Computing and Applying Atomic Regulons to Understand Gene Expression and Regulation

José P. Faria[1, 2, 3, 5*], James J. Davis[1, 2], Janaka N Edirisinghe[1,2], Ronald C. Taylor[4], Pamela B. Weisenhorn[5], Robert D. Olson[1,2], Rick L. Stevens[1,2,6], Miguel Rocha[3], Isabel Rocha[3], Aaron A. Best[7], Matthew DeJongh[8], Nathan L. Tintle[9], Bruce Parrello[2,10], Ross Overbeek[1,2,10], Christopher S. Henry[1,5*]

[1]Computation Institute, University of Chicago, Chicago, Illinois, USA
[2]Computing, Environment and Life Sciences, Argonne National Laboratory, Argonne IL, USA
[3]Centre of Biological Engineering, University of Minho, Campus de Gualtar, Braga, Portugal
[4]Computational Biology & Bioinformatics Group, Pacific Northwest National Laboratory (U.S. Dept. of Energy), Richland, Washington, USA
[5]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA
[6]Department of Computer Science, Ryerson Physical Laboratory, University of Chicago, Chicago, IL 60637, USA
[7]Biology Department, Hope College, Holland, Michigan, USA.
[8]Computer Science Department, Hope College, Holland, Michigan, USA
[9]Department of Mathematics, Statistics and Computer Science, Dordt College, Sioux Center, IA 51, USA
[10]Fellowship for Interpretation of Genomes, Burr Ridge, IL, USA

*** Correspondence:** The University of Chicago, Computation Institute, 5735 South Ellis Avenue, Chicago, IL, 60637, USA.
jplfaria@anl.gov
chenry@mcs.anl.gov

Keywords: Atomic Regulon, clustering, gene expression analysis, transcriptomic data, *Escherichia coli*, hierarchical clustering, CLR, k-means clustering

Abstract

Understanding gene function and regulation is essential for the interpretation, prediction and ultimate design of cell responses to changes in the environment. An important step toward meeting the challenge of understanding gene function and regulation is the identification of sets of genes that are always co-expressed. These gene sets, *Atomic Regulons* (ARs), represent fundamental units of function within a cell and could be used to associate genes of unknown function with cellular processes and to enable rational genetic engineering of cellular systems. Here, we describe an approach for inferring ARs that leverages large-scale expression data sets, gene context, and functional relationships among genes. We computed ARs for *Escherichia coli* based on 907 gene expression experiments and compared our results with gene clusters produced by two prevalent data-driven methods: hierarchical clustering and k-means clustering. We compared ARs and purely data-driven gene clusters to the curated set of regulatory interactions for *E. coli* found in RegulonDB, showing that ARs are more consistent with gold standard regulons than are data-driven gene clusters. We further examined the consistency of ARs and data-driven gene clusters in the context of gene interactions predicted by Context Likelihood of Relatedness (CLR) analysis, finding that the ARs show better agreement with CLR predicted interactions. We determined the impact of increasing amounts of expression data on AR construction and find that while more data improve

ARs, it is not necessary to use the full set of gene expression experiments available for *E. coli* to produce high quality ARs. In order to explore the conservation of co-regulated gene sets across different organisms, we computed ARs for *Shewanella oneidensis*, *Pseudomonas aeruginosa*, *Thermus thermophilus*, and *Staphylococcus aureus*, each of which represents increasing degrees of phylogenetic distance from *E. coli*. Comparison of the organism-specific ARs showed that the consistency of AR gene membership correlates with phylogenetic distance, but there is clear variability in the regulatory networks of closely related organisms. As large scale expression data sets become increasingly common for model and non-model organisms, comparative analyses of atomic regulons will provide valuable insights into fundamental regulatory modules used across the bacterial domain.

Introduction

The inference of gene function and regulation represent intertwined challenges in Systems Biology. (Kitano, 2002). Regulon content often leads to the inference of functions for genes contained in the regulon, and assigned functions are often used to refine regulon structure. Microarray technologies (Young, 2000) accelerated the study of gene regulation by facilitating the production of thousands of expression datasets (Edgar et al., 2002), and next generation sequencing technologies massively increased the availability of reference genomes while also enabling the estimation of relative gene expression (Wang et al., 2009). Despite these advances, we still lack a complete understanding of gene function and regulation even in the most well studied bacterium, *Escherichia coli* (Kochanowski et al., 2013), which is the subject of thousands of phenotype experiments, gene expression datasets, and multiple regulation databases (Huerta et al., 1998;Salgado et al., 2013;Karp et al., 2014).

A key step in the inference of gene regulatory networks, and a valuable step in the functional annotation of genes, is the decomposition of a genome into sets of co-expressed genes. Today, three general methods exist for identifying sets of co-expressed genes: (i) clustering methods; (ii) transcription factor binding-site (TFBS) analysis (Rodionov, 2007); and (iii) *de novo* reverse engineering from expression data (De Smet and Marchal, 2010). Classic clustering methods, such as hierarchical clustering (Murtagh, 1985) and the centroid k-means clustering (Lloyd, 1982), aim to group sets of objects based on some criteria; when applied to the analysis of gene expression data, the aim is to group genes with similar expression profiles. TFBS tools, such as the popular RegPredict (Novichkov et al., 2010), infer regulons based on the presence of conserved upstream regions of DNA, which are presumed to be *cis*-regulatory elements. *De novo* reverse engineering methods use expression data to infer gene-to-gene regulatory interactions. One of the most commonly used methods, Context Likelihood of Relatedness (CLR), has been successfully applied to infer novel regulatory interactions (Faith et al., 2007).

Algorithms for computing co-expressed gene sets produce two different types of output: regulons comprising a transcription factor and an associated set of regulated genes, or a set of co-expressed genes. The first type of output is produced by TFBS binding-site analysis and *de novo* reverse engineering methods to produce gene sets consistent with the classical definition of a regulon – genes are merged together into a set only if they respond to a common transcription factor. It is possible for a gene to appear in multiple sets if it responds to multiple transcription factors. This type of regulon information is valuable as a building block for assembling transcriptional regulatory networks and can be used, for instance, in deriving constraints to represent regulation in metabolic

This is a provisional file, not the final typeset article

models (Shlomi et al., 2007;Chandrasekaran and Price, 2010). However, the overlap in their gene content and the resulting complexity in their interpretation make them less ideal for other applications of co-expressed gene sets.

Purely data-driven algorithms, such as hierarchical clustering or k-means clustering, can be used to produce the second type of output, sets of co-expressed genes that are not necessarily associated with a transcription factor. We propose to call the sets of co-expressed genes that are always ON or OFF together, *Atomic Regulons* (ARs). We define an AR as a set of genes that have essentially identical expression patterns, indicating a strong likelihood that they are functionally related (i.e., the genes are *always* expressed as a set). Each gene can be a member of only one AR; some ARs are represented by a single gene. Thus, a genome can be thought of as being comprised of ARs, with ARs considered to be the fundamental functional units of the cell. As the cell transitions from one functional state to another, it will activate some ARs, and deactivate others, with the functional states being defined by the set of active ARs. Cell states can be thought of as being organized hierarchically, with the ARs that represent core functions being constitutively expressed and the ARs that represent peripheral functions being expressed under specific conditions. In this way, analyzing expression patterns of ARs provide insights about gene functions and relationships among cellular systems.

