BMC Systems Biology

Methodology article

Reliable transfer of transcriptional gene regulatory networks between taxonomically related organisms

Jan Baumbach*[1], Sven Rahmann[2] and Andreas Tauch[3]

Address: [1]International Computer Science Institute, Berkeley, CA 94704, USA, [2]Bioinformatics for High-Throughput Technologies, Technical University of Dortmund, D-44227 Dortmund, Germany and [3]Institute for Genome Research and Systems Biology, Center for Biotechnology, Bielefeld University, D-33594 Bielefeld, Germany

Email: Jan Baumbach* - jbaumbac@icsi.berkeley.edu; Sven Rahmann - Sven.Rahmann@tu-dortmund.de; Andreas Tauch - Andreas.Tauch@Genetik.Uni-Bielefeld.DE

* Corresponding author

Published: 15 January 2009

BMC Systems Biology 2009, **3**:8 doi:10.1186/1752-0509-3-8

Received: 27 October 2008
Accepted: 15 January 2009

This article is available from: http://www.biomedcentral.com/1752-0509/3/8

Abstract

Background: Transcriptional regulation of gene activity is essential for any living organism. Transcription factors therefore recognize specific binding sites within the DNA to regulate the expression of particular target genes. The genome-scale reconstruction of the emerging regulatory networks is important for biotechnology and human medicine but cost-intensive, time-consuming, and impossible to perform for any species separately. By using bioinformatics methods one can partially transfer networks from well-studied model organisms to closely related species. However, the prediction quality is limited by the low level of evolutionary conservation of the transcription factor binding sites, even within organisms of the same genus.

Results: Here we present an integrated bioinformatics workflow that assures the reliability of transferred gene regulatory networks. Our approach combines three methods that can be applied on a large-scale: re-assessment of annotated binding sites, subsequent binding site prediction, and homology detection. A gene regulatory interaction is considered to be conserved if (1) the transcription factor, (2) the adjusted binding site, and (3) the target gene are conserved. The power of the approach is demonstrated by transferring gene regulations from the model organism *Corynebacterium glutamicum* to the human pathogens *C. diphtheriae*, *C. jeikeium*, and the biotechnologically relevant *C. efficiens*. For these three organisms we identified reliable transcriptional regulations for ~40% of the common transcription factors, compared to ~5% for which knowledge was available before.

Conclusion: Our results suggest that trustworthy genome-scale transfer of gene regulatory networks between organisms is feasible in general but still limited by the level of evolutionary conservation.

Background

In the post genome era we observe a continuously growing, vast amount of sequenced organisms spread over all domains of life. Besides the identification and annotation of functional sites within the emerging nucleic acid sequences, an important task in molecular genetics, biotechnology, and human medicine is to unravel the regulation of these sites. DNA-binding transcription factors

(TFs) are the most important components of the cell's regulatory machinery [1]. They recognize specific operator sequences close-by the promoter regions of the controlled target genes, referred to as transcription factor binding sites (TFBSs), and thereby influence the amount of produced proteins. Although inevitable for the understanding of the cell's handling of changing environmental conditions, the wet-lab reconstruction of the resulting transcriptional regulatory networks is cost-intensive, time-consuming, and impossible to perform for any species separately [2,3]. Even for prokaryotic model organisms, such as *Escherichia coli* or *Corynebacterium glutamicum* the monumental task of deciphering transcriptional regulatory networks for whole species is far from being complete. The current knowledge is brought together and stored in reference databases, such as RegulonDB [4] and CoryneRegNet [5]; see [6] for a more detailed analysis of such platforms.

