RegCyanoDB: a database of regulatory interactions in cyanobacteria

Ajay Nair[123*] , Madhu Chetty[4], and Nguyen Xuan Vinh[5]

[1]IITB-Monash Research Academy, Indian Institute of Technology Bombay, Mumbai, 400 076, India
[2]Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia,
[3]Chemical Engineering Department, Indian Institute of Technology Bombay, Mumbai, 400 076, India,
[4]Faculty of Science and Technology, Federation University, VIC 3841, Australia,
[5]School of Computing and Information Systems, The University of Melbourne, VIC 3010, Australia.

Email: *ajaynair@iitb.ac.in;

*Corresponding author

Abstract

Background: Cyanobacteria are photoautotrophic organisms with environmental, evolutionary, and industrial importance. Knowledge of its regulatory interactions are important to predict, optimise, and engineer their characteristics. However, at present, very few of their regulatory interactions are known. The regulatory interactions are known only for a few model organisms such as *Escherichia coli* due to technical and economical constraints, which are unlikely to change soon. Thus, mapping of regulatory interactions from well-studied organisms to less-studied organisms by using computational techniques is widely used. Reverse Best Hit (RBH), with appropriate algorithm parameters, is a simple and efficient method for detecting functional homologs.

Description: We predict the regulatory interactions in 30 strains of cyanobacteria using the known regulatory interactions from the best-studied organism, *E. coli*. RBH method with appropriate parameters is used to identify the functional homologs. An interaction is mapped to a cyanobacterial strain if functional homologs exist for a known transcription factor and its target gene. The confidence of the detected homologs and interactions are also provided. Since RBH is a conservative method, homolog-grouping is performed to recover lost putative interactions. A database of the predicted interactions from all the 30 strains of cyanobacteria is constructed.

Conclusion: RegcyanoDB contains 20,280 interactions with confidence levels for 30 cyanobacterial strains. The predicted regulatory interactions exhibit a scale free network topology as observed in model organisms. The interacting genes in *E. coli* and cyanobacteria are mostly found to have the same gene annotation. This database can be used for posing novel hypotheses and validation studies in wet-lab and computational domains.

The database is available at http://www.che.iitb.ac.in/grn/RegCyanoDB/

Keywords

Cyanobacteria, *Escherichia coli*, gene regulatory network, transcriptional regulation, transcription factors, regulatory interactions, regulatory interaction mapping, Bioinformatics.

1 Introduction

Cyanobacteria, or the blue-green bacteria, are photoautotrophic organisms credited with changing the Earth's atmosphere to oxygen rich condition. It is believed that the photosynthetic machinery in plants and algae, the chloroplast, has evolved from cyanobacteria by endosymbiosis. Cyanobacteria are the model organism for studying photosynthesis, as well as nitrogen and carbon assimilation. They are believed to play an important role in marine nitrogen-fixing cycle [1, 2]. Presently, there is enormous interest in using cyanobacteria for biofuel and hydrogen production [3–5]. The algal species and cyanobacteria have the highest biofuel production per unit area [3] and have higher growth and photosynthetic rates [5]. N_2-fixing cyanobacterial strains have simple growth conditions and have the simplicity of prokaryotic genome compared to eukaryotic algae [5]. Thus, they are a promising sustainable alternative to our energy requirements.

For biofuel applications, understanding the metabolic and genetic factors involved in maximising the productivity in an organism are of great interest [3, 6]. Organisms respond to varying environmental conditions by regulating the cellular protein production. In prokaryotes, the regulation of the proteins is largely carried out by the transcriptional networks. Thus, understanding the transcriptional regulatory network is crucial in optimising the culture conditions and for metabolic or genetic manipulation.

Although around 2000 completely sequenced genomes are reported in GOLD database [7], their functional characterization and regulatory interactions are lagging behind. Even for the model organism *Escherichia coli* K-12, whose regulatory network is best understood of all the living organisms [8], only about one-third of the genes have experimentally validated interactions (RegulonDB database, June 2012). The main reason for this limited information is that there are many technical and organizational issues associated with finding regulatory interactions experimentally [9]. Further, compared to metabolic networks, using comparative genomics in regulatory networks is more challenging as they are less conserved, very plastic, and the transcription factors evolve fast [10–13]. As a result, our current knowledge of the regulatory networks in prokaryotes is limited to only a few model organisms. These are available in public databases such as RegulonDB [8] and EcoCyc [14] for *E. coli*; DBTBS [15] for *Bacillus subtilis*; MtbRegList [16] and MycoRegNet [17] for *Mycobacterium tuberculosis*; and CoryneRegNet [18] for corynebacteria. RegTransBase [19] contains manually-curated, experimentally-verified interactions for 128 microbes, while PRODORIC [20] contains the regulatory information of many prokaryotes but mainly *E. coli*, *B. subtilis*, and *Pseudomonas aeruginosa*. Due to difficulties involved in obtaining regulatory interactions experimentally, mapping regulatory interactions from model organisms to others using computational techniques is widely used [9, 10, 16–18, 21–23]. While generic databases like RegTransBase [19] and PRODORIC [20] contain interactions for cyanobacteria, these interactions are very few. Hence, it has become imperative to develop a database which contains computationally predicted regulatory interactions for this organism.

