2014

THE EVOLUTIONARY DYNAMICS OF TRANSCRIPTION FACTORS, OPERATORS, AND THEIR TARGET GENES ACROSS PROKARYOTES

Marc del Grande
Wilfrid Laurier University, delx8580@mylaurier.ca

THE EVOLUTIONARY DYNAMICS OF TRANSCRIPTION FACTORS, OPERATORS, AND THEIR TARGET GENES ACROSS PROKARYOTES

A Thesis Submitted to
Wilfrid Laurier University

by

MARC DEL GRANDE

In Partial Fulfillment
for the Degree of Master of Science
in the Department of Biology
Waterloo, Ontario

Abstract

In prokaryotes, transcriptional regulation commonly involves a transcription factor (TF) binding to a particular conserved sequence of nucleotides (operator). Binding elicits a transcriptional response, either activation or repression. The evolution of gene regulation has been identified as a primary driver of species diversity, making it an important area of research. This work examined the dynamics of the interactions between TFs and operators, and TFs and their primary target genes in attempt to assess the rapid evolution of transcriptional regulatory networks (TRNs) across a diverse set of prokaryotes. Using software packages, operator sequences from *Escherichia coli* K12 were compared to every bacterial and archaeal genome within the NCBI's RefSeq database. This revealed that, based on genome composition, native TFs have a greater probability of interacting with sequences within their host's genome than those of other species, indicating that appropriate operators may form spontaneously, and often, within a genome. TFs and target genes were assessed through co-occurrence patterns. Recently, research has shown that repeated co-occurrence of two genes is evidence for a functional interaction. Co-occurrence can be observed and quantified in phylogenetic profiles by measuring mutual information (MI); this is a metric of how often two genes co-occur adjusted for what is expected by chance. By measuring MI for all two-gene combinations from a subset of genomes from NCBI's RefSeq database, results showed that, in $> 97\%$ of the organisms observed, TFs form looser functional interactions than other genes, indicating that TFs do not form lasting associations on the evolutionary time scale. These results suggest regulatory interactions are not as specific or conserved as those between most other gene products. Together, these results suggest that TRNs evolve rapidly across most, if not all prokaryotes.

Acknowledgements

I want to thank Dr. Gabriel Moreno-Hagelsieb for his patience, encouragement, and for the freedom he gave me to make my own mistakes. As well, for their patience, willingness to approach unfamiliar territory, and still offer great feedback, my committee members: Drs. Allison McDonald and Christine Dupont. Also to Dr. Ernesto Pérez-Rueda for his invaluable feedback and willingness to share his work with me. I want to offer special thanks to all current and past members of the lab of Computational ConSEQUENCES who have offered their support. Of special note: Dr. Michael Lynch and Gregory Vey for always lending a critical eye as well as much needed computational support; Elisabeth Kell and Dr. Luis David Alcaraz for being outstanding friends and fellow scientists that I hope to work with again in the future, un abrazo.

Table of Contents

Abstract **ii**

Acknowledgements **iii**

1 General Introduction **1**

1.1 Transcriptional regulation 2

1.2 Transcription factor binding sites 3

1.3 Co-occurring genes and functional interactions 4

2 Operator sequence specificity from the *Escherichia coli* perspective **7**

2.1 Materials and Methods 7

2.1.1 Sequence Data 7

2.1.2 Consensus and Patser implementation 8

2.1.3 MEME and MAST implementation 9

2.1.4 Statistics and Graphics 9

2.2 Results and Discussion 10

2.2.1 Patser results reveal the need for a more specific search tool . 10

2.2.2 Motif Alignment and Search Tool shows high rates of false operators in *Escherichia coli* 12

2.3 Conclusion 16

2.3.1 Recommendations 16

3 Loose evolutionary relationship between transcription factors and other gene products **18**

3.1 Materials and Methods 18

3.1.1 Phylogenetic profiles and mutual information 18

3.1.2 Profiles of phylogenetic profiles, the p-cubic analysis 19

3.1.3 P-cubic differences, the $\Delta P3$ 19
3.1.4 Transcription factor identification 21
3.1.5 Transcription factor prediction 21
3.1.6 Codon adaptation index 22
3.2 Results and Discussion 22
3.2.1 Protein family hidden Markov models are adequate for predicting transcription factors 22
3.2.2 P-cubics show that transcription factors have less conserved interactions than other genes 23
3.2.3 Low codon adaptation indices suggest frequent horizontal gene transfer 28
3.3 Conclusion 30
3.3.1 Recommendations 30

4 Summary **32**

References **36**

A False positive transcription factor predictions **42**

List of Tables

A.1 NCBI data for predicted, non-currated transcription factors in *Escherichia coli* 44

A.2 NCBI data for predicted, non-currated transcription factors in *Bacillus subtilis* 45

List of Figures

1.1 Weight matrix and sequence logo for cyclic-AMP response protein-bound operators 4

2.1 Schematic of Consensus/Patser sequence processing 8
2.2 Patser positive alignments for operators bound by global transcription factors 11
2.3 Proportion of positive alignments from Motif Alignment and Search tool for operators bound by global transcription factors 14
2.4 Motif Alignment and Search Tool results for operator motifs in RefSeq genomes. 15

3.1 Schematic for calculating ΔP3 Values 20
3.2 Transcription factors as identified by manual curation, semi-automated curation, and Protein Family hidden Markov models 24
3.3 Comparison p-cubics for curated and predicted transcription factors in *Escherichia coli* and *Bacillus subtilis* 26
3.4 Cumulative proportion of $\Delta P3$ between transcription factor-coding genes and non transcription factor-coding genes 27
3.5 Comparison of the codon adaptation indices of genes coding transcription factors and other genes 29

A.1 Decision tree for determining validity of potentially falsely predicted transcription factors 43

1 General Introduction

Transcriptional regulatory networks (TRNs) are vastly interconnected networks with three primary components: transcription factors (TFs), TF binding sites, and target genes. Recently there have been efforts to characterize TRNs in many organisms across both prokaryotes [1, 2] and eukaryotes [3]. The elucidation of such networks aids in the understanding of how transcriptional regulation (TR) is carried out globally (for example, in response to immediate environmental changes), as well as how these networks can adapt, or reorganize, on an evolutionary timescale when compared across species.

The question of how TRNs rewire has become a targeted area of study. This is primarily due to the recent paradigm that variation in gene expression serves as a driving force in species diversity [3, 4]. It has already been hypothesized that TRNs rapidly evolve [5, 6, 7], however this has only been demonstrated by comparing small reconstructions of a TRN between well-studied model organisms and other closely related species. Though these results support the hypothesis, they have not demonstrated the possibility for rapidly evolving TRNs across diverse sets of organisms.

The objectives of this work were to characterize the evolutionary dynamics of the primary components of a TRN: TFs, their binding sites, and their target genes for a diverse set of prokaryotes, including some Archaea. It is hypothesized here that if TRNs rapidly evolve in all prokaryotes, then TF binding sites should be non-specific enough to facilitate novel binding of TFs to new promoter regions and homologous TFs, on average, should not be observed to regulate homologous genes. By assessing these components of a TRN, it is possible to surmount the daunting task of reconstructing them, making it possible to test this hypothesis over a large set of prokaryotes.

Ultimately, reorganization of TRNs can serve as a possible source of variation for organisms in unstable environments. This work, though computational in nature, is assessing the success of this predicted evolutionary strategy and thus attempts to answer questions beyond those asked with traditional bioinformatics. Aspects of phylogenetics, evolutionary biology, bacterial ecology and genetics, and theoretical biology are integrated here alongside bioinformatic approaches to discover if the ability for a TRN to rapidly evolve is a trait observed for all currently sequenced prokaryotes.

1.1 Transcriptional regulation

The regulation of gene expression is a phenomenon ubiquitous to all known life. TFs are proteins that play an important role in regulating expression by modulating the rate of transcription of an open reading frame by binding to a specific stretch of DNA and affecting the action of RNA polymerase. Though not all domains of life regulate gene expression in identical fashions, the existence and action of TFs is uniform.

There are two classes of TFs currently identified in prokaryotes, global and local [8]. Global TFs are generally responsible for regulating and co-regulating a large number of genes. Typically these genes include genes coding for other TFs, and other genes that are diverse in terms of their protein products. For example, in *Escherichia coli* K12, there are 7 clearly identified global TFs. Together these TFs are involved in half of all known regulatory interactions involving TFs [1, 9]. Cyclic-AMP response protein (CRP) is one of *E. coli*'s best studied global TFs. According to recent databases [1, 10, 11, 12] CRP is involved in the regulation or co-regulation of 473 genes including 245 operons. The set of genes regulated by CRP, also called the CRP regulon, includes other TFs, membrane-bound transport proteins, enzymes involved in aerobic and anaerobic respiration, and many more [1, 10]. Local regulators control a small number of genes and sometimes act as co-regulators with a global regulator. An example of a classic local regulator is LacI, known to regulate the lac operon,

responsible for lactose metabolism in *E. coli* [13].

This dichotomy of TFs is also observed outside of *E. coli*. For example, CodY is a global TF found to regulate 151 genes in *Bacillus subtilis* [11, 14]. These global TFs have been termed "regulatory hubs" [15] and it has been proposed that the existence of these hubs enables TRNs to be particularly robust to random perturbations [16]. Such perturbations could include gene duplication, loss, mutation, and transfer within or between organisms.

1.2 Transcription factor binding sites

When Jacob and Monod first elucidated the regulatory mechanisms of the lac operon, they called the site where lacI bound, the operator [13]; this classic term is still widely used today, especially when referring to the lac operon. The term "operator" will be used here to refer to any TF binding site.

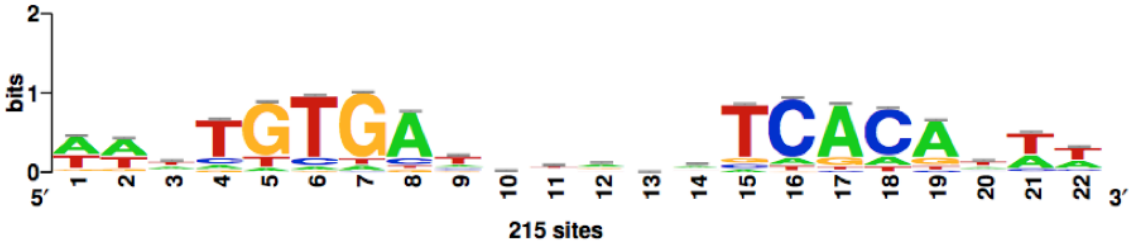
Operators are typically short DNA motifs between 12 and 30 base pairs in length and are known for not having a highly specific consensus sequence; because of this, they are often referred to as "fuzzy" in nature [17]. Due to their short length and inherent fuzzy properties, motif search algorithms are constantly being developed and improved upon in order to facilitate the identification and prediction of operators [18].

Another consequence of these inherent properties of operators is that they have low information content (IC) [17, 18]. The IC of a sequence of DNA is dependent on the length of the motif, as well as the background frequency of each nucleotide [19]. Simply, IC is calculated by aligning similar motifs and counting the occurrences of each nucleotide at each position into a weight matrix (see Figure 1.1 for example with CRP-bound operators) and using that matrix to calculate how frequently another motif would be expected to occur, by chance, in a DNA sequence where the *a priori* nucleotide frequencies are known [19]. Short, variable sequences result in low ICs, i.e. they are likely to exist in the search sequence often, just by chance.

Since operators intrinsically possess low ICs, and they exist as target sites for TF

A |	106	107	72	21	15	13	16	159	41	44	39	90	44	90	13	23	164	33	150	56	73	68
G |	16	24	32	14	162	7	174	14	30	48	37	53	65	34	19	6	29	4	33	28	6	16
C |	11	8	27	28	4	24	6	27	35	73	50	27	49	38	17	169	9	157	13	36	21	25
T |	82	76	84	152	34	171	19	15	109	50	89	45	57	53	166	17	13	21	19	95	115	106

(a) Weight matrix for CRP-bound operators



(b) Sequence logo for CRP-bound operators

Figure 1.1: Weight matrix (a) constructed from 215 operator sequences recognized by the TF, cyclic AMP response protein (CRP). CRP is a global TF in *E. coli*; this matrix was constructed by aligning experimentally determined CRP binding sites and counting the occurrence of each nucleotide at each position. Sequence logo (b) is a graphical representation of the CRP weight matrix that uses IC to determine the height of the nucleotides at the corresponding positions; the height of a letter is representative of how often that nucleotide occurs at its corresponding position, adjusted for what would be expected.

binding, it is important to ask how a genome has adapted in response to possibly frequent binding of TFs to false operators. There exists the possibility though that the genome has not adapted, and instead a high rate of false positives persists. If this is so, small mutations will cause *de novo* binding sites to appear frequently. The existence of such a property would facilitate the evolution of new operators and the reshuffling of a gene's TF profile (the set of TFs involved in regulating a gene) [17], thus providing a means for the rapid evolution of a TRN.

