DOOR: A microbial operon database for gene and genome organization and function discovery

Huansheng Cao, Qin Ma, Xin Chen, and Ying Xu

[1]Huansheng Cao was a Post-Doctoral Researcher in the Department of Biochemistry and Molecular Biology at the University of Georgia and is a Research Assistant Professor at Arizona State University. His interests include systems biology and multi-omics integration.

Qin Ma…

Xin Chen was a PhD student in the Computational Systems Biology Laboratory in the Department of Biochemistry and Molecular Biology at the University of Georgia and is an Assistant Professor in the Center for Applied Mathematics at Tianjin University, China

Ying Xu…..

Corresponding author: xyn@uga.edu

ABSTRACT

The rapid increase of fully sequenced prokaryotic genomes provides unprecedented information for discovery of novel biological knowledge. The organization and function of genes and genomes can be revealed by mining these data with appropriate computational technologies. Here we present the Database of prOkaryotic OpeRons (DOOR), which contains 6,975,454 conserve operons in 2,072 complete genomes. Based on these identified operons and other omic data (e.g., RNA-seq and ChIP-seq data), we have also developed multiple algorithms and tools that can identify transcription units (TUs) under a specific condition/environment, analyze co-expression relationship among operons, and *de-novo* predict *cis*-regulatory motifs. Based on above functionalities, more advanced insights have been derived: the global arrangement of operons in a bacterial genome is largely influenced by the tendency with which a bacterium keeps its operons encoding the same biological pathways in genomic vicinity.

Bacterial genomes are partitioned into a set of folding domains such that the total unfolding/refolding events of these domains is minimal. These results establish a strong link between the global genomic arrangement of encoded biological pathways and transcriptional activation efficiency.

Key words: DOOR database, operon organization, gene regulation, prokaryotic genomes

INTRODUCTION

The number of fully sequenced prokaryotic genomes has been increasing rapidly in the past decades. In the NCBI Genome database, there has been a total of 6,917 complete genomes and 40,257 draft genomes as of February 9, 2017. In the Integrated Microbial Genome and Microbiome Samples of Joint Genome Institute (JGI IMG/M), there are 5,415 complete prokaryotic (bacterial and archaeal) genomes along with 43,348 draft genomes (https://img.jgi.doe.gov/cgi-bin/m/main.cgi) as of January 25, 2017 [1]. Based on the JGI IMG/M, there are over 175 million predicted genes, including ca. 171 million (97.7%) protein-coding genes and ca. four million (2.3%) RNA-coding genes. A major challenge in knowledge discovery from these genomes is gene functional annotation, only about 60% gene in a microbial genome are assigned with functions [2-4]. Current gene annotation is mainly based on homology search, using, e.g., BLAST [5] or Hmmer [6, 7] and relies on a repertoire of well-characterized genes, proteins, and probabilistic hidden Markov models (HMMs) as reference from a dozen model organisms. Given the limitation of these methods, further improvement will be needed to explore information beyond sequence level. Following radiative evolution from common ancestors, the genes the genes of similar functions are evolutionarily conserved in sequence homology (e.g., Clusters of Orthologous Groups (COGs) [8]) or statistical sequence models (e.g., HMMs in Pfam [9]), but also the genes of associated but different functions may be physically constrained in the genomes (e.g., gene linkage [10]). These functionally related genes may be organized into clusters in different positions of the genomes [11], which can be used in gene function inference/annotation.

The basic type of such gene clusters is *operon*, a static unit consisting of one or more consecutive genes in the genomes, which share one promoter upstream of the first gene and are usually transcribed into a single transcription unit (TU) (Figure 1A) [12]. In the review, to differentiate the terms operon and TU, TUs are only used to represent transcriptional units of mRNA while operons represent all non-TU organizations. Besides these basic transcriptional components, prokaryotic genomes also consist of regulatory elements such as transcription start sites, untranslated regions (UTRs), and terminators [13]. These functional information can be used for gene function annotation, as most functionally related genes tend to be placed in the same operons and share regulatory landscape (Figure 1B) [14]. Beside main promoters, there may be secondary promoters leading to alternative TUs, providing a dynamic structure of operon (Figure 1B) [13]. Given this dynamic expression of operons into TUs, TUs may overlap each other and share common genes, giving rise to overlapping TU clusters in bacteria (Figure 1C) [15]. Functionally, genes within a TU are more related than those within a cluster of TUs and likely to be more frequently co-expressed under certain conditions (Figure 1D). therefore, the organization of operon are not only evolutionarily conserved (Figure 1A and 1D) but also functionally constrained (Figure 1E). For example, the current arrangement of operons in most of the bacterial genomes tend to minimize the overall distance between consecutive operons of a same pathway across all pathways encoded in the genomes [14]. Such optimization also minimizes the total number of supercoil unfolding events in the folded chromosomes, needed to transcriptionally activate all the obligatory pathways between growth conditions (Figure 1E) [16]. In summary, the rich genome resources can be mined to gain biology on gene organization and genome functions.

