Published online 10 August 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 22* ***7877–7884***
doi:10.1093/nar/gkq699

SURVEY AND SUMMARY

On the power and limits of evolutionary conservation—unraveling bacterial gene regulatory networks

Jan Baumbach[1,2,3,*]

[1]Algorithms Group, International Computer Science Institute, Berkeley, USA, [2]Computational Systems Biology Group, Max Planck Institute for Informatics and [3]Cluster of Excellence for Multimodel Computing and Interaction, Saarland University, Germany

Received May 27, 2010; Revised June 28, 2010; Accepted July 22, 2010

ABSTRACT

The National Center for Biotechnology Information (NCBI) recently announced '1000 prokaryotic genomes are now completed and available in the Genome database'. The increasing trend will provide us with thousands of sequenced microbial organisms over the next years. However, this is only the first step in understanding how cells survive, reproduce and adapt their behavior while being exposed to changing environmental conditions. One major control mechanism is transcriptional gene regulation. Here, striking is the direct juxtaposition of the handful of bacterial model organisms to the 1000 prokaryotic genomes. Next-generation sequencing technologies will further widen this gap drastically. However, several computational approaches have proven to be helpful. The main idea is to use the known transcriptional regulatory network of reference organisms as template in order to unravel evolutionarily conserved gene regulations in newly sequenced species. This transfer essentially depends on the reliable identification of several types of conserved DNA sequences. We decompose this problem into three computational processes, review the state of the art and illustrate future perspectives.

INTRODUCTION

The National Center for Biotechnology Information (NCBI) recently announced '1000 prokaryotic genomes are now completed and available in the Genome database'. In fact, we may download the genome annotations of 1024 fully sequenced microbial organisms from the NCBI database (1). Thanks to the next-generation sequencing techniques, the cost of DNA sequencing was reduced by over two orders of magnitude and has thus become a routine and widespread method to unravel the genetic repertoire of numerous species (2,3). The increasing trend will provide us with thousands of sequenced organisms over the next years. This genomic revolution in molecular biology leaves us with complete genomes of numerous microbes with varied ecological, economic and medical significance (4).

The availability of genome sequences, however, is only the first step in understanding how cells survive, reproduce and adapt their behavior while being exposed to varying environmental conditions. One major control mechanism is transcriptional gene regulation. The most important components of the cell's transcriptional regulatory apparatus are the so-called transcription factors (TFs)—DNA-binding proteins that are able to detect intra- and extracellular signals. By binding to operator sequences, the transcription factor binding sites (TFBSs), they repress or stimulate the expression of other genes (target genes, TGs) and thereby decisively influence genetic programs like growth, survival and reproduction (5). Supplementary Figure S1 illustrates this transcriptional machinery. Depending on the surrounding and internal conditions of a cell, certain fractions of the total set of TFs are operating to control the expression of their TGs. Some regulators only control the expression of a single gene, whereas others organize the activation or repression of numerous TGs. Regulatory networks emerge. They are modeled as graphs, where nodes correspond to genes and directed edges represent transcriptional regulatory interactions. The reconstruction of these networks, i.e. the identification of the spatial and temporal regulatory interactions between TFs and their targets, is one of the most important goals in molecular systems biology (6).

*To whom correspondence should be addressed. Tel: +49 681 302 70880; Fax: +49 681 9325 399; Email: jbaumbac@mpi-inf.mpg.de

Nowadays, high-throughput experimental techniques exist for the wet-lab reconstruction of gene regulatory networks. The two primary methods, genome-wide measurement of mRNA expression levels and the identification of TFBS locations, are widely available. The application of microarray and RNA-seq technology has opened the way to investigate organism-wide gene expression under different conditions in order to provide hypotheses about putative transcriptional gene regulatory interactions (7). Especially, studying genetic expression in response to the deletion of TF-encoding genes has been successfully utilized to identify potential TGs for numerous TFs in many microbial organisms. Subsequent identification of the TF binding-site location in the promoter regions of the putative TGs provides further evidence for the respective TF-TG interactions. Wet-lab determination of TFBS locations is done by electrophoretic mobility shift assays (EMSA) (8), DNAse footprinting (9), ChIP-chip (10) or ChIP-seq (11). By combining gene expression studies and TFBS location analysis, transcriptional gene regulatory interactions are reconstructed and the emerging networks are stored in publicly available databases (12,13). For prokaryotes, popular reference databases are RegTransBase (14), RegulonDB (15) and EcoCyc (16) for *Escherichia coli*, DBTBS (17) for *Bacillus subtilis*, MtbRegList (18) for *Mycobacterium tuberculosis*, PRODORIC (19) mainly for *Pseudomonas aeruginosa* but also *E. coli* and *B. subtilis* and CoryneRegNet (20) for corynebacteria (mainly *Corynebacterium glutamicum*).

