JB Journals.ASM.org

COMMENTARY

Every Site Counts: Submitting Transcription Factor-Binding Site Information through the CollecTF Portal

Ivan Erill

Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, Maryland, USA

Experimentally verified transcription factor-binding sites represent an information-rich and highly applicable data type that aptly summarizes the results of time-consuming experiments and inference processes. Currently, there is no centralized repository for this type of data, which is routinely embedded in articles and extremely hard to mine. CollecTF provides the first standardized resource for submission and deposition of these data into the NCBI RefSeq database, maximizing its accessibility and prompting the community to adopt direct submission policies.

In order to tackle the complexity of living systems, research in biology has become increasingly reliant on the sharing, availability, and reuse of experimental data. The usefulness of experimental data depends on its specific nature and quality, underlying documentation and standards, availability, and range of applicability. The results of DNA sequencing assays were recognized early in the history of bioinformatics as a precise and broadly usable type of data, leading to the creation of open repositories, the definition of worldwide standards for the submission of DNA sequence data, and eventually, the mandatory requirement for its submission to open repositories when publishing results based on its analysis (1–3). The policy of direct submission by authors has since been adopted for the results of many other experimental protocols. Due to the sheer volume of data generated, policies, standards, and dedicated repositories for direct submission have been created primarily for high-throughput assays such as DNA arrays, high-throughput RNA sequencing (RNA-seq), chromatin immunoprecipitation-DNA sequencing (ChIP-seq), X-ray crystallography, or metagenomic sequencing (4–8). In contrast, data compilation for many other types of experimental results often relies on the manual or automated curation of published literature and lacks central repositories and standards, limiting the availability, visibility, and applicability of a large volume of experimental research in biology.

The identification of transcription factor (TF)-binding sites in the *Bacteria* domain provides an illustrative example of the drawbacks originating from the lack of standards, submission policies, and central repositories that still applies to many experimental protocols. An experimentally validated TF-binding site is a high-quality, information-dense annotation element that subsumes several lines of experimental evidence and prior knowledge to define with base pair resolution the location, span, and sequence of a region bound by a transcription factor on a bacterial DNA molecule. Identifying and validating the precise binding locations of a transcription factor and the regulatory effects of such binding is a time-consuming process that typically combines several experimental steps (e.g., DNase footprinting, electromobility shift assays, beta-galactosidase reporter assays, and ChIP-PCR) with inference based on prior knowledge (e.g., protein structure, binding motif, or promoter architecture). Due to their high specificity and precision and their mapping to a high-quality reference (genome sequence), bacterial TF-binding site annotations can be effectively reused in multiple settings. Known TF-binding sites on a given genome can for instance complement and provide a reference for other data sources (e.g., transcriptomic data) in inferring regulatory networks, and collections of aligned binding sites for a given transcription factor can be used to perform prospective searches for additional genomic targets or leveraged in comparative genomics analyses of regulatory networks (9–15).

The significant investment required for the experimental identification of TF-binding sites and their manifest applicability in subsequent research make them ideal targets for systematic annotation and dissemination using centralized repositories and standards for direct submission. In practice, however, the lack of these resources means that reports of TF-binding sites are typically embedded in articles. The variety of formats in which TF-binding sites are reported (e.g., as coordinates in a table or as boxed sequences in a figure) and the multiple sources of evidence, prior knowledge, and lines of inference usually employed in their identification make it extremely difficult to create automated mining algorithms capable of extracting and documenting this information in a reliable manner. As a result, efforts to compile bacterial TF-binding site information must rely on the manual curation of published literature by trained biocurators, with recent efforts at automation focusing only on prefetching of relevant articles (16–22). With limited funding, reliance on time-consuming manual curation of published literature imposes severe restrictions on the breadth and completeness of databases dedicated to annotating TF-binding sites. Given these constraints, many databases have opted to limit their scope to specific model organisms, allowing them to focus the curation effort and enhance the quality and completeness of the data compiled (16, 17, 19, 20). RegulonDB, which annotates transcriptional regulatory data for *Escherichia*

Accepted manuscript posted online 26 May 2015

Citation Erill I. 2015. Every site counts: submitting transcription factor-binding site information through the CollecTF portal. J Bacteriol 197:2454–2457. doi:10.1128/JB.00031-15.

