CHAPTER V

SwmB: a highly repetitive 1.12 MDa protein that is required for non-flagellar swimming motility in *Synechococcus*

Abstract

An exceptionally large ORF (>33 kb) was identified in sequencing the genome of the marine unicellular cyanobacterium *Synechococcus sp.* strain WH8102. Both random and directed mutagenesis demonstrate that this gene, called *swmB*, is required for the unique non-flagellar swimming motility exhibited by these cells. The sequence of *swmB* is highly repetitive, with 4 domains of tandem repeats comprising over 60% of the protein. The genomic region encoding *swmB* and several other motility genes, including its putative cognate transporter, has an exceptionally low % G+C content relative to the genome average. This portion of the chromosome is not present in two sequenced non-motile strains suggesting acquisition of these genes by horizontal gene transfer. Gel electrophoresis confirms that the translated protein is approximately 1 megadalton in size. SwmB co-purifies with the outer membrane fraction and is also found in the culture medium. Inactivation of this gene does not appear to disrupt the proper positioning of at least one other known component of the motility apparatus, SwmA, although mutants do appear to be impaired in the generation of both torque and thrust.

Introduction

While the mechanism of non-flagellar swimming motility in *Synechococcus* is still unexplained, at least one structure and several genes required for this process have been determined (18, 19). One protein required for motility in these cells, SwmA, is found in abundance in motile strains (4). SwmA is a glycosylated protein that contains repeated Ca^{2+} binding motifs. Inactivation of *swmA* results in complete loss

of motility, but cells can still generate torque as observed in fortuitously attached mutant cells that retain the ability to spin like wild-type cells. Microscopic analysis of wild-type and *swmA* mutant cells revealed that motile strains possess a para-crystalline S-layer that is absent in the *swmA* mutant, suggesting that SwmA is the S-layer protein (19).

While complete genomic sequence information did not identify other components of the motility apparatus (21), development of a transposon mutagenesis technique allowed for the identification of three chromosomal regions involved in non-flagellar swimming motility (18). One of these regions is particularly interesting, as it possesses a dramatically reduced % G+C content suggesting that this genetic material has been acquire through horizontal gene-transfer. Moreover, included in this chromosomal region is an exceptionally large and repetitive open reading frame (ORF). Two non-motile strains were obtained from independent transposon insertions within this ORF. These two transposon insertions occurred at 12 bp and 5237 bp downstream of the predicted start codon respectively, with both insertions completely eliminating motility. A directed inactivation of *swmB* was constructed by insertional mutagenesis, confirming the non-motile phenotype. This ORF has been named *swmB* (for swimming motility). No additional ORFs are found downstream of *swmB* on this coding strand and thus the loss of motility can be ascribed to inactivation of this gene and not to a polar effect of insertion. Repeated attempts were made to locate fortuitously attached cells exhibiting the attached spinning behavior of *swmA* mutants, but this behavior has yet to be observed in *swmB* mutant cells. Presented here is further characterization of this unusual bacterial protein.

Materials and Methods

Bacterial strains and growth conditions. *Synechococcus sp.* strains WH8102 (30) and its isogenic *swmB* mutant strain Swm-2 (inactivated with suicide plasmid pJM20 as previously described (18)) were grown in SN medium (29) prepared with either local seawater from the Scripps Pier (Scripps Institution of Oceanography, La Jolla, CA) or synthetic ocean water (SOW) prepared as according to Price *et. al* (23) except components were not treated with chelex. Cultures were maintained as either 50-ml cultures in 125-ml glass Ehrlenmeyer flasks, or 1-L cultures in 2.8 L Fernbach flasks grown without shaking. Cultures were incubated at 25°C with constant illumination of 25 $\mu E \cdot m^{-2} \cdot sec^{-1}$. Kanamycin was added to a final concentration of 20 $\mu g \cdot ml^{-1}$ for Swm-2, to select for and maintain mutational insertion.

Sequence analysis. The complete genomic sequence of *Synechococcus sp.* strain WH8102 and annotation was recently reported (21) and is available at the Joint Genome Institute website (http://spider.jgi-psf.org). BLAST-P analysis (1) was conducted using the non-redundant database at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/). Repeats were identified using the MEME/MAST motif discovery and search tools available through the San Diego Supercomputer Center (http://www.meme.sdsc.edu). Additionally, motif searches and transmembrane predictions were conducted using proteomic and sequence analysis tools including ScanProsite, Motif Scan, FingerPRINTScan, and HMMTOP, which are all available through the ExPASy website (http://us.expasy.org). Additionally secondary structure predictions were performed using the Robetta server (16). Due to

the large size of the *swmB* sequence, the ORF was subdivided into subregions according to repeat domains (see Fig. 1) and subjected to the above bioinformatic searches as the full-length coding sequence, as individual domains, and as individual repeats.

Protein purification. Outer membrane (OM) proteins were isolated from WH8102 and Swm-2 strains as described by Brahamsha (4) with some modifications. Briefly, exponentially growing cultures were centrifuged, washed once with 30 ml sterile SN medium, and resuspended at ~125× concentration in ice-cold stripping buffer (50 mM Tris HCl + 10 mM Na_2EDTA + 15% sucrose, pH 8.0) to strip off the outer membranes. After a 30 minute incubation on ice, cells were removed by centrifugation for 10 minutes at 6277 × *g*. The resulting material was subjected to a high speed spin of 100,466 × *g* for 90 minutes at 4°C to pellet the membrane fraction (HSP) and yield a supernatant containing soluble proteins not associated with the pelleted membranes (HSS). Proteins were routinely concentrated and subjected to buffer exchange using Amicon ultra 30,000 or 100,000 molecular weight cut-off (MWCO) centrifugal filters (Millipore, Bedford, MA) as directed by the manufacturer. Following removal of cells from media by centrifugation, proteins from spent media were recovered and concentrated using an ultra-filtration cell (Amicon) with 30,000 MWCO filters and further concentrated with 30,000 MWCO centrifugal filters. Gel electrophoresis was conducted using Nu-PAGE Novex Tris-Acetate 3-8% and Novex Tricine 10-20% gradient gels as recommended by the manufacturer (Invitrogen, Carlsbad, CA). Silver staining (Bio-Rad, Hercules, CA) and SYPRO Ruby staining (Sigma, St. Louis, MO) of gels was conducted as recommended by manufacturers.

Gels recorded on a ChemiImager 5500 (Alpha Innotech, San Diego, CA) and densitometry of SYPRO stained gels performed with AlphaEase FC version 3.2.2 software (Alpha Innotech).

Antibodies to SwmB were raised against full-length SwmB obtained from spent media and gel purified on 3-8% gradient gels. SwmB is several times larger than the next largest protein present in *Synechococcus* sp. WH8102 and is thus well separated from nearby bands on these gels. Multiple lanes were loaded identically and one lane cut off and stained to locate the position of the SwmB band on this gel. The equivalent portion of the remaining unstained gel containing SwmB, was excised for use as antigen in rabbit polyclonal antibody production by Strategic Biosolutions (Newark, DE). Raw sera were partially purified to remove antibodies that cross-react with other *Synechococcus* proteins utilizing French-pressed cell lysates from Swm-2 cells to adsorb non-specific antibodies. Briefly, approximately $5x10^9$ cells were collected by centrifugation for 10 minutes at 7500 × *g* and resuspended in 3 ml lysis buffer containing 3× PBS (PBS: 9.56 mM Na_2HPO_4,145 mM NaCl, pH 7.5), 4× Complete Protease Inhibitor Cocktail (Roche, Indianapolis, IN). Cells were lysed with 4 passes through an Aminco french press mini-cell (Thermo Spectronic, Rochester, NY) at a pressure of 20,000 PSI. To this cell lysate solution 1 ml raw serum was added and incubated overnight at 4°C. Following incubation, debris and adsorbed antibodies were spun out of solution by centrifugation for 10 minutes at 15000 × *g*. SwmA was gel-purified by separating 800 μg of protein from an OM soluble fraction preparation on a 7.5% Tris-glycine SDS gel followed by staining with copper using a Bio-Rad copper-staining kit (Bio-Rad Laboratories, Richmond, CA). The band

containing SwmA was excised from the gel and shipped to HRP, Inc (now Covance Research Products, Denver, PA) where it was used to prepare rabbit polyclonal antiserum.

