TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Development of computational models of the chlamydial interactomes and bacterial secreted proteins

Roland Christian Arnold

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. C. Schwechheimer

Prüfer der Dissertation:

1. Univ.-Prof. Dr. H.W. Mewes
2. Univ.-Prof. Dr. D. Frischmann

Die Dissertation wurde am 27.09.2010 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 08.12.2010 angenommen.

Abstract

This thesis, entitled *'Development of computational models of the chlamydial interactomes and bacterial secreted proteins'*, comprises two parts: the prediction and analysis of interactomes of different species of *Chlamydiae* and the prediction of virulence factors transported by the Type III secretion system by their amino-acid sequence.

The bacterial genus *Chlamydiae* comprises important agents of human and animal diseases as infertility and trachoma. The genomes of these obligate intra-cellular organisms are heavily reduced and contain many specific genes of unknown function. In addition to the pathogenic species, several environmental *Chlamydiae* have been sequenced. These environmental species exhibit less reduced genomes. A functional comparison to pathogenic strains reveals insights in the evolution of *Chlamydiae* in higher eukaryotes. Since no genetic manipulation system in *Chlamydiae* exist which would allow to investigate protein function in these species, function prediction by bioinformatics approaches is of high relevance to understand the biology of *Chlamydiae* and to guide further experiments. By integration of different bioinformatics approaches, I built up a chlamydia specific interaction network. This network has been further analyzed to delineate functional sub-systems (called 'functional modules'). The use of these modules for the identification of virulence related genes and for inference of annotation have been assessed and used for function prediction of uncharacterized genes and to identify virulence related proteins. The evolution of these modules due to the adaptation level on their hosts has been investigated and revealed a non-random pattern indicating a preferred loss of complete functionalities.

The major transport route of effector proteins from many Gram(-) bacterial into their host cells is the Type III secretion apparatus. The identification of effector candidates is a short-cut to identify novel virulence factors, since systematic screens in the bacteria of interest are difficult to achieve. Until this work, no general applicable bioinformatics approach to detect effectors has been proposed. I could show, that the recognition signal that leads to specific transport is taxonomically universal and can be modeled computationally using a machine-learning approach. The resulting software-package (EffectiveT3) allows the generation of high confident effector candidate lists from sequence information that will accelerate the characterization of novel effectors.

Zusammenfassung

Die vorgelegte Arbeit *"Development of computational models of the chlamydial interactomes and bacterial secreted proteins"* gliedert sich in zwei Abschnitte: der erste Teil beschreibt die Berechnung von Interaktionsnetzwerken der Chlamydien und ihre weitere Analyse. Der zweite Teil beschreibt eine sequenzbasierte Methode zur Vorhersage von Effektor-Proteinen, welche durch das Type III Sekretionssystem sezerniert werden.

Die Gruppe der Chlamydien beinhaltet relevante Mensch- und Tierpathogene, welche u.a schwere Erkrankungen des Auges und des Genitaltraktes hervorrufen, speziell sind Chlamydieninfektionen eine häufige Ursache von Infertilität. Die Genome dieser Chlamydien sind stark reduziert und die Funktionen vieler ihrer Gene ist unbekannt. Die Genome etlicher der pathogenen Vertreter der Chlamydien liegen sequenziert vor. Chlamydien, welche aus Umweltisolaten gewonnen wurden, zeigen eine geringere Reduzierung iher Genome. Ein systematischer Vergeich der Genomdaten kann Aufschluss über die funktionelle Evolution in Bezug auf die Anpassung an einen spezifischen Wirt geben. Für Chlamydien existiert kein genetisches Manipulationssystem, und die Charakterisierung vieler ihrer Genprodukte ist daher eingeschränkt. Daher besitzt die Bioinformatik in der Erforschung der Chlamydien einen hohem Stellenwert, insbesondere zur Funktionsvorhersage. Durch die Integration verschiedener Methoden der Bioinformatik habe ich eine (funktionelle) Interaktionskarte der Chlamydien erstellt. In folgenden Analysen wurden zelluläre Systeme (sog. funktionelle Module) aus diesen Netzwerken abgeleitet und Strategien zur Identifizierung unbekannter Virulenzfaktoren und zur Funktionsannotation mit Hilfe dieser Module entwickelt und für Vorhersagen eingesetzt. Des weiteren konnte gezeigt werden, dass die Evolution der funktionellen Module zwischen humanpathogenen Chlamydien und den Umweltchlamydien Regelmäßigkeiten unterworfen ist und bevorzugt komplette Module und nicht einzelne Proteine bei der Genomreduktion verloren gehen. Ein wichtiger Weg zur Sekretion von Effektorproteinen ist das Type III Sekretionssystem. Eine Vorhersage der transportierten Proteine ist von hoher Relevanz, da die Sekretion nur mit hohem Aufwand im Labor gezeigt werden kann. Bisher existierte kein generelles Vorhersageverfahren, welches aus Sequenzdaten wahrscheinliche Substrate des Type III Weges identifizieren konnte. Ich konnte zeigen, dass das Signal zum Transport in verschiedenen Bakterien universell ist und zur generellen Vorhersage von Type III transportierten Effektoren nutzbar ist. Die Vorhersage-Software EffectiveT3 kann eingesetzt werden, um mit hoher Verlässlichkeit Effektorkandidaten aus Sequenzdaten vorherzusagen und so die Suche nach noch unbekannten Effektoren zu beschleunigen.

Acknowledgements

This work would not have been possible without the vivid scientific environment which I encountered at the Wissenschaftszentrum Weihenstephan, the Institute for Bioinformatics and Systems Biology at the Helmholtz-Zentrum and with our cooperating groups in Vienna and Graz. Successful scientific projects need exchange of thoughts and discussion as well as cooperation between different scientists and fields.

I want to thank especially Prof. Dr. Hans-Werner Mewes who gave me the opportunity to do this PhD project as well as to work in different projects for many years at the IBIS. I want to thank my supervisor, Prof. Dr. Thomas Rattei for constant support and valuable advice in many occasions.

I also want to thank the whole teams at Weihenstephan and in Neuherberg for support and a nice and friendly scientific environment. Especially I want to mention Patrick Tischler, Marc-Andre Jehl, Dr. Andreas Kirschner, Florian Blöchl, Florian Büttner, Jan Krumsiek, Prof. Dr. Dmitrij Frishman, Dr. Karamfilka Nenova, Dr. Volker Stümpflen, and Dr. Martin Münsterkötter for many helpful discussions and support in technical and scientific questions. I want to thank Prof. Dr. Matthias Horn and Prof. Dr. Michael Wagner in Vienna and Prof. Dr. Ellen Zechner in Graz for the interesting collaborative projects and discussions. From these groups, I want to thank Silvia Lang, Eva Heinz, Sabine Kienesberger, Dr. Alexander Loy, Dr. Stephan Schmitz-Esser and especially Dr. Astrid Horn for direct support in the biological interpretation of data. I want to thank Stefan Brandmaier, Frederick Kleine, Sebastian Behrens, Antti Niinikoski, Anna Dieckmann, and Sebastian Ullher which valuable supported my projects during internships and student projects.

I also want to appreciate the constant support of my friends and family, thanks a lot to You all.

Contents

1 Introduction

The kingdom of *Bacteria* can be seen as the most successful one in terms of biomass and diversity. *Bacteria* participate in almost every important biological process on earth as, among many others, the fixation of nitrate or the degradation of dead biomass. So, they intensively shape the ecological systems on earth. Many *Bacteria* have developed symbiotic or parasitic life-styles and live in close relationship with diverse hosts. In amoeba, for example, several bacterial species can be found affiliated or as intra-cellular symbionts including *Chlamydiae* and *Bacteroidetes* [1, 2]. Other bacteria are important agents of human and animal infections causing a wide range of diseases as *Yersinia* or *Salmonella* [3, 4]. Many of these pathogenic and symbiotic bacteria exhibit a facultative or obligate intra-cellular life-style. The genomes of many intra-cellular bacteria exhibit a reduction in their gene content. Driving forces of this reduction are the restricted ability to acquire genetic material and therefore to compensate deleterious mutations, and the possibility to acquire nutrients from the host, a process that allows to lose the corresponding anabolic pathway in the bacterium [5, 6]. An important aspect of the interaction between bacterial and eukaryotic host cells is the direct manipulation of the host by the bacterium using effector proteins which are actively transported into the target cells by different transport systems. Knowledge on these effector proteins is therefore the basis to understand the mechanisms of bacterial virulence and a possible starting point for the development of novel antibiotic drugs.

An interesting bacterial clade comprise the *Chlamydiae*, pathogens with an obligate intra-cellular life-style. *Chlamydiae* are an important target for bionformatic's research: despite of their clinical relevance, the availability of complete genome sequences from a variety of different chlamydial species allows comparative studies to investigate the evolution of host adaptation. Most importantly, *Chlamydiae* are difficult to assess by wet-lab standard procedures as genetic screens which makes bioinformatics analyses especially important to the *Chlamydiae* research community to short-cut experimental costs and time.

The molecular mechanisms of biological processes such as bacterial virulence cannot be understood by looking at one protein or gene alone. The availability of a plethora of data as complete genome sequences (thousands genome reached) and exhaustive protein-protein interaction screens allows to investigate the interplay of proteins on a large scale and to identify groups of genes which work together, commonly named functional modules. These modules allow a view on cellular functionality which is not gene-centric but describes a cell in terms of sub-systems.

The first part of this work deals with the prediction of chlamydial interactomes and functional modules. The use of this information to detect novel virulence factors and to annotate proteins is assessed. In addition, the evolution of functional modules between chlamydial species with differently strong host adaptation and genome reduction is described. The second part comprises a new approach to detect bacterial virulence factors transported by the Type III secretion apparatus, called EffectiveT3. The thesis starts with an introduction to the biological background and bioinformatics' concepts used in the work.

1.1 Type III secretion

Many bacteria can manipulate their environment by the secretion of proteins which enable, e.g the utilization of nutrients and defense against competitors or the immune response of a respective host [7, 8, 9]. Most prominent is the interaction of symbiotic and pathogenic bacteria with their respective host cells [10]. Seven different secretion systems (Type I–VII) have been described until today. Three of them (Type III, IV and VI) allow penetrating host cell membranes and injecting proteins into the cytosol of host cells. Among them, the Type III secretion apparatus (TTSS) is encoded in the genomes of many, mainly pathogenic or symbiotic, Gram-negative bacteria and is a key factor for the virulence of pathogens.

1.1.1 The Type III secretion system

Structure and Function The TTSS is evolutionary related to the flagellar system and well conserved across a wide range of taxa [11]. It spans both bacterial membranes and enters the host cytosol and is formed by 20–30 different proteins. Several proteins form the exporter, which envelopes the proteins of the basal body. The basal body spans the inner membrane, followed by the secreton which spans the outer membrane and is followed by the needle subunit. The needle is topped by the translocon, which upon contact to the host cell forms a translocation pore within the eukaryotic membrane. Proteins of the translocon are transported by the TTSS itself. The TTSS transports a variety of proteins directly into the host cell in a specific and energy dependent manner. These transported proteins interact with proteins in the host yielding an effect suitable for the bacterium and are therefore called *effectors*. The specific transport implies a targeted recognition of effectors as substrate for transport, the mechanisms of this specific recognition are merely unknown and could only be fully understood with comprehensive knowledge of the TTSS on the structural level. However, up to now, only a small number of TTSS related structures have been resolved. Especially structures of effectors bound to components of the TTSS could not be crystallized yet which would be informative on the substrate recognition and transport. Some recognition particles

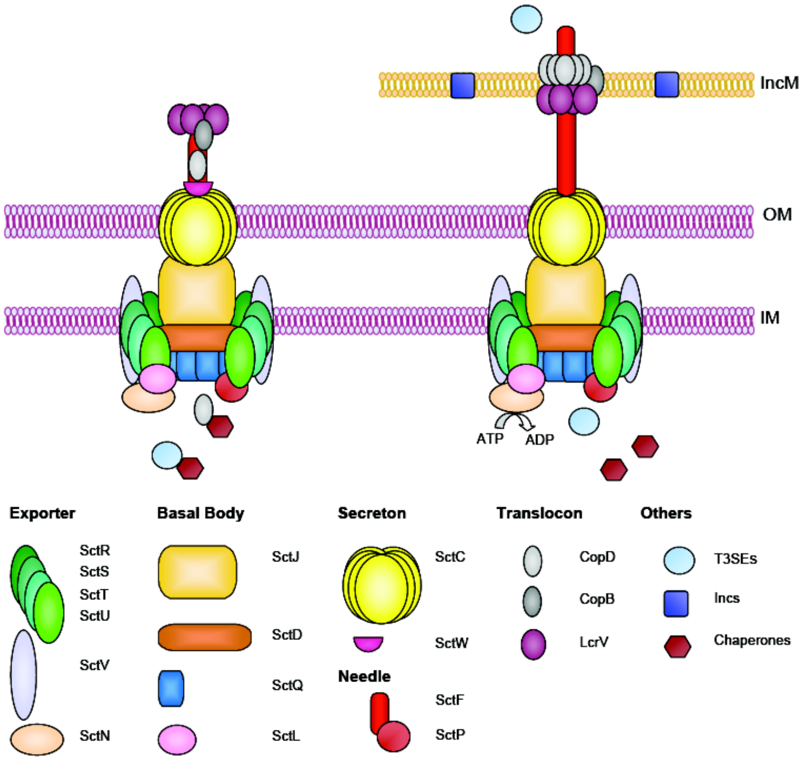


Figure 1.1: The Type III secretion system in *Chlamydiae* as described in [12]. Figure from 'Bacterial secretion systems with an emphasis on the chlamydial Type III secretion system', Beeckman, D. S. and Vanrompay, D. C, 2010. On the left side, an inactive T3SS is shown, while the right side depicts the T3SS after activation. IM, bacterial inner membrane; OM, bacterial outer membrane; IncM, inclusion membrane.

within the TTSS could be identified: in *Escherichia coli*, a direct interaction between the effector Tir and the ATPase EscN has been reported. The latter one is part of the exporter component of TTSS. The chaperon CesT interacts with both Tir and EscN while enhancing secretion [13] and the ATPase might therefore play a role in initial substrate binding. The *Yersinia* proteins YscP and YscU both influence the regulation of translocon and effector secretion. YscP was found to control the needle length and also works as a substrate specificity switch. It is suggested to stop secretion of the translocon YscF while activating Yop secretion upon contact to the host cell [14]. Mutations that trigger conformational changes in the exporter component inhibit recognition and transport of translocon components. The same mutations do not affect export of the Yop effectors [15]. This is an evidence that the recognition differs for transport between substrates which build up the system and effector proteins with destiny to the host cell differ.

TTSS related chaperons Related to the TTSS system itself, several TTSS related chaperons exist. These chaperons play important roles for recognition and regulation, covering several different aspects within the TTSS machinery. Several effectors depend on the association with a specific chaperon for efficient translocation and secretion. For example, the transport of the *Yersinia* effector YopE is activated through binding to its chaperon SycE [16]. In *Y. pestis*, the Yop effectors are mediated by their individual corresponding Sycs (specific Yop chaperones)[17], probably introducing an hierarchy and temporal order for secretion among them [18]. Such chaperons are not restricted to *Yersinia*: in *Salmonella*, the effectors SptP and SopE contain a chaperon binding domain essential for secretion into the host cell via the TTSS machinery [19]. Many effector-binding chaperons have a common fold and effector-binding mode [20], while others show no structural similarity at all [21]. Before completing the secretion step, the effector–chaperon association is released via the SctN component of the TTSS as shown for the SctN homolog InvC in *Salmonella* [22]. Although many findings support the importance of these chaperons for secretion, no general mechanism based on the interaction of effectors with chaperons can be deduced. The binding sites of different chaperons show only weak sequence similarity over different effector molecules [20] and can therefore not be employed for the general detection of novel effectors.

1.1.2 Effectors

The characterization of effector proteins gives hints on the strategies bacteria use to manipulate host cells: some interact with cell signaling pathways to suppress immune response by inducing apoptosis in macrophages as the *Yersina* effector YopJ or the Salmonella effector SipB [23, 24]. Other known effectors manipulate the cytosceleton by actin re-arrangements as described for the Salmonella effector SipA [25]. The arsenal of known effectors varies widely between different bacterial species due to adaptation to different hosts and different survival strategies [26] and even between different strains of the same organism as shown for *Pseudomonas syringae* [27]. Effectors are very diverse and show no typical folds or domain composition, which could be used to identify them with certainty. This picture is congruent with the diverse functions in different hosts and environments that these sequences have to fulfill. Nevertheless, the N-termini share an unusual amino acid composition in the first 20–50 residues, in which e.g. Serine is enriched compared to arbitrary sequences [28]. For several organisms, specific promoters as the *Pseudomonas* HrP box have been described [29], which allow co-regulation of effectors with the TTSS or secretion related chaperons. Since the effectors often interact with complex eukaryotic specific pathways as cell signaling, they have to implement eukaryotic like functional domains. These can be either analogous replacements of an eukaryotic functionality (and act as molecular mimicry [30]), or could be acquired by gene transfer as initially eukaryotic sequence. Many pathogens comprise proteins with a detectable eukaryotic-like domains as the amoeba symbiont *Candidatus Amoebophilus asiaticus* [1] or the environmental *Chlamydium*. An example of a known interaction between the host and effectors with such characteristics comprise the *Yersinia* YopJ effector which inhibits NF-κB by a eukaryotic like SH2 domain [31]. A typical pattern of attack used by several bacteria is to mimic parts of the ubiquitin proteasome by effector proteins [32].

1.1.3 Experimental derived knowledge on the TTSS secretion signal and modes of transport

In the case of the Type III secretion system, the exact mode of signal recognition is unknown. Several experimental evidences could give hints on the nature of the signal and are shortly reviewed here and further insights that could be gained by a computational analyses are discussed .

Location of the signal The N-terminal location of the signal has been assessed by the use of fusion proteins consisting of a N-terminus of an effector protein (the putative signal) and a reporter gene which allows to identify the (secreted) protein outside the bacterial cell. It has been initially studied on the *Yersinia* effector proteins (Yersinia outer proteins, Yops), namely YopH, YopQ, and YopE by Michiels et al. [33] using such hybrid protein assays and has been localized on the first 48–98 N-terminal residues of the Yops. Even shorter N-terminal regions between 10 and 25 residues are sufficient for transportation as shown in several fusion experiments of reporter genes with effector N-termini [34, 35] in different organisms. For example, the minimal signal length of the Yersinia YscP effector has been determined by Riordan et al. to be only 10 residues long. The N-terminal signal can even be exchanged between different effectors of *E. coli* without abolishing function [36]. These findings imply an N-terminal signal location for the proteins under investigation. A computational method that captures the secretion signal can give further insights here: the generality of the N-terminal location can be shown by tests on several effectors, and the length and size of the signal can be deduced as the areas with the highest discriminative power between effectors and non-effectors.

Mode of encoding and transport Whereas the N-terminal localization of this signal is commonly accepted, there is a debate whether the signal is encoded in the translated peptides or in the underlying mRNAs. The mRNA signal hypothesis introduced by Schneewind, Anderson and co-workers (reviewed in [37]) is based on the observation that point as well as frame-shift mutations of several Yop N-termini do not abolish transport, whereas silent mutations on the underlying mRNA have influence on transport [38, 39, 40]. In addition to the proposed N-terminal mRNA signal, Blaylock et al. found a second putative mRNA born signal in the *Yersinia* effector YopR mRNA at codon position 131–149, which is sensitive to silent mutations. For this second signal, a secondary structure containing a stem-loop could be modeled. Interestingly, a mutation in the mRNA, which was predicted not to alter this secondary structure, kept this signal functional whereas mutations abolishing the structure did not [41]. A study with mutants of the *Pseudomonas* effectors AvrB and AvrPto indicates an mRNA signal which can be detected by a Yersinia TTSS [42] and could also be a functional signal in *Pseudomonas*. This theory implies that translation does not occur before recognition by the TTSS and the effectors are synthesized into the TTSS during transport. This is contradicting to the existence of effector-binding chaperons, which can only act on translated sequences. Translation before translocation has been explicitly shown for some effectors as for the Salmonella SopE and SipA [43, 44] the Shigella IpaB and IpaC proteins [45], and the Yersinia YopE [46] proteins. In case of IpaB, IpaC, and SipA, the existence of a large pool of proteins in the bacterial cell and its rapid translocation during infection has been shown by time-lapse microscopy [44, 45]. Even more, there is evidence that effectors can exist in folded state prior to secretion and must be unfolded before transport [47]. A successful computational modeling of the signal based on the amino-acid sequences would be a strong hint in favor to the peptide-based hypothesis but no general proof: the protein sequence is dependent on its' coding RNA and from a theoretical point of view, a signal (encoded by regularities in either the mRNA or amino-acid sequence) could be mutually detected in both sequences due to their correlation by translation as if the mRNA encoded recognition sequence would be strong enough to alter the amino-acid sequences in an detectable fashion. Both theories are picturized in Figure 1.2.

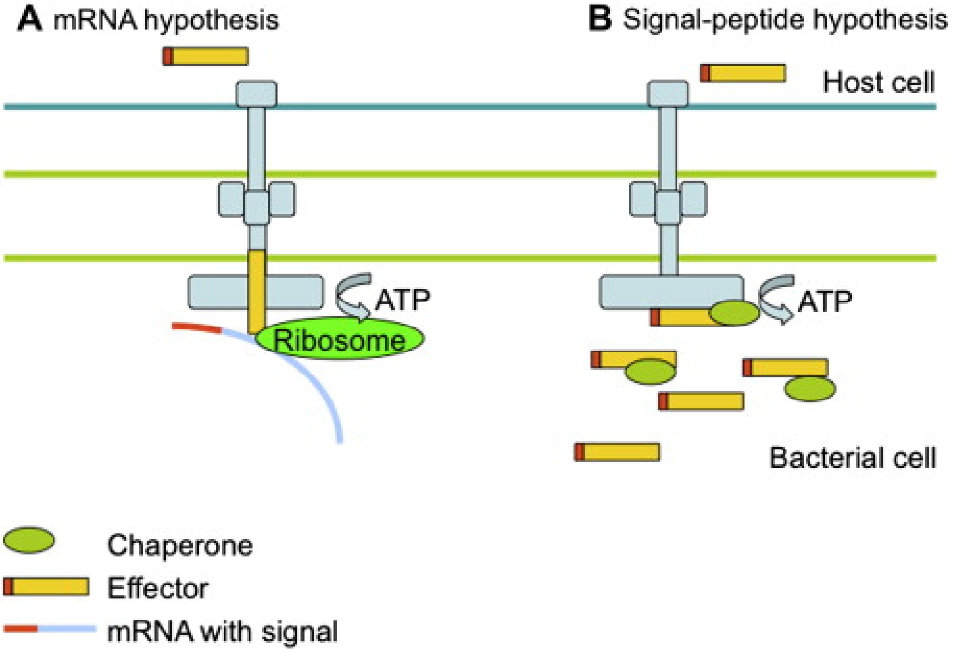


Figure 1.2: Schematic illustration of the two hypotheses of the location of the N-terminal secretion signal: mRNA-based (A) and peptide-based (B). In (A), the effector mRNA, which carries the signal, is synthesized into the TTSS during transport. In (B), the effector is translated in the bacterial cytosol and recognized by a peptide born N-terminal signal. Chaperones play different roles, as enhancing signals or holding the protein in an unfolded, transportable state.

Generality As reported above, the N-termini of different effectors could be exchanged without loss of transportation. As a consequence, no dependency between the functional part of the protein and the signal exists for several *E. coli* effectors [36]. The use of an heterologous assay used in the study of Subtil et al. also indicates that the signal of chlamydial effectors can be recognized by the *Shigella* TTSS. A certain generality of the signal can be expected since the TTSS itself is highly conserved. A computational approach could prove this hypothesis further if unseen instances can be predicted in species that did not participate in the initial deduction of the method.

1.1.4 Approaches to detect Type III secreted proteins by computational methods

Computational analyses have been employed to short-cut screens for effector proteins in several studies which are shortly reviewed here. The perhaps most widely used approach to detect proteins with certain traits is to search for homologs of known proteins comprising those traits. Homology searches against a database of known effectors has been successfully applied by Tobe et al. as initial step to create a candidate list of more than 60 putative effectors in enterohemorrhagic *E. coli* (EHEC) O157:H7, from which for 39 proteins secretion could be shown [35]. Vinatzer et al. and Studholme et al. [48, 49] compared the effector repertoire between different *Pseudomonas* strains using BLAST as initial step. This approach is applicable since a plethora of *Pseudomonas* effectors are known and therefore the variance between different closely related *Pseudomonas* strains can be estimated by homology information. Since virulence factors must act in a concerted manner, the detection of co-regulation of effectors with TTSS components could lead to the identification of effectors since special transcriptional control of effectors has been described in several species [50, 51]. Their specific regulatory elements have been described in *Pseudomonas* [52], *Xanthomonas* [53], *Escherichia* [54, 55], and *Salmonella* [56]. In *P. syringae*, the virulence specific hypersensitive response and pathogenicity sigma factor HrpL activates the pathogenicity regulon, including the TTSS, as well as of some effectors. Fouts et al. [57] created a sensitive Hidden Markov Model detecting HrpL binding sites (called Hrp boxes in *Pseudomonas* [52]) in *P. syringae pv tomato* DC3000 and detected twelve effectors or other virulence related genes in the downstream region of the specific promoter binding sites. Jiang et al. detected 47 novel effectors in *Xanthomonas campestris pv campestris*, screening for a motif of the plant-inducible promoter (PIP) described as binding site for the HrpX regulatory protein in *Xantomonas*

[58]. Chaperons play a mediating role for the translocation of several effectors, and an initial screen on them can give hints to yet unknown effectors. Panina et al. screened *Bordetella bronchiseptica* and several other organisms for putative TTSS related chaperons [59], identified using predicted characteristics as molecular weight fold. Proteins co-localized to these putative chaperons on the chromosome have been further filtered to exclude unlikely candidates. Chaperon–effector pairs have been identified, comprising previously known as well as unknown of these pairs. Hu et al. modeled the interaction of the effector YopE with the chaperon SycE using a docking approach starting with a predicted structure of YopE unbound to the chaperon [60]. The model turned out to be in good agreement with the experimentally solved structure of the complex. This approach does not screen for novel candidates directly, but might be a starting point for an novel bioinformatics approach predicting specific effector/chaperon pairs. Due to the large amount of experimental evidence for a protein born N-terminal signal, it should be straightforward to identify novel effectors by their N-terminal sequences, as by sequence similarity to N-termini known to be transported. Unfortunately, the N-termini of known effectors are very diverse and show no apparent evolutionary conservation between different effector orthologs and therefore do not allow the deduction of a meaningful alignment. In consequence, classical bioinformatics approaches as e.g. deriving sequence motifs, which rely on an initial multiple alignment of a domain, are not applicable to model the signal. Lloyd et al. investigated the first eight residues of the signal peptide by a mutation analysis (directed towards an enrichment of Serine residues) of the Yersinia YopE effector [61] and deduced a model of the signal using linear regression. This ‘synthetically’ derived model has been found to be descriptive for real and putative effectors from different species. This model described the signal as an amphipathic character and an enrichment of Serine in the N-termini is (up to five Serines in the first eight amino acids). Schechter et al. use a composition based rule derived from known *Pseudomonas* effectors to create an initial candidate list for novel effectors [62]. When analyzing the first 50 residues of *Pseudomonas* effectors, they found Serine and Proline enriched, comprising together $>$10% of the N-terminus, and at position 3 or 4 always an aliphatic amino acid. In addition, no acidic amino acid could be found within the first 12 residues Petnicki-Ocwieja [63] used a very similar pattern derived from *Pseudomonas* effectors comprising an high amount of Serine and Proline ($>$10% in the first 50 residues). The screen with these patterns revealed several novel effectors, but the applicability of the model to other species is unclear.

These bioinformatics approaches can be broadly categorized into ’non-signal’ and ’signal’ based, depending on whether they assess and model the N-terminal signal directly or identify effectors by other information. All 'non-signal' based approaches are limited due to their generality and/or applicability, and for the signal based approaches described so far, their applicable generality in terms of independence of the organism has not been shown. Homology based approaches can only detect effectors which are members of known effector families, and these are mostly specific for certain well-known bacterial species. This approach lacks the ability to detect novel effector families where no initial query sequence is available. This drawback is rather severe, since many effectors are specific to a certain clade of bacteria, as e.g. the inclusion proteins of Chlamydia [64], since they depend on the evolutionary history, ecological niche and host adaptation of the bacteria and substantially differ even between closely related species. Thus, the method is particularly not applicable for species with no or only few well studied relatives. Nevertheless, sequence similarity searches are an important tool for comparative genomics studies in closely related species Approaches using transcriptional co-regulation need knowledge about a TTSS effector specific promoters which have not yet been described for most bacteria possessing a TTSS. The unusual amino acid composition in the effector N-termini has to date only been described and exploited in screens in *P. syringae*. Chromosomal co-localization is only applicable if effectors and TTSS related proteins or chaperons are clustered in genomic proximity as described for the pathogenicity islands in *Salmonella* [65]. However, these pathogenicity islands are absent in other bacteria known to harbor a TTSS such as the *Chlamydiae*, for which the genes encoding known effectors are scattered around the genome [66, 67]. The applicability of genomic proximity has been further assessed in this work, as well as the use of 'genomic context methods' to detect functional relationships between effector and secretion system proteins. In addition, the co-membership of chlamydial effectors and TTSS components in virulence related functional modules has been investigated as described in Chapter **??**. Analyses based on co-regulation are restricted to organisms for which a special regulation for virulence is known and an identification of the related regulatory elements is feasible by bioinformatics methods, which is a difficult problem in many cases due to the degenerated nature of their short sequences within the promoter. Approaches by this concept lack specificity to TTSS mediated secretion as they may also detect virulence related genes in general, but could be complemented by additional, effector specific information as an unusual amino acid content in the N-termini of candidates, as utilized by Fouts et al. [57]. The approach to detect pairs of effectors and their cognate chaperons is very elegant since it is species-independent. It is constraint by the actual need of a chaperon for transport, which has only been shown for a couple of effectors. Large-scale screens using the very N-terminal end of an effector [34, 35, 62] strongly indicate chaperon independent substrate recognition which might be regularly the case *in vivo*. This leads to a high false negative rate, since effectors need not to be co-localized to a chaperon.

1.1.5 Complementary approaches to detect secreted and effector proteins of other pathways

Effector proteins (or, more generally, proteins that act outside the bacterial cell) can be delivered not only by the TTSS, but also by other secretion systems. The general secretion pathway recognizes an N-terminal signal which contains a cleavage-site that is removed during transport. The complete signal and especially the cleavage-site can be computationally modeled and used for the prediction of proteins that are substrate to this system. The most prominent implementation of such a prediction software is SignalP [68] which is based on a neural net to predict cleavage site and secretion probability. As aforementioned, many effectors mimic eukaryotic domains. This observation can be used to identify effector proteins rather by their eukaryotic like function as by their recognition sequence. These eukaryotic like proteins can be identified by domain signatures that occur on proteins participating in typically eukaryotic functionalities as in signal transduction pathways. Since they nevertheless exist in bacteria as effectors, they are named 'eukaryotic like domains'. This idea has been applied in several studies as by Angot and co-workers [32] or as feature in an integrative approach for the detection of *Legionella* effectors by Burstein and co-workers [69]. This approach is fruitful, since for most secretion systems no prediction tool is available (as by the Type IV system). In a project of our group, Andre Jehl systematically identified eukaryotic like domains using a simple statistical frame-work and exhaustive comparisons of bacterial and eukaryotic genomes. The aim of the project is to identify domain signatures enriched in pathogenic and depleted in non-pathogenic bacteria that are also present in eukaryotes. The system revealed several known examples as well as novel domain candidates.

1.2 Chlamydiae

The first part of this work deals explicitely with *Chlamydiae*, therefore they are shortly introduced in this Section.

1.2.1 Biology and clinical relevance

Chlamydiae, firstly described as *Chlamydozoa* by Halberstaedter and von Prowazek in 1907 [70] as agent of trachoma, are obligate intra-cellular bacteria. They exibit a unique biology with a bi-phasic life-style and form separate inclusions in their host cell. They are causing agent of several diseases and *Chlamydia* infection is the most frequent sexually transmitted diseases in developed countries. Several different species of the phylum have been described and the host range of *Chlamydiae* covers species as different as amoebae [71, 72], birds [73], small ruminents [74], fish [75], frog [76], cattle [77], and human [78, 79], whereby the individual species and strains are typically found in a specific sets of hosts [73].

Clinical and economic relevance In humans, *Chlamydiae* cause several diseases including pelvic inflammatory disease leading to infertility (estimiate of cases of chlamydia caused infertility in the USA: 40000 p.a. [80]) and urinary tract infections [81], diseases of the respiratory tract [82, 83], and infection of the eye [84]. The complete economic costs are estimated by Washington and co-workers as around 1.5 billion USD p.a. in the United States alone [85]. The impact of *Chlamydiae* based infection on animals cannot be neglected: a specific strain of *C. pneumoniae* [86] could be related to the decline in the population of the koala *Phascolarctos cinereus* [87] and chlamydial infection of animals is also an economic issue due to high abort rates caused by *Chlamydia abortus* in sheep, goat and cattle [88, 77].

Taxonomy The taxonomy of the Phylum *Chlamydiae* is a topic of vivid discussion in the scientific community. Broadly, the known chlamydia-like species can be separated into 'pathogenic' species which are very specialized to an higher eukaryotic host, and 'environmental' species, often symbionts in amoebae, that are supposed to have a broader host range and that are isolated from environmental samples. A molceular based taxonomy has been introduced recently by Bush and co-workers by the creation of robust trees from five different proteins (the major outer membrane complex MOMP, the GroeL chaperone, a KDO transferase, a small cystein rich lipoprotein, and another cysten rich protein) [89]. The resulting consensus-tree turned out to be discriminative between the nine species used in this work. The authors propose a phylogeny in which the 'pathogenic' *Chlamydiae* are split into two genus, the *Chlamydophila* and *Chlamydia* which comprise the family of *Chlamydiaceae.* The 'environmental' species comprise their own families *Simkaniaceae, Waddliaceae* and *Parachlamydiaceae.*

Chlamydial genomes: reduction due to intra-cellular life-style The amount of coding sequences in the genomes of the *Chlamydiae* differs between 895 genes in the smallest pathogenic and genes 2854 in the largest genome of the environmental species. Such relatively small amounts of genes are typical for bacteria with intra-cellular life-style [90], for which different evolutionary forces lead to genome reduction in intra-cellular pathogens as the inability to acquire novel genetic material by horizontal gene transfer [90, 91]. This variance in genome sizes within the environmental and especially in comparison to the pathogenic *Chlamydiae* can be interpreted as differently strong adaptation to the respective hosts as indicated by greater metabolic capabilities and a wider host range of the environmental samples [92, 93, 94, 95]. In general, loss of genes due to an intra-cellular life-style eleminates complete cellular functionalities [96] which must be complemented by (parasitic) interaction with the host [97, 98, 1, 99].

Biphasic life-cycle and inclusion The life-cycle of *Chlamydiae* comprises two different cell states: the elementary body (EB), and the reticulate body (RB). Both differ in their morphology and general functionalities. The only other knwon bacteria with a bi-phasic life-cycle are the *Ricketsiales* which differ substantially from that of *Chlamydiae* [100]. The EB comprise the extracellular, infectious form of Chlamydiae. This form is metabolic inactive, however the Type III secretion system seems to be active as needed for the invasion of the host cell [101]. The EB attaches to the host cell and initiates the uptake of the *Chlaymdium* which is at least partly mediated by the Type III secreted effector TARP that recrutes Actin from the host [102]. After entry, the EB differentiates to RB as metabolic active form, a process called 'primary differentiation'. This process includes the activation of several genes (named 'early' genes) and the building of a separate area, the inclusion, a vacuole like structure typical for *Chlamydiae.* Belland and co-workers identified genes expressed in this phase by exprexssion analysis and found at least eight chlamydia-specific genes activated in this phase [103]. After a lag phase, the bacterial cells start to replicate while the inclusion is expanded and the cells' metabolic activity is at its maximum. Within this time-phase several effectors are secreted by the Type III system but also by other tranport routes as in the case of CPAF [104]. The *Chlamydiae* infer with the Golgi apparatus to obtain lipids from the host [105]. The 'secondary differentiation' creates EBs while the chromosomes are condensed and the metabolism switches to an rather inactive state. This process involves at least 70 genes as identified by Nickolson et al. [106]. Both, the pathogenic as well as the environmental *Chlaymdiae* exhibit this bi-phasic life-cycle. A special feature of the *Chlamydiae* is the inclusion, a vacuole like structure that bears the chlamydial cells and separates them from the host cell. This structure is membrane bound and built up by lipids from the host, mainly sphingolipids. The inclusion contains several chlamydia specific proteins (termed Inc proteins), which are relatively diverse but share a certain hydrophobicity pattern. Many of these are transported by the Type III secretion system. The IncA protein [107], for example, is required for the fusion of 'stolen' vesicles from the exocytic pathway into the inclusion membrane. However, the function of most inclusion proteins are unknown, as well as how nutriens are acquired from the host cell.

1.2.2 The role of computational biology for chlamydia research

Since the *Chlamydiae* exhibt their special life-style, their genomes cannot be genetically manipulated since they are either inactive or unaccessible in the host cell. This fact renders experiments based on genetic manipulation useless [108] and encourages predictive analyses by bioinformatics' means. Important questions to solve include the identification of the genes that participate in the development cycle and infection as by expression analyses and, importantly, by the detection of effector candidates. The amount and function of proteins which populate the inclusion or are effctors is unknown and screens for further candidates included an initial computational analysis [34, 64]. Many chlamydial proteins are of unknown function since the species are relatively distant to well studied organism from which inferences could be made. Especially, the environmental species provide a large fraction of genes which exhibit no detectable orthology to known proteins. This finding motivates an annotation prediction that is not solely dependent on homology searches. The phylogenetic relationships of the *Chlamydiae* to other clades as the *Planctomycetes* is not yet resolved and could give, with the help of bioinformatics, insights into the general evolution of *Bacteria* [109]. Many yet unknown *Chlamydiae* will be isolated from environmental samples by second generation sequencing, which will encourage the development of useful assembly strategies for these species. Importantly, bioinformatics plays a role in the sub-typing of different chlamydial strains due to their pathologic behaviour [110]. In general, bioinformatics can provide shortcuts bye.g. proposing effector candidates. In return, information gained specific for *Chlamydiae* could give insight in the application of bioinformatics' tools when dealing with non-standard organisms.

1.3 Concepts used in this work

In this section, concepts and methods that are used in this work are introduced. For the modeling of the Type III secretion signal, machine learning algorithms (i.e. binary classification algorithms) are employed. Machine Learning also deals with the field of clustering that is needed in this work to group proteins by common origin (in order to get orthologous groups) or by function (in order to delineate functional modules from interaction networks). Different methods to measure and to predict interactions are outlined, as well as their integration into a final, predicted interaction network as generated for the chlamydial genomes in this work. General properties of biological networks are introduced as well as strategies for the delineation of functional modules. The prediction of protein-function using the functional modules is one application used herein, in consequence, basics of function prediction as well as advanced module and network-based concepts on this topic are introduced. The detection of orthologs plays a role in several steps of the network based analyses.

1.3.1 Machine learning

In this section, mathematical and algorithmic concepts used within this work are introduced. These can be broadly subsumed as machine learning algorithms.

Machine learning The field of machine learning comprises statistical, mathematical and algorithmic concepts to extract non-trivial knowledge from data. The applications of algorithms invented in this field are numerous and cover any domain that deals with large amounts of empirical data as the analyses of customers' behavior, prediction of the development of stock prices, automatic analyses of satellite images, text-mining to deduce relevant relationships between entities, and localized services on cellular phones and many more. Many bioinformatics' approaches comprise machine learning techniques as gene prediction, the detection of gene families by clusterings, several classification approaches as automated functional annotation and many more. Two main categories of machine learning techniques exist: supervised and un-supervised learning. Learning means here the generalization of concepts from data which are predictive and/or descriptive. The unsupervised methods analyze the structure of data and report knowledge, e.g. a partitioning or association rules, that has been unknown and for which no *a priori* assumption has been made. Examples of un-supervised learning are clustering algorithms which deduce groups of similar objects from data and hereby report classes in the data that have been unknown before. Supervised methods learn by a given training set which is defined in advance. Binary classification algorithm are prominent examples of this case: they abstract rules from positive and negative training instances of a certain class of interest in a training step, and predict if unseen instances belong to this class. The classification into supervised/un-supervised approaches is fluent and machine learning based analyses may contain both aspects. In this work, representatives of both concepts are used: functional modules are deduced by un-supervised clustering, and Type III effector proteins are predicted by supervised learning.

Clustering Clustering techniques group objects due to their similarity or distance and result in a projection of the objects to groups, then called clusters. The members within a group are ideally more similar to each other as to instances of other clusters. Clustering is one of the key procedures in bioinformatics, and any introduction cannot be complete: as in September 2010, the bioinformatics' journal 'BMC Bioinformatics' lists 1302 papers when querying the term 'clustering algorithm', the older journal 'Bioinformatics' 6262, and the general journal 'Nature' 828. Prominent examples are the delineation of protein families, groups of orthologs, and the detection of functional modules from interaction networks.

Distance measures Initial to clustering a certain set of objects, relationships between the objects must be determined. These can either be defined as distance or as correlation and the definition of a suitable measure is basic to the meaningful application of cluster algorithms. In the example of the clustering of protein families, a suitable measure is sequence similarity that reflects the evolutionary relationships of proteins. In the example of functional modules, the distance or correlation must reflect functional dependency which is, for example, given by the confidence of a measured or predicted functional or physical interaction.

Principal classes of clustering approaches The amount of available clustering algorithms is huge, and many specialized solutions for certain domains of interest exists. The algorithms can be broadly grouped by following properties: approaches with and without *a priori* assumption on the amount of clusters, hierarchical or grouping approaches, and network-based or general approaches. Clusterings which need a given amount of clusters are the K-means clustering and expectation maximization algorithms. In many questions of computational biology, the amount of clusters is unknown and must be estimated in advance when using these algorithms, as in the case of the delineation of modules or protein families. Hierarchical methods do not create groupings but build a tree by joining iteratively objects to their most similar cluster. Prominent examples are provided by the family of linkage clusterings. There a several applications of these approaches for biological data, since many biological entities can be related in an hierarchical structure as evolutionary relations between genes (as by phylogenetic trees) or protein families as in the SYSTERS project [111]. Other problems are solved by partitioning but not by an hierarchy, as the delineation of orthologous groups [112] or most module clustering approaches. Network based methods analyze not only the binary relations between objects but take network properties into account (either directly or indirectly) since this information is meaningful in real-world networks [113]. They are applied if dense regions in a network should be delineated and often rely on principles as information flow or local topology as clustering coefficients. The Markov Clustering algorithm, for example, has been successfully applied in several domains of bioinformatics, including the detection of modules and protein families [114, 115].

Binary classification Approaches of binary classification are supervised methods which abstract knowledge of training instances in order to predict unseen instances. They are therefore often referred to as classification algorithms or classifiers. Their use in bioinformatics is widespread: applications comprise the identification of signal peptides (this work) [116, 117], the identification of tumor sub-types [118], prediction of the origin of EST sequences [119], sub-cellular localization [120], and function [121, 122]. Several different classification algorithms exist based on different mathematically principles. The most simple classificator is the k-nearest neighbor approach which classifies instances related by some distance/similarity measure by the majority vote of k next instances of the training. Neural networks are an analog to the biological neurons and comprise different layers (input, hidden, and output layers, depending on the used variant) with transition probabilities between them. The training of the networks adjusts these transition probabilities to the target function (the classification found in the training). Unseen instances presented the input layer lead then to a prediction of the output layer. Support vector machines (SVM) find the most separating hyperplane in the space of instances defined by the 'supporting vectors'. Unseen instances are then judged by their position in regard of the trained hyperplane. A principle often used with SVMs is the use of so-called 'kernel' functions that allow to project the feature space into a coordinate-system in which instances are separable if this is not sufficiently possible in the original case.

The naive Bayesian classifier is based on the Bayesian theorem and judges the odds ratio of the conditional probabilities to see a certain feature in the positive and in the negative class. The approach is called naive due to the simplification, that different kinds of features are treated as independent from each other in the Bayesian mathematical frame-work, allowing to judge sparse data and saving computational costs. The WEKA toolbox provides implementations of many classification algorithms ready-to-use [123].

Performance measures The performance of a binary classifier must be assessed using independent test sets, which is mostly done by cross-validation. This process divides the set of known instances into equally sized training and test sets, trains with the former one and computes performance measures with the latter one. These measures are based on the counts of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) cases. Typical performance measures computed are Sensitivity $\frac{TP}{(TP+FN)}$ (also called Recall), Selectivity $\frac{TN}{TN+FP}$, and Precision computed as $\frac{TP}{(TP+FP)}$. The receiver operating statistic (ROC analysis) has been originally invented to judge the performance of radar systems in detecting aircrafts causing its' name. A ROC curve is obtained by plotting the true positive versus false positive rate which is commonly obtained by varying a threshold above test instances are classified as positive. So, this plot describes the behavior of the classifier in more detail. The area under the curve (AUC) gives a measure of the classifiers general performance and can be used to compare different approaches on the same data.

1.3.2 Orthology and orthologous groups

The term orthology in the context of genomics describes the descent of a certain protein common to a pair of species by the speciation of their common ancestor in contrast to duplication events (which provides paralogs) or horizontal transfers. Conceptually, two orthologs comprise 'the same gene' in different organisms and, although defined in the context of evolution, two orthologs are often regarded to implement the same functionality in the two different species. This point of view is clearly not correct in all cases since the orthologs might have adapted novel functionalities due to different evolutionary pressures. However, an ortholog is the most probable functional equivalent of a protein in another species. The topic of functional equivalence is reviewed by Eugene Koonin, and in a test on orthologs between *E. coli* and *B. subtilis*, no clear case of functional difference could be found [124]. However, examples exist as for the DnaG primase which acts in different cellular processes in archeal and bacterial cells as described in this review. This functional equivalence indicates the importance to detect orthologous relationships to facilitate function transfer and comparative studies between different organisms. In this work, the identification of functional interactions by genomic context methods as well as a comparative study between the functional equipment of pathogenic and environmental *Chlamydiae* directly refer to methods of ortholog recognition. The main difficulty in the detection of orthologs are *a*) the determination of the correct ortholog in case of several candidates introduced by horizontal transfer and duplication events, *b*) the recognition of absence of an ortholog due to gene loss, and *c*) the detection of orthologs in distant species which are difficult to detect due to the lack of sequence similarity. The most reliable method to cope these problems is a careful analyses of the evolutionary relationship by phylogeny [125]. However, this approach is time consuming and computational costly. The bi-directional best hit method relies on a complete matrix of exhaustive pairwise all-against-all similarity searches and is based on the rational that a reciprocal best hit is the most probable ortholog. This criterion fails in the case of gene loss, paralogs, and HGTs and therefore, problem *a*) and *b*) is tackled by this procedure. To extend this idea, a triangle criterion has been introduced using three instead of two species by Tatusov [112] when creating groups of orthologs. This criterion should further reduce the rate of miss-assignments and is also the basis to create cluster of orthologs which span, in the extreme case, the whole tree of life. The building of clusters of orthologous groups tackles problem *c*) since transitive relations are introduced which join distant orthologs by cluster co-membership without the need of significant sequence similarity between them. Several resources of orthologous groups exist as the NCBI COG database [112] and eggNOG [126]. Other approaches deal with the concept of paralogy by detecting duplication events which are younger as the speciation event as done by Inparanoid [127].

1.3.3 Networks

Networks are a structure that can be found everywhere: the World-Wide-Web, the layout of underground connections, our social networks and many more entities have a network structure. It is therefore not astonishing that networks gained attention in various areas of research from mathematics to economics. In biology, networks are ubiquitous: ecosystems can be described by them as well as the molecular interplay in the cell. The network representation of biological entities gave rise to the field of systems biology by shifting the paradigm away from the interest in only one gene or protein towards their interplay in systems.

Mathematical representation as graphs Networks can be mathematically represented as a graph containing a set of nodes (vertices) that represent objects and a set of edges which form their relationships. Edges might be directed or undirected depending on the kind of relationship they describe and may have either an uniform weight or describe the strength of a relationship by different weights. A graph might comprise only one kind of nodes and edges or several different classes of them. This representation can be directly transfered into suitable data structures which are easily accessible by algorithms on that graphs as adjacency matrices.

Network properties Networks are a vivid subject of research in mathematics, since Erdös introduced this topic in 1960 by investigating random graphs [128]. However, most natural networks do not follow the characteristics of these random model. Barabasi and co-workers developed a mathematical frame-work to describe properties of networks from various domains [129]. Many of the real world networks share general properties [113, 130, 131, 132] and can be described by a scale-free architecture which follows a power law distribution. 'Scale free' in this context means the absence of some network measure which could be used to characterize the network as 'by a typical node' [129]. A very good overview of network properties and their relation to principal patterns of behavior (as robustness and inner structure) of biological networks are summarized in a review of Barabasi and Oltvai [113] which are shortly outlined here: important measures to characterize properties of real-world networks are (beside others) the degree distribution and the average clustering coefficient. The degree (amount of incident edges to a node) distribution describes the probability to get a certain amount of direct partners if randomly choosing a node from the network. In many natural networks, this distribution follows a power law of the form of $P(k)\tilde{}k^{-\gamma}$, but a Poisson distribution in random networks. Typical for the scale free networks is the existence of 'hubs', nodes which connect to many other nodes in the network. Furthermore, these networks are 'small world' networks indicating that each node can be reached from any starting point in only a few steps. The latter property can be related to the exponent γ of the power law distribution and in cases of γ between 2-3, the network is even 'ultra-small' and any node can be reached in only 2-3 steps [129], a property found in many biological networks. As rule of thumb, the properties of scale-free networks (existence of hubs, small world phenomenon, robustness etc.) are valid if the exponent γ is smaller then three. The average clustering coefficient describes the inner structure between the nodes, i.e. their tendency to form highly connected sub-groups. This measure is dependent on the network composition (amount nodes and edges). High average clustering coefficients indicate the existence of dense regions and a modular structure of the network. A similar measure, the function $C(k)$ that describes the cluster-coefficients for nodes of degree k can give further hints to the network structure: hierarchical structures within the network are reflected by a power law distribution of $C(k)$.

Biological networks By principle, any relationship between molecules can be represented as graphs. However, the concept is especially fruitful for certain biological aspects that should be investigated by the interplay of their components. Metabolic pathways, for example, can be modeled as graphs: enzymes and substrates provide nodes, and the reactions comprise (directed) edges in the network. Such networks have found to be scale free [133] when comparing the organization of these networks in 43 different species, indicating a common design principle which is evolutionary conserved. The regulation of cellular processes can be represented as directed graphs since regulation has a direction. Lee an co-workers modeled an regulatory network for yeast and analyzed the existence of network motifs, small patterns consisting of few edges which appear above random [134]. These motifs can be seen as building blocks of functionality in such networks since they implement a certain kind of regulatory behavior. Prominent examples are the bi-fan motif or feed-forward loops. Such motifs can be find in different networks and are characteristic for the kind of network under investigation [135]. The focus in this work is on protein-protein interaction networks. These networks are commonly modeled as undirected graph since the interaction descriptions are undirected (two proteins interact or do not interact), however directions could be introduced to model interactions that lead to modifications. The interaction networks result either from measurements or predictions as described separately. The applications of these networks are various and include function transfer (compare Section 1.3.6) and the deduction of functional modules identified as dense regions in the interaction graph (see Section 1.3.5). These interaction networks can be combined with additional information. Jensen and co-workers integrated the prediction of specific phosphorylation sites into a network of functional interactions. This system, called NetworKIN, allows to identify the probable substrate of a kinase by evaluating the functional neighborhood in the interaction graph. The combined method increases the specificity of the phosphorylation prediction by 2.5 fold [136]. Combination of network information with data of essentiality and expression revealed the importance of proteins that comprise bottlenecks in the interaction network [137] and the combination with structural information revealed different types of hubs (party and date) which interact either concurrently or consecutive with many other proteins since they provide several binding interfaces or only few [138].

1.3.4 Protein-protein interactions

Protein-protein interactions play a key role in the implementation of cellular functionality as most processes comprise an interplay of different proteins. In the first part of this work, interactions between chlamydial proteins are predicted exhaustively from genomic data and provides the basis to deduce functional modules. The backgrounds for interaction measurement and prediction are presented in this section.

Detection of protein-protein interactions in the laboratory The laboratory methods to detect protein-protein interactions (PPIs) are manifold, and the Microbial Protein Interaction Database (MPIDB) knows, at the moment, over 70 different kinds of methods. Among these, the two-hybrid and tandem affinity purification based methods are the two most important contributors of information in this database. The basic principle of the two-hybrid assays is to reconstruct the GAL4 transcription factor which has been divided into two parts (a promotor binding and an polymerase activation domain). These parts are fused to two proteins suspected to interact (the hybrids). Only if these two interacts, the two domains come into proximity and resemble the transcription factor while activating the transcription of a reporter gene. The method can be used for large scale screens using one protein as 'bait' and detecting unknown interactors as 'prey' and comprehensive maps of PPI interactions have been created as for yeast [139, 140]. The two-hybrid assay is capable to detect transient interactions and, especially detect interactions occurring in *in vivo*. Tandem affinity purification (TAP) uses a two step chromatographic procedure to co-purify interactors with their bait protein. To use chromatographic purification of proteins, the latter must be fused to a tag sequence which binds the protein to molecules fixed on a column. In the TAP procedure, two different tags are used which are separated by a protease cleavage site. The first tag is highly affine to the column but the binding can only be resolved while denaturating the protein complex. This problem is solved by the specific protease domain which is used to cleave the detected complex from the column. The second tag allows mild release of the found complexes and is used to wash out the protease. The method can identify indirect interactions since whole complexes are detected and allows to assess the stoichiometry of the found proteins in natural conditions. The former aspect might also be a drawback if only direct interactions should be investigated. To resolve this problem, the MPIDB database adds only interactions predicted between a pair of proteins by this method. Large scale screens with complete proteomes based on variants of this method have been performed for several species including yeast [141] and *E. coli* [142, 143].

Resources of protein-protein interactions The Munich Information Center for Protein Sequences (MIPS now IBIS) provides comprehensive and curated interaction data of *S. cerevisiae* [144] as well as a resource for mammalian complexes [145]. The database of Interacting Proteins (DIP) comprises interactions from various species including two procaryotic set for *E. coli* and *Helicobacter pylori.* BIND, the Biomolecular Interaction Network Database, comprises interactions of various kinds of biologically relevant molecules including proteins [146]. The Microbial Protein Interaction Database (MPIDB) provides comprehensively interactions of bacterial proteins. The resource is based on two principal data-sources: interactions have been curated from literature and bacterial related interactions have been extracted from other data collections [147, 148].

Prediction of protein-protein interactions Functional protein-protein interactions (FPPI; proteins contributing to the same cellular functions, e.g. to a metabolic pathway) often result in an evolutionary coupling of the involved proteins. By systematic comparisons between genome sequences, FPPI can be computationally predicted. Typically, three different methods are subsumed under the term genomic context methods: the detection of gene fusion [149] or fission events, the (conserved) genomic neighborhood method [150, 151, 152], and phylogenetic profiling (also called cooccurrence method) [153]. These methods are in principle applicable to any completely sequenced genome, with the limitation that genomic neighborhood is only well conserved in prokaryotes. Crucial to all these methods is the identification of the 'same' gene in different organisms, which is presumed to carry out the same or a similar functionality in most of the organisms under investigation. Commonly, these corresponding genes are identified in different organisms by orthology-relationships. The most intuitive genomic context method is based on the detection of gene fusion or fission events. A pair of proteins occurring in a number of genomes is predicted to interact if their genes are found fused into one single gene in another number of genomes. If these genomes are not too closely related, the evolutionary conservation of the fused form provides strong evidence for functional coupling and even for physical interaction of the pair of not fused proteins. The (conserved) neighborhood method is based on the observation that functionally interacting genes in prokaryotes are often found in genomic proximity (even in not closely related genomes). Co-regulation of genes in operons and the horizontal transfer of genomic fragments are assumed to evolutionary maintain this co-localization of interacting genes. The phylogenetic profile method is based on the finding that genes, fulfilling a common cellular function, often share the same evolutionary fate. If an organism gains or loses certain functionality, the majority of genes related to this function will also be gained or lost. This common fate is detectable by comparing the phylogenetic distribution of genes in many genome sequences. Functionally coupled proteins are then detected by comparing the pattern of their presence and absence in the genomes. Notably, the information content of these profiles differs, as e.g. house-keeping genes existent in the majority of organisms cannot be functionally distinguished by this method. Profiles with low information content must therefore be filtered out in such an analysis. Since these methods base on different phenomena (common translation, genomic organization, evolutionary fate), it is not surprising that they differ in the amount and quality of prediction. The gene fusion method predicts the fewest number of functional links. However, these are highly reliable due to the strength of the evidence. The conserved neighborhood method returns the highest amount of links, but is only applicable to prokaryotes. The phylogenetic profile method predicts fewer links as the neighborhood method and cannot be applied to ubiquitous or very specific genes. In all three methods the reliability of the prediction depends on the number and phylogenetic diversity of genome sequences supporting the interaction. In addition to these classic genomic context methods that rely on regularities of equivalent genes in different genomes, some other methods exist: the problem of operon prediction is closely related to the prediction of functional coupling since the operon indicates an unit of common regulation implying a common functionality. Operons can be predicted as genes with small intergenic distances on the same strand which form 'gene clusters' [154]. The 'interolog' concept is based on the idea, that two homologs of a pair of known interacting proteins have some probability to interact. Pairs of candidate proteins are detected by homology searches [155] or orthologous assignments [156] of two known interactors in a species of interest and the interaction is transfered. A score which can be further benchmarked can be computed by a function of the sequence similarities found in the interolog compared to the query [155]. It has been found, that several interactions can be meaningfully transfered between different organisms and report known and unknown interactions, even between distant species as yeast an fly [155]. The method relies, as the others, on completely sequenced genomes to detect the most similar pair of proteins. Interactions can also be defined rather between functional domains as between proteins.

Several approaches exist to find pairs of likely interacting domains as, e.g. defined as pair of Pfam domains. Such pairs can be mapped to proteins carrying the domains and are therefore suitable to predict protein-protein interactions by this second step. DIMA **??** is a resource which comprises such domain-domain interactions calculated by different methods including the Domain Pair Exclusion Method (DPEA) [157], the domain-profile method (DPROF), and iPfam [158]. The three mentioned methods are based on different principles. The DPEA method computed a log likelihood ratio of the frequency of pairs of domains found in known interactions against their frequency in different protein interactions. DPROF is based on the same idea as the protein based cooccurrence method: interacting domains are under the evolutionary pressure to be maintained concerted resulting in detectable relationships of the phyletic pattern of their occurrence. iPfam analyses PDB structures and extracts domain pairs which are in strong contact in these structures. Another approach to detect PPIs computationally is the use of text-mining, most commonly by judging the co-occurrence of gene identifiers in abstracts statistically [159]. This approach recovers 'known' knowledge in the sense, that both interaction partners are explicitly mentioned in several studies. However, for eukaryotic species, this is by far the most contributing method in the STRING database (see below) since the neighborhood methods cannot be applied in these organisms [159].

Integration of heterogeneous protein-protein interaction data For most organism, the available interaction data is sparse, and even for the very well studied species as *S. cerevisiae* not all interactions are known [160]. The prediction of functional interactions as well as the deduction by experiments generates data which is partly complementary and could give additional confidence when combined. The challenge in such kind of analyses is the integrative step, i.e. how to make different kinds of measures comparable, and how to integrate them into one resulting network. Lee and co-workers proposed a conceptual framework based on an unifying scoring scheme and demonstrated its' application to create an interaction network of *S. cerevisiae* [161]. The authors observe an increase in accuracy when adding methods and the final integrated network outperforms all the individual networks when tested against an independent test-set of known interactions. This work can be seen as blueprint for similar approaches: This scheme is based on log-likelihood ratios which express the confidence for each method integrated. The likelihood ratios are estimated using a standard of truth (comprising a 'gold-set' of positive and negative interactions) derived independently of the employed prediction methods. Each method results in an evidence which is assessed against these gold sets resulting in a conditional probabilities $P(+) = P(Gold|Evidence)$ and $P(-) = P(\neg Gold|Evidence)$. The gold sets are deduced from KEGG pathways: a link is true positive, if both proteins share the pathway, negative if they are found in two different ones. A prior estimate of seeing a certain interaction is computed as the amount of linkages in a certain pathway against the amount between pathways: $Prior = P(L)/P(\neg L)$. The posterior score can then be expressed as $\frac{P(+)}{P(-)} \cdot Prior$, and is logarithmized to obtain the log-likelihood score. This scoring scheme makes the predictions directly comparable. Several different methods to predict interactions have been integrated including text-mining, co-expression, phylogenetic profiles, and experimental data. The different methods have been integrated using a weighted integration scheme which takes into account dependencies between the different methods (described in the supplemental material) [161]). In general, the likelihood ratios can be integrated by simple multiplication under the naive Bayesian assumption if they are merely independent from each other. Jansen and co-worker showed that even very poor predictors for protein interaction (as co-localization or co-essentiality) can be used to successfully model interaction networks with larger coverage as by the integration of measures PPIs alone. They also showed, that the Bayesian integration outperforms a 'voting' approach in which the highest scoring method defines the interaction [162]. Date and Stoeckert modeled the interactome of *P. falciparum* [163] by integrating co-expression data and phylogenetic profiles under the Bayesian assumption. The resulting network covers 68% of the genome and includes >2000 proteins of unknown function for which functional inferences could be made. They compare the interactome with a prediction for a relative, the *Plasmodium yoelli.* Differences in both networks give hints to the genes responsible for the different phenotypes of the two species. Strong et al. combine different genomic context methods (especially different kind of operon predictions) to predict interactions for the bacterium *Mycobacterium tuberculosis* [164]. In this work, the different evidences are not integrated by a probabilistic approach but intersections of the different methods are computed in order to extract high confidence links. The authors use the genomic context methods to assess different operon prediction methods. The performance of each genomic context method is assessed by common key-word recovery of a predicted pairs of interacting proteins against SWISSPROT key-words. In a further article, the predicted links are correlated with their organization on the chromosome by representing them in a scatter plot of the bacterial chromosome [165]. Hu and co-workers employed the Bayesian integration approach to create a tissue specific interaction network for the retina of the mouse [166] by integrating PPI data, co-expression, and GO term similarity. The nodes of their network have been chosen by literature search and by expression in the retina. In their analyses, they identified candidates for retina related disease genes by linkage to known examples and provide functional predictions. The STRING database [167] integrates interaction predictions from experiments, databases, literature mining, genomic context methods (namely, the cooccurrence method, gene fusion, and conserved neighborhood), and co-expression analyses. The network comprised in STRING has not been established for single organisms but between clusters of orthologous groups [167] which are also used in the prediction process. These orthologous groups are steadily expanded when novel species are added to the system and are maintained and processed in the eggNOG project [126]. The interactions can be projected to a genome of interest by the orthologous relationships. A convenient web-interface allows the navigation through the interaction network either due to the orthologous group network or in the species projection. The scoring system in STRING returns an integrated score of all predictions as well as for each method. The scores are integrated by fitting the accuracy due to the recovery of KEGG pathways as function of the raw scores to Hill equations which are related to each other by a function that describes the equivalence between (i.e. which score of method A results in the same accuracy as score B). The integrated score is then obtained by multiplication of the individual scores for not interacting (resulting in $P(\neg)$) as $1 - P(\neg)$ [168] under the naive Bayesian assumption of independence [167]. The downloadable versions of the database also provide the raw scores as returned by the different methods which, for example, indicate the amount of evidences as for the gene-fusion or neighborhood method or the correlation scores of the phylogenetic profiles. A similar system is the Prolinks database [154]. The main difference between the two system is that in Prolinks, the interaction predictions are computed on the basis of proteins and not on orthologous groups.

1.3.5 Functional modules

Cellular functionalities are mostly implemented by an interplay of several proteins in a modular fashion [169]. These might interact directly as by modification of or binding to each other, or indirectly by a chain of common intermediates. A group of proteins which function together in that way is commonly named a 'functional module'. Possible examples of functional modules are the transcription apparatus, transport systems, and metabolic pathways. A synonymous term sometimes found in the literature is 'cellular sub-system' which mostly refers to the same concept as a part of the cell distinguishable as system from the rest. In addition to the definition as sub-system, the term 'functional module' includes the aspect of a distinct functionality fulfilled by the module. By principle, a module could be defined not only by its' participating proteins but by all additional intermediates of its' related process, however this is rarely done in the scientific community and the focus is set on the proteins as main carrier of cellular functionality. Several definitions of a functional module exist which emphasize different aspects of functional modularity. These aspects may comprise the common functional process, the concerted interplay to fulfill that certain function, the modular structure, and a common evolutionary fate. A possible definition which indirectly includes several of the aspects mentioned beforehand is the detectability of modules as dense regions (clusters) in comprehensive biological networks implying functional and evolutionary constraints on the module. In this work, the definition of a functional module as evolutionary conserved entity comprising several proteins, detectable as natural cluster in biological (i.e. functional interaction) network, and fulfilling a certain biological function is followed. The term 'functional modules' has some homonymous meanings describing different concepts as protein domain composition, network motifs, or cassettes of transcriptional regulatory units which are not regarded herein

The obvious general benefit of the module concept is a possible description of the cell by its general functional components beyond single proteins. Hereby, the clustering of entities (proteins) into groups offers a reduction in complexity since the amount of modules is naturally smaller as of proteins. Since the modules are derived from networks representing large-scale interaction measurements or predictions, they immediately give a reduction in complexity the other way round: proteins can be identified which participate on a certain process and other can be discarded.