1.3 Co-occurring genes and functional interactions

When there exists two or more genes whose protein or RNA products rely on each other to function, the genes' homologs will tend to either be present all together in one organism, or they will all be absent [20, 21]. An example of two genes whose products rely on each other for proper functioning is recA and recF. RecA is a protein involved in homologous recombination, specifically in response to errors in DNA replication [22]. RecA-dependent recombination is stabilized by, among other

proteins, RecF [23, 24]. RecF knockout mutants show a decrease in RecA mediated DNA repair, thereby hindering the organism's ability to repair DNA damage during replication [25]. To contrast, FtsH is a protein involved in the degradation of improperly folded cytoplasmic and membrane proteins [24, 26]. FtsH is not dependent on or required for the functioning of either RecA or RecF, thus homologs of ftsH and recA or recF are under no constraints to be co-present in an organism, whereas there exists selective pressure for recA and recF homologs to co-occur.

The presence and absence of genes and their homologs is documented into matrices called phylogenetic profiles (PPs) where presence and absence across a range of species are recorded as a 1 and 0 respectively. Using PPs, the extent to which two genes co-occur can be directly observed and quantified. Co-occurrence is measured here by assessing how often two homologs are co-present and co-absent across all observed species above what is expected by chance using a metric called mutual information (MI). If the homologs of two genes appear in nearly all species, it is expected that they will co-occur frequently by chance; however, if a different gene pair is only present in half of the species, then frequent co-occurrence may indicate that the gene pair act together functionally, structurally, or in the same chemical pathway, as described above. MI is maximal if a gene pair is present in 50% of species and the pair always occur together [27]. There is no maximum score for mutual information, however the minimum score is 0 and this occurs when the gene pair is present or absent in all species.

Giving support to using MI as a metric for predicting functional interactions, it has recently been demonstrated that pairs of genes for proteins involved in the same metabolic pathways tend to possess higher MI than those genes for proteins which have no known interactions [20]; the same can be said for genes contained within the same operon [28]. Drawing from the previous example with recA, recF, and ftsH: the MI score for recA and recF as gene pairs is six orders of magnitude greater than when calculating pairs with ftsH (unpublished research). Using this principle of co-occurring genes, the strength of functional interactions between TFs and their target genes was assessed.

In a flexible TRN, there is frequent reshuffling of TF to target gene pairs [6]. As a consequence, these gene pairs in one organism will not be under evolutionary constraints to consistently co-occur in other related species. This will result in low MI between the gene for the TF and its target gene. If the properties of rapidly evolving TRNs are truly ubiquitous, then low MI should be observed for most TF and target gene pairs in all organisms.

2 Operator sequence specificity from the *Escherichia coli* perspective

Operators, short segments of a genome recognized by TFs, play a pivotal role in the regulation of gene expression by directing the binding of a TF, which then elicits an activation or repression response in the transcription of an open reading frame. Since operators are short sequences, the likelihood of one occurring by chance in a genome is higher than that of a long sequence, such as an open reading frame. The frequency of *Escherichia coli* operator occurrence in a diverse set of non-redundant, Prokaryotic genomes is analyzed here to determine if there is a bias towards high rates of false positives in closely related organisms. It is hypothesized that operators are not markedly unique sequences compared to their host genome and "near-operators" occur often, by chance. These results may have impacts in the understanding of how regulatory networks are able to reorganize in short evolutionary time scales.

2.1 Materials and Methods

2.1.1 Sequence Data

Operator sequence data for sites recognized by 175 different TFs (2841 sequences) in *E. coli* str. K12 substr. MG1655 were retrieved from the most recent release of RegulonDB, a manually curated database which collects its data from literature sources [1, 10]. Gene expression analysis, site mutagenesis, binding of cellular extracts or purified proteins and, inference based on consensus sequence techniques were used together or separately to identify all of these operators [1, 10]. Operators inferred from consensus sequence techniques alone are unverified computational predictions and thus were not included in the analysis. Any TF that had fewer than 10 oper-

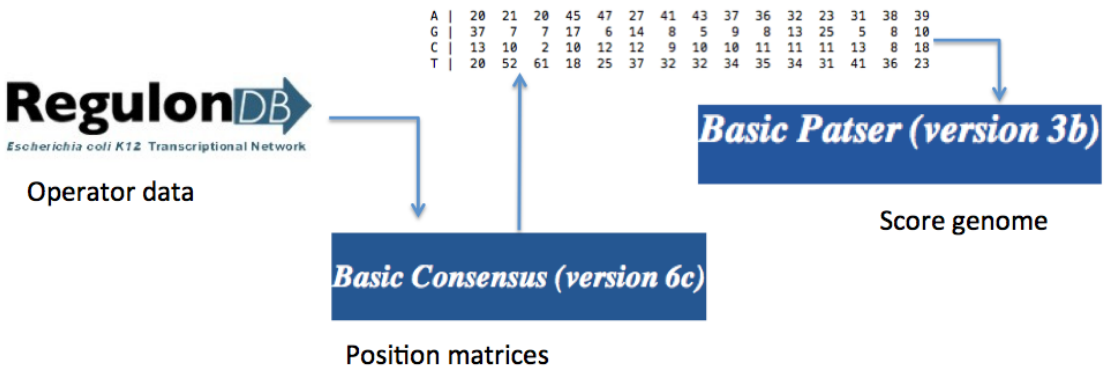


Figure 2.1: Schematic of Consensus/Patser sequence processing. Sequence data is collected from RegulonDB, Consensus creates PSWM which are then used in Patser to compare to query genomes.

ators associated with it was not included in the analysis. Genome sequences were collected from NCBI's RefSeq database [29, 30, 31] by April 2013.

2.1.2 Consensus and Patser implementation

Using PYTHON, a programming language, each set of operators recognized by a single TF in *E. coli* was processed through existing software packages, Consensus and Patser [32, 33, 34] available at <`ftp://beagle.colorado.edu/pub/consensus`>. Consensus was used to create weight matrices for each set of operator sequences and Patser queried input genomes (Figure 2.1) by aligning the weight matrices to every possible position and returning a p-value for the null hypothesis that the aligned sequence is dissimilar to the weight matrix. Python scripts are available at <`https://github.com/marc-delgrande/masters_thesis`>.

The organisms selected for analysis were a subset of those used by Price and colleagues in their 2008 study on the frequency of horizontally transfer of genes for TFs [9]. The operator sequences used to construct the weight matrices were those that are recognized by global TFs. The proportion of positive matches for a given operator was recorded as the number of alignments that returned a p-value less than 0.05 normalized to the length, in base pairs, of the query genome.

2.1.3 MEME and MAST implementation

A pipeline similar to that of the Consensus/Patser analysis was scripted using the software packages MEME (Multiple Em for Motif Elicitation) and MAST (Motif Alignment & Search Tool) [35, 36, 37] available at <http://www.sdsc.edu/MEME>. MEME constructed weight matrices from the operator sequence data for *E. coli* that were used to search for significant alignments with MAST in query genomic sequences. Data were first collected on the same set of genomes used previously and later on the entire set of genomes available through NCBI's RefSeq database, excluding genomes smaller than 2.5 Mbp in size. These reduced genomes belong largely to obligate parasites and symbionts, which are known to contain a reduced number of genes; as the number of genes decreases, the number of TF families appears to decrease quadratically and thus these organisms do not have a sufficient number of TFs to obtain meaningful results [38]. All genomes were grouped into one of five taxonomic categories: Enterobacteria, Gammaproteobacteria, Proteobacteria, all other Bacteria, and Archaea. The taxonomy was determined using NCBI's taxonomic database. Python scripts are available at <https://github.com/marc-delgrande/masters_thesis>.

The MEME/MAST software packages required first-order input hidden Markov models (HMM) for each query genome; these were generated using the previously developed software package `fasta-get-markov` [39] and incorporated into the pipeline. First-order HMMs provide information regarding individual nucleotide as well as dinucleotide frequencies.

2.1.4 Statistics and Graphics

All graphics and statistics were produced using R, a statistical analysis package with ggplot2 libraries [40, 41]. A Shapiro-Wilks and F-test were used to determine test for normality and equal variance respectively. Where assumptions were met, the student t-test and ANOVA were used to test the null hypothesis that the average proportion of binding sites is equal between genomes. For non-parametric data, the

Wilcoxon rank-sum and Kruskal-Wallis tests were used [42].

2.2 Results and Discussion

2.2.1 Patser results reveal the need for a more specific search tool

A first analysis of the taxonomic distribution of potential operators with Patser (Figure 2.2) revealed no substantial difference in the proportion of predicted operators except in those genomes with a high GC content. *Streptomyces coelicolor, Xanthomonas campestris,* and *Methylococcus capsulatus* all have genomic GC contents $> 63\%$, compared to 50.8% in *E. coli* and 33% in all the operators recognized by global TFs combined.

Patser returns a potential operator when it aligns with a sequence within the query genome that increases the IC of the weight matrix. It follows then that most alignments in genomes with high GC contents would fail to be reported as potential operators when compared to a weight matrix of low GC with Patser. Indeed this is an acceptable observation, however Patser is unable to distinguish possible differences between genomes with similar GC, demonstrated by heavy overlapping in Figure 2.2.

When calculating IC (and thus the significance of an alignment), Patser's algorithm assumes an independent, multinomial distribution of nucleotides [32, 33]. This means that two genomes with equal GC content could return the same number of potential operators regardless of taxonomic relationship, thereby not truly evaluating if the differences in frequency of potential operators has any taxonomic relevance.

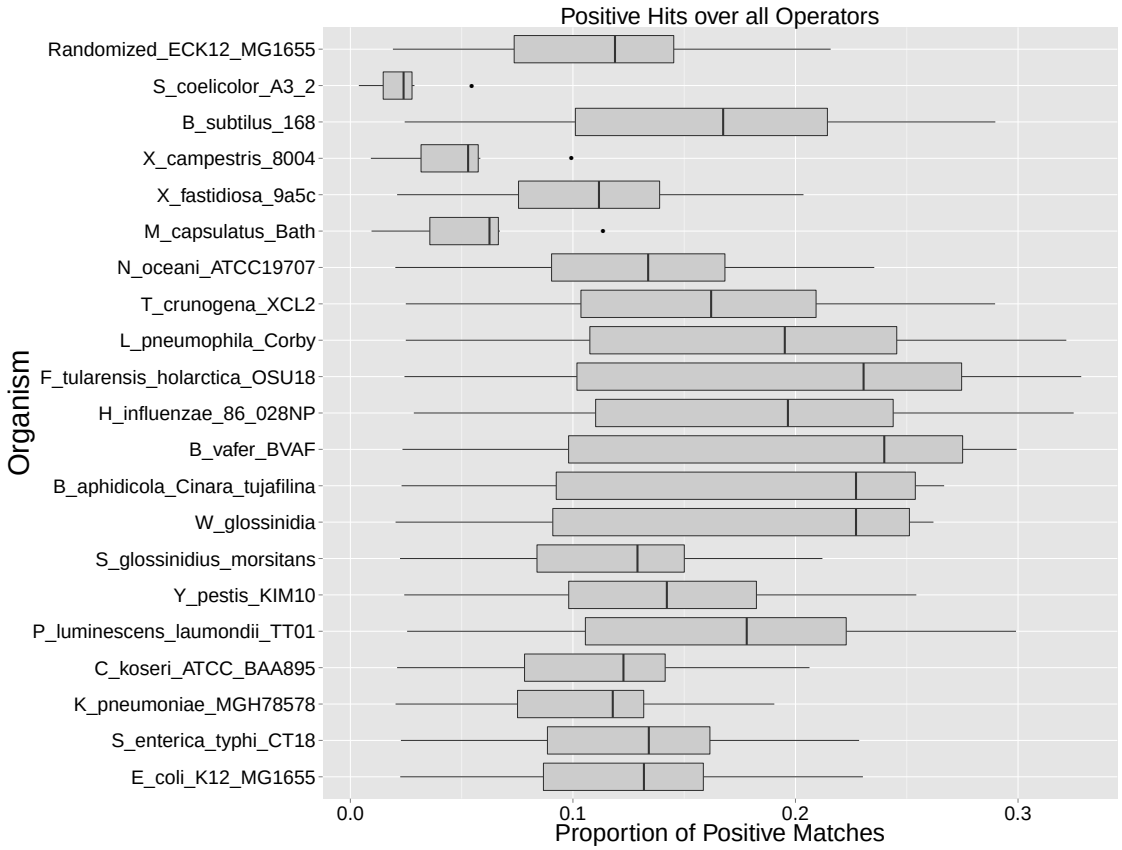


Figure 2.2: Propotion of Patser positive alignments for operators bound by global TFs. Each box plot represents 7 data points (one for each global TF in *E. coli*). Number of matches normalized to number of base pairs in the query genome to yield the proportion of positive matches.

