Tools for functional genomics applied to Staphylococci, Listeriae, Vaccinia virus and other organisms

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades der
Bayerischen Julius-Maximilians-Universität Würzburg

Vorgelegt von
Chunguang Liang, M.Sc.
aus Shandong, V.R.C

Würzburg 2009

Introduction

Problem and challenges

We live in the new age of post-genomic biology. This means that already a large amount of data is known, most importantly many genome sequences. “Post-genomics tries to put these data into context, for instances connecting data sets from metabolism, proteomics and transcriptomics to these pre-existent genomics data.” Furthermore, new efforts and methods allow to deal with biological systems as a whole and in a systematic manner to reveal their inbuild system properties. This is the advent of systems biology, which encompasses the full scale of complexity inherent in the just mentioned data sets to reveal the hidden interdependencies of the system and how it reacts and adapts to the environment. This causes a requirement for additional data and individual data-sets because only massive data on all levels of the system make a quantitative description of biological systems possible and their different possibilities to react to stimuli, stress and environmental factors. Current research on such large, often genome-scale data set faces thus a couple of prominent difficulties: Due to large data sets, there often is an extremely long time of calculation. This is already the case for the most basic calculations which are possible in genomics, for instance sequence comparisons. Only with the advent of the modern versions of fast heuristics for such rapid alignments such as BLAST (Altschul et al. 1997), it is possible to compare new sequences to the millions of entries and the very long DNA sequences stored (billions of nucleotides). However, these queries and sequence alignment calculations pose typical challenges to computer systems which can be dealt by a combination of hardware and a heuristic, i.e. a non exact search. However, there are more and more challenging data sets in current studies, i.e., the time-consumption is not in a linear relationship to the data size, and mostly exponential. This applies to data sets of transcriptomics, for instance regarding RNA folding, searches for matches of RNA structures in different sequences or even genomes and in general analyses of RNA structures. Similarly, structural comparisons in proteins lead to long calculations which quite often have exponential requirements for calculation and storage space, require heuristics for rapid and applied

solutions and are, in fact, for the full protein folding problem NP-hard problems: There is no polynome which can describe the required search time for the calculation algorithm. This theme appears again in various other proteomics data sets and calculation problems in large-scale data sets. They all are NP-hard and require not only very fast hardware and calculation algorithms but in fact depend on a clever heuristic to cope with these time for calculation problems. Moreover, interdependencies between system components drastically increase the calculation time, most evident in calculations regarding genome-scale metabolic networks, metabolic fluxes and involved enzymes and pathways. Here, even best strategies may lead the elementary flux modes in the network incalculable, the so-called combinatorial explosion (Schuster et al. 2002). The interdependencies between data sets thus dramatically increase the modelling complexity. Some of these different challenges of modern genome-scale biological data are dealt with in this thesis. Specifically, we follow the flow of information from the genome via transcriptome to proteome and metabolites and devise here a number of different software applications which always greatly enhance our capability to deal with the calculation challenges involved. First, inGeno allows rapid genome-scale comparisons for many prokaryotic genomes. Second, JANE deals with the challenge of mapping large numbers of ESTs to a template genome and even with the additional challenge that this genome template is not yet known, so mapping has to be done to a surrogate template (important for instance in single cell sequencing). We then turn our research on metabolic modelling and calculate the physiological changes according to metabolite data and proteomic data. This challenge has no publicly available software can support, we promote the algorithm and implement in the YANAsquare package (Schwarz et al. 2007), calculate the null-space. The hard-convergence problem due to large number of modes is solved using YANAvergence (a non-linear optimization routine implemented in GNU R). Besides this research flowchart, a platform (GENOVA for functional genomics, Tardigrade workbench and JANE) is prepared for extending our research also on other species (*S. aureus* RN1, NCTC 8325) and organisms (Tardigrada), transcriptome data and genome data is analyzed. Vaccinia virus provides a promising way to selectively lyse the tumor cells, and thus we sequence and conduct the analysis on potential genes related to oncolysis

and tissue-tropisms. The optimized strains, GLV-1h68, GLV-ONC1 have been in a clinical trial state of cancer therapy.

Genus *Listeria*

Listeriae are gram-positive, motile, facultative anaerobic rod-like bacteria that are ubiquitously occurring in nature. *Listeria* are members of a group of bacteria with low G+C DNA content that includes species of the genera *Bacillus*, *Clostridium*, *Enterococcus*, *Streptococcus*, and *Staphylococcus*. Currently the genus *Listeria* consists of six different species: *Listeria monocytogenes*, *Listeria ivanovii*, *Listeria innocua*, *Listeria welshimeri*, *Listeria seeligeri*, and *Listeria grayi* (Sallen et al. 1996, Collins et al. 1991).

Two species among them, *L. monocytogenes* and *L. ivanovii* have already been recognized currently as pathogen, which can lead to listeriosis, an opportunistic infection of humans as well as animals. The most often-seen symptoms are abortion, septicemia, meningitis, meningoencephalitis, perinatal infections and gastroenteritis (Vazquez-Boland et al. 2001). The natural habitat of *Listeria* is thought to be the surface layer of soil rich in decaying plant matter. From this, they gain opportunity to enter vertebrate host via the oral route after consuming contaminated food. The infectious host range of *L. monocytogenes* is relatively wide including mammals and birds, while *L. ivanovii* can only be pathogenic to ruminant species. Both *L. monocytogenes* and *L. ivanovii* are typically facultative intracellular parasites (Farber and Peterkin 1991, Meier and Lopez 2001). They have capabilities of proliferating within macrophages and a variety of normally nonphagocytic cells, such as epthelial cells, endothelial cells and hepatocytes. In all these cell types, pathogenic *Listeria* develops a characteristic intracellular life cycle spreading even from one infected cell to neighboring cells illustrated in the figure 1 (adapted from Tilny and Portnoy 1989). It begins with *Listeria* escaping from the phagocytic vacuole, entering the cytoplasm of the host cells. Following replication in the cytoplasm, they can move by action polymerization. Occasionally, formed protrusions on the cell membrane can excavate a passage to the neighbouring cell with help of phagocytosis

process. The up-taken pathogen will initiate the next cycle in the new host cell (Dabiri et al. 1990, Tilney et al. 1992).

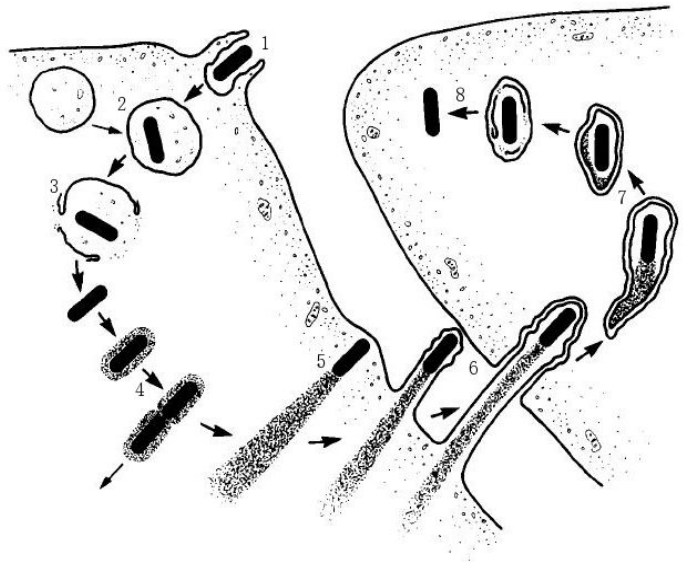

Figure 1: Schema of intracellular life cycle of pathogenic *Listeria*.

1 Entry into host cells; 2 Survival within the phagocytic vacuole; 3 Disruption of phagosomal membranes and escape into the cytosol; 4 Replication in the cytosol; 5 Actin-based motility; 6 Direct cell-to-cell spread; 7 Survival in secondary; 8 Escape from secondary phagosomes and initiation of the next cycle.

In addition to these two pathogens, there are four non-pathogenic species of *Listeria*, i.e., *L. innocua*, *L. welshimeri*, *L. seeligeri*, and *L. grayi*. Phylogenetic analyses, based on the 16S and 23S rRNA genes and the *iap*, *prs*, *vclA*, *vclB*, and *ldh* genes, indicate that *L. innocua* is highly related to *L. monocytogenes*. The second group contains *L. ivanovii*, together with *L. seeligeri*, while *L. welshimeri* is more distant, exhibiting the deepest branching of this group. *L. grayi* seems to be distant from these two groups (Schmid et al. 2005, Cossart and Portnoy 2001). *L. welshimeri* (SLCC5334, CIP8149, and Welshimer V8) was first isolated from decaying plants and is a serovar 6b strain; other serovars (1/2a, 1/2b, 6a, 4c, and 4f) have also been reported for this species (Welshimeri 1968, Rocourt and Seeliger 1985). Like other species of *Listeria*, *L. welshimeri* bacteria are small (0.5 to 2.0 µm), nonspore-forming, gram-positive rods which are motile below 30°C by means of peritrichous flagella. Growth at low temperatures (4°C) proceeds within 5 days. Results from a CAMP test with *Staphylococcus aureus* and *Rhodococcus*

equi were negative, and strains of the species also tested negative for oxidase were positive for catalase activity. Acid production occurs by fermentation of D-xylose and alpha-methyl-D-mannoside but not from L-rhamnose and D-mannitol (Rocourt et al. 1983). These biochemical properties are used to distinguish *L. welshimeri* from other *Listeria* species. We have already inspected the virulence islands (Listeria Pathogenicity Island-1: LIPI-1) and the comparison reveals different structure of these loci (Vazquez-Boland et al. 2001). In this project, the metabolism system of *L. welshimeri* is comprehensively investigated and compared to the pathogen strain *L. monocytogenes* (Hain et al. 2006)

Staphylococcus aureus

The gram-positive bacterial pathogen, *Staphylococcus aureus*, leads to severe nosocomial infections, life-threatening also in the large community (Steinberg et al. 1996, Lowy 1998). Its major habitats are the nasal membranes and skin of warm-blooded animals or human, in which/whom it causes a range of disease, such as skin infections, food poisoning, even sepsis, osteomyelitis and pneumonia. The invaded pathogen is able to produce a variety of enterotoxins. It became most notorious since it developed resistance to some of most potent antibiotics, in particular to methicillin and vancomycin (Walsh and Howe 2002). Although the mechanisms of antibiotic resistance and infection have been elucidated, only few studies have focused on the basic cellular physiology of *S. aureus*. Meanwhile, the availability of limited extensive proteomic (Hecker et al. 2003, Cordwell et al. 2002, Fuchs et al. 2007) and transcriptional data (Dunman et al. 2004, Kuroda et al. 2003) enables to approach staphylococcal physiology with systems biological methods and predict novel potential targets for the future therapy of *aureus* infections. In this study, the construction and merits of different *Staphylococcus aureus* strains, i.e., NCTC 8325, RN1 mutant and curated strains, are discussed and virulence factors are investigated using the GENOVA software. The physiological data analyzed with the genome-scale constructed network has been approached separately under aerobic or hypoxia condition, however some results have not been included in this thesis since they have not been published yet.

Vaccinia virus

Vaccinia virus (VACV) is the prototype virus of the orthopoxvirus (OPV) genus in the poxvirus family. It has a linear, approximately 190-kb, double-stranded DNA genome, encoding more than 200 ORFs. VACV strains have been used extensively as vaccines and have played a central role in the eradication of the smallpox disease. More recently, research has focused on the potential of using VACV as an oncolytic virus for therapy of tumors. Many features associated with VACV are considered valuable for viral therapy, including the large cloning capacity, the natural tumor colonization capability, a short life cycle with strong lytic ability, and the capability to infect and replicate in human cells without causing natural disease in humans (Shen and Nemunaitis 2005). Several VACV vectors have shown remarkable antitumor and antimetastases results in preclinical studies; however, the significant level of infection in other organs remains a safety concern for systemic administration (Thorne et al. 2005).

VACV strains exhibit variations in virulence, as well as variations in host and tissue specificity or tissue tropism. Sequence analysis of VACV genomes has improved our understanding of the potential functions of viral gene products and host-virus interactions (Goebel et al. 1990, Upton et al. 2003). It is also known that several nonessential genes, such as J2R (thymidine kinase TK: Buller et al. 1985), C11R (secreted epidermal growth factor-like: Buller et al. 1988), A56R (hemagglutinin HA: Shida et al. 1988), and B8R (soluble interferon-gamma receptor-like: Verardi et al. 2001) result in reduced virulence when deleted or disrupted experimentally. LIVP is a VACV vaccine strain originated from the Lister strain, which was adapted to calf skin in the Institute of Viral Preparations, Moscow, Russia (AI'tschtein et al. 1985). Western Reserve (WR: Henderson and Moss 1999) is derived from the New York City Board of Health (NYCBH) strain by repeated passages in the mouse brain.

