Genlight: Interactive high-throughput sequence analysis and comparative genomics

Michael Beckstette[1*], Jens T. Mailänder[2*¤], Richard J. Marhöfer[2], Alexander Sczyrba[1], Enno Ohlebusch[1#], Robert Giegerich[1], Paul M. Selzer[2‡]

[1] Technische Fakultät Universität Bielefeld, Postfach 100 131, D-33501 Bielefeld, Germany, {mbeckste, asczyrba, robert}@techfak.uni-bielefeld.de

[2] Akzo Nobel, Intervet Innovation GmbH, BioChemInformatics, Zur Propstei, D-55270 Schwabenheim, Germany, {richard.marhoefer, paul.selzer}@intervet.com

Abstract

With rising numbers of fully sequenced genomes the importance of comparative genomics is constantly increasing. Although several software systems for genome comparison analyses do exist, their functionality and flexibility is still limited, compared to the manifold possible applications. Therefore, we developed *Genlight*, a Client/Server based program suite for large scale sequence analysis and comparative genomics. *Genlight* uses the object relational database system PostgreSQL together with a state of the art data representation and a distributed execution approach for large scale analysis tasks. The system includes a wide variety of comparison and sequence manipulation methods and supports the management of nucleotide sequences as well as protein sequences. The comparison methods are complemented by a large variety of visualization methods for the assessment of the generated results. In order to demonstrate the suitability of the system for the treatment of biological questions, *Genlight* was used to identify potential drug and vaccine targets of the pathogen *Helicobacter pylori*.

Availability: The *Genlight* system is publicly available for non-commercial use on http://piranha.techfak.uni-bielefeld.de. To retrieve a personal user account, please use the contact address mentioned on this page.

---

[*] Both authors contributed equally to this work

[¤] Present address: Lindenhofstraße 70, 68163 Mannheim, Germany, jens_mailaender@web.de

[#] Present address: Fakultät für Informatik, Universität Ulm, Oberer Eselsberg, D-89081 Ulm, Germany, eo@informatik.uni-ulm.de

[‡] Corresponding author

Introduction

A major key for the discovery of molecular mechanisms necessary for the machinery of an organism is the field of comparative genomics [1,2]. Moreover, genome comparison is an excellent method for target finding in the drug discovery process [3]. By April 2004 142 bacterial, 18 archaeal, and 26 eukaryal genomes were completed and not less than 490 prokaryotic and 415 eukaryotic genome sequencing projects are underway (Genomes OnLine Database) [4].

With increasing numbers of complete genome sequences, tasks are shifting from single gene to complete genome or proteome analyses, and many new questions regarding similarities and differences between the sequenced organisms arise in multiple genome comparison approaches. One of these new, challenging questions is the differentiation between species specific and common genes [5,6]. This is one of the fundamental questions in the target-based approach, for example in the development of either narrow-spectrum or broad-spectrum antibiotics. Differential comparative genomics, especially in combination with motif analyses, has proven to be a powerful approach for the screening of new drug targets [7]. However, the number and flexibility of systems in the field, suited for this approach, is limited. Although the integration of various bioinformatics methods and automated sequence homology searches are widely used techniques in genome annotation systems, such as Magpie [8,9], PEDANT [10], and GenDB [11], only three systems exist, that support automated differential genome analyses: FindTarget [12], Difftool [13], and Seebugs [14]. All three systems are very limited in the available comparison methods.

In this article we present *Genlight*, a versatile and powerful software system to address a wide spectrum of tasks in genome scale sequence analysis, with a special focus on differential comparative genome analysis. The system is designed for (i) the discovery of potential new drug targets by comparative genome analysis, (ii) automatic genomic scale analyses in reasonable time, without the need for specialized hardware or large and expensive cluster systems, (iii) the integration of various bioinformatics analysis methods, whose results are stored in a structured, reusable, and queryable way, and (iv) dynamic result presentation and visualization through an easy to use, but still flexible interface. The *Genlight* system is multi-user capable, suited for high-throughput analysis of biomolecular data, and connects the advantages of an object relational database management system with a distributed client/server approach for large scale compute tasks and a powerful web interface. Besides the concepts of *Genlight* and its architecture, we present an exemplary scientific application of the *Genlight* system.

System architecture and functionality

System architecture

The *Genlight* system consists of four major parts as shown in Figure 1: (i) a web-based user interface for the communication with the system, (ii) the *Genlight* server, (iii) client components to carry out various bioinformatics analysis tasks in an asynchronous way, and (iv) a database component for storing, modifying, and accessing data. The underlying relational database system allows easy access to the generated data, even from external applications using standard SQL queries.

The structured storage and reusability of generated results is a critical point for the protocol based step by step modeling of more complex experiments/workflows [15]. In *Genlight* the re-use of derived results is a central concept. It is achieved with a set oriented data model with only two basic data structures: seq-sets and hit-sets. A seq-set is a collection of sequences of one type, either nucleic acid or protein. A hit-set is a set of sequence pairs, defined by a comparison operation between two seq-sets and its parameterization, e.g. the set of all sequence pairs detected by a homology search between two seq-sets.