2.2.2 Motif Alignment and Search Tool shows high rates of false operators in *Escherichia coli*

For the reasons mentioned above, this analysis was conducted again with MAST, a more statistically robust alignment tool. MAST contrasts with Patser in two primary ways: (i) in calculating the p-value of an alignment and (ii) MAST considers not only the *a priori* nucleotide frequencies, but a HMM trained to the search sequence as well.

As previously mentioned, Patser returns a positive match if the alignment increases the IC of the weight matrix. The IC is used to calculate a p-value which is used to determine the significance of the alignment [32]. The most significant problem with this method is that there are no standard procedures for calculating a p-value from IC [33]. Further, considering only independent nucleotide frequencies does not provide a detailed enough description of a genome. The predicted order of nucleotides has, in recent years, been considered a very important descriptor between genomes with otherwise similar nucleotide frequencies [43].

MAST uses a combined approach when calculating the significance of an alignment. First by calculating the significance of the weight matrix using the *a priori* HMM as the null, and second, when a matching sequence is found, by calculating the probability that a random sequence of the same length would match just as well or better (see Algorithm section in [36]). This combined approach measures both the significance of the weight matrix as well as the matching sequence. These two p-values are combined using established methods to determine the significance of the alignment [36].

By using HMMs to determine the significance of an alignment, MAST does not assume a multinomial distribution of nucleotides. Instead, the HMM provided a more descriptive breakdown of a genome by also evaluating the relative frequency of dinucleotide sequences. It was reasoned that, by using HMMs in combination with more robust techniques for determining significant alignments, differences in the frequencies of predicted operators could be more appropriately evaluated between

species with similar GC content allowing for elucidation of possible taxonomically relevant differences.

MAST results for the analysis of the same groups of genomes and operator sets as conducted with Patser are illustrated in Figure 2.3. The numbers of potential operators returned by MAST were several orders of magnitude less than those returned by Patser so the matches were normalized to the number of open reading frames in the search genome (data collected from NCBI by April 2013) rather than the number of base pairs. This analysis revealed a possible trend that is illustrated by the apparent drop in significant matches as the organisms become less taxonomically related to *E. coli*.

The same analysis was extended further to a subset of NCBI's Prokaryotic, RefSeq genomes (see section 2.1.3). Figure 2.4a shows the results for operator motifs recognized by *E. coli*'s 7 global TFs (outliers defined as greater or less than 1.5 times the interquartile ranges indicated by the whiskers; outliers not shown). A Kruskal-Wallis multiple comparison (KMC) test confirmed that the median number of positive matches identified in the enterobacteria was greater than that of any other group ($\chi^2 = 2642.243$, $p < 2.2 \times 10^{-16}$). Similar results are reported for operator motifs bound by 42 of *E. coli*'s local TFs, shown in Figure 2.4b ($\chi^2 = 6534.329$, $p < 2.2 \times 10^{-16}$).

These data indicate that the incidence of false positives is more frequent within the Enterobacteria, the group to which the regulators that recognize these operators are typically native, than other groups. This is consistent with the idea that operator sequences are not markedly unique when compared to their host genome sequence. There is enough information present in operator sets to distinguish them from the sequence of distantly related organisms, however when compared to their host's or a closely related organism's genome, the IC is low enough that the sites occur often by chance.

If operators were highly unique compared to the host genome or possessed sufficient IC, false positives would not be observed at such high rates. Further, the rate of false positives would be relatively consistent across most taxa. As a result, this

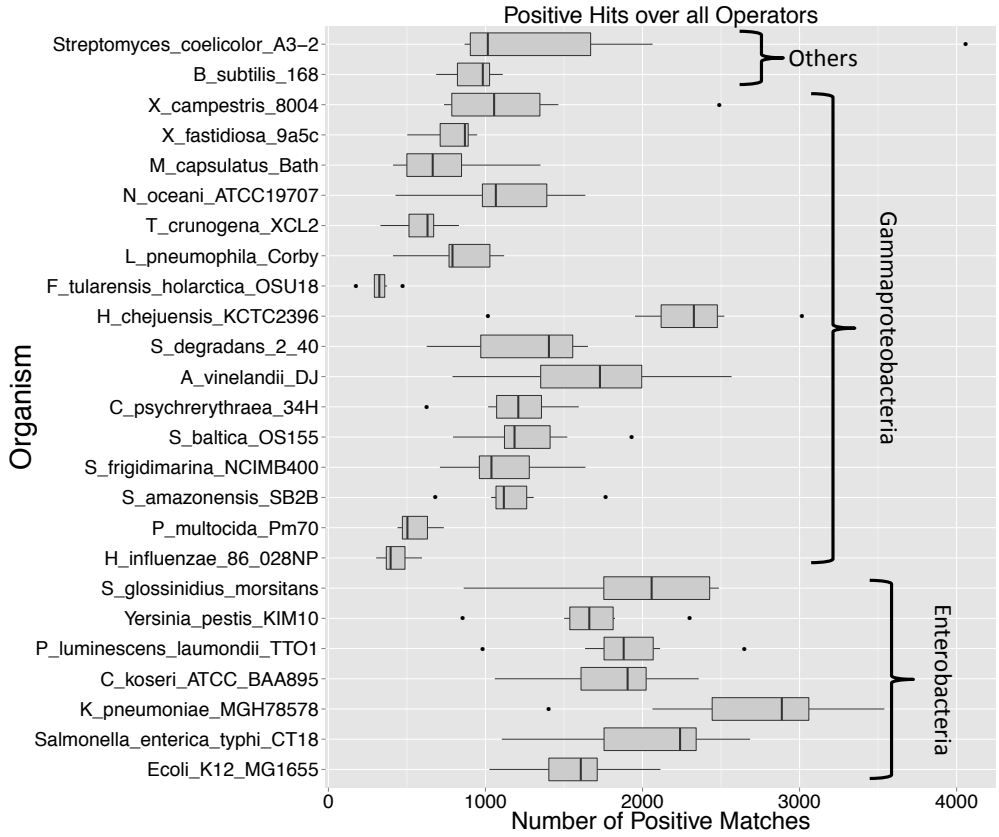
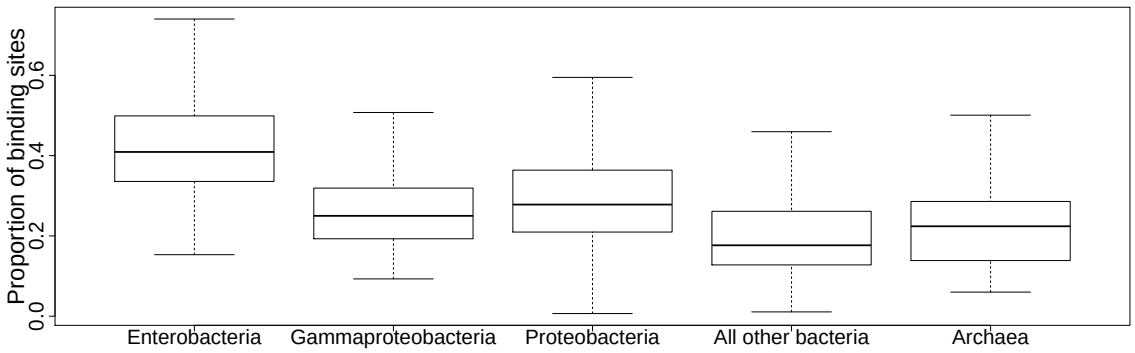
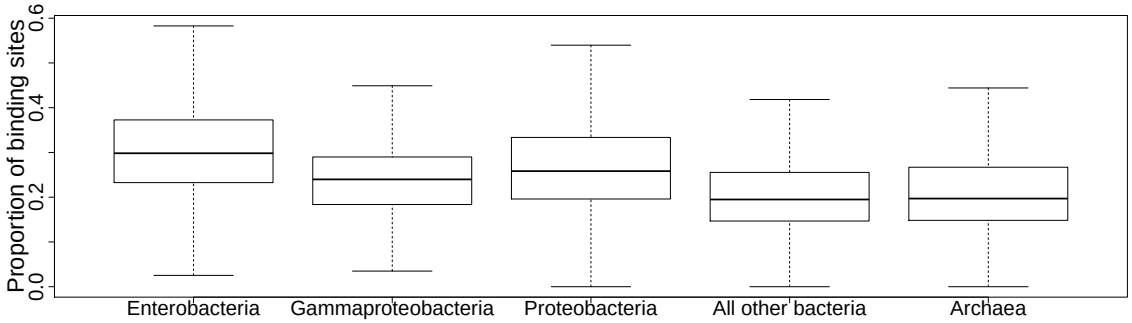


Figure 2.3: Proportion of MAST positive alignments for operators bound by global TFs. Each box plot is composed of 7 data points (one for each global TF in *E. coli*). Number of matches normalized to number of identified genes in the query genome. Organisms identified as, "reduced genomes" were not included in any further analyses (see Materials and Methods).



(a) Global operator motifs



(b) Local operator motifs

Figure 2.4: MAST results for operator motifs in RefSeq genomes. (a) Results for global operator motifs, (b) results for local operator motifs. KMC test confirmed that the number of potential binding sites identified in Enterobacteria is significantly higher than all other categories. Outliers not shown.

may indicate that high incidences of potential operators do not serve as a disadvantage in prokaryotes. They may facilitate an organism's ability to rapidly reorganize its regulons and thus adapt to changing environments.

2.3 Conclusion

A comparison of MAST and Patser as tools for motif search revealed that the use of HMMs and robust calculations of p-value significantly influence the number of possible results by, in some cases, several orders of magnitude. While identifying the frequency of false positives was the goal of this study, Patser was not specific enough in its criteria for identifying matching sequences and provided too many results to draw meaningful conclusions. The use of MAST provided increased specificity and allowed for a more accurate estimation of the frequency of false positives.

Using sets of operators from *E. coli*, the frequency of possible matches was much higher within the Enterobacteria than all other taxonomic groups indicated here. Since the increase in the frequency of matches was so large, especially for operators recognized by global operators, it is unlikely that the increase in matches was due to true positive binding sites alone; recall the number of matches was normalized to the number of open reading frames, thus a 1% increase in frequency represents, on average, an additional 40-50 matches in the Enterobacteria.

This high rate of false positives could be partially due to the low IC of operators. With low IC, the possibility of a binding site occurring by chance is very high. It could be that maintaining operators with non-unique sequences provides a mechanism by which regulons are able to rewire rapidly. A trait that can be advantageous upon encountering new environmental pressures.

2.3.1 Recommendations

Further work into developing motif finding software is still needed as models to describe genetic sequences are continuously being improved [18]. Strong evidence for determining taxonomic relationships based on HMMs would further substantiate the

conclusions made here regarding results from the MAST software. Current research [43] shows that there is some taxonomic relevance to simple HMMs, but more work is still required in this area as it is not clear what mechanisms primarily dictate higher-order HMMs.

This work has been conducted using operator sequence data from *E. coli* and has been limited by the currently available operator sequence data. To further substantiate these conclusions, this work should be repeated from the perspective of many other organisms. This, however, is limited by the rate of experimental determination of operator sequences. Currently it may be possible to echo this work from the *B. subtilis* perspective, however far fewer operator datasets would be available. Extending this analysis as-is to eukaryotes may prove to be unrealistic without considering the architecture of the Eukaryotic genome (euchromatin versus heterochromatin) as well as the many different classes and actions of TFs not witnessed in prokaryotes. Further, the size of a Eukaryotic genome would present additional computational challenges in terms of time spent on the analysis as well as time spent parsing through the large amounts of data that would be returned for meaningful information.