The gathered information about substantial parts of the transcriptional regulatory apparatus is used to study conserved network structures, sensing mechanisms, and to uncover hidden architectures behind gene regulatory networks [7,8]. In addition, specialized approaches, based on the evolutionary conservation of the responsible transcription factors and the controlled target genes, are used to transfer knowledge on gene regulatory networks between different organisms but aim to provide more general, qualitative conclusions (trends) across many species [9]. The main problem, however, is the neglect of the fact that orthologous regulators and target genes not necessarily are involved in conserved regulations. Another complicacy is the dependency on reliable homology detections. Other approaches utilize annotated transcription factor binding sites to compute mathematical models for further TFBS predictions; where the by far most widely used model for TFBSs are position weight matrices (PWMs) [10]. Here, the major intricacy lies in the comparatively low level of TFBS conservation between different organisms [11], even for essential factors such as the bacterial SOS response and DNA damage regulator LexA [12]. Hence, the consideration of PWM-based predictions apart from further evidence is not very meaningful. Moreover, there is a hidden problem with PWM calculations: The determination of the position to which a transcription factor binds is difficult and time-consuming. It is normally performed through electrophoretic mobility shift assays, DNAse footprinting, ChIP-to-chip assays, or mutations of putative TFBSs [13-15]. With all of these methods a precise identification that is accurate to one base pair is problematic. Furthermore, since TFs bind the double-stranded DNA it is a matter of interpretation which strand of the DNA sequence is annotated and stored in the database. This causes a practical problem when a motif from either strand based on approximate knowledge of its position is used for PWM construction.

Methods

In the past years, we extensively studied the transcriptional regulatory repertoire of the model organism *C. glutamicum* and other corynebacteria important in human medicine and biotechnology. We gathered all publicly available data, combined it with own wet-lab findings and developed the reference database and analysis platform CoryneRegNet [5,11,12,16]. Here we introduce an integrative bioinformatics approach that aims for a reliable transfer of gene regulatory interactions, which combines both of the above introduced major approaches: homology detection and DNA binding site prediction. Instead of studying general trends and conserved network motifs across hundreds of organisms we are interested in high-quality predictions with just few or even no false-positives for *C. diphtheriae*, *C. efficiens*, and *C. jeikeium* based on evidenced observations from the model organism *C. glutamicum*.

The success of the approach relies on the optimal interplay of the used bioinformatics components. On a very general level, the workflow is depicted in figure 1. An integrated database system is used for data fusion of annotated nucleotide and amino acid sequences together with evidenced transcriptional regulatory relationships and corresponding sequence features. By having all required data at hand we can re-adjust inaccurately determined TF binding sequences by shifting some motifs by some positions and by assigning a strand annotation, if necessary. Sequence motif discovery tools [17] may be utilized for that purpose or publicly available special purpose tools, such as MoRAine [18]. After computing a PWM for each transcriptional regulator of the model organism from the re-annotated TFBSs, motif matching tools (such as PoSSuMsearch [19] or MATCH [20]) are used to predict binding sites in the target organism assuming that the DNA-binding motif of the regulator is sufficiently conserved. This is done for any pair of the conserved regulators and target genes. It is obvious that the TF itself has to be conserved in the target organism; but the second condition (homologous target genes) is very important as well to provide reliable, high-quality predictions since a high-scoring PWM-based sequence match alone is not very meaningful [10]. This is the only way to reduce the huge amount of false-positives without further background knowledge even for comparatively restrictive thresholds that miss most of the true-positives. An application example illustrating the problem for corynebacteria may be found in ref. [11]. The detection of conserved, orthologous proteins based on the given amino acid sequences alone was a long-standing challenge in computational biology. It emerged that clustering approaches utilizing

