Database

CoryneRegNet: An ontology-based data warehouse of corynebacterial transcription factors and regulatory networks

Jan Baumbach[1,2], Karina Brinkrolf[1,3], Lisa F Czaja[3], Sven Rahmann[2] and Andreas Tauch*[3]

Address: [1]International NRW Graduate School in Bioinformatics and Genome Research, Centrum für Biotechnologie, Universität Bielefeld, Universitätsstraße 25, D-33615 Bielefeld, Germany, [2]Algorithms and Statistics for Systems Biology Group, Genome Informatics, Technische Fakultät, Universität Bielefeld, Universitätsstraße 25, D-33615 Bielefeld, Germany and [3]Institut für Genomforschung, Centrum für Biotechnologie, Universität Bielefeld, Universitätsstraße 25, D-33615 Bielefeld, Germany

Email: Jan Baumbach - Jan.Baumbach@CeBiTec.Uni-Bielefeld.DE; Karina Brinkrolf - Karina.Brinkrolf@Genetik.Uni-Bielefeld.DE; Lisa F Czaja - liczaja@Genetik.Uni-Bielefeld.DE; Sven Rahmann - Sven.Rahmann@CeBiTec.Uni-Bielefeld.DE; Andreas Tauch* - Andreas.Tauch@Genetik.Uni-Bielefeld.DE

* Corresponding author

Published: 14 February 2006

Received: 11 November 2005
Accepted: 14 February 2006

Abstract

Background: The application of DNA microarray technology in post-genomic analysis of bacterial genome sequences has allowed the generation of huge amounts of data related to regulatory networks. This data along with literature-derived knowledge on regulation of gene expression has opened the way for genome-wide reconstruction of transcriptional regulatory networks. These large-scale reconstructions can be converted into in silico models of bacterial cells that allow a systematic analysis of network behavior in response to changing environmental conditions.

Description: CoryneRegNet was designed to facilitate the genome-wide reconstruction of transcriptional regulatory networks of corynebacteria relevant in biotechnology and human medicine. During the import and integration process of data derived from experimental studies or literature knowledge CoryneRegNet generates links to genome annotations, to identified transcription factors and to the corresponding cis-regulatory elements. CoryneRegNet is based on a multi-layered, hierarchical and modular concept of transcriptional regulation and was implemented by using the relational database management system MySQL and an ontology-based data structure. Reconstructed regulatory networks can be visualized by using the yFiles JAVA graph library. As an application example of CoryneRegNet, we have reconstructed the global transcriptional regulation of a cellular module involved in SOS and stress response of corynebacteria.

Conclusion: CoryneRegNet is an ontology-based data warehouse that allows a pertinent data management of regulatory interactions along with the genome-scale reconstruction of transcriptional regulatory networks. These models can further be combined with metabolic networks to build integrated models of cellular function including both metabolism and its transcriptional regulation.

Background

Microorganisms continuously have to handle changing environmental conditions to maintain their functional homeostasis and to overcome stress situations with detrimental consequences for growth and survival [1]. Therefore, they evolved mechanisms to sense alterations within their environmental surroundings and developed molecular strategies co-ordinated by complex transcriptional regulatory networks to manage unfavourable conditions. The complexity of such regulatory networks results from the interaction of numerous transcription units consisting of a transcription factor and a defined set of regulated target genes [2]. The most important components of these units are apparently the DNA-binding transcription factors. They are responsible for sensing environmental and intracellular signals to control cellular reproduction and growth [3], and they include a DNA-binding domain that possesses a secondary structure to recognize the operator sequences of regulated genes [4]. Depending on the growth conditions of a bacterial cell certain fractions of the total set of transcription factors are operating [5]. Some of them only control the expression of a single gene whereas others organize the activation or repression of numerous target genes [2].

The availability of whole genome sequences provides the opportunity to define the total set of DNA-binding transcription factors of an organism [6,7]. This is a first step not only in understanding the regulatory complexity of a certain bacterial cell but also for reconstructing the global connectivity of a regulatory network to theoretically describe and deduce gene expression pattern of a microorganism [8]. From a set of complete genome sequences it has been deduced that large genomes include more transcription factors per gene than small genomes [9]. The increase of genomic complexity is thus associated with a more complex regulation of gene expression since the additional genetic information has to be integrated into the existing regulatory network basically operating in a bacterial cell. The transcriptional regulatory network of *Escherichia coli* so far is one of the best characterized regulatory systems of a single cell. The total number of about 320 transcriptional regulators of *E. coli* K-12 were classified into eight distinct regulatory modules with defined physiological functions [5]. Additional bioinformatics studies suggested a hierarchical and modular structure of the regulatory network, excluding circular feedback loops on transcriptional level for this organism [10].

The genus *Corynebacterium* comprises a number of human pathogens, like *Corynebacterium diphtheriae* and *Corynebacterium jeikeium*, as well as the non-pathogenic soil bacteria *Corynebacterium glutamicum* and *Corynebacterium efficiens* that are widely used in biotechnological production processes of food and feed additives [11,12]. Because of their relevance in biotechnology and medicine the genome sequences of *C. glutamicum* ATCC 13032, *C. efficiens* YS-314, *C. diphtheriae* NCTC 13129, and *C. jeikeium* K411 have recently been determined [13-16]. First comparative analysis revealed a high-level conservation of orthologous genes in these genome sequences, indicating that the corynebacterial species have rarely undergone genome rearrangements and thus largely retained their ancestral genome structure [17]. An initial step in understanding the transcriptional regulatory machinery of corynebacteria was the bioinformatics identification of the encoded transcription factors [7]. A collection of 127 DNA-binding transcription factors was detected in the genome sequence of *C. glutamicum*, whereas 103 regulators were identified in *C. efficiens*, 63 in *C. diphtheriae* and 55 in *C. jeikeium*. The relation between these numbers agrees well with the assumption that the quantity of transcription factors of an organism is correlated to the genome size and the environmental surrounding a bacterial cell is exposed to [9]. Accordingly, the physiological versatility of *C. glutamicum* results in a considerably higher number of transcriptional regulators, and in consequence in a more complex regulatory network by integrating and co-ordinating additional regulatory subnetworks. According to amino acid comparisons and protein structure predictions the repertoire of DNA-binding transcription factors of *C. glutamicum*, *C. efficiens*, *C. diphtheriae*, and *C. jeikeium* were further on divided into 25 families of regulatory proteins. A common set of only 28 regulators was encoded by all of the four genome sequences and thus presumably includes the core set of DNA-binding transcription factors of these bacteria [7]. Despite the progress in bioinformatics prediction of transcription factors, the reconstruction of regulatory networks is generally hindered by the relatively low level of evolutionary conservation of other molecular network components, for instance of the cognate operator sequence of a DNA-binding transcription factor. However, developments in DNA microarray technology have allowed the generation of genome-wide data sets characterizing experimentally the regulatory networks of corynebacteria [18-20].

