Complete genome sequence of *Caulobacter crescentus*

William C. Nierman*†, Tamara V. Feldblyum†, Michael T. Laub‡, Ian T. Paulsen†, Karen E. Nelson†, Jonathan Eisen†, John F. Heidelberg†, M. R. K. Alley§, Noriko Ohta¶, Janine R. Maddock‖, Isabel Potocka§, William C. Nelson†, Austin Newton¶, Craig Stephens**, Nikhil D. Phadke‖, Bert Ely††, Robert T. DeBoy†, Robert J. Dodson†, A. Scott Durkin†, Michelle L. Gwinn†, Daniel H. Haft†, James F. Kolonay†, John Smit‡‡, M. B. Craven†, Hoda Khouri†, Jyoti Shetty†, Kristi Berry†, Teresa Utterback†, Kevin Tran†, Alex Wolf†, Jessica Vamathevan†, Maria Ermolaeva†, Owen White†, Steven L. Salzberg†, J. Craig Venter†§§, Lucy Shapiro‡, and Claire M. Fraser†

†The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850; ‡Department of Developmental Biology, Beckman Center, Stanford University School of Medicine, Stanford, CA 94305; §Department of Biochemistry, Imperial College of Science, Technology and Medicine, London SW7 2AY, United Kingdom; ¶Department of Molecular Biology, Princeton University, Princeton, NJ 08544; ‖Department of Biology, University of Michigan, 830 North University, Ann Arbor, MI 48109-1048; **Biology Department, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053; ††Department of Biological Sciences, University of South Carolina, Columbia, SC 29208; ‡‡Department of Microbiology and Immunology, no. 300, 6174 University Boulevard, University of British Columbia, Vancouver, BC V6T 1Z3 Canada; and §§Celera Genomics, 45 West Gude Drive, Rockville, MD 20850

Contributed by Lucy Shapiro, January 17, 2001

The complete genome sequence of *Caulobacter crescentus* was determined to be 4,016,942 base pairs in a single circular chromosome encoding 3,767 genes. This organism, which grows in a dilute aquatic environment, coordinates the cell division cycle and multiple cell differentiation events. With the annotated genome sequence, a full description of the genetic network that controls bacterial differentiation, cell growth, and cell cycle progression is within reach. Two-component signal transduction proteins are known to play a significant role in cell cycle progression. Genome analysis revealed that the *C. crescentus* genome encodes a significantly higher number of these signaling proteins (105) than any bacterial genome sequenced thus far. Another regulatory mechanism involved in cell cycle progression is DNA methylation. The occurrence of the recognition sequence for an essential DNA methylating enzyme that is required for cell cycle regulation is severely limited and shows a bias to intergenic regions. The genome contains multiple clusters of genes encoding proteins essential for survival in a nutrient poor habitat. Included are those involved in chemotaxis, outer membrane channel function, degradation of aromatic ring compounds, and the breakdown of plant-derived carbon sources, in addition to many extracytoplasmic function sigma factors, providing the organism with the ability to respond to a wide range of environmental fluctuations. *C. crescentus* is, to our knowledge, the first free-living α-class proteobacterium to be sequenced and will serve as a foundation for exploring the biology of this group of bacteria, which includes the obligate endosymbiont and human pathogen *Rickettsia prowazekii*, the plant pathogen *Agrobacterium tumefaciens*, and the bovine and human pathogen *Brucella abortus*.

Caulobacter crescentus, a Gram-negative bacterium that grows in dilute aquatic environments, is a member of the α-subdivision of proteobacteria. *C. crescentus* invariably differentiates and divides asymmetrically at each cell cycle. Asymmetric cell division and differentiation are recurring themes that underline cellular diversity in multicellular organisms. *C. crescentus* is a simple and highly manipulable single-celled model system to study cellular differentiation, asymmetric division, and their coordination with cell cycle progression (1, 2). *Caulobacter* does all that with less than 4,000 genes, allowing full genome-wide studies of a single differentiating cell.

This stalked bacterium adheres to solid surfaces via a holdfast at the tip of the stalk. It also has a motile swarmer cell stage during its life cycle. The stalked cell acts like a stem cell, continually giving rise to a new swarmer cell at each division (Fig. 1) (3). The production of a swarmer cell with an obligatory motile period minimizes competition during growth in a dilute environment by ensuring that the progeny cell will colonize in a new location (1). The swarmer cell is unable to initiate chromosome replication until it differentiates into a stalked cell. The regulation of cell cycle progression in *C. crescentus* occurs at several levels (4): temporally controlled transcriptional activation and repression, differential phosphorylation of two-component system regulatory proteins, and proteolysis of regulatory and structural proteins. The basic paradigm of cell cycle control used by eukaryotic cells, temporally controlled transcription, phosphorylation of regulatory factors, and targeted proteolysis, has been conserved in *C. crescentus*, although the proteins involved in these processes are different. Recent observations of regulatory proteins dynamically localized to defined cellular addresses at specific times in the *C. crescentus* cell cycle suggest that the three-dimensional organization of this cell adds yet another layer of control (5). The completion of the full genome sequence of this organism provides access to the complete signal transduction network that controls differentiation and cell cycle progression within the context of a unicellular organism growing in a dilute nutrient environment.