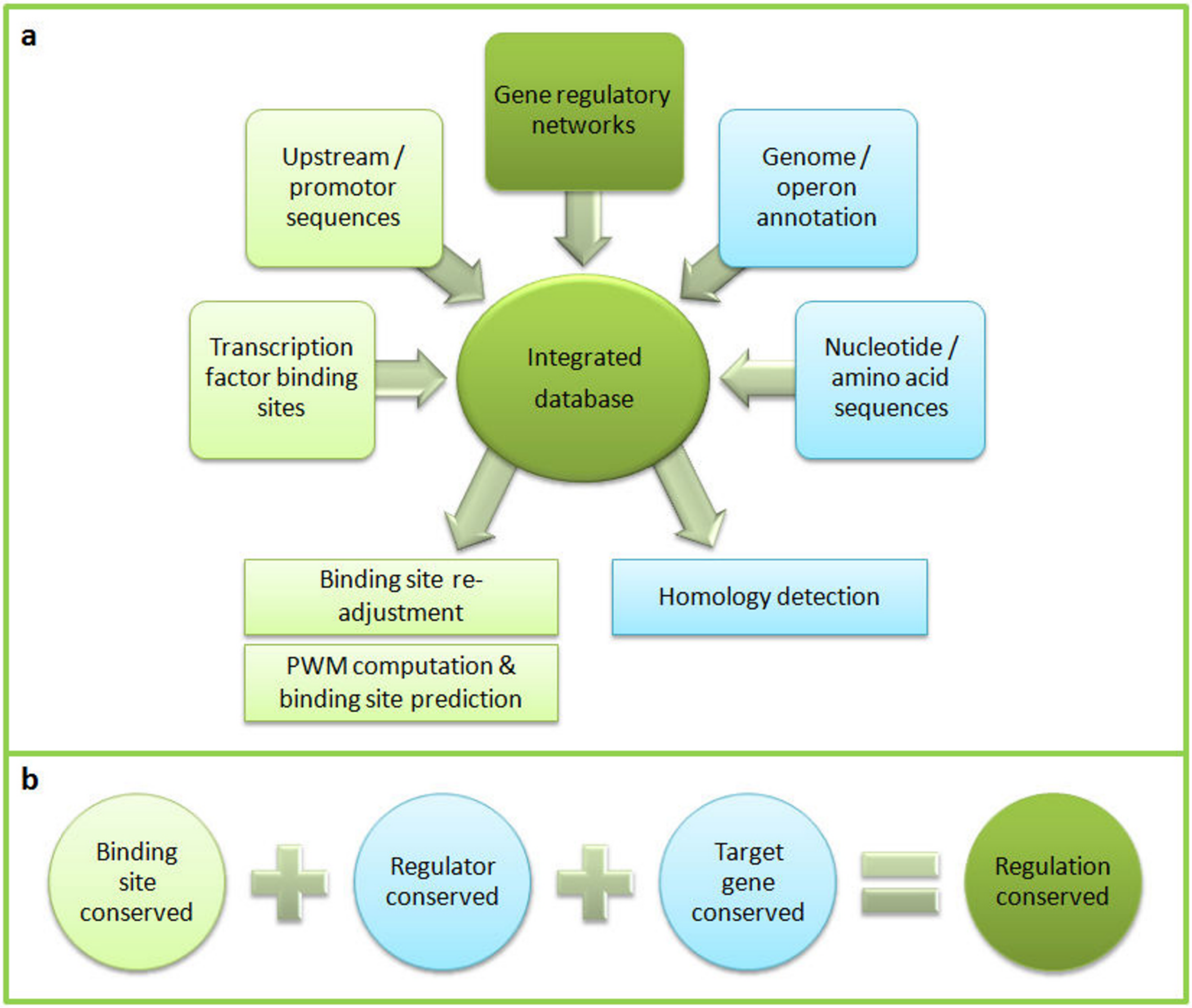
a
Gene regulatory networks
Upstream / promotor sequences
Genome / operon annotation
Transcription factor binding sites
Integrated database
Nucleotide / amino acid sequences
Binding site re-adjustment
PWM computation & binding site prediction
Homology detection
b
Binding site conserved
Regulator conserved
Target gene conserved
Regulation conserved

Figure 1
Scheme of the transfer workflow. a, Simplified structure of a typical transcriptional gene regulatory interaction database. Using genetic upstream sequences and transcription factor binding site annotations the TFBSs can be re-adjusted and modeled as PWMs for subsequent TFBS predictions. Sequence clustering tools can be applied to the stored genome annotation and gene/protein sequences to gain information about homologous genes/proteins. **b**, A regulatory interaction can be transferred from a model organism to a closely related species if the regulator as well as the target gene are orthologous and a matching TFBS can be found in the upstream sequence of the orthologous target gene.

BLAST[21]-based pairwise sequence similarity measures attack the problem comparatively well. Popular approaches are Markov clustering [22], spectral clustering [23] and graph cluster editing [24,25].