The ambition of our current post-genomic approaches is to decipher and reconstruct the transcriptional regulatory network of *C. glutamicum*. Here, we propose a hierarchical and modular scheme of the regulatory network, separating the repertoire of DNA-binding transcription factors into five well-defined and functionally distinct modules with respect to the physiological role of the regulated target genes. This biological concept was applied to design and implement the ontology-based data warehouse CoryneRegNet that provides a solid basis for further regulatory network studies in the field of systems biology. As an application example of CoryneRegNet we reconstructed and visualized the functional module "SOS and stress

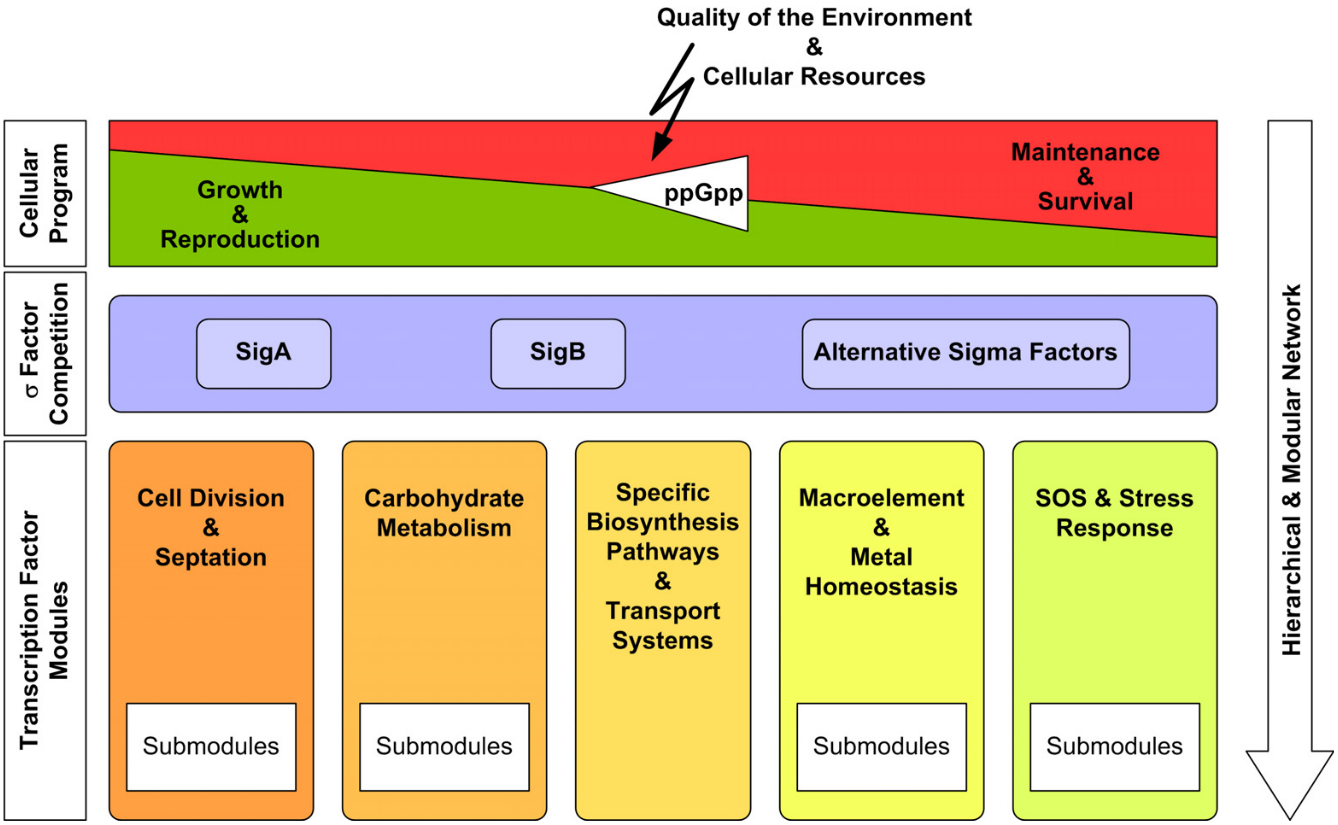


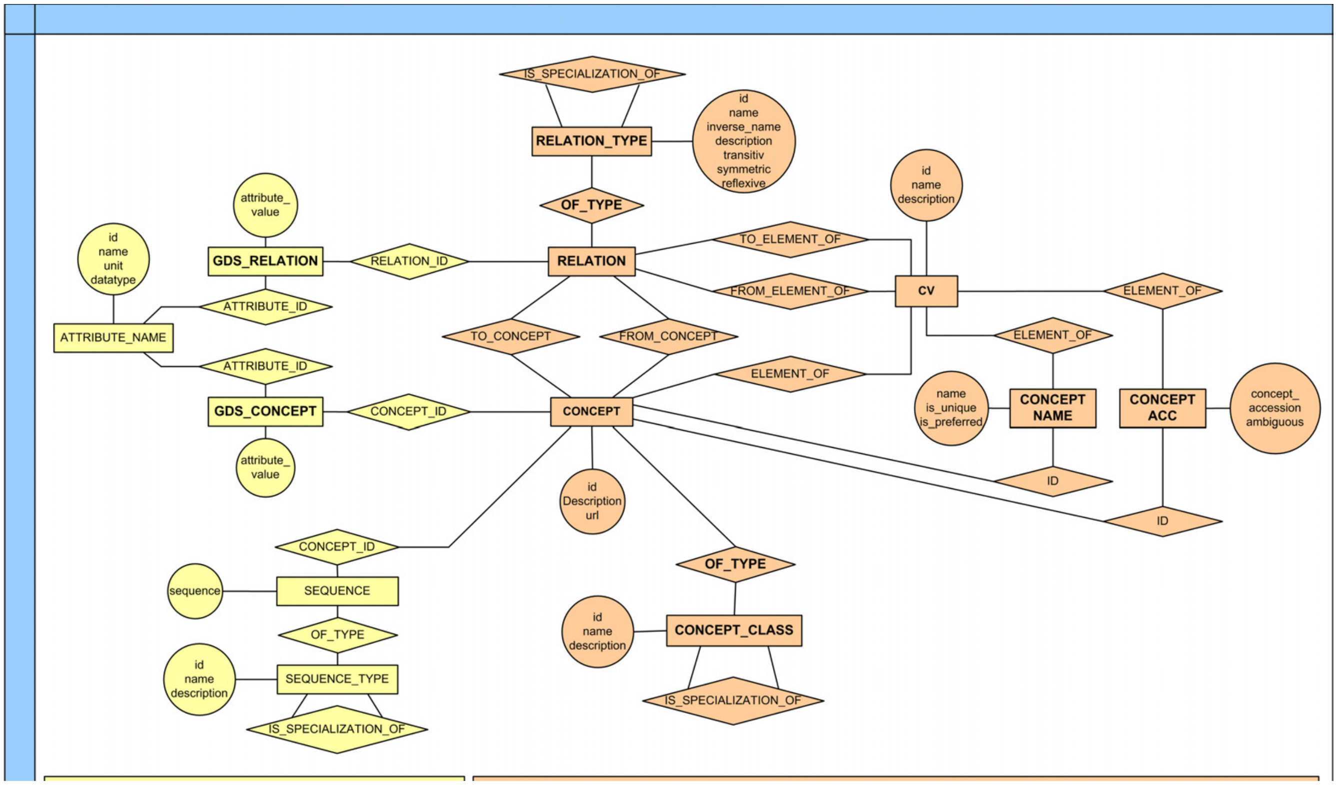
Figure 1
The biological concept of CoryneRegNet. The model presents the hierarchical and modular network structure of transcriptional regulatory interactions in *C. glutamicum*. It consists of five distinct transcription factor modules and a module containing the main and alternative sigma factors involved in differential gene expression by sigma factor competition [7]. A top level regulator is the hyperphosphorylated guanosine nucleotide ppGpp, involved in sensing the quality of the environment and the cellular resources [7]. The amount of ppGpp determines the cellular program and the role of sigma factor competition in global regulation of gene expression.

response" of *C. glutamicum*, revealing a multi-layered, hierarchical and modular structure of the respective transcriptional regulatory interactions.

