ANALYSIS OF DNA STRUCTURE-RELATED LOCAL SEQUENCE PATTERNS AND REGULATORY MOTIFS IN PROKARYOTIC GENOMES

by

YONGJIE HUANG

(Under the Direction of Jan Mrázek)

ABSTRACT

Prokaryotic genomes are diverse in terms of their nucleotide and oligonucleotide composition as well as presence of various sequence features that can affect physical properties of DNA molecule. We present a survey of local sequence patterns which have a potential to promote non-canonical DNA conformations (i.e., different from standard B-DNA double helix) and interpret the results in terms of relationships with organisms' habitats, phylogenetic classifications, and other characteristics. Our work differs from earlier similar surveys not only by investigating a wider range of sequence patterns in a large number of genomes but also by using a more realistic null model to assess significant deviations. Our results show that simple sequence repeats and Z-DNA-promoting patterns are generally suppressed in prokaryotic genomes, whereas palindromes and inverted repeats are overrepresented. Representation of patterns that promote Z-DNA and intrinsic DNA curvature increases with increasing optimal growth temperature (OGT), and decreases with increasing oxygen requirement. The observed relationships with environmental characteristics, particularly OGT, suggest possible evolutionary scenarios of structural adaptation of DNA to particular environmental niches.

As a natural next step, we develop software for identification of specific occurrences of the structure-related patterns and regulatory motifs that are under selective constraints, which would be indicative of a physiological role of such patterns. This is achieved by two major steps. First, the program finds orthologous sites matching the given sequence pattern in a collection of related genomes; the level of pattern conservation is subsequently evaluated by comparison of information entropy within each pattern occurrence and its immediate flanking sequences in the multiple sequence alignment of the orthologous sites. The new tools have been demonstrated in several pilot studies, including analysis of palindromic sequence patterns and intrinsically curved segments in *Campylobacter* and σ^{54} binding site motifs in *Salmonella* and *E. coli*. Our methodology for investigation of evolution of regulatory motifs is an important step towards understanding the evolution of regulatory networks and how organisms adapt to changing conditions or environments. The program for detection of sequence patterns that are under selective constraint can serve as an exploratory and hypothesis-generating tool, which can be of significant interest to the scientific community.

INDEX WORDS: Sequence patterns; DNA conformation; Z-DNA; DNA curvature; sequence repeats; palindromes; DNA sequence motifs; Sigma factor; protein binding sites; ortholog; prokaryotic genomes; comparative genomics

ANALYSIS OF DNA STRUCTURE-RELATED LOCAL SEQUENCE PATTERNS AND REGULATORY MOTIFS IN PROKARYOTIC GENOMES

by

YONGJIE HUANG

B.E., University of Shanghai for Science and Technology, China, 2006

M.S., The University of South Carolina, 2009

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2015

ANALYSIS OF DNA STRUCTURE-RELATED LOCAL SEQUENCE PATTERNS AND REGULATORY MOTIFS IN PROKARYOTIC GENOMES

by

YONGJIE HUANG

Major Professor:	Jan Mrázek
Committee:	Liming Cai Paul Schliekelman William B. Whitman

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
May 2015

DEDICATION

To my parents, for their endless love and support.

ACKNOWLEDGEMENTS

My First and most earnest acknowledgment must go to my major advisor Dr. Jan Mrázek. Jan has been instrumental in ensuring my academic, professional and financial wellbeing ever since I jointed his group. In every sense, none of this work would have been possible without him. Many thanks also go to my committee members Drs. William B. Whitman, Paul Schliekelman and Liming Cai for their valuable time and suggestions.

I would like to thank Drs. Anna Karls and Timothy Hoover who provided me helpful advice. I also want to thank all members from the CMBL group and Dr. Whitman's lab. They provided me a great working environment.

My final and most heartfelt acknowledgement is to my husband Haiwei Luo and my daughter Annie, for their constant encouragement and motivation.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

While analyses of genomes are generally focused on functions of genes, proteins or ncRNAs, and regulatory or metabolic networks, they rarely include considerations of DNA structure or other properties of the DNA itself. There are still substantial gaps in our understanding how the nucleoid structure relates to DNA sequence, how it varies among different bacteria, and how structural heterogeneity of nucleoid affects physiological process in the cells. This work introduces new methodology and tools for exploration of possible roles of sequence-encoded structural elements (e.g., particular DNA sequence patterns) and regulatory motifs (e.g., transcription factor binding sites) in bacteria and archaea by computational techniques.

Structure-related sequence patterns in prokaryotic genomes

For the purpose of this work, we define sequence patterns in DNA sequences are short sequences (typically up to tens of bp) that have some characteristic properties that distinguish them from general DNA sequences. Our study mainly involves sequence patterns that have a potential to generate non-canonical DNA conformations and sequence patterns promoting conformational transitions in DNA.

Although physiological functions of such sequence patterns are not well understood, there is increasing evidence of their important roles in organisms. For example, simple sequence repeats (SSRs) are tandem iterations of short repetitive units, which can expand or contract via

slip-strand mutations (Mrazek, Guo et al. 2007). They have been association with phase variation, a reversible switching between phenotypes that arises from reversible activation or deactivation of genes (Moxon, Rainey et al. 1994; van der Woude and Baumler 2004). Palindromes and inverted repeats can form stem loops in RNA, which may function as transcription terminator and riboswitches. They can also promote DNA cruciform structures which are fundamentally important for a wide range of biological processes, including replication, regulation of gene expression and recombination (Brazda, Laister et al. 2011). Some DNA mirror repeats, oligopurine or oligopyrimidine tracts can promote formation of triple-helical H-DNA, which can influence the regulation of gene expression (Sinden 1994; Wang and Vasquez 2006). Similar roles are suggested for quadruplex G-DNA which are formed by specific guanine-rich patterns (Rawal, Kummarasetti et al. 2006). Alternating purine-pyrimidine patterns can promote transitions to Z-DNA conformation under favorable conditions (Trifonov, Konopka et al. 1985). Transient formation of left-handed Z-DNA can influence transcription and lead to genome instability (Wang and Vasquez 2006). The representations of alternating purine-pyrimidine patterns are found to be largely different in prokaryotic genomes and the patterns may be related to optimal growth temperature (OGT) and pathogenicity, possibly as a result of structural adaptations of DNA of these organisms (Bohlin, Hardy et al. 2009) .

Intrinsic DNA curvature is primarily caused by A-tracts (short runs of As or Ts) periodically spaced with the DNA helical period of about 10.5bp (Figure 1.1). Other sequence patterns may also contribute to DNA intrinsic curvature, but the periodically spaced A-tracts have a dominant effect (Trifonov 1985). It has been speculated that such A-tracts are involved in the relationship between DNA sequence and chromatin structure in eukaryotes. They are also wide-spread in bacteria and may contribute to DNA compaction in the nucleoid (Tolstorukov,

Virnik et al. 2005). The amount of intrinsic DNA curvature in promoter regions was found to be related to OGT(Kozobay-Avraham, Hosid et al. 2006). Our own analysis of more than 1000 prokaryotic chromosomes suggested a large variance among different genomes in DNA curvature-related 10-11 bp sequence periodicity, which could reflect differences in chromosomal strucrture(Mrazek, Guo et al. 2007).

These results indicate that understanding roles of structure-related sequence patterns can be crucial to understanding physiological processes within the cell. The present work mainly targets at developing hypotheses about the evolutionary roles of intrinsic bends as well as other sequence patterns that promoting irregular DNA conformations.

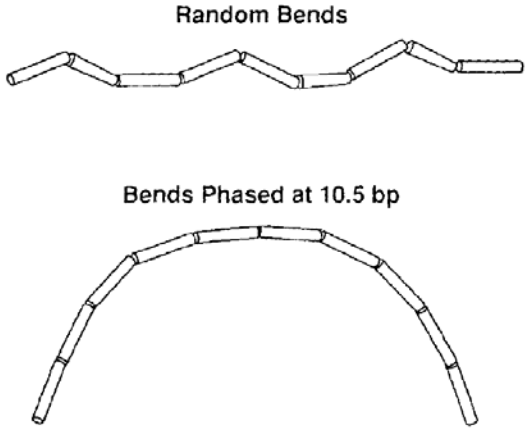

Figure 1.1 Significance of phasing in DNA bending. An A tract will introduce a small bend or deflection of the helix axis in DNA. Only when bends are phased by 10.5 bp is the stable curvature. This figure is adapted from reference (Sinden 1994)

Regulatory motifs in prokaryotic genomes

Regulatory motifs are short DNA sequences that affect expression of genes and binding sites for regulatory proteins (e.g., RNA polymerase sigma factor). They function at the RNA transcription stage and determining under which condition the associated gene is turned on or off (Joanne Willey 2008). Identification of regulatory motifs is non-trivial. They are generally short (up to several of tens of bp) and although they usually involve some conserved features, individual binding sites for a specific protein can tolerate significant sequence variance. There are experimental methods to identify regulatory motifs, such as ChIP-chip and ChIP-seq techniques, but their application is time consuming and have been used for only a small fraction of known regulatory proteins (Buck and Lieb 2004; Park 2009; Pepke, Wold et al. 2009). Comparative genomics offer various methods for finding regulatory motifs. The simplest way is to find significantly conserved regions by performing cross-species sequence alignment or phylogenetic footprinting (McGuire, Hughes et al. 2000; McCue, Thompson et al. 2001). To date, tools of finding significant nucleotide sequence motifs have been successfully established for various tasks (Thompson, Rouchka et al. 2003; Mrazek, Xie et al. 2008; Bailey, Boden et al. 2009; Tran and Huang 2014).

In this dissertation project, we focus on investigating the conservation properties of known regulatory motifs and possibly predict functions of highly conserved motifs in a collection of prokaryotic genomes. There are many motif profiling and motif comparison tools. For example, GOMO performs Genome Ontology (GO) term association with DNA motifs (Buske, Boden et al. 2010); TomTom is used to search DNA motifs against a DNA motif database and produce an alignment for each significant match (Gupta, Stamatoyannopoulos et al. 2007). However, existing software for analysis of evolutionary roles of regulatory motifs are not existent or very

limited. In this work, we aim to develop a new method for analysis of evolution of motifs in prokaryotic genomes that automatically finds sets of orthologous motif sites in a collection of related genomes and evaluates the evolutionary conservation of individual motif occurrences. The results are used to quickly identify motif sites that could be under positive or negative selective constraints.

Overview of dissertation chapters

In Chapter 2, a survey of DNA structure-related local sequence patterns in more than 1500 complete microbial genomes is described. This work differs from earlier similar surveys not only by investigating a wider range of sequence patterns in a large number of genomes but also by using a more realistic null model to assess significant anomalies. Our null model reflects the genome-specific nearest neighbor preferences, codon biases, and heterogeneity of the DNA sequence; therefore, deviations from expected occurrences are more likely to reflect the functional significance of the investigated sequence patterns. Results of this survey are then interpreted in terms of relationships with characteristics of the organisms, such as habitats and taxonomical classifications.

In Chapter 3, we develop a new method for analysis of evolution of motifs in prokaryotic genomes, with a particular focus on transcription factor binding sites. Given a set of known regulatory motif sequences, our program gives out groups of orthologous motifs in the analyzed genomes. Additional information, such as motif scores, measures of evolutionary conservation, surrounding genes and multiple sequence alignment of the orthologous motifs, are provided by the program to facilitate the analysis. As a pilot study, the new tool is used to investigate RpoN

σ^{54} binding site motifs in *Salmonella* and *E. coli*. Results are compared to previous studies and new findings are discussed.

In Chapter 4, software developed in Chapter 3 is adjusted to study evolution of structure-related local sequence patterns. The primary goal is to identify DNA structure-related sequence patterns that may be subject to selective constraints by comparing the conservation of the sequence matching the pattern with its immediate flanking sequences. The methodology and examples of its application are described.

References

Bailey, T. L., M. Boden, et al. (2009). "MEME SUITE: tools for motif discovery and searching." Nucleic Acids Res **37**: W202-W208.

Bohlin, J., S. P. Hardy, et al. (2009). "Stretches of alternating pyrimidine/purines and purines are respectively linked with pathogenicity and growth temperature in prokaryotes." Bmc Genomics **10**.

Brazda, V., R. C. Laister, et al. (2011). "Cruciform structures are a common DNA feature important for regulating biological processes." BMC Mol Biol **12**: 33.

Buck, M. J. and J. D. Lieb (2004). "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments." Genomics **83**(3): 349-360.

Buske, F. A., M. Boden, et al. (2010). "Assigning roles to DNA regulatory motifs using comparative genomics." Bioinformatics **26**(7): 860-866.

Gupta, S., J. A. Stamatoyannopoulos, et al. (2007). "Quantifying similarity between motifs." Genome Biology **8**(2).

Joanne Willey, L. S., Chris Woolverton (2008). Prescott, Harley and Klein's Microbiology.

Kozobay-Avraham, L., S. Hosid, et al. (2006). "Involvement of DNA curvature in intergenic regions of prokaryotes." Nucleic Acids Res **34**(8): 2316-2327.

McCue, L. A., W. Thompson, et al. (2001). "Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes." Nucleic Acids Res **29**(3): 774-782.

McGuire, A. M., J. D. Hughes, et al. (2000). "Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes." Genome Research **10**(6): 744-757.

Moxon, E. R., P. B. Rainey, et al. (1994). "Adaptive evolution of highly mutable loci in pathogenic bacteria." Curr Biol **4**(1): 24-33.

Mrazek, J., X. X. Guo, et al. (2007). "Simple sequence repeats in prokaryotic genomes." Proc Natl Acad Sci U S A **104**(20): 8472-8477.

Mrazek, J., S. Xie, et al. (2008). "AIMIE: a web-based environment for detection and interpretation of significant sequence motifs in prokaryotic genomes." Bioinformatics **24**(8): 1041-1048.

Park, P. J. (2009). "ChIP-seq: advantages and challenges of a maturing technology." Nature Reviews Genetics **10**(10): 669-680.

Pepke, S., B. Wold, et al. (2009). "Computation for ChIP-seq and RNA-seq studies." Nature Methods **6**(11): S22-S32.

Rawal, P., V. B. R. Kummarasetti, et al. (2006). "Genome-wide prediction of G4 DNA as regulatory motifs: Role in Escherichia coli global regulation." Genome Research **16**(5): 644-655.

Sinden, R. R. (1994). DNA Structure and Function, Academic Press.

Thompson, W., E. C. Rouchka, et al. (2003). "Gibbs Recursive Sampler: finding transcription factor binding sites." Nucleic Acids Res **31**(13): 3580-3585.

Tolstorukov, M. Y., K. M. Virnik, et al. (2005). "A-tract clusters may facilitate DNA packaging in bacterial nucleoid." Nucleic Acids Res **33**(12): 3907-3918.

Tran, N. T. L. and C. H. Huang (2014). "A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data." Biology Direct **9**.

Trifonov, E. N. (1985). "Curved DNA." CRC Crit Rev Biochem **19**(2): 89-106.

Trifonov, E. N., A. K. Konopka, et al. (1985). "Unusual frequencies of certain alternating purine-pyrimidine runs in natural DNA sequences: relation to Z-DNA." FEBS Lett **185**(1): 197-202.

van der Woude, M. W. and A. J. Baumler (2004). "Phase and antigenic variation in bacteria." Clin Microbiol Rev **17**(3): 581-611, table of contents.

Wang, G. and K. M. Vasquez (2006). "Non-B DNA structure-induced genetic instability." Mutat Res **598**(1-2): 103-119.

CHAPTER 2

ASSESSING DIVERSITY OF DNA STRUCTURE-RELATED SEQUENCE FEATURES IN PROKARYOTIC GENOMES[1]

[1] Huang, Y. and J. Mrazek, *DNA Res*, 2014. **21**(3): p. 285-97.

Abstract

Prokaryotic genomes are diverse in terms of their nucleotide and oligonucleotide composition as well as presence of various sequence features that can affect physical properties of the DNA molecule. We present a survey of local sequence patterns which have a potential to promote non-canonical DNA conformations (i.e., different from standard B-DNA double helix) and interpret the results in terms of relationships with organisms' habitats, phylogenetic classifications, and other characteristics. Our present work differs from earlier similar surveys not only by investigating a wider range of sequence patterns in a large number of genomes but also by using a more realistic null model to assess significant deviations. Our results show that simple sequence repeats and Z-DNA-promoting patterns are generally suppressed in prokaryotic genomes, whereas palindromes and inverted repeats are overrepresented. Representation of patterns that promote Z-DNA and intrinsic DNA curvature increases with increasing optimal growth temperature (OGT), and decreases with increasing oxygen requirement. Additionally, representations of close direct repeats, palindromes and inverted repeats exhibit clear negative trends with increasing OGT. The observed relationships with environmental characteristics, particularly OGT, suggest possible evolutionary scenarios of structural adaptation of DNA to particular environmental niches.

Key words: Sequence patterns, Z-DNA, DNA curvature, sequence repeats, palindromes

Introduction

Prokaryotic genomes are extremely diverse in terms of their nucleotide and oligonucleotide composition as well as presence of various forms of sequence repeats and patterns that can affect physical properties of the DNA molecule. For example, Ussery and

coworkers (Ussery, Soumpasis et al. 2002) analyzed oligopurine/oligopyrimidine runs and alternating purine/pyrimidine patterns in prokaryotic genomes and reported large differences among different organisms. Alternating purine/pyrimidine patterns promote Z-DNA conformation, whereas oligopurine/oligopurimidine runs can facilitate formation of A-DNA or H-DNA (Sinden 1994). They found that differences among different prokaryotes were related to optimal growth temperature and pathogenicity, possibly as a result of structural adaptations of DNA of these organisms (Bohlin, Hardy et al. 2009). In another example, potential G-DNA-forming sequences were found to occur at vastly different frequencies in different prokaryotic genomes (Rawal, Kummarasetti et al. 2006). The amount of intrinsic DNA curvature in promoter regions was found to be related to optimal growth temperature (Kozobay-Avraham, Hosid et al. 2006). Our own analysis of more than 1000 prokaryotic chromosomes suggested a large variance among different genomes in DNA curvature-related 10-11 bp sequence periodicity, which could reflect differences in chromosomal structure (Mrázek 2010). Presence of long simple sequence repeats in a genome is strongly correlated with the organism's dependence on a eukaryotic host (Mrázek, Guo et al. 2007). These results indicate a significant variance of general DNA properties among different prokaryotes, which in some cases appear to be related to the organisms' habitats and lifestyles.

Advances in DNA sequencing technologies over the past decades led to a situation where complete genomes are available for many microbes about which very little is known apart from the information that can be derived from the genomic DNA sequence. In particular, despite the diversity of DNA properties mentioned above, our knowledge about chromosome structure and organization in the cell is limited to studies of a few model organisms. While it is reasonable to assume that the general model of bacterial nucleoid composed of dynamic, supercoiled DNA

loops stabilized by nucleoid-associated proteins is probably universal(Thanbichler, Wang et al. 2005; Dillon and Dorman 2010), the differences in DNA sequence properties suggest that subtle variations may exist among bacteria with completely sequenced genomes. Such differences in physical characteristics of the chromosomes could play roles in the organisms' physiology and adaptations to their particular environments.

We present a survey of local sequence patterns in prokaryotic genomes with a potential to generate local irregularities in DNA structure. Our goal was to assess diversity of the prokaryotic genomes in terms of abundance of sequence patterns indicative of possible structural transitions in the DNA molecule. While physiological functions of sequence patterns promoting non-canonical DNA conformations and sequence patterns promoting conformational transitions in DNA are not well understood, there is increasing evidence that they can play important roles in both eukaryotes and prokaryotes. For example, simple sequence repeats have been implicated in phase variation, a mechanism that promotes reversible switching of phenotypes (Moxon, Rainey et al. 1994; van der Woude and Baumler 2004). Palindromes and close inverted repeats form stem-loop structures in RNA, which can function in transcription terminators and riboswitches, and they can promote formation of cruciform structures in DNA, which influence replication, regulation of gene expression, and recombination (Brazda, Laister et al. 2011). Close direct and inverted repeats have mutagenic effects on DNA(Levinson and Gutman 1987; Chuzhanova, Abeysinghe et al. 2003; Lovett 2004; Dutra and Lovett 2006). Specific guanine-rich patterns can promote formation of quadruplex G-DNA(Shafer and Smirnov 2000; Burge, Parkinson et al. 2006; Vorlíčková, Bednářová et al. 2007). The G-DNA formation in telomeres is well documented but G-DNA may also play a role in gene expression, replication, recombination, and integration of viruses(Sundquist and Heaphy 1993; Shafer and Smirnov 2000; Arthanari and

Bolton 2001; Schaeffer, Bardoni et al. 2001; Pan, Shi et al. 2006). Alternating purine-pyrimidine patterns can under favorable conditions promote transitions to the Z-DNA conformation while oligopurine/oligopyrimidine runs promote formation of A-DNA or triple-stranded H-DNA(Rich, Nordheim et al. 1983; Belotserkovskii, Veselkov et al. 1990; Sinden 1994; van Holde and Zlatanova 1994; Rustighi, Tessari et al. 2002; Zain and Sun 2003). Transient formation of left-handed Z-DNA can influence transcription and promote genome instability(van Holde and Zlatanova 1994; Herbert and Rich 1999; Rich and Zhang 2003; Wang and Vasquez 2007), and indirect evidence suggests similar roles for triple-helical H-DNA(Zain and Sun 2003; Jain, Wang et al. 2008). Intrinsic DNA curvature is largely related to periodic spacings of A-tracts and can influence DNA-protein interactions in regulatory regions, aid DNA compaction in the nucleoid, and possibly promote a particular mode of supercoiling (Herzel, Weiss et al. 1998; Tolstorukov, Virnik et al. 2005; Kozobay-Avraham, Hosid et al. 2006). Although the physiological significance of the unusual DNA conformations in different organisms is still poorly understood it is reasonable to ask how common the sequence patterns that promote their formation are in different genomes because their unusually high or low occurrence may indicate its functional importance.

Our present work differs from earlier similar surveys not only by investigating a wider range of sequence patterns in a large number of genomes but also by using a more realistic null model to assess significant anomalies. Significant deviations from expected pattern occurrences could indicate that the pattern *per se* is subject to selective constraint and therefore functionally significant in a given organism but the pattern over- or under-representation could also arise as a tolerated artifact of other biases affecting the DNA sequence. Our null model reflects the genome-specific nearest neighbor preferences, codon biases, and heterogeneity of the DNA

sequence; therefore deviations from expected occurrences are more likely to reflect their functional significance of the investigated sequence patterns. We interpret our results in terms of relationships with organisms' habitats, phylogenetic classifications, and other characteristics.

Materials and Methods

Data sets

Complete prokaryotic genomes were downloaded from the National Center for Biotechnology Information (NCBI) ftp server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Pattern representations were assessed for the complete genomes, as well as for 1 Mb segments randomly selected from each genome. The purpose of the latter approach was to reduce the effect of statistical artifacts from comparing sequences of different sizes. The results presented in this paper are based on the 1 Mb segments while the data for the complete genomes are shown in the supplemental Excel files. Genomes smaller than 1 Mb were excluded from the analysis. Our final dataset included 1424 complete genomes of 941 species, 519 genera, and 37 phyla or subphyla. We used the existing annotation (the "CDS" keywords) to differentiate protein-coding and noncoding sequences.

The optimal growth temperature (OGT) and oxygen requirement classifications for each genome were obtained from the Genomes Online Database (http://www.genomesonline.org)(Pagani, Liolios et al. 2012). Among the 1424 complete genomes, 1378 genomes have the OGT classification available and 1304 genomes have the oxygen requirement classification available. To eliminate sampling biases towards genera represented by many completely sequenced genomes, we performed statistical assessments at the level of genera rather than individual genomes. Accordingly, we apply a single classification to

each genus, which is determined by the majority of species within the genus. This procedure yielded OGT classification for 498 genera (including 18 psychrophiles and psychrotolerant organisms, 382 mesophiles, 72 thermophiles, and 26 hyperthermophiles) and oxygen requirement classification for 465 genera (159 anaerobes, 202 aerobes, 95 facultative, and 9 microaerophilic organisms).

Patterns of interest

Table 2.1 provides a list of sequence patterns whose occurrences in prokaryotic genomes were investigated in this work. The patterns were selected based on the available information about sequences that promote various forms of structural transitions or mutations in DNA under favorable circumstances. However, because exact rules governing structural transitions in DNA are not fully understood we use multiple forms of similar patterns, in most cases pertaining to varying length of the pattern and number of tolerated mismatches. The specific parameters used in the definition of the patterns were in part dictated by the size of the analyzed genomes in order to obtain sufficient data sample for statistical evaluations.

It is worthwhile to note that we are not aiming to predict accurately every site in the genome that is likely to undergo a conformational transition under favorable conditions. We are only asking how common are sequences favoring certain structural transitions in each genome and whether their frequency could be indicative of selective constraints acting on such sequences. At the same time, the extensive simulations with randomized genomes described below require that the patterns are sufficiently simple that they can be identified quickly. Given this limitation, we selected the sequence patterns to be approximately representative of the sequences known to undergo the specific structural transitions under favorable conditions.

Evaluation of pattern representations

Pattern Locator (Mrazek and Xie 2006) was used to count loci in each analyzed sequence that matched the sequence pattern at hand (the observed count). The expected count of matching loci was determined as the average number of matching loci in 20 randomized sequences. The randomized sequences were generated by the "m1c1" model of the Genome Randomizer software previously developed in our laboratory (Mrázek 2006) (available for download at http://www.cmbl.uga.edu/software.html). In this model, the analyzed sequence is first divided into segments corresponding to annotated protein-coding genes and intergenic regions. For each intergenic segment, a random sequence of the same length is generated as the first order Markov chain, thus preserving the nucleotide and dinucleotide composition of that specific intergenic segment. Analogously, a random sequence is generated for each gene as a first order Markov chain using the codon alphabet. The final randomized genome is constructed by reassembling the randomized genes and intergenic segments in their original order. The resulting randomized sequence mimics not only the overall nucleotide, dinucleotide, and codon frequencies of the complete genome but rather of each individual gene and intergenic region. Thus the null model incorporates the compositional heterogeneity of the genomic sequence at the scale of individual genes. Because we factor the codon usage biases, dinucleotide preferences, and differences between protein-coding and noncoding sequences into the null model we are more likely to detect anomalous usage of sequence patterns that arises from direct selection on the sequence patterns as opposed to anomalous usage that is a simple consequence of codon or dinucleotide biases characteristic of the particular genome or its various segments. As a result, our assessments are more likely to reflect selective constraints and functional significance of the analyzed sequence pattern than assessments utilizing a simpler null model.

The pattern representations are classified into 9 levels, from -4 (extremely under-represented), through 0 (normally represented) to +4 (extremely over-represented) based on the p-value and the observed/expected ratio (Table 2.S1). We use the combination of the two criteria because the observed/expected ratio is independent of the sample size and measures the deviation from the null model in more absolute terms but does not directly provide assessment of statistical significance. Using p-value alone could emphasize small deviations in large samples (e.g., large genomes or more frequent patterns), whereas using the observed/expected ratio alone could lead to over-interpretation of results that lack statistical significance. Assigning representation categories based on both criteria is a compromise designed to avoid these potential pitfalls. The p-values are assessed based on an assumption that the counts of matching loci follow the Poisson distribution. This is a reasonable approximation for long sequences and patterns that occur at low frequencies.

Statistical assessments could be skewed by inclusion of observations that are not independent, such as several closely related genomes (e.g., different strains of the same species or closely related species), which are likely to feature similar levels of representations of various sequence patterns simply as a result of insufficient evolutionary divergence. To reduce the effect of dependent observations resulting from insufficient divergence we combined the results for all genomes of the same genus into a single ‘observation’ by averaging pattern representation categories for all available genomes belonging to that genus.

Results

Simple Sequence Repeats

A strong avoidance of tandem repeats of mono-, di-, and tri-nucleotides spreads over all phyla, especially in the whole genome and protein-coding regions, while tandem repeats of longer oligomers are generally normally represented and sometimes weakly overrepresented (Table 2.2). However, tandem repeats of mono-, di-, and tri-nucleotides are generally less suppressed in the intergenic regions when compared to protein-coding regions (Tables 2.S10 and 2.S11). *Trichodesmium erythraeum* and *Methanosphaera stadtmanae* are extreme examples with strongly overrepresented mono-nucleotide repeats 1n8 (a single nucleotide repeated ≥8 times) in the intergenic regions ($p<10^{-12}$), but normally represented in the whole genome and extremely underrepresented in the protein-coding regions. Most Chlorobi (6 species out of 10) also tend to have overrepresented 1n8 pattern in the intergenic regions. Tandem repeats of longer oligonucleotides (6-11 bp) are extremely overrepresented in intergenic regions as well as complete genomes of *Methanococcus voltae*, *Methanococcus aeolicus, Natrialba magadii* (and to a lesser extent in several other genomes) but not in the protein coding segments.

Mono- and dinucleotide repeats are about equally suppressed in psychrophiles, mesophiles, thermophiles and hyperthermophiles (Table 2.3). However, the suppression of trinucleotide tandem repeats is more pronounced in thermophiles and hyperthermophiles than in mesophiles and psychrophiles (significant at $p<10^{-5}$; the p-value is based on the Fisher's exact test for 2-by-2 contingency tables). Oxygen requirement appears to have no relationship to tandem repeat representations.

Close Direct Repeats

In general, close direct repeats are normally represented or slightly overrepresented in complete genomes of most prokaryotes and in the intergenic regions, and they are mostly normally represented in genes (Tables 2.S9-2.S11). The cd10g50 repeats (two copies of the same 10-mer separated by ≤ 50 nucleotides) exhibit the strongest trends among the investigated direct repeat structures. They are mostly overrepresented in the intergenic regions, weakly underrepresented in protein-coding regions, and normally represented in the whole genome. It is interesting to note the contrast in representations of close pairs of 10-bp repeats (cd10g50) and multiple close copies of shorter oligonucleotides (4n6g12, 6n6g24, 8n4g24). While the former are often strongly overrepresented in intergenic regions and normally represented to weakly underrepresented in genes, the latter are normally represented or moderately overrepresented in both genes and intergenic regions (Tables 2.S10 and 2.S11).

Representations of close direct repeats exhibit clear negative trends with increasing optimal growth temperature (OGT) (Table 2.3). For example, clustered repeats 8n4g24 (at least four copies of the same octamer separated by gaps of no more than 24 bp) are over-represented in about 50% of psychrophiles and mesophiles but in few thermophiles and virtually no hyperthermophiles (Figure 2.1; the difference is significant at $p<10^{-6}$). Significant decrease in pattern representations with increasing OGT is seen also for patterns 6n6g24 and8n4g24, indicating a general trend of lower local repetitiveness in thermophiles and hyperthermophiles. There is no clear relationship between the representation of close direct repeats and oxygen requirement.

Palindromes and Inverted Repeats

Palindromes and inverted repeats are strongly over-represented in the complete genomes across almost all bacterial phyla but less so in archaea and Aquificales (Table 2.2). There are few organisms that strongly avoid palindromes in protein-coding regions and the complete genomes, including cyanobacteria *Synechosystis* and *Thermosynechococcus*. Representations of palindromes and close inverted repeats in protein-coding sequences vary significantly among different genera, from extremely overrepresented (e.g., *Cyanobacterium*, *Thermoanaerobacter*, *Allivibrio*, *Azobacteroides*) to strongly suppressed (e.g., *Rhodomicrobium*, *Phenylobacterium*, *Jannaschia*, *Clavibacter*). Many α-proteobacteria exhibit some level of palindrome suppression. *Clavibacter* is an outlier among Actinobacteria, which generally tend to have palindromes normally represented and in some cases overrepresented.

A comparison between genes and intergenic regions shows that inverted repeats are almost invariably overrepresented in the intergenic regions and generally normally represented in the protein-coding regions. The high concentration of palindromes and inverted repeats in the intergenic region is not necessarily surprising because of their function in transcription termination and as regulatory elements, which are generally located outside the protein coding segments of genes.

There is a general trend of decreasing representations of palindromes and inverted repeats with increasing OGT (Figure 2.1, Table 2.3). In contrast, there is only a weak or no relationship between representations of palindromes and oxygen requirement.

Oligopurine/oligopyrimidine runs and triplex DNA-promoting patterns

H-DNA triplexes form under favorable conditions in oligopurine/oligopyrimidine sequences with a mirror symmetry (Belotserkovskii, Veselkov et al. 1990; Zain and Sun 2003)

whereas some oligopurine/oligopyrimidine segments can also promote formation A-like DNA (Sinden 1994). Mirror repeats and extended oligopurine/oligopyrimidine stretches are generally normally represented in complete genomes of most phyla (Table 2.2), with moderate overrepresentations in some Cyanobacteria (*Anabaena, Nostoc*, *Microcystis, Trichodesmium*), Chloroflexi (*Chloroflexus* and *Roseiflexus*), Planctomycetes (*Isosphaera*) and Verrucomicrobia (*Methylacidiphilum*). However, oligopurine and oligopyrimidine stretches, mainly patterns R15 and R30e3, are suppressed in complete genomes of some phyla. Separate analysis of protein-coding and noncoding regions shows that oligopurine/oligopyrimidine stretches tend to be normally represented in intergenic regions and that anomalous representations of these patterns in some phyla arise from biases in protein-coding segments (Tables 2.S10 and 2.S11).

Suppression of oligopurine/oligopyrimidine stretches, especially R15, is most common among hyperthermophiles whereas mesophiles and psychrophiles exhibit normal representations of these patterns (Table 2.3). The R15 pattern is also more likely to be underrepresented in anaerobes than in aerobes (Table 2.3 and Figure 2.S1). On the other hand, mirror repeats cm10g50 and mirs12g20 are more likely to be overrepresented in the coding regions of anaerobes than aerobes (Table 2.S4 and Figure 2.S2).

Guanine-rich patterns and G-DNA-promoting sequences

Formation of intrastrand G-DNA quadruplex is promoted by clustered short runs of guanine(Shafer and Smirnov 2000; Burge, Parkinson et al. 2006; Vorlíčková, Bednářová et al. 2007). We investigated several forms of such G-rich clusters (G-patterns; Table 2.1), among which the GGG4g6 pattern (four G-triplets separated by gaps of no more than 6 nucleotides) represents best the sequences known to form G-quadruplexes(Sinden 1994; Vorlíčková, Chládková et al. 2005; Lane, Chaires et al. 2008). In general, the G-patterns are normally

represented in prokaryotic genomes (Table 2.2). GG dimer clusters (GG8g4, eight or more GG dimers separated by ≤4 nucleotides from each other) tend to be moderately overrepresented in protein coding regions of α-, β- and δ-proteobacteria, Actinobacteria, Cyanobacteria, Deincoccus-Thermus group, and Planctomycetes (especially for *Isosphaera*) (Table 2.S2). The other guanine patterns are generally normally represented with most notable exceptions among Cyanobacteria where species of *Microcystis*, *Anabaena* and *Nostoc* feature extreme overrepresentation of all forms of G-patterns. The planctomycete *Isosphaera pallida* also has G-patterns strongly overrepresented. With these few exceptions, the GGG4g6 pattern, which is most directly related to G-DNA formation, is mostly normally represented or slightly underrepresented, suggesting that if G-DNA-promoting sequences have significant physiological roles in bacteria such roles are probably limited to only a few species or genera.

Most organisms with overrepresented GG8g4 pattern are aerobes whereas virtually no anaerobes have excess of GG8g4 (Figure 2.S1). However, overrepresented GG8g4 pattern could also reflect excess of glycine-rich segments (encoded by the GGN codons) or proline-rich segments (encoded by CCN with GG dinucleotides in the complementary strand) in proteins, whereas the GGG4g6 does not exhibit a relationship with oxygen requirement. There is no apparent relationship between G-pattern representations and the optimal growth temperature (Table 2.3).

Z-DNA-promoting patterns

The left-handed Z-DNA conformation is most commonly adopted by runs of alternating G-C and more generally by alternating purine-pyrimidine (RY) patterns. The RY patterns are often underrepresented but to different extent in different phyla – most strongly in α-, β- and γ-proteobacteria, Actinobacteria, Cyanobacteria, Chlorobi, Chloroflexi, and Planctomycetes (Table

2.2). In contrast, the RY patterns are normally represented in Fusobacteria, Aquificales and Thermotogae and even slightly overrepresented in Chlamydiae. In general, the suppression of RY patterns is stronger in protein-coding regions although several phyla exhibit significant RY pattern suppression in intergenic regions as well (Tables 2.S10 and 2.S11). Interestingly, *Treponema pallidum* and *Treponema paraluiscuniculi* exhibit a strong overrepresentation of all forms of RY patterns whereas other spirochaetes, including other species of *Treponema*, have RY patterns normally represented or weakly underrepresented. Similar extreme overrepresentation of RY patterns applies to *Helicobacter felis* and to a lesser extent to *Helicobacter bizzozeronii* but not to other *Helicobacter* species. The archaeon *Thermofilum pendens* also exhibits a strong overrepresentation of RY patterns and several other genomes show a weaker RY-pattern overrepresentation.