The actual entities used as functional modules in different studies vary depending on the set-up, the available data, and the underlying biological question. Possible entities that are not derived by network analysis include (without claim of completeness) pathway definitions and modules generated by text-book knowledge, known operons and transcriptional units defined by common regulatory elements, and measured complexes. Modules have been obtained from network analyses of interaction data predicted by genomic context methods, co-expression, high-throughput interaction screens (as by Yeast-2-hybrid screens), text-mining, regulatory networks, genetic interactions and many more. Often diverse sources of data have been combined int one network model before the clustering procedure. Depending on the input data, clustering procedure, and the actual input data, the 'meaning' of a module varies from functional interaction groups (i.e complexes) to very broadly functionally related entities (as e.g.

larger pathways).

Studies on and with functional modules Meanwhile, a multitude of studies employing the concept of the functional module exist. Where different detection methods are discussed in the next paragraph, some analyses and findings covering different aspects of module related biology are outlined here.

Yang and co-workers identified co-expression modules in mouse and human expression data using a specialized algorithm (the Iterative Signature Algorithm) introduced by Bergmann and co-workers for this purpose [170]. They related the modules to different tissues and compared corresponding modules between mouse and man. The identified tissue related modules showed a great variance in composition (size and members) between the two mammals indicating functional differences in the interplay of proteins between human and its' main used model organism, the mouse [171]. From a conceptual point of view, an interesting aspect of this work is the conditional clustering of modules due to tissue and expression. The clustering of gene expression data has been found to provide meaningful modules by itself [172], and the combination with other data is a logical next step. Tornow and Mewes joined the analysis of expression data in yeast with the comprehensive interaction network available for this organism using the super-paramagnetic clustering algorithm [173]. The study revealed significant correlation of several complexes/interactions with their co-expression while reporting modules which exhibit confidence according to both types of data. While in that study, an integration of expression and PPI data is initially used to detect reliable modules, the consecutive application of expression data to modules is also fruitful: Lichtenberg et al. combined modules from the yeast interaction network with expression data from the yeast cell cycle [174]. The expression data has been screened for genes periodically expressed during the cell cycle. The modules have then been modified according to the appearance of their member genes. The analysis revealed, that the composition of many modules changes during the different stages of the cycle and that some modules are built or modified by an 'just in time assembly' depending on the genes expressed at a certain time-point. In a study of Suthram and co-workers [175] the correlation of different human disease related modules has been investigated by the analysis of gene expression data from cell states of 54 different diseases. The work-flow included the delineation of modules by clustering of high-throughput interaction data, the identification of genes related to a disease by a systematic evaluation of the expression levels different from the normal cell state, and a quantitative correlation of the modules to each disease by

an 'Module Response Score'. In further steps, the authors show that many diseases are correlated by their participating modules and that 'multi-player' modules involved in many diseases are also enriched in drug-targets known to be versatile. Even in the analysis of meta-genomes representing entire microbial communities, functional module can be applied. In a study of Gianoulis and co-workers [176], combinations of modules and pathways have been identified which correlate with environmental traits of the communities. These correlations are called 'metabolic footprints', and are characteristic (i.e. discriminating) for the environments where the samples came from. Li and co-workers assessed the existence of parallel modules (functionally equivalent modules with partly analogous members) in ten different organisms comprising bacteria, archaea, and eukaryota [177]. The analysis revealed the existence of at least thirteen modules which have a parallel counterpart including known and unknown parallel modules. The study shows how the concept of the functional module can be used in comparative studies on data from individual genomes and is a good starting point to understand functional aspects of paralogy and gene duplication.
Mering and co-workers investigated the relationship between modules derived from a functional interaction network and known pathways of the small metabolite metabolism in *E. coli*. Many modules could be uniquely assigned to a certain pathway as represented in the EcoCyc database [178]. The modules tend to cover unbranched pathways with higher accuracy as the branching points of the metabolic network and give hints to possible pathway extensions and yet unknown pathways. Furthermore, some modules participate on several pathways and therefore propose connections between them. Tanay and co-workers created functional modules for *S. cerevisiae* by integrating several kind of comprehensive information including not only PPI data, but also transcriptional and phenotype information [179]. They could show, that modules can be found consistent in these different type of networks. Furthermore, they found a higher structure in the network spanned by the modules (by connecting them due to shared module members) which turned out to be modular in its organization again implying an hierarchical organization of functional modules. In their study, they also benchmarked the performance of module based annotation transfer and predicted GO terms for several proteins.

The evolution of functional modules The evolution of modules has been investigated in three major studies. They key question in all of them is to which extent a module is not only a functional, but also a evolutionary unit. This question has been tackled by comparing the (sometimes partial) existence of a module in several species. Theoretically, two extreme observations could be made: firstly, a module is either completely present or absent as a whole, or secondly, the modules' member proteins are randomly existent in each species. A tendency to the former case is named (evolutionary) cohesive behavior. Snel and Huynen [180] compared corresponding modules in 110 organisms. These modules have been derived from different sources, including MIPS complexes, known operons, known pathways, transcriptional modules. The measurement used for cohesive behavior has been computed as deviation of the fraction of present module members in an organism from the average observation in all species. They found for around 20% of the investigated modules a cohesive behavior indicating a flexible evolution of many functional entities. The influence of noise in the data as introduced, for example, by insufficient orthologous resolution (i.e. by paralogs) has been assessed. The reduction of such influences increased the amount of evolutionary cohesive modules for some sets (as for the pathway modules) but did not change the general trend. The authors conclude, that the observed cohesiveness is therefore an inherent property of modules and flexible modules are not caused by flaws in their detection or by noise in the input data. Campillos et al. determine a module's cohesiveness by assessing the most parsimonious explanation for the joined evolution of the module members given a fixed phylogeny. The algorithm returns the fraction of joined evolutionary events (gene birth and death) and a normalized parsimony score (normalized costs). The distribution of these two variables is used to compute P-Values using a Monte-Carlo approach [181]. This P-Value is then used to categorize the module as 'cohesive' or 'not-cohesive'. Campillos found around 40% of the modules cohesive. Further analyses revealed certain tendencies of cohesive modules as they are larger, enriched in certain molecular processes, and often deal with processes that mediate interaction with the environment as e.g. transporters, but are less frequent horizontally transfered as none cohesive ones. In summary, the study revealed that cohesiveness is indeed a signal which can be correlated to general properties of modules. Fokkens and Snel investigated the cohesive behavior of eukaryotic modules which could be differ from the prokaryotic case [182]. As modules, they investigated pathways and complexes, and for the detection of cohesiveness, they adapted and compared several measures to exclude a possible influence of varying cohesiveness definitions. They found at least 27% of the modules significantly cohesive, depending on the module definitions. As trend, pathways revealed a higher cohesiveness as complexes. An interesting question when dealing with eukaryotes is the influence of paralogs. The authors report a less cohesive behavior for modules comprising several paralogs indicating a negative correlation between cohesiveness and functional divergence introduced by e.g. neo-functionalization of paralogs.

Methods to detect functional modules Functional modules can be detected in biological networks as dense regions. To detect the functional modules *de novo* from these networks, clustering algorithms are employed. The amount of different possible cluster techniques is large, and only a few are introduced here. Many 'classical' clustering approaches as K-means, UPGMA, or hierarchical method have been successfully applied to the problem as by Mering and co-workers [114]. These approaches have been developed to cluster diverse kind of objects given a certain distance measure or metric but do not take any network properties into account unless coded in the distance measure. They are applicable to the problem since simple distance measures (as the edge weight) between the objects in the network can be easily deduced. In the same study, the Markov clustering algorithm (MCL) is tested as alternative [183]. This algorithm is especially dedicated to the delineation of natural clusters' from real-world networks. The algorithm investigates the information flow in the network given by the network topology and enforces edges with an high 'information current' while diminishing edges of low current. While iterating this process, the cluster structure of the network becomes apparent. (For mathematical details, compare the publication [183]). The MCL algorithm has been shown to be very successful in several areas of application including the detection of protein families [115, 184], and also for the delineation of functional modules [114, 185, 186] and can meanwhile be seen as a standard tool in computational biology. MCODE (short for Molecular complex detection) is an algorithm which has been dedicatedly designed for the extraction of complexes from interaction networks [187]. The algorithm evaluates the density of a network around a given 'seed' protein and iteratively expands the seed by neighbors with high clustering structure until a certain cut-off is reached. The candidate neighbors are hereby judged by their participation on dense regions by a measure termed core-clustering coefficient (for details, compare the publication [187]). The method successfully reveals a high number of complexes in different sets. Cfinder [188] is based on the Clique Percolation Method [189] which detects overlapping communities in real-world graphs. These communities reflect modules if applied to interaction networks. An interesting feature of the approach is a possible creation of overlapping clusters. Pereira-Leal and co-workers assessed the use of a line-graph transformation [186] to enable the same effect: due to the line-graph transformation (which can be seen as transformation of edges into nodes), each node of the input graph can participate on several modules with a maximum of the amount of edges it initially had. The line graph is subjected to MCL clustering to deduce the actual modules. The approach has been successfully applied to the yeast interaction network. Tanay [179] et al. employ a bi-clustering approach in an weighted bipartite graph. Such a graph represents proteins and properties (as co-expression or interaction) as different kind of nodes and allows to detect significant clusters (sub-graphs) according to all kinds of properties used (compare Tanay et al. [179] for details). A pruning procedure removes overlapping clusters above a threshold to avoid a grouping to be reported several times while still allowing overlapping modules.

A complementary approach to the *de novo* detection of functional modules by clustering is the application of 'network-alignment' tools. These are applicable if a module itself is already known and should be searched in another biological network. Examples are the Path-Blast tool for linear modules (i.e. pathways) [190] which needs as input a sub-network comprising the module, and a target network with sequence information to search against. It integrates similarity searches with the given topology information given by the input and returns an ordered list of the best matching pathways. The hits are determined by a dynamic programming approach which finds the most probable path alignment with the lowest cost analogous to a pairwise sequence alignment. In a succeeding work, the authors describe a related method to handle not only linear pathways but also complexes [191]. Torque (the topology-free querying algorithm) [192] is an extension to the idea which does not need an input network to return reliable results in the target network.

Curated Module and Pathway definitions and resources To assess the quality and properties of functional modules, defined functional entities must be employed. Commonly used data-compilations for these purposes comprise curated module data-sets, pathway definitions, or compilations of known complexes. Curated pathway models for several organisms can be found in the KEGG database [193]. The pathways are initially defined by their participating orthologs and descriptions of the reactions between them. KEGG provides an own system of orthologous groups which is used to map the pathway definitions onto an organism of interest. All ortholog and pathway definitions are subject of constant improvement of the KEGG annotation team. Recently, the resource (which also contains more detailed information on metabolites, reaction types and drug classifications) has been extended by functional module definitions. Both, the pathway as well as the module definitions have been used in several related studies [176, 180]. An alternative database of metabolic pathways is EcoCyc [178] which is a comprehensive

resource for the small metabolite metabolism of *E. coli.* MetaCyc is an extension of the EcoCyc system to a large variety of microorganisms. Both resources are constantly curated [194].
The Seed project [195] provides a data-source for pre-defined cellular subsystems that are used for annotation. These sub-systems are conceptually near to functional modules. Albeit neither the modules in KEGG nor in the Seed are extracted automatically from a network but are defined by human curators. In consequence, they should exhibit significant overlap but no perfect coincidence with automatically extracted modules from biological networks. The DICS repository [196] provides an example of a special purpose database based on modules: it contains human disease related modules and genes which have been found co-clustered with known disease factors in functional modules. Curated sets of protein complexes are also of importance for the research on functional modules, since the complexes can be seen as functional entities of different proteins working functionally closely together. The CORUM database, for example, provides curated mammalian complex data which is by hand curated and exhibits experimental evidence [145]. Yeast complexes can be found in the MIPS yeast database which is one of the most comprehensive resource for curated complex and interaction data so far [144, 197].

1.3.6 Annotation of proteins

The functional annotation of proteins is one of the key applications of bioinformatics since the large amount of novel sequences cannot be characterized by laboratory methods. In this work, the use of functional modules for the annotation of chlamydial proteins is assessed. In this section, concepts and approaches of functional annotations are introduced.

Controlled vocabularies and ontologies The definition of function in the context of annotation is principally broadly defined and may include any level of resolution from molecular mechanisms to general cellular functionalities. Where it is necessary to attach any of such information to a protein sequence in order to make this information available to the scientific community, an unstructured annotation induces problems: annotations cannot be compared between different sets of proteins which renders comparative analyses impossible. Furthermore, computational approaches cannot make use of unstructured annotation since algorithm cannot 'understand' free text. A solution of this problem is the use of controlled vocabularies and biological ontologies. By principle, controlled vocabularies consists of a defined set of terms describing a domain of interest (i.e. protein sequences). Ontologies are an extension of this concept introducing relations between the terms of a vocabulary. In many cases, these relations reflect hierarchical dependencies of objects, but this is no obligate rule. In the field of protein annotation, a frequently used controlled vocabulary is the compilation of SWISSPROT keywords used as part of the annotation process by the UNIPROT consortium [198]. This vocabulary comprises 1063 terms (as from 19th of August 2010) that are assigned to ten major categories listed in Table 6.4. These categories do not imply a hierarchy or define term relationships but qualify the general kind of the described entities, therefore, these keywords are a controlled vocabulary but not an ontology. Twenty-three functional categories have been defined by Tatusov an co-workers to functionally categorize the Clusters of Orthologous Groups (COGs) [?, 199] describing different cellular abilities as 'Transcription' or 'Defense mechanism' listed in Table 6.2.

Other classification systems provide relationships between their entities and represent ontologies rather than controlled vocabularies. The Enzyme Commission numbers (EC numbers) describe the domain of enzymatic reactions encoded in a string consisting of several numbers [200]. Each position of the string represents a certain level of functional granularity separated by a delimiter and therefore reflects an hierarchical order. The Gene Ontology (GO) initiated by the Gene Ontology consortium is a community based project to create a comprehensive ontology of gene products. The GO ontology terms are categorized int three major areas: cellular component, biological process, and molecular function. GO terms might be interconnected to any other related term building up an acyclic graph of term relationships. At the moment (as from 19th of August 2010), the GO ontology comprises 32168 different entries. These amount of terms is huge and is a result of the community approach used to built up the ontology. Reduced sets (named GO slim) exist containing a sub-set of GO terms. Some of them are taxon specific (as for yeast or plants) or developed for special purposes (as UniProtKB-GOA, created for annotation in UNIPROT). The MIPS Functional Catalog (FunCat) is an hierarchical ontology which contains 1083 categories in 28 main categories, listed in the supplementary Table 6.3 (version 2.0). The FunCat categories are hierarchically organized from less to more specific terms with a maximum of six levels of different granularity. This hierarchy is encoded in the FunCat identifier: for each level, two digits are reserved which refer to a certain entry in the FunCat schema. In the encoding, these levels are separated by a delimiter as in the case of EC numbers. The FunCat has been initially developed during the yeast genome project and has been extended to other organisms by adding categories not needed for yeast annotation. Reference annotations for following organisms exist: *S. cerevisiae*, *Neurospora crassa*, *Arabidopsis thaliana*, *Homo sapiens*, the microbial genomes of *Thermoplasma acidophilum*, *Bacillus subtilis*, *Helicobacter pylori*, *Listeria innocua*, *Listeria monocytogenes* [201, 202], and for *P. amoebophila* UWE25 created during its' genome project [71]. The strict hierarchical structure of the FunCat allows to chose adequate levels of granularity due to the problem tackled e.g. for the analysis of functional modules. Another advantage is the much smaller amount of different terms than it can be found i.e the GO ontology. This smaller size eases the annotation process since the selection of the correct category is simpler, and leads to a clearer and more coherent annotation since the FunCat is merely free of synonymous entries. For a recent essay on the topic, see Jensen and Bork [203].

Concepts in bioinformatics for functional annotation A newly sequenced gene does only provide its' primary sequence which is not informative in terms of the functionality performed by the protein encoded in it. The knowledge about a protein sequence can be enriched by gathering information that is either already known and accessible in the literature or is deduced by appropriate computational analyses. This process, commonly described as functional annotation, can either be done by human experts with the aid of diverse bioinfomatics tools and databases or in a completely automatic fashion by an annotation software. Common to both approaches is the deduction of properties by comparison to already characterized proteins. In general, the annotation by human experts results in a better quality annotation as automatic methods but is very cost and time intensive compared to the automatic methods. For an excellent review on this topic, see the article by Dmitrij Frishman [204]

The most used principle is the annotation transfer (i.e. learning from a related sequence) due to homology since common origin often implies common or similar function. Homologs are detected by similarity searches: if two protein sequences share significant more similarity as by chance they should be evolutionary related and this homology implies a common or similar function [205]. Several programs to detect similar sequences from a sequence database exist, most commonly, the BLAST [206] heuristic or its' derivate PSI-BLAST is used. Pre-calculated all-against-all similarities computed by a community grid can be obtained from the SIMAP [207, 208] database saving computational costs. Since this database has been used at several occasions in this work, it is explained in more detail below. An extension of this approach is the use of orthologous groups: orthologs (genes of common origin aroused by a speciation event) often represent 'the same gene in different organisms' which makes a correct function transfer more likely [199, 127]). Orthologous groups merge orthologs from several organisms and improve the benefits of the orthology concept for the application in several kind of analyses including functional annotation transfer. Another valuable resource for function prediction are domain descriptions as, for example, contained in the Pfam database [209] which can be detected over large evolutionary distances using profile or Hidden-Markov model representations of known domains. The Interpro project [210] unifies several of such domain databases providing an excellent resource for domain based annotation. However, these approaches can only be applied if *a priori* knowledge of either homologs or domain signatures is available. In consequence, sequences with novel properties (orphans) cannot be assessed by them. Functional links as well as measured interactions can give hints to a proteins' functionality by the 'guilt by association' assumption: if two proteins functionally interact they are likely to participate on the same cellular process. In consequence, an appropriate function can be transfered if significant (functional) interactions between an uncharacterized and a known protein exist [211, 159, 212, 185, 213]. The same principle implies the use of functional modules instead of links to define which proteins are functionally related and provide information for annotation [173, 185, 114]. The module as well as the link based approach cannot reach the functional resolution of a individual annotation by e.g. a functional domain: homology or domain based approaches may describe molecular mechanisms (as, e.g. a kinase activity), functional links/module based approaches do, by definition, detect relations between proteins, they are therefore useful to relate unknown proteins to functional units as metabolic pathways and other cellular sub-systems or processes and to transfer more general terms of function. Some approaches to make use of functional interaction networks and modules for annotation are described below, and an excellent review on this topic has been written by Roded Sharan [214].

SIMAP: a comprehensive resource of pre-calculated sequence similarities and domains The **Si**milarity **Ma**trix of **P**roteins [207, 208] comprises pre-calculated similarity searches of all publicly available genomes. The precalculated data comprises pairwise alignments computed using the Fasta3 algorithm [215] and comparisons to domain-signatures as they can be found in Interpro. The database allows fast retrieval of this data via a Java based middle-ware or Web-Services. SIMAP allows to define search-spaces that comprise an user defined sub-set of the sequence-space by selecting certain primary data-sources or taxonomic clade. For example, a *Chlamydiae* specific search-space can be defined which contains only sequences that are existent in these species. Only alignments in the search-space are then reported when querying SIMAP and database-dependent scores as the alignment E-Value are re-calculated according to the search-space on-the-fly. SIMAP also contains the compareDB, an exhaustive repository of bi-directional best hits which can be used to process own clusterings of orthologous groups as it has been done in this work in Section 2.2.1. Although not a solely dedicated to function annotation, SIMAP supports the processes of automated and by hand curated annotation as it short-cuts similarity and Interpro searches.

Annotation and interaction networks The use of the 'guilt by association' principle to infer protein function from interaction data can be extended in manifold ways. Joshi and co-workers propose a system called 'Gene Function Annotation System' (GeneFAS) which incorporates *S. cerevisiae* interaction data of several sources (complex data, co-expression, physical interaction screens) by an Bayesian integration approach while judging their reliability due to different GO categories [216]. This is done by computing a reliability score as the product of *a priori* probabilities for each source of interaction according to a shared GO term at a certain level of granularity and propose for two third of the unknown proteins in yeast functional assignments. This method exploits the network information locally since only the direct neighbors are taken into account. Vazquez and co-workers propose a simulated annealing based approach to optimize function transfer in a network globally [217]. Due to the global optimization, the method captures transient information which is not encoded in the direct neighbors but in information of remote nodes of the network. The simulated annealing procedure automatically results in different near optimal solutions which offers a convenient way for multiple functional assignments for a protein. The authors show, that the reliability of a functional transfer (although globally optimized), is depending on the degree of the node (i.e. amount of interacting partners, but not necessarily the amount of annotated direct interactors) and the more neighbors of a protein exist (i.e. the more local structure in the network), the better the prediction performs. In any case of degree >1, they reach an accuracy of 60%-70% in the Yeast interactome. Deng [218] and co-workers use Markov Random Fields as mathematical backbone to infer function by network-propagation. The method models a belief propagation network which represents the initial probability of a functional labeling and uses Gibbs sampling to obtain a posterior distribution (i.e. the predictions) of the labeling due to the initial labeling after seen the interaction data. The key feature of the approach is that, although modeled globally, nodes nearby in the interaction network have greater influence on the annotation as more distant nodes. Initially investigated on the yest interaction network only, the authors extended the approach for the use on multiple networks [219]. Nabieva et al. [220] iteratively simulates a step-wise flow of information in the interaction network. This procedure automatically takes into account the network topology while propagating information about annotation from annotated to unknown proteins through the network. The method slightly outperforms others as majority voting and a variant of the simulated annealing approach. The authors show that incorporation of additional networks (e.g. genetic interaction networks) and the modeling of the edges due to their reliability improves the annotation accuracy of their method.

Annotation and functional modules The functional modules are entities which reflect proteins functionally more related to each other as to other parts of the proteome. Therefore they are a resource that could intuitively be used in annotation. Modes of such possible use could comprise the correction of miss-annotated proteins, the detection of proteins related to a certain process, and the use in an automated annotation transfer from known to unknown genes. In the literature, the use of functional modules for annotation is often implied but rarely a main focus. Only few studies exist which explicitly deal with the topic. Tanay and co-workers [179] uses module information deduced by their SAMBA software to propose general GO terms to more than 800 yet not annotated proteins in *S. cerevisiae.* The annotation work-flow in this study comprised following steps: firstly, the detection of modules with certain GO terms enriched, secondly, the assessment of their predictional power by a five-fold cross-validation, and finally the extraction of annotation for yet unknown proteins. The detection of modules with significant enrichment automatically determines the set of GO annotations useful for annotation transfer by these modules while discarding non-informative GO terms. The resulting predictions refer to more generally GO terms as ribosome biogenesis or sporulation as it can be seen on the SAMBA web-page. This can be interpreted as in congruence with the module concept which itself describes function more general as for a single protein. The cross-validation resulted in partly high specificity of the predictions with 40%-100% depending on the functional class. The prediction for five unknown proteins as sporulation related has been tested in the laboratory, and in four cases, the prediction could be verified. Song and co-workers [185] compared the performance of module clustering algorithms in terms of function prediction and recovery of known complexes/modules, in physical interactomes of *S. cerevisiae.* Furthermore they compared the performance of module based function prediction with a simple network based approach. In the study, different yeast interactomes (one containing genetic and physical interactions, one with all physical interactions, one based on yeast-2-hybrid screens, one on other high-throughput techniques) have been subjected to six different clustering algorithms which are commonly used to detect functional modules (Network-Blast [221], MCL [183], Cfinder [188], DPClus [222], Mcode [187] and a spectral clustering method [223]). The comparison of the performance for each method revealed that no method is optimal in every combination, for example, MCL outperforms the other methods on sparse networks but is slightly worse in others. The authors introduce a framework to assess the performance of clusterings in a comparative manner and introduce quality measures in terms of the ability to predict function and to recall known functional groupings. The comparison with a simple network based approach revealed a better performance of the latter one compared to any module clustering indicating that direct neighbors in the interaction network are more informative. Other studies do not explicit assess function annotation but imply the value of modules for such purposes. For example, the study of Mering and co-workers can be used to improve the annotation of metabolic pathways as supposed by the authors [167]. The concept of functional modules can be not only used by the generation of candidates (i.e. function prediction) but also as direct annotation aid: in the Seed annotation project [195], the task of functional annotation of complete bacterial genomes is split up in sub-tasks defined by typical subsystems. These sub-systems from several species can then be curated in a much faster manner by an expert in the field of the sub-system system, a principle adapted from industrial production lines [224]. This allows a fast annotation of high quality in a community based project and the technical platform is therefore called RAST (Rapid Annotation using Subsystems Technology).

2 Creation and application of chlamydial functional networks and modules

2.1 Motivation

Due to the relatively small size of the scientific community dealing with *Chlamydia*, their evolutionary distance to the well studied model organisms like *E. coli*, and their restricted accessibility by laboratory manipulation methods, many of their molecular mechanisms are unclear. Especially, the function of many proteins and most of their interactions are unknown. For example, the Microbial Protein Interaction Database lists only 17 chlamydial interactions, where for other species exhaustive screens are available [147]. This motivates the creation of a comprehensive (functional) interaction network for *Chlamydiae* for the generation of prediction based hypotheses on the interplay of their proteins.

This has been accomplished the implementation of a general pipeline to integrate different functional prediction methods for unseen genomes and arbitrary kinds of functional link prediction approaches. This system should enable the prediction of functional interaction networks of *Chlamydiae* which are reliable and cover a large fraction of the proteomes in order to deduce functional modules for further analyses. Furthermore, it is intended to study a possible way to distinguish between *functional* and *physical* interactions by the use of different background models.

The concept of functional modules allows an as well comprehensive as condensed view on an organisms' functional equipment. Furthermore, the functional grouping of proteins by the modules can be applied to characterize proteins by their module co-membership. In this section, the application of these principles to relevant questions in chlamydial research are described. The first two parts describe the delineation and benchmarking of chlamydial specific functional modules including an assessment on the recovery of known groupings from the KEGG database. The other parts present chlamydia-specific applications of the functional modules. These tackle their use to annotate unknown chlamydial proteins, to detect virulence related entities, and to enlighten the evolution of function due to different stages of host adaptation in *Chlamydiae.*

2.2 Chlamydia specific interaction networks

2.2.1 A Pipeline to create functional interaction networks by bayesian integration

The integration of several function prediction methods into one network has been implemented using a Bayesian integration method as used in several other studies described in Section 1.3.4 [225, 162, 163] by a pipeline which allows easy integration of additional sources of information and can theoretically be used for any groups of genomes as long as sufficient data is available. The pipeline is pictured in Figure 2.1 and the single steps are described in the next paragraphs. The pipeline includes following steps:

I) **Creation of the background models**. These models are used to score the performance of different interaction prediction methods against known knowledge. A background model comprises negative and positive instances of prior functional and physical interactions.

II) **The prediction component.** This component of the pipeline retrieves interaction prediction scores for a certain genome. This includes data from following sources: the genomic context methods and expression data from STRING, and domain-domain based interaction predictions. This step also includes the mapping of new protein sequences to orthologous groups as found in eggNOG [126] and the creation of novel groups comprising the proteins of the environmental *Chlamydiae.*

III) **Binning and scoring procedure**. Continous scores reported by the prediction methods are binned into discrete values as needed for the input to the Bayesian integration. The values are then re-scored according to gold standard sets resulting in scoring schemata for each method. The re-scored values are then integrated into a final score under the naive Bayesian assumption. These steps includes the definition of a prior probability assumption that defines the cut-off used to prune

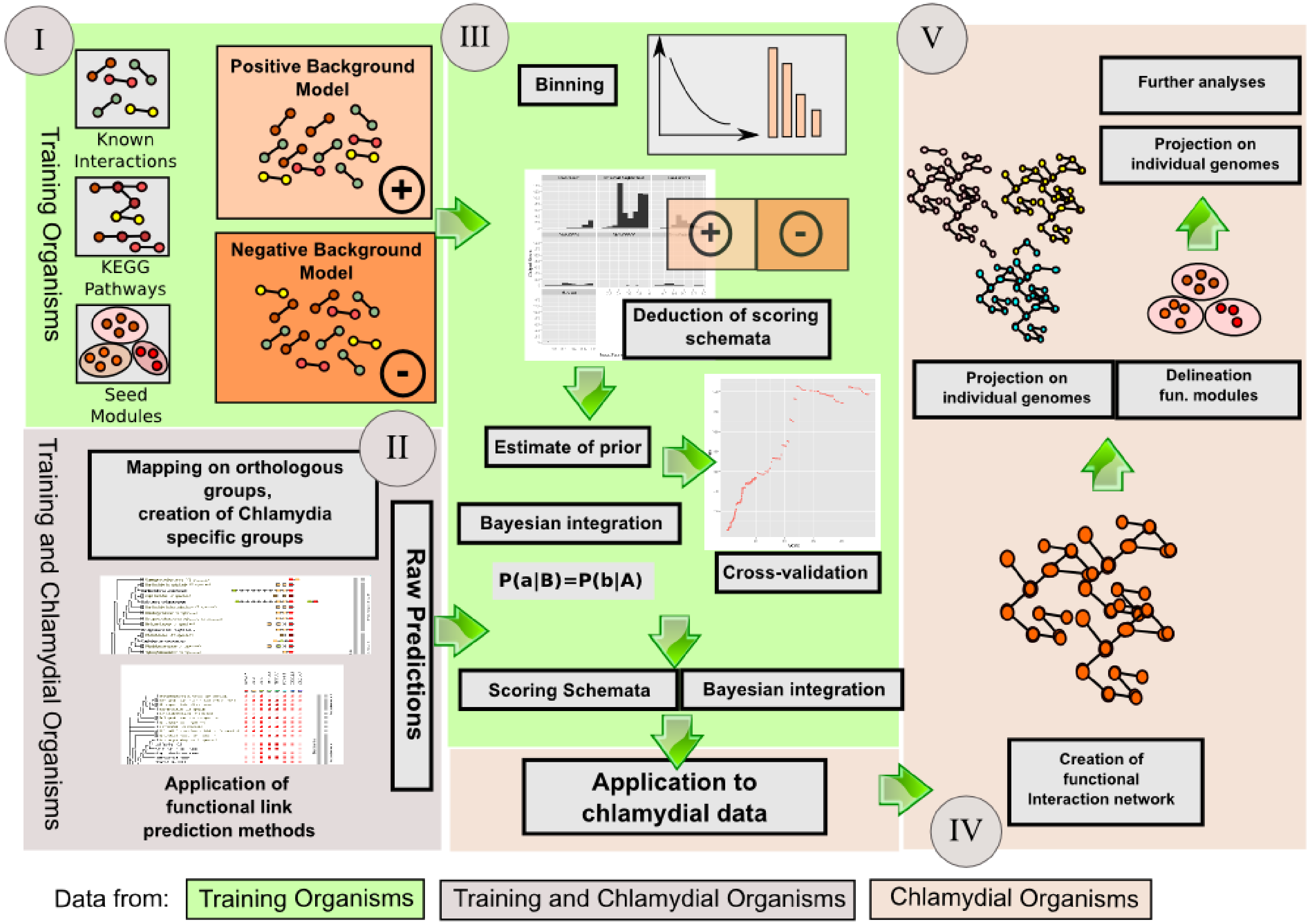

I
Training Organisms
Known Interactions
KEGG Pathways
Seed Modules
Positive Background Model
+
Negative Background Model
-
II
Training and Chlamydial Organisms
Mapping on orthologous groups, creation of Chlamydia specific groups
Application of functional link prediction methods
Raw Predictions
III
Binning
+
-
Deduction of scoring schemata
Estimate of prior
Cross-validation
Bayesian integration
P(a|B)=P(b|A)
Scoring Schemata
Bayesian integration
Application to chlamydial data
IV
Creation of functional Interaction network
V
Projection on individual genomes
Delineation fun. modules
Projection on individual genomes
Further analyses
Data from:
Training Organisms
Training and Chlamydial Organisms
Chlamydial Organisms

Figure 2.1: The functional network generation pipeline. The roman literals and the arrows represent the order of steps, the background-color the used organism set, boxes the intermediate results (for detailed explanation, compare text)
.

poor predictions. The overall performance is investigated by a five-fold cross-validation with predictions from the organisms comprising the background models for validation.

IV) **Prediction of the actual chlamydial functional interaction networks**. This part of the pipeline generates the chlamydia-specific networks using the scoring schemta and chlamydia specific predictions.

V) **Further investigations and processing**. The interaction network are investigated due to their network properties and then further processed by clustering into functional modules.

Material and methods

Used data sets and software The publicly available chlamydial genomes have been downloaded from RefSeq version of May 2010 (The used data-sets have been updated regularily, the given dates refer to the latest of recent update). For the non-public genomes, the gene-prediction pipeline that was implemented in-house by Patrick Tischler has been applied. This pipeline integrates different intrinsic prediction methods with extrinsic homology information to improve the annotation accuracy of gene-starts, the resolution of overlapping genes, and the detection of shadow-ORFs (for details, please refer to PhD thesis of Patrick Tischler). Physical interaction data from diverse bacteria has been obtained from the microbial protein interaction database (MPIDB) [147] as from January 2009. The data has been filtered as follows: data from organisms which contribute less than 100 interactions have been discarded since they would not result in sufficient coverage for the delineation of backgrounds described below. Only species which are also processed in the STRING database have been regarded. This filtering results in a set of interactions from four species (some interactions have multiple evidences): *Helicobacter pylori* (2068 measurements, 1650 interactions), *Escherichia coli* (3038 measurements, 2262 interactions), *Campylobacter jejuni*, (17390 measurements, 11858 interactions), and *Bacillus subtilis* (4366 measurements, 242 interactions). Orthologous groups and KEGG pathway information have been extracted from the STRING database (version 8.1) as from January 2009 [159], as well as the raw scores of functional interaction predictions from the fusion, the phylogenetic profile, the co-expression, and the conserved neighborhood methods. Notably, the orthologous groups in STRING are identical to the groups found in its' 'sister-project', eggNOG

[126]. Module data has been downloaded from the Seed project[195] as from October 2009. The Seed module data has been curated by removing ten entries that do not clearly refer to a functional module as indicated by their description. Further redundancies in the Seed modules which might occur due to the community based annotation approach of the Seed-project have been checked and modules with an overlap >75% to another module have been removed. The entries in the Seed data have been mapped onto the orthologous groups in STRING by sequence identity to at least one orthologous group member to allow the retrieval of orthologous modules for any species in STRING. This procedure enables to use Seed module definitions organisms not yet or insufficiently annotated in the Seed.
Predicted domain-domain interactions defined as pairs of Pfam identiers have been downloaded from the DIMA 2.0 resource (version as from January 2010). Pfam domains for each protein have been retrieved from SIMAP [207, 208] (version as from January 2010, as processed by the Interpro pipeline of SIMAP). Properties of the resulting networks have been investigated using the NetworkAnalyzer plug-in in Cytoscape [226] plug-in [227] and the tYNA web-server [228].

2.2.2 Detection of orthologous relationships

Since in this study several non public genomes not yet represented in the orthologous grouping in STRING are processed, the orthologous relationships of them must be determined in order to make use of the data from the genomic context methods as well as for comparing the resulting data between the chlamydial species. This has been achieved in two steps: clear orthologs of proteins in already known orthologous groups have been assigned to the latter. Proteins which could not be assigned have been tested if they would form orthologous groups of their own.

Extending eggNOG The genomes of the pathogenic *Chlamydiacea* are merely processed in the used STRING version and could be directly mapped to the orthologous groups. Proteins which might be specific to certain strains not in STRING (as the koala related *Chlamydophila pneumoniae* LPCoLN [86] as well as the non-public data-sets have been separately processed. One of the basic concepts in the delineation of orthologous groups is the use of bi-directional best hits (BBH) between complete sets of proteins of two species. Since the test on the BBH criterion for each pair of chlamydial and species in eggNOG would be very costly, the concept has been adapted for the mapping procedure using the following short-cut: the detection of the best hit in the complete set of eggNOG sequences represents the sequence which is the closest relative. This closest relative has the highest probability of being a bidirectional best hit compared to all other sequences in eggNOG and the corresponding orthologous group could be extended by the novel chlamydial sequence. The sequence of the best hit is then tested if the bi-directionality criterion holds in respect to the primary query sequence. This approach can be described as using the eggNOG database as a 'ueber-genome' comprising all possible sequences, an approach that allows to incrementally update the orthologous groups without the need for any re-clustering. In-paralogs have been detected as sequences which are more similar to the main ortholog as to any sequence in eggNOG. Conflicts which are introduced by in-paralogous sequences with BBH assignments (in total, 5 cases of them could be detected) have been resolved. The cut-off set to find the best hit in eggNOG and vice versa are $>$80% sequence coverage and an E-value better $10e^{-5}$. The mapping has been implemented using SIMAP by defining own search-spaces (compare Chapter 1.3.6).

Chlamydia specific orthologous groups Proteins that could not be assigned in the first step to eggNOG have been subjected to a simple procedure to detect possible groups specific to the environmental *Chlamydiae*. For this purpose, bidirectional best hits between their genomes and one representative of the pathogenic species (*C. trachomatis*) as provided by the CompareDB section of SIMAP have been subjected to a connected component clustering using the in-house Superphyler software (settings: E-value better $10e^{-5}$, triangular criterion turned off, $>$50% coverage of the sequence lengths, no initial out-group pruning) The resulting groups have been compared to the eggNOG assignments and novel groups with no connection to eggNOG have been kept separately, the others have been merged with the corresponding eggNOG clusters.

2.2.3 Detection of Functional Links

The derivation of predicted functional interactions used in the pipeline can be divided in two strategies: firstly, mapping of predicted interactions from STRING, and secondly, by the deduction of functional interactions by direct application of methods not provided in STRING.

Mapping interactions from STRING The STRING database comprises functional links predicted by several methods (compare Chapter 1.3.4). Since the use of the intended interaction network is the delineation of functional modules as detectable functional and evolutionary units (compare the module definition given in Chapter 1.3.5), methods as text-mining or pathway co-membership that rely on curated knowledge have been discarded from this analysis. In addition, methods based on data used to built up the background-models (physical interactions providing Interologs or KEGG pathway co-membership) have been discarded since the background models must be independent from the prediction methods assessed by them. The methods that fulfill this criterion are: the fusion method, the conserved neighborhood method, the cooccurrence, and the co-expression method.

The local installation of the STRING database allows to deduce raw scores as reported by the individual prediction methods. These scores are not altered by the STRING scoring scheme and therefore suitable for the processing in the pipeline using own scoring mechanisms.

The interactions have been extracted as pairs of orthologs with the respective raw scores. Using the orthologous assignments deduced as described above, the interactions have been projected on the genomes used in this study (all *Chlamydiae* and the organisms used to create the background models). This projection automatically discards interactions which cannot be implemented by the individual genomes if one interaction partner is missing.

Chlamydia-specific neighborhoods The method that generally provides most links of all genomic context methods is conserved neighborhood (ratio to the next best method 'Cooccurence' is around 4:1). It is therefore probable that the environmental *Chlamydiae* which are not processed in STRING could contribute yet unknown links (which should loosely refer to chlamydiae specific operons). I implemented a simple neighborhood approach to delineate functional links which have only evidence in the environmental *Chlamydiae*. By principle, the resulting links can only have weak statistical support in comparison to functional links in STRING which cover several hundred genomes. However, some predicted links which, i.e. these detectable in all four environmental species, could lead to low scoring but valid functional links in the resulting network. This is even more likely if combined with the raw counts from STRING.

To deduce such links, I implemented following procedure:

- for each gene g in each genome, enumerate the neighborhood N with distance n on the same strand. The distance is counted as gene coding regions (i.e. n=1 are the two neighbored genes of g).
- for each gene g find orthologous genes g_1, g_2... in each other genome.

- for each gene g' in g_1, g_2..., get neighborhood N' of distance n.

- detect the overlap between each neighborhood N' with N. If overlap exists (s by genes a,b..), report an evidence for interaction between g and a, and g and b.

- summarize the found evidences into a report which can be further processed.

The found evidences have been added to the counts found in STRING in cases where the detected interaction has already been known, or has been added to the as novel prediction. The algorithm is parameterized by the distance n which has been set to n=2. This choice of the parameter allows the existence of one 'intervening' gene and should be rather restrictive. The detected conserved neighborhoods are only a signal for functional couplng if the genomes under consideration are not too closely related and have undergone significant re-shuffling of their gene order. To inspect this, I plotted the genomic sequences for each combination of chlamydial genomes using the Gepard Dotter software package. Stretches of conserved genes appear as lines in the dot-plots. A manual inspection revealed a high conservation of gene order in several cases of the pathogenic but not between the environmental *Chlamydiae* and also not between any pathogenic and environmental species. Example plots for both cases are given in the supplementary Figures a and b. In consequence, the simple neighborhood approach should be meaningfully applicable to the environmental set. The algorithm has been applied using the orthologous groups mapped and delineated as described above on a set comprising the environmental species and *C. muridarum* Nigg. The latter has been added as representative of the pathogenic *Chlamydiae*.

Deduction of domain-interaction based predictions Interactions can not only be defined between protein but also between functional domains which often mediate a PPI. Data from prediction methods of such domain-domain interactions have been collected in the DIMA resource. These domain interactions are defined as interactions between two Pfam domains and can be used to predict interacting pairs of proteins by selecting all pairs of proteins that carry one of each Pfam domain (self-interactions were excluded). The resulting pairs have been scored by the score of the predicted domain interaction. The predicted link has also been established between the orthologous groups corresponding to the participating proteins. Following methods from DIMA 2.0 have been used: the domain pair exclusion data (called DIMA DPEA) [229], the profile based method [230] (DIMA DProf), iPfam [158], and 3did [231]. The latter two have been combined into one method since they reflect the same prediction principle (called

DIMA STRUC). Since the obtained values for these methods from the DIMA resource are either '1' (prediction exist) or '0' (does not exist), the combination of the latter two has been done by counting the support by both methods (1=3did or iPfam prediction, 2=predicted by both). For the other, the initial score has been further used.

2.2.4 Creation of scoring schemata

Gold standard sets The definition of gold-standard sets of interacting and non-interacting proteins is not trivial due to the limited knowledge of the real interactomes. A typical approach in eukaryotes is the use of proteins located in different compartments of the cell which are very unlikely interaction partners, as, for example done by Jansen and co-workers [162]. Clearly, this cannot be applied on bacterial data since the prokaryotic cell has no compartments.
Herein, a stratgey to create gold standard sets based on pathway (from the KEGG database) and module definitions (from the Seed project), as well as known interaction data (as comprehensively collected in the MPIDB database) is proposed. This approach generates background models which are further used to delineate the gold sets. Since the amount of validated interaction data in *Chlamydiae* is very sparse, and only 18 examples are reported in the MPIDB database, positive gold sets that show sufficient overlap with the predictions cannot be meaningfully deduced for *Chlamydiae*. In consequence, data from bacterial species which have sufficient interaction data has been employed to create (bacterial specific) gold standard sets. Negative gold-sets are created by the assumption that proteins in different pathways and functional modules are less likely to interact. based on these considerations, two different background models, a “functional module” model (FM) and a “physical interaction” model (PI) have been created. Each background model consists out of true positive binary physical or functional protein-protein interactions (TPI) and true negative non-interactions (TNNI). To define TPI and TNNIs, the used input data has been processed to obtain groups of most likely interactors and most likely 'non-interactors'. This has been done by following procedure: for each species with sufficient high amount of known interactions, three input-graphs have been built up (a Seed-module, a KEGG pathway, and a PPI interaction graph). Each protein represents a node in the graph. A binary interaction with uniform weight is drawn if two nodes share a common interaction in the respective set defined as a measured interaction for the PPI graph, or, a common module or pathway membership in the two other cases. The graphs have been clustered using the Markov clustering implemented in the MCL package with different parameters obtaining fine cluster with few members (Inflation

parameter=5) or broad groups (Inflation parameter=2.0). In general, this clustering should delineate probable complexes from the physical interaction networks which can be used to enlarge the set of likely interacting proteins by additional pairs drawn from the them. In the case of the KEGG/Seed graph, the clustering should mainly retain the original modules while combining these entities which share great overlap and are wrongly split into different entities in the respective databases.

The resulting groups of the fine grained clustering represent likely true positive complexes (PPI) or pathways/modules (KEGG/Seed) with each protein-protein combination within a group forming a TPI of the gold standard set. Pairs of proteins which are separated between the large clusters should be examples of non-interacting proteins. This property is more likely between different pathways and modules as between complexes and the former two have been used to define the negative part of the background models. Two different background models have been created:

- The *physical* background model: this model comprises positive instances defined by the complex data, and negative instances by possible interactions which share neither a Seed nor a KEGG entity.
- The *functional* background model: this model comprises positive instances defined by the Seed module co-membership, and negative instances by possible interactions which share neither a Seed nor a KEGG entity.

This definition of background models is pictured in Figure 2.2. The positive and negative gold sets have been sampled from these models and are further used to assess the performance of the different prediction methods.

Binning procedure In order to apply the Bayesian theorem on the continuous prediction values these have to be altered into discrete units by binning. Two different binning procedures have been implemented: simple binning using fixed intervals over the value space and a partitioning into bins with equal amount of members. Bins with very few data-points are not statistically reliable but might occur when using the first strategy. The second binning strategy avoids such malformed bins and is therefore suited for methods which show only few predictions overlapping with the gold set. This is illustrated in the Figure 2.3.

Creation of scoring schemata and Bayesian integration Interactomes can be modeled by integrating different functional PPI prediction methods into one (functional) interaction network using a simple Bayesian approach. Mathematically, the problem can

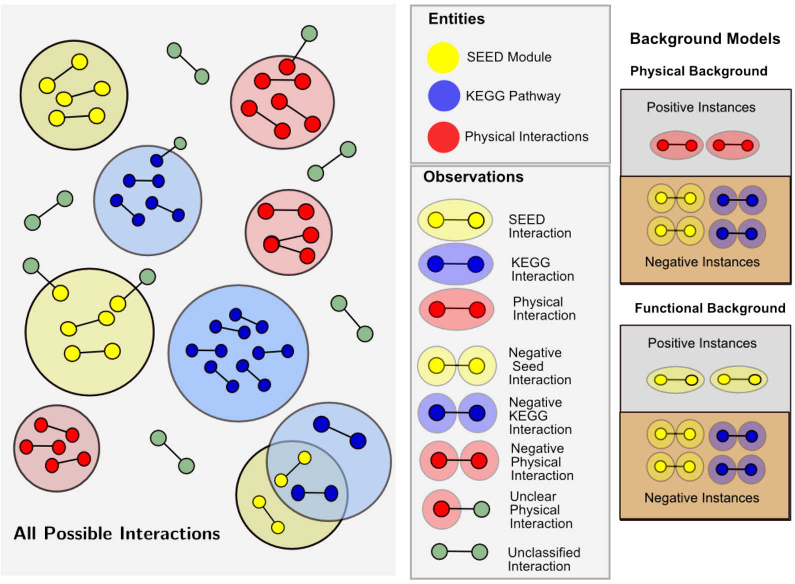


Figure 2.2: The derivation of background models. Two background models (a 'Functional' and a 'Physical' one) have been created as pictured on the right side of the illustration. The pre-processed entities (*bona fide* complexes, KEGG pathways, Seed modules) which define all 'known' possible interactions provide different kinds of possible interactions as listed in the box 'Observations'. The compiled background models provide negative and positive gold-sets which are derived from the entities with certainty in respect to the entities, i.e. by excluding unclear and unclassified interactions.

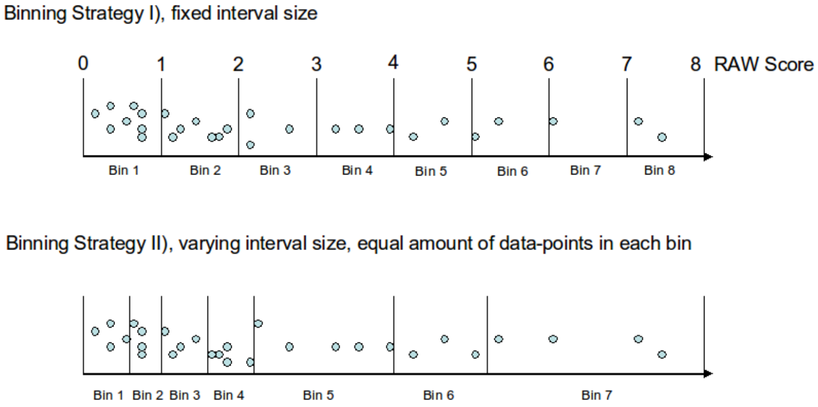

Binning Strategy I), fixed interval size
0
1
2
3
4
5
6
7
8
RAW Score
Bin 1
Bin 2
Bin 3
Bin 4
Bin 5
Bin 6
Bin 7
Bin 8
Binning Strategy II), varying interval size, equal amount of data-points in each bin
Bin 1
Bin 2
Bin 3
Bin 4
Bin 5
Bin 6
Bin 7

Figure 2.3: The two alternative binning strategies. Depending on the coverage obtained by the overlap with the gold-sets, one of these two strategies has been chosen. Strategy I) is based on fixed intervals and is used for methods which produce sufficient data-points over the whole spectrum of their reported scores. Strategy II) creates intervals with fixed amount of data-points by setting the borders accordingly. This strategy returned useful binnings in case of sparse data-points.

be defined as estimating the probability of an interaction between two proteins given an observed combination of evidences from different kind of measurements. The conditional probability of the evidence given a real interaction is determined by its frequency in a gold standard set of interactions/non-interactions. The probability to see an interaction given a certain evidence can then be computed using the Bayes' rule of computing the posterior probability. This approach remains tractable if the different measures are conditionally independent from each other and the full Bayesian approach can be simplified to the naive Bayesian formula. In most implementations for Interactomes, this independence is assumed. However, an only "sufficient" mutual independence can be tolerated as long as the predicted values are not skewed too much. This enables the integration of measures with only few data-points for which a full (not naive) Bayesian network cannot be computed.

I followed the approach used by Jansen and co-workers [162], which compute Likelihood ratios for each binned value of each measure. The same approach has been followed by by Shailesh and co-workers when modeling the interactome of *Plasmodium falciparum* [163].

For each method integrated in the network:

let be for each bin B and the positive and negative gold standards (named G_+ and G_-):

$$P(G_+|b_{pos}) = P(B|G_+); \tag{2.1}$$

the conditional probability of the bin regarding the positive gold set, and

$$P(G_-|b_{pos}) = P(B|G_-); \tag{2.2}$$

the conditional probability of the bin regarding the negative gold set. Both are estimated from contingency tables with the overlap of the bins with the gold-sets. (Compare Jansen et al. [162]). Then the likelihood ratio is defined as:

$$LR(B) = P(G_+|b_{pos})/P(G_-|b_{pos}); \tag{2.3}$$

The values of the individual bins for each method and background model have been saved as *scoring schemata*. These schemata also provide information on the amount of positive and negative instances which have been used to deduce the likelihood ratios. They have been manually inspected due to malformed bins (i.e. providing only few or no evidence) and the 'optimal' scoring scheme for each combination of method and back-

ground model has been chosen. 'Optimal' schemata are those with no malformed bins and a sufficient resolution to distinguish between 'poorly' and 'well' scoring intervals of the input values. If a method does not result in any optimal schema it must be discarded since no predictive power in regard to the background models could be observed.
The overal likelihood ratio for a link with n measures is obtained as:

$$LR(Link) = LR(B)_{method\_1} \cdot LR(B)_{method\_2} ... \cdot LR(B)_{method\_n}; \tag{2.4}$$

According to the Bayesian theorem, the posterior likelihood can be computed using a Prior belief L_prior about the ratio of interacting/non-inetracting proteins:

$$LR_{posterior}(Link) = LR(link) * L_{prior}; \tag{2.5}$$

This last step is used to introduce cut-offs for the reliability of the predictions which are defined by the prior believe (since a likelihood ratio >=1 is desired to classify a prediction as interaction) [163]. This step is discussed below.

Estimation of priors Estimating an exact prior probability of an interaction is not tractable since the amount of interactions in a cell is unknown. Estimates given in the literature vary: Bader and Hogue estimate the amount of interactions in a yeast cell at around 30000 [232], Mering at around 20000 [233]. Jansen et al. use a prior based on the assumption of 30000 interactions which yields a $\frac{1}{600}$ *a priori* chance that two proteins interact in yeast [162]. Following this argumentation, a typical chlamydial genome with around 900 genes should reveal a prior probability of $\frac{1}{30}$. However, the amount of interactions in a bacterial cell might be lower due to the lower complexity of bacteria compared to bakers yeast. These values refer to estimates of physical interactions, but many more functional interactions can be supposed, leading to much higher prior probabilities. Date and Stoeckert use different priors to filter their predicted *Plasmodium* network to obtain different coverages and found in their data a 40% coverage of the genome with a prior of around $\frac{1}{6}$ (likelihood threshold 6), and a 50% coverage at $\frac{1}{5}$ (likelihood threshold around 5).
As indication of an useful prior, I investigated the probability of getting positive interactions when comparing the amount of possible instances which share a KEGG pathway to the amount of possible interactions in disjunct pathways in the background models. This resulted in proposed prior of 0.18, the same test on the Seed-modules revealed an prior of 0.37. In the case of putative complexes, the value is $\frac{1}{25}$. I tested these indi-

cated priors due to their influence on network coverage and prediction performance. The networks resulting from the cut-offs according to these Priors are further named 'high confidence' for the KEGG based one, 'medium confidence' for the Seed based one, and 'complex' for the physical interaction ba sed case.

Cross-validation A five-fold cross validation has been used to asses the performance of the procedure. For this purpose, the initial background sets are split into five equally sized parts. For each, the complete procedure of background modeling is performed using the other parts as training. The resulting scoring schemes are used to build up a network with predictions between proteins from the test-set. These networks are then iterative filtered using different cut-offs on their edge weights and the behavior of the method is measured by counting true positive predictions TP and false positive predictions FP. The ratio TP/FP is plotted for different cut-offs as done by Jansen et al [162] and Date et al.[163]. Furthermore Recall, Precision, and F-measure have been computed as:

$$Recall = \frac{TP}{(TP + FN)}; \tag{2.6}$$

$$Precision = \frac{TP}{(TP + FP)}; \tag{2.7}$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{(Precision + Recall)}; \tag{2.8}$$

2.2.5 Properties of the deduced networks

The background models create valuable scoring schemata The manual inspection of the generated scoring schemata revealed for all except the structural based domain-domain interaction method (DIMA STRUC) suitable models. In case of the DIMA STRUC, the resulting model has been found not informative in respect to the background models since no bin resulted in an clear positive odds ratio (best found <1.2). Since sufficient data-points have been obtained for this method, the most probable explanation of this behavior is the absence of a linear score (which is not provided by DIMA for these methods) which could be used to categorize the predictions more fine-grained. The 3DiD method as provided in the original resource provides Z-scores for each domain interaction. However, these are interpreted in terms of specificity of the interaction but not as indicator of its' reliability by the authors [231]. The iPfam resource does not provide any scoring as checked in the literature [158] and on the data on the Sanger web-site. In conclusion, the prediction of DIMA STRUC could not be improved and the method has been discarded for the integration of the final network. The behavior of each schema is plotted in Figure 2.4.

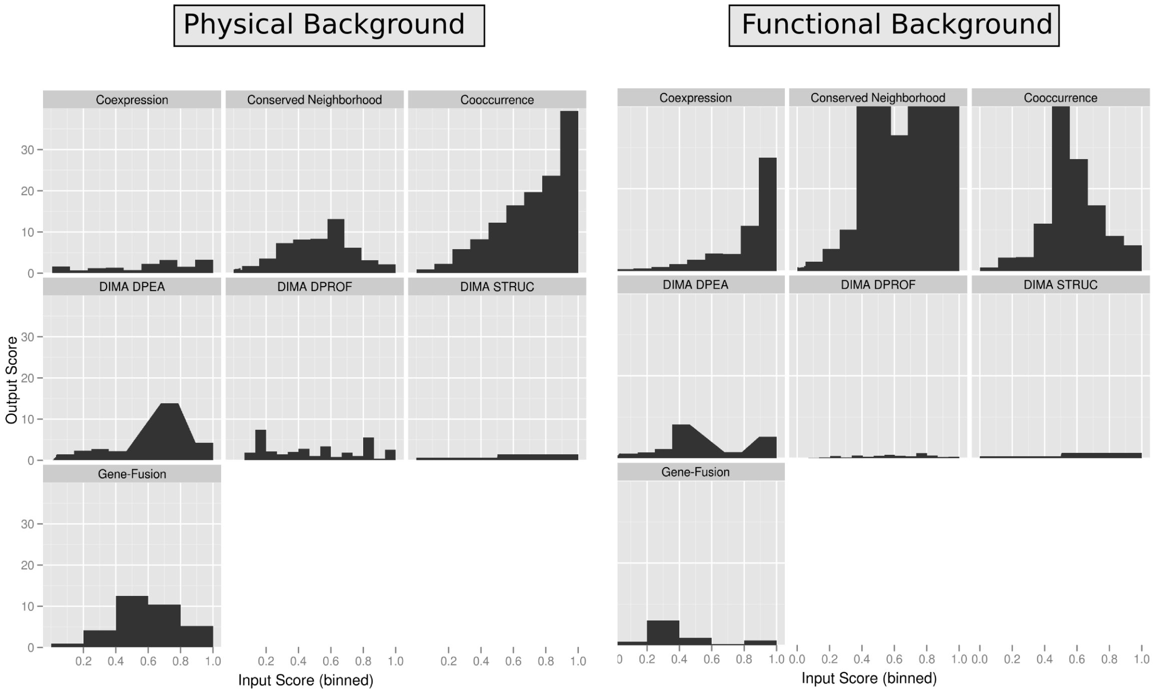


Figure 2.4: The performance of the different interaction prediction methods due to the physical and functional background models plotted as the output-scores for each Bin. For clarity and comparablity, very high values have been cut.

Different performance of different methods for physical and functional interactions The plots of the scoring schemata according to the background models in Figure 2.4 provide not only valuable insights into the predictive power of each method, but give also evidence that the applied methods exhibit individual characteristics according to the functional and physical backgrounds. In consequence, the scoring schemata allow to delineate networks which differ according to the background-models. In addition, they directly show if a method predicts rather functional than physical interactions or *vice versa*. These specific characteristics should be (at least partially) explainable due to the methods' underlying principles and are discussed in further detail in this paragraph:
From all methods, the conserved neighborhood has the strongest predictive power in

terms of the resulting scores. The method produces much smaller odds-ratios in the physical background compared to the functional case. However, compared to the other methods, the performance is still high and the method is also a good predictor for physical interactions, but with a clearly better power in the functional case. This is intuitively clear since many of these links reflect conserved operon structures which then reflect conserved regulation (and therefore a tendency to be related to the same biological process) rather then direct interaction. The performance in the physical background exhibits a drop off for higher input scores. This can explained by the effect that a common pathway membership is statistically much more frequent as a common physical interaction, a principle which leads to a decline in the rate of true positive physical interactions for high counts. The fusion method performs better in the physical case. This has been expected since the fusion event leads to co-translation into one protein which would be deleterious for the functionality of the interaction of the non-fused equivalents if that would not be a strong physical interaction. The relative low odds ratio compared to the neighborhood method is surprising since the observation of a fusion event should indicate a certain interaction and is probably an artifact of the relatively few instances in the positive training examples. Common gene expression implies co-regulation and therefore an rather functional coupling than a physical one, a principle reflected in the schemata. The cooccurrence method exhibits two different maxima for both background models. In the functional case, the maximum is in the area of mean scores, in the physical case the rule 'the higher the better' can be observed. This reflects a tendency for functionally coupled genes to have a common evolutionary fate, but with a less strong coupling as for physical interactors. The extreme examples in the latter case are complexes which need each sub-unit to fulfill their function and are rendered useless if one sub-unit fails leading to a very cohesive evolutionary fate detected as high cooccurrence-scores.

All DIMA methods perform better in the physical case (except DIMA STRUC, as discussed above). This can be interpreted as a recovery of domains dedicated to the mediation of physical interactions by the methods. This is especially meaningful for the domain pair exclusion method since its' training source are domains which can be found enriched in physically interacting proteins. However, the method has still some predictive power in the functional case contributing to the fact that a physical is also a functional interaction (what is true for any method). The DIMA profile method is also subjected to the same 'common fate of physical interactors' effect as the cooccurrence method but exhibits almost no predictive power in the functional case.

The predicted networks exhibit good performance and coverage The integrated networks exhibit a high ratio of TP/FP over a wide range of cut-off L_{cut}. In Figure 2.5, these rates are plotted against the cut-off for the functional and physical network. In general, both exhibit a similar behavior with increasing ratios: the higher the cut-off is chosen, the more confident is the prediction as indicated by increasing TP/FP rates. In case of the physical background, the behaviour of the data is more scattered as between L_{cut} 15-20, and the physical background model generates overall smaller ratios of TP/FP as inidicated by the absence of data-points with a TP/FP rate >40. This can be interpreted as a general lower confidence of a predicted physical interaction compared to the functional case. However, the test-set of positive physical interactions is much smaller than of the positive functional ones since many physical interactions are unknown. So, the performance in the physical case might be underestimated and obliged to a stronger statistical variation as in the functional case. The drop off in the functional

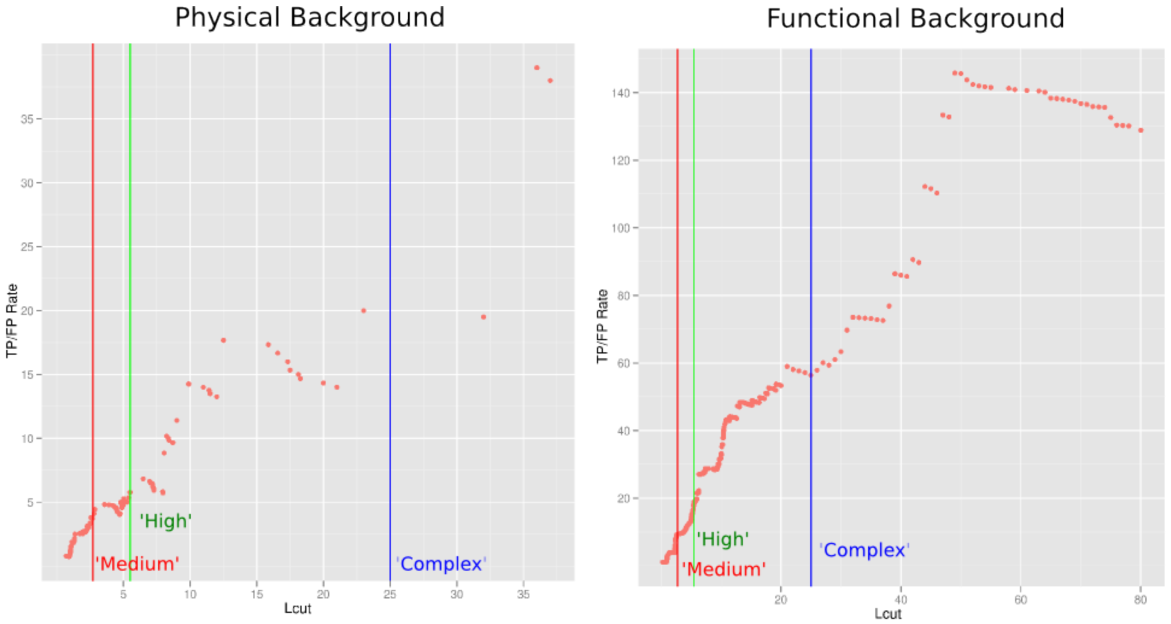


Figure 2.5: The performance of the integrated predictions in terms of TP/FP rate depending on L_{cut}, for both networks ('Functional' on the right, and 'Physical' on the left). The TP/FP ratio is computed according to Jansen et al. [162] as function of L_{cut}: $TP(L_{cut})/FP(L_{cut}) = \sum_{L>L_{cut}} pos(L)/\sum_{L>L_{cut}} neg(L)$ Where $pos(L)$ and $neg(L)$ are the numbers of positives and negatives in the respective gold standard (physical or functional background-model) with a given likelihood-ration L. The cut-offs introduced by the different prior assumptions are indicated by vertical lines (from left to right: 'medium confidence' cut-off at 2.7, 'high confidence' at 5.5, and 'Complex', the restrictive cut-off at 25). In the case of the functional background, spourious data-points with $L_{cut} > 80$ have not been plotted due to clarity.

case for very high values of L_{cut} (>50) is an artefact due to the existance of around ten false positive predictions with very high scores. Since the amount of nodes decreases with higher cut-offs, the ratio of TP/FP gets slightly worse as these ten false positives examples are not pruned but less true positives can be counted. The performance due to the chosen cut-off has also been measured in terms of Recall and Precision which is shown for the functional network in Figure 2.6. Both measures diverge rapidly with growing cut-off and saturate for high cut-offs. The small Recall values are contributed to the fact that most interactions defined in the background are not recovered by any method leading to a high false negative rate. Due to the scissor like behavior of the two values and the constantly very low Recall values, the F-measure also performs weakly and cannot give a hint to an optimal compromise between Recall and Precision. In both plots, the cut-offs according to the different prior assumptions as estimated above are indicated by vertical lines. The actual values of the different performance measures at these cut-offs are listed in Table 2.1. All cut-offs are set in the area of low Recall but high Precision, and already in the case of the 'Medium confidence' network, Precision of 0.9 is reached. Shortly before this first cut-off at $L_{cut} = 2.7$, a rapid transition from Precision of 0.79 to 0.9 (+0.11) can be observed while the sensitivity drops down by a decrease of only 0.04 indicating a meaningful pruning of false positive hits at this point. Date and co-workers used the prior to adjust the desired coverage of their interaction network. In Table 2.2 the resulting coverages due to the different prior assumptions done herein are summarized. The coverages reached for the environmental *Chlamydiae* are generally lower and contribute to the large fraction of their proteins that are orphans and cannot be covered by the genomic context methods. At maximum, 65% of the *C. muridarum* and 46% of *P. amoebophila* are covered. The values for the higher cut-offs indicate a still high coverage > 40%. These coverages are comparable to those found in

Lcut	**2.7**	**5.5**	**25**
Name	Medium	High	Interaction
Precision	0.9	0.95	0.98
Recall	0.06	0.05	0.03
F-Measure	0.12	0.09	0.06
TP/FP	9.34	18.7	56.4

Table 2.1: Performance measures using different prior assumptions. 'Lcut' is the cut-off corresponding to a prior assumption, 'Name' the name given to the networks,'Precision' precision computed as $\frac{TP}{(TP+FP)}$, 'Recall': recall computed as $\frac{TP}{(TP+FN)}$, 'F-measure' computed as weighted harmonic mean of Precision and Recall. 'TP/FP' is the rate of True positive/False positive predictions.