Construction and content

The biological concept of CoryneRegNet

As a prerequisite for database design a biological concept of global transcriptional regulation of *C. glutamicum* was developed that later on was converted into a generalized and ontology-based data structure. In a first step, the total set of 127 DNA-binding transcription factors identified in the genome sequence of *C. glutamicum* [7] was classified into five functionally distinct modules (Figure 1). Using these categories along with the genome annotation, 68.3% of the transcription factors were grouped into the modules according to their proposed or evolutionary conserved function [7]. Transcription factors that did not fit into a module remained unclassified. Consequently, each transcription factor module includes several regulatory units composed of a transcription factor and a specific number of regulated genes. These units might be linked to form physiologically distinct submodules within the functional classes that might in turn assemble into a larger network including regulatory interactions between different transcription factor modules. A higher level of regulation of gene expression is represented by the seven sigma factors of *C. glutamicum* [14] that might exert their function in differential gene expression by the molecular mechanism of sigma factor competition [21,22]. As subunits of the RNA polymerase holoenzyme they recognize specific promoter sequences and alter the transcriptional profile of a cell in response to changing environmental conditions. Thereby sigma factors can influence the expression of certain target genes as well as of genes encoding transcription factors or even sigma factors. When adapting information from an *E. coli* model on hierarchical regulation of gene expression, an important environmental signal input into transcriptional regula-



Entity Relationship (ER) diagram representing the data structure used for the construction of CoryneRegNet. The ER was implemented in the DBMS MySQL and is divided into two main parts: the generalized data structure (*GDS*) and the ontology-based data structure. Rectangles represent entities, rhombi represent relations between two entities and circles represent attributes of entities. The entities *Concept* and *Relation*, which are the main components of the ontology-based data structure, are located in the center of the ER diagram. They store all essential data on genes, proteins and functional modules as well as every linkage between them. They are typed (*Concept_class* and *Relation_type*) and link to the controlled vocabulary (*CV*) they have been extracted from. Furthermore, they link to their generalized attributes (*GDS_relation* and *GDS_concept*) and to associated sequences (entity *Sequence*). Alternative names and accessions are stored in the tables of the entities *Concept_name* and *Concept_accession*.

tion of *C. glutamicum* occurs at the top level of the regulatory cascade that comprises the cellular programs "growth and reproduction" and "maintenance and survival" (Figure 1). Which of these programs dominates the actual physiological state of a cell is dependent on environmental conditions and the cellular resources and is inversely correlated with the cellular amount of the hyperphosphorylated guanosine nucleotide, ppGpp [21,23]. An accumulation of ppGpp occurs for instance in response to amino acid depletion triggering the so-called stringent response, but varying amounts of ppGpp were also observed in consequence to other environmental stresses [23]. In *C. glutamicum*, an increase in ppGpp synthesis is inversely correlated with growth rate and energy production of the cell [24,25]. However, the role of ppGpp as global regulator of gene expression in *C. glutamicum* awaits experimental verification. Respective studies in *E. coli* revealed that ppGpp directly intervenes with sigma factor competition by binding to the RNA polymerase core enzyme [23]. Consequently, this biological concept displays a hierarchical and modular structure composed of (i) the components determining the cellular program, (ii) the components of the sigma factor module involved in sigma factor competition and (iii) the five defined functional modules containing the complete repertoire of DNA-binding transcription factors predicted for *C. glutamicum* (Figure 1).

The database concept of CoryneRegNet

Generally, any kind of biological data can be considered as an ontology, which consists of concepts that are linked through relations. Accordingly, the goal was to integrate heterogenous data related to transcriptional regulation into a database in such a way that they fit into a single

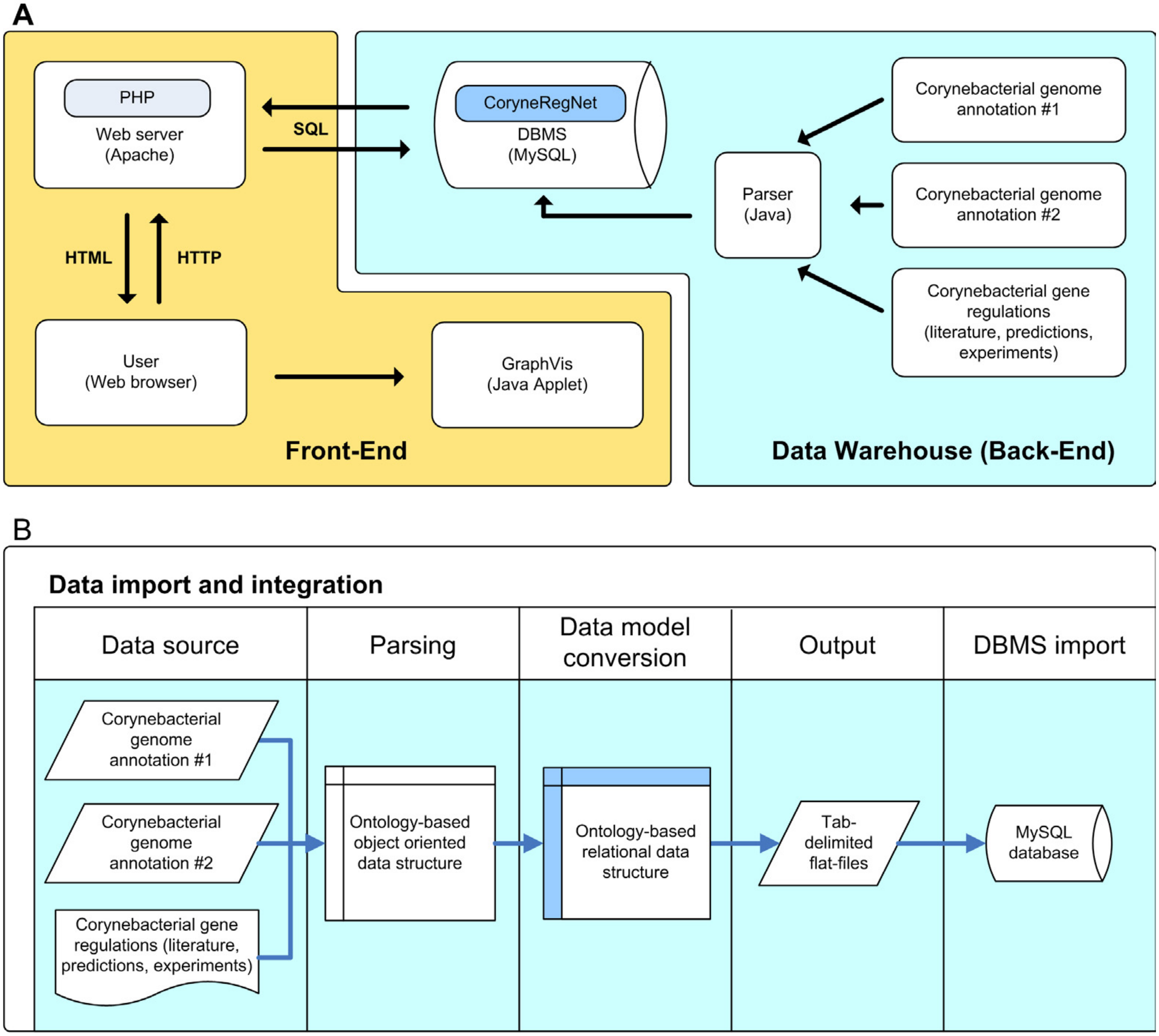
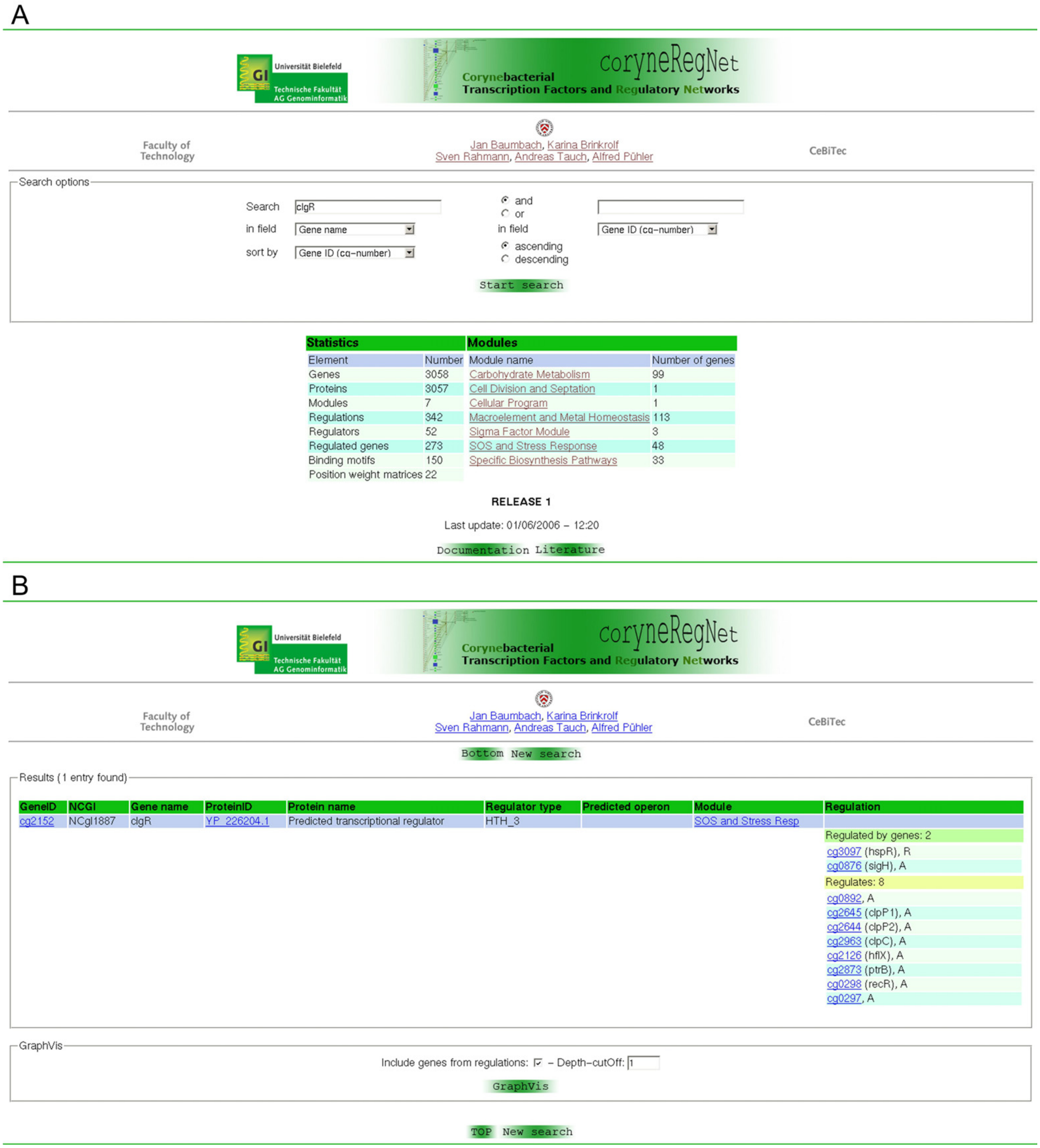

