8-2012

Application of Bioinformatics to Protein Domain, Protein Network, and Whole Genome Studies.

Kirill Andreyevic Borziak
kborziak@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Kirill Andreyevic Borziak entitled "Application of Bioinformatics to Protein Domain, Protein Network, and Whole Genome Studies.." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Igor B. Jouline, Major Professor

We have read this dissertation and recommend its acceptance:

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Application of Bioinformatics to Protein Domain, Protein Network, and Whole Genome Studies.

A Thesis Presented for the

Doctoral of Science

Degree

The University of Tennessee, Knoxville

Kirill Andreyevic Borziak

August 2012

ACKNOWLEDGEMENS

I would like to thank Igor B. Jouline (Zhulin) not only for his guidance and support over the last six years, but also for his dedication to research and passion for science which have greatly influenced me as a scientist. I would also like to acknowledge my other committee members - Jefferey Becker, Jerome Baudry, Mircea Podar, and Elias Fernandez - for their time, comments, suggestions, and discussions. The faculty and staff of the Genome Science and Technology program also deserve acknowledgement for contributing to such an outstanding graduate program. I would also like to thank past and present graduate students and postdocs within the lab, especially Davi Ortega for being a constant companion and an inspiration. Finally, I would like to thank my parents, Andrei and Natalia, and my fiancée Amanda Eiriksson for their love, support, and persistent encouragement.

ABSTRACT

Bioinformatics primarily focuses on the study of sequence data. Analyzing both nucleotide and protein sequence data provides valuable insight into their function, evolution, and importance in organism adaptation. For this dissertation, I have applied bioinformatics to the study sequence data on three levels of complexity: protein domain, protein network, and whole genome.

In the protein domain study, I used sequence similarity searches to identify a novel FIST (F-box and intracellular signal transduction proteins) domain. The domain was found to exist in all three kingdoms of life, pointing to its functional importance. Due to its presence exclusively with transducer and output domains, it was deduced that FIST functions as an input/sensory domain involved in signal transduction. Further functional characterization revealed FIST's proximity to amino acid metabolism and transport genes. This suggested that FIST functions as a small ligand sensor.

In the protein network study, I examined the evolution of the chemotaxis system within the clade of *Escherichia*. Our study confirmed previous results demonstrating that many urinary pathogenic *Escherichia coli* have lost two of their five chemotaxis receptors. However, sequence analysis demonstrates that this loss occurred as an ancestral event and was not a result of adaptive evolution. The retention of the core of the system in the vast majority of *Escherichia* confirms that chemotaxis is important for survival in all of *Escherichia*'s habitats. However analysis of the loss and gain of chemotaxis receptors suggests that the array of compounds that *Escherichia* needs to sense often does not require all 5 canonical receptors.

In the genome study, I used comparative genomic analysis to examine the evolutionary history of *Azospirillum*, agriculturally important plant growth-promoting bacteria. Taxonomic and genomic studies have revealed that *Azospirillum* are very distinct from their closest relatives in both habitat and genome structure. Comparative genomic analysis revealed that *Azospirillum* had undergone massive horizontal gene transfer. Among acquired genes were many of those implicated in survival in the rhizosphere and in plant growth-promotion. It is proposed that this bacteria's unique

genome plasticity and ability to uptake large amounts of foreign DNA allowed azospirilla to transition from an aquatic to terrestrial environment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ATTACHMENTS

INTRODUCTION

Introduction

Computational biology is the application of computer science, statistics, applied mathematics, and information technology to the study of biology and biological problems. The field of computational biology includes data analysis, molecular modeling, prediction, and simulation. Its interdisciplinary approach allows for unique approaches and solutions to biological problems. Specifically, the area of biological data analysis, bioinformatics, has been of great importance, as the ever expanding capacity of DNA sequencing technologies produces vast volumes of data. Bioinformatics, a combination of statistics and applied mathematics approaches, is used to gain biological insight into sequence data. The ability to process large volumes of data has lent bioinformatics to use in many biological fields, including comparative sequence analysis, genomics, biological literature analysis, macromolecular sequence analysis, metagenomic studies, phylogenetic studies, sequence motif analysis, and transcriptional regulation. Bioinformatics, properly coupled with high throughput biology, was able to transform biological research in those areas [1, 2].

The purpose of bioinformatics is to extract novel biological information from sequence data. Currently with the wide availability of sequence data, we can gain previously impossible insights into evolution of protein universe, on the domain, network, and genome levels. Domain level studies allow us to gain insight into the fundamental building blocks of proteins, domains. Characterization of novel domains provides useful information about their function and evolution, which forms the basis of our understanding of protein function. Network level studies allow us to understand the forces of evolution that affect changes and conservation of protein networks. Understanding the evolutionary pressures that affect protein networks provides insight into their functional importance for various organisms, including how they affect the organism's competitiveness. Genome level studies, such as whole-genome comparison, allow us to better understand the evolutionary forces that shape adaptation and survival. Analyzing the proteome of an organism can provide important insight into its adaptation to its environment. This is critical for understanding the evolved traits of organisms, such as plant growth-promotion and pathogenicity. Using the same bioinformatic

foundation, I was able to study biological problems of three distinct levels of complexity. In this dissertation, I present the insight that was gained using bioinformatics to study domain, network, and genome problems.

History and current state of bioinformatics

From its inception, the focus of bioinformatics was extracting biological insights from sequence data. Development of bioinformatics paralleled advancements in sequencing technology. The origins of bioinformatics can be traced back to the first protein sequence elucidated, insulin, by Frederick Sanger in the 1950s [3-6]. This impressive task was accomplished using a range of chemical and enzymatic techniques and took several years to accomplish. The understanding of biochemical function of proteins that peptide sequences provided was already evident [7]. Consequently, the sequences of many proteins were soon elucidated. As the first sequences were becoming available, groundwork for bioinformatics was being laid with seminal discoveries in biology and computer science. These fundamental discoveries included the structure of DNA [8], the encoding of genetic information for proteins [9], the evolution of biochemical pathways [10] and gene regulation [11]. In parallel, fundamental computer science needed for bioinformatics began to emerged during the 1950s and 1960s with the theory of computation [12], information theory [13], the definition of grammars [14], the theory of games [15] and cellular automata [16].

In the early 1960s, Margaret Dayhoff was amongst the first to appreciate the value of biological sequences, specifically to gain insight into evolutionary relationships [17, 18]. To facilitate further research, she collected and published all protein sequences available at the time in the *Atlas of Protein Sequence and Structure* [19]. With the availability of sequence databases, bioinformatics became possible [20]. The beginnings of bioinformatics approaches combined computational and experimental information to gain insights into the evolution of genes and proteins [21-23], information properties of DNA [24] and proteins [25], sequence alignment [26], construction of phylogenetic trees [27], and processes of molecular evolution [28]. This was followed by

further key developments in sequence alignment algorithms [29, 30], models for selection-free molecular evolution [31], the preferential substitution of amino acid residues in protein sequences [32, 33], derivation of preferences for amino acid residues in secondary structures [34, 35], studies of the origins of life [36], and the theory of evolution by gene duplication [37]. These works laid the foundations for modern bioinformatics.

Despite advances in protein sequence determination, sequencing nucleotide molecules remained complicated due to their size and difficulty of purification. After the development of a new sequencing technology during the 1970s, which became the Sanger method, sequencing of large DNA molecules became possible [38-40]. First demonstrated on Bacteriophage φX 174 [41], this method was quickly adapted to sequence even longer nucleotides, including human mitochondrial DNA and Bacteriophage λ [42, 43]. In bioinformatics, prominent theoretical advancements included the merging of classical population genetics with molecular evolution [44, 45] to produce the theory of neutral evolution [46], the molecular clock hypothesis [47, 48], the development of string and sequence alignment theory [49], evolutionary tree analysis and construction [50], and the evolution of the bacterial genome [51].

With these advances in sequencing, ever larger data sets were becoming available to biologists and bioinformaticians. However, central reference data and software resources and the means to access them were missing. This began the bottleneck shift from data production to data management and analysis. As Gengeras and Roberts wrote in 1980, "*the rate limiting step in the process of nucleic acid sequencing is now shifting from data acquisition towards the organization and analysis of that data*" [52]. This realization led to the development of centralized data banks to manage the growing sequence data. European Molecular Biology Laboratory (EMBL) in Heidelberg was first to set up a public data library, EMBL-Bank [53], with the first release in June 1982 containing 568 sequences. The goal was not only to make sequence data available but also to encourage standardization and free exchange of data [53]. The National Center for Biotechnology Information (NCBI) GenBank [54] was created later the same year, and the DNA Data Bank of Japan (DDBJ) began data bank activities in 1986. Since 1987, these three data banks have collaborated as

the International Nucleotide Sequence Database Collaboration (INSDC) to standardize and share nucleotide data amond the three data banks [55]. By 1980, it had become clear that bioinformatical analysis of nucleotide sequences was essential to the understanding of biology [52]. Developed were key algorithms, such as the Smith–Waterman sequence alignment algorithm [56, 57], the FASTA family of algorithms for database searching [58], and methods for tree-based alignment [59, 60]. Important protein motifs were beginning to be identified through the application of sequence analysis, including of the ATP-binding motif [61], the zinc-finger motif [62], homology of bacterial sigma factors [63], and the signal peptide [64]. Protein evolution had also become a key area of research [65, 66], with discoveries such as the coordinated changes of key residues [67] and the definition of homology [68]. Key analyses of protein families and domains were also performed, including globins [69], bacterial ferredoxins [70], phosphorylases [71], the ribonucleases [72]. The theory and practice of phylogenetic tree construction matured into PHYLIP program [73], and phylogenetic analysis yielded significant discoveries in genome evolution, such as the relationships between life forms [74, 75] and the dynamics of genome structure [76-79].

In the early 1990s two new technological developments, highthroughput DNA sequencing and the Internet, allowed for an overwhelming explosion of biological data and its global dissemination. As a result of the former, whole-genome sequencing was made feasible. In quick succession, the genomes of bacteria, *Haemophilus influenzae* [80] and *Mycoplasma genitalium* [81], in 1995, an archea, *Methanococcus jannachii* [82], and a yeast, *Saccharomyces cerevisiae* [83], in 1996, a nematode, *Caenorhabditis elegans* [84], in 1998, the fruit fly, *Drosophila melanogaster* [85], in 2000, and finally a human, *Homo sapiens* [86-88], in 2001 were sequenced. Since, thousands of genomes have been sequenced. This trend was further continued in the 2000s with the development of next generation sequencing technologies, such as 454 pyrosequencing [89], Illumina reversible dye-terminator sequencing [90], and SOLiD sequencing by ligation [91]. These methods were able to significantly lower the price of sequencing, making genome sequencing widely available to biologists and further increasing the growth of sequencing data [92].

As a consequence of the massive genomic sequencing, more data than could realistically be managed or annotated was beginning to be generated. This wealth of data, however, provided the perfect resources to gain further information into protein composition and function, protein network and genome evolution, and organism adaptation. As sequencing technology flourished, so did the field of bioinformatics. The local alignment search program BLAST was developed to improve the usability of the growing dataset [93]. Due to its heuristic methods BLAST performed significantly better than the previously established Smith-Waterman process. Further improvements created gapped BLAST and Position-Specific Iterative BLAST (PSI-BLAST) [94]. Multiple sequence alignment was also being improved with development of more efficient and accurate programs, including CLUSTAL W [95], T-Coffee [96], and MAFFT, based on the fast Fourier transformation, [97]. Phylogenetics programs were also significantly improved with innovative applications of maximum-likelihood estimation [98], such as PhyML [99] and RaxML [100], and Bayesian inference [101], such as MrBayes [102]. Hidden Markov models [103] were also implemented to model and search protein profiles with the HMMER tool [104].

Currently the focus of the field remains the same: extracting new biological information and insight from the ever growing collection of sequence data. The nr database has over 17 million protein sequences alone, excluding metagenomic samples. New approaches are constantly being developed to take advantage of this increasing wealth of data. Subjects that have been studied from the beginning of the bioinformatics and genomics revolutions, such as novel domain identification and domain and protein function prediction [105, 106], protein system analysis [107], and genome sequencing and analysis [108, 109], are all being redeveloped as more sequence data and increased computational power become available. In this dissertation, bioinformatics was used to identify and functionally characterize the FIST domain, to study the evolution of the *Escherichia* chemotaxis system and its relationship to *E. coli* pathogenicity, and to examine the genomic evolution of *Azospirillum* and its relationship to *Azospirillum*'s niche transition.

Protein domains

Protein domains are compact regions within the protein's structure that possess a distinct function. Each domain also forms a three-dimensional structure that is independently stable and folded. Thus, domains are considered the fundamental units of protein structure, folding, function, and evolution. The average length of a protein domain is approximately 120 amino acids, but they can vary in length from 25 to 500 amino acids. Most proteins consist of several domains [110]. Individual domains, as building blocks, appear in a variety of different proteins. Through the processes of evolution, domains are recombined in various arrangements, creating proteins that possess distinct functions. New domain combinations are typically adapted from pre-existing domain combinations rather than through invention of novel domains. Domains are genetically mobile units. Often, the C and N termini of domains are close together in space. This allows them to be easily inserted into other protein sequences during the process of evolution, creating novel protein architectures. Domains, however, do not form random combinations, since only a limited number of combinations have been seen with few that are abundant [111]. Speaking to their versatility and functional importance, many domain families are found in all three kingdoms of life, and protein families of diverse function [112].

Domain prediction is an important step in the annotation of protein sequences, providing a functional background for annotation [113]. Since domains are often associated with specific cellular functions, domain identification can either predict or refine protein functional predictions [106]. Domains with known structures can also be used to infer protein structure [105]. Domain prediction is also fundamental to a range of other more sophisticated analyses, including comparative genomics of domain families [114], evolution of protein and domain structure and function [115, 116], and protein-protein interaction [117].

Domains are conserved sequential and structural motifs that act as building blocks of proteins above the amino acid sequence level [118]. Due to their importance to understanding the complex nature of protein function, several approaches have been

developed to define and classify domains. Domains have been delimited based on structure and sequence clustering [119, 120]. The Pfam database is the most widely used domain identification and curation database [121]. Pfam domain assignment is straightforward. For each domain, manually selected representative sequences are used to build Hidden Markov Models that are used for annotation of the whole the protein sequence space.

Critically important insight into the function and evolution of novel domains cannot, however, be readily obtained through such high throughput computational approaches and still require stringent analysis of genomic data. Identification and functional characterization of novel domains provides important insight into the functions of proteins and their impact on the ecology of the organism as a whole. Thus, characterization of novel domains is fundamental to the full understanding of protein function. Specifically, identification and functional characterization of sensory domains is critical for understanding how cells sense and respond to their environment.

Pathogenic *Escherichia coli*

Escherichia coli are not only the best studied model bacteria but are also important human pathogens. *E. coli* typically colonize the gastrointestinal tract of infants within the first hours after birth. Typically, *E. coli* coexist with the human host with mutual benefit. These commensal *E. coli* strains cause disease only when normal defenses are compromised, such as in peritonitis or immunocompromised patients. Commensal *E. coli* typically inhabit the mucous layer of the large intestine. They are a highly successful, comprising the most abundant facultative anaerobe. Despite the enormous body of literature on *E. coli*, the mechanisms of its symbiosis are poorly understood.

There are, however, many highly adapted *E. coli* clones with acquired virulence traits, which allow for a broad spectrum of disease. These virulence attributes are frequently encoded on mobile genetic elements that allow for their transfer between

different strains, creating novel combinations of virulence factors, or on genetic elements that were once mobile. The most successful combinations of virulence factors have persisted and are classified as specific pathotypes of *E. coli*, capable of causing disease in healthy individuals. These pathotypes cause one of three general clinical syndromes: enteric/diarrhoeal disease, urinary tract infections (UTIs) and sepsis/meningitis. The intestinal pathogens are further subdivided into six well-described categories: enteropathogenic *E. coli* (EPEC), enteroaggregative *E. coli* (EAEC), diffusely adherent *E. coli* (DAEC), enterohaemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), enteroinvasive *E. coli* (EIEC) [122]. UTIs, caused by uropathogenic *E. coli* (UPEC), are the most common extraintestinal *E. coli* infection. The pathotype responsible for meningitis and sepsis is meningitis-associated *E. coli* (MNEC). The *E. coli* pathotypes implicated in extraintestinal infections are collectively referred to as extraintestinal pathogenic *E. coli* (ExPEC) [123]. Pathogenic *E. coli* can also cause disease in animals using many of the same virulence factors used in human pathogenesis and host specific colonization factors. A specific animal pathotype, avian pathogenic *E. coli* (APEC), causes extraintestinal infections in birds. The various pathotypes of *E. coli* are often characterized by shared O (lipopolysaccharide), K (capsule), and H (flagellin) antigens, which define serogroups and serotypes [124]. However, many cases of horizontal acquisition of pathogenicity in *E. coli* also exist [125]. Pathogenic *E. coli* strains use a scheme similar to that of other mucosal pathogens, which consists of colonization of a mucosal site, evasion of host defenses, multiplication, and host damage [126].

Of particular interest are UPEC strains. The urinary tract is a common site of bacterial infection with *E. coli* being most common infecting agent. UPEC form a distinct group of *E. coli*, differing from commensal and intraintestinal pathogenic *E. coli*. Six O groups are responsible for 75% of UTIs [127]. Although many UTI isolates are clonal, there is no single clade of *E. coli* that contains all UPEC. Specific adhesins and fimbriae aid in colonization [128]. The virulence factors, including adhesins and several toxins, are variably distributed among UPEC and are mainly located on pathogenicity islands unique to UPEC strains [129, 130].

Pathogenesis of UPEC begins with the colonization of the bowel, which serves as the primary reservoir for further infection of the urinary tract. After colonization of the periurethral area, UPEC ascend the urethra to the bladder. Within 24 hours after infection, UPEC begin expression of type 1 fimbriae, which play an important role in attachment to and invasion of epithelial cells [131]. Attachment triggers apoptosis and exfoliation of the epithelial cells, in part through the action of cytotoxic necrotizing factor type 1 [132, 133].

In strains that cause cystitis, type 1 fimbriae are continually expressed. As a result, the infection is confined to the bladder [134]. In pyelonephritis strains, type 1 fimbriae and flagella are alternatively expressed [135]. This allows UPEC to ascend the ureters to the kidneys, where attachment to the kidney epithelium is mediated by P fimbriae [136, 137]. In the kidneys, toxins haemolysin and Sat cause damage to the renal epithelium and glomeruli, respectively [138, 139]. If the endothelial cell barrier of the proximal tubules is breached, UPEC is then able to enter the bloodstream, resulting in bacteraemia.

Bacterial motility, specifically, is a trait associated with virulence of UPEC [136, 140]. Motility is achieved through the function of complex surface structures, flagella, and controlled by a signal transduction system, the chemotaxis system. Although flagellar genes, and the motility phenotype, are poorly expressed during chronic infection, synthesis of flagella coincides with ascension of bacteria from the bladder to the kidneys [136, 141]. Thus, flagella contribute to the ascension of UPEC and their competitiveness over non motile strains. Interestingly, the composition of the chemotaxis system differs between many UPEC and intestinal *E. coli* [142]. UPEC generally lack two of the five receptors commonly found in *E. coli* [142]. Investigating these changes in chemotaxis and motility behavior in UPEC from a genomic and evolutionary perspective can shed new light on the importance of chemotaxis and motility in the pathogenicity of UPEC. The impressive availability of sequenced *E. coli* genomes, over 200, provides vast amounts of genomic information to investigate the origin of changes in the chemotaxis system in *E. coli* and gain insight into how those changes have affected the development of UPEC, or vise versa.

Chemotaxis

Bacteria can sense a vast range of environmental signals, from the concentrations of nutrients and toxins to oxygen levels, pH, osmolarity, temperature, and the intensity and wavelength of light. These behaviors are controlled through complex signaling pathways. The best understood of such pathways, the chemotaxis system, regulates flagellar motility behavior. Chemotaxis is the phenomenon of organisms directing their movements according to specific stimuli in their environment. These stimuli range from chemical compounds such as amino acids and salts, to physical properties such as temperature [143, 144]. Chemotaxis functions as part of a complex network of signals that produce a physiological response to a specific environment. Among complex bacterial behaviors that are dependent on chemotaxis are pathogenicity, symbiosis, and biofilm formation [145-147].

Bacterial chemotaxis is the biasing of movement towards regions that contain higher concentrations of beneficial chemicals or lower concentrations of toxic chemicals. The involved signaling pathway has been extensively studied in the model bacteria *Escherichia coli* and *Salmonella enterica* serovar Typhimurium [148]. Studies have also provided structural and biochemical details for the entire pathway, making chemotaxis one of the best understood sensory pathways. These studies have also shed light on the diversity of this system [107]. Motility and chemotaxis are wide spread throughout bacteria and some archaea indicating their importance to adaptation and survival.

The canonical chemotaxis system in *E. coli* consists of a set of 11 proteins: 5 methyl-accepting chemotaxis proteins (MCPs) act as the receptors; CheA, CheW, and CheY act to transmit the signal to the flagellar motor; and CheB, CheR, and CheZ act in signal adaptation. Canonically, *E.coli* posses 5 chemoreceptors: Tsr, Tar, Trg, Tap, and Aer [149]. Tsr senses the attractant serine. Tar senses the attractants aspartate, and maltose, through maltose periplasmic binding protein (PBP), and repellants nickel and cobalt. Trg senses attractant ribose, through ribose PBP, and galactose/glucose, through a galactose/glucose PBP. Tap senses dipeptides, through dipeptide PBP, and pyrimidines [150]. Aer senses FAD as a measure of redox potential.

Bacteria's main propellers are helical filaments, known as flagella. Rotation is driven by the flagellar motor, powered by electrochemical H+ or Na+ gradient across the cytoplasmic membrane [151]. The chemotaxis pathway controls flagellar rotation by transducing the signal from the chemoreceptors to the flagellar motor using a two-component signaling pathway (Figure 1). *E. coli* swim by rotating their flagella anticlockwise, causing them to come together into a bundle and propel the cell forwards. Switching to clockwise rotation of some motors disrupts the bundle and causes random tumbling. When all the motors return to anticlockwise rotation, the cell is reoriented and begins swimming in a new direction [152]. Bacteria are too small to sense concentration gradients along their length, spatial sensing, and therefore use temporal sensing to bias their movement [153]. The *E. coli* chemotaxis pathway, however, is sensitive enough to detect a change in concentration of a few molecules in background concentrations varying five orders of magnitude [154]. This is accomplished through a system of feedback inhibition.

The receptors, MCPs, are typically dimeric transmembrane proteins with periplasmic ligand-binding domains and cytoplasmic coiled coil signaling domains [155, 156]. Ligand binding induces a conformational change that is transmitted through the cytoplasmic domain to the histidine kinase CheA [156]. In *E. coli* and other bacterial species, chemoreceptors arrange into large clusters that are usually located at the cell poles [157]. In the clusters, the chemoreceptors are organized into "trimers of dimers", which form ternary signaling complexes with CheA and CheW, the linker protein. Allosteric interactions in these clusters allow for signal amplification that results in high sensitivity over a wide range of background concentrations [154].

In response to a decreasing attractant concentration, *E. coli* chemoreceptors induce CheA autophosphorylation [159]. CheA functions as soluble dimer that phosphorylates the response regulator domains of CheY and CheB. The canonical CheA has five structural domains, P1–P5. P1 is the phosphotransfer domain that transfers the phosphate to the response regulators; P2 binds the response regulators; P3 is the dimerization domain; P4 is the kinase domain that binds ATP and autophosphorylates the P1 in the other CheA subunit; and P5 is the regulatory domain that controls kinase activity and binds CheW and MCPs. After phosphorylation, CheY-P

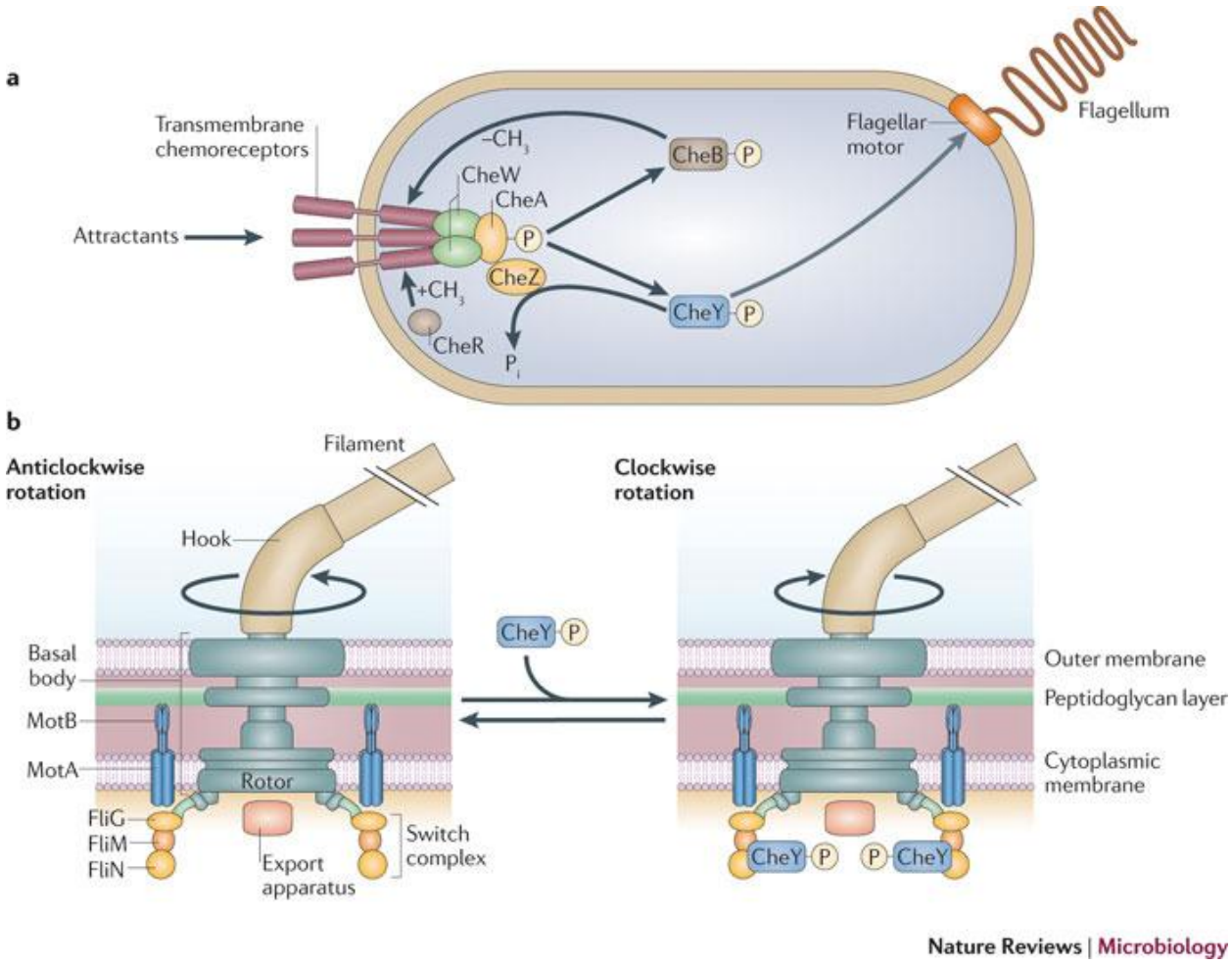

Figure 1: Canonical *E. coli* chemotaxis system.

a) Process of signal transduction through the chemotaxis system. After attactants bind the chemoreceptors, the signal is transduced through CheW to CheA, which phosphorylates CheB and CheY. CheB, CheR, and CheZ act as an adaptation system. b) Phosphorylated CheY binds FliM and FliN, which causes the motor to switch rotation and induce tumbling and a change of direction. Taken from Porter et al. [158].

diffuses to the flagellar motor, where it binds FliM and FliN proteins of the switch complex, promoting a switch in the rotational direction from anticlockwise to clockwise [160] (Figure 1b). The probability of switching increases cooperatively as more subunits, approximately 34 per motor, bind CheY-P [161]. Signal termination occurs by

dephosphorylation of CheY-P, which returns the motor to anticlockwise rotation. This is a rapid process to allow for continuous gradient sensing. In *E. coli*, dephosphorylation is catalyzed by the phosphatase CheZ [162].

Adaptation modifies the signaling state of the pathway by adjusting for the time-averaged attractant and repellant concentrations [163]. This allows bacteria to compare current and previous concentrations and acts as primitive memory. Biochemically, adaptation is achieved through reversible methylation of specific glutamic acid residues in the cytoplasmic signaling domain of MCPs. In *E. coli*, methylation is achieved by the action of methyltransferase CheR. Methylation increases the ability of the chemoreceptors to activate CheA autophosphorylation. Demethylation is achieved by the action of the methylesterase CheB. Demethylation decreases the ability of chemoreceptors to active CheA. While methylation by CheR is relatively constant, methylesterase activity of CheB is increased ~100-fold when CheB is phosphorylated by CheA [164]. Persistent negative stimulus, such as reduction in attractant concentration, results in chemoreceptor activation of CheA autophosphorylation, which leads to elevated levels of CheY-P and tumbling. Additionally, adaptation occurs as CheB-P demethylates active chemoreceptors, reducing their ability to activate CheA, leading to decreased CheY-P levels, and less tumbling. Adaptation restores pre-stimulus tumble bias and thus allows bacteria to be sensitive to a broad spectrum of stimuli concentrations.

Understanding, from an evolutionary perspective, how this important and highly conserved protein network has changed over the course of *E. coli* speciation will provide valuable insight into its importance in adaption and pathogenicity. In uropathogenic *E.coli*, specifically, where chemotaxis has been shown as important for colonization and dissemination [136], evolutionary insight into the history of the chemotaxis system can provide deeper understanding of its effects on virulence. Greater understanding of the origin and effects of the loss of *tap* and *trg* in UPEC will provide additional insight into the development of UPEC as a pathogen and into the importance of chemotaxis and motility in the ascension of UPEC to the kidneys.

Azospirillum

The rhizosphere is the region of the soil influenced by plant roots, their secretions, and plant root associated soil microorganisms. Primarily due to the abundance of nutrients origination from the roots, the rhizosphere has a composition unique from other soils. The abundance of nutrients creates a rich diversity of microorganisms that colonize the rhizosphere. Among them are plant growth-promoting bacteria (PGPB), which are able to influence root formation, plant growth, and crop yield through various factors [165, 166]. This growth-promotion is accomplished either directly through production and secretion of plant growth substances, biological fixation of nitrogen, and/or solubilization of phosphorus [165], or indirectly as a biocontrol agent [167].

Azospirillum are free-living PGPB belonging to the *Rhodospirillales* order and capable of affecting growth and yield of numerous agriculturally significant plants. Azospirilla were first isolated and labeled as *Spirillum lipoferum* by Beijerinck in the 1925 [168]. However, they received little attention until 1974 when Döbereiner and Day isolated them from the roots of tropical grasses in Brazil and was found capable of nitrogen fixation [169]. Azospirilla are typically vibrio- or spirillum-shaped rods capable of chemotaxis and producing both polar and peritrichous flagella. Two main characteristics that define the genus are an ability to fix atmospheric nitrogen and and ability to produce phytohormones [170]. From the beginning of *Azospirillum* plant interaction studies, these two features were considered as cornerstones of *Azospirillum*'s effect on plant growth and yield. Outside rhizobia, *Azospirillum* is the best studied PGPB and has reached commercialization as a biofertilizer in several countries [171, 172]. Although 16 strains of *Azospirillum* have so far been identified, the majority of experimental work is conducted on strains of *Azospirillum brasilense* and *Azospirillum lipoferum* [173, 174]. Considerable knowledge on *Azospirillum*-plant interaction has been accumulated over the past 38 years of study that illustrates the great complexity of this interaction (Figure 2).

In spite of intensive molecular biology and physiology studies, the exact process of plant growth promotion by *Azospirillum* is not much more understood than it was decades ago [174]. However, three important revelations on the nature of *Azospirillum*-plant interaction have emerged. Most *Azospirillum* strains are capable of nitrogen fixation and *Azospirillum* colonization of the plant results in increased nitrogen acquisition [173], but *Azospirillum* derived nitrogen is not a major source of nitrogen in plants [174-176]. Many *Azospirillum* strains are capable of phytohormone production *in vitro* and some were found to produce phytohormones in association with plants [177], and phytohormones have been shown to promote plant growth [178]. However, production of phytohormones alone does not explain all aspects of plant growth-promotion [179, 180]. Although a general improvement in growth is seen in many plant species after inoculation with *Azospirillum*, agricultural yield improvement is not always evident [172, 181]. These revelations show that our understanding of *Azospirillum* as plant growth-promoting bacteria is not yet completely understood.

