Briefings in Bioinformatics Advance Access published December 12, 2008

Towards the integrated analysis, visualization and reconstruction of microbial gene regulatory networks

Jan Baumbach, Andreas Tauch and Sven Rahmann

Submitted: 26th April 2008; Received (in revised form): 6th November 2008

Abstract

To handle changing environmental surroundings and to manage unfavorable conditions, microbial organisms have evolved complex transcriptional regulatory networks. To comprehensively analyze these gene regulatory networks, several online available databases and analysis platforms have been implemented and established. In this article, we address the typical cycle of scientific knowledge exploration and integration in the area of procaryotic transcriptional gene regulation. We briefly review five popular, publicly available systems that support (i) the integration of existing knowledge, (ii) visualization capabilities and (iii) computer analysis to predict promising wet lab targets. We exemplify the benefits of such integrated data analysis platforms by means of four application cases exemplarily performed with the corynebacterial reference database CoryneRegNet.

Keywords: *data integration; systems biology; gene regulation; transcription factor, database*

INTRODUCTION

Microbial organisms 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]. Sensing of the surroundings and an appropriate response is triggered by complex molecular strategies coordinated by the transcriptional regulatory networks of the cell. The complexity of such regulatory networks results from the interaction of numerous transcription units consisting of a transcription factor (TF) and a set of controlled target genes [2]. The most important components of these units are apparently the DNA-binding TFs. They are responsible for sensing environmental and intracellular signals to control cellular reproduction and growth [3–5]. Depending on the growth conditions of a cell certain fractions of the total set of transcription units are operating [6]. Some TFs only control the expression of a single gene, whereas others organize the activation or repression of numerous target genes [2]. TFs contain DNA-binding domains that recognize the operator sequences of controlled target genes [7], referred to as TF binding motifs (TFBMs or shortly BMs).

The reconstruction of the complete transcriptional regulatory program of microorganisms, i.e. the identification of the spatial and temporal regulatory interactions between TFs and their target genes, is an important goal in molecular biology. The genome-wide analysis of gene regulatory networks from genomic data by bioinformatics strategies has

Corresponding author. Jan Baumbach, International Computer Science Institute, Berkeley, USA.

Jan Baumbach graduated in 2005 in Bioinformatics at Rothamsted Research in Harpenden (UK) and Bielefeld University. In his doctoral thesis he developed the microbial gene regulatory network platform CoryneRegNet at the Center for Biotechnology, Bielefeld. Currently, he works as postdoctoral fellow in the Algorithms group at the International Computer Science Institute, Berkeley.
Andreas Tauch is private lecturer at the Institute for Genome Research and Systems Biology of Bielefeld University, Germany. His area of research covers genome sequencing and postgenomics of corynebacteria and the reconstruction of their gene regulatory networks.
Sven Rahmann is professor for Bioinformatics for High-Throughput Technologies at TU Dortmund. Between 2004 and 2007, he was a group leader at Bielefeld University, focusing on transcriptional regulation networks. His doctoral thesis was about oligonucleotide design for microarrays and written at the Max Planck Institute for Molecular Genetics in Berlin.
E-mail: jbaumbac@icsi.berkeley.edu

been hindered, however, by the relatively low level of evolutionary conservation of their molecular components [8]. Comprehensive data sets on gene regulations deduced from bioinformatics and high-throughput experimental approaches provide the basis for the large-scale reconstruction of gene regulatory networks and evolutionary studies [9]. Although the monumental task of reconstructing gene regulatory networks for whole organisms is far from complete, the knowledge about substantial parts of the bacterial transcription regulation apparatus provides insights into conserved network structures, different regulatory modes, environmental sensing mechanisms of TFs and the interplay between various components and modules of the gene regulatory network, as demonstrated for the model organisms *Escherichia coli* [10]. In addition, specialized approaches to compare regulatory network structures and regulations between organisms or to transfer knowledge between networks of different organisms becomes feasible [5]. For instance, the known transcriptional regulatory network of *E. coli* was analyzed to detect the conservation patterns of this network across 175 prokaryotic genomes, and to predict components of the gene regulatory networks for these organisms. This resulted in the observation that organisms with similar lifestyles across a wide phylogenetic range tend to conserve equivalent interactions and network motifs. Also regulons of orthologous TFs can be compared between species to detect core regulatory interactions conserved across phylogenetic boundaries [11]. This integrated data helps to understand the complex pattern of differential gene expression in a microorganism that can be monitored on the genome scale by transcriptomic strategies [9].

To effectively and comprehensively analyze transcriptional regulatory networks, any kind of available and relevant data has to be combined. In this article, we address publicly available systems, which provide a user-oriented software platform that supports (i) the integration of existing knowledge, (ii) visualization capabilities and (iii) the *in silico* generation of novel hypotheses. To address these points, several approaches have been implemented and established. A list of five popular online available databases for microbial gene regulatory interactions is given in Table 1.

All the listed platforms follow a similar structure for data integration, analysis and reconstruction of microbial transcriptional regulations, depicted in Figure 1. First, external databases and publications (points 1 and 2) are preprocessed and imported into an integrative database (point 3). A web interface (point 4) provides all functionalities necessary to explore the database content. Coupled to the interface are integrated visualization and computer prediction capabilities (point 5). They can be used to generate new hypotheses (point 6) to provide promising targets for further wet lab studies (point 7).

In this review, we first present a list of features for the visualization of gene regulatory networks and the computational identification of promising wet lab targets. Subsequently, we briefly introduce the five popular platforms from Table 1 with respect to these features. In the last part of this article, we present four application cases exemplarily performed with CoryneRegNet to illustrate how these systems can be used to generate new hypotheses.

FEATURES

In order to support data integration, visualization and computer predictions, the following features are desirable and supported by most of the existing systems. In the application examples section, we will refer back to them and demonstrate their impact on transcriptional regulatory network analysis.

Web-based user interface: web-based user interfaces to biological databases should support the

Table I: Popular platforms for the storage and web-based analysis of microbial gene regulatory networks

Database	URL	Citations	TFs	TFBMs	PWMs
RegulonDB	http://regulondb.ccg.unam.mx	[56–62]	157	1491	66
MtbRegList	http://www.usherbrooke.ca/vers/mtbreglist	[63]	26	121	22[1]
PRODORIC	http://prodoric.tu-bs.de	[27, 64]	246	2517	182
DBTBS	http://dbtbs.hgc.jp	[65–67]	120	1309	45
CoryneRegNet	http://www.coryneregnet.de	[11, 44, 68, 69]	220	1549	134

Shown is the database name, the URL, publications about the database system, the number of known and characterized TFs, the number of annotated TFBMs, and the number of PWMs that have been computed from these TFBMs. [a]MtbRegList utilizes regular expression patterns instead of PWMs.

following standard tasks: (i) browsing by listing database entries, (ii) searching by restriction-based identification of database entries and (iii) visualizing by presenting a visual representation of the data [12].

Genome browser: a genome browser visualizes the genomic context of a gene of interest, ideally along with known sequence features (TFBMs, gene start/stop positions, etc.). Most online repositories provide such a feature and utilize an own implementation that is specialized for the respective database structure in the back-end.

Network visualization: considering genes as nodes and transcriptional interactions between genes as labeled, directed edges, interactions can be regarded as networks (graphs). These can be visualized by utilizing adequate graph layout algorithms.

TFBM prediction: the prediction of putative TFBMs allows (i) the identification of further gene regulatory interactions of a TF of interest, and thus (ii) the transfer of regulatory networks of a model organism to other closely related species. A variety of tools have been developed in the last years to address this problem. These tools model BMs of a TF as position weight matrices (PWMs [13,14]) and subsequently use them to scan the upstream sequences of putative target genes for matching sequence motifs.

Data exchange methods: usually, interconnections between different data sources are realized by HTML-links to other web pages or by regular, manual downloads and a subsequent integration of the corresponding data. This is both time-consuming and error-prone. By using SOAP-based web services [15], data can be exchanged in a well-structured manner, postprocessed more easily, and presented in a different way; but most importantly, the data are always up to date. Using web services, the end-user may not even recognize that the data are downloaded from another service.

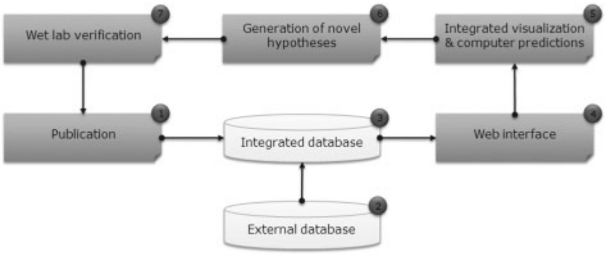


Figure 1: Typical cycle of computer-assisted data exploration, hypotheses generation and knowledge discovery. The typical starting point is a list of publications and one or more external databases. The integrated platform helps to provide promising targets for further wet lab studies.

Network and microarray analysis: bacterial gene regulatory networks tend to show a hierarchical structure that is mostly conserved between closely related species [3, 5, 16]. Hence, graph comparison capabilities (for both known and predicted networks) are a highly desirable feature. Such a function assists scientists with the cross-species knowledge transfer and hence with the identification of novel promising targets for wet lab analysis. The projection of gene expression levels (measured genome wide with DNA microarrays) onto graphs helps to gain a first overview of experimental data. Simple contradictions or inconsistencies in the context of known or predicted gene regulatory networks could become obvious within seconds or can be detected automatically on a large scale.

Homology detection: in order to provide graph comparison functionality at the front-end side, it is necessary to have implemented a mapping strategy for homologous proteins/genes. One way would be the integration of corresponding data from the COG [17] or the SCOP [18] databases. Unfortunately, both standard repositories omit data on many microbial organisms. Hence, the integration of a software for the detection of homologous proteins is desirable.

Promoter prediction: the prediction of transcription start sites (TSS) may help to identify whether a TF activates or represses a target gene.

Operon prediction: in procaryotes, many genes are grouped into transcription units (operons). Since an operon annotation is often missing, the integration of corresponding tools is desirable.

Statistical summaries: advanced statistical summaries and plots of the database content provide more general insights into microbial gene regulatory networks; such as the distribution of the TFBM lengths or connectivity plots, which illustrate the scale-free, hierarchical structure of these networks.

POPULAR PLATFORMS

All databases store predicted and experimentally validated TFBMs, along with genome annotations and operon organizations. Most data are extracted from literature and curated manually. The typical starting point are a list of publication abstracts, obtained from PubMed and filtered by pertinent keywords. The manually extracted data on transcriptional regulatory interactions are stored in a database management system and frequently updated and checked for inconsistencies. Here, we present a brief introduction to the most popular publicly available databases and analysis platforms for microbial gene regulatory interactions.

'RegulonDB' is an internationally recognized and established reference database for the procaryotic model organism *E. coli*. The amount of manually curated knowledge on the gene regulatory network of *E. coli* is the largest, currently available for any living organism. All data are gathered manually by the RegulonDB curation team. An online interface allows querying the database content. A visualization of the genomic context is offered, including several sequence features (genome browser). A network display tool allows a circular, graph-based visualization of the immediate neighbors of the gene of interest within the global regulatory network. RegulonDB integrates a homology detection [19] and offers a promoter detection tool [20] as well as an advanced statistical database summary. It does not provide structured data exchange methods.

'MtbRegList' is a database dedicated to the gene regulation of the human pathogenic bacterium *Mycobacterium tuberculosis*. The genome annotation has been obtained from the TubercuList [21] and NCBI [22] databases. As for RegulonDB, an online interface allows querying the database content. For all genes the web interface provides hyperlinks to corresponding COG database entries, where possible. Also, similar to RegulonDB, one can graphically navigate the genomic context, given a gene or a genetic region as a starting point (genome browser). Network visualization is not supported. A method to scan for TFBMs is provided: the user can enter a 'signature' (similar to a regular expression) to retrieve a list of matching DNA sites. MtbRegList does not provide data exchange methods.

'PRODORIC' aims at the storage and analysis of procaryotic gene regulations. The data structure used in the back-end is based on the TRANSFAC database [23–25], but is extended to specific procaryotic characteristics. PRODORIC includes all NCBI genome annotations of procaryotic organisms, even so no information on transcriptional regulations is available. Mainly *Bacillus subtilis*, *E. coli* and *Pseudomonas aeruginosa* are supported (more than 15 TFBMs). As the other systems, PRODORIC also provides a web interface for querying the database content, and moreover it allows to execute analysis tasks. PRODORIC supports links to COG and to SWISS-PROT [26]. PWMs constructed from stored TFBMs can be used as input for the integrated TFBM matching software Virtual Footprint [27] to predict further TF–DNA interactions *in silico*. As in RegulonDB, a genome browser provides a graphical representation of the genomic context at sequence level. A network visualization is supported for *B. subtilis*, *E. coli* and *P. aeruginosa* by using the ProdoNet [28] feature. With JVirGel [29,30] PRODORIC provides the visualization of virtual 2D protein maps and with PrediSi [31] the prediction of signal peptides. Furthermore, an automated functional analysis of sets of differentially expressed genes is supported with JProGO [32]. As the other systems, it does not provide data exchange methods.

'DBTBS' is the database of transcriptional regulation in the Gram-positive model organism *B. subtilis*. It is essentially a compilation of TFs along with their regulated genes and the recognition sequences (TFBMs). Annotated genes are linked to the BSORF database [32]. DBTBS also supports the prediction of putative TFBMs within a given input sequence by using PWMs and consensus patterns. Furthermore, DBTBS contributes to comparative genomics by detecting potentially orthologous TFs in other procaryotic genomes. It provides a genome browser and a terminator prediction [34], but no network visualization capabilities. It does not provide data exchange methods.

'CoryneRegNet' is the reference database and analysis platform for corynebacterial gene regulatory networks. Initially designed for the storage and analysis of transcriptional regulatory interactions of *Corynebacterium glutamicum*, it has been extended with data on *C. diphtheriae*, *C. efficiens*, *C. jeikeium*, and *E. coli*. PWM-based predictions of TF-DNA interactions are provided at the web front-end via the integrated PoSSuMsearch software [35,36] and MoRAine, a tool for the automatic re-adjustment of TFBMs [37]. A genome browser is supported as well as network visualization and analysis functionalities for all included organisms. Data exchange methods are offered by SOAP-based web services, which also link to the genome annotation system GenDB [38] and to the microarray analysis platform EMMA [39]. Via this inter-linked CoryneCenter [40], web links to several external databases are offered for each database entry, for instance to COG and NCBI. It also offers an interface to the Cytoscape biological graph analysis software [41]. Furthermore, CoryneRegNet contributes to comparative genomics by predicting homologous proteins in CoryneRegNet and in other databases by using the FORCE software [42,43]. Access to an advanced statistical database summary is offered through the web interface.