Results and discussion

To demonstrate our combined approach we decided to integrate MoRAine [18] for TFBS readjustment, PoSSuMsearch [19] for binding motif matching, and FORCE [25] for homology detection and applied it to corynebacterial data gained from CoryneRegNet [5] (see additional file 1). The goal was to use proven knowledge from the model organism *C. glutamicum* (*CG*) to predict reliable gene regulatory networks in *C. diphtheriae* (*CD*), which is a human pathogen, *C. jeikeium* (*CJ*) that is pathogen and extremely antibiotic resistant, and the sister organism of *CG* namely *C. efficiens* (*CE*), which is import for biotechnological production processes. First, the homology detection identified the common repertoire of conserved transcription factors between *CG* and *CD*, *CE*, and *CJ* (49, 77, and 31 regulators respectively). This observation of an average common TF set of ~70% fits well with the previously published study on the individual and common repertoire of corynebacteria [26]. For the next step in the pipeline the known transcription factor binding sites for each of the 69 characterized TFs of *CG* have been re-adjusted resulting in an average improvement of the mean information content of the underlying position frequency matrices of ~23% (see additional file 2). Subsequently, PWMs are computed and matched with all upstream sequences of orthologous target genes in *CD*, *CE*, and *CJ* to predict putative TFBSs. In the case of ambiguity of the homology prediction, the pipeline also considers the neighborhood of the respective genes in the target organism and chooses the candidate gene with the maximal number of homologous genes in the genetic surrounding. If the regulator, the target gene, and the binding site are conserved sufficiently, the regulation is considered to be conserved and added to the pipeline's output. Although these restrictions are very strict, we found 530 reliable novel gene regulations for *CD*, *CE*, and *CJ* and thereby increased the database content of the reference database CoryneRegNet considerably (factor 4.2). For the three organisms we identified reliable transcriptional gene regulations for ~40% of the common transcription factors, compared to ~5% for which knowledge was available before. Table 1 statistically summarizes the original and the transferred database content. Figure 2a exemplarily shows a network visualization of the PcaR regulon as known from *CG* compared to the transferred one of *CE*. It is obvious that all of the 11 target genes including the regulator itself are orthologous between *CG* and *CE*. Since the binding sites are conserved too, as indicated by the sequence logos in figure 2b/c, the pipeline considered all regulations as transferable from *CG* to *CE*. This observation fits well with the known PcaR regulon of *C. efficiens* [27].

The whole data analysis procedure is very time-efficient (< 1 min computing time) and we added the new datasets into the CoryneRegNet 5.0 database for public access. Although it is very likely that our predictions are highly reliable we separated the evidenced from the predicted datasets within the front-end and now provide two subversions of CoryneRegNet.

In the application described above, we combined the software packages MoRAine, FORCE, and PoSSuMsearch into a data fusion workflow that is responsible for the data transfer between model and target organisms. One could also think about a combination of other computational biology tools designed for the same purpose. For instance FORCE could be replaced by TribeMCL [22] and PoSSuMsearch by MATCH [20]. We decided to use PoSSuMsearch since it provides statistically sound p-value computations within reasonable response times [19]. FORCE has been included into the pipeline since it has been shown to outperform other clustering approaches on prokaryotic datasets [25].

With Regulogger [28], Alkema *et al*. presented a pipeline that mainly focuses on the reconstruction of conserved regulatory networks for *Staphylococcus aureus*. Regulogger mainly concentrates on the detection and characterization of conserved sequences in promoter regions (phylogenetic footprinting) of orthologous genes. For the determination of these orthologous genes, the COG database [29] is used, which is impracticable at least for our target species since no COG annotations are available for other corynebacteria than *C. glutamicum*. In the near future, novel ultra-fast sequencing technologies will provide much more data on organisms that are not included in COG. In [30] and a subsequent follow-up study [31], the TRACTOR_DB database [32] was used to identify conserved regulatory interactions between *E. coli* and thirty gamma-proteobacteria. Here, pairwise BLAST Bidirectional Best Hits (BBHs) are used directly to identify orthologous genes, which may result in suboptimal predictions [25]. However, the presented results support our strategy to combine orthology information with binding site detection. Besides, the conservation of gene regulatory networks between corynebacterial organisms has never been studied before and the integration of the presented results into the CoryneRegNet reference database provides a powerful tool for further network studies.