We used Pattern Locator and associated software tools(Mrazek and Xie 2006; Mrázek, Xie et al. 2008) (http://www.cmbl.uga.edu/software/patloc.html) to investigate in detail the distribution of RY patterns in several specific genomes, including the above-mentioned genomes with overrepresented RY patterns. We did not find any significant anomalies in the distribution of the RY patterns with respect to the origin or terminus of replication, with respect to the 3' ends of genes (stop codons), or any strong association with a particular class of genes. However, we noted increased numbers of RY patterns overlapping with start codons, such that the ATG/GTG start codons are embedded in RY patterns which sometimes extend several codons deep into the protein-coding region (Figure 2.S3).This tendency appears to be widespread among prokaryotic genomes. However, comparison between ATG and GTG triplets that function as start codons and those that are not translation start sites shows no significant difference in fractions of ATG/GTG triplets overlapping with extended RY patterns between the two groups

(Table 2.S9). This suggests that the increased number of RY patterns overlapping with translation start sites are not directly related to the translation but rather a simple consequence of ATG and GTG themselves having the form of a short RY pattern (RYR), thus increasing the likelihood of finding a longer RY pattern at the same site.

Representations of RY patterns tend to increase with increasing OGT (Table 2.3). For example, the pattern RY12 is under-represented in 60% of psychrophiles, 50% of mesophiles, about 25% of thermophiles but only a single hyperthermophilic genus, *Thermaerobacter* (Figure 2.1; significant at $p<10^{-6}$).

Among all analyzed sequence patterns, the RY patterns exhibit the most notable trend with respect to the level of oxygen requirement (Table 2.3). Interestingly, facultative species exhibit the strongest suppression of RY patterns followed by aerobes, whereas RY pattern representations in anaerobes and microaerophilic organisms are close to normal (see Figure 2.S1 (c) for pattern RY12). The same trend was observed when the comparisons between aerobes and anaerobes were restricted to mesophiles, or bacteria (Tables 2.S6 and 2.S17), indicating that the relationship between RY patterns and oxygen requirement cannot be attributed to increased number of anaerobes among thermophiles or different representations of anaerobes and aerobes among bacteria and archaea.

Patterns contributing to DNA curvature

Sequence patterns related to DNA curvature (the bend-patterns) range generally from normal to over-represented. The over-representation is most pronounced in ε-proteobacteria, Fusobacteria, Deferribacteres and Thermotogae (Table 2.2). This trend applies to both protein-coding and intergenic regions, although differences among phyla are more prominent in protein-coding regions (Tables 2.S9 and 2.S11). At the opposite extreme, *Mycoplasma haemofelis*,

Mycoplasma penetrans, Cytophaga hutchinsonii, and *Pelagibacter sp* show moderate to strong suppression of DNA bends in their genomes. The genus *Mycoplasma* is particularly interesting, featuring species with extreme overrepresentation of DNA bends as well as strong underrepresentation. This is consistent with our previous report that various genome properties among *Mycoplasma* vary more than in other genera(Mrázek 2006). In addition, *Cytophaga hutchinsonii* and *Pelagibacter* sp feature extreme suppression of bend-patterns in protein coding regions but not in intergenic regions.

With respect to OGT, intrinsic DNA bends tend to be more over-represented in hyperthermophiles and thermophiles than in mesophiles and psychrophiles (Table 2.3 and Figure 2.1). Surprisingly, the number of genera with overrepresented bend-patterns in protein-coding segments increases with increasing OGT in protein-coding regions but decreases for bend-patterns in non-coding regions (Figure 2.S4). The same trend is shown in Figure 2.S5, where genomes of thermophiles and hyperthermophiles tend to have more protein-coding bends than intergenic bends, whereas the opposite is true for most psychrophiles. Statistical significance of this trend has been confirmed by Mann-Whitney U test (Table 2.S10).

With respect to oxygen requirement, representation of DNA bending patterns tends to decrease with the increasing level of oxygen (Table 2.3). Specifically, the bend-patterns tend to be overrepresented in anaerobes and microaerophiles but not in aerobes. For example, the pattern bend60w100 is over represented in half of the anaerobes, but less than 20% of aerobes (Figure 2.S1; significant at $p<10^{-6}$). This trend holds when the data are restricted to mesophiles or thermophiles, suggesting that the trend with respect to oxygen requirement is independent of the trend with respect to OGT (Tables 2.S6 and 2.S7).

Discussion

Comparisons with previous work

The work presented here differs from similar earlier surveys in scope (both the number of genomes used in the analysis and the number of different types of sequence patterns investigated) as well as methodology. The most important methodological difference is in the null model used to assess whether a sequence pattern is anomalously represented in a given genome. Our null model takes into account the nearest-neighbor biases and codon usage propensities of all individual genes and intergenic regions, which likely have separate underlying causes not related to potential functional significance of the investigated sequence patterns. In terms of our results, one significant difference relative to earlier work concerns representations of alternating purine-pyrimidine patterns. Bolin *et al.* (Bohlin, Hardy et al. 2009) reported that alternating RY patterns were generally suppressed across different phyla except β-proteobactria, where they were mostly overrepresented, with a particularly strong surplus of RY patterns in *Burkholderia*. These authors used the i.i.d. model (independently drawn and independently distributed letters) as a benchmark. Using our more realistic null model, we found that β-proteobactria including *Burkholderia* species suppress the RY patterns to the same extent as other bacteria (Tables 2.2 and 2. S9). We therefore conclude that the increased amount of RY patterns in β-proteobactria arises from two opposite evolutionary constraints: biased codon and dinucleotide usages that favor RY-patterns and are specific for β-proteobactria, which are partially offset by suppression of long RY-patterns, which is rather universal among bacterial genomes. Similarly, the overrepresentation of oligopurine/oligopyrimidine stretches reported by Bolin *et al.* for some phyla arises from the biases at the level of dinucleotide and codon usages, as we found the

oligopurine/oligopyrimidine stretches normally represented or underrepresented in all phyla (Table 2.2).

For simple sequence repeats (SSR), our results are consistent with previous works including our own survey of SSRs in prokaryotic genomes (Mrázek, Guo et al. 2007). In particular, our present data confirm that SSR comprised of tandem repeats of very short units (mono-, di-, and trinucleotides; patterns 1n8, 2n5, and 3n4) tend to be strongly suppressed in prokaryotic genomes while repeats of longer units (tetranucleotides through 11-mers) tend to be normally represented or weakly overrepresented (Table 2.2). This strong difference points to likely functional difference between tandem repeats of very short (1-3 bp) and longer (4-11 bp) units. Our present results differ from the earlier data in that we previously included repeats of tetranucleotides in the same group as mono-, di, and trinucleotides (Mrázek, Guo et al. 2007), while the present results indicate that tetranucleotide SSRs may be more accurately included in the same group as SSRs composed of pentanucleotides and longer units, which are generally not suppressed. This result suggests that tandem repeats of tetranucleotides and longer units may not have the same harmful effects as repeats of mono-, di-, and trinucleotides. The difference between tandem repeats of very short units (1-3 bp) and longer units (≥4 bp) could also arise from properties of the methyl-directed repair pathway, which is very efficient in repairing heterologous loops of 1-3 bp but its efficiency dramatically drops as the length of the loop increases(Parker and Marinus 1992; Fang, Wu et al. 1997).

Ladoukakis and Eyre-Walker (Ladoukakis and Eyre-Walker 2008) reported a small but significant excess of short inverted repeats (6-9 bp) in protein coding sequences and proffered that the inverted repeats could arise by sequence-directed mutagenesis where an imperfect palindrome or inverted repeat converts to a perfect one due to template switching during

replication (Lovett 2004). Their result is consistent with that of Katz and Burge (Katz and Burge 2003), who found that native mRNA sequences in many bacterial genomes possess a significantly higher potential to form stable RNA secondary structures than random sequences preserving the codon usage, dinucleotide frequencies and the protein sequences encoded by the native mRNAs. Katz and Burge proposed that the excess of base pairing in mRNA molecules is due to selection for stable mRNAs. We analyzed occurrences of slightly larger palindromes than Ladoukakis and Eyre-Walker but with a similar result that palindromes and close inverted repeats are strongly overrepresented in both protein-coding and noncoding regions of genomes of most prokaryotic phyla.

The excess of palindromes is weaker in the archaea and in the Aquificae than in other bacterial phyla. This result is similar to that of Katz and Burge, who reported a significant excess of mRNA secondary structures in most bacteria but few archaea, and not in *Aquifex aeolicus* (note that their study was based on a much smaller sample of genomes available at that time and *A. aeolicus* was a sole representative of Aquificae in their data set). The weaker overrepresentation of palindromes in archaea and Aquificae is related to general decrease of palindrome representations with increasing optimal growth temperature (Figure 2.1). The decrease in palindrome representations with increasing OGT might seem counterintuitive if selection for mRNA stability drives the excess of palindromes in mRNA sequences. One possible explanation is that higher temperatures prevent formation of stable mRNA structures even with an increased content of inverted repeats, thus making selection for inverted repeats moot, or perhaps higher temperature decreases the efficiency of palindrome conversion via sequence-directed mutagenesis. Yet another potential explanation for weaker overrepresentation of inverted repeats in thermophiles stems from the work by Paz *et al*. (Paz, Mester et al. 2004),

who noted that thermophilic mRNA sequences have increased purine-to-pyrimidine ratios compared to mesophiles, as well as excess of short oligopurine tracts. These authors proposed that purine-loading of mRNA sequences could increase their thermal stability. The resulting purine-pyrimidine imbalance could decrease the base-pairing potential in the mRNA sequences and the purine bias could therefore contribute to lower excess of palindromes and inverted repeats in thermophiles.

Relationship with optimal growth temperature (OGT) and oxygen requirement

Extreme temperature as well as presence of oxygen (via reactive oxygen species) cause damage to DNA and require cellular mechanisms to prevent or correct the damage in order to keep the cells viable. Such protection can be enzymatic (e.g., detoxification pathways) but could also involve adaptations in DNA composition and structure. One particularly puzzling question is how thermophiles prevent their DNA from denaturing. The simple explanation that DNA of thermophile is more GC-rich and thus more stable was mostly rejected as more data became available (Haney, Badger et al. 1999; Suhre and Claverie 2003). Kawashima and coworkers(Kawashima, Amano et al. 2000) proffered excess of RR and YY dinucleotides over RY and YR dinucleotides as a characteristic of thermophiles but this has later been shown a poor indicator of thermophily(Suhre and Claverie 2003). We were therefore interested in finding whether any of the sequence patterns we analyzed could be related to differences in OGT and oxygen requirement, and possibly contribute to the protection of DNA from the effects of extreme temperature and/or oxygen damage.

One of our more surprising results concerns the relationship of representations of intrinsic DNA bends with OGT. Bolshoy and coworkers (Bolshoy and Nevo 2000; Kozobay-Avraham, Hosid et al. 2006) reported an excess of intrinsically bent DNA in intergenic regions of

prokaryotic genomes, including segments containing putative promoters and transcription terminators. Notably, the amount of bent DNA in intergenic regions was significantly weaker in thermophiles than in mesophiles. In a separate work, Tolstorukov *at al*. (Tolstorukov, Virnik et al. 2005) found intrinsic DNA bends distributed widely throughout bacterial genomes, including both intergenic and protein-coding regions, and they proposed that the intrinsic bends could play a role in compaction of the bacterial nucleoid. Our present data show that predicted intrinsic DNA bends are overrepresented in many phyla in both protein-coding and noncoding segments. Interestingly, the protein-coding and noncoding regions exhibit opposite trends with respect to OGT. For noncoding sequences, our results are consistent with the decrease of curved DNA in thermophiles observed by Kozobay-Avraham *et al*. (Kozobay-Avraham, Hosid et al. 2006). However, the representation of DNA bends in protein-coding sequences tends to increase with increasing OGT.

We propose that the opposite trends in protein-coding and noncoding DNA segments reflect different roles of DNA bends. Intergenic bends are often associated with regulatory elements and could facilitate opening of the DNA double helix at the transcription initiation sites (Jauregui, Abreu-Goodger et al. 2003; Olivares-Zavaleta, Jauregui et al. 2006). Decreased amount of intrinsic DNA bends in intergenic regions of thermophiles is consistent with experiments that showed that anomalous gel mobility characteristic of intrinsically curved DNA disappears at increased temperatures (Ussery, Higgins et al. 1999). If sequence-directed intrinsic DNA bending is ineffective at high temperatures then selection for intrinsic bends at regulatory sites could be weak or absent. It is also possible that DNA bending is less important to transcription initiation at high temperatures. Surprisingly, the excess content of intragenic bends, which are less likely to have direct regulatory functions but may contribute to establishing and

maintaining the nucleoid structure, increases with increasing OGT. The latter result could possibly indicate that the intrinsic bends maintain their role in stabilizing the nucleoid structure even at high temperatures and that high OGT may require increased amount of intrinsic DNA bending to maintain adequate structural stability of the chromosome.

In addition to palindromes and close inverted repeats discussed above, some direct repeat structures also feature decreasing representations with increasing OGT (Table 2.3 and Figure 2.1). We speculate that higher temperature might make the thermophile DNA more susceptible to illegitimate recombination, DNA polymerase slippage, or other forms of mutations facilitated by short repeats, and that reducing the amount of repeats, both direct and inverted, could be a strategy to counteract increased recombination rates. Along these lines, thermophiles were reported to have lower mutation rates than mesophiles, possibly driven by selection to maintain thermostability of the encoded proteins(Drake 2009). Because close repeats are often sources of mutations, suppression of close repeats could be a part of the strategy to decrease overall mutation rates in thermophiles.

Alternating RY patterns exhibit trends related to both OGT and oxygen requirement and these trends are independent of each other. The relationship with oxygen requirement is particularly intriguing. Facultative organisms suppress RY patterns more strongly than both aerobes and anaerobes. The main known structural effect of RY patterns is that they can facilitate transitions to left-handed Z-DNA under favorable conditions. The B-to-Z transition in such sequences can be induced by torsional stress arising from processes that require DNA unwinding, such as transcription or replication (van Holde and Zlatanova 1994; Wang and Vasquez 2007). The general suppression of RY patterns in prokaryotes suggests that B-to-Z DNA transitions are generally undesirable. It is intriguing to speculate that Z-DNA could be more detrimental to

facultative organisms perhaps due to interference with effective regulation of a diverse ensemble of metabolic pathways related to the ability to grow in both aerobic and anaerobic conditions. However, the specific mechanism of such interference between Z-DNA formation and transcriptional regulation is not clear.

Overall, our data indicate potential new mechanisms how genome properties adapt to particular environments. Most of these mechanisms are related to adaptations to growth at high temperatures, which appear to be accompanied by reduction of overall repetitiveness of the DNA sequence, reduced excess of palindromes, and reduced DNA curvature in regulatory regions accompanied by general increase of intrinsically curved DNA in protein-coding sequences.

Acknowledgements

We wish to thank Dr. Barny Whitman and Hao Tong for stimulating suggestions, and Orion Mantione for his help in writing the program to predict DNA bends. This work is supported by the grant number DBI-0950266 from National Science Foundation.

Supplementary data

Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

References

Arthanari, H. and P. H. Bolton (2001). "Functional and dysfunctional roles of quadruplex DNA in cells." Chem Biol **8**(3): 221-230.

Belotserkovskii, B. P., A. G. Veselkov, et al. (1990). "Formation of intramolecular triplex in homopurine-homopyrimidine mirror repeats with point substitutions." Nucleic Acids Res **18**(22): 6621-6624.

Bohlin, J., S. P. Hardy, et al. (2009). "Stretches of alternating pyrimidine/purines and purines are respectively linked with pathogenicity and growth temperature in prokaryotes." Bmc Genomics **10**.

Bolshoy, A. and E. Nevo (2000). "Ecologic genomics of DNA: upstream bending in prokaryotic promoters." Genome Res **10**(8): 1185-1193.

Brazda, V., R. C. Laister, et al. (2011). "Cruciform structures are a common DNA feature important for regulating biological processes." BMC Mol Biol **12**: 33.

Burge, S., G. N. Parkinson, et al. (2006). "Quadruplex DNA: sequence, topology and structure." Nucleic Acids Res **34**(19): 5402-5415.

Chuzhanova, N., S. S. Abeysinghe, et al. (2003). "Translocation and gross deletion breakpoints in human inherited disease and cancer II: Potential involvement of repetitive sequence elements in secondary structure formation between DNA ends." Hum Mutat **22**(3): 245-251.

Dillon, S. C. and C. J. Dorman (2010). "Bacterial nucleoid-associated proteins, nucleoid structure and gene expression." Nat Rev Microbiol **8**(3): 185-195.

Drake, J. W. (2009). "Avoiding dangerous missense: thermophiles display especially low mutation rates." PLoS Genet **5**(6): e1000520.

Dutra, B. E. and S. T. Lovett (2006). "Cis and trans-acting effects on a mutational hotspot involving a replication template switch." J Mol Biol **356**(2): 300-311.

Fang, W., J. Y. Wu, et al. (1997). "Methyl-directed repair of mismatched small heterologous sequences in cell extracts from Escherichia coli." J Biol Chem **272**(36): 22714-22720.

Haney, P. J., J. H. Badger, et al. (1999). "Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species." Proc Natl Acad Sci U S A **96**(7): 3578-3583.

Herbert, A. and A. Rich (1999). "Left-handed Z-DNA: structure and function." Genetica **106**(1-2): 37-47.

Herzel, H., O. Weiss, et al. (1998). "Sequence periodicity in complete genomes of archaea suggests positive supercoiling." J Biomol Struct Dyn **16**(2): 341-345.

Jain, A., G. Wang, et al. (2008). "DNA triple helices: biological consequences and therapeutic potential." Biochimie **90**(8): 1117-1130.

Jauregui, R., C. Abreu-Goodger, et al. (2003). "Conservation of DNA curvature signals in regulatory regions of prokaryotic genes." Nucleic Acids Res **31**(23): 6770-6777.

Katz, L. and C. B. Burge (2003). "Widespread selection for local RNA secondary structure in coding regions of bacterial genes." Genome Res **13**(9): 2042-2051.

Kawashima, T., N. Amano, et al. (2000). "Archaeal adaptation to higher temperatures revealed by genomic sequence of Thermoplasma volcanium." Proc Natl Acad Sci U S A **97**(26): 14257-14262.

Kozobay-Avraham, L., S. Hosid, et al. (2006). "Involvement of DNA curvature in intergenic regions of prokaryotes." Nucleic Acids Res **34**(8): 2316-2327.

Ladoukakis, E. D. and A. Eyre-Walker (2008). "The excess of small inverted repeats in prokaryotes." Journal of Molecular Evolution **67**(3): 291-300.

Lane, A. N., J. B. Chaires, et al. (2008). "Stability and kinetics of G-quadruplex structures." Nucleic Acids Res **36**(17): 5482-5515.

Levinson, G. and G. A. Gutman (1987). "Slipped-strand mispairing: a major mechanism for DNA sequence evolution." Molecular Biology and Evolution **4**(3): 203-221.

Lovett, S. T. (2004). "Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences." Mol Microbiol **52**(5): 1243-1253.

Moxon, E. R., P. B. Rainey, et al. (1994). "Adaptive evolution of highly mutable loci in pathogenic bacteria." Curr Biol **4**(1): 24-33.

Mrázek, J. (2006). "Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes." Molecular Biology and Evolution **23**(7): 1370-1385.

Mrázek, J. (2010). "Comparative analysis of sequence periodicity among prokaryotic genomes points to differences in nucleoid structure and a relationship to gene expression." J Bacteriol **192**(14): 3763-3772.

Mrázek, J., X. X. Guo, et al. (2007). "Simple sequence repeats in prokaryotic genomes." Proc Natl Acad Sci U S A **104**(20): 8472-8477.

Mrázek, J., S. Xie, et al. (2008). "AIMIE: a web-based environment for detection and interpretation of significant sequence motifs in prokaryotic genomes." Bioinformatics **24**(8): 1041-1048.

Mrazek, J. and S. H. Xie (2006). "Pattern locator: a new tool for finding local sequence patterns in genomic DNA sequences." Bioinformatics **22**(24): 3099-3100.

Olivares-Zavaleta, N., R. Jauregui, et al. (2006). "Genome analysis of Escherichia coli promoter sequences evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated." Genomics **87**(3): 329-337.

Pagani, I., K. Liolios, et al. (2012). "The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata." Nucleic Acids Res **40**(Database issue): D571-579.

Pan, B., K. Shi, et al. (2006). "Base-tetrad swapping results in dimerization of RNA quadruplexes: implications for formation of the i-motif RNA octaplex." Proc Natl Acad Sci U S A **103**(9): 3130-3134.

Parker, B. O. and M. G. Marinus (1992). "Repair of DNA heteroduplexes containing small heterologous sequences in Escherichia coli." Proc Natl Acad Sci U S A **89**(5): 1730-1734.

Paz, A., D. Mester, et al. (2004). "Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes." Proc Natl Acad Sci U S A **101**(9): 2951-2956.

Rawal, P., V. B. Kummarasetti, et al. (2006). "Genome-wide prediction of G4 DNA as regulatory motifs: role in *Escherichia coli* global regulation." Genome Res **16**(5): 644-655.

Rich, A., A. Nordheim, et al. (1983). "Stabilization and detection of natural left-handed Z-DNA." J Biomol Struct Dyn **1**(1): 1-19.

Rich, A. and S. Zhang (2003). "Timeline: Z-DNA: the long road to biological function." Nat Rev Genet **4**(7): 566-572.

Rustighi, A., M. A. Tessari, et al. (2002). "A polypyrimidine/polypurine tract within the Hmga2 minimal promoter: a common feature of many growth-related genes." Biochemistry **41**(4): 1229-1240.

Schaeffer, C., B. Bardoni, et al. (2001). "The fragile X mental retardation protein binds specifically to its mRNA via a purine quartet motif." EMBO J **20**(17): 4803-4813.

Shafer, R. H. and I. Smirnov (2000). "Biological aspects of DNA/RNA quadruplexes." Biopolymers **56**(3): 209-227.

Sinden, R. R. (1994). DNA structure and function. San Diego, Academic Press.

Suhre, K. and J. M. Claverie (2003). "Genomic correlates of hyperthermostability, an update." J Biol Chem **278**(19): 17198-17202.

Sundquist, W. I. and S. Heaphy (1993). "Evidence for interstrand quadruplex formation in the dimerization of human immunodeficiency virus 1 genomic RNA." Proc Natl Acad Sci U S A **90**(8): 3393-3397.

Thanbichler, M., S. C. Wang, et al. (2005). "The bacterial nucleoid: a highly organized and dynamic structure." J Cell Biochem **96**(3): 506-521.

Tolstorukov, M. Y., K. M. Virnik, et al. (2005). "A-tract clusters may facilitate DNA packaging in bacterial nucleoid." Nucleic Acids Res **33**(12): 3907-3918.

Ussery, D., D. M. Soumpasis, et al. (2002). "Bias of purine stretches in sequenced chromosomes." Comput Chem **26**(5): 531-541.

Ussery, D. W., C. F. Higgins, et al. (1999). "Environmental influences on DNA curvature." J Biomol Struct Dyn **16**(4): 811-823.

van der Woude, M. W. and A. J. Baumler (2004). "Phase and antigenic variation in bacteria." Clin Microbiol Rev **17**(3): 581-611, table of contents.

van Holde, K. and J. Zlatanova (1994). "Unusual DNA structures, chromatin and transcription." Bioessays **16**(1): 59-68.

Vorlíčková, M., K. Bednářová, et al. (2007). "Intramolecular and intermolecular guanine quadruplexes of DNA in aqueous salt and ethanol solutions." Biopolymers **86**(1): 1-10.

Vorlíčková, M., J. Chládková, et al. (2005). "Guanine tetraplex topology of human telomere DNA is governed by the number of (TTAGGG) repeats." Nucleic Acids Res **33**(18): 5851-5860.

Wang, G. and K. M. Vasquez (2007). "Z-DNA, an active element in the genome." Front Biosci **12**: 4424-4438.

Zain, R. and J. S. Sun (2003). "Do natural DNA triple-helical structures occur and function in vivo?" Cell Mol Life Sci **60**(5): 862-870.

Table 2.1 List of sequence patterns investigated in this work

Pattern	Code	Meaning	Example [a]
Simple sequence repeats	1n8	A single nucleotide repeated 8+ times in a row	GCAAAAAAAAATA
	2n5	A dinucleotide repeated 5+ times in a row	ACCACACACACATA
	3n4	Analogous to the two examples above	
	4n4		
	5n4		
	6n3		
	7n3		
	8n3		
	9n2		
	10n2		
Close direct repeats	4n6g12	A tetranucleotide repeated 6+ times with gaps ≤12 nt	TACCATGCTCCATTACCATAGCCAT…
	6n6g24	A 6-mer repeated 6+ times with gaps ≤24 nt	
	8n4g24	An 8-mer repeated 4+ times with gaps ≤24 nt	
	cd8g6	An 8-mer repeated within 6 nt	CTTAGGCATCACCTTAGGCA
	cd10g50	A 10-mer repeated within 50 nt	
Palindromes & inverted repeats	cp8g6	Inverted repeat of an 8-mer separated by no more than 6 nt	CTTAGGCATCACTGCCTAAG
	cp10g50	Inverted repeat of a 10-mer separated by ≤50 nt	
	pals9	9 nt inverted repeat (no separation) OR 12 nt inverted repeat allowing 1 mismatch OR 15 nt inverted repeat allowing 2 mismatches OR … (one mismatch added for every 3 nt length)	CTGGATCAGGCTAAA⁞TTCAGCCTCATCCAG
	pals9g12	Like pals9 but allowing separation up to 12 bp	
	pals12g20	Analogous to the example above	
H-DNA-related patterns	cm8g6	Mirror repeat of an 8-mer separated by ≤6 nt	CTTAGGCATCACACGGATTC
	cm10g50	Mirror repeat of a 10-mer separated by ≤50 nt	
	mirs9	9 nt mirror repeat (no separation) OR 12 nt mirror repeat allowing 1 mismatch OR 15 nt mirror repeat allowing 2 mismatches OR … (one mismatch added for every 3 nt length)	CTGGATCAGGCTAAA⁞AACTCGGACTGGGTC
	mirs9g12	Like mirs9 but allowing separation up to 12 bp	
	mirs12g20	Analogous to mirs9g12	
	R15	Run of ≥15 purines or pyrimidines	AAGGGAGGGAGGAGA
	R30	Run of ≥30 purines or pyrimidines	
	R30e3	Run of ≥30 purines or pyrimidines allowing ≤3 errors	
	R45e6	Analogous to the example above	
	R60e9		
G-DNA-related patterns	GG8g4	8 or more GG dimers separated by ≤4 nt from each other	GGAGGCTGGCGGGGCGGTGGGG
	GGG4g6	Analogous to the example above	
	GGGG4g6		
Z-DNA-related patterns	GC6	Alternating G-C, ≥6 nt length	GCGCGC
	GC8	Alternating G-C, ≥8 nt length	
	RY12	Alternating R-Y, ≥12 nt length	TGTACGTGTGCA
	RY12e1	Like RY12 but allowing 1 error	TGTACGAGTGCA
	RY18e2	Alternating R-Y, ≥18 nt length, ≤2 errors	
	RY24e3	Alternating R-Y, ≥24 nt length, ≤3 errors	
DNA bending	bend45w60	Predicted bend of ≥45° within a ≤60 bp segment	
	bend60w100	Predicted bend of ≥60° within a ≤100 bp segment	
	bend90w120	Predicted bend of ≥90° within a ≤120 bp segment	

[a] Segments matching the sequence pattern are underscored, mismatches are shaded and symmetrical segments are separated by a vertical dashed line.

Table 2.2 Representation of sequence patterns in different phyla

Pattern name	Pattern code	AlPr	BePr	GaPr	DePr	EpPr	Firm	Acti	Cyan	Bact	Chlb	Chlf	Dein	Fuso	Chla	Spir	Acid	Verr	Defe	Plan	Aqui	Ther	Eury	Cren
		58	39	76	23	11	61	63	12	34	5	8	6	5	6	4	4	4	4	4	8	5	44	16
Simple sequence repeats	1n8	-4.00	-3.00	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	-1.67	-4.00	-4.00	-4.00	-3.79	-3.75	-3.00	-3.00	-4.00	-3.50	-4.00	-4.00	-4.00	-4.00
	2n5	-2.21	-2.00	-3.00	-3.00	-4.00	-4.00	-3.00	-3.63	-4.00	-3.00	-3.00	-2.60	-4.00	-4.00	-3.50	-2.00	-2.50	-4.00	-2.00	-4.00	-4.00	-3.00	-3.00
	3n4	-0.78	-2.00	-2.09	-1.00	-3.00	-3.00	-2.00	-1.50	-2.67	-2.00	-1.50	-1.50	-4.00	-3.00	-3.17	-0.25	-1.00	-3.50	-1.00	-3.50	-4.00	-3.00	-3.00
	4n4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5n4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	6n3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	7n3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.50	1.00	2.00	0.00	0.00	0.00	0.00
	8n3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00
	9n2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	10n2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
	11n2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.59	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Close direct repeats	4n6g12	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	1.25	0.00	0.00	0.00	0.00
	6n6g24	0.42	0.25	0.88	1.00	0.00	0.30	2.00	1.50	1.00	1.00	1.00	0.00	0.00	0.00	0.98	1.50	0.00	1.50	1.25	0.00	0.00	0.10	0.00
	8n4g24	0.50	1.00	1.14	2.00	0.00	1.00	3.00	3.00	1.50	2.00	1.00	0.00	0.00	0.00	1.25	2.25	0.50	1.50	2.75	0.00	0.00	1.00	0.00
	cd8g6	0.00	0.00	0.00	1.00	1.00	0.00	2.00	1.92	0.00	0.00	0.75	1.20	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	1.00	0.60	1.00
	cd10g50	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.60	0.00	0.00	0.00	0.00	0.00	-0.19	0.15	0.00	0.00	-1.00	1.00	0.00	0.00	0.00	0.00
Palindromes & inverted repeats	cp8g6	3.00	3.00	4.00	3.50	3.00	4.00	3.00	1.81	3.00	2.00	4.00	2.63	3.00	4.00	2.86	3.50	4.00	3.00	1.50	2.00	4.00	2.00	2.00
	cp10g50	3.00	4.00	4.00	4.00	3.83	4.00	4.00	3.54	4.00	4.00	4.00	3.75	4.00	4.00	3.69	3.50	2.00	3.00	3.50	1.00	4.00	1.00	1.00
	pals9	1.09	2.00	3.35	3.00	2.00	4.00	4.00	2.00	4.00	3.00	3.50	3.00	3.00	3.96	2.17	1.50	2.00	1.50	1.00	0.00	3.50	0.10	0.00
	pals9g12	3.66	4.00	4.00	4.00	3.72	4.00	4.00	3.71	4.00	4.00	4.00	3.75	4.00	4.00	3.77	4.00	4.00	3.00	4.00	1.00	4.00	2.00	1.00
	pals12g20	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	3.77	4.00	4.00	3.00	4.00	0.00	4.00	1.00	0.22
H-DNA-related patterns	cm8g6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cm10g50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.34	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mirs9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mirs9g12	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.17	0.00	0.00	0.34	0.00	0.00	0.00	0.25	0.50	0.00	0.00	0.25	0.00	0.00	0.00	0.00
	mirs12g20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.50	0.00	0.25	0.00	0.00	0.00	0.00
	R15	-0.20	0.00	0.00	0.00	-1.00	-0.20	0.00	-1.07	0.00	0.00	0.25	-1.00	-4.00	0.00	-0.90	-0.50	1.00	-3.00	0.00	-2.00	-3.00	-1.00	-1.00
	R30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R30e3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-2.00	-0.25	-0.13	0.00	0.50	-1.00	0.00	-0.50	-1.50	0.00	0.00
	R45e6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-2.00	0.00	0.15	0.00	0.50	0.00	0.50	-0.50	0.00	0.00	0.00
	R60e9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00

Table 2.2 (continued) Representation of sequence patterns in different phyla

Pattern name	Pattern code	AlPr	BePr	GaPr	DePr	EpPr	Firm	Acti	Cyan	Bact	Chlb	Chlf	Dein	Fuso	Chla	Spir	Acid	Verr	Defe	Plan	Aqui	Ther	Eury	Cren
		58	39	76	23	11	61	63	12	34	5	8	6	5	6	4	4	4	4	4	8	5	44	16
G-DNA-related patterns	GG8g4	1.00	1.33	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	2.70	0.00	0.00	0.54	1.00	0.00	0.00	1.50	0.00	0.00	0.00	0.00
	GGG4g6	0.00	0.00	0.00	0.00	0.00	-0.63	0.00	0.00	0.00	0.00	-0.31	1.00	0.00	0.00	-0.25	0.00	0.00	0.00	0.00	-0.50	0.00	-0.42	0.00
	GGGG4g6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Z-DNA-related patterns	GC6	-1.00	-1.00	-2.00	-2.00	-1.00	-1.00	-2.00	-2.00	-1.00	-4.00	-2.00	-1.20	0.00	0.00	-0.77	-1.50	-1.00	-0.50	-2.75	-0.25	0.00	-2.00	-1.00
	GC8	-0.50	-1.00	-1.00	-1.00	0.00	0.00	-3.00	-0.25	0.00	-2.00	-1.34	-1.25	0.00	0.00	0.00	-0.50	0.00	0.00	-3.00	0.00	0.00	0.00	0.00
	RY12	-2.50	-2.00	-3.00	-1.00	-0.89	-0.33	-2.00	-2.00	-0.35	-2.00	-3.25	-0.25	0.00	0.42	-0.09	-0.75	0.00	-1.00	-3.00	0.00	0.00	-0.72	0.00
	RY12e1	-2.00	-1.40	-3.00	-1.00	-1.00	-0.25	-2.00	-3.00	0.00	-3.00	-2.25	0.00	0.00	0.84	0.05	-0.75	0.00	-1.00	-3.00	0.50	0.00	-0.64	-0.98
	RY18e2	-2.65	-2.00	-2.53	-1.00	0.00	0.00	-2.00	-2.00	0.00	-2.00	-3.00	-1.00	0.00	0.75	0.00	-0.75	0.00	0.00	-3.50	0.00	0.00	0.00	0.00
	RY24e3	-1.03	-1.50	-1.00	0.00	0.00	0.00	-1.00	-1.00	0.00	-1.00	-1.25	0.00	0.00	0.00	-0.17	-1.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00
DNA Bending	bend45w60	0.28	0.00	0.07	1.00	3.00	2.00	0.00	1.50	2.00	1.00	0.17	0.45	2.00	1.50	0.50	0.00	0.00	2.50	0.00	2.00	2.00	2.00	0.00
	bend60w100	0.00	0.00	0.00	0.50	2.97	2.00	0.00	1.00	2.00	0.00	0.00	0.00	2.00	0.75	0.25	0.00	0.00	3.00	0.00	2.25	3.00	2.00	0.00
	bend90w120	0.00	0.00	0.00	0.00	3.00	1.00	0.00	1.00	1.00	0.00	0.00	0.00	3.00	0.00	0.20	0.00	0.00	3.50	0.00	2.00	3.00	1.00	0.00

Numbers in the table refer to the medians of pattern representation among all genera of the corresponding phylum. The pattern representations were categorized into 9 categories from -4 (extremely under-represented), through 0 (normally represented), to +4 (extremely over-represented). See Materials and Methods for details. Codes in the second column refer to specific sequence patterns (see Table 2.1). Columns represent different phyla abbreviated as follows: AlPr, α-proteobacteria; BePr, β-Proteobacteria; GaPr, γ-Proteobacteria; DePr, δ-Proteobacteria; EpPr, ε-Proteobacteria; Firm, Firmicutes; Acti, Actinobacteria; Cyan, Cyanobacteria; Bact, Bacteroidetes; Chlb, Chlorobi; Chlf, Chloroflexi; Dein, Deinococcus-Thermus; Fuso, Fusobacteria; Chla, Chlamydiae; Spir, Spirochaetes; Acid, Acidobacteria; Verr, Verrucomicrobia; Defe, Deferribacteres; Plan, Planctomycetes; Aqui, Aquificales; Ther, Thermotogae; Eury, Euryarchaeota; Cren, Crenarchaeota. Numbers in the second row indicate the number of genera available for each phylum. Only phyla represented by three or more genera are shown.