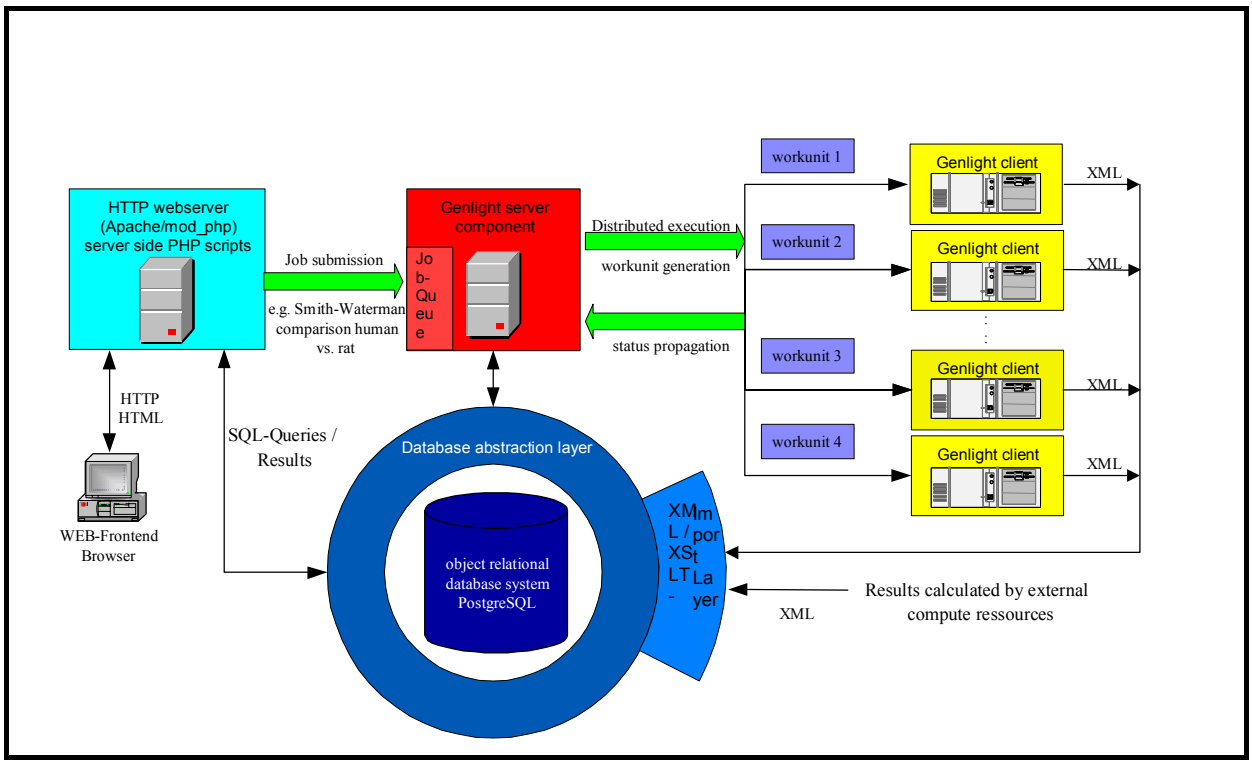

Figure 1: A schematic view of the Genlight system architecture.

Genlight supports various operations that can be applied to hit-sets or seq-sets, where each operation results in a new seq-set or hit-set. A hit-set filter, which can be pre-defined or user defined, generates a new hit-set with sequence pairs satisfying the respective filter condition. Filters can be based on Boolean combinations of arbitrary attribute values stored in the hit-set e.g. method specific alignment scores or significance values. Sequence filters generate new seq-sets and extraction operations convert a hit-set to a new seq-set depending on specified criteria (see Table 1). This procedure follows the software engineering concept of *compositionality* and allows an interactive step by step modeling of complex workflows as schematically drafted in Figure 2. With *Genlight* it is therefore possible to compare two, three or even more genomes to each other. Using a combination of filters and extraction operations three proteomes, for instance proteomes A, B, and C, can be easily screened for proteins common to the proteomes A and B but nonexistent in proteome C. Moreover, all possible intersections of A, B, and C can be calculated. Evidence of proteins with similar function can be defined by different homology search results (e. g. unidirectional best hits or bidirectional best hits), even generated by different homology search methods, like FASTA, BLAST or Smith-Waterman. Further on, the results of different sequence comparison methods can be combined with Boolean operators. With this concept the results of different alignment methods can be taken into account as evidence factors for the detection of homologous genes and weaknesses in the heuristics of one single method, which result in a false negative detection of homologous sequences, can be balanced.

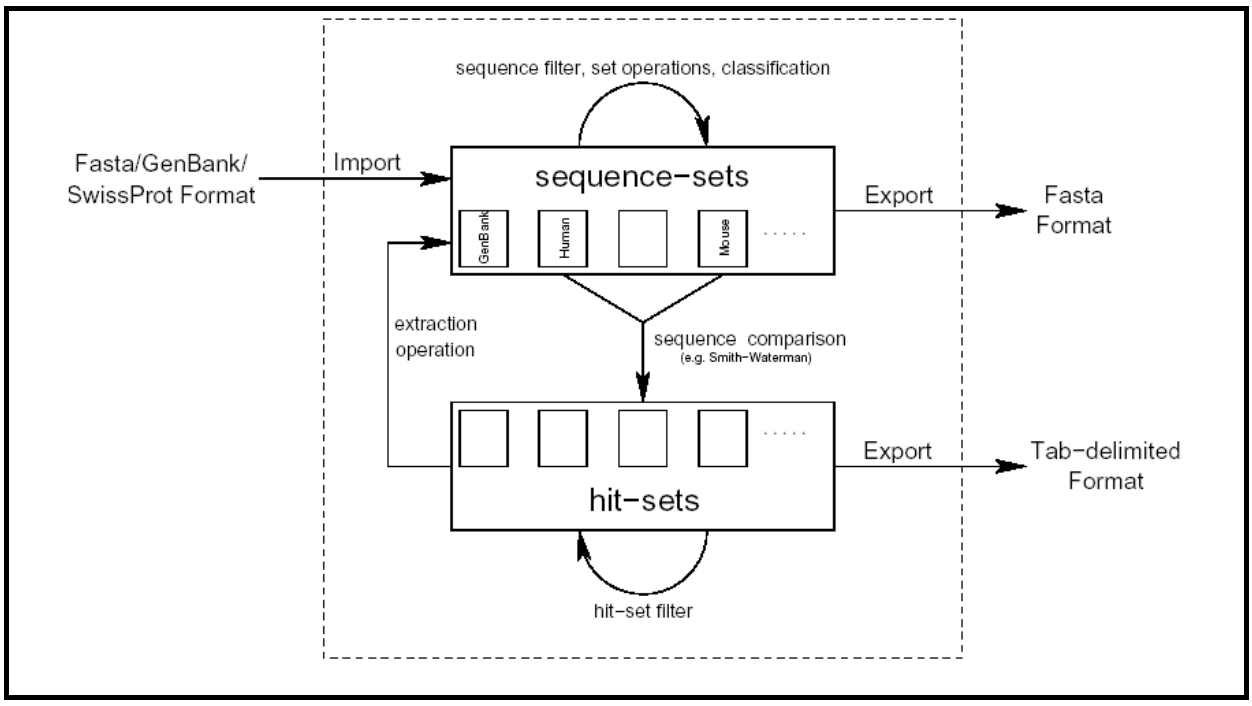

Figure 2: The set-oriented concept: Basic data structures and their compositionality

A project management, providing fundamental access control features, allows to store seq-sets and hit-sets on a per-user basis. Frequently used seq-sets and hit-sets, like the GenBank database, the SpTrembl database, model organism comparisons, etc., can be made available system-wide. The administrative features are complemented by a quota system, which allows to assign resources on a per-user and per-method basis. It is therefore possible to restrict the number of seq-sets and hit-sets in a project or to limit the size of a seq-set in a comparison operation.

