TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Comparative genomics of microbial genomes and development of a comprehensive chlamydiae genome database

Patrick Severin Tischler

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

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. W. Liebl
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr. H.-W. Mewes
	2. Univ.-Prof. Dr. M. Horn (Universität Wien/Österreich)

Die Dissertation wurde am 16.02.2011 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 05.10.2011 angenommen.

Acknowledgements

This is a great opportunity to thank the people who supported me during my PhD. Without their help, guidance and encouragement this study would not have been possible.

First of all I would like to thank my supervisor Univ.-Prof. Dr. Thomas Rattei for supporting me and my work. His valuable advices and constructive discussions helped me to grow personally as well as professionally. I want to express my gratitude that he supported me during the writing phase of the thesis even though he already was working in Vienna at that time.

I am very grateful to Prof. Dr. Hans-Werner Mewes for giving me the opportunity to acquire the PhD under his guidance at his institute. Prof. Mewes showed me that nothing makes sense in bioinformatics except in the light of biology.

I would like to thank Univ.-Prof. Dr. Matthias Horn for giving me the opportunity to develop a database for genomes of members of the phylum *Chlamydiae*. Thanks for the valuable and constructive discussions!

At this place I would like to thank my former colleagues at the Chair of Genome Oriented Bioinformatics at the Technische Universität München (TUM) and at the Munich Information Center for Protein Sequences (MIPS) at the Institute of Bioinformatics and Systems Biology (IBIS) at the Helmholtz Zentrum München. Thank you all for the fruitful and inspiring discussions. A special thank goes out to Roland Arnold who was an irreplaceable support during the whole time. Thomas Weinmaier supported me especially with the gene prediction. Thank you! André Jehl supported me with eukaryotic domains, and Philipp Pagel whenever I had statistical questions. I want to thank Mathias Walter for support with PEDANT, Gabi Kastenmüller for support with metabolic pathways, Volker Stümpflen, Richard Gregory, and Karamfilka Nenova for technical support with portal servers and Thorsten Barnickel for advices about text mining. I would also like to thank Andreas Kirschner, Angelika Fuchs, Martin Sturm, Nadia Latif, Sindy Neumann and all the others not explicitly mentioned here.

A special thank goes out to the office in Weihenstephan to Claudia, Frauke and Leonie who helped me with all non-scientific matters.

I would also like to thank the people that worked with me during their practical semesters or during internships. A special thank goes out to Franz Hamberger for his support with the stylesheets for the SIMAP portal, and to Bianca Weißbrich and Kyialbek Temirbaev for their support for the evaluation of the PhyloGenie software. I would also like to thank Christoph Loderer for his work on orthologous groups. Thank you also to everybody not especially mentioned here.

A large part of this work would not have been possible without the support from other groups and the projects together with other groups. Therefore I would like to thank all collaborators for their trust in me. Thank you to Angelika Lehner and Roger Stephan from the Institute for Food Safety and Hygiene, Vetsuisse Faculty University of Zürich, to Alexander Siegl, Astrid Horn, Barbara Sixt, Elena Tönshoff, Eva Heinz, Holger Daims, Roland Hatzenpichler, Sebastian Lücker, Silvia Knab, Stephan Schmitz-Esser and Michael Wagner from the Department of Microbial Ecology, University of

Vienna, to Anja Spang and Christa Schleper from the Department of Genetics in Ecology, University of Vienna, to Draženka Selesi, Franz Bergmann, Nidal Abu Laban and Rainer Meckenstock from the Institute of Groundwater Ecology, Helmholtz Zentrum München, and to Charles David from the Department of Cell and Developmental Biology, Ludwig-Maximillians-Universität München. And finally I would like to thank all the other partners not explicitly mentioned here.

I would like to thank all the users of the SIMAP portal and of ChlamydiaeDB for their feedback which is the basis for the success of these resources.

But I also had support outside of science. A very special thank goes to my girlfriend Gesine for her trust in me, for her constant support and for keeping me grounded. Thank you to my family who always supported me and thanks to all my friends.

Abstract

It is a long time since bacteria are investigated as they play important roles in biotechnology, disease management, biodefense and chronic diseases of humans. Nevertheless it was only feasible to analyze a small number of model organisms as *Escherichia coli* or *Bacillus subtilis* in depth in the laboratory as this research is time consuming and expensive. The knowledge gained in model organisms is transferred to novel organisms if significant sequence homology between the genetic elements can be detected. Therefore, with the availability of the sequences of many prokaryotic genomes due to novel sequencing technologies, the impact of comparative genomics on the microbiological research has grown over the last 15 years. Bioinformatics provides means to handle and compare the data and by that generates hypotheses that can be checked in experiments in the laboratory. As many bioinformatics analyses can only be conducted with considerable effort for every novel genome sequence by bioinformaticians, the automation of bioinformatics analyses is as essential as the preparation of data for non-bioinformaticians working in the laboratories.

Therefore the aims of this work were the automation and improvement of bioinformatics analysis methods for prokaryotic genomes, and to make the capability of comparative genomics easily available for non-bioinformaticians.

This work describes the results of collaborations with different scientists working on prokaryotes in the laboratory. Bioinformatics standard analyses like the search for best BLAST hits have been applied to the genome sequences of the organisms and comparative genomics has been especially successful adressing various biological issues. An example important for all prokaryotic genome projects is the improvement of the gene prediction in prokaryotic genomes. Different conflicts can occur as gene finders can predict different possible gene starts, and gene models can overlap. The automatic gene prediction pipeline ConsPred has been set up, that is able to resolve such conflicts in unambiguous cases by the integration of knowledge about conserved sequences in other organisms. This minimizes the manual effort necessary for the gene prediction in prokaryotic genomes. An example for a detailed description of a prokaryotic genome and thereby the characterization of a prokaryote is the comprehensive analysis of the genome of the Gram-negative opportunistic foodborne pathogen *Cronobacter turicensis* LMG 23827, that is known as rare but important cause of live-threatening neonatal infections. Several features could be identified that suggest an originally plant-associated lifestyle of *Cronobacter* spp. 44 potential horizontally transferred genes closely related to sequences in non-enterobacterial often plant-associated bacteria could be detected. Additionally pathways typical for plant-associated organisms could be identified. Supplementary it is already known, that *Cronobacter* spp. are generally capable to utilize a wide variety of compounds as a sole carbon source. Some of them such as L-arabinose, D-xylose, D-cellobiose and palatinose are known to be produced and potentially exudated by plants. The presence of a Type IV and a Type VI secretion system as well as the presence of an array of putative eukaryotic protein domains give a possible explanation for the potential of transferring DNA and effector proteins from the bacterial to host cells as a mechanism of interaction with a eukaryotic host. An example for prokaryotes, in which comparative genomics plays an very specific role, are *Chlamydiae*, obligate intracellular bacteria and major pathogens of humans. The ability to specifically inactivate and reactivate single genes is central to show gene functions, e.g. in knockout experiments. As it is not possible to genetically manipulate *Chlamydiae* bioinformatics and comparative genomics play an essential role in the research on *Chlamydiae*. In order to allow non-bioinformaticians to work with state-of-the-art bioinformatics methods and the available data of *Chlamydiae*, ChlamydiaeDB, a novel multi-genome database, was specifically developed for members of the phylum *Chlamydiae*. It facilitates the interactive analysis of all available chlamydial genomes by comprehensively integrating heterogeneous information from diverse sources in one place, structuring genes in clusters of orthologs, providing unique tools for comparative and functional genomics, and manual annotation possibilities. The available data comprises automatic annotations, as well as data from experiments, e.g. SNP and transcript data. A *Chlamydiae* specific textmining procedure has been developed. The possibility to retrieve all information about a group of proteins, feature enrichment in a group of proteins as well as a graphical KEGG pathway comparison make the resource a valuable tool for scientists working with genomic data of *Chlamydiae*. ChlamydiaeDB is easily maintainable and extensible (`http://www.chlamydiaedb.org`).

Zusammenfassung

Bakterien sind seit längerem Gegenstand der Forschung, weil sie wichtige Rollen in der Biotechnologie, im Disease Management, in der Biowaffenabwehr und bei chronischen Erkrankungen des Menschen spielen. Dennoch war es lediglich möglich eine geringe Anzahl von Modellorganismen wie *Escherichia coli* oder *Bacillus subtilis* eingehend im Labor zu untersuchen, da diese Forschung zeit- und kostenintensiv ist. Das Wissen, das in den Modellorganismen gewonnen wurde, wird auf neue Organismen übertragen, wenn signifikante Sequenzhomologie zwischen den genetischen Elementen gefunden werden kann. Daher wuchs mit der Verfügbarkeit von prokaryotischen Genomsequenzen aufgrund neuer Sequenzierungstechnologien im Laufe der vergangenen 15 Jahre der Einfluss von vergleichender Genomik in der mikrobiellen Forschung. Die Bioinformatik stellt die Mittel zu Umgang und Vergleich der dabei anfallenden Daten bereit und erzeugt dadurch Hypothesen, die in Experimenten im Labor überprüft werden können. Da viele Bioinformatikanalysen nur mit erheblichem Aufwand für jede neue Genomsequenz von Bioinformatikern durchgeführt werden können, ist die Automation von Bioinformatikanalysen genauso wichtig wie die Aufbereitung der Daten für Nicht-Bioinformatiker, die in den Laboren arbeiten.

Deshalb waren die Ziele dieser Arbeit die Automation und Verbesserung von bioinformatischen Analysemethoden für prokaryotische Genome, und die einfache Bereitstellung des Potentials von vergleichender Genomik für Nicht-Bioinformatiker.

Die vorliegende Arbeit beschreibt die Ergebnisse von Kollaborationen mit verschiedenen Wissenschaftlern, die sich im Labor mit Prokaryoten beschäftigen. Für die Genomsequenzen der Organismen wurden Standard-Bioinformatikanalysen wie die Suche nach den besten BLAST-Hits durchgeführt, wobei die vergleichende Genomik besonders bei der Beantwortung verschiedenster komplexer biologischer Fragestellungen erfolgreich war. Ein Beispiel für eine solche Fragestellung, die für alle prokaryotischen Genomprojekte wichtig ist, ist die Verbesserung der Genvorhersage in prokaryotischen Genomen. Bei der Genvorhersage können verschiedene Konflikte auftreten, weil die Vorhersageprogramme unterschiedliche Genstarts und überlappende Genmodelle vorhersagen können. Die automatische Genvorhersagepipeline ConsPred wurde implementiert, welche solche Konflikte in eindeutigen Fällen durch die Integration von Wissen um konservierte Sequenzen in anderen Organismen lösen kann. Dadurch wird der manuelle Aufwand für die Genvorhersage in prokaryotischen Genomen reduziert. Ein Beispiel fuer die detaillierte Beschreibung eines prokaryotischen Genoms und dadurch des Prokaryoten selbst, ist die umfassende Analyse des Gram-negativen opportunistischen durch Lebensmittel übertragbaren Pathogens *Cronobacter turicensis* LMG 23827, das als seltene aber wichtige Ursache von lebensbedrohlichen Infektionen bei Neugeborenen bekannt ist. Es konnten mehrere Merkmale gefunden werden, die auf eine möglicherweise pflanzenassoziierte Lebensweise hindeuten. 44 potentiell horizontal übertragene Gene nah verwandt zu Sequenzen in nicht-enterobakteriellen häufig pflanzenassoziierten Bakterien konnten gefunden werden. Zusätzlich wurden metabolische Pfade gefunden, die typisch für pflanzenassoziierte Organismen sind. Darüber hinaus ist bereits bekannt, dass *Cronobacter* Spezies im Allgemeinen dazu in der Lage sind ein breites Spektrum von

Komponenten als alleinige Kohlenstoffquelle zu nutzen. Von einigen dieser Komponenten wie L-Arabinose, D-Xylose, D-Cellobiose und Palatinose ist bekannt, dass sie von Pflanzen produziert und möglicherweise exsudiert werden. Das Vorhandensein eines Typ IV und eines Typ VI Sekretionssystems als auch das Vorhandensein einer Reihe von putativen eukaryotischen Proteindomänen geben eine mögliche Erklärung für das Potential DNA und Effektorproteine aus den Bakterien- in die Wirtszellen übertragen zu können, als möglicher Mechanismus zur Interaktion mit einem eukaryotischen Wirt. Ein Beispiel für Prokaryoten, für die vergleichende Genomik eine ganz besondere Rolle spielt, sind Chlamydien, obligat intrazelluläre Bakterien und bedeutende Pathogene des Menschen. Die Fähigkeit spezifisch einzelne Gene zu inaktivieren und zu reaktivieren ist zentral um die Funktion von Genen nachzuweisen, beispielsweise in Knockout-Experimenten. Da es nicht möglich ist Chlamydien genetisch zu manipulieren, spielen die Bioinformatik und vergleichende Genomik eine essentielle Rolle bei der Forschung an Chlamydien. Um Nicht-Bioinformatikern zu ermöglichen mit aktuellen Bioinformatikmethoden und den verfügbaren Daten von Chlamydien zu arbeiten, wurde ChlamydiaeDB, eine neuartige Multi-Genom-Datenbank speziell für Organismen des Phylum *Chlamydiae* entwickelt. Sie ermöglicht die interaktive Analyse aller verfügbaren Chlamydiengenome durch die umfassende Integration heterogener Informationen aus diversen Quellen an einem Ort, durch die Strukturierung von Genen in Cluster von Orthologen, durch die Bereitstellung einzigartiger Werkzeuge für vergleichende und funktionelle Genomik und die Möglichkeit zur manuellen Annotation. Die verfügbaren Daten umfassen automatische Annotationen sowie Daten aus Experimenten, z.B. SNP und Transkriptionsdaten. Außerdem wurde ein chlamydienspezifisches Textmining entwickelt. Die Möglichkeit des Downloads aller Informationen für eine Gruppe von Proteinen, das Feature Enrichment für eine Gruppe von Proteinen sowie ein grafischer KEGG Pathwayvergleich machen the Ressource zu einem wertvollen Werkzeug für Wissenschaftler, die mit genomischen Daten von Chlamydien arbeiten. Die ChlamydiaeDB ist einfach wartbar und erweiterbar (`http://www.chlamydiaedb.org`).

Contents

1 Introduction

The topics of this work are the analysis of prokaryotic genomes using bioinformatics and the development of a resource for genomic data of members of the phylum *Chlamydiae*. Therefore the necessary basics are described in this chapter.

1.1 Sequencing of bacterial genomes

In the last years the speed of conventional sequencing increased while the prices decreased (Figure 1.1). This made it possible even for relatively small institutions or individual research groups to sequence whole bacterial genomes.

Additionally the companies 454 [1] and Solexa [2] introduced “Next generation sequencing” that simplifies, speeds up and cheapens the DNA sequencing even more dramatically, however at the cost of reduction in read length. As an example using Illumina sequencing a finished base pair costs about 0.01 cent. This is about 100 times cheaper compared to about 0.01$ per finished base pair in 2005 (Figure 1.1).

The combination of next generation sequencing with the classically sequenced reference genomes in the public databases opens potential applications ranging from personalized genome-based medicine to microbial strain optimization and bioterrorism prevention. Especially the use of genomic information for the diagnosis of diseases will help to be able to provide tailor-made medicaments for patients, which is more effective for the patient on the one hand and costs less money on the other hand which is of advantage to the society.

1.2 Primary annotation of genome sequences

1.2.1 Motivation

The genome sequence of an organism consists of one or more chromosomes, each of them consisting of a long stretch of deoxyribonucleic acid (DNA). Such a DNA as provided by sequencing companies is just a text file with the letters A, T, C, G standing for the purines adenine (A) and guanine (G) and for the pyrimidines thymine (T) and cytosine (C). These sequences are incomprehensible and mainly useless unless meaningful biological facts are associated with them in the course of genome annotation.

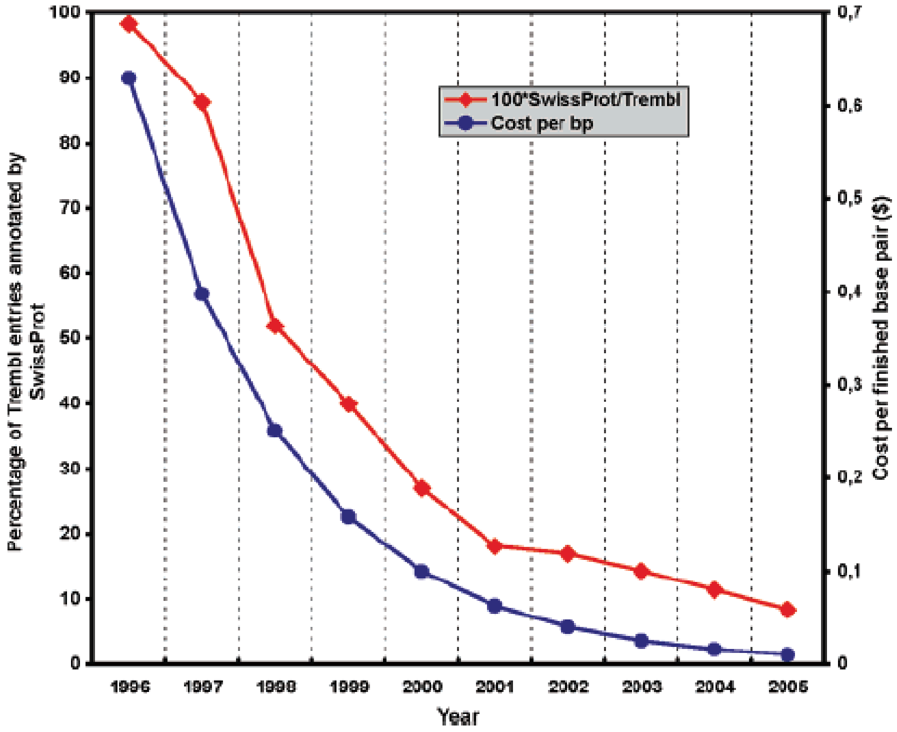


Figure 1.1: Costs per finished base pair Costs per finished base pair [3] and percentage of manually annotated sequences. The latter is estimated as the ratio of the number of sequences in the UniProtKB/Swiss-Prot database and all known sequences available in the UniProtKB/TrEMBL database. (adapted from [4])

Therefore one of the first goals of genome analysis for every newly sequenced genome is the detection of the location of genes and signals in it. This first step in genome analysis is crucial as it determines the quality of all subsequently applied methods. While false positive predictions result in the overestimation of the genomic repertoire of a genome by the prediction of proteins without homology to conserved protein sequences in other organisms, false negative predictions lead to missing genes or signals, e.g. to incomplete metabolic pathways that would wrongly suggest the existence of isoenzymes.

For the primary annotation of a genome sequence protein coding genes, genes of untranslated RNAs and signals of regulatory function have to be considered. The goal of the primary annotation of genome sequences is the determination of a preferably exact and complete catalog of locations in the genome. Each of these locations is assigned to one of the known genetic element types.

1.2.2 Prediction of coding sequences - gene prediction

1.2.2.1 Prokaryotic genes

The knowledge about protein coding genes plays an essential role for the understanding of the functions of an organism. Therefore the accurate prediction of protein coding

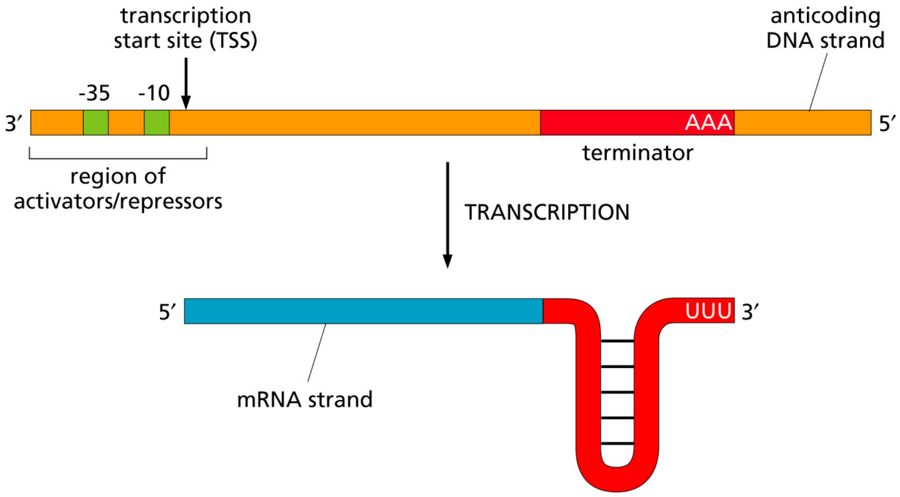


Figure 1.2: The start and stop signals for prokaryotic transcription The signals to start transcription are short nucleotide sequences that bind transcription enzymes. The signal to stop transcription is a short nucleotide sequence that forms a loop structure preventing the transcription apparatus from continuing. (Source: [16])

genes is an important step in the genome analysis.

A prokaryotic gene consists of a region coding for a protein and regulatory regions (Figure 1.2). The promoter region upstream of the coding region is available for almost all genes and is recognized by the transcription machinery when a gene is to be transcribed. Prokaryotic genes can have various promoters that can be activated or inactivated depending on the influence of regulatory factors like activators, enhancers or even DNA packaging. In a transcript sequencing experiment in *Chlamydia trachomatis* L2b several genes with two distinct transcription start sites could be identified which indicates a differential regulation of gene expression by variation in 5' UTR length [5]. Even though weak and strong promoters can be distinguished there exists no common strong consensus sequence. The promoter sequences are dependent on the dissociable subunit σ of the RNA polymerase. This subunit that plays an important role in the promoter recognition and confers promoter specificity [6].

But not every gene has its own promoter. Particularly with the availability of a large number of genomes it has been observed, that genes in microbial genomes tend to form clusters, which are conserved during evolution [7, 8, 9]. Members of gene clusters are often cotranscribed as operons [10], or coregulated as division of a biochemical network [11, 12, 13]. The coregulation is associated with similar or related function of the genes of an operon and is thereby biologically meaningful [8, 14, 15]. Accordingly an operon can be defined as a series of adjacent same-stranded unidirectional genes in a genome. The genes are transcribed into a single mRNA molecule sharing a common regulation. Operons share one common promoter and terminator for all the genes within the operon [11] (see Figure 1.3).

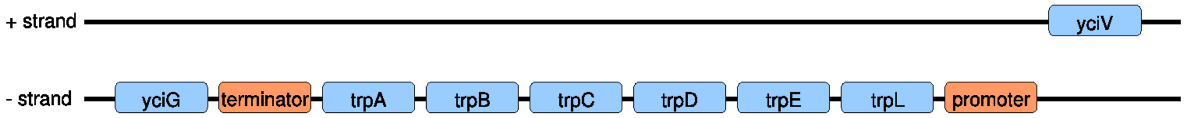


Figure 1.3: The trp operon and its neighboring genes in *Escherichia coli* K12 The trp operon consists of the genes trpL, trpE, trpD, trpC, trpB and trpA [17]. The intergenic distances between the genes within the trp operon are short. trpA and trpB as well as trpD and trpE even overlap. The reason that these short distances and overlaps are possible is that the genes of the trp operon share a common promoter in front of the operon.

1.2.2.2 Features usable for gene prediction

Generally intrinsic and extrinsic information can be used for the prediction of protein coding genes in genomic sequences. Intrinsic information is derived solely from the genome in that the gene prediction is performed, extrinsic information is information gained by comparisons to other genomes.

The following features of prokaryotic genes can be used for their prediction:

- **Length of open reading frames (ORFs)** In prokaryotic genomes the most important intrinsic information is the length of open reading frames (ORFs) that are significantly longer for coding sequences than in intergenic regions. As ORFs can be contained in all six possible reading frames on the genome (Figure 1.4) there is the possibility that long ORFs can be detected in parallel in different frames (Figure 1.5). As overlaps between genes are rare in prokaryotic genomes not all ORFs exceeding a specific length are true genes. This problem is essentially critical for short genes down to 120 nt length as typical e.g. for several ribosomal proteins. Therefore the detection of all ORFs within a DNA sequence is not sufficient for a reliable gene prediction.

- **DNA composition statistics** There are two reasons that the DNA composition of genes can be used for their detection. The first reason is that it has been shown that the frequency of single nucleotides and of oligonucleotides up to a length of six residues is significantly different in coding genes and intergenic regions [19, 20]. These frequencies are specific for every genome (Figure 1.6) and are constantly changing over evolutionary time. The second reason is that it is known, that synonymous codons do not occur with equal frequency in all organisms [21]. It is also well known that preferred codons tend to correspond to the tRNAs that have the highest concentrations in cells [22, 23]. Nevertheless codons with no corresponding tRNA can still be translated using wobble base pairing [24].

- **Promoter signals** Prokaryotic genes have three distinct recognition sequences: the -10 and -35 regions upstream of the translation start and a translation termination sequence 3' of the stop codon (Figure 1.2). The σ subunits of the RNA polymerase are important for transcriptional regulation: (i) They ensure the recognition of core promoter elements, (ii) They position the RNA polymerase at the

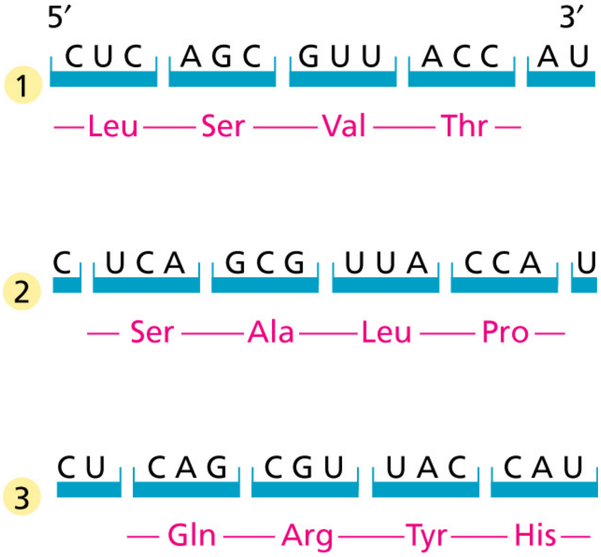
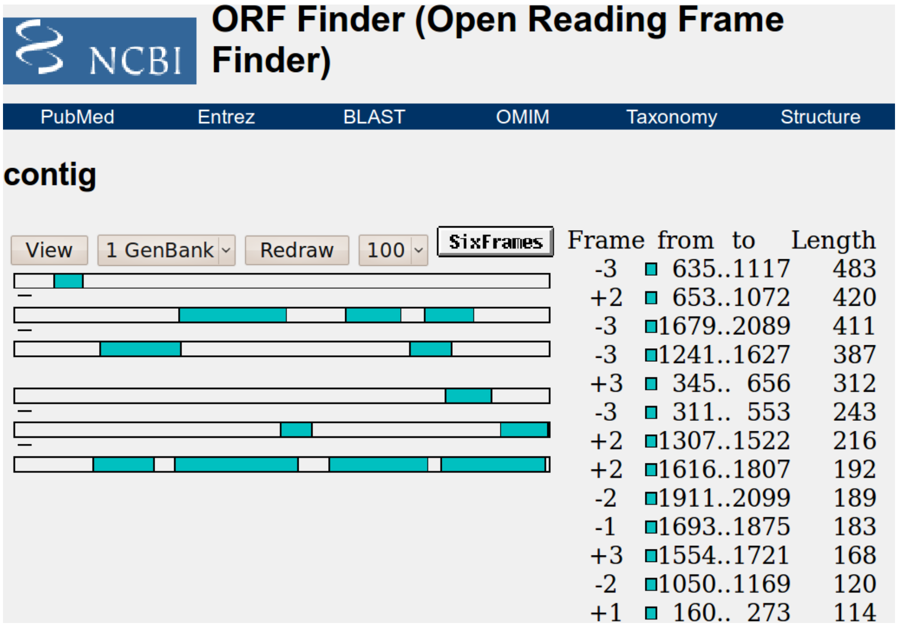


Figure 1.4: The three reading frames of a strand of DNA Each reading frame starts one nucleotide further giving rise to a different protein sequence. The other direction of the DNA also contains three reading frames so that there are six reading frames in total. (adapted from [18])



Frame	from to	Length
-3	□ 635..1117	483
+2	□ 653..1072	420
-3	□1679..2089	411
-3	□1241..1627	387
+3	□ 345.. 656	312
-3	□ 311.. 553	243
+2	□1307..1522	216
+2	□1616..1807	192
-2	□1911..2099	189
-1	□1693..1875	183
+3	□1554..1721	168
-2	□1050..1169	120
+1	□ 160.. 273	114

Figure 1.5: ORFs may overlap in prokaryotic genomes The picture shows all ORFs ≥ 100 nucleotides lying on the six possible reading frames within a part of the chromosome of *Cronobacter turicensis*. It can be seen that some ORFs overlap. Therefore the extraction of ORFs exceeding a specific length is not enough for a reliable gene prediction.

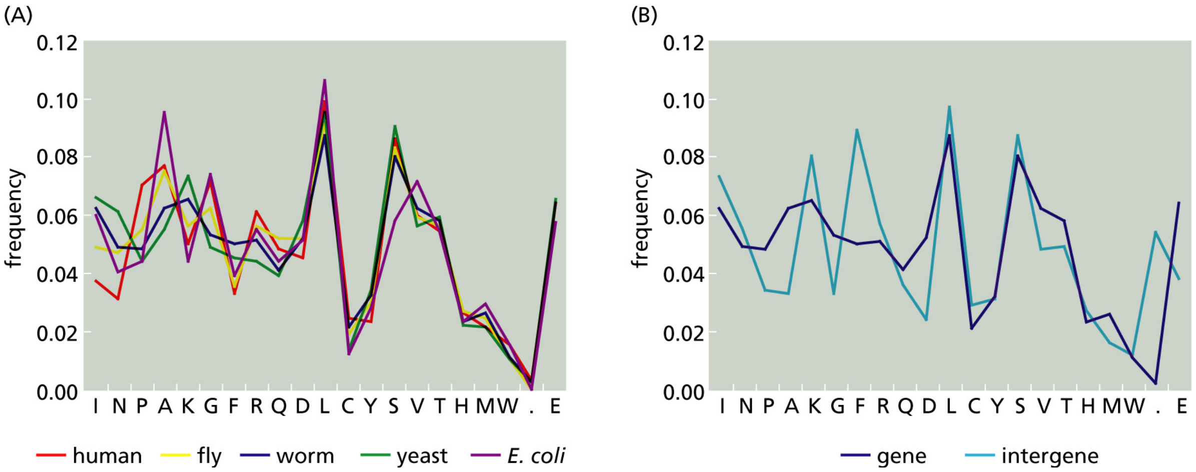


Figure 1.6: Plot showing the frequency of occurrence of different amino acid codons in genes and intergenic DNA The amino acids are represented by their one-letter code, and the three stop codons are combined and represented by the dot. **A:** shows a comparison of the amino acid codon frequencies in genes of humans, *D. melanogaster* (fly), *C.elegans* (worm), *S. cerevisiae* (yeast), and *E. coli* (E. coli). The amino acids have been ordered to show most difference between species on the left side of the graph. **B:** shows the amino acid codon frequencies for *C. elegans* in genes and intergenic DNA. Clearly some of the amino acid codons show a considerable difference in occurrence in coding and non-coding segments. (Adapted from [25]) (Source: [16])

target promoter, and (iii) they unwind the DNA near the transcription start site [26]. Bacterial housekeeping σ-factors are similar to the *Escherichia coli* σ^{70} 70-kDa σ-factor [27, 28] and typically bind to the -35 and -10 DNA sequence elements. The promoter sequences are quite conserved hexanucleotide sequences with the consensus sequences TTGACA at position -35 and TATAAT at position -10 [29] and can therefore be searched upstream of potential genes. (For a nice review about control logic in prokaryotic transcriptional regulation see van Hijum et al. [30])

- **Sequence homology** As there is selective pressure on protein coding regions they show a higher conservation on the sequence level as intergenic regions. Therefore the knowledge about similar regions to a region in the genome in comparison to other genomes can be used for the gene prediction.

1.2.2.3 Intrinsic methods

A huge amount of gene finders for prokaryotic genomes has been published, that solely use intrinsic information, that is information contained within the sequence for which genes should be predicted. This section contains the programs used within this work representing the whole group of intrinsic gene finders.

1.2.2.3.1 General procedure of intrinsic gene finders Intrinsic gene finders build models for coding regions, non-coding regions, and some of them for sites near the gene start, e.g. ribosomal binding sites (RBS). This is done explicitly for the genome of the organism in which the prediction should be made. The resulting models are then used for the prediction of the coding regions within the genome.

For each position in the genome and for each of the six possible reading frames it has to be decided whether the nucleotide is coding in one of the six possible reading frames or if the position is in a non-coding region. The distribution of codons and dicodons differs in coding and non-coding regions and varies from one organism to the other. Therefore the statistics about the distributions of codons or dicodons in each of the frames and in non-coding regions have to be determined by training. As training data for coding regions often long ORFs exceeding a specific cutoff are used as these ORFs are very likely genes. The length cutoff is determined depending on features like the GC content of the genome as it influences the probability that an ORF is long by chance. The higher the GC content of a genome the less probable are the AT rich stop codons. As training data for non-coding regions either regions not containing sufficiently long ORFs can be used or alternative reading frames of the long ORFs. After the training for each position in the sequence a likelihood score that the given position is coding or non-coding can be determined and if it is coding in which frame.

For the prediction of genes the genome sequence is scanned for stretches of DNA sufficiently long and exceeding a specific score.

As the prediction of gene starts is difficult information about sites near the gene starts are often used for their determination. In many cases ribosomal binding sites (RBS) and their distances from the gene starts are used. The model for the sites is often a two-component statistical model consisting of a positional frequency matrix for the sequence motif and the spacer length between predicted gene start and motif.

1.2.2.3.2 GeneMarkS GeneMarkS [31] combines models for coding regions, non-coding regions and sites near the gene start in an iterative Hidden Markov Model based algorithm.

GeneMark [32] and GeneMark.hmm [33] are the predecessors of GeneMarkS and can be seen as intermediate steps. Therefore the summary for all three programs can be found in this section.

For describing the models, GeneMark uses inhomogenous Markov chains and dicodon statistics, that is the sixth base x_6 in a sequence is dependent on its preceding five positions x_1 to x_5. So the probability that $x_6 = a$ is then $P(a|x_1, x_2, x_3, x_4, x_5)$. It is assumed that the dicodon statistic is different for each of the six possible reading frames so that it is treated for each of the reading frames separately. For the non-coding regions it is assumed that there is only one statistic for all dicodons.

GeneMark.hmm uses a semi-Markov model, HMM with duration, or explicit state duration hidden Markov model. GeneMark.hmm incorporates the distribution of gene lengths into the gene prediction. The idea is that on entering a specific state in a HMM the length distribution is used to determine the duration for this particular visit.

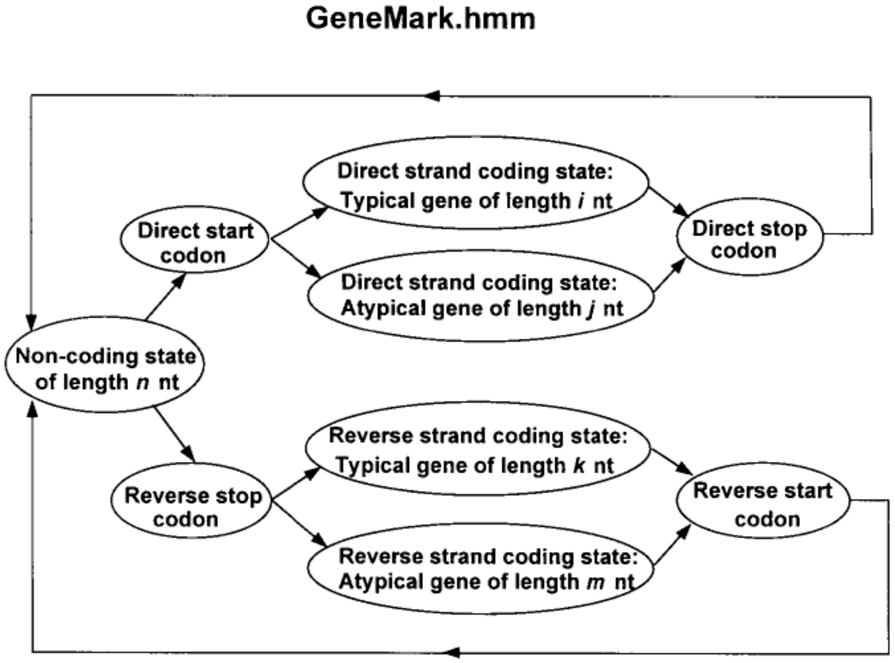

GeneMark.hmm
Direct strand coding state: Typical gene of length *i* nt
Direct start codon
Direct stop codon
Direct strand coding state: Atypical gene of length *j* nt
Non-coding state of length *n* nt
Reverse strand coding state: Typical gene of length *k* nt
Reverse stop codon
Reverse start codon
Reverse strand coding state: Atypical gene of length *m* nt

Figure 1.7: Hidden Markov model of a prokaryotic nucleotide sequence used in the GeneMark.hmm algorithm The hidden states of the model are represented as ovals in the figure, and arrows correspond to allowed transitions between the states. (Source: [33])

An example is the length of a particular gene, where the state is the "direct strand coding state" and length i is the duration (Figure 1.7). The model uses state emission probabilities to emit the respective number of bases, before a transition to another state of the model is made.

As sites near gene starts show a higher variability than previously thought, GeneMarkS creates a two-component statistical model of a conserved in evolution site upstream of the predicted gene start that is not only restricted to the detection of a ribosomal binding site (RBS). The parameters for the model are derived by applying Gibbs sampling to the multiple alignment of DNA sequences situated upstream of annotated translation starts. So GeneMarkS cannot only take into account known RBS sites but also discover possibly new sites upstream of gene starts.

GeneMarkS first determines heuristic initial parameters for the prediction based on features like the GC content of a genome and runs the first prediction for a genome using the GeneMark.hmm program (Figure 1.8). After this first run the upstream regions of genes are used to determine the first gene models specific for the organism. Afterwards GeneMark.hmm is run again using the newly determined models. This is done until convergence, that is until the change in the sequence parses, obtained in two subsequent iterations, is less than some predefined small value.

1.2.2.3.3 Glimmer 3.0 Glimmer 3.0 [34] uses interpolated Markov models (IMM) for the six reading frames for coding regions, one model for non-coding regions, and a model for sites near the gene start.

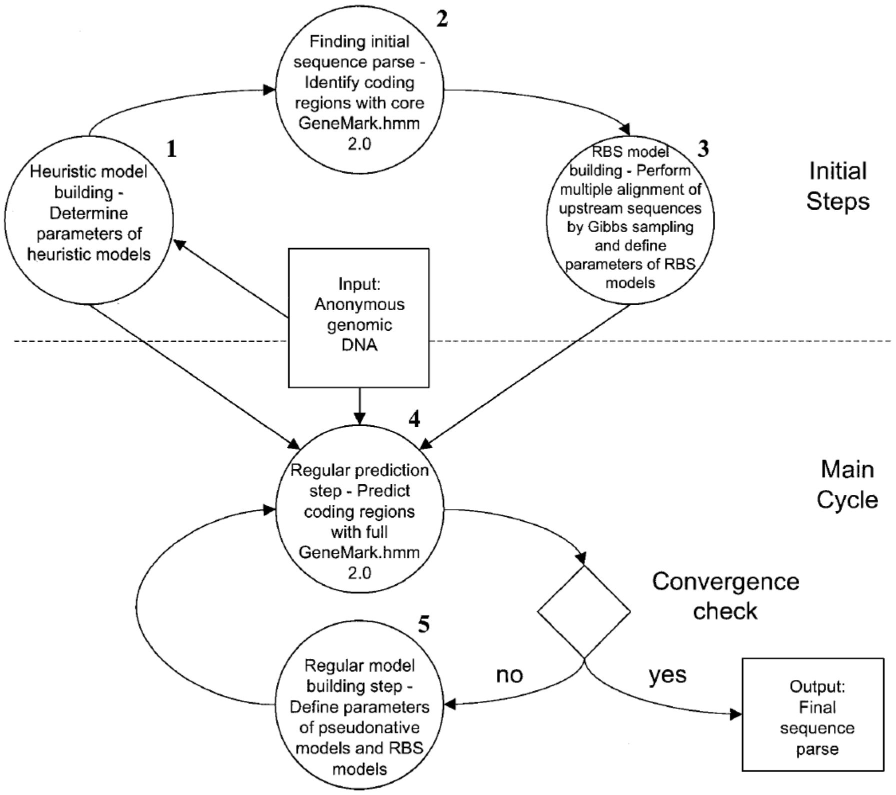

2
Finding initial sequence parse - Identify coding regions with core GeneMark.hmm 2.0
1
Heuristic model building - Determine parameters of heuristic models
3
RBS model building - Perform multiple alignment of upstream sequences by Gibbs sampling and define parameters of RBS models
Initial Steps
Input: Anonymous genomic DNA
4
Regular prediction step - Predict coding regions with full GeneMark.hmm 2.0
Main Cycle
Convergence check
5
Regular model building step - Define parameters of pseudonative models and RBS models
no
yes
Output: Final sequence parse

Figure 1.8: Step-by-step diagram of the GeneMarkS procedure (Source: [31])

ORFs longer than a specified cutoff are used for training, determined by the value that maximizes the number of non-overlapping ORFs produced and by that maximizing the amount of data in the training set. This works quite well for genomes with low GC content as the probability is small that an ORF is long by chance. For genomes with high GC content the length cutoff needs to be adjusted in order to get meaningful long ORFs. A distribution of amino acid distributions in genes in a series of organisms is compared against the ORFs created in the previous step. ORFs not corresponding to the distribution are removed from the training set as these are likely not coding. A set of non-coding ORFs is created from alternative reading frames of these long ORFs.

Using the training data IMMs from zeroth to eighth order are created. The nucleotides do not necessarily have to be located directly upstream of the current position but are also allowed to be further away. This window is called the context. If only less than 400 observations can be made in the genome sequence for a model of order x then it is checked whether model x contains significant information compared to the lower level x-1. If yes then the model of order x can be used otherwise the model of order x-1 is the model providing the longest context. Since Glimmer 3.0 the context is computed in reverse direction, that is that the context is built beginning from the stop to the start of a potential gene as the context is then referring to the coding region of the gene and not to the non-coding region. The score for the ORF is then the sum of the probability of each base conditioned on a context windows on its 3' side and the score of the ORF

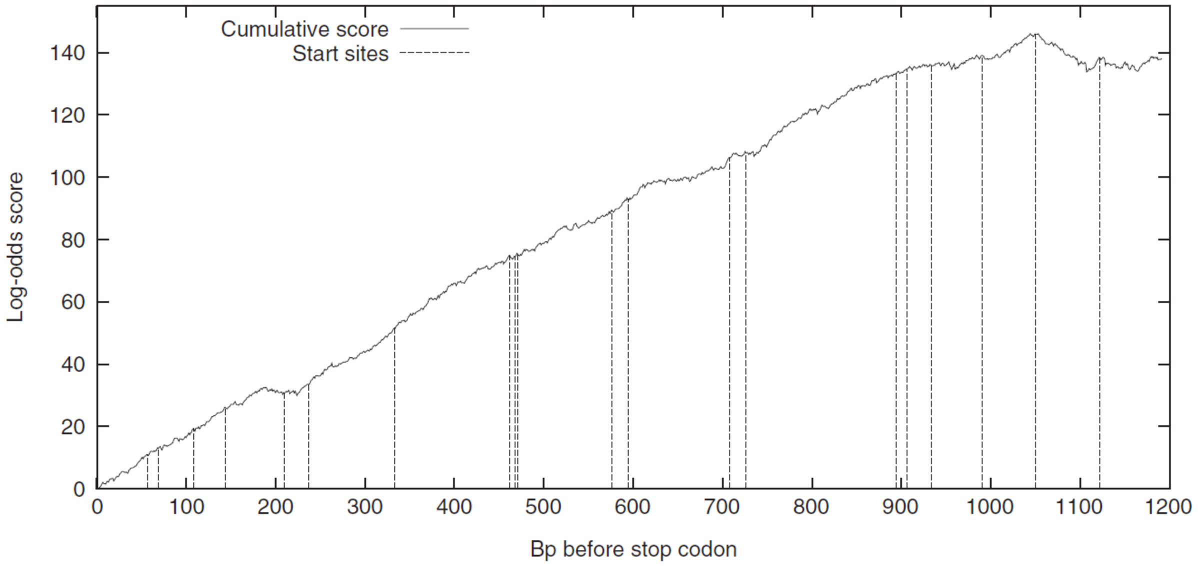


Figure 1.9: Scoring an open reading frame from the stop codon backwards in Glimmer 3.0. The stop codon is at position 0 on the X-axis and the cumulative log-odds score is plotted as the solid line. Positions of possible start codons are indicated by vertical dashed lines. This ORF contains the fructose bis-P aldolase gene in *Escherichia coli* EG14062 and the current Ecogene verified start site is at position 1050, near the peak score. This position is an update for the originally annotated start at position 1122. (Adapted from [34])

being the log-likelihood sum of the bases contained in the ORF. The score is computed incrementally as a cumulative sum at every position in the ORF. When plotting the scores for an ORF there is often a maximal peek visible located near the true gene start (Figure 1.9).

The resulting IMMs involve the weighted sum of terms of models from all orders. If a higher order model could be identified all lower order models are ignored. The IMM is then used to determine coding and non-coding regions in the genome.

The RBS could be integrated in Glimmer versions before Glimmer 3.0 by the use of the standalone program, RBSfinder, that can be run as a post-processor on the results of Glimmer's analysis. Glimmer 3.0 contains the integration of RBS detection in form of the ELPH software (`http://cbcb.umd.edu/software/ELPH`). ELPH uses Gibbs sampling and identifies likely shared motifs upstream of the predicted gene starts which can then be used within Glimmer 3.0.

1.2.2.4 Intrinsic & extrinsic methods

The problem with intrinsic gene finders is that the quality of their predictions varies depending on the genome [35, 36, 37]. Additionally the codings sequences (CDSs) predicted by intrinsic gene finders often need further time-consuming manual refinement, especially the gene starts. The major point of criticism is that intrinsic gene finders do not take into consideration a very valuable information: the knowledge about the conservation of existing protein coding genes in other organisms.

Therefore automatic methods have been developed that integrate intrinsic and extrinsic information like ORPHEUS [38], CRITICA [39] or EasyGene [40, 41]. The programs work similar.

CRITICA searches for stretches of conserved DNA within the genome in comparison to other genomes. If the translation of the aligned sequences has greater amino acid identity than expected for the observed percentage nucleotide identity, this is interpreted as evidence for coding as such excess identity provides evidence of amino acid conservation and, hence, translation. Additionally intrinsic features like the relative frequencies of hexanucleotides in coding frames versus other contexts (i.e., dicodon bias) are derived. These two types of information are integrated into a score. Regions of the sequence having higher than random scores are predicted as coding. The RBS is taken into account for the determination of the best gene starts.

EasyGene searches for significant matches of all ORFs exceeding a length threshold within the genome against UniProtKB/Swiss-Prot and trains a hidden Markov model (HMM) with states for coding regions as well as RBS. The HMM is then used to score all the ORFs in the genome. For each ORF the respective score is converted to a measure of significance R which is the expected number of ORFs that would be predicted in one megabase of random DNA. The lower R is, the higher is the probability that the ORF is coding.

ORPHEUS will be described as a representative for this group of programs in more detail.

1.2.2.4.1 ORPHEUS ORPHEUS [38] is a software with the goal to identify gene candidates with an emphasis on optimal prediction of gene starts. The gene finder is based on the assumption that information about coding regions derived from similarity searches is in principle more reliable than statistical data.

Therefore the first step is the search for regions in the genome significantly related to known proteins. These regions constitute initial homology based gene models. In order to be able to fill the regions of the genome where no initial homology based gene models could be detected, the homology regions are used to compute codon and base frequencies.

These frequencies are used for the determination of coding regions in regions of the genome where there are no initial homology based gene models. The normalized coding potential for the sequence segment $x_m...x_{m+3n-1}$, n codon starting at base x_m, is then defined as

$$R(x_m...x_{m+3m-1}) = \frac{Q(x_m...x_{m+3n-1}) - \mu n}{\sigma\sqrt{n}} \tag{1.1}$$

$Q(x_m...x_{m+3n-1})$ is called the coding potential. The normalized coding potential is computed for each ORF and its alternative reading frames. The difference of the highest coding potential to the alternative coding potentials is named coding quality. If the coding quality exceeds a specific threshold and if the length of the ORF producing the highest coding potential is longer than a specific cutoff (e.g. 100 bp), then this ORF is a candidate ORF. The candidate ORFs are extended to the closest of the possible starts in frame as long as this elongation does not produce overlaps greater than 6 bp with other ORFs and if the coding quality of the new candidate exceeds another threshold value. In case that two candidate ORFs should overlap the one with the higher coding quality is kept.

In order to determine 5' ends of those ORFs where alternative starts are present ORFs with unambiguous starts are used for the definition of a scoring matrix for ribosomal binding sites (RBS). ORFs with unambiguous starts are ORFs having the next upstream ORF within 30 bases. 20 bases upstream of each of the candidates are used in an iterative manner to locate the conserved pattern in each of the upstream sequences. This model is then used for the determination of the gene start in the ambiguous cases.

This way homology is used as major information for the gene prediction and the gaps between the homology genes are filled with intrinsic predictions including a translation start when possible.

1.2.2.5 Pipelines integrating various information

It can be observed that intrinsic gene finders are very commonly used in many studies and that also the predictions of gene finders including intrinsic & extrinsic information differ. Therefore pipelines have been developed that mainly decide about the best of the proposed gene models from intrinsic and intrinsic & extrinsic gene finders and return a consensus prediction. Examples are YACOP [42] or CONSORF [43].

YACOP integrates predictions from CRITICA [39], Glimmer 2.02 [44], Glimmer 2.10 [45], ORPHEUS [38], and ZCurve [46]. However it is neither described in the publication nor in the supplementary material how the methods are combined exactly in order to get a consensus prediction.

CONSORF will be described as a representative for this group of programs in more detail.

1.2.2.5.1 CONSORF CONSORF [43] is a consensus prediction software that integrates the intrinsic predictions from GeneMark [32], GeneMark.hmm [33] and GLIMMER [45] as well as extrinsic information by the integration of similarities produced by FASTX [47]. Additionally to a consensus gene prediction CONSORF provides prediction reliability scores, predicted frameshifts, alternative start sites and best pair-wise match information against other prokaryotes.

CONSORF first detects potential genes by searching for homologies to known protein sequences and then the remaining gaps on the genome, where no gene model has been identified this way, are filled with intrinsic algorithm-based predictions. This approach is similar to ORPHEUS [38]. An overview over the workflow of CONSORF can be seen in Figure 1.10.

First a pair-wise genome-to-proteome comparison between the DNA of the genome to be predicted and known protein coding genes in other genomes are performed. In case of overlapping hits not supporting the same ORF, the hit with the lower bitscore is removed. Then for each ORF a reliability score is computed consisting of the number of hits and the sum of the bitscores of the hits from the FASTX alignments with the other organisms. The alignment with the highest bitscore is used as so called "representative alignment", that is used for the determination of possible gene starts. The resulting ORFs are called "candidate homology ORFs". In the case that two candidate homology ORFs overlap by more than 10% of their lengths the one with the lower reliability score is removed. This results in the set of final "homology CDSs".

Secondly, for the case that ORFs specific just to a small group of organisms are discarded in the previous overlap resolvement so called "alternative CDSs" are created. They are determined by the best FASTX alignment among the comparisons with available proteome sets.

Thirdly, ab-initio algorithm based predictions are conducted for the genome. Then the number of predictions and the sum of the nucleotide lengths of the predicted CDSs with the same stop and start codon positions are used to compute their consensus reliability scores. There is no overlap resolvement done in order to stay as sensitive as possible. Each ORF gets all probable starts between the longest possible start and the most probable start assigned. The most probable start is the start with the largest number of occurrences over all intrinsic predictions. If two alternative starts are predicted by the same number of methods then the start producing the shortest gene product is selected.

Fourthly, the ab-initio gene models and homology gene models are merged into a consensus prediction. The "homology CDSs" with a sum of bitscores greater than a given cutoff are chosen first. Then "ab initio CDSs" with a sum of CDS lengths greater than another score cutoff are chosen, removing CDSs overlapping significantly with the existing "homology CDSs" or with another "ab initio CDS" with the higher consensus reliability score.

The last step is the determination of the most likely start site for each of the previously determined "integrated CDSs". For each of the CDSs all pairwise FASTX alignments containing N-terminal residue matches are analyzed and the start site overlapping with most of the alignments is assigned as new shortest start. The resulting CDS provide the final "representative CDSs".

1.2.2.6 Benchmarking gene predictions

There have several programs for gene finding in prokaryotic genomes been introduced above. In order to be able to assess the performance of the respective methods the predictions of each of the methods have to be compared against sets of validated genes.

Generally the existence of genes can be measured on transcript and on peptide basis. Transcript data has previously been investigated using microarrays (e.g. [48, 49]) and the advent of next generation sequencing technologies has recently even opened the possibility for “deep-sequencing” of prokaryotic transcriptomes (reviewed in [50]). Peptides can be measured in proteome experiments (e.g. [51]).

All these approaches have limitations. When measuring the RNA transcripts or proteins of an organism not all genes are necessarily expressed at the time point of the analysis. Additionally too low levels of transcripts or proteins may not be measurable.

All approaches have in common that they do not contain non-genes as they can only

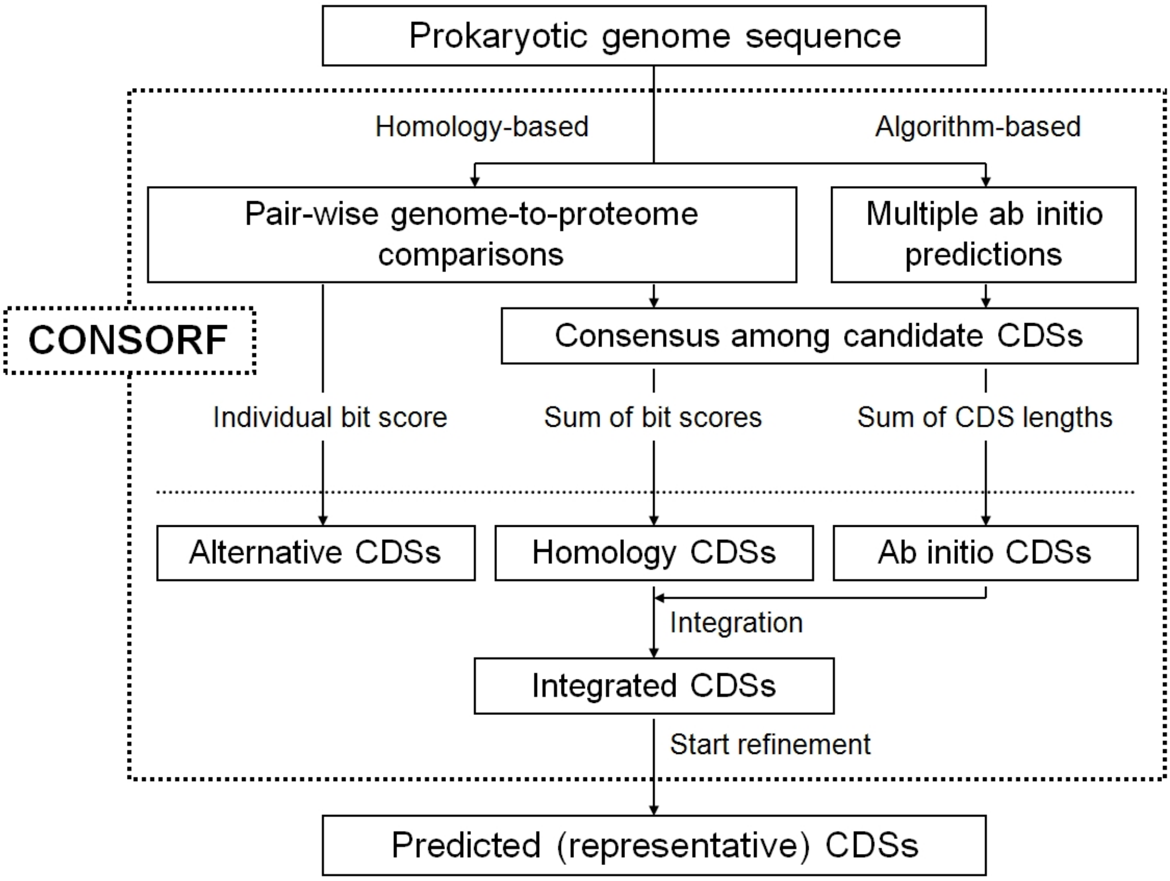


Figure 1.10: CONSORF workflow From a prokaryotic genome sequence, the CONSORF system predicts CDSs in two complementary approaches: homologybased and algorithmbased. In the homologybased approach, pairwise genome to proteome comparisons via the FASTX [47] program are performed to generate both "homology CDSs" and "alternative CDSs", while multiple ab initio predictions are conducted to provide "ab initio CDSs" in the algorithmbased approach. "Homology CDSs" are determined from the representative FASTX alignment with the highest sum of bitscores in consensus analyses regarding stop, start, and frame change positions, while "ab initio CDSs" are determined from the consensus of the algorithmbased CDSs with the highest sum of CDS nucleotide lengths in the consensus analyses regarding only stop and start positions. On the contrary, "alternative CDSs" are directly determined from the FASTX alignments with the highest individual bitscore across all the pairwise comparisons. By integrating the complementary "homology CDSs" and "ab initio CDSs", avoiding a significant positional overlap on the genome, the "integrated CDSs" were predicted with high accuracy. To determine the more likely start site among candidate starts, the "integrated CDSs" aligned exactly with the Nterminal end of a library protein in the pairwise FASTX comparisons were inspected to provide the final "representative CDSs". (Adapted from supplementary material for [43])

make statements whether a specific transcript or protein was measurable at the time of the experiment, but not whether the transcript or protein does actually exist or not.

Data of validated genes is very limited at the moment. The genes published in RefSeq [52] are not a gold set of experimentally verified gene starts. To my knowledge the only available set of experimentally verified genes consists of 195 genes of *E. coli* [53] as used for the evaluation of GeneMarkS by Besemer et al [31]. While these N-termini have been proven by Edman sequencing there exist two studies that provide support for specific gene starts but do not prove that these are the correct gene starts as the data originates from proteome experiments. These studies are a set of 606 genes of *Halobacterium salinarum* (strain R1, DSM 671) [54], of 328 genes of *Natromonas pharaonis* (strain Gabara, DSM 2160) [54], and of 278 genes of *Deinococcus deserti* VCD115 [55]. The datasets from the latter studies are not suited for benchmarking of gene predictions.

What is missing is an almost complete set of genes for the genome of an organism or even a set of genes of taxonomically diverse organisms, as a gene finder performing well on the genome of one organism does not necessarily have to perform well on other genomes.

Basically there are two levels at which the performance of gene finders can be evaluated. On the one hand it is possible to classify the prediction for each nucleotide of the genome. This level is the nucleotide or base level (Figure 1.11). On the other hand it is possible to evaluate the predictions on the protein level (Figure 1.12).

The following measures are known from other areas in which predictions have to be evaluated.

The sensitivity (sens) is a measure how many of the actual genes have been predicted as genes.

$$sensitivity = \frac{TP}{TP + FN} \tag{1.2}$$

The specificity (spec) is a measure how many of the actual non-genes have been predicted as non-genes.

$$specificity = \frac{TN}{TN + FP} \tag{1.3}$$

The positive predictive value (PPV) is a measure how many of the predicted genes are actual genes.

$$PositivePredictiveValue = \frac{TP}{TP + FP} \tag{1.4}$$

Genes can be expressed at different timepoints and at different levels that may not be measurable anymore. Therefore it is almost impossible to identify a trusted set of true negatives, that is non-existing genes. For that reason specificity (Formula 1.3) is not a suited measure for the performance of gene prediction methods and the positive predictive value (PPV) (Formula 1.4) is typically used for measuring the performance of gene predictions [56].

In order to get a more detailed overview over the prediction results additional entities can be introduced besides true positive (TP), false positive (FP) and false negative (FN). The first measure is the number of genes agreeing in the gene stop and strand

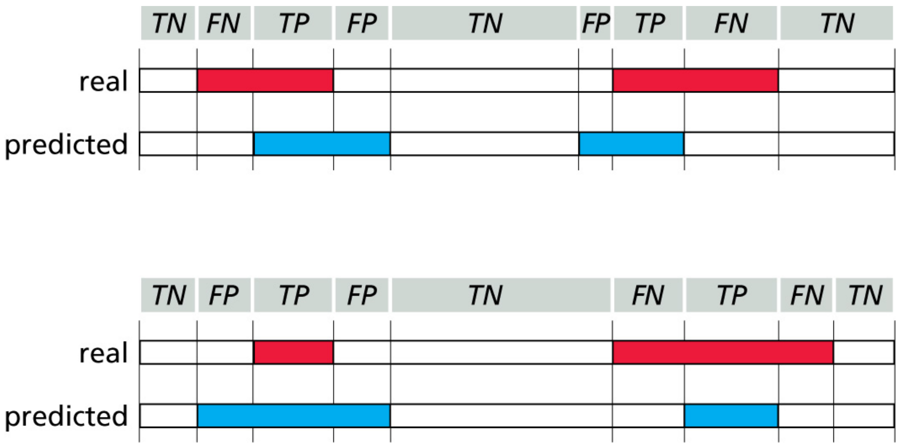


Figure 1.11: Comparison at the base level of actual and predicted genes Red bars are entire actual genes. Blue bars are entire predicted genes. Regions of the sequence are assigned as true positive (TP), true negative (TN), false positive (FP), or false negative (FN). (Adapted from [16])

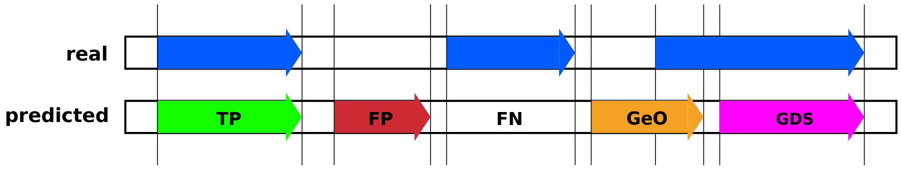


Figure 1.12: Comparison at the protein level of actual and predicted genes Actual genes are shown as blue arrows on the top, the predicted genes are shown below as colored arrows. The arrow heads are the gene stops. **green:** true positive (TP) **dark red:** false positive (FP) **white (labeled with FN):** the prediction for this gene is missing, false negative (FN) **orange:** gene overlapping with the published gene model but not sharing the same stop coordinate (Genes Overlapping, GeO) **lilac:** gene agreeing in the gene stop and strand but disagreeing in the gene start (Genes with Different Start, GDS)

but disagreeing in the gene start (Genes with Different Start, GDS) (Figure 1.12). If GDS are observed they are counted as TP in the formulas. This makes it possible to evaluate the gene prediction without the influence of the gene starts, which are difficult to predict. Additionally a measure for genes overlapping with the actual gene model but not sharing the same stop coordinate (Genes Overlapping, GeO) can be defined (Figure 1.12).

1.2.3 Non-coding sequences

1.2.3.1 Overview

Besides the coding sequences of a prokaryotic genome that are transcribed into messenger RNA (mRNA) which is then finally translated into the amino acid sequence of a protein, there exists a number of non-coding sequences.

With the technology of next generation sequencing RNA-sequencing experiments (e.g. [5]) become feasible and extended knowledge also about non-coding RNAs available.

The first type of non-coding sequences are non-coding functional RNA sequences that are not translated into a protein. Among these sequences are ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs). rRNAs constitute the major components of the bacterial ribosomes, which are composed of several subunits. rRNAs are very well conserved and are present in all extant species [57]. They presumably date back to the earliest forms of life and thus can be used to compute the evolutionary relationships between all species on earth [58]. Therefore rRNA can be used to taxonomically classify an organism and to estimate the rates of species divergence. Thousands of rRNAs have been determined, also for species where no complete genome sequence is available. The rRNA sequences are stored in specialized resources like the Ribosomal Database Project RDP (`http://rdp.cme.msu.edu/`) or SILVA (`http://www.arb-silva.de/`). The space between the subunits of the ribosomes is occupied by the transfer RNAs (tRNAs). Their anticodons base pair with mRNA codons in the 30S subunit, whereas their 39-CCA ends, which carry the growing polypeptide chain and the incoming amino acid, reach into the 50S subunit, the location of the peptidyl transferase center, where peptide bond formation is catalyzed [59].

The second type of non-coding sequences are Cis-regulatory elements that control the transcription of genes. Promoters are such Cis-elements and are typically located upstream of the protein coding region while enhancers may be located more distantly away from the gene start and may interact with the promoter by building loops [60].

The third type of non-coding sequences are DNA sequences, related to known genes, that have lost their protein-coding ability or are otherwise no longer expressed in the cell. These genes are called pseudogenes and can originate by the disruption of a reading frame or promoter regions by point mutations, frameshifts, or the integration of transposable elements [61]. If such mutations occur in genes that are not subject to selective pressure (anymore), that is they are no longer required, they will accumulate mutations and can only be maintained in the genome for some time but are gradually degraded and eliminated by deletions [62, 63, 64].

The fourth type of non-coding sequences are transposons, sequences of DNA able to move to different positions within a genome. By this process called transposition they may cause the disruption of genes when inserted into it. There can be two kinds of transposons distinguished, class I retrotransposons and class II DNA transposons. The DNA transposons normally need the enzyme transposase. The transposase binds to a target site in the genome and cuts the target site in such a way that it produces sticky ends. Then the transposon is cut out and ligated into the target site and the gaps resulting from the sticky ends are filled by a DNA polymerase and the sugar-phosphate backbone is closed by DNA ligase. The result is that the insertion site is surrounded by short direct repeats followed by inverted repeats. This feature can be used for the detection whether a gene duplication was caused by a transposon. The duplications are typical events in the evolution of genomes and can lead to neo-functionalization of one of the copies and are thus an important process in the development of new functionality [65, 66].

Repetetive DNA constitutes another type of non-coding sequence and is ubiquitous in microbial genomes [67]. Various kinds of repeated DNA have been identified in the genomes of many prokaryotes, whereas the repeats can be included in genes, in intergenic sequences, or in transposable elements [68]. DNA elements such as insertion sequences (ISs) and transposons are major evolutionary actors in the genome since they mediate genome rearrangements, plasmid integration, and gene transfer [69, 70]. By doing that repeats are representative of important evolutionary mechanisms that allow bacteria to adapt faster to environmental changes. Repeats can be a hint towards the recent integration of transposons in a genome [71].

1.2.3.2 Determination of selected non-coding sequences used in this work

The detection of non-coding elements within a prokaryotic genome is part of the primary annotation. Some of the analyses are integrated into automatic pipelines like PEDANT [72], other analyses need to be done specifically for every novel prokaryotic genome.

1.2.3.2.1 rRNAs The fact that rRNAs are very well conserved between organisms even on the DNA level can be used for their detection. The easiest way to detect rRNAs within a genome sequence on the computer is to use the 5S, 16S and 23S rRNA sequences of the most closely related species and to execute a BLASTN [73] search against the genome sequence of the organism.

1.2.3.2.2 tRNAs As the sequence of tRNAs is not as good conserved as the sequence of rRNAs it is not possible to detect tRNAs only by similarity to known tRNAs. Fichant et al. therefore introduced the software tRNAscan [74] which uses two characteristic features of tRNA to predict them within a DNA sequence: firstly the local potential hairpin and stem structures consistent with the cloverleaf secondary structure motif of the typical tRNA sequence [75] and secondly the presence of several invariant or semi-invariant bases that define two conserved regions [75, 76] (Figure 1.13).

1.2.3.2.3 Repeats Due to the role of repeats in genome stability, gene transfer, and antigenic variation [68] they are subject to investigation.

There have several tools been proposed for the repeat detection in genomes (for example [78, 79, 80]). At the time of the implementation of the REPuter [80] software all of these tools had a limitation of the maximal length of the allowed input sequence. In the meanwhile REPuter is one of the standard tools for the detection of (in-)exact repeats in genomes. REPuter distinguishes between forward repeats, palindromic repeats, reverse repeats and complemented repeats (Figure 1.14).

1.2.3.2.4 Pseudogene detection Various properties of pseudogenes can be used for their detection. On the one hand the systematical analysis of alignments for the detection of truncated coding sequences is possible. This approach is used by the $\Psi - \Phi$ software [61] by Lerat et al for example. On the other hand the analysis of non-synonymous

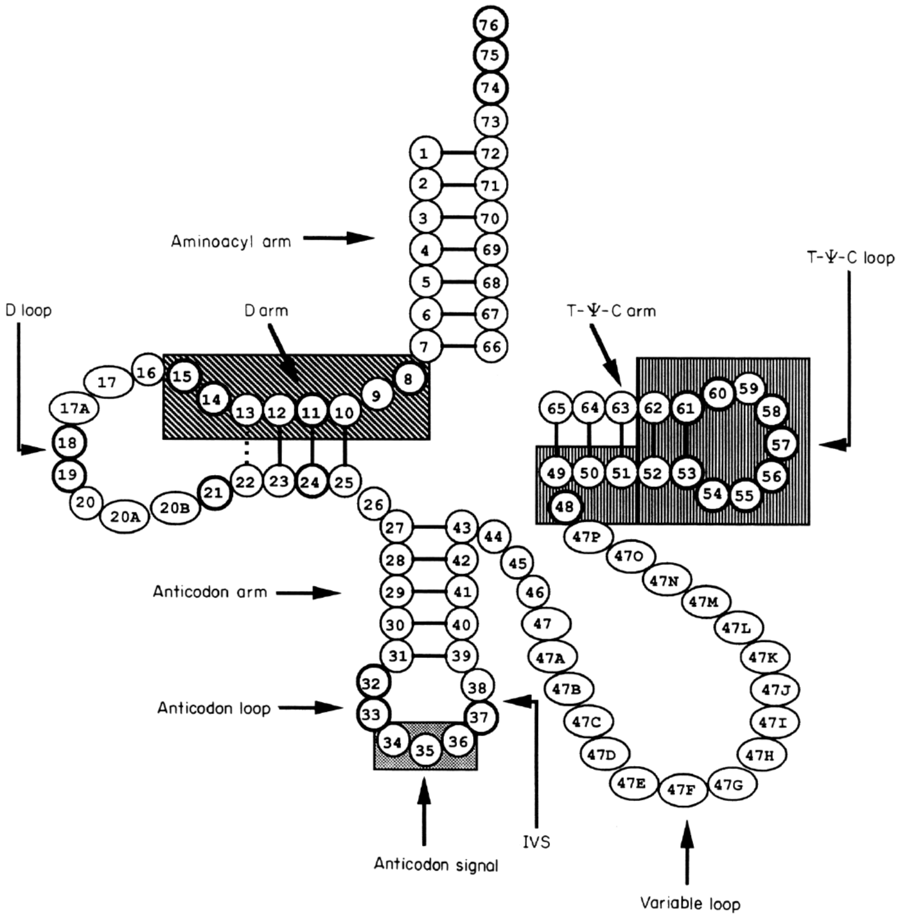


Figure 1.13: Cloverleaf secondary structure of a tRNA sequence The standard system of numbering tRNA sequences is given [77]. Circles represent nucleotides that are always present; among them, the thick-edged circles denote invariant or semi-invariant nucleotides. Ovals represent nucleotides that are not present in each tRNA sequence. The boxed positions correspond to the $T - \Psi - C$ (from 48 to 62) and D (from 8 to 15) signals as defined for use in our algorithm. IVS, intervening sequence. (source: [74])

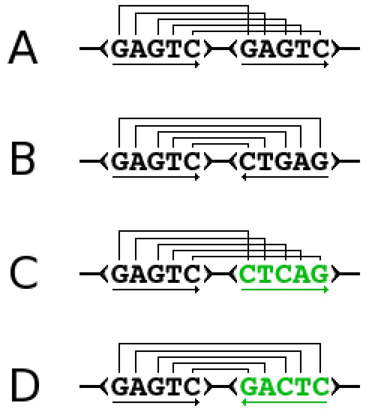


Figure 1.14: Repeats detected by REPuter A: forward (direct) match **B:** reverse match **C:** complement match **D:** palindromic match

(K_a) in relation to the synonymous (K_s) substitution rates can be analyzed, the K_a/K_s ratio, which is a measure whether there is selective pressure on a specific gene.

The precondition is in both cases that there are comparison genomes related closely enough available so that their sequences produce significant long alignments in which mutations can be counted [61].

The identification of the optimal method for pseudogene prediction is difficult as there exists to my knowledge no genome wide gold standard for pseudogenes which could be used to evaluate the predictions. Additionally there exist different definitions of pseudogenes. Existing programs like $\Psi - \Phi$ therefore produce lists of hypothetical pseudogenes that need to be analyzed manually afterwards.

The $\Psi - \Phi$ software [61] mentioned above is one way to create pseudogene candidates. The software has the drawback that it can only use quite similar informant genomes. In order to use a broader range of sequences for the analysis of truncated genes the best BLAST hits for all protein sequences of the query genome against a non-redundant database of publicly available protein sequences can be extracted. Protein sequences producing hits with a specific similarity (e.g. percent identity $\geq$ 30%) and a specific length difference between query and hit sequence (e.g. ratio cutoff between 50 and 79%) are pseudogene candidates. A third way to create pseudogene candidates is to run a geneprediction on the genome that also accepts quite short gene models. If two neighboring genes on the same strand show similarities to the same entries of the sequence database then a frameshift within the genomic sequence is probable.

1.3 Function annotation

After the previous steps the putative genes on the genome are known. In order to determine the molecular functions (e.g. what kind of enzyme is the protein) and biological functions (what is the function of the protein in the organism) of the encoded proteins, a multitude of methods has to be applied in the laboratory, e.g knockout experiments. The determination of the functions of all proteins of an organism is very laborious and expensive. Therefore a nearly comprehensive analysis of the functions of genes is only available for a few model organisms like *Escherichia coli* or *Saccharomyces cerevisiae*.

Therefore bioinformatics is often used for the prediction of functions. These predictions can then be used for the design of specific experiments in the laboratory.

1.3.1 Ontologies

A comparable and consistent vocabulary of function annotations is necessary in order to be able to compare, predict and evaluate functions of proteins. Ruepp et al. [81] state that such a vocabulary should own the properties of human usability, computer readability, independence on organism, breadth and depth of coverage, stability and extendibility. Ontologies are such a vocabulary. They define formal and explicit specifications of the terms used and the relationships between the terms.

The strength of ontologies is that they offer the possibility to automatically annotate genes using a computer. An overview over the ontologies used in this work is given below.

1.3.1.1 mips Functional Catalogue (FunCat)

By April 24, 1996, the completely sequenced genome of *Saccharomyces cerevisiae* was available [81]. The Munich Information Center for Protein Sequences (MIPS) served as informatics centre and annotated and stored the incidental data in databases. While processing the data a hierarchically structured classification system for the functional description of proteins from any organism was developed. This classification system is named Functional Catalogue (FunCat) [81]. The early versions of the functional catalogue contained only the categories required for the description of the *Saccharomyces cerevisiae* biology [82, 83], later the ontology has been extended to plants, prokaryotes and animals [84, 85, 86].

As an example, the UniProtKB/Swiss-Prot entry NADE_YEAST from *Saccharomyces cerevisiae* is annotated with FunCat category "01.01.03.01.02". As the FunCat is a hierarchical classification the protein is automatically annotated with FunCat "01" (metabolism), "01.01" (amino acid metabolism), "01.01.03" (assimilation of ammonia, metabolism of the glutamate group), "01.01.03.01" (metabolism of glutamine), and "01.01.03.01.02" (degradation of glutamine).

A protein can be member of more than one category in order to account for the different functions of a protein.

The big advantage of FunCat is, that it has only a limited set of categories that the annotators can keep at the back of their minds.

FunCat has often been used in bioinformatics and machine learning studies [87, 88, 89]. In 2008 Tetko et al. analyzed the BFAB [90] gold set of manual FunCat annotations and developed an approach for the reliable prediction of FunCat annotations for proteins of unknown functions. The annotation tool is called FUNcat Annotation Tool (FUNAT) [91].

1.3.1.2 Gene Ontology (GO)

The Gene Ontology (GO) [92] is a collaborative project across many laboratories to provide a controlled vocabulary of genes and gene-associated information. It was introduced by the Gene Ontology Consortium and mainly focuses on eukaryotes. GO is not strictly hierarchical but organized as acyclic graphs. The Gene Ontology is divided into three subontologies: "molecular function", "cellular component" and "biological process". Each entry in the GO consists of a number and an associated name which is member of one of the three subontologies.

As an example, the UniProtKB/Swiss-Prot entry GNPAT_HUMAN from *Homo sapiens* is member of the "biological process" GO categories "GO:0006631" (fatty acid metabolism), "GO:0009887" (organ morphogenesis) and of the "molecular function" GO category "GO:0008415" (acyltransferase activity).

There exist several tools that assign GO terms to novel protein sequences. Blast2GO [93] offers GO assignments based on similarity searches with statistical analysis and highlighted visualization on the directed acyclic graphs. Blast2GO was adapted to high throughput analyses for the SIMAP database [94].

1.3.1.3 Enzyme Commission numbers (EC numbers)

The Enzyme Commission numbers (EC numbers) have been introduced in 1956 as a numerical, hierarchical scheme describing the catalyzing functions of enzymes [95]. Additionally the EC nomenclature has been widened to a nomenclature scheme for membrane transport proteins (TC system) [96].

As an example, the UniProtKB/Swiss-Prot entry KIN28_YEAST from *Saccharomyces cerevisiae* is annotated with EC "2.7.11.23". As EC numbers are hierarchical the protein is automatically annotated with: "2" (transferase), "2.7" (transferring phosphorus-containing groups), "2.7.11" (protein-serine/threonine kinase), and "2.7.11.13" ([RNA-polymerase]-subunit kinase).

1.3.2 Transfer of annotation by sequence homology

It has been observed that the relationship of sequences implies sequence, structural and functional similarity [97]. Such related sequences are called “homologs” as they share a common origin. If sequences are related closely enough this property can be used to transfer structure and function from one protein sequence to the other.

The connection between sequence similarity and conserved function and structure is used in bioinformatics methods to transfer protein annotations.

1.3.2.1 Determining sequence similarity

There have been the Needleman-Wunsch algorithm for the detection of global [98] and the Smith-Waterman algorithm for the detection of local [99] alignments of sequences been published. These two dynamic programming algorithms both find the optimal alignments. But they are not appropriate for fast sequence comparisons against large sets of sequences like they exist in databases. Therefore the faster heuristic approaches for local similarity searches BLAST [73] and FASTA [100] have been introduced and are widely used for the determination of sequence alignments and thereby the similarity of sequences.

1.3.2.2 Uni-directional sequence similarity

The easiest and fastest way to transfer annotations is to search for the most similar sequence in public databases to a query sequence. The annotations of the detected protein are then transferred to the query protein.

There are several difficulties associated to the transfer of the annotations of the best hit. The first one is that only annotations of proteins with high quality annotations should be used for the transfer of annotations. An example are proteins from the

UniProtKB/Swiss-Prot database. Otherwise spurious annotations might be transferred from one protein to the other. Additionally only annotations of sequences sharing enough similarity should be used for the annotation transfer [97]. For enzymes it has been suggested that 40 to 70% sequence identity is necessary for functional prediction with 90% accuracy [101, 102].

One of the main problems of uni-directional sequence similarity becomes apparent when looking at the wide range of 40 to 70% necessary sequence identity mentioned before: as the conservation of protein families differs from each other, there can no global threshold be given that assures that the annotations can be reliably transferred for each protein family.

1.3.2.3 Bi-directional sequence similarity and orthologous groups

1.3.2.3.1 Bi-directional sequence similarity - orthology In order to overcome the threshold problem of uni-directional best hits, orthology is often used for the transfer of annotations. Two genes are orthologous if they directly evolved from a single gene in the last common ancestor [103]. Orthologs are most likely to share the same functions.

With the assumption that orthologous genes have evolved by divergent evolution orthologous genes are detectable by bidirectional best hits (BBHs) [104, 105, 106] as the similarity between the orthologous genes should be very high on the amino acid level, assumed that the genes kept their functions. As an example, be the best hit for protein "a" from organism A in organism B protein "b". If the best hit of protein "b" from organism B in organism A is also the protein "a" then "a" and "b" are bidirectional best hits between the organisms A and B. The annotations of the orthologous gene can then be transferred to the other gene.

Two orthologous genes that are members of a protein family that is subject to fast evolution may not show a very high degree of similarity anymore, but when comparing the whole proteomes against each other the two genes will still show the highest similarity to each other. Therefore the cutoff problem of uni-directional sequence similarity is understated.

BBHs are more difficult to use as uni-directional hits. Additionally the age of the last common ancestor of two BBH organisms influences the transferability of annotations. If the proteome of a member of the phylum *Chlamydiae* should be compared to a member of the phylum *Proteobacteria* for example, then the probability that functions of housekeeping genes can still be transferred is given, while the probability that the annotations of phylum specific genes can be reliably transferred is quite low.

1.3.2.3.2 Orthologous groups Therefore the transfer of annotations from only one comparison genome is extended to many comparison genomes in the concept of orthologous groups. Pairs of genes of different species connected by BBHs (=orthologs) can be grouped into orthologous groups by joining them if pairs have genes in common. It has been observed that such groups of orthologous genes most often share the same or similar functions [104].

Each of these orthologous groups constitutes a protein family. The determination of the homogenity of the annotations in a orthologous group is a measure for the transferability of annotations between the members. If the annotations are very homogeneous then it is very likely that also the unannotated proteins share the same function. If the homogenity is not very high then the transfer of annotations is not reliably possible.

Orthologous groups provide the best possibility to transfer annotations. Nevertheless uni-directional best hits are often used as they are easier to handle.

1.3.2.3.2.1 Existing resources Probably the first resource for orthologous groups were the Clusters of Orthologous Groups of proteins (COGs) [104, 105, 107]. Orthologous groups have been built by bidirectional best hits between at least three phylogenetic lineages. It seems that the initial clusters have been analyzed manually. The latest version from the year 2003 contains 66 genomes of unicellular organisms.

The evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) [108] is the follow-up approach for orthologous groups of genes. The clusters have been constructed similar to the above methodology through identification of reciprocal best BLAST matches and triangular linkage clustering. Each group is automatically annotated with a functional description derived from the functional annotations of its members. The current version contains 630 complete genomes (529 bacteria, 46 archaea and 55 eukaryotes). eggNOG also contains extended versions of the older COGs.

The KEGG Orthology (KO) System is consisting of manually defined orthologous groups that correspond to KEGG pathway nodes and BRITE hierarchy nodes. The orthologous groups are constructed using bidirectional best hits [109]. The value of the KEGG orthologous groups lies within the fact that they are connected to the KEGG pathways and that the annotations for selected organisms are checked manually.

1.3.2.3.2.2 ComparDB - (not only) a resource for building own orthologous groups The existing resources do not contain all available complete genomes and especially no private genomes, that is genomes not published yet. If one wants to work with orthologous groups nevertheless the protein coding regions of the new genome have either to be assigned to existing orthologous groups by similarity to cluster members or own orthologous groups have to be built.

The creation of orthologous groups based on bidirectional best hits (BBHs) is the most commonly used procedure [104, 105, 107]. One should always keep in mind that BBHs can by definition only be computed between complete genomes. But in order to be able to work with BBHs the similarities of all against all protein sequences need to be computed. As this is extremely expensive in time and CPU power SIMAP [94] contains the matrix of all precomputed all-against-all similarities between proteins with an opt-score ≥ 80.

The extraction of BBHs between all complete organisms from SIMAP has been accelerated by a database containing all bidirectional and unidirectional best hits between all complete non-eukaryotic genomes, the ComparDB. This database allows easy and fast retrieval of BBHs and best hits for the computation of clusters of orthologs and inparalogs [106].

1.3.2.4 Phylogenomics

It is common for groups of genes similar in sequence to have diverse although usually related functions [110]. Therefore the identification of sequence similarity is frequently not enough to assign a predicted function to an uncharacterized gene.

Phylogenomics assigns known functions to the evolutionary tree of the homologs of a query protein. The functions of the uncharacterized gene are then inferred by their phylogenetic position relative to the characterized genes (e.g. [111, 112]). This allows to choose the genes with known functions for the annotation that are most likely to have the same functions as the query gene.

Phylogenomics is especially precious when the amounts or rates of change vary between lineages as it allows for evolutionary branches to have different lengths. Additionally the multiple sequence alignments allow for masking, that is regions of genes in which sequence similarity is likely to be "noisy" or misleading rather than a biologically important signal can be excluded from the analysis. Pairwise alignments as used for the similarity searches mentioned previously cannot mask out regions [110].

1.3.3 Transfer of annotation by domain homology

If no sequences similar enough can be identified on the whole sequence level it might still be possible to detect known protein domains within the sequences. Protein domains can give hints towards possible functions of proteins.

1.3.3.1 Protein domains

Protein domains can be defined differently. In structural biology, a domain is defined as a spatially distinct, compact and stable protein structural unit that could conceivably fold and function in isolation [113]. Sequence-based domain definitions often define domains as distinct regions of protein sequence that are highly conserved throughout evolution. The sequences carrying these domains are described as sequence homologs and are often present in different molecular contexts [114].