3 Loose evolutionary relationship between transcription factors and other gene products

It has been suggested that nearly half of the TFs in *Escherichia coli* have been recently acquired via horizontal gene transfer (HGT) [9]. It has also been suggested that the network of genes regulated by any TF in *E. coli* and other well-characterized organisms is both flexible and subject to rapid evolution [5, 6, 7]. Using curated and predicted TFs, these claims are assessed here across a large, diverse set of prokaryotes.

3.1 Materials and Methods

3.1.1 Phylogenetic profiles and mutual information

Phylogenetic profiles (PP) and mutual information (MI) scores were previously constructed and calculated as described by Moreno and Jokic [20]. The PPs used a non-redundant subset of genomes ($GSSa = 0.90$) [44, 45] available through NCBI's RefSeq database [29, 30, 31] by the beginning of June 2013 totalling 920 genomes. Orthologs were determined as reciprocal best hits (RBH) as described in [46]. The detection of a RBH was indicated with a 1 and the absence with a 0.

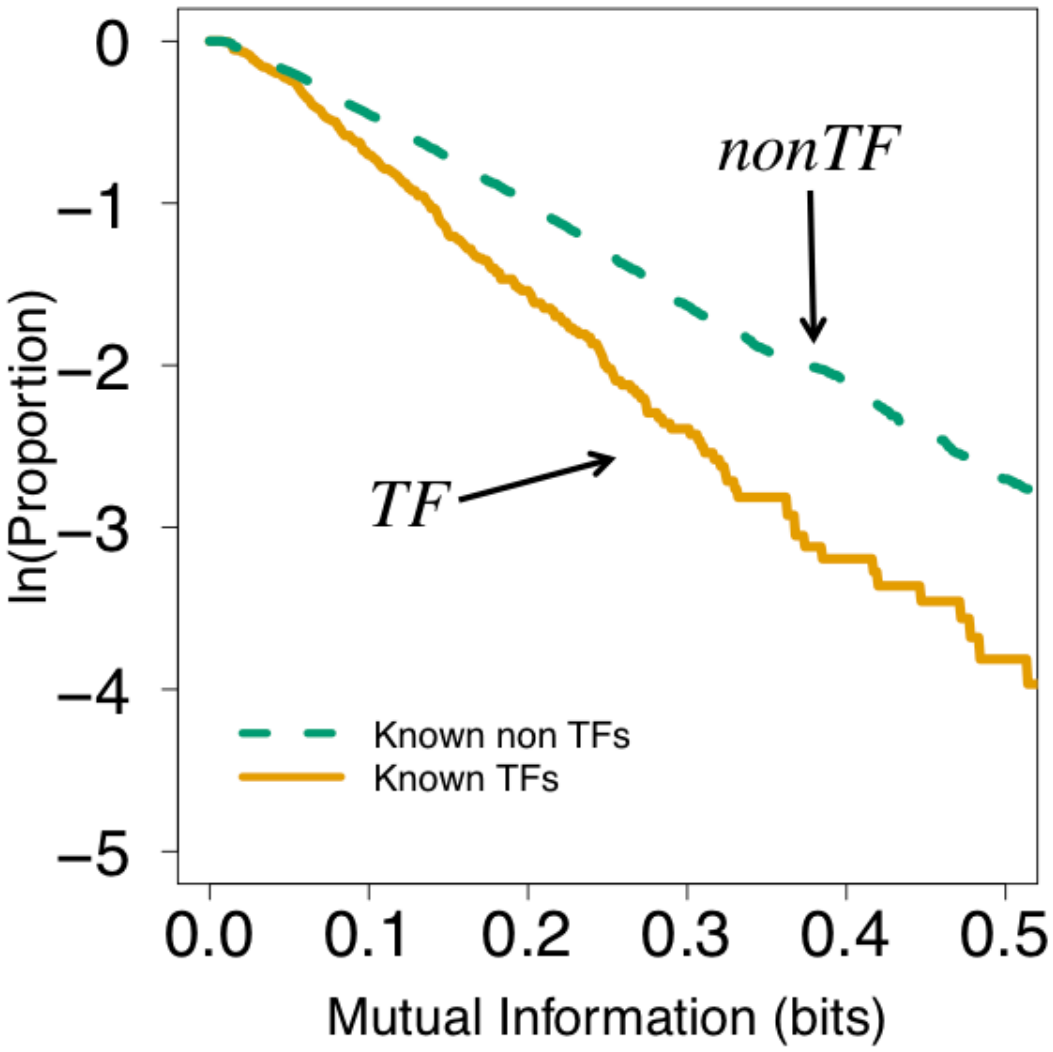
MI was used as a measure of similarity between the PP of a pair of genes [27, 44]. Pairs of genes that co-occur often have higher MI scores than pairs that seem to appear independent of one another. It has been shown that gene pairs with high MI scores tend to be those involved in some form of functional interaction [20, 27, 44]. The MI data was filtered so that only the top scoring gene pair for each unique gene remained.

3.1.2 Profiles of phylogenetic profiles, the p-cubic analysis

Profiles of phylogenetic profiles (p-cubic) were constructed from the remaining MI data for organisms with ≥ 100 identified TFs (see 3.1.5), reducing the set of genomes to 597. In short, a p-cubic is an inverse cumulative representation of the relative abundance of genes with MI scores above a given threshold (see Figure 3.3 for examples). As the MI score threshold increases along the x-axis the curve drops, indicating fewer gene pairs with the given MI score or higher. MI scores were incremented (or "binned") from $[0 - 1]$ by 0.001. Two or more p-cubic curves can be compared: the position of one curve relative to another indicates the abundance of high or low scoring gene pairs. Curves that lie low and drop quickly are those groups in which there was a lower abundance of high MI scores for the gene pairs. For additional details on p-cubic analysis see Figure 2 in [20].

3.1.3 P-cubic differences, the $\Delta P3$

To evaluate the difference in p-cubic between TF-coding and non TF-coding gene interaction pairs for each organism, the sum of the difference of the y values for each curve were calculated and these values were divided by the number of bins to yield a difference of p-cubics ($\Delta P3$); see Figure 3.1. For p-cubics in which the curve representing genes coding for TFs fell below the curve for all other genes, indicative of regulatory genes forming looser associations than all other genes, $\Delta P3$ was > 0. Values of $\Delta P3 \leq 0$ indicated either equal association strength ($\Delta P3 = 0$) or regulatory genes forming stronger associations ($\Delta P3 < 0$). After filtering, this analysis was performed on 597 prokaryotes, including some Archaea.



$$\frac{1}{n}\left(\sum_{i=1}^{n} nonTF_i - TF_i\right) = \Delta P3$$

Figure 3.1: Schematic for calculating ΔP3 Values. The sum of the difference of y values between TF interactions and non-TF interactions is normalized to the number of MI threshold increments (bins) used to construct the p-cubics. Arrows on the graph indicate for which curve the $nonTF_i$ and TF_i values are derived from.

3.1.4 Transcription factor identification

Verified sets of TFs for *E. coli* str. K12 substr. MG1655 and *Bacillus subtilis* substr. 168 were downloaded from RegulonDB [1, 10] and the Database of Transcription regulation in *Bacillus subtilis* (DBTBS) [47] respectively. The gene IDs from the verified lists were used to distinguish TFs from all other genes in the MI content datasets. These data were used to benchmark analyses on predicted TF sets.

Lists of TFs were also collected for *E. coli* and *B. subtilis* from the HAMAP (High-quality Automated and Manual Annotation of Proteins) database [48]. The HAMAP database, as its name suggests, uses a combination of manual and automated curation while also maintaining cross-references to more than 60 other databases. Using these cross-references, an identifier from the Gene Ontology (GO) database [49] was used to extract proteins that had been submitted to HAMAP and annotated as DNA-dependent transcriptional regulators (GO:0006355). Since the time of writing, the GO database has undertaken an initiative to completely redesign their methods for annotating TFs and has begun to phase out the identifier used here [50, 51]. For this reason, GO identifiers were not used beyond *E. coli* and *B. subtilis*.

3.1.5 Transcription factor prediction

Transcription factor prediction for the same non-redundant set of genomes as mentioned previously was conducted using 147 hidden Markov models (HMMs) from the Protein Families database (Pfam) [52] downloaded from DNA-binding domain database (DBD), a comprehensive and accurate database of predicted TFs [53]. Using a combination of `hmmer` [54] and `hmmfetch` [52]processed together using PYTHON [34], the HMMs were compared against all annotated proteins for each genome with the resulting datasets filtered for organisms with ≥ 100 matches of $\geq 80\%$ coverage of a HMM, reducing the dataset from 920 to 597 organisms. In recent work [55], a cutoff of $\geq 60\%$ was used when matching to Pfam HMMs; it was reasoned however, that this allowed for too many false-positives so the increased 80% cutoff was used.

3.1.6 Codon adaptation index

As a representative set of highly expressed and highly conserved genes, ribosomal-protein (r-protein) coding genes were used to construct codon usage tables (CUTs) for each genome. R-protein coding regions were found for each organism available in NCBI's RefSeq database [29, 30, 31] using COG and arCOG identifiers [56] provided by Yutin *et al.* [57], available at <`ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/`>. Reconstructions of phylogenetic trees using r-protein sequences produces trees that match closely with commonly accepted taxonomy [30, 57], thus horizontal transfer of the r-proteins used here is likely a rare event. This makes r-protein coding genes a reliable source for constructing CUTs.

EMBOSS packages `cusp` and `cai` [58] were used to construct codon usage tables derived from r-protein coding genes for each organism and to compare the codon usage of other protein-coding regions to those of r-protein coding genes. From this comparison, a codon adaptation index (CAI) is calculated for each protein-coding region. A low CAI is indicative of possible horizontal gene transfer (HGT) [59].

A normalized difference in the CAI of genes coding for TFs was calculated as the difference in average CAI for TF genes and average CAI for all genes, normalized to the average CAI for all genes:

$$\frac{\overline{x}_{TF} - \overline{x}_{G}}{\overline{x}_{G}} \tag{3.1}$$

Where $\overline{x}_{TF}$ is the average CAI for the set of genes coding TFs, the TF CAI, and $\overline{x}_{G}$ is the average CAI for the set of all genes, the genomic CAI.

3.2 Results and Discussion

3.2.1 Protein family hidden Markov models are adequate for predicting transcription factors

To benchmark the methods used for identifying TFs across all prokaryotes, a 3-way comparison was carried out between verified TFs, HAMAP identified TFs, and the

TFs recovered from Pfam HMMs in *E. coli* and *B. subtilis* (Figure 3.2). In *E. coli* there were only 7 false positives, 3 of which are considered TFs by HAMAP. In *B. subtilis* there were 43 false positives, 32 of which have been identified as TFs as indicated by HAMAP.

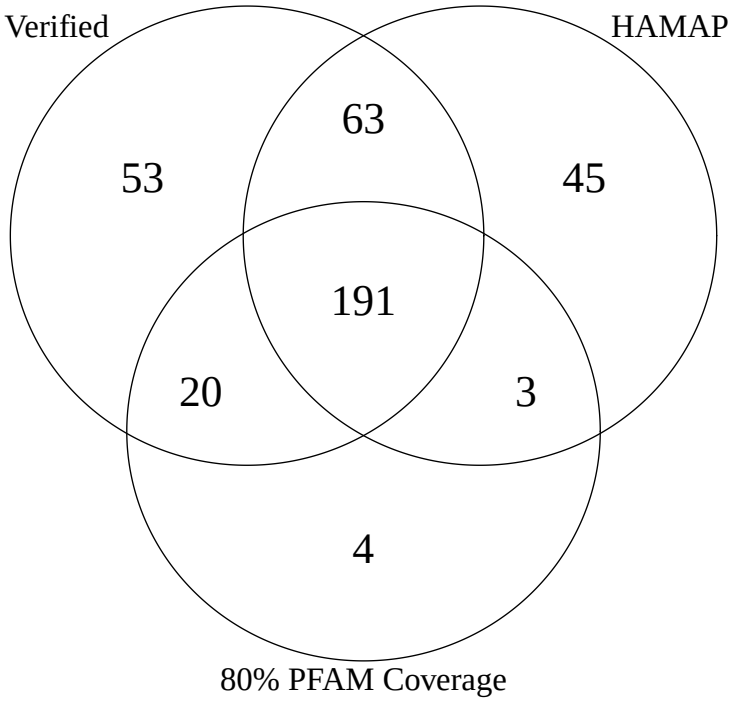
In *E. coli* only 4 of the 7 false positives were true false positives, one of which belongs to a family of proteins involved in TR (GI: 16129176; SirB family), however the role of this particular protein has not been definitively determined. In*B. subtilis*, only 5 of the 43 were true false positives. Of these, 3 are cold shock proteins and though they have not been experimentally implicated in TR, it is known that this family of cold shock proteins is involved in regulating gene expression [30]. These cold shock proteins also each contain a DNA binding site implying they may be DNA-binding transcriptional regulators. Tables A.1 and A.2 outline information collected from NCBI regarding the false positives, including those indicated as TFs by HAMAP. Figure A.1 illustrates the decision tree used to determine a true TF.