Editor: I. B. Zhulin

Address correspondence to erill@umbc.edu.

doi:10.1128/JB.00031-15

The views expressed in this Commentary do not necessarily reflect the views of the journal or of ASM.

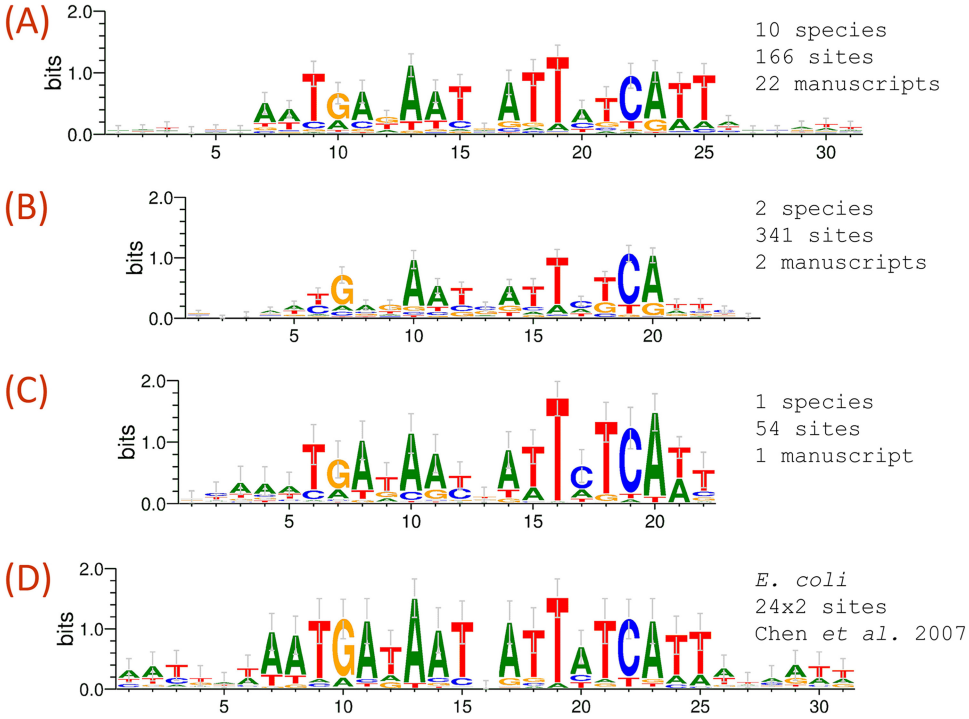

FIG 1 Example of the customizable query system implemented in CollecTF. (A to D) Dynamically generated sequence logos (28) for queries on Fur-binding sites detected through EMSA or DNase footprinting in *Gammaproteobacteria* (A), reported by chromatin immunoprecipitation with microarray technology (ChIP-chip) or ChIP-seq in *Gammaproteobacteria* (B), reported by ChIP-chip or ChIP-seq in *Vibrio cholerae* (C) and as reported for *E. coli* (D) in a reference manuscript (29). To generate logos, sites matching the query are dynamically aligned using LASAGNA to define site orientation and the window of conservation above background (30). All the site data and metadata used to generate the logos are available for download in a variety of export formats (e.g., FASTA, CSV, or PSSM).

coli, is a prime exponent of this strategy, and the use of RegulonDB data sets in hundreds of published research papers illustrates the broad applicability of high-quality TF-binding site data (20). By definition, however, organism-based databases omit the substantial body of experimental research performed on nonmodel organisms, providing a fragmented picture of transcriptional regulation in *Bacteria*. Furthermore, the available data remain scattered across several databases, and the use of different standards complicates the compilation of multispecies data sets for comparative studies. Some attempts at domain-wide, unified annotation of bacterial TF-binding sites have been made, but the reliance on manual curation by a relatively small team of curators means that these resources eventually struggle to keep up with an increasing amount of experimental research (18, 21). Efforts directed at expanding the pool of curators by involving the research community in the annotation process, such as the ORegAnno database on eukaryotic regulatory elements, have also met with limited success due to the substantial amount of time required for proper curation and the lack of clear incentives for participation (23).