Although inevitable for understanding the behavior and the complexity of microbial cells, the reconstruction of transcriptional gene regulatory networks is far from being complete. Even for the model organism *E. coli*, with the largest currently available experimentally validated knowledge of any free-living organism (21), we have some information about the transcriptional regulation of only around one-third of the genes (15). Network reconstruction and standardized data access is complicated by several problems: technical and procedural difficulties comprise, for example, the fabrication of TF-deletion mutants, the noise in gene expression data, the identification of concealed combinatorial effects caused by co-acting TFs and the determination of TFBS locations accurately to one base pair. Consequently, economic aspects arise: Replicated experiments are inevitable to provide statistical significance but drastically increase the amount of necessary temporal and monetary resources. Finally, successfully discovered gene regulatory interactions are published in scientific journals. This is an organizational issue since it requires curation teams to find and extract this knowledge from the literature manually instead of having it available in online repositories for direct, well-structured data access (22). In the light of these technical, monetary and structural difficulties, we conclude that a wet-lab reconstruction of gene regulatory networks is impossible to perform for any sequenced prokaryote separately. Striking is the direct juxtaposition of the six abovementioned reference repositories for *E. coli*, *B. subtilis*, *M. tuberculosis*, *P. aeruginosa* and *C. glutamicum* to the 1000 microbial genomes, which we may download from the NCBI. Recent advances in high-throughput genome sequencing will further widen this gap drastically.

However, recently developed bioinformatics approaches have proven to be helpful here. The similarity of the gene regulatory networks between two organisms correlates with the grade of evolutionary and taxonomical conservation between them (23,24). Hence, the main idea is to use the gene regulatory network of one of the few reference organisms as template (source) in order to unravel evolutionarily conserved gene regulations in newly sequenced species (targets). This transfer of transcriptional regulatory interactions between source and target organisms essentially depends on the reliable discovery of conserved DNA sequence patterns. In the following, we decompose this process into three computational aspects: (i) orthology detection, (ii) TF binding-site prediction and (iii) a combination of both. We investigate recently published studies to illustrate the state of the art. Finally, we will identify open challenges and suggest future directions.

NETWORK TRANSFER

Conserved genes

In first studies, scientists concentrated on the conservation of the most apparent genetic elements: the genes. The assumption is that orthologous transcription factors regulate orthologous target genes. Babu *et al.* (25) used bi-directional best BLAST hits (BBHs) (26,27) as orthology detection to transfer the gene regulatory network of *E. coli* (112 TFs, 755 TGs, 1295 interactions) to 175 fully sequenced prokaryotes. They claim that 'it is now generally accepted that in the majority of cases' a transcriptional gene regulatory interaction is conserved if the participating genes, i.e. the TF and the TG, are conserved. For eukaryotes, Yu *et al.* (28) came to the same conclusion. They found this method to be 'fairly robust' in their studies.

However, this topic is discussed controversially in the scientific community. Price *et al.* studied whether putative orthologous TFs, identified by BBHs, have evolutionary conserved functions, i.e. whether they regulate conserved TGs (29). They showed that, especially for distantly related species, TFs identified as orthologs via BBHs often have different functions, respond to different signals and regulate different TGs. In conclusion, Price *et al.* finally suggest utilizing phylogenetic trees for the identification of putative orthologs, rather than BBHs.

Figure 1 illustrates the general problem by means of the two regulators DtxR (30) and PcaR (31) of *C. glutamicum* and the taxonomically closely related organism *C. efficiens*. The regulons as well as putative orthologous genes of the two TFs were extracted from the CoryneRegNet database (32,33). All 12 TGs of PcaR are conserved in both organisms. In contrast, DtxR regulates 64 genes in *C. glutamicum* but only 27 in *C. efficiens*. From these TGs, only nine are clearly evolutionarily conserved, the others cannot be assigned unambiguously to exactly one putative homologous partner gene in the other organism (34).

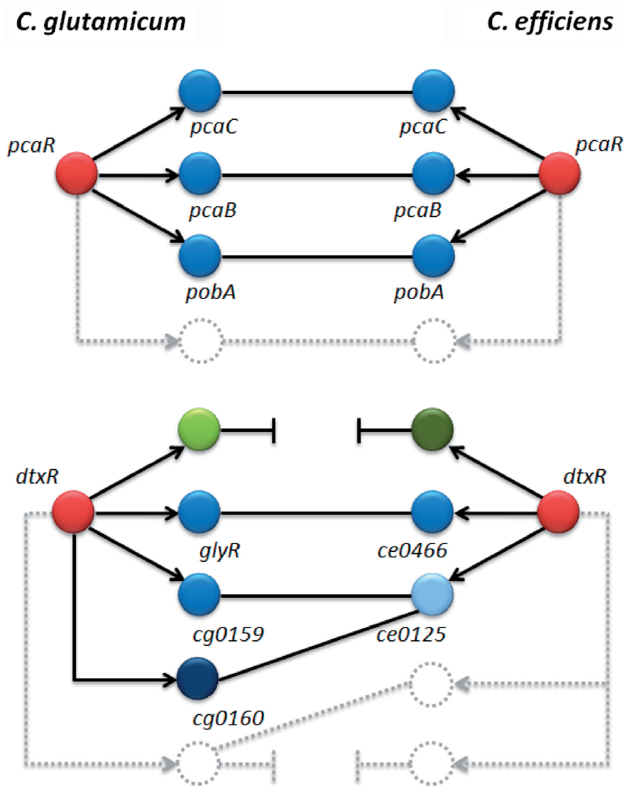

Figure 1. Illustration of the orthology detection problem. To demonstrate this problem, we compare the regulons of the transcription factors *pcaR* and *dtxR* of *Corynebacterium glutamicum* (CG) and *Corynebacterium efficiens* (CE). The red nodes represent the respective regulators, the others their target genes. Directed edges correspond to transcriptional regulatory interactions. Undirected edges symbolize putative orthologies due to sequence-based similarity. While for *pcaR* all 12 target genes are conserved in both organisms, for *dtxR* multiple problems occur: *dtxR* regulates 64 genes in CG but only 27 in CE. From these target genes, only nine are clearly evolutionarily conserved, i.e. one-to-one relationship, such as *glyR* and *ce0466*. The others are either inhomologous (green nodes) or show multiple, ambiguous sequence-based similarities, i.e. one-to-many or many-to-many relationship; *cg0159*, *cg0160* (CG), and *ce0125* (CE) may serve as an example here.