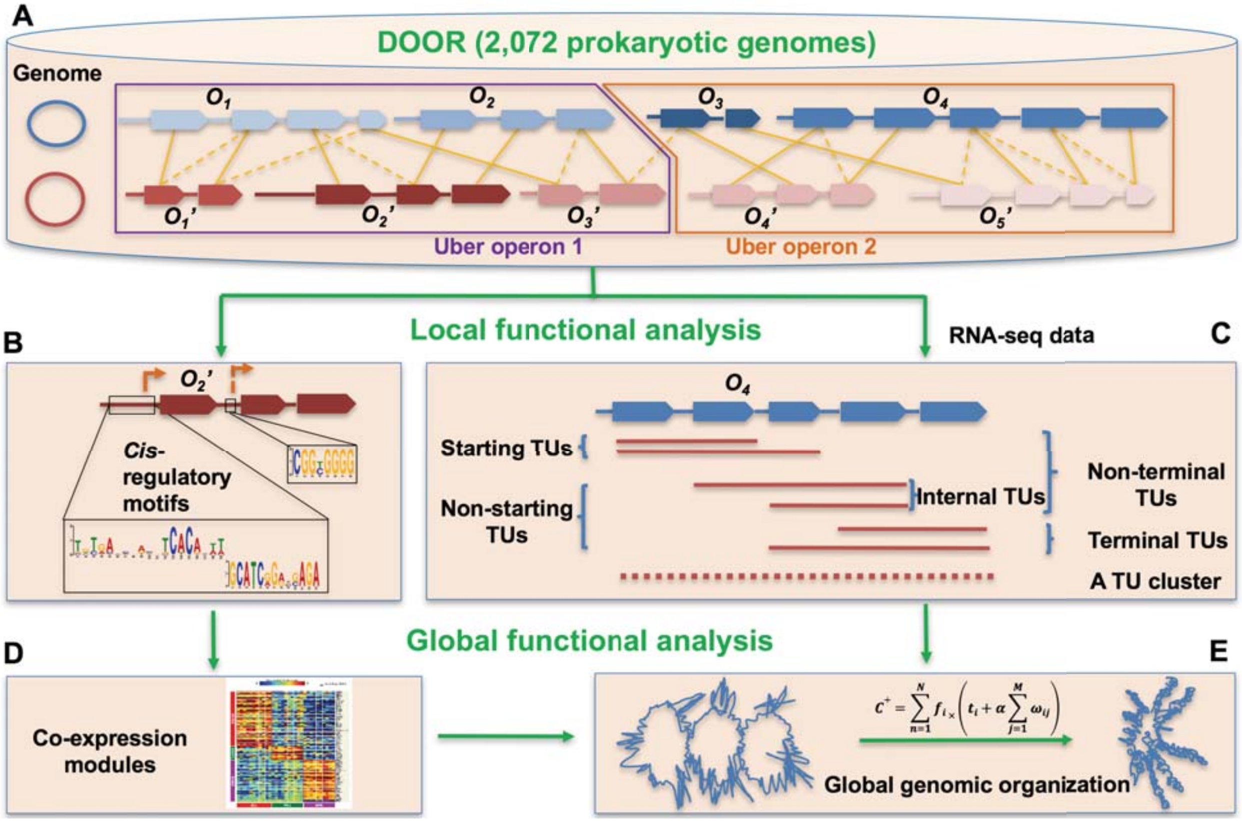

Figure 1. The prokaryotic operon database DOOR and derived analysis of genome organization and function. DOOR contains 2,072 prokaryotic genomes and the operons in which are associated through orthologous relationship and form uber operons (uber operons are explained in the text). For example, the two genomes in (A) each has four (*O1*-*O4*) and five (*O1'*-*O5'*) operons. Their orthologous associations are indicated by solid edges and paralogous associations are indicated by dotted lines. Uber operon 1 and 2 are only connected through one paralogous link (dotted line). We have developed a program (DIMINDA) that can predict *cis*-regulatory motifs (B). Combined with RNA-seq data, TUs and TU clusters can also be identified (C). These TU clusters suggest that genes in operons can be expressed in different combinations, leading to starting, internal, or terminal TUs. Genes and operons can be co-expressed under some but not all the tested conditions and can be identified by biclustering tools such as QUBIC we developed (D). These TUs and co-expressed modules results from optimized gene regulation. This regulatory mechanism we found to be that the global organization of operons in the prokaryotic chromosomes tends to minimize the events of unfolding and refolding chromosomes (E). C^+ is a function representing the partition of a genome into a set of folding domains (a.k.a., supercoils), which is minimized to optimize the folding and unfolding events and facilitate transcription efficiency.