APPLICATION EXAMPLES

We show how the above mentioned platforms visualize their respective integrated knowledge and how they can be used to generate new hypotheses. Visualization comprises (i) standard table-based database content listings, (ii) statistical plots, (iii) specialized graphical data representations (such as sequence logos or a genome browser) and (iv) graph-based network visualization. The generation of novel insights is based either on conspicuous, eye-catching coherences uncovered by an appropriate visualization or by using computational models. In the second case, the most widely used computer prediction tools for the reconstruction of gene regulatory networks aim to identify TF binding sites and homologous genes (to provide further evidence when moving from one organism to another). In some cases, visualization capabilities may be combined with computer predictions to present a compact overview of the predicted data in comparison to the known data in order to deduce promising wet lab targets.

In the following, we present use cases exemplarily performed with the corynebacterial reference database CoryneRegNet. Here, we mainly rely on previously performed studies published in [11,40,41,44–46].

DtxR—the diphtheria toxin repressor

DtxR, the diphtheria toxin repressor of the human pathogen *C. diphtheriae*, is conserved in all sequenced corynebacteria. Over the last years, DtxR was subject to several genetic studies and the orthologous protein of *C. glutamicum* has been characterized. It regulates 64 genes in *C. glutamicum* with function in iron transport and utilization as well as in central carbohydrate metabolism and in transcriptional regulation [45]. The prediction of the regulatory network of DtxR of *C. diphtheriae* utilizing CoryneRegNet is reported in [44]. Based on experimentally verified DtxR binding sites of *C. glutamicum*, CoryneRegNet's TFBM prediction tool is used to calculate a PWM and to search for putative DtxR binding sites in the noncoding regions of the genome sequence of *C. diphtheriae*. Since the *C. diphtheriae* network is already known from [45], the approach can be evaluated by varying the *P*-value cutoff. By using a threshold of 10^{-6}, 7 of 32 known regulations (22%) have been found for *C. diphtheriae* with one false positive. A less restrictive cutoff (10^{-4}) yields 24 of 32 (75%) of the DtxR network in *C. diphtheriae* but produces 59 potential false positive candidates. The high number of false positives is caused by two problems: (i) true binding sites do not behave according to a simple probabilistic model (PWMs); (ii) the TFBMs from one species (*C. glutamicum*) have been used for predictions in another one (*C. diphtheriae*), where the TFBMs might be different. Hence, TFBM matching alone works only for taxonomically very closely related microorganisms, if at all.

LexA—the handling of DNA damage

The sequence logos in Figure 2 depict the price we have to pay when trying to move from one organism to a different one. It graphically illustrates the TFBMs of the DNA damage response regulator LexA of *C. glutamicum* and *E. coli*. Generally, the response of bacteria to DNA damage is an increased expression of a number of genes necessary for the cell's DNA repair machinery, which is negatively regulated by LexA. The TFBMs of LexA, termed SOS box, are conserved among taxonomic closely related bacterial species, but are different in distantly related microorganisms [47]. This affects the gene content of the LexA regulon, which varies among bacterial species. Supplementary Figure 1 shows the

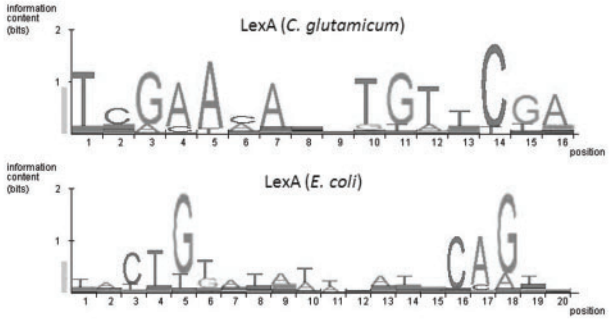


Figure 2: Sequence logos of the DNA binding site (SOS box) of the LexA protein generated by CoryneRegNet for *C. glutamicum* and *E. coli*. The stack heights indicate the information content at each position of the motif in bits, while the bar to the left of graphic indicates the mean information content (also refer to [11]).

small common set of six LexA-regulated genes in *C. glutamicum* and *E. coli* visualized with CoryneRegNet by using the homology-based comparative network layout feature.

Escherichia coli is taxonomically too distant from *C. glutamicum*, but for instance the multi-resistant human pathogen *C. jeikeium* is not. If we use the PWM matching feature of CoryneRegNet to transfer LexA TFBMs of *C. glutamicum* to *C. jeikeium* with a *P*-value cutoff of 10^{-6} a list of 16 putative target genes in *C. jeikeium* is presented (see Supplementary Figure 2). Although all of them are good candidates for further wet lab studies because of the restrictive threshold, further evidence is provided for five of them by the integrated homology detection tool. The genes *jk0181*, *jk1717*, *dnaE2*, *recA* and *modA* of *C. jeikeium* have been identified as putative homologs of *C. glutamicum* genes that are known to be under transcriptional control of LexA in *C. glutamicum*. The regulation of these five genes is a high potential new hypothesis and hence a good starting point for further wet lab verifications, since the combination of TFBM prediction and homology detection drastically decreases the amount of false positives.

GlxR—the first regulatory hub in *C. glutamicum*

Another example, recently published in [46], shows how the typical cycle of computer-assisted data exploration, hypotheses generation and wet lab verification, as illustrated in Figure 1, works. The goal was to unravel the network of the first regulatory hub in *C. glutamicum, GlxR*. The GlxR protein has been identified as cAMP-binding transcriptional regulator of the CRP/FNR protein family, showing homology to the global regulator CRP from *E. coli* [48]. In earlier studies it was shown that GlxR recognizes the promoter regions of five genes (see [46] for more details), which marked the starting point for the subsequent cycle of bioinformatics hypotheses generation and wet lab verification. Based on the sequence logo of the CRP TF of *E. coli*, the motif discovery tools MEME and Bioprospector [49,50] have been configured to detect five conserved 16-bp palindromic TFBMs in the corresponding promoter regions. After a wet lab verification of GlxR binding, the five TFBMs have been used for genome-wide bioinformatics detection of further GlxR binding sites in *C. glutamicum*. The *P*-value threshold has been set in such a way that all five initial TFBMs have been found. Thus, 31 putative TFBMs were identified, of which 28 were experimentally verified. By utilizing these TFBMs, another genome-wide TFBM prediction scan revealed 20 additional potential TFBMs, among which 18 were verified in the wet lab. By using predicted TSS, suggestions were made about whether GlxR activates or represses a certain target gene based on the relative position of the corresponding TFBM. Taken together, 46 out of 51 (90%) predicted TFBMs were experimentally validated. All together, the combined workflow of *in silico* hypotheses generation and wet lab evaluation provided knowledge on 51 (46 + 5) validated transcriptional regulatory interactions of GlxR in *C. glutamicum*.

Glucose versus acetate feeding conditions

The next step in visualizing and analyzing gene regulatory networks is to integrate gene expression data with known and/or predicted networks. One example was published in [40]: The comparison of the transcriptome of acetate-grown *C. glutamicum* cells to that of glucose-grown cells, using microarray hybridization results. The expression data was stored in the microarray analysis platform EMMA [39] and imported into the graph visualization tool of CoryneRegNet by utilizing SOAP-based web services (see [40] for technical details). The combined visualization (Supplementary Figure 3) of the known gene regulatory networks of RamA, RamB and GlxR together with the gene expression data immediately provides the following results: (i) the RamA network shows a consistent answer to the stimulus; (ii) interestingly, the transcription level of the RamA target gene *ramB* was unaffected in the experiment; and (iii) in addition, transcript levels of most RamB target genes are not significantly altered, which is in accordance to the unchanged *ramB* transcription. However, the decreased transcript levels of the RamB target genes *ptsS* and *ptsG* could not be explained and suggest additional regulatory interactions active under acetate or glucose feeding. One potential candidate is the TF SugR [51], which is slightly overexpressed and known to repress the transcription of *pts* genes in the absence of sugarphosphates. This regulation was recently verified experimentally [52]. Similar consistency checks can be performed automatically and results are presented in a table-based style (see [41] for technical details).

Having all relevant data integrated with appropriate bioinformatics tools allows to perform the above introduced use cases with just a few mouse clicks. Besides these special tasks, integrated platforms may also help to address general questions regarding the nature of gene regulatory networks. For instance, the distribution of the number of TFs regulating a gene (Supplementary Figure 4) suggests a scale-free network structure, while e.g. the distribution of PWM lengths suggest that TFBMs typically are between 12 bp and 22 bp in length.

CONCLUSION AND OUTLOOK

In this article, we briefly introduced five online platforms dedicated to the integration, visualization and analysis of gene regulatory networks of microbes: RegulonDB, MtbRegList, PRODORIC, DBTBS and CoryneRegNet. We described 10 main features. Most of them should be addressed adequately in order to present and exchange transcriptional regulatory data in a way to efficiently generate new hypotheses. We illuminated the impact of an appropriate integration of bioinformatics tool by means of four application examples. All of them have been exemplarily performed with the CoryneRegNet platform.

We conclude that most of the necessary features are provided by all of the mentioned platforms and can be utilized to analyze the stored data. One exception is the lack for well-structured data exchange methods. Offering, for instance, SOAP-based web services that would highly facilitate data integration by external bioinformatics platforms and tools. One example was demonstrated in the feeding conditions use case. At the front-end level, the integration of TFBM discovery tools could be beneficial (refer to the GlxR application case). Such a feature is not supported by any platform. Given a set of co-regulated genes in an expression study, it could help to find overrepresented motifs in the upstream sequences of the corresponding genes; see [53,54] for a review of sequence motif discovery tools.

However, the impact and usability of an integrated platform is limited by the quality and quantity of the included data itself. Most of the data to be processed are scattered over numerous publications, most of them stored in the PubMed library. These are structured and curated manually for inclusion in each special-purpose database, which is both time-consuming and error-prone. One way to reduce the necessary amount of manual intervention is to encourage authors of research papers to submit the essence of their results in a formal language. This can be parsed and checked automatically and corresponding data may be stored in a specialized central repository, e.g. one for gene regulatory interactions; just as it is now a generally accepted requirement to submit DNA sequences to GenBank (or alternative databases) before publication. While such a development would certainly ease automatic data integration, it would be a highly political venture to establish standards for such repositories. It may actually impede highly creative research, such as the development of new theories [55]. At present, it therefore appears that integrative systems require a certain amount of manual curation to ensure high quality.

Key Points

- Systems dedicated to the storage, analysis and reconstruction of microbial gene regulatory interactions provide a user-oriented software platform that supports (i) the integration of existing knowledge, (ii) visualization capabilities and (iii) the computational generation of new hypotheses.
- We consider the following features as necessary: (i) web-based user interface, (ii) genome browser, (iii) network visualization capabilities, (iv) online TFBM prediction, (v) well-structured data exchange methods, (vi) network and microarray analysis, (vii) comparative genomics/homology detection, (viii) promoter prediction, (ix) operon prediction and (x) advanced statistical database summaries.
- Popular examples of platforms for microbial gene regulatory networks are RegulonDB (*EC*), MtbRegList (*M. tuberculosis*), PRODORIC (*B. subtilis*, *EC* and *P. aeruginosa*), DBTBS (*B. subtilis*), and CoryneRegNet (corynebacteria).

Acknowledgments

We wish to thank Kenta Nakai (University of Tokyo) and Richard Munch (TU Braunschweig, Germany) for providing statistical data on DBTBS and PRODORIC. J.B. wishes to thank the German Academic Exchange Service for funding.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

References

1. Matic I, Taddei F, Radman M. Survival versus maintenance of genetic stability: a conflict of priorities during stress. *Res Microbiol* 2004;**155**:337–41.
2. Teichmann SA, Babu MM. Gene regulatory network growth by duplication. *Nat Genet* 2004;**36**:492–6.

3. Babu MM, Luscombe NM, Aravind L, *et al*. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* 2004;**14**:283–91.
4. Babu MM, Teichmann SA. Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res* 2003;**31**:1234–44.
5. Babu MM, Teichmann SA, Aravind L. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* 2006;**358**:614–33.
6. Resendis-Antonio O, Freyre-Gonzalez JA, Menchaca-Mendez R, *et al*. Modular analysis of the transcriptional regulatory network of E. coli. *Trends Genet* 2005;**21**:16–20.
7. Pabo CO, Sauer RT. Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem* 1992;**61**:1053–95.
8. Herrgard MJ, Covert MW, Palsson B. Reconstruction of microbial transcriptional regulatory networks. *Curr Opin Biotechnol* 2004;**15**:70–7.
9. Balaji S, Iyer LM, Babu MM, Aravind L. Comparison of transcription regulatory interactions inferred from high-throughput methods: what do they reveal? *Trends Genet* 2008;**24**:319–23.
10. Balaji S, Babu MM, Aravind L. Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of E. coli. *J Mol Biol* 2007;**372**:1108–22.
11. Baumbach J, Wittkop T, Rademacher K, *et al*. CoryneRegNet 3.0-an interactive systems biology platform for the analysis of gene regulatory networks in corynebacteria and Escherichia coli. *J Biotechnol* 2007;**129**: 279–89.
12. Garwood K, McLaughlin T, Garwood C, *et al*. PEDRo: a database for storing, searching and disseminating experimental proteomics data. *BMC Genomics* 2004;**5**:68.
13. Rahmann S, Mueller T, Vingron M. On the power of profiles for transcription factor binding site detection. *Stat Appl Genet Mol Biol* 2003;**2**.
14. Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics* 2000;**16**:16–23.
15. Curbera F, Duftler M, Khalaf R, *et al*. Unraveling the Web Services Web: an introduction to SOAP, WSDL, and UDDI. *IEEE Internet Computing* 2002;**6**:86–93.
16. Lozada-Chavez I, Janga SC, Collado-Vides J. Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res* 2006;**34**:3434–45.
17. Tatusov RL, Fedorova ND, Jackson JD, *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003;**4**:41.
18. Andreeva A, Howorth D, Brenner SE, *et al*. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 2004;**32**:D226–9.
19. Janga SC, Collado-Vides J, Moreno-Hagelsieb G. Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. *Nucleic Acids Res* 2005;**33**:2521–30.
20. Huerta AM, Collado-Vides J, Francino MP, Investigators SMBETNY. Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. Positional conservation of clusters of overlapping promoter-like sequences in enterobacterial genomes. *Mol Biol Evol* 2006;**23**: 997–1010.
21. Camus JC, Pryor MJ, Médigue C, Cole ST. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology* 2002;**148**(**Pt 10**):2967–73.
22. Wheeler DL, Barrett T, Benson DA, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2007;**35**:D5–12.
23. Matys V, Kel-Margoulis OV, Fricke E, *et al*. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006;**34**: D108–10.
24. Wingender E. TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In Silico Biol* 2004;**4**:55–61.
25. Wingender E, Chen X, Fricke E, *et al*. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res* 2001;**29**:281–3.
26. Bairoch A, Apweiler R, Wu CH, *et al*. The Universal Protein Resource (UniProt). *Nucleic Acids Res* 2005;**33**: D154–9.
27. Münch R, Hiller K, Grote A, *et al*. Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics* 2005;**21**:4187–9.
28. Klein J, Leupold S, Münch R, *et al*. ProdoNet: identification and visualization of prokaryotic gene regulatory and metabolic networks. *Nucleic Acids Res* 2008;**36**:W460–4.
29. Hiller K, Grote A, Maneck M, *et al*. JVirGel 2.0: computational prediction of proteomes separated via two-dimensional gel electrophoresis under consideration of membrane and secreted proteins. *Bioinformatics* 2006;**22**: 2441–3.
30. Hiller K, Schobert M, Hundertmark C, *et al*. JVirGel: Calculation of virtual two-dimensional protein gels. *Nucleic Acids Res* 2003;**31**:3862–5.
31. Hiller K, Grote A, Scheer M, *et al*. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res* 2004;**32**:W375–9.
32. Scheer M, Klawonn F, Munch R, *et al*. JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using Gene Ontology information. *Nucleic Acids Res* 2006;**34**:W510–5.
33. Fawcett P, Eichenberger P, Losick R, Youngman P. The transcriptional profile of early to middle sporulation in Bacillus subtilis. *Proc Natl Acad Sci USA* 2000;**97**:8063–8.
34. de Hoon MJL, Makita Y, Nakai K, Miyano S. Prediction of transcriptional terminators in Bacillus subtilis and related species. *PLoS Comput Biol* 2005;**1**:e25.
35. Beckstette M, Homann R, Giegerich R, Kurtz S. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* 2006; **7**:389.
36. Beckstette M, Strothmann D, Homann R, Giegerich R, Kurtz S. PoSSuMsearch: fast and sensitive matching of position specific scoring matrices using enhanced suffix arrays. *GI Lecture Notes Inform* 2004;**P-53**:53–64.
37. Baumbach J, Wittkop T, Weile J, *et al*. MoRAine – a web server for fast computational transcription factor binding motif reannotation. *J Integr Bioinform* 2008;**5**:91.
38. Meyer F, Goesmann A, McHardy AC, *et al*. GenDB-an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res* 2003;**31**:2187–95.