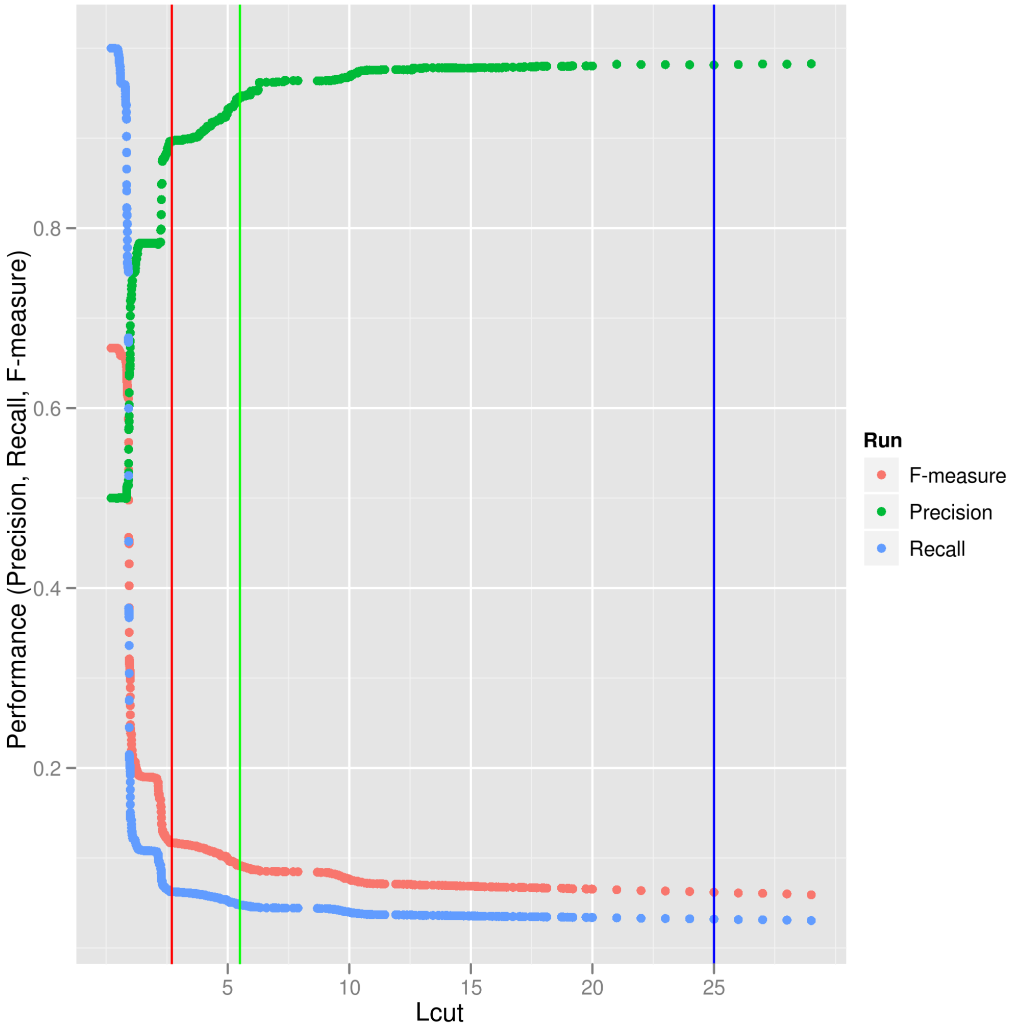


Figure 2.6: The performance of the integrated predictions in terms of Precision and Recall depending on L_{cut}. 'Lcut' is the cut-off corresponding to a prior assumption, 'Name' the name given to the networks,'Precision' precision computed as $\frac{TP}{(TP+FP)}$, 'Recall': recall computed as $\frac{TP}{(TP+FN)}$, 'F-measure' computed as weighted harmonic mean of Precision and Recall. (from left to right: medium confidence cut-off at 2.7, 'high confidence' at 5.5, and 'Complex', the restrictive cut-off at 25).

other studies.

The integration of additional methods extends the networks The different methods predict a different amount of functional interactions. To assess the basic information gain in terms of extension to the network by each method, the amounts of unique links and nodes contributed by each method have been determined. These amounts are sum-

Lcut	**Edges**	**Cov. patho.**	**Cov. env.**
Functional			
2.7	7083	0.65	0.46
5.5	4771	0.58	0.38
25	3296	0.52	0.34
Physical			
2.7	2725	0.47	0.31
5.5	2587	0.47	0.31
25	1757	0.41	0.26

Table 2.2: Coverage depending on different prior assumptions. 'Lcut' is the cut-off corresponding to a prior assumption, 'Name' the name given to the networks, 'Edges' the amount of edges remaining in the network, 'Cov. patho.' the coverage of the genome of *Chlamydia muridarum* Nigg in terms of the fraction of proteins of this organism which participate on the network, after applying Lcut, 'Cov. env.' the same value for *P. amoebophila* UWE25.

marized in Table 2.3. Notably, these counts refer to all predictions made by each method after the initial scoring, many of them might be pruned after applying the bayesian integration step including the application of the prior based cut-off. In addition, the median and maximal degree of the unique nodes has been determined as well as the degree after pruning according to the high and medium confidence cut-offs. By principle, the contributions of predicted interactions with evidence from one method only correlates with the amount of links they predict: the neighborhood method is the most powerful followed by the cooccurrence method. The DIMA DPEA method produces almost as many predictions as the cooccurrence method but is obviously less correlated to the other methods as the latter one. This is indicated by the high amount of additional nodes which are introduced to the network by this method (616 for DIMA DPEA in contrast to 412 for the neighborhood method). This reflects the definition of interacting domains in this method (compare above) which is truly orthogonal to the other methods since no evolutionary aspect, especially no orthologous relationships have been employed in the definition and application of this method. The predicted interactions of DIMA DPEA are merely of low score as indicated by the median values, but also several high scoring edges are predicted that lead to extensions of both, the high and medium confidence networks.

The additional deduction of conserved neighborhood links from the environmental *Chlamydiae* (with support $\geq$3) revealed 571 new links of which 117 have not been covered by the mapping from STRING. These comprise 10 interactions between known orthologous groups and 106 involving environmental specific orthologous groups as summarized in

BG	Method	Unique Edges	Unique Nodes	Median	Max	Medium	High
C	Coexpression	9078	3	1.02	2.45	0	0
C	Fusion	138	0	0.0	0.0	0	0
C	Neighborhood	152525	412	0.95	8.28	26	14
C	Dpea	24784	616	0.98	13.84	492	62
C	Dprof	152	9	1.12	2.60	0	0
C	Cooccurrence	29524	3	0.91	0.91	0	0
M	Coexpression	9078	3	1.06	18.47	10	4
M	Fusion	138	0	0.0	0.0	0	0
M	Neighborhood	152525	412	0.92	50.15	32	20
M	Dpea	24784	616	0.99	5.21	360	0
M	Dprof	152	9	1.09	1.81	0	0
M	Cooccurence	29524	3	0.94	0.94	0	0

Table 2.3: The statistics of newly introduced links due to the different prediction methods. 'BG' is the background-model with C=physical/complex and M=functional/module. 'Method' refers to the prediction method (compare text). 'Unique Edges' gives the amount of edges which are solely predicted by the method. 'Unique Nodes' gives the amount of nodes which would have no link without that method. The former two values are trivially the same for both background models. 'Median' is the median of the unique edge weights after the scoring step. 'Max' the maximal score reached in this set, 'Medium' the amount of links which can be found in the 'medium confidence network', 'High' the respective value in the 'high confidence' network.

Table 2.4. The sub-network spanned by these new links has been further tested on structure by assessing the average clustering coefficient *cl* in comparison with the same but randomized network. This test revealed a higher structure of the 117 links as expected by random (*cl* 0.046 against 0.025). In consequence, the found sub-network should be non-random and therefore biologically meaningful.

The resulting networks are scale-free and exhibit an high inner structure The distributions of the degree probabilities and the cluster-coefficients give insights into general properties of natural networks as introduced in Chapter 1.3.3. A manual inspection of the degree distribution showed, that the networks follows a power law distribution in all networks generated in this study. Example plots of the distributions from the high confidence networks are given in Figure 2.7. The distribution of the clustering coefficients $C(k)$ did not reveal a clear tendency to follow a distribution of the form $C(k)\tilde{}k^{-1}$ but more $C(k) = const$ and are quite scattered. So, they are independent of the choice of k. In consequence, a strong hierarchical behavior cannot be stated. Due to the observations, the networks are clearly scale free. The overall clustering coefficients of the

Partner 1	Partner 2	Amount
COG	COG	5
COG	NOG	5
COG	Env	65
NOG	Env	17
Env	Env	13

Table 2.4: Previously unknown 'neighborhoods' detected using the environmental *Chlamydiae*. Pairs of proteins which appeared at least three times in close neighborhood in the environmental *Chlamydiae* (compare text) have been investigated due to their ability to introduce new links. Links which have been unknown in the STRING database have been classified by their composition due to orthologous groups. 'COG' is a COG orthologous group member, 'NOG' a non-supervised orthologous group member, and 'Env' a member of an orthologous group detected solely in the environmental *Chlamydiae*. 'Partner 1' and 'Partner 2' refer to this classification, 'Amount' gives the occurrences found for links between the respective orthologous groups.

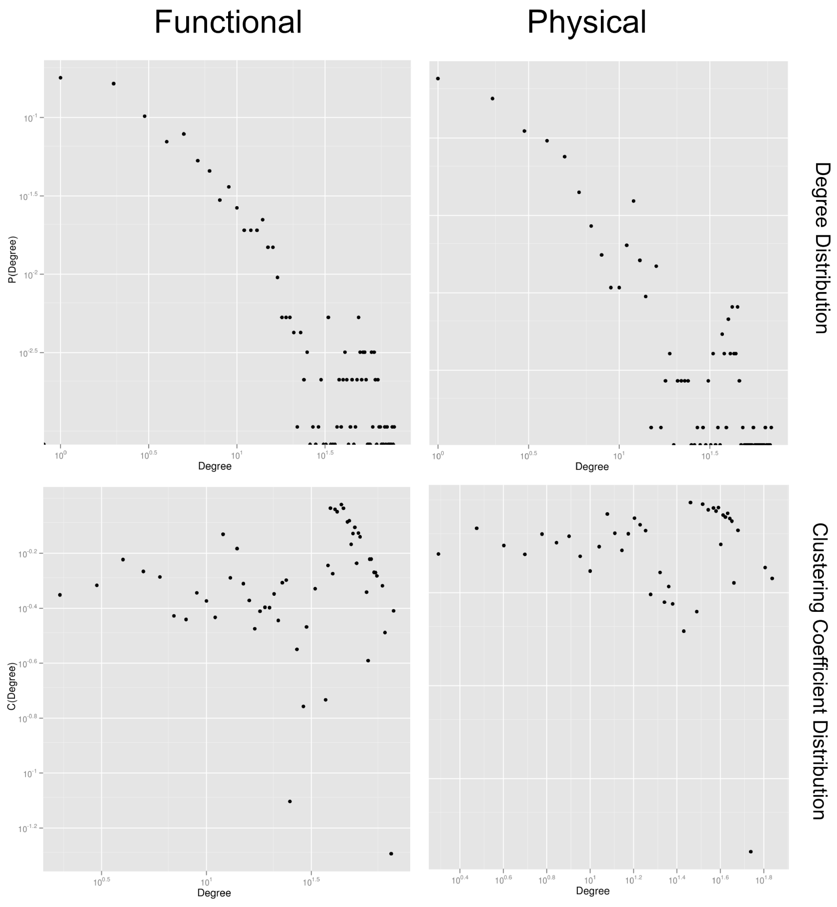


Figure 2.7: The distributions of the clustering coefficients and the degree probabilities in the case of the high confidence networks (log-log scale).

networks are high ranging from 0.4-0.5, and in every case much higher as a compared random case, as listed in Table 2.5.

Name	Background	Cl	Cl(rand)	Ratio
Complex	Physical	0.50	0.015	34.02
Medium	Physical	0.47	0.007	67.77
Medium	Functional	0.43	0.013	32.86
High	Physical	0.46	0.009	53.58
High	Functional	0.40	0.0095	42.38

Table 2.5: Cluster coefficients of the different networks. 'Name' the name of the network, 'Background' the background-model used, 'Cl' the clustering coefficient, 'CL (rand)' the clustering coefficient of an equivalent randomized network, 'Ratio' the ratio of CL and CL(rand).

The 'complex' and the 'functional' network are highly correlated The different background models influence the resulting networks. To assess how different both networks are, they must be compared due to the specific amount of nodes. Especially, I asked if the physical networks comprises parts that are not apparent in the functional one. The tYNA web-server allows to illustrate such dependencies between networks by using one network as predictor for the other one. A test using the functional network as predictor of the physical revealed a Precision of 0.69 and a Recall of 0.95. In consequence, the physical networks is merely a sub-set of the functional one, but with different edge weights due to the different scoring schemata. Due to this correlation between the functional and the physical network (especially due to the high Recall), the different background models do not offer a simple way to distinguish between functional and physical links. To derive a list with the most probable candidates for physical interactions, I determined the links that exist only in the 'complex' physical network but not in the functional one. 85 instances of such interactions exist comprising 56 nodes. The interactions are listed in the supplementary Table 6.5.

Discussion The integration of different types of interaction prediction leads to networks with high TP/FP rates while not loosing too much coverage in terms of nodes. Especially the integration of the DIMA DPEA method added extra knowledge, therefore, the extention of STRING has been valuable for the project. The background-models lead to differences in the predictive behaviour of the individual methods, however, the resulting networks are similar in terms of their node and edge content. The resulting functional network is the basis for the delineation of *Chlamydia* specific functional modules which is described in the next section.

2.2.6 Delineation of functional modules from the chlamydia specific networks

Introduction With the functional interaction networks at hand, functional modules can be deduced by clustering. As first step, the optimal clustering parameters are determined that result in a functional homogeneous clustering. This can be done by a parameter exploration which employs an performance measure of the clusters' functional homogeneity. Such a procedure is a trade-off between two concepts: on the one hand, modules should be detectable solely by their traces in the data as they are evolutionary conserved entities. On the other, a poorly performing detection method can only be identified by the comparison of the resulting modules against a given, external, functional classification. A good compromise is to judge the result of a module detection method by rather broad functional categories from the COG classification in the first place. A strong deviation according to such a general functional categorization (i.e. a large amount of modules functionally not coherent) would indicate a failure of the method and an optimization on them does not bias the module detection method into the direction of man made (but perhaps not correct) definitions of individual modules.

Despite of a general parameter exploration, a possible improvement of the clustering could be to sub-cluster large modules. This idea is based on the observation of very large entities reported when initially clustering these networks. These large components reflect the existence of a giant component in many biological and other 'real-world' networks [234, 235, 236] and the clusters may contain meaningful sub-clusters that cannot be extracted from this dense structure. In this chapter, a simple concept to divide large modules into sub-modules is tested for its' ability to improve the overall clustering quality.

Material and methods

Preparation of Input Data The mode of the three different input graphs (the complete, medium, and high confidence, compare section x) have been labeled using the COG functional categories. For each node, multiple assessments of categories have been allowed as found in the COG database.

Clustering procedure The input graphs have been clustered using by Markov clustering as implemented in the MCL package, version 1.006 [183]. MCL has been chosen since it has been shown to produce good results in several studies and performs well for input graphs with different statistics [114][185][186]. The resulting clusterings comprise three different sets of modules, the 'complete', the 'medium confidence ', and the 'high confidence' set, correspondingly named by their input graphs. As anticipated above, the clusterings exhibit some large components that could be further divided into meaningful sub-clusters. The large component could exhibit a different internal structure (i.e. higher degrees and edge weights) as the global network. If taken alone, the sub-networks defined by each cluster could be further divided into smaller entities by the Markov clustering approach. To experimentally test this hypothesis, big clusters have been tested for their ability to be sub-divided into smaller ones by creating sub-networks that only contain nodes (and edges between them) appearing in the cluster.

The Inflation parameter of the MCL algorithm has been varied between 1.2 (resulting in less but larger clusters) and 5 (resulting in more but smaller clusters) in steps of 0.2. For each clustering, the resulting modules have been re-clustered using a second setting for the Inflation parameter I_2 which has been set to one of 1.2, 3.0, and 5.0. This has been repeated for clusters above a certain size. This size threshold has been varied between 10 and the maximum cluster size for each clustering found in the initial clustering.

The modules derived from the interaction network should be functional homogeneous and the distribution of function over the clusters should be non random. For the assessment of these properties, a homogeneity measure judging the clustering as a whole must be employed. I adopted a functional homogeneity measure introduced by Loganantharaj and co-workers [237] originally used to assess the quality of clusters derived from expression data. This measure incorporates the structure of the clustering (i.e. the modules) as well as the used functional categories and is based on Shannon's entropy [238]. It measures the separation of functional categories by the clustering as well as the inner cluster homogeneity (compare [237]). The non-randomness is then evaluated by the computation of Z-scores. This overall procedure has been adopted from Hu and co-workers [143] (Supplementary protocol S6)

The measure is defined as follows:

Let be $C = \{c_1, c_2, c_3 ... c_m\}$ the clustering with groups c_i of size n,

Let be $F = \{f_1, f_2, f_3 ... f_n\}$ the grouping of orthologs with a certain functional label f_i of size m,

Let be p_{fc} the frequency of a functional category f in cluster c, the cohesiveness $Cc(c)$ of a cluster c concerning a category f is then defined as:

$$Cc(f) = -\sum_{c=1}^{m} p_{fc} log_2(p_{fc}); \tag{2.9}$$

and for the complete clustering, the total cluster cohesiveness CC is defined as:

$$CC = -\sum_{f=1}^{n}\sum_{c=1}^{m} p_{fc} log_2(p_{fc}); \tag{2.10}$$

This equation reflects the functional cohesiveness regarding the cluster. In order to judge the performance due to the ability to separate function between the clusters, Logananthara introduced the cohesiveness in respect to the functional groups as follows: let be b_{ir} the frequency of a clustering in a functional category computed as: for a cluster r with x entries annotated by a functional category n with N_i members in total, let be

$$b_{ir} = x/N_i; \tag{2.11}$$

The cohesiveness of a functional category is then defined as:

$$Cf(c) = -\sum_{i=1}^{n} b_{ir} log_2(b_{ir}); \tag{2.12}$$

and for the complete set of functional categories, the total functional cohesiveness CF is defined as:

$$CF = -\sum_{r=1}^{f}\sum_{i=1}^{m} b_{ir} log_2(b_{ir}); \tag{2.13}$$

The total performance P can then be expressed as the sum of both terms:

$$P = CC + CF; \tag{2.14}$$

The smaller this term, the more functional homogeneous is the clustering. In order to assess the non-randomness of a clustering, the resulting values of actual clusterings have been compared with randomized data using Z-score statistics. For this purpose, the functional labels of the proteins have been randomized while retaining the cluster structure, and the homogeneity scores of 1000 runs have been computed. After inspection of the density distribution of the random values which follow a normal distribution, the absolute values of the Z-scores have been computed as follows:

$$Z - score = |\frac{\times - \mu}{\sigma}|; \tag{2.15}$$

where μ is the score average of the random runs, x the actual score of the clustering and σ the standard deviation of the random runs.

Creation of organism specific modules The resulting modules have been projected to each chlamydial data-set using the assignments of the proteins to the orthologous groups as delineated before. Notably, modules might contain paralogs of member proteins.

Results The majority of modules revealed by the clustering of the functional modules has a size smaller than twenty members, while some exhibit larger sizes, especially in the case of the complete network. The size distributions for all three networks resulting from a clustering with inflation parameter of 3, up to 100 members are plotted in Figure 2.8. The existence of large modules motivates the re-clustering approach. In Figure 2.9, the impact of the re-clustering procedure is shown. In this plot, the size distributions of the complete set before and after re-clustering (of all components with size $\leq$ 10) are shown. Changes in the size distribution illustrate the ability of the re-clustering to sub-divide modules. However, the very large component is not split into a bunch of sub-modules but has been divided in one still large component and few additional small modules as indicated by the still large cluster existing in the re-clustering (at size 250). In Figure 2.10 the performance of the three different networks (complete, medium, and high confidence) during the parameter exploration (without sub-clustering) are plotted. The general performance of all clustering is clearly above random. The complete network results in generally lower Z-scores where the high confidence network achieves the best scores.

The quality of the clustering seems rather independent of the chosen parameters, as only a small increase of Z-scores for higher inflation values (i.e. higher granularity of the clustering) can be observed. The parameter exploration in terms of re-clustering of large modules did not reveal an increase of performance for the medium or high confidence networks as indicated by the Z-score as the maximum values could be found for clustering without re-clustering with Z-score=35.5 for the medium (Inflation=5), and Z-score=38.5 for the high confidence clustering (inflation=2.7). Only the clustering of the complete set could be improved by re-clustering. The maximum Z-score found is 24.5 when clustering with an inflation value of 5 in the initial step and re-clustering components with more

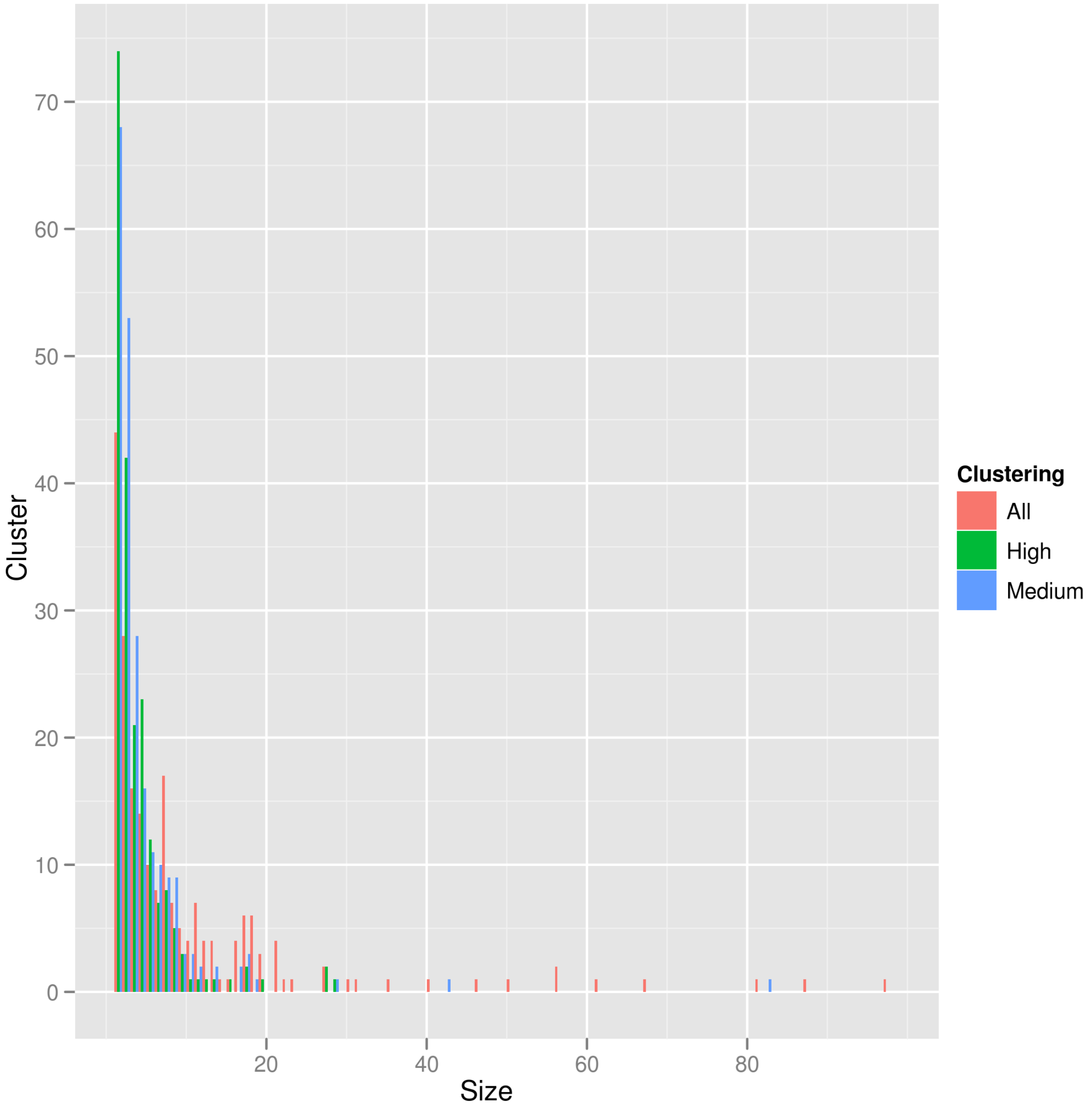


Figure 2.8: Histogram of size distributions resulting in clustering of the three different input networks ((High=high confidence, Medium=medium confidence , All=the complete network) with inflation parameter set to 3.0. Modules with a size>100 not shown.

than 220 members with an inflation parameter of 3.0. This finding is illustrated in the additional Figure 6.1 in which the performance in terms of Z-scores are plotted against the re-clustering of modules above a certain size: while the individual Z-scores vary little due to the parameter I_2 of the re-clustering, a general trend for an increased performance in relation to the re-clustering cannot be seen since the data-points merely lie on horizontal lines. The coverage of the chlamydial proteomes (listed in Table 2.6

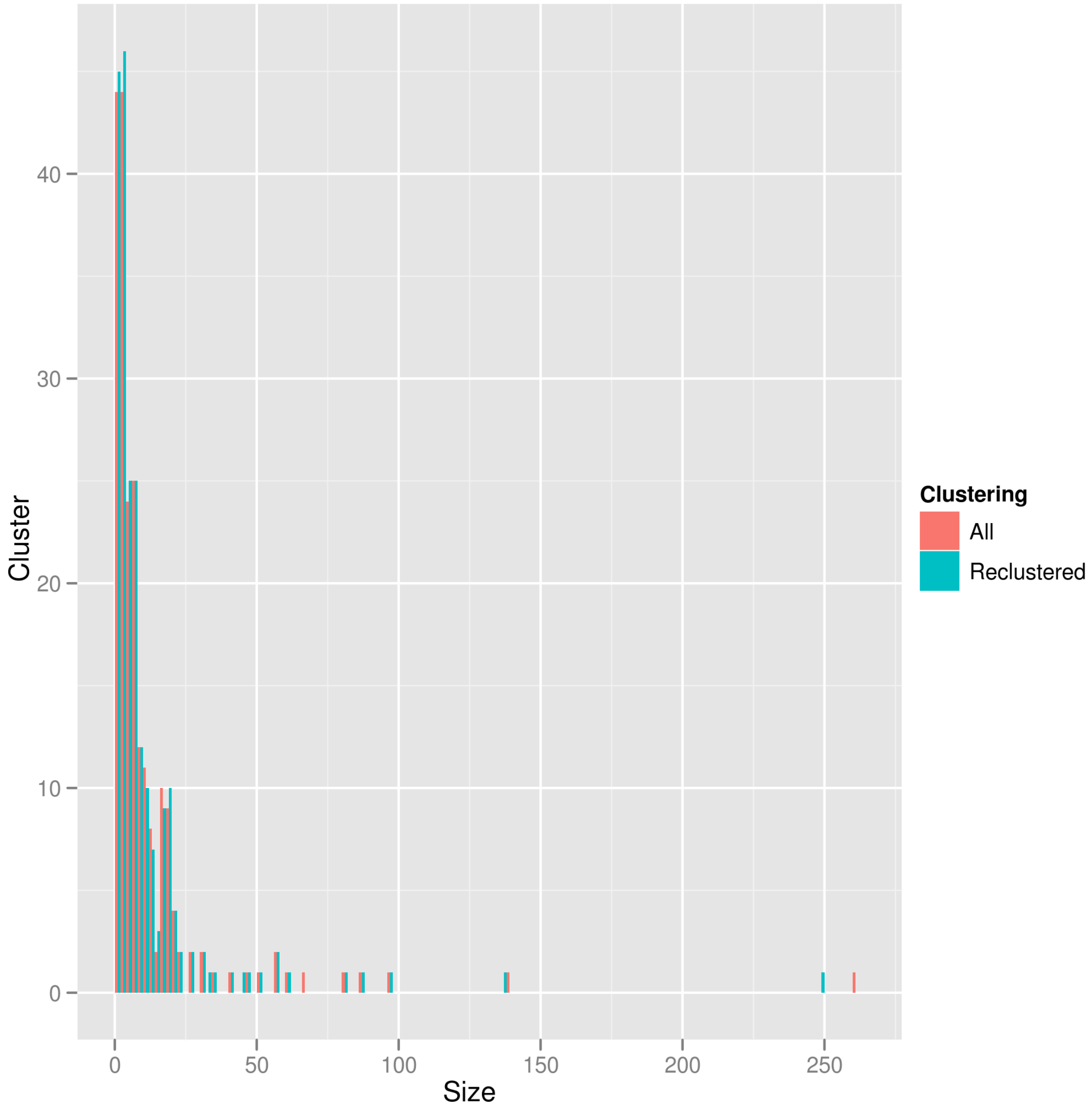


Figure 2.9: Histogram of size distributions resulting in clustering of the complete network before and after the application of the re-clustering procedure (with inflation parameter set to 3.0).

) by the resulting modules varies between 64% for *Waddlia chondrophila* and 67% for *Chlamydophila trachomatis* (medium confidence clustering), the values resulting from the high confidence clustering are somewhat smaller.

Discussion The modules resulting from the cluster analyses exhibit a functional homogeneity much above random and should therefore represent meaningful functional entities. A tolerance in terms of functional homogeneity can be observed in the param-

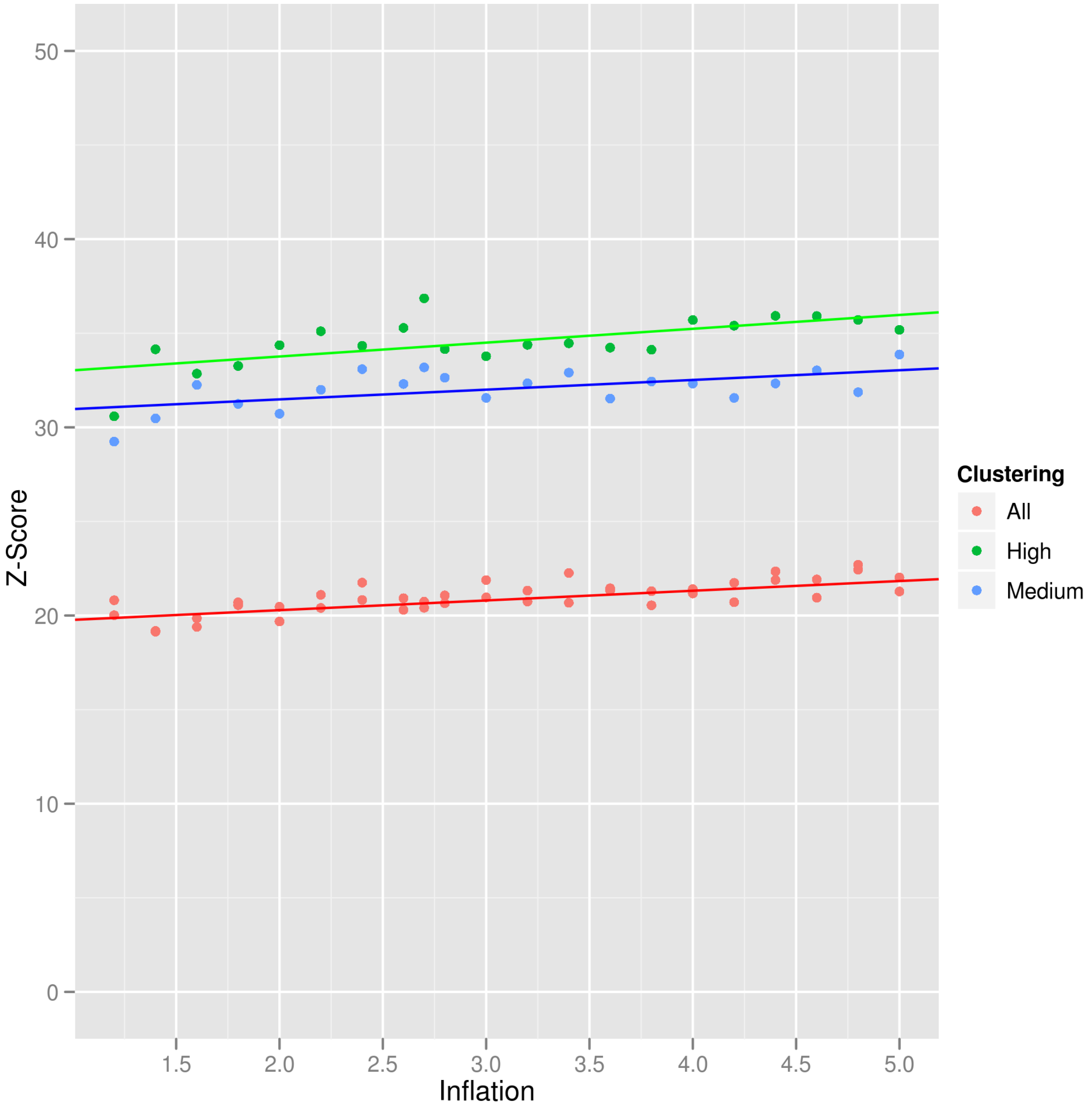


Figure 2.10: Z-scores resulted in the parameter exploration of runs without re-clustering and varying Inflation parameter. Values for all three networks (High=high confidence, Medium=medium confidence , All=the complete network) shown. On the x-axis: the Z-score as defined by $|\frac{x-\mu}{\sigma}|$ of the cohesiveness measure, on the y-axis: value of the MCL inflation parameter.

eter exploration, a finding in congruence with von Mering et al. [114] who detected the same robustness for changes in the parameter settings for several clustering algorithms indicating a strong signal of functional modularity in the networks. The coverage of the chlamydial proteomes is dependent on the used input network, in maximum, two-third of the proteins can be assigned to a module with size>1. The re-clustering approach

Organism	Number proteins	Cov. medium	Cov. high	Num. medium	Num. high
C. abortus S26/3	932	65%	57%	603	535
C. trachomatis A/HAR-13	919	65%	58%	593	530
C. muridarum Nigg	911	64%	56%	580	514
C. pneumoniae LPCoLN	1105	56%	50%	621	552
C. trachomatis D/UW-3/CX	895	66%	58%	587	523
S. negevensis	2509	39%	33%	988	839
W. chondrophila	2070	46%	39%	944	810
C. Protochlamydia amoebophila UWE25	2030	49%	42%	999	854
C. trachomatis L2b/UCH-1/proctitis	874	67%	59%	582	520
C. trachomatis 434/Bu	874	67%	59%	582	520
C. trachomatis B/TZ1A828/OT	880	66%	59%	579	516
P. acanthamoebae UV7	2854	41%	35%	1174	996
C. pneumoniae TW-183	1113	55%	49%	612	543
C. pneumoniae J138	1069	57%	51%	614	544
C. felis Fe/C-56	1013	61%	54%	619	548
C. pneumoniae CWL029	1052	58%	51%	607	537
C. pneumoniae AR39	1112	54%	48%	604	535
C. caviae GPIC	1005	60%	54%	606	540

Table 2.6: Coverage of chlamydial proteomes by modules. Column 'Organism' contains the Organism name, 'Size' is the number of proteins in this organism, 'Cov. medium' the medium confidence network as input for the clustering, 'Cov. high' the high confidence network as input. 'Num. medium', and 'Num. high' give the absolute numbers of covered proteins in the respective clustering.

can divide large modules further but did not significantly change the quality of the clustering in terms of functional homogeneity. This observations reflects a principally hierarchical structure of the module space as it has been described earlier [179]. Due to the projection of the orthologous groups on the individual protein instances, multiple paralogs may populate a module. This might be seen as a small but not avoidable draw-back: paralogs might have adapted to a different functionality and should be part of another module and, in some cases, duplications of complete functional entities might occur. By principle, a functional division of paralogs is not feasible without further in-deep analyses and their occurrence limits the 'orthologous resolution' in the analyses based on the delineated functional modules. Such analyses could comprise the evaluation of neighbored genes in the individual genomes that could allow to differentiate paralogs by their operon-membership or phylogenetic analyses. These steps would have to be done prior to the delineation of functional interactions by the genomic context methods and would result in a more fine-grained definition of orthologous groups. The strategy to cluster orthologous groups first and to project them on genomes of interest in a second step has been adapted from Mering and co-workers [114]. The advantage of this procedure is the utilization of the additional information provided by functional links to orthologs missing in a genome of interest. Furthermore, this approach allows to compare the completeness and existence of modules between the different species without a need for complicated mapping procedures. The opposite argument is that links between orthologs which do partially not exist cannot be implemented by the genome under investigation. The approach used herein is a compromise between both aspects by the use of a taxon specific network that contains all general orthologous groups which can be found in at least one *Chlamydium* as well as purely chlamydia-specific ones. The resulting modules comprise dense regions of the predicted interaction network and divide the latter into highly connected components. So chlamydia specific cellular subsystems have been generated comprehensively. This data represents an inventory of cellular functions existing in chlamydial cells that can be used, for example, to analyze comprehensively the evolution of cellular functionalities between the different *Chlamydiae*.

2.3 Analyses using functional modules

2.3.1 Functional modules reveal KEGG pathway and modules

Introduction The detected modules describe the functional equipment of the *Chlamydiae* as it could be delineated due to evolutionary constraints by the prediction methods. These 'natural' modules might differ from manually defined pathways or modules since a man-made ontology of modules/pathways can differ from the picture found in the chlamydial data. To investigate this point, the recovery of cellular machineries (as pathways, transport systems and others) is further assessed by systematically comparing the delineated functional modules with manually defined pathways and modules from the KEGG database. It is unclear, to which extend the KEGG definitions of modules and pathways match the modules delineated from the network. This information is necessary to judge the potential of KEGG to describe the chlamydial modules. This analysis comprises two steps: firstly, the creation of individual KEGG maps for each *Chlamydium* not existent in KEGG, and secondly an assessment of the mutual recovery of both concepts, modules and pathways.

Material and Methods KEGG pathways, modules, and sequence data has been downloaded from the KEGG web-site on the 30th of May 2010. Each chlamydial genome has been mapped on the KEGG sequence data using SIMAP (best hit, E-value cut-off 10e-3). The publicly available genomes are completely processed in the KEGG database and match with 100% identity. In case of the non-public genomes, hits with a coverage less than 60% in length and with less 20% identity have been discarded to avoid spurious hits. This mapping procedure resulted in individual KEGG maps (pathways and modules) for each organism. For each chlamydial module the best matching KEGG

pathway/module has been determined by the maximal Jaccard index introduced by Song and co-workers [185]. This measure judges the overall coincidence of two partitions (in this case, of KEGG definitions and the module clustering) and is therefore well suited to assess the overall recovery of KEGG pathway and modules. This measure is called the 'overall Jaccard index'. This index incorporates two initial measures, the 'normalized average Jaccard index in respect to the modules (JaccardM), and to the KEGG groups (JaccardK). The overall Jaccard index is defined as harmonic mean of JaccardM and JaccardK.
The Jaccard index between two sets A and B is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.16}$$

let be K the set of KEGG pathways or modules, M a module clustering, The normalized average JaccardM and JaccardK are then defined as: Let be k_i a kegg module/pathway in K with $|K|$ pathways, m a module of clustering M with $|M|$ modules, then the best matching pair is defined as:

$$JaccardC_i = max(Jaccard(k_i, m_j)); \tag{2.17}$$

Let be $|M_i|$ the size of the maximal overlapping module and $|K_j|$ the size of the corresponding Kegg pathway/module, JaccardM is then defined by:

$$JaccardM(M,K) = \frac{\sum_{i=1}^{|M|} |M_i| \cdot JaccardC_i}{\sum_{i=1}^{|M|} |M_i|}; \tag{2.18}$$

and JaccardK analogously

$$JaccardK(M,K) = \frac{\sum_{j=1}^{|K|} |K_j| \cdot JaccardC_i}{\sum_{j=1}^{|K|} |K_j|}; \tag{2.19}$$

and the overall Jaccard index of a module clustering as the harmonic mean:

$$Jaccard(M,K) = \frac{2 \cdot JaccardC(M,K) \cdot JaccardG(M,K)}{JaccardC(M,K) + JaccardG(M,K)} \tag{2.20}$$

These measures have been determined for each KEGG mapping (pathway and modules) and the *Chlamydial* modules resulting from the high and medium confidence clustering. To assess the recovery of KEGG pathways in a random background, the individual

KEGG maps have been randomized by changing pathway-protein relationships while retaining the structure of the KEGG modules and pathways in terms of size distributions. For each module in each organism the best matching KEGG pathway and module have been determined. Modules which participate on several KEGG pathways/modules have been identified: a module has been classified to join two KEGG entities if the overlap to both is greater than one module member. As additional criterion, the overlapping parts of the module must be disjunct since several KEGG entities share protein entries which would be trivially recovered. In the same way, KEGG entities which match several modules have been identified.

Results Additional table 6.6 lists the amount of individual modules matching either KEGG pathways or modules with a certain maximal Jaccard index. For each chlamydial organism and module clustering (medium,high) the amount of modules is listed which have a maximal $JaccardC_i$ (defined in 2.17) of > 0.5, ≤ 0.5 but > 0, and 0 which means no overlap to any KEGG instance. In *P. amoebophila* UWE25, for example, 19 KEGG modules-functional module overlaps result in a $JaccardC_i$ >0.5, 89 cases exhibit overlap, but with low score, and 100 modules do not overlap with KEGG modules at all. In comparison to pathways, more KEGG modules are recovered by chlamydial modules with a score ≥ 0.5 and the overall amount of established relationships is higher: only three modules have a $JaccardC_i$>0.5 with a KEGG pathway, but as many as 134 hit a pathway in contrast to 108 for the KEGG modules. The results do not differ much between the medium and the high confidence clustering in terms of recovered KEGG pathways or modules. The amount of modules that do not have any assignment to KEGG are more frequent in the medium clustering: for example, in *C. abortus* S26/3 115 high-confident modules do not have a counterpart in KEGG where even 136 cannot be assigned in the medium confidence case. The tendency of better (i.e. higher scoring) recovery of KEGG modules instead of pathways is further supported by the overall Jaccard indices which are listed in the additional Table 6.7. In this table, the JaccardK, JaccardC, and the overall Jaccard as well as the overall Jaccard in the randomized case are listed for each organism and module set (high, medium confidence). The KEGG modules reveal higher scores in all measures. The overall Jaccard of KEGG modules has its' maximum at 0.22 for *C. caviae* strain GPIC in the high confidence modules, contrarily, the equivalent value for the KEGG pathways is 0.15. In all cases, the overall Jaccard index is at least twice as high as in the random case. In Table 2.7 all examples of *P. amoebophila* UWE25 modules are listed that exhibit a high scoring

Module	KEGG-Module	Description	Jaccard
module 150	M00007	Pentose phosphate pathway, non-oxidative phase, fructose 6P	0.6
module 99	M00368	D-Methionine transport system	1.0
module 91	M00660	RuvABC complex	1.0
module 83	M00105	dTDP-Sugar biosynthesis, Glc-1P => dTDP-Glc => dTPD-Rha	1.0
module 82	M00008	Entner-Doudoroff pathway, glucose-6P => glyceraldehyde-3P +	0.67
module 78	M00033	Shikimate pathway, phosphoenolpyruvate + erythrose-4P =>	0.6
module 74	M00114	Lipopolysaccharide biosynthesis, KDO2-lipid A	0.67
module 117	M00366	Polar amino acid transport system	1.0
module 114	M00318	Sulfonate/nitrate/taurine transport system	1.0
module 67	M00284	Complex IV (Cytochrome c oxidase), cytochrome o ubiquinol	1.0
module 113	M00658	RecBCD complex	1.0
module 64	M00369	Peptides/nickel transport system	0.6
module 103	M00051	Ectoine biosynthesis	0.67
module 52	M00288	V-type ATPase (Prokaryotes)	0.8
module 43	M00671	Sermidine/putrescine transport system	0.8
module 28	M00293	ATP synthase	0.6
module 195	M00285	Complex IV (Cytochrome c oxidase), cytochrome c oxidase,	1.0
module 166	M00340	Putative ABC transport system	1.0
module 165	M00380	Lipopolysaccharide transport system	1.0

Table 2.7: Modules matching KEGG modules with Jaccard >0.5 in *P. amoebophila* UWE25. 'Module' is the functional module. 'Kegg-Module' the module accession in KEGG, 'Description' is the module description in KEGG, 'Jaccard' is the $JaccardC_i$ index of Kegg-Module and module.

hit with KEGG modules ($JaccardC_i \geq 0.5$). Some KEGG modules are detected exactly (with $JaccardC_i = 1$) in the chlamydial modules such as the KEGG module M00658 (the RecBCD complex, with three member proteins, recovered by module # 113) or the ABC like transporter system M00318 (the Sulfonate/nitrate/taurine transport system, with three members, recovered by module # 114). In the majority of cases, the modules can be uniquely assigned to a certain KEGG pathway. However, some modules exhibit overlap with several pathway and module definitions and 'join' them. Examples found in *P. amoebophila* UWE25 are listed in in Table 2.8 (multiple KEGG modules joined

Module	KEGG entities	Descriptions	Amount
module 88	M00119, M00114	CMP-Kdo biosynthesis, Lipopolysaccharide biosynthesis, KDO2-lipid A	2
module 5	M00192, M00095	C5 isoprenoid biosynthesis, non-mevalonate pathway, Pyrimidine deoxyribonuleotide biosynthesis, CDP/CTP =>	2
module 4	M00309, M00308	Ribosome, archaea, Ribosome, bacteria	2
module 3	M00309, M00308	Ribosome, archaea, Ribosome, bacteria	2
module 77	M00210, M00649, M00648	Phosphatidylglycerol biosynthesis, CDP-diacylglycerol =>, uvrBC complex, uvrA2B2 complex	3
module 66	M00373, M00372	Manganese/iron transport system, Zinc transport system	2
module 58	M00050, M00059, M00679	Lysine degradation, lysine => saccharopine =>, Isoleucine degradation, isoleucine => propionyl-CoA, Pyruvate oxidation, pyruvate => acetyl-CoA	3
module 37	M00012, M00011	Glyoxylate cycle, Citrate cycle, second carbon oxidation	2
module 35	M00194, M00308	C15 isoprenoid biosynthesis, Ribosome, bacteria	2
module 15	M00159, M00160	Fatty acid biosynthesis, initiation, Fatty acid biosynthesis, elongation	2
module 10	M00309, M00308	Ribosome, archaea, Ribosome, bacteria	2

Table 2.8: Modules matching several KEGG modules in *P. amoebophila* UWE25. 'Module': the module identifier, 'KEGG entities': the KEGG modules that have overlap with the module, 'Descriptions': the descriptions of these KEGG modules, 'Amount': the amount of KEGG modules joined by the functional module.

by a functional module) and in the additional Table 6.8 (multiple pathways joined by a module). The amount of different KEGG entities joined by a single module is generally low, in most cases two entities can be found joined. In the case of the KEGG modules, eleven cases can be detected. Nine of them connect two KEGG modules, two three of them. Fifteen functional modules are part of several KEGG pathways with a maximum of five pathways connected by 'module 6'. Many of these observed joins are biologically meaningful as the participating KEGG entities are related to a similar process in the cell. Examples for this interpretation are the 'module 10' which covers two fatty acid metabolism related modules and 'module 71' joining the KEGG pathways 'Ribosome' and 'Aminoacyl-tRNA biosynthesis'. Other observations can be explained by the internal structure of KEGG which, for example, defines two different modules for microbial ribosomes ('M00309', the 'archeal ribosome' and 'M00308', the 'bacterial ribosome') which are connected by the ribosome related 'module 3' since orthologs in both KEGG definitions appear in *Chlamydiae*. Other examples result from insufficient orthologous resolution which cannot distinguish certain types of sub-functionalization as in the case of 'module 66' which joins two metal related ABC transporters. Some examples contain connections which cannot easily be meaningfully interpreted, as for the 'module 35' which connects 'M00012 C15 isoprenoid biosynthesis' and 'M00308 Ribosome, bacteria'. KEGG pathways and modules which join different functional modules are listed in tables 2.9 (modules) and 6.9 (pathways, in the appendix). The KEGG pathway definitions are rather broad and often subsume different functional entities in one map. This implies, that for these kinds of maps the modules cover only a part of the KEGG pathway which is in fact the case: for example, eight modules cover different parts of the ABC transporter overview as they represent different ABC transporters which are distinctly recovered, and three the map KO3070 comprising a compilation of bacterial transport systems (compare Table 6.9). In general, the functional modules comprise subsystems of KEGG entities. A good example is the Type III secretion system (M00713 in KEGG) which is covered by two modules or the ribosomes (M00309, M00308).

Discussion The recovery of KEGG pathways and modules by the functional modules is partial and the two concepts exhibit communalities and differences alike. Many modules do not hit any KEGG pathway or module as they represent entities that are either chlamydia specific or not yet existent in the KEGG definitions. This picture might change with the progress of curation by the KEGG annotation team, i.e. by their completion of the module definitions. The recovery of the pathway definitions given by KEGG differs from case to case. Metabolic pathways are recovered partially, whereas cellular sub-systems as transport systems or certain complexes result in a good coincidence. Christian von Mering and co-workers found modules derived from STRING revealing many pathways with high coincidence in *E. coli* when comparing their module clustering with curated sets of EcoCyc pathways (84% pathway specificity in the case of enzyme related modules) [114]. Notably, the two studies are not directly comparable: despite of differences in the design of the studies (as the used measures and the pre-processing of the functional modules, and benchmarking only proteins existent in

KEGG entity	Description	Modules	Amount
M00273	Complex I (NADH dehydrogenase), NADH dehydrogenase I	module 18, module 26	2
M00114	Lipopolysaccharide biosynthesis, KDO2-lipid A	module 88, module 74	2
M00309	Ribosome, archaea	module 3, module 10, module 4	3
M00308	Ribosome, bacteria	module 29, module 3, module 71, module 10, module 35, module 4	6
M00369	Peptides/nickel transport system	module 132, module 64	2
M00713	Type III secretion system	module 115, module 81	2
M00192	C5 isoprenoid biosynthesis, non-mevalonate pathway	module 5, module 168	2
M00679	Pyruvate oxidation, pyruvate => acetyl-CoA	module 159, module 58	2
M00159	Fatty acid biosynthesis, initiation	module 22, module 15, module 122	3
M00246	Heme biosynthesis, glutamate => protoheme/siroheme	module 140, module 162, module 90	3
M00597	DNA polymerase III complex	module 94, module 185	2

Table 2.9: KEGG module matching several functional modules in *P. amoebophila* UWE25.

the pathway definitions) the *E. coli* data-sets are much more complete due to the study bias towards this model organism and the used EcoCyc database, which is especially dedicated to describe the metabolic pathways of *E. coli*. The chlamydial solutions of certain processes may differ from the common pathway descriptions found in the pathway databases due to their evolutionary distant relationship to typical model organisms and their highly reduced genomes. Although this reduction in chlamydial genomes has been taken account of as far as possible by initially projecting the KEGG pathways to the individual genomes, the reality of chlamydial metabolism might still be less adequately captured by KEGG as it would be the case for e.g. *E. coli*. However, most of the KEGG pathway to module relationships are meaningful and could serve as basis for a curated definition of chlamydial pathways and cellular machineries.

2.3.2 Annotation based on functional modules

Introduction In this analysis, the value of the delineated functional modules for the annotation of still unknown proteins in *Chlamydiae* is assessed. This investigation is motivated by the large amount of uncharacterized proteins in the *Chlamydiae*, and several of them do not exhibit sufficient sequence homology to known proteins in other species that would allow an annotation transfer. The functional ontology used herein is the MIPS functional catalog (FunCat) described in section **??**. This ontology has been chosen for two reasons: firstly, its' hierarchical structure allows to chose these levels of functional granularity that are suitable to describe the function of a module (and hereby of its' member proteins). Secondly, a manually training set is available for *P. amoebophila*. Such a set is necessary to assess the performance of a function prediction approach and for no other *Chlamydium* a by hand curated set of any functional ontology exists. The basic concept of module based annotation is simple and have already been employed in other studies [185][179]: annotations are transfered from annotated to not yet annotated module members. As first step in this analysis, I analyzed the possible information gain (i.e. FunCat predictions) which can be obtained by a module based approach. Secondly, two strategies to delineate FunCat categories of a functional module have been assessed: a *sensitive approach* which generates all possible hypotheses of a modules functional categories and a *selective approach* which deduces the most probable functional category. The performance of both approaches was benchmarked using a cross-validation procedure. Possible causes which might negatively influence the result (i.e the choice of clustering parameters and improper functional categories of the ontology) are further investigated. To compare the performance of the module based approach with a network based one, the same two strategies have been implemented using the functional network neighbors instead of modules.

Material and methods

Used data sets In this analysis, the FunCat 2.0 schema has been used as an ontology. Categories which are obviously not existent in bacteria as well as the categories 98, "classification unclear" and 99 "unclassified protein" have been filtered out resulting in a sub-set of the FunCat which is used further on in the analysis. The set of *P. amoebophila* annotations has been filtered due to this scheme resulting in a gold-standard set of annotations. The projections of the modules and networks on the proteomes of the *chlamydiae* as described above 2.2.6 have been labeled by the functional categories while allowing multiple labels per protein. After generating stop-lists of FunCat categories performing poorly (see below), additional sets of modules and networks have been created by pruning the FunCat scheme according to the stop-lists and by re-labeling of the proteins with the new schemata.

Annotation status of *Chlamydiae* The annotation status of *Chlamydiae* can be defined by the amount of uncharacterized proteins as indicated by their annotation provided in the description lines of each protein. For each publicly available *Chlamydial* genome, the amount of protein entries which are annotated as 'hypothetical protein' or 'conserved hypothetical protein' in RefSeq has been counted. For each organism, the amount of such proteins that cluster in modules have been determined. To assess the local network properties of uncharacterized proteins, a Wilcoxon Rank Sum test on the degree and the weight distributions from all unknown and all other proteins to direct neighbors has been performed (in the predicted *P. amoebophila* selective interaction network, singletons without functional neighbor were not counted).

Implementation of annotation strategies The implementation of the sensitive approach is straightforward: all functional categories provided by any module member are transfered to all other members if they do not already carry this annotation. The selective approach is based on the rationale to transfer only the functional category that is most informative, i.e. is the most probable explanation of a modules' function. This most probable function is determined empirically as the most frequent category. If several top scoring categories exist, the category with the highest information content in terms of functional specificity is chosen. This category has been determined as follows:

- For each category assigned to a protein, add all higher level categories: if, for example, the annotation 10.10.20 is given, add 10.10 and 10.

- Find the most frequent FunCat annotations by counting

- If several categories are equally frequent, chose the most specific one: if, for example, categories 20.10 and 20 are equally frequent, chose 20.10 since it is more specific due to the hierarchical organization of the catalog.

The same two strategies as described above for the module based approach have been applied using the direct network neighbors instead of the module co-members as input.

Assessment of performance The performance of the method is assessed using each protein instance as test instance and the rest as input for the training (leave-one-out cross-validation). In the cross-validation process, only proteins clustered in modules and providing FunCat assignments are counted. Notably, a true positive is detected if the predicted FunCat number is prefix of the test instance, i.e. more general terms are counted if the test comprises a more specific term from the same category.
In the sensitive approach, the performance can be described by the amount of true positive protein-FunCat pairs (existent in both sets) TP, false positive (existing only in the predictions) FP, and false negative FN (prediction missing protein-FunCat pair). As performance measures, recall and precision are computed as follows:

$$Recall = \frac{TP}{(TP + FN)}; \tag{2.21}$$

$$Precision = \frac{TP}{(TP + FP)}; \tag{2.22}$$

In the selective approach the amount of false negatives predictions cannot be meaningfully deduced since only one category is transferred while the others are ignored. Therefore, the performance in the selective case is measured as rate of correct and incorrect assignments: a correct prediction is reported if the transferred category exist for the test protein, an incorrect one otherwise. The rate of correct predictions is by its' way of computation equivalent to precision. Since the precision in the sensitive approach is computed regarding all protein-funcat pairs but in the selective only one instance is counted, both are differently named in this analysis to avoid confusion.

Let N be the amount of proteins predicted,

$$Rate\ correct = \frac{TP}{N}; \tag{2.23}$$

$$Rate\ incorrect = \frac{FP}{N}; \tag{2.24}$$

For comparison, the performance of each method on a random but equally structured module set is determined. This is accomplished by the perturbation of the protein-annotation relationships: each protein gets the set of annotations from a randomly chosen other one. This randomization does not change the module structure nor the distribution of joined annotated functional categories and is therefore rather conservative.

Assessment of systematic negative influences Three major causes may affect the performance of the annotation process:

a) certain categories of the FunCat annotation scheme might be inappropriate to describe functional modules.

b) the clustering might be sub-optimal and does wrongly combine or split 'real' functional entities.

c) the coverage and quality of the gold set might be poor.

Where point c) cannot be assessed by definition without further manual curation, point a) and b) can be investigated further. To assess point a), functional categories that frequently contribute to false predictions have been excluded. For this purpose, five different stop-lists have been tested:

- a stop-list with the two top categories that produce most often false positive predictions in random predictions ('Freq. rand.')
- a stop list with the two top categories that produce most often false positive predictions in random predictions and their sub-categories ('Freq. rand. (all)')
- a stop list with the categories producing 0% percent correct predictions (0% correct)
- a stop list with the categories producing less than 33% percent correct predictions (<34% correct)
- a stop list with the categories producing less than 66% percent correct predictions (<64% correct)

These values have been derived using the selective annotation approach and the high confidence module clustering. To assess point b), the clusterings which resulted from the initial parameter exploration are re-evaluated due to their predictive behavior in the cross-validation procedure.

Results Comparing the local network-properties of *P. amoebophila* proteins with unknown function against the complete set of proteins revealed significant shifts towards lower values for both, degree and weight, in the high confidence networks (significance: P-value degree shift: 0.005951, weight shift: $< 2.2e\text{-}16$) and a large fraction of them does not cluster into modules. However, several instances of unknown proteins have been found to be module members. Therefore, the modules provide candidates for annotation transfer as summarized in Table 2.10. The amount of uncharacterized sequences in modules is on average 60 per organism, with a maximum of 243 proteins (23% of the entire amount of sequences) in *P. amoebophila*. The pathogenic *Chlamydiae* also exhibit a certain amount of hypothetical protein entries in the modules, ranging from 40 to 70 entries.

Organism	Unknown (modules)	Unknown (not modules)	Prot. in modules	Genome size
C. trachomatis D/UW-3/CX	43	234	523	895
C. trachomatis 434/Bu	41	193	520	874
C. trachomatis L2b/UCH-1/proctitis	41	193	520	874
C. muridarum Nigg	66	273	514	911
C. trachomatis A/HAR-13	51	234	530	919
C. abortus S26/3	70	215	535	932
C. caviae GPIC	51	328	540	1005
C. pneumoniae AR39	79	456	535	1112
C. pneumoniae CWL029	51	377	537	1052
C. pneumoniae J138	53	377	544	1069
C. pneumoniae TW-183	58	421	543	1113
C. felis Fe/C-56	40	283	548	1013
C. Protochlamydia amoebophila UWE25	243	998	854	2030
C. pneumoniae LPCoLN	42	353	552	1105
C. trachomatis B/TZ1A828/OT	42	193	516	880
W. chondrophila	0	0	810	2070
S. negevensis	0	0	839	2509
P. acanthamoebae UV7	0	0	996	2854

Table 2.10: Annotation status of chlamydial proteomes. Column 'Organism' contains the organism name, 'Unknown (modules)' is the amount of hypothetical proteins in modules, 'Unknown (not modules)' the amount of hypothetical proteins which do not cluster into modules, 'Prot. in modules' the overall amount of proteins which cluster in modules, 'Genome size' the total amount of proteins.

The functional categories transfered in the selective approach (high confidence clustering) are listed in the additional Table 6.11 and the functional categories transfered in the randomized selective approach (high confidence clustering) are listed in Table 6.12 in the appendix. These lists have been used to determine the stop lists as described above. In the randomized case, the categories '01' (Metabolism) and '16' (Protein with binding function or cofactor requirement, structural or catalytic) are the most frequent to produce false positive hits (with 496 and 123 cases) and represent clear outliers compared to the other categories. In consequence,they were added to the stop-list 'Freq. rand.' and with all their sub-categories to 'Freq. rand. (all)'. The amount of correctly transferred entities per functional category varies widely. Categories which completely fail mostly comprise only few counted transfers (1-7) which are caused by small modules. Contrarily, small and medium sized modules exist with high success rates, even with 100% correct predictions. Categories which participate on many predictions (>20) show success rates between 15% and 70%.