Table 2.3 Representation of sequence patterns in different OGT and oxygen requirement classes

Pattern name	Pattern code	Psychrophile	Mesophile	Thermophile	Hyperthermophile	Anaerobe	Aerobe	Facultative	Microaerophile
		18	382	72	26	159	202	95	9
Simple sequence repeats	1n8	-3.83	-3.54	-3.89	-4.00	-3.73	-3.61	-3.65	-3.67
	2n5	-3.15	-2.89	-3.40	-3.45	-3.30	-2.74	-3.02	-3.11
	3n4	-1.93	-1.92	-2.83	-3.09	-2.56	-1.75	-2.15	-2.33
	4n4	-0.06	0.01	-0.06	-0.08	-0.04	0.01	0.00	-0.10
	5n4	0.00	0.07	0.00	0.00	0.02	0.10	0.01	0.11
	6n3	0.20	0.19	0.06	-0.03	0.05	0.25	0.08	0.31
	7n3	1.27	0.48	0.45	0.00	0.42	0.59	0.35	0.89
	8n3	0.56	0.30	0.12	0.01	0.25	0.27	0.20	0.33
	9n2	-0.08	0.09	0.22	0.04	0.09	0.19	-0.11	-0.11
	10n2	0.09	0.25	0.25	0.28	0.27	0.24	0.16	0.00
	11n2	0.21	0.32	0.30	0.40	0.42	0.26	0.24	0.03
Close direct repeats	4n6g12	0.45	0.58	0.48	0.82	0.55	0.73	0.31	0.50
	6n6g24	1.35	1.16	0.71	0.34	1.00	1.21	0.81	1.39
	8n4g24	2.02	1.61	0.80	0.28	1.27	1.69	1.18	1.13
	cd8g6	0.39	0.73	0.80	1.20	0.87	0.82	0.41	0.70
	cd10g50	0.46	0.41	0.05	-0.07	0.25	0.60	0.01	0.07
Palindromes & inverted repeats	cp8g6	3.57	2.78	2.81	1.99	2.98	2.39	3.21	3.22
	cp10g50	3.56	3.26	2.98	1.64	3.15	3.03	3.45	3.20
	pals9	3.39	2.58	2.13	0.71	2.47	2.27	2.89	2.47
	pals9g12	3.83	3.46	3.05	1.79	3.33	3.22	3.55	3.64
	pals12g20	3.72	3.58	2.97	1.17	3.16	3.47	3.62	3.00
H-DNA-related patterns	cm8g6	0.02	0.20	0.21	0.33	0.24	0.24	0.15	0.04
	cm10g50	0.02	0.16	0.10	0.15	0.16	0.18	0.10	0.00
	mirs9	0.00	0.13	0.19	0.17	0.14	0.18	0.08	-0.02
	mirs9g12	0.15	0.30	0.35	0.44	0.33	0.38	0.20	0.00
	mirs12g20	0.15	0.27	0.28	0.20	0.26	0.33	0.16	0.22
	R15	-0.47	-0.52	-0.96	-1.77	-0.91	-0.55	-0.34	-1.78
	R30	0.03	0.07	0.08	0.11	0.10	0.04	0.07	0.00
	R30e3	0.08	-0.01	-0.18	-0.66	-0.23	0.00	0.14	-0.66
	R45e6	0.17	0.08	-0.03	-0.32	-0.06	0.10	0.15	-0.02
	R60e9	0.15	0.13	0.11	-0.03	0.14	0.10	0.12	0.00
G-DNA-related patterns	GG8g4	0.22	0.72	0.24	0.14	0.20	1.03	0.48	0.44
	GGG4g6	-0.08	-0.26	-0.78	-0.65	-0.67	-0.18	-0.26	-0.22
	GGGG4g6	0.00	0.10	-0.06	-0.12	0.01	0.11	0.02	0.11
Z-DNA-related patterns	GC6	-1.37	-1.50	-1.35	-0.95	-1.41	-1.38	-1.71	-1.42
	GC8	-0.75	-1.10	-0.84	-0.18	-0.47	-1.42	-1.13	-0.37
	RY12	-1.79	-1.58	-0.88	-0.16	-0.78	-1.73	-2.03	-0.99
	RY12e1	-1.81	-1.40	-0.71	-0.31	-0.85	-1.43	-1.93	-0.44
	RY18e2	-1.66	-1.45	-0.72	-0.04	-0.58	-1.60	-2.07	-0.37
	RY24e3	-0.60	-0.78	-0.25	0.05	-0.30	-0.79	-1.12	-0.44
DNA Bending	bend45w60	0.60	0.93	1.51	1.11	1.57	0.59	0.85	1.70
	bend60w100	0.32	0.79	1.40	1.17	1.42	0.50	0.74	1.48
	bend90w120	0.28	0.58	1.23	0.94	1.11	0.34	0.66	1.44

Numbers in the table refer to the average significance category for all genera within each class of organisms. Numbers below the class description indicate numbers of available genera of each class. Anaerobe includes both obligate anaerobes and anaerobes; Aerobe includes both obligate aerobes and aerobes. See Table 2.S5 for colored version of this table.

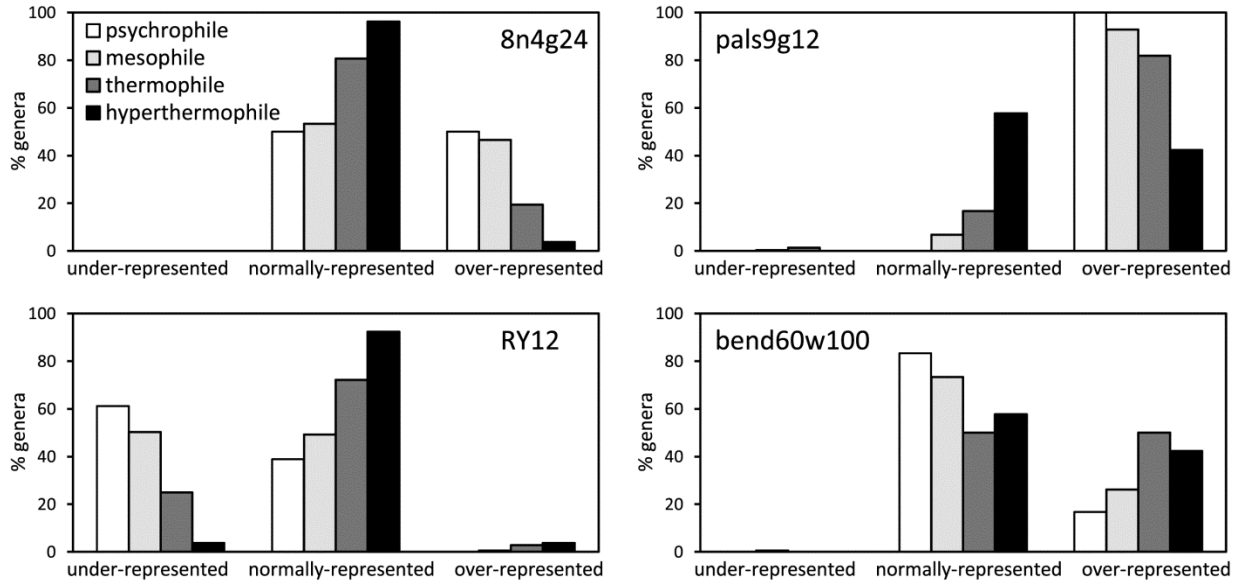

Figure 2.1 Comparison of representations of selected patterns in different OGT classes.

Bars show the percentage of species in each OGT class which have the given pattern under-represented, normally represented, or over-represented. The pattern is considered over-represented if the *p*-value is less than 10^{-4} and observed to expected ratio >1.10 (representation level 2 or higher) for majority of the complete genomes available for that genera, it is deemed under-represented if the *p*-value is less than 10^{-4} and observed to expected ratio <0.91, and normally represented otherwise. See Methods and Table 2.S1. The four patterns for which the data are shown are representative of close repeat structures (8n4g24, top left), palindromes and close inverted repeats (pals9g12, top right), potential Z-DNA-promoting patterns (RY12, bottom left), and DNA bending pattrens (bend60w100, bottom right). See Table 2.1 for description of the pattern codes.

Table 2.S1 Pattern representation classes

General level	Detailed level	p-value	r = observed/expected
Under-represented	-4	$p \leq 10^{-12}$	$r \leq 0.67$
	-3	$p \leq 10^{-7}$	$r \leq 0.80$
	-2	$p \leq 10^{-4}$	$r \leq 0.91$
normally-represented	-1	$p \leq 10^{-2}$	$r \leq 0.95$
	0	$p > 10^{-2}$	$0.95 < r < 1.05$
	1	$p \leq 10^{-2}$	$r > 1.05$
Over-represented	2	$p \leq 10\text{-}^{4}$	$r > 1.10$
	3	$p \leq 10^{-7}$	$r > 1.25$
	4	$p \leq 10^{-12}$	$r > 1.50$

The sequence patterns are classified into three general categories and nine detailed categories based on their representation in the genome using the p-value and the observed/expected ratio. A patern is assigned the most extreme category for which it qualifies by both criteria. See Materials and Methods for details and justification.

Table 2.S2 Representation of sequence patterns in protein-coding regions for different phyla

Pattern name	Pattern code	AlPr	BePr	GaPr	DePr	EpPr	Firm	Acti	Cyan	Bact	Chlb	Chlf	Dein	Fuso	Chla	Spir	Acid	Verr	Defe	Plan	Aqui	Ther	Eury	Cren
		58	39	76	23	11	61	63	12	34	5	8	6	5	6	4	4	4	4	4	8	5	44	16
Simple sequence repeats	1n8	-3.00	-3.00	-4.00	-3.20	-4.00	-4.00	-3.50	-4.00	-4.00	-2.58	-4.00	-4.00	-4.00	-4.00	-3.75	-3.00	-3.00	-4.00	-3.00	-4.00	-4.00	-4.00	-4.00
	2n5	-2.00	-1.00	-2.96	-2.50	-4.00	-3.45	-3.00	-3.00	-4.00	-3.00	-2.50	-2.00	-4.00	-4.00	-3.43	-1.00	-2.50	-4.00	-1.75	-4.00	-4.00	-3.00	-3.00
	3n4	-0.97	-2.00	-2.00	-1.00	-3.00	-2.88	-2.00	-1.09	-2.00	-2.00	-1.44	-1.50	-4.00	-2.00	-2.59	-1.00	-1.50	-3.00	-0.50	-3.00	-3.00	-3.00	-3.00
	4n4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5n4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	6n3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	7n3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	8n3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	9n2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.22	0.00	0.00	0.00	0.00	0.00	-0.50	0.00	0.00
	10n2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	11n2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Close direct repeats	4n6g12	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00
	6n6g24	0.00	0.00	0.00	0.00	0.00	0.00	2.00	1.75	0.00	0.92	0.50	0.00	0.00	0.00	0.47	0.00	0.00	0.00	0.50	0.00	0.00	0.30	0.00
	8n4g24	0.04	0.33	0.39	0.50	0.00	0.00	2.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.47	0.00	0.00	0.00	1.50	0.00	0.00	0.42	0.00
	cd8g6	0.00	0.00	0.00	0.00	0.50	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	-0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13
	cd10g50	-0.33	-0.17	-1.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.81	-0.47	0.00	-0.50	-1.00	0.00	-1.00	-1.29	-0.50	0.00
Palindromes & inverted repeats	cp8g6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.50	0.15	-0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cp10g50	-0.25	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.17	0.00	-0.50	0.00	0.25	0.00	1.00	0.00	0.00
	pals9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	pals9g12	0.00	0.00	0.03	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.75	0.00	0.00	1.50	0.92	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00
	pals12g20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.00	-0.50	0.00	0.50	0.00	0.00	0.00	0.00
H-DNA-related patterns	cm8g6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.25	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
	cm10g50	1.00	2.00	2.00	2.00	1.50	2.00	0.00	1.22	3.00	4.00	2.17	0.00	2.00	2.00	2.15	0.75	2.00	1.00	1.25	2.00	0.14	1.71	0.18
	mirs9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mirs9g12	0.00	0.00	0.33	1.00	0.00	0.39	0.00	0.00	0.00	1.42	0.71	0.00	1.00	0.04	0.42	0.00	1.00	0.00	0.25	0.25	0.00	0.00	0.00
	mirs12g20	1.00	2.00	1.67	2.00	1.00	2.00	0.00	1.16	2.25	3.00	1.75	0.00	2.00	1.84	1.79	0.50	1.50	2.00	0.25	1.25	1.00	1.50	0.00
	R15	0.00	0.00	0.00	0.00	-1.00	-0.60	0.00	-0.75	0.00	0.00	0.00	-1.00	-4.00	0.00	-0.72	-0.50	0.00	-2.50	0.00	-2.25	-3.00	-1.00	-1.00
	R30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
	R30e3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-3.00	0.00	-0.37	0.00	0.50	-1.00	0.00	-0.50	-2.00	0.00	0.00
	R45e6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-2.00	0.00	0.15	0.00	1.00	-0.50	0.00	0.00	-0.50	0.00	0.00
	R60e9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.04	0.00	1.00	0.00	0.00	0.00	-0.50	0.00	0.00
G-DNA-related patterns	GG8g4	1.00	1.33	0.00	0.94	0.00	0.00	1.00	0.75	0.00	0.00	0.50	2.60	0.00	0.00	0.34	0.50	0.00	0.00	2.00	0.00	0.00	0.00	0.00
	GGG4g6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	GGGG4g6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Z-DNA-related patterns	GC6	0.00	0.00	-1.59	-0.57	-0.50	-0.58	-2.00	-2.00	0.00	-3.25	-1.88	-1.20	0.00	0.00	-0.22	-0.75	-1.00	-0.50	-1.50	0.00	0.00	-1.00	-0.50
	GC8	0.00	0.00	-0.54	0.00	0.00	0.00	-3.00	0.00	0.00	-1.42	-0.67	-0.88	0.00	0.00	0.00	-0.50	0.00	0.00	-2.00	0.00	0.00	0.00	0.00
	RY12	-2.00	-1.93	-2.00	-0.50	-0.67	0.00	-1.00	-1.00	0.00	-2.00	-2.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	-2.00	0.00	0.00	-0.66	0.00
	RY12e1	-2.00	-1.33	-2.71	-1.00	-0.42	-1.00	-1.00	-3.00	0.00	-3.00	-2.00	0.00	-1.00	1.34	0.00	-0.75	0.00	-0.50	-3.00	0.50	0.00	-0.45	-1.00
	RY18e2	-2.00	-1.50	-2.00	0.00	0.00	0.00	-1.00	-0.50	0.00	-2.00	-2.59	-0.25	0.00	1.00	0.00	-1.25	0.00	0.00	-2.50	0.00	0.00	0.00	-0.54
	RY24e3	-1.00	-1.00	-1.00	0.00	0.00	0.00	-0.44	0.00	0.00	-1.00	-1.00	-0.25	0.00	0.00	0.00	-0.50	-0.50	0.00	-0.50	0.00	0.00	0.00	0.00
DNA Bending	bend45w60	0.00	0.00	0.00	0.00	2.50	1.00	0.00	1.00	1.00	0.00	0.00	0.00	2.00	0.00	-0.02	-1.00	0.00	2.00	-0.75	2.00	2.00	1.00	0.00
	bend60w100	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	-0.17	0.00	0.00	2.50	-0.50	2.00	2.00	0.65	0.00
	bend90w120	0.00	0.00	0.00	0.00	1.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	-0.79	0.00	0.00	3.00	0.00	0.75	2.00	0.00	0.00

Same as Table 2.2 but showing the data for protein-coding regions only. Protein-coding regions are defined as all segments annotated as protein coding sequences (the “CDS” feature in the GenBank entry), whereas all other segments are considered non-coding. That is, untranslated regions of genes as well as RNA genes are considered “intergenic” for the purpose of this analysis.

Table 2.S3 Representation of sequence patterns in the intergenic regions for different phyla

Pattern name	Pattern code	AlPr	BePr	GaPr	DePr	EpPr	Firm	Acti	Cyan	Bact	Chlb	Chlf	Dein	Fuso	Chla	Spir	Acid	Verr	Defe	Plan	Aqui	Ther	Eury	Cren
		58	39	76	23	11	61	63	12	34	5	8	6	5	6	4	4	4	4	4	8	5	44	16
Simple sequence repeats	1n8	-3.00	-2.00	-2.00	-100	-2.00	-3.00	-3.00	-3.00	-2.00	100	-2.50	-3.00	-2.00	-0.34	-1.79	-150	-2.00	-150	-2.00	-3.00	-4.00	-3.00	-3.75
	2n5	-100	-1.75	-2.00	-125	-3.00	-2.50	-2.00	-2.00	-3.00	-2.00	-1.69	-2.00	-3.00	-2.79	-2.29	-0.75	-2.00	-3.00	-100	-2.00	-3.00	-2.00	-2.00
	3n4	0.00	0.00	0.00	0.00	-100	-0.50	0.00	-0.17	-100	0.00	0.00	0.00	-100	-0.59	-0.72	0.00	0.00	-100	0.00	0.00	-100	0.00	0.00
	4n4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5n4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	6n3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	7n3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.17	0.00	0.00	0.00	0.00	0.00	1.50	2.00	0.25	0.00	0.00	0.00	0.00
	8n3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.34	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00
	9n2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	10n2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	11n2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.41	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.75	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Close direct repeats	4n6g12	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.03	0.00	0.00	2.00	0.00	0.00	0.00	0.31	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	6n6g24	0.00	0.00	0.16	0.47	0.00	0.00	0.00	0.00	0.75	0.00	0.50	0.00	0.00	0.00	0.68	0.50	0.00	2.00	0.00	0.00	0.00	0.00	0.00
	8n4g24	0.00	0.00	0.50	1.00	0.00	0.00	2.00	0.54	1.25	1.42	2.00	0.00	0.00	0.00	0.88	1.00	0.00	2.00	0.50	0.00	0.00	0.00	0.00
	cd8g6	0.47	1.00	0.56	1.00	1.00	1.00	2.00	3.50	1.00	1.00	2.84	1.23	0.00	0.00	0.69	2.00	1.50	2.00	0.25	0.00	1.50	1.00	1.00
	cd10g50	1.27	1.00	1.30	2.00	1.00	1.25	3.73	2.50	3.00	1.00	4.00	1.55	0.00	0.00	1.07	3.50	1.50	0.50	1.50	0.00	2.00	2.00	1.00
Palindromes & inverted repeats	cp8g6	4.00	4.00	4.00	4.00	4.00	4.00	4.00	3.21	4.00	4.00	4.00	4.00	4.00	4.00	3.63	4.00	4.00	4.00	4.00	3.50	4.00	3.75	2.00
	cp10g50	4.00	4.00	4.00	4.00	3.50	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	2.50	4.00	3.00	1.88
	pals9	1.00	1.00	1.54	1.00	0.00	2.00	2.00	0.84	2.00	1.00	1.00	1.00	1.00	1.00	0.62	0.25	1.00	0.50	0.00	0.00	1.50	0.00	0.00
	pals9g12	4.00	3.75	4.00	4.00	2.00	4.00	4.00	3.50	4.00	4.00	4.00	3.90	4.00	4.00	3.42	4.00	3.50	2.50	4.00	1.00	4.00	1.75	0.50
	pals12g20	1.88	1.00	2.00	2.00	0.00	3.00	4.00	1.00	3.00	2.00	2.00	1.50	2.00	1.63	1.07	1.50	1.50	0.00	1.50	0.00	1.50	0.00	0.00
H-DNA-related patterns	cm8g6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cm10g50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mirs9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mirs9g12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mirs12g20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.42	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R30e3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R45e6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R60e9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G-DNA-related patterns	GG8g4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	GGG4g6	0.00	0.00	0.00	0.00	0.00	0.00	-0.52	0.00	0.00	0.00	0.25	-0.10	0.00	0.00	-0.07	0.00	-100	0.00	0.00	0.00	0.00	0.00	0.00
	GGGG4g6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Z-DNA-related patterns	GC6	-0.50	-1.06	-100	-100	0.00	-0.25	-2.00	-0.25	-100	-100	-1.23	0.00	0.00	0.00	-0.50	-125	-150	0.00	-2.50	0.00	0.00	-100	0.00
	GC8	0.00	-0.97	0.00	0.00	0.00	0.00	-1.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.25	0.00	0.00	-100	0.00	0.00	0.00	0.00
	RY12	-100	-100	-100	0.00	0.00	0.00	-100	-0.34	0.00	-0.58	-1.50	0.00	0.00	0.00	0.11	-0.25	0.00	0.00	-100	0.00	0.00	0.00	0.00
	RY12e1	-188	-150	-2.45	-0.27	0.00	0.00	-2.00	-3.00	0.00	-0.58	-2.00	0.00	0.00	0.00	0.52	0.00	-0.50	-0.50	-3.00	0.00	0.00	0.00	0.00
	RY18e2	-0.75	-100	-100	0.00	0.00	0.00	-100	-0.34	0.00	-100	-1.25	0.00	0.00	0.00	0.17	-0.25	0.00	0.00	-1.50	0.00	0.00	0.00	0.00
	RY24e3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DNA Bending	bend45w60	0.00	0.00	0.94	0.00	2.50	1.70	0.00	1.42	1.00	0.25	0.00	0.00	1.00	0.71	1.00	0.00	0.00	1.00	0.75	0.00	1.00	1.00	0.00
	bend60w100	0.00	0.00	0.82	0.00	2.00	1.00	0.00	1.09	1.00	0.00	0.00	0.00	1.00	0.00	0.93	0.25	0.00	1.00	0.50	0.00	0.50	1.00	0.00
	bend90w120	0.00	0.00	0.08	0.00	1.50	1.00	0.00	0.09	1.00	0.00	0.00	0.00	1.00	0.00	0.69	0.00	0.00	1.00	0.00	0.00	0.50	0.00	0.00

Same as Table 2.2 but showing the data for 'intergenic' regions only, that is, all regions not annotated as 'CDS' in the GenBank files.

Table 2.S4 Representation of sequence patterns in different OGT and oxygen requirement classes restricted to protein coding regions

Pattern name	Pattern code	Psychrophile	Mesophile	Thermophile	Hyperthermophile	Anaerobe	Aerobe	Facultative	Microaerophile
		18	382	72	26	159	202	95	9
Simple sequence repeats	1n8	-3.76	-3.39	-3.83	-3.92	-3.73	-3.34	-3.44	-3.78
	2n5	-2.91	-2.47	-3.21	-3.09	-2.96	-2.32	-2.67	-3.00
	3n4	-1.95	-1.82	-2.71	-2.91	-2.43	-1.65	-2.07	-2.30
	4n4	0.00	0.00	-0.01	0.00	-0.02	0.01	0.00	-0.05
	5n4	0.00	0.01	0.00	0.00	0.00	0.02	0.01	0.00
	6n3	0.00	0.11	0.00	0.01	0.01	0.14	0.02	0.11
	7n3	0.00	0.07	0.07	0.00	0.07	0.08	0.04	0.00
	8n3	0.00	0.04	0.00	0.00	0.02	0.04	0.03	0.00
	9n2	-0.20	-0.02	-0.04	0.05	-0.03	0.04	-0.19	-0.11
	10n2	0.00	0.04	0.01	0.05	0.03	0.03	0.05	0.11
	11n2	0.02	0.04	0.01	0.00	0.05	0.03	0.05	0.02
Close direct repeats	4n6g12	0.07	0.29	0.19	0.53	0.22	0.40	0.20	0.07
	6n6g24	0.73	0.72	0.26	0.26	0.46	0.83	0.40	0.70
	8n4g24	1.20	0.92	0.40	0.19	0.62	1.07	0.63	0.46
	cd8g6	-0.32	0.19	0.33	0.70	0.27	0.32	-0.07	0.28
	cd10g50	-0.90	-0.36	-0.70	-0.66	-0.52	-0.20	-0.73	-0.47
Palindromes & inverted repeats	cp8g6	0.31	0.09	0.19	0.45	0.18	0.03	0.14	0.12
	cp10g50	0.28	-0.04	0.24	0.38	0.04	-0.06	0.15	0.48
	pals9	0.33	0.15	0.16	0.25	0.14	0.16	0.17	0.22
	pals9g12	0.67	0.28	0.41	0.41	0.37	0.25	0.35	0.52
	pals12g20	0.33	0.07	0.09	0.31	0.00	0.11	0.17	0.56
H-DNA-related patterns	cm8g6	0.11	0.22	0.18	0.27	0.25	0.24	0.19	0.13
	cm10g50	1.55	1.65	1.32	1.13	1.89	1.29	1.73	1.17
	mirs9	0.00	0.06	0.03	0.09	0.05	0.08	0.05	0.00
	mirs9g12	0.30	0.52	0.44	0.50	0.63	0.43	0.53	0.37
	mirs12g20	1.50	1.49	1.39	1.19	1.78	1.23	1.55	1.13
	R15	-0.40	-0.50	-1.12	-1.81	-1.09	-0.44	-0.25	-1.63
	R30	0.00	0.03	0.07	0.12	0.08	0.02	0.02	0.00
	R30e3	0.00	-0.15	-0.35	-0.82	-0.47	-0.08	0.03	-0.52
	R45e6	0.08	0.00	-0.05	-0.31	-0.14	0.04	0.08	-0.03
	R60e9	0.17	0.06	0.03	-0.07	0.03	0.06	0.08	0.00
G-DNA-related patterns	GG8g4	0.32	0.75	0.38	0.42	0.24	1.07	0.57	0.44
	GGG4g6	-0.06	-0.09	-0.39	-0.27	-0.38	0.02	-0.07	-0.11
	GGGG4g6	0.00	0.04	-0.02	-0.04	0.00	0.06	0.01	0.00
Z-DNA-related patterns	GC6	-1.05	-1.20	-1.09	-0.66	-1.10	-1.10	-1.44	-0.94
	GC8	-0.53	-0.92	-0.61	-0.25	-0.34	-1.17	-1.05	-0.13
	RY12	-1.14	-1.22	-0.65	-0.20	-0.60	-1.27	-1.71	-0.52
	RY12e1	-1.59	-1.27	-0.74	-0.43	-0.87	-1.24	-1.83	-0.05
	RY18e2	-0.93	-1.21	-0.45	-0.20	-0.56	-1.17	-1.75	-0.11
	RY24e3	-0.33	-0.60	-0.22	-0.11	-0.26	-0.55	-0.95	-0.22
DNA bending	bend45w60	0.20	0.36	1.12	1.03	0.91	0.21	0.38	1.02
	bend60w100	0.02	0.26	1.08	0.98	0.82	0.15	0.17	1.12
	bend90w120	-0.25	0.15	0.85	0.71	0.62	0.04	0.14	0.75

Same as Table 2.3 but showing data for protein-coding regions only. Protein-coding regions are defined as segments annotated as protein coding sequences ("CDS" feature in the GenBank entry), whereas all other segments are considered non-coding. That is, untranslated regions of genes as well as RNA genes are considered "intergenic" for the purpose of this analysis.

Table 2.S5 Representation of sequence patterns in different OGT and oxygen requirement classes restricted to the intergenic regions

Pattern name	Pattern code	Psychrophile	Mesophile	Thermophile	Hyperthermophile	Anaerobe	Aerobe	Facultative	Microaerophile
		18	382	72	26	159	202	95	9
Simple sequence repeats	1n8	-1.65	-2.14	-3.10	-3.66	-2.39	-2.48	-2.17	-2.78
	2n5	-2.05	-1.74	-2.37	-2.35	-2.08	-1.70	-1.88	-2.38
	3n4	-0.37	-0.29	-0.60	-0.61	-0.62	-0.17	-0.28	-0.35
	4n4	0.06	0.05	-0.02	0.01	0.04	0.02	0.04	0.00
	5n4	0.06	0.04	0.00	0.00	0.02	0.03	0.02	0.00
	6n3	0.36	0.18	0.07	0.00	0.09	0.18	0.14	0.44
	7n3	1.23	0.47	0.38	0.00	0.36	0.60	0.33	0.78
	8n3	0.56	0.27	0.09	0.01	0.20	0.28	0.17	0.44
	9n2	0.18	0.34	0.35	0.33	0.36	0.33	0.19	0.22
	10n2	0.15	0.29	0.32	0.19	0.31	0.28	0.20	0.00
	11n2	0.19	0.39	0.43	0.34	0.45	0.41	0.25	0.01
Close direct repeats	4n6g12	0.73	0.50	0.48	0.42	0.57	0.53	0.29	0.66
	6n6g24	1.15	0.68	0.43	0.04	0.57	0.74	0.54	0.69
	8n4g24	1.32	0.99	0.62	0.04	0.84	1.05	0.72	0.78
	cd8g6	1.54	1.33	1.26	1.25	1.56	1.31	0.93	1.16
	cd10g50	2.13	1.84	1.64	1.05	1.88	1.88	1.52	1.46
Palindromes & inverted repeats	cp8g6	3.89	3.78	3.49	2.36	3.56	3.68	3.82	3.89
	cp10g50	3.89	3.83	3.37	1.94	3.49	3.76	3.79	3.67
	pals9	2.29	1.40	0.94	0.20	1.18	1.26	1.79	0.88
	pals9g12	3.72	3.49	2.72	1.06	2.96	3.38	3.65	2.79
	pals12g20	2.84	2.04	1.28	0.31	1.62	1.95	2.40	1.02
H-DNA-related patterns	cm8g6	0.10	0.20	0.17	0.26	0.26	0.17	0.12	0.23
	cm10g50	-0.11	0.15	0.14	0.03	0.16	0.14	0.07	0.15
	mirs9	0.00	0.01	0.07	0.00	0.04	0.01	0.01	0.00
	mirs9g12	0.06	0.10	0.09	0.04	0.12	0.09	0.03	0.00
	mirs12g20	0.00	0.01	0.04	0.00	0.04	0.01	0.00	0.00
	R15	-0.15	0.08	0.28	0.07	0.21	-0.05	0.16	-0.18
	R30	0.00	0.03	0.02	0.04	0.05	0.02	0.05	0.00
	R30e3	0.24	0.17	0.17	0.25	0.19	0.12	0.19	0.55
	R45e6	0.21	0.14	0.18	0.09	0.12	0.12	0.19	0.33
	R60e9	0.06	0.09	0.19	0.10	0.10	0.09	0.11	0.33
G-DNA-related patterns	GG8g4	0.08	0.12	0.05	0.12	0.03	0.15	0.09	0.33
	GGG4g6	-0.06	-0.17	-0.54	-0.55	-0.28	-0.30	-0.12	-0.11
	GGGG4g6	0.06	0.07	0.00	-0.04	0.03	0.08	0.05	0.00
Z-DNA-related patterns	GC6	-1.03	-1.17	-0.79	-0.30	-0.71	-1.36	-1.03	-0.73
	GC8	-0.29	-0.48	-0.37	0.01	-0.14	-0.74	-0.36	-0.04
	RY12	-1.09	-0.64	-0.18	-0.19	-0.16	-0.83	-0.76	-0.48
	RY12e1	-1.86	-1.20	-0.22	-0.06	-0.22	-1.48	-1.57	-0.59
	RY18e2	-0.95	-0.57	-0.21	0.06	-0.05	-0.78	-0.81	-0.47
	RY24e3	-0.08	-0.11	0.02	0.18	0.08	-0.16	-0.21	-0.35
DNA bending	bend45w60	1.02	0.92	0.77	0.28	1.03	0.66	0.94	1.72
	bend60w100	0.85	0.74	0.56	0.36	0.86	0.53	0.76	1.29
	bend90w120	0.41	0.52	0.40	0.28	0.68	0.29	0.54	0.89

Same as Table 2.3 but showing the data for 'intergenic' regions only, that is, all regions not annotated as 'CDS' in the GenBank files.

Table 2.S6 Representation of sequence patterns in different oxygen requirement classes restricted to mesophilic organism

Pattern name	Pattern code	All Mesophile	Anaerobe	Aerobe	Facultative	Microaerophile
		382	98	168	81	5
Simple sequence repeats	1n8	-3.54	-3.59	-3.58	-3.60	-3.40
	2n5	-2.89	-3.13	-2.70	-2.98	-3.20
	3n4	-1.92	-2.25	-1.63	-2.10	-2.40
	4n4	0.01	0.00	0.01	0.00	-0.18
	5n4	0.07	0.03	0.11	0.02	0.20
	6n3	0.19	0.08	0.26	0.07	0.37
	7n3	0.48	0.41	0.62	0.30	0.80
	8n3	0.30	0.33	0.30	0.20	0.20
	9n2	0.09	0.07	0.17	-0.12	0.00
	10n2	0.25	0.27	0.25	0.17	0.00
	11n2	0.32	0.40	0.28	0.23	0.05
Close direct repeats	4n6g12	0.58	0.53	0.74	0.30	0.71
	6n6g24	1.16	1.20	1.27	0.82	1.71
	8n4g24	1.61	1.61	1.81	1.21	1.43
	cd8g6	0.73	0.89	0.78	0.41	0.66
	cd10g50	0.41	0.49	0.60	0.00	0.53
Palindromes & inverted repeats	cp8g6	2.78	3.03	2.34	3.34	3.20
	cp10g50	3.26	3.42	3.01	3.65	3.77
	pals9	2.58	2.87	2.29	3.00	2.84
	pals9g12	3.46	3.63	3.23	3.73	3.94
	pals12g20	3.58	3.54	3.54	3.84	3.80
H-DNA-related patterns	cm8g6	0.20	0.24	0.23	0.14	0.07
	cm10g50	0.16	0.19	0.19	0.10	0.00
	mirs9	0.13	0.11	0.20	0.04	-0.03
	mirs9g12	0.30	0.30	0.38	0.16	0.00
	mirs12g20	0.27	0.27	0.35	0.12	0.20
	R15	-0.52	-0.51	-0.54	-0.40	-2.00
	R30	0.07	0.12	0.04	0.03	0.00
	R30e3	-0.01	-0.02	0.02	0.06	-0.99
	R45e6	0.08	0.04	0.11	0.11	0.16
	R60e9	0.13	0.20	0.12	0.08	0.00
G-DNA-related patterns	GG8g4	0.72	0.26	1.11	0.52	0.20
	GGG4g6	-0.26	-0.55	-0.13	-0.27	0.01
	GGGG4g6	0.10	0.07	0.16	0.02	0.20
Z-DNA-related patterns	GC6	-1.50	-1.63	-1.34	-1.73	-2.35
	GC8	-1.10	-0.60	-1.37	-1.25	-0.47
	RY12	-1.58	-0.92	-1.78	-2.21	-1.38
	RY12e1	-1.40	-0.97	-1.47	-2.15	-1.20
	RY18e2	-1.45	-0.71	-1.64	-2.32	-0.67
	RY24e3	-0.78	-0.39	-0.84	-1.31	-0.80
DNA bending	bend45w60	0.93	1.52	0.62	0.81	2.06
	bend60w100	0.79	1.30	0.52	0.70	1.66
	bend90w120	0.58	0.94	0.35	0.63	1.40

Only mesophiles were included in this table in order to assess the independence of trends with respect to OGT and oxygen requirement. See **Error! Reference source not found.**2.3.

