Title	Comparative genomics of bacterial sequence elements associated with chromosome structure
Sub Title	バクテリアゲノムにおける染色体構造関連配列の比較解析
Author	河野, 暢明(Kono, Nobuaki) 冨田, 勝(Tomita, Masaru)
Publisher	慶應義塾大学湘南藤沢学会
Publication year	2012-04
Jtitle	博士論文
JaLC DOI	
Abstract	The structural characteristics and properties that have been observed in various circular bacterial chromosomes have become increasingly clear based on the decipherment of the complete genome sequences of approximately three thousand bacteria. A consideration of chromosome structure includes various symmetries at many scales, ranging from base composition to the gene strand bias, which are based on the replication origin and a terminal axis, and the supercoiling involved in the nucleoid structure. In this study, I focused on the chromosome structure and attempted to shed light on the mechanisms related to the formation and preservation of the chromosomes. First, I performed a comprehensive prediction of replication termination sites based on sequence, and I determined that the sequence is widely conserved in almost all bacteria. Based on this result, I proposed a model of replication termination from the standpoint of bioinformatics. The use of this newly defined symmetric structure made it possible to investigate how strongly symmetric structures have been maintained, and I analyzed the relationship between symmetric structures and insertion events. Furthermore, I found that the association of nucleoid binding sites with protein was controlled by cod on usage in the host bacterial genomes. Finally, I designed and implemented a versatile browser to facilitate the visualization of symmetric structures that cannot be fully captured by purely numeric information, allowing an intuitive grasp of these structures. Throughout this dissertation, I performed comprehensive analyses of the genomic functionality, characteristics, and structural evolution of bacterial chromosome, which serves as a medium of living systems. Based on these analyses, I discuss the means by which chromosome composition and structure have adapted to biological mechanisms in the course of evolution.
Notes	先端生命科学プロジェクト 冨田勝研究会2011年度
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=0302-0000-0662

ISBN 978-4-87762-257-2
SFC-DT 2012-001

博士論文

Comparative Genomics of Bacterial Sequence Elements Associated with Chromosome Structure

2011年度

河野 暢明 政策・メディア研究科 博士課程

先端生命科学プロジェクト

慶應義塾大学湘南藤沢学会

Keio SFC Academic Society

推薦のことば

ゲノムDNAは生命の設計図であると同時に、遺伝情報を次世代へ伝える媒体としての大きな役割を担っている。しかしながら進化の過程においてこの媒体を維持するための物理構造的な制約については未だよく解明されていない。本論文は、生体内機構によって構造的な制約を受けて来た遺伝情報の媒体であるゲノム構造を、計算機的なアプローチによって解析した、極めて独創的で優秀な博士論文である。

慶應義塾大学
環境情報学部教授
冨田　勝

Comparative Genomics of Bacterial Sequence Elements Associated with Chromosome Structure

Dissertation by
Nobuaki Kono

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
in the Graduate School of Media and Governance

Keio University
Fujisawa, Japan

February 2012

Keio University

To my family.

Abstract

The structural characteristics and properties that have been observed in various circular bacterial chromosomes have become increasingly clear based on the decipherment of the complete genome sequences of approximately three thousand bacteria. A consideration of chromosome structure includes various symmetries at many scales, ranging from base composition to the gene strand bias, which are based on the replication origin and a terminal axis, and the supercoiling involved in the nucleoid structure. In this study, I focused on the chromosome structure and attempted to shed light on the mechanisms related to the formation and preservation of the chromosomes. First, I performed a comprehensive prediction of replication termination sites based on sequence, and I determined that the sequence is widely conserved in almost all bacteria. Based on this result, I proposed a model of replication termination from the standpoint of bioinformatics. The use of this newly defined symmetric structure made it possible to investigate how strongly symmetric structures have been maintained, and I analyzed the relationship between symmetric structures and insertion events. Furthermore, I found that the association of nucleoid binding sites with protein was controlled by codon usage in the host bacterial genomes. Finally, I designed and implemented a versatile browser to facilitate the visualization of symmetric structures that cannot be fully captured by purely numeric information, allowing an intuitive grasp of these structures. Throughout this dissertation, I performed comprehensive analyses of the genomic functionality, characteristics, and structural evolution of bacterial chromosome, which serves as a medium of living systems. Based on these analyses, I discuss the means by which chromosome composition and structure have adapted to biological mechanisms in the course of evolution.

Keywords: Bacterial circular chromosome, Genome replication, Genomic/chromosomal structure, Nucleoid, Bioinformatics, Comparative genomics

論文題目

「バクテリアゲノムにおける染色体構造関連配列の比較解析」

論文要旨

ハイスループットなゲノム解読技術の発展により約3,000にも及ぶバクテリアのコンプリートゲノムが解読されてきた近年，多種多様なバクテリアの環状染色体に見られる構造の特徴とその性質が明らかになってきた．染色体上にある構造には，複製開始・終結点を軸とする塩基組成から遺伝子座に至るスケールで観察される対称構造や，核様体に関わる超螺旋形成などが知られている．そこで本研究ではバクテリアの持つ環状染色体において，これら染色体構造に着目し，その誕生から維持に関するメカニズムの解明を行うことを目的とした．まず対称構造の起点となる複製終結関連因子の網羅的な予測を行い，ほぼすべてのバクテリアで共通して保存されていることを解明した．そしてこの結果を元に，計算機的なアプローチから複製終結機構モデルの提唱を行い，対称構造の中心となる複製開始・終結点の定義を改めて行った．この様に定義されたゲノムの対称構造を用いることで，対称構造と水平伝播との関係性の解析が可能になったため，バクテリアが如何にこれまで染色体の構造を維持してきたかの観察を行った．さらに，核様体関連タンパク質の結合部位がホストバクテリアによってコントロールされている事を解明した．また，これまで議論されてきた配列要素や対称構造を視覚的な理解に支えられた直感的な着想で捉える事を可能にするため，汎用的なブラウザの設計と実装を行った．本学位論文においては，上に挙げた多次元的なアプローチから得られた解析結果をもとに，生命活動が遺伝情報の媒体である染色体へ与える影響を議論する．

キーワード：バクテリア環状染色体，ゲノム複製，ゲノム・染色体構造，核様体，バイオインフォマティクス，比較ゲノム解析

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AIMS	Architecture-imparting sequences
AJAX	Asynchronous JavaScript + XML
Ala	Alanine
B. subtilis	*Bacillus subtilis*
CAI	Codon Adaptation Index
CDR	Chromosome dimer resolution
ChIP	Chromatin immunoprecipitation
E. coli	*Escherichia coli*
F. alni	*Frankia alni*
GCSI	GC skew index
GEIs	Genomic islands
H-NS	Histone-like nucleoid structuring protein
HGT	Horizontal gene transfer
HMM	Hidden Markov model
KOPS	FtsK-orienting polar sequences
Leu	Leucine
NAP	Nucleoid-associated protein
Pro	Proline
RMSE	Root mean square error
SOAP	Simple Object Access Protocol
SVG	Scalable Vector Graphics
Ser	Serine
Val	Valine
ZUI	Zoomable user interface
bp	base pair
ssDNA	single strand DNA
tAI	tRNA Adaptation Index

Acknowledgments

First, I sincerely wish to express my great appreciation to Professor Masaru Tomita and Assistant Professor Kazuharu Arakawa. Professor Masaru Tomita, who provided an excellent entree to the beginning of life as a scientist and have always given me a great deal of freedom. Therefore, I was always able to conduct studies freely, and I came to think of science as the ultimate game. Assistant Professor Kazuharu Arakawa has not only been an adviser who gives valuable advice to me throughout my studies but has also become my role model as a scientist. My existence has centered around him, and I have adopted his scientific philosophy as my own. I owe a very important debt to the members of my dissertation committee, Professor Mitsuhiro Itaya, Associate Professor Yasuhiro Naito, Professor Akio Kanai, and Professor Tomoyoshi Soga, for their extremely helpful and scintillating comments regarding this dissertation. I also wish to express my gratitude to a number of respected people: Assistant Professor Rintaro Saito (University of California, USA), Dr. Yoshiaki Ohashi (Human Metabolome Technologies Inc., Japan), Project Research Associate Katsuyuki Yugi, Assistant Professor Hitomi Sano-Itoh, Dr. Ayako Yachie-Kinoshita, Dr. Nozomu Yachie, Dr. Yoshihiro Toya, Dr. Kosuke Fujishima, Dr. Koshi Imami, Vincent Piras, Yoshiteru Negishi, Mikiko Hattori, Dr. Yuka Iwasaki-Watanabe, Dr. Yukino Ogawa, Hikaru Taniguchi, Kousaku Shinoda, Dr. Junichi Sugahara, Motomu Matsui, Taiko Nishino, Chikako Arakawa-Oki, Saeka Tani, and Shinya Murata. I will never forget the wonderful time I spent with these individuals.

I appreciate the cheerful days spent with several members of the Institute for Advanced Biosciences, Keio University during the seven years of my studies. Kahori Takane, Atsuko Shinhara, Kaoru Sugahara-Kikuta, Kentaro Hayashi, Haruna Imamura, Miho Tanaka, Keisuke Morita, Daiki Yamada, Seiya Fujimoto, Kazuki Oshita, Shinnosuke Murakami, Fujitaka Baba, Keiko Iino, Takayuki Ebi, Mana Ogawa-Nakagawa, Yuriko Hasebe, Hitoshi Iuchi, Keisuke Wada, Mizuki Sata, Hanae Shimo, Yuka Hirose, Chikako Okubo, Shiori Komine, Seiko Nakatsuka, Tamami Toki, Jun Yamakubo, Gembu Maryu, Kalyn Kawamoto, Kyoko Ishino, Taiyo Miyahara, Hidetoshi Itaya, and Shiori Hashimoto have served as tremendous emotional supports to me, and I am deeply grateful to each one of them. In particular, Kiyofumi Hamashima, Toshinori Tsuboi, Keita Ikegami, Tadasu Nozaki, Yuki Shindo and Norikazu Saiki have provided a stimulus-rich environment and witty conversations, friendships and meaningful discussions with them have been essential for my personal growth during my college life.

My deepest appreciation goes to my peers, Ryu Ogawa, Satoshi Tamaki, Hiroyuki Nakamura, Ayaka Hiroe, Tsuyoshi Akuzawa, Hiroaki Suzumura, Maria Iga-Takeuchi, Mio Iwasaki, Akihito Kawasaki, Daiki Nakatsu, and Ryoji Yanashima, who entered the Graduate School of Media and Governance, Keio University, at the same time that I did, and who provided very precious experiences. We have talked about the future, science and life and passed the days happily. Mere words could never express my gratitude, yet I will try to explain how grateful I am for everything that they have done.

I would like to express special thanks to several dear and irreplaceable people. Mariko Nakai has contributed to my progress with a dedicated and thoughtful heart and has always emotionally supported and encouraged me throughout my doctoral program. Satoru Nishida and Shingo Iwata have been very understanding and close friends throughout all my college years, and they have supported me with their love and good humor.

Finally, but most importantly, I would like to offer warm thanks to my dearest parents, Tetsuaki Kono and Kimie Kono. They have made innumerable contributions to my wellbeing throughout my life, and I feel deep gratitude for their genuine love. Their thoughtfulness and kindness will always be remembered.

CHAPTER 1

Introduction

"Omnis cellula e cellula."

-Rudolf Virchow

1.1 Definition of life

What is life? This question is a classical and practical proposition that requires not only a philosophical answer but also the following biological answers for the 'definition' and 'recognition' of life. Is a symbiont alive? Is a virus alive? Moreover, are cells or biomolecules alive? The definition of life is an essential challenge in life science and serves as a fundamental pillar for the recognition of life. However, there is still no broadly accepted definition of life (Chyba and McDonald 1995). In 1944, the Nobel laureate and physicist Erwin Schrödinger wrote a book with this title, 'What is Life?', and attempted to define life through physics and chemistry. In the famous chapter entitled 'Order, disorder and entropy', he made the following statement about living organisms: "*What an organism feeds upon is negative entropy. Or, to put it less paradoxically, the essential thing in metabolism is that the organism succeeds in freeing itself from all the entropy it cannot help producing while alive.*". He suggested that energy exchange, known as metabolism, in the living organism is important for avoiding decay (Schrödinger 1944). Another famous definition of life, which has been accepted within the origins-of-life community, is the 'chemical Darwinian' definition. The Darwinian definition states that "*Life is a self-sustained chemical system capable of undergoing Darwinian evolution*" (Joyce 1994). Living things should be able to reproduce and evolve through natural selection. In addition to these two definitions, as discussed above, many other definitions of life have been proposed. Although Pályi *et al.* (Pályi *et al.* 2002) and Popa (Popa 2004) listed approximately 100 different definitions that have been proposed over many years, some of these lists have been refined to small categories using two approaches (Malaterre 2010). One approach lists the elements of a sufficient or necessary system for a living organism. Examples include dynamic low entropy systems (von Bertalanffy 1968), a hierarchical organization of open systems (Prigogine 1980; Prigogine and Stengers 1984), and an unformalizable or non-computational system; this approach is referred to 'the list-based definition'. According to this approach, de Duve and Koshland propounded the following seven pillars as the principle items for the definition of life: (1) program (DNA: deoxyribonucleic acid), (2) improvisation (response to environment), (3) compartmentalization, (4) energy, (5) regeneration, (6) adaptability, and (7) seclusion (chemical control and selectivity); these are abbreviated as PICERAS (de Duve 1991; Koshland 2002). The second approach defines life based on a model that describes the functions of living systems (Maturana and Varela 1980; Gánti 2003). Ruiz-Mirazo and coworkers ascribed an autonomous system with open-ended evolutionary capacities to a living model and propounded that this system must have four properties: a semi-permeable boundary (a membrane), an energy transduction or conversion apparatus (a set of energy currencies), and two types of interdependent macromolecular components, some of which perform and directly coordinate self-construction processes (catalysts), while others store and transmit information (records) (Ruiz-Mirazo *et al.* 2004). Given these definitions of life from various perspectives, it would appear difficult to discover common traits to redefine life clearly. However, when I observe such definitions at the molecular level, the genome actually underlies life, and the chromosome, which is the medium for the genome, meets an important requirement of life. For example, enzymes or proteins that are involved in energy transduction or compartmentalization, which are related to the self-construction process, are encoded by the genome as genes. Evolution, adaptation, or response to stimuli begins with mutations that occur within the protein coding regions or structures of the genome. Furthermore, all of these events occur on the chromosome, which is the medium for the genome. Heredity, growth, regeneration, and recording are performed by the replication system of the chromosome. This suggestion that the genome underlies life can also be observed in the context of the central dogma. The central dogma is a concept of molecular biological systems in living organism that was proposed in 1958 by Nobel laureate and molecular biologist Francis Crick (Crick 1958; Crick 1970), one of the co-discoverers of the structure of DNA in 1958 (Watson and Crick 1953). This dogma asserts that genetic information is transferred from DNA to RNA to proteins through several processes. These processes are 'DNA replication' (DNA to DNA), 'transcription' (DNA to RNA), and 'translation' (protein production). DNA replication produces a copy of the original genome to transmit the information in the chromosome. According to this flow, the process of DNA replication, which is associated with the chromosome, plays a primary role for passing the genetic information on to the subsequent generations. Consequently, the chromosome has a major responsibility as the medium for

the effective transmission of hereditary information to future generations and is dense with the fundamentals of life. In other words, an understanding of the chromosome is necessary for the primordial recognition of life, and this recognition leads to another aspect of 'What is life?'.

1.2 Exponential growth of biological knowledge

Because the chromosome has a major responsibility as the medium of the genome, it is thought that evolutionary footprints are still evident in chromosomal structure. To observe the traces of evolution, genome sequencing is an essential step in molecular biology. More than a decade has passed since the first complete bacterial genome sequences, *Haemophilus influenzae* and *Mycoplasma genitalium*, were reported (Fleischmann *et al.* 1995; Fraser *et al.* 1995), and with the completion of the human genome in 2001 (Venter *et al.* 2001), genome sequencing technologies achieved a remarkable development. Parallel sequencing technology, which is referred to as 'next-generation sequencing', has contributed to the generation of unprecedented levels of data from complete genome sequences by rapidly decreasing sequencing costs and raising efficiency. Such sequencers have many kinds of platforms, such as the Roche GS-FLX 454 Genome Sequencer, the Illumina HiSeq 2000, the ABI SOLiD analyzer, the Ion Torrent Personal Genome Machine (PGM), and the Helicos HeliScope. Each platform uses various schemes: pyrosequencing, sequencing by DNA synthesis, oligonucleotide ligation, among other exemplary approaches. The read length or aptitude for *de novo* or re-sequencing depends on the platform (see review: Metzker 2010). According to the information released by the Genomes Online Database (GOLD; Pagani *et al.* 2011), as of January 2012, there were 12,260 genome projects, including 3,046 complete genome projects, and the number of projects has nearly doubled over the past two years (Liolios *et al.* 2010). In particular, the number of bacterial genome sequence projects has reached approximately 9,000.

The rapid growth of materials has been expected to accelerate the bioscience field, and a bioinformatics approach is required to process these data. When researchers extract valuable knowledge or examine a large quantity of information, the computational approach is an essential process. Currently, various bioinformatics applications and databases have been provided to manipulate biological data (Wieser *et al.* 2011; Galperin and Fernandez-Suarez 2012). Additionally, the field of scientific visualization has been an integral part of the biosciences and has also been an indispensable approach to heuristically comprehend and instinctively understand large-scale data in molecular biology (Figure 1.1). A large number of tools have been developed for information visualization and have contributed to the recognition of biological information, including biomolecular structures, expression profiles, genome annotation and sequence alignments, molecular pathways, ontology, taxonomy and phylogeny (Tao *et al.* 2004).

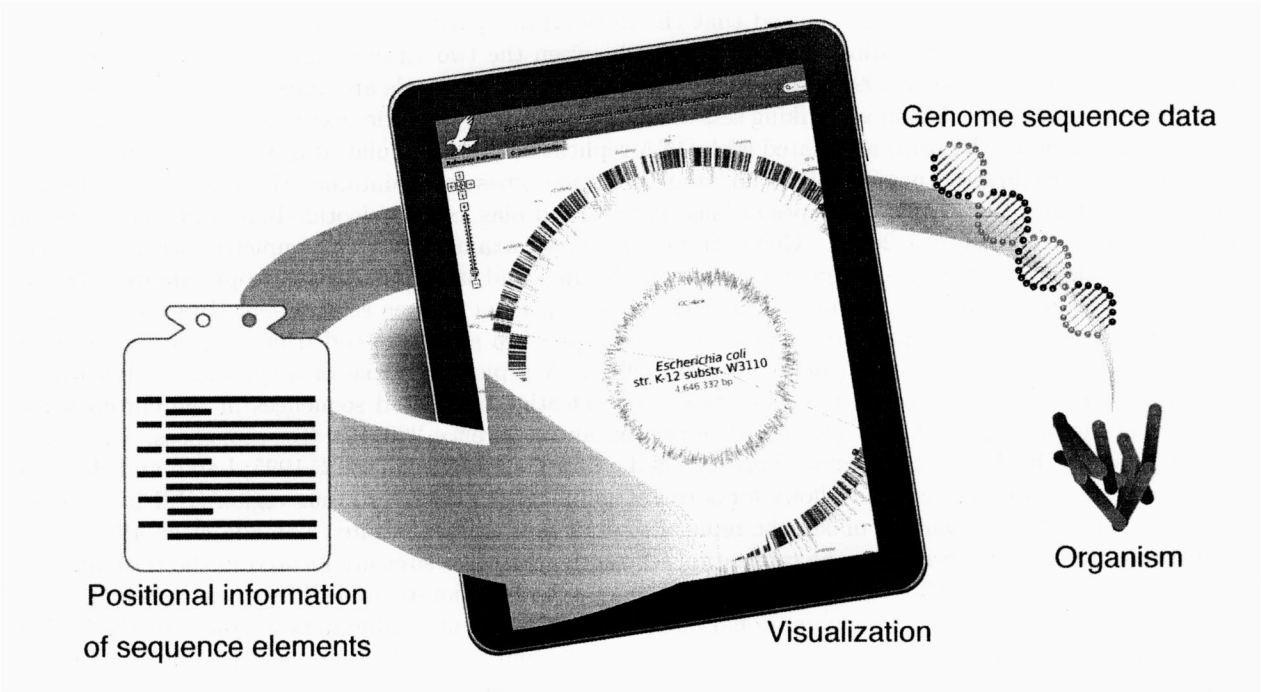

Figure 1.1: The role of visualization.
The genome sequences are obtained from organisms, and the numeric position data of sequence elements, *e.g. dif* sites, Ter sites, or protein binding sites, are detected from genome sequences. The visualization of such positional information by mapping it onto the illustrated chromosome map enables an understanding of the positional relationships.

1.3 Structures of bacterial chromosomes and genomes

In the investigation or research of the basic principles of the chromosome, the eubacteria are recognized as superior model organism because almost all eubacteria have only a single circular chromosome, one copy of each chromosome, and a small genome. Furthermore, as indicated above, there are over 2,728 publicly available bacterial genome sequences because complete genome sequencing, assembly, and annotation can be performed in less than 24 hours (Flicek and Birney 2009; Reeves *et al.* 2009). According to the accumulation of knowledge of bacterial genome sequences, the structural features in chromosome as the media of genetic information have come to be known. The nucleotide composition of genome made up of the nucleotides ATGC conforms to the Chargaff's first parity rule (A $\approx$ T and G $\approx$ C, Chargaff 1951). Interestingly enough, Chargaff also stated that this parity can apply within a single strand DNA, known as the Chargaff's second parity rule (Rudner *et al.* 1968). Sueoka further theorized and reported that the intra-strand parity rule is expected to hold under no strand bias conditions. If the mutation rates are similar between the two strands, these mutations counteract the effect of the selections (Sueoka 1995). However, violations of the parity rule are universally observed in local region of genomic sequences. The most striking sets of compositional asymmetries exist throughout the genome, generated by biological mechanisms associated with DNA replication. In particular, it is well known that various symmetric or asymmetric structures for bacterial circular chromosomes are maintained (Figure 1.2). For typical chromosome structures, there are three types of bias: gene strand bias, oligonucleotide bias, and compositional strand bias (Rocha 2004; Rocha 2008). Gene strand bias is an example of an asymmetric structure. The distributions of bacterial genes are different between the leading and lagging strands. Approximately 78% of genes are in the leading strand in Firmicutes bacteria or *Mycoplasma* (Fraser *et al.* 1995; Kunst *et al.* 1997; Rocha 2002), and 85% of ribosomal protein genes or high-expression genes are coded in the leading strand in *Bacillus subtilis* and *Escherichia coli* (McLean *et al.* 1998). A typical example of a symmetric structure is oligonucleotide bias. There are many recombination- or replication-associated sequences in the chromosome, *e.g.* architecture-imparting sequences (AIMS; Hendrickson and Lawrence 2006) or Ter sequences. In *E. coli*, Tus protein binds to the Ter sites to forms a barrier called a fork-trap (Horiuchi *et al.* 1995; Labib and Hodgson 2007), and acts as an antihelicase and allows forks to enter but not exit the terminus region (Hill *et al.* 1987; Hill 1992). As a result, this complex makes the replication fork stall at the Ter site (Kamada *et al.* 1996; Wake and King 1997). In order to terminate the replication efficiently, most Ter sites are located in the terminus half of the genome (Neylon *et al.* 2005; Mulcair *et al.* 2006). The recombination-related Chi sequence (Schultz *et al.* 1981; Kuzminov 1995), which is a recombinational hotspot, is the third most abundant oligomer in the leading strand (Blattner *et al.* 1997). During replication, when a recombination event occurs an odd number of times, the replicated chromosome is not properly segregated into two daughter chromosomes but instead produces a concatenated dimer (Sherratt 2003; Lesterlin *et al.* 2004). Therefore, many bacteria harbor highly conserved chromosome dimer resolution (CDR) machinery to separate the dimer chromosome into two monomer daughter chromosomes. In *E. coli*, chromosome dimers are resolved by two tyrosine recombinases, XerC and XerD, by the addition of a crossover at a specific 28 bp sequence called the *dif* site, which is located in the replication termination region of the chromosome (Clerget 1991; Blakely *et al.* 1993). The reaction is coordinated to the last stages of cell division by an essential cell division protein, FtsK, which functions as a septum-located DNA translocase (Steiner *et al.* 1999; Barre *et al.* 2000; Bigot *et al.* 2004; Kennedy *et al.* 2008; Dubarry and Barre 2010). FtsK moves along the chromosome unidirectionally towards the *dif* sequence, thanks to polar and oriented sequences, the KOPS (FtsK-orienting polar sequences; Saleh *et al.* 2004; Bigot *et al.* 2005; Bigot *et al.* 2006). The density of KOPS is higher in the leading strand, with their distribution biased toward the replication terminus (Bigot *et al.* 2005; Levy *et al.* 2005; Bigot *et al.* 2006), and KOPS are symmetrically conserved in the two replichores with the same frequencies (Mrazek and Karlin 1998; Salzberg *et al.* 1998). The most notable feature of the asymmetric structure, which is provided by the replication symmetry, is the nucleotide compositional strand bias, an excess of G over C in the leading strand, known as the GC skew (Lobry 1996a; Rocha *et al.* 2006). Although the localization of GC content can also be observed in eukaryotic organisms as an isochore (Bernardi 1989; Bernardi 1995), it is not as biased as the bacterial GC skew. Although the details

of the GC skew are described briefly in later chapter, when DNA is replicated according to the direction of the polymerase, the leading strand is synthesized continuously, but the lagging strand is synthesized discontinuously. As a result, the leading strand has a longer single strand time than the lagging strand, and this difference in replication mechanism provides a bias toward G and C bases (Coulondre *et al.* 1978; Lobry 1996a Reyes *et al.* 1998; Lobry and Sueoka 2002; Mackiewicz *et al.* 2003). These symmetric or asymmetric structures depend on the balance of replication arms because a pair of a replication origin and terminus defines the leading and lagging strands.

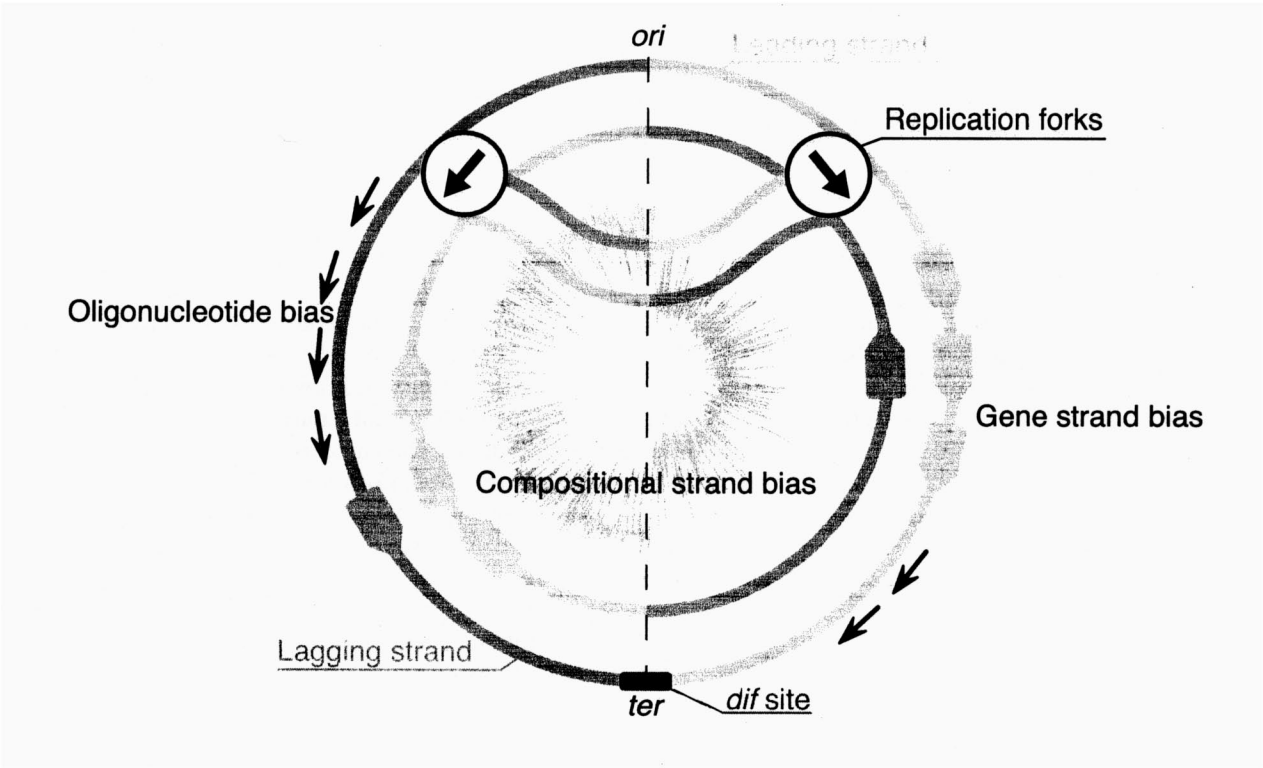

Figure 1.2: Symmetric/asymmetric structure.
The blue line represents a leading strand, and the yellow line represents a lagging strand. In almost all bacteria, the circular chromosome is symmetrical about the replication origin and terminus (gray box) and oligonucleotide bias (black arrow icons) and is asymmetrical about the compositional strand bias (green lines) and gene strand bias (blue and yellow arrows).

The bacterial replication mechanism consists of 3 steps: initiation, elongation, and termination. Although similar to eukaryotes, the mechanisms differ at several points (Rudolph *et al.* 2010). In eukaryotic cells, the initiation of replication occurs at multiple replication origins per chromosome, and the frequency is tightly regulated to exactly once per cell cycle (Diffley 2004). The cell cycle can be divided in G0 (a quiescent stage of cell), G1 (DNA synthesis), S (DNA replication), G2 (cell growth), and M phases (mitosis), and almost all eukaryotic cells progress through these phases. The replication complex is assembled before entry into S phase (Diffley 2004; DePamphilis *et al.* 2006). In S phase, the replication forks move bi-directionally from the origins until they encounter other forks originating from neighbor origins (Edenberg and Huberman 1975). The replication termination sites appear randomly within a 4 kb zone between the replication origins (Greenfeder and Newlon 1992; Zhu *et al.* 1992). By contrast, the initiation of bacterial replication occurs at a single replication origin. The replication of the chromosome proceeds bi-directionally from the origin to the terminus (Prescott and Kuempel 1972; Hirose *et al.* 1983; Schaper and Messer 1995; Schaeffer *et al.* 2005). While bacterial

replication is also regulated once per cell cycle (Nielsen and Lobner-Olesen 2008), there is a high frequency of replication, which permits bacterial cells to grow more rapidly than eukaryotic cells and is also responsible for the difference in cell sizes. For instance, *E. coli* cells can over-replicate prior to the completion of the previous replication, resulting in a generation time (20 min.) that is half of the normal replication time (Simmons *et al.* 2004; Haeusser and Levin 2008).

Furthermore, there is not only structure at the sequence level but also topological structure in the bacterial chromosome. Although bacteria do not have nuclei and the size of the chromosome is smaller than the eukaryotic chromosome, bacterial chromosomes are packaged into a nucleoid structure. The nucleoid is organized by the actions of supercoiling RNA and NAPs, which regulate DNA topology (Dame 2005). In gram-negative bacteria, at least 12 distinct types of nucleoid-associated proteins have been reported, and each type has its own DNA-binding preferences (Azam and Ishihama 1999).

1.4 Organization

As described thus far, a myriad of mutational and selectional pressures affects the structures of bacterial genomes according to their requirements to be effective vehicles of genetic information. In this dissertation, therefore, I show how such structures in bacterial circular chromosomes have evolved, and how the acquired symmetries in genomes are maintained, through a series of comprehensive computational analyses. Firstly, in order to clearly define the regions of symmetry, I have comprehensively predicted the chromosome dimer resolution site *dif*, using a novel phylogenetic pattern finding approach. Based on this information, I have then conducted computer simulations to test the quantitative contribution rates of different evolutionary pressures that have shaped the characteristic genome structure. Secondly, I have tested intrinsic and extrinsic causes that may disrupt the evolutionary structured symmetry in the genomes, and how the genome maintains its order despite the existence of disorders. Specifically, I discuss the evolutionary traces of the maintenance of genomic symmetry in light of numerous horizontal gene transfer events, and the intrinsic mechanism of the genome to protect itself using topological chromosome structure with the control of nucleoids. Lastly, I present a sophisticated software system for the visualization of these numerous evolutionary pressures and structural elements within the genomes, in order to aid the intuitive understanding of such systematic and complex phenomenon.

The bacterial chromosome has a wide variety of structures. Although these structural phenomena have been thoroughly observed, it has not been confirmed how these structures are formed or how bacteria maintain and develop these structures. Therefore, I used computational genomics approaches to address these questions by focusing on the sequence elements in bacterial chromosomes. This dissertation comprises 7 chapters, including an introduction section (here, Chapter 1). I developed a comprehensive prediction method for a chromosome dimer resolution site, known as *dif*, *in silico* using a phylogenetic prediction approach based on iterated hidden Markov modeling (Chapter 2). During the replication process in bacteria with circular chromosomes, an odd number of homologous recombination events results in concatenated dimer chromosomes that cannot be partitioned into daughter cells. However, the *dif* sequence has only been identified in a few bacteria. Additionally, the *dif* sequence is suggested to occur at a site other than the *dif* site. Accordingly, I aimed to obtain the sequences comprehensively and predict the *dif* sequences. Using this phylogenetic approach, *dif* sites were identified in 641 organisms among 16 phyla, with a 97.64% identification rate for single-chromosome strains. The *dif* sequence positions were shown to be strongly correlated with the GC skew shift-point that is induced by replication-associated mutation/selection pressure, but the difference in the positions of the predicted *dif* sites and the GC skew shift-points did not correlate with the degree of replication-associated mutation/selection pressure. This comprehensive identification of unique *dif* candidates can provide materials to elucidate the appropriate machinery for replication termination (Kono *et al.* 2011). Using these predicted *dif* sequences, in the next chapter, I validated the bacterial replication termination models using a computational simulation of

genomic mutations in terms of the compositional strand bias and suggested an appropriate model (Chapter 3). In bacterial circular chromosomes, replication is known to be terminated when any of the following occurs: the forks progressing in opposite directions meet at the distal end of the chromosome, the replication forks become trapped by Tus proteins bound to Ter sites, or the termination occurs at a single definite site, *dif*. To understand this difference between the known replication machinery and the genomic compositional bias, I undertook a simulation study of genomic mutations and reported here how different replication termination models contribute to the generation of replication-related genomic compositional asymmetry. This study could confirm how chromosome structures are formed and how the determination of the replication termination point enables the definition of a symmetric or asymmetric median line in the circular chromosome. From these insights, the symmetry of the circular chromosome was defined. Then, to investigate how the defined symmetric chromosome structure has been maintained, I considered computationally whether the symmetry of the bacterial circular chromosome is disrupted by genomic islands (Chapter 4). The term symmetric chromosome here means the compositional strand bias, known as the GC skew. The GC skew is widely used as an *in silico* method for the prediction of the replication origin and terminus (Frank and Lobry 2000; Worning *et al.* 2006). However, the strength of the GC skew is extremely variable. Some bacteria have only weak biases (Zhang *et al.* 2003; Worning *et al.* 2006; Arakawa *et al.* 2009a), and there are many bacteria that have an imbalance in GC skew. Therefore, I hypothesized that the balance of the GC skew may be affected by mutations in the genome. Among mutations, genomic islands are considered to have the possibility to change the base composition on a large scale. Genomic islands are large foreign regions of approximately 10 Kbp to 1 Mbp in the bacterial genome (Rocha 2008), which were most likely acquired by horizontal gene transfer (HGT; Gogarten and Townsend 2005; Juhas *et al.* 2009). In Chapter 4, I compared the symmetric structures of natural genomes and artificial genomes in which the genomic islands were deleted computationally and investigated the disruption effects and the strength of the GC skew. In Chapter 5, I extended the analysis to the macroscopic level and observed the side effects of the nucleoid, which physically controls the bacterial chromosome upon an acquisition of new function (Chapter 5). H-NS, known as the nucleoid-associated protein, changes the topology of DNA and acts as a transcriptional regulator for the silencing of horizontally acquired genes. Although this transcriptional regulation is well studied, no definitive reason has been provided for the evolution of this function. In this chapter, I investigated codons in the motifs of H-NS binding regions and found that these codons were used more often in horizontally acquired genes than in other regions. Furthermore, they were infrequently used in highly expressed and essential genes. Accordingly, it is likely that the immune-like role of H-NS developed as a byproduct, to prevent the disruption of the bacterial life cycle. Finally I developed a web-based browser for molecular biology with a zoomable user interface (Chapter 6). An understanding of the positional relationships between gene directions, oligonucleotide sequence sites, base compositional biases, foreign regions, protein binding sites and replication origins and termini are important for understanding bacterial chromosome structures. Although the localization of genes or positions of some sequence elements are known numerically for one organism, such numeric does not necessarily lead to instinctive recognitions. Therefore, I implemented a viewer with a design such that each genomic element is mapped on a circular chromosome map to gain an instinctive understanding and inspiration for subsequent investigations. This system was developed as an additional function of MEGU (Kono *et al.* 2006) and Pathway Projector (Kono *et al.* 2009). The Pathway Projector is a biochemical pathway browser and was implemented based on the requirements for a versatile browser: (1) comprehensive search features and data access; (2) data mapping and the ability to edit and annotate maps; (3) an intuitive user experience without the requirement for installation and regular maintenance. I took advantage of the versatility of the Pathway Projector to add the circular chromosome maps. The circular map enabled the visual observation of living organisms from multiple viewpoints. In Chapter 7, I provided a brief summary of the analyses and perspectives derived from the studies described in this dissertation in the concluding chapter.

Throughout this dissertation, I performed comprehensive analyses of the genomic functionality, characteristics, and evolution of bacterial chromosome structure which serves as a medium of living systems, and would like to discuss how the chromosome composition and structures adapted to biological mechanism in the course of evolution.

CHAPTER 2

Comprehensive prediction of chromosome dimer resolution sites in bacterial genomes

"We can only see a short distance ahead, but we can see plenty there that needs to be done."

-Alan Turing 'Computing Machinery and Intelligence'

2.1 Background

In bacteria, replication fork arrest is mainly repaired by homologous recombination (Michel *et al.* 2004). When such a recombination event occurs an odd number of times in one DNA replication event of circular chromosomes, the replicated chromosome is not properly segregated into two daughter chromosomes but instead produces a concatenated dimer (Sherratt 2003; Lesterlin *et al.* 2004). Therefore, many bacteria harbor highly conserved chromosome dimer resolution (CDR) machinery to separate the dimer chromosome into two monomer daughter chromosomes.

In *Escherichia coli*, chromosome dimers are resolved by two tyrosine recombinases, XerC and XerD, by the addition of a crossover at a specific 28 bp sequence called the *dif* site, which is located in the replication termination region of the chromosome (Clerget 1991; Blakely *et al.* 1993). The *dif* sequence contains a pair of palindromic sequence motifs that correspond to the binding domains of XerC and XerD. The reaction is coordinated to the last stages of cell division by an essential cell division protein, FtsK, which functions as a septum-located DNA translocase (Steiner *et al.* 1999; Barre *et al.* 2000; Bigot *et al.* 2004; Kennedy *et al.* 2008; Dubarry and Barre 2010). FtsK moves along the chromosome unidirectionally towards the *dif* sequence, thanks to polar and oriented sequences, the KOPS (Saleh *et al.* 2004; Bigot *et al.* 2005; Bigot *et al.* 2006). CDR is initiated when FtsK reaches *dif* and its extreme C-terminal domain directly interacts with the C-terminal domain of XerD (Aussel *et al.* 2002; Yates *et al.* 2003; Massey *et al.* 2004; Yates *et al.* 2006; Bonne *et al.* 2009). The *dif*/XerCD chromosome dimer resolution system seems widely conserved. *In vivo* experimental evidence for its conservation has been obtained in *Xanthomonas campestris*, *Caulobacter crescentus* and *Vibrio cholerae* (Yen *et al.* 2002; Val *et al.* 2008; Wang *et al.* 2006). *In vitro* characterization of Xer recombinases and *dif* sites has also been carried in *Haemophilus influenzae* and *Bacillus subtilis* (Neilson *et al.* 1999; Sciochetti *et al.* 2001). However, the importance of *dif*/XerCD for the fitness of bacteria has only been demonstrated in *E. coli* and *V. cholerae* (Cornet *et al.* 1996; Val *et al.* 2008). In some other bacteria, like *Lactococci* and *Streptococci*, chromosome dimer resolution is resolved by single tyrosine recombinases that act at specific *dif* site (Le Bourgeois *et al.* 2007; Nolivos *et al.* 2010). In this case, dimer resolution still depends on FtsK and *dif* is still located opposite the origin of replication between oriented polar sequences (Campo *et al.* 2004). Several filamentous phages are known to hijack this site-specific recombination machinery of *dif*/XerCD for their integration into the host chromosome, containing pseudo-*dif* sequences within these phage genomes (Lin *et al.* 2001; Huber and Waldor 2002; Campos *et al.* 2003; Val *et al.* 2005; Derbise *et al.* 2007; Campos *et al.* 2010; Das *et al.* 2010). However, the *dif* sequence remains intact during such recombination process to ensure the integrity of chromosome dimer resolution machinery (Blakely 2004; McLeod and Waldor 2004). The *dif*-like sequences in phages often contain more variable central region that is longer than the canonical 6 bp (Val *et al.* 2005; Campos *et al.* 2010; Das *et al.* 2010), and the XerD binding arm is considerably degenerate (Lin *et al.* 2001).

Because there is only one origin of replication on bacterial circular chromosomes, replication generally terminates in a specific region of the chromosome. This can be followed by the existence of a GC skew on the two replichore arms of the chromosomes with a shift-point opposite the origin of replication (Lobry 1996a). Based on the observation that *dif* sites are generally located at or near the GC skew shift-point, Hendrickson and Lawrence proposed that replication might generally terminate at *dif*, which coordinate replication and chromosome dimer resolution (Hendrickson and Lawrence 2007). In *E. coli*, the replication process usually terminates at a narrow region that includes approximately 5% of the genome length and is located directly opposite the replication origin (Louarn *et al.* 1977; de Massy *et al.* 1987; Hill *et al.* 1987). This is partly due to the existence of the Ter/Tus replication fork-trap (Hill *et al.* 1987). *dif* is located within the replication fork-trap but termination occurs precisely at the Tus site, not at *dif* (Duggin and Bell 2009) and *dif* is active when displaced outside of the replication termination region if it is still within the zone where KOPS converges (Cornet *et al.* 1996). Nevertheless, the lack of universal conservation of the Tus protein may suggest that replication terminated at

dif sites until the relatively recent takeover by the Ter-Tus system (Duggin *et al.* 2008). We reasoned therefore that the comprehensive identification of *dif* sites and of their location with respect to the GC skew shift-point in hundreds of complete genomes might provide clues to the evolution of the CDR machinery and its possible link with the replication termination mechanism in bacterial species.

Prediction of the *dif* sequences has been reported by several groups with different approaches. Hendrickson and Lawrence showed that sequence skew can be used to predict the locations of *dif* sites, and they identified putative *dif* sequences in 25 bacteria based on sequence similarity (Hendrickson and Lawrence 2007). Le Bourgeois and colleagues reported a new type of tyrosine recombinase, named XerS, which is responsible for CDR in *Streptococci* and *Lactococci* and this recombinase targets a 31 bp sequence element named dif_{SL} (Le Bourgeois *et al.* 2007). For comparison, they predicted *dif* sequences in 22 Firmicutes based on their similarity to that of *B. subtilis* with Megablast (Zhang *et al.* 2000) and on the fact that the *dif* sequence occurs only once per genome. Val and colleagues identified that *V. cholerae* chromosome II, whose many features are plasmid-like, has an original *dif* sequence independently, and therefore it has FtsK-dependent CDR (Val *et al.* 2008). For this purpose, they predicted *dif* sequences in five α-Proteobacteria and ten β-Proteobacteria that harbor multiple chromosomes, and discussed a conserved FtsK-dependent CDR on multiple chromosomes based on the close relative distance of the positions of *dif* sequences and the GC skew shift-points. Their prediction method is based on a HMMER (Eddy 1998) score ($< 10^{-5}$) with a profile built from 27 aligned *dif* sequences in the largest chromosomes of γ-Proteobacteria species, with manual checking for 6 bp spacing between two XerC and XerD binding motifs.

Carnoy and Roten reported the most comprehensive predictions to date, identifying putative *dif* sequences in 204 chromosomes in 137 Proteobacteria strains, discussing the high conservation of *dif*/XerCD systems and the possible loss of *dif* sequences in endosymbionts, with suggestions for other CDR mechanisms (Carnoy and Roten 2009). Here, the prediction was based on BLAST searches and YASS alignment (Noe and Kucherov 2005) with the *dif* sequences of *E. coli* and *B. subtilis*, and candidates were selected based on their proximity to the GC skew shift-points and a single occurrence per chromosome. Previous predictions were therefore limited to three bacterial phyla: Proteobacteria, Firmicutes and Actinobacteria.

