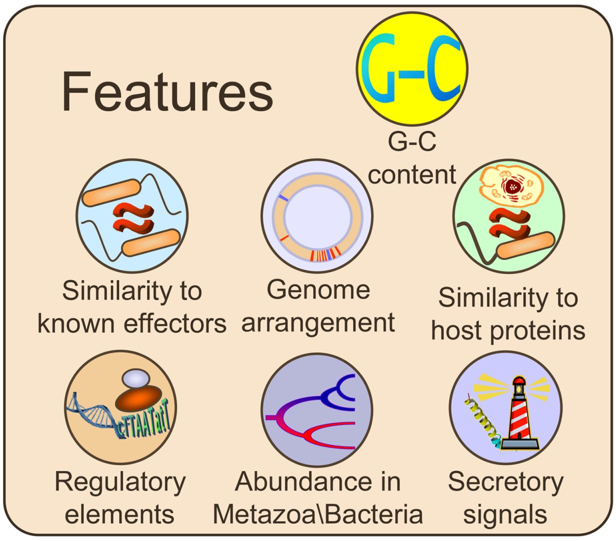

Features
G-C
G-C content
Similarity to known effectors
Genome arrangement
Similarity to host proteins
cTTAATaT
Regulatory elements
Abundance in Metazoa\Bacteria
Secretory signals

Figure 1.12: ***Overview of features used for prediction of effector proteins.*** *Features used by genomic and function based effector prediction approaches [45].*

pathogens with an accuracy of $> 95\%$, calculating distinctive features from the complete primary sequence, such as amino acid composition and position specific scoring matrix profiles [325].

Type III secretion signal For the T3SS of gram-negative bacteria exist several *in silico* methods to predict secreted effectors [20, 249, 176, 299, 64]. These approaches are based on modeling a signal in the N-terminal amino acid sequence of the effector proteins, the part that is responsible for guiding the effector protein through the T3 secretion machinery. The exact signal peptide is not known and T3 effectors show very low sequence similarity in these 100 amino acids of the N-terminal region. Classification algorithms using k-mere frequencies are able to extract a signal and EffectiveT3 is able to predict T3 effector candidates with a sensitivity and selectivity of up to 85% and provides explicit evidence for the taxonomic universality of the predicted signal.

Despite the successful modeling of a T3 signal, with more T3 effector sequences becoming available, limitations of these approaches become visible. Dehoux et al. found that only 64% of secreted Inc proteins in *C. trachomatis* were predicted to possess a T3 secretion signal by at least one of three different T3 effector prediction softwares, [73] These results for the family of chlamydial Inc proteins could hint to different, yet unidentified families within T3 signal peptides.

Sec/Tat secretion pathway Contrary to e.g. T3SS mediated secretion, for Sec and Tat transportation there is no direct association with secretion of effector proteins. The vast number of proteins transported by these pathways in Gram-negative bacteria end up in the periplasmic space within the bacterium. To distinguish these proteins from effectors that are further secreted by Sec-dependent secretion systems T2SS and T5SS, detection of an additional signal is be necessary. E.g. the chlamydial proteases Cpaf (CT858) and Tsp (CT441) with evidence for secretion into the host cytosol both contain a signal peptide for Sec transport and are therefore assumed to be further secreted via T2SS [31]. As the identification of Type II and Type IV signal peptides is not yet feasible, Sec/Tat secretion is of minor importance to the prediction of effector proteins.
The signals for the general secretion (Sec) pathway and Twin-arginine (Tat) pathway can be modeled with high accuracy. Several prediction programs predict a N-terminal cleavage site within the proteins amino acid sequence, compare figure 1.13.

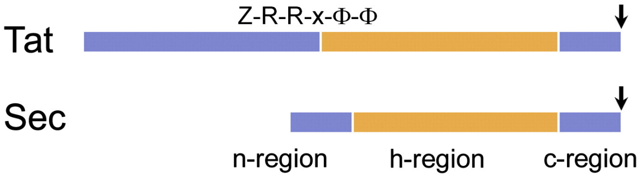


Figure 1.13: ***Schematic comparison of typical Tat and Sec pathway substrates.*** *The polar n-regions have a positive net charge, the hydrophobic h-regions are uncharged. Short polar c-regions contain a Type I signal peptidase cleavage site (AxA) [201].*

SignalP is based on a neural network to detect both N- and C-terminal cleavage sites. Thereby it successfully distinguishes between signal peptides and N-terminal transmembrane helices that share similar N-terminal features, a pitfall of other programs [224]. Effectors secreted by Sec-dependent secretion systems contain this defined cleavage sites. The Sec signal therefore could become informative for effector prediction with future progress in modeling of the specific signals of secretion systems Type II and V.

Chaperone binding sites Chaperones play a vital role for protein folding during T3SS secretion [6]. There is evidence that chaperones also mediate secretion by other bacterial secretion systems, e.g. T4SS [261]. A chaperone binding motif of weak sequence similarity was identified in effector proteins of seven bacterial pathogens which could make T3 chaperones important for the prediction of new T3SS effectors [64].

1.5.2 Homology based approaches

The results of experimental effector protein characterization makes it possible to apply homology searches to identify additional effector candidates in closely related genomes. This approach is applicable within the typical limits of homology based methods. Candidates are solely identified belonging to the few known effector protein families. Tobe et al. identified 65 effector candidates clustering into more than 20 protein families in enterohemorrhagic E.coli (EHEC), many of them shown to be secreted [288]. Not anticipated are also functional changes and cases of evolutionary effector invention by genome rearrangements, e.g. observed for 32% T3SS secreted effectors in analyzed Gram negative bacteria by Stavinides et al. [272].

1.5.3 Genomic and function based approaches

These approaches use more general features of effector proteins to detect substrates that do not contain any detectable secretion signal and the route of transport is unknown. Genomic arrangement, organization into pathogenicity islands or GC content as well as phylogenetic patterns and similarity to host proteins are features that can enable effector detection. Especially the identification of eukaryotic-like domains to predict effector candidates, additionally points to possible functions inside the host cell [19]. Eukaryotic-like protein domains were identified for single bacteria in several studies but these methods are not applicable on a large scale basis.

1.6 Prediction of pathogen host protein-protein interactions

While the prediction of intra-species protein–protein interactions is a well-investigated problem addressed by several computational methods, most of these methods cannot directly be applied to the prediction of pathogen-host protein-protein interactions (PH-PPIs). Besides, the data on experimentally derived PH-PPIs is much more limited than for single-organism PPIs due to the higher costs and effort necessary. Few public resources on experimental protein interactions include bacterial PH-PPIs, mainly the Pathogen Interaction Gateway (PIG) [84], Phidias [313] and the database for Pathogen-Host Interactions (PHI-base) [308]. In this section, important methods for the prediction of PH-PPIs are discussed.

1.6.1 Interolog based methods

An interolog is the conserved PPI between two proteins that have interacting homologs or orthologs in another organism. The interolog method for PH-PPI prediction is based on the assumption that an interaction observed between a pair of proteins will be conserved among similar sequences of different organisms while keeping its function. Thereby known interactions can be mapped onto homologous or orthologous proteins in different species. Krishnadev et al. applied a homology based approach to predict PH-PPIs between human and three pathogens *E. coli*, *Salmonella enterica typhimurium* and *Yersinia pestis* [160]. A similar approach was applied to predict the host-pathogen interactome of human and *Helicobacter pylori* [293]. With increasing evolutionary distance genes share less sequence similarity and are more likely to undergo neo-functionalization. This could lead to erroneous orthologous assignments, weakening the interolog based method. Interactions that were acquired through horizontal gene transfer (HGT) are easily spotted by the method. On the other hand, several PH-PPIs might be unique for particular pathogens and have evolved during the co-evolutionary arms race between pathogen and host. Since a reliable mapping of interactions depends on high sequence similarity (usually above 30%) [12, 317], homologs of many fast evolving effectors might not be detected in the original approach. Additional properties such as molecular characteristics and gene expression data are therefore considered to diminish the rate of false positive predictions [312].

1.6.2 Sequence based approaches

General properties of the amino-acid sequences of effectors and host proteins could be used in training classification approaches to detect pathogen-host interactions. A machine learning approach based on the chemico–physical properties of short amino-acid segments was applied by Dyer et al. to predict the HIV-human protein interaction network [85]. In a subsequent study, the authors extended the approach to include domain information, sequence k-mer and host network properties to achieve 70% precision in the prediction of the host-viral interactome [86]. The considerable performance in both studies depends fundamentally on a reliable training and test set. Composition of these sets was possible due to the high number of experimentally validated host-pathogen/pathogen host protein interactions (PH-PPIs) that are available for HIV. In the case of bacterial pathogens limited experimental data puts major constraints on learning approaches for PH-PPI interactome prediction.

1.6.3 Domain-based prediction

In the study by Dyer et al., the information derived from domain profiles proved as an important feature for the successful prediction of PH-PPIs in HIV [86]. Several domain based methods to predict protein protein interactions are described for the intra-species case. These methods are based on the observation that similar domains in otherwise differing proteins can mediate the interaction with the same substrates. In Psibase, domain–domain interactions (DDIs) are predicted by the Protein Structural Interactome map (PSIMAP) algorithm from known Protein Data Bank (PDB) structures [121]. For assignment of structural domains, PSIMAP uses domain models generated by PSI-BLAST and protein folds as classified in the SCOP database [197]. A comprehensive web resource on domain–domain interactions using Pfam domain signatures is the Domain Interaction Map (DIMA) [178]. DIMA 3.0 integrates four different computational methods to predict domain domain interactions (DDIs): CMM (correlated mutations), DIPD (domain interaction prediction in a discriminative way), DPEA (domain pair exclusion algorithm) and DPROF (domain phylogenetic profiling), compare figure 1.14. The DIPD machine learning approach first constructs domain combinations from both PPI as well as non-PPI data. All possible domain pairs are treated as features. A feature selection algorithm subsequently discriminates between informative and uninformative domain pairs for any given PPI to derive a minimum representative set of interacting domains [320]. The Domain Pair Exclusion Method (DPEA) analyzes the frequency of co-occurring domain pairs in known protein protein interactions [239]. The correlated mutations (CMM) approach uses the idea that evidence for co-evolution between interacting proteins can be observed as functional constraints on the domain level. Correlated mutations are identified and scored based on three different algorithms. The Domain Profile method (DPROF) also makes use of the idea that interacting domains undergo evolutionary constraints [213]. Two domains that depend on each other for an important cellular function generally need to be present together or not at all in a given genome. Interacting domains are recognized by analyzing the phylogenetic distribution of domain pairs. Two domains are reported as interacting domains if they show similar phylogenetic profiles.

PPI and non-PPI datasets required for the DIPD approach are based on protein interactions in the IntAct database [148]. Input data for CMM and DPEA are PPIs from IntAct as well as functionally linked orthologous groups of proteins (COGs) extracted from the STRING database [280]. DPROF uses complete genomes and functional protein annotations from PEDANT [297] and SIMAP [235]. Results of the four computational methods are compared against a reference set for non-interacting protein pairs listed in

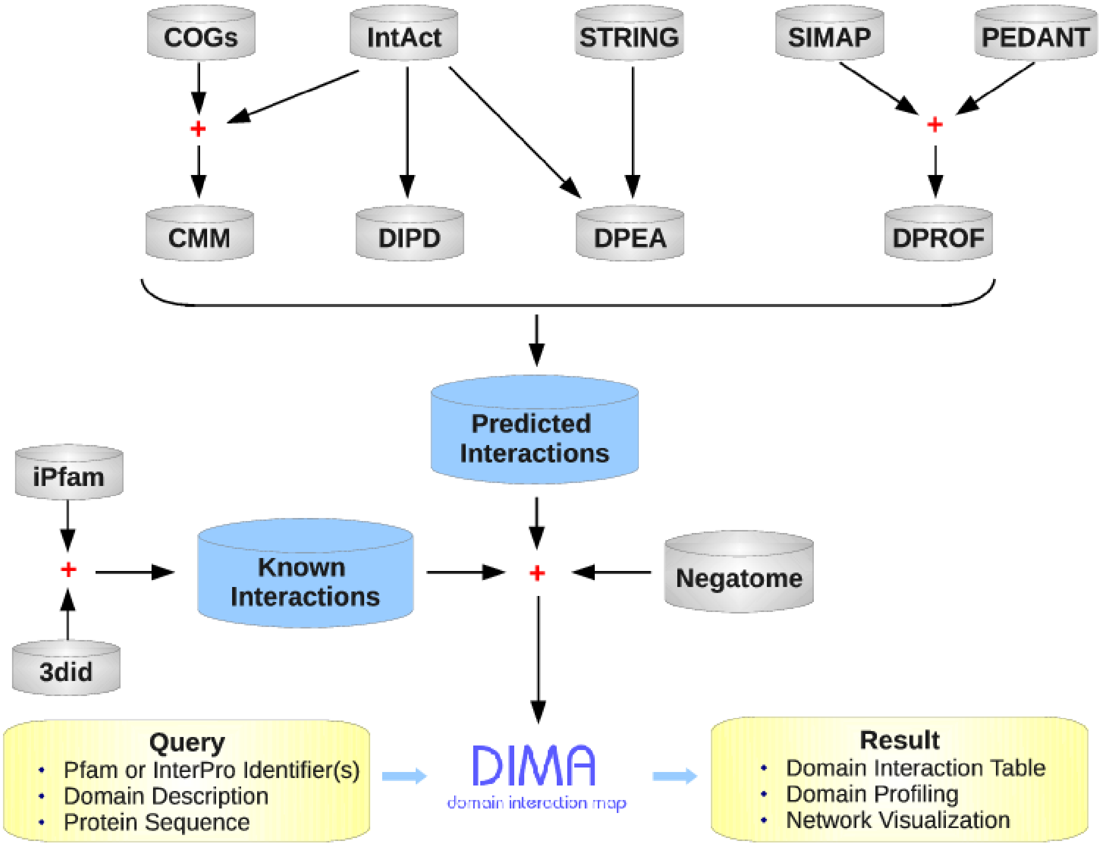


Figure 1.14: ***Overview of the DIMA database.*** *DIMA 3.0 integrates four different computational methods to predict domain domain interactions: CMM (correlated mutations), DIPD (domain interaction prediction in a discriminative way), DPEA (domain pair exclusion algorithm) and DPROF (domain phylogenetic profiling) [178].*

the Negatome database [266]. In this additional step highly unlikely physical interactions are filtered out. Domain-domain interactions that could be derived from protein complexes based on close contact of domains in these structures are directly imported from the iPfam [99] and 3did [274] databases.

Predictions by iPfam were also used in the study by Tyagi et al. modeling the pathogen-host interactome of *Helicobacter* and human [293]. By integrating data from iPfam, PSIMAP with interolog-based predictions, Kim et al. predicted a set of 3400 pathogen-host PPIs for the rice pathogen *Xanthomonas oryzae* [152]. Among the predicted interactors, over-represented GO annotations describe functions known to be relevant for pathogenicity and immune response of the host system, for example regulation of actin cytoskeleton, NF-κB signaling or cytokine production [85].

The presented methods in general are accepted to allow for meaningful predictions. To predict interactions of a specific pathogen, domain-based methods might offer a more sensitive approach than the interologs based method to predict interactions over long evolutionary distances. On the other hand, the modeling of pathogen-host interactomes by domain-based approaches like DIMA could also lead to over-sensitive predictions. The prediction of numerous interaction candidates on the host side could make additional filtering steps necessary. The performance of these methods is still difficult to measure due to the incompleteness of domain and structural databases and a lack of data on experimentally validated interactions.

1.7 Concepts used in this work

General concepts and methods applied in this work are introduced here. Supervised machine learning approaches can be successfully applied to handle complex classification problems. A naive Bayes classification approach allows for a simple and powerful separation of classes on the basis of the a-priori probabilities of input data. Classification performance can be evaluated by several performance measures based on results from validation procedures like 10-fold cross-validation.

1.7.1 Machine learning

In machine learning, statistical and and algorithmic concepts are used to construct systems that are able to extract non-trivial knowledge from data. Machine learning techniques find applications in any field were evaluating large amounts of empirical data is essential. Domains dependent on these algorithms are as ample as stock market analysis, search engine performance, analysis of satellite images or credit card fraud detection. In genome oriented bioinformatics, machine learning approaches are applied in gene prediction, functional annotation, clustering of protein families, classification and many more, compare overview 1.15. Machine learning methods can be categorized into supervised and unsupervised methods. Unsupervised methods, e.g. clustering algorithms find structural features within the data without any prior information. Supervised methods first generalize rules on a set of training instances to be able to make statements on unseen data. Classification approaches are an example of supervised machine learning techniques.

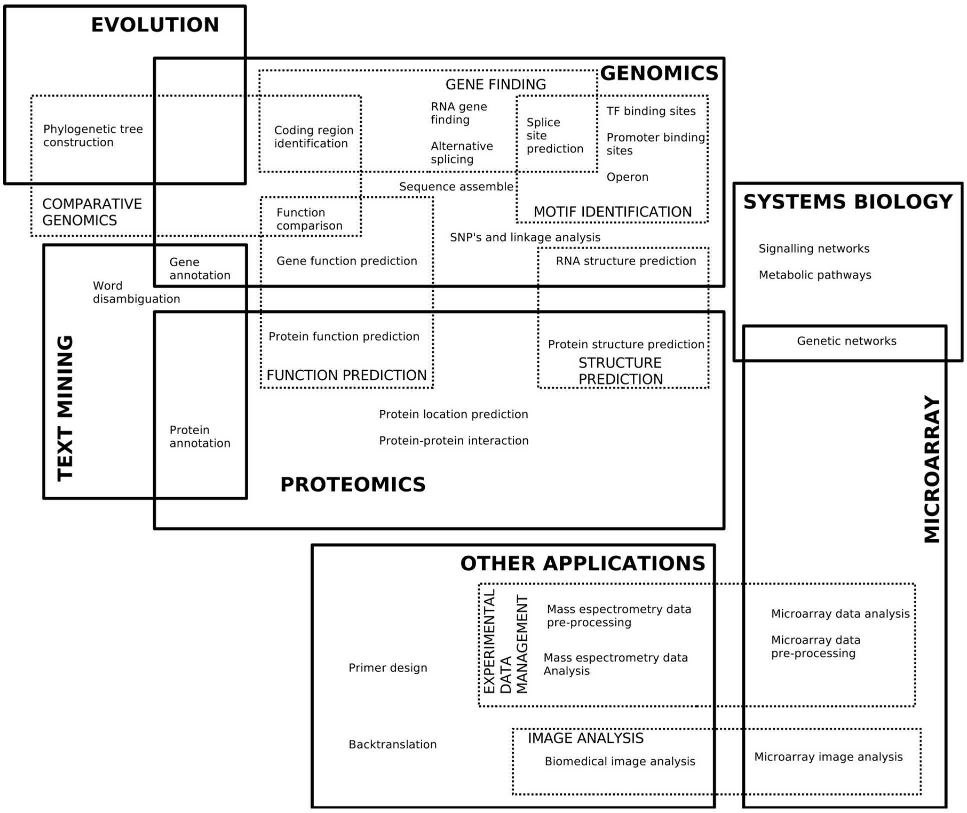


Figure 1.15: ***Scheme of areas in computational biology where machine learning methods are applied.*** *Domains in bioinformatics where machine learning techniques are used for knowledge extraction from large data sets e.g. include genomics, proteomics, microarrays, systems biology, evolution and text mining [164].*

Classification approaches Classification approaches learn abstract rules on a set of training instances in order to predict the correct class of other, unseen instances. Binary classification algorithms abstract features from training instances that are positive and negative regarding a certain class of interest. Several classification algorithms or classifiers were developed, based on different mathematically principles. An example for a non-probabilistic binary classifier are support vector machines (SVM). A SVM model represents training data instances as points in space. The aim thereby is to maximize the gap that separates examples of different classes. The particular class of unseen data points is then predicted by mapping these instances into the space and to check on which side of the gap they end up [44]. Naive Bayes is an example of a probabilistic classifier. The naive Bayes classification approach allows for a simple and powerful separation of classes on the basis of a-priori probabilities of the input data. It uses the odds ratio of conditional probabilities for seeing a certain feature in the positive and in the negative class to predict an instances class label. It is based on the Bayesian theorem and its probabilistic model and combines this model with a decision rule. E.g. the maximum a posteriori decision rule is to chose the most probable hypothesis. The classifier is called "naive" because it assumes general independence of input features. Despite of this simplification, the classification proves nearly optimal for certain applications, especially if conditional independence of features is given [318].

1.7.2 Statistical measures

Beside a Z-score based measure for statistical significance, typical evaluation measures and procedures that are relevant for this work are subsumed in the following.

Z-score The Z-score is a projection of an instance into a normal distribution to decide whether or not the instance is a member of the normal distribution (the null hypothesis). It thereby measures the relationship of a score to the mean in a group of scores in terms of standard deviations. It generally assumes parametric data and can be positive or negative, indicating whether it is above or below the mean and by how many standard deviations 1.16. Scores can be considered significant for values of 3 - 5 [269].

Typical performance measures Several measures are used to capture and compare classifier performance based on the counts regarding the classification of instances in the test data. The number of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) predictions allows for calculation of typical measures:
Accuracy, the percentage of correct predictions

$$Accuracy = (TP + TN)/(TP + FP + TN + FN)$$

precision, the percentage of positive predictions that are correct

$$Precision = TP/(TP + FP)$$

sensitivity, the percentage of positive labeled instances that were predicted as positive

$$Sensitivity = TP/(TP + FN)$$

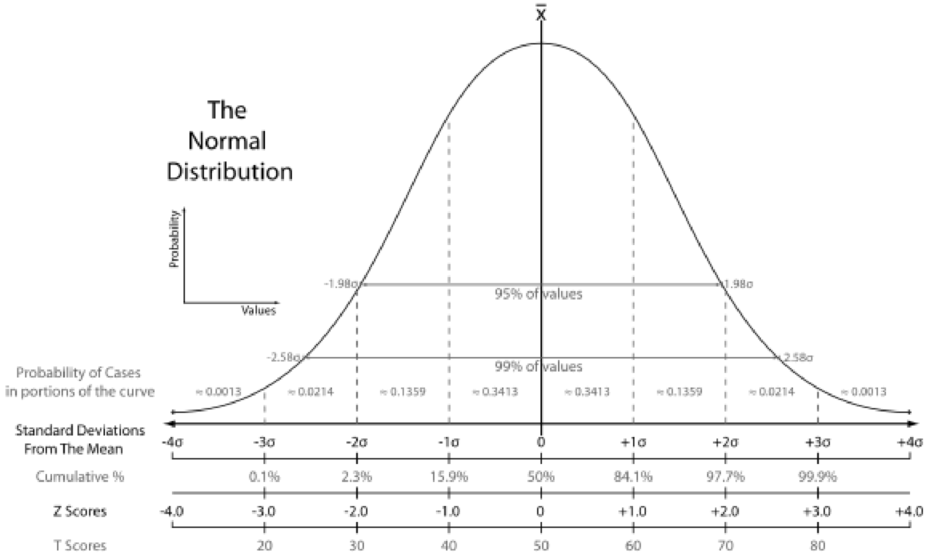


Figure 1.16: ***Statistics on normal distributed data.*** *A graph of a normal bell curve showing statistics used in standardized testing. Among others, the scales include standard deviations, cumulative percentages and Z-scores.*

and selectivity, the percentage of negative labeled instances that were predicted as negative

$$Selectivity = TN/(TN + FP)$$

.

Receiver operating statistics (ROC) Receiver operating statistics (ROC) are a common method to evaluate classifier performance. ROC analysis has originally been invented to estimate the error rate in broadcasting performance of radar systems. In a ROC curve, the true positive rate is plotted against false positive rate for varying a threshold above test instances that are classified as positive. The area under the ROC curve (AUC) is an established single-number performance measure [136]. It can be used to compare different approaches on the same data.

Cross-validation Cross-validation is a powerful method to asses the performance of binary classification approaches [157]. E.g. for a 10-fold cross-validation, the initial data is split into 10 sets of equal size. Each of these sets is used as test data once in 10 independent runs. For each run, all other 9 sets are combined to constitute the instances for training the classifier. Performance measures are calculated on combined predictions of all runs. Cross-validation allows for testing classifiers on a small population of samples simultaneously guaranteeing independence of test and training instances [18].

1.7.3 Web portal/resource development

Basic guidelines and concepts for the development of web portals and resources are addressed in the following.

Basic requirements General basic requirements of a functional web resource include:

- fast download time of web pages
- clean page and content layout
- easy to use and navigate
- comfortable access to all relevant information
- regular updates of page content
- versioning of archived data
- easy maintainability

Use cases A use case is the generalization of different scenarios in a list of steps that occur during the interaction between a system and a user to achieve a specific goal. In software engineering, these steps are typically depicted in Unified Modeling Language (UML), a modeling language designed to provide a standard way to visualize the design of a system [39]. Thereby, use cases are a way to define and document the requirements that need to be implemented during systems development.

Portal architecture An optimal software pattern for web resources is a 3-tier architecture model. It implements independent levels for data storage, processing and presentation [87].

In the 3-tier architecture, presentation tier, application tier and data tier are separate

from each other. The presentation tier handles the interaction with the user. The application tier contains the application logic and is responsible for the integration of different methods, while the data tier is responsible for data storage and retrieval. All three tiers communicate with each other using specified interfaces for data query and retrieval. The effect of changes in one of the layers is limited to this layer and does not affect other layers, as long as the interfaces between them stay unchanged. The result is a highly flexible solution that is easy to extend and maintain.
The three-tier architecture has the following three tiers:

- Presentation Tier: Interaction with the user and display of data is organized by this layer, e.g. in a web browser. Data is retrieved by communication with the lower tiers.
- Application Tier: This tier contains the business logic and is responsible for data processing and functionality of the application. This also includes conversion of data into specific formats that are accepted by the presentation tier for displaying information.
- Data Tier: In this tier, storage and basic retrieval of information is organized, e.g. in a relational database.

Data storage and database design Data storage is possible in the form of files on a filesystem or in a database. Database management systems (DBMS) are optimized for processing complex queries and information retrieval. It consists of the database and a software providing efficient solutions for data access, security, backup and other features. While several types of DBMS exist, the most common type is the relational database model. A relational database usually consists of normalized tables that are connected by relations. Information thereby is not stored in one big table but spread over several tables each addressing different properties to avoid redundancies. To retrieve the complete information spread over several tables, these tables are joined for a particular query. Join processes are optimized for performance by the database management system. Several normalization procedures were developed, dealing with the generation of non-redundant tables in databases [166].

2 A secretion system independent approach to predict bacterial effector proteins

2.1 Motivation

Bacterial secreted proteins, so called effectors, are key to an understanding of the interactions between a pathogen and its host. The analysis of effector proteins offers insights into the biological processes involved in initiating and maintaining infection. While data on experimentally characterized effector proteins is limited, *in silico* prediction approaches are an effective way to select for experimental candidates for further analysis to increase our knowledge on the mechanisms behind bacterial pathogenicity. Signal-based prediction approaches are modeling the secretion system specific signal peptide that is necessary for guiding the effector to the secretion machinery. Still many challenges are to be addressed in this area of research. Up to now, successful modeling of the signal sequence is limited only to a subset of characterized secretion systems. Furthermore, even for the best modeled signal, the Type III secretion signal, experimental results suggest the existence of strong signal diversity and still unknown classes of the T3SS signal [73]. Experimentally validated secretion of eukaryotic-like proteins in *Legionella* strains has shown that the signal based prediction criteria are currently not provide a comprehensive identification of effector proteins and complementary approaches are needed [241]. We developed a function based, secretion signal independent method for the prediction of secreted bacterial proteins. Aim of the approach is an automatic, large-scale detection of effector candidates based on function-specific criteria derived from sequence analysis. The discovery of eukaryotic-like domains in the effector proteins of several pathogenic bacteria is starting point for the development of this method. The concept of eukaryotic-like protein domains (ELDs) is thereby generalized to identify bacterial effector proteins independently of any secretion system mechanism in a large-scale, function-based approach. A methodology is derived from the analysis of representative examples of eukaryotic-like effector domains in the human pathogen *Legionella pneumophila*.

2.2 Exemplary analysis of the eukaryotic F-Box protein domain in bacteria

Buchrieser et al analyzed individual cases of eukaryotic-like domain containing proteins in *Legionella pneumophila*. These proteins were found to directly and effectively alter host functions to the pathogens advantage [50]. In non-hostassociated bacteria, this functionality is not given and eukaryotic-like domain signatures might be acquired e.g. by horizontal gene transfer but are less evolutionary conserved. Eukaryotic-like domains are therefore expected to be a phenomenon primarily observed in pathogenic and symbiotic bacteria. A uneven distribution of these protein domains among bacterial genomes of different lifestyle could provide a basic concept for the large-scale identification of proteins involved in bacteria-host interactions.

Material and methods

For exemplary testing of this assumption, the eukaryotic F-box domain discovered in *Legionella* effectors by Buchrieser et al was selected. The protein domain is characterized for its functional mimicry in divers pathogens [4]. The taxonomic distribution of the F-box domain signature (PF00646) as described by the PFAM domain database [230] was investigated manually. Bacterial genomes with annotated F-box domains where identified using the information provided on the webportal of the Pfam protein domain database. Phenotypes of the bacterial organisms with F-box domain containing proteins where determined by literature search.

Results

While the domain is frequent in eukaryotic organisms, it is only observed in proteins of 17 bacteria. Furthermore, a closer look at the phenotypes of these bacteria reveals that they are all pathogenic. Besides a conservation in most *Legionella* species, the domain is present in strains of *Pseudomonas*, *Xanthomonas* and *Ralstonia*, compare table 2.1. No occurence of the domain was detected in any non-pathogenic, non-host-interacting bacteria. The eukaryotic F-Box protein domain is found exclusively in pathogenic/symbiotic bacteria.

Organism	**phenotype**	**protein accessions**	**protein descriptions**
Legionella longbeachae NSW150	p	YP_003454966.1	choline kinase
Legionella pneumophila str. Lens	p	YP_125603.1	hypothetical protein
Legionella pneumophila str. Paris	p	YP_124791.1,YP_122576.1	hypothetical protein
Legionella pneumophila ATCC 43290	p	YP_005184559.1	FBOX-containing protein
Legionella pneumophila Philadelphia 1	p	YP_094225.1,YP_007569602.1	hypothetical protein
Pseudomonas stutzeri A1501	s	YP_001171179.1	hypothetical protein
Pseudomonas syringae pv. phaseolicola 1448A	p	YP_273421.1	hypothetical protein
Ralstonia solanacearum GMI1000	p	NP_519478.1	GALA protein 5
Ralstonia solanacearum Po82	p	YP_006029672.1	type III effector gala6
Xanthomonas albilineans GPE PC73	p	YP_003375967.1	hypothetical protein
Xanthomonas axonopodis pv. citri str. 306	p	NP_641107.1	hypothetical protein
Xanthomonas campestris pv. vesicatoria str. 85-10	p	YP_362537.1	hypothetical protein
Xanthomonas oryzae pv. oryzicola BLS256	p	YP_005627194.1	type III effector XopI
Burkholderia rhizoxinica HKI 454	s	YP_004029534.1,YP_004030488.1, YP_004028664.1	hypothetical protein
Candidatus Amoebophilus asiaticus 5a2	s	YP_001958190.1,YP_001958411.1, YP_003572975.1,YP_001957289.1, YP_001958017.1	hypothetical protein
Candidatus Protochlamydia amoebophila UWE25	p	YP_008448.1,YP_007728.1, YP_007732.1,YP_007733.1, YP_008193.1	hypothetical protein

Table 2.1: ***Eukaryotic F-box domain signatures in bacterial proteins.*** *Listed are bacterial genomes with annotated proteins containing an eukaryotic F-box domain signature (PF00464). The bacterial phenotype is p for pathogenic and s for symbiotic. For all domain containing proteins, accession numbers and according descriptions are given.*

Discussion

The comparative genomics analysis of the F-box domain of *Legionella pneumophila* effectors reveals a special feature of this eukaryotic-like domain. The F-box domain is conserved exclusively in eukaryots and pathogenic bacteria. The observed over-representation in pathogenic/symbiotic compared to non-hostassociated, non-pathogenic bacteria is likely to be an evolutionary consequence of the vital function of the F-Box containing effector proteins for virulence of the specific pathogens. The protein domain distribution in bacteria observed for the eukaryotic-like F-box domain captures a pattern which is likely to be common to effector proteins of diverse pathogens. The following large-scale approach uses this uneven distribution to predict and evaluate bacterial effector proteins based on the occurrence of eukaryotic-like domain signatures. A non-signal based, complementary prediction approach might help to realize a comprehensive *in silico* identification of effector candidates to increase time and cost-efficiency of experiments in the wet lab.

2.3 Setup of a genome repository for comparative analysis

The large scale *in silico* identification of eukaryotic-like protein domains depends on a comprehensive computational resource. A genome repository is implemented to preprocess all necessary information and to make data easily accessible for further analysis.
Included information and data sources are discussed in detail, an overview is given in figure 2.1. Information in the genome repository comprehends

- protein sequences of completely sequenced genomes
- protein domain signatures
- phenotype annotations of bacterial organisms
- gram status of bacterial organisms
- genomic annotation of bacterial secretion systems

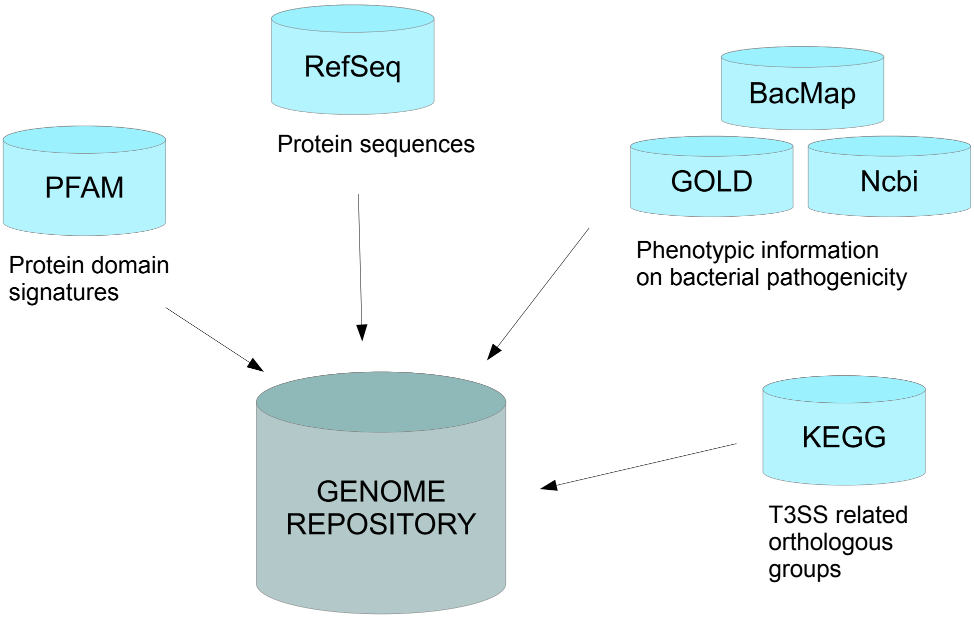

RefSeq
Protein sequences
BacMap
GOLD
Ncbi
Phenotypic information on bacterial pathogenicity
PFAM
Protein domain signatures
GENOME REPOSITORY
KEGG
T3SS related orthologous groups

Figure 2.1: ***Data sources and information processed in the genome repository.*** *The genome repository combines data from several public databases. Protein sequences of all completely sequenced bacterial and eukaryotic genomes are downloaded from the RefSeq database. PFAM domain signatures for all genomes listed by Interpro are retrieved from the SIMAP resource of protein similarities.*

Protein sequence data of completely sequenced bacterial/eukaryotic genomes

The NCBI Reference Sequence Database (RefSeq) is chosen as a comprehensive resource for completely sequenced genomes [229]. All completely sequenced bacterial and eukaryotic genomes have been retrieved from RefSeq release 55. Bacterial genomes were selected by taxonomic lineage according to the NCBI taxonomy [93]. Within the genome repository, the bioprojectid provided by RefSeq serves as unique genome identifier.

A comprehensive collection of protein domain signatures

Protein domain signatures are compared instead of considering overall similarity of complete proteins. This allows for increased sensitivity and accommodates the high evolutionary divergence of bacteria included in the analysis. The PFAM database is a large and comprehensive collection of taxonomically universal protein families [230]. As the basis for analysis, PFAM protein domain signatures have been chosen. PFAM annotations have been taken from the Interpro version 40.0 [138], provided by the SIMAP database [235] (release november 2012).

2.3.1 Phenotype annotations of completely sequenced bacterial genomes

Information about the phenotype and lifestyle of bacteria is available for a large fraction of completely sequenced bacterial organisms. Thereby, phenotype annotations rely on manual revision of experimental results, collected in environmental and host meta-data by scientists of the particular sequencing projects.

Material and methods Phenotype annotations are available for a large number of bacterial organisms. In this work, annotations were collected and combined from three public resources:

- BacMap [68] (actual release on september 2012)
- GOLD [212] (version 4.0 on september 2012)
- NCBI microbial genomes list [303] (actualized in 2011, ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/lproks_0.txt).

Retrieval and processing of phenotype information is implemented and adjusted to each resource individually:

BACMAP: BacMap is not offering options for direct download, so the data was retrieved via scripts directly from HTML web-pages. Retrieved features include the organism name and taxonomy id as well as information regarding pathogenicity, host-association and gram staining. Entries were mapped to completely sequenced bacterial genomes via taxonomy id.

For consistent phenotype annotation of the bacterial organisms, a pathogenicity rule was implemented as follows: According to the respective terms, a bacterium is labeled

non-pathogenic,
if (pathogenicity="No" & habitat!="Multiple")

```
symbiontic,
    if (habitat="Host-associated")
pathogenic,
    if (pathogenicity="Yes" | pathogenicity="Probable" | pathogenicity="Rarely")
and NA,
    otherwise.
```

Manual inspection of the term combination pathogenicity="No" and habitat="Multiple" suggests a symbiotic phenotype in most cases. Nevertheless, this assumption cannot be distinctly determined automatically without a manual literature search and is not suitable for an automated update process. Because it is not possible for these cases to distinguish between non-pathogenic/non-host-associated and symbiotic organisms in an automated manner, combination of these terms is classified as "no phenotype information available" ("NA").

GOLD database: From the GOLD database, phenotype information is retrieved by manual download of all GOLD-stamp tables. Categories are organism name, taxonomy id, habitat, disease, host name, phenotype and gram staining.
The GOLD pathogenicity rule annotates a bacterial phenotype as

```
non-pathogenic,
    if (phenotype="Non-Pathogen" &
        (hostname="" | habitat!="Host" | habitat!="symbiont"))
symbiotic,
    if (phenotype="Non-Pathogen" &
        (hostname!="" | habitat="Host" | habitat="symbiont")) |
        (phenotype="Pathogen" & habitat="Host" & disease="") |
        (habitat="Host" & disease="" ) |
        (hostname!="" | disease="")
pathogenic,
    if (phenotype="Pathogen" & (disease!="None" | disease!="")) |
        (habitat="Host" & disease!="") |
        (hostname!="" & disease!="")
and NA,
    otherwise.
```

NCBI phenotypes list: The NCBI phenotypes list was downloaded from the NCBI FTP-server. Mapping of phenotype information to RefSeq genomes is implemented via NCBI taxonomy id. The pathogenicity rule for the NCBI phenotypes list annotates a bacterial phenotype as

```
pathogenic,
  if !(host="None" or host="No" or host="") |
    !(disease="None" or disease="No" or disease="")
non-pathogenic,
  if (host="None" or host="No") |
    (disease="None" or disease="No") |
    (associated="Multiple" &
    (host="No" or host="None") &
    (disease="None" or disease="" or disease="No"))
symbiotic,
  if (associated="Host-associated" &
    (disease="None" or disease="No" or disease=""))
and NA,
  if (associated="Multiple" &
    host="" &
    (disease="None" or disease="" or disease="No")) |
  otherwise
```

Combination of phenotypic data from different resources Inconsistencies of the phenotype information from the three public resources are solved by the pathogenicity rule:

$$pathogenic > symbiotic > non - pathogenic > NA \tag{2.1}$$

The underlying assumption is that a pathogenic/symbiotic phenotype is very likely to be annotated by any researcher due to direct evidence, e.g. an observed interaction or based on experiment. Simultaneous assignment of a non-pathogenic annotation of the particular organism in an other resource was found to be more likely the result of missing information about a pathogenic/symbiotic context.

Results The majority of completely sequenced bacterial genomes in RefSeq are represented in the microbial databases. For 1706 of 1920 genomes phenotypic information is available. The according 214 genomes without information on pathogenicity are excluded from the genome repository. Of the three public resources, the GOLD database offers the most comprehensive phenotypic annotation, see 2.2. In 414 cases, GOLD offers the only available information. BacMap and NCBI phenotypes include phenotypes of 58 genomes not listed in GOLD and contribute annotations for 1234 genomes.

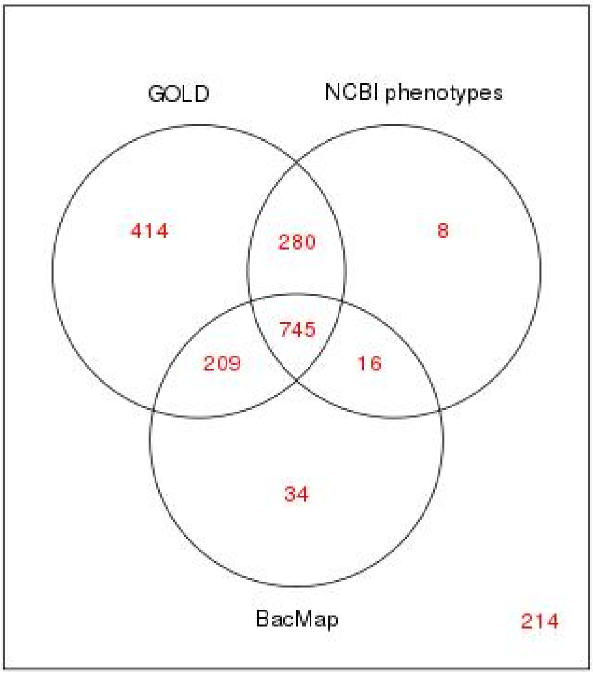


Figure 2.2: ***Publicly available phenotype information on the pathogenicity of completely sequenced bacterial organisms.*** *Shown is the number of sequenced bacterial organisms for which information on pathogenicity is available in any of the public databases GOLD, BacMap and NCBI phenotypes. For 214 bacterial genomes in the RefSeq database, there is no pathogenicity annotation available. The genomes are excluded from the genome repository.*

For 586 genomes, there is consistent information in all three databases. 858 genomes are consistently annotated in any two databases or have information on pathogenicity of the organism in at least one resource. GOLD and NCBI phenotypes show a general high agreement on individual phenotypes. All three resources contribute information to the comprehensiveness of the phenotype annotation. For 251 bacterial genomes in RefSeq, there are inconsistencies between phenotype information for identical strains 2.4. Extensive manual inspection of these cases shows that application of the combined pathogenicity rule 2.1 provides assignment of the correct pathogenic phenotype. The combined pathogenicity rule corrects inconsistent phenotypes in all of the inspected cases. For example for the opportunistic pathogen *Novosphingobium aromaticivorans DSM 12444* [282]. Bacmap has no information on *Novosphingobium*, the NCBI phenotypes list offers the misleading non-pathogenic annotation while the GOLD database provides a comprehensive characterization (Habitat="Fresh water,Host", Host="Homo sapiens") leading to the correct pathogenic annotation.