VACV GLV-1h68 may be a potential improvement for medical therapies, we sequenced the genomic DNAs of wt-LIVP and GLV-1h68. Therefore its attenuation basis is studied here first by

triple genome comparison concerning overall tissue tropism and key effects of engineered mutations. The strain-specific genes contributed to reduced virulence and loci involved in tissue tropism, host-range restriction, immunomodulation are identified and discussed in the study. The reduced pathogenicity of GLV-1h68, while remaining replication competent in tumors, was further confirmed in more tumor models, and the contribution of inactivated F14.5L to the attenuation of GLV-1h68 was investigated by comparing the virulence of GLV-1h68 with its F14.5L- null and revertant viruses (Zhang et al. 2009).

Other organism: Tardigrade

Tardigrades are small metazoans resembling microscopic bears ("water-bears", 0.05 mm to 1.5 mm in size) and live in marine, freshwater and terrestrial environments, especially in lichens and mosses (Marcus and Dahl1928; Marcus 1928; Nelson 2002). They are a phylum of multi-cellular animals capable of reversible suspension of their metabolism and entering a state of cryptobiosis (Keilin 1959; Ramazzotti and Maucci 1983). A dehydrated tardigrade, known as anhydrobiotic tun-stage (Baumann 1922), can survive for years without water. Moreover, the tun is resistant to extreme pressures and temperatures (low/high), as well as radiation and vacuum (Horikawa et al. 2006; Hengherr et al. 2009a; Hengherr et al. 2009b; Jönsson et al. 2008; Jönsson and Schill 2007; Wright 2001).

Well known species include *Hypsibius dujardini* which is an obligatory parthenogenetic species (Ammermann 1967). The tardigrade *H. dujardini* can be cultured continuously for decades and can be cryopreserved. It has a compact genome, a little smaller than that of *Caenorhabditis elegans* or *Drosophila melanogaster*, and the rate of protein evolution in *H. dujardini* is similar to that of other metazoan taxa (Gabriel et al. 2007). *H. dujardini* has a short generation time, 13-14 days at room temperature. Embryos of *H. dujardini* have a stereotyped cleavage pattern with asymmetric cell divisions, nuclear migrations, and cell migrations occurring in reproducible patterns. Molecular data are sparse but include the purinergic receptor occuring in *H. dujardini* (Bavan et al. 2009). *Milnesium tardigradum* is an abundant and ubiquitous terrestrial tardigrade

species in Europe and possibly worldwide (Kinchin and Dannis 1994). It has unique anatomy and motion characteristics compared to other waterbears. Most water-bears prefer vegetarian food, *M. tardigradum* is more carnivorous, feeding on rotifers and nematodes. The animals are really tough and long-living, one of the reasons why *M. tardigradum* is one of the best-studied species so far.

Questions of general interest are: How related are tardigrade proteins to each other? Which protein families provide tardigrade-specific adaptations? Which regulatory elements influence the mRNA stability? To study them, we need a platform to warehouse the new sequenced ESTs as well as published entries, in addition, incorporated software allowing motif searches and functional clustering are urgently demanded.

Ultrafast DNA sequencing technology

Sanger sequencing method, which is a dideoxy chain termination method, has been predominant for more than 20 years. It was used to construct draft genome sequences of many species, i.e. *Home sapiens*, *Arabidopsis thaliana* and a lot of prokaryotic organisms. However, this traditional method suffers a length-limit of around 1000bps per read, this obstacle makes the analysis of complete genomes extremely time-consuming and expensive. With the advent of the new sequencing concept, the top position of the Sanger method has been challenged recently (Goldberg et al. 2006). A Solexa system developed by Illumina (Bennett 2004) and pyrosequencing system initiated by 454 Life science (Margulies et al. 2005) arose and became popular around the world. These high throughout methods can read as many as one billion base pairs per run (Solexa sequencing) and are remarkably faster and less expensive than conventional methods (Bennett et al. 2005).

These two systems are based on the principle of shotgun sequencing. They remove the need to clone the DNA fragments into appropriate vectors for amplification, and therefore avoid the introduction of errors that are derived from the base composition bias of the host genome (Hall

2007). With Solexa, genomic DNA is randomly fragmented using ultra-sonators, the resulting fragments and adapters are ligated to both ends of the fragments, afterwards they are bound to a solid surface, where amplification is performed to form DNA clusters. Labelled reversible terminators are added together with DNA polymerase to initiate the first cycle of the chemical synthesis. The incorporated base is identified by laser scanning, and then the blocked terminus is removed and the next synthesis cycle is commenced. This process is repeated until the entire base composition of the DNA fragments has been obtained. The 454 system involves a similar working principle as that of Solexa. The difference is that 454 technique lacks a protecting group for the termination reaction. Base incorporation is monitored by a chemiluminescence reaction that is caused by the pyrophosphate that is released during the sequencing process. The sequenced length is monitored by measuring the amount of released pyrophosphate. The 454 platform can read 400,000 fragments of more than 100 bases per run, whereas the Solexa system can measure 40 million readings of approximately 36 bases. Besides these two commercial high-throughput sequencing technologies, other methods are also under development, i.e., (1) polony sequencing (Shendure et al. 2005), which has rather lower error rate (below 1/1,000,000); (2) a microfabricated bioprocessor for integrated nanoliter-scale Sanger sequencing (Blazej et al. 2006); (3) the SOLiD sequencing system, which is based on a ligation reaction instead of the traditional polymerization reaction and is able to produce raw data of 8Gb (sequence) and valid data in 4Gb (Blow 2007); (4) a single-molecule sequencing technique, which excludes the need for PCR amplification (Harris et al. 2008).

These two high-throughput sequencing systems have their different pros and cons (Hutchison 2007). The Solexa system is better suited to investigate SNPs by measuring the occurrence frequency of certain sequences. In contrast, the 454 system can produce longer DNA fragments, which facilitates fragment assembly. However, we have to note the 454 system is relatively prone to generate errors in estimating the sequences due to lengths, thus is restricted in its reliability to detect the mutations, i.e., insertion or deletion of a single bp (Hutchison 2007), in particular when there are lower complex regions (sequencing of homopolymers) (Bentley 2006; Moore et al. 2006). For example, errors occurred in the detection of homopolymers when the

454 system was used to scan the genome of the Bacteroidetes species *Sulcia muelleri*. These errors were possibly to be avoided using the Solexa system (McCutcheon and Moran 2007). However, to have a precise assembly of shorter Solexa reads, it takes much more efforts than of 454 reads. There are a couple of groups tackling this problem using a reference genome. They perform efficiently when dealing with short reads (36bp). However, none of them can handle reads of variable lengths (from bp to kbs), which derive from multiple sources including solexa system, 454 and Sanger sequencing reads. Therefore, we develop an algorithm which prevails this challenge.

Motivation

General purpose: designing of these organisms regarding virulence and therapeutic potential agents or establishment of genome-scale metabolic network models is a challenge for current algorithms and published software. To enable this research using genome-wide analysis and modelling, a flexible platform and extensive analysis routines are required. These are the achieved centre pieces of my thesis. However to better understand the numerous biological applications of this effort, the following chapters explain the organisms investigated in this thesis: Listeriae (Hain et al. 2006, Liang and Dandekar 2006), Staphylococci (Schwarz and Liang et al. 2007), Tardigrades (Förster and Liang et al. 2009), vaccinia virus (Zhang and Liang et al. 2009)

Challenges: The high throughput sequencing techniques provide us abundant genomic data. To utilize them, software GENOVA (Liang et al. 2009) is developed to offer a platform to edit the sequence as well as the features. The convenient genome comparison software is limited, hence we implemented a platform-independent software, inGeno (Liang and Dandekar 2006). However, to solve the problems that some genomes of interest have not been completely sequenced, only some ESTs (Transcriptome) and contigs are available, in addition, it is more challenging because no any ideal genome reference is present, we developed an algorithm incorporated in the online software, JANE (Liang et al. 2009), to map and assemble them into a

virtual genome. The comprehensive network analysis software package, YANAsquare (Schwarz, Liang et al. 2007) is written to utilize this information to reconstruct the network (Metabolome), calculate the properties, flux distributions and estimate the robustness. In this project, We tackled the difficulties from the network calculation, in particular the combinatorial explosions during elementary mode analysis, the mapping puzzles when correlating proteomic data into *in silico* metabolic processes. To estimate the fluxes and calculate the other metabolites including intermediary metabolites, we measures external metabolites, 2D gel electrophoresis data (Proteome) to fit the network. The new convergence routine is able to solve the problems whereas the conventional methods lack the capability.

Materials & Methods

Hardware requirements

The major calculations were carried out on a computer with SuSE 10.2 Linux operating system, Intel Core2 6600 CPU (2.40GHz) and 3Gbyte DDR2 memory. A Paracel GeneMatcher2 system and a HPC cluster system having 160 CPUs and 426.3GB memory in total are applied to assist our research.

Databases used

These included all major public databases, starting from primary sequence information (NCBI, EMBL, DDBJ and SwissProt) as well as specific secondary databases such as Protecs (Bernhardt et al.), SMART, listilist, Poxvirus and various novel database mined and warehoused using Postgres locally (e.g. tardigrade workbench, vaccinia proteins). Enzyme databases were generally referred to, including KEGG, Puma, MetaCyc, BRENDA, ExPASy -Proteomic server and -Enzyme.

Transcriptome data of tardigrade

RNA extraction was performed using the QIAGEN RNeasy®Mini kit (Qiagen, Hilden, Germany). The cDNA synthesis was reversed transcribed using 1g total RNA by the Creator™ SMART™ cDNA Library Construction Kit (Clontech-Takara Bio Europe, France). The resulting cDNA was amplified following the manufacturers protocol and cloned into pDNR-Lib cloning vector. The resulting plasmids were used to transform Escherichia coli by electroporation. Sequencing of the cDNA-library was done by ABI 3730XL capillary sequencer (GATC Biotech AG, Konstanz, Germany). All obtained EST sequences were deposited with Genbank including dbEST databank. Nucleotide sequences from other tardigrades were collected from Genbank. For *H. dujardini*, the best represented species, we composed 5,235 ESTs. We stored *H. dujardini* as well as all

published sequences of other tardigrade species (e.g. *T. stephaniae*, *E. testudo*, *M. tardigradum*, *R. coronifer*) in a database (10,787 sequences including translated sequences).

Measuring external metabolites

All data for external metabolites were acquired from the supernatants of respective strains as described earlier (Liebeke et al. 2009). Briefly, cell suspension at mentioned time points of growth were steril filtered and stored at -20°C until 1H-NMR measurements. Measurement was done with a 600 MHz Bruker(R) Avance II spectrometer controlled by TOPSPIN 2.0 (Bruker(R) Biospin), qualitative and quantitative data were obtained with AMIX Version 3.2.

Measuring proteomic data

S. aureus COL (Shafer et al. 1979) was grown at 37°C in a synthetic medium (Gertz et al. 1999) with the following modifications: all amino acids were added with a final concentration of 1mM, glycine and MOPS were omitted from the medium. Samples were taken at mid-log phase (OD_{500nm} 0.5), at the onset of stationary phase (transient point), and 30, 60 and 120 min after transient point. For isolation of intracellular proteins 90 ml of the bacterial culture were harvested and the cell pellets were washed with and resolved in TE buffer. The cells were added to 500 µl glass beads and disrupted using the Precellys 24 homogenizator (PeqLab, Germany). Cell debris and glass beads were separated from the proteins by centrifugation for 10 min at 4°C at 15000 rpm followed by a second centrifugation step to remove insoluble and aggregated proteins (30 min, 4°C, 15000 rpm). The concentration of the protein extracts was determined using the Ninhydrin assay (Starcher, 2001). 80µg of total protein extract were loaded onto commercially available IPG strips with a pI range of 4-7 (GE Healthcare). 2D gel preparation was carried out as described earlier (Büttner et al. 2001). The resulting gels were stained with fluorescent dye Krypton according to the manufacturer's instructions (Thermo Scientific). Quantitative analyses of the gels were performed using the Delta2D software (Decodon GmbH, Germany).

Methods involved in online software

CLANS clustering

For a systematic overview on tardigrade specific adaptations we first clustered all published tardigrade nucleotide sequences into functional clusters using the Cluster analysis of sequences (CLANS) algorithm (Frickey and Lupas 2004). All sequences were clustered in 3D space using 0.001 as an E-value cut-off for TBLASTX all-against-all searches.

Identification of regulatory elements

The ESTs of *H. dujardini* and *M. tardigradum* were systematically screened using the software UTRscan. This software screens 30 regulatory elements for RNA regulation with a focus on 3' UTR elements and stability of mRNA. The default settings for batch mode were used and all reported elements were collected.

COGMaster: COG clustering and identification

In order to acquire a systematic overview of the functionalities, we used the latest version of COG/KOG databases (ftp://ftp.ncbi.nih.gov/pub/COG) and the BLAST hits from both, nucleotide search and protein search, were clustered according to their COG ID. Searches were carried out in parallel on all the tardigrade species including *M. tardigradum*, *H. dujardini*, *E. testudo*, *T. stephaniae* and *R. coronifer*. The results are summarized in a table shown in the tardigrade workbench, the background color from cold to warm (blue to red) indicates the cluster size, which enables an easy comparison. Moreover, users are able to click the COG ID and the hit number. The software then reports the corresponding sequence ID, description, conservation and the homologous entries recorded in the database. The server with its data is automatically updated weekly according to the latest tardigrade sequences.