In a another study about the human pathogen *M. tuberculosis*, Balazsi *et al.* (35) reconstructed the largest known gene regulatory network of this organism. All interactions have been integrated from the MtbRegList database, by extensive literature research, and by transferring data from *E. coli* to *M. tuberculosis* based on the evolutionary conservation of TFs and TGs in both organisms. For the last approach, they found that only 54 of the 410 orthology-based links match with the 581 interactions known from the literature. Additionally, Venancio and Aravind recently observed a lack of successfully identified potential transcription factor encoding genes (4), at least in *M. tuberculosis*. Different publications mention different numbers of TFs [150 and 194 in refs. (35,36)] while Vanancio and Aravind's 'careful profile-based searches' suggest 235 TFs (4). In contrast, Wilson *et al.* predicted 172 TFs for *M. tuberculosis* by using their profiles to construct the DBD database (37). In any case, we still do not even have much knowledge about the 150 definite TFs. Besides, note that this method strongly depends on highly accurate genome annotations. These are often based on computer predictions and subsequently uploaded to the NCBI genome database, a risky procedure. For instance, Bakke *et al.* (38) compared three different genome annotation systems and found that only 47.7% of the predicted protein-coding genes were covered by all three systems. Furthermore, most approaches concentrate on the identification of conserved genes amongst different organisms but neglect genome shuffling and reorganization effects. Here, a major problem is gene duplication resulting in multiple putative orthologs in the target genome. One could avoid this by incorporating surrounding genes in the comparison, for instance with gene cluster detection; see e.g. (39,40).

We conclude that utilizing information about conserved genes between different organisms may be enough for studying general evolutionary dynamics of gene regulatory networks; but using this information alone may lack reliability for detailed reconstructions and subsequent analyses of the cell's ability to organize dynamic behavior by means of finely controlling gene expression. Still, the identification of putative orthologous genes is one major step toward an automatic inter-species network transfer.

Conserved binding sites

A different approach is to utilize knowledge about identified TF binding sites in the source organism. These TFBSs may be converted into computational models for subsequent profile-based predictions of gene regulatory interactions of orthologous TFs in the target organisms. This process is complicated by the comparably small length of the TFBSs (5–50 bp) resulting in computational difficulties regarding the statistical significance of detected putative TFBSs (41). One disadvantage is the necessity of knowledge about TFBSs for the respective TFs in the source organism. However, the main advantage is the potential to unravel regulatory interactions in target organisms that were not previously observed in the source organism, i.e. the TGs do not have to be conserved. However, it is known that orthologous TFs may regulate orthologous TGs through divergent TFBSs, especially in taxonomically distantly related organisms (4). In Figure 2, again the transcriptional regulators PcaR (for *C. glutamicum* and *C. efficiens*) and DtxR (for *C. glutamicum*, *C. efficiens*, *C. diphtheriae* and *C. jeikeium*) are used to illustrate the problem of TFBS conservation. To provide some numbers: Baumbach *et al.* (42) employed known TFBSs to move with the regulatory network of DtxR from *C. glutamicum* to the human pathogen *C. diphtheriae*. For the later bacterium, they pretended not to know the DtxR binding sites and target genes in order to evaluate the bioinformatics prediction performance. For a restrictive significance threshold they found three out of 32 TFBSs (9%) in *C. diphtheriae* with no false positive and, for an intermediate threshold, seven TFBSs (22%) with one false positive. With a comparably

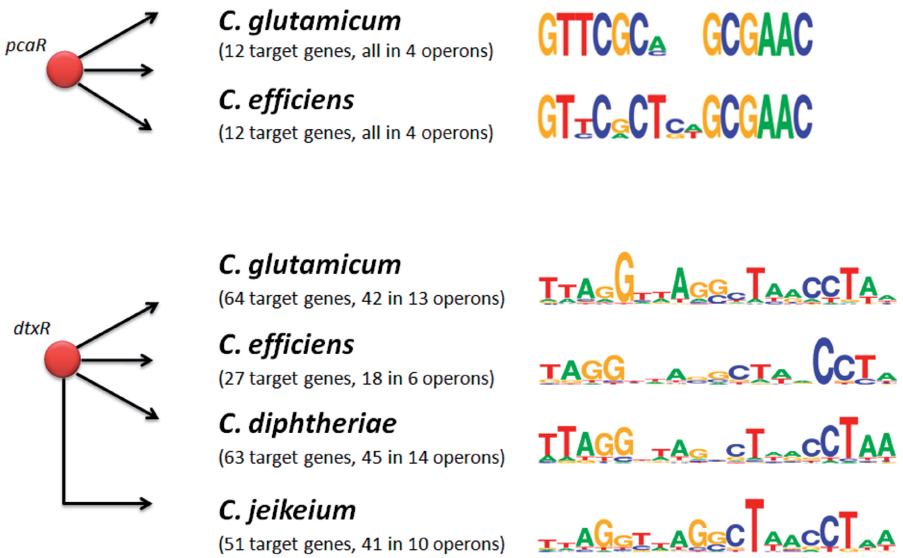
pcaR
C. glutamicum
(12 target genes, all in 4 operons)
C. efficiens
(12 target genes, all in 4 operons)
dtxR
C. glutamicum
(64 target genes, 42 in 13 operons)
C. efficiens
(27 target genes, 18 in 6 operons)
C. diphtheriae
(63 target genes, 45 in 14 operons)
C. jeikeium
(51 target genes, 41 in 10 operons)