It was further investigated if certain habitats (e.g. as "Freshwater" or "Sediment") are prone to harbor exclusively non-pathogenic bacteria. For habitats with a representative high number of bacteria annotated (cutoff >= 20), the results indicate that there is no habitat completely free from pathogenic bacteria.

The taxonomic distribution of genomes in the genome repository for different taxonomic levels is shown in appendix table 8.1. Regarding the phenotype annotation of bacterial genomes, the genome repository is biased towards pathogenic/symbiotic bacteria, see appendix table 8.2. The final genome repository holds phenotype annotation of 1706 bacterial genomes. 796 are of pathogenic phenotype, 335 symbiotic and 575 non-pathogenic (2.3).

Discussion What makes a bacterial organism pathogenic is still an issue discussed extensively by microbiologists [221]. Pathogens and symbionts depend on similar mechanisms for interacting with hosts and show parallel trends in genome evolution [208]. Furthermore, by the ongoing discovery of novel pathogenic relationships, it can be assumed that the current data does not allow for a comprehensive classification. Even for many known bacterial organisms a mutualistic or pathogenic relationship to a specific eukaryotic host might exist but has not yet been discovered experimentally [60]. Up to now, there is no gold standard for separating bacterial organisms into pathogenic, symbiotic and non-pathogenic phenotypes that would enable automated classification. The integration of the different microbal databases offers a comprehensive annotation of bacterial phenotypes for the majority of completely sequenced bacterial genomes. In a considerable number of instances, pathogenic and symbiotic annotations are ambiguous. These inconsistencies reflect the general issue in the annotation of bacterial phenotypes [48]. Considering these challenges and the currently available data, pathogenic and symbiotic annotations are consequently combined for further analysis regarding the identification of eukaryotic-like protein domains. The quality of the phenotypic data is thereby adequate for the intended

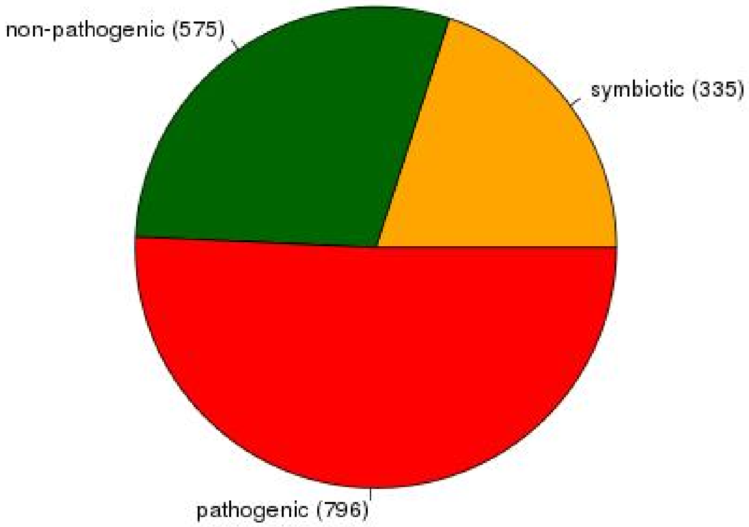


Figure 2.3: ***Distribution of pathogenicity phenotype annotations of completely sequenced bacterial organisms in the genome repository.*** *Shown are the fractions of pathogenic, symbiotic and non-pathogenic phenotypes in all completely sequenced bacterial genomes in the genome repository.*

analysis. Pathogenic and symbiotic bacteria are defined as host-interacting bacteria, while the class of non-pathogenic organisms comprises those bacteria that are non-host associated and not known to cause infectious symptoms in any host. This allows to distinguish non-host associated bacteria from mutualistic symbionts that interact but do not trigger an immune response in the host.

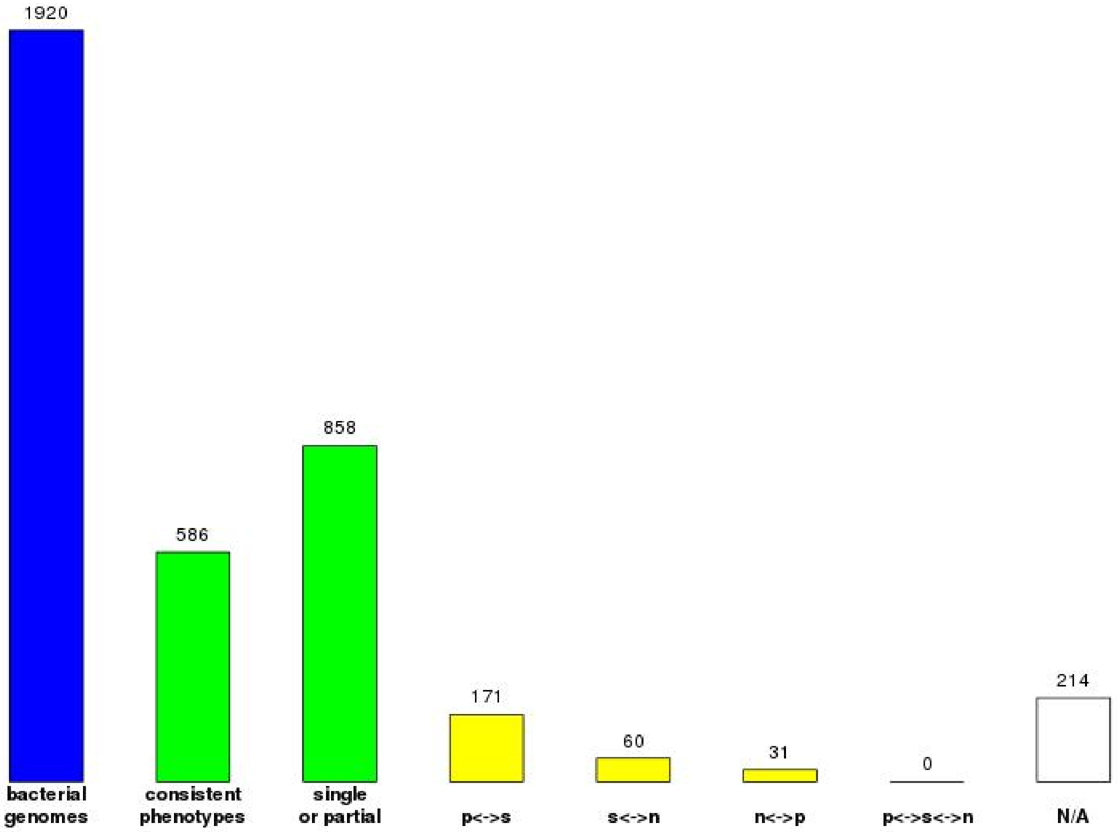


Figure 2.4: ***Consistency of phenotype information on pathogenicity in the public databases GOLD, BacMap and NCBI phenotypes for all completely sequenced bacterial genomes in RefSeq.*** *Green and yellow bars indicate the number and consistency of genomes with phenotype information from different resources, were p is for pathogen, s for symbiont and n for non-pathogen phenotype annotation. Conflicting annotations are solved according to rule 2.1. For 214 organisms there are no terms annotated or annotation terms do provide no relevant information on pathogenicity.*

2.3.2 Bacterial contamination of eukaryotic genome sequences

Contamination of sequencing projects by bacterial sequences is a serious challenge to the correct annotation of genomes [304, 257, 188]. This contamination is different from contamination by bacterial vector sequences and is commonly due to laboratory conditions or the amplification of bacteria that life in association with the sequenced eukaryotic organism. Bacterial vector sequences amplified during the sequencing process are commonly compared against a predefined vector libraries and filtered out with high accuracy. That approach is not feasible regarding bacterial contamination due to environmental factors. Contaminated genomes are expected to include protein sequences of bacterial origin which are otherwise not found in eukaryotic genomes. Protein domain signatures that are found only in a very low number of eukaryotic genomes are likely to account for that phenomenon, due to the evolutionary stability of eukaryotic genomes. The occurrence of domain singletons among eukaryotic genomes is therefore expected to reflect bacterial contamination. In the planned identification of eukaryotic-like protein domains this could be a critical source for false positive predictions and is addressed in the following.

Material and methods For 121 eukaryotic genomes in the genome repository, PFAM protein domain annotations are downloaded as listed by Interpro (version 40.0). 7411 different PFAM domain signatures are detected among the 121 eukaryotic genomes. To analyze bacterial contamination, the occurrence of domain signatures among eukaryotic genomes is evaluated with a regard to protein domain singletons. Domain singletons are therefore defined as protein domains that are detected only in one single eukaryotic genome and bacterial genomes otherwise. Especially a high accumulation of domain singletons in very few eukaryotic genomes could be expected not to be due to genomic properties. It could be a strong indicator of contamination by sequences of bacterial origin. Available literature on identified candidate genomes is examined further to confirm bacterial contamination.

Results The 7411 protein domain signatures show a wide range of frequencies among eukaryotic genomes (see figure 2.5). Noticeable is the high number of domain singletons. These domains are found only in proteins of one or two eukaryotic genomes. The distribution of domain singletons among eukaryotic genomes shows that domain singletons accumulate in 5 eukaryotic genomes in particular (see table 2.2). This is a strong indication for contamination. These genomes are therefore excluded. In Figure 2.5 A) and B), the distribution of domains in eukaryotic genomes is shown before and after the exclusion of the 5 contaminated eukaryotic genomes. A reduction of domain singletons is clearly visible, while domain singletons still remain the largest fraction.

Eukaryotic genome	domains (frequency 1)	domains (frequency 2)
Caenorhabditis japonica	***181***	***104***
Populus trichocarpa	***168***	***103***
Physcomitrella patens	***55***	***26***
Nematostella vectensis	***48***	***32***
Caenorhabditis remanei	***32***	***27***
Trichomonas vaginalis	23	8
Dictyostelium discoideum	20	10
Tetrahymena thermophila	16	6
Plasmodium yoelii	15	7
Trypanosoma brucei	13	14
Trichoplax adhaerens	13	14
Oryza sativa Indica	12	10
Gibberella zeae	12	5
...	...	...
Lachancea thermotolerans	1	1
Homo sapiens	1	0
Gasterosteus aculeatus	1	3
Gallus gallus	1	4
Cryptosporidium hominis	1	1
Canis lupus familiaris	1	1

Table 2.2: ***Eukaryotic genomes and the number of domain singletons (exclusively found in that particular genome /in two genomes).*** *Shown are eukaryotic genomes and the number of domains which are annotated in the specific genome but in no other eukaryotic genome (domain frequency of 1 over all 121 analyzed eukaryotic genomes), as well as the number of domains which are annotated in only two eukaryotic genomes (domain frequency of 2).*

Contrary to the analysis of domains singletons in eukaryotic organisms, singletons observed in bacterial genomes show a different distribution and are not due to a small number of genomes. As shown in table 2.3, domain singletons in bacteria are uniformly distributed over all genomes.

To further reduce the influence of bacterial contamination, an additional frequency cutoff for eukaryotic domains is applied : A domain is only considered for further analysis if it is annotated in at least 3 of the remaining 116 eukaryotic genomes.

Altogether, these pre-processing steps exclude 9.3% of domain signatures in eukaryotic genomes.

Bacterial genome	domains (frequency 1)	domains (freq 2)
Legionella longbeachae NSW150	6	0
Stigmatella aurantiaca DW4/3-1	6	2
Sorangium cellulosum 'So ce 56'	5	3
Planctomyces brasiliensis DSM 5305	4	0
Marinitoga piezophila KA3	4	0
Carboxydothermus hydr. Z-2901	4	0
Bacteriovorax marinus SJ	3	0
Haliangium ochraceum DSM 14365	3	1
Arcobacter sp. L	2	0
Terriglobus roseus DSM 18391	2	0
...	...	...

Table 2.3: ***Bacterial genomes and the number of domains which are exclusively found in that particular genome / in two genomes.***

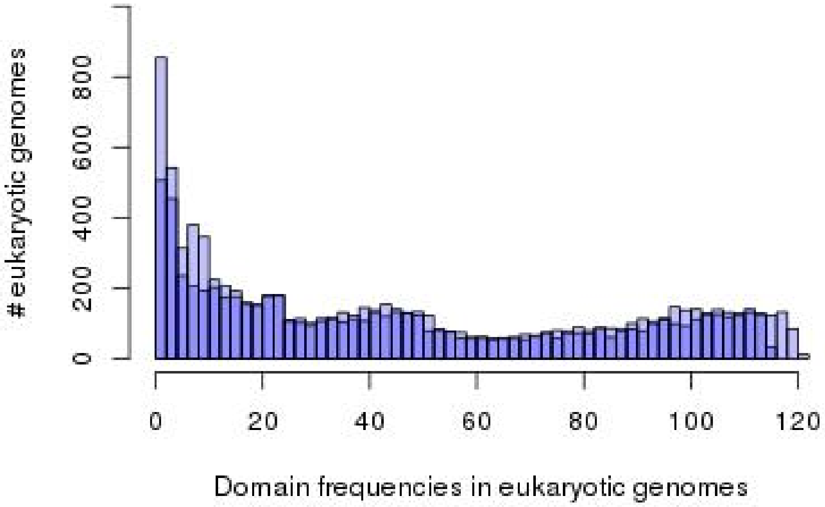


Figure 2.5: ***Frequencies of protein domain signatures in eukaryotic genomes.*** *Domain frequencies and their occurence before (in light blue) and after (in dark blue) removal of 5 vector-contaminated eukaryotic genomes.*

Discussion Domain singletons indicate bacterial contamination of several eukaryotic genomes. Assuming bacterial contamination of sequencing projects to be a rare event, contamination is expected to affect a small number of eukaryotic genomes. As expected, domain singletons were found to accumulate in five eukaryotic genomes which are subsequently removed from the data. Additionally, genomes were filtered for rare domains that could also be a result of contamination as well as lateral gene transfer between bacteria and eukaryots. The remaining protein domain signatures form the basis for a comparative analysis in eukaryotic genomes. With these quality filtering steps applied, 6869 different protein domains in 116 eukaryotic genomes are considered for further analysis.

2.3.3 Criteria for the prediction of a functional T3SS machinery

Published secretion system based effector prediction methods can model the T3 secretion signal quite accurately. On the other hand, these approaches were shown to predict T3 secreted proteins in any bacterial genome, independent of the organisms ability to secrete effectors via a Type III secretion system (T3SS) [20]. A comparison of existing effector prediction approaches with results of the subsequent analysis is only feasible for bacteria with a T3 secreting phenotype. Part of the annotation of genomes in the genome repository therefore is to identify bacteria with a functional Type III secretion system (T3SS). Even while being the best studied secretion system, the T3SS machinery is far from being fully understood [233]. Up to now, there are no comprehensive experimental studies on the exact assembly of the system. But there is evidence that not all characterized molecular T3SS components are required for successful transport of substrates [80]. The minimal set of key components necessary for a fully functional system were not yet determined experimentally.

In general, bacterial gene–phenotype predictions can be addressed using cross-species distributions of genes and phenotypes/traits [265]. Applications like CPAR are successful in predicting microbial traits (e.g. aerobe, anaerobe, thermophile or gram status) from genome data [181]. Thereby, genotype-phenotype association rules are learned from a matrix of clusters of orthologous groups (COGs). Necessary for unsupervised learning techniques is a high number of characterized training instances. A bacterial T3 secreting phenotype is experimentally characterized for few microorganisms in single case studies [81]. Extensive literature search as well as experimental studies might be needed to establish a sufficiently large training set of fully sequenced T3 secreting bacteria. This is not within the scope of the analysis at hand. With the current lack of experimental evidence on T3 secreting bacteria, functional genomics could offer another way to predict the correct microbial trait. On the basis of a small number of genomes, empirical criteria to predict a functional T3SS could be deduced from genes encoding specific T3SS components. In the following, a small collection of reference genomes that are assumed to harbor a functional T3SS are analyzed to predict empirical criteria for bacterial Type 3 secretion.

Material and methods A functional Type III secretion system (T3SS) is investigated for all genomes of Gram-negative bacteria with annotated phenotypes in the genome repository. Information on the Gram status of bacterial organisms is derived from meta-data of the GOLD database [212].
To chose the most suitable resource regarding completeness and sensitivity on T3SS related orthologous groups, different public databases are compared. The KEGG database holds comprehensive information about orthologous proteins involved in bacterial secretion systems [146]. The T3SS reference pathway KO03070 lists 15 different protein subunits which take part in the assembly of this macro-molecular machinery. The automatically generated clusters of orthologous groups in the eggNOG database [196] are considered as well and compared to the semi-automatically generated KOs in KEGG. Relevant KOs in KEGG (current version of august 2013) as well as according TTSS related orthologous groups in eggNOG (Version 3.0) are downloaded. Based on these 15 orthologous groups, a comparison of KEGG and eggNOG data is conducted to determine a suitable and comprehensive data source. Mapping between groups in KEGG and eggNOG considers the assignment of identical sequences to groups in both databases. To evaluate the empirical criteria to determine T3SS harbouring bacteria, the reference list for organisms with functional T3SS system published by Arnold et al [20] is used to construct a set of reference genomes encoding a functional T3SS machinery. Phenotype annotations are retrieved for all genomes. Non-pathogenic phenotype annotations are additionally verified by manual literature search.

Results KEGG KOs as well as eggNOG orthologous groups cover all Gram-negative bacteria with pathogenic/symbiotic phenotype in the genome repository. Comparing KEGG KOs and eggNOG orthologous groups, the representation of T3SS component in KEGG KOs and orthologous groups of eggNOG reveals differences important for further analysis.

For several T3SS components, eggNOG COGs and KEGG KOs are in good accordance and comprise respective proteins well. Clusters of the T3SS components yscW, yscL and yscQ reveal differences. Proteins are split onto several clusters and the major eggNOG

T3SS comp. in KEGG	# orthologs in KEGG	# split on clusters in eggNOG	cluster with highest coverage in eggNOG	coverage in eggNOG	cluster size in eggNOG
yscC	174	1	COG1450 (Type II secretory pathway, PulD)	174	676
yscV	188	1	COG4789 (Type III secretory pathway, EscV)	188	188
yscJ	190	2	COG4669 (Type III secretory pathway, EscJ)	189	189
yscN	188	2	COG1157 (Flagellar biosynthesis/ type III secretory pathway ATPase)	187	758
yscR	190	3	COG4790 (Type III secretory pathway, EscR)	173	179
yscT	192	4	COG4791 (Type III secretory pathway, EscT)	186	189
yscU	190	4	COG4792 (Type III secretory pathway, EscU)	99	99
yscS	184	5	COG4794 (Type III secretory pathway, EscS)	161	163
yscQ	172	6	COG1886 (Flagellar motor switch/ type III secretory pathway protein)	129	790
yscL	98	8	COG1317 (Flagellar biosynthesis/ type III secretory pathway protein)	60	438
yscW	88	5	NOG04987 (Protein involved in negative regulation of protein secretion)	40	40
yscF	86	7	NOG08886 (Protein involved in pathogenesis)	25	25
yscO	23	3	NOG149544 (Type III secretion protein)	18	18
yscP	14	4	NOG41949 (no description)	10	14
yscX	22	3	NOG14224 (Type III secretion protein)	17	17

Table 2.4: ***Representation of T3SS component KOs in eggNOG orthologous groups.*** *Coverage of the 15 KOs in the KEGG database that resemble the T3SS machinery. Given is the coverage of these KOs by genomes of pathogenic/symbiotic bacteria as well as the representation of the respective proteins in the orthologous groups of the eggNOG database. Identifier and description of the cluster with the highest coverage as provided by eggNOG version 3.0.*

cluster in each case contains only a small number of the respective KEGG orthologs. For yscF, yscO and yscX there are no COGs constructed in Eggnog, the proteins are split on diverse non-supervised orthologous groups (NOGs) in eggNOG. Besides proteins of the respective T3SS components, the eggNOG COGs COG1886, COG1317 and COG1157 include many more proteins that are part of the flagellar apparatus. The KEGG KOs on the other hand do distinguish between T3SS components and homologs of the evolutionary related flagellum. Eggnog also does not resolve the relationship of T3SS component yscC with other bacterial transport systems. COG1450 combines homologous proteins of yscC and a component of the Type II secretion system, PulD as well as the general secretion pathway protein D. For an overview of the mapping between clusters of orthologous groups in KEGG and eggNOG, see table 2.4. The reference set contains genomes encoding a functional T3SS machinery. T3SS component KO coverage by proteins of 87 reference genomes are shown in figure 2.6. Several KOs are represented in a low number of genomes. Only 8 of the 15 KOs are represented in the majority of genomes: yscC(K03219), yscJ(K03222), yscN(K03224), yscR(K03226), yscS(K03227), yscT(K03228), yscU(K03229) and yscV(K03230).
The minimal set of KOs that comprise all reference genomes while being maximally specific is an arbitrary selection of 7 out of these 8 KOs.
Applying this empirical criteria for a functional T3SS to all genomes in the genome repository, 165 Gram-negative bacteria of pathogenic and 23 of symbiotic phenotype are identified to encode a functional T3SS. All pathogenic/symbiotic bacteria of the reference set are included. Furthermore, also 14 non-pathogenic bacteria were identified to encode a functional T3SS.

Of the 1706 genomes in the genome repository, according to KEGG Orthologous Groups, 202 are predicted to encode a functional T3SS. The vast majority are of pathogenic phenotype (73), 9 organisms are symbionts. 14 organisms are annotated with non-pathogenic phenotypes, but manual literature search reveals a symbiotic relationship for 7 of them. Among the bacteria with functional T3SS and a manually verified non-host interacting phenotype annotation extracted from literature are Anaeromyxobacter sp, Myxococcus Xanthus DK 1622, Shewanella violacea strain DSS12, see table 2.5. The marine bacterium Hahella chejuensis is an algicidal pathogen [142].

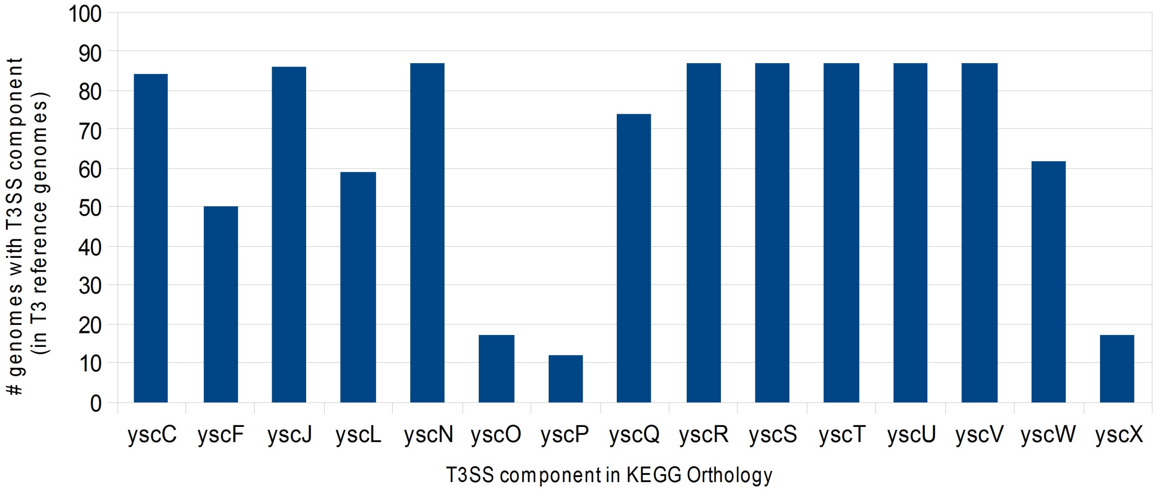


Figure 2.6: ***Coverage of Type III Secretion System (T3SS) components in organisms.*** *Groups of orthologous proteins are named by the according KEGG KO identifier.*

Organism	Essential T3SS components
Anaeromyxobacter dehalogenans 2CP-1	7
Anaeromyxobacter dehalogenans 2CP-C	7
Anaeromyxobacter sp. Fw109-5	7
Anaeromyxobacter sp. K	7
Myxococcus xanthus DK 1622	7
Burkholderia thailandensis E264	8
Collimonas fungivorans Ter331	8
Hahella chejuensis KCTC 2396	8
Marinomonas mediterranea MMB-1	8
Shewanella baltica OS155	8
Shewanella baltica OS195	8
Shewanella violacea DSS12	8
Variovorax paradoxus EPS	8
Vibrio sp. Ex25	8

Table 2.5: ***Bacteria without evidence for a host-dependent lifestyle in public databases that encode a majority of essential T3SS components.*** *Shown are bacteria that are annotated as non-pathogenic and encode all essential components of a Type III secretion system. Literature suggests a possible host interaction for some organisms.*

Discussion Limited experimental knowledge on details of the T3SS secretion mechanism makes the prediction of empirical criteria for the assignment of a functional Type III secretion system necessary.

For addressing the gene-phenotype association problem, the use of clusters in the KEGG Orthology (KOs) revealed an advantage over orthologous groups in eggNOG regarding the analysis at hand. The resolution of the approach used to generate KEGG KOs was found to be more optimal, due to the broad evolutionary scale applied in the construction of orthologous groups in eggNOG. EggNOG reconstructs the last universal common ancestor (LUCA) of the kingdom of bacteria for COG construction. At this evolutionary stage, the T3SS did not yet evolve from the closely related flagellum and components of both systems are combined in single common clusters. KEGG KOs are an accepted resource for molecular systems and subsystems. The coverage of T3SS component KOs identified by KEGG provided the possibility of an automatic classification of bacterial genomes regarding T3SS secretion. Based on empirical criteria, a functional T3SS is assigned to Gram-negative bacteria. An arbitrary selection of 7 out of 8 specific components of the T3SS are predicted to be essential for a functional Type III secretion system. On a structural level, this could imply that the T3SS machinery is robust towards missing protein subunits. Functionality seems still granted in the absence of several peripheral components.

An interesting observation can be discussed regarding the phenotype of microorganisms with T3SS encoding genomes. Certainly, not all Gram-negative host-associated bacteria can be expected to encode a functional Type III secretion system. To establish host contact, several pathogens/symbionts for example use the twin-arginine translocation (Tat) pathway [71], a prokaryotic transport system found among all bacterial phenotypes [82]. Several host-interacting bacteria also encode other secretion systems that are commonly associated with virulence (T4SS, T6SS, T7SS). On the other hand, the T3SS can be assumed to be very costly for a microorganism. According to current knowledge, the evolutionary pressure to conserve this complex machinery within a bacterial genome is the result of close interaction with eukaryotic host cells. A functional T3SS could therefore be assumed to be an indicator of a microbial pathogenic/symbiotic lifestyle. Regarding T3SS based prediction in Gram-negative bacteria, several organisms were found to be non-pathogens without evidence for any host-interaction while encoding a functional T3 secretion system. It is not likely to find a functional T3SS in non-host-interacting microorganisms but cannot be excluded and is tolerable for a small number of genomes. While this phenomena to some extend also might be the result of missing information about the bacterial lifestyle, it also implies that even the best studied bacterial secretion system is not yet a genetic basis for the accurate assignment of a host-interacting bacterial phenotype. The assignment of a pathogenic/symbiotic bacterial phenotype based on genomic features is addressed further in chapter 5.

2.4 Phenotype specificity of eukaryotic-like protein domains in bacteria

The eukaryotic-like subset of protein domains in bacteria can only be defined with strong respect to the biological background. A number of domain signatures are exclusive to pathogenic/symbiotic bacteria and do not appear any bacteria annotated as non-pathogenic. Eukaryotic-like domains like the F-Box motif which are exclusive to pathogenic/ symbiotic bacteria are the most promising candidates for a general role in effector proteins and widespread infection strategies. In general, the boundaries might be less clear. Horizontal gene transfer (HGT) is considered one of the sources of eukaryotic-like domains [179]. Besides between bacteria and eukaryots, HGT events are observed between bacteria of divers phenotypes. Considering the genomic flexibility of bacteria, it is likely that domains are frequently transferred between organisms of different phenotypes [321]. The observed signal of protein domains in bacteria that are exclusive to a particular phenotype therefore is expected to be only an episodic. Because of bacterial genome flexibility, the phenomenon of phenotype exclusive protein domains might be largely due to the limited number of sequenced genomes. The more genomes are considered, the less domains are exclusive to any given phenotype. If the signal is episodically, an increase of fully sequenced bacterial genomes would lead to a decrease of protein domains exclusive to pathogens/symbionts.

Methods To investigate the nature of this phenomenon, domain signatures and their frequencies among proteins in bacteria of different phenotype are inspected. Protein domain signatures are determined in genomes for all bacteria with given information about a annotated phenotype. Domains in pathogenic/symbiotic bacteria are compared to domain signatures in non-pathogenic/non-host-interacting bacteria. During comparison, the number of considered non-pathogenic bacteria is continuously increased, simulating the expected increase in available fully sequenced genomes. In each iteration, 50 genomes of non-host-interacting bacteria are randomly selected and used for comparison. The number of protein domains/eukaryotic-like domains which are exclusive to host-interacting bacteria is determined in each iteration. In each iteration, domains which appear only in genomes of pathogenic/symbiotic phenotype are separated into domains that are only found in bacteria and domains that are also detected in several eukaryotic genomes. The development of both trends is analyzed over all iterations.

Results Considering all bacterial genomes with annotated phenotype, of the protein domain signatures found to be frequent in eukaryots, 176 do occur only in pathogens/ symbionts and not in non-pathogens. In figure 2.7, the development of these pathogen/ symbiont specific domains is visualized for increasing the number of considered bacterial genomes gradually. For an increase of considered bacterial genomes, the number of phenotype specific protein domains shows a steady decline. This trend is strong in the beginning and still visible when including all bacterial genomes with annotated phenotype to date.

Investigation of specific effector domains does provide further insights to define appropriate criteria for the identification of eukaryotic-like protein domains. For example, the Sec7 domain (PF01369) is a guanine-nucleotide-exchange-factor (e.g. of ADP ribosylation factor) [140], present in proteins of 54 eukaryots. In the kingdom of bacteria it is well conserved in characterized effectors of Legionella and Rickettsia species [204]. Cox et al proposed that bacterial Sec7 domain-containing proteins result from two horizontal transfer events: the first one from eukaryotes to bacteria, and the second between Legionella and Rickettsia [67]. In the case of Sec7, the eukaryotic-like domain is observed exclusively in proteins of pathogenic and symbiotic bacteria. For many other domains in bacterial virulence factors with a suggested eukaryotic background suggested by literature, the observed frequencies between phenotypes are less explicit. Several of these domain signatures are present in a high number of pathogenic/symbiotic bacteria with occurrences also in non-pathogenic/non-host-interacting bacteria.

Ankyrin repeat domain, PF00023: The Ankyrin repeat binding domain is well characterized in eukaryotic-like effector proteins of divers pathogens [214]. It can interact with the host cytoskeleton and enables a variety of functions other in the host cell [223, 301]. But also in non-pathogenic, free-living bacteria Ank-containing proteins are widespread and serve as an unspecific protein binding motif [7]. The ankyrin signature PF00023 is recognized in 186 pathogens/symbionts as well as in 132 non-pathogenic bacteria.

Ribosome inactivating protein domain, PF00161: Ribosome inactivating proteins with domain signature PF00161 act as toxins in eukaryotic cells by inhibiting protein synthesis and subsequent apoptosis [277]. They are found in bacteria and plant [183]. The shiga toxin protein released by several E.coli strains contains this domain signa-

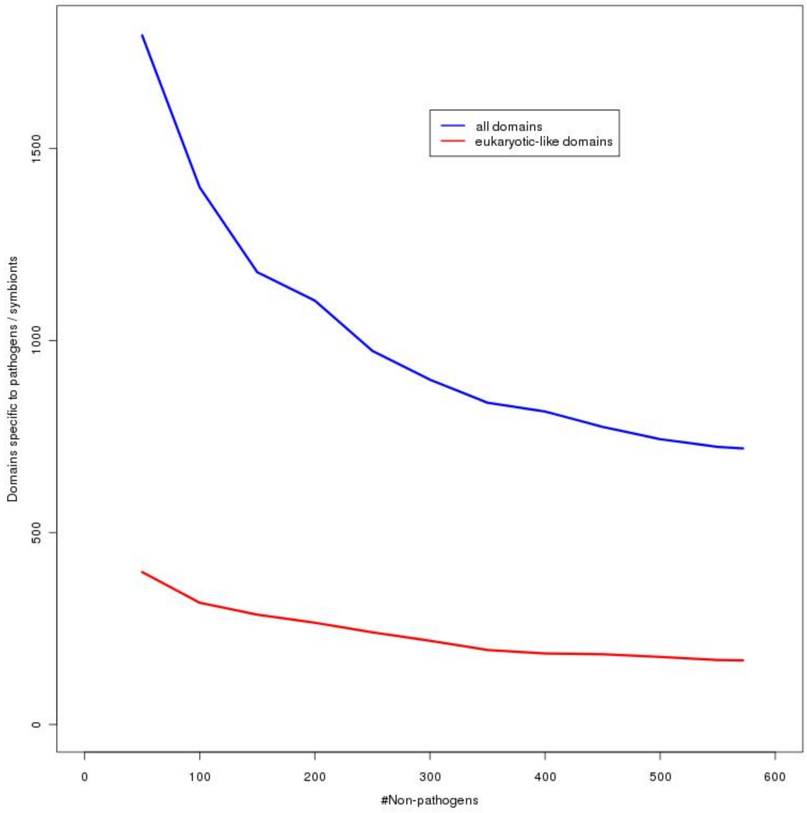


Figure 2.7: ***Trend of the number of pathogen/symbiont specific protein domains for an increasing number of considered non-pathogenic genomes.*** *Shown is the number of domains which occur only in pathogenic and symbiotic bacteria while considering protein domains from a gradually increasing number of non-pathogenic bacteria. The blue line reflects all pathogen/symbiont specific protein domains, the red line visualizes the development of the subset which is found in eukaryotic genomes as well.*

ture [289]. The domain is annotated in the genomes of 24 pathogens/symbionts and one non-pathogen: The *Streptomyces coelicolor A3(2)* strain, taxonomically a member of the *Streptomyces violaceoruber* genus, is a non-pathogenic filamentous soil bacterium [28]. **Ubiquitin carboxyl-terminal hydrolase domain, PF00443:** The Ubiquitin carboxyl-terminal hydrolase domain (UCH) is well conserved among Burkholderia species as well as several other pathogenic and symbiotic bacteria. Memisevic et al analyzed UCH-containing virulence factors in *Burkholderia Mallei* and linked it to specific host-pathogen protein interactions [187]. A protein containing the UCH domain signature is also annotated in the free-living, nonpathogenic bacterium *Chitinophaga pinensis DSM 2588* [116]. **Pentatricopeptides, PF01535** Eukaryotic-like pentatricopeptides (PPR, PF01535) are involved in virulence of 59 pathogens, well characterized for example in TTSS secreted effectors of the soil-borne plant pathogen *Ralstonia solanacearum* [194]. Furthermore, PPRs are conserved in proteins of 5 non-pathogenic bacteria, e.g. in strains of the thermophilic bacterium *Rhodothermus marinus*, isolated from submarine hot springs [10]. Exclusive phenotype specificity of eukaryotic-like domains in bacteria is an episodic phenomenon. Several protein domains of characterized bacterial virulence factors that are frequent in genomes of eukaryotic organisms are not exclusive to the genomes of bacterial pathogens and symbionts. These domains are also found in non-pathogenic bacteria.

2.4.1 Discussion

Most eukaryotic-like protein domains of characterized bacterial virulence factors are not exclusive to the genomes of pathogens and symbionts. These domains are also found in non-pathogenic bacteria. Possible reasons for the non-exclusivity of many characterized eukaryotic-like effector domains are supposed to be bacterial genomic flexibility combined with the limited access to genomic data. To address the influence of these observed characteristics in an automated large scale approach to identify eukaryotic-like protein domains, such a method has to measure not only exclusivity but phenotypic enrichment of protein domains. Also the influence of mistakes in the bacterial phenotype annotations that cannot be excluded as a possible error source completely, are handled by the enrichment approach. It is therefore proposed to consider not only protein domains exclusive to pathogenic/symbiotic bacteria but the enrichment of eukaryotic-like domains in these host-interacting bacteria.

2.5 Identification of eukaryotic-like domains enriched in pathogenic/symbiotic bacteria

2.5.1 Methods

Eukaryotic-like domain score calculation The background model for each domain is estimated by calculating the average and standard deviation of its frequencies in all non-pathogenic genomes. For each genome, the eukaryotic-like domain (ELD) score S of a domain is calculated as the number of standard deviations σ in which the domain frequency n in that particular genome differs from the average background frequency μ

in non-pathogen genomes:

$$S = (n - \mu)/\sigma \tag{2.2}$$

where μ is the mean and σ is the standard deviation of n. S represents the distance between the actual domain frequency and the mean in units of the standard deviation [159]. It is positive when above the mean. Thereby the ELD score directly reflects the enrichment of a particular eukaryotic-like domain in proteins of a particular genome. Although the distribution of domain occurrences across genomes has varying shapes, manual inspection of ELD scores shows that the ELD scores typically show the characteristics of Z-scores. Scores can be considered significant for values greater or equal to 4 [269].

The score of an effector protein candidate is equivalent to the ELD score of the contained eukaryotic-like domain. Regarding multi-domain proteins, the score of an effector protein candidate is equal to the ELD score of the highest scoring domain.

Additionally, a similar analysis was conducted considering domain distributions in orthologous groups of proteins. These reduced proteomes, only containing evolutionary conserved protein sequences have been analyzed by the same approach. Frequencies where determined accordingly and scores calculated for all genomes included in the eggNOG Clusters of Orthologous Groups [196].

2.5.2 Results

Eukaryotic-like domains in proteins of pathogenic/symbiotic genomes In 1103 genomes of pathogenic/symbiotic bacteria, 2504 different eukaryotic-like domains achieve an ELD score equal or above 4 in at least one genome. Many domains occur in proteins widespread among distantly related organisms, indicating long evolutionary history and broad functionality. Others are specific to and well conserved only among strains of particular pathogenic/symbiotic species, compare 2.6. Predictions result in 142934 effector proteins that have been predicted by the eukaryotic-like domain approach, containing one or more eukaryotic-like protein domains. The median average of effector candidates per genome is 37. In pathogenic and symbiotic bacteria, on average in 1.7% of the proteome eukaryotic-like domains are identified. In non-pathogenic bacteria, on average 1.5% of annotated proteins per genome do contain eukaryotic-like domain signatures. The surprisingly high number of positive predictions in these genomes could be due to several reasons. Besides false positive predictions, lateral gene transfer between bacteria of different phenotypes as well as undiscovered host-associations of symbiotic organisms annotated with a non-pathogenic phenotype cannot be completely excluded.

General features of the predicted ELDs and resulting effector protein candidates are discussed in the following regarding the eukaryotic-like protein domain content in strains of the model pathogen *Chlamydia trachomatis*.

Complementary results of signal and function-based prediction approaches For a comprehensive effector prediction analysis, results of the signal-based approaches EffectiveT3 (for predicting effectors secreted by the TTSS mechanism) and SignalP (capturing the signal for the Sec-pathway) are considered besides the eukaryotic-like domain method. TTSS signal peptides are predicted only for all 188 genomes of Gram-negative bacteria that encode a probably functional type III secretion system. For all 1103 pathogenic/ symbiotic genomes, 641225 effector candidates in total are predicted by at least one of the methods. As to be expected from the limited coverage of any of the prediction methods, many predictions are only supported by one or two methods, compare figure 2.8.

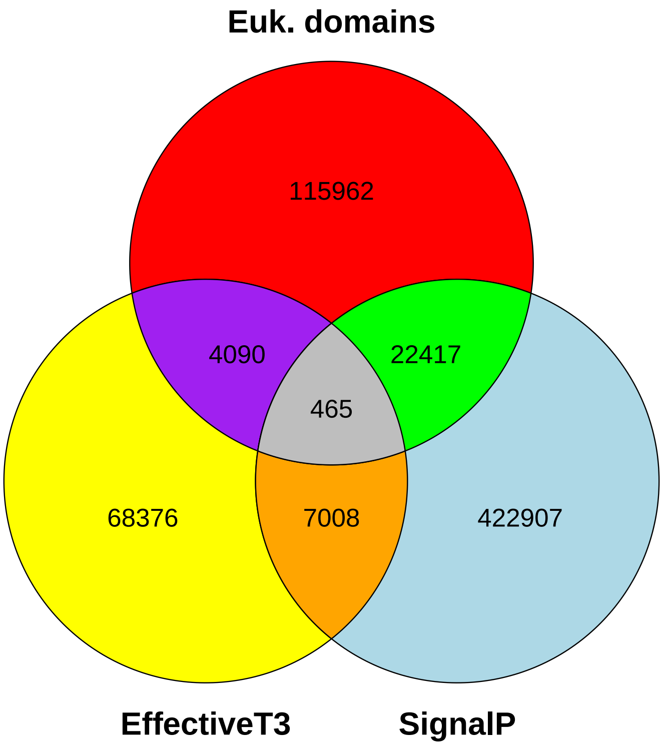


Figure 2.8: ***Numbers of predicted effectors in 1103 pathogenic and symbiotic genomes, indicated by supporting method.*** *EffectiveT3 has been applied to 188 genomes of Gram-negative bacteria encoding a Type III secretion system.*

For many proteins containing eukaryotic-like domains no secretion signal for the Sec and TTSS pathways is detectable. On the other hand, many proteins containing a secretion signal do not contain any well conserved protein domains, preventing the detection of eukaryotic-like function. This limitation is an intrinsic consequence of signature-based approaches.

The high overlap between the two signal-based approaches is unexpected, as signal peptides of the TTSS and Sec pathways should be incompatible to each other. Both signals are located in the N-terminus of the amino acid protein sequence and show no similarity. But since our understanding of the molecular recognition of effector proteins by the secretion systems is still limited, we are not yet capable of discarding one signal due to higher confidence of the other.

A surprisingly high number of predictions is only supported by SignalP (422 907) or EffectiveT3 (68 376). As the real number of secreted proteins is unknown for all genomes of pathogens and symbionts, the amount of false positives contributing to these predictions, cannot be reliably determined. Considering the limited accuracy of both programs, subsequent filtering (e.g. discarding functionally well-annotated, probably not-secreted proteins) could dramatically improve predictions. The ongoing improvement of bioinformatic tools for the identification of signal peptides will probably improve the situation within the next years.

2.5.3 Reported experimental validation of effector candidates predicted in the genomes of *Chlamydiae sp.*

In strains of the obligate intracellular human pathogen *Chlamydia trachomatis*, several eukaryotic-like domains (ELDs) where identified by the presented approach, results listed in table 2.6. These ELDs are found widespread among eukaryotic organisms and are enriched in divers other pathogenic/symbiotic bacteria outside of the genus *Chlamydia*. In an experimental study, Gehre et al. evaluated effector candidates predicted by Effective for the genomes of *C. trachomatis*, *C. caviae* and *C. pneumoniae*. In the lab of Agathe Subtil at the Institut Pasteur, eukaryotic-like domain containing effector candidates conserved in the *Chlamyida* species were tested for secretion by the Type III secretion system (TTSS), the predominant system for virulence in *Chlamyidales* [193]. The screening was performed in a heterologous TTSS of *Shigella flexneri* that had been shown to be functional for chlamydial effectors [278].