For Western analysis, proteins were transferred to Invitrolon PVDF (Invitrogen) membranes in NuPAGE transfer buffer (Invitrogen) + 10% MeOH with 110V constant current for 2 hours. Following overnight blocking at 4°C in BLOTTO (15), membranes were incubated for 1.5 hours at room temperature with primary antibodies diluted in BLOTTO (1:500,000 and 1:50,000 for anti-SwmA and anti-SwmB respectively). Membranes were then washed 4 × 15 minutes in PBS + 0.05% Tween 20 (Fisher Scientific, Fair Lawn, New Jersey). Following washes, membranes were incubated 1.5 hours at room temperature with a peroxidase-conjugated anti-rabbit IgG (Sigma, St. Louis, MO) diluted 1:40000 in BLOTTO. Secondary antibody incubation was followed with another 4 × 15 minute washes in PBS + 0.05% Tween 20 and detected with Super Signal West Dura (Pierce, Rockford, IL) as recommended by the manufacturer. Periodic acid-Schiff staining was performed as described (26) to detect glycosylation.

Gel filtration and anion exchange chromatography were performed using an ÄKTA FPLC system with Superose 6 and HiTrap Q XL pre-packed columns, respectively (Amersham, Piscataway, NJ). Gel filtration was performed with 150 mM NaCl in 50 mM Tris-HCl pH 8.0. For anion exchange, a linear gradient of 0 to 1.5 M NaCl in 20 mM Tris-HCl pH 8.0 was used.

Sucrose density gradient centrifugation was performed with a Beckman L8-70M centrifuge and a SW-41 rotor. Stepwise gradients from 18% - 40% sucrose (w/v)

in 4% increments were centrifuged at 288,000 × *g* for 23 hours at 4°C. Ultracentrifugation was carried out with a Beckman TL-100 ultracentrifuge and a TLA 100.1 rotor at 436,000 × *g* for 20 minutes at 4°C.

Electron microscopy. Partially purified SwmB was applied to freshly glow-discharged carbon-formvar coated grids (Ted Pella Inc., Redding CA) and incubated for 10 minutes at room temperature to allow for adsorbtion to carbon film. Grids were then washed twice on drops of milli-Q (Millipore, Billerica, MA) water, and negatively stained for 20 seconds on a drop of 2% uranyl acetate (Polysciences, Inc., Warrington, PA). Similarly, partially purified SwmB was applied to glow-discharged grids for immuno-gold labeling. Following adsorption, grids were floated on drops of blocking solution (PBS + 1% BSA + 5% Normal Goat Serum (Sigma)) for 1.5 hours at room temperature. Grids were transferred directly to drops of primary antibody diluted 1:1000 in blocking solution and incubated 1.5 hours at room temperature. Grids were washed 3 × 8 minutes on drops of PBS before incubating 1.5 hours on drops of 10 nm-gold-conjugated anti-rabbit IgG (Ted Pella Inc.) diluted 1:100 in blocking solution. Grids were washed again 3 × 8 minutes on drops of PBS before negatively staining as described. Samples were visualized and recorded at an acceleration voltage of 80kV on a JEOL 1200EX transmission electron microscope (TEM).

Immuno-localization. For gold labeling, cells were fixed directly in SN medium for 30min with EM grade glutaraldehyde (Sigma) at a final concentration of 0.5%. Following fixation cells were centrifuged 5 minutes at 6500 × *g* to collect cells, washed once for 5 minutes with PBS, then incubated for 15 minutes in blocking

solution of PBS + 1% γ-globulins (Sigma). Cells were then incubated for 2 hours in 1:50 dilution of a primary antibody in blocking solution, followed by two washes with blocking solution. Secondary incubation with a 1:100 dilution of 10-nm gold conjugated goat anti-rabbit IgG (Ted Pella Inc., Redding, CA) in blocking solution was carried out for 1 hour at room temperature followed by 1 washes in blocking solution and one wash in PBS. Cells were post-fixed in 0.5% glutaraldehyde for 1 hour at 4°C. After fixation cells were applied to glow-discharged carbon-formvar coated grids (Ted Pella Inc., Redding CA), allowed to adsorb for 5 minutes and then briefly stained by floating grid for 10 seconds on a drop of 2% (w/v) uranyl acetate and visualized by TEM as described above.

For fluorophore labeling whole cells were fixed directly in SN medium for 30 minutes with EM grade glutaraldehyde (Sigma) at a final concentration of 0.5%. After 30 minutes of fixation cells were applied to poly-L-lysine (Sigma) coated coverslips and incubated for another 30 minutes to adhere cells to coverslip for antibody incubations and washes. Coverslips were washed 3× with PBS followed by blocking for 1 hour at room temperature with PBS + 1% γ-globulins (Sigma) + 1% normal goat serum (Sigma). Coverslips were incubated overnight at 4°C with primary antibodies diluted 1:25 in blocking solution. Following primary antibody incubation coverslips were washed 9× with PBS and 1× with blocking solution. Coverslips were then incubated with AlexaFluor 488-conjugated anti-rabbit IgG (Molecular Probes, Carlsbad, CA) diluted 1:50 in blocking solution for 2.5 hours at room temperature. Following another 8 washes with PBS, samples were equilibrated for 5 minutes in

Slow-Fade light equilibration buffer (Molecular Probes). Following equilibration, 10 µl Slow-Fade Light (Molecular Probes) was applied to each coverslip prior to mounting. Paired images were collected on an Applied Precision Optical sectioning microscope (Issaquah, WA) equipped with a rhodamine filter set (Ex: 555/28 Em: 617/73) to detect fluorescence from chlorophyll and a FITC filter set (Ex: 490/20 Em: 528/38) to detect that from Alexa-488. Images were processed with softWoRx v3.3.6 software.

Mass spectroscopy. Analysis was conducted at the University of California San Diego, Chemistry & Biochemistry Mass Spectrometry Facility. For gel-purified bands, these bands were cut from a protein gel, reduced, alkylated, and extracted for subsequent protein digestion with trypsin (12, 25, 27). Additionally, preparations of partially purified SwmB were digested in solution for subsequent analysis. The resulting fragments were analyzed on an Applied Biosystems (Foster City, CA) QSTAR hybrid quadrupole-TOF mass spectrometer utilizing a Proxeon Biosystems (Denmark) nanospray source. Peptide masses and partial sequence information were matched against those predicted from genomic sequence information.

Results

Sequence analysis. *swmB* is 32.38 kb in length and encodes a predicted protein of 10,791 amino acids with a molecular mass of 1.126 MDa and a pI of 3.98. This ORF is by far, the largest in the genome. *swmB* is almost 5 times larger than the next largest ORFs (conserved hypotheticals SYNW0985 and SYNW2303, both with similarity to RTX proteins, see discussion below) and comprises 1.33% of the entire

genome. In addition to its large size, SwmB is exceptional due to its highly repetitive primary structure. Greater than 60% of the predicted amino acid sequence is comprised of nearly identical repeats that are tandemly arrayed. Four repeat domains, each consisting of distinct repeats, are present (Figs.1 and 2). Repeat domain A consists of 28 highly conserved tandem repeats of 117 residues (type A repeats). Domain A repeats can be subdivided into three distinct types of nearly perfect repeats. A_I and A_{II} share 96.6% sequence identity and these repeats share 71.4% and 70.6% identity with type A_{III} respectively. The three subtype repeats within domain A are then built into larger blocks arranged in consecutive order (A_I - A_{II} - A_{III}) which itself is repeated multiple times (Fig. 1). The 14^{th} repeat at the middle of this tandem array and the 28^{th} repeat at the end, while still clearly related to the A repeat consensus, are less well conserved. Following domain A there is a short 225-residue region followed by a second array of 19 highly conserved tandem repeats of 127 residues each. Domain B repeats are nearly 100% identical with the exception of the first and last repeats, which have 55% and 66% identity with the consensus repeat respectively (Fig. 2). While domain A and domain B repeats do not share clear sequence homology, compositional analysis shows that these domains share similarly skewed amino acid usage (Table 1). These regions are especially rich in asparagine and threonine (highest 99% quantile in the Swiss-Prot database as analyzed by SAPS (6)) while deficient in methionine, arginine, and proline (lowest 5% - 1% quantile).

Additional repeats are present towards the C-terminus: domain C consists of 5 repeats of approximately 225 amino acids, and domain D contains 4 repeats of approximately 52 amino acids. The repeats within these domains are less well

conserved and do not exhibit the near identical nature seen in domains A and B, but are similar in that the first and last repeats of each tandem array are more degenerate. Additionally, these repeats show the same distinctive bias in amino acid composition similar to domains A and B.

In addition to its large size and repetitive sequence, *swmB* has a strikingly different % G+C content as compared to the rest of the *Synechococcus sp.* strain WH8102 genome (18). The genome average content of guanine and cytosine is 59.41%, while the sequence of *swmB* is only 42.91% G+C. In addition to this strikingly different % G+C content, *swmB* codon usage shows significant variation from the rest of the genome. A comparison of relative synonymous codon usage (RSCU) highlights differences between codon usage for this single gene as compared to the whole genome (Table 2). A Chi-square test of percent usage of each codon shows statistically significantly different ($p<.001$) codon usage for all amino acids except for glutamate, cysteine and lysine. The low % G+C region that contains *swmB* also encompasses 10 other ORFs, two of which have been identified as motility genes by transposon mutagenesis (18). Comparison of this chromosomal region with the homologous regions in two non-motile *Synechococcus* strains, for which complete genomes are available, shows extensive conservation of both gene content and synteny in the sequence immediately flanking the low % G+C region (Fig. 3).