Methods

ORF Prediction and Gene Identification. ORFs were identified by using GLIMMER (6). Annotation of the identified ORFs was accomplished by manual curation of the outputs of a variety of similarity searches. Searches of the predicted coding regions were performed with BLASTP, as previously described (7). The protein–protein matches are aligned with blast_extend_repraze, a modified Smith–Waterman (8) algorithm that maximally extends regions of similarity across frameshifts. Gene identification is facilitated by searching against a database of nonredundant bacterial proteins (nraa) developed at TIGR and curated from the public archives GenBank, Genpept, PIR, and SwissProt. Searches matching entries in nraa have the corresponding role, gene common name, percent identity and similarity of match, pairwise sequence alignment, and taxonomy associated with the match assigned to the predicted

Abbreviations: HMM, hidden Markov model; HPK, histidine protein kinase; RR, response regulator.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AE005673).

*To whom reprint requests should be addressed. E-mail: wnierman@tigr.org.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

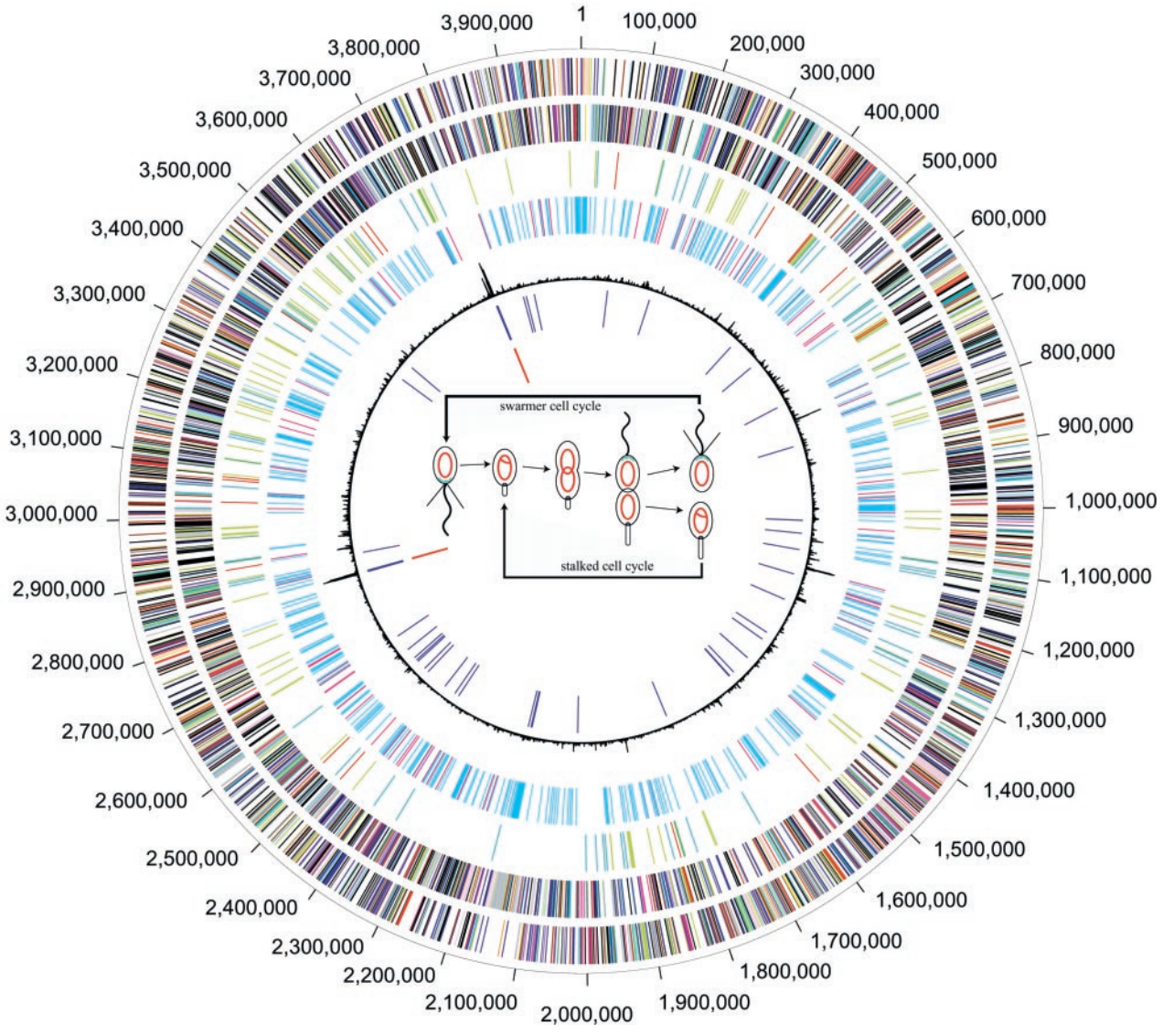

Fig. 1. Circular representation of the *C. crescentus* genome. Coordinate markers around the outside of the circle are in base pairs. First circle, predicted coding regions on the plus strand color coded by role category: violet, amino acid biosynthesis; light blue, biosynthesis of cofactors, prosthetic groups, and carriers; light green, cell envelope; red, cellular processes; brown, central intermediary metabolism; gold, DNA metabolism; light gray, energy metabolism; magenta, fatty acid and phospholipid metabolism; pink, protein synthesis/fate; orange, purines, pyrimidines, nucleosides, nucleotides; olive, regulatory functions; dark green, transcription; teal, transport and binding proteins; salmon, plasmid, phage, and transposon functions; blue, unknown function, hypothetical and conserved hypothetical proteins. Second circle, predicted coding regions on the minus strand color coded by role category. Third circle, genes involved in chemotaxis and motility color coded by role category: olive, two-component regulatory genes; red, methyl-accepting chemotaxis genes; dark green, extracellular function sigma factors; teal, TonB and the TonB-dependent receptors. Fourth circle, cell cycle-regulated genes (2). Fifth circle, atypical nucleotide composition curve. Sixth circle, tRNAs. Seventh circle, rRNAs. The center of the circle contains a schematic of the *C. crescentus* cell cycles. Within the cells, the red circles indicate the nonreplicating chromosome, and the red theta structures indicate replicating the chromosome.