RegCyanoDB is thus, the first database of regulatory interactions in cyanobacteria. The regulatory interactions are mapped using *E. coli* as the reference organism. This database also provides the confidence level of the predicted interactions based on the quality of the sequence alignment.

2 Computational methods for regulatory interaction mapping

Several studies have characterized the two main assumptions in computational transfer of regulatory interactions: (i) the function of a new protein can be predicted using its sequence similarity to a known protein; and (ii) for a known transcription factor(TF) and its target gene(TG) in the 'source' organism, the interaction is conserved in the 'target' organism if there exist functional homologs for both the TF and its corresponding TG. The concept of "interologs", the orthologous pair of interacting proteins, was reported [24] for transferring protein-protein interactions between organisms. This concept was extended to a large scale study [23] that introduced the concept of "regulog", the conserved protein-DNA interactions in different organisms. It was shown that if a TF has a homolog in another organism with 30-60% or better sequence identity, the binding site of the homolog is conserved and for identities above 80%, all the protein-protein interactions are conserved. The sequence identity values reported in [23] also matched observations in [25], which noted that the pairwise alignment of two sequences correlated the structural alignment when the sequence identity

is above 25-30%.

Benchmark studies [26–28] and reviews [29, 30] have analysed the various ortholog detection methods. Recent benchmark studies have shown that reverse best hit (RBH) is as good or even superior to other methods [26, 27]. Note that RBH is also referred to as reciprocal best hit, or bidirectional best hit (BBH), or symmetrical best hit (SymBeT). RBH is a pairwise sequence alignment method that uses the concept of orthologous genes which is operationally defined as the gene pair having the best sequence similarity between all the genes in two genomes [13,31]. Thus, if a protein $P1_x$ in the first organism picks protein $P2_y$ as its best hit in a sequence similarity search against all the proteins in the second organism, and if $P2_y$ picks $P1_x$ as its best hit among all the proteins in the first organism, then $P1_x$ and $P2_y$ are called the RBH of each other. It is important to note that the sequence similarity between the two proteins should have sufficient statistical significance [31]. Among the different methods for ortholog detection, sequence similarity based methods like RBH are strong predictors of functional relatedness [26] and appropriate choice of algorithm parameters yield good results [29, 32, 33]. Orthologs are generally accepted to be functionally equivalent [13, 30, 31].

Difficulties for functional prediction using RBH arise when there is domain shuffling, presence or absence of domains, gene duplication, gene loss, and horizontal gene transfer [25, 30]. The possible error due to changes in protein domains are addressed by considering the coverage of the pairwise alignment. Different implementations have used different coverage criteria, such as, 80% coverage [23, 34], or 70% coverage along with protein domain information [12], or 60% coverage with 60% identity in the alignment region [10], or 50% coverage with relevant e-value cut-off [33].

Many-to-one or one-to-many relations caused by gene deletions and duplications cannot be considered in RBH method and low sequence similarity of alignment will require additional methods like conserved gene neighbourhood analysis [29, 30, 35]. Since RBH considers only the best hit in both the directions, it is considered as a conservative method. Therefore, clustering or grouping of homologous proteins is used to recover the false negatives [10, 35]. Additional constraints, like minimum sequence identity and coverage are considered to minimize false positives [10, 12, 23, 25, 33].

BLAST [36, 37] is by far the most popular method for sequence similarity searches. Ortholog detection using BLAST with different parameter cut-offs such as e-value, raw-score, bit-score, and identity give different but essentially overlapping results [29]. The cut-off for sequence similarity should be statistically significant but not too stringent [31]. Smith-Waterman implementation with e-value cut-off was found to be a good measure of structural similarity between the proteins [32] and Smith-Waterman alignment with soft-filtering in BLAST was reported to give best results for RBH [33].

These results form the basis for parameter selection for the algorithms and for predicting the confidence of regulatory interactions in this work.

3 Construction and content

The regulatory interactions in 30 strains of cyanobacteria were mapped from the known regulatory network in *E. coli* K-12. The database construction procedure for a single cyanobacterial strain is shown in figure 1. The experimentally characterised regulatory interactions in *E. coli* were obtained from RegulonDB [8] which had 3920 interactions, 183 TFs, and 1563 unique mRNA genes (as on June 2012). Interactions with other gene products like tRNA, rRNA, and ncRNA were ignored. The genes reported in the interactions were identified mainly using RSAT [38, 39] and the rest from NCBI RefSeq [40] and UniProtKB [41]. Protein sequences of *E. coli* were obtained mainly from NCBI RefSeq and the rest from UniProtKB. Complete protein sequences of the 30 cyanobacterial strains were obtained from NCBI RefSeq. The list of strains in Table 1, include all the major orders such as Chroococcales (21 strains), Nostocales (4 strains), Prochlorales (2 strains), Gloeobacteria (1 strain), and Oscillatoriales (1 strain).

RBH was implemented in standalone BLAST (version 2.2.26+) from NCBI with the following parameters. The e-value cut-off 1e-04 was decided by the database size [32] of ∼6000 protein sequences for a single BLAST run. Smith-Waterman alignment and soft-filtering were enabled for optimal alignment results [32, 33].

