國立交通大學

生物科技系所

博士論文

細菌致病因子的基因體學分析

A Genome-Wide Study on the Virulence Determinants in Bacteria

研究生：陳盈璁 (8928802)
Student: Ying-Tsong Chen
指導教授：彭慧玲 博士
Advisor: Hwei-Ling Peng Ph.D.

中華民國九十三年九月
September, 2004

摘要

這個論文由三個針對細菌致病因子的研究構成。首先在第一個部份，是克雷白氏菌(*Klebsiella pneumoniae*)裡一個重要的致病性巨型質體(plasmid) pLVPK的核酸定序以及基因註解(annotation)工作。我們定序了這個全長21.9萬鹼基配對(base pair)的巨型質體並且在其中註解出251個開讀框架(ORF; open reading frame)。在這些可能的基因之中，我們發現了許多明顯可能與細菌致病性有關的毒性基因(virulence gene)，其中包括了控制莢膜多醣體(CPS)合成的*rmpA2*以及它的同源基因*rmpA*，各種的抓鐵系統基因如*iucABCD-iutA*, *iroBCDN*, *fepBC*, *fecIRA*等。此外，我們也在這個質體的核酸序列中發現了與其他細菌中負責產生對於銅、銀、鉛、碲等的抗性基因組(gene-cluster)相似的數個基因組。在質體上主要的基因組之間，我們總共發現了十三個由插入序列(insertion sequence, *IS*)組成的區域，這些插入序列使我們相信這個大質體極有可能是經由一系列水平轉移(horizontal-transferred)基因組的重組而形成。

在論文的第二部份，我們在綠膿桿菌(*Pseudomonas aeruginosa*)的全基因體序列中針對一類重要的毒性因子—雙分子調控系統基因(2CS; two-component system)—的演化進行分析。我們分析了這些基因的排列方式以及它們所編碼(encode)的蛋白質序列的功能區塊(domain)組成。再經由比對感受子(sensor)與反應子(regulator)這兩個雙分子調控系統的組成單元的演化樹，我們可以推論至少

有一半以上的雙分子調控系統它所包含的兩個單元之間有很明顯的共演化(co-evolution)現象。我們也發現，以上的共演化特徵，在一群帶有與OmpR相似反應子的雙分子調控系統中顯得特別的明顯。相反的在一些其它的雙分子調控系統中，特別是那些帶有與NarL相似反應子的，這種特徵就比較不明顯。在分別針對感受子與反應子所進行的分組之間找到的關聯性，也支持了以上的結果。此外，證據也顯示六群雙分子調控系統極有可能分別演化來自六個共同的來源。從鄰近基因的功能看來，這些雙分子調控系統基因非但是經由基因組整個的重製(duplication)而產生，它們還很可能在重製之後仍然保有相同的功能。我們更進一步分析並比較它們開讀框架前端的未轉錄序列(untranslated sequence)時，發現細菌可能對這些基因組採取不同的基因轉錄調控來避免功能的重疊。

最後一個章節中，我們則利用HMMER從38個已經完成基因體序列定序的微生物中找出並且分析其中OmpR家族雙分子調控系統的演化。OmpR家族雙分子調控系統在這些物種之中的分布也支持了雙分子調控系統源自細菌而後傳播至其他物種的理論。一般而言，OmpR家族的雙分子調控系統，它們的基因排列都是呈現一個從反應子到感受子(RS)的排列方式。我們分別分析了屬於RS與SR(與RS恰好相反的基因排列順序)的兩群雙分子調控系統，發現它們感受子上，接受磷酸基修飾(phosphorylated)的組胺酸殘基(histidine residue)附近的序列都非常的相似。這意味著雙分子調控系統的兩個組成單元之間的蛋白質－蛋白質作用很可能約束(constrain)了蛋白質上特定功能區域的演化。這個現象也同時說明了感受

子與反應子之間的作用應該發生於兩個單元在演化歷史中組合成為成RS或SR基因組之前。在這些基因組之間，還保留著相當相似的基因排列順序及蛋白質功能區域，這成為雙分子調控系統共演化的強力佐證。

Abstract

The thesis covered three major approaches aimed to identify the property of bacterial virulence determinants. In the first part, we determined the entire DNA sequence of pLVPK, a 219-kb virulence plasmid harbored in *Klebsiella pneumoniae*. A total of 251 open reading frames were annotated. The obvious virulence-associated genes carried by the plasmid are the capsular polysaccharide synthesis regulator *rmpA* and its homolog *rmpA2*, and multiple iron-acquisition systems, including *iucABCD-iutA* and *iroBCDN* siderophore gene clusters, *fepBC* ABC-type transporter, and *fecIRA* that encodes iron uptake regulatory system. In addition, several gene clusters homologous with copper, silver, lead, and tellurite resistance genes of other bacteria were also identified. The presence of thirteen insertion sequences located mostly at the boundaries of the aforementioned gene clusters suggests that pLVPK was derived from a sequential assembly of various horizontally-acquired DNA fragments. In the second part, we analyzed the complete genome sequence of *P. aeruginosa* PAO1 to unravel the evolution of a group of important virulence factors, the two component systems. Gene organization and functional motif analyses of the 123 two component system (2CS) genes in *Pseudomonas aeruginosa* PAO1 were carried out. By comparing the phylogenetic trees built respectively for the two

components, we showed that more than half of the sensor-regulator gene pairs, especially the 2CSs with OmpR-like regulators, are derivatives of a common ancestor, and have most likely co-evolved through gene pair duplication, while several of the 2CS pairs, especially those with NarL-like regulators, appeared to be relatively divergent. Correlation of the classification of sensor kinases and response regulators further provides support for these models. We have identified six congruent clades, which represent the group of the most recently duplicated 2CS gene pairs. Sequence comparison showed that certain paralogous 2CS pairs may carry a redundant function even after a gene duplication event. However, comparative analysis of the putative promoter regions of the paralogs suggested that functional redundancy could be prevented by a differential control. Finally in the third part, we analyzed 38 completed genomes using HMMER to identify the putative 2CS components and investigate the evolution of the 2CSs of OmpR-family. The distribution of OmpR-like response regulators among different genomes of different taxonomy groups also supported the hypothesis that 2CSs are originated in the last common ancestor of bacteria and subsequently passed to the other species. Mostly, the 2CS genes containing an OmpR-like regulator-encoding gene were found in the order of regulator-to-sensor (RS). The amino acid sequences around the phosphorylated histidine residue of the sensor kinases from either RS or SR (sensor-to-regulator) 2CSs were nearly identical.

This suggested that the interaction of 2CS component may have constrained the sequences of the interacting domain between sensor kinase and the cognate response regulator while the ancestral components were brought together during evolution into a RS or SR gene cluster, where coordinated transcriptional control may be economically favored. The nearly invariant gene order and the conservation of catalytic domains of these 2CSs provide strong evidence for the co-evolution of 2CSs.

Acknowledgments

若沒有遇到像彭慧玲博士這樣的導師，我不會有機會在這裡大模大樣的寫這個致謝辭。能圓滿走完這個求學生涯的重要階段，感謝的人很多，然而一切都沒有如老師的無私付出與苦心教誨一般，能在這四年的時光裡深深的影響了我。我由衷感謝我的指導教授彭慧玲博士，在這段期間以無比的耐心與愛包容我的青澀與固執，許多次帶領我、驅策我走出陰影。在這裡研究與學習，克竟難關完成學業，實在是我的福氣與榮幸。我要謝謝曾經一起在交大打拼的夥伴們：靖婷、盈蓉、榕華、怡欣、騰逸、巧韻、致翔、定宇、美甄、婉君、珮瑄、平輝、新耀、祐俊、健誠、育盛、欣穎、志凱、心瑋、佳融、睿瑜、昀錚、頌瑾… 謝謝清大無比認真的張晃猷教授、賴怡琪博士，Nadini、韶智、Jaya、文玲、志宇、貞儀、莉芳…謝謝大家給我許多的協助以及美好的回憶。謝謝交大生科所有的老師們，特別是盧錦隆老師、邱顯泰老師和楊昀良老師，先後不辭辛苦的擔任我的口試委員。還要感謝毛仁淡院長，在我最低潮的時候給我下了很多猛藥。謝謝許芳榮老師，林耀鈐老師和蔡英德老師在生物資訊上的寶貴指點。還要謝謝陽明大學的蔡世峰教授，吳克銘，在定序與基因體研究上給了我許多的指導與協助。也要感謝我的爸爸媽媽姊姊一直支持我放心的投入我的學業。最後，感謝我摯愛的純沂與樂融，感謝愛妻的無悔相許跟扶持，願以我小小的成就與快樂，與你們共榮。

陳盈璁 9/9/2004

Contents

List of Tables

List of Figures

Abbreviations

2CS	two-component system
Ap	Ampicillin
bp	base pair(s)
BLAST	basic local alignment search tool
CPS	capsular polysaccharide
COG	cluster of orthologous groups
DNA	Deoxyribonucleic acid
EDTA	Ethylenediamine-tetraacetic acid
G+C	GC content of gene
GC3	GC of silent 3rd codon position
Hpt	histidine-containing phosphotransfer
HMM	hidden Markov model
IS	insertion sequence
I; T; R; O	input-, transmitter-, receiver, output-domains
kb	kilobase(s)
mb	millionbase(s)
kD	kiloDalton(s)
LB	Luria Bertani
Nc	number of codons used
mM	millimolar
ng	nanogram
ORF	open reading frame
ori	origin of replication
OTU	operational taxonomic units
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PAGE	polyacrylamide gel electrophoresis
RNA	ribonucleic acid
rpm	revolutions per minute
SDS	sodium dodecyl sulfate
SR; RS	sensor-histidine kinase and response regulator gene pairs
Tris	Tris(hydroxymethyl)-aminomethane
tRNA	transfer RNA

Introduction

Microbial infections are one of the major causes of fatality worldwide. In 1998, the most common cause of death among children, defined by the World Health Organization as aged between 0 and 4 years, was infectious diseases, which accounted for 63% of all fatalities (http://www.who.org). According to the report of CDC (Centers for Disease Control and Prevention in the United States) report, microbial agents are the 4th actual cause of the deaths in the United States, 2000 (http://www.cdc.gov). Finlay and Falkow had discussed the definitions of microbial pathogenicity and the idea that pathogens can be distinguished from their non-virulent counterparts by the presence of virulence genes (Finlay and Falkow, 1997). In general, bacterial virulence factors can be divided into several groups. These include the adherence and colonization factors, invasins, capsules and surface components, endotoxins, exotoxins, siderophores, the secretion systems for toxin transport, and the two component systems, by which the expression of many virulence genes are controlled (Krogfelt, 1991; Merritt and Hol, 1995; Payne, 1993; Miller et al., 1989; Finlay and Falkow, 1997).

Beginning from the 1970s, the development of molecular genetics and recombinant DNA technology facilitated the infectious disease research. With the advent of whole-genome sequencing, a revolution in infectious disease research has begun and entered its large-scale production. Genomics, taking advantage of the completed genome sequences, is a top-down approach to study of the genes and their functions in a genome. The completed genome sequences allowed us to decipher the entire microbial physiology and the underlying evolutionary process based on the information encoded in the DNA (Vázquez-Boland et al., 1999). Searches for virulence genes can be achieved on a genome-wide scale by a variety of bioinformatic and genetic techniques. The accumulation of the genome sequences also empowered a new scientific discipline, molecular evolution, in which the evolutionary history of genes and organisms can be reconstructed.

The first genome, the genome of RNA bacteriophage MS2, was sequenced in 1976 (Fiers et al., 1976). This was followed by the genome of bacteriophage ϕX174, with the aid of the rapid sequencing method developed by Walter Gilbert and Fred Sanger (Maxam and Gilbert, 1977; Sanger et al., 1977). These are some of the known smallest genomes with only four and ten genes, respectively. Subsequently, in 1982, Sanger announced the sequence of a relatively large genome, the genome of bacteriophage λ, which has 48,502 bases of genomic DNA and ~70 known and predicted protein- coding genes (Sanger et al., 1982). Although the authors have done meticulous computational analyses of open reading frames, particularly the prediction of translation starts and codon usage, the word "homolog" was not used. The first protein sequence database, the Protein Identification Resource, was launched by Margaret Dayhoff in 1965, long before the genomics has even become conceivable (Dayhoff et al., 1965). However, It is not until the beginning of the sequencing era, a considerable number of completed genomes have been amassed, the time was ripe for the birth of comparative genomics (Koonin and Galperin., 2002).

The announcement of complete genome sequence of the parasitic bacterium *Haemophilus influenzae* (Fleischmann et al., 1995) was greatly facilitated by the whole-genome shotgun approach pioneered by Craig Venter, Hamilton Smith, and Leroy Hood (Venter et al., 1996). Since then, completed genome sequences of bacteria and archaea have been accumulating steadily. The second genome sequencing paper on the *Mycoplasma genitalium* (Fraser et al., 1995) genome inevitably became a comparative-genomics study. Comparison of this genome to that of *Haemophilus influenzae* was carried out and the profound differences in physiology and metabolic capacity between these two bacteria were correlated with the differences in genome content (Fraser et al., 1995). The phenomena of lineage-specific gene loss, a common type of event during genome evolution, were identified also by genome comparison. For example, the genome of *Mycoplasma pneumoniae*, contains all the 480 genes of *M. genitalium*, and 197 additional genes.

Mycobacterium leprae, a closely related species to *M. tuberculosis*, however, has at least 1,200 fewer genes (Cole et al., 2001). As the list of completed genomes rapidly becomes outdated, periodically updated listings of both finished and unfinished genome sequencing projects are available at the web sites Genomes On Line Database (GOLD, http://www.genomesonline.org/) (Kyrpides, 1999). By August 2004, more than a thousand of genome projects were executed and hundreds of completely sequenced genomes are available in public databases. The largest prokaryotic genomes (*Streptomyces avermitilis* among the eubacteria, *Methanosarcina acetivorans* among the archaea) sequenced only recently promise many interesting discoveries yet to come (Ikeda et al., 2003; Galagan et al., 2002).

The massive influx of information from the genome sequencing projects is revolutionizing the science of bacterial pathogenesis, ranging from understanding the most basic aspects of gene content and pathogen genome organization, to elucidating the regulatory networks of virulence gene expression, and to investigating the global patterns of host response to infection. The methods of comparative genomics have made headway in addressing these issues for specific bacterial pathogens. *Helicobacter pylori,* for example, colonizes the human stomach where it can cause a wide spectrum of diseases ranging from asymptomatic gastritis to ulcers to gastric cancer (Censini et al., 1996). The most severe disease is associated with the strains harboring a specific DNA segment, called a pathogenicity island (PAI), which includes a cytotoxin together with a bacterial type IV secretion system that delivers the toxin into host cells (Censini et al., 1996). Genome-sequence comparison was applied to the search of strain-specific genes for the first time when genome sequences of two unrelated *H. pylori* clinical isolates were compared (Alm et al., 1999). The results showed that 6%~7% of the genes appeared to be strain-specific genes encoding the cell-surface proteins that are most likely contributing to the persistence of bacteria during long-term infections (Alm et al., 1999; Salama et al., 2000).