The concept of atomic regulons has many useful applications. ARs are commonly used to provide insights into functions of orphan genes using the guilt-by-association principle, most prominently in resources such as STRING (von Mering et al., 2005). ARs are also used to plug gaps in metabolic reconstructions and models (Benedict et al., 2014). In addition, we recently applied ARs in the curation of regulatory network models to map regulons to stimuli (Faria et al., 2016). ARs make a statistical inference based on input data and are uniquely suited to these applications because: (i) they do not overlap in their gene content; (ii) all of their genes are co-expressed; (iii) they do not exclude genes that are co-expressed; and (iv) they may be computed reliably from tractable amounts of data.

Here we describe a new algorithm for computing ARs, which combines many of the advantages of the existing data-driven approaches, but integrates new evidence types including chromosome location and functional relationships to more quickly converge on a complete set of biologically meaningful ARs. We apply the new atomic regulon inference method to *Escherichia coli*, which has large amounts of expression data that represent many environmental conditions. We compare the new *E. coli* ARs with those produced by existing data-driven approaches, curated sets of regulons mined from literature (RegulonDB), and with CLR-derived gene clusters. We then compute ARs for a set of four bacteria with increasing phylogenetic distance from *E. coli* to begin to understand the nature of conserved and unique ARs across organisms.

Materials and Methods

Algorithm for computing atomic regulons

Our atomic regulon inference algorithm is unique from other approaches in that it begins by constructing draft atomic regulons (gene sets) using a combination of operon predictions and SEED subsystem technology (Overbeek et al., 2005). A subsystem is a set of related functional roles that represents the group of proteins involved in a biological process or pathway (e.g., protein

This is a provisional file, not the final typeset article

biosynthesis or TCA cycle). In contrast to purely data-driven approaches, which start by forming gene clusters based only on available expression data, our approach applies expression data after initial gene clusters are formed in order to either extend or divide gene clusters to ensure they contain all co-expressed genes.

Thus, the computation of ARs is derived from gene context information, functional annotation, and estimates of gene ON and OFF states from the expression data. The estimation of gene ON and OFF states is pre-computed as a separate step (see "Estimation of gene ON/OFF states" in Materials and Methods). Our AR inference process consists of six steps (**Figure 1**).

Step 1. Generate Initial Atomic Regulon Gene Sets

Compute a set of hypotheses in the form:

Genes G1 and G2 should be in the same atomic regulon

Initial clusters are proposed using two independent mechanisms: (i) gene clustering within putative operons; and (ii) membership of genes within SEED subsystems.

Step 2. Process Gene Expression Data and Calculate Pairwise Expression Profile Similarities

Integrate all available gene expression data for the genome, load the normalized data, and compute Pearson correlation coefficients (PCCs) for all possible gene pairs. PCCs are computed to provide a quantitative assessment of how similar the expression profiles are for each gene pair.

Step 3. Expression Informed Splitting of Initial Atomic Regulon Gene Sets

Split operon and subsystem-based clusters using the criterion that genes in a set must have pairwise expression data profiles greater than a PCC of 0.7. This ensures that the initial clusters contain genes that all share a substantial level of co-expression.

** Figure 1**

Step 4. Restrict Gene Membership to One Atomic Regulon Gene Set

Merge the clusters built from operons and subsystems as, at this point, genes may be members of more than one cluster. We convert each cluster into a set of binary connections between all genes in the cluster. We then use the binary connections to form a single set of large clusters using transitive closure (e.g., if A is connected to B and B is connected to C, then A is connected to C). This leads to a condition in which any two genes that are connected are in the same cluster. This also ensures that no gene is a member of more than one cluster.

Step 5. Filter Atomic Regulon Gene Sets to Remove Low Correlation Genes

This is a provisional file, not the final typeset article

Split the merged clusters based on a distance computed between every pair of genes using the formula

$$Distance = \frac{(2 - (PCC + 1))}{2}$$

This corrects for genes with a low PCC value that may have been placed in a common cluster. New sub-clusters are formed by taking the two closest genes (based on the above defined distance) within the initial merged cluster and adding other genes to the growing sub-cluster. At each point, the gene with the minimum average distance to genes in the growing sub-cluster is added to the sub-cluster, until no such gene exists with an average distance less than or equal to 0.25. If this simple accretion algorithm produces a single sub-cluster, no splitting is required. If not, the sub-clusters become the close to final AR gene sets.

Step 6. Generate Final Set of Atomic Regulons

Estimate the ON/OFF status of each cluster in any specific experimental sample by a simple voting algorithm using the ON/OFF estimates for the genes that make up the AR (see Estimation of gene ON/OFF states). We then merge these AR gene sets if they have identical ON/OFF expression profiles across all experimental conditions. It is important to note that the resulting set of atomic regulons is not comprehensive (i.e., not all genes are placed into an AR), but this set attempts to capture many of the operational groups of genes. This merged set becomes the final set of ARs.

The source code for the AR inference algorithm is available on GitHub (https://github.com/janakagithub/atomic_regulons). Additionally, a service for AR inference has been implemented as a tool named "Compute Atomic Regulons" in the DOE Systems Biology Knowledgebase (KBase). Compute Atomic Regulons can be accessed at https://narrative.kbase.us. This service allows users to upload expression datasets and compute ARs for their genomes of interest. A tutorial detailing how to compute ARs in KBase is available in supplementary material.

Estimation of gene ON/OFF states

Our AR inference algorithm requires us to compute the correlation of expression for all genes across all available expression datasets. In this computation, we use expression values to assign all genes in all conditions to one of three possible states: ON, OFF, and UNKNOWN. These gene states are calculated for all genes in two steps.

Step 1. Initial Estimates of ON/OFF Calls Using an A Priori Identified Set of "Always ON" Genes

i) Determine the threshold for a gene to be considered ON based on the normalized expression of genes annotated with functions that are expected to be universally active. In total, we identified 80 functional roles from the SEED as universally active, largely from translation and transcription (see supplementary Table S1). We then consider the genes that implement these roles as ON.

This is a provisional file, not the final typeset article

ii) Empirically set initial ON/OFF calling thresholds for each experiment. The "on" threshold for experiment *i*, called N_i, is set as the 10th percentile of observed gene expression values for "always ON" genes in experiment *i*. The "OFF" threshold for experiment *i*, called F_i, is set as the 80th percentile of observed gene expression values that are below N_i in experiment *i*.

iii) Update initial ON/OFF calling thresholds for each experiment by computing the difference, $D_i = N_i - F_i$, in thresholds for each experiment, then finding the 25th percentile of D_i across all experiments, D_{25th}. For any experiment *i* where $D_i < D_{25th}$, set $F_i = N_i - D_{25th}$. This ensures that the "ON" and "OFF" calling thresholds are never too close together for a particular experiment.

iv) Using the updated ON/OFF calling thresholds for each experiment (N_i and F_i), make initial ON/OFF calls for each gene in experiment *i* by classifying any gene expression value less than F_i as OFF, greater than N_i as ON, and between F_i and N_i as UNKNOWN.