As the conservation of these protein domains is higher than the conservation of complete protein sequences, the detection of conserved protein domains in a novel protein sequence can help to elucidate its function. The use of domains is more sensitive than sequence similarity based on the complete protein sequence (Figure 1.15) and can even characterize novel protein sequences with no known counterparts in other genomes but with protein domains of known functions.

Domains on the sequence level can be described as Hidden Markov Models (HMMs), sequence motifs or patterns. InterPro [115] is a database integrating the contents of various databases representing the domains in different ways. It contains domain signatures from PROSITE [116], PRINTS [117], Pfam [118], ProDom [119], SMART [120], TIGRFAMs [121], PIRSF [122], SUPERFAMILY [123], PANTHER [124], Gene3D [125], and

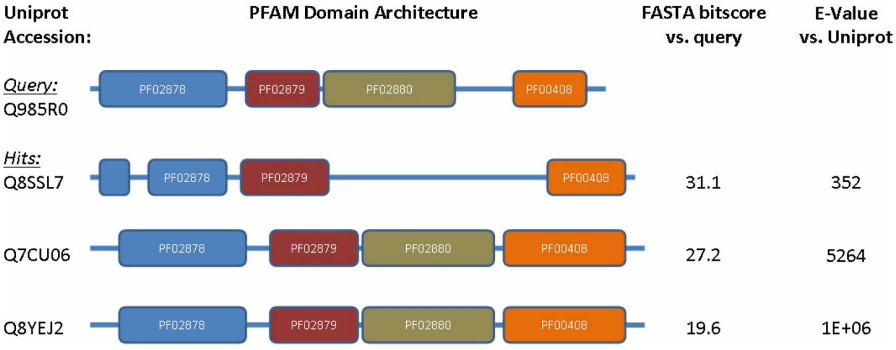


Figure 1.15: Example of remote homologs retrieved by the "Domain similarity" tool of SIMAP When searching the query sequence in the UniProtKB database, high E-Values result from low bitscores. Thus, these proteins show insufficient pair wise sequence homology to the query and would not be found by database searches which are typically restricted to a maximal E-Value of 10. However, the similar domain architectures suggest a common ancestry of these proteins. (Adapted from [128])

HAMAP [126]. When signatures from different member databases of InterPro match the same set of proteins in the same region on the sequence, it is assumed that they describe the same functional family, domain or site and are placed into a single InterPro entry.

All InterPro domains for all sequences are available precalculated in the SIMAP [94] database.

1.3.3.2 Domain homology

There can be cases identified where protein sequences sharing almost no similarity on the sequence level nevertheless share a common or very similar protein domain structure (Figure 1.15). Therefore Lin et al. [127] developed a possibility to compare the domain structures of proteins. The number of shared domains between two proteins, the domain order and domain duplications are incorporated into a similarity score.

The use of domain homology allows to detect even remote homologs lacking sufficient sequence similarity on the sequence level using BLAST or FASTA.

1.3.4 SIMAP

The annotation by sequence as well as domain similarity and many other methods in computational biology rely on the analysis and comparison of protein sequences and therefore perform similarity searches using BLAST or FASTA. But even if BLAST and FASTA are faster than the Smith-Waterman algorithm the increasing volume of publicly available protein sequences produces a problem for all analyses relying on a complete all-against-all comparison between all sequences. Therefore SIMAP [94] was set up and aims to provide the automatically incrementally computed similarities between all publicly available protein sequences. SIMAP allows to retrieve similarities very efficiently using

EJBs or Web Services and allows for analyses that would not be possible without a complete similarity matrix between all proteins. One example for such an analysis is the large scale clustering of the protein sequences.

1.4 Horizontally transferred genes

1.4.1 Definition and importance

Horizontally transferred genes (HTGs), sometimes referred to as laterally transferred genes, are genes in an organism that originate not by vertical inheritance from the ancestor of the organism but by a horizontal gene transfer (HGT) event from a possibly not directly related organism.

HGTs are fundamental for the rapid adaptation of prokaryotic genomes to changing environmental conditions [129]. They are quite common in pathogens and responsible e.g. for acquiring resistance against antibiotics (e.g. [130]). Additionally the knowledge about the origin of HTGs allows to draw conclusions about the natural habitat and the other organisms living closely together with the organism.

1.4.2 Detection of horizontally transferred genes

1.4.2.1 Concepts

A horizontally transferred gene of an organism is characterized by a taxonomic tree for the gene that differs from the taxonomic tree for the whole organism. Figure 1.16 shows an example in which a bacterial gene is transferred into metazoa. This difference between the trees can be used for the detection of HTGs.

The easiest and most commonly applied detection method for HTGs is the analysis of the taxonomic distribution of the best hits of a protein sequence against a non-redundant database of protein sequences. If the closest hits of the protein sequence originate from organisms that are not directly neighbored in the species tree, then this protein is a potential HTG.

As it has been shown that the closest BLAST hits are often not the nearest neighbors of a protein [131] due to differences between the pairwise alignments used for BLAST and the multiple alignments used for the computation of trees, another possibility is the direct comparison of the phylogenetic gene trees with the species trees.

In order to be able to automatically analyze the taxonomy an electronic taxonomy concept is necessary. The NCBI taxonomy is such a taxonomic hierarchical tree structure containing all organisms represented in the genetic databases of NCBI with at least one nucleotide or protein sequence (`http://www.ncbi.nlm.nih.gov/taxonomy`).

1.4.2.2 Alien Index

The Alien Alien Index (AI) introduced by Gladyshev et al. [132] analyzes the taxonomic distribution of the best hits of a protein sequence.

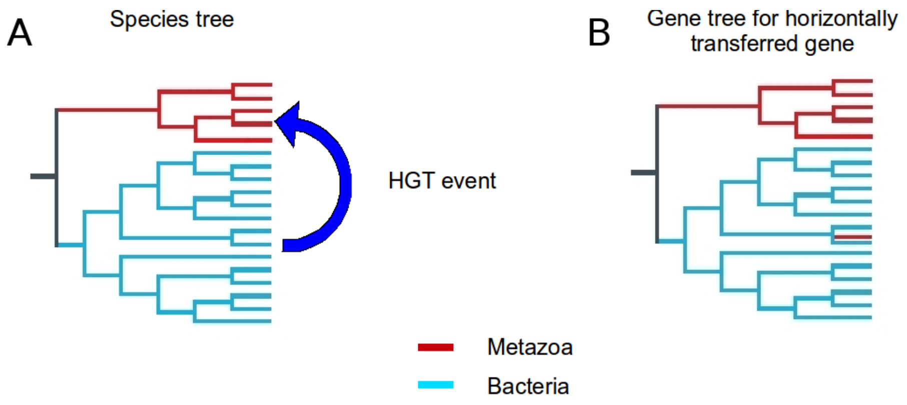


Figure 1.16: Discrepancy between the species tree and the gene tree of a horizontally transferred gene A: Species tree for *Bacteria* (blue) and *Metazoa* (red). The two groups are clearly separated in the tree. A single gene is transferred by a horizontal gene transfer (HGT) from the *Bacteria* to the *Metazoa*. **B:** The gene tree for the horizontally transferred gene shows that the former *Bacteria* gene, which is now a *Metazoa* gene, groups with the *Bacteria* genes. This is a discrepancy between the species tree and the gene tree. (Thanks to Thomas Weinmaier for providing the figure)

HTG candidates are detected as follows:

1. A homology search for the protein against all publicly available sequences of cellular organisms is performed. Only hits fulfilling an E-value threshold are kept.

2. The homologs are grouped according to their taxonomy into ingroup and outgroup hits. The outgroup consists of organisms that are potential donors of the HTG.

3. The AI is then calculated as:

$$AI = \ln((\text{best E-value ingroup}) + 10^{-200}) - \ln((\text{best E-value outgroup}) + 10^{-200}) \tag{1.5}$$

4. An AI > 0 indicates that the protein shows higher similarity to the outgroup than to the ingroup and an AI < 0 indicates that a protein shows higher similarity to the ingroup than to the outgroup. The higher the AI, the bigger is the difference between the best E-value of the outgroup and the ingroup.

   Proteins with an AI ≥ 30 (based on experience from other projects) are considered HTG candidates.

1.4.2.3 PhyloGenie

PhyloGenie [133] compares phylogenetic gene trees with species trees. The PhyloGenie [133] pipeline is a method that automates sequence selection, alignment, and the computation, phylogenetic inference and analysis of the trees. Starting from a set of created trees it is possible to identify trees that match specific topological constraints. PhyloGenie can therefore be used for the detection of HGTs but is not restricted to it.

1.5 Conserved neighborhood

It has been observed that genes occurring repeatedly in each other's proximity on genomes tend to encode functionally interacting proteins [134, 135, 8, 136, 7, 137]. Even divergently oriented gene pairs show to be indicative of functional linkage with somewhat lower confidence [138].

An explanation for this observation is that functionally associated proteins need to be maintained and regulated in the genome together. Therefore they share the same selection pressures as they need to interact with each other in order to build complexes or to facilitate complex functions. This selective pressure leads to joint transfers of genes between genomes [139, 140], concerted gene loss [141], gene fusion events [142], coregulation of genes through common regulatory elements [143], and the creation and maintenance of operons containing nonhomologous but cotranscribed genes [144, 145].

The interactions between the genes are called “protein-protein-interactions”. Likely interacting proteins can be predicted by using the concepts of conserved neighborhood, gene fusion events and cooccurrence of proteins (see [146]).

Scientists working with prokaryotic genomes have used conserved genomic neighborhood of genes for a long time to infer functional linkage, assuming that such arrangements reflect polycistronic transcription units (operons) [147]. The ChlamydiaeDB introduced in this work allows the display of the genomic neighborhood of genes as this is the most straightforward and important of the above concepts for a comparative genome database.

1.6 Bacterial secretion

The understanding of bacterial secretion played a special role in this work as the existence or non-existence of secretion systems gives insights into the biology of an organism. For example, many Gram-negative bacteria live in close association with humans, animals, or plants. These pathogenic or symbiotic interactions between bacteria and host are often mediated or even made possible by the secretion of bacterial effector proteins into the host cells (e.g. in some members of the genus *Rhizobium* [148, 149, 150]).

1.6.1 Bacterial secretion systems

Secretion systems are specialized systems that facilitate the transport of effector proteins to the bacterial supernatant or host cell cytoplasm (for a review see [151]). Gram-negative as well as Gram-positive bacteria have evolved these systems that play an important role in the virulence of bacterial pathogens [151, 152, 153, 154, 155, 156, 157, 158] as often secreted or surface-exposed bacterial proteins play important roles in the interaction of the pathogens with their host cells.

In Gram-positive bacteria four different kinds of pathways are mainly responsible for the protein export from the cytoplasm. The largest number of proteins is translocated using the general protein secretion (Sec) pathway [159]. The twin-arginine translocation

(Tat) pathway translocates folded proteins containing a highly conserved twin-arginine motif in their signal peptide [159]. A sec-independent pathway consisting of type IV prepilin-like proteins transports another class of proteins in *B. subtilis* [159]. Additionally there exist ATP-binding cassette transporters that can be regarded as "special-purpose" pathways, through which only a few proteins are transported [159].

There are seven major secretion systems known in Gram-negative bacteria today [160, 161, 162] (see Figure 1.17 for an overview). The Type I, III and VI secretion are independent of the "general secretory pathway" (sec-pathway). The Type I secretion systems predominantly secrete toxins, proteases and lipases into the extracellular milieu. The Type III secretion systems secrete toxins, proteases, lipases and virulence proteins into the host cell. The Type VI secretion is often involved in interaction with eukaryotic hosts. The Type IV secretion system mediates the secretion of single proteins, protein-protein complexes and protein-DNA complexes across the double membrane to bacterial or eukaryotic cells, requiring direct cell-to-cell contact. It consists of 12 parts that build a needle that can inject these molecules into the host cell. Autotransporters (Type V secretion system) form a pore through the outer membrane. The chaperone/usher pathway consists of an outer membrane protein, termed an usher, and a periplasmic chaperone, guiding proper folding and preventing premature interactions. The Type II secretion system consists of complexes of 12-16 proteins that mediate the transport of extracellular enzymes and toxins: a pilus-like structure of four inner membrane proteins pushes the proteins, delivered to the periplasm by the Sec or Tat system, through an outer membrane pore.

1.6.2 Prediction of Type-III secreted effector proteins

The Type III secretion system (TTSS) is one of the best studied cellular machineries for the secretion of proteins and despite of the importance of this system for bacterial pathogenesis, recognition and targeting of type III secreted proteins is only poorly understood.

Methods for the experimental identification of effectors rely on translocation assays using fusion proteins of a putative effector with a reporter gene [165, 166, 167, 168] or detection of effectors in the culture supernatant [165]. As effector screens by fusion experiments are intractable for all genes of an organism bioinformatics plays an important role in providing candidate lists of putative effector proteins [165, 169, 170].

There have different bioinformatics methods been applied in order to limit the amount of candidates for experimental analyses in the past. These methods include homology to known effector proteins [165], chromosomal co-localization of putative effectors with TTSS related chaperons [171], common transcriptional regulation of effector proteins with elements of the TTSS [167, 172], and an unusual amino acid composition in the N-termini of effectors [172, 170, 173]. Nevertheless, none of these methods is either exhaustive or generally applicable.

Therefore an approach that is exhaustively and generally applicable is desirable. The straightforward way to detect novel effector proteins would be the identification of a general molecular signal which leads to specific recognition of effector proteins by the

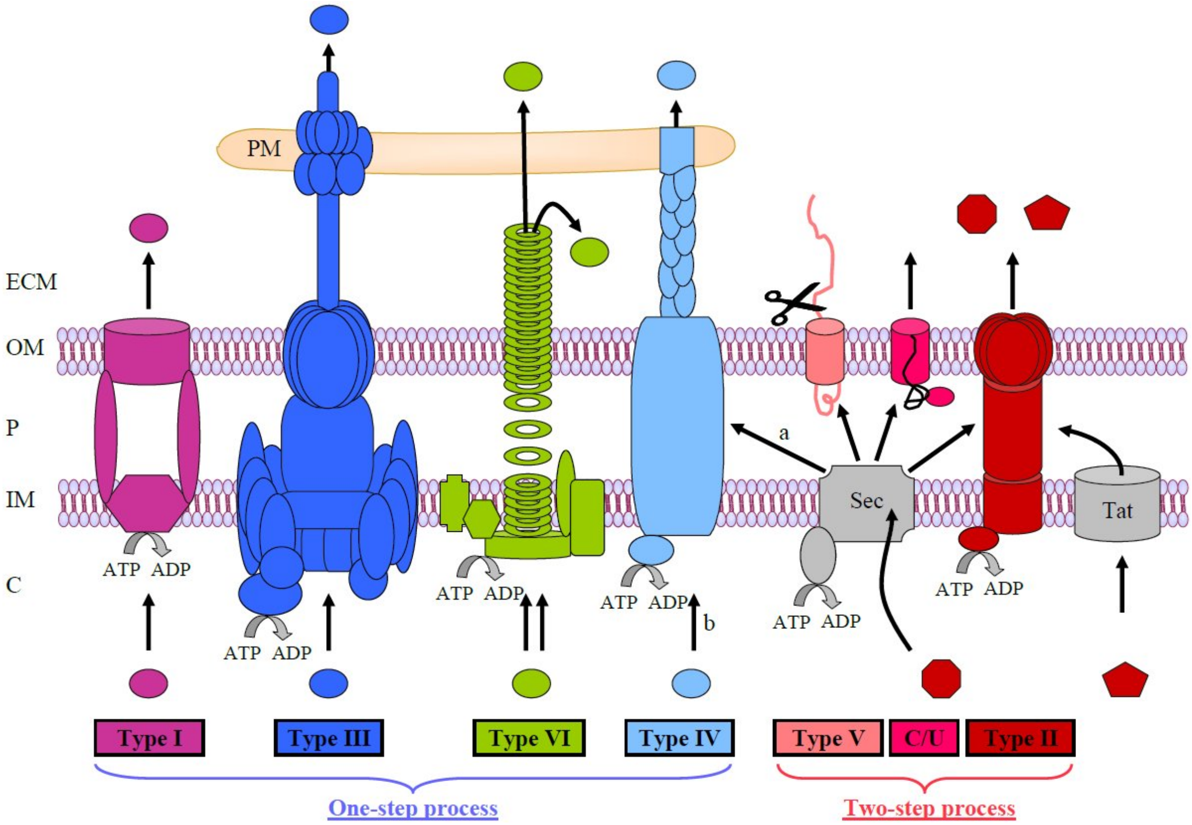


Figure 1.17: Major protein-secretion systems in Gram-negative bacteria Four of these protein secretion pathways depend on the Sec system for protein transport across the inner membrane. Type I, III and VI secretion are Sec-independent. **Type I secretion systems** predominantly secrete toxins, proteases and lipases into the extracellular milieu, whereas **Type III secretion systems** also translocate virulence proteins into the host cell. **Type VI secretion** mediated by a novel kind of a complex multi-component secretion machine, is often involved in interaction with eukaryotic hosts. **Type IV secretion systems** mediate the transport of DNA and proteins across the double membrane to bacterial or eukaryotic cells, requiring direct cell-to-cell contact. Examples are the Bordetella pertussis toxin (Sec-dependent, a) or the Agrobacterium tumefaciens VirB/D4 system for transport of T-DNA-protein complexes (Sec-independent, b). **Autotransporters** form a pore through the outer membrane and are therefore classified among the **Type V secretion systems**. The **chaperone/usher pathway (C/U)** consists of an OM protein, termed an usher, and a periplasmic chaperone, guiding proper folding and preventing premature interactions. Complexes of 12-16 proteins mediate transport of extracellular enzymes and toxins in **Type II secretion**: a pilus-like structure of four IM proteins pushes the proteins, delivered to the periplasm by the Sec or Tat system, through an OM pore. Legend: C, bacterial cytoplasm; IM, bacterial inner membrane; P, bacterial periplasm; OM, bacterial outer membrane; ECM, extracellular milieu. PM, host cell plasma membrane. When appropriate, coupling of ATP hydrolysis to transport is highlighted. Arrows indicate the route followed by transported proteins. Adapted from [163, 164]. (Source: [151])

TTSS. As the N-termini of known effectors are very diverse and show no apparent evolutionary conservation between different effectors [174] classical bioinformatics approaches as deriving sequence motifs are not applicable to model the signal. Therefore machine learning techniques were used for the deduction of properties that distinguish secreted from non-secreted proteins. There is no prior knowledge about the mechanism necessary and the derived features can be used to train a binary classifier that predicts whether an unknown protein sequence is likely to be predicted or not.

Three recent approaches utilize binary classifiers to predict effector proteins.

Löwer et al. [175] trained a neural network with string representations of the amino acid composition of the 30 N-terminal residues of effectors from various studies.

Samudrala et al. [176] 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. The study uses a support vector machine and indicates a common signal in the two used organisms *P. syringae* and *Salmonella typhimurium.*

Arnold et al. [177] implemented the method EffectiveT3 taking into account a representation of the first 25 residues, which includes amino acid frequencies but also frequencies of certain amino acid properties and small combinations of them, as short stretches of hydrophobic residues. The most discriminating features have been extracted using a greedy hill-climbing search with correlated feature selection. A naïve Bayesian classifier is used. Effectors with specific experimental evidence originating from various organisms have been used and the taxonomic generality of the signal is shown. Arnold et al. provide a standalone version of the prediction so that it is applicable also to unpublished data on large scale.

1.6.3 Proteins carrying eukaryotic like protein domains

Some other secretion systems do also use signal peptides. An example is the Type IV secretion system for which our group could show that the C-terminal sequences of Type IV effectors are suited for the discrimination between secreted and non-secreted proteins.

Another way how to predict effectors, no matter by which secretion system they might be transported, is the search for protein domains related to virulence. Proteins carrying eukaryotic like protein domains are of special interest as they could be able to mimic and alter functions in the host cell [178] and therefore play an important role in virulence.

Eukaryotic like protein domains are domains occurring mainly in eukaryotes but also occurring in prokaryotes. The "interesting" domains are those that occur more frequently in pathogenic than in non-pathogenic bacteria (see also the Effective database [179]). The fact that a protein carries such a domain allows to characterize it as likely to play a role in the interaction with the host.

1.7 Methods for automatic genome analysis

1.7.1 Text-Mining

The number of available databases containing biomedical data and the data therein has been constantly growing over the last years. The largest fraction of biomedical knowledge is unfortunately not contained in specialized databases where the information can be easily retrieved and for example used with computers but contained in the literature. As the biomedical literature is growing at a exponential pace [180] it is no longer possible for a researcher to keep up-to-date with all the relevant literature manually [181]. Therefore computer supported text mining systems are essential for speeding up the information retrieval from the literature and the automatic extraction of useful information.

A text can be analyzed at three different levels: The lexical level consists of the words, also called "lexemes". This is the most basic level. The next higher level is the syntax level consisting of the grammar or syntax of a language. The syntax defines the positioning of words and also contains the classification of words into different lexical categories like nouns, verbs, adjectives or adverbs. The highest level is the semantic level. This level is the meaning of the sentence depending on the context (knowledge domain). As some words will have different meanings depending on the context, the context has to be taken into account when analyzing sentences. As example “date” may be an appointment in the business context or may be a fruit in the food context.

The most important source of biomedical texts is MEDLINE/PubMed. It consists of approximately 20 million citations for biomedical literature from MEDLINE, life science journals, and online books from the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and preclinical sciences (`http://www.pubmed.org`. PubMed is developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH). Academic users can retrieve a free license of PubMed/MEDLINE XML files and gain full access to all citations in PubMed with several updates every week.

1.7.2 Genome annotation pipelines and on-line resources

Many of the standard analyses like the detection of the best hits in public sequence databases and function annotations like the assignment of UniProtKB/Swiss-Prot keywords [182], Gene Ontology (GO) [92] and FunCat [81] are daily business and need to be repeated for every genome project. Therefore the automation of these tasks is desirable.

Probably the first genome analysis system was the predecessor of GeneQuiz [183]. Later other systems have been developed, among them MAGPIE [184], PEDANT [72], Genotator [185] and AceDB [186]. They also introduced the first online genome databases that could be accessed over the spreading internet.

Databases as WIT/PUMA [187, 188], KEGG [189] and MetaCyc [190] arose for the purpose of metabolic reconstruction from genome data.

With the availability of complete eukaryotic genomes the next generation of web portals was made available that integrate all available information for the genomes. Examples are the UCSC Genome Browser [191] and Ensemble [192].

Today there exist many different annotation systems in parallel. They differ from each other with respect to used software technology, type of data, organism types, scientific questions they are designed to answer and target group [4]. While the Swiss-Prot team uses the annotation pipeline HAMAP [193] for entire microbial proteomes for example, there also exist systems for small research groups like GANESH [194].

It is even possible to upload complete genome sequences to online annotation services that provide the user with an automatic annotation of the sequence (JCVI Annotation Service (J. Craig Venter Institute) `http://www.jcvi.org/cms/research/projects/annotation-service/`, BASys server `http://basys.ca/` [195]).

In order to make the annotation pipelines very easily customizable some of the major genome analysis systems, such as PEDANT, have been equipped with workflow-based process management. This allows to create standard workflows that can be executed for every genome in the same way, but is still open to extensions or specific settings.

1.7.3 Data from experiments beyond the genome sequence

Bioinformatics is important when handling and processing data from experiments beyond the genome sequences. On the one hand hypotheses can be generated by bioinformatics, on the other hand it can assist in gaining knowledge from data from experiments. This data plays an important role for the characterization of regulation, transcription, translation and characterization of gene products of an organism and is therefore essential for the understanding of an organism.

1.7.3.1 Single Nucleotide Polymorphisms (SNPs)

The advent of next-generation sequencing technologies allowed the unexpensive resequencing of whole bacterial genomes and the detection of polymorphisms between populations (e.g. [196]). The degree of genetic variation between the isolates or populations can be measured and it can be searched for correlations between the variations and specific traits.

A single nucleotide polymorphism (SNP) is a difference between two DNA sequences in one nucleotide. SNPs can be contained within coding regions or in intergenic regions. The SNPs in coding regions can be divided into non-synonymous SNPs if they alter the amino acid sequence of the resulting protein or synonymous SNPs if they do not alter the amino acid sequence of the resulting protein. Not all detectable SNPs may result in phenotypic traits but may just be different genotypes.

1.7.3.2 Transcript data

The transcriptome of an organism is the sum of all RNAs including mRNAs and other types of RNAs. The knowledge about the expression levels of genes under certain conditions helps to gain knowledge about the function of gene products, e.g. genes involved in the reaction to heat stress. The availability of transcript information is therefore important for the precise individual biological interpretation of the functions of genes. Additionally transcript data provides means to identify which of the predicted genes are actually expressed.

The knowledge about the differential expression of genes can even be used for the determination of biomarker candidates for a specific condition. These candidates need to be checked in the laboratory.

In order to evaluate gene expression patterns transcriptome studies using microarrays [197] and reverse transcription PCR [198] can be applied. In order to perform these studies a priori knowledge in the form of the predicted CDSs has been used.

Small regulatory RNAs (sRNAs) have been detected by computational predictions combined with experimental verification [199, 200, 201]. Other ways to identify new sRNAs are by cDNA cloning of small-sized RNA species [202, 203], detection on tiling arrays [204, 205, 206, 207], and the co-precipitation of sRNAs with Hfq, a conserved sRNA-binding protein in bacteria [208], and the subsequent identification of Hfq-associated transcripts on whole genome microarrays [209] or by deep sequencing of cDNA [210].

Recently it could be shown that deep sequencing approaches of RNA are feasible and able to provide an unbiased picture of the RNAs within a cell [211, 212, 213].

Transcription data therefore can contain information about previously missed coding genes, gene structure, gene expression patterns, and non-coding RNAs. This is even possible in genetically inaccessible organisms such as obligate intracellular bacteria [5].

1.7.3.3 Proteome data

The proteome is the complete set of proteins of an organism. In order to be able to describe structure, function and control mechanisms of a biological system the knowledge about the proteome is essential as proteins are involved in almost all biological activities.

Proteins are almost always the effectors of biological functions. The protein levels do not only depend on the levels of the transcripts of the genes but also on a host of translational controls and regulated degradation [214, 215]. The data set comprehensively characterizing a biological system is therefore the expression level of all proteins.

The knowledge of existence and expression levels of specific proteins can be used to identify putative virulence factors and to gain insights into the physiology and metabolic versatility of an organism [51]. Additionally the analysis of the proteome can contribute to the detection of diagnostic biomarkers, construction of vaccines or the development of novel antimicrobial therapies [216, 217, 218].

Biological mass spectrometry (MS) in combination with protocols to handle small amounts of biological samples, the ability to rapidly identify peptides by matching their MS fragmentation spectra to sequence databases, and the direct analysis of very complex protein mixtures [219, 220] make it a suited solution for the analysis of proteomes.

1.7.4 Enrichment of annotations and protein domains

In many experiments groups of genes can be identified that share a specific property under a specific condition. Examples are genes that are overexpressed under a specific condition or genes containing SNPs.

The question which kinds of proteins with which functions are involved in the reaction to the condition is addressed by looking at the available annotations of these genes. Even if specific functions can be detected in this set of proteins it is not clear yet whether this is special when compared to other conditions or to the annotations of the rest of the proteome.

In order to analyze which properties distinguish one protein set from the other all annotations in both sets need to be compared to each other. For each of these annotations it can be counted how many proteins in the first group have the feature, how many proteins in the first group do not have the feature, how many proteins in the second group have the feature, and how many proteins in the second group do not have the feature. In order to be able to distinguish between significant and non-significant differences between the two sets a statistical test using these four numbers and producing some significance measure is necessary. The significance of the difference between the two sets can be assessed with a variety of statistical tests including Pearson's chi-square test, the G-test, Fisher's exact test, and Barnard's test. Fisher's exact test [221, 222] is a robust test that is also suited for small sample sizes, that is if only a small number of proteins has a specific annotation in one of the sets. The test returns the significance of the deviation from the null hypothesis, that there can be no enrichment or depletion for an annotation be found. As it is desired to test the significance of enrichment and depletion the two-tailed test needs to be applied.

In order to account for the problem of multiple testing, the p-Value obtained from the two-tailed Fisher's exact test needs to be corrected. The Bonferroni correction is probably the most commonly used correction for multiple testing. It is the multiplication of the p-Value with the number of performed tests.

1.7.5 Submission of sequences to public databases

After the annotation has been done the newly gained knowledge is often published in scientific journals. If the paper contains novel sequences it is necessary to submit these sequences to either the DNA Data Bank of Japan (DDBJ) (`http://www.ddbj.nig.ac.jp`), the EMBL Nucleotide Sequence Database (EMBL-Bank) (`http://www.ebi.ac.uk/embl`), or the GenBank database (`http://www.ncbi.nlm.nih.gov/genbank/`) prior to submission of the paper. These three institutions take part in the International Nucleotide Sequence Database Collaboration (INSDC). If the publication contains a novel genome it is necessary to register the genome project.

The consortium provides the infrastructure for the registration of DNA sequences and genome projects. This can be done at each of the member databases. A locus_tag prefix can also be registered at this time (`http://www.ebi.ac.uk/embl/Documentation/locus_tag_usage.html`). This prevents the case that locus_tags in two genomes have the same name. Locus_tags should be assigned to all protein coding and non-coding genes such as structural RNAs and consist of a prefix of at least 3 characters length and a tag value separated by an underscore. An example is the locus_tag Ctu_0010. "Ctu" stands for *Cronobacter turicensis* and 0010 is the sequential number of the genetic element on the chromosome. Locus_tags are generally in sequential order on the genome, but it is allowed to leave gaps when initially assigning locus_tags and fill in new annotation with tag values that are between the gaps.

A submission consists of a file either in GenBank or EMBL format that contains the annotations of all protein coding and structural RNAs. If a description for the coding elements is assigned it is also necessary to provide information what the reference for this annotation transfer is. Therefore a typical workflow involves the sequencing of the genome, the registration of the genome project, registration of the locus_tag prefix, prediction of protein coding regions and structural RNAs, the naming of the genetic elements with locus_tags, the transfer of descriptions by sequence homology searches against a trusted set of proteins (e.g. UniProtKB/Swiss-Prot), the creation of the submission files and the upload of the created files.

1.7.6 Main organisms occurring in this work

1.7.6.1 *Cronobacter turicensis*

Cronobacter spp. (*Enterobacter sakazakii*) are Gram-negative opportunistic, foodborne, pathogenic bacteria of the family *Enterobacteriaceae*. They are known as rare but important cause of life-threatening neonatal infections as brain abscesses, meningitis, necrotizing enterocolitis and systemic sepsis [223, 224] with fatal mortality rates varying from 40 to 80% [225]. Neonates and infants under 2 months, born prematurely or with low birthweight (<2500 g) suffer from the highest infection risk [224] most commonly by *Cronobacter* sp. contaminated powdered infant milk formulas. *Cronobacter* sp. is remarkably resistant to desiccation, osmotic stress and is able to survive up to two years in milk powder [226, 227, 228].

Up to now little is known about the mechanisms of pathogenicity in *Cronobacter* sp. It is a common trait of microbial pathogens to express adherence factors responsible for recognizing and binding to specific receptor moieties of cells, thus enabling the bacteria to resist host strategies that would impede colonization. Up to now, only few studies have described aspects of the interaction of *Cronobacter* sp. with human cells [229, 230, 231, 232, 233]. Several putative virulence factors involved in adhesion, invasion and biofilm formation, iron acquisition, protection against reactive oxygen species and protein secretion and transport mechanisms were recently described in *Cronobacter turicensis* using a proteomic approach [51].

The genome sequence of the strain *Cronobacter turicensis* LMG 23827 was used in this work. This strain caused the death of two new-born children in a Children Hospital in Zürich in 2005.

1.7.6.2 *Chlamydiae*

The members of the phylum *Chlamydiae* (from the Greek, $\chi\lambda\alpha\mu\epsilon\sigma$ meaning "cloak") [234] are obligate intracellular bacteria that show a broad host spectrum (Table 1.18) and are major pathogens of humans. *Chlamydiae* have a characteristic developmental cycle consisting of two states, the metabolically inert elementary bodies (EBs) and the actively dividing reticulate bodies (RBs), existing in a host-derived vacuole termed inclusion [235].

At the moment there are 16 completely sequenced genomes available for the phylum *Chlamydiae*: six members of the species *Chlamydia trachomatis*, one member of the species *Chlamydia muridarum Nigg*, one member of the species *Chlamydophila felis*, one member of the species *Chlamydophila caviae*, one member of the species *Chlamydophila abortus*, four members of the species *Chlamydophila pneumoniae*, one member of the species *Waddlia chondrophila*, and one member of the species *Candidatus Protochlamydia amoebophila*.

The two most studied chlamydial species are *Chlamydia trachomatis* and *Chlamydophila pneumoniae* and are responsible for several severe diseases in humans.

Chlamydia trachomatis causes trachoma, an infectious eye disease, that affects about 84 million people, of whom about 8 million are visually impaired as a consequence [236]. Additionally they are the most common cause of sexually transmitted diseases, with over 90 million new cases each year [237].

Chlamydophila pneumoniae is another member of the phylum *chlamydiae* and is a causative agent of pneumonia, which has also been associated with a number of chronic diseases such as atherosclerosis, asthma, and Alzheimer's disease [238].

The ability to specifically inactivate and reactivate a single gene is central to show gene functions [239], e.g. in knockout experiments. The developmental cycle of *Chlamydiae* poses obstacles in generating the tools needed to perform these genetic analyses and to define the genes that are important to the biology, pathogenicity, or transmission of Chlamydia [240]. Therefore it is not possible to genetically manipulate *Chlamydiae* e.g. by transformation using circular plasmids that can be easily manipulated like in *Escherichia coli* [241]. Therefore bioinformatics plays an essential role in the research about *Chlamydiae*.

There is an ongoing discussion about the division of the family *Chlamydiaceae* into the two genera *Chlamydia* and *Chlamydophila* as proposed by Everett et al. 1999 [242]. In this work the NCBI taxonomy is used, that is also used in GenBank. As NCBI also distinguishes *Chlamydia* and *Chlamydophila* it is consistently also distinguished in this work.

1.7.7 Databases specific for genomes of the phylum *Chlamydiae*

The general repositories containing the DNA sequences of *Chlamydiae* are Genbank, EMBL and DDBJ.

There exist several databases that provide specific information for a specific member of the phylum *Chlamydiae*:

		Chlamydia trachomatis	*Chlamydia muridarum*	*Chlamydia suis*	*Chlamydophila abortus*	*Chlamydophila caviae*	*Chlamydophila felis*	*Chlamydophila pecorum*	*Chlamydophila pneumoniae*	*Chlamydophila psittaci*	"*Candidatus* Clavochlamydia salmonicola"	*Neochlamydia hartmannellae*	*Neochlamydia* sp.	*Parachlamydia acanthamoebae*	*Parachlamydia* sp.	*Protochlamydia amoebophila*	*Protochlamydia naegleriophila*	"*Candidatus* Fritschea eriococci"	"*Candidatus* Fritschea bemisiae"	"*Candidatus* Piscichlamydia salmonis"	*Simkania negevensis*	*Waddlia chondrophila*	*Waddlia malayensis*	"*Candidatus* Rhabdochlamydia crassificans"	"*Candidatus* Rhabdochlamydia porcellionis"	*Rhabdochlamydia* sp.	Unidentified *Chlamydiae*
Vertebrates																											
Mammals	**Human** (*Homo sapiens*)	■							■					▨			■				■	▨				▨	
	Cat (*Felis silvestris catus*)						■						▨														
	Pig (*Sus* sp.)			■	■			■							▨												
	Cattle (*Bos taurus*)				■			■		▨					■							■					■
	African buffalo (*Syncerus caffer*)									■																	
	Water buffalo (*Bubauis* sp.)									■																	
	Chamois (*Rupicapra rupicapra*)							■																			
	Sheep (*Ovis aeries*)				■			■																			
	Arabian oryx (*Oryx leucoryx*)									▨																	
	Blackbuck (*Antilope cervicapra*)																										▨
	Fallow deer (*Cervus dama*)									▨																	
	Fallow deer (*Dama dama*)																										▨
	Red deer (*Cervus elaphus*)																										▨
	Reindeer (*Rangifer tarandus*)																										▨
	Mule deer (*Odocoileus hemionus*)																										■
	Mouflon (*Ovis mussimon*)																										▨
	Spanish ibex (*Capra pyrenaica*)																										▨
	Goat (*Capra aegagrus hircus*)				■			■																			
	Mouse (*Mus musculus*)		■																								
	Hamster (*Mesocricetus auratus*)		■																								
	Horse (*Equus caballus*)								■	■																	
	Guinea pig (*Cavia porcellus*)					■																					
	Fruit bat (*Eonycteris spelaea*)																						■				
Marsupials	**Koala** (*Phascolarctos cinereus*)							■	■																		
	Great glider (*Petauroides volans*)							▨																			▨
	Mountain brushtail possum (*Trichosurus caninus*)							▨																			▨
	Greater gilby (*Macrotis lagotis*)														▨												▨
	Western barred bandicoot (*Perameles bouganville*)							▨																			▨
	Gilbert's potoroo (*Potorous gilbertii*)																					▨					
Amphibians	**Great barred frog** (*Mixophyes iteratus*)								■																		
	African clawed frog (*Xenopus tropicalis*)								■																		
	Blue Mountains tree frog (*Litoria citropa*)								■																		
	Common frog (*Rana temporaria*)			▨	▨				▨																		
	African clawed frog (*Xenopus laevis*)			▨																							▨
Birds	>100 species									■																	
	Chicken (*Gallus gallus*)									■																	▨
Reptiles	**Green sea turtle** (*Chelonia mydas*)				■				■				■														
	Burmese python (*Python mulurus bivittatus*)				▨				▨																		
	Puff adder (*Bitis arietans*)				■				■																		
	Snake (unspecified)								■																		▨
	Chelonian (unspecified)								■																		▨
	Lizard (unspecified)																										▨
	Chameleon (*Chameleo dilepis*)								■																		
	Iguana (*Iguana iguana*)						■		■																		
	Nile crocodile (*Crocodylus niloticus*)									■																	
Fish	**Atlantic salmon** (*Salmo salar*)										■									■							
	Wild trout (*Salmo truttar*)										■																
	Leafy seadragon (*Phycodurus eques*)																									■	
	Silver perch (*Bidyanus bidyanus*)												■														
	Barramundi (*Lates calcarifer*)																										■
	Arctic charr (*Salvelinus alpinus*)												■														
Invertebrates																											
Insects	**Sweetpotato whitefly** (*Bemisia tabaci*)																		■								
	Imported elm bark louse (*Eriococcus spurius*)																	■									
	Oriental cockroach (*Blatta orientalis*)																							■			
Crustaceans	**Rough woodlouse** (*Porcellio scaber*)																								■		
Molluscs	**Pacific oyster** (*Crassostrea gigas*)				▨																						
Protozoa	*Acanthamoeba* sp.													■		■											
	Hartmannella vermiformis											■															
	Naegleria lovaniensis																■										

Figure 1.18: Host range of the phylum *Chlamydiae* Evidence for the presence of *chlamydiae* by 16S rRNA analysis in combination with microscopic analyses (immunofluorescence, electron microscopy, histology) or for the recovery of the respective organism is indicated by dark blue boxes. Evidence for the presence of *chlamydiae* by only 16S rRNA analysis or serology without microscopic data is indicated by light blue boxes. Due to the revision of the chlamydial taxonomy in 1999 [242], evidence for *Chlamydophila psittaci* (formerly *Chlamydia psittaci*) could also refer to *Chlamydophila abortus*, *Chlamydophila felis*, or *Chlamydophila caviae*. (Source: [243])

The resource for *Chlamydophila pneumoniae* J138 at http://kantaro2.grt.kyushu-u.ac.jp/microb/J138/ provides the genome sequence, genes and groups with determined function for *C. pneumoniae* J138.

The Proteome 2D-PAGE Database [244] provides proteomics data for *C. pneumoniae* from two studies at http://web.mpiib-berlin.mpg.de/cgi-bin/pdbs/2d-page/extern/menu_frame.cgi.

The genome database for *Protochlamydia amoebophila* UWE25 at http://mips.gsf.de/genre/proj/uwe25/ provides various information like InterPro protein domains, taxonomy of the closest homologs, structure predictions, cluster memberships of proteins, manual functional annotations, and much more.

The Chlamydia Interactive Database (CIDB) [245] at http://www3.it.deakin.edu.au:8080/CIDB/ contains five types of data sources:

- Quantitative RT-PCR expression data for 66 genes from two developmental time points (24 and 48 h) under both normal growth conditions [246] and also under gamma-interferon induced persistence conditions [247].

- Microarray gene expression profiles for *C. trachomatis* serovar D [197] which contains 901 gene expression patterns for six time points (1, 3, 8, 16, 24 and 40 h). Microarray gene expression profiles for *C. trachomatis* L2 [198], which contain microarray data for 890 genes at two time points (24 and 48 h).

- Promoter data which includes a list of genes for which the promoters have been predicted [248, 247]. These are arranged into three categories (sigma-66, sigma-54 and sigma-28).

- Proteomic data for 14 genes which are arranged into three categories (up-regulated, down-regulated and unchanged) [248].

- Genomic data is accessed via the Berkeley Genome web site (http://chlamydia-www.berkeley.edu:4231) and enables provision of the predicted gene function and the gene arrangement maps.

Unfortunately it seems not to be maintained anymore and is not any longer available.

The Predicted Chlamydia Outer Membrane Proteins (pCOMP) database [249] contains a collection of predicted chlamydial outer membrane proteins for *Chlamydia trachomatis*, *Chlamydia muridarum*, *Chlamydophila pneumoniae*, *Chlamydophila caviae*, and *Protochlamydia amoebophila*.

There exist two more comprehensive databases, allowing for comparisons between genomes:

The Genome Information Broker [250] at http://gib.genes.nig.ac.jp provides GC-Plots, sequence similarity searches, full text searches, exploration of functional categories and display of members as well as codon usage and download capabilities for all *Chlamydiae*. There are no comparative genomics capabilities.

The Microbial Genome Database for Comparative Analysis (MBGD) at `http://mbgd.genome.ad.jp/` [251, 252, 253] is a database for comparative analysis of completely sequenced microbial genomes. It aims to facilitate comparative genomics from various points of view such as ortholog identification, paralog clustering, motif analysis and gene order comparison. It contains information about the pathogenic chlamydiae *Chlamydia muridarum* MoPn, *Chlamydia trachomatis* 434/Bu, *Chlamydia trachomatis* A/HAR-13, *Chlamydia trachomatis* Jali20, *Chlamydia trachomatis* D/UW-3/CX, *Chlamydia trachomatis* UCH-1, *Chlamydophila abortus* S26/3, *Chlamydophila caviae* GPIC, *Chlamydophila felis* Fe/C-56, *Chlamydophila pneumoniae* AR39, *Chlamydophila pneumoniae* CWL029, *Chlamydophila pneumoniae* J138, *Chlamydophila pneumoniae* TW-183 and the environmental chlamydium *Protochlamydia amoebophila* UWE25.

1.7.8 Technical basics for the implementation of genome databases

1.7.8.1 The three-tier architecture

The three-tier [254] (Figure 1.19) is a client-server architecture in which the user interface, functional process logic ("business rules"), computer data storage and data access are developed and maintained as independent modules, most often on separate platforms. The replacement, upgrade or update of elements in one of these modules does not affect the other layers so that it is possible to react to upcoming requirements with little effort.

The user interface typically runs on a desktop PC, the functional process logic may consist of one or more separate modules running on an application server (section 1.7.8.2), and a database management system on a database server (section 1.7.8.5) contains the computer data storage logic.

The three-tier architecture has the following three tiers:

- **Presentation Tier** This layer is mainly responsible for the interaction with the user and the display of results, e.g. in a web browser. The results are retrieved by communication with the other tiers.
- **Application Tier (business logic, logic tier, data access tier, middle tier)** This tier controls an application's functionality by performing detailed processing. An example is the retrieval of information about the next neighbors of a gene, their assignments to orthologous groups and the integration of this information into one specific format that the presentation tier can use for the display of information.
- **Data Tier** This layer is responsible for storage and retrieval of data from databases or file systems. It keeps data neutral and independent from the other tiers.

1.7.8.2 Application server

An application server is a software framework that can be used for the construction of dynamic web content and that allows to execute programs on them. Application servers

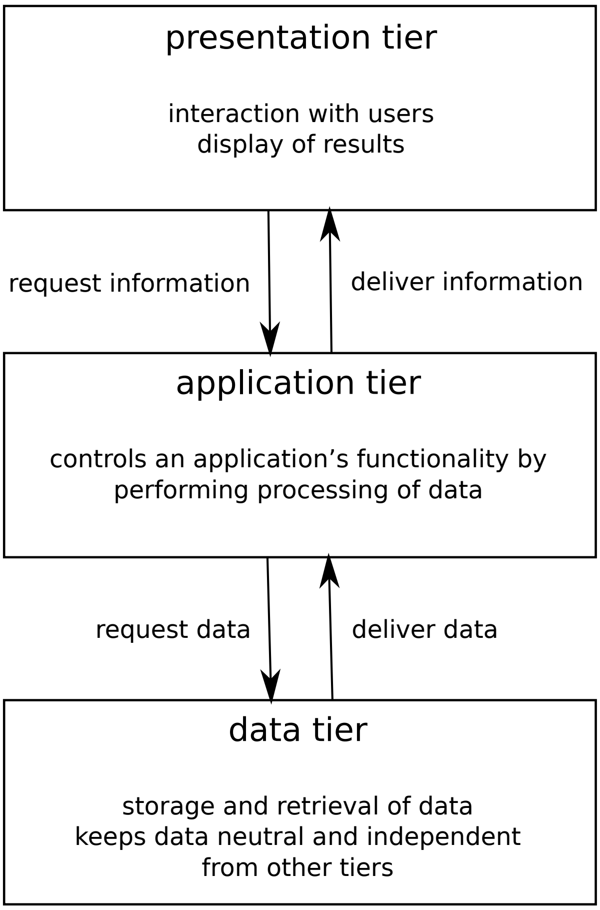


Figure 1.19: The three-tier architecture The three-tier [254] is a client-server architecture in which the user interface (presentation tier), functional process logic (application tier), computer data storage and data access (data tier) are developed and maintained as independent modules. The replacement, upgrade or update of elements in one of these modules does not affect the other layers so that it is possible to react to upcoming requirements with little effort.

offer advantages like encapsulation of data sources (e.g. database connection pooling), data and code integrity, performance and transaction support. In many cases they also implement services like clustering, fail-over and load-balancing. Application servers are often used for the execution of the methods of the business logic (see section 1.7.8.1).

For the projects of this work the application servers Apache Tomcat (`http://tomcat.apache.org/`) and JBOSS (`http://jboss.org/`) have been used.

1.7.8.3 Java & Jave EE

Java is among other things a programming language developed by Oracle (`http://java.oracle.com/`). It is a modern object oriented language that is simple, independent on the operating system, high performance, robust and secure.

The Java Platform, Enterprise Edition (Java EE) is a widely used platform for server programming in the Java programming language. It provides the functionality of fault-tolerant, distributed, multi-tier Java software, based largely on modular components running on an application server (section 1.7.8.2).

Some of the application programming interfaces (APIs) of Java EE have been used for the projects within this work: Enterprise Java Beans (EJBs), JavaServer Pages (JSPs), and Web Services (WS).

1.7.8.4 XML

The extensible Markup Language (XML) is a simple, very flexible text format that makes it easy to exchange and process data. XML is a set of rules for encoding documents in machine-readable form and is widely used for the representation of arbitrary data structures, for example in Web Services (section 1.7.8.6.2). There exist libraries for the handling of XML in probably any programming language which makes processing XML easy.

The example in Figure 1.20 shows a shortened MEDLINE/PubMed XML file for the well-known publication "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid" of Watson & Crick [255].

1.7.8.5 Data storage in databases

Data can either be stored in files on a filesystem or in some kind of database system. Files are on first sight the easiest way to store data but have drawbacks at second sight in comparison to databases.

A database in the context normally used is actually a database management system (DBMS). It consists of two parts, a software that provides storage, access, security, backup and other facilities, and the database itself. Some of the advantages of DBMS are that they allow parallel access to the data for several users, that modifications on the data can be seen by all users, that DBMSs provide multiple user interfaces, storage structures for efficient query processing, mechanisms for controlling redundancies, backup and recovery possibilities, powerful query mechanisms and that they allow for a fine granular rights management.

There can several types of DBMS be distinguished. The most prominent type is the relational model. MySQL (`http://www.mysql.org`) is such a relational database management system (RDBMS) that runs as a server providing multi-user access to a number of databases. A relational database consists of tables containing information connected by relations (Figure 1.21). In order to retrieve data spread over different tables of the database it is necessary to join tables, that is to merge information belonging together from several tables using one or more columns common to several tables.

If data is stored in just one big table then there might occur redundancies. If for example proteins with their protein names and the respective name of the organism should be stored in such a table, then the name of the organism would have to be stored for every single protein redundantly. If the name of an organism changes then the organism name needs to be changed for all proteins of this organism in the table. If

```
<PubmedArticle>
   <MedlineCitation Owner="NLM" Status="MEDLINE">
      <PMID>13054692</PMID>
      <DateCreated>
         <Year>1953</Year>
         <Month>12</Month>
         <Day>01</Day>
      </DateCreated>
      <Article PubModel="Print">
         <Journal>
            <ISSN IssnType="Print">0028-0836</ISSN>
            <JournalIssue CitedMedium="Print">
               <Volume>171</Volume>
               <Issue>4356</Issue>
               <PubDate>
                  <Year>1953</Year>
                  <Month>Apr</Month>
                  <Day>25</Day>
               </PubDate>
            </JournalIssue>
            <Title>Nature</Title>
            <ISOAbbreviation>Nature</ISOAbbreviation>
         </Journal>
         <ArticleTitle>Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.</ArticleTitle>
         <Pagination>
            <MedlinePgn>737-8</MedlinePgn>
         </Pagination>
         <AuthorList CompleteYN="Y">
            <Author ValidYN="Y">
               <LastName>WATSON</LastName>
               <ForeName>J D</ForeName>
               <Initials>JD</Initials>
            </Author>
            <Author ValidYN="Y">
               <LastName>CRICK</LastName>
               <ForeName>F H</ForeName>
               <Initials>FH</Initials>
            </Author>
         </AuthorList>
         <Language>eng</Language>
         <PublicationTypeList>
            <PublicationType>Journal Article</PublicationType>
         </PublicationTypeList>
      </Article>
   </MedlineCitation>
</PubmedArticle>
```

Figure 1.20: Example of a shortened XML file of the Watson & Crick paper Various information available for the publication "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid" of Watson & Crick [255] is contained within the XML file.

not all affected entries are changed then this results in inconsistencies. Therefore tables in relational databases are usually normalized, that is big tables are split into several smaller tables so that every information is only stored once within the database. In the example this can be done by moving the organism name into table 2 and by inserting the information about where the organism name stands in table 2 into table 1. This way the organism name can be looked up in table 2.

If big tables are normalized this might result in many tables. In order to retrieve the full information spread over these tables the tables need to be joined. This process

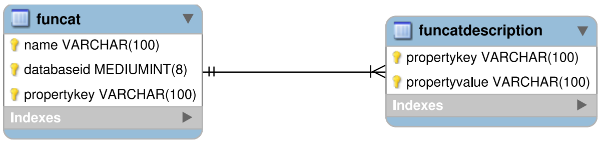


Figure 1.21: Example of two tables and their relation within the ChlamydiaeDB The figure shows two tables, the table "funcat" and the table "funcatdescription". "funcat" contains information which proteins are annotated with which Functional Catalogue (FunCat) categories. name is the proteinname (e.g. GI:10957571), databaseid is an internal databaseid referencing another table not shown here (e.g. 711 for *Chlamydia muridarum* Nigg) and propertykey is a FunCat category (e.g. 11.02.03). The description for the FunCat categories is stored in the table "funcatdescription". propertykey is a FunCat category (e.g. 11.02.03) and propertyvalue is the associated FunCat description (e.g. mRNA synthesis). The propertykey of "funcat" and the propertykey of "funcatdescription" reference each other and define a relation: using the propertykey of "funcat" the description can be retrieved using the same propertykey in table "funcatdescription". This separate storage of FunCat categories and FunCat descriptions makes it possible that the FunCat description has not to be stored for each of the annotated proteins in the table "funcat". In order to retrieve the complete information for a protein a join between the two tables has to be done using the common column ”propertykey”.

can become quite slow when many big tables are involved. If retrieval speed plays a role, as it is the case in most applications in the praxis, then data warehouses are used. Data warehouses store data redundantly in a general scheme in order to facilitate efficient reporting and analysis of the whole stored data or big parts of it. The redundant storage of information results in increased query speed as joins are avoided at the cost of disk space. Additionally aggregated information like for example the numbers of proteins of an organism can be precomputed and retrieved fast. The programmer using a data warehouse has to take care of possible inconsistencies by herself or himself.

A very specialized approach has recently been introduced. It is a system for very large tables and is called Hbase (`http://hbase.apache.org/`) and is similar to Google’s Bigtable [256]. HBase is the Hadoop database and allows for random, realtime read/write access to Big Data.

1.7.8.6 Technologies for the retrieval of information from genome databases

1.7.8.6.1 EJBs If data is available on local servers the best way to retrieve information are Enterprise JavaBeans (EJBs) (`http://www.oracle.com/technetwork/java/index-jsp-140203.html`). They are the server-side component architecture for Java EE (section 1.7.8.3). The EJB technology enables rapid and simplified development of distributed, transactional, secure and portable applications based on Java technology. EJBs belong to the logic tier of the three-tier architecture. An EJB is a software component which runs within an EJB container on a specific application server (e.g. JBOSS (`http://www.jboss.org/`) or GlassFish (`http://www.oracle.com/technetwork/java/javaee/community/index.html`). Simply spoken an EJB is an application running on an application server, that returns a result for a specific request and returns the result e.g. to the presentation tier. These results can be Java objects or XML objects.

1.7.8.6.2 Web Services If data is available on remote servers then Web Services can be used for the retrieval of information. They are typically application programming interfaces (API) or web APIs accessible via Hypertext Transfer Protocol (HTTP) and are executed on remote systems hosting the requested services. Therefore Web Services can interact with clients anywhere in the world. Web Services use XML messages for the communication with the clients, that is requests are sent to Web Services using a specific XML message and the results are sent back as another specific XML message.

1.7.8.7 Presentation of information to the user

1.7.8.7.1 Content management systems and web portals

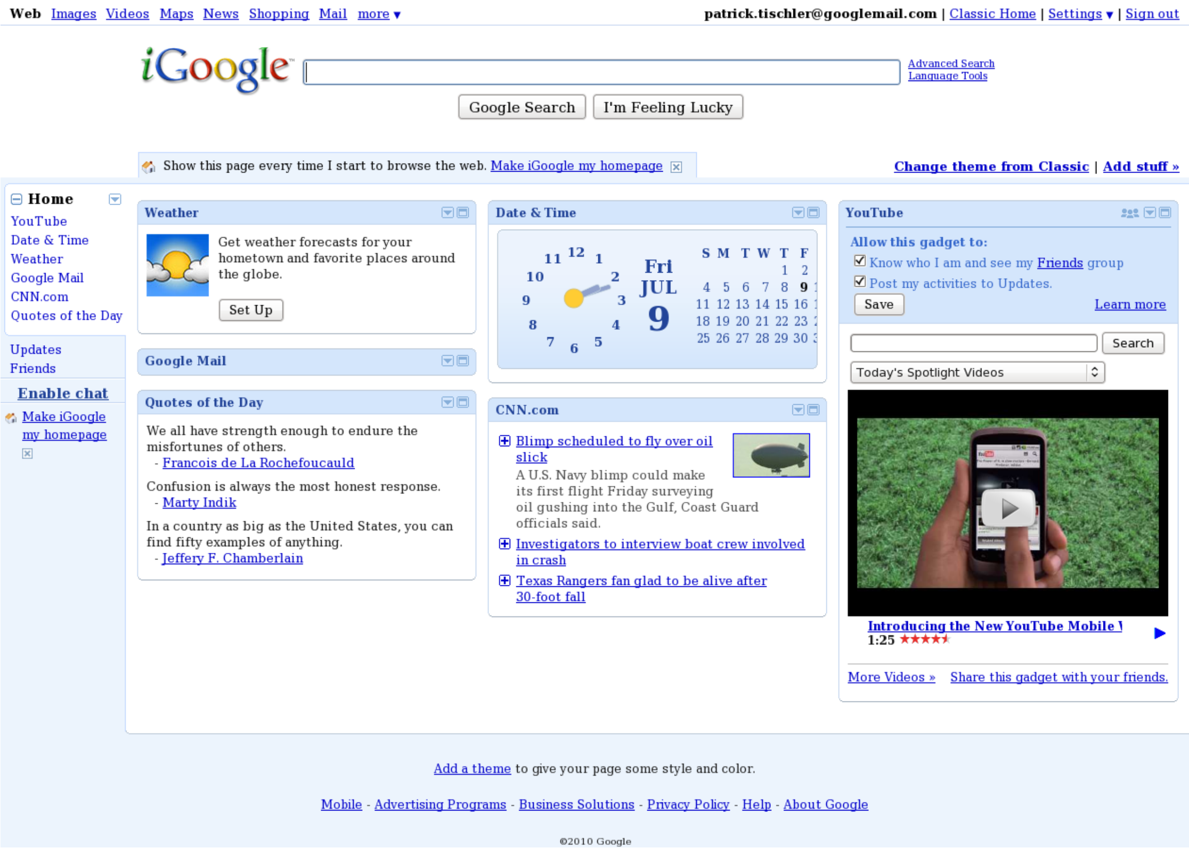
1.7.8.7.1.1 OpenCms OpenCms (`http://www.opencms.org`) is an open source Content Management System (CMS). Content management systems allow the creation and the management of content such as web sites, text and pictures. The systems provide tools for authoring of entries for users with little or no knowledge of a programming or markup language to create and manage content with relative ease. OpenCms and also some other CMS provide WYSIWYG (What You See Is What You Get) editors, that is the user sees at the time of editing how the page will look like finally.

OpenCms is based on Java and XML technology and can be deployed in an open source environment.

1.7.8.7.1.2 GenRE GenRE is the Munich Information Center for Protein Sequences (MIPS) Genome Research Environment (`http://mips.helmholtz-muenchen.de/genre/proj/genre`). It is implemented as a modular and multi-tiered architecture based on Java EE middleware and is designed to provide the possibility to reuse components together with existing software, hence simplifying the integration of various bioinformatics resources. GenRE is not a single piece of software but a framework allowing for midrange data management and processing. OpenCms is used for the presentation tier as it provides very convenient administration possibilities and the easy usage of EJBs and XML/XSL for the visualization of the results.

1.7.8.7.1.3 Portlets on web portals and the Liferay portal server Portlets are pluggable user interface components that are managed and displayed in a web portal. Portlets produce fragments of markup code that are aggregated into a portal page. There exist portlet standards that enable software developers to create portlets that can be plugged in to any portal server supporting the standards.

A web portal provides access to information from diverse sources in a unified way in the World Wide Web (WWW). Portals provide a consistent look and feel with access



Show this page every time I start to browse the web. Make iGoogle my homepage

Get weather forecasts for your hometown and favorite places around the globe.

☑ Know who I am and see my Friends group

☑ Post my activities to Updates.

We all have strength enough to endure the misfortunes of others.
- Francois de La Rochefoucauld

Confusion is always the most honest response.
- Marty Indik

In a country as big as the United States, you can find fifty examples of anything.
- Jeffery F. Chamberlain

⊞ Blimp scheduled to fly over oil slick
A U.S. Navy blimp could make its first flight Friday surveying oil gushing into the Gulf, Coast Guard officials said.

⊞ Investigators to interview boat crew involved in crash

⊞ Texas Rangers fan glad to be alive after 30-foot fall

Add a theme to give your page some style and color.

Figure 1.22: The iGoogle portal page There are different programs visible. These programs (portlets) are Weather, Date and Time, Youtube, CNN.com, Google Mail, Quotes of the Day. The iGoogle portal provides a consistent look and integrates the different programs into a single page.

control and procedures for multiple applications (portlets), which otherwise would have been different entities altogether. Examples for web portals are iGoogle (`http://www.google.com/ig`) (Figure 1.22) and Yahoo (`http://www.yahoo.com`).

The Liferay Portal server (`http://www.liferay.com/`) is an enterprise web platform for building customized web portals and is widely used for business solutions.

1.7.8.7.2 Display of information There are several ways how information can be displayed on a browser page. The easiest way is HTML. JavaScript makes pages dynamic and CSS is used to define the layout. JSPs are a technology creating dynamic pages using a mixture of HTML and Java while the transformation of XML data using XSL stylesheets provides another way how to display dynamic data. The following sections will provide a short introduction into these terms.

1.7.8.7.2.1 HTML The HyperText Markup Language (HTML) is the most common markup language for web pages and provides a means to create structured documents by denoting structural semantics for text such as headings, paragraphs, lists, links, quotes and other items. Images, objects and scripts in languages as JavaScript can be embeded.

HTML documents contain the description of the page but not the exact look as the display is done by the particular web browser. This is also the reason why pages may look different using different web browsers.

1.7.8.7.2.2 JavaScript JavaScript is a scripting language typically used to provide enhanced user interfaces and dynamic web sites. One example are fields on web sites that only appear if a checkbox is clicked, another example is information that pops up if the mouse is over a specific part of a web site.

1.7.8.7.2.3 CSS Cascading Style Sheets (CSS) are a declarative style sheet language used to define the look and the formatting of a document written in a markup language (e.g. HTML, XHTML). The HTML defines the content to be displayed on a page, CSS defines how the content is displayed.

1.7.8.7.2.4 XSL The Extensible Stylesheet Language (XSL) is a specific transformation language describing how XML files should be formatted or transformed. This makes it easy to produce different views on the same data (XML) using different stylesheets (XSL). The use of XSL for the display of information therefore allows to use the same XML from the application tier for completely different representations.

1.7.8.7.2.5 JavaServer Pages JavaServer Pages (JSPs) is a Java technology that can serve dynamically generated web pages based on HTML, XML, or other document types. Internally JSPs are text documents consisting of static text and dynamic elements, the JSP elements. JSPs are converted into servlets on the application server (e.g. Apache Tomcat).

2 Methods & Results

2.1 Computational genome analysis of prokaryotic genomes

2.1.1 Genome projects with collaborators

The sequencing of bacterial genomes became affordable also for small research groups without their own bioinformatics facilities within the last years (section 1.1). Small analyses can be done manually by the experimental groups but as soon as analyses need to be done on a larger scale most of the groups cannot cope with the amount of data anymore or do not have the experience with more sophisticated methods. Additionally ready to use software often does not exist for specific problems, the input format is different from the formats accepted or programs need to be implemented to be able to interpret the results. Additionally, a risk often not taken into account is the problem of submitting unpublished data to public webservers on which the submitter has no control of what might happen with the data. Therefore offline analyses need to be performed if it is unclear whether the webserver is trustworthy.

This is why many groups rely on bioinformaticians who are experienced in the handling of large amounts of data, use stand of the art datasources and programs or even implement programs if necessary.

During the last years the contact between our group and collaboration partners with different backgrounds and investigated organisms could be established. An overview over these projects can be seen in Table 2.1.

2.1.2 Annotation with the PEDANT system

Many of the standard analyses as the detection of best hits in public sequence databases and function annotations like the assignment of UniProtKB/Swiss-Prot keywords [182], Gene Ontology (GO) [92] and FunCat [81] are daily business and need to be repeated for every project. Therefore many applications have been proposed for the automation of these tasks (section 1.7.2).

The PEDANT [72] system was selected as best solution for the cooperation partners, as it automates many of the standard analyses, its web interface provides access to all publicly available RefSeq [52] genomes and it allows convenient access to private organisms in a password protected area. PEDANT offers the possibility to access all

cooperation partners	institution	organism	taxonomy	comment	publication
Angelika Lehner, Roger Stephan	Institute for Food Safety and Hygiene, Vetsuisse Faculty University of Zürich	Cronobacter turicensis LMG 23827	family: Enterobacteriaceae	(foodborne) pathogenic bacterium	[257], in preparation
		Enterobacter helveticus	family: Enterobacteriaceae	various BACs of Enterobacter helveticus	in preparation
		Enterobacter sp. 638	family: Enterobacteriaceae	internal database available before submission of public version	-
		Escherichia coli L1000	family: Enterobacteriaceae	natural strain inhibiting both antibiotic resistant and sensitive Salmonella isolates in vitro	[258]
Astrid Horn	Department of Microbial Ecology, University of Vienna	Candidatus Protochlamydia amoebophila UWE25	family: Parachlamydiaceae	environmental chlamydia, first sequenced member of the family Parachlamydiaceae	-
Stephan Schmitz-Esser	Department of Microbial Ecology, University of Vienna	Candidatus Amoebophilus asiaticus 5a2	no rank: unclassified Bacteroidetes	obligate intracellular amoeba symbiont	[259]
Sebastian Lücker, Holger Daims	Department of Microbial Ecology, University of Vienna	Leptospirillum sp. Group II	genus: Nitrospirae	Iron-oxidizing bacterium from acidic-mine drainage	-
		Leptospirillum sp. Group III	genus: Nitrospirae	Iron-oxidizing bacterium from acidic-mine drainage	-
Astrid Horn, Alexander Siegl, Elena Tönshoff, Matthias Horn	Department of Microbial Ecology, University of Vienna	Candidatus Clavochlamydia salmonicola	family: Clavochlamydiaceae	associated with gills of salmonid fish	in preparation
Sebastian Lücker, Holger Daims	Department of Microbial Ecology, University of Vienna	Candidatus Nitrospira defluvii	genus: Nitrospirae	Nitrite-oxidizing bacterium enriched from activated sludge	[260], submitted
Thomas Penz, Elena Tönshoff, Matthias Horn	Department of Microbial Ecology, University of Vienna	Symbiont TTL1 of the Adelges nordmannianae/piceae complex	family: Enterobacteriaceae	-	in preparation
Astrid Horn, Matthias Horn	Department of Microbial Ecology, University of Vienna	Parachlamydia acanthamoebae UV7	family: Parachlamydiaceae	environmental chlamydia	submitted
		Simkania negevensis Z	family: Simkaniaceae	environmental chlamydia	submitted
		Waddlia chondrophila 2032/99	family: Waddliaceae	environmental chlamydia	submitted
Anja Spang, Christa Schleper	Department of Genetics in Ecology, University of Vienna	Acidianus filamentous virus AFV10	family: Lipothrixviridae	-	in preparation
		Nitrosofabula viennensis	genus: Nitrososphaera	-	-
Roland Hatzenpichler, Anja Spang, Christa Schleper & Michael Wagner	Department of Genetics in Ecology, University of Vienna & Department of Microbial Ecology, University of Vienna	Candidatus Nitrososphaera gargensis	genus: Nitrososphaera	-	[261]
Nidal Abu Laban, Draženka Selesi, Rainer Meckenstock	Institute of Groundwater Ecology, Helmholtz Zentrum München	Benzol Ferrihydrite culture	meta genome	-	[262]
Franz Bergmann, Draženka Selesi, Rainer Meckenstock	Institute of Groundwater Ecology, Helmholtz Zentrum München	deltaproteobacterial enrichment culture N47	meta genome	-	[263], [264], in preparation
Thomas Weinmaier, Charles David	Cell and Developmental Biology, Ludwig-Maximillians-Universität München	uncultured Curvibacter sp.	family: Comamonadaceae	bacterium found together with Hydra	[265]

Table 2.1: Genome projects with different cooperation partners The list of partners within each project is ordered alphabetically, the last positions are reserved for the workgroup leaders or professors.

information available for a genetic element on the one hand and to access all elements having a specific feature on the other hand. This makes PEDANT precious for the annotation of genomes. Integrated tools like BLAST [73] searches of the sequences of the contigs or coding sequences against different databases allow the users to answer many questions by themselves. Additionally a workflow system allows to use standard workflows for the handling of new genomes and to specifically adapt the workflows for the single genome. These points make the PEDANT system ideal for the collaboration with experimental groups.

Additionally the easy access to data from PEDANT on the database level or via Web Service allows the implementation of more sophisticated analyses.

As stated before (section 1.7.5) genomic data needs to be submitted to the DDBJ, EMBL-Bank or the GenBank database prior to the publication of a paper. PEDANT offers the possibility to export all information for an organism into the EMBL or GenBank formats selectively. Not all information is suited to be published in the public databases, for example easy reproducible or changing information like InterPro protein domains or best BLAST hits. Therefore Thomas Weinmaier and myself implemented a program that allows to remove the unnecessary parts from these files and to add additional information like the protein the product was inferred from.

The program has been successfully applied within various genome projects for the submission of sequence data.

2.1.3 Gene prediction in prokaryotic genomes

2.1.3.1 Observations in genome projects

It is common to use different intrinsic gene finders like Glimmer 3.0 [34] or GeneMarkS [31] for the gene prediction in prokaryotes. This is done to increase the sensitivity of the gene finding process. Usually it is decided manually about the best gene models afterwards using information about homology to known gene products.

When I did the gene finding for the genomes of the collaboration partners, I observed that there can occur some difficulties. These issues will be described in the following.

2.1.3.1.1 Different gene starts from different intrinsic gene finders A prokaryotic gene is always limited by a start codon in the beginning and a stop codon in the end. While it is not possible that a stop codon exists in the same reading frame within the coding region, several possible gene starts may exist. Therefore the determination of the correct gene start is one of the major problems in prokaryotic gene prediction (see also Frishman et al [38]). The decision about the “correct” gene starts and by that the best gene models can often be made manually by the inspection of hits to known gene products for each of the competing gene models.

2.1.3.1.2 Overlapping gene models from different intrinsic gene finders Overlaps between gene models can often clearly be solved if only one of the two models has significant similarities to known protein sequences as the model without similarity is likely a false positive gene prediction. If both gene models show similarities to known gene products the decision is more difficult. In this case the model with the higher sum of bitscores of its hits can be selected as done in CONSORF (section 1.2.2.5.1).

A first version of my own gene prediction pipeline implemented this decision. When Thomas Weinmaier and I mapped the predicted genes for the *Curvibacter* putative symbiont of *Hydra magnipapillata* onto KEGG [189] maps in order to learn about the metabolic capabilities of this organism, we could identify several cases where the KEGG maps for the symbiont differed from the KEGG maps of its closely related species. In some of these cases the symbiont was lacking enzymes of pathways. When we did a TBLASTN of the affected protein sequence of the related species against the genomic DNA of the symbiont we could identify the proteins encoded on the DNA. The reason for the missing genes was that they had been removed erroneously due to only slight differences in the sum of bitscores for the overlapping gene models. Unfortunately the wrong model had a slightly higher sum of bitscores produced by hits against hypothetical proteins and therefore was kept while the correct model was deleted.

These cases showed that the comparison of the sum of bitscores of overlapping genes alone seems not to be enough for the decision about the better gene model.

2.1.3.1.3 Problematic similarities to incomplete sequences in public databases I conducted gene predictions for *Nitrososphaera* genomes using similarities against all sequences from UniRef100 [266, 267] as extrinsic evidences. When Anja Spang from the Department of Genetics in Ecology from the University of Vienna analyzed genes of interest in more detail she discovered that some of them were annotated too short.

An example is the Ammonia monooxygenase amoA from an uncultured ammonia-oxidizing beta proteobacterium (UniProtKB/Swiss-Prot accession AMOA_NITEU). An alignment against the most homologous sequences (Figure 2.1) shows that the first six hits support the full length of amoA, then there are a few shorter hits and then again hits supporting the full length. The best hit not supporting the gene start is ABN12960, the ammonia monooxygenase subunit A from the same protein of the uncultured ammonia-oxidizing beta proteobacterium. The alignment starts at position twelve of the query protein, and the two sequences have 100% sequence identity. The corresponding Genbank entry leads to a publication [268] that gives the reason for this short sequence: it was identified by primers located within the coding region of the protein (Figure 2.2). Therefore only the sequence between the primers has been amplified, identified and submitted.

The usage of hits against incomplete sequences can cause problems. This is why only hits to complete sequences should be used for the determination of gene starts.

2.1.3.1.4 Error propagation by hits to too closely related species When Elena Tönshoff from the Department of Microbial Ecology from the University of Vienna inspected the most similar hits of genes of the symbiont TTL1 of the *Adelges nordmannianae/piceae* complex she could see that the same genes in closely related species were annotated with different lengths. The upstream regions of the genes were still conserved

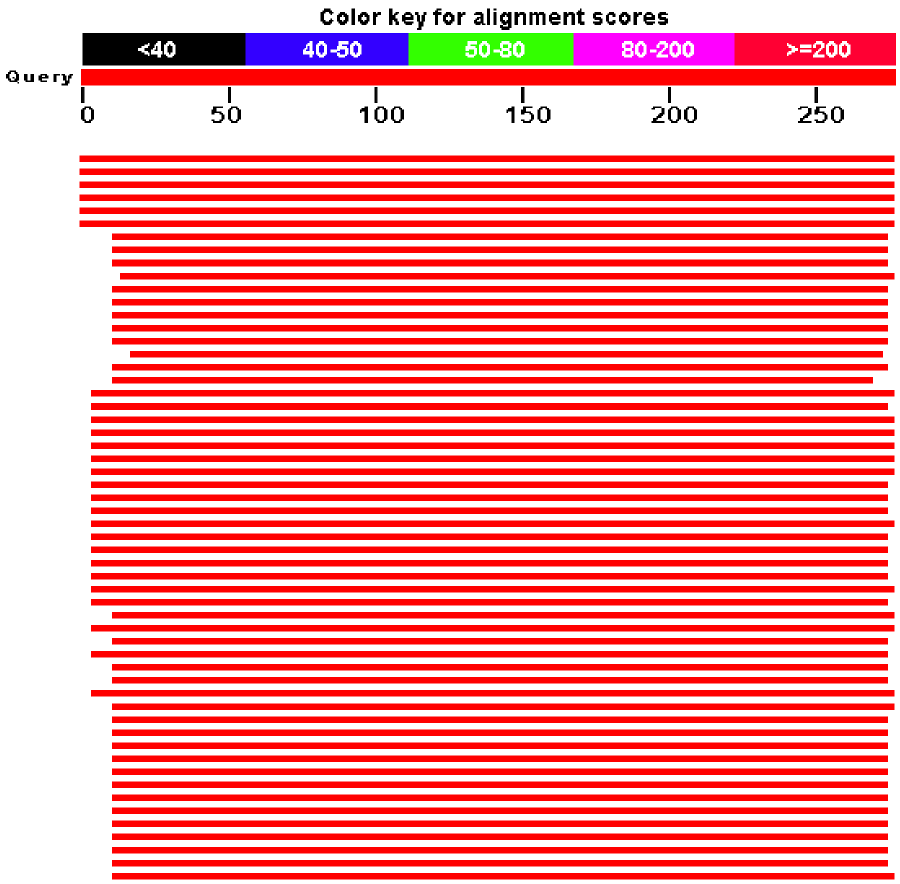


Figure 2.1: Pairwise sequence alignments of the UniProtKB/Swiss-Prot protein AMOA_NITEU and its homologs It can be seen that the first six hits support the full length of the protein, then there are a few hits not supporting the full length and then there are hits supporting the full length of the protein again. The hits not supporting the full length are the same proteins missing an N-terminal part due to primers located within the coding sequence used for their identification (Figure 2.2). When these alignments are used for the determination of gene starts this can cause problems.

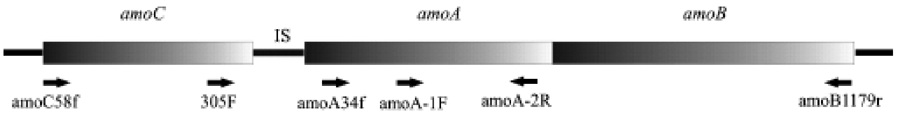


Figure 2.2: Structure of the amoCAB operon The structure of the amoCAB operon (on the top) and the positions of the primers amoC58f, 305F, amoA34f, amoA-1F, amoA-2R, amoB1179r used in this study are shown. It can be seen that all primers lie within amoA, amoB and amoC respectively. (Figure adapted from [268])

and it was not clear whether these similar regions existed due to the close relationship to the symbiont or because the genes were recently shortened in some relatives.

Therefore close relatives should be excluded from the BLAST search in order to prevent the propagation of possibly falsely annotated gene starts in closely related species.

When hits to more distantly related species can be identified it is more likely that conserved regions are conserved due to conservation of coding regions.

2.1.3.1.5 Overlaps of intrinsic gene predictions and non-coding genetic elements rRNAs and tRNAs are very well conserved genetic elements. During the annotation of the genome of *Cronobacter turicensis* LMG 23827 I discovered that some of the gene models overlapped with rRNAs and tRNAs. Therefore these obviously wrong gene models had to be removed.

The problem with that is that it cannot be excluded that one of these removed genes had led to the deletion of other correct genes overlapping with them.

Therefore the knowledge about the position of non-coding regions within genomes should be integrated into the gene prediction process.

2.1.3.2 How to decide about the best intrinsic gene models manually

In the following the typical manual procedure is described how we and our cooperation partners normally annotate genes in prokaryotic genomes, bearing in mind the previously described observations. It is not intended to be a hands-on-tutorial but describes the reasons for the decisions and the course of action.

Bearing in mind the previously described observations in the following the typical manual procedure is described how we and our cooperation partners normally annotate genes in prokaryotic genomes. This is not intended to be a hands-on-tutorial but describes the reasons for the decisions and the course of action.

2.1.3.2.1 Execution of intrinsic gene finders First the intrinsic gene finders are executed. In order to cover genome specific features a training precedes the prediction. Specific properties of sequences upstream of genes like ribosomal binding sites should be switched on, if possible, as this often improves the gene predictions.

2.1.3.2.2 Execution of BLAST against known protein sequences Hits to known protein sequences are searched using BLAST [73] as they serve as valuable information for the decision about the best gene models and can help to improve the sensitivity of the gene prediction as no genes with sufficiently good similarity to known protein sequences are missed.

2.1.3.2.3 Determination of non-coding sequences Noncoding sequences like rRNAs and tRNAs are searched within the genomic sequences as these elements can help to dissolve overlaps.

2.1.3.2.4 Grouping of gene models by same strand and stop coordinate All intrinsic and extrinsic information is available at this time point. This is the first step of the actual decision process about the best gene models:

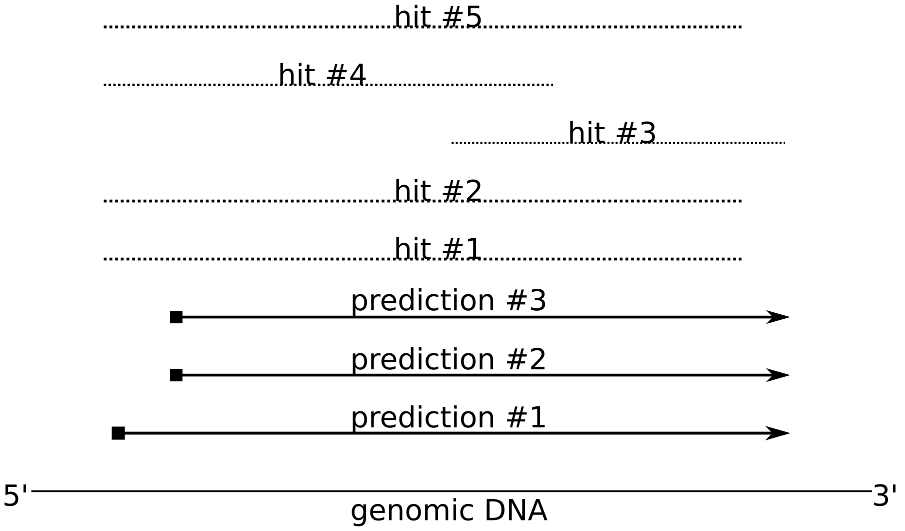


Figure 2.3: Intrinsic and extrinsic information grouped by strand and stop coordinate The line at the bottom is the genomic DNA of a prokaryote, above there are three gene predictions differing in the gene start (squares are gene starts, arrow heads are gene stops). Dashed lines are extrinsic information in form of BLAST hits to known protein sequences. It can be seen by the BLAST hits that prediction #1 offers the most probable gene start.

In order to be able to identify gene models that represent the same genes, all gene models from the various predictions are grouped by same strand and stop coordinate. This is done as the gene stop is a reliable feature in contrast to the gene start and each of these groups represents a potential gene with one gene stop and at least one potential gene start (Figure 2.3).

2.1.3.2.5 Selection of a representative gene model for each group After the gene models have been grouped by same strand and stop coordinate for each group representing one gene, a representative is chosen. That means basically the selection of the most probable gene start for this gene. This is done by looking at the available evidences for the gene models.

If no evidences are available for the gene then the gene model supported by the highest number of intrinsic predictions is chosen. If there is no gene start supported by the majority of the predictions then the prediction of the intrinsic prediction program the annotator trusts most is selected. This is not always necessarily the longest gene model.

When there are evidences it is first checked whether there exist evidences that support one specific gene start. If this is the case, then this start is probably the correct one and has to be chosen (see Figure 2.3). If this is not the case then the region supported by homology to known sequences is searched. The gene start producing the shortest gene and including all the evidences is the best start then. There should a certain overlap between gene start and evidences be allowed due to possible gaps in the alignments. An allowed overlap of 10 amino acids length has shown to be reasonable.

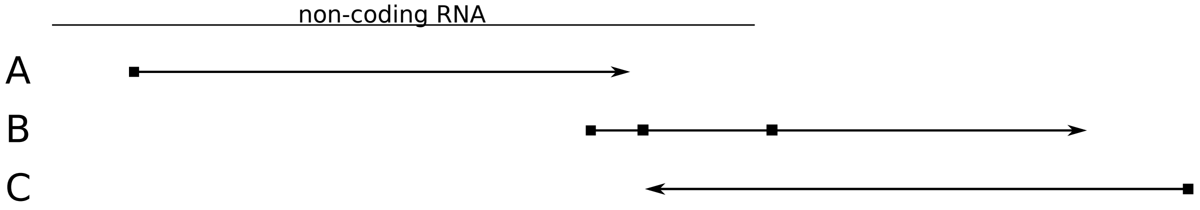


Figure 2.4: Kinds of overlaps between non-coding RNA and gene predictions The line at the top is a non-coding RNA lying on the contig of a prokaryotic genome. The gene predictions are shown as arrows, squares are gene starts, arrow heads are gene stops. **A:** The gene model lies within a non-coding sequence and has to be removed **B:** The N-terminus of the gene model overlaps with a non-coding sequence. It has to be checked whether the overlap can be resolved by shorting the gene model or whether the gene model has to be removed **C:** The C-terminus of the gene model overlaps with a non-coding sequence. Therefore the gene model has to be removed.

As described in section 2.1.3.1.4, too closely related sequences can cause problems in the gene prediction process. For this reason only hits to not too closely related species are used. The exclusion of hits to sequences of the same taxonomic family has shown to give quite reasonable results.

After this step there is only one gene model left for each gene.

2.1.3.2.6 Resolving overlaps with non-coding sequences Overlaps with non-coding sequences are the clearest cases that can be resolved within the gene prediction as rRNAs and the structures of tRNAs are very well conserved. There can three cases be distinguished:

1. **The gene model lies within a non-coding sequence** The gene model has to be removed as it is most probably a false positive gene prediction. (Figure 2.4 A)

2. **The N-terminus of the gene model overlaps with a non-coding sequence** It has to be checked whether the overlap can be resolved by shorting the gene model or whether the gene model has to be removed. First it is searched for alternative start codons within the sequence of the gene model. If another start can be identified that results in a gene model not overlapping more than 6 nt with the RNA then the new start is selected. The new gene start resulting in the longest gene and allowing a small overlap of 6 nt for alignment artifact reasons is selected. If no other possible start can be identified within the gene model then the gene model has to be removed. (Figure 2.4 B)

3. **The C-terminus of the gene model overlaps with a non-coding sequence** The gene model has to be removed as it is most probably a false positive gene prediction. (Figure 2.4 C)

2.1.3.2.7 Resolving overlaps between genes A certain overlap between the genes can be observed in many bacterial genomes. This is also taken into account within intrinsic gene finders like GeneMark.hmm. If the length of the overlap between two gene models is greater than 16% of the lengths of the proteins, then the overlap has to be resolved. First all overlaps between genes overlapping with their C-termini are dissolved as they are clear problem cases. The removal of these cases can solve other overlaps, which can then be handled in the second round.

1. **Genes overlapping with their C-termini** The basic idea is to remove the gene model having much less evidence than the other gene model. This can be measured for each of the gene models by the number of hits, by the quality of the alignments, and by the existing or non-existing annotations of the hits. The gene model with the better evidence is retained, the other gene model is removed. Figure 2.5 A shows two genes without evidence overlapping with their C-termini. The gene model of the intrinsic method the annotator trusts most is retained, the other one is deleted. Figure 2.5 B shows two genes with evidence overlapping with their C-termini. The gene model having the better evidence (the left gene) is retained, the other one is deleted.