Gehre et al. reported the following observations [114]:

- 17/25 candidates are secreted by a TTSS, 8/25 were not secreted
- 8/25 tested proteins had additionally a predicted TTS signal, under which 6/8 proteins were indeed positive for secretion
- 4/6 proteins with an ELD score of 4 were secreted
- all 7 candidates that reached the maximum ELD score of 10 000 were secreted
- when homologs of different chlamydial species were tested, they always showed consistent results
- 70% (17/25) of the chlamydial proteins that had an enrichment of a eukaryotic-like domains were secreted by a TTSS. We cannot exclude the secretion of the remaining 30% of candidates by another secretion pathway.

Candidate proteins for the experimental validation of the eukaryotic-like domain based prediction of effector proteins were selected independently of Type III secretion. Candidates were chosen based only on results of the large scale eukaryotic-like domain identification (ELD) approach for *Chlamydia* species. This provides further evidence, that a secretion system independent prediction approach can be used to predict bacterial effector proteins.

2.5.4 Discussion

Experimental evaluation of effector candidates predicted by the eukaryotic-like domain (ELD) based approach shows that the prediction of secreted effector proteins works. Bacterial effectors from several organisms of the phylum *Chlamyidales* were predicted correctly by the unspecific, large-scale prediction approach based on eukaryotic-like protein domains. Wieder et al. investigated the evolutionary origin and phylogenetic relationship of eukaryotic-like protein domains in 70 randomly chosen bacterial pathogens/symbionts [307]. Especially the role of horizontal gene transfer (HGT) events for the emergence of eukaryotic-like protein domains was investigated, eukaryots functioning as donor and pathogens as acceptor organisms. A significant connection between horizontal gene transfer events and eukaryotic-like domains was stated for several pathogenic/symbiotic bacteria. On the level of domain similarities, evidence for HGTs being involved in the evolution of eukaryotic-like domains was detected for the majority of pathogens/symbionts. Also revealed was the fact that often the HGT donor of the domain sequence is not necessarily the direct eukaryotic host organism. The phylogenetic analysis provides additional reason to assume a general applicability of the ELD based effector prediction approach regarding

domain id	**description**	**ELD score** **in *C.trachomatis***	**# in p/s**	**# in n**	**# in euk**	**effector candidates** ***C.trachomatis***
PF02902	Peptidase C48, Ulp1 family	10000	88	0	108	CT867,CT868
PF07720	Tetratricopeptide repeat	4	154	11	15	CT576,CT862
PF09825	Biotin-protein ligase	4	68	17	37	CT035
PF01823	MACPF protein superfamily	7	44	12	55	CT153
PF01496	V-type ATPase	4	175	69	103	CT305
PF03690	UPF0160, Uncharacterised family	5	59	19	97	CT386
PF02201	SWIB/MDM2 domain	6	61	33	102	CT460,CT643

Table 2.6: ***Overview of eukaryotic-like protein domains identified in Chlamydia trachomatis sp.*** *For each eukaryotic-like domain the ELD score as well as the frequency in proteins of pathogenic/symbiotic bacteria (p/s), non-pathogenic bacteria (n) and eukaryots (euk) is given. Effectors candidates in* ***C.trachomatis*** *that were identified by the particular domain are provided by locus tags.*

bacterial pathogens/symbionts of diverse clades.

As an intrinsic consequence of the domain identification step, the predictive power of the method is limited to the subset of effectors that shows conserved domain sequence similarity. The ELD based effector prediction approach therefore is a powerful tool that is preferentially applied in combination with other, e.g. signal-based prediction methods.

3 The Effective web portal - Development of a comprehensive resource for the prediction of secreted bacterial proteins

3.1 Motivation

Recognition and characterization of effector proteins is key to the understanding of bacterial virulence and symbiosis. To select candidates for experimental analysis, *in silico* effector prediction approaches were shown to be an efficient way. Simultaneous application of complementary approaches additionally narrows down the range of effector candidates to be considered. Molecular biologists in this field of research could profit immensely from a comprehensive public resource for the prediction of bacterial effector proteins.
Several resources momentarily implement signal-based prediction methods for T3 secreted proteins [20, 248, 299] and the Sec pathway [225]. Up to now, there exists no public resource that provides a taxonomically universal, function-based effector prediction approach.
Such a resource should on the one hand side enable analysis of all currently characterized bacterial pathogens/symbionts by providing precalculated effectome data. On the other hand, it should provide tools to predict effector proteins in private, unpublished sequence data, based on a comprehensive set of prediction methods.
The Effective web portal represents a resource that combines both complementary function- and signal-based prediction approaches in one framework, to allow users a state-of-the-art comprehensive prediction of bacterial effectomes (http://effectors.org).

3.2 Functionality and requirements

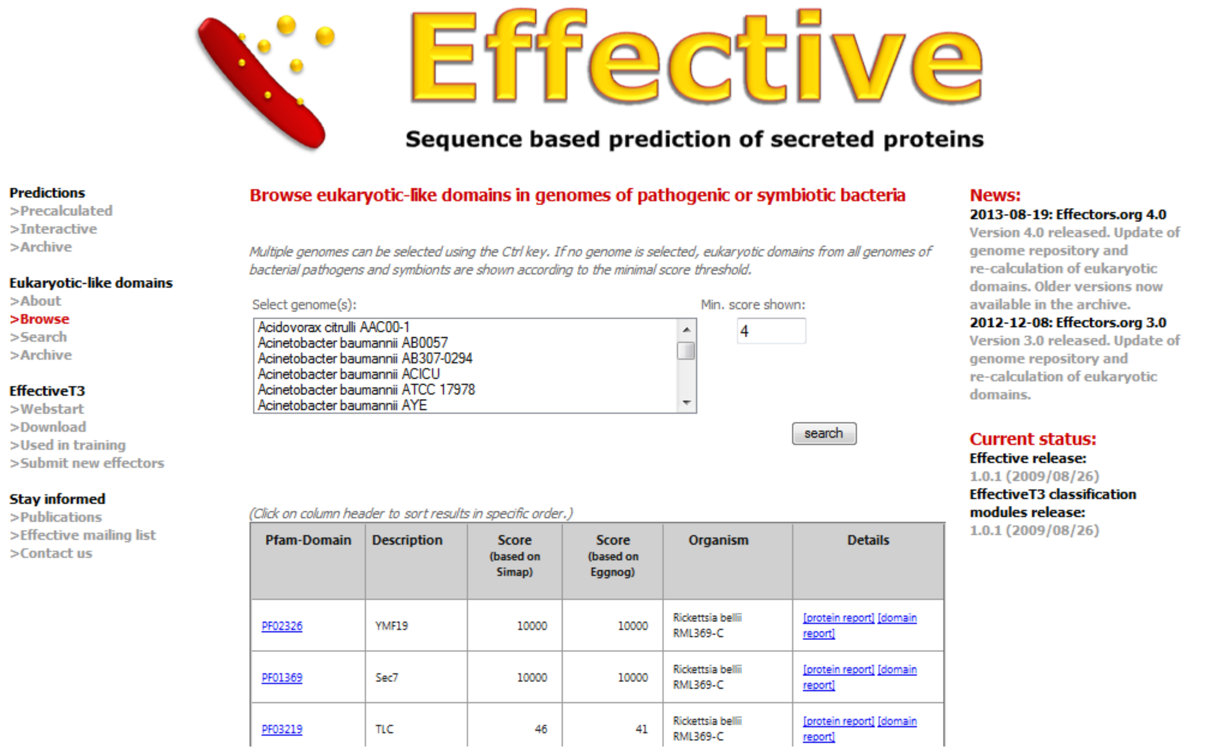
To be of maximal value to the research community, a comprehensive resource for the prediction of bacterial effectomes needs to meet the following requirements:

- **Complementary effector prediction approaches** are combined in a state-of-the-art comprehensive set of methods.
- **Precalculated effector candidates** accessible for all completely sequenced bacteria with pathogenic and symbiotic phenotype.
- **Interface to predict effectors on-the-fly** in sequence data provided by the user.
- **Userfriendliness** - the web site is easy to use and navigate, while comfortably providing all relevant information.
- **Up to date** - regular automated updates while keeping outdated versions accessible in an archived format.
- **Basic technical requirements and standard procedures** necessary for a comprehensive functional web resource, e.g. maintainability, extensibility and versioning.

In the following, the realization of these requirements in the implementation of the Effective web portal is described.

3.3 Concepts and implementation

Effective (http://effectors.org) represents a comprehensive resource of predicted bacterial secreted proteins. It provides information as well as interactive tools, e. g. for the selection of effector candidates for experimental analysis. The identification of eukaryotic-like protein domains (ELDs), based on the ELD score that results from the taxonomic occurrence of each domain, is a unique feature of the Effective web portal. Precomputed information within Effective is easily accessible for all publicly available completely sequenced genomes of bacterial pathogens/symbionts.
All descriptions of the web portal, the data content and integrated methods provided in this chapter refer to version 1.0 of the Effective database.



Browse eukaryotic-like domains in genomes of pathogenic or symbiotic bacteria

Multiple genomes can be selected using the Ctrl key. If no genome is selected, eukaryotic domains from all genomes of bacterial pathogens and symbionts are shown according to the minimal score threshold.

Select genome(s):
Acidovorax citrulli AAC00-1
Acinetobacter baumannii AB0057
Acinetobacter baumannii AB307-0294
Acinetobacter baumannii ACICU
Acinetobacter baumannii ATCC 17978
Acinetobacter baumannii AYE

Min. score shown: 4

(Click on column header to sort results in specific order.)

Pfam-Domain	Description	Score (based on Simap)	Score (based on Eggnog)	Organism	Details
PF02326	YMF19	10000	10000	Rickettsia bellii RML369-C	[protein report] [domain report]
PF01369	Sec7	10000	10000	Rickettsia bellii RML369-C	[protein report] [domain report]
PF03219	TLC	46	41	Rickettsia bellii RML369-C	[protein report] [domain report]

News:
2013-08-19: Effectors.org 4.0
Version 4.0 released. Update of genome repository and re-calculation of eukaryotic domains. Older versions now available in the archive.
2012-12-08: Effectors.org 3.0
Version 3.0 released. Update of genome repository and re-calculation of eukaryotic domains.

Current status:
Effective release:
1.0.1 (2009/08/26)
EffectiveT3 classification modules release:
1.0.1 (2009/08/26)

Figure 3.1: ***Screenshot of the Effective web portal.*** *The News section provides information about the current status and version of the Effective database. The current mainframe shows the interface for displaying genome specific information on eukaryotic-like domains.*

3.3.1 Software model of the Effective web portal

The software pattern realized in the Effective portal is a 3-tier architecture model. Realization of the 3 tier model in the Effective web portal is illustrated in figure 3.2.

To implement each of the individual layers, different technical concepts are applied:

Presentation tier: Java Server Pages (JSP) are responsible for creating HTML user interface pages with enhanced functionality by Javascript elements.

Business tier: The business logic is implemented in the Java programming language. Java classes are responsible for data access. These objects use Java Database Connectivity (JDBC) to query the underlying database.

Data tier: Data is stored in a relational database implemented in the MySQL database query language (http://www.mysql.org).

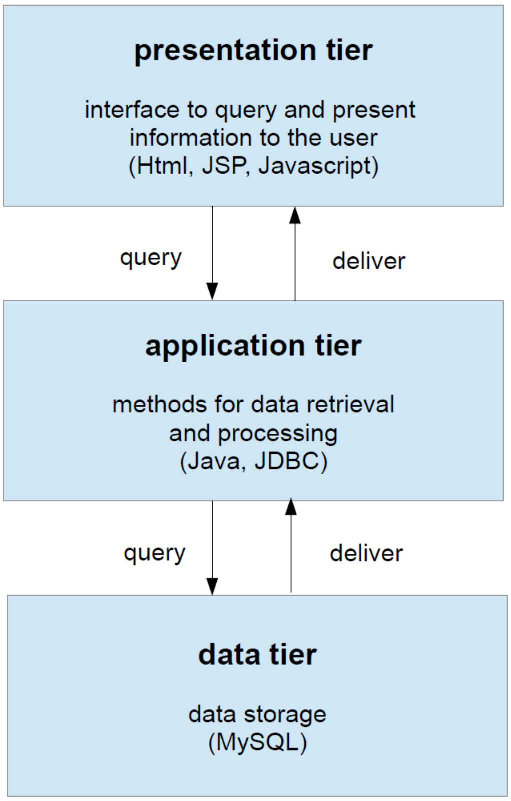


Figure 3.2: ***3-tier-architecture of the Effective web portal.*** *For each layer, required functionality and the concepts used for implementation are indicated. The model enables data querying from top to bottom, while data delivery between layers is possible in the reverse direction.*

Structure of the Effective relational database

A relational database stores all precalculated information about predicted bacterial effector proteins. The database uses a normalized table structure to minimize redundancies [62]. All tables of the Effective database and their relationships are visualized in the entity relationship model of figure 3.3.

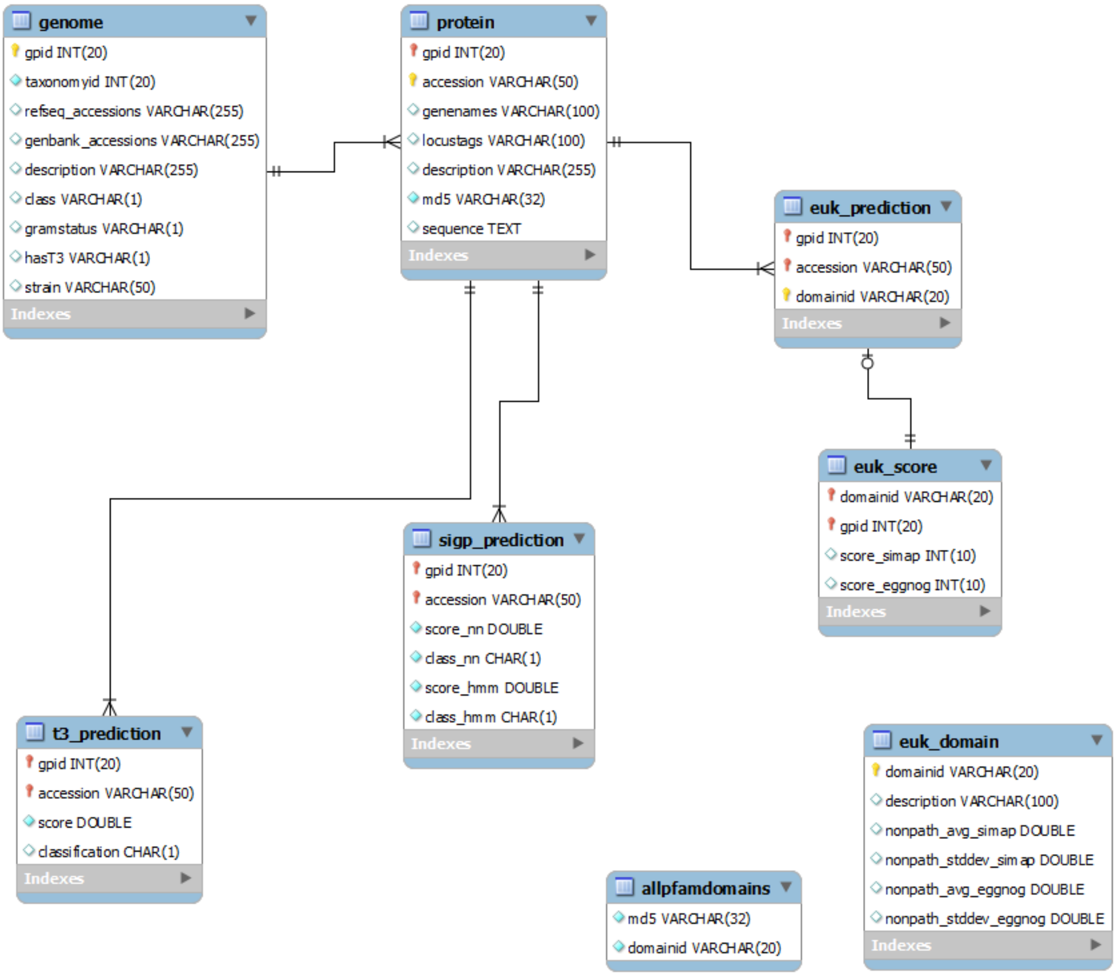


Figure 3.3: ***Entity relationship model of the Effective database.*** *The Effective database for data storage and retrieval by queries through the web portal. The euk_ domain and allpfamdomains tables are not accessible by the portal and used for calculation of eukaryotic-like domain scores and for the automatic update process.*

3.3.2 Combination of complementary effector prediction approaches

The portal provides easy to use tools for the analysis of characterized bacterial genomes and user specific protein sequences, e. g. proteins annotated in the genome sequence of a novel sequenced bacterial organism. Implemented are two complementary prediction strategies for protein secretion: the identification of eukaryotic-like protein domains (ELDs) and the recognition of signal peptides for TTSS and Sec-pathway in amino acid sequences.

Type 3 secretion signal - EffectiveT3: Type 3 secreted effector proteins are predicted via the EffectiveT3 method. From the numerous published T3 prediction softwares, this is the only one for which taxonomic universality of the predicted signal peptide is explicitly shown [20]. The software was implemented as a Java based stand-alone tool by Tanja Bieber and integrated into the Effective portal in a common effort together with Roland Arnold.

Signal for Sec secretion pathway - SignalP: The Sec-secretion pathway transfers proteins into the periplasm of gram-negative bacteria and is generally not considered as a mechanism important for bacterial virulence [24]. In the Effective portal, it is mainly used as a negative control in combination with the T3SS prediction approach. It facilitates the recognition of unspecific effects in the prediction by EffectiveT3. The secretion mechanisms of the Sec-pathway and T3SS serve different purposes in the bacterial cell. The Sec-pathway secrets into the periplasm, while T3SS targets explicitly the eukaryotic host cytoplasm. Furthermore, both signals are located in the N-terminal part of the proteins amino acid sequence and should be exclusive. Proteins secreted by both pathways are not observed in any bacteria and regarded as highly unlikely. The signal for Sec secretion is modeled with high accuracy and therefore serves as a fast and simple method to enhance prediction accuracy of T3SS predicted effector proteins and can be also used to restrict selection of effector candidates potentially secreted by other membrane spanning secretion systems like T4SS and T6SS. As our understanding of the secretion mechanisms is still limited, the Sec-pathway signal is not used as a strict filtering criteria but offered to the user as additional information to rate the confidence of individual predictions.

The most established software for prediction of Sec-pathway proteins is SignalP, provided by Nielsen et al [225]. The SignalP stand-alone software package (version 4.1) was downloaded from the SignalP webpage and integrated via shell scripting into the Effective prediction interface.

Function based approach - Identification of eukaryotic-like protein domains: Identification of eukaryotic-like domains (ELDs) is based on all PFAM-A protein domain hidden markov models (HMM) provided by Pfam [230]. To search for eukaryotic-like domain signatures in user provided protein sequences, the website uses the program Hmmsearch, as implemented in the Hmmer 3.1b1 software package [98].

For on-the-fly identification of ELDs in user provided protein sequences, no individual scores are calculated. The output indicates each recognized eukaryotic-like domain signature that achieves a significant enrichment score in at least one precalculated pathogenic/ symbiotic genome. Calculation of an individual enrichment score requires a complete genome with all protein sequences given, which is not necessarily the case for user provided protein sets. Eukaryotic-like domains with a significant ELD score in at least one precalculated genome provide a sensitive result for further manual inspection by the user. The prediction of protein secretion by Effective does not require any consensus between the complementary prediction approaches but considers any single positive prediction to be of biological relevance.

3.3.3 Automatic updates of the Effective database

To keep the Effective web portal up to date, a semi automatic update process is implemented.

Updates of the genome repository and all precalculated prediction results in the Effective database can be executed with little manual intervention. Update scripts in Unix shell and python programming language implement data retrieval and processing. This facilitates the performance of regular, e.g. yearly update cycles. Furthermore, the Effective web portal can anticipate sudden changes in the processed information and can quickly respond to major updates in the public resources it depends upon.

3.3.4 Implementation of basic technical aspects

Several technical aspects are addressed in the implementation of the portal:

- **Usability/Userfriendliness** The portal is accessible from any operating systems and is tested with all common web browsers. Overall design of the web portal as well as presentation of content is optimized for fast visual processing by the user. A large help section provides the user with all necessary information to make use of the portal.
- **Easy extensibility** The resource is designed in a flexible way to allow easy maintainability and extensibility regarding data types and methods. The portal architecture especially facilitates integration of new prediction methods for secreted proteins that might become available in short future, e.g. for the identification of signal sequences in type-IV secreted proteins.
- **Versioning** Chronological tracking of outdated versions is implemented. Information about the current version is given in the “News” section of the portal. Old versions are accessible in an “Archive” section. Updates of the Effective database are also announced to users via the “News” box and the Effective mailing list.

3.3.5 The Effective web portal provides precalculated effector candidates for all sequenced pathogens and symbionts

The Effective web portal provides user-friendly tools for browsing and retrieving comprehensive precalculated predictions for completely sequenced bacterial genomes. The general outline of steps involved in the use case regarding a basic analysis of a predicted effectome is visualized in 3.4. Eukaryotic-like domains and signal peptides of the Sec pathway can be retrieved from the Effective database for each of the 1160 genomes of pathogenic and symbiotic bacteria. In order to provide consistent scores across genomes, domain enrichment scores based on the total number of annotated proteins are always provided.

The applicability of the prediction approaches implemented in Effective is different: whereas Sec dependent secretion is a widespread feature of pathogens and symbionts and eukaryotic-like domains can be encoded in any of their genomes, type-III secreted proteins can only be expected in genomes of Gram-negative bacteria encoding a type-III secretion system that is likely to be functional. Therefore, precalculations by the different methods are made accessible for the appropriate genome subsets. For all 188 genomes of Gram-negative bacteria that encode a probably functional type III secretion system, the results of EffectiveT3 predicting type-III effectors are included in the precalculated files available for download. These results contain the accessions and descriptions of all annotated proteins in the respective genomes, the EffectiveT3 scores and the secretion prediction for each protein according to the selective default threshold and standard prediction model. By selecting the organism of interest, effector candidates identified in any pathogen/symbiont by all applied prediction methods can be comfortably downloaded as tab-delimited files. To facilitate further exploration and biological interpretation of eukaryotic-like protein domains, additional features are provided regarding the eukaryotic-like domain based prediction approach. Precalculated results are provided online through interactive web pages that link any predicted domain to a status report page. Thereby, the user has the possibility to

- **browse eukaryotic-like domains per sequenced pathogen/symbiont:** Each completely sequenced genome can be selected from a drop down list to access its genome report page, also compare 3.1. On this page, all eukaryotic-like domains in proteins of the specific genome are listed with the according enrichment scores and additional information. If a pathogenic/symbiotic genome is contained in the eggNOG database of Clusters of Orthologous Groups, additional domain enrichment scores according to the conserved proteome are given.

- **browse comprehensive information on eukaryotic like domains:**
  Specific domain report pages are provided for each protein domain that has been detected in at least one pathogenic genome with a significant domain enrichment score of 4 or higher. For any domain, the numbers and lists of pathogenic, non-pathogenic and eukaryotic genomes encoding at least one protein containing this domain are indicated. As the frequencies of the domain in these organisms determine the domain enrichment score, this information also allows the user to understand why Effective has identified the particular domain as eukaryotic-like. A link to descriptions of the particular domain in the Pfam domain database offers additional information to the user.

- **browse sequence information on effector candidates:**
  All effector candidates containing a particular eukaryotic-like domain with a significant enrichment score are provided on protein report pages. Besides protein descriptions and accession numbers, the protein report pages provide links to the according protein entry in the SIMAP database of protein similarities, providing information on e.g. sequence similarity and Gene Ontology functional annotation.

3.3.6 Use case: Selection of effector candidates for pulldown analysis to investigate virulence strategies of *Legionella pneumophila*

The human/animal pathogen *Legionella pneumophila* uses functional processes established by the host cell to maintain infection [50]. It serves as a model organism for the identification of eukaryotic-like protein domains (ELDs) and is exemplarily analyzed in this use case. A diverse set of Type IV secreted effector proteins are suggested to orchestrate the complex pathogen-host interactions of *Legionella* [103]. To closer investigate the molecular interactions of *Legionella pneumophila* with its host, cost and labor intensive experimental analyses are performed [302]. E.g. in pulldown assays identify protein-protein interactions by probing an interaction between a protein of interest that is expressed as a fusion protein (bait) and the potential interacting partners (prey) [1]. These experiments reveal possible functions which could be displayed by these proteins once they reach the host cell. Reliable pre-selection of high-quality effector candidates is crucial.

In the following, the possible workflow of an effector candidate selection via the Effective web portal is illustrated. We retrieve comprehensive information about the predicted

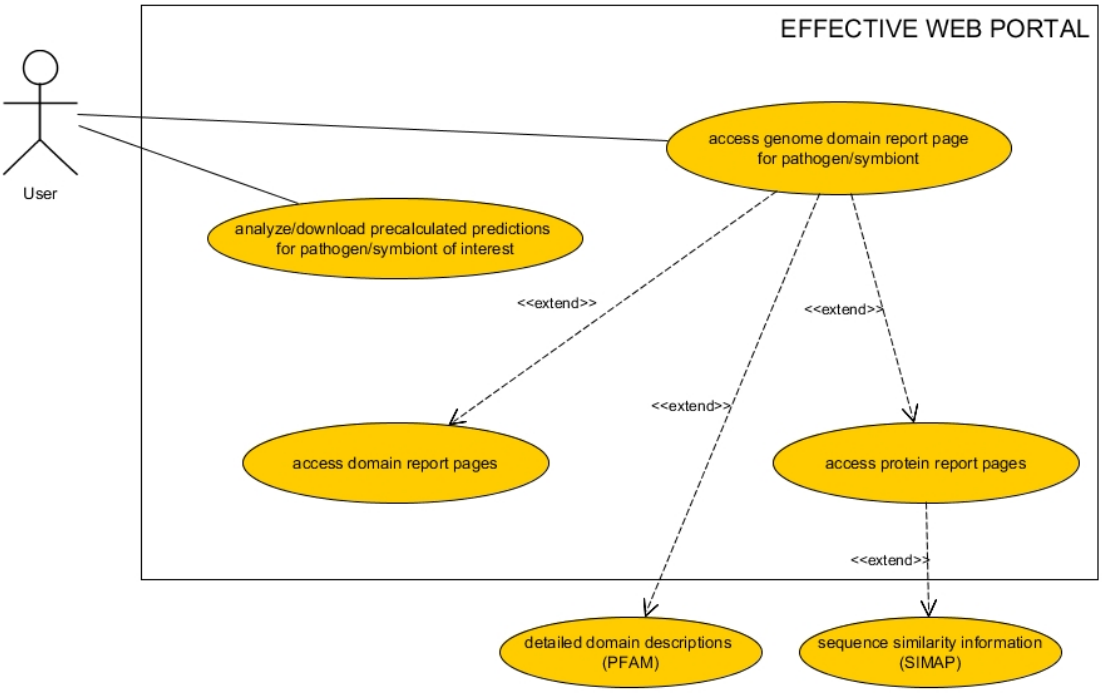

EFFECTIVE WEB PORTAL
User
access genome domain report page for pathogen/symbiont
analyze/download precalculated predictions for pathogen/symbiont of interest
<<extend>>
<<extend>>
<<extend>>
access domain report pages
access protein report pages
<<extend>>
detailed domain descriptions (PFAM)
sequence similarity information (SIMAP)

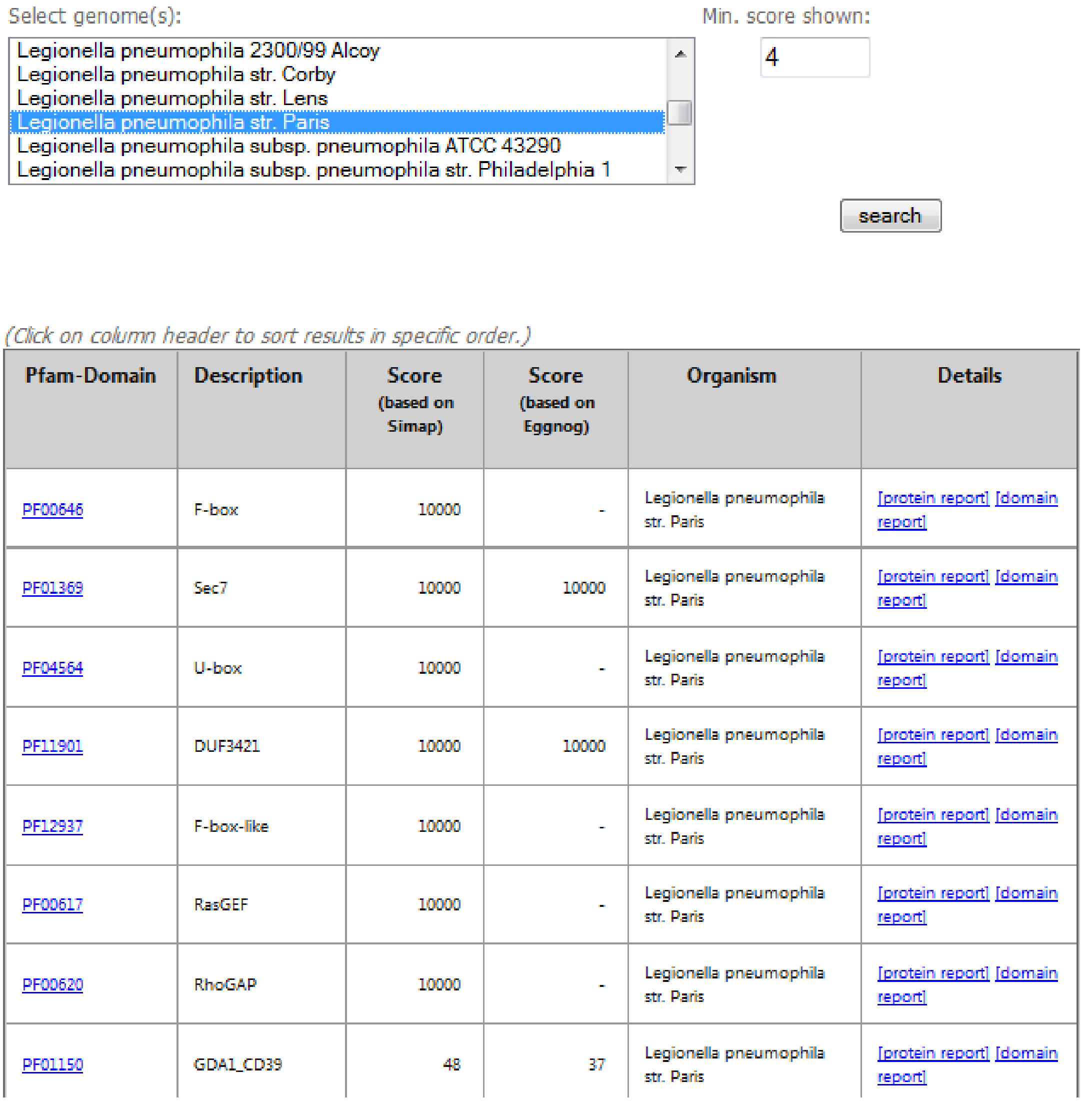
Figure 3.4: ***UML description for retrieving precalculated information for all sequenced pathogens and symbionts. .***

effectome of the completely sequenced pathogen *Legionella pneumophila str. Paris*, resulting in a clearly laid out list of probable effector candidates.

On the Effective start page, under the menu entry "Precalculated", for each organism a link leads to a tab-separated file containing comprehensive effector prediction results for all annotated proteins of *Legionella pneumophila str. Paris*. From that file, candidate effectors based on all applied effector prediction methods can be selected for further experimental analysis. For *Legionella*, effectors are predicted by the eukaryotic-like domain approach and the signal for the Sec-pathway. EffectiveT3 was not applied as the *Legionella* genome does not encode a functional Type 3 secretion system. For investigation of the effector candidates based on the identification of eukaryotic-like domains, the "Browse" section of the portal is used which allows to comfortably browse detailed information on ELDs and ELD containing proteins.

48 protein domains in *Legionella pneumophila str. Paris* were identified as eukaryotic-like domains (ELDs) in 108 protein sequences. Seven ELDs in 8 proteins achieved the maximum ELD score of 10 000 3.5. The score indicates that these ELDs are exclusively found

in eukaryots and pathogenic/symbiotic bacteria. Starting from these highly probable hits, domain report pages are questioned for additional information, e.g. for the F-Box domain shown in figure 3.6. Experimental candidates are chosen on the basis of ELD scores. Manual inspection of literature for the final set of selected proteins completes the candidate selection workflow. The literature search reveals that all of the high-scoring candidates were previously identified in *Legionella*-specific studies as eukaryotic proteins [118, 314]. The majority of predicted candidates are well characterized *Legionella* effector proteins, showing again the high value of the unspecific eukaryotic-like domain approach for the prediction of effector proteins. The proposed functional contexts of these characterized effectors within the host cell are listed below: The predicted *Legionella* protein lpp2082 that contains domain PF12937 (F-box-like) is the characterized effector AnkB, shown to modulate ubiquitination of the host protein parvin B [175]. It functions as a general platform for the docking of polyubiquitinated proteins to the Legionella-containing vacuole (LCV) to enable intravacuolar proliferation [228]. The high-scoring eukaryotic-like U-box domain (PF04564) is identified in the *Legionella* effector LubX (lpg2830) [199]. The LegG2 effector (lpg0276) contains a RasGEF domain signature (PF00617) that allows the pathogen to interfere with GDP-to-GTP exchange in GTP-binding host Ras proteins [254]. Lpp1761 is a RhoGAP domain containing homolog to MavU, substrate of the Type IV secretion system and involved in GTP-protein interactions inside the host cell [14]. A high-scoring domain of unknown function (DUF3421, PF11901) is conserved in proteins of all *Legionella* and *Coxiella* strains. Additionally candidates can be selected based on the ELD score ranking. E.g. manual inspection of the two homologs containing the high scoring domain signature of PF01150 of the GDA1/CD39 (nucleoside phosphatase) family (ELD score of 48) were shown to be critical for *Legionella* infection and virulence [155]. All proteins with ELDs of maximum score make ideal candidates for experimental analysis and are probably secreted by the Type IV secretion system [137]. Contrary to former *Legionella*-specific studies, the presented candidate selection was conducted without the integration of prior knowledge about *Legionella pneumophila* and calculated based on the general criteria described in chapter 2. The presented use case regarding the identification of ELDs *Legionella* also shows that the eukaryotic-like domain method can reproduce earlier findings [309].



Select genome(s):

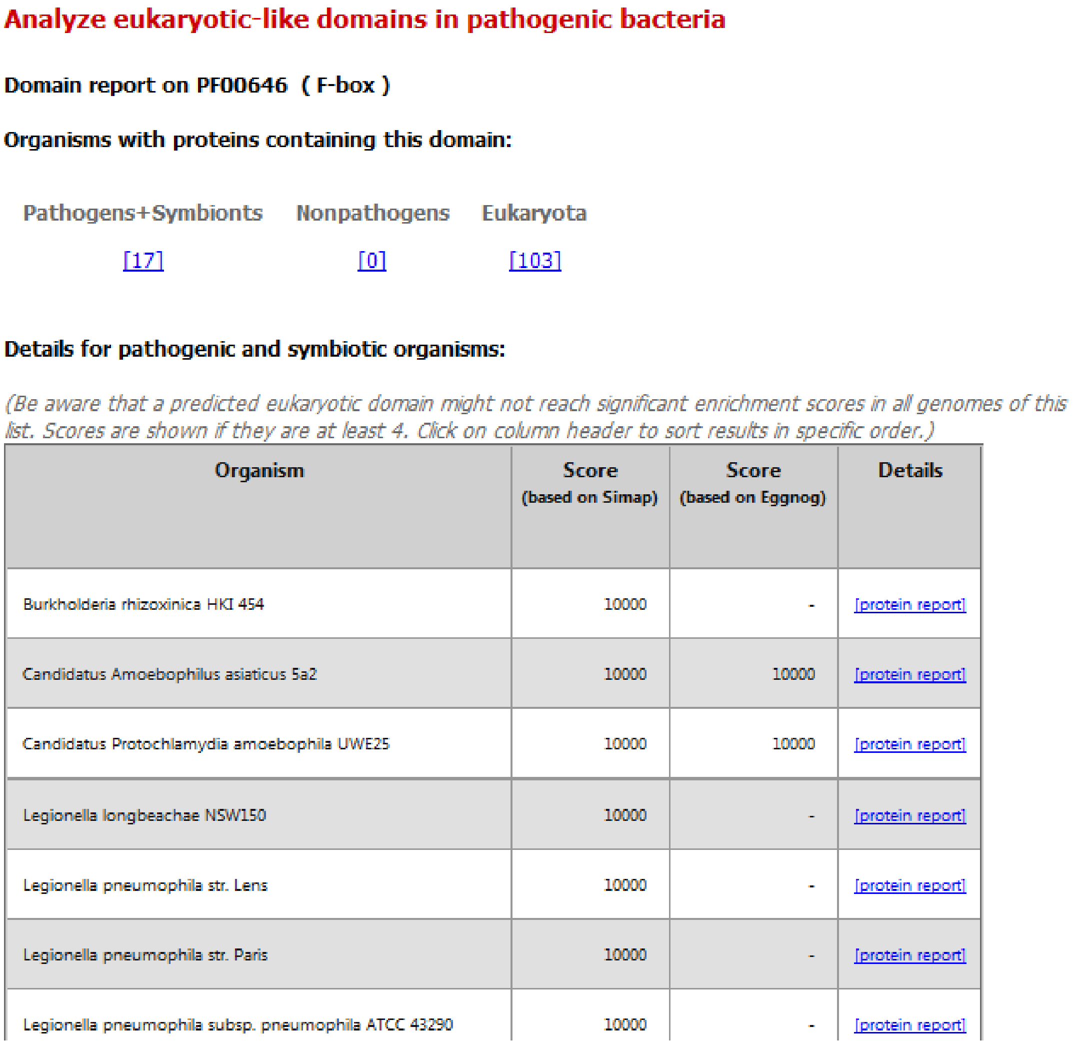
- Legionella pneumophila 2300/99 Alcoy
- Legionella pneumophila str. Corby
- Legionella pneumophila str. Lens
- Legionella pneumophila str. Paris
- Legionella pneumophila subsp. pneumophila ATCC 43290
- Legionella pneumophila subsp. pneumophila str. Philadelphia 1

Min. score shown: 4

(Click on column header to sort results in specific order.)

Pfam-Domain	Description	Score (based on Simap)	Score (based on Eggnog)	Organism	Details
PF00646	F-box	10000	-	Legionella pneumophila str. Paris	[protein report] [domain report]
PF01369	Sec7	10000	10000	Legionella pneumophila str. Paris	[protein report] [domain report]
PF04564	U-box	10000	-	Legionella pneumophila str. Paris	[protein report] [domain report]
PF11901	DUF3421	10000	10000	Legionella pneumophila str. Paris	[protein report] [domain report]
PF12937	F-box-like	10000	-	Legionella pneumophila str. Paris	[protein report] [domain report]
PF00617	RasGEF	10000	-	Legionella pneumophila str. Paris	[protein report] [domain report]
PF00620	RhoGAP	10000	-	Legionella pneumophila str. Paris	[protein report] [domain report]
PF01150	GDA1_CD39	48	37	Legionella pneumophila str. Paris	[protein report] [domain report]

Figure 3.5: ***Screenshot of the Effective domain list output, presenting the list of high-scoring ELDs in Legionella pneumophila str. Paris.*** *For each genome, specific pages list all eukaryotic-like domains with according enrichment scores and links for further analysis. Here, the ranked list of high-scoring ELDs identified in Legionella pneumophila str. Paris are shown that were analyzed in detail.*



Analyze eukaryotic-like domains in pathogenic bacteria

Domain report on PF00646 (F-box)

Organisms with proteins containing this domain:

Pathogens+Symbionts	Nonpathogens	Eukaryota
[17]	[0]	[103]

Details for pathogenic and symbiotic organisms:

(Be aware that a predicted eukaryotic domain might not reach significant enrichment scores in all genomes of this list. Scores are shown if they are at least 4. Click on column header to sort results in specific order.)

Organism	Score (based on Simap)	Score (based on Eggnog)	Details
Burkholderia rhizoxinica HKI 454	10000	-	[protein report]
Candidatus Amoebophilus asiaticus 5a2	10000	10000	[protein report]
Candidatus Protochlamydia amoebophila UWE25	10000	10000	[protein report]
Legionella longbeachae NSW150	10000	-	[protein report]
Legionella pneumophila str. Lens	10000	-	[protein report]
Legionella pneumophila str. Paris	10000	-	[protein report]
Legionella pneumophila subsp. pneumophila ATCC 43290	10000	-	[protein report]

Figure 3.6: ***Screenshot of the Effective report page for F-Box domain (PF00646).*** *For each protein domain found to be enriched in any bacterial pathogen/symbiont, an individual domain report page provides detailed information, e.g. about the taxonomic distribution in all organisms.*

3.3.7 The Effective web portal allows on-the-fly prediction of effector proteins in user provided sequence data

The Effective web portal provides a user-friendly interface for on-the-fly prediction of effectors in user-provided protein sequences. Sets of sequences from single proteins to the size of complete genomes can be uploaded in multi-FASTA format.

For the analysis of input sequence data, the user can choose from any combination of the three different prediction methods: prediction of Sec pathway secreted proteins, prediction of TTSS secreted proteins and identification of eukaryotic-like protein domains (ELDs). Typical steps of an interactive effector prediction are visualized in 3.7.

The user can provide his own protein sequences in multi-sequence FASTA format. The interface allows for a prediction of both function- and signal-based effector prediction methods with adjustable parameter settings. Three different methods can be chosen in any combination: Eukaryotic-like domains, EffectiveT3 and SignalP. During the on-the-fly identification, the Effective method assigns a eukaryotic-like domain enrichment (ELD) score to each protein domain, reflecting the maximal enrichment of that domain in any pathogen or symbiot, compared to the background frequency of the protein domain in non-pathogenic bacteria. A high ELD score equals strong enrichment of the protein domain in pathogenic/symbiotic bacteria. Parameter settings for all prediction methods can be adjusted to individual needs. The size of the sequence dataset can be up to the typical size that can be found in complete bacterial genomes. Comparable to the precalculated data, this tool allows the user to control all prediction settings and to integrate the results from the different methods into one sorted table. The configurable parameters of these methods are explained in a detailed "Help" section on the website. Calculation time for predicting effectors in a bacterial genome of about 1000 protein sequences by all three methods is about 4 minutes. In this typical example, time is distributed on the different methods as follows: EffectiveT3 (20 sec), SignalP (40 sec) and Eukaryotic-like domains (180 sec). For user-provided data, sequences are scanned for Pfam domain signatures using HmmSearch [98]. Detected domains are evaluated based on the precalculated domain enrichment scores in the Effective database, considering the maximum score that is achieved by the particular domain in all pathogen genomes. Proteins that have received positive predictions from at least one selected method are provided in tabular form for further visual inspection on the website and for download in Excel format (figure 3.8).

A standard use case for this functionality of the Effective portal is the prediction of effector candidates in user provided protein sequence data of a novel sequenced pathogen or symbiont.