The values for recall and precision in the sensitive approach are summarized in Table

2.11. The table lists the results for all used stop lists in both, medium and high confidence modules. For each combination, the performance as Recall and Precision for the actual clusterings and the randomized case are given. The maximum Recall that could be reached is 0.74 using the <64% stop-list. The reached Precision is generally low (∼ 0.11 in the medium and ∼ 0.16 in the high confidence modules). The performance of the approach is clearly above random as indicated by the generally higher values for Recall and Precision in the non-randomized case. However, the Recall in the randomized case turned out to be surprisingly high (with a maximum at 0.33). This effect is caused

C.	Stoplist	R	Recall (M)	Precision (M)	Recall (N)	Precision (N)
M	-	-	0.43	0.11	0.48	0.10
M	<0.64% correct	-	0.74	0.13	0.53	0.10
M	<0.34% correct	-	0.58	0.12	0.45	0.08
M	0% correct	-	0.55	0.10	0.44	0.08
M	Freq. rand.	-	0.43	0.11	0.48	0.10
M	Freq. rand. (all)	-	0.45	0.11	0.48	0.09
M	-	+	0.22	0.03	0.35	0.04
M	<0.64% correct	+	0.29	0.02	0.36	0.04
M	<0.34% correct	+	0.26	0.03	0.35	0.04
M	0% correct	+	0.28	0.03	0.30	0.03
M	Freq. rand.	+	0.21	0.03	0.36	0.04
M	Freq. rand. (all)	+	0.18	0.02	0.32	0.03
H	-	-	0.44	0.16	0.45	0.16
H	<0.64% correct	-	0.73	0.16	0.58	0.15
H	<0.34% correct	-	0.62	0.14	0.52	0.13
H	0% correct	-	0.59	0.13	0.50	0.12
H	Freq. rand.	-	0.44	0.16	0.45	0.16
H	Freq. rand. (all)	-	0.46	0.17	0.47	0.16
H	-	+	0.20	0.03	0.26	0.04
H	<0.64% correct	+	0.28	0.02	0.31	0.04
H	<0.34% correct	+	0.23	0.03	0.26	0.04
H	0% correct	+	0.24	0.03	0.23	0.03
H	Freq. rand.	+	0.20	0.03	0.26	0.04
H	Freq. rand. (all)	+	0.17	0.02	0.24	0.03

Table 2.11: Performance of module based annotation, sensitive approach. 'C.' is the clustering (M: from medium confidence, H: high confidence network). 'Stoplist' is the stop-list employed (compare list given in the text , sub-paragraph 'Assessment of systematic negative influences' in 2.3.2), 'R' defines whether the initial annotations are randomized (+) or not (-). 'Recall (M)' the recall computed as TP/(TP+FN) for the module based approach, Precision (M) the precision computed by TP/(TP+FP) in the module based approach. 'Recall (N)' ad 'Recall (N)' are the values for the corresponding network based approach.

by the principle to transfer any annotation found within the module which increases the probability to generate a true positive prediction by chance. This effect does not influence the Precision (which can be found always ~ 0.03 in the randomized case) since much more incorrect than correct annotations are transferred. This observation holds also true in the non-random case: many wrong annotations are transferred that lead to poor values for the Precision in every case. A general observation is that the modules from the high confidence network exhibit a little better performance. In terms of Recall and Precision, the use of stop-lists turns out to be fruitful and to increase the performance. The results of the selective approach are summarized in Table 2.12. Some general observations made in the sensitive case are also valid for the selective one: the high confidence modules exhibit a better performance (best percentage True positive predictions 48% compared to 40%) and the influence of the stop lists increases the rate of true positive predictions. The performance in the randomized case is partially high with up to 28% correct predictions. This high amount is a side effect of the strategy to determine the most probable functional category: in the randomized case, this category is more frequently one of the high level categories as the only intersection between the annotations in a module. Although these assignments have few support in the module they are assigned and counted as true positive if the functional category of the test instance is equal or a descendant of the predicted category. The initial clustering of modules could influence the performance of the function prediction. The results of an assessment of a possible influence are summarized in Figure 2.11. In this figure, the performance for different Inflation parameters are plotted for the selective function prediction approach. The experiment has been done for the function prediction without a stop list and with the 'Freq. rand.' stop list. An obvious tendency towards a better performance for fine grained clustering (i.e. larger Inflation values) can be seen. A further test on the re-clustered modules from the initial parameter exploration supported this finding since the values for the coarse clusterings increased after re-clustering into smaller entities (compare 6.3 in the appendix). In comparison with the module based annotation procedure, the performance of the network-based analysis differs in some aspects: in the selective procedure, the rate of correct predictions is increased compared to the module based approach for each combination of network/clustering and stop-list. In the sensitive approach, the Precision in the network is generally a little worse than for the modules. The values of the Recall exhibit differences: in general the Recall of the network approach is higher when no stop list is used. The picture turns when stop-lists are employed and the combination module based plus the '<0.64%' stop list clearly

Clustering	Stoplist	R	True (M)	True (N)
M	-	-	0.38	0.46
M	<0.64% correct	-	0.35	0.44
M	<0.34% correct	-	0.40	0.56
M	0% correct	-	0.40	0.46
M	Freq. rand.	-	0.38	0.47
M	Freq. rand. (all)	-	0.36	0.44
M	-	+	0.21	0.34
M	<0.64% correct	+	0.09	0.14
M	<0.34% correct	+	0.22	0.45
M	0% correct	+	0.23	0.34
M	Freq. rand.	+	0.10	0.12
M	Freq. rand. (all)	+	0.10	0.18
H	-	-	0.46	0.53
H	<0.64% correct	-	0.40	0.53
H	<0.34% correct	-	0.48	0.60
H	0% correct	-	0.48	0.55
H	Freq. rand.	-	0.42	0.52
H	Freq. rand. (all)	-	0.41	0.49
H	-	+	0.21	0.32
H	<0.64% correct	+	0.10	0.18
H	<0.34% correct	+	0.24	0.37
H	0% correct	+	0.22	0.33
H	Freq. rand.	+	0.07	0.12
H	Freq. rand. (all)	+	0.11	0.12

Table 2.12: Performance of module based annotation, selective approach. 'Clustering' is the clustering (M: from medium confidence, H: high confidence network). 'Stoplist' is the stop-list employed (compare list given in the text , sub-paragraph 'Assessment of systematic negative influences' in 2.3.2), 'R' defines whether the initial annotations are randomized (+) or not (-). 'True (M)' fraction of correctly annotated proteins by the module approach, 'True (N)' the fraction of correct annotations predicted by the network based approach of the corresponding network (high, medium) confidence during cross-validation.

outperforms the network based equivalent (Recall 0.74 against 0.53). In addition, the random runs in the network based annotation produce higher Recalls as their module counterparts (compare Table 2.11 and Table 2.12).

From the 243 proteins which are annotated as 'unknown' in *P. amoebophila*, for 178 a FunCat category can be proposed (no stop-list, high confidence clustering set) by the selective procedure. The proposed FunCats are listed in the supplementary Table 6.10. A check of Funcat predictions for 30 proteins by a biologist (Dr. Astrid Horn, University of Vienna) by careful annotation using database and literature research revealed

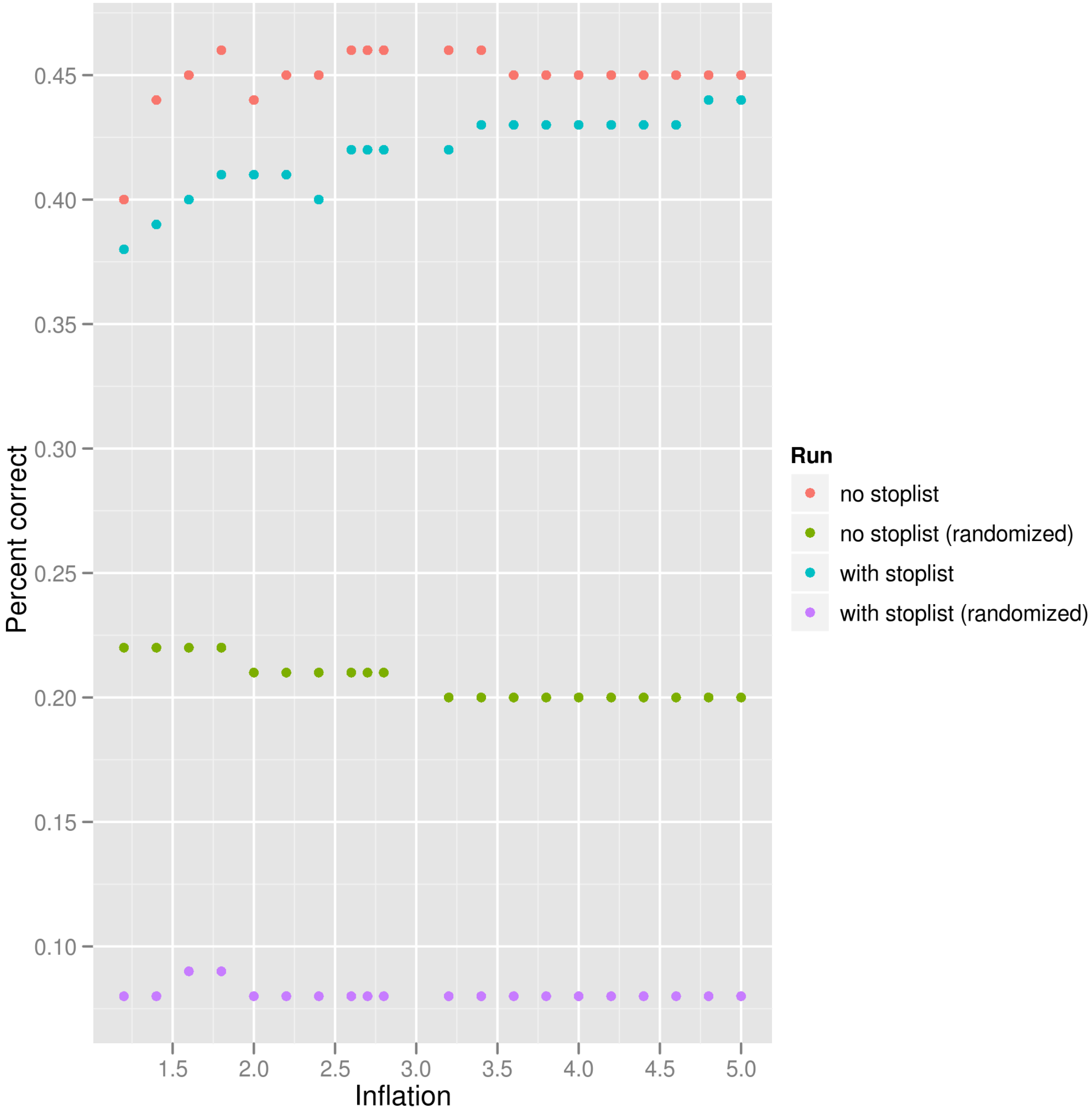


Figure 2.11: Performance of the selective function prediction for different clusterings (with no re-clustering) with and without stop list (stop-list contains FunCat '01' and '16'). 'Inflation' is the inflation value used to cluster the set. 'Percent correct' is the rate of correct predictions. The different runs comprise: 'no stop-list: complete set of functional categories, with stop-list: categories 01 and 16 excluded. Both runs have been repeated with randomized FunCat labels, denoted as 'no stop-list (randomized)' and 'with stop-list (randomized)'.

in eight cases a coincidence with the proposed annotation, in 17 cases the annotation could not be verified. In five cases, no statement could be made. These are listed in the supplementary Table 6.13.

Discussion Many unknown proteins of *Chlamydiae* have a tendency to populate parts of the network with low local structure as inidicated by the investigation of the local network-properties. Therefore, they have a lower chance to be clustered into a functional module, the amount of hypothetical proteins in modules is sufficiently high to justify a module based annotation approach. For a large fraction of them a functional prediction by module co-membership or by an analysis of the neighborhood in the functional interaction network can be made. These unknown proteins have not been annotated due to the lack of clear homologous relationships during the primary annotation process of *P. amoebophila*. Thus, this analysis points out the potential of the module/network based approaches for annotation. Two strategies have been proposed, the rather trivial sensitive approach and the selective approach. The latter one is based on an adaptation of the commonly used majority vote criterion to the hierarchical structure of the FunCat. The procedure reveals several examples where a highly specific FunCat could be transfered if the module is in majority specific to a sharply defined cellular process that is also fine grained described in the FunCat scheme. The procedure automatically adapts to a less fine description of a module if either the module is functionally more diverse or its' functionality is not described in a high (i.e. specific) level of resolution in the FunCat. The performance of the network and the module base approach differ due to the annotation strategy used. Song and co-workers [185] discuss the performance of similar approaches as used herein applied on the yeast interaction network. They found the performance of a module based approach is outperformed by a network based one in case of the *S. cerevisiae* interactome. The findings for the functional chlamydial interaction network in this study must be further differentiated. As found by Song et al., the general performance of the network based approach is better, but this finding depends on the pruning of the used ontology for general terms (i.e. the employed stop-lists) and the employed annotation strategy (sensitive or selective). The use of the stop lists could clearly improve the module based approach in the sensitive approach, whereas this has not been possible for the network based one by this extent. So, a combination of the module based procedure that is restricted to functional categories that are derived as 'predictive powerful' should be employed. This conclusion can also be interpreted as observation that the modules reflect cellular machineries better as enzymatic pathways since the most erroneous category is FunCat 01 'Metabolism'. However, sub classes of metabolism with a good specificity could be be found as, e.g., for the biosynthesis of Cystein (100%), amino-acid metabolism (75%), or the poly-saccharide biosynthesis (80%), albeit with only few candidates (compare Table 6.11 in the appendix) indicating their good representation by functional modules. In the study of Tanay and co-workers on yeast modules using GO annotations, a dependency of the performance due to the chosen GO term and module could also be observed: certain GO terms related to broader processes as sporulation or amino-acid metabolism could be transferred with a high varying specificity between 40%-100% [179] (compare text and supplementary Figure 9 of the publication), reflecting the same effect found here. The procedure proposed herein can serve as a basic frame-work for an automated annotation system using network and module based information of prokaryotic genomes. Such a system could be employed after a first round of homology based annotation to identify proteins related to certain cellular processes. Another application would be the use as a system for target prioritization for further experiments by evaluating modules in which the function prediction performs poorly or results in very unspecific predictions due to lacking primary annotations: the characterization of a few proteins of such modules would give hints to the function of all module members and could therefore save time and costs for experiments.

2.3.3 Detection of virulence genes by module co-membership

Introduction Virulence factors are generally detected *in silico* by two classes of approaches: firstly, by homology to known virulence related genes, and secondly by sequence properties as in the identification of secreted proteins which are identified by the detection of a N-terminal secretion signal. In the first case, factors which are members of yet unknown protein families cannot be detected. The second approach cannot detect virulence related genes that are not secreted as additional secretion system components or proteins involved in virulence related transcriptional control. Furthermore, not for all kinds of secretion systems a sequence based prediction method exists and the question arises if virulence factors (i.e. effector proteins) can be detected by co-membership in modules which are related to virulence. The aim of this analysis is to assess the ability of the idea to uncover chlamydial virulence factors using the functional modules delineated before. This includes two steps: the identification of relevant modules, and the extraction of possible candidates.

Material and methods In this analysis, the modules of the high and medium confidence as well as modules originating from the unfiltered networks have been used in their projection to the individual species. The member proteins have been labeled due to their relatedness to virulence using a list of known and predicted virulence related proteins.

Compilation of a list of virulence related proteins In this analysis data from one representative strain of each species of the pathogenic *Chlamydiaceae* as well as the environmental *Chlamydium*, *P. amoebophila* has been used. A list of *C. pneumoniae* virulence factors compiled from SWISSPROT [198] and the Virulence database [239] , as well as a list of all *Chlamydial* proteins carrying eukaryotic like domains (compare Chapter x) have been kindly provided by Marc Andre Jehl. An additional list of known virulence factors from different *Chlamydiae* has been obtained from Dr. Astrid Horn, University of Vienna. This list has been extended by orthologs from all used *Chlamydiae*. Known Type III effector proteins have been extracted from the EffectiveT3 training set and close orthologs in strains of the same organism have been added to this list [28]. A list of proteins belonging to transport-systems found in the *Chlamydiacea* according to KEGG [193] have been determined by membership to orthologous groups of these systems. These comprise the Type II, III, IV transport systems as well as the Sec-dependent pathway. Putative Type III secreted proteins have been predicted with EffectiveT3 using sensitive settings (cut-off 0.95). Proteins predicted to be transported by the Sec dependent pathway have been determined using SignalP (standard settings) for each chlamydial proteome. All chlamydial proteins have been labeled according to these lists as member of a transport system, as effector or virulence factor, eukaryotic like protein, or predicted by either EffectiveT3 or SignalP. All other proteins have been labeled as 'not related to virulence'
To investigate differences in local network properties for the different kinds of virulence related genes, the degree and a representative edge weight has been computed for each protein. The degree has been defined as the number of direct edges, the typical edge-weight has been defined by the median weight of the direct edges (both in the complete functional network instances for each proteome).

Selection of candidate modules As the first step in this analysis modules that comprise virulence factors must be identified. Since it could be initially observed that the overall amount of virulence related proteins as labeled above is small in comparison to the complete set of proteins, interesting modules have been identified by computing a P-value reflecting these rare events:

- For all proteins encoded in each *Chlamydial* genome, chose the label which is most informative in terms of virulence. Following order has been applied: part of transport system > known Effector > virulence factor, predicted TTSS substrate > predicted substrate of the Sec-dependent pathway > not related to virulence

- let be $L = \{$Transport system component, Effector, Virulence Factor, predicted TTSS substrate, predicted Sec substrate, not related to virulence$\}$ and N the amount of proteins encoded in the genome,

  $\forall x \in L$: compute:

$$freq(x) = \frac{x}{N}; \tag{2.25}$$

The score $S(m)$ for each module M is then computed as:

- let p be a protein with label x_p,

$$S(m) = \prod_{\forall p \in M} freq(x_p); \tag{2.26}$$

The smaller the score $S(m)$, the more likely the module is virulence related. For the numerical computation of the P-Value of an observed module z with score $S(z)$, a Monte-Carlo approach has been used by comparing the resulting score to scores out of several rounds of randomization. The modules are randomized by shuffling the protein identifiers while retaining the module topology (sizes and amounts of modules). The scores for all modules with the same size as module z are computed and the amount of random modules r with $S(r) \leq S(z)$ are counted as c. This has been repeated $n = 10000$ times. The actual P-value is then computed as:

$$Pvalue(m) = \frac{c}{n}; \tag{2.27}$$

Modules with a P-Value ≤ 0.1 have been chosen for the deduction of virulence related candidates.

Creation of candidate lists The delineation of candidates is intuitive: all proteins not yet known as virulence related due to the initial classification are extracted from the virulence related modules. The labels attached to the proteins give evidence if a candidate is probably secreted (with either SignalP or Effective prediction), could serve as possible effector (carrying a eukaryotic like domain and/or is predicted by EffectiveT3), or is a candidate of being generally virulence related if it carries no label.

Depending on these evidences, the candidate lists can be prioritized for experiments depending on the research interest. In order to detect effectors, for example, proteins with an eukaryotic domain and a EffectiveT3 prediction would be good candidates to start.

Results The amounts of virulence related modules per organism is listed in Table 2.14 ranging from four instances in *C. felis* to twelve in *P. amoebophila* (if referred to modules derived from the high confidence network). In *P. amoebophila*, the amount of proteins in the virulence related modules is substantially larger as in the pathogenic *Chlamydiaceae*. This effect contributes to additional modules detected for this organism as well as to additional proteins in modules common to several *Chlamydiales* due to its' larger genome size.

Where the amounts of virulence modules do not drastically differ between modules delineated from the different networks (high and medium confidence, and the complete set), the clustering from the complete network comprise some additional module instances. Many of the additional modules mainly contain unknown proteins. Due to their uncertainty of correctness and their low annotation status, their relevance for virulence cannot be easily judged. However, since the known effector proteins exhibit only weak and few functional interconnections, these modules may give valuable hints to novel virulence related proteins. Only in the clustering of the complete network four characterized effectors can be found in co-membership to other virulence related genes. In *C. muridarum*, the inclusion proteins IncE, IncF, and IncD can be found clustered together in 'module 3'. This module also contains some instances of the GroEL chaperonine family. These are identified as eukaryotic like proteins and several hypothetical proteins. One of the hypothetical proteins found in the module is GI:15835013 (locus tag: TC0394) which is homologous (SIMAP E-value = 6.45^{-30}, alignment over full length) to the inclusion protein IncG in *C. trachomatis*. This protein is supposed to hinder apoptosis of the host cell in *C. trachomatis* [240] and is therefore a good candidate for an effector protein in *C. muridarum*. The effector protein CopN (O34020_CHLCV) can be found in 'module 91' of *Chlamydophila caviae* GPIC. It clusters together with two Type III secretion system related proteins from the list of virulence factors: GI:29840222, a protein of the hrpY/hrcU family and GI:29840190,a flhA homolog, both involved in TTSS mediated secretion due to their annotation in the Pedant database. No other known effector protein nor proteins from the list of virulence factors could be found in module co-membership to a transport system. The amount of not yet characterized proteins which are co-localized in a module with a transport systems is generally low: in the example of *Chlamydia muridarum*, three candidates can be found connected to a transport system by module co-membership (candidates in 'Module 115' and 'Module 16'). The candidate in 'Module 115', TC0092 is already annotated as 'Type III secretion system protein' and has been missing in the initial set of transporter components. TC0780 in Module 16 is probably no virulence factor since it is annotated as 'DNA polymerase'. The same applies to the other candidate of this module, TC0779, which comprises a dephospho-CoA kinase. Therefore, the modules in this settings did not generate suitable candidates of unknown virulence factors. The results for *Chlamydia muridarum* strain Nigg (modules from high confidence network) are listed in Table 2.15. Many modules can be found enriched by components of the transport systems such as 'Module 81' and 'Module 115' which comprise components of the TTSS system or 'Module 16' which harbors several components of the general secretion pathway (Sec dependent pathway). 'Module 1' contains three chaperons related to TTSS transport (TC0055, TC0865 and TC0055), two of them have not been in the list of known virulence factors. The module comprises twelve virulence candidates, three of them with a eukaryotic like domain.

Since not too many effector proteins could be recovered, the local network properties of the virulence related proteins have been computed to assess whether these do not tend to cluster in comparison to arbitrary proteins. In Table 2.13, the averages of local network-properties (degree and edge weights) from proteins of transport systems, known virulence factors, characterized effector proteins, and predicted effector candidates (by either eukaryotic domain, SignalP, or EffectiveT3) are compared against the complete set of proteins. Obviously, the known effector proteins populate areas of the functional interaction networks which have weak structure: their average degree is much smaller (9.07 compared to 226.24) compared to the average of all proteins, however, their average median weight is not such drastically different from the average of all proteins (0.62 compared to 0.65). Known virulence factors, transporter components as well as proteins with an eukaryotic like domain exhibit more structure as indicated by higher averages for degree and weight as the set of all proteins. Proteins predicted to be secreted follow the trend of known effectors and show smaller average degree and weight.

Discussion The Monte-Carlo based approach to detect virulence related modules automatically revealed several modules comprising known virulence related sub-systems, i.e. transporter systems. This basic concept of using a Monte-Carlo simulation to detect significantly enriched sets of difficult or unknown distribution (which comprises the first

Set	Avg. degree	P-Value	Avg weight	P-Value
All proteins	226.24	-	0.65	-
Secretion system	266.04 (+)	5.838E-11	0.86 (+)	2.2E-16
Virulence factor	276.24 (+)	3.973E-16	0.82 (+)	2.2E-16
Effector	9.07 (-)	7.518E-9	0.62 (-)	0.0783
Eukaryotic domain protein	231.16 (+)	0.08265	0.78 (+)	2.2E-16
Predicted T3 substrate	143.78 (-)	2.2E-16	0.53 (-)	2.2E-16
Predicted Sec-pathway substrate	166.00 (-)	2.2E-16	0.59 (-)	0.006235

Table 2.13: Average degree and average median of edge weights of virulence related proteins in comparison with all proteins. (+) indicates higher, (-) lower degree/weight in the set as for all proteins. To judge the significance of these distribution shifts, P-Values have been computed using an one side Wilcoxon rank sum test. 'Secretion system' comprises proteins of the Type II, III, and IV secretion systems. 'Virulence factor' comprises a set of different virulence related genes, 'Effector' comprises known Type III effectors, 'Eukaryotic domain protein' all proteins with a eukaryotic like domain, 'Predicted T3 substrate' comprise proteins predicted by EffectiveT3, 'Predicted T2 substrate' comprise proteins predicted by SignalP

Organism	mod. (high)	prot. (high)	mod. (medium)	prot. (medium)	mod. (all)	prot. (all)
C. muridarum	5	33	5	62	9	129
C. trachomatis	6	36	6	64	9	115
C. abortus	5	37	5	40	8	118
C. caviae	5	50	5	75	9	153
C. pneumoniae	5	29	5	27	7	58
C. felis	4	21	3	15	8	122
P. amoebophila	12	157	9	125	14	199

Table 2.14: Amount of virulence related modules detected by the enrichment analysis of virulence related proteins. 'Organism' is the species, 'mod.(high)' the amount of modules from the high confidence network with a P-Value ≤ 0.1, 'prot.(high)' the amount of proteins covered by these modules. The counts for the medium confidence and the complete networks are given in 'mod.(medium)','prot.(medium)','mod.(all)','prot.(all)'

step in this analysis) has been used in a similar way in several studies [241, 242, 196, 243, 244]. However, an initial knowledge of the systems is necessary to perform the enrichment analysis which can then be used for the candidate generation. The identified modules are limited in their ability to predict virulence factors, especially effector proteins, since these are poorly connected by functional links to these modules and do not cluster together. The connectivity of effector proteins in the functional interaction networks appeared to be very low, especially in terms of functional neighbors as indicated by their low degree. These observations reflect an independent evolution of effector

Locus-tag	Descr.	Se. Sys.	Effector	V. Factor	Eu.	Eff.	Sig.	Candidate
Organism:	Chlamydia muridarum Nigg							
module 81								
TC0853	type III secretion inner membrane protein SctT	+	-	-	-	-	-	
TC0852	type III secretion inner membrane protein SctS	+	-	-	-	-	-	
TC0365	type III secretion inner membrane protein SctV	+	-	-	+	-	-	
TC0848	type III secretion protein SctJ	+	-	-	+	-	-	
module 1								
TC0073	hypothetical protein	-	-	-	-	-	-	*
TC0309	deoxycytidine triphosphate deaminase	-	-	-	-	-	-	*
TC0535	ABC transporter, ATP-binding protein	-	-	-	-	-	+	*
TC0252	type III secretion chaperone	-	-	-	+	-	-	*
TC0390	hypothetical protein	-	-	-	-	-	-	*
TC0217	hypothetical protein	-	-	-	+	-	-	*
TC0771	hypothetical protein	-	-	-	-	-	-	*
TC0379	hypothetical protein	-	-	-	-	-	-	*
TC0055	type III secretion chaperone, putative	-	-	-	-	-	-	*
TC0767	hypothetical protein	-	-	-	-	-	-	*
TC0410	hypothetical protein	-	-	+	-	+	-	
TC0546	type III secretion chaperone, putative	-	-	-	-	-	-	*
TC0918	UDP-N-acetylglucosamine pyrophosphorylase GlmU-related enzyme	-	-	-	-	-	-	*
TC0865	type III secretion chaperone SycD	-	-	+	+	-	-	
module 115								
TC0040	type III secretion system ATPase	+	-	-	-	-	-	
TC0850	type III secretion system protein	+	-	-	-	-	-	
TC0092	type III secretion system protein	-	-	-	+	-	-	*
TC0090	type III secretion system ATPase	+	-	-	-	-	-	
module 113								
TC0302	exodeoxyribonuclease V, alpha subunit	-	-	-	+	-	-	*
TC0021	exodeoxyribonuclease V, alpha subunit, putative	-	-	-	+	-	-	*
TC0008	exodeoxyribonuclease V, gamma subunit, putative	-	-	-	+	-	-	*
TC0007	exodeoxyribonuclease V, beta chain, putative	-	-	-	+	-	-	*
module 16								
TC0858	hypothetical protein	+	-	-	-	-	+	
TC0780	DNA polymerase I	-	-	-	-	-	-	*
TC0779	dephospho-CoA kinase	-	-	-	-	-	-	*
TC0861	general secretion pathway protein D	+	-	-	-	-	-	
TC0045	type III secretion protein SctC	+	-	-	-	-	-	
TC0860	general secretion pathway protein E	+	-	-	-	-	-	
TC0859	general secretion pathway protein F	+	-	-	-	-	-	

Table 2.15: Candidate virulence factors detected by module co-membership in *Chlamydia muridarum* strain Nigg. 'Locus tag' locus tag of protein, 'Descr.' description line, 'Se. Sys.' indicates whether the protein is a known part of a transport system (+) or not (-), 'V. factor' indicates if the protein is a known virulence factor (+) or not (-). 'Eu.' indicates presence (+)/absence (-) of an eukaryotic like domain, likewise 'Eff.' indicates an EffectiveT3 prediction, 'Sig.' a SignalP prediction. 'Candidate' marks proteins of interest by an asterix that are not yet categorized.

proteins for most pathogenic bacteria due to their individual modes of host interaction. These special adaptations imply many effectors to be orphan genes with no or only few orthologs in closely related species. In consequence, they cannot be detected with certainty (i.e. sufficient score) by the genomic context methods which rely on orthology information and regularities between several species. The picture resulting from this analysis matches expectations implied by the biology of virulence factors: the transport systems are well conserved cellular machineries and recovered as such, whereas the effectors should not cluster together with each other since they interact with host proteins but not with other secreted proteins. Interaction with the transport systems could not be frequently detected and is mediated by N-terminal secretion signals (in case of Type III and Sec dependent secretion), a kind of interaction not modeled in the network. Some effector and predicted effector candidates cluster together with chaperons which have been shown to play a role to mediate Type III secretion. Transported proteins are therefore not supposed to be computationally detectable physical interaction partners of the secretion systems, but their functional relatedness to transport systems could possibly bear a sufficent signal for module co-membership. Although, this is not frequently the case, some examples could be detected in the complete network, i.e. one module harboring some of the inclusion proteins. This inclusion proteins play a major role in the intra-cellular survival of the *Chlamydiae* [245] and are therefore indeed virulence related. This module should serve further candidates of inclusion related proteins and could serve as starting point for further experiments to find proteins related to virulence. The low recovery of known effector proteins outlines the importance of dedicated methods to detect them, especially by the identification of their secretion signal or by a detectable function it could play in the host. Both approaches have been tackled in our group: the second part of this thesis deals with EffectiveT3, a software to detect Type III secreted proteins. The systematic detection of eukaryotic like domains has been assessed by Marc Andre Jehl with my participation and represents an additional, secretion system independent approach (data from both analyses have already been used in this chapter).

2.3.4 Evolution of *Chlamydiae* in terms of functional modules

Introduction The availability of the genomes and, in consequence, the predicted functional modules of representatives of the pathogenic and the environmental species, allows to study differences in the equipment and composition of cellular sub-systems reflecting their adaptations to different ecological niches. Following questions are tackled by this analysis: firstly, I asked which modes of module evolution play a role during the adaptation from variable hosts to a very specialized host. For example, the modules could be preferably lost completely or, in contrast, regularily reduced in their size. The environmental *Chlamdiae* exhibit less genome reduction. To get a clue on which functionalities the environmental *Chlaymdiae* still rely but are lost in the pathogenic ones, I investigated which functional modules are lost in the pathogens but still exist in the environmental *Chlamydiae*. Furthermore, three individual example modules are discussed in detail according to their composition in the individual genomes.

Functional modules may contain sub-groups of proteins which are evolutionary more correlated to each other as to the rest of the module, reflecting common evolutionary pressure on parts within the module. For these sub-modules I introduce the term 'cohesive core(s)'. Contrary, proteins which exhibit an irregular pattern of evolution compared to other module members are named 'shell' proteins. In this chapter, a strategy to detect cohesive cores is proposed and the existence of them in the chlamydial modules is assessed. The classification of module members into evolutionary coherent groups allows a more detailed investigation of genome reduction due to the host adaption process in cases where it affect the modules by size reduction and not by complete loss. With the concept of the cohesive cores such reductions can be described in three different ways. These three alternatives are:

- reduction of cohesive and shell parts alike ('irregular' reduction)
- loss of shell parts while retaining cohesive cores ('purifying' reduction)
- loss of cohesive cores while retaining shell proteins ('cohesive' reduction)

The three alternatives are pictured in Figure 2.12. Since the genomes of the environmental *Chlamydiae* are less reduced and their host-adaptation is more variable, their module interior should be more similar to the last common ancestor as it is the case for the pathogenic ones. Therefore, the different events should be detectable when comparing the environmental and the pathogenic *Chlamydiae*.

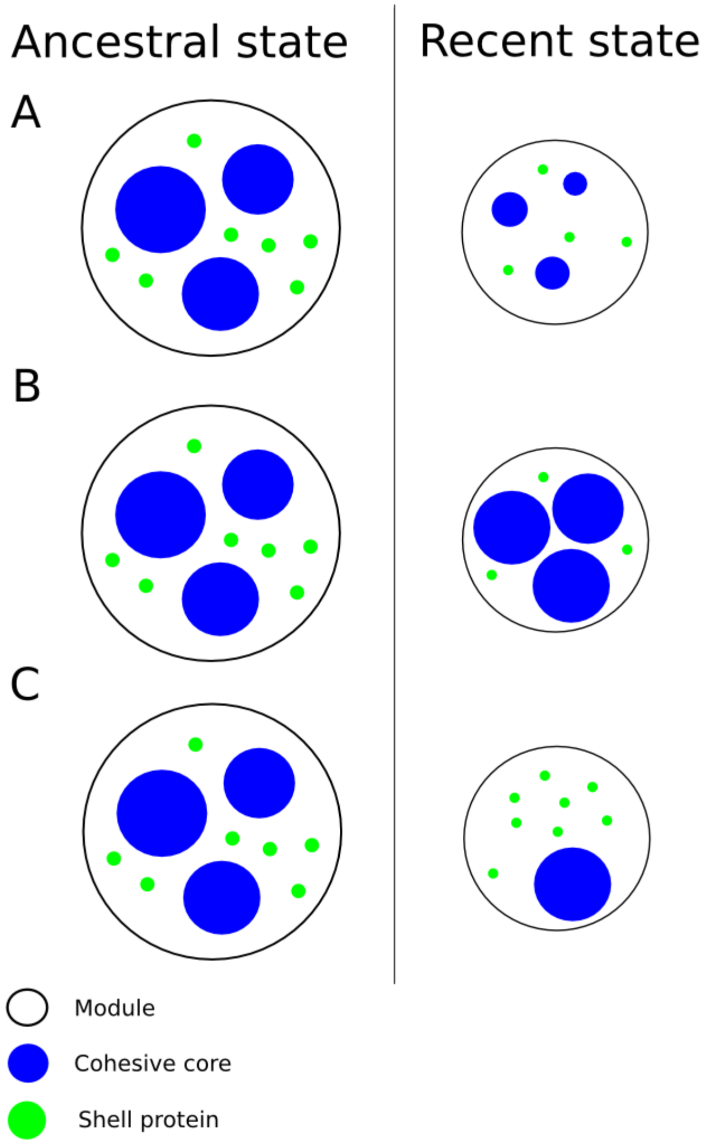


Figure 2.12: Possible scenarios of module reduction due to host adaptation and genome reduction. A: 'irregular reduction' B: 'purifying reduction' C : 'cohesive reduction'

Material and Methods

Data used The organisms used in this study comprise the environmental *Chlamydiae* and one representative strain for each pathogenic *Chlamydiacea.* To avoid circular reasoning, a module set has been created based on a the high-confidence network without links from the co-occurrence method (using the same cut-off and clustering parameters as for the initial high confidence network). This set has been projected to the genomes by orthology assignments as described in Chapter 2.2.6. As phylogenetic tree, the NCBI taxonomy tree as from March 2010 has been used. The orthologous groups as created in Chapter x) have been employed to create phylogenetic profiles covering all prokaryotic genomes in the STRING database and the *Chlamydiae* used in this study.

Detection of cohesive modules I used an algorithm proposed by Campillos et al. [181] to compute a cohesiveness score for a given module. The score is based on the idea to describe a modules' evolution by the most parsimonious explanation of the pairwise evolution of module members. Module members can appear (gene birth) or disappear (gene death), and costs for each of these events can be defined. Given a fixed phylogeny and a profile of the phyletic distribution of the recent module members, the most probable scenario of events can then be computed as the one with the lowest costs [246]. To make tis concept applicable to cohesiveness, the costs are lowered if a *joined* event occurs reflecting the *a priori* assumption of favored cohesive evolution. The resulting tree of events is then scanned for the amount of joined vs. single events and the fraction of joined events, and the average (normalized) costs, are determined. These two variables, which describe the cohesiveness of a module, follow a multivariate distribution with parameters depending on the modules' size, the used phylogeny, and on the used organisms. With this distribution at hand, P-Values can be computed reflecting the probability to see the cohesive behavior of a given module by chance.
The probability function of the multivariate normal distribution is given by:

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} exp[\frac{-1}{2(1-\rho^2)}[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}]]; \tag{2.28}$$

where x=fraction of joined events, and y=normalized parsimony costs. P-Values are computed numerically using a simple Monte-Carlo method as the fraction of random modules with a probability equal or less the observed module probability (10000 data-

points for each module size):
let be s the probability of a module m, $f(x_m, y_m)$, let be c the amount of random modules with $f(x_c, y_c) \leq s$;
the P-Value is then computed as:

$$Pvalue(m) = \frac{c}{10000}; \tag{2.29}$$

An implementation of this algorithm has been kindly received from Sebastian Ullherr who implemented the algorithm during an internship under my co-supervision. In his internship-project, the same data-set as used in the original publication has been used to validate the correctness of the algorithm. To use the algorithm with new data, the parameters of the two dimensional multivariate normal distribution function had to be re-estimated according to the used phylogenetic tree, the set of used organisms, and the used modules which all three differ from the initial study. The parameters have been re-estimated from 50000 randomly shuffled modules (under preservation of the clustering topology). Notably, the cohesiveness is computed over the whole available set of complete prokaryotic genomes and not on the chlamydial genomes alone. This is necessary to gain sufficient resolution as in the related phylogenetic profile method to detect functional links (compare Chapter 1.3.4). The observed cohesiveness may therefore differ from the picture seen when looking at the phylogenetic distributions of a module in chlamydial genomes only. Modules have been classified as 'cohesive' if their P-Value is smaller or equal 0.01, 'variable' otherwise as done in the literature [181, 182].

A strategy to detect cohesive cores I implemented a strategy to split up a module into 'cohesive cores' and 'variable shells'. This procedure is an extension of the approach of Campillos and co-workers described above [181] in which the pairwise cohesiveness scores of module members are used to establish joined groups of proteins which share significant cohesiveness. A module may comprise, beside shell proteins, several cores or none of them. In the cases where shell proteins are absent, the module is completely cohesive itself.

The algorithm to split a module in core and shell is implemented straightforward using graph clustering:

- Create an empty graph G with nodes $n_1..n_m$ where m is the size of the module representing the member proteins.
- Enumerate all pairs (a, b) of module members with $a \neq b$. For each possible pair

c of module-members, compute $p = Pvalue(m)$ as described above.

- If $p \leq 0.01$, establish an unweighted link in G between the two module members.
- Cluster the resulting graph, i.e. by deep-first search for connected components.
- Proteins in components with a size ≥ 2 are labeled as 'core' proteins.
- Proteins remaining singletons are labeled as 'shell' proteins.

A test with two different graph cluster algorithms (MCL and a simple connected component clustering) did not revealed clear differences in the outcome of the clustering since the structure of the input graph is trivial in most cases. In this analysis, the connected component clustering has been further used. Notably, the resulting cores might be not cohesive by themselves (P-Value $\geq$ 0.01). These are treated as *bona fide* cores since they are detectable by the pair-wise cohesiveness criterion. A manual inspection of the reported cores indicated, that they often reflect meaningful sub-units as partners of a complex.

Classification of modules with reduced size In order to quantify the different modes of reduction during host adaptation of the pathogenic *Chlaymdiacea* described in Figure 2.12, the following strategy has been employed:

- all modules categorized as both, 'with core' and 'with shell', have been listed as initial candidates.
- for each of these candidate modules, the environmental *Chlamydium* with the maximal amount (E_{max}) of proteins in it has been chosen. This maximum allocation can be interpreted as most probable representation of the ancient module in the last common ancestor that can be observed in the data of the available recent genomes.
- for each of these candidates the pathogenic *Chlamydium* with the minimal amount of proteins in the module (P_{min}) has been chosen to represent the maximal observable reduced state after host adaptation.
- modules that do not show reduction (i.e. with $E_{max} \leq P_{min}$) have been discarded since they are not informative.
- each module has been labeled as 'cohesive' if it has lost one of its' cores.

- each module has been labeled as 'purifying' if it has lost shell proteins while retaining one of its' cores.

- each module has been labeled as 'irregular' if it has lost shell and core proteins alike but no complete core.

Since the modules may contain several cores, multiple labels are possible.

Assessing the differences in the functional inventories of environmental and pathogenic *Chlamydiae* To compare the functional inventory of the environmental and pathogenic *Chlamydiae* in terms of modules, following aspects are tackled: firstly, a clear definition of a modules' presence or absence in the data of an species of interest must be defined. Secondly, general trends in the differences of the modules found in the pathogenic and environmental *Chlamydiae* in terms of general functional categories are stated by an enrichment analysis. Thirdly, the modules lost in the pathogenic species are annotated and inspected by literature and database re-search on the modules' members. An additional mode of evolution could be variation in the copy number reflecting paralogous modules. The existence of such cases is assessed.

Stating the existence of a functional module in a species The assessment of the presence or absence of a module is not simple in any case: in the case of complete loss, a module absence from a species can be clearly stated, but the definition of a module's existence is somewhat arbitrary if the module is only partially lost, since it is not clear to which point of reduction a module remains functional. The definition of cohesive cores gives an additional criterion which is used to judge the presence of a module. A module is stated as existent in a genome if it is either completely present (i.e. all orthologs of a module have at least one instance encoded), or all of it's cohesive cores can be found. In both cases, one missing protein is accepted to compensate possible missing annotation.

Detection of paralogous modules To assess the existence of paralogous modules, copy numbers of modules have been determined by counting the paralogs of each member of a module. The existence of a certain copy number n has been judged analogous as for the presence criterion described above (all proteins or all cores must be frequent ($\geq n$) , with one member tolerance).

Functional categorization The most frequent functional category in the proteins of a module has been assigned to the module as functional annotation. Enrichment and depletion of functional categories have been assessed between different sets of modules using Fisher's exact test with Bonferroni correction for multiple testing.

Results

Cohesive cores About 20% of the modules show a cohesive behavior as complete module, additional 27% are not cohesive but provide a cohesive core (overall, 91 of 195 modules have a cohesive part). Most modules (162 of 195) provide shell proteins. The average amount of cores/module is 1.2 (only modules with core counted) with a maximum of four cores. From in total 138 cores, 126 exhibit a better P-Value of cohesiveness as their corresponding module indicating that the detection method is reliable.

The genome reduction of pathogenic *Chlamydiacea* An initial comparison of the module sizes in the environmental and pathogenic set is shown in Figure 2.13. In this plot, the maximal observed module sizes in each set are plotted against each other. Many modules (90 instances) can be found on or nearby (with at most one protein missing in the maximum of one of the sets) the diagonal (red line in the plot). In these cases, the module can be found complete in both sets. 49 modules are completely absent in the pathogenic set (the modules at positions x=0), none are completely absent in the environmental set. Modules which cannot be found at either the diagonal or at the axes exhibit a more complicated evolution by reduction (left from the diagonal) or expansion (right from the diagonal) in the pathogenic *Chlamydiacea*. Despite of the total loss of modules, several cases of module reduction can be observed: in 49 cases, the module is reduced in the pathogenic set and in two cases in the environmental. Only thirteen modules cannot be found rather complete in one of the two sets. This picture can also be be observed when comparing single pathogenic and environmental genomes. An example is given in the supplementary Figure 6.2. Notably, this example contains instances at position (0,0) indicating the absence of modules in both sets. This picture indicates two existing aspects of the genome reduction: complete loss exists as well as size reduction. For a further investigation of these aspects, the modules are classified by their cohesiveness and the existence of cohesive cores. The results of this classification are summarized in Table 2.16. The modules comprising cohesive cores are candidates for the analysis of the size reduction of modules and have been classified according to the three possible observations defined in Figure 2.12. In Table 2.17 an example of cohesive

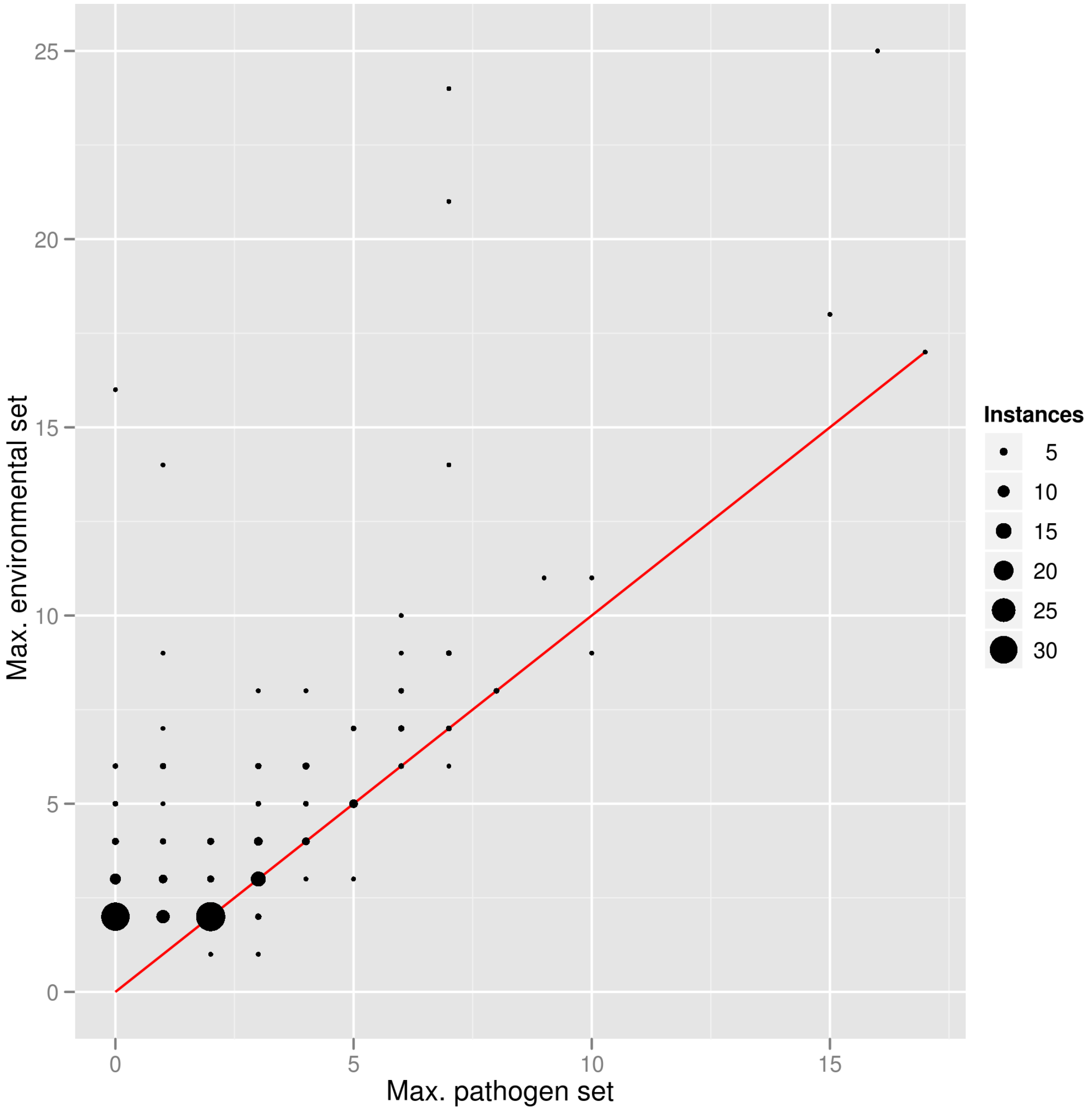


Figure 2.13: General trends in the evolutionary behavior of *Chlamydial* modules. The maximal size of each module detected in the environmental *Chlamydiae* (max.environmental) is plotted against the maximum size in the pathogen set (max.pathogen). The size of the points reflect the number of modules found at each data-point. The red line indicates the diagonal (x=y).

module reduction is shown. Twenty-one modules have been labeled as solely 'purifying', four show loss of a complete core (while retaining another) and a reduced amount of shell proteins and are therefore labeled as 'purifying' and 'cohesive'. Four modules have lost solely a complete core and are labeled as 'cohesive', two have lost a complete core in addition to some shell-proteins and proteins of another core (cohesive+irregular).

Class	Amount modules	Amount proteins
Cohesive	39	182
Not cohesive	156	738
With core	70	548
With cohesive cores	62	495
Cohesive or cohesive cores	91	566
With shell	162	835

Table 2.16: Evolutionary behavior of chlamydial modules. Column 'Set' contains the classification of evolutionary behavior (see text), 'Number modules' the amount of modules in the respective set, 'Number proteins' the amount of proteins in the set.

Profile	Description	Shell/Core
1 1	COG0762 Predicted integral membrane protein	shell
0 0	COG1957 Inosine-uridine nucleoside N-ribohydrolase	shell
1 1	COG0605 Superoxide dismutase	shell
0 1	COG0345 Pyrroline-5-carboxylate reductase	core I
0 1	COG0325 Predicted enzyme with a TIM-barrel fold	core I

Table 2.17: Example for a cohesive module reduction between environmental and pathogenic *Chlamydiae* 'Profile': the reduced phylogenetic profile, the first entry indicates presence (1) or absence (0) in the pathogenic *Chlamydium* that shows the smallest coverage of the module, the second entry indicates presence (1) or absence (0) in the environmental *Chlamydium* that shows the highest coverage of the module. 'Description' COG identifier and description of the orthologous group, 'Shell/Core': the classification of an orthologous group as shell or cohesive core element in the module.

Another seven show no clear tendency and are therefore examples of 'irregular' module reduction. Eleven reduced modules did not contain both, shell and core proteins, and are therefore not counted in this inventory of events. All individual classifications can be found in the supplementary Table 6.15 in the appendix. Examples for each case are given in Tables 2.17,2.18, and 2.19. The first example presents a cohesive reduction of a module. The module comprises one core with two members and three shell entries. One of them could not be found in the two species (the chosen minimal pathogen and maximal environmental *Chlamydium*) (COG1957) but the two others are conserved in both. Contrarily, both core proteins are jointly lost in the pathogen as indicated by the first row of the reduced profile. Due to the presence/absence criterion defined above, this module would, by its' loss of the core, be classified as absent in the pathogen.
The second example shows purifying behavior: in this probably cell-wall related module, the core with four members is conserved but the two shell proteins (a porin and a predicted thioesterase) of the environmental species are lost in the pathogen.

Profile	Description	Shell/Core
0 1	COG3203 Outer membrane protein (porin)	shell
0 1	COG0824 Predicted thioesterase	shell
0 0	COG3064 Membrane protein involved in colicin uptake	shell
1 1	COG2885 Outer membrane protein and related peptidoglycan-associated (lipo)proteins	core I
1 1	COG0823 Periplasmic component of the Tol biopolymer transport system	core I
1 1	COG0848 Biopolymer transport protein	core I
1 1	COG0811 Biopolymer transport proteins	core I

Table 2.18: Example for a purifying module reduction between environmental and pathogenic *Chlamydiae*. An explanation of the columns is given in Table 2.17.

Profile	Description	Shell/Core
1 1	COG0039 Malate/lactate dehydrogenases	shell
0 1	COG2079 Uncharacterized protein involved in propionate catabolism	shell
0 1	COG2513 PEP phosphonomutase and related enzymes	shell
1 1	COG0045 Succinyl-CoA synthetase, beta subunit	core I
0 1	COG0372 Citrate synthase	core I
1 1	COG0074 Succinyl-CoA synthetase, alpha subunit	core I

Table 2.19: Example for an irregular module reduction between environmental and pathogenic *Chlamydiae* Example for a purifying module reduction between environmental and pathogenic *Chlamydiae*. An explanation of the columns is given in Table 2.17.

In the third module (representing a part of the central metabolism and the Citrate-Cycle) shell and core proteins are lost, the module is therefore an example of irregular reduction.

The additional functional equipment of environmental *Chlamydiae* The counts for modules conserved in either the pathogenic, environmental, or in both sets using the criterion that either the complete module or at least the cohesive cores are present is summarized in Table 2.20. Notably, these counts differ from the values given above due to the core completeness criterion. Also, not the maximal module sizes in one of the sets but the instances in all *Chlamydiae* are counted. The majority of detected modules exist in all *Chlamydiae* and define a set of functionalities which are basic to all organisms investigated (89 modules comprising 338 proteins). No modules can be found as specific loss or gain in a single pathogenic *Chlamydium* but two modules are exclusively existent in the pathogenic *Chlamydiaceae*. One of them is related to the COG category 'Repli-

Set	Number modules	Number proteins
Complete	89	338
Both	6	43
Environmental only	63	315
Pathogen only	2	8
Specific loss to a pathogen	0	0
Specific loss to an environmental	5	35
Specific to a pathogen	0	0
Specific to an environmental	19	85

Table 2.20: Distribution of modules in the *Chlamydiae*. Column 'Set' contains the classification, with:

- Complete: Modules existing in every *Chlamydium* under investigation
- Both: Modules existing in pathogenic and environmental *Chlamydiae* under investigation, but not in every species
- Environmental only: Modules existing only in several environmental *Chlamydiae*
- Pathogen only: Modules existing only in several pathogenic *Chlamydiacea*
- Specific loss to a pathogen: Modules only missing in one pathogenic *Chlamydium*
- Specific loss to an environmental: Modules only missing in one environmental *Chlamydium*
- Specific to a pathogen: Modules only existing in one pathogenic *Chlamydium*
- Specific to an environmental: Modules only existing in one environmental *Chlamydium*

'Number modules' the amount of modules in the respective classification, 'Number proteins' the amount of proteins in the classification.

cation, recombination and repair' but a clear function cannot be stated. The other one is related to 'Intracellular trafficking, secretion, and vesicular transport' and provides two flagellar components (FlhA, FlhB) of which at least on can be related to Type III secretion (FlhA [247]) and a sigma factor (COG1191, FliA) that controls flagellar genes in *E. coli* [248] and *B. subtilis* [249]. The fraction of modules which are features of the environmental species only is quite high as expected due to their larger genome sizes and 63 modules comprise a set of modules conserved in the environmental and lost in the pathogenic *Chlamydiae*. The enrichment analyses on the functionality of these sets of modules gives a first insight in the functional differences between the pathogenic and the environmental *Chlamydiae* which are listed in Table 2.21 . The Modules related to translation, carbohydrate metabolism, and related to secretion are commonly conserved and appear therefore enriched in the set of modules common to all *Chlamydiae* ('Both' and 'Complete'). In comparison to the complete set, the environmental species have more abilities for energy production and conversion. Modules which exists in the environmental set only are enriched in the amount of modules related to the production

Function	**P-Value corrected**	**enriched/depleted**
Complete		
Intracellular trafficking, secretion, and vesicular transport	0.0147	+
Translation, ribosomal structure and biogenesis	2.20^{-4}	+
Carbohydrate transport and metabolism	0.0290	+
Amino acid transport and metabolism	0.0395	-
Both		
Translation, ribosomal structure and biogenesis	7.59^{-4}	+
Environmental only		
Cell motility	0.0263	-
Translation, ribosomal structure and biogenesis	9.70^{-6}	-
Energy production and conversion	3.40^{-7}	+
Specific to an environmental		
Signal transduction mechanisms	0.0044	+
Inorganic ion transport and metabolism	0.0180	+
Cell wall/membrane/envelope biogenesis	0.0218	-
Translation, ribosomal structure and biogenesis	0.0013	-
Amino acid transport and metabolism	1.91^{-4}	+
Specific loss to an environmental		
Translation, ribosomal structure and biogenesis	9.64^{-4}	+

Table 2.21: Enrichments and depletions of COG functional categories in the different sets of phyletic distributions. Each set is compared against all other modules. P-Value corrected is the corrected P-Value, 'enriched/depleted' indicates whether the functional category can be found enriched (+) or depleted (-) in the set. Only significant results are shown (with P-Value corrected ≤ 0.05).Column 'Set' contains the classification,
with:

- Complete: Modules existing in every *Chlamydium* under investigation
- Environmental only: Modules existing only in several environmental *Chlamydiae*
- Specific loss to an environmental: Modules only missing in one environmental *Chlamydium*
- Specific to an environmental: Modules only existing in one environmental *Chlamydium*

Classifications which did not return significant enrichments/depletions are not shown.

of energy reflecting an improved capability to gain energy. The translational functionalities appear depleted in this set since these functionalities are merely existent in all *Chlamydiae*. Single environmental *Chlamydiae* have increased capabilities related to signal transduction and amino-acid metabolism as these categories can be found enriched in the set of specific environmental modules. Modules related to cell-wall creation and translation are depleted in this set since they do not appear specific in one species but are implemented in several *Chlamydiae*. *Waddlia chondrophila* lacks some ribosomal genes and a translation elongation factor rendering three modules related to 'Translation, ribosomal structure and biogenesis' absent in the set 'Specific loss environmental'. The absence of these genes in *Waddlia* is rather improbable and the observed picture could be the effect of missing gene annotation since the genome is not closed yet. The broad functional profiling by the COG annotations does not provide a detailed information of the functional repertoire that is commonly lost in the pathogenic species. In order to get a condensed overview of these functionalities, the modules classified as 'environmental only' have been further annotated. The result of this annotation is listed in Table 2.23. For each module an annotation of a probable function and a phyletic pattern describing the presence/absence of the module as defined above are given. Two modules (module 140 and 80) could be related to the stress repair response and exist in all environmental species, several are related to amino-acid metabolism of Histidine, Glutamate, Cystein, and Glycine which are not present in *S. negevensis*. As indicated by the enrichment analysis, several modules dedicated to energy production (modules 8,40,5,44,85) are present in the set. A part of the Type IV secretion system is covered by module 186 and is present in *P. amoebophila* and *S. negevensis* as described earlier [71] (citations needed!), as well as the Twin-Arginine transport system which exists in *P. amoebophila* and *P. acanthamoebae*. Five ABC transporters are absent in the pathogenic set: their putative substrates cover proline and glycine, spermidine and putrescine, Fe3+-siderophores, polysaccharide-polyol phosphate, and a multidrug transporter. Several modules with no clear function are present in all environmental *Chlamydiae* (modules 154,99,51,131,21,186,44,166). Two modules with transposable elements can also be ex-

Index	**NCBI taxonomy-id**	**Name**
1	243161	Chlamydia muridarum
2	315277	Chlamydia trachomatis
3	218497	Chlamydophila abortus
4	227941	Chlamydophila caviae
5	115711	Chlamydia pneumoniae
6	264202	Chlamydophila felis
7	264201	Protochlamydia amoebophila
8	71667	Waddlia chondrophila only Waddlia contigs final
9	83561	Simkania negevensis final
10	83552	Parachlamydia acanthamoebae UV7 final

Table 2.22: Explanation of the phylogenetic profiles use in Table 2.25,2.26,2.27. 'Index' is the index in the profile. 'NCBI taxonomy-id' the NCBI id, 'Name' the name of the organism.

clusively found in the environmental set (module 184 and 189): module 184 contains COG2801 containing proteins with viral Integrase domain (as described in the Interpro entry IPR001584) and COG2963 providing proteins with a IS3/IS911 type Transposase described in Interpro in entry IPR002514. Module 189 provides proteins with the Interpro entry IPR001959, the Transposase IS891/IS1136/IS134 and IPR002686, IS200-like Transposase.

Name	Phyletic Pattern	Category	Annotation
module 8	1101	C	ATPase and ATP-synthase
module 40	1111	C	Hemecopper-type cytochromequinol oxidase
module 5	1001	C	NADH:ubiquinone oxidoreductase
module 44	1111	C	perhaps related to Fe-S cluster assembly [250]
module 85	1011	C	probably part of glutamine degradation (contains malic enzyme) and Pta-Ack pathway [251]
module 115	0110	C	related to glycerol utilisation
module 143	0101	C	related to Krebs cycle
module 111	0101	C	unclear function
module 119	0110	E	ABC-type proline glycine betaine transport systems
module 46	1111	E	ABC-type spermidine putrescine transport system
module 97	1110	E	contains Type V autotransporter
module 72	1111	E	function unclear, contains several dehydrogenases (for aldehyde, choline, alcohol, and proline)
module 26	1101	E	Glycine cleavage system
module 83	0101	E	part of cystein metabolism
module 114	0101	E	part of histidine metabolism: Histidine, Urocanate, Imidazolonepropionase
module 174	1111	E	protease related
module 134	0101	E	related to glutamate metabolism
module 181	0101	E	related to ornithine metabolism
module 124	0100	F	part of purine de novo synthesis pathway [252][253]
module 172	1010	G	ABC transporter (export) of polysaccharide-polyol phosphate
module 141	1101	H	part of pyridoxine biosynthesis
module 112	0111	H	related to nicotinamid metabolism
module 146	1101	H	related to SAM metabolism
module 91	0101	H	unclear function
module 21	1111	I	Biotin related, exact function unclear
module 186	1111	I	membrane associated
module 6	0101	I	related to lipid metabolism, exact function unclear
module 20	1111	J	related to translation
module 71	1111	J	wobble (contains Queuine related enzymes, by description)
module 108	1101	K	unclear function, pseudouridine pathway related
module 140	1111	L	contains RecN, a SOS gene [254]
module 80	1111	L	contains SOS response gene and DNA polymerases, probable DNA repair module
module 116	1001	L	DNA methylation related
module 184	1100	L	Transposase and inactivated derivatives
module 189	0110	L	transposases
module 165	1111	L	unclear function
module 39	1101	M	contains syalic acid synthase, related to cell wall [255]
module 84	1011	M	dTDP sugar metabolism
module 2	1111	M	unique function unclear, many membrane related proteins
module 9	1101	O	function unclear, contains chaperones and proteases
module 173	0101	P	ABC-type Fe3+-siderophore transport system (incomplete)
module 66	1101	P	Cbb3-type cytochrome oxidase
module 155	0111	P	K+ transport system (Trk-type K+ transport system)
module 127	1000	P	related to iron conservation (ferritin),
module 138	1011	R	unclear function
module 37	1001	R	unclear function
module 79	1111	R	unclear function
module 32	1111	R	unclear function
module 82	1101	R	unclear function
module 23	0001	R	unclear function, contains two chaperons (chaperone families: HSP10, HSP60) and two Aspartate, tyrosine, and aromatic aminotransferases
module 154	1111	S	unclear function
module 99	1111	S	unclear function
module 51	1111	S	unclear function
module 104	1101	S	unclear function
module 1	0111	T	dedicated function unclear
module 156	1001	T	might be related to virulence (stress response protein UspA [256] and a Kef-type K+ efflux system)
module 168	1010	U	part of Type IV transport system
module 153	0101	U	Sec-independent secretory pathway, Twin-Arginine system
module 130	1111	V	partial ABC-type multidrug transport system
module 74	1010	W	probable module of transport and virulence (adhesin transporter and bacteriocin exporter)
module 131	1111	-	peptide methionine sulfoxide reductase, related to oxidative stress response [257] and associated protein
module 166	1111	-	unclear function
module 193	0110	-	probably related to antioxidant defense, contains Peroxiredoxin

Table 2.23: Modules specific to the environmental *Chlamydiae*. 'Name' the module identifier, 'Phyletic Pattern' the phylogenetic occurrences in the environmental set (positions as described in Table 2.22, only the positions 7-10 of the environmental *Chlamydiae* are used.), 'Category' the most frequent NCBI functional category in the module, 'Annotation' a by hand annotation of module functionality.

The evolution of the pathogenic *Chlamydiae* may include variations due to the copy numbers of a module reflecting duplication events. In our data, five cases of a reduction in copy number in the pathogenic in comparison to the environmental species, and one pathogenic specific duplication could be detected. These modules are listed in Table 2.24.

The variation comprises mainly (four of the six cases) ABC transporter, one module of unclear function, and one related to RNA processing.

Name	Copy Numbers	Annotation	F
98	111111 2125	ABC transporter, substrate: antimicrobial peptides	+
109	111111 0002	ABC transporter, substrate: (polar) amino-acids	+
120	111111 2101	ABC transporter, substrate: nitrate, sulfonate, bicarbonate	+
36	222222 2111	ABC transporter, substrate: dipeptide, oligopeptide, nickel	-
178	111111 0022	unclear function	+
182	111111 2112	RNA processing	+

Table 2.24: Modules with copy number variation between the environmental and pathogenic set. 'Name' the module identifier, 'Copy Numbers' the count of occurrences of each module member in the organisms as profile (positions as described in Table 2.22), 'Annotation' a by hand annotation of module functionality, 'F' indicates whether the module is more frequent in the environmental (+) or pathogen set (-).

Individual evolutionary fate of modules: three case studies To illustrate individual differences of modules between environmental and pathogenic *Chlamydiae* that do affect parts of modules rather than complete gains or losses, I discuss three examples in more detail: the module covering the F0F1-type and the archaeal/vacuolar (V-type) ATPase, a module describing the Tol-Pal system, and a metal ion ABC transporter.

Example 1: two kinds of ATPases The module 'module 8' listed in Table 2.25 comprises two evolutionary related complexes organized in separate cores. One, the archaeal or vacuolar type H+ ATPase is existent in all *Chlamydiae*, the other one, the F0F1 type ATP synthase, is absent in the pathogenic *Chlamydiaceae* and in *Simkania negevensis*. Both entities could be seen as separated modules. However, they are interconnected by a common component, the orthologous group COG0636 which represents subunit C of the F0F1 type ATP synthase as well as the subunit K of the V-type H^+-ATPase reflecting their common evolutionary origin [258][259]. It is unclear whether the V-type ATPase acts as ATP synthase in *Chlamydiae* or holds up a proton gradient in the membrane [66]. The additional F0F1 type ATP synthase in the three environmental species suggest the ability to create additional ATP by oxidative phosphorylation [71] and might reflect a less host dependent life-style in terms of energy parasitism. The orthologous group COG0545 represents FKBP type peptidyl-prolyl cis-trans isomerases which act as chaperons [260]. This entry is probably attached to the module by a false positive link since no link to energy production or ATPases could be found in the literature.