coding region and stored in the database. ORFs were also analyzed with two sets of hidden Markov models (HMMs) constructed for a number of conserved protein families from PFAM (9) and TIGRFAM (10). Regions of the genome without ORFs and ORFs without a database match are reevaluated by using BLASTX as the initial search, and ORFs are extrapolated from regions of alignment. Finally, each putatively identified gene is assigned to one of 113 role categories adapted from Riley (11).

Construction of Paralogous Families. Paralogous families were built in stages. First, for each PFAM HMM scoring above the cutoff to two or more proteins, an alignment of matching regions was constructed. In some cases, the alignment was discarded, trimmed, or used to generate a new global HMM and improved hit region alignment. Second, all peptide sequence outside of the accepted HMM hit regions was subject to automated domain clustering and alignment by MKDOM (12). Several MKDOM cluster alignments were rejected as insignificant or trimmed. The 678 nonoverlapping paralogous domain alignments include 259 supported by PFAM HMMs and 419 created by MKDOM. The alignments include 2,893 regions from 1,801 different proteins of 3,767.

Dinucleotide Signatures Analysis. The data for dinucleotide signature analysis were computed by the method described by Karlin *et al.* (13) with a window size of 100,000 bp and a granularity of 100 bp. χ^2 analysis: the distribution of all 64 trinucleotides (3 mers) was computed for the complete genome in all 6 reading frames, followed by the 3-mer distribution in 2,000 bp windows. Windows

Table 1. General Features of the *C. crescentus* genome

Size, bp	4,016,942
G+C percent	67.2
Total no. ORFs	3,767
ORF size, bp	969
Percent coding	90.6
No. rRNA operons (16S-23S-5S)	2
No. tRNA	51
No. similar to known proteins	2,030 (53.9%)
No. similar to proteins of unknown function	721 (19.2%)
No. hypothetical proteins	1,012 (26.9%)
No. of ORFs in paralogous families	1,801 (47.8%)

overlapped by 1,000 bp. For each window, the χ^2 statistic on the difference between its 3-mer content and that of the whole genome was computed.

General Features of the Genome. The genome sequence of *C. crescentus* CB15 was determined by the whole genome random sequencing method (14). The genome consists of a circular chromosome of 4,016,942 bp with an average G + C content of 67.2%. A total of 3,767 predicted ORFs were identified, of which 2,030 (53.9%) are assigned putative functions, 725 (19.2%) have matches to hypothetical proteins, and 1,012 (26.9%) have no database match (Fig. 4, published as supplemental data on the PNAS web site, www.pnas.org). Coding regions comprise 90.6% of the chromosome. Approximately 1/2 of the proteins (1,801) are members of 678 paralogous families. The largest protein families are response regulators (71 proteins); TonB-dependent outer membrane channels (65 proteins); histidine kinases (61 proteins); and ATP-binding cassette domain transporters (45 proteins) (7). Base pair 1 of the chromosome was assigned within the experimentally determined origin of replication (15), which was also revealed by GC skew analysis (G-C/G + C) (16). Two nontandem ribosomal RNA operons and 51 tRNAs representing all 20 amino acids are present (Table 1).

The genome contains 23 insertion sequences, which consist of 5 multicopy and 3 single-copy elements. Two of the multicopy elements, IS511 and IS298, have been previously described (17); the remaining elements, to our knowledge, are novel (Table 2). There are two nontandem identical 2.2-kb regions containing *N*-acetylglucosamine phosphotransferase system (PTS) components. The two PTS systems presumably reflect a very recent gene duplication event. The 5′ portion of the DNA damage repair gene *radC* is split in the genome from the 3′ portion by a 60-kb insert. This insert contains genes for an extracytoplasmic function sigma factor, a TonB-dependent channel, a putative methyl-accepting chemotaxis homologue, and a metal ion efflux protein. It also contains genes for transposases, conjugal transfer proteins, and several hypothetical proteins that are concentrated toward the outer boundaries of the insertion. Trinucleotide skew analysis identified these outer portions of the insert as different from the genome at large, suggesting that the disrupted *rad*C gene resulted from a plasmid insertion and subsequent recombination events.

Cell Cycle. The control of cell cycle progression in *C. crescentus* has been shown to depend, in large measure, on the differential availability and activation by phosphorylation of the two-component system response regulator CtrA and the CckA histidine kinase. Both of these regulators are essential for viability and control the time of chromosome replication initiation, DNA methylation mediated by the CcrM DNA methyltransferase, cell division, and flagella and pili biogenesis (18, 19). Another essential response regulator, DivK, functions via two nonessential histidine kinases, DivJ and PleC, to coordinate cell division with polar differentiation events (20–22). Critical to cell cycle progression is the proteolysis of CtrA~P at the G1-S transition. Access to the *C. crescentus* genome sequence has now allowed a global approach to the regulatory mechanisms that include targeted proteolysis, signaling protein activation, DNA methylation, and differential gene transcription that allows cell differentiation within the context of the cell cycle.