Escherichia coli O157:H7 is a cause of food- and water- borne illness that is now a public health problem worldwide. The first genome sequence of the pathogenetic *E. coli* O157:H7, which was isolated from the major outbreak in Sakai, Japan in 1996, was completed by Hayashi et al. (Hayashi et al., 2001). At about the same time, a second, near-complete sequence of *E. coli* O157:H7 isolated from the hamburger meat was reported (Perna et al., 2001), which has been implicated as the culprit for the first outbreak in North America (Riley et al., 1983). Sequences of the two O157:H7 genomes appeared to be very similar, however, dramatically different from that of the laboratory strain *E. coli* K-12 while the sequences compared using MEM, a system for rapidly aligning entire genomes (Perna et al., 2001; Blattner et al., 1997). Over 1.4

Mb larger than the K-12 genome, the O157:H7 carried strain-specific DNA in which ~10% were assumed to have virulence-related functions (Hayashi et al., 2001). The O157:H7 strain-specific DNA are organized into ~180 separate regions in the genome and were referred to as O-islands (Perna et al., 2001). Several of these O-islands include virulence determinants such as the bacteriophage-associated Shiga toxin (*stx*) genes (O'Brien et al., 1992), fimbrial biosynthesis systems, iron uptake and utilization clusters, and putative non-fimbrial adhesins (Perna et al., 2001). The LEE (locus of enterocyte effacement) pathogenicity island, is one of the O-islands which contains ~40 genes encoding the proteins required for the close attachment of bacterial cells to the intestinal epithelium. The acquisition of the LEE island and the *stx* genes were recognized as two of the critical steps in the evolution of *E. coli* O157:H7 (Reid et al., 2000).

Bacterial plasmids, bacteriophages, and other mobile genetic elements play a key role in a haploid world as the seminal effectors of metabolic diversity and specialization. These mobile elements are often essential components of the bacterial pathogenicity (Finlay and Falkow, 1997). It is well established that many pathogenicity factors (and antibiotic-resistance genes) engoding genes on the plasmid are often clustered and flanked by a number of repetitive sequences or transposable elements, such as the ~70 kb virulence plasmid in *Yersinia pestis, Yersinia* *pseudotuberculosis,* and all pathogenic *Yersinia enterocolitica.* The plasmid contains the Yop virulon, which produces several secreted proteins, Yops, to cause damage of the host cells and paralyzing of the phagocytic cells, and a type III secretion system to mediate the translocation of the Yops (Rosqvist et al., 1995).

Salmonella is the causative agent of food-borne gastroenteritis and typhoid fever. The presence of numerous pathogenicity islands has conferred their ability to cause disease. The *Salmonella* genus was divided into two species, *S. bongori* and *S. enterica.* The latter contained the serovars Typhi and Typhimurium, which is responsible for 99% of human infections (Whittam and Bumbaugh, 2002). A comparative genome analysis of the pathogenicity islands (PAI) of *Salmonella enterica* serovars Typhi and Typhimurium, and *E. coli* revealed that the PAI-associated tRNA loci appeared to be species-specific and were horizontally acquired (Hansen-Wester and Hensel, 2002). The differences in the distribution of these tRNA-associated elements likely conferred that bacteria unique pathogenic potentials, such as the restriction of host range or the type of disease (Hansen-Wester and Hensel, 2002).

It has also been proposed that certain pathogenic bacteria were evolved from related nonpathogenic organisms by genetically acquiring relatively large blocks of virulence factors encoding genetic material (Finlay and Falkow, 1997). Many of the virulence determinants, particularly toxins and adherence factors, were found on mobile genetic elements, which can be distributed to other bacteria by transformation, conjugation, and transduction. This kind of genetic spread is readily seen in the spreading of R-plasmids and the transposition of antibiotic resistance genes. The studies on plasmids are undeniably one of the central issues in the investigation of bacterial pathogenesis.

The geneticist Theodosius Dobzhansky believed “Nothing in biology makes sense except in the light of evolution” (Lewontin et al., 2003). Comparative genomics allowed explanation of the most common and important types of events that occurred during genome evolution, including genome rearrangement and gene duplication. The major impact of comparative genomics in genome evolution has shown "genomes in flux", which changed the classic concept that genomes are relatively stable and evolve through gradual changes and spread through vertical inheritance (Snel et al., 2002). The comparative study of the proteins of close-related genomes also improved our understanding of the evolutionary pressure on the molecules involved in the emergence of certain infectious diseases (Whittam and Bumbaugh, 2002).

The upcoming chapters include the study on the virulence determinants in bacteria. The first chapter illustrated the sequencing, annotation and partial characterization of a large plasmid, pLVPK, that contribute significantly to the virulence of *Klebsiella pneumoniae*, an important opportunistic pathogen of human. The second chapter focused on the molecular evolution of a group of virulence factors, the two component systems (2CS), in another important human pathogen, *Pseudomonas aeruginosa* PAO1. 2CSs are the means that bacteria sense the environmental stimuli and make physiological responses correspondingly. By performing the analysis of gene organization and functional motif analysis, together with the comparative phylogenetic, hypotheses on the evolution of the 123 2CS genes were proposed. The third chapter, the evolutionary constraint in the protein interacting domains and the conservation of gene organization of the 2CSs were identified through analysis of the genomes from 38 organisms. The era of complete genomes holds promise to the study of bacterial pathogenesis and also fresh prospective on the molecular evolution of virulence factors. The comparative genomics and evolutionary biology will surely lead to more interesting discoveries yet to come concerning the strategy for a bacterium to become a pathogen.

Chapter 1

Sequencing and analysis of the large virulence plasmid pLVPK of *Klebsiella pneumoniae* CG43

1.1 Introduction

Klebsiella pneumoniae is an important cause of community-acquired bacterial pneumonia, occurring particularly in chronic alcoholics and commonly results in a high fatality rate if untreated. Nevertheless, the vast majority of *K. pneumoniae* infections are associated with hospitalization. It has been estimated that *K. pneumoniae* causes up to 8% of all nosocomial bacterial infections in developed countries, and its colonization in hospitalized patients appears to be associated with the use of antibiotics (Schaberg et al., 1991). Recently, the prevalence of multiple-drug resistant *K. pneumoniae* strains has significantly restricted the availability of antibiotics for effective treatment of the bacterial infections.

Despite its significance, our knowledge of the pathogenicity of the bacterium is rather limited. Clinically isolated *K. pneumoniae* usually produces large amounts of capsular polysaccharides (CPS) as reflected by the formation of glistening mucoid colonies. The CPS provides the bacterium an anti-phagocytic ability and prevents the bacteria from being killed by serum bactericidal factors (Simmons-Smit et al., 1986). Additional virulence associated factors identified so far in *K. pneumoniae* include lipopolysaccharides, several adhesins, and iron-acquisition systems (Simmons-Smit et al., 1986; Nassif and Sansonetti, 1986). The small numbers of known virulence-associated factors rather limit the possible targets for drug development, thus making the intervention of bacterial infection rather difficult.

Several strategies including in vivo expression technology, subtractive DNA hybridization, and signature-tagged mutagenesis have been adopted to identify virulence-associated genes in *K. pneumoniae*. These efforts have allowed the identification of many novel genes that might be important for the bacterium to infect humans. For instance, by using the in vivo expression technology, we have identified the presence of a plasmid-borne iron acquisition gene cluster in *K. pneumoniae* that is primarily expressed in the hosts (Lai et al., 2001). Nevertheless, further investigation of the functional roles of these novel sequences has been significantly hampered by the lack of the complete genome sequence of *K. pneumoniae.*

Most of the blood isolates of *K. pneumoniae* harbor a large plasmid of 200 kb in size (Peng et al., 1991). The plasmid has been demonstrated to contain the aerobactin siderophore biosynthesis genes and curing of the plasmid would result in an avirulent phenotype (Nassif and Sansonetti, 1989). In our laboratory, we also found that the loss of pLVPK, a plasmid of the similar size harbored in *K. pneumoniae* CG43, a highly virulent clinical isolate of K2 serotype (Lai et al., 2003) resulted in a loss of colony mucoidy, the ability to synthesize aerobactin, and a 1000-fold decrease of virulence. It is conceivable that the plasmid is likely to carry many additional virulence-associated genes and complete sequencing of the plasmid would hence be the most straightforward way for their identification. We herein report the 219-kb sequence and annotation of this large virulence plasmid from *K. pneumoniae* CG43.

1.2 Materials and Methods

Sequencing of pLVPK

The DNA of pLVPK was isolated from *K. pneumoniae* CG43 by using a Qiagen Plasmid Purification kit and fragmented by sonication. The DNA fragments were then resolved on a 0.7% low melting point agarose gel and DNA of size ranging from 2.0 to 3.0 kb were recovered, blunt repaired by *Bal*31 nuclease, and subsequently cloned into the pUC18 vector. A total of 2,304 clones were sequenced from both ends to achieve approximately 11-fold coverage of the plasmid. Sequences were assembled initially using the Phred/Phrap program (Ewing et al., 1998) with optimized parameters and the quality score was set to >20. When all the sequences assembled into 11 major contigs (>20 reads; >2 kb), the Consed program (Gordon et al., 1998) was then used for the final sequence closure (auto-finishing). Finally, several gaps among contigs were closed either by primer walking on selected clones, which were identified by analysis on the forward and the reverse links of each of the contigs, or by sequencing the DNA amplicons generated by PCR.

Gene prediction and Annotation

GLIMMER 2.02 (Delcher et al., 1999), a program that searches for protein coding regions, was used to identify those open reading frames (ORFs) possessing more than 30 codons. Overlapping and closely clustered ORFs were manually inspected. The predicted polypeptide sequences were used to search the protein database with the BLAST (NCBI database), and the clusters of orthologous groups of proteins (COGs) database was used to identify families to which the predicted proteins were related. Mobile elements and repetitive sequences were identified using pair-wise comparison with the known insertion sequences. The presence of tRNA sequences was identified by the program tRNAscan-SE (Lowe and Todd, 1997). The G+C nucleotide composition analysis was made by GCWin of the G-Language package (Arakawa et al, 2003).

Drug susceptibility assay

Tellurite, copper, silver and lead susceptibility for the strains were determined essentially as described (Menoharan et al., 2003). *E. coli*, *K. pneumoniae* CG43, and its derivatives were propagated at 37 ^{o}C in Luria-Bertani (LB) broth. The overnight-grown cells were spread onto LB plates and the 3MM paper discs (5 mm diameter) impregnated with aliquots of a serial dilution of K_2TeO_3, $CuSO_4$, $AgNO_3$, and $Pb(NO_3)_2$ solutions were placed on top of each of the plates. The plates were then incubated at 37 ^{o}C for another 12 h and the inhibition zone measured. Iron acquisition activity was assayed using iron-deprived M9 plates (with 200 μM 2,2’-dipyridyl) and the paper discs impregnated with a serial dilution of $FeCl_3$ solution. After spreading the overnight-grown bacteria onto the plates, the iron-loaded discs were then placed on top of each of the plates. The plates were incubated at 37^{o}C for 12 h and the growth zones around the paper discs were measured.

Nucleotide sequence accession number

The nucleotide sequences reported in this paper have been submitted to GenBank under the accession no. **AY378100**.

1.3 Results and Discussions

General overview

The entire DNA sequence consists of 219,385 bp forming a circular plasmid (Figure 1). The size and the predicted restriction enzyme cutting sites are consistent with the experimental findings using pulse-field gel electrophoresis. The plasmid contains 251 ORFs, as determined by the Glimmer program (Appendix I). The possible functions of these ORFs were subsequently analyzed by comparing the sequence to the current non-redundant protein database of the National Center for Biotechnology Information using BLAST software through the Internet. Approximately 37% of the 251 ORFs have significant amino acid sequence similarity (>60%) with the genes of known function in GenBank or with protein domains or motifs in protein databases. Despite their lack of homology to the known genes, the deduced amino acid sequences of 31% of the ORFs matched the hypothetical genes in the database. The remaining 32% had lower or no significant sequence similarities (<20%) with those in the database and their functions could not be assigned.

The average G+C content of the plasmid is 50.35%, which is somewhat lower than that of the *K. pneumoniae* MGH78578 genome (G+C = ~55%). The G+C content plotted along the pLVPK sequence with a window size of 1000 bp is shown in Figure 2. Four regions (Box 1~4) with a significant high G+C content in comparison with the average of the whole plasmid sequence were identified. The Box 1 consists of 9 ORFs showing 56~90% sequence similarity to an unknown gene cluster in *Burkholderia fungorum* genome. The second and third high G+C regions contain two iron acquisition systems: *iut* and *iro* genes, respectively. The fourth box covered the lead-resistant *pbr* gene cluster and its nearby transposase gene. Two low G+C content regions are also marked in Figure 2, which include the two mucoidy regulator encoding genes, *rmpA* (34.6%) and *rmpA2* (31.9%). The values of G+C at the third codon are even lower with 29.2% for *rmpA* and 28% for *rmpA2*.

Virulence-associated genes

The BLAST search revealed an 18-kb region, which is highly similar to the SHI-2 pathogenicity island (PAI) of *Shigella flexneri* (Moss et al., 1999). The SHI-2 like region includes the iron acquisition genes *iucABCDiutA*, *vagCD*, the unknown function ORF *shiF*, and *rmpA2*, a known virulence-associated gene in *K. pneumoniae* (Lai et al., 2003). Elsewhere the PAI-like region, a *rmpA2* homolog, *rmpA,* and two additional gene clusters associated with iron metabolism were also found.

One interesting finding in pLVPK is the presence of *rmpA* and *rmpA2,* two genes encoding regulatory proteins for CPS synthesis in *K. pneumoniae*. CPS has been known to be a major virulence factor in *K pneumoniae* that protects the bacterium from the bactericidal activity of serum complements and macrophages (Simmons-Smit et al., 1986). The gene *rmpA* was first identified in *K. pneumoniae* as a determinant controlling the CPS biosynthesis (Nassif et al., 1989). The gene *rmpA2*, which was named because of its high similarity with *rmpA*, was identified later (Wacharotayankun et al., 1996). Since the major difference between these two gene products is that the RmpA2 has an extended N-terminal region, it has been generally thought that *rmpA* and *rmpA2* are the same gene, and the *rmpA* reported earlier by Nassif and colleagues was a truncated form of *rmpA2*. Our sequencing result shows that *rmpA* and *rmpA2*, which share 81% nucleotide sequence homology (78% identity in the 194 comparable amino acids), are actually two independent loci 29 kb apart (Figure 3a). Southern hybridization analysis of the plasmid using an *rmpA2* probe also confirmed the presence of two copies of the gene (Figure 3b). The finding not only clarified that *rmpA* is not a part of *rmpA2,* but also demonstrated that both the genes are plasmid-borne. Our laboratory has recently found that RmpA2 protein directly interacts with the promoters of the K2 CPS biosynthesis genes through its carboxyl terminal helix-turn-helix motif-containing portion (Lai et al., 2003). Thus, we believe that RmpA could also interact with the *cps* gene promoter, although how it activates the *cps* gene expression and the interplay between these two Rmp proteins remain to be investigated.

