Published in final edited form as:
J Mol Biol. 2007 October 12; 373(1): 11–26.

Structure and Function of the *E. coli* Protein YmgB: a Protein Critical for Biofilm Formation and Acid-resistance

Jintae Lee[1,$], **Rebecca Page**[2,$], **Rodolfo García-Contreras**[1], **Jeanne-Marie Palermino**[3], **Xue-Song Zhang**[1], **Ojus Doshi**[2], **Thomas K. Wood**[1,4], and **Wolfgang Peti**[3*]

1*Artie McFerrin Department of Chemical Engineering, Texas A & M University, College Station, TX 77843-3122*

2*Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, RI, 02912*

3*Department of Molecular Pharmacology, Physiology and Biotechnology, Brown University, Providence, RI, 02912*

4*Department of Biology, Texas A & M University, College Station, TX 77843-3122*

Abstract

The *Escherichia coli* gene cluster *ymgABC* was identified in transcriptome studies to play a role in biofilm development and stability. In this study we show that YmgB represses biofilm formation in rich medium containing glucose, decreases cellular motility, and protects the cell from acid indicating that YmgB plays a major role in acid-resistance in *E. coli*. Our data also shows that these phenotypes are potentially mediated through interactions with the important cell signal indole. In addition, gel shift assays suggest that YmgB may be a non-specific DNA-binding protein. Using nickel-enrichment DNA microarrays, we show that YmgB binds, either directly or indirectly via a second protein, genes important for biofilm formation. To advance our understanding of the function of YmgB, we used X-ray crystallography to solve the structure of the protein to 1.8 Å resolution. YmgB is a biological dimer that is structurally homologous to the *E. coli* gene regulatory protein Hha, despite its low sequence identity of only 5%. This supports our DNA microarray data that YmgB is a gene regulatory protein. Therefore, this protein, which clearly has a critical role in acid-resistance in *E. coli*, has been renamed as AriR for regulator of acid-resistance influenced by indole.

Introduction

Bacteria seldom live individually[1; 2], and swarming and biofilm formation are two important examples of bacterial multicellular behavior. This multicellular behavior is significant because it enhances their chance for survival in competitive environments[3; 4]. However, it can also cause serious problems, such as infection and biofouling. In fact, 80% of human bacterial infections involve biofilms[5]. In recent years, a number of studies have been initiated with the aim of unraveling the development and regulation of these multicellular structures in more detail. Specifically, genes important for biofilm formation and propagation have been identified using gene chips (for *Pseudomonas aeruginosa*6 and *Escherichia coli*[7; 8; 9; 10]), proteome

Address correspondence to: Wolfgang Peti, PhD, Laboratories of Molecular Medicine, 70 Ship Street, GE-3, Providence, RI, 02912. Fax: 401-863-6087 ; Email: wolfgang_peti@brown.edu.
[$]these authors contributed equally

analysis[11], and classical knockout studies and include genes that code for proteins involved in bacterial motility, quorum sensing, and the induction of polysaccharide synthesis. Notably, a significant subset of these identified genes code for hypothetical proteins of unknown function[12]. One of these hypothetical proteins is the 88 amino acid *E. coli* protein YmgB that we have now functionally and structurally characterized using a powerful cross-disciplinary approach.

Suggestions of the function of the YmgB protein were first obtained from transcriptome studies of the *ymgABC* gene cluster. Specifically, we found that the hypothetical *E. coli* gene cluster *ymgABC* was differentially expressed in biofilm cells. Additionally, we found that this expression was influenced by cell signaling [10; 13; 14; 15]. For example, the furanosyl borate diester or derivative quorum signal autoinducer 2 (AI-2)[16] repressed *ymgAB* 3-fold[14]. In contrast, the biofilm inhibitor furanone from the algae *Delisea pulchra*, which masks autoinducer 2 (AI-2) signaling, induced *ymgA* 2-fold[14]. Furthermore, deleting the AI-2 transporter gene *tqsA* also affected *ymgABC* expression. Specifically, it repressed *ymgBC* 4-fold[13], it induced *ymgABC* 14-fold at 15 h relative to 7 h biofilms[10], and the stationary-phase biofilm signal indole repressed *ymgABC* 2- to 5-fold[17]. Therefore, these results all strongly suggested that the *ymgABC* gene cluster, and thus likely the YmgB protein itself, plays an important role in *E. coli* biofilm formation as a result of AI-2 or indole signaling.

Using similar transcriptome studies, we also found that gene cluster *ymgABC* plays a significant role in acid-resistance[17]. Interestingly and not currently understood, the transcription of a number of acid-resistance genes is down-regulated upon biofilm formation. One example is the *gadABC* operon, which is regulated by GadE. GadABC protects *E. coli* under extremely acidic conditions (pH 2 and below) and allows the bacterium to colonize the gastrointestinal tract[18]. Using biofilm array studies, we showed that biochemical signals correlated with enhanced biofilm formation result in the repression of *gadABC*, and vice versa. Specifically, AI-2 signaling, which increases biofilm formation 30-fold[19], represses *gadABC* 7- to 12-fold[14], while furanone does the opposite, repressing biofilm formation and inducing *gadB* transcription 3-fold[14]. Also, the deletion of the *tqsA* gene, which encodes the AI-2-export protein, causes biofilm formation to increase 7000-fold while *gadABCX* expression is repressed 11- to 13-fold[13]. In addition, when the biofilm inhibitor/stress regulator *bhsA* is deleted, which leads to a 5-fold increase in biofilm formation, *gadABCE* are also repressed 3- to 6-fold[20]. Similar observations are made upon the deletion of *bssR*[21], and *trpE*[17]. Finally, indole repressed *gadABCEX* 2- to 4-fold[17].

The transcription levels of other identified acid-resistance genes, such as *hdeABD* (which function as chaperones to prevent aggregation of periplasmic proteins under extremely acidic conditions[22]) vary in a manner similar to that of *gadABCEX*. *hdeABD* are also repressed 3- to 5-fold by indole[17], 6- to 18-fold[14] by AI-2 and in all mutants (*tqsA, bssR, trpE*) that lead to elevated biofilm formation[13; 17; 21]. In addition, the deletion of *bhsA*, which increases biofilm formation 5-fold, induces *hdeABD* expression 3- to 5-fold[20]. Since the uncharacterized gene cluster *ymgABC* is regulated in these data sets[10; 14; 15; 16; 20] in an almost identical manner to those of the known acid-resistance genes *gadABCEX* and *hdeABD*, we deduced that this locus most likely encodes an acid-resistance locus.

Here we show that YmgB, one protein of this gene cluster, represses biofilm formation in rich medium containing glucose, protects the cell from acid, and decreases cell motility. In addition, our data shows that all of these phenotypes may be mediated through interactions with the cellular signaling molecule indole[17; 23]. However, the biological function of YmgB is unknown and, in fact, is annotated as a ‘hypothetical protein with no assigned function’. Because a 3-dimensional structure of protein is more highly conserved than its sequence, it can frequently lead to an identification of its biological function. Therefore, we used X-ray

crystallography to determine the high resolution structure of YmgB. Based on our structure, we were able to determine that YmgB is structurally and functionally similar to the well-characterized *E. coli* protein Hha despite having only *5% sequence identity*. This prediction, based on structural analysis, was then confirmed using *in vitro* functional gel shift experiments. Thus, using these powerful combined genetic and structural studies, we show that YmgB functions as a regulatory protein which binds DNA promoter sequences that regulate genes important in acid-resistance and biofilm formation.

Results

YmgB and its function in biofilm formation and acid-resistance

YmgB regulation in biofilms and motility—Microarray experiments have shown that the *ymg* locus is highly regulated in response to biofilm formation[10; 13; 14; 17]. Therefore we tested the effect of deleting the *ymg* locus on biofilm formation, and compared these results with those observed for other known acid-resistant genes mutants. Two independent cultures of 10 isogenic *E. coli* K-12 mutants (four mutants—*ycgZ, ymgA, ymgB,* and *ymgC*—along with six *known* acid-resistance mutants—*gadA, gadB, gadE, hdeA, hdeB,* and *hdeD*—as controls) were tested for biofilm formation using the 96-well crystal violet assay in LB and LB glu media. The most significant changes occurred in LB glu medium, which showed significant increases in biofilm formation for all the known acid-resistance mutants except the *gadB* mutant (Figure 1A). These results confirm the importance of acid-resistance genes in biofilm formation[14]. Importantly, deletion of *ymgA, ymgB,* or *ymgC* also increased biofilm formation (9-fold, 4-fold and 6-fold, respectively). Finally, biofilm formation of the *ymgB* mutant was completely complemented by expressing YmgB from a plasmid. Specifically, biofilm formation of the *ymgB* mutant (absorbance at 540 nm of 0.212 ± 0.039) was reduced to 0.029 ± 0.009 for *ymgB*/pCA24N-*ymgB*$^{+}$ with 1 mM IPTG which is comparable to wild-type levels (0.034 ± 0.013 for BW25113/pCA24N + 1 mM IPTG).

Since motility positively influences biofilm formation in *E. coli*[13; 24], we also used the motility halo assay to investigate whether motility plays a role in the observed enhanced biofilm formation of the *ymgABC* mutants. Deletion of *ymgA, ymgB* or *ymgC* significantly increased motility (Figure S1, supporting information), agreeing well with the observed increase in biofilm formation of these mutants (Figure 1A). Also, as previously observed with K-12 wild-type, *trpE, tnaA,* and *tnaC* mutants[17], the addition of 500 μM indole decreased motility of the *ymgA, ymgB,* and *ymgC* mutants by 35 to 65% (Figure S1, supporting information). These results support that indole decreases biofilm formation in part by reducing cellular motility[17].

YmgB controls acid-resistance via indole—To determine which genes are regulated by YmgB and to confirm our hypothesis that *ymgB* may be an acid-resistance locus, a whole transcriptome analysis was performed on biofilm (24 h) formed by either an isogenic *E. coli ymgB* mutant or wild-type strain. As shown in Table S2 (Supporting Information), *gadABCE* and *hdeB,* two well-known acid-resistance gene loci, were repressed 2- to 3-fold in the *ymgB* mutant compared to the wild-type strain. This suggested that the YmgB protein itself may mediate acid-resistance.

To verify these microarray results, we characterized the acid-resistance of the *ymgABC* locus by quantifying the number of cells of the *ymgA, ymgB, ymgC,* and *ycgZ* mutants that survived dilution into an acidic LB medium (pH 2.5; Figure 1B). Significantly, the *ymgB* mutant showed 40-fold less survival at pH 2.5 compared to K-12 wild-type. As expected, the positive controls, cells with *hdeA, hdeB, hdeD, gadA, gadB, gadC,* and *gadE* mutations, displayed 2- to 3400-fold less survival in acidic conditions. These results clearly demonstrate that *ymgB* encodes a protein important for the bacterial acid-resistance phenotype. Surprisingly, however, YmgB

over-expression does not complement the acid-resistance as it does for biofilm formation. The cell survival rate was decreased 10-fold further with 1 mM IPTG induction in the YmgB complementation strain, BW25113 Δ*ymgB*/pCA24N-*ymgB*$^+$ (0.0021 ± 0.0009%). Because pCA24N-*ymgB*$^+$ is under the control of the strong T5-lac promoter, YmgB is robustly expressed. Therefore, the lack of the ability of this protein to complement the *ymgB* deletion mutant for the acid-resistance phenotype may be due, in part, to a difference in the YmgB endogenous expression levels. Alternatively, expression of YmgB from a plasmid may interfere with the three other acid-resistance systems[25], since the deletion of *ymgB* repressed *gadABCE* and *hdeB* as shown in Table S2 (Supporting Information).

