doi:10.1016/j.jmb.2007.07.037 J. Mol. Biol. (2007) 373, 11–26

JMB

Available online at www.sciencedirect.com
ScienceDirect

ELSEVIER

Structure and Function of the *Escherichia 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, USA*

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

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

[4]*Department of Biology Texas A & M University College Station TX 77843-3258, USA*

Received 23 April 2007; received in revised form 12 July 2007; accepted 16 July 2007
Available online 2 August 2007

Edited by J. Karn

The *Escherichia coli* gene cluster *ymgABC* was identified in transcriptome studies to have a role in biofilm development and stability. In this study, we showed that YmgB represses biofilm formation in rich medium containing glucose, decreases cellular motility, and protects the cell from acid indicating that YmgB has a major role in acid-resistance in *E. coli*. Our data show that these phenotypes are potentially mediated through interactions with the important cell signal indole. In addition, gel mobility-shift assays suggest that YmgB may be a non-specific DNA-binding protein. Using nickel-enrichment DNA microarrays, we showed that YmgB binds, either directly or indirectly, *via* a probable ligand, 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 having only 5% sequence identity. This supports our DNA microarray data showing 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.

Keywords: biofilm; protein structure; gene regulatory protein; acid resistance; *E. coli*

Introduction

Bacteria seldom live individually,[1,2] and swarming and biofilm formation are two important examples of bacterial multicellular behavior, which is significant because it enhances the chance for survival in competitive environments.[3,4] However, it can 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

*Corresponding author. E-mail address: wolfgang_peti@brown.edu.

† J.L. and R.P. contributed equally to this work.

Abbreviations used: β-OG, β-octyl-β-D-glucopyranoside; EMSA, electrophoretic mobility-shift assay; MALDI-TOF MS, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry.

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*[7–10]), proteome 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 residue *E. coli* protein YmgB, which 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.[10,13–15] For example, the furanosyl borate diester or derivative quorum signal autoinducer 2 (AI-2)[16] repressed *ymgAB* threefold.[14] In contrast, the biofilm inhibitor furanone from the alga *Delisea pulchra*, which masks autoinducer 2 (AI-2) signaling,[17] induced *ymgA* twofold.[14] Furthermore, deleting the AI-2 transporter gene *tqsA* also affected *ymgABC* expression. Specifically, it repressed *ymgBC* fourfold.[13] Also, *ymgABC* were induced 14-fold at 15 h relative to 7 h biofilms,[10] and the stationary-phase biofilm signal indole repressed *ymgABC* two- to fivefold.[15] Therefore, these results all strongly suggested that the *ymgABC* gene cluster, and thus likely the YmgB protein itself, has an important role in *E. coli* biofilm formation as a result of AI-2 or indole signaling.

Using similar transcriptome studies, we present in this manuscript that the gene cluster *ymgABC* has a significant role in acid-resistance. 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* sevenfold to 12-fold,[14] while furanone does the opposite, repressing biofilm formation and inducing *gadB* transcription threefold.[14] Also, the deletion of the *tqsA* gene, causes biofilm formation to increase 7000-fold, while *gadABCX* expression is repressed 11-fold to 13-fold.[13] In addition, when the biofilm inhibitor/stress regulator *bhsA* is deleted, which leads to a fivefold increase in biofilm formation, *gadABCE* are also repressed three- to sixfold.[20] Similar observations were made upon the deletion of *bssR*.[21] Finally, indole repressed *gadABCEX* two- to fourfold.[15]

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 three- to fivefold by indole,[15] sixfold to 18-fold[14] by AI-2 and in all mutants (*tqsA, bssR*) that lead to elevated biofilm formation.[13,21] In addition, the deletion of *bhsA,* which increases biofilm formation fivefold, induces *hdeABD* expression three- to fivefold.[20] Since the uncharacterized gene cluster *ymgABC* is regulated in these data sets in a manner almost identical with those of the known acid-resistance genes *gadABCEX* and *hdeABD*, we deduced that this locus most likely encodes an acid-resistance locus.[10,13–15]