2. **Other overlaps** If none of the two genes has evidence then one of the gene models has to be selected. This is the gene model of the prediction the annotator trusts most, the other model is removed (Figure 2.5 C).

   If only one of the two genes has evidence it is tried to resolve the overlap by shorting the gene without evidence (Figure 2.5 D). First all possible gene starts are searched in the nucleotide sequence of the gene. If another gene start of an intrinsic prediction method provides a gene model that does not produce an overlap then this model is selected. Otherwise the start producing the longest gene model without producing an overlap is selected. If no suitable start can be identified then the gene without evidence is removed.

   If both genes have evidence then the evidences need to be examined more closely. As it has been made sure that the gene model has been selected for each gene that is supported best by evidences previously (section 2.1.3.2.5), none of the two genes can be shortened further. Therefore the same idea is applied as in the case for C-terminal overlaps: The gene model having much less evidence than the other gene model is removed. This can be measured for each of the gene models by the number of hits, by the quality of the alignments, and by the existing or non-existing annotations of the hits. The gene model with the better evidence is retained, the other gene model is deleted (Figure 2.5 E).

After these steps a non-overlapping set of genes is available.

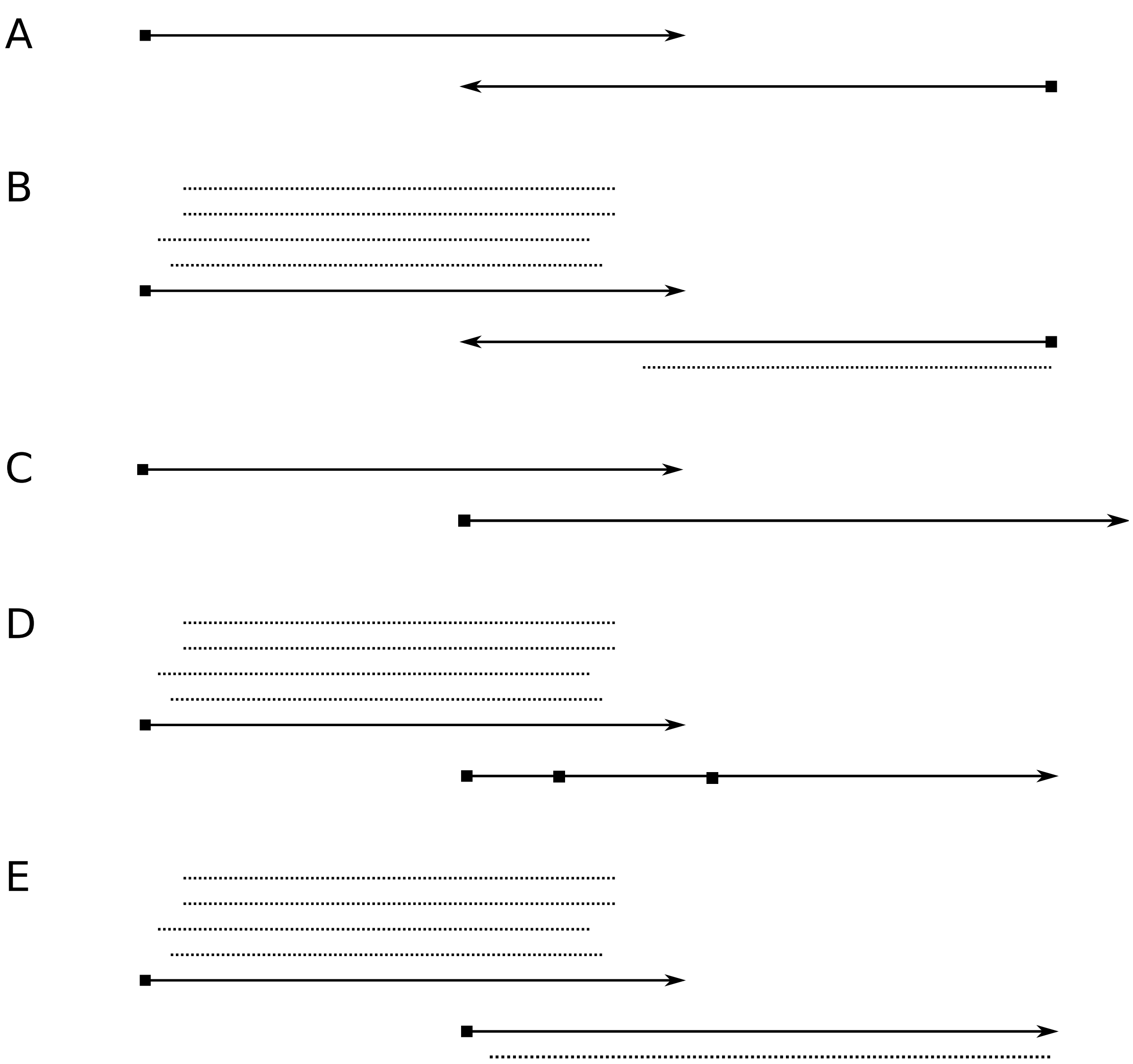


Figure 2.5: Resolvement of overlaps between gene models The gene predictions are shown as arrows, squares are gene starts, arrow heads are gene stops. **A:** Two genes without evidence overlap with their C-termini. The gene model of the intrinsic method the annotator trusts most is retained, the other one is deleted. **B:** Two genes with evidence overlap with their C-termini. The gene model having the better evidence or evidence at all (the left gene) is retained, the other one is deleted. **C:** Two genes overlap and none of the two genes has evidence. The gene model of the method the annotator trusts most is selected, the other model is removed. **D:** Only one of the two genes has evidence. It is tried to resolve the overlap by shorting the gene without evidence. If this is possible the gene without evidence is shortened, if it is not possible then the gene without evidence is removed. **E:** Both genes have evidence and cannot be shortened anymore as they are already shortened so that they represent the evidence best in previous steps. The gene model having much less evidence than the other gene model is removed, the other one is retained (the left one).

2.1.3.3 Existing tools for gene prediction in prokaryotic genomes integrating intrinsic and extrinsic information

The decisions about the "correct" gene starts and the best gene models can be made manually as described above. This has been done for whole genomes (e.g. [269]). With the increasing sequencing speed, the decreased costs for sequencing and the amounts of sequence data produced, the manual annotation of whole genomes is not feasible anymore.

Therefore an automated approach is desirable that supports the human annotator with the gene prediction for bacterial genomes.

There have been some solutions proposed integrating intrinsic and extrinsic information for the improvement of prokaryotic gene prediction.

ORPHEUS [38], CRITICA [39] and YACOP [42] are not suited for the previously described gene prediction as they are far from representing the procedure how a human annotator would decide.

CONSORF [43] integrates intrinsic predictions from established gene finders and extrinsic information is integrated as FASTX [47] similarity of potential gene candidates against a database of known gene products. Even though CONSORF is closer to the decisions described before, there are several points of criticism.

In the case that hits overlap more than 10% of the alignment length the hit with the lower bitscore is removed. But the longer a sequence is the higher is the probability that it gets a high bitscore by chance. This gives a higher weight to long ORFs and potentially very well conserved shorter ORFs might not be detected anymore as all their short hits are removed due to overlaps with a single long hit.

Each ORF gets several potential gene starts assigned. The most probable or shortest start is the first possible start upstream of the hit with the highest bitscore. The problem with this is that this might be a long hit with a high bitscore supporting the C-terminal part of the protein but not the N-terminus anymore. The gene could therefore be predicted too short.

When the gene start is determined, it is checked which potential start is supported by the most evidences. There are several problems with this approach: First no intrinsic information about coding potential or RBS is used and second several weak hits in distantly related species can cause the decision for a wrong gene start.

The question was whether CONSORF is suited to be adapted so that it makes the same decisions as a human annotator. CONSORF is written in Groovy, a dynamic language for the Java Platform. As CONSORF automates all steps necessary for the gene prediction for many genomes at once and therefore has lots of dependencies and as the implementation of the decisions for the overlap resolvement and gene start decision would have required a complete restructuring of the whole software it was necessary to implement an own new gene prediction pipeline from scratch.

2.1.3.4 ConsPred - a new gene prediction pipeline

2.1.3.4.1 Selection of intrinsic gene finders Two of the most established intrinsic gene finders are GeneMarkS [31] and GLIMMER 3.0 [34]. The fact that their predictions are well accepted and established is also reflected by the fact that their predictions are available for all RefSeq [52] genomes. Both programs are under constant development, many years of experience and manpower have been put into the gene finders.

For the new gene prediction pipeline GeneMarkS 2.6r and GLIMMER 3.0 were used.

2.1.3.4.2 Preparation of a BLAST database The gene prediction is designed for gene finding in bacteria, archaea and viruses. Therefore it would make no sense to include other sequences than sequences from these taxonomic groups in a BLAST database as other sequences might produce meaningless hits. Additionally by using only all protein sequences of all publicly available complete prokaryotic, archaeal and virus organisms in NCBI's Reference Sequence (RefSeq) [52] database, the database size can be reduced. Exact duplicates of protein sequences are replaced by one representative sequence possessing the taxonomic information of all the sequences it represents. For the reason that incomplete sequences can cause problems (see section 2.1.3.1.3) only complete genomes from RefSeq are used for the gene prediction.

Each sequence in the database contains the NCBI taxonomy ids of the organisms it is contained in in the description line of the fasta file. This makes it possible to use the same BLAST database for various gene prediction runs and to use flexible filtering of the BLAST hits during the runtime of the gene prediction pipeline in order to filter out too closely related hits.

2.1.3.4.3 Determination of non-coding RNAs as input for the gene prediction The RNAs are determined as described in section 1.2.3.2. The coordinates of RNAs can be supplied as a file to the gene prediction pipeline.

2.1.3.4.4 The gene prediction pipeline ConsPred In the following the steps of the new gene prediction pipeline ConsPred (Consensus Prediction) are outlined. Figure 2.6 shows a complete overview over the pipeline.

ConsPred has been implemented in Java.

1. **Extraction of all open reading frames (ORFs)** All open reading frames (ORFs) between two stop codons with at least 150 nt length are extracted from each of the contigs, using coding table 11, applying getorf from the EMBOSS package [270]. The sequences do not necessarily start with one of the possible start codons at this point yet. This extraction of all possible ORFs is done in order to cover possibly conserved genes that cannot be identified by the intrinsic gene prediction methods.

2. **Retrieval of similarities to known protein sequences** BLASTP [73] is executed for all extracted and translated ORFs from getorf against the previously

prepared BLAST database. An E-value cutoff of 10^{-10} is applied in order to avoid spurious hits. The pre-processed BLAST database allows to remove those hits from the result lists that are too closely related to the query genome.

3. **Removal of ORFs without evidence** All ORFs not having any similarity to known protein sequences are removed as these are most likely not coding for proteins.

4. **Shortening of ORFs using the evidence** Until this step the ORFs do not necessarily have a valid start codon. It can occur that only some regions of the ORFs are supported by similarities to known protein sequences. Therefore the ORFs are shortened so that they have a valid gene start and that most of the ORF is supported by similarities to known protein sequences. First the region supported by BLAST hits is determined. This is done by marking each position of the ORF supported by at least one evidence as evidence region. Then potential start codons are searched upstream of the evidence region. If a potential ORF start has been identified that lies less than 30 nt upstream of the evidence start, this ORF start is taken as granted. If the potential ORF start is further away or no potential ORF start can be identified upstream of the evidence start, potential ORF starts are identified downstream of the evidence start. If a potential downstream ORF start is closer to the evidence start than the upstream gene start then this is the new gene start. If neither upstream nor downstream of the evidence a gene start can be identified then the ORF is deleted.

5. **Removal of ORFs overlapping with RNAs** All ORFs overlapping more than 6 nt with a RNA are removed as these ORFs most probably comprise false positive gene predictions. The truncation of the models is not possible anymore as this has been done in the previous step.

6. **Dissolve conflicts between overlapping ORFs** This process is done in two steps. First the cases of ORFs overlapping with their C-termini are dissolved, then other overlaps. This is done as two ORFs overlapping with their C-termini are clear problem cases. Other cases might be dissolved automatically after these C-terminal cases have been dissolved.

   As discussed before the ORF with better evidence (see section 2.1.3.2.7) is kept. One of the points criticized about CONSORF is that the decision about overlaps is based only on the sum of bitscores without weighting. It happens in almost every genome that the sums of bitscores of two overlapping ORFs differ by just a few bits. In these cases a manual inspection is inevitable, in other cases where the sum of bitscores varies by a specific factor the cases are quite clear and the ORF with much less evidence can be removed.

   Firstly ORFs overlapping with their C-termini are dissolved, then the other overlaps: If more than 16% of the lengths of one of the two ORFs is overlapping with the other ORF this conflict has to be dissolved. The dissolving of the overlaps is done in two rounds with two different cutoffs. First a ratio of 10.0 is applied, then of 6.0. If the sum of bitscores of one ORF is 10.0 or rather 6.0 times higher than of the other ORF then the ORF with the lower sum of bitscores is removed, otherwise a warning is printed out and the overlap cannot be dissolved automatically but needs manual curation.

After these steps a set of non-overlapping gene models deferred by homology exists. The cases with warnings need to be checked manually later.

7. **Execution of intrinsic gene finders** GeneMarkS [31] gene predictions are created for each of the contigs. For this a self trained prokaryotic model is created based on the contig sequence. Then GeneMarkS is executed using RBS information and the combined model file, integrating heuristic information (based on the GC content of the contig) and the previously specifically trained model. If the contig is not long enough for training then only the heuristic model is used.

   Glimmer 3.0 [34] gene predictions are created for each of the contigs. First the coordinates of long ORFs of at least 500 nt length are extracted. If this is not possible it is tried to extract ORFs shorter than 500 nt going down in 10 nt steps until it is possible. The sequences of the long ORFs are extracted from the contig files using the previously determined coordinates of the long ORFs and a model for the gene prediction is trained on them. Afterwards Glimmer 3.0 is executed using a maximum overlap length of 50 nt (overlaps between genemodels this short or shorter are ignored), a minimum gene length of 110 nt, and a standard threshold score for gene calling of 30.

8. **Creation of the consensus prediction** Gene models deferred by homology as well as gene models from intrinsic gene finders are available at this point. The best gene starts need to be determined and potential conflicts between the models need to be dissolved. This is done in several steps:

   a) **Grouping of gene models** Gene models from the various predictions are grouped by strand and stop coordinate as the gene stop is a reliable feature in contrast to the gene start. Each of these groups represents a potential gene with one gene stop and at least one potential gene start (see Figure 2.3).

   b) **Determination of most probable gene starts** For each of the groups the decision has to be made which of the possible starts is chosen.

      If there is no homology deferred gene model for this group available then the gene model of the intrinsic method the annotator trusts most is selected. In the current version GeneMarkS is chosen as more reliable as judged by the gene prediction in *Escherichia coli* K12 compared to Glimmer 3.0 (Table 2.2).

      If there exists a gene model deferred by homology for this group and the best 10 hits of this gene model agree in one start coordinate and all these 10 hits have more than 50% sequence similarity with the gene model, then this start is taken in accordance with the evidences. This rule has been introduced in order to reduce the influence of weak hits on the determination of the gene start.

If not all the evidences agree in a start or an evidence did not fulfill the percent identity cutoff, then a search for the closest start of an intrinsic gene prediction to the evidence is executed. This is done as the intrinsic methods use features of the promoter region of genes for the prediction. Therefore not the nearest possible start to the evidence but the nearest start predicted by an intrinsic gene finder is selected as gene start. If the overlap with the evidence is smaller than 15 nt then the intrinsic start is taken, otherwise the nearest possible start to the evidence is taken as this is the best solution in that case. After this step there is a set of gene models grouped by strand and stop coordinate and each of these groups has one representative gene model.

c) **Dissolving of overlaps between gene models** Now the eventually existing conflicts between the gene models have to be dissolved. This is done in several steps:

   - Gene models shorter than 100 nt are removed in order to avoid false positive predictions.
   - Overlaps of gene models with RNAs are dissolved as described previously.
   - ORFs overlapping with their C-termini are dissolved, then all other overlaps are resolved: If more than 16% of the lengths of one of the two ORFs is overlapping with the other ORF this conflict has to be dissolved. The dissolving of the overlaps is done in two rounds with two different cutoffs. First a ratio of 10.0 is applied, then of 6.0. If the sum of bitscores of one ORF is 10.0 or rather 6.0 times higher than of the other ORF then the ORF with the lower sum of bitscores is removed, otherwise a warning is printed out and the overlap cannot be dissolved automatically but needs manual curation.

9. **Output of the predicted sequences and of the cases to be checked manually** The final gene models have been created at this timepoint and are outputted into a multifasta file. The description line contains information about the way the gene model has been built, e.g. which intrinsic prediction was used and how evidence influenced the gene start decision and also the information whether this gene is complete. The cases to be checked manually are outputted into a log file.

10. **Identification of possible frame shifts** As last step a scan over the whole prediction is performed to find neighboring gene models with shared protein hits. The same hits in adjacent genes can be a hint towards frame shifts that might have split the original gene.

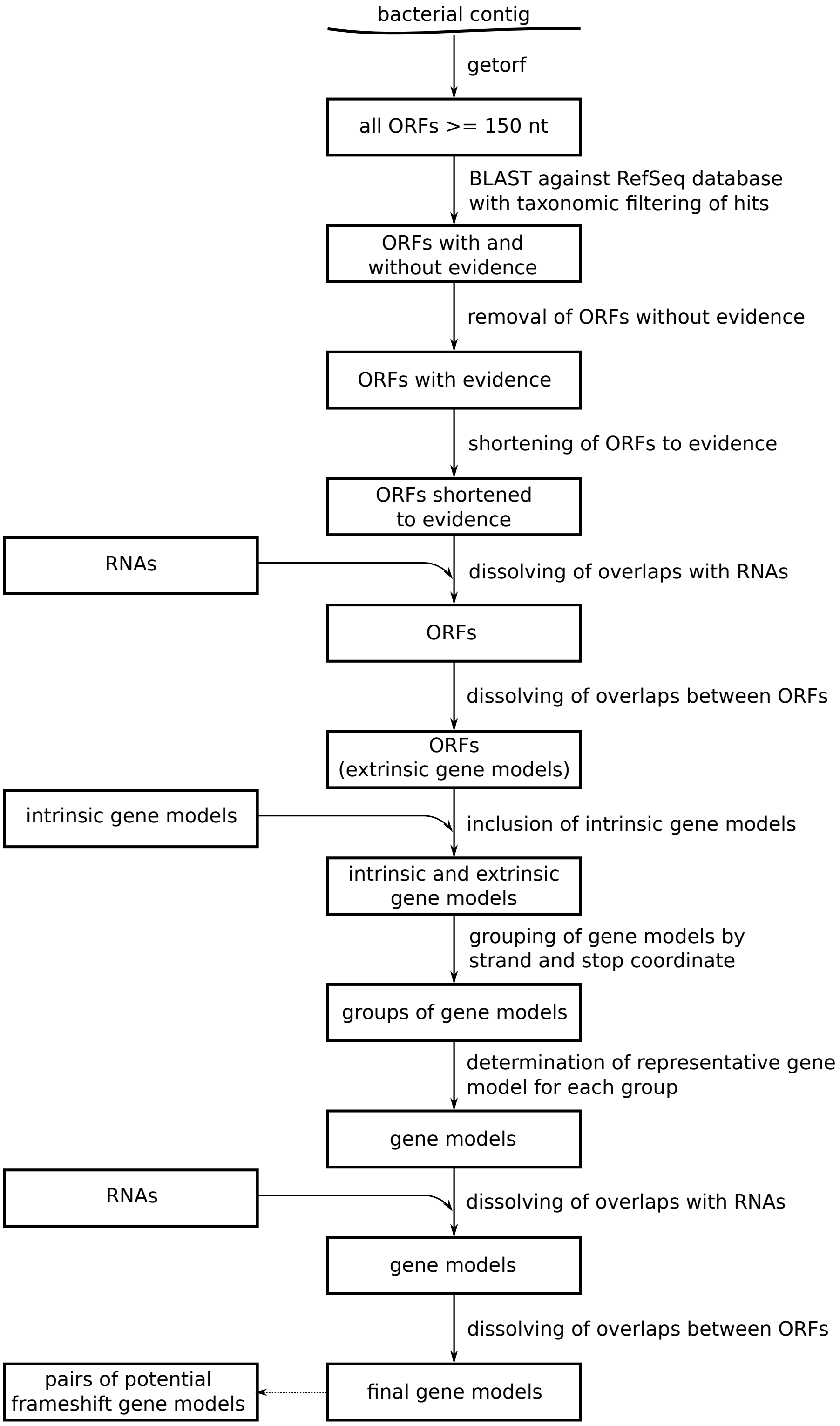

bacterial contig
getorf
all ORFs >= 150 nt
BLAST against RefSeq database
with taxonomic filtering of hits
ORFs with and
without evidence
removal of ORFs without evidence
ORFs with evidence
shortening of ORFs to evidence
ORFs shortened
to evidence
RNAs
dissolving of overlaps with RNAs
ORFs
dissolving of overlaps between ORFs
ORFs
(extrinsic gene models)
intrinsic gene models
inclusion of intrinsic gene models
intrinsic and extrinsic
gene models
grouping of gene models by
strand and stop coordinate
groups of gene models
determination of representative gene
model for each group
gene models
RNAs
dissolving of overlaps with RNAs
gene models
dissolving of overlaps between ORFs
pairs of potential
frameshift gene models
final gene models

Figure 2.6: Overview over the gene prediction pipeline

prediction method	number of predicted genes with correct gene start and stop	number of predicted genes with wrong gene start and correct gene stop
Glimmer 3.0	70	9
GeneMarkS	72	7
ConsPred	72	7

Table 2.2: Benchmark of gene predictions for verified genes of *Escherichia coli* K12 MG1655 The table shows the concordance of the gene predictions of Glimmer3 [34], GeneMarkS [31] and the new gene prediction pipeline ConsPred with 79 experimentally verified genes [53]. GeneMarkS and ConsPred perform equally well on the test set.

2.1.3.5 Evaluation

In order to evaluate the performance of different gene prediction programs, the gene prediction was performed for *Escherichia coli* K12 MG1655 using Glimmer 3.0 [34], GeneMarkS [31] and ConsPred.

As standard of truth the experimentally verified genes of *E. coli* were used [53]. As the genome sequence changed since the publication of the experimentally verified genes and only gene coordinates are available, only the first 79 of the experimentally verified genes can be used for the benchmark.

The results for the predictions in numbers can be seen in Table 2.2. GeneMarkS and ConsPred perform equally well and better than Glimmer 3.0 on the dataset.

As the amount of data about verified gene starts is very limited and as the gene models deposited in public databases are biased towards the gene finders used by the submitters, it is unfortunately currently not possible to judge the true performance of any gene prediction method. Even if a prediction method predicts the true genes these true genes might be falsely annotated in the public databases and a benchmark against this data is not biologically meaningful.

This is the reason why Thomas Weinmaier from the Department for Computational Systems Biology at the University of Vienna is continuing the ConsPred project. The goals are to prepare a comprehensive set of validated genes throughout many organisms and to adapt ConsPred so that it outperforms currently existing methods in the *E. coli* as well as in the comprehensive dataset.

Even if a comprehensive evaluation is missing due to insufficient data, many genes predicted by ConsPred have been carefully manually examined in collaborations (section 2.1.3.6) and ConsPred has shown to provide reasonable results.

2.1.3.6 Application of the gene prediction pipeline

ConsPred has been applied to a number of genomic sequences ranging from BACs over bacterial and archaeal genomes to metagenomes. Table 2.3 presents an overview over the numbers of predicted genes in the respective genomes as well as an overview over the differences between the predictions using Glimmer 3.0, GeneMarkS and the new prediction pipeline ConsPred.

In 11 of 18 cases the new pipeline predicts more genes than the other two methods as it combines the sensitivity of the homology based approach with the models of the two intrinsic gene finders. Nevertheless there are also three cases where the number of ConsPred predictions lies between the numbers of predicted genes of Glimmer 3.0 and GeneMarkS and in four cases ConsPred even predicts less genes than the other two methods, probably due to the conflict dissolving.

All three prediction methods predict differing gene starts, Glimmer 3.0 and GeneMarkS differ in 822 cases averaged over all genomes, ConsPred and Glimmer 3.0 differ in 1240 and ConsPred and GeneMarkS in 926 cases. This clearly shows the need for an improved gene start prediction.

ConsPred was precious for the annotation of the genomes of the collaborations as only a small number of conflicts had to be inspected prior to the submission of the data to public databases.

2.1.3.7 Discussion

ConsPred has been implemented in order to automatically resolve clear cases of differences between the gene predictions of various gene finders consisting of differing gene starts and overlapping gene models. These conflicts between the gene models can be dissolved in many cases by the integration of extrinsic information in the form of BLAST hits to published protein sequences. Only cases not clearly resolvable are left for manual annotation. Thus ConsPred minimizes the manual effort necessary for the gene prediction in prokaryotic genomes and comes to the same decisions as a human annotator in clear problem cases.

ConsPred supported us with gene predictions in many genome projects and performs as good on a dataset with validated gene starts as GeneMarkS and better than Glimmer 3.0. A reliable examination of the performance of ConsPred in organisms besides *Escherichia coli* could not be done yet, due to limited data on validated genes, especially on validated gene starts. Therefore the next steps will be the compilation of a comprehensive set of genes with validated gene starts. But also with such a set there remains the problem that some genes can have different gene starts depending on their regulation, so that there might be not just one correct gene start.

2.1.4 Orthologous groups of proteins

2.1.4.1 Motivation

The knowledge about the membership of a specific protein in an orthologous group (section 1.3.2.3.2) can help to transfer annotations from proteins with annotations to the other members of the same group. This can be done as proteins within such a group often share the same functions. Therefore orthologous groups are of interest for the annotation of the genomes of the collaboration partners. As these genomes are

organism	number genes Glimmer 3.0	number genes GeneMarkS	number genes ConsPred	number identical genes between all three predictions	number identical genes between Glimmer 3.0 and GeneMarkS	number identical genes between Glimmer 3.0 and ConsPred	number identical genes between GeneMarkS and ConsPred	number changed starts Glimmer 3.0 vs GeneMarkS	number changed starts ConsPred vs Glimmer 3.0	number changed starts ConsPred vs GeneMarkS	number genes not in Glimmer 3.0 compared to GeneMarkS	number genes not in Glimmer 3.0 compared to ConsPred	number genes not in GeneMark compared to Glimmer 3.0	number genes not in GeneMarkS compared to ConsPred	number genes not in ConsPred compared to Glimmer 3.0	number genes not in ConsPred compared to GeneMarkS
AciFil	70	72	78	45	52	53	54	16	17	18	4	8	2	6	0	0
BenFer	12158	16651	15503	3258	5543	3859	6987	3339	4795	4499	7285	6454	2792	4011	3109	5159
ClaSal	1033	1050	1059	571	759	611	770	202	369	250	85	75	68	39	49	30
CroTur	4440	4339	4454	2194	3250	2550	2697	885	1757	1520	203	146	304	237	132	122
CurHyd	4091	3947	4017	1350	2276	1733	1960	1323	2039	1699	343	241	487	358	315	288
EntHel	176	172	162	92	116	109	114	42	46	40	12	7	16	8	21	18
LepII	1127	1116	1085	372	592	447	622	341	524	374	183	114	194	89	156	120
LepIII	2927	2691	2668	586	876	1023	1431	794	1045	595	1020	599	1256	642	858	665
N47	5154	5216	5244	2354	3509	2732	3143	1114	2076	1701	586	431	524	400	341	372
NitDef	4268	4093	4229	1685	2826	1895	2263	1110	2209	1734	157	125	332	232	164	96
NitGar	3517	3451	3238	1585	1817	1846	2746	1361	1210	358	268	181	334	133	460	346
NitVie	3081	3135	3210	1184	1553	1414	2151	1297	1529	883	273	255	219	176	126	101
ParAca	2818	2795	2854	1576	2187	1733	1913	443	984	797	165	137	188	144	101	85
ProAmo	2080	2115	2159	1123	1593	1247	1417	345	764	618	177	148	142	124	69	80
SimNeg	2475	2497	2509	1376	1841	1475	1771	520	924	668	135	109	113	70	75	58
TTL1	58	62	77	30	40	37	43	9	16	15	13	24	9	19	5	4
WadCho	2026	2022	2070	919	1335	1056	1284	505	853	639	168	148	172	147	104	99
YerEnt	4890	4912	5008	3147	3264	3414	4523	1152	1175	261	448	374	426	223	256	127

Table 2.3: Overview over the gene predictions in various organisms. AciFil Acidianus filamentous virus AFV10, **BenFer** Benzol Ferrihydrite culture, **ClaSal** Candidatus Clavochlamydia salmonicola, **CroTur** Cronobacter turicensis LMG 23827, **CurHyd** Curvibacter putative symbiont of Hydra magnipapillata, **EntHel** Enterobacter helveticus (BACs), **LepII** Leptospirillum sp. Group II, **LepIII** Leptospirillum sp. Group III, **N47** Deltaproteobacterial enrichment culture N47, **NitDef** Candidatus Nitrospira defluvii, **NitGar** Candidatus Nitrososphaera gargensis, **NitVie** Nitrosofabula viennensis, **ParAca** Parachlamydia acanthamoebae UV7, **ProAmo** Protochlamydia amoebophila UWE25, **SimNeg** Simkania negevensis Z, **TTL1** Symbiont TTL1 of the Adelges nordmannianae/piceae complex, **WadCho** Waddlia chondrophila 2032/99, **YerEnt** Yersinia enterocolitica W22703.

unpublished and therefore not contained in the publicly available clusters, there are two possible solutions. The first one is to search for homologous sequences to a query protein in the public orthologous groups and to assign the protein to the existing orthologous group according to specific criteria, e.g. assignment to the cluster with the most similar hit. Another way is the creation of own orthologous groups.

One of the reasons for the construction of own orthologous groups is that the presence of the sequences of the unpublished organisms at the time of the creation of the clusters can have influence on the initial orthologous groups. It can happen that clusters unique for the unpublished organisms might be split when only assigning their sequences to existing clusters. Another reason for own orthologous groups is that orthologous groups sometimes should not only be built between all organisms but only between a distinct group of organisms.

2.1.4.2 Construction of own orthologous groups

Because of the already stated reasons own orthologous groups have been constructed. As the bidirectional best hits (BBHs) and unidirectional best hits of all against all organisms are already precomputed and available in the ComparDB (section 1.3.2.3.2.2) no expensive computations had to be done.

We wanted to be able to make statements about the specificity of a cluster in relation to other groups of organisms, that is to have the information how many members of a group of organisms that should be clustered are contained in each cluster and how many hits to organisms not contained in this group exist. The less hits to other organisms exist the more specific is a cluster for the group of organisms that should be clustered.

I developed the first prototype for the clustering approach in the scripting language Perl. Marc-André Jehl and myself started with the implementation of the production version in Java and Thomas Rattei refactored and extended the Java version so that it was possible to work with the huge amounts of data needed for analyses in other projects.

A short overview over the algorithm is given below. The following definitions are necessary for the understanding:

Ingroup Ingroups contain organisms for which a full clustering should be applied. Uni- and bidirectional hits of all proteins from within the ingroup against all proteins within the ingroup are needed for these organisms.

Outgroup Outgroups contain organisms against which only bidirectional best hits (BBHs) should be computed from an ingroup protein to an outgroup protein. BBHs between outgroup organisms are not necessary.

The ingroup genomes and outgroup genomes can be grouped for the later taxonomic evaluation. If one wants to know how specific clusters for the phyla A, B, C are and how many members of other phyla are covered in the clusters of these three phyla for example, then A, B, C would be the ingroups with their respective organisms and all other phyla would be outgroups with their respective organisms.

- **Input** The input are ingroups and outgroups. If only ingroups are given and each ingroup consists of one organism then this is the same as a conventional clustering based on BBHs.

- **Retrieval of BBHs** BBHs with a certain E-value and length-ratio cut-off are retrieved from ComparDB. First all BBHs between organisms of the ingroups are extracted, then BBHs of members of the ingroups with members of the outgroups. BBHs between members of the outgroups are not necessary.

- **Creation of the ingroup clusters** BBHs between proteins from ingroup organisms are merged to form one cluster if they share at least one protein.

- **Addition of outgroup proteins to the clusters** Outgroup proteins forming BBHs to ingroup proteins are added to the clusters
- **Addition of in-paralogs** In-paralogs (i.e. paralogs that arose after diversification) [106] are added to the clusters if the respective protein shows a higher similarity to the BBH protein from the same organism than to proteins from other ingroup organisms.
- **Taxonomic analysis of the clusters** Each cluster is analyzed for its taxonomic composition, that is how many and which of the ingroups are contained in each of the clusters and how many and which of the outgroups are covered in each of the clusters. This allows to determine how specific a cluster is for the ingroups.
- **Output** The output is a file with clusters and their taxonomic composition.

2.1.4.3 Application of the orthologous groups

The results of this program have been used in several projects. One example is *Cronobacter turicensis* (section 2.2), the other example are members of the phylum *Chlamydiae* (section 2.1.4.3.1).

2.1.4.3.1 Prediction of outer membrane proteins in the phylum *Chlamydiae* using orthologous groups Even though chlamydial outer membrane proteins (OMP) play an important role for attachment to and entry into host cells, only few had been described. Therefore Eva Heinz from the Department of Microbial Ecology at the University of Vienna developed a comprehensive, multiphasic in-silico approach to predict OMPs [249].

Eva Heinz applied her prediction pipeline (Figure 2.7) to five chlamydial proteomes from two human pathogens (*Chlamydia trachomatis* D/UW3/CX, *Chlamydophila pneumoniae* AR39), two animal pathogens (*Chlamydia muridarum* Nigg, *Chlamydia caviae* GPIC) and an amoeba symbiont (*Protochlamydia amoebophila* UWE25) [249]. As she observed that the membrane predictions were in general more heterogeneous and less well defined for chlamydial outer membrane proteins as for outer membrane proteins of *Escherichia coli*, the idea arose to use the predictions for members of the same orthologous group in order to resolve uncertain predictions and by that to make the predictions more reliable.

I built orthologous groups using bidirectional best hits (BBHs) with an E-value cut-off of 10^{-08} and a length ratio cut-off of 0.5. The E-value cutoff has been determined empirically by manual inspection of known protein families in the resulting clusters by Eva Heinz and Matthias Horn from the Department of Microbial Ecology at the University of Vienna. All *Chlamydiae* including the at that time point unfinished genomes of *Parachlamydia acanthamoebae* UV7, *Simkania negevensis* Z, and *Waddlia chondrophila* 2032/99 were defined as ingroups. Additionally clusters for a selection of *Proteobacteria* including *E. coli K12* as “ingroup” were created. 438 and 427 representatives of other bacterial lineages were considered “outgroup“ organisms, respectively.

When clustering was applied using the *Chlamydiae* as ingroup, 1911 clusters were obtained in total, from which 190 contained at least one protein predicted as outer membrane protein. 81 of these clusters included two or more proteins from the five analyzed *Chlamydiae*, but not all of these proteins were predicted to be located in the outer membrane.

Eva conducted OMP predictions for the chlamydial proteins, evaluated the predictions within the orthologous groups, and investigated the phylogentic conservation of the identified membrane proteins. In total, 312 chlamydial outer membrane proteins and lipoproteins in 88 orthologous clusters could be identified, including 238 proteins not previously recognized to be located in the outer membrane. The analysis of the taxonomic distribution of the clusters revealed an evolutionary conservation among *Chlamydiae*, *Verrucomicrobia*, *Lentisphaerae* and *Planctomycetes* as well as lifestyle-dependent conservation of the chlamydial outer membrane protein composition. This is in compliance with the observation that *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* and sister phyla comprise a superphylum [271].

2.1.5 Pseudogene detection in prokaryotic genomes

2.1.5.1 Motivation

The knowledge about pseudogenes (section 1.2.3.1) can give a hint towards the stability of a genome and whether the genome currently undergoes an adaption to a different environment, for example. The pseudogene prediction using the $\Psi - \Phi$ [61] software has been important in several genome projects, among them *Amoebophilus asiaticus* 5a2 and *Cronobacter turicensis* LMG 23827.

2.1.5.2 Pseudogene detection in *Amoebophilus asiaticus* 5a2

Amoebophilus asiaticus 5a2 belongs to the phylum *Bacteroidetes*. In order to learn about ongoing genome evolution of this organism pseudogenes should be predicted. As the most straightforward method to detect pseudogenes is the search for truncated coding sequences, the published $\Psi - \Phi$ software [61] was applied, that uses a set of informant genomes for the detection of these truncated coding genes. The next related genomes have less than 85% similarity on the 16S rRNA level which is very different as these next related organisms do not belong to the same taxonomic family anymore. For the pseudogene detection using $\Psi - \Phi$ all available genomes from the phylum have been used. As of December 2007 these genomes were *Bacteroides fragilis* YCH46, *Bacteroides thetaiotaomicron* VPI-5482, *Cytophaga hutchinsonii* ATCC 33406, *Gramella forsetii* KT0803, *Porphyromonas gingivalis* W83 and *Salinibacter ruber* DSM13855.

$\Psi - \Phi$ evaluates two parameters for the decision whether two protein sequences are regarded as homologous. These two parameters are the E-value threshold and the percentage identity of the two sequences. The original paper [61] used an E-value cutoff of 10^{-15} and a percentage of protein identity of $> 79\%$. If these settings are used, only 2 genes are predicted as potential pseudogenes in *A. asiaticus*. The most probable reason

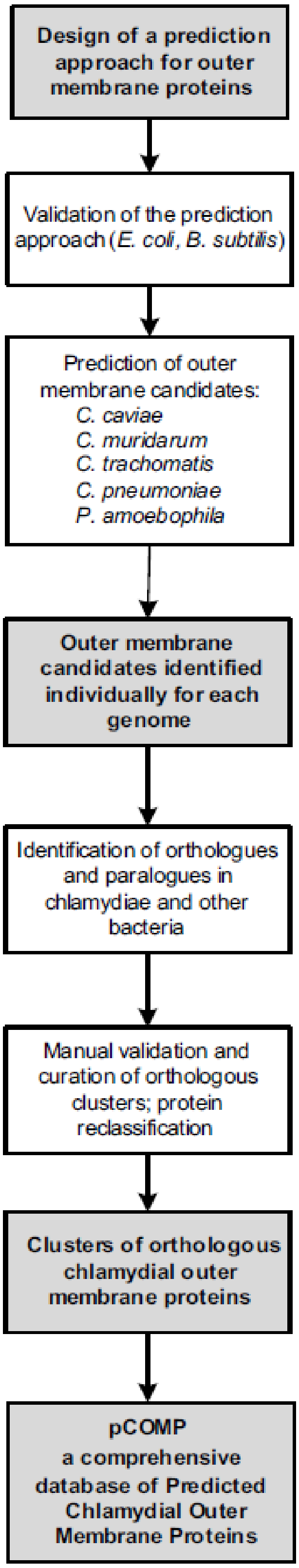

Design of a prediction approach for outer membrane proteins
Validation of the prediction approach (*E. coli*, *B. subtilis*)
Prediction of outer membrane candidates:
C. caviae
C. muridarum
C. trachomatis
C. pneumoniae
P. amoebophila
Outer membrane candidates identified individually for each genome
Identification of orthologues and paralogues in chlamydiae and other bacteria
Manual validation and curation of orthologous clusters; protein reclassification
Clusters of orthologous chlamydial outer membrane proteins
pCOMP
a comprehensive database of Predicted Chlamydial Outer Membrane Proteins

Figure 2.7: The approach to identify chlamydial outer membrane proteins (Source: [249])

organism	chromosome size	number coding genes	number putative pseudo-genes
Amoebophilus asiaticus 5a2	1,884,364	1557	222
Sulcia muelleri GWSS	245,530	228	-
Azobacteroides pseudotrichonymphae CFP2	1,114,206	758	22
Flavobacterium psychrophilum JIP02/86	2,861,988	2432	20
Bacteroides thetaiotaomicron VPI-5482	6,260,361	4779	-

Table 2.4: Number of pseudogenes in *Amoebophilus asiaticus* 5a2 and other members of the *Bacteroidetes*. Adapted from [259]

is that the used *Bacteroidetes* species are more distantly related to each other than the used *Escherichia coli* strains in the paper. In the follow up $\Psi - \Phi$ paper [272] the authors analyzed the inventory of pseudogenes in bacterial pathogens from multiple taxonomic groups. In order to get meaningful results they set the E-value threshold to 10^{-10} and the percentage of protein identity to $> 49\%$.

As the comparison genomes for *A. asiaticus* are even more distantly related from each other than the comparison organisms in the latter paper these settings resulted in still only 15 predicted pseudogenes. We therefore tested various settings and decided to better be more sensitive and to do a manual analysis of the pseudogene candidates afterwards. We chose a percentage of identity of $> 29\%$ and an E-value of $\leq 10^{-15}$ for the prediction using $\Psi - \Phi$ as we still wanted meaningful hits. This resulted in a list of 113 pseudogene candidates.

In order to review the pseudogene candidates from $\Psi - \Phi$, the best BLAST [73] hits for all protein sequences of *A. asiaticus* against a non-redundant database of publicly available protein sequences were extracted. Hits with a percent identity $> 29\%$ and a length ratio $> 79\%$ like for the $\Psi - \Phi$ run were considered possible pseudogene candidates. The advantage of the BLAST approach is that the hit sequences do not necessarily have to originate from the informant genomes used within $\Psi - \Phi$, so that this approach should be more sensitive. The alignments of all candidates from $\Psi - \Phi$ as well as the best BLAST hit candidates were manually checked by Stephan Schmitz-Esser. This resulted in the final list of 222 pseudogene candidates [259], that is 14.26% of all coding sequences. This is much in comparison to the numbers in other members of the phylum *Bacteroidetes* (Table 2.4).

The relatively high number of predicted pseudogenes is not astonishing as the genome of *A. asiaticus* shows a massive proliferation of insertion sequence (IS) elements (24% of all genes). The spreading of IS elements has been reported to result in proliferation of pseudogenes, genome rearrangements, and finally genome reduction [273, 274]. Despite the high percentage of IS elements the genome has not been extensively reshuffled recently but rather has remained stable for an extended evolutionary time period.

2.1.5.3 Pseudogene detection in *Cronobacter turicensis* LMG 23827

As *Cronobacter turicensis* LMG 23827 is a member of the *Gammaproteobacteria* with many closely related genomes available $\Psi-\Phi$ [61] was applied with standard parameters. This resulted in 122 pseudogene candidates (Table 2.5). This number is comparable to the number of pseudogenes in other members of the family *Enterobacteriaceae*.

name	description	name	description
Ctu_00460	2-dehydro-3-deoxy-6-phosphogalactonate aldolase	Ctu_22180	Glycine betaine/carnitine/choline transport ATP-binding protein opuCA
Ctu_00470	D-galactonate dehydratase	Ctu_22220	Inner membrane transport protein ynfM
Ctu_00570	6-phospho-alpha-glucosidase	Ctu_23110	Uncharacterized protein ycjR
Ctu_01420	Formate dehydrogenase-O major subunit	Ctu_23530	Ribosomal large subunit pseudouridine synthase B
Ctu_01430	hypothetical protein	Ctu_24010	Respiratory nitrate reductase 1 alpha chain
Ctu_02050	Uncharacterized protein yifB	Ctu_24680	Bactoprenol glucosyl transferase homolog from prophage CPS-53
Ctu_02070	Acetolactate synthase isozyme 2 large subunit	Ctu_24710	hypothetical protein
Ctu_02800	Trk system potassium uptake protein trkH	Ctu_24720	hypothetical protein
Ctu_03140	Bifunctional purine biosynthesis protein purH	Ctu_25170	Uncharacterized protein yebT
Ctu_03170	Putative amino-acid ABC transporter permease protein yhdX	Ctu_25340	Glucose-6-phosphate 1-dehydrogenase
Ctu_03360	UPF0190 protein ESA_03641	Ctu_25460	UPF0082 protein ESA_01378
Ctu_03390	unknown protein	Ctu_25500	Isochorismatase family protein yecD
Ctu_03520	Uncharacterized protein yhcN	Ctu_25710	Flagellar biosynthesis protein flhA
Ctu_03710	Monofunctional biosynthetic peptidoglycan transglycosylase	Ctu_25910	Alpha,alpha-trehalose-phosphate synthase [UDP-forming]
Ctu_04040	Dihydropteroate synthase	Ctu_25950	L-arabinose-binding periplasmic protein
Ctu_04070	Protein-export membrane protein secG	Ctu_26250	RNA polymerase sigma factor for flagellar operon
Ctu_04350	Galactitol-1-phosphate 5-dehydrogenase	Ctu_26550	Flagellar M-ring protein
Ctu_04510	Uncharacterized protein yqjG	Ctu_26740	Inner membrane protein yedI
Ctu_05350	Glycerol dehydrogenase	Ctu_27450	Phosphomannomutase
Ctu_05860	DNA polymerase III subunit psi	Ctu_27530	Putative colanic acid polymerase
Ctu_06000	Lipoate-protein ligase A	Ctu_27560	Putative colanic acid biosynthesis glycosyl transferase wcaA
Ctu_06820	Protein apaG	Ctu_27710	Multidrug resistance protein mdtB
Ctu_06940	L-arabinose isomerase	Ctu_27730	Putative multidrug resistance protein mdtD
Ctu_06960	Arabinose operon regulatory protein	Ctu_28190	Galactose/methyl galactoside import ATP-binding protein mglA
Ctu_07080	Probable HTH-type transcriptional regulator leuO	Ctu_29390	3-octaprenyl-4-hydroxybenzoate carboxy-lyase
Ctu_07640	Hypoxanthine phosphoribosyltransferase	Ctu_29940	Inner membrane protein ypdA
Ctu_08750	unknown protein	Ctu_30190	Sulfate transport system permease protein cysW
Ctu_09160	hypothetical protein	Ctu_30830	Phosphoribosylglycinamide formyltransferase
Ctu_10070	Ubiquinol oxidase subunit 1	Ctu_31260	Chaperone protein hscA
Ctu_10080	Ubiquinol oxidase subunit 2	Ctu_31650	L-aspartate oxidase
Ctu_10270	Protein crcB homolog	Ctu_32070	hypothetical protein
Ctu_10710	HTH-type transcriptional regulator acrR	Ctu_32150	Late control gene D protein
Ctu_11000	Acyl-CoA thioesterase I	Ctu_32200	Major tail sheath protein
Ctu_12440	Uncharacterized zinc-type alcohol dehydrogenase-like protein ybdR	Ctu_32840	Multidrug resistance protein A
Ctu_12890	Protein nagD	Ctu_34380	Peptide chain release factor 2
Ctu_13080	Ornithine decarboxylase, inducible	Ctu_34970	Probable quinol monooxygenase ygiN
Ctu_13100	KDP operon transcriptional regulatory protein kdpE	Ctu_35050	Malonate decarboxylase acyl carrier protein
Ctu_13110	Sensor protein kdpD	Ctu_36200	hypothetical protein
Ctu_13130	Potassium-transporting ATPase B chain	Ctu_36460	UPF0131 protein ytfP
Ctu_13150	Uncharacterized protein ybfA	Ctu_37380	Acetyl-coenzyme A synthetase
Ctu_14590	Oxygen-insensitive NADPH nitroreductase	Ctu_37510	Single-stranded DNA-binding protein
Ctu_15030	Outer-membrane lipoprotein carrier protein	Ctu_37990	Uncharacterized protein yjbB
Ctu_15320	Chromosome partition protein mukB	Ctu_38020	Isocitrate dehydrogenase kinase/phosphatase
Ctu_15550	ABC transporter ATP-binding protein uup	Ctu_38040	Malate synthase A
Ctu_16160	Curli production assembly/transport component csgG	Ctu_38050	Homoserine O-succinyltransferase
Ctu_16540	Flagellar basal-body rod protein flgF	Ctu_38870	Putative fructoselysine transporter frlA
Ctu_16600	Flagellar hook-associated protein 3	Ctu_39300	HTH-type transcriptional regulator malT
Ctu_17140	Isocitrate dehydrogenase [NADP]	Ctu_39520	Gamma-glutamyltranspeptidase
Ctu_17500	unknown protein	Ctu_39550	sn-glycerol-3-phosphate import ATP-binding protein ugpC
Ctu_17690	hypothetical protein	Ctu_40340	Cellulose synthase catalytic subunit [UDP-forming]
Ctu_17900	unknown protein	Ctu_40730	Xylose isomerase
Ctu_18160	Succinylornithine transaminase	Ctu_40780	Xylose operon regulatory protein
Ctu_18750	Cysteine desulfurase	Ctu_40830	Inner membrane symporter yicJ
Ctu_19090	hypothetical protein	Ctu_40980	Uncharacterized zinc-type alcohol dehydrogenase-like protein yahK
Ctu_19100	Formate dehydrogenase H	Ctu_41760	Uncharacterized sugar isomerase yihS
Ctu_19680	Multidrug resistance protein mdtK	Ctu_41930	Probable acyltransferase yihG
Ctu_20110	Uncharacterized oxidoreductase ydgJ	Ctu_42030	D-ribose-binding periplasmic protein
Ctu_20560	Inner membrane ABC transporter permease protein ydcU	Ctu_42050	Ribose import ATP-binding protein rbsA
Ctu_20810	unknown protein	Ctu_1p01200	unknown protein
Ctu_22020	Inner membrane transport protein yjjL	Ctu_3p00350	unknown protein
Ctu_22060	Spermidine N(1)-acetyltransferase	Ctu_3p00500	Probable sensor protein pcoS

Table 2.5: Pseudogene candidates as predicted by $\Psi - \Phi$ for *Cronobacter turicensis* LMG 23827.

2.2 Comparative analysis of the whole genomes of *Cronobacter turicensis* and *Cronobacter sakazakii*, two opportunistic pathogens

2.2.1 Motivation

A few years ago a cooperation with Angelika Lehner and Roger Stephan from the Institute for Food Safety and Hygiene, Vetsuisse Faculty, University of Zurich could be established. We already had colaborated on several projects when Roger Stephan and Angelika Lehner offered us the opportunity to annotate the genome of *Cronobacter turicensis* LMG 23827. *C. turicensis* is a Gram-negative opportunistic foodborne pathogen and known as rare but important cause of live-threatening neonatal infections. As *Cronobacter* sp. infections can lead to severe disease manifestations and only little is known about the mechanisms of pathogenicity and persistence in dry environments of *Cronobacter* spp. the whole genome of *Cronobacter turicensis* LMG 23827 had been sequenced.

2.2.2 *Cronobacter turicensis*

2.2.2.1 The genome of *Cronobacter turicensis*

In order to get an overview over the general features of the genome of *C. turicensis* a variety of techniques needed to be applied.

2.2.2.1.1 Methods

2.2.2.1.1.1 Genome sequences The genome sequences of *Cronobacter turicensis* LMG 23827 have been sequenced using the 96-capillary 3730xl DNA Analyzer from Applied Biosystems. The sequences are available under RefSeq [52] accessions NC_013282 - NC_013285 and GenBank accessions FN543093 - FN543096.

As the sequencing of the rRNA operons was difficult in *C. turicensis*, firstly only one of the rRNA operons had been sequenced and copied to the positions of the six other rRNA operons. Some of the tRNAs in *C. turicensis* as well as in its next relative *C. sakazakii* are encoded in the spacers between the single member rRNAs of the rRNA operons. Therefore at first no tRNA for glutamic acid could be detected. As it is very unlikely that an organism lacks a tRNA for glutamic acid and the predicted as well as the measured proteins [51] contained glutamic acid, a resequencing of all seven rRNA operons was done. The newly sequenced rRNA operons have been integrated into the genome sequence. The tRNA for glutamic acid is contained within the 6th rRNA operon.

2.2.2.1.1.2 Gene prediction The gene prediction was performed using the novel gene prediction pipeline ConsPred (see section 2.1.3.4). Hits against proteins from the family

Enterobacteriaceae were excluded as BLAST [73] hits in the similarity search against known protein sequences.

2.2.2.1.1.3 Automated annotation of protein sequences The PEDANT [72] system was used for many of the automated annotations. For the UniProtKB/Swiss-Prot [182] keyword annotation and the assignment of EC classifications an E-value cutoff of 10^{-05} against the proteins in UniProtKB/Swiss-Prot (October 2009) was used. Annotations with the MIPS Functional Catalogue 2.1 (FunCat) [81] were done using an E-value cutoff of 10^{-03} against the FunCat database of PEDANT. Signal peptides were predicted using SIGNALP [275] for Gram-negative organisms. GO terms were assigned as described in [276] using an E-value cutoff of 10^{-100} for the BLAST searches against sequences with assigned GO and for the BLAST searches against sequences with assigned EC or UniProtKB/Swiss-Prot keywords.

2.2.2.1.1.4 Homologies The determination of the number of proteins with homologies has been done using the SIMAP [94] database. For each of the protein sequences of the analyzed organisms, similar sequences belonging to bacteria (NCBI taxonomyid 2) [277] within UniProtKB (October 2009) [182] have been searched with an E-value cutoff of 10^{-10}. Hits against the genus *Cronobacter* (NCBI taxonomyid 413496) have been excluded in order to avoid hits to sequences from too closely related species.

The assignment of UniProtKB/Swiss-Prot hits for the proteins has been done performing a homology search against all sequences belonging to *Bacteria* (NCBI taxonomyid 2) within UniProtKB/Swiss-Prot (October 2009) and an E-value cutoff of 10^{-10}. Additionally the ratio of the alignment length to sequence length of each of the proteins had to be ≥ 0.5 and the sequences had to have ≥ 40.0 percent identity. These strict criteria were applied in order to make sure that no spurious annotations would be transferred.

2.2.2.1.1.5 tRNAs tRNAs have been detected using tRNAscan-SE 1.23 [278] using parameter -B for prokaryotic tRNA detection.

2.2.2.1.1.6 rRNAs The rRNA operons have been identified using BLASTN [73] and the rRNA sequences of *C. sakazakii* as queries. Hits with an E-value $\leq 3 \cdot 10^{-57}$ have been used for the 5S rRNAs and hits with an E-value ≤ 0.0 for 16S and 23S rRNAs.

2.2.2.1.1.7 Selection of representative *Enterobacteriaceae* In order to remove the bias towards genera with many sequenced genomes within the family *Enterobacteriaceae* one representative organism for each genus has been determined (Table 2.6) from the set of all complete publicly available bacterial genomes within RefSeq [52] (October 2009). This was needed for the comparison of *Cronobacter* spp. to other *Enterobacteriaceae* (e.g. for the determination of gene losses). The representative organism ideally should have as many proteins in common with as many other members of its genus as possible. For that reason clusters of orthologous groups were built for each genus (section 2.1.4.2) using an E-value cutoff of 10^{-04}. The member of the genus with the highest sum of

organism	NCBI taxonomy ID
Enterobacter sp. 638	399742
Erwinia carotovora subsp. atroseptica SCRI1043	218491
Erwinia tasmaniensis Et1/99	465817
Escherichia coli 55989	585055
Escherichia fergusonii ATCC 35469	585054
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	272620
Photorhabdus luminescens subsp. laumondii TTO1	243265
Proteus mirabilis HI4320	529507
Salmonella enterica subsp. enterica serovar Enteritidis str. P125109	550537
Salmonella typhimurium LT2	99287
Serratia proteamaculans 568	399741
Shigella boydii Sb227	300268
Shigella dysenteriae Sd197	300267
Shigella flexneri 2a str. 2457T	198215
Shigella sonnei Ss046	300269
Sodalis glossinidius str. morsitans	343509
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	36870
Yersinia enterocolitica subsp. enterocolitica 8081	393305
Yersinia pestis Antiqua	360102
Yersinia pseudotuberculosis PB1/+	502801

Table 2.6: Representative *Enterobacteriaceae*

the following conditions has been selected: Firstly the percentage of proteins in clusters covering at least 50% percent of the other species and secondly the percentage of clusters with proteins of the potential representative and at least 50% percent of the other species contained in the cluster.

2.2.2.1.1.8 Pseudogenes Pseudogenes have been identified using the software $\Psi - \Phi$ [61] (section 2.1.5). The representative *Enterobacteriaceae* genomes were used as informant genomes for BLAST [73].

2.2.2.1.1.9 Operons Operons have been predicted using the operon finding software (OFS) [279]. Among others OFS uses information about conserved neighborhood. For this, informant genomes to BLAST [73] against have to be supplied. In the publication the parameter β for the selection of informant genomes for the operon prediction in *Escherichia coli* K12 was chosen so that no *Gammaproteobacteria* (NCBI taxonomyid 1236) were included in the set of informant genomes. As the two *Cronobacter* spp. are also *Gammaproteobacteria* all genomes of publicly available bacteria from RefSeq

[52] (July 2009) except *Gammaproteobacteria* have been used as informant genomes. The cutoff for the operon borders was set to 0.79 as this proved to perform best when compared with validated operons from the literature.

2.2.2.1.1.10 Enrichment and depletion of annotations The enrichment and depletion of annotations was identified comparing the annotations of proteins of a subset to the annotations of the full set of all proteins annotated with this feature (see section 1.7.4). A double-sided Fisher's exact test was used to calculate the significance of the enrichment or depletion. The resulting p-value was Bonferroni corrected by multiplying the p-value with the number of tests to account for multiple testing. Annotations with a corrected p-value ≤ 0.01 were presumed to be significantly enriched or depleted.

2.2.2.1.2 Results The genome of *C. turicensis* consists of a circular chromosome with a size of 4,384,464 bp and three plasmids with sizes of 138,339 bp, 22,448 bp, 53,842 bp. Altogether 4455 coding sequences were identified of which 9.27% ($n = 413$) did not show similarities to other proteins in public sequence databases and therefore remain unknown proteins. With regard to its general features, the genome of *C. turicensis* is typical for an *Enterobacteriaceae* genome (see Table 2.7). The genome encodes 84 tRNAs with 40 different codons for 21 amino acids including selenocysteine. Seven ribosomal RNA operons could be found in the genome, which is comparable to many other *Enterobacteriaceae*. The remarkable number of 122 coding sequences has been predicted as putative pseudogenes (see Table 2.5). Almost 95% of these probably not transcribed genes retained detectable homology to annotated genes in other organisms, allowing to investigate which particular functions are putative targets of gene degradation. Significant enrichments of MIPS Functional Category [81] "01.05 C-compound and carbohydrate metabolism" (Bonferroni corrected p-Value $7.8 \cdot 10^{-04}$) and the UniProtKB/Swiss-Prot keyword [182] "Selenocysteine" (pValue corrected Bonferroni $2.8 \cdot 10^{-03}$) could be detected in these putative pseudogenes compared to all other genes in the genome. The genes similar to other proteins annotated with the keyword "Selenocysteine" are format dehydrogenases H and O and two hypothetical proteins, that are extremely well conserved in other *Enterobacteriaceae* genomes. The number of contigs, the GC content, the coding density and the number of predicted operons in the *C. turicensis* genome are typical for members of the family *Enterobacteriaceae* (Table 2.7) as well.

2.2.3 *Cronobacter turicensis* and *Cronobacter sakazakii*

2.2.3.1 Comparison of the genomes

Cronobacter turicensis LMG 23827 and its next relative *Cronobacter sakazakii* ATCC BAA-894 are the only publicly available completely sequenced genome sequences of the genus *Cronobacter*.

In order to assess their similarity on DNA level dotplots between the two *Cronobacter* species have been created. For the creation of the dotplots the software Gepard [280] with standard settings was used. Non-conserved regions have been identified by manual

	Cronobacter turicensis **LMG 23827**	***Cronobacter sakazakii*** **ATCC BAA-894**	***Escherichia coli*** **K12**
Contigs	4 (chromosome Ctu, plasmids Ctu_1p, Ctu_2p, Ctu_3p)	3 (chromosome ESA, plasmids pESA2, pESA3)	1 (chromosome)
Size [nt]	4384464 (Ctu), 138339 (Ctu_1p), 22448 (Ctu_2p), 53842 (Ctu_3p)	4368373 (ESA), 131196 (pESA3), 31208 (pESA2)	4639675 (chromosome)
GC content	57.4% (Ctu), 56.1% (Ctu_1p), 49.2% (Ctu_2p), 50.0% (Ctu_3p)	56.8% (ESA), 56.8% (pESA3), 51.6% (pESA2)	50.8% (chromosome)
rRNAs	7 operons	7 operons	7 operons
tRNAs	84	82	89
Coding genes	4455	4420	4131
Genes with homology to Uniprot	90.73%	90.07%	97.41%
Coding density	87,86%	89,20%	86,62%
Operons	3015	3167	2843

Table 2.7: Genome features of *Cronobacter* spp. and *Escherichia coli* K12 Features of the genomes of *Cronobacter turicensis* LMG 23827, *Cronobacter sakazakii* ATCC BAA-894 and *Escherichia coli* K12 are shown. The number of contigs, the GC content, the coding density and the number of predicted operons in the *C. turicensis* genome are typical for members of the family *Enterobacteriaceae*.

inspection of the DNA dotplots. Regions longer than 17kbp in the *C. turicensis* chromosome Ctu, longer than 1.7kbp in the *C. turicensis* plasmid Ctu_1p, longer than 1.4kbp in the *C. turicensis* plasmid Ctu_2p, longer than 2.4kbp in the *C.turicensis* plasmid 3 Ctu_3p, longer than 15kbp in the *C. sakazakii* chromosome ESA, longer than 800bp in the *C. sakazakii* plasmid pESA2 and longer than 1.4kbp in the *C. sakazakii* plasmid pESA3 have been examined.

The two genomes show a high degree of synteny (Figure 2.8). Only 11 non-conserved regions in *C. turicensis* and 17 non-conserved regions in *C. sakazakii* could be detected (Figure 2.8). The fraction of these regions unique to *C. turicensis* span 211 genes (4.7% of the proteome) and the regions unique to *C. sakazakii* span 287 genes (6.5% of the proteome).

It is striking that the five consecutive genes Ctu_3p00360, Ctu_3p00370, Ctu_3p00380, Ctu_3p00400, Ctu_3p00410 in region27 on plasmid 3 of *C. turicensis* are homologous to five genes of an arsenical resistance operon. These five genes can be found in the ars operon of the *Escherichia coli* conjugal plasmid R773 [281, 282, 283]. The dotplot of the two *Cronobacter* genomes shows no synteny at this position, but the proteins in this region can still be identified using protein sequence alignments. All members of the operon are available in *C. turicensis*, the two regulatory genes *ars*RD as well as the genes responsible for arsenical resistance, *ars*ABC. *C. sakazakii* on the other hand lacks *arsA*. *arsA* and *arsB* encode the subunits of an ATP-driven arsenite pump [284]. *arsA* encodes the catalytic subunit of the pump [285], while *arsB* encodes the membrane sector [286, 287]. In genomes lacking *arsA* it has been shown that the gene product of *arsB* is sufficient for providing partial arsenical resistance [288, 289]. As the conservation on DNA level is not detectable anymore and *arsA* is split in *C. turicensis* (Ctu_3p00380, Ctu_3p00390; none of their gene products has been detected in [51]) this DNA region is probably not subject to evolutionary selection anymore.

Plasmid 3 of *C. turicensis* shows partial synteny to a region of the chromosome of *C. sakazakii* (Figure 2.8, D), in which copper resistance genes are encoded. Other *C. sakazakii* specific regions contain a complete Tellurium resistance operon and may play a role in the tolerance of antimicrobials [290].

2.2.3.2 Repeats

Repeats are present in very different amounts in prokaryotic genomes and can be a hint towards the recent integration of transposons in a genome [71]. For this reason the repeat contents of the two *Cronobacter* spp. were examined.

For the repeat detection the ready to use software REPuter [80] was used. All exact maximal repeats (-f forward repeats, -p palindromes, -r reverse repeats, -c complemented repeats) with a length of at least 150 nucleotides (-l 150) were searched. REPuter had been successfully applied to the genome of *Amoebophilus asiaticus* 5a2 previously [259].

C. turicensis has a repeat content of 0.94% which is less than half of the repeat content of *C. sakazakii* with 1.90%. This difference can be mainly explained by the existence of a 42kb long tandem repeat in the genome of *C. sakazakii*, visible in the DNA dotplot (Figure 2.9). The detailed examination of the gene contents in these repetitive regions reveals that some of the two copies of the respective genes are already missing or mutated, so that this region is probably subject to fast evolution. No hints on the involvement of transposons, like inverted repeats at the ends of the duplication, could be identified that would explain the origin of these mobile elements.

2.2.3.3 Gene duplications

Gene duplications are typical events in the evolution of genomes and can lead to neofunctionalization of one of the copies and are thus an important process in the development of new functionality [65, 66]. The number of gene duplications is therefore an indicator of ongoing genome enlargement and the adaptation of a genome to a changing environment.

Gene duplications in the two *Cronobacter* spp. have been identified as follows. First the best match of each protein of an organism against the proteome of the organism besides the protein itself has been searched with an E-value cutoff of 1. This cutoff is very sensitive and ensures that no hit is missed in the first step. Then a single linkage clustering was applied. A link between two proteins has been established if their sequence similarity fulfilled the following criteria: The alignment of the two sequences had an E-value $\leq 10^{-10}$, the ratio of the alignment length to sequence length of each of the proteins was ≥ 0.5 and the sequences had ≥ 40.0 percent identity. This selective E-value cutoff ensures that the two sequences are highly probable not just similar to each other by chance, the length cutoff for the alignment is used to avoid hits between multi domain proteins sharing only a single protein domain, and the high percentage

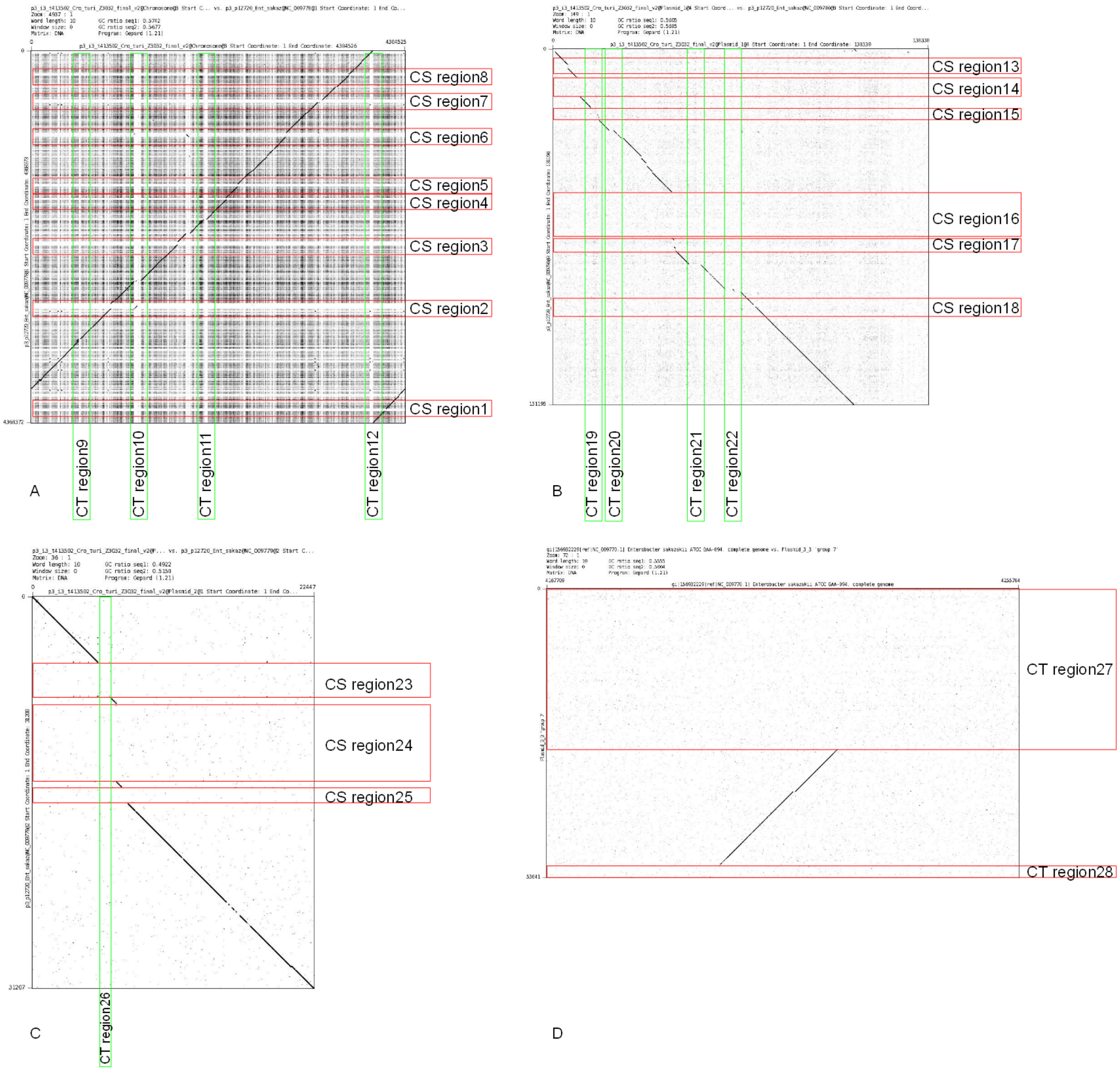


Figure 2.8: Comparisons of *Cronobacter* spp. contigs on DNA level Non-conserved regions are marked by red and green boxes. The following contigs correspond to each other: the chromosome of *C. turicensis* and the chromosome of *C. sakazakii* (Figure A), the plasmid 1 of *C. turicensis* and plasmid pESA3 of *C. sakazakii* (Figure B), plasmid 2 of *C. turicensis* and plasmid pESA2 of *C. sakazakii* (Figure C). **A:** DNA dotplot of *C. turicensis* chromosome against *C. sakazakii* chromosome **B:** DNA dotplot of *C. turicensis* plasmid 1 against *C. sakazakii* plasmid pESA3 **C:** DNA dotplot of *C. turicensis* plasmid 2 against *C. sakazakii* plasmid pESA2 **D:** DNA dotplot of *C. turicensis* plasmid 3 against *C. sakazakii* chromosome

of identity makes it reasonable that the two proteins do not just originally have similar but the same sequences.

Gene duplications could be detected in both *Cronobacter* genomes. *C. turicensis* shows slightly less gene duplications (555 proteins in 211 groups) as *C. sakazakii* (573

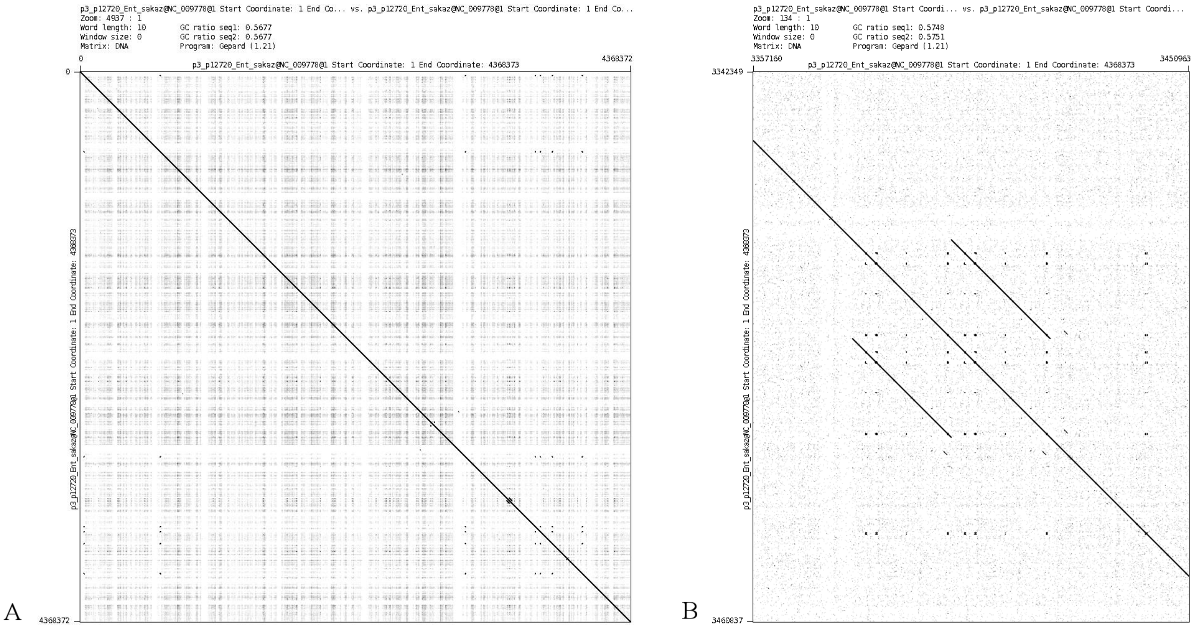


Figure 2.9: DNA dotplot of *C. sakazakii* chromosome against itself. A: The dotplot shows the comparison of the chromosome of *C. sakazakii* plotted against itself. It can be seen that there are no directly visible larger rearrangements or larger duplications. **B:** The dotplot shows the comparison of the chromosome of *C. sakazakii* plotted against itself in the region from (3342000-3451000 bp). The tandem repeat is directly visible as the two lines parallel to the diagonal.

proteins in 219 groups), which can be explained by the tandem repeat in the latter genome. The two *Cronobacter* spp. show a lower fraction of gene duplications than most of the other *Enterobacteriaceae* except for *Erwinia*, *Proteus* and endosymbionts (Table 2.8). The genomes of the two *Cronobacter* spp. can therefore be interpreted to be evolutionarily relatively stable.

2.2.3.4 Summary

Even if the DNA dotplots show a high similarity between the genomes of the two *Cronobacter* spp., there exist slight differences between the two genomes and the encoded proteins. These range from single specific genes to whole operons. The reason for these differences might be the adaption to slightly different ecological niches.

2.2.4 *Cronobacter* and *Enterobacteriaceae*

2.2.4.1 Overall similarity of the proteomes of the genus *Cronobacter* and other *Enterobacteriaceae*.

To quantify the sizes of the predicted proteomes that are shared between the two *Cronobacter* spp. and representative proteomes from other genera of the family *Enterobacteriaceae* in the RefSeq [52] database a matrix of pair-wise comparisons between

organism	number of genes in duplications	complete number of genes	percentage of duplicated genes
Candidatus Blochmannia floridanus	0	583	0.000
Candidatus Blochmannia pennsylvanicus str. BPEN	2	610	0.003
Buchnera aphidicola str. APS (Acyrthosiphon pisum)	2	564	0.004
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	4	611	0.007
Erwinia tasmaniensis Et1/99	402	3622	0.111
Proteus mirabilis HI4320	410	3662	0.112
Cronobacter turicensis LMG 23827	555	4455	0.125
Enterobacter sakazakii ATCC BAA-894	573	4420	0.130
Citrobacter koseri ATCC BAA-895	717	5008	0.143
Escherichia fergusonii ATCC 35469	612	4266	0.143
Yersinia enterocolitica subsp. enterocolitica 8081	582	4051	0.144
Yersinia pseudotuberculosis PB1/+	609	4237	0.144
Salmonella enterica subsp. enterica serovar Enteritidis str. P125109	645	4206	0.153
Enterobacter sp. 638	659	4240	0.155
Erwinia carotovora subsp. atroseptica SCRI1043	699	4472	0.156
Sodalis glossinidius str. morsitans	414	2516	0.165
Yersinia pestis Antiqua	740	4364	0.170
Salmonella typhimurium LT2	776	4525	0.171
Serratia proteamaculans 568	940	4942	0.190
Escherichia coli 55989	955	4763	0.201
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	1186	5185	0.229
Photorhabdus luminescens subsp. laumondii TTO1	1090	4683	0.233
Shigella flexneri 2a str. 2457T	971	4061	0.239
Shigella sonnei Ss046	1181	4457	0.265
Shigella boydii Sb227	1167	4282	0.273
Shigella dysenteriae Sd197	1440	4494	0.320

Table 2.8: Number and percentage of genes in duplications in representative *Enterobacteriaceae*

the proteomes has been created (Figure 2.10). About 20% of the predicted proteome of *C. turicensis* could be identified in previous work by Carranza et al [51].

The matrix was created as follows. First all bidirectional best hits (BBHs) between all pairs of proteomes have been determined. In order to be sensitive an E-value cutoff of 10^{-02} has been applied. Then for each pair of organisms the fraction of proteins of the first organism in BBH pairs with the other organism has been computed. The matrix has been colored from blue (lowest coverage) to red (highest coverage) (Figure 2.10).

As expected, the two *Cronobacter* spp. share a larger fraction (83 - 84%) of their proteomes with each other than with the proteomes of all other *Enterobacteriaceae*. Consistent with the synteny analysis discussed above, the major part of the proteome differences between the *Cronobacter* spp. can be explained by insertions or deletions in the genomes. Smaller non-conserved regions and evolutionary diverged proteins, which are not detectable by the BBH method used in the analysis, account for the remaining differences.

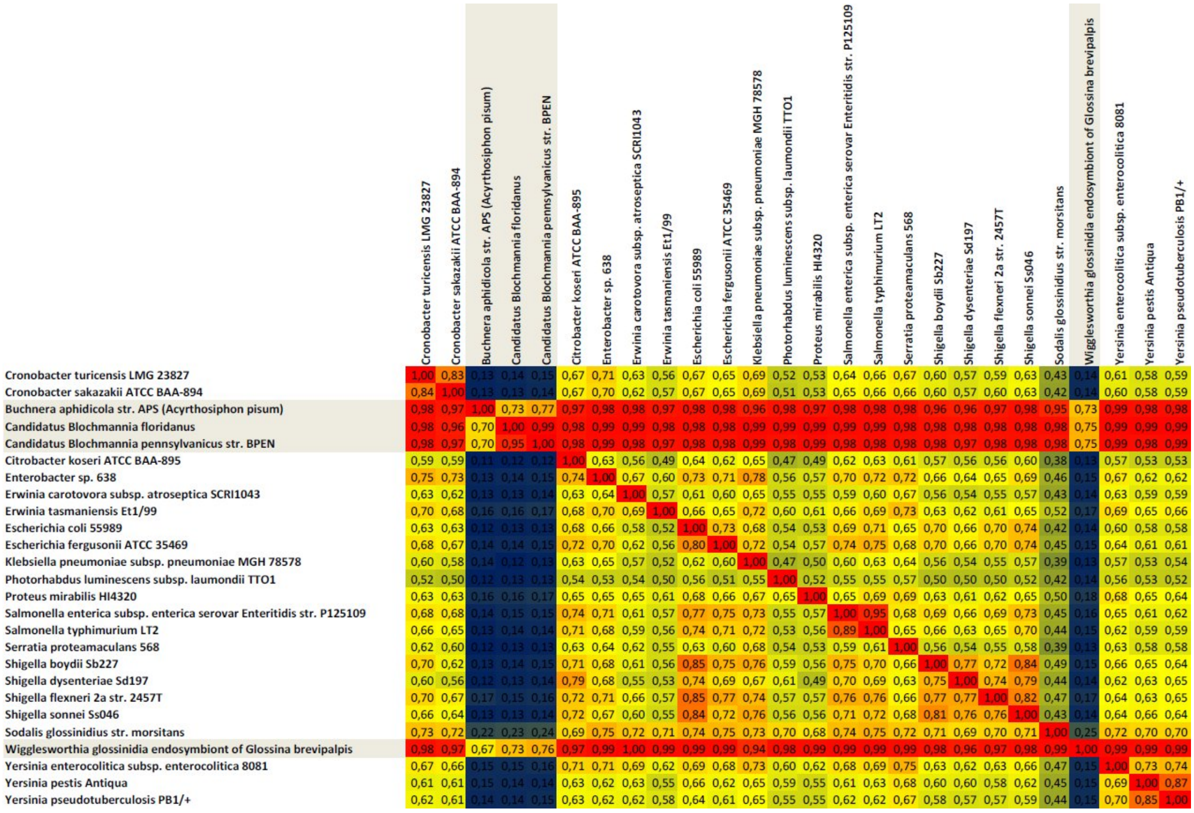
Enterobacteriaceae from different genera typically share 50% of their proteomes. The matrix contains 4 reduced genomes of intracellular symbionts having extremely reduced genomes (see shaded organisms in Figure 2.10). In contrast to them, the *Cronobacter* proteomes share similar fractions with other proteomes as the remaining, non-minimal genomes.

Clusters of orthologous groups have been built as described in section 2.1.4.2 using all *Enterobacteriaceae* as ingroups. Orthologs and inparalogs were grouped if they fulfilled an E-value cutoff of 10^{-04} and at least 50% of the lengths of both proteins of each of the comparisons were involved in the alignments. Genes were considered as specific for the two *Cronobacter* spp., if the clusters of orthologous groups contained only proteins of the two *Cronobacter* spp. and no protein from other *Enterobacteriaceae*.

In comparison to all other publicly available *Enterobacteriaceae*, 359 genes specific for the two *Cronobacter* spp. could be identified. However it should be noted that even if these 359 genes are not orthologous to other *Enterobacteriaceae*, 127 of these potentially coding genes show similarities to published proteins so that they may be genes that were subject to neofunctionalization.

A gene was defined as lost in *Cronobacter* spp. if a cluster of orthologous groups contained no protein from either of the two *Cronobacter* spp. and at least 50% of the genera of the family *Enterobacteriaceae* were covered by the cluster. There could 179 losses of genes be detected. The gene losses consist mainly of Type-III secretion related proteins, general secretion system pathway related proteins, parts of ABC transporters or hypothetical proteins. The fact that Type-III secretion related proteins as well as parts of ABC transporters are missing is in accordance with the findings that the two *Cronobacter* spp. lack a Type-III secretion system and some ABC transporters (discussed in detail in sections 2.2.5.2 and 2.2.6.1).