After obtaining the protein sequences, BLAST of *E. coli* proteins against proteins of a selected strain

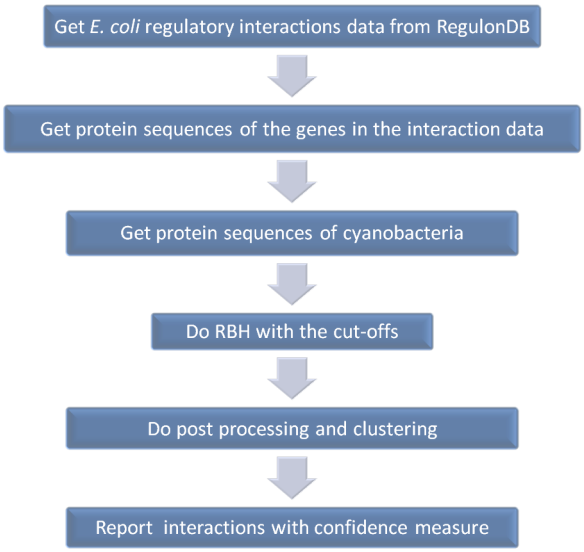

Figure 1: Procedure for obtaining regulatory interactions for a cyanobacterial strain

of cyanobacteria and the reverse BLAST of the cyanobacterial proteins against the *E. coli* proteins were performed. The RBH proteins were selected from the BLAST results and graded for functional homolog confidence. The classification of confidence level is shown in Table 2. For a pairwise sequence alignment, if the identity was at least 80% and alignment length covered at least 80% of the two proteins, the homolog confidence was reported 'high'. As discussed earlier, at this sequence identity level the protein-protein interactions and protein-DNA interactions are expected to be conserved [23]. If the identity is between $60-80\%$ and coverage is at least 80%, the homolog confidence was assigned 'good', as protein-DNA interactions are expected to be conserved and many protein-protein interactions are also conserved [23]. Similar cut-off had been reported to give optimal results for regulatory interaction mapping [10]. Homolog confidence was reported as 'moderate' for identity between $25-60\%$ and coverage greater than 60%. At this identity levels, protein-DNA interactions could be conserved [23] and sequence alignment correlates structural alignment [25]. All other RBH cases were reported as 'low' confidence homologs as the confidence in their function cannot be decided by quality of sequence alignment alone.

For the proteins in *E. coli* which did not have a RBH, 'homolog-groups' were created. In this procedure, homolog confidence was computed for all the hits of an *E. coli* protein PE_x against the selected cyanobacteria. All the homolog proteins in the hit having confidence of 'moderate' or better were listed together in '$Cluster_x$'as the putative functional homologs of PE_x. The confidence level of the homolog-group was assigned as the highest homolog confidence shown by any member in the group. This procedure is similar in concept to the reported clustering methods [10, 35] and is described as follows. RBH cannot extract the functional homologs in the presence of gene deletions and duplications. However, proteins with sufficient identities and coverage in sequence alignment can be functionally equivalent. So, when RBH is not present, all the homolog proteins that show sufficient identity and coverage in pairwise sequence alignment are reported as putative functional homologs. Thus, false negatives from the conservative RBH method can be recovered. However, since there are chances of false positives also being present, the results with homolog-group procedure are reported separately.

After identifying the functional homologs using RBH and homolog-grouping, the TF-TG interactions were obtained using the interactions reported in RegulonDB database. For a TF-TG interaction reported in RegulonDB, corresponding interaction is considered as conserved in the selected cyanobacteria, if a homolog or a homolog-group can be identified for both the TF and TG in the cyanobacteria. The confidence of the interaction is assigned as the lowest of the homolog confidence level between the homolog TF and

Table 1: The 30 strains of cyanobacteria used in the study

Order	Strain
Chroococcales	
	Cyanothece sp. ATCC 51142
	Cyanothece sp. PCC 7424
	Cyanothece sp. PCC 7425
	Cyanothece sp. PCC 7822
	Cyanothece sp. PCC 8801
	Cyanothece sp. PCC 8802
	Microcystis aeruginosa NIES-843
	Synechococcus elongatus PCC 6301
	Synechococcus sp. CC9311
	Synechococcus sp. CC9605
	Synechococcus sp. CC9902
	Synechococcus sp. JA-3-3Ab
	Synechococcus sp. PCC 7002
	Synechococcus sp. RCC307
	Synechococcus sp. WH 7803
	Synechococcus sp. WH8102
	Synechocystis sp. PCC 6803 substr. PCC-N
	Synechocystis sp. PCC 6803 substr. PCC-P
	Synechocystis sp. PCC 6803 substr. GT-I
	Thermosynechococcus elongatus BP-1
	Cyanobacterium UCYN-A
Gloeobacteria	
	Gloeobacter violaceus PCC 7421
Nostocales	
	Anabaena sp. PCC 7120/Nostoc sp. PCC7120
	Anabaena variabilis ATCC 29413
	Nostoc azollae 0708
	Nostoc punctiforme ATCC 29133
Oscillatoriales	
	Trichodesmium erythraeum IMS101
Prochlorales	
	Prochlorococcus marinus str. MIT 9215
	Prochlorococcus marinus SS120

the TG. For example, if the TF homolog confidence was 'moderate' and the TG homolog confidence was 'high', the interaction confidence is taken as 'moderate'. Wherever possible, the homolog proteins identified in cyanobacteria were also verified by their annotations to their *E. coli* protein annotations. The results obtained from RBH and obtained by both RBH and homolog-grouping are reported separately in database. This procedure was repeated for all the other strains of cyanobacteria to create complete database.