A
PHP
Web server
(Apache)
SQL
CoryneRegNet
DBMS
(MySQL)
Corynebacterial genome annotation #1
Parser
(Java)
Corynebacterial genome annotation #2
Corynebacterial gene regulations (literature, predictions, experiments)
HTML
HTTP
User
(Web browser)
GraphVis
(Java Applet)
Front-End
Data Warehouse (Back-End)
B
Data import and integration
Data source
Parsing
Data model conversion
Output
DBMS import
Corynebacterial genome annotation #1
Corynebacterial genome annotation #2
Corynebacterial gene regulations (literature, predictions, experiments)
Ontology-based object oriented data structure
Ontology-based relational data structure
Tab-delimited flat-files
MySQL database

Figure 3
Overview of the CoryneRegNet system (A) and the data import and integration process (B). The front-end consists of an Apache Web server, which queries the CoryneRegNet database back-end and constructs the browser-sided user interface and GraphVis, a Java applet that visualizes a queried result. The back-end is a data warehousing system that cross-links two corynebacterial genome annotations together with gene regulations and integrates the respective data into a single database. Data import and integration is illustrated in more detail in (B).

ontology-based data structure. In principle, technical and semantic data integration can be performed during data import. If a mechanism exists that ensures the correct semantics of the relations, then different data sources from different levels of biological hierarchy can be integrated into the same database scheme [26]. The data that have to be imported can be regarded as a set of structured and named concepts and the respective data sources are thus so-called controlled vocabularies (CVs). In CoryneRegNet, the data are first imported into a data repository, thereby creating a dataset concept for each biological entity, for instance genes, proteins or transcription factors, and a dataset relation for any connection between two concepts. Figure 2 shows the Entity Relationship diagram of CoryneRegNet, which was implemented in MySQL and which is similar to other ontological data structures, such



Jan Baumbach, Karina Brinkrolf
Sven Rahmann, Andreas Tauch, Alfred Pühler

Search options

Search: clgR ● and ○ or
in field: Gene name in field: Gene ID (cg-number)
sort by: Gene ID (cg-number) ● ascending ○ descending

Start search

Statistics		Modules	
Element	Number	Module name	Number of genes
Genes	3058	Carbohydrate Metabolism	99
Proteins	3057	Cell Division and Septation	1
Modules	7	Cellular Program	1
Regulations	342	Macroelement and Metal Homeostasis	113
Regulators	52	Sigma Factor Module	3
Regulated genes	273	SOS and Stress Response	48
Binding motifs	150	Specific Biosynthesis Pathways	33
Position weight matrices	22		

RELEASE 1

Last update: 01/06/2006 – 12:20

Documentation Literature

Jan Baumbach, Karina Brinkrolf
Sven Rahmann, Andreas Tauch, Alfred Pühler

Bottom New search

Results (1 entry found)

GeneID	NCGI	Gene name	ProteinID	Protein name	Regulator type	Predicted operon	Module	Regulation
cg2152	NCgl1887	clgR	YP_226204.1	Predicted transcriptional regulator	HTH_3		SOS and Stress Resp	Regulated by genes: 2 cg3097 (hspR), R cg0876 (sigH), A Regulates: 8 cg0892, A cg2645 (clpP1), A cg2644 (clpP2), A cg2963 (clpC), A cg2126 (hflX), A cg2873 (ptrB), A cg0298 (recR), A cg0297, A