Step 2. Updating Gene-Specific ON/OFF Calls Using Gene Sets to Ensure Maximal Consistency

i) Construct draft sets of genes that are expected to be co-expressed with a high degree of confidence. Sets are constructed from: (i) operons; and (ii) subsystems.

ii) Vote within each gene set to determine the ON/OFF status of the entire set in each experiment based on majority rule. For example, if a set of four genes has two genes initially called ON, one UNKNOWN, and one initially called OFF, we update the calls for all genes in the set to ON since that is the majority of the initial calls. Ties (e.g., two ON and two OFF or all UNKNOWN) are classified as UNKNOWN.

Gene expression data

Gene expression data were collected from the Gene Expression Omnibus (GEO) (Edgar et al., 2002) and M3D databases (Faith et al., 2008), as well as from the laboratory of Dr. Paul Dunman in the case of *S. aureus*. Expression data were downloaded in the form of Affymetrix GeneChip® cell intensity (CEL) files. For each organism, the expression data from the CEL files were background corrected, normalized and summarized using Robust Multichip Averaging (Irizarry et al., 2003) as implemented in R/Bioconductor (http://www.bioconductor.org/) using the *rma* function default settings. In addition, the probe sets for each Affymetrix GeneChip were mapped to gene identifiers in the SEED genome database (Tintle et al., 2012;Overbeek et al., 2014).

Computation of hierarchical and k-means clusters

In order to compare the performance of ARs with k-means and hierarchical clustering analyses, we applied these methods to the normalized gene expression values across all experiments to generate 646 clusters, the same number of clusters identified using our AR approach for the purpose of illustration. K-means clusters were calculated with k set to 646 clusters using the Lloyd-Forgy algorithm (Lloyd, 1982) in R (Team, 2015). This approach is a non-standard use of the k-means clustering algorithm; the value of k was set to match the number of ARs for the sake of comparison. This clustering approach defines clusters by minimizing the Euclidian distance between individual points and cluster centers and is sensitive to variations in true cluster size and variance distributions.

This is a provisional file, not the final typeset article

For hierarchical clustering analysis, a Euclidean distance matrix was calculated between genes based on their normalized expression values across all experiments. We performed hierarchical clustering on this distance matrix using the average algorithm in R (Murtagh, 1985), which is equivalent to the unweighted pair group method using arithmetic mean (UPGMA). This clustering approach begins with each gene assigned to its own cluster and proceeds through an iterative process of joining the two nearest clusters together until all genes are linked into a hierarchical tree. We cut this tree at a given height in order to yield 646 clusters.

Assessing similarity of gene sets

We performed comparisons between gene sets produced by our AR inference algorithm and standard data-driven clustering algorithms to regulons in RegulonDB. We use the Jaccard coefficient, which measures similarity between finite sample sets. This coefficient is defined as the size of the intersection divided by the size of the union of the sample sets

$$Jaccard\ coefficient\ (RegDB, Rx) = \frac{|RegDB \cap Rx|}{|RegDB \cup Rx|}$$

where *RegDB* is a set of genes comprising a single regulon from the RegulonDB database, and *Rx* is a gene cluster inferred by one of the clustering algorithms being evaluated.

Additionally, we performed comparisons of ARs for *E. coli* with ARs for four different organisms, limiting our comparison to functional roles contained in SEED subsystems that occurred both in *E. coli* and the other organisms. For each atomic regulon in *E. coli* (*Ae*), we computed the set of relevant functional roles occurring in *Ae*, calling this set *Re*. Then, for each atomic regulon occurring in one of the other genomes (*Ax*), we considered the set of relevant roles occurring in *Ax*, calling this set *Rx*. We then computed the Jaccard coefficient for R*e* and R*x*.

Context Likelihood of Relatedness (CLR) algorithm

We performed a check on the gene contents of each regulon using the *Context Likelihood of Relatedness (CLR)* algorithm developed by the Gardner group at Boston University. Given a set of gene expression data, CLR predicts transcriptional regulatory relationships (Faith et al., 2007). The CLR algorithm belongs to a category of regulatory network inference algorithms that uses mutual information (MI) to analyze correlations in gene expression. In brief, the higher the MI score between two genes, the greater the information we derive on the expression states of the first gene from the pattern of states in the other, and therefore the greater the likelihood that one of the genes is directly or indirectly regulating the other, or that both genes are being regulated together. Our CLR calculations were done on the DeGNServer website, where the CLR algorithm has been parallelized to handle large gene sets (Li et al., 2013). A CLR score was recorded for each possible unique gene pair in *E. coli* (each possible gene-to-gene connection), resulting in a total of 9,367,956 CLR scores.

We discarded all but the top 0.765% of the CLR scores, representing all scores with values that were at least four standard deviations above the mean score. All gene pairs associated with these top CLR scores are then said to have support from CLR. We used these regulatory interactions inferred by CLR to validate and assess the ARs inferred by our algorithm and gene sets produced by the clustering algorithms at the same level of granularity. In this validation, we calculated the fraction of

This is a provisional file, not the final typeset article

all possible gene pairs in each AR that have a regulatory interaction also predicted by CLR. For example, if an AR includes three genes, then it has three possible gene pairs (i.e., AB, AC, CB). If CLR predicted a regulatory interaction between two of these pairs (i.e., AB, AC), then the fraction of supported interactions would be 0.667 (67% support). See Supplementary Information for a full treatment of the method.

Prediction of operons from genome sequence and gene calls

We predict operons from genomic data to serve as initial gene sets for our AR inference algorithm. In our operon prediction approach, sets of genes in the same strand within 200 base pairs up and down stream of each other were placed into the same operon. This mirrors the existing standard algorithms for operon prediction (Salgado et al., 2000).

Results

Characteristics of *E. coli* K-12 Atomic Regulons

Atomic regulons were computed for *E. coli* using the AR inference algorithm and expression data from the M3D dataset. *E. coli* was selected for initial construction of atomic regulons because it has the largest compendium of consistent gene expression data that is currently available, comprising microarray data from 907 experiments. The computation was performed in approximately 12 minutes on a single dual-core CPU. Four metrics were used to characterize the set of ARs produced for a genome: (i) the total number of genes assigned to ARs; (ii) the number of ARs computed; (iii) the number of genes that were found to be always ON; and (iv) the number of genes found to be always OFF across all experiments.

A summary of the results of the inference pipeline for *E. coli* is shown in **Table 2**, and the full set of atomic regulons is available in Supplementary Table S2. For *E. coli*, 2604 genes are assigned to 646 multi-gene ARs, corresponding to approximately 60% of its genome. The largest ARs in the set contain 292 and 69 genes, representing the set of genes that are always expressed or are never expressed in the experimental conditions represented by the expression array data, respectively. The largest atomic regulon with variable expression in our available data is comprised of 52 genes, which are primarily related to the functions of motility and chemotaxis. The remaining 644 ARs contain less than 52 genes each, with an average size of 3.48. Approximately 85% of multi-gene ARs contain 5 or less genes, with 328 ARs containing only 2 genes. Again, we deployed our AR computation algorithm into KBase, and we demonstrate this method at work in computing ARs for *E. coli* in a KBase Narrative (https://narrative.kbase.us/narrative/ws.14533.obj.1).

Comparison to clusters produced by hierarchical clustering and k-means clustering

The AR inference algorithm was compared to data-driven clustering algorithms by assessing the consistency of gene sets to both a curated set of regulons present in the database RegulonDB and to interactions predicted by the CLR algorithm. Each of the gene sets was compared to the reference set of regulons for *E. coli* found in RegulonDB through the calculation of a Jaccard coefficient (details in the Materials and Methods). The analysis revealed that the co-expressed gene sets generated by our AR inference algorithm had a higher level of similarity to the gold standard regulons in RegulonDB than the gene sets generated by either hierarchical or k-means clustering (**Figure 2a**).