The *ymg* locus is also repressed by indole[17] and it is known that indole represses biofilms[10; 17; 20; 21]. Therefore we investigated whether indole plays a role in mediating acid-resistance by adding 2 mM indole to *E. coli* K-12. This indole supplement decreased acid survival at pH 2.5 by 350- to 650-fold. However, addition of 2 mM indole to the *ymgB* mutant did not appreciably change acid survival (1.7-fold decrease for the *ymgB* mutant). Therefore, these results demonstrate that indole plays a role in acid-resistance and that the effect of indole on acid-resistance is mediated by YmgB.

YmgB also confers hydrogen peroxide (oxidant) resistance—The biofilm related gene *bhsA*, which confers resistance to H_2O_2[20], is induced 26-fold upon the addition of H_2O_2[26]. *ymgB* is induced 20-fold by the same treatment[26], therefore we suspected that *ymgB* also confers resistance to H_2O_2. The *ymgB* mutant was incubated with 30 mM H_2O_2 and tested for survival. As expected, the deletion of the *ymgB* gene increased the sensitivity of the bacterium by 400% to H_2O_2 compared to the K-12 wild-type strain (Figure S2, supporting information). As a negative control the *gadE* mutant was also tested for H_2O_2 sensitivity (deletion of the *gadE* gene had the largest effect of all genes tested on *acid* survival compared to WT, Figure 1B). The *gadE* mutant showed only 20% less survival than the wild-type strain; therefore, besides its role in acid-resistance, YmgB is also important for the survival of the cells against H_2O_2, a role independent of its importance in acid-resistance. As observed for the biofilm formation phenotype, H_2O_2 resistance of the *ymgB* mutant was completely complemented with YmgB overexpression (plasmid pCA24N-*ymgB*$^+$ and 1 mM IPTG) (Figure S2, supporting information). This is an interesting difference to acid resistance, where complementation with YmgB was not successful.

Assignment of a function to the hypothetical protein YmgB

The aforementioned biofilm, acid-, H_2O_2-resistance, and indole experiments clearly demonstrate the significant role for YmgB in *E. coli* biofilm formation and acid-resistance. However, as is a regular occurrence in the post-genomic era, no function has yet been assigned to YmgB (P75993) and the protein is classified as a hypothetical protein. Because tertiary structure is exceedingly better conserved than primary structure, solving the 3-dimensional structure of a protein often leads to the determination of its function. The sequence of YmgB is less than *30% identical* to any protein whose coordinates have been deposited in the PDB. Therefore, we used protein X-ray crystallography to determine its high resolution structure in order to obtain insights into its biological function.

YmgB is an all α-helical protein dimer—The structure of YmgB was determined to 1.8 Å resolution using the multiple anomalous dispersion (MAD) method. YmgB forms a dimer in solution, as determined using size exclusion chromatography. It also crystallized as a head-to-head dimer. The final model includes two protein molecules (each containing residues 25-86), 31 water molecules and two b-OG moieties; no electron density was observed for residues 87-88. The Matthews coefficient (Vm) for YmgB is 2.45 and the estimated solvent content is 45.7%[27; 28]. The Ramachandran plot, produced by MolProbity[29], shows that 100%

of the residues are in the most favored regions. Data collection, model and refinement statistics are summarized in Table 1.

Each subunit of the YmgB dimer consists of three α-helices, spanning residues 27-44 (α1), 50-62 (α2) and 67-84 (α3) (Figure 2a). Helices α2 and α3 are oriented in a near perfect anti-parallel fashion with respect to one another with helix α1 crossing in front of them at nearly a 90° angle. The tertiary structure of the monomer is maintained by an extensive network of hydrophobic interactions consisting almost exclusively of leucine, isoleucine and valine residues (Figure 2b). The peripheral residues of the protein are primarily polar and charged.

The dimerization contact is mediated predominantly by residues in helix α1, including Ser31, Leu34, Gly35, Val38, Thr39, Val48 and Met42, and results in the burial of 1326 $Å^2$ of solvent accessible surface (Figure 2c; calculated using the GetArea1.1 program[30]). *Se*Met42, in addition to interacting with residues from helix α1 of the second monomer, also interacts extensively with the terminal residues from helix α3 of the second monomer, especially Ile79 and Tyr83. Because the remainder of the protein crystal contacts buries significantly smaller amounts of solvent accessible surface area (ASA), and because 1326 Å of ASA is within the range observed for other dimerization interactions[31], we predict that this is the biologically relevant interface. Finally, the electrostatic surface of the YmgB dimer is characterized by an extensive acidic patch, which extends around the surface of the dimer interface, punctuated by much small basic patches (Figure 2d).

The N-terminus of YmgB is subject to proteolysis—The N-terminal ~24 amino acids of YmgB were not observed in the electron density map. Using MALDI-TOF mass spectrometry, we demonstrated that the crystallized YmgB protein was ~3 kD less than that predicted based on its full-length amino acid sequence (Figure S3, supporting information). Three major peaks were observed with MWs of 7153, 7210, and 7298. The YmgB protein fragments that correspond to these fragments are residues 25-88, 24-88, and 23-88, respectively. Thus, the proteolytic fragments of YmgB crystallized, explaining why only residues 25-86 were observed in the electron density map. Crystallization of a proteolytic fragment has been previously observed[32]. However, in most cases, the crystals that form such heterogeneous samples typically diffract to low resolution (~3.0 Å). Thus, the proteolytic fragment is often re-cloned and re-crystallized in the hopes that higher resolution data will be obtained. In our case, the YmgB crystals diffract to high resolution (1.8 Å), with residues 25-86 clearly visible and well-ordered in the electron density map. Therefore, YmgB fragment 25-88 was not re-cloned and re-crystallized. Follow-up studies demonstrated that the time course of proteolysis is slow (>4 weeks) and that YmgB fragments 23- to 26-88 are stable for more than six months without further degradation. Finally, additional experiments showed that full-length YmgB crystallized only in the *absence* of b-OG and never diffracted to resolutions higher than 7 Å while, the YmgB C-terminal fragment formed crystals readily in the presence of b-OG and typically diffracted to better than 2.0 Å. Thus, these studies show that YmgB has a highly stable C-terminal domain and an N-terminus which is sensitive to proteolysis suggesting it is highly flexible.

The three-helical dimerization domain of YmgB is conserved within other bacterial proteins—We use the position-specific iterated BLAST search algorithm[33] to identify proteins from other organisms with sequences similar to that of YmgB. Only four distinct sequences were identified, all of which were from bacterial species: *S. typhimurium*, *Enterobacter sp. 638*, *E. carotovora*, and *S. flexneri*. As can be seen from the sequence alignment in Figure 3, their sequences are most similar in the region corresponding to the crystallized YmgB dimerization domain (residues 26-88; yellow, identical; blue, similar).

YmgB is a structural homolog of the E. coli protein Hha and the dimerization domain of thermolysin—In order to further investigate the function of YmgB, we used the ProFunc[34] server, which is used to help identify the likely biochemical function of a protein based on its three-dimensional structure. The 3D template searches of ProFunc indicate that YmgB is not likely to be an enzyme, not expected to bind ligands and not likely to bind specific sequences of DNA. Finally, it is also not predicted to be a helix-turn-helix DNA-binding protein.

Three-dimensional structure alignment showed that the two proteins in the PDB with the most similar structures to YmgB are the *E. coli* protein Hha (PDBID 1JW2[35]; Figure 4a) and the C-terminal dimerization domain of thermolysin (PDBID 1TRL[36]; Figure 4b). The sequence identity with these proteins is extremely low, only 5% and 7%. However, their structures are highly similar. Hha is a small (~8.5 kD) protein that is a thermo- and osmomodulator of the expression of the toxin α-hemolysin. Other studies have demonstrated that Hha interacts with the nucleoid-associated protein H-NS, a general negative regulator of transcription that binds preferentially to curved DNA (geometric recognition)[37; 38; 39]. Like YmgB, Hha is an all α-helical protein although it has four helices while YmgB has three. The N-terminal domain of Hha (residues 7-56, excluding 24-26) superimposes with YmgB with an RMSD of 6.3 Å (Figure 4a). The primary differences between the two proteins (and the source of the high RMSD) are that helix α1 in YmgB is two full turns longer than that of Hha and that Hha has a fourth helix, α4, which interacts with helix α3 on the opposite side of that occupied by helix α1. Notably, under the conditions used for structure determination by NMR spectroscopy, Hha was shown to form a monomer and not a dimer, as is observed for YmgB. Superposition of YmgB with Hha shows that the Hha loop that connects helices α3 and α4 is located at a position that, if present in YmgB, would block the dimerization contact, consistent with the observation that Hha is a monomer in solution. Notably, however, Hha is proposed to function as a specialized homolog of the amino-terminal oligomerization domain of H-NS[40]. In fact, the first three helices of Hha that overlap with YmgB can functionally replace the oligomerization domain of H-NS, resulting in a chimeric protein that can compensate for some of the *hns*-induced phenotypes in *E. coli*[41].

The importance of this 3-helical structure as functioning as a dimerization domain is further supported by the observation that YmgB is highly similar to the dimerization domain of thermolysin. In contrast to Hha, the dimerization domain of thermolysin contains only three helices and forms a dimer in solution. The superposition of YmgB with the dimerization domain of thermolyisin is shown in Figure 4b. As can be observed, the two proteins superimpose well. The most important difference between the two proteins is the relative shift in the orientations of the helices with respect to one another, which result in the N-terminal helix of the thermolysin dimerization domain to be shifted upwards compared to that of YmgB. This has important consequences for the position of the second dimer of both proteins. In particular, there is a substantial shift in the dimerization interface between these two proteins, such that the interaction surface of the thermolysin dimerization domain is much more extensive than that of YmgB. This is evident in the much larger amount of thermolysin ASA, 2587 $Å^2$, buried upon dimerization.

YmgB regulates expression of biofilm genes—The availability of the 3-dimensional structure of YmgB and its observed high structural similarity to Hha suggested key follow-up experiments to test YmgB function. To test the hypothesis that YmgB is a gene regulatory protein that, either alone or in complex with a binding partner, binds DNA, we performed a nickel enrichment DNA microarray assay[42]. In this assay, formaldehyde cross-linking is used to covalently tether his-tagged YmgB to the DNA fragments to which YmgB is bound. The YmgB:DNA cross-linked species are then purified using his_6-tag chromatography. Following proteolytic and RNase digestion of the proteins and RNA, respectively, the DNA is purified

and used for microarray analyses. If a binding partner for YmgB is required for the interaction with DNA, it is present during this experiment and thus the identified promoters are physiologically relevant.