To this end, we describe comprehensive predictions for *dif* sequences based on a machine learning approach, tracing the phylogenetic conservation patterns of XerCD recombinases and using an iterative hidden Markov modeling method. Furthermore, we observed the relationship between predicted *dif* sequence positions and GC skew shift-points, and investigated whether replication termination occurs at the *dif* site.

2.2 Materials and Methods

2.2.1 Software and sequences

All analyses in this study were conducted using programs written in Perl with the G-language Genome Analysis Environment, version 1.8.10 (Arakawa *et al.* 2003; Arakawa and Tomita 2006; Arakawa *et al.* 2008). Hidden Markov modeling and searching was conducted with HMMER, version 2.3.2 (Eddy 1998). The *dif* sequence is the binding site of the XerCD recombinase; therefore, we first selected 734 circular bacterial chromosomes among 658 species/strains according to their conservation of XerCD using the KEGG (Kyoto Encyclopedia of Genes and Genomes) Orthology database (KO; Kanehisa *et al.* 2010). We obtained these sequences from the NCBI FTP Repository (2009, http://www.ncbi.nlm.nih.gov/Ftp). The following experimentally confirmed (*E. coli* and *B. subtilis*) or computationally predicted (*F. alni*) *dif* sequences were used as seed sequences for subsequent searches and machine learning:

E. coli 5'-GGTGCGCATAATGTATATTATGTTAAAT-3' (Blakely and Sherratt 1994)

B. subtilis 5'-ACTTCCTAGAATATATATTATGTAAACT-3' (Sciochetti *et al.* 2001)

F. alni 5'-CACGCCGATAATGCACATTATGTCAAGT-3' (Hendrickson and Lawrence 2007)

2.2.2 Iterated hidden Markov modeling

XerCD conservation does not immediately imply *dif* sequence conservation (Val *et al.* 2008). Therefore, to determine the phylogenetic conservation patterns of XerCD, we first aligned all XerCD amino acid sequences in the 734 genomes analyzed in this work with those in organisms with the above-mentioned *dif* sequences using ClustalW (Thompson *et al.* 1994). The average distances of XerC and XerD sequences that were calculated from this alignment were used to infer phylogenetic conservation patterns among phyla.

Based on the phylogenetic conservation patterns of XerCD, we iteratively created the hidden Markov models (HMM) for the accurate prediction of *dif* sequences, seeded with the previously described *dif* sequences (Figure 2.1A). Iterated HMM is shown to be able to build a more diverse and potentially more sensitive models than regular HMM, by incorporating distant homologous sequences while avoiding the contamination of non-homologous sequences into the model (Johnson *et al.* 2010), and thus iterative HMM has been frequently utilized in bioinformatics and computational biology (Altschul *et al.* 1997; Karplus *et al.* 1998; Schäffer *et al.* 2001; Scheeff and Bourne 2006). In this work, the first profile hidden Markov model was created from the *dif* sequences identified in genomes belonging in the same genus as the genome harboring the seed sequence. For example, in Proteobacteria, the seed sequences came from *E. coli*; therefore, the *dif* sequences were searched in 28 genomes belonging to the genus *Escherichia* by means of fuzzy matching with the seed sequences of *E. coli* K12 using Perl module String::Approx 3.26 (http://search.cpan.org/~jhi/String-Approx-3.26/Approx.pm). For fuzzy matching, the maximum numbers of insertions, deletions, and substitutions were previously determined to be 0 bp, 0 bp, and 8 bp, respectively (Blakely and Sherratt 1994). Likewise, initial profiles were created for Firmicutes based on 24 genomes in the genus *Bacillus* and for Actinobacteria based on two genomes in the genus *Frankia*. Based on these initial profile hidden Markov models, *dif* sequences were predicted in the genomes of the closest genus to the seed genus according to the amino acid sequences of XerCD proteins.

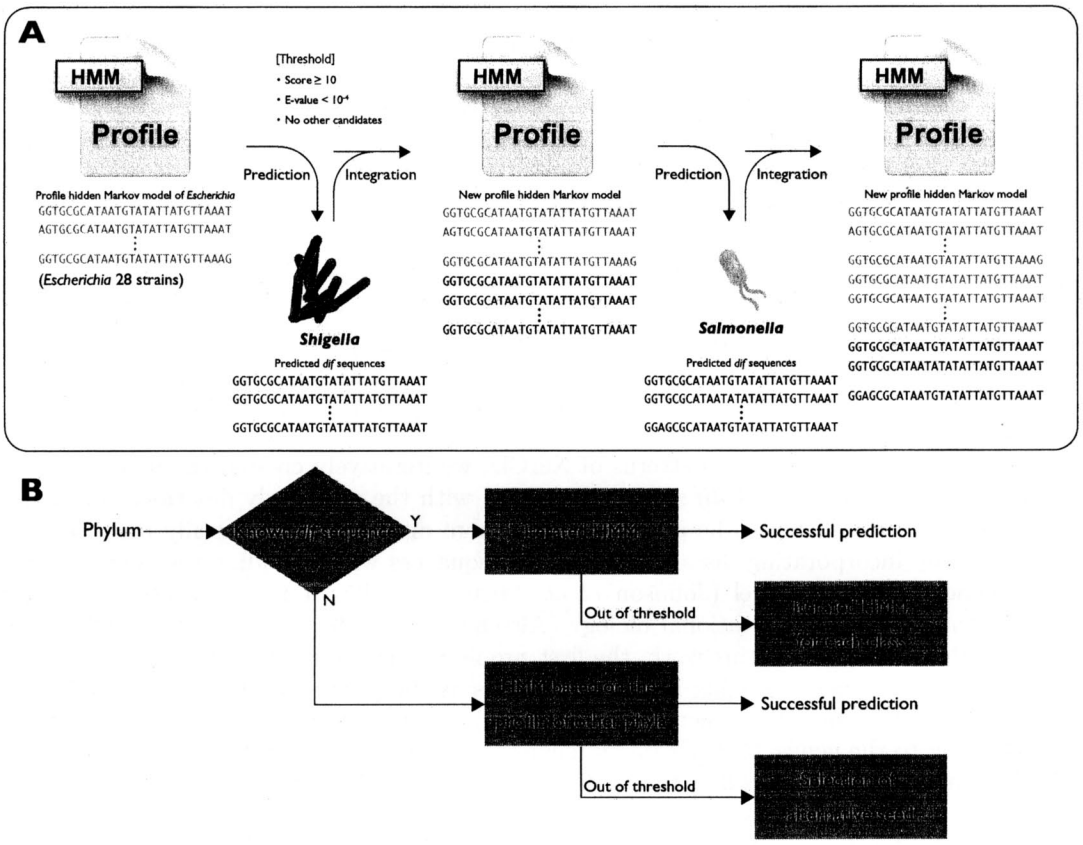

Figure 2.1: Prediction strategy.

A: Example of the iterated HMM in Proteobacteria. The first seed profile hidden Markov model is created from the seed *dif* sequence of *Escherichia coli*, by searching for *dif* sequences in 28 genomes belonging to the genus *Escherichia* by means of fuzzy matching. Based on this initial profile hidden Markov model, *dif* sequences were predicted in the genomes of the closest genus to the *Escherichia* genus (in this case, *Shigella*) according to XerCD amino acid sequences. Subsequently, a new profile is created using the previous profile and the newly predicted *dif* sequences, and this new profile is used to predict in the second closest genus (in this case, *Salmonella*). In this way, profile creation and *dif* sequence prediction were repeated recursively in decreasing order of similarity of XerCD from the *Escherichia* sequence. The iterated HMM is conducted for each phylum. B: Flow chart of the overall strategy.

In the case of Proteobacteria, an initial profile was created using genomes belonging to the genus *Escherichia*, and this profile was used to predict *dif* sequences in the genus *Shigella*. Subsequently, a new profile was created using the previous profile and the newly predicted *dif* sequences, and this new profile was used to predict the second nearest genus (in the case of Proteobacteria, *Salmonella*). In this way, profile creation and *dif* sequence prediction were iterated in decreasing order of similarity of XerCD from the seed sequences; thus, iterated HMM was conducted for each phylum. Because no *dif* seed sequences were available for phyla other than the three described above, the three profile hidden Markov models obtained by iterated HMM in Proteobacteria, Firmicutes, and Actinobacteria were used as the initial profiles. At each iterated HMM, predicted candidates were validated according to the following criteria: 1) HMMER score ≥ 10 and e-value $< 10^{-4}$, 2) leave-one-out cross-validation using the new profiles, and 3) conservation of the palindromic structure. For cross-validation, each time a new profile was created in the iterated HMM, we tested the validity of the training set by leaving out one of the *dif* sequences from the accumulated set of *dif* sequences and checking that the prediction of the left-out sequence by training with all of the other *dif* sequences is always above the threshold for all *dif* sequences collected up to that iteration. For the palindromic structure, positions 7-12 bp and 17-22 bp of *dif* sequences, corresponding to the binding sites of XerC and XerD, were checked for complementarities. For example, the palindromic structure of *E. coli dif* sequences in bracket notation is "--(--- ((((((-()-))))))) ---)--", and the conservation threshold is set to more than four pairs of complementarities within the 7-12 bp and 17-22 bp positions of the predicted *dif* sequences.

Although iterated HMM is based on phyla, this taxonomic unit is sometimes too diverse to accurately follow phylogeny with recursive means. Therefore, prediction was separately conducted in classes instead of phyla for 60 strains, harboring 130 chromosomes for classes α-, β- and γ-Proteobacteria. Similarly, sometimes, a species is highly phylogenetically distant from the seed organism, making it the case that utilization of profile hidden Markov models from other phyla is more suitable than own phyla's profile. When iterated HMM fails in such cases, an alternative seed profile is created using the *dif* sequences from the top three genomes with the closest XerCD sequences, as determined by alignment using ClustalW (Figure 2.1B).

GC skew's shift-point, calculated as $(C - G)/(C + G)$, was computed using the "find_ori_ter" function of the G-language GAE (Arakawa *et al.* 2003), based on the cumulative GC skew (Grigoriev 1998) at 1 bp resolution. Although GC skew is widely observed in bacterial species, a number of genomes do not exhibit notable compositional bias (Arakawa and Tomita 2007a; Arakawa *et al.* 2007). To determine the presence of genomic nucleotide compositional bias, the GC skew index (GCSI) was calculated for all genomes, and GCSI ≥ 0.05 was used as the threshold (Arakawa and Tomita 2007a; Arakawa *et al.* 2009a). GCSI quantifies the degree of GC skew using the compositional distance between the leading and lagging strands and the spectral amplitude of 1 Hz signal of GC skew graph using Fast Fourier Transform. In this study, the replication origin is defined based on the cumulative GC skew at 1 bp resolution using the G-language GAE.

2.2.3 Calculation of the conservation quantity of *dif* sequences

Conservation quantity was calculated based on the nucleotide variance in each position of *dif* sequences. Firstly, we calculated the position-specific base composition of all *dif* sequences in a group (phylum or class). Subsequently, variance of the most frequent base in that position is calculated from the base composition. For example, when a group with 100 *dif* sequences has nth base composition of (A, T, G, C = 100, 0, 0, 0) or (A, T, G, C = 25, 25, 25, 25), the variance is 2,500 or 0, respectively. Hence, if the position-specific base composition is biased toward any one base, its high variance indicates high degree of conservation. These values are normalized to percentages for comparison with other groups. In the case of multiple chromosomes, since these conservation quantities were calculated in each strain, the average value was used for normalization.

2.3 Results

2.3.1 Overview of *dif* sequence prediction

We first analyzed the phylogenetic conservation patterns of XerC and XerD in bacterial species by calculating the distances of their amino acid sequences from those in the seed organisms with known *dif* sequences (experimentally confirmed: *E. coli* and *B. subtilis* and computationally predicted: *Frankia alni*). As depicted in Figure 2.2 and Figure 2.3, sequence similarity distributions were clearly distinguished by phylum. Sequences belonging to different phyla always showed ClustalW distances of ≥ 0.3, and based on this phylogenetic distribution pattern, we separately trained and predicted the *dif* sequences in each phylum using iterated HMM. By this phylogenetic prediction approach, we predicted *dif* sequences in 578 genomes out of 592 that harbor the XerCD recombinase. The same prediction method was applied for 66 organisms with multiple chromosomes, totaling 142 chromosomes, where we could predict *dif* sequences in 63 organisms with 137 chromosomes (Table 2.1).

All of these predictions resulted in unique hits above the threshold, and their validity was further confirmed through leave-one-out cross-validation. On the other hand, predictions below the threshold (score < 10 and e-value $> 10^{-4}$) often resulted in multiple candidates with insufficient scores. When the initial prediction using the strict threshold failed, we manually checked the predicted sequences for the conservation of palindromic structure in the 7-12 bp and 17-22 bp positions, and candidates that were located close to the origin of replication were removed because the displacement of a *dif* sequence near the origin significantly reduces the growth rate (Cornet *et al.* 1996).

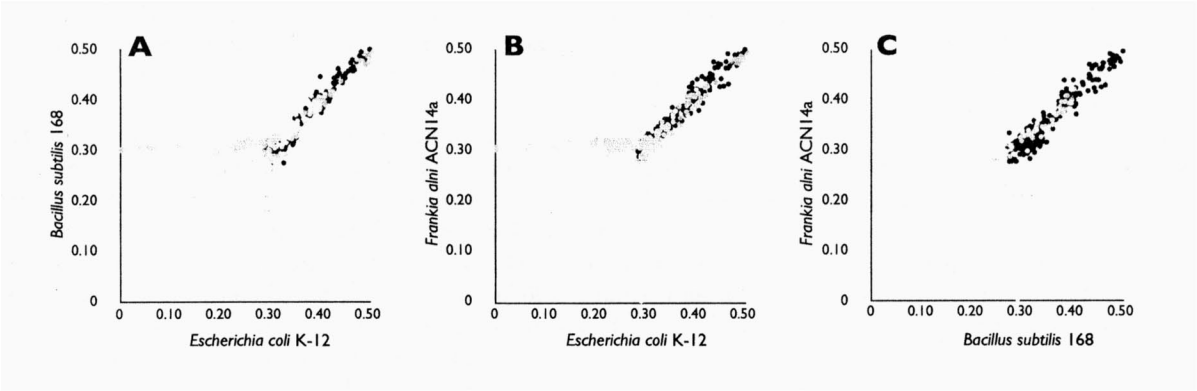

Figure 2.2: The phylogenetic distance of XerCD in each organism.
The phylogenetic distances of bacterial genomes to three seed organisms, *Escherichia coli* (Proteobacteria), *Bacillus subtilis* (Firmicutes) and *Frankia alni* (Actinobacteria), were calculated as the average of phylogenetic distances of XerC and XerD. Detailed example is given in Figure 2.3. A to C are scatter plots of the distances of these genomes to the seed organisms. Axes represent average distances as calculated by ClustalW. A: Distances from *Escherichia coli* K-12 and *Bacillus subtilis* 168; B: distance from *Escherichia coli* K-12 and *Frankia alni* ACN14a; and C: distance from *Bacillus subtilis* 168 and *Frankia alni* ACN14a. Blue represent the genomes of Proteobacteria, green represent Firmicutes, yellow represent Actinobacteria, and the gray marks represent other phyla. All phyla show strong preferences for seeds from the same phylum.

2.3.2 Prediction results of each phylum

In Proteobacteria, fuzzy matching in 28 *Escherichia* strains based on the *dif* sequence of *E. coli* K12 for the creation of an initial seed profile hidden Markov model yielded a unique *dif* sequence in each of the 28

Table 2.1: Prediction result overview.

chr : chromosomes

Single Chromosome	Organism	Predicted	%
Proteobacteria	362	357	98.61
Firmicutes	100	97	97.00
Actinobacteria	66	66	100.00
Bacteroidetes	19	19	100.00
Chlamydiae	14	14	100.00
Chlorobi	11	11	100.00
Acidobacteria	3	3	100.00
Verrucomicrobia	3	3	100.00
Chloroflexi	3	3	100.00
Gemmatimonadetes	1	1	100.00
Nitrospirae	1	1	100.00
Elusimicrobia	1	1	100.00
Tenericutes	1	1	100.00
Spirochaetes	1	1	100.00
Cyanobacteria	5	0	0.00
Planctomycetes	1	0	0.00
Total	592	578	97.64

Multiple Chromosomes	Organism (chr)	Predicted (chr)	% (chr %)
Proteobacteria	60 (130)	57 (125)	95.00 (96.15)
Spirochaetes	6 (12)	6 (12)	100.00 (100.00)
Total	66 (142)	63 (137)	94.45 (96.48)

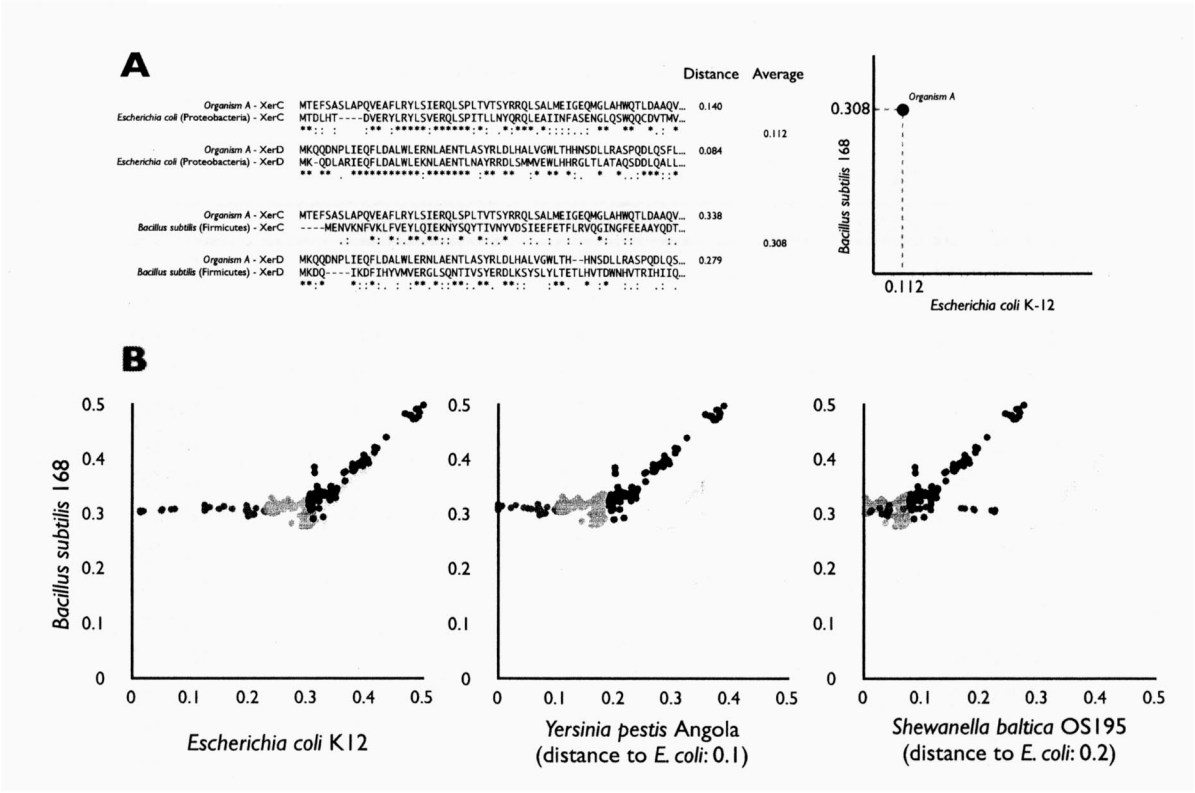

Figure 2.3: An example for the plot shown in Figure 2.2.

This phylogenetic distance was based on XerC and XerD amino acid sequence alignment in each other strain, and we calculated this distance using the average distance matrix that is generated from the pairwise scores. A: In this case of *organism A* in Proteobacteria, the average phylogenetic distances between *organism A* and *Escherichia coli* is 0.112, and that between *organism A* and *Bacillus subtilis* is 0.308. According to this scores, the *organism A* was plotted as the diagram. Since *organism A* belongs in Proteobacteria (represented as blue in Figure 2.2), it has shorter distance to *E. coli* than to *B. subtilis*. B: These graphs are shown by other spectrums. The left figure is Proteobacteria (blue) plot in Figure 2.2A, and colored in each distance (distance 0-0.1, 0.1-0.2, 0.2-0.3 and other are represented by purple, red, orange and gray respectively). The middle and right figures are used *Yersinia pestis* Angola (distance = 0.1) and *Shewanella baltica* OS195 (distance = 0.2) as x-axis respectively. The observation by such spectrum changes shows that the XerCD amino acid sequences are distinguished in each strain.

strains. Iterated HMM using this seed profile resulted in unique predictions over the validation threshold in 306 genomes. An additional 137 chromosomes in 69 genomes were predicted with iterated HMM separated by classes, and 10 distant genomes were predicted using an alternative seed profile created with the 3 most similar genomes. The predicted *dif* sequences totaled 482 in 414 organisms, with a prediction rate of 98.61% for single-chromosome strains and 95.00% for multiple-chromosome strains. Predictions failed in eight organisms and ten chromosomes, namely, *Agrobacterium tumefaciens* str. C58, *Paracoccus denitrificans* PD1222 chromosome I, II (α-Proteobacteria), *Burkholderia phytofirmans* PsJN chromosome I, *Burkholderia* sp. 383 chromosome I, III, *Nitrosospira multiformis* ATCC 25196 (β-Proteobacteria), *Desulfotalea psychrophila* LSv54 (δ-Proteobacteria), *Sulfurimonas denitrificans* DSM 1251 and *Nitratiruptor* sp. SB155-2 (ϵ-Proteobacteria).

For Firmicutes, fuzzy matching in 17 *Bacillus* strains (based on the *dif* sequence of *B. subtilis* str. 168 for the creation of the initial seed profile hidden Markov model) yielded a unique *dif* sequence in each of the 17 strains. Iterated HMM using this seed profile resulted in unique prediction over the validation threshold for 79 chromosomes in 79 genomes. The *dif* sequences are predicted in a total of 97 organisms, with a prediction rate of 97.00%. Prediction failed in three genomes, namely, *Clostridium perfringens* str. 13, *C. beijerinckii* NCIMB 8052 (Clostridia), and *Lactobacillus helveticus* DPC 4571 (Lactobacillales).

Although no experimentally confirmed *dif* sequence is available for Actinobacteria, that of *F. alni* is suggested to be 5'-CACGCCGATAATGCACATTATGTCAAGT-3' (Hendrickson and Lawrence 2007). Therefore, we used this sequence for fuzzy matching in two genomes, *Nocardia farcinica* IFM 10152 and *Mycobacterium avium* subsp. paratuberculosis K-10, whose XerCD amino acid sequences were most similar to those of *F. alni*. Iterated HMM using this seed profile resulted in successful predictions above the validation threshold in all 66 genomes.

In Chlorobi, an initial seed profile was created with predicted *dif* sequences in *Chlorobaculum parvum* NCIB 8327 and *Prosthecochloris aestuarii* DSM 271 that scored above the validation thresholds using the Firmicutes profile, which resulted in the highest scores compared to those of Proteobacteria and Actinobacteria. Likewise, the profile of Firmicutes yielded the highest scores in Chlamydiae, where the initial seed profile was created from predicted *dif* sequences in *Chlamydophila pneumoniae* CWL029 and *Protochlamydia amoebophila* UWE25, which were below the validation thresholds, but contained palindromic structure and were located within 0.01-1.48 degrees from the shift-points of GC skew. Using these seed profiles, iterated HMM successfully predicted *dif* sequences in all 11 genomes in Chlorobi and 14 genomes in Chlamydiae.

Because the number of genomes is very small in all of the other phyla, we utilized the profiles of Proteobacteria, Firmicutes, Actinobacteria, Chlorobi, and Chlamydiae that were created thus far instead of applying iterated HMM based on specific seed profiles, and all of the following candidates were confirmed based on scores, palindromic structure, and position. In Elusimicrobia and Tenericutes, all profiles showed high HMMER scores, and predictions using the profiles of Firmicutes and Chlamydiae predicted identical *dif* sequences. Similarly, the profiles of Firmicutes, Chlamydiae, and Proteobacteria predicted identical *dif* sequences in Nitrospirae, and predictions based on the profiles of Proteobacteria and Chlorobi were identical in Gemmatimonadetes.

In Spirochaetes, predictions using the profiles of Firmicutes, Chlamydiae and Proteobacteria profiles resulted in unique *dif* sequences in species with single chromosomes, and the profiles of Firmicutes were used for the predictions of 12 chromosomes in 6 species with multiple chromosomes, all with HMMER scores above the validation thresholds. The most suitable profiles varied among species in other phyla. In Acidobacteria, the *dif* sequence of *Acidobacterium capsulatum* ATCC 51196 was predicted by the profiles of Firmicutes, Chlamydiae, and Chlorobi *dif* sequences, and other species were predicted using the profile of Firmicutes only. In Verrucomicrobia, profiles based on Proteobacteria, Firmicutes and Chlorobi predicted *Methylacidiphilum infernorum* V4, and that of Proteobacteria and Firmicutes predicted *Opitutus terrae* PB90-1 and *Akkermansia muciniphila* ATCC BAA-835. In Chloroflexi, the Chlorobi profile was suitable for *Dehalococcoides* sp. BAV1 and *Dehalococcoides* sp. CBDB1, and that of Actinobacteria was used in *D. ethenogenes* 195 *dif* sequences. *dif* sequences were

predicted in 14 Bacteroidetes strains using the profile of Proteobacteria, and those in five strains were predicted using alternative profiles created with the three most similar genomes. In this way, we successfully predicted *dif* sequences in most phyla, although the prediction failed in the phyla Cyanobacteria and Planctomycetes.

2.3.3 Correlation of the *dif* sequence position and the GC skew shift-points

Using the predicted *dif* sequences, we compared their positions within the genome to the shift-points of the GC skew. Firstly, we analyzed the distributions of relative genomic distances of *xerC*, *xerD* and *ftsK* genes from the predicted *dif* sites. As a result, *xerC* genes were mostly located near the *dif* sites, *xerD* genes were near the replication origin, and *ftsK* genes were located mostly in between *xerC* and *xerD* genes (Figure 2.4). The comparison of positions between predicted *dif* sites and the shift-points of the GC skew showed that the *dif* sequences predicted in the phyla Proteobacteria and Firmicutes correlated significantly with the GC skew shift-points that are highly likely to be located within the terminus region (Spearman's rank-correlation coefficients: $\rho = 0.844$ and 0.715, respectively; Figure 2.5A). The differences among these positions fell to within 0.00-1.39% of the genome for $\pm$ 1σ, and outliers did not exceed 3% in distance relative to the genome size (Figure 2.6). The above results confirm that chromosome replication and CDR are related, and that show the accuracy of the predictions described in this work.

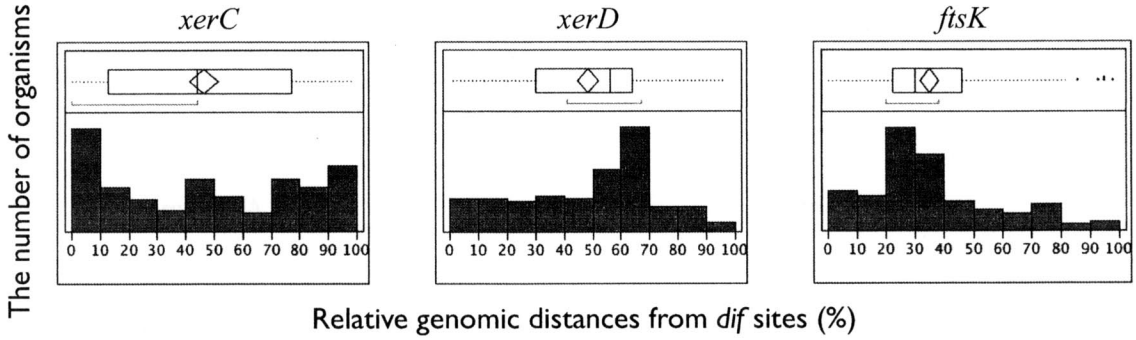

Figure 2.4: Distribution of the genomic distances of *xerC*, *xerD* and *ftsK* gene from predicted *dif* sites. These histograms represent the distributions of relative genomic distances of *xerC*, *xerD* and *ftsK* genes from *dif* sites. X-axis represents the normalized genomic distance (%), and y-axis represents the number of organisms in each group. The boxplots represent the variance.

To further investigate whether replication terminates at the *dif* site, by observing the overall contribution of the genomic selection/mutation pressures of the replication machinery to the collinearity of the *dif* sequence positions and GC skew shift-points, we plotted the distances between them against the GC skew index (GCSI) of genomes to quantify the degree of replicational mutation/selection pressures. GCSI is an index that quantifies the degree of GC skew of a given genome, which can be used as a comparative measure of the accumulated replicational mutation/selection pressures (Arakawa and Tomita 2007a). Since the strength of the GC skew is speculated to partly correlate with the growth rate of bacteria (Worning *et al.* 2006), high replication mutation/selection rate indicated by GCSI implies a greater number of replication events in these organisms. Therefore, if the replication terminates at or around the *dif* site, even allowing for statistical fluctuations, we can assume that the increasing number of replication events should shape GC skew shift-points closer to the *dif* site by the central limit theorem and by the law of large numbers. Hence, genomes with higher GCSI should have closer relative distance between the GC skew shift-points and *dif* sites, if replication terminates at the *dif* site. However, as depicted in Figure 2.5B, we observed no correlation between these two variables (Spearman's rank-correlation coefficients in Proteobacteria and Firmicutes: $\rho = -0.046$ and 0.112, respectively).

2.4 Discussion

In this study, we first demonstrated that the conservation of XerCD genes follows phylogenetic conservation patterns that are specific to each bacterial phylum (Figure 2.2). Based on this principle, we comprehensively predicted the *dif* sequences in hundreds of completely sequenced genomes using a recursive strategy that iteratively models and predicts these sequences using profile hidden Markov models. As a result, we obtained unique candidate *dif* sequences in 715 chromosomes in 641 strains that were validated through multiple means, resulting in the largest collection of predicted *dif* sequences assembled to date. In comparison to previous work by Carnoy and Roten, which predicted *dif* sequences in 228 genomes, our predictions coincided with their results in 208 genomes and we added 507 genomes, including *Aromatoleum aromaticum* str. EbN1, which Carnoy and Roten reported to lack the *dif*/XerCD system. Excluding strains or chromosomes we could not predict, namely, *A. tumefaciens* str. C58, *Burkholderia* sp. 383 chromosome I, II, *D. psychrophila* LSv54, *N. multiformis* ATCC 25196, *P. denitrificans* PD1222 chromosome I, II and *S. denitrificans* DSM 1251, the predicted *dif* sequences in this study differed in 12 chromosomes in comparison to the results of Carnoy and Roten: *C. crescentus* CB15, *Granulibacter bethesdensis* CGDNIH1, *Pseudoalteromonas haloplanktis* TAC125 chromosome II, *Ralstonia eu-*

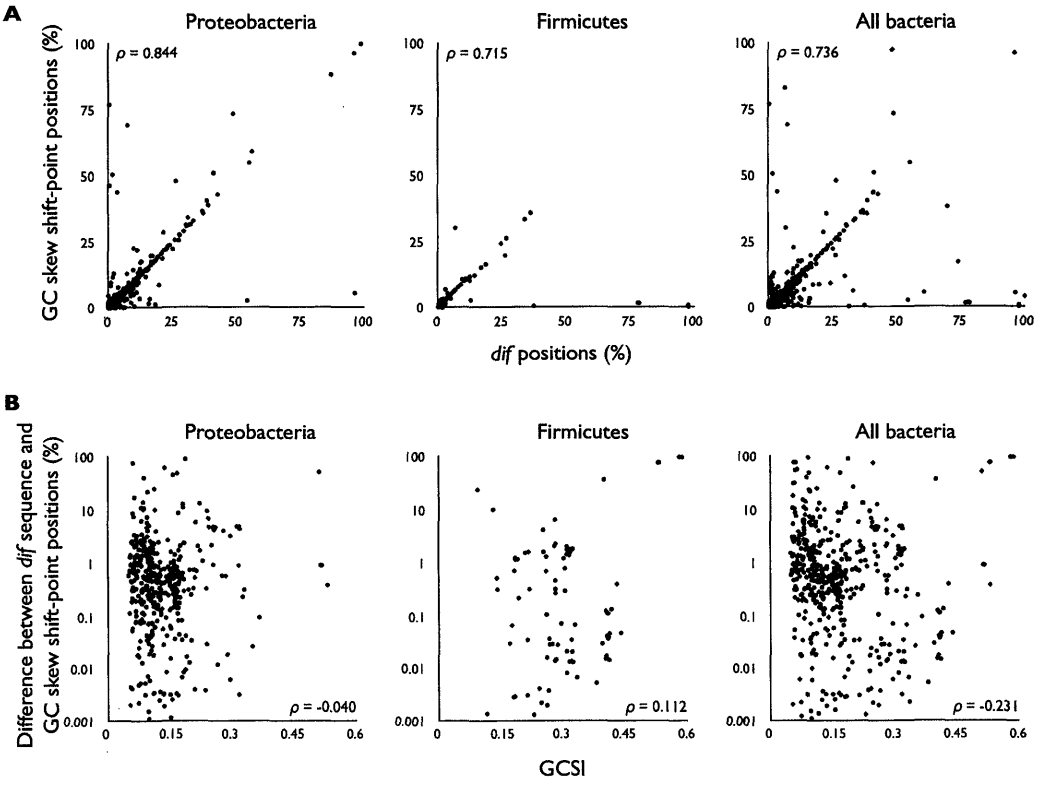

Figure 2.5: The relationship between *dif* sites and GC skew.
A: Correlation of the GC skew shift-point (corresponding to the replication terminus region, y-axis) and the locations of *dif* sequences (x-axis) for genomes with predicted *dif* sequences. Genomes with no visible GC skew, as indicated by GC skew index (GCSI) ≤ 0.05, are omitted. Both axes are shown as the relative distance in percentage of half of the genome size (replichore size), from the position directly opposite of the replication origin. For example, 0% means that the position is directly opposite of the replication origin identified by the GC skew shift-point, and 100% means that it is at the replication origin. In other words, the higher the percentage, the closer the distance to the replication origin. Here the positions of GC skew shift-points and *dif* sites are strongly correlated in all three phyla. B: Lack of correlation between the difference in the positions of GC skew shift-points and *dif* sites (y-axis) and the GCSI (x-axis). GCSI is a quantitative measure of the degree of GC skew, where GCSI = 0 is no observable skew, and GCSI = 1 is extremely pronounced skew. Typically GC skew is visible at GCSI ≥ 0.1, and it is pronounced when GCSI ≥ 0.3. Since we see no correlation in these plots, stronger replication-related mutation bias (*i.e.* larger GCSI) does not necessarily result in closer positions of the GC skew shift-point and the *dif* site. These results suggest that the replication termination occurs near the *dif* site, but not at the *dif* site. The number of *dif* sites is 517 in all bacteria, 438 in Proteobacteria and 97 in Firmicutes. The ρ in this figure is Spearman's rank-correlation coefficient.

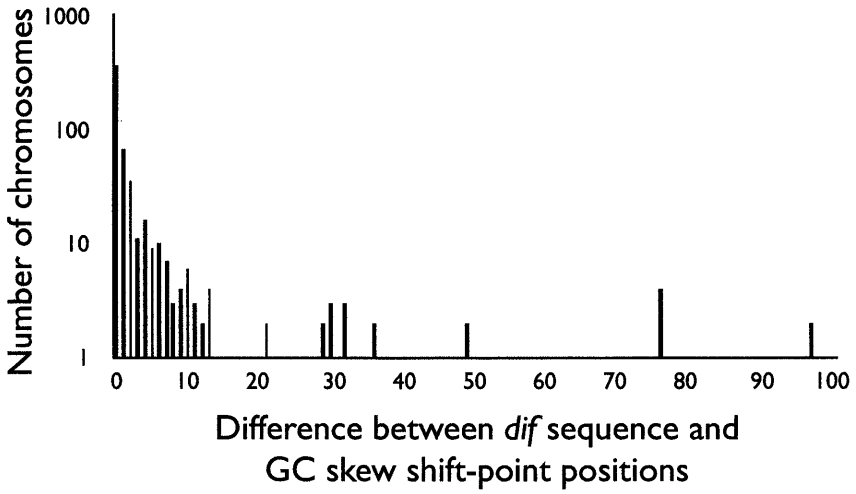

Figure 2.6: The difference between *dif* and GC skew shift-point positions.
The frequency distribution is in a logarithmic scale of the number of chromosomes grouped by the difference between the positions of the *dif* sequence and the GC skew shift-point.

tropha H16 chromosome II, *Rhodobacter sphaeroides* 2.4.1 chromosome I, *R. sphaeroides* 2.4.1 chromosome II, *Rickettsia bellii* OSU 85-389, *R. conorii*, *R. felis* URRWXCal2, *R. prowazekii*, *R. typhi* Wilmington, and *Shewanella* sp. ANA-3. For *R. eutropha* H16 chromosome II and *P. haloplanktis* TAC125 chromosome II, both studies predicted positions that were symmetric from the origin of replication, and although experimental confirmation is required to confirm which candidates function *in vivo*, the palindromic structures of the XerCD binding sites are more conserved in the candidates predicted by our method. Therefore, overall, our results were identical with those of Carnoy and Roten for 92% of the genome analyzed (208/228), and 11/12 mismatch resulted in candidates with more conserved XerCD binding sites, with the addition of 507 genomes among numerous phyla. Carnoy and Roten noted that some *Vibrio* species contain two *dif* sites both located at the vicinity of the GC skew shift-points. Therefore, we further tested whether the predicted *dif* sites in multiple chromosomes are all located near the GC skew shift-points. Using 5% genomic distance as a threshold, 45 out of 54 strains with two chromosomes, including *Vibrio* species, and 6 out of 9 strains with three chromosomes showed such agreement of the positions.

There are four factors that may explain the advantages of our results. First, the selection of bacterial strains in the study by Carnoy and Roten was limited to genomes harboring XerCD that were identified by their similarity to those of *E. coli*, whereas we used all genomes with XerCD orthologs as identified by the KEGG Orthology database. While there is a little time-delay until the sequences are annotated and incorporated into the KEGG Orthology database, use of this database provides a more generic and comprehensive starting point. Second, similarity searches using software tools such as BLAST are not suitable for short sequence motifs that undergo mutation, and the difficulty in identifying only those *dif* sequences with sequence similarity has been shown for *C. crescentus* (Jensen 2006) and several classes of Proteobacteria (Val *et al.* 2008). Third, *dif* sequences require two binding motifs of XerC and XerD to be functional (Hayes and Sherratt 1997); therefore, the conservation of palindromic structure at the 7-12 bp and 17-22 bp positions should be confirmed for each predicted candidate. Finally, the use of iterated HMM allowed *dif* sequence prediction using the profiles of closely related species for each iteration, following the phylogenetic conservation pattern of XerCD.

The high predictability shown in this study suggests that the *dif*/XerCD system of chromosome dimer resolution is highly conserved among bacterial species and that *dif* sequences are almost always conserved when XerCD is present within the genome. In fact, according to the KEGG Orthology database, XerC and XerD are conserved in approximately 60-70% of bacterial species, which is a higher percentage than is found for the replication termination protein Tus (Kobayashi *et al.* 1989) and for universal genes such as the SOS response repressor LexA (Winterling *et al.* 1997). In light of the remarkable conservation of the *dif*/XerCD system, although it is beyond the scope of this study, explorations of alternative CDR machinery in species that lack the *dif*/XerCD machinery would be an interesting area of future research. Chromosome dimer resolution pathways are suggested to be present in species that lack the *dif*/XerCD system, and several alternative pathways have been reported and suggested. Le Bourgeois *et al.* reported an unconventional CDR pathway involving only one recombinase (XerS) in *Streptococci* and *Lactococci*, along with a 31 bp *dif* sequence (Le Bourgeois *et al.* 2007). Similarly, through computational analysis, Carnoy and Roten suggested the existence of another pathway, termed XerH, in ϵ-Proteobacteria in place of XerCD and XerS and discussed the likelihood of the existence of *dif* analogues in these species (Carnoy and Roten 2009; Nolivos *et al.* 2010). The basic strategy of iterated HMM should be applicable in predicting *dif* analogues in these species when defined seed sequences and detailed positions of recombinase binding sites are elucidated.

Although we limited our analysis to strains containing XerCD orthologs, our predictions failed in several species. In Proteobacteria, we could not identify *dif* sequences in five organisms and seven chromosomes, including species with single chromosomes (*Nitratiruptor* sp. SB155-2 and *S. denitrificans* DSM 1251) that are ϵ-Proteobacteria, where an alternative CDR mechanism involving XerH is suggested (Carnoy and Roten 2009), and species with multiple chromosomes (*P. denitrificans* PD1222 chromosome I, *P. denitrificans* PD1222 chromosome II, *B. phytofirmans* PsJN chromosome I, and *Burkholderia* sp. 383 chromosome I and III). Among these, *B.*

phytofirmans PsJN and *Burkholderia* sp. 383 contained *dif* sequences in other chromosomes, indicating that the *dif*/XerCD system is conserved in these strains. Similarly, in Firmicutes, we could not determine *dif* sequence in *L. helveticus* DPC 4571, *C. perfringens* str. 13 or *C. beijerinckii* NCIMB 8052. Among these strains, *L. helveticus* DPC 4571 has an alternative CDR recombinase XerS in its genome, indicating that the *dif*/XerCD system may not be functional. This is an intriguing example of possible evolutionary intermediate with the co-existence of two systems, presumably resulting from a horizontal gene transfer event. While we are unable to find a *dif* sequence corresponding to the XerS machinery, *xerS* gene in this species is located close to the GC skew shift-point (*xerC*: 1,031,814 bp, *xerD*: 1,055,574 bp, *xerS*: 1,228,715 bp, and GC skew shift-point: 1,225,733 bp), which is indicative of its functionality as shown in previous works (Le Bourgeois *et al.* 2007; Carnoy and Roten 2009; Nolivos *et al.* 2010). *C. perfringens* str. 13 and *C. beijerinckii* exhibit highly biased GC contents (28.57% and 29.86%, respectively), and hidden Markov profiling of AT-rich *dif* sequences may have failed due to the background AT-richness of the genome. Comparative studies of *dif*/XerCD systems using close relatives of these genomes may provide evolutionary clues regarding the acquisition and loss of CDR machinery. For example, mapping the types of CDR machinery to the phylogenetic tree of ϵ-Proteobacteria obtained using 16S rRNA sequences with the dnaml program in the PHYLIP package (Felsenstein 1989) shows that a XerH type of CDR machinery may have diverged at an early stage within this phylum. The XerCD type of CDR seems to be absent in the *Campylobacter* and *Helicobacter* genera, except for *Helicobacter hepaticus*, which suggests the existence of the XerH type of CDR in the common ancestor of these species (Figure 2.7). The *dif* candidate in *H. hepaticus* was predicted with iterated HMM only marginally above the threshold, with a score of 10.2 and an e-value of 5.5×10^{-5}. Further analysis is required to identify whether this species actually contains *dif*/XerCD or XerH-type machinery.

Predictions failed in all species belonging to the phylum Cyanobacteria. Although XerCD is present in these species, the sequence similarity distance of XerCD in Cyanobacteria to those of other phyla was high (average of 0.358 ± 0.0159, N = 540), with a minimum distance of 0.322 to *Actinosynnema mirum* (Actinobacteria), which exceeded the 0.3 threshold that was shown in Figure 2.1. Therefore, this divergence of XerCD in Cyanobacteria from those of other phyla implies low applicability of the iterated HMM approach, which utilizes the phylogenetic conservation pattern of XerCD. One possible explanation for the prediction failure in this phylum is that the *dif* sequences and XerCD are highly divergent in Cyanobacteria, preventing their identification with sequence profiles. The replication origin in Cyanobacteria is yet to be identified, and GC skew is weak in these species, implying low degree of replicational mutation/selection pressures, which could also be a reason for the failure of prediction in these species.

Predicted *dif* sequences largely existed in non-coding regions (93.92%). More than half of these coding regions that contained *dif* sequences were hypothetical, with no functional annotation. Furthermore, we found two *dif* sequences included in phage ORF in *Vibrio* and *Xanthomonas*. While these sequences may be integrated with the phages by their hijacking of the host recombination machinery, these sequences are speculated to be the functional *dif* sites, due to 1: their unique occurrence within the genome opposite of the replication origin, and 2: their similarity as identified by our phylogenetic modeling approach. As previously shown in Proteobacteria (Carnoy and Roten 2009), the XerC binding site is more variable and the XerD binding site is more conserved in all phyla (Figure 2.8), both for genomes with single chromosomes and for those with multiple chromosomes, presumably due to the interaction between XerD and FtsK for the initiation of first strand exchange (Aussel *et al.* 2002). The *dif* sequences in α-Proteobacteria with single chromosomes showed higher variation compared to these of other classes and phyla, but this variation was correlated with variations in genomic GC content (Figure 2.9). These differences between variations partly explain the failure of our prediction in extremely AT-rich genomes, such as those found in *C. perfringens* and *C. beijerinckii*.