This is a provisional file, not the final typeset article

The data driven clustering methods produce gene sets that are much less consistent; less than 8% of gene sets show 50% or more similarity to RegulonDB regulons. In contrast, approximately 50% of the RegulonDB regulons show a >50% similarity with the gene sets generated by our AR inference algorithm, and approximately 20% have a similarity >70%.

****Figure 2****

Additionally, we performed a similar analysis to investigate the impact of the inclusion of SEED subsystems in the AR inference algorithm. We compared AR similarity to RegulonDB by computing ARs with and without SEED subsystems using three different subsets of the available experimental data: (i) Full dataset of 907 experiments; (ii) 403 randomly selected experiments (50% of the data); and (iii) 91 randomly selected experiments (10% of the data) (**Figure 2**b). With the entire dataset of E. coli experiments, we observe only a small improvement in the similarity with RegulonDB with the inclusion of SEED subsystems. For the subset comprising of 50% of the data, we observe a slightly larger improvement with the inclusion of SEED subsystems when compared to the full dataset. These results are corroborated by the analysis conducted on **Figure 4**, in which we conclude that starting at 60% of the data, the improvements in our AR computation grow markedly slower.

For the subset comprising only 10% of the E. coli data, we see the largest difference in similarity with RegulonDB. In this case, the results are mixed. When the regulons have a larger degree of similarity with RegulonDB, they tend to be more accurate when generated with subsystems. When regulons have a lower degree of similarity with RegulonDB, they tend to be less accuracy when generated with subsystems. This result is due to the fact that the use of SEED subsystems will generally increase the average size of the regulons, as well as the fraction of the genome that is included in ARs. Note for example, that we have 150 ARs with at least some overlap with RegulonDB when subsystems are used, versus only 130 without. Thus, part of the observed decline in similarity at the lower end of the comparison curve is due to having more (and larger) regulons used in the comparison. This also emphasizes the vital role that the expression data plays in the AR algorithm. When limited data is available, it could make sense to apply our algorithm with and without subsystems and compare the results.

CLR is a mutual-information-based approach for the inference of gene regulatory networks from gene expression data (Faith et al., 2007). The CLR algorithm has been applied extensively to validate *E. coli* regulatory interactions and to identify missing links in the *E. coli* regulatory network. Also, CLR has performed well in DREAM challenges (dreamchallenges.org/publications/) when compared against other inference algorithms. We applied CLR to the same set of expression data that was used to build the ARs and clustering algorithm gene sets for *E. coli*, computing CLR scores for every possible pair of genes in *E. coli*. These scores quantify the level of mutual information found between the expression profiles of each pair of genes, with higher scores meaning more mutual information and thus a greater chance that one gene in the pair is co-regulated with or regulates the other gene in the pair.

****Figure 3****

The ARs generated by our algorithm and gene sets generated from hierarchical clustering and k-means clustering were assessed by binning ARs and gene sets into different ranges of support from CLR (**Figure 3a**). The highest category of support for an AR or gene set is one in which all of the

This is a provisional file, not the final typeset article

gene pairs in the set have a high CLR mutual information score with at least one other member of the gene set; 62% of ARs fall into this category, whereas 26% and 21% of hierarchical and k-means clustering gene sets fall into this category, respectively. We evaluated if a bias existed for the agreement between CLR and our AR inference algorithm based on AR size (**Figure 3b**). Given that larger ARs involve more possible gene pairs and thus more possible interactions that need to be validated by CLR, it was possible that there would be lower CLR support for the larger ARs. However, all 15 of the ARs with 13 or greater gene members were found in the two highest categories of CLR support (80-100% and 100%). More details on this analysis are available in the supplementary material. In particular, see supplementary Tables CS-1, CS-2A, CS-2B, and CS-2C.

Sensitivity of AR inference algorithms to the amount of available expression data

While advances in sequencing technology have led to the ability to produce expression data for non-model organisms, it remains the case that *E. coli* has the largest amount of expression data available among bacteria to date. The expression data have also been obtained in a large number of experimental conditions. This sets up a seemingly ideal case for the calculation of atomic regulons. However, most organisms will not have as much data available as *E. coli*, and it may be possible to obtain meaningful ARs with less expression data. Thus, we studied the impact of decreasing amounts of expression data on the inference of atomic regulons. We randomly selected different subsets of the *E. coli* data, repeated the AR analysis with each data subset, and compared the end results. For our subsets, we selected all increments of 10% of the available data (i.e., considering 10%, 20%, 30%, etc.). We repeated the analysis 100 times for each data size, selecting a different random subset of data with each simulation.

We evaluated the atomic regulons computed in each simulation based on four metrics. The first metric was the *number of genes* assigned to ARs of size two or greater. If insufficient data are available, natural noise in gene expression patterns will overwhelm any correlations that exist between genes, preventing the consolidation of ARs based on correlation of gene expression. Thus, the expectation is that fewer genes will be integrated together into multi-gene ARs. The second metric was the *number of multi-gene ARs* computed. As before, the signal to noise ratio that occurs with smaller amounts of data will prevent some multi-gene ARs from forming. Additionally, with fewer experiments, fewer genes will be differentially expressed. This will prevent some large ARs from being broken up into smaller ARs. The third and fourth metrics are the *number of genes that are always ON* and the *number of genes that are always OFF*, respectively. Fewer experiments may represent fewer experimental conditions and fewer differentially expressed genes, leading to an expectation of more genes that are either always ON or always OFF.

****Figure 4****

The results of the random sensitivity analysis support the expectations. As the amount of available data increases, the number of genes in ARs (**Figure 4a**) and the total number of ARs increase (**Figure 4b**). Additionally, the numbers of always ON genes (**Figure 4c**) and always OFF genes (**Figure 4d**) decrease with increasing amounts of expression data. Interestingly, in all cases, large improvements in each of the metrics are observed as the amount of data used increases from 10% to ~60% of the available data. Continued improvements in all metrics are observed until 100% of the data is used, but the improvements grow markedly smaller as more than 60% of the data is

This is a provisional file, not the final typeset article

considered. We also performed a simple 2-fold cross validation of the data, randomly splitting the 907 experiments for *E. coli* into equal non-overlapping sets (Table 1). When we compare the regulons computed from these two data subsets, the average Jaccard coefficients (0.83 ± 0.31 and 0.80 ± 0.35, mean and standard deviation) were nearly identical to the comparison of atomic regulons computed from the full dataset to either of the subsets (0.81 ± 0.35 and 0.80 ± 0.37 respectively). This result shows that the atomic regulons are very similar when only half the available experimental data is used.

Table 1

These results show that there is enough existing gene expression data for *E. coli* to enable the generation of high quality ARs, capturing a large portion of the atomic regulon space for this organism. Further, it is possible to compute ARs for *E. coli* with approximately the same quality using a much smaller amount of data.