Table 1: An excerpt of available operations on seq-sets and hit-sets

Category	Operation	Result
seq-set filters	filter by domain/motif occurrence or composition	All sequences with a specified motif or combination of motifs detected by screenings vs. the integrated motif databases (e.g. .Pfam, TIGRfam, SMART or CDD). Evidence for motif occurrence can be defined by user/method specified cutoffs (E-value, Alignment Score, percent-identity, etc.).
	SCOP filter	All sequences with user defined similarity to the structural classification of proteins (SCOP) classification hierarchy. Selection of sequences can be based on similarity to class, fold, superfamily, or family.
	taxonomy based filter	All sequences that belong to a given taxon (if taxonomy information is available).
	filter by length	All sequences that satisfy a sequence length constraint.
	homology based filtering	All sequences that have at least one homolog / no homolog in one or more user defined sequence sets. Evidence for homology can be determined by different sequence comparison methods.
hit-set filters	filter by attribute values	All pairs satisfying the filter conditions. Stored attributes of a hit set entry are the alignment score, the bit score, the E-value, percent-identity, percent-positive, coverage rate, etc.) Filter condition can be a Boolean expression combining different attribute values.
	best hit filter	Selects the best hit from a hit-set depending on method specific rankings.
	bidirectional best hit filter	Selects bidirectional best (two way best) hit pairs depending on method specific rankings.
	text pattern filter	Selects all pairs that contain a given pattern (exact or regular expression) in the query/hit annotations.
	full length filters	Select all pairs with an aligned region length equal to the length of the query or hit sequence.
extraction operations (convert a hit-set to a sequence-set)	extract sequences with homologs	Generates a new sequence-set containing sequences that have a homolog in one or multiple hit-sets.
	extract sequences with NO homologs	Generates a new sequence-set containing sequences that have no homolog in one or multiple hit-sets.

The interactive character of *Genlight* and its need to store calculated data on demand requires more complexity in the implementation of the data model than it is the case in systems with pre-calculated, static data. *Genlight* uses the ORDBMS PostgreSQL for data storage and access, and makes use of its object oriented features. Seq-sets and hit-sets are stored in database tables, reflecting the method-specific attributes. Seq-sets or hit-sets, generated by import or through the application of one of the operations described above, generate a child table by inheritance from the method specific template table.

This newly generated table, which can be seen as an instance of the templates, is then unambiguously referenced by a catalog table entry that stores additional parameters (e.g. generating method, parameterizations of the method, etc.).

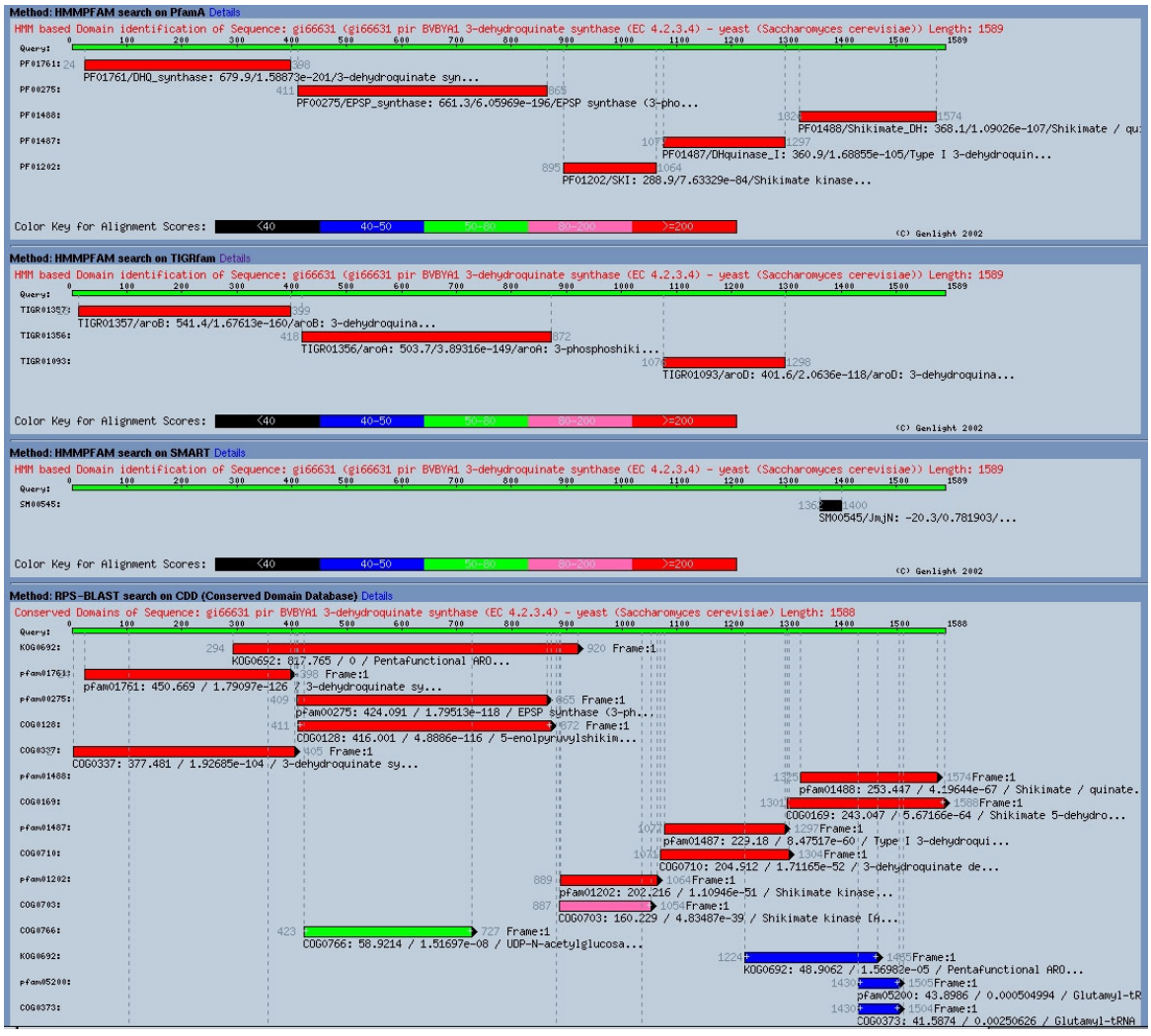

Figure 3: The web based User Interface: Visualization of the results of conserved motif screenings in different databases for the 3-dehydroquinate synthase of *S. cerevisiae*, a typical multi domain protein.