Although *dif* sequences are expected to be located near the shift-point of the GC skew, we did not use this feature to predict and validate *dif* sequences with iterated HMM; therefore, using the comprehensively predicted *dif* sequences across numerous phyla, we were able to directly compare the positions of predicted *dif* sequences with

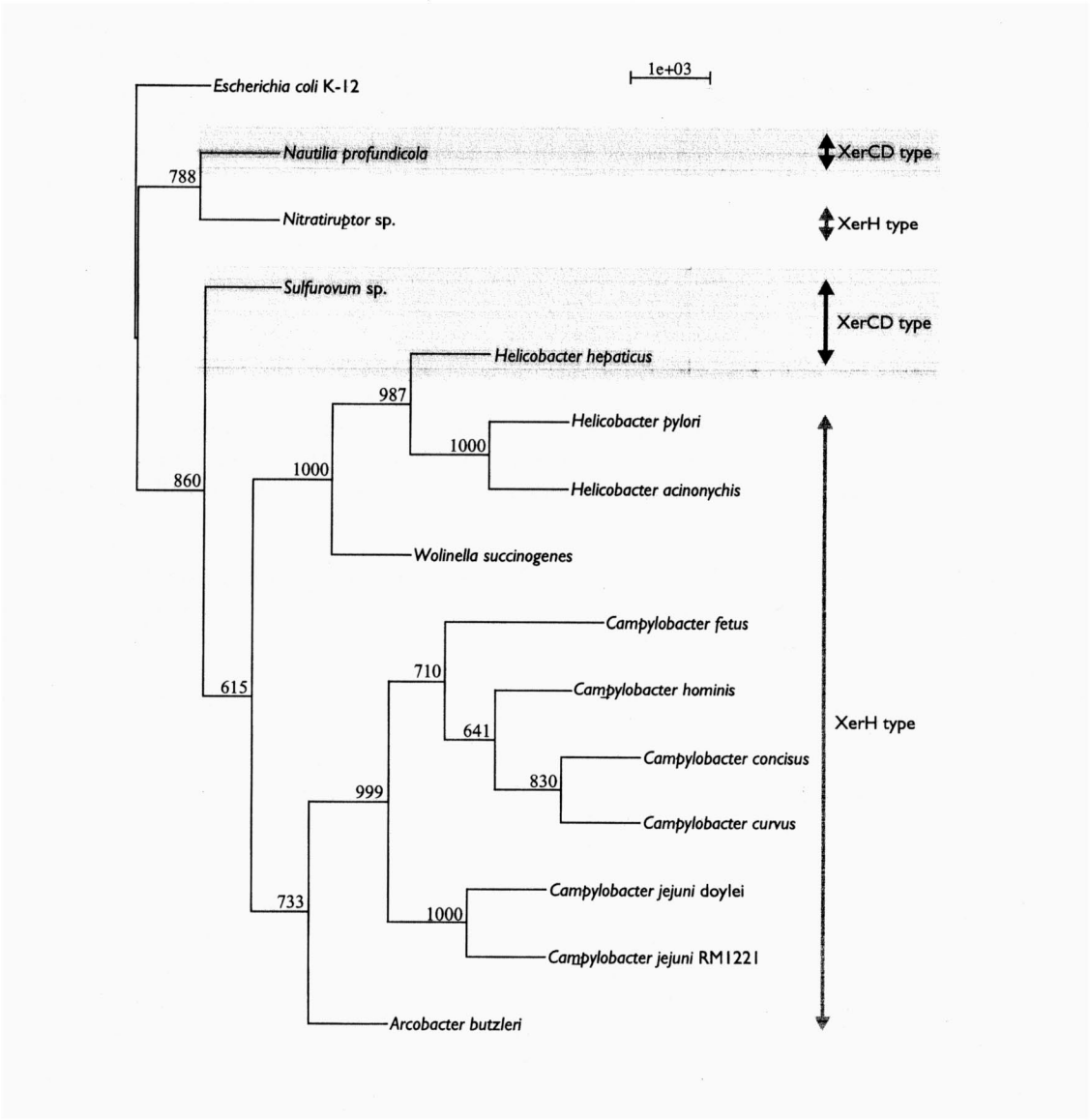

Figure 2.7: Phylogenetic tree based on rRNA for the comparison of XerCD- and XerH-containing genomes.
This phylogenetic tree is constructed using the maximum-likelihood method and is based on 16S rRNAs of 14 organisms in ϵ-Proteobacteria, whose *dif* sequences are predicted in this study. The outgroup is *Escherichia coli* K12.

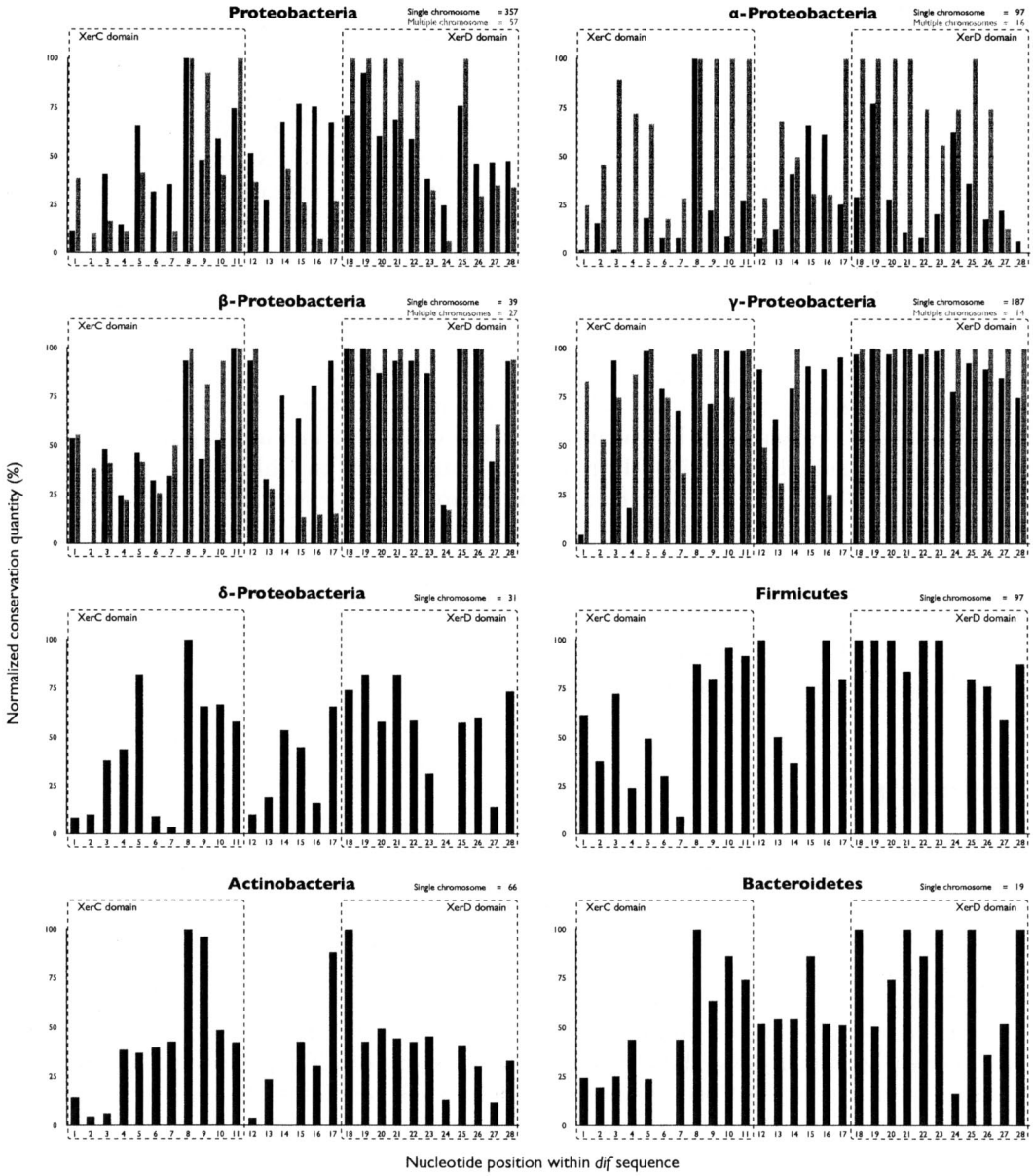

Figure 2.8: The conservation of *dif* sequences.
This figure shows the conservation quantities at each position of *dif* sequence in each phylum or class (Proteobacteria, Firmicutes, Actinobacteria, Bacteroidetes, α-Proteobacteria, β-Proteobacteria, γ-Proteobacteria, and σ-Proteobacteria). The black bars represent the degree of conservation in single-chromosome genomes, and the gray bars represent that of organisms harboring multiple chromosomes. The labels "XerC domain" and "XerD domain" in these graphs represent the binding sites of these proteins. The x-axis represents the nucleotide positions in the *dif* sequence, and the y-axis represents the nucleotide conservation quantity. Y-axis values were normalized to percentages.

those of the GC skew shift-points to analyze their relationships. As expected, these two positions are highly correlated in terms of genomic loci, confirming a previous work (Hendrickson and Lawrence 2007). In this respect, because GC skew is the cumulative result of replicational selection/mutations, the degree of conservation of the CDR machinery is presumably in concordance with the degree of replication selection/mutation pressures (*i.e.* GC skew), which is partly characterized by the difference in the replication machinery and partly characterized by the growth rate (Rocha 2002). On the other hand, as shown in Figure 2.5B, the differences in the positions of the GC skew shift-point and the strength of the GC skew, as quantified by GCSI, were not correlated. If replication termination occurs at the *dif* site, as proposed by Hendrickson and Lawrence (Hendrickson and Lawrence 2007), a stronger GC skew that is generated by a larger number of replication events and/or a higher mutation rate should statistically bring the GC skew shift-point closer to the *dif* site by the central limit theorem and law of large numbers. In fact, the overall correlation of these loci leads to the proposal that the *dif* site is the replication termination point. However, because a stronger degree of replication mutation/selection pressures does not bring these two loci closer to each other, they are not in a causal relationship. Therefore, although the *dif* sequence is located near the replication termination site for efficient CDR, the replication termination site is suggested to be at a site other than the *dif* site, as was recently shown *in vivo* (Duggin and Bell 2009). On the other hand, the *dif* sequences in Firmicutes are more conserved in various phyla because the profile of Firmicutes was the best suited as the initial profile of iterated HMM in Chlorobi, Acidobacteria, Gemmatimonadetes, Nitrospirae, Elusimicrobia, Tenericutes, and Spirochaetes, where initial seed sequences were not available, and those in Proteobacteria were more variable, as shown by the requirement to predict by iterated HMM in classes instead of phyla. Tus proteins, which are shown to terminate replication *in vivo*, are more conserved in Proteobacteria and are not widely conserved in other, partly supporting the possible change in replication termination mechanism by a relatively recent takeover by the Ter-Tus system (Duggin *et al.* 2008). On the other hand, to the best of our knowledge, Tus analogues have not been comprehensively searched in other phyla, and therefore further analysis is required in order to fully support this hypothesis.

2.5 Conclusion

By taking the phylogenetic iterated HMM approach and validating predicted candidates through a combination of HMMER score thresholds, conservation of palindromic structure, and cross-validation, we achieved a comprehensive identification of unique *dif* candidates in hundreds of genomes. As the result, we obtained unique candidate *dif* sequences in 715 chromosomes in 641 strains that were validated through multiple means, resulting in the largest collection of predicted *dif* sequences assembled to date. All of the predicted *dif* sequences described in this study, as well as visualizations of *dif* locations on circular genome maps, are freely available in an online database at http://www.g-language.org/data/repter/. The locations of *dif* sequences can be useful for studies of the regions surrounding the replication terminus, for phylogenetic studies of the replication termination and chromosome dimer resolution mechanisms, and can serve as supporting evidence for GC skew analyses.

Furthermore, we compared the positions of predicted *dif* sequences with those of the GC skew shift-points to understand the relationship between *dif* sequence and replication terminus using GCSI. As the result, although these two positions were highly correlated in terms of genomic loci, the differences in the positions of the GC skew shift-point and the GCSI were not correlated. Therefore, despite the *dif* sequence is located near the replication termination site for efficient CDR, the replication termination site is suggested to be at a site other than the *dif* site.

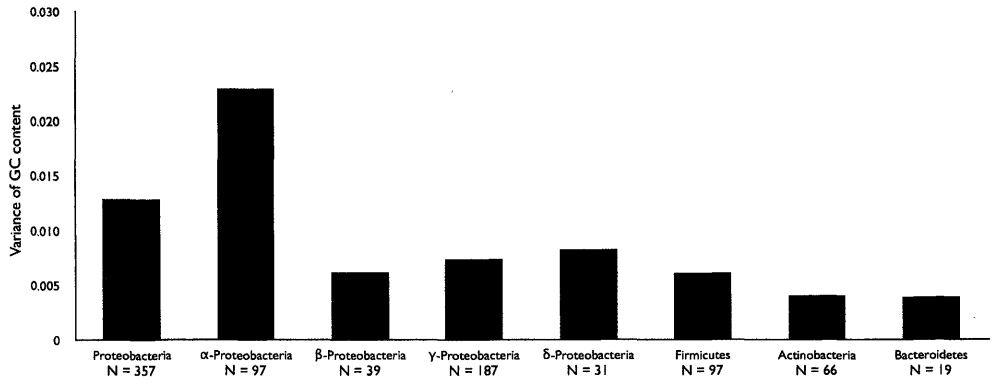

Figure 2.9: Variance of GC content distribution.
This graph shows the GC content variance of organisms in each phylum or class (Proteobacteria, Firmicutes, Actinobacteria, Bacteroidetes, α-Proteobacteria, β-Proteobacteria, γ-Proteobacteria, and σ-Proteobacteria). The x-axis represents the names of the phyla or classes and the number of included organisms. The y-axis represents the variance of GC content.

CHAPTER 3

Validation of bacterial replication termination models using simulation of genomic mutations

"Mind fully to command more than a small specialized portion of it."

-Erwin Schrödinger 'What is Life?'

3.1 Introduction

A circular bacterial chromosome has both a replication origin and a terminus, and replication of the chromosome proceeds bi-directionally from the origin to the terminus (Prescott and Kuempel 1972; Hirose *et al.* 1983; Schaper and Messer 1995; Schaeffer *et al.* 2005). Although the replication termination mechanism is not as well studied as replication initiation (see Scholefield *et al.* 2011 for review), extensive studies have yielded insight into replication termination in organisms such as *Escherichia coli* and *Bacillus subtilis*. The collision of two opposing replication forks at a region approximately opposite the origin was initially suggested to be the predominant mechanism of termination in these organisms (Edlund *et al.* 1976); however, the finding that moving the replication origin does not change the replication terminus in *E. coli* (Kuempel *et al.* 1977; Louarn *et al.* 1977) led to the identification of a fork-trapping mechanism involving the 36 kDa Tus protein in *E. coli* (Mulcair *et al.* 2006), and the 14.5 kDa RTP protein in *B. subtilis*, bound to Ter elements (Sahoo *et al.* 1995; Wilce *et al.* 2001). Tus or RTP protein binds to the Ter sites (in *E. coli*, at the sequence 5'-AGNATGTTGTAAYKAA-3': Coskun-Ari and Hill 1997; in *B. subtilis*, at 5'-KMACTAANWNNWCTATGTACYAAATNTTC- 3': Wake 1997) and forms a barrier called a fork-trap (Horiuchi *et al.* 1995; Labib and Hodgson 2007). This fork-trap acts as an antihelicase and allows forks to enter but not exit the terminus region (Hill *et al.* 1987; Hill 1992). As a result, this complex makes the replication fork stall at the Ter site (Kamada *et al.* 1996; Wake and King 1997). In *E. coli*, most Ter sites are located in the terminus half of the genome (Neylon *et al.* 2005; Mulcair *et al.* 2006).

The *B. subtilis* RTP protein differs from the *E. coli* Tus protein in both sequence and structure, and these systems are not broadly conserved except in species closely related to *E. coli* or *B. subtilis*. These observations suggest a relatively recent introduction of the fork-trap termination mechanism (Duggin *et al.* 2008). Wang and coauthors recently constructed a stain of *E. coli* harboring two origins such that one termination occurred at a Ter site, whereas another terminated speculatively through fork-collision (Wang *et al.* 2011). Similarly, theta-replicating plasmids without fork-trap machinery may terminate by fork-collision; hence, the fork-collision model remains a plausible mechanism for replication termination, especially for species without Tus/RTP analogues.

The bi-directional replication machinery of circular bacterial chromosomes subdivides the genome into two replicating arms, or replichores, with the leading and lagging strands on opposite strands of the DNA duplex. These two replichores experience asymmetric replication-related mutation pressures due to continuous and discontinuous strand synthesis in the leading and lagging strands that results in an excess of G over C in the leading strand (Lobry 1996a; Lobry and Sueoka 2002). Such strand compositional asymmetry is typically visualized using a GC skew plot with moving windows along the genomic sequence. GC skew is calculated as $(C - G)/(C + G)$, and therefore, its polarity shifts near the replication origin and near the terminus, where the leading and lagging strands switch roles (Lobry 1996a; Lobry 1996b; Grigoriev 1998). The cause for this mutational shift from C to G in the leading strand is likely to be multifactorial, and it is still debated (Rocha *et al.* 2006) with several hypotheses having been proposed to date (see details: Francino and Ochman 1997; Reyes *et al.* 1998; Frank and Lobry 1999; Lobry and Sueoka 2002; Rocha *et al.* 2006; Rocha 2008).

The most widely accepted hypothesis is that cytosine deamination occurs in the single stranded DNA (ssDNA), resulting in a decrease in C in ssDNA (Reyes *et al.* 1998; Frank and Lobry 1999) because the leading strand exists as ssDNA for a longer time during the replication of the Okazaki fragments in order to serve as lagging strand template (Marians 1992). Another mutation mechanism that has been proposed is asymmetric transcription-coupled repair (Francino *et al.* 1996), which is based on the strand-specific positioning of transcriptionally active genes (Hanawalt 1991) and their asymmetric distributions (McLean *et al.* 1998). Nevertheless, strand compositional asymmetry, a type of "footprint" of replication-related mutations, is commonly utilized for *in silico* predictions of the replication origin and terminus (Frank and Lobry 2000; Worning *et al.* 2006; Touchon and Rocha 2008). Whereas the GC skew shift-point accurately correlates with the origin of replication in

most bacterial genomes (Arakawa *et al.* 2007; Gao and Zhang 2007), the terminus shift-points are often closer to the chromosome dimer resolution (CDR) site *dif* than to the Ter sites (Hendrickson and Lawrence 2007; Higgins 2007). The 28 bp *dif* sequences are widely conserved in bacteria (Val *et al.* 2008; Carnoy and Roten 2009; Kono *et al.* 2011) and play a central role in CDR as the binding sites of two tyrosine recombinases, XerC and XerD. In the circular bacterial chromosome, when a recombination event occurs an odd number of times in one DNA replication process, the replicated chromosome forms a concatenated dimer that cannot be segregated into two daughter chromosomes (Sherratt 2003; Lesterlin *et al.* 2004). Therefore, many bacteria have the CDR machinery to separate the dimer chromosome via homologous recombination by XerCD into two monomer daughter chromosomes. The *dif* sites are located near the terminus region (Clerget 1991; Blakely *et al.* 1993), but this greater correlation of the *dif* sites with the GC skew remains enigmatic. With their detailed computational study of the skewed oligonucleotides, Hendrickson and Lawrence further confirmed that the skew switch point is closer to the *dif* site than the Ter site. Based on these observations of their "bioinformatically optimized" skew shift-point, they speculated that replication termination is most likely to occur (or had occurred in the course of evolution prior to the introduction of the Ter/Tus system) near the *dif* sites in γ-Proteobacteria, Firmicutes and Actinobacteria, to avoid failure of the CDR system (Hendrickson and Lawrence 2007). In *E. coli*, previous studies clearly show that the replication forks travel through the *dif* site to reach Ter sites *in vivo* (Breier *et al.* 2005; Duggin *et al.* 2008; Duggin and Bell 2009); however, the existence of an unknown replication termination mechanism near the *dif* site remains a possibility in species where the fork-trap associated proteins (Tus or RTP) are not conserved.

We conducted a simulation study to elucidate the relationships between replication termination mechanisms and the genomic compositional bias formed by the replication process. By computationally modeling the above-mentioned replication termination models, namely the fork-trap, fork-collision, and *dif*-stop models, in 65 proteobacterial strains (which have circular genomes, Ter/Tus complexes, and *dif* sites) and in 30 Firmicutes strains (which do not have Ter/Tus complexes), we tested the ability of each model to reconstruct the GC skew graph of existing bacterial genomes. In this chapter, we refer to the GC skew calculated from the published genome sequences as "natural GC skew" to distinguish them from artificially constructed GC skew.

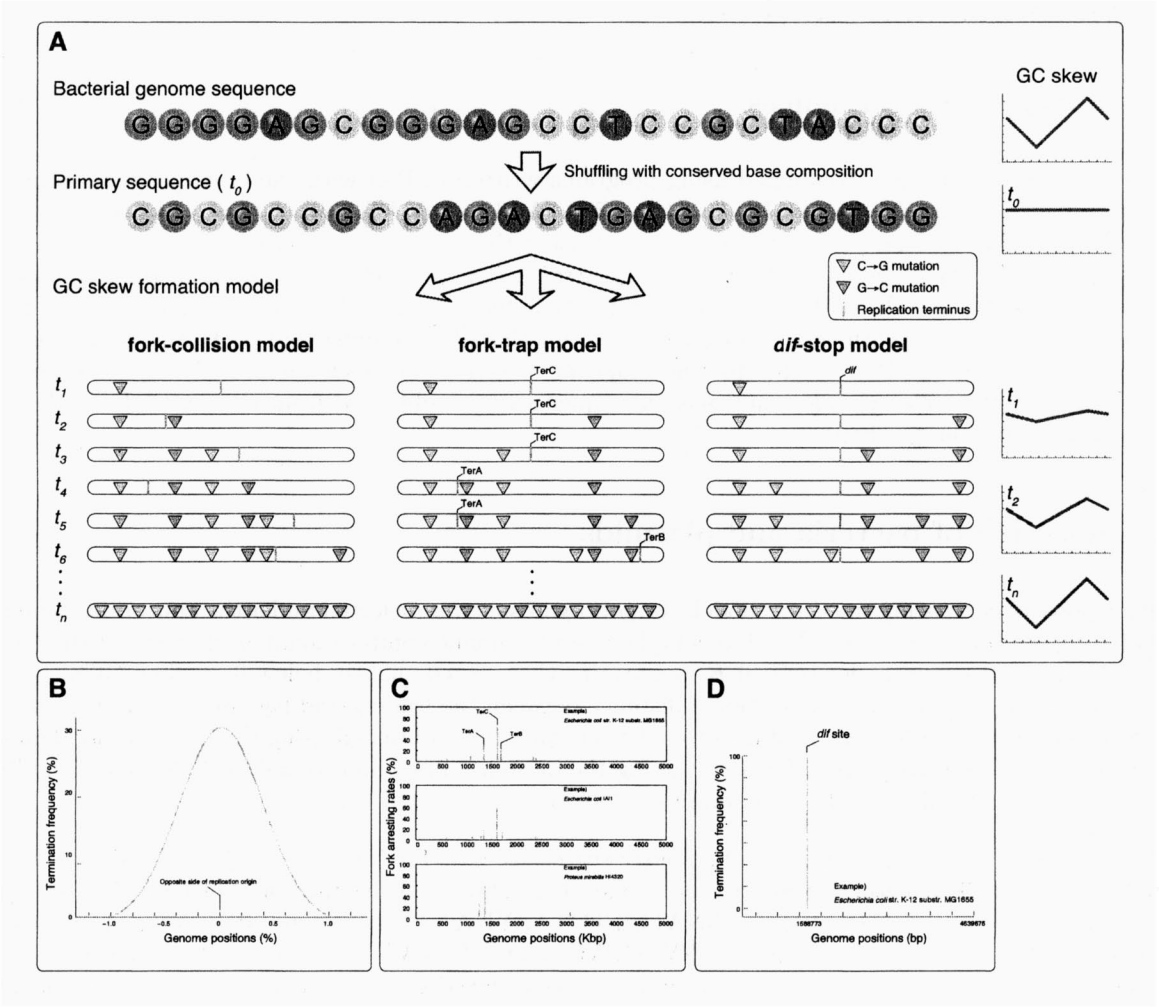

Figure 3.1: Scheme of GC skew reconstruction simulation.

A: A schematic representation of the GC skew reconstruction simulation. The primary sequence was generated based on the shuffled bacterial genome sequence, which had the same base composition as the original sequence. The green and yellow triangles represent the locations of C → G mutations in the leading strand (or G → C in the lagging strand). Graphs on the right show the typical GC skew shape at each simulated time point (t_i). The blue bars represent the replication termini. B: Frequency distribution of replication termini in the fork-collision model. Here, replication terminates near a locus directly opposite the origin, and the position probabilistically fluctuates according to a Gaussian distribution. The distribution was empirically derived from plasmid sequences that are likely to be terminated by fork-collision mechanisms. C: Frequency distribution of replication termini in the fork-trap model in *Escherichia coli* str. K-12 substr. MG1655, *Escherichia coli* IAI1, *Proteus mirabilis* HI4320. Here, replication termination occurs at Ter sites, but different Ter sites have different rates of fork arrest. D: Frequency distribution of replication terminus in the *dif*-stop model in *Escherichia coli* str. K-12 substr. MG1655. Here, all replication terminates at a single finite locus *dif*.

3.2 Materials and Methods

3.2.1 Software and sequences

All analyses in this study were conducted using programs written in Perl with the G-language Genome Analysis Environment, version 1.8.13 (Arakawa *et al.* 2003; Arakawa and Tomita 2006; Arakawa *et al.* 2008). Statistical analysis and visualizations were performed using the R statistics package, version 2.10.0 (www.R-project.org). This study targeted 65 Proteobacteria strains that have circular genomes, Ter sequences, Tus proteins and *dif* sites, as well as 30 Firmicutes strains that have no Ter/RTP homologues. The existence of Ter sequence was confirmed with the "oligomer_search" function of the G-language GAE, and RTP homologues were determined using the KEGG (Kyoto Encyclopedia of Genes and Genomes) Orthology database (KO; Kanehisa *et al.* 2010). The genomic and plasmid sequences were obtained from the NCBI FTP Repository (ftp://www.ncbi.nlm.nih.gov/Ftp).

3.2.2 Selection of bacteria and plasmids

For the purposes of comparing the three models, target organisms were selected under the appropriate conditions for circular chromosomes, *dif* sites, Ter/Tus complexes and genomic compositional asymmetry of the GC skew indexes (GCSIs) ≥ 0.1 (except for several *E. coli* strains that scored slightly below 0.1). The GCSI quantifies the degree of GC skew from the compositional distance between the leading and lagging strands and the extent to which the GC skew graph shape conforms to a discrete sine curve obtained using the Fast Fourier Transform. A threshold of 0.1 is relatively strict for ascertaining the existence of compositional bias (Arakawa and Tomita 2007a; Arakawa and Tomita 2007b; Arakawa *et al.* 2009a).

The Ter and *dif* sites were identified by a homology search using a Ter consensus sequence and by a recursive hidden Markov modeling method, respectively (Kono *et al.* 2011). In bacteria harboring a Tus protein and a replication terminus protein (RTP), 5'-AGNATGTTGTAAYKAA-3' (allows mutations at 1, 4 and 16 bases; Coskun-Ari and Hill 1997) and 5'-KMACTAANWNNWCTATGTACYAAATNTTC-3' (Wake 1997) were used as the Ter consensus sequence. For the set of plasmids used to derive distribution parameters for the fork-collision model, plasmids must have been replicated bi-directionally. Therefore the plasmids with theta replication machinery were selected according to the following criteria: 1) they must be larger than 10 Kbp with sufficient GCSI (window size: 64, spectral amplitude ≥ 1000; Arakawa *et al.* 2009a), 2) they must contain neither Ter nor *dif* sites, 3) they must lack the *repC* gene, which is essential for rolling circle replication (Khan 2000), and 4) no iteron sequences (Haines *et al.* 2006) are located near 5% region from putative replication origin predicted by GC skew shift-point.

3.2.3 Simulation of GC skew formation

The simulation of GC skew formation involves the following steps: 1) shuffling the genome sequence to create an unbiased initial sequence for simulation, while maintaining the same nucleotide composition, 2-a) definition of the leading and lagging strands based on a replication termination model and the position of the replication origin, 2-b) mutation of one random C to a G in the leading strand, 2-c) repeating from 2-a until the maximum simulation cycle is reached, and 3) validation of the simulated GC skew by comparison with the original genome sequence. The shuffled initial sequence was generated with the "shuffleseq" function of the G-language GAE, which is based on the Fisher-Yates algorithm (Fisher and Yates 1948). All simulations used the same randomized

sequences in each organism to avoid errors associated with shuffling. The maximum simulation cycle number was determined by the absolute difference in GC content between the whole genome and the leading strand. The replication origin was defined using the "find_ori_ter" function of the G-language GAE, which is based on the cumulative GC skew (Grigoriev 1998) at 1 bp resolution. The similarities between the simulated and natural GC skews were calculated using the root mean square error (RMSE).

3.2.4 Replication termination models

Four replication termination models were constructed: fork-collision, fork-trap, *dif*-stop, and a control model terminating at the GC skew shift-point, as described in Eqs. 3.1-4. In these equations, X_i represents the replication terminus in bacteria i. In the fork-collision model, the positions of fork-collision were empirically determined to follow a Gaussian distribution based on observations of the GC skew shift-points in plasmids that lack fork-trap machinery and *dif* sites. The mean of this distribution (μ) was a locus directly opposite the replication origin, and the variance was σ^2. Both of these values were normalized by the genome size (Eq. 3.1). The termini in the fork-trap model were defined by the locations of Ter sites in each bacterium, $\{t_1, t_2, t_3, ...t_n\}$, each weighted with certain probabilities (Eq. 3.2). The termini in the *dif*-stop and control models were represented by the constant positions of *dif* sites (C_d) or GC skew shift-points (C_s) (Eq. 3.3, 3.4).

fork-collision model:

$$X_i = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \tag{3.1}$$

fork-trap model:

$$X_i \in \{t_1, t_2, t_3, ..., t_n\} \tag{3.2}$$

dif-stop model:

$$X_i = C_d \tag{3.3}$$

control model:

$$X_i = C_s \tag{3.4}$$

Assuming a simple model where all replication terminates with the fork-trap mechanism and where all replication forks progress continuously without stalling, replication should always terminate at a furthermost Ter site from the origin. In *E. coli*, this is TerC located at 1,607,184 bp, where position directly opposite from origin is at 1,603,784 bp and the *dif* site is at 1,588,773 bp. Second farthest Ter in the other replichore, namely TerA in the right replichore of *E. coli*, is only encountered if replication fork stalls a sufficient time (hereafter referred to as δ) in the right replichore for the replisome in the left replichore to over-travel to reach TerA. Since TerA is located 264,013 bp apart from a site directly opposite from the origin, and since the average speed of replisome is around 1,000 bp/s (Hirose *et al.* 1983), δ in *E. coli* is calculated to be around 5 min. This is in accord with *in vivo* and *in vitro* findings, that stalling by supercoiling tension, protein blocking, and replisome assembly requires around 4-6 min to restart (Possoz *et al.* 2006; McGlynn and Guy 2008; Mirkin and Mirkin 2007). Such long stalling is known to occur *in vivo* in around 20% of replication events (Maisnier-Patin *et al.* 2001). Stalling event should randomly and thus evenly occur in each replichore, and therefore, in *E. coli*, furthermost TerC is first encountered in 80% (without long replisome stalling) +10% (long replisome stalling in the same replichore), and TerA is first encountered in the remaining 10% of replication events. Furthermore, we considered the "leakiness"

rate of each Ter site, which is approximately 80% as observed *in vivo* (Sharma and Hill 1995). As a result, in *E. coli*, given the farthest inverted Ter sites from replication origin are TerC and TerA, followed by TerB, TerD, TerE and etc, pausing rate at each Ter site is TerC = 72%, TerA = 10.5%, TerB = 16%, TerD = 1.152%, TerE = 0.230%. The probability of having long enough stalling time δ so that the second furthest Ter site is utilized (20% in *E. coli*), is different in other species, due to the different distances of second farthest Ter sites from the region directly opposite of the origin. Assuming normal distribution of fork stall durations, this probability is calculated using δ of each species.

To validate these pausing rates, we further determined the optimized pausing ratio that best reconstructs the natural GC skew, by means of parameter tuning. For this parameter tuning, the patterns of the fork arrest ratios in each bacterium were tested in 5% increments, but since the comprehensive parameter searching in a bacterium harboring ten Ter sites requires the calculation of 10,015,005 patterns and is not computationally realistic, the calculated combinations were limited to those having a sum total of fork arrest rates over 80%, with four Ter sites located farthest from the origin, based on *in vivo* observations (Duggin and Bell 2009). As a result, the pausing rates calculated based on the stalling rates and Ter leakiness were very similar with the optimized pausing rates ($\rho = 0.725$, Spearman's rank-correlation coefficient).

3.3 Results

3.3.1 GC skew formation simulation

Because GC skew represents the evolutionary footprint of a replication-related mutational bias, we attempted to elucidate the contributions of different replication termination models by computationally reconstructing the GC skew pattern using simulations of strand-biased mutations. Although the specific substitution types and mechanisms are likely to be multifactorial, compositional replication strand bias, with only few exceptions, is strongest for G > C in the leading strand of prokaryotes (Rocha and Danchin 2001; Rocha *et al.* 2006). Hence, we took the simplest approach to simulating the evolutionary formation of GC skew. We started with shuffled sequence that had no replication strand bias, and we iteratively introduced C → G mutations in the leading strand until the GC compositional bias between the leading and lagging strands was equal to that of existing genomes (Figure 3.1A). The relative amounts of complementary bases should theoretically reach equilibrium when there is no strand bias (Lobry 1995; Sueoka 1995); therefore, replication strand bias should be reconstructed using only the replication-related mutation bias. Our simulation involves three principal sets of variables: 1) the initial sequence with no strand bias, 2) the number of simulated mutations (simulation cycles), and 3) the locations of the replication origins and termini. Although many prokaryotic genomes exhibit significant replication strand bias, the relative amounts of complementary bases are close to equilibrium across the entire genomic sequence. We generated an artificial genomic sequence with no replication strand bias by shuffling the observed sequence while maintaining its overall composition. The number of simulated mutations, or the number of simulation cycles, was determined as the absolute difference between the number of G and C bases in the leading/lagging strands across the whole genome. For example, given an imaginary genome sequence of 1 Mbp with equal amounts of all four bases, the genomic G or C content would be 250,000 bp each. Because the leading strand of this genome would be biased toward G, the quantities of G and C bases would be 260,000 bp and 240,000 bp, respectively. Here, the absolute difference in G or C content, 10,000, is the number of C → G mutations required to reconstruct the GC skew, and this number also represents the number of simulation cycles. The last of the three sets of variables, the location of the replication terminus, is the most central part of our simulation study. The replication strand bias predominantly causes enrichment of G in the leading strand, but the definitions of the leading and lagging strands change under different replication termination models. This is because the locations of replication termination vary according to the models. For example, the

fork-collision model results in probabilistic termination within the region approximately 180 degrees opposite the origin, whereas the fork-trap model involving the Ter/Tus system terminates at multiple but defined finite locations. Likewise, if replication termination occurs near the *dif* site or near the GC skew shift-point, the replication terminus becomes a single finite location. In this simulation study, we assess the reproducibility of the GC skew graph using varying replication termini inferred by different replication termination models.

We first tested the applicability of such simulations using the *E. coli* K-12 genome. In *E. coli*, the numbers of G and C bases in the whole genome were 1,176,923 bp and 1,179,554 bp, respectively, and the numbers of G and C in the leading strand were 1,216,043 bp and 1,140,434 bp, respectively. Therefore, the number of simulation cycles was determined to be 39,120 based on the difference between the two compositions. Shuffled initial sequence with no replication strand bias was generated while maintaining the overall genomic base composition (A: 24.62%, T: 24.59%, G: 25.37% and C: 25.42%). In this first validation, the replication terminus was defined at a finite location at the GC skew shift-point (1,550,412 bp). This was performed to observe whether this simplistic simulation could reconstruct the GC skew graph. The similarities between the artificial GC skew and the natural GC skew graphs were evaluated by root mean square error (RMSE) as well as by the GC skew index (GCSI), which quantifies the degree of GC skew. GC skew is generally visible when GCSI > 0.05 (Arakawa and Tomita 2007a). Although the GC skew shape in the initial sequence ($t = 0$) was almost completely flat and had a high RMSE value (GCSI $= 0.007$ and RMSE $= 6.982$), the GC skew-like shape was gradually formed as the simulation cycles progressed. When the simulation reached 39,120 cycles (the maximum number of iterations), the artificial GC skew shape showed least difference from the natural GC skew as calculated by RMSE (artificial and natural GCSIs were 0.098 and 0.097, respectively, and the RMSE between them was 0.025). The GC skew shapes found after different numbers of simulated cycles ($t = 0$, 10,000, 15,000, 20,000, 25,000 and 39,120) are described in Figure 3.2. Probabilistic errors (or standard deviations) associated with the Monte Carlo simulation procedure used for sequence shuffling and simulating mutations were negligible because the standard deviation was less than 0.0256 (Figure 3.3).

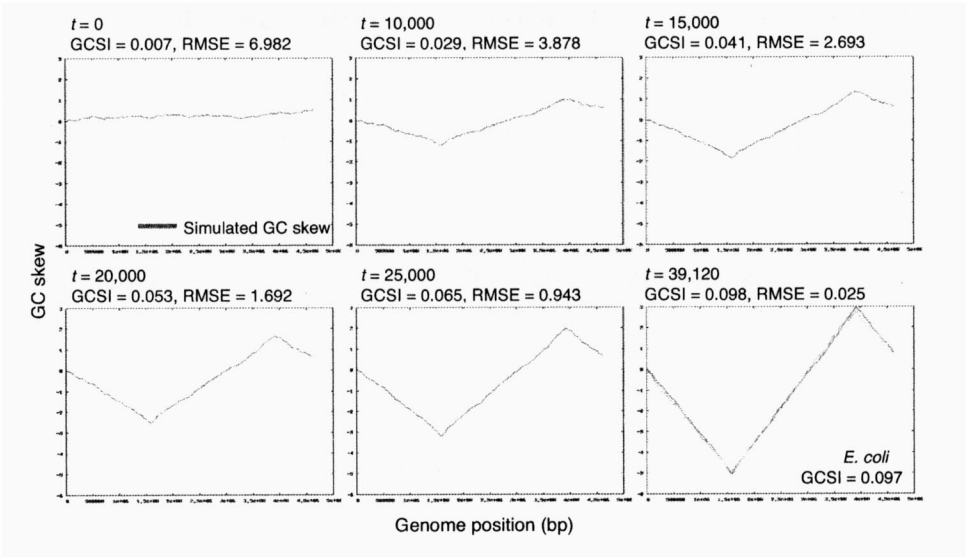

Figure 3.2: Example of GC skew reconstruction simulation.
These figures are simulated GC skews when the simulated cycles (t) were 0, 10,000, 15,000, 20,000, 25,000 and 39,120 (the maximum simulated cycle in *E. coli*). The GCSIs and RMSEs were described in the upper left of each graph. When the simulated cycle reaches 39,120 (the bottom-right corner), red line (simulated GC skew) and green line (natural *E. coli* GC skew) almost completely overlap.

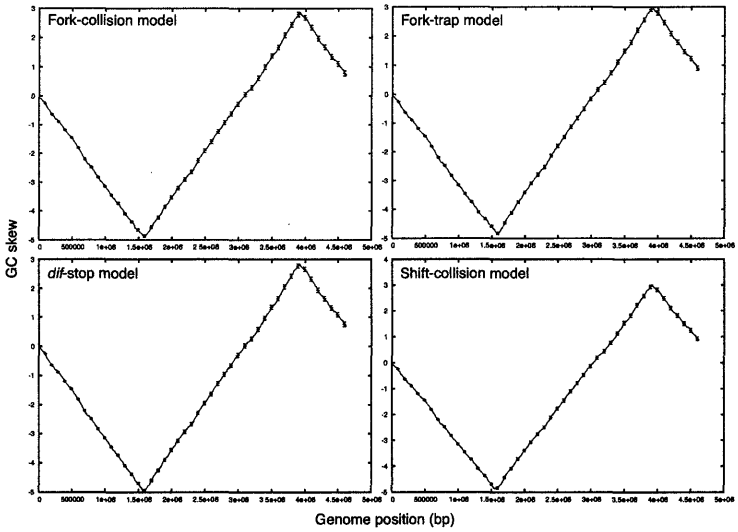

Figure 3.3: Probabilistic error rates.
These figures show the probabilistic simulation error rates in 1,000 iterations. Each error bar represents the standard deviation, with negligible average ≤ 0.0256.

3.3.2 Construction of three replication termination models

Our simulation study involves three replication termination models: fork-collision, fork-trap, and *dif*-stop. As described above, these models define the positions of the leading and lagging strands, and they are mathematically modeled based on the existing knowledge of replication termination, with parameters empirically derived from genomic data.

In the fork-collision model, replication terminates when the two opposite replication forks meet by chance at the far end of the circular chromosome. Because the collision occurs randomly, the termination positions should follow a probabilistic distribution. We derived the distribution by observing the positions of the GC skew shift-points in replicons that are highly likely to be terminating solely by fork-collision: namely, plasmids that have been replicated bi-directionally with theta replication machinery and lack the Ter/Tus complex (for fork-trap model) and the *dif*/XerCD system (for *dif*-stop model). Using 98 plasmids, the distribution was fit to a Gaussian distribution ($p = 0.295$ by Kolmogorov-Smirnov test) centered close to the part of the genome opposite from the origin (Figure 3.1B). The distribution was thus derived and normalized to the genome size (see Materials and Methods for detailed parameters), which was used to define the termination position in each simulated cycle of the fork-collision model.

In the fork-trap model, replication terminates specifically at Ter sites (the sites where Tus proteins bind), but each Ter site individually allows a certain fraction of the incoming replication forks to pass with different rates. We therefore needed to obtain the probabilistic ratio of fork-trapping at each Ter site. Based on the time and probability of accidental stalling of replication forks at sites other than Ter, on the positional relationship among different Ter sites, and on the leakiness of each Ter site (see Materials and Methods), we could calculate the frequency distribution and computationally determined the fork-trap rates at each Ter site (Figure 3.1C).

Unlike the two models described above, the *dif*-stop model involves predictable termination at a single finite position without any probabilistic fluctuations. We sought to determine the exact positions of the *dif* sites in bacterial genomes using computational predictions (Figure 3.1D). We have previously reported an accurate and comprehensive prediction of *dif* sites in 641 bacterial genomes using a recursive hidden Markov model method (Kono *et al.* 2011), and all positions of *dif* sites used in this work were obtained from the database accompanying that previous study (http://www.g-language.org/data/repter/). Similarly, as a control, we implemented a model that terminates at the GC skew shift-point instead of at the *dif* site.

3.3.3 Evaluation of the replication termination models

We tested the validity of the aforementioned models with 65 proteobacterial genomes, including those of *E. coli* strains and others that have circular chromosomes, Ter/Tus systems, *dif* sites and XerCD homologues as well as a compositional bias of GCSIs ≥ 0.1. Typical examples of the simulated GC skew graphs are provided in Figure 3.4 (all simulation results in target organisms are shown in http://web.sfc.keio.ac.jp/~ciconia/AppendixFigure1.zip). Whereas there was no significant difference between the *dif*-stop and fork-collision models ($p = 0.069$, Wilcoxon test), the fork-trap model showed significant differences from other models (*dif*-stop model and fork-collision model, $p = 0.011$ and $p = 0.007$, respectively, Wilcoxon test; Figure 3.5). Interestingly, even the control model scored significantly lower than the fork-trap model ($p = 0.022$, Wilcoxon test; Figure 3.5), and the control model, by naive expectation, should best reproduce the GC skew graph because it terminates replication at the GC skew shift-point. Of the three models tested, the fork-trap model seems to best explain the existing GC skew shapes.

According to above result, it was confirmed that the fork-trap model is appropriate to form the GC skew. However, this result was observed under conditions that the termination machinery exists solely. Although one type of termination machinery may be dominant in the existing genomes, other machineries could co-exist at a much lower prevalence. In order to examine the contribution ratio of each model to construct the GC skew, we conducted further evaluations of the replication termination models in a hypothetical combination: probabilistic combination, where the termination models is assumed to coexist under certain probabilistic preferences (Figure 3.6A).

To determine the time of appearance and duration of each type of machinery, we tested all possible combinations using the three models. For computational efficiency, durations of the models were incremented by units of 10% of the total number of simulated cycles, and consequently, 36 patterns were assessed. In this case, none of the different combinations significantly affected the reproducibility of the GC skew (Figure 3.6B). Nevertheless, combination model often resulted in lower RMSE compared to simulations using only one of the three termination models independently.