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 show that all of these phenotypes may be mediated through interactions with the cellular signaling molecule indole.[15,23] However, the biological function of YmgB is unknown and, in fact, is annotated as a hypothetical protein with no assigned function. Because a 3-D structure of a protein is more highly conserved than its sequence, it can frequently lead to identification of its biological function. Therefore, we used X-ray crystallography to determine the high-resolution structure of YmgB. On the basis of 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 mobility-shift experiments and *in vivo* nickel-enrichment DNA microarrays. Thus, using these powerful combined genetic and structural studies, we show that YmgB functions as a regulatory protein that 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.[10,13–15] Therefore, we tested the effect of deleting the *ymg* locus on biofilm formation, and compared these results with those observed for other known acid-resistance gene mutants. Two independent cultures of ten isogenic *E. coli* K-12 mutants (mutants *ycgZ, ymgA, ymgB*, and *ymgC*, along with the known acid-eresistance 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 1(a)). These results confirm the importance of acid- resistance genes in biofilm formation.[15] Importantly, deletion of *ymgA, ymgB,* or *ymgC* increased biofilm formation (ninefold, fourfold and sixfold, 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 (Supplementary Data Figure S1), agreeing well with the observed increase in biofilm formation of these mutants (Figure 1(a)). Also, as observed previously with K-12 wild-type, *trpE*, *tnaA*, and *tnaC* mutants,[15] the addition of 500 μM indole decreased motility of the *ymgA*, *ymgB*, and *ymgC* mutants by 35–65% (Supplementary Data Figure S1). These results provide evidence that indole decreases biofilm formation, in part, by reducing cellular motility.[15]

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 Supplementary Data Table S2, *gadABCE* and *hdeB*, two well-known acid-resistance gene loci, were repressed two- to threefold 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 1(b)). Significantly, the *ymgB* mutant showed 40-fold less

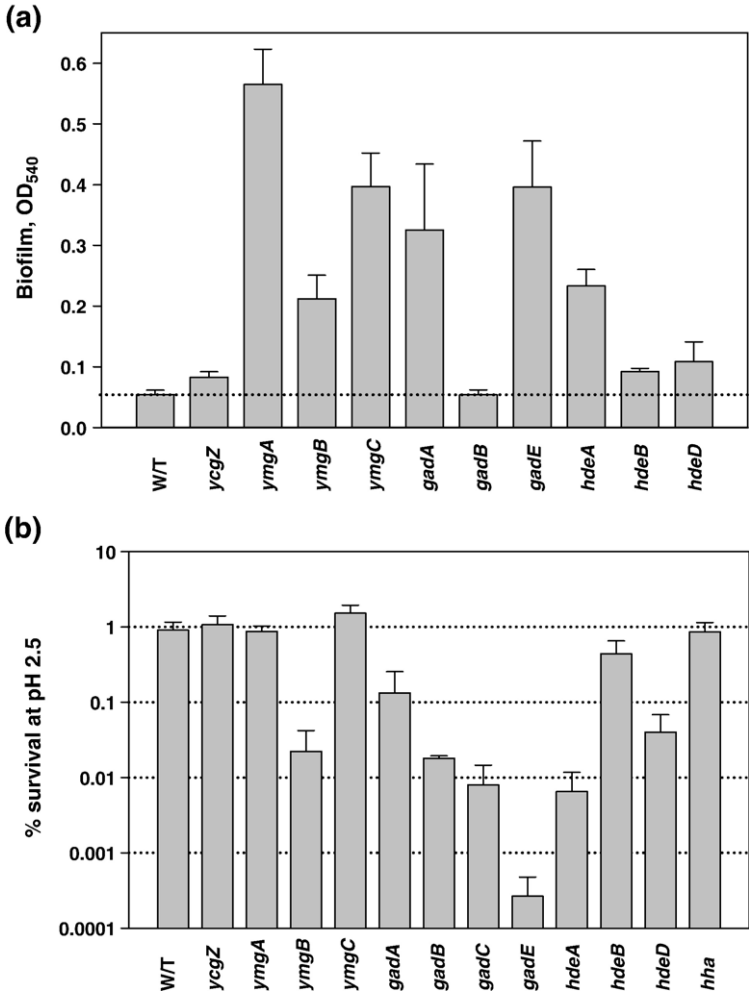

Figure 1. (a) Biofilm formation of BW25113 wild-type (W/T) and various knockout mutants in LB glu medium at 24 h at 37 °C. For both graphs, each experiment was repeated 2–4 times and the experimental standard deviations are shown. (b) Acid-resistance of *E. coli* BW25113 wild-type (W/T) and various knockout 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 are used for comparison.