The *K. pneumoniae vagCD* products exhibit 94% and 84% amino acid sequence identities with that of the VagC and VagD on pR64 of *Salmonella enterica* serovar Dublin. Like the *vagCD* of pR64, the two genes are also overlapped by one nucleotide. It has been proposed that VagC and VagD might be involved in the coordination of plasmid replication and cell division and disruption of the *vagC* locus would reduce the bacterial virulence (Pullinger and Lax, 1992). The high sequence similarity suggests that *vagCD* genes on the pLVPK also participate in the maintenance of the plasmid stability. Interestingly, the G+C content of the *vagCD* genes (~70%) is significantly higher than that of the *rmpA2* (31.9%), which is located only 1.1 kb away, implying that *rmpA2* and *vagCD* were recruited onto pLVPK independently.

Iron acquisition systems

The capability of iron acquisition is generally a prerequisite for a pathogen to establish infection when entering the hosts. In pLVPK, two siderophore-mediated iron acquisition systems, *iucABCDiutA* and *iroBCDN*, were identified. The *iucABCDiutA* operon, which was first reported on pCoIV-K30 in *E. coli* (Ambrozic et al., 1998), consists of five genes responsible for synthesis and transport of the hydroxymate siderophore aerobactin. The presence of the aerobactin synthesis and utilization genes has also been reported for *Salmonella*, and *Shigella* spp., indicating that the genes are freely transferable within the Enterobacteriaceae. This notion is also consistent with the finding that the *iucABCDiutA* gene cluster is flanked by two transposable elements, IS*630* and IS*3*, and 3' sequences of *E. coli* K12 tRNA^{Lys} and tRNA^{Trp}, which have been proposed to play a role in the horizontal transfer of PAIs between bacterial pathogens (Hou, 1999).

The *iroBCDEN* gene cluster, first described in *Salmonella enterica,* is known to participate in the uptake of catecholate-type siderophores. Recently, similar gene cluster contained in a PAI was also found either on the chromosome or a transmissible plasmid in the uropathogenic *E. coli* (Sorsa et al., 2003). It should be mentioned that the *iro* gene cluster in pLVPK lacks *iroE* gene. Nevertheless, the absence of *iroE* gene probably would not affect the utilization of catecholate siderophore by the bacterium since it has been demonstrated in *E. coli* that an *iroE* mutation does not hinder the siderophore utilization activity (Sorsa et al., 2003).

A two-gene operon that encodes a ABC-type transporter related to *Mesorhizobium loti* FepBC was noted on pLVPK at nucleotide positions 77450..80256. The identity between the pLVPK genes and FepBC is 38% and 44%, respectively. These genes also share significant homology with many ABC transporters mediating translocation of iron, siderophores, and heme (Koster, 2001). Although the contribution of this putative ABC transporter in the uptake of iron remains unclear, it is undoubtedly advantageous for the bacteria to have multiple iron acquisition systems in order to obtain iron from the frequently changing environment.

Finally, a gene cluster similar to *E. coli fecIRA*, which is responsible for regulating the uptake of ferric citrate in a Fe^{2+}-Fur dependent manner was identified approximately 3 kb upstream of the *iroBCDN*. In *E. coli*, *fecIR* genes are within a large gene cluster with *fecABCDE* that are the structural genes for iron citrate uptake and are thought to be the target of FecIR regulatory system (Braun et al., 2003). However, there is no observable *fecABCDE* homologs in pLVPK. This phenomenon is not that unusual. As shown in Figure 4, the homologs of *fecIRA*, but not *fecBCDE*, have been identified experimentally in *Bordetella spp.* as well as in several other bacterial species. It is not clear what the target genes are for these FecIRA-like regulatory systems in these bacteria (Braun et al., 2003). One possibility is that a *fecABCDE* gene cluster could be located on *K. pneumoniae* chromosome. Alternatively, the FepBC-like ABC-type iron transporter encoding genes on pLVPK could be the target gene of the FecIRA regulators.

It should be pointed out here that the pLVPK *fecR* open reading frame is disrupted by an in-frame termination codon. FecR is an inner membrane protein that senses whether FecA, the outer membrane ferric citrate receptor, is bound to the substrate, and in response activates FecI, which is known as a transcription factor. Deletion analysis of the *fecR* in *E. coli* has shown that a minimum of 59 amino acids in length of the FecR N-terminal derivative is still able to activate the FecI and subsequently a constitutive expression of the downstream target genes (Ochs et al., 1995). Thus, despite the presence of an internal stop codon, the *fecR* of pLVPK may still be capable of encoding a truncated but functional product and may result in a constitutive iron acquisition phenotype in *K. pneumoniae* CG43.

The hydroxamate-bioassay with the aerobactin indicator strain *E. coli* LG1522 showed that the plasmid-cured strain, CG43-101, loses the aerobactin activity in comparison with its parental strain CG43. In addition, the iron acquisition activity assay revealed that CG43-101 apparently has a smaller growth zone around the iron-loaded disc. These results indicated that the iron-acquisition capability of the bacteria could mostly be attributed to the plasmid pLVPK.

Genes related to metal resistance

Heavy metals at certain concentrations in the cell may form unspecific complex compounds leading to a toxic effect. Many genes for the maintenance of the heavy metal ion homeostasis have been identified in bacteria. Three physically linked gene clusters, as shown in Figure 5 (152306..177234 bp), were identified in the pLVPK that are related to metal resistance phenotype in *K. pneumoniae*. These gene clusters include homologs of the lead-resistance genes *pbrRSABC* of *Ralstonia metallidurans* CH34 (Borrenmans et al., 2001), the copper-resistance genes *pcoEABCDRS* of *E. coli* plasmid pRJ1004 (Brown et al., 1995), and the silver-resistance gene cluster *silCBAPsilRSE* of *S. enterica* serovar Typhimurium (Gupta et al., 1999). By using disk diffusion assay, we have found that the resistance against silver and copper ions between *K. pneumoniae* CG43 and a plasmid-cured strain, CG43-101 remain the same.

A putative lead resistance gene cluster, *pbrRABC,* showed a 63~71% deduced amino acid sequence identity with that of the *R. metallidurans pbrTRABCD genes*. The *R. metallidurans* lead resistance operon, carried on a large plasmid, pMOL30, contains *pbrT* for Pb^{2+} uptake; *pbrA*, for Pb^{2+} efflux; *pbrB* for a putative integral membrane protein; *pbrC* for a putative prolipoprotein signal peptidase; *pbrD* that confers lead sequestration; and *pbrR* that regulates the transcription of *pbrABCD*

(Borremans et al., 2001). Unlike that of the *R. metallidurans,* the *pbr* gene clusters of pLVPK contains only the efflux system (*pbrABC*) and regulator encoding genes (*pbrR*) (Figure 5), which suggest a simple lead-efflux mechanism similar to that of the CadA ATPase of *Staphylococcus aureus* and the ZntA ATPase of *E. coli* (Rensing et al., 1998). In contrast to the indifference of copper and silver ion resistance, the lead susceptibility increased in the disk diffusion assay after curing of the plasmid. The *pbr* genes in the pLVPK may contribute to the adaptation of *K. pneumoniae* in lead polluted human inhabitants.

A gene cluster encoding *E. coli terZABCDE* homolog was also identified. The *terZABCDE* has been shown previously to be a part of a PAI, which also contains integrase, prophage, and urease genes in *E. coli* EDL933 (Taylor et al., 2002). This gene cluster also provides the resistance to bacteriophage infection as well as resistance to pore-forming colicins. Although *terBCDE* are sufficient for the tellurite resistance property, the functions of each of these genes are unknown. The 14.7 kb region (19890..34588 bp) containing *terZABCDE* genes and 12 putative ORFs of pLVPK are comparable to the *ter* genes-containing region in the *E. coli* O157 genome. The homology is interrupted downstream of the *terZABCDE* region by an *E. coli* pTE53 tellurite resistance *terF* homolog and IS*903* gene (Figure 6a). A recent study suggests that the Te^r-containing pathogenicity island in enterohemorrhagic *E. coli* isolates was acquired from plasmid. With considerable degree of sequence homology (75~98% amino acid sequence similarity respectively with that of the *E. coli* O157 *terZABCDE*), the *ter* genes of the pLVPK are likely horizontally acquired. It has been speculated that the *ter* system most likely plays other functional roles such as protection against host defenses so as to be stably maintained in the bacterium (Taylor et al., 2002).

A chromosomally located ORF which showed 77% amino acid sequence identity with the *E. coli* tellurite resistant gene *tehB* (Taylor et al., 2002) has also been recently isolated in our laboratory from *K. pneumoniae* CG43. Deletion of the *tehB*-like gene had no apparent effect on tellurite resistance of the bacteria (Figure 6b) suggesting that the tellurite resistance of the bacteria is determined by the *ter* gene cluster of pLVPK rather than the *tehB* homolog.

Replication and plasmid maintenance

DNA sequence analysis also revealed a single plasmid replication region of 1,756 bp (217448..219203 bp), which consists of *repA* and sequence elements with characteristics of plasmid replicons that employ an iteron-based replication initiation and control mechanism (Chattoraj, 2000). The *repA* product showed a high sequence similarity to a number of plasmid replication initiation proteins, including RepFIB of *Salmonella enterica* serovar Typhi R27 plasmid (60% identity), RepFIB of *E. coli* O103:H2 (43% identity), RepA of *Yersinia pestis* KIM plasmid pMT-1 (42% identity), and RepA of *S. enterica* serovar Typhi plasmid pHCM2 (42% identity). As shown in the multiple sequence alignment in Figure 7a, RepA appears to be an initiator for plasmid replication, which is able to bind the flanking repeated sequences through its DNA binding structures, a winged-helix domain and a leucine-zipper motif (Chattoraj, 2000). We have also found two sets of iterons, four 21 bp and thirteen 42 bp direct repeats, located respectively at the upstream and downstream of the *repA* locus (Figure 7b). The sequences are most likely the specific binding sites for the RepA protein to initiate replication of the plasmid and also control the plasmid copy number (Chattoraj, 2000).

Segregation control machineries

A region (203493..203994 bp) consisting of 11 copies of a 43-bp repeat (5'-gggaccacggtcccacctgcatcgtcgtttaggttttcagcct-3'), is believed to be required for segregation control of the plasmid. Next to the 43-bp direct repeat pattern are positioned the genes encoding *sopA* and *sopB* homologous. The organization is comparable to that of the *sop* operon which governs the partition of the F plasmid (Yates et al., 1999). In addition to *sopAB,* genes showing sequence similarity with *parAB* of *E. coli* P1 phage were identified. It has been shown previously that the corresponding partitioning site in the P1 *parAB* system is composed of direct or inverted repeats (Davis and Austin, 1988). We also noted that a 66-bp direct repeat upstream of the *parAB* homologs is found, which indicates that they also contribute to the partitioning control of the pLVPK. It is reasonable that such a large plasmid has meticulous maintenance systems. Nevertheless, how these two partitioning systems contribute to the maintenance of pLVPK remaine to be confirmed.

Heterogeneity

Pathogenic bacteria have obtained a significant proportion of their genetic diversity by acquisition of DNA from other organisms. Many of the gene clusters identified in pLVPK are homologous to the unknown gene clusters in the other organisms. Although with unknown functions, the homologs of the gene clusters contained in the 9 kb region from nucleotide 2522 to 11618 and the 5.9 kb region from nucleotide 13997 to 19886 were found respectively in the genome of *Burkholderia fungorum* and *Yersinia pestis* KIM. A gene cluster which encodes a putative ABC transporter system (117432..113670 bp) is also identified for which the deduced amino acid sequences are similar to those of the putative ABC transporter system of *Streptomyces coelicolor* A3. A region (46979..51336 bp) comparable to the phage infection inhibition *pif* region of *E. coli* F plasmid was also identified. The boundary sequences of these gene clusters, as well as that of the PAI-like region, are mobile elements including insertion sequences and short pieces of 3'-sequences of tRNA genes. With the involvement of the transposons and the tRNA sequences, horizontal gene transfers have made possible these gene clusters to be introduced into the plasmid and hence affect the ecological and pathological characteristics of bacteria.

Chapter 2

Evolutionary analysis of the two-component systems in *Pseudomonas aeruginosa* PAO1

2.1 Introduction

The two component system (2CS) is the means by which bacteria commonly regulate an adaptive response to versatile environments. A 2CS often comprises of a sensor histidine kinase and a response regulator (Stock et al. 1990). The sensor kinase consists of at least one signal recognition (input) domain coupled to an autokinase (transmitter) domain. Signals binding to the input domain cause activation of the autokinase and thereby, hydrolysis of an ATP molecule to phosphorylate a conserved histidine residue (Stock et al. 1989). The phosphate group is subsequently transferred to the conserved aspartate residue at the receiver domain of a response regulator.

Most sensor kinases contain one *i*nput domain and one *t*ransmitter domain and are hence called classical (IT-type) sensors. Some sensors contain both the sensor kinase signature and a *r*eceiver domain of the response regulator and are thus referred to as hybrid or ITR-type sensory kinases (Ishige et al. 1994). A smaller fraction of the hybrid sensors possesses an additional *o*utput domain at the carboxyl terminus and are referred to as ITRO-type or unorthodox sensor kinases. The response regulator, in most cases, is a transcription factor for genes whose expressions correspond to the input signal. Phosphorylation of the aspartate residue activates the output domain to modulate an appropriate expression of the target gene (Parkinson 1993). Response regulators other than transcription factors have also been reported. For example, the CheY response regulator, after being phosphorylated by the kinase CheA, binds to a flagella motor to promote a clockwise rotation of the flagella (Macnab, 1996).

The tight connection between the functionally coupled bacterial genes and their chromosomal vicinity is a common feature of bacterial genomes (Overbeek et al. 1999; Dandekar et al. 1998). Most of the 2CS genes encoding functionally coupled sensors and regulators are also physically linked as an operon in the genomes. Two models, the co-evolution and recruitment models have been proposed to explain the evolution of 2CS genes. The co-evolution model proposes that the majority of the 2CS genes in a genome have been aroused by gene duplication and a subsequent differentiation of the ancestral 2CSs (Koretke et al. 2000). This is supported by the fact that many of the coupled 2CS genes are concurrent in a genome. On the other hand, the recruitment model suggests that some of the 2CS operons have evolved as a result of an assembly of a sensor gene and a regulator gene from heterologous 2CSs. Signal transductions between 2CSs encoded by distantly located genes have been reported in the sporulation system of *B. subtilis* (Kobayashi et al., 1995) and *E. coli* hybrid sensor kinases respectively (Hoch and Sihavy, 1995). It is conceivable that, in the recruitment model, further assembly of such distantly located 2CS genes into an operon would be beneficial for a coordinate control of the system.