The current study is strongly limited by the level of phylogenetic conservation between the reference organism and the target species. Since remodeling of transcriptional gene regulation is a crucial strategy used by bacteria to evolve and regulate novel biochemical features, (1) specific regulatory pathways may have been altered and (2) unique transcriptional regulatory mechanisms may have been developed. The following three limitations with our approach for inter-species network transfer arouse: (1) Interactions that do not evolve in the reference species cannot be detected in a specific target organism. (2) Utilizing pure sequence-based similarities for *in silico* orthology detections neglect that proteins with comparably high overall amino acid sequence similarity may have different specific functions within the cell although they are predicted as putative homologs and vice versa. (3) In our approach, we assume that a conserved transcription factor in the target organism interacts with a binding site that is very similar to that in the reference organism. This is not necessarily the case and may result in both, false positive and false negative predictions. However, the rapidly increasing amount of fully sequenced organisms without much background knowledge about their gene regulatory repertoire strongly restricts our alternatives to computa-

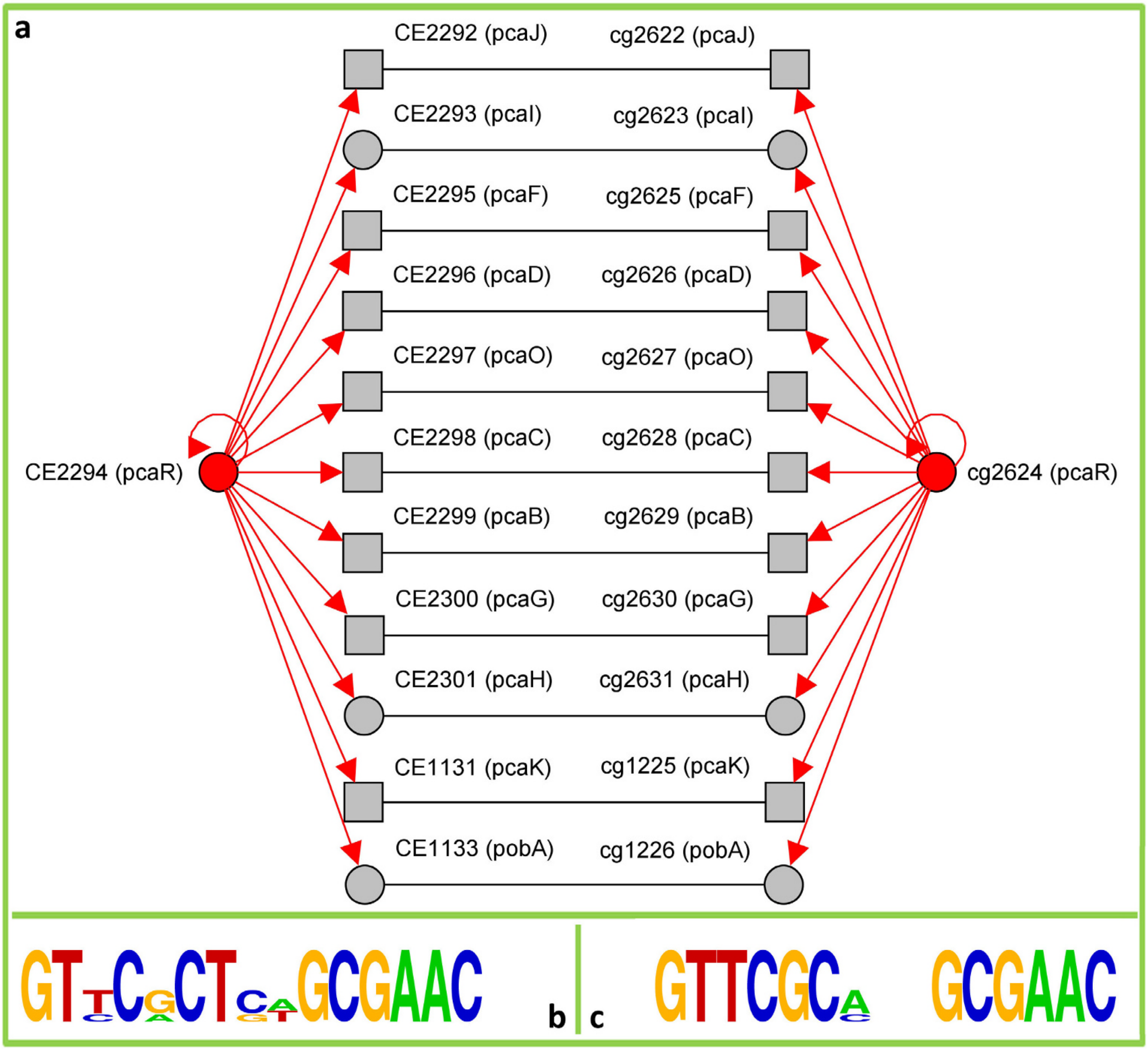

Figure 2
Illustration of the gene regulatory network for PcaR. **a**, A comparative visualization of the known gene regulatory network of PcaR in the model organism *C. glutamicum* (right) transferred to *C. efficiens* (left). Nodes correspond to genes, directed (red) edges to negative transcriptional regulatory interactions, and undirected (black) edges to a sequence-based similarity that indicates a putative homology. **b/c**, The sequence logos computed from the PcaR binding sites in *C. efficiens* (predicted, left) and in *C. glutamicum* (evidenced, right).

tional methods that utilize sequence-based evolutionary conservation.

Conclusion

Besides the immediate advantages for the worldwide medical and biotechnological corynebacteria community, we anticipate the presented results to be a starting point for an integrated analysis of gene regulatory networks in the light of a combined analysis of orthologous genes conjointly with conserved DNA binding sites. Although we tested the presented strategy with corynebacteria, it is of general interest and can be applied to many other organisms as well. We conclude that trustworthy transfers of gene regulatory networks between organisms on a