Figure 4

Computation of atomic regulons across taxa

Studies of the evolution of bacterial transcriptional regulatory networks show conservation of regulatory modules/motifs and that gene co-regulation tends to be more conserved than regulatory genes and mechanisms. This conservation is observed across large phylogenetic distances for organisms with similar lifestyles (Madan Babu et al., 2006). Atomic regulons should be well suited to explore these trends. Thus, the AR inference algorithm was applied to study the consistency of gene co-regulation across five diverse genomes. ARs were computed for four additional organisms that have sufficient high-quality expression data available: *Shewanella oneidensis* MR-1, *Pseudomonas aeruginosa* PAO1, *Thermus thermophilus* HB8 *and Staphylococcus aureus* subsp. aureus Mu50 (organisms are ordered approximately from the closest to the farthest in terms of phylogenetic distance). To minimize noise, all the experimental datasets were selected by considering the same microarray platform with the same process of applied data normalization (see Materials and Methods). However, it should be noted that any source of expression data that is converted into normalized ON/OFF estimates can be used to infer ARs, ensuring that the technique is applicable to the increasing number of RNA-Seq experiments available for model and non-model organisms and to considering emerging data in combination with the rich amount of existing microarray expression data. The basic metrics describing ARs for each of the organisms are shown in **Table 2**. We demonstrate the application of our method to computing ARs for these genomes in a KBase Narrative (https://narrative.kbase.us/narrative/ws.14533.obj.1).

****Table 2****

The descriptive statistics for the ARs from the five organisms suggest interesting trends. It appears that there is a correlation between the number of expression experiments available for an organism and the percentage of genes included in a multi-gene AR. For both *T. thermophilus* and *S. aureus*, the same proportion of genes were included as for *E. coli*. Each of these have a number of experiments corresponding to greater than 60% of the data available for *E. coli*, consistent with the analysis above where subsets of the data were used to calculate ARs for *E. coli*. The organisms with fewer expression experiments (~26% of the experiments available for *E. coli*) have a lower

percentage of genes included in ARs. Given the caveat that the number of experiments does not necessarily reflect the number of unique environmental conditions represented in the set of experiments, it is interesting to speculate that more experimentation in these organisms would lead to inclusion of more genes in the calculated ARs. However, it is also apparent that the number of genes considered to be ON or OFF varies greatly, and that two of the three genomes with the highest number of ON genes have the highest number of experiments associated with them. This could reflect the diversity of the conditions represented by the experiments that leads to genes failing to be differentially expressed, the quality of the available expression data, or limitations of establishing thresholds for the ON/OFF state of genes in those organisms.

In order to compare the ARs inferred for the different organisms, Jaccard coefficients were computed for each AR in *E. coli* versus all ARs in each of the four other genomes. The distribution of these computed coefficients for each genome analyzed reveals that regulation appears to be more similar in genomes that are closer to *E. coli* both phylogenetically and in terms of lifestyle (**Figure 5a**). The only Gram positive genome we included in our study, *S. aureus*, was, as expected, a distant genome in terms of AR variation. The most distant genome in terms of AR variation was *T. thermophilus*, which, despite being "Gram negative", is phylogenetically distant and found in environments that are highly distinct from that of *E. coli*. Another interesting observation is how much variation exists in AR content between *E. coli* and the closest genomes analyzed, *S. oneidensis* and *P. aeruginosa*. Although these genomes are close to *E. coli* phylogenetically, only a small fraction of their atomic regulons have high compositional similarity. In contrast to expectations, the composition of co-regulated gene sets appears to be highly variable among closely related organisms. This result could support the notion that regulation is a highly adaptable system in the cell, but experimental studies specifically dedicated to this type of comparative analysis are needed in order to confirm this result.

Figure 5

Conservation of atomic regulons across taxa

In addition to conducting pairwise comparisons between *E. coli* and four other genomes, we focused on the specific atomic regulons that display the greatest consistency across the greatest number of genomes. To do this, we compared the Jaccard coefficients computed for each *E. coli* AR across all combinations of four, three, and two genomes. For each regulon, we retain the lowest Jaccard coefficients among the genomes being compared, with that coefficient being indicative of the compositional similarity of the AR among the full set of genomes (**Figure 5b**).

The analysis revealed 35 ARs with 100% compositional similarity between *E. coli* and at least one other genome in our set (the distribution of these 35 ARs across all genomes compared can be estimated from **Figure 5a**). This number is larger than the highest number of 100% identical ARs from our pair-wise AR comparisons (18), which indicates that the ARs with 100% similarity between *E. coli* and each individual genome have significant differences (i.e., different ARs were identical in each species). No identical regulons across *E. coli* and the four other genomes were observed, consistent with the large number of differences found between the ARs of *E. coli* and the ARs for *T. thermophilus* and *S. aureus*.

This is a provisional file, not the final typeset article

There were nine atomic regulons with at least 75% similarity across *E. coli* and two other genomes (**Table 3**). All ARs in **Table 3** were conserved with and were most similar to ARs in *S. oneidensis*, the organism with the smallest phylogenetic distance to *E. coli* in this study. The two largest ARs conserved among *E. coli* and two other genomes are both functionally related to cellular respiration. The largest was compositionally identical across three genomes (*E. coli*, *S. oneidensis*, and *P. aeruginosa*) comprising 13 genes, most of which are annotated with functions related to the NADH-ubiquinone oxidoreductase chain. All genes in this AR are members of the SEED Subsystem "Respiratory Complex I". These functions have been characterized in *E. coli* and are responsible for proton translocation in the electron transport chain (Weidner et al., 1993). The gene order is also highly conserved when compared to other bacterial genomes. Studies report these functional roles to be present in *T. thermophilus* (Yano et al., 1997), however, these functions were grouped into a much larger AR in this study, possibly due to the lack of expression data needed to decompose this large AR. The second largest identical AR is comprised of seven functional roles involved in the biogenesis of c-type cytochromes. These functional roles have been described in *E. coli* (Thony-Meyer et al., 1995) and each were present in the *S. oneidensis* and *P. aeruginosa* atomic regulons. None of the respiratory chain associated functional roles in the two largest ARs identified through this analysis were present in *S. aureus*, consistent with known diversity of respiratory chain components among bacteria.

The third and fourth largest ARs shared among *E. coli* and two other genomes are both shared with *S. oneidensis* and *S. aureus*. One AR comprising six functional roles is associated with tryptophan synthesis. In *T. thermophilus* and *P. aeruginosa*, only 40% of the functional roles are present in the corresponding AR. These results are consistent with research on the evolution and dynamics of the tryptophan biosynthesis pathway that found some functional roles are conserved across multiple organisms, while others were lost due to events such as operon splitting or gene fusions (Xie et al., 2003). The fourth largest AR contains functional roles associated with phosphate transport in *E. coli* (Amemura et al., 1985), and it was 100% identical to the AR identified in *S. oneidensis* and 80% in *S. aureus*. All results from the comparative analysis of ARs are reported in supplementary Table S3.

****Table 3****

To aid in genome annotation efforts, we integrated the computed ARs for this study into the SEED database. They can be accessed in the interface showing information for all genes that are members of a computed AR: http://pubseed.theseed.org/?page=AtomicRegulon&genome=all. Additionally, we implemented the AR inference algorithm in KBase (www.kbase.us), allowing the computation of AR for any genome of interest given an expression data set (see Material and Methods).