Due to the extreme length of this protein, similarity searches were conducted using the entire sequence of *swmB* as well as each region and each repeat separately. BLAST-P analysis (1) of SwmB was conducted and the predominance of hits come from genome sequencing projects with most of these annotated as hypothetical or

conserved hypothetical proteins. A few trends are observed however. Hits predominantly align to either domain B or to domain C with some aligning to the non-repetitive portion that precedes domain C as well. Several of these BLAST hits have reported similarity to the RTX (*R*epeats in *T*o*X*in) group of exotoxins, which are secreted, calcium-binding proteins that all share a common nonapeptide repeat (31). The sequence of SwmB however does not contain this RTX repeat.

While BLAST results yield no clear homologs to SwmB, examples of other bacterial proteins with some similarities to SwmB have been identified. A group of cell surface proteins involved in *Staphylococcus aureus* host-cell adhesion called MSCRAMMs (for Microbial Surface Components Recognizing Adhesive Matrix Molecules) are similarly large and repetitive (10). One member of this group is Ebh, a megadalton-sized protein from various *Staphylococcus aureus* strains which contains 44 × 126-residue tandem repeats and is responsible for cellular adhesion (7). LapA is a 900 kDa protein from *Pseudomonas fluorescens* that contains two regions of tandem repeats (9 × 100-residues and 29 × 220-residues) and is required for surface attachment and biofilm formation (13). Lastly, rOmpA is is a 190 kDa protein from *Rickettsia rickettsii* containing 13×72-residue tandem repeats (2) which are required for cell adhesion (17). These proteins are similar to one another in their tandemly repetitive primary structure, large size, extracellular localization, and function. Additionally, these proteins have atypical amino acid usage within their repeated regions that is similar to that of SwmB (Table 1). Whether the function of SwmB is similar to that of these proteins remains to be determined but these similarities have helped to direct efforts at cellular localization of SwmB.

Transmembrane prediction algorithms did not recognize any potential transmembrane helices and motif searches failed to identify any known prokaryotic motifs within the predicted amino acid sequence of *swmB*. SwmB also does not contain any apparent secretion signal sequences. Secondary structural predictions using the Robetta server (16), performed independently on each domain and repeat indicate that SwmB should fold into a predominantly β-sheet conformation.

Protein identification and purification. SDS-PAGE analysis of whole cells, soluble OM fractions, and proteins concentrated from spent media of motile strain WH8102 all show the presence of a high molecular weight band (Fig. 4A). While high molecular mass proteins do not penetrate far into the gel, using rabbit muscle proteins as a relative molecular mass standard, SwmB is observed to be over 1 MDa, as predicted by genomic sequence data. Mass spectrometry analysis of this band excised from a gel identified four unique peptides present within the SwmB sequence (residues 3698-3707, 6720-6728, 8379-8389, and 8602-8612) confirming that this band is SwmB. Periodic acid-Schiff staining did not detect glycosylation of SwmB (results not shown). Insertional inactivation yields cells that do not produce any detectable SwmB, as observed both on gels and by western analysis (Fig. 4B). Swm-2 cells do still produce SwmA at wild-type levels and with wild-type localization (Fig. 4B).

SwmB was partially purified both from spent medium and from whole cells. For the former, spent medium was concentrated approximately 75 fold with an ultrafiltration cell using a 30,000 MWCO filter followed by a further 12 fold concentration with a 30,000 MWCO centrifugal filter. This material was then purified

by sucrose density centrifugation to yield nearly pure SwmB (96.3% of total protein by densitometry) (Fig. 5A). Three contaminanting bands with apparent molecular weights of 100 kDa, 80 kDa, and 58 kDa respectively (determined on Nu PAGE 3-8% tris acetate gels) are present in low amounts as detected by SYPRO staining. Similary, SwmB was purified from whole cells by stripping the outer membrane (OM) with a modified EDTA treatment (24). The material released from cells by EDTA treatment is then subjected to high-speed centrifugation to pellet outer membranes (HSP) and yield a supernatant (HSS) fraction containing SwmB. This HSS was first purified by sucrose density gradient centrifugation followed by a variety of secondary purification steps based on size (gel filtration chromatography), density (ultracentrifugation), charge (anion exchange chromatography and differential ammonium sulfate precipitation). The same three bands were present as minor contaminants (Fig. 5B) in all of these preparations. Mass spectrometry analysis has identified the two largest of these bands as SYNW1565 (a conserved hypothetical protein) and SYNW0406 respectively, of which the latter is a major component the outer membrane insoluble fraction (4). Mass spectrometry analysis of the third contaminating band inconclusively identified this protein as the β-chain subunit of phycoerythrin: a structural component of phycobilisomes. Moreover another mass spectrometry analysis of the total contents in solution identified several phycobilisome components contaminating the sample.

Given its large size and repetitive primary sequence it was hoped that direct visualization of purified SwmB by electron microscopy might be possible. While a pure preparation of SwmB was not obtained, highly enriched samples were visualized

by TEM following heavy metal negative staining. Several distinct structures are visible in these preparations: hexameric rings with a diameter of 14 nm, thin filaments approximately 4 nm width (Fig. 6A, inset), and large bundles of filaments (Fig. 6). Individual filaments are highly variable in length with an average length of 623 nm for the 77 individual filaments measured. Several individual filaments of over 2000 nm were observed. These individual filaments appear quite flexible as many filaments are sharply bent and twisted yet still intact. The size of filament bundles was also highly variable with some greater than 200 μm in length observed. Neither the ring-like or filamentous structures can be attributed to SwmB however as an identical preparation of material from Swm-2 cells also displays both of these structures. Anti-SwmB immunogold labeling of these preparations shows gold labeling is associated preferentially with these filament bundles however (Fig. 6B). Some labeling was observed unassociated with filaments as well, and this appears to be specific as compared to the pre-immune control (Fig. 6C).

While attempts to localize SwmB on whole cells using immunogold labeling and TEM visualization were unsuccessful, immunofluorescent labeling analyzed by deconvolution microscopy did reveal the subcellular localization of SwmB. SwmB is found on the cell surface with an irregular, punctate distribution (Figs. 7A and B). While SwmB has a punctate distribution, it is not localized exclusively to any one part of the cell and appears to be relatively evenly distributed across the entire cell surface. The sub-cellular distribution of SwmB revealed by immunofluorescence contrasts with that of SwmA, which appears as a bright homogenous layer surrounding the cell (Fig.

7E). Additionally, immunofluorescent localization reveals that Swm-2 cells possess SwmA with wild-type distribution (Fig. 7F).

Discussion

Genome sequencing of *Synechococcus sp.* strain WH8102 was anticipated to provide insight to the novel motility exhibited by this bacterium (21). Complete genome sequence information failed to identify more than a few genes suspected to be involved in motility. The most promising candidates were several ORFs with homology to *pilT,* a motor protein involved in pilus retraction (20, 28). While WH8102 does not have the full complement of genes required for pilus formation and microscopic analyses have never observed pili in these cells, the presence of a motility motor protein was reason enough to generate inactivations of these genes but in no case was motility abolished (McCarren and Brahamsha, unpublished results). Consequently a method for random mutagenesis, utilizing a modified Tn5 transposon, was developed to identify genes involved in swimming motility (18). Among other motility genes identified, one extremely large and repetitive protein was discovered to be involved in swimming motility.

At the time of sequencing, this ORF was the largest prokaryotic gene identified to date that we could identify in the literature. Since that time, even larger prokaryotic ORFs have been sequenced but translated proteins corresponding to these ORFs have not been identified. For example, the incomplete genome sequences of *Magnetococcus* strain MC-1 (www.jgi.doe.gov) contains two extremely large ORFs of 15245 amino acids and 11699 amino acids respectively, but putative functions for

these ORFs have not been assigned nor have translated products for ORFs of this size been demonstrated. Additionally multiple strains of *Synechococcus* are currently being sequenced, one of which possesses an ORF nearly three times larger than SwmB (D. Scanlan, personal communication). Experimental work has shown several other bacterial proteins in the megadalton range that are produced and transported to the cell surface. All of the multiple strains of *Staphylococcus aureus* that have been sequenced to date (7) possess a gene, or genes, named *ebh* with predicted products ranging up to 1.13 MDa. Ebh is associated with the cell envelope and expression of a partial fragment of its gene shows host extracellular matrix binding activity (7). Similarly, multiple strains of *Pseudomonas fluorescens* have a gene encoding a large (approximately 900 kDa) cell-surface protein termed LapA that has been shown to be important for substrate attachment and biofilm formation (13).