Table 2.S7 Representation of sequence patterns in different oxygen requirement classes restricted to thermophilic organism

Pattern name	Pattern code	Thermophile	Anaerobe	Aerobe	Facultative	Microaerophile
		112	61	26	13	4
Simple sequence repeats	1n8	-3.89	-3.94	-3.77	-3.92	-4.00
	2n5	-3.40	-3.59	-3.01	-3.24	-3.00
	3n4	-2.83	-3.15	-2.27	-2.48	-2.25
	4n4	-0.06	-0.10	0.00	0.00	0.00
	5n4	0.00	0.00	0.00	0.00	0.00
	6n3	0.06	0.00	0.08	0.04	0.25
	7n3	0.45	0.34	0.12	0.35	1.00
	8n3	0.12	0.07	0.04	0.12	0.50
	9n2	0.22	0.12	0.38	0.00	-0.25
	10n2	0.25	0.31	0.23	0.12	0.00
	11n2	0.30	0.50	0.12	0.24	0.00
Close direct repeats	4n6g12	0.48	0.57	0.69	0.42	0.25
	6n6g24	0.71	0.63	0.50	0.42	1.00
	8n4g24	0.80	0.66	0.66	0.70	0.75
	cd8g6	0.80	0.84	1.14	0.50	0.75
	cd10g50	0.05	-0.17	0.58	-0.14	-0.50
Palindromes & inverted repeats	cp8g6	2.81	2.79	2.06	2.12	3.25
	cp10g50	2.98	2.67	3.07	2.00	2.50
	pals9	2.13	1.73	1.80	1.88	2.00
	pals9g12	3.05	2.80	2.89	2.08	3.25
	pals12g20	2.97	2.52	2.98	1.92	2.00
H-DNA-related patterns	cm8g6	0.21	0.26	0.27	0.08	0.00
	cm10g50	0.10	0.12	0.13	0.04	0.00
	mirs9	0.19	0.19	0.16	0.23	0.00
	mirs9g12	0.35	0.37	0.38	0.29	0.00
	mirs12g20	0.28	0.24	0.25	0.27	0.25
	R15	-0.96	-1.58	-0.60	0.18	-1.50
	R30	0.08	0.09	0.01	0.32	0.00
	R30e3	-0.18	-0.65	-0.07	1.02	-0.25
	R45e6	-0.03	-0.24	0.06	0.45	-0.25
	R60e9	0.11	0.03	0.06	0.30	0.00
G-DNA-related patterns	GG8g4	0.24	0.07	0.62	0.27	0.75
	GGG4g6	-0.78	-0.90	-0.50	-0.59	-0.50
	GGGG4g6	-0.06	-0.07	-0.12	0.04	0.00
Z-DNA-related patterns	GC6	-1.35	-1.11	-1.66	-1.66	-0.25
	GC8	-0.84	-0.29	-1.94	-0.44	-0.25
	RY12	-0.88	-0.50	-1.54	-0.26	-0.50
	RY12e1	-0.71	-0.60	-1.09	-0.13	0.50
	RY18e2	-0.72	-0.32	-1.45	-0.14	0.00
	RY24e3	-0.25	-0.13	-0.51	0.10	0.00
DNA bending	bend45w60	1.51	1.70	0.49	1.59	1.25
	bend60w100	1.40	1.63	0.44	1.56	1.25
	bend90w120	1.23	1.42	0.34	1.31	1.50

Same as Table 2.S6 but including only thermophiles.

Table 2.S8 Representation of sequence patterns in different oxygen requirement classes restricted to bacteria

Pattern name	Pattern code	Anaerobe	Aerobe	Facultative	Microaerophile
		122	189	90	9
Simple sequence repeats	1n8	-3.73	-3.60	-3.64	-3.67
	2n5	-3.27	-2.72	-3.01	-3.11
	3n4	-2.48	-1.70	-2.10	-2.33
	4n4	-0.03	0.01	0.00	-0.10
	5n4	0.02	0.10	0.01	0.11
	6n3	0.05	0.24	0.08	0.31
	7n3	0.52	0.61	0.37	0.89
	8n3	0.26	0.27	0.22	0.33
	9n2	0.03	0.18	-0.11	-0.11
	10n2	0.19	0.23	0.16	0.00
	11n2	0.36	0.25	0.25	0.03
Close direct repeats	4n6g12	0.45	0.74	0.33	0.50
	6n6g24	1.07	1.19	0.84	1.39
	8n4g24	1.38	1.70	1.22	1.13
	cd8g6	0.74	0.80	0.42	0.70
	cd10g50	0.31	0.59	0.05	0.07
Palindromes & inverted repeats	cp8g6	3.24	2.41	3.31	3.22
	cp10g50	3.65	3.16	3.58	3.20
	pals9	3.00	2.41	3.03	2.47
	pals9g12	3.76	3.32	3.65	3.64
	pals12g20	3.79	3.65	3.77	3.00
H-DNA-related patterns	cm8g6	0.15	0.23	0.16	0.04
	cm10g50	0.12	0.18	0.11	0.00
	mirs9	0.06	0.19	0.08	-0.02
	mirs9g12	0.24	0.38	0.21	0.00
	mirs12g20	0.21	0.33	0.17	0.22
	R15	-0.78	-0.49	-0.36	-1.78
	R30	0.10	0.03	0.04	0.00
	R30e3	-0.13	0.02	0.12	-0.66
	R45e6	-0.03	0.10	0.14	-0.02
	R60e9	0.15	0.10	0.10	0.00
G-DNA-related patterns	GG8g4	0.23	1.10	0.52	0.44
	GGG4g6	-0.73	-0.13	-0.23	-0.22
	GGGG4g6	-0.01	0.13	0.02	0.11
Z-DNA-related patterns	GC6	-1.45	-1.33	-1.70	-1.42
	GC8	-0.52	-1.40	-1.14	-0.37
	RY12	-0.93	-1.73	-2.11	-0.99
	RY12e1	-1.03	-1.42	-2.01	-0.44
	RY18e2	-0.73	-1.63	-2.14	-0.37
	RY24e3	-0.39	-0.83	-1.18	-0.44
DNA bending	bend45w60	1.51	0.59	0.79	1.70
	bend60w100	1.33	0.49	0.69	1.48
	bend90w120	1.01	0.32	0.63	1.44

Archaeal genomes were excluded from the data used to generate this table. See Table 2.3 for legend.

Table 2.S9 Number of ATG and GTG start codons embedded in RY-patterns in selected genomes

Species	N_{ATG_RY}	N_{ATG}	Proportion	$N_{\overline{ATG\_RY}}$	$N_{\overline{ATG}}$	Proportion	N_{GTG_RY}	N_{GTG}	Proportion	$N_{\overline{GTG\_RY}}$	$N_{\overline{GTG}}$	Proportion
Chlamydophila pecorum	57	813	7.01%	2,301	35,043	6.57%	0	83	0.00%	1,135	18,798	6.04%
Escherichia coli	154	3,702	4.16%	7,189	149,521	4.81%	11	307	3.58%	5,836	127,565	4.57%
Helicobacter pylori	32	1,282	2.50%	2,955	55,264	5.35%	0	105	0.00%	1,592	34,012	4.68%
Helicobacter felis	96	1,361	7.05%	4,574	55,901	8.18%	11	149	7.38%	4,110	46,662	8.81%
Helicobacter acinonychis	39	1,216	3.21%	3,144	52,483	5.99%	6	143	4.20%	1,811	34,295	5.28%
Helicobacter bizzozeronii	76	1,365	5.57%	4,336	61,445	7.06%	8	227	3.52%	3,581	52,866	6.77%
Thermofilum pendens	89	1,073	8.29%	2,344	28,562	8.21%	37	562	6.58%	3,208	44,316	7.24%
Treponema brennaborense	194	2,285	8.49%	5,337	69,231	7.71%	8	164	4.88%	6,169	56,986	10.83%
Treponema pallidum	78	604	12.91%	4,740	31,469	15.06%	40	322	12.42%	7,590	42,860	17.71%

N_{ATG_RY} is the number of start codons ATG embedded in RY patterns (RY12e1, RY18e2 or RY24e3); $N_{\overline{ATG\_RY}}$ is the number of ATG triplets that are not start codons non-start codon ATG embedded in RY patterns; N_{ATG} is the total number of start codon ATG in the genome; $N_{\overline{ATG}}$ is the number of non-start codon ATG triplets presented in the RY patterns. Same as N_{GTG_RY}, $N_{\overline{GTG\_RY}}$, N_{GTG} and $N_{\overline{GTG}}$. 'Proportion' is the percentage of all ATG/GTG in a given category that are embedded in RY patterns.

Table 2.S10 Mann–Whitney U-test for ratio of intrinsic DNA bends in protein-coding and non-coding regions among different OGT groups

bend60w100 \ bend45w60	**Hyperthermophile**	**Thermophile**	**Mesophile**	**Psychrophile**
Hyperthermophile		0.8943	0.0001	0.0203
Thermophile	0.8857		$< 10^{-9}$	0.0008
Mesophile	0.0001	$< 10^{-9}$		0.6303
Psychrophile	0.0036	0.0001	0.3907	

The ratio of the protein coding and non-coding bend was assessed for each genome and subsequently compared among the genomes belonging to the four OGT classes. The p-values assesed by the Mann-Whitney U test are shown in the table. Low p-values indicate significant differences in protein-coding to noncoding bend ratios between the two classes sompared. Results for the bend45w60 pattern are shown in the upper right triangle whereas those for the bend60w100 patterns are shown in the lower left side triangle.

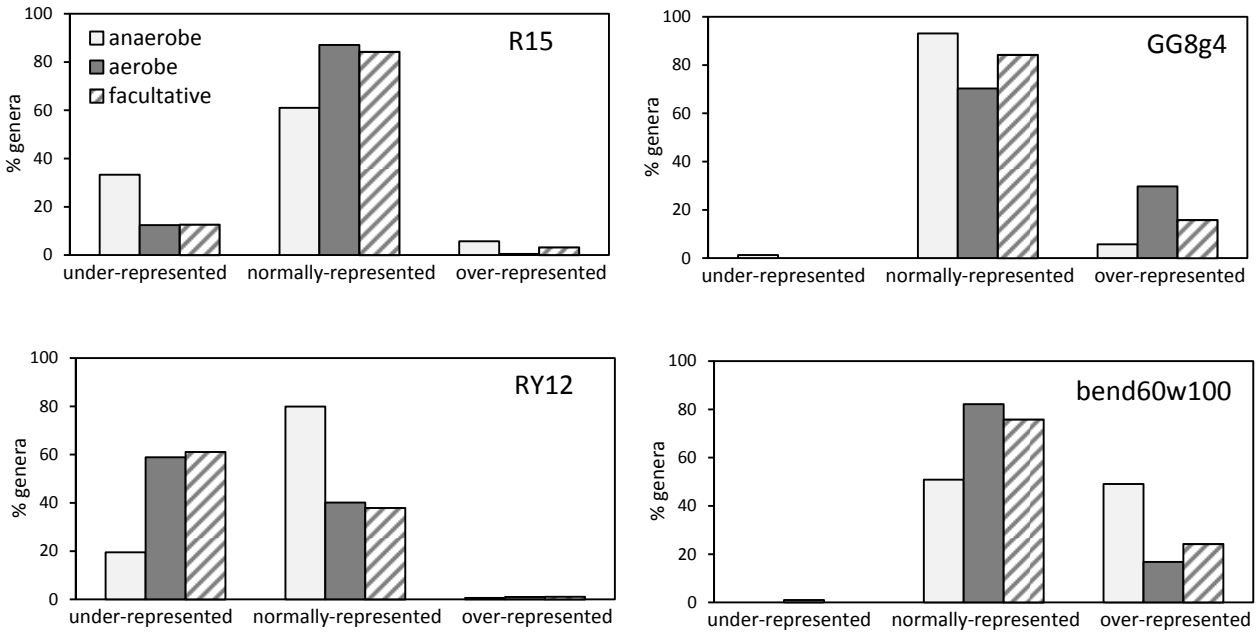

Figure 2.S1 Comparison of pattern representations in different oxygen requirement classes for selected patterns. Bars show the percentage of species in each class which have the given pattern under-represented, normally-represented or over-represented. See legend to Figure 2.1.

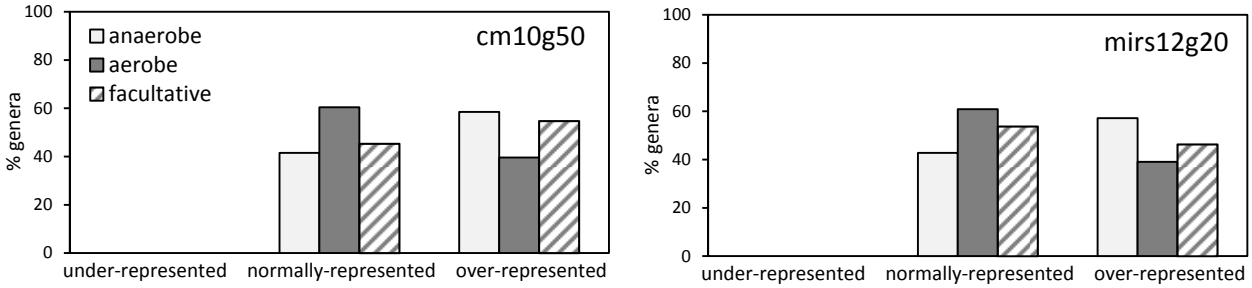

Figure 2.S2 Comparison of representations of two forms of mirror repeats in the protein-coding regions for different oxygen requirement classes. Bars show the percentage of species in each class which have the given pattern under-represented, normally-represented or over-represented. See legend to Figure 2.1.

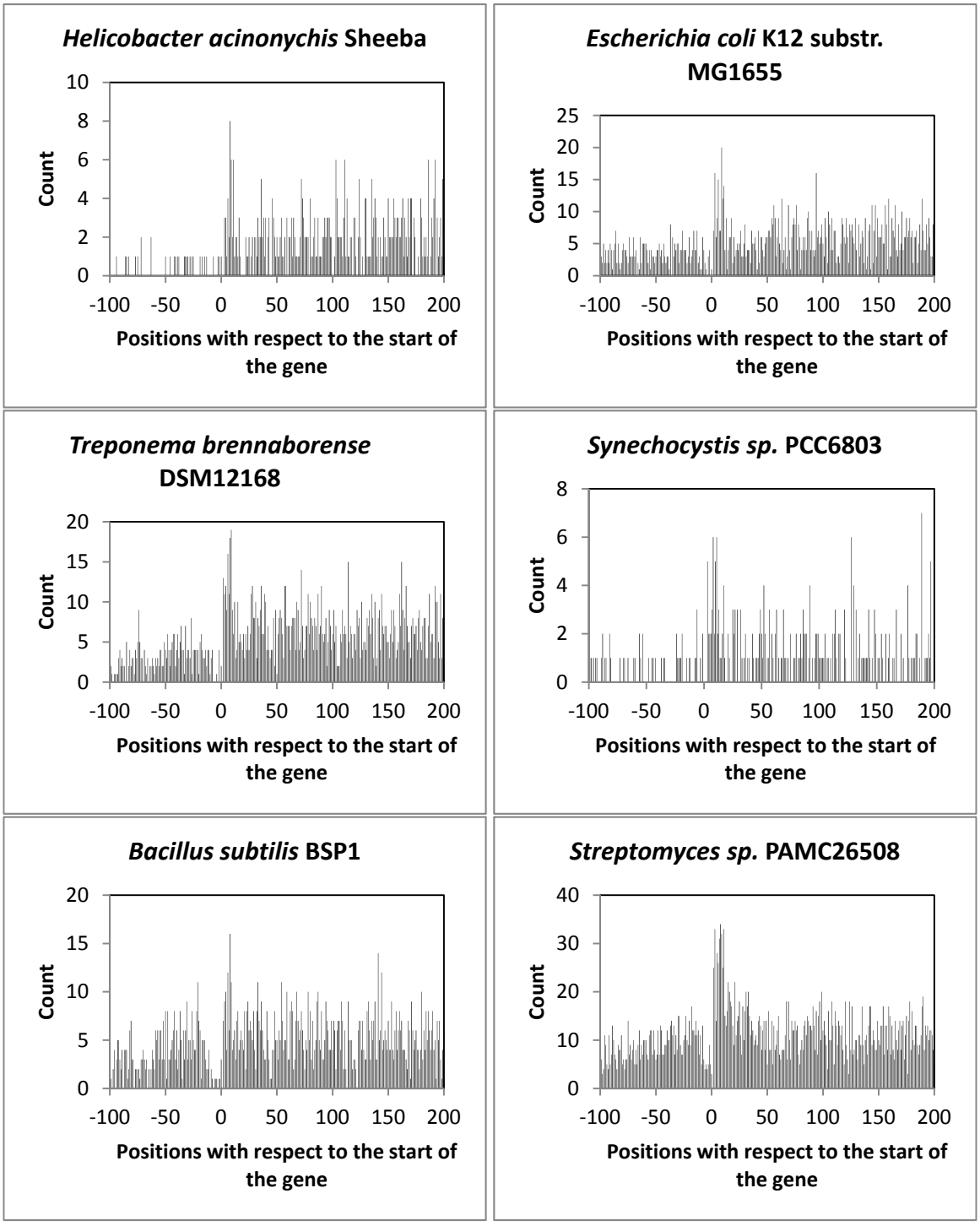

Figure 2.S3 (a) Distribution of RY-patterns with respect to the start of the gene in selected genomes. Position zero refers to the first base of the start codon. The ordinate shows the counts of RY-patterns with the right end located at the position indicated by the abscissa.

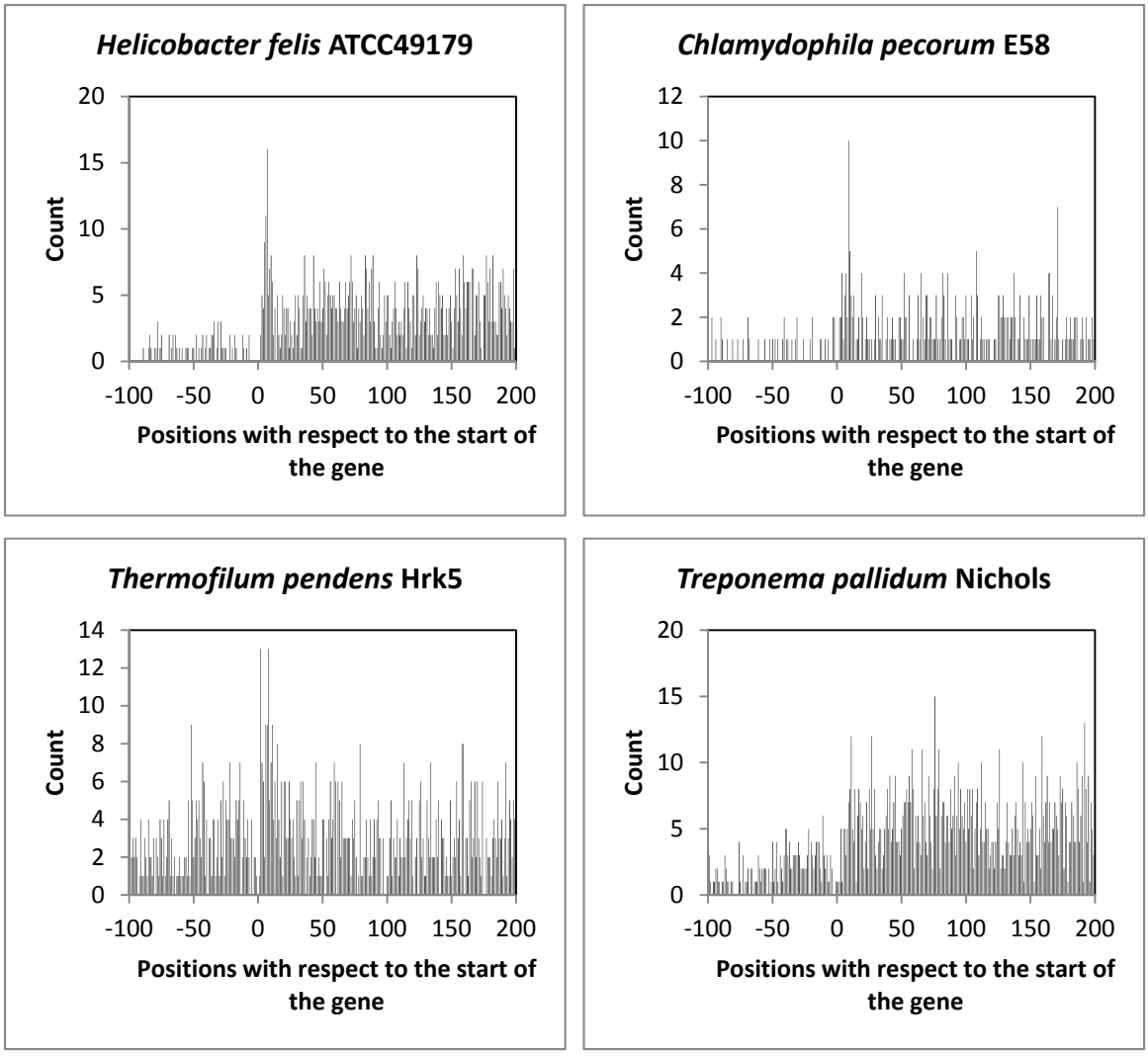

Figure 2.S3 (b) Distribution of RY patterns with respect to the start of the gene in the genomes with the most overrepresented RY patterns.

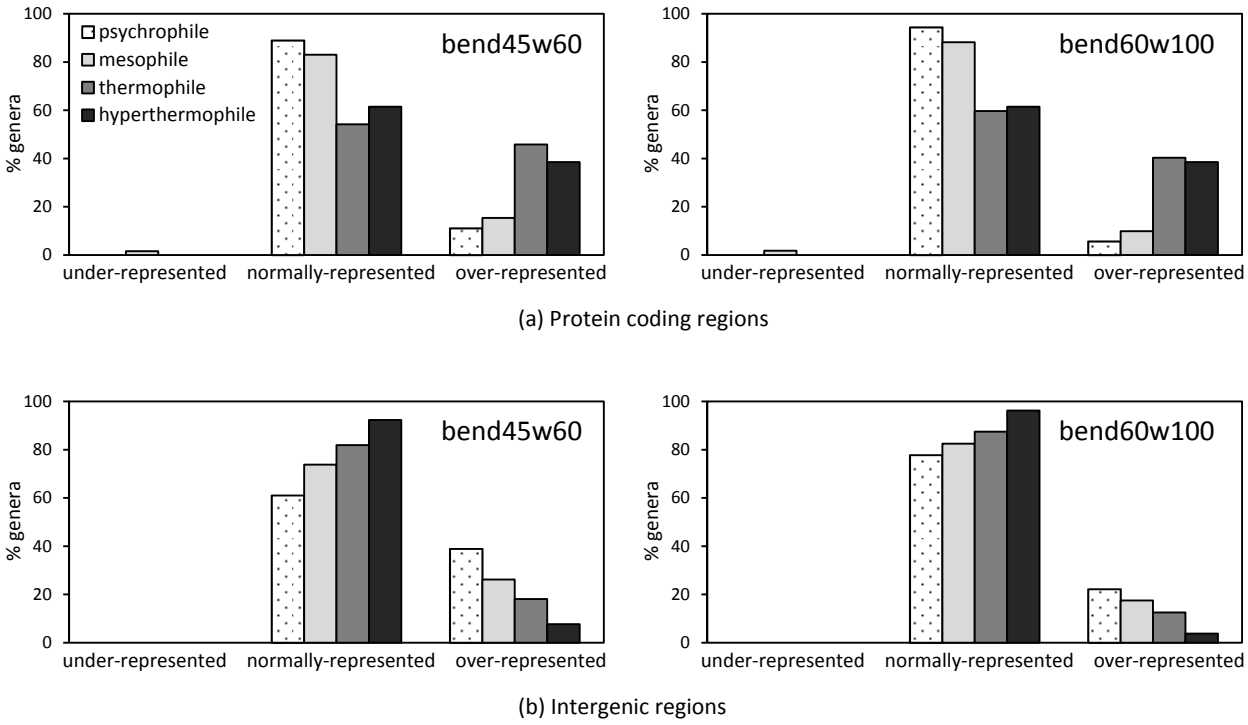

Figure 2.S4 Representations of intrinsic bends in different temperature classes (a) in protein coding regions and (b) in non-coding regions.

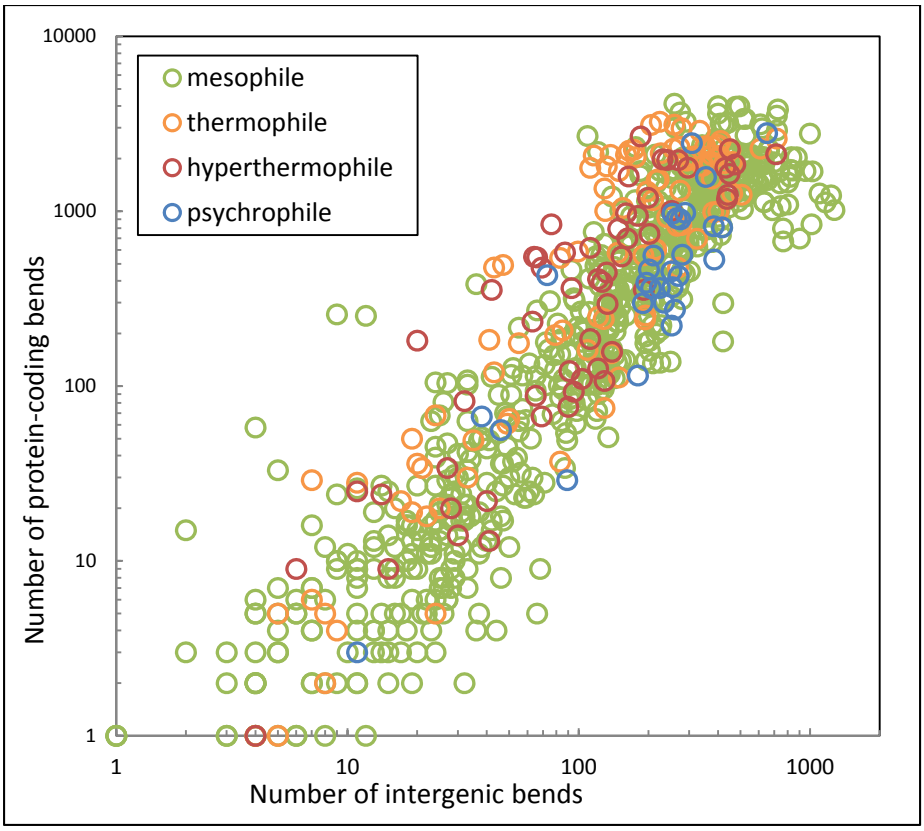

Figure 2.S5 Comparison of number of the bend45w60 patterns in protein-coding regions and non-coding regions. Each circle represents one species and different colors refer to different OGT classes.

CHAPTER 3

DESIGN OF SOFTWARE TO INVESTIGATE EVOLUTION OF REGULATORY MOTIFS IN PROKARYOTIC GENOMES[1]

[1] Huang, Y. and J. Mrazek. To be submitted.

Abstract

We developed a new software work bench for the investigation of evolution of regulatory motifs in prokaryotic genomes. This new tool is aimed to assess the evolutionary conservation of regulatory motifs and identify possible selective constraints that influence their distribution. The evaluation of conservation is based on the multiple sequence alignment of orthologous motif sites in a collection of closely related genomes and comparison of conservation of the motif sequence with its flanking sequences. As a pilot study, the software is used to investigate the evolution of RpoN (σ^{54}) regulons in 107 *Salmonella* and *E. coli* genomes.

Introduction

The development of computational and statistical methods to explore DNA regulatory motifs is one of the central problems in computational biology. Tools for finding significant sequence motifs in both nucleic acid and protein sequences have been extensively established for various tasks (Thompson, Rouchka et al. 2003; Mrazek, Xie et al. 2008; Bailey, Boden et al. 2009; Tran and Huang 2014). There are also many motif profiling and motif comparison tools. For example, GOMO performs Genome Ontology (GO) term association with DNA motifs (Buske, Boden et al. 2010); TomTom is used to search DNA motifs against a DNA motif database and produce an alignment for each significant match (Gupta, Stamatoyannopoulos et al. 2007). However, existing software for analysis of evolutionary roles of regulatory motifs are very limited.

In this work, we aim to develop a new method for analysis of evolution of motifs in prokaryotic genomes, with a particular focus on transcription factor binding sites. First, regulatory motif sites are detected in all analyzed genomes using a standard PSSM (Position-

Specific Score Matrix) technique for supervised motif finding (Mrazek 2009). Orthologous regulatory sites are subsequently identified by BLAST search (Altschul, Gish et al. 1990) and a clustering algorithm that groups together orthologous sites in multiple genomes. To assess potential selective constraints affecting the regulatory motif sites, the level of motif conservation is then evaluated by comparison of information entropy within each motif occurrence and its immediate flanking sequences in the multiple sequence alignment of the clustered orthologs and from the variance of the motif PSSM scores. As a pilot study, the new tool is used to investigate σ^{54} binding site motifs in *Salmonella* and *E. coli* genomes. Results are compared to previous studies and new findings are discussed. Our methodology for investigation of evolution of regulatory motifs is an important step towards understanding the evolution of regulatory networks and how organisms adapt to changing conditions or environments.

Methods

Step 1: Detection of regulatory motifs in individual genomes

Given a set of aligned DNA motif sequences, for example a set of binding sites of a particular transcription factor, other occurrences of the motif can be identified by the Motif Locator (Mrazek, Xie et al. 2008). Motif Locator employs the PSSM method and a cutoff score to find qualified motifs (Mrazek 2009). The PSSM, which is converted from the aligned motif sequences, is a $n \times 4$ matrix consisting of log-odds scores assigned to each nucleotide at each positon of the alignment, where n is the length of the motif. The log-odds score is defined as $s_{i,j} = log(p_{i,j}/q_i)$, where $p_{i,j}$ is the probability of finding nucleotide i at position j and q_i is the probability of finding nucleotide i anywhere in the related genome. Then the PSSM score for a given motif sequence with length n is defined as $S = \sum_{j=1..n} S_{i_j,j}$, where i_j is the nucleotide at

position *j* in the sequence at hand. The PSSM score is a measure of how a sequence is similar to the seed set of known motif sequences, such as verified or high-confidence binding sites for a particular transcription factor. Throughout this chapter, we refer to motif sequences that are found by the Motif Locator as *query motifs*, because they are used as query sequences for Blast search in the following step.

Step 2: Identification and clustering of orthologous regulatory sites in all compared genomes

The next step is to use BLAST to find homologs of each query motif in all investigated genomes. Since the sequence motifs are generally short (most often between 10 and 30 bp), the query motif together with its flanking regions are used as the query sequence for BLAST search (Figure 3.1). By default, 300 bp of the flanking regions on each side is used (see Implementation and Availability).

A query sequence, which consists of the motif plus flanking sequences, is "BLASTed" against the database that includes all investigated genomes except the query sequence genome. Hits are recorded if they satisfy criteria of e-value and bit score cutoff. Only the best hit and one or two other hits whose scores are very close are recorded by the program. Furthermore, hits are filtered based on their BLAST output alignments (between the query sequence and the hit sequence). If the query motif is not included in the BLAST pairwise alignment, i.e., the homology between the query and hit sequence is limited to the motif flanking regions but does not cover the motif itself, the corresponding hit is removed.

After the homologous sites are obtained for each individual query motif (referred to as a *group*), the subsequent clustering step is designed to combine any groups that share one or more motif sites. If any two groups share at least one homolog, where the shared homolog has to be related to a query motif, the program will combine these two groups together, resulting in a

larger new group, called a *cluster*. Note that the occurrence of shared homologs among different groups arises from applying BLAST searches for each query motif; a motif locus in genome A may have homologs in genomes B and C and, consequently, will be included in a group identified by a query motif from genome B as well as the group identified by the query motif from genome C. The process of clustering homologous groups not only reduces the output redundancy but also combines more distantly related homologs that are present in different genomes.

When BLAST is used to find orthologs of the query motifs, paralogs could also be included in the same cluster. Paralogs are excluded in the following way: for each genome in the cluster, if query motif and/or best hit from the BLAST output is present, all other homologs from that genome are discarded; otherwise, the homolog with highest blast bit score among all homologs from that genome is retained whereas other remaining homologs are discarded (Figure 3.1).

Step 3: Multiple sequence alignment

For each cluster, our program will align the homologous motifs plus part of the flanking sequences using ClustalW2 (Larkin, Blackshields et al. 2007). The length of the flanking sequences should be long enough to facilitate a reliable alignment; we use the length of 200 bp by default (see Implementation and Availability). Then, positions of homologous sites of query motifs are adjusted based on the multiple sequence alignment, since those positions are previously predicted only according to the pairwise alignment from the blast output (Figure 3.1).

For the purpose of using surrounding regions as reference to evaluate the level of conservation of the motif, we are only interested in short flanking regions adjacent to the motif itself. Note that there are three different lengths of motif flanking regions used at different stages

of the data processing: largest flanks (e.g., 300 bp on each side) are used as query in the blast search, shorter flanks (typically 200 bp) are used in the multiple sequence alignment, and only part of the multiple alignment centered around the motif and short flanks (typically 30 bp) is used to evaluate the conservation of the motif and its flanking sequences. If the final (shortened) alignment includes identical sequences from the same genome, the involved homologs are considered as paralogs and the one with lower average BLAST score to all other sequences in the cluster (from the original blast results involving long flanks) is removed. The step of multiple sequence alignment is then repeated to reach an alignment output that is not biased by inclusion of identical sequences from the same genome.

Potential problems arise if the homologous sequences in a cluster are too dissimilar because the multiple sequence alignment is likely to be inaccurate, resulting in errors in the subsequent data processing. To counter this issue, our program divides clusters including sequences of low similarity into two or more sub-clusters: First, for each cluster, the program calculates the identity scores I of sequence pairs in the multiple sequence alignment and the minimum identity score I_{min} among the pairs in the cluster. The identity score I between homologs i and j is defined as

$$I_{i,j} = \frac{N_{match}}{N_{match} + N_{mismatch}} \times 100\%$$

where N_{match} is the number of matching nucleotides between sequences i and j in the multiple sequence alignment, and $N_{mismatch}$ is the number of mismatched nucleotides including gaps. Next, if I_{min} of a cluster is lower than a given threshold I_0 (typically 50%), the cluster is split into two. Homologs m and n, whose identity score is equal to I_{min}, are used as seeds for the resulting sub-clusters, and each sequence in the cluster is assigned to the sub-cluster with the more similar seed. The sub-cluster are split recursively until the final sub-clusters have $I_{min} \geq I_0$

or until they contain less than two sequences or no query motif. At the end, multiple sequence alignments are recreated for each new cluster. Note that majority of clusters of homologous motif sites generated in step 2 do not include dissimilar sequences and the sub-clustering procedure described above applies to only a small fraction of the clusters. In the pilot study described below, only 10 out of 2342 clusters identified in step 2 were subsequently divided into sub-clusters.

Step 4: Post processing: calculation of motif scores and sequence entropy

Each motif ortholog is assigned a PSSM score. For orthologous motif sites identified by BLAST search that do not contain a query motif (i.e., a motif with a PSSM score above cutoff value), the PSSM score is calculated based on the sequence that is aligned to the query motif in the multiple sequence alignment. If the length of the aligned sequence is different from that of the query motif (due to an insertion or deletion), the highest PSSM score for any sequence that is close to the aligned sequence is assigned to the ortholog.

The evolutionary conservation of the motif is evaluated using Shannon's information entropy method. First, Shannon entropy is employed as the index of variation for each individual site in the multiple sequence alignment. In a cluster of orthologs that consists of n aligned orthologous sequences, the Shannon entropy score at nucleotide position j is defined as follows.

$$h_j = -\sum_{i=a,t,c,g} p_{i,j} \log_2 p_{i,j} = -\sum_{i=a,t,c,g} \frac{n_{i,j}}{n} \log_2 \frac{n_{i,j}}{n},$$

where $p_{i,j}$ is the probability of finding nucleotide i at position j and $n_{i,j}$ is the number of sequences that have nucleotide i at position j. Position j is excluded from the evaluation if gaps are found at position j in more than $n/3$ sequences. The entropy ranges from 0 to 2 with 0 corresponding to completely conserved sites (all sequences having the same nucleotide at site j)

and 2 corresponds to all four nucleotides being equally frequent at the site. Consequently, conserved nucleotide positions have low entropy scores, whereas variable nucleotide positions have high entropy scores.

For each cluster, the entropy scores are divided into two groups: one for the alignment positions that overlap with query motif sequences, or $H_1 = h_m, h_{m+1}, \dots, h_n$ and the other for the flanking regions of those sites in the multiple sequence alignment,
or $H_2 = h_1, h_2, \dots, h_{m-1}, h_{n+1}, h_{n+2}, .. h_l$, where h_i is the Shannon entropy score for position i in the multiple sequence alignment and the query motif spans from site m to site n. Then the Mann-Whitney U test is conducted on H_1 and H_2, giving out a p-value for each cluster. In addition, the mean value $\bar{h}$ of H_1 and mean value $\bar{h}'$ of H_2 are used to divide the clusters into three groups: a) $\bar{h} < \bar{h}'$; b) $\bar{h} = \bar{h}'$; and c) $\bar{h} > \bar{h}'$. In the primary output, clusters are sorted in the order of p-value, in an increasing order for group a) followed by group b) and group c) sorted by decreasing p-values. Thus, clusters with conserved motifs (relative to their flanks; possibly indicating negative selection) are near the top of the list, clusters with variable motifs (possibly indicative of positive selection) are near the end of the list, and clusters where the motif is neither significantly conserved nor significantly variable comprise the central part of the output. In other word, the ranking of the list of clusters reflects the level of motif conservation from the most conserved to the most variable.