Discussion

Atomic regulons represent fundamental regulatory units of a cell, namely, the sets of genes that are always co-regulated. We have demonstrated a new method for the inference of atomic regulons that outperforms purely data-driven clustering methods for deriving sets of co-regulated genes. The approach relies on the formation of putative atomic regulons based upon operonal organization of genes and the highly curated sets of associated functions for genes (subsystems) used in the SEED and PATRIC databases (Overbeek et al., 2014;Wattam et al., 2014). These serve as high quality starting points that allow for the efficient determination of biologically meaningful co-regulated sets of genes. The inference algorithm is impacted by the current estimate of SEED Subsystems that are

This is a provisional file, not the final typeset article

continuously improved by the SEED annotation team. We apply SEED subsystems to build and refine our initial gene clusters, meaning that errors in the assignment of genes to subsystems can potentially lead to errors in initial clusters (which is mitigated by the concurrent use of expression data). More critically, clusters involving primarily genes that are not assigned to SEED subsystems will not benefit from this aspect of the algorithm. Additionally, our algorithm for setting the "ON" threshold for gene activity depends on the identification of a set of genes in the genome that are expected to be universally active. We currently identify these genes based on the functional roles assigned to them by the SEED. If these genes are not accurately annotated by the SEED, it will impact our ability to set an accurate "ON" threshold for gene activity.

Assessments of the gene sets produced from the AR inference algorithm and two popular clustering algorithms suggest that using a biologically meaningful starting point rather than clustering based solely on expression data produce co-regulated gene sets that are much more consistent with independent measures of gene regulatory interactions. This includes the calculation of the Jaccard coefficient for each AR and that of an independent information-theoretic method, CLR; atomic regulons were shown to be more consistent with gold-standard regulons found in RegulonDB and with mutual information coefficients between pairs of genes as determined by CLR. Further, the AR inference algorithm is robust to varying amounts of expression data and can be applied to diverse organisms. These properties make the described AR inference algorithm attractive for addressing questions of the regulation of fundamental cellular processes, the interactions amongst those fundamental processes, and the conservation of the processes across cellular life.

The atomic regulons inferred for *E. coli* constitute a large portion of the genes in the organism, with most ARs comprising 3 genes, on average. The two largest ARs also represent the two gene sets whose genes are always ON and always OFF in the set of expression experiments studied; we refer to these as *static ARs*. The 292 genes comprising the static AR that is universally expressed are primarily annotated with functions related to core cellular machinery. A large fraction of the genes code for ribosomal proteins and tRNA synthetases. These functions are considered to be constitutive, and were selected to set the threshold for active gene expression used in our ON/OFF calling algorithm. The remaining genes were classified as constitutively expressed and merged into the static AR as part of the inference algorithm. These functions are associated with core cellular machinery, such as DNA synthesis and ATP synthetases. Additionally, multiple genes with unknown functions are members of this atomic regulon. The 69 genes comprising the static AR that are not expressed are poorly annotated, with 63% annotated as hypothetical, putative or uncharacterized. These genes are inactive in all conditions studied, meaning they likely contribute little to cell fitness in any of the conditions tested. It is expected that the static ARs would break apart as more conditions are represented in the expression experiments. Our sensitivity analysis revealed that the ARs were somewhat insensitive to random reductions in the *E. coli* dataset, until those reductions exceed 50% of the dataset. This said, it is the diversity of experimental conditions included in the dataset that has the most significant impact on the accuracy of predicted regulons (Nicolas et al., 2012), and a simple count of the number of datasets is not necessarily an effective measure for this diversity. Unfortunately, we lack sufficient metadata for available expression datasets to rigorously explore the relative impact of experimental diversity on regulon quality.

An initial assessment of the conservation of atomic regulons across a diverse set of bacterial organisms reveals several interesting observations. First, it is clear that there is great variability in the membership of atomic regulons for different organisms. None of the ARs were fully conserved

This is a provisional file, not the final typeset article

among *E. coli* and the four organisms studied. This supports the notion that many strategies for the regulation of sets of functions by organisms exist. Second, despite variability among the ARs in different organisms, some ARs are well conserved over both short and long phylogenetic distances. Nine ARs were identified as conserved above a level of 75% of shared genes among *E. coli* and three of the organisms. Two of these were absolutely conserved with respect to gene membership; both were related to respiratory functions and each of the organisms are members of the Gamma Proteobacteria. Conservation was also identified for *E. coli*, *S. oneidensis*, and *S. aureus*, for two ARs associated with tryptophan amino acid biosynthesis and transport of phosphate. Third, having different levels of expression data for each organism has an impact on the quality of the ARs inferred. This can be seen as fewer genes being included in ARs and a higher proportion of genes being grouped into the static ARs.

We are continuing to investigate more statistically rigorous ON/OFF calling algorithms, which, instead of dichotomizing gene calls, yield a confidence metric that the gene is ON/OFF in a given condition. A calibrated continuous approach to ON/OFF calling may allow for more robustness in the calls. As new technologies are developed, and as current next generation sequencing technologies become more widely used and affordable, expression data for more experimental conditions will become available. This will lead to an improvement of the inference of ARs for the organisms considered in this study and any organism for which sufficient expression data are produced.

While improvements in the AR inference algorithm can be made, the current implementation of the method has been successfully used to extend our knowledge of the *B. subtilis* regulatory network (Faria et al., 2016). In that work, the atomic regulon inference approach was used to propose new additions to the *B. subtilis* regulatory network and led to the proposal of new functional annotations for genes. We integrated the ARs computed for this study into the SEED, allowing users to see the co-regulation of all genes in the ARs. These data complement the set of comparative genomics tools available in the SEED system. This information can be exceptionally valuable for researchers and curators, as co-regulation can group genes with well-known annotated functions with hypothetical genes, providing clues for new functional annotations. The AR inference algorithm and underlying gene ON/OFF calling algorithm have been included as a tool in KBase (www.kbase.us), which allows users to analyze their expression data in a feature rich platform designed to aid the discovery of new biological insights through the integration of multiple data types. We anticipate that a readily accessible implementation for the calculation of atomic regulons will enable many researchers to explore the relationships among these fundamental units of cellular function not only within their organism of choice, but also across the bacterial domain.

This is a provisional file, not the final typeset article

Funding

JF acknowledges funding from [FRH/BD/70824/2010] of the FCT (Portuguese Foundation for Science and Technology) PhD program. CSH and PW were supported by the National Science Foundation under grant number EFRI-MIKS-1137089. RT was supported by the Genomic Science Program (GSP), Office of Biological and Environmental Research (OBER), U.S. Department of Energy (DOE), and his work is a contribution of the Pacific Northwest National Laboratory (PNNL) Foundational Scientific Focus Area. This work was partially supported by an award from the National Science Foundation to MD, AB, NT and RO (NSF ABI-0850546). This work was also supported by the United States National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Service [Contract No. HHSN272201400027C].

Acknowledgements

We would like to acknowledge Emily M. Dietrich for her efforts and suggestions on editing and clarifying the manuscript.

References

Amemura, M., Makino, K., Shinagawa, H., Kobayashi, A., and Nakata, A. (1985). Nucleotide sequence of the genes involved in phosphate transport and regulation of the phosphate regulon in Escherichia coli. *J Mol Biol* 184**,** 241-250.

Benedict, M.N., Mundy, M.B., Henry, C.S., Chia, N., and Price, N.D. (2014). Likelihood-Based Gene Annotations for Gap Filling and Quality Assessment in Genome-Scale Metabolic Models. *Plos Computational Biology* 10. doi: ARTN e1003882 10.1371/journal.pcbi.1003882.