Profile	Gene description	Core/Shell
111111 1111	COG0636 F0F1-type ATP synthase, subunit c/Archaeal/vacuolar-type H+-ATPase, subunit K	shell
111111 1111	COG0545 FKBP-type peptidyl-prolyl cis-trans isomerases 1	shell
111111 0111	COG1390 Archaeal/vacuolar-type H+-ATPase subunit E	shell
111111 1111	COG1269 Archaeal/vacuolar-type H+-ATPase subunit I	core I
111111 1111	COG1155 Archaeal/vacuolar-type H+-ATPase subunit A	core I
111111 1111	COG1394 Archaeal/vacuolar-type H+-ATPase subunit D	core I
111111 1111	COG1156 Archaeal/vacuolar-type H+-ATPase subunit B	core I
000000 1101	COG0711 F0F1-type ATP synthase, subunit b	core II
000000 1101	COG0055 F0F1-type ATP synthase, beta subunit	core II
000000 1101	COG0356 F0F1-type ATP synthase, subunit a	core II
000000 1101	COG0224 F0F1-type ATP synthase, gamma subunit	core II
000000 1101	COG0712 F0F1-type ATP synthase, delta subunit (mitochondrial oligomycin sensitivity protein)	core II
000000 1101	COG0056 F0F1-type ATP synthase, alpha subunit	core II
000000 1100	COG0355 F0F1-type ATP synthase, epsilon subunit (mitochondrial delta subunit)	core II

Table 2.25: Example module with cohesive cores I. The module comprises two ATP related complexes. 'Profile' describes the phylogenetic pattern in *Chlamydiae*. The first six entries describe the existence of orthologs in the pathogenic, the second four in the environmental set (0=does not exist, 1=does exist). 'Gene description' comprises COG entry and description, 'Core/Shell' the classification into core and shell. To distinguish different cores, they are numbered with roman numerals.

Example 2: A Tol-Pal system related module The module 'module 35', listed in Table 2.26, comprises components of the Tol-Pal system. This system is related to cell division and membrane stability and colicin (a proteinaceous toxin of bacterial origin which attacks other bacteria) transport, and necessary for infection by phages in

Vibrio cholerea, and has been shown to be related to virulence in *Erwinia chrysanthemi* [261][262][263][264]. The units in the cluster comprise following components of the system: COG3064 is TolA, COG0848 is TolR, COG0823 is TolB, COG0811 is TolQ, members of COG0824 (YbgC) carry a domain which is annotated as Tol/Pal system associated in Interpro (IPR014166 Tol-Pal system-associated acyl-CoA thioesterase). COG3203 is OpcP an outer membrane porin, COG2885 a membrane protein carrying an OmpA domain. The phyletic pattern in *Chlamydiae* shows complete conservation of the core module (TolB,TolR,TolQ, and COG2885) in all species. The TolA component, YbgC, and OpcP are classified as shell-proteins and completely absent in pathogenic *Chlamydiaceae*. Their appearance in the environmental set exhibits an irregular pattern. This might reflect either differences in cell-division and the maintenance of membrane stability in the environmental species, or in other functionalities as virulence. An interesting hypothesis would be that these shell components are responsible to enable transport of colicin or other bacteriocins. In fact, many colicins need the TolA component to be transported in *E. coli* as summarized in the review of Lazzaroni and co-workers [263]. A TolA ortholog could be found only in the genome of *Parachlamydia acanthamoebae UV7*. The TolA component also plays a major role in the uptake of filamentous phage DNA [265]. Lateral exchange by mobile phage DNA has been shown to exist for the obligate intra-cellular bacterium *Wolbachia* with other bacteria in co-infected insect cells [266]. However, both aspects, the transport of colicines as well as phage uptake, should play a minor role in the chlamydial life-style which actively

Profile	**Gene description**	**Core/Shell**
000000 0100	COG3203 Outer membrane protein (porin)	shell
000000 0111	COG0824 Predicted thioesterase	shell
000000 0001	COG3064 Membrane protein involved in colicin uptake	shell
111111 1111	COG2885 Outer membrane protein and related peptidoglycan-associated (lipo)proteins	core I
111111 1111	COG0823 Periplasmic component of the Tol biopolymer transport system	core I
111111 1111	COG0848 Biopolymer transport protein	core I
111111 1111	COG0811 Biopolymer transport proteins	core I

Table 2.26: Example module with cohesive cores II. The module comprises the Tol-Pal system. 'Profile' describes the phylogenetic pattern in *Chlamydiae*. The first six entries describe the existence of orthologs in the pathogenic, the second four in the environmental set (0=does not exist, 1=does exist). 'Gene description' comprises COG entry and description, 'Core/Shell' the classification into core and shell. To distinguish different cores, they are numbered with roman numerals.

happens mainly in the separated inclusion. If these functionalities exist, they should be found more likely in the environmental species since the less developed immune defense of their hosts favors the encounter of other bacteria in comparison with the immune system of vertebrates.

Example 3: A nickel ABC transporter The module 'module 63' shown in Table 2.27 represents an ABC transporter system of metal ions. Although sequence analysis can give insights into the substrate specificity of metal-ion transporters [267], the orthologous grouping of the ABC transporter components as well as the corresponding regulators does not necessarily resolve substrate specificity. However, the most probable substrate should be zinc since this key-word appears in most of the COG descriptions. In consequence, I describe the system as *bona fide* zinc ABC transporter system. The maintenance of Zn2+ homeostasis is crucial for the bacterial cell to avoid toxic concentrations of this metal ion on the one hand and to provide sufficient support of zinc which is part of many proteins on the other hand [268]. The uptake of sufficient zinc (and other metal ions) is a special problem for intra-cellular bacteria since the concentrations of these in the eukaryotic cells are low. Especially the concentration of iron is reduced in eukaryotic host cells as part of the immune defense against parasites [269]. Deletion of the zinc uptake abilities has been shown to diminish the ability of infection for *Salmonella* by Ammendola et al. [270]. In many bacteria, Zur proteins, homologs of the iron dependent Fur proteins regulate the expression of this transporter [271, 268]. The absence of obvious Fur-like (and therefore of Zur) regulators is a known feature of *Chlamydia trachomatis* and one Fur like protein has been detected by careful analyses of functionally unassigned open reading frames [272]. As summarized by the example of iron uptake in a review of Rodriguez and Smith [273], Fur and TolR can be seen as functional homologs for metal uptake regulation. The module contains an ABC transporter (annotated as specific for Zn2+, Mn2+) as core and in its shell two related regulatory proteins (the orthologs of Fur like regulators represented by COG0735, existent in *P. amoebophila* and *W. chondrophila*, and the orthologs of TroR regulators COG1321 apparent only in *P. acantomoebae*). The phyletic pattern of the two corresponding orthologous groups in the environmental *Chlamydiae* indicates alternative use of the systems in the different species. The absence of both kinds of regulator proteins in the pathogenic *Chlamydiacea* and in *S. negevensis* can be explained either by the use of remote homologs or analogs of Zur as shown for the example of Fur in *C. trachomatis* or another functional replacement. The observation could also be explained by the minor

Profile	Gene description	Core/Shell
000000 1100	COG0735 Fe2+/Zn2+ uptake regulation proteins	shell
000000 0001	COG1321 Mn-dependent transcriptional regulator	shell
111111 1111	COG0803 ABC-type metal ion transport system, periplasmic component/surface adhesin	core I
111111 1111	COG1108 ABC-type Mn2+/Zn2+ transport systems, permease components	core I
111111 1111	COG1121 ABC-type Mn/Zn transport systems, ATPase component	core I

Table 2.27: Example module with cohesive cores III. The module comprises a nickel ABC transporter. 'Profile' describes the phylogenetic pattern in *Chlamydiae*. The first six entries describe the existence of orthologs in the pathogenic, the second four in the environmental set (0=does not exist, 1=does exist). 'Gene description' comprises COG entry and description, 'Core/Shell' the classification into core and shell. To distinguish different cores, they are numbered with roman numerals.

need to regulate the zinc uptake due to more stable environmental conditions as they exist in a defined environment as a certain host cell.

Discussion Around 20% of the chlamydial modules can be found cohesive in this analysis, much less compared to around 40% of cohesive modules in a general bacterial set of modules reported by Campillos et al. [181]. This discrepancy may be biologically interpreted as disproportionally higher loss of cohesive modules than of non-cohesive ones in *Chlamydiae* compared to a 'complete' set of prokaryotic modules. This interpretation is in congruence with the concept of module cohesiveness: a trend to common evolutionary fate of the module members implies that non-cohesive modules should be less vulnerable to partial loss of members and only 'essential' cohesive modules remain. In consequence, non-cohesive modules should be more frequently observed in the reduced genomes of the intra-cellular *Chlamydiae*. This is in congruence with the domino-like loss observed by Dagan and co-workers [91] in other intra-cellular bacteria.

Modules that are reduced in size in the pathogenic set exhibit a certain tendency towards purifying and cohesive reduction which can be found in together 31 modules compared to 18 modules with irregular pattern. Both, purifying and cohesive losses are signs of a cohesive behavior since the purifying reduction releases those parts of modules which are indispensable for the survival of the organism and are therefore cohesively kept. In consequence the remaining conserved cores appear jointly (i.e cohesively) kept. This observation supports the definition of a modules' presence as the existence of its' cohesive cores which has been applied in this study. The amount of modules common to all *Chlamydiae* comprises 338 proteins which is less than the smallest bacterial genome sequenced so far (*Mycoplasma genitalium*) with 470 genes [274], but more as the theoretical minimal genome with 256 genes proposed by Mushegian and co-workers [275]. The amount of chlamydial core modules and genes is therefore in the expected order of magnitude: *Chlamydiae* comprise functionalities which are necessary for their life-style (as genes related to virulence) which are not part of a 'minimal' genome. The functional variance between the environmental and pathogenic species leads to a reduction of the amount of common modules covering less genes in their intersection as can be found in *Mycoplasma*. The pathogenic *Chlamydiae* have only few specific modules which could not be found in the environmental ones indicating a reduced ability to acquire novel genetic material by horizontal gene transfer. Contrarily, the Type IV secretion system which can transport DNA molecules as well as two transposon related modules exist in the environmental species, all indicating a possible participation on the exchange of DNA. Such an exchange could likely happen in an host environment comprising co-infection with other bacteria or phages as in the amoeba [276, 72]. The functionalities found exclusively or in higher copy number in the environmental species comprise several ABC transporter systems indicating a more versatile interplay of the environmental *Chlamydiae* with the environment as the multidrug transport system indicating the need to tolerate the existence of multiple toxins [277], or a Fe^{+}-siderophore transporter as additional way to acquire iron in order to hold iron homeostasis. The environmental species might differ in the modes of SOS stress response induced by UV radiation and other agents of stress such as starvation [278][279] as indicated by their two additional modules that are related to this process.

3

Sequence based prediction of Type III secreted proteins

3.1 Motivation

Among the different secretion systems existing in bacteria, the Type III secretion system is of special interest, as it is encoded in the genomes of many, mainly pathogenic or symbiotic, Gram-negative bacteria and is a key factor for the virulence of pathogens. The identification of novel effectors had to be done experimentally e.g. by translocation assays using fusion proteins of a putative effector with a reporter gene [35, 34, ?, 62] or detection of effectors in the culture supernatant [35]. In many of these studies, prior information is derived computationally from the genome or from protein sequences to create candidate lists of putative effectors before testing them in an appropriate assay [?, 35]. However, none of these methods is either exhaustive or generally applicable and until recently, no sequence based, general method to identify its substrates have been available since the signal underlying the substrate recognition has been unknown. Such a method is desired to short-cut the detection of novel effectors and gives the opportunity to learn the molecular properties comprising the secretion signal. In this chapter, EffectiveT3 is described, a prediction software to detect proteins transported by the TTSS, as well as novel insights into the molecular shape of the recognition signal. EffectiveT3 has been published in Plos Pathogens [28].

3.2 EffectiveT3- a software to predict and investigate TTSS substrates

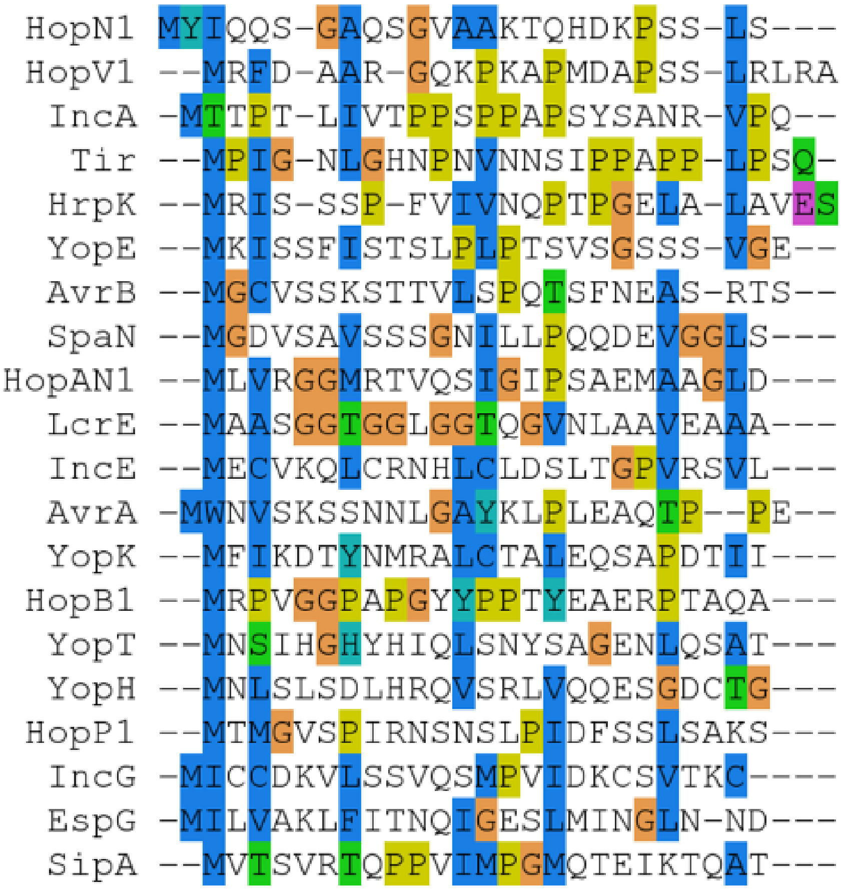
3.2.1 Common features of known effector proteins

As first step to create a prediction method for TTSS substrates, common traits describing the signal must be derived from known effector proteins. These traits could describe features of the signal which are detectable and predictive for unseen effector proteins.

Material and methods As data basis, all known type-III effector proteins have been collected manually from the literature which have been shown to be specifically transported by the TTSS excluding proteins that are part of the TTSS needle complex although some of them are transported by the TTSS and data from large scale screens. The latter have been excluded since the data from these screens might contain a certain rate of false positives. The resulting data set contains 48 proteins comprising the taxa Chlamydia (17 sequences), Salmonella (9 sequences), Yersinia (15 sequences), Escherichia (7 sequences). A representation of this set with only one member of each orthologous group has been created separately. The effectors are listed in the supplementary Table 6.14. The sequences were downloaded from SWISSPROT/UNIPROT [280] (version as downloaded on 07/30/2008) or, if not contained there, downloaded from RefSeq [281] (version as downloaded on 07/30/2008). These sequences have in common that they are effectors in eukaryotic host cells and are further referred to as 'animal pathogen set'. A separate 'plant symbiont set' consisting of 52 known *Pseudomonas* effector proteins has been downloaded from the *Pseudomonas syringae* Genome Resources database [282] (Hop virulence protein/gene database, downloaded on 07/30/2008). Since the N-termini should play an important role in the deduction of a signal, their correctness must be stated. This has been done for the cases, where homologous sequences can be found by validating the gene starts by manual inspection of multiple sequence alignments with their homologs. Negative training sets of non-effectors have been created by randomly choosing proteins from the organisms represented in the animal pathogen and plant symbiont sets not containing the known effectors. Theoretically, the negative set might therefore comprise unknown effectors. I did not filter the negative set to further criteria (as using only 'housekeeping' genes less likely to be effectors) in order not to introduce any bias into the negative set which could be recovered wrongly later. Each negative set is twice as large as its corresponding positive set. This procedure has been repeated five times in order to enable investigations on the influence of the negative set on the prediction. Protein sequences from completely sequenced genomes of the species comprising effectors as well as of other gram(-) and gram(+) *Bacteria*, and *Archaea* were downloaded from RefSeq (version as downloaded on 07/30/2008) [281]. The data sets were classified into organism with and without TTSS by manual search in the literature for the case of gram(−) bacteria or generally classified as "without TTSS" in the case of gram(+) bacteria and archaea. A list of proteins building the TTSS system has been obtained by full-text searches against the SIMAP [283] databases using the gene-names of the TTSS compounds as given by KEGG [284]. The training set might comprise redundancy in terms of proteins of the same orthologous group. This might lead to an over-optimistic behavior of a prediction method in the end if orthologs are easily recovered due to their evolutionary relatedness but not the actual secretion signal. Therefore, an all-against-all comparison of the full length-sequences using the Smith-Waterman algorithm [285] as implemented in the Jaligner package was performed [286]. For each pair, a similarity score S_{ratio} by dividing the alignment score by the self-score is computed and sequences are iteratively grouped if they show a S_{ratio} value greater or equal 0.15. This measure is similar to the measure used by Lerat et al. in a study of genome repertoires in bacteria [287] and has been adjusted to maximal sensitivity in the detection of putative orthologs. This procedure allowed to group equivalent proteins and from each of these groups, only one representative is chosen when needed. The signal in the N-termini could be encoded in their secondary structure. To asses this possibility, secondary structure predictions have been performed using the PSIpred-software [288] applied to the whole sequences. PSIpred can be applied using alignments to conserved sequences as extrinsic information using PSI-BLAST [289]. Since the N-termini did not show similarity to sufficient known proteins that could be used as extrinsic information for PSIpred, only the *ab initio* prediction without alignment information has been employed. The fraction for each predicted class in the N-termini have been counted as possible input feature. To assess possible existing conserved residues in the N-terminal sequences, multiple alignments have been created using two different methods: ClustalW (Version 2.0.5) [290], and Muscle (Version 3.7) [291] with standard parameters. Several rounds of ten randomly chosen sequences from the sets of known effectors have been aligned and manually inspected. Enrichments and depletions of amino acid properties (frequency, frequency of its representations in a reduced alphabet, frequency of secondary structure properties) have been performed by an one sided Mann Whitney test with $p<=0.5$ using the Prompt software (Protein Mapping and Comparison Tool

[292], which employs the statistic software R [293]. Effector proteins could be linked by predicted functional interactions. Predicted functional interactions between orthologous groups containing effector sequences and selected TTSS sequences were obtained from the STRING database [294] (Version 7.1 as downloaded on 10/03/2007). Links from genomic context methods (conserved neighborhood, gene fusion, phylogenetic profiles) were used, the others were discarded in order not to uncover known knowledge as from literature mining. Links with a confidence score less than 0.5 have been discarded and the connected proteins were grouped. To investigate the ability to find effectors by genomic proximity (which is different from conserved neighborhood), complete genome and proteome data for the known effectors has been downloaded from the KEGG database [284] (release 2009/01/19). Components of the TTSS have been identified by their association to the KEGG Orthologous Groups (KO) belonging to the TTSS reference pathway KO03070 (K03219..K03230). Genomic neighbors of a certain distance to known effectors have been extracted from the KEGG data and grouped by their associated KO.

Results A systematic comparison of multiple alignments of the effector N-termini did not reveal any regularities, i.e no conserved residues. The presence of conserved positions would be indicative of a common sequence motif or domain signature. Such a well conserved entity can therefore be excluded as possible signal. An example alignment of 20 effector N-termini is given in Figure 3.1. In Figure 3.2, the functional neighborhoods of two TTSS related proteins (the TTSS related chaperon SicA from *Salmonella* and the IncA effector from *Chlamydia trachomatis*) as found online in the STRING database are pictured. The first example exhibits two instances of effectors which could be detected as functionally coupled with the chaperon and other TTSS related proteins (either components or specific regulatory elements). In the example of IncA, only few links can be found which connect it to proteins which are either of unclear function or member of other pathways and therefore unlikely TTSS related. The observation of these two examples motivates to assess the amount of effector proteins which could be found as direct neighbors to other TTSS related genes which has been further analyzed. In some cases, functional coupling between TTSS components/chaperons and the effectors from the training could be found, but most effectors do not functionally co-evolve with the TTSS. The found partnerships between TTSS components and effectors are listed in the supplementary Table 6.16. This finding is in congruence with the module based analysis of virulence related groups of proteins, which resulted in the same picture. Although no



```
HopN1  MYIQQS-GAQSGVAAKTQHDKPSS-LS---
HopV1  --MRFD-AAR-GQKPKAPMDAPSS-LRLRA
IncA   -MTTPT-LIVTPPSPPAPSYSANR-VPQ--
Tir    --MPIG-NLGHNPNVNNSIPPAPP-LPSQ-
HrpK   --MRIS-SSP-FVIVNQPTPGELA-LAVES
YopE   --MKISSFISTSLPLPTSVSGSSS-VGE--
AvrB   --MGCVSSKSTTVLSPQTSFNEAS-RTS--
SpaN   --MGDVSAVSSSGNILLPQQDEVGGLS---
HopAN1 --MLVRGGMRTVQSIGIPSAEMAAGLD---
LcrE   --MAASGGTGGLGGTQGVNLAAVEAAA---
IncE   --MECVKQLCRNHLCLDSLTGPVRSVL---
AvrA   -MWNVSKSSNNLGAYKLPLEAQTP--PE--
YopK   --MFIKDTYNMRALCTALEQSAPDTII---
HopB1  --MRPVGGPAPGYYPPTYEAERPTAQA---
YopT   --MNSIHGHYHIQLSNYSAGENLQSAT---
YopH   --MNLSLSDLHRQVSRLVQQESGDCTG---
HopP1  --MTMGVSPIRNSNSLPIDFSSLSAKS---
IncG   -MICCDKVLSSVQSMPVIDKCSVTKC----
EspG   -MILVAKLFITNQIGESLMINGLN-ND---
SipA   --MVTSVRTQPPVIMPGMQTEIKTQAT---
```

Figure 3.1: Example alignment of effector N-termini. The colors refer to standard ClustalX coloring schema.

conserved functional relationship could be deduced as general rule, the effectors might be joined to TTSS components by common regulation or common acquisition by an horizontal gene transfer. To check this, Dr. Thomas Rattei performed an analysis of the genomic neighborhoods of effectors and TTSS components alike. This has been done by firstly checking the genomic neighborhood of known effectors using an statistical enrichment analysis of all co-localized proteins in different distances. The highest significance of this enrichment has been observed within the range of 30 proteins up- and downstream. Within these neighbors, 7 structural TTSS proteins show individual enrichment of statistical significance as listed in the supplementary Table 6.17. However, particularly in genomes encoding the TTSS on the chromosome as e.g. *Chlamydiae*, the majority of effectors cannot be found in genomic proximity to components of the TTSS, as can be seen in the supplementary Table 6.18.

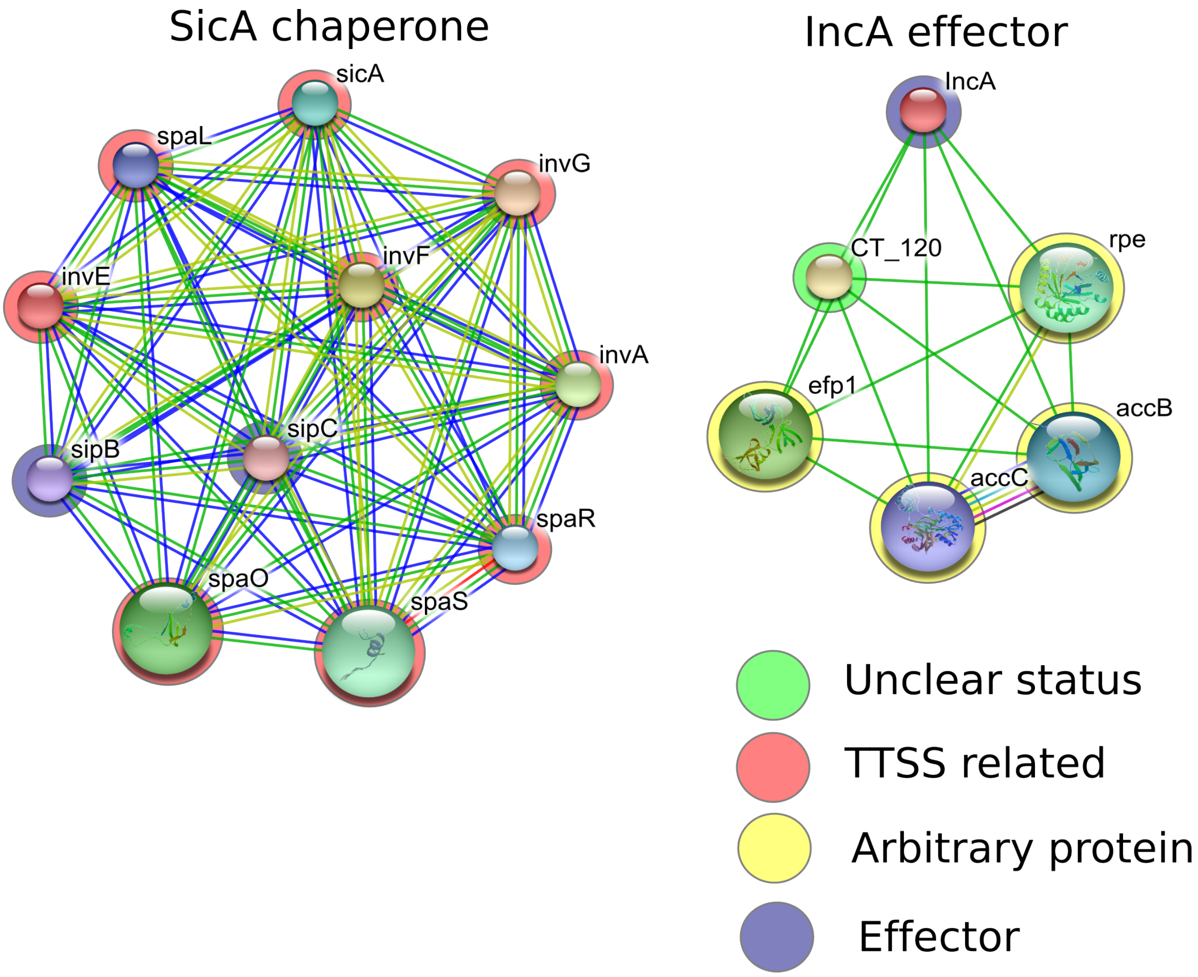


Figure 3.2: Example of two functional networks of TTSS related proteins as found in STRING [159]. The chaperon SicA is part of a dense module which is TTSS related, contrarily, the effector IncA cannot be related to TTSS mediated transport by its' functional neighbors. The proteins have been classified by curated information from databases and literature.

A prediction of secondary structure elements on the effector sequences revealed differences between arbitrary proteins and the effectors: I counted the structural features (coil, α-helix, β-sheet) at each residue within the first 25 amino acids. In the known TTSS effectors, 51% coil, 39% α-helix and 10% β-sheet have been predicted. In randomly selected proteins (not known to be secreted via a TTSS) 39% coil, 45% α-helix and 16% β-sheet have been found, which indicates that coiled regions are enriched in the N-termini of TTSS effectors. The secondary structure prediction has been applied without the use of extrinsic information due to a lack of sufficient homologs in the effector N-termini and might therefore be less accurate as possible. In addition, the prediction might fail at the very N-terminal end since not sufficient information on neighbored residues could be used by the method. However, since these aspects are the same for both sets (effector and non-effectors alike), this should not flaw the observed differences. In conclusion, the effector N-termini exhibit less secondary structure. This lack on secondary structure elements is a first hint on the signals nature: an alien composition in these sequence regions could lead to a smaller fraction of regular structure elements which are existent and predictable. Such an alien nature of the N-termini fits well with data from *P. syringae*, a well-studied plant pathogen, for which an unusual amino acid composition in the N-termini of effectors has been reported [295], [296], [63]. Indeed, we found the amino-acid composition of the effector N-termini unusual in comparison with arbitrary sequences: Stefan Brandmaier and Frederick Kleine tested a possible use of frequency counts of each amino-acid during a practical course and found them to be predictive for the identification of effector proteins. I statistically tested the differences of the amino-acid frequencies in the effector N-termini against arbitrary sequences by an Mann-Whitney test. This test revealed significant enrichments and depletions of certain amino acids in sequences from animal pathogens and plant symbionts compared to arbitrary sequences. This effect has been found when comparing the first 25 residues of effector sequences against complete non-effector and effector sequences, but also when comparing the N-terminal ends of effectors and non-effectors. I also compared complete effector sequences with arbitrary sequences and found the same tendency but to less extend. Since this effect is particularly strong in the N-terminal end, this composition bias could reflect an exploitable signal of TTSS mediated transport in congruence with the N-terminal position proposed by the fusion experiments. The enrichments in the first 25 against arbitrary sequences are plotted in Figure 3.3. The most significant enrichment in the N-termini of effectors of animal pathogens and plant symbionts is that of Serine. Threonine and Proline are significantly enriched in the effectors of animal pathogens, and leucine is depleted in both animal and plant effector proteins. Notably, the enrichment of proline could explain the enrichment of coiled regions in the N-termini as this amino acid is known to be less frequent in α-helices and β-sheets. Interestingly, these experiments revealed both commonalities and differences between the N-terminus of effector proteins from plant and animal pathogens, respectively.

Discussion Since the analysis of conserved functional coupling as well as of the genomic proximity did not recover relationships for most of the effectors, these approaches cannot be applied for an exhaustive search of effector proteins. However, in some cases, co-evolution of certain effectors with each other and the co-localization of several effectors

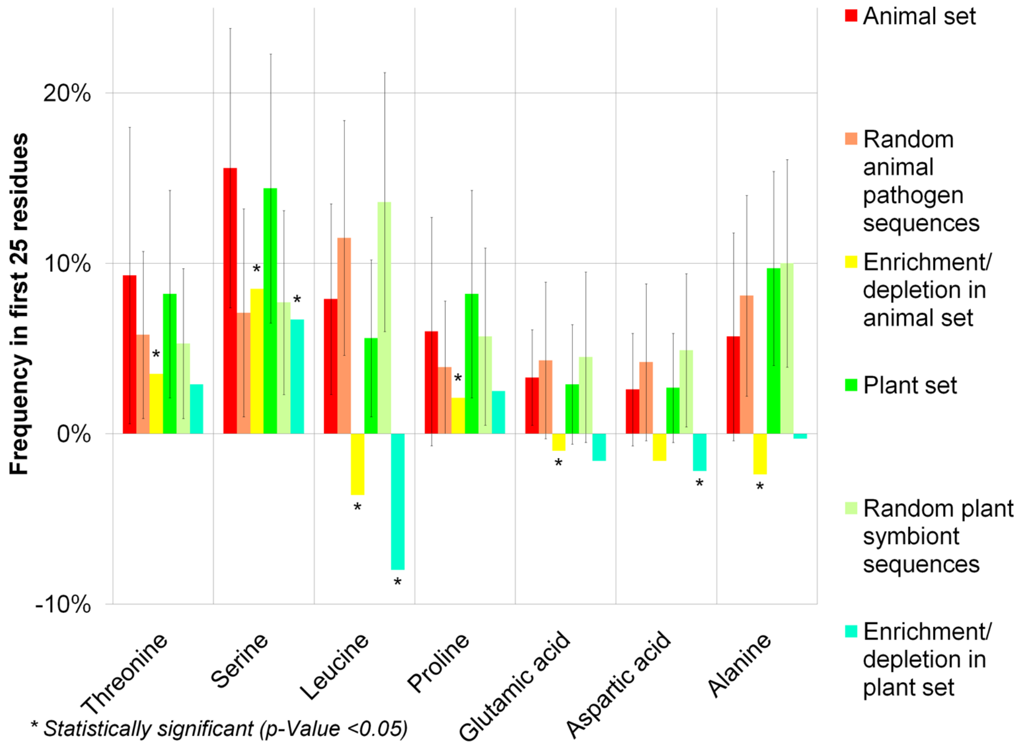


Figure 3.3: Amino acids that are significantly enriched or depleted in the first 25 residues of effectors from the animal pathogen effector set and from the plant symbiont effector set (p-Value<0.05 in the one sided Mann-Whitney test in at least one of the sets). Frequencies are given as percentage of amino acids within the 25 first residues. Error bars represent one standard deviation in plus and one standard deviation in minus directions.

with TTSS components and chaperons can be observed. So, these methods are valuable for situations if such effectors or chaperons are already known or if the TTSS is encoded on a plasmid or on a genomic pathogenity island. The initial examination of the N-termini by multiple sequence alignment revealed no conserved positions and a putative signal is therefore not detectable as conserved domain. Where the latter finding could indicate evidence for an mRNA based signal, additional analyses detected hints to a peptide encoded signal: the differences in secondary structure (i.e. the absence of merely regular structures as α helices or β sheets indicate a different composition of the N-termini. This can be shown more detailed by the enrichment analyses: the N-termini show an alien distribution in comparison with arbitrary sequences. This initial findings encourage further modeling of the signal computationally.

3.2.2 Modeling of the N-terminal TTSS signal peptide using a machine learning approach

Motivation The main difference between effector and non-effector proteins has been found in their N-terminal amino-acid composition. The commonalities of the N-termini which lead to the compositional differences are not encoded as detectable conserved residues. Taken both observations together, the signal could be comprised of a combination of certain amino-acids without clear order. Such a signal is not necessarily restricted to individual amino-acids but might comprise combinations of groups of amino-acids with the same physical-chemical properties (i.e. groups which represent likely exchanged amino-acids). Furthermore it may contain short combinations of the length two or three. Taken together, the space of all combinations referring to this properties is huge and the identification of the combination of these possible features which is optimal to describe the signal cannot be easily obtained. Especially, there is not sufficient *a priori* knowledge which would help to identify the relevant sub-sets of these properties. Machine learning techniques allow for deduction of non-trivial relationships from training data without the need of an exact *a priori* description of the properties describing a certain class of instances. A promising approach is therefore the application of binary classifiers trained on known effectors against arbitrary sequences, which could be used to discriminate unseen instances (in this case, protein N-termini), being probably TTSS secreted or not secreted, after deducing the most discriminating properties between the two classes from a training set. The performance of any classifier can be assessed by rigorous cross-validation and the predictive power on novel data can be estimated in comparison to a random classification.

Material and methods The aforementioned properties can be represented as 'reduced' alphabets to which the initial sequence of amino-acids is mapped. Two alphabets have been employed: firstly, the amino-acids have been mapped to an alphabet of amino acid properties, and secondly, to a hydrophobic/hydrophilic alphabet. Each amino acid is only added to one of the property classes, although some would fit to several classes. In this case, the amino acid has been added to the more specific (smaller) class. The feature mapping is listed in Table 3.1. From these representations, the frequencies of di- and tri-peptides from each of the alphabets have been computed. From these features, I discarded all these which did not occur at least two times in either the positive or the negative data set, since these features would lead to the adaptation of

Property	**Amino-Acids**
Hydrophobic, 1st alphabet	A, G, I, L, M, V
Hydrophilic; 1st alphabet	P, H, U
Aromatic	F, W, Y
Polar	N, Q, S, T
Acidic	D, E
Alkaline	K, L, R
Ionisable	C, Y
Hydrophilic; 2nd alphabet	S, F, T, N, K, Y, E, Q, C, W, P, H, D, R, U
Hydrophobic; 2nd alphabet	V, M, L, A, I,G

Table 3.1: The mapping of amino acids to the two reduced alphabets (amino acid property alphabet and hydrophilic/hydrophobic alphabet) maps each amino acid to exactly one letter of the respective alphabet.

the classifiers to individual sequences (over-fitting). This procedure typically reveals 70 features, depending on the negative set employed. The frequencies of these features range typically between 2 and 5 and have been taken as input to the classification algorithms without further discretization. A list of all features is given in the supplementary Table 6.19. To detect the most influential features, I applied two feature selection strategies, a greedy hill-climbing search (the BestFirst algorithm) (parameters: look-up-cache size = 1, 5 iterations) in combination with Correlated Feature Selection [297] (parameters: locally predictive = true, missing values = false) as provided by WEKA (version 3.5.6) [123]. Implementations of several classification algorithms from the WEKA machine learning package have been tested five times using different negative sets (see used data sets) by a 10-fold cross-validation procedure as provided by WEKA. For cross-validation, the positive and negative sequence sets have been partitioned into 10 subsamples. In each of the 10 passes, a single subsample was retained as validation data for testing the model which has been trained using the remaining 9 subsamples. I systematically aligned each N-terminus of the training set with each other using the Smith-Waterman algorithm with a BLOSUM62 substitution matrix. If two sequences showed S_{ratio} (see above)$>$0.1 over the whole sequence or more than 0.3 in the area of the signal, one of them was discarded from the training set. This has been done to avoid learning protein-families instead of the signal. Sensitivity has been computed as $\frac{TP}{(TP+FN)}$, Selectivity as $\frac{TN}{TN+FP}$, with TP = amount true positive predictions, FN = amount false negative predictions, TN = amount true negative predictions, FP = amount false positive predictions. Receiver Operating Statistics to determine the AUC value had been created using the WEKA-toolbox. Precision and Recall are computed separately for both classes, where the

Name	Description
IB1	First Nearest Neighbor, Euklidian Distance, no weighting [298]
Logistic	Logistic regression [299]
Voted Perceptron	Extension of the Perceptron Algorithm (a simple Feed Forward Neural Network) [300]
SVM	Support Vector Machine with Polynomial Kernel (n=1) [301]
Naïve Bayes	Complement Naïve Bayes classifier, (An adaption for skewed training data) [302]
Naïve Bayes	Naïve Bayes classifier with Multinomial Model [303]
Naïve Bayes	Naïve Bayes classifier [304]

Table 3.2: Classification algorithms used in this study.

AUC describes the overall performance of the classifier. The classification algorithms employed are listed in Table 3.2.

To assess the optimal position and length of the signal, the pipeline (training and testing) has been applied to different starting positions and for different signal lengths starting from the N-terminus. The scan for the optimal position has been done in steps of five residues and a window of fifteen residues has been tested as signal. For each selection of length and position, the complete feature creation, training and testing procedure has been repeated repeated.

Results The rationale in this analysis is that a successful training of a binary classifier should result in a model of the signal which is the better, the more the classifier is able to distinguish between real effectors (positive testing instances) and arbitrary proteins. The performance is measured as the "Area Under the Curve" (AUC) value of the Receiver Operating Statistic Curve (ROC). This represents the performance of a classifier describing the trade-off between sensitivity and selectivity by varying over the classifier's parameter space. The AUC summarizes this overall performance: an ideal classifier yields an AUC of 1.0, whereas a completely random prediction results in a value of 0.5. Values above 0.5 indicate a prediction above random. The 'animal pathogen' and the 'plant symbiont' set may perform differently since the biological signal might differ between both sets. So, each set has been tested separately, as well as combined. The signal should be captured by several different classification algorithms above random. This would indicate, that the result is not by accident due to a certain classifier/data combination but inherent in the data. A systematic comparison of different classification algorithms on the TTSS effector sets from animal pathogens and plant symbionts, respectively, resulted in a performance far above random for all classifiers tested, with an maximal AUC of 0.85 for the animal pathogen set and an AUC of 0.86 for the plant symbiont set, achieved by the complement naive Bayesian algorithm. Both sets combined together achieved their best AUC (0.86) with the Naive Bayesian classifier 3.3. Training the classifier solely on the predicted secondary structure alphabet of the combined set performed well with an AUC value of 0.8. However, adding this alphabet to the sequence derived features did neither improve nor reduce the performance significantly: the test revealed an AUC of 0.87 with and 0.86 without the secondary structure features. The trained classifier capture therefore the underlying signal, but do not describe it detailed since they merely act as black box. To get a better understanding of the signal, the most discriminating features have been derived by a feature selection procedure (see methods). This feature-selection resulted in a reduced list of properties, comprise not only the Serine, proline and Threonine frequencies as already indicated by the amino acid composition analysis, but also depletion of acidic and single alkaline residues and patterns such as the enrichment of two consecutive alkaline residues or the pattern “polar-hydrophobic-polar”. This main components of the recognition signal are listed in Table 3.4. As prove of the concept of the N-terminal signal peptide, C-termini should have no predictive power. The performance for several classifiers has been evaluated using exactly the same feature selection, training and test procedure as used for the N-termini. 5 runs with different negative sets have been performed. The resulting AUC values tend to a random prediction (AUC near 0.5), indicting an absence of any commonalities in the C-termini, i.e. the absence of a signal. However, it is not clear which part of the effector sequences are the most discriminative ones. Additional signals could be possible, as well as a longer N-terminal stretch as initially used. To resolve this question, two tests have been employed by scanning different possible lengths of the signal starting from the N-terminus (to detect the optimal length), and by a sliding window over the whole sequence lengths (to detect the optimal position). The results for these two experiments are shown in Figure 3.4. High AUC values are reported over a wide range of N-terminal peptide lengths, with only a slight maximum peak at length 30 in the animal pathogen and length 50 in the plant symbiont set, the actual length of the signal is difficult to determine. This effect can be explained by the fact, that the very N-terminal end is always part of the prediction and (if it comprises the signal) positively influences the prediction. The position scan revealed that the most discriminating positions are indeed at the N-terminus followed by a region with less predictive power. The

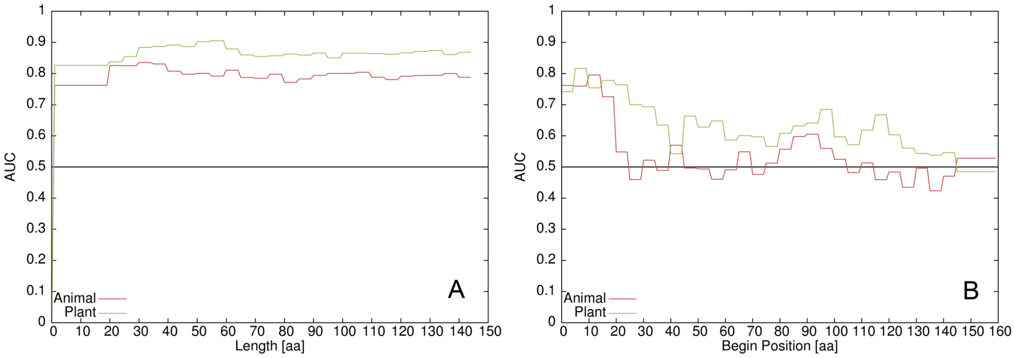


Figure 3.4: Exploration of optimal length of the signal (A) and begin position of a 15 amino acid long window (B). The AUC value for each length and begin position is plotted for the animal pathogen set (red) and the plant symbiont set (green).

best performance was achieved with the residues 0–30 in the plant symbiont and 0–50 in the animal pathogen set of effector proteins. Notably, also the selection 0–15 in both sets gives a good discriminative power. Some other positions (e.g., residues 90–105 and 120–135 in the plant symbiont set) also show (an indeed weaker) predictive power which could hint to an additional signal or at least regularity in these regions. The majority of positions, however, have no predictive power due to AUC values between 0.4–0.6.

Algorithm	**Sens. (a)**	**sd**	**Sel. (a)**	**sd**	**AUC (a)**	**sd**	**AUC (p)**	**sd**	**AUC (b)**	**sd**
Naive Bayes complement	0.77	0.02	0.79	0.04	0.78	0.02	0.78	0.03	0.79	0.02
1 nearest neighbour	0.54	0.09	0.81	0.04	0.68	0.07	0.69	0.04	0.68	0.04
Logistic regression	0.57	0.07	0.75	0.07	0.72	0.08	0.73	0.03	0.74	0.02
Naive Bayes	0.71	0.03	0.85	0.04	0.85	0.03	0.84	0.01	0.86	0.02
Naive Bayes multinomial	0.76	0.03	0.81	0.04	0.85	0.02	0.85	0.02	0.86	0.04
Support vector machine	0.57	0.05	0.86	0.04	0.71	0.04	0.74	0.03	0.77	0.03
Voted perceptron	0.24	0.04	0.97	0.02	0.78	0.01	0.79	0.04	0.83	0.02

Table 3.3: Performance of the different classifiers. Values given: 'AUC (a)' the AUC on the animal pathogen set, 'Sens.': Sensitivity and 'Sel.': Selectivity on this set, 'AUC (p)' performance on the plant symbiont set, and 'AUC (b)' the overall performance.

Pattern	**Enriched/Depleted**
polar-hydrophobic-polar	enriched
alkaline-alkaline	depleted
Threonine	enriched
Serine	enriched
Proline	enriched
polar	enriched
alkaline	depleted
acidic	depleted
hydrophobic-alkaline	depleted
polar-polar	enriched

Table 3.4: The most discriminating features between effectors/non-effectors found by the feature selection procedure.

Algorithm	**AUC**	**Standard Deviation**
Perceptron	0,54	0,04
1 Nearest Neighbor	0,48	0,02
Logistic Regression	0,52	0,02
Support Vector machine	0,49	0,02
Naïve Bayes, Multinomial	0,55	0,03
Naïve Bayes Complement	0,53	0,03
Naïve Bayes	0,52	0,04

Table 3.5: Performance of C-termini.

Discussion The machine learning approach trained by features delineated by this compositional bias successfully captures a non-trivial signal that can be used to predict effector proteins far better than random. This indicates that a signal is highly probably encoded in the N-termini and this finding favors the peptide signal hypothesis over the mRNA hypothesis. The signal is independently captured by several different algorithms and is therefore a signal in the data and no artifact. The worst performance (which is still significantly above random) has been obtained by the 1-nearest neighbor approach with an AUC in the complete set of 0.68. This classifier looks for the next data-point given a query in the Euclidean space spanned by the training instances. The poorer performance indicates, that the signal is complex and can be described better by classifier that weight properties (as the Bayesian classifiers) as by the distance to the next similar training sequence. The signal could be described by the most discriminating features. Lloyd et al. found Serine-rich N-termini to be secreted [61], a finding with is supported by the analysis herein as Serine has been found among the most discriminative features. The machine learning approach as well as the trained classifier can be used for further investigations on the signal as on the generality and size of the signal.

3.2.3 The signal is robust against point mutations and even tolerates frame shifts

The secretion signal should be tolerant against single mutations as long as they do not affect residues which are obligate for recognition. The proposed model comprises traits (as an enrichment of Serine) which are strong contributers to the signal. An *in silico* mutation analysis should reveal a behavior congruent with this findings: the signal should break down fast, if such influential residues are changed but should be more resistant against random mutations. Schneewind and coworkers [305] showed that frame shift mutations in the mRNA altering the N-terminal peptide sequence did not abolish transport of three TTSS effector proteins of Yersinia species. This seems to contradict the N-terminal signal peptide hypothesis but could be explained, if the frame shifts lead to altered amino acids in the N-terminus, which nevertheless retained the characteristic features of the TTSS signal. This has been investigated by looking for examples in which a non-sense mutation retained or created a new signal.

Material and methods To detect a general robustness of the model, residues have been accumulatively changed by random in each effector previously predicted as (true) positive. In a second experiment, these changes have been guided by the most discriminative features and the amount of still positively predicted instances after each round of mutation has been assessed. For example, the amount of Serine and Threonine has been depleted by exchanging them in favor of arbitrary residues. The paper of Ramamurthi et al. [40] provides data of three Yersinia effector proteins with three frame shift mutants for each. To test these sequences, the classifier has been retrained using the first 15 amino acids instead of the first 25, since only the first 15 residues of the mutants are given in the paper. Simulation of frame shifts has been done by shifting the DNA by one (+1) and two (+2) positions. In order to get a sufficient amount of sequences with sufficient length, appearing stop codons have been replaced by methionine. Instances have been selected that exhibit a positive prediction with restrictive parameters (probability for class “secreted” >0.95 as reported by the Naive Bayesian Classifier). As negative control, randomly selected sequences from the same organisms which are covered by the positive set have been chosen and positively predicted instances have been filtered out (probability not secreted >0.95 as reported by the Naive Bayesian Classifier). Signals conserved after frame shift were detected with the same settings as in the selection procedure.

Results In the first experiment the signal turned out to be robust when changing arbitrary residues: after one point mutation 97% after five 75% and after ten 54% of the effector proteins still have a detectable signal. In the second experiment (that introduces mutation on the proposed signals’ core), the signal rapidly breaks down: after one mutation 93% of the effectors, but only 27% after five and 2% after ten mutations carry a detectable signal. The resultes of both experiments are plotted in Figure 3.5. Nine example frame shifts are given in this study which did not abolish secretion. One Yersinia protein (YopQ) could not be predicted as effector and thus represents a false negative prediction. From the remaining six frame shifts in two proteins (YopE and YopN), only the −2 frame shift of the YopN N-terminus did not lead to a loss of the TTSS signal. The same behavior has been shown for the Salmonella effector InvJ which tolerates +1 and −1 frame shifts [306]. In the case of the +1 frame shift the signal is still revealed by EffectiveT3, whereas no signal can be detected for the −1 frame shift. In order to assess the sensitivity of the TTSS signal towards frame shift mutations in a more systematic manner, I artificially introduced all possible frame shift mutations into the 74 known and positively predicted effectors. As control, the same procedure has been applied to a set of 199 randomly selected and negatively predicted control

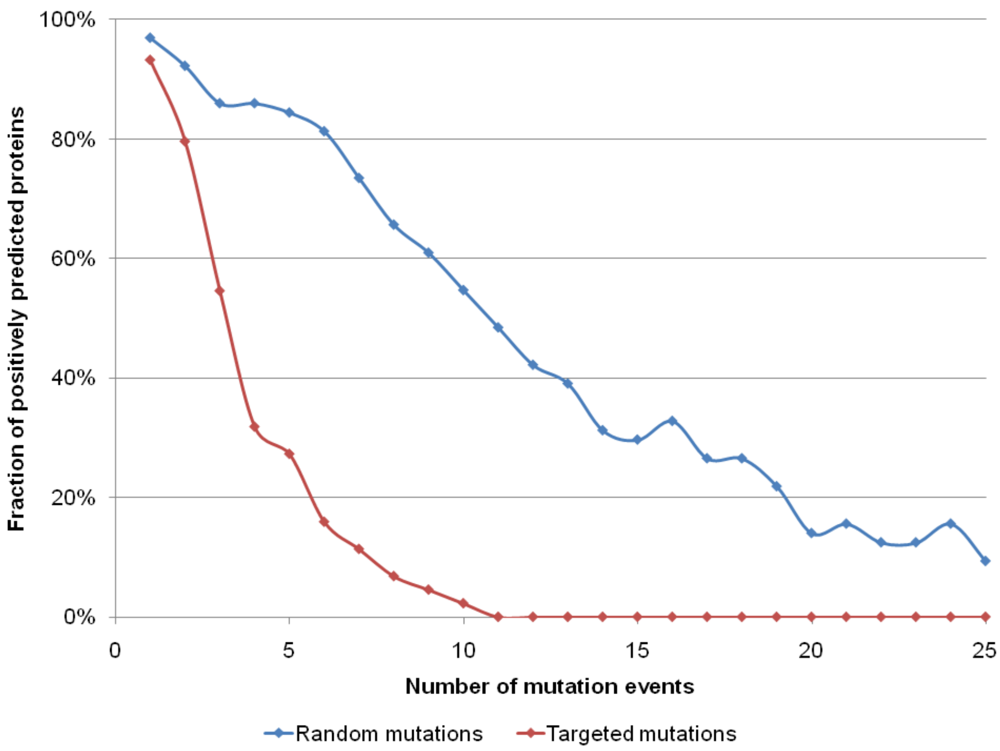


Figure 3.5: Robustness of the TTSS secretion signal against point mutations. The diagram depicts the percentage of positively predicted TTSS signals after accumulation of point mutations. The non-targeted mutation strategy exchanged residues accumulatively by random. The targeted mutation strategy favored to exchange these features, which have the strongest influence on the signal due to the trained model. For both experiments all positively predicted proteins from the animal pathogen and plant symbiont training sets have been used.

sequences. In 15 cases (10%) of the effector mutants, the signal was preserved as it can be seen in Table 3.6, in contrast to 31% of the control sequences (data not shown).

Discussion The initial two experiments on mutations show, that the signal is robust against single and multiple point mutations as long as the significant enrichments and depletions of certain amino acids are not altered. This experiment gives a direct hypothesis testable in the laboratory by guiding targeted mutations which show a possible loss of TTSS mediated transport of the tested sequences.

The introduction of frame shift mutations should abolish a peptide born signal. Contrarily, some mutants are still predicted as secreted. In agreement with the mRNA signal hypothesis [307], [39], three effector sequences are resistant to both kinds of shifts, the +1 and +2 mutations. This finding explains the observed mutation tolerance observed in the latter experiments without the need to abolish the peptide- based signal hypothesis in some cases. The experiment of the frame-shift mutations on the control set (the non-effector mutants) revealed an unexpectedly high rate of predicted signals. These sequences have been found to exhibit patterns of amino acid enrichments and depletions which are very similar to the characteristics of TTSS effectors 3.6. The observed behavior indicates a possible way to acquire an (perhaps week) initial signal from intergenic space during evolution with only few mutations.

Name	**Mutation**
Q9Z8P7 CHLPN	+1,+2
Q9Z8P6 CHLPN	+1
INCA1 CHLTR	+2
AAO54130	+2
AF458396	+1,+2
AAO54892	+1,+2
AJ277494	+2

Table 3.6: Effector sequences which show toleration of frame-shift mutations. The mutations were introduced by either shifting the DNA sequences by one or two bases to the left, stop codons where replaced by Methionine.

3.2.4 The TTSS signal peptide is taxonomically universal

Although the Type III secretion system is well conserved and exists taxonomically seen widespread in very different taxa either inherited or introduced to a genome by horizontal gene transfer [308]. The organisms harboring such a system may differ completely in their host specificity: *Pseudomonas*, for example, is a plant pathogen where the *Chlamydiae* enter eukaryotic cells as their hosts. Although the different modes of host interaction are merely implemented by the effector proteins and their controlled activation, it cannot be excluded that the substrate recognition could substantially differ between different taxa. A certain generality of the system has been uncovered by the application of heterologous screens which transported effector proteins of another species. In addition, the initial analysis of the amino-acid frequencies indicate universality of the signal. The cross-validation procedure itself could not prove this universality, since in each round of the validation, taxon specific features might be introduced. In the following analysis, I show this generality explicitly: by systematically excluding genomes from training, they cannot contribute to the feature selection process. When then used as test, congruence of the signal in the excluded with the signal in the training sequences is indicated by high AUC values.

Material and methods In order to test the universality of the signal, each taxon has been excluded (*Yersinia*, *Salmonella*, *Escherichia*, *Chlamydia*, *Pseudomonas*) from the training and feature-selection procedure. The classifiers performance with this taxon as separate test set has been assessed by the AUC value. For both sets, negative sets twice as large are randomly created from these organisms, which are also in the respective positive set. The values for the AUC have been computed using the WEKA-toolbox. The same data has been computed for the same amount of randomly chosen sequences of the excluded taxon as taxon specific null-model.

Results The results of the analysis for the Bayesian classifier are shown in Figure **??**.

High AUC-values between 0.83 and 0.89 were observed for all tested combinations using the Bayesian classificator. Notably, it was possible to predict effectors from the animal pathogen set when trained by the plant symbiont set and vice versa, yielding an AUC of 0.86 and 0.83 respectively and similar results have been found using other classification algorithms.

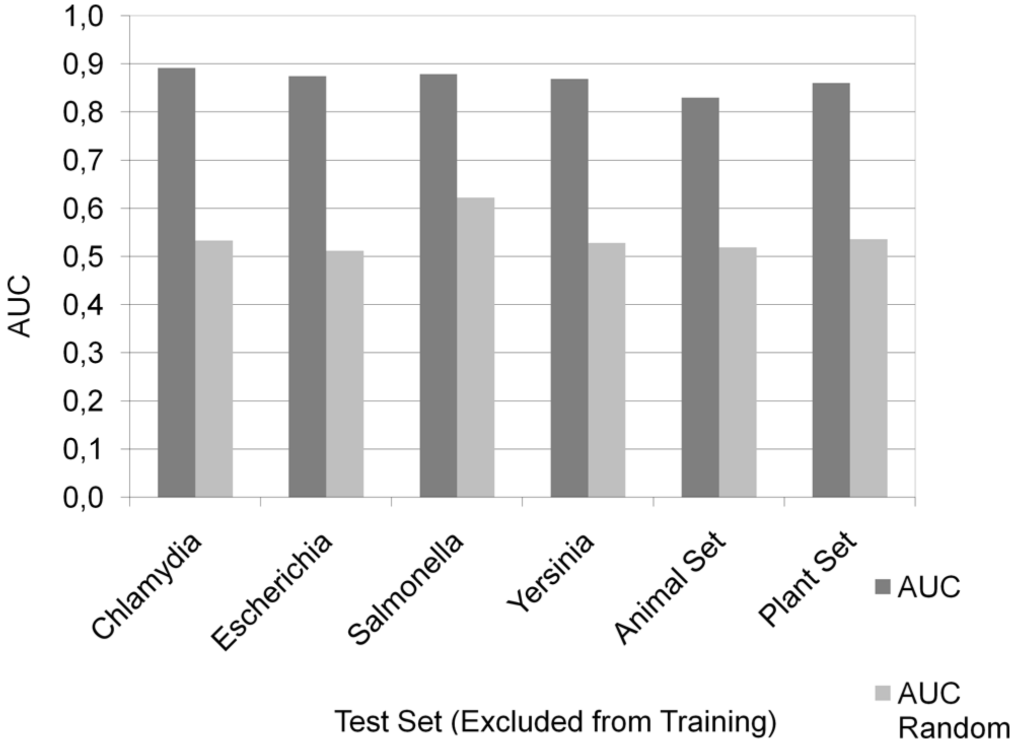


Figure 3.6: Test on the generality of the signal. The y-axis denotes the achieved AUC value of EffectiveT3 when trained without the positive and negative samples from the taxonomic group denoted at the bottom of the x-axis and tested against this set. The performance on a randomly chosen set of positives and negatives having the same taxonomic composition is given for comparison.

Discussion Since a high AUC has been achieved in any combination, the captured signal is not organism specific but must be taxonomically universal. In consequence, the basic recognition process must also be conserved or only slightly different. It can be observed, that the AUC values is smallest when excluding the animal set from training indicating that its' composition of diverse effectors from different species contributes positively to the general performance. Notably, this test encourages the use of Effective in unseen genomes of other gram(-) bacteria.

3.2.5 Evolutionary history of the TTSS signal peptide

In contrast to the secretion apparatus, the effectors are not well conserved and the repertoire in different bacteria is different. This implies the question how bacteria can invent effector proteins *de novo* in order to cope with changing host specificity and in the arms race with the immune response of their respective hosts. Novel effectors can be acquired by horizontal transfer as shown in *Pseudomonas* [309]. However, at some time-point in evolution, novel effector sequences must arise, e.g. by turning an arbitrary protein into a secretable effector. Stavrinides et al. [310] proposed a mechanism called terminal re-assortment which describes the acquisition of signal peptides by genome re-arrangements leading to chimeric sequences, which could be detected for 32% of the investigated effector families in contrast to 7% in arbitrary sequences. Interestingly, intergenic regions show – if translated – a similar amino acid composition as the signal predicted by EffectiveT3. This indicates a simple way of signal acquisition through a $5'$ shift of the start codon followed by few subsequence point mutations. An additional scenario is the turn of an arbitrary protein into an effector by random point mutations which lead to an initial signal. All scenarios are pictured in Figure 3.7. Probably, both processes might contribute to the evolutionary de-novo invention of effectors. To assess this question, I investigated together with Sebastian Behrens how many changes in the N-termini of known effectors occurred in comparison to their orthologs in an organism without TTSS.

Material and methods Orthologous groups have been obtained from the eggNOG database [126] for each effector protein. Proteins from organisms other than Gammaproteobacteria have been filtered out. The remaining proteins where labeled as “effector” if in training set, “putative effector” if from an organism with TTSS or “non-effector” if from an organism without TTSS. In 10 cases, orthologs of effectors in non-TTSS species could be found. In order to investigate the N-terminal ends with the help of alignments, the C-terminal functional parts of the sequences must be cut since they would dominate the alignment. To find the most probable start of the functional part of the proteins, I searched for the first functional domain as detected by Pfam [311] (as contained in InterPro Release 17.0 [312]), cut at the start of the domain and created multiple alignments of the remaining N-terminal fragments. Then, regular N-terminal extensions of effector or putative effector proteins compared with non-effectors by manual inspection in the case of the multiple alignments. Also pair-wise alignments of effector/non-effector sequences from the same orthologous group have been created. I counted elongations

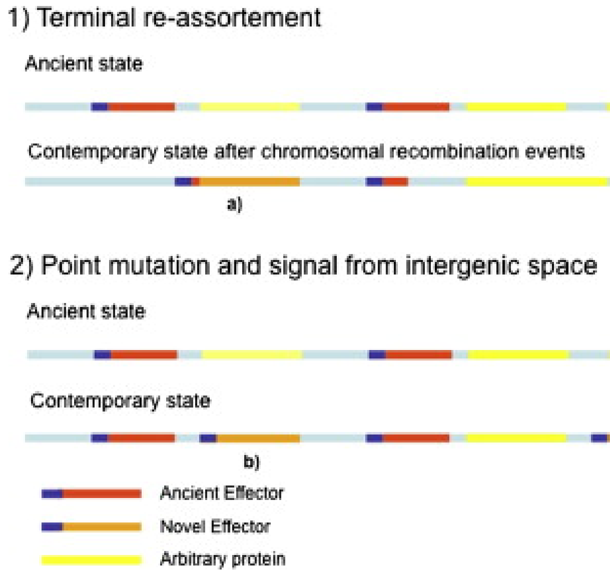

1) Terminal re-assortement
Ancient state
Contemporary state after chromosomal recombination events
a)
a)
2) Point mutation and signal from intergenic space
Ancient state
Contemporary state
b)
c)
Ancient Effector
Novel Effector
Arbitrary protein

Figure 3.7: Schematic illustration of possible evolutionary mechanisms leading to new effectors. (1) Terminal re-assortment. This process is based on the shuffling of genomes creating novel signal-protein hybrids. If the hybrid furthers the fitness of the organism, it is evolutionary retained as novel effector (a). (2) De-novo generation by mutations. Point mutations could lead to a weak initial signal which is sufficient for secretion. If the novel effector increases the species' fitness, the signal will be strengthened by selection (b). If the point mutation leads to an N-terminal elongation by mutating the start codon, the protein is elongated (c) and the additional N-terminus can adapt towards a full secretion signal without changes to the proteins original functional parts.

(alignment start of the effector greater than of the non-effector) and truncations within one group. Examples are given in the supplement 6.4. If the difference between the alignment starts was smaller than 15 residues, we counted the alignment as having the same length. The same procedure has been repeated without aligning the sequences by just comparing the lengths before the start of the functional domain. Multiple alignments were built using ClustalW (Version 2.0.5) [290], Muscle (Version 3.7) [291], with standard parameters, pair wise alignments were calculated with the Smith Waterman algorithm as implemented in the Jaligner package using the BLOSUM62 substitution matrix.

Results The results of the principal evolutionary events (truncation, elongation, no change in size) are summarized in Table 3.7. A manual inspection of the multiple alignments did not reveal a clear pattern which would support regular fusion events between a functional protein and a 'signal domain' This result is further supported by the pair-wise analysis: Elongations of the effector sequences compared to non-effectors

Effector	+	%	-	%	=	%
AAO57557	3	75,00%	1	25,00%	0	0,00%
AAO54387	3	10,71%	4	14,29%	21	75,00%
YPKA YERPS	18	50,00%	18	50,00%	0	0,00%
YOPT YERPS	3	11,11%	22	81,48%	2	7,41%
Q9Z7W9 CHLPN	1	12,50%	7	87,50%	0	0,00%
Q9RPH0 SALTY	0	0,00%	2	100,00%	0	0,00%
YOPM YERPE	7	58,33%	5	41,67%	0	0,00%
YPKA YERPE	18	50,00%	18	50,00%	0	0,00%
YOPT YERPE	3	11,11%	22	81,48%	2	7,41%
AAO58488	0	0,00%	6	100,00%	0	0,00%
TARP1 CHLTR	0	0,00%	1	100,00%	0	0,00%
Q663L9 YERPS	1	16,67%	4	66,67%	1	16,67%

Table 3.7: Evolutionary events. Truncations, elongations and conservations of the N-terminal length until the first functional domain are listed according to the effector protein (first column) compared to orthologs from non-TTSS bearing organisms. '+': elongation of the effector, '-': truncation, '=' equal length.

are less frequent (30%) than truncations (57%), whereas a similar length of effector and non-effector occurs in 13% of all pairs. All three events can be detected within the same orthologous group. HopAK1, a *Pseudomonas syringae* effector, is the only example which is more often elongated (three cases) than truncated (one case). A similar picture can be seen when only the length of the N-terminal regions before the first common functional domain of effector and non-effector orthologs were compared: N-terminal regions with equal lengths can be found in 4%, shorter lengths for the effector in 39% and longer lengths for the effector in 57% of cases

Discussion In this analysis, a regular elongation of effectors in comparison to non-effectors cannot be stated, however, a certain amount of those events exist. The observed data favors the mutation accumulation theory since most examples are not elongated but more often of equal size or smaller compared to their on TTSS relatives. The examples of elongations refer to either the process of terminal re-assortment as proposed by Stavrinides et al. [310], or to the acquisition of intergenic space. These two cannot be distinguished here. Overall, a mixed picture can be seen and the absence of any of the mechanisms proposed cannot be stated.