Pseudomonas aeruginosa is a flexible Gram-negative bacterium that grows in a variety of environmental habitats. Patients with cystic fibrosis, burn victims, and patients requiring extensive hospitalization are particularly at risk of *P. aeruginosa* infections (Goldberg et al. 2000). The complete genome sequence of *P. aeruginosa* PAO1 has been determined and published (Stover et al. 2000). The 6.3-Mb genome contains 5,570 predicted genes, of which 123 2CSs were annotated according to the most recently updated database of the Pseudomonas Genome Project. The number of 2CS genes in *P. aeruginosa* genome is relatively high in comparison with that in the *E. coli* and *Bacillus* genomes, which is likely to be advantageous for the bacteria to adapt to different environments. Nevertheless, the function of approximately two-thirds of the 2CS genes has not been characterized. In this study, we have performed analyses of the phylogenetic relationship of the 2CS genes in *P. aeruginosa* PAO1 in the hope that it might reveal some implication of their functions.

2.2 Materials and Methods

Nucleotide Sequence Source and Sequence Analysis

The known and putative 2CS genes annotated by *Pseudomonas aeruginosa* Community Annotation Project (PseudoCAP) were obtained from the web site http://www.pseudomonas.com. The sequences of sensor kinase genes, hybrid sensor genes, and response regulator genes were collected and processed into FASTA format. Analysis of the 2CS was performed by homology search using the BLAST programs provided by the National Center of Biotechnology Information through the Internet.

Multiple Sequence Alignment and Phylogenetic Estimation

Neighbor-Joining (NJ) trees built with the deduced amino acid sequences for sensor kinases and response regulators were done by CLUSTAL W 1.81 (Tompson et al. 1994). Default substitution matrix (Gonnet) was used for alignments, and the positions with gaps were excluded in the tree construction. The resultant trees were visualized by TreeView 1.6 (Page, 1996) and MEGA2 (Kumar et al. 2001).

For the maximum likelihood analysis, multiple sequence alignments of the amino acid sequences of sensors and regulators from homologous gene clusters were performed, also using CLUSTAL W. The positions containing alignment gaps were subsequently excluded manually using BioEdit 4.8.6 (Hall, 1999). Pair-wise distances were analyzed by the PROML algorithm (with JTT amino acid change model) in PHYLIP 3.6 (Felsenstein, 1993) and 1000 replications of bootstrap sampling were performed for each analysis. Graphical representations of the multiple amino acid sequence alignments, the sequence logos, are presented using WebLogo (Crooks et al., 2004).

GC%, G+C Content in the 3rd Position of Synonymous Codons (GC3s), and The Effective Number of Codons Used in a Gene (Nc)

The GC% and GC% in the 3rd position of synonymous codons (GC3s) were calculated using CodonW (Peden 1999). Nc, the measure of overall codon bias in a gene, was calculated by using the CHIPS program with Wright's Nc statistic for an effective number of the codons used (Wright 1990).

2.3 Results and Discussions

Organization of 2CS encoding gene clusters

The 123 annotated 2CSs in the *P. aeruginosa* PAO1 genome, including 64 sensor and 59 regulator genes, were chosen for this study. The discrepancy in the numbers of sensor kinases and regulator genes as compared to the earlier reports of 64 and 63 sensor kinases and regulator genes respectively (Rodrigue et al. 2000) is most likely due to the recent refinement of the annotation by the Pseudomonas Genome Project.

All these 2CS genes were first classified according to their relative location, gene organization, and transcription orientation. As shown in Table 1, each sensor gene was found to be located adjacent to a regulator gene by either direct linkage or separated by less than 3 open reading frames (ORFs) except for 14 sensor genes which were assigned as orphan sensors in Group IV. The most common type of gene organization as represented by the 29 2CS gene pairs in Group I is that the regulator gene was located upstream to the sensor gene. Two 2CS gene clusters within this group contained an additional 2CS gene (a sensor gene in Group Ib, and a regulator gene in Group Ic), which is transcribed in an opposite direction to that of the paired regulator and sensor genes. The Group Id contains 4 gene clusters with one or three non-2CS ORFs located in between the regulator and the sensor genes respectively.

Group II contains 16 pairs of 2CS genes with the gene order of sensor followed by regulator. There are four 2CS gene clusters in Group III, where the regulator and the sensor genes are transcribed divergently. The rest of the 2CS genes, including 14 sensors and 8 regulator genes, are not physically linked to any 2CS gene and were hence referred to as orphan sensors and regulators respectively.

Analysis of the 2CS genes based on functional motifs

These 2CS genes were further analyzed on the basis of the functional motifs of their gene products. The average length of the response regulator genes was approximately 850 bp. Twenty four of the regulators are members of the OmpR transcription factor family, which forms the largest group of the 2CS response regulators in *P. aeruginosa* PAO1. Apart from the 11 NarL-, 8 NtrC- and 5 CheY-type regulators, the rest of the 11 regulators with signal-receiving motifs, were found not to contain the conserved C-domain for classification and therefore were listed as unclassified (Table 1).

In contrast to that of the regulator genes, the size of the sensor genes varies greatly, ranging from 650 bp to 7418 bp. Classification of the 64 sensor genes was as follows - 42 IT (classic), 12 ITR (hybrid), 5 ITRO (unorthodox), and 4 CheA-type based on the functional motifs of their gene products (Rodrigue et al., 2000). The 650-bp

PA0471, which encodes the iron sensor Fur, is the only unclassified sensor gene (Table 1).

Combining the analysis of gene organization and the structural motifs of these gene products, several interesting features were noted as follows:

(1) Almost all (20 out of 23) of the Group Ia gene clusters carry an OmpR-like regulator, and most of the OmpR-like regulators (22 out of 24) have accompanying classical (IT) sensor genes located adjacently downstream. This observation indicates that the gene order of a regulator-to-classical sensor is preferred by the family of OmpR-like regulators. It is likely that most of the regulator-sensor pairs in Group Ia were co-evolved by duplication from an ancestral OmpR-IT pair so that the gene organization remained unchanged. Moreover, 15 out of the 20 OmpR-IT 2CS gene clusters consist of the regulator gene overlapped with the downstream sensor gene which also supports the co-evolution model. In the *E. coli* K12 MG1655 genome, 11 of the 14 2CSs of the OmpR-like family exert the gene order of regulator-to-sensor according to the KEGG database (Kanehisa et al., 2002). A similar phenomenon is also observed in the genome of *Bacillus subtilis*, where all of the 14 2CS genes of the OmpR-family are in a regulator-to-sensor organization (Fabret et al., 1999). It is likely that most of the 2CS gene clusters of the OmpR-family have originated from a common progenitor before the speciation of the proteobacteria and Gram-positive

bacteria.

(2) The NarL-like regulator genes classified in either of the Group I, II, III or IV appeared to link to the corresponding genes of either IT-, ITR-, or ITRO-type sensors suggesting a different strategy from co-evolution. Instead, they are probably recruited components during evolution.

(3) 10 out of 11 ITR-type hybrid sensors are orphans. An exception is PA1396, which is located next to a NarL-like regulator PA1397 in a divergently transcription orientation. The ITR-type sensor, also referred to as the hybrid sensor kinase, contains a regulator-like receiver domain following the input and transmitter domains. Interestingly, most of the ITRO-type sensors, which carried an additional Hpt (histidine-containing phosphotransfer) domain in comparison with that of the ITR-type sensors, are adjacent to a response regulator. It has been demonstrated that the phosphorelay specificity of the ITRO-type sensors, such as the BvgS of *Bordetella pertussis* and EvgS of *E. coli*, was determined by the Hpt domain (Perraud et al., 1998). The phosphorelay between the ITR-type kinases and the corresponding response regulators in *Vibrio harveyi* and *Saccharomyces cerevisiae* also occurred through Hpt modules, which are however encoded by genes distantly located to the 2CS genes (Freeman et al., 1999; Posas et al., 1996). In the *P. aeruginosa* PAO1 genome, three such Hpt module-encoding genes have been identified (Rodrigue et al.

2000). It is likely that without a combined Hpt domain, the ITR sensors act in concert with the Hpt modules to perform the multiple-step phosphorelay in a manner similar to that of the ITRO systems. It has been proposed that a receiver or receiver-Hpt domain may be fused to an IT-type kinase to yield hybrid or unorthodox sensor kinases (Grebe and Stock, 1999). Recruitment of these domains may confer on these systems an additional flexibility as compared to the classical two-component signal transduction.

Phylogenetic analysis of the 2CS genes

The evolutionary relationships among the sensor and regulator genes were estimated by multiple sequence alignment of their deduced amino acid sequences using CLUSTAL W followed by the neighbor-joining method of tree construction. The 2CSs including four CheA-type sensors (PA0178, PA0413, PA1458, PA3704), 5 sensors (PA1396, PA3078, PA3878, PA4197, PA5262, PA0471) with extraordinary length, and one unclassified regulator PA4843 were poorly aligned with the rest of sensors and thus were excluded from the NJ tree construction.

As shown in Figure 8A, the ITR- and ITRO-type sensors apparently formed a group of sub-trees except for the branches h1 and h2. Their close association in the tree suggests that both ITR- and ITRO-type sensors share a common ancestor.

However, only 4 ITRO- and 1 ITR-type sensors were observed in the 49 sensor-regulator pairs. The question of why the multi-step phosphorelaying system is not favored in the bacteria remains to be answered.

Most OmpR-like and NtrC-like regulator encoding genes were found to form a cluster in the tree (Figure 8B), whereas genes encoding the members in the NarL family and the unclassified regulators were scattered in the tree indicating a lower sequence similarity. In order to analyze the historical associations between sensors and regulators, each node in the sensor tree was first assigned an association with the cognate regulator. Each terminal node of the sensor tree was therefore represented by its cognate regulator, while each of the internal nodes was represented by the union of the cognate regulators of all the descendants of that node. For each node in the sensor tree, the corresponding node in the regulator tree with the same descendent regulators was searched for. Subsequent to this, six clades (congruent monophyletic groups) composed of 11 sensor-regulator pairs designated as clade A to F respectively were identified (Figure 8A, 8B). As shown in Table 2, the distances of the 2CS pairs in each clade calculated based on PRODIST were apparently shorter in contrast to those of the 2CS pairs containing NarL-like regulators. The 2CS pairs in each of the clades appeared to be the most closely related and most likely to be derived from a recent gene cluster-duplication event. This is coincident with the report of an extensive co-evolution of the 2CSs in 20 different genomes (Koretke et al., 2000).

In order to further assess the co-evolution relationships, maximum-likelihood (ML) estimation of the phylogeny was subsequently carried out for the specific groups of OmpR-, NtrC-, and NarL-like regulator-containing 2CSs (Figure 9). As shown in Figure 9A and 9B, the 2CS pairs of OmpR and NtrC families also form congruent clades whereas, the ML trees for the 2CS pairs of NarL-like regulators appeared to show different topologies. In Figure 9C, the PA1397 in the regulator tree is only one branch away from PA3879 (*narL*), however, their corresponding sensors PA1396 and PA3878 (*narX*) are distantly located from each other. This is supportive to the recruitment model that the 2CS pairs are assembly products of a sensor and regulator. It is consistent with the distance-based analysis that the distances between these NarL-group 2CS pairs are relatively long or beyond determination (Table 2).

To measure the dissimilarities between the sensor and regulator trees, the resolved and different quartets were determined (Estabrook et al., 1985). For trees of the 23 sensor and regulator pairs of the OmpR- group, the quartet dissimilarity is 4,251. For the 8 sensor - regulator pairs of NarL- and NtrC-groups, the values are 41 and 31 respectively. To compare the tree dissimilarities of different groups, random trees with the same number of OTUs are generated and thence, the dissimilarities measured. The Figure 10 shows the measurement of tree dissimilarities from the set of random trees. Fewer than 3.57% and 5.65% of the random trees have smaller quartet dissimilarties than the data of the OmpR- and NtrC- groups respectively. In contrast, fewer than 19.49% of the random trees have smaller quartet dissimilarities than that calculated for the NarL-group (Figure 10). The results conclusively show that the tree congruencies of the OmpR- and NtrC- group are relatively higher than that of the NarL- group.

Analysis of the sequences around the phosphorelated histidine of the sensor kinases

It has been shown in the *B. subtilis* 2CSs that, classification of the sequences flanking the histidine in the kinase could be correlated to their cognate regulators (OmpR, NarL, etc.) (Fabret et al., 1999). The sequences around the phosphorylated histidine residue were used to further classify the sensor kinases. We have found that the histidine containing motifs could be classified in to three homologous groups and their sequence logos are as shown (Figure 11). The 4 CheY-type regulators were paired with the Class I kinases. Seven out of nine NtrC-type regulators were paired with the Class II kinases. The two exceptions were PA5484 with a regulator-to-sensor transcription order, and PA4293 with two ORFs located in between the regulator and its cognate sensor. Most interstingly, all the OmpR-type regulators were paired with the kinases classified in Class III. Except for a few of the NarL-type regulators which were paired with the Class II and III sensors, most of the others were paired with the sensors showing low sequence similarity around the histidine residue. This is supporting evidence to the hypothesis that sensors paired with their cognate regulators of the NtrC- or OmpR-types have co-evolved as a unit from a common ancestor.

Functional analysis of the most recently duplicated 2CS sensor-regulator pairs

In order to assess functions of the 2CSs identified in each of the congruent clades, we compared these 2CSs with those of the known functions identified in other species and also their adjacent genes with the characterized properties. Several interesting findings were noted:

(1) The 2CS genes in one clade may contain a similar function. For instance, in clade A, the two 2CS gene pairs *pirS* (PA0930)/*pirR* (PA0929) and *pfeS* (PA2687)/*pfeR* (PA2686) are parts of the operons *pirRSA* and *pfeRSA,* respectively (Figure 12A). Both operons encode siderophore-mediated iron uptake systems and are under the control of the Fur protein (Ochsner and Vasil, 1996). This indicates that the paralogous groups continue to carry out a similar function after gene duplication. Functional redundancy is also seen in the members of clade C: PA3044/PA3045 and PA3946/PA3948, which are likely virulence-related 2CS paralogs (Figure 12B). Both the 2CS gene pairs exhibit significant sequence homology with those of *Bordetella*

parapertussis bvgAS that has been demonstrated to participate in regulating the synthesis of many virulence factors (Bock and Gross 2001). Moreover, the regulator gene PA3947 has been reported to encode a homologue with a 45% sequence similarity to *Vibrio cholerae* virulence-related protein VieA (Lee et al., 1998) and to the regulator PvrR that controls antibiotic susceptibility and biofilm formation in *P. aeruginosa* PA14 (Drenkard and Ausubel 2002). The clustering structure suggests a related function since gene clusters in a bacterial genome may possess the same function (Overbeek et al., 1999).