Proteolysis. Previous work in *C. crescentus* has demonstrated that bacteria, like eukaryotes, regulate proteolysis of specific proteins to control cell cycle progression and morphogenesis (23, 24). For both the response regulator CtrA and the chemoreceptor McpA, residues at or near the C terminus are necessary, although not sufficient, for proper cell cycle-dependent turnover. CtrA ends in a double alanine (AA), and McpA ends in a string of small hydrophobic residues followed by WEEF. Searching the predicted proteome of *C. crescentus* reveals that nearly 20% of all response regulators, and more than 50% of all cell cycle-regulated response regulators and histidine kinases, have C-terminal residues of AA, IA, or VA. The chemoreceptors McpB, McpC, McpD, McpE, like McpA, are all cell cycle-regulated and end in a short string of hydrophobic amino acids such as alanines and valines followed by WEEF. Those hydrophobic residues have recently been found to be essential for turnover (I.P. and M.R.K.A., unpublished results). Of the 12 predicted chemoreceptors lacking the WEEF motif, 8 end in AA or VA. Additionally, 8 of the 23 other predicted chemotaxis genes end in AA or VA. The similarity of C-terminal residues in large families of proteins for which proteolysis has been shown to play a role in cell cycle-regulated turnover suggests that regulated

Table 2. Multicopy insertion sequences

Element	Family	copies	Length	DR*	Structure	Similarity	species	Terminal inverted repeats (5′-3′)
IS511	IS3	4	1266	4	orfA/orfB	self[†]	*C. crescentus*	TGACCTGCCCCTGATTTTTT
								TGACCTGCCTCTGATCTTTC
IS298	IS5	4	845	4	orfA/orfB	self[‡]	*C. crescentus*	GTGGTGTGGACTCTAAGGAT
								GCGGTGTGGACACTTATCGC
ISCc1	IS5	5[§]	848	4	orfA/orfB	IS298	*C. crescentus*	GCCGTAGTGACGATTTAGGA
								GTGGCGGTGACCATTTAGCT
ISCc2	IS110	4[¶]	1140	2	orfA	IS492	*Pseudomonas atlantica*	TATCTGGATTGCAGCGCCAT
								TGTCTGGATCGTCAAGCGGC
ISCc3	IS3	3	1514	2	orfA/orfB	ISD1	*Desulfovibrio vulgaris*	TGTCCGCCGTCAGCGCCAAT
								TGTACGTCGTCAGAAGTTTT

*Size in base pairs of the element (Length) and the direct repeat (DR) generated by insertion into the chromosomal target site.
[†]IS511 gi/1103856/gb/U39501.1/CCU39501[1103856].
[‡]IS298 gi/4836363/gb/AF117124.1/AF117124[4836363].
[§]One copy of ISCc1 is truncated.
[¶]All four copies of ISCc2 occur at the same position in an abundant DNA repeat and may represent a single insertion event.

proteolysis is a significant component of the progression of the cell cycle.

Two-Component Signal Transduction Proteins. *C. crescentus* has the largest number of signal transduction proteins of any sequenced bacterium when adjusted for genome size. In light of the role played by these signal transduction proteins in the control of cell differentiation during cell cycle progression (20) and growth in a dilute aquatic environment, the large number of newly identified members of this family of proteins provides a valuable resource for understanding the complete signaling pathways that control these processes. Analysis of the genome sequence revealed 34 histidine protein kinase (HPK) genes, 44 response regulator (RR) genes, and 27 hybrid HPK/RR genes. Of these, 22 HPKs, 21 RRs, and 26 hybrids were newly identified by the sequence analysis. Approximately 1/3 of the HPK genes are located adjacent to RR genes on the chromosome and are likely to be functional pairs involved in responses to environmental changes. The role of these cognate pairs contrasts with that of the dispersed HPKs and RRs, many of which function in cell cycle regulation (4, 20). A number of these cell cycle-regulated HPKs and RRs are essential for cell viability (4, 18–21). The transcription of at least 35 of the 105 genes encoding two-component signal transduction proteins has recently been shown to be temporally regulated during the cell cycle and to include all those that are essential except DivL (2). Further, of the four studied by fluorescence microscopy in living cells, all were found to be dynamically localized during the cell cycle (19, 22, 23). The signaling proteins whose genes were found to be cell cycle-regulated but are otherwise uncharacterized are thus candidates for having regulatory roles in cell cycle progression.

Eleven of the hybrid HPK/RR proteins are predicted to be cytoplasmic, as are 13 of the nonhybrid HPKs. Surprisingly, many of these cytoplasmic kinases (14 proteins) contain PAS domains, which are often involved in sensing changes in cellular energy levels, oxygen levels, or redox potential. In contrast, few PAS domains are found in the membrane-associated kinases. Thus, it appears that *C. crescentus* relies heavily on molecular networks that sense and respond to intracellular oxygen and redox state.