For illustration, the predicted regulatory interactions of *Cyanothece* sp. ATCC 51142 and *Acaryochloris marina* MBIC11017, visualized using Cytoscape [42], are shown in figures 2 and 3. It can be seen that the network approximates a scale-free topology in which a small number of TFs regulate a large number of TGs, forming network hubs, and a large number of TFs control only a small number of TGs. Further, a large number of TGs are regulated by only a small number of TFs and a small number of TGs are controlled by many TFs. Other cyanobacterial strains also showed similar network topology. This is similar to the gene regulatory network topology reported for other well studied organisms [43, 44].

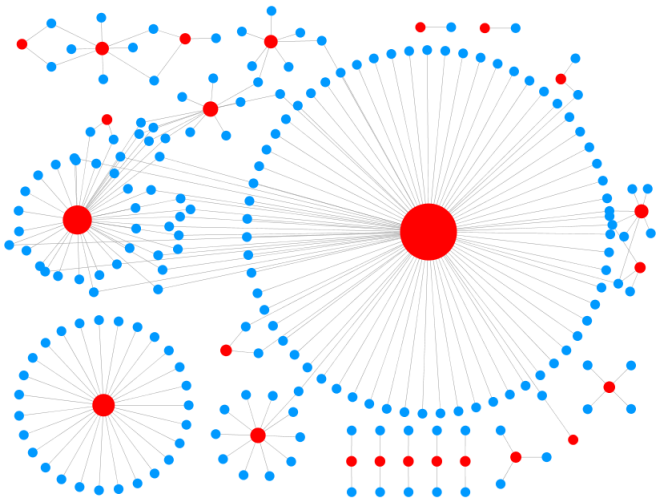

Figure 2: Regulatory network of *Cyanothece* sp. ATCC 51142. The red-nodes represent transcription factors in which the node-size is proportional to the out-degree. The blue-nodes represent target genes and edges represent the interactions. The network topology shows a scale free structure.

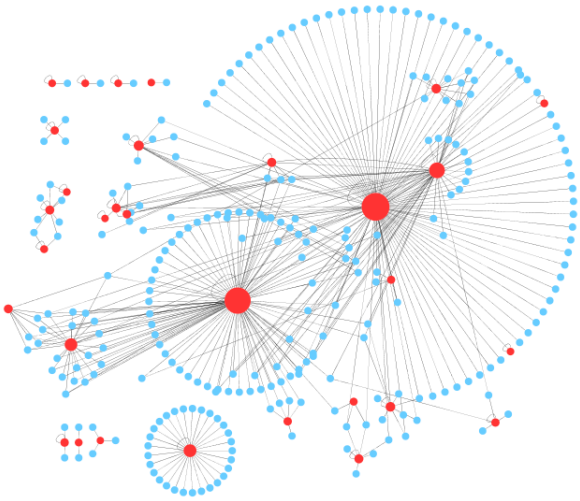

Figure 3: Regulatory network of *A. marina* MBIC11017. The red-nodes represent transcription factors whose node-size is proportional to the out-degree. The blue-nodes represent target genes and edges represent the interactions. The network topology shows a scale free structure.

Table 2: Confidence levels for a functional homolog using RBH. The intervals are based on the pairwise alignment parameters of sequence identity (SI)% and the coverage of the alignment or the alignment length (AL) in % for both proteins from an RBH.

Confidence level	Parameters
High	SI $>$ 80% AL $>$ 80%
Good	SI $>$ 60% AL $>$ 80%
Moderate	SI $>$ 25% AL $>$ 60%
Low	Other RBH

Gene regulatory interactions in different strains of cyanobacteria predicted using RBH method and a combination of RBH and homolog-grouping method are shown in figure 4. The absolute number of interactions predicted for the cyanobacterial strains are shown in figure 4(A). The relative number of interactions predicted for the different strains are shown in figure 4(B) which can be used for comparison of the conservation of interactions among the different strains. It is the ratio of the number of interactions predicted in a strain to the average number of interactions obtained for all the strains, in each method. It can be observed that the strains of Prochlorales and *Cyanobacterium* UCYN-A, which have the smallest genomes among all strains, have the lowest number of the detected interactions. *Nostoc azollae* 0708, which is a symbiont to fern *Azolla*, has the lowest number of interactions among the Nostocales strains. This probably represents the gene loss during symbiosis [45]. Interactions are relatively fewer in Synechococcus strains.

4 Utility and Discussion

4.1 RegCyanoDB user interface

The regulatory interactions in different cyanobacterial strains are available for download from the database website. Selecting 'Downloads' from the main page, and then choosing a specific strain (such as *Cyanothece* sp. ATCC 51142), display links to text files containing regulatory interactions. These files are formatted to aid manual or computational analysis. The 'Downloads' page for *Cyanothece* sp. ATCC 51142 is shown in figure 5.