Though this method of predicting TFs returns false positives, it appears that many of these are not true false positives. For this reason, the TFs recovered by Pfam HMMs were determined robust enough to use for the remainder of this analysis.

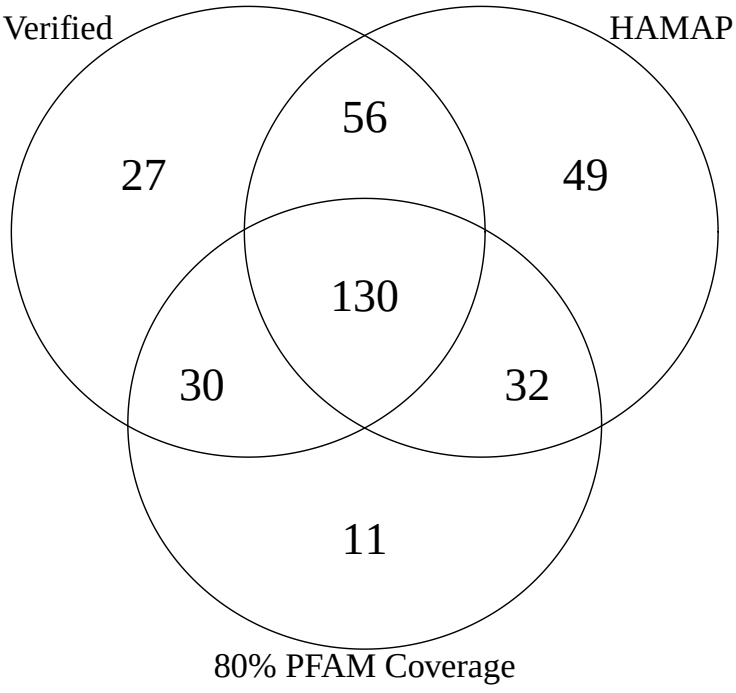
3.2.2 P-cubics show that transcription factors have less conserved interactions than other genes

Upon identifying TFs in 597 prokaryotes, it would have then been ideal to assess the p-cubic for the set of genes that are primarily transcribed by these TFs for each organism. Such a task may have been possible with a very small number of well-characterized organisms, however automating such a job over a diverse group was not possible within a reasonable time frame. Instead, p-cubics for the most conserved associations among TFs, as determined by MI content, were used in place of a TF's experimentally identified primary target gene. These p-cubics were compared with the most conserved associations for all other gene pairs for each respective organism to evaluate the conservation of TF associations.

P-cubics for verified TFs in *E. coli* and *B. subtilis* (Figures 3.3a and 3.3c) show



(a) *E. coli*



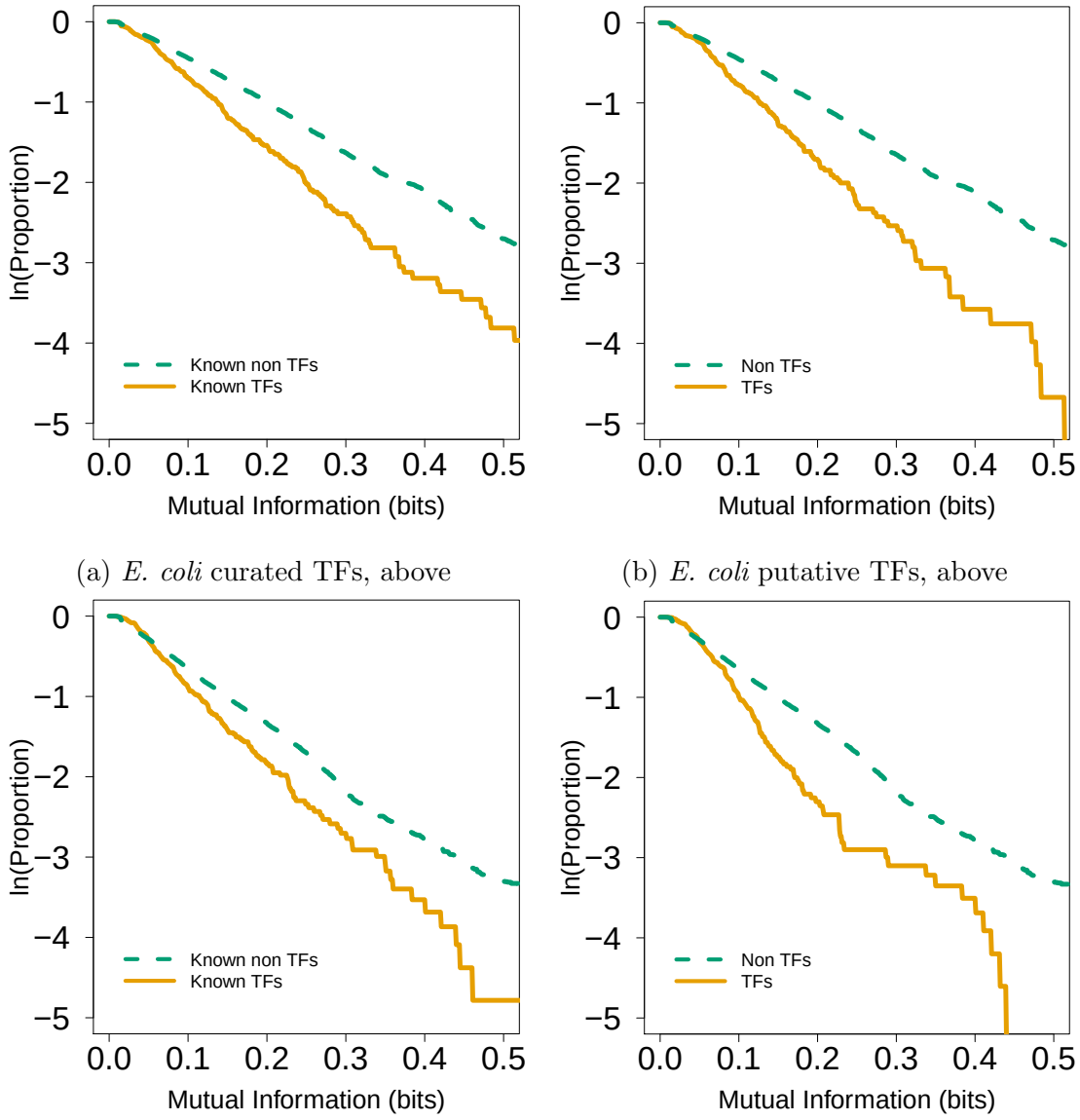
(b) *B. subtilis*

Figure 3.2: TFs identified by manual curation (verified), semi-automated curation (HAMAP), and Pfam HMMs (80% PFAM coverage) in *E. coli* (a) and *B. subtilis* (b). Though each dataset has discrepancies, there is a large intersection in datasets for both organisms.

that the conservation of associations within the these TFs are consistently less than that of all other gene pairs. P-cubics for TFs identified by Pfam HMMs (Figures 3.3b and 3.3d) also produced similar results. In each case, the results show that associations formed by TFs are not as well conserved as those formed by all non TFs. This is reflected in the $\Delta P3$ values: 0.859 and 0.766 for *E. coli* and 0.486 and 0.511 for *B. subtilis*, known and predicted TFs respectively. Further, though the datasets for predicted TFs do not completely match those of the literature-based sets, the verified and predicted datasets agree with each other, demonstrating that the predictions still provide with adequate data to test the conservation of associations for TFs.

The cumulative proportion plot in Figure 3.4 summarizes the $\Delta P3$ values collected from each organism's p-cubic analysis with the predicted TF datasets. Almost all organisms examined here exhibited p-cubics in which genes for TFs tended to form less conserved associations than those formed between all non-TF gene pairs (97.8% with $\Delta P3 > 0$).

Other works have suggested that regulatory networks evolve rapidly [5, 6, 7, 9], however these inferences were made based on observations from a small number of organisms by comparing only well characterized networks. The results here present an overwhelming absence of co-occurring, conserved interactions between TFs and their possible target genes across a large number of prokaryotes, demonstrating that these interactions are indeed rapidly evolving across most, if not all, prokaryotes.



(a) *E. coli* curated TFs, above

(b) *E. coli* putative TFs, above

(c) *B. subtilis* curated TFs, above

(d) *B. subtilis* putative TFs, above

Figure 3.3: Comparison of p-cubics for curated (RegulonDB/DBTBS) and predicted (Pfam HMM) TFs in *E. coli* (a, c) and *B. subtilis* (b, d). In both the curated and predicted sets the p-cubic for TFs falls below that of non TFs, showing that TFs have less conserved co-occurrences than non TFs. Since the predicted datasets produce similar results as the curated sets, predicted TFs may be adequate for analyzing the p-cubic for TFs versus non TFs in other prokaryotes.

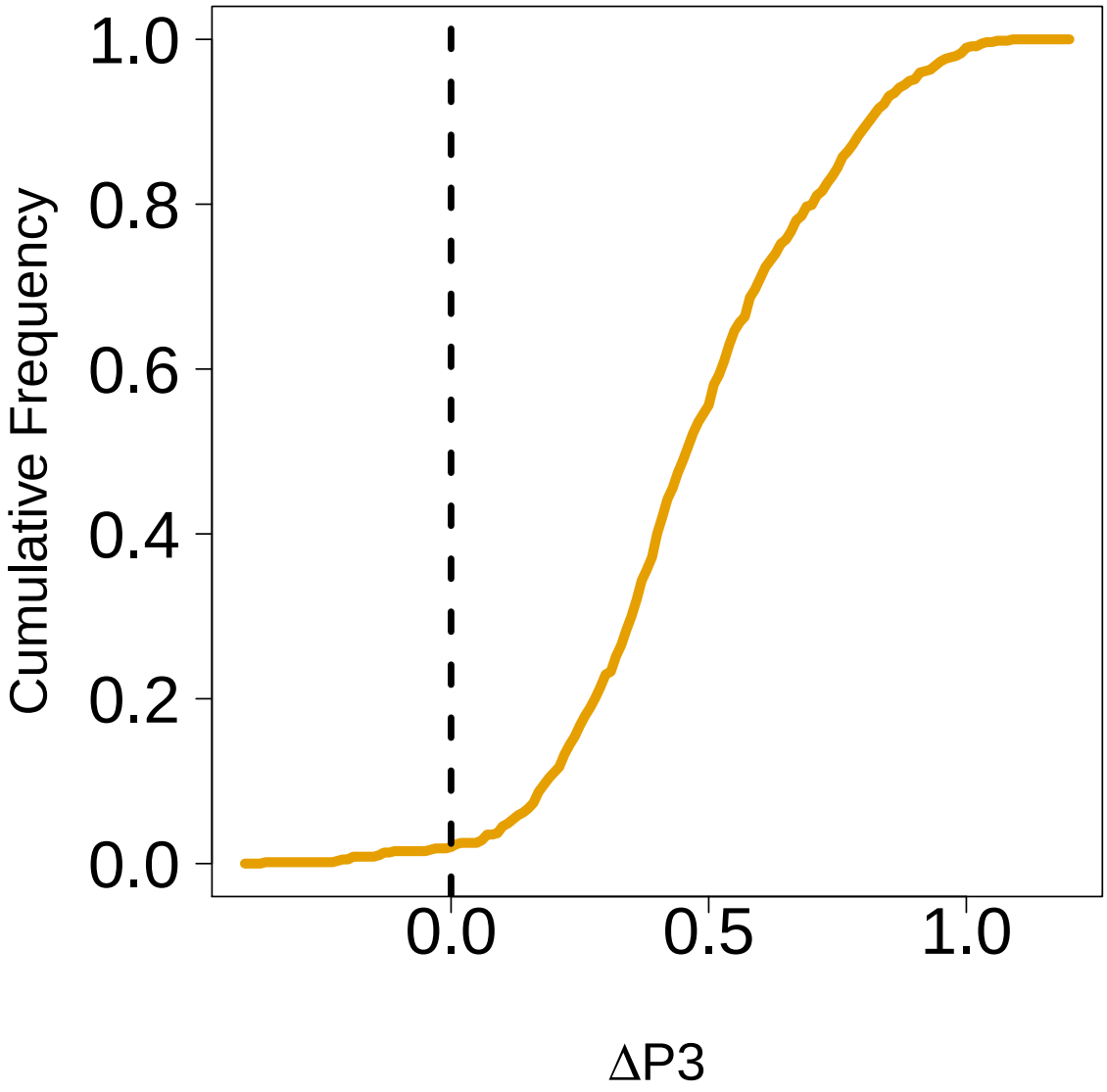


Figure 3.4: Cumulative proportion of $\Delta P3$ between TF-coding genes and non TF-coding genes across prokaryotic genomes with ≥ 100 predicted TF-coding genes. The majority of $\Delta P3$s are > 0 indicating a lack of conserved co-occurring interactions for TFs.