Pattern searches

Pattern-search is an important method to identify the domains/motifs, particular moon-lighting catalytic motifs. In this research, pattern searches are carried out in both, the traditional way (regular expression search) and Versatile Search Processor Array (VeSPA) accelerated search. Tardigrade workbench and JANE software package provide user-friendly interfaces enabling the

conventional regular expression search, in addition, PROSITE expressions can be also accepted. They are readily manipulatable for most researchers. Moreover, we promote the capability of the server using VeSPA chip (unpublished), which is designed in a massively parallel systolic array architecture. This parallelism is two-dimensional: It simultaneously performs the comparison of all motif positions to a window with a length equal to the motif length. For each character position, it considers all alternative characters defined by the complex search motif simultaneously. The VeSPA board is currently located in EMBL Heidelberg, however, we develop a VeSPAoIP protocol as well as a flexible driver to enable this communication via a TCP/IP socket. This VeSPA acceleration benefits us not only parallel searches (4 queries per time), but also flexibility, the ambiguous-expression searches are supported in a hardware level, which can be completed in a same speed to explicit-expression searches. The second feature makes it a perfect candidate as a RNA element searcher besides a protein motif searcher. In our researches, we also carried out a benchmarking test between VeSPA search and NCBI-BLAST (sequence alignment) or regular-expression searches (motif searching) in Perl, the new routine demonstrates to be 2x-10x faster in speed than these conventional methods.

Methods involved in standalone software

In the following I give an overview on key methods used in the standalone software we developed. By their rapid performance they allow analysis of genome-scale data achieved by me during programming work for the thesis.

Genome analysis

Genomes are known to be dynamic and undergo various types of changes in their evolution: gene duplications result in paralogs, whereas gene deletions induce loss of functionality; recombination causes genome rearrangements and horizontal transfer introduces genetic materials into bacterial chromosomes, enabling the organism to recruit novel metabolic enzymes and consequently to survive in a different environment (Hacker and Carniel 2000). Early comparison methods to evaluate genome differences such as Needleman-Wunsch global alignment (Needleman and Wunsch 1970) and Smith-Waterman (SW) local alignment (Smith

and Waterman 1981) were designed to identify sequence differences on a small scale. The methods use dynamic programming algorithms and have been proven to be sensitive to find an optimal alignment between sequences. The increasing number of organisms whose genomes have been completely sequenced demands sensitive and precise methods for aligning long DNA sequences. Local alignments have been generally used to anchor global alignments. A variety of approaches such as MUMmer (Delcher et al. 1999), WABA (Kent and Zahler 2000), AVID (Bray et al. 2003), MAUVE (Darling et al. 2004) and ACT (Carver et al. 2005) have been developed for genome visualization. Several other programs have been developed for specialized purposes, for instance, Sim3 (Chao et al. 1997) uses a greedy algorithm to investigate highly similar input sequences and works well even for long sequences. The LAGAN series (Brudno et al. 2003) and MultiPipMaker (BLASTZ) are designed for dealing with genome rearrangements during the alignment process (Schwartz et al. 2003). The alignment comparison among these applications suggests the software derived from the Smith-Waterman algorithm prevails in revealing the genetic mutation events, though relatively slower, which is helpful to understand phenotype differences, metabolic diversities and boost the further experimental researches. Hence a program compatible with both NCBI-BLAST and Smith-Waterman alignment applications, in particular an accelerated version running on Paracel GeneMatcher 2 system, is highly required for the genome pairwise comparisons.

We implemented the program inGeno (Liang and Dandekar 2006), that offers a user-friendly platform, not only to parse the generated comparison report but also to analyze and visualize the sequence. It recognizes and illustrates the functional relationships between orthologous genes and strain-specific genome islands. It benefits from BioJava and BioJavaX toolkit (Mangalam 2002, Holland et al. 2008), thus the generic interface accepts all major standard sequence input formats (GenBank, EMBL, Fasta, GenBankXML, SwissProt). A dotplot analysis is performed to filter out the strain-specific genes of interest using a user-specified similarity threshold and to plot a comparison map with an interactive interface (according to user choices, e.g. zooming in/out; genome representation style). The modules for information retrieval aid the user: Annotation keywords, logical combinations and concatenations of these, genome

similarities and differences are identified from plain-text annotations then summarized and sorted by occurrences and functional categories (as color coded bars along the genome or as text reports).

Paracel's GeneMatcher is a high-throughput system for sequence similarity analysis that offers unmatched sensitivity and selectivity. Many types of hardware accelerated search algorithms are available through the GeneMatcher, e.g., Smith-Waterman, Profile and Hidden Markov Model searches. With the help of FDF (Fast data finder) system - systolic array designed for text string pattern matching, the hardware accelerated version allows genomic and protein databases to be searched hundreds to thousands of times faster than possible on general purpose computers and achieve the same level sensitivity and precision (Shpaeer et al. 1996). The framework located in University Würzburg is a Paracel GeneMatcher 2 system, which consists of a massively-parallel architecture in custom application-specific integrated circuit (ASIC) technology. 16 nodes with 32 CPUs performing a full range of dynamic programming algorithms are additionally involved in the system accelerating the comparison programs. It contains 31.5GB memory and 1722 GB disk storage for sequence databases, which fits perfectly for the requirement of modern bioinformatics research. The GeneMatcher2 system plays a crucial role in this project because it accelerates genome/proteome sequence alignments greatly, as well as the result format it generates is available for investigation using inGeno software.

InGeno is an efficient comparison application, to manage and investigate the features, in particular to simulate and curate the mutation events. The key idea of the algorithm is to locate syntenic islands from the linear correlation between the closely-related prokaryotic genomes, meanwhile the significant strain-specific genes can be identified. Another Java-based software Genome visualization and analysis (GENOVA), is developed for the purpose of managing, editing, visualizing, storing the genome sequences (Liang et al. 2009). On a feature level, all the genes are categorized by their function and plotted with different colours in a map. Support tools include a sequence format conversion module, specific feature extraction and a sequence

content analysis tool. Importantly, GENOVA has a number of helpful tools to aid in functional genomics: Any desired genomic feature is drawn to scale and compared not only graphically but different feature lists are compiled. On the sequence level, the genes can be studied by comprehensive operations. Genetic removal, modification or insertion of desired sequence fragments and features is rapidly simulated. In the system level, a specific feature column for each gene is used to map the gene to the corresponding process node for metabolic analysis.

Metabolite analysis

Metabolism is the chemical engine important for an organism's survival. Through the utilization of a vast repertoire of enzymatic reactions and transport processes, unicellular and multi-cellular organisms can process and convert thousands of organic compounds into the various bio-molecules which are necessary to support their living. All these reactions can form a rather complex metabolic pathway network, mainly including carbohydrate metabolism pathway, energy metabolism pathway, lipid metabolism pathway, amino acid metabolism pathway and nucleotide metabolism pathway. During a long period, it is a challenge to make a quantification study of metabolic pathways, in particular after network-based definitions of biochemical pathways emerged in recent years. The traditional approach is to have mathematic or computer mode, e.g., a dynamic simulator, which is based on fundamental physicochemical laws and principles. However, there is a great disparity between the power of available molecular biological techniques and the ability to rationally analyze biochemical networks. The explanation is due to the networks' evolution for over millions of years, during which the cell developed complex regulations via interconnected pathways, that definitely lead to inadequate dynamic models.

Figure 2: Brief history of the analysis methods for network-based pathways.

Currently with advent of the post-genomic era, there is a wealth of completely sequenced and annotated microbial genomes publicly available. Therefore, metabolic pathway reconstruction is becoming increasingly important for assessing the inherent network properties in biochemical networks predicted from genome analysis (Dandekar and Sauerborn 2002). The developmental history of network-based pathway analysis is summarized in the figure 2 (adapted from Papin et al. 2004). Of the two most promising definitions for pathway analysis, one relies on elementary flux modes (EM: Schuster et al. 2000) and the other on extreme pathways (EP: Schilling et al. 2000). These concepts are closely related because both methods return edges of the calculated cone as pathways. The elementary mode algorithm collects all possible non-decomposable routines through the whole network as flux modes, which do not

violate the steady intracellular metabolite rule. Extreme pathways are actually a subset of elementary modes and any elementary mode can be interpreted by one or combinations of extreme pathways. The major differences have been discussed in a review by Klamt and Stelling 2003. However, note that the redundant cycles and futile cycles can only be visualized using elementary mode analysis. Some cycles in a certain case may be essential for a system, regarding the energy consumption if the turnover rate is not omitted.

Elementary mode analysis (EMA)

The conceptual framework of these two modeling methods are identical, that can be represented by a stoichiometric matrix (N), the vector of reaction rates (v) and the vector of concentrations of internal metabolites, i.e., metabolites with variable concentrations (in contrast, concentrations of external metabolites are able to keep constant). Any set of biochemical transformations can be described mathematically by a system of ordinary differential equations (eq. 1):

$$\frac{d[X_i]}{dt} = \sum N_{i,j}\, v_j \qquad \text{(eq. 1)}$$

The rows of N correspond to the compounds (e.g. metabolites) in a reaction network. The columns of N correspond to the reactions in a network, with elements corresponding to stoichiometric coefficients of the associated reactions. At steady state, mass balance in a network can be simplified by:

$$N \cdot v = 0 \qquad \text{(eq. 2)}$$

The irreversibility constraints can always be written as a non-negative condition, since in case of negative flux we can reverse the orientation of reaction without loss of generality. By

decomposing the flux vector into two sub-vectors, $\boldsymbol{v}_{rev}$ and $\boldsymbol{v}_{irr}$, which include the fluxes of the reversible and irreversible reactions respectively, the condition can be expressed as:

$$\boldsymbol{v}_{irr} \geq 0 \qquad \text{(eq. 3)}$$

These flux vectors can be used to characterize enzymes which derive from the annotations of detected ORFs. According to the realizability whether these flux vectors obey stationary relationships, we can predict whether the enzymes form a functionally coherent set in metabolism. The generating vectors can be chosen so as to have a simplicity property in that as many components as possible are zero. This is of interest because the set of identified enzyme genes is incomplete in most cases, so that it is sensible to test whether at least one functional metabolic route is realized by this set. This leads to the concept of elementary mode.

The computation of elementary flux modes for biochemical reaction systems is relatively complex. The pioneering research applying a stoichiometric matrix on network stability analyses was carried out twenty years ago (Clarke et al. 1988). However, the two most prominent algorithms computing EMs are the method relying on canonical basis approach (CBA: Schuster et al. 2000), and a latter algorithm introducing a null-space approach (NSA: Wagner 2004). These two computational efforts have been summarized and compared in an article (Gagneur and Klamt 2004).

As programs the algorithm has already been implemented in Smalltalk (EMPATH, by John Woods in Oxford, available from ftp://bmshuxley.brookes.ac.uk/pub/mca/software/ibmpc), METATOOL in C, available from http://www.biozentrum.uni-wuerzburg.de/metatool.html (Pfeiffer et al. 1999) and MAPLE (METAFLUX, by Klaus Mauch at Stuttgart). The most popular one is METATOOL, since it can be compiled by GNU gcc, resulting in executable both on windows and linux/unix platforms. Moreover, the consuming time of the program is less than one second for the system considered as below on a general purpose computer. The program starts from a plain-text input file consisting of a list of reaction equations, a declaration of

reversible and irreversible reactions using enzyme names and of internal and external metabolites in abbreviation form. The output is a text file including the number of reactions and internal, external metabolite declarations, a generated stoichiometry matrix to describe the reactions, a null-space matrix which records all the steady fluxes, the enzyme subsets given both in the form of a matrix and as a list of enzyme names, including the information about whether these subsets correspond to a reversible or irreversible transformation, another stoichiometry matirix of reduced reaction system with the enzyme subsets taken as combined reactions, the overall stoichiometries of the enzyme subset referring to a group of enzymes that operate together in fixed flux proportions. In addition, the calculation will generate a convex basis both in the form of matrix and as a list of enzymes, including the information about reversibility of the basis vectors, the elementary modes given both in the form of a matrix and as a list of enzymes with the reversible properties, the overall stoichiometries of the EMs in terms of the external metabolites.