The best probabilistic combination differed among bacterial species. We extracted patterns that performed well across all of the 65 genomes used in this work, among the 36 probabilistic combinations tested. The best pattern of the probabilistic combinations was 10%-70%-20%, in the order of fork-collision, fork-trap and *dif*-stop models. The probabilistic combination model showed less RMSE values than *dif*-stop and fork-trap models ($p < 0.001$, Wilcoxon test; Figure 3.7 and http://web.sfc.keio.ac.jp/~ciconia/AppendixFigure2.zip).

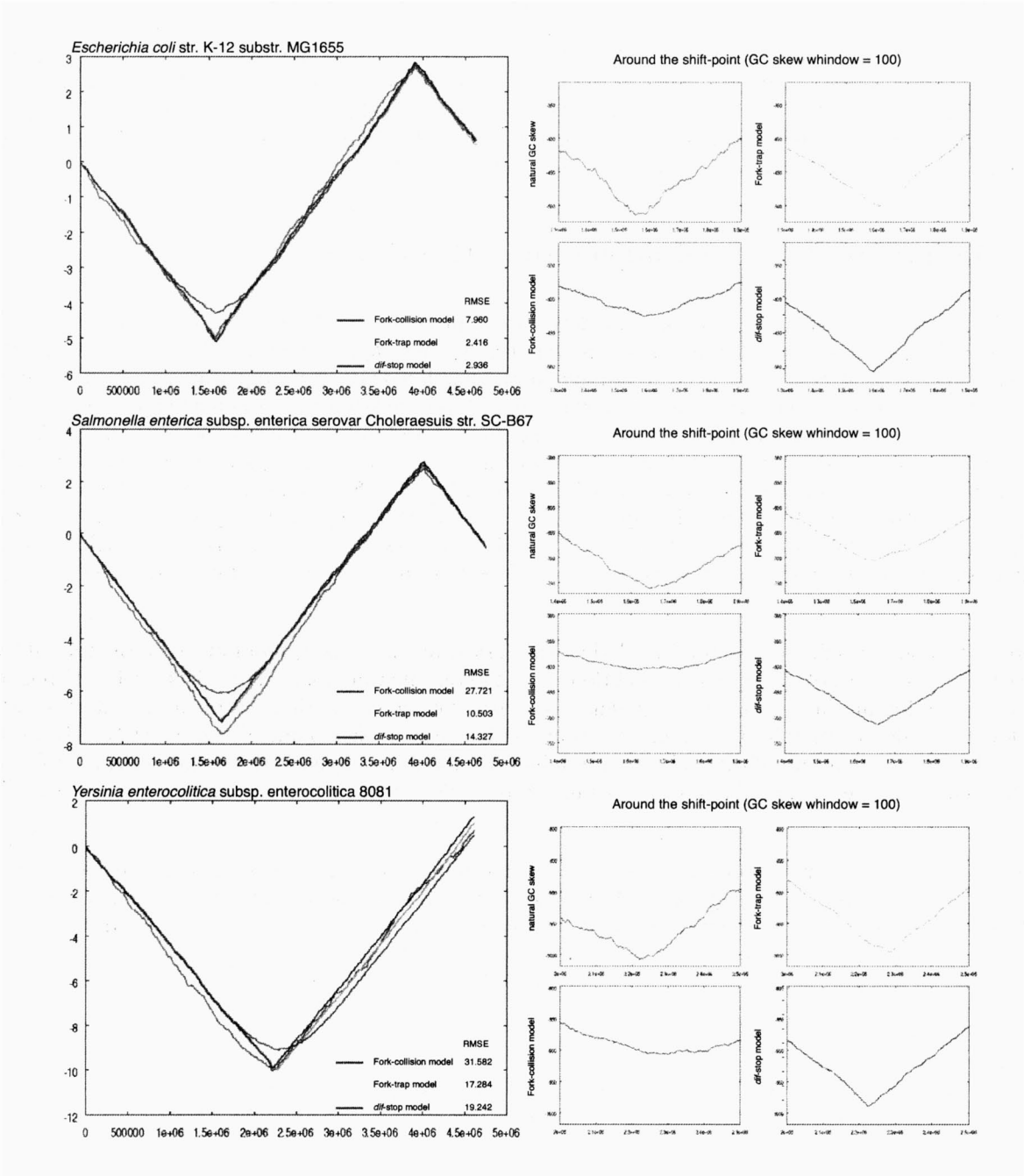

Figure 3.4: Examples of simulated GC skew.

Examples of the overall shapes around the GC skew shift-points (see http://web.sfc.keio.ac.jp/~ciconia/AppendixFigure1.zip for comprehensive results from all organisms used in this work). The left figures show the overall GC skew graph, and close-ups of the regions around the shift-point are shown to the right. In the right set of graphs, red, green, blue and purple lines show the natural GC skew, fork-trap model, fork-collision model and *dif*-stop model, respectively.

3.3.4 Simulations in species lacking fork-trap machinery

Lastly, we conducted the same analysis for species in other phyla to confirm the observed model preferences. For these analyses, we used 30 Firmicutes species that lack Tus and RTP homologues (and therefore are presumed to lack fork-trap machinery). As a result, significant differences were found between the *dif*-stop and probabilistic combination models ($p < 0.001$, Wilcoxon test; Figure 3.8).

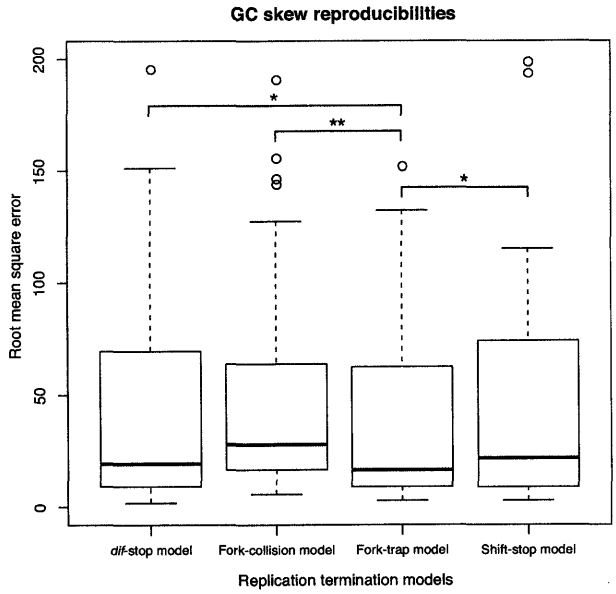

Figure 3.5: Comparison of RMSE scores in four models.
Boxplot of the RMSE scores for four models, representing the similarities between simulated and natural GC skews in the four models (in 65 bacteria). The p-values were calculated by a Wilcoxon test, * $p < 0.05$, ** $p < 0.01$.

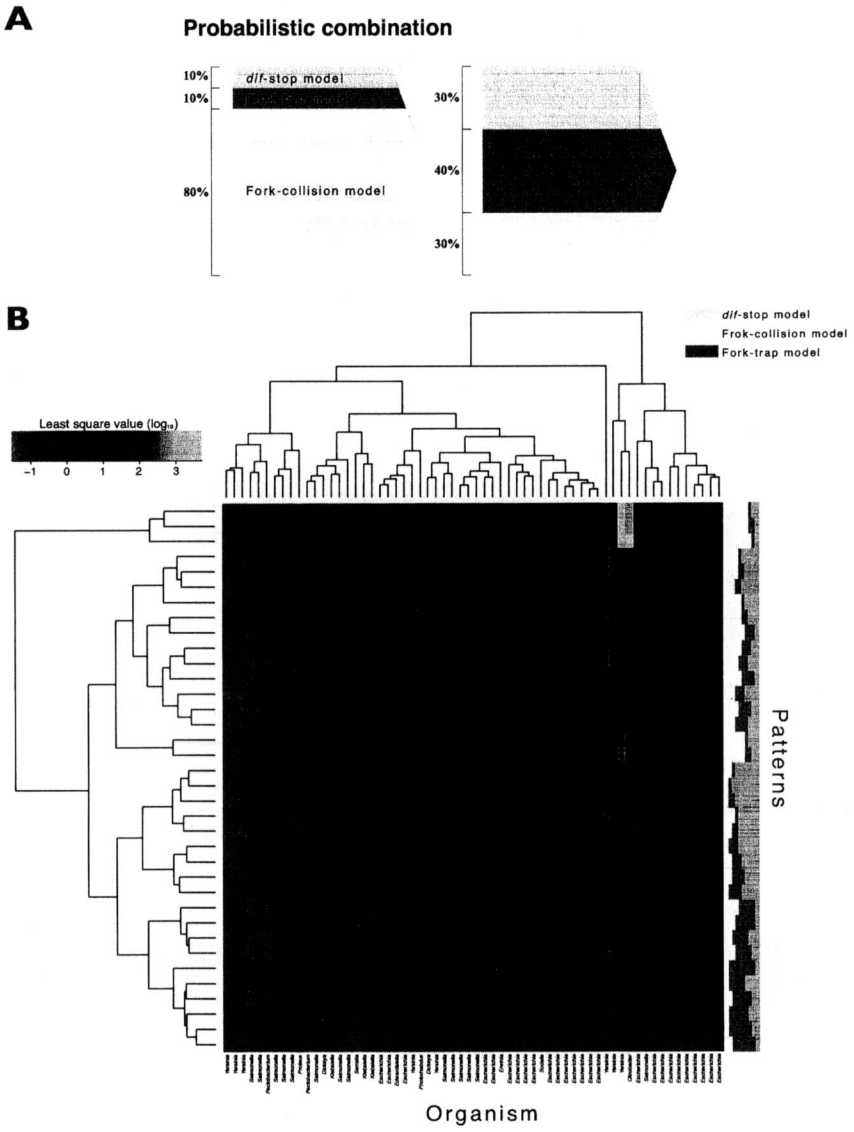

Figure 3.6: Heat map of RMSE scores for probabilistic combination model.
A: The conceptual scheme for probabilistic combination of replication termination models. B: The heat map of RMSE scores. The x-axis represents the 65 organisms, and the y-axis represents the combination patterns. Each color represents one of the three models (blue = *dif*-stop model, yellow = fork-collision model and red = fork-trap model), and the width of colored regions represents their probability. The scales are logarithmic.

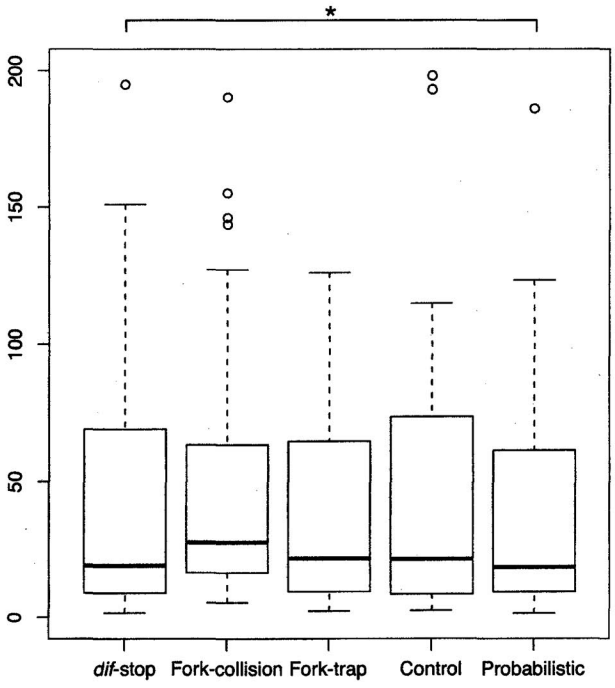

Figure 3.7: Boxplot of RMSE of all simulated models.
The x-axis represents the models (*dif*-stop, fork-collision, fork-trap, and shift-stop (control) models as well as probabilistic combination) and the y-axis represents the RMSE values. * $p < 0.01$, Wilcoxon test.

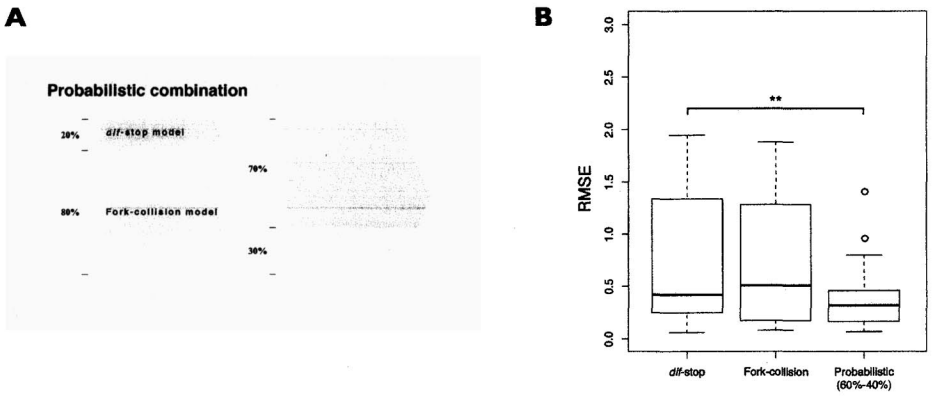

Figure 3.8: Boxplot of RMSE of simulated models in Firmicutes.
A: The conceptual schemes for probabilistic combination of replication termination models. B: The x-axis represents the models (*dif*-stop, fork-collision, and probabilistic combination) and the y-axis represents the RMSE values. * $p < 0.05$, ** $p < 0.001$, Wilcoxon test.

3.4 Discussion

In circular bacterial chromosomes, *in vivo* studies clearly show that replication is terminated by fork-trap mechanisms involving the Ter/Tus system, which impedes fork progression at specific sites. However, the genomic compositional bias shaped by replication-related mutation bias, which is an evolutionary footprint of the replication machinery, has a shift-point of compositional polarity at a site closer to *dif* than Ter. In this study, we took a theoretical approach to elucidate this paradoxical relationship between the replication-related genomic compositional bias and the replication termination mechanism in bacteria. To that end, we conducted a simulation study employing multiple replication termination models. Three main models, namely fork-collision, fork-trap, and *dif*-stop, as well as one control model that assumes replication termination at the GC skew shift-point were tested by computationally reconstructing the GC skew shape in 65 bacteria. Different combinations of these models were also analyzed. Based on the results, the reproducibility of simulated GC skew was highest in the fork-trap and fork-collision models (in comparison to that of original genome sequence). Surprisingly, it was much lower for the *dif*-stop model and the control model. Our result therefore supports previous *in vivo* studies (Duggin and Bell 2009) that favor the fork-trap model as the working replication termination model. Although not intuitively obvious at first sight, the probabilistic usage distributions of the Ter sites better explain the current GC skew shape than the location of the *dif* site.

The simulation method for GC skew reconstruction used in this work was based on the most simplistic approach. The procedure mutates a C to a G in the leading strand for each simulation cycle. We have two justifications for this approach. First, although the specific types and causes of mutations introduced by the replication process are likely to be multifactorial and complex, the resultant compositional bias is predominantly in the direction of C $\rightarrow$ G in most bacteria (Rocha *et al.* 2006), as observed in existing genomes. Second, previous discussions regarding the positioning of Ter, *dif*, and the GC skew shift-point were based on the GC skew graph, which does not contain any information about AT composition. Therefore, we have limited our discussions to the reconstruction of the GC skew graph, which only requires the consideration of C $\rightarrow$ G mutations. However, one other factor that should be considered is the positions of the coding regions. Coding strand bias is as high as approximately 78% in the leading strand in Firmicutes or *Mycoplasma* (Fraser *et al.* 1995; Kunst *et al.* 1997; Rocha 2002), and the GC skew is mostly pronounced only in the third codon positions (McLean *et al.* 1998). On the other hand, the 65 Proteobacteria used in this work have relatively little coding strand bias (averaging 58% in the leading strand), and mutations do not avoid the coding region; they occur all over the genome in these species (Rocha and Danchin 2001). In this work, we have simulated the GC skew formation using the whole genome sequence, without excluding any sequences. This is because, in theory, strand bias effects of mutations induced by other mechanisms than replication should cancel out, unless the mechanism itself is related to replication (Sueoka 1995). In *E. coli* and γ-Proteobacteria utilized in this work, gene orientation bias is almost even (54.43% in the leading strand in *E. coli* K12), and therefore transcription/translation-related mutation bias should have minimal effect on the GC skew in these species. On the other hand, local regions of genomes and especially the coding sequences are nonetheless subject to other types of mutations than replication, and therefore we have conducted additional validations to confirm such effects. For this purpose, we have repeated all three simulations (*dif*-stop, fork-collision, and fork-trap model) using only the third positions of the codons and intergenic regions (hereafter referred to as GC skew (GC3/non-coding)), in addition to the GC skew using whole genome sequences: GC skew (all). As expected, in both simulations, whether using the whole genome or only GC3/non-coding regions, the overall results did not change. The RMSE values showed similar tendencies, where the RMSE medians were 34.980, 37.516, and 1.493, for *dif*-stop, fork-collision, and fork-trap model, respectively in GC skew (GC3/non-coding), whereas those of GC skew (all) were 19.243, 27.772, and 16.439, respectively. Figure 3.9 shows the GC3/non-coding version of Figure 3.5. Overall, both simulations show that the fork-trap model can better explain the existing GC skew shape, rather than the *dif*-stop model.

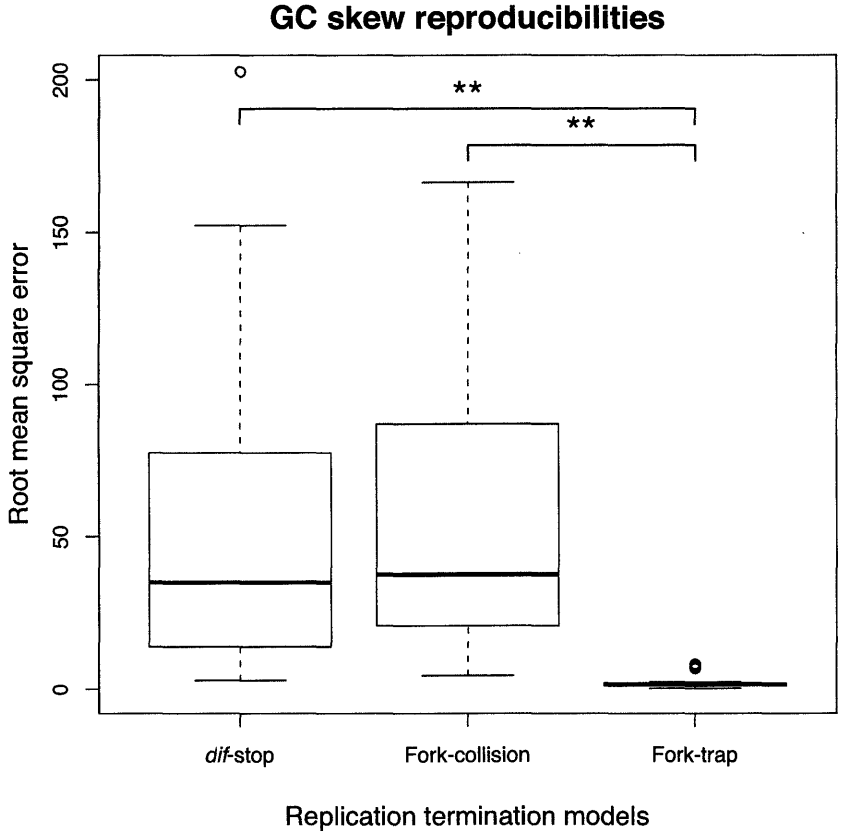

Figure 3.9: Validation of simulations using only the third codon positions and non-coding sequences. This figure shows the boxplot of the RMSE scores for the three replication termination models, representing the similarities between simulated and natural GC skews (in 65 bacteria). In comparison to Figure 3.5, here the GC skews were calculated and simulated only in the third codon positions and non-coding regions. The overall tendencies are identical to Figure 3.5. ** $p < 0.01$, Wilcoxon test.

For the fork-collision model, we determined the positions where the forks collide by observing the fluctuations of the GC skew shift-point in plasmids. Plasmids were used rather than chromosomes for several reasons. First, the chromosomal sequences are not suitable for determining these parameters because replication termination in these replicons involves mechanisms other than fork-collision. Moreover, long chromosomal sequences also undergo large-scale restructuring, typically by horizontal gene transfer or inversion (Rocha 2008). Inversions disrupt gene order and the orientations of oligonucleotides (Hill and Gray 1988; Liu *et al.* 2006), and the genomic islands acquired through horizontal gene transfer likewise change the genomic structure; they can be as large as 10,000 bp up to 1 Mbp (Gogarten and Townsend 2005; Juhas *et al.* 2009). We selected bacterial plasmids that depend on the host replication machinery based on the absence of the *repC* gene, which is required for rolling circle replication (van Passel *et al.* 2006) and based on the lack of Ter or *dif* sites. In these plasmids, the putative locations of frequent fork-collisions obey a clear Gaussian distribution centered at a position directly opposite that of the putative origin, as described in Figure 3.1B, suggesting that replication termination occurs probabilistically through collision and not by the action of specific terminating proteins. The speed of fork progression in both replichores seems to be similar, and the replichores show almost identical base compositions ($R = 0.994$).

The probabilistic distributions of the rates of fork-trapping at each Ter site in each bacterium were calculated from three biochemical evidences: the time and probability of accidental stalling of replication forks at sites other than Ter, the positional relationship among different Ter sites, and the leakiness of each Ter site. Based on these evidences, we could calculate the pausing ratio at each Ter site. Furthermore, in order to validate such pausing rates, we compared these biochemical parameters with a computationally determined pausing ratio by means of parameter search that best reconstructs the natural GC skew using all possible patterns of fork pausing at various Ter sites (see Materials and Methods). As a result, the calculated pausing rates based on experimental data were very similar with the optimized pausing rates ($\rho = 0.725$, Spearman's rank-correlation coefficient). Fork-trap model scored best among other replication termination models using either of these parameters.

The locations of *dif* sites strongly correlate with those of the GC skew shift-points ($\rho = 0.736$) (Kono *et al.* 2011), and these distances are closer than the nearest Ter sites and the loci directly opposite the replication origin (the average distance from the GC skew shift-point to a *dif* site is 0.39%, to the nearest Ter site = 0.68%, to the side opposite the origin = 2.61% in 65 targeted bacteria). Therefore, by naive expectation, replication should terminate near the *dif* sites to produce the GC skew graph seen in existing genomes. However, our simulation study shows that replication termination at a single finite locus cannot accurately reconstruct the GC skew shape. In fact, a single finite termination model results in a highly acute shift-point, but the actual shift-point is less acute and more rounded. Such a shape can only be reproduced with probabilistic models (the fork-trap and fork-collision models) (Figure 3.4). Therefore, the probabilistic balance of replication termination results in the current shift-point position, and the *dif* sites seem to be co-evolving and taking advantage of the genomic compositional bias to be near this probabilistic center of replication termination loci (which allows for efficient CDR). In fact, FtsK translocase locates the *dif* site and recruits XerCD recombinase to the site through the guidance of a highly skewed G-rich oligomer, known as the KOPS (Saleh *et al.* 2004; Bigot *et al.* 2005; Bigot *et al.* 2006), taking advantage of the genomic compositional skews and the distribution of the skewed oligomers (Salzberg *et al.* 1998; Hendrickson and Lawrence 2006). Therefore, our simulation study suggests that *dif* sites are not shaping the GC skew by terminating replication at this specific locus, but rather, the GC skew shift-point shaped by the replication termination machinery is affecting the location of *dif* sites. This is in agreement with *in vivo* studies (Duggin *et al.* 2008; Duggin and Bell 2009) and with our previous *in silico* study, showing that the distance between the *dif* site and GC skew shift-point is not correlated with GC skew strength (Kono *et al.* 2011). Finally, we confirmed the contribution ratio of each model to construct the GC skew using probabilistic combination model. The most optimal combination validated by RMSE was the 10-70-20% (fork-collision, fork-trap, and *dif*-stop model, respectively) in probabilistic combination. In previous studies, it has been indicated that the replication fork arrest occurs in 18 to 50% of replication cycles with

several factors, including transcription-replication collisions, fork-trap with Ter/Tus complex, or by inactivation proteins (Kogoma 1997; McGlynn and Lloyd 2002; Xu and Marians 2003; Michel *et al.* 2007). In addition to these studies, Maisnier-Patin *et al.* reported an estimate of at least 20% of all replication forks are stalled and require replisome reassembly during the replication process (Maisnier-Patin *et al.* 2001). Furthermore, Hendrickson and Lawrence speculate that the cleavage of *dif* might occasionally block the progression of forks (Hendrickson and Lawrence 2007). Therefore, our probabilistic combination simulation yielding 10-70-20% ratios for fork-collision, fork-trap, and *dif*-stop model seems to fit reasonably well to explain the contributions of different fork-termination mechanisms.

CHAPTER 4

The relationship between the symmetry of bacterial circular genomes and genomic islands

"We are like a little child entering a huge library. The walls are covered to the ceilings with books in many different tongues. The child knows that someone must have written these books. It does not know who or how. It does not understand the languages in which they are written. But the child notes a definite plan in the arrangement of the books—a mysterious order which it does not comprehend, but only dimly suspects."

-Albert Einstein

4.1 Introduction

Various symmetrical/asymmetrical structures are maintained in the bacterial circular chromosome (Rocha 2008). A typical example of a symmetrical structure is the orientation of an oligomer nucleotide. The sequences that are associated with recombination or replication–*e.g.* the Chi sequence (Kowalczykowski *et al.* 1994), replication-related KOPS (Bigot *et al.* 2005; Bigot *et al.* 2006), and AIMS (Hendrickson and Lawrence 2006)–are symmetrically conserved in two replichores with the same frequency (Mrazek and Karlin 1998; Salzberg *et al.* 1998). An example of an asymmetrical structure is gene strand bias. The distributions of bacterial genes differ between the leading and lagging strands. Approximately 78% of genes are in the leading strand in Firmicutes bacteria or *Mycoplasma* (Fraser *et al.* 1995; Kunst *et al.* 1997; Rocha 2002) and 85% of ribosomal proteins are encoded in the leading strand in *Bacillus subtilis* and *Escherichia coli* (McLean *et al.* 1998). High-expression genes in particular tend to cluster in the leading strand (McLean *et al.* 1998). These symmetrical/asymmetrical structures are associated with the balance of the replication origin and terminus. For example, the replication origin and terminus pair is symmetrical in *E. coli*, and replication proceeds bi-directionally from the origin to a terminus (Prescott and Kuempel 1972; Hirose *et al.* 1983; Schaper and Messer 1995; Schaeffer *et al.* 2005). The most notable feature of an asymmetrical structure that is provided by replication symmetry is a GC skew (Lobry 1996a; Rocha *et al.* 2006). The GC skew can be calculated as (C − G)/(C + G) and is visualized as the compositional bias towards G and C bases. The shift-points of a GC skew are known to correlate with the replication origin and terminus positions, as the GC skew is believed to be a consequence of the replication process. Although DNA is replicated according to the direction of the polymerase, a leading strand is synthesized continuously, whereas the lagging strand is synthesized discontinuously. Consequently, the single strand duration of the leading strand is longer than for the lagging strand, and this difference in the replication mechanism results in a GC skew (Coulondre *et al.* 1978; Reyes *et al.* 1998; Mackiewicz *et al.* 2003). Based on this finding, analyzing the GC skew has become a common *in silico* method for predicting the origin and terminus (Frank and Lobry 2000; Worning *et al.* 2006). However, the degree of GC skew is extremely variable, and certain bacteria have only a weak skew (Zhang *et al.* 2003; Worning *et al.* 2006; Arakawa *et al.* 2009a) (Figure 4.1). Therefore, to select the bacteria for which a prediction of the replication origin and terminus can be made, a quantitative index to measure the strength of the GC skew has been developed (Arakawa *et al.* 2009a). Based on the increased ability of *ori/ter* prediction using GC skew, many bacteria with imbalanced *ori/ter* structures have been identified. It is highly unlikely that these detected imbalanced *ori/ter* structures were maintained in the actual bacterial genomes. Therefore, the drastic effects in the base composition may perturb the accuracy of *ori/ter* prediction and incorrectly indicate imbalances. Among these mutations, GEIs (genomic islands) are considered to have the potential to invoke large-scale changes in base composition. GEIs are large foreign regions of approximately 10 Kbp - 1 Mbp in the bacterial genome (Rocha 2008) and were most likely acquired via horizontal gene transfer (HGT) (Gogarten and Townsend 2005; Juhas *et al.* 2009). If a large GEI is inserted into a bacterial genome, and if the base compositions of these inserts differ substantially from the host, the insertion will likely reduce the accuracy of predicting the GC skew. Here, we compared the symmetrical structures between natural genomes and artificial genomes whose GEIs were computationally deleted and investigated the disruptive effects that were caused by GEIs.

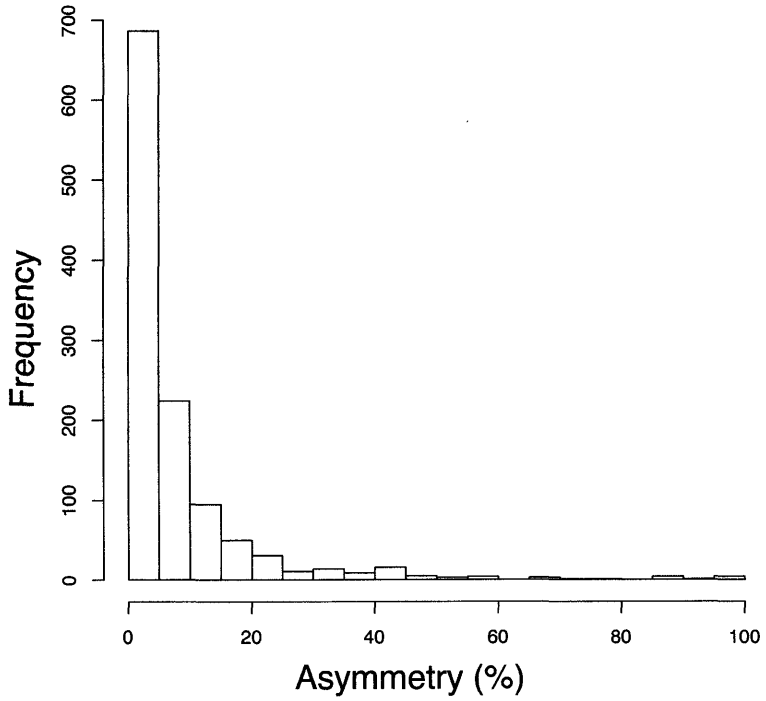

Figure 4.1: Bacterial asymmetries.
This histogram indicates the asymmetry in 1,164 bacteria harboring circular chromosomes. The x-axis represents the extent of asymmetry as a percentage. When asymmetry is 0%, the bacterial replication terminus is located precisely opposite to the replication origin.

4.2 Materials and Methods

4.2.1 Genome sequences and software

All of the analyses in this study were conducted using programs that were written in Perl with the G-language Genome Analysis Environment, version 1.8.13 (Arakawa *et al.* 2003; Arakawa and Tomita 2006; Arakawa *et al.* 2008). Statistical analyses and graphic presentations were performed using the R project, version 2.10.0 (www.R-project.org). Bacteria that contain a circular chromosome, have a genome larger than 5 Kbp, and are included in the GEIs dataset were selected as the target bacteria (486 strains). These bacterial genome sequences were obtained from the NCBI FTP Repository (ftp://www.ncbi.nlm.nih.gov/Ftp).

4.2.2 Dataset

For this study, a set of 486 GEIs was obtained from the IslandPath-DIMOB dataset (Hsiao *et al.* 2005), which was curated using IslandViewer (Langille and Brinkman 2009). The IslandPath-DIMOB dataset was screened by measuring the dinucleotide bias and the presence of mobility genes.

4.2.3 The calculation of asymmetries

The replication terminus and origin were predicted using the "find_ori_ter" function of the G-language GAE, which is based on the cumulative GC skew (Grigoriev 1998) at a resolution of 1 bp. The asymmetries between the replication origin and terminus are represented as percentages. This asymmetry is 0% when the replication terminus is precisely opposite of the origin, and the percentage increases as the terminus approaches the origin. In this analysis, if the replication terminus is found to be located on the left replichore, the asymmetry score will be negative. The extent of change in asymmetry between the wild-type genome and the GEI-deleted genome was calculated as $dA = AP_w - AP_d$ (delta value) and $IR = |AP_w| - |AP_d|$ (improvement rate), where AP_w and AP_d are the asymmetries of the wild-type and GEI-deleted genomes, respectively. If the origin and terminus are closer to the symmetrical position, the IR value will be higher. The asymmetries of the GEI-deleted genome were determined using the replication origin and terminus that were re-predicted based on the GC skew. However, in the calculation of the fixed asymmetry, the replication origin and terminus in the GEI-deleted genome were not re-predicted. We initially marked the positions of the replication origin and terminus in the wild-type genome and calculated the asymmetries for each position of the genomes whose GEIs were computationally deleted.

4.2.4 Random deletions

A randomization test was applied to each bacterium to measure the GEI-specific effects. The parameters including only the positions of GEIs, and the number, length, and replichore in which the GEIs were located were maintained. For each bacterium, the randomization test was run using 100 iterations.

4.3 Results

4.3.1 Assessment of the effects of GEIs

To assess the effects of GEIs on the symmetrical structures of bacterial genomes, we performed an *in silico* experiment for 486 bacteria that were predicted to have GEIs according to IslandPath-DIMOB. Specifically, we computationally deleted the GEI sequences from the bacterial genomes and re-predicted the replication origins and termini in the GEI-deleted genomes. The asymmetries were determined by a comparison between the newly predicted *ori/ter* and the original *ori/ter*. Although no significant correlation was observed between GEI length and the asymmetries ($R = 0.199$, Pearson correlation coefficient), the presence of GEIs had a substantial effect on the genome structure regardless of island size (Figure 4.2A). Interestingly, a strongly correlated group (the gray symbols in Figure 4.2A, $R = 0.990$, Pearson correlation coefficient) was found using an uncorrelated scatter plot. The genome sizes of the bacteria belonging to the gray group were smaller than those of the other bacteria. The average genome sizes were 3,682,798 bp and 2,547,252 bp for the black and gray groups, respectively. A significant difference was measured between the black and gray groups ($p < 0.0001$, t-test, Figure 4.2B).

4.3.2 GEIs affect base composition

In bacteria that have a small genome, GEIs had an effect on the genomic structure. However, it is difficult to determine whether this effect was caused by changes in base composition. Because the replication origins and termini in both the original and GEI-deleted genomes were predicted using the GC skew, the validity of the prediction method cannot be tested. Therefore, we used an additional method to detect the replication origins and termini. This alternate detection method is based on the fixed positions of the original replication origin and terminus. First, we recorded the *ori/ter* positions in the original genome, and then we used the positions in the GEI-deleted genomes without re-prediction. Therefore, if the deletion of the GEIs does not affect the base composition, there will be no significant difference between the re-predicted *ori/ter* positions and the fixed positions. Using this approach, although a strong correlation between the fixed and re-predicted asymmetry was measured ($R = 0.86039$, Pearson correlation coefficient), a few outlier bacteria were detected (Figure 4.3).

4.3.3 Improvement rates from GEIs deletion

According to Figure 4.3, because differences were observed between the re-predicted and fixed *ori/ter* positions in the outliers (the red symbols in Figure 4.3), deleting the GEI may have affected their base compositions. Therefore, we calculated the improvement rates for these outliers. Because the previous asymmetries included only the extent of change between the *ori/ter* positions in the original genomes relative to the respective GEI-deleted genomes, this approach was unable to confirm the negative or positive effects of the GEIs. The improvement rates can validate the extent to which the symmetries in the GEI-deleted genomes were improved relative to the original genomes. The results show that a greater extent of asymmetry in the original genomes corresponded to higher improvement rates in the GEI-deleted genomes. Figure 4.4 shows the improvement rates that were calculated using the re-predicted *ori/ter* positions, fixed *ori/ter* positions, and randomly deleted artificial GEIs.

Moreover, this result also shows that the improvement rates were significantly higher based on the asymmetries in the original genome than for the random artificial GEIs deleted genomes.

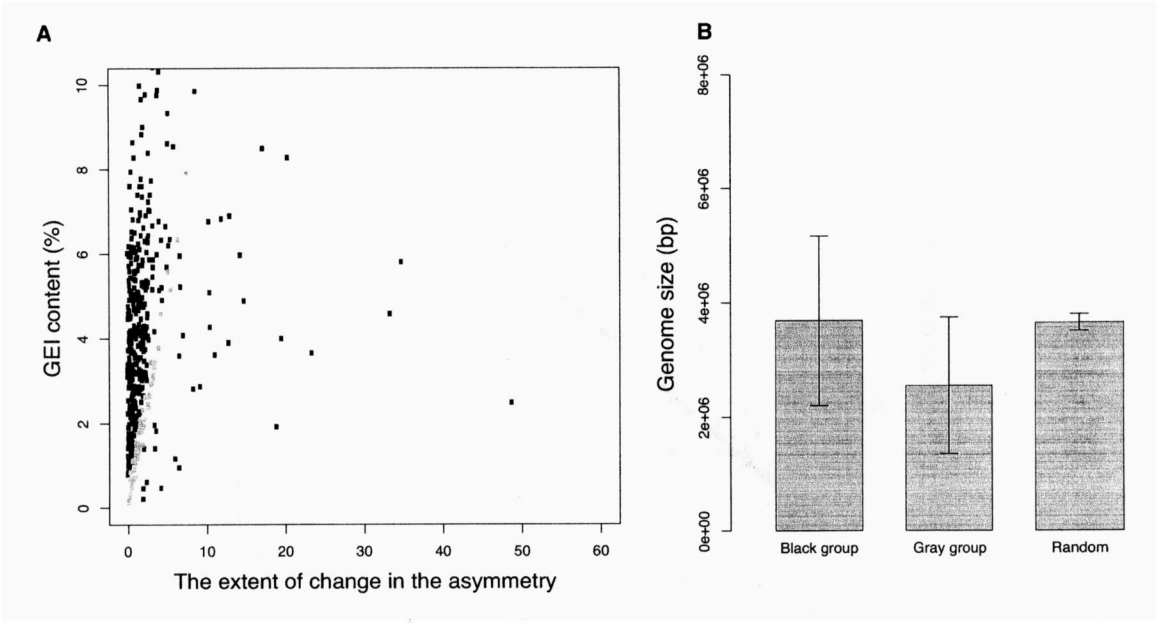

Figure 4.2: The extent of change in asymmetry with genome size.
A: The x-axis indicates the extent of change in the asymmetry, and the y-axis shows the GEI content (%) of each genome. The gray symbols indicate bacteria with a genome size that is comparatively smaller than the others. B: The mean values of the genome sizes in poorly correlated (represented by the black symbols in A) and strongly correlated (the gray symbols in A) bacteria. The "Random" bar is the mean of a random sample of 100 genomes. The error bars show the standard deviation.

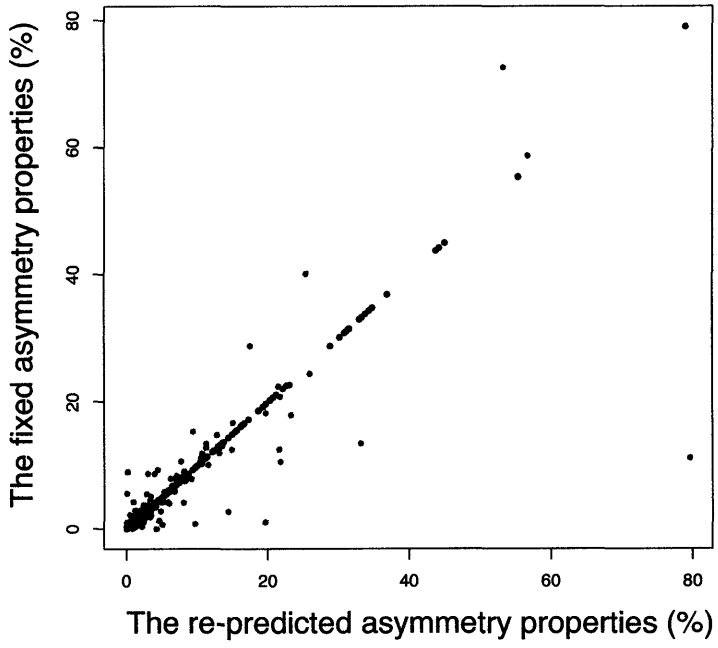

Figure 4.3: The correlation of between the asymmetries of the re-predicted and fixed replication origins and termini.
The x-axis indicates the re-predicted asymmetries, and the y-axis shows the fixed asymmetries. $R = 0.860$, Pearson correlation coefficient. The red symbols indicate the bacteria that differed between their re-predicted and fixed *ori/ter* positions.

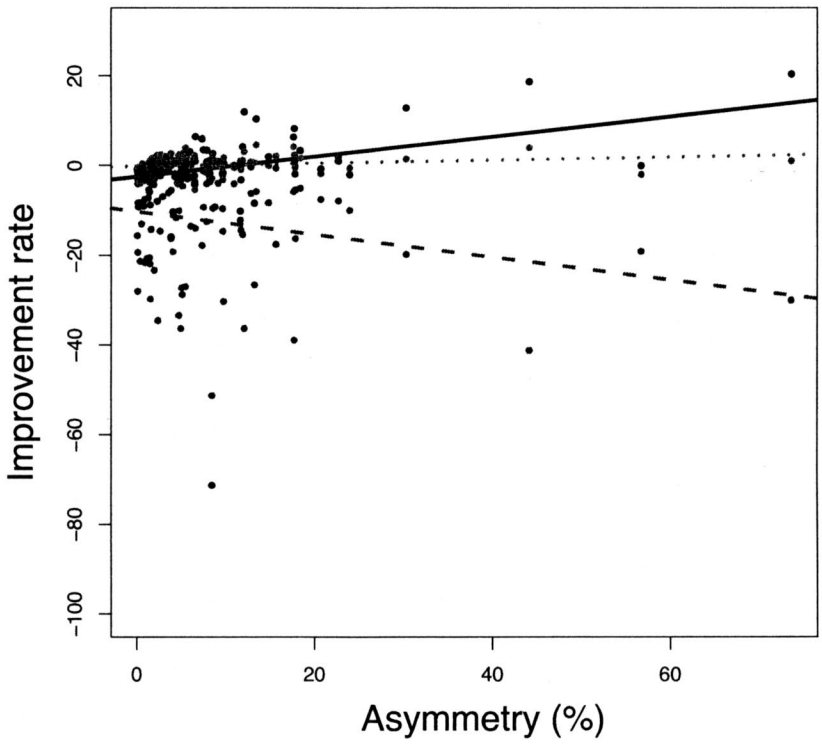

Figure 4.4: Improvement rates in three cases.
This plot shows the improvement rates for each bacterium. The y-axis indicates the improvement rate, and the x-axis is the asymmetry of the original genomes. Each color represents the improvement rate that was calculated using the re-predicted *ori/ter* positions (black), the fixed *ori/ter* positions (red), or the randomly deleted artificial GEIs (blue). Each line is a linear regression fit of the data. The solid, dashed and dotted lines are $p < 0.01$, 0.05, and 0.1, respectively; F-test.

4.4 Discussion

The objective of this research was to validate the effects of GEIs on the symmetry of the structure of bacterial genomes. Using computational deletions of GEIs, we observed a specific relationship between GEIs and genomic structures. Regardless of GEI length, a subset of the GEI-deleted genomes exhibited changes in their asymmetries. In particular, bacteria with small genomes were affected by the GEIs. It is logical to expect that changes in the symmetrical structure due to the positional distance between the replication origin and terminus will be shorter or longer based on the deletions of specific regions from the genome. Indeed, there were many shifts in the positions of the replication origins and/or termini in the GEI-deleted genomes. However, in certain bacteria, when we re-predicted the positions of the replication origin and terminus after deletion of the GEIs, the distances of movement were different from the fixed positions, regardless of GEI length of GEIs. This result suggests that the *ori/ter* prediction results were altered by changes in the GC skew that were caused by the deletion of the GEIs. Furthermore, when the original genome had an imbalanced symmetric structure, large improvement rates were observed in these bacteria.

However, there are a few particular aspects that merit further discussion. The first point is the possibility that the GEIs, including their base composition, might adapt to their new genomic environment. Foreign genomic sequences have been altered progressively by mutations during the long course of their evolution (Lawrence and Ochman 1997). For example, bacteria have acquired numerous GEIs by horizontal events, and although the detection algorithm can identify relatively recently acquired segments, this method may be not able to detect older imported segments due to their adaptations (Mira *et al.* 2001; Cropp *et al.* 2002); in other words, many of the GEIs that can be detected in various bacterial genomes may not affect the symmetrical structure.

The second point is that any imbalances in the symmetrical structure may have other causes such as inversions, replication mechanism including the differences in leading strand and lagging strand, or a programmed system into bacteria for something other purpose. In the Introduction, we presumed that GEIs have the potential to change the base composition on a large scale and that the disruption of symmetrical structures may be caused by GEIs. In particular, the asymmetrical replication origin and terminus may be an incorporated system in the original genome. Because the base composition bias is caused by a mutation that is associated with the replication process (Frank and Lobry 2000; Rocha 2004), the skew is inseparably related to the replication origin and terminus. Therefore, if the replication origin and terminus in its pure state (*i.e.* without GEIs) were not symmetrical and were instead biased towards one side or the other, this asymmetrical structure might represent a deflection, and order can be maintained.