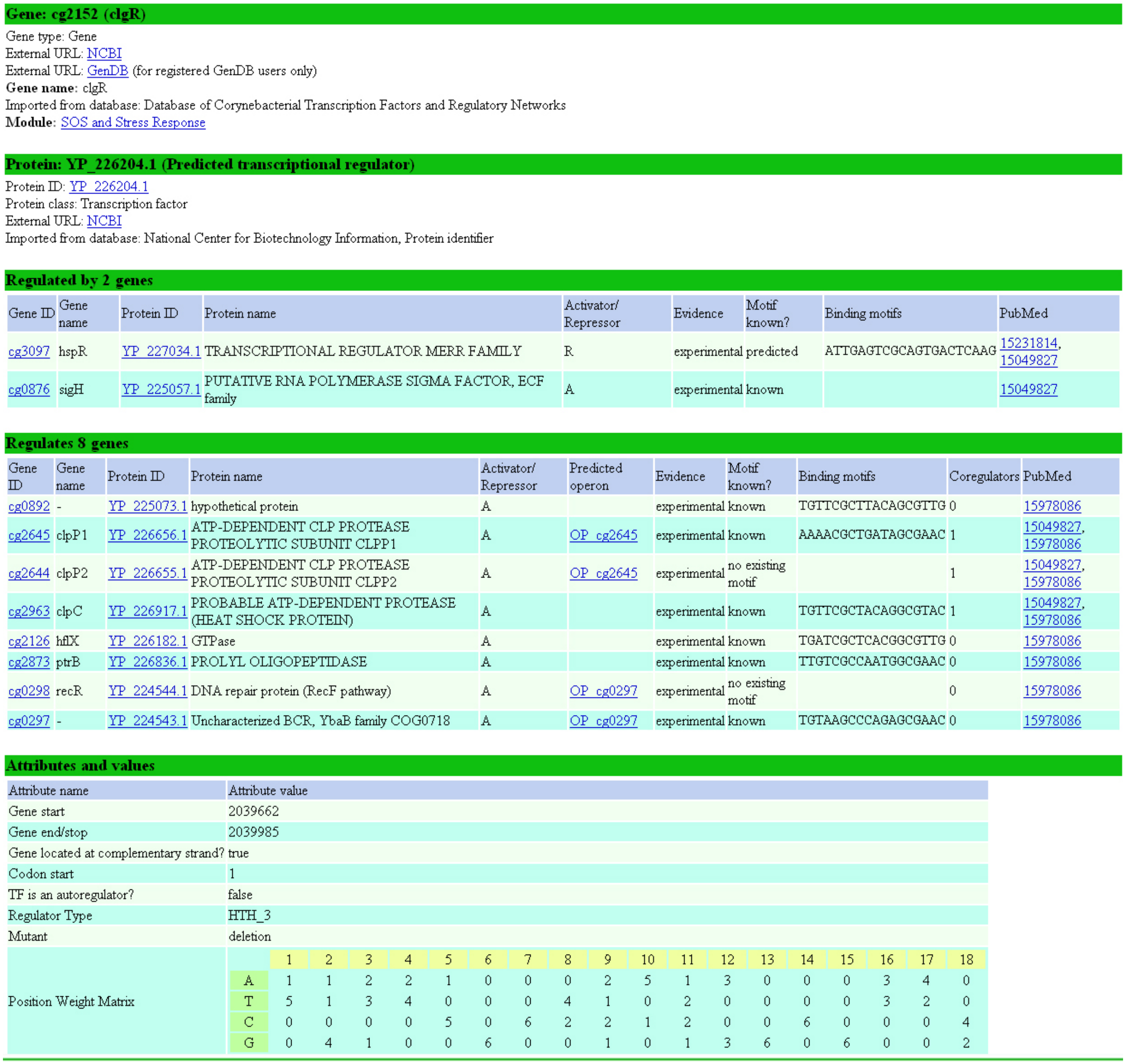
GraphVis

Include genes from regulations: ☑ – Depth-cutOff: 1

GraphVis

TOP New search

Figure 4
Web pages of CoryneRegNet. (A) Entry page of CoryneRegNet containing statistics and a search form. The database currently includes data on 52 transcriptional regulators exerting regulations on 273 genes. The regulators are classified into distinct physiological modules. The database can be searched by several criteria. (B) Exemplary results page representing relevant data of a searched item. Visualization of the respective transcriptional regulatory interactions by the GraphVis Java applet is applicable.



Gene: cg2152 (clgR)

Gene type: Gene
External URL: NCBI
External URL: GenDB (for registered GenDB users only)
Gene name: clgR
Imported from database: Database of Corynebacterial Transcription Factors and Regulatory Networks
Module: SOS and Stress Response

Protein: YP_226204.1 (Predicted transcriptional regulator)

Protein ID: YP_226204.1
Protein class: Transcription factor
External URL: NCBI
Imported from database: National Center for Biotechnology Information, Protein identifier

Regulated by 2 genes

Gene ID	Gene name	Protein ID	Protein name	Activator/ Repressor	Evidence	Motif known?	Binding motifs	PubMed
cg3097	hspR	YP_227034.1	TRANSCRIPTIONAL REGULATOR MERR FAMILY	R	experimental	predicted	ATTGAGTCGCAGTGACTCAAG	15231814, 15049827
cg0876	sigH	YP_225057.1	PUTATIVE RNA POLYMERASE SIGMA FACTOR, ECF family	A	experimental	known		15049827

Regulates 8 genes

Gene ID	Gene name	Protein ID	Protein name	Activator/ Repressor	Predicted operon	Evidence	Motif known?	Binding motifs	Coregulators	PubMed
cg0892	-	YP_225073.1	hypothetical protein	A		experimental	known	TGTTCGCTTACAGCGTTG	0	15978086
cg2645	clpP1	YP_226656.1	ATP-DEPENDENT CLP PROTEASE PROTEOLYTIC SUBUNIT CLPP1	A	OP_cg2645	experimental	known	AAAACGCTGATAGCGAAC	1	15049827, 15978086
cg2644	clpP2	YP_226655.1	ATP-DEPENDENT CLP PROTEASE PROTEOLYTIC SUBUNIT CLPP2	A	OP_cg2645	experimental	no existing motif		1	15049827, 15978086
cg2963	clpC	YP_226917.1	PROBABLE ATP-DEPENDENT PROTEASE (HEAT SHOCK PROTEIN)	A		experimental	known	TGTTCGCTACAGGCGTAC	1	15049827, 15978086
cg2126	hflX	YP_226182.1	GTPase	A		experimental	known	TGATCGCTCACGGCGTTG	0	15978086
cg2873	ptrB	YP_226836.1	PROLYL OLIGOPEPTIDASE	A		experimental	known	TTGTCGCCAATGGCGAAC	0	15978086
cg0298	recR	YP_224544.1	DNA repair protein (RecF pathway)	A	OP_cg0297	experimental	no existing motif		0	15978086
cg0297	-	YP_224543.1	Uncharacterized BCR, YbaB family COG0718	A	OP_cg0297	experimental	known	TGTAAGCCCAGAGCGAAC	0	15978086

Attributes and values

Attribute name	Attribute value
Gene start	2039662
Gene end/stop	2039985
Gene located at complementary strand?	true
Codon start	1
TF is an autoregulator?	false
Regulator Type	HTH_3
Mutant	deletion
Position Weight Matrix	

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	1	1	2	2	1	0	0	0	2	5	1	3	0	0	0	3	4	0
T	5	1	3	4	0	0	0	4	1	0	2	0	0	0	0	3	2	0
C	0	0	0	0	5	0	6	2	2	1	2	0	0	6	0	0	0	4
G	0	4	1	0	0	6	0	0	1	0	1	3	6	0	6	0	0	2

Figure 5
CoryneRegNet data page of a gene encoding a DNA-binding transcription factor. The page summarizes data relevant for the reconstruction of regulatory networks. It is linked to other databases such as NCBI Entrez, GenDB and PubMed. The example illustrates data for ClgR that is involved in stress response of *C. glutamicum* [39]. The DNA-binding motifs of ClgR are used to deduce a position weight matrix.

as the ONDEX (ONtological inDEXing) data structure [26]. The most important entities located in the center of the diagram are *Concept* and *Relation* (Figure 2). These two entities are connected by the relations *From_concept* and *To_concept*. *Relations* and *Concepts* are typed and accordingly there are the entities *Concept_class* and *Relation_type* that reference themselves by the relationship *Is_specialization_of*. For instance *Transcription factor* is a specialization of *Protein* or the relation *Activates* is a specialization of *Binds_to*. Additionally, every concept and every relation links to the CV it has been extracted from. In particular, there are references such as *Element_of*, *From_element_of* and *To_element_of* that link the concept item and relation item to the entity *CV*, respectively (Figure 2). Furthermore, accessions are unique identifiers that unambiguously determine a concept but solely within a specific CV. *Concepts* might have different names and accessions in different databases. This information is stored in tables for the entities *Concept_name* and *Concept_accession*. Moreover, concepts and relations might have attached values, for instance genes have start and stop positions within the genome sequence, whereas transcription factors can be associated with position weight matrices deduced from the cognate DNA-binding sites. Thus, it is necessary to have also a generalized data structure, the entity *GDS*, to store concept or relation specific values in a generalized way (Figure 2).