Table 1: Comparison of the original and the transferred database content of CoryneRegNet.

	TFs	TFs_C	TFs^K		$\text{TFs}_C{}^K$	Regulations	
CG	128		69			530	
			original	transferred	transferred	original	transferred
CD	63	49 (77.8%)	2 (3.2%)	20 (×10)	20 (40.1%)	46	193 (×4.2)
CE	103	77 (74.8%)	5 (4.9%)	28 (×5.6)	28 (36.4%)	64	348 (×5.4)
CJ	55	31 (56.4%)	1 (1.8%)	13 (×13)	13 (41.9%)	51	150 (×2.9)
Av		69.6%	3.3%	x9.5	39.7%		x4.2

Abrev.: CG = The model organism *Corynebacterium glutamicum*, CD = *C. diphtheriae*, CE = *C. efficiens*, CJ = *C. jeikeium*, Av = Average, TFs = Transcription factors, TFs_C = Common transcription factors with *C. glutamicum*, TFs^K = Transcription factors with knowledge, $\text{TFs}_C{}^K$ = Transcription factors common with *C. glutamicum* with knowledge, Regulations = Transcriptional regulatory interactions, original = Original database content, transferred = Database content after network transfer. Percentages/factors: TFs_C = relative to column TFs, TFs^K/original = relative to column TFs, TFs^K/transferred = relative to column TFs^K/orig., $\text{TFs}_C{}^K$ = relative to column TFs_C, Regulations/transferred = relative to column Regulations/original.

genome-scale are feasible but still limited by the level of evolutionary conservation.

Availability and requirements

Project name: CoryneRegNet 5.0

Project home page: http://coryneregnet.cebitec.uni-bielefeld.de/v5

Operating system(s): Platform independent

Programming language: PHP, Java 6

License: Academic Free

License (AFL)

Any restrictions to use by non-academics: No.

Authors' contributions

JB designed and implemented the data analysis and transfer pipeline. SR and AT co-supervised the project, host the web services, and curate the database content.

Appendix

Here, we briefly introduce the integrated bioinformatics tools that have been utilized for this article.

MoRAine – Regulatory binding site re-adjustment

MoRAine is a software that re-assesses and re-annotates transcription factor binding sites. Each TFBS sequence with experimental evidence underlying a PWM model is compared against each other. MoRAine heuristically solves a combinatorial optimization problem to readjust TFBSs by possibly switching their strands and shifting them a few positions to the left or to the right in order to maximize the mean information content of the resulting PWM. In [18] we validated and confirmed the improvement of the PWM-based TFBS prediction performance for *E. coli* by using MoRAine as pre-processing step prior to PWM computation. For this article, we applied MoRAine to adjust corynebacterial TFBSs for computing the PWM model being the input for PoSSuMsearch.

