PRODORIC2: the bacterial gene regulation database in 2018

Denitsa Eckweiler[1,*,†], Christian-Alexander Dudek[2,†], Juliane Hartlich[1], David Brötje[1] and Dieter Jahn[1]

[1]Institute of Microbiology and Braunschweig Integrated Centre of Systems Biology (BRICS), Technische Universität Braunschweig, Rebenring 56, Braunschweig D-38106, Germany and [2]Department of Bioinformatics and Biochemistry and Braunschweig Integrated Centre of Systems Biology (BRICS), Technische Universität Braunschweig, Rebenring 56, Braunschweig D-38106, Germany

ABSTRACT

Bacteria adapt to changes in their environment via differential gene expression mediated by DNA binding transcriptional regulators. The PRODORIC2 database hosts one of the largest collections of DNA binding sites for prokaryotic transcription factors. It is the result of the thoroughly redesigned PRODORIC database. PRODORIC2 is more intuitive and user-friendly. Besides significant technical improvements, the new update offers more than 1000 new transcription factor binding sites and 110 new position weight matrices for genome-wide pattern searches with the *Virtual Footprint* tool. Moreover, binding sites deduced from high-throughput experiments were included. Data for 6 new bacterial species including bacteria of the *Rhodobacteraceae* family were added. Finally, a comprehensive collection of sigma- and transcription factor data for the nosocomial pathogen *Clostridium difficile* is now part of the database. PRODORIC2 is publicly available at http://www.prodoric2.de.

INTRODUCTION

The adaptation process of a bacterial cell to its environment is often controlled at the transcriptional level (1,2). In this context, promoters of target genes are usually controlled by various transcriptional regulators in response to environmental stimuli. For many of these proteins DNA binding in the promoter region is required to influence transcription. Usually, a well-defined Transcription Factor Binding Site (TFBS) composed of a conserved DNA sequence is utilized for this process.

Traditional experimental techniques used to study TFBSs include DNAse footprinting (3), EMSA (4) or SELEX (5). With the emergence of microarrays and Next Generation Sequencing (NGS) techniques, numerous TFBSs can be efficiently recovered in a single experimental setup and bioinformatically mapped back to the genome. Such experiments are time- and resource efficient and generate vast amounts of data that can be easily processed by the current computational pipelines. Consequently, the traditional methods to study protein-DNA interaction are stepwise replaced by high-throughput methods like DNase-seq (6), FAIRE-seq (7) or ChIP-seq (8).

Initially, databases on transcriptional regulation in prokaryotes mainly relied on TFBSs annotated from the literature. However, now databases on transcriptional regulation have correspondingly started the curation and inclusion of TFBSs recovered from high-throughput experiments. Examples are the CollecTF database (9) that offers TFBSs collected across the *Bacteria* domain, as well as the *Escherichia coli* model organism database RegulonDB (10). Another resource, focused mostly on eukaryotic TFBSs, is the footprintDB (11) that integrates data from various other databases including RegulonDB and DBTBS (12), a database on the model organism *Bacillus subtilis*. Further typical model organism databases focused on prokaryotic gene regulation are the MycoRegNet (13) for *Mycobacterium tuberculosis* and CoryneRegNet (14) for *Corynebacteria*. The RegNet collection has been extended by the EhecRegNet (15) database on the human pathogenic *E. coli*. In contrast to these mainly manually curated resources, some databases (16) offer entirely bioinformatically propagated regulons, predicted using transcription factor position weight matrices (PWMs) compiled from experimental evidence (17). In conclusion, although a significant improvement of knowledge was achieved by advanced experimental approaches in the past years, there still remains much to learn about gene regulatory networks in bacteria. As recently described (18), it was estimated that 37% of the gene regulatory interactions of *E. coli* have already been dis-