The core of *Genlight*, i.e. the server and client components, is written in the C programming language and accesses the PostgreSQL ORDBMS through a database abstraction layer. This allows an easy adaptation to other database management systems. The web based user interface is written in the server sided scripting language PHP and makes use of the GD graphics library for dynamic data visualization. Calculated results are presented in graphical as well as in textual/tabular form (see Figure 3 and 4). *Genlight* was developed and intensively tested on the Solaris operating system (Sun Inc.) as well as on Irix (SGI Inc.) and Linux. It should be easily portable to other UNIX systems.

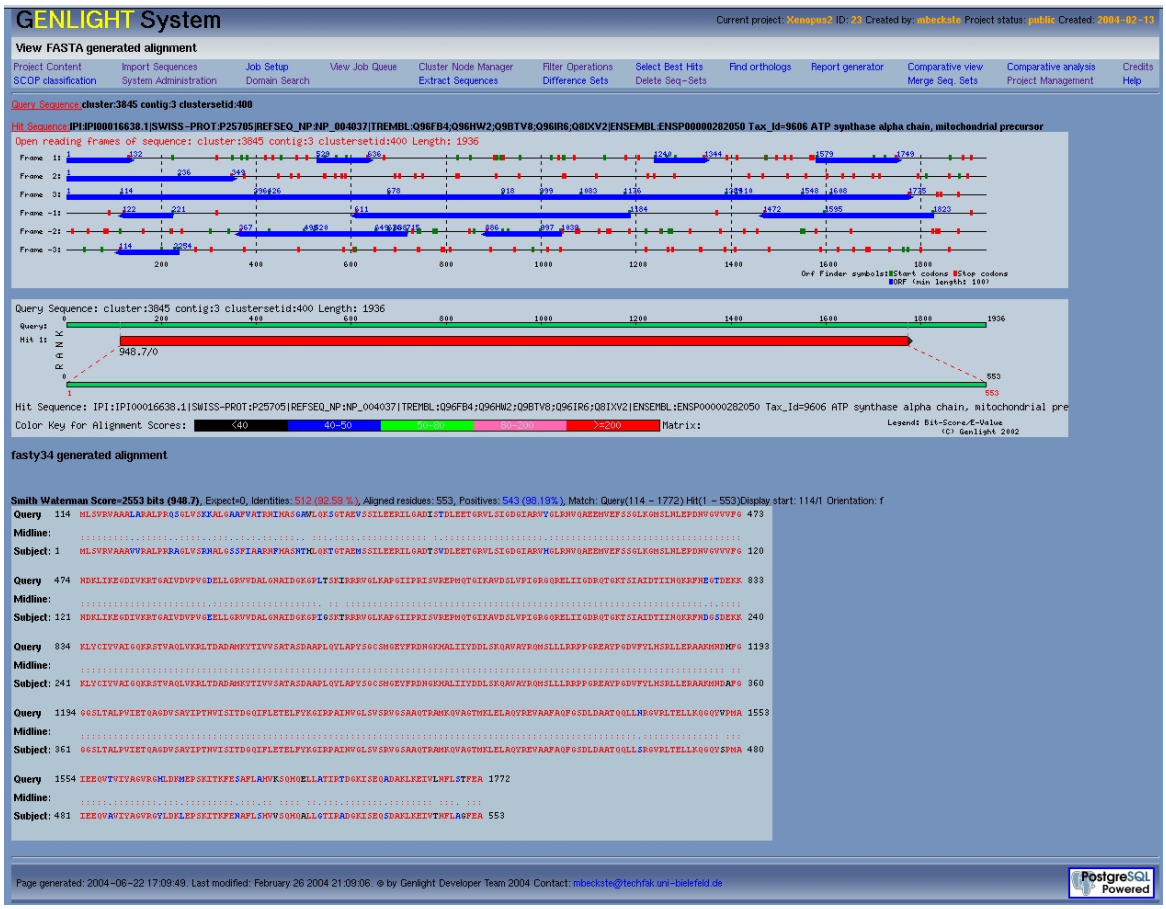

Figure 4: Screenshot showing graphical and colored textual representation of a FASTY alignment. Open reading frames (top, blue boxes) and the matching region in query and target sequences (bottom) are presented graphically. The corresponding textual alignment is shown at the bottom of the page.

Integrated sequence analysis methods and databases

Genlight can handle nucleotide and protein sequences. Almost all algorithms of the BLAST and FASTA family [16-19] as well as the traditional Smith-Waterman [20] algorithm are integrated (see Table 2). For the discovery of conserved sequence motifs, the following motif databases are integrated: Pfam [21], Tigrfam [22], Conserved Domain Database [23] and SMART [24] with their specific search routines hmmpfam [25] and rps-blast (reverse posi-

tion specific blast). For the functional/structural classification of sequences the COG [26], KOG [27] and SCOP [28] databases are integrated. Moreover, Gene Ontologies (GOs) can be assigned by the system, inferred from the respective assignment of the integrated databases mentioned above. In addition, all sequence databases of any size, which are available in Fasta, Genbank or SwissProt format (e.g. Genbank, Swissprot or PIR) can be imported. These databases are handled as normal seq-sets and can be made available as a system-wide resource by the *Genlight* administrator. Analysis-specific pre-formatting of seq-sets is automatically done by the system. This modular architecture of the system and the flexible data model allows the straightforward integration of new analysis methods.