CHAPTER 5

Codon usage is a selection pressure for the H-NS binding sites

"Ce que je vois là n'est qu'une écorce. Le plus important est invisible."

-Antoine de Saint-Exupéry 'Le Petit Prince'

5.1 Introduction

Bacterial chromosome does not form a chromatin structure using histones as eukaryotic chromosome. However, the bacterial chromosome is packaged into a nucleoid structure in order to compact the genome. The nucleoid is organized by the act of supercoiling, RNA and nucleoid-associated proteins (NAPs), which regulate DNA topology (Dame 2005). Similar DNA packaging machinery is also known for mitochondria (Friddle *et al.* 2004) and chloroplasts (Sekine *et al.* 2002). The formation of nucleoid is mediated by the interactions of DNA and various proteins known as NAPs, and wraps the DNA in less than half of the intracellular volume (Cunha *et al.* 2001; Wiggins *et al.* 2010). The DNA-NAPs interactions have several styles as follows, bending, wrapping or bridging (Dorman and Kane 2009; Dillon and Dorman 2010). In gram-negative bacteria, at least 12 types of NAPs have been reported, and the most well-researched NAP is H-NS. This protein was discovered by isolating bacterial homologs of eukaryotic histones (Varshavsky *et al.* 1977), and plays an important role in modifying the DNA topology to form the nucleoid (Falconi *et al.* 1988). The overexpression of H-NS leads to dimers in solution, and the H-NS proteins can organize DNA-H-NS-DNA bridges. Recently, it was reported that the H-NS protein not only contributes to the genome packaging, but also controls the bacterial transcriptions along with DNA sequences and range of transacting factors (Browning and Busby 2004). Furthermore, H-NS is called as 'genome sentinel' because the regulation by H-NS has a key role in selectively silencing horizontally-acquired genes (Dorman 2007). The function of gene regulation was highlighted by the recent genome wide analyses of H-NS binding sites with a DNA microarray or a chromatin immunoprecipitation (ChIP) method in *Escherichia coli*, *Salmonella* and *Yersinia* (Lucchini *et al.* 2006; Navarre *et al.* 2006; Cathelyn *et al.* 2007; Banos *et al.* 2008; Grainger *et al.* 2006; Oshima *et al.* 2006; Gordon *et al.* 2011). In *E. coli*, Grainger *et al.* and Oshima *et al.* observed the binding sites for H-NS genome widely by using ChIP-on-chip approach, and revealed that the H-NS binding regions were more AT-rich than the resident genome. Additionally, they found that there were somewhat correlations between RNA polymerases and H-NS binding sites in both at promoter and coding regions (Oshima *et al.* 2006; Grainger *et al.* 2006). These results suggest that the trapping of RNA polymerase by H-NS causes the gene silencing. In *Salmonella*, Lucchini *et al.* and Navarre *et al.* found that the binding sites of H-NS were essentially restricted to AT-rich region in common with *E. coli* by ChIP-on-chip approach (Lucchini *et al.* 2006; Navarre *et al.* 2006). Moreover, Lucchini and colleagues showed that the H-NS of *Salmonella enterica* inhibits the interactions of RNA polymerase and DNA, and functions as a gene silencer for horizontally-acquired genes (Lucchini *et al.* 2006). These characteristics were very similar to H-NS of *E. coli*. Additionally, with increase of the genome wide analyses of binding sites for H-NS, Lang *et al.* discovered that an abundant AT-rich motif were conserved at the high affinity binding regions from the genome wide ChIP-on-chip experiments. According to their study, this motif was significantly present in pathogeny gene and suggested that the H-NS guards the host genome from pathogeny new genetic materials (Lang *et al.* 2007). Thus, the features of binding sites for H-NS were ascertained in genome wide. Despite that H-NS homologous genes are conserved among gram-negative bacteria, the evolutionary trend of the gene silencing machinery has been observed in several other species (Dorman 2004). Here we analyzed how the silencing machinery evolved. In order to understand the acquisition of gene regulator function, we compared the features of the H-NS target sites.

5.2 Materials and Methods

5.2.1 Software and genome sequences

All analyses in this study were conducted in use of programs written in Perl with the G-language Genome Analysis Environment, version 1.8.13 (Arakawa *et al.* 2003). For all statistical analyses, we used the R statistical

package 2.10.0 (www.R-project.org). This study used 113 Proteobacteria strains which have H-NS homologous gene of *E. coli* K-12 substr. MG1655 and *Salmonella enterica* serovar Typhimurium str. LT2 based on sequence similarity matches (blastp; e-value $\leq 10^{-20}$). We obtained these genome sequences from NCBI FTP Repository (ftp://www.ncbi.nlm.nih.gov/Ftp). Horizontally-acquired regions were predicted by Alien_Hunter 1.7 which is implemented based on the interpolated variable order motifs (Vernikos and Parkhill 2006). The essential gene data sets were identified based on DEG 6.5 (Zhang and Lin 2009).

5.2.2 Discovering motifs and prediction of binding sites from ChIP data

The H-NS binding locations were acquired from two previous reports, Kahramanoglou *et al.* (Kahramanoglou *et al.* 2011) and Lucchini *et al.* (Lucchini *et al.* 2006). The centers of binding signal were identified as the binding site in accordance with the methods from each paper. In order to discover the binding site motifs for H-NS, the 201 bp regions from the upstream 100 bp to downstream 100 bp including the signal site were defined as the binding regions. The motifs in these determined binding regions were searched by using MEME 4.6.1 (Bailey and Elkan 1994). The parameters are as follows: the type of sequence model to use was zero or one occurrence per sequence; motif width ranges were from 8; background distribution file contained the mono- and di-nucleotide frequencies of the *E. coli* and *S. enterica* chromosomes. Using the searched motifs, the binding sites in other bacteria were predicted by MAST 4.6.1 (Bailey and Gribskov 1998). The threshold was set to $p < 0.001$, e-value $< 10^{-10}$.

5.2.3 The codon usages among horizontally-acquired genes and core genes

The tribasic motifs, which are observed in the H-NS binding site motifs with approximately 2-bits (Figure 5.1), code mainly three amino acids (Leucine (Leu), Serine (Ser), and Valine (Val)), and the synonymous codon usages are compared among horizontally-acquired genes and core genes. In particular, we calculated the occurrence rates of the tribasic codons in horizontally-acquired genes and core genes. At this time, the tribasic motif codons were GTA in Val, AGT in Ser, and CTA in Leu. Alanine (Ala) and Proline (Pro), whose codons are most and least used in *E. coli* and *S. enterica* and are four-fold degenerate, were used as the negative controls. This calculation ignored the start and the stop codons to avoid the extreme bias.

5.2.4 The binding ratio in each region

The frequencies of H-NS binding sites were calculated by using the number of bound sites divided by the total length of each gene. This method was applied to high expression gene or essential gene. The screening of high expression genes was done based on Codon Adaptation Index (CAI) value (Sharp and Li 1987) and tRNA Adaptation Index (tAI) value (dos Reis *et al.* 2004). Similarly, the screening of genes whose codon usages are similar to essential gene was also done based on the CAI value. At this time, the w-value was calculated from the essential genes.

5.2.5 Phylogenetic analysis

A phylogenetic tree was constructed using a neighbor-joining method with 1,000 bootstrap replicates and alignment of 16S rRNA in γ-Proteobacteria. These sequences were aligned by MAFFT 6.859 (Katoh and Toh 2008). The cladogram tree was drawn by using FigTree 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

5.3 Results

5.3.1 Discovering H-NS targeting motifs

We searched the H-NS binding motifs in *E. coli* and *S. enterica* to verify the features of binding sites which were conserved between different species. In two previous studies, the binding sites for H-NS were identified by genome wide ChIP approaches in *E. coli* and *S. enterica* (Lucchini *et al.* 2006; Kahramanoglou *et al.* 2011). Therefore we used these data to compare the motifs. The discovery of the binding motifs was implemented by MEME software (see detail in Materials and Methods about the parameters). As a result, the detected motifs were very AT-rich and similar to each other in *E. coli* and *S. enterica* (Figure 5.1). Additionally, these motifs had characteristic tribasic motifs A, T and A at 4 bp, 6 bp and 7 bp. Since these target motifs are very short, here we considered that the tribasic motifs may represent some codons. Therefore, we investigated the distribution of such codon usages in horizontally-acquired genes and core genes. As a result, every tribasic motif codons were significantly more frequently used at the horizontally-acquired genes than core genes in *E. coli* (Figure 5.2A, $p < 0.0001$, Wilcoxon rank sum test) and *S. enterica* (Figure 5.2B, $p < 0.00001$, Wilcoxon test), whereas the negative controls Ala and Pro could not show the same tendency. Altogether, H-NS seems to distinguish the horizontal and core genes by using this tribasic motif.

5.3.2 The motif conservation in related bacteria

Subsequently, we applied the same analyses to the related 113 γ-Proteobacteria. In almost all bacteria, the tribasic motif codons were significantly more frequently used in horizontally-acquired genes than core genes (Figure 5.3 and Figure 5.4). According to the plots, it confirmed that the tribasic motif codons usages were significantly biased toward horizontally-acquired regions in almost all bacteria ($p \leq 0.0001$, Welch t-test). However, the biases were not observed in all bacteria. 12 bacteria (hereafter regarded as the "dissimilar group"; Table 5.1) did not show the significant differences (gray plots in Figure 5.3A, B, and C). In the dissimilar group, in contrast, the usage of tribasic motif codons in core gene is significantly higher than horizontally-acquired gene instead of other tendency. As an example, we showed the bar chart for *Buchnera aphidicola* str. APS

similar to Figure 5.2 (Figure 5.3D).

Table 5.1: List of the dissimilar group.

Bacteria	Genome size (bp)	AT (%)	Binding rate (%)[a]
Escherichia coli str. K-12 substr. MG1655	4,639,675	49.210	0.074
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	4,857,432	47.778	0.080
Baumannia cicadellinicola str. Hc	686,194	66.761	0.375
Buchnera aphidicola str. 5A	642,122	73.714	0.681
Buchnera aphidicola str. APS	640,681	73.688	0.678
Buchnera aphidicola str. Tuc7	641,895	73.708	0.680
Candidatus *Blochmannia pennsylvanicus* str. BPEN	791,654	70.435	0.580
Candidatus *Hamiltonella defensa* 5AT	2,110,331	59.672	0.210
Photorhabdus luminescens subsp. laumondii TTO1	5,688,987	57.175	0.180
Proteus mirabilis HI4320	4,063,606	61.100	0.247
Vibrio fischeri ES114	2,897,536	61.047	0.165
Vibrio fischeri MJ11	2,905,029	61.116	0.164
Vibrio harveyi ATCC BAA-1116	3,765,351	54.453	0.052
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	697,724	77.521	0.938

[a] This percentage is the ratio between the number of predicted binding sites and genome size.

5.3.3 Index to recognize the binding target

In this section, we attempted to explain why the H-NS selects the motifs including tribasic motifs as the target sites. If H-NS binds to any sites, it has an undeniable effect on the bacterial growth because the expressions of bound genes are suppressed. In fact, since H-NS binding sites are twice more frequent in the coding region than in the intergenic region (covering 0.056% and 0.061% of coding regions, and 0.028% and 0.032% of intergenic regions) in *E. coli* and *S. enterica* , an H-NS binding site therefore needs to be selected based on the expression level as well as the coding requirements. In order to assess the gene expression level, we used CAI values with ribosomal genes as the reference set to assess the gene expression in related bacteria. In *E. coli*, actually, the H-NS did not bind to the high CAI genes in comparison to the low CAI genes (Figure 5.5). As a result, there was significant relationship between the CAI values and the frequencies of H-NS binding frequencies in all bacteria but dissimilar group (Figure 5.6A). The higher the CAI values, the lower the binding frequency in genes. In the dissimilar group, however, the relationship was not observed (Figure 5.6A). Similarly, the tAI values also showed the significant relationship (Figure 5.6B). The higher tAI values, the lower the binding frequency in genes. In the dissimilar group, however, the relationship was not observed (Figure 5.6B). Hence, the H-NS may avoid the genes that are translationally optimized. Furthermore, the relationship among the codon usage or base composition of essential genes and H-NS binding rates in *E. coli* were analyzed. The result revealed that the H-NS did not bind to genes having similar codon usage as essential genes (CAI $\geq$ 0.8). On the other hand, this binding frequency was only observed at the codon usage level, and not at the level of similar base composition (AT% $\pm$ 0.01%) (Figure 5.7). Finally, we observed the phylogenetic relationship among the target bacteria (Figure 5.8). The phylogenetic tree resulted that the dissimilar group bacteria diverged, and they were located far from *E. coli* and *S. enterica*.

5.4 Discussion

In this research, we aimed to understand how H-NS acquired the genome sentinel function by analyzing the features of H-NS binding motifs. As a result, it was suggested that the genome sentinel function might be caused as a byproduct of avoiding the silencing of important genes.

We first confirmed that the H-NS binding site motifs were common to *E. coli* and *S. enterica* (Figure 5.1), and were very AT-rich. The appropriateness of these confirmed motifs was also supported by previous studies, for example, Navarre *et al.* and Lang *et al.* reported the H-NS target site is 78% AT composition (Navarre *et al.* 2006; Lang *et al.* 2007), and Gordon *et al.* showed that the H-NS binds to AT-rich minor groove (Gordon *et al.* 2011). In this study, two independently obtained ChIP data were used to discover binding motifs, namely, that in *E. coli* using ChIP-seq by Kahramanoglou and coworkers (Kahramanoglou *et al.* 2011), and that in *S. enterica* using ChIP-on-chip by Lucchini and colleagues (Lucchini *et al.* 2006). Although they did not use the same target bacteria and methods, the same motifs could be detected from their binding data sets. Previously, in γ-Proteobacteria, it is known that there is a relationship between H-NS structures and genome similarities (Tendeng and Bertin 2003). We could confirm that using experimental data. This fact attributes that the binding site motifs for their H-NS are tightly conserved because of these strong sequence similarities.

However, since such detected target motifs were common and short segments, this carries a risk of having the bindings and inhibitions everywhere on the chromosome. Thereby, we focused on the tribasic motifs which were dominantly conserved with nucleotides at 4 bp, 6 bp, and 7 bp positions in target motif. These tribasic motifs showed distributions throughout the whole genome. Especially, the tribasic motif codons were present in horizontally-acquired genes than the core genes. Hence, these codons may be the important key to recognize the foreign genes. In order to understand the selection reason of these codons, we considered that the codon usage in each gene is a selection pressure of H-NS binding motifs. In *E. coli*, actually, H-NS less frequently binds to genes having similar codon usage to essential genes, because the essential genes should be avoided by the gene repression (Figure 5.7). Furthermore, in related bacteria, the H-NS binding motifs were not found in genes which have similar codon usage to ribosomal genes (Figure 5.6). Hence, H-NS avoided to bind to the essential or ribosomal genes.

This fact was found in almost all related bacteria, but some species did not show the same tendency. According to the bacterial phylogenetic tree (Figure 5.8), it seemed that some bacteria were branched early (we called such bacteria as dissimilar group). These different results in dissimilar group are suspected to be due to a property of each H-NS protein structure. In previous study, it was suggested that the "QGR" motif may be used by H-NS family proteins to bind DNA (Gordon *et al.* 2011). Therefore we observed the amino acid sequences in target bacteria, and the alignment result showed that there were variability of sequences between dissimilar group and others (Figure 5.9). Although the "QGR" motifs were conserved in almost all H-NS proteins, the dissimilar group sequences showed different trends around 100th residue. If the same H-NS binding site is maintained in dissimilar group bacteria, the H-NS will not function as the genome sentinel, because the tribasic motif codon usages of foreign genes were different from core genes. The genomes in dissimilar group evolved independently, and might have changed the H-NS target site. On the other hand, in *B. aphidicola*, other hypothesis about the evolutionary change of H-NS is also considerable. *Buchnera* is an insect endosymbiont and has nearly minimal gene set and small genome size, approximately 0.6 Mbp, because of interdependence with host species (Brinza *et al.* 2009). The *B. aphidicola* genome sequences used in this study were three strains, 5A (NC_011833), APS (NC_002528) and Tus7 (NC_011834), and these strains have *hns* genes. However, the *hns* homologous genes were not found in other genome sequenced strains, *B. aphidicola* Bp (NC_004545), Sg (NC_008513), and Cc (NC_004061), using blastn (e-value $\leq 10^{-5}$). Therefore, in such endosymbiont genome, H-NS functions may be lost.

Thus, H-NS targeted the tribasic motif codons which were more frequently used in horizontally-acquired genes than core genes. H-NS regulates DNA topology and organizes the nucleoid structure to compact bacterial chromosomes like an eukaryotic chromatin structure. In order to package the chromosome, H-NS should bind to many positions in bacterial chromosome. However, because gene densities in bacterial genomes are high (average gene density in targeted bacteria: 84.580%), H-NS might avoid the risk of having a critical effect on bacteria to select target genes, for example essential or ribosomal genes. Accordingly, the tribasic motif codons were chosen in consequence of the avoidance, and the genome sentinel function is acquired as a byproduct of genome packaging mechanism.

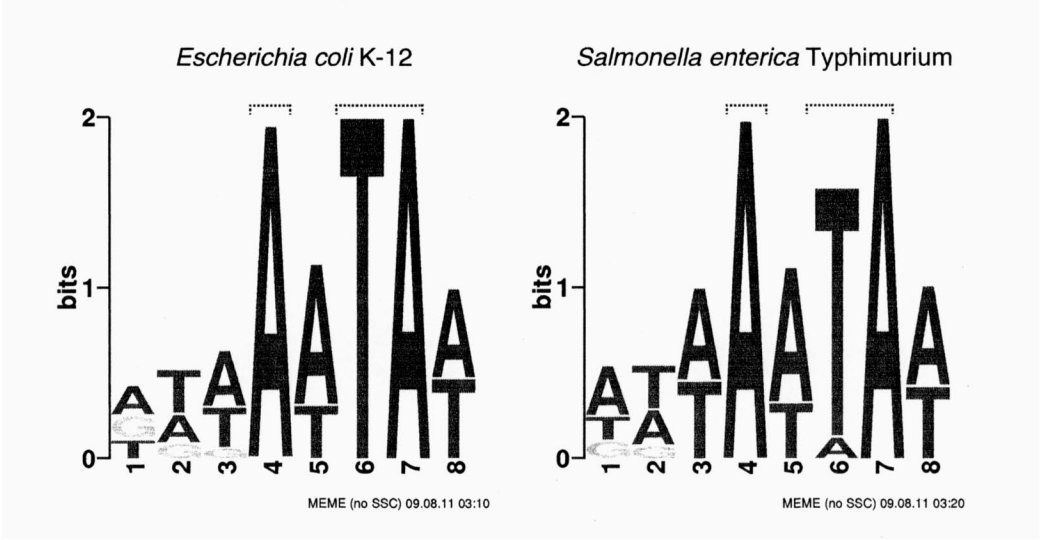

Figure 5.1: Sequence logo of H-NS bind sites.
The sequence logos of the H-NS binding site motifs in *E. coli* and *S. enterica*. The regions marked by dashed line are the prominent nucleotides (tribasic motifs).

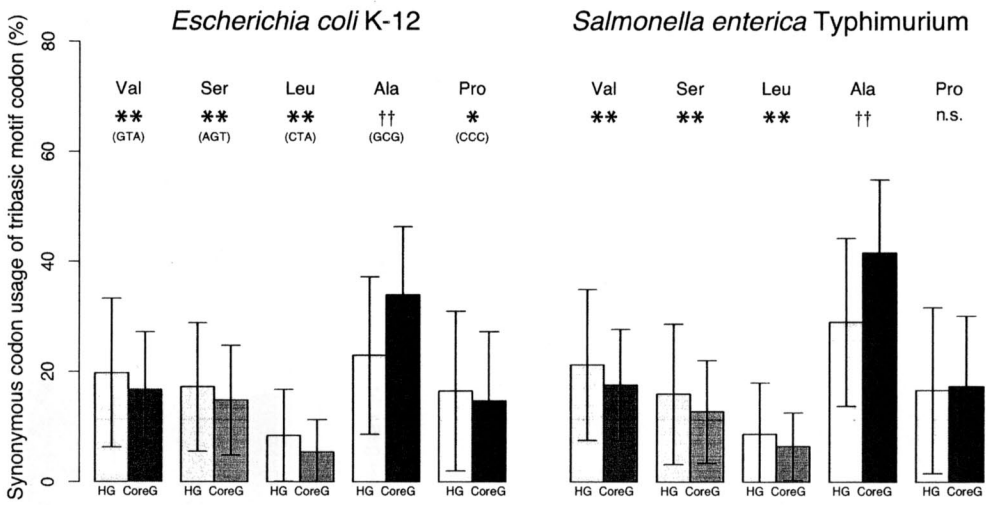

Figure 5.2: Synonymous codon usage of tribasic motif codon.
The synonymous codon usages in each horizontally-acquired gene (HG: light colors) and core gene (CoreG: dark colors). In *E. coli* and *S. enterica*, colored three bars represent the codon usage in each amino acid (Val: blue, Ser: green, and Leu: yellow) in HG and CoreG. The light and dark gray bars are the negative controls (Ala and Pro). For example, in blue bar (Val), the tribasic motif codons (GTA) were over 19% of total Val codons in HG (light blue), but in CoreG (dark blue bar), the usage was about 16%. The error bars indicate each standard deviation. The usage of tribasic motif codon in HG is significantly higher than CoreG (* $p < 0.01$, ** $p < 0.0001$). On the other hand, the usage of tribasic motif codon in CoreG is significantly higher than HG (†† $p < 0.0001$). n.s. (no significance, Wilcoxon rank sum test).

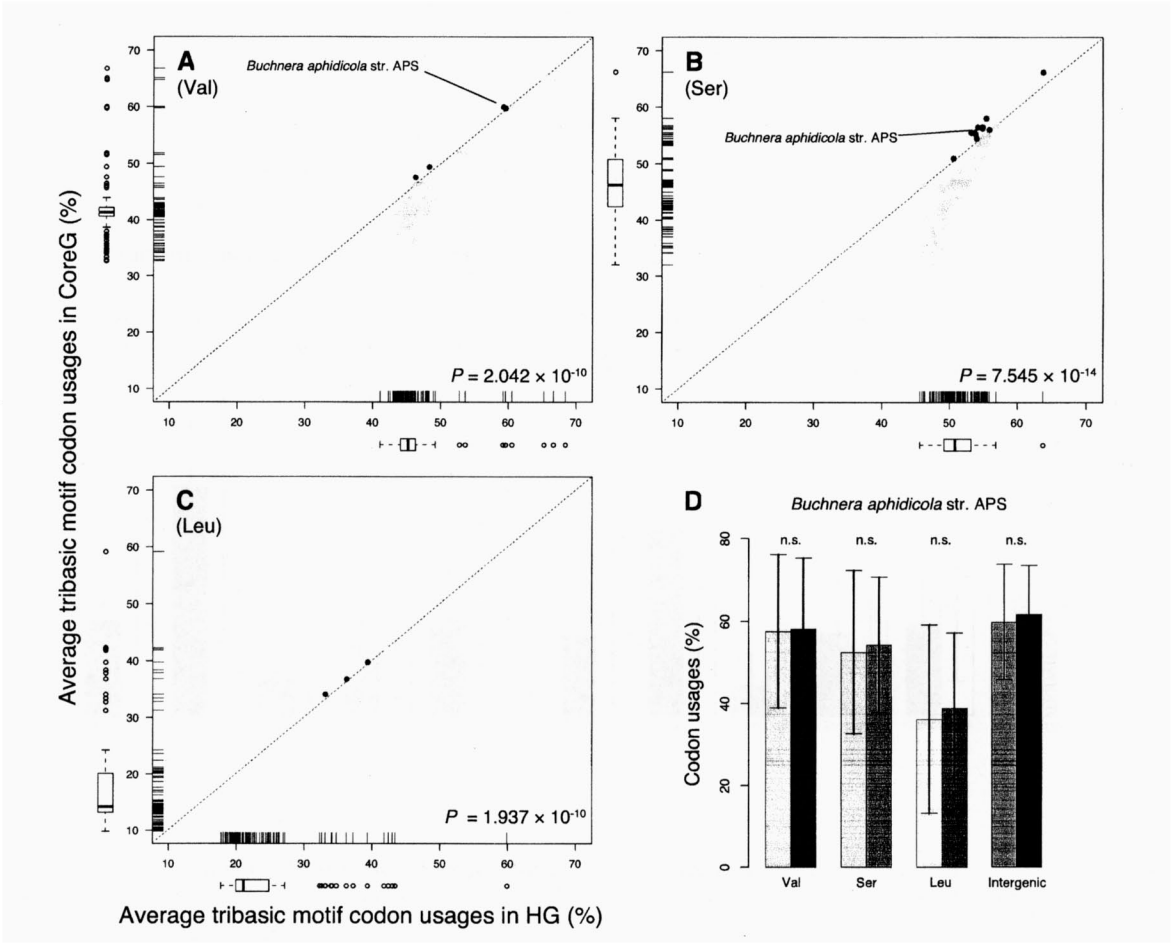

Figure 5.3: Average tribasic motif codon usages.

These scatter plots A - C represent the average usages of tribasic motif codons in horizontally-acquired genes (HG) and core genes (CoreG). D: This figure is the bar chart about *Buchnera aphidicola* str. APS similar to Figure 5.2. The x-axes are the average content percentages of tribasic motif codons in HG, and the y-axes are the percentages in CoreG in each bacterium. Colored plots represent each bacterium, and the Val (A), Ser (B), and Leu (C) are drawn as blue, green and yellow respectively. The gray plots are the bacteria which do not have significant differences. These tribasic motif codons were significantly more frequently used at the HG than CoreG in almost all bacteria, but there were 12 exception bacteria including *Buchnera*. n.s. (no significance, Wilcoxon rank sum test).

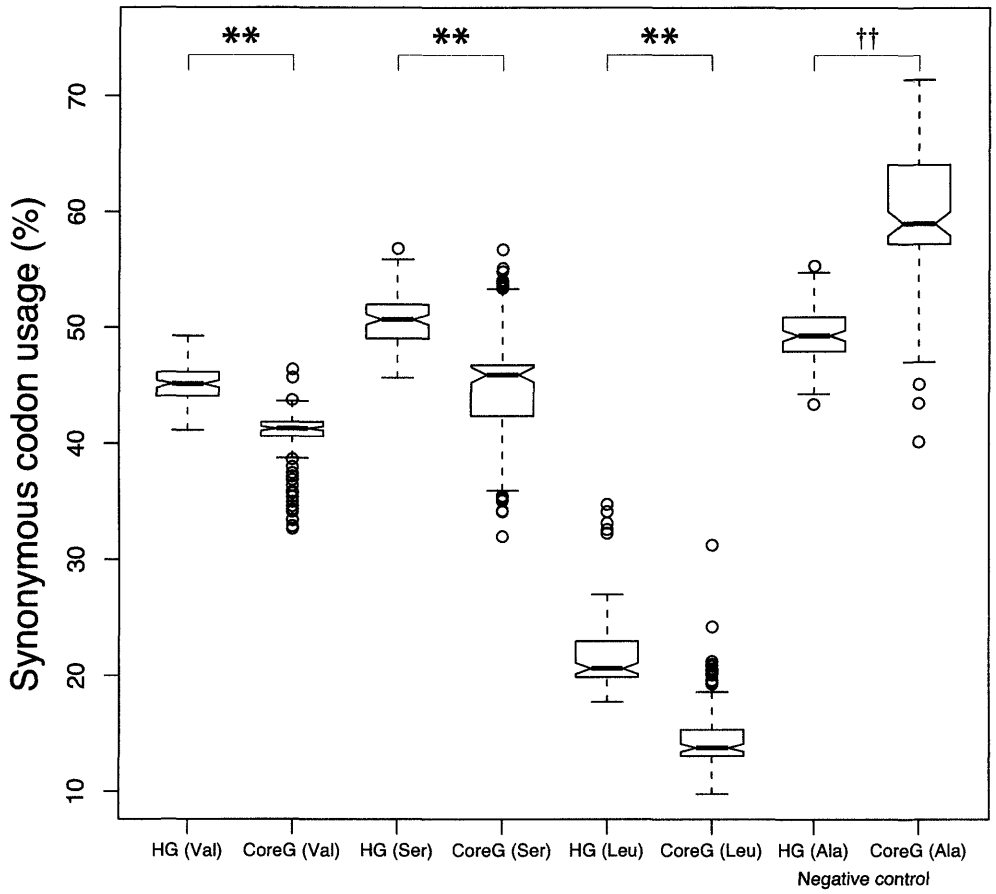

Figure 5.4: The average of synonymous codon usage in 113 bacteria.
The x-axis is the synonymous codon usages in each bacterium. The usage of tribasic motif codon in horizontally acquired gene (HG) is significantly higher than core gene (CoreG) (** $p < 0.0001$). On the other hand, the usage of tribasic motif codon in CoreG is significantly higher than HG (†† $p < 0.0001$).

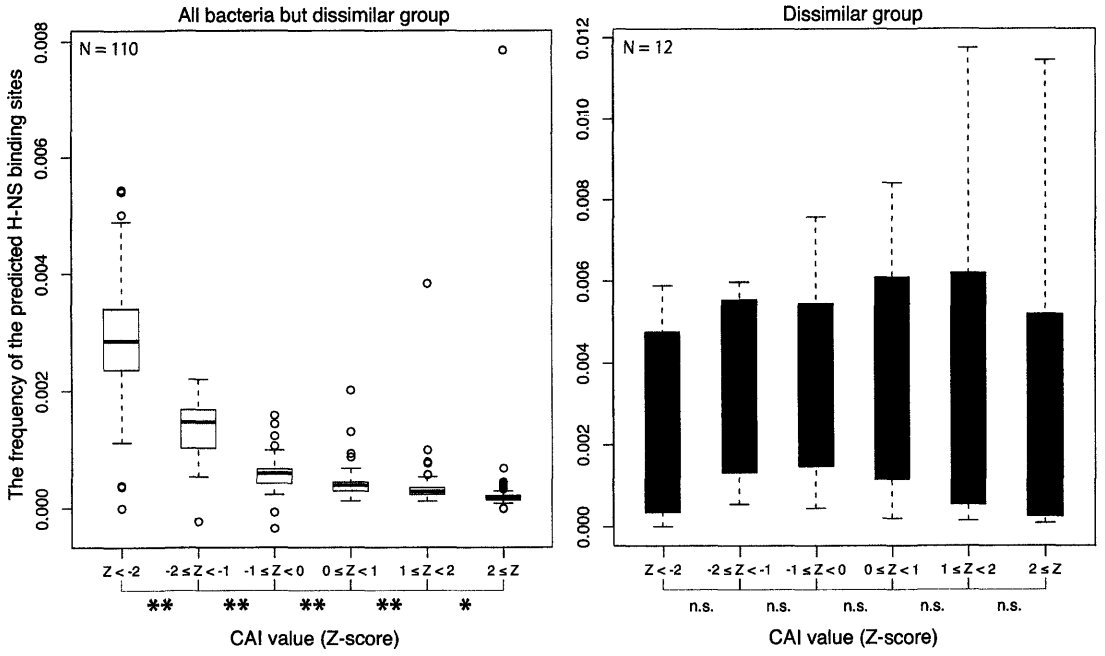

Figure 5.5: The number of H-NS binding sites.
The distribution of the H-NS binding sites by using *E. coli* ChIP-seq experimental data (Kahramanoglou *et al.* 2011). The x-axes represent the CAI values. The y-axes are the number of H-NS binding sites per one gene. The number of H-NS binding sites per one genes showed lower as CAI values were higher. * $p < 0.01$, ** $p < 0.0001$, n.s. (no significance), Welch t-test.

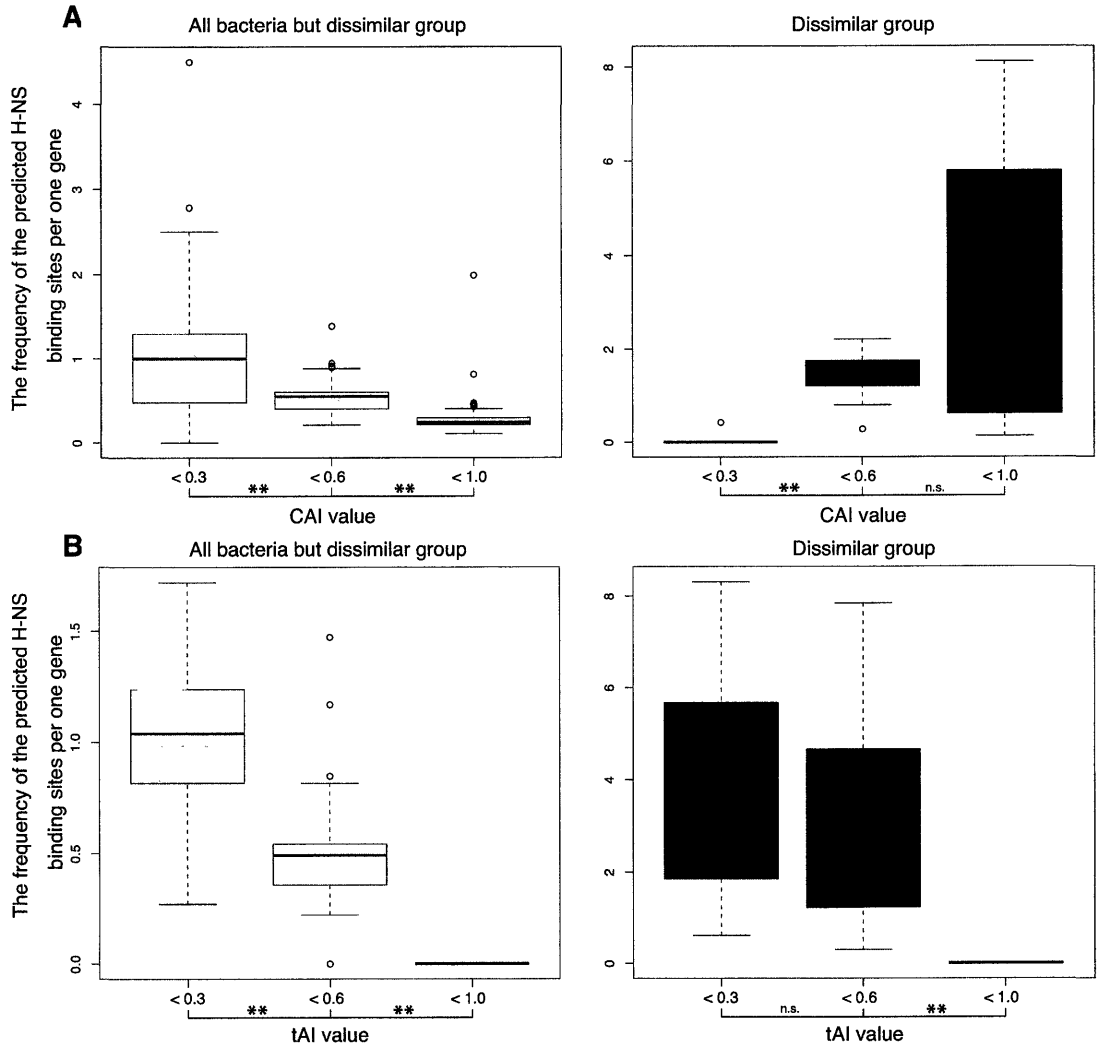

Figure 5.6: The frequency of binding sites.
The x-axes represent the CAI values (A) and tAI values (B), and the y-axes are the frequencies of the predicted H-NS binding sites. These gray bars are all bacteria but dissimilar group, and black bars are the dissimilar group. ** ($p < 0.0001$), n.s. (no significance), Welch t-test.

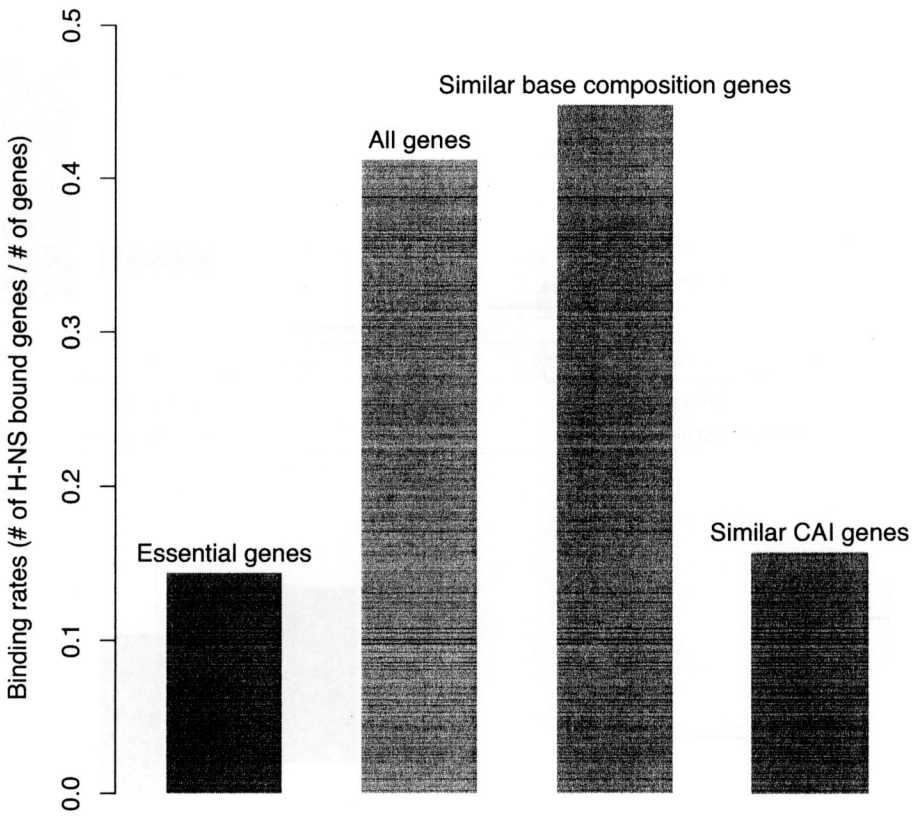

Figure 5.7: Binding rates in several type genes.
The y-axis is the predicted H-NS binding rates. The bars mean each gene group, essential gene group (red), all gene group (blue), similar base composition to essential genes group (green), similar codon usage to essential genes group (orange). The essential gene data sets were identified based on the database of essential genes (DEG 6.5; Zhang and Lin 2009).

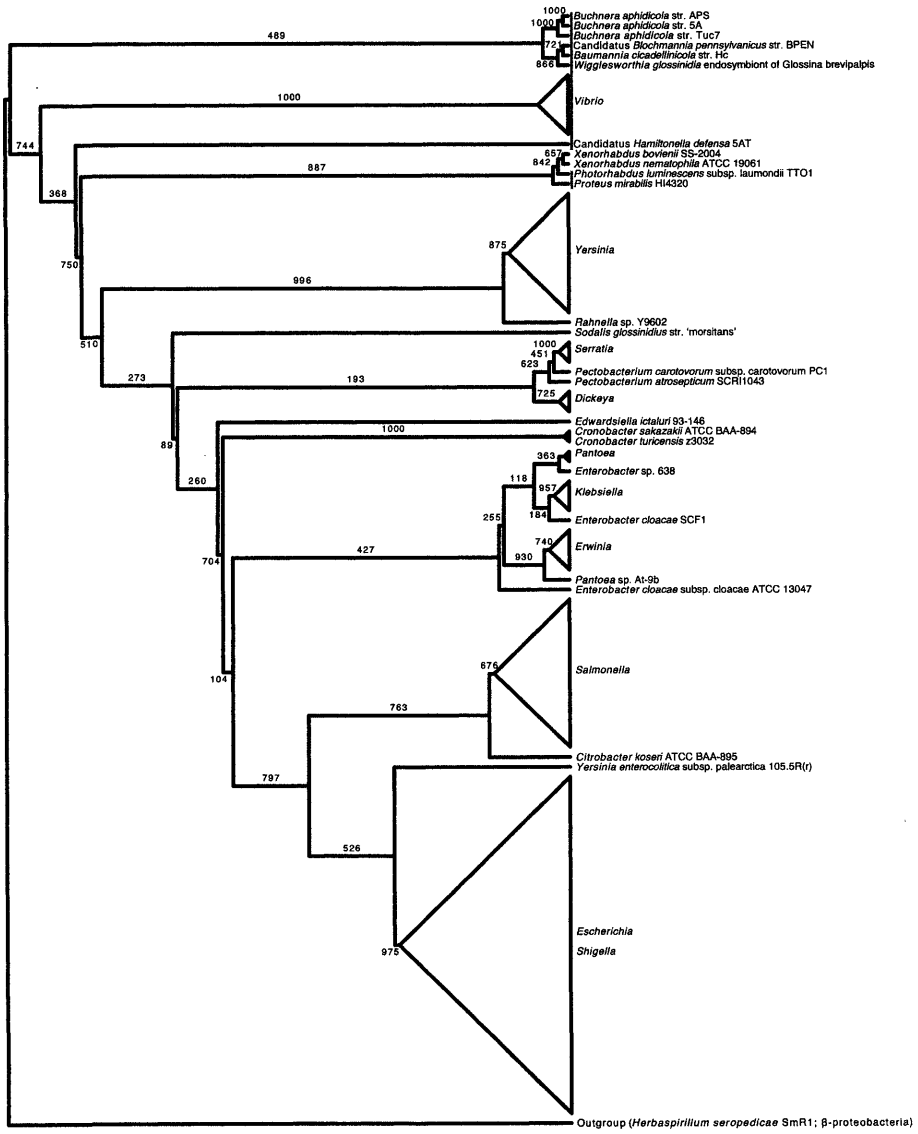

Figure 5.8: Phylogenetic tree.
This phylogenetic tree was built based on 16S rRNA sequences of 113 Proteobacteria including *E. coli* K-12 and *S. enterica* LT2 using a neighbor-joining method with 1,000 bootstrap replicates. The dissimilar group bacteria are shown by red. *Herbaspirillum seropedicae* SmR1 (β-Proteobacteria) was used as an outgroup.

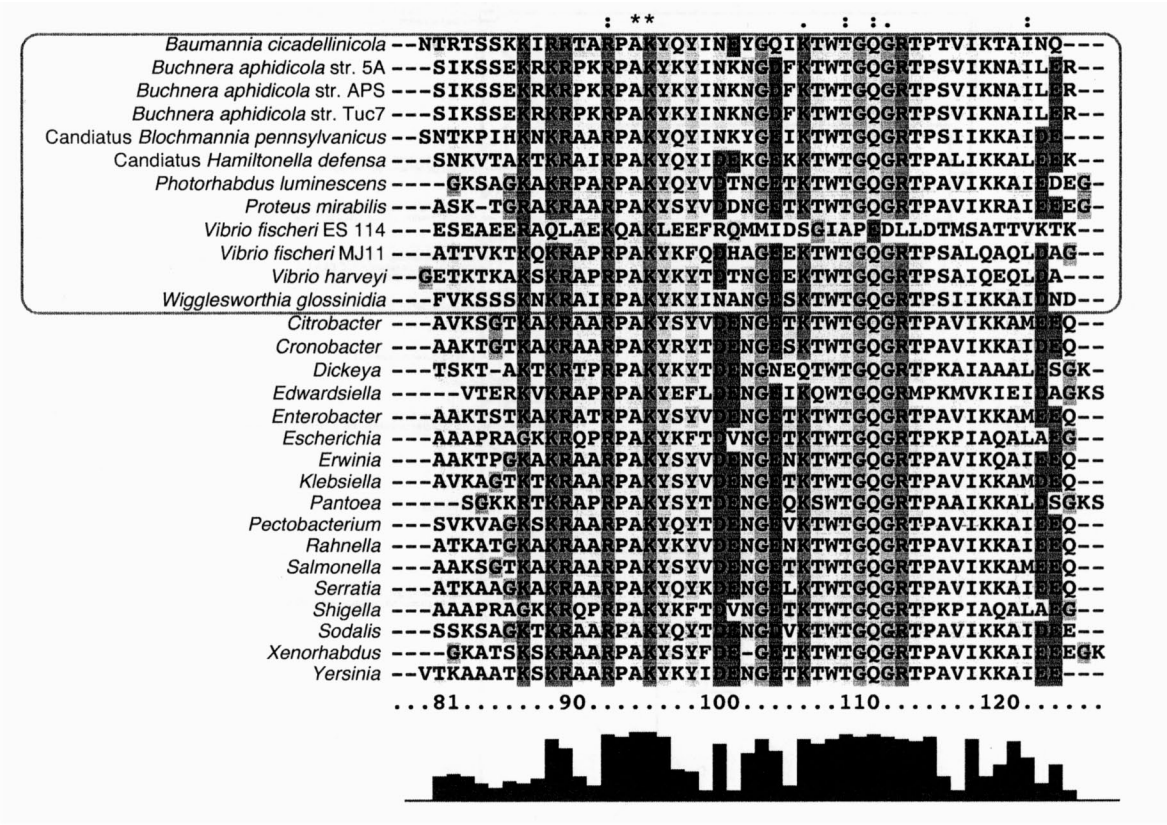

Figure 5.9: Alignment result of H-NS in dissimilar group.
This figure shows the alignment result of H-NS amino acid sequences in a part of target bacteria with ClustalX (Larkin *et al.* 2007). The organisms surrounded by a red border are dissimilar group. The dissimilar group sequences showed different trends around the 100th residue.