For the nickel enrichment DNA microarray assays, biofilm cells were cultured for 24 h at 37° C in LB glu medium. Under these conditions (the same as those used for the DNA microarray studies), the *ymgB* mutant produces 4-fold more biofilm than the wild type strain. The DNA purified using nickel enrichment from this culture was then used for microarray analysis. The Ni-enriched DNA microarray assay (Figure 5) revealed several possible genes and intergenic (IG) regions (Table 2) to which YmgB binds. These targets are involved in several physiological functions, a subset of which may be involved in biofilm formation. Most notable is the important cyclic AMP regulator protein CRP, which controls global gene expression based on glucose catabolism[43], and is a gene that was, until now, not known to be regulated by any transcription factors other than itself[44]. Another is *lsrF*, which is a putative aldolase that is predicted to be involved in AI-2 catabolism[45] and is part of the *lsrACDBFG* operon that participates in the regulation of the uptake of autoinducer-2 (AI-2). Other genes include *yhjQ*, which encodes an important cell division protein (chromosome partitioning ATPase) that affects morphology, cell size, and chromosome partitioning[46], and *bcsA*, which encodes a putative cellulose synthase that is part of the *bcsABC* operon and is involved in biosynthesis of cellulose. Notably, cellulose is one of the major constituents of biofilm matrices and, although laboratory *E. coli* K-12 derivative strains do not produce it, the synthesis of cellulose is a biofilm determinant of some natural *E. coli* isolates[47]. Thus the regulation of *bcsA* may be one way that YmgB directly regulates biofilm formation. Finally, another key gene identified in these studies was that of *rpsV*, which encodes a 30S ribosomal subunit-associated protein that is induced at stationary phase[48]. This gene is co-transcribed with *bdm* ("biofilm-dependent modulation") whose expression is reduced in biofilms[49] and induced by the Rcs signaling system in response to low temperature growth in glucose and to high external zinc concentrations[50]. The *rpsV* gene was also identified in our regular DNA-microarray of *ymgB* vs. wild-type (repressed 2.3-fold), which further supports the likelihood that YmgB directly regulates this gene. Thus, these nickel enrichment DNA microarray studies show that YmgB, either alone or in complex with a second protein binding partner, binds the genes and/or intergenic regions of a number of genes that play central roles in biofilm formation and acid-resistance.

YmgB exhibits non-specific DNA binding in vitro—Due to the 3-dimensional structural resemblance of YmgB to Hha, it was important to determine whether or not YmgB forms a nucleoprotein complex similar to Hha, which participates in the control of gene expression in biofilms and the acid-resistance response. In order to further probe YmgB-DNA binding activity, we performed a series of electrophoretic mobility shift assays (EMSA) using PCR amplified promoter regions of two genes that were found to be repressed in our DNA microarray studies. These are the acid-resistance gene *gadA* (repressed 2.5-fold) and the ribosomal protein gene *rpsV* (repressed 2.3-fold). The EMSA assays results showed that both the truncated (7.7 kDa) and full-length protein (9 kDa) (Figure 5), either directly or indirectly via an as yet, despite several efforts, unidentified DNA-binding protein (potentially analogous to HNS of the Hha-H-NS complex) bind both DNA targets, suggesting YmgB may be a nucleoid-associated protein. However, when we assayed YmgB binding to other DNA fragments of similar size using genes that were not identified by the DNA-microarray studies, such as a fragment of the *bhsA* (*ycfR*) promoter (232 bp), the protein also exhibited binding. This indicates that the DNA binding activity of YmgB *in vitro* is not sequence dependent, but is most likely, in manner similar to Hha-H-NS, geometry dependent[37; 38]. In addition, as expected from the nickel enrichment DNA microarray assay results, both the full-length and truncated versions of the YmgB protein associate with fragments of the *crp* promoter (216 bp) and the *lsrF* gene (202 bp). Although YmgB is similar to its structural homolog Hha, deletions

of *hha* and *ymgB* in *E. coli* K-12 produce distinct phenotypes. Therefore, it is probable that the *in vivo* functions of these proteins are different. This is also strongly supported by DNA microarrays of the *hha* deletion mutant (biofilm cells, 15 h, LB glu), which exhibit a distinct gene expression pattern compared to the *ymgB* deletion mutant, with very few coincident genes. Remarkably, the acid-resistance genes *gadBC* and *hdeAB*, two of the genes that were common to both arrays, were induced 4- to 5-fold by the *hha* deletion and repressed 2- to 3-fold by the deletion of *ymgB* (unpublished data).

Discussion

To summarize, we have identified that the 88 residue protein YmgB is critical for biofilm formation and acid-resistance in *E. coli*. Figure 6 illustrates the experimental relationship between acid-resistance and biofilm formation, as it was previously identified for the *gadABC* and *hdeABD* genes. In this manuscript we demonstrate that the *ymgABC* locus also follows these relationship rules and thus too plays a significant role in acid-resistance and biofilm formation in *E. coli*. In addition, deletion of *ymgB* affects *gadABCE* and *hdeB* expression, further emphasizing the overall importance of the YmgB protein as a functional regulator of these acid-resistance processes. *E. coli* has at least four complex strategies for acid resistance: 1) a glucose-repressed system, 2) a glutamate-dependent system, 3) an arginine-dependent system, and 4) now, as reported in this manuscript, the YmgB system. These four systems, with the possibility that additional systems are likely to be detected in the next few years, help the bacterium to survive in extreme acid environments. The genetic regulation of these four acid-resistance systems is most likely highly interconnected, as has been previously proposed in the literature[25].

To understand the how the hypothetical protein YmgB protein functions to accomplish these biological processes we carried out a sequence analysis of this protein using NCBI blast (http://www.ncbi.nlm.nih.gov) and ExPASy blast (http://expasy.org/sprot). This analysis did not reveal any previously identified conserved domains and instead only identified a limited number of homologs, all of which were bacterial. *E. coli* YmgB (88 aa) shares 27-51% identity with several unknown-function putative proteins in Gram-negative plant/animal pathogenic bacteria such as *Enterobacter sp.* 638, *Salmonella typhimurium*, *S. choleraesuis*, *S. paratyphi*, *S. typhi*, *Erwinia carotovora* sub sp. atroseptica, *Shigella sonnei*, and *Sh. flexneri*. Therefore, to gain additional insights in the biological function of YmgB, we elucidated its 3-dimensional structure[51] and used this structure to identify *structural* homologs of the YmgB protein.

Using the 3-dimensional structure of YmgB as our search scaffold in MATRAS (http://biunit.naist.jp/matras/)[52], we identified the well-known *E. coli* protein Hha as the closest structural neighbor. This close structural similarity is remarkable (Figure 4), because YmgB and Hha share only 5% sequence identity. Hha is a small (~10 kDa) protein, slightly basic in nature. It is also a nucleoid-associated protein[39] that has been shown to interact with H-NS[37; 38], a 2-domain protein, which is known to regulate the expression of a large number of genes (~5%) in Gram-negative bacteria, such as *E. coli*[53; 54]. Most of these regulated genes are related to environmental conditions and/or virulence. The N-terminal domain of H-NS is responsible for dimerization/multimerization, critical for protein:protein interactions, while the C-terminal domain is essential for DNA binding. Interestingly, recent reports also show that also the N-terminal domain has modulatory functions on the DNA interaction. Hha, which binds H-NS via its N-terminal domain, is also responsible for H-NS modulation[55]. Thus, the Hha/H-NS interaction is most likely regulated by different multimerization and oligomerization states of H-NS and possibly Hha[39]. Finally, H-NS does not show sequence specificity when binding DNA; rather it preferably recognizes curved DNA sequences, a

process commonly referred to as geometric recognition. This is similar to what we have observed for YmgB (Figure 5).

Taken together, our transcriptome, structural and subsequently follow-up nickel enrichment microarray and gel shift assay data strongly suggest that YmgB is a nucleoid-type protein that functions to regulate gene expression in a manner similarly to that of Hha. Notably, the most significant structural difference between these proteins is that YmgB forms a dimer in solution while Hha is monomeric. However, since it is known that a multimer-oligomer balance of Hha/H-NS is critical for the subtle regulation of gene expression observed in response to a variety of differing environmental conditions, the observed YmgB dimer most likely represents a biologically relevant oligomer that plays a critical role in its ability to regulate gene expression, by means of interaction with a H-NS-like protein.

Finally, despite the striking similarity of the 3-dimensional structure between YmgB and Hha that strongly indicates a similar function, our data shows convincingly that they fulfill different regulatory roles. First, the *hha* mutation does not change acid survival in *E. coli* and while Hha is known to regulate numerous environmental conditions, it does not regulate acid-resistance. In addition, deletion of *hha* produces a change in pH of LB glu medium to ~4.5 after 24 h incubation whereas deletion of *ymgB* shows same pH (~8.3) after 24 h as wild type strain. Therefore, while both proteins, as judged by the similarity of their structures, use similar functional mechanisms, the physiological roles of Hha and YmgB are distinct as shown by both the acid survivability and pH phenotypes. This is also strongly supported by follow-up DNA microarray studies of a *hha* deletion mutant that show a very different gene expression pattern compared to the one of the *ymgB* deletion mutant, with only very few coincident genes. Therefore, we predict that a currently unidentified H-NS-like protein, which is similar in its functional behavior of H-NS but distinct in the processes it regulates, plays a major role in the regulation of acid-resistance and oxidative stress.

Material and Methods

Bacterial strains

Strains and plasmids used are listed in Table S1 (supporting information). LB[56] was used to pre-culture all the *E. coli* cells. Indole was obtained from Acros Organics (Geel, Belgium).

Acid-resistance assay

This assay was adapted from Masuda and Church[22]. Overnight cultures grown for 19 h in LB were re-grown to mid-log phase in LB (turbidity at 600 nm of 1), and the culture was diluted 40-fold into phosphate-buffered saline (pH 7.2) or LB (pH 2.5) at 37°C. *E. coli* in LB (pH 2.5) was incubated at 37°C for 1 h without shaking. The percentage of cells surviving the acid treatment was calculated as the number of colonies remaining per mL after acid treatment divided by the initial colonies per mL. Each experiment was performed using two or four independent cultures.

Hydrogen peroxide resistance and complementation assays

Overnight cultures grown for 18 h in LB were re-grown to mid-log phase in LB (turbidity at 600 nm of 1), and 1 mL of each culture was incubated with H_2O_2 at a final concentration of 30 mM at 37°C for 15 min without shaking. The percentage of cells surviving the H_2O_2 treatment was calculated as the number of colony forming units/mL remaining after H_2O_2 treatment divided by the initial colony forming units/mL at time zero. Two independent experiments were conducted.

Crystal violet biofilm assay

This assay was adapted from Pratt and Kolter[24]; cells were grown in polystyrene 96 well plates at 37°C for one day without shaking in LB and LB glu (0.2%) medium. Each data point was averaged from 12 to 24 replicate wells (six wells from each independent culture and two or four independent cultures were used). The data for biofilm assay were analyzed with a Student t-test, and only those with a *p*-value less than or equal to 0.05 were chosen as significant[57].

Motility assay

LB overnight cultures were used to assay motility in plates containing 1% (w/v) tryptone, 0.25% (w/v) NaCl, and 0.3% (w/v) agar[58]. The motility halos were measured at 8 h. When the effect of indole on motility was tested, indole (500 μM) dissolved in dimethylformamide (DMF) was added to the motility agar. DMF (0.1%, v/v) was added as the negative control. Each experiment was performed using two or four independent cultures with each culture evaluated using triplicate plates. Also, 0.1% (v/v) DMF has no effect on motility.

Biofilm total RNA isolation

For the microarray experiments, 10 g of glass wool (Corning Glass Works, Corning, NY) were used to form biofilms[8] in 250 mL in 1 L Erlenmeyer shake flasks which was inoculated with overnight cultures that were diluted 1:167 in LB glu medium. Cells were shaken at 250 rpm at 37°C for 24 h to form biofilms on the glass wool, and RNA was isolated from the biofilm cells as described previously[8].

DNA microarrays

The *E. coli* Genechip antisense genome array (P/N 900381, Affymetrix, Santa Clara, CA) which contains probe sets for all 4290 open reading frames (ORF), rRNA, tRNA, and 1350 intergenic regions was used to study the differential gene expression profile for the *ymgB* mutant compared to the isogenic wild-type K-12 in a mature biofilm as described previously[19]. Hybridization was performed for 16 h and the total cell intensity was scaled automatically in the software to an average value of 500. The data were inspected for quality and analyzed according to the procedures described in Data Analysis Fundamentals which includes using premixed polyadenylated transcripts of the *B. subtilis* genes (*lys*, *phe*, *thr*, *dap*) at different concentrations. Also, as expected, there was insignificant *ymgB* mRNA signal in the biofilm of the *ymgB* mutant (549-fold change), and the completely-deleted *E. coli* K-12 BW25113 genes *araA* and *rhaA* showed insignificant mRNA levels. Genes were identified as differentially expressed if the expression ratio was greater than 2 and the change in p-value was less than 0.05 since the standard deviation for all the genes was 1.2. The gene functions were obtained from the National Center for Biotechnology Information database (http://www.ncbi.nlm.nih.gov/) and from the EcoCyc database[59] (http://biocyc.org/ECOLI/). The expression data for the biofilm samples have been deposited in the NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO Series accession number GSE6925[60; 61].