Table 2: Supported sequence analysis methods

Method	Explanation
BLASTN	Nucleotide Blast: Nucleotide query vs. nucleotide DB
BLASTP	Protein Blast: Protein query vs. protein DB
BLASTX	Translated nucleotide query vs. protein DB
TBLASTN	Protein query vs. translated nucleotide DB
TBLASTX	Translated query vs. translated nucleotide DB
PSIBLAST	Position specific iterated Blast: Protein query vs. protein DB
FASTA	Nucleotide query vs. nucleotide DB or protein query vs. protein DB
FASTX/Y	Nucleotide query vs. protein DB
TFASTA	Translated nucleotide query vs. translated nucleotide DB
SSEARCH	Smith-Waterman algorithm: Nucleotide query vs. nucleotide DB or protein query vs. protein DB
RPS-BLAST	Reverse position specific Blast: Protein query vs. CDD models
HMMPFAM	Hidden Markov Model based approach: Protein query vs. Pfam, TigrFam, or Smart models

Large scale sequence analysis

The comparison of whole genomes/proteomes or their use as query sets for searches in large databases like the Genbank-database or the SwissProt-database is a challenging and time consuming task. To compare, for instance, the mouse proteome to the human proteome by pairwise sequence comparison, approximately 41,000 (International Protein Index (IPI), February 2004) single homology searches with programs like BLAST or FASTA versus the human

proteome set (42,000 entries, IPI, February 2004) have to be performed. To handle such comparison tasks in an interactive system, the individual comparison calculations have to be done

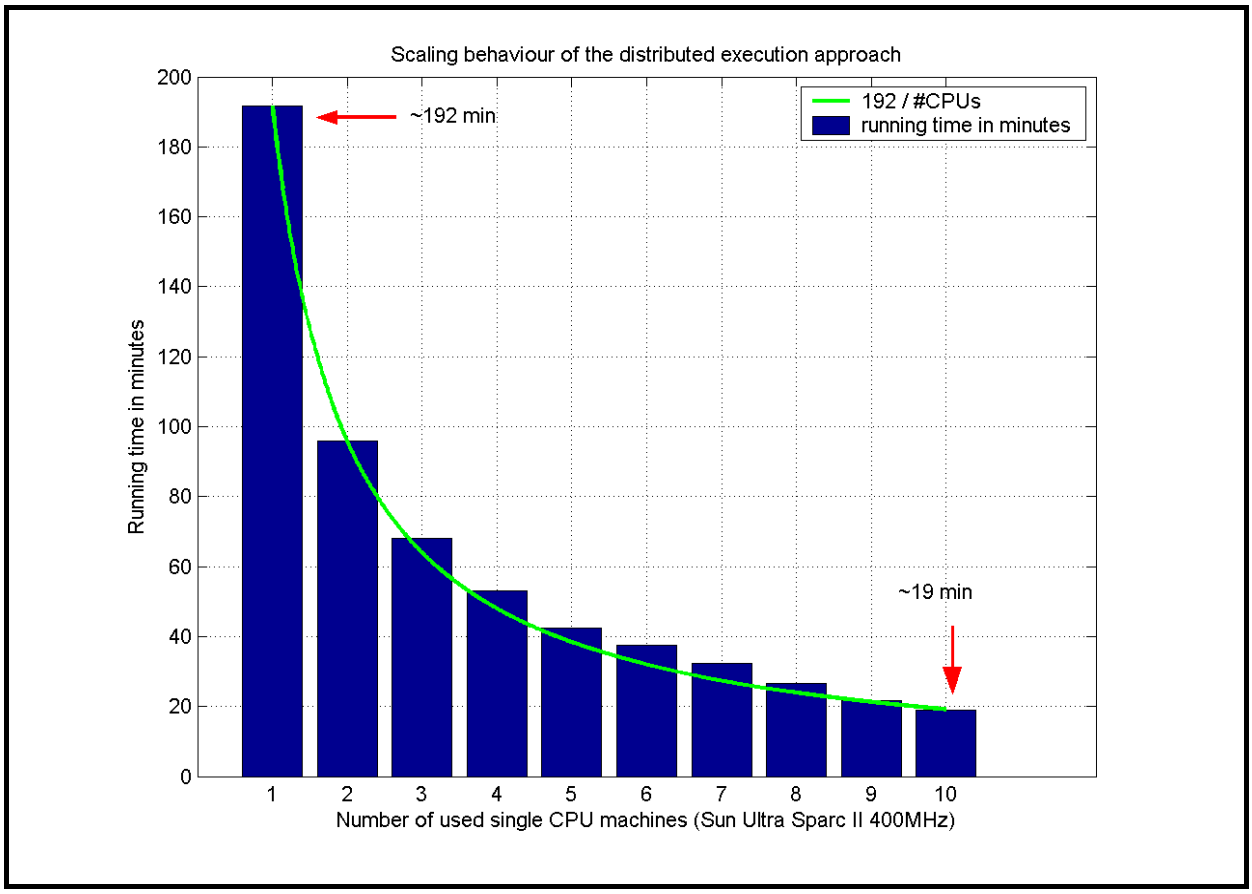

Figure 5: Scaling behavior of the distributed computing approach. Running times for a BLASTP comparison of the *H. pylori J99* proteome (1487 proteins) and the SwissProt protein database (134803 sequences) depending on the number of used CPUs.

asynchronously. Therefore, the *Genlight* server component contains a queuing mechanism for all analysis tasks. To process queued entries, the system has its own scheduling component, which allows a parallel, distributed execution of comparison jobs and can form a virtual cluster system of regular workstations for high throughput analysis tasks. The two major strengths of this approach are the complete integration into one system and a high robustness of the system, achieved by methods to insure data integrity during distributed execution. Compute nodes can be added to the virtual cluster and deleted from the virtual cluster at any time, by starting or stopping the *Genlight* client component on a workstation. The virtual cluster is completely manageable from the web based user interface. Due to the flexibility of the virtual cluster it is possible to temporarily exclude departmental workstations during working hours, while using idle compute-power during the night.

Table 3: Running times for different comparison methods using the *Genlight* virtual cluster system with 25 Sun Ultra SparcII CPUs on different workstations.

Query Set	DB Set	Method	Running time [hh:mm:ss]
H. pylori	*H. influenzae*	BLASTP	00:00:32
H. pylori	*V. cholerae*	PSIBLAST (10 iterations)	00:03:22
L. innocua	*L. monocytogenes*	BLASTN	00:00:27
H. pylori	CDD	RPS-BLAST	00:03:41
S. thyphimurium	SCOP40	BLASTP	00:00:42
S. cerevisiae	*A. thaliana*	BLASTP	00:03:48
H. pylori	Pfam	HMMPFAM	04:41:33
H. sapiens	*M. musculus*	BLASTP	02:17:30

This approach scales very well. The overall running time is nearly inversely proportional to the number of CPUs used (see Figure 5). The system design allows comparisons of complete genomes or proteomes and of large public databases in relative short times. Even CPU intensive tasks like Hidden Markov Model based approaches are processed in reasonable running times (see Table 3).