CHAPTER 6

Pathway Projector: Web-based Zoomable Browser using Google Maps API

"Imagination is more important than knowledge."

-Albert Einstein

6.1 Introduction

With a long tradition of being a descriptive discovery science, the field of scientific visualization has been an integral part of biosciences and has also been an indispensable approach for understanding complex, large-scale data in molecular biology. Numerous approaches for information visualization have been successfully utilized and have contributed to the understanding of genomic information, including those for the protein 3D structure, sequence alignment, and phylogenetic trees (Tao *et al.* 2004). Genome browsers, such as Gbrowse (Stein *et al.* 2002), UCSC Genome Browser (Kuhn *et al.* 2009), and Ensembl (Hubbard *et al.* 2009), have been a particular success because they provide a visual context (Mangan *et al.* 2008). Genome browsers show gene structures and their locations within the genome, and they can also be used to map novel knowledge and experimental data to display them in a genomic context. Systems biology approaches (Kitano 2002; Zhu *et al.* 2009) attempt to understand cellular processes as a system of molecular interactions. In post-genomic research, these approaches demand another context for biochemical pathways in order to understand biological information. A biochemical pathway is a series of reactions that consists of enzymes, proteins, and molecular compounds (Papin *et al.* 2003), and is a useful context for understanding how gene disruptions or alterations of conditions associate with a phenotype (Ekins *et al.* 2007). For example, in microarray or proteomic experiments, researchers can map their experimental data through pathway mapping systems, such as ArrayXPath II (Chung *et al.* 2005), GenMAPP (Salomonis *et al.* 2007), and Pathway Explorer (Mlecnik *et al.* 2005), to gain a comprehensive understanding of cellular regulation and to explore the existence of alternative pathways after gene deletions or change in conditions. Therefore, visualization approaches allow for an intuitive understanding of a large quantity of data that is inherently difficult to comprehend, while biochemical pathways provide a suitable context for observing the systematic cellular behavior that is analyzed through "-omics" experiments (Adriaens *et al.* 2008). Pathway browsers will thus enhance systems biology research.

6.1.1 Individual pathway map

We first have developed a web application that visualizes complex omics data of multiple layers simultaneously for individual pathway map, including transcriptome, proteome, and metabolome, onto an integrated pathway diagram derived by connecting the individual KEGG pathway maps; the mapped images are generated in Scalable Vector Graphics (SVG) for easy editing by hand or with computer programs or drawing software. The web application is available at http://megu.iab.keio.ac.jp/.

We present a new generic pathway visualization tool primarily to aid the heuristics of researchers during the observation of system-level regulation of interacellular interactions. For this purpose, here we define the requirements as follows: 1) ability to map data from multiple layers of omics, 2) visualization based on familiar pathway map but with a cell-wide view, 3) flexibility in editing with manual methods and with computer-friendly format and semantics, and 4) interactivity to connect with external data.

The software system was developed using the generic bioinformatics workbench, G-language Genome Analysis Environment (Arakawa *et al.* 2003; Arakawa and Tomita 2006; Arakawa *et al.* 2008), and was implemented as a web-service. SVG is supported in most browsers either natively with Firefox and Opera, or with free plug-in available from Adobe as in Internet Explorer or Safari. Individual pathway diagrams are drawn in SVG using the coordinates in KEGG Markup Language (KGML), and the integrated pathway diagram is drawn manually by connecting the individual pathway diagrams. For example, the integrated map for all carbohydrate metabolism encompasses 17 pathways including glycolysis, the TCA cycle, the pentose phosphate cycle, and many other metabolic pathways, in order to capture the cell-wide activity at a glance. Cellular compartmentalization is also based on the KEGG pathway representation.

The SVG format has several advantages over other graphic formats, including object-based vector representation for easy editing, resolution-free zooming suitable for the observation of large integrated pathways, interaction and animation features for richer interfaces, and data in regular text format as a subset of an eXtensible Markup Language (XML) that can be easily handled by computer programs or text editors. SVG files can also be loaded and edited as objects with common graphic software such as GIMP and Adobe Illustrator, so that the users can add new graphics and components (that are not in KEGG, for example) onto the mapped pathway image, and customize the colors and values manually as desired.

To map specific data onto the pathway diagrams, users are required only to select the target pathway from the pull-down menu and input the data in the text-area at our web site. The input data are presented as a list of comma-delimited "Name,Value" pairs, where each line represents one molecular entity. "Name" is the name of an entry to be mapped onto the pathway, which can be either a canonical or standard name for genes and mRNAs, the EC number for proteins, and the KEGG compound ID for compounds. "Value" is the intensity of the color used for mapping represented by an integer between 1 and 100. Values for genes and proteins are represented by the range of colors between red and green, and those for metabolites between blue and green. Sample data and demonstrations are available at the software web page as working examples.

Since this system is based on the KEGG pathway representation, labels sometimes overlap. In order to avoid this, information windows can be toggled by pressing the key 1 to 4 to show/hide the labels. Pressing the '1' key shows compound data ('2' key hides these data), '3' key shows gene data ('4' key hinds these data) and toggles the Information Window, which contains information on the color levels, names, and EC numbers of the nodes displayed directly beside the pathway nodes. Furthermore, most of the objects are also draggable. All nodes are linked to external resources in KEGG; the hyperlinks are activated by clicking the nodes.

The primary objective of this system is in the pathway mapping using familiar pathway diagrams of the KEGG database, for omics experimental data. Therefore, our primary focus is in the simultaneous mapping of the quantitative information about biomolecules including the genes, mRNAs, proteins, and metabolites on a single pathway map. Currently MEGU focuses on the visualization of the high-throughput experimental data such as microarray and metabolome, and has limitation in the visualization of the molecular interaction networks.

Using MEGU, we generated pathway graphics of integrated carbohydrate metabolism mapped (Figure 6.1) with the microarray data of 38 knockout mutants of the two-component regulatory system in *Escherichia coli* K-12 W3110 (Oshima *et al.* 2002).

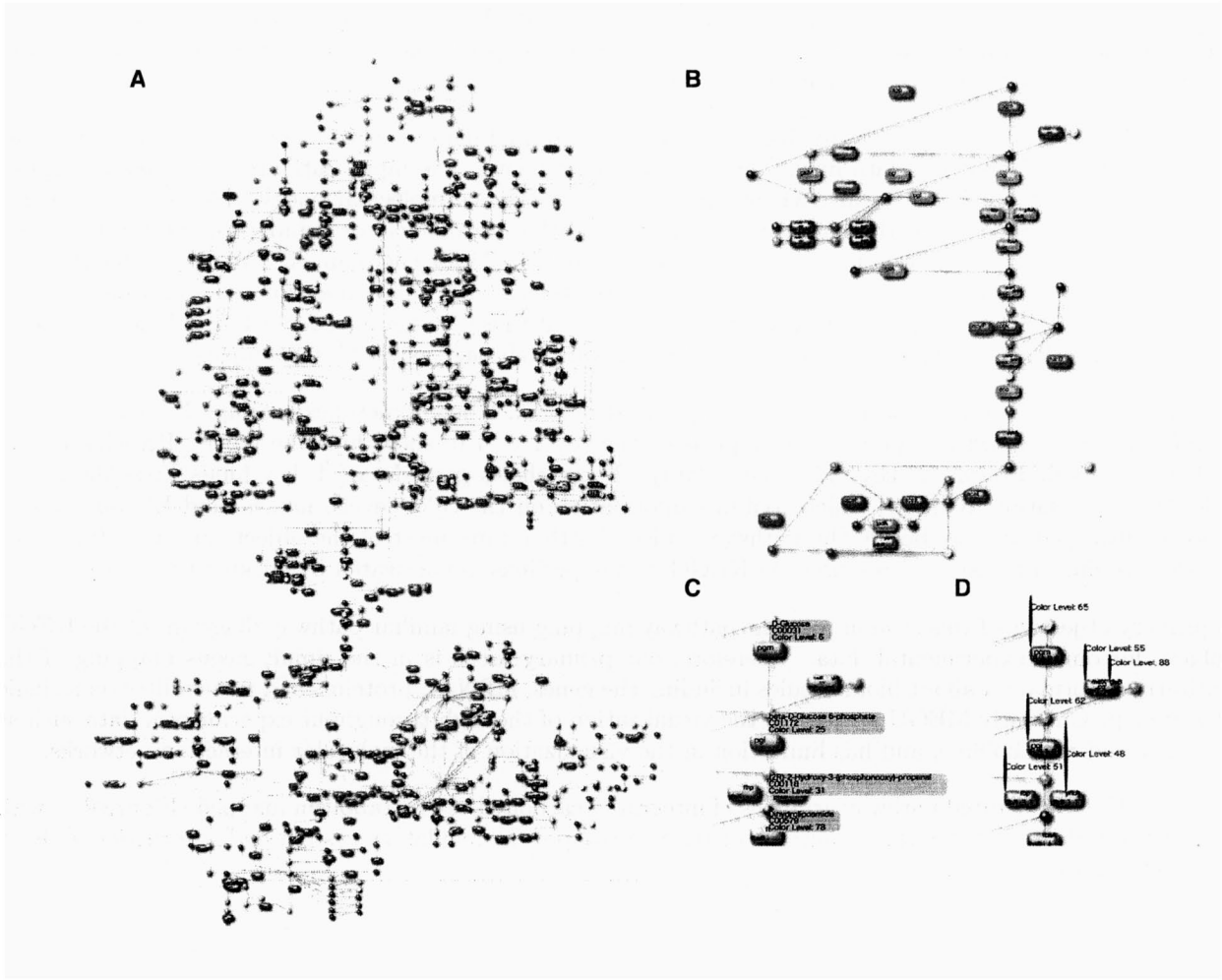

Figure 6.1: Output example of MEGU.
A: Integrated diagram of carbohydrate metabolism pathways mapped and generated with MEGU. The rectangular nodes correspond to mRNAs or proteins and circles correspond to metabolites. The visualized pathway diagram based on the coordinates of KEGG database is generated in SVG format from the web-service, so that any part of the pathway can be easily enlarged without affecting the resolution of image, as shown in (B) Glycolysis Pathway. The map can be further toggled to show information windows (C) and bar graphs of the molecular amount (D) with interactive and animated manipulations, and the objects are hyperlinked to external resources. Corresponding entries of KEGG can be accessed by activating the hyperlink on each of the nodes.

6.1.2 Integrated pathway map

Most existing pathway maps have been provided as parts of major public pathway databases at their websites. These maps are subdivided into individual pathways, in part due to technical limitations in manipulating large images on the World Wide Web. Given that pathways are essentially connected *in vivo* and that highly comprehensive experimental data that encompass a wide variety of pathways is readily available, arbitrary partitioning of pathways is often not useful for the mapping and observation of comprehensive experimental data. For instance, the glycolysis/gluconeogenesis pathway (map00010) in KEGG (Kanehisa *et al.* 2008) links to five pathways: the citrate cycle (map00020), the pentose phosphate pathway (map00030), starch and sucrose metabolism (map00500), carbon fixation in photosynthetic organisms (map00710), and propanoate metabolism (map00640). Users have to constantly switch back and forth between the maps to observe reactions that encompass multiple pathways. Therefore, with the advancement in web development technologies (Zhang *et al.* 2009), several pathway databases have started to release integrated pathway maps that allow comprehensive viewing. For example, the KEGG Atlas (Okuda *et al.* 2008), iPath (Letunic *et al.* 2008) and the new beta version of Reactome (Matthews *et al.* 2009) display comprehensive integrated pathway maps without page transitions, which have been implemented as zoomable and scalable maps. The Omics Viewer (Paley and Karp 2006) in BioCyc (Karp *et al.* 2005) implements this feature with pop-ups upon mouse-over action. These interface technologies that enable the continuous display of a large image at different scales without page transitions are collectively known as the zoomable user interface (ZUI). ZUI is successfully utilized for the representation of geographical information as typified by Google Maps (http://maps.google.com/), as well as for the visualization of gene networks and the implementation of genome browsers (Uchiyama *et al.* 2006; Arakawa *et al.* 2009b; Itoh and Watanabe 2009; Obayashi *et al.* 2009).

Despite the recent availability of several integrated pathway maps, the abstraction level of represented entities in these maps is often not sufficient to map experimental data, which is primarily due to the objectives of each pathway database. For example, the KEGG Atlas and Omics Viewer do not show genes and enzymes as nodes, but instead represent them as reaction edges. The Reactome only shows enzymes as nodes, which limits its applicability for the mapping of microarray data, since many enzymes exist as heteromers that are comprised of several proteins. Similarly, when only reactions are represented in the map, data from metabolomic experiments cannot be mapped. To the best of our knowledge, there is currently no pathway browser that can map experimental data for genes, enzymes, and metabolites simultaneously using a comprehensive integrated pathway map.

In this chapter, we present Pathway Projector, a pathway browser that allows the mapping of multi-omics information on an integrated pathway map through an intuitive user interface and ZUI. The integrated pathway map of Pathway Projector is based on the widely used layout of the KEGG Atlas with the addition of nodes for genes and enzymes for the mapping of experimental data from transcriptomic, proteomic, and metabolomic experiments. We also identify and discuss the requirements for an ideal pathway browser.

6.2 Materials and Methods

6.2.1 Requirement analysis

As a result of a close collaboration between our bioinformatics group and experimental biologists, we have first identified the requirements for a pathway browser. We have analyzed the requirement especially in consideration of the current situations facing the biologists working with large-scale omics data with systems biology approaches, where they require a comprehensive understanding of cellular workings. Therefore, the pathway

browser should be continuous and global, covering the omics layers of genome, transcriptome, proteome, and metabolome, while being intuitive and not requiring too many user interactions to allow rapid navigation for the day-to-day heuristic usage. The requirements are categorized in five groups: (R1) pathway representation, (R2) data access, (R3) mapping and editing, (R4) data export and exchange, and (R5) availability.

6.3 Implementation

6.3.1 Pathway map

Several pathway maps are already commonly used by biologists. We therefore chose to utilize the layout of an existing familiar pathway map as the basis of our pathway browser rather than creating a new one. The KEGG Atlas was selected for several reasons: (1) a global integrated map is provided; (2) it is a part of one of the most popular pathway databases (Arakawa *et al.* 2005; Kono *et al.* 2006; Werner 2008); and (3) the reference pathway layout can be utilized for the representation of pathways in a wide variety of organisms. The KEGG database provides various tools and a wealth of pathway-related data that are curated with controlled identifiers and external references. These identifiers and references are useful for the implementation of many functions of the pathway browser and for the interoperability of the tool. Since the KEGG Atlas only represents metabolites as nodes *per se*, all related gene and enzyme nodes have been automatically added on the midpoint of reaction edges of the KEGG Atlas pathway map and a SVG file has been generated in Perl. To calculate the midpoints of edges, all quadratic bézier curves used for the representation of reaction edges were converted into polylines for computational efficiency. Following the automatic positioning of enzyme nodes, several nodes were manually curated using Adobe Illustrator CS3 13.0.2 for layout optimization. Enzyme nodes were partitioned into multiple compartments when several genes comprised the heteromeric enzyme, with a horizontal layout for genes ≤ 6 or with two rows of up to six columns for genes ≤ 12. A list of gene names only are shown for genes > 12. As a result, the reference pathway map contains 1,572 metabolite nodes and 1,813 enzyme nodes. As examples of the organism specific pathways, there are 1,365 gene nodes in *E. coli* and 2,883 in human.

In order to provide the pathway browser as cross platform software without the need for user maintenance and installation, Pathway Projector is implemented as a web application using HTML and JavaScript with the AJAX (Asynchronous JavaScript + XML) web development paradigm. The Ext JS 2.0 (http://extjs.com/) library was utilized to implement the overall user interface framework. ZUI for the large pathway map is implemented by using the Google Maps API for intuitive and familiar user interface and to take advantage of many online tools that are associated with the API. Since large images need to be split up into image tiles that are 256 × 256 pixels for use with the Google Maps API, the original large pathway map of 8,192 × 8,192 pixels was split using the "generateGMap" function available in the G-language GAE v.1.8.8 (Arakawa *et al.* 2003; Arakawa and Tomita 2006; Arakawa *et al.* 2008) to produce five zoom levels. The number of tiles increases by the power of 4 depending on the zoom level: 1 in level 0, 4 in level 1, 16 in level 2, 64 in level 3, 256 in level 4, and 1,024 in the maximum zoom level 5. The pathway map has been made clickable by sending the coordinate information and asynchronously retrieving related information upon the user mouse click event, which is displayed through the InfoWindow function of Google Maps API as an information window. The information window contains the retrieved annotation of each component, including the common name, identifier, structural formula or chemical equation, and links to external databases such as KEGG, PubChem (Sayers *et al.* 2009), ChEBI (Degtyarenko *et al.* 2008), and MSDchem (Dimitropoulos *et al.* 2006) for metabolites and, ExPASy (Kiefer *et al.* 2009), MetaCyc (Caspi *et al.* 2008), Brenda (Chang *et al.* 2009), IntEnz (Fleischmann *et al.* 2004), PUMA2 (Maltsev *et al.* 2006), and IUBMB (McDonald *et al.* 2009) for enzymes (see http://web.sfc.keio.ac.jp/~ciconia/AppendixTable1.eps for a complete listing for organism-specific database references). Although the default reference pathway only contains external links to enzymes, when an

organism specific pathway map is opened from the "Organism Selection" tab, the information window on the nodes also shows gene-centric links to suitable databases for the organism (*e.g.* EcoGene: Rudd 2000; MGI: Collins *et al.* 2007a; Collins *et al.* 2007b). Nodes within the pathway map show their corresponding labels (common names for genes and metabolites, EC number for enzymes) at zoom levels higher than 4, utilizing the semantic zooming capabilities of Google Maps API.

Organism-specific pathway maps for 843 species, including both eukaryotes and prokaryotes, were subsequently generated using the reference pathway map based on KEGG Orthology.

6.3.2 Search and data retrieval

Pathway Projector has four types of search functionalities: (1) by keyword and identifiers; (2) by molecular mass; (3) by computation of possible routes between two metabolites; and (4) by sequence similarity. Searches by keyword, identifier, and molecular mass were established by searching through a server-side database reconstructed from KEGG flatfile distributions. The results are listed in a search result panel and are also visually highlighted by red markers placed onto the respective components on the pathway map. Additional gray polylines are drawn on corresponding reaction edges when the search result points to enzymes or genes. The red markers can be clicked to invoke an information window to display more detailed annotations. The route search and similarity search capabilities are implemented as wrappers of PathComp (http://www.genome.ad.jp/kegg-bin/mk_pathcomp_html) and KEGG BLAST Search (http://blast.genome.jp/) to take advantage of existing KEGG services. Since PathComp calculates every possible route between two given metabolites from all combinations in the KEGG database, only the paths that are available within the KEGG Atlas layout are displayed in the results. For the BLAST search, sequence type (nucleotide or protein) and the type of BLAST program (blastn, blastp, blastx, tblastx), are automatically interpreted by the system, in which sequences comprising 80% A, T, G, C, or N are considered to be nucleotide sequences. The Pathway prediction tool reconstructs the pathway from given multi-FASTA sequence files using BLAT and SwissProt (Bairoch *et al.* 2009) using the GEM System (Arakawa *et al.* 2006).

6.3.3 Mapping and editing

The Pathway mapping tool modifies the SVG map based on user input, by changing the visibility, size, color, and labels of edges and nodes, and subsequently creates an overlay. Users can also place predefined icons or any image available in the World Wide Web by URLs and show the directionality of reactions by adding arrowheads to reaction edges. When values for time-series or multiple conditions are specified for nodes, graphs or charts generated by the Google Chart API (http://code.google.com/intl/en/apis/chart/) are displayed on the nodes. Quikmaps (http://quikmaps.com/) was utilized to implement manual annotation and editing capabilities.

6.4 Results and Discussion

6.4.1 Requirement analysis

The requirements are categorized in five groups: (R1) pathway representation, (R2) data access, (R3) mapping and editing, (R4) data export and exchange, and (R5) availability (see Table 6.4.2 for a summary).

(R1) Pathway representation

A biochemical pathway is often subdivided into smaller maps harboring specific biological processes, such as glycolysis, the TCA cycle, and the pentose phosphate pathway. Capturing the large systematic picture of cellular dynamics that spans several of these specific pathways, however, is difficult, especially in light of the availability of large-scale, comprehensive experimental data from high-throughput "-omics" measurements that encompass various pathways. An integrated pathway map that connects all subdivided pathways into a single large pathway is more suitable and intuitive as a context for information mapping rather than rotating through hundreds of specific maps.

While several pathway components are conserved among different organisms, each species also has its own specific genes, metabolites, and enzymes, and therefore has some unique set of pathways. While a reference pathway map with all known pathway components regardless of the organism, such as the one that is available in the KEGG database, may be a useful gateway, a researcher focusing, for example, on *E. coli* is primarily interested in the pathways of that organism, which makes all the other components dispensable. The availability of organism-specific pathway maps is therefore essential for this purpose, as well as for comparative studies among different organisms. Furthermore, it is desirable for a user to be able to reconstruct a pathway map from his/her own genomic sequence to keep up with the rapid availability of new genomes, both for comparative study and for the functional annotation of the genome.

Although the majority of existing pathway maps belong to the metabolic pathways, numerous aspects of cellular processes are formulated as pathways, including signal transduction, genetic information processing and maintenance, gene regulation, and the cell cycle. The availability of a drug resistance pathway map, for example, will facilitate drug discovery.

(R2) Data access

A pathway is a gateway to a wide variety of biological information, including genomic sequences, the functional annotations of genes, the biochemistry of enzymatic reactions, and the chemistry of metabolites. A pathway browser should, therefore, provide links to associated data in external public databases such as PathCase (Elliott *et al.* 2008). Various search capabilities are also essential for such highly complex and large-scale information in addition to simple query types, such as keywords and identifiers. The primary outcome from genomic, transcriptomic, and proteomic experiments is the sequence information, which demands sequence similarity searches that incorporate nucleotide or amino acid sequences. Likewise, metabolomic experiments produce spectrograms with molecular mass and retention times, from which corresponding compounds need to be identified.

When observing experimental data in the context of pathways, a biologist is often interested in a subset path or route within a complex pathway map. For example, in gene knockout experiments, biologists are often interested in the change in flux or gene expression within a given route as well as in the activation of alternative paths. Hence, computation and searches for possible routes between two given components are desirable for the observation of alternative paths and for the prediction of unidentified intermediate molecules that are difficult to measure by standard means.

(R3) Mapping and editing

An essential feature of a pathway browser is the ability to map and edit experimental data, which can involve changing the colors, sizes, and shapes of pathway graphics according to the experimental data or placing graphs of data onto the locations of respective entities, to visualize the data in the context of the biochemical pathway. Since high-throughput measurement technologies are available, large-scale experimental data from transcriptomics, proteomics, and metabolomics should be supported. For this reason, molecular components,

including genes, proteins, enzymes, and metabolites, should be represented and be available for mapping within the pathway map. While genes, proteins, and enzymes can usually be represented as single nodes or sometimes simply as edges in most existing pathway maps, heteromeric enzymes, which are formed from several proteins and, therefore, from several genes, require dedicated nodes to correctly map data from microarray experiments. Furthermore, experimental measurement is often performed on multiple conditions or in a time series, which requires the simultaneous mapping of multiple data to observe the changes and differences. These can be visualized, for example, by using animations that show pathway changes according to time as in the Omics Viewer (Paley and Karp 2006) or by displaying graphs on each object as in VANTED (Junker *et al.* 2006).

Naive mapping that is based upon a predefined pathway map does not allow novel pathways or entities to be mapped, which limits the applicability of a pathway browser to the evolving knowledge in molecular biology. For example, noncoding RNAs, as typified by micro RNAs (miRNA) or small nucleolar RNAs (snoRNA), or phosphorylated isoforms of proteins are emerging areas of researches that are actively being explored and could benefit from interpretation within a biochemical context. Therefore, the pathway map should be able to be freely edited and annotated by adding nodes, edges or data, as in WikiPathways (Pico *et al.* 2008).

(R4) Data export and exchange

For interoperability and data management, pathway data and mapping results should be downloadable in a standard XML image format, such as SBML (Systems Biology Markup Language: Hucka *et al.* 2004) or BioPAX (http://www.biopax.org).

(R5) Availability

In order to be platform-independent and interoperable without maintenance and installation efforts, a pathway browser should be available as a web-based application that requires no registration fees and that is freely available for academic users.

6.4.2 System overview

The Pathway Projector was implemented according to the aforementioned requirements, including the availability of a large-scale comprehensive pathway map, pathways from a wide variety of organisms, and searching and mapping capabilities. The software is freely available for academic users without any registration at http://www.g-language.org/PathwayProjector/. Since it has been implemented as a web application, this software is cross-platform and requires no installation or maintenance. Moreover, use of the AJAX web development paradigm provides an intuitive user experience similar to that of desktop applications. The main pathway map of Pathway Projector was reconstructed from the popular KEGG Atlas layout by adding nodes for enzymes and genes. As an example, the map for *E. coli* K12 contains 1,365 genes, 1,813 enzymes, and 1,572 metabolites (Figure 6.2). Circular and rectangular nodes represent metabolites and genes/enzymes, respectively, and the names of genes, enzymes, and metabolites are displayed within the map at high zoom levels by means of semantic zooming. Reaction edges are color coded according to the following pathway categories: aqua represents glycan biosynthesis and metabolism, blue represents carbohydrate metabolism, green represents lipid metabolism, red represents nucleotide metabolism, purple represents energy metabolism, yellow represents amino acid metabolism, pink represents metabolism of cofactors and vitamins, dark red represents biosynthesis of secondary metabolites, orange represents metabolism of other amino acids, and magenta represents biodegradation and metabolism of xenobiotics. Pathway Projector utilizes the Google Maps API for the implementation of ZUI and enables smooth navigation through panning and zooming without page transitions using a mouse scroll wheel or double clicks. Every component in the pathway map shows more detailed information by clicking on the nodes, which opens up an information window containing annotations, such as chemical and structural

formulas and links to external public databases (Figure 6.3). Detailed information about the nodes can be alternatively accessed through "Mouse Over Mode" that can be toggled from the right-most search result panel, with which users can simply move the mouse over to the nodes to show information in the sub-window located in the bottom panel. Users can therefore use Pathway Projector as a generic browser and as a gateway for various pathway-related resources.

The default pathway map of Pathway Projector is the reference pathway map, but organism-specific pathway maps can be selected from a list of 870 organisms available from the Organism Selection tab located in the upper left section of the user interface (Figure 6.3A). This list of organisms can be searched incrementally or can be sorted by species names, domains, kingdoms, and subphyla by clicking on the column headers. An organism-specific map is opened as a closable tab next to the reference pathway map tab upon double clicking on the desired row in the list. These maps show gene names within enzyme nodes, and the information window also contains organism specific database links, such as EcoCyc for *E. coli* (Keseler *et al.* 2009) and SGD for yeast (Hong *et al.* 2008). Since tab-browsing has been adopted as an effective approach in navigating the World Wide Web, as seen in numerous web browsers with this function, tab-browsing was utilized for the navigation of different organism-specific maps, which allows the user to quickly switch between species for comparative study.

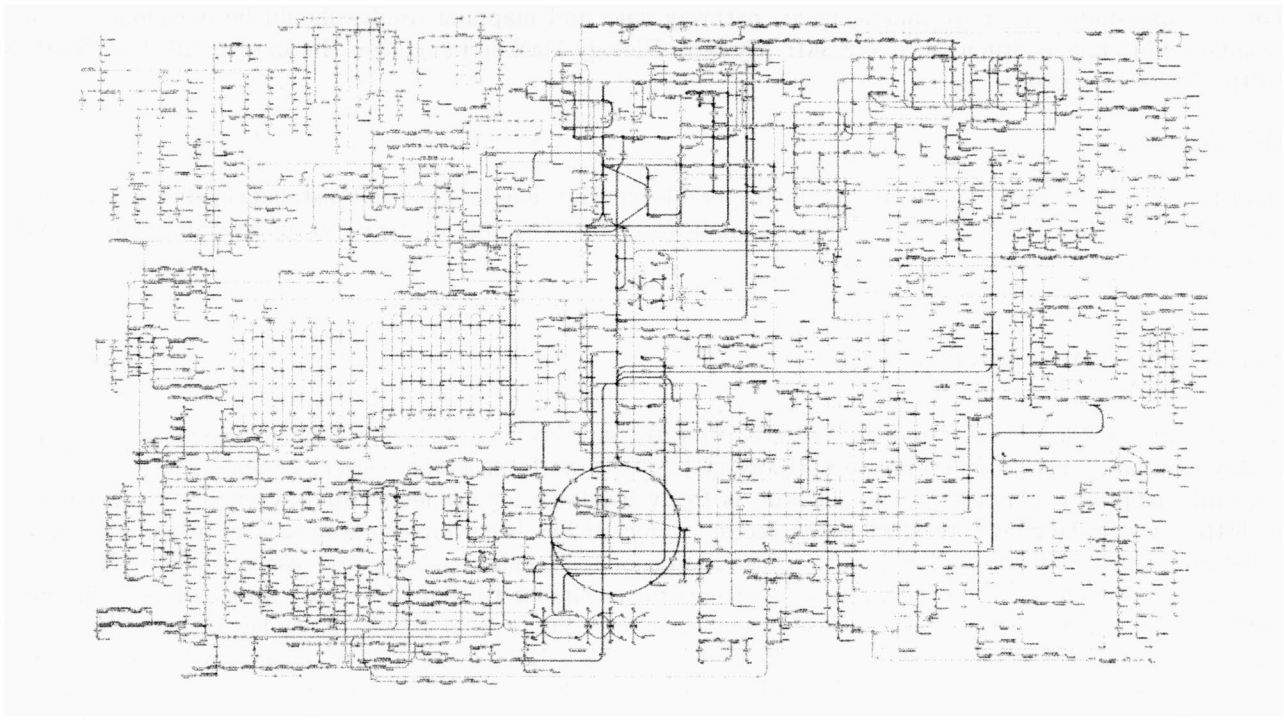

Figure 6.2: Reference pathway map.
Pathway Projector provides an integrated pathway map that is based upon the KEGG Atlas, with the addition of nodes for genes and enzymes. Circles represent metabolites, and rectangles represent enzymes that are further subdivided into several compartments indicating the composite genes for heteromeric enzymes. Nodes are labeled with names or EC numbers at high zoom levels.

Table 6.1: Comparison of existing pathway-related software and databases according to the requirement analysis.

	R1				R2					R3								R4		R5		
Software	Integrated pathway	Variety (coverage) of pathway	Availability of organism-specific pathway maps	Original pathway map / generate by genome	Keyword search	Search by sequence homology	Search by molecular weight	Route search	Links to external DB	Microarray data	Proteome data	Metabolome data	Color code	Multiple conditions and time-series data	Adding annotation by text	Adding drawings and diagrams	Export to BioPAX or SBML	Save as image file	Installation-free	Free for academics	No registration
Pathway Projector (this work)	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●
ArrayXPath II (Chung *et al.* 2005)			○						●	●			●	●		●		●	●	●	●
BioCarta *1		●	○		●				●							○		●	●	●	●
BioCyc (Karp *et al.* 2005)	●	●	●		●	●	●		●	○	●	●	●	○			●	●	●	●	●
BioCyc: Pathway Tools Software (Paley and Karp 2006)	●	●	●	●	●	●	●	●	●	○	●	●	●	○	●	●	●	●		●	
ExPASy Biochemical Pathway (Kiefer *et al.* 2009)	●	○			●				●										●	●	●
GenMAPP (Salomonis *et al.* 2007)		●	○	○	●				●	●	●		●	○			●	●		●	
Genome Projector (Arakawa *et al.* 2009b)	●		○	○	●	●			●	○	●		●						●	●	●
Ingenuity Pathway Analysis *2		●	○	○	●				●	●	●	●	●	●	●	●	●	●	●		
iPath (Letunic *et al.* 2008)	●		●						○	●	●	●	●					●	●	●	●
KEGG (Kanehisa *et al.* 2008)	●	●	●	●	●	●		●	●	○	●	●	●				●	●	●	●	●
KEGG Atlas (Okuda *et al.* 2008)	●		●		●				●									●	●	●	●
MEGU (Kono *et al.* 2006)	○	●	●	○					●	●	●	●	●					●	●	●	●
PathCase (Elliott *et al.* 2008)		●	●	○	●				●							●	●	●	●	●	●
Pathway Explorer (Mlecnik *et al.* 2005)		●	○		●				●	●	●		●					●	●	●	●
Reactome (Matthews *et al.* 2009)	●	●	○		●			●	●	○	●		●				●	●	●	●	●
Reactome β (Matthews *et al.* 2009)	○	●			●				●										●	●	●
VANTED (Junker *et al.* 2006)		●	●	●	●					●	●	●	●	●	●	●	●	●		●	●
WikiPathways (Pico *et al.* 2008)		●	○	○	●											●		●	●	●	○

This table summarizes the functions of existing pathway tools according to the requirements identified in this work: (R1) pathway representation, (R2) data access, (R3) mapping and editing, (R4) data export and exchange, and (R5) availability. Closed circles indicate satisfactory implementations, and open circles represent partial implementations.

*1 http://www.biocarta.com/

*2 http://www.ingenuity.com/

6.4.3 Search and retrieval

Users can search through the pathway components from the search box located in the upper right corner (Figure 6.3D) using keywords and identifiers for genes, enzymes, pathways, and metabolites, or with the molecular mass of metabolic compounds (default range within ± 10 mass number). Search results are directly shown within the pathway map, marked with red pins on the nodes as well as gray lines highlighting the corresponding edges for reactions. A list of search results is also available in the right-most panel (Figure 6.3E), where the search range for molecular mass can be adjusted. Clicking on a marker or an entry in the result list invokes an information window.

In order to identify the paths or the existence of alternative pathways, possible routes between two metabolites can be searched in Pathway Projector. Starting and ending metabolites for the route search can be selected from within the information window. The search results are displayed as lists in the search result panel, similar to keyword searches, and routes are highlighted after clicking on the route number. This feature is especially useful when observing the change in the flux distribution upon gene knockout or over-expression experiments to identify the existence of alternative pathways or for the prediction of the concentration of immeasurable metabolites from the changes in neighboring compounds.

A sequence similarity search using nucleotide or amino acid sequences has been implemented based upon KEGG BLAST and is displayed in a pop-up window that is opened by clicking on the "Tools" button located next to the search box (Figure 6.3C). The system automatically interprets the type of sequence (nucleotide or protein) and subsequently chooses the appropriate program (blastn, blastp, blastx, tblastx); therefore, the user only needs to paste in a sequence of interest to the text area and choose the type of database to run the BLAST search. The search results are marked with highlighted edges on the reference pathway and are also listed in the search result panel with KEGG Orthology identifiers, species names, e-values, and links to organism-specific pathway maps with the BLAST result. The "Pathway Prediction" tab, which is another tool that can be found in the "Tools" window, reconstructs the pathway from given multi-FASTA amino acid files and draws the resulting pathway map accompanied by corresponding e-values. Users can use this feature to analyze novel organisms that are not included in Pathway Projector or to analyze pathways in any given gene set, such as those included in an environmental metagenome database.

6.4.4 Pathway mapping of experimental data

The mapping feature of Pathway Projector is available from the "Tools" window, which allows full customization of pathway diagrams by changing the color, size or width, labels, and node image (specified by preset icons or URLs of images). In addition, directionality can be indicated by arrow heads, and graphs of multiple conditions or time series can be displayed (Figure 6.4). Users can map data from transcriptomic, proteomic, and metabolomic experiments, and multiple "-omics" data can be simultaneously represented on a single map. Because the node representing an enzyme is subdivided into multiple compartments when the enzyme is heteromeric and, therefore, comprised of several genes, transcriptome data can be correctly mapped onto individual genes. Entities are specified using the KEGG identifier (*e.g.* C00010), EC number (*e.g.* 1.1.1.1), and gene names, while the basal pathway map is specified by the KEGG Organism identifier (*e.g.* "eco" for *E. coli* K12). The graph of time-series or multi-condition data can also be visualized on top of the corresponding node by specifying the values as comma-separated vectors, and the graphs can be viewed at higher resolution by invoking the information window. Users can also alternatively upload a tab-delimited data file, where the first rows are component identifiers, and the columns are experimental data. By placing these graphs on the metabolic pathway, researchers can easily interpret complex multiple experiments in the context of biochemical pathways and subsequently identify the systematic response to perturbations. Map generation for Google Maps

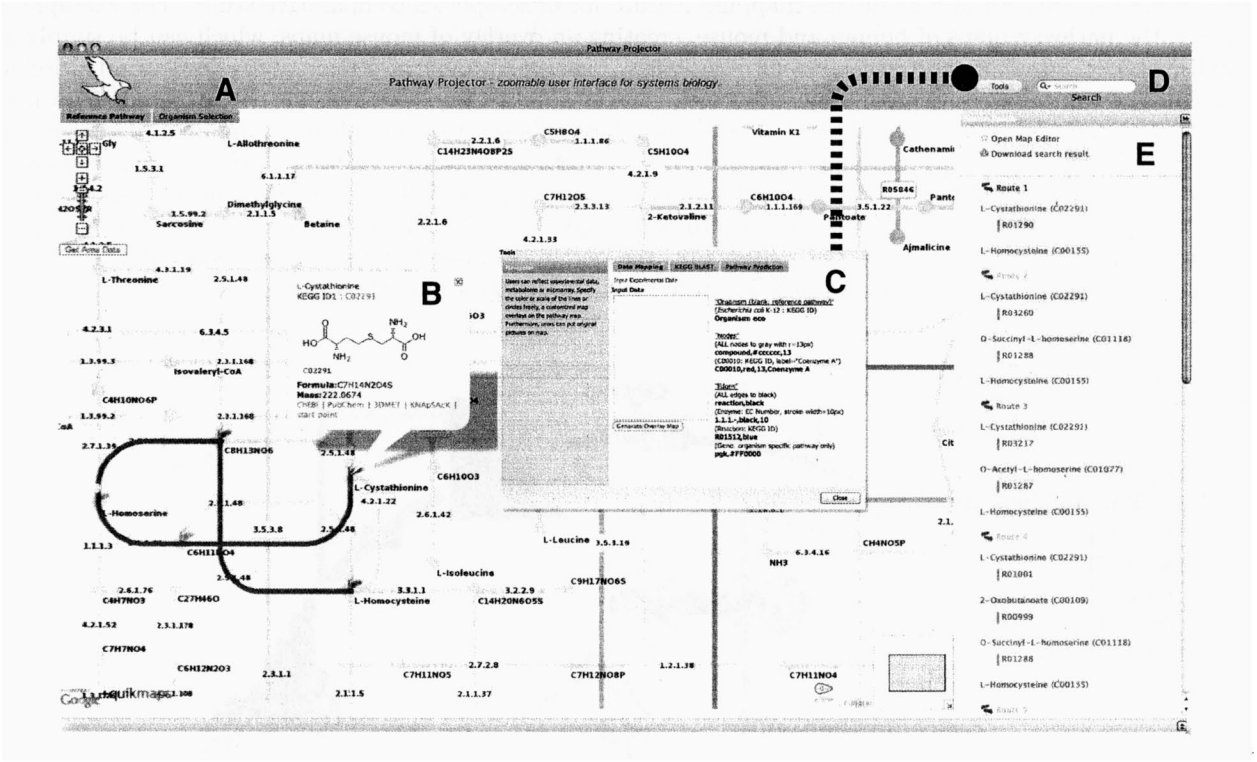

Figure 6.3: User interface.

A: Organism selection tab lists all available organism-specific pathways, which are opened as new tabs upon selection. B: The information window is opened by clicking on the entities represented in the map or on the markers that are shown as search results. This window shows detailed information about the selected entity, including names, images of structures, and molecular weight, and provides links to external databases. Furthermore, the selection of two metabolites as starting and ending compounds through this window results in the computation of possible paths between the two selected compounds. The result of path search is displayed in the right-most result panel and as highlighted lines on the map. C: Data mapping, sequence similarity searches, and pathway reconstructions based on sequence data, are available in a pop-up window that can be invoked from the "Tools" button. D: The search box located in the top-right corner automatically interprets the given query type and searches accordingly based on keywords, molecular mass, or identifiers. E: This panel displays the search results as a list. Users can locate the entities by opening an information window, which automatically moves the map to show the selected object in the center. Links to downloadable pathway images and editing and annotation palettes are located above the search results.

generally requires several minutes of computation time since thousands of tiled images must be prepared on the server; however, through multi-thread and multi-core optimization, Pathway Projector is able to generate the mapped image as Google Maps ZUI in less than 20 seconds on a dual quad-core CPU node. The mapped image is available as a downloadable image or as a transparent overlay, which can be toggled to display by pressing on the "Customized" button located in the upper right corner of the map. Because the mapping feature can specify the basal organism map and because it is implemented as an overlay of existing pathway maps, users can take advantage of this mapping feature for cross-species comparative study. For example, to compare the pathway maps of human and mouse, creating an overlay of mouse maps, which can be simply accomplished by creating a mapping layer with "Organism:mmu", on top of the human map allows rapid switching between the two pathways. Simple Object Access Protocol (SOAP) web service is also available for data mapping for interoperability with other tools. Users can access this service through programming languages such as Perl, Python and Ruby, or through intuitive SOAP clients with graphical user interfaces such as Taverna workbench (Hull *et al.* 2006). Web Service Description Language (WSDL) file is available at http://www.g-language.org/PathwayProjector/PathwayProjector.wsdl and detailed documentation as well as sample scripts for several programming languages are available in the online documentation, section annotation, subsection SOAP service (http://www.g-language.org/PathwayProjector/annotation.html#soap).

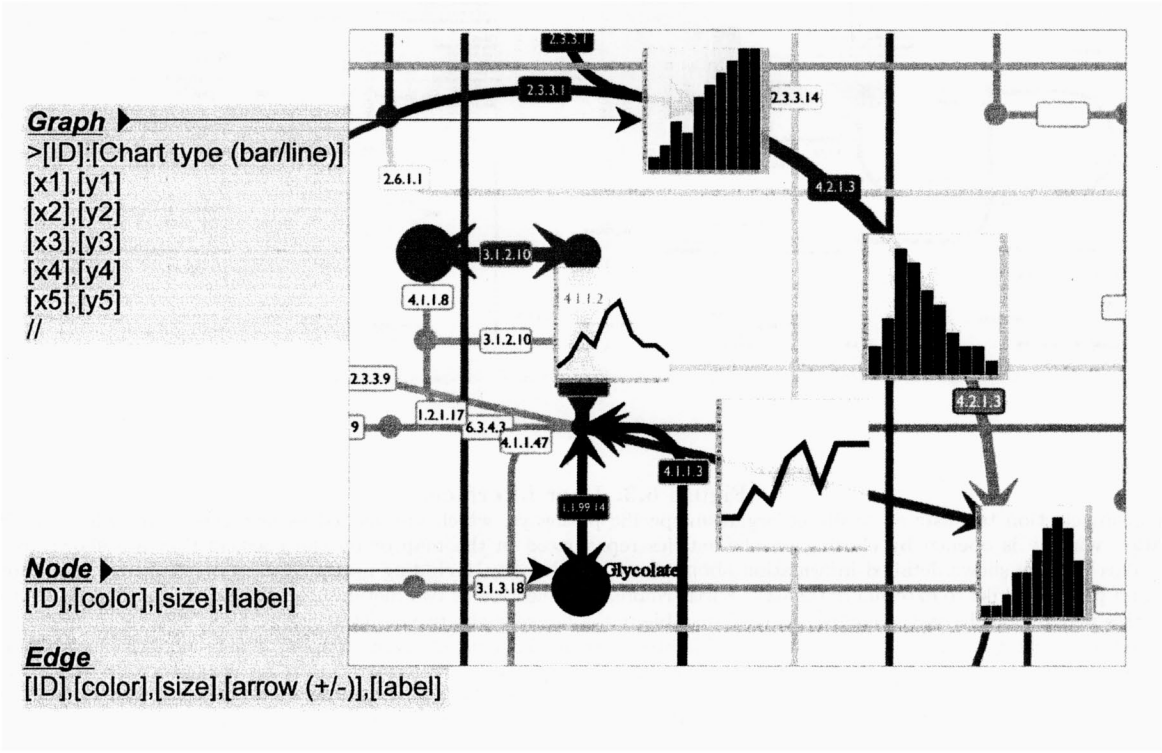

Figure 6.4: Data mapping.
An example result and required data formats for mapping are shown. For graph mapping, the target compound ID and chart type should be defined with a colon "ID:line". The graph data is written on the next line and "//" is placed on the final line. For node or edge mapping, user-generated data can be input line-by-line following such order by comma, ID, color, size, arrow, and label. Furthermore, details of the graph picture are shown when the graph is clicked.