3.2.6 A substantial fraction of proteomes is predicted as secreted

The amount of sequences predicted as secreted could give insights into the amount of a proteins provided by an organism as effector arsenal and can be compared between different taxa.

Material and methods Complete genome and proteome data of prokaryotic genomes has been downloaded from the KEGG database [284] (release 2009/01/19). Components of the TTSS have been identified by their association to the KEGG Orthologous Groups (KO) belonging to the TTSS reference pathway KO03070 (K03219..K03230). Genomes in which at least 9 of these 12 KO are present have been considered as genomes with TTSS. Genomes in which less than 6 of these 12 KO are present have been considered as genomes without TTSS. All genomes in which between 6 and 8 of these 12 KO are present have been excluded from this analysis to avoid uncertainty. Additionally, all bacterial genomes have been excluded from this analysis for which no information on cell wall type (Gram-positive vs. Gram-negative) was available at the NCBI Entrez Genome Project Organism Info database [313]. For the remaining 739 proteomes, EffectiveT3 predictions have been calculated using a selective parameter setting (probability for class "secreted" >0.99 using the Naive Bayesian Classifier). To estimate the enrichment of TTSS effector-like sequences in the N-termini of the proteomes, a genome-wide Z-Score is calculated for every proteome: $Z = \frac{(N-A)}{SD}$, whereas N denotes the number of positives in the N-termini of the real proteome. A and SD are derived from 50 repetitions predicting positives in randomly chosen segments of 25 aa length (one segment per protein), whereas A corresponds to the average number of positives in the 50 runs and SD to their standard deviation.

Results The application on the multitude of different genome provides pre-calculated data which can be further combined with other knowledge and guide experiments for the detection of novel effectors. The pre-calculated data for the gram(-) bacteria can be downloaded from www.effectors.org.

In organisms encoding a TTSS, a substantial fraction of proteins is predicted as secreted, varying between 2% and 7% percent with an average of 4% of all proteins. In species without a TTSS, smaller percentages should be found. However, the recovered (false) positive rate in these species is quite high and varies between the taxonomic groups. Interestingly the *Deinococci* (6%) and the Gram-positive *Actinobacteria* (up to 10%) exhibit high percentages of positives despite the differences in cell wall composition and the absence of a TTSS. Contrarily, *Archaea* and *Firmicutes* exhibit a very low amount of positives with 1%, respectively 2% on average. Between more closely related bacteria, similar percentages of predicted TTSS effectors were found in different strains of e.g. *S. enterica* (on average 3%) and *E. coli* (3%). The amoebae symbiont *Protochlamydia amoebophila* exhibits a slightly higher percentage (6.1%) compared to its chlamydial relatives, which are pathogens of animals and humans (on average 5%). Since the former genome encodes roughly the double amount of sequences as in the pathogenic species, the total counts of predicted effectors are much higher than for the pathogenic relatives. This could reflect a more diverse arsenal of effectors employed by the amoeba symbiont due to an adaptation to more diverse hosts which implies more flexibility in the modes of chlamydi-host interactions. A direct comparison of several *Gammaproteobacteria* with and without TTSS shows a little less amount of predicted effectors in the latter case. The pathogenic *Escherichia coli* strain O157:H7 EDL933, for example, comprises 3.9% of its' genome as possible effectors, more as the harmless K12 strain with 2.9%. The high rate of false positives has been further investigated by systematically comparing the composition of the N-termini with the functional parts of the proteins using Z-score statistics. A high average Z-score for a genome (>1) indicates a prevalence of true effector like sequences (many N-termini with signal like composition differing from the rest of the protein), otherwise, the genome provides many unusual sequences *per se* resulting in a high false positive rate. These Z-scores can therefore be used to classify genomes due to the ability of EffectiveT3 to distinguish between effector and non-effectors. As further investigated together with Dr. Thomas Rattei, the rate of false positives correlates with the CG content in Gram(+) and archeal genomes. We found the correlation in case of the gram(-) weaker (R^2=0.05 against R^2=0.17) for species harboring a TTSS indicating a certain evolutionary pressure against signals in non-effectors.

Discussion The correlation of the false positive rate of GC content can be explained by elongated genes with N-termini that have a similar compositional characteristic as would be produced by the intergenic regions which have been found to bear a signal in many cases in the frame-shift experiments. Such elongated gene starts can comprise either real evolutionary events or correspond to wrongly predicted gene starts. The latter ones have been shown to correlate positively with the CG content [314], a finding in congruence with the poitive correlation of false positive effector predictiosn and GC content. In the absence of a TTSS, the evolutionary pressure for arbitrary proteins not to encode effector like N-termini is also absent, which could lead to effector like proteins in some cases and therefore to false positive predictions in these genomes. Some effectors could be transported by the flagellar system as the recognition seems to be similar: Badea et al. showed that FliC which is normally secreted by the flagellar apparatus to form the flagellar hook can be transported by the TTSS in absence of the flagellum [315]. Interestingly, EffectiveT3 predicts FliC as TTSS substrate further supporting this finding. In consequence, flagellar systems could serve as alternative route for TTSS effectors in species with no detected TTSS.

3.2.7 Comparison to other methods

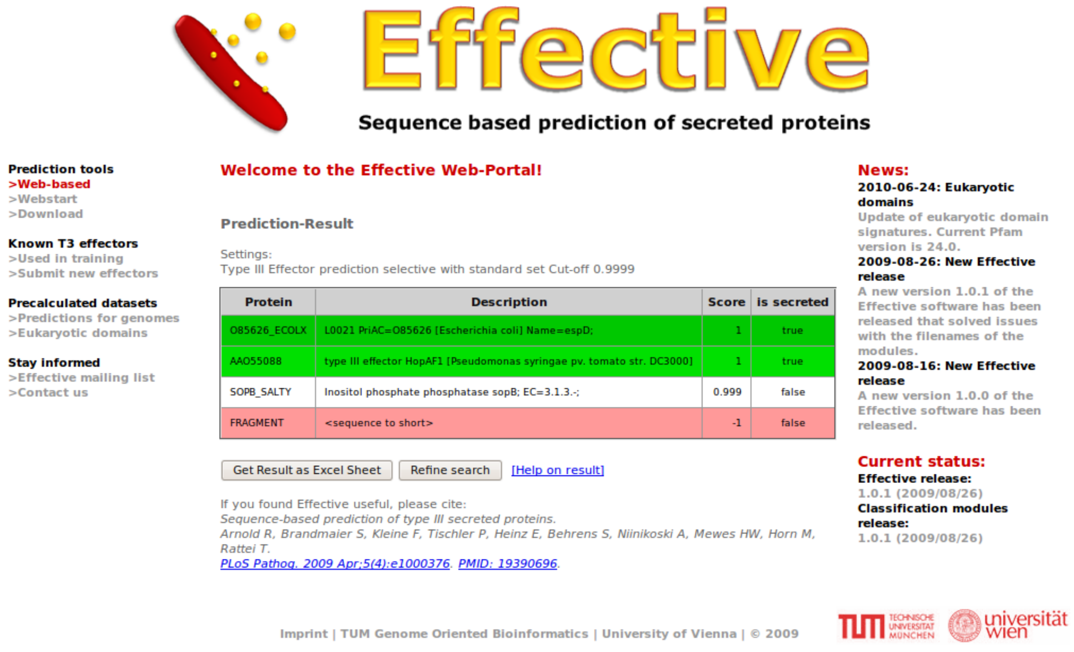
Three approaches, which utilize binary classifier to predict effector proteins, have been published recently. Löwer et al.[116] uses a neural network trained with string representations of the amino acid composition of the first 30 residues. The positive training set contained effectors of different organisms and from various studies including large-scale screens. Samudrala et al. [117] combined the amino acid composition of the first 20 amino acids with additional information as nucleotide composition of the gene, phylogenetic distribution of orthologs and the overall conservation of the protein as initial features and extracted the most discriminating features using recursive feature elimination. In this study, sequences from *P. syringae* and *Salmonella typhimurium* have been used as training for a support vector machine and the performance has been assessed by cross-validation as well as by a comparison between *Pseudomonas* and *Salmonella* indicating a common signal in both organisms. The EffetiveT3 method is the only one of them that is available as stand-alone tool and online-prediction server. Löwer and co-workers found the results of their study in good agreement with the predictions of EffectiveT3 [116]. Samudrala et al found as most discriminating features to be an enrichment of Serine and other polar residues and a depletion in charged residues. In this study found Serine, Threonine, Proline and polar residues as well as the amphipathic pattern “polar-hydrophobic-polar” enriched, whereas alkaline and acidic patterns were depleted in the N-terminal sequences. Overall, these studies are in good agreement with each other: all three detect the signal in the N-termini and can successfully distinguish effector and non-effector proteins.

3.2.8 The EffectiveT3 software package and Web-Page

To increase the impact of the Effective classification software in the scientific community, easy to use distributions of it are needed which can be used offline with private data and online by a web-interface. Tanja Bieber implemented a graphical user-interface during her Bachelor thesis in Java$^{\text{TM}}$(under my and Prof. Dr. Thomas Rattei supervision). This software allows to process own protein sequences either in graphical or batch mode. Different classification schemata (as complete, animal-pathogen, or plant-symbiont training set) can be loaded dynamically. This enables easy update of future classification schemata by the user.
I adapted the software for the use in a Web-Portal which allows the on-line prediction of TTSS substrates. In addition, I adapted the EffectiveT3 user interface to run as Java Web-Start$^{\text{TM}}$application. Both cases implied some modification of the original code concerning the dynamic loading of the classification modules in order to run in the Web-Server environment, the Web-Start environment, and as normal stand-alone software alike. The portal (http://www.effectors.org) has been implemented by Prof. Dr. Thomas Rattei, Marc-Andre Jehl, and me in a common effort using Java Server Pages. The portal also comprises predicted eukaryotic like domains (this part has been implemented by Marc-Andre Jehl, compare Chapter x.), pre-calculated predictions, supplementary data, and feedback forms to allow the submission of novel effector sequences and subscription to a newsletter. The intention of the portal is to create a platform for methods and data-sets related to effector candidates and known effectors. A screen-shot of an example prediction is shown in Figure 3.8.
Until today, the software has been cited in several publications, either in studies on novel effectors [316] or as part of analysis pipelines [317][318] indicating the actual need of the provided implementations of the method which are publicly accessible.
A paper describing the Web-Portal and the underlying database of eukaryotic like domains has been submitted.

3.2.9 Can a similar approach detect Type IV secreted proteins?

As in the case of Type III secretion, the mode of effector recognition of the Type IV system [319] is unknown and no general prediction software for its substrates exist that is based on a system specific recognition signal. The positive outcome of the Effective pipeline in the Type III case motivated to test the same machine-learning procedure on Type IV substrates. This has been done by Sebastian Behrens during his



Welcome to the Effective Web-Portal!

Prediction-Result

Settings:
Type III Effector prediction selective with standard set Cut-off 0.9999

Protein	Description	Score	is secreted
O85626_ECOLX	L0021 PriAC=O85626 [Escherichia coli] Name=espD;	1	true
AAO55088	type III effector HopAF1 [Pseudomonas syringae pv. tomato str. DC3000]	1	true
SOPB_SALTY	Inositol phosphate phosphatase sopB; EC=3.1.3.-;	0.999	false
FRAGMENT	<sequence to short>	-1	false

If you found Effective useful, please cite:
Sequence-based prediction of type III secreted proteins.
Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes HW, Horn M, Rattei T.
PLoS Pathog. 2009 Apr;5(4):e1000376. PMID: 19390696.

News:
2010-06-24: Eukaryotic domains
Update of eukaryotic domain signatures. Current Pfam version is 24.0.
2009-08-26: New Effective release
A new version 1.0.1 of the Effective software has been released that solved issues with the filenames of the modules.
2009-08-16: New Effective release
A new version 1.0.0 of the Effective software has been released.

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

Imprint | TUM Genome Oriented Bioinformatics | University of Vienna | © 2009
TUM TECHNISCHE UNIVERSITAT MÜNCHEN
universität wien

Figure 3.8: Screen-shot of a Effective Web-Portal session. The report of an example prediction is shown.

Bachelor thesis. The position of a recognition signal for Type IV secretion has been described for several *Legionella* effectors at the C-terminal end. Analyses equivalent to the Type III case revealed a predictive power above random for C-termini of known Type IV secreted effectors (AUC of 0.83). However, the sequences exhibit an alien composition over the whole sequence length that is as discriminative and the signal cannot be stated to be purely C-terminal. In conclusion, the classifier captures a general property of the sequences and a C-terminal signal could not be clearly stated. The training-set comprised in comparison with the Type III case only few instances, mainly *Legionella* effectors. This introduces two drawbacks: firstly, the compositional bias in the amino-acids which has been captured by the classificator could be explained by the existence of pathogenicity islands in *Legionella* which might exhibit different GC content as the genome average and influence the amino-acid composition [320]. Secondly, the taxonomic generality of the method could not be stated. In consequence, the resulting classifier must be seen as experimental but comprises a good starting point for further research.

In a cooperative project with the chair of Prof. Ellen Zechner of the University of Graz, I tested the trained Type IV classificator on the helicase TraI from *E. coli* which is an substrate of the Type IV secretion system. This sequence has been suspected to comprise internal secretion signals that are not C-terminal. Interestingly, a sliding window approach on the sequence revealed to positions in coincidence with experimental findings of the group. However, the significance of the result could not be stated. Further analyses on these position revealed their origin by a duplication event from a recD2 domain. A paper describing the internal secretion signal identified in the laboratory and our analyses on the phylogenetic relationships has been submitted.

4 Summary and conclusions

In this work, two aspects have been investigated: firstly, the prediction and use of functional interaction networks to investigate intra-cellular pathogens, namely the *Chlamydiae* by using their genome data. Secondly, I computationally modeled the secretion signal that leads to specific transport by the Type III secretion apparatus. This model provides insight into the molecular nature of the signal as well as a prediction method available to the scientific community. The key findings of this work are summarized in the following.

The prediction of Chlamydiae interactomes from their genomic sequences revealed a comprehensive map of the functional interplay of chlamydial proteins. A comprehensive knowledge of physical and functional interactions is crucial to understand the biology of an organism. Cellular functions are fulfilled not by single proteins alone but by the interplay of several proteins. Only few interactions of chlamydial proteins have been measured so far and no exhaustive protein-protein interaction screen is available. A computational prediction of the interactomes by genomic context and related methods is feasible as soon as the complete genome sequences are available. While such predicted interaction networks are available for publicly available genomes in the STRING database, this has not been the case for most of the environmental species. A pipeline to process the chlamydial genomes has been implemented that is able to integrate any computational interaction prediction method. The resulting data predicted by the implemented pipeline provides several extensions to the data found in similar resources as STRING[159]: it comprises chlamydia specific proteins that have no orthologous counterpart in other taxa and employs additional prediction methods that do not initially rely on the detection of orthologs in other species but on domain signatures. Furthermore, the pipeline employs different scoring schemata that weight the prediction methods according to their ability to predict physical or functional interactions. As an extension, the pipeline identifies cellular sub-systems, called functional modules. The predicted interactomes cover large fractions of the chlamydial genomes. Confident hypotheses on interactions have been generated for up to 65% of the predicted genes for the pathogenic *Chlamydiae*. Even in the case of the much larger genomes of the environmental species, it has been possible to predict interactions for a substantial fraction of proteins.

A substantial gain of information can be reached by integrating various methods. The integration of different prediction methods into a single scoring scheme has been shown to be generally fruitful in terms of increased confidence and coverage in several studies [114, 162, 163]. The genomic context methods (conserved gene neighborhood, the detection of fused proteins, and the co-occurrence method) contribute predictions which rely on the identification of orthologs between a plethora of species. In case of the chlamydia, several proteins exist which are taxon specific. The integration of domain based methods and the delineation of chlamydial specific conserved neighborhoods increased the coverage of the network since these taxon-specific proteins could be processed and several hundred proteins could be added to the interaction networks with confidence, especially by the domain based methods. Using the approaches followed herein, the coverage and quality of the interaction networks will increase with more chlamydial genomes sequenced which could contribute to detectable conserved gene neighborhoods reflecting chlamydia specific operons.

Functional modules of *Chlamydiae* reveal known pathways but also differ. The delineated functional modules have been tested on their broad functional coherence which can be stated to be much above random. As a consequence, these modules are highly probable meaningful functional entities. This picture is further supported by the recovery of known pathways and modules as defined in the KEGG database. However, the recovery is often partial and for many modules no assignment to a pathway could be made. The only partial recovery of known pathways is not surprising due to the reduced genomes of the *Chlamydiae*. The large fraction of modules completely uncovered indicates a need for additional, chlamydia specific pathway and module definitions not yet reflected in KEGG. The data generated in this analysis can give a basis for such an extension since it points out missing entities and chlamydia specific differences.

Functional modules can be successfully applied to predict functions of unknown proteins. Many proteins in the chlamydial genomes are of unknown function as the *Chlamydiae* are difficult to assess by standard laboratory techniques and are only distantly related to the typical model-organisms hindering an exhaustive automated function transfer by homology. The functional interaction networks and modules offer a convenient way to predict function by the guilt by association principle (i.e. by module co-membership or inference form the direct interaction neighbors). In this work, different strategies to employ this idea have been tested for *P. amoebophila* UWE25 and the FunCat annotation schema and for 178 unknown proteins, a function could be predicted. The quality of the predictions returned by the approaches (module co-membership and inference from neighbors in the interaction network) depends on the choice of adequate functional categories which have been automatically identified during the analysis. This information can serve as basis for an automated function prediction procedure in *Chlamydiae* that can complement i.e. homology based approaches.

Functional modules give hints to yet unknown virulence related genes. A necessary key feature of the *Chlamydiae* is their ability to manipulate their host cells by the use of secreted effector proteins. These (and other virulence related genes) could be identified by co-membership with virulence factors in functional modules. Although recovering a couple of known virulence factors such as Inclusion proteins and the effector CopN, the approach turned out to be not generally applicable since most known effector proteins do not cluster with their cognate transport system into modules. In general can be observed that effector proteins exhibit less predicted interactions of weaker confidence as arbitrary proteins. This indicates that most of them are individual inventions which are taxon or even species specific and their interaction with a transport system is therefore not detectable by the used methods in many cases. This finding motivates the development of a sequence based prediction method as it has been done in this work for Type III secreted proteins.

The genome reduction of pathogenic Chlamydiae in terms of functional modules: cohesive fate in most occasions. In contrast to the pathogenic species, the environmental *Chlamydiae* exhibit a weaker genome reduction. This allows to investigate how the functional equipment of the *Chlamydiae* changes in the process of adaptation to higher eukaryotes as host. This has been investigated on the level of functional modules. The further reduction of the pathogenic cases can be either explained by a concerted loss of functionally coupled proteins (cohesive losses), by a preference to keep them (purifying losses), or by a random reduction by arbitrary proteins (irregular behavior). The systematic investigation of these events revealed the existence of non-cohesive scenarios in some cases that indicates tolerance of single gene losses in these cases which must be complemented by functional alternatives. However, the majority of cases behaves cohesive and the examples of purifying gene loss exhibit the functional cores of the modules. One plausible interpretation is a loss of sub-functionalities of a module not needed in the more specific environment of the pathogenic *Chlamydiacea*. Cohesive loss of complete modules or cores indicate the absence of evolutionary pressure to maintain these functionalities if not complemented otherwise as by host parasitism. An interesting question for further research would be to ask if remaining sub-modules in the reduced cases neo-functionalize since they are de-coupled from the original functionality of the module or are found conserved because they comprise general functionalities not specific to their module.

Additional needs of the environmental *Chlamydiae*: functional differences to the pathogenic *Chlamydiae* due to a more variable environment. The comparison of the functional equipment between the environmental and pathogenic *Chlamydiae* indicated no detectable gain of functional modules in the latter case, but several modules existing in the former one. These functionalities comprise different transporter as an ABC multidrug and siderophore transporters, the Type IV secretion system, modules that are related to the exchange and repair of DNA, as well as modules of unclear function. This additional functional equipment indicates a more variable environment for the environmental species which is less stable as the eukaryotic host cell, a finding in congruence with the life-styles of the different *Chlamydiae*.

The recognition signal for Type III secretion is encoded in amino-acid N-terminus of effector proteins and can be computationally modeled. The Type III secretion system is a major transport route employed by many Gram(-) bacteria to deliver effector proteins into evolutionary host cells an the identification of novel effectors is crucial to understand the mechanisms of virulence. The general signal that leads to specific transport of effectors has been unknown and no prediction method to identify novel effectors from their sequence information existed. In this work it could be shown, that the model comprises an unusual pattern of amino-acids that can be used to model the signal computationally by a machine learning approach. The model allows systematic investigations on the nature of the signal which could be found encoded in the N-terminal peptide sequence.

The signal is taxonomically universal. Although the Type III secretion system is a well conserved entity, it has been not clear if the recognition signal is conserved in several organisms. An analysis using a taxon-specific validation procedure showed that the signal is indeed general and a computational prediction method based on it is generally applicable.

The signal is resistant against mutations: implications for the interpretation of frame-shift experiments and the evolution of novel effectors. The computational model has been used for *in silico* mutation experiments. Although the signal could be destroyed quickly if targeting key features of the signal, it turned out to be resilient against arbitrary point mutations. In some cases, the signal turned out to be immune against frame-shift mutations, a behavior that had been observed for some effector proteins [321, 322] and lead to the hypothesis of a mRNA encoded signal. The examples found herein give an alternative interpretation of these findings without the need to abolish the peptide-born signal hypothesis which is in general more likely due to several observations as the existence of translated effectors inside the bacterial cell. Furthermore, I found a weak initial signal for some intergenic regions in front of the coding sequences. This observation implies a novel possible mechanism of the *de novo* invention of effectors with low evolutionary costs by extending the coding sequence of an arbitrary protein into the intergenic region.

EffectiveT3, a software to predict Type III secreted proteins. Application and availability. The computational model can be used to predict novel effector candidates with high accuracy. A test on several genomes resulted in 2%-7% of a genome in cases of an existent Type III secretion system. Furthermore, possible sources of false positive predictions are discussed in this work. The software EffectiveT3 incorporates the computational model and is freely available as stand-alone software as well as in a web-interface.

Proposals for further experiments The computational model of the Type III secretion signal can guide point mutation experiments to further investigate the molecular mechanisms of Type III mediated transport. An interesting experiment would be to show secretion of a fusion protein consisting of an arbitrary gene fused on the N-terminus with a possible initial signal found in the nearby intergenic space. The transcriptional regulation of bacterial genes turned out to be much more complex than anticipated and even the organization of operon structures turns out to be flexible and dependent on the cells' current state (for a review, see [323]). Data resulting from transcriptomics studies as from deep RNA sequencing and tiling arrays will provide information on the transcriptional control of the TTSS, effectors, and related chaperons. This data will help to refine the prediction of candidate lists and to create dynamic models of the infection process. An interesting direction of research would be to determine commonalities and differences in substrate recognition between the Flagellum that has secretory capabilities [315] and the TTSS, as there might be common patterns between the substrate recognition of both systems congruent with their tight evolutionary relationship. The identification of a “switch-mechanism” between flagellar and TTSS mediated transport, which might be under expressional or chaperon based control remains an open question. Computational models will aid the investigation of this topic by determining differences in the signal, and by dynamic, time dependent computational models of infection and the cell cycle. Further research towards a better understanding of the TTSS and its molecular recognition of substrates is needed since effectors might be highly specific drug targets and novel types of antibacterial drug molecules could directly attack these effectors outside the bacterial cell, circumventing many of the bacterial resistance strategies.

5 Bibliography

[1] 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. *J Bacteriol*, 192(4):1045–1057, Feb 2010.

[2] Matthias Horn. Chlamydiae as symbionts in eukaryotes. *Annu Rev Microbiol*, 62:113–131, 2008.

[3] Brendan W Wren. The yersiniae–a model genus to study the rapid evolution of bacterial pathogens. *Nat Rev Microbiol*, 1(1):55–64, Oct 2003.

[4] E. A. Groisman and H. Ochman. How salmonella became a pathogen. *Trends Microbiol*, 5(9):343–349, Sep 1997.

[5] Nancy A Moran. Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, 108(5):583–586, Mar 2002.

[6] A. Mira, H. Ochman, and N. A. Moran. Deletional bias and the evolution of bacterial genomes. *Trends Genet*, 17(10):589–596, Oct 2001.

[7] Luís J Mota and Guy R Cornelis. The bacterial injection kit: type iii secretion systems. *Ann Med*, 37(4):234–249, 2005.

[8] Laure Journet, Kelly T Hughes, and Guy R Cornelis. Type iii secretion: a secretory pathway serving both motility and virulence (review). *Mol Membr Biol*, 22(1-2):41–50, 2005.

[9] Roman G Gerlach and Michael Hensel. Protein secretion systems and adhesins: the molecular armory of gram-negative pathogens. *Int J Med Microbiol*, 297(6):401–415, 2007.

[10] Daniela Büttner and Ulla Bonas. Common infection strategies of plant and animal pathogenic bacteria. *Curr Opin Plant Biol*, 6(4):312–319, Aug 2003.

[11] Mark J Pallen, Scott A Beatson, and Christopher M Bailey. Bioinformatics, genomics and evolution of non-flagellar type-iii secretion systems: a darwinian perspective. *FEMS Microbiol Rev*, 29(2):201–229, 2005.

[12] D. S. Beeckman and D. C. Vanrompay. Bacterial secretion systems with an emphasis on the chlamydial type iii secretion system. *Curr Issues Mol Biol*, 12(1):17–41, 2010. Journal Article England.

[13] Annick Gauthier and B. Brett Finlay. Translocated intimin receptor and its chaperone interact with atpase of the type iii secretion apparatus of enteropathogenic escherichia coli. *J Bacteriol*, 185(23):6747–6755, 2003.

[14] Petra J Edqvist, Jan Olsson, Moa Lavander, Lena Sundberg, Ake Forsberg, Hans Wolf-Watz, and Scott A Lloyd. Yscp and yscu regulate substrate specificity of the yersinia type iii secretion system. *J Bacteriol*, 185(7):2259–2266, 2003.

[15] Isabel Sorg, Stefanie Wagner, Marlise Amstutz, Shirley A Müller, Petr Broz, Yvonne Lussi, Andreas Engel, and Guy R Cornelis. Yscu recognizes translocators as export substrates of the yersinia injectisome. *EMBO J*, 26(12):3015–3024, Jun 2007.

[16] L. W. Cheng, D. M. Anderson, and O. Schneewind. Two independent type iii secretion mechanisms for yope in yersinia enterocolitica. *Mol Microbiol*, 24(4):757–765, 1997.

[17] P. Wattiau, B. Bernier, P. Deslée, T. Michiels, and G. R. Cornelis. Individual chaperones required for yop secretion by yersinia. *Proc Natl Acad Sci U S A*, 91(22):10493–10497, Oct 1994.

[18] Nikhil A Thomas, Wanyin Deng, Noel Baker, Jose Puente, and B. Brett Finlay. Hierarchical delivery of an essential host colonization factor in enteropathogenic escherichia coli. *J Biol Chem*, 282(40):29634–29645, 2007.

[19] Sang Ho Lee and Jorge E Galán. Salmonella type iii secretion-associated chaperones confer secretion-pathway specificity. *Mol Microbiol*, 51(2):483–495, Jan 2004.

[20] Carina R Büttner, Guy R Cornelis, Dirk W Heinz, and Hartmut H Niemann. Crystal structure of yersinia enterocolitica type iii secretion chaperone syct. *Protein Sci*, 14(8):1993–2002, Aug 2005.

[21] Artem G Evdokimov, Joseph E Tropea, Karen M Routzahn, and David S Waugh. Three-dimensional structure of the type iii secretion chaperone syce from yersinia pestis. *Acta Crystallogr D Biol Crystallogr*, 58(Pt 3):398–406, 2002.

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

[23] D. Hersh, D. M. Monack, M. R. Smith, N. Ghori, S. Falkow, and A. Zychlinsky. The salmonella invasin sipb induces macrophage apoptosis by binding to caspase-1. *Proc Natl Acad Sci U S A*, 96(5):2396–2401, 1999.

[24] Sarit Lilo, Ying Zheng, and James B Bliska. Caspase-1 activation in macrophages infected with yersinia pestis kim requires the type iii secretion system effector yopj. *Infect Immun*, 76(9):3911–3923, Sep 2008.

[25] J. H. Brumell, O. Steele-Mortimer, and B. B. Finlay. Bacterial invasion: Force feeding by salmonella. *Curr Biol*, 9(8):R277–R280, 1999.

[26] Imke Hansen-Wester, Bärbel Stecher, and Michael Hensel. Analyses of the evolutionary distribution of salmonella translocated effectors. *Infect Immun*, 70(3):1619–1622, 2002.

[27] 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, 2006.

[28] 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 Pathog*, 5(4):e1000376, Apr 2009.

[29] Y. Xiao, S. Heu, J. Yi, Y. Lu, and S. W. Hutcheson. Identification of a putative alternate sigma factor and characterization of a multicomponent regulatory cascade controlling the expression of pseudomonas syringae pv. syringae pss61 hrp and hrma genes. *J Bacteriol*, 176(4):1025–1036, Feb 1994.

[30] C. E. Stebbins and J. E. Galán. Structural mimicry in bacterial virulence. *Nature*, 412(6848):701–705, Aug 2001.

[31] K. Schesser, A. K. Spiik, J. M. Dukuzumuremyi, M. F. Neurath, S. Pettersson, and H. 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. *Mol Microbiol*, 28(6):1067–1079, Jun 1998.

[32] Aurélie Angot, Annette Vergunst, Stéphane 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, Jan 2007.

[33] T. Michiels and G. R. Cornelis. Secretion of hybrid proteins by the yersinia yop export system. *J Bacteriol*, 173(5):1677–1685, Mar 1991.

[34] Agathe Subtil, Cédric Delevoye, Maria-Eugenia Balañá, Laurence Tastevin, Stéphanie 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. *Mol Microbiol*, 56(6):1636–1647, 2005.

[35] 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. *Proc Natl Acad Sci U S A*, 103(40):14941–14946, 2006.

[36] Xavier Charpentier and Eric Oswald. Identification of the secretion and translocation domain of the enteropathogenic and enterohemorrhagic escherichia coli effector cif, using tem-1 beta-lactamase as a new fluorescence-based reporter. *J Bacteriol*, 186(16):5486–5495, 2004.

[37] Joseph A Sorg, Nathan C Miller, and Olaf Schneewind. Substrate recognition of type iii secretion machines-testing the rna signal hypothesis. *Cell Microbiol*, 7(9):1217–1225, 2005.

[38] D. M. Anderson and O. Schneewind. A mrna signal for the type iii secretion of yop proteins by yersinia enterocolitica. *Science*, 278(5340):1140–1143, 1997.

[39] D. M. Anderson and O. Schneewind. Yersinia enterocolitica type iii secretion: an mrna signal that couples translation and secretion of yopq. *Mol Microbiol*, 31(4):1139–1148, 1999.

[40] Kumaran S Ramamurthi and Olaf Schneewind. Yersinia yopq mrna encodes a bipartite type iii secretion signal in the first 15 codons. *Mol Microbiol*, 50(4):1189–1198, 2003.

[41] Bill Blaylock, Joseph A Sorg, and Olaf Schneewind. Yersinia enterocolitica type iii secretion of yopr requires a structure in its mrna. *Mol Microbiol*, 70(5):1210–1222, Dec 2008.

[42] 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. *Proc Natl Acad Sci U S A*, 96(22):12839–12843, 1999.

[43] M. H. Karavolos, A. J. Roe, M. Wilson, J. Henderson, J. J. Lee, D. L. Gally, and C. M A Khan. Type iii secretion of the salmonella effector protein sope is mediated via an n-terminal amino acid signal and not an mrna sequence. *J Bacteriol*, 187(5):1559–1567, 2005.

[44] Markus C Schlumberger, Andreas J Müller, Kristin Ehrbar, Brit Winnen, Iwan Duss, Bärbel Stecher, and Wolf-Dietrich Hardt. Real-time imaging of type iii secretion: Salmonella sipa injection into host cells. *Proc Natl Acad Sci U S A*, 102(35):12548–12553, Aug 2005.

[45] Jost Enninga, Joëlle Mounier, Philippe Sansonetti, and Guy Tran Van Nhieu. Secretion of type iii effectors into host cells in real time. *Nat Methods*, 2(12):959–965, Dec 2005.

[46] S. A. Lloyd, M. Norman, R. Rosqvist, and H. Wolf-Watz. Yersinia yope is targeted for type iii secretion by n-terminal, not mrna, signals. *Mol Microbiol*, 39(2):520–531, 2001.

[47] Gottfried Wilharm, Verena Lehmann, Wibke Neumayer, Janja Trcek, and Jürgen Heesemann. Yersinia enterocolitica type iii secretion: evidence for the ability to transport proteins that are folded prior to secretion. *BMC Microbiol*, 4:27, 2004.

[48] Boris A Vinatzer, Gail M Teitzel, Min-Woo Lee, Joanna Jelenska, Sara Hotton, Keke Fairfax, Jenny Jenrette, and Jean T Greenberg. The type iii effector repertoire of pseudomonas syringae pv. syringae b728a and its role in survival and disease on host and non-host plants. *Mol Microbiol*, 62(1):26–44, 2006.

[49] David J Studholme, Selena Gimenez Ibanez, Daniel MacLean, Jeffery L Dangl, Jeff H Chang, and John P Rathjen. A draft genome sequence and functional screen reveals the repertoire of type iii secreted proteins of pseudomonas syringae pathovar tabaci 11528. *BMC Genomics*, 10:395, 2009.

[50] A. P. Boyd, I. Lambermont, and G. R. Cornelis. Competition between the yops of yersinia enterocolitica for delivery into eukaryotic cells: role of the syce chaperone binding domain of yope. *J Bacteriol*, 182(17):4811–4821, 2000.

[51] Marc Valls, Stéphan Genin, and Christian Boucher. Integrated regulation of the type iii secretion system and other virulence determinants in ralstonia solanacearum. *PLoS Pathog*, 2(8):e82, 2006.

[52] R. W. Innes, A. F. Bent, B. N. Kunkel, S. R. Bisgrove, and B. J. Staskawicz. Molecular analysis of avirulence gene avrrpt2 and identification of a putative regulatory sequence common to all known pseudomonas syringae avirulence genes. *J Bacteriol*, 175(15):4859–4869, Aug 1993.

[53] S. Fenselau and U. Bonas. Sequence and expression analysis of the hrpb pathogenicity operon of xanthomonas campestris pv. vesicatoria which encodes eight proteins with similarity to components of the hrp, ysc, spa, and fli secretion systems. *Mol Plant Microbe Interact*, 8(6):845–854, 1995.

[54] Sunao Iyoda and Haruo Watanabe. Clpxp protease controls expression of the type iii protein secretion system through regulation of rpos and grlr levels in enterohemorrhagic escherichia coli. *J Bacteriol*, 187(12):4086–4094, Jun 2005.

[55] Sunao Iyoda, Nobuo Koizumi, Hitomi Satou, Yan Lu, Takehito Saitoh, Makoto Ohnishi, and Haruo Watanabe. The grlr-grla regulatory system coordinately controls the expression of flagellar and lee-encoded type iii protein secretion systems in enterohemorrhagic escherichia coli. *J Bacteriol*, 188(16):5682–5692, Aug 2006.

[56] K. H. Darwin and V. L. Miller. Type iii secretion chaperone-dependent regulation: activation of virulence genes by sica and invf in salmonella typhimurium. *EMBO J*, 20(8):1850–1862, 2001.

[57] Derrick E Fouts, Robert B Abramovitch, James R Alfano, Angela M Baldo, C. Robin Buell, Samuel Cartinhour, Arun K Chatterjee, Mark D'Ascenzo, Michelle L Gwinn, Sondra G Lazarowitz, Nai-Chun Lin, Gregory B Martin, Amos H Rehm, David J Schneider, Karin van Dijk, Xiaoyan Tang, and Alan Collmer. Genomewide identification of pseudomonas syringae pv. tomato dc3000 promoters controlled by the hrpl alternative sigma factor. *Proc Natl Acad Sci U S A*, 99(4):2275–2280, 2002.

[58] Wei Jiang, Bo-Le Jiang, Rong-Qi Xu, Jun-Ding Huang, Hong-Yu Wei, Guo-Feng Jiang, Wei-Jian Cen, Jiao Liu, Ying-Ying Ge, Guang-Hua Li, Li-Li Su, Xiao-Hong Hang, Dong-Jie Tang, Guang-Tao Lu, Jia-Xun Feng, Yong-Qiang He, and Ji-Liang Tang. Identification of six type iii effector genes with the pip box in xanthomonas campestris pv. campestris and five of them contribute individually to full pathogenicity. *Mol Plant Microbe Interact*, 22(11):1401–1411, Nov 2009.

[59] Ekaterina M Panina, Seema Mattoo, Natasha Griffith, Natalia A Kozak, Ming H Yuk, and Jeff F Miller. A genome-wide screen identifies a bordetella type iii secretion effector and candidate effectors in other species. *Mol Microbiol*, 58(1):267–279, 2005.

[60] Xin Hu, Michael S Lee, and Anders Wallqvist. Interaction of the disordered yersinia effector protein yope with its cognate chaperone syce. *Biochemistry*, 48(47):11158–11160, Dec 2009.

[61] Scott A Lloyd, Michael Sjöström, Sara Andersson, and Hans Wolf-Watz. Molecular characterization of type iii secretion signals via analysis of synthetic n-terminal amino acid sequences. *Mol Microbiol*, 43(1):51–59, 2002.

[62] Lisa M Schechter, Kathy A Roberts, Yashitola Jamir, James R Alfano, and Alan Collmer. Pseudomonas syringae type iii secretion system targeting signals and novel effectors studied with a cya translocation reporter. *J Bacteriol*, 186(2):543–555, 2004.

[63] Tanja Petnicki-Ocwieja, David J Schneider, Vincent C Tam, Scott T Chancey, Libo Shan, Yashitola Jamir, Lisa M Schechter, Misty D Janes, C. Robin Buell,

Xiaoyan Tang, Alan Collmer, and James R Alfano. Genomewide identification of proteins secreted by the hrp type iii protein secretion system of pseudomonas syringae pv. tomato dc3000. *Proc Natl Acad Sci U S A*, 99(11):7652–7657, May 2002.

[64] A. Subtil, C. Parsot, and A. Dautry-Varsat. Secretion of predicted inc proteins of chlamydia pneumoniae by a heterologous type iii machinery. *Mol Microbiol*, 39(3):792–800, Feb 2001.

[65] K. Al-Hasani, B. Adler, K. Rajakumar, and H. Sakellaris. Distribution and structural variation of the she pathogenicity island in enteric bacterial pathogens. *J Med Microbiol*, 50(9):780–786, 2001.

[66] R. S. Stephens, S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, R. L. Tatusov, Q. Zhao, E. V. Koonin, and R. W. Davis. Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. *Science*, 282(5389):754–759, Oct 1998.

[67] Jan Peters, David P Wilson, Garry Myers, Peter Timms, and Patrik M Bavoil. Type iii secretion Ã la chlamydia. *Trends Microbiol*, 15(6):241–251, 2007.

[68] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak. Improved prediction of signal peptides: Signalp 3.0. *J Mol Biol*, 340(4):783–95, 2004. Comparative Study Journal Article Research Support, Non-U.S. Gov't England.

[69] 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, Jul 2009.

[70] L Halberstädter and SV Prowazek. Über zelleinschlüsse parasitärer natur beim trachom. *Arbeiten aus dem Kaiserlichen Gesundheitsamte, Berlin*, 26:44?47, 1907.

[71] Matthias Horn, Astrid Collingro, Stephan Schmitz-Esser, Cora L Beier, Ulrike Purkhold, Berthold Fartmann, Petra Brandt, Gerald J Nyakatura, Marcus Droege, Dmitrij Frishman, Thomas Rattei, Hans-Werner Mewes, and Michael Wagner. Illuminating the evolutionary history of chlamydiae. *Science*, 304(5671):728–730, Apr 2004.

[72] Matthias Horn and Michael Wagner. Bacterial endosymbionts of free-living amoebae. *J Eukaryot Microbiol*, 51(5):509–514, 2004.

[73] H. Fukushi and K. Hirai. Genetic diversity of avian and mammalian chlamydia psittaci strains and relation to host origin. *J Bacteriol*, 171(5):2850–2855, May 1989.

[74] Silke Ruhl, Genevieve Goy, Nicola Casson, Rudolf Thoma, Andreas Pospischil, Gilbert Greub, and Nicole Borel. Parachlamydia acanthamoebae infection and abortion in small ruminants. *Emerg Infect Dis*, 14(12):1966–1968, Dec 2008.

[75] Andrew Draghi, Vsevolod L Popov, Melissa M Kahl, James B Stanton, Corrie C Brown, Gregory J Tsongalis, A. Brian West, and Salvatore Frasca. Characterization of "candidatus piscichlamydia salmonis" (order chlamydiales), a chlamydia-like bacterium associated with epitheliocystis in farmed atlantic salmon (salmo salar). *J Clin Microbiol*, 42(11):5286–5297, Nov 2004.

[76] L. Berger, K. Volp, S. Mathews, R. Speare, and P. Timms. Chlamydia pneumoniae in a free-ranging giant barred frog (mixophyes iteratus) from australia. *J Clin Microbiol*, 37(7):2378–2380, Jul 1999.

[77] N. Borel, R. Thoma, P. Spaeni, R. Weilenmann, K. Teankum, E. Brugnera, D. R. Zimmermann, L. Vaughan, and A. Pospischil. Chlamydia-related abortions in cattle from graubunden, switzerland. *Vet Pathol*, 43(5):702–708, Sep 2006.

[78] Massimiliano Don, Mika Paldanius, Lolita Fasoli, Mario Canciani, and Matti Korppi. Simkania negevensis and pneumonia in children. *Pediatr Infect Dis J*, 25(5):470–1; author reply 471–2, May 2006.

[79] K. D. Everett. Chlamydia and chlamydiales: more than meets the eye. *Vet Microbiol*, 75(2):109–126, Jul 2000.

[80] L. Weström, R. Joesoef, G. Reynolds, A. Hagdu, and S. E. Thompson. Pelvic inflammatory disease and fertility. a cohort study of 1,844 women with laparoscopically verified disease and 657 control women with normal laparoscopic results. *Sex Transm Dis*, 19(4):185–192, 1992.

[81] M. J. Hare and R. N. Thin. Chlamydial infection of the lower genital tract of women. *Br Med Bull*, 39(2):138–144, Apr 1983.

[82] G. D. Fang, M. Fine, J. Orloff, D. Arisumi, V. L. Yu, W. Kapoor, J. T. Grayston, S. P. Wang, R. Kohler, and R. R. Muder. New and emerging etiologies for

community-acquired pneumonia with implications for therapy. a prospective multicenter study of 359 cases. *Medicine (Baltimore)*, 69(5):307–316, Sep 1990.

[83] Hammerschlag. The role of chlamydia in upper respiratory tract infections. *Curr Infect Dis Rep*, 2(2):115–120, Apr 2000.

[84] Y. Wang. Etiology of trachoma: a great success in isolating and cultivating chlamydia trachomatis. *Chin Med J (Engl)*, 112(10):938–941, Oct 1999.

[85] A. E. Washington, R. E. Johnson, and L. L. Sanders. Chlamydia trachomatis infections in the united states. what are they costing us? *JAMA*, 257(15):2070–2072, Apr 1987.

[86] Candice M Mitchell, Kelley M Hovis, Patrik M Bavoil, Garry S A Myers, Jose A Carrasco, and Peter Timms. Comparison of koala lpcoln and human strains of chlamydia pneumoniae highlights extended genetic diversity in the species. *BMC Genomics*, 11:442, 2010.

[87] K. A. McColl, R. W. Martin, L. J. Gleeson, K. A. Handasyde, and A. K. Lee. Chlamydia infection and infertility in the female koala (phascolarctos cinereus). *Vet Rec*, 115(25-26):655, 1984.

[88] H. Chanton-Greutmann, R. Thoma, L. Corboz, N. Borel, and A. Pospischil. [abortion in small ruminants in switzerland: investigations during two lambing seasons (1996-1998) with special regard to chlamydial abortions]. *Schweiz Arch Tierheilkd*, 144(9):483–492, Sep 2002.

[89] R. M. Bush and K. D. Everett. Molecular evolution of the chlamydiaceae. *Int J Syst Evol Microbiol*, 51(Pt 1):203–220, Jan 2001.

[90] Eugene V Koonin and Yuri I Wolf. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res*, 36(21):6688–6719, Dec 2008.

[91] Tal Dagan, Ran Blekhman, and Dan Graur. The "domino theory" of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. *Mol Biol Evol*, 23(2):310–316, Feb 2006.

[92] Claire Bertelli, François Collyn, Antony Croxatto, Christian Rückert, Adam Polkinghorne, Carole Kebbi-Beghdadi, Alexander Goesmann, Lloyd Vaughan, and

Gilbert Greub. The waddlia genome: a window into chlamydial biology. *PLoS One*, 5(5):e10890, 2010.

[93] R. Michel, M. Steinert, L. Zoeller, B. Hauroeder, and K. Henning. Free-living amoebae may serve as hosts for the chlamydia-like bacterium waddlia chondrophila isolated from an aborted bovine foetus. *Acta Protozoologica*, 43:37–42, 2004.

[94] Geneviève Goy, Antony Croxatto, and Gilbert Greub. Waddlia chondrophila enters and multiplies within human macrophages. *Microbes Infect*, 10(5):556–562, Apr 2008.

[95] K. M. Kocan, T. B. Crawford, P. M. Dilbeck, J. F. Evermann, and T. C. McGuire. Development of a rickettsia isolated from an aborted bovine fetus. *J Bacteriol*, 172(10):5949–5955, Oct 1990.

[96] Kishore R Sakharkar, Pawan Kumar Dhar, and Vincent T K Chow. Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. *Int J Syst Evol Microbiol*, 54(Pt 6):1937–1941, Nov 2004.

[97] J. Tjaden, H. H. Winkler, C. Schwöppe, M. Van Der Laan, T. Möhlmann, and H. E. Neuhaus. Two nucleotide transport proteins in chlamydia trachomatis, one for net nucleoside triphosphate uptake and the other for transport of energy. *J Bacteriol*, 181(4):1196–1202, Feb 1999.

[98] Stephan Schmitz-Esser, Nicole Linka, Astrid Collingro, Cora L Beier, H. Ekkehard Neuhaus, Michael Wagner, and Matthias Horn. Atp/adp translocases: a common feature of obligate intracellular amoebal symbionts related to chlamydiae and rickettsiae. *J Bacteriol*, 186(3):683–691, Feb 2004.

[99] Ilka Haferkamp, Stephan Schmitz-Esser, Nicole Linka, Claude Urbany, Astrid Collingro, Michael Wagner, Matthias Horn, and H. Ekkehard Neuhaus. A candidate nad+ transporter in an intracellular bacterial symbiont related to chlamydiae. *Nature*, 432(7017):622–625, Dec 2004.

[100] Yasser M Abdelrahman and Robert J Belland. The chlamydial developmental cycle. *FEMS Microbiol Rev*, 29(5):949–959, Nov 2005.

[101] K. A. Fields, D. J. Mead, C. A. Dooley, and T. Hackstadt. Chlamydia trachomatis type iii secretion: evidence for a functional apparatus during early-cycle development. *Mol Microbiol*, 48(3):671–683, May 2003.

[102] D. R. Clifton, K. A. Fields, S. S. Grieshaber, C. A. Dooley, E. R. Fischer, D. J. Mead, R. A. Carabeo, and T. Hackstadt. A chlamydial type iii translocated protein is tyrosine-phosphorylated at the site of entry and associated with recruitment of actin. *Proc Natl Acad Sci U S A*, 101(27):10166–10171, Jul 2004.

[103] Robert J Belland, Guangming Zhong, Deborah D Crane, Daniel Hogan, Daniel Sturdevant, Jyotika Sharma, Wandy L Beatty, and Harlan D Caldwell. Genomic transcriptional profiling of the developmental cycle of chlamydia trachomatis. *Proc Natl Acad Sci U S A*, 100(14):8478–8483, Jul 2003.

[104] Ding Chen, Lei Lei, Chunxie Lu, Rhonda Flores, Matthew Delisa, Tucker C Roberts, Floyd E Romesberg, and Guangming Zhong. Secretion of the chlamydial virulence factor cpaf requires sec-dependent pathway. *Microbiology*, Jun 2010.

[105] Dagmar Heuer, Anette Rejman Lipinski, Nikolaus Machuy, Alexander Karlas, Andrea Wehrens, Frank Siedler, Volker Brinkmann, and Thomas F Meyer. Chlamydia causes fragmentation of the golgi compartment to ensure reproduction. *Nature*, 457(7230):731–735, Feb 2009.

[106] Tracy L Nicholson, Lynn Olinger, Kimberley Chong, Gary Schoolnik, and Richard S Stephens. Global stage-specific gene regulation during the developmental cycle of chlamydia trachomatis. *J Bacteriol*, 185(10):3179–3189, May 2003.

[107] T. Hackstadt, M. A. Scidmore-Carlson, E. I. Shaw, and E. R. Fischer. The chlamydia trachomatis inca protein is required for homotypic vesicle fusion. *Cell Microbiol*, 1(2):119–130, Sep 1999.

[108] Dagmar Heuer, Christoph Kneip, André P Mäurer, and Thomas F Meyer. Tackling the intractable - approaching the genetics of chlamydiales. *Int J Med Microbiol*, 297(7-8):569–576, Nov 2007.

[109] Michael Wagner and Matthias Horn. The planctomycetes, verrucomicrobia, chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr Opin Biotechnol*, 17(3):241–249, Jun 2006.

[110] Thomas Rattei, Stephan Ott, Michaela Gutacker, Jan Rupp, Matthias Maass, Stefan Schreiber, Werner Solbach, Thierry Wirth, and Jens Gieffers. Genetic diversity of the obligate intracellular bacterium chlamydophila pneumoniae by genome-wide

analysis of single nucleotide polymorphisms: evidence for highly clonal population structure. *BMC Genomics*, 8:355, 2007.

[111] Antje Krause, Jens Stoye, and Martin Vingron. Large scale hierarchical clustering of protein sequences. *BMC Bioinformatics*, 6:15, 2005.

[112] R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin. The cog database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, 29(1):22–28, Jan 2001.

[113] Albert-László Barabási and Zoltán N Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, Feb 2004.

[114] Christian von Mering, Evgeny M Zdobnov, Sophia Tsoka, Francesca D Ciccarelli, Jose B Pereira-Leal, Christos A Ouzounis, and Peer Bork. Genome evolution reveals biochemical networks and functional modules. *Proc Natl Acad Sci U S A*, 100(26):15428–15433, Dec 2003.

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

[116] M. Lower and G. Schneider. Prediction of type iii secretion signals in genomes of gram-negative bacteria. *PLoS One*, 4(6):e5917, 2009. Journal Article Research Support, Non-U.S. Gov't United States.

[117] 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. GM068152/GM/NIGMS NIH HHS/United States Y1-AI-4894-01/AI/NIAID NIH HHS/United States Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. United States.

[118] C. H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub. Molecular classification of multiple tumor types. *Bioinformatics*, 17 Suppl 1:S316–S322, 2001.

[119] Caroline C Friedel, Katharina H V Jahn, Selina Sommer, Stephen Rudd, Hans W Mewes, and Igor V Tetko. Support vector machines for separation of mixed plant-pathogen est collections based on codon usage. *Bioinformatics*, 21(8):1383–1388, Apr 2005.

[120] Qinghua Cui, Tianzi Jiang, Bing Liu, and Songde Ma. Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. *BMC Bioinformatics*, 5:66, May 2004.

[121] Yu-Dong Cai and Andrew J Doig. Prediction of saccharomyces cerevisiae protein functional class from functional domain composition. *Bioinformatics*, 20(8):1292–1300, May 2004.

[122] Igor V Tetko, Igor V Rodchenkov, Mathias C Walter, Thomas Rattei, and Hans-Werner Mewes. Beyond the 'best' match: machine learning annotation of protein sequences by integration of different sources of information. *Bioinformatics*, 24(5):621–628, Mar 2008.

[123] I. H. Witten and E. Frank. Data mining: Practical machine learning tools and techniques. 2005.

[124] Eugene V Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, 39:309–338, 2005.

[125] O. Eulenstein, B. Mirkin, and M. Vingron. Duplication-based measures of difference between gene and species trees. *J Comput Biol*, 5(1):135–148, 1998.

[126] Lars Juhl Jensen, Philippe Julien, Michael Kuhn, Christian von Mering, Jean Muller, Tobias Doerks, and Peer Bork. eggnog: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*, 36(Database issue):D250–D254, Jan 2008.

[127] M. Remm, C. E. Storm, and E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5):1041–1052, Dec 2001.

[128] P. Erdős and A. Rényi. *On the evolution of random graphs.* Citeseer, 1960.

[129] Barabasi and Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct 1999.

[130] Yuri I Wolf, Georgy Karev, and Eugene V Koonin. Scale-free networks in biology: new insights into the fundamentals of evolution? *Bioessays*, 24(2):105–109, Feb 2002.

[131] Eugene V Koonin, Yuri I Wolf, and Georgy P Karev. The structure of the protein universe and genome evolution. *Nature*, 420(6912):218–223, Nov 2002.

[132] M. Schnegg. Reciprocity and the emergence of power laws in social networks. *Arxiv preprint physics/0603005*, 2006.

[133] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, Oct 2000.

[134] Tong Ihn Lee, Nicola J Rinaldi, François Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, Christopher T Harbison, Craig M Thompson, Itamar Simon, Julia Zeitlinger, Ezra G Jennings, Heather L Murray, D. Benjamin Gordon, Bing Ren, John J Wyrick, Jean-Bosco Tagne, Thomas L Volkert, Ernest Fraenkel, David K Gifford, and Richard A Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, Oct 2002.

[135] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, Oct 2002.

[136] Rune Linding, Lars Juhl Jensen, Gerard J Ostheimer, Marcel A T M van Vugt, Claus Jørgensen, Ioana M Miron, Francesca Diella, Karen Colwill, Lorne Taylor, Kelly Elder, Pavel Metalnikov, Vivian Nguyen, Adrian Pasculescu, Jing Jin, Jin Gyoon Park, Leona D Samson, James R Woodgett, Robert B Russell, Peer Bork, Michael B Yaffe, and Tony Pawson. Systematic discovery of in vivo phosphorylation networks. *Cell*, 129(7):1415–1426, Jun 2007.

[137] Haiyuan Yu, Philip M Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, 3(4):e59, Apr 2007.

[138] Philip M Kim, Long J Lu, Yu Xia, and Mark B Gerstein. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314(5807):1938–1941, Dec 2006.

[139] Haiyuan Yu, Pascal Braun, Muhammed A Yildirim, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, Tong Hao, Jean-François Rual, Amélie Dricot, Alexei Vazquez, Ryan R Murray, Christophe Simon, Leah Tardivo, Stanley Tam, Nenad Svrzikapa, Changyu Fan, Anne-Sophie de Smet, Adriana Motyl, Michael E Hudson, Juyong Park, Xiaofeng Xin, Michael E Cusick, Troy Moore, Charlie Boone, Michael Snyder, Frederick P Roth, Albert-László Barabási, Jan Tavernier, David E Hill, and Marc Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, Oct 2008.

[140] K. Young, S. Lin, L. Sun, E. Lee, M. Modi, S. Hellings, M. Husbands, B. Ozenberger, and R. Franco. Identification of a calcium channel modulator using a high throughput yeast two-hybrid screen. *Nat Biotechnol*, 16(10):946–950, Oct 1998.

[141] Anne-Claude Gavin, Markus Bösche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jörg Schultz, Jens M Rick, Anne-Marie Michon, Cristina-Maria Cruciat, Marita Remor, Christian Höfert, Malgorzata Schelder, Miro Brajenovic, Heinz Ruffner, Alejandro Merino, Karin Klein, Manuela Hudak, David Dickson, Tatjana Rudi, Volker Gnau, Angela Bauch, Sonja Bastuck, Bettina Huhse, Christina Leutwein, Marie-Anne Heurtier, Richard R Copley, Angela Edelmann, Erich Querfurth, Vladimir Rybin, Gerard Drewes, Manfred Raida, Tewis Bouwmeester, Peer Bork, Bertrand Seraphin, Bernhard Kuster, Gitte Neubauer, and Giulio Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, Jan 2002.

[142] Gareth Butland, José Manuel Peregrín-Alvarez, Joyce Li, Wehong Yang, Xiaochun Yang, Veronica Canadien, Andrei Starostine, Dawn Richards, Bryan Beattie, Nevan Krogan, Michael Davey, John Parkinson, Jack Greenblatt, and Andrew Emili. Interaction network containing conserved and essential protein complexes in escherichia coli. *Nature*, 433(7025):531–537, Feb 2005.

[143] Pingzhao Hu, Sarath Chandra Janga, Mohan Babu, J. Javier Díaz-Mejía, Gareth Butland, Wenhong Yang, Oxana Pogoutse, Xinghua Guo, Sadhna Phanse, Peter Wong, Shamanta Chandran, Constantine Christopoulos, Anaies Nazarians-Armavil, Negin Karimi Nasseri, Gabriel Musso, Mehrab Ali, Nazila Nazemof, Veronika Eroukova, Ashkan Golshani, Alberto Paccanaro, Jack F Greenblatt, Gabriel Moreno-Hagelsieb, and Andrew Emili. Global functional atlas of es-

cherichia coli encompassing previously uncharacterized proteins. *PLoS Biol*, 7(4):e96, Apr 2009.

[144] H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Güldener, G. Mannhaupt, M. Münsterkötter, P. Pagel, N. Strack, V. Stümpflen, J. Warfsmann, and A. Ruepp. Mips: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*, 32(Database issue):D41–D44, Jan 2004.

[145] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H-Werner Mewes. Corum: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res*, 38(Database issue):D497–D501, Jan 2010.

[146] Gary D Bader, Doron Betel, and Christopher W V Hogue. Bind: the biomolecular interaction network database. *Nucleic Acids Res*, 31(1):248–250, Jan 2003.

[147] Johannes Goll, Seesandra V Rajagopala, Shen C Shiau, Hank Wu, Brian T Lamb, and Peter Uetz. Mpidb: the microbial protein interaction database. *Bioinformatics*, 24(15):1743–1744, Aug 2008.

[148] Seesandra V Rajagopala, Johannes Goll, N. D Deve Gowda, Kumar C Sunil, Björn Titz, Arnab Mukherjee, Sharmila S Mary, Naresh Raviswaran, Chetan S Poojari, Srinivas Ramachandra, Svetlana Shtivelband, Stephen M Blazie, Julia Hofmann, and Peter Uetz. Mpi-lit: a literature-curated dataset of microbial binary protein–protein interactions. *Bioinformatics*, 24(22):2622–2627, Nov 2008.

[149] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–3, 1999. P01 GM 31299/GM/United States NIGMS Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. United states.

[150] R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, 96(6):2896–2901, Mar 1999.

[151] Jan O Korbel, Lars J Jensen, Christian von Mering, and Peer Bork. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotechnol*, 22(7):911–917, Jul 2004.

[152] T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23(9):324–328, Sep 1998.

[153] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–4288, Apr 1999.

[154] Peter M Bowers, Matteo Pellegrini, Mike J Thompson, Joe Fierro, Todd O Yeates, and David Eisenberg. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol*, 5(5):R35, 2004.

[155] 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 Res*, 14(6):1107–1118, Jun 2004.

[156] Magali Michaut, Samuel Kerrien, Luisa Montecchi-Palazzi, Franck Chauvat, Corinne Cassier-Chauvat, Jean-Christophe Aude, Pierre Legrain, and Henning Hermjakob. Interoporc: automated inference of highly conserved protein interaction networks. *Bioinformatics*, 24(14):1625–1631, Jul 2008.

[157] Monica Riley, Takashi Abe, Martha B Arnaud, Mary K B Berlyn, Frederick R Blattner, Roy R Chaudhuri, Jeremy D Glasner, Takashi Horiuchi, Ingrid M Keseler, Takehide Kosuge, Hirotada Mori, Nicole T Perna, Guy Plunkett, Kenneth E Rudd, Margrethe H Serres, Gavin H Thomas, Nicholas R Thomson, David Wishart, and Barry L Wanner. Escherichia coli k-12: a cooperatively developed annotation snapshot–2005. *Nucleic Acids Res*, 34(1):1–9, 2006.

[158] Robert D Finn, Mhairi Marshall, and Alex Bateman. ipfam: visualization of protein-protein interactions in pdb at domain and amino acid resolutions. *Bioinformatics*, 21(3):410–412, Feb 2005.

[159] Christian von Mering, Lars J Jensen, Michael Kuhn, Samuel Chaffron, Tobias Doerks, Beate Krüger, Berend Snel, and Peer Bork. String 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, 35(Database issue):D358–D362, Jan 2007.

[160] G. Traver Hart, Arun K Ramani, and Edward M Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biol*, 7(11):120, 2006.

[161] Insuk Lee, Shailesh V Date, Alex T Adai, and Edward M Marcotte. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555–1558, Nov 2004.

[162] Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F Greenblatt, and Mark Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, Oct 2003.

[163] Shailesh V Date and Christian J Stoeckert. Computational modeling of the plasmodium falciparum interactome reveals protein function on a genome-wide scale. *Genome Res*, 16(4):542–549, Apr 2006.

[164] Michael Strong, Parag Mallick, Matteo Pellegrini, Michael J Thompson, and David Eisenberg. Inference of protein function and protein linkages in mycobacterium tuberculosis based on prokaryotic genome organization: a combined computational approach. *Genome Biol*, 4(9):R59, 2003.

[165] Michael Strong, Thomas G Graeber, Morgan Beeby, Matteo Pellegrini, Michael J Thompson, Todd O Yeates, and David Eisenberg. Visualization and interpretation of protein networks in mycobacterium tuberculosis based on hierarchical clustering of genome-wide functional linkage maps. *Nucleic Acids Res*, 31(24):7099–7109, Dec 2003.

[166] Jianfei Hu, Jun Wan, Laszlo Hackler, Donald J Zack, and Jiang Qian. Computational analysis of tissue-specific gene networks: application to murine retinal functional studies. *Bioinformatics*, 26(18):2289–2297, Sep 2010.

[167] Christian von Mering, Lars J Jensen, Berend Snel, Sean D Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A Huynen, and Peer Bork. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33(Database issue):D433–D437, Jan 2005.

[168] Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic Acids Res*, 31(1):258–261, Jan 2003.

[169] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–C52, Dec 1999.

[170] S. Bergmann, J. Ihmels, and N. Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical review E*, 67(3):31902, 2003.

[171] Ruolin Yang and Bing Su. Characterization and comparison of the tissue-related modules in human and mouse. *PLoS One*, 5(7):e11730, 2010.

[172] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868, Dec 1998.

[173] Sabine Tornow and H. W. Mewes. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res*, 31(21):6283–6289, Nov 2003.

[174] U. de Lichtenberg, L.J. Jensen, S. Brunak, and P. Bork. Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724, 2005.

[175] Silpa Suthram, Joel T Dudley, Annie P Chiang, Rong Chen, Trevor J Hastie, and Atul J Butte. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput Biol*, 6(2):e1000662, 2010.

[176] Tara A Gianoulis, Jeroen Raes, Prianka V Patel, Robert Bjornson, Jan O Korbel, Ivica Letunic, Takuji Yamada, Alberto Paccanaro, Lars J Jensen, Michael Snyder, Peer Bork, and Mark B Gerstein. Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A*, 106(5):1374–1379, Feb 2009.

[177] Huiying Li, Matteo Pellegrini, and David Eisenberg. Detection of parallel functional modules by comparative analysis of genome sequences. *Nat Biotechnol*, 23(2):253–260, Feb 2005.

[178] Peter D Karp, Monica Riley, Milton Saier, Ian T Paulsen, Julio Collado-Vides, Suzanne M Paley, Alida Pellegrini-Toole, César Bonavides, and Socorro Gama-Castro. The ecocyc database. *Nucleic Acids Res*, 30(1):56–58, Jan 2002.

[179] Amos Tanay, Roded Sharan, Martin Kupiec, and Ron Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A*, 101(9):2981–2986, Mar 2004.

[180] Berend Snel and Martijn A Huynen. Quantifying modularity in the evolution of biomolecular systems. *Genome Res*, 14(3):391–397, Mar 2004.