*To whom correspondence should be addressed. Tel: +49 531 391 55290; Fax: +49 531 391 5854; Email: d.eckweiler@tu-braunschweig.de
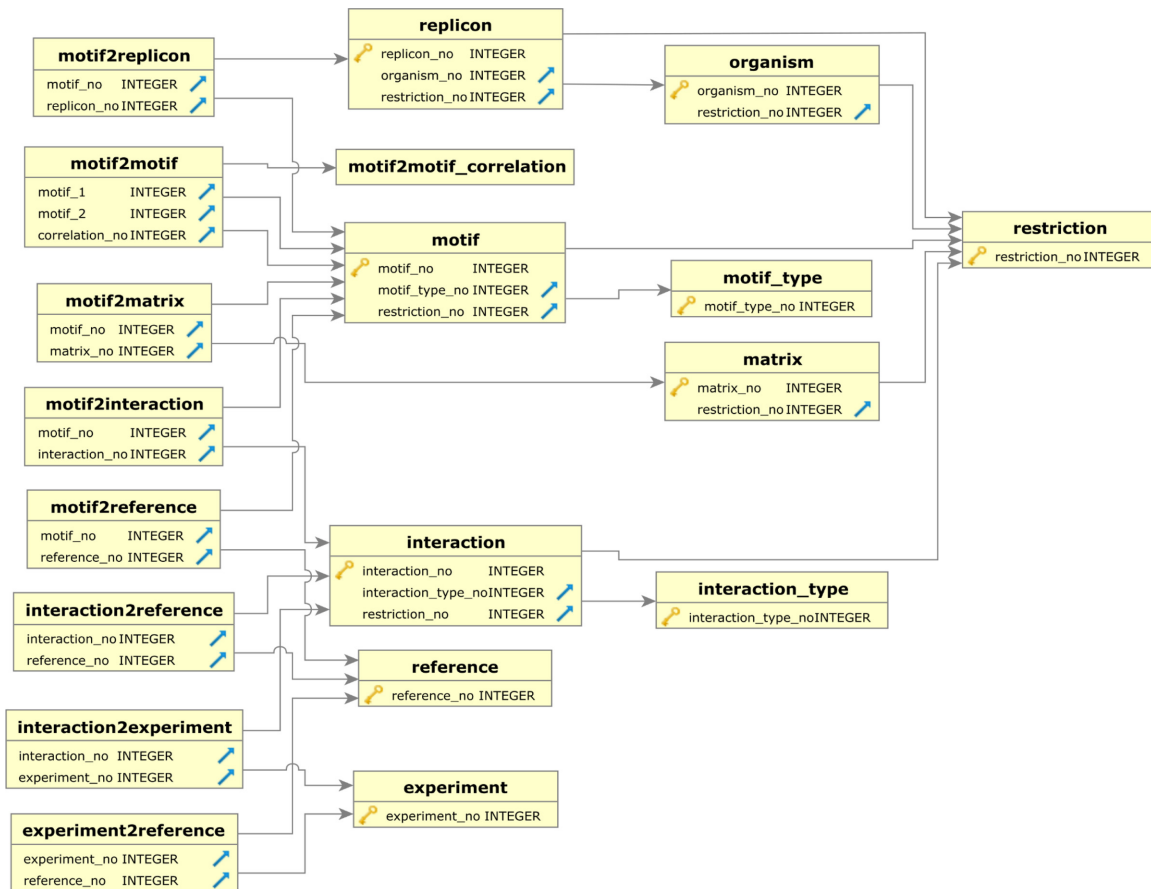†These authors contributed equally to this work as first authors.

Figure 1. Scheme of the PRODORIC2 database showing 19 of its 20 tables.

covered. In contrast only 24% of the corresponding interactions are known for *B. subtilis*.

The PRODORIC database was first introduced in 2003 (19). It initially offered gene regulation data for a few model organisms including *E. coli*, *B. subtilis* and *Pseudomonas aeruginosa*. Since 2005, PRODORIC has been associated with the prediction tool *Virtual Footprint* (20) that uses transcription factor- and species-specific aligned TFBSs compiled in position weight matrices (PWMs) to search prokaryotic genomes. Since the last PRODORIC update in 2009 (21) we have completely redesigned the database and the associated website. Here, we summarize the new features and the extended database content including TFBSs from the newly introduced species *Clostridia spp.* and *Dinoroseobacter shibae*.