Chandrasekaran, S., and Price, N.D. (2010). Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* 107**,** 17845-17850. doi: 10.1073/pnas.1005139107.

De Smet, R., and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews Microbiology* 8**,** 717-729. doi: 10.1038/nrmicro2419.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30**,** 207-210.

Faith, J.J., Driscoll, M.E., Fusaro, V.A., Cosgrove, E.J., Hayete, B., Juhn, F.S., Schneider, S.J., and Gardner, T.S. (2008). Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res* 36**,** D866-870. doi: 10.1093/nar/gkm815.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 5**,** e8. doi: 10.1371/journal.pbio.0050008.

Faria, J.P., Overbeek, R., Taylor, R.C., Conrad, N., Vonstein, V., Goelzer, A., Fromion, V., Rocha, M., Rocha, I., and Henry, C.S. (2016). Reconstruction of the Regulatory Network for Bacillus subtilis and Reconciliation with Gene Expression Data. *Front Microbiol* 7**,** 275. doi: 10.3389/fmicb.2016.00275.

This is a provisional file, not the final typeset article

Huerta, A.M., Salgado, H., Thieffry, D., and Collado-Vides, J. (1998). RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res* 26**,** 55-59.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4**,** 249-264. doi: 10.1093/biostatistics/4.2.249.

Karp, P.D., Weaver, D., Paley, S., Fulcher, C., Kubo, A., Kothari, A., Krummenacker, M., Subhraveti, P., Weerasinghe, D., Gama-Castro, S., Huerta, A.M., Muniz-Rascado, L., Bonavides-Martinez, C., Weiss, V., Peralta-Gil, M., Santos-Zavaleta, A., Schroder, I., Mackie, A., Gunsalus, R., Collado-Vides, J., Keseler, I.M., and Paulsen, I. (2014). The EcoCyc Database. *Ecosal Plus* 2014. doi: 10.1128/ecosalplus.ESP-0009-2013.

Kitano, H. (2002). Systems biology: a brief overview. *Science* 295**,** 1662-1664. doi: 10.1126/science.1069492.

Kochanowski, K., Sauer, U., and Chubukov, V. (2013). Somewhat in control--the role of transcription in regulating microbial metabolic fluxes. *Curr Opin Biotechnol* 24**,** 987-993. doi: 10.1016/j.copbio.2013.03.014.

Li, J., Wei, H., and Zhao, P.X. (2013). DeGNServer: deciphering genome-scale gene networks through high performance reverse engineering analysis. *Biomed Res Int* 2013**,** 856325. doi: 10.1155/2013/856325.

Lloyd, S.P. (1982). Least squares quantization in PCM. *Information Theory, IEEE Transactions on* 28**,** 129-137.

Madan Babu, M., Teichmann, S.A., and Aravind, L. (2006). Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* 358**,** 614-633. doi: 10.1016/j.jmb.2006.02.019.

Murtagh, F. (1985). Multidimensional clustering algorithms. *Compstat Lectures, Vienna: Physika Verlag, 1985* 1.

Nicolas, P., Mader, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M., Aymerich, S., Becher, D., Bisicchia, P., Botella, E., Delumeau, O., Doherty, G., Denham, E.L., Fogg, M.J., Fromion, V., Goelzer, A., Hansen, A., Hartig, E., Harwood, C.R., Homuth, G., Jarmer, H., Jules, M., Klipp, E., Le Chat, L., Lecointe, F., Lewis, P., Liebermeister, W., March, A., Mars, R.A., Nannapaneni, P., Noone, D., Pohl, S., Rinn, B., Rugheimer, F., Sappa, P.K., Samson, F., Schaffer, M., Schwikowski, B., Steil, L., Stulke, J., Wiegert, T., Devine, K.M., Wilkinson, A.J., van Dijl, J.M., Hecker, M., Volker, U., Bessieres, P., and Noirot, P. (2012). Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. *Science* 335**,** 1103-1106. doi: 10.1126/science.1206848.

Novichkov, P.S., Rodionov, D.A., Stavrovskaya, E.D., Novichkova, E.S., Kazakov, A.E., Gelfand, M.S., Arkin, A.P., Mironov, A.A., and Dubchak, I. (2010). RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Research* 38**,** W299-W307. doi: 10.1093/nar/gkq531.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., and Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33**,** 5691-5702. doi: 33/17/5691 [pii]

This is a provisional file, not the final typeset article

10.1093/nar/gki866.

Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A.R., Xia, F., and Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* 42**,** D206-214. doi: 10.1093/nar/gkt1226.

Rodionov, D.A. (2007). Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chemical Reviews* 107**,** 3467-3497. doi: 10.1021/cr068309+.

Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., and Collado-Vides, J. (2000). Operons in Escherichia coli: Genomic analyses and predictions. *Proceedings of the National Academy of Sciences of the United States of America* 97**,** 6652-6657. doi: DOI 10.1073/pnas.110147297.

Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muniz-Rascado, L., Garcia-Sotelo, J.S., Weiss, V., Solano-Lira, H., Martinez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernandez, S., Alquicira-Hernandez, K., Lopez-Fuentes, A., Porron-Sotelo, L., Huerta, A.M., Bonavides-Martinez, C., Balderas-Martinez, Y.I., Pannier, L., Olvera, M., Labastida, A., Jimenez-Jacinto, V., Vega-Alvarado, L., Del Moral-Chavez, V., Hernandez-Alvarez, A., Morett, E., and Collado-Vides, J. (2013). RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res* 41**,** D203-213. doi: 10.1093/nar/gks1201.

Shlomi, T., Eisenberg, Y., Sharan, R., and Ruppin, E. (2007). A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol Syst Biol* 3**,** 101. doi: 10.1038/msb4100141.

Team, R.C. (2015). R: A language and environment for statistical computing. Vienna, Austria; 2014. *URL* http://www/. *R-project. org*.

Thony-Meyer, L., Fischer, F., Kunzler, P., Ritz, D., and Hennecke, H. (1995). Escherichia coli genes required for cytochrome c maturation. *J Bacteriol* 177**,** 4321-4326.

Tintle, N.L., Sitarik, A., Boerema, B., Young, K., Best, A.A., and DeJongh, M. (2012). Evaluating the consistency of gene sets used in the analysis of bacterial gene expression data. *BMC bioinformatics* 13**,** 193.

von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research* 33**,** D433-D437. doi: 10.1093/nar/gki005.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10**,** 57-63. doi: 10.1038/nrg2484.

Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E.K., Olson, R., Overbeek, R., Pusch, G.D., Shukla, M., Schulman, J., Stevens, R.L., Sullivan, D.E., Vonstein, V., Warren, A., Will, R., Wilson, M.J., Yoo, H.S., Zhang, C., Zhang, Y., and Sobral, B.W. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 42**,** D581-591. doi: 10.1093/nar/gkt1099.

Weidner, U., Geier, S., Ptock, A., Friedrich, T., Leif, H., and Weiss, H. (1993). The gene locus of the proton-translocating NADH: ubiquinone oxidoreductase in Escherichia coli. Organization of the 14 genes and relationship between the derived proteins and subunits of mitochondrial complex I. *J Mol Biol* 233**,** 109-122. doi: 10.1006/jmbi.1993.1488.