39. Dondrup M, Goesmann A, Bartels D, *et al*. EMMA: a platform for consistent storage and efficient analysis of microarray data. *J Biotechnol* 2003;**106**:135–46.

40. Neuweger H, Baumbach J, Albaum S, *et al*. CoryneCenter - an online resource for the integrated analysis of corynebacterial genome and transcriptome data. *BMC Syst Biol* 2007;**1**:55.

41. Baumbach J, Apeltsin L. Linking Cytoscape and the corynebacterial reference database CoryneRegNet. *BMC Genomics* 2008;**9**:184.

42. Rahmann S, Wittkop T, Baumbach J, *et al*. Exact and heuristic algorithms for weighted cluster editing. *Comput Syst Bioinformatics Conf* 2007;**6**:391–401.

43. Wittkop T, Baumbach J, Lobo F, Rahmann S. Large scale clustering of protein sequences with FORCE -A layout based heuristic for weighted cluster editing. *BMC Bioinformatics* 2007;**8**:396.

44. Baumbach J, Brinkrolf K, Wittkop T, *et al*. CoryneRegNet 2: An integrative bioinformatics approach for reconstruction and comparison of transcriptional regulatory networks in prokaryotes. *J Integr Bioinform* 2006;**3**:24.

45. Brune I, Werner H, Huser A, *et al*. The DtxR protein acting as dual transcriptional regulator directs a global regulatory network involved in iron metabolism of Corynebacterium glutamicum. *BMC Genomics* 2006;**7**:21.

46. Kohl TA, Baumbach J, Jungwirth B, *et al*. The GlxR regulon of the amino acid producer Corynebacterium glutamicum: in silico and in vitro detection of DNA binding sites of a global transcription regulator. *J Biotechnol* 2008;**135**:340–50.

47. Mazon G, Erill I, Campoy S, *et al*. Reconstruction of the evolutionary history of the LexA-binding sequence. *Microbiology* 2004;**150**(**Pt 11**):3783–95.

48. Kim HJ, Kim TH, Kim Y, Lee HS. Identification and characterization of glxR, a gene involved in regulation of glyoxylate bypass in Corynebacterium glutamicum. *J Bacteriol* 2004;**186**:3453–60.

49. Bailey TL, Williams N, Misleh C, *et al*. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* 2006;**l34**:W369–73.

50. Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* 2001;**6**:127–38.

51. Engels V, Wendisch VF. The DeoR-type regulator SugR represses expression of ptsG in Corynebacterium glutamicum. *J Bacteriol* 2007;**189**:2955–66.

52. Gaigalat L, Schlüter JP, Hartmann M, *et al*. The DeoR-type transcriptional regulator SugR acts as a repressor for genes encoding the phosphoenolpyruvate:sugar phospho-transferase system (PTS) in Corynebacterium glutamicum. *BMC Mol Biol* 2007;**8**:104.

53. Li N, Tompa M. Analysis of computational approaches for motif discovery. *Algorithms Mol Biol* 2006;**1**:8.

54. Tompa M, Li N, Bailey TL, *et al*. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005;**23**:137–44.

55. Philippi S, Köhler J. Addressing the problems with life-science databases for traditional uses and systems biology. *Nat Rev Genet* 2006;**7**:482–8.

56. Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, *et al*. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* 2008;**36**:D120–4.

57. Huerta AM, Salgado H, Thieffry D, Collado-Vides J. RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res* 1998;**26**:55–9.

58. Salgado H, Gama-Castro S, Martinez-Antonio A, *et al*. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Res* 2004;**32**:D303–6.

59. Salgado H, Gama-Castro S, Peralta-Gil M, *et al*. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 2006;**34**:D394–7.

60. Salgado H, Santos A, Garza-Ramos U, *et al*. RegulonDB (version 2.0): a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res* 1999;**27**:59–60.

61. Salgado H, Santos-Zavaleta A, Gama-Castro S, *et al*. RegulonDB (version 3.0): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Res* 2000;**28**:65–7.

62. Salgado H, Santos-Zavaleta A, Gama-Castro S, *et al*. RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Res* 2001;**29**:72–4.

63. Jacques PE, Gervais AL, Cantin M, *et al*. MtbRegList, a database dedicated to the analysis of transcriptional regulation in Mycobacterium tuberculosis. *Bioinformatics* 2005;**21**: 2563–5.

64. Muench R, Hiller K, Barg H, *et al*. PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res* 2003;**31**:266–9.

65. Ishii T, Yoshida K, Terai G, *et al*. DBTBS: a database of Bacillus subtilis promoters and transcription factors. *Nucleic Acids Res* 2001;**29**:278–80.

66. Makita Y, Nakao M, Ogasawara N, Nakai K. DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Res* 2004;**32**:D75–7.

67. Sierro N, Makita Y, de Hoon M, Nakai K. DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res* 2008;**36**:D93–6.

68. Baumbach J. CoryneRegNet 4.0 - A reference database for corynebacterial gene regulatory networks. *BMC Bioinformatics* 2007;**8**:429.

69. Baumbach J, Brinkrolf K, Czaja L, *et al*. CoryneRegNet: an ontology-based data warehouse of corynebacterial transcription factors and regulatory networks. *BMC Genomics* 2006;**7**:24.

Integrated analysis and reconstruction of microbial transcriptional gene regulatory networks using CoryneRegNet

Jan Baumbach[1], Tobias Wittkop[1,2,3], Christiane Katja Kleindt[2,4], Andreas Tauch[5]

[1] Algorithms Group, International Computer Science Institute, Berkeley, CA 94704, USA

[2] International Graduate School in Bioinformatics and Genome Research, Bielefeld University, D-33594 Bielefeld, Germany

[3] Graduiertenkolleg Bioinformatik, Bielefeld University, D-33594 Bielefeld, Germany

[4] Department of Plant and Microbial Biology, University of California, Berkeley, CA 94704, USA

[5] Institute for Genome Research and Systems Biology, Center for Biotechnology, Bielefeld University, D-33594 Bielefeld, Germany

ABSTRACT

CoryneRegNet is the reference database and analysis platform for corynebacterial gene regulatory networks. It provides web-based access to integrated data on gene regulatory interactions of: corynebacteria relevant to human medicine and biotechnology; *Escherichia coli*; and *Mycobacterium tuberculosis*. To facilitate the analysis and reconstruction of the corresponding networks, CoryneRegNet provides user-friendly interfaces for bioinformatics analysis and network visualization tools. This protocol describes four major workflows: (1) querying the regulatory network of a gene of interest, (2) prediction and inter-species transfer of gene regulatory interactions, (3) visualization and comparison of predicted or known networks, and (4) integration of gene expression data analysis and visualization. The presented protocol guides the user through the most important features of CoryneRegNet and takes 45-60 minutes to complete.

INTRODUCTION

Sensing and handling changing environmental conditions is crucial for any living organism. Microorganisms have evolved complex molecular strategies to adapt to their surroundings and to maintain growth, survival, and reproduction. The most important elements of these mechanisms are DNA-binding proteins called transcription factors (TFs). They recognize and bind to certain sequences within the DNA double helix, the transcription factor binding sites (TFBSs), and thereby regulate gene transcription. Complex regulatory networks emerge. Depending on the growth and survival conditions of a cell, certain fractions of the total set of transcription units are operating. While some TFs control the transcription of only a single gene, others regulate the activation or repression of numerous target genes and may represent regulatory hubs within the network structure[1-5].

An important goal in Systems Biology is the reconstruction of the complete regulatory network of microorganisms. Current research focuses on the identification and analysis of the regulatory interaction between TFs and the target genes they control. The resulting gene regulatory networks may be transformed to computational models, which allow an *in silico* study of the cell's behavior in response to external stimuli. However, bioinformatics reconstruction of these networks is hindered mainly by the relatively low level of evolutionary conservation of their molecular components, in particular of the TFBSs. Nevertheless, comprehensive data sets on transcriptional gene control deduced from high-throughput wet-lab experiments and bioinformatics prediction strategies provide the basis for an integrated large-scale reconstruction of gene regulatory networks[6-9].

Although required for the understanding of the microorganism's transcriptional regulatory response to changing environmental conditions, the wet-lab reconstruction of the resulting networks is cost-intensive, time-consuming, and thus impossible to perform for any species separately. Recent advances in ultra-fast sequencing technologies will provide more data on fully-sequenced organisms that has to be analyzed in the context of transcriptional gene regulation. The monumental task of deciphering whole species transcriptional regulatory networks is far from being complete, even for bacterial model organisms, such as *Escherichia coli*[10] (Gram-negative model) or *Corynebacterium glutamicum*[11] (high GC Gram-positive model).

The current knowledge is brought together and stored in reference databases. It is subsequently combined with bioinformatics analysis, prediction, and visualization features. A number of these online platforms are publicly available for integrated microbial transcriptional gene regulatory network investigation and reconstruction. Popular systems are CoryneRegNet[12-15] (corynebacteria, *M. tuberculosis*, and *E. coli*), MtbRegList[16] (*M. tuberculosis*), PRODORIC[17,18] (mainly *Pseudomonas aeruginosa, E. coli, and Bacillus subtilis*), DBTBS[19-21] (*B. subtilis*), and RegulonDB[22-28] (*E. coli*); refer to Table 1. A review may be found in ref. [8].

In the past years, we extensively studied the transcriptional regulatory repertoire of the model organism *C. glutamicum* (used world-wide for industrial amino acid and food additive production) and other closely related organisms important in human medicine and biotechnology: the human pathogens *Corynebacterium diphtheriae*[29] (the etiological agent of diphtheria) and *Corynebacterium jeikeium*[30] (multiple antibiotic resistant pathogen), and *Corynebacterium efficiens*[31] (used for biotechnological production processes). We gathered all publicly available data, combined it with our wet-lab findings and computational predictions, and developed the reference database and analysis platform CoryneRegNet, which can be accessed at http://www.coryneregnet.de. Further data can be included into CoryneRegNet upon request. Subsequently, we also integrated data on the model organism *E. coli* and the human pathogen *M. tuberculosis*[32].

CoryneRegNet is a user-oriented software platform that supports functionalities for the following major tasks: (1) integration of existing knowledge, (2) graph-based network visualization, and (3) the *in silico* generation of novel hypotheses.

Recently, we analyzed and compared methods that are necessary to perform the above mentioned aims in an integrated manner[8]. Most important is successful data integration[33]. Based on that, the integrated networks can be visualized and investigated together with whole-genome expression data. The most prominent computational challenges for the *in* *silico* generation of novel hypothesis are (1) TFBS prediction and (2) homology detection. For the first task, known TFBSs are most widely modeled as so-called position weight matrices[34] and subsequently used for genome-wide scans. In this protocol, we used PoSSuMsearch[35] since it is fast enough to provide exact p-value computations and reasonable response times; other tools may be found elsewhere[18,36]. The long-standing problem of computational remote homology detection based on sequence information is performed by means of clustering approaches. Here, we use the weighted graph cluster editing software FORCE[37,38], which outperforms other typical approaches for protein homology identification, such as TribeMCL[39], Spectral Clustering[40], or Affinity Propagation[41].

In this protocol paper, we introduce the user interface of CoryneRegNet from an end-user point of view. We guide users through four case studies that illustrate how to deal with common questions that biologists typically address when using CoryneRegNet to generate novel hypotheses and promising wet-lab targets.

Structure of the CoryneRegNet user interface

The web front-end of CoryneRegNet is hierarchical in its organization (also refer to Figure 1):

Query: The query interface of CoryneRegNet offers five main options to start gene regulatory network investigation:

1. A basic summary of the database content and a link to special purpose statistical network analyses
2. A list of integrated organisms as a starting point for database browsing
3. A typical search mask for query specification
4. A link to the CoMa feature, which checks for inconsistencies in gene expression data, given the known gene regulatory networks stored in the back-end
5. A link to the TFBScan feature providing an interface to the integrated transcription factor binding site prediction

The structure of CoryneRegNet allows for browsing the database content at every step of an ongoing analysis by clicking on an organism or gene of interest.

Investigation and network transfer: The details page of a selected gene is organized in a table-based style. It is structured in separate modules providing preprocessed graphical, statistical, and table-based access to the stored information for a gene of interest (see Figure 2 for an example, and Box 1 for a detailed description of the modules). From this page the user has two options for accessing the integrated binding site prediction feature TFBScan: (1) The gene of interest encodes for a transcriptional regulator with known TFBSs. These may be used to scan for further TFBSs in the promoter regions in the genes of all organisms. (2) The binding motifs of all TFs in the database may be used to search for TFBSs within the upstream sequence of the investigated gene. Note that this feature, together with integrated homology detection, may be utilized for cross-species transfer of gene regulatory networks (see Box 2 and option C in the protocol). For a more intuitive, graphical analysis of the regulatory network of the gene of interest, the details page links directly to the graph visualization feature GraphVis.