DNA Methylation. Chromosome methylation on the N-6 adenine of the sequence GAnTC is catalyzed by the CcrM DNA methyltransferase (25). The transcription of the *ccrM* gene is under tight cell cycle control; the CcrM protein is present only in the predivisional cell, when it is available to bring the two newly replicated chromosomes from the hemi- to the full methylation state. CcrM is essential for viability, and its expression at inappropriate times in the cell cycle causes defects in cell division and DNA replication. Thus, temporally regulated methylation of GAnTC site is a component of cell cycle progression. We therefore examined the number of GAnTC sites in the genome and their location with respect to coding sequences. Given the size of the *C. crescentus* genome, if the occurrence of these sites were random, we would expect there to be approximately 12,000 GAnTC sites. In fact, there are only 4,496 of these sites in the genome, and 22% are located between the ORFs that comprise 90.6% of the genome, supporting the argument that the methylation of these sites plays an important regulatory role. Knowledge of the genome position of these sites now opens the way for understanding their function in the regulation of the cell cycle.

Transcription. There are 16 putative RNA polymerase sigma factors in the *C. crescentus* genome, only three of which, *rpoD* (26), *rpoH* (27, 28), and *rpoN* (29), have been previously identified. The 13 new sigma factors revealed by the genome analysis are all extracytoplasmic function (ECF) sigma factors, which typically act to couple periplasmic or extracellular stimuli to changes in gene expression. Two of these new ECF sigma factors, SigT and SigU, are specifically transcribed at the swarmer-to-stalked cell transition and are components of the genetic network that controls cell cycle progression

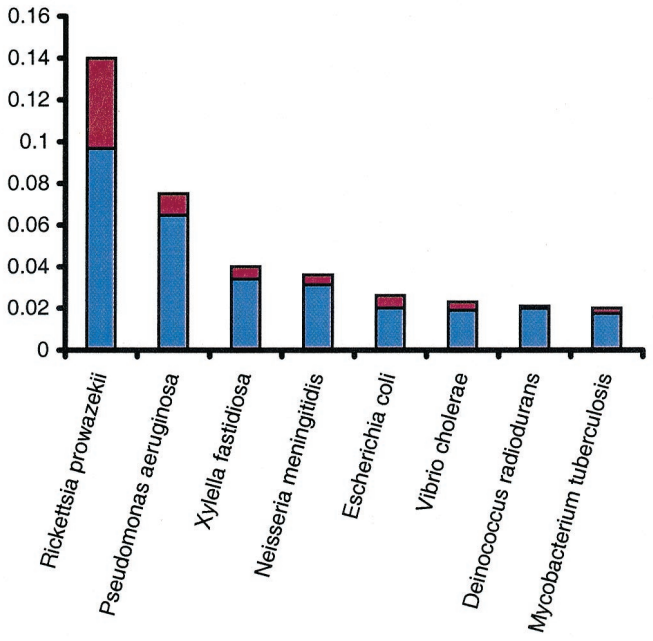

Fig. 2. Comparison of the *C. crescentus* strain CB15 ORFs to those of other completely sequenced organisms. The sequences of all proteins from each completely sequenced genome were retrieved from the National Center for Biotechnology Information and TIGR databases. All *C. crescentus* ORFs were searched against the ORFs from all other genomes with FASTA3. The number of ORFs with the highest similarity ($P < 10-5$) to an ORF from a given species is shown as a proportion of the total number of ORFs in that species. The red portion of each organism's bar represents the percentage of genes that were also found to be cell cycle-regulated in *C. crescentus* (2). Only the organisms with the most hits, after adjustment for genome size, are presented.

(2). Thus, the newly identified complement of sigma factors will clearly contribute to understanding the control of the 19% of *C. crescentus* genes whose transcription has been shown to be cell cycle regulated (2). The RNA polymerase holoenzyme containing RpoN (sigma 54) is used in *C. crescentus* for the transcription of genes involved in cell differentiation events, such as flagellar biogenesis (30). Accessory factors that control transcriptional activation, such as those required to work in concert with sigma-54, are candidate mediators of differential gene expression during the cell cycle. There are only four putative RpoN activators in the *C. crescentus* genome, compared with 12 sigma-54 dependent activators in *Escherichia coli* (31), and at least one of these four *C. crescentus* RpoN activators is temporally regulated (2, 30).

Adaptation to Dilute Aquatic Conditions. Genome analysis identified a large number of genes that would enable utilization of dilute carbon sources and provides a comprehensive picture of the strategies used by *C. crescentus* for survival in nutrient-limiting conditions. Unlike *E. coli* and *Vibrio cholerae*, *C. crescentus* has no OmpF-type outer membrane porins that allow the passive diffusion of hydrophilic substrates across the outer membrane. However, it does possess 65 members of the family of TonB-dependent outer membrane channels that catalyze energy-dependent transport across the outer membrane. This is more than any other organism thus far characterized, with the next highest being 34 in *Pseudomonas aeruginosa* (32), and with no other sequenced proteobacteria possessing more than 10. *C. crescentus* has substantially fewer cytoplasmic membrane transporters relative to genome size than either *E. coli* or *V. cholerae* (33). Given *C. crescentus* ' low nutrient habitat, it is surprising that PTS or ATP-binding cassette domain transporters, which usually have high affinity for their substrates, are not overly represented compared with low-substrate affinity