The major growth-promoting effect on plants inoculated with *Azospirillum* is significant changes in the plants' root structure and physiology. Inoculation have been shown to promote root elongation [178, 183], development of both lateral and adventitious roots [184, 185], increased root hair formation and branching [186-188], thus significantly increasing and improving the structure of the root system. It is generally accepted that these developmental responses in root morphology are triggered by phytohormones, possibly aided by their associated molecules [177, 189]. Additional plant growth-promoting effects in inoculated plants include increased absorption of minerals and water, increased vigor, and enhanced growth [190-193]. Several mechanisms have been suggested to explain plant growth-promotion [173, 194], however there is no consensus explanations.

Complete understanding of the physiological properties of plant-growth promotion and the mechanisms behind them is fundamental to understanding the complex phenomonona of rhizosphere survival and *Azospirillum*-plant interaction. Identifying the responsible mechanisms and protein networks is the desired goal to not only better understand plant growth-promotion but also to improve agricultural production. This question is the driving force in *Azospirillum* research, since a clearer

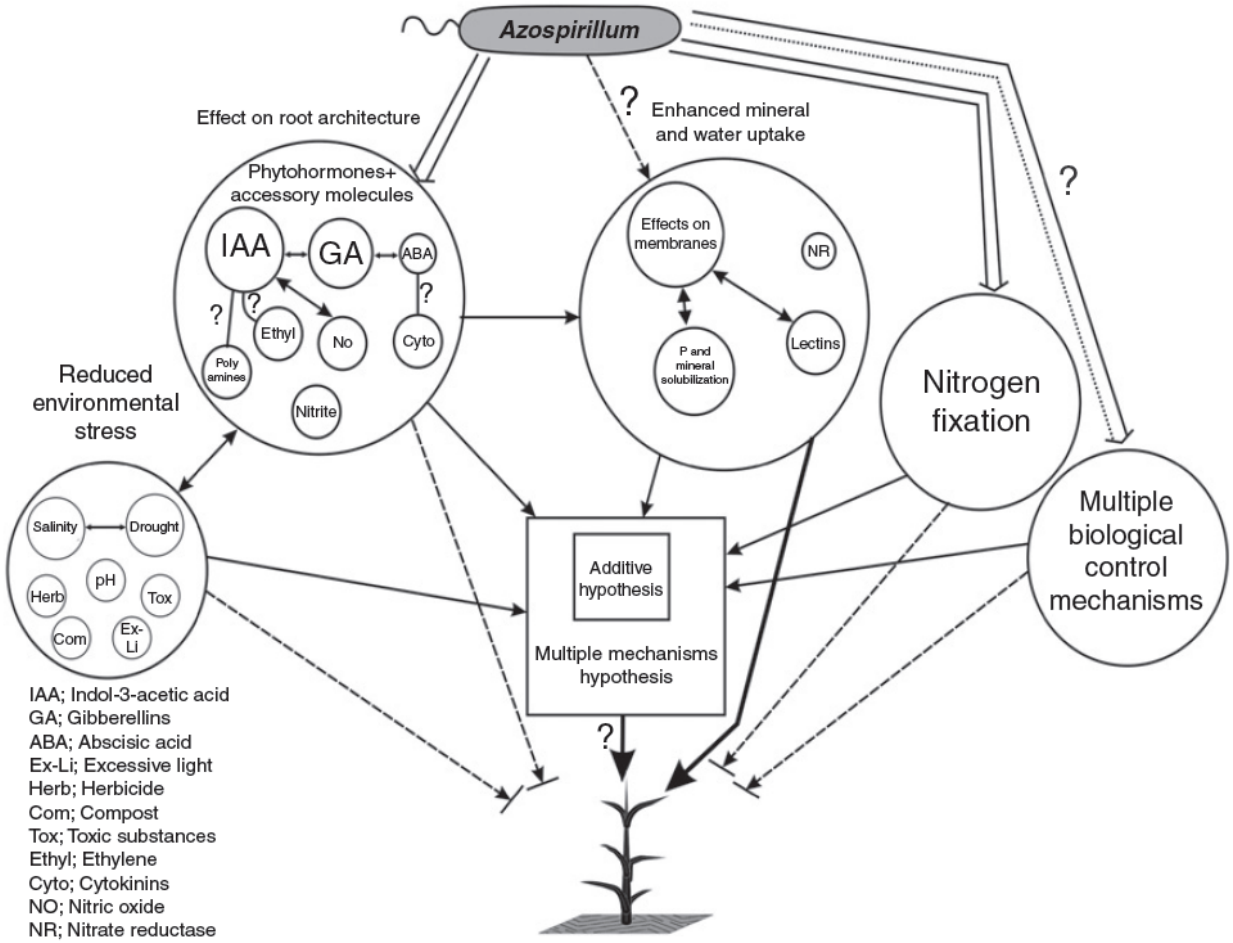

Figure 2: Plant Growth-Promoting Properties of *Azospirillum*.

Properties of *Azospirillum* that may affect plant growth-promotion, grouped by biological processes. Circles represent properties confirmed by experimental data. Squares represent theories. Circle size reflects relative importance based on experimental data. Solid arrows represent experimentally confirmed growth-promotion. Dashed arrows represent unconfirmed growth-promotion. Question marks represent unproven pathways. Taken from Bashan and de-Bashan [182].

understanding of how the bacterium interacts with its host will allow for engineering improved interactions and improve *Azospirillum*'s role as a biofertilizer.

However, with the lack of sequenced genomes for this genus, knowledge of *Azospirillum*-plant interaction has little genomic and evolutionary perspective to help understand the importance of the various mechanisms involved in plant growth-promotion. The genome structure of the genus *Azospirillum* has been examined using molecular techniques, providing evidence of unique genomic architectures [195]. The genomes *Azospirillum* have a great variability size from 4.8 Mbp in *Azospirillum irakense* to 9.7 Mbp in *A. lipoferum* [196]. The genomes of *Azospirillum* species have also been analyzed by pulsed-field gel electrophoresis and horizontal Eckhardt-type gel electrophoresis. *Azospirillum* genomes were shown to possess multiple megareplicons that range in size from 0.2 to 2.7 Mbp and contain ribosomal operons and other ribosomal features [195, 196]. In contrast, genomes of other sequenced members of the class *Rhodospirillaceae* contain no large plasmids. Although the genomes of *Azospirillum* are not unique in containing megareplicons and secondary chromosomes, their existence relative to closely related bacteria provides further incentive to sequence and understand this large complex genome. Evolutionary comparisons of *Rhodospirillaceae* genomes can provide unique insight into the origin of multiple large replicons of *Azospirillum* and advance our understanding of genome dynamics. The size and structure of *Azospirillum* genomes could also be one explanation for the exceptional ecological distribution, metabolic flexibility, and plant-growth promoting properties of this genus [195, 196]. Studying the evolutionary history of *Azospirillum* through genome sequencing and analysis will provide insight into *Azospirillum*'s origin and unparalleled ability as a PGPB, and aid in the attempts to fully understand its plant growth-promoting properties.

Scope of dissertation

This dissertation will describe bioinformatics studies of three biological problems: FIST domain discovery and analysis, evolutionary study of the *Escherichia* chemotaxis system, and adaptation of *Azospirillum* to the terrestrial plant growth-promoting niche. The focus of this dissertation is how bioinformatics and the availability of large

sequence repositories are able to provide unique insight into biological problems of varying scale, from protein domain, to protein network, to whole genome. Chapter 1 will provide a detailed description of the discovery and functional annotation of a novel sensory domain, FIST. The chapter provides a systematic approach to identifying and characterizing highly divergent domains. Chapter 2 will describe the analysis of the *Escherichia* chemotaxis system and provide insight into the changes observed as they relate to the evolution and pathogenicity of *Escherichia*. The focus will be the evolution of this system within *Escherichia* and how the system is maintained by various phenotypes. Specifically, changes in the chemotaxis systems of urinary pathogenic *Escherichia coli*, where chemotaxis system has been shown as important for efficient virulence, will be analyzed from an evolutionary perspective. Chapter 3 will cover the sequencing of two *Azospirillum* strains and their genomic analysis. The focus of the genomic analysis is *Azospirillum*'s transition from an aquatic to a terrestrial habitat and its acquisition of plant growth-promoting traits. We show that these processes were a result of massive horizontal gene transfer, which *Azospirillum* was able to accomplish due to its high genome plasticity. The conclusion will summarize this work and provide future direction.

CHAPTER I

FUNCTIONAL ANALYSIS OF A NOVEL DOMAIN

A version of this chapter was originally published by Kirill Borziak and Igor B. Zhulin in *Bioinformatics*:

Borziak K and Zhulin IB. (2007) **FIST: a sensory domain for diverse signal transduction pathways in prokaryotes and ubiquitin signaling in eukaryotes.** *Bioinformatics* 23(19): 2518-2521

Conceived and designed the experiments: IBZ. Performed the experiments: KB. Analyzed the data: KB IBZ. Wrote the paper: KB IBZ.

Abstract

Motivation

Sensory domains that are conserved among Bacteria, Archaea and Eucarya are important detectors of common signals detected by living cells. Due to their high sequence divergence, sensory domains are difficult to identify. We systematically look for novel sensory domains using sensitive profile-based searches initiated with regions of signal transduction proteins where no known domains can be identified by current domain models.

Results

Using profile searches followed by multiple sequence alignment, structure prediction and domain architecture analysis, we have identified a novel sensory domain termed FIST, which is present in signal transduction proteins from Bacteria, Archaea and Eucarya. Chromosomal proximity of FIST-encoding genes to those coding for proteins involved in amino acid metabolism and transport suggest that FIST domains bind small ligands, such as amino acids.

Introduction

Signal transduction systems are information processing pathways that link environmental cues to adaptive responses in all living organisms. Major prokaryotic and eukaryotic signal transduction systems are different at the level of their principal components and complexity. Prokaryotic signal transduction pathways consist of simple one- and two-component systems [197], whereas signal transduction in eukaryotes often involves multi-protein, branched cascades [198]. However, all living cells react to many similar signals that are present in the environment and inside cells (e.g. small molecules, such as amino acids and carbohydrates), therefore, some level of similarity is expected in their signal input elements. Indeed, several universal input domains have been described in sensory receptors from both prokaryotes and eukaryotes. The most abundant such domain is PAS (named after eukaryotic Per, ARNT and SIM proteins, where it was first described), which is found in diverse signal transduction pathways in organisms ranging from all major prokaryotic clades to humans [199, 200]. PAS domains serve as detectors of small molecules, redox potential, oxygen, light and other important parameters [201]. Other input, sensory elements that are found in both branches of life include small-ligand binding GAF [202], Cache [203] and CHASE [204, 205] domains. On the other hand, many other input domains are limited to bacterial signal transduction, e.g. MHYT [206], NIT [207], CHASE2 through CHASE6 [208] and 4HB_MCP [209], etc. Identification of novel sensory domains is a difficult task due to their extreme sequence variation, both in composition and in length. Sensitive similarity search methods, such as Position-Specific-Iterative (PSI) BLAST [94] must be employed to detect relationships between the domain family members, which must be further verified through a careful analysis of a multiple alignment of the domain family.

Recently developed MiST (Microbial Signal Transduction) database [210] enables rapid identification of sensory receptors where current domain models implemented in leading domain databases Pfam [211] and SMART [212] do not detect any input domains. We carry out systematic similarity searches using protein regions of receptor proteins implicated in signal transduction that contain no identifiable domain. In many cases, such regions do contain known domains; however, current domain models

are not sensitive enough to detect them. In other cases, such regions contain potentially new domain. In this report, we describe the identification of such novel domain, which we termed FIST, which is implicated in sensory reception in diverse signal transduction pathways in all three domains of life: Bacteria, Archaea and Eucarya.

Domain Identification

During systematic analysis of microbial sensory proteins, we focused on a methyl-accepting chemotaxis protein (MCP) blr4191 from *Bradyrhizobium japonicum* (gi27352453), which was predicted to be a cytoplasmic chemotaxis receptor. All known chemotaxis receptors contain a sensory domain in their N-terminus followed by a conserved C-terminal signaling domain [155], which can be identified by a Pfam domain model MCPsignal (accession PF00015). The blr4191 protein contained the full-length C-terminal MCPsignal domain, but lacked any detectable domain in its large (more than 400 amino acid residues) N-terminus. Exhaustive PSI-BLAST searches (*E*-value cutoff 0.01, composition-biased statistics on, filter on) against the NCBI nr database (1 March 2006) were initiated with residues 1–449 that include the entire N-terminal region up to the first residue of the MCPsignal domain and continued until no new homologs were identified. Duplicate sequences were excluded from analysis. Profile searches yielded the list of domain family, which we termed FIST (F-box and intracellular signal transduction proteins), containing 176 proteins. Of these, 155 were full length matches, and 15 matched the C-terminal half of the domain (FIST_C) only. This search also revealed an overlap of the newly proposed domain with the Pfam domain of unknown function DUF1745 (also found in the COG database as COG3287, ‘uncharacterized conserved protein’), which was detected in more than a half of proteins identified by PSI-BLAST searches.

Sequences that matched only the FIST_C subdomain all belong to F-box-containing eukaryotic proteins. When the region of these proteins that did not match the N-terminal half of the FIST domain (FIST_N) was subjected to a PSI-BLAST search, the only profile matches returned were from the same subfamily of F-box proteins.

However, searches initiated with FIST_C from these proteins returned most of full-length FIST domain proteins. Based on results of PSI-BLAST searches and predicted secondary structure, we built domain models from multiple alignments for FIST_N and FIST_C, as well as for the entire FIST domain, and carried out searches using the HMMER program to retrieve all homologs. Resulting sets contained 253 proteins with full-length FIST, 4 proteins with FIST_N only and 30 proteins with FIST_C only (nr database, 1 May 2007). Representative and complete multiple alignments of subdomains are provided as Figure 3 and File 1, correspondingly.

Domain Features and Architecture

Multiple alignment of full-length FIST domain sequences was built using ClustalX [213] with default parameters and edited with the VISSA technique [209], which utilizes PSIPRED [214] to guide the editing process using predicted secondary structure. The final fold profile consists of 20 beta strands and 7 alpha helices. Multiple alignment of the FIST domain family revealed several highly conserved residues. The most conspicuous motif of the FIST domain family, which contains highly conserved cysteine and arginine, is found in the beta-strand 19 and the following loop (Figure 3B). It may represent a site for ligand binding or protein–protein interactions. Using SMART [212] and Pfam [211] domain architecture tools, we determined that in more than 70% of analyzed sequences the FIST domain comprises a single-domain protein; in the remaining sequences it is present in association with well-known signal transduction domains (Figure 4) including the sensory PAS domain, GGDEF and EAL domains involved in the turnover of cyclic di-GMP, the MCP signaling domain, HisKA and HATPase_c domains of sensor histidine kinases, the DNA-binding helix-turn-helix TrmB domain and F-box domains of the ubiquitin signaling pathway. The search for transmembrane regions and signal peptides using Phobius [215] showed that most FIST-containing proteins analyzed in this study are predicted to be intracellular; however, signal peptides were predicted for a few stand-alone FIST-domain proteins indicating that they are likely to be extracellular.

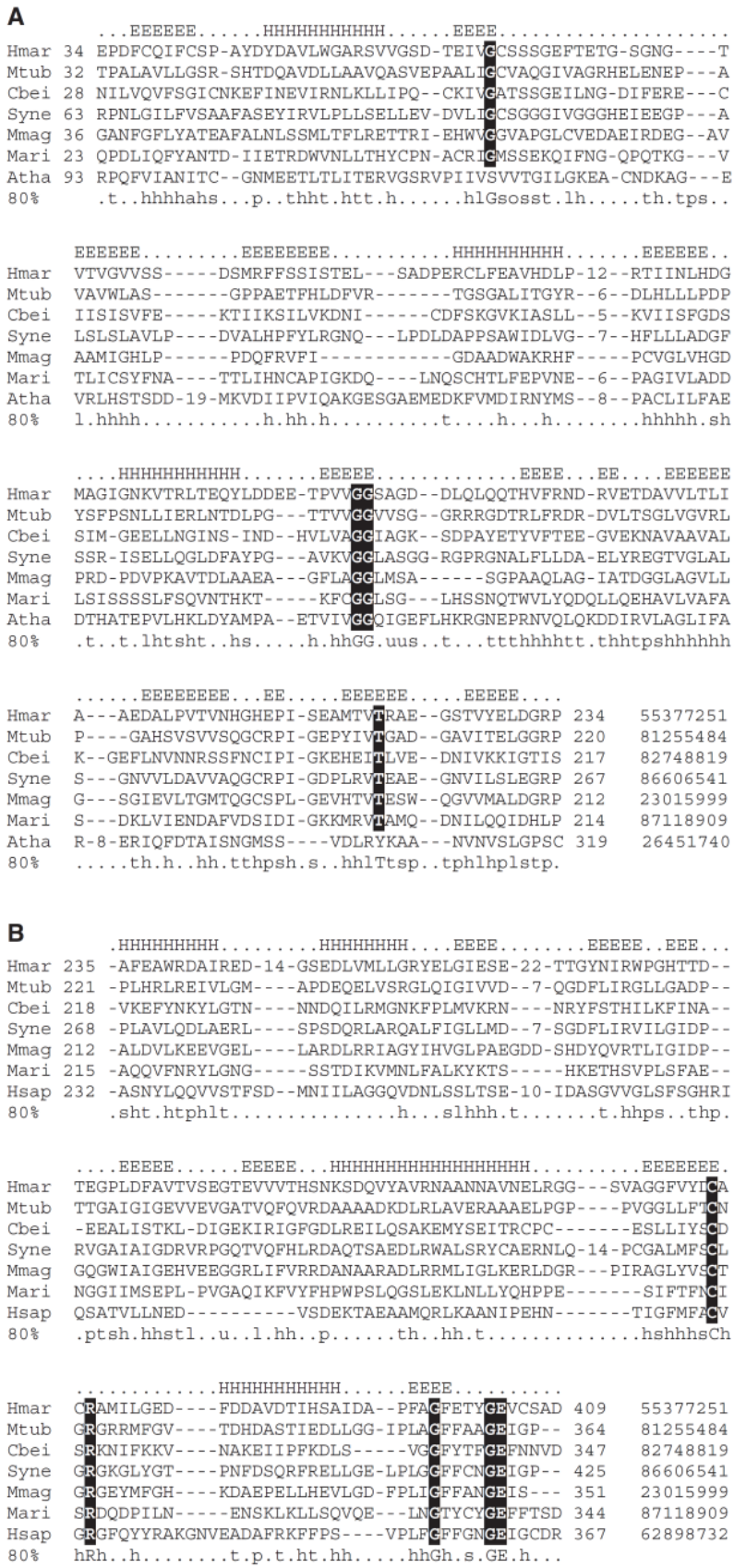
A
...EEEEEE.....HHHHHHHHHH.....EEEE.....
Hmar 34 EPDFCQIFCSP-AYDYDAVLWGARSVVGSD-TEIVGCSSSGEFTETG-SGNG----T
Mtub 32 TPALAVLLGSR-SHTDQAVDLLAAVQASVEPAALIGCVAQGIVAGRHELENEP---A
Cbei 28 NILVQVFSGICNKEFINEVIRNLKLLIPQ--CKIVGATSSGEILNG-DIFERE---C
Syne 63 RPNLGILFVSAAFASEYIRVLPLLSELLEV-DVLIGCSGGGIVGGGHEIEEGP---A
Mmag 36 GANFGFLYATEAFALNLSSMLTFLRETTRI-EHWVGGVAPGLCVEDAEIRDEG--AV
Mari 23 QPDLIQFYANTD-IIETRDWVNLLTHYCPN-ACRIGMSSEKQIFNG-QPQTKG---V
Atha 93 RPQFVIANITC--GNMEETLTLITERVGSRVPIIVSVVTGILGKEA-CNDKAG---E
80% .t..hhhhahs...p..thht.htt.h.....hlGsosst.lh.....th.tps..
EEEEEE.....EEEEEEEE.....HHHHHHHHHH.....EEEEEE..
Hmar VTVGVVSS-----DSMRFFSSISTEL---SADPERCLFEAVHDLP-12--RTIINLHDG
Mtub VAVWLAS-------GPPAETFHLDFVR-------TGSGALITGYR--6--DLHLLLPDP
Cbei IISISVFE-----KTIIKSILVKDNI------CDFSKGVKIASLL--5--KVIISFGDS
Syne LSLSLAVLP----DVALHPFYLRGNQ---LPDLDAPPSAWIDLVG--7--HFLLLADGF
Mmag AAMIGHLP------PDQFRVFI-----------GDAADWAKRHF-----PCVGLVHGD
Mari TLICSYFNA----TTLIHNCAPIGKDQ----LNQSCHTLFEPVNE--6--PAGIVLADD
Atha VRLHSTSDD-19-MKVDIIPVIQAKGESGAEMEDKFVMDIRNYMS--8--PACLILFAE
80% l.hhhh.....h.hh.h.....t....h...h.....hhhhh.sh
....HHHHHHHHHH.....EEEEE.....EEEE...EE....EEEEEE
Hmar MAGIGNKVTRLTEQYLDDEE-TPVVGGSAGD--DLQLQQTHVFRND-RVETDAVVLTLI
Mtub YSFPSNLLIERLNTDLPG---TTVVGGVVSG--GRRRGDTRLFRDR-DVLTSGLVGVRL
Cbei SIM-GEELLNGINS-IND--HVLVAGGIAGK--SDPAYETYVFTEE-GVEKNAVAAVAL
Syne SSR-ISELLQGLDFAYPG---AVKVGGLASGG-RGPRGNALFLLDA-ELYREGTVGLAL
Mmag PRD-PDVPKAVTDLAAEA---GFLAGGLMSA------SGPAAQLAG-IATDGGLAGVLL
Mari LSISSSSLFSQVNTHKT-----KFCGGLSG---LHSSNQTWVLYQDQLLQEHAVLVAFA
Atha DTHATEPVLHKLDYAMPA--ETVIVGGQIGEFLHKRGNEPRNVQLQKDDIRVLAGLIFA
80% .t..t.lhtsht..hs.....h.hhGG.uus..t...ttthhhhtt.thhtpshhhhhh
......EEEEEEEE...EE.....EEEEEE.....EEEEE....
Hmar A---AEDALPVTVNHGHEPI-SEAMTVTRAE--GSTVYELDGRP 234 55377251
Mtub P----GAHSVSVVSQGCRPI-GEPYIVTGAD--GAVITELGGRP 220 81255484
Cbei K--GEFLNVNNRSSFNCIPI-GKEHEITLVE--DNIVKKIGTIS 217 82748819
Syne S---GNVVLDAVVAQGCRPI-GDPLRVTEAE--GNVILSLEGRP 267 86606541
Mmag G---SGIEVLTGMTQGCSPL-GEVHTVTESW--QGVVMALDGRP 212 23015999
Mari S---DKLVIENDAFVDSIDI-GKKMRVTAMQ--DNILQQIDHLP 214 87118909
Atha R-8-ERIQFDTAISNGMSS----VDLRYKAA---NVNVSLGPSC 319 26451740
80% .....th.h..hh.tthpsh.s..hhlTtsp..tphlhplstp.
B
.HHHHHHHHH.....HHHHHHHH....EEEE.....EEEEE..EEE...
Hmar 235 -AFEAWRDAIRED-14-GSEDLVMLLGRYELGIESE-22-TTGYNIRWPGHTTD--
Mtub 221 -PLHRLREIVLGM----APDEQELVSRGLQIGIVVD--7-QGDFLIRGLLGADP--
Cbei 218 -VKEFYNKYLGTN----NNDQILRMGNKFPLMVKRN----NRYFSTHILKFINA--
Syne 268 -PLAVLQDLAERL----SPSDQRLARQALFIGLLMD--7-SGDFLIRVILGIDP--
Mmag 212 -ALDVLKEEVGEL----LARDLRRIAGYIHVGLPAEGDD-SHDYQVRTLIGIDP--
Mari 215 -AQQVFNRYLGNG-----SSTDIKVMNLFALKYKTS-----HKETHSVPLSFAE--
Hsap 232 -ASNYLQQVVSTFSD--MNIILAGGQVDNLSSLTSE-10-IDASGVVGLSFSGHRI
80% .sht.htphlt.....h...slhhh.t.....t.hhps..thp.
....EEEEE......EEEEE...HHHHHHHHHHHHHHHHHH.....EEEEEEE.
Hmar TEGPLDFAVTVSEGTEVVVTHSNKSDQVYAVRNAANNAVNELRGG---SVAGGFVYLCA
Mtub TTGAIGIGEVVEVGATVQFQVRDAAAADKDLRLAVERAAAELPGP----PVGGLLFTCN
Cbei -EEALISTKL-DIGEKIRIGFGDLREILQSAKEMYSEITRCPC-------ESLLIYSCD
Syne RVGAIAIGDRVRPGQTVQFHLRDAQTSAEDLRWALSRYCAERNLQ-14-PCGALMFSCL
Mmag GQGWIAIGEHVEEGGRLIFVRRDANAARADLRRMLIGLKERLDGR---PIRAGLYVSCT
Mari NGGIIMSEPL-PVGAQIKFVYFHPWPSLQGSLEKLNLLYQHPPE------SIFTFNCI
Hsap QSATVLLNED----------VSDEKTAEAAMQRLKAANIPEHN-------TIGFMFACV
80% .ptsh.hhstl..u..l.hh..p......th..hh.t.....hshhhsCh
.....HHHHHHHHHH.....EEEE.....
Hmar CRAMILGED----FDDAVDTIHSAIDA--PFAGFETYGEVCSAD 409 55377251
Mtub GRGRRMFGV----TDHDASTIEDLLGG-IPLAGFFAAGEIGP-- 364 81255484
Cbei SRKNIFKKV----NAKEIIPFKDLS-----VGGFYTFGEFNNVD 347 82748819
Syne GRGKGLYGT----PNFDSQRFRELLGE-LPLGGFFCNGEIGP-- 425 86606541
Mmag GRGEYMFGH----KDAEPELLHEVLGD-FPLIGFFANGEIS--- 351 23015999
Mari SRDQDPILN-----ENSKLKLLSQVQE---LNGTYCYGEFFTSD 344 87118909
Hsap GRGFQYYRAKGNVEADAFRKFFPS----VPLFGFFGNGEIGCDR 367 62898732
80% hRh..h.....t.p.t.ht.hh.....hhGh.s.GE.h...

Figure 3: Representative multiple alignments.

Representative multiple alignments of the FIST_N (**A**) and FIST_C (**B**) subdomains. Representatives of Archaea, Bacteria and Eucarya are included. Predicted secondary structure is shown above the alignment and consensus (80%) calculated for all members of the subdomain family is shown below. GI identification numbers are shown at the end of each sequence. Highly conserved residues are highlighted. Species abbreviations: Atha, *Arabidopsis thaliana*; Cbei, *Clostridium beijerincki*; Hmar, *Haloarcula marismortui*; Hsap, *Homo sapiens*; Mari, *Marinomonas* species; Mmag, *Magnetospirillum magnetotacticum*; Mtub,*Mycobacterium tuberculosis* and Syne, *Synechococcus* species.

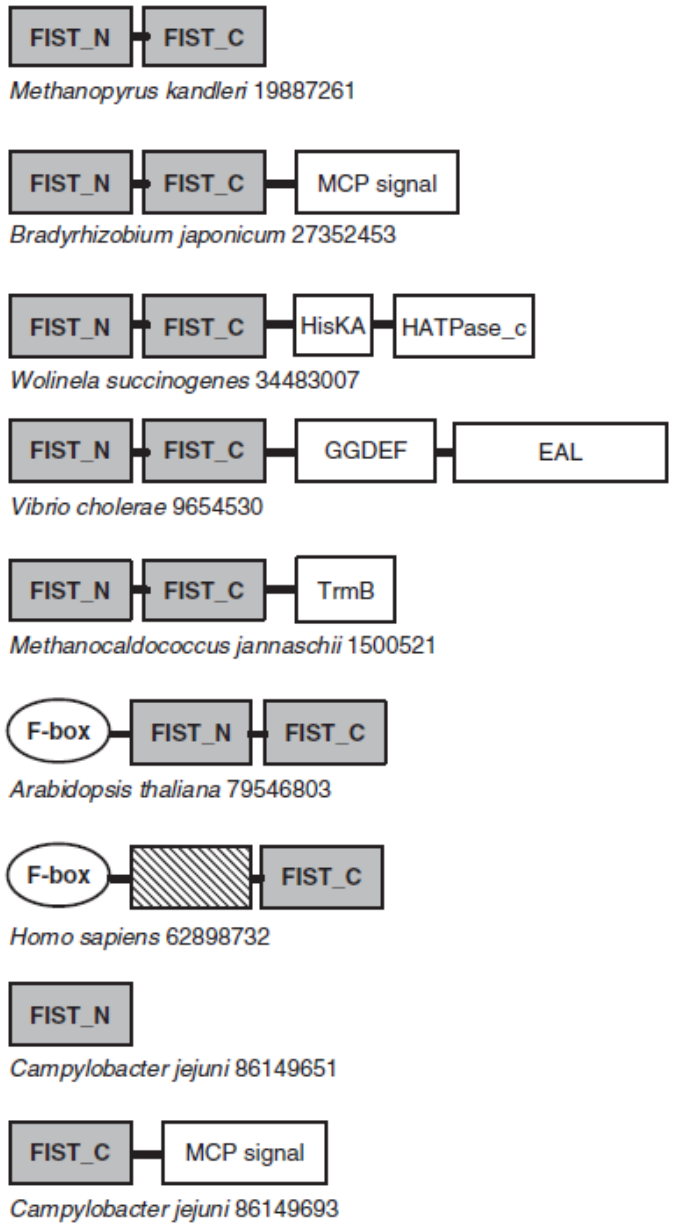

Figure 4: Domain architecture of representative proteins containing the FIST domain.

Species names and GenBank identification numbers are shown. Domain designations (Pfam accession numbers in parentheses): MCPsignal (PF00015), MCP C-terminal signaling domain; GGDEF (PF00990), di-guanylate cyclase; EAL (PF00563), di-guanylate esterase; HisKA (PF00512), histidine kinase dimerization domain; HATPase_c (PF02518), histidine kinase domain ATP-binding domain; TrmB (PF01978),

helix-turn-helix DNA-binding domain; F-box (PF00646), protein-protein interaction domain and PAS (PF00989), the PAS sensory domain.

Biological Function

FIST represents a new sensory (input) domain in signal transduction. As many other sensory domains, it is found either as a single-domain protein or exclusively in a combination with other domains that transmit signals from the sensory domain down the regulatory pathway, namely transducers (MCPsignal, HATPase_c domains) and output domains (GGDEF, EAL, TrmB). The FIST domain was found in four major classes of prokaryotic signal transduction: methyl-accepting chemotaxis proteins, sensor histidine kinases, di-guanylate cyclases and diesterases, and transcriptional regulators.

Further evidence suggesting FIST involvement in signal transduction comes from the analysis of the genome context [216] of FIST-encoding genes. Using the MIST database [210], we have determined that in more than 30% cases, adjacent (immediately upstream or downstream) to the FIST-encoding gene there is a gene coding for a known signal transduction protein. Overall, signal transduction genes in 198 prokaryotic genomes that contain FIST domains comprise <7% of the total number of genes (http://genomics.ornl.gov/mist). Thus there is a significant enrichment in signal transduction genes next to FIST-encoding genes. As a sensory domain, FIST is predicted to bind small molecule ligands. In eukaryotes, FIST is found in F-box proteins that are involved in ubiquitin mediated degradation of regulatory proteins, a frequent means of controlling progression through signaling pathways [217]. Interestingly, genes encoding FIST-containing proteins are also often found next to genes coding for enzymes involved in amino acid metabolism (peptidases, aspartamine synthase) and amino acid transporters. In plants, F-box proteins are known to bind a plant hormone auxin, which is a derivative of tryptophan [218]. This further suggests that small molecule ligands detected by FIST might be amino acids and their derivatives. Detecting amino acid concentration both inside and outside the cell is important for many physiological processes, and it is likely that FIST-containing sensors play a

significant role in converting these signals into changes in transcription, metabolism, development and behavior.

FIST domains are found in 16 phyla (File 2) representing all three kingdoms of life. This broad phyletic distribution suggests the ancient origin of this domain and further underscores its importance as a ubiquitous sensor module in diverse signal transduction pathways.

CHAPTER II

EVOLUTIONARY ANALYSIS OF THE CHEMOTAXIS SYSTEM

This chapter was taken from a manuscript in preparation:

Borziak K, Fleetwood A, and Zhulin IB. **Pathogenicity as the driving force of evolution of chemotaxis in *Escherichia***. Manuscript in preparation.

Conceived and designed the experiments: IBZ. Performed the experiments: KB AF. Analyzed the data: KB AF IBZ. Wrote the paper: KB AF IBZ.