Network visualization: The integrated network visualization GraphVis provides various options for graph based analyses and visualization of the gene regulatory network of a specific gene or of whole organisms. Various graphical layouts are available for the displayed network. Besides using a gene of interest as starting point for an investigation, it is also possible to use GraphVis for the visualization of predicted networks. These visualizations may be constructed after carrying out binding site prediction, described above. Cross-species comparison can be performed by using the integrated comparative network layouts. GraphVis further allows for visual analysis of gene expression data. A detailed description of the features of GraphVis can be found in Box 3.

Gene expression data analysis: Once a gene regulatory network is displayed as a graph in GraphVis, the user may query microarray data directly from the stimulon database of CoryneRegNet, via the integrated Web Service to EMMA[42], or by uploading their own data files. This expression data is subsequently projected onto the graph by altering the node sizes according to the relative differential expression levels of the corresponding genes. The user can graphically investigate the effect and impact of a certain stimulus to the underlying regulatory network. An inconsistent network response may be seen easily, indicating hitherto unknown regulatory interactions.

An automatic table-based analysis of the same type can be performed with the CoMa feature, which can be accessed from the query interface. The user input is a gene expression study and the output is a list of contradictory regulatory sub-network responses (option E in the protocol).

Overview of the Procedure

Options A and B - Gene regulatory network and gene expression data visualization: The general aim here is: Given a gene of interest, show all available information for this gene, visualize its regulatory network as a graph together with gene expression data, and compare it to the network of a homologous gene of a closely related species.

Option A and B of the protocol guide users through the basic analysis of the gene regulatory network of a specific gene using the global iron-dependent repressor gene *dtxR*[43] as an example. CoryneRegNet provides a table-based overview of the database content for *dtxR*. A subsequent graphical visualization of the network provides a visual overview, extended by a projection of gene expression data, to unravel regulatory network response to a stimulus. Putative contradictions to known regulations, such as differentially expressed genes within one operon, may become obvious. Another example of a conflict that may occur is the up-regulation of a repressor and the up-regulation of its target gene. These effects can be studied and potentially explained by extending the network's depth. Furthermore, the *dtxR* network of *C. glutamicum* can be visually compared to that of a closely related species (here we use *C. efficiens* as an example) by using comparative graph layout algorithms. Thus, the focus of this section is to introduce the user to the web interface and the network visualization tool GraphVis.

Option C - Inter-species transfer of regulatory networks: The general aim here is: Given a transcription factor with known binding sites, predict further gene regulatory interactions in a closely related species, visualize this network as graph, and compare it with the known network of the reference organism.

Transfer of knowledge from a model organism to closely related species is a common way to gain information about organisms that cannot be analyzed efficiently with wet lab experiments[8] - such as *Corynebacterium jeikeium*, which is highly resistant to numerous antibiotics, and has been recognized as a serious nosocomial pathogen[30]. Knock-out experiments on multidrug-resistant organisms for gene expression studies are hard to perform in a wet lab, since they often rely on marker genes that introduce an additional antibiotic resistance into the organism. Nevertheless it is important to gain knowledge about the gene regulatory networks of such human pathogens.

For cross-species network transfers, CoryneRegNet utilizes the integrated binding motif search tool TFBScan combined with a graphical visualization of the obtained results. Since potentially conserved regulations may be targets for further wet lab experiments, this kind of analysis helps to reduce the number of experiments. Subsequent filtering of the results based on homology between the target genes of the reference and the target organism results in high-probability candidates for accurately predicted interactions.

In option C, we describe such an analysis by studying the gene *glxR* of *C. glutamicum*, a regulatory hub, which controls at least 150 genes directly. After predicting putative binding sites in the promoter regions of genes in *C. jeikeium*, the predicted network will be visualized and analyzed with GraphVis. The integrated homology graph layout emphasizes the similarities between the original (*C. glutamicum*) and the predicted (*C. jeikeium*) network.

<u>Option D - Transcription factor binding site prediction:</u> The general aim here is: Given a transcription factor that is not characterized so far and hence not stored in the database, but some of its binding sites are known. Predict further candidates for binding sites in the same organism to generate new hypothesis for wet-lab studies.

TFBS search is crucial for cross-species network transfer but also for the prediction of further gene regulatory interactions within the same organism. Therefore CoryneRegNet provides an easy-to-use web interface (TFBScan) to the integrated TFBS prediction software PoSSuMsearch. TFBScan allows the user to upload either upstream regions or known binding motifs. See Box 2 for further details.

To improve the TFBS prediction quality, CoryneRegNet links to the motif re-assessment software MoRAine. Often, binding motifs are annotated routinely 5' -> 3' relative to the target gene, which may lead to suboptimal position weight matrices (PWMs) that are subsequently used for the TFBScan feature. MoRAine can automatically identify these cases and provide optimized binding motifs by shifting a site by a few positions or switching its strand if beneficial[44].

Option D of the protocol describes an example workflow through the integrated TFBS prediction framework.

<u>Option E - Inconsistencies in gene expression studies:</u> Here, the general aim is: Given a gene expression experiment, show all those regulations that contradict with the known gene regulatory network of the corresponding organism. These regulations are targets for further wet lab studies since there must be a trigger for the unexpected response of the network to the stimulus, i.e. a gene regulatory interaction that controls the respective target gene is unknown and may be identified in the wet-lab.

The analysis of gene expression studies is the key to the study of gene regulatory networks. One crucial step is to validate the results against the knowledge base. The integrated feature CoMa automatically searches for contradictions given a known regulatory network and an expression profile[12]. All genes involved in the corresponding experiment will be separated into three groups: (1) genes with putative regulatory contradictions, (2) with no regulatory contradiction, and (3) those with unknown regulations. Similarly, all operons are split into two groups: with and without contradictions. CoMa allows the user to upload their own gene expression files, ‘copy and paste’ them directly in a text field, or to use the CoryneRegNet stimulon database. Option E introduces the user to the integrated application of CoMa.

MATERIALS

EQUIPMENT

Web browser: For CoryneRegNet we recommend to use at least Mozilla FireFox 2.0 or MS Internet Explorer 6.0 under Linux, SunOS, or Windows (XP, ME, 2000, or VISTA).

Screen resolution: Although not necessary, we further recommend a screen resolution of at least 1024x768.

Java Script: For the details pages of the genes, Java Script is used for the modules. It has to be enabled with the web browser.

GraphVis and Java: The graph visualization feature GraphVis is a Java Applet. At least Java version 1.6 has to be installed and configured properly for usage with the web browser. Java is publicly available for free download at http://www.java.com. Ensure to have Java enabled in the web browser.

PROCEDURE

1| Start the web browser and navigate to http://www.coryneregnet.de. To enter the CoryneRegNet online interface, click on “Experimental” or “Predicted” to choose the evidence level. The experimental version of CoryneRegNet only contains gene regulations with experimental evidence. In the other version, we additionally store predicted gene regulatory interactions without wet lab proof[45]. For this protocol, click on “Experimental”.

2| In the following choose between five options: study of gene expression profiles (option A); comparative study of networks (option B); an interspecies network transfer (option C); prediction of TFBSs (option D); identification of inconsistencies in gene expression data (option E).

(A) Combined visualization of gene regulatory networks and gene expression data (TIMING ~ 10 min)

(i) To search for a specific gene, enter the gene name in the “Search” field. For this particular protocol, enter “*dtxR”* and chose “Gene name” in the next dropdown menu (“in field”).

(ii) Select to search either all organisms for the gene of interest (default) or a specific organism. *dtxR* exists only in corynebacteria, therefore use default settings and start the search by pressing “Start search”. The *dtxR* query results in four database entries for corynebacteria*,* which can be seen by scrolling up and down the page. The results page displays the following information for each entry: gene ID, alternative gene ID, gene name, protein ID, protein name, regulator type (if applicable), predicted operon (if any), organism, functional module, and regulated target genes (if any). The column “Regulation” provides an overview of all genes which are regulated by DtxR in the four corynebacterial organisms.

TROUBLESHOOTING

(iii) To obtain detailed information for the gene of interest (here that of *Corynebacterium glutamicum* ATCC 13032), click on the relevant entry (here "*cg2103"*) in the "Gene ID" column. CoryneRegNet loads the details page of *dtxR.* Various modules provide preprocessed tabular and graphical details for your gene of interest, for instance a genome browser.

(iv) Click "+"/"-" to expand each module to receive more detailed information, such as links to external platforms. For a description of the details page see Box 1 and Figure 2.

(v) Scroll to the bottom of the web page and locate the "GraphVis" settings. In this example we don't want to include genes from regulations, so uncheck the respective box and use the default "depth cut-off" of "1".

(vi) To obtain a graph-based visualization of the gene regulatory network of the DtxR transcription factor, start GraphVis by clicking on the "GraphVis" button. You will be asked if you wish to import homology information for the genes within the graph-representation. For this particular protocol choose "No". The GraphVis window appears and displays two panels. The "Details" panel at the top provides the user with details about selected genes, a legend to explain symbols, and a help page. The panel below contains the graph-representation of the queried network. Here the user may change the layout of the graph, extend it, import homologies, or import and study gene expression data (for a detailed description of all GraphVis features see Box 3). CRITICAL STEP: The browser may ask you to accept a certificate for the signed Applet; accept it.

TROUBLESHOOTING

(vii) To study gene expression data, zoom into the network view and select the part of the regulatory network of interest (use the "Shift" key for multiple selections or keep the first mouse button pressed for "rectangular selection"). In this example, select the whole regulatory network. Afterwards, go to "Tools" → "Gene expression data" → "Import expression data from database (stimulons)". Press "OK" to request stimulon data for all selected genes.

(viii) This will open the import wizard. To select a stimulon, click on one entry at the left site. Choose "delta_dtxr" for this protocol. All genes stimulated by the respective stimulon are shown at the right side of the window. For each gene, the gene ID and the gene name are shown, as is the M-value and if the gene was previously selected in the graph. Now select the genes that you are interested in from this table (use the "Shift" or "Ctrl" keys for multiple selections). Here, select all displayed genes, by clicking on any position in the right table and pressing "Ctrl" + "A".

(ix) Before proceeding with the next step, the user may decide between "Import genes that are stimulated but not in the selected part of the graph" and "Just consider previously selected genes" at the bottom of the page. In this protocol we don't want to import all stimulated genes, therefore select "Just consider previously selected genes" and click "OK". The results are presented in a table and GraphVis projects the gene expression data onto the displayed network as illustrated in Figure 3a. A closer inspection of the graph provides an example of different regulation types. Red arrows indicate repressions, while green arrows represent activations. Circles symbolize regulated target genes, which are preceded by a transcription factor binding site. Boxes indicate regulated target genes that are part of an operon and not preceded by a transcription factor binding site. Circles or boxes surrounded by a red dotted line denote down-stimulated genes, while a green dotted border refers to up-stimulated genes. Descriptions of the symbols and colors can be reviewed, by clicking at the "Legend" tag at the very top of the "Details" window (also see Box 3).

(x) To go on with additional studies remove the expression data from the displayed graph. Go to "Tools" → "Gene expression data" → "Clear expression data from graph (reset)".

(B) Comparative visualization of gene regulatory networks (TIMING ~ 10 min)

(i) In order to compare two networks they have to be visualized first. Therefore, carry out the steps (i) - (vi) of option A.

(ii) To compare two regulatory networks additional data has to be imported. Here, the aim is to compare the DtxR network of *C. glutamicum* to that of the homologous regulator in *C. efficiens*. Extend the network with *dtxR* data from *C. efficiens*: Go to "Tools" → "Extend the graph" → "Extend graph using wizard".

(iii) The import wizard provides a search field on the upper left side. Type in a gene ID, alternative gene ID, gene name, or protein ID. Afterwards, change the "in field" box respectively. Here, type in "*dtxR"*, set "in field" to "Gene name", and click "Search".

(iv) Again, a query for *dtxR* results in four database entries for corynebacteria. To compare homologous genes between different organisms, select a gene from the organism of interest on the left site of the window and click on "Add" to import the selection. The selected gene will be added to the import table (right side). Click "Remove" to reverse a previous selection. For this protocol: "Add" *dtxR* from *C. efficiens* YS-314 (entry *CE1812*) to the right import list and press "OK". The subsequently displayed window asks to choose a depth cutoff. Press "OK" to use the default value "1". You will be informed that the database is requested and the graph will be re-layouted circularly afterwards. Confirm with "OK".
TROUBLESHOOTING

(v) After the network import process has finished, you are asked to import candidates for homologous proteins. Click "Yes" and "OK". Wait for the homology information to be imported.

(vi) GraphVis now displays the depth-1-networks of *dtxR* of *C. glutamicum* and *C. efficiens*. To visualize these two networks in a comparative way, select both *dtxR* genes. Therefore, click one central node, hold "Shift" pressed on your keyboard and click on the other *dtxR* node. Go to "Graph layouts" → "Homology Layouter" → "Run homology layouter" and press "OK" to change the layout.

(vii) Figure 3b shows the two DtxR-networks in comparative layout style. Genes with a homologous counterpart in the other network are located in the middle and connected by undirected black edges. These edges reflect the homology and may later be removed via "Tools" → "Homologies" → "Clear all homology edges". All nodes without any homologous relation are located in a semi-circle left and right to the *dtxR* nodes, respectively. Exemplarily, investigate the two genes *cg0527* of *C. glutamicum* and *CE0466* of *C. efficiens*; both are down-regulated by the corresponding *dtxR* genes. Since they are putative homologs they are located in the middle and connected by a black edge. Note that clicking on one gene highlights all potential homologs in the visualized graph.
TROUBLESHOOTING

(viii) Optional step: GraphVis further offers an alternative comparative graph layouter. To apply the second homology layout, select both *dtxR* nodes as in step (vi), go to "Graph layouts" → "Force based homology layouter" and click "OK". In contrast to the previous layout, all nodes without a homologous counterpart are located in a semi-circle below the central nodes. Two homologous genes are located at the same position relative to the central node in the upper part of the displayed graph. For a better overview you can add/remove the homology edges as described in step (vii). For more optional features of GraphVis refer to Box 3.
TROUBLESHOOTING

(C) Inter-species transfer of regulatory networks (TIMING ~ 20 min)

(i) CoryneRegNet provides an easy-to-use interface for the prediction of transcription factor binding sites for a gene of interest. Enter "cg0350" in the search field of the query page. The default search option for "in field" is set to "Gene ID", so no more changes have to be performed. You can either search all or one specific organism for the gene ID of interest. Since *cg0350* is a unique gene Id for *glxR* of *C. glutamicum* ATCC 13032 use the default setting and press the "Start search" button.

(ii) A search for *cg0350* results in one database entry. To access the details page of the gene of interest, click on the gene Id in the “Gene ID” column. Here, click on “cg0350”.
TROUBLESHOOTING

(iii) As described in Box 1, the next page provides detailed information about the *glxR* gene. We are interested in the prediction of transcription factor binding sites for inter-species transfer to *C. jeikeium* in this protocol. Expand the module “Binding site prediction (for this regulator in other upstream sequences)” as shown in Figure 4a. Note: The user can also predict binding sites within the upstream region of *glxR* by expanding “Binding site prediction (in the upstream sequence of this gene)”; see Box 2 for more details.