MAJOR NEW DEVELOPMENTS AND IMPROVEMENTS

New database structure

The PRODORIC version from 2009 was a PostgreSQL object-relational database and consisted of 126 tables. Omitting static annotation (genome, gene and protein sequences) and experimental data available elsewhere (22) allowed us to design a platform-independent, portable (in its current version ~20MB large) SQLite database. The new structure is shown in Figure 1. A further advantage of the

new architecture is that it supports programmatic access under various programming languages without additional backend software such as servers. This makes the database easily distributable, portable and easy to embed into R packages.

All data associated with the 2009 update is also available on the original PRODORIC website (http://www.prodoric.de), to retain some of the original features of PRODORIC until they are ported to PRODORIC2. For the new PRODORIC2 version (http://www.prodoric2.de) we extracted the essentials of the existing database—TFBSs, interactions between TF and its TFBSs, PWMs, operon/promoter structures and all connected literature and organized them in 20 new tables in a completely redesigned SQLite architecture. The positions of the TFBSs were transferred on the new replicon versions as of 2015.

New database website

In parallel to the development of a new database architecture, the website of the PRODORIC database was updated correspondingly to meet its new demands. The new PRODORIC2 website uses a decent color scheme according to the corporate design of the Technische Universität Braunschweig. The new website uses PHP 7 as technical back-end and Ajax and jQuery as front-end support. The
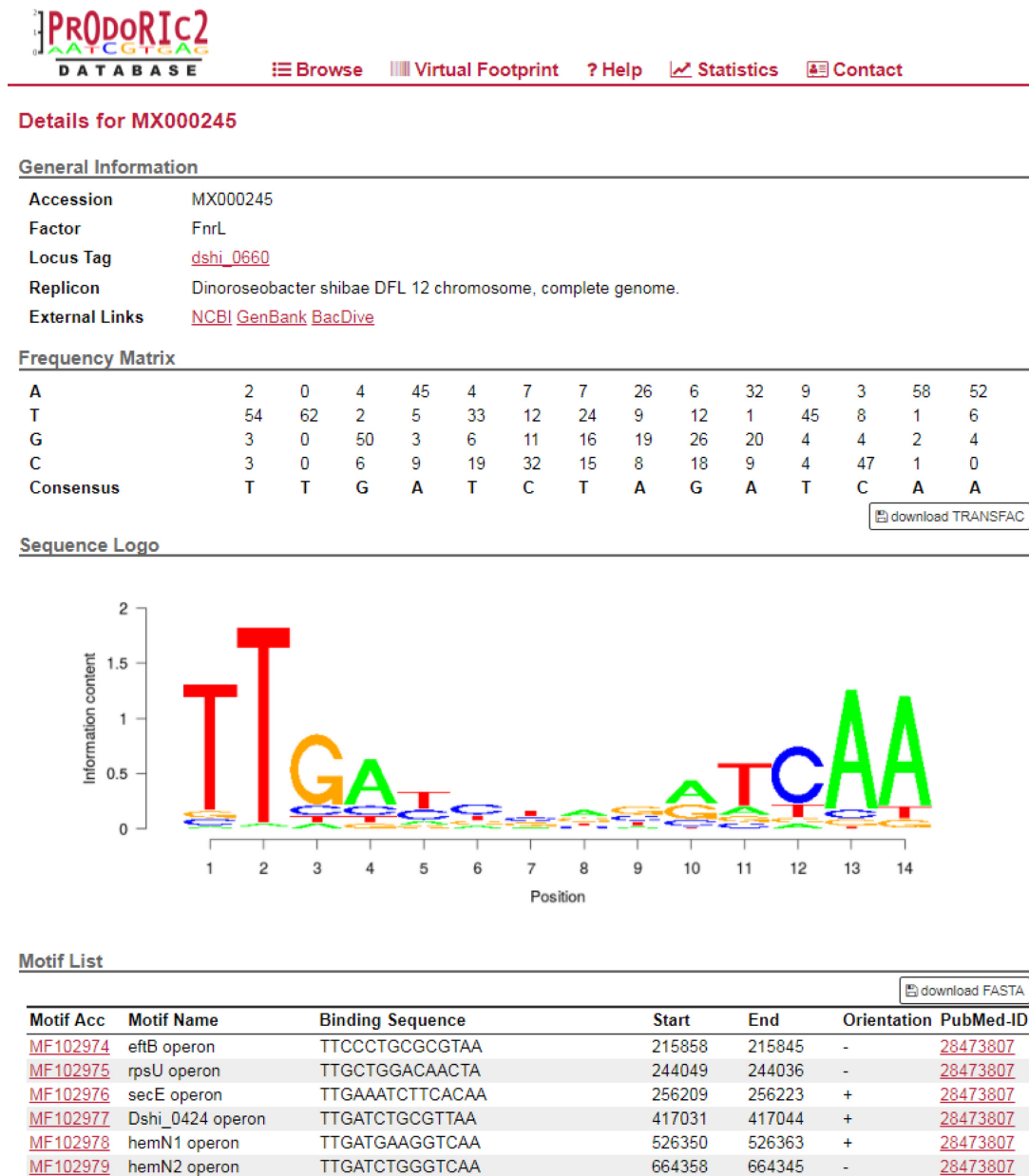
Figure 2. A matrix report page for the transcription factor FnrL from *D. shibae*. Links providing additional information about the reference strain and the transcription factor are depicted at the top of the page. The position weight matrix is shown in tabulated form and the corresponding binding sites (here truncated) are shown below the matrix as they appear in the original manuscript. Their genomic coordinates, strand orientation and the original citation(s) are provided as well. The accession numbers of the TFBSs are hyperlinked to the NCBI genome browser.