Abstract

Chemotaxis allows bacteria to more efficiently colonize their environment and find optimal growth conditions, and is consequently under strong evolutionary pressures. The chemotaxis system of *Escherichia coli* is the best-studied model system and *E. coli* is the most abundantly sequenced organism to date. The *Escherichia* clade encompasses a variety of commensal and pathogenic strains, which inhabit different habitats across a wide range of hosts. Chemotaxis has been implicated in allowing *E. coli* to colonize its host. However, the evolutionary history of chemotaxis in *E. coli* has not been examined from the genomic perspective. Here we show that the core components of chemotaxis have remained intact in the majority of sequenced strains, but accessory MCPs have undergone ancestral loss and recent gain events. The historically non-motile *Shigella* were seen to have a greater number of mutations in their chemotaxis genes, however *Shigella* strains with intact chemotaxis systems were also seen. The previously noted losses of *trg* and *tap* MCPs in UPEC were found to be ancestral events that occurred prior to the divergence of the B2 phylogroup from other *E. coli*. Since the B2 phylogroup contains the majority of extra-intestinal pathogenic *E. coli*, the losses suggest that the resultant decrease in competitiveness in the intestinal tract prompt colonization of other habitats such as the urinary tract. MCP acquisitions were found to be recent, plasmid-born events, indicating their minor relative evolutionary importance. Our results demonstrate the possible changes in the highly conserved chemotaxis system on a small evolutionary time scale, mainly through the force of gene loss. Changes in the chemotaxis system of *Escherichia* play an important

role in the evolution of pathogenicity of *Escherichia*. Due to the reticulate evolution of *Escherichia* pathogenicity, further understanding of how chemotaxis affects virulence is important for its full evolutionary understanding.

Introduction

Pathogenic bacteria present ongoing challenges to both human and animal health, however, the processes of virulence evolution remain incompletely understood, even in the model bacteria *Escherichia coli. E. coli* is ubiquitous and is a common environmental bacteria, but most strains are commensal colonizers of the intestines of mammals and birds. The *Escherichia* clade includes pathogens of global significance responsible for epidemic dysentery, hemolytic uremic syndrome, gastroenteritis, neonatal meningitis, urinary tract infections, and other diseases. Chemotaxis and motility play an important role in the colonization of both commensal and pathogenic *E. coli,* as well as the pathogenesis of the latter [219-221]. However, the evolutionary trends affecting the chemotaxis system and their effects on pathogenicity have not been studied. Addressing such questions requires a global genomic overview of how chemotaxis system has evolved within the *Escherichia* clade.

E. coli typically colonizes the gastrointestinal tract of humans within the first few hours after birth [222, 223]. Usually, *E. coli* and its host coexist with mutual benefit. However, there are several highly adapted *E. coli* clones that have acquired virulence traits that increase their adaptability to new niches and allow them to cause a broad spectrum of disease [224]. Three general clinical syndromes can result: enteric/diarrheal disease, urinary tract infections (UTIs), and sepsis/meningitis. Intestinal pathogens include enteropathogenic *E. coli* (EPEC), enterohemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAEC), enteroinvasive *E. coli* (EIEC), adherent-invasive *E. coli* (AIEC), diffusely adherent *E. coli* (DAEC), and Shiga toxin-producing *E. coli* (STEC). The *E. coli* pathotypes implicated in extraintestinal infections are collectively called ExPEC [225]. UTIs are the most common extraintestinal *E. coli* infections and are caused by uropathogenic *E. coli* (UPEC). An

increasingly common cause of extraintestinal infections is the pathotype responsible for meningitis and subsequent sepsis — meningitis-associated *E. coli* (MNEC). An additional animal pathotype, avian pathogenic *E. coli* (APEC), causes extraintestinal infections of poultry.

Phylogenetic methods have shed light on the processes of genomic evolution of this extraordinarily diverse genus and the origins of pathogenic *E. coli* [226-228]. One must also include the genus *Shigella* when discussing *E. coli*, because *Shigella* is phylogenetically indistinguishable from *E. coli* and retains its name due to historical reasons [229]. The genus *Escherichia* also includes *E. albertii* [230] and *E. fergusonii* [231]. *E. coli* strains were further classified into phylotypes using multilocus enzyme electrophoresis (MLEE) [232] and later multi-locus sequence typing (MLST) [226, 227, 233]. The initial attempt [232], the ECOR collection, subdivided *E. coli* into four groups, designated A, B1, B2 and D, plus a minor group E that has largely been ignored because it clustered inconsistently in subsequent analyses. More recently, group F has been delineated from a subset of group D [226]. Phylogenetic analysis suggests that the B2 and D phylotypes were ancestral to A and B1 [226, 233]. However, other work suggests reticulate evolution over clonality [227]. *Shigella* were found to have arisen independently and repeatedly within several lineages of *E. coli* [227].

The ability to respond and adapt to changing environment is important for bacterial survival. Chemotaxis allows bacteria to migrate towards favorable chemicals (attractants) and away from unfavorable chemicals (repellents). Chemotaxis also plays a key role in the virulence of many pathogens, including *Escherichia coli* [219, 220].

In the typical *E. coli* chemotaxis system, the methyl-accepting chemotaxis proteins (MCPs) act as receptors, which activate the CheA kinase, resulting in an increase in phosphorylation of the CheY response regulator. Phosphorylated CheY in turn binds to the switching proteins at the flagellar motor, causing a switch in the motor rotation and a tumbling response. While repellents promote phosphorylation and increased tumbling, attractants have the opposite, inhibitory, effect, resulting in longer periods of swimming. The chemotaxis system responds to changes in the concentration of effectors, rather than to their absolute levels. Signal termination occurs by

dephosphorylation of CheY by a specific phosphatase, CheZ, to allow continuous gradient sensing.

Specific glutamate residues in the conserved signaling domains of the receptors are subject to methylation by a specific methyltransferase, CheR, and demethylation by a specific methylesterase, CheB. The methylesterase is coupled with a regulatory domain, allowing its activity to be modulated by CheA, increasing methylesterase activity through phosphorylation. Methylation of the glutamate residues tends to increase kinase activation, while demethylation has the opposite effect and acts as a negative feedback mechanism. Methylation provides a robust mechanism that maintains a constant steady-state swimming behavior under a wide range of different environmental conditions.

Five canonical MCPs are found in *E. coli*, which fall into two classes, major and minor receptors [234]. The two major receptors in *E. coli*, Tar and Tsr, are distinguished from the other receptors by their ability to undergo adaption and signal transmission independent of other receptors, and their greater abundance. The three minor receptors, Trg, Tap, and Aer, are dependent, at least in part, on the major receptors to undergo adaptation and signal transmission, and are present at one-tenth the level of the major receptors. Tsr senses serine; Tar senses aspartate and maltose, while nickel and cobalt serve as repellants [235]. Trg senses ribose and galactose/glucose. Tap senses dipeptides and pirimidines [236]. Aer senses FAD as a measure of redox potential. Maltose, ribose, galactose/glucose, and dipeptides are sensed indirectly by the MCPs through interaction with respective periplasmic binding proteins, MalE, RbsB, MglB, and DppA. The majority of chemotaxis proteins are found on two adjacent operons, *mocha* (*motA, motB, cheA, cheW*) and *meche* (*tar, tap, cheR*, *cheB, cheY, cheZ*) [237], while other MCPs and accessory components are distributed across the chromosome. MotA and MotB form the stator of the flagellar motor.

Specifically in UPEC, chemotaxis and motility have been shown to play important roles in dissemination and efficient host colonization [136, 219, 238, 239]. UPEC cause more than 70% of UTIs among healthy individuals, which makes understanding the genomic basis of particular medical importance [240]. The majority of UPEC were

previously seen to have lost their *trg* and *tap* genes, and it was postulated that this was a result of a lack of selective pressure in the urinary tract [142]. We wanted to examine this assertion from an evolutionary perspective. Additionally, we analyzed the evolutionary forces affecting the chemotaxis system in all currently sequenced *E. coli*. Further understanding of the importance of chemotaxis in UPEC and *E. coli* in general will provide important insight into pathogenesis of UTIs and commensal colonization of the host.

In this study we examine the evolutionary histories of chemotaxis system components of all available *Escherichia* genomes. The data presented provide insight into the evolutionary forces affecting the chemotaxis system as well as its implication on the ecology and pathogenicity of *Escherichia*. We show that the chemotaxis system is undergoing loss of accessory receptors in many clones and, in rare cases, the loss of the whole chemotaxis system. Losses of *trg* and *tap* were ancestral events that occurred prior to the divergence of the B2 phylogroup. This contradicts the previous notions of the losses as adaptations of UPEC to the uretic environment, and instead suggests that the loss of receptors prompted B2 *E. coli* to colonize extra-intestinal environments due to decreased competitiveness in the intestinal environment.

Results

We obtained 219 sequenced genomes of publicly available *Escherichia* and *Shigella*. Other than the closely related *E. coli* and *Shigella* genomes, our set included genomes of *E. fergusonii* and *E. albertii*, which served as outgroups. Of those, 55 were complete genomes, and 164 were draft genomes (File 3). First, all *Escherichia* genomes were classified based on pathotype and phylotype (Figure 5, File 3). The phylotype assignment agreed with previously established assertions that the B2 and D phylogroups were ancestral to the A and B1 phylogroups [233, 241]. Additionally, ExPEC strains were found to belong only to B2, D, and F phylogroups, as was previously seen [228, 242, 243].

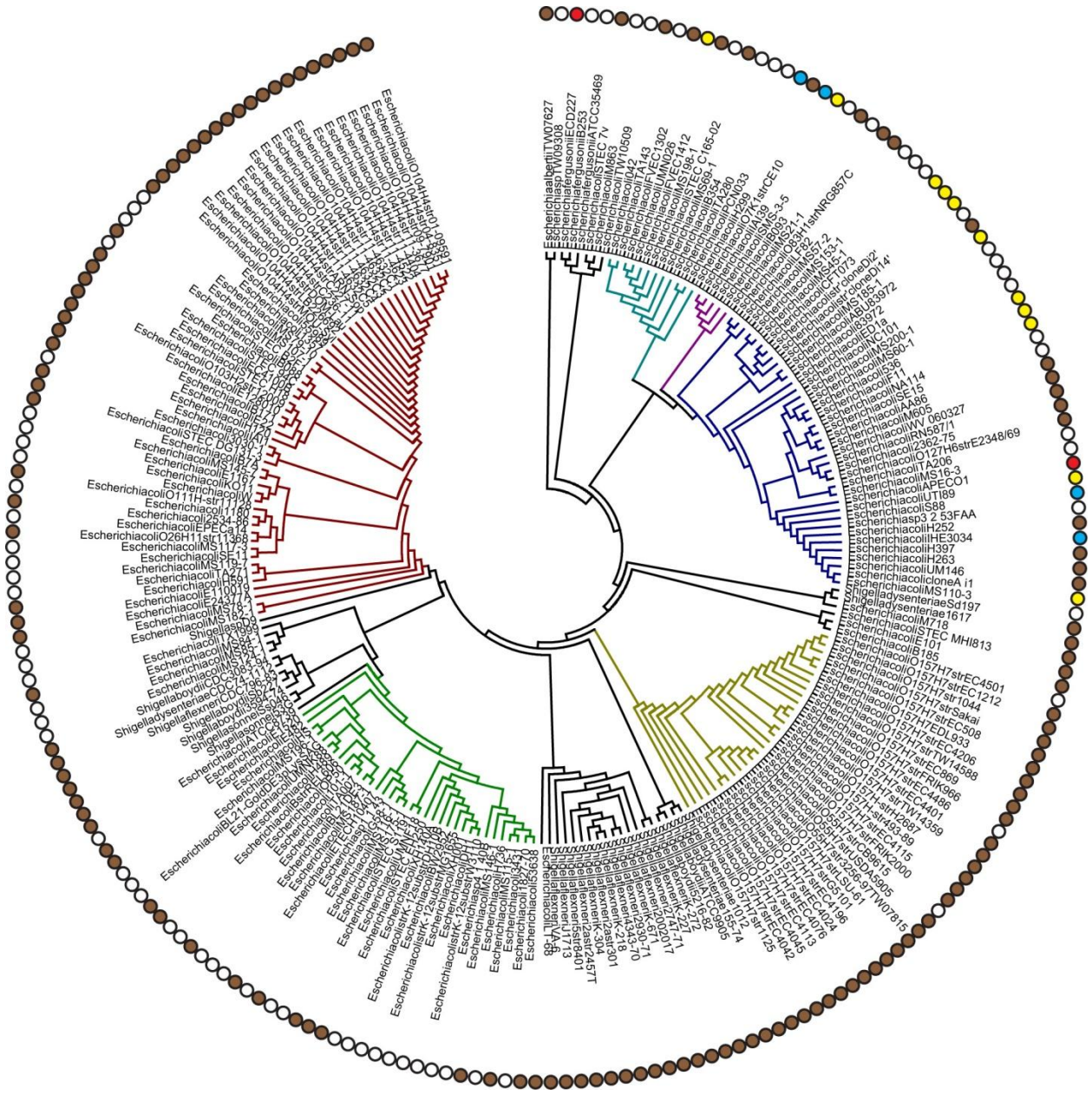

Figure 5: Phylogenetic tree of *Escherichia*.

Phylogenetic tree constructed using *arcA*, *aroE*, *icd*, *mdh*, *mtlD*, *pgi*, and *rpoS* genes [244]. Branches are colored according to phylogroup: (teal) D phylogroup; (purple) F phylogroup; (blue) B2 phylogroup; (yellow) E phylgroup; (green) A phylogroup; (red) B1 phylogroup. Outer circle is colored according to pathotype: (red circle) APEC; (blue circle) MNEC; (yellow circle) UPEC; (brown circle) intestinal pathogen; (white circle) nonpathogenic/commensal.

Our study on chemotaxis of *E. coli* focused on the *mocha* and *meche* operons, as well as orphan MCPs found outside the chemotaxis operons: Tar, Trg, and Aer. Our results show that the genomic context of the chemotaxis operons and orphan MCPs is very well conserved. Due partly to the variation in sequencing and annotation approaches used, 10% (approximate) of genomes had deletions, frameshifts, or inaccurate start codon predictions in the amino acid sequences of their chemotaxis proteins. As expected, all such genomes were in draft status.

Due to conflicting observations on *Shigella* motility [245, 246], we first examined their chemotaxis systems for defects (Figure 6, File 3). 7 of the 28 *Shigella* had significant deletions in the *mocha*/*meche* operons, compared to only 1 *E. coli* strain and the *E. albertii* strain. These deletions were present in both closed and draft genomes, thus this finding is unlikely to be due to sequencing error. 10 *Shigella* strains had no obvious defects in their chemotaxis operons. It was previously found that many *Shigella*, including those with seemingly healthy chemotaxis operons have significant mutations in their flagellar genes [247], suggesting that the chemotaxis system is involved in other vital processes. *E. albertii* was also missing its chemotaxis system. Since it is also pathogenic, the notion that chemotaxis and motility are not strict requirements for pathogenicity is further reinforced. The *Shigella* receptors were overall more prone to mutation than the other chemotaxis proteins: 10 frameshifts or deletions in *tar*, 13 in *tsr*, 7 in *trg*, 10 in *tap*, and 12 in *aer*. Of the 10 *Shigella* strains with no defects in the chemotaxis operons, 7 strains have deletions or frameshifts in their *aer* genes. Only *Shigella* sp. D9 has a complete set of canonical chemotaxis proteins. Its position on the phylotype tree suggests that it is likely a misnamed *E. coli*.

Also of note were several laboratory strains that showed marked chemotaxis operon deficiencies that suggest severe motility defects or a total loss of motility. The widely used high transformation efficiency strain *Escherichia coli* K12 DH10B, for instance, has a frameshift mutation in the *cheA* gene and has lost the major MCP *tsr.* Furthermore, *Escherichia coli* OP50, commonly used as food for *C. elegans*, possesses many chemotaxis proteins that contain not one but two stop codon insertions, indicating a drastic and understandable loss in selective pressure. In stark contrast, “wild type” strains K12 W3110 and MG1655 show no observable deficits. Within the relatively

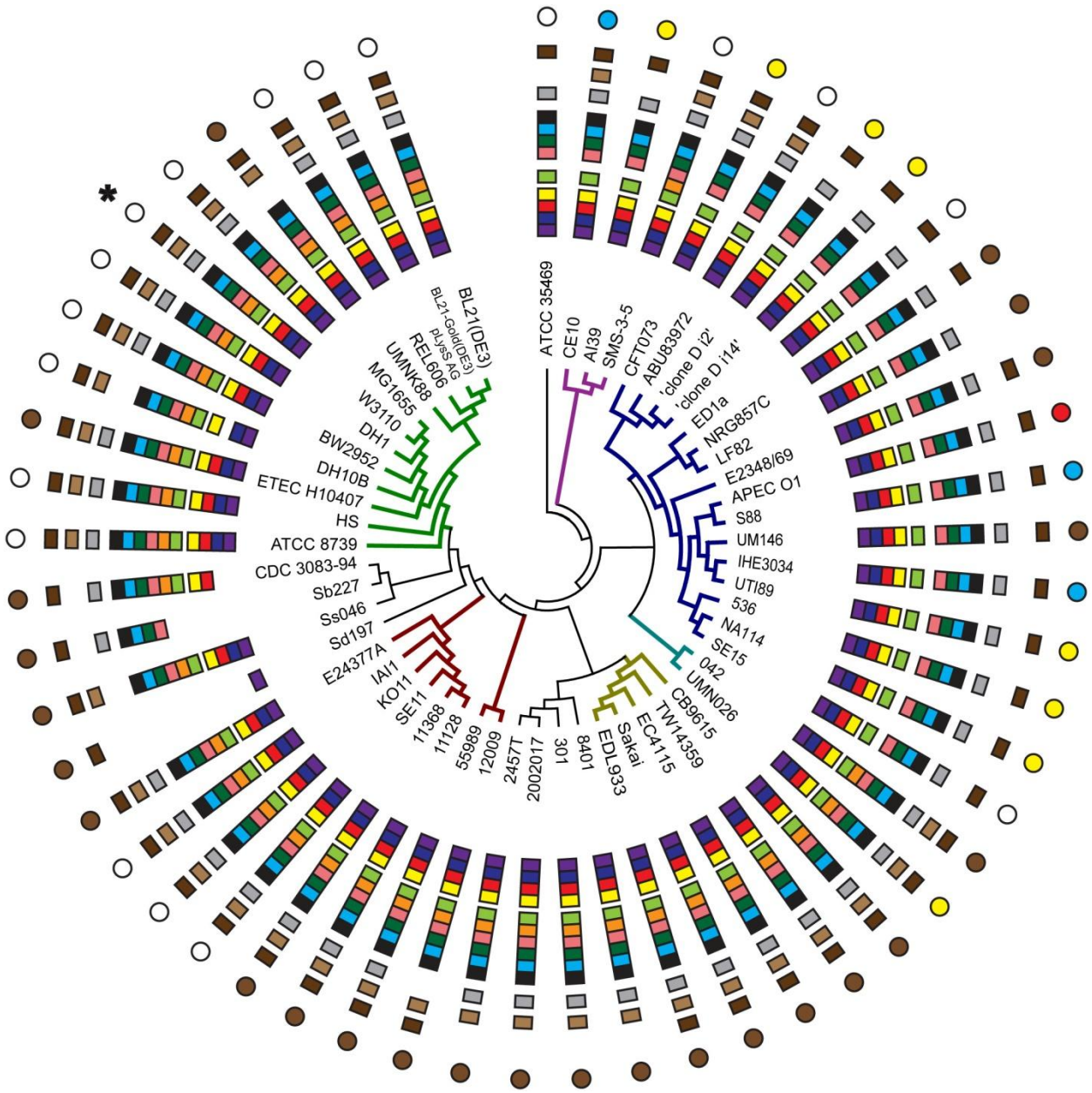

Figure 6: Presence of chemotaxis genes in complete *Escherichia* genomes.

Branches are colored according to phylogroup: (teal) D phylogroup; (purple) F phylogroup; (blue) B2 phylogroup; (yellow) E phylgroup; (green) A phylogroup; (red) B1 phylogroup. The rectangles, from inner to outer represent intact chemotaxis proteins: (purple) MotA, (dark blue) MotB, (red) CheA, (yellow) CheW, (lime) Tsr, (orange) Tap, (pink) CheR, (dark green) CheB, (light blue) CheY, (black) CheZ, (maroon) Tar, (tan) Trg, (brown) Aer. Duplicate MCPs (gray) appear adjacent to the ancestral copies. Outer circle is colored according to pathotype: (red circle) APEC; (blue circle) MNEC; (yellow

circle) UPEC; (brown circle) intestinal pathogen; (white circle) nonpathogenic/commensal. * marks *E. coli* W3110, the most commonly strain used for the study of chemotaxis.

short (evolutionarily-speaking) period of modern biological investigation, one can potentially observe profound effects on the chemotaxis system.

Although all five MCPs were found in a majority of *E. coli*, several clades of *Escherichia* were found to have lost various MCPs (Figure 6). Most strikingly, all members of the B2 phylogroup and 3 of the 5 members of the F phylogroup underwent a deletion in the *tap* gene. Due to the identical nature of the deletions (Figure 7, Figure 8a), the parsimonious explanation suggests that the event occurred ancestral to the B2 clade speciation. The majority (33 of 38) of B2 phylogroup members have also undergone a deletion within their *trg* gene. Only *E. coli* WV_060327 has a complete *trg* gene, while the other 4 B2 strains all possess frameshift mutations within that gene. Similarly to the deletion of *tap*, the symmetrical nature of the *trg* deletion (Figure 7, Figure 8b) suggests that the loss was an ancestral event. The 5 *trg* genes without the deletion from B2 phylogroup strains share higher similarity with the *trg* genes of A and B1 phylogroup strains rather than the more closely related D, E, and F phylogroup strains, suggesting recombination. The fact that the loss of MCPs is not due to the lack of nutrients sensed by *trg* and *tap* is further evidenced by the presence and intact nature of the periplasmic binding proteins RbsB, MglB, and DppA in all but 1 B2 *E. coli* and in all ExPEC. Additionally, all strains of *E. fergusoniii* underwent deletion of *tap* and *trg*. Of those *Escherichia* classified as extraintestinal pathogens, only *E. coli* PCN033 and *E. coli* H299, both belonging to clade D, have intact *tap* and *trg* genes. Loss of MCPs appears to be a common trend among *Escherichia* with only 57% of analyzed genomes containing all 5 intact, canonical MCPs.

The overall trend of MCP loss is countered by sporadic events of MCP acquisition. In 9 strains, 1 *E. fergusonii* and 8 *E. coli*, a horizontally transferred MCP was found (Figure 6, File 3). All acquired MCPs were plasmid-borne. *E. fergusonii*

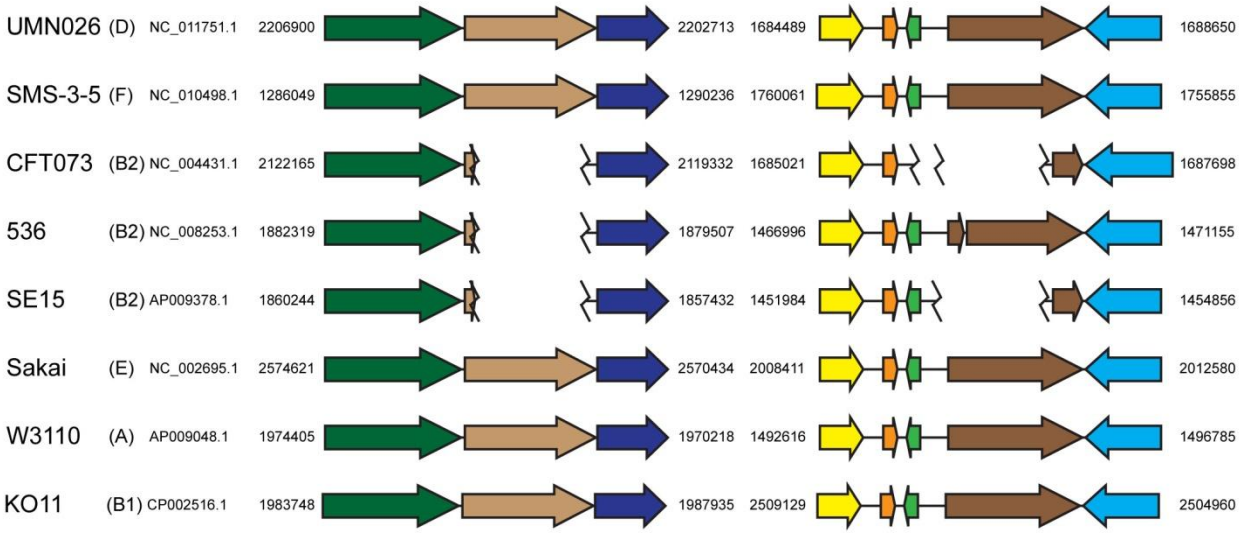

Figure 7: Representative gene neighborhoods of *E. coli tap* and *trg* genes.

The gaps show symmetrical deletions that likely occurred in the ancestor of the B2 phylogroup. *E. coli* 536 shows a recombination event to partially restore the *trg* sequence. (Dark green) *tsr*, (tan) *tap*, (dark blue) *cheR*, (yellow) *cybB*, (orange) hypothetical protein, (light green) *mokB*, (brown) *trg*, (light blue) *ydcI*.

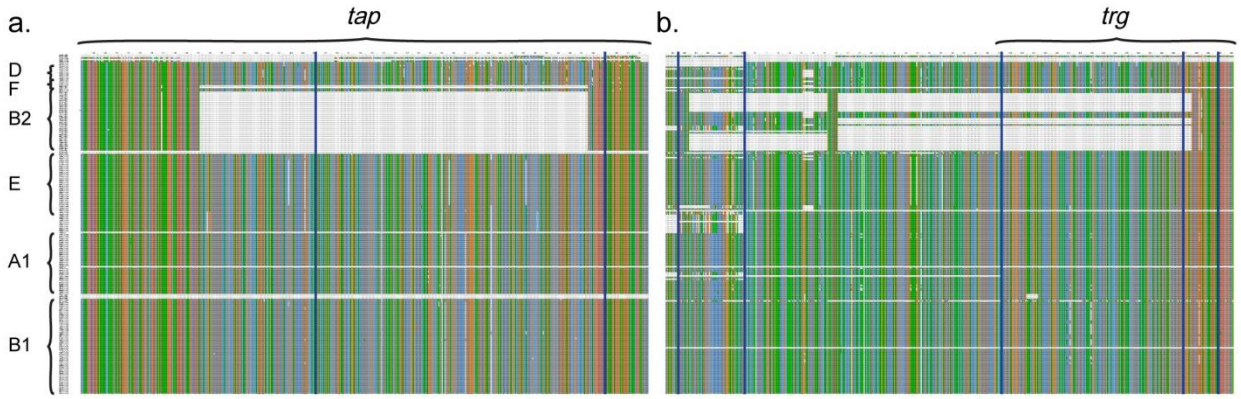

Figure 8: Multiple sequence alignments of *E. coli tap* and *trg* gene neighborhoods.

a) *tap* multiple sequence alignment. b) *trg* multiple sequence alignment. The gaps show symmetrical deletions that likely occurred in the ancestor of the B2 phylogroup. Sequences appear in the order found on the phylogenetic tree (Figure 5,

File 3). Multiple sequence alignment is colored according to the Clustalx color scheme. Blue lines represent alignment truncations for ease of viewing.

ECD227 acquired a *tar*-like gene from *Salmonella enterica*. *E. coli* O157:H7 str. EC4024 acquired a *trg*-like gene from an *Enterobacteria* strain. The MCP is found neighboring a sucrose metabolism gene cluster, suggesting a role as a sucrose sensor. Additionally 7 strains were found to possess an *aer*-like MCP acquired from *Aeromonas punctata*, which is also known to cause gastroenteritis. 6 of the acquired MCPs are identical, suggesting recent plasmid acquisition and dispersal. Additionally, 4 of the 7 strains belong to the A phlotype while the remaining 3 are not yet typed. Interestingly, the 6 strains with identical *aer*-like MCPs were isolated from different sources.

Discussion

Chemotaxis behavior remains important in *Escherichia*. Since the majority of the genomes (84%) retain an intact chemotaxis system, the ability to undergo chemotaxis confers an evolutionary advantage to the majority of *Escherichia* strains. The importance of flagellar motility and by extension chemotaxis to efficiently colonize both the intestines and the urinary tract further supports this assertion. However, while the potential for chemotaxis is retained, chemotaxis sensory proteins are largely dispensable. Of the 184 genomes with intact chemotaxis proteins, only 113 (61%) had all 5 receptors intact. Although loss of receptors was noted in all *Escherichia* clades, the majority of strains with MCP loss were from the B2 clade, which underwent ancestral deletions within their *trg* and *tap* receptors. Since the deletions occurred in the same place in all B2 strains, it is highly unlikely that the deletions were a recent event. The major receptors *tar* and *tsr* were the most consistently maintained. This trend of minor receptor loss demonstrates that the environmental cues that *Escherichia* needs to sense are fewer than the canonical set of 5 receptors allows. It is also known that the

major receptors allow for other taxis behavior in *E. coli*, including osmotaxis, thermotaxis, and pH taxis [144, 248, 249]. These additional properties likely make *tar* and *tsr* more indispensable than *tap* and *trg*. The aerotaxis receptor *aer*, although a minor receptor, is also highly retained in *E. coli*. Interestingly, *aer* was primarily lost in only *Shigella*. Due to *Shigella*'s nature as an exclusive intracellular pathogen, it is likely that this constant intracellular environment has relaxed evolutionary pressure on motility and chemotaxis. This is evidenced by the loss of flagella in *Shigella*. However, 30% of *Shigella* strains retain intact *mocha* and *meche* operons, suggesting that the ability to sense its environment is still adaptive to *Shigella* even if the function of this signal transduction system has been changed from regulating motility.

While primarily gene loss shapes the chemotaxis system of *Escherichia*, gain of new receptors has been seen in 9 strains. All new receptors were a result of horizontal gene transfer, rather than duplication. New sensors would theoretically allow the strains that possessed them to take advantage of new niches. 7 of the acquired receptors were intracellular *aer*-type PAS domain containing MCPs. It is possible that these new receptors are also involved in sensing the cell's metabolism status or oxidative stress that could be the result of host immune cells. None of the acquired MCPs were found on the chromosomes, however. Instead, they were all plasmid borne. The lack of non canonical genomic MCPs further reinforces the point of the chemotaxis system undergoing reduction in order to increase efficiency. Heavy reliance upon a host organism is the most likely reason for reduced pressure on the minor receptors, and this trend will most likely continue, resulting in further diversification and the potential development of novel pathotypes.

Of most interest was investigating the origin of *trg* and *tap* loss in UPEC as well as the importance of chemotaxis and motility in UPEC pathogenicity [142, 238]. This loss does not appear to be a result of relaxed environmental pressure to maintain sensors to sugars and dipetides. Since the deletions that result in the loss of these two receptors seems to be identical in all B2 phylogroup strains, to which most UPEC belong, the most parsimonious explanation is a deletion of *trg* and *tap* in the ancestral B2 strain. The presence of intact *tap* and *trg* in urinary pathogen *E. coli* UMN026 shows that these receptors can be advantageous or neutral to UPEC fitness. Further, the

presence of ribose, galactose/glucose, and dipeptide PBPs which mediate the sensing of those compounds through *trg* and *tap* indicates that UPEC still metabolize those compounds at some stage in their life cycle, even though they are absent in the urine of healthy individuals with normally functioning kidneys. Since the colon appears to be the major source of UPEC, their ability to metabolize sugars and peptides present in the colon is still advantageous, even if they are unable to chemotax toward them. The same pattern was found for MNEC. Again, the majority of MNEC were from the B2 phylogroup. The only exceptions were two strains from the D phylogroup, uropathogenic *E. coli* UMN026 and the porcine meningital *E. coli* PCN033, which both retain intact *trg* and *tap*. However, since the majority of ExPEC strains are from the B2 phylogroup, it is possible that the ancestral loss of *trg* and *tap* predisposed those strains to adapt to niches outside the colon. Loss of *trg* and *tap* could cause a decrease in colon colonization efficiency, promoting those strains to seek another nearby and fairly uninhabited niche to occupy.

Materials and Methods

Bioinformatics software and computer programming environment

The following software packages were used in this study: HMMER v3.0 [104], Jalview [250], MAFFT v6.847b [251], MEGA v4.0 [252], PhyML v3.0 [253], and BLAST+ v2.2.4+ [254]. All multiple sequence alignments were built in MAFFT with its l-INS-i algorithm. All maximum likelihood phylogenetic trees were built in PhyML with standard parameters and subtree pruning and regrafting topology search. All computational analyses were performed in a local computing environment (including high-performance computing), and custom scripts for data analysis were written in PHP.