(iv) The module “Binding site prediction (for this regulator in other upstream sequences)” allows the user to select a target organism and to decide which nucleotide content should be used as background model: the complete nucleotide content; the nucleotide content of the coding regions; or of the noncoding regions. Furthermore the p-value threshold has to be specified (for a full description of all parameters and their interpretation refer to Box 2). Select “C. jeikeium K411” as target organism, set “Nucleotide content of the noncoding regions” as background model, select a comparably moderate p-value threshold of “1e-6 (10^-6)”, make sure that “genes in operons” is checked, and start your search by clicking on “Start search” (refer to Figure 4a for a graphical illustration of the user-interface to TFBScan).

(v) Wait for the results page to appear. Figure 4b shows the results page for this step of the protocol. See the PWM computed for the used TFBSs and the background nucleotide content model. Scroll down to investigate the table below. It summarizes the prediction results: For each target gene that is preceded by a binding site conforming to the chosen p-value threshold, the following information is shown (most important only): the target gene ID; the target gene name; the corresponding predicted operon; the gene’s strand in the DNA; the p-value of the potential TFBS; the TFBS sequence; and a list of putative homologs of genes in the verified list of *glxR* target genes in *C. glutamicum*. Note that especially those genes are high-potential candidates to be regulated by GlxR in *C. jeikeium*, since the regulator is conserved, the transcriptional recognition site is conserved, and the target gene is conserved as well.
TROUBLESHOOTING

(vi) From this page, GraphVis may be executed (note the “prediction mode” hint). Click on the corresponding “GraphVis” button to start this application. Press “No”, when being asked to import homologies at this point. Note that in the prediction mode, the directed edges of predicted regulations are colored green since there is no information to distinguish between up- or down-regulation.
TROUBLESHOOTING

(vii) To extend the network for a further comparative analysis with the reference network the original network (origin) has to be imported. Here, this is the *glxR*-network of *C. glutamicum*. Therefore, navigate to “Tools” → “Import original database network” and press “OK”. Again the user is asked whether potential homologies should be imported. Here, press “Yes” and “OK” to start a corresponding database query.

(viii) GraphVis offers two layouts to highlight the similarities between two networks: in this case between the predicted *glxR* network of *C. jeikeium* and the known *glxR* network of *C. glutamicum*. As in the previous section (option B, step (vi)), select both center nodes (*glxR*) and go either to “Graph layouts” → “Homology Layouter” → “Run homology layouter” or ”Graphs layouts” → “Force based homology layouter” and press “OK” to re-layout the graph. The user is essentially provided with the same information as with the table at the end of step (v), but the graphical representation eases the results interpretation.
TROUBLESHOOTING

(D) Transcription factor binding site prediction (TIMING ~ 10 min)

(i) After entering the query interface of CoryneRegNet, click on the “TFBScan” button at the bottom of the page to access the binding site prediction feature.

(ii) Two modules refer to the two possible applications of TFBScan as described in Box 2. Given some known TFBSs of a regulator, the aim is to predict further binding sites within the same organism. Therefore, scroll down to the second module. Now open supplementary file 1 with a text editor. It contains some of the known binding sites of the transcription factor DtxR of *C. glutamicum*. Copy+paste all binding motif sequences (press “Ctrl” + “A” to select all → “Ctrl” + “C” to copy) into the text field (“Ctrl” + “V” to paste) of the module.

(iii) Optional step: PWM-based prediction performance may be improved by readjusting the stored TFBS annotations. Therefore, use the binding motif re-assessment tool MoRAine to improve the information content of the respective PWM. See Box 4 for further details.

(iv) The dropdown menu “Organism” allows choosing the target organism to search for putative binding sites. Select “*Corynebacterium glutamicum* ATCC 13032“ for this particular protocol.

(v) Choose “Nucleotide content of the noncoding regions” in the next dropdown menu as background model (also refer to Box 2 for more details).

(vi) Select “1e-5 (10^-5)” as a moderate cut-off for this particular protocol. The p-value threshold highly influences the results of the search (see Box 2 for details).

(vii) Decide either to allow or prohibit searching for reverse motifs and complementary motifs. We recommend using either both or none of these options. The last choice (“genes in operons”) is to determine whether genes in an operon should be displayed in the results table. It does not influence the search procedure but the amount of genes presented. Use the default parameters here, which are: no reverse motifs; no complementary motifs; report genes in operons.

(viii) To call the prediction function click on the button “Start search”. Wait for the results page to appear. Here the resulting PWM (as computed from your input TFBSs) is displayed together with the chosen nucleotide content (background model). At the bottom of this page a table lists the prediction results: all genes with subsequences in their promoter regions that are conform to the previously selected p-value threshold. These genes represent further putative target genes of DtxR in *C. glutamicum*. Besides the names and IDs of the reported genes, the table shows further important information: the corresponding p-values, the matched sequences in the upstream regions, and the positions of these.
TROUBLESHOOTING

(E) Inconsistencies in gene expression studies (TIMING ~ 10 min)

(i) For the detection of inconsistencies in gene expression studies given the known regulatory network stored in the database, start the CoMa feature by clicking on the button “CoMa” at the bottom of the query page.

(ii) There are three different possibilities to provide CoMa with gene expression data: copy+paste expression data into a text field; upload an expression data file; import from the stimulon database. Choose one by clicking on the corresponding link. For this particular protocol use “Text input”.

(iii) The main page of CoMa explains the required data format. Download supplementary file 2, which contains a *dtxR*-knock-out gene expression experiment[43] in the correct format.

(iv) To filter insignificantly differentially expressed genes, choose the absolute of the “mValue threshold” by typing a number greater than 0 into the corresponding text field. We recommend using a value greater or equal to 1. For this protocol, use the default parameter “1”.

(v) Check the next box (“ignore autoregulations?”) to decide to exclude autoregulations from further analyses.

(vi) Since expression levels may be provided as ratios instead of M-values, the user can specify this in the next dropdown menu. In supplementary file 2, M-values are used. Hence, no changes are necessary at this point.

(vii) To provide CoMa with input data, open the downloaded supplementary file 2 with a text editor and copy+paste the whole content into the text field (press “Ctrl” + “A” → “Ctrl” + “C” in the text editor and “Ctrl” + “V” in the text field).

(viii) To start the analysis click on the button “Start search”. Wait for the calculation to be finished and the appearance of the results page. The results are presented in several tables. “Putative regulatory contradictions” occur if a repressor is up (down) regulated and its target gene as well or if an activator is up (down) regulated and its target gene is down (up) regulated. “Putative operon contradictions” occur if not all genes within an operon are regulated in the same direction, i.e. all up or all down. Note that the second last column (“Non-contradictory regulations”) of the first table (“Putative regulatory contradictions”) lists potential explanations for the observed inconsistencies, i.e. further transcriptional regulators that interact with the respective target gene. The same analysis may be performed with Cytoscape[46] by using the CoryneRegNet and CoMa plug-ins[47] (see Box 5 for further details).

TIMING

The time required to execute this protocol is also related to the querying time of GraphVis for the DtxR and GlxR networks. With a 1.5 Mbs ADSL internet connection under favorable operating conditions, it takes approximately up to 30 sec to download the GraphVis Java Applet (the first time GraphVis is called), and up to 2 min for each of the two networks. Extending the two graphs may take another 2 min, as does the import of homology information for each network. Running times for COMA, TFBScan, and graph layouting can be neglected (less than 10 sec). An experienced user can execute the protocol within 45 - 60 min.

Option A - Combined visualization of gene regulatory networks and gene expression data: ~ 10 min

Option B - Comparative visualization of gene regulatory networks: ~ 10 min

Option C - Inter-species transfer of regulatory networks: ~ 20 min

Option D - Transcription factor binding site prediction: ~ 10 min

Option E - Inconsistencies in gene expression studies: ~ 10 min

TROUBLESHOOTING

Troubleshooting advice can be found in Table 2.

ANTICIPATED RESULTS

Here, we discuss the concrete results that can be achieved by executing the five options in this protocol.

Options A and B - Gene regulatory network and gene expression data visualization

The goal is to visualize and analyze all available transcriptional regulatory information for a gene of interest together with gene expression data, and to cross-species compare the network with that of a taxonomically related species.

The regulatory network of the global iron-dependent repressor DtxR of *C. glutamicum* consists of d*txR* and 64 target genes (57 repressed, 7 activated); refer to Figure 2 for a screenshot of the *dtxR* details page. In the investigated (*dtxR* knock-out) gene expression study [43], 261 genes are differentially expressed, 47 of them are genes within the DtxR network (39 up-stimulated, 8 down-stimulated). The screenshot in Figure 3a illustrates these results after projection onto the graph-visualization of the corresponding network with GraphVis. No inconsistencies in the DtxR-network response to the stimulon occur, i.e. there is either no response, or all activated target genes are down-stimulated while all repressed target genes are up-stimulated.

The phylogenetic comparison of the *dtxR* network of *C. glutamicum* to that of *C. efficiens* is illustrated in Figure 3b by using the comparative graph layout visualization feature of GraphVis. In *C. efficiens*, DtxR regulates 27 genes (22 repressions, 5 activations). Here, 28 DtxR target genes in *C. glutamicum* are potential homologs to 18 target genes in *C. efficiens*. Note that the homology information is deduced from sequence-based similarities. Hence, we sometimes observe clusters of putative homologous genes instead of one-to-one relationships (e.g. *cg0159*, *cg0160*, and *CE0125*; refer to the network zoom in Figure 3b). Here the user has to decide, which of the two *C. glutamicum* target genes represents the *CE0125* homolog, based on further background information, such as genomic context conservation.

Option C - Inter-species transfer of regulatory networks

Starting with the known binding sites of a transcription factor, the aim is to predict further gene regulatory interactions in a closely related species and to compare it with the known network of the reference organism.

The GlxR network of *C. glutamicum* consists of 150 genes (106 repressed, 44 activated), 89 binding sites are known and used to build a model for TFBScan. By using a p-value threshold of 10^{-6} for predictions in *C. jeikeium*, we find 14 putative TFBSs (refer to Figure 4b). Six of them are high-potential target genes for further wet-lab studies since they show a sequence-based similarity to known *glxR* targets in *C. glutamicum* (last column), i.e. (1) the regulator is conserved, (2) the binding site is conserved, and (3) the target genes is conserved as well. Note that for the putative homology *cg2640* (*benD*) / *jk0230* (*fabG1*), the BLAST E-value is comparably poor (10^{-13}) and further evidence may be necessary. Further note the similarity of *jk0416* (*rpfA*) to *cg0936* (*rpf1*) and *cg0937* (*rpf2*). Here the sequence of *rpfA* compares much better to that of *rpf1* than to *rpf2* (E-value of 10^{-44}) suggesting a homology *rpfA*/*rpf1* (also refer to [48]).

Option D - Transcription factor binding site prediction

Here, the goal is to utilize known binding sites of a transcription factor of interest to predict further candidate sites in the same organism.

Assuming that we only know 15 arbitrarily selected DtxR binding sites of *C. glutamicum* (supplementary file 1), we can scan for further binding sites (i.e. gene regulatory interactions) within the same organism. Since we know the true TFBSs, we evaluate the prediction performance in the following. Within the listed 33 putative TFBSs, 14 of the 15 known TFBSs used to compute the PWM model can be found. This indicates that the chosen p-Value was reasonable. From the 33 hits, we can remove those that are between two genes lying next to each other in the genome but on opposite strands (they share the upstream sequence with the hit), if one of them is a true positive (i.e. known to be regulated by DtxR). This results in 28 predicted TFBSs, with 4 putative false positives: *cg0923*, *cg0957*, *cg1070*, and *cg1580* (14% false positive rate). A closer investigation of *cg1580* (*argC*) with the *dtxR* knock-out study (see options A and B in the protocol) unravels a response of this gene to a DtxR-loss-of-function stimulus. Hence, the predicted binding site for *argC* may be a true positive and should be studied in further wet-lab experiments.

Option E - Inconsistencies in gene expression studies

The goal in this section is to identify unexpected responses of the network to an external stimulus.

Here, we study the consistency of the network response to the *dtxR* knock-out stimulus[43], but this time on a large scale, i.e. for the whole *C. glutamicum* regulatory network. CoMa detects potential contradictions for 29 gene regulations (for 13 transcription units). For all but three, other regulatory interactions are known that could possibly explain the inconsistent target gene expression level. No further transcription factors regulating *cg1813*, *cg1814*, *cg2810*, and *cg3138* are known that could unravel their contradicting gene expression levels. Hence, they are good candidates for further *in silico* and wet-lab investigations, since there must be an unknown regulatory interaction triggering the unexpected response of the network to the stimulus.

ACKNOWLEDGMENTS

JB thanks the German Academic Exchange Service (DAAD) for funding his work at ICSI, Berkeley. TW is grateful for financial support of the International Graduate School for Bioinformatics and Genome Research (IGS, Bielefeld, Germany) and the Boehringer Ingelheim Fonds. CKK thanks the IGS as well as the Arabidopsis Functional Genomics network (AFGN) – 2010 Young Researcher Exchange Program for funding her work at UC Berkeley.

Competing Financial Interests

The authors declare no competing financial interests.

BOXES

Box 1 | Interpretation of the details page

Depending on the database content and the type of gene, the following interactive modules are displayed on the details page. Click “+”/”-“ to expand/hide.

- Genomic context: A genome browser visually summarizes the genomic as well as the regulatory sites context. Click at a specific gene or binding site to browse there.
- Gene details: Basic information, such as gene name, ID, and the corresponding organism are shown; if available, links to NCBI, GenDB[49], and RegulonDB are provided. Click on the respective link to browse the internal and external database entries.
- Further gene annotations (GenDB web service): The integrated Web Service client to the genome annotation systems GenDB provides further up-to-date gene annotations and links to external databases. Click on them to browse the respective entries.
- Protein details: Protein name, ID, and links to NCBI are shown within this module. Click on the links to browse the respective database entries.
- Regulations: regulated by: A table of all transcription factors that regulate the gene of interest shows the (1) TF gene ID, (2) TF gene name, (3) TF protein ID, (4) TF protein name, (5) whether the TF is an activator or repressor, (6) an evidence code, (7) whether a binding site exists and if it is known or predicted, (8) the binding sites recognized by the TF, and (9) a link to the corresponding PubMed entry. Click on the links to browse the respective internal database entries or to go to PubMed.
- Regulations: regulates: A table similar to that of the previous module lists those genes that are regulated by the gene of interest, including the operon memberships. Click on one of the links to browse the respective entry.
- Regulations: regulates pathways: A table with links to KEGG pathways that contain at least one target gene. Click on the “KEGG PathwayID” to browse to the respective external web page.
- Stimulons: If it is known that the gene is differentially expressed within a gene expression study, a table provides information about the corresponding stimulon. Click at one of the gene links or a PubMed ID to explore the respective site.
- Attributes, position weight matrix, and sequence logo: If available, this module presents information about (1) the gene start, (2) stop, and (3) DNA strand, (4) the start codon position, (5) the nucleotide content, (6) the regulator type, (7) if the TF is an autoregulator, (8) the position weight matrix shown as count matrix table, (9) a sequence logo of the TFBSs, and (10) histograms visualizing the distribution of the binding site distances from the target gene start.
- Binding site prediction (for this regulator in other upstream sequences): An interface for the integrated binding site prediction feature TFBScan allows to search for putative binding sites in upstream regions of all genes within the database, if the gene of interest encodes for a transcriptional regulator with known TFBSs. See Box 2 for more details.