PoSSuMsearch – Statistically sound binding site prediction

The *in silico* prediction of TFBSs is a long-standing challenge in computational biology and several software tools exist for this purpose. Here we used PoSSuMsearch. It provides a combination of (non-permuted) lookahead scoring and efficiently searching an enhanced suffix array that previously has been created from corynebacterial upstream sequences. The scores of putative matches are compared to a threshold that is computed based on the tolerable frequency of hits in random sequences (p-value) by an efficient and exact lazy-evaluation method [19]. To our knowledge, PoSSuMsearch is the only available bioinformatics tool that provides a statistically sound TFBS match evaluation together with reasonable response times. We used PoSSuMsearch to predict putative binding sites for conserved corynebacterial regulators.

FORCE – Transitivity clustering of protein sequences

Another long-standing challenge in bioinformatics is the detecting of homologous proteins based on amino acid sequence similarity. For this purpose, we recently developed a clustering approach based on weighted graph cluster editing (or weighted transitive graph projection). With FORCE, we presented a heuristic that solves the respective NP-hard graph-modification problem. In [25], we demonstrated the ability of FORCE to cluster hundreds of thousands of protein sequences efficiently and accurately. Our evaluations with gold standard databases show that it outperforms other tools at least in terms of accuracy. In particular its ability to handle huge datasets makes it the ideal candidate for the study introduced in this article. We used FORCE as homology detection software to identify conserved transcriptional regulators as well as orthologous target genes.

Additional material

Additional File 1

Data fusion workflow as used for the presented study with CoryneRegNet. *We used all validated datasets of* Corynebacterium glutamicum *from the CoryneRegNet database. All known transcription factor binding sites were re-adjusted by using MoRAine (no position shifts, but strand annotation; method: Cluster growing/Motif-seed similarity) for subsequent computations of position weight matrices (PWMs). We utilized PoSSuMsearch to scan the upstream/promoter sequences of all transcription units of* C. efficiens, C. diphtheriae, *and* C. jeikeium, *extracted from CoryneRegNet, to scan for putative transcription factor binding sites by using the MoRAine-adjusted binding sites of* C. glutamicum. *All amino acid sequences of* C. efficiens, C. diphtheriae, C. glutamicum, *and* C. jeikeium *were extracted from CoryneRegNet and grouped into clusters of orthologous/conserved corynebacterial proteins by using the FORCE software. We consider two genes as orthologous/conserved (1) if the corresponding proteins are in the same FORCE cluster and (2) if at least one of the surrounding genes is also "FORCE-conserved". We consider a gene regulatory interaction as conserved between* C. glutamicum *and another corynebacterium if (1) the transcription factor is conserved (2) its binding sites are conserved, and (3) the putative target genes are conserved as well. The corresponding gene regulation is added to the CoryneRegNet 5.0 p database. Refer to the CoryneRegNet web site for an interactive version of this picture including links to the corresponding tools.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-3-8-S1.jpeg]

Additional File 2

Original and re-adjusted corynebacterial transcription factor binding sites. *The RAR-file contains a list of files, two for each regulator of* C. glutamicum. *They contain the original and the MoRAine-adjusted transcription factor binding sites in FASTA-format.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-3-8-S2.rar]

Acknowledgements

JB thanks the German Academic Exchange Service (DAAD) for financial support.