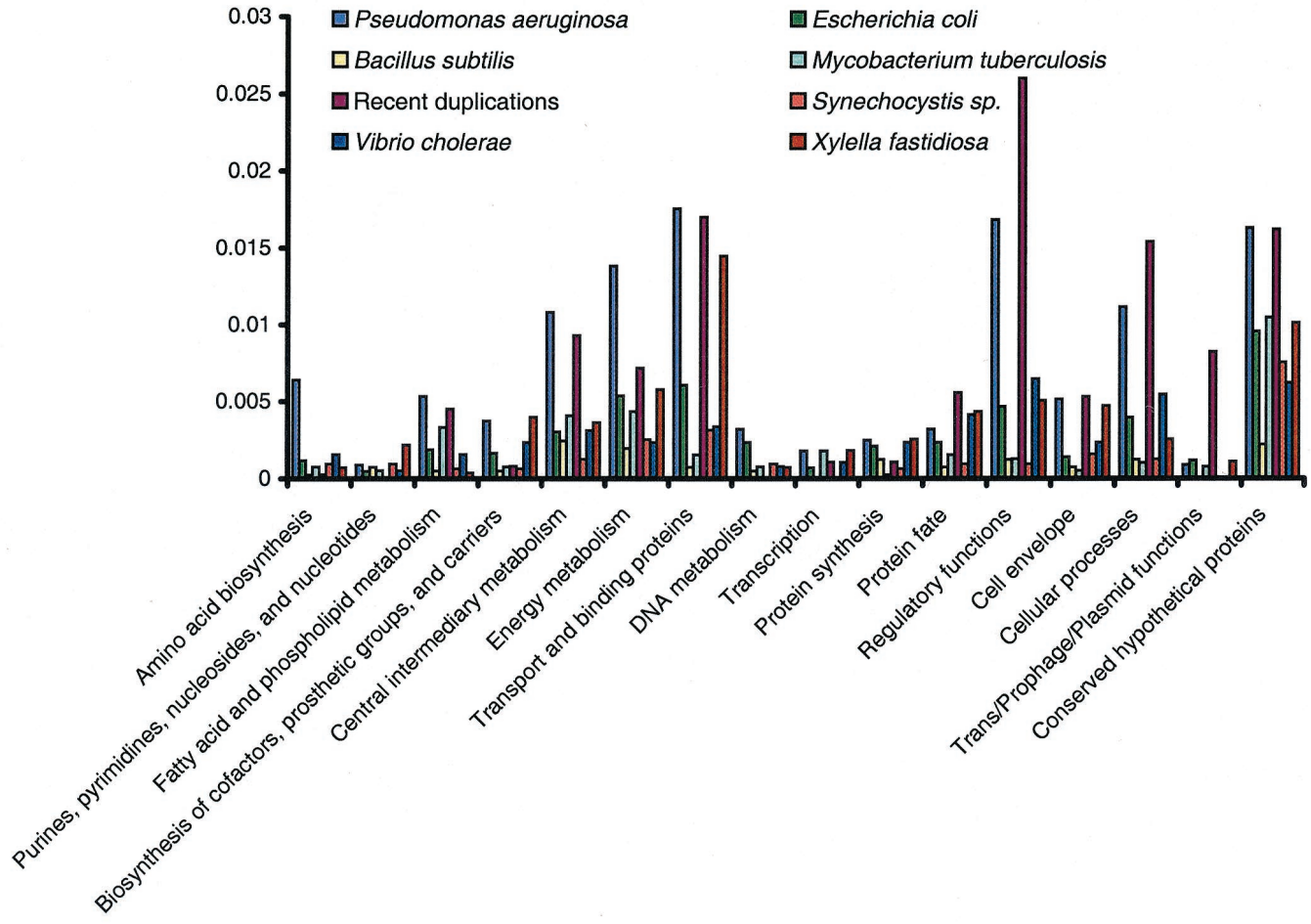

Fig. 3. Comparison of the *C. crescentus* strain CB15 ORFs to those of other completely sequenced organisms by major biological role categories. The number of ORFs in *C. crescentus* that are most similar to other completed genomes was size adjusted, and only those organisms with over 100 significant ($P < 10–5$) most similar hits are presented. ORFs with best hits to *R. prowazekii* are not included to allow detection of the lower number of similarities not caused by the shared common lineage.

transporters. This transporter configuration, energy-gated outer membrane channels for specific substrates and lower-affinity cytoplasmic membrane transporters, may be essential for *C. crescentus* nutrient scavenging, compared with other organisms that use passive diffusion of substrates across their outer membrane by using nonspecific porins and high-affinity inner membrane transporters.

A variety of efflux systems are predicted in *C. crescentus.* This bacterium displays a paracrystalline S-layer on its outer surface composed of a single protein, and this system has been exploited for the heterologous expression of proteins and peptide fragments (34). The S-layer protein is secreted by an ATP-binding cassette (ABC) domain transporter, together with membrane fusion proteins (MFP) and outer membrane factor (OMF) family constituents (35). At least one other ABC-type protein secretion system and a PTS complex polysaccharide extrusion system are present in *C. crescentus*, which may also play roles in secreting cell adhesion products. Representatives are also present from all known prokaryotic families of multidrug efflux systems, as well as four amino acid efflux systems and four resistance-nodulation-cell division- (RND) type metal ion efflux systems. Thus, *C. crescentus* is equipped to carry out active scavenging and secretion processes when growing in extreme dilute environments.