Pathway Projector is also equipped with editing capabilities, such as drawing lines, placing icons and text labels, and adding annotations, to add novel findings that are not included in the reference pathway (Figure 6.5). The

editing tool is opened by clicking on the "Open Map Editor" link in the right-most panel. The line tool and scribble tool are available for drawing lines and curves, respectively, and the brush size and color for these tools can be configured. Text labels and icons can be dragged and dropped onto the map, and these objects can be clicked to show an editable information window in which the users can add custom annotations. The edited and annotated pathway map data are downloadable as XML files that can be saved and exchanged. This XML file is based on the Quikmaps format, and contains the editing log including the texts, coordinates, color and size of lines or icons. Hence, users can recapture the manually edited map by pasting this XML log and then clicking on the "import" button in a new session. This XML file can be shared among researchers to share the manual annotations.

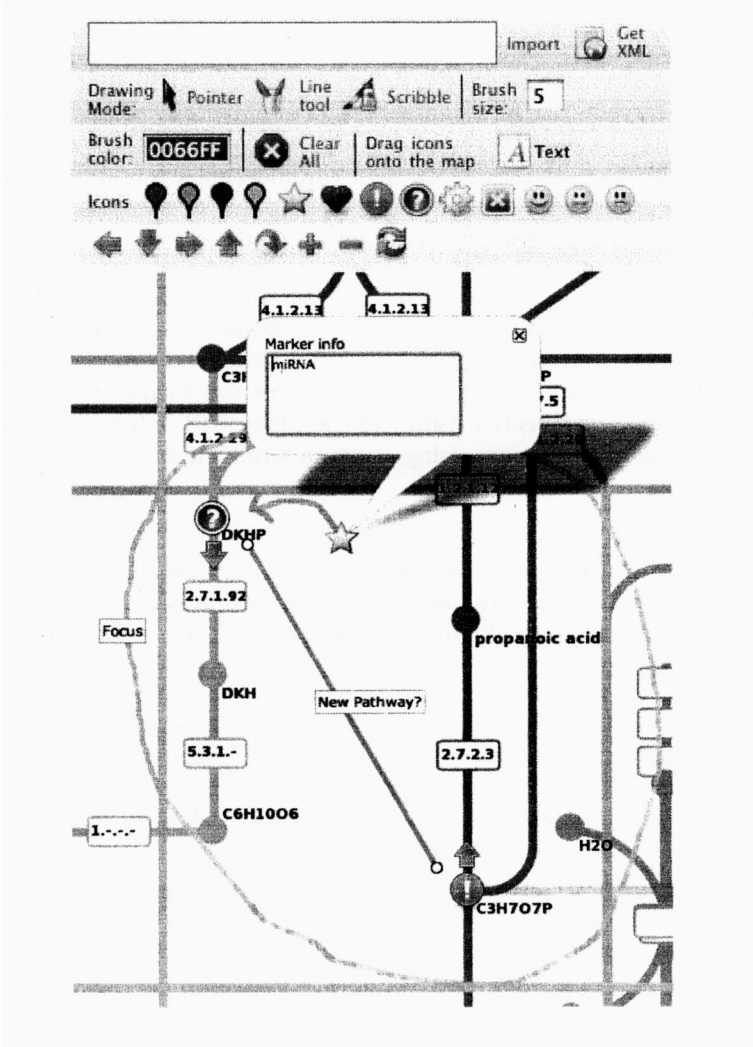

Figure 6.5: Manual editing and annotation of pathway maps.
The pathway editing and annotation palette can be invoked from the link located at the top of the search results panel. 21 pre-set icons are available to be dragged and dropped anywhere in the map, functioning as original markers. Users can freely move these markers around the map and can also add annotations and comments by clicking on the markers. Several other drawing features are available, including the line tool for drawing lines, scribble tool for free drawing, and text tool for placing labels. Brush color and size can be customized. Users can export the edited map as an XML file from the "Get XML" button, and this file can be shared and imported by other users to recapture the edited map.

6.5 Chromosome map

We developed a chromosome map browser as an extra feature of the Pathway Projector. Recognition of the positional relationship between gene directions, oligonucleotide sequence sites, the base compositional bias, and replication origin and terminus is important for understanding bacterial chromosome structures. However, since the localizations of genes or the positions of some sequence elements are identified numerically, only such numbers can not lead the instinctive recognition. Therefore, we implemented a viewer designed to map each genomic element on a circular chromosome map in order to develop an instinctive understanding and inspiration for subsequent research (Figure 6.6). This system is available at http://ws.g-language.org/g4/Repter.

(R'1) Chromosome map representation

Chromosome maps need to be large in size for the observation of individual genes because bacterial genomes are at least approximately 160,000 bp in size. However, the important point is that the minimum unit needed here is a gene, not a nucleotide. This is because the study of chromosome structure does not focus on promoters, start/stop codons, or other short segments that consist of just a few nucleotides. Therefore, the map requires a resolution of 1 base per 1 pixel.

Thus, we generated a large chromosome map adapted to the average number of genes to meet this requirement. For example, the number of genes in *E. coli* is approximately 4,000. Hence, the radius of the circular chromosome becomes 3,000 pixels, calculated as 4 pixels per gene. However, although researchers may be able to get a quick overview of such large images, they cannot circumstantially recognize each element in these maps. Therefore, these chromosome maps were also represented using Google Maps API, which enables smooth navigation via panning and zooming without page transitions using a mouse scroll wheel or double clicks.

(R'2) Data annotation

In Chapter 2, the chromosome structure was discussed. The term "chromosome structure" includes the gene strand bias, the base composition bias, and the positional relationship between several sequence elements (*dif* or Ter sites). Therefore, these annotations should be represented on the chromosome maps using many color variations.

In order to clarify the gene strand bias, blue bars representing the genes on the leading strand and orange bars representing the genes on the lagging strand were arranged in each independent concentric circle. The GC skew was represented by a red wavy line and located near the center of the circle. Additionally, several sequence elements were mapped onto the chromosome map using a red marker (*dif* site) or a yellow or purple marker (Ter sites) in each genome position. Finally, the replication origin and terminus were connected with a black line. As a result, this circular chromosome map enables the distinction of replichores, and users can intuitively observe the symmetric/asymmetric structure constructed by the gene strand bias, base composition bias, replication origin and terminus, and sequence elements.

(R'3) Map editing

Since the chromosome maps are generated based on GenBank data, the maps do not allow novel annotations or sequence elements to be mapped. For example, other sequence elements (*e.g.* KOPS, Chi sequence, AIMS, or other important sequences), some protein binding sites, horizontally acquired regions, and inverted regions represent information emerging from my research. Therefore, the chromosome map should be freely editable by adding icons, text, or drawings, as in Pathway Projector (R3).

This chromosome map browser is also equipped with editing capabilities to fulfill requests such as drawing

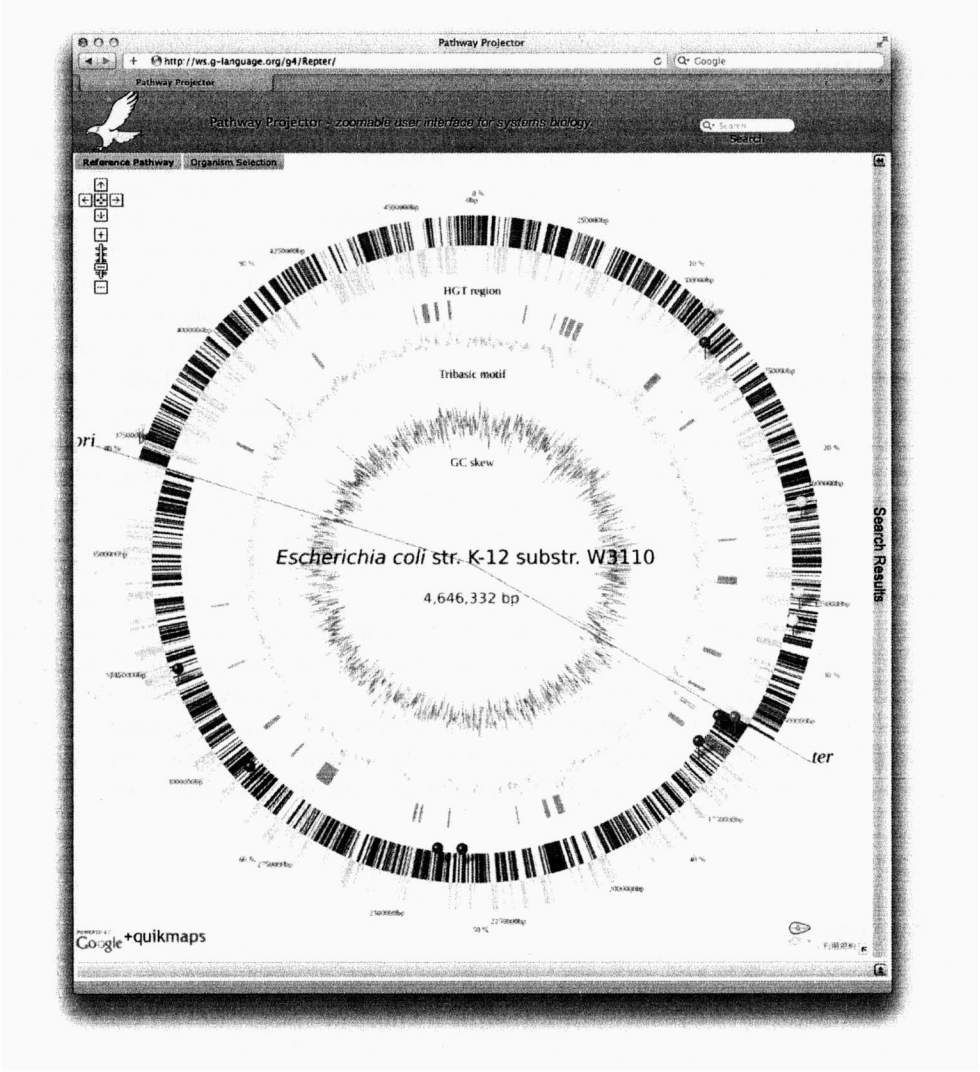

Figure 6.6: Chromosome map viewer.
This figure shows an overview of a circular chromosome map. The red pin indicates a *dif* site, the yellow pins indicate direct Ter sequences, and the purple pins indicate complement Ter sequences. The blue and orange lines represent genes in the leading and lagging strands, respectively. The green line represents the frequency of the tribasic motif, and the pink boxes represent horizontal gene transfer (HGT) regions. The *dif* sequence is located near the replication terminus (*ter*), the boundary line of the gene-strand bias, and the shift-point of the GC skew (red lines), as well as between the inverted Ter sequences (indicated by the yellow and purple pins).

lines, placing icons, and inserting text labels, as well as adding annotations, to add novel findings that are not included in the original browsed map. The detailed operative procedure for this editing function can be found in the subsection "Pathway mapping of experimental data."

6.6 Limitations

Pathway Projector is currently limited to the metabolic pathway, due to the availability of the global map in KEGG Atlas. Therefore, pathways such as signaling networks, gene regulation, and cell cycle are not supported by our software. Since our software tool is semi-automated to import information from KEGG, large-scale maps for these particular pathways, if they become available, will be relatively easy to integrate into Pathway Projector. Secondly, the KEGG pathway maps are generic implementations, for which the reference pathway serves as the net sum of all known reactions from a multitude of organisms. There are cases for which a specific pathway map dedicated to just a single species of interest is instead more desirable, such as the PMN (http://www.plantcyc.org/) for plants, EcoCyc for *E. coli*, and MGD (Bult *et al.* 2008) for mouse. Although links are provided to these databases for organism-specific information, we may need to consider using organism-specific maps for well-curated pathways in addition to those of KEGG, when integrated pathways become available in these websites. Nevertheless, the basic user interface frameworks based on the AJAX technology for searching and mapping, as well as the ZUI based on Google Maps API, are applicable for other pathway maps as long as the coordinate information is provided along with the map image. Finally, Pathway Projector does not provide pathway data in an exchangeable format such as SBML and BioPAX, as noted in the requirement analysis section. However, since our system is based on the KEGG maps, KGML (KEGG Markup Language) can be readily converted into the BioPAX format, and these BioPAX file can be downloaded from ftp://ftp.genome.jp:/pub/db/community/biopax. The map editing features provided in this software are intended for quick note taking for information exchange along researchers. While this feature allows free editing and annotations on the pathway maps, these editings are overlays of additional information, and are not true modifications to the existing pathway map. For the generation of static pathway map images, users should use the mapping tool to generate a custom pathway map, and download this image to make permanent modifications.

6.7 Conclusion

In light of the advent of high-throughput measurement technologies among several layers of "-omics," understanding the systematic workings of the intracellular activities is critical, and biochemical pathways provide an indispensable context for this purpose. Since pathways are essentially connected *in vivo*, the use of a global map is desirable, especially for the mapping of comprehensive experimental data. Pathway Projector is an intuitive browser that has been designed for this purpose by providing a large-scale metabolic pathway map based on the popular KEGG Atlas layout, with the addition of gene and enzyme nodes. Pathway Projector has been implemented with user-friendly interfaces while also providing its software as web-based application for cross-platform availability. Integrated search capabilities, as well as mapping and editing capabilities, facilitate exploratory and heuristic data analysis. Therefore, Pathway Projector can serve as a useful gateway for pathway information analysis.

CHAPTER 7

Conclusions

"Science is what you know, philosophy is what you do not know."

-Bertrand Russell

7.1 The role of scientific visualization

We humans can capture topological structures, such as nucleoids, with atomic force microscopy (AFM) or scanning force microscopy (SFM: van Noort *et al.* 2004; Dame 2005). However, we cannot capture biochemical pathways or genomic structures if they are not illustrated computationally as a map. The visualization of genomic information has enormous potential: it will contribute to the debate regarding the structures of chromosomes, particularly the symmetric structures. The symmetric structures, in this dissertation, are a phenomenon that uses the replication origin and terminus as a line-symmetric axis, and various types of elements are scattered along the rule of line-asymmetry. By marking the position of each element on a circular chromosome map, I was able to capture these relative positional relationships immediately. Thus, I implemented a specialized browser for my research that incorporated various genomic information (Chapter 6), and I was able to confirm several known or unknown findings. For example, the symmetric structures are not entirely conserved in all of the bacterial chromosomes. In particular, there is a rich diversity in the intensity of the GC skew and the gene-strand bias. Although this knowledge has been reported by previous researchers (Arakawa *et al.* 2009a), the illustrated chromosome map enabled us to view the degree of imbalance of the GC skew shift-points and gene localizations holistically. Additionally, when the predicted *dif* sites are mapped onto the chromosome map, I observed that almost all of the *dif* sites are located at or near the replication terminus, the boundary line of gene-strand bias, the shift-point of GC skew, or between the inverted Ter sequences (Figure 7.1). However, these positional relationships are not observed in all bacteria, and several *dif* sequences are located far from these positions. Based on this visual information, I questioned whether bacterial replication has terminated at close *dif* sites, or where it has occurred, at least during the process of forming GC skews. These feedbacks and interactions are among the results and ideas that may be generated by the extraction of information on an intuitive level from increasing amounts of data.

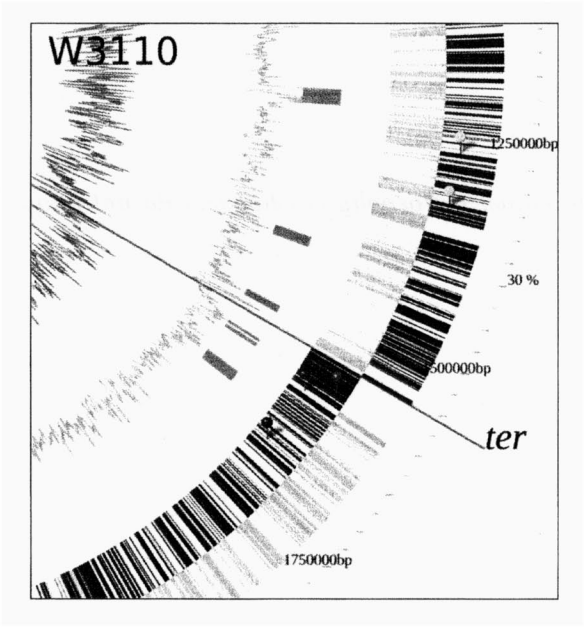

Figure 7.1: *dif* sites in a chromosome map.
This figure is a section of a circular chromosome map of *E. coli*. The red pin indicates a *dif* site, the yellow pins indicate direct Ter sequences, and the purple pins indicate complement Ter sequences. The blue and orange lines represent genes in the leading and lagging strands, respectively. The *dif* sequence is located near the replication terminus (*ter*), the boundary line of gene-strand bias, and the shift-point of GC skew (red lines), as well as between the inverted Ter sequences (indicated by the yellow and purple pins).

7.2 Chromosome structural aspects of bacterial evolution

The high predictability obtained using a recursive hidden Markov model suggested that the *dif*/XerCD system of chromosome dimer resolution was highly conserved among bacterial species and that the *dif* sequences were almost always conserved when XerCD was present within the genome. There may be many factors contributing to chromosome dimer resolution, but I presume that the circular structure of the chromosome is one of the most important considerations. When a recombination event occurs an odd number of times in a single DNA replication event, the replicated chromosome is not properly segregated into two daughter chromosomes but instead produces a concatenated dimer (Sherratt 2003; Lesterlin *et al.* 2004). However, if the bacteria have linear chromosomes, this problem does not occur. The bacteria with linear genomes do not need to have a chromosome dimer resolution mechanism because there are topological differences between circular and linear chromosomes. Therefore, because the bacterial chromosome forms a closed circular structure, a chromosome dimer resolution mechanism may become necessary.

Given this disadvantage, why do bacteria have circular chromosomes? A few bacteria have linear chromosomes (*e.g. Streptomyces* species; Volff and Altenbuchner 2000, and *Borrelia burgdorferi*; Ferdows and Barbour 1989), but almost all bacterial chromosomes are circular structures. In this study, I hypothesized that this structural evolution might be associated with the accuracy of copying the genetic information. The primal mechanism for the duplication of genetic information is a simple replication reaction that synthesizes nucleotides using other oligonucleotides as templates. The reaction between the nucleotides may be catalyzed by the substrate itself, and such enzymes were reported as yielding self-sustained replication (Lincoln and Joyce 2009). However, this primal duplication reaction is certainly not semiconservative replication based on self-replication. Semiconservative replication requires nucleotides that are complementary to the template segment when the hybridization of nucleotides occurs, but the hybridized nucleotides in the primal duplication reaction do not necessarily have the same length as the template oligonucleotide. Indeed, pyranosyl-RNA could be replicated by a shorter oligonucleotide than the template base sequence (Bolli *et al.* 1997). If the template oligonucleotides have certain important gene regions, such as polymerase or ligase (of course, in the primary genome, the 'gene' might not be established as a protein-coding region and might be nothing more than an architectural feature, similar to an active site of a ribozyme), then the primal organisms are exposed to the risk that part or all of these regions will not be duplicated. Hence, this system is insufficient for the accurate inheritance of genetic information, and the circular structure may be employed to resolve such length-independent problems. As a result of adopting a circular structure, because the oligonucleotides are always duplicated with the same length, the accuracy of the inheritance is ensured.

7.3 Biological validations

The *dif* sequences were conserved in almost all bacteria. This universal conservation indicated that bacterial replication may terminate at the *dif* site. It is desirable to terminate at the *dif* site for an effective procedure for resolving dimeric chromosomes. Moreover, this hypothesis is supported by many computational aspects based on the directions of oligonucleotide sequences and the correlations between GC skew shift-points and *dif* sites (Hendrickson and Lawrence 2007; Higgins 2007). Based on these previous studies, the *dif*-stop model was suggested as one of several replication termination models. Indeed, our comprehensive prediction method confirmed that the comparison of positions between the predicted *dif* sites and the shift-points of the GC skew showed that the *dif* sequences predicted in the phyla Proteobacteria and Firmicutes correlated significantly with the GC skew shift-points, which are highly likely to be located within the terminal region. However, although these two positions were highly correlated in terms of genomic loci, the differences in the positions of the GC skew shift-point and the GCSI were not correlated. Therefore, although the *dif* sequence is located near the

replication termination site for efficient CDR (chromosome dimmer resolution), the replication termination site was suggested to be at a site other than the *dif* site (Kono *et al.* 2011). Thus, the appropriate replication termination model needed to be validated.

The most fundamental problem in the understanding of the replication termination mechanism was that the positional information of the *dif* sequences had been unknown. In the three replication termination models described in Chapter 3, the fork-trap, fork-collision, and *dif*-stop models, it was not difficult to detect the termination positions predicted by the fork-trap and fork-collision models. The replication termini in the fork-trap model are Ter sites (the sites where Tus proteins bind). The Ter sites could be identified by their consensus sequence. Similarly, in the fork-collision model, replication terminates when the two oppositely directed replication forks meet by chance at the far end of the circular chromosome. Because the collision occurs randomly, the termination positions should follow a probabilistic distribution centered on the site opposite from the origin of replication. However, because the *dif* sites cannot be detected by a simple homology search, previous predictions have been limited to a few bacterial phyla (Val *et al.* 2008; Carnoy and Roten 2009). Therefore, the comprehensive prediction of *dif* sites in bacterial genomes enabled the validation of the bacterial replication termination models using simulations of genomic mutations. Accordingly, from the viewpoint of the replication-related mutational bias, I suggest that the fork-trap model is the most appropriate model for the replication termination system, not the *dif*-stop model. The validations of the replication termination models were also performed in the case in which the three models function together. When the probabilistic combination models were performed, the best probabilistic combination differed among bacterial species. Therefore, it was suggested that replication termination is multifactorial and associates with the genome size, gene density, and/or growth speed of each bacterium. This suggestion was also supported by the validation in Firmicutes, the members of which do not contain the Ter/Tus complex.

These investigations were limited to Proteobacteria or Firmicutes, but the results may also apply to other bacteria harboring clear GC skews. The most typical bacterium in this study is *E. coli*, which does not have a high GCSI score. Although the GCSI is not very high, the *dif*-stop model was not observed to be the most appropriate. A GCSI of 0.9 is a borderline score for accepting the existence of a clear GC skew. However, among the completely sequenced genomes, approximately 70% of the bacteria show GCSI scores greater than 0.9. Therefore, the only bottleneck of this restricted result is the Ter sequence, and our results which, the *dif*-stop model is not appropriate, may apply to almost all bacteria.

I had conducted rigorous examinations of the calculation methods for determining each termination rate, cautious constructions of the three termination models, and statistical comparisons. However, the construction of the *dif*-stop model should consider another role of the *dif* site, namely, its involvement in the integration and excision of exogenous DNA. At the *dif* site, two tyrosine recombinases, XerCD, resolve the dimeric chromosome (Clerget 1991; Blakely *et al.* 1993). At this time, XerCD can also function in the integration and excision of foreign DNA (Huber and Waldor 2002). The integration acts on bacteriophages, such as CTXΦ (Val *et al.* 2005) or VGJΦ (Das *et al.* 2011), and is promoted by XerCD. As a result, the integrations have the potential to result in the disruptions of chromosome proportion, regardless of the replication mechanism. Therefore, the integrated regions will be considered when researchers investigate the areas around *dif* sites. To resolve this problem, it is necessary to exclude the integration-reported bacteria from the target bacteria, and my research did not use such bacteria (*e.g. Vibrio* genus).

Not only these *dif*-related phage integrations but also HGTs are detected in several genomic positions. These regions are called GEIs and are considered to have the potential to invoke large-scale changes in base composition because they are large foreign regions of approximately 10 Kbp - 1 Mbp in the bacterial genome (Rocha 2008). Given a quantitative index to measure the strength of the genomic symmetries based on the GC skew, many bacteria with imbalanced *ori/ter* structures have been identified. It is highly unlikely that these detected imbalanced *ori/ter* structures were maintained in the actual bacterial genomes. Therefore, we investigated the

drastic effects on the base composition. However, regardless of the lengths or position of the HGTs, very few of the bacteria showed the symmetry disruptions caused by HGT. This result suggested the robustness of the genomic symmetries in bacterial genomes.

7.4 Concluding remarks

I have researched the symmetric structure defined by DNA replication and the topological structure defined by gene expression throughout this dissertation. What I began to see is that the living organism has conducted an 'adaptation to biological mechanism' as a part of the ways in which it changes throughout its evolution. The most important point is that this 'adaptation to biological mechanism' is different from the response to the environment and from the adaptability in PICERAS, an acronym that is known as the definition of life (Koshland 2002). The trigger is an internal environment rather than an external one. Replication and gene expression are key mechanisms of fundamental biological systems. Replication can only progress unidirectionally because it requires a free 3'-OH (hydroxyl) to attach the 5'-phosphate of the incoming nucleotide in both strands. Accordingly, the DNA synthesis mechanism introduces a difference between the leading and lagging strands, and there was a bias of the accumulative rates of mutation between strands. This bias results in GC skew as a part of the formation of the base compositional bias. Similarly, in the replication mechanism, the bacterial genomes have adapted to increase the efficiency of FtsK translocation on the DNA. FtsK progresses toward the replication termination region prior to the completion of cell division and functions as a key player in chromosome dimer resolution (Lesterlin *et al.* 2004). In exchange for the circular chromosome structure, which enhances the accurate inheritance of the genetic information, bacteria accept the risk of the formation of dimeric chromosomes by recombination. Because of that risk, almost all bacterial circular chromosomes would be expected to gain the *dif* site and XerCD in their genomes. Although the adaptational evolution of chromosome structure in biological mechanism does not yet occur at this point, such adaptation might occur to control FtsK translocation. Currently, this FtsK translocation has been oriented by KOPS (FtsK-orienting polar sequences). The KOPS are highly skewed in the genome (Bigot *et al.* 2005) and constitute a typical example of oligonucleotide bias. There are also contributions from other biological mechanisms, including octamer skews (Salzberg *et al.* 1998). The Chi sequence is known as a recombinational hot spot, and it interacts with RecBCD. When properly oriented, the Chi sequence controls recombinational repair (Taylor *et al.* 1985; Kuzminov 1995) in conjunction with the conformational changes that switch RecBCD function from exonuclease to recombinase (Dixon and Kowalczykowski 1995; Singleton *et al.* 2004; Handa *et al.* 2005). Therefore, approximately 75% of Chi sequences are oriented toward the direction of replication (Blattner *et al.* 1997). In gene expression, I found that the nucleoid-associated proteins avoid the frequently used codons in essential or highly expressed genes as their target binding sites. Although the repressive function of H-NS toward the expression of particular genes has been described as a genome sentinel (Dorman 2007), I considered the possibility that the genome sentinel function might be caused as a byproduct of avoiding the silencing of important genes. Of course, the influence of the avoidance alters the binding site of NAPs and affects the topological chromosome structure. These adaptations to biological mechanism are as slow as the adaptability and response to environment seen in PICERAS but will certainly accumulate gradually. The observation of the adaptation to biological mechanism will support the understanding of the importance of each mechanism and lead to the 'recognition' of life.

Bibliography

Adriaens, M.E., Jaillard, M., Waagmeester, A., Coort, S.L., Pico, A.R. and Evelo, C.T. (2008) The public road to high-quality curated biological pathways. *Drug Discov Today* **13**:856-862.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**:3389-3402.

Arakawa, K., Kono, N., Yamada, Y., Mori, H. and Tomita, M. (2005) KEGG-based pathway visualization tool for complex omics data. *In Silico Biol* **5**:419-423.

Arakawa, K., Mori, K., Ikeda, K., Matsuzaki, T., Kobayashi, Y. and Tomita, M. (2003) G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. *Bioinformatics* **19**:305-306.

Arakawa, K., Saito, R. and Tomita, M. (2007) Noise-reduction filtering for accurate detection of replication termini in bacterial genomes. *FEBS Lett* **581**:253-258.

Arakawa, K., Suzuki, H. and Tomita, M. (2008) Computational genome analysis using the G-language system. *Genes, Genomes and Genomics* **2**:1-13.

Arakawa, K., Suzuki, H. and Tomita, M. (2009a) Quantitative analysis of replication-related mutation and selection pressures in bacterial chromosomes and plasmids using generalised GC skew index. *BMC Genomics* **10**:640.

Arakawa, K., Tamaki, S., Kono, N., Kido, N., Ikegami, K., Ogawa, R. and Tomita, M. (2009b) Genome Projector: zoomable genome map with multiple views. *BMC Bioinformatics* **10**:31.

Arakawa, K. and Tomita, M. (2006) G-language System as a platform for large-scale analysis of high-throughput omics data. *J Pestic Sci* **31**:282-288.

Arakawa, K. and Tomita, M. (2007a) The GC skew index: a measure of genomic compositional asymmetry and the degree of replicational selection. *Evol Bioinform Online* **3**:159-168.

Arakawa, K. and Tomita, M. (2007b) Selection effects on the positioning of genes and gene structures from the interplay of replication and transcription in bacterial genomes. *Evol Bioinform Online* **3**:279-286.

Arakawa, K., Yamada, Y., Shinoda, K., Nakayama, Y. and Tomita, M. (2006) GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes. *BMC Bioinformatics* **7**:168.

Aurrecoechea, C., Brestelli, J., Brunk, B.P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J.C., Kraemer, E., Li, W., Miller, J.A., Nayak, V., Pennington, C., Pinney, D.F., Roos, D.S., Ross, C., Stoeckert, C.J., Jr., Treatman, C. and Wang, H. (2009) PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res* **37**:D539-543.

Aussel, L., Barre, F.X., Aroyo, M., Stasiak, A., Stasiak, A.Z. and Sherratt, D. (2002) FtsK Is a DNA motor protein that activates chromosome dimer resolution by switching the catalytic state of the XerC and XerD recombinases. *Cell* **108**:195-205.

Azam, T.A. and Ishihama, A. (1999) Twelve species of the nucleoid-associated protein from Escherichia coli. Sequence recognition specificity and DNA binding affinity. *J Biol Chem* **274**:33105-33113.

Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**:28-36.

Bailey, T.L. and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**:48-54.

Bairoch, A., Bougueleret, L., Altairac, S., Amendolia, V., Auchincloss, A., Argoud-Puy, G., Axelsen, K., Baratin, D., Blatter, M.C., Boeckmann, B., Bolleman, J., Bollondi, L., Boutet, E., Quintaje, S.B., Breuza, L., Bridge, A., deCastro, E., Ciapina, L., Coral, D., Coudert, E., Cusin, I., Delbard, G., Dornevil, D., Roggli, P.D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gehant, S., Farriol-Mathis, N., Ferro, S., Gasteiger, E., Gateau, A., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hulo, N., James, J., Jimenez, S., Jungo, F., Junker, V., Kappler, T., Keller, G., Lachaize, C., Lane-Guermonprez, L., Langendijk-Genevaux, P., Lara, V., Lemercier, P., Le Saux, V., Lieberherr, D., Lima, T.O., Mangold, V., Martin, X., Masson, P., Michoud, K., Moinat, M., Morgat, A., Mottaz, A., Paesano, S., Pedruzzi, I., Phan, I., Pilbout, S., Pillet, V., Poux, S., Pozzato, M., Redaschi, N., Reynaud, S., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A.L., Yip L., Zuletta, L., Apweiler, R., Alam-Faruque, Y., Antunes, R., Barrell, D., Binns, D., Bower, L., Browne, P., Chan, W.M., Dimmer, E., Eberhardt, R., Fedotov, A., Foulger, R., Garavelli, J., Golin, R., Horne, A., Huntley, R., Jacobsen, J., Kleen, M., Kersey, P., Laiho, K., Leinonen, R., Legge, D., Lin, Q., Magrane, M., Martin, M.J., O'Donovan, C., Orchard, S., O'Rourke, J., Patient, S., Pruess, M., Sitnov, A., Stanley, E., Corbett, M., di Martino, G., Donnelly, M., Luo, J., van Rensburg, P., Wu, C., Arighi, C., Arminski, L., Barker, W., Chen, Y., Hu, Z.Z., Hua, H.K., Huang, H., Mazumder, R., McGarvey, P., Natale, D.A., Nikolskaya, A., Petrova, N., Suzek, B.E., Vasudevan, S., Vinayaka, C.R., Yeh, L.S. and Zhang, J. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* **37**:D169-174.

Banos, R.C., Pons, J.I., Madrid, C. and Juarez, A. (2008) A global modulatory role for the Yersinia enterocolitica H-NS protein. *Microbiology* **154**:1281-1289.

Barre, F.X., Aroyo, M., Colloms, S.D., Helfrich, A., Cornet, F. and Sherratt, D.J. (2000) FtsK functions in the processing of a Holliday junction intermediate during bacterial chromosome segregation. *Genes Dev* **14**:2976-2988.

Bernardi, G. (1989) The isochore organization of the human genome. *Annu Rev Genet* **23**:637-661.

Bernardi, G. (1995) The human genome: organization and evolutionary history. *Annu Rev Genet* **29**:445-476.

Bigot, S., Corre, J., Louarn, J.M., Cornet, F. and Barre, F.X. (2004) FtsK activities in Xer recombination, DNA mobilization and cell division involve overlapping and separate domains of the protein. *Mol Microbiol* **54**:876-886.

Bigot, S., Saleh, O.A., Cornet, F., Allemand, J.F. and Barre, F.X. (2006) Oriented loading of FtsK on KOPS. *Nat Struct Mol Biol* **13**:1026-1028.

Bigot, S., Saleh, O.A., Lesterlin, C., Pages, C., El Karoui, M., Dennis, C., Grigoriev, M., Allemand, J.F., Barre, F.X. and Cornet, F. (2005) KOPS: DNA motifs that control E. coli chromosome segregation by orienting the FtsK translocase. *EMBO J* **24**:3770-3780.

Blakely, G., May, G., McCulloch, R., Arciszewska, L.K., Burke, M., Lovett, S.T. and Sherratt, D.J. (1993) Two related recombinases are required for site-specific recombination at dif and cer in E. coli K12. *Cell* **75**:351-361.

Blakely, G.W. (2004) Smarter than the average phage. *Mol Microbiol* **54**:851-854.

Blakely, G.W. and Sherratt, D.J. (1994) Interactions of the site-specific recombinases XerC and XerD with the recombination site dif. *Nucleic Acids Res* **22**:5613-5620.

Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) The complete genome sequence of Escherichia coli K-12. *Science* **277**:1453-1462.

Bolli, M., Micura, R. and Eschenmoser, A. (1997) Pyranosyl-RNA: chiroselective self-assembly of base sequences by ligative oligomerization of tetranucleotide-2',3'-cyclophosphates (with a commentary concerning the origin of biomolecular homochirality). *Chem Biol* **4**:309-320.

Bonne, L., Bigot, S., Chevalier, F., Allemand, J.F. and Barre, F.X. (2009) Asymmetric DNA requirements in Xer recombination activation by FtsK. *Nucleic Acids Res* **37**:2371-2380.

Bowes, J.B., Snyder, K.A., Segerdell, E., Gibb, R., Jarabek, C., Noumen, E., Pollet, N. and Vize, P.D. (2008) Xenbase: a Xenopus biology and genomics resource. *Nucleic Acids Res* **36**:D761-767.

Breier, A.M., Weier, H.U. and Cozzarelli, N.R. (2005) Independence of replisomes in Escherichia coli chromosomal replication. *Proc Natl Acad Sci U S A* **102**:3942-3947.

Brinza, L., Vinuelas, J., Cottret, L., Calevro, F., Rahbe, Y., Febvay, G., Duport, G., Colella, S., Rabatel, A., Gautier, C., Fayard, J.M., Sagot, M.F. and Charles, H. (2009) Systemic analysis of the symbiotic function of Buchnera aphidicola, the primary endosymbiont of the pea aphid Acyrthosiphon pisum. *C R Biol* **332**:1034-1049.

Browning, D.F. and Busby, S.J. (2004) The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2**:57-65.

Bruford, E.A., Lush, M.J., Wright, M.W., Sneddon, T.P., Povey, S. and Birney, E. (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res* **36**:D445-448.

Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E. and Blake, J.A. (2008) The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res* **36**:D724-728.

Campo, N., Dias, M.J., Daveran-Mingot, M.L., Ritzenthaler, P. and Le Bourgeois, P. (2004) Chromosomal constraints in Gram-positive bacteria revealed by artificial inversions. *Mol Microbiol* **51**:511-522.

Campos, J., Martinez, E., Izquierdo, Y. and Fando, R. (2010) VEJphi, a novel filamentous phage of Vibrio cholerae able to transduce the cholera toxin genes. *Microbiology* **156**:108-115.

Campos, J., Martinez, E., Suzarte, E., Rodriguez, B.L., Marrero, K., Silva, Y., Ledon, T., del Sol, R. and Fando, R. (2003) VGJ phi, a novel filamentous phage of Vibrio cholerae, integrates into the same chromosomal site as CTX phi. *J Bacteriol* **185**:5685-5696.

Carnoy, C. and Roten, C.A. (2009) The dif/Xer recombination systems in proteobacteria. *PLoS ONE* **4**:e6531.

Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S.Y., Shearer, A.G., Tissier, C., Walk, T.C., Zhang, P. and Karp, P.D. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **36**:D623-631.

Cathelyn, J.S., Ellison, D.W., Hinchliffe, S.J., Wren, B.W. and Miller, V.L. (2007) The RovA regulons of Yersinia enterocolitica and Yersinia pestis are distinct: evidence that many RovA-regulated genes were acquired more recently than the core genome. *Mol Microbiol* **66**:189-205.

Chambaud, I., Heilig, R., Ferris, S., Barbe, V., Samson, D., Galisson, F., Moszer, I., Dybvig, K., Wroblewski, H., Viari, A., Rocha, E.P. and Blanchard, A. (2001) The complete genome sequence of the murine respiratory pathogen Mycoplasma pulmonis. *Nucleic Acids Res* **29**:2145-2153.

Chang, A., Scheer, M., Grote, A., Schomburg, I. and Schomburg, D. (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res* **37**:D588-592.

Chargaff, E. (1951) Structure and function of nucleic acids as cell constituents. *Fed Proc* **10**:654-659.

Chisholm, R.L., Gaudet, P., Just, E.M., Pilcher, K.E., Fey, P., Merchant, S.N. and Kibbe, W.A. (2006) dictyBase, the model organism database for Dictyostelium discoideum. *Nucleic Acids Res* **34**:D423-427.

Chung, H.J., Park, C.H., Han, M.R., Lee, S., Ohn, J.H., Kim, J. and Kim, J.H. (2005) ArrayXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids Res* **33**:W621-626.

Chyba, C.F. and McDonald, G.D. (1995) The origin of life in the solar system: current issues. *Annu Rev Earth Planet Sci* **23**:215-249.

Clerget, M. (1991) Site-specific recombination promoted by a short DNA segment of plasmid R1 and by a homologous segment in the terminus region of the Escherichia coli chromosome. *New Biol* **3**:780-788.

Collins, F.S., Finnell, R.H., Rossant, J. and Wurst, W. (2007a) A new partner for the international knockout mouse consortium. *Cell* **129**:235.

Collins, F.S., Rossant, J. and Wurst, W. (2007b) A mouse for all reasons. *Cell* **128**:9-13.

Cornet, F., Louarn, J., Patte, J. and Louarn, J.M. (1996) Restriction of the activity of the recombination site dif to a small zone of the Escherichia coli chromosome. *Genes Dev* **10**:1152-1161.

Coskun-Ari, F.F. and Hill, T.M. (1997) Sequence-specific interactions in the Tus-Ter complex and the effect of base pair substitutions on arrest of DNA replication in Escherichia coli. *J Biol Chem* **272**:26448-26456.

Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. (1978) Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274**:775-780.

Crick, F. (1970) Central dogma of molecular biology. *Nature* **227**:561-563.

Crick, F.H. (1958) On protein synthesis. *Symp Soc Exp Biol* **12**:138-163.

Cropp, S., Boinski, S. and Li, W.H. (2002) Allelic variation in the squirrel monkey x-linked color vision gene: biogeographical and behavioral correlates. *J Mol Evol* **54**:734-745.

Cunha, S., Odijk, T., Suleymanoglu, E. and Woldringh, C.L. (2001) Isolation of the Escherichia coli nucleoid. *Biochimie* **83**:149-154.

da Silva, A.C., Ferro, J.A., Reinach, F.C., Farah, C.S., Furlan, L.R., Quaggio, R.B., Monteiro-Vitorello, C.B., Van Sluys, M.A., Almeida, N.F., Alves, L.M., do Amaral, A.M., Bertolini, M.C., Camargo, L.E., Camarotte, G., Cannavan, F., Cardozo, J., Chambergo, F., Ciapina, L.P., Cicarelli, R.M., Coutinho, L.L., Cursino-Santos, J.R., El-Dorry, H., Faria, J.B., Ferreira, A.J., Ferreira, R.C., Ferro, M.I., Formighieri, E.F., Franco, M.C., Greggio, C.C., Gruber, A., Katsuyama, A.M., Kishi, L.T., Leite, R.P., Lemos, E.G., Lemos, M.V., Locali, E.C., Machado, M.A., Madeira, A.M., Martinez-Rossi, N.M., Martins, E.C., Meidanis, J., Menck, C.F., Miyaki, C.Y., Moon, D.H., Moreira, L.M., Novo, M.T., Okura, V.K., Oliveira, M.C., Oliveira, V.R., Pereira, H.A., Rossi, A., Sena, J.A., Silva, C., de Souza, R.F., Spinola, L.A., Takita, M.A., Tamura, R.E., Teixeira, E.C., Tezza, R.I., Trindade dos Santos, M., Truffi, D., Tsai, S.M., White, F.F., Setubal, J.C. and Kitajima, J.P. (2002) Comparison of the genomes of two Xanthomonas pathogens with differing host specificities. *Nature* **417**:459-463.

Dame, R.T. (2005) The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Mol Microbiol* **56**:858-870.

Das, B., Bischerour, J. and Barre, F.X. (2011) VGJphi integration and excision mechanisms contribute to the genetic diversity of Vibrio cholerae epidemic strains. *Proc Natl Acad Sci U S A* **108**:2516-2521.

Das, B., Bischerour, J., Val, M.E. and Barre, F.X. (2010) Molecular keys of the tropism of integration of the cholera toxin phage. *Proc Natl Acad Sci U S A* **107**:4377-4382.

de Duve, C. (1991) *Blueprint for a Cell: The Nature and Origin of Life.* Carolina Biological Supply Company, Burlington.

de Massy, B., Bejar, S., Louarn, J., Louarn, J.M. and Bouche, J.P. (1987) Inhibition of replication forks exiting the terminus region of the Escherichia coli chromosome occurs at two loci separated by 5 min. *Proc Natl Acad Sci U S A* **84**:1759-1763.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M. and Ashburner, M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* **36**:D344-350.

DePamphilis, M.L., Blow, J.J., Ghosh, S., Saha, T., Noguchi, K. and Vassilev, A. (2006) Regulating the licensing of DNA replication origins in metazoa. *Curr Opin Cell Biol* **18**:231-239.

Derbise, A., Chenal-Francisque, V., Pouillot, F., Fayolle, C., Prevost, M.C., Medigue, C., Hinnebusch, B.J. and Carniel, E. (2007) A horizontally acquired filamentous phage contributes to the pathogenicity of the plague bacillus. *Mol Microbiol* **63**:1145-1157.

Diffley, J.F. (2004) Regulation of early events in chromosome replication. *Curr Biol* **14**:R778-786.

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

Dimitropoulos, D., Ionides, J. and Henrick, K. (2006) Using MSDchem to search the PDB ligand dictionary. *Curr Protoc Bioinformatics* **Chapter 14**:Unit14.3.

Dixon, D.A. and Kowalczykowski, S.C. (1995) Role of the Escherichia coli recombination hotspot, chi, in RecABCD-dependent homologous pairing. *J Biol Chem* **270**:16360-16370.

Dorman, C.J. (2004) H-NS: a universal regulator for a dynamic genome. *Nat Rev Microbiol* **2**:391-400.

Dorman, C.J. (2007) H-NS, the genome sentinel. *Nat Rev Microbiol* **5**:157-161.

Dorman, C.J. and Kane, K.A. (2009) DNA bridging and antibridging: a role for bacterial nucleoid-associated proteins in regulating the expression of laterally acquired genes. *FEMS Microbiol Rev* **33**:587-592.

dos Reis, M., Savva, R. and Wernisch, L. (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**:5036-5044.

Dubarry, N. and Barre, F.X. (2010) Fully efficient chromosome dimer resolution in Escherichia coli cells lacking the integral membrane domain of FtsK. *EMBO J* **29**:597-605.

Duggin, I.G. and Bell, S.D. (2009) Termination structures in the Escherichia coli chromosome replication fork trap. *J Mol Biol* **387**:532-539.

Duggin, I.G., Wake, R.G., Bell, S.D. and Hill, T.M. (2008) The replication fork trap and termination of chromosome replication. *Mol Microbiol* **70**:1323-1333.

Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* **14**:755-763.

Edenberg, H.J. and Huberman, J.A. (1975) Eukaryotic chromosome replication. *Annu Rev Genet* **9**:245-284.

Edlund, T., Gustafsson, P. and Wolf-Watz, H. (1976) Effect of thymine concentration on the mode of chromosomal replication in Escherichia coli K-12. *J Mol Biol* **108**:295-303.