Figure 2. Illustration of the binding-sites detection problem. Here, we demonstrate the problem when moving from one organism to another by investigating the evolutionary conservation of transcription factor binding sites. As in Figure 1, we study the transcriptional regulators *pcaR* in *Corynebacterium glutamicum* (CG) and *Corynebacterium efficiens* (CE) as well as the regulator *dtxR* in CG, CE, *Corynebacterium diphtheriae* (CD) and *Corynebacterium jeikeium* (CJ). For *pcaR*, all 12 target genes are conserved as are the transcription factor binding sites (TFBSs), depicted by the sequence logos (74) at the right side. It is more complicated with *dtxR*. The regulons are not conserved, ranging from 27 target genes in CE to 64 targets in CG. The sequence logos for DtxR are also slightly different for CG, CE, CD, and CJ.

weak significance cutoff, they found 24 of the 32 TFBSs (75%) but paid a high price: 59 false positives. The statistics suggest that generally one should be able to find more true regulated TGs in *C. diphtheriae* (coverage), but we need to keep in mind that we are using the *C. glutamicum* TFBSs for predictions of binding sites in *C. diphtheriae*, where the DtxR binding motif is slightly different (see Figure 2). This is the price to be paid for moving from one organism to a different one with having the source organism's TFBSs as only information source. On top of that, real TFBSs do not behave according to probabilistic sequence models, and therefore the expected coverage prediction can only be true up to an order of magnitude (42). Another well-studied example for the evolutionary divergence of binding sites is the DNA damage-response regulator LexA. Its TFBSs, termed SOS box, are similar among taxonomically closely related species but different in others (43). Hence, for instance, the LexA regulons of *C. glutamicum* [48 TGs (44)] and *E. coli* [25 TGs (15)] only share six orthologous genes (6).

To sum up, TFBS prediction has the potential to provide us with knowledge about new gene regulatory interactions. However, we still need to know some TFBSs of a particular TF from the source organism. The major drawback is the poor trade-off between sensitivity and specificity.

Combining both, orthologous genes and conserved binding sites

With the insights gained through the above-introduced studies, we now concentrate on recent approaches that combine both, the identification of orthologous genes as well as the detection of conserved TF binding sites. Under the assumption that a TF-DNA binding within the promoter region of a TG generally affects the co-transcription and co-expression of all genes within the TG's operon, we may further extend our predictions. For a given conserved TF-TG regulation, we extend the set of TGs in the respective regulon of the target organism by all genes within the TG's operon (35). Apparently, we need careful operon predictions for this step; refer to (45) for a summary of the state of the art. An overview of the combined inter-species network transfer procedure that utilizes orthology detection and TFBS prediction together with operon extension is depicted in Supplementary Figure S2. We start with the genome annotation data for source and target organisms. Together with the template regulatory network and the known TFBSs from the source organism, we may compute potentially conserved TFs, TGs and TFBSs. In the next step, we assume a TF-TG regulatory interaction to be conserved if the TF, the TG and the TFBS are evolutionarily conserved. In addition, if a TG encodes for the first gene within an operon in the target organism, we extend the regulon of the TF-ortholog by all the genes within the operon. The main advantage of combining TFBS prediction with orthology detection is the drastically decreased false positive rate.

The bioinformatics tool Regulogger serves as our first example here, where specificity was increased up to 25-fold over approaches that solely rely on the identification of conserved TF binding sites (46). Alkema *et al.* predicted 125 conserved regulogs in *Staphylococcus aureus*, i.e. sets of co-regulated genes with conserved regulatory sequence across multiple species. They utilized the COG (47) database as orthology detection and a combination of Gibbs Sampling (48) and the TFBS (49) software for binding site predictions. The promoter region is defined as the sequence 250 bp upstream of a putative target gene. Operons are defined as genes with the same orientation and with an intergenic distance of <50 bp, following a suggestion from ref. (50). Note that using the COG database may be impracticable for future studies since COG annotations for newly sequenced species are not available.

In ref. (51) and a subsequent follow-up study (52), the TRACTOR_DB (53) database was used to study conserved regulatory networks in 30 gamma-proteobacteria by using the network of *E. coli* as template. The number of predicted interactions (regulons) ranges from 6 (3) for *Xanthomonas axonopodis* to 1901 (69) for *Salmonella typhimurium*. Here, BBHs were utilized for the detection of orthologous genes. The promoter region was defined as the sequence ranging from −400 to +50 bp relative to the putative target gene start site. For the prediction of operons and TFBSs, the TRACTOR_DB (53) database and PATSER (54) were used.