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 two- 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 tenfold further with 1 mM IPTG induction in the YmgB complementation strain, BW25113 *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 Supplementary Data Table S2.

The *ymg* locus is also repressed by indole,[15] and it is known that indole represses biofilms.[10,15,20,21] Therefore, we investigated whether indole has 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-fold 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 resistance to hydrogen peroxide (oxidant)

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 to H_2O_2 by 537% compared to the K-12 wild-type strain (Supplementary Data Figure S2). As a negative control, the *gadE* mutant was tested for H_2O_2 sensitivity (deletion of the *gadE* gene had the largest effect of all genes tested on acid survival compared to wild-type, Figure 1(b)). 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 over-expression (plasmid pCA24N-*ymgB*$^+$ with 1 mM IPTG) (Supplementary Data Figure S2).

Assignment of a function to the hypothetical protein YmgB

The aforementioned biofilm, acid resistance, H_2O_2 resistance, and indole experiments demonstrate clearly 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 been assigned to YmgB (P75993) and the protein is classified as a hypothetical protein. Because tertiary structure is far better conserved than primary structure, solving the 3-D structure of a protein often leads to the determination of its function. The sequence of YmgB is less than 30% identical with any protein whose coordinates have been deposited in the RCS Protein Data Bank (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, and it crystallized as a head-to-head dimer. The final model includes two protein molecules (each containing residues 25–86), 31 water molecules and two β-octyl-β-D-glucopyranoside (β-OG) moieties; no electron density was observed for residues 87–88. The Matthews coefficient (V_m) for YmgB is 2.45 and the estimated solvent content is 45.7% (v/v).[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 2(a)). 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 an angle of nearly 90°. 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 2(b)). 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 2(c); calculated using the GetArea1.1 program.[30] SeMet42, in addition to interacting with residues from helix α1 of the second monomer, 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 bury significantly smaller amounts of solvent-accessible surface area (*ASA*), and because 1326 Å^2 of *ASA* is within the range observed for other dimerization interactions,[31] we predict that this is the biologically relevant interface. Finally, the

Table 1. Summary of crystal parameters, data collection, and refinement statistics for YmgB

Space group		$C222_1$	
Unit cell parameters			
a (Å)		69.92	
b (Å)		69.95	
c (Å)		55.0	
α = β = γ = (deg.)		90	
A. *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
No. observations	164,403	164,876	192,240
No. reflections	10,129	10,124	12,802
Completeness (%)	99.9 (99.6)+	99.9 (99.6)+	99.6 (96.8)+
Mean $I/\sigma(I)$	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
B. *Model and refinement statistics*			
Resolution range (Å)		20.0–1.80	
No. reflections (total)		12,174	
No. reflections (test)		625	
Completeness (% total)		99.7	
Data set used in refinement		λ^3	
Cutoff criterion		$\|F\| > 0$	
R_{cryst}		0.215	
R_{free}		0.238	
C. *Stereochemical parameters*			
Restraints (RMS observed)			
Bond lengths (Å)		0.016	
Bond angles (deg.)		1.67	
Average isotropic *B*-value (Å^2)		20.25	
Protein residues/ atoms		124/966	
No. solvent molecules		31	
β-Octyl glucoside		2	

+ Highest resolution shell.

$R_{sym} = \sum | I_i - \langle I_i \rangle | | / \sum | I_i |$ where *Ii* is the scaled intensity of the ith measurement, and $\langle I_i \rangle$ is the mean intensity for that reflection.

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

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

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 smaller basic patches (Figure 2(d)).