website runs on all major internet browsers in their latest versions. The user can start using PRODORIC2 with a search for a PWM or for a functional element. Here, functional elements are genes, promoters or operons. This is a user-friendly simplification of the previous PRODORIC search mask that offered multiple direct access to the information stored in the previous database version.

The user can now choose to type the name of the functional element of interest or can alternatively search for all functional elements/PWMs associated with a given bacterial species. The Ajax/jQuery search mask queries interactively the database and returns all hits containing the information entered by the user. The extracted information can be comfortably downloaded as a CSV-formatted file, which can be opened under any operating system. The database can be searched using the full-text search for motifs and matrices, which allows to search for accessions, organisms, transcription factors and locus tags. The user can also browse all PWMs and all non-artificial TFBSs associated with them. The PWM list can also be downloaded in CSV format for later use. The matrix report has been kept similar to that in the PRODORIC 2009 version. Figure 2 shows the matrix report page for one of the newly introduced PWMs of *D. shibae*. We have improved the presen-

Virtual Footprint Job IGZV0O

Sequence: Clostridium difficile 630 uid57679
Matrix: MX000296

Download as .csv

Start	Stop	Strand	Sequence	Score	Core Score	Locus Tag	Gene	Distance to ATG
15998	16035	+	AAATAGAATATAAAGTCAATATTGATTTATAATAAGAT	12.1636	6.64117	CD630_00120	CD630_00120	-75
17939	17976	+	AGGAGGAATTAATATGTTAGATATTAAAAGAATAAGAG	11.5783	6.64117	CD630_00140	serS1	-13
30093	30130	+	TTGTCTTATAAATAGGCAATATGTGGTTATAATAGCAA	12.0677	6.64117	CD630_r0060	5s rRNA	-22
34860	34897	+	ATAGAGTATGCTTAAAAAGTTTAATAAAATACTAGTGG	11.3031	6.64117	CD630_00210	pycA	-7
48016	48053	+	TTCTTGAATATAATATACACTTTGGTCTATAATAGATA	12.9066	6.64117	CD630_00290	CD630_00290	-60
52533	52570	+	ATTATAAATAGTAAATTTATAATGTTCAATATTAAAAA	11.3442	6.64117	CD630_00320	CD630_00320	-191
61609	61646	+	AATATGATTTGCTTAATTAAAATATAAGAAAATAAATA	11.3634	6.64117	CD630_00390	acoL	-50
69669	69706	+	TGGGAGGATTTAATTAATGTTATTAGAGAAATTAAGAA	11.9759	6.64117	CD630_00450	CD630_00450	-16
70443	70480	+	TAATAGTATATAAATTCATTTAAGACTAAAAATAATAA	12.5275	6.64117	CD630_00460	CD630_00460	-66
76841	76878	+	TATGAAATTAAAACTAGGATTTTCTGATATAATAAACA	11.3246	6.64117	CD630_00510	gltX	-156
78885	78922	+	TTGAAAGATAGTACGAAAAATGATAGTAATACTATTAT	11.3128	6.64117	CD630_00520	cysS	-114
85098	85135	+	TTTAAAGATATTATATTCACATATATTAACAATATATA	11.3378	6.64117	CD630_00581	rpmG	-69
89964	90001	+	ATATGGATTTTATATTGGAAAATTGTACATAATACATA	12.3768	6.64117	CD630_00660	rpoB	-196

Figure 3. *Virtual Footprint* output generated with a search using the *C. difficile* SigG PWM. The number of hits has been shortened. Strain information is available as a hyperlink to NCBI.

tation of the TFBSs by showing their original sequences as extracted from literature together with their genomic coordinates and strand orientation. This offers a more intuitive data presentation without the need to click on a particular TFBS in order to see detailed information. If the TFBSs related to a particular PWM have not been already aligned by the source publication, they were aligned manually with the MAFFT tool (https://www.ebi.ac.uk/Tools/msa/mafft/). In cases where the DNA sequence of a minority of TFBSs was longer than the rest, these were truncated after alignment and prior to PWM computation. The PWM can be downloaded in the TRANSFAC (23) format. The TFBSs can be exported as a Multi-FASTA file. Upon clicking on the TFBS motif ID, the NCBI genome browser opens as a new window and the TFBS sequence can be interactively inspected in the browser.

The *Virtual Footprint* tool is also accessible from the new PRODORIC2 website. To keep it user-friendly, only the most essential options of *Virtual Footprint* are offered. The user can choose among ~5200 locally stored genome sequence files in GenBank (22) format or can alternatively upload own sequence files in FASTA format. Regular strain updates are planned every 3–4 months. The options to limit the search for TFBSs to intergenic regions or to show only hits found within a certain distance to the start of the coding sequence can only be used in combination with GenBank files. The computational performance of *Virtual Footprint* has not changed in comparison to the previous release. Figure 3 shows a sample *Virtual Footprint* output generated using one of the new *C. difficile* PWMs. The output has been kept fairly similar to the previous *Virtual Footprint* version—the genomic coordinates and orientation of the particular hit are provided together with its score and core score values. Description of the scores can be found on the Help page. If a GenBank file has been selected as input, the locus tag, gene name, and distance to the start codon (ATG) are provided. The distance to ATG can be easily changed on the Virtual Footprint input page and the user can exclude possible (palindromic) hits on the reverse strand that have most probably lower genomic significance (option *Hide hits without genomic context*). The results can be downloaded in CSV format.