SwmB is readily apparent in whole cell protein extracts as a large molecular weight band. Although accurate determination of molecular weight is difficult for large proteins that do not enter far into the gel, SwmB is clearly in the megadalton range. This finding suggests that transcription and translation of the entire reading frame likely occurs. Periodic acid-Schiff staining indicates this protein is not glycosylated as is the case for other *Synechococcus* cell surface proteins such as SwmA and a 70 kDa protein (4). Much like SwmA, SwmB copurifies with the outer membrane and does not appear to be an integral outer membrane protein as it purifies with the soluble fraction of outer membrane preparations. Furthermore, both SwmA and SwmB are found in abundance in spent medium. Their location on the cell surface and lack of membrane anchoring may make these proteins more susceptible to

being shed by living cells. These characteristics may provide a clue as to the function of these proteins in swimming motility. Perhaps for the surface wave generation model proposed by Ehlers et al. (9), proper functioning of the motility apparatus requires dynamic and more loosely attached cell envelope layers. It therefore would follow that SwmA and SwmB, as cell surface components of the motility apparatus, are easily shed and build up in the medium.

Inactivation of *swmB* results in a loss of motility but it does not affect the attachment of SwmA to the cell surface as determined by fractionation and immunolocalization experiments. Both WH8102 and Swm-2 strains contain SwmA in the outer membrane preparations (Fig. 4). Furthermore, spent medium from both wild-type and *swmB* mutant strains contain comparable amounts of SwmA (Fig. 4). Lastly, immunofluorescent localization of Swm-2 cells reveals identical distribution of SwmA as wild-type cells (Figs. 7E and F). Clearly SwmB is not just a structural protein involved in the attachment of SwmA to the cell surface. We have searched for fortuitously attached Swm-2 cells to see if these mutants retain the ability to produce torque as is seen in *swmA*$^-$ mutants. Such spinning cells have yet to be observed suggesting that torque production has been eliminated in this mutant. While this behavior is not uncommon in wild-type cells, such spinning cells are infrequently observed in *swmA*$^-$ cells. Thus we cannot rule out the possibility that *swmB*$^-$ cells still produce torque as we may have just not yet observed it.

Purification of SwmB by a variety of methods resulted in highly enriched preparations but several minor contaminants were never completely eliminated. One of these contaminants is a 70 kDa polypeptide that is particularly abundant in EDTA

stripped OM preparations (4), suggesting that following multiple purification steps, pieces of membrane still remain. Whether this is due to specific or non-specific interactions remains to be determined. Mass spectrometry also detected components of the light harvesting phycobilisome structures. Moreover, electron microscopic analysis of these samples revealed the presence of ring-like structures resembling phycobilisome discs in both size and shape (11). While the filamentous structures observed in these preparations are evidently not SwmB, there does appear to be an interaction between SwmB and these filaments as observed by immuno-gold labeling. Due to their extremely large size (several hundred times the size of an individual cell) it seems unlikely that the large bundles of filaments observed by TEM could correspond to actual structures found *in situ.* More likely these are accumulations of individual filaments that occur due to concentration effects. SwmB does associate with these filament bundles however, and it is tempting to speculate that the accumulation of individual filaments into large bundles may be mediated by SwmB.

While the function of SwmB remains uncertain, the origin of this gene poses interesting questions as well. The strikingly different % G+C content of this gene and flanking sequence, as compared to the genome, implies that this piece of DNA has been acquired by horizontal gene transfer. Even more convincing is a comparison of the homologous chromosomal region in two other sequenced marine *Synechococcus* strains: non-motile, oligotrophic strain CC9605 and non-motile, coastal strain CC9902 (genome.jgi-psf.org/mic_home.html). Immediately outside of the low % G+C region encompassing *swmB* and several flanking genes, the gene content and synteny is highly conserved across all three genomes (Fig. 3). Other clusters of motility genes

present on separate regions of the chromosome do not exhibit the altered % G+C content of *swmB* (18). Apparently all the genes required for motility in marine *Synechococcus* were not gained by a single acquisition.

Sequence similarity provides few clues as to the function of SwmB. Of the few significant similarities found, many are to RTX proteins. RTX proteins, most specifically the HlyA hemolysin of *E. coli,* are the prototype substrate for type I secretion pathway across the gram-negative cell envelope (3). Type I secretion relies on a multi-component system comprised of an ABC transporter, a periplasm-spanning membrane fusion protein (MFP), and an outer membrane protein (OMP). While the subtrates for type I secretion are quite varied (small peptides to proteins of varying molecular weights from 19 – 800 kDa, β-glucans, polysaccharides, and sialic acid (14)) several generalizations can be made. Type I secreted proteins are typically very acidic with a pI around 4, these substrates have very few or no cysteine residues, and many transported proteins that do not contain the actual RTX nonapeptide repeat still contain other types of repeats (8). All of these characteristics apply to SwmB.

Notably, an ABC transporter (SYNW0959) of the protein-1 exporter (Prot1E) family (http://www.tcdb.org/tcdb) and an MFP (SYNW0958) are present on the low % G+C region containing *swmB*. Transposon mutagenesis suggests the requirement of these genes for motility (18). Moreover Hinsa *et al.* have shown that the gene cluster adjacent to *lapA* (which encodes another extremely large cell surface protein) contains *lapEBC,* which encodes an OMP, a Prot1E family ABC transporter, and an MFP respectively. Their results show that this multi-component transporter is required for the correct localization of LapA on the cell surface. The sequence characteristics of

SwmB indicates it is exported by a type I secretion apparatus. The presence of both a Pro1E family ABC transporter and an MFP encoded on the same piece of DNA implies that SwmB and the ability to transport such a large protein were acquired together in a single step.

SwmA also contains multiple RTX repeats and may well be transported by a type I secretion mechanism as well. There is a another set of genes encoding a second copy of a Prot1E family ABC transporter (SYNW0193) and an MFP (SYNW0194) that are also required for motility. Mutations in either SYNW0193 or SYNW0194 abolishes motility and these cells do not produce any SwmA (detailed in the following chapter). Lastly, there is one *tolC*-like OMP gene (SYNW2187) present in the *Synechococcus* strain WH8102 genome, completing the genetic complement required for type I secretion.

While the cell surface location of this exceptionally large protein has been determined and a likely mechanism for its transport is suggested, the function of SwmB, once it is in place, remains a mystery. Due to its highly repetitive primary structure, SwmB presumably has a repetitive tertiary structure. Whatever structure one domain assumes, each subsequent repeat should have a similar fold, resulting in a repetitive structure for the complete folded protein. Nevertheless, such a structure was not observed by TEM negative staining. Perhaps a repetitive structure, like that presumed for SwmB, is important for interaction with the highly repetitive S-layer formed by SwmA. If such an interaction occurs, conceivably conformational changes in SwmB could result in structural changes in the S-layer (*i.e*. the mechanical deformations (22) or regions of localized contractions (9) previously proposed). If

these structural changes were manifested and organized in a wave traveling the length of the cell it could result in the motility observed. While this model is largely hypothetical at this point it does provide queries for further research such as do SwmA and SwmB specifically interact, does SwmB undergo conformational changes, and under what conditions?

Acknowledgements

Thanks to Ross Hoffman and Justin Torpey of the UCSD Biochemistry department mass spectrometry facility for skilled mass spectrometry analysis, Kit Pogliano and Maryann Martone for providing facilities and help with microscopic investigations, Charlie Strauss for help with the Robetta server secondary structure predictions, Patrick Chain and JGI for completing genome sequencing and especially for resolving the sequence of *swmB,* and the DOE for financial support.

References

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

2. **Anderson, B. E., G. A. McDonald, D. C. Jones, and R. L. Regnery** 1990. A protective protein antigen of *Rickettsia rickettsii* has tandemly repeated, near-identical sequences. Infect. Immun. **58:**2760-2769.

3. **Blight, M. A., C. Chervaux, and I. B. Holland** 1994. Protein secretion pathways in *Escherichia coli* Curr. Op. Biotechnol. **5:**468-474.

4. **Brahamsha, B.** 1996. An abundant cell-surface polypeptide is required for swimming by the nonflagellated marine cyanobacterium *Synechococcus.* Proc. Natl. Acad. Sci. USA. **93:**6504-6509.

5. **Brahamsha, B.** 1996. A genetic manipulation system for oceanic cyanobacteria of the genus *Synechococcus.* Appl. Environ. Microbiol. **62:**1747-1751.