References

1. Teichmann SA, Babu MM: **Gene regulatory network growth by duplication.** *Nat Genet* 2004, **36(5):**492-496.
2. Matic I, Taddei F, Radman M: **Survival versus maintenance of genetic stability: a conflict of priorities during stress.** *Res Microbiol* 2004, **155(5):**337-341.
3. Pabo CO, Sauer RT: **Transcription factors: structural families and principles of DNA recognition.** *Annu Rev Biochem* 1992, **61:**1053-1095.
4. Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muniz-Rascado L, Martinez-Flores I, Salgado H, *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:D120-124.
5. Baumbach J: **CoryneRegNet 4.0 A-reference database for corynebacterial gene regulatory networks.** *BMC Bioinformatics* 2007, **8(1):**429.
6. Baumbach J, Tauch A, Rahmann S: **Towards the integrated analysis, visualization, and reconstruction of microbial gene regulatory networks.** *Briefings in Bioinformatics* 2008 in press.
7. 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(4):**1108-1122.
8. Balaji S, Iyer LM, Aravind L, Babu MM: **Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks.** *J Mol Biol* 2006, **360(1):**204-212.
9. Madan Babu M, Teichmann SA, Aravind L: **Evolutionary dynamics of prokaryotic transcriptional regulatory networks.** *J Mol Biol* 2006, **358(2):**614-633.
10. Rahmann S, Müller T, Vingron M: **On the power of profiles for transcription factor binding site detection.** *Statistical Applications in Genetics and Molecular Biology* 2003, **2(1):**Article 7.
11. 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* 2006, **3(2):**24.
12. Baumbach J, Wittkop T, Rademacher K, Rahmann S, Brinkrolf K, Tauch A: **CoryneRegNet 3.0–an interactive systems biology platform for the analysis of gene regulatory networks in corynebacteria and Escherichia coli.** *J Biotechnol* 2007, **129(2):**279-289.
13. Galas DJ, Schmitz A: **DNAse footprinting: a simple method for the detection of protein- DNA binding specificity.** *Nucleic Acids Res* 1978, **5(9):**3157-3170.
14. Hellman LM, Fried MG: **Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions.** *Nature Protocols* 2007, **2(8):**1849-1861.
15. Sun LV, Chen L, Greil F, Negre N, Li TR, Cavalli G, Zhao H, Van Steensel B, White KP: **Protein-DNA interaction mapping using genomic tiling path microarrays in Drosophila.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100(16):**9428-9433.
16. Baumbach J, Brinkrolf K, Czaja LF, Rahmann S, Tauch A: **CoryneRegNet: An ontology-based data warehouse of corynebacterial transcription factors and regulatory networks.** *BMC Genomics* 2006, **7(1):**24.
17. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, *et al.*: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nature Biotechnol* 2005, **23(1):**137-144.
18. 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* 2008, **5(2):**91.
19. 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.
20. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCH: A tool for searching transcription factor binding sites in DNA sequences.** *Nucleic Acids Res* 2003, **31(13):**3576-3579.
21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17):**3389-3402.
22. Enright AJ, Kunin V, Ouzounis CA: **Protein families and TRIBES in genome sequence space.** *Nucleic Acids Res* 2003, **31(15):**4632-4638.
23. Paccanaro A, Casbon JA, Saqi MA: **Spectral clustering of protein sequences.** *Nucleic Acids Res* 2006, **34(5):**1571-1580.

24. Rahmann S, Wittkop T, Baumbach J, Martin M, Truss A, Böcker S: **Exact and heuristic algorithms for weighted cluster editing.** *Comput Syst Bioinformatics Conf* 2007, **6:**391-401.
25. Wittkop T, Baumbach J, Lobo FP, Rahmann S: **Large scale clustering of protein sequences with FORCE – A layout based heuristic for weighted cluster editing.** *BMC Bioinformatics* 2007, **8(1):**396.
26. Brune I, Brinkrolf K, Kalinowski J, Pühler A, Tauch A: **The individual and common repertoire of DNA-binding transcriptional regulators of Corynebacterium glutamicum, Corynebacterium efficiens, Corynebacterium diphtheriae and Corynebacterium jeikeium deduced from the complete genome sequences.** *BMC Genomics* 2005, **6(1):**86.
27. Brinkrolf K, Brune I, Tauch A: **Transcriptional regulation of catabolic pathways for aromatic compounds in Corynebacterium glutamicum.** *Genet Mol Res* 2006, **5(4):**773-789.
28. Alkema WB, Lenhard B, Wasserman WW: **Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus.** *Genome Res* 2004, **14(7):**1362-1373.
29. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al.*: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
30. Espinosa V, Gonzalez AD, Vasconcelos AT, Huerta AM, Collado-Vides J: **Comparative studies of transcriptional regulation mechanisms in a group of eight gamma-proteobacterial genomes.** *J Mol Biol* 2005, **354(1):**184-199.
31. Gonzalez Perez AD, Gonzalez Gonzalez E, Espinosa Angarica V, Vasconcelos AT, Collado-Vides J: **Impact of Transcription Units rearrangement on the evolution of the regulatory network of gamma-proteobacteria.** *BMC Genomics* 2008, **9:**128.
32. Perez AG, Angarica VE, Vasconcelos AT, Collado-Vides J: **Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes.** *Nucleic Acids Res* 2007:D132-136.