Besides local variations in the composition of structural genes and dynamic expression of operons, operon organization are also constrained at the global level: the location of one operon also depends on the locations of other operons in the same biological pathways, and together they tend to form proximate clusters so as to maximize transcription efficiency [16]. Such functionally related organization also contributes to the overall organization of prokaryotic chromosome [17], which is also a key target of selection revealed by long-term evolution [18]. In this review, we introduce our database for prokaryotic operons—DOOR—and its computational functionalities of operon prediction and analysis, as an example of how to make knowledge discovery from rapidly increasing genome sequences. Based on these resources and derived insights on operon organization, advanced applications in revealing global operon organization and optimization are also addressed.

DEVELOPMENT OF THE DOOR DATABASE

Operon databases and DOOR

Given its central importance in many fields of microbiology, operon as a basic transcription unit has attracted substantial attention. For example, there are several operon databases built by different research groups: OperonDB [19], ProOpDB [20], ODB [21], rrnDB [22], etc. Our DOOR database not only stores curated operons in prokaryotic genomes, but also has a set of programs that performs a series of analyses on operon organization and expression. When first developed in 2006, DOOR has operon information for only 675 complete prokaryotic genomes predicted by our own algorithm, which was ranked as the best [23] in terms of its outperforming accuracy (greater than 90% in *Bacillus subtilis* and *Escherichia coli*) [24]. Besides curating predicted operons and providing general statistic summary for each genome and associated literature, DOOR allows users to search for operons by the genes (e.g., gene name *lacZ*) they contain, an operon ID (e.g., operon ID = 4015) or similar operons. The selected operons can be further used to predict *cis*-regulatory motif with embedded MEME [25] or CUBIC [26] programs upon request. Lastly, DOOR has an operonWiki (http://ecoliwiki.net/colipedia/index.php/Database_of_prOkaryotic_OpeRons) to facilitate interactions between users and the developers.

DOOR2: 2,072 genomes and RNA-seq datasets

As more complete genomes became available, we updated DOOR to DOOR2, which now has a total of 2,072 complete genomes (three times of what were initially included):

1,939 bacteria and 133 archaea, with 2,205 chromosomes and 1,645 plasmids [27]. For these genomes, a total of 1,323,902 multi-gene operons are predicted, averaging 583 such operons per chromosome and 24 operons per plasmid, along with 2,578,949 single-gene operons. Besides, 6,408 verified transcription factor-binding sites (TFBS) for 203 prokaryotic genomes, 3,456,718 Rho-independent terminators for 2,072 genomes, and 6,975,454 conserved operons are also predicted. Given the availability of RNA-seq data for some complete genomes, TUs are also predicted which can contain part of the genes in an operon or span at least two operons [13, 28]. All these operons, TUs, and regulatory elements are stored in the relational MySQL DOOR2 database, which can be queried with multiple terms such as species name, operon ID, or gene name, etc. A genome browser is also created to support visualization of selected operon data along with dynamic TU structures under multiple conditions if the RNA-seq datasets are available [27].

DOOR2 is implemented as a web portal server with a multi-layer architecture. Technically, our online operon prediction requires three input files, specifying gene locations in a genome (e.g., .gff or .gtf files), protein sequences (.faa files), and nucleotide sequences (.fna files) of an entire genome, respectively. The representation and the logic layers are implemented using the Web 2.0 technology (HTML5, CSS3 and Javascript language along with jQuery library) and PHP server-side scripting language. The keyword-based search engine is implemented based on the Sphinx Open Source Search Server (http://sphinxsearch.com), and the genome browser is implemented based on JBrowse Genome Browser (http://jbrowse.org) [29]. To this point, DOOR2 has become a real integrated database and tool for identification of one-stop operon prediction and analysis webserver (Figure 1).

KEY OPERON-RELATED COMPUTATIONAL FUNCTIONALITIES

To make discoveries from genomic and transcriptomic data, appropriate algorithms and tools need to be designed and developed. Here we summarize the in-house available programs in DOOR, which can be potentially used to facilitate related objectives through data mining (Figure 1B-1E).

Computational prediction of operons

This operon prediction program was developed through training genomic features in a classification model and achieves an average accuracy of 90% and viewed as the most accurate for operon prediction [23] (Figure 1A). A total of five features are included for prediction: intergenic distance, conserved gene neighborhood, phylogenetic distances between adjacent genes, length ratio of two adjacent genes, and frequencies of specific DNA motifs in the intergenic regions [30]. Our analyses showed that the length of the intergenic region between a pair of adjacent genes is the most reliable indicator of whether it is an operon pair or a boundary pair. For genomes with a substantial number of operons, our (non-linear) decision tree-based classifier can predict operons in a prokaryotic genome with a high accuracy level. For example, the prediction accuracy of our program can reach 90.2 and 93.7% on *Bacillus subtilis* and *Escherichia coli* genomes, respectively. Without known operon information, our (linear) logistic function-based classifier can reach the prediction accuracy at 84.6 and 83.3% for *E. coli* and *B. subtilis*, respectively.