According to this principle, CoryneRegNet is designed as a web-based software environment (Figure 3A) that is publicly available. At the front-end, the scripting language PHP 4.3.2 http://www.php.net/ is used for the development of a user interface. CoryneRegNet runs on an Apache HTTP server 2.0.49 http://www.apache.org/ and queries the open source database management system MySQL 4.1.9 http://www.mysql.com/, which is used for data storage. The program enabling data import and data integration is implemented in Java 1.4.2 http://java.sun.com/ as is the graph-based network visualization tool GraphVis, which uses an academic licence version of the yFiles JAVA graph library http://www.yworks.com/[27]. The entire system was developed and runs on servers configured with the operating system Solaris 9/SunOS 5.9.

For the reconstruction of corynebacterial regulatory networks, the complete genome sequence of *C. glutamicum* along with the genome annotation [14] was downloaded from NCBI [28] in GenBank format and imported into CoryneRegNet. Subsequently, the gene identifiers were mapped to a second *C. glutamicum* genome sequence and annotation [29] to enable scientists working with either of the two annotations the efficient usage of CoryneRegNet. Furthermore, biological data relevant to transcriptional regulations were imported into the database as derived from literature knowledge (included in the database as PubMed link) [30], computer predictions [7] or experimental studies [19,20] (Figure 3B). The data import process was realized by running a parser that was implemented in Java. The parser software additionally integrates the imported data into a single ontology-based data structure and converts it into a relational data model (Figure 2). The output are tab-delimited flat-files that in turn are input files for the MySQL built-in import procedure and finally used to fill the CoryneRegNet database (Figure 3B).

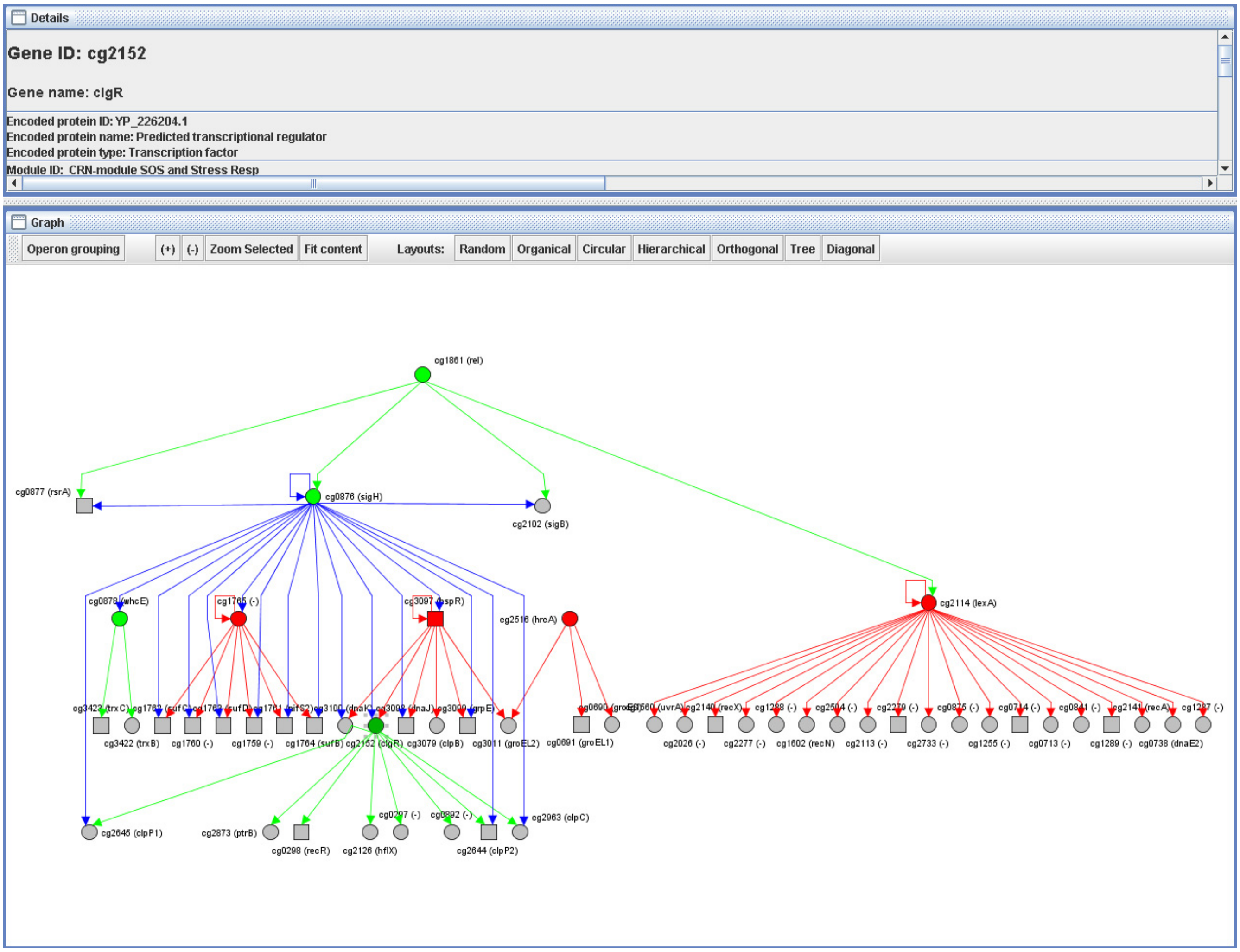
Utility and discussion

The user interface of CoryneRegNet

Web-based user interfaces to biological databases often support the following tasks: (i) *browsing* by listing or navigating through database entries, (ii) *searching* by identifying entries based on restrictions on the values of data fields within the database, (iii) *visualizing* by presenting a visual representation of the data, and (iv) *querying* by specifying a special search using a query building interface [31]. As well as other gene regulatory databases, such as PRODORIC [32,33], CoryneRegNet also emphasizes browsing, searching and visualizing. The entry page of CoryneRegNet shows a statistical summary of the data currently integrated into the database and provides the possibility to browse the functional modules of transcription factors (Figure 4A). Alternatively, the user can start searching the database using criteria that were obtained through a requirements analysis with potential users. The criteria are implemented following the typical search mask style (Figure 4A) of other gene regulatory databases, such as PRODORIC or TRANSFAC [32,34]. The search results are presented in a table-based style (Figure 4B) including gene and protein identifiers and names, the regulator type (if the specific protein is a transcription factor), the functional module the gene belongs to and the transcriptional regulations the gene is involved in. The user may acquire additional information on specific elements by clicking on them. A typical detailed view of data regarding a transcription factor gene is presented in Figure 5. It is possible to navigate to other entries of CoryneRegNet, to the genome annotation system GenDB [35] and to the NCBI Entrez Gene database [30] by following the respective links.

Graphical visualization of regulatory interactions

The user can visualize a transcriptional regulatory network at every navigation point using a result table or a detailed frame as starting point. The user has to define a graph depth cut-off and whether genes from hierarchical regulations should be included into the graph (Figure 4B). Graph construction starts with the selected set of genes, propagates through the regulatory network and adds more genes into the graph until the depth cut-off has been reached. The respective algorithm was implemented in PHP and generates a HTML file that in turn starts the Java applet GraphVis. Due to the security restrictions of the Java Virtual Machine, the whole graph is transferred by using named "PARAM tags" inside the "APPLET tag" before the GraphVis applet appears. Once the applet has been started, it is not able to "leave" the virtual machine. Accordingly, the whole graph along with all additional information on specific elements, for instance genes and proteins, is currently generated and transferred to the applet at start-up. Figure 6 shows an exemplary GraphVis Java applet window. The user obtains the same details on genes, proteins and regulatory interactions as in the browser-based view of CoryneRegNet. The main advantage is the graphical overview of the reconstructed regulatory network, where nodes in the graph represent genes



Gene ID: cg2152

Gene name: clgR

Encoded protein ID: YP_226204.1
Encoded protein name: Predicted transcriptional regulator
Encoded protein type: Transcription factor
Module ID: CRN-module SOS and Stress Resp

Figure 6
GraphVis Java applet showing the reconstruction of the SOS and stress response module of *C. glutamicum*. The graph was generated by means of the yFiles JAVA graph library using the hierarchical layout mode. Nodes represent genes included in this functional module. Color code: red node and line, repressor and repressing regulatory interaction; green node and line, activator and activating regulatory interaction; green node and blue line, sigma factor and sigma factor interaction; gray node, regulated target gene preceded by a transcription factor binding site; gray box, regulated target gene that is part of an operon and not preceded by a transcription factor binding site. The top level regulation of gene expression is indicated by the green node of the *rel* gene that is responsible for the cellular amount of ppGpp.

and edges represent regulatory relationships. The user can zoom into the graph, layout the graph by using different styles, remove selected elements from the graph or retrieve detailed information on selected genes (Figure 6).