(2) Gene rearrangement may have occurred after duplication of the co-evolved 2CS gene cluster. The virulence-associated 2CS gene PA0928 (*lemA*) is adjacent to the *pirRSA* operon (Figure 12A). In the *Pseudomonas syringae* genome, the *lemA* gene is clustered with the cysteine synthase encoding gene *cysM* (PA0932) in a divergently transcriptional direction suggesting that the region has been subject to gene rearrangement during speciation of *P. syringae* and *P. aeruginosa*. PA0928, which resides relatively distant to the sensor PA0930 in the tree, appears to be recruited later by the 2CS pair PA0929/0930. As shown in Figure 12B, the 2CS pair PA3045 and PA3044 in clade C are co-transcribed. However, the PA3946 and PA3948 in this clade are transcribed divergently, which is probably indicative that the co-evolved 2CS gene clusters have undergone a rearrangement event after the gene

duplication.

(3) A group of the functionally related 2CSs are probably required in controlling the translocation of metabolites and ions. As shown in Figure 13A, all three gene pairs (PA1335/PA1336, PA5165/PA5166, and PA5511/PA5512) in clade D appear to be homologs of *Rhizobium meliloti* DctBD, which controls the transportation of C4-dicarboxylic acids in *R. meliloti* (Wang et al. 1989). The downstream genes clustered with PA5165/PA5166 are homologs of DctPQM (67%, 47% and 72% sequence similarities) that are essential for transportation of the C4-dicarboxylates in *Rhodobacter capsulatus* (Shaw et al., 1991). Several genes encoding homologs of glutaminase-asparaginase (88% sequence similarity) of *Pseudomonas* 7A (Holcenberg et al., 1997), *E. coli* glutamyltranspeptidase (62% sequence similarity) (Suzuki et al., 1988), and *E. coli* glutamate-aspartate ABC transporters (> 68% sequence similarities) (Oshima et al., 1996) respectively were found upstream of the PA1335/PA1336. Moreover, a putative *S. typhimurium* amino acid permease-encoding gene was found adjacent to PA5511/PA5512. The regulator of the four 2CS gene pairs in clade E showed significant similarities (>74% sequence similarities) to the transcriptional regulator IrlR of *Burkholderia pseudomallei*, (Jones et al., 1997) and CopR of *P. syringae* (Mills et al., 1994), which are related to Zn/Cd and Cu resistance respectively. As shown in Figure 13B, the genes PA2523/PA2524,

are part of *czcSRCBA* gene cluster, which have been reported as a Zn and Cd-resistant determinant in *P. aeruginosa* and *R. eutropha* (Nies et al., 1995). A consolidation of all these above findings together with the fact that the 2CS genes are in part of the gene cluster with a functional similarity of being membrane-bound or small molecule transporting proteins according to the Peudomonas Genome Project suggests that even though, divergent evolution is imminent, a conservation of function persists.

Comparative promoter analysis

To elucidate whether the 2CSs which appeared to be resulted from gene pair duplication also shared the same transcriptional control, the upstream non-translated sequence for each of the 2CS gene clusters of the six clades was collected and the sequences were searched using BLASTN against the DPInteract database in which *E. coli* transcription factor binding sites were collected (Robinson and Church, 1994). In the iron transport operons, *pfeRS* and *pirRS* of the clade A, a *fep* signature (Wang and Church, 1992) was identified in the upstream non-translated region of the *pirR* (Figure 12A). However, the signature was not found in that of the homolog *pfeR* suggesting a different regulation for the two gene clusters. This finding is supported by the report that different levels of Fur proteins were required for the regulatory control (Dean et al., 1996; Ochsner and Vasil, 1996). Furthermore, a *P. syringae*

CopR regulator binding site P*copH* (Mills et al., 1994) identified upstream of the CopR gene homolog *czcR* is not found in the other members of the clade E, which also suggested a differential control of the gene expression for copper resistance (Figure 13B).

We compared further the sequences with that of their homologous 2CSs in other *Pseudomonadaceae* including *P. aeruginosa, P. fluorescens, P putida, P. syringae and Azotobacter vinelandii.* Surprisingly, sequence similarity was observed only when *P. aeruginosa* PA14 sequences were compared. The conserved sequences upstream of these 2CSs suggest a preserved response to regulatory signals in *P. aeruginosa*. Nevertheless, the degeneracy of the transcriptional factor binding sequences and the faster rate of accumulation of sequence variations in non-coding regions should be taken into consideration. It is conceivable that retaining functionally redundant genes is not economically favorable for the bacteria. The comparative promoter analysis indicated that the duplicated gene clusters have most likely evolved in order to perform similar functions under a different regulatory control.

GC%, GC3s and Nc of the 2CS genes

After analysis and comparison of the GC% and Nc for the 2CS genes, two findings emerged (i) None of the 123 2CS genes were found in the regions with typical low GC content of the PAO1 genome. (ii) No signs of recent horizontal gene transfer events were evident on comparison with the average GC% and codon usage. In view of these findings, and the fact that *P. aeruginosa* has a higher number of 2CS genes relative to its genome size, it is reasonable to speculate that this is a result of expansion of bacterial genome and hence its ubiquity in nature.

In summary, we have analyzed the *P. aeruginosa* 2CS genes by the criteria of gene organization, the functional motifs and sequence similarity determined by phylogenetic tree construction and their similarity measurements. Using these approaches, we have identified evidence that support both the co-evolution and recruitment models for the evolution of 2CSs.

Chapter 3

Protein interaction specificity and conservation of the gene organization of the two-component systems – An analysis of the OmpR-family

3.1 Introduction

Two-component signal transduction system (2CS) is commonly seen in the regulation of adaptive responses in bacteria. This system often contains a sensor histidine kinase and a response regulator protein. Signals activate the sensor kinase by causing an autophosphorylation on the conserved histidine residue. The phosphorylated sensor then transfers this phosphate group to a conserved aspartic acid residue of the corresponding response regulator, which thereby modulates the subsequent cellular responses (Hoch, 2000). Most of the sensor kinases consist of a highly variable membrane sensor domain that connected to the conserved histidine phosphotrasferase region with a cytoplasmic linker segment. The response regulator, in most cases, consists of an amino-terminal receiver domain followed by a linker region and a carboxyl-terminal DNA-binding domain (Fabret et al., 1999; Hoch, 2000).

Analysis of the 2CSs in *Bacillus subtilis* revealed that the classification of regulator family could be correlated with the sequences surrounding the phosphorylated histidine residue of the cognate kinase, which suggested that the catalytic domain of kinase and DNA-binding domain of the response regulator co-evolved as a unit (Fabret et al., 1999). Similar result has also been reported for the

2CS genes in *Enterococcus faecalis* (Hancock and Perego, 2002). On the basis of sequence homology in the DNA binding domain of the response regulators (Mizuno, 1997), two major families, OmpR and NarL, were identified in *Pseudomonas aeruginosa* PAO1. We have shown that the gene order of regulator-to-sensor transcription unit is preserved within the 2CSs of OmpR-family in the genome. The conserved sequences surrounding the active-site histidine of the sensor kinases were shown also correlated to the classification of their corresponding response regulators (Chen et al., 2004), which implied that the protein interaction specificity and gene organization were preserved in the evolution of the 2CSs of OmpR family.

The OmpR-like proteins are one of the most widespread transcriptional regulators and constituted the largest family of the response regulators (Chang and Stewart, 1998). The family contains at least 15 proteins with extensive amino acid sequence similarities to that of OmpR (Mizuno and Tanaka, 1997), including PhoP, PhoB, KdpE, ArcA, and CreB. The primary osmosensor EnvZ and the response regulator OmpR are the prototype of the 2CS required for transducing an osmotic stress signal in bacteria. Upon activation, OmpR functions as a transcriptional factor to modulate the expression of *ompF* and *ompC* genes that encode porins for controlling the cell membrane permeability (Roberts et al., 1994; Delgado et al., 1993). Other members in the OmpR family appeared to be involved in a diverse functions such as virulent activity (PhoQ/PhoP), phosphate regulation (PhoR/PhoB), and anaerobic nitrite reduction (ResD/ResE) (Hoch and Silhavy, 1995).

Bacteria generally possess multiple 2CSs. To ensure a specific signal transduction, it is conceivable that a preserved specificity of the interacting doamin is present between the kinase and the corresponding response regulator. It is of our interest to identify the constraint through evolution between the functional coupled 2CS components, the sensor kinase and the cognate regulator. We report herein an extended analysis, concerning the gene order and the sequence conservation in the interacting domain, of the 2CSs of OmpR-family from 38 species, chosen to represent the diversity. Our results indicated that the specificity between the 2CS components may have formed before assembly of the genes to from an operon.

3.2 Materials and Methods

Database searches

The completed genomic sequences of archaea (5 species), gamma-proteobacteria (10 species), beta-proteobacteria (5 species), *Chlorobi* (1 species), proteobacteria delta-epsilonsubdivision (1 species), alpha-proteobacteria (1 species), Gram-positive bacteria (5 species), *Actinobacterium* (4 species), *Deinococcus* (1 species), *Cyanobacteria* (3 species), *Aquificae* (1 species), and *Thermotogae* (1 species) were collected to search for the proteins containing respectively the sensor kinase transmitter domain, response regulator receiver domain, and the DNA-binding domain of OmpR-type regulator. The HMM for the transmitter domain (HisKA.hmm), receiver domain (Response_reg.hmm), and OmpR C-terminal (Trans _reg_C.hmm) were obtained from the pfam database (http://www.sanger.ac.uk/Software/ Pfam/) (Bateman et al. 2004). To look for these domains in the query sequences of total ORFs from the collected genomes, HMMSEARCH (HMMER version 2.2) (Durbin et al., 1998) was used.

The ORFs that contain both the receiver domain and OmpR C-terminal domain were designated as “OmpR-type” regulators. The regulator encoding genes that appeared to have accompanied ORFs nearby bearing a histidine kinase transmitter domain were noted as “paired 2CS”. Alternatively, the regulator encoding genes without the physically linked sensor gene were nominated as ‘orphan regulators’.

Multiple sequence analysis and estimation of phylogeny

The phylogeny of the selected organisms was produced based on maximum-likelihood analysis of the 16S rRNA, which were obtained from the European ribosomal RNA database (Wuyts et al., 2002). Sequences alignment were performed by ClustalW (Thompson et al., 1997) and phylogenetic analyses were performed using PAUP* 4.0 (Swofford, 1998). A maximum likelihood tree was obtained based on a general time reversible (GTR) model. The tree was rooted with the eukaryotic lineage of Arabidopsis and Yeast. The genome sequence and ORFs of *Klebsiella pneumoniae* str. NTUH-K2044 was kindly provided by Dr. Shih-Feng Tsai at NHRI, Taiwan (http://genome.nhri.org.tw/kp/).

Sequence logos

The amino acid sequences around the phosphorylated histidine residue of each histidine kinases were identified after multiple sequence alignment, which was performed using ClustalW with the BLOSUM62 (Henikoff and Henikoff 1992) similarity matrix and gap opening and extension penalties of 10.0 and 0.05, respectively according to the parameters used (Koretke et al., 2000). The sequences in length of 14 amino acids were collected and graphical representations of the multiple sequence alignments, the sequence logos, were presented using WebLogo (Crooks et al., 2004).

3.3 Results and Discussions

Compilation of the 2CSs of the OmpR family

The ORFs from 38 species, including 5 archaea, and 33 eubacteria were collected from GenBank. As shown in the maximum-likelihood analysis based on the 16S rRNA sequences, the phylogeny built for the chosen species were placed in the clades that correspond to the major taxons, which include the archaea, purple bacteria (proteobacteria), Gram-positives, Actinomycetes, Cyanobacteria, Deinococcus, Thermotogales (Figure 14). Corresponding to the canonical view of bacterial taxonomy, such phylogeny indicated that the chosen species represent a wide diversity of different archaea and eubacteria. We firstly searched the ORFs using HMMER (Durbin et al., 1998) for the three domains including the transmitter domain of sensor kinase, the receiver domain of response regulator, and the C-terminal DNA-binding domain of OmpR response regulator. As shown in Table 3, a bacterium that carries more HisKA (histidine kinases) generally contains more Receivers (response regulators). Most of the closely-related species, for example, those of the *Salmonella* genus, appeared to have similar number of 2CSs. Except the 5 archaeal genomes, the other bacteria appeared to carry a number of OmpR-like regulators. These OmpR-like regulator-encoding genes were accompanied with a sensor kinase gene mostly in a regulator-to-sensor transcriptional orientation. Some of the OmpR-like regulator genes do not have a linked kinase gene and hence named orphan regulators (OP). A relatively small portion of the OmpR-like regulator encoding genes revealed a different transcriptional orientation with their accompanied sensor genes. Interestingly, more orphan OmpR-like regulators were found in the cyanobacteria, including *Synechococcus*, *Synechocystis* and *Prochlorococus*.

Conventional view of the universal tree is that archaea and eukaryotes are sister groups rooted in the bacteria, and the three were separated into monophyletic groups (Brown and Doolittle, 1997). The genes encoding 2CSs were identified in all three domains of life (Woese et al., 1990), however, mostly present in bacteria. Among eukaryotes, only yeast, fungi, slime molds, and plants appeared to contain 2CSs. Only a single histidine kinase and three response regulators were found in the completed yeast *Saccharomyces cerevisiae* genome (Koretke et al., 2000). In our analysis, no OmpR-like response regulator was found in the five archaea genomes, which suggested that the OmpR 2CS genes were originated from a common ancestor in bacteria thereby spread to most, but not all, of the bacteria species. These genes were also passed to some of the eukaryotes, probably through horizontal gene transfer events. Conservation of the interacting domain of the sensor kinases suggested that, although co-evolution of the paired 2CS genes of the OmpR-faimly is the major scheme, recruitment of the interacting components from distantly located sensor and response regulator genes may still be important for a flexible regulation in bacteria.

Correlation of the 2CSs number with the genome size

As shown in Figure 15, our analysis indicated that the numbers of 2CS components including sensor kinase, response regulators, and OmpR-like regulators were positively correlated with their genome sizes. The *P. aeruginosa* genome appeared to carry the largest number of 2CS genes among the chosen species. This may explain the flexible habitats of the bacteria that, as opportunistic pathogens, they are not only able to survive in various environments, but also adapt rapidly to the changing host conditions (Rodrigue et al., 2000). The alpha proteobacteria *Rhodopseudomonas palustris*, which is among the most metabolically versatile bacteria capable of photoautotrophic, photoheterotrophic, chemoheterotrophic, or chemoautotrophic growth (Larimer et al., 2004), also harbored a large number of 2CS genes. The *Streptomyces coelicolor A3*, known as the most numerous and ubiquitous soil bacteria (Bentley et al., 2002), likewise, carried more than hundred of 2CS genes. This could be concluded that the relatively large number of the 2CS genes in a genome allow the bacteria to grow in many different environments.