At the end, the program flags questionable clusters where non-orthologous motifs are possibly included, based on distances of the motif from adjacent genes. For a cluster that contains intergenic orthologous sites, the program calculates the standard deviation (*std*) of distances from the center of the motif to its closest genes on both sides. If the *std* is more than a cutoff value (e.g. >100 bp), the corresponding cluster is flagged as questionable. The number of

such questionable clusters can be used as a quality measure for the output, yet the percentage of such clusters is generally low (e.g. ≤5%; Table 4.2).

Implementation and availability

Software

The program was written in C and python and executed in Linux RedHat environment. It includes 8075 lines of code. The requisite software includes BLAST 2.2.29+, CLUSTAL 2.1 (CLUSTAL W2), python 2.6 and R 3.0.3.

Input and options

Given a list of locations of Genbank files for genomes of interest, our program reads the genome information from the corresponding files. High-confidence DNA sequence motifs to be used in construction of the PSSM should be stored in another file, also one motif each line, with no spaces or gaps and all of the same length.

We provide a set of options for the user to customize the program for their needs (Table 3.1). The default values of parameters were selected based on tests with σ^{54} binding sites in collections of *Salmonella* and *E coli* genomes. It turns out that altering values of parameters only changes the output very slightly, implying the good robustness of the program (Table 3.2).

Output

The program generates a large spreadsheet that contains all final clusters. It is organized in such a way that clusters containing the most conserved motif sites are listed near the top of the file, while least conserved motif sites are near the bottom. Clusters are also numbered by their rankings. Within each cluster, orthologs are sorted in the same order as the multiple sequence

alignment guide tree (Table 3.3). Detailed information for each ortholog is listed, including motif score, multiple sequence alignment, the genome, information about adjacent genes, DNA strand and the genome coordinates. There is a summary line printed at the end of each cluster about the statistics of sequence similarity, motif scores, entropies inside and outside the motif, and Mann-Whitney U-test comparing the entropy within the motif with that of the flanking sequences.

Limitations

Our program employs the PSSM method to identify initial query motifs. Therefore, the aligned sequences in the training set must have fixed length and no gap is allowed. Moreover, the analyzed genomes have to be sufficiently closely related in order to identify orthologous motif sites with high accuracy. Because regulatory motifs often occur in intergenic regions, all alignments are done with nucleotide sequence, which diverge faster than amino acid sequences.

In another aspect, our program is limited by the available memory space. In order to reach a faster speed, the program stores sequences of all genomes in the cache. For a machine with 4GB spare memory, the total number of genomes to be analyzed should be less than 400. The number of orthologs is less constrained by the memory space because they are segmentally processed if needed.

Availability

At this time, the program is available from the author upon request. The source codes will be available for download from a web server or public depository upon publication. The source code is distributed under the conditions of the GNU General Public License.

Pilot study: σ^{54} binding site motif in *Salmonella* and *E. coli*

We used the methodology and software described above to investigate the evolution of RpoN (σ^{54}) regulon in *Salmonella* and *E. coli*. RpoN is an alternative σ factor found widely in enterobacteria. It regulates genes that are involved in various cellular processes, including responses to nitrogen starvation, transport and metabolism of carbon substrates and responses to phage shock or other stresses that compromise the cell membrane (Ames and Nikaido 1985; Weiner, Brissette et al. 1991; Gruber and Gross 2003; Niehus, Gressmann et al. 2004). The RpoN regulon in *Salmonella* Typhimurium LT2 has been characterized (Samuels, Frye et al. 2013), but there are still remaining questions about how the RpoN regulon evolves among related genomes and how changes in RpoN binding sites relate to differences in the organisms' physiology, nutrition and environment.

The training set for construction of the PSSM includes 75 high-confidence σ^{54} binding sites in *Salmonella*, which was originally generated from a ChIP-chip experiment for *Salmonella Typhimurium* LT2 (Samuels, Frye et al. 2013). The motif contains two conserved sections, TGGCA followed by TGC, separated by 7 nucleotides that show a lower level of conservation (Figure 3.2). Comparison of motif PSSM scores with the intensity of signal in the ChIP-chip data shows that a motif score ≥ 16 has strong indication of a σ^{54} binding site, and motif score from 14 to 16 has a slightly weaker indication, while sequence with motif score from 12 to 14 is still possibly a σ^{54} binding site (Samuels, Frye et al. 2013). Thus, the motif score cutoff for Motif Locator is set to at 12, while the post analysis is mainly focused on clusters that have maximum motif scores ≥ 14.

As for the input genomes, we have downloaded all available *Salmonella* and *E. coli* genomes from NCBI database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/ in March 2015). This

dataset consists of 107 complete genomes: including 45 complete *Salmonella* genomes and 62 complete *E. coli* genomes. The comparison was performed for the whole dataset of 107 genomes together as well as the *Salmonella* and *E. coli* genomes separately. The reason *E. coli* was included is that it is closely related to *Salmonella* and expected to share many aspects of their regulatory networks. Moreover, the σ^{54} binding sites are similar in the two genomes, justifying the use of the same training set for all genomes under investigation.

Results and Discussion

The analysis of σ^{54} binding sites in 107 *Salmonella* and *E. coli* genomes yielded 2332 clusters of orthologs. The default values of parameters were used (Table 3.4). The number of clusters is reduced to 732 when the maximum motif score cutoff is raised to 14; that is, all clusters that do not include at least one motif site with PSSM score > 14 are excluded. When the maximum motif score cutoff is set to 16 there are only 204 clusters of orthologs; as stated above, scores >16 give high confidence of a true σ^{54} binding sites motif (Table 3.4). The total number of orthologs that are found in the set of 107 genomes is more than the sum of those found in the two individual datasets of *Salmonella* and *E. coli* genomes separately (Table 3.4). It demonstrates that increasing the number of genomes may contribute to finding additional orthologs of the query motif. The distribution of number of orthologs in each cluster is shown in Figure 3.3. Most clusters include about 100 orthologs (e.g. Clusters 1656 and 2315 shown in Tables 3.13 and 3.14), that is, almost all genomes contain the orthologous motif site.

Relationship between motif score and level of conservation

The maximum motif score of a cluster is decreasing with the ranking in all three sets of genomes (Figure 3.4 and Table 3.5). As noted above, the ranking reflects the level of

conservation of the motif sites, with most conserved binding sites at the top and the most variable at the bottom. The higher cutoff of the maximum motif score yields stronger correlation between maximum motif score and ranking. The relationship between average motif score (as opposed to the maximum motif score) and ranking is similar but stronger (Figure 3.5 and Table 3.5). The association between the motif score and ranking suggests that binding sites with higher scores (that is, more likely to be true binding sites) tend to be more conserved than those with low scores. This is expected because stronger motifs are more likely under negative selection and therefore more conserved.

Clusters belonging to the intergenic regions generally rank higher (average ranking is 133 out of 356) than those in the intragenic regions (average = 195), with $p < 10^{-7}$ (from the Mann-Whitney U test on rankings of intergenic and intragenic motif sites) (Figure 3.6). Also, intergenic sites have stronger association with the ranking than those in genes (Figure 3.6 and Table 3.6). This is not surprising because motifs located in the intergenic regions are more likely to have regulatory functions and therefore under negative selection and more conserved.

Orthologs of known σ^{54} binding sites in Salmonella

Table 3.7 summarizes the results of orthologous motifs that include five known σ^{54} binding sites (See complete data in Additional File 1). The listed operons are known to be regulated by σ^{54} in *Salmonella Typhimurium* (Ames and Nikaido 1985; Hirschman, Wong et al. 1985; Weiner, Brissette et al. 1991; Klose and Mekalanos 1997; Palacios and Escalante-Semerena 2000). As expected, the orthologous motifs are all found within 100 bp upstream of genes in almost all investigated genomes (*E. coli* K12 DH10B and *E. coli* BW2952 do not have *prpBCDE* operon, which was verified by protein BLAST search). It retroactively confirms that the σ^{54} proteins in *Salmonella* and *E. coli* tend to bind to the same DNA sequence. The average motif scores are

very high (≥16, except for *argT*), indicating the true σ^{54} binding sites. The motif entropy is generally lower than the flanking entropy, that is, the motif sites have less variation than the flanking regions. In particular, motifs related to *prpBCDE, argT* and *pspABCDE* have significantly lower entropy than the flanking sequences, with $p<0.03$, which indicates that these sites are probably under negative selection and that the regulation by σ^{54} is maintained across all these genomes.

Additional high-scoring motifs

We investigated in detail clusters with the minimum motif score above 17 (the complete list of clusters is provided in Additional File 2). Cluster 692 is an example of such a high score cluster (Table 3.8). The motif is located 63 bp upstream of gene *nac*, a nitrogen regulatory protein. It has been known that *nac* is regulated by *RpoN* in *E. coli* (Atkinson, Blauwkamp et al. 2002). However, this regulatory motif is absent in all *Salmonella* genomes, which is likely due to the loss of the *nac* gene (verified by protein BLAST search).

There is another interesting cluster where a 20 bp insertion occurs upstream of a high-score motif in some sequences in the alignment (Table 3.9). The 20 bp length corresponds to approximately two turns of the DNA double helix. Although the 20 bp length of the insertion could be a coincidence, it might also arise from constraints on the promoter topology. It is known that σ^{54}-RNA polymerase must interact with an activator protein that is bound to an enhancer sequence at distance from the promoter such that the intervening DNA sequence is looped-out; insertions of sizes that are multiples of the DNA helical period (~10.5 bp) change the distance between the enhancer and the promoter, but not the relative angle with respect to the DNA double helix (Dixon and Kahn 2004). Thus, the insertions of 20 bp are unlikely to disrupt the topology of the complex.

High score motifs in Clusters 1, 196 and 1560 are all found upstream of gene *gltI* (Tables 3.10 - 3.12). They actually belonged to a single cluster that was subsequently divided into three sub-clusters due to their low sequence similarity (sequence identity < 50%; see Step 3 in the Methods). Note that the total number of orthologs in these three clusters is 107, that is, one ortholog in each investigated genome. It is interesting that the motif sequences in these three sub-clusters have accumulated several mutations but they still maintain very high motif scores of above 16 (Figure 3.7). It suggests that although the sequence at this site is variable the high affinity of this site for σ^{54} is conserved, which could be indicative of negative selection and the significance of σ^{54} in the regulation of *gltI* (Zimmer, Soupene et al. 2000). It is also worth mentioning that the motif in Cluster 1 is significantly more conserved than the flanking regions, with p-value = 0.001 (Table 3.11). It again indicates the important function of the motif to the organisms over the evolutionary process.

Unconserved motif sites

In Cluster 1656, the orthologous motif sites are located inside the gene *holB*, a DNA polymerase III subunit, and they all have the same distances to the start and the end of gene, which confirms their orthology (Table 3.13 and Additional File 3). However, high score motifs (orthologs 1 to 40) only exist in most *E. coli* genomes, whereas orthologs in *Salmonella* (orthologs 63 to 107) have negative motif scores. It may indicate that this motif is conserved in *E. coli* but lost its function in *Salmonella*, or possibly the other way round. It is interesting that motif orthologs 41 to 62 in Cluster 1656, which are all from *E. coli* genomes, have moderate motif scores (<9) in the direct strand but relatively high motif scores (> 13) in the opposite strand, or opposite orientation (Table 3.13). This particular binding site is an imperfect palindrome and a

single substitution (from G to T at the second nucleotide position of motif) decreases the score in one orientation but improves the motif in the opposite orientation.

Cluster 2315 is located in the intergenic region, with about 160 bp to the start of the gene *malS* and about 156 bp to the start of the gene *bax* (Table 3.14 and Additional File 3). The orientation of the binding site points to its possible role in regulation of *bax*, which encodes a hypothetical protein of unknown function, rather than *malS*. The orthologous binding sites at this location in *Salmonella enterica* have fairly high motif scores (between 10 and 15; orthologs 63 to 105 in Table 3.14). However, the binding site is absent in *E. coli* due to single nucleotide substitutions in several key positions, yielding strongly negative PSSM scores. While the flanking regions are highly conserved, the nucleotides within the motif are significantly more variable (p-value = 0.01). It may suggest that positive selection led to the divergence between *Salmonella* and *E.coli* in this case.

Supplementary data:

Additional files are available at http://www.cmbl.uga.edu/downloads/data_sets/2015/Huang_dissertation/. All clusters listed in the files were extracted from the results of finding orthologous σ^{54} binding sites in 107 *E. coli* and *Salmonella* genomes.

Additional_File_1_Ch3_orthologs_of_known_binding_sites.xlsx

Additional file 1 contains five clusters of orthologous motifs that are involved in the regulation of five known σ^{54} dependent genes (Table 3.7).

Additional_File_2_Ch3_high_score_motifs.xlsx

Additional file 2 contains 26 clusters of high score motifs (minimum motif score > 17).

Additional_File_3_Ch3_unconserved_motifs.xlsx

Additional file 3 contains two clusters of highly unconserved motifs.

References

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Ames, G. F. L. and K. Nikaido (1985). "Nitrogen Regulation in Salmonella-Typhimurium - Identification of an Ntrc Protein-Binding Site and Definition of a Consensus Binding Sequence." Embo Journal **4**(2): 539-547.

Atkinson, M. R., T. A. Blauwkamp, et al. (2002). "Activation of the glnA, glnK, and nac promoters as Escherichia coli undergoes the transition from nitrogen excess growth to nitrogen starvation." Journal of Bacteriology **184**(19): 5358-5363.

Bailey, T. L., M. Boden, et al. (2009). "MEME SUITE: tools for motif discovery and searching." Nucleic Acids Res **37**: W202-W208.

Buske, F. A., M. Boden, et al. (2010). "Assigning roles to DNA regulatory motifs using comparative genomics." Bioinformatics **26**(7): 860-866.

Crooks, G. E., G. Hon, et al. (2004). "WebLogo: A sequence logo generator." Genome Research **14**(6): 1188-1190.

Dixon, R. and D. Kahn (2004). "Genetic regulation of biological nitrogen fixation." Nat Rev Microbiol **2**(8): 621-631.

Gruber, T. M. and C. A. Gross (2003). "Multiple sigma subunits and the partitioning of bacterial transcription space." Annual Review of Microbiology **57**: 441-466.

Gupta, S., J. A. Stamatoyannopoulos, et al. (2007). "Quantifying similarity between motifs." Genome Biology **8**(2).

Hirschman, J., P. K. Wong, et al. (1985). "Products of Nitrogen Regulatory Genes Ntra and Ntrc of Enteric Bacteria Activate Glna Transcription Invitro - Evidence That the Ntra Product Is a Sigma-Factor." Proc Natl Acad Sci U S A **82**(22): 7525-7529.

Klose, K. E. and J. J. Mekalanos (1997). "Simultaneous prevention of glutamine synthesis and high-affinity transport attenuates Salmonella typhimurium virulence." Infection and Immunity **65**(2): 587-596.

Larkin, M. A., G. Blackshields, et al. (2007). "Clustal W and clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

Mrazek, J. (2009). "Finding sequence motifs in prokaryotic genomes-a brief practical guide for a microbiologist." Briefings in Bioinformatics **10**(5): 525-536.

Mrazek, J., S. Xie, et al. (2008). "AIMIE: a web-based environment for detection and interpretation of significant sequence motifs in prokaryotic genomes." Bioinformatics **24**(8): 1041-1048.

Niehus, E., H. Gressmann, et al. (2004). "Genome-wide analysis of transcriptional hierarchy and feedback regulation in the flagellar system of Helicobacter pylori." Molecular Microbiology **52**(4): 947-961.

Palacios, S. and J. C. Escalante-Semerena (2000). "pryR, ntrA, and ihf functions are required for expression of the prpBCDE operon, encoding enzymes that catabolize propionate in

Salmonella enterica serovar Typhimurium LT2." Journal of Bacteriology **182**(4): 905-910.

Samuels, D. J., J. G. Frye, et al. (2013). "Use of a promiscuous, constitutively-active bacterial enhancer-binding protein to define the sigma(54) (RpoN) regulon of Salmonella Typhimurium LT2." Bmc Genomics **14**.

Thompson, W., E. C. Rouchka, et al. (2003). "Gibbs Recursive Sampler: finding transcription factor binding sites." Nucleic Acids Res **31**(13): 3580-3585.

Tran, N. T. L. and C. H. Huang (2014). "A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data." Biology Direct **9**.

Weiner, L., J. L. Brissette, et al. (1991). "Stress-Induced Expression of the Escherichia-Coli Phage Shock Protein Operon Is Dependent on Sigma-54 and Modulated by Positive and Negative Feedback Mechanisms." Genes & Development **5**(10): 1912-1923.

Zimmer, D. P., E. Soupene, et al. (2000). "Nitrogen regulatory protein C-controlled genes of Escherichia coli: Scavenging as a defense against nitrogen limitation." Proc Natl Acad Sci U S A **97**(26): 14674-14679.

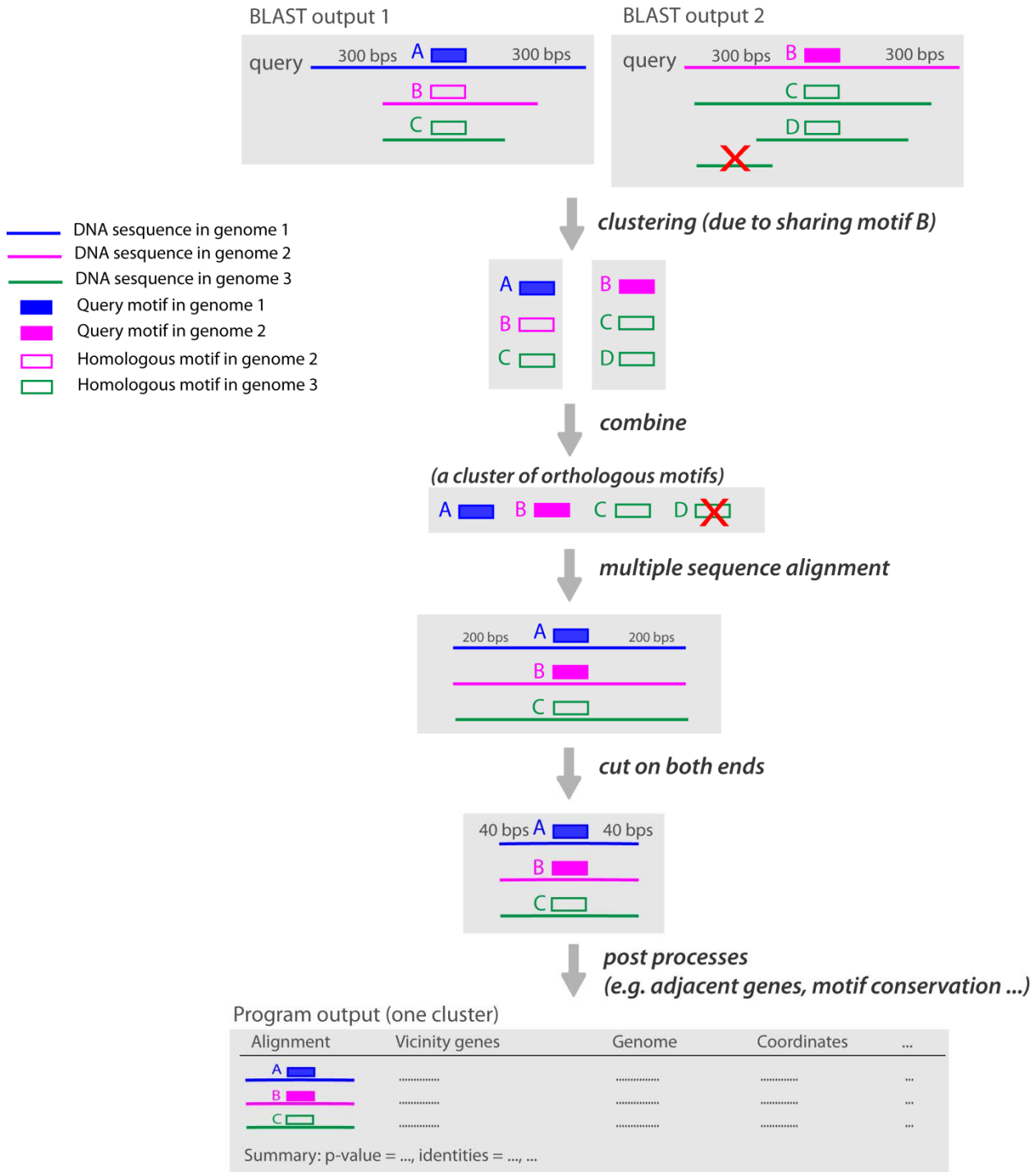

Figure 3.1 Illustration of finding orthologous motifs in three closely related genomes.

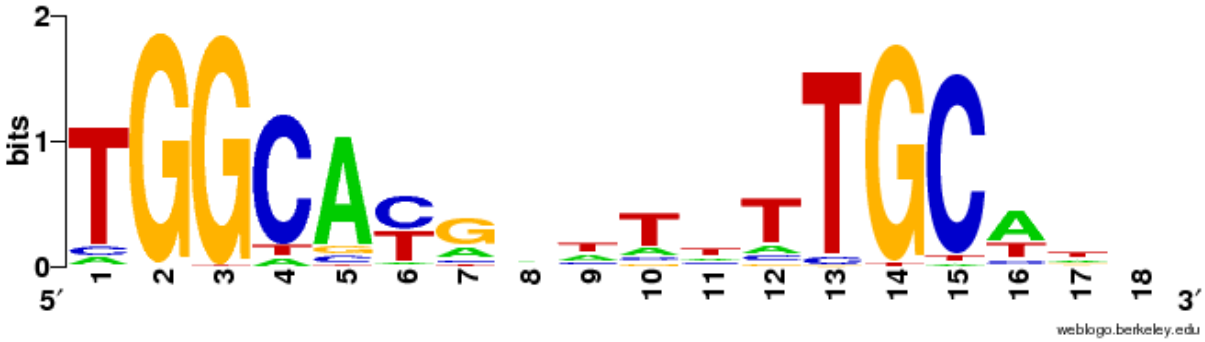

Figure 3.2 Sequence logo from the σ^{54} binding site motif in *Salmonella.* This was generated by WebLogo (Crooks, Hon et al. 2004) from the collection of 75 seed sequences used in this work(Samuels, Frye et al. 2013).

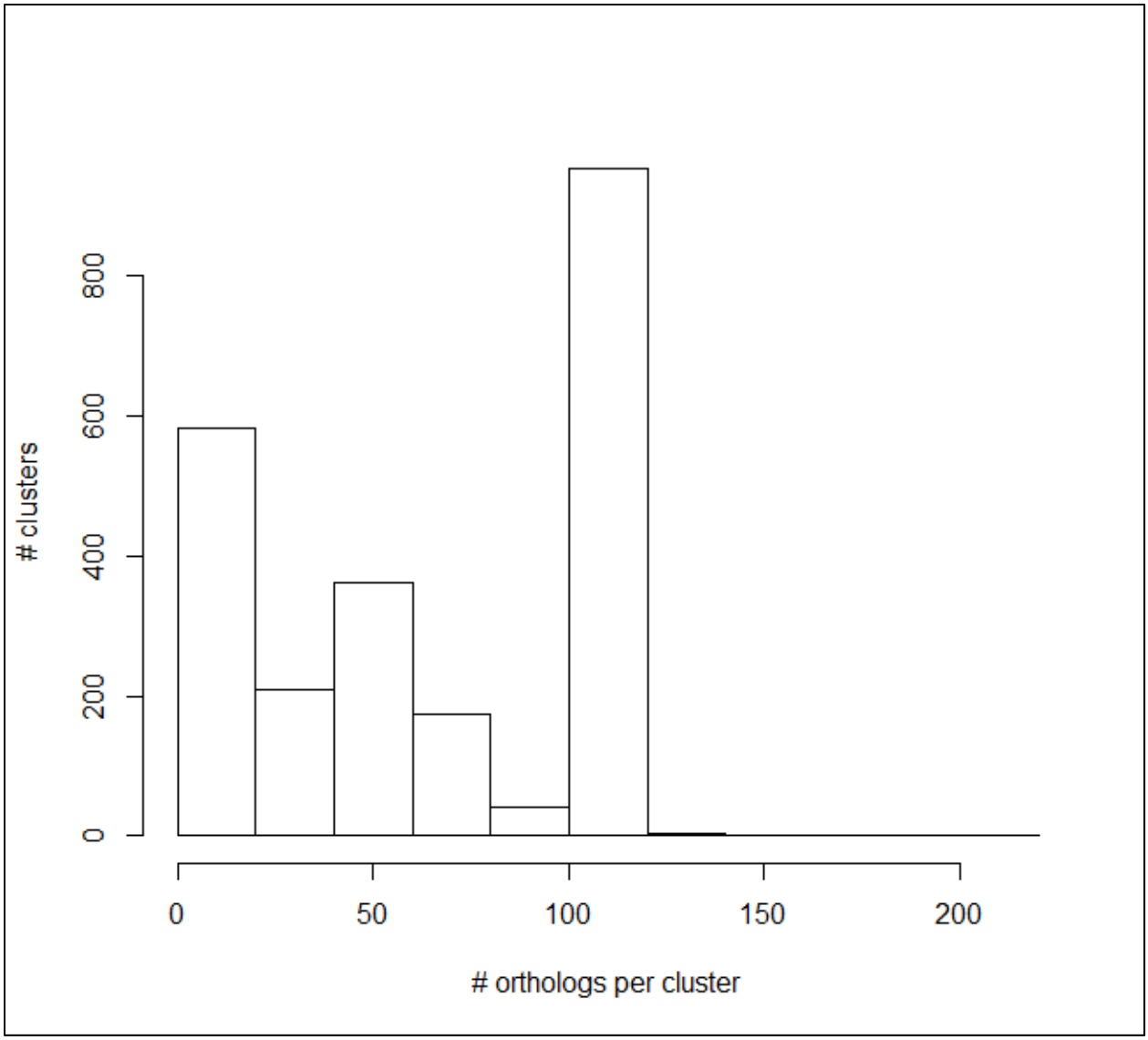

Figure 3.3 Histogram of number of orthologs per cluster in the analysis of σ^{54} binding sites in 107 *E. coli and Salmonella* genomes. The maximum motif score cutoff is set to be 12.

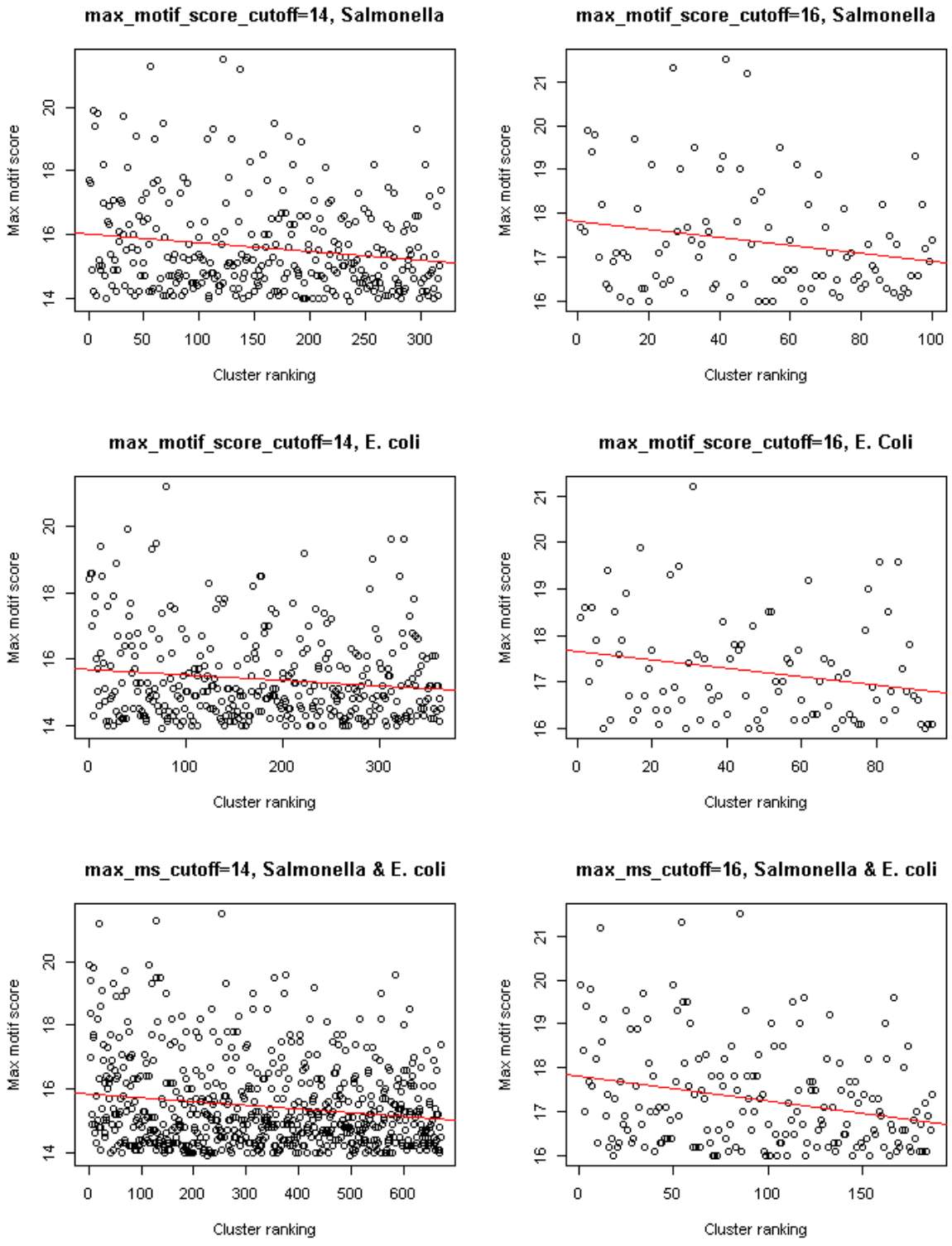

Figure 3.4 Comparisons of cluster maximum motif scores and its conservation ranking. Note that max_ms_cutoff is the maximum motif score cutoff for the output clusters. Each small circle represents one cluster. The red line is the linear regression model of maximum motif score on clustering ranking.

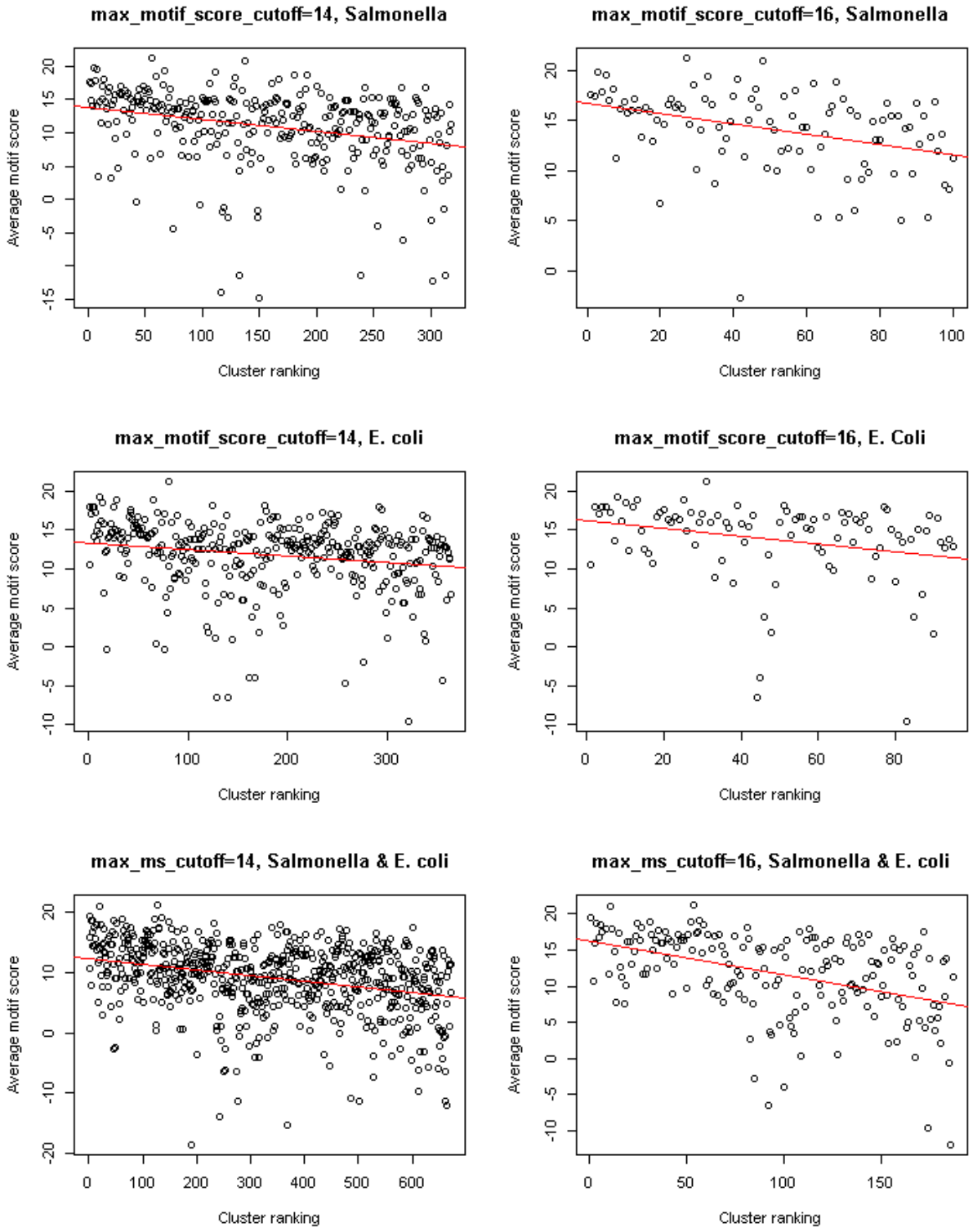

Figure 3.5 Comparisons of cluster average motif scores and its conservation ranking. Note that max_ms_cutoff is the maximum motif score cutoff for the output clusters. See Figure 3.4 for legend.

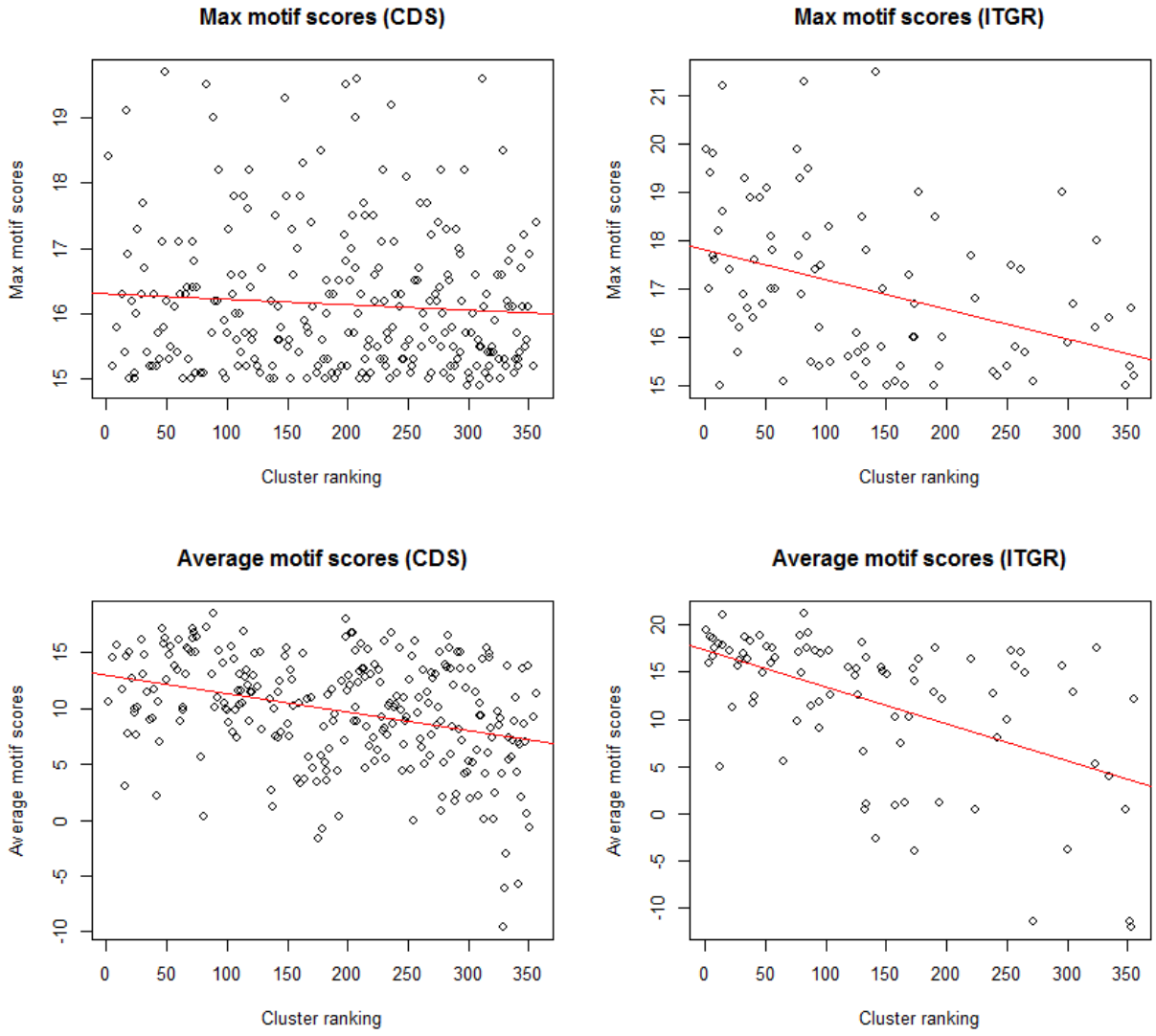

Figure 3.6 Comparisons of cluster motif score and its conservation ranking, for coding region clusters and intergenic region clusters separately. The maximum motif score cutoff is 15 and there are 107 *Salmonella* and *E coli* genomes under test. See Figure 3.4 for legend.