In a feasibility study for taxonomically closely related species, four corynebacteria, the attempt to transfer data from *C. glutamicum* to *C. efficiens*, *C. jeikeium* and *C. diphtheriae* yielded 530 new gene regulations (55). The database content of the underlying CoryneRegNet database was increased by factor 4.2 for the three target organisms. Reliable knowledge for ~40% of the common transcription factors was made available, compared with ~5% for which knowledge was available before. Here, a promoter region was defined −560 to +20 bp relative to the putative target gene start site. The software packages PoSSuMsearch (56) and Transitivity Clustering (57,58) were used for TFBS predictions and orthology detections, respectively. A disadvantage is the usage of the operon database VIMSS (59). Since the update frequency is limited by technical restrictions, there is a delay for operon annotations of newly sequenced species. Table 1 summarizes the results of the transfer exemplarily for the transcriptional regulators GlxR (60), LexA (44), RamB (61), McbR (62) and DtxR (30). For the latter case, we know the regulons of all four organisms, the source as well as the three targets. Here, the transfer pipeline reconstructed almost half of the DtxR regulons with no false positives.

Table 1. Examples for regulons transferred between corynebacteria[a]

	GlxR	LexA	RamB	McbR	DtxR	
CG	99	20	47	46	64	
					TP	FP
CD	35	9	27	11	25 of 63 (40%)	0
CE	104	14	22	26	18 of 27 (67%)	0
CJ	33	4	13	12	21 of 51 (41%)	0

[a]The table shows the number of known and predicted target genes for five transcription factors that are conserved among the species *C. glutamicum* (CG), *C. diphtheriae* (CD), *C. efficiens* (CE), and *C. jeikeium* (CJ). CG served as source organism, while CD, CE, and CJ are the target organisms. A combination of orthology detection, binding-site conservation and operon extension has been used for the inter-species transfer procedure(55). The DtxR regulons of CD, CE and CJ have been known in advance allowing us to judge the prediction performance, i.e. we may give numbers for true positives (TP) and false positives (FP).

Note that the presented list of case studies and examples is explicitly not claimed to be exhaustive. We concentrated on specific examples that highlight genome-wide approaches and provide clear and easily interpretable results allowing us to receive an impression of the state of the art. For instance the usage of corynebacterial data throughout this work is not biased but based on practical considerations: We have some proven knowledge about four taxonomically closely related organisms (32), which allowed us to judge and investigate typical problems.

In summary, we conclude that solely using orthologous TFs and TGs is too unreliable. It overestimates the inter-species conservation of TF-TG interactions and underestimates the amount of new regulations that have not been observed in the reference organism before. The TFBS-based approach is capable of identifying new regulations in the target organisms but suffers from high false positive rates if used in isolation. The combination of both is reliable. Essentially, we filter the comparably unspecific TFBS-based results by adding further evidence predicated on conserved TGs. Still, we neglect the underestimation of new regulations of a specific TF in the target organism, i.e. conserved TFBSs but no conserved TGs; something to be discussed in the next section.

OPEN CHALLENGES AND FUTURE DIRECTIONS

The major problem with all approaches is the dependency on a successful discovery of evolutionarily conserved sequences. In contrast to the promoter sequences, TFBSs are comparably short. Furthermore, their variation is comparably high, even within the same organism. This may result in low information content, i.e. an unfavorable signal-to-noise ratio, and is one of the main reasons for the high false positive rates of computational TFBS-identification methods. In addition, we observe a reduced sensitivity when moving from one organism to another one, even if closely related (42). Orthology detection methods are integrated to reduce these error rates, which generally hinders unraveling regulatory interactions in target organisms with many inorthologous TGs. New gene regulations for unconserved TGs may not be identified anymore. We propose an additional step, depicted at the bottom right of Supplementary Figure S2, to counter this problem. After the identification of conserved interactions between source and target, we should not use the TFBSs of the source organism but use the conserved TFBSs in the target organism. These are expected to be more precise since they are putative true binding sites from the target organism itself. Revised computational profiles, constructed from these TFBSs, could subsequently be utilized to scan for further TFBSs in the target organism. Note that we still risk a number of false positive predictions. This can be reduced by applying restrictive significance thresholds and by re-adjusting (fine-tuning) the TFBSs by using motif discovery tools (see below). One possible tool to integrate with a network transfer pipeline may be PhyloGibbs. It identifies conserved sequences motifs and additionally accounts for phylogenetic distances (63,64). We may further decrease the number of false positives by not scanning upstream sequences with fixed positions relative to a TGs start sites. Instead, we might want to use more reasonable promotor sequences by integrating software dedicated to the discovery of transcription start-sites (65).

Another problem with TFBSs is the annotation procedure, where data is transferred manually from the literature to the reference databases. In a recent study about the TFBSs of seven TFs from *E. coli*, Keilwagen *et al.* found that 34.5% of the 536 TFBS annotations are questionable; 51 are suggested to be removed, 134 to be shifted by some base pairs (66). The incorporation of so-called sequence motif discovery tools helps with identifying such annotation problems, subsequent TFBS readjusting and finally with the fine-tuning and discovery of new binding motifs in the target organism. A summary and review of corresponding tools is available in a paper from Tompa *et al.* (67), newer tools may be found e.g. in (66,68,69).