In order to be able to describe differences between the two *Cronobacter* proteomes, all species specific protein coding genes have been identified by searching for proteins in each of the two *Cronobacter* proteomes with no counterpart in the other *Cronobacter* proteome. There have 160 species specific genes for *C. turicensis* and 178 species specific genes for *C. sakazakii* been identified. Unfortunately, most of the genes exhibit



	Cronobacter turicensis LMG 23827	Cronobacter sakazakii ATCC BAA-894	Buchnera aphidicola str. APS (Acyrthosiphon pisum)	Candidatus Blochmannia floridanus	Candidatus Blochmannia pennsylvanicus str. BPEN	Citrobacter koseri ATCC BAA-895	Enterobacter sp. 638	Erwinia carotovora subsp. atroseptica SCRI1043	Erwinia tasmaniensis Et1/99	Escherichia coli 55989	Escherichia fergusonii ATCC 35469	Klebsiella pneumoniae subsp. pneumoniae MGH 78578	Photorhabdus luminescens subsp. laumondii TTO1	Proteus mirabilis HI4320	Salmonella enterica subsp. enterica serovar Enteritidis str. P125109	Salmonella typhimurium LT2	Serratia proteamaculans 568	Shigella boydii Sb227	Shigella dysenteriae Sd197	Shigella flexneri 2a str. 2457T	Shigella sonnei Ss046	Sodalis glossinidius str. morsitans	Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	Yersinia enterocolitica subsp. enterocolitica 8081	Yersinia pestis Antiqua	Yersinia pseudotuberculosis PB1/+
Cronobacter turicensis LMG 23827	1,00	0,83	0,13	0,14	0,15	0,67	0,71	0,63	0,56	0,67	0,65	0,69	0,52	0,53	0,64	0,66	0,67	0,60	0,57	0,59	0,63	0,43	0,14	0,61	0,58	0,59
Cronobacter sakazakii ATCC BAA-894	0,84	1,00	0,13	0,13	0,14	0,67	0,70	0,62	0,57	0,67	0,65	0,69	0,51	0,53	0,65	0,66	0,66	0,60	0,57	0,60	0,63	0,42	0,14	0,60	0,58	0,59
Buchnera aphidicola str. APS (Acyrthosiphon pisum)	0,98	0,97	1,00	0,73	0,77	0,97	0,98	0,98	0,97	0,98	0,98	0,96	0,98	0,97	0,98	0,98	0,98	0,96	0,96	0,97	0,98	0,95	0,73	0,99	0,98	0,98
Candidatus Blochmannia floridanus	0,98	0,96	0,70	1,00	0,99	0,98	0,99	0,99	0,98	0,98	0,98	0,99	0,99	0,99	0,99	0,99	0,98	0,98	0,98	0,98	0,98	0,98	0,75	0,99	0,99	0,99
Candidatus Blochmannia pennsylvanicus str. BPEN	0,98	0,97	0,70	0,95	1,00	0,98	0,99	0,98	0,97	0,98	0,98	0,99	0,98	0,99	0,98	0,98	0,98	0,98	0,97	0,98	0,98	0,98	0,75	0,99	0,98	0,99
Citrobacter koseri ATCC BAA-895	0,59	0,59	0,11	0,12	0,12	1,00	0,63	0,56	0,49	0,64	0,62	0,65	0,47	0,49	0,62	0,63	0,61	0,57	0,56	0,56	0,60	0,38	0,13	0,57	0,53	0,53
Enterobacter sp. 638	0,75	0,73	0,13	0,14	0,15	0,74	1,00	0,67	0,60	0,73	0,71	0,78	0,56	0,57	0,70	0,72	0,72	0,66	0,64	0,65	0,69	0,46	0,15	0,67	0,62	0,62
Erwinia carotovora subsp. atroseptica SCRI1043	0,63	0,62	0,13	0,13	0,14	0,63	0,64	1,00	0,57	0,61	0,60	0,65	0,55	0,55	0,59	0,60	0,67	0,56	0,54	0,55	0,57	0,43	0,14	0,63	0,59	0,59
Erwinia tasmaniensis Et1/99	0,70	0,68	0,16	0,16	0,17	0,68	0,70	0,69	1,00	0,66	0,65	0,72	0,60	0,61	0,66	0,69	0,73	0,63	0,62	0,61	0,65	0,52	0,17	0,69	0,65	0,66
Escherichia coli 55989	0,63	0,63	0,12	0,13	0,13	0,68	0,66	0,58	0,52	1,00	0,73	0,68	0,54	0,53	0,69	0,71	0,65	0,70	0,66	0,70	0,74	0,42	0,14	0,60	0,58	0,58
Escherichia fergusonii ATCC 35469	0,68	0,67	0,14	0,14	0,15	0,72	0,70	0,62	0,56	0,80	1,00	0,72	0,54	0,57	0,74	0,75	0,68	0,70	0,66	0,70	0,74	0,45	0,15	0,64	0,61	0,61
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	0,60	0,58	0,14	0,12	0,13	0,63	0,65	0,57	0,52	0,62	0,60	1,00	0,47	0,50	0,60	0,63	0,64	0,56	0,54	0,55	0,57	0,39	0,13	0,57	0,53	0,54
Photorhabdus luminescens subsp. laumondii TTO1	0,52	0,50	0,12	0,13	0,13	0,54	0,53	0,54	0,50	0,56	0,51	0,55	1,00	0,52	0,55	0,55	0,57	0,50	0,50	0,50	0,52	0,42	0,14	0,56	0,53	0,52
Proteus mirabilis HI4320	0,63	0,63	0,16	0,16	0,17	0,65	0,65	0,65	0,61	0,68	0,66	0,67	0,65	1,00	0,65	0,69	0,69	0,63	0,61	0,62	0,65	0,50	0,18	0,68	0,65	0,64
Salmonella enterica subsp. enterica serovar Enteritidis str. P125109	0,68	0,68	0,14	0,15	0,15	0,74	0,71	0,61	0,57	0,77	0,75	0,73	0,55	0,57	1,00	0,95	0,68	0,69	0,66	0,69	0,73	0,45	0,16	0,65	0,61	0,62
Salmonella typhimurium LT2	0,66	0,65	0,13	0,14	0,14	0,71	0,68	0,59	0,56	0,74	0,71	0,72	0,53	0,56	0,89	1,00	0,65	0,66	0,63	0,65	0,70	0,44	0,15	0,62	0,59	0,59
Serratia proteamaculans 568	0,62	0,60	0,12	0,13	0,13	0,63	0,64	0,62	0,55	0,63	0,60	0,68	0,54	0,53	0,59	0,61	1,00	0,56	0,54	0,55	0,58	0,39	0,13	0,63	0,58	0,58
Shigella boydii Sb227	0,70	0,62	0,13	0,14	0,15	0,71	0,68	0,61	0,56	0,85	0,75	0,76	0,59	0,56	0,75	0,70	0,66	1,00	0,77	0,72	0,84	0,49	0,15	0,66	0,65	0,64
Shigella dysenteriae Sd197	0,60	0,56	0,12	0,13	0,14	0,79	0,68	0,55	0,53	0,74	0,69	0,67	0,61	0,49	0,70	0,69	0,63	0,75	1,00	0,74	0,79	0,44	0,14	0,62	0,63	0,65
Shigella flexneri 2a str. 2457T	0,70	0,67	0,17	0,15	0,16	0,72	0,71	0,66	0,57	0,85	0,77	0,74	0,57	0,57	0,76	0,76	0,66	0,77	0,77	1,00	0,82	0,47	0,17	0,64	0,63	0,65
Shigella sonnei Ss046	0,66	0,64	0,13	0,13	0,14	0,72	0,67	0,60	0,55	0,84	0,72	0,76	0,56	0,56	0,71	0,72	0,68	0,81	0,76	0,76	1,00	0,43	0,14	0,64	0,66	0,64
Sodalis glossinidius str. morsitans	0,73	0,72	0,22	0,23	0,24	0,69	0,75	0,72	0,71	0,74	0,75	0,73	0,70	0,68	0,74	0,75	0,72	0,71	0,69	0,70	0,71	1,00	0,25	0,72	0,70	0,70
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	0,98	0,97	0,67	0,73	0,76	0,97	0,99	1,00	0,99	0,99	0,99	0,94	0,98	0,99	0,99	0,99	0,99	0,98	0,96	0,97	0,98	0,99	1,00	0,99	0,99	0,99
Yersinia enterocolitica subsp. enterocolitica 8081	0,67	0,66	0,15	0,15	0,16	0,71	0,71	0,69	0,62	0,69	0,68	0,73	0,60	0,62	0,68	0,69	0,75	0,63	0,62	0,63	0,66	0,47	0,15	1,00	0,73	0,74
Yersinia pestis Antiqua	0,61	0,61	0,15	0,14	0,14	0,63	0,62	0,63	0,55	0,66	0,62	0,65	0,59	0,55	0,61	0,63	0,68	0,60	0,60	0,58	0,62	0,45	0,15	0,69	1,00	0,87
Yersinia pseudotuberculosis PB1/+	0,62	0,61	0,14	0,14	0,15	0,63	0,62	0,62	0,58	0,64	0,61	0,65	0,55	0,55	0,62	0,62	0,67	0,58	0,57	0,57	0,59	0,44	0,15	0,70	0,85	1,00

Figure 2.10: Orthologs matrix between representative proteomes of the family *Enterobacteriaceae* The heat plot shows similarities between representative *Enterobacteriaceae* proteomes as identified by bidirectional best hits (BBHs). For each pair of organisms the fraction of proteins of the first organism (left) in BBH pairs with the other organism (top) is shown. Blue indicates the smallest, red the highest percentage of the shared proteome of the organism on the left. Reduced proteomes of endosymbionts are shaded.

homologies only to uncharacterized putative proteins (true for both organisms).

2.2.4.2 Repeat contents

While the repeat content of *C. sakazakii* (1.90%) is comparable to the repeat contents of other *Enterobacteriaceae* (Table 2.9), *C. turicensis* shows a very low repeat content of 0.94%. Only the genomes of the endosymbionts *Buchnera aphidicola* 5A (*Acyrthosiphon pisum*), *Candidatus Blochmannia pennsylvanicus* BPEN and *Candidatus Blochmannia flordidanus* contain fewer repeats than *C. turicensis*. If the tandem repeat of *C. sakazakii* is not considered, its repeat content is with 0.94% the same as for *C. turicensis*. The small repeat contents in comparison to other *Enterobacteriaceae* further support the conclusion that there occurs no massive re-organisation of the genomes at the moment and that the genomes of the two *Cronobacter* spp. are evolutionary quite stable.

organism	repeat percentage of chromosome
Buchnera aphidicola str. APS (Acyrthosiphon pisum)	0.0000
Candidatus Blochmannia pennsylvanicus str. BPEN	0.0007
Candidatus Blochmannia floridanus	0.0009
Cronobacter turicensis LMG 23827	0.0094
Serratia proteamaculans 568	0.0103
Enterobacter sp. 638	0.0106
Salmonella enterica subsp. enterica serovar Enteritidis str. P125109	0.0123
Escherichia fergusonii ATCC 35469	0.0129
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	0.0135
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	0.0138
Erwinia carotovora subsp. atroseptica SCRI1043	0.0140
Proteus mirabilis HI4320	0.0152
Erwinia tasmaniensis Et1/99	0.0153
Salmonella typhimurium LT2	0.0170
Cronobacter sakazakii ATCC BAA-894	0.0188
Citrobacter koseri ATCC BAA-895	0.0214
Yersinia pseudotuberculosis PB1/+	0.0215
Yersinia enterocolitica subsp. enterocolitica 8081	0.0232
Escherichia coli 55989	0.0311
Photorhabdus luminescens subsp. laumondii TTO1	0.0530
Yersinia pestis Antiqua	0.0595
Sodalis glossinidius str. morsitans	0.0683
Shigella flexneri 2a str. 2457T	0.0793
Shigella sonnei Ss046	0.1014
Shigella boydii Sb227	0.1028
Shigella dysenteriae Sd197	0.1266

Table 2.9: Repeat contents of members of the family *Enterobacteriaceae* The repeat content of *C. sakazakii* is comparable to the repeat contents of other *Enterobacteriaceae*. *C. turicensis* shows a very low repeat content. Only the genomes of the endosymbionts *Buchnera aphidicola* 5A (*Acyrthosiphon pisum*), *Candidatus Blochmannia pennsylvanicus* BPEN and *Candidatus Blochmannia floridanus* contain fewer repeats than *C. turicensis*. The small repeat contents of the *Cronobacter* spp. in comparison to other *Enterobacteriaceae* further support the conclusion that there occurs no massive re-organization of the genomes at the moment and that the genomes of the two *Cronobacter* spp. are evolutionary quite stable.

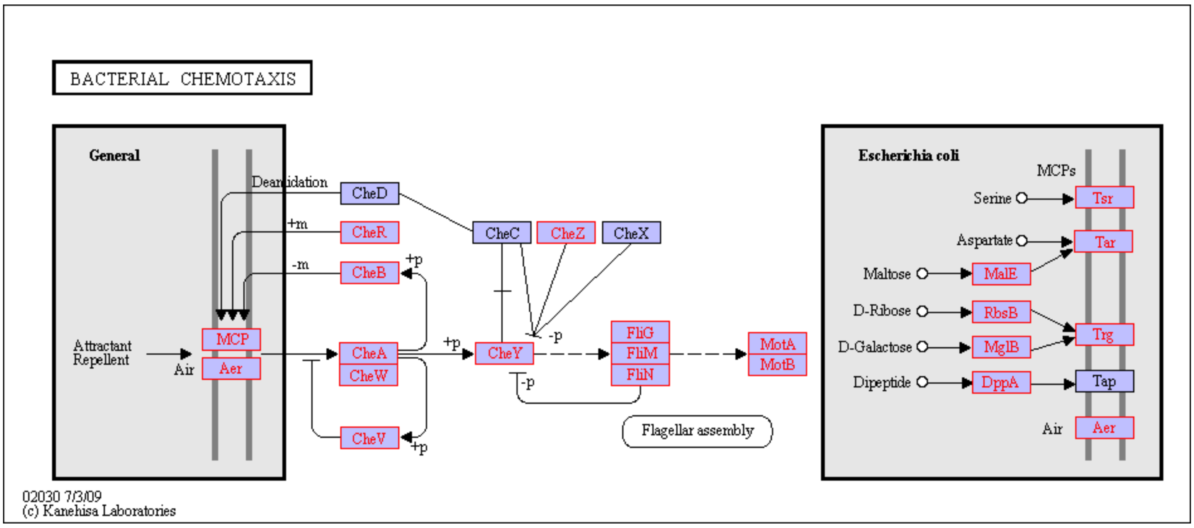


Figure 2.11: KEGG pathway of the bacterial chemotaxis of the two *Cronobacter* spp. The figure shows the colored KEGG pathway of bacterial chemotaxis of the two *Cronobacter* spp. Red bordered boxes indicate enzymes of the two *Cronobacter* spp. that could be mapped to KEGG orthologous groups. Black borders indicate enzymes of the pathway that do not exist in the two *Cronobacter* spp. A blue background indicates that the enzyme is member of a KEGG orthologous group.

2.2.4.3 Over/underrepresented protein domains in *Cronobacter* spp.

The comparison of the whole proteomes of the *Enterobacteriaceae* revealed that the two *Cronobacter* spp. slightly differ from each other and more from other *Enterobacteriaceae*. In order to further functionally characterize these differences, the enrichment and depletion of protein domains (the structural and functional building blocks of proteins [291]) have been analyzed.

All protein domains as predicted by InterProScan [292] using InterPro 23.0 have been counted in the proteomes of the two *Cronobacter* spp. in comparison to all other proteomes of the family *Enterobacteriaceae*. This information was used to perform a double-sided Fisher's exact test with Bonferroni correction for multiple testing (section 1.7.4). Domains with a corrected p-value ≤ 0.01 were presumed to be significantly enriched or depleted.

Both *Cronobacter* proteomes show a significant depletion of transposition related domains and a significant enrichment of chemotaxis related domains. The depletion of transposition related domains is in compliance with the fact that the two *Cronobacter* spp. show low repeat contents. The fact that chemotaxis related domains are identified as enriched is in compliance with the fact that all proteins of the KEGG pathway "Flagellar assembly" and almost all proteins of the KEGG pathway "Bacterial chemotaxis" are encoded in the genomes of the two *Cronobacter* spp. (Figure 2.11). This is in accordance with the observation that members of the genus *Cronobacter* are able to direct their movements towards chemicals in their environment.

2.2.4.4 Horizontal gene transfers

Horizontal gene transfers are fundamental for the rapid adaptation of prokaryotic genomes to changing environmental conditions [129]. They are quite common in pathogens and responsible e.g. for acquiring resistance against antibiotics (e.g. [130]). Therefore potentially horizontally transferred genes have been searched in *C. turicensis* with the goal to detect signs for a recent adaptation of this organism to its environment or hints on organisms living closely together with this organism.

A screen for putative horizontally transferred genes (HTGs) having their origin outside of the class *Gammaproteobacteria* was therefore performed for the complete proteome of *C. turicensis*. As Thomas Weinmaier is a specialist for horizontal gene transfers [265] I conducted the HGT analysis in collaboration with him. The methods Alien Index [132] and PhyloGenie [133] have been applied.

2.2.4.4.1 Detection using the Alien Index The Alien Index (AI) introduced by Gladyshev et al. [132] is a method to identify horizontally transferred genes (HTGs) relying on pair-wise sequence alignments. The following steps have been performed:

1. A homology search for each protein of *C. turicensis* against all sequences of cellular organisms (NCBI [277] taxonomyid 131567) in SIMAP [94] was performed. Only hits with an E-value $\leq 10^{-03}$ were kept.

2. Hits to the genus *Cronobacter* (NCBI taxonomyid 413496) were excluded from the hits in order to identify genus *Cronobacter* specific HTGs and not *C. turicensis* specific HTGs.

3. The homologs were grouped into *Gammaproteobacteria* (=ingroup) and not-*Gammaproteobacteria* (=outgroup) according to their taxonomy.

4. The AI was calculated from the E-value of the best member of the ingroup (*Gammaproteobacteria*) and the best member of the outgroup (non-*Gammaproteobacteria*). If there was no member in ingroup or outgroup, the E-value was set to 1.

   The AI was calculated as:

   $$AI = \ln((\text{best E-value ingroup}) + 10^{-200}) - \ln((\text{best E-value outgroup}) + 10^{-200}) \tag{2.1}$$

5. An AI > 0 indicates that the protein shows higher similarity to the outgroup than to the ingroup and an AI < 0 indicates that a protein shows higher similarity to the ingroup than to the outgroup. The higher the AI, the bigger is the difference between the best E-value of the outgroup and the ingroup.

   Proteins with an AI ≥ 30 were considered as HTG candidates based on experience from other projects and were used for further manual analyses.

2.2.4.4.2 Detection using PhyloGenie PhyloGenie [133] is a fully automated software for the calculation of phylogenetic trees. The reference database for PhyloGenie was generated from the nonredundant protein database NCBI nr [277], in which taxon names were edited in order to remove characters that control the structure of tree files in Newick format. The NCBI taxonomy database name file was adapted in a similar manner. The PhyloGenie software was executed for each query protein using default parameters with the following modifications: For BLAMMER -taxid f was used in order to not use GI-numbers for the analyses as those were not available for the not yet published *C. turicensis* proteins. Additionally -getdissim f was used so that the most similar sequences were used for the creation of the hidden Markov model (HMM) as otherwise the query sequence could sometimes not be detected with the HMM anymore. This can happen when the query sequence is very different from its nearest homologs. For the tree selection tool PHAT -showtrees 0 was used in order to suppress the tree output, -verbose 0 was used for less verbose output, and -checkquery true was used to ensure that the select statement related to the query and not to another protein of *C. turicensis* anywhere in the tree. NCBI BLAST version 2.2.19 was used for the similarity searches.

The trees of potential HTGs were selected in two steps. In the first step all trees containing a *C. turicensis* protein, as next neighbor a *C. sakazakii* protein and as next neighbor not a *Gammaproteobacteria* protein were searched with the statement “(((Cronobacter turicensis & Cronobacter sakazakii) & !(*Gammaproteobacteria)) & (*cellular organisms))". In the second step all trees in which a *C. turicensis* protein directly groups with a non-*Gammaproteobacteria* protein (these are the *C. turicensis* specific proteins) were searched. In order to obtain the list of HTG candidates the set union was built.

2.2.4.4.3 Manual selection of trees The AlienIndex method identified 44 HTGs, whereas PhyloGenie identified 57 HTGs. The two sets overlap by 21 HTGs.

In order to evaluate the results from the automated tools a manual analysis of all 80 HTG candidates was performed. Firstly Neighbor-Joining trees were computed:

1. a homology-search of every HTG candidate against RefSeq [52] was carried out using SIMAP [94] and an E-value cutoff of 10^{-10}.
2. a multiple alignment was calculated using MUSCLE [293] with standard parameters
3. a distance matrix for every alignment was calculated using protdist [294] with standard parameters
4. neighbor joining trees were calculated using neighbour [294] with standard parameters

The trees were then examined manually using iTOL [295] in order to validate the results from the automated tools. The potential donors of the HTGs were extracted and it was decided from case to case whether the trees were suited for a decision about a horizontal gene transfer, that is whether the other proteins in the tree were close enough to the query protein so that this conclusion could be drawn.

This resulted in a filtered set of 44 putative HTGs (Table 2.10). These proteins show statistically significant enrichments of “C-4 compound metabolism” and EC 1.1.100 “3-oxoacyl-[acyl-carrier-protein] reductase” in comparison to the whole proteome. However, an unexpected high number of 15 HTGs are localized in local clusters on the chromosome. The most striking examples are the four genes Ctu_24550-24580 (Table 2.11). Even though the genes do not represent an operon, as they are localized on opposite strands, it is possible that they have been transferred in a single horizontal gene transfer (HGT) event. The most probable donor for all 4 genes are *Burkholderia* spp. as determined by inspecting the protein trees. Overall, the most common donors for all potential HTGs are the orders *Burkholderiales* (19 HTGs) and *Rhizobiales* (9 HTGs). Members of both orders are partly plant-associated bacteria.

In a recent study it has been shown that members of the genus *Cronobacter* can be readily isolated from plant roots, that clinical as well as plant isolates are capable of developing epiphytic and endophytic colonization of tomato and maize roots, and that *Cronobacter* spp. can produce factors potentially beneficial to plant growth [296]. The fact that most of the HTGs originate from possibly plant-associated bacteria provides further evidence for plants as the original natural habitat of these two *Cronobacter* spp. The practical implication of this is that plant-related materials, such as starches, are a potential source of contamination of infant formula production facilities.

2.2.5 Pathogenicity determinants

Many of known or potential determinants for pathogenicity are shared among the *Enterobacteriaceae* and are common across a wider range of bacteria. These include flagella and motility, lipopolysaccharides (LPS) exopolysaccharides (EPS) and O-antigens, enterobacterial common antigen (ECA), fimbriae, the ability to acquire iron and resistance to antibiotics. Many of these determinants are secreted proteins and thus the various secretion types, autotransporters and two component secretion systems also constitute accessory pathogenicity determinants in the broadest sense [297].

2.2.5.1 Proteins with homologies to proteins with annotations related to virulence

This analysis has been conducted in collaboration with Angelika Lehner from the Institute for Food Safety and Hygiene, Vetsuisse Faculty, University of Zürich, as Angelika is a *Cronobacter* spp. expert.

Among all proteins of *C. turicensis*, 22 show strong homology to proteins annotated with the UniProtKB/Swiss-Prot keyword [182] “virulence” and 41 are annotated with the UniProtKB/Swiss-Prot keyword “antibiotic resistance”. However, not all known virulence factors can be detected by UniProtKB/Swiss-Prot keyword annotations. As an example, the outer membrane protein A (ompA, Ctu_15640) has no virulence related

gene	gene description	AlienIndex candidates	PhyloGenie candidates	manual inspection	order of potential donor
Ctu_00950	hypothetical protein	1	0	0	
Ctu_01050	hypothetical protein	1	0	0	
Ctu_01060	hypothetical protein	1	0	1	Sphingobacteriales
Ctu_01080	hypothetical protein	1	1	1	Burkholderiales
Ctu_01150	hypothetical protein	1	0	0	
Ctu_05110	hypothetical protein	0	1	0	
Ctu_05130	hypothetical protein	1	0	1	Burkholderiales
Ctu_05190	hypothetical protein	0	1	0	
Ctu_05310	unknown protein	0	1	0	
Ctu_05340	unknown protein	0	1	0	
Ctu_05360	unknown protein	0	1	0	
Ctu_05620	Cytosine deaminase	0	1	1	Burkholderiales
Ctu_06100	hypothetical protein	0	1	1	Desulfovibrionales
Ctu_08840	hypothetical protein	0	1	0	
Ctu_08850	hypothetical protein	1	0	1	Sphingobacteriales
Ctu_11930	hypothetical protein	0	1	0	
Ctu_11960	Uncharacterized oxidoreductase ykvO	0	1	1	Rhizobiales
Ctu_11970	hypothetical protein	1	1	1	Burkholderiales
Ctu_12240	hypothetical protein	0	1	0	
Ctu_13220	hypothetical protein	0	1	0	
Ctu_16070	Protein ydeP	0	1	0	
Ctu_17550	hypothetical protein	0	1	0	
Ctu_17720	hypothetical protein	0	1	0	
Ctu_17760	hypothetical protein	1	1	1	Rhizobiales
Ctu_18420	hypothetical protein	0	1	0	
Ctu_18800	Glycosyltransferase tibC	1	0	1	Burkholderiales
Ctu_19240	hypothetical protein	1	1	1	Rhizobiales
Ctu_19360	hypothetical protein	0	1	0	
Ctu_19380	hypothetical protein	0	1	0	
Ctu_19400	hypothetical protein	0	1	0	
Ctu_19410	unknown protein	0	1	0	
Ctu_19500	hypothetical protein	0	1	0	
Ctu_19640	hypothetical protein	1	1	?	
Ctu_19660	hypothetical protein	1	1	1	Burkholderiales
Ctu_1p00150	hypothetical protein	0	1	?	
Ctu_1p00700	hypothetical protein	1	1	1	Chlorobiales
Ctu_1p00710	hypothetical protein	1	1	1	Desulfuromonadales
Ctu_1p00720	hypothetical protein	0	1	1	Burkholderiales
Ctu_1p01180	hypothetical protein	0	1	?	
Ctu_20240	hypothetical protein	1	0	1	Rhizobiales
Ctu_20250	Uncharacterized oxidoreductase ykvO	1	1	1	Rhizobiales
Ctu_20620	hypothetical protein	0	1	0	
Ctu_20640	hypothetical protein	0	1	1	Chroococcales
Ctu_20900	hypothetical protein	0	1	0	
Ctu_20910	hypothetical protein	0	1	0	
Ctu_20940	hypothetical protein	1	0	0	
Ctu_21280	Uncharacterized HTH-type transcriptional regulator yqhC	1	1	1	Methylophilales
Ctu_21890	hypothetical protein	0	1	1	Rhodocyclales
Ctu_22580	hypothetical protein	1	0	1	Caulobacterales
Ctu_22790	hypothetical protein	1	1	1	Actinomycetales
Ctu_22800	Uncharacterized oxidoreductase yhdF	1	0	1	Planctomycetales
Ctu_22820	hypothetical protein	1	1	1	Rhizobiales
Ctu_24550	Methyl-accepting chemotaxis protein II	1	0	1	Burkholderiales
Ctu_24560	hypothetical protein	1	0	1	Burkholderiales
Ctu_24570	hypothetical protein	1	1	1	Burkholderiales
Ctu_24580	Uncharacterized HTH-type transcriptional regulator yhjC	1	0	1	Burkholderiales
Ctu_24700	hypothetical protein	0	1	0	
Ctu_26530	hypothetical protein	1	0	1	Rhizobiales
Ctu_26820	Uncharacterized oxidoreductase yvaG	0	1	1	Burkholderiales
Ctu_27320	hypothetical protein	0	1	0	
Ctu_27370	hypothetical protein	0	1	0	
Ctu_30970	hypothetical protein	1	1	1	Bacillales
Ctu_31210	hypothetical protein	1	0	0	
Ctu_32590	hypothetical protein	1	0	1	Desulfovibrionales
Ctu_32600	hypothetical protein	1	0	1	Desulfovibrionales
Ctu_32660	hypothetical protein	1	1	1	Burkholderiales
Ctu_32670	hypothetical protein	0	1	0	
Ctu_33040	hypothetical protein	1	1	1	Burkholderiales
Ctu_34710	hypothetical protein	1	1	1	Neisseriales
Ctu_35370	hypothetical protein	1	1	1	Burkholderiales
Ctu_35740	hypothetical protein	1	0	1	Burkholderiales
Ctu_36050	hypothetical protein	1	0	1	Burkholderiales
Ctu_36070	Uncharacterized HTH-type transcriptional regulator ydgC	1	1	1	Bacillales
Ctu_36150	hypothetical protein	0	1	1	Burkholderiales
Ctu_36340	Methyl-accepting chemotaxis protein II	1	0	1	Burkholderiales
Ctu_37100	unknown protein	0	1	0	
Ctu_37820	hypothetical protein	1	0	1	Rhizobiales
Ctu_40080	hypothetical protein	1	0	0	
Ctu_40870	hypothetical protein	1	1	1	Rhizobiales
Ctu_40950	hypothetical protein	1	1	0	

Table 2.10: Potentially horizontally transferred genes of *C. turicensis* from outside the *Gammaproteobacteria* For each gene and each method it is shown whether the gene is a potential horizontally transferred gene (HTG) according to the method (1) or not (0). The "?" indicates that a decision could not clearly be made. The taxonomic order of the potential donor is given in the cases of positively manually identified HTG candidates.

gene	description	best UniProtKB/Swiss-Prot hit	best hit description
Ctu_24550	Methyl-accepting chemotaxis protein II	MCP2_SALTY	RecName: Full=Methyl-accepting chemotaxis protein II; Short=MCP-II; AltName: Full=Aspartate chemo.. PriAC=P02941 [Salmonella typhimurium] Name=tar; Synonyms=cheM; OrderedLocusNames=STM1919;
Ctu_24560	hypothetical protein	A7MKD7_ENTS8 *	SubName: Full=Putative uncharacterized protein PriAC=A7MKD7 [Enterobacter sakazakii (strain ATCC BAA-894)] OrderedLocusNames=ESA_01472;
Ctu_24570	hypothetical protein	YAJO_ECOLI	RecName: Full=Uncharacterized oxidoreductase yajO; EC=1.-.-.- PriAC=P77735 SecAC=Q2MC07 [Escherichia coli (strain K12)] Name=yajO; OrderedLocusNames=b0419, JW0409;
Ctu_24580	Uncharacterized HTH-type transcriptional regulator yhjC	YHJC_ECOLI	RecName: Full=Uncharacterized HTH-type transcriptional regulator yhjC PriAC=P37641 SecAC=Q2M7I3 [Escherichia coli (strain K12)] Name=yhjC; OrderedLocusNames=b3521, JW3489;

Table 2.11: Example of four consecutive genes probably horizontally transferred from *Burkholderia* Within the candidates of horizontally transferred genes the biggest cluster of consecutive genes is composed of the four genes Ctu_24550-24580. Even though the genes do not represent an operon, as they are localized on opposite strands, it is possible that they have been transferred in a single horizontal gene transfer event. The most probable donor for all 4 genes are *Burkholderia* spp. (* No UniProtKB/Swiss-Prot hit could be detected so this is the best UniProtKB/TrEMBL hit)

keyword in UniProtKB/Swiss-Prot. Nevertheless it has been shown that *E. sakazakii* (*Cronobacter*) expresses outer membrane protein A (OmpA) [51] and that it is crucial for the invasion of brain endothelial cells [303].

The following proteins showing homology to virulence related genes have been identified in *C. turicensis*: MviN, the two-component system PhoP-PhoQ, Hfq, IgaA, VacJ, Wza and Wzb, wecA.

Moreover a number of penicillin binding proteins were detected: penicillin and cefalosporine binding protein 1B (Ctu_07800), the penicillin binding protein 2 (MrdA, Ctu_12680), and the AmpC (Ctu_21110), a class C beta-lactamase which hydrolyses broad and extended-spectrum cephalosporins [298, 299, 300].

Interestingly two loci, one chromosomal (tehB, Ctu_22340) and one plasmid borne (tehA, Ctu_1p00320) were identified, both putatively involved in conferring resistance to the oxidative reagent tellurite. In *E. coli* tehAB constitute an operon and were originally thought to be plasmid encoded [301, 302].

2.2.5.2 Secretion systems

In Bacteria, 7 different secretion systems have been described so far (see also section 1.6.1). They facilitate the export of DNA and/or proteins from the inside of the bacterial cell into the host cell and therefore play an important role in the virulence of bacterial pathogens [151, 152, 153, 154, 155, 156, 157, 158]. Proteins transported by secretion systems are also called effectors and are likely to interact with molecules of the host cell. Thus effector proteins are of particular interest for the analysis of the pathogenic *Cronobacter* spp.

As Roland Arnold is experienced with bacterial secretion systems due to his PhD (e.g. [177]) the analysis of secretion systems was conducted in collaboration with him.

In order to detect transport related genes, homologs of proteins of both *Cronobacter* spp. to transport proteins in datasets of known transporters were searched using BLAST with an E-value cutoff of 10^{-10} and 50% of both sequences needed to be involved in the alignment to avoid single domain hits. For the ABC transporters and the Type II-VI secretion systems, the according KEGG [189] modules as downloaded at 20th of October 2009 have been used. Bi-directional best hits (BBHs) have been determined against the full set of orthologous groups in KEGG and the best matching KEGG ortholog has been chosen. The coloring of the KEGG maps was done identifying the best hits in the protein sequences in the KEGG database using BLAST with an E-value cutoff of 10^{-10} and at least 50% of the length of the query and hit protein involved in the alignment. The KEGG orthologous group of the best hit was then transferred and colored in the KEGG maps.

The genome of *C. sakazakii* encodes 779, the genome of *C. turicensis* 810 proteins associated to transport and secretion and both species comprise a similar amount of the different transport systems. The high number of involved proteins can be explained by the sensitive approach that was used to detect them.

The Type II secretion pathway is completely absent in both genomes, only a *gspO* homolog could be detected in *C. turicensis.*

The Type III secretion system that plays an important role in the pathogenicity of other members of the family *Enterobacteriaceae* (e.g. *Yersinia* or *Shigella*) is not encoded in the two *Cronobacter* genomes (Table 2.12).

The Type IV secretion system mediates the secretion of single proteins, protein-protein complexes and protein-DNA complexes. It consists of 12 parts that build a needle that can inject these molecules into the host cell. The genes encoding the proteins of this system are completely apparent in both *Cronobacter* genomes, only *VirB7* and *VirB3* are missing. However, the role of *VirB3* in Type IV mediated transport is unclear [304]. *VirB7* interacts with and stabilizes *VirB9* at the outer membrane component of the pore and its absence could weaken the ability to build up the Type IV secretion system core complex which spans both membranes. This secretion system is located on plasmid 2 of *C. turicensis*. In conclusion both *Cronobacter* spp. carry the potential to exchange genetic material and to translocate proteins employing their Type IV secretion system.

The genes encoding proteins for the SecA dependent pathway for protein export are completely existent in both species missing the SecDF fusion protein whereby harboring the *secD* and *secF* as separate genes. Therefore all members of this pathway are encoded in the two *Cronobacter* genomes and it is by that potentially fully functional.

The recently described Type VI transport system is also encoded in the genomes of both species. There is only little known about architecture and function of this secretion system. A homolog of *vrG* is missing as well as homologs of the *stpA1* (serine/threonine Phosphatase) and *ppkA* (serine/threonine Kinase). Therefore the Type VI secretion system is partially complete with five existing (*Hcp*, *Lip*, *IcmF*, *DotU*, *ClpV*) and four missing components (*VgRG*, *PpkA*, *Fha1*, *PppA*) (Figures 2.12, 2.13).

The Twin-Arginin System which is able to transport proteins in their fully folded conformation [305] is completely encoded in both *Cronobacter* spp.

2.2.5.3 Prediction of putative secreted proteins

There are three secretion systems able to transport DNA or proteins into the host cell: Type-III, Type-IV and Type-VI (reviewed in [151]). As both *Cronobacter* genomes encode Type IV and Type VI secretion systems that are probably functional, effector proteins translocated by these machineries are likely to be encoded in the genome. In contrast to the secretion systems these effectors are known to be species specific or poorly conserved even in closely related organisms [306], hindering their identification by homology searches. Therefore the detection of secreted effectors is a non-trivial and important task for the characterization of the virulence of the two *Cronobacter* spp.

We screened the complete *Cronobacter* proteomes for InterPro protein domains known to be typical for secreted effectors [307]. *C. turicensis* and *C. sakazakii* both encode 31 potentially secreted effector proteins. This number is rather typical for pathogenic bacteria than for host-associated non-pathogenic bacteria (Table 2.13).

2.2.5.4 Proteins having eukaryotic like protein domains.

Proteins with eukaryotic-like protein domains acquired by bacterial pathogens could be able to mimic and alter functions in the host cell [178] and therefore play an important role for the virulence of *Cronobacter*.

As Marc-André Jehl was currently working on the investigation of these domains the following analysis was conducted in collaboration with him.

organism	bacterial secretion systems						
	Type I (3 proteins)	Type III (15 proteins)	Type II (13 proteins)	Type IV (12 proteins)	VI (9 proteins)	Sec-SRP (10 proteins)	Twin arginine targeting (Tat) (4 proteins)
Buchnera aphidicola str. APS (Acyrthosiphon pisum)	0	0	0	0	0	9	0
Candidatus Blochmannia floridanus	0	0	0	0	0	8	0
Candidatus Blochmannia pennsylvanicus str. BPEN	0	0	0	0	0	10	0
Citrobacter koseri ATCC BAA-895	1	0	12	10	2	10	4
Cronobacter sakazakii ATCC BAA-894	1	0	1	10	6	10	4
Cronobacter turicensis	1	0	1	10	5	10	4
Enterobacter sp. 638	1	0	0	0	3	9	4
Erwinia carotovora subsp. atroseptica SCRI1043	1	11	13	10	6	10	4
Erwinia tasmaniensis	1	11	12	9	6	10	4
Escherichia coli 55989	1	3	12	0	6	10	4
Escherichia fergusonii ATCC 35469	1	0	12	0	6	10	4
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	1	0	12	0	5	10	4
Photorhabdus luminescens subsp. laumondii TTO1	3	15	1	0	6	10	3
Proteus mirabilis HI4320	3	11	1	10	6	10	3
Salmonella enterica subsp. enterica serovar Enteritidis str. P125109	1	11	1	0	1	10	4
Salmonella typhimurium LT2	1	11	1	0	5	10	4
Serratia proteamaculans 568	1	0	13	0	6	10	4
Shigella boydii Sb227	1	0	9	0	0	10	4
Shigella dysenteriae Sd197	1	11	11	0	0	10	4
Shigella flexneri 2a str. 2457T	1	0	0	1	0	10	4
Shigella sonnei Ss046	1	11	0	0	6	10	4
Sodalis glossinidius str. morsitans	1	11	0	1	1	10	3
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	0	0	0	0	0	8	0
Yersinia enterocolitica subsp. enterocolitica 8081	3	15	13	1	5	10	4
Yersinia pestis Antiqua	3	15	12	0	6	10	4
Yersinia pseudotuberculosis PB1/+	3	15	12	0	6	10	4

Table 2.12: Overview over bacterial secretion systems in *Enterobacteriaceae* The table shows how many of the parts of each secretion system as annotated in the KEGG maps for "bacterial secretion" are existing in the respective organism for each of the representative *Enterobacteriaceae*.

In order to detect eukaryotic like protein domains that could be the reason for the pathogenicity of *C. turicensis*, protein domains occurring in protein sequences of *C. turicensis* were searched that are more frequent in pathogenic than in non-pathogenic bacterial organisms.

For this fully sequenced pathogenic and non-pathogenic organisms from RefSeq [52] were identified by their phenotypes listed at NCBI (August 2009) [277]. This resulted in 388 pathogenic and 306 non-pathogenic bacterial organisms. Furthermore, 148 available completely sequenced eukaryotic genomes from RefSeq were used. For each of the genomes, all annotated proteins were extracted and analyzed for domain signatures from the PFAM [118] protein family database. Domains were only considered if they occurred in at least 10 different eukaryotic genomes. This prevents misannotations caused by low-quality eukaryotic genome sequences with bacterial contaminations. The frequency of these domains in pathogenic bacteria was compared to the frequency of the occurrence in non-pathogenic bacteria (see [179]). The most overrepresented domains were used for

organism	PF00023	PF00086	PF00149	PF00515	PF00560	PF00614	PF00646	PF01145	PF01222	PF01734	PF09335	PS50011	sum of domains	classification *
Escherichia coli 55989	1	0	9	5	1	5	0	5	0	3	7	2	38	p
Escherichia fergusonii ATCC 35469	2	0	12	3	0	4	0	6	0	2	7	2	38	p
Shigella sonnei Ss046	1	0	9	3	5	5	0	4	0	2	7	1	37	p
Shigella boydii Sb227	0	0	10	4	5	3	0	4	0	2	7	1	36	p
Shigella dysenteriae Sd197	1	0	7	1	5	4	0	4	0	2	7	1	32	p
Cronobacter turicensis	2	0	9	4	0	5	0	4	0	1	5	1	31	p
Enterobacter sakazakii ATCC BAA-894	2	0	9	4	0	4	0	4	0	1	5	2	31	p
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	1	0	10	5	0	3	0	3	0	2	7	0	31	p
Salmonella typhimurium LT2	0	0	11	2	1	5	0	4	0	2	5	0	30	p
Citrobacter koseri ATCC BAA-895	0	0	8	5	0	5	0	3	0	2	6	0	29	p
Enterobacter sp. 638	1	0	8	5	0	4	0	3	0	2	6	0	29	h
Erwinia carotovora subsp. atroseptica SCRI1043	2	0	12	4	0	3	0	3	0	1	4	0	29	p
Salmonella enterica subsp. enterica serovar Enteritidis str. P125109	0	0	11	2	1	4	0	4	0	2	5	0	29	p
Serratia proteamaculans 568	2	0	9	4	0	3	0	3	0	2	5	1	29	p
Shigella flexneri 2a str. 2457T	0	0	7	2	1	5	0	5	0	2	7	0	29	p
Yersinia pestis Antiqua	1	0	8	2	2	3	0	5	0	2	4	2	29	p
Yersinia enterocolitica subsp. enterocolitica 8081	1	0	9	3	1	2	0	3	0	2	4	1	26	p
Yersinia pseudotuberculosis PB1/+	1	0	6	2	3	2	0	4	0	2	4	2	26	p
Erwinia tasmaniensis Et1/99	2	0	6	3	0	3	0	3	0	1	5	1	24	h
Proteus mirabilis HI4320	0	0	8	3	0	2	0	5	0	1	4	0	23	p
Photorhabdus luminescens subsp. laumondii TTO1	2	0	7	1	0	3	0	3	0	1	4	0	21	p
Sodalis glossinidius str. morsitans	1	0	5	1	0	2	0	3	0	0	3	1	16	h
Candidatus Blochmannia floridanus	0	0	2	0	0	2	0	2	0	0	1	0	7	h
Candidatus Blochmannia pennsylvanicus str. BPEN	0	0	2	0	0	2	0	2	0	0	1	0	7	h
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	0	0	2	0	0	2	0	2	0	0	1	0	7	h
Buchnera aphidicola str. APS (Acyrthosiphon pisum)	0	0	1	0	0	1	0	2	0	0	1	0	5	h

Table 2.13: Number of potential effector proteins in *Enterobacteriaceae* The table shows how many proteins contain protein domains known to be typical for secreted effectors for each of the representative *Enterobacteriaceae*. The number of 31 potentially secreted effector proteins of the two *Cronobacter* spp. is rather typical for pathogenic bacteria than for host-associated non-pathogenic bacteria. (* p = pathogenic, h = host-associated)

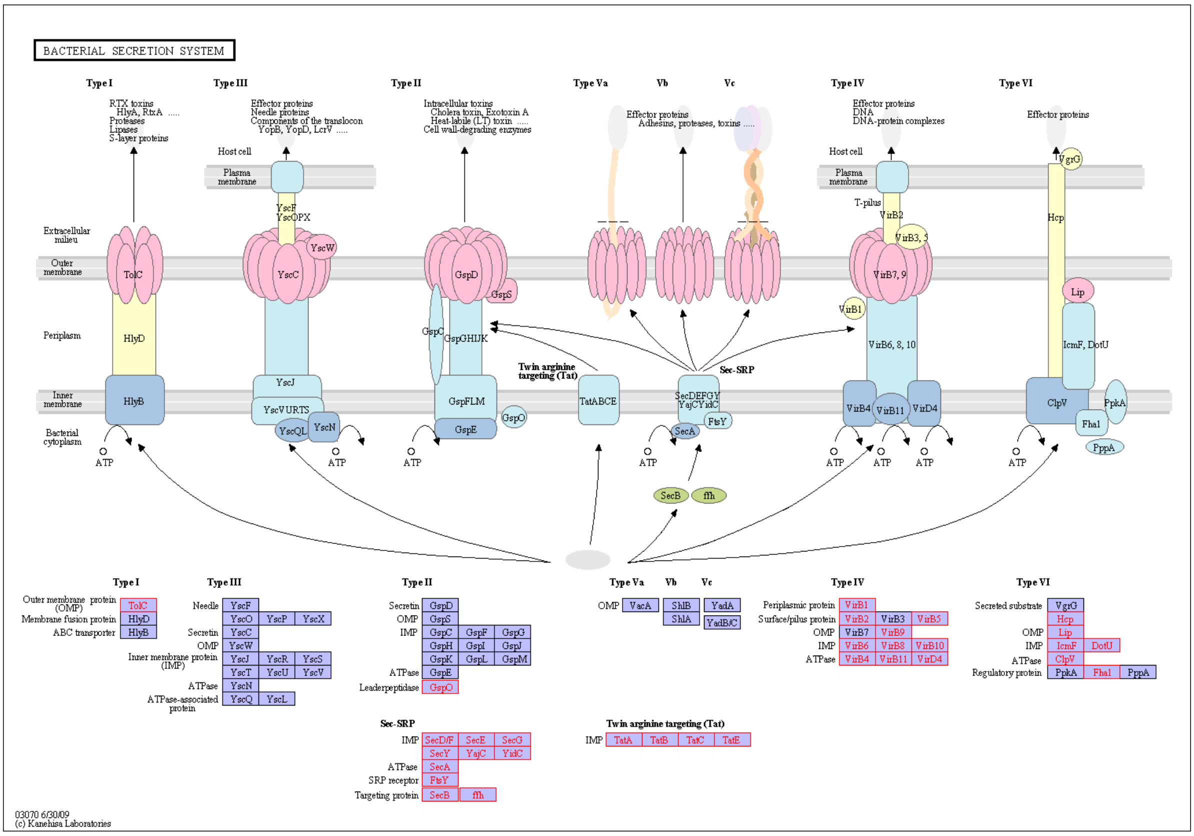


Figure 2.12: KEGG pathway of the bacterial secretion system of *Cronobacter sakazakii* ATCC BAA-894 The figure shows the colored KEGG pathway of the bacterial secretion system of *Cronobacter sakazakii* ATCC BAA-894. Red bordered boxes indicate enzymes of *Cronobacter sakazakii* ATCC BAA-894 that could be mapped to KEGG orthologous groups. Black borders indicate enzymes of the pathway that do not exist in *Cronobacter sakazakii* ATCC BAA-894. A blue background indicates that the enzyme is member of a KEGG orthologous group.

further investigation.

There could 25 eukaryotic domains be identified in the two *Cronobacter* spp. The domains are rather typical for pathogenic than for nonpathogenic bacteria, whereas 3 domains are contained in *C. turicensis* but not in *C. sakazakii* (PF01234, PF04488, PF10022). The domain PF01234 describes a family of methyltransferases that is only present in *C. turicensis* (Ctu_31510) and in plant-pathogenic/symbiontic organisms (*Frankia alni*, *Agrobacterium vitis*, *Stackebrandtia nassauensis*, *Manganese-oxidizing bacterium*) and in 44 taxonomically distributed eukaryotes. This is a further hint on the plant association of *C. turicensis*. No gene carrying this domain could be identified in *C. sakazakii*. The trehalase domain PF01204 can be found in both *Cronobacter* spp., which is in accordance with the fact, that the accumulation of trehalose inside the *E. sakazakii* cells plays a role in desiccation resistance [226]. 2 genes are coding for a trehalase in each of the two *Cronobacter* spp. (*C. turicensis*: Ctu_21560, Ctu_24460, *C. sakazakii*:

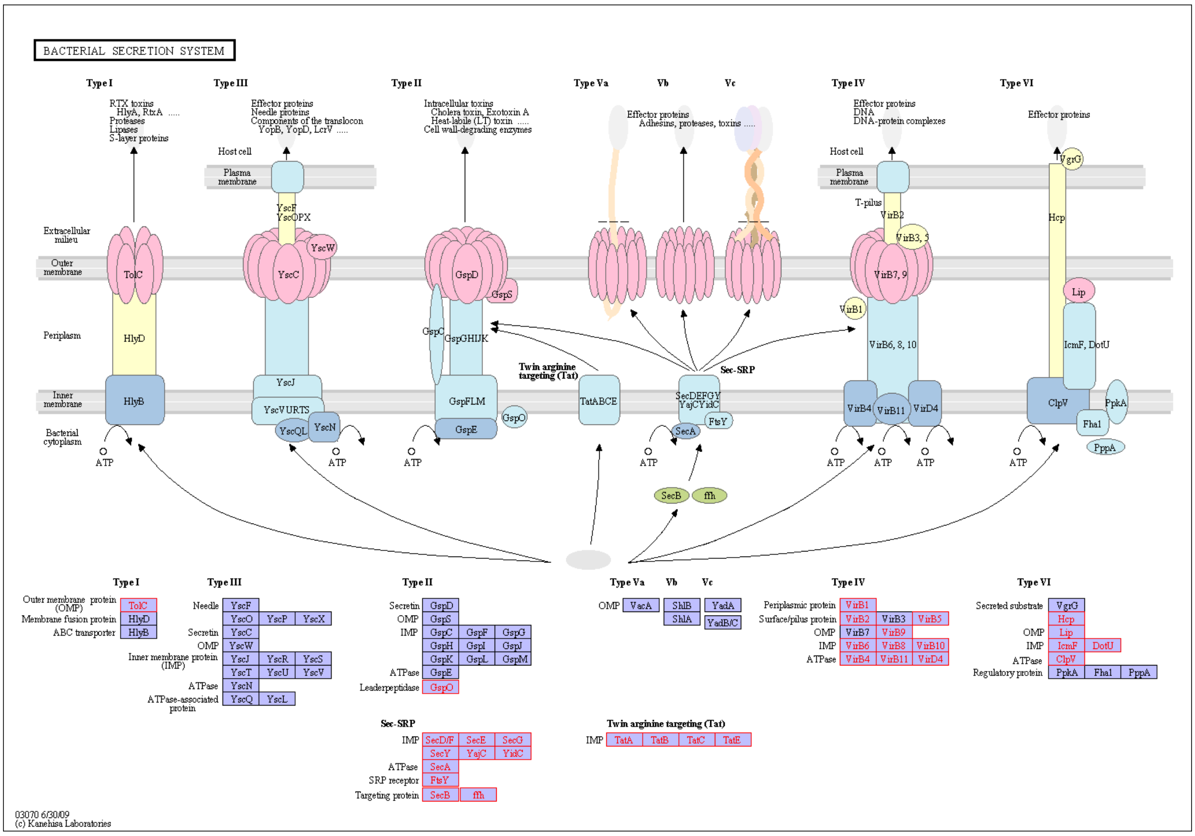


Figure 2.13: KEGG pathway of the bacterial secretion system of *Cronobacter turicensis* LMG 23827 The figure shows the colored KEGG pathway of the bacterial secretion system of *Cronobacter turicensis* LMG 23827. Red bordered boxes indicate enzymes of *Cronobacter turicensis* LMG 23827 that could be mapped to KEGG orthologous groups. Black borders indicate enzymes of the pathway that do not exist in *Cronobacter turicensis* LMG 23827. A blue background indicates that the enzyme is member of a KEGG orthologous group.

GI:156933657, GI:156934000) whereas a signal peptide can be identified in Ctu_24460 and GI:156933657 using SignalP [275].

2.2.6 Survival and persistence in diverse environments

In order to elucidate the survival and persistence of *Cronobacter* spp. transporters and metabolic pathways typical for plant-associated bacteria have been investigated.

A transporter is a membrane protein involved in the transportation of ions, small molecules, or macromolecules as proteins across a biological membrane. ABC transporters play important roles in multidrug resistance of pathogenic bacteria [308].

The transporters have been analyzed in collaboration with Roland Arnold from the Department of Genome Oriented Bioinformatics, Technische Universität München, and in collaboration with Angelika Lehner from the Institute for Food Safety and Hygiene, Vetsuisse Faculty, University of Zurich. The metabolic pathways have been analyzed in collaboration with Gabi Kastenmüller from the Institute of Bioinformatics and Systems Biology at the Helmholtz Zentrum München.

2.2.6.1 ABC transporter systems

The genomes of both *Cronobacter* spp. encode 68 bacterial ABC transporter systems (counted as amount of ABC transport specific ATPases), a typical amount for bacterial genomes of this size [309]. The colored KEGG maps of the ABC transporters for the *Cronobacter* spp. can be seen in Figures 2.14 and 2.15. Only slight differences in the completeness of the systems could be detected, which might either be explained by gene-loss or gene-gain in one species or by miss-assignments by the bidirectional best-blast hit criterion against the KEGG [189] data-set.

Most of the systems are encoded as single copy instances. Multi-copy ABC transporters comprise the osmoprotectant transport system (two copies of the complete system and four copies of the permease component), the maltose/maltodextrin transporter (at least two copies due to the substrate binding component), the lipoprotein-releasing system (two copies), the sulfonate/nitrate/taurin transport system (four copies), the antibiotic transport system (three permeases, one ATPase), the iron complex transport system (three copies), the peptide/nickel transporter (six copies), the D-methionine transport system (three copies), and the branched-chain amino acid transporter (two copies).

The ABC transporter protein YojL (Ctu_28680) represents an efflux pump for the antimicrobial peptide microcin [310] and the undecaprenyl-diphosphatase (EC 3.6.1.27), also referred to as bacitracin resistance protein (Ctu_35260) conferring resistance to this peptidoglycan synthesis inhibitor [311]. Possible modifications of the phosphoethanolamine transferase EptA (Ctu_19850) may result in resistance to polymyxin (inferred by similarity) and the bicyclomycin resistance proteins Bcr (Ctu_28490) as well as DHPS (Ctu_04040) may confer resistance to folate biosynthesis inhibitors (sulfonamides) [312, 313].

2.2.6.2 Other transport systems

The analysis revealed 18 genes unique for *C. turicensis* and 7 for *C. sakazakii* related to other transport systems. Some members of the fimbrial usher family [314] are conserved differently in the two *Cronobacter* spp.: The Type π fimbrial usher PapC is only detectable in *C. turicensis* (Ctu_36430), whereas the Type β fimbrial usher YhcD can only be found in *C. sakazakii* (GI:156935650).

Interestingly, *C. turicensis* harbors three transport systems for galactose derivates which are missing in *C. sakazakii*: a putative N-acetyl galactosamine (GalNAc or Aga) porter Ctu_07680, a putative galactitol porter Ctu_04380, and a putative N-acetyl galactosamine (GalNAc or Aga) porter Ctu_07680 and a putative two component malonate transport operon (Ctu_35020 and Ctu_35030). *C. sakazakii* uniquely provides one Alanin transporter GI:156932523.

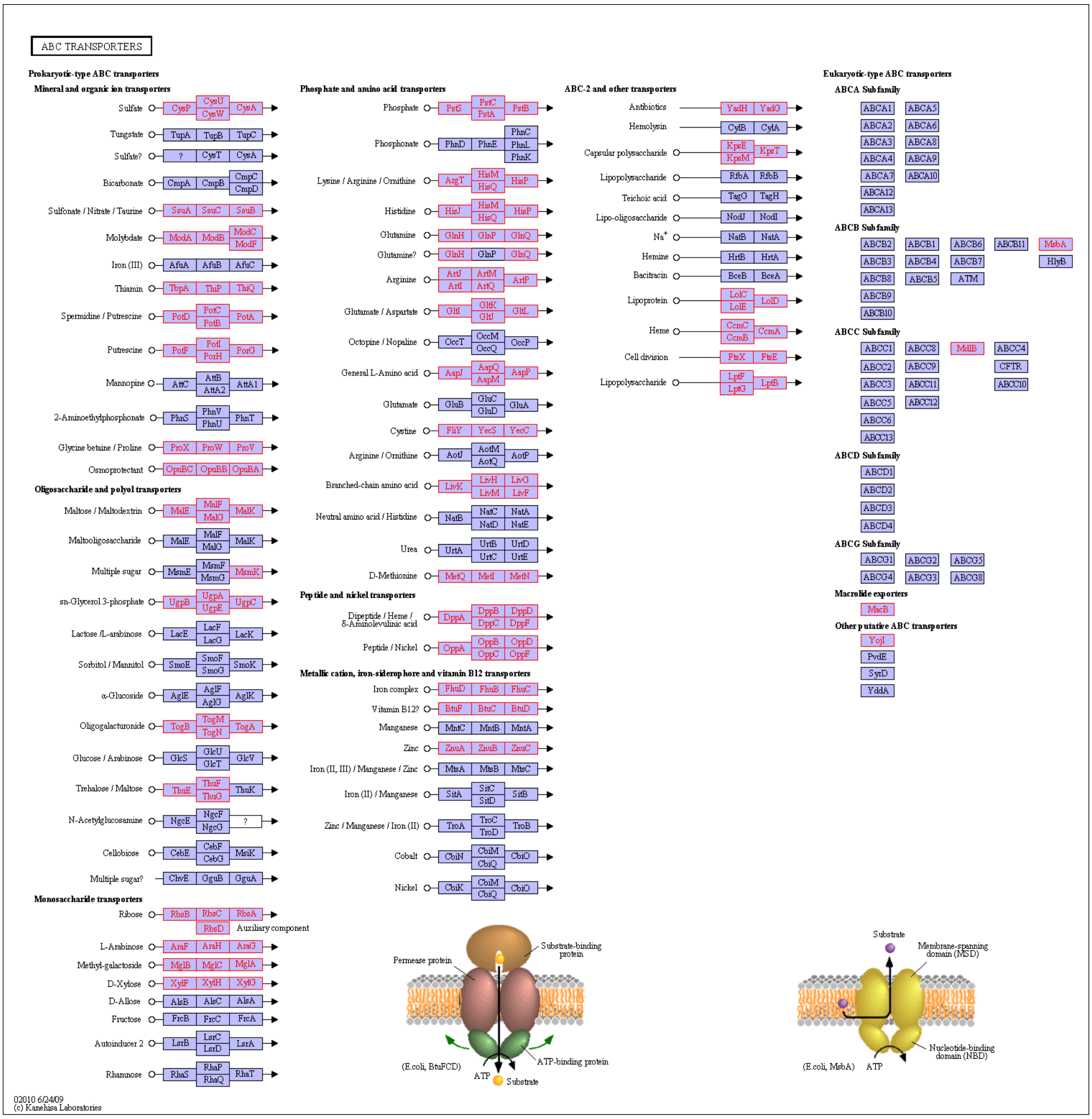

ABC TRANSPORTERS
Prokaryotic-type ABC transporters
Mineral and organic ion transporters
Sulfate CysP CysU CysW CysA
Tungstate TupA TupB TupC
Sulfate? ? CysT CysA
Bicarbonate CmpA CmpB CmpC CmpD
Sulfonate / Nitrate / Taurine SsuA SsuC SsuB
Molybdate ModA ModB ModC ModF
Iron (III) AfuA AfuB AfuC
Thiamin TbpA ThiP ThiQ
Spermidine / Putrescine PotD PotC PotB PotA
Putrescine PotF PotI PorH PorG
Mannopine AttC AttB AttA2 AttA1
2-Aminoethylphosphonate PhnS PhnV PhnU PhnT
Glycine betaine / Proline ProX ProW ProV
Osmoprotectant OpuBC OpuBB OpuBA
Oligosaccharide and polyol transporters
Maltose / Maltodextrin MalE MalF MalG MalK
Maltooligosaccharide MalE MalF MalG MalK
Multiple sugar MsmE MsmF MsmG MsmK
sn-Glycerol 3-phosphate UgpB UgpA UgpE UgpC
Lactose /L-arabinose LacE LacF LacG LacK
Sorbitol / Mannitol SmoE SmoF SmoG SmoK
α-Glucoside AglE AglF AglG AglK
Oligogalacturonide TogB TogM TogN TogA
Glucose / Arabinose GlcS GlcU GlcT GlcV
Trehalose / Maltose ThuE ThuF ThuG ThuK
N-Acetylglucosamine NgcE NgcF NgcG ?
Cellobiose CebE CebF CebG MsiK
Multiple sugar? ChvE GguB GguA
Monosaccharide transporters
Ribose RbsB RbsC RbsA
RbsD Auxiliary component
L-Arabinose AraF AraH AraG
Methyl-galactoside MglB MglC MglA
D-Xylose XylF XylH XylG
D-Allose AlsB AlsC AlsA
Fructose FrcB FrcC FrcA
Autoinducer 2 LsrB LsrC LsrD LsrA
Rhamnose RhaS RhaP RhaQ RhaT
Phosphate and amino acid transporters
Phosphate PstS PstC PstA PstB
Phosphonate PhnD PhnE PhnC PhnL PhnK
Lysine / Arginine / Ornithine ArgT HisM HisQ HisP
Histidine HisJ HisM HisQ HisP
Glutamine GlnH GlnP GlnQ
Glutamine? GlnH GlnP GlnQ
Arginine ArtJ ArtI ArtM ArtQ ArtP
Glutamate / Aspartate GltI GltK GltJ GltL
Octopine / Nopaline OccT OccM OccQ OccP
General L-Amino acid AapJ AapQ AapM AapP
Glutamate GluB GluC GluD GluA
Cystine FliY YecS YecC
Arginine / Ornithine AotJ AotM AotQ AotP
Branched-chain amino acid LivK LivH LivM LivG LivF
Neutral amino acid / Histidine NatB NatC NatD NatA NatE
Urea UrtA UrtB UrtC UrtD UrtE
D-Methionine MetQ MetI MetN
Peptide and nickel transporters
Dipeptide / Heme / δ-Aminolevulinic acid DppA DppB DppC DppD DppF
Peptide / Nickel OppA OppB OppC OppD OppF
Metallic cation, iron-siderophore and vitamin B12 transporters
Iron complex FhuD FhuB FhuC
Vitamin B12? BtuF BtuC BtuD
Manganese MntC MntB MntA
Zinc ZnuA ZnuB ZnuC
Iron (II, III) / Manganese / Zinc MtsA MtsB MtsC
Iron (II) / Manganese SitA SitC SitD SitB
Zinc / Manganese / Iron (II) TroA TroC TroD TroB
Cobalt CbiN CbiM CbiQ CbiO
Nickel CbiK CbiM CbiQ CbiO
ABC-2 and other transporters
Antibiotics YadH YadG
Hemolysin CylB CylA
Capsular polysaccharide KpsE KpsM KpsT
Lipopolysaccharide RfbA RfbB
Teichoic acid TagG TagH
Lipo-oligosaccharide NodJ NodI
Na+ NatB NatA
Hemine HrtB HrtA
Bacitracin BceB BceA
Lipoprotein LolC LolE LolD
Heme CcmC CcmB CcmA
Cell division FtsX FtsE
Lipopolysaccharide LptF LptG LptB
Eukaryotic-type ABC transporters
ABCA Subfamily
ABCA1 ABCA5
ABCA2 ABCA6
ABCA3 ABCA8
ABCA4 ABCA9
ABCA7 ABCA10
ABCA12
ABCA13
ABCB Subfamily
ABCB2 ABCB1 ABCB6 ABCB11 MsbA
ABCB3 ABCB4 ABCB7 HlyB
ABCB8 ABCB5 ATM
ABCB9
ABCB10
ABCC Subfamily
ABCC1 ABCC8 MdlB ABCC4
ABCC2 ABCC9 CFTR
ABCC3 ABCC11 ABCC10
ABCC5 ABCC12
ABCC6
ABCC13
ABCD Subfamily
ABCD1
ABCD2
ABCD3
ABCD4
ABCG Subfamily
ABCG1 ABCG2 ABCG5
ABCG4 ABCG3 ABCG8
Macrolide exporters
MacB
Other putative ABC transporters
YojI
PvdE
SyrD
YddA
Permease protein
Substrate-binding protein
ATP-binding protein
(E.coli, BtuFCD)
ATP
Substrate
Substrate
Membrane-spanning domain (MSD)
Nucleotide-binding domain (NBD)
(E.coli, MsbA)
ATP
02010 6/24/09
(c) Kanehisa Laboratories

Figure 2.14: KEGG pathway of the ABC transporters of *C. sakazakii* The figure shows the colored KEGG pathway of the ABC transporters of *C. sakazakii*. Red bordered boxes indicate enzymes of *C. sakazakii* that could be mapped to KEGG orthologous groups. Black borders indicate enzymes of the pathway that do not exist in *C. sakazakii*. A blue background indicates that the enzyme is member of a KEGG orthologous group.

Active efflux is a recognized virulence mechanism in *Enterobacteriaceae*, contributing to survival in the host‘s gastro intestinal tract. The membrane associated pumps involved in this mechanism extrude a range of xenobiotic compounds from the cell, including bile salts, antibiotics, disinfectants, sanitizers and dyes [315]. Antibiotic resistance proteins which cannot be found in *C. sakazakii* are homologs to the H+ an-

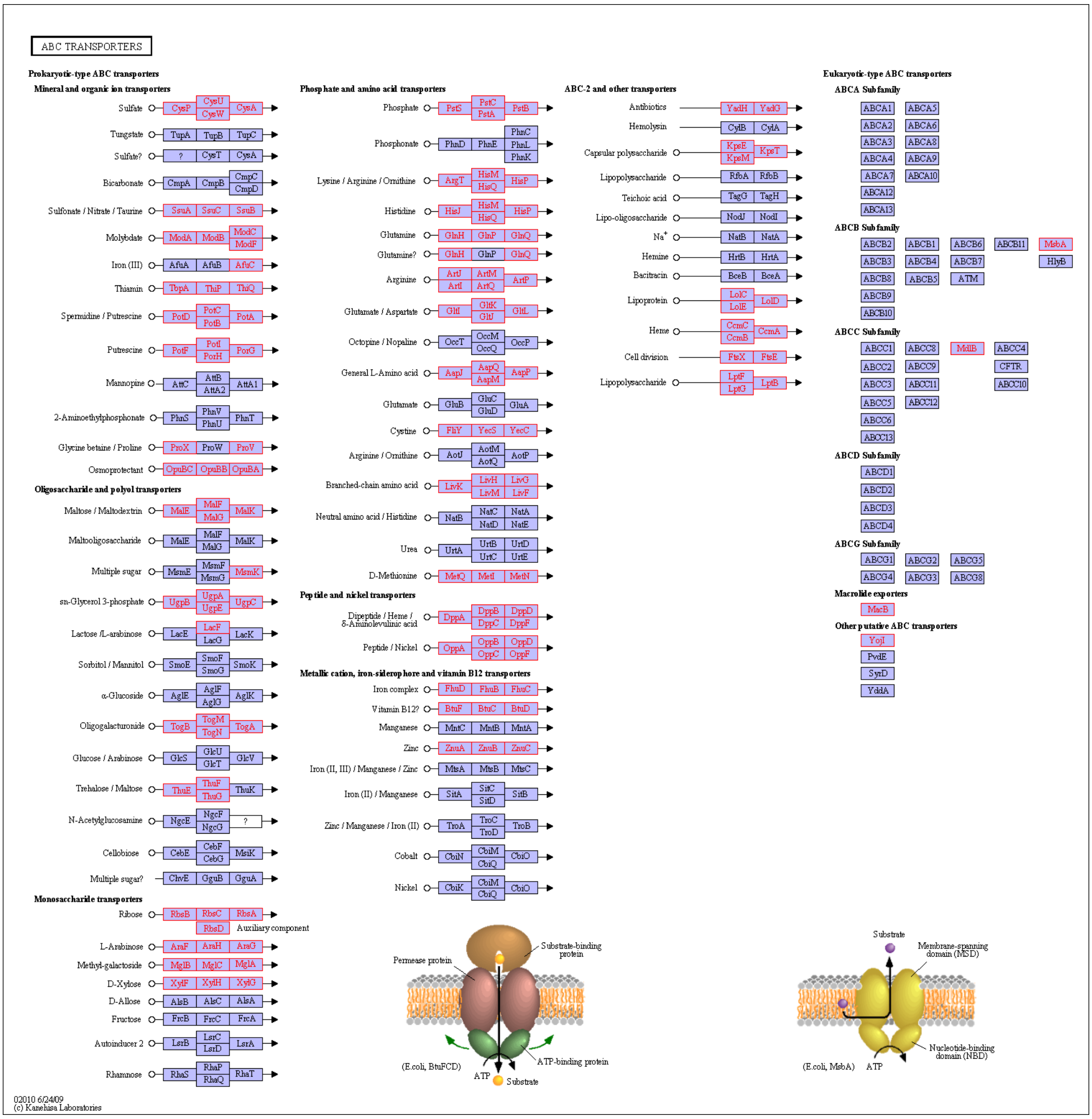

ABC TRANSPORTERS
Prokaryotic-type ABC transporters
Mineral and organic ion transporters
Sulfate CysP CysU CysW CysA
Tungstate TupA TupB TupC
Sulfate? ? CysT CysA
Bicarbonate CmpA CmpB CmpC CmpD
Sulfonate / Nitrate / Taurine SsuA SsuC SsuB
Molybdate ModA ModB ModC ModF
Iron (III) AfuA AfuB AfuC
Thiamin TbpA ThiP ThiQ
Spermidine / Putrescine PotD PotC PotB PotA
Putrescine PotF PotI PotH PotG
Mannopine AttC AttB AttA2 AttA1
2-Aminoethylphosphonate PhnS PhnV PhnU PhnT
Glycine betaine / Proline ProX ProW ProV
Osmoprotectant OpuBC OpuBB OpuBA
Oligosaccharide and polyol transporters
Maltose / Maltodextrin MalE MalF MalG MalK
Maltooligosaccharide MalE MalF MalG MalK
Multiple sugar MsmE MsmF MsmG MsmK
sn-Glycerol 3-phosphate UgpB UgpA UgpE UgpC
Lactose /L-arabinose LacE LacF LacG LacK
Sorbitol / Mannitol SmoE SmoF SmoG SmoK
α-Glucoside AglE AglF AglG AglK
Oligogalacturonide TogB TogM TogN TogA
Glucose / Arabinose GlcS GlcU GlcT GlcV
Trehalose / Maltose ThuE ThuF ThuG ThuK
N-Acetylglucosamine NgcE NgcF NgcG ?
Cellobiose CebE CebF CebG MsiK
Multiple sugar? ChvE GguB GguA
Monosaccharide transporters
Ribose RbsB RbsC RbsA RbsD Auxiliary component
L-Arabinose AraF AraH AraG
Methyl-galactoside MglB MglC MglA
D-Xylose XylF XylH XylG
D-Allose AlsB AlsC AlsA
Fructose FrcB FrcC FrcA
Autoinducer 2 LsrB LsrC LsrD LsrA
Rhamnose RhaS RhaP RhaQ RhaT
Phosphate and amino acid transporters
Phosphate PstS PstC PstA PstB
Phosphonate PhnD PhnE PhnC PhnL PhnK
Lysine / Arginine / Ornithine ArgT HisM HisQ HisP
Histidine HisJ HisM HisQ HisP
Glutamine GlnH GlnP GlnQ
Glutamine? GlnH GlnP GlnQ
Arginine ArtJ ArtI ArtM ArtQ ArtP
Glutamate / Aspartate GltI GltK GltJ GltL
Octopine / Nopaline OccT OccM OccQ OccP
General L-Amino acid AapJ AapQ AapM AapP
Glutamate GluB GluC GluD GluA
Cystine FliY YecS YecC
Arginine / Ornithine AotJ AotM AotQ AotP
Branched-chain amino acid LivK LivH LivM LivG LivF
Neutral amino acid / Histidine NatB NatC NatD NatA NatE
Urea UrtA UrtB UrtC UrtD UrtE
D-Methionine MetQ MetI MetN
Peptide and nickel transporters
Dipeptide / Heme / δ-Aminolevulinic acid DppA DppB DppC DppD DppF
Peptide / Nickel OppA OppB OppC OppD OppF
Metallic cation, iron-siderophore and vitamin B12 transporters
Iron complex FhuD FhuB FhuC
Vitamin B12? BtuF BtuC BtuD
Manganese MntC MntB MntA
Zinc ZnuA ZnuB ZnuC
Iron (II, III) / Manganese / Zinc MtsA MtsB MtsC
Iron (II) / Manganese SitA SitC SitD SitB
Zinc / Manganese / Iron (II) TroA TroC TroD TroB
Cobalt CbiN CbiM CbiQ CbiO
Nickel CbiK CbiM CbiQ CbiO
ABC-2 and other transporters
Antibiotics YadH YadG
Hemolysin CylB CylA
Capsular polysaccharide KpsE KpsM KpsT
Lipopolysaccharide RfbA RfbB
Teichoic acid TagG TagH
Lipo-oligosaccharide NodJ NodI
Na+ NatB NatA
Hemine HrtB HrtA
Bacitracin BceB BceA
Lipoprotein LolC LolE LolD
Heme CcmC CcmB CcmA
Cell division FtsX FtsE
Lipopolysaccharide LptF LptG LptB
Eukaryotic-type ABC transporters
ABCA Subfamily
ABCA1 ABCA5 ABCA2 ABCA6 ABCA3 ABCA8 ABCA4 ABCA9 ABCA7 ABCA10 ABCA12 ABCA13
ABCB Subfamily
ABCB2 ABCB1 ABCB6 ABCB11 MsbA ABCB3 ABCB4 ABCB7 HlyB ABCB8 ABCB5 ATM ABCB9 ABCB10
ABCC Subfamily
ABCC1 ABCC8 MdlB ABCC4 ABCC2 ABCC9 CFTR ABCC3 ABCC11 ABCC10 ABCC5 ABCC12 ABCC6 ABCC13
ABCD Subfamily
ABCD1 ABCD2 ABCD3 ABCD4
ABCG Subfamily
ABCG1 ABCG2 ABCG5 ABCG4 ABCG3 ABCG8
Macrolide exporters
MacB
Other putative ABC transporters
YojI PvdE SyrD YddA
Permease protein
Substrate-binding protein
ATP-binding protein
(E.coli, BtuFCD)
ATP
Substrate
Substrate
Membrane-spanning domain (MSD)
Nucleotide-binding domain (NBD)
(E.coli, MsbA)
ATP
02010 6/24/09
(c) Kanehisa Laboratories

Figure 2.15: KEGG pathway of the ABC transporters of *C. turicensis* The figure shows the colored KEGG pathway of the ABC transporters of *C. turicensis*. Red bordered boxes indicate enzymes of *C. turicensis* that could be mapped to KEGG orthologous groups. Black borders indicate enzymes of the pathway that do not exist in *C. turicensis*. A blue background indicates that the enzyme is member of a KEGG orthologous group.

tiporter for Pristinamycin and Rifamycin Q54806 (Ctu_1p00220), and the multidrug resistance pump P24181 (Ctu_03210) which is according to the TCDB [316] a multidrug (acriflavin, doxorubicin, ethidium, rhodamine 6G, SDS, deoxycholate) resistance pump (required for normal chromosomal condensation and segregation as well as cell division) [317]. Contrarily, the genome of *C. sakazakii* contains GI:156936562 and GI:156936563, both TriABC-OpmH homologs which is a triclosan resistance efflux pump.

Analysis of the *C. turicensis* genome using the UniProtKB/Swiss-Prot keyword "antibiotic resistance" revealed a number of multidrug resistance/efflux pumps including MarRA (Ctu_21730, Ctu_21740) and AcrBAR (Ctu_10690, Ctu_10700, Ctu_10710). The latter has been proposed to confer resistance to acriflavin but is has also been reported that the AcrAB-TolC efflux system may expel a broader range of antibiotic classes, detergents, biocides, and dyes [318].

The multidrug efflux pump encoded by the MdtK (Ctu_19680) protein is involved in resistance to many drugs such as certain fluoroquinolones (norfloxacin, enoxacin), tetraphenylphosphonium ion (TPP), deoxycholate, doxorubicin, trimethoprim, chloramphenicol, fosfomycin, acriflavine, ethidium bromide and benzalkonium [319] and the MacAB (Ctu_14880, Ctu_14890) proteins represent a macrolide type 1 excretion system [320].

Additionally, the following proteins were annotated as multidrug resistance proteins: MdtG (Ctu_09560) is involved in resistance to fosfomycin and deoxycholate, MdtH (Ctu_16420) conferring resistance to norfloxacin and enoxacin [319], and the EmrAB proteins (Ctu_32840, Ctu_32850) are responsible for resistance to substances with high hydrophobicity [321].

The polyspecificity of such multidrug efflux transporters represents a general resistance phenotype that can result in the acquisition of additional antimicrobial resistance.

2.2.6.3 Metabolic pathways relevant for plant-association

The computational method [322] of Gabi Kastenmüller from the Institute of Bioinformatics and Systems Biology at the Helmholtz Zentrum München allows to identify metabolic pathways distinguishing two sets of organisms ("relevant paths").

In order to analyze whether *Cronobacter* spp. show pathways typical for a plant associated environment, Gabi compared the predictions made for the metabolism of the *Cronobacter* spp. to the metabolic pathways predicted for species known to be plant associated from the literature. In order to identify pathways relevant in such an environment a computer-based approach that has been successfully applied for various microbial habitat-related traits previously [322] was applied. This analysis can be divided into three major steps:

1. **Analysis of the predicted metabolic pathways** For assessing the metabolisms of 14 plant associated and 38 non-plant associated species, their metabolic capabilities were represented by a comprehensive set of entire metabolic pathways (290) provided in the BioPath database [323]. A score for each pathway was computed indicating its coverage in the species under consideration. Thereby, the score value 1 means that all enzymes needed for the pathway are present in this species according to automatic annotations taken from the genome database PEDANT [72], the score value 0 means that none of the enzymes is predicted to be encoded in the species' genome. The pathway score is determined based on the ratio of the reactions that are predicted to be catalyzed by the species and all the reactions

forming the pathway. In order to take the importance of key enzymes into account, this ratio is additionally weighted by the number of occurrences of these reactions in other pathways within the pathway database. (For more details on this score-based pathway prediction method see [324]).

2. **Detecting metabolic pathways relevant for plant association** In order to detect the pathways differentiating the two groups (excluding the two *Cronobacter* spp.) the program provided in Kastenmüller et al., 2009 [322] was used. Based on the score-based pathway predictions and the information (binary: yes/no) on the species' plant association, this program identified metabolic pathways that are relevant in distinguishing plant associated species from non-plant associated species. For this purpose, the program used supervised machine learning techniques, namely three different attribute selection methods. The application of these methods resulted in rankings of pathways according to their relevance for the distinction of the groups. Only pathways ranked among the ten most relevant pathways in at least one of the three lists and that showed an average rank below 30 were considered. This resulted in 13 relevant pathways for plant association. We supplementary regarded two pathways as relevant that have been highly ranked but did not pass the filtering: "Biosynthesis of betaine" and "Biosynthesis of L-ascorbate".

3. **Grouping of organisms by their pathway profiles** The metabolic profiles of the two *Cronobacter* spp. and the plant associated and non-plant associated species used for the detection of the relevant pathways were clustered using hierarchical clustering.

14 plant-associated and 38 non-plant-associated organisms have been compared and 15 paths distinguishing plant-associated from non-plant-associated bacteria could be detected (paper in preparation). The two *Cronobacter* spp. are clustered near other plant-associated bacteria in the cluster tree at which not all plant-associated bacteria are clustered together (Figure 2.16).

One of the pathways identified as relevant for plant association is the biosynthesis of menaquinone (Vitamin K2), a derivative of naphthoquinone, that can be found in many bacteria in the large intestine [325]. The main function of menaquinone is the electron transfer in the process of anaerobic respiration [326] which creates ATP. *C. turicensis* encodes 4.1.3.36 naphthoate synthase (Ctu_28940) as well as 5.4.4.2 isochorismate synthase (Ctu_28970, Ctu_30470) from the pathway. The naphthoquinone derivate phylloquinone (Vitamin K1) is naturally found in a wide variety of green plants. An example is lawsone, a naphthoquinone derivate of the henna plant *Lawsonia inermis*, which is used as red-orange dye and is also known as antimicrobial drug [327, 328]. The *Cronobacter* spp. are resistant against this drug, even though the reason is not known yet. No transporter for phylloquinones could be identified in the genome. It has been shown that phylloquinone can be converted into menaquinone in mice [329] but it has also been shown that the conversion in rats is not dependent on the availability of gut bacteria [330]. The question whether *C. turicensis* can convert phylloquinone into menaquinone remains to be elucidated.

2.2.7 Conclusion

The sequencing of the genome of a second *Cronobacter* sp., *Cronobacter turicenis* strain LMG 23827, enabled comparisons to the distinctly related species *Cronobacter sakazakii* ATCC BAA-894 and other members of the family *Enterobacteriaceae* on the genome level.

The comparison of the two *Cronobacter* spp. on the DNA level revealed a high degree of synteny, although a region of the chromosome of *C. sakazakii* is encoded on plasmid 3 of *C. turicensis*.

The comparison of the hypothetical proteomes of representative *Enterobacteriaceae* and the two *Cronobacter* spp. showed a conserved “enterobacterial backbone” as anticipated. Nevertheless both *Cronobacter* genomes show remarkable features specific for the genus *Cronobacter* that give insights into the genetic basis for their pathogenicity.

The capability of *Cronobacter* spp. to colonize eukaryotes such as plants and humans and to cause rare but severe infections in neonates and preterm infants raise the question which molecular factors facilitate these lifestyles. The Type IV and a Type VI secretion system as well as an array of proteins with eukaryotic like protein domains are encoded on the genomes and give a potential explanation for the potential of transferring DNA and effector proteins from the bacterial to the host cell as a mechanism of interaction with a eukaryotic host. Additionally both genomes encode diverse transporters that may be responsible for the resistance of *Cronobacter* spp. against several antibiotics.

There is genomic support that plants are the natural habitat of *Cronobacter* spp. Evidence of several horizontal gene transfer events that have resulted in gene acquisition can be detected in both *Cronobacter* spp. Most of these potentially horizontally acquired genes are closely related to sequences in non-enterobacterial, often plant-associated bacteria (e.g. *Burkholderiales* and *Rhizobiales*). Although the functions of some of these horizontally transferred genes are unknown, others are clearly required for a lifestyle in a plant associated environment such as the enriched sequences for C4 compound metabolism and flagellar chemotaxis associated sequences. It is also remarkable that the analysis of the protein domains of *Enterobacteriaceae* and the two *Cronobacter* spp. revealed a significant depletion of transposition related and a significant enrichment of chemotaxis related protein domains. Pathways typical for plant-associated organisms could be detected, and it is already known that *Cronobacter* spp. are in general capable to utilize a wide variety of compounds as a sole carbon source, some of them are known to be produced and potentially exudated by plants such as L-arabinose, D-xylose, D-cellobiose and palatinose.

The sequences of the whole genome of *C. turicensis* establish a powerful platform for further functional genomics research of this organism. This is an important prerequisite towards future development of countermeasures against this foodborne pathogen.

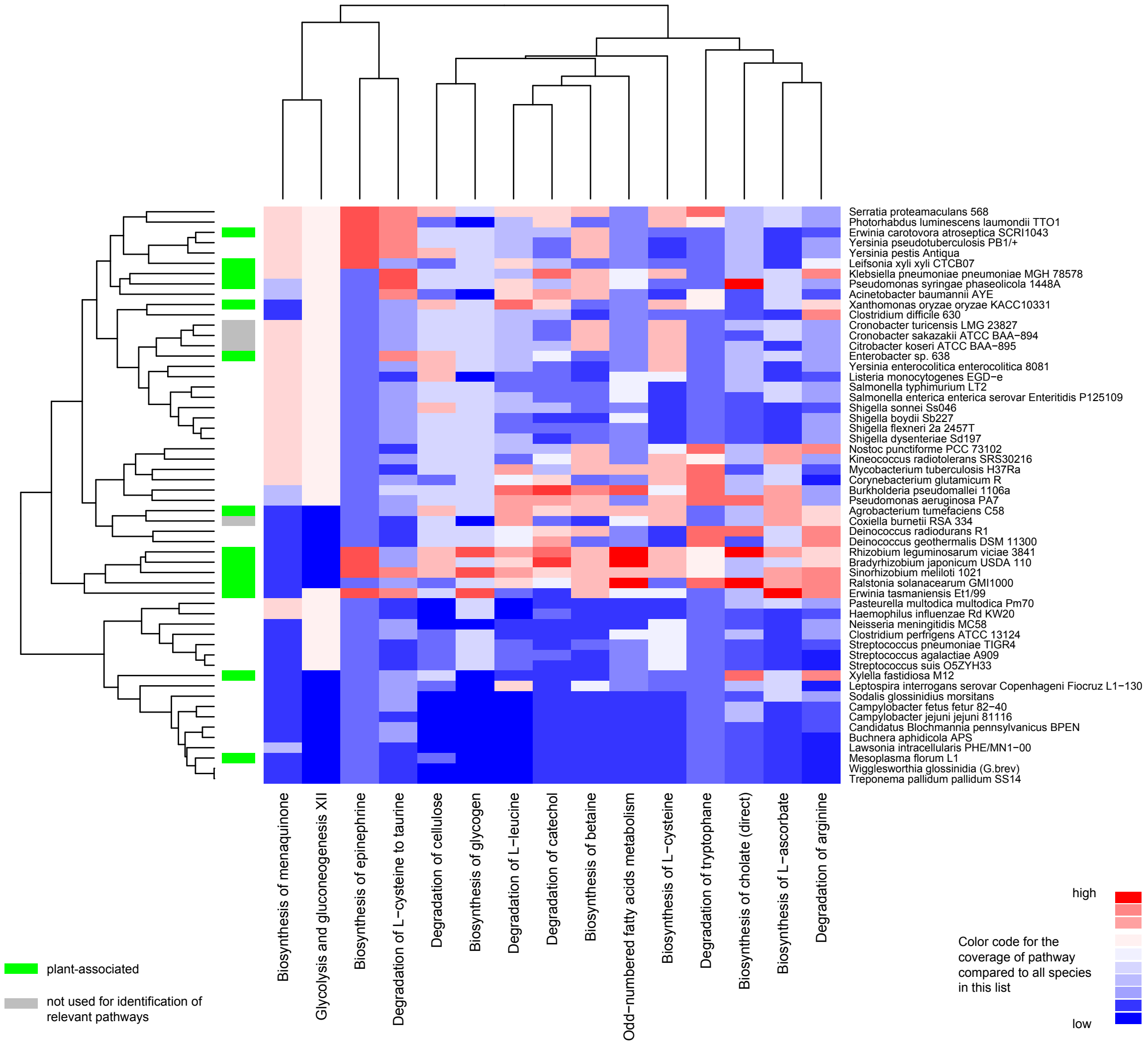


Figure 2.16: Heatplot for the clustering of plant-associated and non-plant associated organisms based on relevant pathways The graphic shows the heatplot for the clustering of plant-associated organisms and organisms that are not associated with plants based on the pathways that have been identified to be most relevant for this distinction. Green bars on the left side show organisms classified as plant-associated (in literature); white color denotes species not associated with plants and grey bars show organisms that have not been taken into account for the detection of the relevant pathways (species for which we could not find enough evidence in literature for being plant-associated or non-plant-associated). The relative coverage (pathway score) of the species considered in this plot is represented by a color code, a color gradient ranging from red to blue via white. Thereby, red fields in the heat matrix correspond to pathways with higher pathway scores (i.e. higher coverage), blue fields correspond to pathways with lower pathway scores in relation to the scores reached by all species shown in this plot.

2.3 Development of a comprehensive chlamydiae genome database

2.3.1 Motivation

Chlamydiae are agents of several sexually transmittable diseases. They occur in a broad range of hosts including animals and humans. Research about their biology is therefore of interest and several of their genomes have been or are currently sequenced. In addition to basic genomic data, data about gene expression, proteomics, metabolic capabilities, and much more has been created and is currently created by the scientific community. This data is mainly contained in the literature or partly in diverse biological databases.

The various kinds of information are available from different sources and are spread all over the internet. To find and to make the data systematically usable is laborious and time consuming, since there exists no central resource that provides the mapping between the contents of the data sources and offers all information about the phylum *Chlamydiae* in one place.