With the advent of high-throughput techniques for mapping TF-binding sites (e.g., ChIP-seq) and the increasing standardization, parallelization, and ability to outsource more traditional methods (e.g., electrophoretic mobility shift assay [EMSA]), it has become increasingly apparent that manual curation of published literature is not a sustainable model for centralizing and making available domain-wide TF-binding site data. As is the case with genomic sequences or DNA array experiments, the only sustainable model for compiling TF-binding site information on a domain-wide level is to promote its direct submission by authors. This entails the creation of standards and interfaces for submission of this information to a stable and highly accessible public repository. CollecTF was born as an effort to provide an open platform for the sustainable annotation of information on TF-binding sites and their regulatory effects across the *Bacteria* domain (24). To achieve this goal, CollecTF implements a dual annotation process that combines in-house curation of published literature by trained biocurators with direct submissions from authors. CollecTF annotates experimentally validated, naturally occurring TF-binding sites in *Bacteria*, documenting the experimental evidence supporting them through a standardized vocabulary and mapping transcription factors, their verified sites, and regulated genes to reference sequence databases. When available, CollecTF also stores links to the primary data sources used in the annotation, such as GEO or ArrayExpress accession numbers for ChIP-seq reads (5, 8). Submitted annotation data are fully accessible, and users can customize data retrieval at multiple levels, such as defining the level of experimental support or aggregating data across taxonomic groups or transcription factor families (Fig. 1). Direct annotation of TF-binding sites by authors is facilitated by means of an easy-to-use graphical interface that automates the mapping process and guides the author through the annotation process, ensuring compliance with defined standards and accurate database cross-references (Fig. 2).

A fundamental component in the genesis of CollecTF was the realization that TF-binding sites represent well-defined elements of genomic sequences: the binding of a protein to a specific location and sequence of the DNA molecule that is often associated with regulatory effects on nearby genes. Hence, much like coding sequences, experimentally verified TF-binding sites and their reg-