Reconstruction of the SOS and stress response module

We used CoryneRegNet to reconstruct and visualize the transcriptional regulatory network of the SOS and stress response module of *C. glutamicum* (Figure 6). The module currently includes six DNA-binding transcription factors and 42 regulated genes. Since sigma factors play a key role in regulating gene expression when the cell is exposed to stress conditions and switches in part to the program "maintenance and survival" [21,36], the regulatory network is apparently linked to components of the sigma factor competition module. Thus, the reconstructed network reveals a hierarchical scheme also including the top level regulator ppGpp, synthesized by the Rel protein and influencing expression of the sigma factors SigH and SigB [21,22]. The reconstructed network allowed us to characterize the transcription factor module "SOS and stress response" in more detail: Several genes are under dual control by a DNA-binding transcription factor and by the alternative sigma factor SigH, whereas the *groEL2* gene is co-regulated by two transcription factors. The network is additionally characterized by a number of autoregulatory loops (Figure 6) in which the transcription factor regulates its own expression. Regarding regulatory network motifs, the presence of feed-forward loops is apparent when considering the regulatory action on gene expression of both a transcriptional regulator (HspR or ClgR) and an alternative sigma factor (SigH). This is consistent with observations in *E. coli* that feed-forward loop motifs tend to be implemented within modules, whereas bi-fan motifs seem to be responsible for the connection between different physiological modules [5]. Two types of feed-forward loops are present in the reconstructed network of the SOS and stress response module, namely the coherent type 1 and the incoherent type 1 motif [37]. In a coherent type 1 feed-forward loop all the regulatory connetions are activating (SigH, ClgR, ClpP1-ClpP2), while in the incoherent type 1 motif one of the regulatory links represses the activity of the target node (SigH, HspR, DnaK). It is also apparent that the reconstructed regulatory network is composed of two distinct submodules reflecting different responses of the cell upon exposure to environmental stresses (Figure 6). The SOS response is induced by DNA damage and under control of the LexA protein, while the heat-shock and oxidative stress response is induced by denaturation and/or inactivation of proteins and is under SigH control [1]. Accordingly, the reconstruction and visualization of the SOS and stress response module of *C. glutamicum* by CoryneRegNet reflects the hierarchical and modular scheme of the cell's transcriptional regulatory system.

Conclusion

With the recent progress made in large-scale postgenomic analysis of complete genome sequences a vast amount of novel data is becoming available. Comprehensive evaluation of postgenomic data asks for user-oriented databases supporting data management and data integration into existing knowledge. The CoryneRegNet database discloses detailed information on DNA-binding transcription factors, the key players in regulation of gene expression, and on transcriptional regulatory interactions of *C. glutamicum* deduced from literature-derived knowledge, computer predictions and global DNA microarray hybridization experiments. A web-based user interface provides access to the database content, allows various queries and supports the reconstruction and visualization of regulatory networks at different hierarchical levels. CoryneRegNet is moreover linked to the NCBI Entrez Gene database to provide direct access to corresponding genomic data. Although CoryneRegNet was developed as a data warehouse of transcriptional regulatory networks of *C. glutamicum*, its ontology-based design along with its programs and scripts is generally applicable to implement other species-specific databases. Consequently, CoryneRegNet is a versatile systems biology tool to support the large-scale analysis of transcriptional regulation of gene expression in microorganisms. The ultimate purpose of CoryneRegNet is to assist in reconstruction of transcriptional regulatory networks and to provide models that can be combined with metabolic networks of the cell to build integrated models including both cellular metabolism and transcriptional regulation. Since comparative computer analyses exploiting transcriptional regulatory data might be helpful to uncover hidden information on regulation of gene expression, transcriptional data of other sequenced corynebacterial species will be integrated into the next release of CoryneRegNet. For the future, we further plan to integrate existing and currently developing bioinformatics tools to perform for instance genome-wide searches for regulatory motifs specified by position weight matrices with sound statistical analysis [38] or to discover new potential motifs based on transcriptional profile analysis and comparative sequence analysis over several related genomes. We would also like to integrate algorithms and visualization techniques for comparing regulatory networks in several species. All of the above areas are active research fields with several new ideas being presented at every bioinformatics conference; therefore we are planning a flexible external tool plug-in concept for CoryneRegNet. Our long-term vision consists of CoryneRegNet proposing new regulatory hypotheses for wet-lab verification. While we expect that it will take some time for this vision to become reality, already now CoryneRegNet is a free open-source central repository and analysis tool for regulatory networks of microorganisms that is easy to extend because of its ontology-based design.

Availability and requirements

The CoryneRegNet database is freely accessible through the website https://www.cebitec.uni-bielefeld.de/groups/gi/software/coryneregnet/. Application of the yFiles JAVA graph library is restricted to academic users. Programs, scripts and information for setting up a species-specific database can be obtained from the authors upon request.

Authors' contributions

JB designed and implemented CoryneRegNet and drafted the manuscript. KB developed the biological concept of CoryneRegNet, evaluated the data and participated in drafting the manuscript. LFC participated in data evaluation of the SOS and stress response module. SR participated in co-ordination and supervision and conceived of the design of CoryneRegNet. AT conceived of the study and participated in data evaluation and supervision. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Iris Brune, Olaf Brockmann-Gretza and Alfred Pühler for helpful discussions and Alexander Goesmann, Ralf Nolte and Peter Serocka for expert technical support.