To overcome this problem, a joint project between the University of Vienna and the Technische Universität München has been launched, the development of the "Comprehensive Chlamydia Genome Database", the "ChlamydiaeDB".

At the conference of the Chlamydia Basic Research Society (CBRS) in Little Rock USA in March 2009 we asked the participants about the value of a web portal for *Chlamydiae* genomes. Of 34 scientists 35% said that such a portal is essential, 56% very important, 9% slightly important, and 0% not important for their work (Figure 2.17, A).

The development of a web portal for *Chlamydiae* is therefore of high importance for the field.

2.3.2 Criteria for a comprehensive chlamydiae genome database

The collaboration with scientists working with genomic data of *Chlamydiae* and the experience with the genome database for the environmental chlamydia *Protochlamydia amoebophila* UWE25 showed the requirements that a comprehensive resource for the genomes of the whole phylum *Chlamydiae* has to fulfill:

- **Content** The following contents should be available in a comprehensive resource for all publicly available genomes of the phylum *Chlamydiae*:
  - **All chlamydial genomes in one place** In order to make the data about *Chlamydiae* comparable it is necessary to provide data for all members of the phylum *Chlamydiae* in one place.
  - **All available data in one place** The different kinds of available data for each of the chlamydial species are spread over a range of databases over the

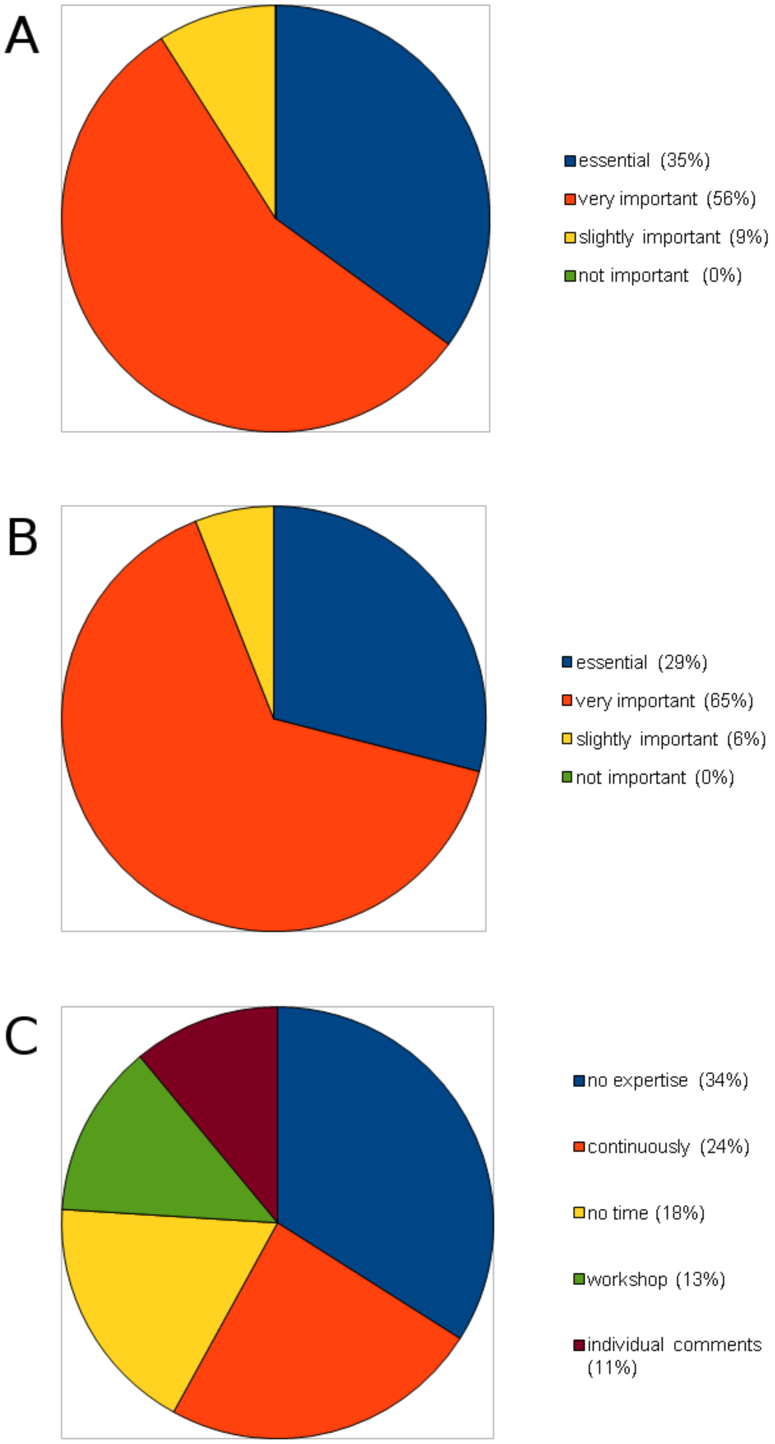


Figure 2.17: Survey about chlamydial genome databases and re-annotation of chlamydial genomes 34 scientists from the chlamydiae field participated in the survey. **A:** Value of a web portal for *Chlamydiae* genomes **B:** Importance of re-annotation of chlamydial genomes **C:** Willingness to contribute to the re-annotation of chlamydial genomes

internet. There are several problems with that, the mapping of different gene identifiers is only one of them. Therefore another goal is to provide all different kinds of data in one single place so that they are represented in a consistent and comparable way.

  - **Access to information from literature** As knowledge is mostly contained in the literature before it is included in other databases especially up-to-date literature should be included in the resource.
  - **Integration of data from experiments** Data from experiments is precious as it is more reliable than knowledge only inferred by homology, for example. Therefore a specific emphasis should be put on the integration of these data into the resource. The data can be of different kinds:
    - **Single Nucleotide Polymorphism (SNP) data** As more and more genomes are being sequenced using next-generation sequencing technologies it becomes important to provide a possibility to deal with SNPs detected in the resequenced genomes in comparison to the reference genomes.
    - **Transcript data** The information about transcripts of chlamydial genomes provides evidence for the transcription of genes, can provide information for the improvement of gene starts, and can provide information about unknown genes and non-coding elements.
    - **Proteome data** Proteome data can provide evidence for the existence of a gene product.

- **Functionality** The following functionality should be provided by a comprehensive resource for all publicly available genomes of the phylum *Chlamydiae*:
  - **Gene centric view** All available information should be available in one place for each genetic element contained in the database.
  - **Group centric view** A protein is not only characterized by the annotations available in the gene centric view, but especially by the protein's context. Therefore instead of reporting data of a single gene the resource should present a gene in its functional and evolutionary context. Each gene ("the query") is connected to different "groups of interest". These groups comprise proteins, that are informative in respect to the query. These are orthologous groups or neighboring genes. The information connected to a group should be presented in a comprehensible view, enriched with annotations.
  - **Tools for the analysis of user defined data sets** The resource should provide easy to use tools for the analysis of user defined data sets. This concept extends current resources, which are merely navigable data-repositories extended with some comparative tools on pre-calculated data, towards a tool which allows to work with user-specific data like sets of proteins identified in an experiment.

  Amongst other functionality the following tools should be provided:

    * **Retrieval of all information about a list of proteins** It should be possible to retrieve all information available for a list of proteins at once without the need to navigate to the report page for each of the proteins separately.

    * **Feature enrichment in a list of proteins** It should be possible to retrieve the features for a user defined set of proteins that are enriched or depleted in comparison to another user defined set of proteins.

    * **Graphical pathway comparison between organisms** As we observed that some of our cooperation partners wanted to have a possibility to color the KEGG maps for their organisms automatically, the resource should also include a possibility to compare pathways between organisms graphically.

  - **Manual annotation possibilities** In order to be able to gain valuable feedback from human annotators it should be possible to annotate every entry in the new resource so that this information eventually even can flow back into the primary data resources like GenBank.

  - **Comprehensive search possibilities** The same genetic element may have different identifiers in different public databases. As scientists work with different kinds of these identifiers it should be possible to search with many kinds of identifiers. It should also be possible to search only in specific organisms, e.g. search for omcB only in pathogenic chlamydia. The same should be possible for the search of the next homologous sequences to a sequence, e.g. search all homologs for omcB in pathogenic chlamydia.

  - **Structuring of search results by integration of taxonomy and orthologous information** It should be possible to structure the search results by their properties, e.g. by their taxonomic distribution or their membership in orthologous groups. This makes it possible to show all eggNOG clusters hit by a search query, for example.

- **Technical** The following technical conditions should be provided by a comprehensive resource for all publicly available genomes of the phylum *Chlamydiae*:

  - **Up-to-dateness with little manual effort** A comprehensive resource is only valuable if it always contains up-to-date information. As every update costs time it should be ensured that this can be reached with as little manual effort as possible.

  - **Easy maintainability** It can be seen in many genome projects that they cannot be continued after the initial funding period as there is no one who would be able to maintain the databases. Therefore the maintenance should be doable with as little manual effort as possible and there should be as little dependencies as possible.

- **Easy extensibility** Often new kinds of information become available that could not be foreseen when designing the resource. Therefore the resource should be designed in a flexible way to allow easy integration of new methods and data types.

2.3.3 Existing resources

First it was checked whether there already existed a resource fulfilling these requirements (section 1.7.7).

The general repositories Genbank, EMBL, DDBJ contain the DNA sequences of all members of the phylum *Chlamydiae* and are primary resources. Their goal is not to be a comprehensive resource for a specific group of organisms.

The resources specialized on information for a specific member or some members of the phylum *Chlamydiae* are not sufficient for the previously stated criteria as they do not contain information about all members of the phylum *Chlamydiae* and are therefore not suited (and designed) for comparative analyses.

The Genome Information Broker [250] has no comparative genomics capabilities and lacks information like orthologous groups or literature. The Microbial Genome Database for Comparative Analysis (MBGD) [251, 252, 253] offers comparative analyses from various points of views. But it does not contain all available data for *Chlamydiae* in one place (e.g. no literature information), it does not allow users to analyze user defined datasets, structuring search results by taxonomy or clusters is not possible, and it offers no possibilities for manual annotations.

The representation of the complete available current knowledge about the phylum *Chlamydiae* in combination with the ability to work with own uploaded data is perfectly new. Furthermore the concept of the “groups of interest” goes beyond typical views in other databases. Therefore a novel solution had to be implemented.

2.3.4 Concept

A three-tier client-server architecture (section 1.7.8.1) should be used for ChlamydiaeDB, as it already fulfills many of the important goals, among them easy maintainability and easy extensibility. Figure 2.18 shows the three tiers of the three-tier: presentation tier, application tier and data tier.

The concept section is structured by topics not by tiers. Therefore the three tiers are contained within the respective section for each topic.

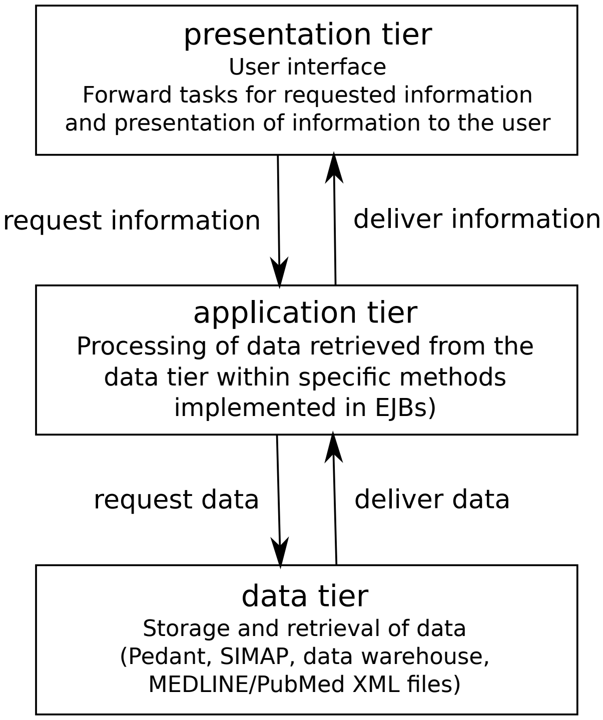

presentation tier
User interface
Forward tasks for requested information
and presentation of information to the user
request information
deliver information
application tier
Processing of data retrieved from the
data tier within specific methods
implemented in EJBs)
request data
deliver data
data tier
Storage and retrieval of data
(Pedant, SIMAP, data warehouse,
MEDLINE/PubMed XML files)

Figure 2.18: Three tier concept of ChlamydiaeDB A three-tier [254] architecture is used as concept for the implementation of the ChlamydiaeDB. The three layers “presentation tier”, “application tier” and “data tier” are separated from each other. A change in one of these layers does not affect the other layers as long as the interfaces between them do not change. This results in a highly flexible, easy extensible and easy maintainable solution. The presentation tier is responsible for the interaction with the user, the application tier is responsible for the integration of various datasources and also contains the application logic. The data tier is responsible for data storage and retrieval. All three tiers communicate with each other using specified interfaces.

2.3.4.1 Integration of heterogenous data sources

The complete and up-to-date representation of the current knowledge needs an update strategy supported by the database. There can be different kinds of data integration distinguished that differ in the amount of data that has to be stored within the new resource:

- **On-the-fly integration of information** Some data can be integrated from the primary data sources using different techniques. One is the retrieval of data from EJBs running on an application server. Examples are the retrieval of similarities, clusters, or protein domains from SIMAP [94]. Another way is the data retrieval from external data sources using Web Services that are provided by many institutes. An example is the PEDANT [72] Web Service.
- **Regularly updated data** Data like XML files containing MEDLINE/PubMed literature can be downloaded on a regular basis, but there is no possibility to use an online service that would allow to retrieve the information on-the-fly. Therefore the data needs to be downloaded, pre-processed and stored locally on a regular basis so that it can be accessed when it is required.
- **Specifically computed information** Some information is specific for the novel resource and needs to be pre-computed. An example is the Type-III secretion prediction for proteins. These predictions need to be done for every new sequence in the database.
- **Data imported once** Some information only needs to be imported once. Examples are transcript and proteome data or information about SNPs.
- **Interactive data** As manual annotation should be possible in the new resource, there has to be a possibility to store data until it is accepted by an authorized annotator and then made available for the public.

2.3.4.2 Data storage in a data warehouse

Some data needs to be stored as already discussed in section 2.3.4.1. Additionally external sources are not always reliably available or not optimized or suited for the kind of requests needed for the new *Chlamydiae* resource.

Therefore a data warehouse should be set up containing own data and computations as well as information not available or not conveniently available from external data sources.

The disadvantage of the data warehouse, that it needs more disk space is not regarded as problematic as the amount of information is limited for ChlamydiaeDB.

2.3.4.3 Initialization and update strategy for ChlamydiaeDB

2.3.4.3.1 Initialization The (re-)initialization of the ChlamydiaeDB should be as automatic as possible. The following steps should be performed:

1. **Delete automatically generated database content** For the case, that this is a re-initialization the previously automatically inserted content is deleted first.

2. **Retrieval of automatically computed information** Get and insert the sequences for all PEDANT [72] organisms belonging to the phylum *Chlamydiae* and insert the information into the data warehouse.

3. **Get synonymous names for genetic elements** Get and insert synonymous names for the genetic elements using information from PEDANT, RefSeq [52] and UniProtKB [267].

4. **Compute and insert sequence related information** Compute and insert sequence related information, e.g. Type-III secretion predictions.

5. **Get annotations** Get annotations for all genetic elements and create indices for better retrieval in the data warehouse.

6. **Insert data from experiments** Insert SNP, transcript and proteome experiment data.

7. **Parse whole literature from MEDLINE/PubMed** Parse the literature from MEDLINE/PubMed and insert matching documents into the database. Only literature related to *Chlamydiae* is saved in order to keep the database performant.

2.3.4.3.2 Update strategy There can four update levels be distinguished:

1. **on-the-fly** This information is always up-to-date and does not need regular updates as it is retrieved via EJBs or Web Services.

2. **daily** One kind of data has to be stored in the data warehouse, e.g. aggregated information about annotation from Pedant (how many proteins have annotation X) or from Blast2GO [93]. On-the-fly queries would be too slow in these cases.

   Another possible kind of data is data available for download and only updated once a day, e.g. MEDLINE/PubMed literature. It is not necessary and reasonable to download the information for every request and to process it every time.

3. **monthly** Once a month or if SIMAP has been updated for example, it has to be checked:

   - whether there are new chlamydial genomes available
   - whether there are new sequences for the already known chlamydial genomes available
   - whether there are new RefSeq and UniProtKB releases available, that is new names for the genetic elements

   These cases are not independent of the other updates so that it might be necessary to also execute “daily” or “if required” updates.

4. **if required** These updates include computations that only need to be done if a new version of the software is available, e.g. Type-III secretion predictions.

The execution of specific scripts on a regular basis ensures the up-to-dateness of ChlamydiaeDB.

2.3.4.4 Easy maintainability and extensibility

Sometimes the primary data sources change their data formats or new kinds of information should be added that could not be foreseen when designing the new resource. Therefore it should be easy to maintain and extend the resource without the need to change everything.

2.3.4.4.1 Three tier architecture As already mentioned before the three-tier architecture allows to separate data storage and retrieval from application logic and presentation. This architecture is optimal for ChlamydiaeDB, that can intensely use existing resources like SIMAP, PEDANT, FUNAT and by that avoid redundant data storage and achieve automatic up-to-dateness. The connection to the resources can be realized using EJBs or Web Services. Specific data will be hosted in the previously mentioned data warehouse.

2.3.4.4.2 OpenCms and GenRE GenRE is based on OpenCms, implements a three-tier and by that provides reusable elements like data from the application tier in XML format and XSL stylesheets for the presentation tier that can be re-used for other applications.

The problem with OpenCms is that it can only handle one XML and one XSL file for a single page. If information from different data sources has to be displayed, then one merged XML file has to be built within the application tier and to be processed by OpenCms in the presentation tier. This is time consuming and violates the postulated separation between application tier and presentation tier. Therefore OpenCms is not flexible enough for the aimed use.

2.3.4.4.3 Portlets in a portal server Portlets in a portal server provide a clear separation of the three layers of the three tier. Portlets are reusable and flexible as the addition or the removal of a portlet to or from a portal page can be easily done. The separation of the three layers is given as each portlet can have its own JSP dynamically creating the page or its own XML containing the data with the corresponding XSL for the visualization. A portlet can be used on other pages where the same information should be displayed.

2.3.4.5 Summary and used software

ChlamydiaeDB will be implemented using a three-tier architecture. The presentation tier will use the data from the application tier and transform the data in XML format into

HTML using XSL stylesheets. The application tier will be an EJB retrieving information from other EJBs, Web Services or integrating data from the specific data warehouse of ChlamydiaeDB. The data tier will implement specific tools that pre-process the data, e.g. daily check for new literature and fill it into the database.

The data warehouse will run on a MySQL server version 5.1, the application tier will be EJB3s running on a JBoss application server 4.0.4 and the presentation tier will be a Liferay Portal 4.2.2 running on an Apache Tomcat 5.5.26.

2.3.5 SIMAP as a prototype of a web portal

In 2007 it became necessary to update the existing web site of SIMAP [94] as knowledge from SIMAP and other sources like PEDANT [72] or Blast2GO [93] should be integrated.

As we switched to EJBs at that time in order to provide our data internally it was convenient to implement a solution for the web site using Java. My former coworker Richard Gregory presented his NGFN portal (`http://portal.ngfn.de`) based on the portal server Liferay using portlets. After I had consulted him I discovered that this is also be the optimal solution for SIMAP and ChlamydiaeDB.

2.3.5.1 Goals

The SIMAP web portal was built as a proof-of-concept in order to find out to what extent portlets on a portal server are practicable for a *Chlamydiae* resource, that even integrates more data sources as SIMAP and contains additional functionalities.

The following goals were defined for the SIMAP portal:

- **General**
  - **Access to all precomputed information for all publicly available protein sequences within SIMAP** All available precomputed information within SIMAP should be aggregated in one place for all publicly available sequences within SIMAP.
  - **Always up to date** The sequences on the web site and within SIMAP should be always up-to-date with the public data repositories (with a little delay for the computation).
  - **On-the-fly integration of heterogenous data sources** As data should always be up-to-date the possibilities to integrate heterogeneous datasources on the fly should be tested.
  - **Userfriendlieness** The web site should be easy to use nevertheless be comprehensive and contain all information without hiding anything.

- **Functionality**
  - **Searching capabilities**
    - **Fulltext search** It should be possible to perform a fulltext search using a search term in the available names and descriptions of all entries within SIMAP.

      The output should be made available:

      - As a list of hits, similar to commonly used search engines
      - Sorted taxonomically so that the user can decide which hits in which taxonomic organisms should be displayed
    - **Sequence search** It should be possible to perform a search using a protein sequence and to search for similar sequences in SIMAP.
    - **Taxonomic search** Additionally to searching for a search term or search sequence in the whole sequence space and restricting the hits to specific taxonomic branches afterwards, it should be possible to perform a search in restricted taxonomic regions, e.g. only in bacteria.
  - **Mapping of protein identifiers** Due to the use of multiple identifiers for the same protein in different databases, an important but time-consuming task in bioinformatics is the translation of the identifiers of a set of proteins into another domain of identifiers. This task is necessary also for proprietary databases that use special identifiers, that should be mapped to recent public databases. Therefore it should be possible to upload a set of sequences and to get all available names for these sequences back.
  - **Clusters** Cluster information should be integrated into SIMAP as clusters allow to structure the sequence space and allow transfer of knowledge between the members of orthologous clusters.
  - **Sequence homologs** One of the most important features of a database of protein sequence similarities is the retrieval of similar sequences to a query sequence. In order to get a better overview over the results the most similar sequences to a protein sequence should be made available in various views:
    - **BLAST like representation** As many users are familiar with the view of the NCBI BLAST output a similar output should be provided. That is that there is a list of the most similar sequences sorted by descending bitscore. A graphical view should visualize the alignments and be clickable so that the user can jump to the respective alignment.
    - **Taxonomical representation** It should be possible to switch to a taxonomic view of the hits. This way the user gets an impression of the taxonomic distribution of the most similar sequences and can restrict the hits that should be displayed taxonomically.

data source	information	data retrieval via
SIMAP	proteins in various public databases	EJB
	AA sequences	EJB
	clusters of protein sequences	EJB
	protein domains (among others InterPro)	EJB
	sequence similarities “all against all”	EJB
	domain similarities	EJB
SIMAP2GO	GO annotations by Blast2GO	EJB
PEDANT	automatic annotations (e.g. EC, FunCat)	Web Service
	genes and their positions in the genomes	SQL
	nucleotide sequences	SQL
FUNAT	Automatically Assigned FunCat Annotation	Web Service

Table 2.14: External datasources integrated into the SIMAP web portal

∗ **Cluster representation** The information about cluster memberships should be used to show the clusters that contain sequences homologous to the query sequence.

– **Domain homologs** Protein domains as contained in the InterPro database, for example, are the building blocks of life. Even if two protein sequences are not similar on the amino acid sequence level they still may contain conserved shared protein domains (Figure 1.15). If two proteins contain the same protein domains they are more likely to have a common function than proteins not sharing protein domains. Therefore the most similar sequences as measured by their domain similarity should be visualized.

– **Protein report** The protein report should be the central place containing all the information available for a single protein no matter from which information source. The report should be available for every protein.

2.3.5.2 Implementation

The three tier concept intended for ChlamydiaeDB was also used for the SIMAP web portal.

2.3.5.2.1 Integrated data sources The various information available in the SIMAP portal is retrieved from different data-sources using direct connections to the databases (SQL), EJBs and Web Services (Table 2.14). The direct connections to PEDANT are necessary as some information is not provided by the PEDANT Web Service.

2.3.5.2.2 Liferay portal server for the presentation The open source portal server Liferay (`http://www.liferay.com`) was used as interface to the users, as Richard Gregory and Karamfilka Nenova, both from the Institute of Bioinformatics and Systems Biology at the Helmholtz Zentrum München, had experience with it. The portal server can be used with almost any web browser and operating system. No extra software needs to be installed on the clients to be able to use it.

2.3.5.2.3 Inter Portlet Communication The original portlet specification JSR-168 (`http://jcp.org/aboutJava/communityprocess/final/jsr168/`) did not contain any support for Inter Portlet Communication, that is the exchange of information between portlets. This feature was introduced with portlet specification JSR-286 (`http://jcp.org/aboutJava/communityprocess/final/jsr286/`), which is supported in Liferay since version 5.

When the SIMAP portal was implemented, Liferay 4.2.2 was current so that Inter Portlet Communication was not natively available. Therefore a solution from Michelle Osmond (available at `http://mus.purplecloud.net/portlets/index.php`) was used that provides an object for each user session in the memory of the application server. The object can be used to transfer information from one portlet to the other. According to Volker Stümpflen from the Institute of Bioinformatics and Systems Biology at the Helmholtz Zentrum München the Inter Portlet Communication does still not work smoothly in JSR-286. This is the reason why his group still uses the approach of Michelle Osmond for portals (personal communication).

An example illustrating Inter Portlet Communication is a link within one portlet linking to the protein report page with lots of other portlets. As soon as the link is clicked the information about the clicked protein is stored in a user specific "MessageBean" object on the server and the request is redirected to the protein report page. This page contains all portlets responsible for the display of information about the protein. At load time each portlet extracts the information about the protein to be displayed from the MessageBean. This way every portlet can retrieve the information relevant for the protein from the application tier and display it.

2.3.5.2.4 eXtensible Stylesheet Language Transformations (XSLT) The visualization in most of the portlets is done using XSL files that convert the information from the XML files from the application tier into a visible interpretation in HTML. This technique allows for a complete separation between the application logic and the visualization within the presentation tier.

I adapted several of the XSL stylesheets existing from the old OpenCms based SIMAP web site and created new stylesheets with extended functionality together with Franz Hamberger in his applied semester.

2.3.5.3 Results

The SIMAP web portal has been finished in 2007 and was published in the same year [128]. All goals have been achieved. In the following the results will be shown in detail.

2.3.5.3.1 News and documentation pages The easiest way to put information online within Liferay is to use its integrated Content Management System (CMS). This capability is used for the news on the start page or documentation pages (Figure 2.19).

2.3.5.3.2 User forms Forms for the handling of user data, e.g. identifier or sequence searches (see Figure 2.20) have been implemented as JSPs. The JSPs create the forms that are displayed on the page when the portlet is loaded.

2.3.5.3.3 Searching capabilities It is possible to search for search terms and sequences within SIMAP. Additionally it is possible to restrict the search to specified subsections of the taxonomic tree. An overview over the search possibilities can be seen in Figure 2.20.

The search results can either be displayed as list of hits or sorted taxonomically (Figure 2.21). This can help to narrow the search results down in the case that it was searched for a commonly used gene name.

If the sequence should not be known to the SIMAP system then parts of the query sequence are searched in a suffix array of all SIMAP sequences generated by VMATCH (`http://www.vmatch.de`). These results are shown as a list.

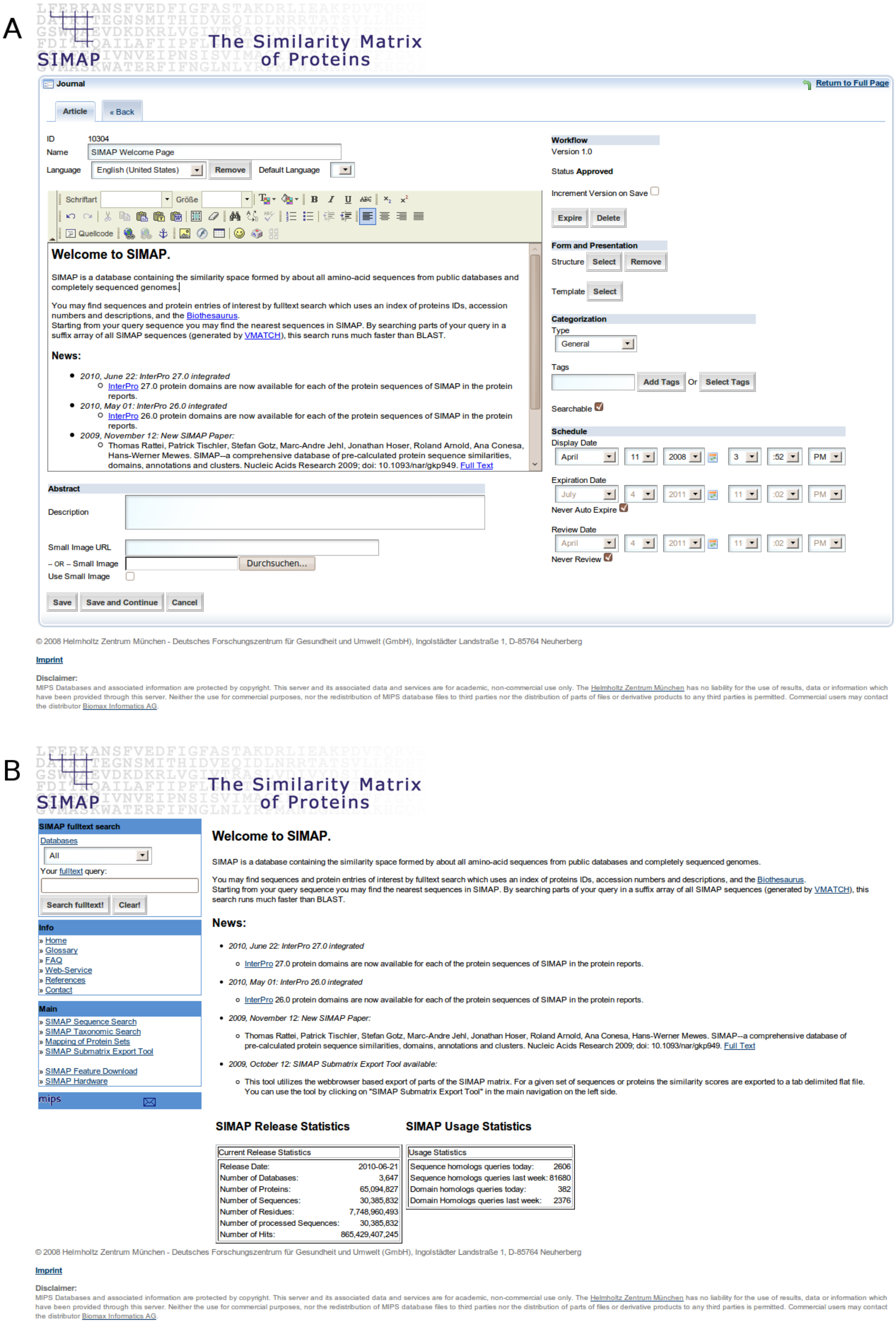
The fulltext search as well as the suffix array search have been implemented by Thomas Rattei within the SIMAP EJBs.

2.3.5.3.4 Mapping of protein sets SIMAP allows to map up to 100 amino acid sequences at once based on the identity of protein sequences by comparison of their MD5 hashes. The output is downloadable as multifasta file containing all known identifiers within SIMAP in the description line of each sequence. This makes mapping of protein sets easy.

2.3.5.3.5 Clusters It is possible to display a cluster and all of the member proteins within SIMAP. The clusters are available via the protein reports (Figure 2.24 C). Besides COG [107], KOG [107] and eggNOG [108] clusters there is also a complete clustering of all SIMAP sequences using Tribe-MCL [331] and a subclustering using domain architecture available (see [128]).

2.3.5.3.6 Sequence homologs Sequence homologs to a query sequence can be reached from search results (Figure 2.21), protein reports (Figure 2.24 B), sequence homologs (Figure 2.22) and domain homologs (Figure 2.23) visualizations. The user can choose between three different visualizations: the BLAST like representation of homologs (Figure 2.22, A), the taxonomic representation of homologs (Figure 2.22, B) and the assignment of homologs to sequence clusters (Figure 2.22, C).

2.3.5.3.7 Domain homologs Domain homologs to a query sequence can be reached from search results (Figure 2.21), protein reports (Figure 2.24 B), sequence homologs (Figure 2.22) and domain homologs (Figure 2.23) visualizations. An example of the



ID 10304
Name SIMAP Welcome Page
Language English (United States) Default Language

Welcome to SIMAP.

SIMAP is a database containing the similarity space formed by about all amino-acid sequences from public databases and completely sequenced genomes.

You may find sequences and protein entries of interest by fulltext search which uses an index of proteins IDs, accession numbers and descriptions, and the Biothesaurus.
Starting from your query sequence you may find the nearest sequences in SIMAP. By searching parts of your query in a suffix array of all SIMAP sequences (generated by VMATCH), this search runs much faster than BLAST.

News:

- *2010, June 22: InterPro 27.0 integrated*
  - InterPro 27.0 protein domains are now available for each of the protein sequences of SIMAP in the protein reports.
- *2010, May 01: InterPro 26.0 integrated*
  - InterPro 26.0 protein domains are now available for each of the protein sequences of SIMAP in the protein reports.
- *2009, November 12: New SIMAP Paper:*
  - Thomas Rattei, Patrick Tischler, Stefan Gotz, Marc-Andre Jehl, Jonathan Hoser, Roland Arnold, Ana Conesa, Hans-Werner Mewes. SIMAP--a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. Nucleic Acids Research 2009; doi: 10.1093/nar/gkp949. Full Text

Abstract

Description

Small Image URL
-- OR -- Small Image
Use Small Image □

Workflow
Version 1.0

Status **Approved**

Increment Version on Save □

Form and Presentation
Structure
Template

Categorization
Type
General

Tags

Searchable ☑

Schedule
Display Date
April 11 2008 3 :52 PM

Expiration Date
July 4 2011 11 :02 PM
Never Auto Expire ☑

Review Date
April 4 2011 11 :02 PM
Never Review ☑

© 2008 Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Ingolstädter Landstraße 1, D-85764 Neuherberg

Imprint

Disclaimer:
MIPS Databases and associated information are protected by copyright. This server and its associated data and services are for academic, non-commercial use only. The Helmholtz Zentrum München has no liability for the use of results, data or information which have been provided through this server. Neither the use for commercial purposes, nor the redistribution of MIPS database files to third parties nor the distribution of parts of files or derivative products to any third parties is permitted. Commercial users may contact the distributor Biomax Informatics AG.

SIMAP fulltext search
Databases
All
Your fulltext query:

Info
» Home
» Glossary
» FAQ
» Web-Service
» References
» Contact

Main
» SIMAP Sequence Search
» SIMAP Taxonomic Search
» Mapping of Protein Sets
» SIMAP Submatrix Export Tool

» SIMAP Feature Download
» SIMAP Hardware

Welcome to SIMAP.

SIMAP is a database containing the similarity space formed by about all amino-acid sequences from public databases and completely sequenced genomes.

You may find sequences and protein entries of interest by fulltext search which uses an index of proteins IDs, accession numbers and descriptions, and the Biothesaurus.
Starting from your query sequence you may find the nearest sequences in SIMAP. By searching parts of your query in a suffix array of all SIMAP sequences (generated by VMATCH), this search runs much faster than BLAST.

News:

- *2010, June 22: InterPro 27.0 integrated*
  - InterPro 27.0 protein domains are now available for each of the protein sequences of SIMAP in the protein reports.
- *2010, May 01: InterPro 26.0 integrated*
  - InterPro 26.0 protein domains are now available for each of the protein sequences of SIMAP in the protein reports.
- *2009, November 12: New SIMAP Paper:*
  - Thomas Rattei, Patrick Tischler, Stefan Gotz, Marc-Andre Jehl, Jonathan Hoser, Roland Arnold, Ana Conesa, Hans-Werner Mewes. SIMAP--a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. Nucleic Acids Research 2009; doi: 10.1093/nar/gkp949. Full Text
- *2009, October 12: SIMAP Submatrix Export Tool available:*
  - This tool utilizes the webbrowser based export of parts of the SIMAP matrix. For a given set of sequences or proteins the similarity scores are exported to a tab delimited flat file. You can use the tool by clicking on "SIMAP Submatrix Export Tool" in the main navigation on the left side.

SIMAP Release Statistics

Current Release Statistics	
Release Date:	2010-06-21
Number of Databases:	3,647
Number of Proteins:	65,094,827
Number of Sequences:	30,385,832
Number of Residues:	7,748,960,493
Number of processed Sequences:	30,385,832
Number of Hits:	865,429,407,245

SIMAP Usage Statistics

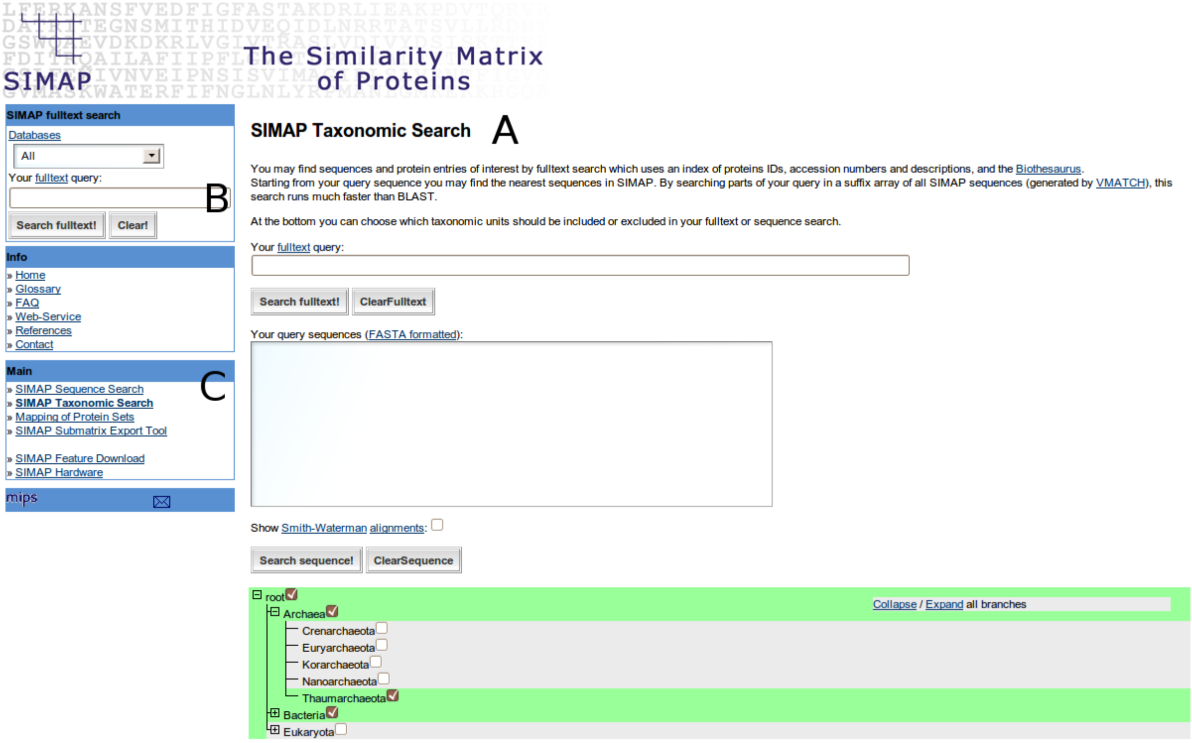
Usage Statistics	
Sequence homologs queries today:	2606
Sequence homologs queries last week:	81680
Domain homologs queries today:	382
Domain Homologs queries last week:	2376

© 2008 Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Ingolstädter Landstraße 1, D-85764 Neuherberg

Imprint

Disclaimer:
MIPS Databases and associated information are protected by copyright. This server and its associated data and services are for academic, non-commercial use only. The Helmholtz Zentrum München has no liability for the use of results, data or information which have been provided through this server. Neither the use for commercial purposes, nor the redistribution of MIPS database files to third parties nor the distribution of parts of files or derivative products to any third parties is permitted. Commercial users may contact the distributor Biomax Informatics AG.

Figure 2.19: Content Management System (CMS) portlet for news within the SIMAP portal A: The Liferay portal provides a graphical editor for writing content displayed on the web sites. The example shows the editor for the news on the start site of the SIMAP portal. **B:** These are the resulting news on the SIMAP start site.



You may find sequences and protein entries of interest by fulltext search which uses an index of proteins IDs, accession numbers and descriptions, and the Biothesaurus.
Starting from your query sequence you may find the nearest sequences in SIMAP. By searching parts of your query in a suffix array of all SIMAP sequences (generated by VMATCH), this search runs much faster than BLAST.

At the bottom you can choose which taxonomic units should be included or excluded in your fulltext or sequence search.

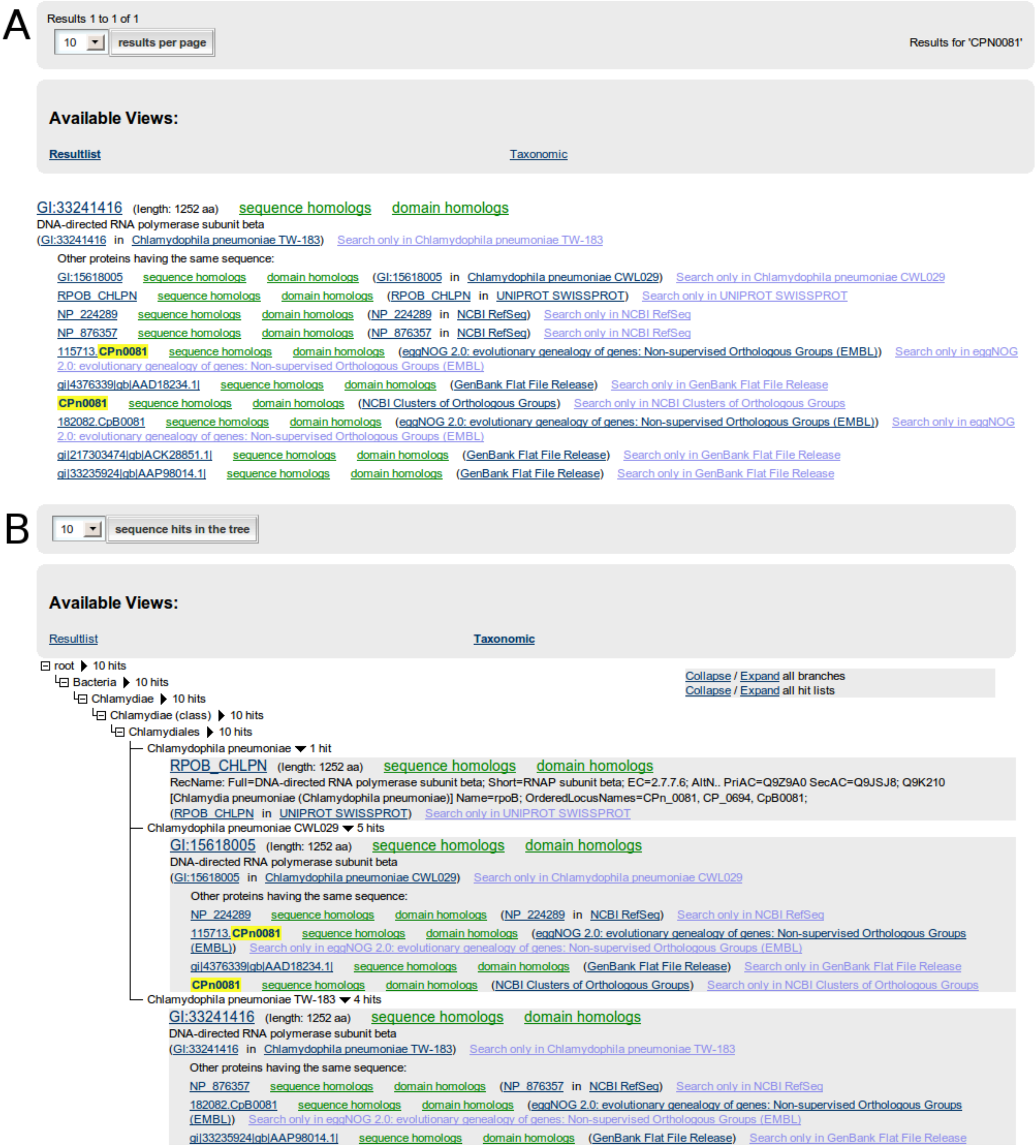
Figure 2.20: Search possibilities in the SIMAP portal A: The taxonomic search can be seen on the right side. It is possible to enter a search term or search sequence and determine in which regions the search should be performed by the selection or deselection of nodes in the taxonomic tree. **B:** This search field is part of the main portlet that is always available on every page of the SIMAP portal. **C:** The sequence search is available using the provided link in the main navigation.

domain homologs visualization can be seen in Figure 2.23. The colored domains indicate different protein domains within the sequences. A protein domain has the same color in all sequences.

2.3.5.3.8 Protein report The protein report is the most feature rich part of the SIMAP web portal. The report combines information from various sources and uses the full potential of portlets on a portal server. Additionally links to sequence and domain homologs (Figure 2.24 B), clusters (Figure 2.24 C), primary datasources of the proteins (Figure 2.24 D), InterPro and its member databases (Figure 2.24 K), PEDANT (Figure 2.24 G, I), GeneOntology (Figure 2.24 H) and FunCat (Figure 2.24 I, J) are included.

Figure 2.24 shows an example of a protein report for protein CPn0081, the DNA-directed RNA polymerase subunit beta, from *Chlamydophila pneumoniae* TW-183.

2.3.5.3.9 Linking to SIMAP The LinkPortlet was specifically implemented to allow linking to SIMAP entries. This portlet allows to link to the protein report, sequence homologs or domain homologs of a specific protein from outside of the portal.



Results 1 to 1 of 1

Results for 'CPN0081'

Available Views:

Resultlist Taxonomic

GI:33241416 (length: 1252 aa) sequence homologs domain homologs
DNA-directed RNA polymerase subunit beta
(GI:33241416 in Chlamydophila pneumoniae TW-183) Search only in Chlamydophila pneumoniae TW-183

Other proteins having the same sequence:

GI:15618005 sequence homologs domain homologs (GI:15618005 in Chlamydophila pneumoniae CWL029) Search only in Chlamydophila pneumoniae CWL029
RPOB_CHLPN sequence homologs domain homologs (RPOB_CHLPN in UNIPROT SWISSPROT) Search only in UNIPROT SWISSPROT
NP_224289 sequence homologs domain homologs (NP_224289 in NCBI RefSeq) Search only in NCBI RefSeq
NP_876357 sequence homologs domain homologs (NP_876357 in NCBI RefSeq) Search only in NCBI RefSeq
115713.CPn0081 sequence homologs domain homologs (eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)) Search only in eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)
gi|4376339|gb|AAD18234.1| sequence homologs domain homologs (GenBank Flat File Release) Search only in GenBank Flat File Release
CPn0081 sequence homologs domain homologs (NCBI Clusters of Orthologous Groups) Search only in NCBI Clusters of Orthologous Groups
182082.CpB0081 sequence homologs domain homologs (eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)) Search only in eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)
gi|217303474|gb|ACK28851.1| sequence homologs domain homologs (GenBank Flat File Release) Search only in GenBank Flat File Release
gi|33235924|gb|AAP98014.1| sequence homologs domain homologs (GenBank Flat File Release) Search only in GenBank Flat File Release

Available Views:

Resultlist **Taxonomic**

Collapse / Expand all branches
Collapse / Expand all hit lists

⊟ root ▶ 10 hits
⊟ Bacteria ▶ 10 hits
⊟ Chlamydiae ▶ 10 hits
⊟ Chlamydiae (class) ▶ 10 hits
⊟ Chlamydiales ▶ 10 hits

Chlamydophila pneumoniae ▼ 1 hit

RPOB_CHLPN (length: 1252 aa) sequence homologs domain homologs
RecName: Full=DNA-directed RNA polymerase subunit beta; Short=RNAP subunit beta; EC=2.7.7.6; AltN.. PriAC=Q9Z9A0 SecAC=Q9JSJ8; Q9K210 [Chlamydia pneumoniae (Chlamydophila pneumoniae)] Name=rpoB; OrderedLocusNames=CPn_0081, CP_0694, CpB0081;
(RPOB_CHLPN in UNIPROT SWISSPROT) Search only in UNIPROT SWISSPROT

Chlamydophila pneumoniae CWL029 ▼ 5 hits

GI:15618005 (length: 1252 aa) sequence homologs domain homologs
DNA-directed RNA polymerase subunit beta
(GI:15618005 in Chlamydophila pneumoniae CWL029) Search only in Chlamydophila pneumoniae CWL029

Other proteins having the same sequence:

NP_224289 sequence homologs domain homologs (NP_224289 in NCBI RefSeq) Search only in NCBI RefSeq
115713.CPn0081 sequence homologs domain homologs (eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)) Search only in eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)
gi|4376339|gb|AAD18234.1| sequence homologs domain homologs (GenBank Flat File Release) Search only in GenBank Flat File Release
CPn0081 sequence homologs domain homologs (NCBI Clusters of Orthologous Groups) Search only in NCBI Clusters of Orthologous Groups

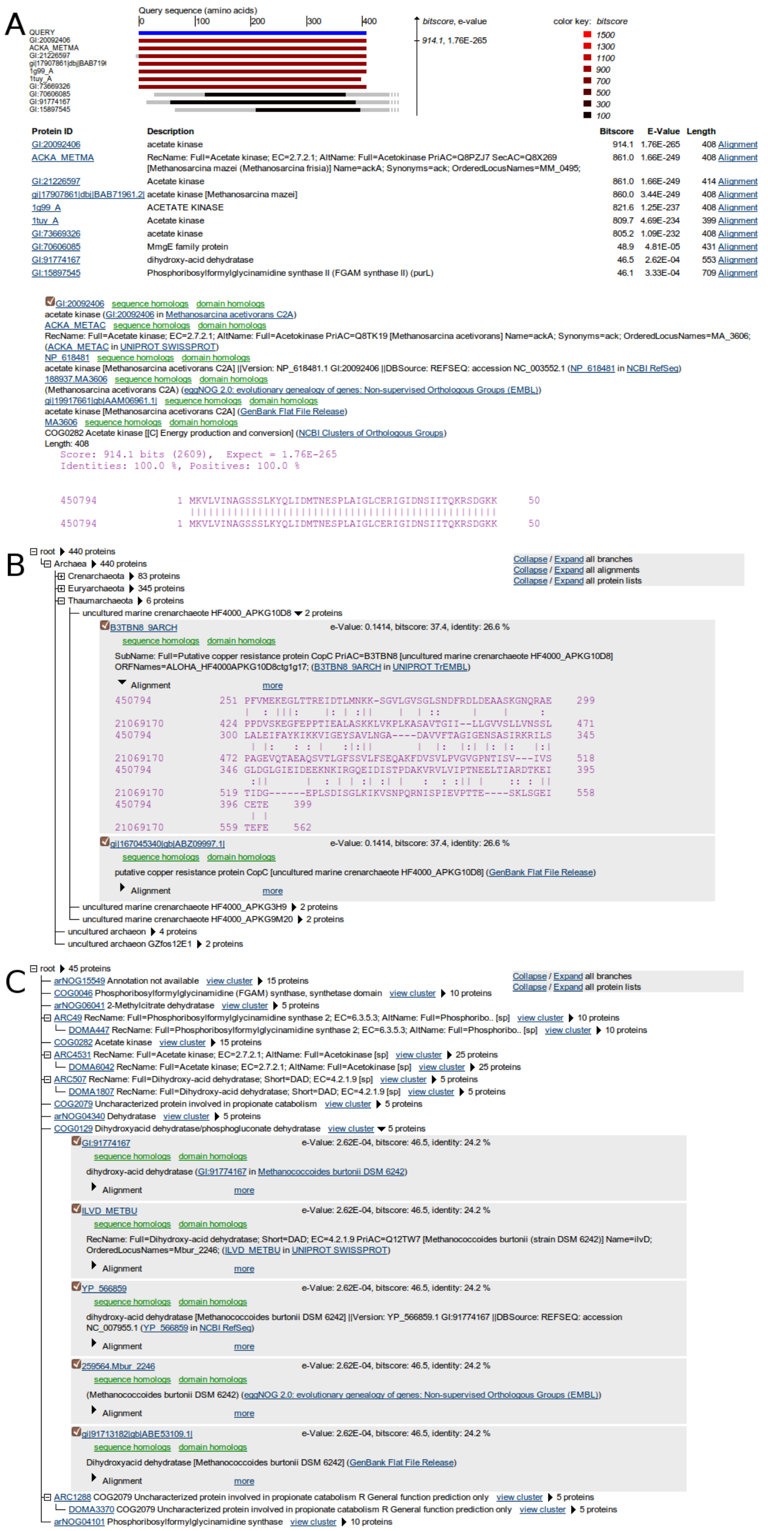
Chlamydophila pneumoniae TW-183 ▼ 4 hits

GI:33241416 (length: 1252 aa) sequence homologs domain homologs
DNA-directed RNA polymerase subunit beta
(GI:33241416 in Chlamydophila pneumoniae TW-183) Search only in Chlamydophila pneumoniae TW-183

Other proteins having the same sequence:

NP_876357 sequence homologs domain homologs (NP_876357 in NCBI RefSeq) Search only in NCBI RefSeq
182082.CpB0081 sequence homologs domain homologs (eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)) Search only in eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)
gi|33235924|gb|AAP98014.1| sequence homologs domain homologs (GenBank Flat File Release) Search only in GenBank Flat File Release

Figure 2.21: Search result visualizations in the SIMAP portal A: List representation of search results **B:** Taxonomic representation of search results



Protein ID	Description	Bitscore	E-Value	Length
GI:20092406	acetate kinase	914.1	1.76E-265	408 Alignment
ACKA_METMA	RecName: Full=Acetate kinase; EC=2.7.2.1; AltName: Full=Acetokinase PriAC=Q8PZJ7 SecAC=Q8X269 [Methanosarcina mazei (Methanosarcina frisia)] Name=ackA; Synonyms=ack; OrderedLocusNames=MM_0495;	861.0	1.66E-249	408 Alignment
GI:21226597	Acetate kinase	861.0	1.66E-249	414 Alignment
gi|17907861|dbj|BAB71961.2|	acetate kinase [Methanosarcina mazei]	860.0	3.44E-249	408 Alignment
1g99_A	ACETATE KINASE	821.6	1.25E-237	408 Alignment
1tuy_A	Acetate kinase	809.7	4.69E-234	399 Alignment
GI:73669326	acetate kinase	805.2	1.09E-232	408 Alignment
GI:70606085	MmgE family protein	48.9	4.81E-05	431 Alignment
GI:91774167	dihydroxy-acid dehydratase	46.5	2.62E-04	553 Alignment
GI:15897545	Phosphoribosylformylglycinamidine synthase II (FGAM synthase II) (purL)	46.1	3.33E-04	709 Alignment

☑GI:20092406 sequence homologs domain homologs
acetate kinase (GI:20092406 in Methanosarcina acetivorans C2A)
ACKA_METAC sequence homologs domain homologs
RecName: Full=Acetate kinase; EC=2.7.2.1; AltName: Full=Acetokinase PriAC=Q8TK19 [Methanosarcina acetivorans] Name=ackA; Synonyms=ack; OrderedLocusNames=MA_3606; (ACKA_METAC in UNIPROT SWISSPROT)
NP_618481 sequence homologs domain homologs
acetate kinase [Methanosarcina acetivorans C2A] ||Version: NP_618481.1 GI:20092406 ||DBSource: REFSEQ: accession NC_003552.1 (NP_618481 in NCBI RefSeq)
188937.MA3606 sequence homologs domain homologs
(Methanosarcina acetivorans C2A) (eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL))
gi|19917661|gb|AAM06961.1| sequence homologs domain homologs
acetate kinase [Methanosarcina acetivorans C2A] (GenBank Flat File Release)
MA3606 sequence homologs domain homologs
COG0282 Acetate kinase [[C] Energy production and conversion] (NCBI Clusters of Orthologous Groups)
Length: 408

```
   Score: 914.1 bits (2609),  Expect = 1.76E-265
   Identities: 100.0 %, Positives: 100.0 %


   450794            1 MKVLVINAGSSSLKYQLIDMTNESPLAIGLCERIGIDNSIITQKRSDGKK     50
                         ||||||||||||||||||||||||||||||||||||||||||||||||||
   450794            1 MKVLVINAGSSSLKYQLIDMTNESPLAIGLCERIGIDNSIITQKRSDGKK     50
```

⊟ root ▶ 440 proteins
└⊟ Archaea ▶ 440 proteins
⊞ Crenarchaeota ▶ 83 proteins
⊞ Euryarchaeota ▶ 345 proteins
⊟ Thaumarchaeota ▶ 6 proteins
— uncultured marine crenarchaeote HF4000_APKG10D8 ▼ 2 proteins

Collapse / Expand all branches
Collapse / Expand all alignments
Collapse / Expand all protein lists

☑B3TBN8_9ARCH e-Value: 0.1414, bitscore: 37.4, identity: 26.6 %
sequence homologs domain homologs
SubName: Full=Putative copper resistance protein CopC PriAC=B3TBN8 [uncultured marine crenarchaeote HF4000_APKG10D8] ORFNames=ALOHA_HF4000APKG10D8ctg1g17; (B3TBN8_9ARCH in UNIPROT TrEMBL)
▼ Alignment more

```
450794          251 PFVMEKEGLTTREIDTLMNKK-SGVLGVSGLSNDFRDLDEAASKGNQRAE    299
                    |   :  |||:     |:  |  :||     |   |  ::      |     |       :
21069170        424 PPDVSKEGFEPPTIEALASKKLVKPLKASAVTGII--LLGVVSLLVNSSL    471
450794          300 LALEIFAYKIKKVIGEYSAVLNGA----DAVVFTAGIGENSASIRKRILS    345
                     | |:    :  :  |    :|:||         |   |:   |:|  |:  |:    |:|
21069170        472 PAGEVQTAEAQSVTLGFSSVLFSEQAKFDVSVLPVGVGPNTISV---IVS    518
450794          346 GLDGLGIEIDEEKNKIRGQEIDISTPDAKVRVLVIPTNEELTIARDTKEI    395
                     :||       |   :  |  |  :|  :|  |    :   : :||  |     ::  :  ||
21069170        519 TIDG------EPLSDISGLKIKVSNPQRNISPIEVPTTE----SKLSGEI    558
450794          396 CETE    399
                     | |
21069170        559 TEFE    562
```

☑gi|167045340|gb|ABZ09997.1| e-Value: 0.1414, bitscore: 37.4, identity: 26.6 %
sequence homologs domain homologs
putative copper resistance protein CopC [uncultured marine crenarchaeote HF4000_APKG10D8] (GenBank Flat File Release)
▶ Alignment more

— uncultured marine crenarchaeote HF4000_APKG3H9 ▶ 2 proteins
└ uncultured marine crenarchaeote HF4000_APKG9M20 ▶ 2 proteins
— uncultured archaeon ▶ 4 proteins
└ uncultured archaeon GZfos12E1 ▶ 2 proteins

⊟ root ▶ 45 proteins
— arNOG15549 Annotation not available view cluster ▶ 15 proteins
— COG0046 Phosphoribosylformylglycinamidine (FGAM) synthase, synthetase domain view cluster ▶ 10 proteins
— arNOG06041 2-Methylcitrate dehydratase view cluster ▶ 5 proteins
⊟ ARC49 RecName: Full=Phosphoribosylformylglycinamidine synthase 2; EC=6.3.5.3; AltName: Full=Phosphoribo.. [sp] view cluster ▶ 10 proteins
└ DOMA447 RecName: Full=Phosphoribosylformylglycinamidine synthase 2; EC=6.3.5.3; AltName: Full=Phosphoribo.. [sp] view cluster ▶ 10 proteins
— COG0282 Acetate kinase view cluster ▶ 15 proteins
⊟ ARC4531 RecName: Full=Acetate kinase; EC=2.7.2.1; AltName: Full=Acetokinase [sp] view cluster ▶ 25 proteins
└ DOMA6042 RecName: Full=Acetate kinase; EC=2.7.2.1; AltName: Full=Acetokinase [sp] view cluster ▶ 25 proteins
⊟ ARC507 RecName: Full=Dihydroxy-acid dehydratase; Short=DAD; EC=4.2.1.9 [sp] view cluster ▶ 5 proteins
└ DOMA1807 RecName: Full=Dihydroxy-acid dehydratase; Short=DAD; EC=4.2.1.9 [sp] view cluster ▶ 5 proteins
— COG2079 Uncharacterized protein involved in propionate catabolism view cluster ▶ 5 proteins
— arNOG04340 Dehydratase view cluster ▶ 5 proteins
— COG0129 Dihydroxyacid dehydratase/phosphogluconate dehydratase view cluster ▼ 5 proteins

Collapse / Expand all branches
Collapse / Expand all protein lists

☑GI:91774167 e-Value: 2.62E-04, bitscore: 46.5, identity: 24.2 %
sequence homologs domain homologs
dihydroxy-acid dehydratase (GI:91774167 in Methanococcoides burtonii DSM 6242)
▶ Alignment more

☑ILVD_METBU e-Value: 2.62E-04, bitscore: 46.5, identity: 24.2 %
sequence homologs domain homologs
RecName: Full=Dihydroxy-acid dehydratase; Short=DAD; EC=4.2.1.9 PriAC=Q12TW7 [Methanococcoides burtonii (strain DSM 6242)] Name=ilvD; OrderedLocusNames=Mbur_2246; (ILVD_METBU in UNIPROT SWISSPROT)
▶ Alignment more

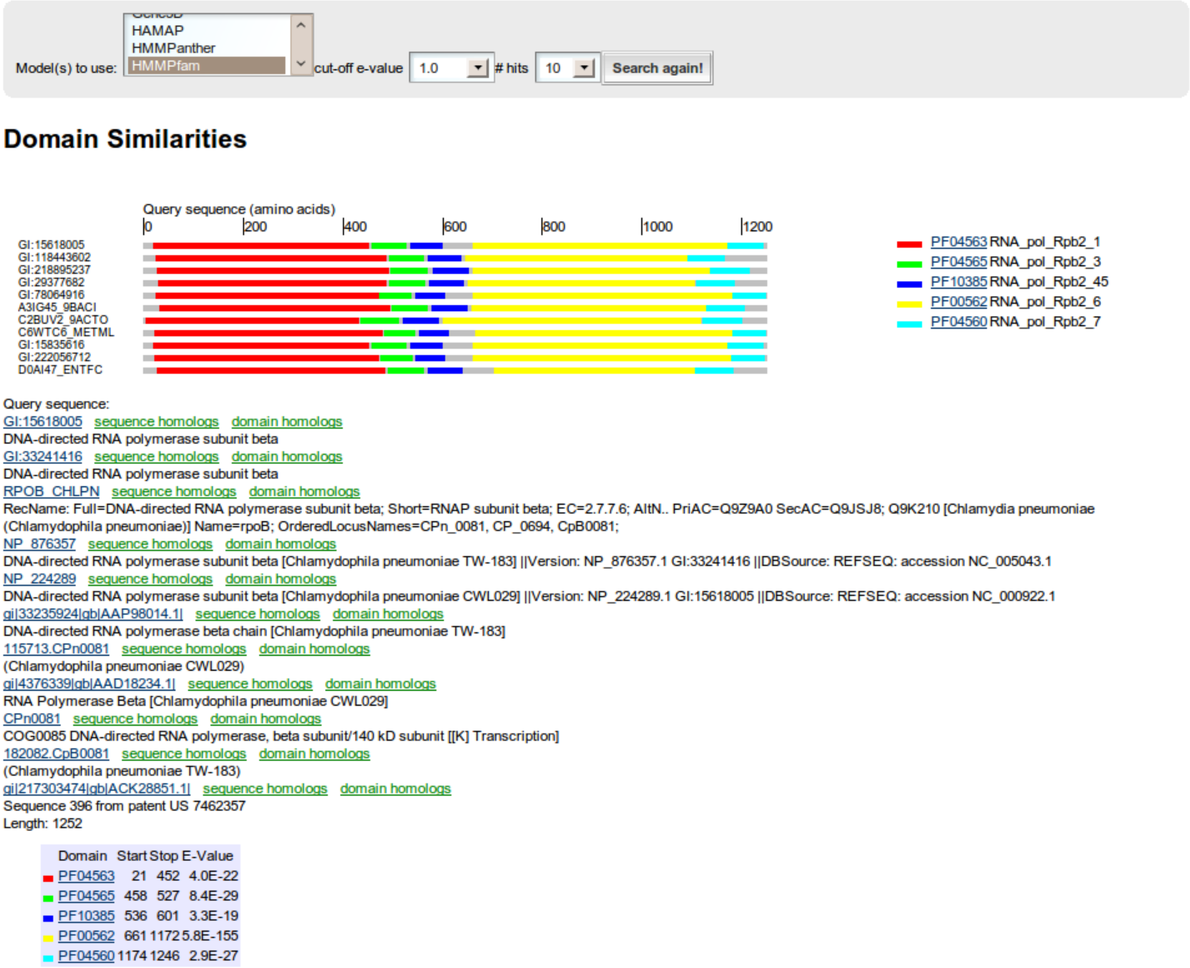
☑YP_566859 e-Value: 2.62E-04, bitscore: 46.5, identity: 24.2 %
sequence homologs domain homologs
dihydroxy-acid dehydratase [Methanococcoides burtonii DSM 6242] ||Version: YP_566859.1 GI:91774167 ||DBSource: REFSEQ: accession NC_007955.1 (YP_566859 in NCBI RefSeq)
▶ Alignment more

☑259564.Mbur_2246 e-Value: 2.62E-04, bitscore: 46.5, identity: 24.2 %
sequence homologs domain homologs
(Methanococcoides burtonii DSM 6242) (eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL))
▶ Alignment more

☑gi|91713182|gb|ABE53109.1| e-Value: 2.62E-04, bitscore: 46.5, identity: 24.2 %
sequence homologs domain homologs
Dihydroxyacid dehydratase [Methanococcoides burtonii DSM 6242] (GenBank Flat File Release)
▶ Alignment more

⊟ ARC1288 COG2079 Uncharacterized protein involved in propionate catabolism R General function prediction only view cluster ▶ 5 proteins
└ DOMA3370 COG2079 Uncharacterized protein involved in propionate catabolism R General function prediction only view cluster ▶ 5 proteins
└ arNOG04101 Phosphoribosylformylglycinamidine synthase view cluster ▶ 10 proteins

Figure 2.22: Sequence homologs visualizations in the SIMAP portal A: NCBI BLAST like representation of sequence homologs **B:** Taxonomic representation of sequence homologs **C:** Assignment of sequence homologs to sequence clusters representation



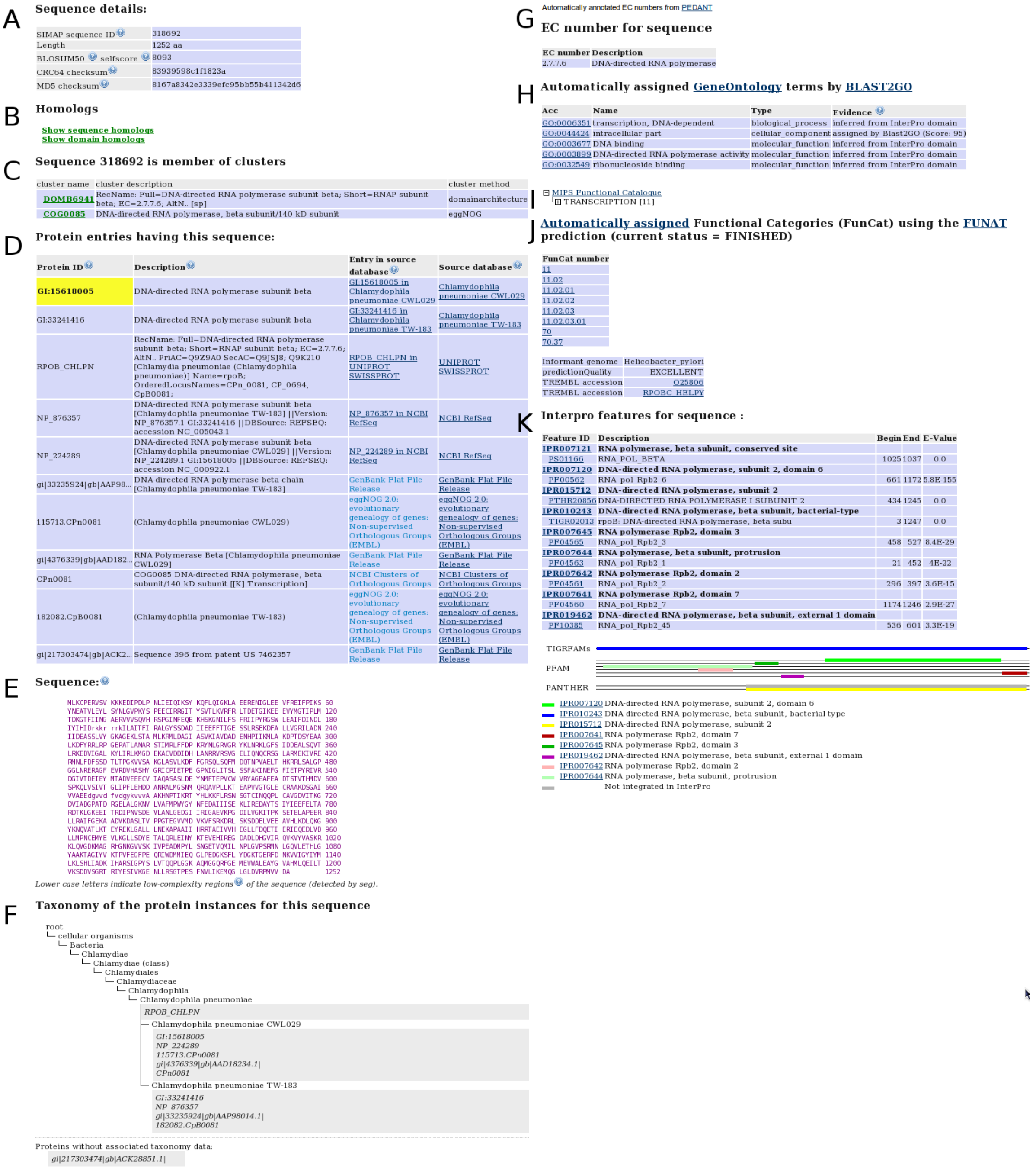
Model(s) to use: HMMPfam cut-off e-value 1.0 # hits 10

Domain Similarities

Query sequence:
GI:15618005 sequence homologs domain homologs
DNA-directed RNA polymerase subunit beta
GI:33241416 sequence homologs domain homologs
DNA-directed RNA polymerase subunit beta
RPOB_CHLPN sequence homologs domain homologs
RecName: Full=DNA-directed RNA polymerase subunit beta; Short=RNAP subunit beta; EC=2.7.7.6; AltN.. PriAC=Q9Z9A0 SecAC=Q9JSJ8; Q9K210 [Chlamydia pneumoniae (Chlamydophila pneumoniae)] Name=rpoB; OrderedLocusNames=CPn_0081, CP_0694, CpB0081;
NP_876357 sequence homologs domain homologs
DNA-directed RNA polymerase subunit beta [Chlamydophila pneumoniae TW-183] ||Version: NP_876357.1 GI:33241416 ||DBSource: REFSEQ: accession NC_005043.1
NP_224289 sequence homologs domain homologs
DNA-directed RNA polymerase subunit beta [Chlamydophila pneumoniae CWL029] ||Version: NP_224289.1 GI:15618005 ||DBSource: REFSEQ: accession NC_000922.1
gi|33235924|gb|AAP98014.1| sequence homologs domain homologs
DNA-directed RNA polymerase beta chain [Chlamydophila pneumoniae TW-183]
115713.CPn0081 sequence homologs domain homologs
(Chlamydophila pneumoniae CWL029)
gi|4376339|gb|AAD18234.1| sequence homologs domain homologs
RNA Polymerase Beta [Chlamydophila pneumoniae CWL029]
CPn0081 sequence homologs domain homologs
COG0085 DNA-directed RNA polymerase, beta subunit/140 kD subunit [[K] Transcription]
182082.CpB0081 sequence homologs domain homologs
(Chlamydophila pneumoniae TW-183)
gi|217303474|gb|ACK28851.1| sequence homologs domain homologs
Sequence 396 from patent US 7462357
Length: 1252

	Domain	Start	Stop	E-Value
🟥	PF04563	21	452	4.0E-22
🟩	PF04565	458	527	8.4E-29
🟦	PF10385	536	601	3.3E-19
🟨	PF00562	661	1172	5.8E-155
🟦	PF04560	1174	1246	2.9E-27

Figure 2.23: Domain homologs visualization in the SIMAP portal The various PFAM domains are visualized by colored bars on the long gray protein sequence lines. There is a color legend on the right side. It can be seen that the next similar sequences on domain level carry the same protein domains in the same number and even the same order.



Sequence details:

SIMAP sequence ID	318692
Length	1252 aa
BLOSUM50 selfscore	8093
CRC64 checksum	83939598c1f1823a
MD5 checksum	8167a8342e3339efc95bb55b411342d6

Homologs

Show sequence homologs
Show domain homologs

Sequence 318692 is member of clusters

cluster name	cluster description	cluster method
DOMB6941	RecName: Full=DNA-directed RNA polymerase subunit beta; Short=RNAP subunit beta; EC=2.7.7.6; AltN.. [sp]	domainarchitecture
COG0085	DNA-directed RNA polymerase, beta subunit/140 kD subunit	eggNOG

Protein entries having this sequence:

Protein ID	Description	Entry in source database	Source database
GI:15618005	DNA-directed RNA polymerase subunit beta	GI:15618005 in Chlamydophila pneumoniae CWL029	Chlamydophila pneumoniae CWL029
GI:33241416	DNA-directed RNA polymerase subunit beta	GI:33241416 in Chlamydophila pneumoniae TW-183	Chlamydophila pneumoniae TW-183
RPOB_CHLPN	RecName: Full=DNA-directed RNA polymerase subunit beta; Short=RNAP subunit beta; EC=2.7.7.6; AltN.. PriAC=Q9Z9A0 SecAC=Q9JSJ8; Q9K210 [Chlamydia pneumoniae (Chlamydophila pneumoniae)] Name=rpoB; OrderedLocusNames=CPn_0081, CP_0694, CpB0081;	RPOB_CHLPN in UNIPROT SWISSPROT	UNIPROT SWISSPROT
NP_876357	DNA-directed RNA polymerase subunit beta [Chlamydophila pneumoniae TW-183] ||Version: NP_876357.1 GI:33241416 ||DBSource: REFSEQ: accession NC_005043.1	NP_876357 in NCBI RefSeq	NCBI RefSeq
NP_224289	DNA-directed RNA polymerase subunit beta [Chlamydophila pneumoniae CWL029] ||Version: NP_224289.1 GI:15618005 ||DBSource: REFSEQ: accession NC_000922.1	NP_224289 in NCBI RefSeq	NCBI RefSeq
gi|33235924|gb|AAP98...	DNA-directed RNA polymerase beta chain [Chlamydophila pneumoniae TW-183]	GenBank Flat File Release	GenBank Flat File Release
115713.CPn0081	(Chlamydophila pneumoniae CWL029)	eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)	eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)
gi|4376339|gb|AAD182...	RNA Polymerase Beta [Chlamydophila pneumoniae CWL029]	GenBank Flat File Release	GenBank Flat File Release
CPn0081	COG0085 DNA-directed RNA polymerase, beta subunit/140 kD subunit [[K] Transcription]	NCBI Clusters of Orthologous Groups	NCBI Clusters of Orthologous Groups
182082.CpB0081	(Chlamydophila pneumoniae TW-183)	eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)	eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)
gi|217303474|gb|ACK2...	Sequence 396 from patent US 7462357	GenBank Flat File Release	GenBank Flat File Release

Sequence:

```
MLKCPERVSV KKKEDIPDLP NLIEIQIKSY KQFLQIGKLA EERENIGLEE VFREIFPIKS 60
YNEATVLEYL SYNLGVPKYS PEECIRRGIT YSVTLKVRFR LTDETGIKEE EVYMGTIPLM 120
TDKGTFIING AERVVVSQVH RSPGINFEQE KHSKGNILFS FRIIPYRGSW LEAIFDINDL 180
IYIHIDrkkr rrkILAITFI RALGYSSDAD IIEEFFTIGE SSLRSEKDFA LLVGRILADN 240
IIDEASSLVY GKAGEKLSTA MLKRMLDAGI ASVKIAVDAD ENHPIIKMLA KDPTDSYEAA 300
LKDFYRRLRP GEPATLANAR STIMRLFFDP KRYNLGRVGR YKLNRKLGFS IDDEALSQVT 360
LRKEDVIGAL KYLIRLKMGD EKACVDDIDH LANRRVRSVG ELIQNQCRSG LARMEKIVRE 420
RMNLFDFSSD TLTPGKVVSA KGLASVLKDF FGRSQLSQFM DQTNPVAELT HKRRLSALGP 480
GGLNRERAGF EVRDVHASHY GRICPIETPE GPNIGLITSL SSFAKINEFG FIETPYRIVR 540
DGIVTDEIEY MTADVEEECV IAQASASLDE YNMFTEPVCW VRYAGEAFEA DTSTVTHMDV 600
SPKQLVSIVT GLIPFLEHDD ANRALMGSNM QRQAVPLLKT EAPVVGTGLE CRAAKDSGAI 660
VVAEEdgvvd fvdgykvvvA AKHNPTIKRT YHLKKFLRSN SGTCINQQPL CAVGDVITKG 720
DVIADGPATD RGELALGKNV LVAFMPWYGY NFEDAIIISE KLIREDAYTS IYIEEFELTA 780
RDTKLGKEEI TRDIPNVSDE VLANLGEDGI IRIGAEVKPG DILVGKITPK SETELAPEER 840
LLRAIFGEKA ADVKDASLTV PPGTEGVVMD VKVFSRKDRL SKSDDELVEE AVHLKDLQKG 900
YKNQVATLKT EYREKLGALL LNEKAPAAII HRRTAEIVVH EGLLFDQETI ERIEQEDLVD 960
LLMPNCEMYE VLKGLLSDYE TALQRLEINY KTEVEHIREG DADLDHGVIR QVKVYVASKR 1020
KLQVGDKMAG RHGNKGVVSK IVPEADMPYL SNGETVQMIL NPLGVPSRMN LGQVLETHLG 1080
YAAKTAGIYV KTPVFEGFPE QRIWDMMIEQ GLPEDGKSFL YDGKTGERFD NKVVIGYIYM 1140
LKLSHLIADK IHARSIGPYS LVTQQPLGGK AQMGGQRFGE MEVWALEAYG VAHMLQEILT 1200
VKSDDVSGRT RIYESIVKGE NLLRSGTPES FNVLIKEMQG LGLDVRPMVV DA         1252
```

Lower case letters indicate low-complexity regions of the sequence (detected by seg).

Taxonomy of the protein instances for this sequence

```
root
└─ cellular organisms
   └─ Bacteria
      └─ Chlamydiae
         └─ Chlamydiae (class)
            └─ Chlamydiales
               └─ Chlamydiaceae
                  └─ Chlamydophila
                     └─ Chlamydophila pneumoniae
                        │ RPOB_CHLPN
                        ├─ Chlamydophila pneumoniae CWL029
                        │     GI:15618005
                        │     NP_224289
                        │     115713.CPn0081
                        │     gi|4376339|gb|AAD18234.1|
                        │     CPn0081
                        └─ Chlamydophila pneumoniae TW-183
                              GI:33241416
                              NP_876357
                              gi|33235924|gb|AAP98014.1|
                              182082.CpB0081
```

Proteins without associated taxonomy data:
gi|217303474|gb|ACK28851.1|

Automatically annotated EC numbers from PEDANT

EC number for sequence

EC number	Description
2.7.7.6	DNA-directed RNA polymerase

Automatically assigned GeneOntology terms by BLAST2GO

Acc	Name	Type	Evidence
GO:0006351	transcription, DNA-dependent	biological_process	inferred from InterPro domain
GO:0044424	intracellular part	cellular_component	assigned by Blast2GO (Score: 95)
GO:0003677	DNA binding	molecular_function	inferred from InterPro domain
GO:0003899	DNA-directed RNA polymerase activity	molecular_function	inferred from InterPro domain
GO:0032549	ribonucleoside binding	molecular_function	inferred from InterPro domain

⊟ MIPS Functional Catalogue
 └⊞ TRANSCRIPTION [11]

Automatically assigned Functional Categories (FunCat) using the FUNAT prediction (current status = FINISHED)

FunCat number
11
11.02
11.02.01
11.02.02
11.02.03
11.02.03.01
70
70.37

Informant genome	Helicobacter_pylori
predictionQuality	EXCELLENT
TREMBL accession	O25806
TREMBL accession	RPOBC_HELPY

Interpro features for sequence :

Feature ID	Description	Begin	End	E-Value
IPR007121	**RNA polymerase, beta subunit, conserved site**			
PS01166	RNA_POL_BETA	1025	1037	0.0
IPR007120	**DNA-directed RNA polymerase, subunit 2, domain 6**			
PF00562	RNA_pol_Rpb2_6	661	1172	5.8E-155
IPR015712	**DNA-directed RNA polymerase, subunit 2**			
PTHR20856	DNA-DIRECTED RNA POLYMERASE I SUBUNIT 2	434	1245	0.0
IPR010243	**DNA-directed RNA polymerase, beta subunit, bacterial-type**			
TIGR02013	rpoB: DNA-directed RNA polymerase, beta subu	3	1247	0.0
IPR007645	**RNA polymerase Rpb2, domain 3**			
PF04565	RNA_pol_Rpb2_3	458	527	8.4E-29
IPR007644	**RNA polymerase, beta subunit, protrusion**			
PF04563	RNA_pol_Rpb2_1	21	452	4E-22
IPR007642	**RNA polymerase Rpb2, domain 2**			
PF04561	RNA_pol_Rpb2_2	296	397	3.6E-15
IPR007641	**RNA polymerase Rpb2, domain 7**			
PF04560	RNA_pol_Rpb2_7	1174	1246	2.9E-27
IPR019462	**DNA-directed RNA polymerase, beta subunit, external 1 domain**			
PF10385	RNA_pol_Rpb2_45	536	601	3.3E-19

- ▬ IPR007120 DNA-directed RNA polymerase, subunit 2, domain 6
- ▬ IPR010243 DNA-directed RNA polymerase, beta subunit, bacterial-type
- ▬ IPR015712 DNA-directed RNA polymerase, subunit 2
- ▬ IPR007641 RNA polymerase Rpb2, domain 7
- ▬ IPR007645 RNA polymerase Rpb2, domain 3
- ▬ IPR019462 DNA-directed RNA polymerase, beta subunit, external 1 domain
- ▬ IPR007642 RNA polymerase Rpb2, domain 2
- ▬ IPR007644 RNA polymerase, beta subunit, protrusion
- ▬ Not integrated in InterPro

Figure 2.24: Protein report in the SIMAP portal A: Basic information about the sequence **B:** Direct links to sequence and domain homologs of this sequence **C:** Link to clusters this sequence is assigned to **D:** List of protein instances having the same sequence as the selected protein. The selected protein is highlighted yellow. **E:** Amino acid sequence of the protein **F:** Taxonomy of proteins having the same sequence **G:** Automatically annotated EC numbers from PEDANT **H:** Automatically annotated GO annotations from Blast2GO **I:** Automatically annotated FunCats from PEDANT **J:** Automatically annotated FunCats from FUNAT **K:** InterPro protein domains of this sequence

2.3.6 Implementation of ChlamydiaeDB

As the utilization of portlets on the portal server Liferay for the SIMAP database was successful it was also applied for the new resource for *Chlamydiae*.

There are many different kinds of information integrated into ChlamydiaeDB in different ways. The following sections will provide an overview over the integrated information and show how data is preprocessed, stored and how it can be retrieved.

The structure of this section is geared to the structure of section 2.3.2 containing the criteria for ChlamydiaeDB.

2.3.6.1 Content

2.3.6.1.1 Data storage in a data warehouse In contrast to SIMAP there is specific information that needs to be stored for ChlamydiaeDB, for example Type-III secretion predictions.

Some information is stored redundantly within a general data scheme that is different from the normalized data scheme. This is done as this pays retrieval speed by avoiding expensive joins between tables. Therefore the ChlamydiaeDB database is also a data warehouse.

2.3.6.1.2 All publicly available data for all chlamydial genomes in one place The demand to keep all publicly available data for all chlamydial genomes in one place requires the integration of external information sources on the one hand and specific developments, so that data becomes easily accessible, on the other hand.

2.3.6.1.2.1 Retrieval of synonymous identifiers One of the key requirements for a comprehensive resource for all *Chlamydiae* is the possibility to be able to handle all kinds of identifiers, in order to be able to search for these names in the literature, for example. Therefore a mapping procedure allowing to retrieve, store and use all synonymous names for the genetic elements was developed.