Cluster 1  motif score = 19

Cluster 196  motif score = 16

Cluster 1560  motif score = 18

Figure 3.7 Comparison of motif sequences from Clusters 1, 196 and 1560 in Tables 3.10 – 3.12. Clusters 1, 196 and 1560 were originally belonging to a single cluster but subsequently split into three sub-clusters due to low sequence similarity.

Table 3.1 Inputs and options for the command-line application of the program

	option	*description and default value*
input	-g	Genome input file that contains a list of genbank file directories
	-p	Motif input file for a set of aligned DNA sequence motifs, all of the same length, one each line, no spaces or gaps
options	-ms	Motif score cutoff for the Motif Locator; default = 12
	-fl	Length of flank sequence on each side of the query motif; used to construct the BLAST query sequence; default = 300 bp
	-e	BLAST e-value cutoff; default = 1e-20
	-bs	BLAST bit score cutoff; default = 150
	-bsr	BLAST bit score cutoff relative to the best hit; range = [0, 1]; default = 0.8
	-f2	Length of flank sequence used for motif conservation evaluation
	-mpi	Minimum sequence pair identity score cutoff; used as a threshold for splitting a cluster; range = (0, 100%); default = 50%
	-gapopen	Gap opening penalty used for ClustalW2 multiple sequence alignment; default = 15

Table 3.2 Summary of parameter testing for finding σ^{54} binding site orthologs in 40 *Salmonella* and *E. coli* genomes

Parameter setting [a]	*21 Salmonella genomes*			*30 E. coli genomes*			*40 Salmonella & E. coli genomes*		
	# orthologs per cluster (average)	*# clusters (total)*	*# questionable clusters* [b]	*# orthologs per cluster (average)*	*# clusters (total)*	*# questionable clusters*	*# orthologs per cluster (average)*	*# clusters (total)*	*# questionable clusters*
BLAST evalue cutoff									
1E-10	17.4	1020	31	21.2	1211	54	25.7	1923	85
1E-15	17.4	1020	31	21.2	1211	54	25.7	1923	85
1E-20	**17.4**	**1020**	**31**	**21.2**	**1213**	**56**	**25.7**	**1923**	**85**
1E-25	17.4	1020	31	21.2	1212	55	25.7	1923	85
1E-30	17.4	1020	31	21.2	1213	55	25.6	1922	86
1E-40	17.4	1018	31	21.2	1211	53	25.5	1916	86
BLAST flank length									
150	17.0	1011	31	21.1	1203	55	24.8	1907	83
200	17.4	1017	32	21.2	1208	57	25.4	1917	84
250	17.4	1019	31	21.2	1209	55	25.6	1919	84
300	**17.4**	**1020**	**31**	**21.2**	**1213**	**56**	**25.7**	**1923**	**85**
350	17.4	1021	31	21.2	1214	58	25.8	1923	85
400	17.4	1022	31	21.2	1216	58	25.8	1922	85
450	17.4	1022	31	21.2	1215	55	25.8	1924	85
500	17.4	1022	31	21.2	1217	54	25.8	1924	86
BLAST score cutoff									
100	17.4	1020	31	21.3	1212	59	25.9	1924	88
125	17.4	1020	31	21.3	1210	55	25.8	1924	87
150	**17.4**	**1020**	**31**	**21.2**	**1213**	**56**	**25.7**	**1923**	**85**
175	17.4	1018	31	21.2	1213	57	25.6	1919	87
200	17.4	1017	29	21.1	1213	54	25.3	1921	83
250	17.3	1016	28	21.0	1213	53	25.0	1918	77
BLAST score cutoff (relative to best hit)									
0.6	17.4	1020	31	21.2	1213	58	25.7	1922	84
0.7	17.4	1020	31	21.2	1213	56	25.7	1922	86
0.8	**17.4**	**1020**	**31**	**21.2**	**1213**	**56**	**25.7**	**1923**	**85**
0.85	17.4	1021	31	21.2	1213	56	25.7	1923	88
0.9	17.4	1021	31	21.2	1213	56	25.7	1923	88
Multiple sequence alignment flank length									
20	17.4	1021	31	21.2	1210	58	25.5	1924	86
30	17.4	1021	31	21.2	1213	56	25.6	1925	85
40	**17.4**	**1020**	**31**	**21.2**	**1213**	**56**	**25.7**	**1923**	**85**
50	17.4	1021	31	21.2	1216	56	25.6	1928	84
60	17.5	1019	31	21.2	1213	56	25.6	1927	84

[a] Tests are conducted by changing one parameter at a time and using default values for all other parameters. Values in **bold** type are selected to be used in the pilot study.

[b] A cluster is flagged as questionable cluster if it contains any non-ortholog motif. See Step 4 in the Method.

Table 3.3 A cluster of orthologs from the results of finding orthologous σ^{54} binding site motifs in 21 *Salmonella* genomes.

Cluster 1302 [a]

# [b]	Motif Score [c]	Multiple sequence alignment [d]	Genome
1	11.02	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_serovar_Typhi_CT18_uid57793
2	11.02	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_serovar_Choleraesuis_SC_B67_uid58017
3	11.06	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_serovar_Paratyphi_A_AKU_12601_uid59269
4	11.04	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_serovar_Enteritidis_P125109_uid59247
5	11.04	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_serovar_Gallinarum_pullorum_RKS5078_uid87035
6	11.04	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_Serovar_Cubana_CFSAN002050_uid212973
7	11.04	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_serovar_Pullorum_S06004_uid214431
8	11.02	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_serovar_Agona_24249_uid230614
9	11.03	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_serovar_Dublin_CT_02021853_uid58917
10	11.03	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_Serovar_Heidelberg_CFSAN002069_uid212974
11	11.04	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_Serovar_Typhimurium_var__5__CFSAN001921
12	11.04	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_serovar_4_5_12_i__08_1736_uid212969
13	11.05	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_serovar_Bareilly_CFSAN000189_uid212971
14	11.04	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_serovar_Bovismorbificans_3114_uid218006
15	11.04	aaccagtttcgccagcgtttcgttaggcaaccaccctgttggaataaataatgcggtgataataccccacgcgataaacgcgccgatattaggcat	Salmonella_enterica_serovar_Thompson_RM6836_uid222802
16	14.86	gacaattttggcgaaatgctcatttggcatccacccggtTGGAATAAATAATGCAGTgataaaaccccaggcaataaaagcgccgatattgggaat	Salmonella_enterica_serovar_Newport_SL254_uid58831
17	14.85	gacaattttggcgaaatgctcatttggcatccatccggtTGGAATAAATAATGCAGTgataaaaccccaggcaataaaagcgccgatattgggaat	Salmonella_enterica_serovar_Schwarzengrund_CVM19633_uid58915
18	14.86	gacaattttggcgaaatgctcgtttggcatccacccggtTGGAATAAATAATGCAGTgataaaaccccaggcaataaaagcgccgatattgggaat	Salmonella_enterica_serovar_Javiana_CFSAN001992_uid190101
19	14.67	gacaattttggcgaaatgctcatttggcatccaccctgtTGGAATAAATAATGCAGTgataaaaccccaggcaataaaagcgccgatattgggaat	Salmonella_enterica_arizonae_serovar_62_z4_z23__uid58191
20	14.65	gacaattttggcgaaatgctcattcggcatccaccctgtTGGAATAAATAATGCAGTaataaaaccccaggcaataaaagcgccaatattgggaat	Salmonella_bongori_NCTC_12419_uid70155
21	14.66	gacaattttggcgaaatgctcattcggcatccaccctgtTGGAATAAATAATGCAGTaataaaaccccaggcaataaaagcgccaatattgggaat	Salmonella_bongori_Sbon_167_uid213088

Cluster 1302 (continued)

#	Vicinity [e]	Strand	Position [f]
1	[Gene] (3970915, 3972831, -, STY4111, mannitol-specific enzyme II of phosphotransferase system)(<--); 1808 ~~~~ 108	+	[3972715, 3972732]
2	[Gene] SC3609(3817816, 3819732, +, mtlA, PTS family, mannitol-specific enzyme IIABC components)(<--); 1808 ~~~~ 108	-	[3817915, 3817932]
3	[Gene] SSPA3302(3658951, 3660867, +, mannitol-specific enzyme II of phosphotransferase system)(<--); 1808 ~~~~ 108	-	[3659050, 3659067]
4	[Gene] SEN3507(3756569, 3758485, +, mtlA, mannitol-specific enzyme II of phosphotransferase system)(<--); 1808 ~~~~ 108	-	[3756668, 3756685]
5	[Gene] SPUL_3880(3920699, 3922615, -, mtlA, mannitol-specific enzyme II of phosphotransferase system)(<--); 1808 ~~~~ 108	+	[3922499, 3922516]
6	[Gene] CFSAN002050_25395(4931884, 4933800, +, PTS mannitol transporter subunit IIABC)(<--); 1808 ~~~~ 108	-	[4931983, 4932000]
7	[Gene] I137_18615(3966942, 3968858, -, PTS mannitol transporter subunit IIABC)(<--); 1808 ~~~~ 108	+	[3968742, 3968759]
8	[Gene] Q786_18015(3760274, 3762190, +, PTS mannitol transporter subunit IIABC)(<--); 1808 ~~~~ 108	-	[3760373, 3760390]
9	[Gene] SeD_A4070(3915515, 3917431, +, 2.7.1.-, PTS system mannitol-specific transporter subunit EIICBA)(<--); 1808 ~~~~ 108	-	[3915614, 3915631]
10	[Gene] CFSAN002069_13615(2830487, 2832403, -, PTS mannitol transporter subunit IIABC)(<--); 1808 ~~~~ 108	+	[2832287, 2832304]
11	[Gene] CFSAN001921_22020(4550246, 4552162, -, PTS mannitol transporter subunit IIABC)(<--); 1808 ~~~~ 108	+	[4552046, 4552063]
12	[Gene] SE451236_02495(521432, 523348, -, PTS mannitol transporter subunit IIABC)(<--); 1808 ~~~~ 108	+	[523232, 523249]
13	[Gene] SEEB0189_01440(215404, 217320, -, PTS mannitol transporter subunit IIABC)(<--); 1808 ~~~~ 108	+	[217204, 217221]
14	[Gene] BN855_37770(3747416, 3749332, +, SBOV37741, pts system mannitol-specific eiicba component)(<--); 1808 ~~~~ 108	-	[3747515, 3747532]
15	[Gene] IA1_17910(3736975, 3738891, +, PTS mannitol transporter subunit IIABC)(<--); 1808 ~~~~ 108	-	[3737074, 3737091]
16	[Gene] SNSL254_A3320(3223586, 3224965, -, 2.7.1.-, PTS system mannitol-specific transporter subunit IICBA)(<--); 1268 ~~~~ 111	+	[3224846, 3224863]
17	[Gene] SeSA_A3253(3142488, 3143867, -, 2.7.1.-, PTS system mannitol-specific transporter subunit IICBA)(<--); 1268 ~~~~ 111	+	[3143748, 3143765]
18	[Gene] CFSAN001992_18165(3761387, 3762766, +, PTS system mannitol-specific transporter subunit IIBC)(<--); 1268 ~~~~ 111	-	[3761489, 3761506]
19	[Gene] SARI_04564(4469279, 4470658, +, hypothetical protein)(<--); 1268 ~~~~ 111	-	[4469381, 4469398]
20	[Gene] SBG_2684(2949563, 2950942, -, PTS family membrane transport system protein)(<--); 1268 ~~~~ 111	+	[2950823, 2950840]
21	[Gene] A464_3101(3095521, 3096900, -, PTS system mannitol-specific cryptic IIB component, PTS system mannitol-specific cryptic II)(<--); 1268 ~~~~ 111	+	[3096781, 3096798]

[Identity] [g]: avg = 90%, min = 77% [Motif score]: max = 14.9, min = 11.0, avg = 12.1, std = 1.7 [Avg Entropy]: pattern = 0.048, flank = 0.235 (0.315 left, 0.156 right) [Mann-Whitney U test] [h]: p = 0.167, w = 402.5

[a] The cluster is numbered by its ranking among all reported clusters in the output, which is based on the level of conservation of the query motif (See the Method). The higher the ranking, the more conserved the motif is expected to be.

[b] The serial numbers of orthologs within the cluster. Orthologs are sorted in the same order as they are in the guide tree of ClustalW2 multiple sequence alignment.

[c] Motif score of the orthologous motif sites (See more in Method).

[d] Multiple sequence alignment of the orthologous motif sites, together with 40 bp of their immediate flanking regions. The alignment was trimmed from a longer multiple sequence alignment, which is about 150 bp longer on both ends. High score motifs that are reported by Motif Locator are uppercased.

[e] Vicinity genes are the most adjacent genes that the ortholog is located in or nearby. [Gene] means the ortholog belongs to a gene; [Intergenic] means the ortholog is in the intergenic region. The arrow represents the transcription orientation of the gene relative to that of the motif ortholog. If the gene has the same orientation as the motif, --> is used; otherwise, <-- is used. The numbers next to ~~ symbol measures the distance from the central of the motif to both ends of the gene if the motif is in gene, or to the nearest gene starts or ends if the motif is in intergenic regions.

[f] Genome coordinates of the ortholog.

[g] Pairwise sequence identity of the multiple sequence alignment in the cluster.

[h] Mann-Whitney U-test on information entropies of sites within the motif v.s. those in the flanking regions in the multiple sequence alignment.

Table 3.4 Summary of pilot study results for finding σ^{54} binding site motif orthologs in *Salmonella* and *E. coli* genomes.

maximum motif score cutoff	***45 Salmonella genomes***		***62 E. coli genomes***		**107 *Salmonella* and *E. coli* genomes**	
	# clusters	# orthologs	# clusters	# orthologs	# clusters	# orthologs
12	1098	39743	1380	56181	2332	146063
14	353	13031	418	17746	732	46933
15	194	7412	216	9344	385	25190
16	110	4351	111	4446	204	12672

Table 3.5 Kendall's rank correlation tests on motif score vs ranking of the cluster.

Genomes	Maximum motif score *vs* Ranking				Average motif score vs Ranking			
	max_ms_cutoff = 14		max_ms_cutoff = 16		max_ms_cutoff = 14		max_ms_cutoff = 16	
	p-value	tau	p-value	tau	p-value	tau	p-value	tau
Salmonella	0.006	-0.10	0.048	-0.11	$< 10^{-10}$	-0.25	$< 10^{-5}$	-0.29
E. coli	0.040	-0.06	0.010	-0.17	$< 10^{-7}$	-0.19	$< 10^{-3}$	-0.24
Salmonella & *E. coli*	0.002	-0.08	0.001	-0.16	$< 10^{-15}$	-0.23	$< 10^{-10}$	-0.32

For the table on the left, correlation tests are conducted for the maximum motif score of the cluster *v.s.* the ranking of the cluster, in the analysis of finding σ^{54} binding site orthologs in 45 *Salmonella* genome, in 62 *E. coli* genomes and in 107 *Salmonella* and *E. coli* genomes. See scatter plots in Figures 3.2 and 3.3. The table on the right is similar but for the average motif score of the cluster (instead of the maximum motif score).

Table 3.6 Kendall's rank correlation tests on motif score vs ranking of in-gene clusters and intergenic clusters.

	Maximum motif score *vs* Ranking		**Average motif score vs Ranking**	
	p-value	**tau**	**p-value**	**tau**
Coding region clusters	0.2	-0.04	$< 10^{-7}$	-0.22
Intergenic region clusters	$< 10^{-4}$	-0.28	$< 10^{-6}$	-0.37

The tests are based on the output of finding orthologous σ^{54} binding sites in 107 *Salmonella* and *E. coli* genomes; the maximum motif score cutoff set to be 15. See scatter plots in Figure 3.4.

Table 3.7 Summary of orthologous motifs for five well-characterized σ^{54} binding sites (from the results of 107 *Salmonella* and *E. coli* genomes).

Known σ^{54}-dependent operons [a]			**Orthologs of σ54 binding site motifs** [b]					
Gene Symbol	*Function*	*Reference*	*Motif location relative to gene start (bp)*	*# orthologs*	*motif score (average)*	*average entropy* [c] *motif*	*flanking*	*p-value* [c]
prpBCDE	Putative propionate catabolism	(Palacios and Escalante-Semerena 2000)	53 - 55	105 [d]	16.7	**0.21**	**0.53**	**0.02**
glnHPQ	Glutamine ABC transporter	(Klose and Mekalanos 1997)	61 - 62	107	19.1	0.11	0.36	0.67
argT	Lysine/arginine/ornithine transporter	(Ames and Nikaido 1985)	75 - 76	107	14.6	**0.01**	**0.25**	**0.03**
pspABCDE	phage shock protein	(Weiner, Brissette et al. 1991)	59-63	107	15.9	**0.02**	**0.51**	**0.01**
glnALG	Glutamine synthetase	(Hirschman, Wong et al. 1985)	91 - 94	107	17.1	0.03	0.17	0.56

[a] The listed operons are known to be σ^{54}-dependent in *Salmonella* Typhimurium, *E. coli* or Enteric bacteria.

[b] For each operon, a cluster of orthologs is identified from the output of the program, which is located at upstream of the operon and have the same orientation.

[c] The average entropies are calculated for sites within the motif and for sites on the flanking regions in the multiple sequence alignment of the orthologs in the cluster. Average entropies are in **bold** type if sites within motif are significantly more conserved than flanking regions, with p-value ≤ 0.03 from the Mann-Whitney U-tests.

[d] No homologous sites are found in *E. coli* K12 DH10B and *E. coli* BW2952.

Table 3.8 Cluster 692 from the results of finding orthologous σ^{54} binding sites in 107 *Salmonella* and *E. coli* genomes.

Cluster 692

#	Motif Score	Sequence	Genome
1	18.09	gacagttggttagcttgtacatcaacaccaaaataaaac-TGGCAAGCATCTTGCAATctggttgtaagtaatggcggcacttgggccgattcttaa	Escherichia_coli_P12b_uid162061
2	18.10	gacagttggttagcttgtacatcaacaccaaaataaaac-TGGCAAGCATCTTGCAATctggttgtaagtaatggcggcacttgggccgattcttaa	Escherichia_coli_K_12_substr__MDS42_uid193705
3	18.08	gacagttggttagcttgtacatcaacaccaaaataaaac-TGGCAAGCATCTTGCAATctggttgtaagtaatggcggcacttgggccgattcttaa	Escherichia_coli_DH1_uid162051
…	…	… …	… …
25	18.08	gacagttggttagcttgtacatcaacgccaaaataaaac-TGGCAAGCATCTTGCAATctggttgtaagtaatggcggcacttgggccgattcttaa	Escherichia_coli_LY180_uid219461
26	17.12	gacagttggttagcttgtacatcaatgccaaaataaaac-TGGCAAGCATCTTGCAGTctggttgtaagtaatggcggcacttgggccgattcttaa	Escherichia_coli_IAI1_uid59377
27	18.08	gacagttggttagcttgtacatcaacgccaaaataaaac-TGGCAAGCATCTTGCAATctggttgtaagtaatggcggcacttgggccgattcttat	Escherichia_coli_UMN026_uid62981
28	18.06	gacagttggttagcttgtacatcaacgccaaaataaaac-TGGCAAGCATCTTGCAATctggttgtaagtaatggcggcacttgggccgattctcaa	Escherichia_coli_O157_H7_EDL933_uid57831
…	…	… …	… …
37	17.11	gacagttggttagcttgtacatcaacgccaaaataaaac-TGGCAAGCATCTTGCAGTctggttgtaagtaatggcggcacttgggccgattctcaa	Escherichia_coli_O26_H11_11368_uid41021
38	17.11	gacagttggttagcttgtacatcaacgccaaaataaaac-TGGCAAGCATCTTGCAGTctggttgtaagtaatggcggcacttgggccgattctcaa	Escherichia_coli_042_uid161985
39	17.11	gacagttggttagcttgtatatcaacgccaaaataaaac-TGGCAAGCATCTTGCAGTctggttgtaagtaatggcggcacttgggccgattctcaa	Escherichia_coli_ETEC_H10407_uid161993
40	18.51	gacagttggttagcttgcgcatcaacgctaaaatagaac-TGGCAAGCATCTTGCATTctggttgtaagtaatggcggcacttgggccgattcttaa	Escherichia_coli_JJ1886_uid226103
41	18.49	gacagttggttagcttgcgcatcaacgctaaaatagaac-TGGCAAGCATCTTGCATTctggttgtaagtaatggcggcacttgggccgattcttaa	Escherichia_coli_CFT073_uid57915
42	18.49	gacagttggttagcttgcgcatcaacgctaaaatagaac-TGGCAAGCATCTTGCATTctggttgtaagtaatggcggcacttgggccgattcttaa	Escherichia_coli_UTI89_uid58541
…	…	… …	… …
61	18.50	gacagttggttagcttgcgcatcaacgccaaaatagaac-TGGCAAGCATCTTGCATTctggttgtaagtaatggcggcacttgggccgactcttaa	Escherichia_coli_O104_H4_2011C_3493_uid176127
62	18.50	gacagttggttagcttgcgcatcaacgccaaaatagaac-TGGCAAGCATCTTGCATTctggttgtaagtaatggcggcacttgggccgactcttaa	Escherichia_coli_O104_H4_2009EL_2071_uid176128
63	17.26	gacggctg-ccatcctgcgt-ctgatgcggtgtgaaaaccTGGCAGGCATTTTGCTAA-taattttaagatgtgacgccatagagg-------ttaa	Salmonella_bongori_NCTC_12419_uid70155

Cluster 692 (continued)

#	Motif Score	Vicinity
1	18.09	[Intergenic] P12B_t0023(1128287, 1128362, -, tRNA-Asn)(<--); 265 ~~~~ 63; P12B_c1029(1128690, 1129607, +, Nitrogen assimilation regulatory protein nac)(-->)
2	18.10	[Intergenic] ECMDS42_1617(1699075, 1699150, +, asnV, tRNA-Asn)(<--); 264 ~~~~ 63; ECMDS42_1616(1697831, 1698748, -, nac, DNA-binding transcriptional dual regulator)(-->)
3	18.08	[Intergenic] ECDH1ME8569_t0034(2046063, 2046138, +, asnV, tRNA-Asn)(<--); 264 ~~~~ 63; ECDH1ME8569_1925(2044819, 2045736, -, nac, nitrogen assimilation transcriptional regulator)(-->)
…	…	… …
25	18.08	[Intergenic] LY180_10305(2153396, 2153471, +, tRNA-Asn)(<--); 264 ~~~~ 63; LY180_10300(2152152, 2153069, -, LysR family transcriptional regulator)(-->)
26	17.12	[Intergenic] ECIAI1_tRNA19(2124895, 2124970, +, tRNA-Asn)(<--); 264 ~~~~ 63; ECIAI1_2069(2123651, 2124568, -, nac, nitrogen assimilation transcriptional regulator)(-->)
27	18.08	[Intergenic] ECUMN_tRNA19(2361703, 2361778, +, tRNA-Asn)(<--); 263 ~~~~ 63; ECUMN_2325(2360460, 2361377, -, nac, nitrogen assimilation transcriptional regulator)(-->)
28	18.06	[Intergenic] Z3149(2808094, 2808169, +, asnV, tRNA-Asn)(<--); 263 ~~~~ 63; Z3147(2806851, 2807768, -, nac, nitrogen assimilation transcriptional regulator)(-->)
…	…	… …
37	17.11	[Intergenic] ECO26_tRNA048(2796974, 2797049, +, asnV, tRNA-Asn)(<--); 263 ~~~~ 63; ECO26_2879(2795731, 2796648, -, nac, nitrogen assimilation transcriptional regulator)(-->)
38	17.11	[Intergenic] EC042_2228(2317726, 2318661, -, erfK, hypothetical protein)(-->); 394 ~~~~ 63; EC042_2227(2316352, 2317269, -, nac, tRNA-Asn)(-->)
39	17.11	[Intergenic] ETEC_t038(2266599, 2266671, +, tRNA-Asn)(<--); 263 ~~~~ 63; ETEC_2099(2265356, 2266273, -, nitrogen assimilation regulatory protein)(-->)
40	18.51	[Intergenic] P423_11150(2228262, 2228337, +, tRNA-Asn)(<--); 264 ~~~~ 63; P423_11145(2227018, 2227935, -, LysR family transcriptional regulator)(-->)
41	18.49	[Intergenic] c5548(2258516, 2258591, -, asnV, tRNA-Asn)(<--); 264 ~~~~ 63; c2446(2258918, 2259835, +, nac, nitrogen assimilation transcriptional regulator)(-->)
42	18.49	[Intergenic] UTI89_C2203(2113924, 2113996, +, asnV, tRNA-Asn)(<--); 264 ~~~~ 63; UTI89_C2202(2112680, 2113597, -, nac, nitrogen assimilation transcriptional regulator)(-->)
…	…	… …
61	18.50	[Intergenic] O3K_t25390(2009796, 2009871, -, tRNA-Asn)(<--); 264 ~~~~ 63; O3K_09625(2010198, 2011115, +, nitrogen assimilation transcriptional regulator)(-->)
62	18.50	[Intergenic] O3O_t25764(2005527, 2005602, -, tRNA-Asn)(<--); 264 ~~~~ 63; O3O_16000(2005929, 2006846, +, nitrogen assimilation transcriptional regulator)(-->)
63	17.26	[Intergenic] SBG_t035(1998980, 1999052, +, tRNA-Asn, tRNA-Asn)(<--); 230 ~~~~ 53; SBG_1839(1998302, 1998697, -, nitrogen assimilation regulatory protein (pseudogene))(-->)

[Identity]: avg = 95%, min = 51% [Motif score]: max = 18.5, min = 17.1, avg = 18.1, std = 0.4 [Avg Entropy]: pattern = 0.110, flank = 0.135 (0.173 left, 0.096 right) [Mann-Whitney U test]: p = 0.12, w = 432.0

See legend in Table 3.3.

Table 3.9 Cluster 348 from the results of finding orthologous σ^{54} binding sites in 107 *Salmonella* and *E. coli* genomes.

Cluster 348

#	Motif Score	Sequence	Genome
1	21.25	ttatttttataagtaattgattatttt----------gataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcgtatttttaccataaacattaa	Salmonella_enterica_serovar_Heidelberg_SL476
2	21.26	ttatttttataagtaattgattatttt----------gataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcgtatttttaccataaacattaa	Salmonella_enterica_serovar_Dublin_CT_02021853
3	21.24	ttatttttataagtaattgattatttt----------gataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcgtatttttaccataaacattaa	Salmonella_enterica_serovar_Gallinarum_287_91
4	21.25	ttatttttataagtaattgattatttt----------gataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcgtatttttaccataaacattaa	Salmonella_enterica_serovar_Enteritidis_P125109
5	21.26	ttatttttataagtaattgattatttt----------gataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcgtatttttaccataaacattaa	Salmonella_enterica_serovar_Heidelberg_B182
6	21.26	ttatttttataagtaattgattatttt----------gataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcgtatttttaccataaacattaa	Salmonella_enterica_serovar_Heidelberg_41578
7	21.22	ttatttttataagtaattgattatttt----------gataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcgtatttttaccataaacattaa	Salmonella_enterica_Serovar_Heidelberg_CFSAN002069
8	19.69	ttatttttataagtaattgattatttt----------gataaaaaatagTGGCATGCCTTTTGCCTTatccc-tgtaaagaatgcgtatttttaccataaacattaa	Salmonella_enterica_Serovar_Cubana_CFSAN002050
9	21.23	ttatttttataagtaattgattatttt----------tataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcggatttttaccataaacattaa	Salmonella_enterica_serovar_Paratyphi_A_ATCC_9150
10	21.23	ttatttttataagtaattgattatttt----------tataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcggatttttaccataaacattaa	Salmonella_enterica_serovar_Paratyphi_A_AKU_12601
11	21.27	ttatttttataagtaattgattatttt----------tataaaaaatagTGGCATGCCTTTTGCTTTatcccctgtaaagaatgcggatttttaccataaacattaa	Salmonella_enterica_serovar_Schwarzengrund…
…	…	… …	… …
37	21.25	ttatttttataagtaattgattatgat----------gataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcggatttttaccataaacattaa	Salmonella_enterica_serovar_Choleraesuis_SC_B67
38	19.68	ttatttttataagtaattgattatgat----------gataaaaaatagTGGCATGCCTTTTGCCTTatccc-tgtaaagaatgcggatttttaccataaacattaa	Salmonella_enterica_serovar_Paratyphi_C_RKS4594
39	21.25	ttatttttataagtaattgattatgat----------gataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcggatttttaccataaacattaa	Salmonella_enterica_serovar_Newport_SL254
40	21.22	ttatttttataagtaattgattatgat----------gataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcggatttttaccataaacattaa	Salmonella_enterica_serovar_Newport_USMARC
41	21.21	ttatttttataagtaattgattatttttataagtaattgattattttgataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcgtatttttaccataaacattaa	Salmonella_enterica_serovar_Gallinarum_pullorum_
42	21.20	ttatttttataagtaattgattatttttataagtaattgattattttgataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcgtatttttaccataaacattaa	Salmonella_enterica_serovar_Pullorum_S06004
43	21.21	ttatttttataagtaattgattatttttataagtaattgattattttgataaaaaatagTGGCATGCCTTTTGCTTTatccc-tgtaaagaatgcgtatttttaccataaacattaa	Salmonella_enterica_serovar_Gallinarum_Pullorum

Cluster 348 (continued)

#	Motif Score	Vicinity
1	21.25	[Intergenic] SeHA_C2603(2517643, 2519070, +, regulatory protein)(<--); 150 ~~~~ 61; SeHA_C2602(2516035, 2517432, -, diaminopimelate decarboxylase)(-->)
2	21.26	[Intergenic] SeD_A2712(2594036, 2595463, +, regulatory protein)(<--); 150 ~~~~ 61; SeD_A2711(2592428, 2593825, -, diaminopimelate decarboxylase)(-->)
3	21.24	[Intergenic] SG2390(2457807, 2459233, +, pseudo,)(<--); 150 ~~~~ 61; SG2389(2456199, 2457596, -, amino acid decarboxylase)(-->)
4	21.25	[Intergenic] SEN2343(2466939, 2468366, +, transcriptional regulator)(<--); 150 ~~~~ 61; SEN2342(2465331, 2466728, -, amino acid decarboxylase)(-->)
5	21.26	[Intergenic] SU5_02956(3205739, 3207166, +, sigma(54)-Dependent Activator)(<--); 150 ~~~~ 61; SU5_02955(3204131, 3205528, -, 4.1.1.20, diaminopimelate decarboxylase)(-->)
6	21.26	[Intergenic] SEEH1578_21080(4266163, 4267590, +, sigma(54)-Dependent Activator)(<--); 150 ~~~~ 61; SEEH1578_21075(4264555, 4265952, -, diaminopimelate decarboxylase)(-->)
7	21.22	[Intergenic] CFSAN002069_20045(4202615, 4204042, -, regulatory protein)(<--); 150 ~~~~ 61; CFSAN002069_20050(4204253, 4205650, +, diaminopimelate decarboxylase)(-->)
8	19.69	[Intergenic] CFSAN002050_18735(3525166, 3526593, +, regulatory protein)(<--); 150 ~~~~ 61; CFSAN002050_18730(3523558, 3524955, -, diaminopimelate decarboxylase)(-->)
9	21.23	[Intergenic] SPA0503(573609, 575036, -, transcriptional regulator)(<--); 150 ~~~~ 61; SPA0504(575247, 576644, +, amino acid decarboxylase)(-->)
10	21.23	[Intergenic] SSPA0467(573643, 575070, -, transcriptional regulator)(<--); 150 ~~~~ 61; SSPA0468(575281, 576678, +, amino acid decarboxylase)(-->)
11	21.27	[Intergenic] SeSA_A2591(2475971, 2477398, +, regulatory protein)(<--); 150 ~~~~ 62; SeSA_A2590(2474362, 2475759, -, diaminopimelate decarboxylase)(-->)
…	…	… …
37	21.25	[Intergenic] SC2363(2487283, 2488710, +, rocR, regulatory protein)(<--); 150 ~~~~ 61; SC2362(2485675, 2487072, -, dcdA, diaminopimelate decarboxylase)(-->)
38	19.68	[Intergenic] SPC_1345(1405913, 1407340, -, transcriptional regulator)(<--); 150 ~~~~ 61; SPC_1346(1407551, 1408948, +, amino acid decarboxylase)(-->)
39	21.25	[Intergenic] SNSL254_A2549(2471843, 2473270, +, regulatory protein)(<--); 150 ~~~~ 61; SNSL254_A2548(2470235, 2471632, -, diaminopimelate decarboxylase)(-->)
40	21.22	[Intergenic] SN31241_34680(3497441, 3498868, +, regulatory protein)(<--); 150 ~~~~ 76; SN31241_34670(3495833, 3497215, -, Diaminopimelate decarboxylase)(-->)
41	21.21	[Intergenic] SPUL_0524(575644, 577070, -, pseudo,)(<--); 170 ~~~~ 61; SPUL_0525(577301, 578698, +, putative amino acid decarboxylase)(-->)
42	21.20	[Intergenic] I137_02415(575660, 577087, -, regulatory protein)(<--); 170 ~~~~ 61; I137_02420(577318, 578715, +, diaminopimelate decarboxylase)(-->)
43	21.21	[Intergenic] SPUCDC_0524(575605, 577031, -, pseudo,)(<--); 170 ~~~~ 61; SPUCDC_0525(577262, 578659, +, putative amino acid decarboxylase)(-->)

[Identity]: avg = 96%, min = 79% [Motif score]: max = 21.3, min = 19.7, avg = 21.2, std = 0.3 [Avg Entropy]: pattern = 0.015, flank = 0.062 (0.104 left, 0.021 right) [Mann-Whitney U test]: p = 6.09e-01, w = 341.0

See legend in Table 3.3.

Table 3.10 Cluster 1 from the results of finding orthologous σ^{54} binding sites in 107 *Salmonella* and *E. coli* genomes.