- Binding site prediction (in the upstream sequence of this gene): A similar interface allows utilizing all known position weight matrices to scan the upstream region of the gene of interest for putative binding sites. See Box 2 for more details.
- Gene/Protein identifiers: Known identifiers from other databases are listed.
- Candidates for homologous genes/proteins: A tabular listing of all genes/proteins that show a sequence-based similarity to the gene of interest and hence candidates for putative homologs. Provided is information about their ID, name, the lowest BLAST E-values (for both directions), and the corresponding organism. Click on a gene, protein, or organism ID to explore the corresponding dataset.
- Protein cluster: A table lists all members of the group of orthologous proteins obtained from the integrated protein sequence clustering software FORCE[38]. Click on a gene, protein, or organism ID to explore the corresponding dataset.
- Gene and protein sequences: The nucleotide sequence of the gene and the amino acid sequence of the encoded protein are shown.
- GraphVis: The interface to the integrated graph-based network visualization tool GraphVis. See Box 3 for details.

END OF BOX 1

Box 2 | Binding site prediction and inter-species network transfer

Generally, there are two ways within the CoryneRegNet web interface to access the integrated transcription factor binding site prediction (TFBScan) feature. First we explain some technical terms, subsequently how to apply TFBScan.

Background model, p-values, reverse/complementary motifs, and operon genes

The nucleotide content serves as background model for the statistical evaluation while the p-value cut-off is a threshold for the p-values (false-positive probabilities) of the predicted TFBSs. Note that using p-values instead of scores for a hit provides the opportunity to directly judge the predictions of TFBS models (PWMs) of different quality. For further details, the reader is referred to the original publication of the integrated PoSSuMsearch software[35]. The “Nucleotide of the non-coding regions” normally is an appropriate background model and p-values of 10^{-3} or 10^{-4} are considered to be comparably bad while 10^{-8} or 10^{-9} indicate good matches. Check “reverse motifs”/”complementary motifs” if you wish to scan the reverse/complementary promoter region as well. Check report “genes in operons” if you wish to include putative target genes within operons (that are not directly preceded by a binding site) in the results list together with the first gene in the corresponding operon. If your scan yields no or too few hits, choose a less restrictive p-value threshold.

Inter-species transfer of a known network

If the regulatory network for the gene of interest is already known and stored in the database, search for the corresponding gene after login and go to the genes details page.

If the gene encodes a transcription factor, open the module “Binding site prediction (for this regulator in other upstream sequences)”, and follow the instructions in option C, step (iv) for an *in silico* transfer of the regulatory network to another organism.

If the gene of interest does not encode for a TF, you may open the module “Binding site prediction (in the upstream sequence of this gene)” to scan for putative TFBSs of all regulators in the database in the promoter region of the displayed gene. Choose a source organism and the minimal number of binding sites that have to be known for the regulators to be included (this significantly affects the prediction quality; we do not recommend to use values below 5). Furthermore, choose the p-value cut-off and whether you wish to include reverse/complementary hits. Click “Start search” and wait for the results page to appear. It presents the used background model (the nucleotide content of the promoter region) and a list of transcription factors of the selected source organism that potentially regulate the gene of interest. Again, note the p-values of the predicted TFBSs. The last column indicates if a regulatory interaction is already known. Click on a TF to navigate to its details page.

Prediction of new networks

After login to CoryneRegNet, click on the “TFBScan” button at the bottom. This opens a new page with two menus.

Given a list of differentially expressed genes in a microarray experiment, one may check which TFs are potential regulators of these genes. Extract the upstream/promoter regions for the corresponding genes and copy+paste them in FASTA format into the first text field (under “Search binding sites in max. 10 sequences …”). Now choose the organism, the minimal number of binding sites that have to be known for the regulators to be included (again, we do not recommend using values below 5), the p-value cut-off, and whether to report reverse/complementary motifs. Click “Start search” and wait for the results page to appear. As above, in addition to the used background model it presents a list of transcription factors that potentially regulate the genes of interest. Again, note the p-values of the predicted TFBSs. The last column indicates if a regulatory interaction is already known. Therefore, the used FASTA-IDs have to match with the gene IDs within CoryneRegNet. Click on a TF to navigate to its details page.

Given a list of known binding sites for a transcription factor of interest, further regulatory sites may be predicted. Go to “Search binding sites by using a PWM” and proceed as described in option D, step (ii).

END OF BOX 2

Box 3 | Graph-based network visualization with GraphVis

CoryneRegNet provides a powerful tool for the graph-based visualization of gene regulatory networks. Nodes in the graph correspond to genes, directed edges to transcriptional regulatory interactions, and undirected edges to homologies (i.e. sequence-based similarities). This box explains how to start GraphVis, how to interpret the graph representation, and how to utilize the main features. The options A and B of the protocol illustrate workflows through these main features.

Starting GraphVis

In the CoryneRegNet web front-end the GraphVis tool may be started from any browsing or search (query) result page as well as from the details page of the gene of interest and the binding site prediction results page. The corresponding “GraphVis” menu is located at the bottom of each web site. Click on it, select a network depth threshold and (if applicable) whether genes from the regulations list should be integrated as well. Note that checking

“Include genes from regulations” is equivalent to increasing the depth cut-off by one. Afterwards click the “GraphVis” button. This will start a Java Applet that imports the corresponding network from the database (this may take a while). Subsequently choose if you would like to import homology data; corresponding relations will be visualized as undirected edges between nodes. GraphVis appears in a separate window and displays the gene regulatory network (refer to Figure 3 for an example).

User interface, symbols, and colors

The GraphVis user interface is separated into two parts (“Details” and “Graph”). Selecting a node in the graph window (bottom) will show the most important details of the corresponding gene in the top window. A click on “Import additional data” (confirm with “OK”) will query the database back-end and display all available information for the selected gene of interest.

As mentioned above, nodes in the graph correspond to genes, directed edges to regulatory interactions, and undirected edges to sequence-based similarities. Multi-nodes represent operons and can be added/removed by clicking on “Tools” and “Operon grouping” in the graph window. We use the following color codes:

Red node	Repressor
Red line	Repressing regulatory interaction
Green node	Activator
Green line	Activating regulatory interaction
Blue node	Dual regulator
Blue line	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 binding site
Black line	Putative homology (sequence based similarity)

If gene expression data has been imported (see below), nodes surrounded by red-dotted borders represent down-stimulated genes, nodes with green-dotted borders up-stimulated genes. This information may also be found in the details window under “Legend”.

Extension of the visualized network

GraphVis may start in two slightly different modes. This affects how the displayed network can be extended.

Normal mode: This is the standard mode, automatically started from all web sites. Click on “Tools” → “Extend the graph” to choose between two options: (1) Click on “Extend graph from selection” to use the currently selected genes as starting point. Enter a depth cut-off, click “OK”, and accept the prompt for confirmation. The extended network is queried from the CoryneRegNet back-end and displayed. Subsequently choose if you would like to import homology data. (2) Click on “Extend graph using wizard” and proceed as illustrated in option B, step (ii).

Prediction mode: This mode can be accessed from the results pages of the TFBS prediction feature. If the known TFBSs of a regulator of interest have been used to predict further binding sites, click the “GraphVis” button below the results table (note the “Prediction mode” hint). Proceed as described in option C, step (vi) of the protocol. When GraphVis was started, note the “%” indicator behind the gene IDs of the “original” network. Now it is possible to visually compare these two networks by using the comparative network layout method (option C, step (viii)). Note that the edges of predicted regulations are colored green since no information is available to distinguish between up- or down-regulation.

Graph layouts

GraphVis provides several possibilities to layout a network. Click on “Graph layouts” and choose one. The most useful options for a general overview are the circular, organic, and orthogonal layouts; standard is circular. The best style reflecting the structure of gene regulatory networks is the hierarchical layout. By selecting one of the options, a new window appears presenting various fine-tuning options for the favored style. First start with the standard options (click “OK”) and subsequently alter them until the graph looks as desired. Furthermore, two special comparative graph layouts have been integrated into GraphVis, namely the “Homology Layouter” and the “Force Based Homology Layouter”. Be sure to have homology information imported. Refer to option C, step (viii) for example applications of these layout algorithms.

Gene expression data visualization

Having a gene regulatory network visualized with GraphVis, one has several possibilities to project gene expression data onto the graph. One way is to use the stimulon database of CoryneRegNet. Option A in the protocol describes this procedure. Additionally, the user can upload own gene expression profiles as tab-delimited flat file (see supplementary file 2) or MS Excel file. Therefore, click on “Tools” → “Gene expression data” → “Import expression data from tab-delimited flat file” or “Import expression data from MS Excel file”. Select a file and click “Open”. Proceed as described in option A, step (viii), but additionally choose if the expression level is given as M-value or ratio. Figure 3a illustrates the results of the application of supplementary file 2 to the DtxR network of *C. glutamicum*. Finally, to remove the projection from the visualization, click “Tools” → “Gene expression data” → “Clear expression data from graph (reset)”.

Homology edges

To display/hide undirected, black edges between nodes, which are putative homologies, click “Tools” → “Show all homology edges”/”Clear all homology edges”. In order to display this information or to utilize it for comparative graph layouts (see above), corresponding information has to be imported from the database by clicking “Tools” → “Import homologies” → “OK”.

Further options

Export as JPG file: The displayed network may be exported to a JPG file. Click on “Tools” → “Export graph to JPG file”, chose a filename, click “Open”, select a resolution, click “OK”, select the view, and click “OK”.

Operon grouping: In order to show operon memberships, click on “Tools” → “Operon grouping” to toggle the clustering of nodes into meta-nodes that group genes of one operon together.

END OF BOX 3

Box 4 | Moraine

Sometimes, the quality (information content) of PWMs is poor. For instance, this may occur if binding motifs are annotated routinely 5' -> 3' relative to the target gene. MoRAine is a tool that can automatically identify these cases and adjust binding sequences by shifting TFBSs or switching their strands. See ref. [44] for more details.

- To perform the optional step (iii) of option D in the protocol, click on the corresponding "MoRAine" link in the upper right corner of the module to access this feature. We recommend opening this site in a separate window to ease coming back to the TFBScan page. Alternatively, open a new window and navigate to http://moraine.cebitec.uni-bielefeld.de/.
- Scroll down and copy+paste the binding motif sequences into the first text box.
- Start the re-assessment by clicking on the "Start" button. Wait for the results page of MoRAine.
- Two sequence logos will be displayed, with respective links to the corresponding sequences located below. On the left side, you see the logo painted from the original sequences that you pasted in option D, step (iii), on the right side the logo painted by using the MoRAine-adjusted TFBSs respectively. Click on the right "(download FASTA file)" link to open the file with the adjusted sequences.
- Mark the whole page (press "Ctrl" + "A") and copy all sequences into the clipboard ("Ctrl" + "C"). To complete optional step (iii) of option D, switch back to TFBScan and paste the sequences into the text field ("Ctrl" + "V") to proceed with the adjusted sequences.

END OF BOX 4

Box 5 | CoryneRegNet plug-ins for Cytoscape

CoryneRegNet provides a SOAP-based Web Service server which allows external tools to query data from the database in a well-structured manner, for instance the Cytoscape plug-ins CoryneRegNetLoader and CoMa[47]. They allow to visualize whole-organism networks with Cytoscape and to check for inconsistencies in gene expression studies within Cytoscape. A protocol on how to install and configure Cytoscape may be found in ref [46] and at the Cytoscape web site http://cytoscape.org. An online tutorial with application examples, screenshots, and step-by-step instructions on how to install and use the CoryneRegNet plug-ins is available through the CoryneRegNet web site by clicking on the "Cytoscape plug-ins" link after login.

END OF BOX 5

FIGURES

Figure 1 | Overview of the CoryneRegNet user interface. This figure illustrates the structure of the web front-end of CoryneRegNet. We split it into four parts reflecting the typical hierarchical major steps during gene regulatory network analysis: (1) querying/browsing to a gene of interest, (2) illustrating details for this gene and transferring its network to a closely related species, (3) visualizing and analyzing the known or the predicted network as graph, and (4) investigating this graph together with gene expression data. This figure also outlines the structure of this protocol. Steps in red are included in the protocol while features in green are useful alternatives or extended features that are not further considered here.

Figure 2 | Details page for the global iron-dependent repressor protein DtxR of *Corynebacterium glutamicum*. This screenshot shows the details web page for the gene *dtxR* (also refer to Box 1). On top of the page is a visual representation of the genomic context, at the bottom the entry point for the graph visualization feature GraphVis (see Box 3 for more details). In the middle, several modules offer preprocessed and graphically prepared detailed information (such as statistical summaries and a sequence logo) as well as the entry points for the integrated regulatory binding site prediction (refer to Figure 4a and Box 2).

Figure 3 | Visualization of the gene regulatory network of DtxR with GraphVis. This screenshot shows the gene regulatory networks of DtxR of *Corynebacterium glutamicum* **(a)** together with imported gene expression data from the stimulon database (or supplementary file 2) and **(b)** compared to that of *Corynebacterium efficiens* using the homology layout style. In the first case, node sizes are altered according to the relative gene expression levels. Refer to section “User interface, symbols, and colors” of Box 3 for symbol and color codes. Box 3 also explains how to use and to interpret the comparative homology layout as well as the gene expression data projection.

Figure 4 | Prediction of GlxR transcription factor binding sites – Network transfer from *Corynebacterium glutamicum* to *Corynebacterium jeikeium*. (a) The binding site prediction user interface as module within the details page of the *glxR* gene of *C. glutamicum* offers access to the integrated TFBScan feature. **(b)** Predicted binding sites of GlxR in *C. jeikeium*. Note the right column that shows putative homologs to known GlxR-regulated genes in *C. glutamicum* and hence indicates high-potential candidates for further wet-lab studies in *C. jeikeium*. Refer to Box 2 for detailed descriptions.