TU prediction based on high-throughput RNA-seq data

TUs are the basic transcriptional units. Therefore, identifying TUs under different conditions offers insight into the gene regulatory strategies in prokaryotes. For that, we have developed a program called SeqTU (Figure 1B), when RNA-seq data is provided [31]. Compared to the static structure of operons, TUs are more dynamic in terms of their composition, with constituent genes being expressed condition-specifically [31, 32]. SeqTU predicts TUs based on two features in a machine learning model measuring the RNA-seq expression patterns across the genome: expression-level continuity and variance, which has been developed into a web server [33]. In *Clostridium thermocellum*, 2590 TUs are predicted based on four RNA-seq datasets using SeqTU; 44% of the TUs have multiple genes. The high precision of SeqTU is also validated with RNA-seq data in *E. coli*.

Cis-regulatory motif analysis and prediction

Elucidating gene regulation in response to environmental stimuli and cellular changes is one of the fundamental goals in biology. Identification of *cis*-regulatory motifs in genomic sequences for operons and TUs is a key step in computational genomics toward complete understanding of the global regulatory network in microbes. With the availability of operon/TU, we can then identify the regulatory motifs, particularly *cis*-regulatory motifs that bind with transcription factors (Figure 1B). The first program we developed in this regard is called BOBRO, which substantially improves the prediction accuracy compared to existing programs [34]. It identifies significant motifs in promoters in two steps: it first assesses the possibility for each position in a given promoter to be the (approximate) start of a conserved sequence motif using a highly effective method; and identifies actual motifs from the accidental ones based on the concept of 'motif closure' [34]. These two key ideas are embedded in a classical framework for motif finding by identifying cliques in a graph but have made this framework substantially more sensitive and more selective in a very noisy background. In an updated version, BoBro2.0, prediction and analysis of *cis*-regulatory motifs are integrated. Besides reliably identification of *cis*-regulatory motifs at a genome scale, it can accurately scan for all motif instances of a query motif in specified genomic regions and provide reliable comparisons and clustering of identified motifs, which takes into consideration the weak signals from the flanking regions of the motifs. Additionally, it can analyze co-occurring motifs in the regulatory regions, in support of elucidation of co-regulation by multiple transcription factors [35]. Finally, we implement all these tools and algorithms in a web server called DMINDA for identification and analysis of regulatory DNA motifs of operons/TUs [36]. Recently, we have developed a phylogenetic footprinting method for *cis*-regulatory motifs identification in prokaryotic genomes, which produces improved results over BOBRO and similar programs [37]. Based on these identified *cis*-regulatory motifs, bacterial regulons can be predicted on a genome-scale [38], which is important in knowledge discovery from genomes.

Co-expression analysis of operons

The genes in the operons and TUs identified above can be co-expressed (with similar expression levels) under different conditions. Identifying the expression patterns provides important information on the functional responses in cells and helps identify higher level functional machineries, e.g., metabolic and regulatory pathways. The first generation of co-expression analysis usually focuses on grouping genes under all given conditions into clusters of similar expression levels or similar changes [39]. With the quick popularity of microarray and RNA-seq, it soon became desirable to identify genes, operons, and TUs that are only co-expressed under some (to-be-identified) conditions, not all conditions. From these 'conditionally' expressed gene biclusters (two-dimensional representation of gene clusters)—clusters of co-expressed genes (rows) under multiple conditions (columns), we can infer local and global gene regulation in cells and how cellular systems respond to different environmental conditions.

We have developed a QUalitative BIClustering algorithm (QUBIC) for the clustering of co-expressed genes across many test conditions (Figure 1D) [40]. Our algorithm first converts gene expression data into a qualitative matrix, and then identifies all biclusters in this matrix one-by-one in a heuristic way, starting with the closest gene pairs as a seed to build an initial bicluster and then iteratively recruiting additional genes into the current bicluster without violating a pre-specified consistency level. Employing a combination of qualitative (or semi-quantitative) measures of gene expression data and a combinatorial optimization technique, the QUBIC algorithm can identify all statistically significant biclusters including biclusters with the so-called 'scaling patterns', a problem considered to be rather challenging and biologically meaningful. Another key feature is that QUBIC solves general biclustering problems very efficiently, capable of solving biclustering problems with tens of thousands of genes under up to thousands of conditions in a few minutes of the CPU time on a desktop computer. This algorithm outperforms other programs in this regard, such as SAMBA [41], ISA [42], BIMAX [43], RMSBE [44] has subsequently been developed into a web server [45], and recently a Bioconductor package [46].