Since the criteria used for histidine kinase identification may have resulted in underestimation of the number, the numbers in Table 3 appeared to be smaller than that of the annotation results in some of the genomes. Wheras, the numbers of response regulator may be slightly overestimated since some of the 2CSs appear as a hybrid in which a protein contains the transmitter domain of a sensor fused with a response regulator receiver domain. However, the hybrid sensors only accommodate for a minor portion of the sensor kinases, which should not result in significant bias to the cross-species comparison.

Gene organization of the 2CS of OmpR-family

We have reported recently that, in the *P. aeruginosa* PAO1 genome, most of the OmpR-type regulator-containing 2CS systems, as an operon, revealed the gene order of regulator-to-kinase (RS) (Chen et al., 2004, in press). This phenomenon was also seen in the analysis of the 2CSs in *B. subtilis* (Fabret et al., 1999). In the 38 genomes analyzed, we have identified 325 ORFs that contained both a response regulator receiver domain and an OmpR-like C-terminal domain. Among them, 258 (79.3%) carried an accompanied sensor kinase in the same transcription direction, 50 appeared to be orphan regulator genes, the other 14 showed different transcriptional directions with the accompanied sensor genes. In the 258 regulator-sensor pairs, 220 (85.3%) formed a gene order of regulator-to-sensor (RS). The majority of the less found sensor-to-regulator (SR) 2CSs are homologs of *kdpDE* and *baeSR* (Polarek et al., 1992; Baranova and Nikaido, 2002), which imply that the majority of these SR-type 2CSs probably also have a shared ancestry.

Analysis of sequences flanking the phosphorylated histidine residue of the sensor kinases

Signal transduction between the sensor and regulator which are encoded by distantly located genes has been reported in the sporulation system of *B. subtilis* and the hybrid sensor systems in *E. coli* (Kobayashi et al., 1995; Hoch and Sihavy, 1995). The 2CSs genes involved in the same signaling pathway may later evolve into an operon for the benefit of a coordinate control. We have shown here the gene order of SR was not uncommon in the bacterial genomes. Most likely, as the abundantly found RS, these SR 2CSs appeared to be evolved from a common ancestor of SR gene order. It has been proposed that, for all the RS-2CSs of OmpR-family in *B. subtilis*, the sequences surrounding the phosphorylated histidine residue of the sensor kinases were co-evolved with the cognate response regulator (Fabret et al., 1999). To clarify that the co-evolved shceme also apply for the SR-2CSs of OmpR-family, the amino acid sequences for RS and SR 2CS sensors from the 38 genomes were aligned using CLUSTAL W with the BLOSUM62 similarity matrix followed to the parameters used by Koretke et al. As shown in Figure 16, the sequence logos derived from the aligned sequences for both RS and SR groups appeared to be nearly identical, which suggested that the interaction of the sensor and regulator components might have preserved in the sequences before the formation of either RS- or SR-2CS.

In contrast to the OmpR-family, 2CSs of the other regulator families such as NtrC and NarL appeared no obvious evidence of co-evolution with the sequences of their cognate sensor kinases (Chen et al., 2004). Their relatively smaller numbers and lacking of uniform gene organization with the accompanied cognate sensor component nearby in the genome probably distorted the clues of co-evolution. By compilation of 2CSs of OmpR-family from 38 species in this study, irrelevant of the gene orders, the constraint in their interacting domains, the domain that contained the phosphorylated histidine residue of the sensor kinase and the output domain of the response regulator, appeared to be preserved. Since the sequences flanking the phosphorylated aspartate residue are highly conserved, the response regulators were classified by the relatedness of their output domains. It has been shown that only a small number of residues around the active sites of *B. subtilis* response regulator Sp0F affected its specificity of interaction, and therefore the sequences at the receiver domain alone could not allow sufficient distinction of the response regulators (Tzeng and Hoch, 1997). Recently, by classification of the response regulators of *B. subtilis* and *E. coli* using surface hydrophobicity of their receiver domains, significant correlation was shown between the receiver domain sub-class and the sensor kinase classifications (Kojetin et al., 2003). In addition, the linker region that joins the receiver domain and output domain was also shown to play an important role both in the phosphorelay and output of the signal (Mattison et al., 2002). Albeit that we were not able to show the correlation of the two components by using the sequence comparison of the receiver domain, it is plausible that, the correlation between the classification of the kinase and the output domain of the response regulator could be supportive to propose the presence of a constraint between the two interacting components during evolution.

Summary

The plasmid pLVPK (GenBank accession no. AY378100) appeared to be the largest (219 kb) ever reported in *K. pneumoniae*. According to the GenBank record at August 2004, it is the 3rd largest sequenced plasmid among the gamma-proteobacteria, and the 6th largest among the 532 sequenced bacterial plasmids. Homology analysis is no doubt the central methodology of genomics that produces the bulk of useful information. However, approximately 2/3 of the ORFs in pLVPK were annotated as "conserved hypothetical" and "hypothetical" proteins, of which there were no functional predictions at all. Apparently, there is ample room for improvement in computational annotation. Several recently developed approaches, known as genome context analysis, have gone beyond sequence or structure comparison. In genome context analysis, the genes that have no experimentally characterized homologs can be assigned to particular cellular systems or pathways based on the associations, such as phyletic profiles of protein families, domain fusions in multi-domain proteins, gene adjacency in genomes, and expression patterns (Huynen et al., 2000; Galperin and Koonin, 2000). Unlike the traditional homology analysis, results produced by these methods are often very intuitive. It is foreseeable that, with more genomic information available, context-based methods will substantially complement the traditional methods and improve the situation.

We have shown that the virulence-related genes encoding the siderophore system and the mucoid regulator constitute two PAI-like regions in pLVPK. As shown in the Appendix I, the first PAI-like region includes the *iuc*, *vag*, *shiF* and *rmpA2* flanking by transposable elements and short sequences similar to that of the 3'-sequences of *E. coli* tRNAs. The second PAI-like region contains *rmpA* and *iro* genes together with a set of insertion sequence genes near the *iro* operon. Since both of the PAI-like regions contain virulence-associated genes and potentially mobile elements, the study focused in both their contribution to virulence and their prevalence among different *K. pneumoniae* isolates is being carried out in our laboratory. Plenty of questions aroused from the sequence analysis await to be answered, such as: How exactly do these gene clusters contained in the PAI-like regions orchestrate in different environmental conditions? Why are the DNA sequences of the two mucoid regulator genes, *rmpA* and *rmpA2*, have such a low GC% and biased codon usage? What are the functions of the putative ORFs in this plasmid? The availability of the completed sequence and annotation information have promised much to the study of the pathogenesis of *K. pneumoniae*.

In the second chapter, our results supported the idea that most of the 2CS gene clusters, especially that of the OmpR-family, are composed of co-evolved sensor and regulator pair in *P. aeruginosa* PAO1. We have shown that similar biological functions were preserved for the closely-related 2CSs in the congruent clade, however, different transcriptional control was suggested to prevent the functional redundancy. The analysis of hybrid kinases, on the other hand, supported the recruitment model. The co-evolution pattern was also reported in the other interacting proteins such as neuropeptide and the receptors (Darlinson and Richter 1999), and the archaeal chaperonin subunits (Archibald et al., 1999), which have attempted to provide a global view of protein linkage to their biochemical relevance in a genome.

In the third chapter, we reported our study in the evolutionary analysis of the 2CSs of OmpR-family among 38 species. We have found the existence of a evolutionary constraint of the interacting 2CS components, the sensor and the cognate regulator. We proposed here that, for the orphan 2CS of the OmpR family, its cognate kinase could be assigned on the basis of the sequences identified. Although several studies have been made to categorize response regulators and sensors by sequence similarities (Grebe and Stock, 1999; Fabret et al., 1999; Volz, 1993), the classification of response regulator, however, remained an unsolved question since the receiver domain containing the phosphorylated aspartate is highly conserved. Nevertheless, investigation on the evolutionary constraints in these interacting modules increases our knowledge in deciphering the evolution of protein-protein interaction.

In conclusion, these studies provide fundamental insights for the research of bacterial pathogenesis and the molecular evolutionary basis for two-component signal transduction systems.

References

Archibald JM, Logsdon JM, and Doolittle WF (1999) Recurrent paralogy in the evolution of archaeal chaperonins. Curr. Biol. 9, 1053–1056.

Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, deJonge BL et al. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 397, 176-180.

Ambrozic J, Ostroversnik A, Starcic M, Kuhar I, Grabnar M, Zgur-Bertok D (1998) *Escherichia coli* CoIV plasmid pRK100: genetic organization, stability and conjugal transfer. Microbiology 144, 343-352.

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

Baranova N, and Nikaido H (2002) The BaeSR two-component regulator system activates transcription of the yegMNOB (mdtABCD) transporter gene cluster in *Escherichia coli* and increases its resistance to novobiocin and deoxycholate. J. Bacteriol. 184, 4168-4176.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, and Eddy SR (2004) The Pfam Protein Families Database. Nucleic Acids Res. Database Issue 32:D138-D141

Bentley SD, Chater KF, Cerdeno-Tarraga A-M, Challis GL, Tomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CA, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang C-H, Kieser T, Larke L. Murphy L, Oliver K, O'neil S, Rabbinowitsch E, Rajandream M-A, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, and Hopwood DA (2002) Complete genome sequence of the model actinomyces *Streptomyces coelicolor* A3(2). Nature 417, 141-147.

Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glassner JD, Rode CK, Mayhew GF et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277, 1453-1462.

Bock A, Gross R (2001) The BvgAS two-component system of *Bordetella* spp.: a versatile modulator of virulence gene expression. Int. J. Med. Microbiol. 291, 119-130.

Borremans B, Hobman JL, Provoost A, Brown NL, van der Lelie D (2001) Cloning and functional analysis of the *pbr* lead resistance determinant of *Ralstonia metallidurans* CH34. J. Bacteriol. 183, 5651-5658.

Braun V, Mahren S, Ogierman M (2003) Regulation of the FecI-type ECF factor by transmembrane signaling. Curr. Opin. Microbiol. 6, 173-180.

Brown JR, and Doolittle WF (1997) Archaea and the prokaryote-to eukaryotes transition. Microbiol Mol Biol Rev. 61, 456-502.

Brown NL, Barrett SR, Camakaris J, Lee BT, Rouch DA (1995) Molecular genetics and transport analysis of the copper-resistance determinant (*pco*) from *Escherichia coli* plasmid pRJ1004. Mol. Microbiol. 17, 1153-1166.

Censini S, Lange C, Xiang Z, Crabtree JE, Ghiara P, Borodovsky M, Rappuoli R, and Covacci A (1996) cag, a pathogenicity island of *Helicobacter pylori*, encodes type I- specific and disease-associated virulence factors. Proc. Natl. Acad. Sci. USA 93, 14648-14653.

Chang C, and Stewart RC (1998) Regulation of diverse signaling pathways in prokaryotes and eukaryotes. Plant Physiol. 117, 723-731.

Chattoraj DK (2000) Control of plasmid DNA replication by iterons: no longer paradoxical. Mol. Microbiol. 37, 467-476.

Chen YT, Chang HY, Lai YC, Pan CC, and Peng HL (2004) Sequencing and analysis of the large virulence plasmid pLVPK of *Klebsiella pneumoniae* CG43. Gene

337, 189-198.

Chen YT, Chang HY, Lu CL, and Peng HL (2004) Evolutionary analysis on the two-component systems of Pseudomonas aeruginosa PAO1. J. Mol. Evol. (in press)

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, et al. (2001) Massive gene decay in the leprosy bacillus. Nature 409, 1007-1011.

Crooks GE, Hon G, Chandonia JM, and Brenner SE (2004) WebLogo: A sequence logo generator. Genome Res. 14:1188-1190. URL http://weblogo.berkeley.edu

Dandekar T, Snel B, HuynenM, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. 23, 324-328.

Darlinson MG, and Richter D (1999) The 'chicken and egg' problem of co-evolution of peptides and their cognate receptors: which came first? Results Probl. Cell Differ. 26, 1–11.

Davis MA, Austin SJ (1988) Recognition of the P1 plasmid centromere analog involves binding of the ParB protein and is modified by a specific host factor. EMBO J. 7, 1881-1888.

Dayhoff MO, Eck RV, Chang MA, and Sochard MR (1965) Atlas of Protein Sequence and Structure. Vol. 1 National Biomedical Research Foundation, Silver Spring, MD.

Dean CR, Neshat S, and Poole K (1996) PfeR, and enterobactin-responsive activator of ferric enterobactin receptor gene expression in *Pseudomonas aeruginosa*. J. Bacteriol. 178, 5361-5369.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27, 4636-4641.

Delgado J, Forst S, Harlocker S, and Inouye M. (1993) Identification of a phosphorylation site and functional analysis of conserved aspartic acid residues of OmpR, a transcriptional activator for *ompF* and *ompC* in *Escherichia coli*. Mol. Microbiol. 10, 1037-1047.

Drenkard E, Ausubel FM (2002) *Pseudomonas* biofilm formation and antibiotic resistance are linked to phenotypic variation. Nature 416, 740-743.

Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids, Cambridge University Press.

Estabrook GF, McMorris FR, and Meacham CA (1985) Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst. Zool. 34, 193-200.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8, 175-185.

Fabret C, Feher VA, and Hich JA (1999) Two-component signal transduction in *Bacillus subtilis*: How one organism sees its world. J. Bacteriol. 181, 1975-1983.

Felsenstein J (1993) PHYLIP 3.5 (Phylogeny inference package). Department of Genetics, University of Washington, Seattle.

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Min Jou W, Molemans F, Raeymaekers A, Van den Berghe A, et al. (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260, 500-507.

Finlay BB, and Falkow S (1997) Common themes in microbial pathogenicity revisited. Microb. Mol. Biol. Rev. 16, 136-169.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496-512.

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult

CJ, Kerlavage AR, Sutton G, Kelley JM, et al. (1995) The minimal gene complement of *Mycoplasma genitalium* . Science 270, 397-403.

Freeman JA, Bassler B (1999) Sequence and function of LuxU: a two-component phosphorelay protein that regulates quorum sensing in *Vibrio harveyi*. J. Bacteriol 181, 899-906.

Galagan JE et al. (2002) The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. Genome Res. 12, 532-42.

Galperin MY, Koonin EV (2000) Who's your neighbor? New computational approaches for functional genomics. Nat. Biotechnol. 18, 609-613.

Goldberg JB, and Pier GB (2000) The role of the CFTR in susceptibility to *Pseudomonas aeruginosa* infections in cystic fibrosis. Trends Microbiol. 8, 514-520.

Gordon D, Abajian C, and Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res. 8, 195-202.