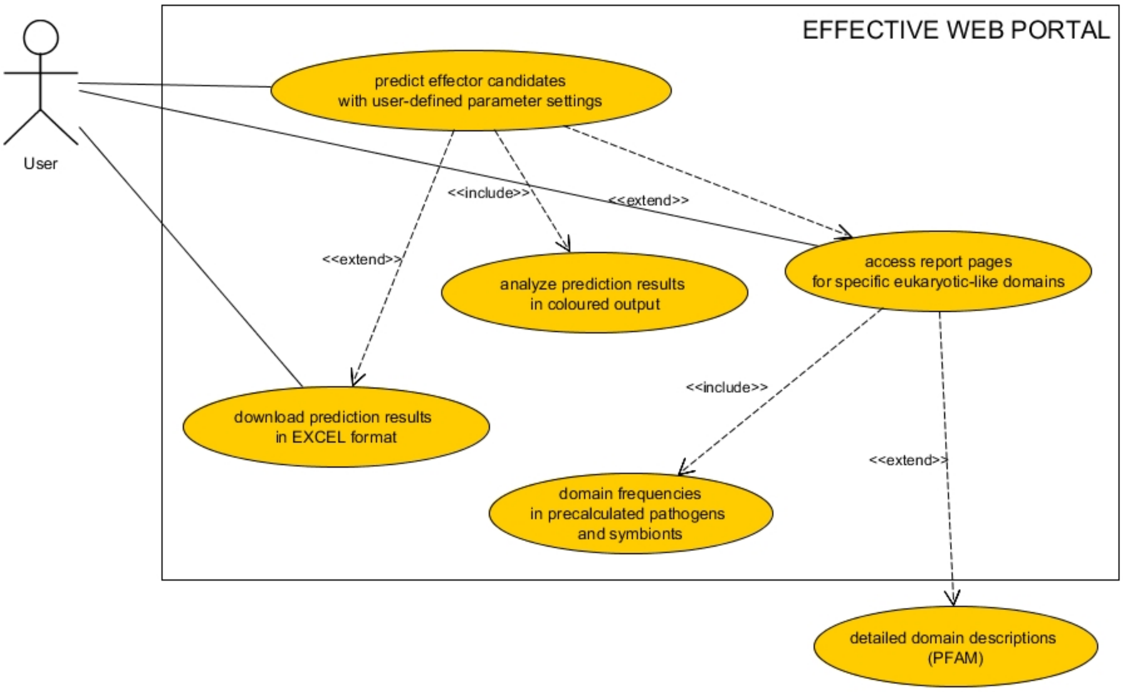
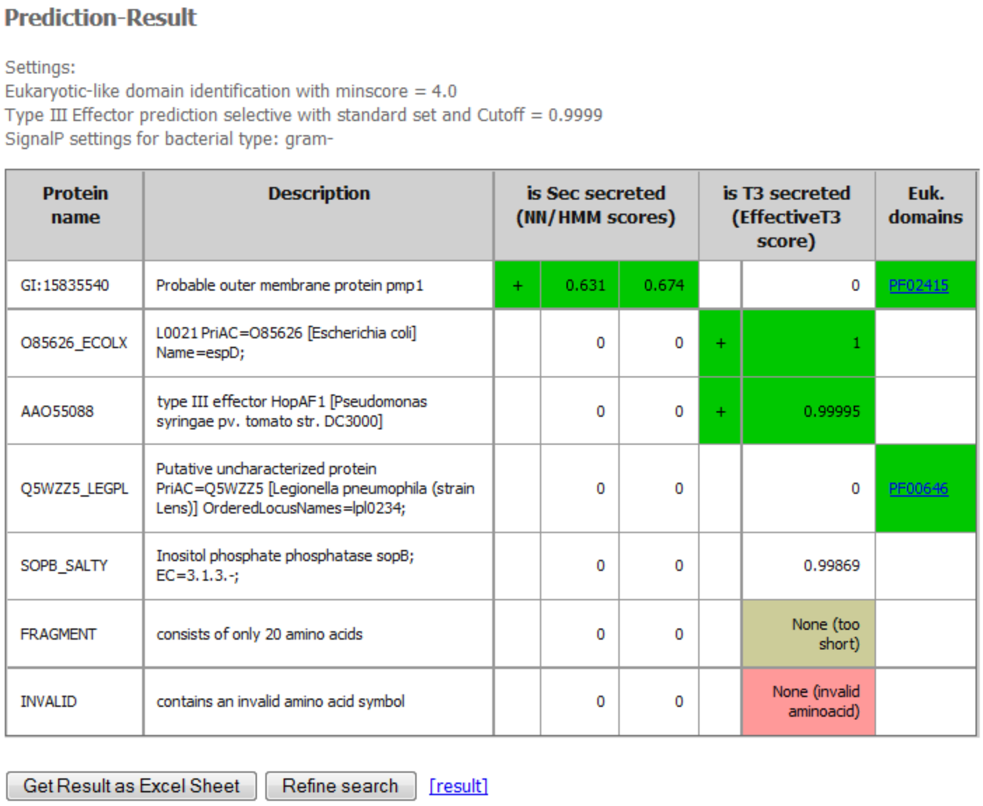

EFFECTIVE WEB PORTAL
User
predict effector candidates with user-defined parameter settings
<<include>>
<<extend>>
<<extend>>
analyze prediction results in coloured output
access report pages for specific eukaryotic-like domains
download prediction results in EXCEL format
<<include>>
<<extend>>
domain frequencies in precalculated pathogens and symbionts
detailed domain descriptions (PFAM)

Figure 3.7: ***UML description for the prediction of effector candidates in user provided data. .***

3.3.8 Use case: Comparative genomics analysis of three *Pantoea ananatis* strains causing diverse host phenotypes

In a recent study, the prediction interface of the Effective web portal was used in a comprehensive investigation of genomic features in the plant pathogen Pantoea ananatis (P. ananantis). The methodology of the analysis constitutes a representative use case of this part of the Effective web-portal and is described in the following.

P. ananatis is an emerging plant pathogen and can be found all over the world in different agricultural plant species, such as rice, onion, maize and many more. Depending on the host, it can cause different diseases and symptoms [115]. Certain strains are not only reported to be non-pathogenic but even plant growth promoting [222]. This makes P. ananatis an interesting candidate to investigate, especially as a means of biological control [65]. Genome analysis of a plant growth promoting strain of P. ananatis indicate that these strains lack traits related to pathogenicity whereas they harbor genes that are involved in plant growth stimulation [151]. Molecular interaction with the host plant is



Prediction-Result

Settings:
Eukaryotic-like domain identification with minscore = 4.0
Type III Effector prediction selective with standard set and Cutoff = 0.9999
SignalP settings for bacterial type: gram-

Protein name	Description	is Sec secreted (NN/HMM scores)			is T3 secreted (EffectiveT3 score)		Euk. domains
GI:15835540	Probable outer membrane protein pmp1	+	0.631	0.674		0	PF02415
O85626_ECOLX	L0021 PriAC=O85626 [Escherichia coli] Name=espD;		0	0	+	1	
AAO55088	type III effector HopAF1 [Pseudomonas syringae pv. tomato str. DC3000]		0	0	+	0.99995	
Q5WZZ5_LEGPL	Putative uncharacterized protein PriAC=Q5WZZ5 [Legionella pneumophila (strain Lens)] OrderedLocusNames=lpl0234;		0	0		0	PF00646
SOPB_SALTY	Inositol phosphate phosphatase sopB; EC=3.1.3.-;		0	0		0.99869	
FRAGMENT	consists of only 20 amino acids		0	0		None (too short)	
INVALID	contains an invalid amino acid symbol		0	0		None (invalid aminoacid)	

Figure 3.8: ***Screenshot of the Effective result page for an interactive exemplary effector prediction.*** *A descriptive visual output allows the user to easily capture important results about the input data. Links provide additional information, e.g. on enriched eukaryotic-like protein domains. All results can be comfortably downloaded in excel-format.*

established via a Type 6 Secretion System (T6SS) that enables Pantoea to secrete proteins into the host cytoplasm [260].

In the selected study, Sheibani et al. compared genotypic differences in three novel strains of P. ananatis regarding the impact on the host plant. Before the *in silico* analysis, these strains were shown to cause different phenotypes in the infected host: Strain X1 led to death of the host plant, being pathogenic; X2 showed no recognizable effect, while X3 stimulated plant growth and therefore is considered beneficial to the plant. Sheibani et al. identified eukaryotic-like protein domains (ELDs) in all three P. ananatis strains (X1, X2 and X3) by applying the prediction framework provided by the Effective web-portal. For each genome, on-the-fly predictions were caculated using the web-interface with adjusted parameter settings. As there is no functional Type 3 Secretion System (TTSS) detected in any of the P. ananatis strains, the option to predict TTSS effectors was disabled. All eukaryotic-like domains with a highly significant ELD score greater or equal to 10 were selected for output. Results were downloaded and compared regarding eukaryotic-like protein domains and ELD containing candidates to investigate the genomic variance of P. annatis strains that cause different phenotypes in the host. For all candidate ELDs, frequencies in other bacterial pathogens/symbionts as well as occurrences in the eukaryotic host genomes were inspected using the domain specific report pages of the Effective portal.

A high overall overlap of predicted ELDs in all 3 three Pantoea strains supports the assumption of high average functional similarity of effector proteins. Despite this vast degree of similarity, a close comparison of eukaryotic-like domains for P. ananatis strains reveals crucial differences. A literature search for the selected candidate ELDs shows them to be highly relevant regarding the pathogenic context of several bacterial pathogens/symbionts or play major roles in the defense mechanisms of plant. The eukaryotic-like domain content of all genomes reveals a varying molecular repertoire of the P. ananatis strains that are suggested to be responsible for the observed differences in phenotypes of infected host plant *(manuscript in preparation)*. Experimental evaluation of all predicted candidates is the next step in an analysis that could lead to profound insights into host-interactions of P.ananatis.

3.4 Discussion and outlook

The Effective database is the first bioinformatics resource that combines two complementary approaches for the prediction of bacterial secreted proteins:

the function-based prediction by identification of eukaryotic-like domains and the prediction based on signal peptides necessary for transport by protein secretion systems. By principle, none of these two strategies can achieve complete coverage. Therefore, their integration in a single resource is beneficial for the comprehensive annotation of putative effectors in genomes and proteomes.

Further advancements of the Effective portal could further enhance user friendliness and automation. Aspects that are predestined to be addressed in the upcoming versions of the portal are improvements regarding the on-the-fly prediction user interface and workflow: In the case of effector prediction in user provided sequence data, currently the eukaryotic-like domain based prediction approach provides ELD scores regarding on all precalculated pathogenic/symbiotic genomes, compare 3.3.2. An enhancement would include an option that indicates that a user provided set of protein sequences comprehends all sequences from a complete bacterial genome. By selecting a checkbox, the user can indicate completeness of the genome and thereby activate individual ELD enrichment score calculation for the provided set of protein sequences. Another advancement in the methodology to analyze user defined complete input genomes could become feasible in the near future. In the current implementation, the Effective portal does not provide any help in the decision which prediction methods are to be applied correctly. E.g. in the case of *Legionella*, the pathogen is well known not to encode a Type III secretion system. To prevent misleading results, effector candidates are therefore are not predicted by the EffectiveT3 method, in the precalculated genomes of all *Legionella* strains. If an inexperienced user checks the button to include the prediction of T3 effector candidates in a user provided set of *Legionella* proteins, results are likely to include positive hits. These positive predictions are expected and in accordance to former findings by Arnold et al. The signal peptide can well occur as it does not damage other protein functions. As there is no evolutionary pressure on the N-termini of Legionella proteins, a secretion signal could be formed by chance [20]. Functional secretion systems are recognized in many well characterized pathogens, as for example in *Legionella*. For many other bacteria of interest, this information is not experimentally verified. Limiting factors are typically the confined knowledge about details of the secretion systems, especially regarding completeness and functionality. E.g., even for the well studied T3SS, the experimental data on the T3SS apparatus is not sufficient enough to enable prediction of its functionality. In consequence, this would need an experimental functional validation for any given organism at hand, a claim that is unworkable in practice. Identification of TxSS components in any pathogen of interest is feasible as demonstrated for the components of the Type III secretion machinery in section 2.3.3. For several systems, the specific mechanisms of transport are still unknown. Identification of all components necessary for successful transport of effector proteins depends on experimental validation. It needs to be evaluated, if empirical criteria can lead to a reliable recognition of functional secretion systems. The integration of such secretion system recognition in user provided sequence data could accommodate the user with a completely automatic preselection of applicable effector prediction methods. The user-friendly web portal of the Effective database offers a versatile toolbox for generating new effector candidates and for target selection toward experimental investigation of putative secreted proteins. As the development and improvement of computational methods for effector prediction is a vital area of research, new methods can be expected to become available within the next years. The Effective database provides a powerful framework for their easy integration and will therefore make relevant new methods accessible to the users of the database.

4 Application of a domain-based approach to predict pathogen-host interactomes

4.1 Motivation

Upon infection, bacterial pathogens secrete effector proteins into the host cytosol to alter functional pathways of the host cell. Comprehensive prediction of a pathogens effectome is a crucial first step towards the understanding of bacterial virulence. For an understanding of the direct pathogen-host interplay, investigation of the molecular interactions taking place between bacterial effectors and targeted host proteins is required. In the EraNet Pathomics project that funded the current study, collaboration partners experimentally investigated pathogen-host protein interactions (HP-PPIs) for proteins of three pathogens: *Pseudomonas aeruginosa*, *Chlamydia pneumoniae* and *Chlamydia trachomatis*. Simultaneously, I investigated computational methods to predict PH-PPI networks. Main aim of this approach was to support the selection of specific candidates for further experimental studies of collaboration partners, e.g. in interaction proteomics. *In silico* prediction of pathogen-host PPIs faces the challenge of sparse experimental data. The experimental data on inter-species protein-protein interactions in general and bacterial pathogen-host/ host-pathogen interactions in particular is very limited. Contrary to intra-species protein complexes, pathogen host protein interactions cannot be directly measured under laboratory conditions in a cultivated cell culture and are much more difficult to investigate experimentally [259]. For example, one of the few characterized interactions, the interaction between chlamydial inclusion membrane protein IncG and host protein 14-3-3β was experimentally discovered by a yeast two-hybrid screen of IncG against human proteins and subsequent confirmation by fluorescence microscopy in infected HeLa cells [255].
The major public resources on experimental PPIs hold only sparse information on a few bacterial pathogens with a strong bias towards viral organisms. Regarding the pathogens

of interest, Phidias [313], PHI-base [308] and the Pathogen Interaction Gateway (PIG) [84] list no experimentally confirmed pathogen-host interactions for *Chlamydia pneumonia* and *Pseudomonas aeruginosa*, and one experimentally verified PH-PPI for *Chlamydia trachomatis*.

Given the sparse experimental data on pathogen-host interactions, cost and time efficient additional methods that could lead towards a better understanding of the pathogen-host interactome are crucial. Up to now, no organism specific PPI prediction approach for the pathogens of interest exists or was implemented. Taxon independent PH-PPI prediction methods are not yet established and difficult to evaluate due to the lack of data on pathogen-host PPIs supported by experimental evidence. Large scale bioinformatics approaches typically rely on training sets containing a sufficiently large number of experimentally validated candidates from diverse organisms to evaluate prediction results. Such comprehensive data is not available in the case of bacterial pathogen-host PPIs. *In silico* approaches to predict pathogen-host interactions typically assess the plausibility of predicted PPIs by different methods. Evaluation of the approach by a comprehensive comparison of predicted PH-PPIs to known interactions is not feasible in most cases, due to the lack of experimental data. The plausibility of the predictions is assessed by analysis of gene expression data and enrichment of general functional properties found to be relevant for pathogen-host interactions [70].

The potential as well as current limitations in the prediction of pathogen-host interactomes are evaluated exemplarily for the obligate intra-cellular pathogen *Chlamyida trachomatis*. The Domain Interaction MAp (DIMA) is chosen as a representative collection of predicted protein domain-domain interactions (DDIs) derived from molecular intra-species interactions. Mapped onto the chlamydial effectome and human proteome, the predicted DDIs are expected to reveal possible interactions between effectors and their targets in the human PPI network. Generally, the number of proteins annotated in the human genome that share the same domain signatures can be very high [170]. The domain based approach can therefore be expected to provide very general results with a large number of interaction partner candidates. Human proteome data on tissue specific protein expression is integrated as additional support and discussed as a possible means to confine interaction candidates. Biological plausibility of all predictions is tested considering the over-representation of functions related to pathogenicity. Host pathways/processes and molecular functions that are primarily affected by the chlamydial effectome are identified by Gene Ontology (GO) term and pathway enrichment analysis. Host proteins that are linked to known strategies applied by *C. trachomatis* to initiate and maintain infection are expected to be over-represented. Novel functions that also might be enriched could

reveal host processes yet unknown to be altered during pathogenic infection.
In this work, the domain based prediction of protein interactions between the secreted effectome of a pathogen and the protein network of the host cell is investigated. It reveals current limitations and chances to the *in silico* prediction of pathogen-host protein interactions. All predicted PH-PPI networks between human and *Chlamydia pneumonia*, *Pseudomonas aeruginosa* as well as *Chlamydia trachomatis* were made available online to provide information on the biological context of experimental protein interaction candidates to all collaboration partners of the EraNet Pathomics project.

4.2 Domain-domain interaction candidates of the *C. trachomatis* effectome

4.2.1 Introduction

Experimental data on protein complexes and functional intra-species protein interactions are basis to several methods that predict domain-domain interactions (DDIs) from these protein interactions. The Domain Interaction MAp (DIMA) combines different methods to predict interactions between protein domains from known and predicted protein protein interactions. The potential of domain interactions identified in intra-species data to predict the protein interaction network between pathogenic effectors and host proteins is investigated. The approach explored in this chapter assumes that characteristics of inter-species interactions can be deduced from intra-species interactions because the underlying molecular chemistry is not changed by the host-pathogen context. With this assumption being valid, DIMA should provide domain interactions for bacterial effectors that link these proteins to host proteins containing the interacting target domain. The domain-based prediction approach is not expected to offer a comprehensive characterization of effector interaction candidates. Major constraints could lie in the intrinsic dependence on recognized protein domain signatures. The obligate intra-cellular pathogen *Chlamyida trachomatis* is chosen as a model organism for its relevance to the EraNet Pathomics project. Furthermore, *C. trachomatis* is a representative of the chlamydial phylum, which has evolved separate from other bacterial divisions invariably containing pathogenic organims. It can be considered a model for the study of bacterial infection strategies and pathogenicity [144]. The predicted DDIs are expected to reflect features of protein interaction networks independent of organism boundaries. This could offer a novel approach to the prediction of pathogen-host protein interactions (PH-PPIs). Furthermore, it was analyzed to which extent targeted interacting domains are enriched for the set of chlamydial effectors. Over-represented domains related to key functions targeted during pathogenic infection serve as an additional indicator that prediction results are biologically plausible.

4.2.2 Material and methods

The candidate effectome of *Chlamydia trachomatis* was identified by collecting data from experimental studies as well as effector prediction approaches. Considered are experimentally validated effectors from manual literature search and effector proteins predicted by EffectiveT3 and the eukaryotic-like domain (ELD) based approach. Furthermore, the results of several T3 secretion assays report a high number of *C. trachomatis* proteins that are shown to be transported by a heterologous Type III secretion system [278, 73, 11]. These proteins can be expected to be secreted and to interact with the host cytosol during infection. They are included in the predicted effectome.

In a post-processing step, candidates were screened for chlamydia specific virulence factors with known non-host specific functions [218, 287]. Eight effector candidates containing the domain signatures for Chlamydia polymorphic membrane protein repeat (PF02415) and two TLC ATP/ADP transporter (PF03219) were removed. Pfam domain signatures for chlamydial protein sequences are retrieved from Simap [235]. All domain-domain interaction (DDI) pairs with calibrated DIMA score >= 0.7 were retrieved from the DIMA 3.0 database [178]. Thereby, the majority of DDIs are supported by structural evidence derived from the public databases 3did [274] and ipfam [99] and can be considered of very high quality. All DDIs are mapped onto the chlamydial effectome. Furthermore, the specificity of predicted pathogen-host domain interactions (PH-DDIs) of the chlamydial effectors is investigated. The enrichment of domain interactions for the chlamydial effectome was analyzed in comparison to interactions of the complete proteome of *C. trachomatis*. P-values for all predicted HP-DDIs are obtained by comparing the number of effectors targeting a particular domain with the distribution of that domain sampled from random shuffling. In the shuffling process, PH-DDIs are subsequently exchanged for randomly drawn interactions from the set of all chlamydial DDIs. The number of proteins with DDIs and the number of diverse DDI targets for each of these proteins are kept constant, preserving characteristics of the original set of predicted PH-DDIs. Permutations of this process (n = 1000) simulate a normal distribution of the numbers of effectors targeting each domain that is hypothetically interacting with proteins from the pathogenic effectome. P-values are calculated for each domain of the host-sided interactome based on mean and standard deviation of the given normal distribution and the number of effectors targeting the particular domain using the R Software for Statistical Computing [232].

4.2.3 Results

The predicted effectome of *C. trachomatis* consists of 156 proteins and has been determined by combining experimentally verified as well as predicted effectors, compare table 8.3. A TTSS signal peptide has been predicted in 102 candidates, while 10 proteins contain eukaryotic-like protein domains. With 33 of 89 proteins, are large fraction of experimental candidates is part of the family of chlamydial inclusion membrane (Inc) proteins. The Inc proteins are included into the analysis, as several proteins were found to be secreted from the inclusion and directly interact with host proteins [246, 271, 255, 76].
Domain signatures were detected in 81 effector proteins (91 different domains in total). For 53 of these effectors, domain domain interactions were predicted, targeting 207 different domains. While the chlamydial effectors have an average of four DDIs and a median of two predicted interacting domains, three predicted effectors have a very high number of predicted domain interactions: CT267 (ihfA, 15 DDIs), CT344 (lon, 21 DDIs) and CT755 (groEL_3, 46 DDIs). These proteins are described in literature to interact with a multitude of bacterial proteins [295]. Compared to the complete chlamydial proteome, specific domain targets are found to be enriched regarding the set of effector proteins. Interacting domains that are enriched and achieve a p-value of less than 0.05 are listed in table 4.1. Several interacting domains that are over-represented reflect functions involved in general bacterial pathogenicity strategies as well as in host response towards infection. The characterized effector protein ChlaDub1(CT867) of *Chlamydia trachomatis* alters host cell physiology and promote bacterial survival in host tissues. The cysteine protease was shown to possess deubiquitinating and deneddylating activity and is involved in inhibition of NFκB degradation [165]. Homologous to ChlaDub1 is the effector protein CT868, also a predicted cysteine protease [189]. Both proteins contain the domain signature of Ulp1 protease family, C-terminal catalytic domain/PeptidaseC48 (PF02902). Two domain-domain interactions are predicted in each of these effectors: One interaction with the Ubiquitin domain (PF00240) and a self hit to the Ulp1 protease domain. Both of these interactions were found to be enriched in the set of effector proteins. Furthermore, domains of the Hsp60 (PF00118), Hsp70 (PF00012) as well as Hsp90 (PF02518) families of heat shock proteins are over-represented in the effector interaction partners. Many of the host proteins containing this domains are ubiquitously expressed and play

key roles in the stress response of cells during pathogenic infection [26, 296, 326, 185]. Several chlamydial effectors were predicted to interact with the conserved region of 14-3-3 proteins (PF00244). Human 14-3-3 proteins regulate the NFκB signaling pathway by controlling the nuclear export of p65-IκBα [5]. During *Chlamydia trachomatis* infection, the host protein 14-3-3β was shown to localize to the surface of the inclusion vacuole where it interacts with the chlamydial inclusion membrane protein IncG [255]. No conserved Pfam domain signatures are recognized in IncG, therefore this interaction could not be reproduced. Interestingly, four other chlamydial effectors (CT205, CT441, CT755 and CT823) are predicted to interact with 14-3-3 proteins. The predicted interactions could yield light on the important role of 14-3-3 proteins for chlamydial pathogenicity that is proposed in literature [240]. The large family of ABC transporters (PF00005) includes members of many different functions. They share an exposed location in the host cell membrane and some members are involved in the transport of defense peptides for antigen presentation [131]. Effects on the host response towards pathogenic infection was shown for AtPDR8, a plasma membrane ABC transporter of *Arabidopsis thaliana* [156]. Several predicted domain-domain interactions are related to ribosomal proteins. Interestingly, for several ribosomal proteins, there is evidence for extra-ribosomal functions such as cell growth and apoptosis [200, 300]). A pathogenic targeting of human ribosomal protein S3 by a bacterial effector during E.coli infection was proven experimentally [112, 111]. Yet, no findings are reported on manipulation of ribosomal proteins during infection by *C. trachomatis*. Further investigation could clarify if the domain interaction partners predicted for ribosomal proteins are an artifact of the applied domain interaction prediction approach or have a biological background.

4.2.4 Discussion

Several domain domain interactions (DDIs) over-represented in chlamydial effectors reflect molecular functions that are related to host processes altered by pathogenic infections or that are related to the host immune response. While these interactions are in accordance with possible functions of effector proteins inside the host cell, several annotated functions of targeted host domains could identify novel processes important to pathogenicity. The domain based prediction could serve as a valuable approach to narrow down functional interactions within the host cell. While enrichment of certain functions might be a strong indicator for a relevance in pathogenicity, predicted interactions that are rare in the set of effector proteins could be of equal importance. For further analysis, all predicted domain-domain interactions are considered. The prediction method is dependent on the conservation of domain motifs within effector protein and target host protein. This restriction to conserved domain signatures puts an intrinsic limitation to the prediction approach. The Pfam protein domain database used in this study resembles a representative collection of domain models. DDI predictions might profit from the integration of domain signatures from additional methods. The Interpro consortium for functional classification of protein sequences list several more public resources. Regarding domain signatures of the complete Interpro database, functional domains are detected in 111 effector proteins, while Pfam domains are annotated in 91 of 156 effectors [235]. To evaluate the integration of additional domain signatures, further domain based prediction methods would need to be considered as the domain interaction prediction approach used in this study is restricted to DDIs between Pfam domain models.

Interacting host domain	Domain description	# interacting effectors
PF00012	Hsp70 protein	8
PF00005	ABC transporter	5
PF00118	TCP-1/cpn60 chaperonin family	4
PF02540	NAD synthase	4
PF00244	14-3-3 protein	4
PF00166	Chaperonin 10 Kd subunit	3
PF00573	Ribosomal protein L4/L1 family	3
PF00227	Proteasome subunit	3
PF02518	GHKL domain	3
PF00156	Phosphoribosyl transferase domain	3
PF01751	Toprim domain	3
PF00583	Acetyltransferase (GNAT) family	2
PF01025	GrpE	2
PF01052	Surface presentation of antigens (SPOA)	2
PF01380	SIS domain	2
PF04055	Radical SAM superfamily	2
PF03572	Peptidase family S41	2
PF00400	WD domain, G-beta repeat	2
PF00240	Ubiquitin family	2
PF02201	SWIB/MDM2 domain	2
PF01245	Ribosomal protein L19	2
PF00542	Ribosomal protein L7/L12 C-terminal domain	2
PF01206	SirA-like protein	2
PF05362	Lon protease (S16) C-terminal proteolytic domain	2
PF00534	Glycosyl transferases group 1	2
PF00376	MerR family regulatory protein	2
PF00132	Bacterial transferase hexapeptide (three repeats)	2
PF00439	Bromodomain	2
PF02441	Flavoprotein	2
PF01433	Peptidase family M1	2
PF00501	AMP-binding enzyme	2
PF02902	Ulp1 protease family, C-terminal catalytic domain	2
PF03129	Anticodon binding domain	2
PF00297	Ribosomal protein L3	2
PF00216	Bacterial DNA-binding protein	2
PF00072	Response regulator receiver domain	2
PF00595	PDZ domain (Also known as DHR or GLGF)	2
PF08241	Methyltransferase domain	2

Table 4.1: ***Over-represented domain interaction targets of the Chlamydia trachomatis effectome***
Shown are all enriched domain-domain interaction targets of chlamydial effector proteins compared to the complete chlamydial proteome (p-value $\leq$ 0.05) as well as the number of effector proteins that are predicted to target the respective domain.

4.3 Prediction of the chlamydial pathogen-host interactome in human

4.3.1 Introduction

The primary function of bacterial effector proteins is to alter molecular processes within the host cell. The domain interactions predicted for these bacterial proteins are expected to reflect the pathogen-host relationship. In this section, the domain interactions (DDIs) predicted for chlamydial effectors are utilized to identify interacting host protein candidates. Generally, the number of proteins annotated in the human genome that share the same domain signatures can be very high [170]. The domain based approach can therefore be expected to provide a large number of interaction partner candidates. For the identification of interaction partners in the human host proteome, additional information is taken into account. Bacterial pathogens affect predominantly specific tissues in the host organism. The host tissue in which the infection is initiated and spreads varies depending on the pathogen. *Chlamydia trachomatis* includes serovars comprising lymphogranuloma venereum (L1–L3), ocular (A–C) and genital (D–K) serovars [3]. Beside physical restrictions on the pathogens access to the human body, reasons for tissue tropism were also found in the molecular pathogen-host interactions [94], were the virulent function of a bacterial effector protein within the host is displayed specific in certain host tissues [207]. Regarding tissue tropism of different *C. trachomatis* serovars, tissue specific protein expression is considered in the prediction of interacting host proteins. The integration of gene expression data offers a possibility to narrow down interaction partner candidates to those proteins with experimental evidence for expression. Several molecular functions as well as biological processes and pathways are expected to be enriched in targeted host proteins that are crucial for pathogenicity, such as regulation of the cell cycle, subversion of the host's DNA replication and transcription machinery, manipulation of host cellular programs such as apoptosis, immune response and NF-κB pathways [85]. The majority of predicted pathways are expected to reflect typical host functions altered during pathogenic infection that are supported by literature. Over-represented novel functions might hint towards additional infection strategies that could prove helpful in candidate selection for experimental studies. Many of the biological processes occurring in *C. trachomatis* infection are orchestrated by a large number of effectors [193]. While different effectors are expected to target identical host proteins, host pathways that are crucial for infection are also likely to be targeted at several different stages [109]. The predicted pathogen-host protein interactions could provide a comprehensive inventory of interaction candidates that allows a sensitive analysis of the interplay between different effector proteins during pathogenic infection.

4.3.2 Material and methods

All human protein sequences as well as Pfam domain signatures of human protein sequences as detected by InterproScan were retrieved from the SIMAP database of protein similarities [235]. A data set on differential protein expression in healthy human tissues collected from public resources was downloaded from the Human Protein Reference Database (HPRD). The HPRD provides a collection of manual curated and high quality protein expression data from different public resources and experiments [149, 117]. The general scope of protein expression data for a number of representative human tissue types is shown in table 4.2. The overall protein sequences integrated in HPRD cover most of the human proteins in the Refseq database. Genes might only be expressed in particular tissues under specific experimental or cellular conditions, any of these experiments therefore lack comprehensiveness. Nevertheless, the integrated data from HPRD, in line with other established approaches, represents a gold standard for tissue specific protein expression in human [226].

The human protein-protein interaction (PPI) network was also retrieved from the Human Protein Reference Database.

Domain-domain interactions (DDIs) link a bacterial effector domain to domains in proteins of the human proteome. Human proteins that include the DDI target domain are predicted to form a pathogen-host protein interaction with the particular chlamydial effector protein. An effector candidate that contains one or more protein domain signatures and may form a DDI with one or more domains from a host protein, then these two proteins are considered to interact with each other and constitute a pathogen-host protein-protein interaction (PH-PPI). These PH-PPIs are determined for all annotated human proteins listed in the HPRD.

Tissues primarily affected during infection of the human host by different *C.trachomatis* serovars include mucosa [43], cervix/endocervix [154] and vagina [154]. Secondary infection tissues are for example prostate [89], testis [141], epididymis [251] and lymph nodes [285]. Proteins that are expressed in infected tissue types are ranked above interaction candidates without experimental evidence for expression. Furthermore, effectors have been shown to interact preferentially with hubs of the host PPI network [195]. We use

Tissue	**#observed expressed proteins**
Brain	5881
Kidney	5498
Liver	5431
Heart	5203
Lung	5097
Placenta	4628
Testis	4347
Pancreas	4333
Skeletal muscle	4178
Spleen	3766
Ovary	3268
Leukocyte	2903
Colon	2753
Prostate	2713
Small intestine	2516
...	...
$\sum$	12227

Table 4.2: *Number of expressed proteins in different human tissues and cell types listed in the HPRD database (581 in total), as well as the total number of observed proteins in any of the tissue/cell types.*

this tendency to rank the proposed interaction candidates. The degree of human proteins in the human PPI network is determined and interaction partners are ranked accordingly in decreasing order. Actual interactors should rank high in the respective lists.

For all protein interaction candidates, a Gene Ontology (GO) enrichment analysis as well as the enrichment of functional KEGG pathways was conducted using FatiGO with default parameters [8]. FatiGO takes two lists of proteins and extracts the according Gene Ontology terms for the list of all human proteins and the list of protein interaction candidates. A Fisher's exact test is used to check for significant over-representation of GO terms in the set of interaction partners.

4.3.3 Results

Regarding all predicted domain interactions (DDIs) for the effectome of *C. trachomatis*, 71 % of all DDIs of chlamydial effectors are recovered by domain signatures annotated in human proteins. This way, interactions to human host proteins were predicted for 43 chlamydial effectors. A total of 3602 human proteins are predicted as targets of these chlamydial effector proteins. For 623 interaction candidates, expression in human tissues prone to chlamydial infection has been directly observed.

Interaction candidates inside the host cell are involved in several molecular functions as well as different biological processes. In these two categories, 157 and 131 unique GO terms were found to be enriched, respectively. Gene ontology (GO) terms over-represented in the set of interacting human proteins are listed in appendix tables 8.5 and 8.6. Several over-represented terms depict functional themes known to be targeted by pathogenic effector proteins: Among them are ubiquitin-protein transferase activity, PDZ domain binding and protein serine/threonine kinase activity. Enriched biological processes that are relevant to pathogenicity include cellular response to stimulus, MAPK cascade, intracellular transport and ER-nucleus signaling. Regarding GO terms referring to the cellular component of the host proteins, the enriched locations that are furthermore targeted by the maximum number of effector proteins are the mitochondrial part (29), insoluble fraction (23), membrane-bounded vesicle (22), cytoplasmic membrane-bounded vesicle (22) and the membrane fraction (22), compare appendix table 8.8. An analysis of KEGG pathways reveals 59 pathways that are enriched among the set of interacting proteins. Many of these pathways are targeted at multiple stages by several different chlamydial effectors, for example the regulation of actin cytoskeleton, ubiquitin mediated proteolysis and the MAPK signaling pathway. Table 8.7 lists KEGG pathways that are over-represented among host protein targets of the *C. trachomatis* effectome. For several of these effectors possible functional roles are proposed and described in recent literature.

During the initiation of infection, a crucial step is the disassembly of the host actin cytoskeleton to gain entry to the host cell [161, 256]. Attachment and entry of the pathogen into the host cell is only enabled by remodeling of the host actin network [47]. Without a switch shifting the balance of actin dynamics from polymerization to disassembly, this physical barrier would prevent internalization of the pathogen [46]. Two chlamydial effectors, CT456 (Tarp) and CT166 seem to play key roles for this important early stage of infection. Effector CT166 is targeting the host GTPase Rac to induce actin re-organization [284], while Tarp is binding to host guanine nucleotide exchange factors to activate the host signaling cascade to recruit actin [162]. In this study, eight more chlamydial effector proteins are predicted to target host substrates participating in the regulation of actin cytoskeleton in the host cell, compare figure 4.1. Thereby, the two proteases CT441 (Tsp) and CT823 (htrA) are both predicted to interact with the human rho guanine nucleotide exchange factor NP_056128. This might indicate that the effector proteins Tarp, Tsp and htrA share similar functions during the early stages of infection, shutting down the actin pro-assembly signaling by inhibition of Rho GTPases.

In table 8.4, targeted host proteins are ranked by the degree in the PPI network of the

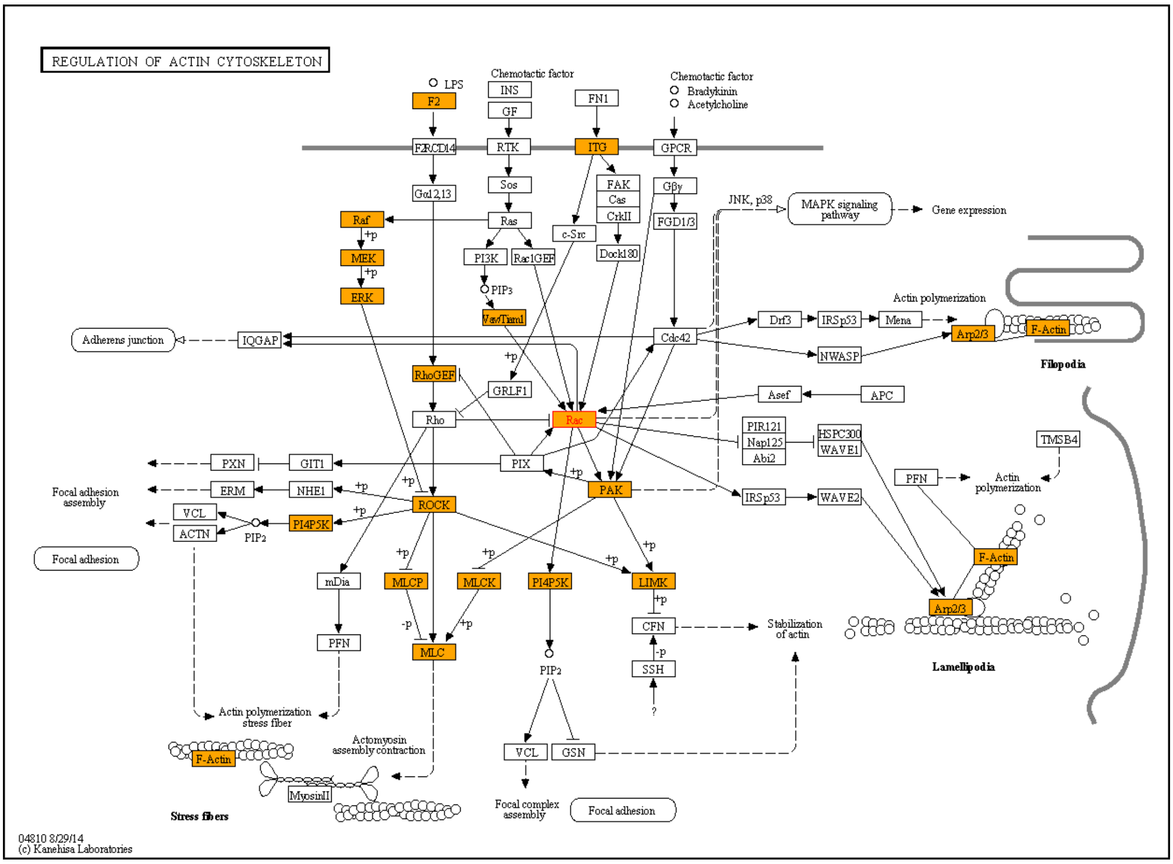


Figure 4.1: *Chlamydial-host PPI targets in the regulation of actin cytoskeleton.* *Shown is the KEGG map for pathway 'Regulation of actin cytoskeleton' (KEGG id: 04810). Indicated are all interaction candidates of chlamydial effector proteins targeting this pathway in human.*

human cell. All interactions between *C. trachomatis* effectors and the expressed host protein with the highest degree are listed. If no interaction candidate was found to be expressed in infected tissues, the annotated protein with the highest degree is considered. Experimental assays could provide additional candidate PPIs and experimental evidence of interacting host proteins. Hence, one aim of this work was to facilitate candidate selection for experimental analysis and to support experimental collaboration partners within the Pathomics project. The predicted pathogen-host interactomes for *Pseudomonas aeruginosa* PAO1, *Chlamydia pneumoniae* and *Chlamydia trachomatis* were made available online to provide easy access to all relevant information for manual inspection and further analysis.

4.3.4 Discussion

Molecular functions, biological processes and cellular pathways that are regulated by the host proteins interacting with chlamydial effectors provide information on the biological context of the chlamydial effectome inside the host cell. The domain-based prediction method represents a possible approach to investigate characteristics and properties of pathogen-host interactions (PH-PPI) and select candidates for further experimental analysis. The domain-based approach is independent of experimentally observed pathogen-host interactions. It is based on the assumption that basic characteristics of inter-species interactions are similar to intra-species PPIs. Without a gold standard of characterized PH-PPIs it is difficult to comprehensively evaluate the prediction results. Due to the limited data on experimentally validated interactions the prediction approach is tested regarding the biological plausibility. Gene Ontology (GO) terms and KEGG pathways enriched in the set of targeted host proteins depict functional the contexts that were reported to be altered by chlamydial infection [31]. Beside several known functions, the enriched terms reveal also novel functional contexts that could be important for *C. trachomatis* pathogenicity. E.g. ATPase activity is enriched in host interaction candidates and targeted by 24 putative chlamydial effectors. While not being addressed experimentally, this could indicate that *Chlamydia* inhibits ATP hydrolysis and proton translocation, a pathogenic strategy applied by other obligate intra-cellular pathogens, e. g. *Legionella pneumophila* [315].

Considering tissue expression data and the degree of interacting proteins within the host PPI network provide a possibility to rank the numerous possible effector targets. Both criteria include strong assumptions about the biological background. The tendency of pathogens to target hubs in the host network might depend on specific properties of the particular pathogen related to bacterial lifestyle. It is suggested that persistent intra-cellular pathogens follow a tendency to reduce their impact on the host cell by targeting host proteins with different properties than extra-cellular pathogens that cause acute symptoms in the host [42]. Bacterial pathogens preferentially infect specific host tissues. The function of a pathogenic effector protein can be specific for a particular host tissue type. Gene expression in cells of different tissue types in the human body undergo high variation [305]. Particular genes may not be expressed at all in a particular tissue, while others are only expressed during certain phases of the cell cycle. On the other hand, apart from differentiation in gene expression, also post-translational regulation and tissue dependent expression of protein isoforms play major roles in tissue differentiation [236]. Experimental evidence shows that tissue specificity is achieved by the precise regulation of protein levels. Besides controlling which protein is expressed, different tissues in the body might acquire their unique characteristics by controlling how much of a protein is produced [226]. The current status in research on mechanisms of tissue specific regulation of protein levels suggest that the influence of differential gene expression in the human body is far from being fully understood. Data from experiments on protein expression as well as the degree of interacting proteins in the host network are applied as criteria to rank pathogen-host PPI candidates. Many biologically reasonable predictions might not fall into one of these categories. Both integration of gene expression data as well as including the position of interactors in the host PPI network need further evaluation based on individual properties of pathogens with different host phenotypes.

Many challenges related to the prediction of pathogen-host PPI networks are still not addressable. On the one hand, there are only few effector proteins functionally characterized. On the other hand, while domain-based approaches are rather unspecific, only a limited very number of pathogen-host PPIs is experimentally verified that could provide a basis for training pathogen specific PPI prediction approaches. The domain-based PH-PPI prediction approach could hint to possible functional roles of still uncharacterized effector proteins. Prediction methods were explored in the scope of the EraNet project in addition to experimental studies on interaction proteomics, determining how predictions could support experimental data. Experimental verification of the predicted interactions between *C. trachomatis* effectors and proteins of the host PPI network would be necessary to further valuate predicted interactions. Within these boundaries, we provide pathogen-host interaction networks that can support candidate generation for further experimental studies.