The N terminus of YmgB is subject to proteolysis

The N-terminal ~24 amino acid residues of YmgB were not observed in the electron density map. Using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS), we demonstrated that the crystallized YmgB protein was ~3 kDa less than that predicted on the basis of its full-length amino acid sequence (Supplementary Data Figure S3). Three major peaks were observed with molecular masses of 7153 Da, 7210 Da, and 7298 Da. 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 reported.[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 are 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 23–88 and 26–88 are stable for more than six months. 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 that 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 to identify proteins from other organisms with sequences similar to that of YmgB.[33] Only four distinct sequences were identified, all of which were from bacterial species: *Salmonella typhimurium*, *Enterobacter* sp. 638, *Erwinia carotovora*, and *Shigella 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 server,[34] which is used to help identify the likely biochemical function of a protein on the basis of its three-D structure. The 3D template searches of ProFunc indicate that YmgB is not likely to be an enzyme, is not expected to bind ligands, and is not likely to bind specific sequences of DNA. Finally, it is 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; Figure 4(a))[35] and the C-terminal dimerization domain of thermolysin (PDBID 1TRL; Figure 4(b)). The sequence identity

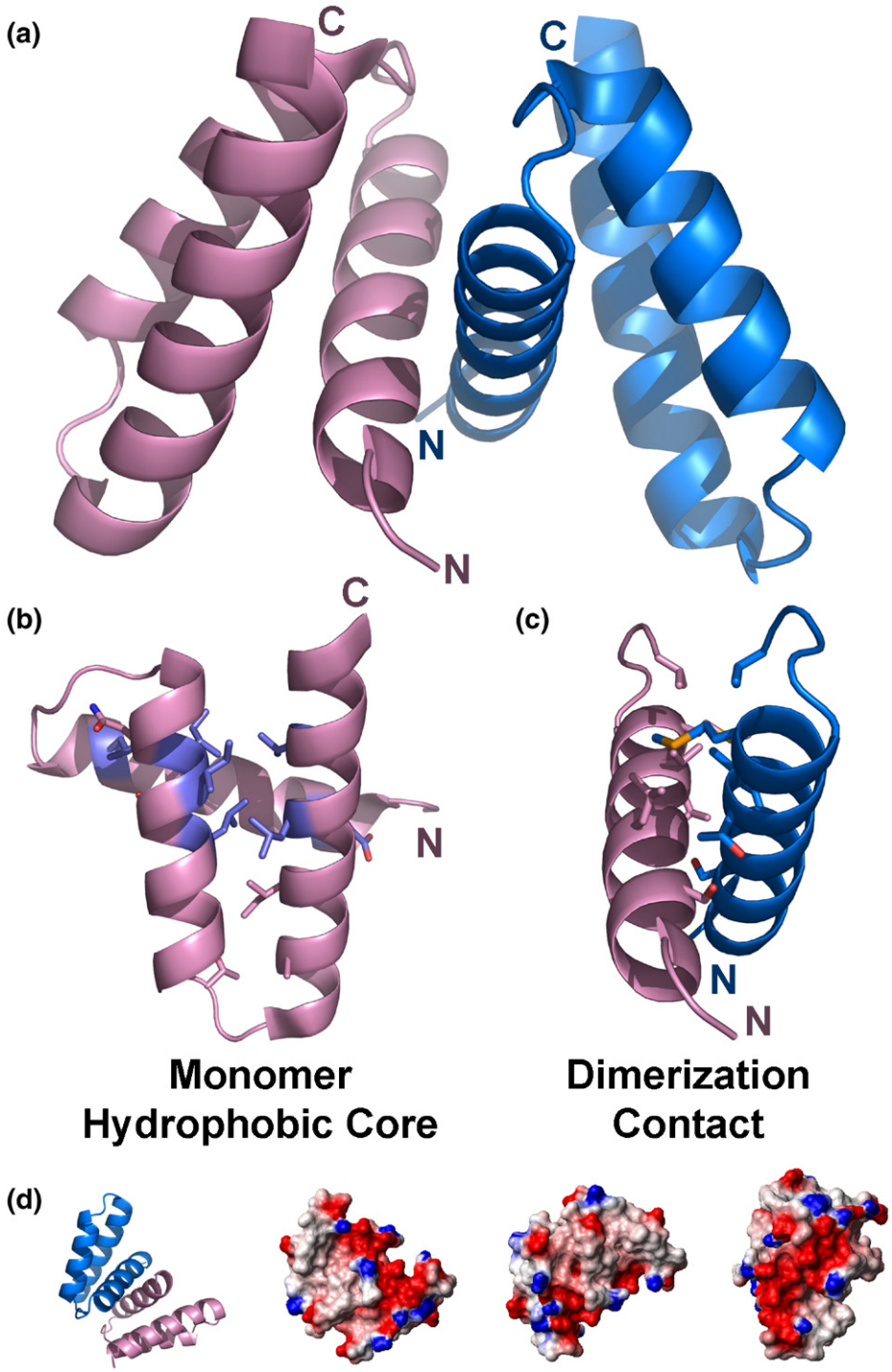