[181] Mónica Campillos, Christian von Mering, Lars Juhl Jensen, and Peer Bork. Identification and analysis of evolutionarily cohesive functional modules in protein networks. *Genome Res*, 16(3):374–382, Mar 2006.

[182] Like Fokkens and Berend Snel. Cohesive versus flexible evolution of functional modules in eukaryotes. *PLoS Comput Biol*, 5(1):e1000276, Jan 2009.

[183] S. Van Dongen. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*, 30:121, 2008.

[184] Thomas Rattei, Patrick Tischler, Roland Arnold, Franz Hamberger, Jörg Krebs, Jan Krumsiek, Benedikt Wachinger, Volker Stümpflen, and Werner Mewes. Simap–structuring the network of protein similarities. *Nucleic Acids Res*, 36(Database issue):D289–D292, Jan 2008.

[185] Jimin Song and Mona Singh. How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics*, 25(23):3143–3150, Dec 2009.

[186] Jose B Pereira-Leal, Anton J Enright, and Christos A Ouzounis. Detection of functional modules from protein interaction networks. *Proteins*, 54(1):49–57, Jan 2004.

[187] Gary D Bader and Christopher W V Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, Jan 2003.

[188] Balázs Adamcsek, Gergely Palla, Illés J Farkas, Imre Derényi, and Tamás Vicsek. Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021–1023, Apr 2006.

[189] I. Derényi, G. Palla, and T. Vicsek. Clique percolation in random networks. *Physical review letters*, 94(16):160202, 2005.

[190] Brian P Kelley, Bingbing Yuan, Fran Lewitter, Roded Sharan, Brent R Stockwell, and Trey Ideker. Pathblast: a tool for alignment of protein interaction networks. *Nucleic Acids Res*, 32(Web Server issue):W83–W88, Jul 2004.

[191] R. Sharan, T. Ideker, B. Kelley, R. Shamir, and R.M. Karp. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *Journal of Computational Biology*, 12(6):835–846, 2005.

[192] S. Bruckner, F. H
"uffner, R.M. Karp, R. Shamir, and R. Sharan. Topology-free querying of protein interaction networks. *Journal of Computational Biology*, 17(3):237–252, 2010.

[193] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

[194] Ron Caspi, Hartmut Foerster, Carol A Fulcher, Rebecca Hopkinson, John Ingraham, Pallavi Kaipa, Markus Krummenacker, Suzanne Paley, John Pick, Seung Y Rhee, Christophe Tissier, Peifen Zhang, and Peter D Karp. Metacyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*, 34(Database issue):D511–D516, Jan 2006.

[195] Ross Overbeek, Tadhg Begley, Ralph M Butler, Jomuna V Choudhuri, Han-Yu Chuang, Matthew Cohoon, Valérie de Crécy-Lagard, Naryttza Diaz, Terry Disz, Robert Edwards, Michael Fonstein, Ed D Frank, Svetlana Gerdes, Elizabeth M Glass, Alexander Goesmann, Andrew Hanson, Dirk Iwata-Reuyl, Roy Jensen, Neema Jamshidi, Lutz Krause, Michael Kubal, Niels Larsen, Burkhard Linke, Alice C McHardy, Folker Meyer, Heiko Neuweger, Gary Olsen, Robert Olson, Andrei Osterman, Vasiliy Portnoy, Gordon D Pusch, Dmitry A Rodionov, Christian Rückert, Jason Steiner, Rick Stevens, Ines Thiele, Olga Vassieva, Yuzhen Ye, Olga Zagnitko, and Veronika Vonstein. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, 33(17):5691–5702, 2005.

[196] S. Dietmann, E. Georgii, A. Antonov, K. Tsuda, and H.W. Mewes. The dics repository: module-assisted analysis of disease-related gene lists. *Bioinformatics*, 25(6):830, 2009.

[197] U. Güldener, M. Münsterkötter, G. Kastenmüller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. García-Martínez, J. E. Pérez-Ortín, H. Michael, A. Kaps, E. Talla, B. Dujon, B. André, J. L. Souciet, J. De Montigny, E. Bon, C. Gaillardin, and H. W. Mewes. Cygd: the comprehensive yeast genome database. *Nucleic Acids Res*, 33(Database issue):D364–D368, Jan 2005.

[198] UniProt Consortium. The universal protein resource (uniprot). *Nucleic Acids Res*, 36(Database issue):D190–D195, Jan 2008.

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

[200] Nomenclature of multiple forms of enzymes. recommendations (1976) iupac-iub commission on biochemical nomenclature (cbn). *J Biol Chem*, 252(17):5939–5941, Sep 1977.

[201] Igor V Tetko, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Gisela Fobo, Andreas Ruepp, Alexey V Antonov, Dimitrij Surmeli, and Hans-Wernen Mewes. Mips bacterial genomes functional annotation benchmark dataset. *Bioinformatics*, 21(10):2520–2521, May 2005.

[202] Andreas Ruepp, Alfred Zollner, Dieter Maier, Kaj Albermann, Jean Hani, Martin Mokrejs, Igor Tetko, Ulrich Güldener, Gertrud Mannhaupt, Martin Münsterkötter, and H. Werner Mewes. The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*, 32(18):5539–5545, 2004.

[203] Lars J Jensen and Peer Bork. Ontologies in quantitative biology: a basis for comparison, integration, and discovery. *PLoS Biol*, 8(5):e1000374, May 2010.

[204] Dmitrij Frishman. Protein annotation at genomic scale: the current status. *Chem Rev*, 107(8):3448–3466, Aug 2007.

[205] C. A. Wilson, J. Kreychman, and M. Gerstein. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol*, 297(1):233–249, Mar 2000.

[206] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990. LM04960/LM/NLM NIH

HHS/United States LM05110/LM/NLM NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. England.

[207] Roland Arnold, Thomas Rattei, Patrick Tischler, Minh-Duc Truong, Volker Stümpflen, and Werner Mewes. Simap–the similarity matrix of proteins. *Bioinformatics*, 21 Suppl 2:ii42–ii46, Sep 2005.

[208] Thomas Rattei, Patrick Tischler, Stefan Götz, 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 Res*, 38(Database issue):D223–D226, Jan 2010.

[209] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman. The pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–22, 2010. WT077044/Z/05/Z/Wellcome Trust/United Kingdom Howard Hughes Medical Institute/United States Journal Article Research Support, Non-U.S. Gov't England.

[210] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. Interpro: the integrative protein signature database. *Nucleic Acids Res*, 37(Database issue):D211–5, 2009. BB/F010508/1/Biotechnology and Biological Sciences Research Council/United Kingdom GM081084/GM/NIGMS NIH HHS/United States Wellcome Trust/United Kingdom Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England.

[211] G. Kolesov, H. W. Mewes, and D. Frishman. Snapping up functionally related genes based on context information: a colinearity-free approach. *J Mol Biol*, 311(4):639–656, Aug 2001.

[212] Y. I. Wolf, I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res*, 11(3):356–372, Mar 2001.

[213] L. Aravind. Guilt by association: contextual information in genome analysis. *Genome Res*, 10(8):1074–1077, Aug 2000.

[214] Roded Sharan, Igor Ulitsky, and Ron Shamir. Network-based prediction of protein function. *Mol Syst Biol*, 3:88, 2007.

[215] W. R. Pearson. Rapid and sensitive sequence comparison with fastp and fasta. *Methods Enzymol*, 183:63–98, 1990.

[216] Trupti Joshi, Yu Chen, Jeffrey M Becker, Nickolai Alexandrov, and Dong Xu. Genome-scale gene function prediction using multiple sources of high-throughput data in yeast saccharomyces cerevisiae. *OMICS*, 8(4):322–333, 2004.

[217] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol*, 21(6):697–700, Jun 2003.

[218] Minghua Deng, Kui Zhang, Shipra Mehta, Ting Chen, and Fengzhu Sun. Prediction of protein function using protein-protein interaction data. *J Comput Biol*, 10(6):947–960, 2003.

[219] Minghua Deng, Zhidong Tu, Fengzhu Sun, and Ting Chen. Mapping gene ontology to proteins based on protein-protein interaction data. *Bioinformatics*, 20(6):895–902, Apr 2004.

[220] Elena Nabieva, Kam Jim, Amit Agarwal, Bernard Chazelle, and Mona Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21 Suppl 1:i302–i310, Jun 2005.

[221] R. Sharan, S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R.M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences of the United States of America*, 102(6):1974, 2005.

[222] Md Altaf-Ul-Amin, Yoko Shinbo, Kenji Mihara, Ken Kurokawa, and Shigehiko Kanaya. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, 7:207, 2006.

[223] M. E J Newman. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, 103(23):8577–8582, Jun 2006.

[224] F. Alizon, S.B. Shooter, and T.W. Simpson. Henry ford and the model t: lessons for product platforming and mass customization. *Design Studies*, 30(5):588–605, 2009.

[225] Minghua Deng, Kui Zhang, Shipra Mehta, Ting Chen, and Fengzhu Sun. Prediction of protein function using protein-protein interaction data. *Proc IEEE Comput Soc Bioinform Conf*, 1:197–206, 2002.

[226] Sarah Killcoyne, Gregory W Carter, Jennifer Smith, and John Boyle. Cytoscape: a community-based framework for network modeling. *Methods Mol Biol*, 563:219–239, 2009.

[227] Yassen Assenov, Fidel Ramírez, Sven-Eric Schelhorn, Thomas Lengauer, and Mario Albrecht. Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284, Jan 2008.

[228] Kevin Y Yip, Haiyuan Yu, Philip M Kim, Martin Schultz, and Mark Gerstein. The tyna platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics*, 22(23):2968–2970, Dec 2006.

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

[230] Philipp Pagel, Matthias Oesterheld, Oksana Tovstukhina, Norman Strack, Volker Stümpflen, and Dmitrij Frishman. Dima 2.0–predicted and known domain interactions. *Nucleic Acids Res*, 36(Database issue):D651–D655, Jan 2008.

[231] Amelie Stein, Robert B Russell, and Patrick Aloy. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res*, 33(Database issue):D413–D417, Jan 2005.

[232] Gary D Bader and Christopher W V Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, 20(10):991–997, Oct 2002.

[233] Christian von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, May 2002.

[234] Berend Snel, Peer Bork, and Martijn A Huynen. The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci U S A*, 99(9):5890–5895, Apr 2002.

[235] Hong-Wu Ma and An-Ping Zeng. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11):1423–1430, Jul 2003.

[236] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proc Natl Acad Sci U S A*, 101(9):2658–2663, Mar 2004.

[237] Raja Loganantharaj, Satish Cheepala, and John Clifford. Metric for measuring the effectiveness of clustering of dna microarray expression. *BMC Bioinformatics*, 7 Suppl 2:S5, 2006.

[238] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[239] 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 Res*, 33(Database issue):D325–D328, Jan 2005.

[240] Philippe Verbeke, Lynn Welter-Stahl, Songmin Ying, Jon Hansen, Georg Häcker, Toni Darville, and David M Ojcius. Recruitment of bad by the chlamydia trachomatis vacuole correlates with host-cell survival. *PLoS Pathog*, 2(5):e45, May 2006.

[241] Alexey V Antonov, Sabine Dietmann, and Hans W Mewes. Kegg spider: interpretation of genomics data in the context of the global gene metabolic network. *Genome Biol*, 9(12):R179, 2008.

[242] Alexey V Antonov, Thorsten Schmidt, Yu Wang, and Hans W Mewes. Profcom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic Acids Res*, 36(Web Server issue):W347–W351, Jul 2008.

[243] Sabine Dietmann, Wanseon Lee, Philip Wong, Igor Rodchenkov, and Alexey V Antonov. Ccancer: a bird's eye view on gene lists reported in cancer-related studies. *Nucleic Acids Res*, 38 Suppl:W118–W123, Jul 2010.

[244] Gabriel F Berriz, Oliver D King, Barbara Bryant, Chris Sander, and Frederick P Roth. Characterizing gene sets with funcassociate. *Bioinformatics*, 19(18):2502–2504, Dec 2003.

[245] K.A. Fields and T. Hackstadt. The chlamydial inclusion: Escape from the endocytic pathway. *Annual review of cell and developmental biology*, 18(1):221–245, 2002.

[246] JA Hartigan. Minimum mutation fits to a given tree. *Biometrics*, 29(1):53–65, 1973.

[247] Chris B Stone, David C Bulir, Jodi D Gilchrist, Raman K Toor, and James B Mahony. Interactions between flagellar and type iii secretion proteins in chlamydia pneumoniae. *BMC Microbiol*, 10:18, 2010.

[248] X. Liu and P. Matsumura. An alternative sigma factor controls transcription of flagellar class-iii operons in escherichia coli: gene sequence, overproduction, purification and characterization. *Gene*, 164(1):81–84, Oct 1995.

[249] L. M. Márquez-Magaña and M. J. Chamberlin. Characterization of the sigd transcription unit of bacillus subtilis. *J Bacteriol*, 176(8):2427–2434, Apr 1994.

[250] L. Zheng, V. L. Cash, D. H. Flint, and D. R. Dean. Assembly of iron-sulfur clusters. identification of an iscsua-hscba-fdx gene cluster from azotobacter vinelandii. *J Biol Chem*, 273(21):13264–13272, May 1998.

[251] H. Kakuda, K. Hosono, K. Shiroishi, and S. Ichihara. Identification and characterization of the acka (acetate kinase a)-pta (phosphotransacetylase) operon and complementation analysis of acetate utilization by an acka-pta deletion mutant of escherichia coli. *J Biochem*, 116(4):916–922, Oct 1994.

[252] H. Momose, H. Nishikawa, and I. Shiio. Regulation of purine nucleotide synthesis in bacillus subtilis. i. enzyme repression by purine derivatives. *J Biochem*, 59(4):325–331, Apr 1966.

[253] A. Aiba and K. Mizobuchi. Nucleotide sequence analysis of genes purh and purd involved in the de novo purine nucleotide biosynthesis of escherichia coli. *J Biol Chem*, 264(35):21239–21246, Dec 1989.

[254] P. W. Finch, P. Chambers, and P. T. Emmerson. Identification of the escherichia coli recn gene product as a major sos protein. *J Bacteriol*, 164(2):653–658, Nov 1985.

[255] Emmanuele Severi, Derek W Hood, and Gavin H Thomas. Sialic acid utilization by bacterial pathogens. *Microbiology*, 153(Pt 9):2817–2822, Sep 2007.

[256] Wen-Tssann Liu, Michail H Karavolos, David M Bulmer, Abdelmounaaïm Allaoui, Raquel Demarco Carlos E Hormaeche, Jeong Jin Lee, and C. M Anjam Khan. Role of the universal stress protein uspa of salmonella in growth arrest, stress and virulence. *Microb Pathog*, 42(1):2–10, Jan 2007.

[257] J. Moskovitz, M. A. Rahman, J. Strassman, S. O. Yancey, S. R. Kushner, N. Brot, and H. Weissbach. Escherichia coli peptide methionine sulfoxide reductase gene: regulation of expression and role in protecting against oxidative damage. *J Bacteriol*, 177(3):502–507, Feb 1995.

[258] Richard L Cross and Volker Müller. The evolution of a-, f-, and v-type atp synthases and atpases: reversals in function and changes in the h+/atp coupling ratio. *FEBS Lett*, 576(1-2):1–4, Oct 2004.

[259] J. P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, and T. Oshima. Evolution of the vacuolar h+-atpase: implications for the origin of eukaryotes. *Proc Natl Acad Sci U S A*, 86(17):6661–6665, Sep 1989.

[260] S. F. Göthel and M. A. Marahiel. Peptidyl-prolyl cis-trans isomerases, a superfamily of ubiquitous folding catalysts. *Cell Mol Life Sci*, 55(3):423–436, Mar 1999.

[261] A. J. Heilpern and M. K. Waldor. Ctxphi infection of vibrio cholerae requires the tolqra gene products. *J Bacteriol*, 182(6):1739–1747, Mar 2000.

[262] Matthew A Gerding, Yasuyuki Ogata, Nicole D Pecora, Hironori Niki, and Piet A J de Boer. The trans-envelope tol-pal complex is part of the cell division machinery and required for proper outer-membrane invagination during cell constriction in e. coli. *Mol Microbiol*, 63(4):1008–1025, Feb 2007.

[263] Jean-Claude Lazzaroni, Jean-François Dubuisson, and Anne Vianney. The tol proteins of escherichia coli and their involvement in the translocation of group a colicins. *Biochimie*, 84(5-6):391–397, 2002.

[264] M. M. Muller and R. E. Webster. Characterization of the tol-pal and cyd region of escherichia coli k-12: transcript analysis and identification of two new proteins encoded by the cyd operon. *J Bacteriol*, 179(6):2077–2080, Mar 1997.

[265] E. M. Click and R. E. Webster. Filamentous phage infection: required interactions with the tola protein. *J Bacteriol*, 179(20):6464–6471, Oct 1997.

[266] Meghan E Chafee, Daniel J Funk, Richard G Harrison, and Seth R Bordenstein. Lateral phage transfer in obligate intracellular bacteria (wolbachia): verification from natural populations. *Mol Biol Evol*, 27(3):501–505, Mar 2010.

[267] J. P. Claverys. A new family of high-affinity abc manganese and zinc permeases. *Res Microbiol*, 152(3-4):231–243, 2001.

[268] Ekaterina M Panina, Andrey A Mironov, and Mikhail S Gelfand. Comparative genomics of bacterial zinc regulons: enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. *Proc Natl Acad Sci U S A*, 100(17):9912–9917, Aug 2003.

[269] E. D. Weinberg. The role of iron in protozoan and fungal infectious diseases. *J Eukaryot Microbiol*, 46(3):231–238, 1999.

[270] Serena Ammendola, Paolo Pasquali, Claudia Pistoia, Paola Petrucci, Patrizia Petrarca, Giuseppe Rotilio, and Andrea Battistoni. High-affinity zn2+ uptake system znuabc is required for bacterial zinc homeostasis in intracellular environments and contributes to the virulence of salmonella enterica. *Infect Immun*, 75(12):5867–5876, Dec 2007.

[271] K. Hantke. Bacterial zinc transporters and regulators. *Biometals*, 14(3-4):239–249, 2001.

[272] S. Wyllie and J. E. Raulston. Identifying regulators of transcription in an obligate intracellular pathogen: a metal-dependent repressor in chlamydia trachomatis. *Mol Microbiol*, 40(4):1027–1036, May 2001.

[273] G. Marcela Rodriguez and Issar Smith. Mechanisms of iron regulation in mycobacteria: role in physiology and virulence. *Mol Microbiol*, 47(6):1485–1494, Mar 2003.

[274] C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, R. D. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J. F. Tomb, B. A. Dougherty, K. F. Bott, P. C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. Hutchison, and J. C. Venter. The minimal gene complement of mycoplasma genitalium. *Science*, 270(5235):397–403, Oct 1995.

[275] A. R. Mushegian and E. V. Koonin. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A*, 93(19):10268–10273, Sep 1996.

[276] Hiroyuki Ogata, Bernard La Scola, Stéphane Audic, Patricia Renesto, Guillaume Blanc, Catherine Robert, Pierre-Edouard Fournier, Jean-Michel Claverie, and Didier Raoult. Genome sequence of rickettsia bellii illuminates the role of amoebae in gene exchanges between intracellular pathogens. *PLoS Genet*, 2(5):e76, May 2006.

[277] H. W. van Veen and W. N. Konings. The abc family of multidrug transporters in microorganisms. *Biochim Biophys Acta*, 1365(1-2):31–36, Jun 1998.

[278] B. Thoms and W. Wackernagel. Regulatory role of recf in the sos response of escherichia coli: impaired induction of sos genes by uv irradiation and nalidixic acid in a recf mutant. *J Bacteriol*, 169(4):1731–1736, Apr 1987.

[279] C. Janion. Some aspects of the sos response system–a critical survey. *Acta Biochim Pol*, 48(3):599–610, 2001.

[280] Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, and Amos Bairoch. Uniprotkb/swiss-prot: the manually annotated section of the uniprot knowledgebase. *Methods Mol Biol*, 406:89–112, 2007.

[281] K. D. Pruitt, T. Tatusova, and D. R. Maglott. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35(Database issue):D61–5, 2007. Journal Article Research Support, N.I.H., Intramural England.

[282]

[283] T. Rattei, P. Tischler, R. Arnold, F. Hamberger, J. Krebs, J. Krumsiek, B. Wachinger, V. Stumpflen, and W. Mewes. Simap–structuring the network of protein similarities. *Nucleic Acids Res*, 36(Database issue):D289–92, 2008. Journal Article Research Support, Non-U.S. Gov't England.

[284] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. Kegg for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–4, 2008. Journal Article Research Support, Non-U.S. Gov't England.

[285] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7, 1981. Journal Article England.

[286] Ahmed Moustafa. Jaligner: Open source java implementation of smith-waterman. Technical report, 2008.

[287] Emmanuelle Lerat, Vincent Daubin, Howard Ochman, and Nancy A Moran. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol*, 3(5):e130, May 2005.

[288] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292(2):195–202, Sep 1999.

[289] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997. LM05110/LM/United States NLM Journal Article Research Support, U.S. Gov't, P.H.S. Review England.

[290] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–8, 2007. Journal Article Research Support, Non-U.S. Gov't England.

[291] R. C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–7, 2004. Comparative Study Evaluation Studies Journal Article England.

[292] T. Schmidt and D. Frishman. Prompt: a protein mapping and comparison tool. *BMC Bioinformatics*, 7:331, 2006. Journal Article Research Support, Non-U.S. Gov't England.

[293] R Development Core Team. R: a language and environment for statistical computing. Technical report, Vienna, Austria: R Foundation for Statistical Computing, 2005.

[294] C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel, and P. Bork. String 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, 35(Database issue):D358–62, 2007. Journal Article Research Support, Non-U.S. Gov't England.

[295] Monica Vencato, Fang Tian, James R Alfano, C. Robin Buell, Samuel Cartinhour, Genevieve A DeClerck, David S Guttman, John Stavrinides, Vinita Joardar, Magdalen Lindeberg, Philip A Bronstein, John W Mansfield, Christopher R Myers, Alan Collmer, and David J Schneider. Bioinformatics-enabled identification of the hrpl regulon and type iii secretion system effector proteins of pseudomonas syringae pv. phaseolicola 1448a. *Mol Plant Microbe Interact*, 19(11):1193–1206, 2006.

[296] Boris A Vinatzer, Joanna Jelenska, and Jean T Greenberg. Bioinformatics correctly identifies many type iii secretion substrates in the plant pathogen pseudomonas syringae and the biocontrol isolate p. fluorescens sbw25. *Mol Plant Microbe Interact*, 18(8):877–888, 2005.

[297] Mark A Hall. *Correlation-based Feature Selection for Machine Learning.* PhD thesis, University of Waikato, 1998.

[298] Aha David and Kibler Dennis. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[299] S. le Cessie and J.C van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[300] Y Freund and R.E Schapire. Large margin classification using the perceptron algorithm. *11th Annual Conference on Computational Learning Theory, New York, NY*, pages 209–217, 1998.

[301] SS Keerthi, Shevade SK, C Bhattacharyya, and Murthy KRK. Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13(3):637–649, 2001.

[302] Jason D. Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. Tackling the poor assumptions of naive bayes text classifiers. *International Conference on Machine Learning*, pages 616–623, 2003.

[303] Andrew Mccallum and Kamal Nigam. A comparison of event models for naive bayes tex. *Classification. In: AAAI-98 Workshop on 'Learning for Text Categorization'*, 1998.

[304] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo*, pages 338–345, 1995.

[305] K. S. Ramamurthi and O. Schneewind. Substrate recognition by the yersinia type iii protein secretion machinery. *Mol Microbiol*, 50(4):1095–102, 2003. Journal Article Review England.

[306] H. Russmann, T. Kubori, J. Sauer, and J. E. Galan. Molecular and functional analysis of the type iii secretion signal of the salmonella enterica invj protein. *Mol Microbiol*, 46(3):769–79, 2002. AI30492/AI/United States NIAID Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. England.

[307] Kumaran S Ramamurthi and Olaf Schneewind. Yersinia enterocolitica type iii secretion: mutational analysis of the yopq secretion signal. *J Bacteriol*, 184(12):3321–3328, 2002.

[308] Uri Gophna, Eliora Z Ron, and Dan Graur. Bacterial type iii secretion systems are ancient and evolved by multiple horizontal-transfer events. *Gene*, 312:151–163, 2003.

[309] Laurence Rohmer, David S Guttman, and Jeffery L Dangl. Diverse evolutionary mechanisms shape the type iii effector virulence factor repertoire in the plant pathogen pseudomonas syringae. *Genetics*, 167(3):1341–1360, 2004.

[310] John Stavrinides, Wenbo Ma, and David S Guttman. Terminal reassortment drives the quantum evolution of type iii effectors in bacterial pathogens. *PLoS Pathogenes*, 2(10):e104, 2006.

[311] R. D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Res*, 34(Database issue):D247–51, 2006. United Kingdom Wellcome Trust Journal Article Research Support, Non-U.S. Gov't England.

[312] Nicola J Mulder, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Virginie Buillard, Lorenzo Cerutti, Richard Copley, Emmanuel Courcelle, Ujjwal Das, Louise Daugherty, Mark Dibley, Robert Finn, Wolfgang Fleischmann, Julian Gough, Daniel Haft, Nicolas Hulo, Sarah Hunter, Daniel Kahn, Alexander Kanapin, Anish Kejariwal, Alberto Labarga, Petra S Langendijk-Genevaux, David Lonsdale, Rodrigo Lopez, Ivica Letunic, Martin Madera, John Maslen, Craig McAnulla, Jennifer McDowall, Jaina Mistry, Alex Mitchell, Anastasia N Nikolskaya, Sandra Orchard, Christine Orengo, Robert Petryszak, Jeremy D Selengut, Christian J A Sigrist, Paul D Thomas, Franck Valentin, Derek Wilson, Cathy H Wu, and Corin Yeats. New developments in the interpro database. *Nucleic Acids Res*, 35(Database issue):D224–D228, Jan 2007.

[313] 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, Wolfgang Helmberg, Yuri Kapustin, David L Kenton, Oleg Khovayko, David J Lipman, Thomas L Madden, Donna R Maglott, James Ostell, Kim D Pruitt, Gregory D Schuler, Lynn M Schriml, Edwin Sequeira, Stephen T Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tugba O Suzek, Roman Tatusov, Tatiana A Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 34(Database issue):D173–D180, Jan 2006.

[314] Maike Tech and Rainer Merkl. Yacop: Enhanced gene prediction obtained by a combination of existing methods. *In Silico Biol*, 3(4):441–451, 2003.

[315] Luminita Badea, Scott A Beatson, Maria Kaparakis, Richard L Ferrero, and Elizabeth L Hartland. Secretion of flagellin by the lee-encoded type iii secretion system of enteropathogenic escherichia coli. *BMC Microbiol*, 9:30, 2009.

[316] Jeffrey E Christensen, Sophia A Pacheco, and Michael E Konkel. Identification of a campylobacter jejuni-secreted protein required for maximal invasion of host cells. *Mol Microbiol*, 73(4):650–662, Aug 2009.

[317] David Goudenège, Stéphane Avner, Céline Lucchetti-Miganeh, and Frédérique Barloy-Hubler. Cobaltdb: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources. *BMC Microbiol*, 10:88, 2010.

[318] Nancy Y Yu, James R Wagner, Matthew R Laird, Gabor Melli, Sébastien Rey, Raymond Lo, Phuong Dao, S. Cenk Sahinalp, Martin Ester, Leonard J Foster, and Fiona S L Brinkman. Psortb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13):1608–1615, Jul 2010.

[319] Eric Cascales and Peter J Christie. The versatile bacterial type iv secretion systems. *Nat Rev Microbiol*, 1(2):137–149, Nov 2003.

[320] G. Segal, J. J. Russo, and H. A. Shuman. Relationships between a new type iv secretion system and the icm/dot virulence system of legionella pneumophila. *Mol Microbiol*, 34(4):799–809, Nov 1999.

[321] D. M. Anderson and O. Schneewind. A mrna signal for the type iii secretion of yop proteins by yersinia enterocolitica. *Science*, 278(5340):1140–1143, Nov 1997.

[322] D. M. Anderson and O. Schneewind. Type iii machines of gram-negative pathogens: injecting virulence factors into host cells and more. *Curr Opin Microbiol*, 2(1):18–24, 1999.

[323] Rotem Sorek and Pascale Cossart. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet*, 11(1):9–16, Jan 2010.

6
Appendix and supplemental material

Set	Name	Proteins
P	Chlamydia muridarum Nigg	911
P	Chlamydia trachomatis A/HAR-13	919
P	Chlamydophila abortus S26/3	932
P	Chlamydophila caviae GPIC	1005
P	Chlamydophila pneumoniae AR39	1112
P	Chlamydophila pneumoniae CWL029	1052
P	Chlamydophila pneumoniae J138	1069
P	Chlamydophila pneumoniae TW-183	1113
P	Chlamydophila felis Fe/C-56	1013
P	Chlamydia trachomatis D/UW-3/CX	895
P	Chlamydia trachomatis 434/Bu	874
P	Chlamydia trachomatis L2b/UCH-1/proctitis	874
P	Chlamydia trachomatis B/TZ1A828/OT	880
P	Chlamydophila pneumoniae LPCoLN	1105
E	Candidatus Protochlamydia amoebophila UWE25	2030
E	Waddlia chondrophila	2049
E	Simkania negevensis	2500
E	Parachlamydia acanthamoebae UV7	2833

Table 6.1: *Chlamydiae* used in this work. 'Set' pathogen (P) or environmental (E), 'Name': the name with strain identifier, 'Proteins': amount of coding proteins as predicted by the gene-prediction software.

COG functional classifications

J	Translation, ribosomal structure and biogenesis
A	RNA processing and modification
K	Transcription
L	Replication, recombination and repair
B	Chromatin structure and dynamics
D	Cell cycle control, cell division, chromosome partitioning
Y	Nuclear structure
V	Defense mechanisms
T	Signal transduction mechanisms
M	Cell wall/membrane/envelope biogenesis
N	Cell motility
Z	Cytoskeleton
W	Extracellular structures
U	Intracellular trafficking, secretion, and vesicular transport
O	Posttranslational modification, protein turnover, chaperones
C	Energy production and conversion
G	Carbohydrate transport and metabolism
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
P	Inorganic ion transport and metabolism
Q	Secondary metabolites biosynthesis, transport and catabolism

Table 6.2: General categories from the COG scheme

Main categories of the funcat
01 Metabolism
02 Energy
04 Storage protein
10 Cell cycle and dna processing
11 Transcription
12 Protein synthesis
14 Protein fate (folding, modification, destination)
16 Protein with binding function or cofactor requirement
18 Protein activity regulation
20 Cellular transport, transport facilitation and transport routes
30 Cellular communication/signal transduction mechanism
32 Cell rescue, defense and virulence
34 Interaction with the cellular environment
36 Interaction with the environment (systemic)
38 Transposable elements, viral and plasmid proteins
40 Cell fate
41 Development (systemic)
42 Biogenesis of cellular components
43 Cell type differentiation
45 Tissue differentiation
47 Organ differentiation
70 Subcellular localization
73 Cell type localization
75 Tissue localization
77 Organ localization
78 Ubiquitous expression
98 Classification not yet clear-cut
99 Unclassified proteins

Table 6.3: Main categories of the MIPS FunCat

Categories of Swissprot keywords
Biological process
Cellular component
Coding sequence diversity
Developmental stage
Disease
Domain
Ligand
Molecular function
PTM (Post translational modification)
Technical term

Table 6.4: General categories used in the controlled vocabulary of Uniprot

Interactor I	Interactor II	Score
COG0178	COG0266	191.67482
COG0459	COG1126	191.67482
COG1124	COG0459	175.41096
COG2227	COG0266	175.41096
COG5563	COG1269	191.67482
COG0130	COG0459	202.87624
COG5563	GI:269302908	191.67482
GI:269302987	GI:269303304	191.67482
COG1131	COG0459	175.41096
COG5563	COG0045	191.67482
COG1136	COG0266	175.41096
GI:269303304	COG0045	191.67482
COG0178	COG0459	175.41096
env1641	COG0266	191.67482
GI:269303158	COG0098	191.67482
GI:269303158	COG0097	191.67482
COG2227	COG0459	191.67482
COG0814	COG0266	191.67482
COG0459	COG0577	175.41096
GI:269303304	GI:269302908	191.67482
COG5361	COG0266	191.67482
COG0459	COG1121	191.67482
COG5563	COG0582	191.67482
COG0097	GI:269303432	191.67482
COG5563	COG0515	191.67482
COG1136	COG0459	175.41096
env1641	COG0459	191.67482
COG5361	COG0459	191.67482
COG0814	COG0459	191.67482
GI:269303304	COG1640	191.67482
COG0459	COG0667	132.0155
COG1135	COG0459	191.67482
COG5563	GI:269303106	191.67482
GI:269303106	GI:269303304	191.67482
COG1127	COG0266	191.67482
COG0266	COG0444	175.41096
COG0658	GI:269303304	191.67482
COG1116	COG0266	175.41096
COG0488	COG0266	191.67482
GI:269303304	COG1269	191.67482
COG1127	COG0459	191.67482
COG5563	COG0658	191.67482
COG1116	COG0459	175.41096
GI:269303304	COG0582	191.67482
COG0577	COG0266	175.41096
GI:269303304	COG0515	191.67482
COG0488	COG0459	202.87624
COG0459	COG2274	175.41096
GI:269303428	GI:269303304	191.67482
COG1126	COG0266	191.67482
COG0443	COG2319	102.32367
COG0266	COG1137	191.67482
962855 960876	gsn.131	191.67482
COG5563	COG1294	191.67482
COG0396	COG0266	191.67482
COG0098	COG1137	191.67482
COG0459	COG0444	175.41096
COG3209	COG0266	191.67482
COG4608	COG0459	175.41096
COG0636	GI:269303304	191.67482
COG0444	COG0098	175.41096
COG1271	GI:269303304	191.67482
COG3209	COG0459	191.67482
COG5563	GI:269302987	191.67482
GI:269303432	COG0098	191.67482
COG1132	COG0266	175.41096
COG5563	COG0636	191.67482
COG5563	COG1640	191.67482
COG0459	COG1137	191.67482
COG2201	COG0266	191.67482
COG0459	COG0396	191.67482
COG0266	COG1121	191.67482
COG3842	COG0459	175.41096
COG5563	GI:269303428	191.67482
COG5563	COG1271	191.67482
COG2201	COG0459	191.67482
COG0130	COG0266	191.67482
COG0171	COG0388	95.54407
COG0459	COG1134	191.67482
COG0459	COG1132	120.81382
GI:269303304	COG1294	191.67482
COG0266	COG0667	175.41096

Continued . . .

Table 6.5: Predicted high probable physical interactions.

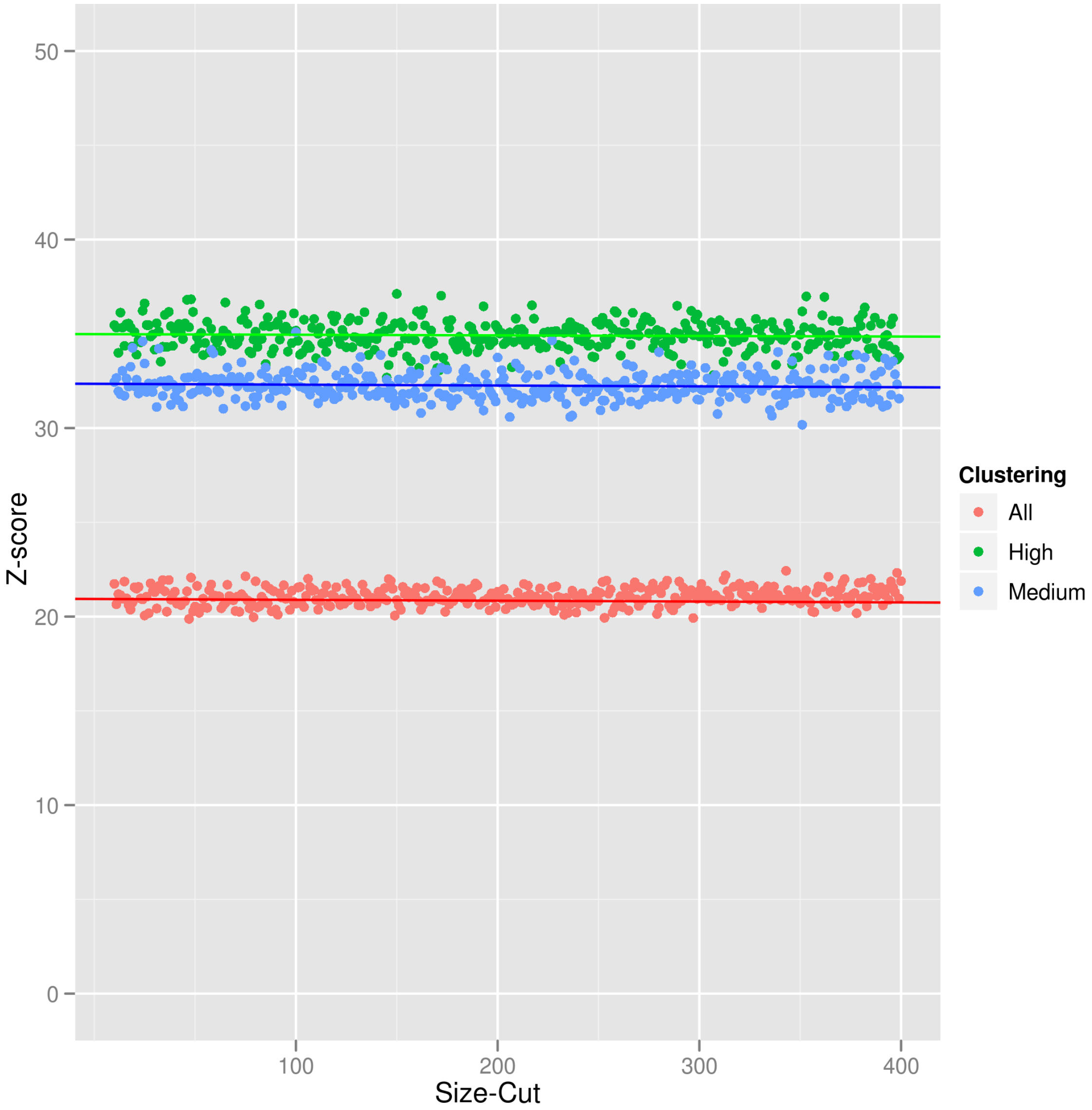


Figure 6.1: Z-scores resulted in the parameter exploration of runs with re-clustering by and varying Inflation parameter. Values for all three networks (High=high confidence, Medium=medium confidence , All=the complete network) shown. On the x-axis: the Z-score as defined by $|\frac{x-\mu}{\sigma}|$ of the cohesiveness measure, on the y-axis: value of the MCl inflation parameter.

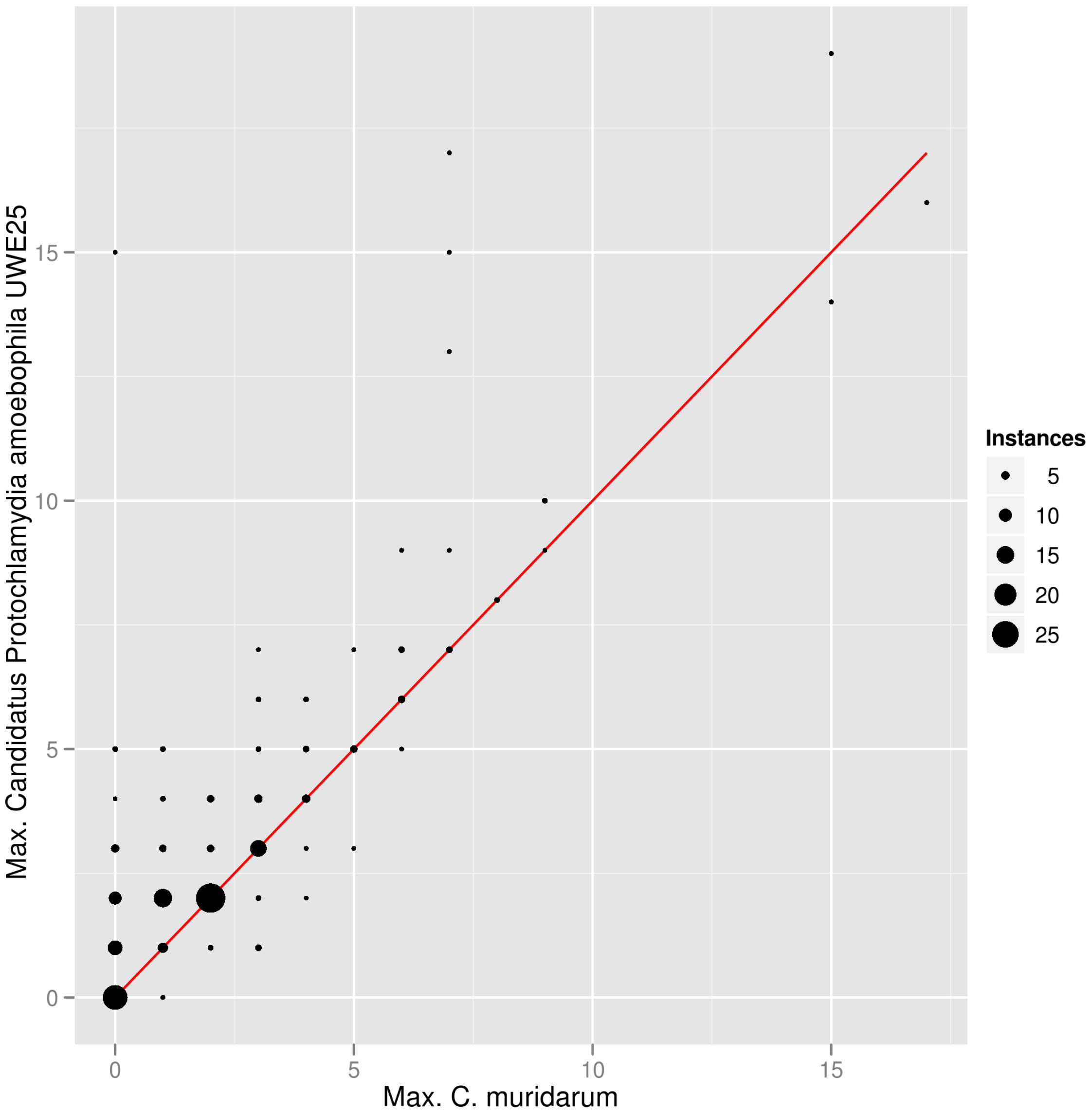


Figure 6.2: General trends in the evolutionary behavior of *Chlamydial* modules. The size of each module detected in the *Candidatus Protochlamydia amoebophila UWE25* is plotted against the maximum size detected in the pathogen *Chlamydia muridarum*, strain Nigg. The size of the points reflect the number of modules found at each data-point. The red line indicates the diagonal (x=y).

Organism	Run	Set	>0.5	<=0.5,>0	0
C. abortus S26/3	high	module	15	78	115
C. abortus S26/3	high	pathway	6	106	96
C. abortus S26/3	medium	module	17	82	136
C. abortus S26/3	medium	pathway	3	121	111
C. caviae GPIC	high	module	15	81	112
C. caviae GPIC	high	pathway	4	110	94
C. caviae GPIC	medium	module	16	86	133
C. caviae GPIC	medium	pathway	3	124	108
C. felis Fe/C-56	high	module	15	82	111
C. felis Fe/C-56	high	pathway	5	110	93
C. felis Fe/C-56	medium	module	17	86	132
C. felis Fe/C-56	medium	pathway	2	126	107
C. muridarum Nigg	high	module	14	80	114
C. muridarum Nigg	high	pathway	4	110	94
C. muridarum Nigg	medium	module	16	84	135
C. muridarum Nigg	medium	pathway	2	125	108
C. pneumoniae CWL029	high	module	14	80	114
C. pneumoniae CWL029	high	pathway	6	108	94
C. pneumoniae CWL029	medium	module	16	84	135
C. pneumoniae CWL029	medium	pathway	3	124	108
P. amoebophila UWE25	high	module	19	89	100
P. amoebophila UWE25	high	pathway	3	131	74
P. amoebophila UWE25	medium	module	18	97	120
P. amoebophila UWE25	medium	pathway	3	146	86
C. trachomatis A/HAR-13	high	module	14	79	115
C. trachomatis A/HAR-13	high	pathway	3	110	95
C. trachomatis A/HAR-13	medium	module	16	83	136
C. trachomatis A/HAR-13	medium	pathway	2	123	110
P. acanthamoebae UV7	high	module	20	98	90
P. acanthamoebae UV7	high	pathway	3	145	60
P. acanthamoebae UV7	medium	module	18	108	109
P. acanthamoebae UV7	medium	pathway	3	164	68
S. negevensis	high	module	16	95	97
S. negevensis	high	pathway	1	134	73
S. negevensis	medium	module	17	103	115
S. negevensis	medium	pathway	1	154	80
W. chondrophila	high	module	21	95	92
W. chondrophila	high	pathway	2	140	66
W. chondrophila	medium	module	20	99	116
W. chondrophila	medium	pathway	2	152	81

Table 6.6: Match statistics of modules to KEGG pathways and modules of representative chlamydial genomes. 'Organism' is the chlamydial genome, 'Set' describes to which KEGG entity (pathways or modules) the modules are compared, 'Run' is the module clustering (from medium or high confidence network).'>0.5' gives the count of modules with maximum Jaccard(Module,Kegg) between]0.5-1] ,'<=0.5,>0' is the count between [0-0.5[, '0' gives the amount of modules which do not match a KEGG entity at all.

Organism	Set	Run	JacK.	JacM.	Jac.	Jac. (rand)
C. trachomatis D/UW-3/CX	module	medium	0.15	0.26	0.19	0.08
C. muridarum Nigg	module	medium	0.15	0.26	0.19	0.09
C. abortus S26/3	module	medium	0.16	0.28	0.20	0.08
C. caviae GPIC	module	medium	0.17	0.28	0.21	0.09
C. pneumoniae AR39	module	medium	0.15	0.27	0.19	0.09
C. felis Fe/C-56	module	medium	0.16	0.28	0.21	0.09
P. amoebophila UWE25	module	medium	0.18	0.23	0.20	0.07
W. chondrophila	module	medium	0.18	0.22	0.20	0.09
S. negevensis	module	medium	0.15	0.20	0.17	0.08
P. acanthamoebae UV7	module	medium	0.17	0.20	0.18	0.08
C. trachomatis D/UW-3/CX	module	high	0.15	0.27	0.20	0.09
C. muridarum Nigg	module	high	0.15	0.27	0.19	0.08
C. abortus S26/3	module	high	0.16	0.29	0.21	0.08
C. caviae GPIC	module	high	0.17	0.29	0.22	0.08
C. pneumoniae AR39	module	high	0.15	0.28	0.20	0.08
C. felis Fe/C-56	module	high	0.17	0.29	0.21	0.09
P. amoebophila UWE25	module	high	0.18	0.24	0.21	0.07
W. chondrophila	module	high	0.19	0.24	0.21	0.09
S. negevensis	module	high	0.16	0.22	0.18	0.08
P. acanthamoebae UV7	module	high	0.17	0.22	0.19	0.08
C. trachomatis D/UW-3/CX	pathway	medium	0.10	0.17	0.13	0.06
C. muridarum Nigg	pathway	medium	0.10	0.17	0.13	0.06
C. abortus S26/3	pathway	medium	0.11	0.18	0.13	0.06
C. caviae GPIC	pathway	medium	0.11	0.18	0.14	0.06
C. pneumoniae AR39	pathway	medium	0.10	0.17	0.13	0.06
C. felis Fe/C-56	pathway	medium	0.11	0.18	0.13	0.07
P. amoebophila UWE25	pathway	medium	0.10	0.14	0.12	0.05
W. chondrophila	pathway	medium	0.10	0.13	0.11	0.06
S. negevensis	pathway	medium	0.09	0.12	0.11	0.05
P. acanthamoebae UV7	pathway	medium	0.09	0.12	0.10	0.05
C. trachomatis D/UW-3/CX	pathway	high	0.11	0.19	0.14	0.06
C. muridarum Nigg	pathway	high	0.11	0.19	0.14	0.06
C. abortus S26/3	pathway	high	0.12	0.19	0.14	0.06
C. caviae GPIC	pathway	high	0.12	0.19	0.15	0.07
C. pneumoniae AR39	pathway	high	0.11	0.19	0.14	0.06
C. felis Fe/C-56	pathway	high	0.12	0.19	0.14	0.07
P. amoebophila UWE25	pathway	high	0.11	0.15	0.13	0.05
W. chondrophila	pathway	high	0.10	0.14	0.12	0.06
S. negevensis	pathway	high	0.10	0.13	0.11	0.05
P. acanthamoebae UV7	pathway	high	0.10	0.14	0.11	0.05

Table 6.7: Recovery of KEGG pathways and modules. 'Organsim' is the chlamydial genome, 'Set' describes to which KEGG entity (pathways or modules) the modules are compared, 'Run' is the module clustering (from medium or high confidence network). 'JacK.' is the average of the Jaccard index of the best matching module/KEGG pairs normalized by the KEGG module/pathway sizes. 'JacM.' is the equivalent normalized by the module sizes, 'Jac.' their harmonic mean, 'Jac. (random)' is the harmonic mean in the randomized case (compare Song et al. [185]).

Module	KEGG entities	Descriptions	Amount
module 94	ko03030, ko03440	DNA replication, Homologous recombination	2
module 6	ko04112, ko00520, ko00471, ko00300, ko00550	Cell cycle - Caulobacter, Amino sugar and nucleotide sugar metabolism, D-Glutamine and D-glutamate metabolism, Lysine biosynthesis, Peptidoglycan biosynthesis	5
module 5	ko00900, ko00230, ko00240	Terpenoid backbone biosynthesis, Purine metabolism, Pyrimidine metabolism	3
module 3	ko00970, ko03010	Aminoacyl-tRNA biosynthesis, Ribosome	2
module 77	ko00564, ko03420	Glycerophospholipid metabolism, Nucleotide excision repair	2
module 2	ko00626, ko02020, ko00720	Naphthalene and anthracene degradation, Two-component system, Reductive carboxylate cycle (CO2 fixation)	3
module 71	ko00970, ko03010	Aminoacyl-tRNA biosynthesis, Ribosome	2
module 58	ko00280, ko00310, ko00010	Valine, leucine and isoleucine degradation, Lysine degradation, Glycolysis / Gluconeogenesis	3
module 37	ko00640, ko00720, ko00020	Propanoate metabolism, Reductive carboxylate cycle (CO2 fixation), Citrate cycle (TCA cycle)	3
module 35	ko00900, ko03010	Terpenoid backbone biosynthesis, Ribosome	2
module 29	ko03030, ko03010	DNA replication, Ribosome	2
module 28	ko00195, ko00190	Photosynthesis, Oxidative phosphorylation	2
module 21	ko00520, ko03018, ko00790	Amino sugar and nucleotide sugar metabolism, RNA degradation, Folate biosynthesis	3
module 13	ko02060, ko00260	Phosphotransferase system (PTS), Glycine, serine and threonine metabolism	2
module 12	ko00710, ko00010	Carbon fixation in photosynthetic organisms, Glycolysis / Gluconeogenesis	2

Table 6.8: Modules matching several KEGG pathways in *P. amoebophila* UWE25. 'Module': the module identifier, 'KEGG entities': the KEGG pathways that have overlap with the module, 'Descriptions': the descriptions of these KEGG pathways, 'Amount': the amount of KEGG pathways joined by the functional module.

KEGG entity	Description	Modules	Amount
ko00300	Lysine biosynthesis	module 179, module 6	2
ko00970	Aminoacyl-tRNA biosynthesis	module 56, module 50, module 3, module 71, module 59, module 124	6
ko00010	Glycolysis / Gluconeogenesis	module 58, module 12, module 96	3
ko00520	Amino sugar and nucleotide sugar metabolism	module 21, module 6	2
ko00260	Glycine, serine and threonine metabolism	module 13, module 25	2
ko03018	RNA degradation	module 14, module 21	2
ko00740	Riboflavin metabolism	module 40, module 44	2
ko00130	Ubiquinone and other terpenoid-quinone biosynthesis	module 118, module 8	2
ko03010	Ribosome	module 29, module 3, module 71, module 137, module 10, module 35, module 4	7
ko00190	Oxidative phosphorylation	module 18, module 24, module 67, module 195, module 26, module 28, module 52	7
ko03070	Bacterial secretion system	module 115, module 81, module 16	3
ko00900	Terpenoid backbone biosynthesis	module 5, module 168, module 35	3
ko00030	Pentose phosphate pathway	module 82, module 150	2
ko00860	Porphyrin and chlorophyll metabolism	module 140, module 162, module 90	3
ko00540	Lipopolysaccharide biosynthesis	module 88, module 42, module 1, module 74	4
ko04112	Cell cycle - Caulobacter	module 9, module 6	2
ko03030	DNA replication	module 29, module 94	2
ko00790	Folate biosynthesis	module 97, module 21	2
ko00500	Starch and sucrose metabolism	module 119, module 46	2
ko00564	Glycerophospholipid metabolism	module 77, module 178	2
ko03440	Homologous recombination	module 91, module 153, module 94, module 113	4
ko02010	ABC transporters	module 38, module 43, module 132, module 165, module 64, module 99, module 66, module 114	8
ko00720	Reductive carboxylate cycle (CO2 fixation)	module 37, module 2	2

Table 6.9: KEGG pathways matching several functional modules in *P. amoebophila* UWE25.

GI	mhc	mmcf	nhc	tnhc
GI:46446426	42.34	42.34	01	01
GI:46446796	42.34	42.34	01	01
GI:46445942	42.34	42.34	01	01
GI:46446833	16.21	16	16.21	01
GI:46446537	16.21	16	16.21	01
GI:46446536	16.21	16	16.21	01
GI:46446849	16.21	16	16.21	01
GI:46447036	16.21	16	16.21	01
GI:46446830	16.21	16	16.21	01
GI:46446183	01.06.01.01	01		01
GI:46445659	01.06.01.01	01	1	01
GI:46446576	01.06.01.01	01	1	01
GI:46446807	14.07.10	16	11.06.02	16
GI:46445991	16.03.10	12	16	16
GI:46446734	16.03.10	16.03.10	16.03.10	16.03.10
GI:46447502	16.03.10	16.03.10	16.03.10	01
GI:46445845	01.03.10	01	01.03.10	16
GI:46445844	01.03.10	01	01.03.10	16
GI:46446113	10.01.10	10.01.10	16	16
GI:46446111	10.01.10	10.01.10	01	01
GI:46447135	11.06.02	01	42.34	01
GI:46447621	11.06.02	01		01
GI:46445887	11.06.02			
GI:46445964	01.07.01	01		
GI:46445905	01.07.01	01	1	01
GI:46447522	32		01	01
GI:46447520	32		32	01
GI:46446923	32	01	1	01
GI:46447012	32	10.01.10	01	01
GI:46447523	32	32	01	01
GI:46447521	32		32	01
GI:46447204	32	10.01.10	01	01
GI:46446268	32	32	01	01
GI:46446964	32	10.01.10	01	01
GI:46447622	32	01.01.10	01.01.10	01
GI:46447369	32	01.01.10	01.01.10	01
GI:46445669	32	01		01
GI:46447013	32	10.01.10	01	01
GI:46446572	32	32	32	01
GI:46447495	32			
GI:46446852	32	32	01	01
GI:46446220	20	20	20	01
GI:46446450	20	20	01	01
GI:46446012	20	20	11.06.02	11.06.02
GI:46445738	20	20	20	16
GI:46447029	16	01	1	01
GI:46446106	16	16	16	16
GI:46446485	16	16		
GI:46445775	16	16	16	01
GI:46446667	16	16	16	16
GI:46446690	16	16	16	01
GI:46446832	16	16	01	01
GI:46447563	16	16	01	01
GI:46446777	16	16	01	01
GI:46447295	16	16	16	16
GI:46446329	16	16	16	16
GI:46445986	16	16	01	01
GI:46446604	16	16	16	16
GI:46446714	16	16	01	01
GI:46446314	16	16	16	16
GI:46447083	16	16	16	16
GI:46446271	16	16	01	01
GI:46446975	16	16	16	16
GI:46445973	16	16	16	16
GI:46446350	16	16	16	01
GI:46445895	16	16	01	01
GI:46446270	16	16	01	01
GI:46446666	16	16	16	16
GI:46447273	16	16	16	16
GI:46446347	16	16	16	16
GI:46447248	16	16	16	16
GI:46447025	16	16	16	01
GI:46445972	16	16	16	01
GI:46445716	16	16	16	01
GI:46445898	16	16	16	16
GI:46447606	16	16	16	16
GI:46446962	16	16	16	01
GI:46446221	16	16		20
GI:46446585	16	16	16.03.10	16
GI:46447333	16	01	16	01
GI:46446348	16	16	16	16
GI:46446344	16	16	16	16
GI:46446343	16	16	16	16
GI:46447245	16	16	16	16
GI:46445788	16	16	16	16
GI:46446345	16	16	16	16
GI:46447089	16	16	16	16
GI:46445672	16	16	16	16
GI:46446486	16	16		
GI:46446219	12		01	01
GI:46445827	12	12	16	01
GI:46446243	12		01	01
GI:46447259	12	12		01.03.10
GI:46446930	12			01
GI:46447190	2	2	2	2
GI:46446425	01	1		01
GI:46447490	01	1	01	01
GI:46445675	01	1	01	01
GI:46447267	01	1	01	01
GI:46446640	01	1	32	32
GI:46445769	01	1	01	01
GI:46445943	01	1	01	01
GI:46446751	01	1	01	01
GI:46445688	01	1	01	01
GI:46447474	01	1	01	01
GI:46446494	01	1	01.05.10	01
GI:46447616	01	1	01	01
GI:46446643	01	1	32	32
GI:46446480	01	1	01	01
GI:46446645	01	1	01	01
GI:46446020	01	1	32	32
GI:46446719	01	1	01	01
GI:46446024	01	1	32	32
GI:46447491	01	1	01	01
GI:46445700	01	1	01	01
GI:46446688	01	1	01	01
GI:46446989	01	1	32	32
GI:46445822	01	1	42.34	01
GI:46446591	01	1		11.06.02
GI:46447180	01	1	32	32
GI:46446103	01		12	01
GI:46446376	01	1	32	01
GI:46446188	01	1	01	01
GI:46447117	01	1		01
GI:46446387	01	1	11.06.02	11.06.02
GI:46445641	01	1	01	01

Continued . . .

GI	mhc	mmcf	nhc	tnhc
GI:46446273	01	1	01.05.10	01
GI:46446460	01	1	32	32
GI:46446966	01	1	01	01
GI:46447555	01	1	32	32
GI:46446148	01	1	01	01
GI:46447176	01	1	32	32
GI:46446955	01	1	32	32
GI:46447591	01	1	01	01
GI:46446226	01	1	20	20
GI:46446086	01	1	01	01
GI:46447578	01	1	32	32
GI:46447268	01		14.07.10	01
GI:46446701	01	1	01	01
GI:46447433	01	1	01	01
GI:46447656	01	1	32	32
GI:46446765	01	1	01	01
GI:46447463	01	1	01.07.01	16
GI:46446294	01	1	01	01
GI:46446842	01	1	32	32
GI:46446753	01	1	01	01
GI:46445983	01	1	01	01
GI:46447234	01	1	01	01
GI:46445978	01	1	10.01.10	01
GI:46446860	01	1	01	01
GI:46445676	01	1	01	01
GI:46447254	01	1	01	01
GI:46445888	01	1	01	01
GI:46446680	01	1	01	01
GI:46446792	01	1	01	01
GI:46447101	01	1	32	32
GI:46447252	01	1	32	32
GI:46445693	01	1	01	01
GI:46446891	01	1	42.34	01
GI:46445691	01	1	01	01
GI:46446282	01	1	01.07.01	16
GI:46446677	01	1		11.06.02
GI:46445690	01	1	32	32
GI:46447422	01	1	01	01
GI:46445685	01	1	01	01
GI:46446728	01	1	01	01
GI:46445692	01	1	01	01
GI:46446733	01	1	42.34	01
GI:46446757	01	1	01	01
GI:46445687	01	1	01	01
GI:46446747	01	1	01	01
GI:46447427	01	1		11.06.02
GI:46446085	01	1	01	01
GI:46445686	01	1	16	16
GI:46446762	01	1	32	32
GI:46445680	01	1	16	16
GI:46445678	01	1	01	01
GI:46445660	01	1		
GI:46447403	01	1	01.01.10	20
GI:46446763	01	1	32	32
GI:46445679	01	1	01	01
GI:46445674	01	1	01	01
GI:46446782	01	1	01	01
GI:46446754	01	1		01
GI:46446310	01	1	32	32
GI:46446541	01	1	42.34	01
GI:46446875		01		01
GI:46446873				01
GI:46446993		42.34		42.34

Continued . . .

GI	mhc	mmcf	nhc	tnhc
GI:46446318		2		2
GI:46446590				01
GI:46445859		01		01
GI:46447509		01		01
GI:46445910		01		01
GI:46446552		16		10
GI:46446995		42.34		42.34
GI:46446971		20		20
GI:46446351		01		16
GI:46446994		42.34		42.34
GI:46446322				01
GI:46445776		01		01
GI:46447619		16		16
GI:46446635				01
GI:46446615		20		01
GI:46446553		01		
GI:46446608		01		01
GI:46446134		01.01.10		01.01.10
GI:46447342				01
GI:46445877		01		01
GI:46446639			01	01
GI:46446386			16	12
GI:46447529			16	16
GI:46447279			01	01
GI:46446568		20		01
GI:46446638			01	01
GI:46447387		20	01	01
GI:46446317		2		2
GI:46446100			16	01
GI:46445733				01
GI:46446922		01		
GI:46447596				01
GI:46446495			16	01
GI:46446016				10.01.10
GI:46447361			01	01
GI:46446546		01		01
GI:46447271		01		01
GI:46447386		20	01	01
GI:46446161				01
GI:46446969		01		01
GI:46447158		01		01
GI:46447285			01	01
GI:46446945		01		01
GI:46446535				01
GI:46447516			16	01
GI:46445701		10	01	01
GI:46445869				01
GI:46446162			01	01
GI:46447195		01		01
GI:46445868			16	01
GI:46446914		01		01
GI:46447008		10.01.10		
GI:46446114		16		01
GI:46446913		01		01
GI:46447118		10.01.10		
GI:46446927				01
GI:46446900		16		16
GI:46446902		01		01
GI:46447446		16		10
GI:46446224		01		32
GI:46446644		01		32
GI:46447114			16	01
GI:46446340			16.01.10	01

Continued . . .

GI	mhc	mmcf	nhc	tnhc
GI:46446634			01	01
GI:46446810		01		01.06.01.01
GI:46447493			16	16
GI:46446457			12	12
GI:46447400			11.06.02	16
GI:46446252			01	01
GI:46447200		42.34		42.34
GI:46446207		01		
GI:46445814		01		42.34
GI:46445768		42.34		01
GI:46446218		01	1	01
GI:46446215		01		01
GI:46446621			01	01
GI:46446620			01	01
GI:46447517				16
GI:46447201		20		01
GI:46447344		01		16
GI:46446850			01	01
GI:46446907			10	10
GI:46446409		01		01
GI:46445999			01	01
GI:46445963		01		01
GI:46446469		01		01
GI:46446683		16		16
GI:46447651			10	10
GI:46447467		10.01.10		
GI:46446684		16		16
GI:46446607		20		20
GI:46447465		16		12
GI:46447638		01		01
GI:46447657			10	10
GI:46446513		01		01
GI:46446187		20		01
GI:46446783			32	01
GI:46447457		16		16
GI:46447451			20	01
GI:46446357		01		01
GI:46446413			10	10
GI:46447242		01		01
GI:46446901		01		
GI:46446279		01		01
GI:46447499		01		
GI:46445929		01		01
GI:46445652			10	10
GI:46446454		01		2
GI:46446689				01
GI:46447284			01	01
GI:46446280		01		01
GI:46447002			16	16
GI:46447597			01	01
GI:46447437		01		01
GI:46446678		01		

Table 6.10: Proposed functional categories for still uncharacterized proteins. 'GI' the genebank identifier, 'mhc' the FunCats resulting from the high confidence module based run, 'mcf' candidate annotations from the medium confidence modules, 'nhc' and 'nhc' the proposed FunCats by the respective network based approach. In each case, the selective approach of FunCat detection has been used.