Data sources

Genomes, proteomes, and genome annotations of all distinct *Escherichia* and *Shigella* strains available in the NCBI *nr* database as of 12th January, 2012 were

collected (219 genomes). Nucleotide and protein sequences were compiled into local BLAST databases. The pathotype information was retrieved from primary literature, were available, or from public information repositories, including IMG, GOLD, and PATRIC. The genome of *E. coli* W3110 was used as the source for all protein and nucleotide for initial BLAST searches due to its status as the model strain for chemotaxis studies.

Construction of a phylotype tree for *Escherichia*

Escherichia phylogenetic tree was constructed using the *arcA*, *aroE*, *icd*, *mdh*, *mtlD*, *pgi*, and *rpoS* genes [244]. The nucleotide sequences for the above genes were retrieved from the genome of *E. coli* W3110 and used as BLAST queries against the genome set. The nucleotide sequence sets for each gene were aligned individually in MAFFT. The alignments were concatenated, and the resulting alignment was used to build a maximum likelihood tree in PhyML. The genomes were assigned to phylogroups based on the presence of previously assigned genomes in their clade [244] (Figure 5).

Identification of chemotaxis and accessory proteins in genomic data sets

Chemotaxis and accessory genes and proteins were retrieved from the genome of *E. coli* W3110 and used as BLAST queries against the genome set. Exhaustive BLAST was performed with retrieved chemotaxis genes to search for any missing and partial genes. Homologs were differentiated based on E-value cutoff, which varied for each gene. Gene neighborhoods were extracted from NCBI genome feature files with custom PHP scripts.

Multiple sequence alignment and phylogenetic analyses

The nucleotide and protein chemotaxis sequence sets (MotA, MotB, CheA, CheW, MCPII, MCPIV, CheR, CheB, CheY, CheZ, MCPI, MCPIII, and MCPV) were individually aligned by MAFFT. The alignments of the core chemotaxis operons, *mocha*

and *meche* (MotA, MotB, CheA, CheW, MCPII, MCPIV, CheR, CheB, CheY, CheZ), were concatenated and used to build a maximum likelihood tree in PhyML.

CHAPTER III

GENOME ANALYSIS OF AZOSPIRILLUM

A version of this chapter was originally published by Florence Wisniewski-Dyé, Kirill Borziak, *et al* in *PLoS Genetics*:

Wisniewski-Dyé F, Borziak K, Khalsa-Moyers G, Alexandre G, Sukharnikov LO, et al. (2011) ***Azospirillum* Genomes Reveal Transition of Bacteria from Aquatic to Terrestrial Environments.** *PLoS Genet* 7(12): e1002430.

Conceived and designed the experiments: FW-D PM AHP IBZ. Performed the experiments: KB GK-M GA GBH FW-D CP-C JSR VB AC ZR SM LOS KW MB VG PS GC GK. Analyzed the data: FW-D KB PM WHM AHP PN CE YD IK IBZ. Contributed reagents/materials/analysis tools: GA GBH VB ZR KW. Wrote the paper: FW-D KB PM AHP IBZ.

Abstract

Fossil records indicate that life appeared in marine environments ~3.5 billion years ago (Gyr) and transitioned to terrestrial ecosystems nearly 2.5 Gyr. Sequence analysis suggests that “hydrobacteria” and “terrabacteria” might have diverged as early as 3 Gyr. Bacteria of the genus*Azospirillum* are associated with roots of terrestrial plants; however, virtually all their close relatives are aquatic. We obtained genome sequences of two *Azospirillum* species and analyzed their gene origins. While most *Azospirillum* house-keeping genes have orthologs in its close aquatic relatives, this lineage has obtained nearly half of its genome from terrestrial organisms. The majority of genes encoding functions critical for association with plants are among horizontally transferred genes. Our results show that transition of some aquatic bacteria to terrestrial habitats occurred much later than the suggested initial divergence of hydro- and terrabacterial clades. The birth of the genus *Azospirillum* approximately coincided with the emergence of vascular plants on land.

Author Summary

Genome sequencing and analysis of plant-associated beneficial soil bacteria *Azospirillum* spp. reveals that these organisms transitioned from aquatic to terrestrial environments significantly later than the suggested major Precambrian divergence of aquatic and terrestrial bacteria. Separation of *Azospirillum* from their close aquatic relatives coincided with the emergence of vascular plants on land. Nearly half of the *Azospirillum* genome has been acquired horizontally, from distantly related terrestrial bacteria. The majority of horizontally acquired genes encode functions that are critical for adaptation to the rhizosphere and interaction with host plants.

Introduction

Fossil records indicate that life appeared in marine environments ~3.5–3.8 billion years ago (Gyr) [255] and transitioned to terrestrial ecosystems ~2.6 Gyr [256]. The lack of fossil records for bacteria makes it difficult to assess the timing of their transition to terrestrial environments; however sequence analysis suggests that a large clade of prokaryotic phyla (termed “terrabacteria”) might have evolved on land as early as 3 Gyr, with some lineages later reinvading marine habitats [257]. For example, cyanobacteria belong to the terrabacterial clade, but one of its well-studied representatives, *Prochlorococcus*, is the dominant primary producer in the oceans [258].

Bacteria of the genus *Azospirillum* are found primarily in terrestrial habitats, where they colonize roots of important cereals and other grasses and promote plant growth by several mechanisms including nitrogen fixation and phytohormone secretion [259, 260]. *Azospirillum* belong to proteobacteria, one of the largest groups of “hydrobacteria”, a clade of prokaryotes that originated in marine environments [257]. Nearly all known representatives of its family *Rhodospirillaceae* are found in aquatic habitats (Figure 9 and Table 1) suggesting that *Azospirillum* represents a lineage which might have transitioned to terrestrial environments much later than the Precambrian split of “hydrobacteria” and “terrabacteria”. To obtain insight into how bacteria transitioned from marine to terrestrial environments, we sequenced two well studied

species, *A. brasilense* and *A. lipoferum*, and a third genome of an undefined *Azospirillum* species became available while we were carrying out this work [261].

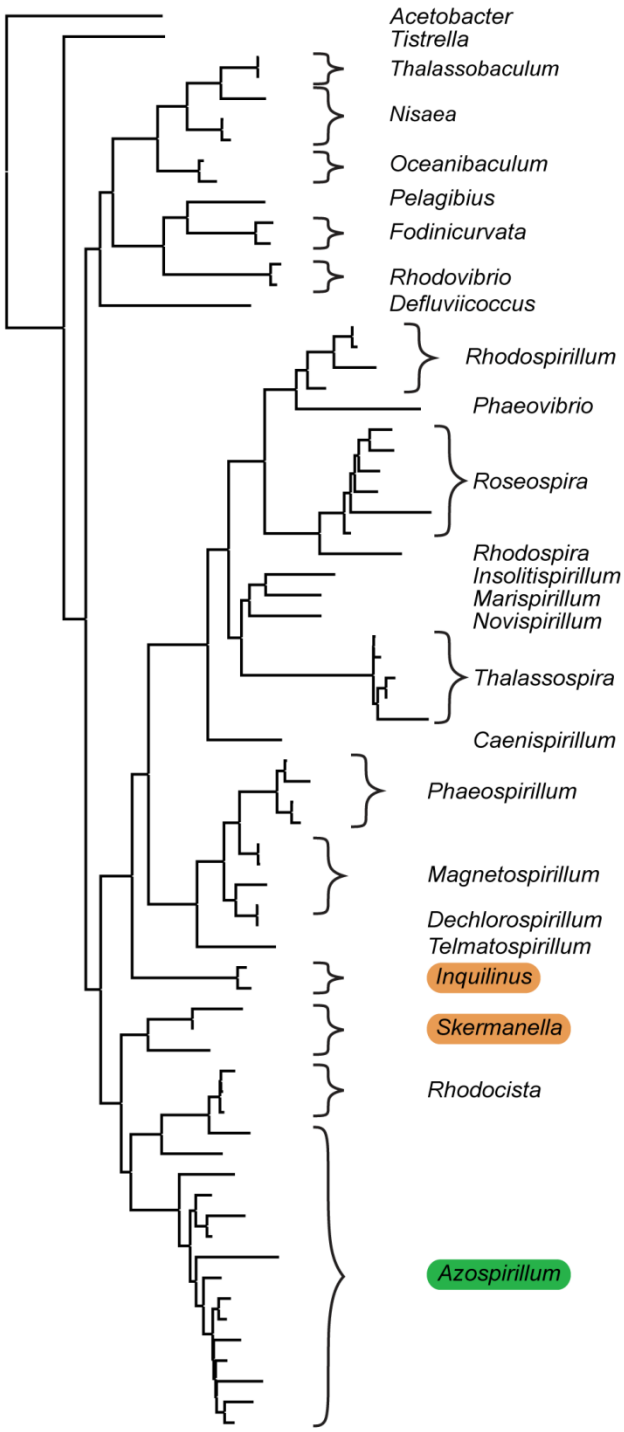

Figure 9: Habitats of *Azospirillum* and its closest aquabacterial relatives.

A maximum-likelihood tree built from 16S rRNA sequences from members of *Rhodospirillaceae*. *Acetobacter acetii*, a member of the same order *Rhodospirillales*, but a different family, *Acetobacteriaceae*, is shown as an outgroup. Aquatic inhabitants are not highlighted; terrestrial are highlighted in brown and plant-associated *Azospirillum* is highlighted in green. See Table 1 for details.

Table 1: Typical habitats of *Rhodospirillaceae*.

Species †	Habitat	Reference
Azospirillum amazonense	roots of maize, sorghum, rice and wheat plants, as well as forage grasses grown around Brazil	[262]
Azospirillum brasilense Sp245	colonizing several plants including cereals, forage grasses, vegetables, legumes, and banana plants	[263]
Azospirillum canadense	corn rhizosphere	[264]
Azospirillum doebereinerae	root of Miscanthus sinensis cv. “Giganteus” and Miscanthus sacchariflorus and also in the rhizosphere soil of these plants grown in Freising, Germany	[265]
Azospirillum halopraeferens	root surface of Kallar grass (Leptochloa fusca) grown in saline-sodic soils in Punjab, Pakistan	[266]
Azospirillum irakense	rhizosphere soil and roots of rice plants grown in the region of Diwaniyah in Iraq	[267]
Azospirillum largimobile	fresh lake water in Australia	[268, 269]
Azospirillum lipoferum 4B	rice field of Camargue (South of France)	[270]
Azospirillum melinis	isolated from molasses grass (Melinis minutiflora Beauv.)	[271]
Azospirillum oryzae	roots of the rice plant Oryza sativa	[272]
Azospirillum palatum	forest soil in Zhejiang province, China	[273]
Azospirillum picis	discarded road tar	[274]
Azospirillum rugosum	oil-contaminated soil sample	[275]
Azospirillum sp. B510	endophytic bacterium isolated from stems of rice plants	[276]
Azospirillum zeae	corn rhizosphere	[277]
Caenispirillum bisanense	Sludge from the wastewater treatment plant	[278]
Dechlorospirillum sp.	Sewage treatment plant	[279]
Defluviicoccus vanus	Wastewater	[280]
Fodinicurvata fenggangensis	salt mine in Yunnan, south-west China	[281]
Fodinicurvata sediminis	salt mine in Yunnan, south-west China	[281]
Inquilinus ginsengisoli	Soil	[282]
Inquilinus limosus	Human respiratory tract	[283]
Insolitispirillum peregrinum	Primary oxidation pond	[284]
Magentospirillum bellicus	bioelectrical reactor (BER) inoculated from creek water	[285]
Magnetospirillum gryphiswaldense MSR-1	freshwater sediment	[286]
Magnetospirillum magneticum AMB-1	Pond water in Tokyo Japan	[287]
Magnetospirillum magnetotacticum MS-1	Microaerobic zones from freshwater sediments	[288]
Marispirillum indicum	deep sea	[289]
Nisaea denitrificans	mediterranean sea	[290]

Table 1. Continued

Species †	Habitat	Reference
Nisaea nitritireducens	mediterranean sea	[290]
Nisaea sp. BAL199	3m depth of Baltic proper	[291]
Novispirillum itersonii	Pond water	[284]
Oceanibaculum indicum	deep sea Indian Ocean	[292]
Oceanibaculum pacificum	hydrothermal sediment of the south-west Pacific ocean	[293]
Pelagibius litoralis	coastal seawater Korea	[294]
Phaeospirillum chandramohanii	freshwater habitat	[295]
Phaeospirillum cystidoformans	freshwater the whole genus	[296]
Phaeospirillum fulvum	stagnant and anoxic freshwater habitats that are exposed to the light	[296]
Phaeospirillum molischianum	stagnant and anoxic freshwater habitats that are exposed to the light	[296]
Phaeovibrio sulfidiphilus	brackish water	[297]
Rhodocista pekingensis	Wastewater	[298]
Rhodocista xerospirillum	Lake water	[299]
Rhodospira trueperi	salt marsh microbial mat	[300]
Rhodospirillum centenum SW	hot spring (hot spring mud) Wyoming, Fresh water	[301]
Rhodospirillum photometricum	Freshwater pond	[302]
Rhodospirillum rubrum ATCC 11170	aquatic environments such as lakes, streams, and standing water	[303]
Rhodospirillum sulfurexigens	freshwater reservoir	[304]
Rhodovibrio salinarum	halophylic, sea water	[296]
Rhodovibrio sodomensis	water/sediment of the Dead Sea	[296]
Roseospira goensis	seawater	[305]
Roseospira marina	sediments, saline springs, microbial mats	[306]
Roseospira mediosalina	sediments, saline springs, microbial mats	[306]
Roseospira navarrensis	sediments, saline springs, microbial mats	[306]
Roseospira thiosulfatophila	microbial mats in French Polynesia	[306]
Roseospira visakhapatnamensis	seawater	[305]
Skermanella aerolata	air	[307]
Skermanella parooensis	water from the Paroo Channel in southwest Queensland	[308]
Skermanella xinjiangensis	desert soil	[309]
Telmatospirillum siberiense	groundwater (mesotrophic fen)	[310]
Thalassobaculum litoreum	coastal seawater	[311]
Thalassobaculum salexigens	mediterranean sea	[312]
Thalassospira lucentensis	mediterranean sea	[313, 314]
Thalassospira profundimaris	deep sea	[315]
Thalassospira tepidiphila	petroleum-contaminated seawater during a bioremediation experiment	[316]
Thalassospira xiamenensis	surface water of a waste-oil pool	[315]
Thalassospira xianhensis	oil-degrading marine bacterium from oil-poluted soil	[317]
Tistrella mobilis	Wastewater, deep sea	[314, 318]

† All currently described members of the family *Rhodospirillaceae* and the habitat of their initial isolation as of January 2011. Species names in bold refer to sequenced strains, both complete and draft genomes.

Results/Discussion

In contrast to the genomes of their closest relatives (other *Rhodospirillaceae*), the three *Azospirillum* genomes are larger and are comprised of not one, but seven replicons each (File 4 and Table 3). Multiple replicons have been previously suggested for various *Azospirillum* strains [196]. The largest replicon in each genome has all characteristics of a bacterial chromosome, whereas the smallest is a plasmid. Five replicons in the genomes of *A. lipoferum* and *Azospirillum* Sp. 510 can be defined as "chromids" (intermediates between chromosomes and plasmids [109]), whereas in *A. brasilense* only three replicons are "chromids" (File 5 and Table 3). While multiple replicons, and chromids specifically, are not unusual in proteobacteria [109, 319], *Azospirillum lipoferum* has the largest number of chromids among all prokaryotes sequenced to date [109] indicating a potential for genome plasticity.

Comparisons among the three genomes reveal further evidence of extraordinary genome plasticity in *Azospirillum*, a feature that has also been suggested by some experimental data [320]. We found very little synteny between replicons of *Azospirillum* species. The genetic relatedness among *Azospirillum* strains is comparable to that of rhizobia, other multi-replicon alpha-proteobacteria (Table 4). Surprisingly, we found substantially more genomic rearrangement within *Azospirillum* genomes than within rhizobial genomes (Figure 10) that are suggested to exemplify genome plasticity in prokaryotes [319]. This could be a consequence of many repetitive sequences and other recombination hotspots (Tables 4 and 5), although the detailed mechanisms underlying such extraordinary genome plasticity remain incompletely understood.

Which genes does *Azospirillum* share with its aquatic relatives, and what is the origin of its additional genes? To answer this question, we developed a robust scheme

Table 2: General features of *Azospirillum* genomes.

	***Azospirillum lipoferum* 4B**	***Azospirillum brasilense* Sp245**
Sequence length	6846400 bp	7530241 bp
GC content (%)	67.67	68.49
Number of Contigs	7	67
Total number of genes	6354	7962
Total number of CDS	6233	7848
Protein coding regions (%)	87.02	85.62
Number of rRNA operons	9	9
Number of tRNA genes	79	81
Genes with functional assignment	4125	4770
Genes with general function prediction only	657	746
Genes of unknown function	1451	2332

Table 3: Identification of chromids in Azospirillum by GC content.

Replicon	**G+C content (%)**	**Chromid GC content cutoff (%) ±0.92% difference with host chromosome**
***A. lipoferum* 4B**		
AZOLI	68.42	69.05 – 67.79
AZOLI_p1	68.52	
AZOLI_p2	68.54	
AZOLI_p3	68.55	
AZOLI_p4	69.13	outside cutoff
AZOLI_p5	68.52	
AZOLI_p6	67.88	
		Chromid GC content cutoff (%) ±0.92% difference with host chromosome
***A. brasilense* Sp245**		
AZOBR	69.31	69.95 – 68.67
AZOBR_p1	69.41	
AZOBR_p2	69.05	
AZOBR_p3	68.70	
AZOBR_p4	69.75	
AZOBR_p5	68.00	outside cutoff
AZOBR_p6	67.98	outside cutoff
		Chromid GC content cutoff (%) ±0.92% difference with host chromosome
***A.* sp. B510**		
AZL	68.39	69.02 – 67.76
AZL_a	68.31	
AZL_b	68.08	
AZL_c	68.02	
AZL_d	68.74	
AZL_e	68.12	
AZL_f	66.28	outside cutoff

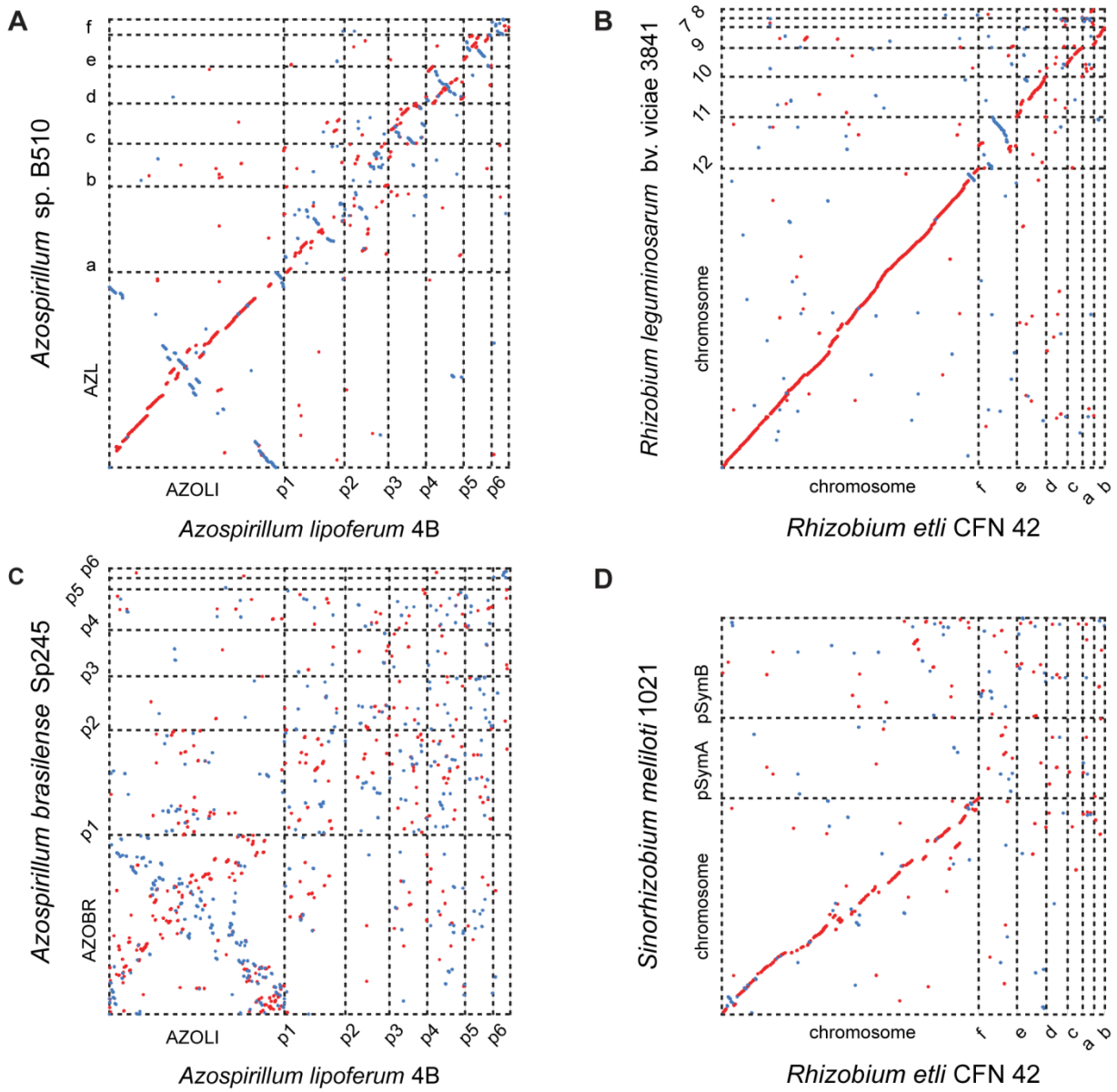

Figure 10: Whole-genome alignments for *Azospirillum* and related multi-replicon rhizobial species.

Relative distances between genomes (calculated from a concatenated ribosomal protein tree):*A. lipoferum* 4B to *Azospirillum* sp.510 – 0.01; *Rhizobium etli to Rhizobium leguminosarum* – 0.02; *A. lipoferum* 4B to *A. brasilense* Sp245 – 0.10; *Rhizobium etli* to *S. meliloti* – 0.11.

Table 4: ANI analysis of *Azospirillum* and rhizobial genomes.

Pair of strains*	Number of MUMs	MUMs (bp)	ANIm (%)	Coverage (%)	Genetic Distance †
4B vs B510	1964	4 782 709	91	71	0.0114
4B vs Sp245	1637	2 012 936	89	33	0.0972
CFN42 vs Rl3841	649	2 796 109	89	43	0.0215
CFN42 vs Sm1021	590	745403	84	11	0.110

* 4B, *A. lipoferum* ; B510, *Azospirillum* sp. ; Sp245, *A. brasilense* ; CFN42, *Rhizobium etli* ; Rl3841, *Rhizobium leguminosarum* biovar *viciae* ; Sm1021, *Sinorhizobium meliloti*.

† Genetic Distance based on concatenated ribosomal protein tree.

Table 5: Recombination hotspots in *Azospirillum* genomes.

Strains[a]	Direct repeats (>80bp)	Palindromic repeats (>80bp)	IS[d] elements (potentially active)	CRISPR[e]
4B	497	412	99 (55)	126
B510	1720	1406	310 (176)	153
Sp245	283	256	ND	12

[a]4B, *A. lipoferum* ; B510, *Azospirillum* sp. ; Sp245, *A. brasilense.*

[b]IS, Insertion sequences. ND, not determined.

[c]CRISPR, Clustered Regularly Interspaced Short Palindromic Repeats.

for detecting ancestral and horizontally transferred (HGT) genes (Figure 11) using bioinformatics tools, then classified most protein coding genes in the *Azospirillum* genomes as ancestral or horizontally transferred with quantified degrees of confidence (Figure 12A and File 6). Remarkably, nearly half of the genes in each *Azospirillum* genome whose origins can be resolved appeared to be horizontally transferred. As a control, we subjected the genomes of other *Rhodospirillaceae* to the same analysis,

finding a substantially lower HGT level in aquatic species, while the number of ancestral genes in all organisms was comparable (Figure 12B). Horizontally transferred genes are frequently expendable, whereas ancestral genes usually serve ‘house-keeping’ functions and are conserved over long evolutionary distances [321]. To further validate our classifications, we determined functional assignments of genes in each of the two categories using the COG database [322]. The ‘ancestral’ set primarily contained genes involved in ‘house-keeping’ functions such as translation, posttranslational modification, cell division, and nucleotide and coenzyme metabolism (Figure 13). The HGT set contained a large proportion of genes involved in specific dispensable functions, such as defense mechanisms, cell wall biogenesis, transport and metabolism of amino acids, carbohydrates, inorganic ions and secondary metabolites (Figure 13 and File 6). This is consistent with the role of HGT in adaptation to the rhizosphere, an environment rich in amino acids, carbohydrates, inorganic ions and secondary metabolites excreted by plant roots [323].

Such an extraordinary high level of HGT in *Azospirillum* genomes leads us to hypothesize that it was a major driving force in the transition of these bacteria from aquatic to terrestrial environments and adaptation to plant hosts. This process was likely promoted by conjugation and transduction, as *Azospirillum* hosts phages and notably a Gene Transfer Agent [324]; this should have also resulted in loss of ancestral ‘aquatic’ genes that are not useful in the new habitat. Indeed, one of the defining features of *Rhodospirillaceae*, photosynthesis (responsible for the original taxonomic naming of these organisms – purple bacteria) is completely absent from *Azospirillum*. We have analyzed origins of genes that are proposed to be important for adaptation to the rhizosphere and interactions with the host plant [260, 325]. Consistent with our hypothesis, the majority of these genes were predicted to be horizontally transferred (Figure 14 and File 7). It is important however to stress that plant-microbe interactions involve a complex interplay of many functions that are determined by both ancestral and horizontally acquired genes.

What was the source of horizontally transferred genes? A large proportion of genes that we assigned as HGT show relatedness to terrestrial proteobacteria, including representatives of *Rhizobiales* (distantly related alpha-proteobacteria) and

Figure 11: Scheme for detecting ancestral and horizontally transferred genes.

See Materials and Methods for details.

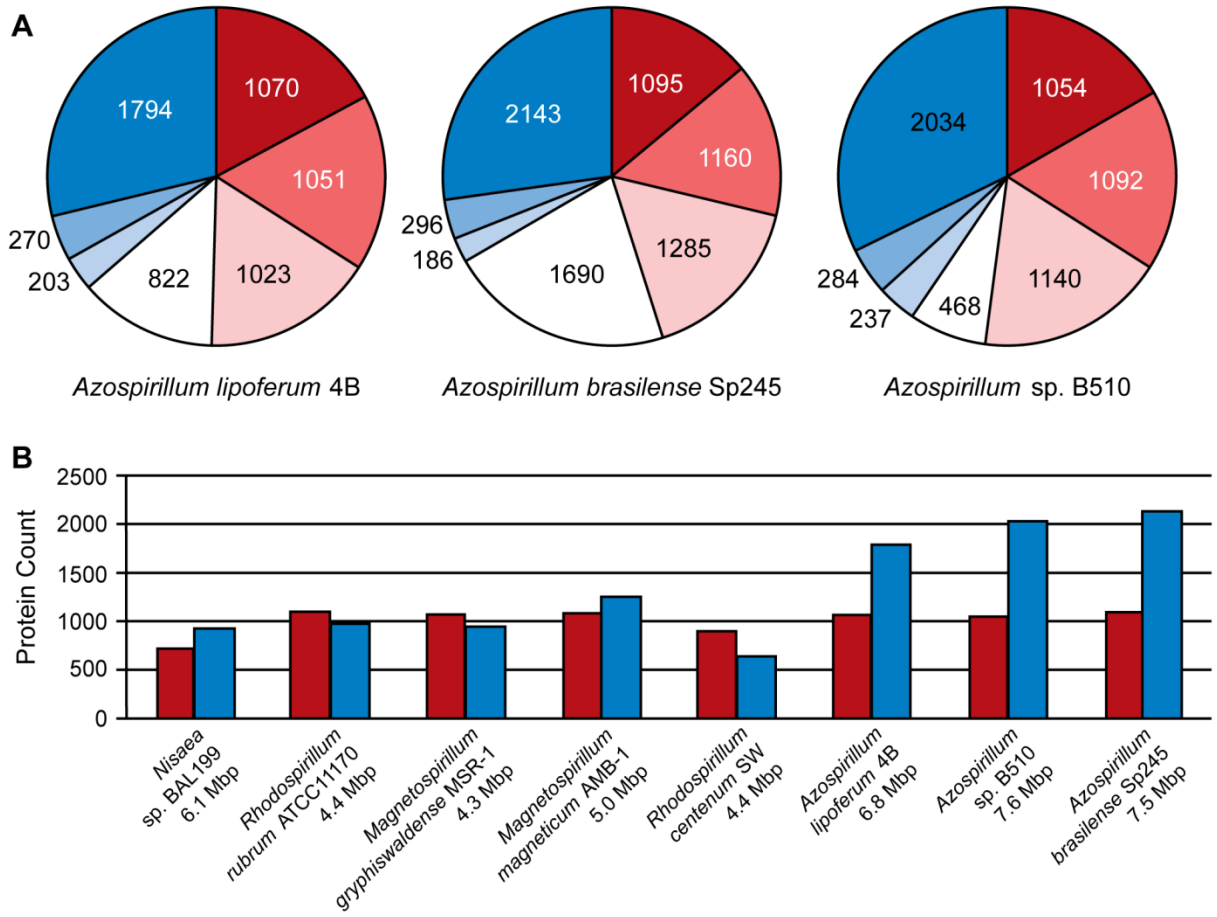

Figure 12: Ancestral (red) and horizontally transferred (blue) genes in *Azospirillum*.

(A) Proportion of ancestral and horizontally transferred genes predicted in three *Azospirillum* genomes with varying confidence: intensity of color shows high (dark), medium (medium) and low (light) levels of confidence for predictions (see Materials and Methods). Genes that cannot be assigned using this protocol are shown in white. Majority of these genes are unique to each species and have no identifiable homologs; thus, they are likely the result of HGT. (B) Proportion of ancestral and horizontally transferred genes in genomes of *Rhodospirillaceae*. Only genes that were predicted with high confidence are shown.

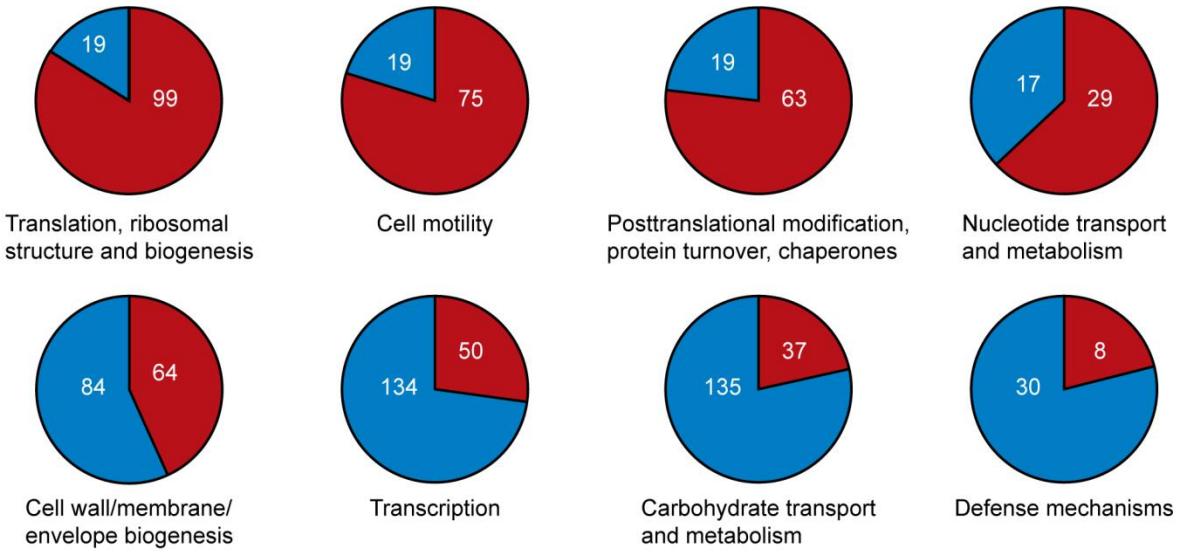

Figure 13: Functional categories for *A. lipoferum* 4B genes enriched in ancestral (top) and horizontally transferred (bottom) genes.

Only genes that were predicted with high confidence are shown.