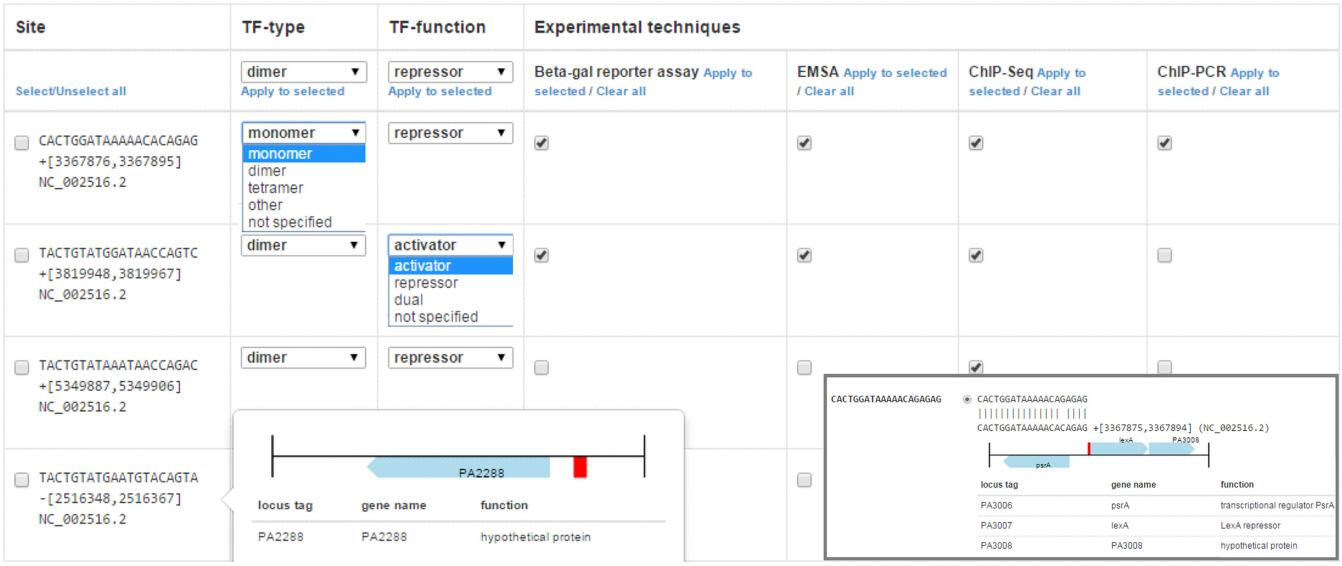

FIG 2 Details of the CollecTF site annotation step, where submitters select the specific techniques used to identify sites, as well as the mode of action and conformation, if known, of the transcription factor. Annotation is performed on sites entered as sequences or coordinates and previously mapped to the RefSeq genome record (inlet). Extensive documentation for the curation process is available on the CollecTF website (http://www.collectf.org/static/CollecTF_submission_guide.pdf) (24).

ulatory effects can and should be annotated as features in genome records. In collaboration with the NCBI RefSeq team, CollecTF has developed standards for the inclusion of TF-binding site information in RefSeq genome sequences using the "protein_bind" feature identifier (Fig. 3). These records define the genomic location of the TF-binding site as well as the bound protein, and capture the experimental evidence for TF-binding sites using "/experiment" tags that include the PubMed identifier of the publications providing such evidence. Data deposited in CollecTF are periodically submitted to the NCBI to populate RefSeq genomes. In addition, CollecTF is currently working to incorporate gene ontology (GO) and evidence ontology (ECO) terms in its annotations to provide interoperability with the EBI UniProt database (25–27). Hence, in contrast to previous initiatives, CollecTF operates both as a conventional database and as a portal for submission of TF-binding site annotations to NCBI. As a database, CollecTF facilitates direct and highly customizable access to annotated TF-binding site data for its application in subsequent research. As a portal, CollecTF provides the means to maximize the visibility and guarantee the long-term accessibility of submitted data by incorporating it into a centralized, stable, and widely accessed reference resource for genomic data.

At CollecTF we firmly believe that the future of biological research hinges on its ability to effectively disseminate and reuse experimental data. Experimental evidence for TF-binding sites is currently not deposited in any centralized repository, and consequently, the significant investment required for its generation can yield only limited returns in terms of scientific impact. The CollecTF team is committed to maintaining its curation effort to make bacterial TF-binding site annotations widely available, but the ultimate success of this initiative depends on the adoption of a culture of direct submission by experimental researchers. *Journal of Bacteriology* articles currently represent more than one-third of curated manuscripts in CollecTF, and the Editorial Board of the *Journal of Bacteriology* has graciously endorsed CollecTF as a vehicle for the public deposition of TF-binding site information to the NCBI. On behalf of the CollecTF team, I urge the readership of this journal to consider submitting the results of upcoming experiments on transcriptional regulation to CollecTF. There are obvious advantages to submitting your results to CollecTF and having it populate NCBI RefSeq genomes, such as increased vis-

```
protein_bind    complement(1559178..1559199)
                /experiment="ChIP-chip [PMID:19682264]"
                /experiment="ChIP-PCR [PMID:19682264]"
                /experiment="DNA-array expression analysis [PMID:19682264]"
                /experiment="DNAse footprinting [PMID:15505212,18045385]"
                /experiment="EMSA [PMID:19682264,15505212,18045385]"
                /note="Transcription factor binding site for NP_250121.;
                 Evidence of regulation for: PA1431, PA1432"
                /bound_moiety="LasR"
                /db_xref="CollecTF:EXPSITE_00008b40"
```

FIG 3 Details of a CollecTF "protein_bind" feature extracted from the *Pseudomonas aeruginosa* PAO1 genome sequence (NC_002516.2). The transcription factor (LasR) is identified as the "/bound_moiety," and its protein accession number is provided in the "/note" field, together with regulated genes. The experimental support for this LasR-binding site comes several lines of evidence reported using the "/experiment" tag. The PubMed identifiers (PMID) for the scientific papers providing such evidence are listed next to the evidence description. A "/db_xref" field provides a link to the CollecTF record to explore the data integration and curation process for the reported site.

ibility that may result in additional citations and spur collaborations, but the key incentive for submission should lie in the knowledge that by making your results broadly accessible, you are driving forward research in your field of interest. As the CollecTF team is fond of saying: in transcriptional regulation, every site counts; help us make yours count too.