Grebe TW, and Stock JB (1999) The histidine protein kinase superfamily. Adv. Microb. Phys. 41, 139-227.

Gupta A, Matsui K, Lo JF, and Silver S (1999) Molecular basis for resistance to silver cations in *Salmonella*. Nat. Med. 5, 183-188.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41, 95-98.

Hancock L, and Perego M (2002) Two-Component Signal Transduction in *Enterococcus faecalis*. J. Bacteriol. 184, 5819-5825.

Hansen-Wester I, and Hensel M (2002) Genome-based identification of chromosomal regions specific for *Salmonella* spp. Infect. Immun. 70, 2351-60.

Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, and Murata T et al. (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res. 8, 11-22.

Henikoff S, and Henihoff JG (1992) Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89, 10915-10919.

Hoch J A (2000) Two component phosphorelay signal transduction. Curr. Opin. Microbiol. 3, 165-170.

Hoch J, and Sihavy T (1995) Two-component signal transduction. ASM Press, Washington, D.C.

Holcenberg JS, Ericsson L, and Roberts J (1997) Amino acid sequence of the diazooxonorleucine binding site of *Acinetobacter* and *Pseudomonas* 7A glutaminase--asparaginase enzymes. Biochemistry 17, 411-417.

Hou YM (1999) Transfer RNAs and pathogenicity islands. TIBS. 24, 295-298.

Huynen M, Snel B, Lathe W 3rd, and Bork P (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. 10, 1204-1210.

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, 526-31.

Ishige K, Nagasawa S, Tokishita S, and Mizuno T (1994) A novel device of bacterial signal transducers. EMBO J. 13, 5195-5202.

Jones AL, Deshazer D, and Woods DE (1997) Identification and characterization of a two-component regulatory system involved in invasion of eukaryotic cells and heavy-metal resistance in *Burkholderia pseudomallei*. Infect. Immun. 65, 4972-4977.

Kanehisa M, Goto S, Kawashima S, and Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acid Res. 30, 42-46.

Kobayashi K, Shoji J, Shimizu T, Nakano K, Sato T, and Kobayashi Y (1995) Analysis of a suppressor mutation *ssb* (*kinC*) of *sur0B20* (*spo0A*) mutation in *Bacillus subtilis* reveals that *kinC* encodes a histidine protein kinase. J. Bacteriol. 177, 176-182.

Kojetin DJ, Tompson RJ, and Cavanagh J (2003) Sub-classification of response regulators using the surface characteristics of their receiver domains. FEBS letters 554, 231-236.

Koonin EV, and Galperin MY (2002) Sequence-evolution-function : Computational approaches in comparative genomics. Kluwer Academic Publishers.

Koretke KK, Lupas AN, Warren PV, Rosenberg M, and Brown JR (2000) Evolution of two-component signal transduction. Mol. Biol. Evol. 17, 1956-1970.

Koster W (2001) ABC transporter-mediated uptake of iron, siderophores, heme and vitamin B12. Res. Microbiol. 152, 291-301.

Krogfelt KA (1991) Bacterial adhesion: genetics, biogenesis, and role in pathogenesis of fimbrial adhesins of *Escherichia coli*. Rev. Infect. Dis. 13, 721–735.

Kumar S, Tamura K, Jakobsen IB, and Nei M (2001) MEGA2: Molecular Evolutionary Genetics Analysis software. Bioinformatics 17, 1244-1245.

Kyrpides N (1999) Genomes OnLine Database (GOLD): a monitor of complete and ongoing genome projects world wide. Bioinformatics 15, 773-774.

Lai YC, Peng HL, Chang HY (2003) RmpA2, an activator of capsule biosynthesis in *Klebsiella pneumoniae* CG43, regulates K2 cps gene expression at the transcriptional level. J. Bacteriol. 185, 788-800.

Lai YC, Peng HL, Chang HY (2001) Identification of genes induced in vivo during

Klebsiella pneumoniae CG43 infection. Infect. Immun. 69, 7140-7145.

Larimer FW, Chain P, Hauser L, Lamerdin J, Malfatti S, Do L, Land ML, Pelletier DA, Beatty JT, Lang AS, Tabita FR, Gibson JL, Hanson TE, Bobst C, Torres y Torres JL, Peres C, Harrison FH, Gibson J, and Harwood CS (2004) Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. Nat. Biotechnol. 22, 55-61.

Lee SH, Angelichio MJ, Mekalanos JJ, and Camilli A (1998) Nucleotide sequence and spatiotemporal expression of the *Vibrio cholerae vieSAB* genes during infection. J. Bacteriol. 180, 2298-2305.

Lewontin RC, Moore JA, Provine WB, Wallace B (2003) Dobzhansky's Genetics of Natural Populations (Origins of the Genetics of Natural Populations) Columbia University Press.

Lowe TM, Todd SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955-964.

Macnab RM (1996) Flagella and motility. In Neidhardt FC et al. (eds) *Escherichia coli* and *Salmonella*, Cellular and Molecular Biology. ASM press, Washington DC, pp.123-145.

Mattison K, Oropeza R, and Kenney L (2002) The linkers region plays an important role in the interdomain communication of the response regulator OmpR. J.

Biol. Chem. 277, 32714-32721.

Maxam AM, and Gilbert W (1977) A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA 74, 560-564.

Menoharan A, Pai R, Shankar V, Thomas K, and Lalitha MK (2003) Comparison of disc diffusion and E test methods with agar dilution for antimicrobial susceptibility testing of *Haemophilus influenzae*. Indian J. Med. Res. 117, 81-87.

Merritt EA, and Hol WG (1995) AB5 toxins. Curr. Opin. Struct. Biol. 5, 165-171.

Miller JF, Mekalanos JJ, and Falkow S. (1989) Coordinate regulation and sensory transduction in the control of bacterial virulence. Science 243, 916–922.

Mills SD, Lim CK, and Cooksey DA (1994) Purification and characterization of CopR, a transcriptional activator protein that binds to a conserved domain (*cop* box) in copper-inducible promoters of *Pseudomonas syringae*. Mol. Gen. Genet. 244, 341-351.

Mizuno T (1997) Compilation of all genes encoding two-component phosphotransfer signal transducers in the genome of *Escherichia coli*. DNA Res. 4,161-8.

Mizuno T, and Tanaka I (1997) Structure of the DNA-binding domain of the OmpR family of response regulators. Mol. Microbiol. 24, 665-670.

Moss JE, Cardozo TJ, Zychlinsky A, and Groisman EA (1999) The *selC*-associated

SHI-2 island of *Shigella flexneri*. Mol. Microbiol. 33, 74-83.

Nassif X, Fournier JM, Arondel J, and Sansonetti PJ (1989) Mucoid phenotype of *Klebsiella pneumoniae* is a plasmid-encoded virulence factor. Infect. Immun. 57, 546-552.

Nassif X, and Sansonetti PJ (1986) Correlation of the virulence of *Klebsiella pneumoniae* K1 and K2 with the presence of a plasmid encoding aerobactin. Infect. Immun. 54, 603-608.

Nies DH, and Silver S (1995) Ion efflux systems involved in bacterial metal resistances. J. Ind. Microbial. 14, 186-199.

O'Brien AD, Tesh VL, Donohue-Rolfe A, Jackson MP, Olsnes S, Sandvig K, Lindberg AA, and Keusch GT (1992) Shiga toxin: biochemistry, genetics, mode of action, and role in pathogenesis. Curr. Top. Microbiol. Immunol. 180, 65-94.

Ochs M, Veitinger S, Kim I, Welz D, Angerer A, and Braun V (1995) Regulation of citrate-dependent iron transport of *Escherichia coli*: *fecR* is required for transcription activation by FecI. Mol. Microbiol. 15, 119-132.

Ochsner UA, and Vasil ML (1996) Gene repression by the ferric uptake regulator in *Pseudomonas aeruginosa*: cycle selection of iron-regulated genes. Proc. Natl. Acad. Sci. USA 93, 4409-4414.

Oshima T et al. (1996) A 718-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 12.7-28.0 min region on the linkage map. DNA Res. 3, 137-155.

Overbeek R et al. (1999) The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA 96, 2896-2901.

Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12, 357-358.

Page RD (1993) COMPONENT: Tree comparison software for Microsoft Windows, version 2.0. The Natural History Museum, London.

Parkinson JS (1993) Signal transduction schemes of bacteria. Cell 73, 857-871.

Payne SM (1993) Iron acquisition in microbial pathogenesis. Trends. Microbiol. 1, 66–69.

Peden JF (1999) Analysis of Codon Usage, Ph.D. Thesis, University of Nottingham. http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html

Peng HL, Wang PY, Wu JL, Chiu CT, and Chang HY (1991) Molecular epidemiology of *Klebsiella pneumoniae*. Zhonghua Min Guo Wei Sheng Wu Ji Mian Yi Xue Za Zhi. 24, 264-271.

Perna NT, Plunkett G III, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA et al. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409, 529-533.

Perraud AL, Kimmel B, Weiss V, and Gross R (1998) Specificity of the BvgAS and EvgAS phosphorelay is mediated by the C-terminal Hpt domains of the sensor proteins. Mol. Micorbiol. 27, 875-887.

Polarek JW, Williams G, and Epstein W (1992) The products of kdpDE operon are required for expression of the Kdp ATPase of *Escherichia coli*. J. Bacteriol. 174, 2145-2151.

Posas F et al. (1996) Yeast HOG1 MAP kinase cascade is regulated by a multistep phosphorelay mechanism in the SLN1-YPD1-SSK1 'two-component' osmosensor. Cell 86, 865-875.

Pullinger GD, and Lax AJ (1992) A *Salmonella dublin* virulence plasmid locus that affects bacterial growth under nutrient-limited conditions. Mol. Microbiol. 6, 1631-1643.

Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, and Whittam TS (2000) Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature 406, 64-67.

Rensing C, Sun Y, Mitra B, and Rosen B (1998) Pb(II)-translocating P-type ATPases. J. Biol. Chem. 273, 32614-32617.

Riley LW, Remis RS, Helgerson SD, McGee HB, Wells JG, Davis BR, Hebert RJ, Olcott ES, Johnson LM, Hargrett NT et al. (1983) Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. N. Engl. J. Med. 308, 681-685.

Roberts DL, Bennett DW, and Forst SA (1994) Identification of the site of phosphorylation on the osmosensor, EnvZ, of *Escherichia coli*. J. Biol. Chem. 269, 8728-8733.

Robinson K and Church G (1994) A database on DNA-protein interactions (Unpublished) (1994). http://arep.med.harvard.edu/dpinteract/

Rodrigue A, Quentin Y, Lazdunski A, Mejean V, and Foglino M (2000) Two-component systems in *Pseudomonas aeruginosa*: Why so many? Trends Microbiol. 8, 498-504.

Rosqvist R, Hakansson S, Forsberg A, and Wolf-Watz H (1995) Functional conservation of the secretion and translocation machinery for virulence proteins of *yersiniae*, *salmonellae* and *shigellae*. EMBO J. 14, 4187–4195.

Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, and Falkow S (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. Proc. Natl. Acad. Sci. USA 97, 14668-14673

Sanger F, Coulson AR, Hong GF, Hill DF, and Petersen GB (1982) Nucleotide sequence of bacteriophage lambda DNA. J. Mol. Biol. 162, 729-773.

Sanger F, Nicklen S, and Coulson AR (1977) DNA sequencing with chainterminating inhibitors. Proc. Natl. Acad. Sci. USA 74, 5463-5467.

Schaberg DR, Culver DH, and Gaynes RP (1991) Major trends in the microbial etiology of nosocomial infection. Am. J. Med. 91, 72S-75S.

Shaw JG, Hamblin MJ, and Kelly DJ (1991) Purification, characterization and nucleotide sequence of the periplasmic C4-dicarboxylate-binding protein (DctP) from *Rhodobacter capsulatus*. Mol. Microbiol. 5, 3055-3062.

Simmons-Smit AM, Verweij-van Vught AM, and MacLaren DM (1986) The role of K antigens as virulence factors in *Klebsiella*. J. Med. Microbiol. 21, 133-137.

Snel B, Bork P, and Huynen MA (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res. 12, 17-25.

Sorsa LJ, Dufke S, Heesemann J, and Schubert S (2003) Characterization of an *iroBCDEN* gene cluster on a transmissible plasmid of uropathogenic *Escherichia coli*: evidence for horizontal transfer of a chromosomal virulence factor. Infect. Immun. 71, 3285-3293.

Stock JB, Ninfa AJ, and Stock AM (1989) Protein phosphorylation and regulation of adaptive responses in bacteria. Microbiol. Rev. 53, 450-490.

Stock JB, Stock AM, and Mottonen JM (1990) Signal transduction in bacteria. Nature 344, 395-400.

Stover CK et al. (2000) Complete genome sequence of *Pseudomonas aeruginosa*

PA01, an opportunistic pathogen. Nature 406, 959-964.

Suzuki H, Kumagai H, Echigo T, and Tochikura T (1988) Molecular cloning of *Escherichia coli* K-12 *ggt* and rapid isolation of gamma-glutamyl-transpeptidase. Biochem. Biophys. Res. Commun. 150, 33-38.

Swofford DL (1998) Paup* 4.0: Phylogentic analysis using parsimony (and other methods). Sinauer Associates, Sunderland, MA.

Taylor DE, Rooker M, Keelan M, Ng LK, Martin I, Perna NT, Burland NTV, and Blattner FR (2002) Genomic variability of O islands encoding tellurite resistance in *Enterohemorrhagic Escherichia coli* O157:H7 isolates. J. Bacteriol. 184, 4690-4698.

Tompson JD, Higgins DG, and Gibson TJ (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.

Tzeng YL, and Hoch JA (1997) Molecular recognition in signal transduction: the interaction surfaces of the Spo0F response regulator with its cognate phosphorelay proteins revealed by alanine scanning mutagenesis. J. Mol. Biol. 272, 200–212.

Vázquez-Boland JA, Suarez M, and Rotger R (1999) Microbial pathogenesis.

Internatl. Micorboiol. 2, 131-132.

Venter JC, Smith HO, and Hood L (1996) A new strategy for genome sequencing. Nature 381, 364-366.

Volz K (1993) Structural conservation in the CheY superfamily. Biochemistry. 32, 11741-11753.

Wacharotayankun R, Arakawa Y, Ohta M, Tanaka K, Akashi T, Mori M, and Kato N (1993) Enhancement of extracapsular polysaccharide synthesis in *Klebsiella pneumoniae* by RmpA2, which shows homology to NtrC and FixJ. Infect. Immun. 61, 3164-3174.

Wang YP, Birkenhead K, Boesten B, Manian S, and O'Gara F (1989) Genetic analysis and regulation of the *Rhizobium meliloti* genes controlling C4-dicarboxylic acid transport. Gene 85, 135-144.

Wang MX, and Church GM (1992) A whole genome approach to in vivo DNA-protein interactions in *E. coli*. Nature 360, 606-610.