Cluster 1

#	Motif Score	Sequence	Genome
1	18.77	cgccggatggttcttatccggcctactctccttcgtcattcTGGCACGTCTGTTGCTTTgttatatgtggtaaacccgttttggctatcagtgccacc	Salmonella_bongori_NCTC_12419_uid70155
2	18.76	cgccggatggttcttatccggcctactctccttcgtcattcTGGCACGTCTGTTGCTTTgttatatgtggtaaacccgttttggctatcagtgccacc	Salmonella_bongori_Sbon_167_uid213088
3	19.88	catttatccggcctgaa--aatcctccttttttccccattcTGGCACGTCTATTGCTTTgttaaacatggcaaacctgttctggctgcttgtgcaacc	Salmonella_enterica_serovar_Paratyphi_A_ATCC_9150_uid58201
4	19.88	catttatccggcctgaa--aatcctccttttttccccattcTGGCACGTCTATTGCTTTgttaaacatggcaaacctgttctggctgcttgtgcaacc	Salmonella_enterica_serovar_Paratyphi_A_AKU_12601_uid59269
5	19.88	catttatccggcctgaa--aatcctccttttttcctcattcTGGCACGTCTATTGCTTTgttaaacatggcaaacctgttctggctgcttgtgcaacc	Salmonella_enterica_serovar_Paratyphi_B_SPB7_uid59097
6	19.89	catttatccggcctgaa--aatcctccttttttcctcattcTGGCACGTCTATTGCTTTgttaaacatggcaaacctgttctggctgcttgtgcaacc	Salmonella_enterica_serovar_Thompson_RM6836_uid222802
7	19.89	catttatccggcctgaa--aatcctccttttttccctcattcTGGCACGTCTATTGCTTTgttaaacatggcaaacctgttctggctgcttgtgcaacc	Salmonella_enterica_serovar_Heidelberg_SL476_uid58973
…	…	… …	… …
39	19.89	catttatccggcctgaa--aatcctccttttttcccccattcTGGCACGTCTATTGCTTTgttaaacatggcaaacctgttctggctgcttgtgcaacc	Salmonella_enterica_serovar_Typhimurium_UK_1_uid87049
40	19.89	catttatccggcctgaa--aatcctccttttttcccccattcTGGCACGTCTATTGCTTTgttaaacatggcaaacctgttctggctgcttgtgcaacc	Salmonella_enterica_serovar_Typhimurium_T000240_uid84397
41	19.80	catttatccggcctgaa--aatcctccttttttcccccattcTGGCACGTCTATTGCTTTgttaaacatggcaaacctgttctggctgcttgtgcaacc	Salmonella_enterica_arizonae_serovar_62_z4_z23__uid58191
42	18.55	cagtgcattatccggata-acaatattctcctatccaacttTGGCACATCTATTGCTTTgttatacaaggcaaaccctgaaacagcatcagtacaacc	Escherichia_coli_K_12_substr__MG1655_uid57779
43	18.55	cagtgcattatccggata-acaatattctcctatccaacttTGGCACATCTATTGCTTTgttatacaaggcaaaccctgaaacagcatcagtacaacc	Escherichia_coli_K_12_substr__W3110_uid161931
44	18.55	cagtgcattatccggata-acaatattctcctatccaacttTGGCACATCTATTGCTTTgttatacaaggcaaaccctgaaacagcatcagtacaacc	Escherichia_coli_K_12_substr__DH10B_uid58979
45	18.56	cagtgcattatccggata-acaatattctcctatccaacttTGGCACATCTATTGCTTTgttatacaaggcaaaccctgaaacagcatcagtacaacc	Escherichia_coli_BW2952_uid59391
46	18.55	cagtgcattatccggata-acaatattctcctatccaacttTGGCACATCTATTGCTTTgttatacaaggcaaaccctgaaacagcatcagtacaacc	Escherichia_coli_DH1_uid161951
…	…	… …	… …
60	18.52	cagtgcgttatccggatg-acaaaattcccctatccaacttTGGCACATCTATTGCTTTgttatacaaggcaaagcttgaacccgcatcagtacaacc	Escherichia_coli_O7_K1_CE10_uid162115
61	18.55	cagtgcgttatccggata-acaaaattcccctatccaacttTGGCACATCTATTGCTTTgttatacaaggcaaagcttgaaccagcatcagtacaacc	Escherichia_coli_UMN026_uid62981

Cluster 1 (continued)

#	Motif Score	Vicinity
1	18.77	[Intergenic] SBG_0564(630705, 632243, -, lnt, apolipoprotein N-acyltransferase)(-->); 98 ~~~~ 264; SBG_0562(629417, 630343, -, ybeJ, ABC transporter periplasmic binding protein)(-->)
2	18.76	[Intergenic] A464_632(635044, 636582, -, Apolipoprotein N-acyl transferase)(-->); 98 ~~~~ 282; A464_631(633756, 634664, -, Glutamate Aspartate periplasmic binding protein precursor GltI)(-->)
3	19.88	[Intergenic] SPA2074(2158402, 2159940, +, lnt, apolipoprotein N-acyltransferase)(-->); 73 ~~~~ 264; SPA2075(2160277, 2161203, +, ybeJ, ABC transporter substrate-binding protein)(-->)
4	19.88	[Intergenic] SSPA1927(2153586, 2155124, +, lnt, apolipoprotein N-acyltransferase)(-->); 73 ~~~~ 264; SSPA1928(2155461, 2156387, +, glutamate and aspartate transporter subunit)(-->)
5	19.88	[Intergenic] SPAB_02883(2393248, 2394786, +, lnt, apolipoprotein N-acyltransferase)(-->); 73 ~~~~ 282; SPAB_02884(2395141, 2396049, +, glutamate and aspartate transporter subunit)(-->)
6	19.89	[Intergenic] IA1_03485(750678, 752216, -, lnt, apolipoprotein N-acyltransferase)(-->); 73 ~~~~ 282; IA1_03480(749415, 750323, -, amino acid transporter)(-->)
7	19.89	[Intergenic] SeHA_C0784(778588, 780177, -, lnt, 2.3.1.-, apolipoprotein N-acyltransferase)(-->); 73 ~~~~ 282; SeHA_C0783(777325, 778233, -, gltI, glutamate and aspartate transporter subunit)(-->)
…	…	… …
39	19.89	[Intergenic] STMUK_0671(730668, 732206, -, lnt, apolipoprotein N-acyltransferase)(-->); 73 ~~~~ 264; STMUK_0670(729405, 730331, -, gltI, glutamate and aspartate transporter subunit)(-->)
40	19.89	[Intergenic] STMDT12_C07290(768225, 769763, -, lnt, apolipoprotein N-acyltransferase)(-->); 73 ~~~~ 282; STMDT12_C07280(766962, 767870, -, ABC transporter substrate binding protein)(-->)
41	19.80	[Intergenic] SARI_02276(2204204, 2205742, +, lnt, apolipoprotein N-acyltransferase)(-->); 73 ~~~~ 264; SARI_02277(2206079, 2207005, +, glutamate and aspartate transporter subunit)(-->)
42	18.55	[Intergenic] b0657(688566, 690104, -, lnt, 2.3.1.-, apolipoprotein N-acyltransferase)(-->); 108 ~~~~ 222; b0656(687220, 688236, -, insH1, IS5 transposase and trans-activator)(-->)
43	18.55	[Intergenic] Y75_p0647(689765, 691303, -, lnt, apolipoprotein N-acyltransferase)(-->); 108 ~~~~ 222; Y75_p0646(688419, 689435, -, insH, IS5 transposase and trans-activator)(-->)
44	18.55	[Intergenic] ECDH10B_0726(741158, 742696, -, lnt, apolipoprotein N-acyltransferase)(-->); 108 ~~~~ 222; ECDH10B_0725(739812, 740828, -, IS5 transposase and trans-activator)(-->)
45	18.56	[Intergenic] BWG_0528(591326, 592864, -, lnt, apolipoprotein N-acyltransferase)(-->); 108 ~~~~ 222; BWG_0527(589980, 590996, -, insH, IS5 transposase and trans-activator)(-->)
46	18.55	[Intergenic] EcDH1_2969(3190238, 3191776, +, apolipoprotein N-acyltransferase)(-->); 108 ~~~~ 258; EcDH1_2970(3192142, 3193122, +, transposase IS4 family protein)(-->)
…	…	… …
60	18.52	[Intergenic] CE10_0647(699039, 700577, -, lnt, apolipoprotein N-acyltransferase)(-->); 108 ~~~~ 289; CE10_0646(697734, 698642, -, gltI, glutamate and aspartate transporter subunit)(-->)
61	18.55	[Intergenic] ECUMN_0750(801649, 803187, -, lnt, apolipoprotein N-acyltransferase)(-->); 108 ~~~~ 289; ECUMN_0749(800344, 801252, -, gltI, glutamate and aspartate transporter subunit)(-->)

[Identity]: avg = 78%, min = 53% [Motif score]: max = 19.9, min = 18.5, avg = 19.4, std = 0.6 [Avg Entropy]: pattern = 0.062, flank = 0.576 (0.705 left, 0.447 right) [Mann-Whitney U test]: p = 1.28e-03, w = 519.5

See legend in Table 3.3.

Table 3.11 Cluster 196 from the results of finding orthologous σ^{54} binding sites in 107 *Salmonella* and *E. coli* genomes.

Cluster 196

#	Motif Score	Sequence	Genome
1	17.71	gaaattacaatcgtaattaattcactcctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcgatgcaacc	Escherichia_coli_KO11FL_uid52593
2	17.73	gaaattacaatcgtaattaattcactcctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcgatgcaacc	Escherichia_coli_W_uid162011
3	17.66	gaaattacaatcgtaattaattcactcctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcgatgcaacc	Escherichia_coli_KO11FL_uid162099
4	17.73	gaaattacaatcgtaattaattcactcctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcgatgcaacc	Escherichia_coli_W_uid162101
5	17.73	gaaattacaatcgtaattaattcactcctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcgatgcaacc	Escherichia_coli_LY180_uid219461
6	17.74	gaaattacaatcgtaattaattcacttctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcaaccgcgacgcaacc	Escherichia_coli_HS_uid58393
7	17.69	gaaattacaatcgtaattaattcacccctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcaacgcaaac	Escherichia_coli_CFT073_uid57915
8	17.71	gaaattacaatcgtaattaattcacccctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcaacgcaaac	Escherichia_coli_ABU_83972_uid161975
9	17.70	gaaattacaatcgtaattaattcacccctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcaacgcaaac	Escherichia_coli__clone_D_i2__uid162047
10	17.70	gaaattacaatcgtaattaattcacccctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcaacgcaaac	Escherichia_coli__clone_D_i14__uid162049
11	15.90	gaaattacaatcgtaattaattcacccctaagtatttttTGGCATGTAAGTTGCTCTatttactgtgtaagcaatgaaatcagccgcaacgcaacc	Escherichia_coli_APEC_O1_uid58623
12	15.91	gaaattacaatcgtaattaattcacccctaagtatttttTGGCATGTAAGTTGCTCTatttactgtgtaagcaatgaaatcagccgcaacgcaacc	Escherichia_coli_S88_uid62979
13	17.74	gaaattacaatcgtaattaattcacccctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcaacgcaacc	Escherichia_coli_ED1a_uid59379
14	17.69	gaaattacaatcgtaattaattcacccctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcaacgcaacc	Escherichia_coli_UTI89_uid58541
15	17.71	gaaattacaatcgtaattaattcacccctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcaacgcaacc	Escherichia_coli_IHE3034_uid162007
…	…	… …	… …
42	17.73	gaaattacaatcgtaattaattcacccctaagtatttttTGGCATGTAAGTTGCTTTatttactgtgtaagcaatgaaatcagccgcgacgcaacc	Escherichia_coli_SE11_uid59425

Cluster 196 (continued)

#	Motif Score	Vicinity
1	17.71	[Intergenic] EKO11_3210(3351020, 3352896, -, pseudo,)(<--); 166 ~~~~ 289; EKO11_3211(3353351, 3354259, +, family 3 extracellular solute-binding protein)(-->)
2	17.73	[Intergenic] ECW_m0711(759002, 760878, +, pseudo,)(<--); 166 ~~~~ 289; ECW_m0710(757639, 758547, -, gltI, glutamate and aspartate transporter subunit)(-->)
3	17.66	[Intergenic] KO11_20380(4321174, 4322712, +, lnt, apolipoprotein N-acyltransferase)(-->); 2120 ~~~~ 289; KO11_20395(4325121, 4326029, +, gltI, glutamate and aspartate transporter subunit)(-->)
4	17.73	[Intergenic] WFL_03540(760962, 762500, -, lnt, apolipoprotein N-acyltransferase)(-->); 2120 ~~~~ 289; WFL_03525(757645, 758553, -, gltI, glutamate and aspartate transporter subunit)(-->)
5	17.73	[Intergenic] LY180_03530(752590, 753387, +, rhomboid family protein)(<--); 166 ~~~~ 289; LY180_03525(751227, 752135, -, amino acid transporter)(-->)
6	17.74	[Intergenic] EcHS_A0702(716350, 718226, +, pseudo,)(<--); 166 ~~~~ 32; EcHS_A0701(716030, 716152, +, hypothetical protein)(<--)
7	17.69	[Intergenic] c0741(720260, 722137, +, hypothetical protein)(<--); 165 ~~~~ 32; c0740(719941, 720063, +, hypothetical protein)(<--)
8	17.71	[Intergenic] ECABU_c07040(729264, 731141, +, intramembrane serine protease rhomboid family)(<--); 165 ~~~~ 12; ECABU_c07030(728926, 729087, +, hypothetical protein)(<--)
9	17.70	[Intergenic] i02_0710(721070, 722947, +, hypothetical protein)(<--); 165 ~~~~ 32; i02_0709(720751, 720873, +, hypothetical protein)(<--)
10	17.70	[Intergenic] i14_0710(721070, 722947, +, hypothetical protein)(<--); 165 ~~~~ 32; i14_0709(720751, 720873, +, hypothetical protein)(<--)
11	15.90	[Intergenic] APECO1_1407(663720, 665597, +, hypothetical protein)(<--); 165 ~~~~ 220; APECO1_1408(662358, 663335, -, ybeJ, glutamate and aspartate transporter subunit)(-->)
12	15.91	[Intergenic] ECS88_0691(702153, 704030, +, hypothetical protein)(<--); 165 ~~~~ 289; ECS88_0690(700791, 701699, -, gltI, glutamate and aspartate transporter subunit)(-->)
13	17.74	[Intergenic] ECED1_0646(666816, 668693, +, hypothetical protein)(<--); 165 ~~~~ 289; ECED1_0645(665454, 666362, -, gltI, glutamate and aspartate transporter subunit)(-->)
14	17.69	[Intergenic] UTI89_C0653(662120, 663997, +, hypothetical protein)(<--); 165 ~~~~ 32; UTI89_C0652(661801, 661923, +, hypothetical protein)(<--)
15	17.71	[Intergenic] ECOK1_0660(698796, 700673, +, 3.4.21.-, peptidase, S54 (rhomboid) family)(<--); 165 ~~~~ 32; ECOK1_0659(698477, 698599, +, hypothetical protein)(<--)
…	…	… …
42	17.73	[Intergenic] ECSE_0727(773093, 774970, +, hypothetical protein)(<--); 166 ~~~~ 32; ECSE_0726(772773, 772895, +, hypothetical protein)(<--)

[Identity]: avg = 97%, min = 93% [Motif score]: max = 17.8, min = 15.9, avg = 17.6, std = 0.4 [Avg Entropy]: pattern = 0.015, flank = 0.082 (0.072 left, 0.092 right) [Mann-Whitney U test]: p = 4.11e-01, w = 377.0

See legend in Table 3.3.

Table 3.12 Cluster 1560 from the results of finding orthologous σ^{54} binding sites in 107 *Salmonella* and *E. coli* genomes.

Cluster 1560

#	Motif Score	Sequence	Genome
1	18.02	tcaggcctacaacctctcaccgccaccaaaatctcattcTGGCACGCCAATTGCTTAatataatcaagaggctgcacactccctaccggcgcaacc	Salmonella_enterica_serovar_Typhi_CT18_uid57793
2	18.02	tcaggcctacaacctctcaccgccaccaaaatctcattcTGGCACGCCAATTGCTTAatataatcaagaggctgcacactccctaccggcgcaacc	Salmonella_enterica_serovar_Typhi_Ty2_uid57973
3	18.02	tcaggcctacaacctctcaccgccaccaaaatctcattcTGGCACGCCAATTGCTTAatataatcaagaggctgcacactccctaccggcgcaacc	Salmonella_enterica_serovar_Typhi_Ty21a_uid201427
4	18.02	tcaggcctacaacctctcaccgccaccaaaatctcattcTGGCACGCCAATTGCTTAatataatcaagaggctgcacactccctaccggcgcaacc	Salmonella_enterica_serovar_Typhi_P_stx_12_uid87001

Cluster 1560 (continued)

#	Motif Score	Vicinity
1	18.02	[Intergenic] (706916, 708454, -, STY0711, apolipoprotein N-acyltransferase)(-->); 136 ~~~~ 264; (705590, 706516, -, STY0710, ABC transporter periplasmic binding protein (glutamate/aspartate?))(-->)
2	18.02	[Intergenic] t2207(2272617, 2274155, +, lnt, apolipoprotein N-acyltransferase)(-->); 136 ~~~~ 282; t2208(2274573, 2275481, +, ybeJ, glutamate and aspartate transporter subunit)(-->)
3	18.02	[Intergenic] TY21A_11190(2272617, 2274155, +, lnt, apolipoprotein N-acyltransferase)(-->); 136 ~~~~ 282; TY21A_11195(2274573, 2275481, +, glutamate and aspartate transporter subunit)(-->)
4	18.02	[Gene] STBHUCCB_23320(2269331, 2270920, +, Apolipoprotein N-acyltransferase)(-->); 136 ~~ { [Gene] STBHUCCB_23330(2270930, 2271121, -, hypothetical protein)(<--); 126 ~~~~ 65 } ~~ 282; STBHUCCB_23340(2271338, 2272246, +, Glutamate/aspartate periplasmic-binding protein)(-->)

[Identity]: avg = 100%, min = 100% [Motif score]: max = 18.0, min = 18.0, avg = 18.0, std = 0.0 [Avg Entropy]: pattern = 0.000, flank = 0.000 (0.000 left, 0.000 right)
[both consv] Pattern and its (one-side or both-side) surrounding sequences are identically aligned

See legend in Table 3.3.

Table 3.13 Cluster 1656 from the results of finding orthologous σ^{54} binding sites in 107 *Salmonella* and *E. coli* genomes.

Cluster 1656

#	Motif Score	Motif Score (opposite strand)	Sequence	Genome
1	**16.06**	10.88	atgattaagggccgccagcagcgaataccagtcgcccgaTGGCACGCTATATGCCAAcgcctgacacaatgtttcacgagcctgccagttatctcc	Escherichia_coli_O157_H7_EDL933_uid57831
2	**16.07**	10.88	atgattaagggccgccagcagcgaataccagtcgcccgaTGGCACGCTATATGCCAAcgcctgacacaatgtttcacgagcctgccagttatctcc	Escherichia_coli_O157_H7_uid57781
3	**16.06**	10.88	atgattaagggccgccagcagcgaataccagtcgcccgaTGGCACGCTATATGCCAAcgcctgacacaatgtttcacgagcctgccagttatctcc	Escherichia_coli_SMS_3_5_uid58919
4	**16.06**	10.88	atgattaagggccgccagcagcgaataccagtcgcccgaTGGCACGCTATATGCCAAcgcctgacacaatgtttcacgagcctgccagttatctcc	Escherichia_coli_O157_H7_EC4115_uid59091
5	**16.07**	10.88	atgattaagggccgccagcagcgaataccagtcgcccgaTGGCACGCTATATGCCAAcgcctgacacaatgtttcacgagcctgccagttatctcc	Escherichia_coli_O157_H7_TW14359_uid59235
…	…	…	… …	… …
40	**16.07**	10.88	atgattaagggccgccagcagcgaataccagtcgcccgaTGGCACGCTATATGCCAAcgcctgacacaatgtttcacgagcctgccagttatctcc	Escherichia_coli_LY180_uid219461
41	7.86	**13.29**	atgattaagggccgccagcagcgaataccagtcgcccgATTGCACGCTATATGCCAacgcctgacacaatgtttcacgagcctgccagttatctcc	Escherichia_coli_IAI39_uid59381
42	7.86	**13.27**	atgattaagggccgccagcagcgaataccagtcgcccgATTGCACGCTATATGCCAacgcctgacacaatgtttcacgagcctgccagttatctcc	Escherichia_coli_O7_K1_CE10_uid162115
43	8.39	**13.26**	atgattaagggccgccagcagcgaataccagtcgcccgATTGCACGCTATATGCCAgcgcctgacacaatgtttcacgagcctgccagttatctcc	Escherichia_coli_CFT073_uid57915
44	8.38	**13.27**	atgattaagggccgccagcagcgaataccagtcgcccgATTGCACGCTATATGCCAgcgcctgacacaatgtttcacgagcctgccagttatctcc	Escherichia_coli_ABU_83972_uid161975
45	8.39	**13.27**	atgattaagggccgccagcagcgaataccagtcgcccgATTGCACGCTATATGCCAgcgcctgacacaatgtttcacgagcctgccagttatctcc	Escherichia_coli__clone_D_i2__uid162047
…	…	…	… …	… …
62	8.38	**13.27**	atgattaagggccgctagcagcgaataccagtcgcccgATTGCACGCTATATGCCAgcgcctggcacaatgtttcacgagcctgccagttatctcc	Escherichia_coli_042_uid161985
63	-16.24	-5.90	atgattgagcgccgttaataccgcataccagtcgcccgtatgcaaactgtccattaacgcctgacacagcgcctcacgctgcgcccagcgttcaga	Salmonella_enterica_serovar_Typhimurium_LT2_uid57799
64	-16.24	-5.91	atgattgagcgccgttaataccgcataccagtcgcccgtatgcaaactgtccattaacgcctgacacagcgcctcacgctgcgcccagcgttcaga	Salmonella_enterica_serovar_Typhimurium_SL1344_uid86645
…	…	…	… …	… …
105	-19.48	-14.06	atgattgagcgccgttaataacgcataccagtcgcccgtatgtaaactgtccattaacgcctgacacagcgcctcacgctgcgcccagcgttcaga	Salmonella_enterica_serovar_Newport_USMARC_S3124_1
106	2.54	-3.13	atgattgagcgccgttaataacgcgtaccagtcgcccgtttgcacactgtctgttaacgcctggcacagtgcctcgcgctgcgcccatcgttcaga	Salmonella_bongori_NCTC_12419_uid70155
107	2.54	-3.12	atgattgagcgccgttaataacgcgtaccagtcgcccgtttgcacactgtctgttaacgcctggcacagtgcctcgcgctgcgcccatcgttcaga	Salmonella_bongori_Sbon_167_uid213088

Cluster 1656 (continued)

#	Motif Score	Motif Score (opposite strand)	Vicinity	Strand	Position	Orthologous motifs
1	**16.06**	10.88	[Gene] Z1738(1599943, 1600947, +, holB, 2.7.7.7, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1600600, 1600617]	TGGCACGCTATATGCCAA
2	**16.07**	10.88	[Gene] ECs1477(1514857, 1515861, +, 2.7.7.7, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1515514, 1515531]	TGGCACGCTATATGCCAA
3	**16.06**	10.88	[Gene] EcSMS35_2028(2049761, 2050765, -, holB, 2.7.7.7, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	+	[2050091, 2050108]	TGGCACGCTATATGCCAA
4	**16.06**	10.88	[Gene] ECH74115_1478(1458442, 1459446, +, holB, 2.7.7.7, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1459099, 1459116]	TGGCACGCTATATGCCAA
5	**16.07**	10.88	[Gene] ECSP_1400(1458730, 1459734, +, holB, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1459387, 1459404]	TGGCACGCTATATGCCAA
…	…	…	… …	…	…	…
40	**16.07**	10.88	[Gene] LY180_05700(1211489, 1212493, +, 2.7.7.7, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1212146, 1212163]	TGGCACGCTATATGCCAA
41	7.86	**13.29**	[Gene] ECIAI39_2062(2109930, 2110934, -, holB, 2.7.7.7, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	+	[2110260, 2110277]	ttgcacgctatatgccaa
42	7.86	**13.27**	[Gene] CE10_1179(1227616, 1228620, +, holB, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1228273, 1228290]	ttgcacgctatatgccaa
43	8.39	**13.26**	[Gene] c1371(1299263, 1300267, +, holB, 2.7.7.7, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1299920, 1299937]	ttgcacgctatatgccag
44	8.38	**13.27**	[Gene] ECABU_c13130(1322725, 1323729, +, holB, 2.7.7.7, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1323382, 1323399]	ttgcacgctatatgccag
45	8.39	**13.27**	[Gene] i02_1253(1261443, 1262447, +, holB, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1262100, 1262117]	ttgcacgctatatgccag
…	…	…	… …	…	…	…
62	8.38	**13.27**	[Gene] EC042_1169(1248184, 1249188, +, holB, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1248841, 1248858]	ttgcacgctatatgccag
63	-16.24	-5.90	[Gene] STM1201(1284262, 1285266, +, holB, 2.7.7.7, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1284919, 1284936]	atgcaaactgtccattaa
64	-16.24	-5.91	[Gene] SL1344_1138(1240404, 1241408, +, holB, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1241061, 1241078]	atgcaaactgtccattaa
…	…	…	… …	…	…	…
105	-19.48	-14.06	[Gene] SN31241_22690(2309812, 2310816, +, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[2310469, 2310486]	atgtaaactgtccattaa
106	2.54	-3.13	[Gene] SBG_1039(1147185, 1148189, +, holB, DNA polymerase III subunit delta)(<--); 338 ~~~~ 666	-	[1147842, 1147859]	ttgcacactgtctgttaa
107	2.54	-3.12	[Gene] A464_1135(1160617, 1161621, +, DNA polymerase III delta prime subunit)(<--); 338 ~~~~ 666	-	[1161274, 1161291]	ttgcacactgtctgttaa

[Identity]: avg = 82%, min = 61% [Motif score]: max = 16.1, min = -21.1, avg = 0.1, std = 15.7 [Avg Entropy]: pattern = 0.627, flank = 0.321 (0.240 left, 0.403 right) [Mann-Whitney U test]: p = 0.403, w = 305.0

See legend in Table 3.3.

Table 3.14 Cluster 2315 from the results of finding orthologous σ^{54} binding sites in 107 *Salmonella* and *E. coli* genomes.

Cluster 2315

#	Motif Score	Sequence	Genome
1	-30.45	caaaaaacggctgaattttgcgataacccccacattttctgcgatttagcgccaatctgaatcgttaacacgtgacatagtttcagatttggactat	Escherichia_coli_K_12_substr__MG1655_uid57779
2	-30.44	caaaaaacggctgaattttgcgataacccccacattttctgcgatttagcgccaatctgaatcgttaacacgtgacatagtttcagatttggactat	Escherichia_coli_P12b_uid162061
3	-30.45	caaaaaacggctgaattttgcgataacccccacattttctgcgatttagcgccaatctgaatcgttaacacgtgacatagtttcagatttggactat	Escherichia_coli_HS_uid58393
…	…	… …	… …
61	-30.46	caaaaaacggctgaattttgcgataaacccccacattttctgcgatttagcgccaatcttaatcgttaacacgtgacatagtttcagatttggactat	Escherichia_coli_O104_H4_2009EL_2071_uid176128
62	-30.45	caaaaaacggctgaattttgcgataaacccccacattttctgcgatttagcgccaatcttaatcgttaacacgtgacatagtttcagatttggactat	Escherichia_coli_55989_uid59383
63	10.25	caaaaaaacgcctctatttgcgataaacgccgcattttattggcattattgctgtctgttatcgttaatagttgacatagtttcggattaagactat	Salmonella_enterica_serovar_Typhi_CT18_uid57793
64	10.24	caaaaaaacgcctctatttgcgataaacgccgcattttattggcattattgctgtctgttatcgttaatagttgacatagtttcggattaagactat	Salmonella_enterica_serovar_Typhi_Ty2_uid57973
…	…	… …	… …
70	10.26	caaaaaaacgcctctatttgcgataagcaccacattttattggcattattgctgtctgttatcgttaatagttgacatagtttcggattaagactat	Salmonella_enterica_serovar_Javiana_CFSAN001992_uid190101
71	14.6	caaaaaaacgcctctatttgcgataagcaccacattttatTGGCATTATTGCTGCCTGttatcgttaatacttgacatagtttcggattaagactat	Salmonella_enterica_serovar_Heidelberg_SL476_uid58973
72	14.6	caaaaaaacgcctctatttgcgataagcaccacattttatTGGCATTATTGCTGCCTGttatcgttaatacttgacatagtttcggattaagactat	Salmonella_enterica_serovar_Heidelberg_B182_uid162195
73	14.6	caaaaaaacgcctctatttgcgataagcaccacattttatTGGCATTATTGCTGCCTGttatcgttaatacttgacatagtttcggattaagactat	Salmonella_enterica_serovar_Heidelberg_41578_uid212970
74	14.58	caaaaaaacgcctctatttgcgataagcaccacattttatTGGCATTATTGCTGCCTGttatcgttaatacttgacatagtttcggattaagactat	Salmonella_enterica_Serovar_Heidelberg_CFSAN002069
…	…	… …	… …
104	14.57	caaaaaa-cgcctctatttgcgataagcaccacattttatTGGCATTATTGCTGCCTGttatcgttaatagttgacatagtttcggattaagactat	Salmonella_enterica_serovar_Newport_USMARC_S3124_1
105	13.03	caaaaaaacgcgtctatttgcgataagcgacacattttatTGGCATGATTGCTGTTGGttatcgttaatcaatgacatagtttcggattaagactat	Salmonella_enterica_arizonae_serovar_62_z4_z23__uid58191
106	3.02	caaaaaa-tgtgtctgtttgtgataagtaccacatttaattggcattacttcctttagttatcgttaatgtgcgtcatagtttcggattaagattat	Salmonella_bongori_NCTC_12419_uid70155
107	3.01	caaaaaa-tgtgtctgtttgtgataagtaccacatttaattggcattacttcctttagttatcgttaatgtgcgtcatagtttcggattaagattat	Salmonella_bongori_Sbon_167_uid213088

Cluster 2315 (continued)

#	Motif Score	Vicinity
1	-30.45	[Intergenic] b3571(3735520, 3737550, +, malS, 3.2.1.1, alpha-amylase)(<--); 164 ~~~~ 156; b3570(3734376, 3735200, -, bax, hypothetical protein)(-->)
2	-30.44	[Intergenic] P12B_c3700(4039221, 4041251, +, malS, Alpha-amylase precursor)(<--); 164 ~~~~ 156; P12B_c3699(4038077, 4038901, -, hypothetical protein)(-->)
3	-30.45	[Intergenic] EcHS_A3774(3766800, 3768830, +, malS, 3.2.1.1, periplasmic alpha-amylase precursor)(<--); 164 ~~~~ 156; EcHS_A3773(3765656, 3766480, -, bax, hypothetical protein)(-->)
…	…	… …
61	-30.46	[Intergenic] O3O_24680(182400, 183653, -, avtA, 2.6.1.66, valine--pyruvate transaminase)(<--); 2371 ~~~~ 156; O3O_24665(186180, 187004, +, hypothetical protein)(-->)
62	-30.45	[Intergenic] EC55989_4028(4108073, 4110103, +, malS, 3.2.1.1, periplasmic alpha-amylase)(<--); 164 ~~~~ 156; EC55989_4026(4106929, 4107753, -, bax, hypothetical protein)(-->)
63	10.25	[Intergenic] (3996312, 3998339, -, STY4134, 3.2.1.1, alpha-amylase)(<--); 160 ~~~~ 156; (3998655, 3999479, +, STY4135, putative exported amidase)(-->)
64	10.24	[Intergenic] t3855(3981034, 3983061, -, malS, periplasmic alpha-amylase)(<--); 160 ~~~~ 156; t3856(3983377, 3984201, +, bax, hypothetical protein)(-->)
…	…	… …
70	10.26	[Intergenic] CFSAN001992_15325(3166045, 3168072, -, malS, alpha-amylase)(<--); 160 ~~~~ 156; CFSAN001992_15330(3168388, 3169212, +, hypothetical protein)(-->)
71	14.6	[Intergenic] SeHA_C3987(3863622, 3865649, +, malS, 3.2.1.1, periplasmic alpha-amylase)(<--); 160 ~~~~ 57; SeHA_C3986(3862482, 3863405, -, bax, hypothetical protein)(-->)
72	14.6	[Intergenic] SU5_04141(4521605, 4523632, +, 3.2.1.1, Periplasmic alpha-amylase)(<--); 160 ~~~~ 156; SU5_04140(4520465, 4521289, -, Glucosaminidase)(-->)
73	14.6	[Intergenic] SEEH1578_04405(787224, 789251, +, malS, alpha-amylase)(<--); 160 ~~~~ 156; SEEH1578_04400(786084, 786908, -, hypothetical protein)(-->)
74	14.58	[Gene] CFSAN002069_13720(2855075, 2857102, -, malS, alpha-amylase)(<--); 160 ~~ { [Gene] CFSAN002069_13725(2857102, 2857473, -, hypothetical protein)(<--); 160 ~~~~ 211 } ~~ 408; CFSAN002069_13730(2857670, 2858242, +, mannosyl-glycoprotein endo-beta-N-acetylglucosamidase)(-->)
…	…	… …
104	14.57	[Intergenic] SN31241_2190(225859, 227886, -, Alpha-amylase)(<--); 159 ~~~~ 408; SN31241_2200(228453, 229025, +, Protein bax)(-->)
105	13.03	[Intergenic] SARI_03972(3906867, 3907205, -, hypothetical protein)(<--); 157 ~~~~ 189; SARI_03973(3907551, 3908342, +, hypothetical protein)(-->)
106	3.02	[Intergenic] SBG_3258(3573569, 3575596, +, malS, 3.2.1.1, alpha-amylase)(<--); 159 ~~~~ 156; SBG_3257(3572430, 3573254, -, bax, exported amidase)(-->)
107	3.01	[Intergenic] A464_3745(3730291, 3732318, +, Periplasmic alpha-amylase)(<--); 159 ~~~~ 408; A464_3744(3729152, 3729724, -, BAX protein)(-->)

[Identity]: avg = 84%, min = 63% [Motif score]: max = 16.6, min = -30.5, avg = -12.0, std = 21.7 [Avg Entropy]: pattern = 0.711, flank = 0.273 (0.307 left, 0.239 right) [Mann-Whitney U test]: p = 0.017, w = 226.5

See legend in Table 3.3.

CHAPTER 4

DETECTING ORTHOLOGS FOR STRUCTURE-RELATED LOCAL SEQUENCE PATTERNS IN PROKARYOTIC GENOMES[1]

[1] Huang, Y. and J. Mrazek. To be submitted.

Introduction

In this chapter, we investigate the roles of structure-related sequence patterns in bacteria and archaea by computational techniques. We have adjusted the software developed in Chapter 3 for the assessment of evolutionary conservation of sequence-encoded structural elements, such as DNA intrinsic bends, palindromes and local direct repeats. Our program is intended as an exploratory and hypothesis-generating tool. The primary goal is to identify DNA structure-related sequence patterns that may be subject to selective constraints by comparing the conservation of the sequence matching the pattern with its immediate flanking sequences. The methodology and examples of its application are described below.

Methods

Step 1: Pattern Locator

First, the pattern loci in a collection of genomes are detected by Pattern Locator, a tool for finding local sequences patterns in long DNA sequences (Mrazek and Xie 2006). It performs an exhaustive search for all matches of the pattern and uses an intuitive syntax for pattern description, allowing combinations of specific sequences, direct and inverted repeats, and tandem repeats of sub-patterns. Here, we refer to locus that exactly matches the pattern as a query pattern locus, or a query locus.

Step 2: Identification and clustering of orthologous patterns

Similar to the method of clustering orthologous motifs in Chapter 3 (Method, Step 2), our program uses BLAST to find orthologs of each individual query pattern locus in all investigated genomes and subsequently combine groups of orthologous pattern loci. However, one technical

difficulty arises from the fact that many sequence patterns of interest can have variable length (unlike the regulatory motifs in Chapter 3 which have a fixed length) and the orthologous pattern loci can only partially overlap in the aligned sequences. For example, the intrinsic bend patterns can vary from about 40 to more than 100 bp in length. To accommodate the variable length of the sequence patterns, our program clusters groups of orthologs that contain only partially overlapping orthologous pattern loci as long as the length of the overlapping part satisfies a cutoff value (for example, half of the pattern length).

Step 3: Multiple sequence alignment

This process is the same as in Chapter 3 (Method, Step 3).

Step 4: Evaluation of conservation level

Shannon's information entropy (Shannon 1948) is used to evaluate the evolutionary conservation of each pattern location. For patterns that include gaps, the program treats the gaps as not being part of the pattern for the purpose of evaluating conservation of the pattern. For example, when investigating palindromes that can form stem-loop or cruciform structures, the loop sequences are not considered part of the pattern because substitutions in the loop regions do not affect the palindromic character of the whole sequence. Similarly, in intrinsic DNA bends, which consist of periodically spaced A-tracts, the spacers between the A-tracts are not considered part of the pattern.

Implementation

Input and Output

The program reads two input files. One is the pattern definition file using Pattern Locator syntax (Mrazek and Xie 2006). Another is the genome list file, where locations of genome genbank files are provided. In the output of each ortholog cluster, the orthologous pattern loci are listed in form of multiple sequence alignment. Additional information, such as adjacent genes, sequence identities and sequence entropy, are provided in the output file.

Limitations

This program is designed for the investigation of conservation of local sequence patterns, but not for long patterns with variable lengths (>100 bp). This is in part because Pattern Locator is not efficient in finding patterns with long, variable gaps but also because extensive gaps in the alignment can result in unreliable alignments and inaccurate assessments of sequence conservation.