First the chlamydiae relevant entries from RefSeq [52] are extracted. The entries can be easily identified by their NCBI taxonomy ids, belonging to the phylum *Chlamydiae*. The following information is extracted for each chlamydial protein: The RefSeq version which is identical to the PEDANT identifier, the RefSeq accession, the gene names and the locus_tag. Gene names and locus_tag are extracted from RefSeq and not from UniProtKB as an entry in UniProtKB does not necessarily refer to only one protein of one chlamydial strain so that the entry may contain locus_tags from various *Chlamydiae*. Then a mapping between RefSeq protein accessions and UniProtKB protein accessions provided within RefSeq is used to extract entries from UniProtKB/Swiss-Prot and UniProtKB/TrEMBL containing information relevant for chlamydial proteins. The following information is extracted from the entries: The UniProtKB accession, the UniProtKB name, the UniProtKB protein full name, and the UniProtKB protein short name. An example for the synonymous names for a genetic element can be seen in Table 2.15.

name	kind of name	source
GI:166154924	PEDANT code / RefSeq version	PEDANT / RefSeq
YP_001653179	RefSeq accession	RefSeq
tig	gene name	RefSeq
CTLon_0076	locus_tag	RefSeq
B0BAG2	UniProtKB/Swiss-Prot accession	UniProtKB/Swiss-Prot
TIG_CHLTB	UniProtKB/Swiss-Prot name	UniProtKB/Swiss-Prot
Trigger factor	UniProtKB/Swiss-Prot full name	UniProtKB/Swiss-Prot
TF	UniProtKB/Swiss-Prot short name	UniProtKB/Swiss-Prot

Table 2.15: Synonymous names for UniProtKB/Swiss-Prot entry B0BAG2 Gene name and locus_tag are taken from RefSeq as UniProtKB/Swiss-Prot entries do not necessarily refer to a single strain and therefore may contain several locus_tags.

The information about the genetic elements is stored in the geneticelements table in the data warehouse. Additionally to a unique elementid each entry is characterized by the fields allowstaxonomicclassification, sourcetypeid, isswissprot, pedantelementid. The field allowstaxonomicclassification is true if the name is unambiguously mappable to a specific gene or protein of one chlamydial strain. This feature plays an important role for literature mining. The sourcetypeid references the table geneticelementsourcetype, which contains information about the source and the type of the elements name. The field isswissprot is true if the UniProtKB information is from UniProtKB/Swiss-Prot, false if it is from UniProtKB/TrEMBL. The pedantelementid is important as this easily allows to retrieve all synonymous elements as they share the same pedantelementid. An overview over the involved tables can be seen in Figure 2.25.

2.3.6.1.2.2 Information from PEDANT Most of the information from PEDANT is retrieved on the fly using the Web Service. Some queries are not efficiently accessible, for example the question how many proteins in a chlamydial species are annotated with annotation X. This aggregated information is cached in the data warehouse.

The PEDANT information that should be mirrored in the data warehouse is easily adjustable in the configuration of the maintenance program. The necessary tables are automatically created and filled with content. At the moment FunCat annotations, EC annotations and UniProtKB/Swiss-Prot keywords are cached in the ChlamydiaeDB.

For each kind of annotation two tables are created, one with the annotation itself, e.g. table funcat containing as annotation FunCat “01“, and one with the description of the annotation, e.g. table funcatdescription with description “metabolism“ for FunCat 01. Figure 2.26 shows the table structure for PEDANT information for the example FunCat. Funcat contains the annotations for each of the proteins, funcatdescription contains the descriptions for all FunCat categories once.

The PEDANT annotations are displayed in the protein report for each of the proteins and are used for other analyses like the enrichment analysis (see protein report, Figures 2.36 A, C, I, J, 2.37, 2.46, 2.42, 2.43) .

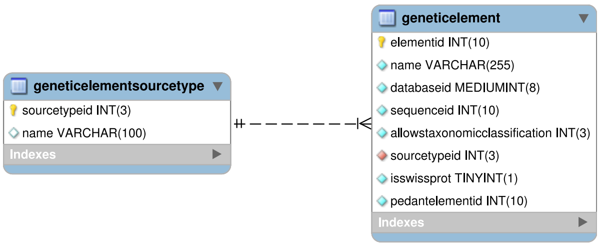


Figure 2.25: Tables involved in storing information about synonymous names within ChlamydiaeDB The table **geneticelementsourcetype** stores information about the kind of name (PEDANT code, RefSeq accession, gene name, locus_tag, UniProtKB accession, UniProtKB name, UniProtKB full name, UniProtKB short name), the table **geneticelement** contains the names and additional information for the genetic element. Each entry is amongst others characterized by the fields allowstaxonomicclassification, sourcetypeid, isswissprot, pedantelementid. The field allowstaxonomicclassification is true if the name is unambiguously mappable to a specific gene or protein of one chlamydial strain. The sourcetypeid references the sourcetypeid of the table geneticelementsourcetype, the field isswissprot is true if the UniProtKB information is from UniProtKB/Swiss-Prot, false if it is from UniProtKB/TrEMBL. The pedantelementid is important as it easily allows to retrieve all synonymous elements as they share the same pedantelementid.

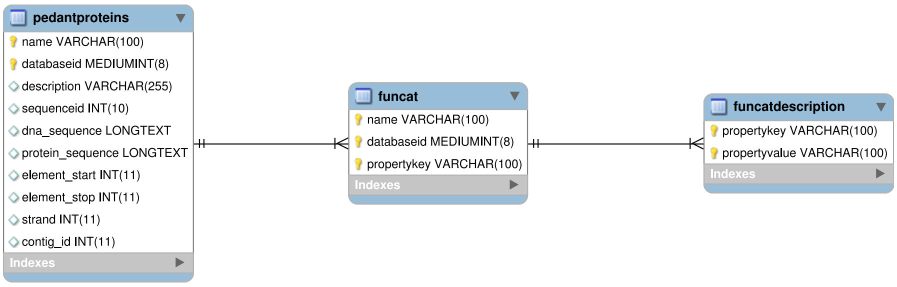


Figure 2.26: Tables involved in storing information mirrored from PEDANT, for the example FunCat The table **pedantproteins** contains all chlamydial proteins within PEDANT. They are uniquely characterized by their name and databaseid. Name and databaseid are referenced from the table **funcat** and **funcatdescription**. **funcat** contains the annotations for each of the proteins, **funcatdescription** contains the descriptions for all FunCat categories.

2.3.6.1.2.3 Information from SIMAP Some information from SIMAP is integrated on the fly using the SIMAP EJBs, for example sequence similarities. Protein domains from InterPro and the associated domain similarities are also integrated on the fly as the application logic for the domain similarities is contained within the SIMAP EJBs.

The assignment of sequences to eggNOG sequence clusters has been cached within ChlamydiaeDB as not all chlamydial genomes are integrated into the sequence clusters. Therefore Thomas Rattei implemented a mapping procedure for the assignment of sequences to the best fitting cluster of a specific clustering approach. As this assignment would have to be done over and over again and it is heavily used within the portal for each sequence (e.g. for the synteny view), it was necessary to cache the information about eggNOG cluster memberships as this makes a more efficient aggregation of

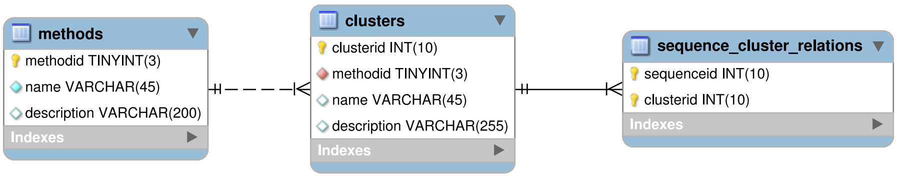


Figure 2.27: Tables involved in storing clusters. methods contains various clustering methods, **clusters** contains information about the specific clusters and **sequence_cluster_relations** contains the assignments of sequences to clusters

knowledge possible. The data warehouse is filled as follows: First for all sequences of all chlamydial organisms within eggNOG, clusters are retrieved and stored within the data warehouse. Then additional cluster assignments for the sequences of organisms not contained in eggNOG are added using the method implemented by Thomas Rattei. The clusters are then stored in the same table structure as in the original simapclusters database (see Figure 2.27).

For the GO annotations an adapted version of Blast2GO [93] for high-throughput SIMAP calculations is used. Firstly a separate EJB was implemented that allowed to retrieve GO annotations for sequences either by the internal SIMAP sequenceid or by the MD5 hash of the protein sequence. As the Blast2GO calculations are connected to the SIMAP releases, the retrieval functionality was transferred into the SIMAP EJBs by Thomas Rattei later.

GO annotations have to be cached within ChlamydiaeDB as typical queries like the number of proteins within an organism having GO X are too time expensive. Therefore an index containing information about the number of occurrences of each GO annotation for each of the chlamydial organisms is stored in the table goindex.

The information from SIMAP is used throughout the whole ChlamydiaeDB, for searches as well as for the protein reports.

2.3.6.1.2.4 Type-III secretion predictions As the Type-III secretion system plays an important role for the pathogenicity of *Chlamydiae* our Type-III secretion predictions [177] for all sequences have been integrated into ChlamydiaeDB.

Firstly also the interactive secretion prediction for new sequences was possible, but in the course of the preparation of the publication of the prediction software the training set as well as the algorithm have been modified and the Effective web portal has been implemented (`http://www.effectors.org`) [179]. This portal is specialized on the prediction of bacterial Type III secreted proteins by their N-terminal sequence but will also offer the predictions of effectors based on eukaryotic domain signatures (see also section 1.6.3). Therefore the interactive prediction was outsourced to effectors.org.

The Type-III secretion predictions are cached within ChlamydiaeDB. The table t3effectorpredictions contains information about whether each of the chlamydial proteins is predicted to be secreted or not. The combined training set with selective settings for the effector predictions are used.

The Type-III secretion predictions are displayed in the protein report (section 2.3.6.2.3.4) and are used for enrichment analyses (section 2.3.6.2.5.2).

2.3.6.1.2.5 Literature mining **Reasons for an own textmining system** As much of biological knowledge is only available in the free text of publications, literature mining is an important field in bioinformatics. There exist very sophisticated solutions for literature mining that are used to analyze the available paper titles, paper abstracts or fulltext papers. As Thorsten Barnickel from the Institute of Bioinformatics and Systems Biology at the Helmholtz Zentrum München was working on his text mining system EXCERBT during his PhD, we discussed whether his system or another existing system would be suited for the aim to extract only entries relevant for *Chlamydiae* from the literature. We discovered that the existing systems do not fit the requirements and would be breaking a fly on the wheel. Therefore a very basic textmining procedure was specifically developed for the ChlamydiaeDB.

Kinds of information that should be identified in literature One problem of datamining in literature is the sheer amount of data that needs to be stored. Therefore I decided to only store the documents relevant for *Chlamydiae* in the database. In order to be able to decide which literature is *Chlamydiae* relevant the following kinds of information should be automatically identified: various gene and protein identifiers (see section 2.3.6.1.2.1), names of chlamydial organisms, e.g. “Chlamydophila pneumoniae” or “C. pneumoniae”, and whitelist names, that is words giving a hint on the relevance for the research field on *Chlamydiae*, e.g. “chlamydiae”, “chlamydial”, “chlamydiales”.

Retrieval of genetic element names It has already been described how the synonymous names are automatically identified (section 2.3.6.1.2.1).

Retrieval of names of chlamydial organisms The names of chlamydial organisms are retrieved automatically as follows. First all organisms belonging to the phylum *Chlamydiae* are extracted from the NCBI taxonomy and are stored in the table db with their “scientific name”. Then the following alternatives to the “scientific name” are extracted and stored in the table “dbnames”: “acronym”, “anamorph”, “common name”, “equivalent name”, “genbank acronym”, “genbank anamorph”, “genbank common name”, “genbank synonym”, “synonym”, “teleomorph”, “misspelling”, “blast name”, “misnomer”, “in-part”.

Retrieval of other names The whitelist has been created manually together with Matthias Horn from the Department of Microbial Ecology at the University of Vienna using terms associated to the phylum *Chlamydiae* and the respective taxonomic families of chlamydiae in the NCBI taxonomy.

Taxonomic and non-taxonomic names There exist taxonomic and non-taxonomic synonymous names for a genetic element. Taxonomic names are names that allow to identify exactly one protein of one strain. Taxonomic names are locus_tags, PEDANT codes and RefSeq accessions. Non-taxonomic names are genenames, UniProtKB accessions, UniProtKB names, UniProtKB full names, and UniProtKB short names that alltogether do not unambiguously refer to only one gene of a specific chlamydial strain. If a taxonomic name is identified in a publication, then the publication is unambiguously *Chlamydiae* related. If a non-taxonomic name is identified in a publication then it is not clear yet whether this is a *Chlamydiae* related article.

Stepwise textmining in the literature As there exist taxonomic and non-taxonomic names I developed a stepwise procedure for the literature mining.

For each title and abstract of every publication the following steps are performed (Figure 2.28):

1. Search for organism names within title and abstract.
2. Search for taxonomic genetic element hits within title and abstract.
3. If no organism name hit and no taxonomic genetic element hits could be identified then search for whitelist hits within title and abstract.
4. If an organism name or a taxonomic genetic element name or a whitelist name has been identified within title and abstract then search for non-taxonomic genetic element hits within title and abstract. This way non-taxonomic genetic elements are only searched within title and abstract when it is already clear that this is a *Chlamydiae* related publication.
5. If an organism name hit or a taxonomic genetic element hit or a whitelist hit has been identified then classify the literature as *Chlamydiae* relevant and store it in the data warehouse of ChlamydiaeDB else reject it.

As of May 18th 2010 MEDLINE/PubMed contained 20082979 documents that needed to be analyzed including updates of articles. In sum there are 121585 names that need to be identified within the literature (68738 non-taxonomic names, 52493 taxonomic names, 333 organism names, 21 whitelist names).

The performance for identifying relevant documents within the literature using string matching or the java.util.regex package with Pattern and Matcher engines was quite slow as it needed about 1 minute for 1000 documents. Therefore HashMap objects are used that store the search terms as keys. The titles and abstracts of each article are broken down into words and it is then checked whether there exists a key that equals one of these words in the HashMaps. This allows to process 1000 articles in about 2.4 seconds.

Storage of literature and hits within the literature The documents are stored separately from the hits. The documents are stored in the table “document”. Every document has a literature datasource. The datasources are stored in the table “documentsource”. At the moment MEDLINE/PubMed is available as datasource but more sources would be possible.

The hits of either genetic elements or organism or whitelist entries in the literature are stored in three tables. The table “document2chlamydiaunspecific” contains hits of whitelist entries in the literature, the table “document2database” contains organism hits in the literature, the table “document2geneticelement” contains genetic element hits in the literature (Figure 2.29).

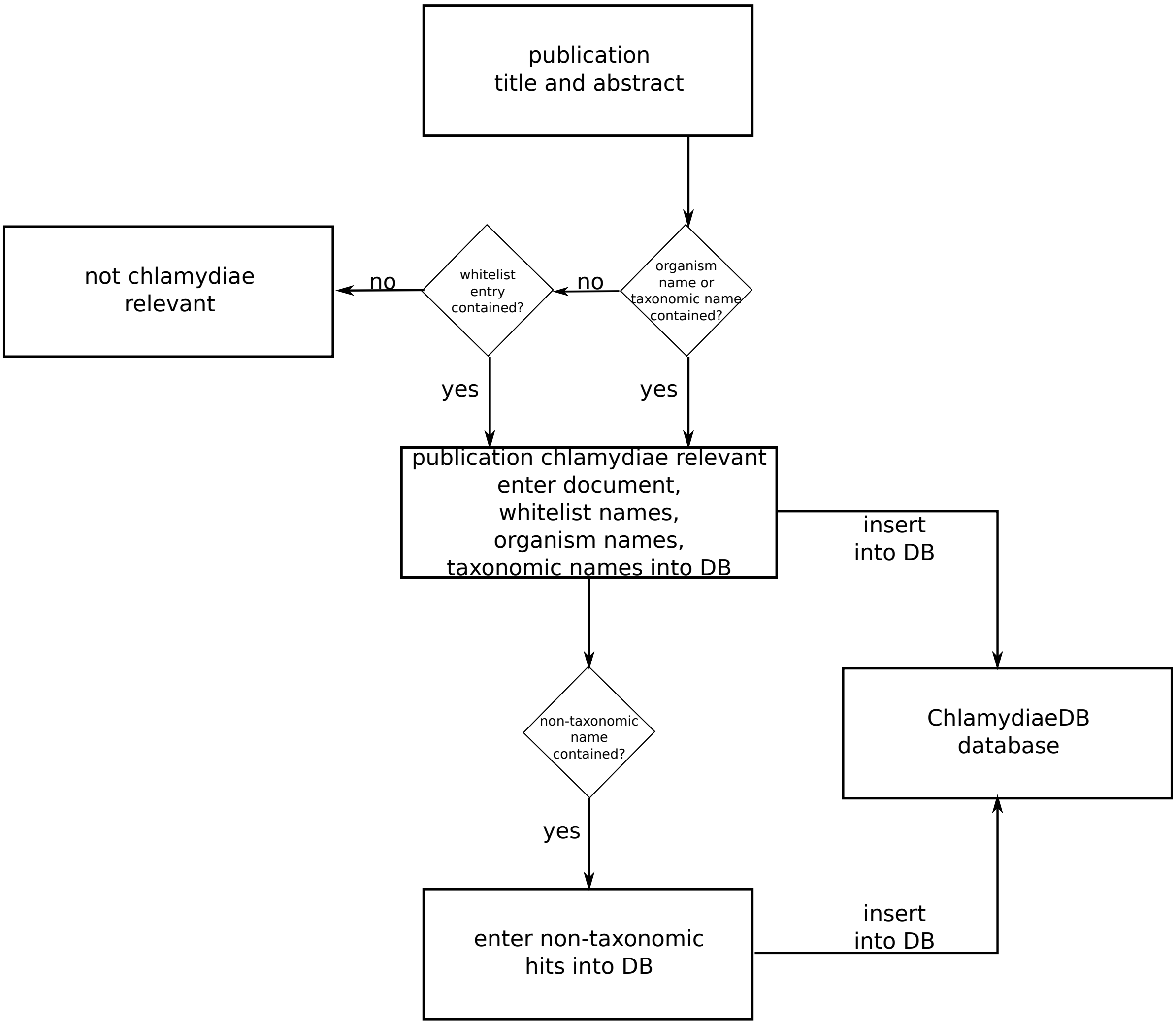


Figure 2.28: Overview over the decision process whether a publication is relevant for *Chlamydiae*

Display of literature information in the ChlamydiaeDB Literature is used in different contexts within the ChlamydiaeDB. It is available for all chlamydial species, namely for specific genetic elements or for specific organism.

The ten most current *Chlamydiae* relevant publications within MEDLINE/PubMed are shown on the start site of ChlamydiaeDB. This allows to stay up to date with the newest literature or to discover new publications (Figure 2.30).

Literature for a specific genetic element is shown in the protein report within the "automatic literature" portlet (section 2.3.6.2.3.2) and in the "information about proteins in the same protein family" portlet (section 2.3.6.2.3.11). It is also used when retrieving aggregated information for a set of proteins (section 2.3.6.2.5.1).

Literature for a specific chlamydial species is displayed on the information page for an organism (Figure 2.31).

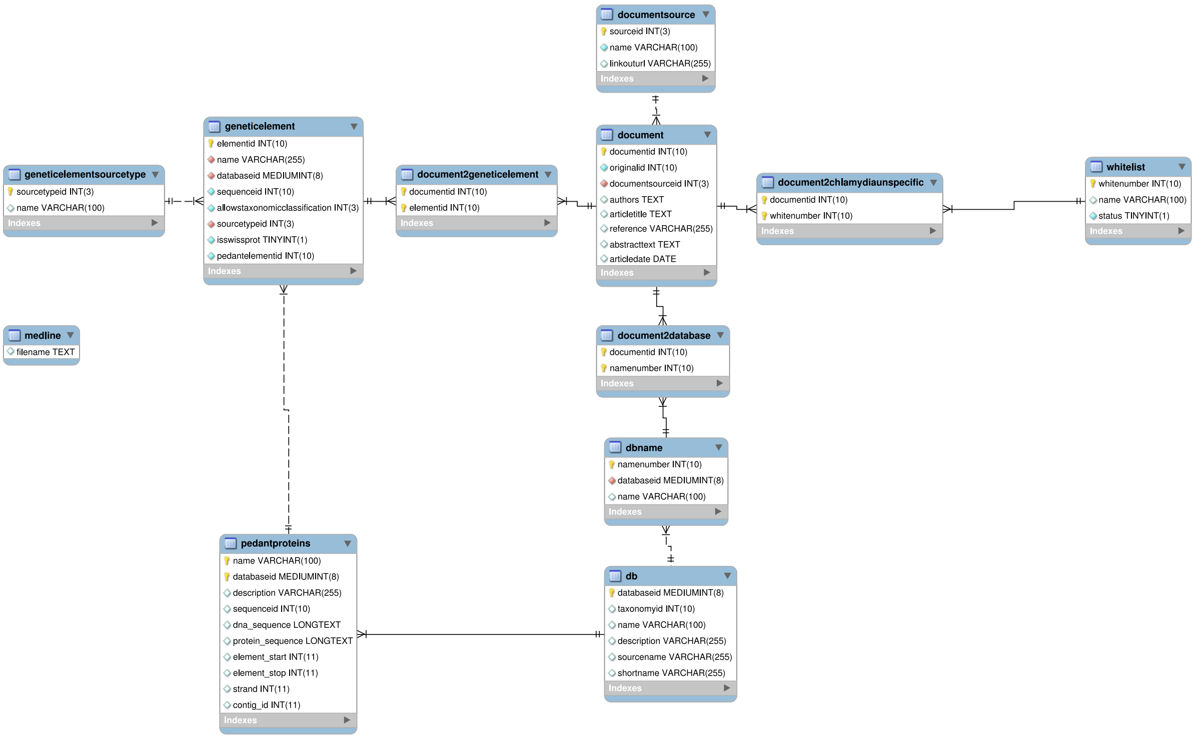
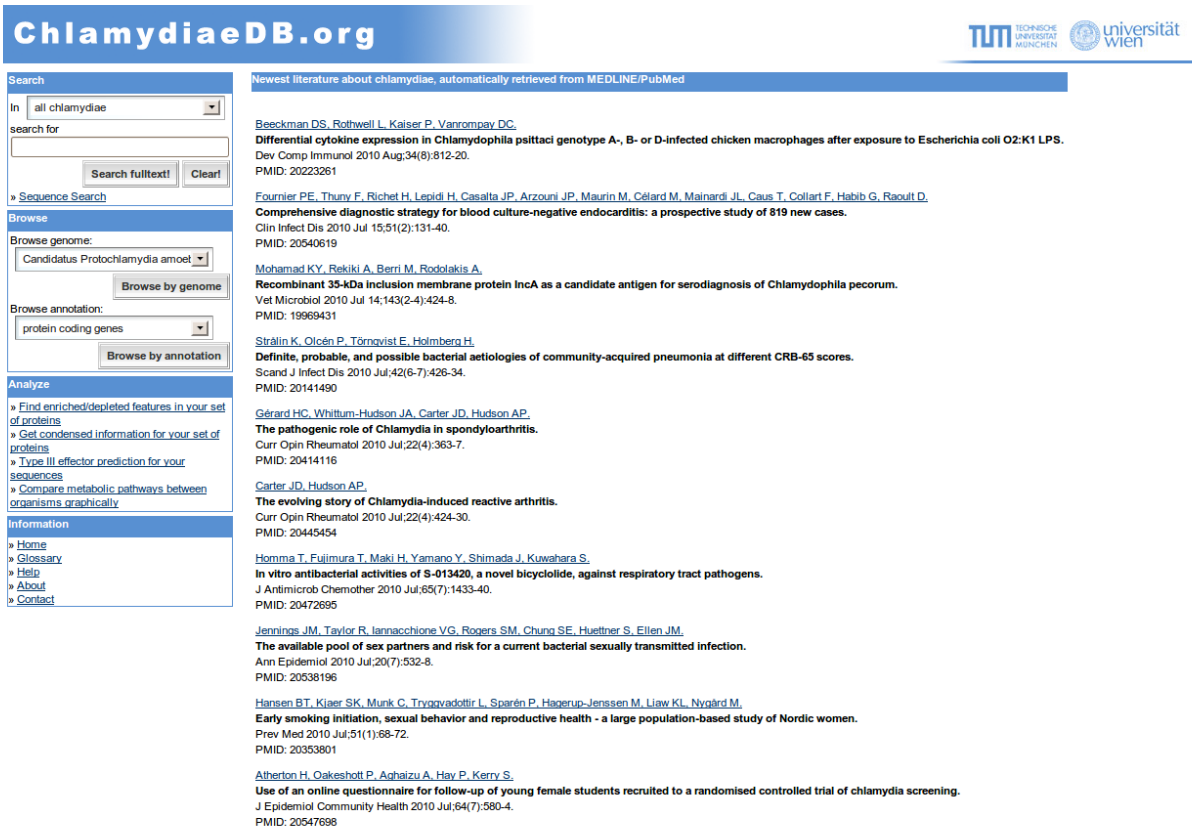


Figure 2.29: Tables involved in storing literature. pedantproteins contains all PEDANT proteins, **geneticelement** contains all PEDANT proteins and all synonyms (see section 2.3.6.1.2.1) and their types are described in **geneticelementsourcetype**. **db** contains all chlamydial organisms and their names and **dbname** contains the additional names for the organisms extracted from the NCBI taxonomy. **document** contains the literature entries from a specific literature datasource, at the moment MEDLINE/PubMed is available. **documentsource** contains the literature datasources so that it would be possible to add literature also from other sources than MEDLINE/PubMed. **whitelist** contains terms that make a literature document chlamydiae specific even if no genetic element name or organism name should be identified in the document. Examples are "chlamydial" or "chlamydiae". The following three tables are for the storage of hits of either genetic elements or organism or whitelist entries in the literature. **document2chlamydiaunspecific** contains hits of whitelist entries in the literature, **document2database** contains organism hits in the literature, **document2geneticelement** contains genetic element hits in the literature. **medline** is used internally in order to administrate already processed XMLs available for download at NCBI.

2.3.6.1.2.6 KEGG pathways In order to get a first overview over the metabolic capabilities of an organism or the metabolic differences between organisms the KEGG pathways annotated with the enzymes available in the respective organisms can be used. KEGG offers the annotations of enzymes and colors them in the pathways for many organisms, but not for all. As the enzymes can be characterized by their EC number the automatically assigned EC numbers from PEDANT can be used to extend the KEGG annotations to all RefSeq genomes within PEDANT. Therefore the goals were to automate the coloring of the KEGG maps also for genomes not already colored


ChlamydiaeDB.org
TECHNISCHE UNIVERSITÄT MÜNCHEN
universität wien
Search
In
all chlamydiae
search for
Search fulltext!
Clear!
» Sequence Search
Browse
Browse genome:
Candidatus Protochlamydia amoet
Browse by genome
Browse annotation:
protein coding genes
Browse by annotation
Analyze
» Find enriched/depleted features in your set of proteins
» Get condensed information for your set of proteins
» Type III effector prediction for your sequences
» Compare metabolic pathways between organisms graphically
Information
» Home
» Glossary
» Help
» About
» Contact

Newest literature about chlamydiae, automatically retrieved from MEDLINE/PubMed

Beeckman DS, Rothwell L, Kaiser P, Vanrompay DC.
Differential cytokine expression in Chlamydophila psittaci genotype A-, B- or D-infected chicken macrophages after exposure to Escherichia coli O2:K1 LPS.
Dev Comp Immunol 2010 Aug;34(8):812-20.
PMID: 20223261

Fournier PE, Thuny F, Richet H, Lepidi H, Casalta JP, Arzouni JP, Maurin M, Célard M, Mainardi JL, Caus T, Collart F, Habib G, Raoult D.
Comprehensive diagnostic strategy for blood culture-negative endocarditis: a prospective study of 819 new cases.
Clin Infect Dis 2010 Jul 15;51(2):131-40.
PMID: 20540619

Mohamad KY, Rekiki A, Berri M, Rodolakis A.
Recombinant 35-kDa inclusion membrane protein IncA as a candidate antigen for serodiagnosis of Chlamydophila pecorum.
Vet Microbiol 2010 Jul 14;143(2-4):424-8.
PMID: 19969431

Strålin K, Olcén P, Törnqvist E, Holmberg H.
Definite, probable, and possible bacterial aetiologies of community-acquired pneumonia at different CRB-65 scores.
Scand J Infect Dis 2010 Jul;42(6-7):426-34.
PMID: 20141490

Gérard HC, Whittum-Hudson JA, Carter JD, Hudson AP.
The pathogenic role of Chlamydia in spondyloarthritis.
Curr Opin Rheumatol 2010 Jul;22(4):363-7.
PMID: 20414116

Carter JD, Hudson AP.
The evolving story of Chlamydia-induced reactive arthritis.
Curr Opin Rheumatol 2010 Jul;22(4):424-30.
PMID: 20445454

Homma T, Fujimura T, Maki H, Yamano Y, Shimada J, Kuwahara S.
In vitro antibacterial activities of S-013420, a novel bicyclolide, against respiratory tract pathogens.
J Antimicrob Chemother 2010 Jul;65(7):1433-40.
PMID: 20472695

Jennings JM, Taylor R, Iannacchione VG, Rogers SM, Chung SE, Huettner S, Ellen JM.
The available pool of sex partners and risk for a current bacterial sexually transmitted infection.
Ann Epidemiol 2010 Jul;20(7):532-8.
PMID: 20538196

Hansen BT, Kjaer SK, Munk C, Tryggvadottir L, Sparén P, Hagerup-Jenssen M, Liaw KL, Nygård M.
Early smoking initiation, sexual behavior and reproductive health - a large population-based study of Nordic women.
Prev Med 2010 Jul;51(1):68-72.
PMID: 20353801

Atherton H, Oakeshott P, Aghaizu A, Hay P, Kerry S.
Use of an online questionnaire for follow-up of young female students recruited to a randomised controlled trial of chlamydia screening.
J Epidemiol Community Health 2010 Jul;64(7):580-4.
PMID: 20547698

Figure 2.30: ChlamydiaeDB start site showing the newest literature relevant for *Chlamydiae*

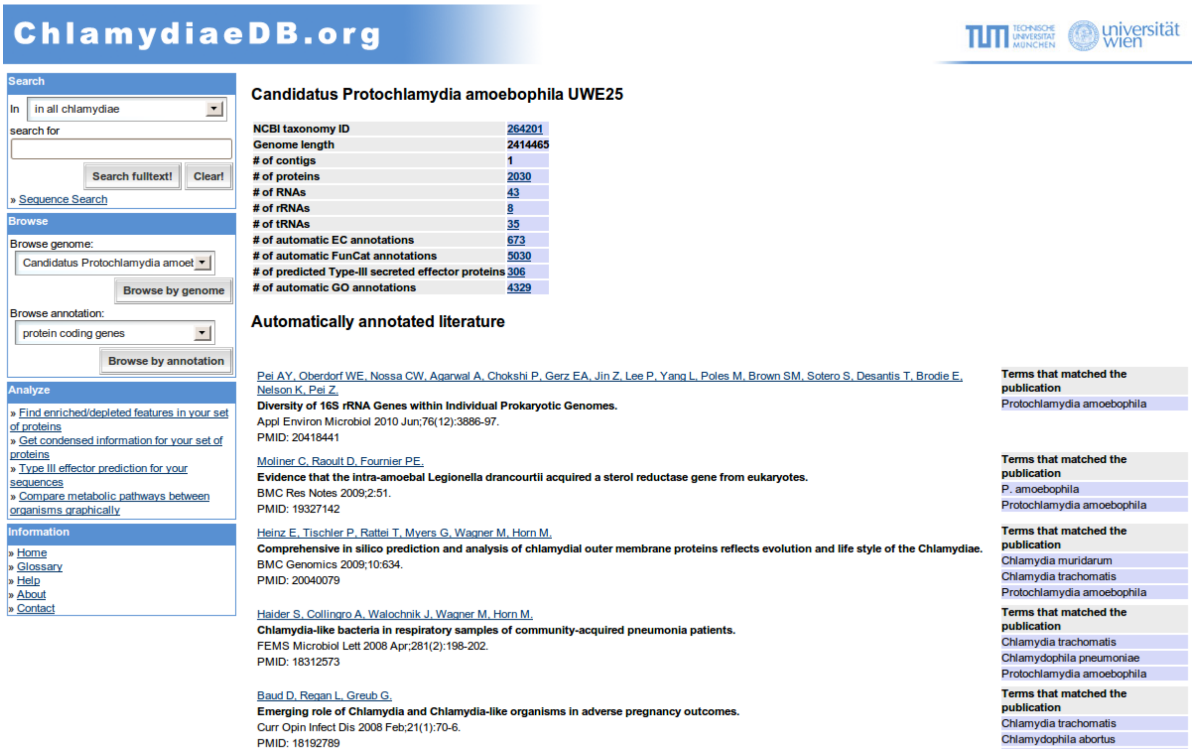
and to even allow to color more than one organism at the same time in order to allow a graphic comparison between various organisms in relation to a specific pathway.

This work was started by Christian Hainzinger during his Diploma thesis, Thomas Weinmaier adapted it to the new version of the KEGG Web Service and Roland Arnold took care of the KEGG data update. I adapted the methods so that they would run within ChlamydiaeDB.

The mapping is done using the EC numbers from PEDANT and the assignments of ECs to KEGG orthologous groups from KEGG. The orthologous groups available within a pathway are then colored respectively.

The information necessary for the graphical pathway comparison is contained within the tables "gene2ncbi", which contains the mapping from the PEDANT GI identifiers to KEGG gene names, "genes2ko", which contains the mapping from KEGG gene names to KEGG orthologous groups, "kegg_mapping" which contains the mapping from KEGG orthologous groups to EC numbers, "kegg_pathway_member", which contains the assignment of KEGG orthologous groups to KEGG pathways, "kegg_pathway_name", which contains the names of the KEGG pathways, and "kegg_pathway_umbrella", which contains the umbrella terms of the KEGG pathways.

An example of a pathway comparison between three chlamydial species can be seen


ChlamydiaeDB.org
TUM TECHNISCHE UNIVERSITÄT MÜNCHEN
universität wien
Search
In
in all chlamydiae
search for
Search fulltext!
Clear!
» Sequence Search
Browse
Browse genome:
Candidatus Protochlamydia amoet
Browse by genome
Browse annotation:
protein coding genes
Browse by annotation
Analyze
» Find enriched/depleted features in your set of proteins
» Get condensed information for your set of proteins
» Type III effector prediction for your sequences
» Compare metabolic pathways between organisms graphically
Information
» Home
» Glossary
» Help
» About
» Contact

Candidatus Protochlamydia amoebophila UWE25

NCBI taxonomy ID	264201
Genome length	2414465
# of contigs	1
# of proteins	2030
# of RNAs	43
# of rRNAs	8
# of tRNAs	35
# of automatic EC annotations	673
# of automatic FunCat annotations	5030
# of predicted Type-III secreted effector proteins	306
# of automatic GO annotations	4329

Automatically annotated literature

	Terms that matched the publication
Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, Jin Z, Lee P, Yang L, Poles M, Brown SM, Sotero S, Desantis T, Brodie E, Nelson K, Pei Z. **Diversity of 16S rRNA Genes within Individual Prokaryotic Genomes.** Appl Environ Microbiol 2010 Jun;76(12):3886-97. PMID: 20418441	Protochlamydia amoebophila
Moliner C, Raoult D, Fournier PE. **Evidence that the intra-amoebal Legionella drancourtii acquired a sterol reductase gene from eukaryotes.** BMC Res Notes 2009;2:51. PMID: 19327142	P. amoebophila Protochlamydia amoebophila
Heinz E, Tischler P, Rattei T, Myers G, Wagner M, Horn M. **Comprehensive in silico prediction and analysis of chlamydial outer membrane proteins reflects evolution and life style of the Chlamydiae.** BMC Genomics 2009;10:634. PMID: 20040079	Chlamydia muridarum Chlamydia trachomatis Protochlamydia amoebophila
Haider S, Collingro A, Walochnik J, Wagner M, Horn M. **Chlamydia-like bacteria in respiratory samples of community-acquired pneumonia patients.** FEMS Microbiol Lett 2008 Apr;281(2):198-202. PMID: 18312573	Chlamydia trachomatis Chlamydophila pneumoniae Protochlamydia amoebophila
Baud D, Regan L, Greub G. **Emerging role of Chlamydia and Chlamydia-like organisms in adverse pregnancy outcomes.** Curr Opin Infect Dis 2008 Feb;21(1):70-6. PMID: 18192789	Chlamydia trachomatis Chlamydophila abortus

Figure 2.31: Overview and literature for an organism, as example *Candidatus Protochlamydia amoebophila* UWE25 within the ChlamydiaeDB The table in the upper part shows information about the number of specific elements and some annotations, the lower part shows the automatically assigned literature.

in Figure 2.32. Each organism has a specific color assigned. If an enzyme is existing in the organism then this enzyme is colored with the the color of the organism. There are also color mixtures possible if an enzyme is contained in more than one organism. The colors can be interpreted using the additive chromatic circle.

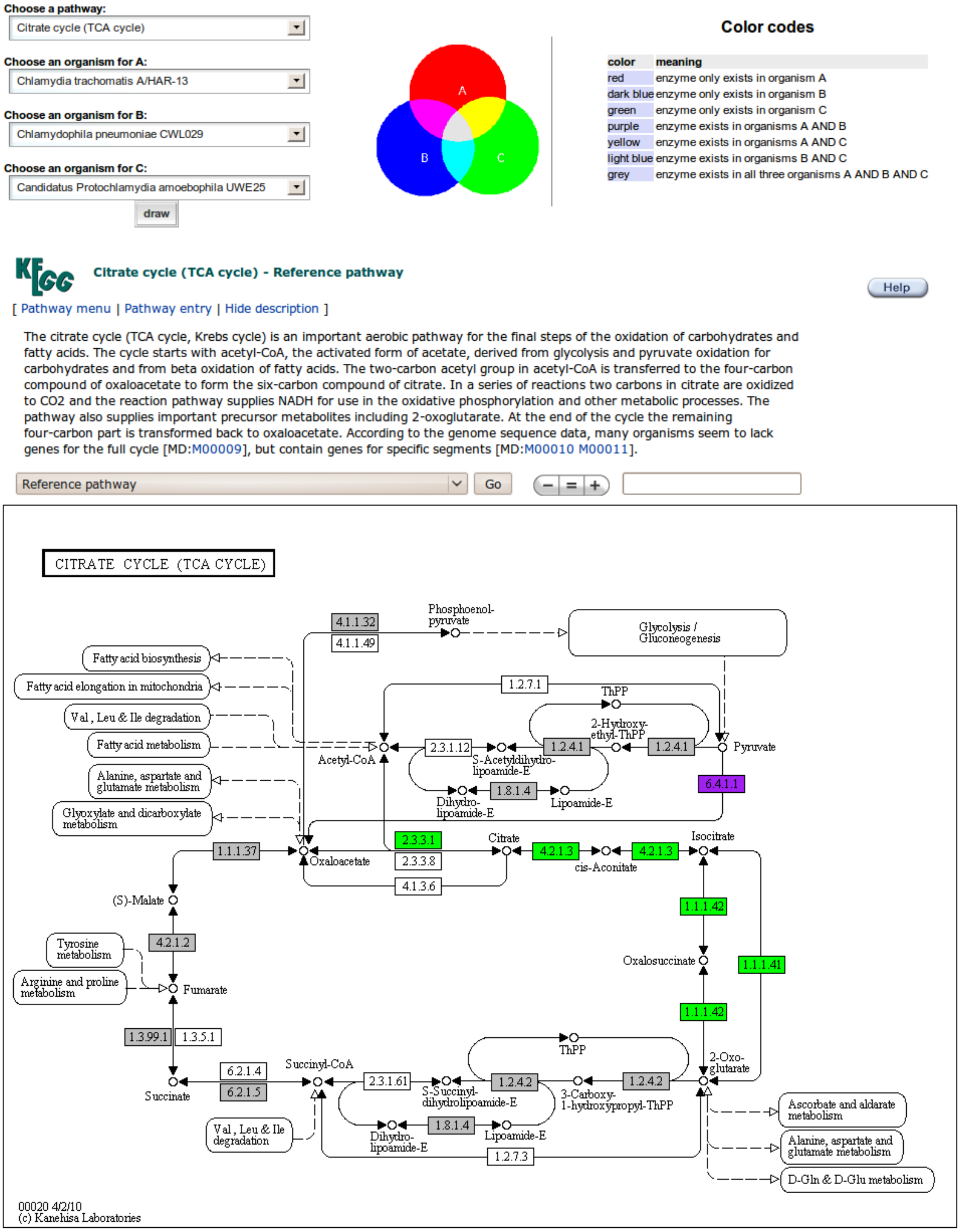
The possibility of graphical pathway comparison has also been provided for private genomes in a password protected section within the ChlamydiaeDB.

The graphical pathway comparison can be of interest for everyone working with RefSeq genomes as the comparison is not only restricted to *Chlamydiae* but is available for all RefSeq genomes contained in PEDANT.

2.3.6.1.3 Integration of data from experiments The integration of data from experiments is essential for the representation of the complete knowledge about *Chlamydiae*. Therefore different kinds of information from experiments can be included in the ChlamydiaeDB.

2.3.6.1.3.1 Single Nucleotide Polymorphisms (SNPs) The ChlamydiaeDB is able to store and display SNPs that have been determined for example in resequencing projects.

SIFT [332] is a tool that uses sequence homology to predict whether a substitution (e.g.



Choose a pathway: Citrate cycle (TCA cycle)

Choose an organism for A: Chlamydia trachomatis A/HAR-13

Choose an organism for B: Chlamydophila pneumoniae CWL029

Choose an organism for C: Candidatus Protochlamydia amoebophila UWE25

Color codes

color	meaning
red	enzyme only exists in organism A
dark blue	enzyme only exists in organism B
green	enzyme only exists in organism C
purple	enzyme exists in organisms A AND B
yellow	enzyme exists in organisms A AND C
light blue	enzyme exists in organisms B AND C
grey	enzyme exists in all three organisms A AND B AND C

Citrate cycle (TCA cycle) - Reference pathway

[Pathway menu | Pathway entry | Hide description]

The citrate cycle (TCA cycle, Krebs cycle) is an important aerobic pathway for the final steps of the oxidation of carbohydrates and fatty acids. The cycle starts with acetyl-CoA, the activated form of acetate, derived from glycolysis and pyruvate oxidation for carbohydrates and from beta oxidation of fatty acids. The two-carbon acetyl group in acetyl-CoA is transferred to the four-carbon compound of oxaloacetate to form the six-carbon compound of citrate. In a series of reactions two carbons in citrate are oxidized to CO2 and the reaction pathway supplies NADH for use in the oxidative phosphorylation and other metabolic processes. The pathway also supplies important precursor metabolites including 2-oxoglutarate. At the end of the cycle the remaining four-carbon part is transformed back to oxaloacetate. According to the genome sequence data, many organisms seem to lack genes for the full cycle [MD:M00009], but contain genes for specific segments [MD:M00010 M00011].

Figure 2.32: Graphical pathway comparison between the citrate cycles of three chlamydial species in the ChlamydiaeDB Each organism has a specific color assigned. If an enzyme is existing in the organism then this enzyme is colored with the the color of the organism. There are also color mixtures possible if an enzyme should be contained in more than one organism. The colors can be interpreted using the the additive chromatic circle on the upper right. The pathway comparison shows that the two pathogenic *Chlamydia C. trachomatis* and *Cp. pneumoniae* have reduced metabolic capabilities in comparison to the environmental chlamydia *P. amoebophila*. Nevertheless the three organisms share a large fraction of the pathway.

SNP) affects protein function. The SIFT predictions can be used as a hint towards how a detected SNP might affect the function of a coding region. Therefore the predictions can also be included in the ChlamydiaeDB.

The information about SNPs and SIFT predictions is stored in the tables "isolate" and "snp" within the data warehouse. The table "isolate" contains information about the isolate including information that makes it possible to linkout, and the file containing the SNPs. The table "snp" contains the information contained in the SNP files and the SIFT predictions.

The SNP files contain tab-separated the name of the reference genome, the position in the reference genome, the nucleotide in the reference genome, the name of the isolate, the nucleotide in the isolate. This format has been used as it was already used within the Master thesis of Jonathan Hoser.

An example of the visualization of SNPs can be seen in Figure 2.33 B. The SIFT prediction for a specific SNP can be seen when the mouse is moved over it.

2.3.6.1.3.2 Transcript data Transcript data provides information about the transcription of genes and can give hints towards correct gene starts. Therefore transcript data of protein coding regions has also been integrated into the ChlamydiaeDB.

Transcript data is stored in the table "transcript". Besides the name of the respective protein and a databaseid that needs to be existing in the table "db", element starts and element stops, the strand, the transcription start site (TSS), abundance, source and linkouturl can be stored.

An example of the visualization of transcript data can be seen in Figure 2.33 C.

2.3.6.1.3.3 Proteome data Proteome data is important as it provides hints towards the existence of a protein.

Proteome data is stored in the table "proteome". Besides the name of the respective protein and a databaseid that needs to be existing in the table "db", the abundance, a remark, the source and a linkouturl can be stored.

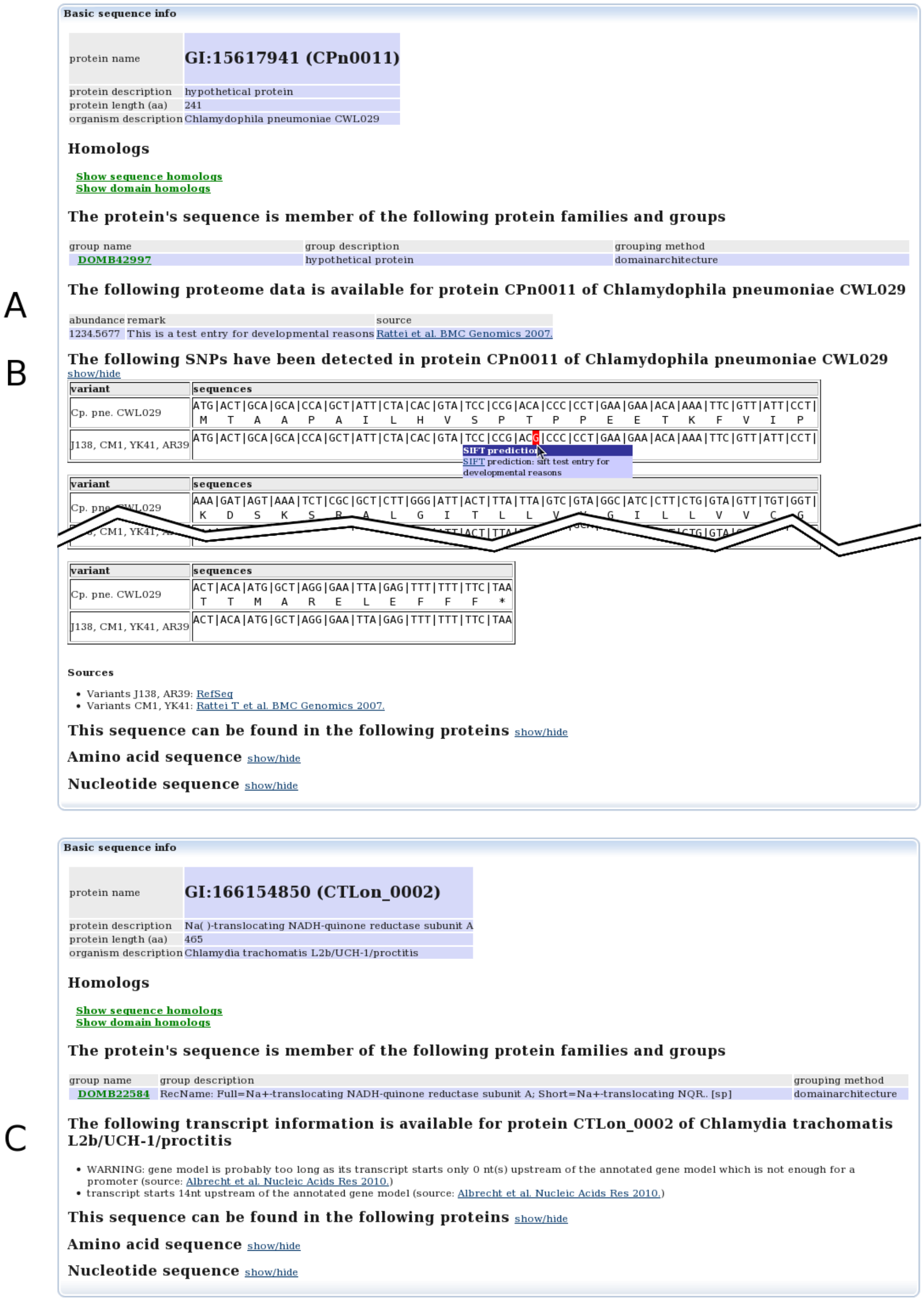
An example of the visualization of proteome data can be seen in Figure 2.33 A.

2.3.6.2 Functionality

In the following the functionality as defined in the criteria section 2.3.2 is shown.

2.3.6.2.1 Comprehensive search possibilities As in the SIMAP web portal it is possible to search for search terms and sequences. An overview over the search possibilities can be seen in Figure 2.34.

2.3.6.2.2 Structuring of search results by the integration of taxonomy and orthologous information The search results can either be displayed as list of hits or sorted taxonomically (see Figure 2.35). In the case that it is searched for a commonly used gene name this can help to narrow the search results down.



Basic sequence info

protein name	**GI:15617941 (CPn0011)**
protein description	hypothetical protein
protein length (aa)	241
organism description	Chlamydophila pneumoniae CWL029

Homologs

Show sequence homologs
Show domain homologs

The protein's sequence is member of the following protein families and groups

group name	group description	grouping method
DOMB42997	hypothetical protein	domainarchitecture

The following proteome data is available for protein CPn0011 of Chlamydophila pneumoniae CWL029

abundance	remark	source
1234.5677	This is a test entry for developmental reasons	Rattei et al. BMC Genomics 2007.

The following SNPs have been detected in protein CPn0011 of Chlamydophila pneumoniae CWL029
show/hide

variant	sequences
Cp. pne. CWL029	ATG|ACT|GCA|GCA|CCA|GCT|ATT|CTA|CAC|GTA|TCC|CCG|ACA|CCC|CCT|GAA|GAA|ACA|AAA|TTC|GTT|ATT|CCT| M T A A P A I L H V S P T P P E E T K F V I P
J138, CM1, YK41, AR39	ATG|ACT|GCA|GCA|CCA|GCT|ATT|CTA|CAC|GTA|TCC|CCG|ACG|CCC|CCT|GAA|GAA|ACA|AAA|TTC|GTT|ATT|CCT|

SIFT prediction
SIFT prediction: sift test entry for developmental reasons

variant	sequences
Cp. pne. CWL029	AAA|GAT|AGT|AAA|TCT|CGC|GCT|CTT|GGG|ATT|ACT|TTA|TTA|GTC|GTA|GGC|ATC|CTT|CTG|GTA|GTT|TGT|GGT| K D S K S R A L G I T L L V V G I L L V V C G
, CM1, YK41,	|ATT|ACT|TTA| |CTG|GTA|

variant	sequences
Cp. pne. CWL029	ACT|ACA|ATG|GCT|AGG|GAA|TTA|GAG|TTT|TTT|TTC|TAA T T M A R E L E F F F *
J138, CM1, YK41, AR39	ACT|ACA|ATG|GCT|AGG|GAA|TTA|GAG|TTT|TTT|TTC|TAA

Sources

- Variants J138, AR39: RefSeq
- Variants CM1, YK41: Rattei T et al. BMC Genomics 2007.

This sequence can be found in the following proteins show/hide

Amino acid sequence show/hide

Nucleotide sequence show/hide

Basic sequence info

protein name	**GI:166154850 (CTLon_0002)**
protein description	Na( )-translocating NADH-quinone reductase subunit A
protein length (aa)	465
organism description	Chlamydia trachomatis L2b/UCH-1/proctitis

Homologs

Show sequence homologs
Show domain homologs

The protein's sequence is member of the following protein families and groups

group name	group description	grouping method
DOMB22584	RecName: Full=Na+-translocating NADH-quinone reductase subunit A; Short=Na+-translocating NQR.. [sp]	domainarchitecture

The following transcript information is available for protein CTLon_0002 of Chlamydia trachomatis L2b/UCH-1/proctitis

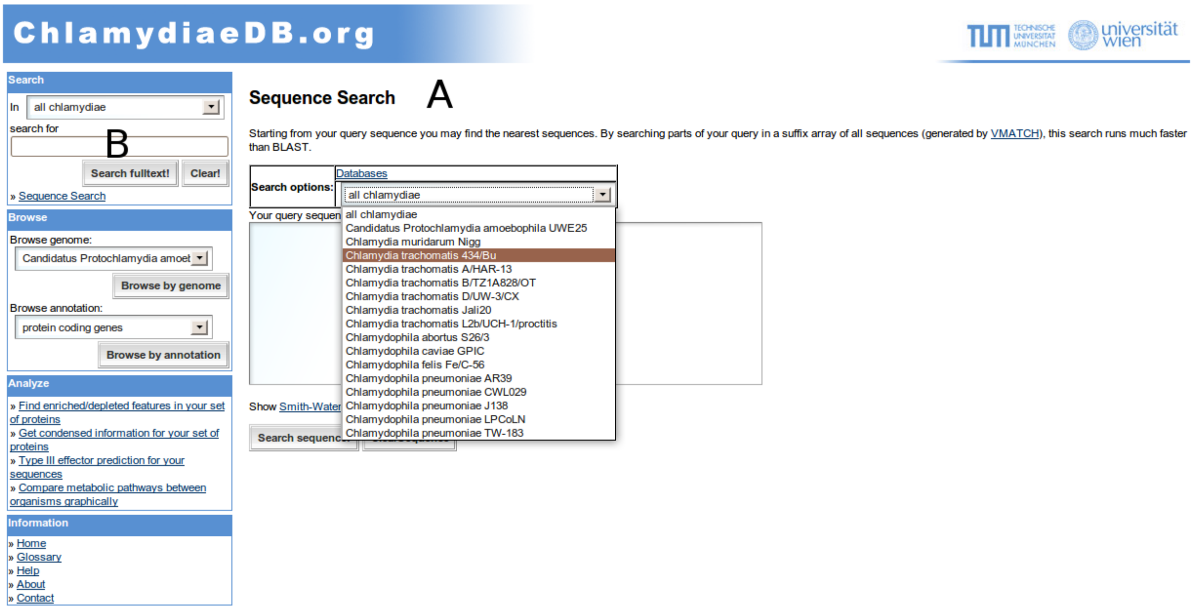
- WARNING: gene model is probably too long as its transcript starts only 0 nt(s) upstream of the annotated gene model which is not enough for a promoter (source: Albrecht et al. Nucleic Acids Res 2010.)
- transcript starts 14nt upstream of the annotated gene model (source: Albrecht et al. Nucleic Acids Res 2010.)

This sequence can be found in the following proteins show/hide

Amino acid sequence show/hide

Nucleotide sequence show/hide

Figure 2.33: Visualization of SNPs, transcripts and proteome data in the ChlamydiaeDB The figure shows extracts of two protein reports, the upper one showing the visualization of proteome data and SNPs, the other one showing the visualization of transcript data. **A:** Shows the visualization of (test) data from a proteome experiment **B:** Shows the visualization of a synonymous SNP in isolates of *Chlamydophila pneumoniae* CWL029. The first sequence is the reference genome, the line below are SNPs in isolates. The same SNPs in different isolates are merged. It is also possible to display SIFT [332] predictions for each SNP, that are displayed on mouse-over for every SNP. **C:** Shows the information available from a transcript sequencing experiment in *Chlamydia trachomatis* L2b/UCH-1/proctitis.



Sequence Search

Starting from your query sequence you may find the nearest sequences. By searching parts of your query in a suffix array of all sequences (generated by VMATCH), this search runs much faster than BLAST.

Search options: Databases: all chlamydiae

ChlamydiaeDB.org
TUM TECHNISCHE UNIVERSITÄT MÜNCHEN
universität wien
A
B
Search
In all chlamydiae
search for
Search fulltext!
Clear!
» Sequence Search
Browse
Browse genome:
Candidatus Protochlamydia amoet
Browse by genome
Browse annotation:
protein coding genes
Browse by annotation
Analyze
» Find enriched/depleted features in your set of proteins
» Get condensed information for your set of proteins
» Type III effector prediction for your sequences
» Compare metabolic pathways between organisms graphically
Information
» Home
» Glossary
» Help
» About
» Contact
Your query sequen
all chlamydiae
Candidatus Protochlamydia amoebophila UWE25
Chlamydia muridarum Nigg
Chlamydia trachomatis 434/Bu
Chlamydia trachomatis A/HAR-13
Chlamydia trachomatis B/TZ1A828/OT
Chlamydia trachomatis D/UW-3/CX
Chlamydia trachomatis Jali20
Chlamydia trachomatis L2b/UCH-1/proctitis
Chlamydophila abortus S26/3
Chlamydophila caviae GPIC
Chlamydophila felis Fe/C-56
Chlamydophila pneumoniae AR39
Chlamydophila pneumoniae CWL029
Chlamydophila pneumoniae J138
Chlamydophila pneumoniae LPCoLN
Chlamydophila pneumoniae TW-183
Show Smith-Water
Search sequenc

Figure 2.34: Search possibilities in the ChlamydiaeDB A: The sequence search can be seen on the right side. It is possible to enter a search sequence and to select in which organism the search should be performed in the pull down menu. **B:** This search field is part of the main portlet which is always available on every page of the ChlamydiaeDB portal. It can be selected in which organism the search for a term, e.g. gene name, should be performed.

These views are quite similar to the search result views in SIMAP (Figure 2.21) and are a good example for the reusability of portlets. The layout of the search result visualization has been slightly modified due to user requests.

If the sequence should not be known to the ChlamydiaeDB then parts of the query sequence are searched in a suffix array of all ChlamydiaeDB sequences generated by VMATCH (`http://www.vmatch.de`). These results are shown as a list. This is the same procedure as for SIMAP.

2.3.6.2.3 "The protein report" - Gene centric and group centric views

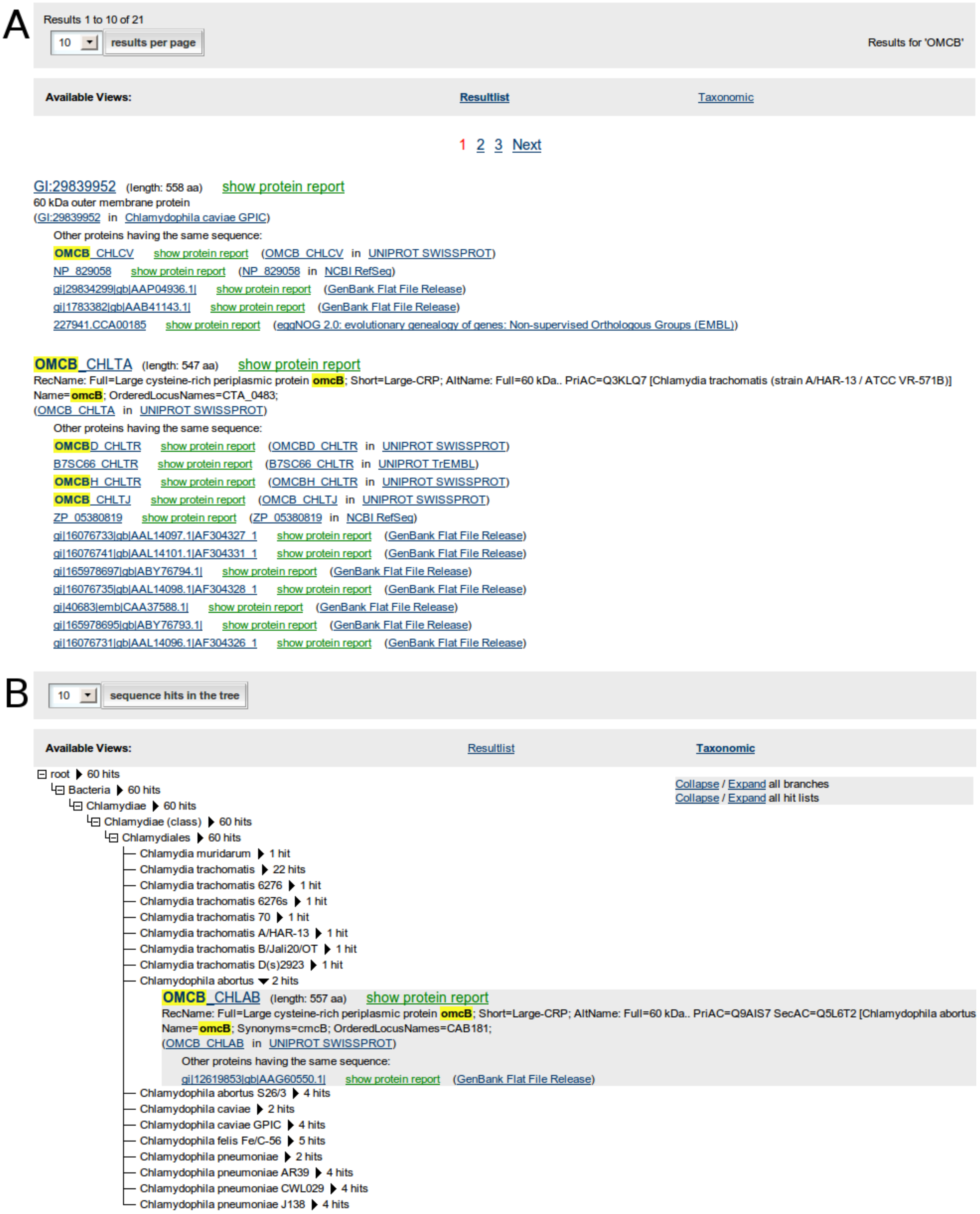
The protein report combines all available information about a genetic element in one place (Figure 2.36). That way the researcher interested in the protein can get an insight into the currently available knowledge about this specific entry. The parts of the protein report are described in more detail in the following sections.

2.3.6.2.3.1 General information

The general information portlet (Figure 2.37) shows general information available for all sequences and additionally specific information like data from experiments.

The following information is always available:

- protein name
- protein description



Results 1 to 10 of 21

Results for 'OMCB'

GI:29839952 (length: 558 aa) show protein report
60 kDa outer membrane protein
(GI:29839952 in Chlamydophila caviae GPIC)

Other proteins having the same sequence:

OMCB_CHLCV show protein report (OMCB_CHLCV in UNIPROT SWISSPROT)
NP_829058 show protein report (NP_829058 in NCBI RefSeq)
gi|29834299|gb|AAP04936.1| show protein report (GenBank Flat File Release)
gi|1783382|gb|AAB41143.1| show protein report (GenBank Flat File Release)
227941.CCA00185 show protein report (eggNOG 2.0: evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL))

OMCB_CHLTA (length: 547 aa) show protein report
RecName: Full=Large cysteine-rich periplasmic protein omcB; Short=Large-CRP; AltName: Full=60 kDa.. PriAC=Q3KLQ7 [Chlamydia trachomatis (strain A/HAR-13 / ATCC VR-571B)]
Name=omcB; OrderedLocusNames=CTA_0483;
(OMCB_CHLTA in UNIPROT SWISSPROT)

Other proteins having the same sequence:

OMCBD_CHLTR show protein report (OMCBD_CHLTR in UNIPROT SWISSPROT)
B7SC66_CHLTR show protein report (B7SC66_CHLTR in UNIPROT TrEMBL)
OMCBH_CHLTR show protein report (OMCBH_CHLTR in UNIPROT SWISSPROT)
OMCB_CHLTJ show protein report (OMCB_CHLTJ in UNIPROT SWISSPROT)
ZP_05380819 show protein report (ZP_05380819 in NCBI RefSeq)
gi|16076733|gb|AAL14097.1|AF304327_1 show protein report (GenBank Flat File Release)
gi|16076741|gb|AAL14101.1|AF304331_1 show protein report (GenBank Flat File Release)
gi|165978697|gb|ABY76794.1| show protein report (GenBank Flat File Release)
gi|16076735|gb|AAL14098.1|AF304328_1 show protein report (GenBank Flat File Release)
gi|40683|emb|CAA37588.1| show protein report (GenBank Flat File Release)
gi|165978695|gb|ABY76793.1| show protein report (GenBank Flat File Release)
gi|16076731|gb|AAL14096.1|AF304326_1 show protein report (GenBank Flat File Release)

```
⊟ root ▶ 60 hits
 └⊟ Bacteria ▶ 60 hits
   └⊟ Chlamydiae ▶ 60 hits
     └⊟ Chlamydiae (class) ▶ 60 hits
       └⊟ Chlamydiales ▶ 60 hits
          ├─ Chlamydia muridarum ▶ 1 hit
          ├─ Chlamydia trachomatis ▶ 22 hits
          ├─ Chlamydia trachomatis 6276 ▶ 1 hit
          ├─ Chlamydia trachomatis 6276s ▶ 1 hit
          ├─ Chlamydia trachomatis 70 ▶ 1 hit
          ├─ Chlamydia trachomatis A/HAR-13 ▶ 1 hit
          ├─ Chlamydia trachomatis B/Jali20/OT ▶ 1 hit
          ├─ Chlamydia trachomatis D(s)2923 ▶ 1 hit
          ├─ Chlamydophila abortus ▼ 2 hits
          │     OMCB_CHLAB (length: 557 aa) show protein report
          │     RecName: Full=Large cysteine-rich periplasmic protein omcB; Short=Large-CRP; AltName: Full=60 kDa.. PriAC=Q9AIS7 SecAC=Q5L6T2 [Chlamydophila abortus
          │     Name=omcB; Synonyms=cmcB; OrderedLocusNames=CAB181;
          │     (OMCB_CHLAB in UNIPROT SWISSPROT)
          │       Other proteins having the same sequence:
          │       gi|12619853|gb|AAG60550.1| show protein report (GenBank Flat File Release)
          ├─ Chlamydophila abortus S26/3 ▶ 4 hits
          ├─ Chlamydophila caviae ▶ 2 hits
          ├─ Chlamydophila caviae GPIC ▶ 4 hits
          ├─ Chlamydophila felis Fe/C-56 ▶ 5 hits
          ├─ Chlamydophila pneumoniae ▶ 2 hits
          ├─ Chlamydophila pneumoniae AR39 ▶ 4 hits
          ├─ Chlamydophila pneumoniae CWL029 ▶ 4 hits
          └─ Chlamydophila pneumoniae J138 ▶ 4 hits
```

Figure 2.35: Search result visualizations in the ChlamydiaeDB These views are quite similar to the representations of the SIMAP web portal (Figure 2.21), only slight changes have been made in comparison to SIMAP due to user requests. **A:** List representation of search results **B:** Taxonomic representation of search results

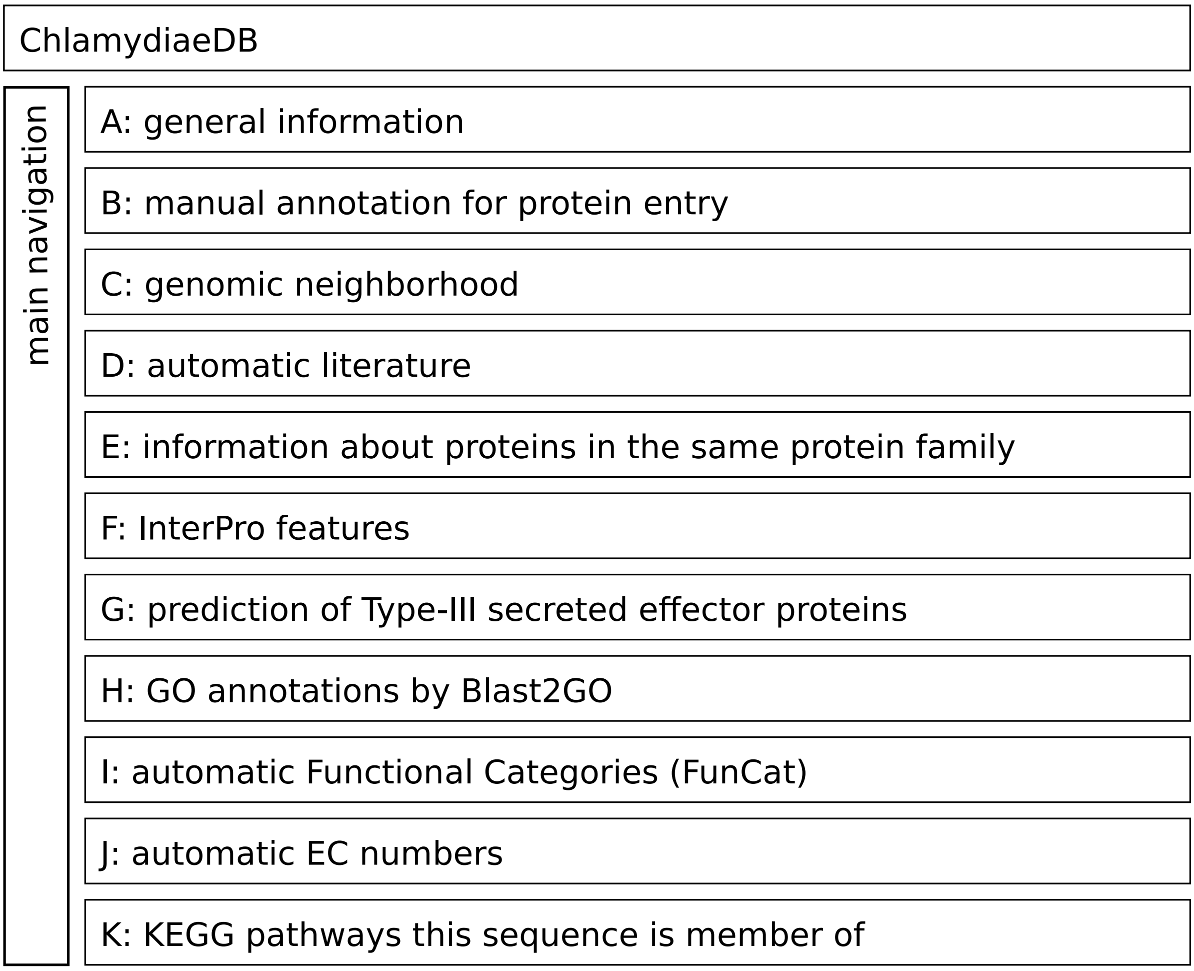

ChlamydiaeDB
main navigation
A: general information
B: manual annotation for protein entry
C: genomic neighborhood
D: automatic literature
E: information about proteins in the same protein family
F: InterPro features
G: prediction of Type-III secreted effector proteins
H: GO annotations by Blast2GO
I: automatic Functional Categories (FunCat)
J: automatic EC numbers
K: KEGG pathways this sequence is member of

Figure 2.36: Overview over information within the protein report of the ChlamydiaeDB The protein report is the central place within the ChlamydiaeDB that combines information from various sources in one place.

- protein length
- organism
- links to sequence and domain homologs
- information about synonymous names for this entry and about the proteins with the same sequence in the same or in other organisms
- amino acid sequence
- nucleotide sequence

Additionally there might be information available about:

- links to clusters the protein is assigned to

- SNPs (Figure 2.33 B)
- transcript data (Figure 2.33 C)
- proteome data (Figure 2.33 A)

2.3.6.2.3.2 Automatic literature Automatically assigned literature for a protein can be seen if available (Figure 2.38). The publications are listed and also the terms that were automatically identified within the publication.

2.3.6.2.3.3 InterPro features InterPro protein domains are displayed as a table that shows the assigned InterPro entries and the corresponding entries in the integrated domain signature databases. Additionally a graphical representation of the domain signatures on the protein sequence is available (Figure 2.39).

2.3.6.2.3.4 Prediction of Type-III secreted effector proteins The prediction result for the Type-III secretion prediction is displayed (Figure 2.40).

2.3.6.2.3.5 GO annotations by Blast2GO The GO annotations as derived by Blast2GO [93] are displayed as a table (Figure 2.41). Links to the GO graph are provided.

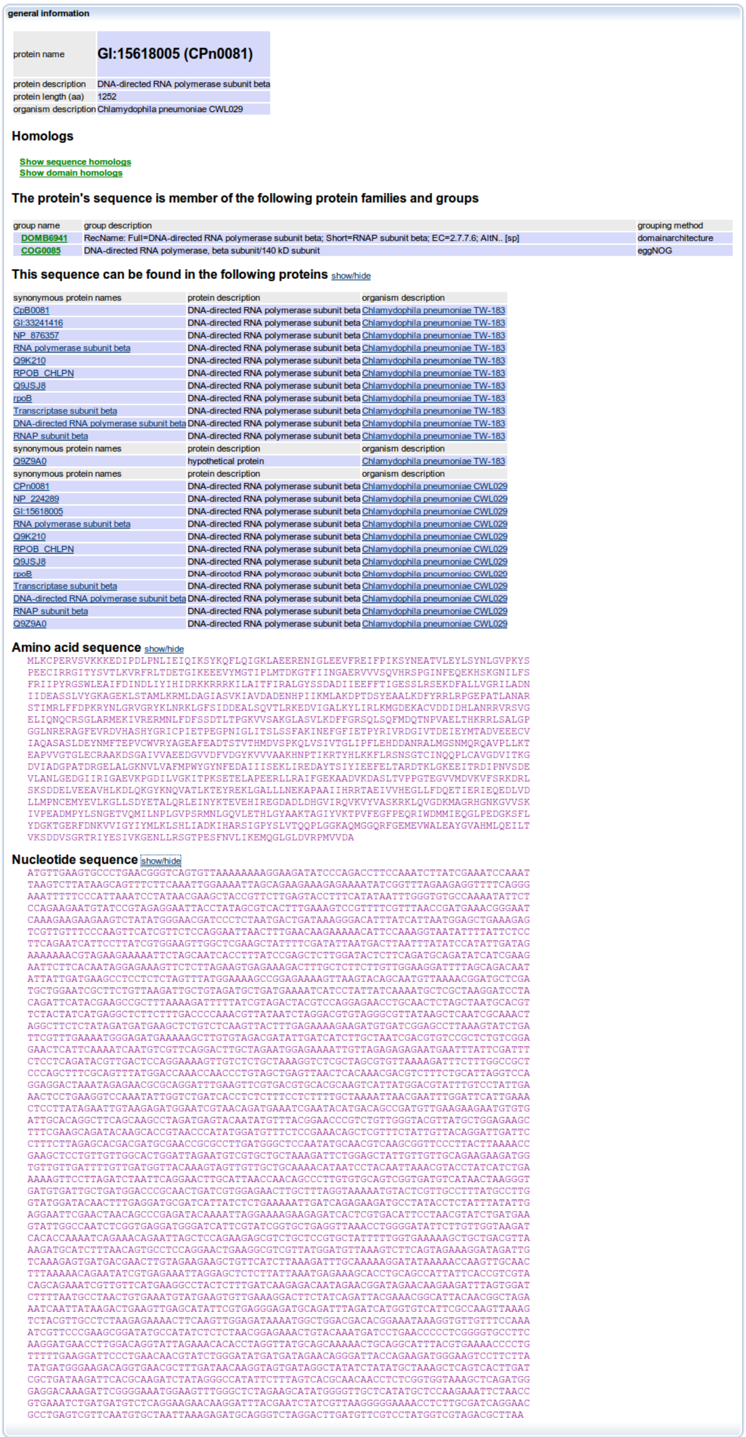
2.3.6.2.3.6 Automatic Functional Categories (FunCat) The hierarchical FunCat annotations as automatically derived from PEDANT [72] are displayed in an interactive tree (Figure 2.42). It is possible to expand and collapse nodes in the tree.

2.3.6.2.3.7 Automatic EC numbers The EC annotations as automatically derived from PEDANT [72] are displayed as a table (Figure 2.43).

2.3.6.2.3.8 KEGG pathways this sequence is member of The assignment of the protein to different KEGG pathways is displayed (Figure 2.44). It is possible to directly show the affected map in the graphical pathway comparison (section 2.3.6.1.2.6 and Figure 2.32).