Nickel-enrichment DNA microarrays

E. coli K-12 AG1/pCA24N and AG1/pCA24N-*ymgB*$^+$ (plasmid pCA24N-*ymgB*$^+$ expresses YmgB with a His_6-tag at the amino terminus[62]) cells were cultured in 250 mL in 1 L Erlenmeyer shake flasks with 10 g glass wool to produce biofilms[8] which were inoculated with overnight cultures that were diluted 1:100. The cells were grown at 37°C with 250 rpm shaking in LB glu with chloramphenicol (30 μg/ mL) to maintain the pCA24N or pCA24N-*ymgB*$^+$ plasmid and 2 mM IPTG to induce YmgB synthesis. After 24 h, formaldehyde (1%, Fisher Scientific Co., Pittsburgh, PA) was added to promote crosslinking between the His-tagged YmgB protein and the DNA to which it was bound[42], and the cultures were incubated another

20 min with 100 rpm shaking at room temperature. The crosslinking was stopped by adding 0.125 M glycine (Sigma, St. Louis, MO) for 5 min with shaking (100 rpm). The glass wool was washed twice with 200 mL of cold (50 mM NaH_2PO_4, 0.5 M NaCl, pH 8) buffer, and the washed cells were sonicated in a bath (Fisher Scientific; Model: FS3) in 200 mL of the same buffer for 2 min. The cells were harvested by centrifugation, and the cell pellets were suspended in 25 mL of the same buffer with 1 mM phenylmethylsulfonyl fluoride (PMSF) (Pierce Biotechnology, Rockford, IL) to preserve the integrity of the proteins. The cells were lysed by sonication at a power level of 10 watts using 3 pulses of 1 min (Model 60 Dismembrator, Fisher Scientific), and the lysate was centrifuged at 10,000 rpm, 4°C for 20 min to pellet the cellular debris. 10 mL of Ni-NTA agarose gel resin (Invitrogen, San Diego, CA) were added to the supernatant which was incubated overnight at 4°C with 150 rpm shaking in order to bind the His_6-tagged YmgB-DNA complex to the resin. The YmgB-DNA complexes were purified following the manufacturer protocol for the regular purification of His_6-tagged proteins (Invitrogen), and the purified complexes were eluted with 10 mL of buffer 50 mM NaH_2PO_4 pH 8, 0.5 M NaCl and 250 mM imidazole (Fisher Scientific). The eluate was incubated 5 h at 70°C with 75 μg/mL RNase A (Fisher Scientific) and 0.5 M NaCl, then the samples were precipitated overnight at −20°C in 2.5 volume of ethanol, centrifuged at 15,000 rpm for 25 min at 4°C and the pellet was dissolved in 1.8 mL TE buffer (50 mM Tris-Cl pH 7.5, 25 mM EDTA) with 100 μg of Proteinase K (Fisher Scientific) and incubated for 2 h at 45°C to degrade protein. The protein- and RNA-free DNA fragments were then purified using the DNA QIAquick kit (Qiagen, Valencia, CA) and resuspended in 150 μL of sterile water; 600 ng of a 357 bp PCR product from the gene *rhaD* was added in order to provide a spiked positive control for the microarrays. The presence of 100 to 1000 bp DNA fragments was confirmed by gel electrophoresis, the DNA fragments were labeled as described previously[8], and DNA microarrays were performed as indicated above.

For analysis of the data, we chose as positive candidates those genes and intergenic regions with at least a 4-fold higher signal than the global average signal (~588) of all the genes and IG regions in the YmgB chip and those that were at least 7-fold enriched with respect to the signal they showed in a control array (using an empty pCA24N vector); this cut off ratio was selected because the standard deviation of the enrichment (YmgB chip signal divided by empty vector signal) was 6.1. The spiked control *rhaD* gene gave a signal of 11,129 (19-fold higher than the average signal) which validates our methodology.

Electrophoretic mobility shift assay (EMSA)

EMSA mobility shift assays were performed as previously reported[20], using purified 9 kDa (full-length) and 7.7 kDa YmgB proteins were used along with PCR-amplified products containing the *gadA* promoter region (294 bp, consisting of 285 bp upstream and 9 bp downstream the start codon of *gadA*), the *rpsV* promoter region (235 bp, beginning 242 bp upstream of the start codon), the *ycfR* promoter region (262 bp, beginning 259 bp upstream of the start codon), the *crp* promoter region (216 bp, beginning 93 bp upstream of the start codon), and a non-promoter region of the *lsrF* gene (202 bp, beginning 48 bp upstream of the start codon); these DNA regions were amplified from genomic DNA of the wild-type strain BW25113 with the primers 5'-GAT GTG GAT GAT ATC GTA-3'/5'-CTG GTC ATT TCG AAC TC-3', 5'-GCA CAG CAT GGT GTT GTC-3'/5'-CCT CAA TCC TGT AGC TAG -3', 5'-GTG TTG AGT CAG TTG CCA-3'/5'-CAT AAT AGT GGC CTT ATG-3', 5'-GCT TGC ATT TTT GCT ACT-3'/5'-GCT ATC AAC TGT ACT GCA -3', and 5'-ATG TGG TTC AGC AAG GCA -3'/5'-TTC TCT TTG TTG AAT ATC -3', respectively.

Expression and purification

The DNA sequence of YmgB (residues 1-88) was subcloned into a His_6-tagged bacterial expression vector ($Thio_6His_6$-TEV-)[63]. This construct was transformed into the *E. coli* strain

BL21-CodonPlus (DE3)-RIL (Stratagene, La Jolla, CA) and a single colony used to inoculate a 100 mL culture of LB containing kanamycin (50 μg/mL) and chloramphenicol (34 μg/mL). The expression of uniformly ^{15}N-labeled protein was carried out by growing freshly transformed cells in M9 minimal medium containing 1 g/L $^{15}NH_4Cl$ as the sole nitrogen source. The expression of selenomethionine-labeled protein was carried out by growing freshly transformed cells in selenomethionine-containing medium. Unlabeled culture was grown overnight at 37°C with shaking at 250 rpm. The next morning the cells were diluted 1:50 into fresh LB medium with appropriate antibiotics. Cells were grown at 37°C until they reached an OD_{600} of 0.6-0.9. The cultures were then placed at 4°C and the shaker temperatures adjusted to 18°C. Cultures were induced with 1 mM IPTG and allowed to express overnight (~18 hours) at 18°C under vigorous shaking (250 rpm). Cultures were harvested by centrifugation and the pellets stored at −80°C until purification.

For purification, the pellets were resuspended in lysis buffer (50 mM Tris pH 8.0, 5 mM imidazole, 500 mM NaCl, 0.1% Triton-X, Complete tabs-EDTA free (Roche Indianapolis, IN)). The cells were lysed by three passes through a C3 Emulsiflex cell cracker (Avestin, Canada) and the cell debris removed by centrifugation (40000 g/40 min/4°C). The clarified lysate was filtered through a 0.22 μm membrane (Millipore, Billerica, MA) and loaded onto a HisTrap HP column (GE Healthcare, Piscataway, NJ) equilibrated with 50 mM Tris pH 8.0, 5 mM imidazole, and 500 mM NaCl. The protein was eluted with a 5-500 mM Imidazole gradient. The fractions containing the target protein were identified by SDS-PAGE gel electrophoresis and were pooled and dialyzed against 50 mM Tris pH 7.5, 250 mM NaCl at 4° C. Tobacco etch virus (TEV) NIa protease was used to cleave the purification tag. Cleavage, verified by SDS-PAGE gel electrophoresis, was achieved overnight at room temperature under steady rocking. The cleaved sample was then dialyzed against 50 mM Tris pH 7.5, 250 mM NaCl overnight and further purified by a second His_6-tag purification step. The cleaved, now untagged YmgB protein was collected in the flow through. The protein was concentrated and diluted 1:1 with crystallization buffer (20 mM Tris, pH 7.8, 100 mM NaCl) and purified using size exclusion chromatography (Sephadex 75 26/60, GE Healthcare) equilibrated in crystallization buffer. The protein elutes as single peak. The eluted fractions were pooled, the sample concentrated to 7.5 mg/mL and either stored at 4°C or used immediately for crystallization trials.

Crystallization

Initial crystallization trials were carried out by the high-throughput crystallization laboratory at the Hauptmann-Woodward Institute[64]. Crystals were observed in 138 of the 1536 crystallization conditions screened. The most obvious trend of these conditions was a pH value between 7 and 10 and the presence of high concentrations of salt. A subset of these conditions was screened using microbatch crystallization (0.5 μL protein + 0.5 μL crystallization condition under 15 μL paraffin oil) at room temperature. Two of these conditions (*a*: 0.1 M bicine, pH 9.0, 1.0 M lithium chloride; *b*: 0.1 M Tris, pH 8.0, 2.0 M sodium chloride) produced well-formed protein crystals, which diffracted to ~7 Å, independent of cryoprotection protocol. Additive screening led to the discovery that crystallization of YmgB in condition *b* in the presence of b-octyl-b-D-glucopyranoside (b-OG) at a final concentration of 0.5% (w/v) resulted in diffraction quality crystals. Final crystallization conditions were 0.09 M Tris, pH 8.0, 1.8 M NaCl, 0.5% (w/v) b-OG. Crystals were transferred into cryoprotection buffer (crystallization solution with 25% (v/v) glycerol) and flash frozen in liquid nitrogen. It was determined later (see below) that diffraction quality crystals were only obtained with proteolytically cleaved protein, in which the N-terminus had been cleaved, resulting in a shorter C-terminal YmgB protein fragment (~7 kD instead of the full-length 9.9 kD).

Data Collection and Structure Determination

Data were collected at NSLS Beamline X6A at 100 K using a 210 CCD detector (Area Detector Systems Corporation, Poway, CA). Data for all crystals were indexed and scaled using HKL2000[65]. Pointless[66] was used to facilitate the identification of the correct space group. It reported virtually identical probabilities for P4222, C2221, P2221 and P42 (0.847, 0.842, 0.842 and 0.713, respectively). However, phases were only obtained only with data processed as C2221 (or P2221) and not in P42 or P4222. In addition, the phased model was later used as a template for molecular replacement against data processed in P42 and P4222; this also failed repeatedly. Thus, it was concluded that the true space group is C2221.