Results and Discussion

Palindromes in Campylobacter genomes

As a pilot study, the conservation of palindromic patterns in 8 *Campylobacter* genomes (Table 4.1) was investigated. The palindromic pattern here is defined as a 9-nucleotide inverted repeat with no errors or a 12-nucleotide inverted repeat allowing 1 mismatch or a 15-nucleotide mirror repeat allowing 2 mismatches, etc. (one mismatch allowance added for every 3 nucleotide), with separation of up to 12 bp between the inverted repeats. This definition of palindromes is identical to the *pal9g12* patterns from Chapter 2 above. For example,

CTGGATCAGGCT… AGCCTCATCCAG is a *pal9g12* pattern (mismatched nucleotides are underscored). Palindromes can form stem-loop structure in RNA and thus may act as transcription terminators and riboswitches. They can also promote formation of cruciform structures in DNA, which can influence replication, regulation of gene expression, and recombination (Sinden 1994; Ussery, Soumpasis et al. 2002). Our previous study has shown that palindromic patterns are generally overrepresented in prokaryotic genomes (Huang and Mrazek 2014). Numbers of *pal9g12* patterns reported by the Pattern Locator in 8 *Campylobacter* genomes are listed in Table 4.1.

The values of parameters that are used in this study were selected based on a series of tests, which are generated by changing one parameter at a time and using default values for other parameters as listed in Table 3.1. For each test, we record the average number of orthologs per cluster, the total number of clusters and the number of questionable clusters (where orthologs have inconsistent vicinity genes information; see chapter 3) (Table 4.2). The total number of output clusters increases as BLAST e-value cutoff is changing from 1E-40 to 1E-10. However, there is little difference on this number when the e-value cutoff is ≥ 1E-20; thus we use 1E-20 for the value of BLAST e-value cutoff. Similarly, the results of the program are only slightly different when the relative BLAST score cutoff ranges from 0.6 to 0.8, or when the multiple sequence alignment flank length ranges from 30 to 60 (Table 4.2). In the tests for BLAST flank length, the results suggest that if the BLAST flank length is too long (> 400 bp), the total number of clusters and the average number of orthologs per cluster increase slowly but the number of questionable clusters increase more quickly; whereas a short BLAST flank length (<250 bp) results in losing too many clusters of orthologs without an adequate decrease in number of questionable clusters (Table 4.2). As for the BLAST score cutoff, we use the value 125 because

the corresponding result includes the highest number of orthologs per cluster among all tests while the number of questionable clusters and the total number of clusters are still acceptable.

Using the inputs and values of parameters described above, our program finds a total number of 10531 orthologous pattern loci in 3126 clusters. The distribution of number of orthologs in each cluster is shown in Figure 4.1. It shows that most clusters have only 2 or 3 orthologs and only less than 10% of clusters have orthologs in all 8 genomes. Orthologs of pattern loci might be absent due to deletions or genome rearrangements that result in a loss or gain of a locus in some of the genomes. Nevertheless, we investigated the reasons for absent orthologous loci in some of the clusters to verify that our method for detection of orthologs did not produce a large number of errors. For example, Cluster 1 (Table 4.4) has orthologs in 6 genomes, but the gene *acnB,* in which the pattern is located, is absent from *Campylobacter lari* or *Campylobacter curvus* (verified by protein BLAST, which is more sensitive to distant homologies than the nucleotide BLAST that is used in our program). In Cluster 105 (Table 4.5), which includes five genomes, the pattern locus is in the intergenic region between divergently oriented genes *cfrA* and *exbB*. Although the three absent genomes all have gene *exbB*, there is gene rearrangement upstream of *exbB*, which results in the absence of the corresponding pattern locus. In Cluster 1090 (Table 4.6; three genomes), the palindrome occupies most of the intergenic region between two transferase genes as well as part of both genes. However, these two genes are rearranged in *Campylobacter curvus, Campylobacter fetus*, and *Campylobacter* 03 427, and one of the two genes is missing in *Campylobacter hominis* and *Campylobacter concisus*. These and other examples suggest that the loci that our program did not find in some of the genomes are indeed absent and the program produced correct results.

One interesting result involves a cluster where the palindromic pattern is significantly variable (with $p<0.001$) in comparison to the flanking regions (Table 4.7). However, although the palindrome sequence is not conserved, the palindromic characteristic is still present in each ortholog, possibly indicating a functional significance of this palindromic pattern. This pattern is close to the end of gene *tuf*, which encodes translation elongation factor *Tu* (a highly expressed essential gene in most bacteria (Karlin and Mrazek 2000)); and the conserved palindrome may be a transcription terminator.

Intrinsic bends in Campylobacter genomes

DNA intrinsic bends may influence DNA-protein interactions in regulatory regions, facilitate DNA compaction in the nucleoid and cause a particular mode of supercoiling (Herzel, Weiss et al. 1998; Tolstorukov, Virnik et al. 2005; Kozobay-Avraham, Hosid et al. 2006). It's believed that DNA intrinsic bends are most often associated with periodically spaced A-tracts (Rozenberg, Rabinovich et al. 1998; Trifonov 1998). In our program, degree of such local DNA curvature is predicted by accumulating bends of A-tracts within a certain length of sequence (Tolstorukov, Virnik et al. 2005). The pattern used in this study is *bend40w60* (Table 2.1), referring to predicted bends of $\geq 45^{o}$ within a segment of ≤ 60 bp. Our study in Chapter 2 shows that *bend40w60* are generally slightly over-represented in prokaryotic genomes (Huang and Mrazek 2014).

The selection of parameters is the same as in the study of *pal9g12* (Table 4.3). There are 2204 orthologs of intrinsic bends in 597 clusters in 8 *Campylobacter* genomes. The average number of orthologs in each cluster is about 2 to 3 (Figure 4.2), which is affected by rearrangement of gene order and mutation of genes, as explained in the previous section. For example, in Cluster 40 of the output (Table 4.8), the pattern is located near the end of gene *murC*.

However, this pattern is absent in the unlisted 5 genomes because the second half of gene *murC* has been largely mutated. It is interesting to find out that the intrinsic bend inside gene *murC* is shifted in one *Campylobacter* genome. We speculate that this bend may play some physiological role which results in selective constraint that maintains the bend but a loss of some A-tracts on one side of the bend may be compensated by emergence of new A-tracts on the other side, as long as they are properly positioned to maintain the ~10.5 bp periodicity. This mechanism could possibly lead to "traveling bends", that is, bends that gradually shift their position by losing A-tracts on one side while gaining new A-tracts on the other.

Supplementary data:

Additional files are available at http://www.cmbl.uga.edu/downloads/data_sets/2015/Huang_dissertation/.

Additional_File_4_Ch4-pals9g12_Campylobacter_8.xlsx

Additional file 4 is the output of finding orthologous pattern loci of *pals9g12* in 8 *Campylobacter* genomes.

Additional_File_5_Ch4-bend45w60_Campylobacter_8.xlsx

Additional file 5 is the output of finding orthologous pattern loci of *bend45w60* in 8 *Campylobacter* genomes.

Reference

Herzel, H., O. Weiss, et al. (1998). "Sequence periodicity in complete genomes of Archaea suggests positive supercoiling." Journal of Biomolecular Structure & Dynamics **16**(2): 341-345.

Huang, Y. and J. Mrazek (2014). "Assessing diversity of DNA structure-related sequence features in prokaryotic genomes." DNA Res **21**(3): 285-297.

Karlin, S. and J. Mrazek (2000). "Predicted highly expressed genes of diverse prokaryotic genomes." Journal of Bacteriology **182**(18): 5238-5250.

Kozobay-Avraham, L., S. Hosid, et al. (2006). "Involvement of DNA curvature in intergenic regions of prokaryotes." Nucleic Acids Res **34**(8): 2316-2327.

Mrazek, J. and S. H. Xie (2006). "Pattern locator: a new tool for finding local sequence patterns in genomic DNA sequences." Bioinformatics **22**(24): 3099-3100.

Rozenberg, H., D. Rabinovich, et al. (1998). "Structural code for DNA recognition revealed in crystal structures of papillomavirus E2-DNA targets." Proc Natl Acad Sci U S A **95**(26): 15194-15199.

Shannon, C. E. (1948). "A Mathematical Theory of Communication." The Bell System Technical Journal **27**: 379-423.

Sinden, R. R. (1994). DNA Structure and Function, Academic Press.

Tolstorukov, M. Y., K. M. Virnik, et al. (2005). "A-tract clusters may facilitate DNA packaging in bacterial nucleoid." Nucleic Acids Res **33**(12): 3907-3918.

Trifonov, E. N. (1998). "3-, 10.5-, 200- and 400-base periodicities in genome sequences." Physica A **249**(1-4): 511-516.

Ussery, D., D. M. Soumpasis, et al. (2002). "Bias of purine stretches in sequenced chromosomes." Comput Chem **26**(5): 531-541.

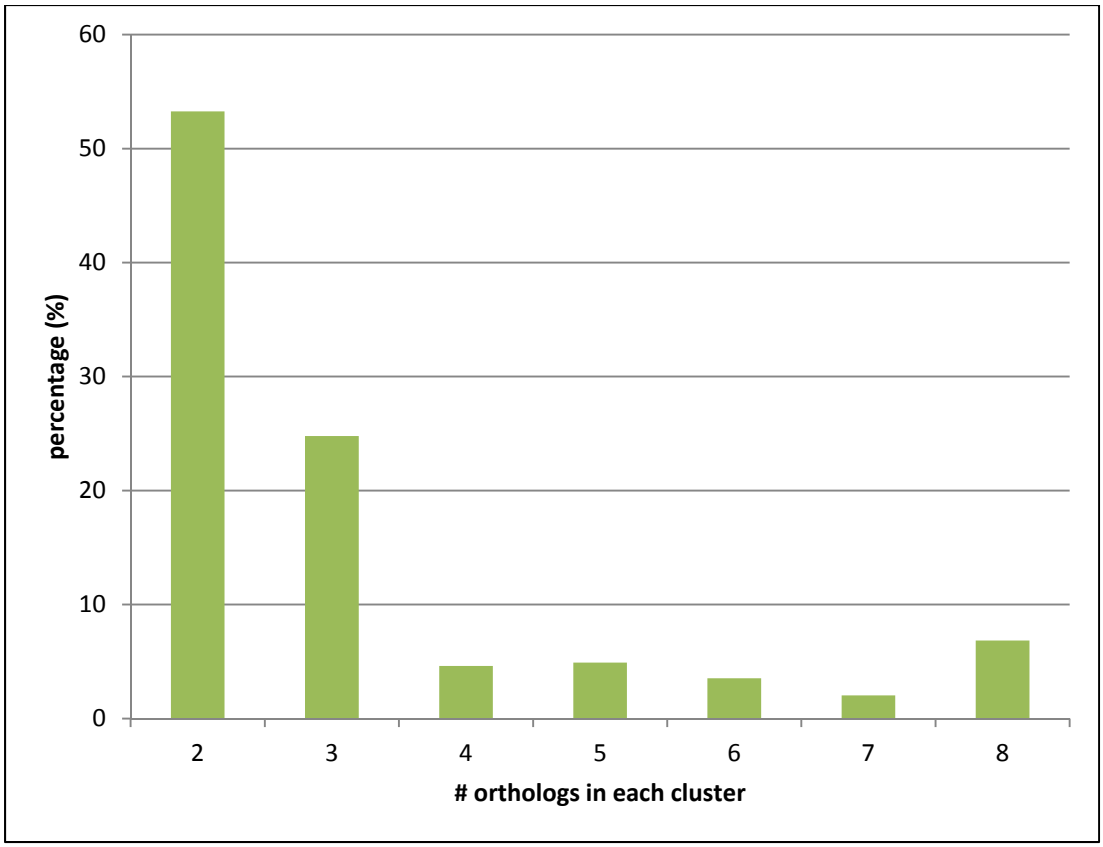

Figure 4.1 Distribution of number of orthologs in each cluster, for the results of finding orthologous pattern loci of *pal9g12* in 8 *Campylobacter* genomes (Table 4.1).

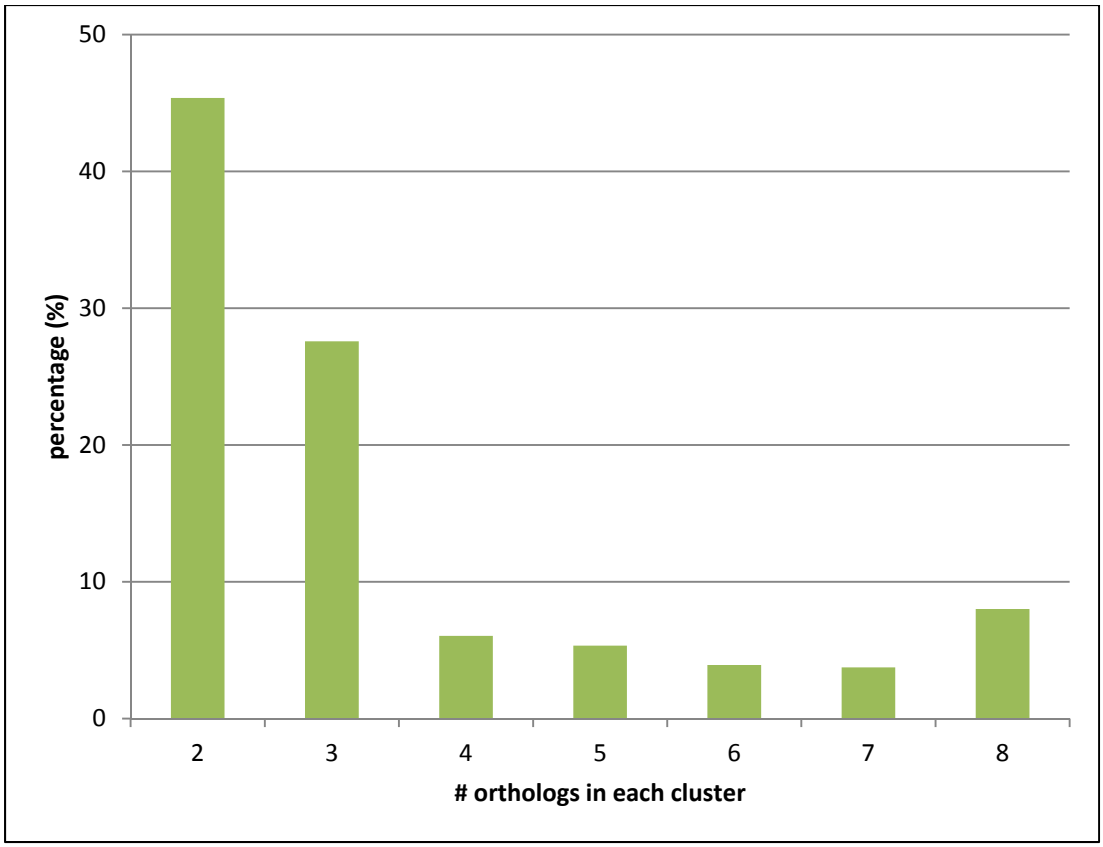

Figure 4.2 Distribution of number of orthologs in each cluster, for the results of finding orthologous pattern loci of *bend45w60* in 8 *Campylobacter* genomes (Table 4.1).

Table 4.1 List of 8 Campylobacter genomes being studied

Genome	**Size (Mbp)**	**# query pattern loci [a] (pal9g12)**	**# query pattern loci [a] (bend45w60)**
Campylobacter 03 427	1.78	706	127
Campylobacter coli 15 537360	1.66	801	119
Campylobacter concisus 13826	2.05	641	83
Campylobacter curvus 525 92	1.97	465	48
Campylobacter fetus 82 40	1.77	723	130
Campylobacter hominis ATCC BAA	1.71	1187	224
Campylobacter jejuni 32488	1.70	931	121
Campylobacter lari RM2100	1.53	882	118

[a] number of pattern loci in the genome that exactly match the pattern and have been reported by Pattern Locator

Table 4.2 Summary of parameter testing results (*pal9g12*).

Parameter setting[a]	# orthologs per cluster (average)	# clusters	# questionable clusters[b]
BLAST evalue cutoff			
1E-10	3.2	2998	14
1E-15	3.1	2998	14
1E-20	**3.1**	**2993**	**14**
1E-25	3.1	2985	13
1E-30	3.1	2975	13
1E-40	3.0	2870	11
BLAST flank length			
150	2.6	2523	8
200	2.9	2772	10
250	3.0	2902	13
300	3.1	2993	14
350	3.2	3044	14
400	**3.3**	**3071**	**14**
450	3.3	3094	15
500	3.4	3115	17
BLAST score cutoff			
100	3.0	3128	16
125	**3.3**	**3067**	**14**
150	3.1	2993	14
175	3.0	2905	12
200	2.9	2825	12
BLAST score cutoff (relative to best hit)			
0.6	3.1	2992	14
0.7	3.1	2994	14
0.8	**3.1**	**2993**	**14**
0.85	3.1	2990	14
0.9	3.1	2989	14
Multiple sequence alignment flank length			
20	3.1	2957	13
30	3.1	2987	14
40	**3.1**	**2993**	**14**
50	3.2	2996	13
60	3.2	2995	14

[a] Values of parameters in **bold** type are used in the pilot study**.**

[b] A cluster is flagged as questionable cluster if it contains any predicted non-ortholog motif. See Method in Chapter 3.

Table 4.3 Summary of parameter testing results (*bend45w60*).

Parameter setting	# orthologs per cluster (average)	# clusters	# questionable clusters
BLAST evalue cutoff			
1E-10	3.4	564	1
1E-15	3.4	564	1
1E-20	**3.4**	**563**	**1**
1E-25	3.4	560	1
1E-30	3.3	558	1
1E-40	3.2	537	1
BLAST flank length			
150	2.8	478	0
200	3.1	514	1
250	3.2	539	1
300	3.4	563	1
350	3.5	574	1
400	**3.6**	**581**	**1**
450	3.6	584	2
500	3.7	589	2
BLAST score cutoff			
100	**4.0**	**591**	**2**
125	3.5	577	1
150	3.4	563	1
175	3.2	541	1
200	3.1	522	1
BLAST score cutoff (relative to best hit)			
0.6	3.4	564	1
0.7	3.4	564	1
0.8	**3.4**	**563**	**1**
0.85	3.4	564	1
0.9	3.4	564	1
Multiple sequence alignment flank length			
20	3.3	560	1
30	3.4	563	1
40	**3.4**	**563**	**1**
50	3.4	562	1
60	3.5	566	1

See legend in Table 4.2.

Table 4.4 A cluster of orthologous palindromic sequence pattern (*pal9g12*) in 8 *Campylobacter* genomes (Example 1).

Cluster 1

#	Sequence	Strain
1	tttattaat-aacatt----ctagcttgaagttcggtactagatTGAAATTTAAATATTTATAAAT-TTCAttttctttttc--cagcaagttttgccggaactatcttcata	*fetus* 82
2	tttattagt-aacact----ctagtttgaagtttggtataagatTGAAATTTAGATATTTATAAAT-TTCAtcttgttttc--cagcaagttttgcaggaactatcttcata	03 427
3	aatttgaattaatttt----caagtttgaaattttcaatctcattaaaatttAAATATCTATAGATATTTgccttatgctcatcacttagtttttcgctgacaatttgtaag	*jejuni* 32488
4	aatttggattaattat----ctagtttaaaattatcaatttcattaaaattaagatatttataaatatttgctttgtgcgcatcgcttaatttatcgcttacgatttgcaaa	*coli* 15
5	aagtttaaaatttttgattatagcgtaaaattcttaacttcattaaaatttaaatatttgtaaat-ttcactctcttttc--ctgctaatttttttaggtacaattttcaga	*hominis* ATTC
6	tattgtaacaaattt-----atatactaaactcgcttatttcattaaaatttaggtatttataaat-ttgatctttgttta--aactaaggctatctcttactatctttttta	*Concisus* 13826

Cluster 1 (continued)

#	Vicinity
1	[Gene] CFF8240_0997(1001644, 1004202, -, acnB, 4.2.1.3, bifunctional aconitate hydratase 2/2-methylisocitrate dehydratase)(←); 47 ~~~~ 2511
2	[Gene] CFT03427_0977(1003776, 1006334, -, acnB, 4.2.1.3, aconitate hydratase 2)(←); 47 ~~~~ 2511
3	[Gene] M635_08460(1638216, 1640762, -, 4.2.1.3, bifunctional aconitate hydratase 2/2-methylisocitrate dehydratase)(←); 48 ~~~~ 2498
4	[Gene] N149_0775(780336, 782882, -, acnB, 4.2.1.99, Aconitate hydratase 2 / 2-methylisocitrate dehydratase)(←); 45 ~~~~ 2501
5	[Gene] CHAB381_1421(1356465, 1359017, -, acnB, 4.2.1.3, bifunctional aconitate hydratase 2/2-methylisocitrate dehydratase)(←); 38 ~~~~ 2514
6	[Gene] CCC13826_1408(1128058, 1130643, -, acnB, 4.2.1.3, bifunctional aconitate hydratase 2/2-methylisocitrate dehydratase)(←); 38 ~~~~ 2547

[Identity]: avg = 63%, min = 56% [Avg Entropy]: pattern = 0.290, flank = 0.793 (0.806 left, 0.779 right) [Mann-Whitney U test]: p = 2.13e-04, w = 773.5

This table was extracted from the results of finding orthologous patterns of *pal9g12* in 8 different *Campylobacter* species. See Table 2.1 for description of pattern *pal9g12* and Table 3.3 for legend of the table.

Table 4.5 A cluster of orthologous palindromic sequence pattern (*pal9g12*) in 8 *Campylobacter* genomes (Example 2).

Cluster 105

#	Sequence	Strain
1	a-tatttttgtaattcgagcattataataaaaacaaactaaatttataattaatatca-ttatattatatataggcataattttgattttatt-ttatatagtaa	*fetus* 82
2	a-tatttttgtgatttgagcattataatgaaaacaaactaaatttataattaatatca-ttatattatatataggtataatttcgattttatt-tt-tatagtta	03 427
3	tatacttataaattttaaaagattatactgtataaaaaataaatttataattaatatcaattatcataatttttaagaaattcttgtttata---gtatttatttt	*jejuni* 32488
4	t----tgattaaatttttggcgattatactacaataaaaataaaATTATAATTAATATTAATTATCATaatttttttaaataatcatactttcacaagcctttggctt	*coli* 15
5	---atcttaaaatttttttgtgattatattgc-taaaaattaaatttttttattaataataattatcataattgatacaataaattcattttaa---attaataatta	*lari* RM2100

Cluster 105 (continued)

#	Vicinity
1	[Intergenic] CFF8240_1649(1620353, 1620790, +, exbB, TonB-system energizer ExbB)(←); 69 ~~~~ 148; CFF8240_1648(1618046, 1620136, -, ferric receptor CfrA)(→)
2	[Intergenic] CFT03427_1601(1625868, 1626305, +, exbB3, TonB system transport protein ExbB)(←); 69 ~~~~ 145; CFT03427_1600(1623564, 1625654, -, cfrA, ferric enterobactin uptake receptor)(→)
3	[Intergenic] M635_03830(753644, 754069, +, biopolymer transporter ExbB)(←); 70 ~~~~ 88; M635_03825(752743, 753486, -, hypothetical protein)(→)
4	[Intergenic] N149_0125(141631, 142056, -, exbB, Ferric siderophore transport system, biopolymer transport protein ExbB)(←); 66 ~~~~ 101; N149_0126(142223, 142960, +, Hypothetical protein)(→)
5	[Intergenic] Cla_0468(444415, 444846, -, exbB2, TonB system transport protein ExbB)(←); 60 ~~~~ 94; Cla_0469(+, iron-regulated outer membrane virulence protein, TonB receptor CfrA)(→)

[Identity]: avg = 63%, min = 54% [Avg Entropy]: pattern = 0.252, flank = 0.748 (0.675 left, 0.831 right) [Mann-Whitney U test]: p = 5.07e-03, w = 658.0

See Table 2.1 for description of pattern *pal9g12* and Table 3.3 for legend of the table.

Table 4.6 A cluster of orthologous palindromic sequence pattern (*pal9g12*) in 8 *Campylobacter* genomes (Example 3).

Cluster 1090

#	Sequence	Strain
1	tgaaatgagtgaaaagttaagaaaggctctaaagagtttttAATAAATTTTGTAA----AAAAATTTATTctttttccatctactataggcacaaataaacattcttctaaa	*jejuni* 32488
2	tgaaatgagcgagaagctaagaaaggtttttaaaagtttTGAATAAATTTTTTATTGATAAAAATTTATTCTtttccatctactatgggtacaaaaagacattcttctaga	*coli* 15
3	tatcccctataagaagcttat----gtttaaaaataattatattaaaattaaaattagtagga--ttattctttaccgtctttgataggcacaaaaagacattcatctaaa	*lari* RM2100

Cluster 1090 (continued)

#	Vicinity
1	[Intergenic] M635_05425(1050979, 1052160, +, acetylornithine aminotransferase)(→); 11 ~~~~ 8; M635_05430(1052179, 1052808, -, protein-L-isoaspartate O-methyltransferase)(←)
2	[Intergenic] N149_0223(241135, 242319, +, argD, 2.6.1.17, Acetylornithine aminotransferase / N-succinyl-L,L-diaminopimelate aminotransferase)(→); 10 ~~~~ 10; N149_0224(242339, 242968, -, pcm, 2.1.1.77, Protein-L-isoaspartate O-methyltransferase)(←)
3	[Intergenic] Cla_1514(1461952, 1462692, +, carbonic anhydrase)(←); 82 ~~~~ 9; Cla_1513(1461229, 1461861, +, pcm, 2.1.1.77, protein-L-isoaspartate O-methyltransferase)(←)

[Identity]: avg = 67%, min = 58% [Avg Entropy]: pattern = 0.506, flank = 0.436 (0.612 left, 0.259 right) [Mann-Whitney U test]: p = 6.14e-01, w = 1218.0

See Table 2.1 for description of pattern *pal9g12* and Table 3.3 for legend of the table.

Table 4.7 A cluster of unconserved palindromic sequence pattern (*pal9g12*) in 8 *Campylobacter* genomes.

Cluster 3121

#	Sequence	Strain
1	actttaaaccaactttaattctcatttcaatttttccttaaaaaatT-AAAGCCAAGCTTTTAAC----------TTGGCTTTATTAAATTATTTAATAAttttagaaacaacacctgaaccaacagttttaccaccttt	*jejuni* 32488
2	attttaaaccaactttaattctcatttatattttccttaaaaaaTTTAAAGCCAAGTATTAAAC----------TTGGCTTT-TAAaattacttaataatttttagaaacaacacctgaaccaacagttttaccaccttt	*coli* 15
3	attttaaaccaactttaattctcatttttatt--ccttaaaAAG--------CGAG-ATCAAAT----------CTCGCTT---ataattatttaataatttttagaaacaacacctgatccaacagtacggccaccttt	*lari* RM2100
4	attttaaaccaactttgattctcattaaatttccttaagGCAAAAGAGCAAAATCGCTCTTTTG----------Ctaaatttta--aattatgcaagtatcttagaaacgacacctgaaccaacagttctgccacccct	*curvus* 525
5	attttaaaccgattttaattctcatatatattccttatagaaaGCCAAAGGGTTTCCCCTTTGG----------Cataattttataaattaaccaagtatttttgaaacaacacctgaaccaacagttctaccaccttt	*concisus* 138
6	acttcaaacctattttaactct-atttgcacttgccatt-taatttccctTTTATATTAAATGCGAGCCAAAG-CTCGCAATTAA--AATTAAgctattatcttagaaacaacacctgaaccaacagttctaccaccct	*fetus* 82
7	acttcaaacctattttaactct-atttgcacttgccatt-taatttccctTTTATATTAAATGCGAGCCAAAG-CTCGCAATTAA--AATTAAgctattatcttagaaacaacacctgaaccaacagttctaccaccct	03 427
8	attttaaaccaattttaactct-atttgcatttttgccataatttttccttaaaattttTTGGGGGAAAATTTCCCCCAAtttataaattaacctaaaattttagaaacaacacctgaaccgactgtgtgtccaccttt	*hominis* 381

Cluster 3121 (continued)

#	Vicinity
1	[Intergenic] M635_06710(1281597, 1281755, +, 50S ribosomal protein L33)(←); 42 ~~~~ 11; M635_06705(1280345, 1281544, +, tuf, 3.6.5.3, elongation factor Tu)(←)
2	[Intergenic] N149_0472(472561, 472719, +, rpmG, LSU ribosomal protein L33p)(←); 35 ~~~~ 18; N149_0471(471309, 472508, +, tuf, Translation elongation factor Tu)(←)
3	[Intergenic] Cla_0441(414076, 414234, +, rpmG, 50S ribosomal protein L33)(←); 25 ~~~~ 15; Cla_0440(412837, 414036, +, tuf, elongation factor Tu)(←)
4	[Intergenic] CCV52592_2190(1399992, 1400069, -, tRNA-Trp)(←); 200 ~~~~ 25; CCV52592_0173(1400294, 1401493, -, tuf, elongation factor Tu)(←)
5	[Intergenic] CCC13826_2234(657732, 657809, +, tRNA-Trp)(←); 206 ~~~~ 25; CCC13826_0166(656302, 657501, +, tuf, elongation factor Tu)(←)
6	[Intergenic] CFF8240_1324(1313847, 1314014, -, rpmG, 50S ribosomal protein L33)(←); 31 ~~~~ 17; CFF8240_1325(1314062, 1315261, -, tuf, 3.6.5.3, elongation factor Tu)(←)
7	[Intergenic] CFT03427_1289(1320222, 1320389, -, rpmG, 50S ribosomal protein L33)(←); 31 ~~~~ 17; CFT03427_1290(1320437, 1321636, -, tuf, translation elongation factor Tu)(←)
8	[Intergenic] CHAB381_1671(1593964, 1594134, -, rpmG, 50S ribosomal protein L33)(←); 30 ~~~~ 19; CHAB381_1672(1594183, 1595382, -, tuf, elongation factor Tu)(←)

[Identity]: avg = 65%, min = 51% [Avg Entropy]: pattern = 1.112, flank = 0.383 (0.545 left, 0.224 right) [Mann-Whitney U test]: p = 1.83e-04, w = 490.5

See Table 2.1 for description of pattern *pal9g12* and Table 3.3 for legend of the table.

Table 4.8 A cluster of intrinsic bend orthologs (*bend45w60*) in the test of 8 *Campylobacter* genomes.

Cluster 40

#	Sequence
1	tgtttttagtgcgggcgaggcttcaaataatatagttttgaaagatgaatttAAA------AAggcgATTTTTgccgAAAAAgttgaacgAAAAgatAATTctatagagttttttgatagttttggagtaaagcataggt
2	tgtttttagtgcgggagaagcttcaaataacatagttttgaaagatgaatttAAA------AAAgcgATTTTTgccgAAAAAgttgaacgAAAAgatAATTctatagagttttttgatagttttggagtaaagcataggc
3	tgtttttgcagccggagaagcgcaaaacggaatagatgtAAAAgacgAATTTaaaggtAAAAATatAATTTTTactcAAAAAgtaaaacgaaacggcgaagcgattgaatttaatgatgaattcggcgtaaaacaccgtg

Cluster 40 (continued)

#	Strain	Vicinity
1	fetus 82	[Gene] CFF8240_0616(625085, 626383, +, murC, 6.3.2.8, UDP-N-acetylmuramate--L-alanine ligase)(→); 1170 ~~~~ 128
2	03 427	[Gene] CFT03427_0619(627545, 628843, +, murC, 6.3.2.8, UDP-N-acetylmuramate-alanine ligase)(→); 1170 ~~~~ 128
3	hominis 381	[Gene] CHAB381_1690(1609531, 1610841, -, murC, 6.3.2.8, UDP-N-acetylmuramate--L-alanine ligase)(→); 1163 ~~~~ 147

[Identity]: avg = 74%, min = 62% [Avg Entropy]: pattern = 0.203, flank = 0.384 (0.370 left, 0.437 inside, 0.370 right) [Mann-Whitney U test]: p = 9.84e-02, w = 804.5

This table was extracted from the results of finding orthologous patterns of *bend45w60* in 8 different *Campylobacter* genera. See Table 2.1 for description of pattern *pal9g12* and Table 3.3 for legend of the table.

CHAPTER 5

CONCLUSION

As described in chapter 2, we conducted a survey of DNA structure-related local sequence patterns in more than 1500 complete microbial genomes and interpreted the results in terms of their relationship with organisms' properties (Huang and Mrazek 2014). Our results show that simple sequence repeats and Z-DNA-promoting patterns are generally suppressed in prokaryotic genomes, whereas palindromes and inverted repeats are over-represented (Table 2.2). Additionally, DNA repeats and palindromes exhibit negative trend with increasing optimal growth temperature (Figure 2.1). Patterns that promote intrinsic DNA curvature increase with the increasing OGT in protein-coding regions, but decrease in the non-coding regions (Tables 2.S4 and 2.S5). Patterns that promote Z-DNA are more suppressed in facultative microbes than in aerobes and anaerobes (Table 2.S6).

All these results indicate that many of the sequence patterns of interest (Table 2.1) are not randomly distributed in the genomes and their usage relates to the organism's habitat, especially with respect to optimal growth temperature. The nonrandom occurrence of these patterns suggests that they could have some biological role and in many instances there is other evidence that this is the case (Sinden 1994; Mrazek, Guo et al. 2007; Bohlin, Hardy et al. 2009). The research conducted in chapters 3 and 4 are the natural next steps where we aim to identify specific occurrences of such sequence patterns that are under selective constraints, which would be indicative of a physiological role of such patterns.

In chapter 3, we developed a new software for analysis of evolution of short sequence motifs in prokaryotic genomes, with a particular focus on transcription factor binding sites. First, regulatory motif sites are detected in all analyzed genomes using a standard PSSM technique for supervised motif finding (Mrazek 2009). Orthologous regulatory sites are subsequently identified by BLAST search (Altschul, Gish et al. 1990) and a clustering algorithm that groups together orthologous sites in multiple genomes. To assess potential selective constraints affecting the regulatory motif sites, the level of motif conservation is then evaluated by comparison of information entropy within each motif occurrence and its immediate flanking sequences in the multiple sequence alignment of the clustered orthologs.

As a pilot study, the new tool was used to investigate σ^{54} binding site motifs in *Salmonella* and *E. coli* genomes. Our first goal was to evaluate the accuracy of the method by verifying that it produced some expected results. In this regard, the binding sites upstream of well-characterized σ^{54}-dependent genes were highly conserved in all analyzed genomes (Table 3.7). The association between motif score and motif conservation ranking implies that high-scoring motifs are more likely to be conserved, which is also expected (Figures 3.4 – 3.6). While some high scoring motifs are significantly conserved (Tables 3.8 – 3.12), we also found some strongly variable σ^{54} binding sites, which might suggest possible positive selection and change of the regulatory network (Tables 3.13 and 3.14). The analysis of evolution of σ^{54} binding sites helps us gain new insights into the evolution of σ^{54} regulon.

The software we developed in chapter 3 has been demonstrated to be capable of detecting orthologous sites of regulatory motifs in a collection of closely related genomes. It is able to identify motifs that are possibly under selective constraints which influence their evolutionary conservation. Our methodology for investigation of evolution of regulatory motifs is an

important step towards understanding the evolution of regulatory networks and how organisms adapt to changing conditions or environments.

In chapter 4, we adapted the software developed in chapter 3 for the assessment of evolutionary conservation of sequence-encoded structural elements, such as DNA intrinsic bends, palindromes, and local direct repeats. As an exploratory and hypothesis-generating tool, the software identifies DNA structure-related sequence patterns that may be subject to selective constraints by comparing the conservation of the sequence matching the pattern with its immediate flanking sequences. In the pilot study of accessing conservation of palindromic patterns and intrinsic bends in 8 *Campylobacter* genomes, our program was demonstrated to be able to find orthologous pattern loci in genomes under investigation and identify possible selective constraints.

References

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Bohlin, J., S. P. Hardy, et al. (2009). "Stretches of alternating pyrimidine/purines and purines are respectively linked with pathogenicity and growth temperature in prokaryotes." Bmc Genomics **10**.

Huang, Y. and J. Mrazek (2014). "Assessing diversity of DNA structure-related sequence features in prokaryotic genomes." DNA Res **21**(3): 285-297.

Mrazek, J. (2009). "Finding sequence motifs in prokaryotic genomes-a brief practical guide for a microbiologist." Briefings in Bioinformatics **10**(5): 525-536.

Mrazek, J., X. X. Guo, et al. (2007). "Simple sequence repeats in prokaryotic genomes." Proc Natl Acad Sci U S A **104**(20): 8472-8477.

Sinden, R. R. (1994). DNA Structure and Function, Academic Press.