Figure 2. Structure of YmgB. (a) Structure of the YmgB dimer. Chain A, pink, chain B, blue. N and C termini are labeled. (b) Residues (shown as sticks) of the YmgB monomer that form the hydrophobic core, as determined by a loss of solvent-accessible surface area (ASA, GETAREA1.1). Residues in are purple are conserved in the YmgB family of proteins (see Figure 3). (c) Residues (shown as sticks) that 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 dimers 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 shown as a ribbon model on the left.

with these proteins is extremely low, only 5% and 7%. However, their structures are highly similar. Hha is a small (~8.5 kDa) 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

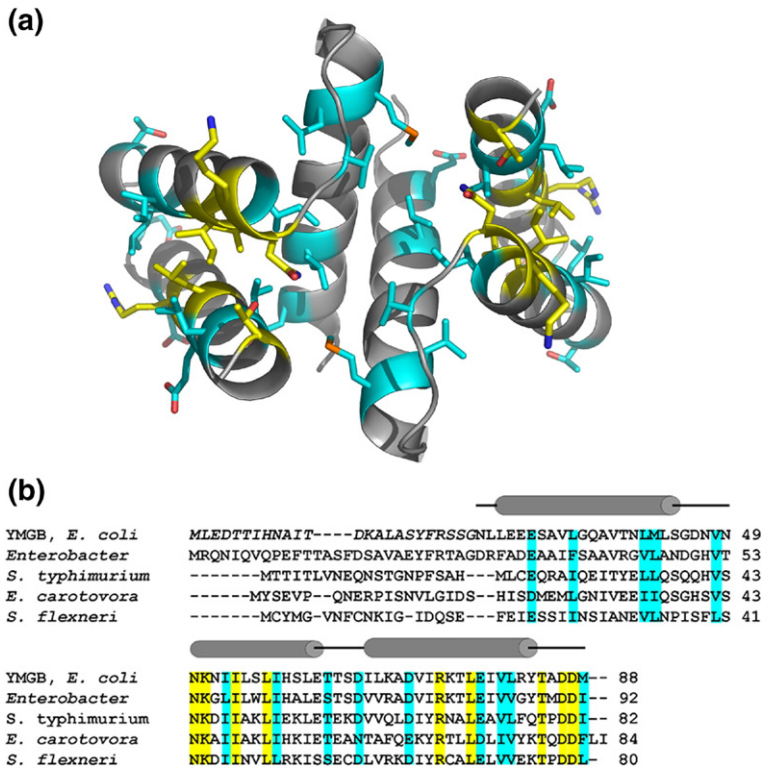

Figure 3. Sequence alignment of the four proteins identified by position-specific iterated (PSI) BLAST to be most similar to YmgB (NR database: all non-redundant GenBank CDS translations, RefSeq proteins, PDB, SwissProt, PIR, PRF databases). Four proteins with similarities to YmgB were identified: (1) conserved hypothetical protein, *Enterobacter* sp. 638, 51% sequence identity; (2) putative cytoplasmic protein, *S. typhimurium LT2*, 41% sequence identity; (3) hypothetical protein, ECA1764 *E. carotovora*, 36% sequence identity; (4) Hypothetical 9.1 kDa protein in spaS 3′region, *S. flexneri*, 27% sequence identity. YmgB α-helices are illustrated with grey cylinders above the sequence. YmgB residues proteolytically removed before 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.

recognition).[36–38] 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 4(a)). 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.[39] 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*.[40]

The importance of this three-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.[41] 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 thermolysin is shown in Figure 4(b). 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-D structure of YmgB and its observed high level of 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.