C. crescentus possesses a large number of genes for sensing and responding to environmental substrates (Fig. 1). Approximately 2.5% of the genome is devoted to swarmer cell motility (chemotaxis and flagellum related genes). Before the determination of the genome sequence, 44 sequenced genes were implicated in assembly or activity of the *C. crescentus* flagellum (30). Nine additional genes are present in the genome whose predicted products are similar to proteins with roles in flagella biogenesis. There are two chemotaxis operons in *C. crescentus.* The previously characterized *mcpA* operon (McpA, CheX, CheYI, CheAI, CheWI, CheRI, CheBI, CheYII, CheD, CheU, CheYIII, and CheE) is essential for chemotaxis (36, 37) and is closely related to the chemotaxis operons from other α-proteobacteria such as *Rhodobacter sphaeroides* (38) and *Sinorhizobium meliloti* (39). The newly revealed second chemotaxis operon (McpK, CheAII, CheWII, CheYIV, CheBII, and CheRII) is most similar to that in the α-proteobacterium, *Rhodospirillum centenum.* In addition to the two defined chemotaxis operons, a large number of *mcp* genes and *cheY* genes are scattered throughout the genome. The genome contains 16 unlinked genes encoding chemoreceptors (MCPs), indicating that *Caulobacter* has the ability to respond to a wide variety of compounds. Six of the MCPs lack membrane-spanning domains and may be involved in sensing cytoplasmic substrates linked to cell cycle events.

Phylogeny. Comparison of the *C. crescentus* proteome to those of all other organisms for which the complete genome sequence is available demonstrated the close relationship between *C. crescentus* and the nonfree-living endosymbiont, *Rickettsia prowazekii,* the only other sequenced α-proteobacterium (40). Analysis of the genome sequence of *Rickettsia* indicated that the obligate endosymbiosis of

this bacterium has led to a dramatic reduction of its genome size and the elimination of large numbers and sets of genes (40). We predict that those genes critical to cell cycle progression in *Rickettsia* are less likely to have been lost during its reductive evolution. This prediction is supported by two comparative global analyses. First, of the *C. crescentus* ORFs whose predicted protein matched closest to a *Rickettsia* homolog, more than 30% were cell cycle regulated in *C. crescentus* (2) (Fig. 2). That same percentage was only 22% for *E. coli*, 17% for *Neisseria meningitidis*, and less than 15% for all others. Second, we generated a complete list of cell cycle-regulated genes from *C. crescentus* that had homologs in the *Rickettsia* genome. This list of around 150 genes included genes known to be critical for cell cycle progression in *C. crescentus,* such as *ctrA, parAB,* and *recA*, but did not include genes required only for extraneous cell cycle processes such as flagellar biogenesis. This type of global comparison between the *C. crescentus* genome and the reduced genome of *Rickettsia* may ultimately help to discriminate between the "core" cell cycle genes required for proper progression of the cell cycle and the "peripheral" genes that are cell cycle-regulated but not needed for full viability in *C. crescentus*.

Comparison of the *Caulobacter* proteome to that of other species reveals that there are more matches to the *P. aeruginosa* proteome, when scaled for genome size, than to any species other than *R. prowazekii* (Fig. 2). There are approximately twice as many best matches to *P. aeruginosa* than to other γ-proteobacteria (Fig. 3). An explanation for this observation is the existence of a shared biology between members of these genera and the opportunity for gene transfer between these two lineages. This is supported by the nonuniform distribution of best matching proteins across role categories; functions such as transcription and translation are underrepresented, whereas peripheral metabolic functions are overrepresented. The presence in *C. crescentus* of a 20-gene cluster for the metabolism of aromatic compounds, a pathway extensively characterized only in soil bacteria including *Pseudomonas* and *Streptomyces* species (41), highlights a shared biology between this aquatic species and various species of soil bacteria. It also suggests that *C. crescentus* may be exposed to diverse substrates of terrestrial origin in its natural habitat. As revealed by comparative genome analysis, this shared biology between *C. crescentus* and soil organisms extends to other cellular processes. The conservation of gene order and the sequence similarity of genes involved in intermediary metabolism again suggests that gene transfer between these species has taken place. Consistent with this concept, it has been experimentally demonstrated that *C. crescentus* is able to integrate, retain, and efficiently express plasmid encoded degradative pathway genes from *Pseudomonas putida* (42). The presence of genes for the breakdown of numerous plant polysaccharides, including cellulose, xylan, lignin, glucan, and pectin, as well as transporter systems for the import of the resulting sugars, suggests that, unexpectedly, plant polymers are a significant source of metabolites for the central intermediary metabolism of this organism.

Conclusion

Caulobacters are the most prevalent organisms adapted solely for survival in nutrient-poor aquatic and marine environments. The completion of the genomic sequence now lays the foundation for understanding, on a molecular level, how this bacterium's obligate differentiation and asymmetric division enable it to thrive in such dilute habitats. Furthermore, the tools developed for genetic manipulation of *C. crescentus* make it an attractive organism for development as a bioremediation agent (J.S., unpublished results).

With the completion of the annotated sequence of the *C. crescentus* genome, a full description of the genetic network that controls its cell differentiation, cell growth, and cell cycle progression is within reach. Cell cycle analysis of global transcription patterns, proteomics of stable and unstable proteins, genetic and biochemical analysis of phosphotransfer to regulatory proteins, and time-lapse fluorescent imaging for spatial tracking of regulatory proteins will generate a comprehensive map of the *C. crescentus* cell cycle genetic circuitry.

This work was supported by U.S. Department of Energy Office of Biological and Environmental Research Cooperative Agreement DEFC0295ER61962.