2.3.6.2.3.9 Manual annotation for protein entry Manual annotations if already accepted by an administrator can be seen (Figure 2.45). These annotations include literature, GO, FunCat, EC and comments. (see also section 2.3.6.2.4)

2.3.6.2.3.10 Genomic neighborhood The synteny or genomic neighborhood portlet shows the neighboring genes of the currently selected gene in the same genome as well as in the other chlamydial genomes (Figure 2.46). The genes are colored by their membership in eggNOG clusters. The protein, for which the protein report is currently displayed, is shown in the first line in the middle colored in red. The other members of



general information

protein name	GI:15618005 (CPn0081)
protein description	DNA-directed RNA polymerase subunit beta
protein length (aa)	1252
organism description	Chlamydophila pneumoniae CWL029

Homologs

Show sequence homologs
Show domain homologs

The protein's sequence is member of the following protein families and groups

group name	group description	grouping method
DOMB6941	RecName: Full=DNA-directed RNA polymerase subunit beta; Short=RNAP subunit beta; EC=2.7.7.6; AltN.. [sp]	domainarchitecture
COG0085	DNA-directed RNA polymerase, beta subunit/140 kD subunit	eggNOG

This sequence can be found in the following proteins show/hide

synonymous protein names	protein description	organism description
CpB0081	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae TW-183
GI:33241416	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae TW-183
NP_876357	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae TW-183
RNA polymerase subunit beta	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae TW-183
Q9K210	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae TW-183
RPOB_CHLPN	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae TW-183
Q9JSJ8	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae TW-183
rpoB	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae TW-183
Transcriptase subunit beta	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae TW-183
DNA-directed RNA polymerase subunit beta	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae TW-183
RNAP subunit beta	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae TW-183
synonymous protein names	protein description	organism description
Q9Z9A0	hypothetical protein	Chlamydophila pneumoniae TW-183
synonymous protein names	protein description	organism description
CPn0081	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae CWL029
NP_224289	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae CWL029
GI:15618005	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae CWL029
RNA polymerase subunit beta	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae CWL029
Q9K210	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae CWL029
RPOB_CHLPN	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae CWL029
Q9JSJ8	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae CWL029
rpoB	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae CWL029
Transcriptase subunit beta	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae CWL029
DNA-directed RNA polymerase subunit beta	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae CWL029
RNAP subunit beta	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae CWL029
Q9Z9A0	DNA-directed RNA polymerase subunit beta	Chlamydophila pneumoniae CWL029

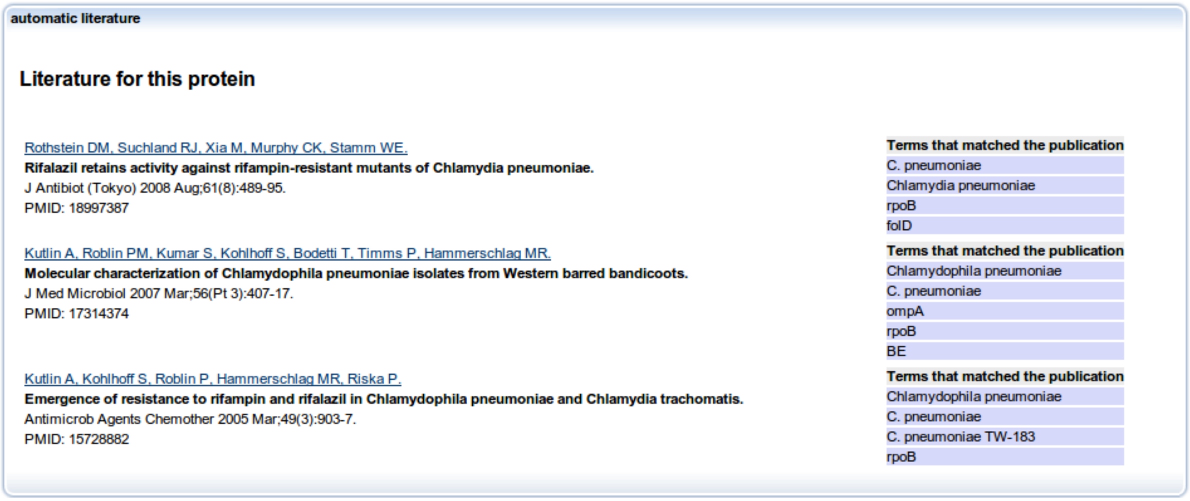
Amino acid sequence show/hide

```
MLKCPERVSVKKKEDIPDLPNLIEIQIKSYKQFLQIGKLAEERENIGLEEVFREIFPIKSYNEATVLEYLSYNLGVPKYS
PEECIRRGITYSVTLKVRFRLTDETGIKEEEVYMGTIPLMTDKGTFIINGAERVVVSQVHRSPGINFEQEKHSKGNILFS
FRIIPYRGSWLEAIFDINDLIYIHIDRKKRRRKILAITFIRALGYSSDADIIEEFFTIGESSLRSEKDFALLVGRILADN
IIDEASSLVYGKAGEKLSTAMLKRMLDAGIASVKIAVDADENHPIIKMLAKDPTDSYEAALKDFYRRLRPGEPATLANAR
STIMRLFFDPKRYNLGRVGRYKLNRKLGFSIDDEALSQVTLRKEDVIGALKYLIRLKMGDEKACVDDIDHLANRRVRSVG
ELIQNQCRSGLARMEKIVRERMNLFDFSSDTLTPGKVVSAKGLASVLKDFFGRSQLSQFMDQTNPVAELTHKRRLSALGP
GGLNRERAGFEVRDVHASHYGRICPIETPEGPNIGLITSLSSFAKINEFGFIETPYRIVRDGIVTDEIEYMTADVEEECV
IAQASASLDEYNMFTEPVCWVRYAGEAFEADTSTVTHMDVSPKQLVSIVTGLIPFLEHDDANRALMGSNMQRQAVPLLKT
EAPVVGTGLECRAAKDSGAIVVAEEDGVVDFVDGYKVVVAAKHNPTIKRTYHLKKFLRSNSGTCINQQPLCAVGDVITKG
DVIADGPATDRGELALGKNVLVAFMPWYGYNFEDAIIISEKLIREDAYTSIYIEEFELTARDTKLGKEEITRDIPNVSDE
VLANLGEDGIIRIGAEVKPGDILVGKITPKSETELAPEERLLRAIFGEKAADVKDASLTVPPGTEGVVMDVKVFSRKDRL
SKSDDELVEEAVHLKDLQKGYKNQVATLKTEYREKLGALLLNEKAPAAIIHRRTAEIVVHEGLLFDQETIERIEQEDLVD
LLMPNCEMYEVLKGLLSDYETALQRLEINYKTEVEHIREGDADLDHGVIRQVKVYVASKRKLQVGDKMAGRHGNKGVVSK
IVPEADMPYLSNGETVQMILNPLGVPSRMNLGQVLETHLGYAAKTAGIYVKTPVFEGFPEQRIWDMMIEQGLPEDGKSFL
YDGKTGERFDNKVVIGYIYMLKLSHLIADKIHARSIGPYSLVTQQPLGGKAQMGGQRFGEMEVWALEAYGVAHMLQEILT
VKSDDVSGRTRIYESIVKGENLLRSGTPESFNVLIKEMQGLGLDVRPMVVDA
```

Nucleotide sequence show/hide

```
ATGTTGAAGTGCCCTGAACGGGTCAGTGTTAAAAAAAAGGAAGATATCCCAGACCTTCCAAATCTTATCGAAATCCAAAT
TAAGTCTTATAAGCAGTTTCTTCAAATTGGAAAATTAGCAGAAGAAAGAGAAAATATCGGTTTAGAAGAGGTTTTCAGGG
AAATTTTTCCCATTAAATCCTATAACGAAGCTACCGTTCTTGAGTACCTTTCATATAATTTGGGTGTGCCAAAATATTCT
CCAGAAGAATGTATCCGTAGAGGAATTACCTATAGCGTCACTTTGAAAGTCCGTTTTCGTTTAACCGATGAAACGGGAAT
CAAAGAAGAAGAAGTCTATATGGGAACGATCCCTCTAATGACTGATAAAGGGACATTTATCATTAATGGAGCTGAAAGAG
TCGTTGTTTCCCAAGTTCATCGTTCTCCAGGAATTAACTTTGAACAAGAAAAACATTCCAAAGGTAATATTTTATTCTCC
TTCAGAATCATTCCTTATCGTGGAAGTTGGCTCGAAGCTATTTTCGATATTAATGACTTAATTTATATCCATATTGATAG
AAAAAAACGTAGAAGAAAAAATTCTAGCAATCACCTTTATCCGAGCTCTTGGATACTCTTCAGATGCAGATATCATCGAAG
AATTCTTCACAATAGGAGAAAGTTCTCTTAGAAGTGAGAAAGACTTTGCTCTTCTTGTTGGAAGGATTTTAGCAGACAAT
ATTATTGATGAAGCCTCCTCTCTAGTTTATGGAAAAGCCGGAGAAAAGTTAAGTACAGCAATGTTAAAACGGATGCTCGA
TGCTGGAATCGCTTCTGTTAAGATTGCTGTAGATGCTGATGAAAATCATCCTATTATCAAAATGCTCGCTAAGGATCCTA
CAGATTCATACGAAGCCGCTTTAAAAGATTTTTATCGTAGACTACGTCCAGGAGAACCTGCAACTCTAGCTAATGCACGT
TCTACTATCATGAGGCTCTTCTTTGACCCCAAACGTTATAATCTAGGACGTGTAGGGCGTTATAAGCTCAATCGCAAACT
AGGCTTCTCTATAGATGATGAAGCTCTGTCTCAAGTTACTTTGAGAAAAGAAGATGTGATCGGAGCCTTAAAGTATCTGA
TTCGTTTGAAAATGGGAGATGAAAAAGCTTGTGTAGACGATATTGATCATCTTGCTAATCGACGTGTCCGCTCTGTCGGA
GAACTCATTCAAAATCAATGTCGTTCAGGACTTGCTAGAATGGAGAAAATTGTTAGAGAGAGAATGAATTTATTCGATTT
CTCCTCAGATACGTTGACTCCAGGAAAAGTTGTCTCTGCTAAAGGTCTCGCTAGCGTGTTAAAAGATTTCTTTGGCCGCT
CCCAGCTTTCGCAGTTTATGGACCAAACCAACCCTGTAGCTGAGTTAACTCACAAACGACGTCTTTCTGCATTAGGTCCA
GGAGGACTAAATAGAGAACGCGCAGGATTTGAAGTTCGTGACGTGCACGCAAGTCATTATGGACGTATTTGTCCTATTGA
AACTCCTGAAGGTCCAAATATTGGTCTGATCACCTCTCTTTCCTCTTTTGCTAAAATTAACGAATTTGGATTCATTGAAA
CTCCTTATAGAATTGTAAGAGATGGAATCGTAACAGATGAAATCGAATACATGACAGCCGATGTTGAAGAAGAATGTGTG
ATTGCACAGGCTTCAGCAAGCCTAGATGAGTACAATATGTTTACGGAACCCGTCTGTTGGGTACGTTATGCTGGAGAAGC
TTTCGAAGCAGATACAAGCACCGTAACCCATATGGATGTTTCTCCGAAACAGCTCGTTTCTATTGTTACAGGATTGATTC
CTTTCTTAGAGCACGACGATGCGAACCGCGCCTTGATGGGCTCCAATATGCAACGTCAAGCGGTTCCCTTACTTAAAACC
GAAGCTCCTGTTGTTGGCACTGGATTAGAATGTCGTGCTGCTAAAGATTCTGGAGCTATTGTTGTTGCAGAAGAAGATGG
TGTTGTTGATTTTGTTGATGGTTACAAAGTAGTTGTTGCTGCAAAACATAATCCTACAATTAAACGTACCTATCATCTGA
AAAAGTTCCTTAGATCTAATTCAGGAACTTGCATTAACCAACAGCCCTTGTGTGCAGTCGGTGATGTCATAACTAAGGGT
GATGTGATTGCTGATGGACCCGCAACTGATCGTGGAGAACTTGCTTTAGGTAAAAATGTACTCGTTGCCTTTATGCCTTG
GTATGGATACAACTTTGAGGATGCGATCATTATCTCTGAAAAATTGATCAGAGAAGATGCCTATACCTCTATTTATATTG
AGGAATTCGAACTAACAGCCCGAGATACAAAATTAGGAAAAGAAGAGATCACTCGTGACATTCCTAACGTATCTGATGAA
GTATTGGCCAATCTCGGTGAGGATGGGATCATTCGTATCGGTGCTGAGGTTAAACCTGGGGATATTCTTGTTGGTAAGAT
CACACCAAAATCAGAAACAGAATTAGCTCCAGAAGAGCGTCTGCTCCGTGCTATTTTTGGTGAAAAAGCTGCTGACGTTA
AAGATGCATCTTTAACAGTGCCTCCAGGAACTGAAGGCGTCGTTATGGATGTTAAAGTCTTCAGTAGAAAGGATAGATTG
TCAAAGAGTGATGACGAACTTGTAGAAGAAGCTGTTCATCTTAAAGATTTGCAAAAAGGATATAAAAACCAAGTTGCAAC
TTTAAAAACAGAATATCGTGAGAAATTAGGAGCTCTCTTATTAAATGAGAAAGCACCTGCAGCCATTATTCACCGTCGTA
CAGCAGAAATCGTTGTTCATGAAGGCCTACTCTTTGATCAAGAGACAATAGAACGGATAGAACAAGAAGATTTAGTGGAT
CTTTTAATGCCTAACTGTGAAATGTATGAAGTGTTGAAAGGACTTCTATCAGATTACGAAACGGCATTACAACGGCTAGA
AATCAATTATAAGACTGAAGTTGAGCATATTCGTGAGGGAGATGCAGATTTAGATCATGGTGTCATTCGCCAAGTTAAAG
TCTACGTTGCCTCTAAGAGAAAACTTCAAGTTGGAGATAAAATGGCTGGACGACACGGAAATAAAGGTGTTGTTTCCAAA
ATCGTTCCCGAAGCGGATATGCCATATCTCTCTAACGGAGAAACTGTACAAATGATCCTGAACCCCCTCGGGGTGCCTTC
AAGGATGAACCTTGGACAGGTATTAGAAACACACCTAGGTTATGCAGCAAAAACTGCAGGCATTTACGTGAAAACCCCTG
TTTTTGAAGGATTCCCTGAACAACGTATCTGGGATATGATGATAGAACAGGGATTACCAGAAGATGGGAAGTCCTTCTTA
TATGATGGGAAGACAGGTGAACGCTTTGATAACAAGGTAGTGATAGGCTATATCTATATGCTAAAGCTCAGTCACTTGAT
CGCTGATAAGATTCACGCAAGATCTATAGGGCCATATTCTTTAGTCACGCAACAACCTCTCGGTGGTAAAGCTCAGATGG
GAGGACAAAGATTCGGGGAAATGGAAGTTTGGGCTCTAGAAGCATATGGGGTTGCTCATATGCTCCAAGAAATTCTAACC
GTGAAATCTGATGATGTCTCAGGAAGAACAAGGATTTACGAATCTATCGTTAAGGGGGAAAACCTCTTGCGATCAGGAAC
GCCTGAGTCGTTCAATGTGCTAATTAAAGAGATGCAGGGTCTAGGACTTGATGTTCGTCCTATGGTCGTAGACGCTTAA
```

Figure 2.37: Visualization of general information about a protein within the ChlamydiaeDB The general information portlet shows the protein name, protein description, protein length, organism name, links to sequence and domain homologs, information about synonymous names for this entry and about the proteins with this sequence occurring in other organisms, amino acid sequence and nucleotide sequence. Additionally if available it shows links to clusters the protein is assigned to, SNPs (Figure 2.33 B), transcript data (Figure 2.33 C) and proteome data (Figure 2.33 A).



automatic literature

Literature for this protein

Rothstein DM, Suchland RJ, Xia M, Murphy CK, Stamm WE.
Rifalazil retains activity against rifampin-resistant mutants of Chlamydia pneumoniae.
J Antibiot (Tokyo) 2008 Aug;61(8):489-95.
PMID: 18997387

Terms that matched the publication
C. pneumoniae
Chlamydia pneumoniae
rpoB
folD

Kutlin A, Roblin PM, Kumar S, Kohlhoff S, Bodetti T, Timms P, Hammerschlag MR.
Molecular characterization of Chlamydophila pneumoniae isolates from Western barred bandicoots.
J Med Microbiol 2007 Mar;56(Pt 3):407-17.
PMID: 17314374

Terms that matched the publication
Chlamydophila pneumoniae
C. pneumoniae
ompA
rpoB
BE

Kutlin A, Kohlhoff S, Roblin P, Hammerschlag MR, Riska P.
Emergence of resistance to rifampin and rifalazil in Chlamydophila pneumoniae and Chlamydia trachomatis.
Antimicrob Agents Chemother 2005 Mar;49(3):903-7.
PMID: 15728882

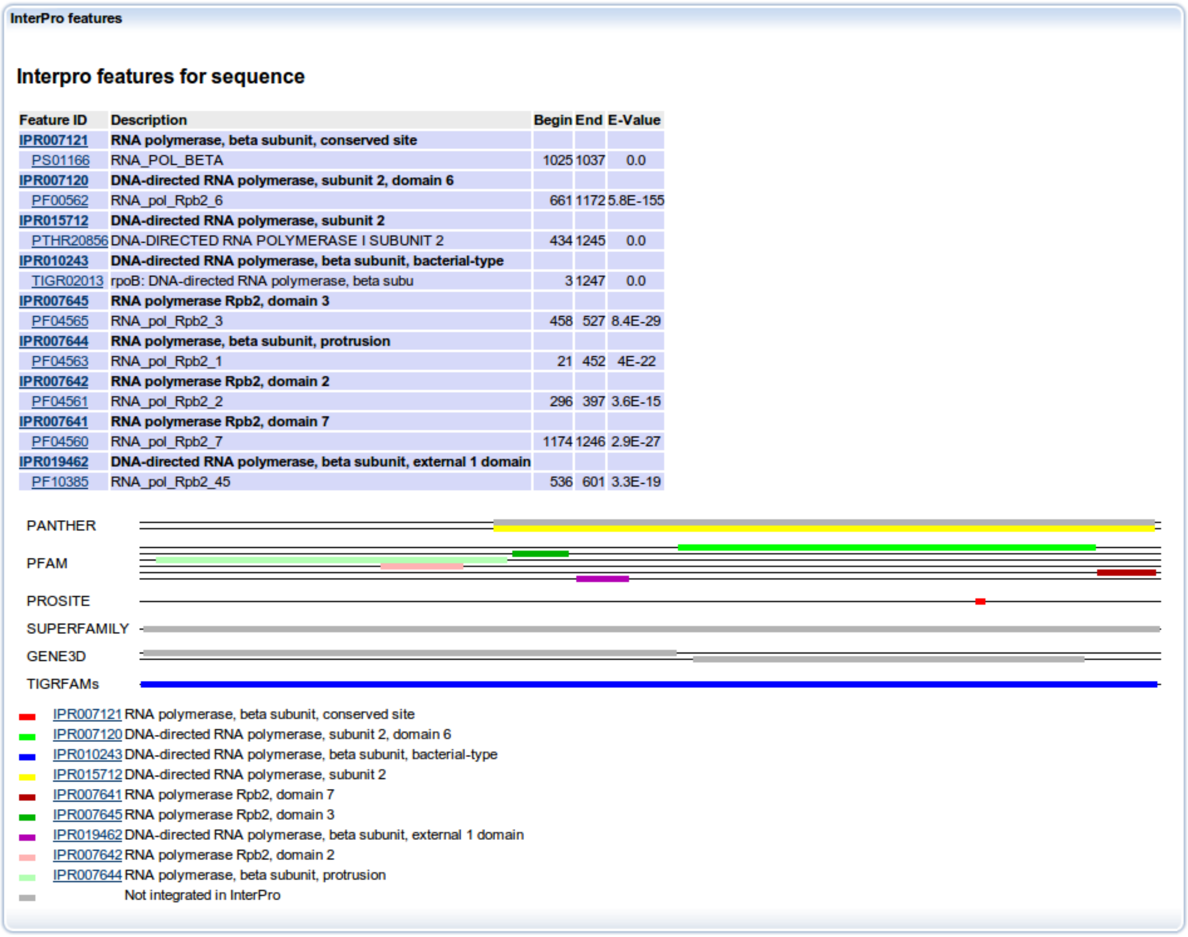
Terms that matched the publication
Chlamydophila pneumoniae
C. pneumoniae
C. pneumoniae TW-183
rpoB

Figure 2.38: Visualization of automatically assigned literature for protein CPn0081 of *Chlamydophila pneumoniae* CWL029 within the ChlamydiaeDB The publications as well as the identified terms within the publications are shown. The list of authors is directly linked to the respective MEDLINE/PubMed entry.

the respective orthologous group (if existing) are below it ordered by descending bitscore in comparison to the selected protein.

The synteny portlet allows to investigate the conservation of the genomic neighborhood of a gene, to detect rearrangements, insertions or deletions by the coloring of the genes. The synteny view is one of the possibilities within the ChlamydiaeDB that easily allows for comparative genomics.

2.3.6.2.3.11 Information about proteins in the same protein family (orthologous group) An orthologous group contains proteins that originate from the same ancestor, come from different *Chlamydiae* species, and probably share the same function. The orthologous group portlet allows to get an overview over information available for the members of an orthologous group (Figure 2.47). The eggNOG cluster for the respective protein is retrieved and all member proteins not belonging to the phylum *Chlamydiae* are removed. Then the available information for each of the proteins is retrieved. Information displayed is whether there exists: manual comments, manual EC annotations, manual FunCat annotations, manual GO annotations, manually annotated literature, transcript data, SNP data, automatically assigned literature, automatic FunCat annotations, automatic GO annotations, automatic EC annotations, predicted type III secreted proteins and whether the protein is contained in UniProtKB/Swiss-Prot. If the information is available then the box is checked, otherwise it is empty. This allows to get an overview over available information in other genomes and by that makes the transfer of knowledge from one organism to the other easily possible. This greatly enhances previously existing genomic resources.



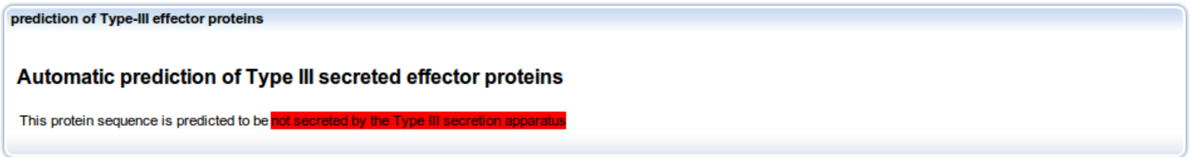
InterPro features

Interpro features for sequence

Feature ID	Description	Begin	End	E-Value
IPR007121	**RNA polymerase, beta subunit, conserved site**			
PS01166	RNA_POL_BETA	1025	1037	0.0
IPR007120	**DNA-directed RNA polymerase, subunit 2, domain 6**			
PF00562	RNA_pol_Rpb2_6	661	1172	5.8E-155
IPR015712	**DNA-directed RNA polymerase, subunit 2**			
PTHR20856	DNA-DIRECTED RNA POLYMERASE I SUBUNIT 2	434	1245	0.0
IPR010243	**DNA-directed RNA polymerase, beta subunit, bacterial-type**			
TIGR02013	rpoB: DNA-directed RNA polymerase, beta subu	3	1247	0.0
IPR007645	**RNA polymerase Rpb2, domain 3**			
PF04565	RNA_pol_Rpb2_3	458	527	8.4E-29
IPR007644	**RNA polymerase, beta subunit, protrusion**			
PF04563	RNA_pol_Rpb2_1	21	452	4E-22
IPR007642	**RNA polymerase Rpb2, domain 2**			
PF04561	RNA_pol_Rpb2_2	296	397	3.6E-15
IPR007641	**RNA polymerase Rpb2, domain 7**			
PF04560	RNA_pol_Rpb2_7	1174	1246	2.9E-27
IPR019462	**DNA-directed RNA polymerase, beta subunit, external 1 domain**			
PF10385	RNA_pol_Rpb2_45	536	601	3.3E-19

🟥 IPR007121 RNA polymerase, beta subunit, conserved site
🟩 IPR007120 DNA-directed RNA polymerase, subunit 2, domain 6
🟦 IPR010243 DNA-directed RNA polymerase, beta subunit, bacterial-type
🟨 IPR015712 DNA-directed RNA polymerase, subunit 2
🟥 IPR007641 RNA polymerase Rpb2, domain 7
🟩 IPR007645 RNA polymerase Rpb2, domain 3
🟪 IPR019462 DNA-directed RNA polymerase, beta subunit, external 1 domain
▨ IPR007642 RNA polymerase Rpb2, domain 2
▨ IPR007644 RNA polymerase, beta subunit, protrusion
▨ Not integrated in InterPro

Figure 2.39: Visualization of InterPro protein domain annotations for protein CPn0081 of *Chlamydophila pneumoniae* CWL029 within the ChlamydiaeDB The InterPro protein domains are displayed as a table on the top and graphically on the bottom. The black lines represent the protein sequence, the colored bars represent domains lying on the sequence.



prediction of Type-III effector proteins

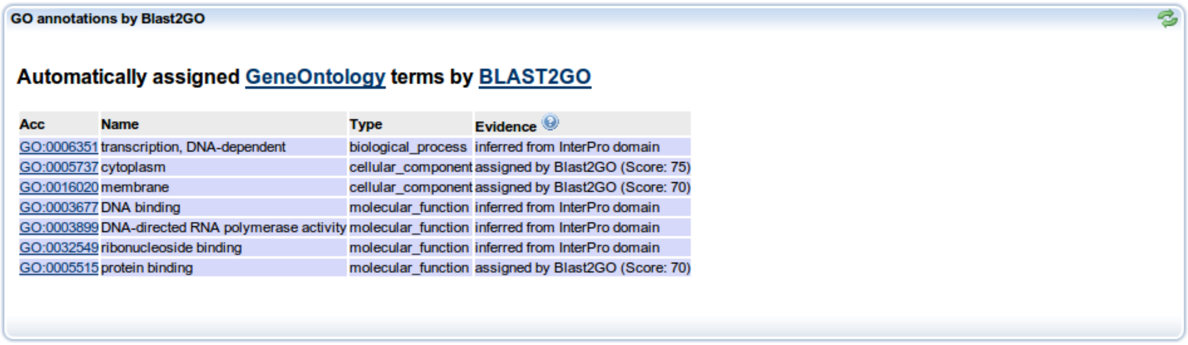
Automatic prediction of Type III secreted effector proteins

This protein sequence is predicted to be not secreted by the Type III secretion apparatus

Figure 2.40: Visualization of the Type-III secretion prediction for CPn0081 of *Chlamydophila pneumoniae* CWL029 within the ChlamydiaeDB The portlet shows whether the protein is predicted to be secreted by the Type-III secretion system or not.

2.3.6.2.4 Manual annotation possibilities Manual annotations of experts are the annotations with highest quality and reliability. Therefore it is essential to provide a possibility for every user to submit novel knowledge to the ChlamydiaeDB. This way the knowledge gained by the usage of ChlamydiaeDB can flow back.

Manual annotation is possible within the protein report page in the manual annotation

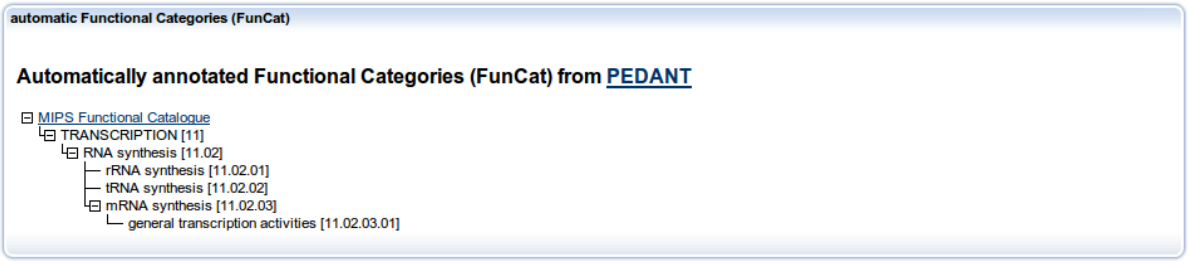


GO annotations by Blast2GO

Automatically assigned GeneOntology terms by BLAST2GO

Acc	Name	Type	Evidence
GO:0006351	transcription, DNA-dependent	biological_process	inferred from InterPro domain
GO:0005737	cytoplasm	cellular_component	assigned by Blast2GO (Score: 75)
GO:0016020	membrane	cellular_component	assigned by Blast2GO (Score: 70)
GO:0003677	DNA binding	molecular_function	inferred from InterPro domain
GO:0003899	DNA-directed RNA polymerase activity	molecular_function	inferred from InterPro domain
GO:0032549	ribonucleoside binding	molecular_function	inferred from InterPro domain
GO:0005515	protein binding	molecular_function	assigned by Blast2GO (Score: 70)

Figure 2.41: Visualization of the automatic Gene Ontology (GO) annotations for CPn0081 of *Chlamydophila pneumoniae* CWL029 within the ChlamydiaeDB The GO annotations are derived by Blast2GO [93].

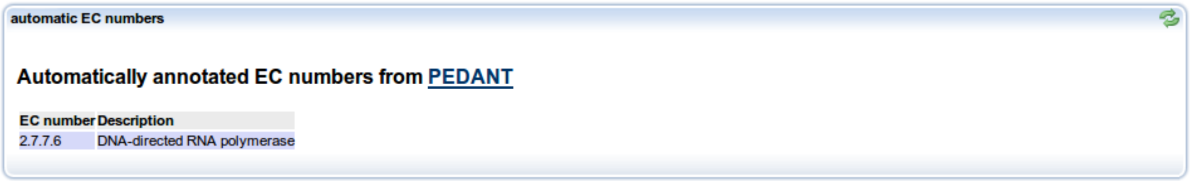


automatic Functional Categories (FunCat)

Automatically annotated Functional Categories (FunCat) from PEDANT

```
⊟ MIPS Functional Catalogue
  └⊟ TRANSCRIPTION [11]
     └⊟ RNA synthesis [11.02]
        ├ rRNA synthesis [11.02.01]
        ├ tRNA synthesis [11.02.02]
        └⊟ mRNA synthesis [11.02.03]
           └ general transcription activities [11.02.03.01]
```

Figure 2.42: Visualization of the automatic Functional Catalogue (FunCat) annotations for CPn0081 of *Chlamydophila pneumoniae* CWL029 within the ChlamydiaeDB The portlet shows the FunCat annotations as derived from PEDANT [72] in an interactive tree.

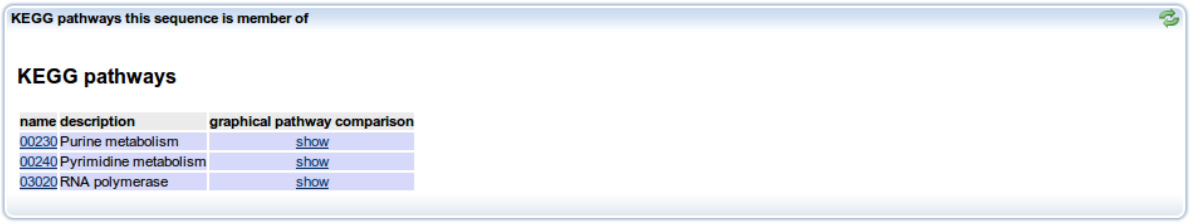


automatic EC numbers

Automatically annotated EC numbers from PEDANT

EC number	Description
2.7.7.6	DNA-directed RNA polymerase

Figure 2.43: Visualization of the automatic EC annotations for CPn0081 of *Chlamydophila pneumoniae* CWL029 within the ChlamydiaeDB The portlet shows the EC annotations as derived from PEDANT [72] in a table.

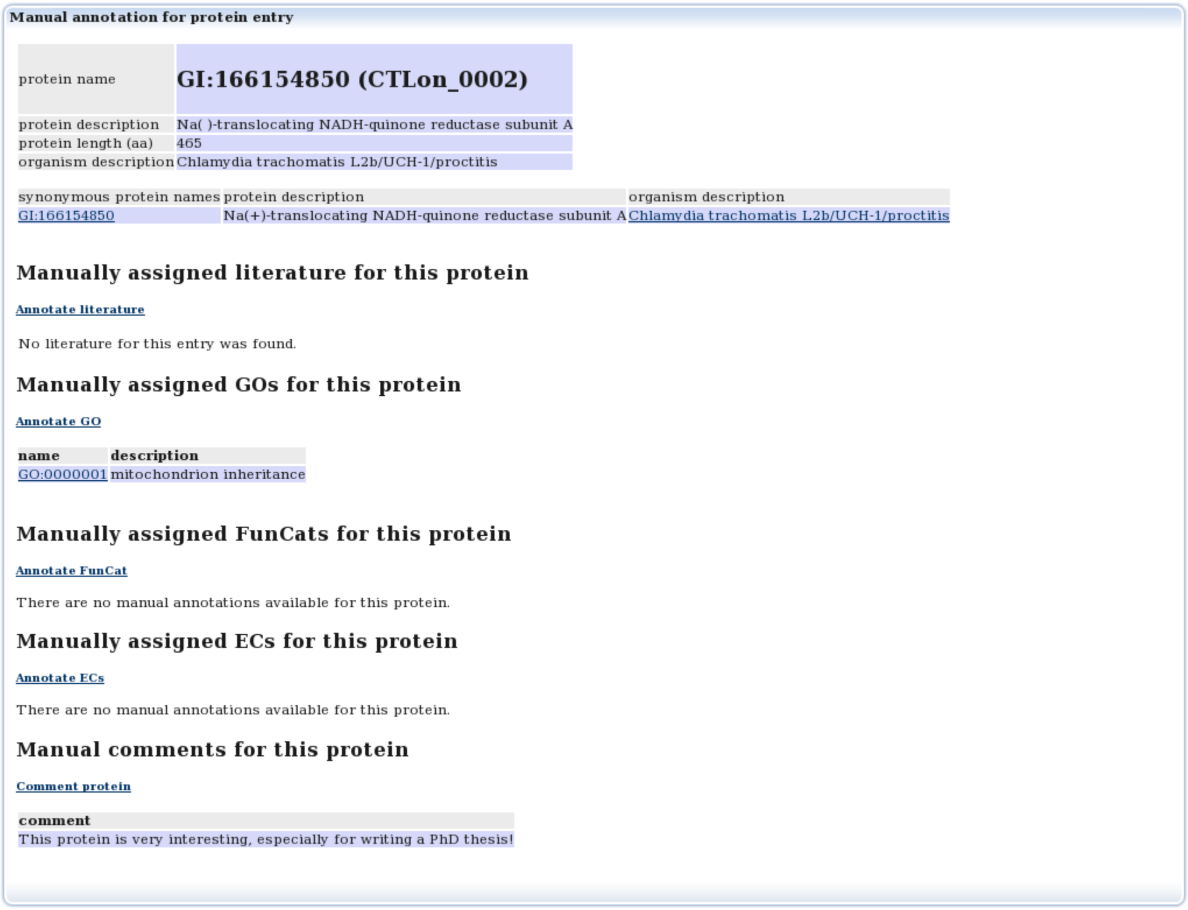


KEGG pathways this sequence is member of

KEGG pathways

name	description	graphical pathway comparison
00230	Purine metabolism	show
00240	Pyrimidine metabolism	show
03020	RNA polymerase	show

Figure 2.44: Visualization of the KEGG pathway annotations for CPn0081 of *Chlamydophila pneumoniae* CWL029 within the ChlamydiaeDB The assignment of the protein to different KEGG pathways is displayed. It is possible to directly show the affected map in the graphical pathway comparison (2.32).



Manual annotation for protein entry

protein name	**GI:166154850 (CTLon_0002)**
protein description	Na( )-translocating NADH-quinone reductase subunit A
protein length (aa)	465
organism description	Chlamydia trachomatis L2b/UCH-1/proctitis

synonymous protein names	protein description	organism description
GI:166154850	Na(+)-translocating NADH-quinone reductase subunit A	Chlamydia trachomatis L2b/UCH-1/proctitis

Manually assigned literature for this protein

Annotate literature

No literature for this entry was found.

Manually assigned GOs for this protein

Annotate GO

name	description
GO:0000001	mitochondrion inheritance

Manually assigned FunCats for this protein

Annotate FunCat

There are no manual annotations available for this protein.

Manually assigned ECs for this protein

Annotate ECs

There are no manual annotations available for this protein.

Manual comments for this protein

Comment protein

comment
This protein is very interesting, especially for writing a PhD thesis!

Figure 2.45: Display of imaginary manual annotations for CTLon_0002 of *Chlamydia trachomatis* L2b/UCH-1/proctitis within the ChlamydiaeDB Manual annotations if already accepted by an administrator can be seen. These annotations include literature, GO, FunCat, EC and comments. Please note that these annotations are imaginary.

portlet (Figure 2.45). Besides viewing manual annotations for each kind of supported information the possibility to suggest new manual annotations is given.

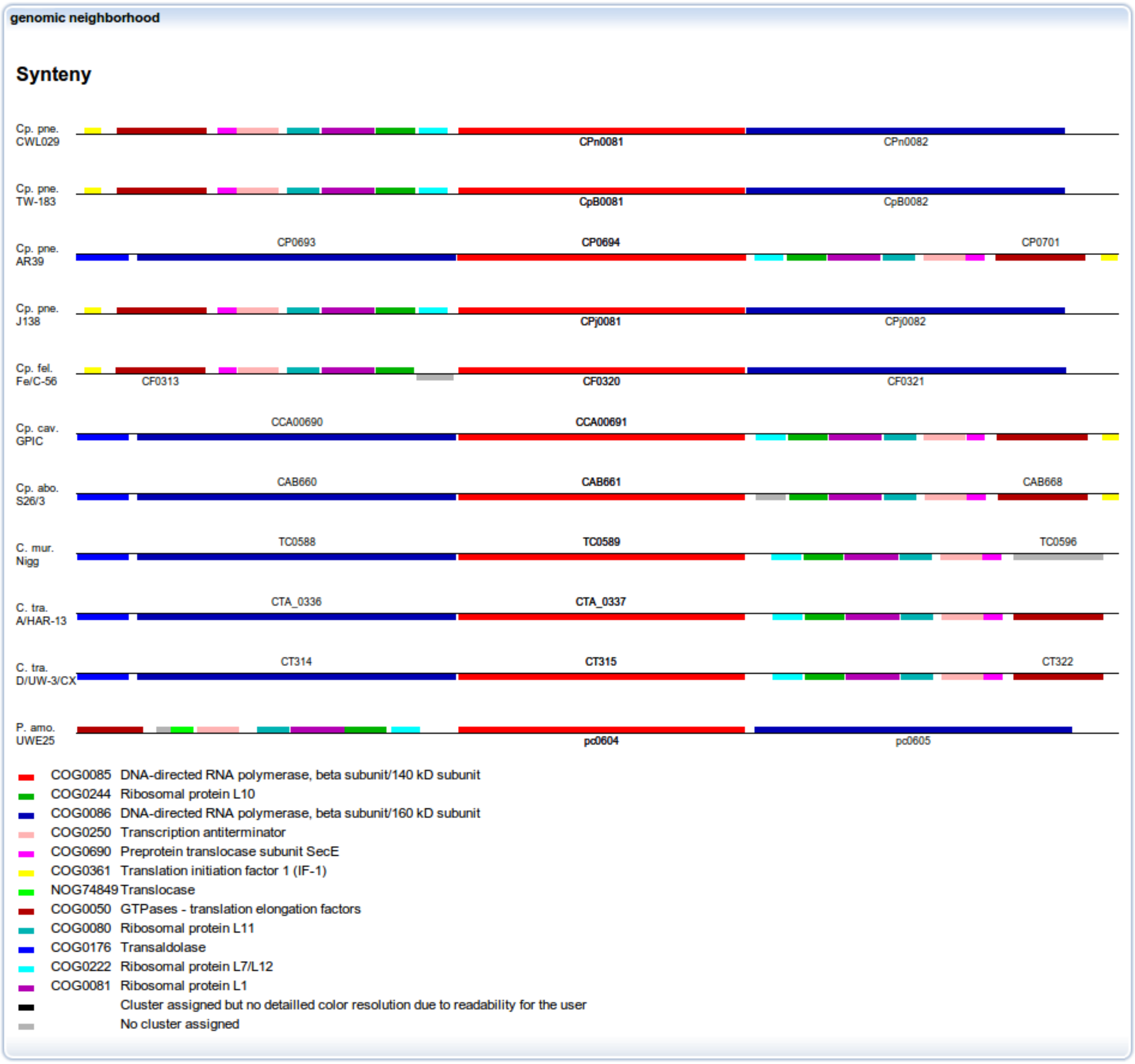
In order to avoid SPAM and nonsense annotation I implemented a multi step procedure for the annotation (Figures 2.48, 2.49):

1. Starting point is the "manual annotation for protein entry" portlet (Figure 2.48 A, example Figure 2.49 A)

2. A ChlamydiaeDB user selects the type of annotation and enters the annotation (Figure 2.48 B , example Figure 2.49 B)

3. The user submits the annotation (Figure 2.48 C , example Figure 2.49 C)

4. An annotation administrator checks the annotation for validity and conclusiveness and can accept or reject the annotation (Figure 2.48 D, example Figure 2.49 D)



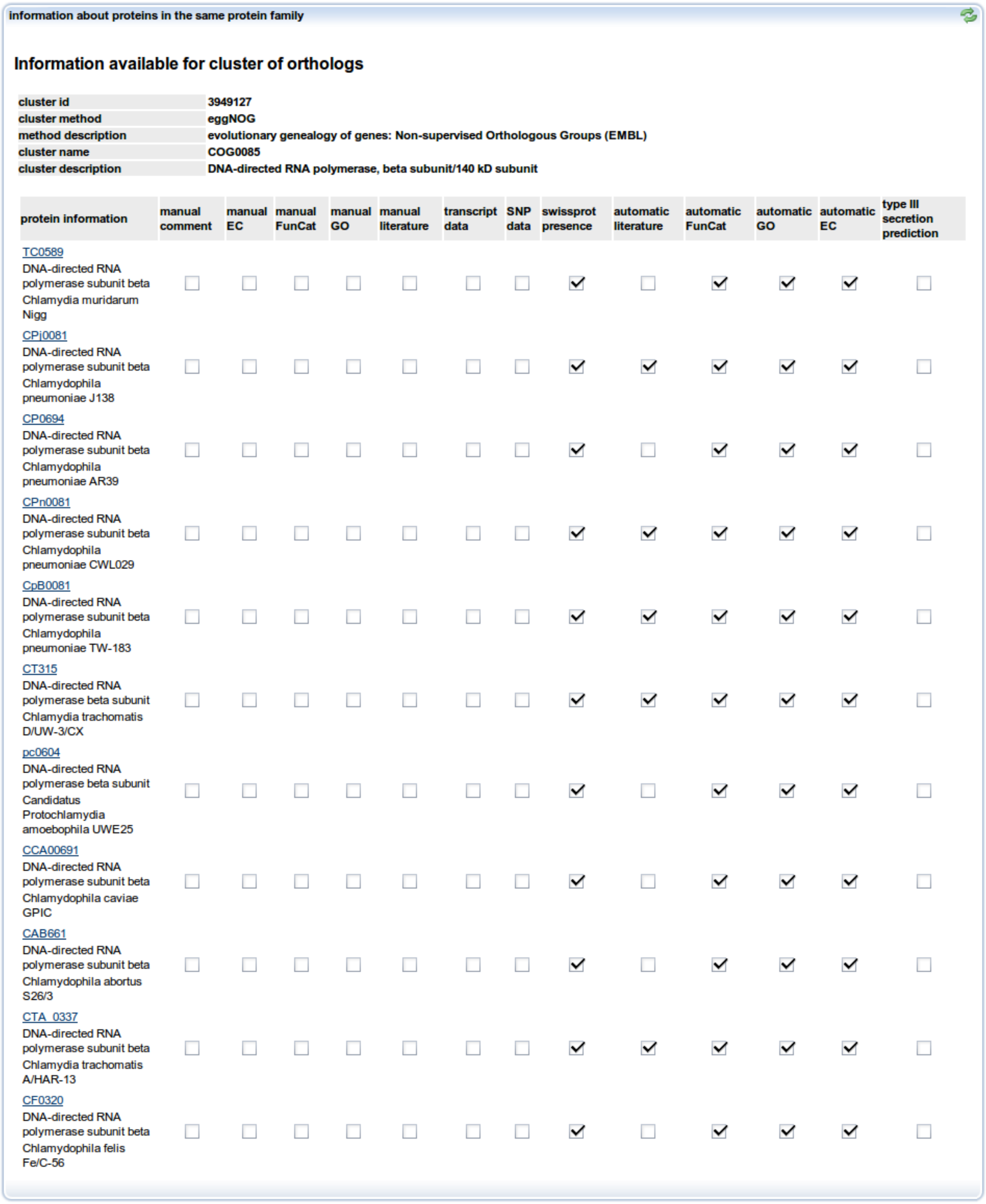
🟥	COG0085	DNA-directed RNA polymerase, beta subunit/140 kD subunit
🟩	COG0244	Ribosomal protein L10
🟦	COG0086	DNA-directed RNA polymerase, beta subunit/160 kD subunit
■	COG0250	Transcription antiterminator
■	COG0690	Preprotein translocase subunit SecE
🟨	COG0361	Translation initiation factor 1 (IF-1)
🟩	NOG74849	Translocase
🟥	COG0050	GTPases - translation elongation factors
■	COG0080	Ribosomal protein L11
🟦	COG0176	Transaldolase
■	COG0222	Ribosomal protein L7/L12
🟪	COG0081	Ribosomal protein L1
⬛		Cluster assigned but no detailled color resolution due to readability for the user
⬜		No cluster assigned

Figure 2.46: Visualization of the genomic neighborhood of CPn0081 of *Chlamydophila pneumoniae* CWL029 within the ChlamydiaeDB Each of the lines of the genomic neighborhood represents a stretch of DNA in one of the chlamydial genomes. The colored bars on them are genes, colored by their membership in eggNOG clusters. The selected gene CPn0081 of *Chlamydophila pneumoniae* CWL029 is shown in the first line in the middle colored in red. The other members of the respective orthologous group are below ordered by descending bitscore in comparison to CPn0081.

5. If the annotation administrator accepted the annotation then the annotation will be visible in the protein report (Figure 2.48 E, example Figure 2.49 E)

The ChlamydiaeDB user is informed about the current status of the annotation by email. That way it is also transparent when and why an annotation was accepted or rejected.

It is part of the concept to keep it quite easy to submit annotations without the need



Information available for cluster of orthologs

cluster id	3949127
cluster method	eggNOG
method description	evolutionary genealogy of genes: Non-supervised Orthologous Groups (EMBL)
cluster name	COG0085
cluster description	DNA-directed RNA polymerase, beta subunit/140 kD subunit

protein information	manual comment	manual EC	manual FunCat	manual GO	manual literature	transcript data	SNP data	swissprot presence	automatic literature	automatic FunCat	automatic GO	automatic EC	type III secretion prediction
TC0589 DNA-directed RNA polymerase subunit beta Chlamydia muridarum Nigg	□	□	□	□	□	□	□	☑	□	☑	☑	☑	□
CPj0081 DNA-directed RNA polymerase subunit beta Chlamydophila pneumoniae J138	□	□	□	□	□	□	□	☑	☑	☑	☑	☑	□
CP0694 DNA-directed RNA polymerase subunit beta Chlamydophila pneumoniae AR39	□	□	□	□	□	□	□	☑	□	☑	☑	☑	□
CPn0081 DNA-directed RNA polymerase subunit beta Chlamydophila pneumoniae CWL029	□	□	□	□	□	□	□	☑	☑	☑	☑	☑	□
CpB0081 DNA-directed RNA polymerase subunit beta Chlamydophila pneumoniae TW-183	□	□	□	□	□	□	□	☑	☑	☑	☑	☑	□
CT315 DNA-directed RNA polymerase beta subunit Chlamydia trachomatis D/UW-3/CX	□	□	□	□	□	□	□	☑	☑	☑	☑	☑	□
pc0604 DNA-directed RNA polymerase beta subunit Candidatus Protochlamydia amoebophila UWE25	□	□	□	□	□	□	□	☑	□	☑	☑	☑	□
CCA00691 DNA-directed RNA polymerase subunit beta Chlamydophila caviae GPIC	□	□	□	□	□	□	□	☑	□	☑	☑	☑	□
CAB661 DNA-directed RNA polymerase subunit beta Chlamydophila abortus S26/3	□	□	□	□	□	□	□	☑	□	☑	☑	☑	□
CTA_0337 DNA-directed RNA polymerase subunit beta Chlamydia trachomatis A/HAR-13	□	□	□	□	□	□	□	☑	☑	☑	☑	☑	□
CF0320 DNA-directed RNA polymerase subunit beta Chlamydophila felis Fe/C-56	□	□	□	□	□	□	□	☑	□	☑	☑	☑	□

Figure 2.47: Overview over information available in the orthologous group of CPn0081 of *Chlamydophila pneumoniae* CWL029 within the ChlamydiaeDB The members of the orthologous group are shown in the rows, the kinds of information available for the members of the orthologous groups are shown in the columns. Information displayed is whether there exists: manual comments, manual EC annotations, manual FunCat annotations, manual GO annotations, manually annotated literature, transcript data, SNP data, automatically assigned literature, automatic FunCat annotations, automatic GO annotations, automatic EC annotations, predicted type III secreted proteins and whether the protein is contained in UniProtKB/Swiss-Prot. If the information is available then the box is checked, otherwise it is empty.

to register in order to keep the gateway hurdle as low as possible.

2.3.6.2.5 Tools for the analysis of user defined data sets One of the features distinguishing ChlamydiaeDB from many other resources is the possibility for the user to apply specific tools to user defined data sets.

2.3.6.2.5.1 Retrieval of all information about a list of proteins It is possible to get all information available for a set of proteins. The user can specify a list of identifiers or sequences and retrieve a table listing all available information for this set of proteins (Figure 2.50). The available information contains all sequence names, manual comments, manual ECs, manual FunCats, manual GOs, manual literature, information about UniProtKB/Swiss-Prot presence, type III effector prediction, automatically annotated literature, automatically annotated literature within the orthologous group, automatic FunCats from PEDANT, automatic FunCats from PEDANT within the orthologous group, automatic GO annotations from Blast2GO, automatic GO annotations from Blast2GO within the orthologous group, automatic ECs from PEDANT, automatic EC from PEDANT within the orthologous group.

2.3.6.2.5.2 Feature enrichment in a list of proteins A common task is that a set of proteins has been identified in an experiment and it should be determined what these proteins have in common, that is which properties are over- or underrepresented in this set of proteins in comparison to another set of proteins.

The ChlamydiaeDB with its very different kinds of data is predestinated for these questions. Therefore I developed the possibility to define own sets of proteins for every user and to find out which features are enriched or depleted in one set of proteins in comparison to the other set of proteins.

The user can define both sets in the “Find enriched/depleted features in your set of proteins” section available in the main navigation (see e.g. Figure 2.30 on the left side). The first set can be defined by sequence identifiers or protein sequences. The second set can be defined by sequence identifiers, protein sequences or by the selection of one or more chlamydial organisms. It is important to note that necessarily the first set has to be a subset of the second set.

Then for each protein in each of the two sets the following annotations are retrieved:

- Automatic FunCat annotations from PEDANT
- Automatic EC numbers from PEDANT
- Automatic UniProtKB/Swiss-Prot keywords from PEDANT
- InterPro protein domains from SIMAP
- GO annotations from Blast2GO (from SIMAP)
- Type-III secretion predictions

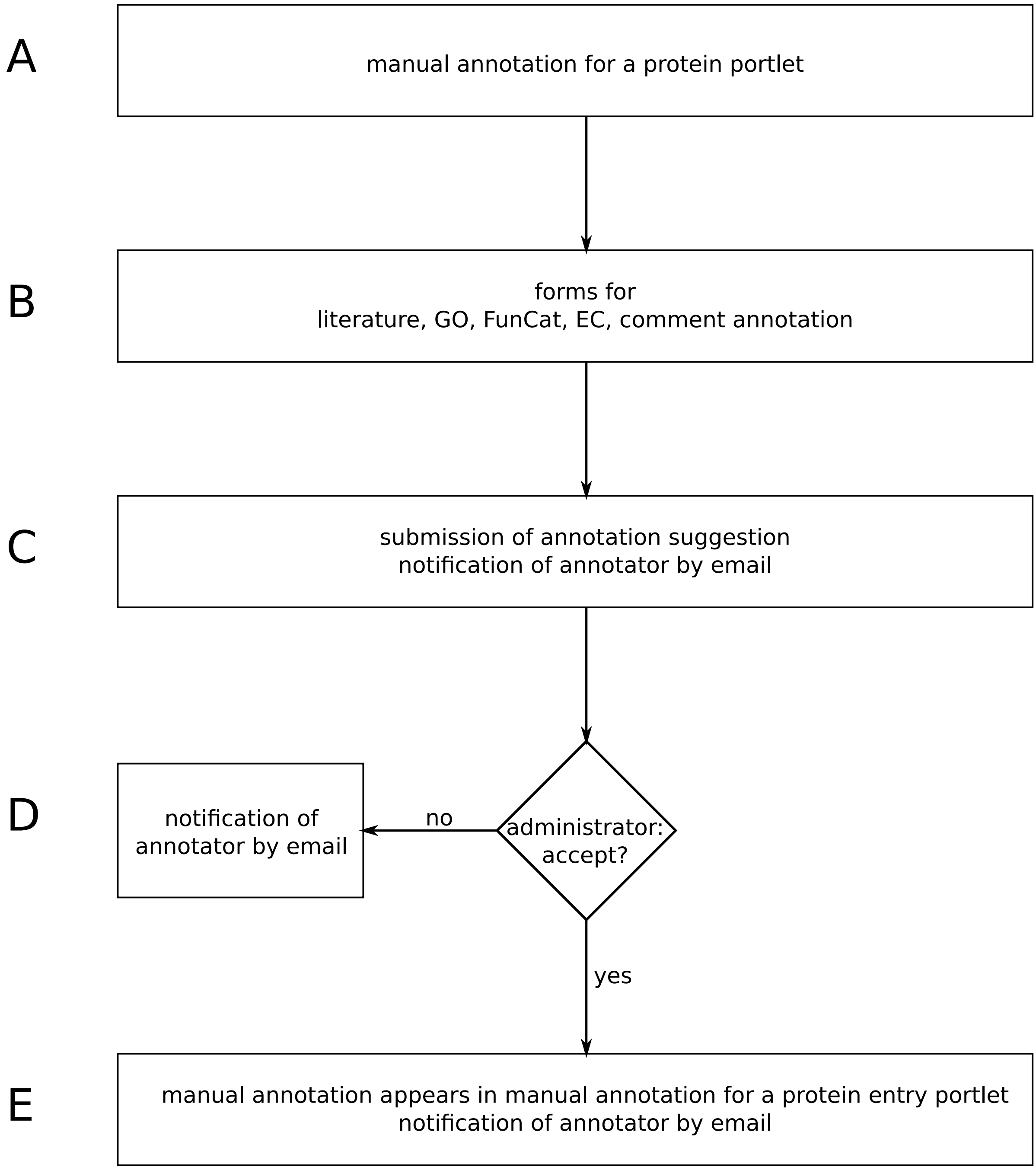
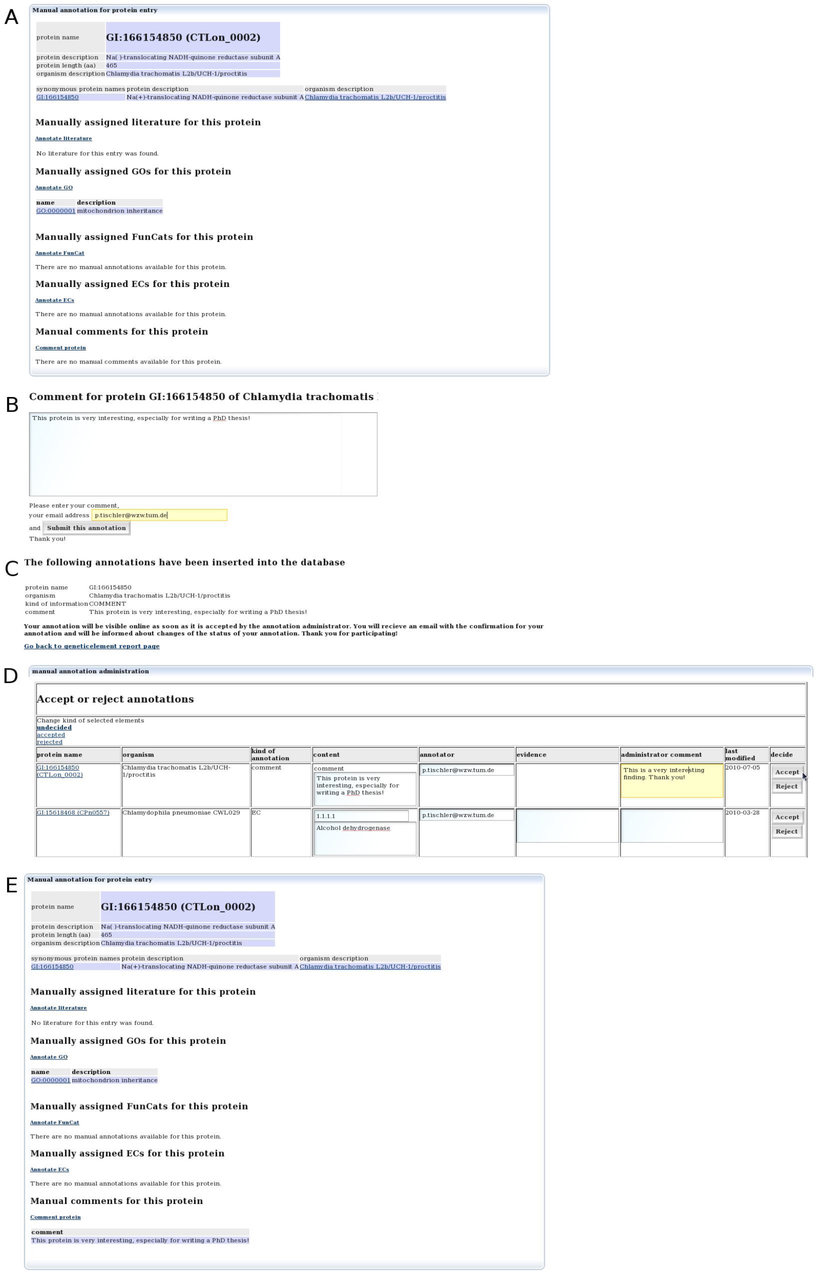

A
manual annotation for a protein portlet
B
forms for
literature, GO, FunCat, EC, comment annotation
C
submission of annotation suggestion
notification of annotator by email
D
notification of annotator by email
no
administrator: accept?
yes
E
manual annotation appears in manual annotation for a protein entry portlet
notification of annotator by email

Figure 2.48: Overview over the manual annotation procedure of the ChlamydiaeDB A: Starting point is the "manual annotation for protein entry" portlet. The user selects the kind of annotation that should be annotated. **B:** The user enters annotations into the respective form for the previously selected kind of annotation. **C:** The annotations are submitted and the user gets an email confirming the annotation. **D:** An annotation administrator checks the annotation for validity and conclusiveness and accepts or rejects the annotation. The annotator has a comment field so that the user knows why an annotation was accepted or rejected. The decision is automatically sent by email to the annotator. **E:** After acception of the annotation it is visible in the "manual annotation for a protein entry" portlet on the protein report page.



Manual annotation for protein entry

protein name	GI:166154850 (CTLon_0002)
protein description	Na( )-translocating NADH-quinone reductase subunit A
protein length (aa)	465
organism description	Chlamydia trachomatis L2b/UCH-1/proctitis

synonymous protein names	protein description	organism description
GI:166154850	Na(+)-translocating NADH-quinone reductase subunit A	Chlamydia trachomatis L2b/UCH-1/proctitis

Manually assigned literature for this protein

Annotate literature

No literature for this entry was found.

Manually assigned GOs for this protein

Annotate GO

name	description
GO:0000001	mitochondrion inheritance

Manually assigned FunCats for this protein

Annotate FunCat

There are no manual annotations available for this protein.

Manually assigned ECs for this protein

Annotate ECs

There are no manual annotations available for this protein.

Manual comments for this protein

Comment protein

There are no manual comments available for this protein.

Comment for protein GI:166154850 of Chlamydia trachomatis

This protein is very interesting, especially for writing a PhD thesis!

Please enter your comment,
your email address p.tischler@wzw.tum.de
and Submit this annotation
Thank you!

The following annotations have been inserted into the database

protein name	GI:166154850
organism	Chlamydia trachomatis L2b/UCH-1/proctitis
kind of information	COMMENT
comment	This protein is very interesting, especially for writing a PhD thesis!

Your annotation will be visible online as soon as it is accepted by the annotation administrator. You will recieve an email with the confirmation for your annotation and will be informed about changes of the status of your annotation. Thank you for participating!

Go back to geneticelement report page

manual annotation administration

Accept or reject annotations

Change kind of selected elements
undecided
accepted
rejected

protein name	organism	kind of annotation	content	annotator	evidence	administrator comment	last modified	decide
GI:166154850 (CTLon_0002)	Chlamydia trachomatis L2b/UCH-1/proctitis	comment	comment This protein is very interesting, especially for writing a PhD thesis!	p.tischler@wzw.tum.de		This is a very interesting finding. Thank you!	2010-07-05	Accept Reject
GI:15618468 (CPn0557)	Chlamydophila pneumoniae CWL029	EC	1.1.1.1 Alcohol dehydrogenase	p.tischler@wzw.tum.de			2010-03-28	Accept Reject

Manual annotation for protein entry

protein name	GI:166154850 (CTLon_0002)
protein description	Na( )-translocating NADH-quinone reductase subunit A
protein length (aa)	465
organism description	Chlamydia trachomatis L2b/UCH-1/proctitis

synonymous protein names	protein description	organism description
GI:166154850	Na(+)-translocating NADH-quinone reductase subunit A	Chlamydia trachomatis L2b/UCH-1/proctitis

Manually assigned literature for this protein

Annotate literature

No literature for this entry was found.

Manually assigned GOs for this protein

Annotate GO

name	description
GO:0000001	mitochondrion inheritance

Manually assigned FunCats for this protein

Annotate FunCat

There are no manual annotations available for this protein.

Manually assigned ECs for this protein

Annotate ECs

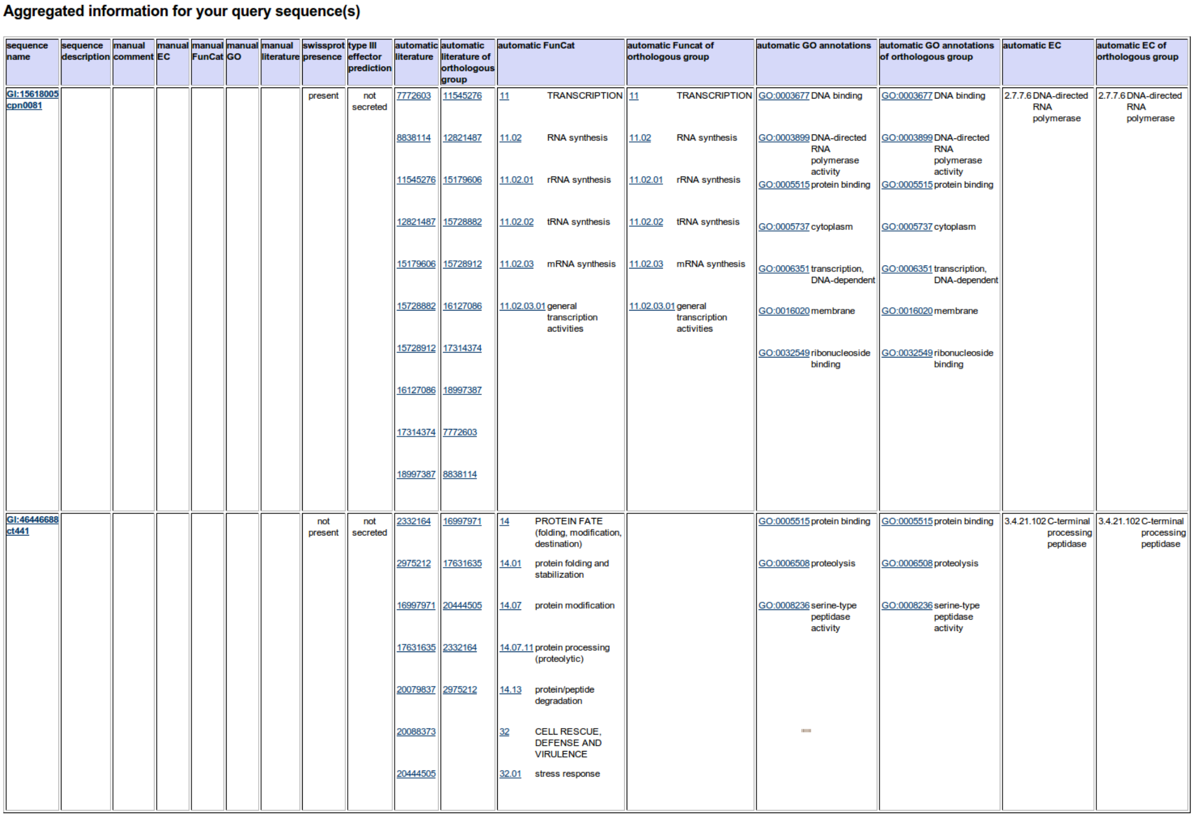
There are no manual annotations available for this protein.

Manual comments for this protein

Comment protein

comment
This protein is very interesting, especially for writing a PhD thesis!

Figure 2.49: Example of manual annotation of protein CTLon_0002 of *Chlamydia trachomatis* L2b/UCH-1/proctitis within the ChlamydiaeDB The example shows the submission of a comment, its acception by an administration administrator and how the comment is displayed after acception. **A:** Starting point is the "manual annotation for protein entry" portlet. There is no comment annotated yet. The user selects the "Comment protein" link. **B:** The user enters a comment. **C:** The user submits the comment. A page showing the information entered is displayed and the user gets an email confirming the annotation.**D:** An annotation administrator checks the annotation for validity and conclusiveness and accepts the annotation. The annotator can attach a comment to his decision so that the user knows why an annotation was accepted or rejected. The decision is automatically sent by email to the annotator. **E:** After acception of the annotation it is visible in the "manual annotation for a protein entry" portlet on the protein report



Aggregated information for your query sequence(s)

sequence name	sequence description	manual comment	manual EC	manual FunCat	manual GO	manual literature	swissprot presence	type III effector prediction	automatic literature	automatic literature of orthologous group	automatic FunCat	automatic Funcat of orthologous group	automatic GO annotations	automatic GO annotations of orthologous group	automatic EC	automatic EC of orthologous group
GI:15618005 cpn0081							present	not secreted	7772603 8838114 11545276 12821487 15179606 15728882 15728912 16127086 17314374 18997387	11545276 12821487 15179606 15728882 15728912 16127086 17314374 18997387 7772603 8838114	11 TRANSCRIPTION 11.02 RNA synthesis 11.02.01 rRNA synthesis 11.02.02 tRNA synthesis 11.02.03 mRNA synthesis 11.02.03.01 general transcription activities	11 TRANSCRIPTION 11.02 RNA synthesis 11.02.01 rRNA synthesis 11.02.02 tRNA synthesis 11.02.03 mRNA synthesis 11.02.03.01 general transcription activities	GO:0003677 DNA binding GO:0003899 DNA-directed RNA polymerase activity GO:0005515 protein binding GO:0005737 cytoplasm GO:0006351 transcription, DNA-dependent GO:0016020 membrane GO:0032549 ribonucleoside binding	GO:0003677 DNA binding GO:0003899 DNA-directed RNA polymerase activity GO:0005515 protein binding GO:0005737 cytoplasm GO:0006351 transcription, DNA-dependent GO:0016020 membrane GO:0032549 ribonucleoside binding	2.7.7.6 DNA-directed RNA polymerase	2.7.7.6 DNA-directed RNA polymerase
GI:46446688 ct441							not present	not secreted	2332164 2975212 16997971 17631635 20079837 20088373 20444505	16997971 17631635 20444505 2332164 2975212	14 PROTEIN FATE (folding, modification, destination) 14.01 protein folding and stabilization 14.07 protein modification 14.07.11 protein processing (proteolytic) 14.13 protein/peptide degradation 32 CELL RESCUE, DEFENSE AND VIRULENCE 32.01 stress response		GO:0005515 protein binding GO:0006508 proteolysis GO:0008236 serine-type peptidase activity	GO:0005515 protein binding GO:0006508 proteolysis GO:0008236 serine-type peptidase activity	3.4.21.102 C-terminal processing peptidase	3.4.21.102 C-terminal processing peptidase

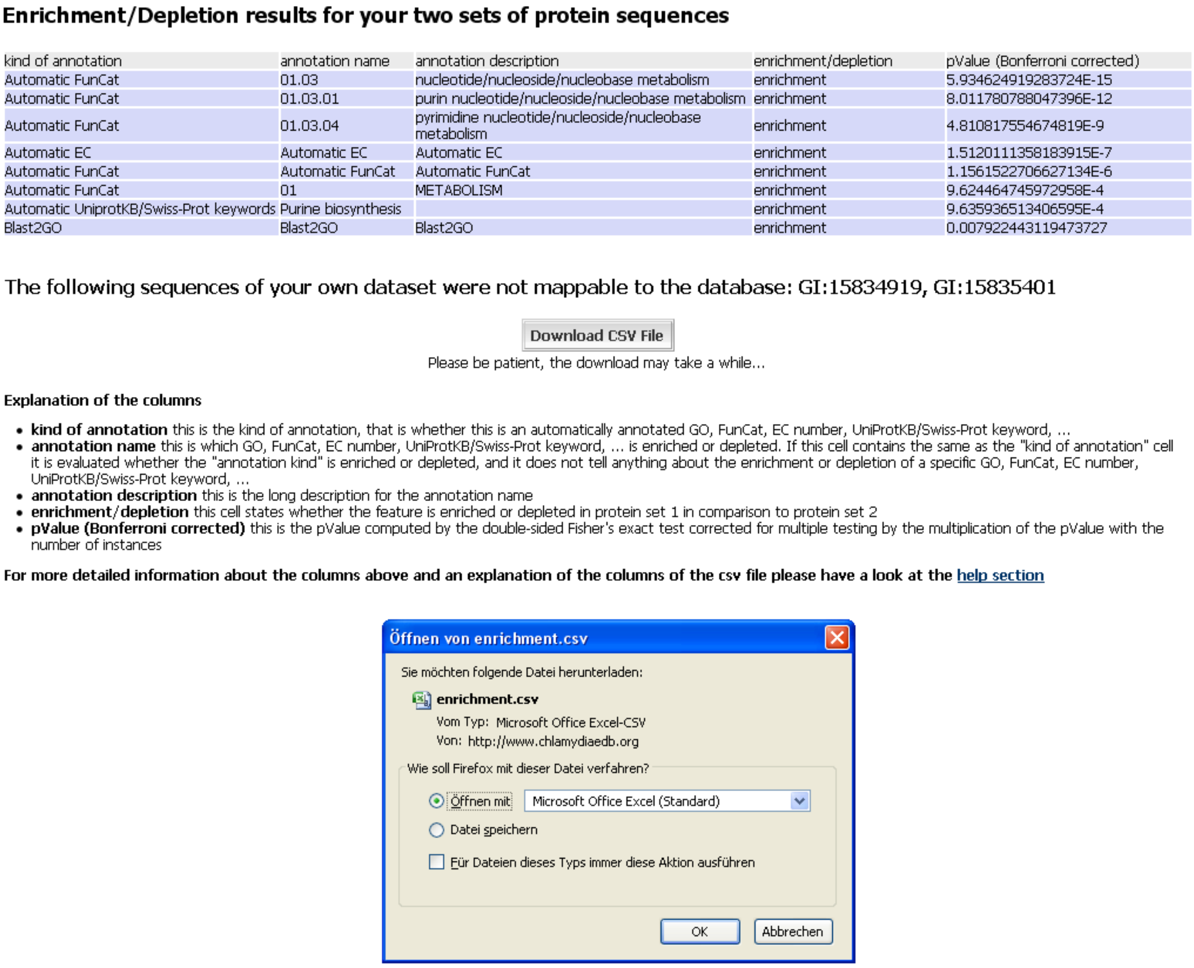
Figure 2.50: Aggregated information for CPn0081 of *Chlamydophila pneumoniae* CWL029 and CT441 of *Chlamydia trachomatis* D/UW-3/CX within the ChlamydiaeDB

For each of these points (FunCat, EC, UniProtKB/Swiss-Prot keywords, InterPro domains, GO, Type-III) and for each of the single annotations within these points (e.g. Funcat 01, Funcat 01.01, ...) the following counts are created:

- number of proteins in set 1 having this annotation
- number of proteins in set 1 not having this annotation
- number of proteins in set 2 having this annotation
- number of proteins in set 2 not having this annotation

Then a two-tailed Fisher's exact test with Bonferroni correction is applied (section 1.7.4) in order to detect significantly enriched or depleted annotations. Annotations with a corrected p-Value ≤ 0.01 have been defined as significantly enriched or depleted.

After the determination of significant enrichments and depletions the user gets a view on the web site showing the results. It is also possible to download the results as an Excel file for further processing (Figure 2.51).



Enrichment/Depletion results for your two sets of protein sequences

kind of annotation	annotation name	annotation description	enrichment/depletion	pValue (Bonferroni corrected)
Automatic FunCat	01.03	nucleotide/nucleoside/nucleobase metabolism	enrichment	5.934624919283724E-15
Automatic FunCat	01.03.01	purin nucleotide/nucleoside/nucleobase metabolism	enrichment	8.011780788047396E-12
Automatic FunCat	01.03.04	pyrimidine nucleotide/nucleoside/nucleobase metabolism	enrichment	4.810817554674819E-9
Automatic EC	Automatic EC	Automatic EC	enrichment	1.5120111358183915E-7
Automatic FunCat	Automatic FunCat	Automatic FunCat	enrichment	1.1561522706627134E-6
Automatic FunCat	01	METABOLISM	enrichment	9.624464745972958E-4
Automatic UniprotKB/Swiss-Prot keywords	Purine biosynthesis		enrichment	9.635936513406595E-4
Blast2GO	Blast2GO	Blast2GO	enrichment	0.007922443119473727

The following sequences of your own dataset were not mappable to the database: GI:15834919, GI:15835401

Please be patient, the download may take a while...

Explanation of the columns

- **kind of annotation** this is the kind of annotation, that is whether this is an automatically annotated GO, FunCat, EC number, UniProtKB/Swiss-Prot keyword, ...
- **annotation name** this is which GO, FunCat, EC number, UniProtKB/Swiss-Prot keyword, ... is enriched or depleted. If this cell contains the same as the "kind of annotation" cell it is evaluated whether the "annotation kind" is enriched or depleted, and it does not tell anything about the enrichment or depletion of a specific GO, FunCat, EC number, UniProtKB/Swiss-Prot keyword, ...
- **annotation description** this is the long description for the annotation name
- **enrichment/depletion** this cell states whether the feature is enriched or depleted in protein set 1 in comparison to protein set 2
- **pValue (Bonferroni corrected)** this is the pValue computed by the double-sided Fisher's exact test corrected for multiple testing by the multiplication of the pValue with the number of instances

For more detailed information about the columns above and an explanation of the columns of the csv file please have a look at the help section

Öffnen von enrichment.csv

Sie möchten folgende Datei herunterladen:

enrichment.csv

Vom Typ: Microsoft Office Excel-CSV

Von: http://www.chlamydiaedb.org

Wie soll Firefox mit dieser Datei verfahren?

● Öffnen mit: Microsoft Office Excel (Standard)

○ Datei speichern

□ Für Dateien dieses Typs immer diese Aktion ausführen

Figure 2.51: Enrichment analysis within the ChlamydiaeDB The enrichment and depletion of various features in an example dataset in comparison to another dataset is shown. It is also possible to download the results as a csv file that can be further processed in Microsoft Excel or OpenOffice.org Calc.

2.3.6.2.5.3 Graphical pathway comparison between organisms The graphical pathway comparison allows the user to easily compare the metabolic capabilities of up to three organisms (Figure 2.32). This has been described previously (section 2.3.6.1.2.6).

2.3.6.3 Technical

2.3.6.3.1 Initialization of ChlamydiaeDB The initialization of the ChlamydiaeDB is fully automatic and ensures that the available knowledge is integrated.

The following steps are performed:

- retrieval and insertion of all complete chlamydial genomes within RefSeq from PEDANT
- retrieval and insertion of genetic element names from RefSeq and UniProtKB

- computation and insertion of Type-III secretion predictions
- retrieval and insertion of the GO index
- retrieval of clusters and assignment of organisms not contained in the clusters to the clusters
- parsing and inserting the available literature from MEDLINE/PubMed

2.3.6.3.2 Up-to-dateness with little manual effort It is easy to keep ChlamydiaeDB up-to-date. There are two levels of updates, daily and manual updates.

2.3.6.3.2.1 Daily updates Daily updates contain the search for current literature about *Chlamydiae* and the update of the cached information within the data warehouse.

2.3.6.3.2.2 Manual updates These updates are dependent on the update of other systems like PEDANT or SIMAP. They can be activated if needed.

These updates comprise:

- The search and insertion of novel available chlamydial genomes
- The search and insertion of novel sequences within already contained genomes
- The search for changed entries within RefSeq and UniProtKB
- The prediction of Type-III secreted proteins when a new version of the software is released

2.3.6.3.3 Easy extensibility Each page in the portal can easily be changed. No longer needed portlets can be removed, new portlets can be added and portlets can be dragged and dropped where they should be displayed.

2.3.7 Stored data in ChlamydiaeDB

2.3.7.1 Genomic sequences

The ChlamydiaeDB contains 16 publicly available complete genomes of the phylum *Chlamydiae* at the moment (Table 2.16). *Candidatus Protochlamydia amoebophila* UWE25 is the only environmental chlamydia, the others are pathogenic.

2.3.7.2 Automatic annotations

ChlamydiaeDB offers the following automatic annotations:

- annotations from PEDANT (FunCat, EC, UniProtKB/Swiss-Prot keywords)
- GO annotations from Blast2GO

organism	number proteins
Candidatus Protochlamydia amoebophila UWE25	2030
Chlamydia muridarum Nigg	911
Chlamydia trachomatis 434/Bu	874
Chlamydia trachomatis A/HAR-13	919
Chlamydia trachomatis B/TZ1A828/OT	880
Chlamydia trachomatis D/UW-3/CX	895
Chlamydia trachomatis Jali20	883
Chlamydia trachomatis L2b/UCH-1/proctitis	874
Chlamydophila abortus S26/3	932
Chlamydophila caviae GPIC	1005
Chlamydophila felis Fe/C-56	1013
Chlamydophila pneumoniae AR39	1112
Chlamydophila pneumoniae CWL029	1052
Chlamydophila pneumoniae J138	1069
Chlamydophila pneumoniae LPCoLN	1105
Chlamydophila pneumoniae TW-183	1113
	$\sum = 16667$

Table 2.16: Contained genomes within ChlamydiaeDB

- automatically determined literature relevant for *Chlamydiae*
- eggNOG clusters of orthologous groups
- KEGG metabolic pathways

2.3.7.3 Data from experiments

Different kinds of data from experiments have already been included into the web portal. The capability to present proteome data has been implemented but there is no data available yet.

2.3.7.3.1 SNP data SNP positions are available for *Chlamydophila pneunomiae* CWL029. 14 isolates are included from Rattei et al. [196]. Another 3 isolates have been imported from RefSeq [52], namely *Chlamydophila pneumoniae* AR39, *Chlamydophila pneumoniae* J138, *Chlamydophila pneumoniae* TW183.

2.3.7.3.2 Transcript data Transcription data for *Chlamydia trachomatis* L2b/UCH-1/proctitis has been included from Albrecht et al. [333].

2.3.8 Application

The ChlamydiaeDB is used by many researchers in the *Chlamydiae* field. On average 300 users access ChlamydiaeDB every month (Table 2.52).

Different features of ChlamydiaeDB are used by researchers. Astrid Horn from the Department of Microbial Ecology from the University of Vienna is using the possibility to get an overview over proteins of interest and Hector Alex Saka from the Duke University Medical Center in Durham USA is using the possibility to retrieve aggregated information for his current research.

In the Masterthesis of Jonathan Hoser the enrichment possibilities have been used for the detection of properties of genes affected by synonymous and nonsynonymous SNPs in *Chlamydiae*. In the genes containing nonsynonymous SNPs proteins with FunCats related to ribosomal proteins and ribosomal biogenesis were significantly depleted, the GO term “GO:0044425 membrane part” was significantly enriched. This is in accordance with the observation that ribosomal proteins are conserved very well throughout all organisms.

2.3.9 Discussion

ChlamydiaeDB is a comprehensive online genome database for members of the phylum *Chlamydiae*. It is available at `http://www.ChlamydiaeDB.org`, provides all available kinds of data for all chlamydial genomes in one place, provides access to information from literature, and provides data from experiments like single nucleotide polymorphism (SNP) data, transcript data, and data from proteome experiments. Annotations for a protein as well as information connected to proteins in the same orthologous group or in the neighborhood on the genome are contained and are instantly visible for the user. The user is supported by tools for the retrieval of all information for a list of proteins, the statistical enrichment and depletion of annotations in a list of proteins in comparison to another list of proteins, and a graphical metabolic pathway comparison between organisms. Every user can add manual annotations to the database, that is always up-to-date with the primary resources, easily maintainable and extensible.

It was agreed that ChlamydiaeDB will be the database used and maintained by the community in the future, at the conference of the Chlamydia Basic Research Society (CBRS) in Little Rock USA in March 2009. A survey showed that 29% of the scientists at the conference thought that the re-annotation of chlamydial genomes is essential and 65% thought that it is very important (Figure 2.17 B). On the other hand only 24% of the scientists would be willing to continuously contribute to the re-annotation and 13% would contribute within a workshop (Figure 2.17 C). It was discussed whether there should be a workshop during the biennial CBRS meeting, in which the proposed annotations should be reviewed and accepted or rejected.

Since the availability of the annotation system in March 2010 not a single annotation has been made. I can only guess the reason for that, probably that the time spent to enter the data is not rewarded in any way.

With more and more isolates beeing sequenced it will not be reasonable anymore

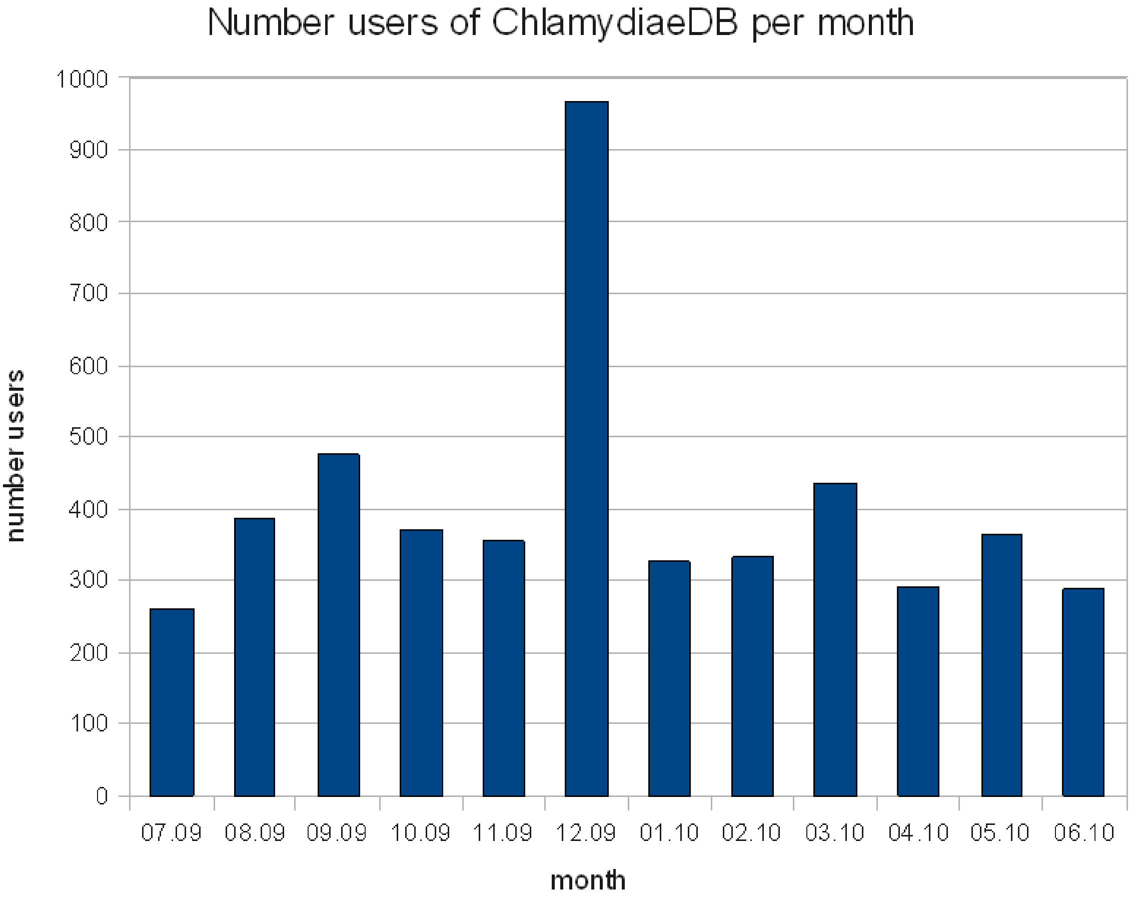


Figure 2.52: Number of users of ChlamydiaeDB per month The diagram shows the number of users of ChlamydiaeDB in the period between 12th of July 2009 and 30st of June 2010.

to treat every genome sequence as a new species. This would result in a huge list of genomes for example within the orthologous groups and the synteny views even though the genomes/isolates may only differ in a few bases. Therefore it will be necessary to decide on one reference genome for each species and to integrate the other species as SNPs into ChlamydiaeDB. Currently all genomes are treated equally in the initialization phase of ChlamydiaeDB. For the future the search for all RefSeq organisms under the NCBI taxonomyid *Chlamydiae* will have to be replaced by a list of taxonomyids for the reference genomes and an additional step for the determination and insertion of SNPs for the non-reference genomes.

3 Conclusion

The availability of more and more bacterial genome sequences opened a whole new dimension of analyses based on the comparison of genomes on nucleotide and protein sequence level. The knowledge gained in time consuming and expensive experiments in the laboratory is transferred to novel organisms if significant sequence homology can be detected between the genetic elements, as protein function depends on protein structure and sequence. Bioinformatics provides means to handle and compare the sequence data and generates hypotheses that can be checked in experiments for the novel organism in the laboratory.

As many bioinformatics analyses have to be conducted with considerable effort for every novel genome by bioinformaticians, the automation of bioinformatics analyses is as essential as the preparation of data for non-bioinformaticians working in the laboratories.

Therefore the aims of this work were the automation and improvement of analysis methods for specific prokaryotic genomes, and to make the possibilities of comparative genomics easily available for non-bioinformaticians.

This work describes the results of several collaborations with different scientists working on prokaryotes in the laboratory. Standard bioinformatics analyses like functional annotations by sequence similarity have been applied to the genome sequences of the organisms, and comparative genomics has been especially successful adressing various biological issues.

An example in which comparative genomics has been extensively used, and that is of interest for all prokaryotic genome projects, is the gene prediction, as it had to be conducted for the genomes of almost all collaborations. For this several gene finders integrating different kinds of intrinsic and extrinsic information were used. As gene predictions from different gene finders are not identical, there may occur overlapping gene models or contradictory gene starts for the same gene. These conflicts can in many cases be dissolved by the consideration of extrinsic evidence in terms of sequence similarities of the conflicting gene models to sequences of published proteins. The gene model or gene start with better support by extrinsic evidences is very likely the correct gene or gene start. We had identified the rules for the manual resolvement of these conflicts and wanted to minimize the time-consuming manual effort for the post-processing of the predictions. As there existed no software coming to the same decisions as we, ConsPred was set up, a novel automatic gene prediction pipeline, that is able to resolve problem cases by the integration of extrinsic information in the form of BLAST hits to published protein sequences in unambiguous cases. Only cases not clearly resolvable are left for manual annotation. Thus ConsPred minimizes the manual effort necessary for the gene prediction in prokaryotic genomes and comes to the same decisions as a human annotator. Due to limited data on validated genes, especially on validated gene starts, it could not reliably be examined yet how the gene start prediction performs in organisms besides *Escherichia coli*. The compilation of a comprehensive set of genes with validated gene starts is therefore the next step towards the validation and improvement of the gene prediction. But also with a set of validated genes there remains the problem that some genes can have different gene starts depending on their regulation, so that there might be not just one correct gene start.

PEDANT databases have been set up for the genomes of almost all collaboration organisms as it automates many annotations, allows the user to browse these annotations and to perform analyses like the search for BLAST hits. Specific analyses, especially comparative analyses, need to be done outside of PEDANT. An example is the detection of pseudogenes in the obligate intracellular bacterium *Amoebophilus asiaticus* 5a2, in order to get an impression of ongoing genome evolution in this organism. As the most straightforward method to detect pseudogenes is the search for truncated coding sequences, the published $\Psi\!-\!\Phi$ software was applied, that uses a set of informant genomes for the detection of these truncated coding genes. In order to review the pseudogene candidates from $\Psi - \Phi$, BLAST searches of the candidates against a non-redundant database of protein sequences have been performed and the alignments controlled manually. As $\Psi - \Phi$ only uses a limited set of informant genomes these BLAST searches have been performed for all potentially coding genes of *A. asiaticus* to be more sensitive. This resulted in the final list of 222 pseudogene candidates, that is 14.26% of all coding sequences. This is much in comparison to other members of the phylum *Bacteroidetes*, that have less than 3% of their coding sequences annotated as pseudogenes. The relatively high number of predicted pseudogenes is not astonishing as the genome of *A. asiaticus* shows a massive proliferation of insertion sequence (IS) elements (24% of all genes). The spreading of IS elements has been reported to result in proliferation of pseudogenes, genome rearrangements, and finally genome reduction. Despite the high percentage of IS elements the genome has not been extensively reshuffled recently but rather has remained stable for an extended evolutionary time period. Therefore one can only speculate about the reasons for the high amount of pseudogenes in *Amoebophilus asiaticus*, whether these genes underwent neofunctionalization for example.

The genome project of *Cronobacter turicensis* LMG 23827 could benefit most from comparative genomics. *Cronobacter* spp. are Gram-negative opportunistic foodborne pathogens. Especially neonates and infants under two months suffer from the highest infection risk and an infection can lead to severe disease manifestations such as brain abscesses, meningitis, necrotizing enterocolitis and systemic sepsis with fatal mortality rates varying from 40 to 80%. As there was only little known about lifestyle and pathogenicity of *Cronobacter* spp., *Cronobacter turicensis* LMG 23827, a strain that caused the death of two newborn children in a Children Hospital in Zürich in 2005, was sequenced and analyzed in depth in-silico. The comparison between *Cronobacter turicensis*, its distinctly related species *Cronobacter sakazakii* ATCC BAA-894, and other members of the *Enterobacteriaceae* on genome level was of central importance. The comparison on the DNA level revealed a high degree of synteny between the two *Cronobacter* spp., although a region of the chromosome of *C. sakazakii* is encoded on plasmid 3 of *C. turicensis*. The regions differing contain among others homologs to an almost complete arsenical resistance operon in *C. turicensis* and homologs to a region containing a complete Tellurium resistance operon in *C. sakazakii*. These differences might reflect adaptations to different habitats and might play a role in a differing tolerance against antimicrobials. In order to determine the conservation on proteome level a search for bidirectional-best-hits (BBHs) between all *Enterobacteriaceae*, including the two *Cronobacter* spp., was conducted. The *Cronobacter* spp. share 83-84% of their proteomes and overall *Enterobacteriaceae* share about 50% of their proteomes with each other, except for symbionts or non-opportunistic pathogens with reduced genomes. The latter organisms share almost all of their proteome with the other *Enteobacteriaceae* but do not cover much of the proteomes of other non-reduced *Enterobacteriaceae*. The *Cronobacter* spp. therefore do not belong to the group of organisms with reduced genomes. In order to determine whether the genomes of the *Cronobacter* spp. were recently subject to genome reorganization, transposases and repeats were searched in the genomes as transposases representing insertion sequence (IS) elements consist of a transposase gene flanked by inverted and/or direct repeats. The facts that *Cronobacter* spp. show a significant depletion of transposition related protein domains in comparison to other *Enterobacteriaceae* and that the repeat contents of *C. sakazakii* (1.90%) are as high and in *C. turicensis* (0.94%) lower than in other *Enterobacteriaceae* except symbionts or non-opportunistic pathogens with even lower repeat contents, suggest that there occurs no massive re-organisation of the genomes at the moment, and that the genomes of the two *Cronobacter* spp. are evolutionary quite stable. There is genomic support that plants are the natural habitat of *Cronobacter* spp. Evidence of 44 potential horizontally transferred genes closely related to sequences in non-enterobacterial often plant-associated bacteria could be detected in both *Cronobacter* spp. Although the functions of some of these horizontally transferred genes are unknown, others are clearly required for a lifestyle in a plant associated environment such as the enriched sequences for C4 compound metabolism and flagellar chemotaxis associated sequences. A plant associated environment is also supported by a significant enrichment of chemotaxis related protein domains compared to the protein domains of all other *Enterobacteriaceae*. 15 pathways typical for plant-associated organisms could be detected in-silico, for example the biosynthesis of menaquinone. Furthermore it is already known that *Cronobacter* spp. are in general capable to utilize a wide variety of compounds as a sole carbon source, some of them are known to be produced and potentially exudated by plants such as L-arabinose, D-xylose, D-cellobiose and palatinose. The capability of *Cronobacter* spp. to colonize eukaryotes such as plants and humans and to cause rare but severe infections in neonates and preterm infants raise the question which molecular factors facilitate these lifestyles. The Type IV and a Type VI secretion system as well as an array of proteins with eukaryotic like protein domains are encoded on the genomes and give a potential explanation for the potential of transferring DNA and effector proteins from the bacterial to the host cell as a mechanism of interaction with a eukaryotic host.

Additionally both genomes encode diverse transporters that may be responsible for the resistance of *Cronobacter* spp. against several antibiotics. The sequences of the whole genome of *C. turicensis* together with the insights gained within this project establish a powerful platform for further functional genomics research of this organism. The possibility to compare *Cronobacter turicensis* with *Cronobacter sakazakii* and other *Enterobacteriaceae* on the genome level allowed to hypothesize about the characteristics of this foodborne pathogen solely based on the genome sequence.