6. **Brendel, V., P. Bucher, I. Nourbakhsh, B. Blaisdell, and S. Karlin** 1992. Methods and algorithms for statistical analysis of protein sequences. Proc. Natl. Acad. Sci. USA. **89:**2002-2006.

7. **Clarke, S. R., L. G. Harris, R. G. Richards, and S. J. Foster** 2002. Analysis of Ebh, a 1.1-megadalton cell wall-associated fibronectin-binding protein of *Staphylococcus aureus.* Infect. Immun. **70:**6680-6687.

8. **Delepelaire, P.** 2004. Type I secretion in gram-negative bacteria Biochim. Biophys. Acta. **1694:**149-161.

9. **Ehlers, K. M., A. D. T. Samuel, H. C. Berg, and R. Montgomery** 1996. Do cyanobacteria swim using traveling surface waves? Proc. Natl. Acad. Sci. USA. **93:**8340-8343.

10. **Foster, T. J., and M. Hook** 1998. Surface protein adhesins of *Staphylococcus aureus.* Trends Microbiol. **6:**484-488.

11. **Glazer, A. N.** 1982. Phycobilisomes: structure and dynamics. Annu. Rev. Microbiol. **36:**173-198.

12. **Hellman U., Wernstedt C., Gonez J., and Heldin C. H.** 1995. Improvement of an in-gel digestion procedure for the micropreparation of internal protein fragments for amino acid sequencing. Anal. Biochem. **224:**451-455.

13. **Hinsa, S. M., M. Espinosa-Urgel, J. L. Ramos, and G. A. O'Toole** 2003. Transition from reversible to irreversible attachment during biofilm formation by *Pseudomonas fluorescens* WCS365 requires an ABC transporter and a large secreted protein. Mol. Microbiol. **49:**905-918.

14. **Holland, I. B., L. Schmitt, and J. Young** 2005. Type 1 protein secretion in bacteria, the ABC-transporter dependent pathway. Mol. Mem. Biol. **22:**29-39.

15. **Johnson, D. A., J. W. Gautsch, J. R. Sportsman, and E. J. A.** 1984. Improved technique utilizing nonfat dry milk for analysis of proteins and nucleic acids transferred to nitrocellulose. Gene Anal. Tech. **1:**3-8.

16. **Kim, D. E., D. Chivian, and D. Baker** 2004. Protein structure prediction and analysis using the Robetta server. Nucl. Acids Res. **32:**W526-31.

17. **Li, H., and D. H. Walker** 1998. rOmpA is a critical protein for the adhesion of *Rickettsia rickettsii* to host cells. Microb. Path. **24:**289-298.

18. **McCarren, J., and B. Brahamsha** 2005. Transposon mutagenesis in a marine *Synechococcus* strain: Isolation of swimming motility mutants. J. Bacteriol. **187:**4457-4462.

19. **McCarren, J., J. Heuser, R. Roth, N. Yamada, M. Martone, and B. Brahamsha** 2005. Inactivation of *swmA* results in the loss of an outer cell layer in a swimming *Synechococcus* strain. J. Bacteriol. **187:**224-230.

20. **Merz, A. J., M. So, and M. P. Sheetz** 2000. Pilus retraction powers bacterial twitching motility. Nature. **407:**98-102.

21. **Palenik, B., B. Brahamsha, F. W. Larimer, M. Land, L. Hauser, P. Chain, J. Lamerdin, W. Regala, E. E. Allen, J. McCarren, I. Paulsen, A. Dufresne, F. Partensky, E. A. Webb, and J. Waterbury** 2003. The genome of a motile marine *Synechococcus.* Nature. **424:**1037-1042.

22. **Pitta, T. P., and H. C. Berg** 1995. Self-electrophoresis is not the mechanism for motility in swimming cyanobacteria. J. Bacteriol. **177:**5701-5703.

23. **Price, N. M., G. I. Harrison, J. G. Hering, R. J. Hudson, P. M. V. Nirel, B. Palenik, and F. M. M. Morel** 1988/89. Preparation and chemistry of the artificial algal culture medium Aquil. Biol. Oceanogr. **6:**443-461.

24. **Resch, C., and J. Gibson** 1983. Isolation of the carotenoid-containing cell wall of three unicellular cyanobacteria. J. Bacteriol. **55:**345-350.

25. **Rosenfeld, J., J. Capdevielle, J. C. Guillemot, and P. Ferrara** 1992. In-gel digestion of proteins for internal sequence analysis after one or two-dimensional gel electrophoresis. Anal. Biochem. **203:**173-179.

26. **Segrest, J. P., and R. L. Jackson** 1972. Molecular weight determination of glycoproteins by polyacrylamide gel electrophoresis in sodium dodecyl sulfate. Meth. Enzymol. **28:**54-63.

27. **Shevchenko, A., M. Wilm, O. Vorm, and M. Mann** 1996. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. Anal. Chem. **68:**850-858.

28. **Skerker, J. M., and H. C. Berg** 2001. Direct observation of extension and retraction of type IV pili. Proc. Natl. Acad. Sci. USA. **98:**6901-6904.

29. **Waterbury, J. B., and J. M. Willey** 1988. Isolation and growth of marine planktonic cyanobacteria. Meth. Enzymol. **167:**100-105.

30. **Waterbury, J. B., J. M. Willey, D. G. Franks, F. W. Valois, and S. W. Watson** 1985. A cyanobacterium capable of swimming motility. Science. **230:**74-76.

31. **Welch, R. A., C. Forestier, A. Lobo, S. Pellett, W. Thomas, Jr., and G. Rowe** 1992. The synthesis and function of the *Escherichia coli* hemolysin and related RTX exotoxins. FEMS Microbiol. Immun. **105:**29-36.

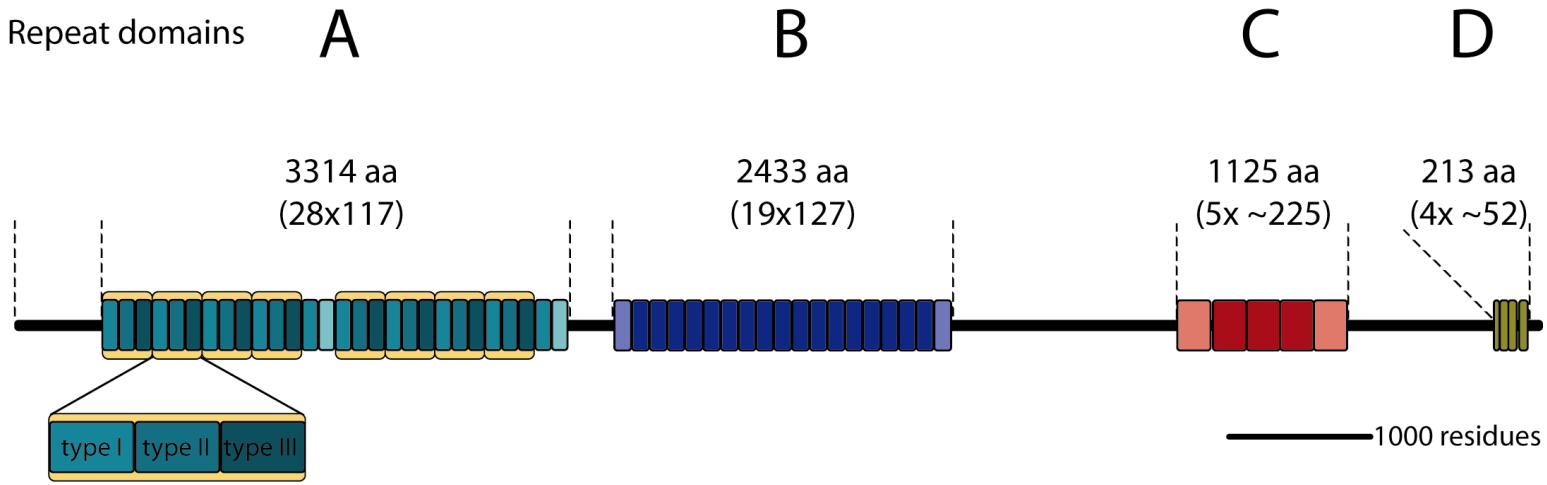

FIG. 1. Diagram of SwmB primary sequence divided according to repetitive domains A-D. Domain A contains three repeat types sharing over 70% identity that are arranged into a larger unit (A_I- A_{II}- A_{III}) which is itself repeated. The central and C-terminal repeats in domain A as well as both terminal repeats in domains B and C are less well conserved than the central core repeats.