1. Brun, Y. V. & Janakiraman, R. (2000) in *Prokaryotic Development*, eds. Brun, Y. V. & Shimkets, L. J. (Am. Soc. Microbiol., Washington, DC), pp. 297–317.
2. Laub, M. T., McAdams, H., Feldblyum,T., Fraser, C. & Shapiro, L. (2000) *Science* **290,** 2144–2148.
3. Stove, J. L. & Stanier, R. Y. (1962) *Nature (London)* **196,** 1189–1192.
4. Hung, D., McAdams, H. & Shapiro, L. (2000) in *Prokaryotic Development*, eds. Brun, Y. V. & Shimkets, L. (Am. Soc. Microbiol., Washington, DC), pp. 361–378.
5. Shapiro, L. & Losick, R. (2000) *Cell* **100,** 89–98.
6. Salzberg, S., Delcher, A., Kasif, S. & White, O. (1998) *Nucleic Acids Res.* **26,** 544–548.
7. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., *et al.* (1995) *Science* **270,** 397–408.
8. Waterman, M. S. (1988) *Methods Enzymol.* **164,** 765–793.
9. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000) *Nucleic Acids Res.* **28,** 263–266.
10. Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, I. T. & White, O. (2001) *Nucleic Acids Res.* **29,** 41–43.
11. Riley, M. (1993) *Microbiol. Rev.* **57,** 862–952.
12. Gouzy, J., Eugene, P., Greene, E. A., Kahn, D. & Corpet, F. (1997) *Comput. Appl. Biosci.* **13,** 601–608.
13. Karlin, S., Campbell, A. M. & Mrazek, J. (1998) *Annu. Rev. Genet.* **32,** 185–196.
14. Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Kickey, E. K., Peterson, J. D., Umayam, L., *et al*. (2000) *Nature (London)* **406,** 477–483.
15. Marczynski, G. T. & Shapiro, L. (1992) *J. Mol. Biol.* **226,** 959–977.
16. Lobry, J. R. (1996) *Mol. Biol. Evol.* **13,** 660–665.
17. Ohta, M., Mullin, D. A., Tarleton, J., Ely, B. & Newton, A. (1990) *J. Bacteriol*. **172,** 236–242.
18. Quon, K., Marczynski, G. & Shapiro, L. (1996) *Cell* **84,** 83–93.
19. Jacobs, C., Domian, I., Maddock, J. R. & Shapiro, L. (1997) *Cell* **97,** 111–120.
20. Ohta, N., Grebe, T. W. & Newton, A. (2000) in *Prokaryotic Development,* eds. Brun, Y. V. & Shimkets, L. J. (Am. Soc. Microbiol., Washington, DC), pp. 341–359.
21. Hecht, G. B., Lane, T., Ohta, N., Sommer, J. N. & Newton, A. (1995) *EMBO J.* **14,** 3915–3924.
22. Wheeler, R. & Shapiro, L. (1999) *Mol. Cell.* **4,** 683–694.
23. Domian, I. J., Quon, K. C. & Shapiro, L. (1997) *Cell* **90,** 415–424.
24. Wright, R. J., Stephens, C. M., Zweiger, G., Shapiro, L. & Alley, M. R. K. (1996) *Genes Dev.* **10,** 1532–1542.
25. Reisenauer, A., Kahng, L. S., McCollum, S. & Shapiro, L. (1999) *J. Bacteriol.* **181,** 5135–5139.
26. Malakooti, J. & Ely, B. (1995) *J. Bacteriol*. **177,** 6854–6860.
27. Reisenauer, A., Mohr, C. D. & Shapiro, L. (1996) *J. Bacteriol.* **178,** 1919–1927.
28. Wu, J. & Newton, A. (1996) *J. Bacteriol.* **178,** 2094–2101.
29. Brun, Y. V. & Shapiro, L. (1992) *Genes Dev.* **6,** 2395–2408.
30. Gober, J. W. & England, J. C. (2000) in *Prokaryotic Development*, eds. Brun, Y. V. & Shimkets, L. (Am. Soc. Microbiol., Washington, DC), pp. 319–339.
31. Blattner F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., *et al*. (1997) *Science* **277,** 1453–1474.
32. Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., *et al*. (2000) *Nature (London)* **406,** 959–964.
33. Paulsen, I. T., Nguyen, L., Rabus, R. & Saier, M. H., Jr. (2000) *J. Mol. Biol*. **301,** 75–101.
34. Bingle, W. H., Nomellini, J. F. & Smit, J. (1997) *Mol. Microbiol.* **26,** 277–288.
35. Awram, P. & Smit, J. (1998) *J. Bacteriol.* **180,** 3062–3069.
36. Alley, M. R. K., Gomes, S. L., Alexander, W. & Shapiro, L. (1991) *Genetics* **129,** 333–341.
37. Tsai, J. W. & Alley, M. R. (2000) *J. Bacteriol.* **182,** 504–507.
38. Ward, M. J., Bell, A. W., Hamblin, P. A., Packer, H. L. & Armitage, J. P. (1995) *Mol. Microbiol.* **7,** 357–366.
39. Greck, M., Platzer, J., Sourjik, V. & Schmitt, R. (1995) *Mol. Microbiol.* **15,** 989–1000.
40. Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C., Podowski, R. M., Naslund, A. K., Eriksson, A. S., Winkler, H. H. & Kurland, C. G. (1998) *Nature (London)* **396,** 133–140.
41. Iwagami, S. G., Yang, K. & Davies, J. (2000) *Appl. Environ. Microbiol.* **66,** 1499–1508.
42. Chatterjee, D. K. & Chatterjee, P. (1987) *J. Bacteriol.* **169,** 2962–2966.