METATOOL has already been successfully applied in a couple of researches and been present as an analysis option in software packages such as GEPASI (Mendes 1993, 1997; Baigent 2001), FluxAnalyzer (Klamt et al. 2003) and SNA (Urbanczik 2006). However, the graphic user interface (GUI) is not convenient for assigning and modifying the model and involved parameters. The original command-driven program for sure requires adequate experiences to be manipulated properly. Therefore we implemented the software YANA and YANAsquare (Schwarz et al. 2005, 2008), which provides comprehensive operation possibilities, more importantly, the framework enables further investigations using the generated stoichiometry matrix and flux modes. The one new module is the robustness evaluation in YANAsquare. In stead of traditional robustness-test methods relying on the linear programming, this module calculates the product rates from quantification study using the elementary modes. Moreover, we offer a rapid reconstruction plan to accelerate researches from the initial setup step. A KEGG Browser was written with capabilities of selectively retrieving metabolites, enzymes and reactions directly from the KEGG online database (Kanehisa and Goto 2000) or a local integrated database. Any abundant metabolite such as H_2O, CO_2, or Phosphate can be optionally filtered by the program, whereas

KEGG identifiers of metabolites and enzymes of over-length or containing illegal characters for METATOOL can be automatically processed in order to ensure that the calculation module can recognize and deal with them. We used version 1.3 of the apache axis library, a Java implementation of the simple object access protocol (SOAP) and web services description Language (WSDL) on top of the HTTP protocol to connect to the database server. We developed a user-friendly graphical interface using the SWING framework to retrieve the information from KEGG. Moreover, a visualization module is responsible to render the metabolic network, a couple of topological field-tested layout algorithms, such as spring-embedded layout, radial tree and sugiyama, to automatically arrange the nodes and edges of the network in a user-friendly way. In our experiences, most of networks in a small size to a moderate size can achieve an ideal layout only after minor adjusting steps.

YANAsquare provides a powerful platform to have the network calculated in a few steps, however, pathway analysis of large of highly entangled networks still meets the problem of combinatorial explosion of possible routes across the networks. A couple of strategies have to be undertaken, i.e., classifying suitable external metabolites, eliminating futile cycles, removing the ambiguity of metabolites, in order to limit the network complexity within an acceptable range. Moreover, the metropolis algorithm can be applied to achieve a stoichastic optimization for larger metabolic networks, e.g., a genome-scale metabolism model (Dandekar et al. 2003).

Flux balance analysis (FBA)

Convex basis calculation leads to many flux modes comprising all possible situations of the system, however, how does the metabolic flux distribute through the whole system remains undetermined. Parameter *w* is the activity vector of each modes, ΔS counts the exchanges between metabolite nodes and the environment pool.

$$\mathrm{N} \cdot v \cdot w = \sum \Delta \mathrm{S}_i$$

The range of *w* can be restricted by *w*(0, 100) when a certain mode is irreversible and *w*ϵ(-100, 100) if reversible. When there are accumulations of extra-cellular metabolites, the concentration S_i of the equation right side can be quantitatively measured by NMR at different time point, thus the accumulations and consumptions of each extra-cellular metabolite can be calculated. We have prepared solution space N using YANAsquare (Schwarz et al. 2007) and thus *v* can be calculated using flux balance analysis. The vector (*w*) denotes the distribution of flux activities, which is going to be approached using YANAvergence routine.

Evolutionary algorithm and a large-scale bound-constrained optimization algorithm, L-BFGS-B have been applied to fit the network and detect the flux distribution. L-BFGS-B is a limited-memory quasi-Newton method for bound-constrained optimization, which fits for problems where the only constraints are of the form $l \leq x \leq u$. The original code is written in R and for most cases, it is able to reach the convergence much more rapidly than the genetic algorithm, if manageable.

When there is accumulation of extra-cellular metabolites, the concentration S of the right side of eq. 3 can be quantitatively measured by NMR at different time points. The accumulation and consumption of each metabolite can be calculated. N can be prepared using YANAsquare (Schwarz et al. 2007) and thus *v* can be calculated. The vector (*w*) denotes the distribution of flux activities, which is calculated using the YANAvergence routine.

Publications and Own contributions

Hereby we confirm major contribution of Chunguang Liang to the following publications:

DNA sequence analysis:
[1] **Liang C** and Dandekar T. **inGeno--an integrated genome and ortholog viewer for improved genome to genome comparisons**. *BMC Bioinformatics* 2006, 7:461.

[2] **Liang C**, Schmid A, Lopez-Sanchez MJ, Moya A, Gross R, Bernhardt J, Dandekar T. **JANE: Efficient mapping of prokaryotic ESTs and variable length sequence reads on related template genome**s. *BMC Bioinformatics* 2009, 10:391.

[3] **Liang C**, Wolz C, Herbert S, Bernhard J, Engelmann S, Hecker M, Götz F and Dandekar T. **GENOVA: A rapid genome visualization and functional genomics software applied to strain comparisons in *Staphylococcus aureus***. *Online Journal of Bioinformatics*, 2009, 10(2): 201-219.

Transcriptome:
[4] **Liang C***, Förster F*, Shkumatov A*, Beisser D, Engelmann JC, Schnölzer M, Frohme M, Müller T, Schill RO and Dandekar T. **Tardigrade workbench: Comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades**. *BMC Genomics* 2009. *shared first authors

Metabolom:
[5] **Liang C***, Schwarz R*, Kaleta C, Kühnel M, Hoffmann E, Kuznetsov S, Hecker M, Griffiths G, Schuster S and Dandekar T. **Integrated network reconstruction, visualization and analysis using YANAsquare**. *BMC Bioinformatics* 2007, 8:313. *shared first authors

Functional genomics:
[6] **Liang C***, Zhang Q*, Yu YA, Chen N, Dandekar T and Szalay AA. **The highly attenuated oncolytic recombinant vaccinia virus GLV-1h68: Comparative genomic features and the contribution of F14.5L inactivation**. *Molecular Genetics and Genomics* 2009, 282(4): 417-35.
*shared first authors

[7] Hain T, Steinweg C, Kuenne CT, Billion A, Ghai R, Chatterjee SS, Domann E, Kärst U, Goesmann A, Bekel T, Bartels D, Kaiser O, Meyer F, Pühler A, Weisshaar B, Wehland J, **Liang C**, Dandekar T, Lampidis R, Kreft J, Goebel W and Chakraborty T. **Whole-genome sequence of *Listeria welshimeri* reveals common steps in genome reduction with *Listeria innocua* as compared to *Listeria monocytogenes***. *J. Bacteriol*. 2006, 188(21):7405-15.

The following publications did not appear yet, but will so during the next months:

[8] **Liang C**, Liebeke M, Schwarz R, Zühlke D, Engelmann S, Sagnitz S, Bernhardt J, Hecker M, Lalk M and Dandekar T. **Input and output analysis of central *S.aureus* metabolism under different growth conditions**. (in preparation)

[9] Krüger B, **Liang C** and Dandekar T. **GoSynthetic: A synthetic biology work bench for cellular processes: Onthology, tests and data from different organisms and technical implications**. (in preparation)

[10] Epstein A, **Liang C**, Dandekar T, Vainshtein E, et al. **The Versatile Search Processor Array (VeSPA) – an Ultrafast Hardware for Bioinformatics**. (in preparation)

[11] Förster F*, Beisser D*, Grohme M*, Liang C, Mali B, Reuner A, Engelmann J, Shkumatov A, Schokarie E, Müller R, Blaxter M, Schnölzer M, Schill RO, Frohme M and Dandekar T. **Transcriptome analyzed at different levels in *Hypsibius dujardini* and *Milnesium tardigradum*: Specific adaptations, motifs and clusters as well as general protective pathways** (submission ready for NAR; *shared first authors)

Contribution details of the authors

[1] Liang C and Dandekar T. **inGeno-an integrated genome and ortholog viewer for improved genome to genome comparisons**. *BMC Bioinformatics*. 2006, 7:461.

Under instructions from my supervisor, the major work in particular regarding coding, debugging work and the comprehensive study of Listeriae were accomplished by me. I contributed approximately 90% of all this work in this publication.

- CL: programming and testing of the genome editor and examples, writing of the manuscript.
- TD: advice, organisation and guidance of the study, writing of the manuscript.

[5] Liang C*, Schwarz R*, Kaleta C, Kühnel M, Hoffmann E, Kuznetsov S, Hecker M, Griffiths G, Schuster S and Dandekar T. **Integrated network reconstruction, visualization and analysis using YANAsquare**. *BMC Bioinformatics*. 2007, 8:313.
*shared first authors

I was involved in the coding work, i.e., the method and interface of the KEGG browser were implemented and tested by me. The staphylococci primary network was reconstructed and analyzed also and I prepared the Figure 1 and the metabolic comparisons on the *S. epidermidis, S. aureus* and *S. saprophyticus* in a genome-scale. I contributed approximately 50% of the work in the study.

- RS: programmed and tested YANAsquare and its visualization routines and drafted the manuscript.
- CL: implemented, tested the KEGG browser module and prepared the staphylococci metabolism model, drafted the manuscript.
- CK: provided the Java implementation of the elementary mode analysis algorithm.
- MK, EH and SK: designed and carried out experiments on the phagosomal phospholipids and verified our *in-silico* predictions.
- MH: provided expert advice on *Staphylococci* and analysis of their metabolism.
- GG: guided the study of the phagosomal phospholipids and provided his expertise on phospholipid networks for this work.
- SS: developed algorithms for elementary mode analysis and aided CK in the efficient implementation of the algorithm in Java.
- TD: advised, organized and guided the present study and drafted the manuscript.

[6] Liang C*, Zhang Q*, Yu YA, Chen N, Dandekar T and Szalay AA. **The highly attenuated oncolytic recombinant vaccinia virus GLV-1h68: Comparative genomic features and the contribution of F14.5L inactivation**. *Molecular Genetics and Genomics* 2009, 282(4): 417-35. *shared first authors

The genome comparison, sequence/motif analysis and phylogenetic analysis were performed by me and the major experimental parts were completed by Qian Zhang. I prepared the figures 1, 2, 7 and the tables 1, 3 in the manuscript, as well as the comprehensive comparison tables in the supplementary material. Moreover, the final genome sequence was prepared and annotated by me. I contributed approximately 80% of the sequence analysis part of this study.

- CL: was involved in programming, analysis of motifs and sequences, preparation and annotation of the genome sequence.
- QZ, YAY and NC: conducted all experiments and functional tests.
- TD: guided the entire study, expertise on the bioinformatical research.
- AAS: organized and directed the entire study.

---

[2] **Liang C**, Schmid A, Lopez-Sanchez MJ, Moya A, Gross R, Bernhardt J and Dandekar T. **JANE: Efficient mapping of prokaryotic ESTs and variable length sequence reads on related template genome**s. *BMC Bioinformatics* 2009, 10:391.

The seed locating and alignment extending algorithm was initiated by me under the supervision of Prof. Thomas Dandekar. The implementation in the online software JANE was accomplished by me. Moreover, I carried out the benchmarking test among different mapping software for the article. I contributed approximately 90% of all this work.

- CL: designing the algorithm, programming and debugging of the JANE software, conceiving the manuscript.
- AS: testing of JANE, mapping of sequence reads, EST function analysis.
- MLS: genome sequencing, EST sequencing, analysis of ESTs and genome.
- AM: expert advice on genome projects, supervision of MLS.
- RG: expert advice on microbiological EST mapping in Blattabacteria and single cell sequencing.
- JB: expert advice on software development;
- TD: advice, organisation and guidance of the study, testing of JANE, writing of the manuscript

[3] **Liang C**, Wolz C, Herbert S, Bernhard J, Engelmann S, Hecker M, Götz F and Dandekar T. **GENOVA: A rapid genome visualization and functional genomics software applied to strain comparisons in *Staphylococcus aureus***. *Online Journal of Bioinformatics*, 2009, 10(2): 201-219.

I implemented the software GENOVA in Java and debugged the program. The software was applied to study Staphylococci and Vaccinia virus, the resulting Vaccinia GLV-1h68 genome were accepted by GenBank meanwhile the study was published in MGG (Zhang and Liang et al. 2009). Both tasks were accomplished by me. I contributed approximately 90% of all this work. I was also writing the manuscript together with TD.

- CL: established the software GENOVA and tested the current version of the program, using it studied the genomic features of Staphylococci and Vaccinia virus.
- CW, SH, FG: expertise in the staphylococcus study.
- JB: gave expert advice on software development.
- SE, MH: advice for the genomic/proteomic study.
- TD: organized and guided the entire study including analysis of data, supervision and manuscript writing.

[4] **Liang C***, Förster F*, Shkumatov A*, Beisser D, Engelmann JC, Schnölzer M, Frohme M, Müller T, Schill RO and Dandekar T. **Tardigrade workbench: Comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades**. *BMC Genomics* 2009.
*shared first authors

The initial workbench was established by me and the additional routines allowing genome comparisons, protein motif searches and RNA element searches were implemented by me. I warehoused Tardigrada nucleotide and protein sequences as different databases in the server. The adaptive “COGmaster” routine was particularly integrated in this workbench, having more functionalities than the original version published within JANE (Liang et al. 2009). The daily maintaining jobs were done by me. I contributed 90% in the coding work of the workbench.

- FF: did tardigrade protein data analysis including CLANS clustering and RNA motif analysis.
- CL: established the current version of the tardigrade workbench including programming new routines, data management and nucleotide motif analysis.
- AS: did the initial setup of the server, of the virtual ribosome and the CLANS clustering.
- DB, JE, MS and MF: participated in tardigrade data analysis.
- TM: gave expert advice and input on statistics.
- RS: gave expert advice on tardigrade physiology and zoology.
- TD: led and guided the study including analysis of data and program, supervision, and manuscript writing.