DATABASE CURATION AND CONTENT

Update with TFBS from new environmental and pathogenic bacteria

Since the last PRODORIC update, there has been a rapid progress in generation of TFBSs by high-throughput techniques that complete the results from classical techniques such as SELEX, EMSA or DNAse footprinting. In the new PRODORIC2 update we also introduce TFBS data obtained by RNA-seq (24), ChIP-seq, and microarrays in combination with bioinformatics based motif search. Binding sites predicted solely by computational approaches are not curated in the database. Where possible, TFBS confirmed additionally by another approach, such as EMSA, have been included. All data have been manually curated before entering it into the database. This involves computational validation of the position and orientation of the TFBS on the corresponding replicon. The mapping does not allow for mismatches. The new as well as the older TFBS data have been mapped to the current bacterial genome versions. The PRODORIC2 database contains now genomic information on 2274 bacterial species and their 5191 replicons without explicitly storing the genomic sequences in the database files as previously done. Instead, genomic sequences and corresponding features are stored locally in GenBank format as input for *Virtual Footprint*. This makes information update much more flexible. It represents a great improvement to the PRODORIC 2009 update, where the *Virtual Footprint* input summed up to total of database stored 696 genomes and their corresponding 1304 replicons.

A summary of the newly included bacterial species, numbers of curated TFBSs and PWMs is offered in Table 1. Considering the fact that 55 TFBSs failed to map to the current genome sequence of *E. coli*, there is a total increase of 25% of TFBSs in this update. The new PRODORIC2 release introduces TFBS data on six new bacterial species of the genus *Bacillus*, *Clostridia* and *Roseobacter*. Here, *Clostridium difficile* is of special clinical interest due to its drastically increased pathogenicity. We have curated PWMs on most of its sigma factors including those previously linked to sporulation. We also provide data on its major transcription factors, Fur, CodY, RgaR and Spo0A. An-

Table 1. Statistics of the PRODORIC2 content (September 2017)

Organism	TFBSs	PWMs[a]	Increase in %[b]
Bacillus licheniformis	8	3	*new*
Bacillus megaterium	7	2	*new*
Bacillus subtilis	907	91(65)	28
Bradyrhizobium japonicum	36	4(3)	33
Campylobacter jejuni	5	2	100
Clostridium acetobutylicum	23	3	*new*
Clostridium beijerinckii	7	2	*new*
Clostridium difficile 630	266	13(12)	*new*
Corynebacterium glutamicum	59	6	500
Dinoroseobacter shibae	255	6	*new*
Escherichia coli CFT073	4	2	100
Escherichia coli str. K-12 substr. MG1655	1562	94(82)	7
Escherichia coli str. K-12 substr. W3110	76	6	50
Helicobacter pylori	24	3	50
Listeria monocytogenes	62	5	150
Mycobacterium tuberculosis	73	5	400
Pseudomonas aeruginosa PAO1	444	41(36)	86
Rhodobacter sphaeroides	38	6	50
Salmonella enterica	6	1	-
Staphylococcus aureus COL	34	2	100
Staphylococcus aureus Newman	2	1	-
Streptococcus agalactiae	5	1	-
Streptococcus pyogenes MGAS5005	4	1	-
Streptococcus pyogenes MGAS8232	10	2	100
Streptococcus pyogenes	3	1	-
Synechococcus elongatus	11	1	-
Synechocystis	16	3	200
Sum	3947	307(261)	31

[a]The non-redundant number of PWMs (in parentheses).
[b]Based on the total numbers.

other pathogenic species, for which there were already 19 non-redundant PWMs available in the PRODORIC 2009 release, is the opportunistic nosocomial pathogen *Pseudomonas aeruginosa*. In the present release we included 14 PWMs for this pathogen that are already available in the CollecTF database (9). Altogether 20 other PWMs associated with the species *Corynebacterium glutamicum*, *Listeria monocytogenes*, *Helicobacter pylori,* and *Mycobacterium tuberculosis* were included from the same database. All of them were linked to the same replicon sequences as realized in CollecTF. The genomic positions of the binding sites were bioinformatically confirmed before importing them into the PRODORIC2 database. Out of the 110 new PWMs and 1026 TFBSs introduced in this update, 34 PWMs and corresponding 359 TFBSs came from CollecTF. Import from CollecTF is acknowledged in the 'Description' field of the corresponding matrix report. Overall, there is a significant increase of 31% in the total number of PWMs in comparison to the 2009 release.