The 'Transcription factor - Target gene interactions' section gives the RBH interactions alone, with no information from the homolog-grouping procedure. This is expected to contain only conservative number of interactions. 'TF-TG' section reports just the reference number of TFs and TGs and their interaction confidence. The 'Detailed information' section gives the protein name, the *E. coli* protein, and the BLAST results.

The interactions detected using the homolog-grouping procedure along with the RBH method is in second row. This section again contains two files, one concise and other detailed, as describe previously.

The details of the proteins present in the homolog-groups and their 'Cluster numbers' are given in the 'Protein homolog-groups' section.

4.2 RegCyanoDB significance

To the best of our knowledge, RegCyanoDB is the first dedicated regulatory interaction database for cyanobacteria. While there are other databases for cyanobacteria that give information about the transcription factor families [22], protein-protein interactions in specific strains [46], operons [47], and genome details [48], none of these provide regulatory interaction information. Currently, only a few hundred regulatory interactions for cyanobacteria are available in the public databases like PRODORIC and RegTransBase.

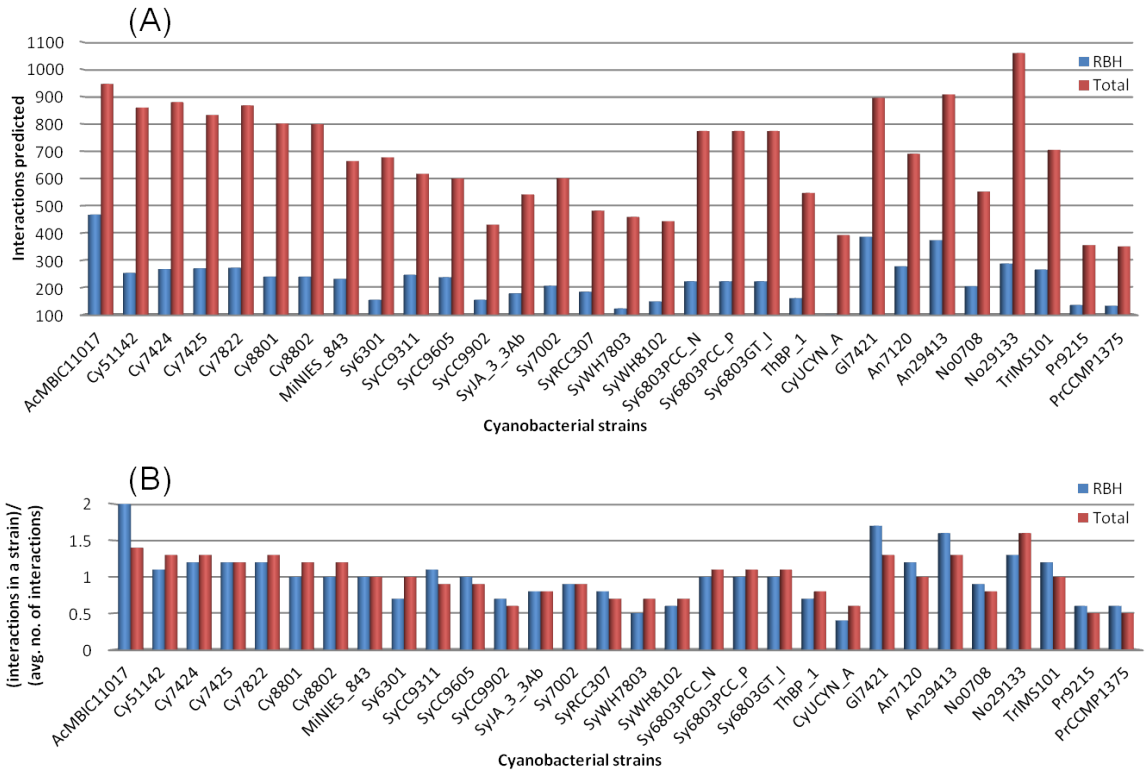

Figure 4: Regulatory interactions predicted in all cyanobacterial strains. All the 30 strains of cyanobacteria are shown in horizontal-axis. Red bars represent the interactions for RBH method and blue bars show total results for both RBH and homolog-grouping methods. (A) The absolute number of interactions predicted for all the 30 strains, using both the methods, are shown. (B) The relative number of interactions predicted in the different strains; vertical-axis is the ratio of interactions in a strain to the average number of interactions obtained for all the strains in each method.