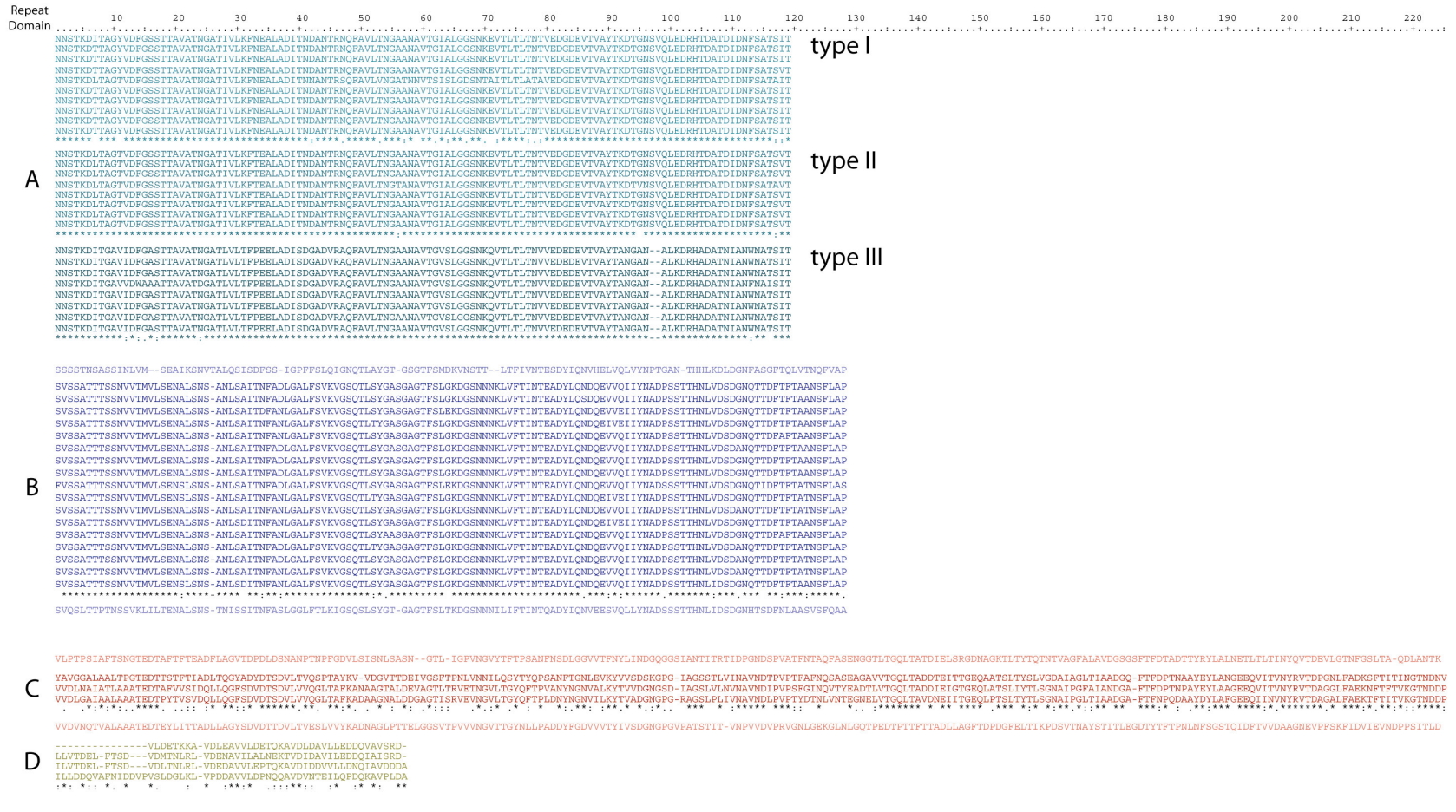

FIG. 2. Amino acid alignments of SwmB domains A-D color coded as in Fig 1. For each domain identical (*), strongly similar (:), and weakly similar (.) residues are marked.

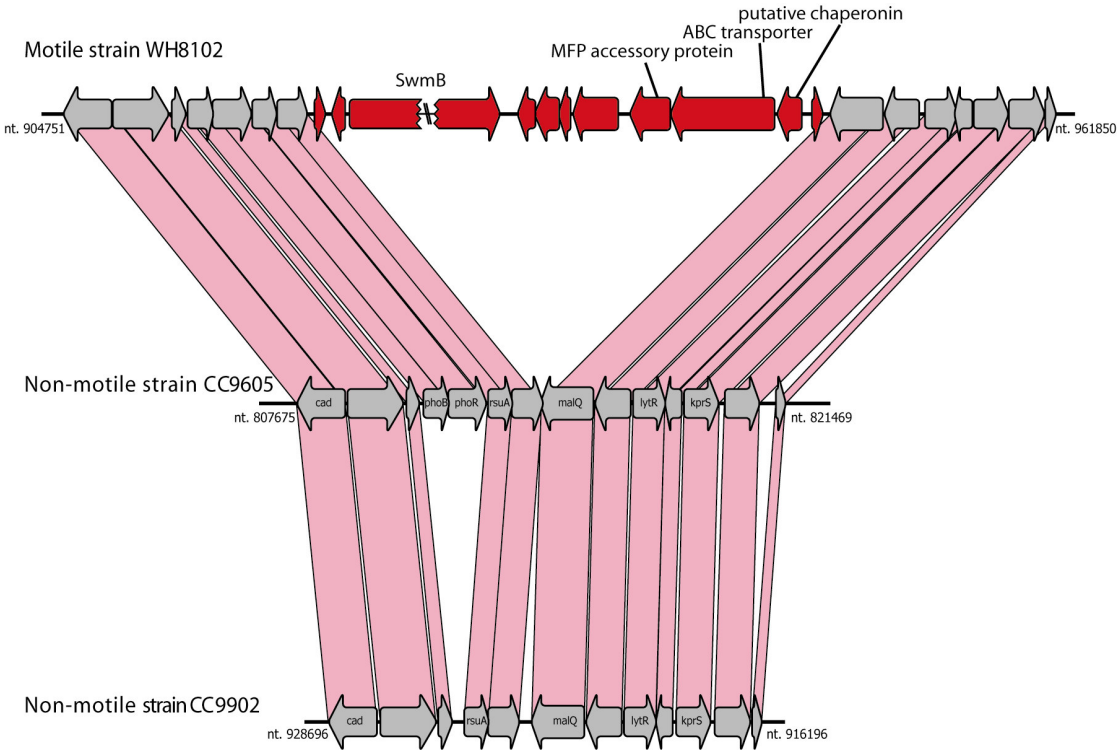

FIG. 3. Chromosomal region containing SwmB and flanking genes. Homologous regions from two non-motile strains CC9605 and CC9902 (www.genome.jgi-psf.org/mic_home.html), show the absence of SwmB and other ORFs (in red) including a multicomponent transport apparatus implicated in motility. The twelve ORFs present only in *Synechococcus sp.* strain WH8102 are contained within a 41.8 kb region of DNA characterized by a % G+C content much lower than the genome average (18).