5 Classification of bacterial phenotypes based on genomic features

5.1 Motivation

In many areas of microbial diagnostics, early recognition of the potentially harmful phenotype of novel microbial agents is a critical task. Traditional methods rely on phenotypic identification by culture and biochemical experiments. Major drawbacks of most traditional phenotypic procedures is the limitation to organisms that can be cultivated in vitro. Furthermore, unknown species could exhibit unique biochemical characteristics that do not fit into known patterns for the detection and characterization of bacterial phenotypes. Novel experimental screening approaches try to overcome some of these obstacles and succeed in creating phenotype-genotype maps of genetically manipulatable bacterial organisms, e.g. the human pathogen *Streptococcus pneumoniae* [286]. The advances in next generation sequencing, open the possibility to enhance or even replace traditional experimental methods by fast and reliable bioinformatics methods. By these approaches, basic genetic features of complex bacterial traits are revealed, e.g. associations with known enzymatic pathways, molecular complexes and signaling pathways [265]. Comparative genomics could also offer approaches to determine the pathogenic/symbiotic phenotype of a bacterial microorganism. Starting point of the analyses is the genome sequence of a novel sequenced bacterium. This chapter explores the possibilities to predict bacterial phenotypes by comparative genomics methods regarding basic molecular mechanisms important for pathogenicity and general interaction with the host cell. Possible features for the prediction of bacterial phenotypes could be the molecular mechanisms to establish an interaction with the host. E.g. several different secretion systems can be encoded in the genome of bacterial organisms. Secretion systems in general are associated with a variety of molecular processes in bacteria of any phenotype while the main function of several systems in particular is to establish a connection to the eukaryotic host cell [25]. The Type III Secretion System (T3SS) is the secretion system most directly linked to virulence and found in many gram-negative pathogenic/symbiontic bacteria [61]. It enables the pathogen to inject bacterial proteins directly into the host cell. As the implementation of the macro-molecular machinery is very costly for the organism, it is expected to be evolutionary conserved only in bacteria whose lifestyle depends on the successful delivery of effector proteins. Considering the necessity of secretion systems to deliver effectors into the host cell, a classification of bacterial phenotypes based on these genomic features could constitute a promising approach. On the other hand, the recognition of functional secretion systems in bacteria which is a prerequisite to an accurate classification is still challenging, compare section 2.3.3. Further limitations of a secretion system based approach to identify bacterial phenotypes are revealed when considering the coverage of these systems. According to results in section 2.3.3, only about 17% (188 of 1131) of sequenced pathogens/symbionts do encode a functional T3SS. An analysis of orthologous groups regarding components of the Type IV secretion system (T4SS) suggests that coverage of the T4SS is in the same range. Apart from the pathogen/symbiont specific secretion systems T3SS and T4SS, many pathogens rely on the Type II secretion system (T2SS) to secrete virulence factors. T2SS is encoded in non-host-interacting bacteria alike and could be considered unspecific for the bacterial phenotype [158]. Also the Type VI secretion system (T6SS) is not confined solely to pathogenic bacteria, orthologs of several T6SS components are widespread in the kingdom of bacteria [35].

The research topic addressed in this chapter is to evaluate a classification of bacterial phenotypes based on another major genomic feature in which pathogens as well as symbionts differ from non-pathogenic, non-host interacting bacteria. Bacterial organisms of host-interacting phenotypes depend on the existence of effector proteins to modulate functional processes inside the host cell [33]. It is not yet investigated to which extent bacterial effectors have discriminatory power to classify bacterial genomes regarding their phenotype. The eukaryotic-like domain (ELD) approach offers the possibility to evaluate an effector based classification of bacterial phenotypes independently of a bacterial secretion systems. The ELD based approach could provide a conceptual advantage and extend a secretion system based classification. Secretion system based classification needs complete genomes, a premise that is not given e.g. in metagenomic samples. A phenotype classification based on effector proteins could also work for contigs/unfinished genome sequences. In the following, a bacterial phenotype classification based on eukaryotic-like domain containing effector proteins is evaluated. Besides the contribution to existing *in silico* methods for the classification of bacterial phenotypes, characteristics of bacterial eukaryotic-like protein domains and the possibilities they provide to predict the phenotype of novel bacterial agents are explored.

5.2 Characteristics of eukaryotic-like protein domains in bacteria of diverse phenotype

5.2.1 Diversity of eukaryotic-like domain profiles between closely related bacterial genomes

Virulence factors are in general associated with the flexible part of the bacterial pangenome which is found to vary for strains of the same species [127]. E.g. in Streptococcus agalactiae, the majority of virulence associated genes are found in this dispensable gene pool of the pangenome [283]. Because of their implications for virulence within the genome of pathogenic bacteria, eukaryotic-like domain containing proteins (ELDPs) are expected to fall mostly into this category. A considerable average differences of eukaryotic-like protein domain (ELD) profiles between closely related pathogens could be a result of their functional context. Also ELDs in non-pathogenic bacteria could be exposed to higher variation than non-ELDs. These domains could have been acquired by lateral gene transfer and serve functions less vital to bacteria of non-pathogenic than of pathogenic or symbiotic lifestyle. Less evolutionary pressure results in lower detectable conservation of protein domains. In this section, the diversity of ELD profiles between bacterial genomes is analyzed and compared to the similarity of non-ELD profiles considering different levels of taxonomic divergence. The diversity of eukaryotic-like domains (ELDs) among bacterial genomes has implications for further classification and testing. High variety of ELDPs between closely related bacterial organisms could account for a well-suited and non-redundant data basis to classification.

Material and methods Domain profiles are analyzed for all genomes with distinct annotated phenotype in the genome repository. Within these 1706 bacterial genomes, strains of several species are over-represented. E. g. over 50 different strains of *E.coli* and *Helicobacter pylori* are listed, respectively. For details on the genome repository and assorting of genomes into classes of different taxonomic levels, see section 2.3. Similarly to the calculation of eukaryotic-like domains (ELDs) for individual bacterial genomes of pathogenic phenotype, identical background frequencies and standard deviations are used to identify pseudo eukaryotic-like domains in all non-pathogenic bacteria. Domain profiles are compared regardless of the phenotype of the particular bacterial organism. The pool of non-eukaryotic-like domains (non-ELDs) are defined as those domains, which are not recognized as eukaryotic-like (with an ELD score $>= 4$) in the given genome. ELD score differences are not taken into account for further analysis. For each pair of genomes, we can estimate the fraction of domains likewise identified as ELDs/non-ELDs in both genomes as well as the number of domains which are identified as ELDs only in any one of them. To estimate the diversity of domain profiles for different genomes, the Jaccard index or Jaccard similarity coefficient (JSC) is used. The JSC measures the similarity between two data sets by dividing the number of intersecting samples through the size of the union of the data sets:

$$J(A,B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid}$$

According to equation 5.2.1, a JSC of 1.00 translates to identical sets, while a JSC of 0.00 represents no overlap of sets A and B.

The Jaccard index and the according average standard deviation are calculated for the different taxonomic classes of species, genus, family, order and phylum. For each level, an all-against-all comparison of ELD and non-ELD profiles is performed for all genomes within that level. The average of these JSCs is a measure for the average similarity of domain profiles within the class, considering a particular taxonomic level. E.g. on the species level, JSCs are calculated for all possible pairs of strains of any one species. The average Jaccard similarity coefficient of the ELD profiles is a measure for the similarity of eukaryotic-like domain content in bacterial genomes on the particular taxonomic level.

Results Eukaryotic-like domain (ELD) profiles show considerable divergence on all taxonomic levels of comparison. The similarity of ELD profiles between different strains of the same species is 0.66 on average. Naturally, similarity decreases for higher taxonomic divergence, to an average of 0.14 for bacterial genomes of the same phyla. Average profile divergence for different taxonomic levels are compared in table 5.1.

The observed similarity between genomic ELD profiles on each level is considerably lower than the according average similarity of non-ELD profiles. Distributions for average Jaccard similarity coefficients (JSC) for ELD and non-ELD similarities over different taxonomic levels are visualized in figure 5.1.

Discussion Eukaryotic-like domain (ELD) profiles of closely related bacterial genomes were shown to have low average similarity. ELD profiles also show lower average similarity between bacterial organisms of the same taxonomic level than the profiles of non-eukaryotic-like domains. These findings are in accordance with expectations regarding the assumed biological background of eukaryotic-like domains. Nevertheless, the underlying

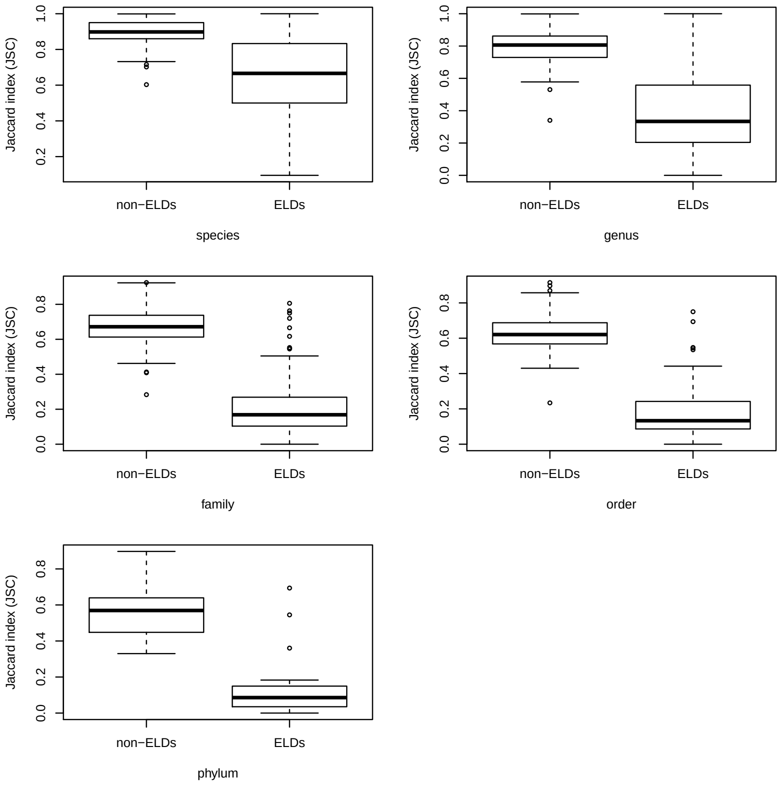


Figure 5.1: ***Average similarity of eukaryotic-like domain (ELD) and non-eukaryotic-like domain profiles between bacterial genomes on different levels of taxonomic divergence.*** *Distributions are shown in standard boxplot format, indicating Jaccard similarity coefficient (JSC) median values and standard deviations.*

calculation procedure for identifying eukaryotic-like protein domains cannot be excluded as a possible reason to enforce these effects. Due to their considerable variation, ELDs of closely related genomes are expected to add valuable information to the classification of bacterial phenotypes. Regarding the high average divergence of ELD profiles, bacterial genomes of the complete genome repository are used in the following test layouts.

Taxonomic divergence	non-ELDs		ELDs	
	JSC	Stdev	JSC	Stdev
species	0.90	0.07	0.66	0.21
genus	0.79	0.10	0.39	0.23
family	0.68	0.11	0.21	0.17
order	0.63	0.11	0.18	0.16
phylum	0.56	0.14	0.14	0.17

Table 5.1: ***Average eukaryotic-like domain (ELD) and non-eukaryotic-like domain (non-ELD) similarities between bacterial genomes for different levels of taxonomic divergence.*** *Listed are average Jaccard similarity coefficient (JSC) for non-eukaryotic-like domains (non-ELDs) and eukaryotic-like domains (ELDs) as well as according standard deviations (Stdev).*

5.2.2 The pathogenicity probability of a eukaryotic-like protein domains

Many pathogens are known to share common themes of virulence. Even infection strategies of bacterial families so divers as Chlamydiaceae [23], Mycoplasmataceae [58], Streptococcaceae [100] and Enterobacteriaceae [191] show common virulence themes [96, 324]. Similar implementations of these strategies on the genomic level are often realized by a limited pool of effector proteins with specialized functionality [108]. E.g. for Pseudomonas aeruginosa, Rahme et al. pointed out a remarkable conservation in virulence mechanisms used to infect hosts of divergent evolutionary origins [234]. The eukaryotic-like domain (ELD) score captures the deviation of a protein domain from the non-pathogenic background in a single pathogenic genome. Based on this calculations, the distribution of an ELD within different pathogens and symbionts compared to non-pathogens could hold additional information that could be integrated in a phenotypic classification.

Material and methods For each eukaryotic-like domain with a certain ELD score ($ELD_{id,score}$), the number of genomes with proteins containing that ELD is estimated for different phenotypes. pathogens/symbionts bacteria with proteins containing this ELD are compared to the number of non-pathogenic bacteria with that ELD and a respective score. The ELD score of a domain in non-pathogenic bacteria is calculated according to the formula for eukaryotic-like domain calculation given in chapter 1. Average occurrence and standard deviation values for score calculation are retrieved according to the distributions in pathogens/symbionts organisms. The pathogenicity probability P_{patho_ELD} of an $ELD_{id,score}$ reflects the ratio between the occurrences of this ELD in genomes of different phenotype:

$$P_{patho\_ELD}(ELD_{id,score}) = \frac{\#G_{pathogens/symbionts}(ELD_{id,score})}{\#G_{non-pathogens}(ELD_{id,score}) + \#G_{pathogens/symbionts}(ELD_{id,score})} \quad (5.1)$$

Where $\#G_{pathogens/symbionts}(ELD_{id,score})$ is the number of pathogens/symbionts genomes containing an instance of $ELD_{id,score}$ and $G_{pathogens/symbionts}(all)$ is the number of all pathogens/symbionts genomes considered in the analysis (for non-pathogens respectively). Pseudo counts in case of an ELD being exclusive to one phenotype are set to 1.0 for probability calculation.

Results Eukaryotic-like domains (ELDs) with a high pathogenicity probability P_{patho_ELD} reflect common virulence themes of divers pathogens. Table 5.2 shows the eukaryotic-like domains with the highest pathogenicity probability. All identified ELDs were considered independent of the particular score, which corresponds to the minimal ELD score cutoff of 4. Many domains are involved in functions commonly associated with pathogenic strategies in literature:

Ulp1, ubiquitin-like protein peptidase The ELD with the highest pathogenicity probabiltiy is the ubiquitin-like protein peptidase (Ulp1) domain. Ulp1 is the catalytic domain of the XopD effector (Xanthomonas outer membrane protein D), a TypeIII-secreted protein in Xanthomonas [134]. The importance for virulence of Ulp1 is also shown for several other pathogens [189, 291]. Considering the strong similarity to ULP1-like proteins in yeast and mammals, the origin of this effector is suggested to be a eukaryotic ULP1-like gene [57, 27].

Glycosyltransferase Family 25 of glycosyltransferases are involved in lipopolysaccharide (LPS) biosynthesis. Lipopolysaccharides are among the major virulence factors of gram-negative bacteria [105] and important for altering host immune response. Glycosyltransferase activity is observed for the effector protein NleB. NleB inhibits NF-κB activation in the host of several attaching/effacing human pathogens [113].

ACT domain The ACT domain is the C-terminal regulatory unit of 3-phosphoglycerate dehydrogenase and has homologs over a wide range of prokaryotic and eukaryotic organisms [56]. The observed enrichment in pathogens/symbionts could be an indication of its role in RelA proteins and their involvement in the NF-κB mediated response to stressful conditions such as pathogenic attack [34, 281].

Fumerase Gaham et al showed that a mutation in the Fumarase gene caused reduced virulence of the intracellular pathogen Listeria monocytogenes [107].

TCP-1/CPN60 The TCP-1/CPN60 chaperonines include T-complex proteins and the

HSP60 family of heat shock proteins. While these proteins can be found in any kingdom of life and they are mainly released by the organism on heat induced stress, they were found to have important secondary functions in symbiotic bacteria. E.g. the chaperonin GroEL has versatile roles in the interactions of bacterial endosymbionts and their hosts [92] and the HSP60 heat shock protein is a virulence factor of Helicobacter Pylori [145].

Pfamid	P_{patho_ELD}	**# in p/s**	**# in n**	**domain description**
PF02902	.97	29	0	Peptidase C48, Ulp1 protease
PF01755	.95	35	2	Glycosyltransferase family 25
PF13710	.94	33	2	ACT domain
PF05681	.94	15	1	Fumarate hydratase (Fumerase)
PF00118	.93	14	1	TCP-1/cpn60 chaperonin family
PF04991	.93	38	3	LicD family
PF03382	.92	24	2	DUF285
PF00982	.92	12	1	Glycosyltransferase family 20
PF01276	.92	12	1	Orn/Lys/Arg decarboxylase
PF13637	.92	11	1	Ankyrin repeats (many copies)
PF01306	.91	20	2	LacY proton/sugar symporter
PF00854	.91	20	2	POT family
PF02916	.91	10	1	DNA polymerase processivity factor
PF00589	.91	10	1	Phage integrase family
...	...	...	...	...

Table 5.2: ***Overview of eukaryotic-like protein domains (ELD) with highest pathogenicity probability (P_{patho_ELD})*** *Besides P_{patho_ELD}, the number of genomes with proteins containing the particular ELD (with ELD-score >= 4) are given for pathogens/symbionts and non-pathogenic bacteria.*

Discussion For many ELDs with high pathogenicity probability there is evidence for leading roles in the virulence of divers pathogens. The ELD containing proteins are likely to trigger molecular functions and biological processes which play major roles during infection. Therefore, specific eukaryotic-like effectors alone that are detected within a bacterial genome could allow for a separation of bacterial phenotypes. Besides the ELD score, the pathogenicity probability of a specific ELD could be used as an additional measure of its contribution to the pathogenicity of a bacterial organism and to increase prediction accuracy of the ELD based effector prediction approach.

5.2.3 Predictive power of a genomic pathogenicity probability derived from eukaryotic-like domains

The pathogenicity probability of a specific eukaryotic-like domain signatures is an indicator for its importance in the interplay of diverse pathogens/symbionts and their hosts. Independent of the particular pathogenic strategy it measures the functional relevance of a domain within a variety of pathogens/symbionts and could be used to measure the organisms potential for interaction as well. Based on a genomic pathogenicity probability, a separation of different bacterial phenotypes could be feasible and is explored in this section.

Methods

Definition of the genomic pathogenicity probability P_{patho_GENOME} To test the applicability of the pathogenicity probability to the classification procedure, the P_{patho_ELD} probability is adjusted to whole genomes. The genomic pathogenicity probability P_{patho_GENOME} is defined as

$$P_{patho\_GENOME}(genomeid) = \max_{ELD(x)\ in\ genome}(P_{patho\_ELD}(ELD(x))) \tag{5.2}$$

The pathogenicity probability of a bacterial genome is the maximum probability P_{patho_ELD} derived from the pool of all ELDs identified in the particular genome. It determines the probability that a given bacterial genome belongs to a pathogens/symbionts organism on the bases of the eukaryotic-like domains. This allows for a simple classification is based on the maximum pathogenicity probability of each genome.

Adjustments on the P_{patho_ELD} pathogenicity probability In general, ELD scores calculated for identical eukaryotic-like domains in different bacterial organisms do differ due to varying genome sizes or their presence in proteins of varying copy-number. With the P_{patho_ELD} calculation described in 5.2.2, ELD score differences of the same eukaryotic-like domain are not considered.

Details in the distribution of several ELDs show that there is more information contained regarding the differentiation of eukaryotic-like domains between pathogens/symbionts and non-pathogenic bacteria. The example of eukaryotic-like leucine rich repeats (PF00560) illustrates this point. This domain is widespread over a range of taxa. Besides in the genomes of 111 eukaryots and 184 pathogens/symbionts bacteria, it can be found also in 66 non-pathogenic bacteria. Due to its wide distribution, it is identified as eukaryotic in only 26 bacteria, in 20 pathogens as well as 6 non-pathogens. This results in a relatively low $P_{patho\_ELD}(PF00560) = 0.77$. Considering different ELD scores, PF00560 has an ELD score above 14 in 5 pathogens/symbionts, while this score is reached in none of the non-pathogens, resulting in a very high pathogenicity probability in the particular pathogens. In accordance, e.g. for Leptospira spp., a high copy number of leucine rich repeats in pathogenic strains against the background of low copy numbers in non-pathogenic strains was observed by Xue et al [316].

On the other hand, a completely rigid calculation considering each individual ELD score of each domain completely independent might also not directly reflect the biological background. E.g. the eukaryotic-like ankyrin repeat domain (PF00023) has an ELD score of 16 in a strain of the endosymbiont Wolbachia sp. wRi as well as in the genome of one non-pathogenic bacterium. For this specific ELD score, $ELD_{PF00023,16}$ is calculated to a low pathogenicity probability of 50%. Simultaneously, 7 other pathogens have ankyrin repeat domains with high ELD scores above 10, while no other non-pathogenic organism has highscoring ankyrin repeat domains.

Both examples show that a flexible approach using ELD score cutoffs could better reflect the biological relevance of this domain towards pathogenicity of a genome. Therefore, different minimum ELD score cutoffs partitions where tested to allow for a higher flexibility. Thereby all genomes are considered for probability calculation that contain proteins with the ELD scoring equal or higher to a particular score cutoff.

P_{patho_ELD} is determined for different ELD score intervals to test the influence on the discriminatory power of the subsequently calculated genomic probability P_{patho_GENOME}. In one attempt, the probabilities for each ELD score of a given domain are calculated separately. In all other settings, different score intervals are considered for calculation. This is achieved by summarizing the counts for subsequent ELD scores of each cutoff (from the particular ELD score to 10000 for each interval).

Evaluation of the discriminatory power of eukaryotic-like domain derived genomic properties The discriminatory power of ELD derived genomic properties is evaluated using Receiver Operating Characteristic (ROC) statistics [41]. To calculate all ROC statistics in this work, I applied the implementation provided in the ROCR-package of the R programming language for statistical data analysis [263]. ROC plots compare the true positive rate vs. the false positive rate of the prediction results at various threshold settings. The according area under the ROC-curve (AUC) is a standard measure for comparing the discriminatory power of different features in binary classification decisions [90]. It captures the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative one. The AUC typically ranges between 0.5 and 1, while values around 0.5 indicate random predictions. The higher the AUC value, the better the discriminatory power.

The evaluation procedure is based on a 10-fold cross-validation. All genomes considered for the analysis are divided randomly into 10 subsets of equal size, each of these sets serving as test data in a 10-fold cross-validation. In each of the 10 runs, the remaining 9 subsets are combined and constitute the training data, respectively. All genomes of the training data are the basis to identify eukaryotic-like domains and calculate the ELD scores. According to section 2.5.1, protein domains frequent in eukaryota and occurring in pathogens of the training data are selected. For these domains, average domain frequencies and standard deviation are calculated based on the genomes of non-pathogens. According to these values, ELD scores for all eukaryotic-like domains in genomes of the test set are calculated. The results of all 10 runs are combined to determine the relevant genomic properties that are used for further evaluation.

Results The genomic probability P_{patho_GENOME} reflects the pathogenicity of the phenotype. Medium levels of ELD score partitions show the highest discriminatory power in the separation of bacterial phenotypes. AUC scores for different ELD score cutoff partitions are listed in table 5.3. The ELD score cutoff partitioning of 4.10.20.50.100.10000 was found to have the highest discriminatory power with an AUC of 0.80 and is used for further analysis.

ELD score cutoffs	AUC
4.5... (separate scores)	0.79
4.5.6.7.8.9.10.50.100.500.10000	0.79
4.10.20.50.100.10000	0.80
4.20.30.40.50.100.500.10000	0.78
4-10000 (no partitions)	0.78

Table 5.3: ***Discriminatory power of*** P_{patho_GENOME} ***for a selection of different ELD score cutoff partitions.***

Other basic features were derived from the eukaryotic-like domains and evaluated. As expected, the discriminative power of the number of eukaryotic-like domain containing proteins in a bacterial genome alone is very low. The according AUC of 0.54 is close to random. The AUC of a separation of bacterial phenotypes based on the highest genomic

ELD score of a protein domain is in the same range with an AUC of 0.57. The pathogenicity probability P_{patho_GENOME} offers a significantly higher discriminatory power.

Discussion The results support initial assumptions on the information content of eukaryotic-like protein domains regarding the bacterial phenotype. Regarding the biological background of eukaryotic-like domains in bacteria of pathogenic/symbiotic phenotype, the information content of individual ELDs could be utilized to offer even higher phenotype prediction accuracy.

5.3 Pathofier - a machine learning approach to predict bacterial phenotypes based on effector protein candidates

The discriminative power of the pathogenicity probability of individual eukaryotic-like protein domains is combined in a machine learning approach to evaluate the prediction of bacterial phenotypes from these genomic features. In recent literature, there is no evidence for distinct inter-dependency and subsequent coevolution of effector proteins. Therefore, eukaryotic-like protein domains in the genomic domain profile of a bacterial organism could be regarded as independent features. The naive bayes approach assumes conditional independence of features and in these scenarios provides nearly optimal performance [319]. Considering the data characteristics, a naive Bayes classifier is expected to recognizes independence and structure of the input features and to constitute a suitable and effective machine learning approach for the challenge at hand.

5.3.1 Material and methods

Implementation of the naive Bayes approach In a first step, for each eukaryotic-like domain (ELD) and each ELD score the according a-priori probabilities $P(domainX, eukscore|isPathogen)$ are calculated. Numbers are derived from the distribution of eukaryotic-like domains in all genomes.
The pathogen probability model of the naive bayes classifier is calculated as

$$P(isPathogen|domainX, eukscore) = \frac{P(domainX,eukscore|isPathogen)*P(isPathogen)}{P(domainX,eukscore)} \tag{5.3}$$

For non-pathogens accordingly. For those eukaryotic-like domains that are only present in genomes of known pathogens, a pseudo count of 1.0 occurrence in non-pathogens is chosen, compare section 5.2.2.

The classification step of the naive bayes approach is realized in

$$Q = \frac{P(isPathogen|domainX, eukscore)}{P(isNonpathogen|domainX, eukscore)} \tag{5.4}$$

e.g. with $Q > 1$: organism is pathogens/symbionts and $Q \leq 1$: organism is non-pathogenic.

The classifiers ability to recognize interdependence and structure of the input features is evaluated. Under certain conditions, classifier performance can be overrated, especially in k-fold cross-validation procedures. These conditions include high dimensionality of input features coupled with low number of data points. The naive Bayes classifier is tested for its potential to correctly generalize beyond the training data for the given data structure.

Permutation test on randomized phenotypic annotations An accurate method is to assess the classifiers performance on permuted class labels in the training data. The observed performance on real data is compared to the performance on artificial data with randomly assigned phenotypic class labels. If standard performance measures do accurately reflect the potential of the classifier to generalize beyond the training data, the AUC of a classification based on randomized class labels is expected to be close to 0.5 [209].

The general conceptual design of the permutation procedure is taken from Good et al. [122]. For all bacterial genomes in the genome repository, phenotypic annotations (pathogenic, symbiotic, non-pathogenic) are randomly permuted. On this randomly labeled input data, the eukaryotic-like domain calculation and subsequent pathofier classification procedure is carried out. The process is iterated three times to compare different permutations.

Selection of representative genomes from taxonomically diverse bacteria On the taxonomic level of species, families and phyla, representative sets of genomes are generated to constitute test and training sets for classification. For example, the genome repository lists several different strains of the same species. When performance is tested on the level of species, only one genome is randomly chosen from several strains, to represent the particular bacterial species. Test and training set partitions for 10-fold cross-validation are constructed from this evaluation set.

Bacterial species Genomes in the genome repository are taxonomically distributed over 1003 species. For each species, one representative genome of each species is chosen randomly for further calculation. All organisms of the same species were found to have either all pathogens/symbionts or all non-pathogenic phenotypes. The input data for classification and 10-fold cross-validation consists of 1003 genomes from diverse bacterial species (529 of pathogens/symbionts and 474 of non-pathogenic phenotype).

Bacterial families Genomes are taxonomically subdivided into 204 bacterial families. The average number of members in a family is 5, with a median of 3. For classification, a maximum of 5 representative genomes of each species per family are selected to avoid over-representation of one group. Evaluation is based on 611 genomes. Separation into test- and training data in 10-fold cross-validation is performed on the family level.

Bacterial phyla The bacterial phyla are unevenly represented in the genome data. For most of the 24 phyla, there are only few genomes, while 447 genomes are listed for the phylum of Proteobacteria alone. For evaluation, a selection of representative phyla is chosen. These phyla include Proteobacteria, Bacteriodetes, Chlamydiae, Actinobacteria, Firmicutes and Cyanobacteria. A maximum of 50 genomes are selected randomly from each phylum, adding up to a total of 239 genomes. The evaluation procedure is a leave-one-out procedure on the phylum level. In each of the 6 cross-validation runs, the test set consists of all genomes of exactly one phylum.

Evaluation and performance measures Basis for measuring classification performance are the completely sequenced bacterial genomes of the genome repository. The classification performance is evaluated by 10-fold cross-validation. Of all genomes included in the analysis, 10 genome subsets are randomly generated. In each iteration, 9 of these sets are merged and used for training of the classifier, while the remaining genomes are used as the test set for evaluation.

In each run, the eukaryotic-like protein domains are determined from genomes of the training set only. Non-pathogenic organisms in the training set form the statistical background of the domains. On that basis, the ELD score of each eukaryotic-like domain is calculated. With average and standard deviations derived from the background of non-pathogenic genomes, the calculation is performed for all genomes in the training set, regardless of the phenotype.

Performance measures are calculated on the combined test set classification results of all 10 cross-validation runs to measure the overall performance of the classifier.

The area under the ROC curve (AUC) is an established single-number measure to evaluate the performance of classifiers [136]. Other measures to capture and compare the classifiers performance are accuracy, the percentage of correct predictions

$$Accuracy = TP + TN/TP + FP + TN + FN$$

precision, the percentage of positive predictions that are correct

$$Precision = TP/TP + FP$$

sensitivity, the percentage of positive labeled instances that were predicted as positive

$$Sensitivity = TP/TP + FN$$

and selectivity, the percentage of negative labeled instances that were predicted as negative

$$Selectivity = TN/TN + FP$$

where TP is the number of true positive, FP the false positive, FN the false negative and TN the true negative predictions.

5.3.2 Results

The classifier recognizes independence and structure of the input features To evaluate the potential of eukaryotic-like effector candidates to discriminate between different bacterial phenotypes, a naive Bayes classifier is implemented and applied to separate bacterial phenotypes on the basis of a-priori probabilities of the input data. The naive Bayes classification based on randomized class labels results in a very low Area Under the Curve (AUC) of 0.51. This reflects a classification performance close to random. Results are consistent with the expectation. The classifier can be assumed not to comprise any inherent, unspecific structure of the input data.

Exemplary classification of the pathogen Mycoplasma conjunctivae The human pathogen *Mycoplasma conjunctivae* HRC/581 has a very small genome. Two different eukaryotic-like protein domains are recognized. Choline/ethanolamine kinase domain (PF01633, eukscore: 6), which is suggested to play an important part in mycoplasma pathogenesis ([244]). And BRCT domain (PF12738, eukscore: 9), which is not found to be associated with virulence in recent literature.

For both domains, the a-priori probabilities are calculated based on their occurrence in the training data. For PF01633, an eukscore of 6 has been calculated for 18 pathogens/ symbionts (529 in total) and 3 non-pathogenic genomes (474 in total).
The a-priori probability for pathogenicity is calculated as

$$P(isPath|PF01633,6) = \frac{P(PF01633,6|isPath)*P(isPath)}{P(PF01633,6)} = \frac{18/529*529/1003}{21/1003} = 0.86$$

and for non-pathogenicity as 0.14, according to 5.3, respectively. PF12738 with an eukscore of 9 has been calculated for 2 pathogens/symbionts and 5 non-pathogenic organisms. The a-priori probability for pathogenicity and non-pathogenicity are calculated as 0.28 and 0.72, respectively. A naive bayes classification using 5.4 therefore resolves to

$$Q = \frac{P(isPath|PF01633,6)*P(isPath|PF12738,9)}{P(isNonpath|PF01633,6)*P(isNonpath|PF12738,9)} = \frac{0.86*0.28}{0.14*0.72} = 2.4$$

Mycoplasma conjunctivae has a score greater than 1 and is correctly classified as bacterial pathogen.

Performance on completely sequenced bacterial genomes A large fraction of bacterial phenotypes are predicted correctly. Bacteria of host-interacting and non-host-interacting phenotypes are classified with an AUC value of 0.84 regarding all completely sequenced genomes with characterized phenotypes 5.2.
Most true positive predictions of the host-interacting phenotype fall into the phyla of proteobacteria, actinobacteria, chlamydia, bacteriodetes, spirochaetes and firmicutes. The predominant phenotype in these phyla is pathogenic/symbiotic. On average, 64% of non-pathogens are predicted correctly in these phyla (279 of 444). Other phyla, for example Chlorobi or Tenericutes are dominated by non-pathogenic phenotypes. The true negative rate for the non-pathogenic phenotype in this phyla is 0.97. As expected, the performance of classifying phenotypes of distantly related bacterial organisms decreases (see table 5.4). Results on the species level have an AUC of 0.8. while predictions for organisms of different bacterial families reach an AUC of 0.7. Predicting phenotypes for bacteria of different phyla offers no predictive power, the performance measures indicate random results.

Tax. level	AUC	Acc.	Prec.	Selectivity	Sensitivity	(Random)
all genomes	0.84	0.8	0.85	0.7	0.84	0.51
species	0.8	0.76	0.75	0.71	0.8	0.51
family	0.7	0.67	0.6	0.69	0.64	0.51
phylum	0.49	0.53	0.52	0.69	0.35	0.48

Table 5.4: ***Performance measures of the pathofier approach for different levels of taxonomic divergence.*** *Measures of the classification of bacterial genomes into pathogens/symbionts and non-pathogenic phenotypes for the particular taxonomic level are given regarding the Area under the ROC-curve (AUC), accuracy (acc), precision (prec), selectivity and sensitivity. The last column shows the control, indicating the AUC of the classification approach on randomized data.*

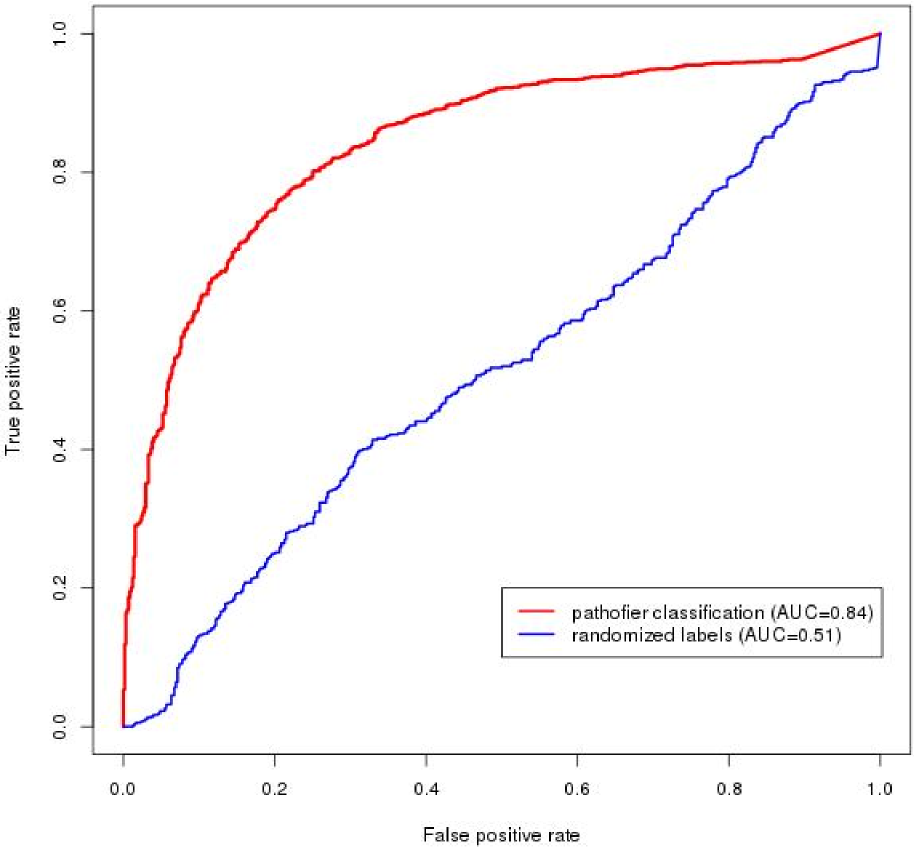


Figure 5.2: ***Performance of the pathofier approach for bacterial genomes in the genome repository.*** *The red ROC curve shows the performance of the classification of bacterial genomes into pathogens/symbionts and non-pathogenic phenotypes for all genomes in the genome repository. The blue curve is the control which indicates the performance of the classification approach on randomized data.*

5.3.3 Discussion

In this chapter, eukaryotic-like protein domains (ELDs) were explored as a novel genomic feature for the prediction of bacterial phenotypes. While the concept was generally shown to work, several issues need to be addressed before a software implementation could be realized to be used under laboratory conditions.

As could be expected, the prediction accuracy of the ELD approach drops considerably with increasing evolutionary divergence. An explanation could be that bacteria of different phyla are evolutionary very distantly related and have evolved divers strategies for initiating and maintaining interaction with host cells. Over these long evolutionary distances, several factors might shape the phenotypic lifestyle of a bacterial organism. Horizontal gene transfer (HGT) events occur infrequently and do not seem to provide considerable genotypic similarity.

Independently of any specific secretion system, the eukaryotic-like domain (ELD) based approach offers the possibility of an effector based classification of bacterial phenotypes. On the other hand, very high accuracy is needed to effectively address phenotype related questions in the field of molecular diagnostics which clearly cannot be attained by this approach alone. The current version of the ELD based approach might need further adjustments to be able to meet these criteria. A possible way could be, to improve classification of bacterial phenotypes by a combination of complementary genomic features. A genomic feature that could be used additionally to predict pathogenic/symbiotic bacteria is the presence of a functional protein secretion system. The ELD based approach could be combined with secretion system based predictions to further increase performance. The Type 3 secretion system (T3SS) is the best studied system and is assumed to be specific for targeting eukaryotic host cells. Many pathogens and symbionts could thereby be distinguished from non-pathogenic bacteria through the identification of genes encoding a functional T3SS within the bacterial genome, compare section 2.3.3. The T3SS was chosen as representative for an analysis of the potential to differentiate pathogenic and symbiotic from non-host-interacting bacteria by the genomic presence of secretion system components. As indicated in section 2.3.3, a classification based on the existence of a functional T3SS is still challenging and covers only a small number of pathogens/ symbionts.

Eukaryotic-like domains and the presence of a functional protein secretion system are two complementing genomic properties relevant for the particular phenotype in pathogens/ symbionts bacteria. To extend the secretion system based approach to other systems including a functional T3SS and to combine it with the prediction of ELDs could enable an accurate classification of bacterial phenotypes based on genomic features.

6 Summary and conclusions

This work focuses on the interaction of bacterial pathogens/symbionts and their eukaryotic hosts. Different aspects regarding these interactions have been investigated, based on genomic sequence data. On the bacterial side, I put the focus on a approach to predict secreted bacterial proteins, the molecular main determinants necessary for bacterial infection. The taxonomically universal method allows the prediction of these effector proteins and was made available to the scientific community via the Effective web portal. Furthermore, I investigated the possibility to utilize effector protein candidates predicted in bacterial genomes for the identification of a pathogenic/symbiotic bacterial phenotype. Regarding the analysis of host sided effects of bacterial infection, I explored the current state in the prediction of pathogen-host interactomes and possible means to narrow down protein interaction candidates for experimental analysis. A summary of the key findings is provided in the following.

A large-scale identification of eukaryotic-like protein domains in bacterial genomes allows for the prediction of secreted bacterial proteins Bacterial infections are orchestrated by so called effectors which are secreted by the microorganism to alter molecular functions and disrupt processes within the host cell. The identification of effector proteins is crucial to the understanding of bacterial virulence. Few effectors are characterized experimentally and existing prediction approaches are limited to the recognition of weakly conserved signal peptides. In this work, I developed a novel, function-based approach to predict effector proteins in genomic data. It was shown in single case studies that several bacterial effector proteins contain protein domains otherwise mainly found in eukaryotic proteins, allowing the pathogen to directly alter host cellular pathways. Based on the characteristic taxonomic distribution of their occurrence, I identified eukaryotic-like protein domains (ELDs) in a large-scale approach to predict effector candidates in the genomes of pathogenic as well as symbiotic bacteria. The method calculates an ELD score for each protein domain, reflecting the enrichment of this domain in pathogenic/symbiotic bacteria. Predicted effector candidates are ranked accordingly which allows to select only the most promising candidates for further experimental investigation. The ELD based method is the first to allow for a taxonomically universal, secretion-system independent prediction of effector proteins and was successfully tested experimentally for the obligate intra-cellular pathogen *Chlamydia trachomatis* by collaboration partners within the EraNet Pathogenomics project.

The Effective web portal provides precalculated effectomes and implements complementary methods to predict secreted bacterial proteins in user defined sequence data To make a state-of-the-art prediction of secreted bacterial proteins available to the scientific community, we implemented the Effective web portal. The Effective web portal represents a database for predicted secreted bacterial proteins. It is the first bioinformatics resource combining two complementary approaches for the prediction of bacterial secreted proteins: the function-based prediction by identification of eukaryotic-like domains and prediction based on signal peptides leading to transport by protein secretion systems. None of the two strategies can, by principle, achieve complete coverage. Therefore, their integration in a single resource is beneficial for the comprehensive annotation of putative effectors in genomes and proteomes. The user interface provides easy access to precalculated effectomes for all completely sequenced pathogens and symbionts. Fully automatic updates are performed on a regular basis and do include a full recalculation of eukaryotic-like protein domains on the actual genome repository. Furthermore, Effective provides a web interface that allows the user to predict effector proteins in their own sequence data, using a complementary set of prediction methods. Computational prediction of secreted bacterial proteins is important to many areas of microbial research and rapid progress in method development can be expected. The framework of the Effective portal allows for an easy integration of upcoming effector prediction approaches, facilitating access to the latest relevant methods.

Predicted domain interactions of a bacterial effectome target host functions playing major roles during pathogenic infection For a comprehensive understanding of the pathogen-host interplay, investigation of the molecular interactions taking place between bacterial effectors and targeted host proteins is required. The experimental investigation of pathogen-host protein interactions (PH-PPIs) is more challenging than for intra-species interactions. The limited availability of experimentally validated interactions increases the importance of *in silico* prediction methods for PH-PPIs. In this work, I explored the potential of domain-domain interactions (DDIs) to predict pathogen-host interactions.

DDIs were deduced mainly from structural and functional data on intra-species protein complexes. In an exemplary study, this information was transferred to the pathogen host inter-species context by mapping domain interactions onto effector proteins of the human pathogen *Chlamydia trachomatis* and the PPI network of the human host. A high number of domain targets are found to be predicted within the human proteome. This supports the initial assumption that inter-species interactions of bacterial effector domains can be deduced from intra-species protein interactions. To a large extent, bacterial effector proteins seem to affect host cellular functions by evolutionary well conserved mechanisms. Assessment of the functional context of human proteins targeted by DDIs of bacterial effector protein domains shows a functional enrichment of host processes and pathways known to be altered during pathogenic infection.

The prediction of pathogen-host interactomes profits from integrating information on the biological context and host PPI network Domain-domain interaction approaches provide numerous protein interaction candidates on the protein level especially for frequent domain signatures. To generate testable hypothesis from the DDI prediction results, the specificity must be increased by taking into account additional information. The tendency of effector proteins to target highly connected proteins of the host PPI network is used to rank predictions to extract the most likely interaction candidates. Bacterial pathogens predominantly infect specific host tissues. The host tissue in which a bacterial infection is initiated and spreads varies depending on the pathogen. Beside physical restrictions on the pathogens access to the host, reasons for tissue tropism were found on the molecular level. Filtering predicted host interactors according to tissue specific gene expression narrows down interaction candidates to proteins with experimental evidence for expression. For several bacterial effectors, a tendency to target hubs of the host PPI network was detected. Considering this observation as a weak signal of PH-PPIs, the degree of interacting host proteins within the host protein network provides the possibility to rank effector targets in a biologically meaningful way. A protocol that is based on the application of domain interaction methods in combination with gene expression and network-level information can reduce the number of PH-PPIs to an experimentally tractable set of predicted interaction candidates.