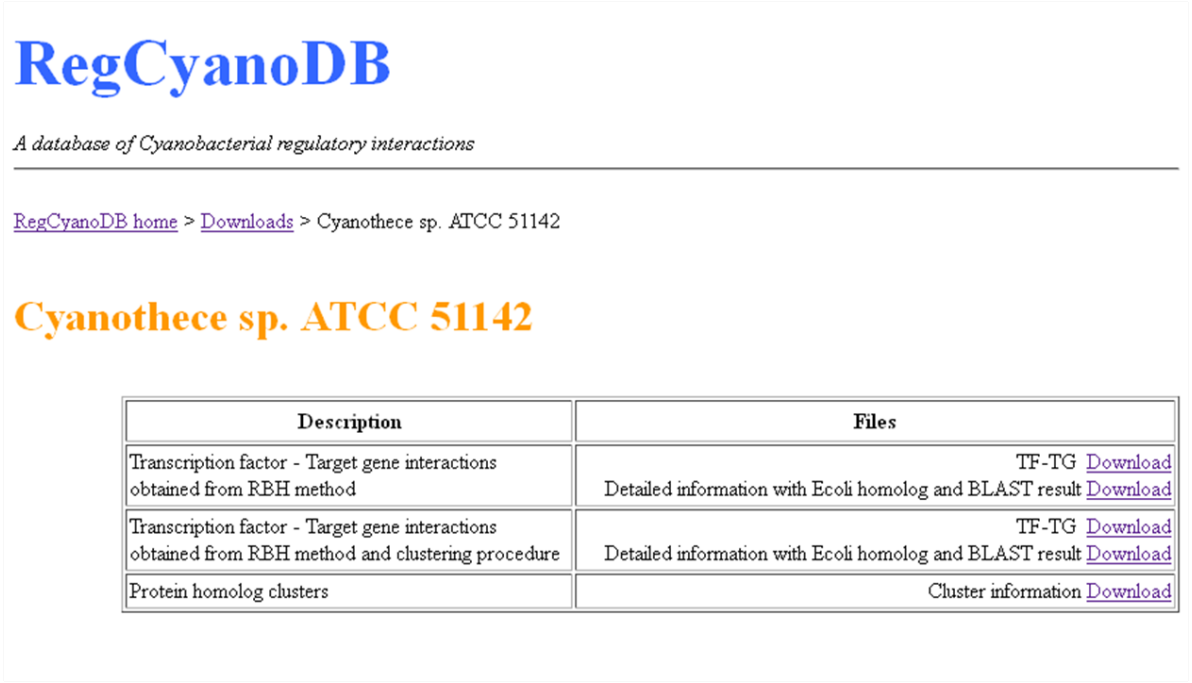

Figure 5: 'Downloads' section for *Cyanothece* sp. ATCC 51142 which shows the regulatory interaction information for cyanobacterial strain *Cyanothece* sp. ATCC 51142 available at the website.

This database predicts 20,280 interactions for the 30 strains of cyanobacteria along with confidence levels in the predicted homologs and interactions.

The regulatory interactions of *E. coli* in RegulonDB database had been used for upto 100 different applications [49], both experimental and computational. Since the cyanobacteria database reported here is computationally predicted, it can serve as the first step for targeted wet-lab experiments studying the regulatory interactions in this organism. Since the microbes predominantly use transcriptional regulation to adapt itself, the information in the website can also be used to generate new hypothesis about the characteristics and phenotype of cyanobacteria [50].

The information from the database will also be useful to understand and analyse the microarray gene expression data, predict upstream binding locations of different TFs, detect the TFBS motifs, analyse protein expression, and study regulatory interactions in different strains of cyanobacteria. Other important investigations in bioinformatic applications such as assigning protein function, uncovering novel interactions, and studying operons [47, 51, 52], will be aided by this database. The structure of the gene regulatory network in cyanobacteria at different levels, e.g. individual interactions, network motifs, and also at the global level can be studied. Current gene regulatory network modelling techniques which are limited due to 'curse of dimensionality', i.e. too many variables and too few genes, will also be benefited as these known interactions and the transcription factors can be used as a-priori knowledge in the modelling process.

5 Conclusions

RegCyanoDB provides the computationally predicted regulatory interactions in cyanobacteria, mapped from the most well-studied organism, *E. coli*. A total of 20,280 interactions with confidence levels, have been reported for the 30 strains of cyanobacteria. These confidence levels will give a better idea for using the interactions in experimental or computational applications. Further, we observe that the regulatory inter-

actions obtained in the cyanobacterial strains approximate a global scale-free network topology as reported for other model organisms.

Competing interests

The authors declare that they have no competing interests.

Author contributions

AN and MC conceptualised the work. AN collected the data, developed the algorithms, carried out the experiments and developed the database. NXV provided critical suggestions. AN, NXV and MC drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

AN performed part of this work in Pramod Wangikar's lab. He thanks Pramod Wangikar for taking part in conceptualisation and some of the discussions on this work. He thanks Dilip Durai for feedback on Perl codes and S Krishnakumar for general discussions.