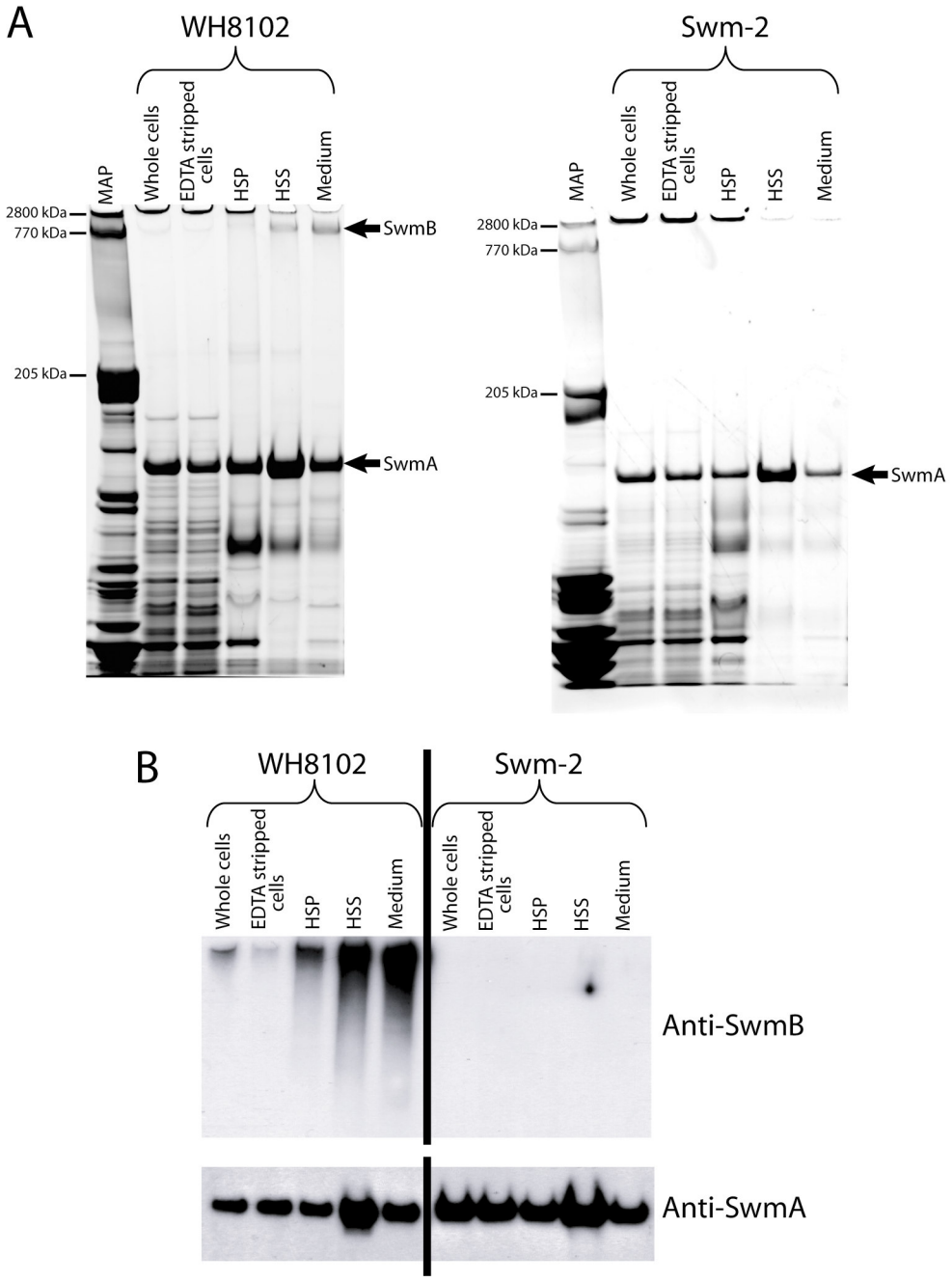

FIG. 4. (A) 3% - 8% SDS-PAGE analysis of motility proteins SwmB and SwmA. MAP, muscle acetone powder containing proteins titin (2800 kDa), nebulin (770 kDa) and myosin (205 kDa) used as molecular weight markers. (B) Western analysis of both motility proteins in cellular fractions from wild-type strain WH8102 and *swmB* mutant strain Swm-2.

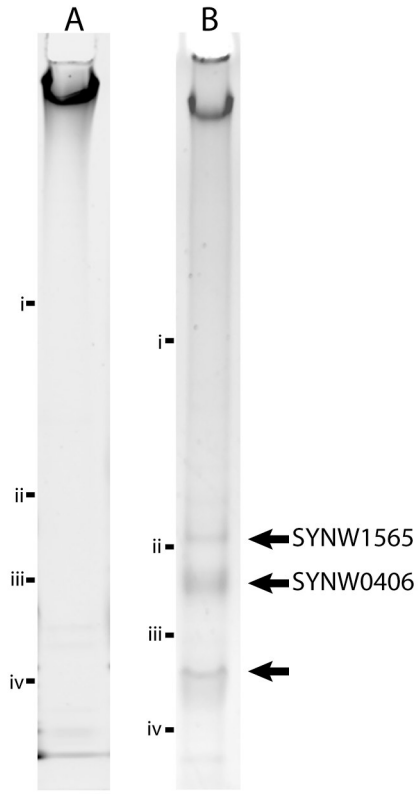

FIG. 5. NuPAGE 3-8% tris-acetate gels of SwmB from spent medium purified by sucrose gradient centrifugation (A). SwmB from HSS fractions purified by sucrose gradient centrifugation followed by ultracentrifugation (B). Identical MW markers used for both samples: 205 kDa (i), 97.4 kDa (ii), 66.2 kDa (iii), and 45 kDa (iv). Three contaminating bands marked with arrows.

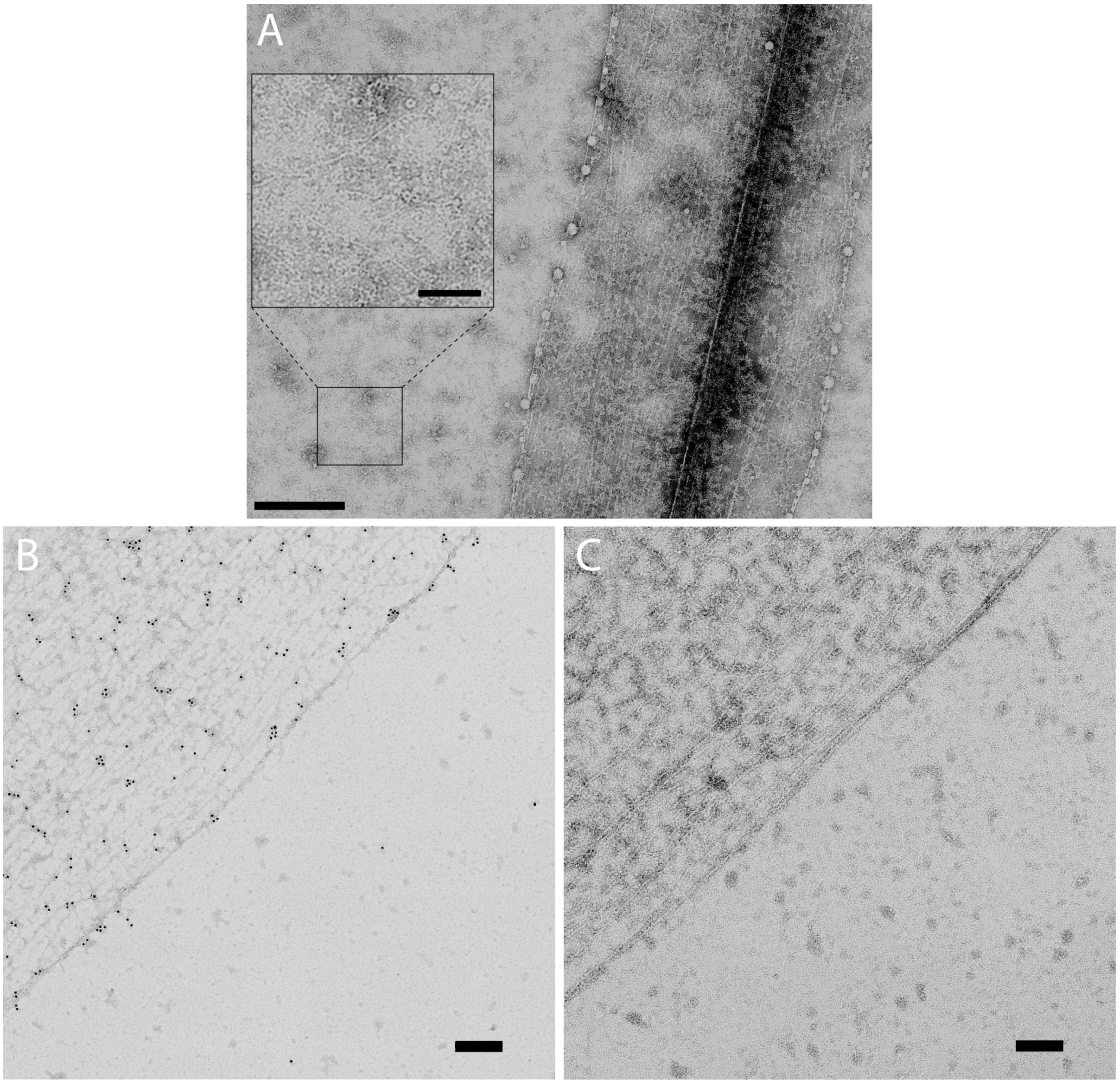

FIG. 6. Partially purified SwmB imaged by negative staining TEM (A, bar, 150 nm). Individual filaments and ring-like structures observed (inset, bar, 100 nm) as well as larger bundles of filaments. Anti-SwmB immunogold labeling of the same material shows gold labeling preferentially associated with filament bundles (B, bar, 200 nm) while the pre-immune control exhibits no labeling (C, bar, 200 nm).