[7] Hain T, Steinweg C, Kuenne CT, Billion A, Ghai R, Chatterjee SS, Domann E, Kärst U, Goesmann A, Bekel T, Bartels D, Kaiser O, Meyer F, Pühler A, Weisshaar B, Wehland J, **Liang C**, Dandekar T, Lampidis R, Kreft J, Goebel W, Chakraborty T. **Whole-genome sequence of *Listeria welshimeri* reveals common steps in genome reduction with *Listeria innocua* as compared to *Listeria monocytogenes***. *J. Bacteriol*. 2006, 188(21):7405-15.

I was involved in the annotation work of *welshimeri* genome, in particular the genome comparison between pathogenic strains and non-pathogenic strains (*Listeria monocytogenes*, *Listeria innocua* and *Listeria welshimeri*). I identified the strain-specific genes and orthologs/paralogs, in the manuscript I prepared the data for Table 1. Moreover our group investigated its metabolism on a genome scale, the resulting Figure 7 (Hain T et al. 2006) was prepared by me. I contributed approximately 50% of these bioinformatical studies.

_____	_____
Thomas Dandekar	Ort, Datum
_____	_____
Chunguang Liang	Ort, Datum

General Discussion

Functional genomics and Visualization

Currently free visualization software is of limited availability, i.e., GeneVito (Vernikos et al. 2003) (linear genome view) and GenomeVis (Ghai et al. 2004) (circular map). Alternative programs include Artemis (Rutherford et al. 2000), a Java program developed at the EBI for annotation of bacterial genomes. GENOVA's main enhancement over this is the condensed display of the complete information of a genome and fast change of the information displayed (Liang et al. 2009). This includes inspection of the genome sequence, GFF format import and translation in several frames with our toolkit. The software Genome2D (Baerends et al. 2004) is also functionally rich and flexible, however, it is more complicated to handle. This list can of course be continued, for instance regarding pathways, there are different pathway tools available, i.e., BioCyc (Caspi et al. 2008), KEGG (Aoki and Kanehisa 2005) and there are different commercial software available for genome comparisons and drawing, e.g. vector NTI. However, GENOVA is free and easy to handle. Furthermore, it offers besides free coloring of any feature in a rich-annotation sequence file (including exact scale drawing, zooming in and out, format conversion and GC content plot) specific functional genomics support such as feature list generation and comparison of individual feature lengths, annotation capabilities and experimental design options (e.g. visualization and analysis of the same genome with or without phage integration). With its capabilities GENOVA allowed to investigate here different *S.aureus* strains and their biological design including pathogenicity features and genome features. Based on these analyses, *Staphylococcus aureus* COL (mild, lab strain) is suggested as a good physiological model from the multitude of pathogenic strains and RN1HG as an *rsbU* restored strain for investigating pathogenic *S.aureus* strains.

Genome Comparisons

With combination of inGeno software (Liang and Dandekar 2006), all comparisons were easy to handle. It identifies orthologs, strain-specific genes, visualizes and inspects genome comparison results in good quality. InGeno can be applied for genome comparisons between various strains, closer and less related species. Its graphical output reveals evolutionary changes, bacterial pathogenicity island and differences in metabolism. Furthermore, annotation search capabilities including logical concatenation of keywords, automatic comparison reports and lists offer further information to the user. By this, the genes of interests can be readily picked out from the genome, in particular the ones related to metabolism.

Modelling Bacteria

Metabolism involves two diverse aspects, one of them is the central metabolism including carbohydrate metabolism as an energy engine which supports fundamental survival and reproduction, the other is complete or restricted functionality of individual pathways, in particular in different pathogens, this determines important differences for *in vivo* or *in vitro* behaviour.

To understand the physiology of pathogenic bacteria, this project studies Staphylococci and Listeriae (Hain et al. 2006). We started from their complete or partially known genome sequences. The availability of complete sequences enabled us to determine the biological components that make up the cell, including important enzymes, regulators and transporters involved in their metabolisms, the cell wall and membranes produced by the specific metabolism. A partially sequenced genome can be completed using our alignment software JANE (Liang et al. 2009) to generate a virtual genome according to a related template of known sequence, to estimate the incomplete new genome. For higher level of cellular processes, software such as GENOVA (functional genomics) and YANAsquare (systems biology) were developed.

Limitations and Challenges: The regulations/functions in a transcriptional level, translational level and catalytic level lead to multi-dimensional annotations. We started from studying metabolism

in a genome-scale network. However, an efficient platform is required to reconstruct and analyze the network. In this thesis, the new discipline of systems biology was approached studying components inside the cell and how they interact to influence physiology. A typical application is the YANAsquare software (Schwarz et al. 2007), which we applied for flux balance analysis and used data from proteins and enzyme concentrations from quantitative proteomic data. Detailed data on staphylococci are going to be published soon. The flux changes in major metabolic pathways were considered in this work.

In recent years, a number of stoichiometric network analyzers have been presented. The most prominent include Metatool (Pfeiffer et al. 1999), FluxAnalyzer (Klamt et al. 2003), Jarnac, a module for the systems biology workbench (Sauro et al. 2003), GEPASI (Mendes 1993, 1997), ScrumPy (Poolman et al. 2004) and COPASI (Hoops et al. 2004). Additionally the recent work by Urbanczik et al. introduces SNA, a stoichiometric analysis package using Mathematica (Urbanczik 2006). These programs include sophisticated implementations of the algorithm for computation of the steady-state flux modes, or link to the efficient METATOOL implementation. Some, for example "FluxAnalyzer", even provide a basic graphical view of the network but may require, like the SNA, a valid MATLAB license to run. With YANAsquare we introduce four new components to the previously published software, namely (i) an integrated EMA algorithm (ii) a browser module for direct access to and download pathways from KEGG, (iii) visualization and layout algorithms for metabolic networks and (iv) a robustness analysis algorithm and apply them all here to different example networks. YANAsquare is a standalone open-source application and a 100% platform-independent modelling suite for steady state analyses of metabolic networks. It directly integrates a graphical visualization of the network using field-tested layout algorithms and powerfully combines this approach with the possibility to query the KEGG database directly from the application. The querying of pathway data is offered in two options, screening a local database or by performing remote calls to the KEGG server via the Internet. It includes a smart editor for reworking of the results as well as a database driven approach to abbreviate the often complicated and non-standardized compound names found in KEGG. Despite its straightforwardness it has to be stressed that importing pathways from KEGG

has to be done with some caution. Even though the resource is a valuable tool and certainly eases the construction of metabolic networks as shown in our examples, networks should always be double-checked manually. The data offered in KEGG is not always reliable and may contain pathway or stoichiometric inconsistencies on which YANA performs only some basic checks. Additionally thorough literature research is mandatory to find out about variant pathways in the organism under study as the KEGG pathways are quite generic. Nevertheless it should be noted that the KEGGBROWSER module was never intended to be a replacement for a well-crafted manually set up metabolic network but to ease and guide the process of network reconstruction by quickly providing the user a sound data basis to work on. Besides these points mentioned, the necessary manual reworking also includes the careful setting of the internal/external status of metabolites and the addition of any needed transport processes (see for example TransportDB (Ren et al. 2004, 2007) which are not part of KEGG at that time. Transport processes are easily incorporated into YANAsquare using reactions such as "Transporter1: substrateext → substrateint" to indicate that transporter1 transports the substrate from outside to inside. Compartimentalization can also be studied by YANAsquare. Different compartments are easily incorporated and defined as different subnetworks in the software. Each subnetwork contains only those enzymatic reactions actually taking place in this compartment. Transport reactions between compartments (e.g. mitchondrion ↔ cytoplasm) are defined analogous to above. The visualization routine allows easy rendering and drawing even of genome-scale networks, including standard editing possibilities like grid-alignment and zooming into interesting parts of the map. The force-directed layout was further improved to avoid overlaps in drawings and sketches. This is one key example, however, a number of other visualization routines were developed by me (JANE, GENOVA) and their limits and advantages are discussed in the attached publications.

Robustness and complexity: the implemented robustness routine in YANAsquare allows to rapidly investigate the effect of enzyme disruptions in terms of network stability (by drug effects or gene impairment). The question arises how significant the found differences are in robustness for the compared three genome-scale networks. A statistical test is difficult: The

resulting elementary mode matrix and the resulting identified differences are strictly deterministic. Instead one would like to investigate stochastic variation of the input stochiometric matrix. As the landscape for the obtained number of modes is very rugged, changes in the matrix require an extremely high number of trials to get sufficient observations for statistical significance. Instead we compare here biological meaningful optimized stochiometric input matrices (decision of compounds and their status as internal or external according to biochemistry). This does not leave much space for variation here (e.g. change of one metabolite from internal to external is in most cases not biochemically justified). Thus we believe that the found differences correspond to realistic differences in robustness. However, in an applied experimental study we would in any case recommend to complement the prediction data from the routine with some genetic data (e.g. how many knockouts are tolerated in one strain compared to others).

However, we still meet a prominent obstacle regarding the computational complexity. We can easily reconstruct and calculate primary pathways, however a complete genome-scale network calculation still suffers from the combinational explosion during the elementary mode (EM) calculation. All the reactions are described by stoichiometric relationships, the whole network can be represented in a concise matrix. However, when its dimension reaches over 300 reactions, in our experience the combination of elementary modes can not be solved. On the other hand, even if all the modes for these cases are considered, there is no method that can really determine all their parameters (flux activities).

We have tackled this problem by a couple of strategies. The first one is to manually simplify the stoichiometric complexity in addition to automatic methods, i.e., elimination of redundant/ futile cycles, switching extra/intra-metabolites, disconnecting certain pathways when regulations are present. In addition we applied different dissection strategy to minimize the combination. This strategy is generally efficient, however, we have to note this is a simple method, modes critical for specific conditions may be covered. Future work here remains to develop a novel computational method of higher capacity. Another effort is to rely on extreme

pathways (EP), which is a minimized subset of all the elementary modes. Only a number of modes just sufficient to describe the cone of permissible solutions are considered. By this the number of modes can be dramatically reduced to a manageable range. We applied this step when studying the influence from gene knockouts. The only drawback from the extreme pathway calculation is that a lot of pathways are split into a couple of EP modes, they are simply invisible, though present, in the result table. We successfully combined thus both strategies during the study of *Staphylococcus aureus* metabolic network, however, for a future comprehensive inspection on the physiology of pathogens, a high-capacity calculation tool is still required. One approach might be distributed computation, we are aware that sequence alignment calculations have benefited a lot from parallel computation in the last twenty years, thus currently most of bioinformatics department have at least one cluster system installed. However, the known applications of computing EM/EP are not ready for parallel computations on many processors/nodes (MPI, PVM). Of course the Sun Grid Engine (SGE) software can not help either, since the task can not be separated and prepared in a batch form. These programs used to be sufficient for studies on a moderate-sized network, but for whole metabolism studies on a genome-scale, algorithms are expected to be improved in future. Other groups in this field have tackled large networks and faced similar challenges trying independent work-around strategies, e.g., the Palsson group applied Monte Carlo sampling of the flux space then with principle component analysis (PCA) using singular value decomposition (SVD) to attain unified flux distributions, the resulting lower-dimensional solution space reduces the complexity to interpret flux modes (Barrett CL et al. 2009, Schellenberger and Palsson 2009), whereas the Schuster group introduces a concept of “elementary flux patterns” to circumvent the combination puzzle. A flux pattern is defined as a set of reactions in a certain subsystem in a large network, which considers all the possible flux routes that the entire system in particular a genome-scale network allows when reaching a steady state. The flux pattern is a “macro”-elementary mode, if it can not be derived from a combination of other flux patterns, this pattern is called elementary. For this effort, it relies on a prerequisite of human knowledge to distinguish subsystems of interest, however, it is a promising way to study flux distributions of

interest, though limited in subsystems, taking into account all the influences from the entire network (Kaleta C et al. 2009).

Applications

***S.aureus* metabolism:** Currently we focused on interactions within a local limited dimension. For our study, we collected extra-cellular metabolite data, proteomic data and have investigated them in two compartments, which are the periplasmic compartment and the cytoplasmic compartment. In future for a better understanding of *S. aureus* or pathogens in general physiology, there are still multiple extra-cellular compartments, including the interactions between the pathogen and host remaining to be investigated. Introducing compartments to an extensive range, the physiological changes caused by shifting a couple of fluxes can be studied and discussed.

We implemented programs that enable to study flux changes corresponding to different physiological states. However, it is just a first milestone of the complete effort to decipher the fluxome. The metabolite approach provides us a clear view on the flux distribution according to different time points and under different stresses. However quantitative data from proteomics tests remain to be discussed, one potential complication is the amount of enzyme (protein) is not exactly correlated to catalytic activity, there is further regulation present. Moreover if there is an impaired enzyme complex, e.g., one subunit absent or changed by sequence mutation. This non-correlated mode may lead to an early stop of the fitting procedure, we call premature: then a local vertex is reached during the calculation, not the global one. However, this effort provides us a highly sensitive method to detect enzyme regulation, even without the prerequisite of certain regulation mechanisms. The correlation of flux activities to the protein amount can be analyzed and visualized statistically to offer more interesting information for researchers to study them.