3.2.3 Low codon adaptation indices suggest frequent horizontal gene transfer

Previous work has suggested that half of all TFs in *E. coli* may have arisen via HGT into the γ-Proteobacteria lineage [9]. This frequent horizontal inheritance may likely be contributing to the rapid evolution of associations involving TFs. To test the possibility of this being true for other prokaryotes, the CAI was calculated for all the genes of the genomes used in the p-cubic analyses. Figure 3.5 shows the cumulative frequency of the normalized difference for the TF CAI (see section 3.1.6, equation 3.1). In $> 97\%$ of organisms, genes coding for TFs had a lower CAI than other genes. After removing the 77 organisms for which the set of TF CAIs were not significantly different than the genomic CAIs (Wilcoxon Rank-Sum test $p \geq 0.05$), $> 98\%$ had TF-coding genes with lower CAIs than other genes.

Of the 77 organisms where there was no notable difference between TF CAI and genomic CAI, they did not belong to a particular taxonomic group, nor did they identify with a particular environment (e.g. soil, aquatic, etc.). It could be that the environments these organisms exist in do not experience perturbations as extreme as others, thus there is a decreased need for HGT as a source of relatively quick genetic variation. It could also be that these organisms do not rely on HGT and instead have evolved another strategy for maintaining adequate genetic variation under changing environmental pressures.

Though assessing CAI alone is not adequate for determining HGT [60], these data show that genes for TFs in nearly all studied organisms have skewed codon usages. This result is an indication that these genes could be frequently involved in HGT across all prokaryotes.

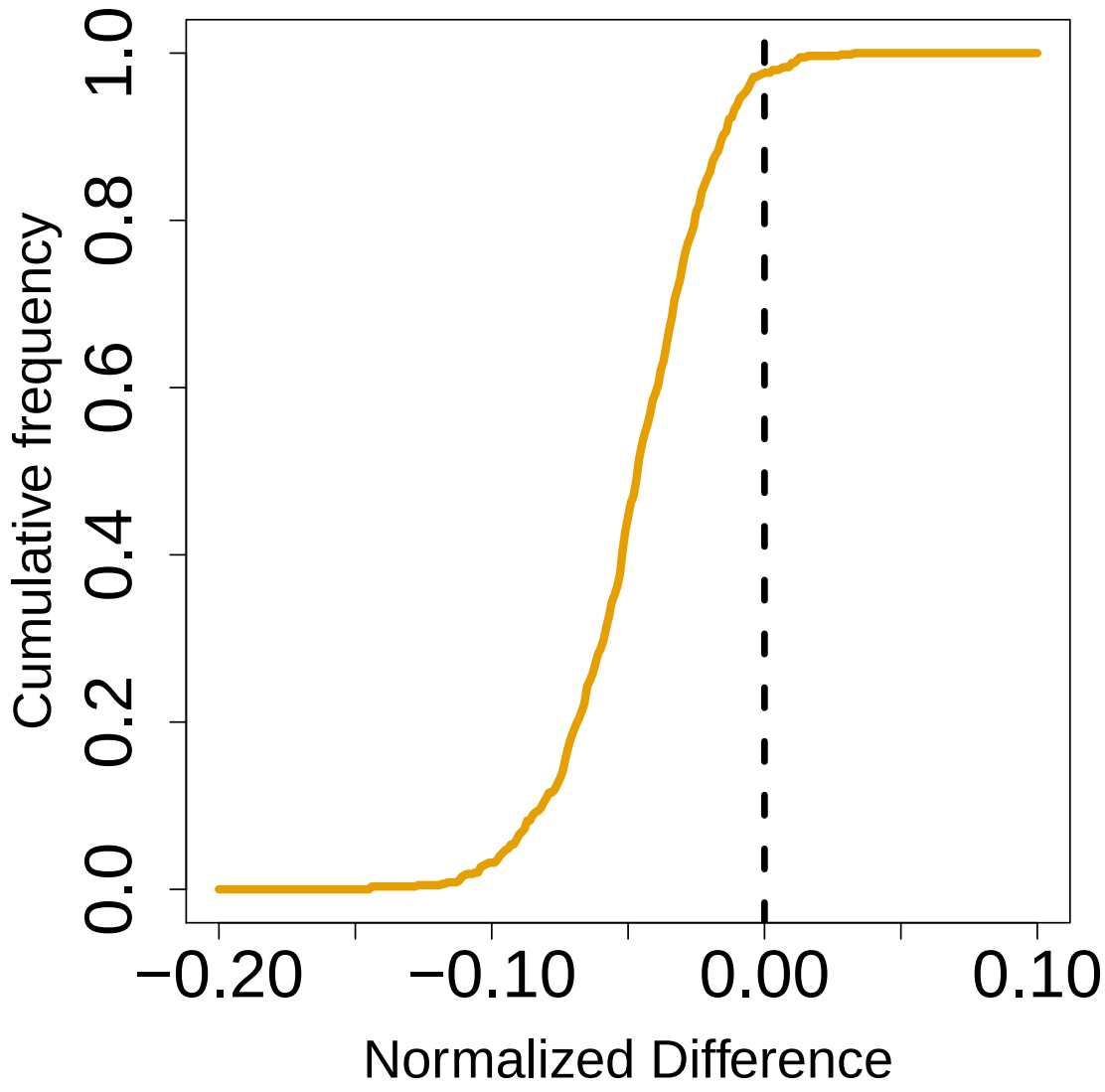


Figure 3.5: Comparison of the TF CAI and genomic CAI. The strong majority of TF-coding genes have a CAI less than that of non TF-coding genes indicating possible frequent HGT that is ubiquitous to TFs across prokaryotes.

3.3 Conclusion

The use of HMMs from Pfam [52] have provided an appropriate method for predicting TFs in prokaryotes. Of the results that did not agree with the curated sets, most were confirmed to be true TFs via manual referencing to NCBI [30]. Further, the p-cubic analyses performed with the putative TFs returned similar results as those performed with the curated sets.

P-cubic analyses revealed that the associations formed by TFs are not as conserved as those formed by all other gene pairs. This result could be due, in part, to frequent horizontal transfer of genes that code for TFs. Comparing the TF CAI to the genomic CAI showed that TF-coding genes often have a skewed codon usage, suggesting that these genes have been subject to recent horizontal transfer.

It has been shown that changes in regulation have led to major phenotypic changes in both eukaryotes [4] and prokaryotes [6, 61]. Thus, understanding the evolutionary dynamics of TFs will provide insight to understanding variation between species [20]. Combined, these results provide new indications that rapid evolution of regulatory networks is a property maintained in possibly all prokaryotes.

That this property of TRNs appears ubiquitous to all sequenced prokaryotes is of significant importance as it likely evolved early in at least this group of organisms. Altering the way genes are accessed is a newly discovered source of variation. This is in contrast with the traditional way that variation is referred to in the evolutionary context, which involves altering the genes, and presents as a newly discovered evolutionary strategy.

3.3.1 Recommendations

The development of a high quality, publicly accessible, and frequently updated set of HMMs that accurately predict TFs will be crucial for future research interested in targeting TFs across a large number of diverse species. Incorporating datasets from recent works [62, 55], existing databases, such as Pfam [52] and SUPERFAMILY [63], and curated sources [1, 10, 47] will be required to construct such a resource. Once

collected, these reference HMMs should be able to facilitate the accurate prediction of TFs in any organism.

With existing datasets and acceptably predicted TFs, a survey on the proportion of TFs that show evidence of recent horizontal transfer would offer greater insight into the possibility of HGT being the cause of loosely conserved interactions between TFs and other gene products. Past works [60, 64] have focused on developing comprehensive methods for detecting genomic islands (portions of genetic sequence that have recently been horizontally inherited [64]). By targeting genes that code for TFs and using these tools to determine evidence of horizontal transfer, it will be possible to predict the proportion of TF-coding genes that have been subject to recent HGT. Further, the proportion of genomic islands that carry TFs with them should be assessed. This may provide some insight into how new genes, and new TFs integrate into a TRN.

4 Summary

By comparing transcriptional regulatory networks between *Escherichia coli, Bacillus subtilis*, and organisms closely related to the two, it has been suggested that these networks evolve rapidly. Within these model organisms, rebuilding such networks is a daunting task both computationally and experimentally; it is currently not feasible to rebuild them for a large, diverse set of organisms. This limits the conclusions that can be drawn regarding the flexibility of transcriptional regulatory networks across all prokaryotes. Further, the contribution of operators to the potential for rapid evolution of these networks has not been substantially explored.

By analyzing the individual components of a transcriptional regulatory network: transcription factors, target genes, and operators, it is possible to assess the evolutionary dynamics of these networks without reconstructing them. If the components of a network appear flexible or to be rapidly evolving, then the entire network itself must be rapidly evolving. This work attempts to assess the evolutionary stability of the components of transcriptional regulatory networks in order to conclude on the stability of them across all prokaryotes.

The rate of false positives for potential operators is assessed here from the *E. coli* perspective by using two different motif alignment softwares on experimentally determined *E. coli* operator sequences and searching available prokaryotic genomes in NCBI's RefSeq database. Initially the motif search tool, Patser, returned an overwhelming number of positive potential operators in each organism for all sets of operators used. Though the purpose of this analysis was to determine a rate of false positives, the results provided by Patser were too non-specific and so meaningful conclusions could not be drawn. Using a new software package, Motif Alignment and Search Tool, which uses a more strict definition for identifying significant alignments, it was possible to assess the frequency of potential operators. It appears that

E. coli operator sequences are particularly noisy within *E. coli* and closely related species. This presents as a possible mechanism for rapid reshuffling of regulons (the set of genes regulated by a transcription factor) and may facilitate the frequent, spontaneous formation of operators within a genome.

The number of operator sequences available for both *E. coli* and other organisms limits this work to only focusing on a few operator sets within one organism. As more operator sequence data becomes available for other organisms, this procedure can be repeated from the perspective of many other species to further assess these claims. *B. subtilis* is likely the next organism to have enough data on operator sequences to repeat this analysis. This work is also currently biased towards culturable organisms, as these are the primary types of organisms that can be sequenced with enough depth to be classified as reference sequences in NCBI's database. This limits the types of species analyzed to primarily ones of significant medical, industrial, or agricultural importance. Difficult to culture organisms are usually recovered from environments that are impossible to recreate in the lab. The environmental stresses these species face may be drastically different than those of the species analyzed here. It will be important to determine how life in these environments affects TRN structures as these species' sequences become available in the future.

It was possible here to predict the conservation of co-occurrence between transcription factors and target genes across a diverse set of species. This was achieved by using protein domain hidden Markov models provided by the Protein Family database to predict genes that code for transcription factors. Putative transcription factors were compared to databases of those confirmed for *E. coli* and *B. subtilis* to verify the competency of the prediction methods. Very low rates of false positives were observed for transcription factor prediction in both organisms.