The host-associated phenotype of novel bacteria is to a considerable extent predicted correctly from virulence related genomic features Recognition of potentially harmful bacteria is a crucial challenge for microbial diagnostics. Bacterial effector proteins and the presence of a functional secretion system are two complementing genomic properties that are relevant for the particular phenotype in pathogenic/symbiotic bacteria. An initial phenotype classification based on the existence of a functional T3SS reveals challenges of the approach and covers only a small number of pathogens/symbionts. I evaluated the potential to predict phenotypic characteristics of bacterial organisms based on effector proteins predicted within the genome sequence. Thereby, I focused on eukaryotic-like protein domains (ELDs) as a particular feature of predicted bacterial effectors. By applying a classification approach based on eukaryotic-like protein domains, a large fraction of host-interacting bacteria are predicted correctly. To enable an accurate identification of pathogenic/symbiotic bacteria, the ELD based classification could be integrated into existing frameworks for genotype-phenotype prediction by multiple genomic features.

7 Bibliography

[1] Detection of protein-protein interactions using the GST fusion protein pull-down technique. *Nature Methods*, 1(3):275–276, December 2004.

[2] Abdallah M Abdallah, Nicolaas C Gey van Pittius, Patricia A DiGiuseppe Champion, Jeffery Cox, Joen Luirink, Christina MJE Vandenbroucke-Grauls, Ben J Appelmelk, and Wilbert Bitter. Type vii secretion—mycobacteria show the way. *Nature reviews microbiology*, 5(11):883–891, 2007.

[3] Hossam Abdelsamed, Jan Peters, and Gerald I Byrne. Genetic variation in Chlamydia trachomatis and their hosts: impact on disease severity and tissue tropism. *Future Microbiology*, 8(9):1129–1146, September 2013.

[4] Yousef Abu Kwaik and Christopher T. Price. Exploitation of host polyubiquitination machinery through molecular mimicry by eukaryotic-like bacterial F-box effectors. *Cellular and Infection Microbiology - closed section*, 1:122, 2010.

[5] Cristina Aguilera, Vanessa Fernandez-Majada, Julia Ingles-Esteve, Veronica Rodilla, Anna Bigas, and Lluis Espinosa. Efficient nuclear export of p65-IkappaBalpha complexes requires 14-3-3 proteins. *Journal of Cell Science*, 119(17):3695–3704, January 2006.

[6] Yukihiro Akeda and Jorge E. Galan. Chaperone release and unfolding of substrates in type III secretion. *Nature*, 437(7060):911–915, October 2005.

[7] Souhaila Al-Khodor, Christopher T. Price, Awdhesh Kalia, and Yousef Abu Kwaik. Functional diversity of ankyrin repeats in microbial proteins. *Trends in Microbiology*, 18(3):132–139, March 2010.

[8] Fatima Al-Shahrour, Pablo Minguez, Joaquin Tarraga, Ignacio Medina, Eva Alloza, David Montaner, and Joaquin Dopazo. FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Research*, 35(Web Server issue):W91–96, July 2007.

[9] John F. Alderete, Kevin W. Millsap, Michael W. Lehker, and Marlene Benchimol. Enzymes on microbial pathogens and Trichomonas vaginalis: molecular mimicry and functional diversity. *Cellular Microbiology*, 3(6):359–370, 2001.

[10] Gudni A. Alfredsson, Jakob K. Kristjansson, Sigridur Hjoerleifsdottir, and Karl O. Stetter. Rhodothermus marinus, gen. nov., sp. nov., a Thermophilic, Halophilic Bacterium from Submarine Hot Springs in Iceland. *Journal of General Microbiology*, 134(2):299–306, January 1988.

[11] Filipe Almeida, Vitor Borges, Rita Ferreira, Maria Jose Borrego, João Paulo Gomes, and Luis Jaime Mota. Polymorphisms in Inc Proteins and Differential Expression of inc Genes among Chlamydia trachomatis Strains Correlate with Invasiveness and Tropism of Lymphogranuloma Venereum Isolates. *Journal of Bacteriology*, 194(23):6574–6585, January 2012.

[12] Patrick Aloy, Hugo Ceulemans, Alexander Stark, and Robert B. Russell. The Relationship Between Sequence and Interaction Divergence in Proteins. *Journal of Molecular Biology*, 332(5):989–998, October 2003.

[13] Christian L. Althaus, Katherine M. E. Turner, Boris V. Schmid, Janneke C. M. Heijne, Mirjam Kretzschmar, and Nicola Low. Transmission of Chlamydia trachomatis through sexual partnerships: a comparison between three individual-based models and empirical data. *Journal of The Royal Society Interface*, page rsif20110131, June 2011.

[14] Whitney M. Amyot, Dennise deJesus, and Ralph R. Isberg. Poison Domains Block Transit of Translocated Substrates via the Legionella pneumophila Icm/Dot System. *Infection and Immunity*, 81(9):3239–3252, January 2013.

[15] D. M. Anderson, D. E. Fouts, A. Collmer, and O. Schneewind. Reciprocal secretion of proteins by the bacterial type III machines of plant and animal pathogens suggests universal recognition of mRNA targeting signals. *Proceedings of the National Academy of Sciences of the United States of America*, 96(22):12839–12843, October 1999.

[16] Deborah M. Anderson and Olaf Schneewind. A mRNA Signal for the Type III Secretion of Yop Proteins by Yersinia enterocolitica. *Science*, 278(5340):1140–1143, July 1997.

[17] Aurelie Angot, Annette Vergunst, Stephane Genin, and Nemo Peeters. Exploitation of Eukaryotic Ubiquitin Signaling Pathways by Effectors Translocated by Bacterial Type III and Type IV Secretion Systems. *PLoS Pathog*, 3(1):e3, January 2007.

[18] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[19] Roland Arnold, Kurt Boonen, Mark G. F. Sun, and Philip M. Kim. Computational analysis of interactomes: current and future perspectives for bioinformatics approaches to model the host-pathogen interaction space. *Methods (San Diego, Calif.)*, 57(4):508–518, August 2012.

[20] Roland Arnold, Stefan Brandmaier, Frederick Kleine, Patrick Tischler, Eva Heinz, Sebastian Behrens, Antti Niinikoski, Hans-Werner Mewes, Matthias Horn, and Thomas Rattei. Sequence-based prediction of type III secreted proteins. *PLoS pathogens*, 5(4):e1000376, 2009.

[21] Roland Arnold, Andre Jehl, and Thomas Rattei. Targeting effectors: the molecular recognition of Type III secreted proteins. *Microbes and Infection*, 12(5):346–358, May 2010.

[22] Nathan L. Bachmann, Adam Polkinghorne, and Peter Timms. Chlamydia genomics: providing novel insights into chlamydial biology. *Trends in Microbiology*.

[23] Robert J. Bastidas, Cherilyn A. Elwell, Joanne N. Engel, and Raphael H. Valdivia. Chlamydial Intracellular Survival Strategies. *Cold Spring Harbor Perspectives in Medicine*, 3(5):a010256, January 2013.

[24] Jon Beckwith. The Sec-dependent pathway. *Research in Microbiology*, 164(6):497–504, July 2013.

[25] DS Beeckman and DC Vanrompay. Bacterial secretion systems with an emphasis on the chlamydial type III secretion system. *Curr Issues Mol Biol*, 12(1):17–41, 2010.

[26] Helen M. Beere, Beni B. Wolf, Kelvin Cain, Dick D. Mosser, Artin Mahboubi, Tomomi Kuwana, Pankaj Tailor, Richard I. Morimoto, Gerald M. Cohen, and Douglas R. Green. Heat-shock protein 70 inhibits apoptosis by preventing recruitment of procaspase-9 to the Apaf-1 apoptosome. *Nature Cell Biology*, 2(8):469–475, August 2000.

[27] Miklos Bekes and Marcin Drag. Trojan Horse Strategies Used by Pathogens to Influence the Small Ubiquitin-Like Modifier (SUMO) System of Host Eukaryotic Cells. *Journal of Innate Immunity*, 4(2):159–167, 2012.

[28] S. D. Bentley, K. F. Chater, A.-M. Cerdeno-Tarraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C.-H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M.-A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren,

A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill, and D. A. Hopwood. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature*, 417(6885):141–147, May 2002.

[29] Inga Benz and M. Alexander Schmidt. Structures and functions of autotransporter proteins in microbial pathogens. *International Journal of Medical Microbiology*, 301(6):461–468, August 2011.

[30] M. Berry and O. M. Kon. Multidrug- and extensively drug-resistant tuberculosis: an emerging threat. *European Respiratory Review*, 18(114):195–197, January 2009.

[31] Helen J Betts, Katerina Wolf, and Kenneth A Fields. Effector protein modulation of host cells: examples in the Chlamydia spp. arsenal. *Current Opinion in Microbiology*, 12(1):81–87, February 2009.

[32] Aparna Bhaduri, Kun Qu, Carolyn S. Lee, Alexander Ungewickell, and Paul A. Khavari. Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics*, 28(8):1174–1175, April 2012.

[33] Amit P. Bhavsar, Julian A. Guttman, and B. Brett Finlay. Manipulation of host-cell pathways by bacterial pathogens. *Nature*, 449(7164):827–834, October 2007.

[34] Erik A. van der Biezen, Jongho Sun, Mark J. Coleman, Mervyn J. Bibb, and Jonathan D. G. Jones. Arabidopsis RelA/SpoT homologs implicate (p)ppGpp in plant signaling. *Proceedings of the National Academy of Sciences*, 97(7):3747–3752, March 2000.

[35] Lewis EH Bingle, Christopher M Bailey, and Mark J Pallen. Type VI secretion: a beginner's guide. *Current Opinion in Microbiology*, 11(1):3–8, February 2008.

[36] Sophie Bleves, Veronique Viarre, Richard Salacha, Gerard PF Michel, Alain Filloux, and Rome Voulhoux. Protein secretion systems in pseudomonas aeruginosa: A wealth of pathogenic weapons. *International Journal of Medical Microbiology*, 300(8):534–543, 2010.

[37] Adam J. Bogdanove and Daniel F. Voytas. TAL Effectors: Customizable Proteins for DNA Targeting. *Science*, 333(6051):1843–1846, September 2011.

[38] Thomas Boller and Sheng Yang He. Innate immunity in plants: An arms race between pattern recognition receptors in plants and effectors in microbial pathogens. *Science (New York, N.Y.)*, 324(5928):742–744, May 2009.

[39] Grady Booch, James Rumbaugh, and Ivar Jacobson. *The Unified Modeling Language User Guide.* Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1999.

[40] Pascale Bourhy, Sylvie Bremont, Farida Zinini, Claude Giry, and Mathieu Picardeau. Comparison of Real-Time PCR Assays for Detection of Pathogenic Leptospira spp. in Blood and Identification of Variations in Target Sequences. *Journal of Clinical Microbiology*, 49(6):2154–2160, January 2011.

[41] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997.

[42] Igor E. Brodsky and Ruslan Medzhitov. Targeting of immune signalling networks by bacterial pathogens. *Nature Cell Biology*, 11(5):521–526, May 2009.

[43] Robert C. Brunham and Jose Rey-Ladino. Immunology of Chlamydia infection: implications for a Chlamydia trachomatis vaccine. *Nature Reviews Immunology*, 5(2):149–161, February 2005.

[44] Christopher J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, June 1998.

[45] David Burstein, Tal Zusman, Elena Degtyar, Ram Viner, Gil Segal, and Tal Pupko. Genome-Scale Identification of Legionella pneumophila Effectors Using a Machine Learning Approach. *PLoS Pathog*, 5(7):e1000508, July 2009.

[46] Rey Carabeo. Bacterial subversion of host actin dynamics at the plasma membrane. *Cellular Microbiology*, 13(10):1460–1469, October 2011.

[47] Reynaldo A. Carabeo, Scott S. Grieshaber, Elizabeth Fischer, and Ted Hackstadt. Chlamydia trachomatis Induces Remodeling of the Actin Cytoskeleton during Attachment and Entry into HeLa Cells. *Infection and Immunity*, 70(7):3793–3803, January 2002.

[48] Arturo Casadevall and Liise-anne Pirofski. Host-Pathogen Interactions: Basic Concepts of Microbial Commensalism, Colonization, Infection, and Disease. *Infection and Immunity*, 68(12):6511–6518, January 2000.

[49] Eric Cascales. The type VI secretion toolkit. *EMBO reports*, 9(8):735–741, August 2008.

[50] Christel Cazalet, Christophe Rusniok, Holger Brüggemann, Nora Zidane, Arnaud Magnier, Laurence Ma, Magalie Tichit, Sophie Jarraud, Christiane Bouchier, François Vandenesch, et al. Evidence in the legionella pneumophila genome for exploitation of host cell functions and high genome plasticity. *Nature genetics*, 36(11):1165–1173, 2004.

[51] CDC. CDC Features - Trends in Foodborne Illness in the United States, 2012.

[52] Jeff H. Chang, Darrell Desveaux, and Allison L. Creason. The ABCs and 123s of Bacterial Secretion Systems in Plant Pathogenesis. *Annual Review of Phytopathology*, 52(1):null, 2014.

[53] Ding Chen, Lei Lei, Chunxue Lu, Rhonda Flores, Matthew P. DeLisa, Tucker C. Roberts, Floyd E. Romesberg, and Guangming Zhong. Secretion of the chlamydial virulence factor CPAF requires the Sec-dependent pathway. *Microbiology*, 156(10):3031–3040, January 2010.

[54] Lihong Chen, Jian Yang, Jun Yu, Zhijian Yao, Lilian Sun, Yan Shen, and Qi Jin. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Research*, 33(Database issue):D325–328, January 2005.

[55] Yi-Shan Chen. Type III Secretion Chaperones in Chlamydia trachomatis: Identification of a New Effector Protein and Insights into Hierarchical Protein Secretion during Early Infection. 2014.

[56] David M Chipman and Boaz Shaanan. The ACT domain family. *Current Opinion in Structural Biology*, 11(6):694–700, December 2001.

[57] Renee Chosed, Diana R. Tomchick, Chad A. Brautigam, Sohini Mukherjee, Veera S. Negi, Mischa Machius, and Kim Orth. Structural Analysis of Xanthomonas XopD Provides Insights into Substrate Specificity of Ubiquitin-like Protein Proteases. *Journal of Biological Chemistry*, 282(9):6773–6782, February 2007.

[58] Christine Citti and Alain Blanchard. Mycoplasmas and their host: emerging and re-emerging minimal pathogens. *Trends in Microbiology*, 21(4):196–203, April 2013.

[59] James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, 4(4):265–270, April 2009.

[60] Adam L. Clayton, Kelly F. Oakeson, Maria Gutin, Arthur Pontes, Diane M. Dunn, Andrew C. von Niederhausern, Robert B. Weiss, Mark Fisher, and Colin Dale. A Novel Human-Infection-Derived Bacterium Provides Insights into the Evolutionary Origins of Mutualistic Insect–Bacterial Symbioses. *PLoS Genet*, 8(11):e1002990, November 2012.

[61] Brian K Coombes. Type 3 secretion systems in symbiotic adaptation of pathogenic and non-pathogenic bacteria. *Trends in microbiology*, 17(3):89–94, 2009.

[62] Carlos Coronel, Steven Morris, and Peter Rob. *Database Systems: Design, Implementation, and Management.* Cengage Learning, November 2009.

[63] Daniele Corsaro, Marcello Valassina, and Danielle Venditti. Increasing Diversity within Chlamydiae. *Critical Reviews in Microbiology*, 29(1):37–78, January 2003.

[64] Sonia CP Costa, Alexa M Schmitz, Fathima F Jahufar, Justin D Boyd, Min Y Cho, Marcie A Glicksman, and Cammie F Lesser. A new means to identify type 3 secreted effectors: functionally interchangeable class ib chaperones recognize a conserved sequence. *MBio*, 3(1), 2012.

[65] Teresa A. Coutinho and Stephanus N. Venter. Pantoea ananatis: an unconventional plant pathogen. *Molecular Plant Pathology*, 10(3):325–335, May 2009.

[66] Antonello Covacci, John L. Telford, Giuseppe Del Giudice, Julie Parsonnet, and Rino Rappuoli. Helicobacter pylori Virulence and Genetic Geography. *Science*, 284(5418):1328–1333, May 1999.

[67] Randal Cox, Roberta J Mason-Gamer, Catherine L. Jackson, and Nava Segev. Phylogenetic Analysis of Sec7-Domain-containing Arf Nucleotide Exchangers. *Molecular Biology of the Cell*, 15(4):1487–1505, April 2004.

[68] J. Cruz, Y. Liu, Y. Liang, Y. Zhou, M. Wilson, J. J. Dennis, P. Stothard, G. Van Domselaar, and D. S. Wishart. BacMap: an up-to-date electronic atlas of annotated bacterial genomes. *Nucleic Acids Research*, 40(D1):D599–D604, December 2011.

[69] Maria da Cunha, Catarina Milho, Filipe Almeida, Sara V. Pais, Vitor Borges, Rui Mauricio, Maria J. Borrego, João P. Gomes, and Luis J. Mota. Identification of type III secretion substrates of Chlamydia trachomatis using Yersinia enterocolitica as a heterologous system. *BMC Microbiology*, 14(1):40, February 2014.

[70] Fred P. Davis, David T. Barkan, Narayanan Eswar, James H. McKerrow, and Andrej Sali. Host–pathogen protein interactions predicted by comparative modeling. *Protein Science : A Publication of the Protein Society*, 16(12):2585–2596, December 2007.

[71] Emmy De Buck, Elke Lammertyn, and Jozef Anne. The importance of the twin-arginine translocation pathway for bacterial virulence. *Trends in Microbiology*, 16(9):442–453, September 2008.

[72] Paul Dean. Functional domains and motifs of bacterial type III effector proteins and their roles in infection. *FEMS Microbiology Reviews*, 35(6):1100–1125, November 2011.

[73] Pierre Dehoux, Rhonda Flores, Catherine Dauga, Guangming Zhong, and Agathe Subtil. Multi-genome identification and characterization of chlamydiae-specific type III secretion substrates: the Inc proteins. *BMC Genomics*, 12(1):109, February 2011.

[74] P. Delepelaire. Type I secretion in gram-negative bacteria. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1694(1–3):149–161, November 2004.

[75] Edward F. DeLong and Norman R. Pace. Environmental Diversity of Bacteria and Archaea. *Systematic Biology*, 50(4):470–478, January 2001.

[76] Isabelle Derre, Rachel Swiss, and Herve Agaisse. The Lipid Transfer Protein CERT Interacts with the Chlamydia Inclusion Protein IncD and Participates to ER-Chlamydia Inclusion Membrane Contact Sites. *PLoS Pathog*, 7(6):e1002092, June 2011.

[77] Les Dethlefsen, Margaret McFall-Ngai, and David A. Relman. An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature*, 449(7164):811–818, October 2007.

[78] S. Diaz-Sanchez, I. Hanning, Sean Pendleton, and Doris D'Souza. Next-generation sequencing: The future of molecular genetics in poultry production and food safety. *Poultry Science*, 92(2):562–572, January 2013.

[79] Xavier Didelot, Rory Bowden, Daniel J. Wilson, Tim E. A. Peto, and Derrick W. Crook. Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics*, 13(9):601–612, September 2012.

[80] Andreas Diepold, Marlise Amstutz, Soeren Abel, Isabel Sorg, Urs Jenal, and Guy R Cornelis. Deciphering the assembly of the Yersinia type III secretion injectisome. *The EMBO Journal*, 29(11):1928–1940, June 2010.

[81] Andreas Diepold and Samuel Wagner. Assembly of the bacterial type III secretion machinery. *FEMS Microbiology Reviews*, pages n/a–n/a, February 2014.

[82] Kieran Dilks, R. Wesley Rose, Enno Hartmann, and Mechthild Pohlschroeder. Prokaryotic Utilization of the Twin-Arginine Translocation Pathway: a Genomic Survey. *Journal of Bacteriology*, 185(4):1478–1483, February 2003.

[83] Gerd Doering. Chronic Pseudomonas aeruginosa Lung Infection in Cystic Fibrosis Patients. In Mario Campa, Mauro Bendinelli, and Herman Friedman, editors, *Pseudomonas aeruginosa as an Opportunistic Pathogen*, Infectious Agents and Pathogenesis, pages 245–273. Springer US, January 1993.

[84] Tim Driscoll, Matthew D. Dyer, T. M. Murali, and Bruno W. Sobral. PIG—the pathogen interaction gateway. *Nucleic Acids Research*, 37(suppl 1):D647–D650, January 2009.

[85] Matthew D Dyer, T. M Murali, and Bruno W Sobral. The Landscape of Human Proteins Interacting with Viruses and Other Pathogens. *PLoS Pathog*, 4(2):e32, February 2008.

[86] Matthew D. Dyer, T. M. Murali, and Bruno W. Sobral. Supervised learning and prediction of physical interactions between human and HIV proteins. *Infection, Genetics and Evolution*, 11(5):917–923, July 2011.

[87] WW Eckerson. Three Tier Client/Server Architectures: Achieving Scalability, Performance, and Efficiency in Client/Server Applications. *Open Information Systems*, 3(20):46–50, 1995.

[88] Nels C Elde and Harmit S Malik. The evolutionary conundrum of pathogen mimicry. *Nature Reviews Microbiology*, 7(11):787–797, 2009.

[89] Poletti F, Medici Mc, Alinovi A, Menozzi Mg, Sacchini P, Stagni G, Toni M, and Benoldi D. Isolation of Chlamydia trachomatis from the prostatic cells in patients affected by nonacute abacterial prostatitis. *The Journal of urology*, 134(4):691–693, October 1985.

[90] David Faraggi and Benjamin Reiser. Estimation of the area under the ROC curve. *Statistics in medicine*, 21(20):3093–3106, October 2002.

[91] J M Farber and P I Peterkin. Listeria monocytogenes, a food-borne pathogen. *Microbiological Reviews*, 55(3):476–511, September 1991.

[92] Mario Ali Fares, Andres Moya, and Eladio Barrio. GroEL and the maintenance of bacterial endosymbiosis. *Trends in Genetics*, 20(9):413–416, September 2004.

[93] Scott Federhen. The NCBI Taxonomy database. *Nucleic Acids Research*, 40(D1):D136–D143, January 2012.

[94] Christine Fehlner-Gardiner, Christine Roshick, John H. Carlson, Scott Hughes, Robert J. Belland, Harlan D. Caldwell, and Grant McClarty. Molecular basis defining human Chlamydia trachomatis tissue tropism. A possible role for tryptophan synthase. *J Biol Chem*, 277(30):26893–26903, July 2002.

[95] K. A. Fields and T. Hackstadt. Evidence for the secretion of Chlamydia trachomatis CopN by a type III secretion mechanism. *Molecular Microbiology*, 38(5):1048–1060, December 2000.

[96] B. B. Finlay and S. Falkow. Common themes in microbial pathogenicity revisited. *Microbiology and Molecular Biology Reviews*, 61(2):136–169, January 1997.

[97] B. Brett Finlay and Pascale Cossart. Exploitation of Mammalian Host Cell Functions by Bacterial Pathogens. *Science*, 276(5313):718–725, February 1997.

[98] Robert D. Finn, Jody Clements, and Sean R. Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl 2):W29–W37, January 2011.

[99] Robert D. Finn, Benjamin L. Miller, Jody Clements, and Alex Bateman. iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Research*, 42(D1):D364–D373, January 2014.

[100] Nahuel Fittipaldi, Mariela Segura, Daniel Grenier, and Marcelo Gottschalk. Virulence factors involved in the pathogenesis of the infection caused by the swine pathogen and zoonotic agent *Streptococcus suis*. *Future Microbiology*, 7(2):259–279, February 2012.

[101] Brian M. Forde and Paul W. O'Toole. Next-generation sequencing technologies and their impact on microbial genomics. *Briefings in Functional Genomics*, 12(5):440–453, January 2013.

[102] Owen E. Francis, Matthew Bendall, Solaiappan Manimaran, Changjin Hong, Nathan L. Clement, Eduardo Castro-Nallar, Quinn Snell, G. Bruce Schaalje, Mark J. Clement, Keith A. Crandall, and W. Evan Johnson. Pathoscope: Species identification and strain attribution with unassembled sequencing data. *Genome Research*, 23(10):1721–1729, January 2013.

[103] Irina S. Franco, Howard A. Shuman, and Xavier Charpentier. The perplexing functions and surprising origins of Legionella pneumophila type IV secretion effectors. *Cellular Microbiology*, 11(10):1435–1443, October 2009.

[104] Remi Fronzes, Peter J. Christie, and Gabriel Waksman. The structural biology of type IV secretion systems. *Nature Reviews Microbiology*, 7(10):703–714, October 2009.

[105] Benjamin N. Fry, Shi Feng, Yuen-Yuen Chen, Diane G. Newell, Peter J. Coloe, and Victoria Korolik. The galE Gene of Campylobacter jejuni Is Involved in Lipopolysaccharide Synthesis and Virulence. *Infection and Immunity*, 68(5):2594–2601, January 2000.

[106] H. Fukushi and K. Hirai. Genetic diversity of avian and mammalian Chlamydia psittaci strains and relation to host origin. *Journal of Bacteriology*, 171(5):2850–2855, January 1989.

[107] Cormac G. M. Gahan and Colin Hill. The use of listeriolysin to identify in vivo induced genes in the Gram-positive intracellular pathogen Listeria monocytogenes. *Molecular Microbiology*, 36(2):498–507, 2000.

[108] Jorge E. Galan. Common Themes in the Design and Function of Bacterial Effectors. *Cell Host & Microbe*, 5(6):571–579, June 2009.

[109] Jorge E Galan and Pascale Cossart. Host-pathogen interactions: a diversity of themes, a variety of molecular machines. *Current opinion in microbiology*, 8(1):1–3, February 2005.

[110] Lian-Yong Gao and Yousef Abu Kwaik. Hijacking of apoptotic pathwaysby bacterial pathogens. *Microbes and Infection*, 2(14):1705–1719, November 2000.

[111] Xiaofei Gao and Philip R. Hardwidge. Ribosomal protein s3: a multifunctional target of attaching/effacing bacterial pathogens. *Front Microbiol*, 2:137, 2011.

[112] Xiaofei Gao, Fengyi Wan, Kristina Mateo, Eduardo Callegari, Dan Wang, Wanyin Deng, Jose Puente, Feng Li, Michael S. Chaussee, B Brett Finlay, Michael J. Lenardo, and Philip R. Hardwidge. Bacterial effector binding to ribosomal protein s3 subverts NF-kappaB function. *PLoS Pathog*, 5(12):e1000708, December 2009.

[113] Xiaofei Gao, Xiaogang Wang, Thanh H Pham, Leigh Ann Feuerbacher, Marie-Luise Lubos, Minzhao Huang, Rachel Olsen, Arcady Mushegian, Chad Slawson, and Philip R Hardwidge. NleB, a bacterial effector with glycosyltransferase activity, targets GAPDH function to inhibit NF-kappaB activation. *Cell host & microbe*, 13(1):87–99, January 2013.

[114] Lena Gehre. Identification and characterization of novel type III secreted proteins in chlamydia infection. 2011.

[115] R. D. Gitaitis, R. R. Walcott, M. L. Wells, J. C. Diaz Perez, and F. H. Sanders. Transmission of Pantoea ananatis, Causal Agent of Center Rot of Onion, by Tobacco Thrips, Frankliniella fusca. *Plant Disease*, 87(6):675–678, June 2003.

[116] Tijana Glavina Del Rio, Birte Abt, Stefan Spring, Alla Lapidus, Matt Nolan, Hope Tice, Alex Copeland, Jan-Fang Cheng, Feng Chen, David Bruce, Lynne Goodwin, Sam Pitluck, Natalia Ivanova, Konstantinos Mavromatis, Natalia Mikhailova, Amrita Pati, Amy Chen, Krishna Palaniappan, Miriam Land, Loren Hauser, Yun-Juan Chang, Cynthia D. Jeffries, Patrick Chain, Elizabeth Saunders, John C. Detter, Thomas Brettin, Manfred Rohde, Markus Goker, Jim Bristow, Jonathan A. Eisen, Victor Markowitz, Philip Hugenholtz, Nikos C. Kyrpides, Hans-Peter Klenk, and Susan Lucas. Complete

genome sequence of Chitinophaga pinensis type strain (UQM 2034t). *Standards in Genomic Sciences*, 2(1):87–95, February 2010.

[117] Renu Goel, Babylakshmi Muthusamy, Akhilesh Pandey, and T. S. Keshava Prasad. Human protein reference database and human proteinpedia as discovery resources for molecular biotechnology. *Molecular Biotechnology*, 48(1):87–95, May 2011.

[118] Laura Gomez-Valero, Christophe Rusniok, Christel Cazalet, and Carmen Buchrieser. Comparative and functional genomics of legionella identified eukaryotic like proteins as key players in host–pathogen interactions. *Frontiers in microbiology*, 2, 2011.

[119] Laura Gomez-Valero, Christophe Rusniok, Christel Cazalet, and Carmen Buchrieser. Comparative and functional genomics of Legionella identified eukaryotic like proteins as key players in host–pathogen interactions. *Frontiers in microbiology*, 2, 2011.

[120] Siqi Gong, Lei Lei, Xiaotong Chang, Robert Belland, and Guangming Zhong. Chlamydia trachomatis secretion of hypothetical protein CT622 into host cell cytoplasm via a secretion pathway that can be inhibited by the type III secretion system inhibitor compound 1. *Microbiology*, 157(4):1134–1144, January 2011.

[121] Sungsam Gong, Giseok Yoon, Insoo Jang, Dan Bolser, Panos Dafas, Michael Schroeder, Hansol Choi, Yoobok Cho, Kyungsook Han, Sunghoon Lee, Hwanho Choi, Michael Lappe, Liisa Holm, Sangsoo Kim, Donghoon Oh, and Jonghwa Bhak. PSIbase: a database of Protein Structural Interactome map (PSIMAP). *Bioinformatics*, 21(10):2541–2543, May 2005.

[122] Phillip Good. *Permutation, parametric and bootstrap tests of hypotheses*. Springer, 2005.

[123] J. Thomas Grayston, Lee Ann Campbell, Cho-Chou Kuo, Carl H. Mordhorst, Pekka Saikku, David H. Thorn, and San-Pin Wang. A New Respiratory Tract Pathogen: Chlamydia pneumoniae Strain TWAR. *Journal of Infectious Diseases*, 161(4):618–625, January 1990.

[124] Jean T Greenberg and Boris A Vinatzer. Identifying type III effectors of plant pathogens and analyzing their interaction with plant cells. *Current Opinion in Microbiology*, 6(1):20–28, February 2003.

[125] Eduardo A. Groisman and Howard Ochman. How Salmonella became a pathogen. *Trends in Microbiology*, 5(9):343–349, September 1997.

[126] Julien Guglielmini, Bertrand Neron, Sophie S. Abby, Maria Pilar Garcillan-Barcia, Fernando de la Cruz, and Eduardo P. C. Rocha. Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Research*, 42(9):5715–5727, January 2014.

[127] Joerg Hacker and Elisabeth Carniel. Ecological fitness, genomic islands and bacterial pathogenicity. *EMBO reports*, 2(5):376–381, May 2001.

[128] Charlotte Harrison. Antibacterial drugs: Overcoming MRSA resistance. *Nature Reviews Drug Discovery*, 11(5):354–354, May 2012.

[129] Richard A. Harvey. *Microbiology.* Lippincott Williams & Wilkins, 2007.

[130] Ian R. Henderson, Fernando Navarro-Garcia, Mickaël Desvaux, Rachel C. Fernandez, and Dlawer Ala'Aldeen. Type V Protein Secretion Pathway: the Autotransporter Story. *Microbiology and Molecular Biology Reviews*, 68(4):692–744, January 2004.

[131] C. F. Higgins. ABC transporters: from microorganisms to man. *Annual Review of Cell Biology*, 8:67–113, 1992.

[132] Anne-Sofie Hobolt-Pedersen, Gunna Christiansen, Evy Timmerman, Kris Gevaert, and Svend Birkelund. Identification of Chlamydia trachomatis CT621, a protein delivered through the type III secretion system to the host cell cytoplasm and nucleus. *FEMS Immunology & Medical Microbiology*, 57(1):46–58, October 2009.

[133] Matthias Horn and Michael Wagner. Bacterial Endosymbionts of Free-living Amoebae1. *Journal of Eukaryotic Microbiology*, 51(5):509–514, September 2004.

[134] Andrew Hotson, Renee Chosed, Hongjun Shu, Kim Orth, and Mary Beth Mudgett. Xanthomonas type III effector XopD targets SUMO-conjugated proteins in planta. *Molecular Microbiology*, 50(2):377–389, 2003.

[135] S. Hower, K. Wolf, and K. A. Fields. Evidence that CT694 is a novel Chlamydia trachomatis T3s substrate capable of functioning during invasion or early cycle development. *Molecular Microbiology*, 72(6):1423–1437, June 2009.

[136] Jin Huang and C.X. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, March 2005.

[137] Andree Hubber and Craig R. Roy. Modulation of Host Cell Function by Legionella pneumophila Type IV Effectors. *Annual Review of Cell and Developmental Biology*, 26(1):261–283, 2010.

[138] Sarah Hunter, Philip Jones, Alex Mitchell, Rolf Apweiler, Teresa K. Attwood, Alex Bateman, Thomas Bernard, David Binns, Peer Bork, Sarah Burge, Edouard de Castro, Penny Coggill, Matthew Corbett, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, Robert D. Finn, Matthew Fraser, Julian Gough, Daniel Haft, Nicolas Hulo, Daniel Kahn, Elizabeth Kelly, Ivica Letunic, David Lonsdale, Rodrigo Lopez, Martin Madera, John Maslen, Craig McAnulla, Jennifer McDowall, Conor McMenamin, Huaiyu Mi, Prudence Mutowo-Muellenet, Nicola Mulder, Darren Natale, Christine Orengo, Sebastien Pesseat, Marco Punta, Antony F. Quinn, Catherine Rivoire, Amaia Sangrador-Vegas, Jeremy D. Selengut, Christian J. A. Sigrist, Maxim Scheremetjew, John Tate, Manjulapramila Thimmajanarthanan, Paul D. Thomas, Cathy H. Wu, Corin Yeats, and Siew-Yit Yong. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research*, 40(D1):D306–D312, January 2012.

[139] Daniel H. Huson, Alexander F. Auch, Ji Qi, and Stephan C. Schuster. MEGAN analysis of metagenomic data. *Genome Research*, 17(3):377–386, January 2007.

[140] Catherine L. Jackson and James E. Casanova. Turning on ARF: the Sec7 family of guanine-nucleotide-exchange factors. *Trends in Cell Biology*, 10(2):60–67, February 2000.

[141] C. Jantos, W. Baumgaertner, B. Durchfeld, and H. G. Schiefer. Experimental epididymitis due to Chlamydia trachomatis in rats. *Infection and Immunity*, 60(6):2324–2328, January 1992.

[142] Haeyoung Jeong, Joung Han Yim, Choonghwan Lee, Sang-Haeng Choi, Yon Kyoung Park, Sung Ho Yoon, Cheol-Goo Hur, Ho-Young Kang, Dockyu Kim, Hyun Hee Lee, Kyun Hyang Park, Seung-Hwan Park, Hong-Seog Park, Hong Kum Lee, Tae Kwang Oh, and Jihyun F Kim. Genomic blueprint of Hahella chejuensis, a marine microbe producing an algicidal agent. *Nucleic acids research*, 33(22):7066–7073, 2005.

[143] Andrey V. Kajava and Alasdair C. Steven. The turn of the screw: variations of the abundant beta-solenoid motif in passenger domains of Type V secretory proteins. *Journal of Structural Biology*, 155(2):306–315, August 2006.

[144] Sue Kalman, Wayne Mitchell, Rekha Marathe, Claudia Lammel, Jun Fan, Richard W. Hyman, Lynn Olinger, Jane Grimwood, Ronald W. Davis, and Richard Stephens. Comparative genomes of Chlamydia pneumoniae and C. trachomatis. *Nature Genetics*, 21(4):385–389, April 1999.

[145] Shigeru Kamiya, Hiroyuki Yamaguchi, and Haruhiko Taguchi. A Virulence Factor of Helicobacter pylori: Role of Heat Shock Protein in Mucosal Inflammation After H. pylori Infection.

[146] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue):D109–D114, January 2012.

[147] Nicole Kapitein and Axel Mogk. Type VI Secretion System Helps Find a Niche. *Cell Host & Microbe*, 16(1):5–6, July 2014.

[148] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, Christine Jandrasits, Rafael C. Jimenez, Jyoti Khadake, Usha Mahadevan, Patrick Masson, Ivo Pedruzzi, Eric Pfeiffenberger, Pablo Porras, Arathi Raghunath, Bernd Roechert, Sandra Orchard, and Henning Hermjakob. The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40(D1):D841–D846, January 2012.

[149] T. S. et al. Keshava Prasad. Human Protein Reference Database–2009 update. *Nucleic Acids Res*, 37(Database issue):D767–D772, January 2009.

[150] Christina Tobin Kåhrstroem. Entering a post-antibiotic era? *Nature Reviews Microbiology*, 11(3):146–146, March 2013.

[151] Hyun Jung Kim, Jin Hee Lee, Beom Ryong Kang, Xiaoqing Rong, Brian B. McSpadden Gardener, Hyung Jin Ji, Chang-Seuk Park, and Young Cheol Kim. Draft Genome Sequence of Pantoea ananatis B1-9, a Nonpathogenic Plant Growth-Promoting Bacterium. *Journal of Bacteriology*, 194(3):729–729, January 2012.

[152] Jeong-Gu Kim, Daeui Park, Byoung-Chul Kim, Seong-Woong Cho, Yeong T. Kim, Young-Jin Park, Hee J. Cho, Hyunseok Park, Ki-Bong Kim, Kyong-Oh Yoon, Soo-Jun Park, Byoung-Moo Lee, and Jong Bhak. Predicting the Interactome of Xanthomonas oryzae pathovar oryzae for target selection and DB service. *BMC Bioinformatics*, 9(1):41, January 2008.

[153] Veljo Kisand and Teresa Lettieri. Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. *BMC Genomics*, 14(1):211, April 2013.

[154] Kiviat NB, Peterson M, Kinney-Thomas E, Tam M, Stamm WE, and Holmes KK. Cytologic manifestations of cervical and vaginal infections: Ii. confirmation of chlamydia trachomatis infection by direct immunofluorescence using monoclonal antibodies. *JAMA*, 253(7):997–1000, February 1985.

[155] Aileen F. Knowles. The GDA1_cd39 superfamily: NTPDases with diverse functions. *Purinergic Signalling*, 7(1):21–45, March 2011.

[156] Yoshihiro Kobae, Tetsuro Sekino, Hirofumi Yoshioka, Tsuyoshi Nakagawa, Enrico Martinoia, and Masayoshi Maeshima. Loss of AtPDR8, a Plasma Membrane ABC Transporter of Arabidopsis thaliana, Causes Hypersensitive Cell Death Upon Pathogen Infection. *Plant and Cell Physiology*, 47(3):309–318, January 2006.

[157] Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 2001.

[158] Konstantin V. Korotkov, Maria Sandkvist, and Wim G. J. Hol. The type II secretion system: biogenesis, molecular architecture and mechanism. *Nature Reviews Microbiology*, 10(5):336–351, May 2012.

[159] Erwin Kreyszig. *Advanced Engineering Mathematics*. John Wiley & Sons, December 2010.

[160] O. Krishnadev and N. Srinivasan. Prediction of protein–protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *International Journal of Biological Macromolecules*, 48(4):613–619, May 2011.

[161] Yadunanda Kumar and Raphael H. Valdivia. Reorganization of the host cytoskeleton by the intracellular pathogen Chlamydia trachomatis. *Communicative & Integrative Biology*, 1(2):175–177, 2008.

[162] B. Josh Lane, Charla Mutchler, Souhaila Al Khodor, Scott S. Grieshaber, and Rey A. Carabeo. Chlamydial entry involves TARP binding of guanine nucleotide exchange factors. *PLoS pathogens*, 4(3):e1000014, March 2008.

[163] Nicolas Lapaque, Ignacio Moriyon, Edgardo Moreno, and Jean-Pierre Gorvel. Brucella lipopolysaccharide acts as a virulence factor. *Current Opinion in Microbiology*, 8(1):60–66, February 2005.

[164] Pedro Larranaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Inaki Inza, Jose A. Lozano, Ruben Armananzas, Guzman Santafe, Aritz Perez, and Victor Robles. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, January 2006.

[165] Gaelle Le Negrate, Andreas Krieg, Benjamin Faustin, Markus Loeffler, Adam Godzik, Stan Krajewski, and John C. Reed. ChlaDub1 of Chlamydia trachomatis suppresses NF-kappaB activation and inhibits IkappaBalpha ubiquitination and degradation. *Cell Microbiol*, 10(9):1879–1892, September 2008.

[166] Heeseok Lee. Justifying database normalization: a cost/benefit model. *Information Processing & Management*, 31(1):59–67, January 1995.

[167] Lutz Eric Lehmann, Klaus-Peter Hunfeld, Martina Steinbrucker, Volker Brade, Malte Book, Harald Seifert, Tobias Bingold, Andreas Hoeft, Heimo Wissing, and Frank Stueber. Improved detection of blood stream pathogens by real-time PCR in severe sepsis. *Intensive Care Medicine*, 36(1):49–56, January 2010.

[168] Lei Lei, Manli Qi, Nicole Budrys, Robert Schenken, and Guangming Zhong. Localization of Chlamydia trachomatis hypothetical protein CT311 in host cell cytoplasm. *Microbial Pathogenesis*, 51(3):101–109, September 2011.

[169] Rudolf M. Lequin. Enzyme Immunoassay (EIA)/Enzyme-Linked Immunosorbent Assay (ELISA). *Clinical Chemistry*, 51(12):2415–2418, January 2005.

[170] Wen-Hsiung Li, Zhenglong Gu, Haidong Wang, and Anton Nekrutenko. Evolutionary analyses of the human genome. *Nature*, 409(6822):847–849, February 2001.

[171] Ziv Lifshitz, David Burstein, Michael Peeri, Tal Zusman, Kierstyn Schwartz, Howard A. Shuman, Tal Pupko, and Gil Segal. Computational modeling and experimental validation of the Legionella and Coxiella virulence-related type-IVB secretion signal. *Proceedings of the National Academy of Sciences*, 110(8):E707–E715, February 2013.

[172] Clare L. Ling and Timothy D. McHugh. Rapid Detection of Atypical Respiratory Bacterial Pathogens by Real-Time PCR. In Mark Wilks, editor, *PCR Detection of Microbial Pathogens*, number 943 in Methods in Molecular Biology, pages 125–133. Humana Press, January 2013.

[173] W. Ian Lipkin. Microbe Hunting. *Microbiology and Molecular Biology Reviews*, 74(3):363–377, January 2010.

[174] Yancheng Liu and Zhao-Qing Luo. The Legionella pneumophila Effector SidJ Is Required for Efficient Recruitment of Endoplasmic Reticulum Proteins to the Bacterial Phagosome. *Infection and Immunity*, 75(2):592–603, February 2007.