References

1. Arrigo KR: **Marine microorganisms and global nutrient cycles**. *Nature* 2005, **437**(7057):349–355.
2. Zehr JP, Waterbury JB, Turner PJ, Montoya JP, Omoregie E, Steward GF, Hansen A, Karl DM: **Unicellular cyanobacteria fix N2 in the subtropical North Pacific Ocean**. *Nature* 2001, **412**(6847):635–638.
3. Georgianna DR, Mayfield SP: **Exploiting diversity and synthetic biology for the production of algal biofuels**. *Nature* 2012, **488**(7411):329–335.
4. Bandyopadhyay A, Stockel J, Min H, Sherman LA, Pakrasi HB: **High rates of photobiological H2 production by a cyanobacterium under aerobic conditions**. *Nature Communications* 2010, **1**:139.
5. Quintana N, Van der Kooy F, Van de Rhee MD, Voshol GP, Verpoorte R: **Renewable energy from Cyanobacteria: energy production optimization by metabolic pathway engineering**. *Applied Microbiology and Biotechnology* 2011, **91**(3):471–490.
6. Peralta-Yahya PP, Zhang F, Cardayre SBd, Keasling JD: **Microbial engineering for the production of advanced biofuels**. *Nature* 2012, **488**(7411):320–328.
7. Pagani I, Liolios K, Jansson J, Chen IMA, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC: **The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata**. *Nucleic Acids Research* 2012, **40**(D1):D571–D579.
8. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muiz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, Garca-Sotelo JS, Lpez-Fuentes A, Porrn-Sotelo L, Alquicira-Hernndez S, Medina-Rivera A, Martnez-Flores I, Alquicira-Hernndez K, Martnez-Adame R, Bonavides-Martnez C, Miranda-Ros J, Huerta AM, Mendoza-Vargas A, Collado-Torres L, Taboada B, Vega-Alvarado L, Olvera M, Olvera L, Grande R, Morett E, Collado-Vides J: **RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units)**. *Nucleic Acids Research* 2011, **39**(Database issue):D98–D105.
9. Baumbach J: **On the power and limits of evolutionary conservation unraveling bacterial gene regulatory networks**. *Nucleic Acids Research* 2010, **38**(22):7877 –7884.
10. Madan Babu M, Teichmann SA, Aravind L: **Evolutionary dynamics of prokaryotic transcriptional regulatory networks**. *Journal of Molecular Biology* 2006, **358**(2):614–633.
11. Herrgard MJ, Covert MW, Palsson BA: **Reconstruction of microbial transcriptional regulatory networks**. *Current Opinion in Biotechnology* 2004, **15**:70–77.

12. Lozada-Chavez I, Janga SC, Collado-Vides J: **Bacterial regulatory networks are extremely flexible in evolution**. *Nucleic Acids Research* 2006, **34**(12):3434–3445.

13. Huynen MA, Bork P: **Measuring genome evolution**. *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**(11):5849–5856.

14. Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muaiz-Rascado L, Bonavides-Martinez C, Paley S, Krummenacker M, Altman T, Kaipa P, Spaulding A, Pacheco J, Latendresse M, Fulcher C, Sarker M, Shearer AG, Mackie A, Paulsen I, Gunsalus RP, Karp PD: **EcoCyc: a comprehensive database of *Escherichia coli* biology**. *Nucleic Acids Research* 2011, **39**(Database issue):D583–D590.

15. 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 Research* 2008, **36**(Database issue):D93–D96.

16. Jacques PA, Gervais AL, Cantin M, Lucier JF, Dallaire G, Drouin G, Gaudreau L, Goulet J, Brzezinski R: **MtbRegList, a database dedicated to the analysis of transcriptional regulation in *Mycobacterium tuberculosis***. *Bioinformatics* 2005, **21**(10):2563–2565.

17. Krawczyk J, Kohl TA, Goesmann A, Kalinowski J, Baumbach J: **From *Corynebacterium Glutamicum* to *Mycobacterium Tuberculosis* towards Transfers of Gene Regulatory Networks and Integrated Data Analyses with MycoRegNet**. *Nucleic Acids Research* 2009, **37**(14):e97–e97.

18. Pauling J, Rattger R, Tauch A, Azevedo V, Baumbach J: **CoryneRegNet 6.0 Updated database content, new analysis methods and novel features focusing on community demands**. *Nucleic Acids Research* 2012, **40**(D1):D610–D614.

19. Kazakov AE, Cipriano MJ, Novichkov PS, Minovitsky S, Vinogradov DV, Arkin A, Mironov AA, Gelfand MS, Dubchak I: **RegTransBase a database of regulatory sequences and interactions in a wide range of prokaryotic genomes**. *Nucleic Acids Research* 2007, **35**(Database issue):D407–D412.

20. Grote A, Klein J, Retter I, Haddad I, Behling S, Bunk B, Biegler I, Yarmolinetz S, Jahn D, Manch R: **PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes**. *Nucleic Acids Research* 2009, **37**(Database issue):D61–D65.

21. Parez AG, Angarica VE, Vasconcelos ATR, Collado-Vides J: **Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes**. *Nucleic Acids Research* 2007, **35**(Database issue):D132–D136.

22. Wu J, Zhao F, Wang S, Deng G, Wang J, Bai J, Lu J, Qu J, Bao Q: **cTFbase: a database for comparative genomics of transcription factors in cyanobacteria**. *BMC Genomics* 2007, **8**:104.

23. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JDJ, Bertin N, Chung S, Vidal M, Gerstein M: **Annotation Transfer Between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs**. *Genome Research* 2004, **14**(6):1107–1118.

24. Walhout AJM, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M: **Protein Interaction Mapping in *C. elegans* Using Proteins Involved in Vulval Development**. *Science* 2000, **287**(5450):116–122.