The structure of the YmgB was phased to 2.0 Å using a 3 wavelength SeMet MAD data set using SOLVE (mean FOM = 0.70)[67]. Only one selenium atom (of the expected three) per monomer was identified, which corresponded to residue *Se*MET52. RESOLVE[68] was immediately used for density modification and automatically built 108 residues of the expected 180 for two copies of YmgB in the asymmetric unit (FOM = 0.77). Careful inspection of the phased and density modified electron density maps suggested that the N-terminal ~25 amino acids of YmgB were not present in protein that had crystallized, as there was no space for these residues in the asymmetric unit. MALDI-TOF mass spectrometry (see below) confirmed that the YmgB protein that crystallized was indeed a mixture of C-terminal proteolytic fragments that included residues 23-88 to 25-88. ARP/wARP[69] was used for automatic model building against the density modified phases and resulted in a model for YmgB with 60 residues in monomer 1 (YmgB 26-85) and 61 residues in monomer 2 (YmgB 27-85). Data statistics are summarized in Table 1. The model was completed by cycles of manual building in Coot[70] coupled with structure refinement using RefMac71 against the 1.8 Å resolution far wavelength dataset. The final model of YmgB includes residues 25-86 of both monomers. Analysis of the stereochemical quality of the models was carried out using MolProbity[29] and the JCSG Structure Validation Central suite (www.jcsg.org) which integrates seven validation tools: Procheck 3.5.4, SFcheck 4.0, Prove 2.5.1, ERRAT, WASP, DDQ 2.0, and Whatcheck. Atomic coordinates of the final model and experimental structure factors of have been deposited with the PDB and are accessible under the code 2OXL.

Mass Spectrometry

Crystals from the same plate as that used for structure determination were harvested and washed (3x) with crystallization buffer. Crystals were then transferred to water, leading to their dissolution. To determine the time course of proteolysis, fresh YmgB protein was prepared as described and split into two samples, one of which contained proteases inhibitors (a protease cocktail (Roche)) and one which did not. At various time points (0, 2, 6, 8, 10, 13, and 38 days), 10 μL were removed, ethanol precipitated and stored at −20°C. Protein was resuspended in water immediately prior to mass spectrometry screening. Samples are diluted 1:1 with matrix solution (10 mg sinapinic acid in 30% acetonitrile with 0.1% trifluoroacetic acid in water) and 1 μL plated on a MALDI laser-etched stainless steel sample plate, allowed to dry and examined using an Voyager DE-Pro matrix-assisted laser desorption ionization time of flight mass spectrometer (MALDI-TOF-MS) (Applied Biosystems, Foster City, CA). BSA was used as a calibration standard.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the NIH (EB003872-0IAI) to T.K.W. and Brown University start-up funds to W.P. and Brown University start-up funds to R.P. The authors would like to thank D. Reid for help with a portion of the

crystallization experiments and G. Stetson with protein purification. W.P. is the Manning Assistant Professor for Medical Science at Brown University.

References

1. Kolter R, Losick R. One for all and all for one. Science 1998;280:226–7. [PubMed: 9565532]
2. Sauer K, Camper AK. Characterization of phenotypic changes in Pseudomonas putida in response to surface-associated growth. J Bacteriol 2001;183:6579–89. [PubMed: 11673428]
3. Nickel JC, Ruseska I, Wright JB, Costerton JW. Tobramycin resistance of Pseudomonas aeruginosa cells growing as a biofilm on urinary catheter material. Antimicrob Agents Chemother 1985;27:619–24. [PubMed: 3923925]
4. Stanley NR, Lazazzera BA. Environmental signals and regulatory pathways that influence biofilm formation. Mol Microbiol 2004;52:917–24. [PubMed: 15130114]
5. Costerton B. Microbial ecology comes of age and joins the general ecology community. Proc Natl Acad Sci U S A 2004;101:16983–4. [PubMed: 15572449]
6. Whiteley M, Bangera MG, Bumgarner RE, Parsek MR, Teitzel GM, Lory S, Greenberg EP. Gene expression in Pseudomonas aeruginosa biofilms. Nature 2001;413:860–4. [PubMed: 11677611]
7. Beloin C, Valle J, Latour-Lambert P, Faure P, Kzreminski M, Balestrino D, Haagensen JA, Molin S, Prensier G, Arbeille B, Ghigo JM. Global impact of mature biofilm lifestyle on Escherichia coli K-12 gene expression. Mol Microbiol 2004;51:659–74. [PubMed: 14731270]
8. Ren D, Bedzyk LA, Thomas SM, Ye RW, Wood TK. Gene expression in *Escherichia coli* biofilms. Appl Microbiol Biotechnol 2004;64:515–24. [PubMed: 14727089]
9. Schembri MA, Kjaergaard K, Klemm P. Global gene expression in *Escherichia coli* biofilms. Mol Microbiol 2003;48:253–67. [PubMed: 12657059]
10. Domka J, Lee J, Bansal T, Wood TK. Temporal gene-expression in *Escherichia coli* K-12 biofilms. Environ Microbiol 2007;9:332–46. [PubMed: 17222132]
11. Sauer K, Camper AK, Ehrlich GD, Costerton JW, Davies DG. Pseudomonas aeruginosa displays multiple phenotypes during development as a biofilm. J Bacteriol 2002;184:1140–54. [PubMed: 11807075]
12. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. The complete genome sequence of Escherichia coli K-12. Science 1997;277:1453–74. [PubMed: 9278503]
13. Herzberg M, Kaye IK, Peti W, Wood TK. YdgG (TqsA) controls biofilm formation in *Escherichia coli* K-12 through autoinducer 2 transport. J Bacteriol 2006;188:587–98. [PubMed: 16385049]
14. Ren D, Bedzyk LA, Ye RW, Thomas SM, Wood TK. Differential gene expression shows natural brominated furanones interfere with the autoinducer-2 bacterial signaling system of *Escherichia coli*. Biotechnol Bioeng 2004;88:630–42. [PubMed: 15470704]
15. Lee J, Jayaraman A, Wood TK. Indole is an inter-species biofilm signal mediated by SdiA. BMC Microbiol 2007;7:42. [PubMed: 17511876]
16. Lombardia E, Rovetto AJ, Arabolaza AL, Grau RR. A LuxS-dependent cell-to-cell language regulates social behavior and development in *Bacillus subtilis*. J Bacteriol 2006;188:4442–52. [PubMed: 16740951]
17. Lee J, Jayaraman A, Wood TK. Indole and acyl-homoserine lactones are is an inter-species Escherichia coli biofilm signals mediated by SdiA. BMC Microbiol. 2007in revision
18. Ma Z, Gong S, Richard H, Tucker DL, Conway T, Foster JW. GadE (YhiE) activates glutamate decarboxylase-dependent acid resistance in *Escherichia coli* K-12. Mol Microbiol 2003;49:1309–20. [PubMed: 12940989]
19. González Barrios AF, Zuo R, Hashimoto Y, Yang L, Bentley WE, Wood TK. Autoinducer 2 controls biofilm formation in *Escherichia coli* through a novel motility quorum-sensing regulator (MqsR, B3022). J Bacteriol 2006;188:305–16. [PubMed: 16352847]
20. Zhang XS, García-Contreras R, Wood TK. YcfR (BhsA) influences *Escherichia coli* biofilm formation through stress response and surface hydrophobicity. J Bacteriol 2007;189:3051–3062. [PubMed: 17293424]

21. Domka J, Lee J, Wood TK. YliH (BssR) and YceP (BssS) regulate *Escherichia coli* K-12 biofilm formation by influencing cell signaling. Appl Environ Microbiol 2006;72:2449–59. [PubMed: 16597943]
22. Masuda N, Church GM. Regulatory network of acid resistance genes in *Escherichia coli*. Mol Microbiol 2003;48:699–712. [PubMed: 12694615]
23. Wang D, Ding X, Rather PN. Indole can act as an extracellular signal in Escherichia coli. J Bacteriol 2001;183:4210–6. [PubMed: 11418561]
24. Pratt LA, Kolter R. Genetic analysis of *Escherichia coli* biofilm formation: roles of flagella, motility, chemotaxis and type I pili. Mol Microbiol 1998;30:285–93. [PubMed: 9791174]
25. Foster JW. Escherichia coli acid resistance: tales of an amateur acidophile. Nat Rev Microbiol 2004;2:898–907. [PubMed: 15494746]
26. Zheng M, Wang X, Templeton LJ, Smulski DR, LaRossa RA, Storz G. DNA microarray-mediated transcriptional profiling of the *Escherichia coli* response to hydrogen peroxide. J Bacteriol 2001;183:4562–70. [PubMed: 11443091]
27. Matthews BW. Solvent content of protein crystals. J Mol Biol 1968;33:491–7. [PubMed: 5700707]
28. Kantardjieff KA, Rupp B. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. Protein Sci 2003;12:1865–71. [PubMed: 12930986]
29. Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins 2003;50:437–50. [PubMed: 12557186]
30. von Freyberg B, Richmond TJ, Braun W. Surface area included in energy refinement of proteins. A comparative study on atomic solvation parameters. J Mol Biol 1993;233:275–92. [PubMed: 7690855]
31. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–98. [PubMed: 9925793]
32. Credle JJ, Finer-Moore JS, Papa FR, Stroud RM, Walter P. On the mechanism of sensing unfolded protein in the endoplasmic reticulum. Proc Natl Acad Sci U S A 2005;102:18773–84. [PubMed: 16365312]
33. 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;25:3389–402. [PubMed: 9254694]
34. Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 2005;33:W89–93. [PubMed: 15980588]
35. Yee A, Chang X, Pineda-Lucena A, Wu B, Semesi A, Le B, Ramelot T, Lee GM, Bhattacharyya S, Gutierrez P, Denisov A, Lee CH, Cort JR, Kozlov G, Liao J, Finak G, Chen L, Wishart D, Lee W, McIntosh LP, Gehring K, Kennedy MA, Edwards AM, Arrowsmith CH. An NMR approach to structural proteomics. Proc Natl Acad Sci U S A 2002;99:1825–30. [PubMed: 11854485]
36. Rico M, Jimenez MA, Gonzalez C, De Filippis V, Fontana A. NMR solution structure of the C-terminal fragment 255-316 of thermolysin: a dimer formed by subunits having the native structure. Biochemistry 1994;33:14834–47. [PubMed: 7993910]
37. Dorman CJ. H-NS: a universal regulator for a dynamic genome. Nat Rev Microbiol 2004;2:391–400. [PubMed: 15100692]
38. Dorman CJ. H-NS, the genome sentinel. Nat Rev Microbiol 2007;5:157–61. [PubMed: 17191074]
39. Madrid C, Balsalobre C, Garcia J, Juarez A. The novel Hha/YmoA family of nucleoid-associated proteins: use of structural mimicry to modulate the activity of the H-NS family of proteins. Mol Microbiol 2007;63:7–14. [PubMed: 17116239]
40. Nieto JM, Madrid C, Miquelay E, Parra JL, Rodriguez S, Juarez A. Evidence for direct protein-protein interaction between members of the enterobacterial Hha/YmoA and H-NS families of proteins. J Bacteriol 2002;184:629–35. [PubMed: 11790731]
41. Rodriguez S, Nieto JM, Madrid C, Juarez A. Functional replacement of the oligomerization domain of H-NS by the Hha protein of Escherichia coli. J Bacteriol 2005;187:5452–9. [PubMed: 16030239]