[175] M Lomma, D Dervins-Ravault, M Rolando, T Nora, HJ Newton, FM Sansom, T Sahr, L Gomez-Valero, M Jules, EL Hartland, et al. The legionella pneumophila f-box protein lpp2082 (ankb) modulates ubiquitination of the host protein parvin b and promotes intracellular replication. *Cellular microbiology*, 12(9):1272–1291, 2010.

[176] Martin Löwer and Gisbert Schneider. Prediction of type III secretion signals in genomes of gram-negative bacteria. *PloS one*, 4(6):e5917, 2009.

[177] Chunxue Lu, Lei Lei, Bo Peng, Lingli Tang, Honglei Ding, Siqi Gong, Zhongyu Li, Yimou Wu, and Guangming Zhong. Chlamydia trachomatis GlgA Is Secreted into Host Cell Cytoplasm. *PLoS ONE*, 8(7):e68764, July 2013.

[178] Qibin Luo, Philipp Pagel, Baiba Vilne, and Dmitrij Frishman. DIMA 3.0: Domain Interaction Map. *Nucleic Acids Research*, 39(suppl 1):D724–D729, January 2011.

[179] Mor N Lurie-Weinberger, Laura Gomez-Valero, Nathalie Merault, Gernot Glöckner, Carmen Buchrieser, and Uri Gophna. The origins of eukaryotic-like proteins in legionella pneumophila. *International Journal of Medical Microbiology*, 300(7):470–481, 2010.

[180] Mor N. Lurie-Weinberger, Laura Gomez-Valero, Nathalie Merault, Gernot Gloeckner, Carmen Buchrieser, and Uri Gophna. The origins of eukaryotic-like proteins in Legionella pneumophila. *International Journal of Medical Microbiology*, 300(7):470–481, November 2010.

[181] Norman J. MacDonald and Robert G. Beiko. Efficient learning of microbial genotype–phenotype association rules. *Bioinformatics*, 26(15):1834–1840, January 2010.

[182] Jose Paolo V. Magbanua, Beng Tin Goh, Claude-Edouard Michel, Aura Aguirre-Andreasen, Sarah Alexander, Ines Ushiro-Lumb, Catherine Ison, and Helen Lee. Chlamydia trachomatis variant not detected by plasmid based nucleic acid amplification tests: molecular characterisation and failure of single dose azithromycin. *Sexually Transmitted Infections*, 83(4):339–343, January 2007.

[183] Amanda Nga-Sze Mak, Yuen-Ting Wong, Young-Jun An, Sun-Shin Cha, Kong-Hung Sze, Shannon Wing-Ngor Au, Kam-Bo Wong, and Pang-Chui Shaw. Structure-function study of maize ribosome-inactivating protein: implications for the internal inactivation region and the sole glutamate in the active site. *Nucleic Acids Research*, 35(18):6259–6267, January 2007.

[184] Seema Mattoo, Yvonne M Lee, and Jack E Dixon. Interactions of bacterial effector proteins with host proteins. *Current Opinion in Immunology*, 19(4):392–401, August 2007.

[185] Annick Mayor, Fabio Martinon, Thibaut De Smedt, Virginie Petrilli, and Juerg Tschopp. A crucial function of SGT1 and HSP90 in inflammasome activity links mammalian and plant innate immune responses. *Nature Immunology*, 8(5):497–503, May 2007.

[186] Jason E McDermott, Abigail Corrigan, Elena Peterson, Christopher Oehmen, George Niemann, Eric D Cambronne, Danna Sharp, Joshua N Adkins, Ram Samu-

drala, and Fred Heffron. Computational prediction of type III and IV secreted effectors in gram-negative bacteria. *Infection and immunity*, 79(1):23–32, 2011.

[187] Vesna Memisevic, Nela Zavaljevski, Rembert Pieper, Seesandra V. Rajagopala, Keehwan Kwon, Katherine Townsend, Chenggang Yu, Xueping Yu, David DeShazer, Jaques Reifman, and Anders Wallqvist. Novel Burkholderia mallei Virulence Factors Linked to Specific Host-Pathogen Protein Interactions. *Molecular & Cellular Proteomics*, 12(11):3036–3051, January 2013.

[188] C. Miller, J. Gurd, and A. Brass. A RAPID algorithm for sequence database comparisons: application to the identification of vector contamination in the EMBL databases. *Bioinformatics*, 15(2):111–121, January 1999.

[189] Shahram Misaghi, Zarine R. Balsara, Andre Catic, Eric Spooner, Hidde L. Ploegh, and Michael N. Starnbach. Chlamydia trachomatis-derived deubiquitinating enzymes in mammalian cells during infection. *Molecular Microbiology*, 61(1):142–150, 2006.

[190] Sarah T. Miyata, Maya Kitaoka, Teresa M. Brooks, Steven B. McAuley, and Stefan Pukatzki. Vibrio cholerae Requires the Type VI Secretion System Virulence Factor VasX To Kill Dictyostelium discoideum. *Infection and Immunity*, 79(7):2941–2949, January 2011.

[191] Denise M Monack. Salmonella persistence and transmission strategies. *Current Opinion in Microbiology*, 15(1):100–107, February 2012.

[192] David M Morens, Gregory K Folkers, and Anthony S Fauci. The challenge of emerging and re-emerging infectious diseases. *Nature*, 430(6996):242–249, July 2004.

[193] K. E. Mueller, G. V. Plano, and K. A. Fields. New Frontiers in Type III Secretion Biology: the Chlamydia Perspective. *Infection and Immunity*, 82(1):2–9, January 2014.

[194] Takafumi Mukaihara, Naoyuki Tamura, and Masaki Iwabuchi. Genome-Wide Identification of a Large Repertoire of Ralstonia solanacearum Type III Effector Proteins by a New Functional Screen. *Molecular Plant-Microbe Interactions*, 23(3):251–262, February 2010.

[195] M. Shahid Mukhtar, Anne-Ruxandra Carvunis, Matija Dreze, Petra Epple, Jens Steinbrenner, Jonathan Moore, Murat Tasan, Mary Galli, Tong Hao, Marc T. Nishimura, Samuel J. Pevzner, Susan E. Donovan, Lila Ghamsari, Balaji Santhanam, Viviana Romero, Matthew M. Poulin, Fana Gebreab, Bryan J. Gutierrez, Stanley Tam, Dario Monachello, Mike Boxem, Christopher J. Harbort, Nathan McDonald, Lantian

Gai, Huaming Chen, Yijian He, Jean Vandenhaute, Frederick P. Roth, David E. Hill, Joseph R. Ecker, Marc Vidal, Jim Beynon, Pascal Braun, and Jeffery L. Dangl. Independently Evolved Virulence Effectors Converge onto Hubs in a Plant Immune System Network. *Science*, 333(6042):596–601, July 2011.

[196] J. Muller, D. Szklarczyk, P. Julien, I. Letunic, A. Roth, M. Kuhn, S. Powell, C. von Mering, T. Doerks, L. J. Jensen, and P. Bork. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Research*, 38(suppl 1):D190–D195, January 2010.

[197] Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, April 1995.

[198] Sandra Muschiol, Gaelle Boncompain, François Vromman, Pierre Dehoux, Staffan Normark, Birgitta Henriques-Normark, and Agathe Subtil. Identification of a Family of Effectors Secreted by the Type III Secretion System That Are Conserved in Pathogenic Chlamydiae. *Infection and Immunity*, 79(2):571–580, January 2011.

[199] Hiroki Nagai and Tomoko Kubori. Purification and Characterization of Legionella U-Box-Type E3 Ubiquitin Ligase. In Carmen Buchrieser and Hubert Hilbi, editors, *Legionella*, number 954 in Methods in Molecular Biology, pages 347–354. Humana Press, January 2013.

[200] H. Naora. Involvement of ribosomal proteins in regulating cell growth and apoptosis: translational modulation or recruitment for extraribosomal activity? *Immunol Cell Biol*, 77(3):197–205, June 1999.

[201] Paolo Natale, Thomas Brueser, and Arnold J. M. Driessen. Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane–distinct translocases and mechanisms. *Biochimica Et Biophysica Acta*, 1778(9):1735–1756, September 2008.

[202] George S. Niemann, Roslyn N. Brown, Jean K. Gustin, Afke Stufkens, Afshan S. Shaikh-Kidwai, Jie Li, Jason E. McDermott, Heather M. Brewer, Athena Schepmoes, Richard D. Smith, Joshua N. Adkins, and Fred Heffron. Discovery of Novel Secreted Virulence Factors from Salmonella enterica Serovar Typhimurium by Proteomic Analysis of Culture Supernatants. *Infection and Immunity*, 79(1):33–43, January 2011.

[203] Krzysztof Niemczuk, Marian Truszczyńsk, and Monika Szymańska-Czerwińska. Chlamydiales – Taxonomy, Pathogenicity, and Zoonotic Potential. *Bulletin of the Veterinary Institute in Pulawy*, 56(3):267–270, 2013.

[204] Shira Ninio and Craig R. Roy. Effector proteins translocated by Legionella pneumophila: strength in numbers. *Trends in Microbiology*, 15(8):372–380, August 2007.

[205] Tamara Nora, Mariella Lomma, Laura Gomez-Valero, and Carmen Buchrieser. Molecular mimicry: an important virulence strategy employed by legionella pneumophila to subvert host functions. *Future Microbiology*, 4(6):691–701, 2009.

[206] Patrice Nordmann, Laurent Poirel, Mark A. Toleman, and Timothy R. Walsh. Does broad-spectrum beta-lactam resistance due to NDM-1 herald the end of the antibiotic era for treatment of infections caused by Gram-negative bacteria? *Journal of Antimicrobial Chemotherapy*, 66(4):689–692, January 2011.

[207] Alexandra Nunes, Maria J. Borrego, and João P. Gomes. Genomic features beyond Chlamydia trachomatis phenotypes: What do we think we know? *Infection, Genetics and Evolution*, 16:392–400, June 2013.

[208] Howard Ochman and Nancy A. Moran. Genes Lost and Genes Found: Evolution of Bacterial Pathogenesis and Symbiosis. *Science*, 292(5519):1096–1099, November 2001.

[209] Markus Ojala and Gemma C Garriga. Permutation tests for studying classifier performance. *The Journal of Machine Learning Research*, 99:1833–1863, 2010.

[210] Kim Orth, Zhaohui Xu, Mary Beth Mudgett, Zhao Qin Bao, Lance E. Palmer, James B. Bliska, Walter F. Mangel, Brian Staskawicz, and Jack E. Dixon. Disruption of Signaling by Yersinia Effector YopJ, a Ubiquitin-Like Protein Protease. *Science*, 290(5496):1594–1597, November 2000.

[211] Jorma Paavonen. Chlamydia trachomatis infections of the female genital tract: State of the art. *Annals of Medicine*, 44(1):18–28, February 2011.

[212] Ioanna Pagani, Konstantinos Liolios, Jakob Jansson, I-Min A. Chen, Tatyana Smirnova, Bahador Nosrat, Victor M. Markowitz, and Nikos C. Kyrpides. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 40(D1):D571–D579, January 2012.

[213] Philipp Pagel, Philip Wong, and Dmitrij Frishman. A Domain Interaction Map Based on Phylogenetic Profiling. *Journal of Molecular Biology*, 344(5):1331–1346, December 2004.

[214] X. Pan, A. Luhrmann, A. Satoh, M. A. Laskowski-Arce, and C. R. Roy. Ankyrin repeat proteins comprise a diverse family of bacterial type IV effectors. *Science*, 320(5883):1651–4, 2008.

[215] M.-È Paradis, D. Haine, B. Gillespie, S. P. Oliver, S. Messier, J. Comeau, and D. T. Scholl. Bayesian estimation of the diagnostic accuracy of a multiplex real-time PCR assay and bacteriological culture for 4 common bovine intramammary pathogens. *Journal of Dairy Science*, 95(11):6436–6448, November 2012.

[216] Kaustubh R. Patil, Peter Haider, Phillip B. Pope, Peter J. Turnbaugh, Mark Morrison, Tobias Scheffer, and Alice C. McHardy. Taxonomic metagenome sequence assignment with structured output models. *Nature Methods*, 8(3):191–192, March 2011.

[217] Ian T. Paulsen, Rekha Seshadri, Karen E. Nelson, Jonathan A. Eisen, John F. Heidelberg, Timothy D. Read, Robert J. Dodson, Lowell Umayam, Lauren M. Brinkac, Maureen J. Beanan, Sean C. Daugherty, Robert T. Deboy, A. Scott Durkin, James F. Kolonay, Ramana Madupu, William C. Nelson, Bola Ayodeji, Margaret Kraul, Jyoti Shetty, Joel Malek, Susan E. Van Aken, Steven Riedmuller, Herve Tettelin, Steven R. Gill, Owen White, Steven L. Salzberg, David L. Hoover, Luther E. Lindler, Shirley M. Halling, Stephen M. Boyle, and Claire M. Fraser. The Brucella suis genome reveals fundamental similarities between animal and plant pathogens and symbionts. *Proceedings of the National Academy of Sciences*, 99(20):13148–13153, January 2002.

[218] Anna Sofie Pedersen, Gunna Christiansen, and Svend Birkelund. Differential expression of Pmp10 in cell culture infected with Chlamydia pneumoniae CWL029. *FEMS Microbiology Letters*, 203(2):153–159, September 2001.

[219] Meghan E. Pennini, Stephanie Perrinet, Alice Dautry-Varsat, and Agathe Subtil. Histone Methylation by NUE, a Novel Nuclear Effector of the Intracellular Pathogen Chlamydia trachomatis. *PLoS Pathog*, 6(7):e1000995, July 2010.

[220] Thomas Penz, Matthias Horn, and Stephan Schmitz-Esser. The genome of the amoeba symbiont "Candidatus Amoebophilus asiaticus" encodes an afp-like prophage possibly used for protein secretion. *Virulence*, 1(6):541–545, December 2010.

[221] Vicente Perez-Brocal, Amparo Latorre, and Andres Moya. Symbionts and Pathogens: What is the Difference? In Ulrich Dobrindt, Joerg H. Hacker, and Catharina Svanborg, editors, *Between Pathogenicity and Commensalism*, number 358 in Current Topics in Microbiology and Immunology, pages 215–243. Springer Berlin Heidelberg, January 2013.

[222] Francisco Perez-Montano, Irene Jimenez-Guerrero, Rocio Contreras Sanchez-Matamoros, Francisco Javier Lopez-Baena, Francisco Javier Ollero, Miguel A. Rodriguez-Carvajal, Ramon A. Bellogin, and M. Rosario Espuny. Rice and bean AHL-mimic quorum-sensing signals specifically interfere with the capacity to form biofilms

by plant-associated bacteria. *Research in Microbiology*, 164(7):749–760, September 2013.

[223] L. L. Peters, K. M. John, F. M. Lu, E. M. Eicher, A. Higgins, M. Yialamas, L. C. Turtzo, A. J. Otsuka, and S. E. Lux. Ank3 (epithelial ankyrin), a widely distributed new member of the ankyrin gene family and the major ankyrin in kidney, is expressed in alternatively spliced forms, including forms that lack the repeat domain. *The Journal of Cell Biology*, 130(2):313–330, July 1995.

[224] Thomas Nordahl Petersen, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):785–786, 2011.

[225] Thomas Nordahl Petersen, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):785–786, 2011.

[226] Fredrik Ponten, F.n, Marcus Gry, Linn Fagerberg, Emma Lundberg, Anna Asplund, Lisa Berglund, Per Oksvold, Erik Bjoerling, Sophia Hober, Caroline Kampf, Sanjay Navani, Peter Nilsson, Jenny Ottosson, Anja Persson, Henrik Wernerus, H.rus, Kenneth Wester, and Mathias Uhlen, M.n. A global view of protein expression in human cells, tissues, and organs. *Mol Syst Biol*, 5:337, 2009.

[227] C. P. Ponting, L. Aravind, J. Schultz, P. Bork, and E. V. Koonin. Eukaryotic Signalling Domain Homologues in Archaea and Bacteria. Ancient Ancestry and Horizontal Gene Transfer. *Journal of Molecular Biology*, 289(4):729–745, June 1999.

[228] Christopher T. D. Price, Tasneem Al-Quadan, Marina Santic, Snake C. Jones, and Yousef Abu Kwaik. Exploitation of conserved eukaryotic host cell farnesylation machinery by an F-box effector of Legionella pneumophila. *The Journal of Experimental Medicine*, 207(8):1713–1726, February 2010.

[229] Kim D. Pruitt, Tatiana Tatusova, Garth R. Brown, and Donna R. Maglott. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*, 40(D1):D130–D135, January 2012.

[230] Marco Punta, Penny C Coggill, Ruth Y Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, et al. The pfam protein families database. *Nucleic acids research*, 40(D1):D290–D301, 2012.

[231] Manli Qi, Lei Lei, Siqi Gong, Quanzhong Liu, Matthew P. DeLisa, and Guangming Zhong. Chlamydia trachomatis Secretion of an Immunodominant Hypothetical Protein

(CT795) into Host Cell Cytoplasm. *Journal of Bacteriology*, 193(10):2498–2509, May 2011.

[232] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013.

[233] Julia Radics, Lisa Koenigsmaier, and Thomas C. Marlovits. Structure of a pathogenic type 3 secretion system in action. *Nature Structural & Molecular Biology*, 21(1):82–87, January 2014.

[234] Laurence G. Rahme, Frederick M. Ausubel, Hui Cao, Eliana Drenkard, Boyan C. Goumnerov, Gee W. Lau, Shalina Mahajan-Miklos, Julia Plotnikova, Man-Wah Tan, John Tsongalis, Cynthia L. Walendziewicz, and Ronald G. Tompkins. Plants and animals share functionally common bacterial virulence factors. *Proceedings of the National Academy of Sciences*, 97(16):8815–8821, January 2000.

[235] Thomas Rattei, Patrick Tischler, Stefan Goetz, Marc-Andre Jehl, Jonathan Hoser, Roland Arnold, Ana Conesa, and Hans-Werner Mewes. SIMAP—a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Research*, 38(suppl 1):D223–D226, January 2010.

[236] Wolf Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425–432, May 2007.

[237] David A. Relman. The Search for Unrecognized Pathogens. *Science*, 284(5418):1308–1310, May 1999.

[238] Relman DA. Metagenomics, infectious disease diagnostics, and outbreak investigations: Sequence first, ask questions later? *JAMA*, 309(14):1531–1532, April 2013.

[239] Robert Riley, Christopher Lee, Chiara Sabatti, and David Eisenberg. Inferring protein domain interactions from databases of interacting proteins. *Genome Biology*, 6(10):R89, September 2005.

[240] Daniel D. Rockey, Marci A. Scidmore, John P. Bannantine, and Wendy J. Brown. Proteins in the chlamydial inclusion membrane. *Microbes and Infection*, 4(3):333–340, March 2002.

[241] Monica Rolando and Carmen Buchrieser. Post-translational modifications of host proteins by legionella pneumophila: a sophisticated survival strategy. *Future microbiology*, 7(3):369–381, 2012.

[242] Bettina Rosner. Epidemiologie des EHEC O104:H4/HUS-Ausbruchs in Deutschland, Mai bis Juli 2011. *Journal fuer Verbraucherschutz und Lebensmittelsicherheit*, December 2011.

[243] Jonathan D. C. Ross. Pelvic inflammatory disease. *Medicine*, 42(6):333–337, June 2014.

[244] Shlomo Rottem. Interaction of mycoplasmas with host cells. *Physiological reviews*, 83(2):417–432, 2003.

[245] Ingrid G. I. J. G. Rours, Margaret R. Hammerschlag, Alewijn Ott, Tjeerd J. T. H. N. De Faber, Henri A. Verbrugh, Ronald de Groot, and Roel P. Verkooyen. Chlamydia trachomatis as a Cause of Neonatal Conjunctivitis in Dutch Infants. *Pediatrics*, 121(2):e321–e326, January 2008.

[246] K. A. Rzomp, A. R. Moorhead, and M. A. Scidmore. The GTPase Rab4 Interacts with Chlamydia trachomatis Inclusion Membrane Protein CT229. *Infection and Immunity*, 74(9):5362–5373, January 2006.

[247] Dor Salomon, Lisa N. Kinch, David C. Trudgian, Xiaofeng Guo, John A. Klimko, Nick V. Grishin, Hamid Mirzaei, and Kim Orth. Marker for type VI secretion system effectors. *Proceedings of the National Academy of Sciences*, 111(25):9271–9276, June 2014.

[248] R. Samudrala, F. Heffron, and J. E. McDermott. Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems. *PLoS Pathog*, 5(4):e1000375, 2009.

[249] Ram Samudrala, Fred Heffron, and Jason E McDermott. Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems. *PLoS pathogens*, 5(4):e1000375, 2009.

[250] Sara F. Sarkar, Jeffrey S. Gordon, Gregory B. Martin, and David S. Guttman. Comparative Genomics of Host-Specific Virulence in Pseudomonas syringae. *Genetics*, 174(2):1041–1056, January 2006.

[251] J. H. Scheibel, J. T. Andersen, P. Brandenhoff, J. P. Geerdsen, A. Bay-Nielsen, B. A. Schultz, and S. Walter. Chlamydia trachomatis in acute epididymitis. *Scandinavian Journal of Urology and Nephrology*, 17(1):47–50, 1983.

[252] Kurt Schesser, Ann-Kristin Spiik, Jean-Marie Dukuzumuremyi, Markus F. Neurath, Sven Pettersson, and Hans Wolf-Watz. The yopJ locus is required for Yersinia-mediated inhibition of NF-kappaB activation and cytokine expression: YopJ contains

a eukaryotic SH2-like domain that is essential for its repressive activity. *Molecular Microbiology*, 28(6):1067–1079, June 1998.

[253] Stephan Schmitz-Esser, Patrick Tischler, Roland Arnold, Jacqueline Montanaro, Michael Wagner, Thomas Rattei, and Matthias Horn. The Genome of the Amoeba Symbiont "Candidatus Amoebophilus asiaticus" Reveals Common Mechanisms for Host Cell Interaction among Amoeba-Associated Bacteria. *Journal of Bacteriology*, 192(4):1045–1057, February 2010.

[254] Gunnar N. Schroeder, Nicola K. Petty, Aurelie Mousnier, Clare R. Harding, Adam J. Vogrin, Bryan Wee, Norman K. Fry, Timothy G. Harrison, Hayley J. Newton, Nicholas R. Thomson, Scott A. Beatson, Gordon Dougan, Elizabeth L. Hartland, and Gad Frankel. Legionella pneumophila Strain 130b Possesses a Unique Combination of Type IV Secretion Systems and Novel Dot/Icm Secretion System Effector Proteins. *Journal of Bacteriology*, 192(22):6001–6016, November 2010.

[255] M. A. Scidmore and T. Hackstadt. Mammalian 14-3-3beta associates with the Chlamydia trachomatis inclusion membrane via its interaction with IncG. *Molecular Microbiology*, 39(6):1638–1650, March 2001.

[256] Marci A. Scidmore. Recent Advances in Chlamydia Subversion of Host Cytoskeletal and Membrane Trafficking Pathways. *Microbes and infection / Institut Pasteur*, 13(6):527–535, June 2011.

[257] G. A. Seluja, A. Farmer, M. McLeod, C. Harger, and P. A. Schad. Establishing a method of vector contamination identification in database sequences. *Bioinformatics*, 15(2):106–110, January 1999.

[258] Stephanie R Shames, Sigrid D Auweter, and B Brett Finlay. Co-evolution and exploitation of host cell signaling pathways by bacterial pathogens. *The International Journal of Biochemistry & Cell Biology*, 41(2):380–389, 2009.

[259] Sagi D. Shapira, Irit Gat-Viks, Bennett O.V. Shum, Amelie Dricot, Marciela M. Degrace, Wu Liguo, Piyush B. Gupta, Tong Hao, Serena J. Silver, David E. Root, David E. Hill, Aviv Regev, and Nir Hacohen. A physical and regulatory map of host-influenza interactions reveals pathways in H1n1 infection. *Cell*, 139(7):1255–1267, December 2009.

[260] Divine Y. Shyntum, Stephanus N. Venter, Lucy N. Moleleki, Ian Toth, and Teresa A. Coutinho. Comparative genomics of type VI secretion systems in strains of Pantoea ananatis from different environments. *BMC Genomics*, 15(1):163, February 2014.

[261] Julie M. Silverman, Danielle M. Agnello, Hongjin Zheng, Benjamin T. Andrews, Mo Li, Carlos E. Catalano, Tamir Gonen, and Joseph D. Mougous. Haemolysin Coregulated Protein Is an Exported Receptor and Chaperone of Type VI Secretion Substrates. *Molecular Cell*, 51(5):584–593, September 2013.

[262] Roxane Simeone, Daria Bottai, and Roland Brosch. ESX/type VII secretion systems and their role in host–pathogen interaction. *Current Opinion in Microbiology*, 12(1):4–10, February 2009.

[263] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, October 2005.

[264] Amit Singh, Somayyeh Poshtiban, and Stephane Evoy. Recent Advances in Bacteriophage Based Biosensors for Food-Borne Pathogen Detection. *Sensors*, 13(2):1763–1786, January 2013.

[265] Noam Slonim, Olivier Elemento, and Saeed Tavazoie. Ab initio genotype–phenotype association reveals intrinsic modularity in genetic networks. *Molecular Systems Biology*, 2(1):n/a–n/a, 2006.

[266] Pawel Smialowski, Philipp Pagel, Philip Wong, Barbara Brauner, Irmtraud Dunger, Gisela Fobo, Goar Frishman, Corinna Montrone, Thomas Rattei, Dmitrij Frishman, and Andreas Ruepp. The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Research*, 38(suppl 1):D540–D544, January 2010.

[267] Emily A. Snavely, Marcela Kokes, Joe Dan Dunn, Hector A. Saka, Bidong D. Nguyen, Robert J. Bastidas, Dewey G. McCafferty, and Raphael H. Valdivia. Reassessing the role of the secreted protease CPAF in Chlamydia trachomatis infection through genetic approaches. *Pathogens and Disease*, pages n/a–n/a, May 2014.

[268] Marie-Paule Sory and Guy R. Cornelis. Translocation of a hybrid YopE-adenylate cyclase from Yersinia enterocolitica into HeLa cells. *Molecular Microbiology*, 14(3):583–594, November 1994.

[269] Richard C. Sprinthall. *Basic statistical analysis*. Prentice Hall, 1990.

[270] Lola V Stamm and Benjamin Mudrak. Old foes, new challenges: syphilis, cholera and TB. *Future Microbiology*, 8(2):177–189, February 2013.

[271] Michael N. Starnbach, Wendy P. Loomis, Pam Ovendale, David Regan, Bruce Hess, Mark R. Alderson, and Steven P. Fling. An Inclusion Membrane Protein from Chlamydia trachomatis Enters the MHC Class I Pathway and Stimulates a CD8+ T Cell Response. *The Journal of Immunology*, 171(9):4742–4749, January 2003.

[272] John Stavrinides, Wenbo Ma, and David S Guttman. Terminal Reassortment Drives the Quantum Evolution of Type III Effectors in Bacterial Pathogens. *PLoS Pathog*, 2(10):e104, October 2006.

[273] John Stavrinides, Honour C. McCann, and David S. Guttman. Host–pathogen interplay and the evolution of bacterial effectors. *Cellular Microbiology*, 10(2):285–292, February 2008.

[274] Amelie Stein, Arnaud Ceol, and Patrick Aloy. 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, 39(suppl 1):D718–D723, January 2011.

[275] M Steinert, U Hentschel, and J Hacker. Symbiosis and pathogenesis: evolution of the microbe-host interaction. *Die Naturwissenschaften*, 87(1):1–11, January 2000.

[276] Richard S. Stephens, Sue Kalman, Claudia Lammel, Jun Fan, Rekha Marathe, L. Aravind, Wayne Mitchell, Lynn Olinger, Roman L. Tatusov, Qixun Zhao, Eugene V. Koonin, and Ronald W. Davis. Genome Sequence of an Obligate Intracellular Pathogen of Humans: Chlamydia trachomatis. *Science*, 282(5389):754–759, October 1998.

[277] F Stirpe and M G Battelli. Ribosome-inactivating proteins: progress and problems. *Cellular and molecular life sciences: CMLS*, 63(16):1850–1866, August 2006.

[278] Agathe Subtil, Cedric Delevoye, Maria-Eugenia Balana, Laurence Tastevin, Stephanie Perrinet, and Alice Dautry-Varsat. A directed screen for chlamydial proteins secreted by a type III mechanism identifies a translocated protein and numerous other new candidates. *Molecular Microbiology*, 56(6):1636–1647, June 2005.

[279] Agathe Subtil, Claude Parsot, and Alice Dautry-Varsat. Secretion of predicted Inc proteins of Chlamydia pneumoniae by a heterologous type III machinery. *Molecular Microbiology*, 39(3):792–800, February 2001.

[280] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(suppl 1):D561–D568, January 2011.

[281] Kosaku Takahashi, Koji Kasai, and Kozo Ochi. Identification of the bacterial alarmone guanosine 5-diphosphate 3-diphosphate (ppGpp) in plants. *Proceedings of the National Academy of Sciences*, 101(12):4320–4324, March 2004.

[282] M. Takeuchi, F. Kawai, Y. Shimada, and A. Yokota. Taxonomic Study of Polyethylene Glycol-Utilizing Bacteria: Emended Description of the Genus Sphingomonas and

New Descriptions of Sphingomonas macrogoltabidus sp. nov., Sphingomonas sanguis sp. nov. and Sphingomonas terrae sp. nov. *Systematic and Applied Microbiology*, 16(2):227–238, July 1993.

[283] Herve Tettelin, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, Jonathan Crabtree, Amanda L. Jones, A. Scott Durkin, Robert T. DeBoy, Tanja M. Davidsen, Marirosa Mora, Maria Scarselli, Immaculada Margarit y Ros, Jeremy D. Peterson, Christopher R. Hauser, Jaideep P. Sundaram, William C. Nelson, Ramana Madupu, Lauren M. Brinkac, Robert J. Dodson, Mary J. Rosovitz, Steven A. Sullivan, Sean C. Daugherty, Daniel H. Haft, Jeremy Selengut, Michelle L. Gwinn, Liwei Zhou, Nikhat Zafar, Hoda Khouri, Diana Radune, George Dimitrov, Kisha Watkins, Kevin J. B. O'Connor, Shannon Smith, Teresa R. Utterback, Owen White, Craig E. Rubens, Guido Grandi, Lawrence C. Madoff, Dennis L. Kasper, John L. Telford, Michael R. Wessels, Rino Rappuoli, and Claire M. Fraser. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–13955, September 2005.

[284] Jessica Thalmann, Katrin Janik, Martin May, Kirsten Sommer, Jenny Ebeling, Fred Hofmann, Harald Genth, and Andreas Klos. Actin re-organization induced by Chlamydia trachomatis serovar D–evidence for a critical role of the effector protein CT166 targeting Rac. *PloS One*, 5(3):e9887, 2010.

[285] Nicholas R. Thomson, Matthew T. G. Holden, Caroline Carder, Nicola Lennard, Sarah J. Lockey, Pete Marsh, Paul Skipp, C. David O'Connor, Ian Goodhead, Halina Norbertzcak, Barbara Harris, Doug Ormond, Richard Rance, Michael A. Quail, Julian Parkhill, Richard S. Stephens, and Ian N. Clarke. Chlamydia trachomatis: Genome sequence analysis of lymphogranuloma venereum isolates. *Genome Research*, 18(1):161–171, January 2008.

[286] Andrew Camilli Tim van Opijnen. A fine scale phenotype-genotype virulence map of a bacterial pathogen. *Genome research*, 2012.

[287] J. Tjaden, H. H. Winkler, C. Schwoeppe, M. Van Der Laan, T. Moehlmann, and H. E. Neuhaus. Two Nucleotide Transport Proteins in Chlamydia trachomatis, One for Net Nucleoside Triphosphate Uptake and the Other for Transport of Energy. *Journal of Bacteriology*, 181(4):1196–1202, February 1999.

[288] Toru Tobe, Scott A. Beatson, Hisaaki Taniguchi, Hiroyuki Abe, Christopher M. Bailey, Amanda Fivian, Rasha Younis, Sophie Matthews, Olivier Marches, Gad Frankel, Tetsuya Hayashi, and Mark J. Pallen. An extensive repertoire of type III secretion effectors in Escherichia coli O157 and the role of lambdoid phages in their dissemination. *Proceedings of the National Academy of Sciences*, 103(40):14941–14946, March 2006.

[289] Seav-Ly Tran, Alan D. Billoud, and Stephanie Schueller. Shiga toxin production and translocation during microaerobic human colonic infection with shiga toxin-producing e.coli o157-h7 and o104-h4. *Cellular Microbiology*, 2014.

[290] Paul Troisfontaines and Guy R. Cornelis. Type III Secretion: More Systems Than You Think. *Physiology*, 20(5):326–339, October 2005.

[291] Jennifer E. Trosky, Amy D. B. Liverman, and Kim Orth. Yersinia outer proteins: Yops. *Cellular Microbiology*, 10(3):557–565, 2008.

[292] Tsai-Tien Tseng, Brett M. Tyler, and João C. Setubal. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiology*, 9(Suppl 1):S2, February 2009.

[293] Nidhi Tyagi, Oruganty Krishnadev, and Narayanaswamy Srinivasan. Prediction of protein–protein interactions between Helicobacter pylori and a human host. *Molecular BioSystems*, 5(12):1630–1635, November 2009.

[294] Simon Urwyler, Yves Nyfeler, Curdin Ragaz, Hookeun Lee, Lukas N. Mueller, Ruedi Aebersold, and Hubert Hilbi. Proteome Analysis of Legionella Vacuoles Purified by Magnetic Immunoseparation Reveals Secretory and Endosomal GTPases. *Traffic*, 10(1):76–87, January 2009.

[295] Paul V. Viitanen, Anthony A. Gatenby, and George H. Lorimer. Purified chaperonin 60 (groEL) interacts with the nonnative states of a multitude of Escherichia coli proteins. *Protein Science*, 1(3):363–369, March 1992.

[296] Robert P. A Wallin, Andreas Lundqvist, Solveig H More, Arne von Bonin, Rolf Kiessling, and Hans-Gustaf Ljunggren. Heat-shock proteins as activators of the innate immune system. *Trends in Immunology*, 23(3):130–135, March 2002.

[297] Mathias C. Walter, Thomas Rattei, Roland Arnold, Ulrich Gueldener, Martin Muensterkoetter, Karamfilka Nenova, Gabi Kastenmueller, Patrick Tischler, Andreas Woelling, Andreas Volz, Norbert Pongratz, Ralf Jost, Hans-Werner Mewes, and Dmitrij Frishman. PEDANT covers all complete RefSeq genomes. *Nucleic Acids Research*, 37(suppl 1):D408–D411, January 2009.

[298] Yejun Wang, Xiaowei Wei, Hongxia Bao, and Shu-Lin Liu. Prediction of bacterial type IV secreted effectors by C-terminal features. *BMC Genomics*, 15(1):50, January 2014.

[299] Yejun Wang, Qing Zhang, Ming-an Sun, and Dianjing Guo. High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics*, 27(6):777–784, 2011.

[300] Jonathan R. Warner and Kerri B. McIntosh. How common are extraribosomal functions of ribosomal proteins? *Mol Cell*, 34(1):3–11, April 2009.

[301] George M. Weinstock, John M. Hardham, Michael P. McLeod, Erica J. Sodergren, and Steven J. Norris. The genome of Treponema pallidum: new light on the agent of syphilis. *FEMS Microbiology Reviews*, 22(4):323–332, October 1998.

[302] Armand G. Ngounou Wetie, Izabela Sokolowska, Alisa G. Woods, Urmi Roy, Katrin Deinhardt, and Costel C. Darie. Protein–protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cellular and Molecular Life Sciences*, 71(2):205–228, January 2014.

[303] David L. Wheeler, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Lewis Y. Geer, Yuri Kapustin, Oleg Khovayko, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, James Ostell, Vadim Miller, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Steven T. Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Roman L. Tatusov, Tatiana A. Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 35(suppl 1):D5–D12, January 2007.

[304] Owen White, Ted Dunning, Granger Sutton, Mark Adams, J. Craig Venter, and Chris Fields. A quality control algorithm for DNA sequencing projects. *Nucleic Acids Research*, 21(16):3829–3838, November 1993.

[305] Andrew Whitehead and Douglas L. Crawford. Variation in tissue-specific gene expression among natural populations. *Genome Biol*, 6(2):R13, 2005.

[306] World Health Organisation (WHO). *World Health Statistics 2012*.

[307] Marcus Wieder. Bacterial effector proteins in the evolution of pathogenic and symbiotic bacteria. 2013.

[308] Rainer Winnenburg, Martin Urban, Andrew Beacham, Thomas K. Baldwin, Sabrina Holland, Magdalen Lindeberg, Hilde Hansen, Christopher Rawlings, Kim E.

Hammond-Kosack, and Jacob Koehler. PHI-base update: additions to the pathogen–host interaction database. *Nucleic Acids Research*, 36(suppl 1):D572–D576, January 2008.

[309] Karl Wooldridge. *Bacterial secreted proteins: secretory mechanisms and role in pathogenesis*. The Publisher, 2009.

[310] Brendan W. Wren. The Yersiniae — a model genus to study the rapid evolution of bacterial pathogens. *Nature Reviews Microbiology*, 1(1):55–64, October 2003.

[311] Xiang Wu, Lei Lei, Siqi Gong, Ding Chen, Rhonda Flores, and Guangming Zhong. The chlamydial periplasmic stress response serine protease cHtrA is secreted into host cell cytosol. *BMC Microbiology*, 11(1):87, April 2011.

[312] Stefan Wuchty. Computational Prediction of Host-Parasite Protein Interactions between P. falciparum and H. sapiens. *PLoS ONE*, 6(11):e26960, November 2011.

[313] Zuoshuang Xiang, Yuying Tian, and Yongqun He. PHIDIAS: a pathogen-host interaction data integration and analysis system. *Genome Biology*, 8(7):R150, 2007.

[314] Li Xu and Zhao-Qing Luo. Cell biology of infection by Legionella pneumophila. *Microbes and Infection*, 15(2):157–167, February 2013.

[315] Li Xu, Xihui Shen, Andrew Bryan, Simran Banga, Michele S. Swanson, and Zhao-Qing Luo. Inhibition of Host Vacuolar H+-ATPase Activity by a Legionella pneumophila Effector. *PLoS Pathog*, 6(3):e1000822, March 2010.

[316] Feng Xue, Jie Yan, and Mathieu Picardeau. Evolution and pathogenesis of Leptospira spp.: lessons learned from the genomes. *Microbes and Infection*, 11(3):328–333, March 2009.

[317] Haiyuan Yu, Nicholas M. Luscombe, Hao Xin Lu, Xiaowei Zhu, Yu Xia, Jing-Dong J. Han, Nicolas Bertin, Sambath Chung, Marc Vidal, and Mark Gerstein. Annotation Transfer Between Genomes: Protein–Protein Interologs and Protein–DNA Regulogs. *Genome Research*, 14(6):1107–1118, January 2004.

[318] Harry Zhang. The Optimality of Naive Bayes. AAAI Press, 2004.

[319] Harry Zhang. Exploring conditions for the optimality of naive bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(02):183–198, March 2005.

[320] Xing-Ming Zhao, Luonan Chen, and Kazuyuki Aihara. A discriminative approach for identifying domain–domain interactions from protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 78(5):1243–1253, 2010.

[321] Olga Zhaxybayeva and W. Ford Doolittle. Lateral gene transfer. *Current Biology*, 21(7):R242–R246, April 2011.

[322] C. E. Zhou, J. Smith, M. Lam, A. Zemla, M. D. Dyer, and T. Slezak. MvirDB–a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res*, 35(Database issue):D391–D394, January 2007.

[323] Daoguo Zhou, Mark S. Mooseker, and Jorge E. Galan. Role of the S. typhimurium Actin-Binding Protein SipA in Bacterial Internalization. *Science*, 283(5410):2092–2095, March 1999.

[324] Peter F. Zipfel, Reinhard Wuerzner, and Christine Skerka. Complement evasion of pathogens: Common strategies are shared by diverse organisms. *Molecular Immunology*, 44(16):3850–3857, September 2007.

[325] Lingyun Zou, Chonghan Nan, and Fuquan Hu. Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics*, 29(24):3135–3142, December 2013.

[326] Ulrich Zuegel and Stefan H. E. Kaufmann. Role of Heat Shock Proteins in Protection from and Pathogenesis of Infectious Diseases. *Clinical Microbiology Reviews*, 12(1):19–39, January 1999.

8 Appendix

Taxonomic level	Different categories
strains	1703
species	1003
genus	487
family	204
order	96
phylum	27

Table 8.1: ***Distribution of bacterial genomes in the genome repository on different taxonomic levels.*** *Shown is the distribution of completely sequenced bacterial genomes with annotated pathogenic/symbiotic/non-pathogenic phenotype in the genome repository.*

Species	**# pathogenic/symb. strains**	**# non-pathogenic strains**
Escherichia coli	49	0
Helicobacter pylori	38	0
Staphylococcus aureus	28	0
Salmonella enterica	25	0
Streptococcus pneumoniae	18	0
Chlamydia trachomatis	18	0
Streptococcus pyogenes	16	0
Neisseria meningitidis	14	0
Corynebacterium pseudotuberculosis	14	0
Corynebacterium diphtheriae	13	0
Clostridium botulinum	13	0
Prochlorococcus marinus	0	12
Mycobacterium tuberculosis	12	0
Listeria monocytogenes	12	0
Buchnera aphidicola	12	0
Yersinia pestis	11	0
Bacillus cereus	10	1
Acinetobacter baumannii	10	0
Streptococcus suis	9	0
Haemophilus influenzae	9	0
Francisella tularensis	9	0
Campylobacter jejuni	9	0
Bifidobacterium longum	9	0
Mycoplasma gallisepticum	8	0
Acetobacter pasteurianus	0	8
Vibrio cholerae	7	0
Treponema pallidum	7	0
Shewanella baltica	0	7
Rhodopseudomonas palustris	0	7
...	...	...