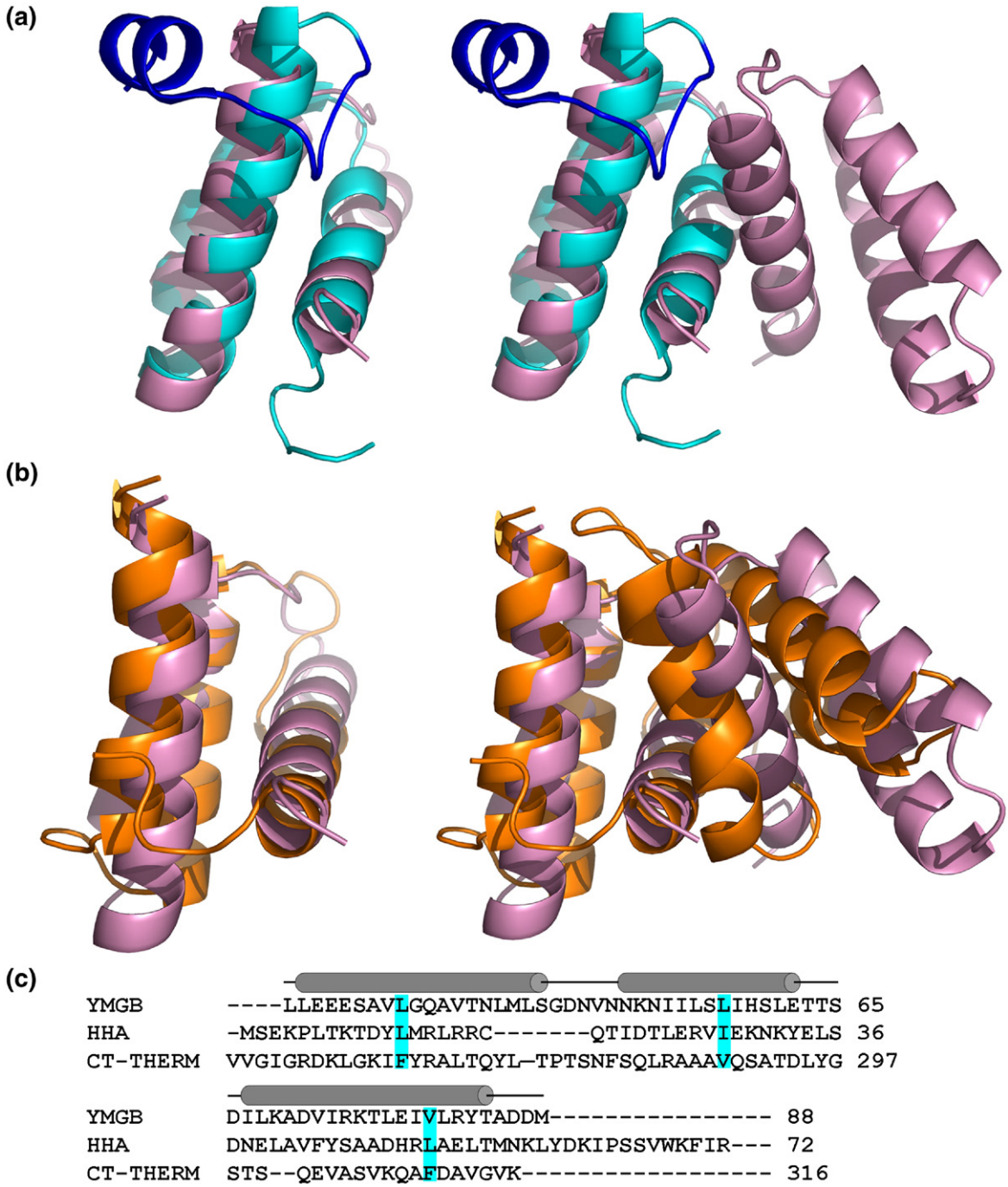

Figure 4. Superposition of YmgB with Hha and the C-terminal dimerization domain of thermolysin (CT-THERM). Left, superposition onto YmgB monomer; right, superposition onto YmgB dimer. (a) YmgB (pink) superimposed onto the structure of the *E. coli* protein Hha (cyan). The C-terminal loop of Hha (dark 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 á1 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 that of 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 is no identical residue between the three proteins.