References

1. Matic I, Taddei F, Radman M: **Survival versus maintenance of genetic stability: a conflict of priorities during stress.** *Res Microbiol* 2004, **155(5):**337-341.
2. Teichmann SA, Madan Babu M: **Gene regulatory network growth by duplication.** *Nat Genet* 2004, **36(5):**492-496.
3. Madan Babu M, Teichmann SA: **Evolution of transcription factors and the gene regulatory network in Escherichia coli.** *Nucleic Acids Res* 2003, **31(4):**1234-1244.
4. Pabo CO, Sauer RT: **Transcription factors: structural families and principles of DNA recognition.** *Annu Rev Biochem* 1992, **61:**1053-1095.
5. Resendis-Antonio O, Freyre-Gonzalez JA, Menchaca-Mendez R, Gutierrez-Rios RM, Martinez-Antonio A, Avila-Sanchez C, Collado-Vides J: **Modular analysis of the transcriptional regulatory network of E. coli.** *Trends Genet* 2005, **21(1):**16-20.
6. Perez-Rueda E, Collado-Vides J: **The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12.** *Nucleic Acids Res* 2000, **28(8):**1838-1847.
7. Brune I, Brinkrolf K, Kalinowski J, Pühler A, Tauch A: **The individual and common repertoire of DNA-binding transcriptional regulators of Corynebacterium glutamicum, Corynebacterium efficiens, Corynebacterium diphtheriae and Corynebacterium jeikeium deduced from the complete genome sequences.** *BMC Genomics* 2005, **6(1):**86.
8. Herrgard MJ, Covert MW, Palsson BO: **Reconstruction of microbial transcriptional regulatory networks.** *Curr Opin Biotechnol* 2004, **15(1):**70-77.
9. Cases I, de Lorenzo V, Ouzounis CA: **Transcription regulation and environmental adaptation in bacteria.** *Trends Microbiol* 2003, **11(6):**248-253.
10. Ma HW, Buer J, Zeng AP: **Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach.** *BMC Bioinformatics* 2004, **5(1):**199.
11. Fudou R, Jojima Y, Seto A, Yamada K, Kimura E, Nakamatsu T, Hiraishi A, Yamanaka S: **Corynebacterium efficiens sp. nov., a glutamic-acid-producing species from soil and vegetables.** *Int J Syst Evol Microbiol* 2002, **52(Pt 4):**1127-1131.
12. Hermann T: **Industrial production of amino acids by coryneform bacteria.** *J Biotechnol* 2003, **104(1-3):**155-172.
13. Tauch A, Kaiser O, Hain T, Goesmann A, Weisshaar B, Albersmeier A, Bekel T, Bischoff N, Brune I, Chakraborty T, Kalinowski J, Meyer F, Rupp O, Schneiker S, Viehoever P, Pühler A: **Complete genome sequence and analysis of the multiresistant nosocomial pathogen Corynebacterium jeikeium K411, a lipid-requiring bacterium of the human skin flora.** *J Bacteriol* 2005, **187(13):**4671-4682.
14. Kalinowski J, Bathe B, Bartels D, Bischoff N, Bott M, Burkovski A, Dusch N, Eggeling L, Eikmanns BJ, Gaigalat L, Goesmann A, Hartmann M, Huthmacher K, Krämer R, Linke B, McHardy AC, Meyer F, Möckel B, Pfefferle W, Pühler A, Rey DA, Rückert C, Rupp O, Sahm H, Wendisch VF, Wiegräbe I, Tauch A: **The complete Corynebacterium glutamicum ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins.** *J Biotechnol* 2003, **104(1-3):**5-25.
15. Cerdeno-Tarraga AM, Efstratiou A, Dover LG, Holden MT, Pallen M, Bentley SD, Besra GS, Churcher C, James KD, De Zoysa A, Chillingworth T, Cronin A, Dowd L, Feltwell T, Hamlin N, Holroyd S, Jagels K, Moule S, Quail MA, Rabbinowitsch E, Rutherford KM, Thomson NR, Unwin L, Whitehead S, Barrell BG, Parkhill J: **The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129.** *Nucleic Acids Res* 2003, **31(22):**6516-6523.
16. Nishio Y, Nakamura Y, Kawarabayasi Y, Usuda Y, Kimura E, Sugimoto S, Matsui K, Yamagishi A, Kikuchi H, Ikeo K, Gojobori T: **Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of Corynebacterium efficiens.** *Genome Res* 2003, **13(7):**1572-1579.
17. Nakamura Y, Nishio Y, Ikeo K, Gojobori T: **The genome stability in Corynebacterium species due to lack of the recombinational repair system.** *Gene* 2003, **317(1-2):**149-155.
18. Hüser AT, Becker A, Brune I, Dondrup M, Kalinowski J, Plassmeier J, Pühler A, Wiegräbe I, Tauch A: **Development of a Corynebacterium glutamicum DNA microarray and validation by genome-wide expression profiling during growth with propionate as carbon source.** *J Biotechnol* 2003, **106(2-3):**269-286.
19. Koch DJ, Rückert C, Albersmeier A, Hüser AT, Tauch A, Pühler A, Kalinowski J: **The transcriptional regulator SsuR activates expression of the Corynebacterium glutamicum sulphonate utilization genes in the absence of sulphate.** *Mol Microbiol* 2005, **58(2):**480-494.
20. Rey DA, Nentwich SS, Koch DJ, Rückert C, Pühler A, Tauch A, Kalinowski J: **The McbR repressor modulated by the effector substance S-adenosylhomocysteine controls directly the transcription of a regulon involved in sulphur metabolism of Corynebacterium glutamicum ATCC 13032.** *Mol Microbiol* 2005, **56(4):**871-887.
21. Nyström T: **Growth versus maintenance: a trade-off dictated by RNA polymerase availability and sigma factor competition?** *Mol Microbiol* 2004, **54(4):**855-862.
22. Ishihama A: **Functional modulation of Escherichia coli RNA polymerase.** *Annu Rev Microbiol* 2000, **54:**499-518.
23. Magnusson LU, Farewell A, Nyström T: **ppGpp: a global regulator in Escherichia coli.** *Trends Microbiol* 2005, **13(5):**236-242.
24. Ruklisha MP, Damberga BE, Shvinka JE: **Stringend control and ppGpp synthesis in Brevibacterium flavum during amino acid starvation.** *Proc Latvian Acad Sci* 1993, **12:**59-70.
25. Ruklisha MP, Viesturs U, Labane L: **Growth control and ppGpp synthesis in Brevibacterium flavum cells at various medium mixing rates and aeration intensities.** *Acta Biotechnol* 1995, **15(1):**41-48.
26. Köhler J, Rawlings C, Verrier P, Mitchell R, Skusa A, Ruegg A, Philippi S: **Linking experimental results, biological networks and sequence analysis methods using Ontologies and Generalised Data Structures.** In *Silico Biol* 2005, **5(1):**33-44.
27. Wiese R, Eiglsperger M, Kaufmann M: **yFiles: Visualization and Auomaic Layout of Graphs, Proceeding.** Springer Verlag; 2001:453 ff.
28. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L: **Database resources of the National Center for Biotechnology.** *Nucleic Acids Res* 2003, **31(1):**28-33.
29. Ikeda M, Nakagawa S: **The Corynebacterium glutamicum genome: features and impacts on biotechnological processes.** *Appl Microbiol Biotechnol* 2003, **62(2-3):**99-109.
30. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2005, **33(Database issue):**D39-45.
31. Garwood K, McLaughlin T, Garwood C, Joens S, Morrison N, Taylor CF, Carroll K, Evans C, Whetton AD, Hart S, Stead D, Yin Z, Brown AJ, Hesketh A, Chater K, Hansson L, Mewissen M, Ghazal P, Howard J, Lilley KS, Gaskell SJ, Brass A, Hubbard SJ, Oliver SG, Paton NW: **PEDRo: a database for storing, searching and disseminating experimental proteomics data.** *BMC Genomics* 2004, **5(1):**68.
32. Münch R, Hiller K, Barg H, Heldt D, Linz S, Wingender E, Jahn D: **PRODORIC: prokaryotic database of gene regulation.** *Nucleic Acids Res* 2003, **31(1):**266-269.
33. Münch R, Hiller K, Grote A, Scheer M, Klein J, Schobert M, Jahn D: **Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes.** *Bioinformatics* 2005.
34. Wingender E: **TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks.** In *Silico Biol* 2004, **4(1):**55-61.
35. Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, Pühler A: **GenDB--an open source genome annotation system for prokaryote genomes.** *Nucleic Acids Res* 2003, **31(8):**2187-2195.
36. Nyström T: **Conditional senescence in bacteria: death of the immortals.** *Mol Microbiol* 2003, **48(1):**17-23.
37. Mangan S, Alon U: **Structure and function of the feed-forward loop network motif.** *Proc Natl Acad Sci U S A* 2003, **100(21):**11980-11985.
38. Rahmann S, Müller T, Vingron M: **On the power of profiles for transcription factor binding site detection.** *Statistical Applications in Genetics and Molecular Biology* 2003, **2(1):**Article 7.

39. Engels S, Schweitzer JE, Ludwig C, Bott M, Schaffer S: **clpC and clpP1P2 gene expression in Corynebacterium glutamicum is controlled by a regulatory network involving the transcriptional regulators ClgR and HspR as well as the ECF sigma factor sigmaH.** *Mol Microbiol* 2004, **52(1):**285-302.