With predicted transcription factor datasets phylogenetic profiles were used to calculate mutual information between gene pairs, taking the most often co-occurring gene pairs as proxy for true transcription factor to target gene interactions. With these predicted interactions compiled, p-cubics comparing the evolutionary stability of interactions between transcription factor and target gene versus all other gene

pairs revealed that, in $> 97\%$ of organisms, interactions involving transcription factors are less evolutionarily stable than the average of all other interactions. The primary advantage to this analysis is that it allows for the assessment of the stability of transcription factor and target gene interactions without the need for computational reconstruction of TRNs and experimental determination for interactions of homologous gene products.

A possible reason for this lack of conserved interactions is that transcription factors may be involved in frequent horizontal gene transfer. This has been suggested by previous research in *E. coli*, but has not been definitively shown for other prokaryotes [9]. As a predictor of horizontal gene transfer, the codon adaptation index of predicted transcription factors of an organism was compared to that for all other genes. Indices were created using codon usage tables derived from ribosomal protein coding genes. It was observed that, for $> 98\%$ of genomes analyzed, the average codon adaptation index of transcription factor coding genes was significantly less than that for all other genes. Low codon adaptation indices, though not definitive evidence, are indicative of horizontal gene transfer. Though this is not decisive evidence, it certainly supports the idea that horizontal gene transfer may be frequently occurring with transcription factors.

Perhaps a reason for widespread horizontal transfer of transcription factors is that genomic islands which contain open reading frames require immediate regulation upon integration in order to be maintained. Thus, for most horizontal transfer events, a transcription factor must be present. This idea has not been explored in recent literature, but is certainly a testable hypothesis with current technologies.

Overall, it is shown here the operators, transcription factors, and target genes have loosely constrained evolutionary dynamics for most of the prokaryotes studied. Recent literature has suggested that TRNs evolve rapidly in a few model organisms and closely related ones. This work not only demonstrates that this is true for a larger set of prokaryotes, but presents a method to test for such dynamics without reconstructing TRNs, which is not only a computationally difficult task, but also requires substantial amounts of data from wet-lab experiments. Rapid reor-

ganization, or evolution, of a TRN presents a method for varying gene expression in the face of changing environmental pressures. That this property is observed in many prokaryotes means this may have evolved early on as a method for accessing variation beyond that involving only gene mutations.

References

[1] Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muiz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, Garca-Sotelo JS, Lpez-Fuentes A, Porrn-Sotelo L, Alquicira-Hernndez S, Medina-Rivera A, Martnez-Flores I, Alquicira-Hernndez K, Martnez-Adame R, Bonavides-Martnez C, Miranda-Ros J, Huerta AM, Mendoza-Vargas A, Collado-Torres L, Taboada B, Vega-Alvarado L, Olvera M, Olvera L, Grande R, Morett E, Collado-Vides J: **RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units)**. *Nucleic Acids Research* 2011, **39**(suppl 1):D98–D105.

[2] Leyn SA, Kazanov MD, Sernova NV, Ermakova EO, Novichkov PS, Rodionov DA: **Genomic Reconstruction of the Transcriptional Regulatory Network in Bacillus subtilis**. *Journal of bacteriology* 2013, **195**(11):2463–2473.

[3] Lavoie H, Hogues H, Mallick J, Sellam A, Nantel A, Whiteway M: **Evolutionary tinkering with conserved components of a transcriptional regulatory network**. *PLoS biology* 2010, **8**(3):e1000329.

[4] Rebeiz M, Jikomes N, Kassner VA, Carroll SB: **Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences**. *Proceedings of the National Academy of Sciences* 2011, **108**(25):10036–10043.

[5] Babu MM, Teichmann SA: **Evolution of transcription factors and the gene regulatory network in Escherichia coli**. *Nucleic Acids Research* 2003, **31**(4):1234–1244.

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

[7] Price MN, Dehal PS, Arkin AP: **Orthologous transcription factors in bacteria have different functions and regulate different genes**. *PLoS computational biology* 2007, **3**(9):e175.

[8] Martınez-Antonio A, Collado-Vides J: **Identifying global regulators in transcriptional regulatory networks in bacteria**. *Current opinion in microbiology* 2003, **6**(5):482–489.

[9] Price MN, Dehal PS, Arkin AP: **Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli**. *Genome Biol* 2008, **9**:R4.

[10] Huerta AM, Salgado H, Thieffry D, Collado-Vides J: **RegulonDB: a database on transcriptional regulation in Escherichia coli**. *Nucleic Acids Research* 1998, **26**:55–59.

[11] Novichkov PS, Laikova ON, Novichkova ES, Gelfand MS, Arkin AP, Dubchak I, Rodionov DA: **RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes**. *Nucleic Acids Research* 2010, **38**(suppl 1):D111–D118.

[12] Novichkov PS, Kazakov AE, Ravcheev DA, Leyn SA, Kovaleva GY, Sutormin RA, Kazanov MD, Riehl W, Arkin AP, Dubchak I, et al.: **RegPrecise 3.0–A resource for genome-scale exploration of transcriptional regulation in bacteria**. *BMC genomics* 2013, **14**:745.

[13] Jacob F, Monod J: **Genetic regulatory mechanisms in the synthesis of proteins**. *Journal of molecular biology* 1961, **3**(3):318–356.

[14] Sonenshein AL: **CodY, a global regulator of stationary phase and virulence in Gram-positive bacteria**. *Current opinion in microbiology* 2005, **8**(2):203–207.

[15] Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA: **Structure and evolution of transcriptional regulatory networks**. *Current opinion in structural biology* 2004, **14**(3):283–291.

[16] Isalan M, Lemerle C, Michalodimitrakis K, Horn C, Beltrao P, Raineri E, Garriga-Canut M, Serrano L: **Evolvability and hierarchy in rewired bacterial gene networks**. *Nature* 2008, **452**(7189):840–845.

[17] van Hijum SA, Medema MH, Kuipers OP: **Mechanisms and evolution of control logic in prokaryotic transcriptional regulation**. *Microbiology and Molecular Biology Reviews* 2009, **73**(3):481–509.

[18] Zia A, Moses AM: **Towards a theoretical understanding of false positives in DNA motif finding**. *BMC bioinformatics* 2012, **13**:151.

[19] D'haeseleer P: **What are DNA sequence motifs?** *Nature biotechnology* 2006, **24**(4):423–425.

[20] Moreno-Hagelsieb G, Jokic P: **The evolutionary dynamics of functional modules and the extraordinary plasticity of regulons: the Escherichia coli perspective**. *Nucleic acids research* 2012, **40**(15):7104–7112.

[21] Cohen O, Ashkenazy H, Levy Karin E, Burstein D, Pupko T: **CoPAP: Coevolution of PresenceAbsence Patterns**. *Nucleic Acids Research* 2013, **41**(W1):W232–W237.

[22] Kuzminov A: **Recombinational Repair of DNA Damage inEscherichia coli and Bacteriophage** λ. *Microbiology and Molecular Biology Reviews* 1999, **63**(4):751–813.

[23] Chow KH, Courcelle J: **RecO acts with RecF and RecR to protect and maintain replication forks blocked by UV-induced DNA damage in Escherichia coli**. *Journal of Biological Chemistry* 2004, **279**(5):3492–3496.

[24] Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martnez C, Fulcher C, Huerta AM, Kothari A, Krummenacker M, Latendresse M, Muiz-Rascado L, Ong Q, Paley S, Schrder I, Shearer AG, Subhraveti P, Travers M, Weerasinghe D, Weiss V, Collado-Vides J, Gunsalus RP, Paulsen I, Karp PD: **EcoCyc: fusing model organism databases with systems biology**. *Nucleic Acids Research* 2013, **41**(D1):D605–D612.

[25] Whitby MC, Lloyd RG: **Altered SOS induction associated with mutations in recF, recO and recR**. *Molecular and General Genetics MGG* 1995, **246**(2):174–179.

[26] Kihara A, Akiyama Y, Ito K: **FtsH is required for proteolytic elimination of uncomplexed forms of SecY, an essential protein translocase subunit**. *Proceedings of the National Academy of Sciences* 1995, **92**(10):4532–4536.

[27] Huynen M, Snel B, Lathe W, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences**. *Genome research* 2000, **10**(8):1204–1210.

[28] Hagelsieb GM, Collado-Vides J: **Operon conservation from the point of view of Escherichia coli, and inference of functional interdependence of gene products from genome context**. *In silico biology* 2002, **2**(2):87–95.

[29] Maglott DR, Katz KS, Sicotte H, Pruitt KD: **NCBIs LocusLink and RefSeq**. *Nucleic acids research* 2000, **28**:126–128.

[30] McEntyre J, Ostell J: **The NCBI handbook** 2002.

[31] Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic acids research* 2007, **35**(suppl 1):D61–D65.

[32] Hertz GZ, III GWH, Stormo GD: **Identification of consensus patterns in unaligned DNA sequences known to be functionally related**. *Computer applications in the Biosciences* 1990, **6**(2):81–92.

[33] Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15**(7):563–577.

[34] Haddock SHD, Dunn CW: *Practical computing for biologists*. Sinauer Associates Sunderland, MA 2011.

[35] Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in bipolymers** 1994.

[36] Bailey TL, Gribskov M: **Combining evidence using p-values: application to sequence homology searches.** *Bioinformatics* 1998, **14**:48–54.

[37] Bailey TL, Williams N, Misleh C, Li WW: **MEME: discovering and analyzing DNA and protein sequence motifs**. *Nucleic Acids Research* 2006, **34**(suppl 2):W369–W373.

[38] Ranea JA, Buchan DW, Thornton JM, Orengo CA: **Evolution of protein superfamilies and bacterial genome size**. *Journal of molecular biology* 2004, **336**(4):871–887.

[39] Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME Suite: tools for motif discovery and searching**. *Nucleic Acids Research* 2009, **37**(suppl 2):W202–W208.

[40] R Core Team: **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria 2013, [`http://www.R-project.org`].

[41] Wickham H: *ggplot2: elegant graphics for data analysis*. Springer Publishing Company, Incorporated 2009.

[42] Fowler J, Cohen L, Jarvis P, Wiley J: *Practical statistics for field biology*. Wiley Chichester 1998.

[43] Skewes AD, Welch RD: **A Markovian analysis of bacterial genome sequence constraints**. *PeerJ* 2013, **1**:e127.

[44] Moreno-Hagelsieb G, Janga SC: **Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles**. *Proteins: Structure, Function, and Bioinformatics* 2008, **70**(2):344–352.

[45] Moreno-Hagelsieb G, Wang Z, Walsh S, ElSherbiny A: **Phylogenomic clustering for selecting non-redundant genomes for comparative genomics.** *Bioinformatics* 2013, **29**(7):947–949.

[46] Moreno-Hagelsieb G, Latimer K: **Choosing BLAST options for better detection of orthologs as reciprocal best hits**. *Bioinformatics* 2008, **24**(3):319–324.

[47] Sierro N, Makita Y, de Hoon M, Nakai K: **DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information**. *Nucleic acids research* 2008, **36**(suppl 1):D93–D96.

[48] Lima T, Auchincloss AH, Coudert E, Keller G, Michoud K, Rivoire C, Bulliard V, de Castro E, Lachaize C, Baratin D, Phan I, Bougueleret L, Bairoch A: **HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot**. *Nucleic Acids Research* 2009, **37**(suppl 1):D471–D478.

[49] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al.: **Gene Ontology: tool for the unification of biology**. *Nature genetics* 2000, **25**:25–29.

[50] Chan J, Kishore R, Sternberg P, Van Auken K: **The gene ontology: enhancements for 2011**. *Nucleic Acids Research* 2012, **40**(D1):D559–D564.

[51] Tripathi S, Christie KR, Balakrishnan R, Huntley R, Hill DP, Thommesen L, Blake JA, Kuiper M, Lgreid A: **Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort**. *Database* 2013, **2013**.

[52] Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD: **The Pfam protein families database**. *Nucleic Acids Research* 2012, **40**(D1):D290–D301.

[53] Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA: **DBD taxonomically broad transcription factor predictions: new content and functionality**. *Nucleic acids research* 2008, **36**(suppl 1):D88–92.

[54] Eddy SR: **Accelerated profile HMM searches**. *PLoS computational biology* 2011, **7**(10):e1002195.

[55] Martínez-Núñez MA, Poot-Hernandez AC, Rodríguez-Vázquez K, Pérez-Rueda E: **Increments and Duplication Events of Enzymes and Transcription Factors Influence Metabolic and Regulatory Diversity in Prokaryotes**. *PLoS ONE* 2013, **8**(7):e69707.