Comparative genomics analysis in support of conserved operons identification among different genomes

Mapping biological pathways across microbial genomes is critical to functional studies of biological systems. Most existing methods mainly rely on sequence-based orthologous gene mapping, which often leads to biased results because sequence-similarity information alone does not contain sufficient information for accurate identification of orthologous relationship. We developed an algorithm for pathway mapping across microbial genomes, combining both sequence similarity and genomic structure information such as operons and regulons (Figure 1A) [47]. Our algorithm is based on the observation that the products of genes in an operon or TU usually perform closely related functions and these genes are more likely to be in the same pathway than those not belonging in the same operons, TUs, or regulons [48-50]. Such a program, P-MAP, solved this constrained optimization problem using the integer-programming algorithm [47]. Our analysis on a number of known homologous pathways shows that using genomic structure information as constraints can greatly improve the pathway-mapping accuracy over methods that use sequence-similarity information alone [47].

Another direct application of the DOOR database and is to increase the accuracy of orthologous gene prediction, which is the most basic function of comparative genomics. Most of the programs developed for this function are based on either phylogeny [51-53] or sequence similarity [49, 54]. In our method GOST (Global Optimization STrategy) for orthologous gene mapping, we utilize groups of functionally or transcriptionally related operons (uber-operons) (Figure 1B), whose gene sets are conserved across the target and multiple reference genomes [55]. Then GOST identifies all the orthologous gene pairs across two genomes with a good ‘enough’ sequence similarity score and the insight that the two genes have homologous working partners in their respective genomes (two genes are defined as homologous if their sequence similarity is below a specific E-value threshold in BLAST). Two genes in a genome are considered as ‘working partners’ if they belong to a common operon/uber-operon [55]. GOST identified 665 more enzyme gene pairs than RBH, 1,901 more than INPARANOID [56] and 2,354 more than OrthoMCL [57]. We believe that the actual performance of GOST is even better than suggested by these comparisons, as the genes tested only represent a small portion of all the orthologous relationships among enzyme-encoding genes across bacterial genomes. Overall, GOST is much more efficient than OrthoMCL and INPARANOID and runs as fast as RBH [49, 55].

ADVANCED KNOWLEDGE DISCOVERY USING DOOR

Evolutionary understanding of operons: Uber operon and operon structures across multiple closely related organisms

Operons are gene sets placed in contiguous blocks on genomes which are co-regulated and co-transcribed to execute related functions. We already show that this structure is not static, but rather dynamic with respect to gene composition particularly during gene expression across various external growth conditions. From an evolutionary perspective, the gene content, gene order, and operon/TU regulation may vary among genomes but such rearrangements tend to conserve individual genes in specific functional and regulatory contexts across genomes, such contexts are called **uber-operons** (Figure 1A) [58]. We have developed an algorithm to predict uber-operons, which can reveal the patterns of operon evolution in prokaryotes (Figure 1A). Our algorithm sets two groups of genomes (target and references) for comparison and then identifies groups of functionally or transcriptionally related operons, whose gene sets are conserved across the target and multiple reference genomes. Using this algorithm, we predicted uber-operons for each of available 91 genomes, using other 90 genomes as references. Totally 158 uber-operons containing 1,830 genes were obtained in *E. coli* K12, and it is found that many of the uber-operons are parts of known regulons or biological pathways or involved in highly related biological processes based on their Gene Ontology (GO) [59] assignments. For some of the predicted uber-operons that are not part of known regulons or pathways, our analyses indicate that their genes are highly likely to work together in the same GO biological processes, suggesting the possibility of new regulons or pathways [60]. Besides the uber operons between distantly related species, we also examined the diversity of operons with high resolution between closely related genomes with a graph-based model, we found that genes in a connected component (a maximal set of genes linked together as a subgraph in the entire graph) are likely to be functionally related and these identified components tend to form treelike topology, such as paths (each path is a chain of connected unique nodes) and stars (each star is a tree with one internal node and many leaves), corresponding to different biological mechanisms in transcriptional regulation [61]. For example, a path-structure component integrates genes encoding a protein complex, such as ribosome; a star-structure component not only groups related genes together, but also reflects the key functional roles of the central node of this component, such as the ABC transporter with a transporter permease and substrate-binding proteins surrounding it. Most interestingly, the genes from organisms with highly diverse living environments, i.e., biomass degraders and animal pathogens of the Clostridium genus, can be clearly classified into different topological groups on some connected components [61].

The global organizing principle of operons in bacterial genomes

Dynamic organization of the bacteria chromosome under different conditions.

Besides the local evolutionary conservation of operons, little is understood about what may determine the global arrangement of bacterial genes in a genome beyond the operon level. It is postulated that the global genomic organization of bacterial genes may be affected and constrained by multiple cellular processes, particularly gene transcription, genome replication, and nucleoid compaction, at both local and global levels [62]. For example, we found at least 40% of the operons in the genomes *E. coli* K-12 and *Bacillus subtilis* strain participate in multiple metabolic pathways [14]. Furthermore, we found that the global arrangement of operons in a bacterial genome is largely influenced by the fact that bacteria tend to keep their operons of the *same* biological pathways in vicinity on genomes and also keep operons in *multiple* pathways close to other fellow operons of these pathways (Figure 1E) [14].