Although the identification of orthologous genes and proteins is a long-standing challenge in computational biology, classical sequence-based approaches neglect to incorporate methods to distinguish between groups of sequences that share common ancestry from groups that share inserted domains but are otherwise unrelated. This protein domain shuffling problem was recently introduced and attacked with a method called Neighborhood Correlation (70,71). However, we suggest performing more research about the discovery of protein domain architecture and its impact on TF-DNA binding and TG conservation; after all, we are still interested in predicting reliable gene regulatory networks here, but not necessarily in unraveling the path of evolution itself.

Besides technical difficulties we also face organizational problems. While nowadays sequenced genomes and their annotations are stored in a well-structured manner, e.g. with the NCBI repositories, gene regulatory interactions, binding sites, operon annotations, homology information, etc., are not. Instead, this data is scattered over numerous publications, not utilizing standardized vocabularies to describe the content for subsequent processing. Text Mining tools are necessary to retrieve relevant literature suggestions (72,73). In addition, even if the data are stored in public databases, it is often not available through standard interfaces. Furthermore, software packages usually need to be compiled, installed and configured locally, often a difficult and time-consuming task. We recommend that the community should follow the advices of Philippi and Köhler (22); primarily, we propose enforcing standardization by making it a requirement for publication in scientific journals.

CONCLUSIONS

Despite all the technical and organizational problems, we conclude that the inter-species transfer of knowledge about gene regulatory networks from model organisms to reference organisms is generally feasible. Reference networks for some prokaryotes are publicly available and can be used for automatic annotations, at least for somewhat related species. In principle, we have all the necessary computational tools available but we are not using them as integral part of standard data-processing pipelines. The performance is limited in terms of sensitivity, which can be improved, for instance, by incorporating phylogenetic sequence motif discovery tools. However, predicted regulations are reliable if the integrated tools are combined appropriately. Hence, we suggest to define standard pipelines similar to the one depicted in Supplementary Figure S2. Furthermore, we motivate their compulsory application to any new genome sequence. Database providers for the reference organism networks could (i) allow uploading whole-genome sequence annotations or (ii) automatically integrate all new genomes from NCBI. After inter-species transferring, potential gene regulations for the target organism may be downloaded, visualized or post-processed. Researchers are automatically provided with new promising wet-lab targets for further studies. This would significantly reduce the gap between existing bacterial genome sequences and the knowledge about gene regulatory networks, a big step in systems biology.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for this work was provided by the German Academic Exchange Service (DAAD), the Cluster of Excellence for Multimodel Computing and Interaction (MMCI) of the German Research Foundation (DFG) and the Boehringer Ingelheim Fonds (BIF). Funding for open acces charge: MMCI.

Conflict of interest statement. None declared.

REFERENCES

1. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **38**, D5–16.
2. Metzker,M.L. (2010) Sequencing technologies – the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
3. Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.
4. Venancio,T.M. and Aravind,L. (2009) Reconstructing prokaryotic transcriptional regulatory networks: lessons from actinobacteria. *J. Biol.*, **8**, 29.
5. Pabo,C.O. and Sauer,R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.*, **61**, 1053–1095.
6. Baumbach,J., Tauch,A. and Rahmann,S. (2009) Towards the integrated analysis, visualization and reconstruction of microbial gene regulatory networks. *Brief Bioinform.*, **10**, 75–83.
7. van Vliet,A.H. (2009) Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiol. Lett.*, **302**, 1–7.