TABLES

Table 1 | Comparison of popular microbial gene regulatory network platforms. A detailed description and review may be found in ref. [8]. Abbreviations: CRN = CoryneRegNet[12-15], MRL = MtbRegList[16], PRD = PRODORIC[17,18], DBTBS = DBTBS[19-21], RDB = RegulonDB[22-28]

Feature	CRN	MRL	PRD	DBTBS	RDB
Web interface	+	+	+	+	+
Genome browser	+	+	+	+	+
Network visualization	+		+		+
Raw data access		+			+
Binding site prediction	+	+	+	+	

Data exchange methods	+				
Network and gene expression data analysis	+		+		
Homology detection	+			+	+
Statistical summaries	+		+		+

Table 2 | Troubleshooting

Step	Problem	Possible reason	Solution
Option A, step (ii) Option B, step (iv) Option C, step (ii)	No database query results	Incorrect “in-field” selection	Make sure you chose the correct “in field” menu entry. Obviously, searching for a certain gene name by using the “in field“-option “protein ID”, cannot provide any hit.
		Mistyped search string	Navigate back and enter the correct search string.
Option C, step (v) Option D, step (viii)	Too small number of predicted binding sites	The quality of the PWM used for the prediction is poor.	Check the sequence logo. The blue bar at the left side indicates the average information content of the calculated PWM. Try MoRAine to optimize the binding site annotation. See Box 4 for details.
		The noise in the scanned upstream regions is too high. Hence, the selected p-value threshold is too restrictive.	Try using a less restrictive threshold. Refer to Box 2 for further details about reasonable thresholds for binding site predictions.
Option B, steps (vii) and (viii) Option C, step (viii)	No homology layout appears.	Homology information was not imported.	Import homology information as described in Box 3, section “Homology edges”.
Option A, step (vi) Option C, step (vi)	GraphVis does not start.	Java may not be installed or inactivated in your web browser.	Install and configure Java from http://www.java.com.
		Certificate was not accepted.	Navigate back and accept the certificate.

		Firewall settings block GraphVis internet connections.	Alter firewall settings to allow Java to access the internet.

SUPPLEMENTARY FILES

Supplementary file 1 | Sample transcription factor binding sites (15 arbitrarily selected) for DtxR of *Corynebacterium glutamicum* as text file in FASTA format. This text file contains 15 arbitrarily selected binding sites of DtxR in FASTA format. They are used as input for the TFBScan feature in option D of the protocol.

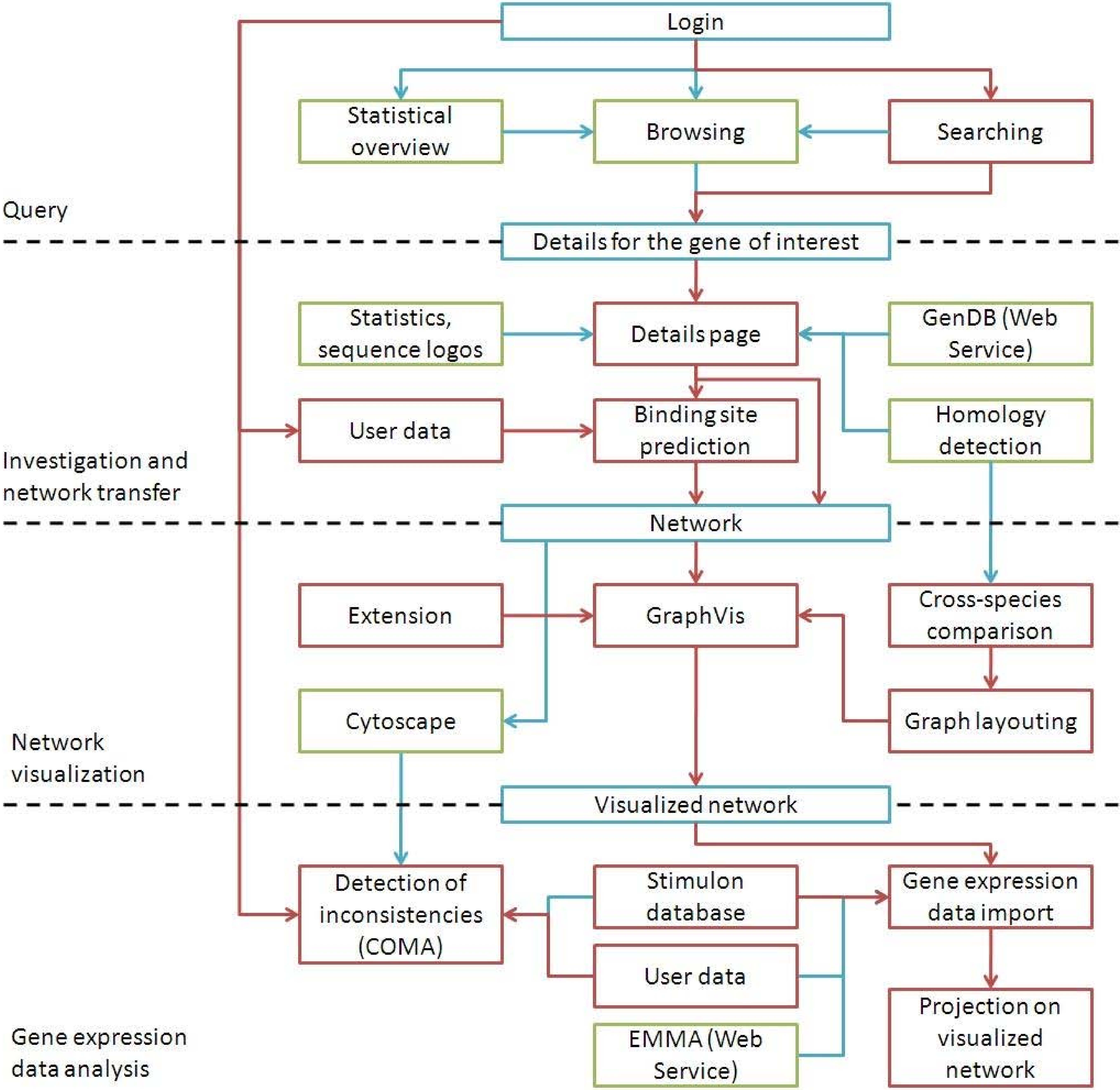
Supplementary file 2 | Sample DtxR knock-out gene expression data for *Corynebacterium glutamicum* as tab-delimited flat-file. This text file contains all significantly differentially expressed genes and M-values for a DtxR knock-out microarray experiment[43]. The gene IDs and M-values are stored tab-delimited, one line for each pair. This file is used in option E of the protocol.

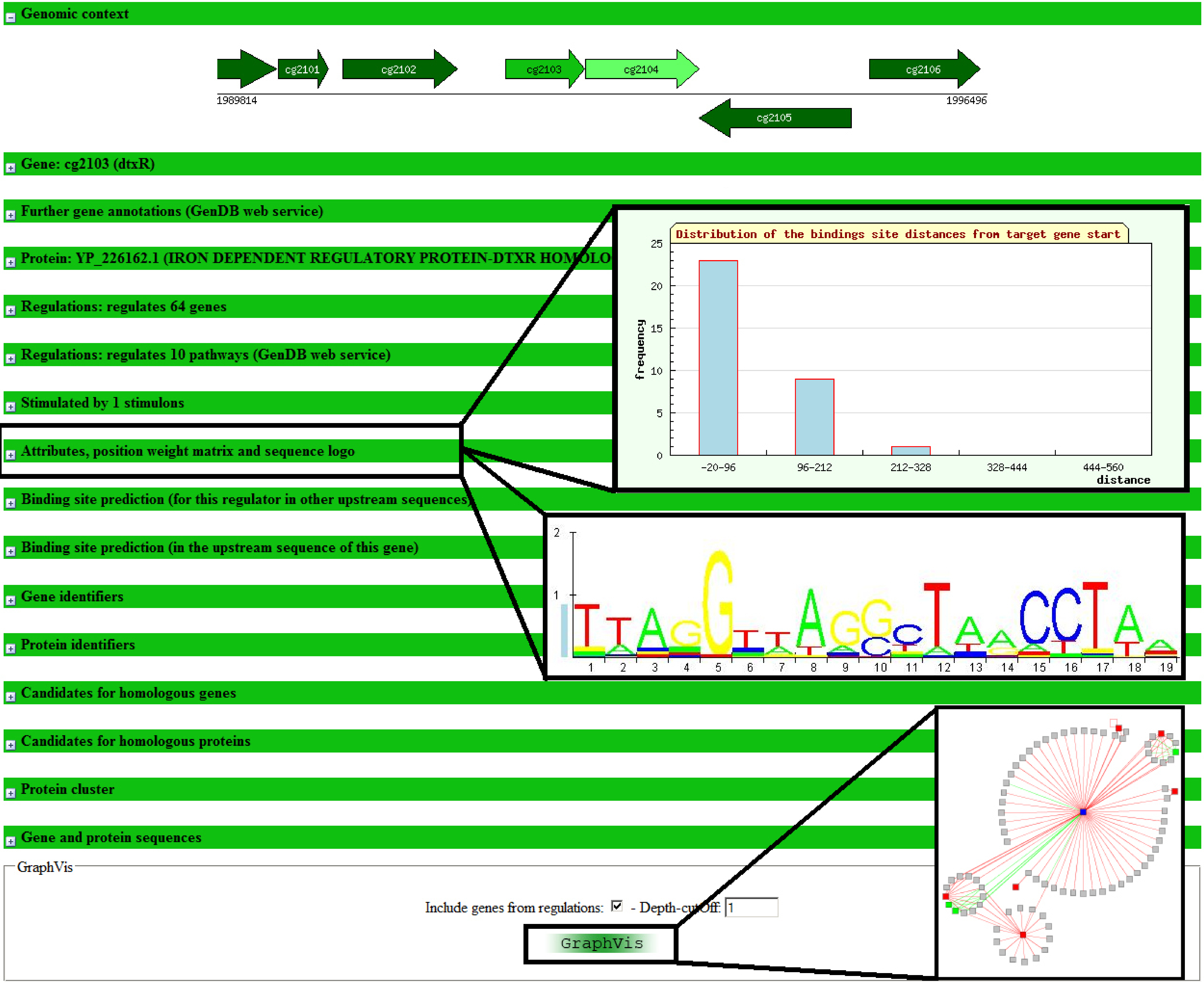
REFERENCES

1. Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M. & Teichmann, S.A. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* **14**, 283-91 (2004).
2. Madan Babu, M. & Teichmann, S.A. Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res* **31**, 1234-44 (2003).
3. Madan Babu, M., Teichmann, S.A. & Aravind, L. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* **358**, 614-33 (2006).
4. Matic, I., Taddei, F. & Radman, M. Survival versus maintenance of genetic stability: a conflict of priorities during stress. *Res Microbiol* **155**, 337-41 (2004).
5. Teichmann, S.A. & Babu, M.M. Gene regulatory network growth by duplication. *Nat Genet* **36**, 492-6 (2004).
6. Balaji, S., Babu, M.M. & Aravind, L. Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of E. coli. *J Mol Biol* **372**, 1108-22 (2007).
7. Balaji, S., Iyer, L.M., Babu, M.M. & Aravind, L. Comparison of transcription regulatory interactions inferred from high-throughput methods: what do they reveal? *Trends Genet* **24**, 319-23 (2008).
8. Baumbach, J., Tauch, A. & Rahmann, S. Towards the integrated analysis, visualization and reconstruction of microbial gene regulatory networks. *Brief Bioinform* **10**, 75-83 (2009).
9. Herrgard, M.J., Covert, M.W. & Palsson, B.O. Reconstruction of microbial transcriptional regulatory networks. *Curr Opin Biotechnol* **15**, 70-7 (2004).
10. Blattner, F.R. et al. The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453-74 (1997).
11. Kalinowski, J. et al. The complete Corynebacterium glutamicum ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. *J Biotechnol* **104**, 5-25 (2003).
12. Baumbach, J. CoryneRegNet 4.0 - A reference database for corynebacterial gene regulatory networks. *BMC Bioinformatics* **8**, 429 (2007).
13. Baumbach, J., Brinkrolf, K., Czaja, L.F., Rahmann, S. & Tauch, A. CoryneRegNet: An ontology-based data warehouse of corynebacterial transcription factors and regulatory networks. *BMC Genomics* **7**, 24 (2006).

14. Baumbach, J., Brinkrolf, K., Wittkop, T., Tauch, A. & Rahmann, S. CoryneRegNet 2: An Integrative Bioinformatics Approach for Reconstruction and Comparison of Transcriptional Regulatory Networks in Prokaryotes. *Journal of Integrative Bioinformatics* **3**, 24 (2006).
15. Baumbach, J. et al. CoryneRegNet 3.0--an interactive systems biology platform for the analysis of gene regulatory networks in corynebacteria and Escherichia coli. *J Biotechnol* **129**, 279-89 (2007).
16. Jacques, P.E. et al. MtbRegList, a database dedicated to the analysis of transcriptional regulation in Mycobacterium tuberculosis. *Bioinformatics* **21**, 2563-5 (2005).
17. Munch, R. et al. PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res* **31**, 266-9 (2003).
18. Munch, R. et al. Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics* **21**, 4187-9 (2005).
19. Ishii, T., Yoshida, K., Terai, G., Fujita, Y. & Nakai, K. DBTBS: a database of Bacillus subtilis promoters and transcription factors. *Nucleic Acids Res* **29**, 278-80 (2001).
20. Makita, Y., Nakao, M., Ogasawara, N. & Nakai, K. DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Res* **32**, D75-7 (2004).
21. Sierro, N., Makita, Y., de Hoon, M. & Nakai, K. DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res* **36**, D93-6 (2008).
22. Gama-Castro, S. et al. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* **36**, D120-4 (2008).
23. Huerta, A.M., Salgado, H., Thieffry, D. & Collado-Vides, J. RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res* **26**, 55-9 (1998).
24. Salgado, H. et al. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Res* **32**, D303-6 (2004).
25. Salgado, H. et al. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* **34**, D394-7 (2006).
26. Salgado, H. et al. RegulonDB (version 2.0): a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res* **27**, 59-60 (1999).
27. Salgado, H. et al. RegulonDB (version 3.0): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Res* **28**, 65-7 (2000).
28. Salgado, H. et al. RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Res* **29**, 72-4 (2001).
29. Cerdeno-Tarraga, A.M. et al. The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129. *Nucleic Acids Res* **31**, 6516-23 (2003).
30. Tauch, A. et al. Complete genome sequence and analysis of the multiresistant nosocomial pathogen Corynebacterium jeikeium K411, a lipid-requiring bacterium of the human skin flora. *J Bacteriol* **187**, 4671-82 (2005).
31. Nishio, Y. et al. Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of Corynebacterium efficiens. *Genome Res* **13**, 1572-9 (2003).
32. Cole, S.T. et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393**, 537-44 (1998).
33. Philippi, S. & Kohler, J. Addressing the problems with life-science databases for traditional uses and systems biology. *Nat Rev Genet* **7**, 482-8 (2006).
34. Rahmann, S., Müller, T. & Vingron, M. On the power of profiles for transcription factor binding site detection. *Statistical Applications in Genetics and Molecular Biology* **2**, Article 7 (2003).
35. Beckstette, M., Homann, R., Giegerich, R. & Kurtz, S. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* **7**, 389 (2006).