ACKNOWLEDGMENTS

CollecTF is supported by the U.S. National Science Foundation award MCB-1158056.

I thank Stacy Ciufo and Tatiana Tatusova from the National Center for Biotechnology Information RefSeq team, as well as Sefa Kılıç, Dinara Sagitova, Elliot White, and all the other members of the University of Maryland Baltimore County (UMBC) CollecTF team.

REFERENCES

1. **Burks C, Fickett JW, Goad WB, Kanehisa M, Lewitter FI, Rindone WP, Swindell CD, Tung CS, Bilofsky HS.** 1985. The GenBank nucleic acid sequence database. Comput Appl Biosci Comput Appl Biosci (Camb) **1:**225–233.
2. **Baker W, van den Broek A, Camon E, Hingamp P, Sterk P, Stoesser G, Tuli MA.** 2000. The EMBL Nucleotide Sequence Database. Nucleic Acids Res **28:**19–23. http://dx.doi.org/10.1093/nar/28.1.19.
3. **Cochrane G, Karsch-Mizrachi I, Nakamura Y.** 2011. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res **39:**D15–D18. http://dx.doi.org/10.1093/nar/gkq1150.
4. **Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE.** 2000. The Protein Data Bank. Nucleic Acids Res **28:**235–242. http://dx.doi.org/10.1093/nar/28.1.235.
5. **Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R.** 2005. NCBI GEO: mining millions of expression profiles–database and tools. Nucleic Acids Res **33:**D562–D566. http://dx.doi.org/10.1093/nar/gki022.
6. **Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA.** 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics **9:**386. http://dx.doi.org/10.1186/1471-2105-9-386.
7. **Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Pillay M, Ratner A, Huang J, Pagani I, Tringe S, Huntemann M, Billis K, Varghese N, Tennessen K, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC.** 2014. IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Res **42:**D568–D573. http://dx.doi.org/10.1093/nar/gkt919.
8. **Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A.** 2015. ArrayExpress update–simplifying data submissions. Nucleic Acids Res **43:**D1113–D1116. http://dx.doi.org/10.1093/nar/gku1057.
9. **Shen-Orr SS, Milo R, Mangan S, Alon U.** 2002. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet **31:**64–68. http://dx.doi.org/10.1038/ng881.
10. **Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS.** 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol **5:**e8. http://dx.doi.org/10.1371/journal.pbio.0050008.
11. **Babu MM, Lang B, Aravind L.** 2009. Methods to reconstruct and compare transcriptional regulatory networks. Methods Mol Biol **541:**163–180. http://dx.doi.org/10.1007/978-1-59745-243-4_8.
12. **Zare H, Sangurdekar D, Srivastava P, Kaveh M, Khodursky A.** 2009. Reconstruction of Escherichia coli transcriptional regulatory networks via regulon-based associations. BMC Syst Biol **3:**39. http://dx.doi.org/10.1186/1752-0509-3-39.
13. **Erill I, O'Neill MC.** 2009. A reexamination of information theory-based methods for DNA-binding site identification. BMC Bioinformatics **10:**57. http://dx.doi.org/10.1186/1471-2105-10-57.
14. **Sanchez-Alberola N, Campoy S, Barbe J, Erill I.** 2012. Analysis of the SOS response of Vibrio and other bacteria with multiple chromosomes. BMC Genomics **13:**58. http://dx.doi.org/10.1186/1471-2164-13-58.
15. **Rodionov DA, Rodionova IA, Li X, Ravcheev DA, Tarasova Y, Portnoy VA, Zengler K, Osterman AL.** 2013. Transcriptional regulation of the carbohydrate utilization network in Thermotoga maritima. Front Microbiol **4:**244. http://dx.doi.org/10.3389/fmicb.2013.00244.
16. **Ishii T, Yoshida K, Terai G, Fujita Y, Nakai K.** 2001. DBTBS: a database of Bacillus subtilis promoters and transcription factors. Nucleic Acids Res **29:**278–280. http://dx.doi.org/10.1093/nar/29.1.278.