Application – Identification of potential drug targets in *Helicobacter pylori*

Helicobacter is a spiral shaped bacterium living in the stomach and duodenum of humans and in other mammalians [29]. Uncontrolled *H. pylori* infections are a major factor for duodenal ulcers, gastric ulcers, stomach cancer, and non-ulcer dyspepsia [30]. The sequencing of the *H. pylori* genome (strains *H. pylori* 26695 and *H. pylori* J99) offers the chance to develop highly specific treatments against *H. pylori* infections [31,32]. With the idea of minimizing toxicological effects, a perfect target protein should have low similarity to eukaryotic proteins. The strategy of this study was therefore to find all *H. pylori* proteins with low similarity to eukaryotes.

The *H. pylori J99* proteome (1,487 protein sequences) as published on the EBI-server (http://www.ebi.ac.uk/proteome/index.html) was compared to eukaryotic proteome sets (see Table 4) using the system-integrated BLASTP [17] method. All proteomes were obtained from the EBI-server in March 2004. To extract the sequences of *H. pylori* proteins having no homologues in any of the considered eukaryotic proteomes, an extraction filter for BLAST hit-sets was applied to the resulting hit-sets. For a stringent operation, the filter cutoff for the bit-score was set to bit score ≥ 30 (scoring matrix: BLOSUM62), and the minimum coverage rate, i.e. the

minimum sequence length covered by the matching pair, was set to zero. This setting results in the extraction of all sequences producing a hit in the eukaryotic sequence sets with a bit score < 30 regardless of a minimum overlap of both aligned sequences. The filter was used as negated filter because we were interested in *H. pylori* proteins with low similarity to eukaryotes. After this initial filtering step 226 *H. pylori* sequences remained.

Table 4: EBI proteome sets used in the comparative analysis

Organism	Number of protein sequences
H. sapiens (IPI)	43,426
M. musculus (IPI)	40,742
R. norvegicus (IPI)	33,028
A. thaliana	26,192
C. elegans	22,439
D. melanogaster	16,106
S. cerevisiae	6,195
P. falciparum	5,257
S. pombe	5,037
E. cuniculli	1,908
G. theta	451
Total	200,781

In a subsequent analysis step the remaining 226 protein sequences were screened for putative drug/vaccine targets using the system-integrated motif databases Pfam, Tigrfam, SMART and CDD. UreI, a well known putative drug target [33,34], which served as an internal control, was detected within this sequence set. UreI encodes an activated urea channel enabling urea access to intrabacterial urease at acidic pH. UreI is necessary for survival of *H. pylori* at pH < 4.0 [35].

Table 5: PFAM accession numbers for protein families involved in cell cycle process in bacteria

PFAM accession number	Description
pfam06160	Septation ring formation regulator, EzrA
pfam07432	Histone H1-like protein Hc1
pfam01098	Cell cycle protein
pfam00493	MCM2/3/5 family
pfam01189	NOL1/NOP2/sun family
pfam03568	Peptidase family C50
pfam07432	Histone H1-like protein Hc1

Vital processes, like the process of cell division, are of special interest for drug development. Proteins involved in these processes are quite often fundamental and therefore are putative drug targets. In order to find such putative targets the remaining 226 protein sequences were screened for several protein families involved in the cell cycle process in bacteria (see Table 5). This search resulted in a potential target, the cell division protein FtsW (pfam01098). FtsW is a polytopic membrane protein that is required for cell division in *E. coli* [36,37] and is present in virtually all bacteria having a peptidoglycan cell wall [38-40]. It is also discussed in the context of chemotherapeutic intervention of *M. tuberculosis* [41].

Table 6: PFAM accession numbers for outer membrane protein families in bacteria

PFAM accession number	Description
pfam02608	Basic membrane protein
pfam01075	Glycosyltransferase family 9 (heptosyltransferase)
pfam03865	Hemolysin activator HlyB
pfam02264	LamB porin
pfam04348	LppC putative lipoprotein
pfam04170	Uncharacterized lipoprotein NlpE involved in copper resistance
pfam02321	Outer membrane efflux protein
pfam03922	OmpW family
pfam04355	SmpA / OmlA family
pfam04932	O-Antigen Polymerase
pfam03895	YadA-like C-terminal region
pfam05244	Brucella outer membrane protein 2
pfam05818	Enterobacterial TraT complement resistance protein
pfam05844	YopD protein
pfam04728	Repeated sequence found in lipoprotein LPP
pfam05101	Type IV secretory pathway, VirB3-like protein
pfam06178	Oligogalacturonate-specific porin protein (KdgM)
pfam06316	Enterobacterial Ail/Lom protein
pfam06604	Bacterial outer membrane lipoprotein omp19
pfam06864	Pilin accessory protein (PilO)
pfam06901	RTX iron-regulated protein FrpC
pfam07012	Curlin associated repeat
pfam07017	Antimicrobial peptide resistance and lipid A acylation protein PagP
pfam03549	Translocated intimin receptor (Tir) intimin-binding domain
pfam07489	Translocated intimin receptor (Tir) C-terminus
pfam07490	Translocated intimin receptor (Tir) N-terminus
pfam00395	S-layer homology domain
pfam03502	Nucleoside-specific channel-forming protein, Tsx

PFAM accession number	Description
pfam03503	Chlamydia cysteine-rich outer membrane protein 3
pfam03504	Chlamydia cysteine-rich outer membrane protein 6
pfam00267	Gram-negative porin
pfam03518	Salmonella/Shigella invasin protein B
pfam00691	OmpA family
pfam01278	Omptin family
pfam03573	outer membrane porin, OprD family
pfam05137	Fimbrial assembly protein (PilN)
pfam04972	Putative phospholipid-binding domain
pfam04333	VacJ like lipoprotein
pfam00263	Bacterial type II and III secretion system protein
pfam01103	Surface antigen
pfam01308	Chlamydia major outer membrane protein
pfam01441	Lipoprotein
pfam03349	Outer membrane protein transport protein (OMPP1/FadL/TodX)
pfam04333	VacJ like lipoprotein
pfam02521	Putative outer membrane protein
pfam03077	putative vacuolating cytotoxin
pfam01856	Outer membrane protein
pfam02253	Phospholipase A1

Surface proteins playing a role in pathogen-host interaction represent potential targets for vaccination [42]. To find such putative targets within the specific *H. pylori* proteins, the 226 protein sequences were analyzed for the appearance of surface exposed proteins using an hmmpfam screening versus the Pfam database (see Table 6). Thirteen potential outer membrane proteins were found in the screening (see Table 7). These proteins could serve as potential candidates for vaccination. As protein families PF02521 and PF01856 have been seeded with *H. pylori* sequences, their occurrence in the results seems to be clear. However, due to significant similarities to eukaryotic proteins some of the *H. pylori* proteins belonging to these families could have been dismissed in the initial filtering step. Indeed, only five out of seven proteins belonging to family PF02521 and only six out of thirty-six protein sequences belonging to family PF1856 passed the filtering process.