Burkholderiales (beta-proteobacteria) (Figure 15) that are soil and plant-associated organisms. In the absence of fossil data, it is nearly impossible to determine the time of divergence for a specific bacterial lineage; however, a rough approximation (1–2% divergence in the 16S rRNA gene equals 50 Myr [326]) suggests that azospirilla might have diverged from their aquatic *Rhodospirillaceae* relatives 200–400 Myr (Table 6). This upper time limit coincides with the initial major radiation of vascular plants on land and evolution of plant roots, to 400 Myr [327, 328]. Grasses, the main plant host for *Azospirillum*, appeared much later, about 65–80 Myr [329], which is consistent with reports that azospirilla can also colonize plants other than grasses.

Using a global proteomics approach we have found that many HGT genes including nearly 1/3 of those that are common to all three *Azospirillum* genomes were expressed under standard experimental conditions and under nitrogen limitation, a

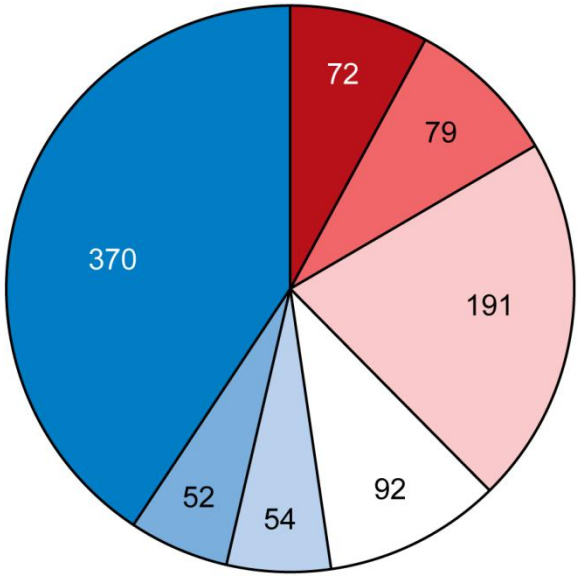

Figure 14: Proportion of ancestral (red) and horizontally transferred (blue) genes involved in adaptation of *Azospirillum* to the rhizosphere and its interaction with host plants (see File 6 for details).

Color intensity indicates high (dark), medium (medium) and low (light) confidence levels for prediction (see Materials and Methods for details).

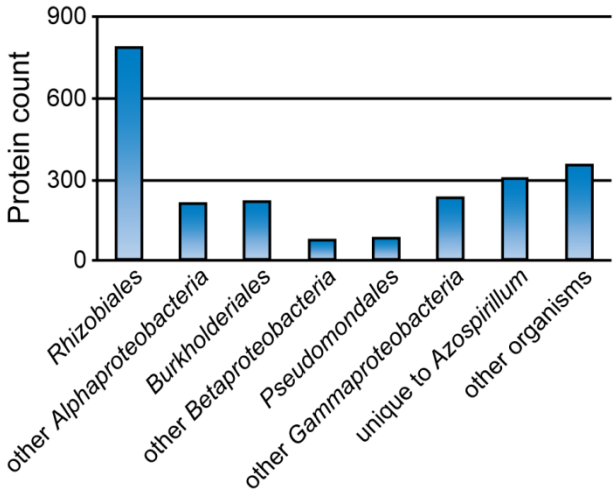

Figure 15: Taxonomic distribution of the best BLAST hits for predicted HGT in *Azospirillum*.

Table 6: Divergence in the 16S rRNA gene between *Azospirillum lipoferum* 4B and other members of*Rhodospirillaceae*.

***Azospirillum lipoferum* 4B**

Azospirillum sp. B510	2.16 %
Azospirillum brasilense Sp245	2.98 %
Rhodospirillum centenum SW	8.03 %
Nisaea sp. BAL199	9.18 %
Magnetospirillum magnetotacticum MS-1	9.34 %
Magnetospirillum magneticum AMB-1	9.46 %
Magnetospirillum gryphiswaldense MSR-1	10.12 %
Rhodospirillum rubrum ATCC 11170	10.93 %

condition usually encountered in the rhizosphere of natural ecosystems (Figure 16 and File 8).

Genes that differentiated the *Azospirillum* species from one another and from their closest relatives are implicated in specializations, such as plant colonization. *Azospirillum* and closely related*Rhodospirillum centenum* both possess multiple chemotaxis operons and are model organisms to study chemotaxis [330, 331]. Interestingly, operon 1, which controls chemotaxis in *R. centenum* [330], plays only a minor role in chemotaxis of *A. brasilense* [332]. All three *Azospirillum* species possess three chemotaxis operons that are orthologous to those in *R. centenum*; however, they also have additional chemotaxis operons that are absent from their close aquatic relative (Figure 17 and File 6 and Table 7). Additional chemotaxis operons have been acquired by azospirilla prior to each speciation event yielding 4, 5 and 6 chemotaxis systems in *A. brasilense* Sp245, *A. lipoferum* 4B and *Azospirillum* sp. 510, respectively. These stepwise acquisitions have made the latter organism an absolute “chemotaxis champion”, with 128 chemotaxis genes, more than any other prokaryote sequenced to

date (data from MiST database [333]). Recent analysis showed the prevalence of chemotaxis genes in the rhizosphere [334]. We have determined that the dominant chemotaxis genes in this dataset belong to a specific chemotaxis class F7 [107] (Figure 18 and Table 8). Strikingly, it is this F7 system that is shared by all *Azospirillum* and is predicted to have been transferred horizontally into their common ancestor.

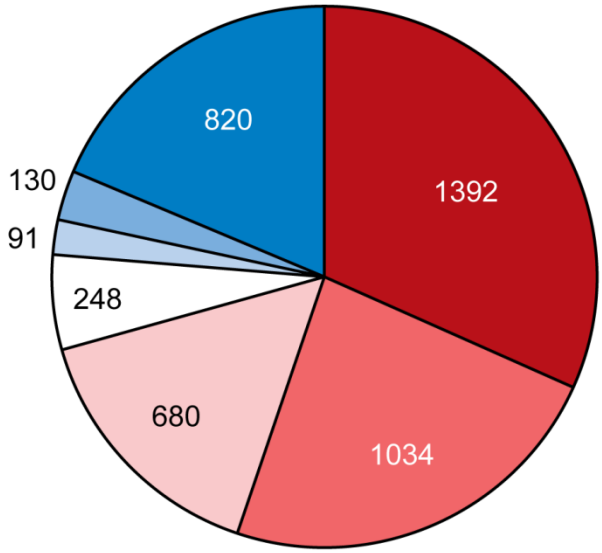

Figure 16: Proportion of ancestral (red) and horizontally transferred (blue) genes in the proteomics data for *A. lipoferum* 4B.

Color intensity indicates high (dark), medium (medium) and low (light) confidence levels for prediction. See Files 6 and 8 for details.

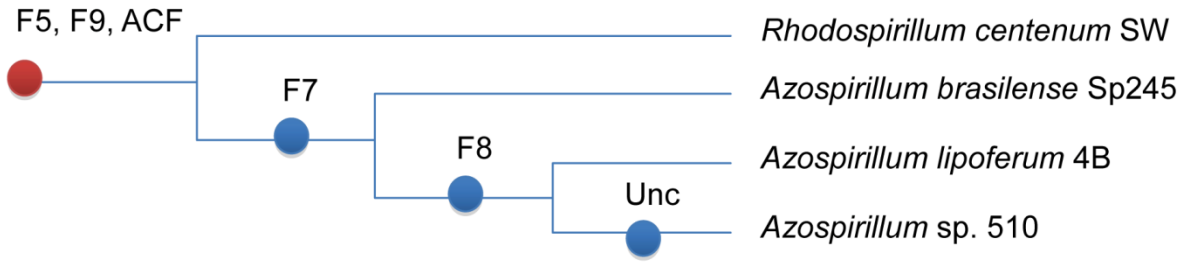

Figure 17: Chemotaxis operons in *Azospirillum*.

F5, F9 and ACF class chemotaxis systems were present in a common ancestor of azospirilla and other *Rhodospirillaceae* (e.g. *Rhodospirillum centenum*) [335, 336]. The F7 system was horizontally transferred to a common ancestor of *Azospirillum*. The F8 system was horizontally transferred to a common ancestor of *Azospirillum lipoferum*. The unclassified chemotaxis system (Unc) was obtained horizontally by *Azospirillum* sp. B510 only. See File 6 and Table 8 for detailed information for each system. Chemotaxis classes were assigned according to previous work by Wuichet & Zhulin [107].

Cellulolytic activity may be crucial to the ability of some azospirilla to penetrate plant roots [337]. All *Azospirillum* genomes encode a substantial number of glycosyl hydrolases that are essential for decomposition of plant cell walls (Figure 19). The total number of putative cellulases and hemicellulases in azospirilla is comparable to that in soil cellulolytic bacteria (Table 8) and most of them are predicted to be acquired horizontally (File 6). We tested three *Azospirillum* species and found detectable cellulolytic activity in *A. brasilense* Sp245 (Figure 20). The *A. brasilense* Sp245 genome contains three enzymes encoded by AZOBR_p470008, AZOBR_p1110164 and AZOBR_150049 (Figure 21) that are orthologous to biochemically verified cellulases. We propose that these and other horizontally transferred genes (*e.g.* glucuronate isomerase, which is involved in pectin decomposition) contributed to establishing *A. brasilense* Sp245 as a successful endophyte [337]. Interestingly, another successful endophytic bacterium, *Herbaspirillum seropedicae*, lacks the genes coding for plant cell wall degradation enzymes [338] indicating that endophytes may use very different strategies for penetrating the plant.

Attachment, another function important for plant association by *Azospirillum*, was also acquired horizontally. Type IV pili is a universal feature for initiating and maintaining contact with the plant host [339, 340]. The genome of *A. brasilense* Sp245 lacks genes coding for Type IV pili, but encodes a set of genes for TAD (tight adhesion) pili that are known to be HGT prone [341]. In our analysis, TAD pili were confidently predicted to be a result of HGT (File 6). We show that a mutant deficient in TAD pili had

Table 7: Orthologous chemotaxis operons in *Azospirillum* and *Rhodospirillum centenum*.

Chemotaxis genes	*R. centenum* SW	*A. brasilense* Sp245	*A. iipoferum* 4B	*Azospirillum* sp. B510
Operon 1 - F5				
CheA	RC1_1758	AZOBR_p1130073	AZOLI_p40444	AZL_d03050
CheW	RC1_1757	AZOBR_p1130074	AZOLI_p40443	AZL_d03040
CheY	RC1_1756	AZOBR_p1130075	AZOLI_p40442	AZL_d03030
CheB	RC1_1755	AZOBR_p1130076	AZOLI_p40441	AZL_d03020
CheR	RC1_1754	AZOBR_p1130077	AZOLI_p40440	AZL_d03010
Operon 2 - F9				
MCP	RC1_0344	AZOBR_p280105	AZOLI_p40518	AZL_d03500
CheW	RC1_0343	AZOBR_p280107	AZOLI_p40517	AZL_d03490
CheB	RC1_0342	AZOBR_p280108	AZOLI_p40516	AZL_d03480
Other (HEAT)	RC1_0341	AZOBR_p280109	AZOLI_p40515	AZL_d03470
CheR	RC1_0340	AZOBR_p280110	AZOLI_p40514	AZL_d03460
CheY	RC1_0339	AZOBR_p280111	AZOLI_p40513	AZL_d03450
CheA	RC1_0338, RC1_0337	AZOBR_p280112	AZOLI_p40512	AZL_d03440
Operon 3 - ACF				
CheY	RC1_2133			
CheW	RC1_2132	AZOBR_p1100030	AZOLI_p20369	AZL_a03160
CheR	RC1_2131, RC1_2130	AZOBR_p1100031	AZOLI_p20368	AZL_a03150
CheW	RC1_2129	AZOBR_p1100032	AZOLI_p20367	AZL_a03140
MCP	RC1_2128	AZOBR_p1100034	AZOLI_p20366	AZL_a03130
CheA	RC1_2127, RC1_2126	AZOBR_p1100035	AZOLI_p20364	AZL_a03120
CheB	RC1_2125	AZOBR_p1100037	AZOLI_p20363	AZL_a03110
RR	RC1_2124	AZOBR_p1100039	AZOLI_p20362	AZL_a03100

Table 7. Continued

Chemotaxis genes	*R. centenum* SW	*A. brasilense* Sp245	*A. iipoferum* 4B	*Azospirillum* sp. B510
Operon 4 - F7				
CheY		AZOBR_200200	AZOLI_2425	AZL_023410
CheA		AZOBR_200201, AZOBR_200202	AZOLI_2426	AZL_023420
CheW		AZOBR_200203	AZOLI_2427	AZL_023430
MCP		AZOBR_200204	AZOLI_2428	
CheR		AZOBR_200205	AZOLI_2429	AZL_023450
CheD		AZOBR_200206	AZOLI_2430	AZL_023460
CheB		AZOBR_200207	AZOLI_2431	AZL_023470
MCP		AZOBR_200208	AZOLI_2432	AZL_023480
Operon 5 - Unc				
MCP			AZOLI_1666	AZL_016690
MCP			AZOLI_1665	AZL_016680
CheR			AZOLI_1664	AZL_016670
CheW			AZOLI_1663	AZL_016660
CheB			AZOLI_1662	AZL_016650
CheY			AZOLI_1661	AZL_016640
CheA			AZOLI_1660	AZL_016630
Operon 6 - F8				
CheY				AZL_a08750
CheA				AZL_a08740
CheW				AZL_a08730
MCP				AZL_a08720
CheW				AZL_a08710
CheR				AZL_a08700
CheB				AZL_a08690
MCP				AZL_a08680

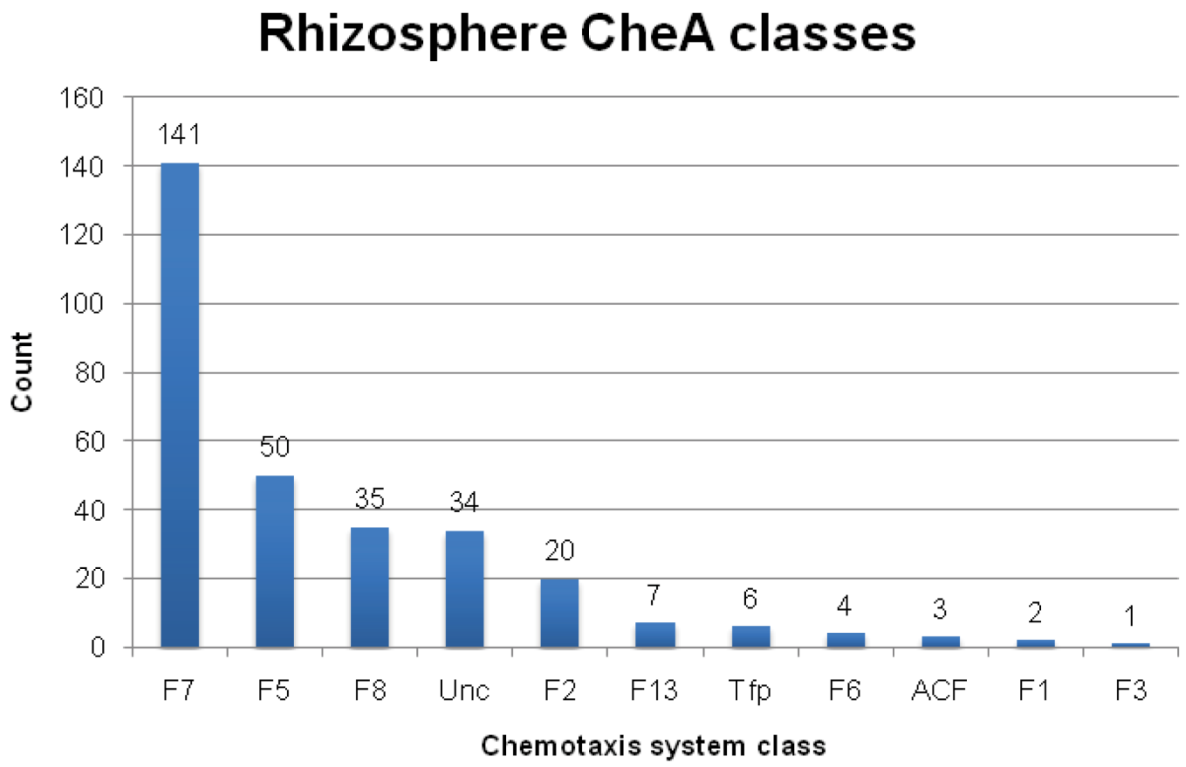

Figure 18: Abundance of the F7 chemotaxis system in the rhizosphere.

Chemotaxis systems were assigned as described in SI Materials and Methods. See Table 9 for detailed information.

a severe defect in attachment and biofilm formation (Figure 22) suggesting a role for TAD in plant-microbe association.

Concluding remarks

Horizontal gene transfer has been long recognized as a major evolutionary force in prokaryotes [321]. Its role in the emergence of new pathogens and adaptation to environmental changes is well documented [347]. While other recent studies indicate

Table 8: Classification of chemotaxis systems in rhizosphere.

Clone name	Assigned chemotaxis class	Clone name	Assigned chemotaxis class	Clone name	Assigned chemotaxis class	Clone name	Assigned chemotaxis class	Clone name	Assigned chemotaxis class
cprootA02	F7	cprootG03	Unc	soilB11N	F7	soilD12N	F7	whtrootC03	F8
cprootA03	F8	cprootG04	F7	soilB12	Unc	soilE02N	Tfp	whtrootC04	F7
cprootA04	F5	cprootG05	F7	soilB12N	F5	soilE03N	F7	whtrootC05	F7
cprootA05	F5	cprootG06	F7	soilC01	F7	soilE04N	F5	whtrootC06	F7
cprootA07	F7	cprootG07	F8	soilC01N	Unc	soilE05N	F2	whtrootC07	F8
cprootA09	F5	cprootG08	F5	soilC02	F8	soilE06N	F7	whtrootC08	F7
cprootA11	F5	cprootG09	F7	soilC02N	F2	soilE07N	F7	whtrootC09	F8
cprootA12	F7	cprootG10	F7	soilC03	F7	soilE08N	F5	whtrootC10	F13
cprootB01	F7	cprootG11	F7	soilC03N	F7	soilE10N	F7	whtrootC12	F7
cprootB02	F7	cprootG12	F5	soilC04	F2	soilE12N	F5	whtrootclone2	F7
cprootB03	F5	cprootH01	F5	soilC04N	F7	soilF02N	Tfp	whtrootclone3	F8
cprootB04	F5	cprootH02	Unc	soilC05	F7	soilF03N	F7	whtrootclone4	F7
cprootB05	F5	cprootH03	F2	soilC05N	F7	soilF04	F13	whtrootclone6	F8
cprootB06	F2	cprootH05	F7	soilC06	F8	soilF05N	F7	whtrootD01	F5
cprootB07	Unc	cprootH06	F5	soilC06N	F7	soilF06N	Unc	whtrootD02	F2
cprootB08	F7	cprootH07	F7	soilC07	F7	soilF07N	F5	whtrootD05	Unc
cprootB09	F5	cprootH08	F5	soilC08	F5	soilF08N	F5	whtrootD07	F7
cprootB10	F7	cprootH09	F5	soilC08N	F7	soilF09N	Unc	whtrootD09	F6
cprootB11	F7	cprootH10	ACF	soilC09	F7	soilF10N	F13	whtrootD10	F1
cprootB12	Unc	cprootH11	F7	soilC10	F7	soilF11N	F7	whtrootD11	F8
cprootC01	F7	cprootH12	F7	soilC11	F2	soilF12N	F5	whtrootD12	F8
cprootC02	F7	soilA01	F7	soilC11N	F7	soilG01N	F7	whtrootE01	F7
cprootC03	F5	soilA01	F7	soilC12	Unc	soilG04N	Unc	whtrootE02	F7
cprootC04	Unc	soilA02	Unc	soilC12N	F7	soilG05N	Unc	whtrootE03	F7
cprootC05	F7	soilA02N	Unc	soilCL1	F7	soilG06	F7	whtrootE04	F8
cprootC06	F5	soilA03	F7	soilCL2	F7	soilG07N	F6	whtrootE05	F7
cprootC07	Unc	soilA03N	F7	soilCL2B	F7	soilG08N	F7	whtrootE06	F8
cprootC08	Unc	soilA04	F5	soilCL3	F7	soilG09N	F1	whtrootE07	F13
cprootC09	F8	soilA04N	F8	soilCL3b	F7	soilG10N	F7	whtrootE09	F8
cprootC10	F5	soilA05	F7	soilCL4b	F5	soilG11N	F7	whtrootE10	F7
cprootC11	F5	soilA05N	Unc	soilCL5b	F2	soilG12	Unc	whtrootE11	F7

Table 8. Continued

Clone name	Assigned chemotaxis class	Clone name	Assigned chemotaxis class	Clone name	Assigned chemotaxis class	Clone name	Assigned chemotaxis class	Clone name	Assigned chemotaxis class
cprootclone1	F3	soilA06	F7	soilclone10	F7	soilH01N	F13	whtrootE12	F8
cprootclone2	F8	soilA06N	F7	soilclone2	F8	soilH02N	F5	whtrootF01	F13
cprootclone3	F5	soilA07	F7	soilclone3	F2	soilH03N	F5	whtrootF02	F7
cprootD01	F5	soilA07N	Unc	soilclone5	F2	soilH05N	F7	whtrootF03	F7
cprootD02	Unc	soilA08	Unc	soilclone6	F8	soilH06N	F7	whtrootF04	F7
cprootD03	F7	soilA08N	F7	soilclone7	F2	soilH07N	Unc	whtrootF05	F8
cprootD04	ACF	soilA09	F7	soilclone8a	F8	soilH08N	F7	whtrootF06	F8
cprootD08	F7	soilA09N	Unc	soilclone9	F7	soilH10N	Unc	whtrootF08	F7
cprootD09	F7	soilA10N	F5	soilD01	F6	soilH11N	F5	whtrootF09	F5
cprootD10	F7	soilA11	Unc	soilD01N	F7	soilH12N	F7	whtrootF10x429bpFOR	F7
cprootD11	F8	soilA11N	F7	soilD02	F2	whtrootA01	F5	whtrootF11	F7
cprootD12	F7	soilA12	F7	soilD02N	F7	whtrootA02	Unc	whtrootF12	F8
cprootE01	F7	soilA12N	Tfp	soilD03	F7	whtrootA03	F7	whtrootG01	F7
cprootE02	F7	soilB01	F7	soilD03N	F5	whtrootA04	F5	whtrootG03	Unc
cprootE03	Tfp	soilB01N	F2	soilD04	F8	whtrootA06	F8	whtrootG04	F7
cprootE04	F2	soilB02	F5	soilD04N	F13	whtrootA07	F7	whtrootG05	F5
cprootE06	F5	soilB03	F5	soilD05	F7	whtrootA08	F8	whtrootG06	Unc
cprootE07	F2	soilB03N	F7	soilD05N	Unc	whtrootA10	F7	whtrootG08	F7
cprootE08	F8	soilB04	F7	soilD06	F7	whtrootA12	F7	whtrootG09	F7
cprootE10	F7	soilB04N	Tfp	soilD06N	F5	whtrootB01	F8	whtrootG10	F2
cprootE12	F8	soilB05	F2	soilD07N	F7	whtrootB05	F7	whtrootG11	F7
cprootF01	F7	soilB05N	Tfp	soilD08	F5	whtrootB06	F7	whtrootG12	F7
cprootF03	F7	soilB06	F6	soilD08a	F7	whtrootB07	F8	whtrootH02	F2
cprootF04	F7	soilB07	F7	soilD09	F7	whtrootB08	Unc	whtrootH03	F8
cprootF07	F7	soilB08N	F7	soilD09N	F7	whtrootB09	F7	whtrootH04	F7
cprootF08	F5	soilB09	F7	soilD10	F5	whtrootB10	F7	whtrootH06	F5
cprootF09	ACF	soilB09N	F5	soilD10N	F7	whtrootB11	F5	whtrootH07	F7
cprootF10	Unc	soilB10	Unc	soilD11	F7	whtrootB12	F7	whtrootH10	F7
cprootF12	F7	soilB10N	F8	soilD11N	F7	whtrootC01	F2	whtrootH12	Unc
cprootG01	F8	soilB11	F2	soilD12	F7	whtrootC02	F7		

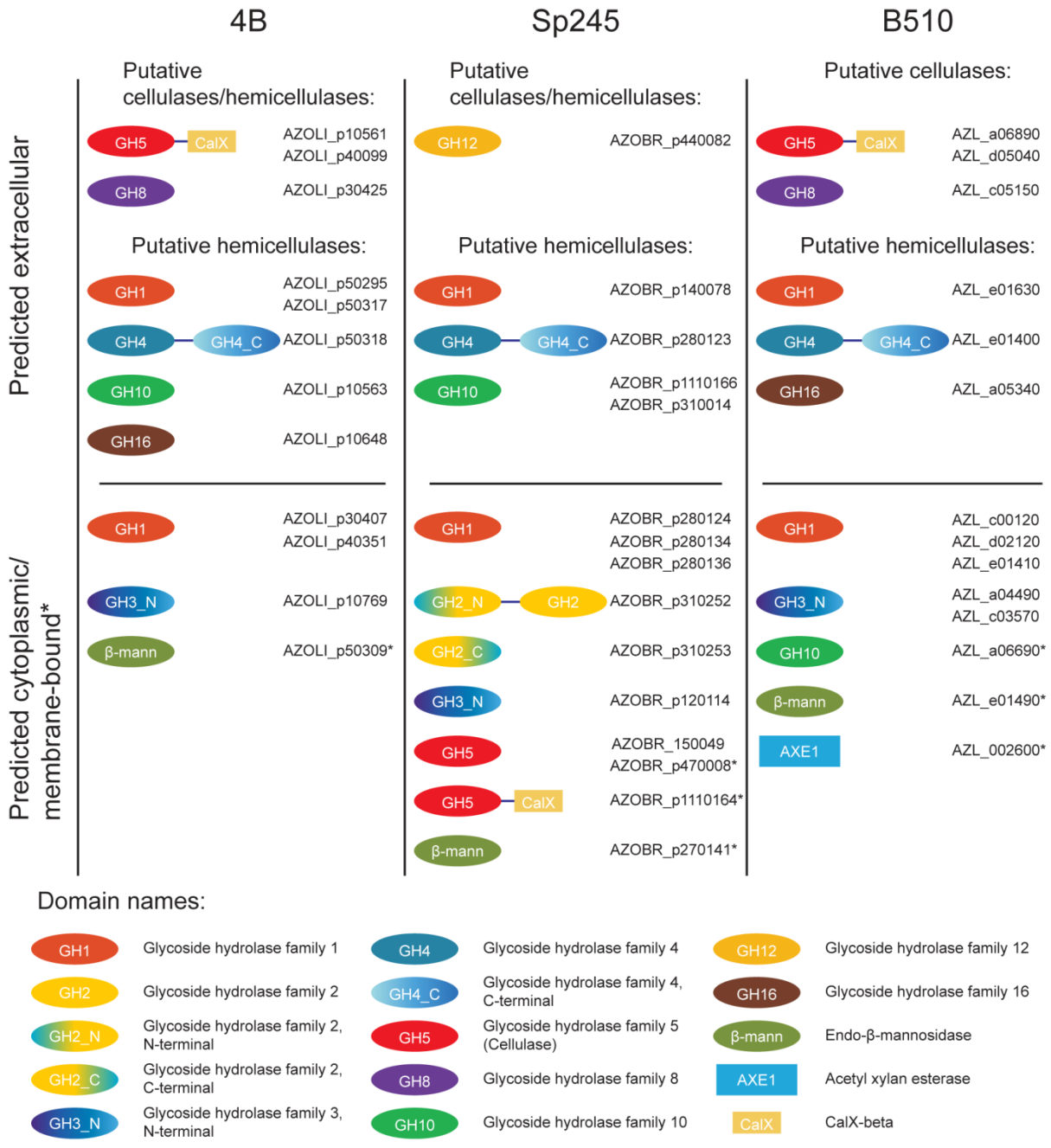

Figure 19: Glycoside hydrolases in *Azospirillum* with a potential to degrade the plant cell wall.

The genomes of *Azospirillum* encode from 26 to 34 glycoside hydrolases that belong to various CAZy [342] families (Table 8). Total number of glycoside hydrolases in *Azospirillum* species is similar to that in a soil cellulolytic bacterium *Thermobifida fusca* [343]. All three species have orthologs of putative cellulases (AZOLI_p10561, AZOLI_p40099; AZOBR_p1110164; AZL_a06890; AZL_d05040) with unique domain architecture: GH_5 – CalX-β. The other two putative cellulases (AZOBR_150049, AZOBR_p470008) are found only in *A. brasilense*. In addition to putative cellulases, *Azospirillum* species encode putative extracellular endoglucanases that may be involved in cellulose/hemicellulose degradation. For example, glycoside hydrolases that belong to family GH8 (AZOLI_p30425, AZL_c05150), which are known for a wide range of cellulose-containing substrates [344-346] and family GH12 (AZOBR_p440082). All three species are predicted to secrete a number of putative hemicellulases, that belong to glycoside hydrolase families GH1 (β-glycosidases), GH4 (glucuronidase/galactosidase), GH10 (endo-xylanases) and GH16 (licheninases) (Table 8). CAZy families were assigned as described in Materials and Methods.

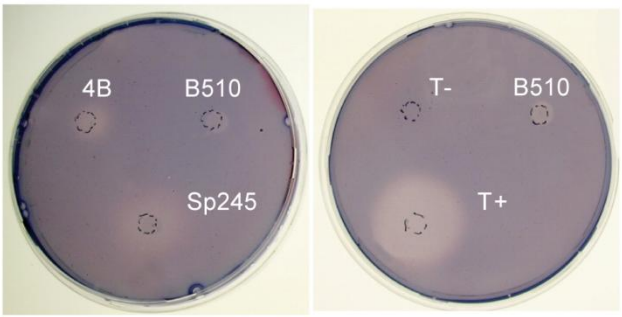

Figure 20: Cellulolytic activity of *A. brasilense* Sp245 cells.

All three *Azospirillum* species are shown on the left panel. Known cellulose degrader (*Dickeya dadantii* 3937, T+) and non-degrader (*Agrobacterium tumefaciens* NT1, T-) are shown as positive and negative controls, respectively.

Table 8: Putative complex carbohydrate-degrading enzymes in three *Azospirillum* species in comparison with a soil cellulolytic bacterium *Thermobifida fusca*.

CAZy family	Putative activity	*A. lipoferum* 4B	*A. brasilense* Sp245	*A.* sp. B510	*T. fusca*
GH 1	β-glucosidase, cellobiase	4	4	4	2
GH 2	β-mannosidase, β-glucruonidase, galactosidase	1	3	1	1
GH 3	Xylosidase, β -N-acetylhexosaminidase	1	1	2	2
GH 4	Glucuronidase, galactosidase, glucosidase	1	1	1	1
GH 5	Mannanase	0	0	0	1
GH 5	Cellulase, endogluconase	2	3	2	2
GH 6	Endogluconase	0	0	0	2
GH 8	Cellulase, endogluconase	1	0	1	0
GH 9	Endogluconase	0	0	0	2
GH 10, GH11	Endoxylanase, xylanase	1	2	1	3
GH 12	Endogluconase	0	1	0	0
GH 13, GH15	Amylase, pullulanase, dextranase	10	14	8	8
GH 16	Lichninase	1	0	1	0
GH 48	1,4-exocellulase	0	0	0	1
GH 17, GH25	Other	4	5	5	12
All	Glycosyl hydrolases	26	34	26	37

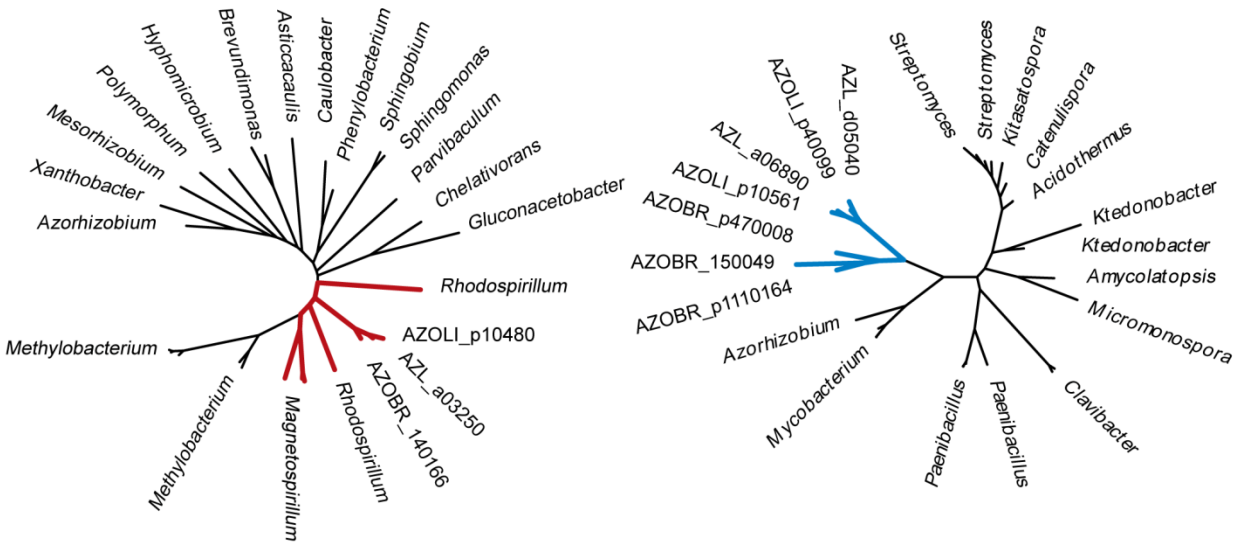

Figure 21: Phylogenetic trees for thiamine synthetase (left) and cellulase (right).