17. **Jacques P-É, Gervais AL, Cantin M, Lucier J-F, Dallaire G, Drouin G, Gaudreau L, Goulet J, Brzezinski R.** 2005. MtbRegList, a database dedicated to the analysis of transcriptional regulation in Mycobacterium tuberculosis. Bioinformatics **21:**2563–2565. http://dx.doi.org/10.1093/bioinformatics/bti321.
18. **Grote A, Klein J, Retter I, Haddad I, Behling S, Bunk B, Biegler I, Yarmolinetz S, Jahn D, Münch R.** 2009. PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. Nucleic Acids Res **37:**D61–D65. http://dx.doi.org/10.1093/nar/gkn837.
19. **Pauling J, Röttger R, Tauch A, Azevedo V, 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. http://dx.doi.org/10.1093/nar/gkr883.
20. **Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, Weiss V, Solano-Lira H, Martínez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernández S, Alquicira-Hernández K, López-Fuentes A, Porrón-Sotelo L, Huerta AM, Bonavides-Martínez C, Balderas-Martínez YI, Pannier L, Olvera M, Labastida A, Jiménez-Jacinto V, Vega-Alvarado L, Del Moral-Chávez V, Hernández-Alvarez A, Morett E, Collado-Vides J.** 2013. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res **41:**D203–D213. http://dx.doi.org/10.1093/nar/gks1201.
21. **Cipriano MJ, Novichkov PN, Kazakov AE, Rodionov DA, Arkin AP, Gelfand MS, Dubchak I.** 2013. RegTransBase–a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. BMC Genomics **14:**213. http://dx.doi.org/10.1186/1471-2164-14-213.
22. **Gama-Castro S, Rinaldi F, Lopez-Fuentes A, Balderas-Martinez YI, Clematide S, Ellendorff TR, Santos-Zavaleta A, Marques-Madeira H, Collado-Vides J.** 2014. Assisted curation of regulatory interactions and growth conditions of OxyR in E. coli K-12. Database (Oxford) **2014:**bau049. http://dx.doi.org/10.1093/database/bau049.
23. **Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, Griffith M, Gallo SM, Giardine B, Hooghe B, Van Loo P, Blanco E, Ticoll A, Lithwick S, Portales-Casamar E, Donaldson IJ, Robertson G, Wadelius C, De Bleser P, Vlieghe D, Halfon MS, Wasserman W, Hardison R, Bergman CM, Jones SJ.** 2008. ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res **36:**D107–D113. http://dx.doi.org/10.1093/nar/gkm967.
24. **Kiliç S, White ER, Sagitova DM, Cornish JP, Erill I.** 2014. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. Nucleic Acids Res **42:**D156–D160. http://dx.doi.org/10.1093/nar/gkt1123.
25. **Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G.** 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25:**25–29. http://dx.doi.org/10.1038/75556.
26. **Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M.** 2014. Standardized description of scientific evidence using the Evidence Ontology (ECO). Database (Oxford) **2014:**bau075. http://dx.doi.org/10.1093/database/bau075.
27. **UniProt Consortium.** 2015. UniProt: a hub for protein information. Nucleic Acids Res **43:**D204–D212. http://dx.doi.org/10.1093/nar/gku989.
28. **Schneider TD, Stephens RM.** 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res **18:**6097–6100. http://dx.doi.org/10.1093/nar/18.20.6097.
29. **Chen Z, Lewis KA, Shultzaberger RK, Lyakhov IG, Zheng M, Doan B, Storz G, Schneider TD.** 2007. Discovery of Fur binding site clusters in Escherichia coli by information theory models. Nucleic Acids Res **35:**6762–6777. http://dx.doi.org/10.1093/nar/gkm631.
30. **Lee C, Huang C-H.** 2013. LASAGNA: a novel algorithm for transcription factor binding site alignment. BMC Bioinformatics **14:**108. http://dx.doi.org/10.1186/1471-2105-14-108.