The detection of UreI, FtsW and outer membrane proteins clearly demonstrates the ability of *Genlight* to identify potential targets via the differential genome comparison approach. UreA, UreB, VacA and other well known pharmaceutical targets were abandoned in the initial filtering step due to their significant similarity to eukaryotic proteins [43]. This finding is in accordance to our strategy to detect only proteins with very low similarity to eukaryotic proteins.

Table 7: Outer membrane proteins of *H. pylori* with no homologs in eukaryotes.

Pfam model accession Number	Model annotation	Matching *H. pylori* protein sequence	*H. pylori* sequence annotation	E-Value	Gene Ontology Assignments
PF02253	Phospholipase A1. Phospholipase A1 is a bacterial outer membrane bound acyl hydrolase with a broad substrate specifity	Q9ZLX5	Putative phospholipase A1	2.8E-223	GO:0004620, GO:0006629, GO:0016020
PF03349	Outer membrane protein transport protein (OMPP1/Fad/TodX)	Q9ZL05	Putative outer membrane protein	3.14E-10	GO:0009279
PF02521	This family consists of putative outer membrane proteins from Helicobacter pylori (campylobacter pylori)	Q9ZL61	Putative outer membrane protein	2.21E-261	GO:0009279
		Q9ZM80	Putative outer membrane protein	3.19E-301	GO:0009279
		Q9ZKT5	Putative outer membrane protein	<1.0E-400	GO:0009279
		Q9ZK48	Putative outer membrane protein	5.56E-278	GO:0009279
		Q9ZL55	Putative outer membrane protein	<1.0E-400	GO:0009279
PF01856	This family seems confined to Helicobacter pylori. It is predicted to be an outer membrane protein based on its pattern of alternating hydrophobic amino acids similar to porins	Q9ZMI2	Putative outer membrane protein	2.31E-94	
		Q9ZLG6	Putative outer membrane protein	6.16E-90	
		Q9ZLD5	Outer membrane protein	6.60E-81	
		Q9ZJ99	Putative	1.59E-50	
		Q9ZLJ8	Putative outer membrane protein	2.72E-42	
		Q9ZJ82	Putative outer membrane protein	3.82E-73	

Conclusion

Genlight is an extremely flexible system for comparative genomics. The underlying object relational database system together with the state of the art data representation allows to integrate various methods for the comparison of whole genomes or proteomes. In addition, the several different filters and operations which are predefined within the system, allow for a fast, easy, and user friendly post analysis of the generated results. The virtual cluster system guarantees reasonable running times even for the comparison of very large sequence-sets.

As shown, *Genlight* is well suited to tackle biological questions like the identification of organism-specific proteins, which could serve as potential target proteins in the drug discovery process. The integrated database lookup methods and the various data extraction and transformation methods, as well as the possibility to access the data even from applications outside *Genlight*, allow to characterize the identified nucleotide or protein sequences very rapidly. The high operation speed of *Genlight*, already confirmed by benchmarks listed in Table 3, could again be verified with the *H. pylori* study. (The BLASTP jobs ran for 25 minutes on seven nodes of an SGI Origin 3200 (MIPS R12K, 400 MHz) machine and five SGI O2 (MIPS R12K, 300 MHz) machines.

In principle, the approach of differential genome comparison exhibits several limitations which necessitate a careful interpretation of the results. Some of these limitations, such as distant gene relationship or multiple domain proteins can be addressed via the experienced usage of the software. Others, such as differential gene expression or uncertainties arising from incomplete genomes can only be addressed in subsequent in-depth analyses. *Genlight's* facility for the integration of user defined filters and its integrated secondary databases (PFAM, TIGRFAM, SMART, CDD, KOG, COG, SCOP) can also support such in-depth analyses.

Acknowledgements

The development of Genlight was supported by Intervet Innovation GmbH, Schwabenheim, Germany. We thank Prof. Terri K. Attwood and Prof. Curtis R. Altmann for many helpful discussions and for critically reading the manuscript.

References

[1] G. M. Rubin, M.D. Yandell, J. R. Wortmann, and G.L. Miklos, et al. Comparative Genomics of the Eukaryotes. Science, 287:2204-2215, 2000.

[2] G. Emilien, M. Ponchon, C. Caldas, O. Isacson, and J. M. Maloteux. Impact of genomics on drug discovery and clinical medicine. QJM, 93:391-423, 2000.

[3] P. Glaser, L. Frangeul, C. Buchrieser, C. Rusniok, et al. Comparative Genomics of Listeria Species. Science, 294:849-852, 2001.

[4] A. Bernal, U. Ear, and N. Kyrpides. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. Nucleic Acids Res., 29:126-127, 2001.

[5] E. Herrero, M. A. de la Torre, and E. Valentin. Comparative genomics of yeast species: new insights into their biology. Int. Microbiol., 6:183-190, 2003.

[6] E. V. Koonin. Comparative genomics, minimal gene sets, and the last universal common ancestor. Nature Reviews Microbiology, 1:127-136, 2003.

[7] T. Dandekar, F. Du, R. H. Schirmer, and S. Schmidt. Medical target prediction from genome sequence: combining different sequence analysis algorithms with expert knowledge and input from artificial intelligence approaches. Computers and Chemistry, 26:15-21, 2001.

[8] T. Gaasterland and C. W. Sensen. Fully Automated Genome Analysis that Reflects User Needs and Preferences - a Detailed Introduction to the MAGPIE System Architecture. Biochemie, 78:302-310, 1996.

[9] T. Gaasterland and C. W. Sensen. MAGPIE: automated genome interpretation. Trends Genet., 12:76-78, 1996.