Tardigrade workbench: The situation where you have a variety of different genomic and proteomic data occurs in many genome projects. Besides studying prokaryotes, I participated in a project studying tardigrades. These are of interest for their remarkable adaptation capabilities against environmental stress.

Bioinformatically, the challenge was here to establish a flexible platform to analyze and process all available large-scale data. The tardigrade workbench features here organism specific BLASTs, translation and prediction of all encoded protein reading frames, specific motif searches in RNA and proteins as well as listings of different protein clusters (COG cluster). At the same time, we developed and made available software for further analysis in this direction (for promotor and motif searches, for protein motifs and for new organism specific data and databanks).

These type of services are partly available also at the NIH (for several tardigrade species organismic or group specific blast is possible), for *Hypsibius* a specific repository used to be available at www.tardigrade.org. However, this is no longer continued. Furthermore, the richness of different organism-specific searches offered at tardigrade workbench is unique. Regarding high speed sequence comparison we developed a specific hardware chip, the VeSPA search. In the proceeding version (unpublished), we are going to release a generic VeSPA driver written in Perl and a convenient online program to ease rapid motif/pattern/sequence searches against multiple databases, not only specific tardigrade and also for other organisms.

The detailed list of clusters of orthologous proteins opens for the first time a full-scale view on the different functional protein clusters in tardigrades. Besides stress-protection there are a number of tardigrade-specific sequence clusters (TSPs). Both are investigated using COG classification and CLAN clustering algorithms, the results can be visualized online and are automatically updated each Monday morning according to the latest database.

Other organisms: Genomics software (GENOVA) was also used for strain selection in *Staphylococcus aureus*, inGeno to identify and compare Listeria strains. Also they prove efficacy

in challenging on vaccinia virus studies, they are readily manageable, not highly sophisticated software, such as Artemis or other genome editors. These software are instrumental for the vaccinia comparison discussed below, e.g., inGeno and GENOVA used for LIVP *vs* GLV-1h68.

Applications for systems biology and synthetic biology

As an example, in a study in close cooperation with Genelux located in San Diego, we sequenced, analyzed and designed an oncolytic vaccinia strain, GLV-1h68. Vaccinia has good replication potential in tumor cells and is well removed in the healthy body parts, a key reason for the choice of this virus. The data suggest it as the promising candidate for cancer therapy since it has a property of tumor-specific replication, lower toxicity and appropriate oncolysis. We show the tissue tropism is contributed by ankyrine-like proteins and a couple of genes, e.g., the thymidine kinase (TK), the genes coding ankyrine-like protein are generally located in an active region between ITR and central conserved region. However the details including their structures and how they evolve remain unclear. Another critical approach is virulence attenuation, our research suggests the effect of one single gene deletion is possibly not significant, the pathogenicity can be caused by combination of a couple of genes. These essential genes are specific to different tissues, one gene is critical in one organ but may play only a minor role in the other organ. In addition, attenuations should be designed practically without compromising the sufficient replication rate. Using this information, it is possible to design a more optimized virus to improve the cancer therapy.

Our only competitor in this advanced therapy, Jennerex Inc., Pennsylvania, developed targeted oncolytic virus product for cancer therapy in a similar way. They investigated the oncolytic potential of another vaccinia strain JX-594, which is a Wyeth derivation strain carrying a deletion of the TK gene and expressing human granulocyte-macrophage colony-stimulating factor (hGM-CSF). Antitumoral effects were observed, including replication and systematic spread into distant tumors, after the virus was delivered via intratumoral injection to the patients (Park BH et al. 2008, Thorne SH et al. 2008, Kim JH et al. 2006). In comparison, we did a more comprehensive study on genes related on reducing the pathogenicity as well as

modulating host-adaptation, there are still interesting loci remained to be studied as a potential open door to further optimize vaccinia agents. All these differences detected are important for our new designed and improved vaccinia virus strains (e.g. GLV-1h68). With the advent of clinical trials (Phase I studies: Sloan Catering cancer center, Royal Marsden hospital London, several studies indented in Germany) and approvals, we believe it will lead to significant advances in cancer therapeutic techniques.

We show the optimized vaccinia virus as our best example for synthetic biology in an infection study, however, the platform we prepared is also optimal for systems biological research on other organisms as mentioned in the introduction part, i.e., *Staphylococcus aureus* COL, *Listeria monocytogenes* EGD-e and *Salmonella Typhimurim* LT2. The genome-scale models have been prepared in our lab and whole metabolic data were successfully tackled using this software, we present methods and applications, but all detailed results for these organisms have not been included in this cumulative dissertation. Some data are still remaining confidential or are unpublished yet.

Conclusion

We bridged from pure sequence analysis over sophisticated algorithm up to medical therapy. The main innovation in this thesis is treatment of genome-sized data by suitable algorithms and software. This includes functional genomics, metabolic modelling, EST mapping and engineering of complete viral genomes. These innovations were possible by a suite of new implementations in Perl and Java with Biojava and Bioperl extensions. In the near future, the latest methods, e.g., elementary flux patterns analysis (Kaleta et al. 2009), is going to be implemented in the next version of our network analysis software. We expect to be able to study more sophisticated genome-scale networks, more dimensions of the living cell (e.g. metabolism, protein complexes and channelling) and will consider more compartments.

Summary

Genome sequence analysis

A combination of genome analysis application has been established here during this project. This offers an efficient platform to interactively compare similar genome regions and reveal loci differences. The genes and operons can be rapidly analyzed and local collinear blocks (LCBs) categorized according to their function. The features of interests are parsed, recognized, and clustered into reports. Phylogenetic relationships can be readily examined such as the evolution of critical factors or a certain highly-conserved region. The resulting platform-independent software packages (GENOVA and inGeno), have been proven to be efficient and easy to handle in a number of projects. The capabilities of the software allowed the investigation of virulence factors, e.g., *rsbU*, strains' biological design, and in particular pathogenicity feature storage and management. We have successfully investigated the genomes of *Staphylococcus aureus* strains (COL, N315, 8325, RN1HG, Newman), *Listeria spp.* (*welshimeri*, *innocua* and *monocytogenes*), *E.coli* strains (O157:H7 and MG1655) and Vaccinia strains (WR, Copenhagen, Lister, LIVP, GLV-1h68 and parental strains).

Metabolic network analysis

Our YANAsquare package offers a workbench to rapidly establish the metabolic network of such as *Staphylococcous aureus* bacteria in genome-scale size as well as metabolic networks of interest such as the murine phagosome lipid signalling network. YANAsquare recruits reactions from online databases using an integrated KEGG browser. This reduces the efforts in building large metabolic networks. The involved calculation routines (METATOOL-derived wrapper or native Java implementation) readily obtain all possible flux modes (EM/EP) for metabolite fluxes within the network. Advanced layout algorithms visualize the topological structure of the

network. In addition, the generated structure can be dynamically modified in the graphic interface. The generated network as well as the manipulated layout can be validated and stored (XML file: scheme of SBML level-2). This format can be further parsed and analyzed by other systems biology software, such as CellDesigner. Moreover, the integrated robustness-evaluation routine is able to examine the synthesis rates affected by each single mutation throughout the whole network. We have successfully applied the method to simulate single and multiple gene knockouts, and the affected fluxes are comprehensively revealed.

Recently we applied the method to proteomic data and extra-cellular metabolite data of Staphylococci, the physiological changes regarding the flux distribution are studied. Calculations at different time points, including different conditions such as hypoxia or stress, show a good fit to experimental data. Moreover, using the proteomic data (enzyme amounts) calculated from 2D-Gel-EP experiments our study provides a way to compare the fluxome and the enzyme expression.

Oncolytic vaccinia virus (VACV)

We investigated the genetic differences between the *de novo* sequence of the recombinant oncolytic GLV-1h68 and other related VACVs, including function predictions for all found genome differences. Our phylogenetic analysis indicates that GLV-1h68 is closest to Lister strains but has lost several ORFs present in its parental LIVP strain, including genes encoding CrmE and a viral Golgi anti-apoptotic protein, v-GAAP. Functions of viral genes were either strain-specific, tissue-specific or host-specific comparing viral genes in the Lister, WR and COP strains. This helps to rationally design more optimized oncolytic virus strains to benefit cancer therapy in human patients. Identified differences from the comparison in open reading frames (ORFs) include genes for host-range selection, virulence and immune modulation proteins, e.g. ankyrin-like proteins, serine proteinase inhibitor SPI-2/CrmA, tumor necrosis factor (TNF) receptor homolog CrmC, semaphorin-like and interleukin-1 receptor homolog proteins. The contribution of foreign gene expression cassettes in the therapeutic and oncolytic virus GLV-

1h68 was studied, including the F14.5L, J2R and A56R loci. The contribution of F14.5L inactivation to the reduced virulence is demonstrated by comparing the virulence data of GLV-1h68 with its F14.5L-null and revertant viruses. The comparison suggests that insertion of a foreign gene expression cassette in a nonessential locus in the viral genome is a practical way to attenuate VACVs, especially if the nonessential locus itself contains a virulence gene. This reduces the virulence of the virus without compromising too much the replication competency of the virus, the key to its oncolytic activity. The reduced pathogenicity of GLV-1h68 was confirmed by our experimental collaboration partners in male mice bearing C6 rat glioma and in immunocompetent mice bearing B16-F10 murine melanoma. In conclusion, bioinformatics and experimental data show that GLV-1h68 is a promising engineered VACV variant for anticancer therapy with tumor-specific replication, reduced pathogenicity and benign tissue tropism.

Zusammenfassung

Genom Sequenz Analyse

Im Zuge der vorliegenden Doktorarbeit wurden verschiedene Programme zur Genomanalyse kombiniert, um eine effiziente Plattform zum interaktiven Vergleich lokaler Ähnlichkeiten bzw. Unterschiede in Genomen bereitzustellen. Damit können Gene und Operons schnell untersucht und "local collinear blocks" entsprechend ihrer Funktion kategorisiert werden. Phylogenetische Beziehungen, wie beispielsweise die Evolution spezifischer Elemente oder stark konservierter Regionen können leicht überprüft werden. Die hierfür entwickelte plattformunabhängige Software (GENOVA und inGeno) hat sich in mehreren Projekten als effizient und leicht handhabbar bewährt. Die Programme erlauben die Untersuchung von Virulenzfaktoren auf Sequenz- oder Annotationsebene. Während der vorliegenden Doktorarbeit konnten so die Genome von verschiedenen *Staphylococcus aureus*, *Listeria spp.*, *Escherichia coli* und *Vaccinia* Stämmen untersucht werden.

Metabolische Netzwerk Analyse

Unser "YANAsquare" Programmpaket bietet eine Oberfläche um schnell metabolische Netzwerke vom genomweiten Anzatz bis hinunter zum Einzelnetzwerk zu analysieren. Dafür greift YANA mit Hilfe des integrierten KEGG-Browsers auf Onlinedatenbanken zu, um die notwendigen Informationen zum metabolischen Reaktionsweg bereitzustellen und reduziert so maßgeblich den Arbeitsaufwand beim Beschreiben von Netzwerke. Die implementierten Methoden zur Berechnung (METATOOL, eigene Implementation in Java) des Netzwerkes liefern exakt alle die möglichen Elementarmoden (EM/EP) für die Metabolite zurück. Durch den Einsatz von fortgeschrittenen Layout Algorithmen wird anschliessend die Darstellung der Netzwerktopologie möglich. Außerdem kann in der grafischen Darstellung das generierte Netzwerklayout dynamisch verändert werden. Das Speichern der Daten erfolgt im XML (SBML level-2) Format und erlaubt so die Weiterverwendung in anderen systembiologischen

Programmen, wie dem “CellDesigner”. Mit Hilfe einer gen-Knockout Simulations Methode kann der Einfluss von einzelnen Mutationen im gesamten Netzwerk auf die Syntheseraten untersucht werden. Wir konnten mit dieser Methode Einzel- sowie Mehrfachgenknockouts und deren Effekte auf die Elementarmoden analysieren.

Die Methode wurde ebenfalls auf Proteomdaten und extrazelluläre Metabolite von Staphylokokken angewandt, um Änderungen bezüglich der Flussverteilung zu untersuchen. Die Simulationen zu verschieden Zeitpunkten und unter verschiedenen Stessbedingungen zeigen große Übereinstimmung mit experimentell erhobenen Daten.