The trees exemplify ancestral and HGT relationships, respectively, that were predicted with high confidence. Trees were built from aligned sequences of the *A. brasilense* Sp245 query and twenty most similar sequences determined by BLAST. The thiamine synthetase set contains only representatives of alpha-proteobacteria including *Rhodospirillaceae* (shown in red). The cellulase set consists of representatives of Actinobacteria, Firmicutes, and Chloroflexi with only one representative of alpha-proteobacteria other than*Azospirillum* (that are shown in blue, highlighting their HGT origin), *Azorhizobium*.

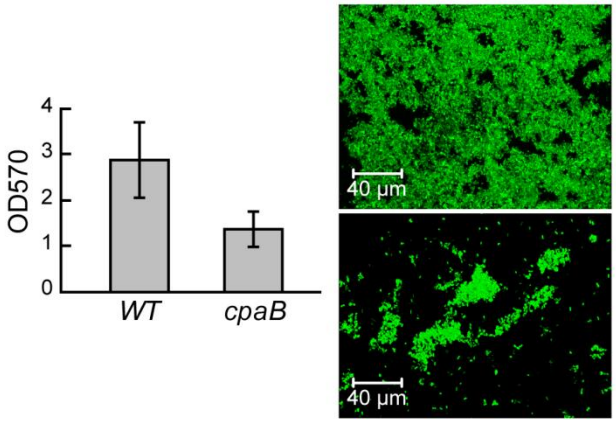

Figure 22: TAD pili in *A. brasilense* are required for biofilm formation.

Quantification of biofilm formed by wild type (wt) and a pili mutant (*cpaB*) on glass using crystal violet staining (left panel) and 3-D-reconstruction of the biofilm formed by wild type (top) and a pili mutant (bottom) by confocal microscopy (right panel).

that HGT levels in natural environments may reach as much as 20% of a bacterial genome [108], our data suggest that HGT has affected nearly 50% of the *Azospirillum* genomes, in close association with dramatic changes in lifestyle necessary for transition from aquatic to terrestrial environments and association with plants. Emergence of these globally distributed plant-associated bacteria, which appear to coincide with radiation of land plants and root development, likely has dramatically changed the soil ecosystem.

Materials and Methods

Genome sequencing and assembly

The genome of *Azospirillum lipoferum* 4B was sequenced by the whole random shotgun method with a mixture of ~12X coverage of Sanger reads, obtained from three different libraries, and ~18X coverage of 454 reads. Two plasmid libraries of 3 kb (A) and 10 kb (B), obtained by mechanical shearing with a Hydroshear device (GeneMachines, San Carlos, California, USA), were constructed at Genoscope (Evry, France) into pcDNA2.1 (Invitrogen) and into the pCNS home vector (pSU18 modified, Bartolome et al. [348]), respectively. Large inserts (40 kb) (C) were introduced into the PmlI site of pCC1FOS. Sequencing with vector-based primers was carried out using the ABI 3730 Applera Sequencer. A total of 95904 (A), 35520 (B) and 15360 (C) reads were analysed and assembled with 504591 reads obtained with Genome Sequencer FLX

(Roche Applied Science). The Arachne “HybridAssemble” version (Broad institute, MA) combining 454 contigs with Sanger reads was used for assembly. To validate the assembly, the Mekano interface (Genoscope), based on visualization of clone links inside and between contigs, was used to check the clones coverage and misassemblies. In addition, the consensus was confirmed using Consed functionalities (www.phrap.org), notably the consensus quality and the high quality discrepancies. The finishing step was achieved by PCR, primer walks and transposon bomb libraries and a total of 5460 sequences (58, 602 and 4800 respectively) were needed for gap closure and quality assessment.

The genome of strain *Azospirillum brasilense* Sp245 was sequenced by the whole random shotgun method with a mixture of ~10X coverage of Sanger reads obtained from three different libraries and ~25X coverage of 454 reads. A plasmid library of 3 kb, obtained by mechanical shearing with a Hydroshear device (GeneMachines, San Carlos, California, USA), were constructed at Plant Genome Mapping Laboratory (University of Georgia, USA) into pcDNA2.1 vector (Invitrogen). Large inserts (40 kb) were introduced into the PmlI site of pCC1FOS. Sequencing with vector-based primers was carried out using the ABI 3730 Applera Sequencer. The Arachne “HybridAssemble” version combining 454 contigs with Sanger reads was used for assembly. Contig scaffolds were created using Sequencher (Gene Codes) and validated using clone link inside and between contigs.

Genome annotation

AMIGene software [349] was used to predict coding sequences (CDSs) that were submitted to automatic functional annotation [350]. The resulting 6233 *A. lipoferum* 4B CDSs and 7848 *A. brasilense*Sp245 CDSs were assigned a unique identifier prefixed with “AZOLI” or “AZOBR” according to their respective genomes. Putative orthologs and synteny groups were computed between the sequenced genomes and 650 other complete genomes downloaded from the RefSeq database (NCBI) using the procedure described in Vallenet et al. [350]. Manual validation of the automatic annotation was performed using the MaGe (Magnifying Genomes) interface. IS finder (www-is.biotoul.fr)

was used to annotate insertion sequences [351]. The *A. lipoferum* 4B nucleotide sequence and annotation data have been deposited to EMBL databank under accession numbers: FQ311868 (chromosome), FQ311869 (p1), FQ311870 (p2), FQ311871 (p3), FQ311872 (p4), FQ311873 (p5), FQ311874 (p6). The *A. brasilense* Sp245 nucleotide sequence and annotation data have been deposited at EMBL databank under accession numbers: HE577327 (chromosome), HE577328 (p1), HE577329 (p2), HE577330 (p3), HE577331 (p4), HE577332 (p5), HE577333 (p6). In addition, all the data (i.e., syntactic and functional annotations, and results of comparative analysis) were stored in a relational database, called AzospirilluScope [350], which is publicly available at http://www.genoscope.cns.fr/agc/mage/microscope/about/collabprojects.php?P_id=39.

Computational genomics/bioinformatics

BLAST searches were performed using NCBI toolkit version 2.2.24+ [94]. Multiple sequence alignments were built using the L-INS-i algorithm of MAFFT [352] with default parameters. Phylogenetic tree construction was performed using PhyML [353] with default parameters unless otherwise specified. 16S rRNA sequences were retrieved from the Ribosomal Database Project [354].

A concatenated ribosomal protein tree was constructed from sequenced members of alpha-proteobacteria with a 98% 16S rRNA sequence identity cutoff to limit overrepresentation. The following ribosomal proteins were used: L3, L5, L11, L13, L14, S3, S7, S9, S11, and S17. The proteins were identified using corresponding Pfam models and HMMER [104] searches against the genomes of sequenced alpha-proteobacteria selected above. The sequences were aligned and concatenated. GBlocks [355] with default parameters was used to reduce the number of low information columns. The tree was constructed using PhyML with the following options: empirical amino acid frequencies, 4 substitution categories, estimated gamma distribution parameter, and NNI tree topology search.

As described by Harrison *et al.*[109], GC content was calculated for all the replicons. A cutoff value was calculated as GC within 0.521 ± 0.399% (mean ± standard deviation) of the host chromosome. Proteins present in all chromid containing genomes, as identified by Harrison *et al.*[109], were used to identify chromids in *Azospirillum* genomes.

Average Nucleotide Identity (ANIm) is calculated from the maximal unique matches (MUMs) determined by the MUMmer 2.1 program in pairwise comparisons [356]. Direct and palindromic repeats were calculated using the repfind application of REPUTER with the default parameters [357].

Assignment of gene ancestry

Protein sequences queries from all 3 *Azospirillum* genomes were used in BLAST searches against the non-redundant microbial genome set constructed by Wuichet and Zhulin [107] supplemented with sequenced members of *Rhodospirillales* absent in the original set (*Acetobacter pasteurianus* IFO 3283-01, *alpha proteobacterium* BAL199, *Magnetospirillum gryphiswaldense* MSR-1, and *Magnetospirillum magnetotacticum* MS-1). E-value cutoff of 10^{-4} was used.

Only the first occurrence of each species was used in ancestry assignment. Proteins were assigned as being ancestral or horizontally transferred, with varying degrees of confidence, based on the presence of members of *Rhodospirillales* and *Rhodospirillaceae* in the top eight BLAST hits. Ancestral assignment was based on the top 8 hits, based on the number of *Rhodospirillaceae* genomes in the database: 2 *Azospirillum*, 3 *Magnetospirillum*, 2 *Rhodospirillum*, and *Nisaea* sp. BAL199, excluding the organism on which ancestry assignment is being performed. High confidence ancestral proteins have at least 6 of the top 8 species belonging to *Rhodospirillales* or all but 1, if the BLAST result had less than 8 species. This rule allows for 1–2 independent events of HGT from *Rhodospirillales* to other distantly related species. Medium confidence ancestral proteins have at least 4 *Rhodospirillaceae* in the top 8. Low confidence ancestral proteins have at least

1 *Rhodospirillaceae* in the top 8, excluding hits to other *Azospirillum* genomes. High confidence horizontally transferred proteins have 0 hits to*Rhodospirillales* in the top 10, excluding hits to other *Azospirillum* genomes. Medium confidence horizontally transferred proteins have 0 hits to *Rhodospirillales* in the top 5, excluding hits to other*Azospirillum* genomes. Low confidence horizontally transferred proteins have 0 hits to *Rhodospirillaceae*in the top 8, excluding hits to other *Azospirillum* genomes. Unassigned proteins either have no BLAST hits outside *Azospirillum*, or simultaneously classify as medium confidence horizontally transferred and medium or low confidence ancestral.

Proteomics

Cell growth

Azospirillum brasilense strain Sp245: Overnight starter cultures (5 mL) were inoculated from fresh plates. Starter cultures were grown overnight at 27°C in a shaking water bath in minimal media containing malate as carbon source and ammonium chloride as nitrogen source. Cells were pelleted from starter cultures and washed with appropriate growth media. Base media for all cultures was minimal media (MMAB) [358] with 20 mM malate as carbon source, ammonium chloride as nitrogen source where appropriate, and molybdate. Starter cultures were resuspended with appropriate media and used to inoculate 250 mL cultures for nitrogen-fixing growth, or 500 ml cultures for non-nitrogen-fixing growth. Nitrogen fixation requires a great deal of energy and continuous optimal oxygen concentration, so growth of nitrogen fixing cells is slower than those growing in nitrogen sufficient conditions. Cells grown under nitrogen fixing conditions exhibit a doubling time of 170 minutes while control (non nitrogen fixing) cells have a doubling time of 120 minutes [331]. Further, OD of cells grown under nitrogen fixing cultures never reaches high levels, tending to level off at or below an OD600 of 0.2–0.3 [331]. Therefore, each growth condition was optimized as follows. For nitrogen-fixing cultures, nitrogen gas was sparged through the head space of the media bottle through the serum port, and sufficient air was injected to give a final oxygen content in the head space of 2%; cultures were grown at 25°C without shaking to early log phase

(OD_{600} = 0.1–0.2) to minimize exposure to high levels of oxygen, as *Azospirillum* species are microaerophilic diazotrophs. Non-nitrogen fixing cultures were grown under optimum growth conditions (shaking and in presence of ammonium) at 25°C on an orbital shaker to mid-log phase (OD_{600} = 0.5–0.6). Cells were harvested by centrifugation at 8000 rpm for 10 minutes, washed twice with 50 mM Tris (pH 7.9), then pelleted by centrifugation at 8000 rpm for 10 minutes, and stored at −80 C. Cell pellets from two biological replicates were pooled for subsequent proteome preparation. *Azospirillum lipoferum*: Growth conditions were as described above for *A. brasilense* Sp245, except that cells were grown in MMAB media supplemented with 1 mg/L D-biotin.

Proteome preparation for LC/LC-MS/MS

Frozen cell pellets (0.1 g for each sample) were resuspended at a rate of 500 µl lysis buffer/0.1 g wet cell pellet weight in lysis buffer of 6 M guanidine hydrochloride, 10 mM DTT solubilized in 50 mM Tris-HCl, 10 mM $CaCl_2$ [359]. Resuspended cells were then further lysed by sonication. Lysate was centrifuged at 18,000 g for 20 minutes to clear cellular debris. Supernatant was collected for tryptic digestion. 10 mM DTT was added and lysate was incubated at 60°C for 1 hour. Lysate was then diluted 6-fold with trypsin digestion buffer (50 mM Tris-HCl, 10 mM $CaCl_2$, 10 mM DTT, pH 7.9) and 20 µg sequencing-grade trypsin (Promega, Madison, WI) was added to each sample. Samples were incubated overnight at 37°C with gentle rotation. An additional 20 µg of trypsin was added the following morning and samples were subsequently incubated for an additional 5–6 hours at 37°C with gentle rotation. Digestion was halted by addition of 5 µl formic acid to the 5 ml lysate. Samples were then desalted using Sep-Pak Plus C-18 solid phase extraction (Waters, Milford, MA) following manufacturer's recommendations, and subsequently concentrated and solvent-exchanged into 100% HPLC-grade H_2O, 0.1% formic acid using vacuum centrifugation (Savant, Thermo Scientific). Samples were aliquoted into 40 µL volumes and stored at −80°C until analysis.

LC/LC-MS/MS analysis

Proteome samples were analyzed via Multi-dimensional Protein Identification Technology (MudPIT) [360-362] with triphasic columns. Columns were individually packed using a pressure cell (New Objective, Woburn, MA). Back columns were loaded in 150 µm ID fused silica capillary tubing first with 3 cm of Luna 5 µm particle diameter strong cation exchange (SCX) resin (Phenomenex, Torrance, CA) followed by 3 cm of Aqua 5 µm C-18 reverse phase resin (Phenomenex). Proteome aliquots (40 µl) were loaded directly onto the back column via pressure cell and subsequently coupled to the front column. Front columns were pulled from 100 µm ID fused silica capillary tubing to a tip with an inside diameter of 5 µm using a P-2000 laser puller (Sutter Instruments, Novato, CA), and packed with a 17 cm long bed of Aqua 5 µm diameter C-18 reverse phase resin. This column acts as the resolving column for peptides eluted from the back column. For analysis, the combined columns were placed directly in-line with an LTQ mass spectrometer (ThermoScientific, San Jose, CA) using a Proxeon source.

Chromatographic separation was accomplished with an Ultimate HPLC system (LC Packings, a division of Dionex, San Francisco, CA) providing a flow rate of 100 µl/minute which was split prior to the resolving column such that the final flow rate through the resolving column was ~300 nl/minute. Twelve two-dimensional (2D) chromatographic steps were done. An initial 1 hour gradient from buffer A (95% water, 5% acetonitrile, 0.1% formic acid) to buffer B (70% acetonitrile, 0.1% formic acid) bumped the peptides from the initial reverse phase column onto the strong cation exchange column. Subsequent cycles included 2 minute salt pulses with varying percentages of 500 mM ammonium acetate (10, 15, 20, 25, 30, 35, 40, 45, 50, 60%) to first elute subsets of peptides from the SCX column according to charge, followed by a 2 hour gradient from buffer A to buffer B, to further separate peptides by hydrophobicity. The final chromatographic step consisted of a 20 minute salt pulse of 100% 500 mM ammonium acetate, followed by a 2 hour A-to-B gradient.

Data collection was controlled by Xcaliber software (ThermoScientific). Data was collected in data-dependent mode with one full scan followed by 6 dependent scans, each with 2 microscans. Dynamic exclusion was employed with a repeat count of 1,

repeat duration of 60 s and exclusion list size of 300 and duration of 180 s. Isolation mass width was set at 3 m/z units.

Data analysis

The Sp245 protein database was constructed from translated CDSs called in the draft genome sequence (http://genome.ornl.gov/microbial/abra/19sep08/). The 4B protein database was constructed from translated CDSs called in the complete genome sequence. A list of common contaminants was appended to the gene call sequences, and all coding sequences, including contaminant sequences, were reversed and appended to the forward sequences in order to serve as distractors. From the number of identifications in the reverse direction, peptide false positive (FP) rates were determined using the formula %FP = 2[No. reverse ID/(no. reverse ID+no. real ID)] [363]; FP rates ranged from 1.4%–4.3%. All MS/MS spectra were searched against the corresponding database using SEQUEST [364], specifying tryptic digestion, peptide mass tolerance of 3 m/z and a fragment ion tolerance of 0.5 m/z. Additionally, search parameters included two dynamic modifications: 1. methylation represented by a mass shift of +14 m/z on glutamate residues, and 2. deamidation followed by methylation represented by a mass shift of +15 m/z on glutamine residues. Output data files were sorted and filtered with DTASelect [365], specifying XCorr filter levels of 1.8 for peptides with a charge state of +1, 2.5 for those with charge state +2 and 3.5 for charge state +3, minimum delta CN of 0.08, semi-tryptic status and 2 peptides per protein identification. In order to determine relative abundance of a given protein in a sample, normalized spectral abundance factors (NSAF) were calculated for each individual protein k using the formula $NSAF_k = (SpC/L)_k/\Sigma\ (SpC/L)_n$, where SpC is the total spectral count for all peptides contributing to protein k, L is the length of protein k, and n is the total number of proteins detected in the sample [366].

Identification of glycoside hydrolases

Bidirectional BLAST was used to identify orthologs of the putative glycoside hydrolase (GH) genes. Phyml package was used to confirm evolutionary relationships and visualize the results. Domain architectures were obtained through Pfam [367] search for each protein. Then information from CAZy [342] and recent analysis [368] was used to assign putative activities of the predicted GHs.

Classification of chemotaxis systems in the rhizosphere

Chemotaxis proteins were identified in genomic datasets as previously described [369]. Using CheA sequences from a recent chemotaxis system classification analysis [107], alignments of the P3–P5 regions of CheA were built for each class and for the entire set of CheA sequences. Each alignment was made non-redundant so that no pair of sequences shared more than 80% sequence identity. Hidden Markov Models (HMMs) were built from each non-redundant alignment and used to create library via the HMMER3 software package (version HMMER 3.0b3) [104] and default parameters.

The rhizosphere CheA sequences from a recent study [334] were run against the CheA HMM library. Unclassified sequences (Unc) are those with top hits to the full CheA set HMM rather than a class-specific HMM. The remaining sequences were assigned to the class of the top scoring HMM.

Cellulase assay

Azospirillum strains and control strains (*Dickeya dadantii* 3937 as a positive control, *A. tumefaciens* NT1 as a negative control) were cultured for 16 h in liquid AB minimal medium [370] containing 0.2% malate and 1 mg/L biotin. An aliquot of 10^7 cells (for *Dickeya dadantii* 3937) or 2.10^7 cells (for all other strains) was deposited on top of AB plates containing 0.1% carboxymethylcellulose instead of malate. Plates were incubated for 5 days before being stained as previously described [371].

Pili mutant and attachment assay

A 211-bp *cpaB* (AZOBR_p460079) internal fragment was amplified by PCR with primers F6678 (GCGTGGACCTGATCCTGAC) and F6679 (GTGACCGTCTCGCTCTGAC) and subcloned into pGEM-T easy (Promega). White colonies were screened by PCR with primers F6678 and F6679 for correct insertion in pGEM-T easy, resulting in pR3.37. The insert of plasmid pR3.37 was digested with *Not*I and cloned into the *Not*I site of pKNOCK-Km [372], resulting in pR3.39 after transfer into chemically-competent cells of *E. coli* S17.1 λ*pir*. pR3.39 was introduced into *A. brasilense* Sp245 by biparental mating. Transconjugants resulting from a single recombination event of pR3.39 were selected on AB medium containing 0.2% malate, ampicillin (100 mg/mL) and kanamycin (40 mg/mL). The correct insertion of pKNOCK into *cpaB* was confirmed by PCR with primers (F6678 and F5595 TGTCCAGATAGCCCAGTAGC, located on pKNOCK) and sequencing of the PCR amplicon.

Sp245 and Sp245*cpaB* were labelled with pMP2444 [373] allowing the constitutive expression of EGFP. The strains were grown in NFB* (Nitrogen free broth containing 0.025% of LB) with appropriate antibiotics in glass tubes containing a cover-slide, under a mild lateral agitation for 6 days. After the incubation, the liquid and the cover-slide were removed from the tubes and the biofilm formed at the air/liquid interface was colored by 0.1% crystal violet. After two washings with distilled water, crystal violet was solubilized by ethanol and quantified by spectrophotometry at 570 nm. The experiment was performed twice in triplicate. In parallel, the colonization of the glass cover-slide was monitored by confocal laser scanning microscopy (510 Meta microscope; Carl Zeiss S.A.S.) equipped with an argon-krypton laser, detectors, and filter sets for green fluorescence (i.e., 488 nm for excitation and 510 to 531 nm for detection). Series of horizontal (*x-y*) optical sections with a thickness of 1 µm were taken throughout the full length of the Sp245 and Sp245*cpaB* biofilms. Three dimensional reconstructions of biofilms were performed using LSM software release 3.5 (Carl Zeiss S.A.S.).

CONCLUSION

The research presented in this dissertation represents the application of bioinformatics to biological problems at three levels of complexity: protein domain, protein network, and whole genome. The problems studied were the identification and functional characterization of the novel sensory FIST domain, evolutionary analysis of the chemotaxis system in *Escherichia*, and comparative genomic analysis of two *Azospirillum* genomes. This range of scales demonstrated how bioinformatics can be effectively applied to a wide variety of problems, providing unique insight at many levels of complexity.

The domain study, described in Chapter 1, showed how bioinformatics can be applied to examine a previously uncharacterized protein region. Bioinformatic analysis of the novel FIST domain provided for its annotation and functional characterization. Sequence analysis tools allowed us to systematically analyze unknown regions in signal transduction proteins, which led to the discovery of the FIST domain. Further analysis showed that this domain was widely distributed in genomes of bacteria, archaea, and eukarya, which implied its functional importance and ancestral origin. Its function as an input domain was proposed due to its exclusive appearance with transducer and output domains. Genomic context analysis showed that the FIST domain containing proteins were found preferentially near amino acid metabolism and transport proteins. This further suggested that FIST functioned as a small ligand binder, possibly an amino acid or amino acid metabolite.

The protein network study, described in Chapter 2, shows how bioinformatics can be used to gain insight into the evolutionary forces affecting a protein network. Analysis of the *Escherichia* chemotaxis system allowed for the forces that affect its conservation to be studied. Additionally, this analysis provides insight into the evolutionary relationship between the chemotaxis system and *E. coli* pathogenicity. Comparative genomic analysis showed that this system is conserved in the majority of sequenced *Escherichia* strains. This further confirmed that the chemotaxis system provides an adaptive advantage to *Escherichia* in all habitats. It was found that some of the traditionally nonmotile *Shigella* retained their chemotaxis system. Since the chemotaxis system is known to be adapted to uses other than switching flagellar motor, it is likely that the chemotaxis system in *Shigella* still functions as a signal transduction system but

with a different output. The finding that a large number of *Escherichia* were losing some of their receptors confirms that many of the receptors are accessory components. This implies that other sensory functions, such as osmotaxis, pH taxis, and thermotaxis, that do not require accessory receptors are adaptively more important that the sensing of chemicals provided by the accessory receptors. This point was further reinforced by the relative lack of receptor acquisition either through duplication or horizontal gene transfer. Additional receptors were found only on plasmids, suggesting their temporary nature.

The genome study, described in Chapter 3, shows how bioinformatics can be applied to gain insight into the evolutionary history of an organism. Sequencing and genomic analysis of two strains of *Azospirillum* was carried out. The analysis confirmed the complex arrangement of replicons previously noticed in *Azospirillum* strains. Also, it was noted that the genomes possessed remarkable plasticity. Comparative genomic analysis revealed that *Azospirillum* acquired almost half of its genes through horizontal gene transfer. Among those acquired genes were many implicated in adaptation to the terrestrial environment, survival in the rhizosphere, and plant growth-promoting properties. It was found that many of the horizontally transferred genes were also expressed in proteomic analyses, providing further evidence of their importance for the adaptation of *Azospirillum*. This level of horizontal gene transfer is unprecedented. This genome provides a unique insight into how bacteria can adapt to a drastically different niche through massive acquisition of foreign DNA.

As these studies illustrate, the application of bioinformatics tools to sequence data allows for the study of a wide variety of biological problems, providing novel insight at those levels of complexity presented in this dissertation.

Future Aims

A progression towards additional integration of advanced bioinformatic principles will allow for novel knowledge to be extracted from sequence data. With the increasing

ability of sequencing technologies to produce cheaper and larger volumes of data, much of the manual bioinformatics analysis that was discussed in this dissertation will soon become impractical. Although automation of bioinformatics applications cannot provide the insights derived from manual work, it will allow for improved data processing. This will be critical for analyzing the increasing protein and nucleotide databases.

The systematic approach for domain identification and functional characterization, while providing valuable insight, is a very manual process. Automation of this process can lead to many additional discoveries and provide functional characterization to many domains of unknown function (DUFs). Manual domain discovery and analysis is already being replaced by automated processes to cover unannotated protein space [374]. However, these methods only identify novel domains without investing any resources into their functional characterization. The result is the abundance of DUFs. While these DUFs provide valuable starting points for both bioinformaticians and biochemists in the pursuit of their functional characterization, additional automation will allow for these ventures to be more fruitful. Incorporating genomic context, as described in Chapter 1, and structural fold information into the available domain databases will allow for increased functional prediction of these DUFs.

Similarly, automated tools and databases exist to help extract information from newly sequenced genomes [350, 375]. However, their analysis does not delve deeply into comparative genomics. Automation of the ancestry assignment scheme presented in Chapter 3 in a framework such as SEED [375] would provide information on the origin of proteins, which would be valuable in understanding the evolutionary history of that organism. Studying the adaptation of organisms to their environment would be greatly aided by knowing the ancestry of their proteome. As more genomes are sequenced, covering more taxomomic space, the scheme for assigning ancestry becomes more powerful and refined. This in turn would increase the confidence of the predictions.

REFERENCES

1. Kafatos FC: **Challenges for European biology.** *Science* 1998, **280:**1327.
2. C.A. O: **Rise and demise of bioinformatics? Promise and progress.** *PLoS Comput Biol* 2012, **8:**e1002487.
3. Sanger F, Tuppy H: **The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates.** *Biochem J* 1951, **49:**463-481.
4. Sanger F, Tuppy H: **The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates.** *Biochem J* 1951, **49:**481-490.
5. Sanger F, Thompson EO: **The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates.** *Biochem J* 1953, **53:**353-366.
6. Sanger F, Thompson EO: **The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates.** *Biochem J* 1953, **53:**366-374.
7. Du Vigneaud V, Ressler C, Trippett S: **The sequence of amino acids in oxytocin, with a proposal for the structure of oxytocin.** *J Biol Chem* 1953, **205:**949-957.
8. Watson JD, Crick FHC: **Genetic-Implications of the Structure of Deoxyribonucleic-Acid.** *Jama-Journal of the American Medical Association* 1993, **269:**1967-1969.
9. Gamow G, Rich A, Ycas M: **The problem of information transfer from the nucleic acids to proteins.** *Adv Biol Med Phys* 1956, **4:**23-68.
10. Horowitz NH: **On the Evolution of Biochemical Syntheses.** *Proc Natl Acad Sci U S A* 1945, **31:**153-157.
11. Britten RJ: **Rates of DNA-Sequence Evolution Differ between Taxonomic Groups.** *Science* 1986, **231:**1393-1398.
12. Chaitin GJ: **On the Length of Programs for Computing Finite Binary Sequences.** *J ACM* 1966, **13:**547-569.
13. Shannon CE, Weaver W: *The mathematical theory of communication.* Urbana,: University of Illinois Press; 1949.
14. Chomsky N: **On certain formal properties of grammars.** *Information and Control* 1959, **2:**137-167.
15. Von Neumann J, Morgenstern O: *Theory of games and economic behavior.* 3d edn. Princeton,: Princeton University Press; 1953.
16. Von Neumann J, Burks AW: *Theory of self-reproducing automata.* Urbana,: University of Illinois Press; 1966.
17. Eck RV, Dayhoff MO: **Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences.** *Science* 1966, **152:**363-366.
18. Mclaughl.Pj, Hunt LT, Barker WC, Dayhoff MO: **Myoglobin Evolution Deduced from Amino Acid Sequences.** *Federation Proceedings* 1971, **30:**1208-&.
19. Dayhoff MO, Silver Spring MNBRF: *Atlas of Protein Sequence and Structure.* 1965.
20. Dayhoff MO: **Computer analysis of protein evolution.** *Sci Am* 1969, **221:**86-95.
21. Zuckerkandl E, Pauling L: **Molecules as documents of evolutionary history.** *J Theor Biol* 1965, **8:**357-366.
22. Ingram VM: **Gene evolution and the haemoglobins.** *Nature* 1961, **189:**704-708.
23. Margoliash E: **Primary Structure and Evolution of Cytochrome C.** *Proc Natl Acad Sci U S A* 1963, **50:**672-679.
24. Gatlin LL: **The information content of DNA.** *J Theor Biol* 1966, **10:**281-300.
25. Nolan C, Margoliash E: **Comparative aspects of primary structures of proteins.** *Annu Rev Biochem* 1968, **37:**727-790.
26. Cantor CR: **The occurrence of gaps in protein sequences.** *Biochem Biophys Res Commun* 1968, **31:**410-416.
27. Fitch WM, Margoliash E: **Construction of phylogenetic trees.** *Science* 1967, **155:**279-284.

28. Nei M: **Gene duplication and nucleotide substitution in evolution.** *Nature* 1969, **221:**40-42.
29. Gibbs AJ, McIntyre GA: **The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences.** *Eur J Biochem* 1970, **16:**1-11.
30. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol* 1970, **48:**443-453.
31. King JL, Jukes TH: **Non-Darwinian evolution.** *Science* 1969, **164:**788-798.
32. Epstein CJ: **Non-randomness of amino-acid changes in the evolution of homologous proteins.** *Nature* 1967, **215:**355-359.
33. Clarke B: **Selective Constraints on Amino-acid Substitutions during the Evolution of Proteins.** *Nature* 1970, **228:**159-160.
34. Ptitsyn OB: **Statistical analysis of the distribution of amino acid residues among helical and non-helical regions in globular proteins.** *J Mol Biol* 1969, **42:**501-510.
35. Pain RH, Robson B: **Analysis of the code relating sequence to secondary structure in proteins.** *Nature* 1970, **227:**62-63.
36. West MW, Ponnamperuma C: **Chemical evolution and the origin of life. A comprehensive bibliography.** *Space Life Sci* 1970, **2:**225-295.
37. Ohno S: *Evolution by gene duplication.* London, New York: Springer-Verlag; 1970.
38. Sanger F, Donelson JE, Coulson AR, Kossel H, Fischer D: **Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA.** *Proc Natl Acad Sci U S A* 1973, **70:**1209-1213.
39. Sanger F, Coulson AR: **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.** *J Mol Biol* 1975, **94:**441-448.
40. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors.** *Proc Natl Acad Sci U S A* 1977, **74:**5463-5467.
41. Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL, Fiddes JC, Hutchison CA, 3rd, Slocombe PM, Smith M: **The nucleotide sequence of bacteriophage phiX174.** *J Mol Biol* 1978, **125:**225-246.
42. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, et al: **Sequence and organization of the human mitochondrial genome.** *Nature* 1981, **290:**457-465.
43. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB: **Nucleotide sequence of bacteriophage lambda DNA.** *J Mol Biol* 1982, **162:**729-773.
44. Kimura M: **The rate of molecular evolution considered from the standpoint of population genetics.** *Proc Natl Acad Sci U S A* 1969, **63:**1181-1188.
45. Ohta T, Kimura M: **Functional Organization of Genetic Material as a Product of Molecular Evolution.** *Nature* 1971, **233:**118-&.
46. Kimura M: *The neutral theory of molecular evolution.* Cambridge Cambridgeshire ; New York: Cambridge University Press; 1983.
47. Jukes TH, Holmquist R: **Evolutionary clock: nonconstancy of rate in different species.** *Science* 1972, **177:**530-532.
48. Kimura M, Ota T: **On some principles governing molecular evolution.** *Proc Natl Acad Sci U S A* 1974, **71:**2848-2852.
49. Waterman MS, Smith TF, Beyer WA: **Some Biological Sequence Metrics.** *Advances in Mathematics* 1976, **20:**367-387.
50. Klotz LC, Komar N, Blanken RL, Mitchell RM: **Calculation of Evolutionary Trees from Sequence Data.** *Proceedings of the National Academy of Sciences of the United States of America* 1979, **76:**4516-4520.