[10] D. Frishman, K. Albermann, J. Hani, K. Heumann, A. Metanomski, A. Zollner, and H.-W. Mewes. Functional and structural genomics using PEDANT. Bioinformatics, 17:44-57, 2001.

[11] F. Meyer, A. Goesmann, A. C. Hardy, D. Bartels, T. Bekel, J. Clausen, J. Kalinkowski, B. Linke, O. Rupp, R. Giegerich, and A. Pühler. GenDB - an open source genome annotation system for prokaryote genomes. Nucleic Acids Res., 31:2187-2195, 2003.

[12] F. Chetouani, P. Glaser, and F. Kunst. FindTarget: a software for substractive genome analysis. Microbiology, 147:2643-2649, 2001.

[13] F. Chetouani, P. Glaser, and F. Kunst. DiffTool: building, visualizing and querying protein clusters. Bioinformatics, 18:1143-1144, 2002.

[14] R. Bruccoleri, T. Dougherty, and D. Davison. Concordance analysis of microbial genomes. Nucleic Acids Res., 26:4482-4486, 1998.

[15] S. Hoon, K. K. Ratnapu, J. M. Chia, B. Kumarasamy, X. Juguang, M. Clamp, A. Stabenau, S. Potter, L. Clarke, and E. Stupka. Biopipe: A Flexible Framework for Protocol-Based Bioinformatics Analysis. Genome Research, 13:1904-1915, 2003.

[16] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. J. Mol. Biol., 215:403-410, 1990.

[17] S. F. Altschul, T. L. Madden, A. A. Schaeffer, J. Zhang, W. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res., 25:3389-3402, 1997.

[18] W. R. Pearson. Using the FASTA program to search protein and DNA sequence databases. Methods Mol. Biol., 24:307-311, 1994.

[19] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA, 4:2444-2448, 1998.

[20] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. J. Mol. Biol., 147:195-197, 1981.

[21] A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. L. Sonnhammer. The Pfam protein families database. Nucleic Acids Res., 30:276-280, 2002.

[22] D. H. Haft, J. D. Selengut, and O. White. The TIGRFAMS database of protein families. Nucleic Acid Res., 31:371-373, 2003.

[23] A. Marchler-Bauer, J. B. Anderson, C. DeWeese-Scott, N. D. Fedorova *et al.* CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res., 31:383-387, 2003.

[24] I. Letunic, R. R. Copley, S. Schmidt, F. D. Cicarelli, T. Doerks, J. Schultz, C. P. Ponting, and P. Bork. SMART 4.0: Towards genomic data integration. Nucleic Acids Res., 32:D142-D144, 2004.

[25] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, and P. Bork. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res., 31:315-318, 2003.

[26] R. L. Tatusov, D. A. Natala, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res., 29:22-28, 2001.

[27] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, *et al.* The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 11:41-55, 2003.

[28] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Cothia, and A. G. Murzin. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res.,32:D226-D229, 2004.

[29] G. R. Turbett, P. B. Hoj, R. Horne, and B. J. Mee. Purification and characterization of the urease enzymes of Helicobacter species from humans and animals. Infect. Immun., 60:5259-5266, 1992.

[30] B. Marshall. Helicobacter pylori: 20 years on. Clin. Med., 2:147-152, 2002.

[31] J. F. Tomb, O. White, A. R. Kerlavage, R. A. Clayton, and G. G. Sutton. The complete genome sequence of the gastric pathogen Helicobacter pylori, Nature, 388:539-547, 1997.

[32] R. A. Alm, L. S. Ling, D. T. Moir, B. L. King, and E. D. Brown. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. Nature, 397:176-80, 1999.

[33] S. Skouloubris, J. M. Thiberge, A. Labigne, and H. De Reuse. The Helicobacter pylori UreI protein is not involved in urease activity but is essential for bacterial survival in vivo. Infect. Immun., 66, 4517-4521, 1998.

[34] S. Bury-Mone, S. Skouloubris, A. Labigne, and H. De Reuse. The Helicobacter pylori UreI protein: role in adaptation to acidity and identification of residues essential for its activity and for acid activation. Mol. Microbiol., 42:1021-34, 2001.

[35] M. Mollenhauer-Rektorschek, G. Hanauer, G. Sachs, and K. Melchers. Expression of UreI is required for intragastric transit and colonization of gerbil gastric mucosa by Helicobacter pylori. Res. Microbiol., 153:659-666, 2002.

[36] M. M. Khattar, K. J. Begg, and W. D. Donachie. Identification of FtsW and characterization of a new ftsW division mutant of Escherichia coli. J. Bacteriol., 176: 7140-7147, 1994.

[37] F. Ishino, H. K. Jung, M. Ikeda, M. Doi, M. Wachi, and M. Matsuhashi. New mutations fts-36, lts-33, and ftsW clustered in the mra region of the Escherichia coli chromosome induce thermosensitive cell growth and division. J. Bacteriol., 171:5523-5530, 1989.

[38] B. Lara and A. Ayala. Topological characterization of the essential Escherichia coli cell division protein FtsW. FEMS Microbiol. Lett., 216:23-32, 2002.

[39] M. Ikeda, T. Sato, M. Wachi, H. K. Jung, F. Ishino, Y. Kobayashi, and M. Matsuhashi. Structural similarity among Escherichia coli FtsW and RodA proteins and Bacillus subtilis SpoVE protein, which function in cell division, cell elongation, and spore formation, respectively. J. Bacteriol., 171:6375-6378, 1989.

[40] A. O. Henriques, P. Glaser, P. J. Piggot, and C. P. Moran Jr. Control of cell shape and elongation by the rodA gene in Bacillus subtilis. Mol. Microbiol., 28:235-247, 1998.

[41] P. Datta, A. Dasgupta, S. Bhakta, and J. Basu. Interaction between FtsZ and FtsW of Mycobacterium tuberculosis. J. Biol. Chem., 277:24983-24987, 2002.

[42] N. Sabarth, S. Lamer, U. Zimney-Arndt, P. R. Jungblut, , T. F. Meyer, and D. Bumann. Identification of surface proteins of Helicobacter pylori by selective biotinylation, affinity purification, and two-dimensional gel electrophoresis. J. Biol. Chem., 277:27896-27902, 2002.

[43] S. Suerbaum and C. Josenhans. Virulence factors of Helicobacter pylori: implications for vaccine development. Mol. Med. Today, 5:32-39, 1999.