42. Tamimi Y, Lines M, Coca-Prados M, Walter MA. Identification of target genes regulated by FOXC1 using nickel agarose-based chromatin enrichment. Invest Ophthalmol Vis Sci 2004;45:3904–13. [PubMed: 15505035]
43. Perrenoud A, Sauer U. Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in *Escherichia coli*. J Bacteriol 2005;187:3171–9. [PubMed: 15838044]
44. Hanamura A, Aiba H. A new aspect of transcriptional control of the *Escherichia coli crp* gene: positive autoregulation. Mol Microbiol 1992;6:2489–97. [PubMed: 1328816]
45. Taga ME, Miller ST, Bassler BL. Lsr-mediated transport and processing of AI-2 in *Salmonella typhimurium*. Mol Microbiol 2003;50:1411–27. [PubMed: 14622426]
46. Kim MK, Park SR, Cho SJ, Lim WJ, Ryu SK, An CL, Hong SY, Park YW, Kahng GG, Kim JH, Kim H, Yun HD. The effect of a disrupted *yhjQ* gene on cellular morphology and cell growth in *Escherichia coli*. Appl Microbiol Biotechnol 2002;60:134–8. [PubMed: 12382054]
47. Zogaj X, Nimtz M, Rohde M, Bokranz W, Romling U. The multicellular morphotypes of *Salmonella typhimurium* and *Escherichia coli* produce cellulose as the second component of the extracellular matrix. Mol Microbiol 2001;39:1452–63. [PubMed: 11260463]
48. Izutsu K, Wada C, Komine Y, Sako T, Ueguchi C, Nakura S, Wada A. *Escherichia coli* ribosome-associated protein SRA, whose copy number increases during stationary phase. J Bacteriol 2001;183:2765–73. [PubMed: 11292794]
49. Francez-Charlot A, Castanie-Cornet MP, Gutierrez C, Cam K. Osmotic regulation of the Escherichia coli bdm (biofilm-dependent modulation) gene by the RcsCDB His-Asp phosphorelay. J Bacteriol 2005;187:3873–7. [PubMed: 15901715]
50. Hagiwara D, Sugiura M, Oshima T, Mori H, Aiba H, Yamashino T, Mizuno T. Genome-wide analyses revealing a signaling network of the RcsC-YojN-RcsB phosphorelay system in *Escherichia coli*. J Bacteriol 2003;185:5735–46. [PubMed: 13129944]
51. Lesk, AM. Introduction into Protein Architecture. Oxford University Press; 2001.
52. Kawabata T. MATRAS: A program for protein 3D structure comparison. Nucleic Acids Res 2003;31:3367–9. [PubMed: 12824329]
53. Juarez A, Nieto JM, Prenafeta A, Miquelay E, Balsalobre C, Carrascal M, Madrid C. Interaction of the nucleoid-associated proteins Hha and H-NS to modulate expression of the hemolysin operon in Escherichia coli. Adv Exp Med Biol 2000;485:127–31. [PubMed: 11109097]
54. Nieto JM, Madrid C, Prenafeta A, Miquelay E, Balsalobre C, Carrascal M, Juarez A. Expression of the hemolysin operon in *Escherichia coli* is modulated by a nucleoid-protein complex that includes the proteins Hha and H-NS. Mol Gen Genet 2000;263:349–58. [PubMed: 10778755]
55. Garcia J, Cordeiro TN, Nieto JM, Pons I, Juarez A, Pons M. Interaction between the bacterial nucleoid associated proteins Hha and H-NS involves a conformational change of Hha. Biochem J 2005;388:755–62. [PubMed: 15720293]
56. Sambrook, J.; Fritsch, EF.; Maniatis, T. Molecular Cloning, A Laboratory Manual. 2 edit. Cold Spring Harbor Laboratory Press; Cold Spring Harbor, NY: 1989.
57. Ross, SM. Introduction to Probability and Statistics for Engineers and Scientists. Elsevier Academic Press; USA: 2004.
58. Sperandio V, Torres AG, Kaper JB. Quorum sensing *Escherichia coli* regulators B and C (QseBC): a novel two-component regulatory system involved in the regulation of flagella and motility by quorum sensing in *E. coli*. Mol Microbiol 2002;43:809–21. [PubMed: 11929534]
59. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD. EcoCyc: a comprehensive database resource for *Escherichia coli*. Nucleic Acids Res 2005;33:D334–7. [PubMed: 15608210]
60. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R. NCBI GEO: mining millions of expression profiles-database and tools. Nucleic Acids Res 2005;33:D562–6. [PubMed: 15608262]
61. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10. [PubMed: 11752295]

62. Kitagawa M, Ara T, Arifuzzaman M, Ioka-Nakamichi T, Inamoto E, Toyonaga H, Mori H. Complete set of ORF clones of *Escherichia coli* ASKA library (A Complete Set of *E. coli* K-12 ORF Archive): Unique Resources for Biological Research. DNA Res 2005;12:291–9. [PubMed: 16769691]

63. Peti W, Page R. Strategies to maximize heterologous protein expression in Escherichia coli with minimal cost. Protein Expr Purif 2007;51:1–10. [PubMed: 16904906]

64. Luft JR, Collins RJ, Fehrman NA, Lauricella AM, Veatch CK, DeTitta GT. A deliberate approach to screening for initial crystallization conditions of biological macromolecules. J Struct Biol 2003;142:170–9. [PubMed: 12718929]

65. Otwinowski Z, Minor W. Processing of X-ray Diffraction Data Collected in Oscillation Mode. Methods in Enzym (part A) 1997;276:307–326.

66. Evans P. Scaling and assessment of data quality. Acta Crystallogr D Biol Crystallogr 2006;62:72–82. [PubMed: 16369096]

67. Terwilliger TC, Berendzen J. Automated MAD and MIR structure solution. Acta Crystallogr D 1999;55:849–861. [PubMed: 10089316]

68. Terwilliger TC. Maximum-likelihood density modification. Acta Crystallogr D 2000;56:965–972. [PubMed: 10944333]

69. Lamzin, VS.; Perrakis, A.; Wilson, KS.; Rossman, MG.; Arnold, E., editors. Vol. F. International Tables for Crystallography. Volume F: Crystallography of biological macromolecules. Kluwer Academic Publishers; Dordrecht, Netherlands: 2001. The ARP/wARP suite for automated construction and refinement of protein models.

70. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 2004;60:2126–32. [PubMed: 15572765]

71. Murshudov GN, Vagin AA, Dodson EJ. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. Acta Crystallogr D 1997;53:240–255. [PubMed: 15299926]

72. Koradi R, Billeter M, Wuthrich K. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 1996;14:51–5. [PubMed: 8744573]29-32

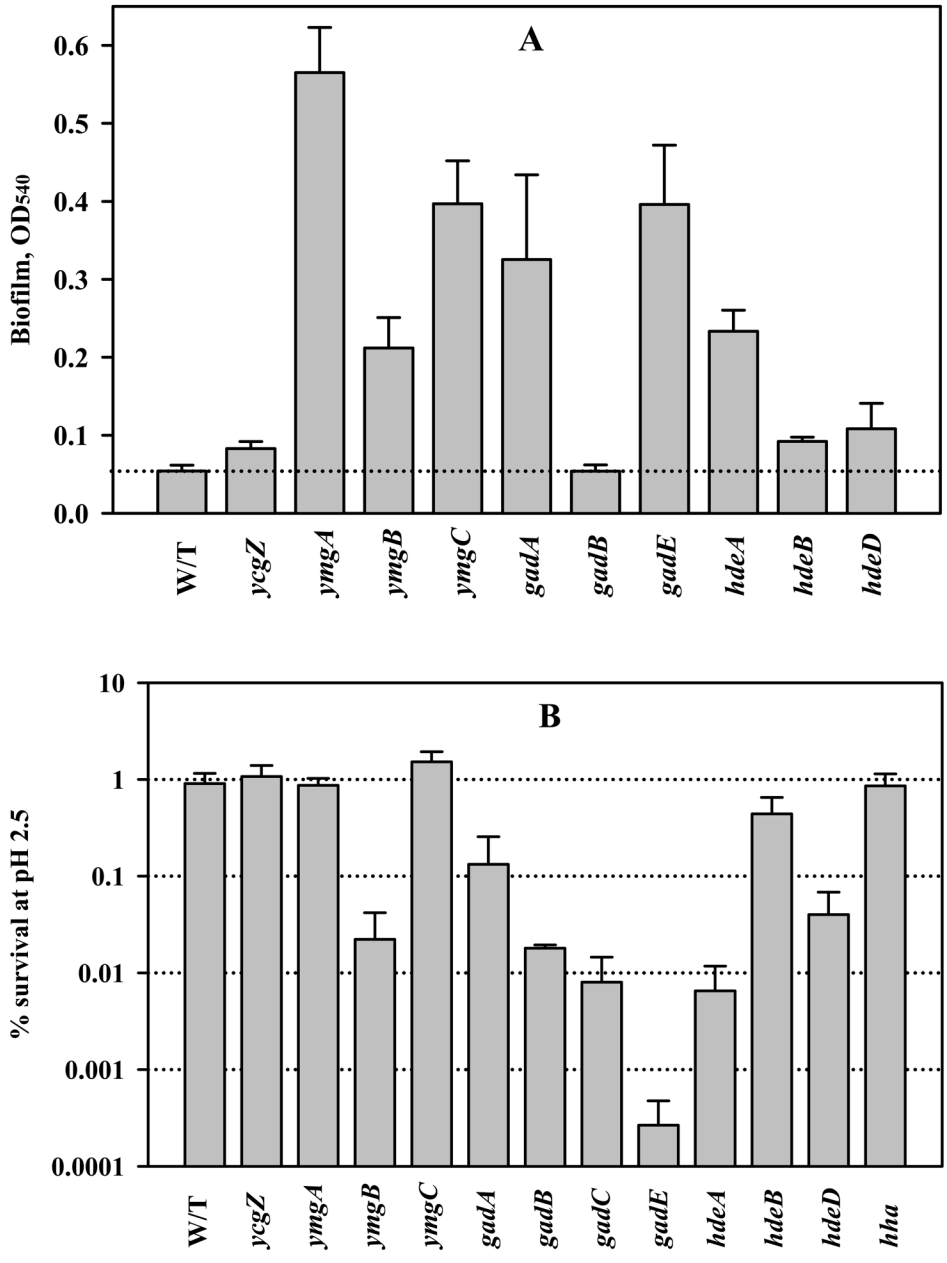

Figure 1.

A) Biofilm formation of BW25113 wild-type (W/T) and various knockout mutants in LB glu medium at 24 h at 37°C. For both graphs, each experiment was repeated between 2-4 times and the experimental standard deviations are shown. B) Acid-resistance of *E. coli* BW25113 wild-type (W/T) and various knockout mutants in LB medium (pH 2.5) at 37°C for 1 h. *hde* and *gad* are known acid-resistance gene loci in *E. coli* and used for comparison.

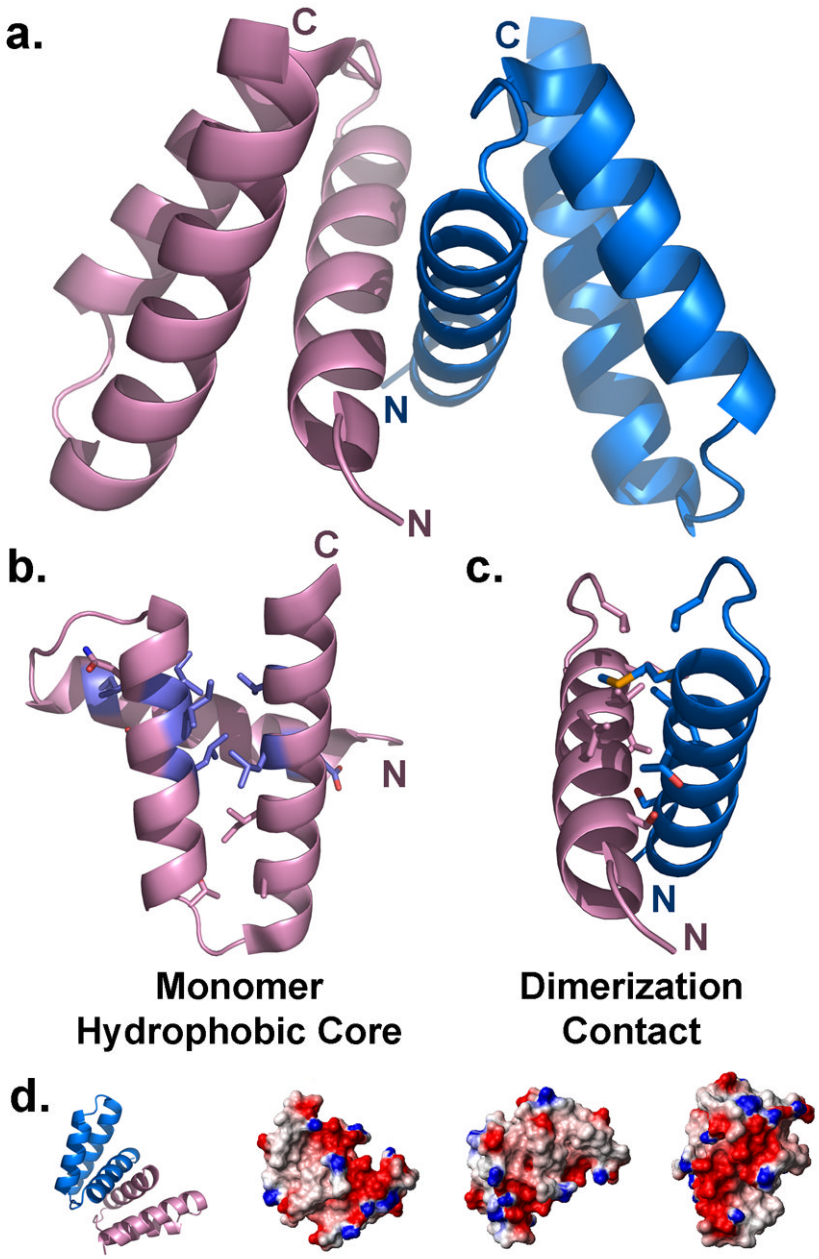

Figure 2.