LINKS TO OTHER RESOURCES

We have completely revised the access to the information available for the single transcription factor of interest. In the PRODORIC 2009 update the particular TF was linked to the corresponding Uniprot entry (25), however, many of those entries have become obsolete in the past years. Therefore, we manually relinked the TFs associated with the 261 non-redundant PWMs to the corresponding gene loci as available in the KEGG (26) database. This has a major advantage with regard to future developments of

PRODORIC2. There we like to integrate metabolic data and link to comprehensive resources on enzymes, such as the BRENDA (27) database. Following the link to KEGG, the user immediately finds all relevant sequence and annotation information related to the TF, including further linking to Uniprot and other major resources. Since the KEGG gene locus report offers a comprehensive genome browser, it was not necessary to include the GBpro browser that is further available on the PRODORIC 2009 website. The link to KEGG is found at the top of the PRODORIC2 matrix report page. There the user can additionally take advantage of the newly included *INSDC*-conforming link to the corresponding replicon sequence stored on the GenBank (22) and the link to the strain entry in the BacDive (28) database. Wherever possible, all TFs and their replicons in PRODORIC2 are connected to the corresponding BacDive entries.

CONCLUSIONS

Redesigning the database content and the website structure greatly improved the appearance and performance of the new PRODORIC2 database. The new version contains more than 1000 new TFBSs, which were used to build 110 new PWMs. These PWMs are readily available for their use in pattern searches employing the *Virtual Footprint* tool, which is accessible on the same page. TFBS and PWM data on six new bacterial species have been introduced in this update, whereas special attention has been devoted to the emerging pathogen *C. difficile*. The new update includes also TFBSs detected by diverse high-throughput

techniques. We have made first steps to curate more PWMs of bacterial species that were rather poorly represented in the PRODORIC 2009 update. Our goal is to continue this effort for example for *Staphylococcus aureus* and *Salmonella* spp. Moreover, we aim to introduce more high-throughput data such as gene expression, ChIP-seq or TSS-seq data in form of webservers connected to PRODORIC2. Overall, PRODORIC2 provides a solid basis for the prediction of gene regulatory networks and in the long run it will offer services for comparing regulatory networks between different species (29). The new structure also provides a solid basis for our future efforts to combine obtained results with experimental high-throughput transcriptome data and integrate those with enzyme and metabolic data.

In future versions, more features and options will be introduced in PRODORIC2. For example, more export options could be implemented, such as Cytoscape (30) compatible JSON format to connect gene regulatory networks to Cytoscape using suitable plugins (31). Furthermore, along with the growing number of PWMs, a tool for their similarity clustering could become important for the database.