36. Kel, A.E. et al. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* **31**, 3576-9 (2003).
37. Rahmann, S. et al. Exact and heuristic algorithms for weighted cluster editing. *Comput Syst Bioinformatics Conf* **6**, 391-401 (2007).
38. Wittkop, T., Baumbach, J., Lobo, F.P. & Rahmann, S. Large scale clustering of protein sequences with FORCE -- A layout based heuristic for weighted cluster editing. *BMC Bioinformatics* **8**, 396 (2007).
39. Enright, A.J., Kunin, V. & Ouzounis, C.A. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res* **31**, 4632-8 (2003).
40. Paccanaro, A., Casbon, J.A. & Saqi, M.A. Spectral clustering of protein sequences. *Nucleic Acids Res* **34**, 1571-80 (2006).
41. Frey, B.J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972-6 (2007).
42. Dondrup, M. et al. EMMA 2--a MAGE-compliant system for the collaborative analysis and integration of microarray data. *BMC Bioinformatics* **10**, 50 (2009).
43. Brune, I. et al. The DtxR protein acting as dual transcriptional regulator directs a global regulatory network involved in iron metabolism of Corynebacterium glutamicum. *BMC Genomics* **7**, 21 (2006).
44. Baumbach, J., Wittkop, T., Weile, J., Kohl, T. & Rahmann, S. MoRAine - A web server for fast computational transcription factor binding motif re-annotation. *Journal of Integrative Bioinformatics* **5**(2008).
45. Baumbach, J., Rahmann, S. & Tauch, A. Reliable transfer of transcriptional gene regulatory networks between taxonomically related organisms. *BMC Syst Biol* **3**, 8 (2009).
46. Cline, M.S. et al. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2**, 2366-82 (2007).
47. Baumbach, J. & Apeltsin, L. Linking Cytoscape and the corynebacterial reference database CoryneRegNet. *BMC Genomics* **9**, 184 (2008).
48. Hartmann, M. et al. The glycosylated cell surface protein Rpf2, containing a resuscitation-promoting factor motif, is involved in intercellular communication of Corynebacterium glutamicum. *Arch Microbiol* **182**, 299-312 (2004).
49. Meyer, F. et al. GenDB--an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res* **31**, 2187-95 (2003).


Login
Statistical overview
Browsing
Searching
Query
Details for the gene of interest
Statistics, sequence logos
Details page
GenDB (Web Service)
User data
Binding site prediction
Homology detection
Investigation and network transfer
Network
Extension
GraphVis
Cross-species comparison
Cytoscape
Graph layouting
Network visualization
Visualized network
Detection of inconsistencies (COMA)
Stimulon database
Gene expression data import
User data
Projection on visualized network
EMMA (Web Service)
Gene expression data analysis



⊟ Genomic context

⊞ Gene: cg2103 (dtxR)

⊞ Further gene annotations (GenDB web service)

⊞ Protein: YP_226162.1 (IRON DEPENDENT REGULATORY PROTEIN-DTXR HOMOLO

⊞ Regulations: regulates 64 genes

⊞ Regulations: regulates 10 pathways (GenDB web service)

⊞ Stimulated by 1 stimulons

⊞ Attributes, position weight matrix and sequence logo

⊞ Binding site prediction (for this regulator in other upstream sequences)

⊞ Binding site prediction (in the upstream sequence of this gene)

⊞ Gene identifiers

⊞ Protein identifiers

⊞ Candidates for homologous genes

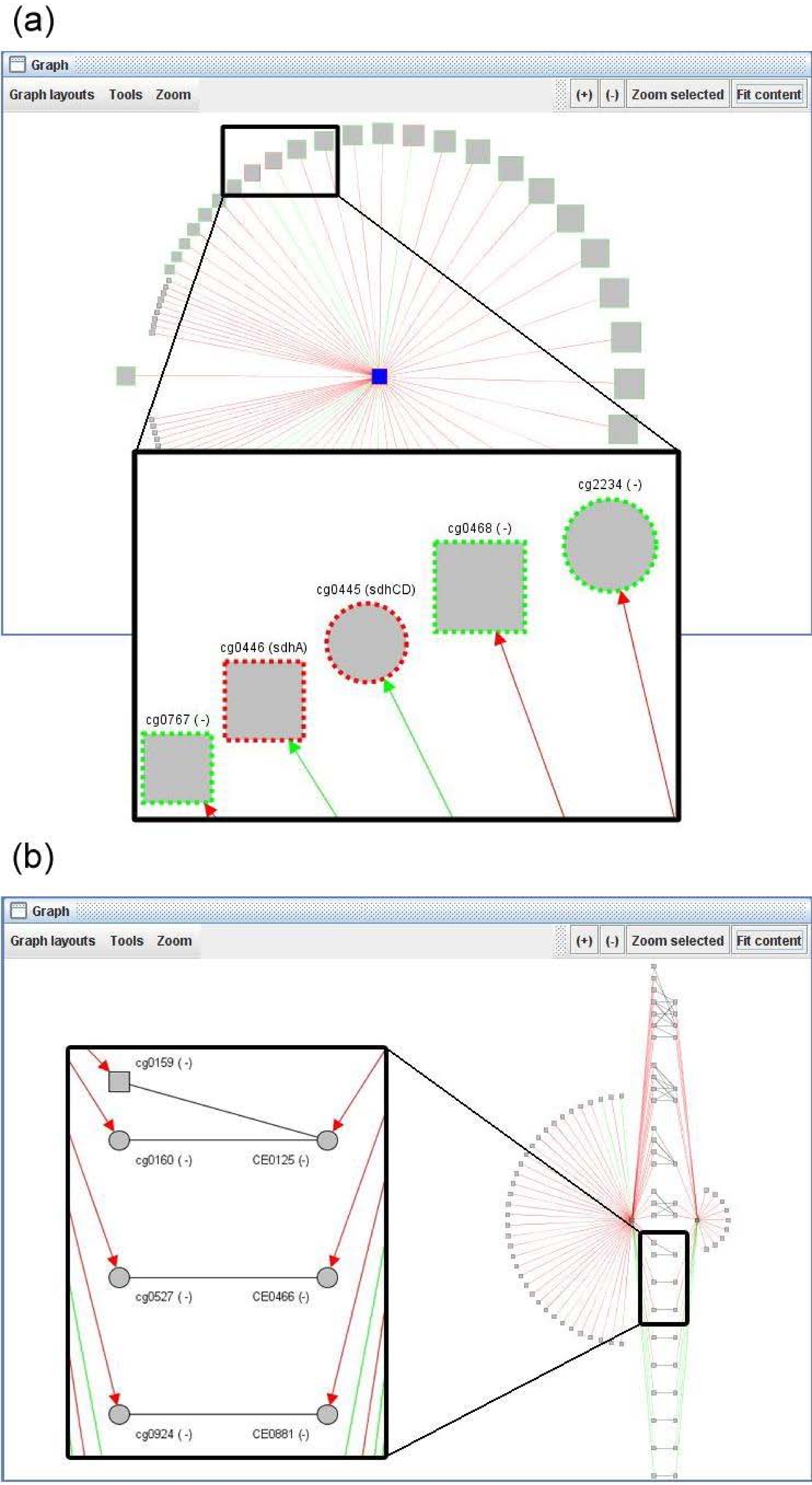
⊞ Candidates for homologous proteins

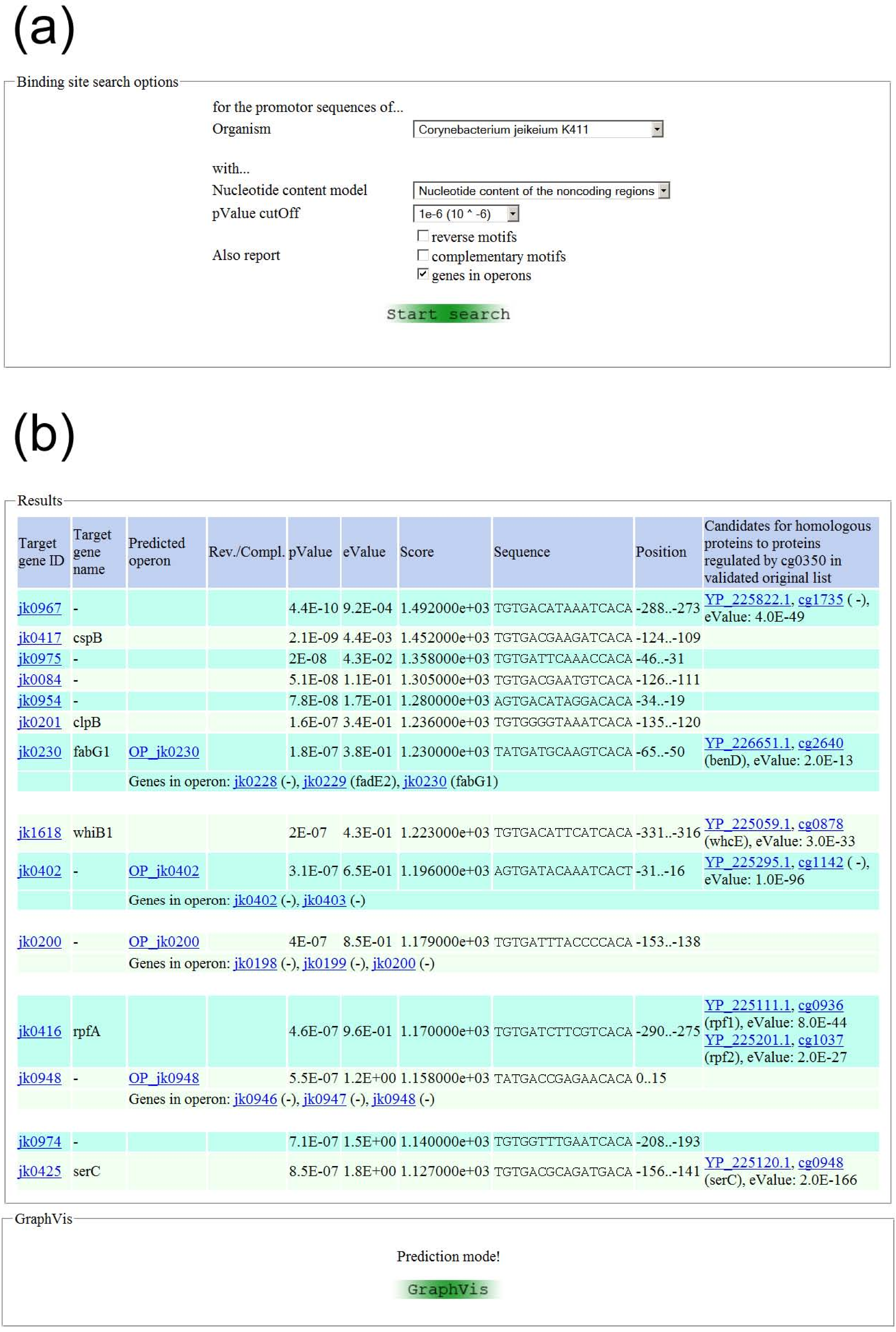
⊞ Protein cluster

⊞ Gene and protein sequences

GraphVis

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


(a)
Graph
Graph layouts
Tools
Zoom
(+)
(-)
Zoom selected
Fit content
cg2234 (-)
cg0468 (-)
cg0445 (sdhCD)
cg0446 (sdhA)
cg0767 (-)
(b)
Graph
Graph layouts
Tools
Zoom
(+)
(-)
Zoom selected
Fit content
cg0159 (-)
cg0160 (-)
CE0125 (-)
cg0527 (-)
CE0466 (-)
cg0924 (-)
CE0881 (-)



(a)

Binding site search options

for the promotor sequences of...	
Organism	Corynebacterium jeikeium K411
with...	
Nucleotide content model	Nucleotide content of the noncoding regions
pValue cutOff	1e-6 (10 ^ -6)
	□ reverse motifs
Also report	□ complementary motifs
	☑ genes in operons

(b)

Results

Target gene ID	Target gene name	Predicted operon	Rev./Compl.	pValue	eValue	Score	Sequence	Position	Candidates for homologous proteins to proteins regulated by cg0350 in validated original list
jk0967	-			4.4E-10	9.2E-04	1.492000e+03	TGTGACATAAATCACA	-288..-273	YP_225822.1, cg1735 (-), eValue: 4.0E-49
jk0417	cspB			2.1E-09	4.4E-03	1.452000e+03	TGTGACGAAGATCACA	-124..-109	
jk0975	-			2E-08	4.3E-02	1.358000e+03	TGTGATTCAAACCACA	-46..-31	
jk0084	-			5.1E-08	1.1E-01	1.305000e+03	TGTGACGAATGTCACA	-126..-111	
jk0954	-			7.8E-08	1.7E-01	1.280000e+03	AGTGACATAGGACACA	-34..-19	
jk0201	clpB			1.6E-07	3.4E-01	1.236000e+03	TGTGGGGTAAATCACA	-135..-120	
jk0230	fabG1	OP_jk0230		1.8E-07	3.8E-01	1.230000e+03	TATGATGCAAGTCACA	-65..-50	YP_226651.1, cg2640 (benD), eValue: 2.0E-13
		Genes in operon: jk0228 (-), jk0229 (fadE2), jk0230 (fabG1)							
jk1618	whiB1			2E-07	4.3E-01	1.223000e+03	TGTGACATTCATCACA	-331..-316	YP_225059.1, cg0878 (whcE), eValue: 3.0E-33
jk0402	-	OP_jk0402		3.1E-07	6.5E-01	1.196000e+03	AGTGATACAAATCACT	-31..-16	YP_225295.1, cg1142 (-), eValue: 1.0E-96
		Genes in operon: jk0402 (-), jk0403 (-)							
jk0200	-	OP_jk0200		4E-07	8.5E-01	1.179000e+03	TGTGATTTACCCCACA	-153..-138	
		Genes in operon: jk0198 (-), jk0199 (-), jk0200 (-)							
jk0416	rpfA			4.6E-07	9.6E-01	1.170000e+03	TGTGATCTTCGTCACA	-290..-275	YP_225111.1, cg0936 (rpf1), eValue: 8.0E-44 YP_225201.1, cg1037 (rpf2), eValue: 2.0E-27
jk0948	-	OP_jk0948		5.5E-07	1.2E+00	1.158000e+03	TATGACCGAGAACACA	0..15	
		Genes in operon: jk0946 (-), jk0947 (-), jk0948 (-)							
jk0974	-			7.1E-07	1.5E+00	1.140000e+03	TGTGGTTTGAATCACA	-208..-193	
jk0425	serC			8.5E-07	1.8E+00	1.127000e+03	TGTGACGCAGATGACA	-156..-141	YP_225120.1, cg0948 (serC), eValue: 2.0E-166

GraphVis

Prediction mode!