Structure of YmgB. a) Structure of the YmgB dimer. Chain A, pink, chain B, blue. N-and C-termini are labeled. b) Residues (shown as sticks) of the YmgB monomer which form the hydrophobic core, as determined by a loss of solvent accessible surface area (ASA, GETAREA1.1). Residues which are purple are conserved in the YmgB family of proteins (see Figure 3). c) Residues (shown as sticks) which are buried upon dimerization and constitute the dimerization interface. 1326 $Å^2$ of ASA is buried upon dimer formation. d) Electrostatic surface of the YmgB dimer as prepared with MolMol[72]. The YmgB dimer is characterized by an extensive acidic patch (red), which extends around the long axis of the dimer. Small basic

patches (blue) punctuate the remainder of the surface. The initial orientation of the dimer surface is show as a ribbon model on the left.

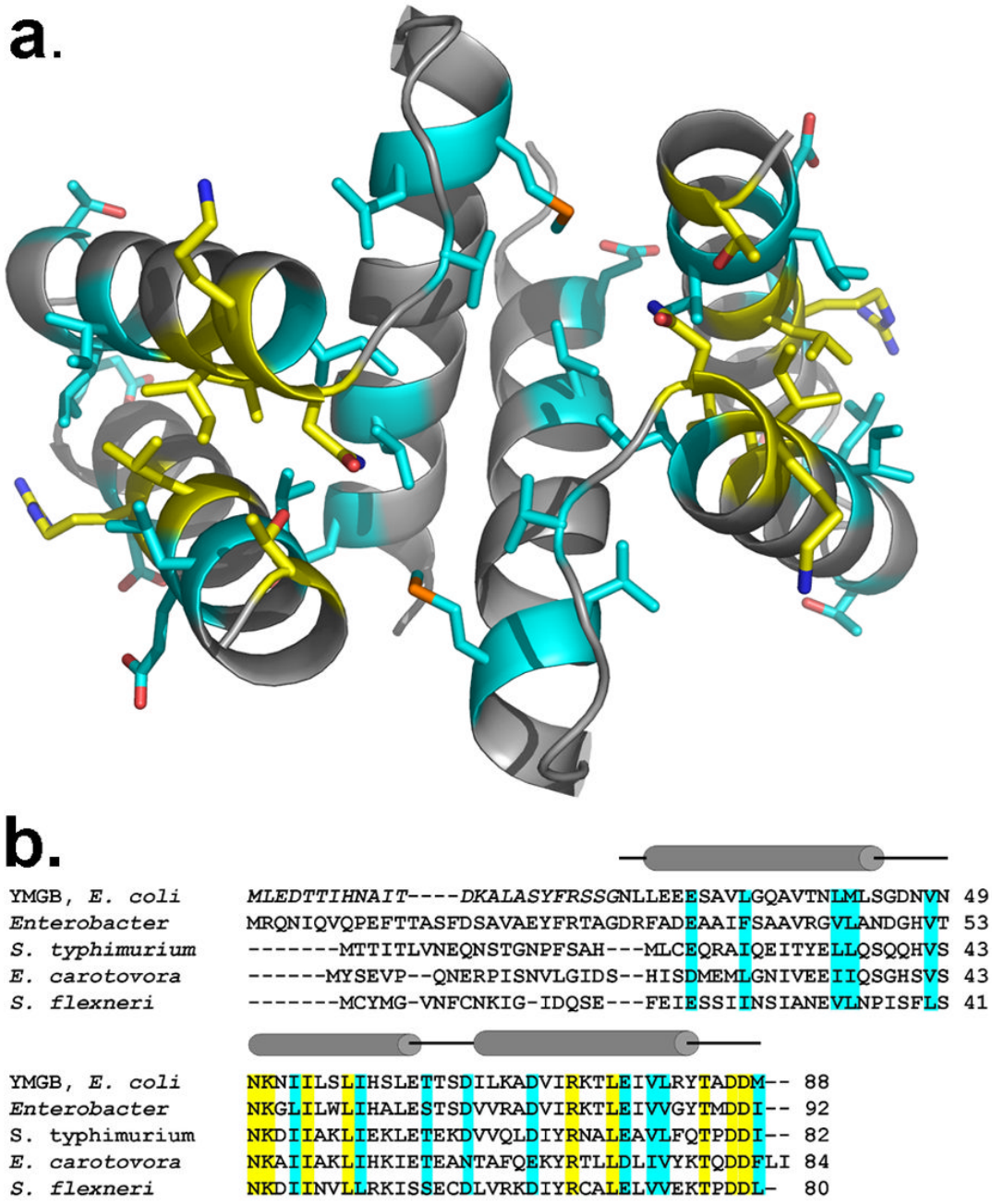

Figure 3.

Sequence alignment of the four proteins identified by position-specific iterated (PSI) BLAST to be most similar to YmgB (NR database: all non-redundant GenBank CDS translations, RefSeq Proteins, PDB, SwissProt, PIR, PRFdatabases). Four proteins with similarities to YmgB were identified: 1. conserved hypothetical protein, *Enterobacter sp. 638*, 51% sequence identity; 2. putative cytoplasmic protein, *S. typhimurium LT2*, 41% sequence identify; 3. hypothetical protein, ECA1764 *E. carotovora*, 36% sequence identify; 4. hypothetical 9.1 kDa protein in spaS 3'region, *S. flexneri*, 27% sequence identify. YmgB α– helices are illustrated with grey cylinders above the sequence. YmgB residues proteolytically removed prior to crystal formation and not present in the structure are italicized. Identical residues are

highlighted in yellow, while similar residues are highlighted in cyan: a) identical and similar residues mapped onto the structure of YmgB. b) sequence alignment.

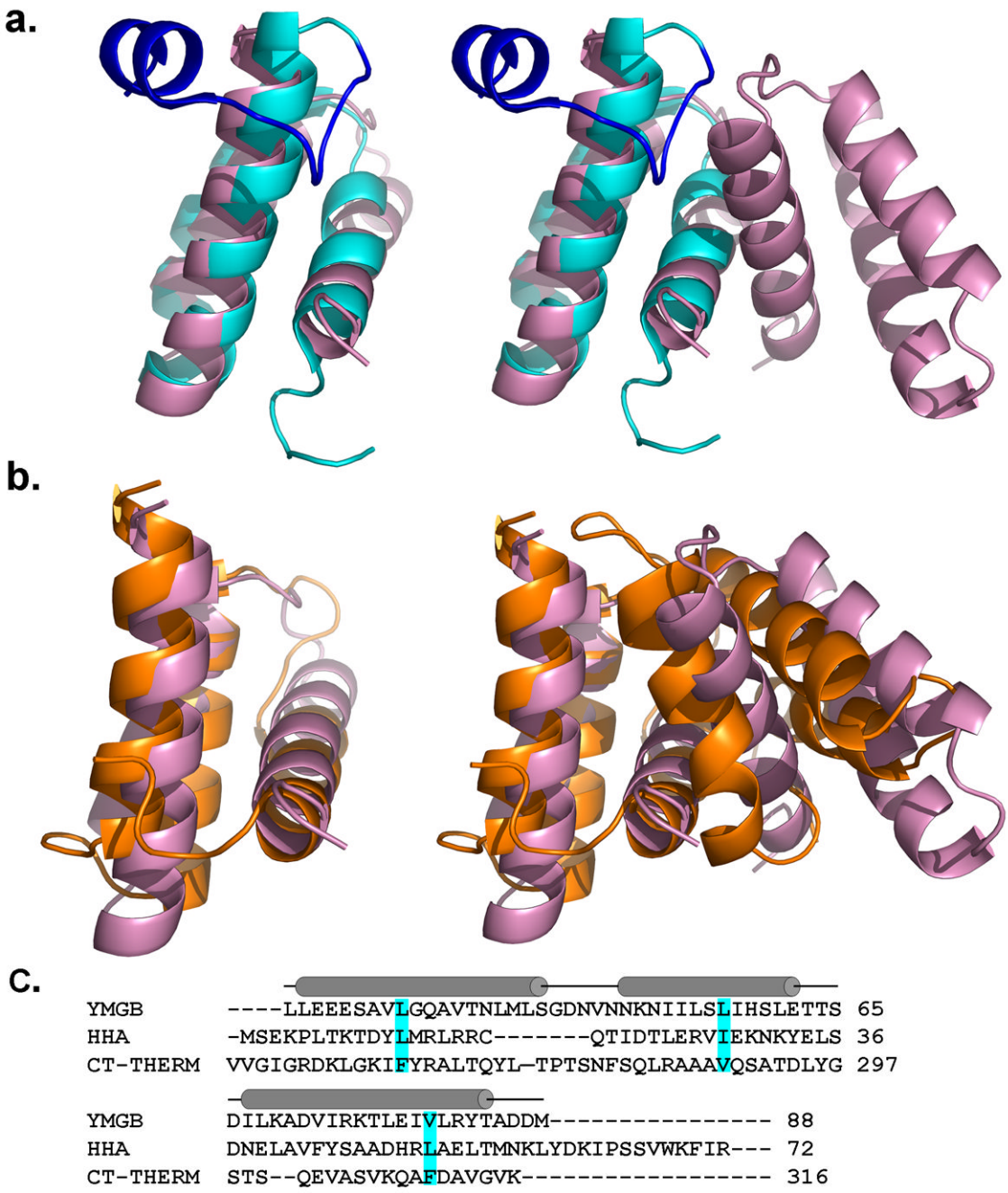

Figure 4.

Superposition of YmgB with Hha and the C-terminal dimerization domain of thermolysin (CT-THERM). Left, superposition onto YmgB monomer; right, superposition onto YmgB dimer. a.) YmgB (pink) superimposed onto the structure of the *E. coli* protein Hha (cyan). The C-terminal loop of Hha (dark blue) is expected to inhibit Hha dimerization (full-length Hha is monomeric) via a steric clash, as the loop would collide with the C-terminal end of helix α1 of the second monomer. b.) YmgB superimposed onto the structure of the CT-THERM (orange). The dimerization interface of thermolysin (over 3000 $Å^2$ ASA) is much more extensive than that of YmgB because the second monomer is associated more intimately with

the first than YmgB, as can be seen from the dimer superposition (left). c. The structure-based sequence alignment of YmgB, Hha and CT-THERM; residues highlighted in cyan are similar. There are no identical residues between the three proteins.