Onkolytischer Vaccinia Virus (VACV)

Wir haben die genetischen Unterschiede zwischen der *de novo* Sequenz des rekombinanten onkolytischen Virus GLV-1h68 und anderen VACVs untersucht und gefundene Unterschiede funktionell charakterisiert. Die phylogenetische Analyse zeigt das GLV-1h68 mit dem Lister Stamm am nächsten verwandt ist. Auffällig ist dabei der Verlust von einigen open reading frames (ORFs), die noch im Eltern LIVP Stamm vorhanden sind (CrmE, v-GAAP). Beim Vergleich der Funktion viraler Gene aus Lister, WR und COP Stämmen treten stamm-, gewebe- und wirtsspezifische Gene auf. Diese Tatsache ermöglicht die Optimierung der onkolytischen Virusstämme für den Einsatz bei humanen Krebstherapien. Die beim Vergleich identifizierten Unterschiede zwischen den ORFs enthalten Gene für die Wirtsselektion, Virulenz und immunmodulierende Proteine (Ankyrin ähnliche Proteine, Serine-Proteinasen Inhibitor SPI-2/CrmA, Tumor Nekrose Faktor (TNF) Rezeptorhomolog CrmC, semaphorinähnliche und Interleukin-1 rezeptorhomologe Proteine). An den Loki F14.5L, J2R und A56R des GLV-1h68 Virus wurden die Vorteile der eingesetzten fremden Genexpressionskassetten untersucht. So zeigt GLV-1h68 mit F14.5L-Inaktivierung gegenüber der F14.5L-Revertanten Viren eine reduzierte Virulenz. Das erlaubt die Schlussfolgerung, dass die Insertion von fremden Genexpressionskassetten in nicht-essentielle Loki zur Verminderung der Virulenz von VACVs führt, besonders, wenn der nicht-essentielle Lokus selbst ein Virulenzgen enthält. Das Replikationsvermögen, welches ausschlaggebend für die onkolytische Aktivität des Virus ist,

wird trotz der verminderten Virulenz nicht eingeschränkt. Die reduzierte Pathogenität des GLV-1h68 Virus wurde durch experimentelle Daten unserer Kollaborationspartner in männlichen Mäusen mit Ratten C6 Gliom und in immunokompetenten Mäusen mit B16-F10 Mausmelanom nachgewiesen. Zusammenfassend zeigen experimentelle und bioinformatisch gewonnene Daten, dass GLV-1h68 eine vielversprechende VACV Variante für die Krebstherapie mit tumorspezifischer Replikation, verringerter Pathogenität und hoher Gewebsspezifität ist.

Bibliography

Al'tshtein AD, Zakharova LG, Loparev VN, Pashvykina GV, Gorodetski SI: Isolation of a recombinant vaccinia virus based on the LIVP strain inducing the surface antigen of the hepatitis B virus. *Dokl Akad Nauk SSSR*. 1985, 285(3):696-9.

Ammermann D: The cytology of parthenogenesis in the tardigrade *Hypsibius dujardini*. *Chromosoma* 1967, 23(2):203-213.

Aoki KF, Kanehisa M: Using the KEGG database resource in: Curr Protoc Bioinfo Unit 1.12, Wiley Publ. 2005.

Baerends RJ, Smits WK, de Jong A, Hamoen LW, Kok J, Kuipers OP: Genome2D: a visualization tool for the rapi d analysis of bacterial transcriptome data. *Genome Biology* 2004, 5(5):R37.

Baigent S:Software review: Gepasi 3.0. *Brief Bioinform* 2001, 2(3):300-2.

Barrett CL, Herrgard MJ, Palsson B: Decomposing complex reaction networks using random sampling, principal component analysis and basis rotation. *BMC Syst Biol.* 2009, 3:30.

Baumann H: Die Anabiose der Tardigraden. *Zool Jahrb* 1922, 45:501-556.

Baumann H: Bemerkungen zur Anabiose von Tardigraden. *Zool Anz* 1927, 72:175-179.

Bavan S, Straub VA, Blaxter ML, Ennion SJ: A P2X receptor from the tardigrade species *Hypsibius dujardini* with fast kinetics and sensitivity to zinc and copper. BMC Evol Biol 2009, 9:17.

Bennett S: Solexa Ltd. *Pharmacogenomics* 2004, 5: 433–438.

Bennett ST, Barnes C, Cox A, Davies L, Brown C: Toward the 1,000 dollars human genome. *Pharmacogenomics* 2005, 6: 373-382.

Bentley DR: Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* 2006, 16: 545-552.

Blazej RG, Kumaresan P and Mathies RA: Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proc. Natl. Acad. Sci. USA* 2006, 103: 7240-7245.

Blow N: Genomics: the personal side of genomics. *Nature* 2007, 449: 627–630.

Bray N, Dubchak I, Pachter L: AVID: A Global Alignment Program. *Genome Res* 2003, 13(1):97-102.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, ISC Comparative Sequencing Program, reen ED, Sidow A, Batzoglou S: LAGAN and multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* 2003, 13:721-31.

Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S: Glocal alignment: Finding rearrangements during alignment. *Bioinformatics* 2003, 19(suppl 1):i54-62.

Buller RM, Smith GL, Cremer K, Notkins AL, Moss B: Decreased virulence of recombinant vaccinia virus expression vectors is associated with a thymidine kinase-negative phenotype. *Nature* 1985, 317:813-815.

Buller RM, Chakrabarti S, Cooper JA, Twardzik DR, Moss B: Deletion of the vaccinia virus growth factor gene reduces virus virulence. *J Virol* 1988, 62:866-874.

Büttner K, Bernhardt J, Scharf C, Schmid R, Mäder U, Eymann C, Antelmann H, Völker A, Völker U, Hecker M: A comprehensive two-dimensional map of cytosolic proteins of *Bacillus subtilis*. *Electrophoresis* 2001, 22:2908-2935.

Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J: ACT: the Artemis Comparison Tool. *Bioinformatics* 2005, 21(16):3422-3.

Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD: The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of pathway/ genome databases. *Nucleic Acids Res*, 2008, 36(Database issue):D623-31.

Chao KM, Zhang J, Ostell J, Miller W: A tool for aligning very similar DNA sequences. *Comput Appl Biosci* 1997, 13(1):75-80.

Clarke BL: Stoichiometric network analysis. *Cell Biophys* 1988, 12:237-253.

Cordwell SJ, Larsen MR, Cole RT, Walsh BJ: Comparative proteomics of *Staphylococcus aureus* and the response of methicillin-resistant and methicillin-sensitive strains to Triton X-100. *Microbiology* 2002, 148:2765-2781.

Collins MD, Wallbanks S, Lane DJ, Shah J, Nietupski R, Smida J, Dorsch M, Stackebrandt E: Phylogenetic analysis of the genus Listeria based on reverse transcriptase sequencing of 16S rRNA. *Int J Syst Bacteriol* 1991, 41:240-246.

Cossart P, Portnoy D, in: Fischetti VA., Novick RP, Ferretti JJ, Portnoy DA, Rood JI(Eds.), Gram-Positive Pathogens, American Society for Microbiology, Washington D.C., 2001:507-515.

Dabiri GA, Sanger JM, Portnoy DA, Southwick FS: *Listeria monocytogenes* moves rapidly through the host-cell cytoplasm by inducing directional actin assembly. *PNAS* 1990, 87(16):6068-6072.

Dandekar T, Moldenhauer F, Bulik S, Bertram H, Schuster S: A method for classifying metabolites in topological pathway analyses based on minimization of pathway number. *Biosystems* 2003, 70(3):255-70.

Darling ACE, Mau B, Frederick RB, Perna NT: Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangements. *Genome Research* 2004, 14(7):1394-1403.

Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL: Alignment of whole genomes. *Nucleic Acids Res* 1999, 27(11):2369-2376.

Dunman PM, Mounts W, MacAleese F, Immermann F, Macapagal D, Marsilio E, MacDougal L, Tenover FC, Bradford PA, Petersen PJ, Projan SJ, Murphy E: Uses of *Staphylococcus aureus* GeneChips in genotyping and genetic composition analysis. *J Clin Microbiol* 2004, 42(9):4275-4283.

Farber JM, Peterkin PI: *Listeria monocytogenes*, a foodborne pathogen. *Microbiol Rev* 1991, 55:476-511.

Frickey T, Lupas A: CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* 2004, 20:3702-3704.

Förster F, Liang C, Shkumatov A, Beisser D, Engelmann JC, Schnölzer M, Frohme M, Müller T, Schill RO, Dandekar T: Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades. *BMC Genomics*. 2009, 10:469.

Fuchs S, Farre JP, Kohler C, Hecker M, Engelmann S: Anaerobic gene expression in *Staphylococcus aureus*. *J Bacteriol.* 2007, 189(11):4275-4289.

Gabriel WN, McNuff R, Patel SK, Gregory TR, Jeck WR, Jones CD, Goldstein B: The tardigrade Hypsibius dujardini, a new model for studying the evolution of development. *Dev Biol* 2007, 312:545-559.

Gagneur J, Klamt S: Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics* 2004, 5:175.

Gertz S, Engelmann S, Schmid R, Ohlsen K, Hacker J, Hecker M: Regulation of *sigmaB*-dependent transcription of *sigB* and *asp23* in two different *Staphylococcus aureus* strains. *Mol. Gen. Genet*. 1999, 261:558–566.

Ghai R, Hain T, Chakraborty T: GenomeViz: visualizing microbial genomes. *BMC Bioinfo.* 2004, 15(5):198.

Goebel SJ, Johnson GP, Perkus ME, Davis SW, Winslow JP, Paoletti E: The complete DNA sequence of vaccinia virus. *Virology* 1990, 179:247-266, 517-563.

Goldberg SMD, Johnson J, Busam D, Feldblyum T, Ferriera S, Friedman R, Halpern A, Khouri H, Kravitz SA and Lauro FM: A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci. USA* 2007, 103: 11240–11245.

Hacker J, Carniel E: Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep* 2000, 2(5):376-81.

Hain T, Steinweg C, Kuenne CT, Billion A, Ghai R, Chatterjee SS, Domann E, Kärst U, Goesmann A, Bekel T, Bartels D, Kaiser O, Meyer F, Pühler A, Weisshaar B, Wehland J, Liang C, Dandekar T, Lampidis R, Kreft J, Goebel W, Chakraborty T: Whole-genome sequence of *Listeria welshimeri* reveals common steps in genome reduction with *Listeria innocua* as compared to *Listeria monocytogenes*. *J. Bacteriol*. 2006, 188(21):7405-15.

Hall N: Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* 2007, 210: 1518–1525.

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H and Xie Z: Single-molecule DNA sequencing of a viral genome. *Science* 2008, 320: 106–109.

Hecker M, Englemann S, Cordwell SJ: Proteomics of *Staphylococcus aureus* - Current state and future challenges. *J Chromatorgr B.* 2003, 787(1):179-195.

Henderson DA, Moss B: Smallpox and vaccinia. *In*: Plotkin S, Orenstein W(eds) Vaccines. Saunders WB, Philadelphia, 1999, pp74-75.

Hengherr S, Worland MR, Reuner A, Brümmer F, Schill RO: Freeze tolerance, supercooling points and ice formation: comparative studies on the subzero temperature survival of limnoterrestrial tardigrades. *J Exp Biol* 2009a, 212:802-807.

Hengherr S, Worland MR, Reuner A, Brümmer F, Schill RO: High-Temperature Tolerance in Anhydrobiotic Tardigrades Is Limited by Glass Transition. Physiol Biochem Zool 2009b, 82(6):749-755.

Holland RCG, Down T, Pocock M, Prlic A, Huen D, James K, Foisy S, Draeger A, Yates A, Heuer M, Schreiber MJ: BioJava: an Open-Source Framework for Bioinformatics. *Bioinfo.* 2008, 24(18): 2096- 2097.

Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U: COPASI-a COmplex PAthway SImulator. *Bioinformatics* 2006, 22(24):3067-3074.

Horikawa DD, Sakashita T, Katagiri C, Watanabe M, Kikawada T, Nakahara Y, Hamada N, Wada S, Funayama T, Higashi S, Kobayashi Y, Okuda T, Kuwabara M: Radiation tolerance in the *tardigrade Milnesium* tardigradum. *Int J Radiat Biol* 2006, 82:843-848.

Hutchison CA: 3rd, DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.* 2007, 35: 6227–6237.

Jönsson KI, Rabbow E, Schill RO, Harms-Ringdahl M, Rettberg P: Tardigrades survive exposure to space in low Earth orbit. *Curr Biol* 2008, 18:R729-R731.

Jönsson KI, Schill RO: Induction of Hsp70 by desiccation, ionising radiation and heat-shock in the eutardigrade Richtersius coronifer. *Comp Biochem Physiol B Biochem Mol Biol* 2007, 146:456-460.

Kaleta C, de Figueiredo LF, Schuster S: Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res.* 2009 (in press).

Kanehisa M, Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000, 28:27-30.

Keilin D: The Leeuwenhoek Lecture: The problem of anabiosis or latent life: History and current concept. *Proc R Soc Lond B Biol Sci* 1959, 150:149-191.

Kent WJ, Zahler AM: Conservation, Regulation, Synteny, and Introns in a Large-scale C. briggsae-C. elegans Genomic Alignment. *Genome Res* 2000, 10(8):1115-25.

Kim JH, Oh JY, Park BH, Lee DE, Kim JS, Park HE, Roh MS, Je JE, Yoon JH, Thorne SH, Kirn D, Hwang TH: Systemic armed oncolytic and immunologic therapy for cancer with JX-594, a targeted poxvirus expressing GM-CSF. *Mol Ther.* 2006,14(3):361-70.

Kinchin I, Dennis R: The biology of tardigrades Portland Press London; 1994.

Klamt S, Stelling J: Two approaches for metabolic pathway analysis? *Trends Biochem Sci* 2003, 21:195-198.