51. Riley M: **Discontinuous-Processes in the Evolution of the Bacterial Genome.** *Evolutionary Biology* 1985, **19:**1-36.
52. Gingeras TR, Roberts RJ: **Steps toward Computer-Analysis of Nucleotide-Sequences.** *Science* 1980, **209:**1322-1328.
53. Hamm GH, Cameron GN: **The Embl Data Library.** *Nucleic Acids Res* 1986, **14:**5-9.
54. Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FI, Rindone WP, Swindell CD, Tung CS: **The Genbank Genetic Sequence Data-Bank.** *Nucleic Acids Res* 1986, **14:**1-4.
55. Cochrane G, Karsch-Mizrachi I, Nakamura Y: **The International Nucleotide Sequence Database Collaboration.** *Nucleic Acids Res* 2011, **39:**D15-18.
56. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147:**195-197.
57. Smith TF: **Comparison of biosequences.** *Journal Name: Advances in Applied Mathematics; (United States); Journal Volume: 2* 1981**:**Medium: X; Size: Pages: 482-489.
58. Lipman DJ, Pearson WR: **Rapid and sensitive protein similarity searches.** *Science* 1985, **227:**1435-1441.
59. Feng DF, Doolittle RF: **Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees.** *J Mol Evol* 1987, **25:**351-360.
60. Higgins DG, Sharp PM: **Clustal - a Package for Performing Multiple Sequence Alignment on a Microcomputer.** *Gene* 1988, **73:**237-244.
61. Walker JE, Saraste M, Runswick MJ, Gay NJ: **Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold.** *EMBO J* 1982, **1:**945-951.
62. Klug A, Rhodes D: **Zinc Fingers - a Novel Protein Motif for Nucleic-Acid Recognition.** *Trends Biochem Sci* 1987, **12:**464-469.
63. Gribskov M, Burgess RR: **Sigma factors from E. coli, B. subtilis, phage SP01, and phage T4 are homologous proteins.** *Nucleic Acids Res* 1986, **14:**6745-6763.
64. von Heijne G: **On the hydrophobic nature of signal sequences.** *Eur J Biochem* 1981, **116:**419-422.
65. Doolittle RF: **Similar amino acid sequences: chance or common ancestry?** *Science* 1981, **214:**149-159.
66. Dayhoff MO, Barker WC, Hunt LT: **Establishing homologies in protein sequences.** *Methods Enzymol* 1983, **91:**524-545.
67. Altschuh D, Vernet T, Berti P, Moras D, Nagai K: **Coordinated amino acid changes in homologous protein families.** *Protein Eng* 1988, **2:**193-199.
68. Reeck GR, de Haen C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH, et al.: **"Homology" in proteins and nucleic acids: a terminology muddle and a way out of it.** *Cell* 1987, **50:**667.
69. Lesk AM, Chothia C: **How Different Amino-Acid-Sequences Determine Similar Protein Structures - Structure and Evolutionary Dynamics of the Globins.** *Journal of Molecular Biology* 1980, **136:**225-&.
70. George DG, Hunt LT, Yeh LSL, Barker WC: **New Perspectives on Bacterial Ferredoxin Evolution.** *J Mol Evol* 1985, **22:**20-31.
71. Hwang PK, Fletterick RJ: **Convergent and Divergent Evolution of Regulatory Sites in Eukaryotic Phosphorylases.** *Nature* 1986, **324:**80-84.
72. Beintema JJ, Schuller C, Irie M, Carsana A: **Molecular Evolution of the Ribonuclease Superfamily.** *Progress in Biophysics & Molecular Biology* 1988, **51:**165-192.
73. Felsenstein J: **Phylogenies from Molecular Sequences - Inference and Reliability.** *Annual Review of Genetics* 1988, **22:**521-565.

74. Cedergren R, Gray MW, Abel Y, Sankoff D: **The Evolutionary Relationships among Known Life Forms.** *J Mol Evol* 1988, **28:**98-112.
75. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T: **Evolutionary Relationship of Archaebacteria, Eubacteria, and Eukaryotes Inferred from Phylogenetic Trees of Duplicated Genes.** *Proceedings of the National Academy of Sciences of the United States of America* 1989, **86:**9355-9359.
76. Loomis WF, Gilpin ME: **Multigene families and vestigial sequences.** *Proc Natl Acad Sci U S A* 1986, **83:**2143-2147.
77. Reanney DC: **Genetic error and genome design.** *Cold Spring Harb Symp Quant Biol* 1987, **52:**751-757.
78. Ohta T: **Simulating evolution by gene duplication.** *Genetics* 1987, **115:**207-213.
79. Sankoff D, Goldstein M: **Probabilistic models of genome shuffling.** *Bull Math Biol* 1989, **51:**117-124.
80. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al: **Whole-Genome Random Sequencing and Assembly of Haemophilus-Influenzae Rd.** *Science* 1995, **269:**496-512.
81. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, et al: **The Minimal Gene Complement of Mycoplasma-Genitalium.** *Science* 1995, **270:**397-403.
82. Bult CJ, White O, Olsen GJ, Zhou LX, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, et al: **Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii.** *Science* 1996, **273:**1058-1073.
83. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al: **Life with 6000 genes.** *Science* 1996, **274:**546-&.
84. **Genome sequence of the nematode C. elegans: a platform for investigating biology.** *Science* 1998, **282:**2012-2018.
85. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al: **The genome sequence of Drosophila melanogaster.** *Science* 2000, **287:**2185-2195.
86. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
87. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
88. **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:**931-945.
89. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.
90. Mardis ER: **Next-generation DNA sequencing methods.** *Annu Rev Genomics Hum Genet* 2008, **9:**387-402.
91. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, et al: **A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning.** *Genome Res* 2008, **18:**1051-1063.
92. Schuster SC: **Next-generation sequencing transforms today's biology.** *Nat Methods* 2008, **5:**16-18.
93. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.

94. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
95. Thompson JD, Higgins DG, Gibson TJ: **Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
96. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302:**205-217.
97. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Res* 2002, **30:**3059-3066.
98. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.** *Molecular Biology and Evolution* 2001, **18:**691-699.
99. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52:**696-704.
100. Stamatakis A, Ludwig T, Meier H: **RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees.** *Bioinformatics* 2005, **21:**456-463.
101. Yang ZH, Rannala B: **Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo method.** *Molecular Biology and Evolution* 1997, **14:**717-724.
102. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754-755.
103. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling.** *J Mol Biol* 1994, **235:**1501-1531.
104. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
105. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J: **SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny.** *Nucleic Acids Res* 2009, **37:**D380-386.
106. Forslund K, Sonnhammer EL: **Predicting protein function from domain content.** *Bioinformatics* 2008, **24:**1681-1687.
107. Wuichet K, Zhulin IB: **Origins and diversification of a complex signal transduction system in prokaryotes.** *Sci Signal* 2010, **3:**ra50.
108. Caro-Quintero A, Deng J, Auchtung J, Brettar I, Hofle MG, Klappenbach J, Konstantinidis KT: **Unprecedented levels of horizontal gene transfer among spatially co-occurring Shewanella bacteria from the Baltic Sea.** *ISME J* 2011, **5:**131-140.
109. Harrison PW, Lower RP, Kim NK, Young JP: **Introducing the bacterial 'chromid': not a chromosome, not a plasmid.** *Trends Microbiol* 2010, **18:**141-148.
110. Chothia C, Gough J, Vogel C, Teichmann SA: **Evolution of the protein repertoire.** *Science* 2003, **300:**1701-1703.
111. Song N, Joseph JM, Davis GB, Durand D: **Sequence similarity network reveals common ancestry of multidomain proteins.** *PLoS Comput Biol* 2008, **4:**e1000063.
112. Campbell ID, Downing AK: **Building protein structure and function from modular units.** *Trends Biotechnol* 1994, **12:**168-172.
113. Stein L: **Genome annotation: from sequence to biology.** *Nat Rev Genet* 2001, **2:**493-503.
114. Coulson RM, Hall N, Ouzounis CA: **Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum.** *Genome Res* 2004, **14:**1548-1554.
115. Fong JH, Geer LY, Panchenko AR, Bryant SH: **Modeling the evolution of protein domain architectures using maximum parsimony.** *Journal of Molecular Biology* 2007, **366:**307-315.

116. Song N, Joseph JM, Davis GB, Durand D: **Sequence similarity network reveals common ancestry of multidomain proteins.** *PLoS Comput Biol* 2008, **4**.
117. Kanaan SP, Huang C, Wuchty S, Chen DZ, Izaguirre JA: **Inferring protein-protein interactions from multiple protein domain combinations.** *Methods Mol Biol* 2009, **541:**43-59.
118. Richardson JS: **The anatomy and taxonomy of protein structure.** *Adv Protein Chem* 1981, **34:**167-339.
119. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247:**536-540.
120. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R: **Pfam: multiple sequence alignments and HMM-profiles of protein domains.** *Nucleic Acids Res* 1998, **26:**320-322.
121. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34:**D247-251.
122. Nataro JP, Kaper JB: **Diarrheagenic Escherichia coli.** *Clinical Microbiology Reviews* 1998, **11:**142-+.
123. Russo TA, Johnson JR: **Proposal for a new inclusive designation for extraintestinal pathogenic isolates of Escherichia coli: ExPEC.** *Journal of Infectious Diseases* 2000, **181:**1753-1754.
124. Neidhardt FC, Curtiss R: *Escherichia coli and Salmonella : cellular and molecular biology.* 2nd edn. Washington, D.C.: ASM Press; 1996.
125. Wirth T, Falush D, Lan RT, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, Achtman M: **Sex and virulence in Escherichia coli: an evolutionary perspective.** *Molecular Microbiology* 2006, **60:**1136-1151.
126. Croxen MA, Finlay BB: **Molecular mechanisms of Escherichia coli pathogenicity.** *Nature Reviews Microbiology* 2010, **8:**26-38.
127. Mobley HLT, Warren JW: *Urinary tract infections : molecular pathogenesis and clinical management.* Washington, D.C.: ASM Press; 1996.
128. Johnson JR: **Virulence Factors in Escherichia-Coli Urinary-Tract Infection.** *Clinical Microbiology Reviews* 1991, **4:**80-128.
129. Johnson JR, Stell AL: **Extended virulence genotypes of Escherichia coli strains from patients with urosepsis in relation to phylogeny and host compromise.** *Journal of Infectious Diseases* 2000, **181:**261-272.
130. Welch RA, Burland V, Plunkett G, 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, et al: **Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli.** *Proc Natl Acad Sci U S A* 2002, **99:**17020-17024.
131. Bahrani-Mougeot FK, Buckles EL, Lockatell CV, Hebel JR, Johnson DE, Tang CM, Donnenberg MS: **Type 1 fimbriae and extracellular polysaccharides are preeminent uropathogenic Escherichia coli virulence determinants in the murine urinary tract.** *Mol Microbiol* 2002, **45:**1079-1093.
132. Mills M, Meysick KC, O'Brien AD: **Cytotoxic necrotizing factor type 1 of uropathogenic Escherichia coli kills cultured human uroepithelial 5637 cells by an apoptotic mechanism.** *Infect Immun* 2000, **68:**5869-5880.
133. Hofman P, Le Negrate G, Mograbi B, Hofman V, Brest P, Alliana-Schmid A, Flatau G, Boquet P, Rossi B: **Escherichia coli cytotoxic necrotizing factor-1 (CNF-1) increases the adherence to epithelia and the oxidative burst of human polymorphonuclear leukocytes but decreases bacteria phagocytosis.** *J Leukoc Biol* 2000, **68:**522-528.
134. Connell I, Agace W, Klemm P, Schembri M, Marild S, Svanborg C: **Type 1 fimbrial expression enhances Escherichia coli virulence for the urinary tract.** *Proc Natl Acad Sci U S A* 1996, **93:**9827-9832.

135. Gunther NWt, Lockatell V, Johnson DE, Mobley HL: **In vivo dynamics of type 1 fimbria regulation in uropathogenic Escherichia coli during experimental urinary tract infection.** *Infect Immun* 2001, **69:**2838-2846.
136. Lane MC, Alteri CJ, Smith SN, Mobley HLT: **Expression of flagella is coincident with uropathogenic Escherichia coli ascension to the upper urinary tract.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104:**16669-16674.
137. Korhonen TK, Virkola R, Holthofer H: **Localization of binding sites for purified Escherichia coli P fimbriae in the human kidney.** *Infect Immun* 1986, **54:**328-332.
138. Smith YC, Rasmussen SB, Grande KK, Conran RM, O'Brien AD: **Hemolysin of uropathogenic Escherichia coli evokes extensive shedding of the uroepithelium and hemorrhage in bladder tissue within the first 24 hours after intraurethral inoculation of mice.** *Infect Immun* 2008, **76:**2978-2990.
139. Guyer DM, Radulovic S, Jones FE, Mobley HL: **Sat, the secreted autotransporter toxin of uropathogenic Escherichia coli, is a vacuolating cytotoxin for bladder and kidney epithelial cells.** *Infect Immun* 2002, **70:**4539-4546.
140. Lane MC, Lockatell V, Monterosso G, Lamphier D, Weinert J, Hebel JR, Johnson DE, Mobley HL: **Role of motility in the colonization of uropathogenic Escherichia coli in the urinary tract.** *Infect Immun* 2005, **73:**7644-7656.
141. Snyder JA, Haugen BJ, Buckles EL, Lockatell CV, Johnson DE, Donnenberg MS, Welch RA, Mobley HL: **Transcriptome of uropathogenic Escherichia coli during urinary tract infection.** *Infect Immun* 2004, **72:**6373-6381.
142. Lane MC, Lloyd AL, Markyvech TA, Hagan EC, Mobley HL: **Uropathogenic Escherichia coli strains generally lack functional Trg and Tap chemoreceptors found in the majority of E. coli strains strictly residing in the gut.** *J Bacteriol* 2006, **188:**5618-5625.
143. Nara T, Lee L, Imae Y: **Thermosensing Ability of Trg and Tap Chemoreceptors in Escherichia-Coli.** *J Bacteriol* 1991, **173:**1120-1124.
144. Li CY, Boileau AJ, Kung C, Adler J: **Osmotaxis in Escherichia-Coli.** *Proceedings of the National Academy of Sciences of the United States of America* 1988, **85:**9451-9455.
145. Butler SM, Camilli A: **Both chemotaxis and net motility greatly influence the infectivity of Vibrio cholerae.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101:**5018-5023.
146. Millikan DS, Ruby EG: **FlrA, a sigma(54)-dependent transcriptional activator in Vibrio fischeri, is required for motility and symbiotic light-organ colonization.** *J Bacteriol* 2003, **185:**3547-3557.
147. Merritt PM, Danhorn T, Fuqua C: **Motility and chemotaxis in Agrobacterium tumefaciens surface attachment and Biofilm formation.** *J Bacteriol* 2007, **189:**8005-8014.
148. Armitage JP: **Bacterial tactic responses.** *Advances in Microbial Physiology, Vol 41* 1999, **41:**229-289.
149. Grebe TW, Stock J: **Bacterial chemotaxis: The five sensors of a bacterium.** *Current Biology* 1998, **8:**R154-+.
150. Liu X, Parales RE: **Chemotaxis of Escherichia coli to pyrimidines: a new role for the signal transducer tap.** *J Bacteriol* 2008, **190:**972-979.
151. Berg HC: **Bacterial flagellar motor.** *Current Biology* 2008, **18:**R689-R691.
152. Turner L, Ryu WS, Berg HC: **Real-time imaging of fluorescent flagellar filaments.** *J Bacteriol* 2000, **182:**2793-2801.
153. Society for General Microbiology. Symposium (46th : 1990 : University of York), Armitage JP, Lackie JM, Society for General Microbiology., British Society for Cell Biology.: *Biology of the chemotactic response.* Cambridge ; New York: Cambridge University Press; 1990.

154. Sourjik V, Berg HC: **Receptor sensitivity in bacterial chemotaxis.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99:**123-127.
155. Alexander RP, Zhulin IB: **Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104:**2885-2890.
156. Hazelbauer GL, Falke JJ, Parkinson JS: **Bacterial chemoreceptors: high-performance signaling in networked arrays.** *Trends Biochem Sci* 2008, **33:**9-19.
157. Briegel A, Ortega DR, Tocheva EI, Wuichet K, Li Z, Chen SY, Muller A, Iancu CV, Murphy GE, Dobro MJ, et al: **Universal architecture of bacterial chemoreceptor arrays.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106:**17181-17186.
158. Porter SL, Wadhams GH, Armitage JP: **Signal processing in complex chemotaxis pathways.** *Nat Rev Microbiol* 2011, **9:**153-165.
159. Borkovich KA, Simon MI: **The Dynamics of Protein-Phosphorylation in Bacterial Chemotaxis.** *Cell* 1990, **63:**1339-1348.
160. Sarkar MK, Paul K, Blair D: **Chemotaxis signaling protein CheY binds to the rotor protein FliN to control the direction of flagellar rotation in Escherichia coli.** *Proc Natl Acad Sci U S A* 2010, **107:**9370-9375.
161. Bai F, Branch RW, Nicolau DV, Pilizota T, Steel BC, Maini PK, Berry RM: **Conformational Spread as a Mechanism for Cooperativity in the Bacterial Flagellar Switch.** *Science* 2010, **327:**685-689.
162. Silversmith RE: **Auxiliary phosphatases in two-component signal transduction.** *Current Opinion in Microbiology* 2010, **13:**177-183.
163. Vladimirov N, Sourjik V: **Chemotaxis: how bacteria use memory.** *Biological Chemistry* 2009, **390:**1097-1104.
164. Anand GS, Stock AM: **Kinetic basis for the stimulatory effect of phosphorylation on the methylesterase activity of CheB.** *Biochemistry* 2002, **41:**6752-6760.
165. Spaepen S, Vanderleyden J, Okon Y: **Plant Growth-Promoting Actions of Rhizobacteria.** *Plant Innate Immunity* 2009, **51:**283-320.
166. Hartmann A, Schmid M, van Tuinen D, Berg G: **Plant-driven selection of microbes.** *Plant and Soil* 2009, **321:**235-257.
167. Weller DM: **Pseudomonas biocontrol agents of soilborne pathogens: Looking back over 30 years.** *Phytopathology* 2007, **97:**250-256.
168. Beijerinck MW: **Uber ein Spirillum, welches frei en Stickstoff binden kann?** *Zentralblatt fur Bakteriologie, Parasitenkunde, Infektionskrankheiten und Hygiene Abteilung II* 1925, **63:**353-359.
169. Dobereiner JD, L. : **Associative symbiosis in tropical grasses: Characterization of microorganisms and dinitrogen fixing sites.** In *Proceedings First International Symposium on Nitrogen Fixation*. Edited by Newton WEN, C.J. Washington State University Press, Pullman; 1976: 518–538.
170. Tien TM, Gaskins MH, Hubbell DH: **Plant-Growth Substances Produced by Azospirillum-Brasilense and Their Effect on the Growth of Pearl-Millet (Pennisetum-Americanum L).** *Appl Environ Microbiol* 1979, **37:**1016-1024.
171. Hartmann A, Bashan Y: **Ecology and application of Azospirillum and other plant growth-promoting bacteria (PGPB) - Special Issue.** *European Journal of Soil Biology* 2009, **45:**1-2.
172. Diaz-Zorita M, Fernandez-Canigia MV: **Field performance of a liquid formulation of Azospirillum brasilense on dryland wheat productivity.** *European Journal of Soil Biology* 2009, **45:**3-11.
173. Bashan Y, Holguin G: **Azospirillum-plant relationships: Environmental and physiological advances (1990-1996).** *Canadian Journal of Microbiology* 1997, **43:**103-121.
174. Bashan Y, Holguin G, de-Bashan LE: **Azospirillum-plant relationships: physiological, molecular, agricultural, and environmental advances (1997-2003).** *Can J Microbiol* 2004, **50:**521-577.

175. Baldani JI, Baldani VLD: **History on the biological nitrogen fixation research in graminaceous plants: special emphasis on the Brazilian experience.** *Anais Da Academia Brasileira De Ciencias* 2005, **77:**549-579.
176. Kennedy IR, PeregGerk LL, Wood C, Deaker R, Gilchrist K, Katupitiya S: **Biological nitrogen fixation in non-leguminous field crops: Facilitating the evolution of an effective association between Azospirillum and wheat.** *Plant and Soil* 1997, **194:**65-79.
177. Tsavkelova EA, Klimova SY, Cherdyntseva TA, Netrusov AI: **Microbial producers of plant growth Stimulators and their practical use: A review.** *Applied Biochemistry and Microbiology* 2006, **42:**117-126.
178. Dobbelaere S, Croonenborghs A, Thys A, Vande Broek A, Vanderleyden J: **Phytostimulatory effect of Azospirillum brasilense wild type and mutant strains altered in IAA production on wheat.** *Plant and Soil* 1999, **212:**155-164.
179. Bothe H, Korsgen H, Lehmacher T, Hundeshagen B: **Differential-Effects of Azospirillum, Auxin and Combined Nitrogen on the Growth of the Roots of Wheat.** *Symbiosis* 1992, **13:**167-179.
180. Spaepen S, Vanderleyden J, Remans R: **Indole-3-acetic acid in microbial and microorganism-plant signaling.** *FEMS Microbiol Rev* 2007, **31:**425-448.
181. Okon Y, Labanderagonzalez CA: **Agronomic Applications of Azospirillum - an Evaluation of 20 Years Worldwide Field Inoculation.** *Soil Biology & Biochemistry* 1994, **26:**1591-1601.
182. Bashan Y, de-Bashan LE: **How the Plant Growth-Promoting Bacterium Azospirillum Promotes Plant Growth-a Critical Assessment.** *Advances in Agronomy, Vol 108* 2010, **108:**77-136.
183. Levanony H, Bashan Y: **Enhancement of Cell-Division in Wheat Root-Tips and Growth of Root Elongation Zone Induced by Azospirillum-Brasilense Cd.** *Canadian Journal of Botany-Revue Canadienne De Botanique* 1989, **67:**2213-2216.
184. Creus CM, Graziano M, Casanovas EM, Pereyra MA, Simontacchi M, Puntarulo S, Barassi CA, Lamattina L: **Nitric oxide is involved in the Azospirillum brasilense-induced lateral root formation in tomato.** *Planta* 2005, **221:**297-303.
185. Molina-Favero C, Creus CM, Simontacchi M, Puntarulo S, Lamattina L: **Aerobic nitric oxide production by Azospirillum brasilense Sp245 and its influence on root architecture in tomato.** *Molecular Plant-Microbe Interactions* 2008, **21:**1001-1009.
186. Jain DK, Patriquin DG: **Root Hair Deformation, Bacterial Attachment, and Plant-Growth in Wheat-Azospirillum Associations.** *Appl Environ Microbiol* 1984, **48:**1208-1213.
187. Okon Y, Kapulnik Y: **Development and Function of Azospirillum-Inoculated Roots.** *Plant and Soil* 1986, **90:**3-16.
188. Hadas R, Okon Y: **Effect of <i>Azospirillum brasilense inoculation on root morphology and respiration in tomato seedlings.** *Biology and Fertility of Soils* 1987, **5:**241-247.
189. Dobbelaere S, Vanderleyden J, Okon Y: **Plant growth-promoting effects of diazotrophs in the rhizosphere.** *Critical Reviews in Plant Sciences* 2003, **22:**107-149.
190. Bashan Y, Harrison SK, Whitmoyer RE: **Enhanced Growth of Wheat and Soybean Plants Inoculated with Azospirillum-Brasilense Is Not Necessarily Due to General Enhancement of Mineral Uptake.** *Appl Environ Microbiol* 1990, **56:**769-775.
191. Bacilio M, Vazquez P, Bashan Y: **Alleviation of noxious effects of cattle ranch composts on wheat seed germination by inoculation with Azospirillum spp.** *Biology and Fertility of Soils* 2003, **38:**261-266.
192. Rodriguez H, Gonzalez T, Goire I, Bashan Y: **Gluconic acid production and phosphate solubilization by the plant growth-promoting bacterium Azospirillum spp.** *Naturwissenschaften* 2004, **91:**552-555.

193. Bacilio M, Rodriguez H, Moreno M, Hernandez JP, Bashan Y: **Mitigation of salt stress in wheat seedlings by a gfp-tagged Azospirillum lipoferum.** *Biology and Fertility of Soils* 2004, **40:**188-193.
194. Bashan Y, Dubrovsky JG: **Azospirillum spp participation in dry matter partitioning in grasses at the whole plant level.** *Biology and Fertility of Soils* 1996, **23:**435-440.
195. Caballero-Mellado J, Lopez-Reyes L, Bustillos-Cristales R: **Presence of 16S rRNA genes in multiple replicons in Azospirillum brasilense.** *Fems Microbiology Letters* 1999, **178:**283-288.
196. Martin-Didonet CC, Chubatsu LS, Souza EM, Kleina M, Rego FG, Rigo LU, Yates MG, Pedrosa FO: **Genome structure of the genus Azospirillum.** *J Bacteriol* 2000, **182:**4113-4116.
197. Ulrich LE, Koonin EV, Zhulin IB: **One-component systems dominate signal transduction in prokaryotes.** *Trends in Microbiology* 2005, **13:**52-56.
198. Citri A, Yarden Y: **EGF-ERBB signalling: towards the systems level.** *Nature Reviews Molecular Cell Biology* 2006, **7:**505-516.
199. Zhulin IB, Taylor BL: **PAS domain S-boxes in Archaea, bacteria and sensors for oxygen and redox.** *Trends Biochem Sci* 1997, **22:**331-333.
200. Ponting CP, Aravind L: **PAS: a multifunctional domain family comes to light.** *Current Biology* 1997, **7:**R674-677.
201. Taylor BL, Zhulin IB: **PAS domains: Internal sensors of oxygen, redox potential, and light.** *Microbiology and Molecular Biology Reviews* 1999, **63:**479-+.
202. Aravind L, Ponting CP: **The GAF domain: an evolutionary link between diverse phototransducing proteins.** *Trends Biochem Sci* 1997, **22:**458-459.
203. Anantharaman V, Aravind L: **Cache - a signaling domain common to animal Ca2+ channel subunits and a class of prokaryotic chemotaxis receptors.** *Trends Biochem Sci* 2000, **25:**535-537.
204. Anantharaman V, Aravind L: **The CHASE domain: a predicted ligand-binding module in plant cytokinin receptors and other eukaryotic and bacterial receptors.** *Trends Biochem Sci* 2001, **26:**579-582.
205. Mougel C, Zhulin IB: **CHASE: an extracellular sensing domain common to transmembrane receptors from prokaryotes, lower eukaryotes and plants.** *Trends Biochem Sci* 2001, **26:**582-584.
206. Galperin MY, Gaidenko TA, Mulkidjanian AY, Nakano M, Price CW: **MHYT, a new integral membrane sensor domain.** *Fems Microbiology Letters* 2001, **205:**17-23.
207. Shu CJ, Ulrich LE, Zhulin IB: **The NIT domain: a predicted nitrate-responsive module in bacterial sensory receptors.** *Trends Biochem Sci* 2003, **28:**121-124.
208. Zhulin IB, Nikolskaya AN, Galperin MY: **Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in Bacteria and Archaea.** *J Bacteriol* 2003, **185:**285-294.
209. Ulrich LE, Zhulin IB: **Four-helix bundle: a ubiquitous sensory module in prokaryotic signal transduction.** *Bioinformatics* 2005, **21:**45-48.
210. Ulrich LE, Zhulin IB: **MiST: a microbial signal transduction database.** *Nucleic Acids Res* 2007, **35:**D386-D390.
211. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34:**D247-D251.
212. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P: **SMART 5: domains in the context of genomes and networks.** *Nucleic Acids Res* 2006, **34:**D257-D260.

213. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25:**4876-4882.
214. McGuffin LJ, Bryson K, Jones DT: **The PSIPRED protein structure prediction server.** *Bioinformatics* 2000, **16:**404-405.
215. Kall L, Krogh A, Sonnhammer ELL: **A combined transmembrane topology and signal peptide prediction method.** *Journal of Molecular Biology* 2004, **338:**1027-1036.
216. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96:**2896-2901.
217. Hershko A, Ciechanover A: **The ubiquitin system.** *Annu Rev Biochem* 1998, **67:**425-479.
218. Parry G, Estelle M: **Auxin receptors: a new role for F-box proteins.** *Current Opinion in Cell Biology* 2006, **18:**152-156.
219. Lane MC, Lockatell V, Monterosso G, Lamphier D, Weinert J, Hebel JR, Johnson DE, Mobley HL: **Role of motility in the colonization of uropathogenic Escherichia coli in the urinary tract.** *Infection and Immunity* 2005, **73:**7644-7656.
220. Giron JA, Torres AG, Freer E, Kaper JB: **The flagella of enteropathogenic Escherichia coli mediate adherence to epithelial cells.** *Mol Microbiol* 2002, **44:**361-379.
221. Macnab RM: **Flagella and motility.** In *Escherichia coli and Salmonella : cellular and molecular biology.* Edited by Neidhardt FC, Curtiss R. Washington, D.C.: ASM Press; 1996: 123-145
222. Park HK, Shim SS, Kim SY, Park JH, Park SE, Kim HJ, Kang BC, Kim CM: **Molecular analysis of colonized bacteria in a human newborn infant gut.** *J Microbiol* 2005, **43:**345-353.
223. Bettelheim KA, Breadon A, Faiers MC, O'Farrell SM, Shooter RA: **The origin of O serotypes of Escherichia coli in babies after normal delivery.** *J Hyg (Lond)* 1974, **72:**67-70.
224. Kaper JB, Nataro JP, Mobley HL: **Pathogenic Escherichia coli.** *Nat Rev Microbiol* 2004, **2:**123-140.
225. Russo TA, Johnson JR: **Proposal for a new inclusive designation for extraintestinal pathogenic isolates of Escherichia coli: ExPEC.** *J Infect Dis* 2000, **181:**1753-1754.
226. Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, Guigon G, Carbonnelle E, Lortholary O, Clermont O, Denamur E, et al: **Phylogenetic and genomic diversity of human bacteremic Escherichia coli strains.** *BMC Genomics* 2008, **9:**560.
227. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M: **Sex and virulence in Escherichia coli: an evolutionary perspective.** *Mol Microbiol* 2006, **60:**1136-1151.
228. Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N, Bingen E, Elion J, Denamur E: **The link between phylogeny and virulence in Escherichia coli extraintestinal infection.** *Infection and Immunity* 1999, **67:**546-553.
229. Chaudhuri RR, Henderson IR: **The evolution of the Escherichia coli phylogeny.** *Infect Genet Evol* 2012, **12:**214-226.
230. Huys G, Cnockaert M, Janda JM, Swings J: **Escherichia albertii sp nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children.** *International Journal of Systematic and Evolutionary Microbiology* 2003, **53:**807-810.
231. Lawrence JG, Ochman H, Hartl DL: **Molecular and Evolutionary Relationships among Enteric Bacteria.** *Journal of General Microbiology* 1991, **137:**1911-1921.
232. Ochman H, Selander RK: **Standard Reference Strains of Escherichia-Coli from Natural-Populations.** *J Bacteriol* 1984, **157:**690-693.
233. Escobar-Paramo P, Clermont O, Blanc-Potard AB, Bui H, Le Bouguenec C, Denamur E: **A specific genetic background is required for acquisition and expression of virulence factors in Escherichia coli.** *Molecular Biology and Evolution* 2004, **21:**1085-1094.