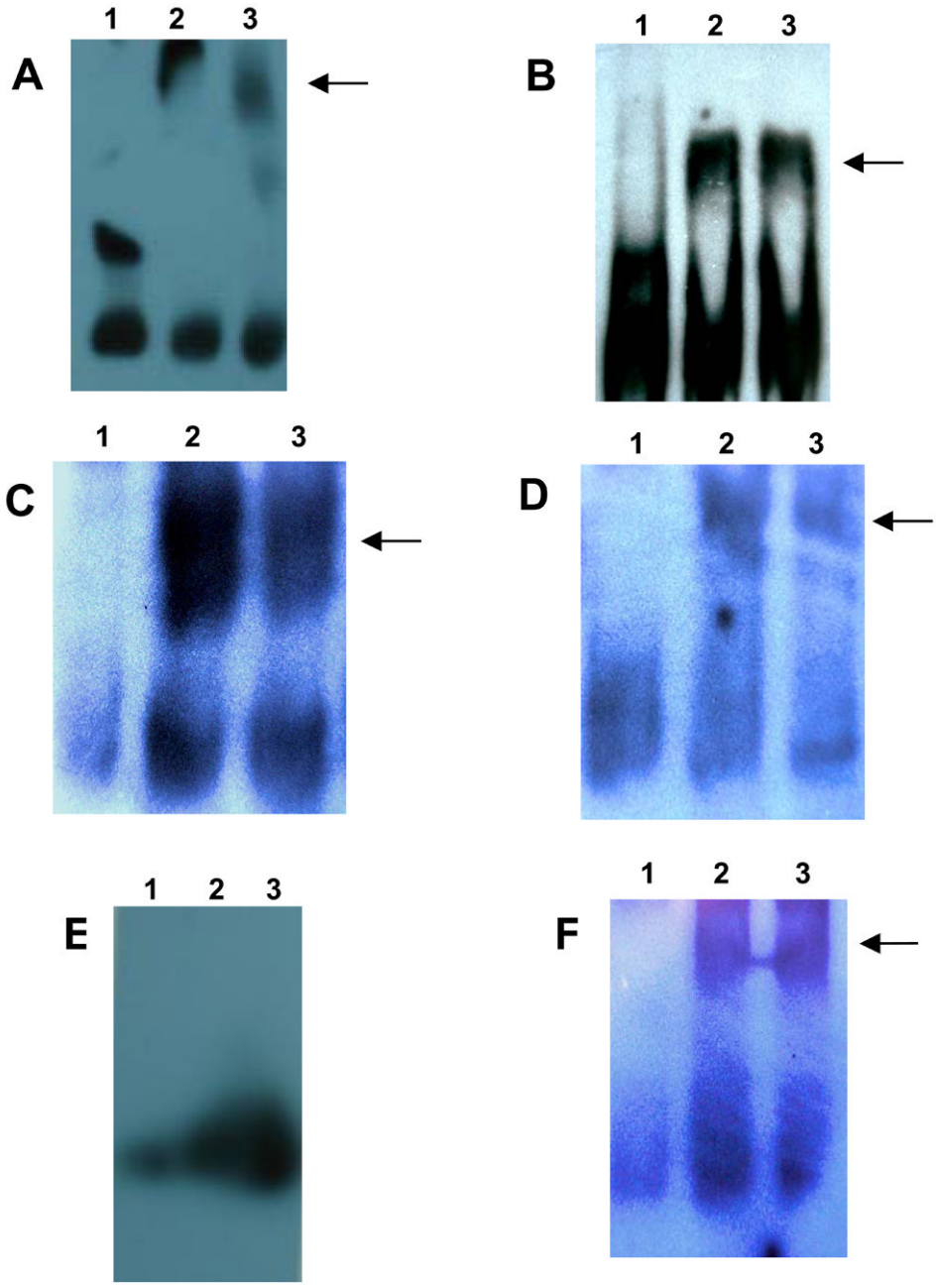

Figure 5.

DNA-binding assays: YmgB (7 and 9.9 kDa, N-terminal proteolytically cleaved and uncleaved, respectively) bind to A) *crp* (candidate identified with nickel-enrichment DNA microarrays); B) *rpsV* (identified with both regular DNA microarrays and nickel-enrichment DNA microarrays); C) *lsrF* (identified with nickel-enrichment DNA microarrays) and D) *gadA* (identified with microarray) promoters. E) YmgB does not bind to the EBNA DNA. The significant difference in these experiments is the length of the DNA fragments used for the binding studies. The EBNA DNA fragment is much shorter (60 bp) than the 200-300 bp fragments used for the experiments in A-D, indicating that the interaction requires a longer DNA fragment, which is characteristic for geometric DNA recognition. Thus, at least a DNA

fragment longer or equal to 60 bp is required for DNA binding. To confirm these results, binding to an unrelated but long DNA fragment was tested: F) YmgB binds the *ycfR* promoter, which has not been identified as an interaction candidate. The length (bp) of the used DNA fragment is similar to the DNA fragments used in experiments A-D, and therefore strongly supports that YmgB binds either direct or indirect unspecific DNA fragments, most likely via a geometric recognition. This is the same method as used by H-NS (details described in the manuscript). In all assays: Lane 1: labeled DNA of *crp*, *rpsV*, *lsrF*, and *gadA etc;* Lane 2: labeled DNA + YmgB-7 kDa (proteolytically cleaved, see Figure S3 and text); Lane 3: labeled DNA + YmgB-9.9 kDa; arrows (←) highlight YmgB bound DNA.

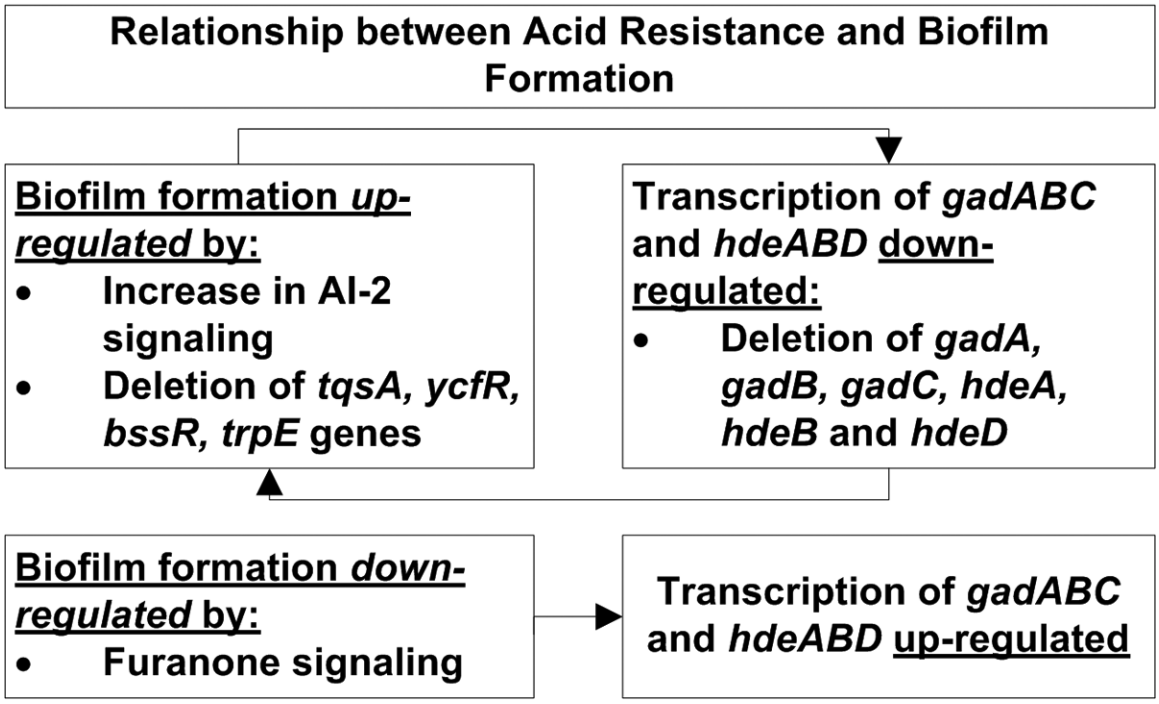

Figure 6.
Relationship between acid-resistance and biofilm formation. Biofilm formation is up-regulated by increased AI-2 signaling and deletion of important biofilm regulatory genes. In parallel, acid-resistance is down-regulated. When genes important for acid-resistance, such as genes in the *gad* and *hde* loci, are deleted, biofilm formation is up-regulated. The same behavior is seen for the *ymg* gene cluster.

Table 1

Summary of crystal parameters, data collection, and refinement statistics for YmgB.

Space group	***$C222_1$***		
Unit cell parameters	a = 69.92, b = 69.95 c = 55.0 Å, α = β = γ = 90°		
Data Collection	λ_1MADSe	λ_2MADSe	λ_3MADSe
Wavelength (Å)	0.9793	0.9790	0.9322
Resolution range (Å)	45.0 – 1.95	45.0 – 1.95	45.0 – 1.8
Number of observations	164,403	164,876	192,240
Number of reflections	10129	10124	12802
Completeness (%)	99.9 (99.6)$^+$	99.9 (99.6)$^+$	99.6 (96.8)$^+$
Mean I/σ(I)	6.8 (4.7)$^+$	6.8 (4.7)$^+$	6.6 (3.6)$^+$
R_{sym} on I	0.043 (0.14)$^+$	0.035 (0.13)$^+$	0.038 (0.380)$^+$
Sigma Cutoff	0.0	0.0	0.0
Highest resolution shell (Å)	1.95 – 2.02	1.95 – 2.02	1.80 – 1.86
Model and refinement statistics			
Resolution range (Å)	20.0 – 1.80	Data set used in refinement	λ_3
No. of reflections (total)	12174	Cutoff criteria	|F| > 0
No. of reflections (test)	625	R_{cryst}	0.215
Completeness (% total)	99.7	R_{free}	0.238
Stereochemical parameters			
Restraints (RMS observed)			
Bond length	0.016Å		
Bond angle	1.67 °		
Average isotropic B-value	20.25 $Å^2$		
Protein residues / atoms	124 / 966		
Solvent molecules	31		
β-Octyl Glucoside	2		

$R_{sym} = \Sigma|I_i - \langle I_i \rangle| / \Sigma|I_i|$ where I_i is the scaled intensity of the i^{th} measurement, and $\langle I_i \rangle$ is the mean intensity for that reflection.

$R_{cryst} = \Sigma| |F_{obs}| - |F_{calc}| | / \Sigma|F_{obs}|$ where F_{calc} and F_{obs} are the calculated and observed structure factor amplitudes, respectively.

R_{free} = as for R_{cryst}, but for 5.0% of the total reflections chosen at random and omitted from refinement.

$^+$highest resolution shell

Table 2

List of DNA binding sites for YmgB identified *in vivo* using nickel-enrichment DNA microarrays.

Gene or IG region	B number	YmgB chip signal	Enrichment *ymgB*/pCA24N signal ratio
crp	b3357	9201	7
rpsV	b1480	8765	21
lsrF	b1517	4584	13
IG region ydeU-ydeK	b1509, b1510	3337	25
yfeZ	b2433	2529	14
ymgG	b1172	2157	28
yhjQ	b3534	2099	28
bcsA	b3533	2099	28
rrsA	b3851	2029	68
ileT	b3852	2029	68