Whittam TS, and Bumbaugh AC. (2002) Inferences from whole-genome sequences of bacterial pathogens. Curr. Opin. Genet. Dev. 12, 719-25.

Woese CR, Kandler O, and Wheelis ML (1990) Towards a natural system of organisms: proposals for the domains archaea, bacteria and eucarya. Proc. Natl. Acad. Sci. USA 87, 4576-4579.

Wright F (1990) The 'effective number of codons' used in a gene. Gene 87, 23-29.

Wuyts J, Perriere G, and Van De Peer Y (2004) The European ribosomal RNA database. Nucleic Acids Res. Database issue: D101-3.

Yates P, Lane D, and Biek DP (1999) The F plasmid centromere, *sopC*, is required for full repression of the *sopAB* operon. J. Mol. Biol. 290, 627-638.

Table 1. Classification of the 123 annotated 2CS genes in *P. aeruginosa* PAO1 genome

Group Ia	Group IIa	Group IV
(R→S→)	(S→R→)	(Orphan S)
0179(CY)-0178(CA)	0600(IT)-0601(Nar)	0471(UC)
0463(Omp)-0464(IT)	1098(IT)-1099(Ntr)	1243(ITR)
0756(Omp)-0757(IT)	1336(IT)-1335(Ntr)	1611(ITR)
1157(Omp)-1158(IT)	1636(IT)-1637(Omp)	1976(ITR)
1179(Omp)-1180(IT)	1979(IT)-1980(Nar)	1992(ITR)
1437(Omp)-1438(IT)	2882(IT)-2881(UC)	2177(ITR)
1799(Omp)-1798(IT)	3878(IT)-3879(Nar)	2583(ITR)
2479(Omp)-2480(IT)	4197(IT)-4196(Nar)	2824(ITR)
2523(Omp)-2524(IT)	4494(IT)-4493(UC)	3271(ITR)
2657(Omp)-2656(IT)	4546(IT)-4547(Ntr)	3462(ITR)
2686(Omp)-2687(IT)	4725(IT)-4726(Ntr)	3974(ITR)
2809(Omp)-2810(IT)	5124(IT)-5125(Ntr)	4117(IT)
3045(Nar)-3044(ITRO)	5165(IT)-5166(Ntr)	4112(ITRO)
3077(Omp)-3078(IT)	5262(IT)-5261(UC)	4856(ITR)
3192(Omp)-3191(IT)	5512(IT)-5511(Ntr)	
4101(Omp)-4102(IT)		Group V
4381(Omp)-4380(IT)	Group IIb	(Orphan R)
4776(Omp)-4777(IT)	(S→X→R)	0034(Nar)
4885(Omp)-4886(IT)	3704(CA)-X-3702(CY)	2798(UC)
4983(Omp)-4982(ITRO)		3346(UC)
5200(Omp)-5199(IT)		3604(Nar)
5360(Omp)-5361(IT)	Group IIIa	3714(Nar)
5483(Ntr)-5484(IT)	(S←→R)	4781(UC)
	1396(ITR)-1397(Nar)	4843(UC)
Group Ib	2571(IT)-2572(UC)	5364(UC)
(S←→R→S)		
0928(ITRO)-0929(Omp)-0930(IT)	Group IIIb	
	(S←X→R)	
Group Ic	4036(IT)-X-X-X-4032(Omp)	
(R←→R→S)	4293(IT)-X-X-4296(Nar)	
3948(Nar)-3947(UC)-3946(ITRO)		
Group Id		
(R→X→S)		
0408(CY)-0409(CY)-X-X-X-0413(CA)		
1456(CY)-X -1458(CA)		
3204(Omp)-X-3206(IT)		
4396(UC)-X-4398(IT)		

The arrows indicate the transcription direction of each gene. The gene numbers are given in accordance with that of PseudoCAP. The abbreviations are: R, regulator; S, sensor; X, any open reading frame; CA, CheA-like; CY, CheY-like; Nar, NarL-like; Ntr, NtrC-like; Omp, OmpR-like; UC, unclassified. The genes without an adjacent 2CS gene are listed as orphan S and orphan R.

Table 2. The collective distance estimated for some of the 2CS pairs

$2CS_i$		$2CS_j$		Distance†		Clade/family
Sensor	Regulator	Sensor	Regulator	S_{ij}	R_{ij}	
0930	0929	2687	2686	0.81209	0.86742	A
1158	1157	1798	1799	1.78970	1.68556	B
3044	3045	3946	3948	1.10422	0.59288	C
1336	1335	5165	5166	1.38551	0.75003	D
5165	5166	5512	5511	1.54328	0.77573	D
1336	1335	5512	5511	1.50176	0.73861	D
1438	1437	2810	2809	1.37453	0.63785	E
2810	2809	4886	4885	1.47017	0.52720	E
2524	2523	2810	2809	1.64221	0.67896	E
2524	2523	4886	4885	1.65854	0.66942	E
1438	1437	4886	4885	1.77727	0.70514	E
1438	1437	2524	2523	1.85489	0.69369	E
2571	2572	2882	2881	1.32648	2.75433	F
0600	0601	1396	1397	ND*	2.57472	NarL
0600	0601	3878	3879	ND*	2.40411	NarL
0600	0601	3946	3948	19.73464	2.54159	NarL
0600	0601	4197	4196	ND*	2.70931	NarL
0600	0601	4293	4296	ND*	3.44983	NarL
1396	1397	1979	1980	ND*	1.62793	NarL
1396	1397	4197	4196	10.11732	2.74051	NarL
1979	1980	3878	3879	ND*	1.46125	NarL
1979	1980	3946	3948	18.40756	1.84689	NarL
1979	1980	4197	4196	ND*	2.44646	NarL
1979	1980	4293	4296	ND*	3.64490	NarL
3878	3879	4197	4196	20.78054	2.55514	NarL
3878	3879	4293	4296	ND*	2.79356	NarL

†The distance was calculated for each two 2CS pairs, in which S_{ij} represents the distance between two sensors S_i and S_j, and R_{ij} is the distance between two regulators R_i and R_j. S_{ij} and R_{ij} were calculated by PRODIST in Phylip using deduced amino acid sequences.

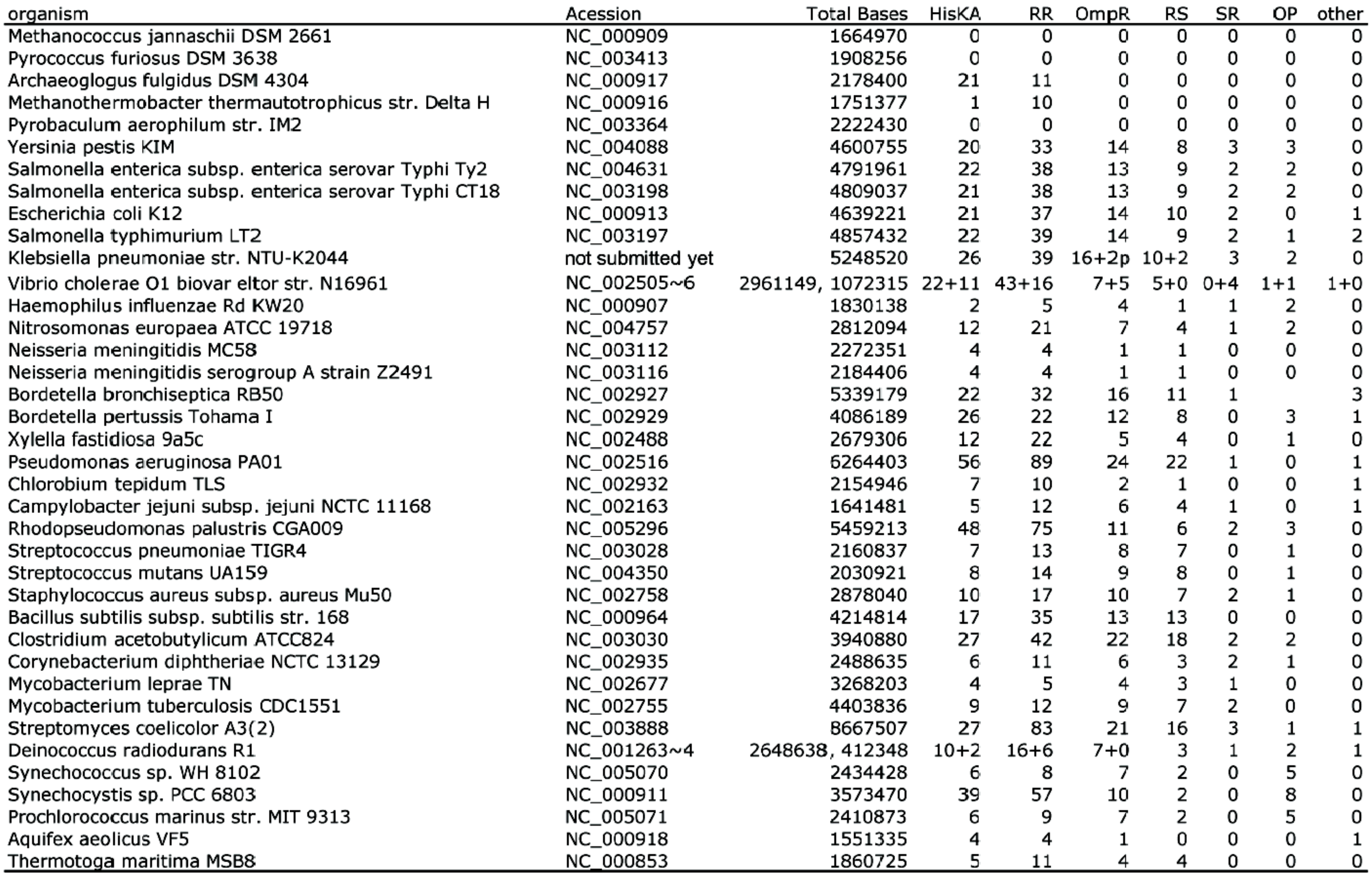
*ND, not determined: For the distances S_{ij} or R_{ij} exceeding the limitation in PRODIST, the distances are designated as ND.

Table 3. Species used in this analysis, the genome sizes, number of 2CS components and their gene organizations.

(see next page)

Note: HisKA, histidine kinase transmitter domain; RR, response regulator receiver domain; OmpR, ORFs that contained both a RR and *E. coli* OmpR C-terminal DNA-binding domain; RS, 2CS gene cluster with a transcriptional orientation in the regulator-to-sensor order; SR, 2CS gene cluster with a transcriptional orientation in the sensor-to-regulator order; OP, orphan OmpR-like regulators; other, OmpR regulators that have accompanied sensor gene in different transcriptional orientation. The *K. pneumoniae* NTUHK2044 genome sequence was obtained kindly from the NHRI, Taiwan. For genomes with more than one chromosome or megaplasmid, the numbers are shown separately in order.

Table 3. (continued)



organism	Acession	Total Bases	HisKA	RR	OmpR	RS	SR	OP	other
Methanococcus jannaschii DSM 2661	NC_000909	1664970	0	0	0	0	0	0	0
Pyrococcus furiosus DSM 3638	NC_003413	1908256	0	0	0	0	0	0	0
Archaeoglogus fulgidus DSM 4304	NC_000917	2178400	21	11	0	0	0	0	0
Methanothermobacter thermautotrophicus str. Delta H	NC_000916	1751377	1	10	0	0	0	0	0
Pyrobaculum aerophilum str. IM2	NC_003364	2222430	0	0	0	0	0	0	0
Yersinia pestis KIM	NC_004088	4600755	20	33	14	8	3	3	0
Salmonella enterica subsp. enterica serovar Typhi Ty2	NC_004631	4791961	22	38	13	9	2	2	0
Salmonella enterica subsp. enterica serovar Typhi CT18	NC_003198	4809037	21	38	13	9	2	2	0
Escherichia coli K12	NC_000913	4639221	21	37	14	10	2	0	1
Salmonella typhimurium LT2	NC_003197	4857432	22	39	14	9	2	1	2
Klebsiella pneumoniae str. NTU-K2044	not submitted yet	5248520	26	39	16+2p	10+2	3	2	0
Vibrio cholerae O1 biovar eltor str. N16961	NC_002505~6	2961149, 1072315	22+11	43+16	7+5	5+0	0+4	1+1	1+0
Haemophilus influenzae Rd KW20	NC_000907	1830138	2	5	4	1	1	2	0
Nitrosomonas europaea ATCC 19718	NC_004757	2812094	12	21	7	4	1	2	0
Neisseria meningitidis MC58	NC_003112	2272351	4	4	1	1	0	0	0
Neisseria meningitidis serogroup A strain Z2491	NC_003116	2184406	4	4	1	1	0	0	0
Bordetella bronchiseptica RB50	NC_002927	5339179	22	32	16	11	1		3
Bordetella pertussis Tohama I	NC_002929	4086189	26	22	12	8	0	3	1
Xylella fastidiosa 9a5c	NC_002488	2679306	12	22	5	4	0	1	0
Pseudomonas aeruginosa PA01	NC_002516	6264403	56	89	24	22	1	0	1
Chlorobium tepidum TLS	NC_002932	2154946	7	10	2	1	0	0	1
Campylobacter jejuni subsp. jejuni NCTC 11168	NC_002163	1641481	5	12	6	4	1	0	1
Rhodopseudomonas palustris CGA009	NC_005296	5459213	48	75	11	6	2	3	0
Streptococcus pneumoniae TIGR4	NC_003028	2160837	7	13	8	7	0	1	0
Streptococcus mutans UA159	NC_004350	2030921	8	14	9	8	0	1	0
Staphylococcus aureus subsp. aureus Mu50	NC_002758	2878040	10	17	10	7	2	1	0
Bacillus subtilis subsp. subtilis str. 168	NC_000964	4214814	17	35	13	13	0	0	0
Clostridium acetobutylicum ATCC824	NC_003030	3940880	27	42	22	18	2	2	0
Corynebacterium diphtheriae NCTC 13129	NC_002935	2488635	6	11	6	3	2	1	0
Mycobacterium leprae TN	NC_002677	3268203	4	5	4	3	1	0	0
Mycobacterium tuberculosis CDC1551	NC_002755	4403836	9	12	9	7	2	0	0
Streptomyces coelicolor A3(2)	NC_003888	8667507	27	83	21	16	3	1	1
Deinococcus radiodurans R1	NC_001263~4	2648638, 412348	10+2	16+6	7+0	3	1	2	1
Synechococcus sp. WH 8102	NC_005070	2434428	6	8	7	2	0	5	0
Synechocystis sp. PCC 6803	NC_000911	3573470	39	57	10	2	0	8	0
Prochlorococcus marinus str. MIT 9313	NC_005071	2410873	6	9	7	2	0	5	0
Aquifex aeolicus VF5	NC_000918	1551335	4	4	1	0	0	0	1
Thermotoga maritima MSB8	NC_000853	1860725	5	11	4	4	0	0	0