234. Grebe TW, Stock J: **Bacterial chemotaxis: the five sensors of a bacterium.** *Curr Biol* 1998, **8:**R154-157.
235. Englert DL, Adase CA, Jayaraman A, Manson MD: **Repellent Taxis in Response to Nickel Ion Requires neither Ni2+ Transport nor the Periplasmic NikA Binding Protein.** *J Bacteriol* 2010, **192:**2633-2637.
236. Liu XX, Parales RE: **Chemotaxis of Escherichia coli to pyrimidines: a new role for the signal transducer tap.** *J Bacteriol* 2008, **190:**972-979.
237. Kalir S, McClure J, Pabbaraju K, Southward C, Ronen M, Leibler S, Surette MG, Alon U: **Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria.** *Science* 2001, **292:**2080-2083.
238. Schwan WR: **Flagella allow uropathogenic Escherichia coli ascension into murine kidneys.** *International Journal of Medical Microbiology* 2008, **298:**441-447.
239. Wright KJ, Seed PC, Hultgren SJ: **Uropathogenic Escherichia coli flagella aid in efficient urinary tract colonization.** *Infection and Immunity* 2005, **73:**7657-7668.
240. Stamm WE, Norrby SR: **Urinary tract infections: disease panorama and challenges.** *J Infect Dis* 2001, **183 Suppl 1:**S1-4.
241. Angiuoli SV, Salzberg SL: **Mugsy: fast multiple alignment of closely related whole genomes.** *Bioinformatics* 2011, **27:**334-342.
242. Bingen E, Picard B, Brahimi N, Mathy S, Desjardins P, Elion J, Denamur E: **Phylogenetic analysis of Escherichia coli strains causing neonatal meningitis suggests horizontal gene transfer from a predominant pool of highly virulent B2 group strains.** *J Infect Dis* 1998, **177:**642-650.
243. Boyd EF, Hartl DL: **Chromosomal regions specific to pathogenic isolates of Escherichia coli have a phylogenetically clustered distribution.** *J Bacteriol* 1998, **180:**1159-1165.
244. Miquel S, Peyretaillade E, Claret L, de Vallee A, Dossat C, Vacherie B, Zineb el H, Segurens B, Barbe V, Sauvanet P, et al: **Complete genome sequence of Crohn's disease-associated adherent-invasive E. coli strain LF82.** *PLoS One* 2010, **5**.
245. Pupo GM, Lan RT, Reeves PR: **Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics.** *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97:**10567-10572.
246. Giron JA: **Expression of flagella and motility by Shigella.** *Mol Microbiol* 1995, **18:**63-75.
247. Tominaga A, Lan R, Reeves PR: **Evolutionary changes of the flhDC flagellar master operon in Shigella strains.** *J Bacteriol* 2005, **187:**4295-4302.
248. Mizuno T, Imae Y: **Conditional Inversion of the Thermoresponse in Escherichia-Coli.** *J Bacteriol* 1984, **159:**360-367.
249. Kihara M, Macnab RM: **Cytoplasmic Ph Mediates Ph Taxis and Weak-Acid Repellent Taxis of Bacteria.** *J Bacteriol* 1981, **145:**1209-1221.
250. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ: **Jalview Version 2-a multiple sequence alignment editor and analysis workbench.** *Bioinformatics* 2009, **25:**1189-1191.
251. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program.** *Briefings in Bioinformatics* 2008, **9:**286-298.
252. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0.** *Molecular Biology and Evolution* 2007, **24:**1596-1599.
253. Guindon S, Dufayard JF, Hordijk W, Lefort V, Gascuel O: **PhyML: Fast and Accurate Phylogeny Reconstruction by Maximum Likelihood.** *Infection Genetics and Evolution* 2009, **9:**384-385.
254. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST plus : architecture and applications.** *Bmc Bioinformatics* 2009, **10**.
255. Mojzsis SJ, Arrhenius G, McKeegan KD, Harrison TM, Nutman AP, Friend CR: **Evidence for life on Earth before 3,800 million years ago.** *Nature* 1996, **384:**55-59.

256. Watanabe Y, Martini JE, Ohmoto H: **Geochemical evidence for terrestrial ecosystems 2.6 billion years ago.** *Nature* 2000, **408:**574-578.
257. Battistuzzi FU, Hedges SB: **A major clade of prokaryotes with ancient adaptations to life on land.** *Mol Biol Evol* 2009, **26:**335-343.
258. Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, et al: **Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*.** *PLoS Genet* 2007, **3:**e231.
259. Okon Y, Labandera-Gonzalez CA: **Agronomic applications of Azospirillum: An evaluation of 20 years worldwide field inoculation.** *Soil Biology & Biochemistry* 1994, **26:**1591-1601.
260. Steenhoudt O, Vanderleyden J: **Azospirillum, a free-living nitrogen-fixing bacterium closely associated with grasses: genetic, biochemical and ecological aspects.** *FEMS Microbiol Rev* 2000, **24:**487-506.
261. Kaneko T, Minamisawa K, Isawa T, Nakatsukasa H, Mitsui H, Kawaharada Y, Nakamura Y, Watanabe A, Kawashima K, Ono A, et al: **Complete genomic structure of the cultivated rice endophyte Azospirillum sp. B510.** *DNA Res* 2010, **17:**37-50.
262. Magalhaes FM, Baldani JI, Souto SM, Kuykendall JR, Doebereiner J: **A new acid-tolerant Azospirillum species.** *An Acad Bras Cienc* 1983, **55:**417-430.
263. Tarrand JJ, Krieg NR, Dobereiner J: **A taxonomic study of the Spirillum lipoferum group, with descriptions of a new genus, Azospirillum gen. nov. and two species, Azospirillum lipoferum (Beijerinck) comb. nov. and Azospirillum brasilense sp. nov.** *Can J Microbiol* 1978, **24:**967-980.
264. Mehnaz S, Weselowski B, Lazarovits G: **Azospirillum canadense sp. nov., a nitrogen-fixing bacterium isolated from corn rhizosphere.** *International Journal of Systematic Bacteriology* 2007, **57:**620-624.
265. Eckert B, Weber OB, Kirchhof G, Halbritter A, Stoffels M, Hartmann A: **Azospirillum doebereinerae sp. nov., a nitrogen-fixing bacterium associated with the C4-grass Miscanthus.** *International Journal of Systematic Bacteriology* 2001, **51:**17-26.
266. Hurek T, Reinhold B, Fendrik I, Niemann EG: **Root-zone-specific oxygen tolerance of Azospirillum spp. and diazotrophic rods closely associated with Kallar grass.** *Appl Environ Microbiol* 1987, **53:**163-169.
267. Khammas KM, Ageron E, Grimont PA, Kaiser P: **Azospirillum irakense sp. nov., a nitrogen-fixing bacterium associated with rice roots and rhizosphere soil.** *Res Microbiol* 1989, **140:**679-693.
268. Skerman VBD, Sly LI, Williamson ML: **Conglomeromonas largomobilis gen. nov., sp. nov., a sodium-sensitive, mixed-flagellated organism from fresh waters.** *International Journal of Systematic Bacteriology* 1983, **33:**300-308.
269. Dekhil SB, Cahill M, Stackebrandt E, Sly LI: **Transfer of Conglomeromonas largomobilis subsp largomobilis to the genus Azospirillum as Azopirillum largomobile comb nov, and elevation of Conglomeromonas largomobilis subsp parooensis to the new type species of Conglomeromonas, Conglomeromonas parooensis sp nov.** *Systematic and Applied Microbiology* 1997, **20:**72-77.
270. Bally R, Thomas-Bauzon D, Heulin T, Balandreau J, Richard C, Ley JD: **Determination of the most frequent N2-fixing bacteria in a rice rhizosphere.** *Canadian Journal of Microbiology* 1983, **29:**881-887.
271. Peng G, Wang H, Zhang G, Hou W, Liu Y, Wang ET, Tan Z: **Azospirillum melinis sp. nov., a group of diazotrophs isolated from tropical molasses grass.** *International Journal of Systematic Bacteriology* 2006, **56:**1263-1271.
272. Xie CH, Yokota A: **Azospirillum oryzae sp. nov., a nitrogen-fixing bacterium isolated from the roots of the rice plant Oryza sativa.** *International Journal of Systematic Bacteriology* 2005, **55:**1435-1438.

273. Zhou Y, Wei W, Wang X, Xu L, Lai R: **Azospirillum palatum sp. nov., isolated from forest soil in Zhejiang province, China.** *J Gen Appl Microbiol* 2009, **55:**1-7.
274. Lin SY, Young CC, Hupfer H, Siering C, Arun AB, Chen WM, Lai WA, Shen FT, Rekha PD, Yassin AF: **Azospirillum picis sp. nov., isolated from discarded tar.** *International Journal of Systematic Bacteriology* 2009, **59:**761-765.
275. Young CC, Hupfer H, Siering C, Ho MJ, Arun AB, Lai WA, Rekha PD, Shen FT, Hung MH, Chen WM, Yassin AF: **Azospirillum rugosum sp. nov., isolated from oil-contaminated soil.** *International Journal of Systematic Bacteriology* 2008, **58:**959-963.
276. Elbeltagy A, Nishioka K, Sato T, Suzuki H, Ye B, Hamada T, Isawa T, Mitsui H, Minamisawa K: **Endophytic colonization and in planta nitrogen fixation by a Herbaspirillum sp. isolated from wild rice species.** *Appl Environ Microbiol* 2001, **67:**5285-5293.
277. Mehnaz S, Weselowski B, Lazarovits G: **Azospirillum zeae sp. nov., a diazotrophic bacterium isolated from rhizosphere soil of Zea mays.** *International Journal of Systematic Bacteriology* 2007, **57:**2805-2809.
278. Yoon JH, Kang SJ, Park S, Oh TK: **Caenispirillum bisanense gen. nov., sp. nov., isolated from sludge of a dye works.** *International Journal of Systematic Bacteriology* 2007, **57:**1217-1221.
279. Bardiya N, Bae JH: **Isolation and characterization of Dechlorospirillum anomalous strain JB116 from a sewage treatment plant.** *Microbiol Res* 2008, **163:**182-191.
280. Maszenan AM, Seviour RJ, Patel BK, Janssen PH, Wanner J: **Defluvicoccus vanus gen. nov., sp. nov., a novel Gram-negative coccus/coccobacillus in the 'Alphaproteobacteria' from activated sludge.** *International Journal of Systematic Bacteriology* 2005, **55:**2105-2111.
281. Wang YX, Liu JH, Zhang XX, Chen YG, Wang ZG, Chen Y, Li QY, Peng Q, Cui XL: **Fodinicurvata sediminis gen. nov., sp. nov. and Fodinicurvata fenggangensis sp. nov., poly-beta-hydroxybutyrate-producing bacteria in the family Rhodospirillaceae.** *International Journal of Systematic Bacteriology* 2009, **59:**2575-2581.
282. Jung HM, Lee JS, Bae HM, Yi TH, Kim SY, Lee ST, Im WT: **Inquilinus ginsengisoli sp. nov., isolated from soil of a ginseng field.** *International Journal of Systematic Bacteriology* 2011, **61:**201-204.
283. Wellinghausen N, Essig A, Sommerburg O: **Inquilinus limosus in patients with cystic fibrosis, Germany.** *Emerg Infect Dis* 2005, **11:**457-459.
284. Yoon JH, Kang SJ, Park S, Lee SY, Oh TK: **Reclassification of Aquaspirillum itersonii and Aquaspirillum peregrinum as Novispirillum itersonii gen. nov., comb. nov. and Insolitispirillum peregrinum gen. nov., comb. nov.** *International Journal of Systematic Bacteriology* 2007, **57:**2830-2835.
285. Thrash JC, Ahmadi S, Torok T, Coates JD: **Magnetospirillum bellicus sp. nov., a novel dissimilatory perchlorate-reducing alphaproteobacterium isolated from a bioelectrical reactor.** *Appl Environ Microbiol* 2010, **76:**4730-4737.
286. Geelhoed JS, Sorokin DY, Epping E, Tourova TP, Banciu HL, Muyzer G, Stams AJ, van Loosdrecht MC: **Microbial sulfide oxidation in the oxic-anoxic transition zone of freshwater sediment: involvement of lithoautotrophic Magnetospirillum strain J10.** *FEMS Microbiol Ecol* 2009, **70:**54-65.
287. Matsunaga T, Okamura Y, Fukuda Y, Wahyudi AT, Murase Y, Takeyama H: **Complete genome sequence of the facultative anaerobic magnetotactic bacterium Magnetospirillum sp. strain AMB-1.** *DNA Res* 2005, **12:**157-166.
288. Schleifer KH, Schuler D, Spring S, Weizenegger M, Amann R, Ludwig W, Kohler M: **The Genus Magnetospirillum Gen-Nov - Description of Magnetospirillum-Gryphiswaldense Sp-Nov and Transfer of Aquaspirillum-Magnetotacticum to Magnetospirillum-Magnetotacticum Comb-Nov.** *Systematic and Applied Microbiology* 1991, **14:**379-385.

289. Lai Q, Yuan J, Gu L, Shao Z: **Marispirillum indicum gen. nov., sp. nov., isolated from a deep-sea environment.** *International Journal of Systematic Bacteriology* 2009, **59:**1278-1281.
290. Urios L, Michotey V, Intertaglia L, Lesongeur F, Lebaron P: **Nisaea denitrificans gen. nov., sp. nov. and Nisaea nitritireducens sp. nov., two novel members of the class Alphaproteobacteria from the Mediterranean Sea.** *International Journal of Systematic Bacteriology* 2008, **58:**2336-2341.
291. Riemann L, Leitet C, Pommier T, Simu K, Holmfeldt K, Larsson U, Hagstrom A: **The native bacterioplankton community in the central baltic sea is influenced by freshwater bacterial species.** *Appl Environ Microbiol* 2008, **74:**503-515.
292. Lai Q, Yuan J, Wu C, Shao Z: **Oceanibaculum indicum gen. nov., sp. nov., isolated from deep seawater of the Indian Ocean.** *International Journal of Systematic Bacteriology* 2009, **59:**1733-1737.
293. Dong C, Lai Q, Chen L, Sun F, Shao Z, Yu Z: **Oceanibaculum pacificum sp. nov., isolated from hydrothermal field sediment of the south-west Pacific Ocean.** *International Journal of Systematic Bacteriology* 2010, **60:**219-222.
294. Choi DH, Hwang CY, Cho BC: **Pelagibius litoralis gen. nov., sp. nov., a marine bacterium in the family Rhodospirillaceae isolated from coastal seawater.** *International Journal of Systematic Bacteriology* 2009, **59:**818-823.
295. Anil Kumar P, Srinivas TN, Takaichi S, Maoka T, Sasikala C, Ramana Ch V: **Phaeospirillum chandramohanii sp. nov., a phototrophic alphaproteobacterium with carotenoid glycosides.** *International Journal of Systematic Bacteriology* 2009, **59:**2089-2093.
296. Imhoff JF, Petri R, Suling J: **Reclassification of species of the spiral-shaped phototrophic purple non-sulfur bacteria of the a-Proteobacteria: description of the new genera Phaeospirillum gen. nov., Rhodovibrio gen. nov., Rhodothalassium gen. nov. and Roseospira gen. nov. as well as transfer of Rhodospirillum fulvum to Phaeospirillum fulvum comb. nov., of Rhodospirillum molischianum to Phaeospirillum molischianum comb. nov., of Rhodospirillum salinarum to Rhodovibrio salexigens.** *Int J Syst Bacteriol* 1998, **48 Pt 3:**793-798.
297. Lakshmi KV, Sasikala C, Ashok GV, Chandrasekaran R, Ramana CV: **Phaeovibrio sulfidiphilus gen. nov., sp. nov., phototrophic alphaproteobacterium isolated from brackish water.** *International Journal of Systematic Bacteriology* 2010.
298. Zhang D, Yang H, Zhang W, Huang Z, Liu SJ: **Rhodocista pekingensis sp. nov., a cyst-forming phototrophic bacterium from a municipal wastewater treatment plant.** *International Journal of Systematic Bacteriology* 2003, **53:**1111-1114.
299. Winkelmann G, Schmidtkunz K, Rainey FA: **Characterization of a novel Spirillum-like bacterium that degrades ferrioxamine-type siderophores.** *Biometals* 1996, **9:**78-83.
300. Pfennig N, Lunsdorf H, Suling J, Imhoff JF: **Rhodospira trueperi gen. nov., spec. nov., a new phototrophic Proteobacterium of the alpha group.** *Arch Microbiol* 1997, **168:**39-45.
301. Favinger J, Stadtwald R, Gest H: **Rhodospirillum centenum, sp. nov., a thermotolerant cyst-forming anoxygenic photosynthetic bacterium.** *Antonie Van Leeuwenhoek* 1989, **55:**291-296.
302. Skerman VBD, Mcgowan V, Sneath PHA: **Approved lists of bacterial names.** *International Journal of Systematic Bacteriology* 1980, **30:**225-420.
303. Reslewic S, Zhou S, Place M, Zhang Y, Briska A, Goldstein S, Churas C, Runnheim R, Forrest D, Lim A, et al: **Whole-genome shotgun optical mapping of Rhodospirillum rubrum.** *Appl Environ Microbiol* 2005, **71:**5511-5522.
304. Anil Kumar P, Aparna P, Srinivas TN, Sasikala C, Ramana Ch V: **Rhodospirillum sulfurexigens sp. nov., a phototrophic alphaproteobacterium requiring a reduced sulfur source for growth.** *International Journal of Systematic Bacteriology* 2008, **58:**2917-2920.

305. Kalyan Chakravarthy S, Srinivas TN, Anil Kumar P, Sasikala C, Ramana Ch V: **Roseospira visakhapatnamensis sp. nov. and Roseospira goensis sp. nov.** *International Journal of Systematic Bacteriology* 2007, **57:**2453-2457.
306. Guyoneaud R, Moune S, Eatock C, Bothorel V, Hirschler-Rea A, Willison J, Duran R, Liesack W, Herbert R, Matheron R, Caumette P: **Characterization of three spiral-shaped purple nonsulfur bacteria isolated from coastal lagoon sediments, saline sulfur springs, and microbial mats: emended description of the genus Roseospira and description of Roseospira marina sp. nov., Roseospira navarrensis sp. nov., and Roseospira thiosulfatophila sp. nov.** *Arch Microbiol* 2002, **178:**315-324.
307. Weon HY, Kim BY, Hong SB, Joa JH, Nam SS, Lee KH, Kwon SW: **Skermanella aerolata sp. nov., isolated from air, and emended description of the genus Skermanella.** *International Journal of Systematic Bacteriology* 2007, **57:**1539-1542.
308. Sly LI, Stackebrandt E: **Description of Skermanella parooensis gen. nov., sp. nov. to accommodate Conglomeromonas largomobilis subsp. parooensis following the transfer of Conglomeromonas largomobilis subsp. largomobilis to the genus Azospirillum.** *International Journal of Systematic Bacteriology* 1999, **49:**541-544.
309. An H, Zhang L, Tang Y, Luo X, Sun T, Li Y, Wang Y, Dai J, Fang C: **Skermanella xinjiangensis sp. nov., isolated from the desert of Xinjiang, China.** *International Journal of Systematic Bacteriology* 2009, **59:**1531-1534.
310. Sizova MV, Panikov NS, Spiridonova EM, Slobodova NV, Tourova TP: **Novel facultative anaerobic acidotolerant Telmatospirillum siberiense gen. nov. sp. nov. isolated from mesotrophic fen.** *Systematic and Applied Microbiology* 2007, **30:**213-220.
311. Zhang GI, Hwang CY, Cho BC: **Thalassobaculum litoreum gen. nov., sp. nov., a member of the family Rhodospirillaceae isolated from coastal seawater.** *International Journal of Systematic Bacteriology* 2008, **58:**479-485.
312. Urios L, Michotey V, Intertaglia L, Lesongeur F, Lebaron P: **Thalassobaculum salexigens sp. nov., a new member of the family Rhodospirillaceae from the NW Mediterranean Sea, and emended description of the genus Thalassobaculum.** *International Journal of Systematic Bacteriology* 2010, **60:**209-213.
313. Lopez-Lopez A, Pujalte MJ, Benlloch S, Mata-Roig M, Rossello-Mora R, Garay E, Rodriguez-Valera F: **Thalassospira lucentensis gen. nov., sp. nov., a new marine member of the a-Proteobacteria.** *International Journal of Systematic Bacteriology* 2002, **52:**1277-1283.
314. Cui Z, Shao Z: **Predominant strains of polycyclic aromatic hydrocarbon-degrading consortia from deep sea of the Middle Atlantic Ridge.** *Wei Sheng Wu Xue Bao* 2009, **49:**902-909.
315. Liu C, Wu Y, Li L, Ma Y, Shao Z: **Thalassospira xiamenensis sp. nov. and Thalassospira profundimaris sp. nov.** *International Journal of Systematic Bacteriology* 2007, **57:**316-320.
316. Kodama Y, Stiknowati LI, Ueki A, Ueki K, Watanabe K: **Thalassospira tepidiphila sp. nov., a polycyclic aromatic hydrocarbon-degrading bacterium isolated from seawater.** *International Journal of Systematic Bacteriology* 2008, **58:**711-715.
317. Zhao B, Wang H, Li R, Mao X: **Thalassospira xianhensis sp. nov., a polycyclic aromatic hydrocarbon-degrading marine bacterium.** *International Journal of Systematic Bacteriology* 2010, **60:**1125-1129.
318. Shi BH, Arunpairojana V, Palakawong S, Yokota A: **Tistrella mobilis gen nov, sp nov, a novel polyhydroxyalkanoate-producing bacterium belonging to a-Proteobacteria.** *J Gen Appl Microbiol* 2002, **48:**335-343.
319. Gonzalez V, Santamaria RI, Bustos P, Hernandez-Gonzalez I, Medrano-Soto A, Moreno-Hagelsieb G, Janga SC, Ramirez MA, Jimenez-Jacinto V, Collado-Vides J, Davila G: **The partitioned**

Rhizobium etli genome: genetic and metabolic redundancy in seven interacting replicons. *Proc Natl Acad Sci U S A* 2006, **103:**3834-3839.
320. Vial L, Lavire C, Mavingui P, Blaha D, Haurat J, Moenne-Loccoz Y, Bally R, Wisniewski-Dye F: **Phase variation and genomic architecture changes in Azospirillum.** *J Bacteriol* 2006, **188:**5364-5373.
321. Koonin EV, Makarova KS, Aravind L: **Horizontal gene transfer in prokaryotes: quantification and classification.** *Annu Rev Microbiol* 2001, **55:**709-742.
322. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29:**22-28.
323. Dennis PG, Miller AJ, Hirsch PR: **Are root exudates more important than other sources of rhizodeposits in structuring rhizosphere bacterial communities?** *FEMS Microbiol Ecol* 2010, **72:**313-327.
324. Boyer M, Haurat J, Samain S, Segurens B, Gavory F, Gonzalez V, Mavingui P, Rohr R, Bally R, Wisniewski-Dye F: **Bacteriophage prevalence in the genus *Azospirillum* and analysis of the first genome sequence of an *Azospirillum brasilense* integrative phage.** *Appl Environ Microbiol* 2008, **74:**861-874.
325. Giraud E, Moulin L, Vallenet D, Barbe V, Cytryn E, Avarre JC, Jaubert M, Simon D, Cartieaux F, Prin Y, et al: **Legumes symbioses: absence of Nod genes in photosynthetic bradyrhizobia.** *Science* 2007, **316:**1307-1312.
326. Kuo CH, Ochman H: **Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria.** *Biol Direct* 2009, **4:**35.
327. Kenrick P, Crane PR: **The origin and early evolution of plants on land.** *Nature* 1997, **389:**33-39.
328. Raven JA, Edwards D: **Roots: evolutionary origins and biogeochemical significance.** *J Exp Bot* 2001, **52:**381-401.
329. Prasad V, Stromberg CA, Alimohammadian H, Sahni A: **Dinosaur coprolites and the early evolution of grasses and grazers.** *Science* 2005, **310:**1177-1180.
330. Jiang ZY, Bauer CE: **Analysis of a chemotaxis operon from Rhodospirillum centenum.** *J Bacteriol* 1997, **179:**5712-5719.
331. Xie Z, Ulrich LE, Zhulin IB, Alexandre G: **PAS domain containing chemoreceptor couples dynamic changes in metabolism with chemotaxis.** *Proc Natl Acad Sci U S A* 2010, **107:**2235-2240.
332. Bible AN, Stephens BB, Ortega DR, Xie Z, Alexandre G: **Function of a chemotaxis-like signal transduction pathway in modulating motility, cell clumping, and cell length in the alphaproteobacterium Azospirillum brasilense.** *J Bacteriol* 2008, **190:**6365-6375.
333. Ulrich LE, Zhulin IB: **The MiST2 database: a comprehensive genomics resource on microbial signal transduction.** *Nucleic Acids Res* 2010, **38:**D401-407.
334. Buchan A, Crombie B, Alexandre GM: **Temporal dynamics and genetic diversity of chemotactic-competent microbial populations in the rhizosphere.** *Environ Microbiol* 2010, **12:**3171-3184.
335. Berleman JE, Bauer CE: **Involvement of a Che-like signal transduction cascade in regulating cyst cell development in Rhodospirillum centenum.** *Mol Microbiol* 2005, **56:**1457-1466.
336. Berleman JE, Bauer CE: **A che-like signal transduction cascade involved in controlling flagella biosynthesis in Rhodospirillum centenum.** *Mol Microbiol* 2005, **55:**1390-1402.
337. Assmus B, Hutzler P, Kirchhof G, Amann R, Lawrence JR, Hartmann A: **In situ localization of Azospirillum brasilense in the rhizosphere of wheat with fluorescently labeled, rRNA-targeted oligonucleotide probes and scanning confocal laser microscopy.** *Appl Environ Microbiol* 1995, **61:**1013-1019.

338. Pedrosa FO, Monteiro RA, Wassem R, Cruz LM, Ayub RA, Colauto NB, Fernandez MA, Fungaro MH, Grisard EC, Hungria M, et al: **Genome of Herbaspirillum seropedicae strain SmR1, a specialized diazotrophic endophyte of tropical grasses.** *PLoS Genet* 2011, **7:**e1002064.
339. Dorr J, Hurek T, Reinhold-Hurek B: **Type IV pili are involved in plant-microbe and fungus-microbe interactions.** *Molecular Microbiology* 1998, **30:**7-17.
340. Ramey BE, Koutsoudis M, von Bodman SB, Fuqua C: **Biofilm formation in plant-microbe associations.** *Curr Opin Microbiol* 2004, **7:**602-609.
341. Tomich M, Planet PJ, Figurski DH: **The tad locus: postcards from the widespread colonization island.** *Nat Rev Microbiol* 2007, **5:**363-375.
342. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics.** *Nucleic Acids Res* 2009, **37:**D233-238.
343. Lykidis A, Mavromatis K, Ivanova N, Anderson I, Land M, DiBartolo G, Martinez M, Lapidus A, Lucas S, Copeland A, et al: **Genome sequence and analysis of the soil cellulolytic actinomycete Thermobifida fusca YX.** *J Bacteriol* 2007, **189:**2477-2486.
344. Fierobe HP, Bagnara-Tardif C, Gaudin C, Guerlesquin F, Sauve P, Belaich A, Belaich JP: **Purification and characterization of endoglucanase C from *Clostridium cellulolyticum*. Catalytic comparison with endoglucanase A.** *Eur J Biochem* 1993, **217:**557-565.
345. Ogura J, Toyoda A, Kurosawa T, Chong AL, Chohnan S, Masaki T: **Purification, characterization, and gene analysis of cellulase (Cel8A) from Lysobacter sp. IB-9374.** *Biosci Biotechnol Biochem* 2006, **70:**2420-2428.
346. Qi M, Jun HS, Forsberg CW: **Characterization and synergistic interactions of Fibrobacter succinogenes glycoside hydrolases.** *Appl Environ Microbiol* 2007, **73:**6098-6105.
347. Handelsman J, Tiedje J, Alvarez-Cohen L, Ashburner M, Cann IKO, Delong EF, Doolittle WF, Fraser-Liggett CMR, Godzik A, Gordon JI, et al: *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet.* Washington, DC: National Academies Press; 2007.
348. Bartolome B, Jubete Y, Martinez E, de la Cruz F: **Construction and properties of a family of pACYC184-derived cloning vectors compatible with pBR322 and its derivatives.** *Gene* 1991, **102:**75-78.
349. Bocs S, Cruveiller S, Vallenet D, Nuel G, Medigue C: **AMIGene: Annotation of MIcrobial Genes.** *Nucleic Acids Res* 2003, **31:**3723-3726.
350. Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, Lajus A, Pascal G, Scarpelli C, Medigue C: **MaGe: a microbial genome annotation system supported by synteny results.** *Nucleic Acids Res* 2006, **34:**53-65.
351. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M: **ISfinder: the reference centre for bacterial insertion sequences.** *Nucleic Acids Res* 2006, **34:**D32-36.
352. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res* 2005, **33:**511-518.
353. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010, **59:**307-321.
354. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM: **The Ribosomal Database Project: improved alignments and new tools for rRNA analysis.** *Nucleic Acids Res* 2009, **37:**D141-145.
355. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Syst Biol* 2007, **56:**564-577.
356. Delcher AL, Phillippy A, Carlton J, Salzberg SL: **Fast algorithms for large-scale genome alignment and comparison.** *Nucleic Acids Res* 2002, **30:**2478-2483.

357. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R: **REPuter: the manifold applications of repeat analysis on a genomic scale.** *Nucleic Acids Res* 2001, **29:**4633-4642.
358. Hauwaerts D, Alexandre G, Das SK, Vanderleyden J, Zhulin IB: **A major chemotaxis gene cluster in Azospirillum brasilense and relationships between chemotaxis operons in alpha-proteobacteria.** *FEMS Microbiol Lett* 2002, **208:**61-67.
359. Thompson MR, Chourey K, Froelich JM, Erickson BK, VerBerkmoes NC, Hettich RL: **Experimental approach for deep proteome measurements from small-scale microbial biomass samples.** *Analytical Chemistry* 2008, **80:**9517-9525.
360. McDonald WH, Ohi R, Miyamoto DT, Mitchison TJ, Yates JR: **Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT.** *International Journal of Mass Spectrometry* 2002, **219:**245-251.
361. Washburn MP, Wolters D, Yates JR: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology.** *Nature Biotechnology* 2001, **19:**242-247.
362. Wolters DA, Washburn MP, Yates JR: **An automated multidimensional protein identification technology for shotgun proteomics.** *Analytical Chemistry* 2001, **73:**5683-5690.
363. Peng JM, Elias JE, Thoreen CC, Licklider LJ, Gygi SP: **Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome.** *Journal of Proteome Research* 2003, **2:**43-50.
364. Eng JK, Mccormack AL, Yates JR: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.** *Journal of the American Society for Mass Spectrometry* 1994, **5:**976-989.
365. Tabb DL, McDonald WH, Yates JR: **DTASelect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics.** *Journal of Proteome Research* 2002, **1:**21-26.
366. Washburn MP, Florens L, Carozza MJ, Swanson SK, Fournier M, Coleman MK, Workman JL: **Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors.** *Methods* 2006, **40:**303-311.
367. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38:**D211-222.
368. Sukharnikov LO, Cantwell BJ, Podar M, Zhulin IB: **Cellulases: ambiguous nonhomologous enzymes in a genomic perspective.** *Trends Biotechnol* 2011, **29:**473-479.
369. Wuichet K, Alexander RP, Zhulin IB: **Comparative genomic and protein sequence analyses of a complex system controlling bacterial chemotaxis.** *Methods Enzymol* 2007, **422:**1-31.
370. Shaw PD, Ping G, Daly SL, Cha C, Cronan JE, Jr., Rinehart KL, Farrand SK: **Detecting and characterizing N-acyl-homoserine lactone signal molecules by thin-layer chromatography.** *Proc Natl Acad Sci U S A* 1997, **94:**6036-6041.
371. Park SR, Cho SJ, Yun HD: **Cloning and sequencing of pel gene responsible for CMCase activity from Erwinia chrysanthemi PY35.** *Biosci Biotechnol Biochem* 2000, **64:**925-930.
372. Alexeyev MF: **The pKNOCK series of broad-host-range mobilizable suicide vectors for gene knockout and targeted DNA insertion into the chromosome of gram-negative bacteria.** *Biotechniques* 1999, **26:**824-826, 828.
373. Bloemberg GV, Wijfjes AH, Lamers GE, Stuurman N, Lugtenberg BJ: **Simultaneous imaging of Pseudomonas fluorescens WCS365 populations expressing three different autofluorescent proteins in the rhizosphere: new perspectives for studying microbial communities.** *Mol Plant Microbe Interact* 2000, **13:**1170-1176.
374. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40:**D290-D301.

375. Disz T, Akhter S, Cuevas D, Olson R, Overbeek R, Vonstein V, Stevens R, Edwards RA: **Accessing the SEED Genome Databases via Web Services API: Tools for Programmers.** *BMC Bioinformatics* 2010, **11**.

VITA

Kirill Borziak was born in Russia. Having moved to Tennessee, he pursued a Bachelor's Degree in Biochemistry/Cellular and Molecular Biology. Now finishing his Ph.D. at the age of 28, he is ready to set out for new adventures in a far away land...