REFERENCES

1. Tielen,P., Schobert,M., Härtig,E. and Jahn,D. (2012) Anaerobic regulatory networks in bacteria. In: Filloux,AAM (ed). *Bacterial Regulatory Networks*. Horizon Academic Press, Norwich, pp. 273–305.
2. Hartig,E. and Jahn,D. (2012) Regulation of the anaerobic metabolism in Bacillus subtilis. *Adv. Microb. Physiol.*, **61**, 195–216.
3. 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.
4. Garner,M.M. and Revzin,A. (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res.*, **9**, 3047–3060.
5. Oliphant,A.R., Brandl,C.J. and Struhl,K. (1989) Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol. Cell. Biol.*, **9**, 2944–2949.
6. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
7. Giresi,P.G., Kim,J., McDaniell,R.M., Iyer,V.R. and Lieb,J.D. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17**, 877–885.
8. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
9. Kilic,S., White,E.R., Sagitova,D.M., Cornish,J.P. and Erill,I. (2014) CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. *Nucleic Acids Res.*, **42**, D156–D160.
10. Gama-Castro,S., Salgado,H., Santos-Zavaleta,A., Ledezma-Tejeida,D., Muniz-Rascado,L., Garcia-Sotelo,J.S., Alquicira-Hernandez,K., Martinez-Flores,I., Pannier,L., Castro-Mondragon,J.A. *et al.* (2016) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.*, **44**, D133–D143.
11. Sebastian,A. and Contreras-Moreira,B. (2014) footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*, **30**, 258–265.
12. 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.
13. Krawczyk,J., Kohl,T.A., Goesmann,A., Kalinowski,J. and Baumbach,J. (2009) From Corynebacterium glutamicum to Mycobacterium tuberculosis–towards transfers of gene regulatory networks and integrated data analyses with MycoRegNet. *Nucleic Acids Res.*, **37**, e97.
14. Pauling,J., Rottger,R., Tauch,A., Azevedo,V. and Baumbach,J. (2012) CoryneRegNet 6.0–Updated database content, new analysis methods and novel features focusing on community demands. *Nucleic Acids Res.*, **40**, D610–D614.
15. Pauling,J., Rottger,R., Neuner,A., Salgado,H., Collado-Vides,J., Kalaghatgi,P., Azevedo,V., Tauch,A., Puhler,A. and Baumbach,J. (2012) On the trail of EHEC/EAEC–unraveling the gene regulatory networks of human pathogenic Escherichia coli bacteria. *Integr. Biol. (Camb)*, **4**, 728–733.
16. Novichkov,P.S., Kazakov,A.E., Ravcheev,D.A., Leyn,S.A., Kovaleva,G.Y., Sutormin,R.A., Kazanov,M.D., Riehl,W., Arkin,A.P., Dubchak,I. *et al.* (2013) RegPrecise 3.0–a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics*, **14**, 745.
17. 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.
18. Rottger,R., Ruckert,U., Taubert,J. and Baumbach,J. (2012) How little do we actually know? On the size of gene regulatory networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **9**, 1293–1300.
19. Munch,R., Hiller,K., Barg,H., Heldt,D., Linz,S., Wingender,E. and Jahn,D. (2003) PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res.*, **31**, 266–269.
20. Munch,R., Hiller,K., Grote,A., Scheer,M., Klein,J., Schobert,M. and Jahn,D. (2005) Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics*, **21**, 4187–4189.
21. 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.
22. Coordinators,N.R. (2017) Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **45**, D12–D17.
23. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
24. Morin,R., Bainbridge,M., Fejes,A., Hirst,M., Krzywinski,M., Pugh,T., McDonald,H., Varhol,R., Jones,S. and Marra,M. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques*, **45**, 81–94.
25. Pundir,S., Martin,M.J. and O'Donovan,C. (2017) UniProt Protein Knowledgebase. *Methods Mol. Biol.*, **1558**, 41–55.
26. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.

27. Placzek,S., Schomburg,I., Chang,A., Jeske,L., Ulbrich,M., Tillack,J. and Schomburg,D. (2017) BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.*, **45**, D380–D388.

28. Sohngen,C., Podstawka,A., Bunk,B., Gleim,D., Vetcininova,A., Reimer,L.C., Ebeling,C., Pendarovski,C. and Overmann,J. (2016) BacDive–The Bacterial Diversity Metadatabase in 2016. *Nucleic Acids Res.*, **44**, D581–D585.

29. Barh,D., Gupta,K., Jain,N., Khatri,G., Leon-Sicairos,N., Canizalez-Roman,A., Tiwari,S., Verma,A., Rahangdale,S., Shah Hassan,S. *et al.* (2013) Conserved host-pathogen PPIs. Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in C. pseudotuberculosis, C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis, and E. coli targeted by Piper betel compounds. *Integr. Biol. (Camb.)*, **5**, 495–509.

30. Su,G., Morris,J.H., Demchak,B. and Bader,G.D. (2014) Biological network exploration with Cytoscape 3. *Curr. Protoc. Bioinformatics*, **47**, 11–24.

31. Baumbach,J. and Apeltsin,L. (2008) Linking Cytoscape and the corynebacterial reference database CoryneRegNet. *BMC Genomics*, **9**, 184.