25. Sjolander K: **Phylogenomic inference of protein molecular function: advances and challenges**. *Bioinformatics* 2004, **20**(2):170–179.

26. Altenhoff AM, Dessimoz C: **Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods**. *PLoS Computational Biology* 2009, **5**:e1000262.

27. Salichos L, Rokas A: **Evaluating Ortholog Prediction Algorithms in a Yeast Model Clade**. *PLoS ONE* 2011, **6**(4):e18755.

28. Chen F, Mackey AJ, Vermunt JK, Roos DS: **Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes**. *PLoS ONE* 2007, **2**(4):e383.

29. Kuzniar A, van Ham RC, Pongor S, Leunissen JA: **The quest for orthologs: finding the corresponding gene across genomes**. *Trends in Genetics* 2008, **24**(11):539–551.

30. Koonin EV: **Orthologs, Paralogs, and Evolutionary Genomics**. *Annual Review of Genetics* 2005, **39**:309–338.

31. Tatusov RL, Koonin EV, Lipman DJ: **A Genomic Perspective on Protein Families**. *Science* 1997, **278**(5338):631–637.
32. Hulsen T, Vlieg Jd, Leunissen JA, Groenen PM: **Testing statistical significance scores of sequence comparison methods with structure similarity**. *BMC Bioinformatics* 2006, **7**:444.
33. Moreno-Hagelsieb G, Latimer K: **Choosing BLAST Options for Better Detection of Orthologs as Reciprocal Best Hits**. *Bioinformatics* 2008, **24**(3):319–324.
34. Price MN, Dehal PS, Arkin AP: **Orthologous Transcription Factors in Bacteria Have Different Functions and Regulate Different Genes**. *PLoS Computational Biology* 2007, **3**(9):e175.
35. Tekaia F, Yeramian E: **SuperPartitions: detection and classification of orthologs**. *Gene* 2012, **492**:199–211.
36. Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic local alignment search tool**. *Journal of Molecular Biology* 1990, **215**:403–410.
37. 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 Research* 1997, **25**(17):3389–3402.
38. Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J: **RSAT 2011: regulatory sequence analysis tools**. *Nucleic Acids Research* 2011, **39**(Web Server issue):W86–W91.
39. Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M, Vervisch E, Brohae S, van Helden J: **RSAT: regulatory sequence analysis tools**. *Nucleic Acids Research* 2008, **36**(Web Server issue):W119–W127.
40. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Research* 2007, **35**(Database issue):D61–D65.
41. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data**. *Database: The Journal of Biological Databases and Curation* 2011, **2011**:bar009.
42. Smoot ME, Ono K, Ruscheinski J, Wang P, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics* 2011, **27**(3):431–432.
43. Kim T, Park PJ: **Advances in analysis of transcriptional regulatory networks**. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 2011, **3**:21–35.
44. Emmert-Streib F, Dehmer M: **Networks for systems biology: conceptual connection of data and function**. *IET Systems Biology* 2011, **5**(3):185–207.
45. Ran L, Larsson J, Vigil-Stenman T, Nylander JAA, Ininbergs K, Zheng WW, Lapidus A, Lowry S, Haselkorn R, Bergman B: **Genome Erosion in a Nitrogen-Fixing Vertically Transmitted Endosymbiotic Multicellular Cyanobacterium**. *PLoS ONE* 2010, **5**(7):e11486.
46. Kim WY, Kang S, Kim BC, Oh J, Cho S, Bhak J, Choi JS: **SynechoNET: integrated protein-protein interaction database of a model cyanobacterium *Synechocystis* sp. PCC 6803**. *BMC Bioinformatics* 2008, **9**(Suppl 1):S20.
47. Memon D, Singh AK, Pakrasi HB, Wangikar PP: **A global analysis of adaptive evolution of operons in cyanobacteria**. *Antonie van Leeuwenhoek* 2012, **103**(2):331-346.
48. Nakao M, Okamoto S, Kohara M, Fujishiro T, Fujisawa T, Sato S, Tabata S, Kaneko T, Nakamura Y: **CyanoBase: the cyanobacteria genome database update 2010**. *Nucleic Acids Research* 2009, **38**(Database):D379–D381.
49. Collado-Vides J, Salgado H, Morett E, Gama-Castro S, Jimnez-Jacinto V, Martnez-Flores I, Medina-Rivera A, Muiz-Rascado L, Peralta-Gil M, Santos-Zavaleta A: **Bioinformatics Resources for the Study of Gene Regulation in Bacteria**. *Journal of Bacteriology* 2009, **191**:23–31.
50. Baumbach J, Tauch A, Rahmann S: **Towards the integrated analysis, visualization and reconstruction of microbial gene regulatory networks**. *Briefings in Bioinformatics* 2009, **10**:75–83.
51. Vinh NX, Chetty M, Coppel R, Wangikar PP: **GlobalMIT: learning globally optimal dynamic Bayesian network with the mutual information test criterion**. *Bioinformatics* 2011, **27**(19):2765–2766.
52. Xuan N, Chetty M, Coppel R, Wangikar P: **Gene regulatory network modeling via global optimization of high-order dynamic Bayesian network**. *BMC Bioinformatics* 2012, **13**:131.