Klamt S, Stelling J, Ginkel M, Gilles ED: FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics* 2003, 19(2):261-9.

Kuroda M, Kuroda H, Oshima T, Takeuchi F, Mori H, Hiramatsu K: Two-component system VraSR positively modulated the regulation of cell-wall biosynthesis pathway in *Staphylococcus aureus*. *Mol Microbiol* 2003, 49(3):807-821.

Liang C, Wolz C, Herbert S, Bernhard J, Engelmann S, Hecker M, Götz F, Dandekar T: GENOVA: A rapid genome visualization and functional genomics software applied to strain comparisons in *Staphylococcus aureus*. *Online Journal of Bioinformatics* 2009, 10(2): 201-219.

Liang C, Dandekar T: inGeno--an integrated genome and ortholog viewer for improved genome to genome comparisons. *BMC Bioinformatics* 2006,7:461.

Liang C, Schmid A, López-Sánchez MJ, Moya A, Gross R, Bernhardt J, Dandekar T: JANE: efficient mapping of prokaryotic ESTs and variable length sequence reads on related template genomes. *BMC Bioinformatics*. 2009, 10:391.

Liebeke M, Brozel VS, Hecker M and Lalk M: Chemical characterization of soil extract as growth media for the ecophysiological study of bacteria. *Appl Microbiol Biotechnol.* 2009, 83(1): 161-173.

Lowy FD: *Staphylococcus aureus* infections. *N Engl J Med.* 1998, 339:520-532.

Mangalam H: The Bio toolkits - a brief overview. *Brief Bioinformatics* 2002, 3(3):296-302.

Meier J, Lopez L: Listeriosis: an emerging food-borne disease, *Clin Lab Sci.* 2001, 14(3):187-92.

Mendes P: GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput Appl Biosci.* 1993, 9(5):563-571.

Mendes P: Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem Sci* 1997, 22(9):361-363.

Marcus E, Dahl F: Spinnentiere oder Arachnoidea IV. Bärtierchen (Tardigrada) Urban & Fischer Bei Elsevier; 1928.

Marcus E: Zur Ökologie und Physiologie der Tardigraden. *Zool Jahrb Abt Phys* 1928, 44:323-370.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF and Rothberg JM: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005, 437: 376–380.

McCutcheon JP, Moran NA: Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proc. Natl. Acad. Sci. USA* 2007, 104: 19392-19397.

Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM and Soltis DE: Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol* 2006, 6: 17.

Needleman S, Wunsch C: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* 1970, 48:443-453.

Nelson DR: Current Status of the Tardigrada: Evolution and Ecology. *Integr Comp Biol* 2002, 42:652-659.

Park BH, Hwang T, Liu TC, Sze DY, Kim JS, Kwon HC, Oh SY, Han SY, Yoon JH, Hong SH, Moon A, Speth K, Park C, Ahn YJ, Daneshmand M, Rhee BG, Pinedo HM, Bell JC, Kirn DH: Use of a targeted oncolytic poxvirus, JX-594, in patients with refractory primary or metastatic liver cancer: a phase I trial. *Lancet Oncol.* 2008, 9(6):533-42.

Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BO: Comparison of network-absed pathway analysis methods. *Trends in Biotech* 2004, 22(8):400-405.

Pfeiffer T, Sanchez-Valdenebro I, Nuno JC, Montero F, Schuster S: METATOOL: for studing metabolic networks. *Bioinformatics* 1999, 15:251-257.

Poolman MG, Venkatesh KV, Pidcock MK, Fell DA: A method for the determination of flux in elementary modes, and its application to *Lactobacillus rhamnosus*. *Biotechnol Bioeng* 2004, 88(5):601-612.

Ramazzotti G, Maucci W: The Phylum Tardigrada. Memorie dell'Istituto Italiano di Idrobiologia, Pallanza 1983, 41:309-314.

Ren Q, Chen K, Paulsen IT: TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res* 2007:D274-D279.

Ren Q, Kang KH, Paulsen IT: TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res* 2004:D284-D288.

Rocourt J, Schrettenbrunner A, Seeliger HP: Biochemical differentiation of the "*Listeria monocytogenes*" (sensu lato) genomic groups. *Ann Microbiol.* 1983, 134A(1):65-71.

Rocourt J, Seeliger HP: Distribution of species of the genus. Listeria. *Zentbl. Bakteriol. Mikrobiol. Hyg.* 1985, 259:317-330.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: Artemis: sequence visualization and annotation. *Bioinformatics*, 2000, 16(10):944-5.

Sallen B, Rajoharison A, Desvarenne S, Quinn F, Mabilat C: Comparative analysis of 16S and 23S rRNA Sequences of *Listeria* Species. *Int J Syst Bacteriol* 1996, 46(3):669-674.

Sauro HM, Hucka M, Finney A, Wellock C, Bolouri H, Doyle J, Kitano H: Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS* 2003, 7(4):355-372.

Schellenberger J, Palsson B: Use of randomized sampling for analysis of metabolic networks. *J Biol Chem.* 2009, 284(9):5457-61.

Schilling CH, Letscher D, Palsson BO: Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol* 2000, 203:229-248.

Schmid MW, Ng EY, Lampidis R, Emmerth M, Walcher M, Kreft J, Goebel W, Wagner M, Schleifer KH: Evolutionary history of the genus *Listeria* and its virulence genes. *Syst Appl Microbiol.*, 2005, 28(1):1-18.

Schuster S, Pfeiffer T, Moldenhauer F, Koch I, Dandekar T: Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae*. *Bioinformatics* 2002, 18(2):351-361.

Schuster S, Fell DA, Dandekar T: A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* 2000, 18(3):326-332.

Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, Smit A; NISC Comparative Sequencing Program, Green ED, Hardison RC, Miller W: MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res* 2003, 31:3518-3524.

Schwarz R, Musch P, von Kamp A, Engels B, Schirmer H, Schuster S, Dandekar T: YANA - a software tool for analyzing flux modes, gene-expression and enzyme activities. *BMC Bioinformatics* 2005, 6:135.

Schwarz R, Liang C, Kaleta C, Kühnel M, Hoffmann E, Kuznetsov S, Hecker M, Griffiths G, Schuster S, Dandekar T: Integrated network reconstruction, visualization and analysis using YANAsquare. *BMC Bioinformatics* 2007, 8:313.

Shafer WM, Iandolo JJ: Genetics of staphylococcal enterotoxin B in methicillin-resistant isolates of *Staphylococcus aureus*. *Infect. Immun.* 1979, 25:902–911.

Shen Y, Nemunaitis J: Fighting cancer with vaccinia virus: teaching new tricks to an old dog. *Mol Ther.* 2005, 11:180-195.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005, 309: 1728–1732.

Shida H, Hinuma Y, Hatanaka M, Morita M, Kidokoro M, Suzuki K, Maruyama T, Takahashi-Nishimaki F, Sugimoto M: Effects and virulences of recombinant vaccinia viruses derived from attenuated strains that express the human T-cell leukaemia virus type I envelope gene. *J Virol.* 1998, 62:4474-4480.

Shpaeer EG, Robinson M, Yee D, Candlin JD, Mines R, Hunkapiller T: Sensitivity and selectivity in protein similarity searches: A comparison of Smith-Waterman in Hardware to BLAST and FASTA, *Genomics* 1996, 38: 179-191.

Smith T, Waterman M: Identification of common molecular subsequences. *J Mol Biol* 1981, 147:195-197.

Starcher B: A ninhydrin-based assay to quantitate the total protein content of tissue samples. *Analytical Biochemistry* 2001, 292:125-129

Steinberg JP, Clark CC, Hackman BO: Nosocomial and community-acquired *Staphylococcus aureus* bacteremias from 1980 to 1983: impact of intravascular devices and methicillin resistance. *Clin Infect Dis.* 1996, 23(2):255-9.

Thorne SH: Oncolytic vaccinia virus: from bedside to benchtop and back. *Curr Opin Mol Ther.* 2008, 10(4):387-92.

Thorne SH, Bartlett DL, Kirn DH: The use of oncolytic vaccinia viruses in the treatment of cancer: a new role for an old ally? *Curr Gene Ther* 2005, 5:429-443.

Tilny LG, Portnoy DA: Actin filament and the growth, movement and spread of the intracellular bacterial parasite, *Listeria monocytogenes*, *J. Cell Biolo* 1989, 109:1597-1608.

Tilney LG, DeRosier DJ, Tilney MS: How *Listeria* exploits host cell Actin to form its own cytoskeleton.I. Formation of a tail and how that tail might be involved in movement. *J Cell Biol.* 1992, 118(1):71-81.

Upton C, Slack S, Hunter AL, Ehlers A, Roper RL: Pox virus orthologous clusters: toward defining the minimum essential pox virus genome. *J Virol.* 2003, 77:7590-7600.

Urbanczik R: SNA-a toolbox for the stoichiometric analysis of metabolic networks. *BMC Bioinformatics* 2006, 7:129.

Vazquez-Boland JA, Kuhn M, Berche P, Chakraborty T, Dominguez-Bernal G, Goebel W, Gonzalez-Zorn B, Wehland J, Kreft J: Listeria pathogenesis and molecular virulence determinants. *Clin Microbiol Rev* 2001, 14:548-640.

Verardi PH, Jones LA, Aziz FH, Ahmad S, Yilma TD: Vaccinia virus vectors with an inactivated gamma interferon receptor homolog gene (B8R) are attenuated in vivo without a concomitant reduction in immunogenicity. *J Virol* 2001,75: 11-18.

Vernikos GS, Gkogkas CG, Promponas VJ, Hamodrakas SJ: GeneViTo: visualizing gene product functional and structural features in genomic datasets. *BMC Bioinformatics*, 2003, 31(4):53.

Wagner C: Null space Approach to Determine the Elementary Modes of Chemical Reaction Systems. *J Phys Chem B.* 2004, 108:2425-2431.

Walsh TR, Howe RA: The prevalence and mechanisms of vancomycin resistance in *Staphylococcus aureus*. *Annu Rev Microbiol.* 2002, 56:657-675.

Welshimer HJ: Isolation of *Listeria monocytogenes* from vegetation. *J. Bacteriol.* 1968, 95:300-303.

Wright JC: Cryptobiosis 300 Years on from van Leuwenhoek: What Have We Learned about Tardigrades? Zoologischer Anzeiger - A Journal of Comparative Zoology 2001, 240:563-582.

Zhang Q, Liang C, Yu YA, Chen N, Thomas Dandekar, Aladar AS: The highly attenuated oncolytic recombinant vaccinia virus GLV-1h68: Comparative genomic features and the contribution of F14.5L inactivation. *Molecular Genetics and Genomics* 2009, 282(4):417-35.

Nomenclatures

ASIC	application-specific integrated circuit
COL	*Staphylococcus aureus* COL strain
COP	vaccinia virus strain Copenhagen (GenBank: M35027)
EMA	elementary mode analysis
EP	extreme pathway
EM	elementary mode
FBA	flux balance analysis
FDF	fast data finder
GLV-1h68	vaccinia virus strain GLV-1h68 (GenBank: EU410304)
HA	hemagglutinin
L-BFGS-B	a version of limited memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm
LCB	local collinear blocks
Lister	vaccinia Lister major strain (GenBank: AY678276)
Newman	*Staphylococcus aureus* Newman strain
N315	*Staphylococcus aureus* N315 strain
RN1HG	*Staphylococcus aureus* RN1HG strain
rRNA	ribosome ribonucleic acid
OPV	orthopox virus
SBML	systems biology markup language
SNP	single nucleotide polymorphism
SPI	serine proteinase inhibitor
SW	smith-waterman algorithm
SOAP	simple object access protocol
TK	thymidine kinase
TNF	tumor necrosis factor
VACV	vaccinia virus
VeSPA	versatile search processor array
WR	vaccinia virus strain Western Reserve (GenBank: AY243312)
WSDL	services description language
2D-Gel-EP	two-dimensional gel electrophoresis

Appendix of original publications

Acknowledgements

This work was undertaken in the Department Bioinformatics, University of Würzburg, with cooperation with University of Tübingen, University of Greifswald, University of Giessen and Genelux Corporation under the guidance of Prof. Dr. Thomas Dandekar.

I wish to thank my supervisor and best friend Thomas for his detailed instruction, kind encouragement and persistent support throughout all my researching work. He as a smart and kind scientist, presented me an excellent platform for academic training, a friendly environment to work efficiently.

I am grateful for my colleagues, without their specific assistance, the present version of the thesis would not have been possible. Special thanks to Frank Förster and Felix Bemm for translating the summary into German, Daniela Beisser and Santosh Nilla for the English proof-reading. The most important acknowledgement is to my wife Jinjin, who has accepted and tolerated many hours of my absence during the research in weekends and the conceiving of manuscripts. Special thanks for the love, care and support.

The financial support from DFG Deutsche Forschungsgemeinschaft (DFG) TR34/A5 and Bundesministerium für Bildung und Forschung (BMBF) Pathogenomik and Land Bavaria (TD) are highly acknowledged.

Particularly dedicated to my grandfather, Prof. Dr. Shangzhi Liang who actively celebrated his 100th birthday on the 20.04.2009.