Ekins, S., Nikolsky, Y., Bugrim, A., Kirillov, E. and Nikolskaya, T. (2007) Pathway mapping tools for analysis of high content data. *Methods Mol Biol* **356**:319-350.

Elliott, B., Kirac, M., Cakmak, A., Yavas, G., Mayes, S., Cheng, E., Wang, Y., Gupta, C., Ozsoyoglu, G. and Meral Ozsoyoglu, Z. (2008) PathCase: pathways database system. *Bioinformatics* **24**:2526-2533.

Falconi, M., Gualtieri, M.T., La Teana, A., Losso, M.A. and Pon, C.L. (1988) Proteins from the prokaryotic nucleoid: primary and quaternary structure of the 15-kD Escherichia coli DNA binding protein H-NS. *Mol Microbiol* **2**:323-329.

Felsenstein, J. (1989) PHYLIP: Phylogeny Inference Package. *Cladistics* **5**:164-166.

Ferdows, M.S. and Barbour, A.G. (1989) Megabase-sized linear DNA in the bacterium Borrelia burgdorferi, the Lyme disease agent. *Proc Natl Acad Sci U S A* **86**:5969-5973.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. and Bateman, A. (2008) The Pfam protein families database. *Nucleic Acids Res* **36**:D281-288.

Fisher, R.A. and Yates, F. (1948) *Statistical tables for biological, agricultural and medical research.* Oliver & Boyd, Edinburgh.

Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K.B., Bairoch, A., Schomburg, D., Tipton, K.F. and Apweiler, R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res* **32**:D434-437.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocyne, J.D., Scott, J., Shirley, R., Liu, L.I., Glodek, A., Kelley J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O. and Venter, J.C. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**:496-512.

Flicek, P. and Birney, E. (2009) Sense from sequence reads: methods for alignment and assembly. *Nat Methods* **6**:S6-12.

Francino, M.P., Chao, L., Riley, M.A. and Ochman, H. (1996) Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science* **272**:107-109.

Francino, M.P. and Ochman, H. (1997) Strand asymmetries in DNA evolution. *Trends Genet* **13**:240-245.

Frank, A.C. and Lobry, J.R. (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**:65-77.

Frank, A.C. and Lobry, J.R. (2000) Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* **16**:560-561.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, R.D., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.F., Dougherty, B.A., Bott, K.F., Hu, P.C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison, C.A., 3rd and Venter, J.C. (1995) The minimal gene complement of Mycoplasma genitalium. *Science* **270**:397-403.

Friddle, R.W., Klare, J.E., Martin, S.S., Corzett, M., Balhorn, R., Baldwin, E.P., Baskin, R.J. and Noy, A. (2004) Mechanism of DNA compaction by yeast mitochondrial protein Abf2p. *Biophys J* **86**:1632-1639.

Galperin, M.Y. and Fernandez-Suarez, X.M. (2012) The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res* **40**:D1-8.

Gama-Castro, S., Jimenez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M.I., Contreras-Moreira, B., Segura-Salazar, J., Muniz-Rascado, L., Martinez-Flores, I., Salgado, H., Bonavides-Martinez, C., Abreu-Goodger, C., Rodriguez-Penagos, C., Miranda-Rios, J., Morett, E., Merino, E., Huerta, A.M., Trevino-Quintanilla, L. and Collado-Vides, J. (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* **36**:D120-124.

Gánti, T. (2003) *The Principles of Life.* Oxford University Press, Oxford.

Gao, F. and Zhang, C.T. (2007) DoriC: a database of oriC regions in bacterial genomes. *Bioinformatics* **23**:1866-1867.

Gogarten, J.P. and Townsend, J.P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* **3**:679-687.

Gordon, B.R., Li, Y., Cote, A., Weirauch, M.T., Ding, P., Hughes, T.R., Navarre, W.W., Xia, B. and Liu, J. (2011) Structural basis for recognition of AT-rich DNA by unrelated xenogeneic silencing proteins. *Proc Natl Acad Sci U S A* **108**:10690-10695.

Grainger, D.C., Hurd, D., Goldberg, M.D. and Busby, S.J. (2006) Association of nucleoid proteins with coding and non-coding segments of the Escherichia coli genome. *Nucleic Acids Res* **34**:4642-4652.

Greene, J.M., Collins, F., Lefkowitz, E.J., Roos, D., Scheuermann, R.H., Sobral, B., Stevens, R., White, O. and Di Francesco, V. (2007) National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics. *Infect Immun* **75**:3212-3219.

Greenfeder, S.A. and Newlon, C.S. (1992) A replication map of a 61-kb circular derivative of Saccharomyces cerevisiae chromosome III. *Mol Biol Cell* **3**:999-1013.

Grigoriev, A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res* **26**:2286-2290.

Haeusser, D.P. and Levin, P.A. (2008) The great divide: coordinating cell cycle events during bacterial growth and division. *Curr Opin Microbiol* **11**:94-99.

Haines, A.S., Akhtar, P., Stephens, E.R., Jones, K., Thomas, C.M., Perkins, C.D., Williams, J.R., Day, M.J. and Fry, J.C. (2006) Plasmids from freshwater environments capable of IncQ retrotransfer are diverse and include pQKH54, a new IncP-1 subgroup archetype. *Microbiology* **152**:2689-2701.

Hanawalt, P.C. (1991) Heterogeneity of DNA repair at the gene level. *Mutat Res* **247**:203-211.

Handa, N., Bianco, P.R., Baskin, R.J. and Kowalczykowski, S.C. (2005) Direct visualization of RecBCD movement reveals cotranslocation of the RecD motor after chi recognition. *Mol Cell* **17**:745-750.

Hayes, F. and Sherratt, D.J. (1997) Recombinase binding specificity at the chromosome dimer resolution site dif of Escherichia coli. *J Mol Biol* **266**:525-537.

Heiges, M., Wang, H., Robinson, E., Aurrecoechea, C., Gao, X., Kaluskar, N., Rhodes, P., Wang, S., He, C.Z., Su, Y., Miller, J., Kraemer, E. and Kissinger, J.C. (2006) CryptoDB: a Cryptosporidium bioinformatics resource update. *Nucleic Acids Res* **34**:D419-422.

Hendrickson, H. and Lawrence, J.G. (2006) Selection for chromosome architecture in bacteria. *J Mol Evol* **62**:615-629.

Hendrickson, H. and Lawrence, J.G. (2007) Mutational bias suggests that replication termination occurs near the dif site, not at Ter sites. *Mol Microbiol* **64**:42-56.

Henrick, K., Feng, Z., Bluhm, W.F., Dimitropoulos, D., Doreleijers, J.F., Dutta, S., Flippen-Anderson, J.L., Ionides, J., Kamada, C., Krissinel, E., Lawson, C.L., Markley, J.L., Nakamura, H., Newman, R., Shimizu, Y., Swaminathan, J., Velankar, S., Ory, J., Ulrich, E.L., Vranken, W., Westbrook, J., Yamashita, R., Yang, H., Young, J., Yousufuddin, M. and Berman, H.M. (2008) Remediation of the protein data bank archive. *Nucleic Acids Res* **36**:D426-433.

Hertz-Fowler, C., Peacock, C.S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K., Parkhill, J., Ivens, A.C., Rajandream, M.A. and Barrell, B. (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res* **32**:D339-343.

Higgins, N.P. (2007) Mutational bias suggests that replication termination occurs near the dif site, not at Ter sites: what's the Dif? *Mol Microbiol* **64**:1-4.

Hill, C.W. and Gray, J.A. (1988) Effects of chromosomal inversion on cell fitness in Escherichia coli K-12. *Genetics* **119**:771-778.

Hill, T.M. (1992) Arrest of bacterial DNA replication. *Annu Rev Microbiol* **46**:603-633.

Hill, T.M., Henson, J.M. and Kuempel, P.L. (1987) The terminus region of the Escherichia coli chromosome contains two separate loci that exhibit polar inhibition of replication. *Proc Natl Acad Sci U S A* **84**:1754-1758.

Hirose, S., Hiraga, S. and Okazaki, T. (1983) Initiation site of deoxyribonucleotide polymerization at the replication origin of the Escherichia coli chromosome. *Mol Gen Genet* **189**:422-431.

Hong, E.L., Balakrishnan, R., Dong, Q., Christie, K.R., Park, J., Binkley, G., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Krieger, C.J., Livstone, M.S., Miyasato, S.R., Nash, R.S., Oughtred, R., Skrzypek, M.S., Weng, S., Wong, E.D., Zhu, K.K., Dolinski, K., Botstein, D. and Cherry, J.M. (2008) Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res* **36**:D577-581.

Horiuchi, T., Nishitani, H. and Kobayashi, T. (1995) A new type of E. coli recombinational hotspot which requires for activity both DNA replication termination events and the Chi sequence. *Adv Biophys* **31**:133-147.

Hsiao, W.W., Ung, K., Aeschliman, D., Bryan, J., Finlay, B.B. and Brinkman, F.S. (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet* **1**:e62.

Hubbard, T.J., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S. and Flicek, P. (2009) Ensembl 2009. *Nucleic Acids Res* **37**:D690-697.

Huber, K.E. and Waldor, M.K. (2002) Filamentous phage integration requires the host recombinases XerC and XerD. *Nature* **417**:656-659.

Hucka, M., Finney, A., Bornstein, B.J., Keating, S.M., Shapiro, B.E., Matthews, J., Kovitz, B.L., Schilstra, M.J., Funahashi, A., Doyle, J.C. and Kitano, H. (2004) Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project. *Syst Biol (Stevenage)* **1**:41-53.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P. and Oinn, T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* **34**:W729-732.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B.A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P.S. and Sigrist, C.J. (2008) The 20 years of PROSITE. *Nucleic Acids Res* **36**:D245-249.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H. and Yeats, C. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**:D211-215.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M. and Omura, S. (2003) Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat Biotechnol* **21**:526-531.

Itoh, M. and Watanabe, H. (2009) CGAS: comparative genomic analysis server. *Bioinformatics* **25**:958-959.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyere, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pe, M.E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.F., Weissenbach, J., Quetier, F. and Wincker, P. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**:463-467.

Jensen, R.B. (2006) Analysis of the terminus region of the Caulobacter crescentus chromosome and identification of the dif site. *J Bacteriol* **188**:6016-6019.

Johnson, L.S., Eddy, S.R. and Portugaly, E. (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**:431.

Joyce, G.F. (1994) *"Foreward" In Origins of Life: The Central Concepts.* Jones and Bartlett Publishers, Boston.

Juhas, M., van der Meer, J.R., Gaillard, M., Harding, R.M., Hood, D.W. and Crook, D.W. (2009) Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol Rev* **33**:376-393.

Junker, B.H., Klukas, C. and Schreiber, F. (2006) VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics* **7**:109.

Kahramanoglou, C., Seshasayee, A.S., Prieto, A.I., Ibberson, D., Schmidt, S., Zimmermann, J., Benes, V., Fraser, G.M. and Luscombe, N.M. (2011) Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic Acids Res* **39**:2073-2091.

Kamada, K., Horiuchi, T., Ohsumi, K., Shimamoto, N. and Morikawa, K. (1996) Structure of a replication-terminator protein complexed with DNA. *Nature* **383**:598-603.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**:D480-484.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. and Hirakawa, M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38**:D355-360.

Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N., Kunin, V. and Lopez-Bigas, N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* **33**:6083-6089.

Karplus, K., Barrett, C. and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**:846-856.

Katoh, K. and Toh, H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* **9**:286-298.

Kennedy, S.P., Chevalier, F. and Barre, F.X. (2008) Delayed activation of Xer recombination at dif by FtsK during septum assembly in Escherichia coli. *Mol Microbiol* **68**:1018-1028.

Keseler, I.M., Bonavides-Martinez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R.P., Johnson, D.A., Krummenacker, M., Nolan, L.M., Paley, S., Paulsen, I.T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A.G. and Karp, P.D. (2009) EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res* **37**:D464-470.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D.S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C.J., Kanth, S., Ahmed, M., Kashyap, M.K., Mohmood, R., Ramachandra, Y.L., Krishna, V., Rahiman, B.A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R. and Pandey, A. (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res* **37**:D767-772.

Khan, S.A. (2000) Plasmid rolling-circle replication: recent developments. *Mol Microbiol* **37**:477-484.

Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L. and Schwede, T. (2009) The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res* **37**:D387-392.

Kitano, H. (2002) Systems biology: a brief overview. *Science* **295**:1662-1664.

Kobayashi, T., Hidaka, M. and Horiuchi, T. (1989) Evidence of a ter specific binding protein essential for the termination reaction of DNA replication in Escherichia coli. *EMBO J* **8**:2435-2441.

Kogoma, T. (1997) Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. *Microbiol Mol Biol Rev* **61**:212-238.

Kono, N., Arakawa, K., Ogawa, R., Kido, N., Oshita, K., Ikegami, K., Tamaki, S. and Tomita, M. (2009) Pathway Projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. *PLoS ONE* **4**:e7710.

Kono, N., Arakawa, K. and Tomita, M. (2006) MEGU: pathway mapping web-service based on KEGG and SVG. *In Silico Biol* **6**:621-625.

Kono, N., Arakawa, K. and Tomita, M. (2011) Comprehensive prediction of chromosome dimer resolution sites in bacterial genomes. *BMC Genomics* **12**:19.

Koshland, D.E., Jr. (2002) Special essay. The seven pillars of life. *Science* **295**:2215-2216.

Kowalczykowski, S.C., Dixon, D.A., Eggleston, A.K., Lauder, S.D. and Rehrauer, W.M. (1994) Biochemistry of homologous recombination in Escherichia coli. *Microbiol Rev* **58**:401-465.

Kuempel, P.L., Duerr, S.A. and Seeley, N.R. (1977) Terminus region of the chromosome in Escherichia coli inhibits replication forks. *Proc Natl Acad Sci U S A* **74**:3927-3931.

Kuhn, R.M., Karolchik, D., Zweig, A.S., Wang, T., Smith, K.E., Rosenbloom, K.R., Rhead, B., Raney, B.J., Pohl, A., Pheasant, M., Meyer, L., Hsu, F., Hinrichs, A.S., Harte, R.A., Giardine, B., Fujita, P., Diekhans, M., Dreszer, T., Clawson, H., Barber, G.P., Haussler, D. and Kent, W.J. (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* **37**:D755-761.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.K., Codani, J.J., Connerton, I.F., Cummings, N.J., Daniel, R.A., Denizot, F., Devine, K.M., Düsterhöft, A., Ehrlich, S.D., Emmerson, P.T., Entian, K.D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.Y., Glaser, P., Goffeau, A., Golightly, E.J., Grandi, G., Guiseppi, G., Guy, B.J., Haga, K., Haiech, J., Harwood, C.R., Hénaut, A., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M.F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.M., Levine, A., Liu, H., Masuda, S., Mauël, C., Médigue, C., Medina, N., Mellado, R.P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.H., Parro, V., Pohl, T.M., Portetelle, D., Porwollik, S., Prescott, A.M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Seror, S.J., Serror, P., Shin, B.S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, T., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanaka, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.F., Zumstein, E., Yoshikawa, H. and Danchin, A. (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature* **390**:249-256.

Kuzminov, A. (1995) Collapse and repair of replication forks in Escherichia coli. *Mol Microbiol* **16**:373-384.

Labib, K. and Hodgson, B. (2007) Replication fork barriers: pausing for a break or stalling for time? *EMBO Rep* **8**:346-353.

Lang, B., Blot, N., Bouffartigues, E., Buckle, M., Geertz, M., Gualerzi, C.O., Mavathur, R., Muskhelishvili, G., Pon, C.L., Rimsky, S., Stella, S., Babu, M.M. and Travers, A. (2007) High-affinity DNA binding sites for H-NS provide a molecular basis for selective silencing within proteobacterial genomes. *Nucleic Acids Res* **35**:6330-6337.

Langille, M.G. and Brinkman, F.S. (2009) IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* **25**:664-665.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**:2947-2948.

Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**:383-397.

Le Bourgeois, P., Bugarel, M., Campo, N., Daveran-Mingot, M.L., Labonte, J., Lanfranchi, D., Lautier, T., Pages, C. and Ritzenthaler, P. (2007) The unconventional Xer recombination machinery of Streptococci/Lactococci. *PLoS Genet* **3**:e117.

Lechat, P., Hummel, L., Rousseau, S. and Moszer, I. (2008) GenoList: an integrated environment for comparative analysis of microbial genomes. *Nucleic Acids Res* **36**:D469-474.

Lefranc, M.P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G. and Duroux, P. (2009) IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res* **37**:D1006-1012.

Lesterlin, C., Barre, F.X. and Cornet, F. (2004) Genetic recombination and the cell cycle: what we have learned from chromosome dimers. *Mol Microbiol* **54**:1151-1160.

Letunic, I., Yamada, T., Kanehisa, M. and Bork, P. (2008) iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci* **33**:101-103.

Levy, O., Ptacin, J.L., Pease, P.J., Gore, J., Eisen, M.B., Bustamante, C. and Cozzarelli, N.R. (2005) Identification of oligonucleotide sequences that direct the movement of the Escherichia coli FtsK translocase. *Proc Natl Acad Sci U S A* **102**:17618-17623.

Lin, N.T., Chang, R.Y., Lee, S.J. and Tseng, Y.H. (2001) Plasmids carrying cloned fragments of RF DNA from the filamentous phage (phi)Lf can be integrated into the host chromosome via site-specific integration and homologous recombination. *Mol Genet Genomics* **266**:425-435.

Lincoln, T.A. and Joyce, G.F. (2009) Self-sustained replication of an RNA enzyme. *Science* **323**:1229-1232.

Liolios, K., Chen, I.M., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V.M. and Kyrpides, N.C. (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **38**:D346-354.

Liu, G.R., Liu, W.Q., Johnston, R.N., Sanderson, K.E., Li, S.X. and Liu, S.L. (2006) Genome plasticity and ori-ter rebalancing in Salmonella typhi. *Mol Biol Evol* **23**:365-371.

Lobry, J.R. (1995) Properties of a general model of DNA evolution under no-strand-bias conditions. *J Mol Evol* **40**:326-330.

Lobry, J.R. (1996a) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13**:660-665.

Lobry, J.R. (1996b) Origin of replication of Mycoplasma genitalium. *Science* **272**:745-746.

Lobry, J.R. and Sueoka, N. (2002) Asymmetric directional mutation pressures in bacteria. *Genome Biol* **3**:RESEARCH0058.

Louarn, J., Patte, J. and Louarn, J.M. (1977) Evidence for a fixed termination site of chromosome replication in Escherichia coli K12. *J Mol Biol* **115**:295-314.

Lucchini, S., Rowley, G., Goldberg, M.D., Hurd, D., Harrison, M. and Hinton, J.C. (2006) H-NS mediates the silencing of laterally acquired genes in bacteria. *PLoS Pathog* **2**:e81.

Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Dudkiewicz, M., Dudek, M.R. and Cebrat, S. (2003) High divergence rate of sequences located on different DNA strands in closely related bacterial genomes. *J Appl Genet* **44**:561-584.

Maisnier-Patin, S., Nordstrom, K. and Dasgupta, S. (2001) Replication arrests during a single round of replication of the Escherichia coli chromosome in the absence of DnaC activity. *Mol Microbiol* **42**:1371-1382.

Malaterre, C. (2010) On what it is to fly can tell us something about what it is to live. *Orig Life Evol Biosph* **40**:169-177.

Maltsev, N., Glass, E., Sulakhe, D., Rodriguez, A., Syed, M.H., Bompada, T., Zhang, Y. and D'Souza, M. (2006) PUMA2–grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res* **34**:D369-372.

Mangan, M.E., Williams, J.M., Lathe, S.M., Karolchik, D. and Lathe, W.C., 3rd (2008) UCSC genome browser: deep support for molecular biomedical research. *Biotechnol Annu Rev* **14**:63-108.

Marians, K.J. (1992) Prokaryotic DNA replication. *Annu Rev Biochem* **61**:673-719.

Massey, T.H., Aussel, L., Barre, F.X. and Sherratt, D.J. (2004) Asymmetric activation of Xer site-specific recombination by FtsK. *EMBO Rep* **5**:399-404.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L. and D'Eustachio, P. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* **37**:D619-622.

Maturana, H. and Varela, F. (1980) *Autopoiesis and Cognition: The Realization of the Living.* D. Reidel Publishing Company, Boston.

McDonald, A.G., Boyce, S. and Tipton, K.F. (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res* **37**:D593-597.

McGlynn, P. and Guy, C.P. (2008) Replication forks blocked by protein-DNA complexes have limited stability in vitro. *J Mol Biol* **381**:249-255.

McGlynn, P. and Lloyd, R.G. (2002) Recombinational repair and restart of damaged replication forks. *Nat Rev Mol Cell Biol* **3**:859-870.

McLean, M.J., Wolfe, K.H. and Devine, K.M. (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol* **47**:691-696.

McLeod, S.M. and Waldor, M.K. (2004) Characterization of XerC- and XerD-dependent CTX phage integration in Vibrio cholerae. *Mol Microbiol* **54**:935-947.

Megy, K., Hammond, M., Lawson, D., Bruggner, R.V., Birney, E. and Collins, F.H. (2009) Genomic resources for invertebrate vectors of human pathogens, and the role of VectorBase. *Infect Genet Evol* **9**:308-313.

Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat Rev Genet* **11**:31-46.

Mewes, H.W., Dietmann, S., Frishman, D., Gregory, R., Mannhaupt, G., Mayer, K.F., Munsterkotter, M., Ruepp, A., Spannagl, M., Stumpflen, V. and Rattei, T. (2008) MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res* **36**:D196-201.

Michel, B., Boubakri, H., Baharoglu, Z., LeMasson, M. and Lestini, R. (2007) Recombination proteins and rescue of arrested replication forks. *DNA Repair (Amst)* **6**:967-980.

Michel, B., Grompone, G., Flores, M.J. and Bidnenko, V. (2004) Multiple pathways process stalled replication forks. *Proc Natl Acad Sci U S A* **101**:12783-12788.

Mira, A., Ochman, H. and Moran, N.A. (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17**:589-596.

Mirkin, E.V. and Mirkin, S.M. (2007) Replication fork stalling at natural impediments. *Microbiol Mol Biol Rev* **71**:13-35.

Mlecnik, B., Scheideler, M., Hackl, H., Hartler, J., Sanchez-Cabo, F. and Trajanoski, Z. (2005) PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res* **33**:W633-637.

Mrazek, J. and Karlin, S. (1998) Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci U S A* **95**:3720-3725.

Mulcair, M.D., Schaeffer, P.M., Oakley, A.J., Cross, H.F., Neylon, C., Hill, T.M. and Dixon, N.E. (2006) A molecular mousetrap determines polarity of termination of DNA replication in E. coli. *Cell* **125**:1309-1319.

Nakamura, Y., Kaneko, T. and Tabata, S. (2000) CyanoBase, the genome database for Synechocystis sp. strain PCC6803: status for the year 2000. *Nucleic Acids Res* **28**:72.

Navarre, W.W., Porwollik, S., Wang, Y., McClelland, M., Rosen, H., Libby, S.J. and Fang, F.C. (2006) Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. *Science* **313**:236-238.

Neilson, L., Blakely, G. and Sherratt, D.J. (1999) Site-specific recombination at dif by Haemophilus influenzae XerC. *Mol Microbiol* **31**:915-926.

Neylon, C., Kralicek, A.V., Hill, T.M. and Dixon, N.E. (2005) Replication termination in Escherichia coli: structure and antihelicase activity of the Tus-Ter complex. *Microbiol Mol Biol Rev* **69**:501-526.

Nielsen, O. and Lobner-Olesen, A. (2008) Once in a lifetime: strategies for preventing re-replication in prokaryotic and eukaryotic cells. *EMBO Rep* **9**:151-156.

Noe, L. and Kucherov, G. (2005) YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res* **33**:W540-543.

Nolivos, S., Pages, C., Rousseau, P., Le Bourgeois, P. and Cornet, F. (2010) Are two better than one? Analysis of an FtsK/Xer recombination system that uses a single recombinase. *Nucleic Acids Res* **38**:6477-6489.

Obayashi, T., Hayashi, S., Saeki, M., Ohta, H. and Kinoshita, K. (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Res* **37**:D987-991.

Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S. and Kanehisa, M. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res* **36**:W423-426.

Oliynyk, M., Samborskyy, M., Lester, J.B., Mironenko, T., Scott, N., Dickens, S., Haydock, S.F. and Leadlay, P.F. (2007) Complete genome sequence of the erythromycin-producing bacterium Saccharopolyspora erythraea NRRL23338. *Nat Biotechnol* **25**:447-453.

Oshima, T., Aiba, H., Masuda, Y., Kanaya, S., Sugiura, M., Wanner, B.L., Mori, H. and Mizuno, T. (2002) Transcriptome analysis of all two-component regulatory system mutants of Escherichia coli K-12. *Mol Microbiol* **46**:281-291.

Oshima, T., Ishikawa, S., Kurokawa, K., Aiba, H. and Ogasawara, N. (2006) Escherichia coli histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase. *DNA Res* **13**:141-153.

Pagani, I., Liolios, K., Jansson, J., Chen, I.M., Smirnova, T., Nosrat, B., Markowitz, V.M. and Kyrpides, N.C. (2011) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40**:D571-579.

Paley, S.M. and Karp, P.D. (2006) The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Res* **34**:3771-3778.

Pályi, G., Zucchi, C. and Caglioti, L. (2002) *Fundamentals of life.* Elsevier, New York.

Papin, J.A., Price, N.D., Wiback, S.J., Fell, D.A. and Palsson, B.O. (2003) Metabolic pathways in the post-genome era. *Trends Biochem Sci* **28**:250-258.

Petersen, G., Johnson, P., Andersson, L., Klinga-Levan, K., Gomez-Fabre, P.M. and Stahl, F. (2005) RatMap–rat genome tools and data. *Nucleic Acids Res* **33**:D492-494.

Pico, A.R., Kelder, T., van Iersel, M.P., Hanspers, K., Conklin, B.R. and Evelo, C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol* **6**:e184.

Popa, R. (2004) *Between necessity and probability: searching for the definition and origin of life.* Springer, New York.

Possoz, C., Filipe, S.R., Grainge, I. and Sherratt, D.J. (2006) Tracking of controlled Escherichia coli replication fork stalling and restart at repressor-bound DNA in vivo. *EMBO J* **25**:2596-2604.

Prescott, D.M. and Kuempel, P.L. (1972) Bidirectional replication of the chromosome in Escherichia coli. *Proc Natl Acad Sci U S A* **69**:2842-2845.

Prigogine, I. (1980) *From Being to Becoming.* W H Freeman & Co, San Francisco.

Prigogine, I. and Stengers, I. (1984) *Order out of chaos: Man's new dialogue with nature.* Bantam New Age Books, London.

Reeves, G.A., Talavera, D. and Thornton, J.M. (2009) Genome and proteome annotation: organization, interpretation and integration. *J R Soc Interface* **6**:129-147.

Reyes, A., Gissi, C., Pesole, G. and Saccone, C. (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol* **15**:957-966.

Rocha, E. (2002) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol* **10**:393-395.

Rocha, E.P. (2004) The replication-related organization of bacterial genomes. *Microbiology* **150**:1609-1627.

Rocha, E.P. (2008) The organization of the bacterial genome. *Annu Rev Genet* **42**:211-233.

Rocha, E.P. and Danchin, A. (2001) Ongoing evolution of strand composition in bacterial genomes. *Mol Biol Evol* **18**:1789-1799.

Rocha, E.P., Touchon, M. and Feil, E.J. (2006) Similar compositional biases are caused by very different mutational effects. *Genome Res* **16**:1537-1547.

Rogers, A., Antoshechkin, I., Bieri, T., Blasiar, D., Bastiani, C., Canaran, P., Chan, J., Chen, W.J., Davis, P., Fernandes, J., Fiedler, T.J., Han, M., Harris, T.W., Kishore, R., Lee, R., McKay, S., Muller, H.M., Nakamura, C., Ozersky, P., Petcherski, A., Schindelman, G., Schwarz, E.M., Spooner, W., Tuli, M.A., Van Auken, K., Wang, D., Wang, X., Williams, G., Yook, K., Durbin, R., Stein, L.D., Spieth, J. and Sternberg, P.W. (2008) WormBase 2007. *Nucleic Acids Res* **36**:D612-617.

Rudd, K.E. (2000) EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic Acids Res* **28**:60-64.

Rudner, R., Karkas, J.D. and Chargaff, E. (1968) Separation of B. subtilis DNA into complementary strands, I. Biological properties. *Proc Natl Acad Sci U S A* **60**:630-635.

Rudolph, C.J., Upton, A.L., Briggs, G.S. and Lloyd, R.G. (2010) Is RecG a general guardian of the bacterial genome? *DNA Repair (Amst)* **9**:210-223.

Ruiz-Mirazo, K., Peretó, J. and Moreno, A. (2004) A universal definition of life: autonomy and open-ended evolution. *Orig Life Evol Biosph* **34**:323-346.

Sahoo, T., Mohanty, B.K., Lobert, M., Manna, A.C. and Bastia, D. (1995) The contrahelicase activities of the replication terminator proteins of Escherichia coli and Bacillus subtilis are helicase-specific and impede both helicase translocation and authentic DNA unwinding. *J Biol Chem* **270**:29138-29144.

Saleh, O.A., Perals, C., Barre, F.X. and Allemand, J.F. (2004) Fast, DNA-sequence independent translocation by FtsK in a single-molecule experiment. *EMBO J* **23**:2430-2439.

Salomonis, N., Hanspers, K., Zambon, A.C., Vranizan, K., Lawlor, S.C., Dahlquist, K.D., Doniger, S.W., Stuart, J., Conklin, B.R. and Pico, A.R. (2007) GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics* **8**:217.

Salzberg, S.L., Salzberg, A.J., Kerlavage, A.R. and Tomb, J.F. (1998) Skewed oligomers and origins of replication. *Gene* **217**:57-67.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Yaschenko, E. and Ye, J. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **37**:D5-15.

Schaeffer, P.M., Headlam, M.J. and Dixon, N.E. (2005) Protein–protein interactions in the eubacterial replisome. *IUBMB Life* **57**:5-12.

Schäffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. and Altschul, S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* **29**:2994-3005.

Schaper, S. and Messer, W. (1995) Interaction of the initiator protein DnaA of Escherichia coli with its DNA target. *J Biol Chem* **270**:17622-17626.

Scheeff, E.D. and Bourne, P.E. (2006) Application of protein structure alignments to iterated hidden Markov model protocols for structure prediction. *BMC Bioinformatics* **7**:410.

Scholefield, G., Veening, J.W. and Murray, H. (2011) DnaA and ORC: more than DNA replication initiators. *Trends Cell Biol* **21**:188-194.

Schrödinger, E. (1944) *What is life? The physical Aspect of the Living Cell.* Cambridge University Press, Cambridge.

Schultz, D.W., Swindle, J. and Smith, G.R. (1981) Clustering of mutations inactivating a Chi recombinational hotspot. *J Mol Biol* **146**:275-286.

Sciochetti, S.A., Piggot, P.J. and Blakely, G.W. (2001) Identification and characterization of the dif Site from Bacillus subtilis. *J Bacteriol* **183**:1058-1068.

Sekine, K., Hase, T. and Sato, N. (2002) Reversible DNA compaction by sulfite reductase regulates transcriptional activity of chloroplast nucleoids. *J Biol Chem* **277**:24399-24404.

Sharma, B. and Hill, T.M. (1995) Insertion of inverted Ter sites into the terminus region of the Escherichia coli chromosome delays completion of DNA replication and disrupts the cell cycle. *Mol Microbiol* **18**:45-61.

Sharp, P.M. and Li, W.H. (1987) The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**:1281-1295.

Sherman, D.J., Martin, T., Nikolski, M., Cayla, C., Souciet, J.L. and Durrens, P. (2009) Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. *Nucleic Acids Res* **37**:D550-554.

Sherratt, D.J. (2003) Bacterial chromosome dynamics. *Science* **301**:780-785.

Simmons, L.A., Breier, A.M., Cozzarelli, N.R. and Kaguni, J.M. (2004) Hyperinitiation of DNA replication in Escherichia coli leads to replication fork collapse and inviability. *Mol Microbiol* **51**:349-358.

Singleton, M.R., Dillingham, M.S., Gaudier, M., Kowalczykowski, S.C. and Wigley, D.B. (2004) Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. *Nature* **432**:187-193.

Sprague, J., Doerry, E., Douglas, S. and Westerfield, M. (2001) The Zebrafish Information Network (ZFIN): a resource for genetic, genomic and developmental research. *Nucleic Acids Res* **29**:87-90.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A. and Lewis, S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res* **12**:1599-1610.

Steiner, W., Liu, G., Donachie, W.D. and Kuempel, P. (1999) The cytoplasmic domain of FtsK protein is required for resolution of chromosome dimers. *Mol Microbiol* **31**:579-583.

Stover, N.A., Krieger, C.J., Binkley, G., Dong, Q., Fisk, D.G., Nash, R., Sethuraman, A., Weng, S. and Cherry, J.M. (2006) Tetrahymena Genome Database (TGD): a new genomic resource for Tetrahymena thermophila research. *Nucleic Acids Res* **34**:D500-503.

Sueoka, N. (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol* **40**:318-325.

Sugita, C., Ogata, K., Shikata, M., Jikuya, H., Takano, J., Furumichi, M., Kanehisa, M., Omata, T., Sugiura, M. and Sugita, M. (2007) Complete nucleotide sequence of the freshwater unicellular cyanobacterium Synechococcus elongatus PCC 6301 chromosome: gene content and organization. *Photosynth Res* **93**:55-67.

Suhre, K. and Claverie, J.M. (2004) FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res* **32**:D273-276.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P. and Huala, E. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res* **36**:D1009-1014.

Tao, Y., Liu, Y., Friedman, C. and Lussier, Y.A. (2004) Information Visualization Techniques in Bioinformatics during the Postgenomic Era. *Drug Discov Today Biosilico* **2**:237-245.

Taylor, A.F., Schultz, D.W., Ponticelli, A.S. and Smith, G.R. (1985) RecBC enzyme nicking at Chi sites during DNA unwinding: location and orientation-dependence of the cutting. *Cell* **41**:153-163.

Tendeng, C. and Bertin, P.N. (2003) H-NS in Gram-negative bacteria: a family of multifaceted proteins. *Trends Microbiol* **11**:511-518.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**:4673-4680.

Touchon, M. and Rocha, E.P. (2008) From GC skews to wavelets: a gentle guide to the analysis of compositional asymmetries in genomic data. *Biochimie* **90**:648-659.

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R. and Zhang, H. (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res* **37**:D555-559.

Twigger, S.N., Shimoyama, M., Bromberg, S., Kwitek, A.E. and Jacob, H.J. (2007) The Rat Genome Database, update 2007–easing the path from disease to data and back again. *Nucleic Acids Res* **35**:D658-662.

Uchiyama, I., Higuchi, T. and Kobayashi, I. (2006) CGAT: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes. *BMC Bioinformatics* **7**:472.

Val, M.E., Bouvier, M., Campos, J., Sherratt, D., Cornet, F., Mazel, D. and Barre, F.X. (2005) The single-stranded genome of phage CTX is the form used for integration into the genome of Vibrio cholerae. *Mol Cell* **19**:559-566.

Val, M.E., Kennedy, S.P., El Karoui, M., Bonne, L., Chevalier, F. and Barre, F.X. (2008) FtsK-dependent dimer resolution on multiple chromosomes in the pathogen Vibrio cholerae. *PLoS Genet* **4**:e1000201.

van Noort, J., Verbrugge, S., Goosen, N., Dekker, C. and Dame, R.T. (2004) Dual architectural roles of HU: formation of flexible hinges and rigid filaments. *Proc Natl Acad Sci U S A* **101**:6969-6974.

van Passel, M.W., Bart, A., Luyf, A.C., van Kampen, A.H. and van der Ende, A. (2006) Compositional discordance between prokaryotic plasmids and host chromosomes. *BMC Genomics* **7**:26.

Varshavsky, A.J., Nedospasov, S.A., Bakayev, V.V., Bakayeva, T.G. and Georgiev, G.P. (1977) Histone-like proteins in the purified Escherichia coli deoxyribonucleoprotein. *Nucleic Acids Res* **4**:2725-2745.

Vasconcelos, A.T., Ferreira, H.B., Bizarro, C.V., Bonatto, S.L., Carvalho, M.O., Pinto, P.M., Almeida, D.F., Almeida, L.G., Almeida, R., Alves-Filho, L., Assuncao, E.N., Azevedo, V.A., Bogo, M.R., Brigido, M.M., Brocchi, M., Burity, H.A., Camargo, A.A., Camargo, S.S., Carepo, M.S., Carraro, D.M., de Mattos Cascardo, J.C., Castro, L.A., Cavalcanti, G., Chemale, G., Collevatti, R.G., Cunha, C.W., Dallagiovanna, B., Dambros, B.P., Dellagostin, O.A., Falcao, C., Fantinatti-Garboggini, F., Felipe, M.S., Fiorentin, L., Franco, G.R., Freitas, N.S., Frias, D., Grangeiro, T.B., Grisard, E.C., Guimaraes, C.T., Hungria, M., Jardim, S.N., Krieger, M.A., Laurino, J.P., Lima, L.F., Lopes, M.I., Loreto, E.L., Madeira, H.M., Manfio, G.P., Maranhao, A.Q., Martinkovics, C.T., Medeiros, S.R., Moreira, M.A., Neiva, M., Ramalho-Neto, C.E., Nicolas, M.F., Oliveira, S.C., Paixao, R.F., Pedrosa, F.O., Pena, S.D., Pereira, M., Pereira-Ferrari, L., Piffer, I., Pinto, L.S., Potrich, D.P., Salim, A.C., Santos, F.R., Schmitt, R., Schneider, M.P., Schrank, A., Schrank, I.S., Schuck, A.F., Seuanez, H.N., Silva, D.W., Silva, R., Silva, S.C., Soares, C.M., Souza, K.R., Souza, R.C., Staats, C.C., Steffens, M.B., Teixeira, S.M., Urmenyi, T.P., Vainstein, M.H., Zuccherato, L.W., Simpson, A.J. and Zaha, A. (2005) Swine and poultry pathogens: the complete genome sequences of two strains of Mycoplasma hyopneumoniae and a strain of Mycoplasma synoviae. *J Bacteriol* **187**:5568-5577.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern,

D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. (2001) The sequence of the human genome. *Science* **291**:1304-1351.

Vernikos, G.S. and Parkhill, J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. *Bioinformatics* **22**:2196-2203.

Volff, J.N. and Altenbuchner, J. (2000) A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol Lett* **186**:143-150.

von Bertalanffy, L. (1968) *General System theory: Foundations, Development, Applications.* George Braziller, New York.

Wake, R. and King, G. (1997) A tale of two terminators: crystal structures sharpen the debate on DNA replication fork arrest mechanisms. *Structure* **5**:1-5.

Wake, R.G. (1997) Replication fork arrest and termination of chromosome replication in Bacillus subtilis. *FEMS Microbiol Lett* **153**:247-254.

Wang, S.C., West, L. and Shapiro, L. (2006) The bifunctional FtsK protein mediates chromosome partitioning and cell division in Caulobacter. *J Bacteriol* **188**:1497-1508.

Wang, X., Lesterlin, C., Reyes-Lamothe, R., Ball, G. and Sherratt, D.J. (2011) Replication and segregation of an Escherichia coli chromosome with two replication origins. *Proc Natl Acad Sci U S A* **108**:E243-250.

Warren, R., Hsiao, W.W., Kudo, H., Myhre, M., Dosanjh, M., Petrescu, A., Kobayashi, H., Shimizu, S., Miyauchi, K., Masai, E., Yang, G., Stott, J.M., Schein, J.E., Shin, H., Khattra, J., Smailus, D., Butterfield, Y.S., Siddiqui, A., Holt, R., Marra, M.A., Jones, S.J., Mohn, W.W., Brinkman, F.S., Fukuda, M., Davies, J. and Eltis, L.D. (2004) Functional characterization of a catabolic plasmid from polychlorinated- biphenyl-degrading Rhodococcus sp. strain RHA1. *J Bacteriol* **186**:7783-7795.

Watson, J.D. and Crick, F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**:737-738.

Werner, T. (2008) Bioinformatics applications for pathway analysis of microarray data. *Curr Opin Biotechnol* **19**:50-54.

Wieser, D., Papatheodorou, I., Ziehm, M. and Thornton, J.M. (2011) Computational biology for ageing. *Philos Trans R Soc Lond B Biol Sci* **366**:51-63.

Wiggins, P.A., Cheveralls, K.C., Martin, J.S., Lintner, R. and Kondev, J. (2010) Strong intranucleoid interactions organize the Escherichia coli chromosome into a nucleoid filament. *Proc Natl Acad Sci U S A* **107**:4991-4995.

Wilce, J.A., Vivian, J.P., Hastings, A.F., Otting, G., Folmer, R.H., Duggin, I.G., Wake, R.G. and Wilce, M.C. (2001) Structure of the RTP-DNA complex and the mechanism of polar replication fork arrest. *Nat Struct Biol* **8**:206-210.

Winsor, G.L., Van Rossum, T., Lo, R., Khaira, B., Whiteside, M.D., Hancock, R.E. and Brinkman, F.S. (2009) Pseudomonas Genome Database: facilitating user-friendly, comprehensive comparisons of microbial genomes. *Nucleic Acids Res* **37**:D483-488.

Winterling, K.W., Levine, A.S., Yasbin, R.E. and Woodgate, R. (1997) Characterization of DinR, the Bacillus subtilis SOS repressor. *J Bacteriol* **179**:1698-1703.

Worning, P., Jensen, L.J., Hallin, P.F., Staerfeldt, H.H. and Ussery, D.W. (2006) Origin of replication in circular prokaryotic chromosomes. *Environ Microbiol* **8**:353-361.

Xu, L. and Marians, K.J. (2003) PriA mediates DNA replication pathway choice at recombination intermediates. *Mol Cell* **11**:817-826.

Xu, P., Widmer, G., Wang, Y., Ozaki, L.S., Alves, J.M., Serrano, M.G., Puiu, D., Manque, P., Akiyoshi, D., Mackey, A.J., Pearson, W.R., Dear, P.H., Bankier, A.T., Peterson, D.L., Abrahamsen, M.S., Kapur, V., Tzipori, S. and Buck, G.A. (2004) The genome of Cryptosporidium hominis. *Nature* **431**:1107-1112.

Yang, J., Chen, L., Yu, J., Sun, L. and Jin, Q. (2006) ShiBASE: an integrated database for comparative genomics of Shigella. *Nucleic Acids Res* **34**:D398-401.

Yates, J., Aroyo, M., Sherratt, D.J. and Barre, F.X. (2003) Species specificity in the activation of Xer recombination at dif by FtsK. *Mol Microbiol* **49**:241-249.

Yates, J., Zhekov, I., Baker, R., Eklund, B., Sherratt, D.J. and Arciszewska, L.K. (2006) Dissection of a functional interaction between the DNA translocase, FtsK, and the XerD recombinase. *Mol Microbiol* **59**:1754-1766.

Yen, M.R., Lin, N.T., Hung, C.H., Choy, K.T., Weng, S.F. and Tseng, Y.H. (2002) oriC region and replication termination site, dif, of the Xanthomonas campestris pv. campestris 17 chromosome. *Appl Environ Microbiol* **68**:2924-2933.

Zhang, C.T., Zhang, R. and Ou, H.Y. (2003) The Z curve database: a graphic representation of genome sequences. *Bioinformatics* **19**:593-599.

Zhang, R. and Lin, Y. (2009) DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res* **37**:D455-458.

Zhang, Z., Cheung, K.H. and Townsend, J.P. (2009) Bringing Web 2.0 to bioinformatics. *Brief Bioinform* **10**:1-10.

Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**:203-214.

Zhu, J., Newlon, C.S. and Huberman, J.A. (1992) Localization of a DNA replication origin and termination zone on chromosome III of Saccharomyces cerevisiae. *Mol Cell Biol* **12**:4733-4741.

Zhu, M., Yu, M. and Zhao, S. (2009) Understanding quantitative genetics in the systems biology era. *Int J Biol Sci* **5**:161-170.

Une fondation réflexive du solidarisme
- Étude philosophique sur la pensée de Léon Bourgeois -

発　行　日　2012年9月5日
著　　者　宮代康丈
発　行　所　慶應義塾大学 湘南藤沢学会
印　刷　所　株式会社 ワキプリントピア

ISBN 978-4-87762-260-2
SFC-RMN 2012-001

Comparative Genomics of Bacterial Sequence Elements
Associated with Chromosome Structure

2012年4月10日　　初版発行

著者　河野暢明

監修　冨田勝　荒川和晴

発行　慶應義塾大学 湘南藤沢学会
〒252-0816　神奈川県藤沢市遠藤5322
TEL:0466-49-3437

Printed in Japan　　印刷・製本　　ワキプリントピア

ISBN 978－4－87762－257－2
SFC-DT 2012-001

■本論文は博士論文において優秀と認められ、出版されたものです。