The circular chromosome of bacteria has been suggested to fold into a collection of sequentially consecutive domains (supercoils), which are dynamically positioned with high precision [63-65]. A domain needs to be unfolded when a biological pathway which contains genes encoded in this DNA segment is transcriptionally activated. We postulated that bacterial genomes are partitioned into a set of folding domains such that the total unfolding/refolding events of these domains is minimal. By testing this hypothesis, we predicted seven distinct sets of such domains along the *E. coli* chromosome under seven growth conditions, namely exponential growth, stationary growth, anaerobiosis, heat shock, oxidative stress, nitrogen limitation, and SOS stress responses [16]. These predicted folding domains are highly stable statistically and are generally consistent with the experimental data of DNA binding sites of the nucleoid-associated proteins that assist the folding of these domains, as well as genome-scale protein occupancy profiles. These results establish a strong link between a folded *E. coli* chromosomal structure and the encoded biological pathways and their activation frequencies [16].

Global genomic arrangement of bacterial operons is closely tied with the total transcriptional efficiency

The above results suggest that microbial genomes are globally folded such that the total number of supercoil unfolding events in the folded chromosome, needed to transcriptionally activate all the obligatory pathways under a specific growth condition, tends to be minimized [16] (Figure 1E). We reasoned this is a result of adaptive evolution, such that the operons are globally arranged in such a way that the total energy is minimized for unfolding (and then refolding) relevant DNA segments needed to make the required genes transcriptionally accessible in response to various stimuli. To this end, we developed a simple model for estimating this total energy cost and found that partitions of the whole genome into 10-100 kb genomic regions offers can minimize such energy cost [16, 17]. Through applying the above quantitative model on 52 *E. coli* genomes, we further investigated the potential underlying principles that dictate the global organization of operons into supercoils in the chromosomes. We found the commonalities and differences in the genomic organizations of genes (and operons) encoding specific pathways across different genomes is largely dictated by the frequencies of the transcription activation of pathways relative to those of the other encoded pathways in an organism and the variation in the activation frequencies of a specific pathway across the related genomes [66].

SUMMARY AND OUTLOOK

Genomes are the complete set of genes or genetic materials present in a cell or organism and contain heredity information which needs to be deciphered to make sense out of them and life. Increasingly available prokaryotic genomes have provided a rich source of ‘raw’ materials for knowledge discovery on the organization and function of genes, operons, and chromosomes, and the evolution of these functional constituents. We have curated an operon database in 2,072 genomes and developed multiple tools to understand the organization of gene, operon, and chromosome. We have demonstrated operon as a local gene-organizing strategy has profound implications in gene function, regulation, and genome organization. On a system level, our findings so far support that genome organization and transcription efficiencies are closely correlated. With more genomes sequenced and more powerful tools, advanced systems knowledge will be gained for bacterial life. This will the main goal of our next version of DOOR3, which is under construction.

FUNDING

This work was supported by a grant from the U.S. Department of Energy (#DE-PS02-06ER64304). The BioEnergy Science Center (BESC) is supported by the Office of Biological and Environmental Research in the DOE Office of Science. This work was also supported in part by the State of South Dakota Research Innovation Center and the Agriculture Experiment Station of South Dakota State University.

KEY POINTS

1. DOOR is a comprehensive database that curates operons in 2,072 prokaryotic genomes that have becoming increasingly available and predicts regulatory motifs of the operons toward biology knowledge discovery. A total of 1,323,902 multi-gene

operons is predicted, averaging 583 such operons per chromosome and 24 operons per plasmid, along with 2,578,949 single-gene operons.

2. Operons are not static, but rather dynamic under different conditions and usually form clusters of transcription units. More importantly, operons of related functions are evolutionary conserved, forming uber operons.
3. The global arrangement of operons in a bacterial genome is largely influenced by the fact that bacteria tend to keep their operons of related functions in vicinity and form supercoils; bacterial genomes are such partitioned to minimize the total unfolding/refolding events of these domains.
4. Our tools developed around operons can predict and analyze transcription units, regulatory motifs, operon evolution, transcription efficiency, and genome organization, which can establish link between organization and function at operon and genome levels.