Table 8.2: ***Species representation of genomes in the genome repository.*** *Shown are bacterial species in the genome repository, ranked by the highest number of representatives. For each species, the number of pathogenic/symbiotic as well as the number of non-pathogenic strains are given.*

Table 8.3: ***The putative effectome of Chlamydia trachomatis***
Shown are all proteins of C. trachomatis that are predicted as effectors by EffectiveT3, contain eukaryotic-like domains or have experimental evidence for secretion. Chlamydial inclusion proteins are marked Ïncïn the gene name column. For proteins with experimental evidence for secretion, the according publication is indicated.

locus	accession	name	protein description	reference
CT005	NP_219507	Inc	hypothetical protein	Almeida [11]
CT008	NP_219510		ribonuclease HIII	*predicted*
CT018	NP_219520		hypothetical protein	*predicted*
CT025	NP_219527		signal recognition particle, subunit FFH/SRP54	*predicted*
CT034	NP_219536		cationic amino acid transporter	*predicted*
CT035	NP_219537		biotin protein ligase	*predicted*
CT036	NP_219538	Inc	hypothetical protein	Almeida [11]
CT046	NP_219549		histone-like protein 2	*predicted*
CT047	NP_219550		hypothetical protein	*predicted*
CT049	NP_219552	Pls1	hypothetical protein	Betts [31]
CT050	NP_219553	Pls2	hypothetical protein	Betts [31]
CT053	NP_219556		hypothetical protein	Cunha [69]
CT058	NP_219561	Inc	hypothetical protein	Dehoux [73]
CT066	NP_219569		hypothetical protein	*predicted*
CT082	NP_219585		hypothetical protein	*predicted*
CT083	NP_219586		hypothetical protein	Subtil [278]
CT087	NP_219590		4-alpha-glucanotransferase	*predicted*
CT088	NP_219591		secretion chaperone	*predicted*
CT089	NP_219592	CopN	low calcium response E	Fields [95]
CT105	NP_219608		hypothetical protein	Cunha [69]
CT112	NP_219615		oligoendopeptidase F	*predicted*
CT115	NP_219618	Inc	inclusion membrane protein D	Subtil [278]
CT116	NP_219619	Inc	inclusion membrane protein E	Subtil [278]
CT118	NP_219621	Inc	inclusion membrane protein G	Subtil [278]
CT119	NP_219622	Inc	inclusion membrane protein A	Subtil [278]

CT133	NP_219636		rRNA methylase	*predicted*
CT135	NP_219638	Inc	hypothetical protein	Almeida [11]
CT138	NP_219641		microsomal dipeptidase	*predicted*
CT142	NP_219645		hypothetical protein	Cunha [69]
CT143	NP_219646		hypothetical protein	Cunha [69]
CT144	NP_219647		hypothetical protein	*predicted*
CT147	NP_219650		hypothetical protein	*predicted*
CT153	NP_219656		MAC/perforin family protein	*predicted*
CT154	NP_219657		phospholipase D endonuclease	*predicted*
CT155	NP_219658		phospholipase D endonuclease	*predicted*
CT156	NP_219659	Lda1	hypothetical protein	Betts [31]
CT157	NP_219660		phospholipase D endonuclease	*predicted*
CT161	NP_219664		hypothetical protein	Cunha [69]
CT163	NP_219666	Lda2	hypothetical protein	Betts [31]
CT164	NP_219667		hypothetical protein	*predicted*
CT165	NP_219668		hypothetical protein	*predicted*
CT166	NP_219669		hypothetical protein	Betts [31]
CT192	NP_219696	Inc	hypothetical protein	Almeida [11]
CT195	NP_219699	Inc	hypothetical protein	Dehoux [73]
CT196	NP_219700	Inc	hypothetical protein	Almeida [11]
CT205	NP_219709		diphosphate–fructose-6-phosphate 1-phosphotransferase	*predicted*
CT214	NP_219718	Inc	hypothetical protein	Almeida [11]
CT222	NP_219727	Inc	hypothetical protein	Almeida [11]
CT223	NP_219728		hypothetical protein	Subtil [278]
CT224	NP_219729	Inc	hypothetical protein	Almeida [11]
CT226	NP_219731	Inc	hypothetical protein	Dehoux [73]
CT227	NP_219732	Inc	hypothetical protein	Almeida [11]
CT228	NP_219733	Inc	hypothetical protein	Dehoux [73]
CT229	NP_219734	Inc	hypothetical protein	Subtil [278]
CT232	NP_219737		inclusion membrane protein B	*predicted*

CT233	NP_219738		inclusion membrane protein C	Subtil [278]
CT249	NP_219754	Inc	hypothetical protein	Dehoux [73]
CT257	NP_219762	Lda4	hypothetical protein	Betts [31]
CT260	NP_219765		hypothetical protein	*predicted*
CT267	NP_219772		integration host factor alpha-subunit	*predicted*
CT275	NP_219780		chromosomal replication initiation protein	*predicted*
CT280	NP_219785		Na(+)-translocating NADH-quinone reductase subunit D	*predicted*
CT288	NP_219793		hypothetical protein	Subtil [278]
CT292	NP_219797		deoxyuridine 5'-triphosphate nucleotidohydrolase	*predicted*
CT298	NP_219803		DNA repair protein RadA	*predicted*
CT300	NP_219805	Inc	hypothetical protein	Almeida [11]
CT309	NP_219814		hypothetical protein	*predicted*
CT311	NP_219816		hypothetical protein	Lei [168]
CT324	NP_219829	Inc	hypothetical protein	Dehoux [73]
CT329	NP_219836		exodeoxyribonuclease VII large subunit	*predicted*
CT338	NP_219845		hypothetical protein	Cunha [69]
CT344	NP_219851		ATP-dependent protease La	*predicted*
CT345	NP_219852	Inc	hypothetical protein	Almeida [11]
CT357	NP_219866	Inc	hypothetical protein	Almeida [11]
CT358	NP_219867	Inc	hypothetical protein	Dehoux [73]
CT362	NP_219871		aspartate kinase	*predicted*
CT365	NP_219874	Inc	hypothetical protein	Almeida [11]
CT366	NP_219875		3-phosphoshikimate 1-carboxyvinyltransferase	*predicted*
CT368	NP_219877		chorismate synthase	*predicted*
CT373	NP_219882		hypothetical protein	Subtil [278]
CT376	NP_219885		malate dehydrogenase	*predicted*
CT383	NP_219893	Inc	hypothetical protein	Dehoux [73]
CT384	NP_219894		hypothetical protein	*predicted*
CT386	NP_219896		metal dependent hydrolase	*predicted*
CT387	NP_219897		hypothetical protein	*predicted*

CT392	NP_219902		hypothetical protein	*predicted*
CT395	NP_219905		HSP-70 cofactor	*predicted*
CT429	NP_219941		hypothetical protein	Cunha [69]
CT440	NP_219952	Inc	hypothetical protein	Dehoux [73]
CT441	NP_219953	tsp	tail-specific protease	Betts [31]
CT442	NP_219954		hypothetical protein	Subtil [278]
CT449	NP_219962	Inc	hypothetical protein	Almeida [11]
CT450	NP_219963		UDP pyrophosphate synthase	*predicted*
CT456	NP_219969	Tarp	hypothetical protein	Betts [31]
CT460	NP_219973		SWIB (YM74) complex protein	*predicted*
CT463	NP_219976		tRNA pseudouridine synthase A	*predicted*
CT467	NP_219980		2-component regulatory system-sensor histidine kinase	*predicted*
CT469	NP_219982		hypothetical protein	*predicted*
CT472	NP_219985		hypothetical protein	*predicted*
CT473	NP_219986	Lda3	hypothetical protein	Betts [31]
CT483	NP_219997	Inc	hypothetical protein	Almeida [11]
CT497	NP_220012		replicative DNA helicase	*predicted*
CT529	NP_220044		hypothetical protein	Subtil [278]
CT531	NP_220046		UDP-N-acetylglucosamine acyltransferase	*predicted*
CT545	NP_220060		DNA polymerase III subunit alpha	*predicted*
CT550	NP_220065		hypothetical protein	Subtil [278]
CT554	NP_220069		amino acid ABC transporter substrate-binding protein	*predicted*
CT565	NP_220080		hypothetical protein	*predicted*
CT575	NP_220090		DNA mismatch repair protein	*predicted*
CT576	NP_220091		low calcium response protein H	*predicted*
CT578	NP_220093		hypothetical protein	Subtil [278]
CT579	NP_220094		hypothetical protein	Subtil [278]
CT606.1	NP_220123		hypothetical protein	Subtil [278]
CT610	NP_220127		hypothetical protein	Subtil [278]
CT613	NP_220130		dihydropteroate synthase	*predicted*

CT618	NP_220135	Inc	hypothetical protein	Almeida [11]
CT620	NP_220137		hypothetical protein	Muschiol [198]
CT621	NP_220138		hypothetical protein	Hobolt [132]
CT622	NP_220139		hypothetical protein	Gong [120]
CT642	NP_220160		hypothetical protein	Subtil [278]
CT643	NP_220161		DNA topoisomerase I/SWI	*predicted*
CT652	NP_220170		exodeoxyribonuclease V alpha chain	*predicted*
CT652.1	NP_220171		hypothetical protein	Subtil [278]
CT656	NP_220175		hypothetical protein	Cunha [69]
CT671	NP_220190		hypothetical protein	Subtil [278]
CT686	NP_220205		ABC transporter permease	*predicted*
CT694	NP_220213		hypothetical protein	Hower [135]
CT695	NP_220214		hypothetical protein	*predicted*
CT711	NP_220230		hypothetical protein	Muschiol [198]
CT712	NP_220231		hypothetical protein	Subtil [278]
CT715	NP_220234		UDP-N-acetylglucosamine pyrophosphorylase	*predicted*
CT718	NP_220237		hypothetical protein	Subtil [278]
CT721	NP_220240		NifS-related protein	*predicted*
CT737	NP_220256	NUE	SET domain containing protein	Pennini [219]
CT738	NP_220257		Zn-dependent hydrolase	Subtil [278]
CT749	NP_220268		alanyl-tRNA synthetase	*predicted*
CT755	NP_220274		molecular chaperone GroEL	*predicted*
CT766	NP_220285		tRNA delta(2)-isopentenylpyrophosphate transferase	*predicted*
CT772	NP_220291		inorganic pyrophosphatase	*predicted*
CT789	NP_220308	Inc	hypothetical protein	Almeida [11]
CT795	NP_220315		hypothetical protein	Qi [231]
CT796	NP_220316		glycyl-tRNA synthetase	*predicted*
CT798	NP_220318	GlgA	glycogen synthase	Lu [177]
CT813	NP_220333	Inc	hypothetical protein	Chen [53]
CT823	NP_220344	cHtrA	DO serine protease	Wu [311]

CT847	NP_220368		hypothetical protein	Betts [31]
CT848	NP_220369		hypothetical protein	Subtil [278]
CT849	NP_220370		hypothetical protein	Cunha [69]
CT850	NP_220372	Inc	hypothetical protein	Dehoux [73]
CT858	NP_220380	CPAF	protease	Betts [31]
CT860	NP_220382		hypothetical protein	Subtil [278]
CT861	NP_220383		hypothetical protein	Subtil [278]
CT863	NP_220385		hypothetical protein	Subtil [278]
CT867	NP_220389	ChlaDub1	hypothetical protein	Betts [31]
CT868	NP_220390	ChlaDub2	hypothetical protein	Betts [31]
CT875	NP_219502	TepP	hypothetical protein	Chen [55]

Table 8.4: ***Predicted pathogen-host interactome in the human cell during C. trachomatis infection***
Shown are predicted interactions between chlamydial effectors and human host proteins. For each interaction, the responsible domain-domain pair is depicted. Besides a description of the host protein, evidence for gene expression in infected tissues is indicated.

effector	**eff. dom**	**target dom**	**host protein**	**description**	**expressed**
CT008	PF01351	PF09468	NP_078846	ribonuclease H2 subunit B isoform 1	-
CT008	PF01351	PF01351	NP_006388	ribonuclease H2 subunit A	-
CT025	PF02978	PF02978	NP_003127	signal recognition particle 54 kDa protein isoform 1	-
CT025	PF02881	PF02881	NP_003127	signal recognition particle 54 kDa protein isoform 1	-
CT025	PF00448	PF00448	NP_003127	signal recognition particle 54 kDa protein isoform 1	-
CT112	PF01432	PF01432	NP_003240	thimet oligopeptidase	+
CT133	PF08241	PF00583	NP_002961	diamine acetyltransferase 1	+
CT133	PF08241	PF00118	NP_002147	60 kDa heat shock protein, mitochondrial	+
CT133	PF08241	PF00576	NP_000362	transthyretin precursor	-
CT133	PF08241	PF08241	NP_059998	Williams Beuren syndrome chromosome region 22 protein isoform 2	+
CT133	PF08241	PF01380	NP_005101	glucosamine–fructose-6-phosphate aminotransferase	+
CT138	PF01244	PF01244	NP_004404	dipeptidase 1 precursor	+
CT205	PF00365	PF00244	NP_036611	14-3-3 protein gamma	-
CT205	PF00365	PF00036	NP_079452	ninein-like protein	+
CT205	PF00365	PF00365	NP_002618	6-phosphofructokinase type C isoform 1	+
CT205	PF00365	PF01433	NP_071745	endoplasmic reticulum aminopeptidase 2	+
CT205	PF00365	PF01026	NP_114415	putative deoxyribonuclease TATDN1 isoform a	-
CT267	PF00216	PF01423	NP_937859	small nuclear ribonucleoprotein-associated proteins B	-
CT267	PF00216	PF00012	NP_002145	heat shock 70 kDa protein 4	+
CT267	PF00216	PF01702	NP_112486	queuine tRNA-ribosyltransferase	-
CT267	PF00216	PF00580	NP_116196	F-box only protein 18 isoform 1	-
CT267	PF00216	PF00313	NP_001123995	cold shock domain-containing protein E1 isoform 3	+
CT267	PF00216	PF02540	NP_060631	glutamine-dependent NAD(+) synthetase	-
CT267	PF00216	PF00343	NP_005600	glycogen phosphorylase, muscle form isoform 1	-
CT267	PF00216	PF04493	NP_775898	endonuclease V isoform 1	-

CT292	PF00692	PF00692	NP_001939	deoxyuridine 5'-triphosphate nucleotidohydrolase, mitochondrial isoform 2	-
CT298	PF06745	PF06745	NP_068602	twinkle protein, mitochondrial isoform A	-
CT298	PF05362	PF05362	NP_113678	lon protease homolog 2, peroxisomal	-
CT329	PF01336	PF01336	NP_002936	replication protein A 70 kDa DNA-binding subunit	-
CT329	PF01336	PF02540	NP_060631	glutamine-dependent NAD(+) synthetase	-
CT344	PF00004	PF00179	NP_003336	SUMO-conjugating enzyme UBC9	+
CT344	PF00004	PF00659	NP_005021	serine/threonine-protein kinase PLK1	+
CT344	PF00004	PF01398	NP_006828	COP9 signalosome complex subunit 5	+
CT344	PF00004	PF00012	NP_002145	heat shock 70 kDa protein 4	+
CT344	PF00004	PF09280	NP_005044	UV excision repair protein RAD23 homolog A	-
CT344	PF00004	PF00004	NP_004144	origin recognition complex subunit 1 isoform 1	+
CT344	PF00004	PF04683	NP_008933	proteasomal ubiquitin receptor ADRM1 precursor	-
CT344	PF00004	PF01851	NP_002799	26S proteasome non-ATPase regulatory subunit 2	-
CT344	PF00004	PF01399	NP_003743	eukaryotic translation initiation factor 3 subunit C	+
CT344	PF00004	PF03399	NP_003897	80 kDa MCM3-associated protein	+
CT344	PF00004	PF01105	NP_006806	transmembrane emp24 domain-containing protein 2 precursor	-
CT344	PF00004	PF01111	NP_001818	cyclin-dependent kinases regulatory subunit 2	-
CT344	PF00004	PF05348	NP_057016	proteasome maturation protein	-
CT344	PF00004	PF00227	NP_002791	proteasome subunit beta type-9 proprotein	+
CT344	PF00004	PF10508	NP_005038	26S proteasome non-ATPase regulatory subunit 5	-
CT344	PF00004	PF04055	NP_057492	CDK5 regulatory subunit-associated protein 1 isoform a	+
CT344	PF00004	PF05160	NP_006295	26S proteasome complex subunit DSS1	-
CT344	PF00004	PF00574	NP_006003	ATP-dependent Clp protease proteolytic subunit, mitochondrial precursor	+
CT344	PF02190	PF02190	NP_001027026	LON peptidase N-terminal domain and RING finger protein 3 isoform 1	-
CT344	PF05362	PF05362	NP_113678	lon protease homolog 2, peroxisomal	-
CT344	PF00004	PF02985	NP_115826	HEAT repeat-containing protein 7A isoform 1	+
CT362	PF00696	PF00696	NP_002851	delta-1-pyrroline-5-carboxylate synthase isoform 1	+
CT366	PF00275	PF00166	NP_002148	10 kDa heat shock protein, mitochondrial	-
CT366	PF00275	PF00005	NP_005682	ATP-binding cassette sub-family C member 9 isoform SUR2A	+
CT366	PF00275	PF02540	NP_060631	glutamine-dependent NAD(+) synthetase	-

CT368	PF01264	PF00005	NP_005682	ATP-binding cassette sub-family C member 9 isoform SUR2A	+
CT368	PF01264	PF00009	NP_056988	eukaryotic translation initiation factor 5B	+
CT368	PF01264	PF03129	NP_036340	probable histidyl-tRNA synthetase, mitochondrial precursor	+
CT376	PF02866	PF02866	NP_005557	L-lactate dehydrogenase A chain isoform 1	+
CT376	PF00056	PF00056	NP_005557	L-lactate dehydrogenase A chain isoform 1	+
CT395	PF01025	PF00012	NP_002145	heat shock 70 kDa protein 4	+
CT395	PF01025	PF00282	NP_000808	glutamate decarboxylase 1 isoform GAD67	+
CT395	PF01025	PF07690	NP_114427	protein spinster homolog 1 isoform 1	+
CT395	PF01025	PF00132	NP_037466	mannose-1-phosphate guanyltransferase beta isoform 1	+
CT395	PF01025	PF01025	NP_689620	grpE protein homolog 2, mitochondrial precursor	-
CT395	PF01025	PF00293	NP_694853	diphosphoinositol polyphosphate phosphohydrolase 3-alpha	+
CT395	PF01025	PF00156	NP_001034180	ribose-phosphate pyrophosphokinase 2 isoform 1	+
CT441	PF00595	PF00244	NP_036611	14-3-3 protein gamma	-
CT441	PF00595	PF00595	NP_001356	disks large homolog 4 isoform 1	+
CT441	PF00595	PF00227	NP_002791	proteasome subunit beta type-9 proprotein	+
CT441	PF03572	PF03572	NP_002891	retinol-binding protein 3 precursor	-
CT450	PF01255	PF00166	NP_002148	10 kDa heat shock protein, mitochondrial	-
CT450	PF01255	PF01255	NP_079163	dehydrodolichyl diphosphate synthase isoform 2	+
CT450	PF01255	PF01553	NP_848934	glycerol-3-phosphate acyltransferase 6 precursor	+
CT450	PF01255	PF00156	NP_001034180	ribose-phosphate pyrophosphokinase 2 isoform 1	+
CT460	PF02201	PF00439	NP_001122321	transcription activator BRG1 isoform A	+
CT460	PF02201	PF02201	NP_002384	protein Mdm4 isoform 1	+
CT463	PF01416	PF00012	NP_002145	heat shock 70 kDa protein 4	+
CT463	PF01416	PF01416	NP_699170	tRNA pseudouridine synthase-like 1	-
CT467	PF00989	PF00989	NP_858045	nuclear receptor coactivator 3 isoform a	+
CT467	PF02518	PF01751	NP_003926	DNA topoisomerase 3-beta-1	+
CT467	PF02518	PF02518	NP_001135858	pyruvate dehydrogenase kinase, isozyme 3 isoform 1 precursor	+
CT467	PF02518	PF00072	NP_003710	cAMP-specific and IBMX-insensitive 3',5'-cyclic phosphodiesterase 8B isoform 1	+
CT472	PF02582	PF07738	NP_056197	E3 ubiquitin-protein ligase HECTD1	+
CT497	PF03796	PF03796	NP_001157284	twinkle protein, mitochondrial isoform B	-

CT531	PF00132	PF00012	NP_002145	heat shock 70 kDa protein 4	+
CT531	PF00132	PF01008	NP_001405	translation initiation factor eIF-2B subunit alpha	-
CT531	PF00132	PF01145	NP_004090	erythrocyte band 7 integral membrane protein isoform a	+
CT531	PF00132	PF04716	NP_004991	NADH dehydrogenase	-
CT531	PF00132	PF00132	NP_037466	mannose-1-phosphate guanyltransferase beta isoform 1	+
CT531	PF00132	PF01025	NP_689620	grpE protein homolog 2, mitochondrial precursor	-
CT531	PF00132	PF01380	NP_005101	glucosamine–fructose-6-phosphate aminotransferase	+
CT545	PF07733	PF00005	NP_005682	ATP-binding cassette sub-family C member 9 isoform SUR2A	+
CT575	PF02518	PF01751	NP_003926	DNA topoisomerase 3-beta-1	+
CT575	PF08676	PF08676	NP_000526	mismatch repair endonuclease PMS2 isoform a	+
CT575	PF02518	PF02518	NP_001135858	pyruvate dehydrogenase kinase, isozyme 3 isoform 1 precursor	+
CT575	PF01119	PF01119	NP_000526	mismatch repair endonuclease PMS2 isoform a	+
CT575	PF02518	PF00072	NP_003710	cAMP-specific and IBMX-insensitive 3',5'-cyclic phosphodiesterase 8B isoform 1	+
CT613	PF00809	PF00809	NP_000245	methionine synthase	-
CT643	PF02201	PF00439	NP_001122321	transcription activator BRG1 isoform A	+
CT643	PF02201	PF02201	NP_002384	protein Mdm4 isoform 1	+
CT643	PF01751	PF01751	NP_003926	DNA topoisomerase 3-beta-1	+
CT643	PF01131	PF01131	NP_003926	DNA topoisomerase 3-beta-1	+
CT643	PF01751	PF02518	NP_001135858	pyruvate dehydrogenase kinase, isozyme 3 isoform 1 precursor	+
CT643	PF01751	PF03129	NP_036340	probable histidyl-tRNA synthetase, mitochondrial precursor	+
CT686	PF01458	PF00005	NP_005682	ATP-binding cassette sub-family C member 9 isoform SUR2A	+
CT715	PF01704	PF01704	NP_003106	UDP-N-acetylhexosamine pyrophosphorylase	+
CT718	PF02108	PF00118	NP_002147	60 kDa heat shock protein, mitochondrial	+
CT721	PF00266	PF00012	NP_002145	heat shock 70 kDa protein 4	+
CT721	PF00266	PF00994	NP_065857	gephyrin isoform 1	-
CT721	PF00266	PF01230	NP_002003	bis(5'-adenosyl)-triphosphatase	+
CT721	PF00266	PF00501	NP_054750	long-chain fatty acid transport protein 6	+
CT721	PF00266	PF02540	NP_060631	glutamine-dependent NAD(+) synthetase	-
CT721	PF00266	PF01842	NP_004170	tryptophan 5-hydroxylase 1	+
CT721	PF00266	PF02441	NP_068595	phosphopantothenoylcysteine decarboxylase	-

CT721	PF00266	PF01592	NP_998760	iron-sulfur cluster assembly enzyme ISCU, mitochondrial ISCU2 precursor	-
CT721	PF00266	PF00266	NP_478059	phosphoserine aminotransferase isoform 1	+
CT737	PF00856	PF00400	NP_378663	F-box/WD repeat-containing protein 1A isoform 1	+
CT737	PF00856	PF00856	NP_006700	histone-lysine N-methyltransferase EHMT2 isoform a	+
CT738	PF00753	PF00753	NP_056303	probable hydrolase PNKD isoform 1 precursor	+
CT749	PF07973	PF07973	NP_065796	alanyl-tRNA synthetase, mitochondrial precursor	+
CT749	PF01411	PF01411	NP_065796	alanyl-tRNA synthetase, mitochondrial precursor	+
CT755	PF00118	PF00244	NP_036611	14-3-3 protein gamma	-
CT755	PF00118	PF00069	NP_002728	protein kinase C alpha type	+
CT755	PF00118	PF00022	NP_001092	actin, cytoplasmic 1	+
CT755	PF00118	PF00583	NP_002961	diamine acetyltransferase 1	+
CT755	PF00118	PF00400	NP_378663	F-box/WD repeat-containing protein 1A isoform 1	+
CT755	PF00118	PF07544	NP_060489	mediator of RNA polymerase II transcription subunit 9	+
CT755	PF00118	PF01839	NP_002196	integrin alpha-5 precursor	+
CT755	PF00118	PF00012	NP_002145	heat shock 70 kDa protein 4	+
CT755	PF00118	PF00118	NP_002147	60 kDa heat shock protein, mitochondrial	+
CT755	PF00118	PF00149	NP_006238	serine/threonine-protein phosphatase 5 isoform 1	+
CT755	PF00118	PF06677	NP_006387	Sjoegren syndrome/scleroderma autoantigen 1	-
CT755	PF00118	PF00166	NP_002148	10 kDa heat shock protein, mitochondrial	-
CT755	PF00118	PF00085	NP_006532	glutaredoxin-3	+
CT755	PF00118	PF01048	NP_000261	purine nucleoside phosphorylase	-
CT755	PF00118	PF00956	NP_005960	nucleosome assembly protein 1-like 4	+
CT755	PF00118	PF00106	NP_057457	WW domain-containing oxidoreductase isoform 1	+
CT755	PF00118	PF00501	NP_054750	long-chain fatty acid transport protein 6	+
CT755	PF00118	PF02525	NP_000894	NAD(P)H dehydrogenase	-
CT755	PF00118	PF00248	NP_003730	aldo-keto reductase family 1 member C3	+
CT755	PF00118	PF00005	NP_005682	ATP-binding cassette sub-family C member 9 isoform SUR2A	+
CT755	PF00118	PF03095	NP_821068	serine/threonine-protein phosphatase 2A activator isoform a	-
CT755	PF00118	PF01494	NP_073602	NEDD9-interacting protein with calponin homology and LIM domains isoform 1	+
CT755	PF00118	PF00464	NP_004160	serine hydroxymethyltransferase, cytosolic isoform 1	-

CT755	PF00118	PF00153	NP_003347	mitochondrial uncoupling protein 3 isoform UCP3L	+
CT755	PF00118	PF00083	NP_003051	solute carrier family 22 member 5	+
CT755	PF00118	PF06293	NP_291028	TP53-regulating kinase	+
CT755	PF00118	PF04055	NP_057492	CDK5 regulatory subunit-associated protein 1 isoform a	+
CT755	PF00118	PF01507	NP_079483	FAD synthase isoform 1	-
CT755	PF00118	PF02441	NP_068595	phosphopantothenoylcysteine decarboxylase	-
CT755	PF00118	PF00483	NP_037466	mannose-1-phosphate guanyltransferase beta isoform 1	+
CT755	PF00118	PF00155	NP_115981	1-aminocyclopropane-1-carboxylate synthase-like protein 1	+
CT755	PF00118	PF00037	NP_002931	ATP-binding cassette sub-family E member 1	-
CT755	PF00118	PF04851	NP_002171	DNA-binding protein SMUBP-2	+
CT755	PF00118	PF04434	NP_872327	E3 ubiquitin-protein ligase ZSWIM2	-
CT755	PF00118	PF02348	NP_061156	N-acylneuraminate cytidylyltransferase	-
CT755	PF00118	PF01370	NP_001491	GDP-mannose 4,6 dehydratase	+
CT755	PF00118	PF00702	NP_777613	sarcoplasmic/endoplasmic reticulum calcium ATPase 3 isoform e	+
CT755	PF00118	PF00561	NP_004181	gastric triacylglycerol lipase isoform 2 precursor	-
CT755	PF00118	PF00156	NP_001034180	ribose-phosphate pyrophosphokinase 2 isoform 1	+
CT755	PF00118	PF08241	NP_059998	Williams Beuren syndrome chromosome region 22 protein isoform 2	+
CT755	PF00118	PF04158	NP_056235	DDB1- and CUL4-associated factor 13	-
CT755	PF00118	PF02527	NP_079027	glutathione S-transferase C-terminal domain-containing protein isoform 2	-
CT755	PF00118	PF02114	NP_076970	phosducin-like protein 3	+
CT755	PF00118	PF01433	NP_071745	endoplasmic reticulum aminopeptidase 2	+
CT755	PF00118	PF01266	NP_079160	L-2-hydroxyglutarate dehydrogenase, mitochondrial precursor	+
CT755	PF00118	PF00534	NP_001006637	glycosyltransferase-like domain-containing protein 1 isoform a	+
CT766	PF01715	PF01715	NP_060116	tRNA dimethylallyltransferase, mitochondrial precursor	-
CT772	PF00719	PF00012	NP_002145	heat shock 70 kDa protein 4	+
CT772	PF00719	PF00254	NP_003968	AH receptor-interacting protein	+
CT772	PF00719	PF00719	NP_066952	inorganic pyrophosphatase	-
CT798	PF00534	PF00118	NP_002147	60 kDa heat shock protein, mitochondrial	+
CT798	PF00534	PF00534	NP_001006637	glycosyltransferase-like domain-containing protein 1 isoform a	+
CT823	PF00595	PF00244	NP_036611	14-3-3 protein gamma	-

CT823	PF00595	PF00595	NP_001356	disks large homolog 4 isoform 1	+
CT823	PF00089	PF00089	NP_000292	plasminogen isoform 1 precursor	+
CT823	PF00089	PF00273	NP_000468	serum albumin preproprotein	-
CT823	PF00595	PF00227	NP_002791	proteasome subunit beta type-9 proprotein	+
CT858	PF03572	PF03572	NP_002891	retinol-binding protein 3 precursor	-
CT867	PF02902	PF00240	NP_004553	E3 ubiquitin-protein ligase parkin isoform 1	+
CT867	PF02902	PF02902	NP_056386	sentrin-specific protease 6 isoform 1	+
CT868	PF02902	PF00240	NP_004553	E3 ubiquitin-protein ligase parkin isoform 1	+
CT868	PF02902	PF02902	NP_056386	sentrin-specific protease 6 isoform 1	+

Table 8.5: *Enriched molecular functions in predicted protein interaction partners of chlamydial effectors in the human cell*
Shown are all enriched Gene Ontology terms describing molecular functions of interacting host protein candidates (p-value <= 0.05) and targeted by 10 or more chlamydial effectors. Beside the description of each GO term, the number of host interaction candidates annotated with this GO term as well as the number of chlamydial effectors targeting these host proteins are given.

GO term	description	# interaction candidates	# interacting effectors
GO:0042623	ATPase activity, coupled	133	24
GO:0016887	ATPase activity	176	24
GO:0019899	enzyme binding	111	21
GO:0051082	unfolded protein binding	42	19
GO:0016874	ligase activity	172	19
GO:0016491	oxidoreductase activity	191	19
GO:0008233	peptidase activity	275	19
GO:0004672	protein kinase activity	629	17
GO:0004175	endopeptidase activity	227	17
GO:0008289	lipid binding	113	16
GO:0003697	single-stranded DNA binding	16	14
GO:0000287	magnesium ion binding	209	13
GO:0019900	kinase binding	41	12
GO:0017171	serine hydrolase activity	177	12
GO:0008236	serine-type peptidase activity	177	12
GO:0005496	steroid binding	66	12
GO:0019842	vitamin binding	61	11
GO:0019787	small conjugating protein ligase activity	96	11
GO:0016881	acid-amino acid ligase activity	100	11
GO:0016746	transferase activity, transferring acyl groups	70	11
GO:0015291	secondary active transmembrane transporter activity	74	11
GO:0004252	serine-type endopeptidase activity	176	11
GO:0048037	cofactor binding	94	10

GO:0019901	protein kinase binding	36	10
GO:0016747	transferase activity, transferring acyl groups other than amino-acyl groups	69	10
GO:0008415	acyltransferase activity	69	10
GO:0008022	protein C-terminus binding	34	10
GO:0004842	ubiquitin-protein transferase activity	69	10
GO:0004674	protein serine/threonine kinase activity	601	10

Table 8.6: ***Enriched biological processes in predicted protein interaction partners of chlamydial effectors in the human cell***

Shown are all enriched Gene Ontology terms describing biological processes of interacting host protein candidates (p-value <= 0.05) and targeted by 10 or more chlamydial effectors. Beside the description of each GO term, the number of host interaction candidates annotated with this GO term as well as the number of chlamydial effectors targeting these host proteins are given.

GO term	description	# interaction candidates	# interacting effectors
GO:0006259	DNA metabolic process	143	29
GO:0046907	intracellular transport	174	27
GO:0051716	cellular response to stimulus	206	26
GO:0044265	cellular macromolecule catabolic process	228	26
GO:0019752	carboxylic acid metabolic process	178	24
GO:0006519	amino acid and derivative metabolic process	95	24
GO:0006082	organic acid metabolic process	180	24
GO:0006066	alcohol metabolic process	134	23
GO:0046483	heterocycle metabolic process	118	22
GO:0033554	cellular response to stress	156	22
GO:0022402	cell cycle process	165	22
GO:0007049	cell cycle	248	22
GO:0050790	regulation of catalytic activity	165	21
GO:0034984	cellular response to DNA damage stimulus	96	20
GO:0009117	nucleotide metabolic process	94	20
GO:0006974	cellular response to DNA damage stimulus	104	20
GO:0006520	cellular amino acid metabolic process	78	20
GO:0006457	protein folding	69	20
GO:0019725	cellular homeostasis	98	19
GO:0000279	M phase	81	18
GO:0051246	regulation of protein metabolic process	162	16
GO:0044255	cellular lipid metabolic process	147	16
GO:0006281	DNA repair	80	16

GO:0051603	proteolysis involved in cellular protein catabolic process	188	15
GO:0051276	chromosome organization	119	15
GO:0044257	cellular protein catabolic process	188	15
GO:0043085	positive regulation of catalytic activity	118	15
GO:0042493	response to drug	68	15
GO:0006605	protein targeting	67	15
GO:0055114	oxidation-reduction process	145	14
GO:0043086	negative regulation of catalytic activity	77	14
GO:0019941	modification-dependent protein catabolic process	177	14
GO:0006725	cellular aromatic compound metabolic process	88	14
GO:0006260	DNA replication	55	14
GO:0051186	cofactor metabolic process	93	12
GO:0032446	protein modification by small protein conjugation	62	12
GO:0006984	ER-nucleus signaling pathway	13	12
GO:0005996	monosaccharide metabolic process	54	12
GO:0000278	mitotic cell cycle	142	12
GO:0044262	cellular carbohydrate metabolic process	82	11
GO:0032269	negative regulation of cellular protein metabolic process	36	11
GO:0031647	regulation of protein stability	20	11
GO:0006839	mitochondrial transport	52	11
GO:0051301	cell division	76	10
GO:0046942	carboxylic acid transport	54	10
GO:0019318	hexose metabolic process	49	10
GO:0018193	peptidyl-amino acid modification	56	10
GO:0009166	nucleotide catabolic process	26	10

Table 8.7: ***Enriched pathways of predicted protein interaction partners targeted by chlamydial effectors in the human cell***
Shown are all KEGG pathways that are enriched in interacting host protein candidates (p-value <= 0.05) and targeted by 3 or more chlamydial effectors. Beside the description of each KEGG pathway, the number of host interaction candidates assigned to this pathway as well as the number of chlamydial effectors targeting these host proteins are given.

KEGG pathway id	description	# interaction candidates	# interacting effectors
hsa03040	Spliceosome	34	12
hsa04010	MAPK signaling pathway	91	11
hsa04310	Wnt signaling pathway	64	10
hsa04810	Regulation of actin cytoskeleton	68	8
hsa04110	Cell cycle	46	8
hsa04722	Neurotrophin signaling pathway	75	7
hsa04666	Fc gamma R-mediated phagocytosis	23	7
hsa04660	T cell receptor signaling pathway	49	7
hsa04530	Tight junction	59	7
hsa03420	Nucleotide excision repair	17	7
hsa04916	Melanogenesis	36	6
hsa04914	Progesterone-mediated oocyte maturation	33	6
hsa04114	Oocyte meiosis	70	6
hsa00250	Alanine, aspartate and glutamate metabolism	12	6
hsa05014	Amyotrophic lateral sclerosis (ALS)	19	5
hsa04520	Adherens junction	24	5
hsa04120	Ubiquitin mediated proteolysis	68	5
hsa03430	Mismatch repair	13	5
hsa02010	ABC transporters	43	5
hsa00520	Amino sugar and nucleotide sugar metabolism	15	5
hsa00360	Phenylalanine metabolism	11	5
hsa04960	Aldosterone-regulated sodium reabsorption	21	4
hsa04910	Insulin signaling pathway	50	4
hsa04670	Leukocyte transendothelial migration	28	4

hsa04664	Fc epsilon RI signaling pathway	33	4
hsa04621	NOD-like receptor signaling pathway	21	4
hsa04614	Renin-angiotensin system	8	4
hsa04360	Axon guidance	34	4
hsa04350	TGF-beta signaling pathway	28	4
hsa04270	Vascular smooth muscle contraction	47	4
hsa04062	Chemokine signaling pathway	44	4
hsa04020	Calcium signaling pathway	54	4
hsa03320	PPAR signaling pathway	21	4
hsa00970	Aminoacyl-tRNA biosynthesis	14	4
hsa00620	Pyruvate metabolism	15	4
hsa04540	Gap junction	25	3
hsa04510	Focal adhesion	71	3
hsa04210	Apoptosis	30	3
hsa04150	mTOR signaling pathway	21	3
hsa03050	Proteasome	40	3
hsa03030	DNA replication	12	3
hsa00650	Butanoate metabolism	15	3
hsa00640	Propanoate metabolism	11	3
hsa00310	Lysine degradation	24	3
hsa00270	Cysteine and methionine metabolism	13	3
hsa00052	Galactose metabolism	9	3

Table 8.8: *Enriched cellular components in predicted protein interaction partners of chlamydial effectors in the human cell*

Shown are all enriched Gene Ontology terms describing cellular components of interacting host protein candidates (p-value <= 0.05). Beside the description of each GO term, the number of host interaction candidates annotated with this GO term as well as the number of chlamydial effectors targeting these host proteins are given.

GO term	description	# interaction candidates	# interacting effectors
GO:0044429	mitochondrial part	161	29
GO:0005626	insoluble fraction	161	23
GO:0031988	membrane-bounded vesicle	105	22
GO:0016023	cytoplasmic membrane-bounded vesicle	103	22
GO:0005624	membrane fraction	156	22
GO:0019866	organelle inner membrane	107	20
GO:0005625	soluble fraction	84	20
GO:0042995	cell projection	140	17
GO:0042598	vesicular fraction	54	14
GO:0015629	actin cytoskeleton	75	13
GO:0012506	vesicle membrane	37	12
GO:0005793	endoplasmic reticulum-Golgi intermediate compartment	12	12
GO:0005777	peroxisome	38	12
GO:0045177	apical part of cell	47	10
GO:0016324	apical plasma membrane	30	9
GO:0046581	intercellular canaliculus	4	7
GO:0043235	receptor complex	42	7
GO:0019861	flagellum	34	6
GO:0008287	protein serine/threonine phosphatase complex	27	6
GO:0043190	ATP-binding cassette (ABC) transporter complex	9	5
GO:0043198	dendritic shaft	9	4
GO:0032391	photoreceptor connecting cilium	5	3
GO:0009288	bacterial-type flagellum	26	3
GO:0005852	eukaryotic translation initiation factor 3 complex	10	3

GO:0005851	eukaryotic translation initiation factor 2B complex	5	3
GO:0005839	proteasome core complex	29	3
GO:0000502	proteasome complex	49	3
GO:0000159	protein phosphatase type 2A complex	21	3
GO:0048179	activin receptor complex	5	1
GO:0008305	integrin complex	27	1
GO:0005964	phosphorylase kinase complex	27	1
GO:0005954	calcium- and calmodulin-dependent protein kinase complex	5	1
GO:0005890	sodium:potassium-exchanging ATPase complex	5	1
GO:0005838	proteasome regulatory particle	6	1

List of Figures

1.1 Estimated change of incidences of pathogenic infections in the United States for 2012 1

1.2 Overview of bacterial infections in human 2

1.3 Illustration of biphasic developmental lifecycle of *Chlamydia trachomatis* . 5

1.4 Circular map of the *L. pneumophila* 130b draft genome 7

1.5 Overview of the 7 secretion systems identified in bacteria 11

1.6 Schematic overview of Sec/Tat translocase secretion pathways in *Escherichia coli* 12

1.7 Schematic figures and electrone microscope image of the T3SS injectisome 13

1.8 Schema illustrating the two hypotheses regarding T3 secretion signal location 14

1.9 Schematic illustration of secretion mechanisms of Type IV secretion system subgroups 16

1.10 Schematic overview of Type V secretion system subtypes 18

1.11 Schematic phylogenetic tree of selected eukaryotic-like domains 22

1.12 Overview of features used for prediction of effector proteins 23

1.13 Schematic comparison of typical Tat and Sec pathway substrates 24

1.14 Overview of the DIMA database 28

1.15 Scheme of areas in computational biology where machine learning methods are applied 30

1.16 Statistics on normal distributed data 32

2.1 Data sources and information processed in the genome repository 39

2.2 Publicly available phenotype information on the pathogenicity of completely sequenced bacterial organisms 43

2.3 Distribution of pathogenicity phenotype annotations of completely sequenced bacterial organisms in the genome repository 45

2.4 Consistency of phenotype information on pathogenicity in the public databases GOLD, BacMap and NCBI phenotypes for all completely sequenced bacterial genomes in RefSeq 46

2.5 Frequencies of protein domain signatures in eukaryotic genomes 49

2.6 Coverage of Type III Secretion System (T3SS) components in organisms . 54

2.7 Trend of the number of pathogen/symbiont specific protein domains for an increasing number of considered non-pathogenic genomes 58

2.8 Numbers of predicted effectors in 1103 pathogenic and symbiotic genomes, indicated by supporting method 61

3.1 Screenshot of the Effective web portal 69

3.2 3-tier-architecture of the Effective web portal 70

3.3 Entity relationship model of the Effective database 71

3.4 UML description for retrieving precalculated information for all sequenced pathogens and symbionts 76

3.5 Screenshot of the Effective domain list output, presenting the list of high-scoring ELDs in *Legionella pneumophila str. Paris* 78

3.6 Screenshot of the Effective report page for F-Box domain (PF00646) 79

3.7 UML description for the prediction of effector candidates in user provided data 81

3.8 Screenshot of the Effective result page for an interactive exemplary effector prediction 82

4.1 Chlamydial-host PPI targets in the regulation of actin cytoskeleton 97

5.1 Average similarity of eukaryotic-like domain (ELD) and non-eukaryotic-like domain profiles between bacterial genomes on different levels of taxonomic divergence 105

5.2 Performance of the pathofier approach for bacterial genomes in the genome repository . . . 117

List of Tables

2.1 Eukaryotic F-box domain signatures in bacterial proteins 37

2.2 Eukaryotic genomes and the number of domain singletons (exclusively found in that particular genome /in two genomes) 48

2.3 Bacterial genomes and the number of domains which are exclusively found in that particular genome / in two genomes 49

2.4 Representation of T3SS component KOs in eggNOG orthologous groups . . 52

2.5 Bacteria without evidence for a host-dependent lifestyle in public databases that encode a majority of essential T3SS components 54

2.6 Overview of eukaryotic-like protein domains identified in *Chlamydia trachomatis sp.* 64

4.1 Over-represented domain interaction targets of the *Chlamydia trachomatis* effectome 92

4.2 Number of expressed proteins in different human tissues and cell types . . 95

5.1 Average eukaryotic-like domain (ELD) and non-eukaryotic-like domain (non-ELD) similarities between bacterial genomes for different levels of taxonomic divergence 106

5.2 Overview of eukaryotic-like protein domains (ELD) with highest pathogenicity probability 108

5.3 Discriminatory power of a genomes pathogenicity probability 111

5.4 Performance measures of the pathofier approach for different levels of taxonomic divergence 117

8.1 Distribution of bacterial genomes in the genome repository on different taxonomic levels 159

8.2 Species representation of genomes in the genome repository 160

8.3 The putative effectome of *Chlamydia trachomatis* 161

8.4 Predicted pathogen-host interactome in the human cell during *C. trachomatis* infection 167

8.5 Enriched molecular functions in predicted protein interaction partners of chlamydial effectors in the human cell 174

8.6 Enriched biological processes in predicted protein interaction partners of chlamydial effectors in the human cell 176

8.7 Enriched pathways of predicted protein interaction partners targeted by chlamydial effectors in the human cell 178

8.8 Enriched cellular components in predicted protein interaction partners of chlamydial effectors in the human cell 180