8. Hellman,L.M. and Fried,M.G. (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat. Protoc.*, **2**, 1849–1861.
9. Galas,D.J. and Schmitz,A. (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.*, **5**, 3157–3170.
10. Sun,L.V., Chen,L., Greil,F., Negre,N., Li,T.R., Cavalli,G., Zhao,H., Van Steensel,B. and White,K.P. (2003) Protein-DNA interaction mapping using genomic tiling path microarrays in Drosophila. *Proc. Natl. Acad. Sci. USA*, **100**, 9428–9433.
11. Jothi,R., Cuddapah,S., Barski,A., Cui,K. and Zhao,K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.
12. Bonneau,R. (2008) Learning biological networks: from modules to dynamics. *Nat. Chem. Biol.*, **4**, 658–664.
13. Herrgard,M.J., Covert,M.W. and Palsson,B.O. (2004) Reconstruction of microbial transcriptional regulatory networks. *Curr. Opin. Biotechnol.*, **15**, 70–77.
14. Kazakov,A.E., Cipriano,M.J., Novichkov,P.S., Minovitsky,S., Vinogradov,D.V., Arkin,A., Mironov,A.A., Gelfand,M.S. and Dubchak,I. (2007) RegTransBase–a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res.*, **35**, D407–D412.
15. Gama-Castro,S., Jimenez-Jacinto,V., Peralta-Gil,M., Santos-Zavaleta,A., Penaloza-Spinola,M.I., Contreras-Moreira,B., Segura-Salazar,J., Muniz-Rascado,L., Martinez-Flores,I., Salgado,H. *et al.* (2008) 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–D124.
16. Keseler,I.M., Bonavides-Martinez,C., Collado-Vides,J., Gama-Castro,S., Gunsalus,R.P., Johnson,D.A., Krummenacker,M., Nolan,L.M., Paley,S., Paulsen,I.T. *et al.* (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res.*, **37**, D464–D470.
17. Sierro,N., Makita,Y., de Hoon,M. and Nakai,K. (2008) DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res.*, **36**, D93–D96.
18. Jacques,P.E., Gervais,A.L., Cantin,M., Lucier,J.F., Dallaire,G., Drouin,G., Gaudreau,L., Goulet,J. and Brzezinski,R. (2005) MtbRegList, a database dedicated to the analysis of transcriptional regulation in *Mycobacterium tuberculosis*. *Bioinformatics*, **21**, 2563–2565.
19. Grote,A., Klein,J., Retter,I., Haddad,I., Behling,S., Bunk,B., Biegler,I., Yarmolinetz,S., Jahn,D. and Munch,R. (2009) PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. *Nucleic Acids Res.*, **37**, D61–D65.
20. Baumbach,J., Wittkop,T., Kleindt,C.K. and Tauch,A. (2009) Integrated analysis and reconstruction of microbial transcriptional gene regulatory networks using CoryneRegNet. *Nat. Protoc.*, **4**, 992–1005.
21. Salgado,H., Gama-Castro,S., Peralta-Gil,M., Diaz-Peredo,E., Sanchez-Solano,F., Santos-Zavaleta,A., Martinez-Flores,I., Jimenez-Jacinto,V., Bonavides-Martinez,C., Segura-Salazar,J. *et al.* (2006) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.*, **34**, D394–D397.
22. Philippi,S. and Kohler,J. (2006) Addressing the problems with life-science databases for traditional uses and systems biology. *Nat. Rev. Genet.*, **7**, 482–488.
23. Babu,M.M., Lang,B. and Aravind,L. (2009) Methods to reconstruct and compare transcriptional regulatory networks. *Methods Mol. Biol.*, **541**, 163–180.
24. Babu,M.M., Luscombe,N.M., Aravind,L., Gerstein,M. and Teichmann,S.A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, **14**, 283–291.
25. Madan Babu,M., Teichmann,S.A. and Aravind,L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J. Mol. Biol.*, **358**, 614–633.
26. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
27. Huynen,M.A. and Bork,P. (1998) Measuring genome evolution. *Proc. Natl. Acad. Sci. USA*, **95**, 5849–5856.
28. Yu,H., Luscombe,N.M., Lu,H.X., Zhu,X., Xia,Y., Han,J.D., Bertin,N., Chung,S., Vidal,M. and Gerstein,M. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.*, **14**, 1107–1118.
29. Price,M.N., Dehal,P.S. and Arkin,A.P. (2007) Orthologous transcription factors in bacteria have different functions and regulate different genes. *PLoS Comput. Biol.*, **3**, 1739–1750.
30. Brune,I., Werner,H., Huser,A.T., Kalinowski,J., Puhler,A. and Tauch,A. (2006) The DtxR protein acting as dual transcriptional regulator directs a global regulatory network involved in iron metabolism of *Corynebacterium glutamicum*. *BMC Genomics*, **7**, 21.
31. Brinkrolf,K., Brune,I. and Tauch,A. (2006) Transcriptional regulation of catabolic pathways for aromatic compounds in *Corynebacterium glutamicum*. *Genet. Mol. Res.*, **5**, 773–789.
32. Baumbach,J. (2007) CoryneRegNet 4.0 - A reference database for corynebacterial gene regulatory networks. *BMC Bioinformatics*, **8**, 429.
33. Baumbach,J. and Apeltsin,L. (2008) Linking Cytoscape and the corynebacterial reference database CoryneRegNet. *BMC Genomics*, **9**, 184.
34. Baumbach,J., Wittkop,T., Rademacher,K., Rahmann,S., Brinkrolf,K. and Tauch,A. (2007) CoryneRegNet 3.0–an interactive systems biology platform for the analysis of gene regulatory networks in corynebacteria and *Escherichia coli*. *J. Biotechnol.*, **129**, 279–289.
35. Balazsi,G., Heath,A.P., Shi,L. and Gennaro,M.L. (2008) The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest. *Mol. Syst. Biol.*, **4**, 225.
36. Guo,M., Feng,H., Zhang,J., Wang,W., Wang,Y., Li,Y., Gao,C., Chen,H., Feng,Y. and He,Z.G. (2009) Dissecting transcription regulatory pathways through a new bacterial one-hybrid reporter system. *Genome Res.*, **19**, 1301–1308.
37. Wilson,D., Charoensawan,V., Kummerfeld,S.K. and Teichmann,S.A. (2008) DBD–taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.*, **36**, D88–D92.
38. Bakke,P., Carney,N., Deloache,W., Gearing,M., Ingvorsen,K., Lotz,M., McNair,J., Penumetcha,P., Simpson,S., Voss,L. *et al.* (2009) Evaluation of three automated genome annotations for Halorhabdus utahensis. *PLoS ONE*, **4**, e6291.
39. Bocker,S., Jahn,K., Mixtacki,J. and Stoye,J. (2009) Computation of median gene clusters. *J. Comput. Biol.*, **16**, 1085–1099.
40. Raghupathy,N. and Durand,D. (2009) Gene cluster statistics with gene families. *Mol. Biol. Evol.*, **26**, 957–968.
41. Rahmann,S., Müller,T. and Vingron,M. (2003) On the power of profiles for transcription factor binding site detection. *Statistical Applications in Genetics and Molecular Biology*, **2**, Article 7.
42. Baumbach,J., Brinkrolf,K., Wittkop,T., Tauch,A. and Rahmann,S. (2006) CoryneRegNet 2: An Integrative Bioinformatics Approach for Reconstruction and Comparison of Transcriptional Regulatory Networks in Prokaryotes. *Journal of Integrative Bioinformatics*, **3**, 24.
43. Mazon,G., Erill,I., Campoy,S., Cortes,P., Forano,E. and Barbe,J. (2004) Reconstruction of the evolutionary history of the LexA-binding sequence. *Microbiology*, **150**, 3783–3795.
44. Jochmann,N., Kurze,A.K., Czaja,L.F., Brinkrolf,K., Brune,I., Huser,A.T., Hansmeier,N., Puhler,A., Borovok,I. and Tauch,A. (2009) Genetic makeup of the Corynebacterium glutamicum LexA regulon deduced from comparative transcriptomics and in vitro DNA band shift assays. *Microbiology*, **155**, 1459–1477.
45. Brouwer,R.W., Kuipers,O.P. and van Hijum,S.A. (2008) The relative value of operon predictions. *Brief Bioinform.*, **9**, 367–375.
46. Alkema,W.B., Lenhard,B. and Wasserman,W.W. (2004) Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus. *Genome Res.*, **14**, 1362–1373.
47. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG

database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
48. Thompson,W., Rouchka,E.C. and Lawrence,C.E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, **31**, 3580–3585.
49. Lenhard,B. and Wasserman,W.W. (2002) TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.
50. Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18(Suppl. 1)**, S329–336.
51. Espinosa,V., Gonzalez,A.D., Vasconcelos,A.T., Huerta,A.M. and Collado-Vides,J. (2005) Comparative studies of transcriptional regulation mechanisms in a group of eight gamma-proteobacterial genomes. *J. Mol. Biol.*, **354**, 184–199.
52. Gonzalez Perez,A.D., Gonzalez Gonzalez,E., Espinosa Angarica,V., Vasconcelos,A.T. and Collado-Vides,J. (2008) Impact of Transcription Units rearrangement on the evolution of the regulatory network of gamma-proteobacteria. *BMC Genomics*, **9**, 128.
53. Perez,A.G., Angarica,V.E., Vasconcelos,A.T. and Collado-Vides,J. (2007) Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes. *Nucleic Acids Res.*, **35**, D132–136.
54. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
55. Baumbach,J., Rahmann,S. and Tauch,A. (2009) Reliable transfer of transcriptional gene regulatory networks between taxonomically related organisms. *BMC Syst. Biol.*, **3**, 8.
56. Beckstette,M., Homann,R., Giegerich,R. and Kurtz,S. (2006) Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics*, **7**, 389.
57. Wittkop,T., Baumbach,J., Lobo,F.P. and Rahmann,S. (2007) Large scale clustering of protein sequences with FORCE – A layout based heuristic for weighted cluster editing. *BMC Bioinformatics*, **8**, 396.
58. Wittkop,T., Emig,D., Lange,S., Rahmann,S., Albrecht,M., Morris,J.H., Bocker,S., Stoye,J. and Baumbach,J. (2010) Partitioning biological data with transitivity clustering. *Nat. Methods*, **7**, 419–420.
59. Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, **33**, 880–892.
60. Kohl,T.A., Baumbach,J., Jungwirth,B., Puhler,A. and Tauch,A. (2008) 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.*, **135**, 340–350.
61. Gerstmeir,R., Cramer,A., Dangel,P., Schaffer,S. and Eikmanns,B.J. (2004) RamB, a novel transcriptional regulator of genes involved in acetate metabolism of Corynebacterium glutamicum. *J. Bacteriol.*, **186**, 2798–2809.
62. Rey,D.A., Nentwich,S.S., Koch,D.J., Ruckert,C., Puhler,A., Tauch,A. and Kalinowski,J. (2005) The McbR repressor modulated by the effector substance S-adenosylhomocysteine controls directly the transcription of a regulon involved in sulphur metabolism of Corynebacterium glutamicum ATCC 13032. *Mol. Microbiol.*, **56**, 871–887.
63. van Nimwegen,E. (2007) Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, **8(Suppl. 6)**, S4.
64. Siddharthan,R., Siggia,E.D. and van Nimwegen,E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67.
65. Mendoza-Vargas,A., Olvera,L., Olvera,M., Grande,R., Vega-Alvarado,L., Taboada,B., Jimenez-Jacinto,V., Salgado,H., Juarez,K., Contreras-Moreira,B. *et al.* (2009) Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. *PLoS ONE*, **4**, e7526.
66. Keilwagen,J., Baumbach,J., Kohl,T.A. and Grosse,I. (2009) MotifAdjuster: a tool for computational reassessment of transcription factor binding site annotations. *Genome Biol.*, **10**, R46.
67. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
68. Baumbach,J., Wittkop,T., Weile,J., Kohl,T. and Rahmann,S. (2008) MoRAine – A web server for fast computational transcription factor binding motif re-annotation. *Journal of Integrative Bioinformatics*, **5**.
69. Linhart,C., Halperin,Y. and Shamir,R. (2008) Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res.*, **18**, 1180–1189.
70. Joseph,J.M. and Durand,D. (2009) Family classification without domain chaining. *Bioinformatics*, **25**, i45–53.
71. Song,N., Joseph,J.M., Davis,G.B. and Durand,D. (2008) Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput. Biol.*, **4**, e1000063.
72. Jensen,L.J., Saric,J. and Bork,P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 119–129.
73. Winnenburg,R., Wachter,T., Plake,C., Doms,A. and Schroeder,M. (2008) Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief Bioinform.*, **9**, 466–478.
74. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.