Xie, G., Keyhani, N.O., Bonner, C.A., and Jensen, R.A. (2003). Ancient origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiol Mol Biol Rev* 67**,** 303-342, table of contents.

This is a provisional file, not the final typeset article

Yano, T., Chu, S.S., Sled, V.D., Ohnishi, T., and Yagi, T. (1997). The proton-translocating NADH-quinone oxidoreductase (NDH-1) of thermophilic bacterium Thermus thermophilus HB-8. Complete DNA sequence of the gene cluster and thermostable properties of the expressed NQO2 subunit. *J Biol Chem* 272**,** 4201-4211.

Young, R.A. (2000). Biomedical discovery with DNA arrays. *Cell* 102**,** 9-15.

Tables and Figures

Tables

Table 1 Average Jaccard similarity coefficient between each set of atomic regulons from 2-fold the cross validation.

	Set1	Set2	All ARs
Set1	1	0.80 ± 0.35*	0.89 ± 0.26
Set2	0.83 ± 0.31	1	0.92 ± 0.19
All ARs	0.81 ± 0.35	0.80 ± 0.37	1

*Mean ± Standard Deviation

Table 2 Atomic regulon statistics for *Shewanella oneidensis* MR-1, *Pseudomonas aeruginosa* PAO1, *Thermus thermophilus* HB8 *and Staphylococcus aureus* subsp. aureus Mu50 *and Escherichia coli* K-12.

	Array datasets	No. ARs	% Genes in ARs	Gene always ON	Gene always OFF	Genome size
Escherichia coli K-12	907	646	2604 (60%)	292	69 (1.6%)	4309
Shewanella oneidensis MR-1	245	335	1559(37%)	265	32 (0.8%)	4167
Pseudomonas aeruginosa PAO1	236	423	2427(43%)	557	78 (1.4%)	5682
Thermus thermophilus HB8	543	196	1422(63%)	692	27 (1.2%)	2239
Staphylococcus aureus subsp. aureus Mu50	852	397	1749(63%)	447	28 (1%)	2770

This is a provisional file, not the final typeset article

Table 3 Atomic regulons similarity >75% across *E.coli* and two other genomes.

AR ID	AR Size	Associated genomes	Functional role summary
15	13	*S. oneidensis* MR-1 *P. aeruginosa* PAO1	NADH-ubiquinone oxidoreductase chain
54	7	*S. oneidensis* MR-1 *P. aeruginosa* PAO1	Biogenesis of c-type cytochromes
57	6	*S. oneidensis* MR-1 *S. aureus*	Tryptophan synthesis
112	5	*S. oneidensis* MR-1 *S. aureus*	Phosphate transport system
316	3	*S. oneidensis* MR-1 *P. aeruginosa* PAO1	Molybdenum transport system
362	2	*S. oneidensis* MR-1 *S. aureus*	Heat shock proteins
398	2	*S. oneidensis* MR-1 *P. aeruginosa* PAO1	Paraquat-inducible proteins
500	2	*S. oneidensis* MR-1 *S. aureus*	Ribonucleotide reductase

Figure legends

Figure 1: Atomic Regulon Inference. Six steps of Atomic Regulon (AR) inference algorithm. Step 1. Generate Initial Atomic Regulon Gene Sets. Initial clusters are proposed using gene clustering within putative operons and membership of genes within SEED subsystems. Step 2. Process Gene Expression Data and Calculate Pairwise Expression Profile Similarities. Integrate gene expression data, load the normalized data, and compute Pearson correlation coefficients. Step 3. Expression Informed Splitting of Initial Atomic Regulon Gene Sets. Split operon and subsystem-based clusters using the criterion that genes in a set must have pairwise expression data profiles greater than a set Pearson correlation coefficient (PCC) threshold. Step 4. Restrict Gene Membership to One Atomic Regulon Gene Set. Merge the clusters built from operons and subsystems then use the binary connections to form a single set of large clusters using transitive closure. This also ensures that no gene is a member of more than one cluster. Step 5. Filter Atomic Regulon Gene Sets to Remove Low Correlation Genes. Split the merged clusters based on a distance computed between every pair of genes. This corrects for genes with a low PCC value that may have been placed in a common cluster. Step 6. Generate Final Set of Atomic Regulons. Estimate the ON/OFF status of each cluster in any specific experimental sample by a simple voting algorithm using the ON/OFF estimates for the genes that make up the AR. This merged set becomes the final set of ARs.

This is a provisional file, not the final typeset article

Figure 2: Comparison of RegulonDB regulons with hierarchical clustering, k-means clustering and atomic regulons. a) The Jaccard coefficient comparing *E. coli* RegulonDB regulons vs. the clustering methods is shown as a percentage of similarity. b) Comparison AR similarity to RegulonDB regulons with and without inclusion of SEED subsystems using 100%, 50% and 10% of experiment data.

Figure 3: Degree of CLR support. a) CLR support compared between our ARs, and ARs produced via k-means clusters and hierarchical clusters. b) CLR support for our AR construction method, broken down by different AR sizes.

Figure 4: Sensitivity analysis of Atomic Regulon inference for *Escherichia coli* K-12. a) Average number of genes in atomic regulons. b) Average number of atomic regulons. c) Average number of genes always ON d) Average number of genes always OFF. Standard deviation error bars represent the variation across 100 data set randomizations from random sampling of experiments.

Figure 5: Comparison of *E. coli* Atomic Regulons versus all other ARs for *Shewanella oneidensis* MR-1, *Pseudomonas aeruginosa* PAO1, *Thermus thermophilus* HB8 *and Staphylococcus aureus* subsp. aureus Mu50. a) The % of similarity is given by the Jaccard coefficient, which is defined as the size of the intersection divided by the size of the union of the sample sets. b) The % of similarity is given by the Jaccard coefficient, which is defined as the size of the intersection divided by the size of the union of the sample sets. Jaccard coefficients computed for each *E. coli* AR across all combinations of four, three, and two genomes. The number of Atomic Regulons in the y-axis is represented in a log2 scale.

Supplementary Material

Supplementary Table S1: 80 functional roles considered to be “always ON”.

Supplementary Table S2: Complete set of atomic regulons computed for *E. coli*.

Supplementary Table S3: Results from comparative analysis.

CLR Supplementary Table CS-1 - CLR Support for Atomic Regulons (ARs). Method: percentage of support for an AR is given by the percent of genes in the AR that have at least one high-scoring CLR edge that links to another gene in the same AR.

CLR Supplementary Table CS-2A - Support for our Atomic Regulons (ARs). Method: percentage of support for an AR is given by the percent of all possible gene-to-gene connections in the same AR that have a corresponding high-scoring (above > SD4) CLR edge. This is a much tougher support criterion to meet than the criterion used in CLR Supplementary Table CS-1.

CLR Supplementary Table CS-2B - Support for k-means clusters. Method: percentage of support for an AR (k-means cluster) is given by the percent of all possible gene-to-gene connections in the same cluster that have a corresponding high-scoring (above > SD4) CLR edge.

This is a provisional file, not the final typeset article

CLR Supplementary Table CS-2C - Support for hierarchical clusters. Method: percentage of support for an AR (hierarchical cluster) is given by the percent of all possible gene-to-gene connections in the same cluster that have a corresponding high-scoring (above > SD4) CLR edge.

This is a provisional file, not the final typeset article