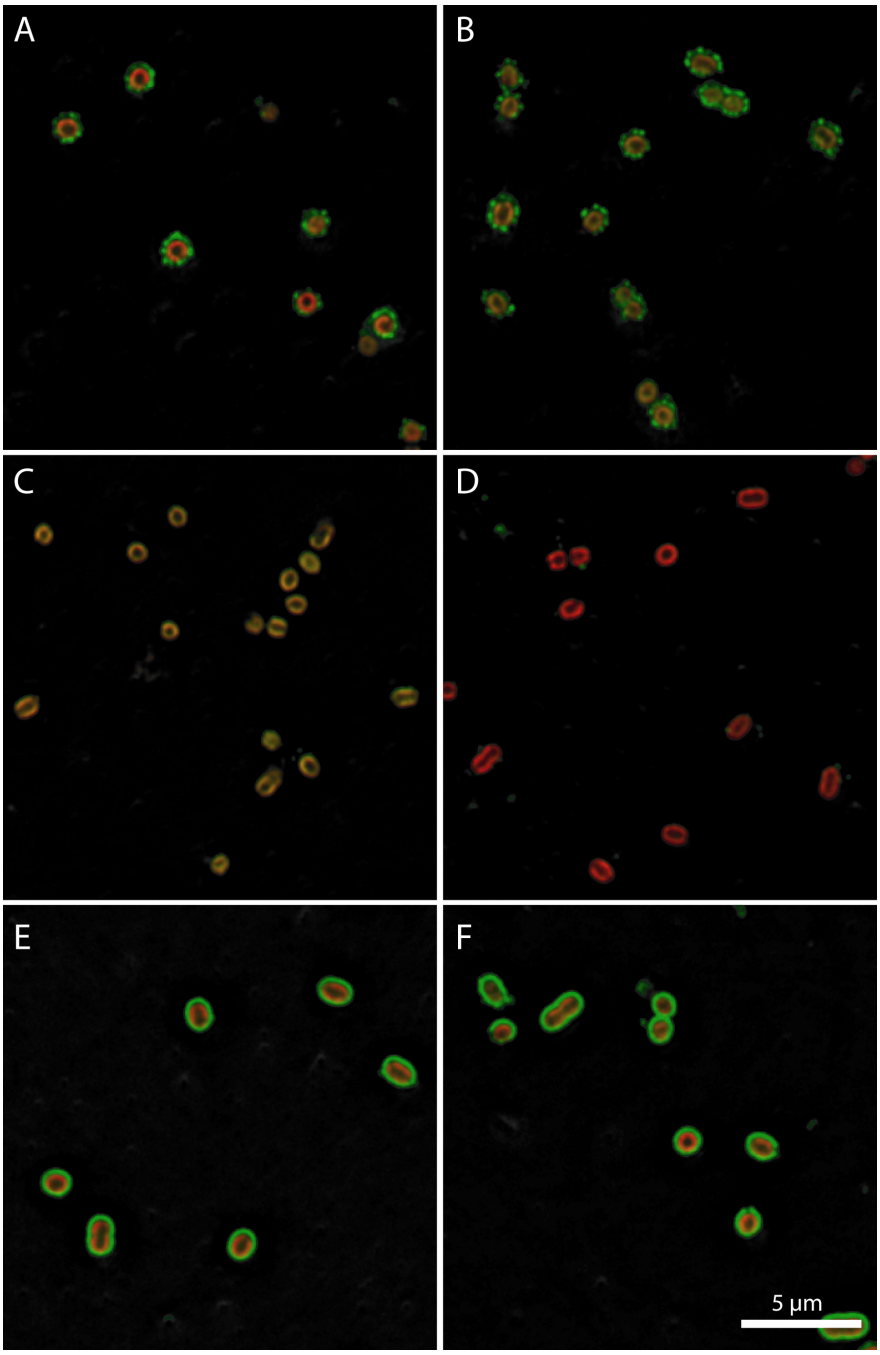

FIG 7. Immunofluorescent labeling of cell surface proteins SwmA and SwmB visualized by deconvolution microscopy. Red autofluorescence from chlorophyll shows the cell shape while immunolabeled proteins are displayed with green fluorescence. Wild-type cells labeled with anti-SwmB (A and B) reveal the punctate, cell-surface distribution of SwmB. Negative controls of wild-type cells labeled with a pre-immune antibody (C) and Swm-2 cells labeled with anti-SwmB do not possess this same distribution. SwmA is detected as a bright, homogenous layer on both wild-type (E) and Swm-2 (F) cells. Bar, 5 µm (all panels).

Table 1. Amino acid usage analysis (6) for several large, repetitive, cell surface prokaryotic proteins

Protein	**Highest 99% quantile**	**Highest 95% quantile**	**Lowest 5% quantile**	**Lowest 1% quantile**
SwmB domain A	N, T	A, D	R	M, P
SwmB domain B	N, S, T		M, P	R
SwmB domain C	T	V, D	K	M, R, H
SwmB domain D	D, V	L	M, R, F	H, G, S
S. aureus Ebh[a] repeats	N	A, Q, T	P, R	F
P. fluorescens LapA domain 2[b]	T	N, V	K, L	H, M, R
P. fluorescens LapA domain 3[b]	T, V			M, R
R. rickettsii rOmpA[c] repeats	N, T, V	A, G	Q, F	E,H,M,P,R,Y
Summary	**N, T, V**		**M, R, P**	

[a]*Staphylococcus aureus* strain COL Ebh (7)
[b]*Pseudomonas. fluorescens* strain WCS365 LapA (13)
[c]*Rickettsia. rickettsii* rOmpA (2)

Table 2. Comparison of Relative Synonymous Codon Usage (RSCU) for the entire *Synechococcus sp.* strain WH8102 genome versus SwmB.

		Entire Genome		SwmB	
Amino Acid	**Codon**	**%**		**%**	**RSCU**
Phe	UUU	33.62	0.67	73.67	1.47
	UUC	66.38	1.33	26.33	0.53
Leu	UUA	2.46	0.15	20.96	1.26
	UUG	15.07	0.90	18.50	1.11
	CUU	10.54	0.63	26.01	1.56
	CUC	19.59	1.18	9.30	0.56
	CUA	2.63	0.16	10.20	0.61
	CUG	49.71	2.98	15.02	0.90
Ile	AUU	25.59	0.77	59.32	1.78
	AUC	70.46	2.11	25.05	0.75
	AUA	3.95	0.12	15.63	0.47
Met	AUG	100.00	1.00	100.00	1.00
Val	GUU	18.12	0.72	50.49	2.02
	GUC	21.63	0.87	17.92	0.72
	GUA	4.12	0.16	21.82	0.87
	GUG	56.13	2.25	9.77	0.39
Ser	UCU	16.02	0.64	47.08	1.88
	UCC	39.89	1.60	14.48	0.58
	UCA	19.52	0.78	24.79	0.99
	UCG	24.57	0.98	13.65	0.55
Pro	CCU	16.62	0.66	46.43	1.86
	CCC	34.25	1.37	16.27	0.65
	CCA	16.87	0.67	28.57	1.14
	CCG	32.26	1.29	8.73	0.35
Thr	ACU	10.84	0.43	32.85	1.31
	ACC	51.87	2.07	18.91	0.76
	ACA	13.75	0.55	28.40	1.14
	ACG	23.54	0.94	19.84	0.79
Ala	GCU	19.02	0.76	37.67	1.51
	GCC	46.27	1.85	14.87	0.59
	GCA	13.51	0.54	32.41	1.30
	GCG	21.20	0.85	15.05	0.60
Tyr	UAU	36.21	0.72	72.91	1.46
	UAC	63.79	1.28	27.09	0.54
His	CAU	43.96	0.88	79.71	1.59
	CAC	56.04	1.12	20.29	0.41
Gln	CAA	25.68	0.51	46.01	0.92
	CAG	74.32	1.49	53.99	1.08
Asn	AAU	35.70	0.71	71.31	1.43
	AAC	64.30	1.29	28.69	0.57
Lys	AAA	35.27	0.71	27.14	0.54
	AAG	64.73	1.29	72.86	1.46
Asp	GAU	52.95	1.06	69.43	1.39
	GAC	47.05	0.94	30.57	0.61
Glu	GAA	42.26	0.85	48.66	0.97
	GAG	57.74	1.15	51.34	1.03
Cys	UGU	29.71	0.59	33.33	0.67
	UGC	70.29	1.41	66.67	1.33
Trp	UGG	100.00	1.00	100.00	1.00
Arg	CGU	20.39	0.82	46.30	1.85
	CGC	41.98	1.68	32.41	1.30
	CGA	11.33	0.45	18.52	0.74
	CGG	26.30	1.05	2.78	0.11
Ser	AGU	26.72	0.53	60.51	1.21
	AGC	73.28	1.47	39.49	0.79
Arg	AGA	41.13	0.82	72.34	1.45
	AGG	58.87	1.18	27.66	0.55
Gly	GGU	24.99	1.00	46.06	1.84
	GGC	42.07	1.68	21.12	0.84
	GGA	15.90	0.64	26.72	1.07
	GGG	17.04	0.68	6.11	0.24

The text of Chapter V, in full, is being prepared for publication. The dissertation author was the primary author, and co-author B. Brahamsha directed and supervised the research, which forms the basis for this chapter.