REFERENCES

1. Chen IMA, Markowitz VM, Chu K et al. IMG/M: integrated genome and metagenome comparative data analysis system, Nucleic Acids Research 2016;45:D507-D516.
2. Brent MR. Genome annotation past, present, and future: How to define an ORF at each locus, Genome Research 2005;15:1777-1786.
3. Golyshev MA, Korotkov EV. Developing of the Computer Method for Annotation of Bacterial Genes, Advances in Bioinformatics 2015;2015:9.
4. Baric RS, Crosson S, Damania B et al. Next-Generation High-Throughput Functional Annotation of Microbial Genomes, mBio 2016;7.
5. Camacho C, Coulouris G, Avagyan V et al. BLAST+: architecture and applications, BMC Bioinformatics 2009;10:421.
6. Overbeek R, Disz T, Stevens R. The SEED: a peer-to-peer environment for genome annotation, Commun. ACM 2004;47:46-51.
7. Eddy SR. Accelerated profile HMM searches, PLoS Comput Biol 2011;7:e1002195.
8. Tatusov RL, Galperin MY, Natale DA et al. The COG database: a tool for genome-scale analysis of protein functions and evolution, Nucleic Acids Research 2000;28:33-36.
9. Finn RD, Coggill P, Eberhardt RY et al. The Pfam protein families database: towards a more sustainable future, Nucleic Acids Research 2016;44:D279-D285.
10. Rudd KE. Linkage Map of Escherichia coli K-12, Edition 10: The Physical Map, Microbiology and Molecular Biology Reviews 1998;62:985-1019.
11. Wu H, Mao F, Olman V et al. Hierarchical classification of functionally equivalent genes in prokaryotes, Nucleic Acids Research 2007;35:2125-2140.
12. Jacob F, Perrin D, Sánchez C et al. The operon: a group of genes with expression coordinated by an operator, C.R.Acad. Sci. Paris 1960;250:1727-1729.

13. Cho B-K, Zengler K, Qiu Y et al. The transcription unit architecture of the Escherichia coli genome, Nat Biotech 2009;27:1043-1049.

14. Yin Y, Zhang H, Olman V et al. Genomic arrangement of bacterial operons is constrained by biological pathways encoded in the genome, Proceedings of the National Academy of Sciences 2010;107:6310-6315.

15. Mao X, Ma Q, Liu B et al. Revisiting operons: an analysis of the landscape of transcriptional units in E. coli, BMC Bioinformatics 2015;16:356.

16. Ma Q, Yin Y, Schell MA et al. Computational analyses of transcriptomic data reveal the dynamic organization of the Escherichia coli chromosome under different conditions, Nucleic Acids Research 2013;41:5594-5603.

17. Ma Q, Xu Y. Global Genomic Arrangement of Bacterial Genes Is Closely Tied with the Total Transcriptional Efficiency, Genomics, Proteomics & Bioinformatics 2013;11:66-71.

18. Crozat E, Philippe N, Lenski RE et al. Long term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection, Genetics 2005;169:523-532.

19. Pertea M, Ayanbule K, Smedinghoff M et al. OperonDB: a comprehensive database of predicted operons in microbial genomes, Nucleic Acids Research 2009;37:D479-D482.

20. Taboada B, Ciria R, Martinez-Guerrero CE et al. ProOpDB: Prokaryotic Operon DataBase, Nucleic Acids Research 2012;40:D627-D631.

21. Okuda S, Katayama T, Kawashima S et al. ODB: a database of operons accumulating known operons across multiple genomes, Nucleic Acids Research 2006;34:D358-D362.

22. Stoddard SF, Smith BJ, Hein R et al. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development, Nucleic Acids Research 2015;43:D593-D598.

23. Brouwer RWW, Kuipers OP, van Hijum SAFT. The relative value of operon predictions, Briefings in Bioinformatics 2008;9:367-375.

24. Mao F, Dam P, Chou J et al. DOOR: a database for prokaryotic operons, Nucleic Acids Research 2009;37:D459-D463.

25. Bailey TL, Williams N, Misleh C et al. MEME: discovering and analyzing DNA and protein sequence motifs, Nucleic Acids Research 2006;34:W369-W373.

26. Olman V, Xu D, Xu Y. CUBIC: identification of regulatory binding sites through data clustering, Journal of Bioinformatics and Computational Biology 2003;01:21-40.

27. Mao X, Ma Q, Zhou C et al. DOOR 2.0: presenting operons and their functions through dynamic and integrated views, Nucleic Acids Research 2014;42:D654-D659.

28. Adhya S. Suboperonic Regulatory Signals, Science's STKE 2003;2003:pe22-pe22.

29. Skinner ME, Uzilov AV, Stein LD et al. JBrowse: A next-generation genome browser, Genome Research 2009;19:1630-1638.

30. Dam P, Olman V, Harris K et al. Operon prediction using both genome-specific and general genomic information, Nucleic Acids Research 2007;35:288-298.

31. Chou W-C, Ma Q, Yang S et al. Analysis of strand-specific RNA-seq data using machine learning reveals the structures of transcription units in Clostridium thermocellum, Nucleic Acids Research 2015;43:e67.

32. Quail MA, Haydon DJ, Guest JR. The pdhR–aceEF–lpd operon of Escherichia coli expresses the pyruvate dehydrogenase complex, Molecular Microbiology 1994;12:95-104.

33. Chen X, Chou W-C, Ma Q et al. SeqTU: A Web Server for Identification of Bacterial Transcription Units, Scientific Reports 2017;7:43925.