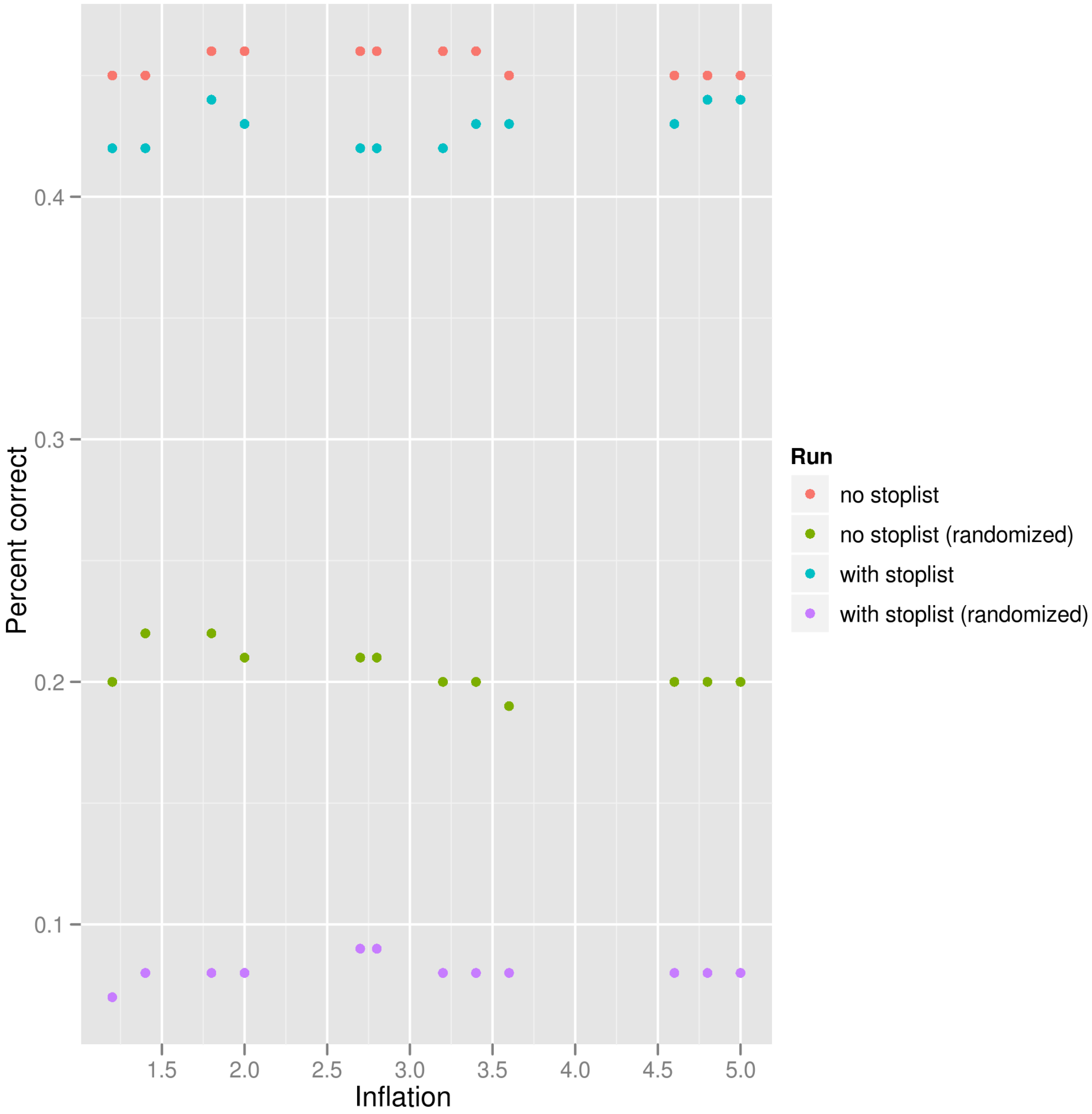


Figure 6.3: Performance of the selective function prediction for different clusterings (with reclustering all modules with size>=10 by Inflation parameter 5.0) with and without stop list (stoplist contains FunCat '01' and '16'). 'Inflation' is the initial inflation value used to cluster the set. 'Percent correct' is the rate of correct predictions. The different runs comprise: 'no stoplist: complete set of functional categories, with stoplist: categories 01 and 16 excluded. Both runs have been repeated with randomized FunCat labels, denoted as 'no stoplist (randomized)' and 'with stoplist (randomized)'.

FunCat identifier	Description	Amount transfers	Correct
01.01.06.01.01	biosynthesis of aspartate	1	0.00%
01.01.06.06.01.01	diaminopimelic acid pathway	1	100.00%
01.01.06.06.01	biosynthesis of lysine	1	100.00%
01.01.09.07.01	biosynthesis of histidine	1	0.00%
01.01.09	metabolism of the cysteine - aromatic group	3	100.00%
01.01	amino acid metabolism	4	75.00%
01.02.01.09.05	cyanate catabolism	1	0.00%
01.03	nucleotide metabolism	4	50.00%
01.05.01.03.02	polysaccharide biosynthesis	5	80.00%
01.05.01	C-compound and carbohydrate utilization	2	0.00%
01.05	C-compound and carbohydrate metabolism	5	40.00%
01.06.01.01	phospholipid biosynthesis	4	0.00%
01.06.01.07	isoprenoid biosynthesis	2	100.00%
01.06.01	lipid, fatty acid and isoprenoid biosynthesis	3	100.00%
01.06	lipid, fatty acid and isoprenoid metabolism	1	0.00%
01.07.01	biosynthesis of vitamins, cofactors, and prosthetic groups	9	33.33%
01.07	metabolism of vitamins, cofactors, and prosthetic groups	1	0.00%
01.20.19.01	biosynthesis of porphyrins	1	0.00%
01	METABOLISM	304	31.58%
02.07.01	pentose-phosphate pathway oxidative branch	1	0.00%
02.07	pentose-phosphate pathway	2	100.00%
02	ENERGY	22	72.73%
10.01.01	cellular DNA uptake	1	0.00%
10.01.03	DNA synthesis and replication	1	0.00%
10.01.05.01	DNA repair	2	50.00%
10.01	DNA processing	14	78.57%
10.03.03	cytokinesis (cell division) /septum formation	1	0.00%
10	CELL CYCLE AND DNA PROCESSING	6	66.67%
11.04	RNA processing	1	0.00%
11.06.02	tRNA modification	7	0.00%
11	TRANSCRIPTION	1	0.00%
12.04.02	translation elongation	3	0.00%
12.04.03	translation termination	2	0.00%
12.04	translation	1	100.00%
12	PROTEIN SYNTHESIS	61	68.85%
14.01	protein folding and stabilization	1	0.00%
14.07.03	modification by phosphorylation, dephosphorylation, autophosphorylation	2	50.00%
14.07	protein modification	6	33.33%
14.13	protein degradation	1	0.00%
14	PROTEIN FATE (folding, modification, destination)	1	0.00%
16.01	protein binding	5	80.00%
16.03.03	RNA binding	1	0.00%
16.03	nucleic acid binding	8	37.50%
16.19.03	ATP binding	2	0.00%
16.19.05	GTP binding	1	0.00%
16.19	nucleotide binding	1	0.00%
16.21.01	heme binding	1	0.00%
16.21	complex cofactor/cosubstrate binding	20	70.00%
16	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)	151	50.33%
20.01.01.01.01	heavy metal ion transport (Cu, Fe, etc.)	1	0.00%
20.01.01.07.05	sulfate transport	1	0.00%
20.03.22	transport ATPases	8	87.50%
20.03.25	ABC transporters	1	0.00%
20.03	transport facilitation	1	0.00%
20.09.16.03	Type III protein secretion system (virulence related secretory pathway, Gram- bacteria specific)	6	83.33%
20	CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES	44	81.82%
32.01.11	nutrient starvation response	1	0.00%
32.01	stress response	2	0.00%
32.05.01	resistance proteins	1	0.00%
32.05.05	virulence, disease factors	2	100.00%
32.07.07	oxygen and radical detoxification	1	0.00%
32.07.09	detoxification by degradation	1	0.00%
32	CELL RESCUE, DEFENSE AND VIRULENCE	20	15.00%
34.01.01.03	homeostasis of protons	2	100.00%
34.01	ionic homeostasis	1	100.00%
34	INTERACTION WITH THE CELLULAR ENVIRONMENT	1	0.00%
38	TRANSPOSABLE ELEMENTS, VIRAL AND PLASMID PROTEINS	1	0.00%
40.01.03	directional cell growth (morphogenesis)	1	0.00%
42.01	cell wall	1	0.00%
42.33	pilus/fimbria	2	100.00%
42.34	prokaryotic cell envelope structures	11	63.64%

Table 6.11: Functional categories transfered in the high confidence clustering. 'FunCat identifier' and 'Description' are the MIPS functional categories with their description, 'Amount transfers' is the number of protein to which the FunCat has been transfered during cross-validation, 'Correct' the percentage of correct transfers.

FunCat identifier	Description	Amount transfers	Correct
32.05	disease, virulence and defense	3	0.00%
32.01	stress response	3	0.00%
01.06.01	lipid, fatty acid and isoprenoid biosynthesis	2	0.00%
20.01.01.07.05	sulfate transport	1	0.00%
16.19.05	GTP binding	1	0.00%
10.01.03	DNA synthesis and replication	1	0.00%
16.19.03	ATP binding	2	0.00%
11.04	RNA processing	1	0.00%
20	CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES	21	4.76%
20.01.07	amino acid transport	1	0.00%
11.06.02	tRNA modification	1	0.00%
16.03	nucleic acid binding	1	0.00%
16.21.17	pyridoxal phosphate binding	2	0.00%
32.01.05	heat shock response	1	0.00%
20.09.16.02	Type II protein secretion system (general secretory pathway, exocytosis)	1	0.00%
16.21.11	thiamin pyrophosphate binding	3	0.00%
43.01.02.05	sporulation and other development of resting stage	1	0.00%
40.01.03	directional cell growth (morphogenesis)	1	0.00%
20.03.25	ABC transporters	1	0.00%
01.07	metabolism of vitamins, cofactors, and prosthetic groups	3	0.00%
14.01	protein folding and stabilization	1	0.00%
01.05	C-compound and carbohydrate metabolism	1	0.00%
16	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)	123	35.77%
01.03	nucleotide metabolism	2	0.00%
14	PROTEIN FATE (folding, modification, destination)	1	0.00%
01.01	amino acid metabolism	1	0.00%
20.03.01.01	ion channels	1	0.00%
12	PROTEIN SYNTHESIS	24	8.33%
20.03	transport facilitation	1	0.00%
11	TRANSCRIPTION	1	0.00%
10	CELL CYCLE AND DNA PROCESSING	5	0.00%
01.07.01	biosynthesis of vitamins, cofactors, and prosthetic groups	4	0.00%
01.05.01.03.02	polysaccharide biosynthesis	2	0.00%
16.21	complex cofactor/cosubstrate binding	5	0.00%
16.21.01	heme binding	1	0.00%
02.13.01	anaerobic respiration	1	0.00%
10.01.05.01	DNA repair	1	0.00%
32.07.09	detoxification by degradation	1	0.00%
34.01.01.03	homeostasis of protons	2	0.00%
32.07.05	detoxification by export	1	0.00%
42.34.07	peptidoglycan layer or other prokaryotic cell wall	1	0.00%
01.06.01.07	isoprenoid biosynthesis	1	0.00%
42.34	prokaryotic cell envelope structures	1	0.00%
42.33	pilus/fimbria	2	0.00%
10.01	DNA processing	5	0.00%
42.34.03	capsule and slime layer	1	0.00%
02	ENERGY	20	5.00%
42.34.01	bacterial outer membrane (only in Gram- bacteria)	4	0.00%
02.07	pentose-phosphate pathway	1	0.00%
01	METABOLISM	496	21.57%
32	CELL RESCUE, DEFENSE AND VIRULENCE	15	6.67%
16.19	nucleotide binding	5	20.00%
02.01	glycolysis and gluconeogenesis	1	0.00%
32.05.05.08	cytolysins	1	0.00%
32.05.01	resistance proteins	2	0.00%
16.03.03	RNA binding	2	0.00%
16.03.01	DNA binding	1	0.00%
01.01.09	metabolism of the cysteine - aromatic group	1	0.00%

Table 6.12: Functional categories transfered in the high confidence clustering after randomization. 'FunCat identifier' and 'Description' are the MIPS functional categories with their description, 'Amount transfers' is the number of protein to which the FunCat has been transfered during cross-validation, 'Correct' the percentage of correct transfers.

Locus Tag	Predicted FunCat module high conf.	Predicted FunCat module medium conf.	Predicted FunCat network high conf.	Predicted FunCat module medium conf.	manual classification	Comment	Classification
GI:46447202	42.34	42.34	01	01	42.34	still hypothetical, could be sugar epimerase and thus involved in cell wall biogenesis	++
GI:46446426	42.34	42.34	01	01	42.34.01	still hypothetical, seems to be an aminotransferase and involved in LPS biogenesis	++
GI:46446796	42.34	42.34	01	01	42.34, 16.21.07, 01.05.01.03.04, 01.05.01.03.04	NAD-binding epimerase, probably involved in sugar (EC 1.1.1.271) metabolism	+
GI:46445942	42.34	42.34	01	01	42.34.01	still hypothetical, seems to be an aminotransferase and involved in LPS biogenesis	++
GI:46446833	16.21	16	16.21	01	38.03, 16.03.01	transposase	-
GI:46446537	16.21	16	16.21	01	38.03, 16.03.01	transposase	-
GI:46446536	16.21	16	16.21	01	38.03, 16.03.01	transposase	-
GI:46446849	16.21	16	16.21	01	38.03, 16.03.01	transposase	-
GI:46447036	16.21	16	16.21	01	38.03, 16.03.01	transposase	-
GI:46446830	16.21	16	16.21	01	38.03, 16.03.01	transposase	-
GI:46446183	01.06.01.01	01	01	20		still hypothetical, has similarity to permeases (involved in any kind of transport)	-
GI:46445659	01.06.01.01	01	01	01	16.17., (32.05.01.01)	still hypothetical, with beta-lactamase domain, but this domain is found in many other enzymes than beta-lactamases	-
GI:46446576	01.06.01.01	01	01	01	16.17.	still hypothetical, with beta-lactamase and rhodanese domain, but this domains are found in many enzymes	-
GI:46446807	14.07.10	16	11.06.02	16		still hypothetical, well conserved in many species, but function completely unknown	
GI:46445991	16.03.10	12	16	16	16.03.01	still hypothetical, but with RNA-binding domain, can have many functions (all the predicted funcats - could also be possible)	+
GI:46446734	16.03.10	16.03.10	16.03.10	16.03.10	10.01.03, 10.01.05.01., 10.01.05.03., 16.01.	DNA polymerase III subunit d (holA, EC 2.7.7.7)	-
GI:46447502	16.03.10	16.03.10	16.03.10	1		still hypothetical, well conserved in many species, but function completely unknown	
GI:46445845	01.03.10	01	01.03.10	16	16.03.01, 10.01.	chromosome segregation protein scpA	-
GI:46445844	01.03.10	01	01.03.10	16	16.03.01, 10.01.	chromosome segregation protein scpB	-
GI:46446113	10.01.10	10.01.10	16	16	10.01.05.	still hypothetical, but well conserved and thought to play a role in DNA recombination	+
GI:46446903	10.01.10	10.01.10	01	01	10.01.05., 16.03.01, 32.01.09.	lexA repressor of SOS response	+
GI:46446111	10.01.10	10.01.10	01	01		still hypothetical, well conserved in many species, but function completely unknown	
GI:46445871	01.01.10	01.01.10	01.01.10	01.01.10	01.07.01.	Pyridoxal biosynthesis lyase pdxS	-
GI:46447135	11.06.02	01	42.34	01	(42.34.)	still hypothetical, could be cell-wall associated and thus play a role in invasion (very speculative+)	-
GI:46446094	11.06.02	11.06.02	01	01	14.07.11, 16.01.	still hypothetcal, any kind of protease	-
GI:46447621	11.06.02	01		01	12.01., 16.03.01., 16.01.	still hypothetical, could be tRNA-dihydrouridine synthase	-
GI:46445887	11.06.02					still hypothetical, but seems to be conserved in Chlamydiae and a few other spp.	
GI:46445964	01.07.01	01			(20.09.16.03, 32.05.05)	still hypothetical, predicted to be T3 secreted	-
GI:46445719	01.07.01	01.07.01	01.07.01	1	01.03.01	was already annotated as putative deoxyribonucleotide triphosphate pyrophosphatase	-
GI:46445905	01.07.01	01	01	01	01.07.01	Uroporphyrinogen-III synthase (hemD, 4.2.1.75)	+

Table 6.13: By hand curated FunCat predictions. 'Classification': rating of the prediction(s), '++' marks good coincidence of manual and predicted annotation, '+' marks mediocre coincidence, '-' wrong prediction. If no of these marks is given, an assessement is not possible. 'Locus Tag' : the locus tag of the protein, 'Predicted FunCat module high conf.','Predicted FunCat module medium conf.', 'Predicted FunCat network high conf.', and 'Predicted FunCat module medium conf.' contain the predictions of the different prediction runs, 'manual classification' contains the annotation by a biologist, unsure assignments in brackets (annotation done by Dr. Astrid Horn, University of Vienna), 'Comment': the biologist's comment on the annotation.

SWISSPROT-Identifier
YOPE_YERPS
YOPE_YEREN
YOPH_YERPS
YOPH_YEREN
A6M3V0_YERPE
Q56935_YERPS
YOPQ_YEREN
YOPJ_YERPS
Q93KQ5_YEREN
YOPM_YERPE
Q663L9_YERPS
A1JU68_YERE8
YOPT_YERPE
YOPT_YERPS
YOPT_YEREN
YOPT1_YEREN
YPKA_YERPE
YPKA_YERPS
Q56921_YEREN
O85239_YEREN
YSCH_YERPE
YSCH_YERPS
YSCH_YEREN
O34020_CHLCV
Q9Z8L4_CHLPN
Q824H6_CHLCV
TARP_CHLTR
Y572_CHLPN
Q46210_CHLCV
INCA1_CHLTR
Q9Z8Z8_CHLPN
O84235_CHLTR
Q9Z8P7_CHLPN
O30783_CHLCV
O84236_CHLTR
Q9Z8P6_CHLPN
INCD_CHLTR
INCE_CHLTR
INCF_CHLTR
INCG_CHLTR
Q9Z9F5_CHLPN
Q9Z7W9_CHLPN
SPAN_SALTY
SIPA_SALTY
B5C6I0_SALET
SIFA_SALTY
B5MXT4_SALET
SOPB_SALTY
SOPD_SALTY
SOPE_SALTY
SOPE2_SALTY
SPTP_SALTY
Q58I88_ECOLX
B7UM94_ECO27
B3BD94_ECO57
B3A274_ECO57
B2PHR1_ECO57
O85646_ECOLX
Q8X2D5_ECO57
A2A0X3_ECOLX
B3A307_ECO57
B3BS59_ECO57
AVRB_PSESG
HOPM1_PSESM
Q888Y7_PSESM
Q52537_PSESX
Q886L1_PSESM
Q88BF6_PSESM
Q889A9_PSESM
Q87V79_PSESM
Q882F0_PSESM
Q8RP03_PSEYM
Q888Y1_PSESM
Q87W07_PSESM
HRMA_PSESY
Q87WF7_PSESM
Q87X57_PSESM
Q87W42_PSESM
Q88A09_PSESM
Q881L7_PSESM
Q9K2L5_PSESH
Q87W46_PSESM
Q88AB8_PSESM
HOAE1_PSEU2
Q7PC42_PSEU2
Q52530_PSESH
Q9L6W4_PSESM
AVRP2_PSESJ
Q52394_PSESH
HPAB1_PSESH
Q888W0_PSESM
Q7PC45_PSEU2
AVRA_PSESG
Q52432_PSESX
AVRD1_PSESH
Q52389_PSESX
Q9JP32_PSESM
HOPAD_PSESM
Q87XS5_PSESM
Q9L6W3_PSESM

Continued ...

Table 6.14: Known effector sequences used in the training-step of EffectiveT3.

Min/max pattern	Behaviour, COG description	Core/Shell
Module 94	purifying	
0 1	COG0412 Dienelactone hydrolase and related enzymes	shell
1 1	COG0136 Aspartate-semialdehyde dehydrogenase	core I
1 1	COG0527 Aspartokinases	core I
Module 93	purifying	
0 1	COG0285 Folylpolyglutamate synthase	shell
1 1	COG0777 Acetyl-CoA carboxylase beta subunit	core I
1 1	COG0825 Acetyl-CoA carboxylase alpha subunit	core I
Module 83	cohesive	
1 1	COG0369 Sulfite reductase, alpha subunit (flavoprotein)	shell
0 0	COG2128 Uncharacterized conserved protein	shell
0 1	COG0626 Cystathionine beta-lyases/cystathionine gamma-synthases	core I
0 1	COG0031 Cysteine synthase	core I
Module 9	cohesive, purifying	
0 0	COG1052 Lactate dehydrogenase and related dehydrogenases	shell
0 1	COG2256 ATPase related to the helicase subunit of the Holliday junction resolvase	shell
0 1	COG0071 Molecular chaperone (small heat shock protein)	shell
1 1	COG0466 ATP-dependent Lon protease, bacterial type	shell
1 1	COG0172 Seryl-tRNA synthetase	shell
0 0	COG0339 Zn-dependent oligopeptidases	shell
1 1	COG0542 ATPases with chaperone activity, ATP-binding subunit	core I
1 1	COG0544 FKBP-type peptidyl-prolyl cis-trans isomerase (trigger factor)	core I
1 1	COG1219 ATP-dependent protease Clp, ATPase subunit	core I
1 1	COG0740 Protease subunit of ATP-dependent Clp proteases	core I
0 1	-	core II
0 1	-	core II
Module 8	cohesive	
1 1	COG0636 F0F1-type ATP synthase, subunit c/Archaeal/vacuolar-type H+-ATPase, subunit K	shell
1 1	COG0545 FKBP-type peptidyl-prolyl cis-trans isomerases 1	shell
1 1	COG1390 Archaeal/vacuolar-type H+-ATPase subunit E	shell
1 1	COG1269 Archaeal/vacuolar-type H+-ATPase subunit I	core I
1 1	COG1155 Archaeal/vacuolar-type H+-ATPase subunit A	core I
1 1	COG1394 Archaeal/vacuolar-type H+-ATPase subunit D	core I
1 1	COG1156 Archaeal/vacuolar-type H+-ATPase subunit B	core I
0 1	COG0711 F0F1-type ATP synthase, subunit b	core II
0 1	COG0055 F0F1-type ATP synthase, beta subunit	core II
0 1	COG0356 F0F1-type ATP synthase, subunit a	core II
0 1	COG0224 F0F1-type ATP synthase, gamma subunit	core II
0 1	COG0712 F0F1-type ATP synthase, delta subunit (mitochondrial oligomycin sensitivity protein)	core II
0 1	COG0056 F0F1-type ATP synthase, alpha subunit	core II
0 1	COG0355 F0F1-type ATP synthase, epsilon subunit (mitochondrial delta subunit)	core II
Module 3	cohesive, purifying	
1 1	COG0081 Ribosomal protein L1	shell
0 1	COG5272 Ubiquitin	shell
1 1	COG0244 Ribosomal protein L10	shell
0 0	COG0423 Glycyl-tRNA synthetase (class II)	shell
1 1	COG0250 Transcription antiterminator	shell
1 1	COG0199 Ribosomal protein S14	shell
1 1	COG0080 Ribosomal protein L11	shell
0 1	COG1597 Sphingosine kinase and enzymes related to eukaryotic diacylglycerol kinase	shell
0 1	COG0017 Aspartyl/asparaginyl-tRNA synthetases	shell
1 0	COG0690 Preprotein translocase subunit SecE	shell
0 1	COG0259 Pyridoxamine-phosphate oxidase	shell
1 1	COG0048 Ribosomal protein S12	core I
1 1	COG0201 Preprotein translocase subunit SecY	core I
1 1	COG0049 Ribosomal protein S7	core I
1 1	COG0480 Translation elongation factors (GTPases)	core I
1 1	COG0100 Ribosomal protein S11	core I
0 1	COG0222 Ribosomal protein L7/L12	core II
0 1	COG0267 Ribosomal protein L33	core II
1 1	COG0087 Ribosomal protein L3	core III
1 1	COG0162 Tyrosyl-tRNA synthetase	core III
Module 77	purifying	
0 1	COG0652 Peptidyl-prolyl cis-trans isomerase (rotamase) - cyclophilin family	shell
1 1	COG0215 Cysteinyl-tRNA synthetase	shell

Continued . . .

Min/max pattern	Behaviour, COG description	Core/Shell
1 1	COG0018 Arginyl-tRNA synthetase	core I
1 1	COG0008 Glutamyl- and glutaminyl-tRNA synthetases	core I
Module 2	irregular	
0 1	COG1596 Periplasmic protein involved in polysaccharide export	shell
0 1	COG2148 Sugar transferases involved in lipopolysaccharide synthesis	shell
0 1	COG2244 Membrane protein involved in the export of O-antigen and teichoic acid	shell
1 1	COG0561 Predicted hydrolases of the HAD superfamily	shell
0 1	COG1208 Nucleoside-diphosphate-sugar pyrophosphorylase involved in lipopolysaccharide biosynthesis/translation initiation factor 2B, gamma/epsilon subunits (eIF-2Bgamma/eIF-2Bepsilon)	shell
1 1	COG0110 Acetyltransferase (isoleucine patch superfamily)	shell
0 1	COG0381 UDP-N-acetylglucosamine 2-epimerase	shell
0 1	COG1040 Predicted amidophosphoribosyltransferases	shell
0 0	COG0639 Diadenosine tetraphosphatase and related serine/threonine protein phosphatases	shell
0 1	COG1215 Glycosyltransferases, probably involved in cell wall biogenesis	shell
0 1	COG0677 UDP-N-acetyl-D-mannosaminuronate dehydrogenase	shell
1 1	COG1132 ABC-type multidrug transport system, ATPase and permease components	shell
0 1	COG1086 Predicted nucleoside-diphosphate sugar epimerases	shell
0 1	COG1216 Predicted glycosyltransferases	shell
0 0	COG3307 Lipid A core - O-antigen ligase and related enzymes	shell
0 1	COG0859 ADP-heptose:LPS heptosyltransferase	shell
0 1	COG0367 Asparagine synthase (glutamine-hydrolyzing)	shell
0 0	COG1442 Lipopolysaccharide biosynthesis proteins, LPS:glycosyltransferases	shell
1 1	COG0717 Deoxycytidine deaminase	shell
0 1	COG0662 Mannose-6-phosphate isomerase	shell
1 0	COG0726 Predicted xylanase/chitin deacetylase	shell
0 1	COG1087 UDP-glucose 4-epimerase	shell
0 1	COG1807 4-amino-4-deoxy-L-arabinose transferase and related glycosyltransferases of PMT family	shell

Continued . . .

Min/max pattern	Behaviour, COG description	Core/Shell
0 1	COG0489 ATPases involved in chromosome partitioning	core I
0 1	COG0438 Glycosyltransferase	core I
0 1	COG0463 Glycosyltransferases involved in cell wall biogenesis	core I
1 1	COG0457 FOG: TPR repeat	core I
1 1	COG0500 SAM-dependent methyltransferases	core I
Module 76	purifying	
0 1	COG0062 Uncharacterized conserved protein	shell
1 1	COG0860 N-acetylmuramoyl-L-alanine amidase	shell
1 1	COG0802 Predicted ATPase or kinase	core I
1 1	COG1214 Inactive homolog of metal-dependent proteases, putative molecular chaperone	core I
Module 1	cohesive, irregular	
1 1	COG0477 Permeases of the major facilitator superfamily	shell
0 1	COG0524 Sugar kinases, ribokinase family	shell
1 0	COG2172 Anti-sigma regulatory factor (Ser/Thr protein kinase)	shell
1 1	COG1366 Anti-anti-sigma regulatory factor (antagonist of anti-sigma factor)	shell
0 1	COG0473 Isocitrate/isopropylmalate dehydrogenase	shell
1 1	COG2204 Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains	shell
0 1	COG0406 Fructose-2,6-bisphosphatase	shell
0 1	COG1566 Multidrug resistance efflux pump	shell
1 1	COG1651 Protein-disulfide isomerase	shell
1 0	COG1816 Adenosine deaminase	shell
1 1	COG0450 Peroxiredoxin	shell
0 1	COG0513 Superfamily II DNA and RNA helicases	shell
0 0	COG1695 Predicted transcriptional regulators	shell
1 1	COG0697 Permeases of the drug/metabolite transporter (DMT) superfamily	shell
1 1	COG0531 Amino acid transporters	shell
0 1	COG0667 Predicted oxidoreductases (related to aryl-alcohol dehydrogenases)	shell
0 1	COG3264 Small-conductance mechanosensitive channel	shell
0 1	COG1741 Pirin-related protein	shell
0 1	COG2259 Predicted membrane protein	shell
1 1	COG0642 Signal transduction histidine kinase	core I
1 1	COG0745 Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	core I

Continued . . .

Min/max pattern	Behaviour, COG description	Core/Shell
1 1	COG0265 Trypsin-like serine proteases, typically periplasmic, contain C-terminal PDZ domain	core I
1 1	COG0204 1-acyl-sn-glycerol-3-phosphate acyltransferase	core I
0 1	COG0583 Transcriptional regulator	core I
1 1	COG0178 Excinuclease ATPase subunit	core I
0 1	COG1048 Aconitase A	core II
0 1	COG0604 NADPH:quinone reductase and related Zn-dependent oxidoreductases	core II
0 1	COG2265 SAM-dependent methyltransferases related to tRNA (uracil-5-)-methyltransferase	core III
1 0	COG0664 cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases	core III
Module 68	purifying	
1 1	COG0441 Threonyl-tRNA synthetase	shell
0 1	COG0073 EMAP domain	shell
1 1	COG0291 Ribosomal protein L35	core I
1 1	COG0292 Ribosomal protein L20	core I
Module 66	cohesive	
1 1	COG2217 Cation transport ATPase	shell
0 1	COG2836 Uncharacterized conserved protein	shell
0 0	COG2608 Copper chaperone	shell
0 1	COG3278 Cbb3-type cytochrome oxidase, subunit 1	core I
0 1	COG2993 Cbb3-type cytochrome oxidase, cytochrome c subunit	core I
Module 64	purifying	
0 0	COG1858 Cytochrome c peroxidase	shell
0 1	COG1482 Phosphomannose isomerase	shell
1 1	COG0363 6-phosphogluconolactonase/Glucosamine-6-phosphate isomerase/deaminase	core I
1 1	COG0364 Glucose-6-phosphate 1-dehydrogenase	core I
1 1	COG0362 6-phosphogluconate dehydrogenase	core I
Module 63	purifying	
0 1	COG0735 Fe2+/Zn2+ uptake regulation proteins	shell
0 0	COG1321 Mn-dependent transcriptional regulator	shell
1 1	COG0803 ABC-type metal ion transport system, periplasmic component/surface adhesin	core I
1 1	COG1108 ABC-type Mn2+/Zn2+ transport systems, permease components	core I
1 1	COG1121 ABC-type Mn/Zn transport systems, ATPase component	core I
Module 60	irregular	
1 1	COG0012 Predicted GTPase, probable translation factor	shell
1 1	COG0193 Peptidyl-tRNA hydrolase	shell
0 1	COG0462 Phosphoribosylpyrophosphate synthetase	shell
0 1	COG1947 4-diphosphocytidyl-2C-methyl-D-erythritol 2-phosphate synthase	core I
1 1	COG1825 Ribosomal protein L25 (general stress protein Ctc)	core I
Module 56	purifying	
1 1	COG0327 Uncharacterized conserved protein	shell
0 1	COG0144 tRNA and rRNA cytosine-C5-methylases	shell
1 1	COG0223 Methionyl-tRNA formyltransferase	core I
1 1	COG1198 Primosomal protein N (replication factor Y) - superfamily II helicase	core I
1 1	COG0242 N-formylmethionyl-tRNA deformylase	core I
Module 54	purifying	
1 1	COG0009 Putative translation factor (SUA5)	shell
1 1	COG2890 Methylase of polypeptide chain release factors	shell
0 1	COG0613 Predicted metal-dependent phosphoesterases (PHP family)	shell
1 1	COG0216 Protein chain release factor A	core I
1 1	COG0254 Ribosomal protein L31	core I
Module 51	cohesive	
1 1	COG0762 Predicted integral membrane protein	shell
0 0	COG1957 Inosine-uridine nucleoside N-ribohydrolase	shell
1 1	COG0605 Superoxide dismutase	shell
0 1	COG0345 Pyrroline-5-carboxylate reductase	core I
0 1	COG0325 Predicted enzyme with a TIM-barrel fold	core I
Module 50	purifying	
1 1	COG0016 Phenylalanyl-tRNA synthetase alpha subunit	shell
1 1	COG0072 Phenylalanyl-tRNA synthetase beta subunit	shell
0 1	COG0789 Predicted transcriptional regulators	shell
1 1	COG0776 Bacterial nucleoid DNA-binding protein	core I
1 1	COG0290 Translation initiation factor 3 (IF-3)	core I
Module 47	irregular, cohesive	
0 1	COG0753 Catalase	shell
0 1	COG2032 Cu/Zn superoxide dismutase	core I

Continued ...

Min/max pattern	Behaviour, COG description	Core/Shell
1 1	COG1830 DhnA-type fructose-1,6-bisphosphate aldolase and related enzymes	core I
0 1	COG1376 Uncharacterized protein conserved in bacteria	core I
0 1	COG0380 Trehalose-6-phosphate synthase	core II
0 1	COG1877 Trehalose-6-phosphatase	core II
Module 45	irregular	
1 1	COG0039 Malate/lactate dehydrogenases	shell
0 1	COG2079 Uncharacterized protein involved in propionate catabolism	shell
0 1	COG2513 PEP phosphonomutase and related enzymes	shell
1 1	COG0045 Succinyl-CoA synthetase, beta subunit	core I
0 1	COG0372 Citrate synthase	core I
1 1	COG0074 Succinyl-CoA synthetase, alpha subunit	core I
Module 43	irregular	
0 1	COG2001 Uncharacterized protein conserved in bacteria	shell
1 1	COG0472 UDP-N-acetylmuramyl pentapeptide phosphotransferase/UDP-N-acetylglucosamine-1-phosphate transferase	shell
1 1	COG0769 UDP-N-acetylmuramyl tripeptide synthase	core I
1 1	COG0770 UDP-N-acetylmuramyl pentapeptide synthase	core I
1 1	COG0772 Bacterial cell division membrane protein	core I
0 1	COG1181 D-alanine-D-alanine ligase and related ATP-grasp enzymes	core I
Module 38	purifying	
1 1	COG0112 Glycine/serine hydroxymethyltransferase	shell
1 1	COG1327 Predicted transcriptional regulator, consists of a Zn-ribbon and ATP-cone domains	shell
0 1	COG0384 Predicted epimerase, PhzC/PhzF homolog	shell
1 1	COG0590 Cytosine/adenosine deaminases	shell
1 1	COG0117 Pyrimidine deaminase	core I
1 1	COG0108 3,4-dihydroxy-2-butanone 4-phosphate synthase	core I
1 1	COG0307 Riboflavin synthase alpha chain	core I
Module 35	purifying	
0 1	COG3203 Outer membrane protein (porin)	shell
0 1	COG0824 Predicted thioesterase	shell
0 0	COG3064 Membrane protein involved in colicin uptake	shell

Continued . . .

Min/max pattern	Behaviour, COG description	Core/Shell
1 1	COG2885 Outer membrane protein and related peptidoglycan-associated (lipo)proteins	core I
1 1	COG0823 Periplasmic component of the Tol biopolymer transport system	core I
1 1	COG0848 Biopolymer transport protein	core I
1 1	COG0811 Biopolymer transport proteins	core I
Module 34	irregular	
1 1	COG1137 ABC-type (unclassified) transport system, ATPase component	shell
0 1	COG1994 Zn-dependent proteases	shell
1 1	COG1413 FOG: HEAT repeat	shell
1 1	COG0442 Prolyl-tRNA synthetase	shell
0 1	COG0516 IMP dehydrogenase/GMP reductase	core I
0 1	COG0518 GMP synthase - Glutamine amidotransferase domain	core I
1 1	COG0517 FOG: CBS domain	core I
Module 31	purifying	
1 1	COG0828 Ribosomal protein S21	shell
1 1	COG0533 Metal-dependent proteases with possible chaperone activity	shell
0 0	COG1610 Uncharacterized conserved protein	shell
0 1	COG2518 Protein-L-isoaspartate carboxylmethyltransferase	shell
0 1	COG2312 Erythromycin esterase homolog	shell
1 1	COG0358 DNA primase (bacterial type)	core I
1 1	COG0568 DNA-directed RNA polymerase, sigma subunit (sigma70/sigma32)	core I
Module 27	purifying	
1 1	COG1523 Type II secretory pathway, pullulanase PulA and related glycosidases	shell
1 1	COG1640 4-alpha-glucanotransferase	shell
1 1	COG0448 ADP-glucose pyrophosphorylase	shell
1 1	COG0297 Glycogen synthase	shell
0 1	COG3280 Maltooligosyl trehalose synthase	shell
0 0	COG0366 Glycosidases	shell
1 1	COG0058 Glucan phosphorylase	core I
1 1	COG0296 1,4-alpha-glucan branching enzyme	core I
Module 26	irregular	
1 1	COG0665 Glycine/D-amino acid oxidases (deaminating)	shell
0 0	COG0352 Thiamine monophosphate synthase	shell
0 0	COG2145 Hydroxyethylthiazole kinase, sugar kinase family	shell

Continued . . .

Min/max pattern	Behaviour, COG description	Core/Shell
1 0	COG1060 Thiamine biosynthesis enzyme ThiH and related uncharacterized enzymes	shell
0 1	COG0404 Glycine cleavage system T protein (aminomethyltransferase)	core I
1 1	COG0509 Glycine cleavage system H protein (lipoate-binding)	core I
0 1	COG0403 Glycine cleavage system protein P (pyridoxal-binding), N-terminal domain	core I
0 1	COG1003 Glycine cleavage system protein P (pyridoxal-binding), C-terminal domain	core I
Module 25	purifying	
1 1	COG4232 Thiol:disulfide interchange protein	shell
1 1	COG0526 Thiol-disulfide isomerase and thioredoxins	shell
1 1	COG0492 Thioredoxin reductase	shell
1 1	COG0623 Enoyl-[acyl-carrier-protein] reductase (NADH)	shell
0 1	-	shell
0 1	COG0755 ABC-type transport system involved in cytochrome c biogenesis, permease component	shell
1 1	COG1674 DNA segregation ATPase FtsK/SpoIIIE and related proteins	core I
1 1	COG1158 Transcription termination factor	core I
Module 24	purifying	
0 1	COG4886 Leucine-rich repeat (LRR) protein	shell
0 0	COG2114 Adenylate cyclase, family 3 (some proteins contain HAMP domain)	shell
1 1	COG1716 FOG: FHA domain	shell
0 0	COG1262 Uncharacterized conserved protein	shell
0 1	COG0666 FOG: Ankyrin repeat	shell
0 1	COG2319 FOG: WD40 repeat	shell
1 1	COG0631 Serine/threonine protein phosphatase	core I
1 1	COG0515 Serine/threonine protein kinase	core I
Module 23	cohesive, purifying	
0 1	COG0266 Formamidopyrimidine-DNA glycosylase	shell
1 1	COG0234 Co-chaperonin GroES (HSP10)	shell
1 1	COG1448 Aspartate/tyrosine/aromatic aminotransferase	shell
0 1	COG0302 GTP cyclohydrolase I	shell
0 1	-	core I
0 1	-	core I
1 1	COG0459 Chaperonin GroEL (HSP60 family)	core II
1 1	COG0436 Aspartate/tyrosine/aromatic aminotransferase	core II
Module 21	cohesive, purifying	
1 1	COG0231 Translation elongation factor P (EF-P)/translation initiation factor 5A (eIF-5A)	shell
0 1	COG4799 Acetyl-CoA carboxylase, carboxyltransferase component (subunits alpha and beta)	shell
0 1	COG1509 Lysine 2,3-aminomutase	shell
1 1	COG0511 Biotin carboxyl carrier protein	core I
1 1	COG0439 Biotin carboxylase	core I
0 1	COG1385 Uncharacterized protein conserved in bacteria	core II
0 1	COG2264 Ribosomal protein L11 methylase	core II
0 1	COG1217 Predicted membrane GTPase involved in stress response	core II
Module 19	purifying	
0 1	COG0790 FOG: TPR repeat, SEL1 subfamily	shell
0 0	COG0402 Cytosine deaminase and related metal-dependent hydrolases	shell
1 1	COG1450 Type II secretory pathway, component PulD	shell
1 1	COG0237 Dephospho-CoA kinase	shell
1 1	COG0258 5-3 exonuclease (including N-terminal domain of PolI)	shell
0 1	COG2814 Arabinose efflux permease	shell
1 1	COG2804 Type II secretory pathway, ATPase PulE/Tfp pilus assembly pathway, ATPase PilB	core I
1 1	COG2165 Type II secretory pathway, pseudopilin PulG	core I
1 1	COG1459 Type II secretory pathway, component PulF	core I
Module 18	irregular (does not fit perfectly)	
1 1	COG1624 Uncharacterized conserved protein	shell
1 1	COG0037 Predicted ATPase of the PP-loop superfamily implicated in cell cycle control	shell
1 1	COG0465 ATP-dependent Zn proteases	core I
1 1	COG0617 tRNA nucleotidyltransferase/poly(A) polymerase	core I
1 1	COG1109 Phosphomannomutase	core II
1 1	COG0449 Glucosamine 6-phosphate synthetase, contains amidotransferase and phosphosugar isomerase domains	core II
0 0	COG1539 Dihydroneopterin aldolase	core III
0 1	COG0801 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase	core III

Continued ...

Min/max pattern	Behaviour, COG description	Core/Shell
1 1	COG0294 Dihydropteroate synthase and related enzymes	core III
Module 17	purifying	
1 0	COG0333 Ribosomal protein L32	shell
0 1	COG1853 Conserved protein/domain typically associated with flavoprotein oxygenases, DIM6/NTAB family	shell
0 1	COG0346 Lactoylglutathione lyase and related lyases	shell
0 1	COG2070 Dioxygenases related to 2-nitropropane dioxygenase	shell
1 1	COG1028 Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)	core I
1 1	COG0331 (acyl-carrier-protein) S-malonyltransferase	core I
1 1	COG0304 3-oxoacyl-(acyl-carrier-protein) synthase	core I
1 1	COG0236 Acyl carrier protein	core I
1 1	COG0332 3-oxoacyl-[acyl-carrier-protein] synthase III	core I
1 1	COG0416 Fatty acid/phospholipid biosynthesis enzyme	core I
Module 14	purifying	
1 1	COG0124 Histidyl-tRNA synthetase	shell
1 1	COG0821 Enzyme involved in the deoxyxylulose pathway of isoprenoid biosynthesis	shell
0 0	COG0647 Predicted sugar phosphatases of the HAD superfamily	shell
1 1	COG0105 Nucleoside diphosphate kinase	shell
1 1	COG1426 Uncharacterized protein conserved in bacteria	shell
0 1	COG0820 Predicted Fe-S-cluster redox enzyme	shell
0 1	COG1236 Predicted exonuclease of the beta-lactamase fold involved in RNA processing	shell
1 1	COG0217 Uncharacterized conserved protein	core I
1 1	COG1160 Predicted GTPases	core I
1 1	COG0173 Aspartyl-tRNA synthetase	core I
Module 11	purifying	
1 1	COG0504 CTP synthase (UTP-ammonia lyase)	shell
1 1	COG0469 Pyruvate kinase	shell
1 1	COG0205 6-phosphofructokinase	shell
0 1	COG1064 Zn-dependent alcohol dehydrogenases	shell
1 1	COG0588 Phosphoglycerate mutase 1	shell
0 1	COG3961 Pyruvate decarboxylase and related thiamine pyrophosphate-requiring enzymes	shell
1 1	COG2877 3-deoxy-D-manno-octulosonic acid (KDO) 8-phosphate synthase	shell
1 1	COG0190 5,10-methylene-tetrahydrofolate dehydrogenase/Methenyl tetrahydrofolate cyclohydrolase	shell
1 1	COG0126 3-phosphoglycerate kinase	core I
1 1	COG0148 Enolase	core I
1 1	COG0057 Glyceraldehyde-3-phosphate dehydrogenase/erythrose-4-phosphate dehydrogenase	core I
Module 10	purifying	
1 1	COG0203 Ribosomal protein L17	shell
0 1	COG0101 Pseudouridylate synthase	shell
1 1	COG0098 Ribosomal protein S5	shell
0 1	COG0657 Esterase/lipase	shell
1 1	COG0256 Ribosomal protein L18	core I
1 1	COG0198 Ribosomal protein L24	core I
1 1	COG0097 Ribosomal protein L6P/L9E	core I
1 1	COG0202 DNA-directed RNA polymerase, alpha subunit/40 kD subunit	core I
1 1	COG0099 Ribosomal protein S13	core I
1 1	COG0200 Ribosomal protein L15	core II
1 1	COG0024 Methionine aminopeptidase	core II

Continued ...

Table 6.15: Classification of reduced modules in the pathogenic *Chlamydiacea*. 'Min/max pattern': the appearance in the maximal environmental module (see text) and in the minimal pathogenic module (1=exist, 0 does not exist). 'COG description' COG identifier and description of the orthologous group the protein belongs to (in case of an environmental specific orthologous group not existent in COG, '-' is set). 'Core/shell': the classification of the module member in core and shell. Different cores are numbered by roman numerals.

Groups of co-evolving proteins	
Identifier	Effector or TTSS
Group I (Evidence exclusively by fusion events)	
YpkA	Effector
SopA	Effector
YopT	Effector
SspH2	Effector
Group II (Evidence by conserved neighbourhood for all interactions, supported by phylogenetic profile in some cases)	
EspD	Effector
Span	Effector
SopB	Effector
SipA	Effector
YscH	Effector
YscN	TTSS
YscL	TTSS
YscU	TTSS
YscC	TTSS
YscQ	TTSS
YscJ	TTSS
YscV	TTSS
YscR	TTSS
YscT	TTSS
YscU	TTSS
YscS	TTSS
YscB	TTSS/Chaperone
YscI	TTSS
YscF	TTSS
YscY	TTSS
Pairs of Effector and Chaperones	
YopN	SycN
YopT	SycT
CopN	SycE

Table 6.16: Groups of co-evolving effector and TTSS proteins and examples of co-localized effector proteins and chaperones based on the STRING database For each group of co-evolving effector and TTSS proteins, gene names of the members are given. The right column indicates, whether the orthologous group comprises effectors, TTSS proteins or TTSS related chaperones. A gene is added to a cluster, if the score of a genomic context method to another member derived from STRING exceeds 0.5. In the last section, examples of co-localized effectors and chaperones are listed.

Component	Description	P-Value
K03229	type III secretion protein SctU	2.11E-005
K03230	type III secretion protein SctV	3.60E-005
K03225	type III secretion protein SctQ	3.12E-004
K04058	type III secretion protein SctW	1.04E-002
K03220	type III secretion protein SctD	2.86E-002
K03227	type III secretion protein SctS	2.90E-002
K03228	type III secretion protein SctT	4.15E-002

Table 6.17: Adjusted p-Value for the enrichment of the KO within 30 neighbours up- and downstream of known effectors. Known effectors and their genomic neighbourhood to TTSS components. The genomic neighbourhood (30 genes up- and downstream) to TTSS components has been evaluated for all known effectors, except on Yersinia pestis KIM due to the absence of the plasmid pCD1 from the KEGG database. The number of effectors which are neighboured to at least one TTSS component is given in the middle column, the remaining effectors are summarized in the right column. 'Component': the TTSS component, 'Description' its' description, 'P-Value': the P-value of the effector enrichment found.

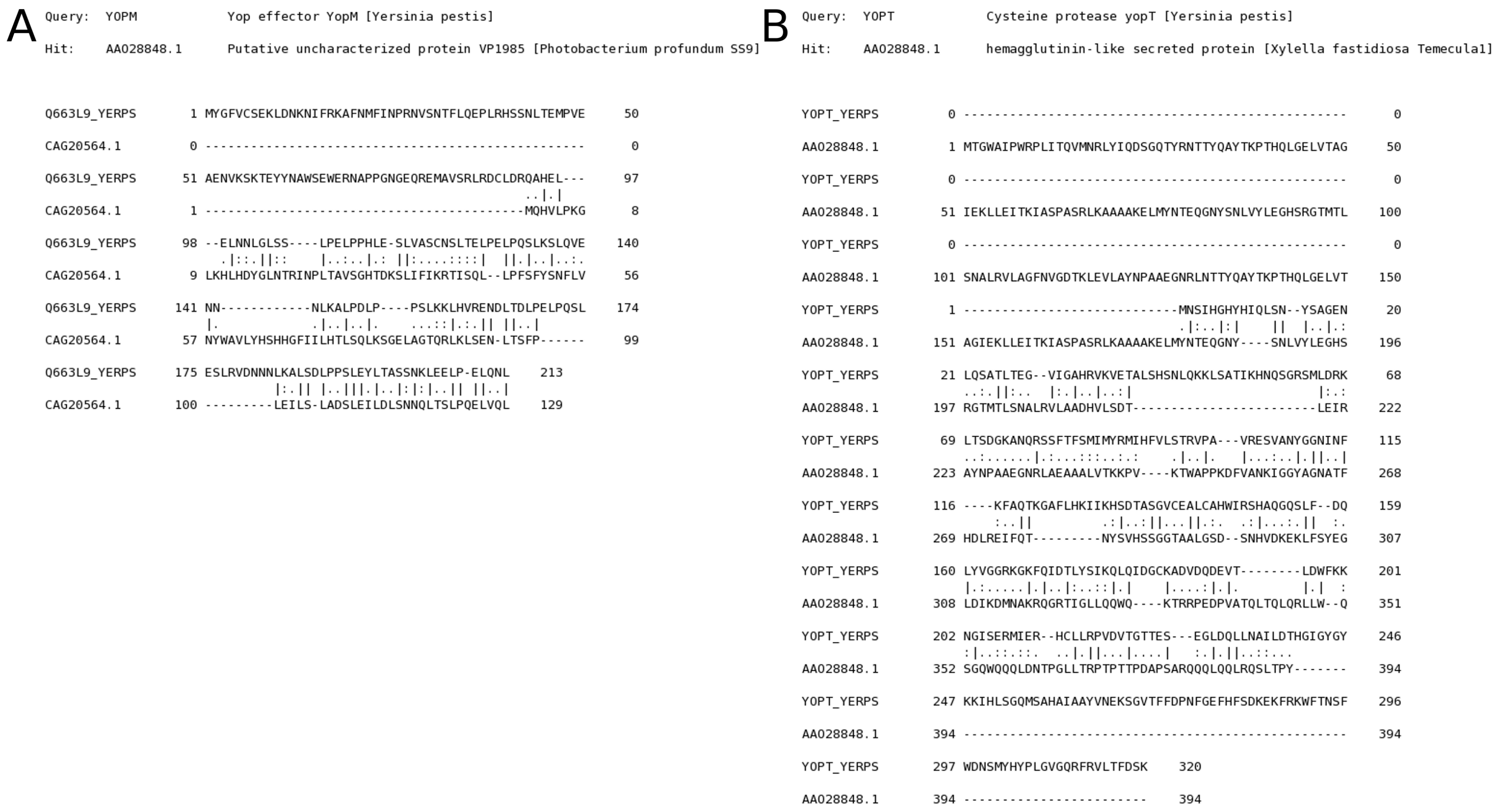
Genome	N1	N2
Chlamydophila caviae GPIC	2	3
Chlamydophila pneumoniae CWL029	2	5
Chlamydia trachomatis A/HAR-13 (serovar A)	2	0
Chlamydia trachomatis D/UW-3/CX (serovar D)	3	3
Escherichia coli O157:H7 EDL933 (EHEC)	5	2
Escherichia coli O127:H6 E2348/69	1	0
Escherichia coli O157:H7 Sakai (EHEC)	0	2
Pseudomonas syringae pv. syringae B728a	0	3
Pseudomonas syringae pv. phaseolicola 1448A	1	4
Pseudomonas syringae pv. tomato DC3000	5	19
Salmonella enterica subsp. enterica serovar Choleraesuis	0	1
Salmonella enterica subsp. enterica serovar Schwarzengrund CVM19633	0	1
Salmonella typhimurium LT2	3	5
Yersinia enterocolitica subsp. enterocolitica 8081	9	0
Yersinia pestis Antiqua	1	0
Yersinia pseudotuberculosis IP32953	8	0

Table 6.18: Enrichment of KEGG orthologous groups within the genomic neighbourhood of known effectors This table lists KEGG orthologous groups (KO), which are significantly enriched (Bonferroni-corrected t-Test p-Value < 0.05) within 30 neighbours up- and downstream of known effectors. 'Genome': the genome from which the effectors originate from, 'N1': the Number of effectors in the genomic neighbourhood of TTSS components, 'N2': Number of effectors not in the genomic neighbourhood of TTSS components

Feature
Amino acid alphabet
Alanine
Arginine
Asparagine
Aspartic Acid
Glutamic Acid
Glutamine
Glycine
Histidine
Isoleucine
Leucine
Lysine
Methionine
Phenylalanine
Proline
Serine
Serine-Leucine
Serine-Serine
Threonine
Threonine-Leucine
Tyrosine
Valine
Hydrophobic/hydrophilic alphabet
hydrophilic-hydrophilic
hydrophilic-hydrophilic-hydrophilic
hydrophilic-hydrophilic-hydrophobic
hydrophilic-hydrophobic-hydrophilic
hydrophilic-hydrophobic-hydrophobic
hydrophobic-hydrophilic-hydrophilic
hydrophobic-hydrophilic-hydrophobic
hydrophobic-hydrophobic-hydrophilic
Amino acid property alphabet
acidic
acidic-hydrophobic
alkaline
alkaline-alkaline
alkaline-hydrophilic
alkaline-hydrophobic
alkaline-hydrophobic-hydrophobic
alkaline-hydrophobic-polar
alkaline-polar
alkaline-polar-polar
aromatic
hydrophilic
hydrophilic-alkaline
hydrophilic-hydrophobic
hydrophilic-polar
hydrophobic
hydrophobic-acidic
hydrophobic-alkaline
hydrophobic-alkaline-hydrophobic
hydrophobic-alkaline-polar
hydrophobic-hydrophilic
hydrophobic-hydrophobic
hydrophobic-hydrophobic-hydrophobic
hydrophobic-hydrophobic-polar
hydrophobic-ionizable
hydrophobic-polar
hydrophobic-polar-hydrophobic
hydrophobic-polar-polar
ionizable
ionizable-polar
polar
polar-acidic
polar-alkaline
polar-alkaline-hydrophobic
polar-alkaline-polar
polar-hydrophilic
polar-hydrophilic-hydrophobic
polar-hydrophobic-hydrophobic
polar-hydrophobic-polar
polar-polar
polar-polar-alkaline
polar-polar-hydrophobic
polar-polar-polar

Continued ...

Table 6.19: Input features of the machine learning algorithms after initial feature selection. This table comprises these features, which are selected from all possible feature combinations using three different alphabets (amino acid alphabet, amino acid property alphabet, hydrophobic/hydrophilic alphabet) with a maximal pattern length of three. In order to avoid over-fitting on the data, only features are selected which are not specific to either the positive or the negative set but exists in both.



```
A  Query:  YOPM          Yop effector YopM [Yersinia pestis]
   Hit:    AA028848.1    Putative uncharacterized protein VP1985 [Photobacterium profundum SS9]

Q663L9_YERPS       1 MYGFVCSEKLDNKNIFRKAFNMFINPRNVSNTFLQEPLRHSSNLTEMPVE     50

CAG20564.1         0 --------------------------------------------------      0

Q663L9_YERPS      51 AENVKSKTEYYNAWSEWERNAPPGNGEQREMAVSRLRDCLDRQAHEL---     97
                                                               ..|.|
CAG20564.1         1 ------------------------------------------MQHVLPKG      8

Q663L9_YERPS      98 --ELNNLGLSS----LPELPPHLE-SLVASCNSLTELPELPQSLKSLQVE    140
                       .|::.||::    |..:..|.: ||:....::::|  ||.|..|..:.
CAG20564.1         9 LKHLHDYGLNTRINPLTAVSGHTDKSLIFIKRTISQL--LPFSFYSNFLV     56

Q663L9_YERPS     141 NN------------NLKALPDLP----PSLKKLHVRENDLTDLPELPQSL    174
                     |.            .|..|..|.    ...::|.:.|| ||..|
CAG20564.1        57 NYWAVLYHSHHGFIILHTLSQLKSGELAGTQRLKLSEN-LTSFP------     99

Q663L9_YERPS     175 ESLRVDNNNLKALSDLPPSLEYLTASSNKLEELP-ELQNL    213
                              |:.|| |..|||.|..|:|:|..|| ||..|
CAG20564.1       100 ---------LEILS-LADSLEILDLSNNQLTSLPQELVQL    129
```

```
B  Query:  YOPT          Cysteine protease yopT [Yersinia pestis]
   Hit:    AA028848.1    hemagglutinin-like secreted protein [Xylella fastidiosa Temecula1]

YOPT_YERPS         0 --------------------------------------------------      0

AA028848.1         1 MTGWAIPWRPLITQVMNRLYIQDSGQTYRNTTYQAYTKPTHQLGELVTAG     50

YOPT_YERPS         0 --------------------------------------------------      0

AA028848.1        51 IEKLLEITKIASPASRLKAAAAKELMYNTEQGNYSNLVYLEGHSRGTMTL    100

YOPT_YERPS         0 --------------------------------------------------      0

AA028848.1       101 SNALRVLAGFNVGDTKLEVLAYNPAAEGNRLNTTYQAYTKPTHQLGELVT    150

YOPT_YERPS         1 ----------------------------MNSIHGHYHIQLSN--YSAGEN     20
                                                 .|:..|:|    ||  |..|.:
AA028848.1       151 AGIEKLLEITKIASPASRLKAAAAKELMYNTEQGNY----SNLVYLEGHS    196

YOPT_YERPS        21 LQSATLTEG--VIGAHRVKVETALSHSNLQKKLSATIKHNQSGRSMLDRK     68
                     ..:.||:..  |:.|..|..:|                        |:.:
AA028848.1       197 RGTMTLSNALRVLAADHVLSDT------------------------LEIR    222

YOPT_YERPS        69 LTSDGKANQRSSFTFSMIMYRMIHFVLSTRVPA---VRESVANYGGNINF    115
                     ..:......|.:...:::..:.:    .|..|.   |...:..|.||..|
AA028848.1       223 AYNPAAEGNRLAEAAALVTKKPV----KTWAPPKDFVANKIGGYAGNATF    268

YOPT_YERPS       116 ----KFAQTKGAFLHKIIKHSDTASGVCEALCAHWIRSHAQGQSLF--DQ    159
                         :..||         .:|..:||...||.:.  .:|...:.||  :.
AA028848.1       269 HDLREIFQT---------NYSVHSSGGTAALGSD--SNHVDKEKLFSYEG    307

YOPT_YERPS       160 LYVGGRKGKFQIDTLYSIKQLQIDGCKADVDQDEVT--------LDWFKK    201
                     |.:.....|.|..|:..::|.|    |....:|.|.        |.|  :
AA028848.1       308 LDIKDMNAKRQGRTIGLLQQWQ----KTRRPEDPVATQLTQLQRLLW--Q    351

YOPT_YERPS       202 NGISERMIER--HCLLRPVDVTGTTES---EGLDQLLNAILDTHGIGYGY    246
                     :|..::.::.  ..|.||...|....|   :.|.||..::...
AA028848.1       352 SGQWQQQLDNTPGLLTRPTPTTPDAPSARQQQLQQLRQSLTPY-------    394

YOPT_YERPS       247 KKIHLSGQMSAHAIAAYVNEKSGVTFFDPNFGEFHFSDKEKFRKWFTNSF    296

AA028848.1       394 --------------------------------------------------    394

YOPT_YERPS       297 WDNSMYHYPLGVGQRFRVLTFDSK    320

AA028848.1       394 ------------------------    394
```

Figure 6.4: Example alignments between effector and non-effector orthologs. To investigate the evolutionary acquisition of the signal peptide, a pair wise sequence alignment study counting individual elongations and truncations between effectors and non-effector orthologs has been performed. This figure shows examples of these alignments. A) demonstrates elongation and B) truncation of effector proteins (upper row) aligned with sure non-effector proteins (lower row).

7

Publication Record

List of Publications

Arnold R, Jehl A, Rattei T.
Targeting effectors: the molecular recognition of Type III secreted proteins.
Microbes Infect. 2010 May;12(5):346-58. Epub 2010 Feb 21. Review.
PMID: 20178857

Rattei T, Tischler P, Götz S, Jehl MA, Hoser J, **Arnold R**, Conesa A, Mewes HW
SIMAP – a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters
Nucleic Acids Res. 2010 Jan;38(Database issue):D223-6. Epub 2009 Nov 1
PMID: 19906725

Schmitz-Esser S, Tischler P, **Arnold R**, Montanaro J, Wagner M, Rattei T, Horn M
The genome of the amoeba symbiont ‘Candidatus Amoebophilus asiaticus’ reveals common mechanisms for host cell interaction among amoeba-associated bacteria
J Bacteriol. 2010 Feb;192(4):1045-57. Epub 2009 Dec 18
PMID: 20023027

Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes HW, Horn M, Rattei T.

Sequence-based prediction of type III secreted proteins
Plos Pathogens 2009 Apr;5(4);
PMID: 19390696

Walter MC, Rattei T, **Arnold R**, Güldener U, Münsterkötter M, Nenova K, Kastenmüller G, Tischler P, Wölling A, Volz A, Pongratz N, Jost R, Mewes HW, Frishman D.
PEDANT covers all complete RefSeq genomes.
Nucleic Acids Res. 2009 Jan;37(Database issue):D408-11
PMID: 18940859

Loy A, **Arnold R**, Tischler P, Rattei T, Wagner M, Horn M.
probeCheck–a central resource for evaluating oligonucleotide probe coverage and specificity.
Environ Microbiol. 2008 Oct;10(10):2894-8;
PMID: 18647333

Rattei T, Tischler P, **Arnold R**, Hamberger F, Krebs J, Krumsiek J, Wachinger B, Stümpflen V, Mewes W.
SIMAP–structuring the network of protein similarities.
Nucleic Acids Res. 2008 Jan;36(Database issue):D289-92
PMID: 18037617

Krumsiek J, **Arnold R**, Rattei T.
Gepard: A rapid and sensitive tool for creating dotplots on genome scale.
Bioinformatics. 2007 Feb 19; PMID: 17309896

Rattei T, **Arnold R**, Tischler P, Lindner D, Stumpflen V, Mewes HW.
SIMAP: the similarity matrix of proteins.
Nucleic Acids Res. 2006 Jan 1;34(Database issue):D252-6.
PMID: 16381858

Rattei T, Walter M, **Arnold R**, Anderson DP, Mewes W.
Using public resource computing and systematic precalculation for large scale sequence analysis.
International Workshop Distributed, High-Performance and Grid Computing in Computational Biology (GCCB 2006) at the ECCB 2006
Lecture Notes in Bioinformatics

Arnold R, Rattei T, Tischler P, Truong MD, Stumpflen V, Mewes W.
SIMAP–The similarity matrix of proteins.
Bioinformatics. 2005 Sep 1;21 Suppl 2:ii42-ii46.
PMID: 16204123

Mewes HW, Amid C, **Arnold R**, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V, Warfsmann J, Ruepp A.
MIPS: analysis and annotation of proteins from whole genomes.
Nucleic Acids Res. 2004 Jan 1;32(Database issue):D41-4.
PMID: 14681354

Schoof H, Zaccaria P, Gundlach H, Lemcke K, Rudd S, Kolesov G, **Arnold R**, Mewes HW, Mayer KF.
MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome.
Nucleic Acids Res. 2002 Jan 1;30(1):91-3.
PMID: 11752263

Publications, submitted

Silvia Lang, Karl Gruber, Sanja Mihajlovic, **Roland Arnold**, Christian J. Gruber, Sonja Steinlechner Marc-Andre Jehl, Thomas Rattei , Kai-Uwe Fröhlich and Ellen L. Zechner
Molecular recognition determinants for type IV secretion of diverse
families of conjugative relaxases (submitted to Molecular Microbiology)

Jehl A, **Arnold R**, Rattei T
Effective- a database of predicted secreted bacterial proteins (submitted to NAR)

Götz S, **Arnold R**, Sebastián P, Tischler P, Jehl MA, Dopazo MA, Rattei T Conesa A
B2G-FAR, a species centered GO annotation repository (submitted)

Refereed Talks

Prediction and analysis of Type III secreted proteins
Sixth meeting of the European society for chlamydia research, Aarhus, Denmark
July 2008

Prediction of Type III secreted proteins
6. Deutscher Chlamydien-Workshop, Ulm, Germany, Februar 2008
(German conference on chlamydial research)

Prediction of Type III secreted proteins by their N-terminal amino-acid sequence
German Conference on Bioinformatics, Potsdam, Germany, September 2007

Software-Demo: The SimpAT Package: Integrating SIMAP into its own Applications
5th European Conference on Computational Biology, Eilat, Israel, January 2007