[56] Tatusov RL, Koonin EV, Lipman DJ: **A Genomic Perspective on Protein Families**. *Science* 1997, **278**(5338):631–637.

[57] Yutin N, Puigbò P, Koonin EV, Wolf YI: **Phylogenomics of prokaryotic ribosomal proteins**. *PLoS One* 2012, **7**(5):e36972.

[58] Rice P, Longden I, Bleasby A: **EMBOSS: the European molecular biology open software suite**. *Trends in genetics* 2000, **16**(6):276–277.

[59] Huang Q, Cheng X, Cheung MK, Kiselev SS, Ozoline ON, Kwan HS: **High-density transcriptional initiation signals underline genomic islands in bacteria**. *PLoS ONE* 2012, **7**(3):e33759.

[60] Langille MG, Hsiao WW, Brinkman FS: **Detecting genomic islands using bioinformatics approaches**. *Nature Reviews Microbiology* 2010, **8**(5):373–382.

[61] Hershberg R, Margalit H: **Co-evolution of transcription factors and their targets depends on mode of regulation**. *Genome biology* 2006, **7**(7):R62.

[62] Charoensawan V, Wilson D, Teichmann SA: **Genomic repertoires of DNA-binding transcription factors across the tree of life**. *Nucleic acids research* 2010, **38**(21):7364–7377.

[63] Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J: **SUPERFAMILYsophisticated comparative genomics, data mining, visualization and phylogeny**. *Nucleic acids research* 2009, **37**(suppl 1):D380–D386.

[64] Langille MG, Brinkman FS: **IslandViewer: an integrated interface for computational identification and visualization of genomic islands**. *Bioinformatics* 2009, **25**(5):664–665.

A False positive transcription factor predictions

The information below is for the potentially falsely predicted TFs. These predicted TFs were not identified as TFs by the manually curated lists from RegulonDB [1, 10] for *E. coli* or DBTBS [47] for *B. subtilis*; TFs were identified only from predictions by protein domain HMMs provided by the Pfam database [52].

Each putative TF was assessed based on the information that could be manually collected from NCBI [30]. If, within the metadata contained by searching the GI, there was enough information to suggest the protein's involvement in DNA-dependent transcriptional regulation as well as at least a putative DNA-binding domain, it was considered a true TF. See Figure A.1 for the decision tree which was used to determine sufficient evidence for a TF.

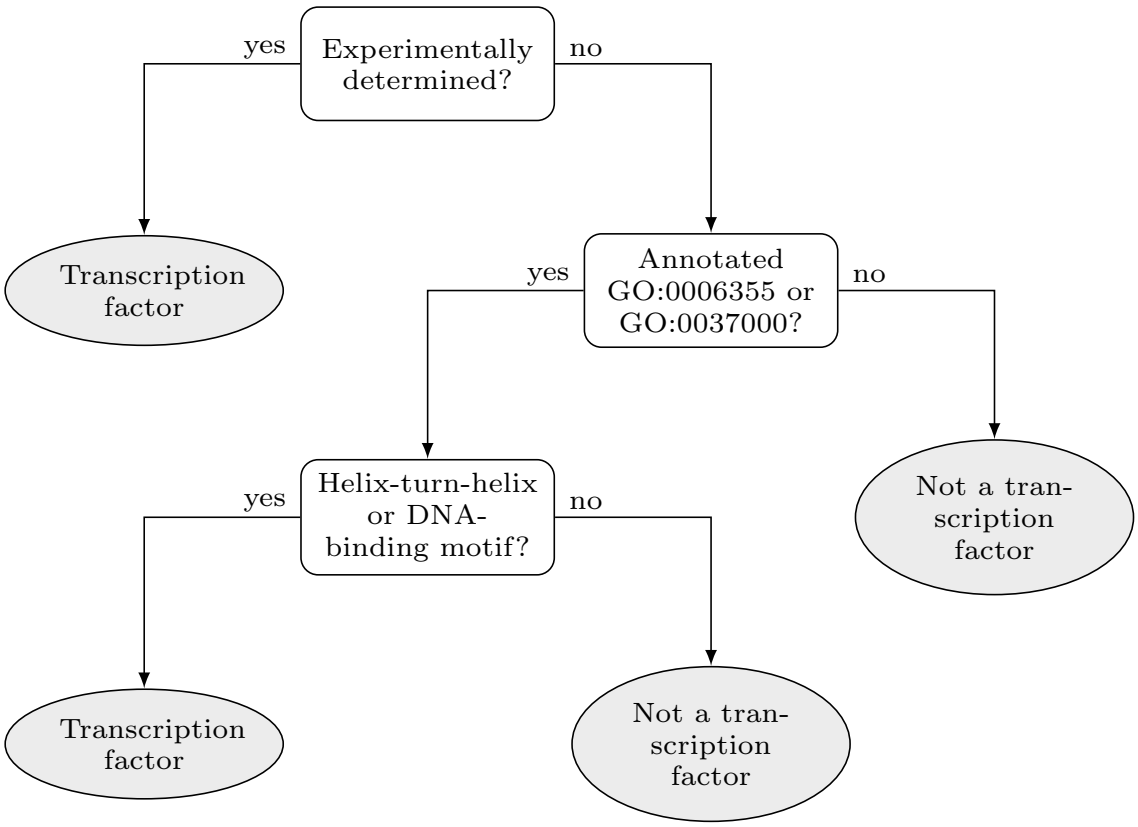


Figure A.1: Decision tree for determining validity of potentially falsely predicted TFs. Gene ontology (GO) annotations 0006355 and 0037000 are specific for transcriptional regulation; to confirm that these proteins are involved in DNA dependent transcriptional regulation, it was also important to verify a helix-turn-helix or DNA binding domain.

Table A.1: NCBI data for non-curated TFs predicted by Pfam HMMs in *E. coli*

GI	NCBI Definition	NCBI Description	Pfam	Regulator
16130977	antitoxinof the HigB-HigA toxin-antitoxin system	Contains non-specific and specific DNA-binding sites; part of XRE family	HTH_3:PF01381.17	Yes
16128444	modulator of gene expression with H-NS	Transcriptional regulator	HHA:PF05321.6	Yes
16130098	putative kinase	Contains a winged HTH and ATP binding domain	HTH_11:PF08279.7	No
16129583	oriC-binding complex H-NS/Cnu; binds 26 bp cnb site; also forms a complex with StpA	See definition	HHA:PF05321.6	No
90111706	putative transcriptional regulator, HxlR-type, DUF24 family	Transcriptional regulator	HxlR:PF01638.12	Yes
90111708	antitoxin of the ChpBS toxin-antitoxin system	Autoregulated	Antitoxin-MazE:PF04014.13	No
16129176	putative inner membrane protein, SIRB family	SirB family proteins are regulators of transcription	SirB:PF04247.7	No

Table A.2: NCBI data for non-curated TFs predicted by Pfam HMMs in *B. subtilis*

GI	**NCBI Definition**	**NCBI Description**	**Pfam**	**Regulator**
16080092	two-component response regulator controlling resistance to antibiotics affecting the envelope YtsB	Transcriptional regulator	Trans_reg_C:PF00486.23	Yes
16079466	transcriptional regulator	Transcriptional regulator	HTH_8:PF02954.14	Yes
16078218	DNA transport protein	competence protein CoiA; likely has a DNA-binding domain	CoiA:PF06054.6	No
16080470	LacI family transcriptional regulator	Transcriptional regulator	LacI:PF00356.16	Yes
16077579	cold-shock protein	Contains a DNA-binding site; part of Csp family that contains other regulators	CSD:PF00313.17	No
16079765	transcriptional regulator	Transcriptional regulator	TrmB:PF01978.14	Yes
255767247	DNA/RNA binding protein	DNA-binding protein YizB; predicted regulator	PadR:PF03551.9	Yes

Continued on next page

Table A.2 – continued from previous page

GI	NCBI Definition	NCBI Description	Pfam	Regulator
16078980	two-component response regulator DesK	Contains HTH DNA-binding domain; LuxR_C_like is a response regulator; Transcriptional regulator	GerE:PF00196.14	Yes
16080564	regulator (stress mediated)	Putative stress-responsive transcriptional regulator	PspC:PF04024.7	Yes
161511067	DNA-binding transcriptional regulator FrlR	Transcriptional regulator	GntR:PF00392.16	Yes
16080420	transcriptional repressor	Transcriptional regulator	HTH_3:PF01381.17	Yes
16080817	regulator of sulfur assimilation CysL, activates cysJI expression	Transcriptional regulator	HTH_1:PF00126.22	Yes
16078903	LysR family transcriptional regulator	Transcriptional regulator	HTH_1:PF00126.22	Yes
16078380	transcriptional regulator sensing organic peroxides	Transcriptional regulator	MarR:PF01047.17	Yes

Continued on next page

Table A.2 – continued from previous page

GI	NCBI Definition	NCBI Description	Pfam	Regulator
16077652	GntR family transcriptional regulator	Transcriptional regulator	GntR:PF00392.16	Yes
255767803	transcriptional regulator	Predicted regulator	PadR:PF03551.9	Yes
16081093	two-component response regulator YycG	Transcriptional regulator	Trans_reg_C:PF00486.23	Yes
255767747	transcriptional regulator	Predicted regulator	HTH_5:PF01022.15	Yes
16079579	negative regulator of gluconeogenesis	Transcriptional regulator	HTH_DeoR:PF08220.7	Yes
16078003	NO-dependent activator of the ResDE regulon	Transcriptional regulator	Rrf2:PF02082.15	Yes
16078431	MarR family transcriptional regulator	Transcriptional regulator	MarR:PF01047.17	Yes
16077343	two-component response regulator NatK	Transcriptional regulator	LytTR:PF04397.10	Yes
255767206	HTH-type transcriptional regulator	Predicted regulator	HTH_3:PF01381.17	Yes

Continued on next page

Table A.2 – continued from previous page

GI	NCBI Definition	NCBI Description	Pfam	Regulator
255767729	transcriptional regulator	Predicted regulator	CodY:PF06018.9	Yes
16078021	copper efflux transcriptional regulator	Transcriptional regulator	MerR:PF00376.18	Yes
16077549	XRE family transcriptional regulator	Transcriptional regulator	HTH_3:PF01381.17	Yes
16080361	two-component response regulator YvqE responding to cell wall stress	Transcriptional regulator	GerE:PF00196.14	Yes
16077975	cold-shock protein	DNA and RNA-binding motifs; probable regulator	CSD:PF00313.17	No
16077886	Mal operon transcriptional activator	Transcriptional regulator	HTH_6:PF01418.12	Yes
16080716	hypothetical protein BSU36630	Predicted regulator	Rrf2:PF02082.15	Yes
255767629	cysteine biosynthesis transcriptional regulator	Transcriptional regulator	Rrf2:PF02082.15	Yes

Continued on next page

Table A.2 – continued from previous page

GI	NCBI Definition	NCBI Description	Pfam	Regulator
16078972	ArsR family transcriptional regulator	Transcriptional regulator	HTH_5:PF01022.15	Yes
16078970	transcriptional regulator	Predicted regulator	HTH_11:PF08279.7	Yes
16077726	hypothetical protein BSU06580	Uncharacterized protein conserved in bacteria	Trp_repressor:PF01371.14	No
16077600	ArsR family transcriptional regulator	Transcriptional regulator	HTH_5:PF01022.15	Yes
16077877	transcriptional regulator	Transcriptional regulator	HTH_8:PF02954.14	Yes
16080432	ArsR family transcriptional regulator	Transcriptional regulator	HTH_5:PF01022.15	Yes
16080516	LacI family transcriptional regulator	Predicted regulator	LacI:PF00356.16	Yes
16081087	NtrC/NifA family transcriptional regulator	Transcriptional regulator	HTH_8:PF02954.14	Yes
16080419	transcriptional repressor	Transcriptional regulator	HTH_3:PF01381.17	Yes
255767639	transcriptional repressor	Transcriptional regulator	HTH_11:PF08279.7	Yes

Continued on next page

Table A.2 – continued from previous page

GI	NCBI Definition	NCBI Description	Pfam	Regulator
16079252	cold-shock protein	DNA and RNA-binding motifs; probable regulator	CSD:PF00313.17	No
16080491	transcriptional regulator	Transcriptional regulator	HTH_3:PF01381.17	Yes