34. Li G, Liu B, Ma Q et al. A new framework for identifying cis-regulatory motifs in prokaryotes, Nucleic Acids Research 2011;39:e42.

35. Ma Q, Liu B, Zhou C et al. An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale, Bioinformatics 2013;29:2261-2268.

36. Ma Q, Zhang H, Mao X et al. DMINDA: an integrated web server for DNA motif identification and analyses, Nucleic Acids Research 2014;42:W12-W19.

37. Liu B, Zhang H, Zhou C et al. An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes, BMC Genomics 2016;17:578.

38. Liu B, Zhou C, Li G et al. Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses, Scientific Reports 2016;6:23030.

39. Eisen MB, Spellman PT, Brown PO et al. Cluster analysis and display of genome-wide expression patterns, Proceedings of the National Academy of Sciences 1998;95:14863-14868.

40. Li G, Ma Q, Tang H et al. QUBIC: a qualitative biclustering algorithm for analyses of gene expression data, Nucleic Acids Research 2009;37:e101.

41. Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data, Bioinformatics 2002;18:S136-S144.

42. Ihmels J, Bergmann S, Barkai N. Defining transcription modules using large-scale gene expression data, Bioinformatics 2004;20:1993-2003.

43. Prelić A, Bleuler S, Zimmermann P et al. A systematic comparison and evaluation of biclustering methods for gene expression data, Bioinformatics 2006;22:1122-1129.

44. Liu X, Wang L. Computing the maximum similarity bi-clusters of gene expression data, Bioinformatics 2007;23:50-56.

45. Zhou F, Ma Q, Li G et al. QServer: A Biclustering Server for Prediction and Assessment of Co-Expressed Gene Clusters, PLoS ONE 2012;7:e32660.

46. Zhang Y, Xie J, Yang J et al. QUBIC: a Bioconductor package for qualitative biclustering analysis of gene co-expression data, Bioinformatics 2016.

47. Mao F, Su Z, Olman V et al. Mapping of orthologous genes in the context of biological pathways: An application of integer programming, Proceedings of the National Academy of Sciences of the United States of America 2006;103:129-134.

48. Dandekar T, Snel B, Huynen M et al. Conservation of gene order: a fingerprint of proteins that physically interact, Trends in Biochemical Sciences 1998;23:324-328.

49. Overbeek R, Fonstein M, D'Souza M et al. The use of gene clusters to infer functional coupling, Proceedings of the National Academy of Sciences 1999;96:2896-2901.

50. Chen Y-M, Zhu Y, Lin ECC. The organization of the fuc regulon specifying l-fucose dissimilation in Escherichia coli K12 as determined by gene cloning, Molecular and General Genetics MGG 1987;210:331-337.

51. Zmasek CM, Eddy SR. RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs, BMC Bioinformatics 2002;3:14.

52. Storm CEV, Sonnhammer ELL. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability, Bioinformatics 2002;18:92-99.

53. Kim KM, Sung S, Caetano-Anollés G et al. An approach of orthology detection from homologous sequences under minimum evolution, Nucleic Acids Research 2008;36:e110-e110.

54. Tatusov RL, Koonin EV, Lipman DJ. A Genomic Perspective on Protein Families, Science 1997;278:631-637.

55. Li G, Ma Q, Mao X et al. Integration of sequence-similarity and functional association information can overcome intrinsic problems in orthology mapping across bacterial genomes, Nucleic Acids Research 2011;39:e150.
56. Remm M, Storm CEV, Sonnhammer ELL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons1, Journal of Molecular Biology 2001;314:1041-1052.
57. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes, Genome Research 2003;13:2178-2189.
58. Lathe Iii WC, Snel B, Bork P. Gene context conservation of a higher order than operons, Trends in Biochemical Sciences 2000;25:474-479.
59. Ashburner M, Ball CA, Blake JA et al. Gene Ontology: tool for the unification of biology, Nat Genet 2000;25:25-29.
60. Che D, Li G, Mao F et al. Detecting uber-operons in prokaryotic genomes, Nucleic Acids Research 2006;34:2418-2427.
61. Zhou C, Ma Q, Li G. Elucidation of Operon Structures across Closely Related Bacterial Genomes, PLoS ONE 2014;9:e100999.
62. Rocha EPC. The Organization of the Bacterial Genome, Annual Review of Genetics 2008;42:211-233.
63. Dillon SC, Dorman CJ. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression, Nat Rev Micro 2010;8:185-195.
64. Browning DF, Grainger DC, Busby SJW. Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression, Current Opinion in Microbiology 2010;13:773-780.
65. Vincenzo GB, Bruno B, Kevin DD et al. Physical descriptions of the bacterial nucleoid at large scales, and their biological implications, Reports on Progress in Physics 2012;75:076602.
66. Ma Q, Chen X, Liu C et al. Understanding the commonalities and differences in genomic organizations across closely related bacteria from an energy perspective, Science China Life Sciences 2014;57:1121-1130.