An example for prokaryotes, for which comparative genomics plays a very specific role, are *Chlamydiae*, obligate intracellular bacteria and major pathogens of humans. *Chlamydia trachomatis* is the most common cause of sexually transmitted diseases, with over 90 million new cases each year, and it can amongst others cause preventable blindness and infertility in women. *Chlamydophila pneumoniae* is a causative agent of pneumonia, which has also been associated with a number of chronic diseases such as atherosclerosis, asthma, and Alzheimer's disease. Generally, in order to be able to determine gene functions of bacteria the ability to specifically inactivate and reactivate single genes is central, e.g. in knockout experiments. As *Chlamydiae* have a characteristic developmental cycle consisting of two states, the metabolically inert elementary bodies (EBs) and the actively dividing reticulate bodies (RBs), existing in a host-derived vacuole termed inclusion, this poses obstacles in generating the tools needed to perform these genetic analyses and to define the genes that are important for the biology, pathogenicity, or transmission of *Chlamydiae*. As it is not possible to genetically manipulate *Chlamydiae*, e.g. by transformation using circular plasmids that can be easily manipulated like in *Escherichia coli*, bioinformatics and comparative genomics play an essential role in the research about *Chlamydiae*.

Even though chlamydial outer membrane proteins (OMP) are important for attachment to and entry into host cells, only few had been described. Therefore Eva Heinz developed a comprehensive, multiphasic in-silico approach to predict OMPs. As she observed that membrane predictions were in general more heterogeneous and less well defined for chlamydial outer membrane proteins as for outer membrane proteins of *Escherichia coli*, the idea arose to use the predictions for members of the same orthologous group in order to resolve uncertain predictions and by that to make the predictions more reliable. Orthologs between the chlamydial species were detected by me using the bidirectional-best-hit (BBH) method and orthologous groups were built using these BBH relations. Eva conducted OMP predictions for the chlamydial proteins, evaluated the predictions within the orthologous groups, and investigated the phylogentic conservation of the identified membrane proteins. This resulted in 88 outer membrane protein orthologous groups, including 238 proteins not previously recognized to be located in the outer membrane. Additionally it could be seen that outer membrane proteins seem to be among the fastest evolving groups of proteins and might therefore have contributed most to the differentiation of lifestyle and host spectrum of *Chlamydiae*.

As *Chlamydiae* are medically relevant organisms there are already many kinds of information available and are currently produced besides the knowledge about clusters of predicted outer membrane proteins. Examples are literature, data from experiments and genomic sequences. As there existed no resource containing the available information for all chlamydial species in a comparable manner in one place, ChlamydiaeDB, a novel multi-genome database was specifically developed for members of the phylum *Chlamydiae*. The goals for this resource were to have all available kinds of data for all chlamydial genomes in one place, to provide access to information from literature, and to provide data from experiments like single nucleotide polymorphism (SNP) data, transcript data, and data from proteome experiments. Annotations for a protein as well as information connected to proteins in the same orthologous group or in the neighborhood on the genome should be contained. The user should be supported by easy to use tools like the retrieval of all information for a list of proteins, the statistical enrichment and depletion of annotations in a list of proteins and a graphical metabolic pathway comparison between organisms. The database should be able to receive feedback from the users in the form of manual annotations, and the resource should finally be up-to-date with little manual effort, easy maintainable and easy extensible.

As a three-tier architecture, a client-server architecture from software engineering, supports the easy maintainability of a resource and is quite common in large web projects it has been used for ChlamydiaeDB. The three-tier architecture logically separates presentation layer, application processing layer, and data storage and retrieval layer. A change in one of the layers does not influence the other layers in principle and it is possible to develop and maintain each of the layers independently from each other. For the creation of the visible presentation the XML/XSLT technology has been used as this allows to create different views (XSL files) within the presentation layer on the same data (XML) from the application layer. A method was necessary that would allow to display many kinds of information on one page and would still be easily extensible. Methodologies as content management systems allow to use one XML and one XSL for a webpage. In order to add additional information to a page, XML (application layer) and XSL (presentation layer) need to be changed, which violates the separation of the layers of the three-tier and is difficult to maintain. Portlets on a web portal are the optimal solution to these problems. Each portlet is a separat building block with its own XML and XSL. Several portlets assemble a portal page, and the portlets can be reused on other portal pages. Portlets can be implemented and tested independently from other portlets on a portal page, but are not as straightforward to implement as Java Server Pages for example, and a portal server adds an additional level of complexity. But the benefits of independent development and maintenance of functionality in the form of portlets, reusability of these building blocks, separation of the layers of the three-tier, outweigh these drawbacks. In order to keep the information in ChlamydiaeDB always up-to-date, data about sequence similarities, protein domains, sequence clusters, and similarities on protein domain level was integrated on the fly from the SIMAP database and the information about coordinates of genes and diverse automatic annotations were integrated from the PEDANT system. The benefits of a direct integration of the primary resources are the up-to-dateness without the need to mirror information, the retrieval of information can be done without the need to implement the access and without knowledge about the storage of the data, and the code for accessing the data must only be maintained once. The retrieval of data has been accomplished using Enterprise Java Beans (EJBs) and Web Services within the application layer.

In order to evaluate the applicability of the three-tier architecture, the on-the-fly integration of information using EJBs and Web Services and the display of information using XML/XSLT, a web portal for SIMAP (`http://mips.gsf.de/simap`) was implemented as a protoype for ChlamydiaeDB. SIMAP is a database of publicly available protein sequences, InterPro protein domains, sequence clusters and precomputed similarities of all against all sequences. The separation of the three layers proved to be useful as the efficient retrieval of information was already provided by SIMAP and PEDANT and did not need to be reimplemented. Therefore only the presentation of the data had to be created. Even though information from different data sources is displayed on the portal pages, especially on the protein report page, the implementation was convenient and extensions to the initial functionality were easily realizable. The portal allows all users to access the sequences, protein domains, protein clusters, sequence similarities, and protein domain similarities of SIMAP in a convenient and easy way. Therefore the applicability of the technical methodologies could be proven.

As ChlamydiaeDB has to be able to provide data not retrievable on the fly from somewhere else, e.g. different types of names for proteins, Type-III secretion predictions, *Chlamydiae* relevant literature, and in order to speed up some analyses like the mapping of non-clustered proteins to clusters of orthologs or the number of proteins of an organism with a specific annotation, a specific datasource had to be developed for the storage of this information. As this resource should primarily provide fast access to information, large parts of it are designed as a data warehouse, meaning that different kinds of information are stored in a general scheme, sometimes redundantly. The decision to cache GO annotations as well as annotations from PEDANT has been made as this allowed to create aggregated results, for example the number of proteins of an organism having or not having a specific annotation. This knowledge is necessary for enrichment analyses for example. As much information is hidden in the free text of scientific publications and it very laborious to detect all *Chlamydiae* related literature manually, it was desirable to have means to detect relevant literature automatically. As there existed no straightforward solution that would easily allow to detect chlamydial gene and protein names as well as chlamydial species names automatically within the literature, a specific pipeline was developed for ChlamydiaeDB. The pipeline extracts different kinds of names for the chlamydial organisms as well as their proteins from the NCBI taxonomy, RefSeq and Uni-Prot and searches for these names within the titles and abstracts of the MEDLINE/PubMed publications available as files downloadable at the NCBI. Relevant publications are stored in the ChlamydiaeDB database and are displayed at several locations on the web portal. As it is checked for new literature daily, ChlamydiaeDB always provides the latest literature about *Chlamydiae*. SNP and transcript data has been integrated from two studies, the possibility to display information from proteome experiments has been prepared. Therefore also data from experiments can easily be used within the resource. The most feature rich page of ChlamydiaeDB is the protein report. This page provides different kinds of information from very different data sources for a selected protein. This information for a protein includes protein sequence, nucleotide sequence, manual annotations, genomic neighborhood, automatically detected literature, cluster membership of the sequence, overview over annotations for other members of the same orthologous group, Interpro protein domain annotations, type III secretion predictions, FunCat annotations, GO annotations, EC annotations, membership of the protein in KEGG pathways. This makes an overview over the features of a protein easily possible. The portlet showing available information about orthologous proteins in other chlamydial species is very powerful as the user can easily see which kinds of information are available for the same protein in other chlamydial species, and this makes knowledge easily transferrable over the borders of a single species. In order to make the retrieval of all information for a list of proteins more convenient for the user, a specific tool was implemented so that the available data can be retrieved at once without the need to open the protein report for every single protein. The KEGG metabolic pathways can be used to get a graphical overview over the metabolic capabilities of an organism, if it is already annotated in KEGG. If the genome is not annotated yet, EC assignments for proteins can be used to map the proteins to KEGG orthologous groups that can be colored in the KEGG maps. As PEDANT contains the EC annotations for all RefSeq organisms, the coloring of the KEGG maps is easily doable. When the metabolic capabilities of more than one organism should be compared this is difficult as the pictures of several colored pathways need to be compared visually. Therefore a novel graphical pathway comparison tool for up to three RefSeq organisms was implemented together with students. A pathway and up to three organisms can be selected and the enzymes in the pathway are colored respectively in a single picture, indicating in which of the three organisms the enzyme is encoded. An extremely powerful tool is the possibility to compare the annotations of two sets of proteins and to detect significant enrichments or depletions of annotations. This tool requires no expert statistical knowledge and allows the detection of significantly over- or underrepresented annotations in one set of proteins in comparison to another set of proteins. All these tools make ChlamydiaeDB a very powerful toolbox for scientists working with genomic data of *Chlamydiae.* In order to be able to get feedback from the scientists, the possibility to submit manual annotations for every protein has been implemented. The annotation consists of a multi-step procedure consisting of the submission of the initial annotation proposal, the evaluation of the proposal by an administrator, and the rejection or approval of the annotation that is visible after acception by the annotator. At the conference of the Chlamydia Basic Research Society (CBRS) in Little Rock USA in March 2009 it was agreed that ChlamydiaeDB will be the database used and maintained by the community in the future. It is currently accessed over 300 times per month.

This work shows the successful application of comparative genomics for the research on prokaryotes. The manual effort for the gene prediction could be reduced, pseudogenes were predicted and clusters of orthologs were used for the detection of novel outer membrane proteins in the Gram-negative bacteria of the phylum *Chlamydiae.* The comprehensive analysis of the genome of *Cronobacter turicensis* profited from the capabilities of comparative genomics. It was shown how the characterization of the genomic features of a pathogenic prokaryote can elucidate its lifestyle and pathogenicity factors, and how this characterization creates novel hypotheses that can be evaluated in the laboratory. These findings can hopefully support the development of countermeasures against these pathogens. Finally ChlamydiaeDB, a novel resource implemented for genomic data of all members of the phylum *Chlamydiae*, shows that state-of-the-art comparative genomics methods can be easily made available and used without the need to be an informatician. That way every scientist working with genomic data of *Chlamydiae* can profit from ChlamydiaeDB.

By providing information to non-bioinformaticians a very important precondition for collaboration is established. All the projects of this work show that the concerted effort of scientists from various fields is essential towards the goal to better understand the biology of organisms. Therefore I am sure that the collaboration of scientists from different fields bears great potential for the future of scientific research in natural sciences in general.

References

[1] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Chun He Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-Bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari A Vogt, Greg A Volkmer, Shally H Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Sep 2005.

[2] Michael A Quail, Iwanka Kozarewa, Frances Smith, Aylwyn Scally, Philip J Stephens, Richard Durbin, Harold Swerdlow, and Daniel J Turner. A large genome center's improvements to the illumina sequencing system. *Nat Methods*, 5(12):1005–1010, Dec 2008.

[3] Francis S Collins, Michael Morgan, and Aristides Patrinos. The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290, Apr 2003.

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

[5] Marco Albrecht, Cynthia M Sharma, Richard Reinhardt, Jörg Vogel, and Thomas Rudel. Deep sequencing-based discovery of the chlamydia trachomatis transcriptome. *Nucleic Acids Res*, Nov 2009.

[6] Preeti Sachdeva, Richa Misra, Anil K Tyagi, and Yogendra Singh. The sigma factors of mycobacterium tuberculosis: regulation of the regulators. *FEBS J*, 277(3):605–626, Feb 2010.

[7] J. Tamames, G. Casari, C. Ouzounis, and A. Valencia. Conserved clusters of functionally related genes in two bacterial genomes. *J Mol Evol*, 44(1):66–73, Jan 1997.

[8] 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.

[9] T. Ettema, J. van der Oost, and M. Huynen. Modularity in the gain and loss of genes: applications for function prediction. *Trends Genet*, 17(9):485–487, Sep 2001.

[10] F. JACOB and J. MONOD. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–356, Jun 1961.

[11] Yu Zheng, Joseph D Szustakowski, Lance Fortnow, Richard J Roberts, and Simon Kasif. Computational identification of operons in microbial genomes. *Genome Res*, 12(8):1221–1230, Aug 2002.

[12] 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.

[13] B. Snel, G. Lehmann, P. Bork, and M. A. Huynen. String: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res*, 28(18):3442–3444, Sep 2000.

[14] I. Yanai, A. Derti, and C. DeLisi. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci U S A*, 98(14):7940–7945, Jul 2001.

[15] 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.

[16] M Zvelebil and JO Baum. *Understanding Bioinformatics.* Garland Science, 2008.

[17] C. Yanofsky, T. Platt, I. P. Crawford, B. P. Nichols, G. E. Christie, H. Horowitz, M. VanCleemput, and A. M. Wu. The complete nucleotide sequence of the tryptophan operon of escherichia coli. *Nucleic Acids Res*, 9(24):6647–6668, Dec 1981.

[18] B Alberts, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. *Molecular Biology of the Cell, 5th ed. New York.* Garland Science, 2008.

[19] G. J. Phillips, J. Arnold, and R. Ivarie. Mono- through hexanucleotide composition of the escherichia coli genome: a markov chain analysis. *Nucleic Acids Res*, 15(6):2611–2626, Mar 1987.

[20] S. Ohno. Universal rule for coding sequence construction: Ta/cg deficiency-tg/ct excess. *Proc Natl Acad Sci U S A*, 85(24):9630–9634, Dec 1988.

[21] P. M. Sharp and W. H. Li. Codon usage in regulatory genes in escherichia coli does not reflect selection for 'rare' codons. *Nucleic Acids Res*, 14(19):7737–7749, Oct 1986.

[22] T. Ikemura. Correlation between the abundance of escherichia coli transfer rnas and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the e. coli translational system. *J Mol Biol*, 151(3):389–409, Sep 1981.

[23] T. Ikemura. Codon usage and trna content in unicellular and multicellular organisms. *Mol Biol Evol*, 2(1):13–34, Jan 1985.

[24] F. H. Crick. Codon–anticodon pairing: the wobble hypothesis. *J Mol Biol*, 19(2):548–555, Aug 1966.

[25] Nathaniel Echols, Paul Harrison, Suganthi Balasubramanian, Nicholas M Luscombe, Paul Bertone, Zhaolei Zhang, and Mark Gerstein. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res*, 30(11):2515–2523, Jun 2002.

[26] M. M. Wösten. Eubacterial sigma-factors. *FEMS Microbiol Rev*, 22(3):127–150, Sep 1998.

[27] Tanja M Gruber and Carol A Gross. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol*, 57:441–466, 2003.

[28] Mark S B Paget and John D Helmann. The sigma70 family of sigma factors. *Genome Biol*, 4(1):203, 2003.

[29] Douglas F Browning and Stephen J Busby. The regulation of bacterial transcription initiation. *Nat Rev Microbiol*, 2(1):57–65, Jan 2004.

[30] Sacha A F T van Hijum, Marnix H Medema, and Oscar P Kuipers. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiol Mol Biol Rev*, 73(3):481–509, Table of Contents, Sep 2009.

[31] J. Besemer, A. Lomsadze, and M. Borodovsky. Genemarks: a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*, 29(12):2607–18, 2001. Journal Article England.

[32] M Borodovsky and J McIninch. Genemark: parallel gene recognition for both dna strands. *Computers & Chemistry*, Vol. 17, No. 19:123–133, 1993.

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

[34] A. L. Delcher, K. A. Bratke, E. C. Powers, and S. L. Salzberg. Identifying bacterial genes and endosymbiont dna with glimmer. *Bioinformatics*, 23(6):673–9, 2007. HHSN266200400038C/PHS HHS/United States R01 LM006845-08/LM/NLM NIH HHS/United States R01 LM007938-04/LM/NLM NIH HHS/United States R01-LM006845/LM/NLM NIH HHS/United States R01-LM007938/LM/NLM NIH HHS/United States Evaluation Studies Journal Article Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. England.

[35] D. Devos and A. Valencia. Intrinsic errors in genome annotation. *Trends Genet*, 17(8):429–431, Aug 2001.

[36] M. Skovgaard, L. J. Jensen, S. Brunak, D. Ussery, and A. Krogh. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet*, 17(8):425–428, Aug 2001.

[37] 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.

[38] D. Frishman, A. Mironov, H. W. Mewes, and M. Gelfand. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res*, 26(12):2941–2947, Jun 1998.

[39] J. H. Badger and G. J. Olsen. Critica: coding region identification tool invoking comparative analysis. *Mol Biol Evol*, 16(4):512–524, Apr 1999.

[40] Thomas Schou Larsen and Anders Krogh. Easygene–a prokaryotic gene finder that ranks orfs by statistical significance. *BMC Bioinformatics*, 4:21, Jun 2003.

[41] Pernille Nielsen and Anders Krogh. Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics*, 21(24):4322–4329, Dec 2005.

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

[43] Sungsoo Kang, Sung-Jin Yang, Sangsoo Kim, and Jong Bhak. Consorf: a consensus prediction system for prokaryotic coding sequences. *Bioinformatics*, 23(22):3088–3090, Nov 2007.

[44] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg. Improved microbial gene identification with glimmer. *Nucleic Acids Res*, 27(23):4636–4641, Dec 1999.

[45] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated markov models. *Nucleic Acids Res*, 26(2):544–548, Jan 1998.

[46] Feng-Biao Guo, Hong-Yu Ou, and Chun-Ting Zhang. Zcurve: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res*, 31(6):1780–1789, Mar 2003.

[47] W. R. Pearson, T. Wood, Z. Zhang, and W. Miller. Comparison of dna sequences with protein sequences. *Genomics*, 46(1):24–36, Nov 1997.

[48] Tomasz A Leski, Anthony P Malanoski, David A Stenger, and Baochuan Lin. Target amplification for broad spectrum microbial diagnostics and detection. *Future Microbiol*, 5(2):191–203, Feb 2010.

[49] N. Haddad, C. Marce, C. Magras, and J-M. Cappelier. An overview of methods used to clarify pathogenesis mechanisms of campylobacter jejuni. *J Food Prot*, 73(4):786–802, Apr 2010.

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

[51] P. Carranza, I. Hartmann, A. Lehner, R. Stephan, P. Gehrig, J. Grossmann, S. Barkow-Oesterreicher, B. Roschitzki, L. Eberl, and K. Riedel. Proteomic profiling of cronobacter turicensis 3032, a food-borne opportunistic pathogen. *Proteomics*, 9(13):3564–79, 2009. Journal Article Research Support, Non-U.S. Gov't Germany.

[52] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott. Ncbi reference sequences: current status, policy and new initiatives. *Nucleic Acids Res*, 37(Database issue):D32–6, 2009. Journal Article Research Support, N.I.H., Intramural England.

[53] A. J. Link, K. Robison, and G. M. Church. Comparing the predicted and observed properties of proteins encoded in the genome of escherichia coli k-12. *Electrophoresis*, 18(8):1259–1313, Aug 1997.

[54] Michalis Aivaliotis, Kris Gevaert, Michaela Falb, Andreas Tebbe, Kosta Konstantinidis, Birgit Bisle, Christian Klein, Lennart Martens, An Staes, Evy Timmerman, Jozef Van Damme, Frank Siedler, Friedhelm Pfeiffer, Joël Vandekerckhove, and Dieter Oesterhelt. Large-scale identification of n-terminal peptides in the halophilic archaea halobacterium salinarum and natronomonas pharaonis. *J Proteome Res*, 6(6):2195–2204, Jun 2007.

[55] Mathieu Baudet, Philippe Ortet, Jean-Charles Gaillard, Bernard Fernandez, Philippe Guérin, Christine Enjalbal, Gilles Subra, Arjan de Groot, Mohamed Barakat, Alain Dedieu, and Jean Armengaud. Proteomics-based refinement of deinococcus deserti genome annotation reveals an unwanted use of non-canonical translation initiation codons. *Mol Cell Proteomics*, 9(2):415–426, Feb 2010.

[56] M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–367, Jun 1996.

[57] S. Smit, J. Widmann, and R. Knight. Evolutionary rates vary among rrna structural elements. *Nucleic Acids Res*, 35(10):3339–3354, 2007.

[58] N. R. Pace. A molecular view of microbial diversity and the biosphere. *Science*, 276(5313):734–740, May 1997.

[59] M. M. Yusupov, G. Z. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. H. Cate, and H. F. Noller. Crystal structure of the ribosome at 5.5 a resolution. *Science*, 292(5518):883–896, May 2001.

[60] VA Bondarenko, YV Liu, YI Jiang, and VM Studitsky. Communication over a large distance: enhancers and insulators. *Biochem Cell Biol*, 81:241–251, 2003.

[61] Emmanuelle Lerat and Howard Ochman. Psi-phi: exploring the outer limits of bacterial pseudogenes.delcher a., harmon d., kasif s., white o., salzberg s. 1999. improved microbial gene identification with glimmer. nucleic acids res. 27:4636-4641.delcher a., harmon d., kasif s., white o., salzberg s. 1999. improved microbial gene identification with glimmer. nucleic acids res. 27:4636-4641. *Genome Res*, 14(11):2273–2278, Nov 2004.

[62] J. O. Andersson and S. G. Andersson. Genome degradation is an ongoing process in rickettsia. *Mol Biol Evol*, 16(9):1178–1191, Sep 1999.

[63] J. O. Andersson and S. G. Andersson. Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev*, 9(6):664–671, Dec 1999.

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

[65] J Zhang. Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6):292–298, 2003.

[66] Linus Sandegren and Dan I Andersson. Bacterial gene amplification: implications for the evolution of antibiotic resistance. *Nat Rev Microbiol*, 7(8):578–588, Aug 2009.

[67] A. van Belkum. Short sequence repeats in microbial pathogenesis and evolution. *Cell Mol Life Sci*, 56(9-10):729–734, Nov 1999.

[68] E. P. Rocha, A. Danchin, and A. Viari. Functional and evolutionary roles of long repeats in prokaryotes. *Res Microbiol*, 150(9-10):725–733, 1999.

[69] J. Mahillon and M. Chandler. Insertion sequences. *Microbiol Mol Biol Rev*, 62(3):725–774, Sep 1998.

[70] M. Syvanen. *Bacterial Genomes.* Chapman & Hall, 1998.

[71] Nicholas Delihas. Small mobile sequences in bacteria display diverse structure/function motifs. *Mol Microbiol*, 67(3):475–481, Feb 2008.

[72] Mathias C Walter, Thomas Rattei, Roland Arnold, Ulrich Güldener, Martin Münsterkötter, Karamfilka Nenova, Gabi Kastenmüller, Patrick Tischler, Andreas Wölling, Andreas Volz, Norbert Pongratz, Ralf Jost, Hans-Werner Mewes, and Dmitrij Frishman. Pedant covers all complete refseq genomes. *Nucleic Acids Res*, 37(Database issue):D408–D411, Jan 2009.

[73] 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.

[74] G. A. Fichant and C. Burks. Identifying potential trna genes in genomic dna sequences. *J Mol Biol*, 220(3):659–671, Aug 1991.

[75] A. Rich and U. L. RajBhandary. Transfer rna: molecular structure, sequence, and properties. *Annu Rev Biochem*, 45:805–860, 1976.

[76] D. H. Gauss and M. Sprinzl. Compilation of sequences of trna genes. *Nucleic Acids Res*, 11(1):r55–103, Jan 1983.

[77] M. Sprinzl, T. Hartmann, J. Weber, J. Blank, and R. Zeidler. Compilation of trna sequences and sequences of trna genes. *Nucleic Acids Res*, 17 Suppl:r1–172, 1989.

[78] J. Devereux, P. Haeberli, and O. Smithies. A comprehensive set of sequence analysis programs for the vax. *Nucleic Acids Res*, 12(1 Pt 1):387–395, Jan 1984.

[79] P. Agarwal and D. J. States. The repeat pattern toolkit (rpt): analyzing the structure and evolution of the c. elegans genome. *Proc Int Conf Intell Syst Mol Biol*, 2:1–9, 1994.

[80] S. Kurtz and C. Schleiermacher. Reputer: fast computation of maximal repeats in complete genomes. *Bioinformatics*, 15(5):426–7, 1999. Journal Article Research Support, Non-U.S. Gov't England.

[81] 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.

[82] H. W. Mewes, K. Albermann, M. Bähr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, S. G. Oliver, F. Pfeiffer, and A. Zollner. Overview of the yeast genome. *Nature*, 387(6632 Suppl):7–65, May 1997.

[83] H. W. Mewes, K. Albermann, K. Heumann, S. Liebl, and F. Pfeiffer. MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res*, 25(1):28–30, Jan 1997.

[84] M. Salanoubat, K. Lemcke, M. Rieger, W. Ansorge, M. Unseld, B. Fartmann, G. Valle, H. Blöcker, M. Perez-Alonso, B. Obermaier, M. Delseny, M. Boutry, L. A. Grivell, R. Mache, P. Puigdomènech, V. De Simone, N. Choisne, F. Artiguenave, C. Robert, P. Brottier, P. Wincker, L. Cattolico, J. Weissenbach, W. Saurin, F. Quétier, M. Schäfer, S. Müller-Auer, C. Gabel, M. Fuchs, V. Benes,

E. Wurmbach, H. Drzonek, H. Erfle, N. Jordan, S. Bangert, R. Wiedelmann, H. Kranz, H. Voss, R. Holland, P. Brandt, G. Nyakatura, A. Vezzi, M. D'Angelo, A. Pallavicini, S. Toppo, B. Simionati, A. Conrad, K. Hornischer, G. Kauer, T. H. Löhnert, G. Nordsiek, J. Reichelt, M. Scharfe, O. Schön, M. Bargues, J. Terol, J. Climent, P. Navarro, C. Collado, A. Perez-Perez, B. Ottenwälder, D. Duchemin, R. Cooke, M. Laudie, C. Berger-Llauro, B. Purnelle, D. Masuy, M. de Haan, A. C. Maarse, J. P. Alcaraz, A. Cottet, E. Casacuberta, A. Monfort, A. Argiriou, M. flores, R. Liguori, D. Vitale, G. Mannhaupt, D. Haase, H. Schoof, S. Rudd, P. Zaccaria, H. W. Mewes, K. F. Mayer, S. Kaul, C. D. Town, H. L. Koo, L. J. Tallon, J. Jenkins, T. Rooney, M. Rizzo, A. Walts, T. Utterback, C. Y. Fujii, T. P. Shea, T. H. Creasy, B. Haas, R. Maiti, D. Wu, J. Peterson, S. Van Aken, G. Pai, J. Militscher, P. Sellers, J. E. Gill, T. V. Feldblyum, D. Preuss, X. Lin, W. C. Nierman, S. L. Salzberg, O. White, J. C. Venter, C. M. Fraser, T. Kaneko, Y. Nakamura, S. Sato, T. Kato, E. Asamizu, S. Sasamoto, T. Kimura, K. Idesawa, K. Kawashima, Y. Kishida, C. Kiyokawa, M. Kohara, M. Matsumoto, A. Matsuno, A. Muraki, S. Nakayama, N. Nakazaki, S. Shinpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, S. Tabata, European Union Chromosome 3 Arabidopsis Sequencing Consortium, Institute for Genomic Research, and Kazusa DNA Research Institute. Sequence and analysis of chromosome 3 of the plant Arabidopsis thaliana. *Nature*, 408(6814):820–822, Dec 2000.

[85] A. Ruepp, W. Graml, M. L. Santos-Martinez, K. K. Koretke, C. Volker, H. W. Mewes, D. Frishman, S. Stocker, A. N. Lupas, and W. Baumeister. The genome sequence of the thermoacidophilic scavenger Thermoplasma acidophilum. *Nature*, 407(6803):508–513, Sep 2000.

[86] James E Galagan, Sarah E Calvo, Katherine A Borkovich, Eric U Selker, Nick D Read, David Jaffe, William FitzHugh, Li-Jun Ma, Serge Smirnov, Seth Purcell, Bushra Rehman, Timothy Elkins, Reinhard Engels, Shunguang Wang, Cydney B Nielsen, Jonathan Butler, Matthew Endrizzi, Dayong Qui, Peter Ianakiev, Deborah Bell-Pedersen, Mary Anne Nelson, Margaret Werner-Washburne, Claude P Selitrennikoff, John A Kinsey, Edward L Braun, Alex Zelter, Ulrich Schulte, Gregory O Kothe, Gregory Jedd, Werner Mewes, Chuck Staben, Edward Marcotte, David Greenberg, Alice Roy, Karen Foley, Jerome Naylor, Nicole Stange-Thomann, Robert Barrett, Sante Gnerre, Michael Kamal, Manolis Kamvysselis, Evan Mauceli, Cord Bielke, Stephen Rudd, Dmitrij Frishman, Svetlana Krystofova, Carolyn Rasmussen, Robert L Metzenberg, David D Perkins, Scott Kroken, Carlo Cogoni, Giuseppe Macino, David Catcheside, Weixi Li, Robert J Pratt, Stephen A Osmani, Colin P C DeSouza, Louise Glass, Marc J Orbach, J. Andrew Berglund, Rodger Voelker, Oded Yarden, Michael Plamann, Stephan Seiler, Jay Dunlap, Alan Radford, Rodolfo Aramayo, Donald O Natvig, Lisa A Alex, Gertrud Mannhaupt, Daniel J Ebbole, Michael Freitag, Ian Paulsen, Matthew S Sachs, Eric S Lander, Chad Nusbaum, and Bruce Birren. The genome sequence of the filamentous fungus Neurospora crassa. *Nature*, 422(6934):859–868, Apr 2003.

[87] Alvaro Mateos, Joaquín Dopazo, Ronald Jansen, Yuhai Tu, Mark Gerstein, and Gustavo Stolovitzky. Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Res*, 12(11):1703–1715, Nov 2002.

[88] A. Clare and R. D. King. Predicting gene function in Saccharomyces cerevisiae. *Bioinformatics*, 19 Suppl 2:II42–II49, Oct 2003.

[89] 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.

[90] 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.

[91] 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.

[92] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.

[93] Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, Sep 2005.

[94] T. Rattei, P. Tischler, S. Gotz, M. A. Jehl, J. Hoser, R. Arnold, A. Conesa, and H. W. Mewes. Simap–a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Res*, 38(Database issue):D223–6, 2010. Journal Article Research Support, Non-U.S. Gov't England.

[95] A. J. Barrett. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). *Eur J Biochem*, 250(1):1–6, Nov 1997.

[96] M. H. Saier. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol Mol Biol Rev*, 64(2):354–411, Jun 2000.

[97] 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.

[98] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, Mar 1970.

[99] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, Mar 1981.

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

[101] Burkhard Rost. Enzyme function less conserved than anticipated. *J Mol Biol*, 318(2):595–608, Apr 2002.

[102] Weidong Tian and Jeffrey Skolnick. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol*, 333(4):863–882, Oct 2003.

[103] W. M. Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, 19(2):99–113, Jun 1970.

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

[105] 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.

[106] 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):1041–52, 2001. Comparative Study Journal Article Research Support, Non-U.S. Gov't England.

[107] Roman L Tatusov, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris Kiryutin, Eugene V Koonin, Dmitri M Krylov, Raja Mazumder, Sergei L Mekhedov, Anastasia N Nikolskaya, B. Sridhar Rao, Sergei Smirnov, Alexander V Sverdlov, Sona Vasudevan, Yuri I Wolf, Jodie J Yin, and Darren A Natale. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, Sep 2003.

[108] J. Muller, D. Szklarczyk, P. Julien, I. Letunic, A. Roth, M. Kuhn, S. Powell, C. von Mering, T. Doerks, L. J. Jensen, and P. Bork. eggnog v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups,

species and functional annotations. *Nucleic Acids Res*, 38(Database issue):D190–D195, Jan 2010.

[109] Yuki Moriya, Masumi Itoh, Shujiro Okuda, Akiyasu C Yoshizawa, and Minoru Kanehisa. Kaas: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*, 35(Web Server issue):W182–W185, Jul 2007.

[110] J. A. Eisen. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*, 8(3):163–167, Mar 1998.

[111] J. A. Eisen, K. S. Sweder, and P. C. Hanawalt. Evolution of the snf2 family of proteins: subfamilies with distinct sequences and functions. *Nucleic Acids Res*, 23(14):2715–2723, Jul 1995.

[112] J. A. Eisen, D. Kaiser, and R. M. Myers. Gastrogenomic delights: a movable feast. *Nat Med*, 3(10):1076–1078, Oct 1997.

[113] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, New York, 1999.

[114] Chris P Ponting and Robert R Russell. The natural history of protein domains. *Annu Rev Biophys Biomol Struct*, 31:45–71, 2002.

[115] 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.

[116] Christian J A Sigrist, Lorenzo Cerutti, Edouard de Castro, Petra S Langendijk-Genevaux, Virginie Bulliard, Amos Bairoch, and Nicolas Hulo. Prosite, a protein domain database for functional characterization and annotation. *Nucleic Acids Res*, 38(Database issue):D161–D166, Jan 2010.

[117] T. K. Attwood, P. Bradley, D. R. Flower, A. Gaulton, N. Maudling, A. L. Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor, A. Uddin, and C. Zygouri. Prints and its automatic supplement, preprints. *Nucleic Acids Res*, 31(1):400–402, Jan 2003.

[118] 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.

[119] Florence Servant, Catherine Bru, Sébastien Carrère, Emmanuel Courcelle, Jérôme Gouzy, David Peyruc, and Daniel Kahn. Prodom: automated clustering of homologous domains. *Brief Bioinform*, 3(3):246–251, Sep 2002.

[120] Ivica Letunic, Tobias Doerks, and Peer Bork. Smart 6: recent updates and new developments. *Nucleic Acids Res*, 37(Database issue):D229–D232, Jan 2009.

[121] Daniel H Haft, Jeremy D Selengut, and Owen White. The tigrfams database of protein families. *Nucleic Acids Res*, 31(1):371–373, Jan 2003.

[122] Anastasia N Nikolskaya, Cecilia N Arighi, Hongzhan Huang, Winona C Barker, and Cathy H Wu. Pirsf family classification system for protein functional and evolutionary analysis. *Evol Bioinform Online*, 2:197–209, 2006.

[123] Derek Wilson, Martin Madera, Christine Vogel, Cyrus Chothia, and Julian Gough. The superfamily database in 2007: families and functions. *Nucleic Acids Res*, 35(Database issue):D308–D313, Jan 2007.

[124] Paul D Thomas, Michael J Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome Res*, 13(9):2129–2141, Sep 2003.

[125] Jonathan Lees, Corin Yeats, Oliver Redfern, Andrew Clegg, and Christine Orengo. Gene3d: merging structure and function for a thousand genomes. *Nucleic Acids Res*, 38(Database issue):D296–D300, Jan 2010.

[126] Tania Lima, Andrea H Auchincloss, Elisabeth Coudert, Guillaume Keller, Karine Michoud, Catherine Rivoire, Virginie Bulliard, Edouard de Castro, Corinne Lachaize, Delphine Baratin, Isabelle Phan, Lydie Bougueleret, and Amos Bairoch. Hamap: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in uniprotkb/swiss-prot. *Nucleic Acids Res*, 37(Database issue):D471–D478, Jan 2009.

[127] Kui Lin, Lei Zhu, and Da-Yong Zhang. An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics*, 22(17):2081–2086, Sep 2006.

[128] 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.

[129] E. V. Koonin and Y. I. Wolf. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res*, 36(21):6688–719, 2008. Journal Article Research Support, N.I.H., Intramural Review England.

[130] Matthew T G Holden, Heidi Hauser, Mandy Sanders, Thi Hoa Ngo, Inna Cherevach, Ann Cronin, Ian Goodhead, Karen Mungall, Michael A Quail, Claire Price, Ester Rabbinowitsch, Sarah Sharp, Nicholas J Croucher, Tran Bich Chieu, Nguyen Thi Hoang Mai, To Song Diep, Nguyen Tran Chinh, Michael Kehoe, James A Leigh, Philip N Ward, Christopher G Dowson, Adrian M Whatmore, Neil Chanter, Pernille Iversen, Marcelo Gottschalk, Josh D Slater, Hilde E Smith, Brian G Spratt, Jianguo Xu, Changyun Ye, Stephen Bentley, Barclay G Barrell, Constance Schultsz, Duncan J Maskell, and Julian Parkhill. Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen streptococcus suis. *PLoS One*, 4(7):e6072, 2009.

[131] L. B. Koski and G. B. Golding. The closest blast hit is often not the nearest neighbor. *J Mol Evol*, 52(6):540–542, Jun 2001.

[132] Eugene A Gladyshev, Matthew Meselson, and Irina R Arkhipova. Massive horizontal gene transfer in bdelloid rotifers. *Science*, 320(5880):1210–1213, May 2008.

[133] Tancred Frickey and Andrei N Lupas. Phylogenie: automated phylome generation and analysis. *Nucleic Acids Res*, 32(17):5231–5238, 2004.

[134] 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.

[135] R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol*, 1(2):93–108, 1999.

[136] A. R. Mushegian and E. V. Koonin. Gene order is not conserved in bacterial evolution. *Trends Genet*, 12(8):289–290, Aug 1996.

[137] H. Watanabe, H. Mori, T. Itoh, and T. Gojobori. Genome plasticity as a paradigm of eubacteria evolution. *J Mol Evol*, 44 Suppl 1:S57–S64, 1997.

[138] 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.

[139] E. V. Koonin, K. S. Makarova, and L. Aravind. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*, 55:709–742, 2001.

[140] J. G. Lawrence. Selfish operons and speciation by gene transfer. *Trends Microbiol*, 5(9):355–359, Sep 1997.

[141] L. Aravind, H. Watanabe, D. J. Lipman, and E. V. Koonin. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci U S A*, 97(21):11319–11324, Oct 2000.

[142] B. Snel, P. Bork, and M. Huynen. Genome evolution. gene fusion versus gene fission. *Trends Genet*, 16(1):9–11, Jan 2000.

[143] Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet*, 29(2):153–159, Oct 2001.

[144] Jeffrey G Lawrence. Shared strategies in gene organization among prokaryotes and eukaryotes. *Cell*, 110(4):407–413, Aug 2002.

[145] W. C. Lathe, B. Snel, and P. Bork. Gene context conservation of a higher order than operons. *Trends Biochem Sci*, 25(10):474–479, Oct 2000.

[146] 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.

[147] Lars J Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, Peer Bork, and Christian von Mering. String 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, 37(Database issue):D412–D416, Jan 2009.

[148] Maarten Fauvart and Jan Michiels. Rhizobial secreted proteins as determinants of host specificity in the rhizobium-legume symbiosis. *FEMS Microbiol Lett*, 285(1):1–9, Aug 2008.

[149] William J Deakin and William J Broughton. Symbiotic use of pathogenic strategies: rhizobial protein secretion systems. *Nat Rev Microbiol*, 7(4):312–320, Apr 2009.

[150] Kumiko Kambara, Silvia Ardissone, Hajime Kobayashi, Maged M Saad, Olivier Schumpp, William J Broughton, and William J Deakin. Rhizobia utilize pathogen-like effector proteins during symbiosis. *Mol Microbiol*, 71(1):92–106, Jan 2009.

[151] Delphine Sylvie Anne Beeckman and Daisy C G Vanrompay. Bacterial secretion systems with an emphasis on the chlamydial type iii secretion system. *Curr Issues Mol Biol*, 12(1):17–42, Jul 2009.

[152] Alan R Hauser. The type iii secretion system of pseudomonas aeruginosa: infection by injection. *Nat Rev Microbiol*, 7(9):654–665, Sep 2009.

[153] Matxalen Llosa, Craig Roy, and Christoph Dehio. Bacterial type iv secretion systems in human disease. *Mol Microbiol*, 73(2):141–151, Jul 2009.

[154] Stefan Pukatzki, Steven B McAuley, and Sarah T Miyata. The type vi secretion system: translocation of effectors and effector-domains. *Curr Opin Microbiol*, 12(1):11–17, Feb 2009.

[155] Jorge E Galán and Hans Wolf-Watz. Protein delivery into eukaryotic cells by type iii secretion machines. *Nature*, 444(7119):567–573, Nov 2006.

[156] Lewis Eh Bingle, Christopher M Bailey, and Mark J Pallen. Type vi secretion: a beginner's guide. *Curr Opin Microbiol*, 11(1):3–8, Feb 2008.

[157] Roxane Simeone, Daria Bottai, and Roland Brosch. Esx/type vii secretion systems and their role in host-pathogen interaction. *Curr Opin Microbiol*, 12(1):4–10, Feb 2009.

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

[159] H. Tjalsma, A. Bolhuis, J. D. Jongbloed, S. Bron, and J. M. van Dijl. Signal peptide-dependent protein transport in bacillus subtilis: a genome-based survey of the secretome. *Microbiol Mol Biol Rev*, 64(3):515–547, Sep 2000.

[160] D. G. Thanassi and S. J. Hultgren. Multiple pathways allow protein secretion across the bacterial outer membrane. *Curr Opin Cell Biol*, 12(4):420–430, Aug 2000.

[161] Joseph D Mougous, Marianne E Cuff, Stefan Raunser, Aimee Shen, Min Zhou, Casey A Gifford, Andrew L Goodman, Grazyna Joachimiak, Claudia L Ordoñez, Stephen Lory, Thomas Walz, Andrzej Joachimiak, and John J Mekalanos. A virulence locus of pseudomonas aeruginosa encodes a protein secretion apparatus. *Science*, 312(5779):1526–1530, Jun 2006.

[162] Stefan Pukatzki, Amy T Ma, Derek Sturtevant, Bryan Krastins, David Sarracino, William C Nelson, John F Heidelberg, and John J Mekalanos. Identification of a conserved bacterial protein secretion system in vibrio cholerae using the dictyostelium host model system. *Proc Natl Acad Sci U S A*, 103(5):1528–1533, Jan 2006.

[163] Daniela Büttner and Ulla Bonas. Port of entry–the type iii secretion translocon. *Trends Microbiol*, 10(4):186–192, Apr 2002.

[164] Alain Filloux, Abderrahman Hachani, and Sophie Bleves. The bacterial type vi secretion machine: yet another player for protein transport across membranes. *Microbiology*, 154(Pt 6):1570–1583, Jun 2008.

[165] 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, Oct 2006.

[166] Agathe Subtil, Cédric Delevoye, María-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, Jun 2005.

[167] Lisa M Schechter, Monica Vencato, Katy L Jordan, Sarah E Schneider, David J Schneider, and Alan Collmer. Multiple approaches to a complete inventory of pseudomonas syringae pv. tomato dc3000 type iii secretion system effector proteins. *Mol Plant Microbe Interact*, 19(11):1180–1192, Nov 2006.

[168] 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, Jan 2004.

[169] Mark J Pallen, Scott A Beatson, and Christopher M Bailey. Bioinformatics analysis of the locus for enterocyte effacement provides novel insights into type-iii secretion. *BMC Microbiol*, 5:9, 2005.

[170] 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, Aug 2005.

[171] 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, Oct 2005.

[172] 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, Nov 2006.

[173] 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.

[174] Roland Arnold, Andre Jehl, and Thomas Rattei. Targeting effectors: the molecular recognition of type iii secreted proteins. *Microbes Infect*, 12(5):346–358, May 2010.

[175] Martin Löwer and Gisbert Schneider. Prediction of type iii secretion signals in genomes of gram-negative bacteria. *PLoS One*, 4(6):e5917, 2009.

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

[177] 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.

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

[179] Marc-André Jehl, Roland Arnold, and Thomas Rattei. Effective–a database of predicted secreted bacterial proteins. *Nucleic Acids Res*, Nov 2010.

[180] Lawrence Hunter and K. Bretonnel Cohen. Biomedical language processing: what's beyond pubmed? *Mol Cell*, 21(5):589–594, Mar 2006.

[181] Lars Juhl Jensen, Jasmin Saric, and Peer Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*, 7(2):119–129, Feb 2006.

[182] UniProt Consortium. The universal protein resource (uniprot) 2009. *Nucleic Acids Res*, 37(Database issue):D169–D174, Jan 2009.

[183] M. Scharf, R. Schneider, G. Casari, P. Bork, A. Valencia, C. Ouzounis, and C. Sander. Genequiz: a workbench for sequence analysis. *Proc Int Conf Intell Syst Mol Biol*, 2:348–353, 1994.

[184] T. Gaasterland and C. W. Sensen. Magpie: automated genome interpretation. *Trends Genet*, 12(2):76–78, Feb 1996.

[185] N. L. Harris. Genotator: a workbench for sequence annotation. *Genome Res*, 7(7):754–762, Jul 1997.

[186] Lincoln D. Stein and Jean Thierry-Mieg. Acedb: A genome database management system. *Computing in Science and Engg.*, 1(3):44–52, 1999.

[187] R. Overbeek, N. Larsen, G. D. Pusch, M. D'Souza, E. Selkov, N. Kyrpides, M. Fonstein, N. Maltsev, and E. Selkov. Wit: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res*, 28(1):123–125, Jan 2000.

[188] Natalia Maltsev, Elizabeth Glass, Dinanath Sulakhe, Alexis Rodriguez, Mustafa H Syed, Tanuja Bompada, Yi Zhang, and Mark D'Souza. Puma2–grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res*, 34(Database issue):D369–D372, Jan 2006.

[189] 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.

[190] 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.

[191] Brooke Rhead, Donna Karolchik, Robert M Kuhn, Angie S Hinrichs, Ann S Zweig, Pauline A Fujita, Mark Diekhans, Kayla E Smith, Kate R Rosenbloom, Brian J Raney, Andy Pohl, Michael Pheasant, Laurence R Meyer, Katrina Learned, Fan Hsu, Jennifer Hillman-Jackson, Rachel A Harte, Belinda Giardine, Timothy R Dreszer, Hiram Clawson, Galt P Barber, David Haussler, and W. James Kent. The ucsc genome browser database: update 2010. *Nucleic Acids Res*, 38(Database issue):D613–D619, Jan 2010.

[192] Paul Flicek, Bronwen L Aken, Benoit Ballester, Kathryn Beal, Eugene Bragin, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Julio Fernandez-Banet, Leo Gordon, Stefan Gräf, Syed Haider, Martin Hammond, Kerstin Howe, Andrew Jenkinson, Nathan Johnson, Andreas Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Gautier Koscielny, Eugene Kulesha, Daniel Lawson, Ian Longden, Tim Massingham, William McLaren, Karine Megy, Bert Overduin, Bethan Pritchard, Daniel Rios, Magali Ruffier, Michael Schuster, Guy Slater, Damian Smedley, Giulietta Spudich, Y. Amy Tang, Stephen Trevanion, Albert Vilella, Jan Vogel, Simon White, Steven P Wilder, Amonida Zadissa, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M Fernández-Suarez, Javier Herrero, Tim J P Hubbard, Anne Parker, Glenn Proctor, James Smith, and Stephen M J Searle. Ensembl's 10th year. *Nucleic Acids Res*, 38(Database issue):D557–D562, Jan 2010.

[193] Alexandre Gattiker, Karine Michoud, Catherine Rivoire, Andrea H Auchincloss, Elisabeth Coudert, Tania Lima, Paul Kersey, Marco Pagni, Christian J A Sigrist, Corinne Lachaize, Anne Lise Veuthey, Elisabeth Gasteiger, and Amos Bairoch. Automated annotation of microbial proteomes in swiss-prot. *Comput Biol Chem*, 27(1):49–58, Feb 2003.

[194] Derek Huntley, Holger Hummerich, Damian Smedley, Sasivimol Kittivoravitkul,

Mark McCarthy, Peter Little, and Marek Sergot. Ganesh: software for customized annotation of genome regions. *Genome Res*, 13(9):2195–2202, Sep 2003.

[195] Gary H Van Domselaar, Paul Stothard, Savita Shrivastava, Joseph A Cruz, AnChi Guo, Xiaoli Dong, Paul Lu, Duane Szafron, Russ Greiner, and David S Wishart. Basys: a web server for automated bacterial genome annotation. *Nucleic Acids Res*, 33(Web Server issue):W455–W459, Jul 2005.

[196] 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.

[197] 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.

[198] 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.

[199] Susan Gottesman. Stealth regulation: biological circuits with small rna switches. *Genes Dev*, 16(22):2829–2842, Nov 2002.

[200] Jörg Vogel and Cynthia Mira Sharma. How to find small non-coding rnas in bacteria. *Biol Chem*, 386(12):1219–1238, Dec 2005.

[201] Shoshy Altuvia. Identification of bacterial small non-coding rnas: experimental approaches. *Curr Opin Microbiol*, 10(3):257–261, Jun 2007.

[202] Jörg Vogel, Verena Bartels, Thean Hock Tang, Gennady Churakov, Jacoba G Slagter-Jäger, Alexander Hüttenhofer, and E. Gerhart H Wagner. Rnomics in escherichia coli detects new srna species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res*, 31(22):6435–6443, Nov 2003.

[203] Mitsuoki Kawano, April A Reynolds, Juan Miranda-Rios, and Gisela Storz. Detection of 5'- and 3'-utr-derived small rnas and cis-encoded antisense rnas in escherichia coli. *Nucleic Acids Res*, 33(3):1040–1050, 2005.

[204] Stephen G Landt, Eduardo Abeliuk, Patrick T McGrath, Joseph A Lesley, Harley H McAdams, and Lucy Shapiro. Small non-coding rnas in caulobacter crescentus. *Mol Microbiol*, 68(3):600–614, May 2008.

[205] Patrick T McGrath, Honglak Lee, Li Zhang, Antonio A Iniesta, Alison K Hottes, Meng How Tan, Nathan J Hillson, Ping Hu, Lucy Shapiro, and Harley H

McAdams. High-throughput identification of transcription start sites, conserved promoter motifs and predicted regulons. *Nat Biotechnol*, 25(5):584–592, May 2007.

[206] D. W. Selinger, K. J. Cheung, R. Mei, E. M. Johansson, C. S. Richmond, F. R. Blattner, D. J. Lockhart, and G. M. Church. Rna expression analysis using a 30 base pair resolution escherichia coli genome array. *Nat Biotechnol*, 18(12):1262–1268, Dec 2000.

[207] Alejandro Toledo-Arana, Olivier Dussurget, Georgios Nikitas, Nina Sesto, Hélène Guet-Revillet, Damien Balestrino, Edmund Loh, Jonas Gripenland, Teresa Tiensuu, Karolis Vaitkevicius, Mathieu Barthelemy, Massimo Vergassola, Marie-Anne Nahori, Guillaume Soubigou, Béatrice Régnault, Jean-Yves Coppée, Marc Lecuit, Jörgen Johansson, and Pascale Cossart. The listeria transcriptional landscape from saprophytism to virulence. *Nature*, 459(7249):950–956, Jun 2009.

[208] Poul Valentin-Hansen, Maiken Eriksen, and Christina Udesen. The bacterial sm-like protein hfq: a key player in rna transactions. *Mol Microbiol*, 51(6):1525–1533, Mar 2004.

[209] Aixia Zhang, Karen M Wassarman, Carsten Rosenow, Brian C Tjaden, Gisela Storz, and Susan Gottesman. Global analysis of small rna and mrna targets of hfq. *Mol Microbiol*, 50(4):1111–1124, Nov 2003.

[210] Alexandra Sittka, Cynthia M Sharma, Katarzyna Rolle, and Jörg Vogel. Deep sequencing of salmonella rna associated with heterologous hfq proteins in vivo reveals small rnas as a major target class and identifies rna processing phenotypes. *RNA Biol*, 6(3):266–275, Jul 2009.

[211] D. R. Yoder-Himes, P. S G Chain, Y. Zhu, O. Wurtzel, E. M. Rubin, James M Tiedje, and R. Sorek. Mapping the burkholderia cenocepacia niche response via high-throughput sequencing. *Proc Natl Acad Sci U S A*, 106(10):3976–3981, Mar 2009.

[212] Karla D Passalacqua, Anjana Varadarajan, Brian D Ondov, David T Okou, Michael E Zwick, and Nicholas H Bergman. Structure and complexity of a bacterial transcriptome. *J Bacteriol*, 191(10):3203–3211, May 2009.

[213] Jane M Liu, Jonathan Livny, Michael S Lawrence, Marc D Kimball, Matthew K Waldor, and Andrew Camilli. Experimental discovery of srnas in vibrio cholerae by direct cloning, 5s/trna depletion and parallel sequencing. *Nucleic Acids Res*, 37(6):e46, Apr 2009.

[214] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold. Correlation between protein and mrna abundance in yeast. *Mol Cell Biol*, 19(3):1720–1730, Mar 1999.

[215] Peng Lu, Christine Vogel, Rong Wang, Xin Yao, and Edward M Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, 25(1):117–124, Jan 2007.

[216] Gaby Haas, Galip Karaali, Karl Ebermayer, Wolfram G Metzger, Stephanie Lamer, Ursula Zimny-Arndt, Susanne Diescher, Ulf B Goebel, Konstanze Vogt, Artur B Roznowski, Bertram J Wiedenmann, Thomas F Meyer, Toni Aebischer, and Peter R Jungblut. Immunoproteomics of helicobacter pylori infection and relation to gastric disease. *Proteomics*, 2(3):313–324, Mar 2002.

[217] Tracie L Williams, Steven R Monday, Sharon Edelson-Mammel, Robert Buchanan, and Steven M Musser. A top-down proteomics approach for differentiating thermal resistant strains of enterobacter sakazakii. *Proteomics*, 5(16):4161–4169, Nov 2005.

[218] Visith Thongboonkerd. Urinary proteomics: towards biomarker discovery, diagnostics and prognostics. *Mol Biosyst*, 4(8):810–815, Aug 2008.

[219] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, Mar 2003.

[220] John R Yates, Annalyn Gilchrist, Kathryn E Howell, and John J M Bergeron. Proteomics of organelles and large cellular structures. *Nat Rev Mol Cell Biol*, 6(9):702–714, Sep 2005.

[221] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[222] Ronald Fisher. *Statistical Methods for Research Workers.* HIPPOCRENE BOOKS, Place, 1970.

[223] B. Bar-Oz, A. Preminger, O. Peleg, C. Block, and I. Arad. Enterobacter sakazakii infection in the newborn. *Acta Paediatr*, 90(3):356–8, 2001. Case Reports Journal Article Norway 1992).

[224] S. M. Townsend, E. Hurrell, J. Caubilla-Barron, C. Loc-Carrillo, and S. J. Forsythe. Characterization of an extended-spectrum beta-lactamase enterobacter hormaechei nosocomial outbreak, and other enterobacter hormaechei misidentified as cronobacter (enterobacter) sakazakii. *Microbiology*, 154(Pt 12):3659–67, 2008. Journal Article England.

[225] K. K. Lai. Enterobacter sakazakii infections among neonates, infants, children, and adults. case reports and a review of the literature. *Medicine (Baltimore)*, 80(2):113–22, 2001. Case Reports Journal Article Review United States.

[226] P. Breeuwer, A. Lardeau, M. Peterz, and H. M. Joosten. Desiccation and heat tolerance of enterobacter sakazakii. *J Appl Microbiol*, 95(5):967–73, 2003. Journal Article England.

[227] S. G. Edelson-Mammel, M. K. Porteous, and R. L. Buchanan. Survival of enterobacter sakazakii in a dehydrated powdered infant formula. *J Food Prot*, 68(9):1900–2, 2005. Journal Article United States.

[228] K. Riedel and A. Lehner. Identification of proteins involved in osmotic stress response in enterobacter sakazakii by proteomics. *Proteomics*, 7(8):1217–31, 2007. Journal Article Research Support, Non-U.S. Gov't Germany.

[229] F. J. Pagotto, M. Nazarowec-White, S. Bidawid, and J. M. Farber. Enterobacter sakazakii: infectivity and enterotoxin production in vitro and in vivo. *J Food Prot*, 66(3):370–5, 2003. Journal Article United States.

[230] M. C. Collado, M. Gueimonde, M. Hernandez, Y. Sanz, and S. Salminen. Adhesion of selected bifidobacterium strains to human intestinal mucus and the role of adhesion in enteropathogen exclusion. *J Food Prot*, 68(12):2672–8, 2005. Journal Article Research Support, Non-U.S. Gov't United States.

[231] J. P. Mange, R. Stephan, N. Borel, P. Wild, K. S. Kim, A. Pospischil, and A. Lehner. Adhesive properties of enterobacter sakazakii to human epithelial and brain microvascular endothelial cells. *BMC Microbiol*, 6:58, 2006. Journal Article England.

[232] M. K. Mohan Nair and K. Venkitanarayanan. Role of bacterial ompa and host cytoskeleton in the invasion of human intestinal epithelial cells by enterobacter sakazakii. *Pediatr Res*, 62(6):664–9, 2007. Journal Article Research Support, U.S. Gov't, Non-P.H.S. United States.

[233] K. P. Kim and M. J. Loessner. Enterobacter sakazakii invasion in human intestinal caco-2 cells requires the host cell cytoskeleton and is enhanced by disruption of tight junction. *Infect Immun*, 76(2):562–70, 2008. Journal Article United States.

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

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

[236] WHO. *Priority eye diseases*, 2008.

[237] WHO. *Global Prevalence and Incidence of Curable STIs.* Genev: WHO, 2001.

[238] MV Kalayoglu and Byrne GI. *The Prokaryotes*, chapter The genus Chlamydia - medical, pages 741–754. New York: Springer, 2006.

[239] Stanley Falkow. Molecular koch's postulates applied to bacterial pathogenicity–a personal recollection 15 years later. *Nat Rev Microbiol*, 2(1):67–72, Jan 2004.

[240] 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.

[241] Rachel Binet and Anthony T Maurelli. Transformation and isolation of allelic exchange mutants of chlamydia psittaci using recombinant dna introduced by electroporation. *Proc Natl Acad Sci U S A*, 106(1):292–297, Jan 2009.

[242] K. D. Everett, R. M. Bush, and A. A. Andersen. Emended description of the order chlamydiales, proposal of parachlamydiaceae fam. nov. and simkaniaceae fam. nov., each containing one monotypic genus, revised taxonomy of the family chlamydiaceae, including a new genus and five new species, and standards for the identification of organisms. *Int J Syst Bacteriol*, 49 Pt 2:415–440, Apr 1999.

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

[244] Klaus-Peter Pleissner, Till Eifert, and Peter R Jungblut. A european pathogenic microorganism proteome database: construction and maintenance. *Comp Funct Genomics*, 3(2):97–100, 2002.

[245] Yan Chen, Peter Timms, and Yi-Ping Phoebe Chen. Cidb: Chlamydia interactive database for cross-querying genomics, transcriptomics and proteomics data. *Biomol Eng*, 24(6):603–608, Dec 2007.

[246] S. Mathews, C. George, C. Flegg, D. Stenzel, and P. Timms. Differential expression of ompa, ompb, pyk, nlpd and cpn0585 genes between normal and interferon-gamma treated cultures of chlamydia pneumoniae. *Microb Pathog*, 30(6):337–345, Jun 2001.

[247] Y.P.P. Chen. *Bioinformatics Technologies*. Springer, 2005.

[248] Robert E Molestina, Jon B Klein, Richard D Miller, William H Pierce, Julio A Ramirez, and James T Summersgill. Proteomic analysis of differentially expressed chlamydia pneumoniae genes during persistent infection of hep-2 cells. *Infect Immun*, 70(6):2976–2981, Jun 2002.

[249] Eva Heinz, Patrick Tischler, Thomas Rattei, Garry Myers, Michael Wagner, and Matthias Horn. Comprehensive in silico prediction and analysis of chlamydial outer membrane proteins reflects evolution and life style of the chlamydiae. *BMC Genomics*, 10:634, 2009.

[250] Masaki Fumoto, Satoru Miyazaki, and Hideaki Sugawara. Genome information broker (gib): data retrieval and comparative analysis system for completed microbial genomes and more. *Nucleic Acids Res*, 30(1):66–68, Jan 2002.

[251] Ikuo Uchiyama. Mbgd: microbial genome database for comparative analysis. *Nucleic Acids Res*, 31(1):58–62, Jan 2003.

[252] Ikuo Uchiyama. Mbgd: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res*, 35(Database issue):D343–D346, Jan 2007.

[253] Ikuo Uchiyama, Toshio Higuchi, and Mikihiko Kawai. Mbgd update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic Acids Res*, 38(Database issue):D361–D365, Jan 2010.

[254] Wayne W Eckerson. Three tier client/server architecture: Achieving scalability, performance, and efficiency in client server applications. *Open Information Systems*, 10, 1995.

[255] J. D. WATSON and F. H. CRICK. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.

[256] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C Hsieh, Deborah A Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E Gruber. Bigtable: A distributed storage system for structured data. *OSDI'06: Seventh Symposium on Operating System Design and Implementation, Seattle, WA*, 2006.

[257] Roger Stephan, Angelika Lehner, Patrick Tischler, and Thomas Rattei. Complete genome sequence of cronobacter turicensis lmg 23827, a food-borne pathogen causing deaths in neonates. *J Bacteriol*, 193(1):309–310, Jan 2011.

[258] A. Zihler, G. Le Blay, T. de Wouters, C. Lacroix, C. P. Braegger, A. Lehner, P. Tischler, T. Rattei, H. Hächler, and R. Stephan. In vitro inhibition activity of different bacteriocin-producing escherichia coli against salmonella strains isolated from clinical cases. *Lett Appl Microbiol*, 49(1):31–38, Jul 2009.

[259] 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.

[260] Frank Maixner, Michael Wagner, Sebastian Lücker, Eric Pelletier, Stephan Schmitz-Esser, Karin Hace, Eva Spieck, Robert Konrat, Denis Le Paslier, and Holger Daims. Environmental genomics reveals a functional chlorite dismutase in the nitrite-oxidizing bacterium 'candidatus nitrospira defluvii'. *Environ Microbiol*, 10(11):3043–3056, Nov 2008.

[261] Anja Spang, Roland Hatzenpichler, Céline Brochier-Armanet, Thomas Rattei, Patrick Tischler, Eva Spieck, Wolfgang Streit, David A Stahl, Michael Wagner, and Christa Schleper. Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum thaumarchaeota. *Trends Microbiol*, 18(8):331–340, Aug 2010.

[262] Nidal Abu Laban, Draženka Selesi, Thomas Rattei, Patrick Tischler, and Rainer U Meckenstock. Identification of enzymes involved in anaerobic benzene degradation by a strictly anaerobic iron-reducing enrichment culture. *Environ Microbiol*, Jun 2010.

[263] Draženka Selesi, Nico Jehmlich, Martin von Bergen, Frank Schmidt, Thomas Rattei, Patrick Tischler, Tillmann Lueders, and Rainer U Meckenstock. Combined genomic and proteomic approaches identify gene clusters involved in anaerobic 2-methylnaphthalene degradation in the sulfate-reducing enrichment culture n47. *J Bacteriol*, Oct 2009.

[264] Franz Bergmann, Draženka Selesi, Thomas Weinmaier, Patrick Tischler, Thomas Rattei, and Rainer U Meckenstock. Genomic insights into the metabolic potential of the polycyclic aromatic hydrocarbon degrading sulfate-reducing deltaproteobacterium n47. *Environ Microbiol*, Dec 2010.

[265] Jarrod A Chapman, Ewen F Kirkness, Oleg Simakov, Steven E Hampson, Therese Mitros, Thomas Weinmaier, Thomas Rattei, Prakash G Balasubramanian, Jon Borman, Dana Busam, Kathryn Disbennett, Cynthia Pfannkoch, Nadezhda Sumin, Granger G Sutton, Lakshmi Devi Viswanathan, Brian Walenz, David M Goodstein, Uffe Hellsten, Takeshi Kawashima, Simon E Prochnik, Nicholas H Putnam, Shengquiang Shu, Bruce Blumberg, Catherine E Dana, Lydia Gee, Dennis F Kibler, Lee Law, Dirk Lindgens, Daniel E Martinez, Jisong Peng, Philip A Wigge, Bianca Bertulat, Corina Guder, Yukio Nakamura, Suat Ozbek, Hiroshi Watanabe, Konstantin Khalturin, Georg Hemmrich, André Franke, René Augustin, Sebastian Fraune, Eisuke Hayakawa, Shiho Hayakawa, Mamiko Hirose, Jung Shan Hwang, Kazuho Ikeo, Chiemi Nishimiya-Fujisawa, Atshushi Ogura, Toshio Takahashi, Patrick R H Steinmetz, Xiaoming Zhang, Roland Aufschnaiter, Marie-Kristin Eder, Anne-Kathrin Gorny, Willi Salvenmoser, Alysha M Heimberg, Benjamin M Wheeler, Kevin J Peterson, Angelika Böttger, Patrick Tischler, Alexander Wolf, Takashi Gojobori, Karin A Remington, Robert L Strausberg, J. Craig Venter, Ulrich Technau, Bert Hobmayer, Thomas C G Bosch, Thomas W Holstein, Toshitaka Fujisawa, Hans R Bode, Charles N David, Daniel S Rokhsar, and Robert E Steele. The dynamic genome of hydra. *Nature*, 464(7288):592–596, Mar 2010.

[266] Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, May 2007.

[267] UniProt Consortium. The universal protein resource (uniprot) in 2010. *Nucleic Acids Res*, 38(Database issue):D142–D148, Jan 2010.

[268] Ok-Sun Kim, Pilar Junier, Johannes F Imhoff, and Karl-Paul Witzel. Comparative analysis of ammonia monooxygenase (amoa) genes in the water column and sediment-water interface of two lakes and the baltic sea. *FEMS Microbiol Ecol*, 66(2):367–378, Nov 2008.

[269] 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.

[270] P. Rice, I. Longden, and A. Bleasby. Emboss: the european molecular biology open software suite. *Trends Genet*, 16(6):276–7, 2000. Journal Article England Tig.

[271] 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.

[272] Emmanuelle Lerat and Howard Ochman. Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res*, 33(10):3125–3132, 2005.

[273] Nancy A Moran and Gordon R Plague. Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev*, 14(6):627–633, Dec 2004.

[274] Patricia Siguier, Jonathan Filée, and Michael Chandler. Insertion sequences in prokaryotic genomes. *Curr Opin Microbiol*, 9(5):526–531, Oct 2006.

[275] 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.

[276] D. M. Martin, M. Berriman, and G. J. Barton. Gotcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, 5:178, 2004. Journal Article Research Support, Non-U.S. Gov't England.

[277] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 37(Database issue):D5–15, 2009. Journal Article England.

[278] T. M. Lowe and S. R. Eddy. trnascan-se: a program for improved detection of transfer rna genes in genomic sequence. *Nucleic Acids Res*, 25(5):955–64, 1997. Journal Article England.

[279] B. P. Westover, J. D. Buhler, J. L. Sonnenburg, and J. I. Gordon. Operon prediction without a training set. *Bioinformatics*, 21(7):880–8, 2005. CDK30292/DK/NIDDK NIH HHS/United States Comparative Study Evaluation Studies Journal Article Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. Validation Studies England.

[280] J. Krumsiek, R. Arnold, and T. Rattei. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, 23(8):1026–8, 2007. Journal Article England.

[281] C. M. Chen, T. K. Misra, S. Silver, and B. P. Rosen. Nucleotide sequence of the structural genes for an anion pump. the plasmid-encoded arsenical resistance operon. *J Biol Chem*, 261(32):15030–8, 1986. AI15682/AI/NIAID NIH HHS/United States AI19793/AI/NIAID NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. United states.

[282] M. J. San Francisco, C. L. Hope, J. B. Owolabi, L. S. Tisa, and B. P. Rosen. Identification of the metalloregulatory element of the plasmid-encoded arsenical resistance operon. *Nucleic Acids Res*, 18(3):619–24, 1990. AI07375/AI/NIAID NIH HHS/United States AI19793/AI/NIAID NIH HHS/United States GM12187/GM/NIGMS NIH HHS/United States Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. England.

[283] J. Wu and B. P. Rosen. The arsd gene encodes a second trans-acting regulatory protein of the plasmid-encoded arsenical resistance operon. *Mol Microbiol*, 8(3):615–23, 1993. AI19793/AI/NIAID NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. England.

[284] S. Dey, D. Dou, and B. P. Rosen. Atp-dependent arsenite transport in everted membrane vesicles of escherichia coli. *J Biol Chem*, 269(41):25442–6, 1994. AI19793/AI/NIAID NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. United states.

[285] C. M. Hsu and B. P. Rosen. Characterization of the catalytic subunit of an anion pump. *J Biol Chem*, 264(29):17349–54, 1989. AI19713/AI/NIAID NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. United states.

[286] S. Dey, D. Dou, L. S. Tisa, and B. P. Rosen. Interaction of the catalytic and the membrane subunits of an oxyanion-translocating atpase. *Arch Biochem Biophys*, 311(2):418–24, 1994. AI19793/AI/NIAID NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. United states.

[287] L. S. Tisa and B. P. Rosen. Molecular characterization of an anion pump. the arsb protein is the membrane anchor for the arsa protein. *J Biol Chem*, 265(1):190–4, 1990. AI19793/AI/NIAID NIH HHS/United States GM12187/GM/NIGMS NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. United states.

[288] D. Dou, S. Dey, and B. P. Rosen. A functional chimeric membrane subunit of an ion-translocating atpase. *Antonie Van Leeuwenhoek*, 65(4):359–68, 1994.

AI19793/AI/NIAID NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. Netherlands.

[289] S. Dey and B. P. Rosen. Dual mode of energy coupling by the oxyanion-translocating arsb protein. *J Bacteriol*, 177(2):385–9, 1995. AI19793/AI/NIAID NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. United states.

[290] D. E. Taylor. Bacterial tellurite resistance. *Trends Microbiol*, 7(3):111–115, Mar 1999.

[291] A. J. Te Velthuis and C. P. Bagowski. Linking fold, function and phylogeny: a comparative genomics view on protein (domain) evolution. *Curr Genomics*, 9(2):88–96, 2008. Journal Article Netherlands.

[292] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. Interproscan: protein domains identifier. *Nucleic Acids Res*, 33(Web Server issue):W116–W120, Jul 2005.

[293] R. C. Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, 2004. Comparative Study Journal Article England.

[294] J. Felsenstein. Phylip – phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[295] I. Letunic and P. Bork. Interactive tree of life (itol): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–8, 2007. Journal Article England.

[296] M. Schmid, C. Iversen, I. Gontia, R. Stephan, A. Hofmann, A. Hartmann, B. Jha, L. Eberl, K. Riedel, and A. Lehner. Evidence for a plant-associated natural habitat for cronobacter spp. *Res Microbiol*, 160(8):608–14, 2009. Journal Article France.

[297] I. K. Toth, L. Pritchard, and P. R. Birch. Comparative genomics reveals what makes an enterobacterial plant pathogen. *Annu Rev Phytopathol*, 44:305–36, 2006. Journal Article Research Support, Non-U.S. Gov't Review United States.

[298] C. Fraipont, F. Sapunaric, A. Zervosen, G. Auger, B. Devreese, T. Lioux, D. Blanot, D. Mengin-Lecreulx, P. Herdewijn, J. Van Beeumen, J. M. Frere, and M. Nguyen-Disteche. Glycosyl transferase activity of the escherichia coli penicillin-binding protein 1b: specificity profile for the substrate. *Biochemistry*, 45(12):4007–13, 2006. Journal Article Research Support, Non-U.S. Gov't United States.

[299] F. Ishino, W. Park, S. Tomioka, S. Tamaki, I. Takase, K. Kunugita, H. Matsuzawa, S. Asoh, T. Ohta, B. G. Spratt, and et al. Peptidoglycan synthetic activities in membranes of escherichia coli caused by overproduction of penicillin-binding

protein 2 and roda protein. *J Biol Chem*, 261(15):7024–31, 1986. Journal Article Research Support, Non-U.S. Gov't United states.

[300] A. Philippon, G. Arlet, and G. A. Jacoby. Plasmid-determined ampc-type beta-lactamases. *Antimicrob Agents Chemother*, 46(1):1–11, 2002. Journal Article Review United States.

[301] E. G. Walter and D. E. Taylor. Plasmid-mediated resistance to tellurite: expressed and cryptic. *Plasmid*, 27(1):52–64, 1992. Journal Article Research Support, Non-U.S. Gov't Review United states.

[302] D. E. Taylor, Y. Hou, R. J. Turner, and J. H. Weiner. Location of a potassium tellurite resistance operon (teha tehb) within the terminus of escherichia coli k-12. *J Bacteriol*, 176(9):2740–2, 1994. Journal Article Research Support, Non-U.S. Gov't United states.

[303] V. K. Singamsetty, Y. Wang, H. Shimada, and N. V. Prasadarao. Outer membrane protein a expression in enterobacter sakazakii is required to induce microtubule condensation in human brain microvascular endothelial cells for invasion. *Microb Pathog*, 45(3):181–91, 2008. AI40567/AI/NIAID NIH HHS/United States R01 AI040567-13/AI/NIAID NIH HHS/United States Comparative Study Journal Article Research Support, N.I.H., Extramural England.

[304] Rémi Fronzes, Peter J Christie, and Gabriel Waksman. The structural biology of type iv secretion systems. *Nat Rev Microbiol*, 7(10):703–714, Oct 2009.

[305] Emmy De Buck, Elke Lammertyn, and Jozef Anné. The importance of the twin-arginine translocation pathway for bacterial virulence. *Trends Microbiol*, 16(9):442–453, Sep 2008.

[306] Nalvo F Almeida, Shuangchun Yan, Magdalen Lindeberg, David J Studholme, David J Schneider, Bradford Condon, Haijie Liu, Carlos J Viana, Andrew Warren, Clive Evans, Eric Kemen, Dan Maclean, Aurelie Angot, Gregory B Martin, Jonathan D Jones, Alan Collmer, Joao C Setubal, and Boris A Vinatzer. A draft genome sequence of pseudomonas syringae pv. tomato t1 reveals a type iii effector repertoire significantly divergent from that of pseudomonas syringae pv. tomato dc3000. *Mol Plant Microbe Interact*, 22(1):52–62, Jan 2009.

[307] Daniel E Voth and Robert A Heinzen. Coxiella type iv secretion and cellular microbiology. *Curr Opin Microbiol*, 12(1):74–80, Feb 2009.

[308] David Parcej and Robert Tampé. Abc proteins in antigen translocation and viral inhibition. *Nat Chem Biol*, 6(8):572–580, Aug 2010.

[309] Amy L Davidson, Elie Dassa, Cedric Orelle, and Jue Chen. Structure, function, and evolution of bacterial atp-binding cassette systems. *Microbiol Mol Biol Rev*, 72(2):317–64, table of contents, Jun 2008.

[310] M. A. Delgado, P. A. Vincent, R. N. Farias, and R. A. Salomon. Yoji of escherichia coli functions as a microcin j25 efflux pump. *J Bacteriol*, 187(10):3465–70, 2005. Journal Article Research Support, Non-U.S. Gov't United States.

[311] M. El Ghachi, A. Bouhss, D. Blanot, and D. Mengin-Lecreulx. The baca gene of escherichia coli encodes an undecaprenyl pyrophosphate phosphatase activity. *J Biol Chem*, 279(29):30106–13, 2004. Journal Article Research Support, Non-U.S. Gov't United States.

[312] B. P. Nichols and G. G. Guay. Gene amplification contributes to sulfonamide resistance in escherichia coli. *Antimicrob Agents Chemother*, 33(12):2042–8, 1989. AI125106/AI/NIAID NIH HHS/United States AI18639/AI/NIAID NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. United states.

[313] J. Berglez, P. Iliades, W. Sirawaraporn, P. Coloe, and I. Macreadie. Analysis in escherichia coli of plasmodium falciparum dihydropteroate synthase (dhps) alleles implicated in resistance to sulfadoxine. *Int J Parasitol*, 34(1):95–100, 2004. 1 R01AI46966-01A1/AI/NIAID NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. England.

[314] Sean-Paul Nuccio and Andreas J Bäumler. Evolution of the chaperone/usher assembly pathway: fimbrial classification goes greek. *Microbiol Mol Biol Rev*, 71(4):551–575, Dec 2007.

[315] T. Touze, J. Eswaran, E. Bokma, E. Koronakis, C. Hughes, and V. Koronakis. Interactions underlying assembly of the escherichia coli acrab-tolc multidrug efflux system. *Mol Microbiol*, 53(2):697–706, 2004. Journal Article Research Support, Non-U.S. Gov't England.

[316] Milton H Saier, Ming Ren Yen, Keith Noto, Dorjee G Tamang, and Charles Elkan. The transporter classification database: recent advances. *Nucleic Acids Res*, 37(Database issue):D274–D278, Jan 2009.

[317] S. Y. Lau and H. I. Zgurskaya. Cell division defects in escherichia coli deficient in the multidrug efflux transporter acref-tolc. *J Bacteriol*, 187(22):7815–25, 2005. 1-R01-AI052293-01A1/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, P.H.S. United States.

[318] L. J. Piddock. Multidrug-resistance efflux pumps - not just for resistance. *Nat Rev Microbiol*, 4(8):629–36, 2006. Journal Article Review England.

[319] K. Nishino and A. Yamaguchi. Analysis of a complete library of putative drug transporter genes in escherichia coli. *J Bacteriol*, 183(20):5803–12, 2001. Journal Article Research Support, Non-U.S. Gov't United States.

[320] N. Kobayashi, K. Nishino, and A. Yamaguchi. Novel macrolide-specific abc-type efflux transporter in escherichia coli. *J Bacteriol*, 183(19):5639–44, 2001. Journal Article Research Support, Non-U.S. Gov't United States.

[321] O. Lomovskaya and K. Lewis. Emr, an escherichia coli locus for multidrug resistance. *Proc Natl Acad Sci U S A*, 89(19):8938–42, 1992. Comparative Study Journal Article Research Support, Non-U.S. Gov't United states.

[322] G. Kastenmuller, M. E. Schenk, J. Gasteiger, and H. W. Mewes. Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes. *Genome Biol*, 10(3):R28, 2009. Journal Article England.

[323] M. Reitz, O. Sacher, A. Tarkhov, D. Trumbach, and J. Gasteiger. Enabling the exploration of biochemical pathways. *Org Biomol Chem*, 2(22):3226–37, 2004. Journal Article Research Support, Non-U.S. Gov't England.

[324] G. Kastenmuller, J. Gasteiger, and H. W. Mewes. An environmental perspective on large-scale genome clustering based on metabolic capabilities. *Bioinformatics*, 24(16):i56–62, 2008. Journal Article England.

[325] R. Bentley and R. Meganathan. Biosynthesis of vitamin k (menaquinone) in bacteria. *Microbiol Rev*, 46(3):241–80, 1982. GM 20053/GM/NIGMS NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. Review United states.

[326] B. A. Haddock and C. W. Jones. Bacterial respiration. *Bacteriol Rev*, 41(1):47–99, 1977. Journal Article Review United states.

[327] A. Singh and D. K. Singh. Molluscicidal activity of lawsonia inermis and its binary and tertiary combinations with other plant derived molluscicides. *Indian J Exp Biol*, 39(3):263–8, 2001. Journal Article Research Support, Non-U.S. Gov't India.

[328] J. E. Kelmanson, A. K. Jager, and J. van Staden. Zulu medicinal plants with antibacterial activity. *J Ethnopharmacol*, 69(3):241–6, 2000. Journal Article Research Support, Non-U.S. Gov't Ireland.

[329] T. Okano, Y. Shimomura, M. Yamane, Y. Suhara, M. Kamao, M. Sugiura, and K. Nakagawa. Conversion of phylloquinone (vitamin k1) into menaquinone-4 (vitamin k2) in mice: two possible routes for menaquinone-4 accumulation in cerebra of mice. *J Biol Chem*, 283(17):11270–9, 2008. Journal Article Research Support, Non-U.S. Gov't United States.

[330] R. T. Davidson, A. L. Foley, J. A. Engelke, and J. W. Suttie. Conversion of dietary phylloquinone to tissue menaquinone-4 in rats is not dependent on gut bacteria. *J Nutr*, 128(2):220–3, 1998. 5 PO1 DK14881/DK/NIDDK NIH HHS/United States 5 T32 DK07665/DK/NIDDK NIH HHS/United States Comparative Study Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. United states.

[331] 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.

[332] P. C. Ng and S. Henikoff. Predicting deleterious amino acid substitutions. *Genome Res*, 11(5):863–874, May 2001.

[333] Marco Albrecht, Cynthia M Sharma, Richard Reinhardt, Jörg Vogel, and Thomas Rudel. Deep sequencing-based discovery of the chlamydia trachomatis transcriptome. *Nucleic Acids Res*, 38(3):868–877, Jan 2010.

List of Figures

List of Tables

Publications

1. Götz S, Arnold R, Sebastían-Léon P, Martín-Rodríguez S, **Tischler P**, Jehl MA, Dopazo J, Rattei T, Conesa A. *B2G-FAR, a species centered GO annotation repository.* Bioinformatics, in press.

2. Bergmann F, Selesi D, Weinmaier T, **Tischler P**, Rattei T, Meckenstock RU. *Genomic insights into the metabolic potential of the polycyclic aromatic hydrocarbon degrading sulfate-reducing Deltaproteobacterium N47.* Environ Microbiol. 2010 Dec 22. doi: 10.1111/j.1462-2920.2010.02391.x. [Epub ahead of print]

3. Stephan R, Lehner A, **Tischler P**, Rattei T. *Complete Genome Sequence of Cronobacter turicensis LMG 23827, a Food-Borne Pathogen Causing Deaths in Neonates.* J Bacteriol. 2011 Jan;193(1):309-10. Epub 2010 Oct 29.

4. Spang A, Hatzenpichler R, Brochier-Armanet C, Rattei T, **Tischler P**, Spieck E, Streit W, Stahl DA, Wagner M, Schleper C. *Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota.* Trends Microbiol. 2010 Aug;18(8):331-40. Epub 2010 Jul 2.

5. Abu Laban N, Selesi D, Rattei T, **Tischler P**, Meckenstock RU. *Identification of enzymes involved in anaerobic benzene degradation by a strictly anaerobic iron-reducing enrichment culture.* Environ Microbiol. 2010 Jun 1. [Epub ahead of print]

6. Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, Rattei T, Balasubramanian PG, Borman J, Busam D, Disbennett K, Pfannkoch C, Sumin N, Sutton GG, Viswanathan LD, Walenz B, Goodstein DM, Hellsten U, Kawashima T, Prochnik SE, Putnam NH, Shu S, Blumberg B, Dana CE, Gee L, Kibler DF, Law L, Lindgens D, Martinez DE, Peng J, Wigge PA, Bertulat B, Guder C, Nakamura Y, Ozbek S, Watanabe H, Khalturin K, Hemmrich G, Franke A, Augustin R, Fraune S, Hayakawa E, Hayakawa S, Hirose M, Hwang JS, Ikeo K, Nishimiya-Fujisawa C, Ogura A, Takahashi T, Steinmetz PR, Zhang X, Aufschnaiter R, Eder MK, Gorny AK, Salvenmoser W, Heimberg AM, Wheeler BM, Peterson KJ, Böttger A, **Tischler P**, Wolf A, Gojobori T, Remington KA, Strausberg RL, Venter JC, Technau U, Hobmayer B, Bosch TC, Holstein TW, Fujisawa T, Bode HR, David CN, Rokhsar DS, Steele RE. *The dynamic genome of Hydra.* Nature. 2010 Mar 14. [Epub ahead of print]

7. Rattei T, **Tischler P**, Götz S, Jehl MA, Hoser J, Arnold R, Conesa A, Mewes HW. *SIMAPâa 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 11.

8. Selesi D, Jehmlich N, von Bergen M, Schmidt F, Rattei T, **Tischler P**, Lueders T, Meckenstock RU. *Combined genomic and proteomic approaches identify gene clus-*

ters involved in anaerobic 2-methylnaphthalene degradation in the sulfate-reducing enrichment culture N47. J Bacteriol. 2010 Jan;192(1):295-306. Epub.

9. Heinz E, **Tischler P**, Rattei T, Myers G, Wagner M, Horn M. *Comprehensive in silico prediction and analysis of chlamydial outer membrane proteins reflects evolution and life style of the Chlamydiae.* BMC Genomics. 2009 Dec 29;10(1):634.

10. 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. 2009 Dec 18.

11. Zihler A, Le Blay G, de Wouters T, Lacroix C, Braegger CP, Lehner A, **Tischler P**, Rattei T, Hächler H, Stephan R. *In vitro inhibition activity of different bacteriocin-producing Escherichia coli against Salmonella strains isolated from clinical cases.* Lett Appl Microbiol. 2009 Jul;49(1):31-8. Epub 2009 Apr 17.

12. 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 Pathog. 2009 Apr;5(4):e1000376. Epub 2009 Apr 24. Erratum in: PLoS Pathog. 2009 Apr;5(4). doi: 10.1371/annotation/78659a32-7869-4b14-91a6-b301a588d937.

13. 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. Epub 2008 Oct 21.

14. 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. Epub 2008 Jul 21.

15. 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. Epub 2007 Nov 23