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 fourfold more biofilm than the wild-type strain. The DNA purified using nickel enrichment from this culture was then used for microarray analysis. The nickel-enriched DNA microarray assay 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

Table 2. 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

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 regulates biofilm formation directly. Finally, another key gene identified in these studies was that of *rpsV*, which encodes a 30 S ribosomal subunit-associated protein that is induced at stationary phase.[48] This gene is co-transcribed with

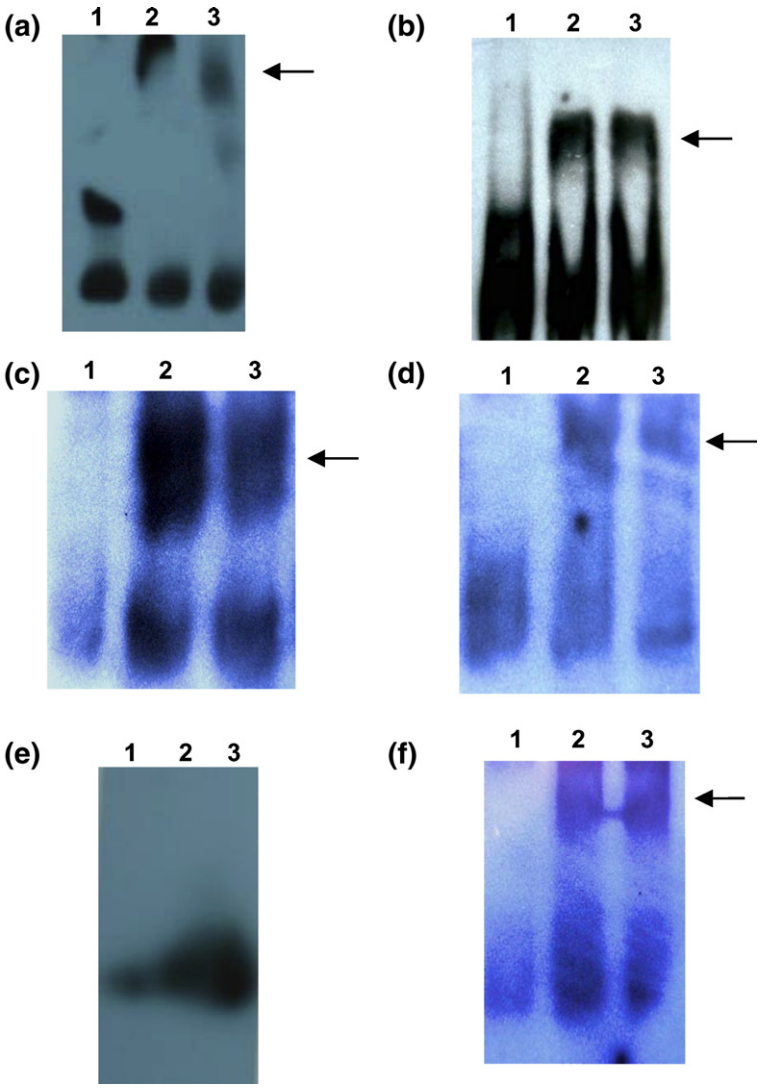

Figure 5. DNA-binding assays: YmgB (7 and 9.9 kDa, N-terminal proteolytically cleaved and uncleaved, respectively) bind to (a) *crp* (candidate identified with nickel-enrichment DNA microarrays); (b) *rpsV* (identified with both regular DNA microarrays and nickel-enrichment DNA microarrays); (c) *lsrF* (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 provides evidence that YmgB binds indirectly unspecific DNA fragments either directly or, most likely *via* a geometric recognition. This method is the same as that used by H-NS (see the text). 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 the text); lane 3, labeled DNA + YmgB-9.9 kDa; arrows (←) highlight YmgB-bound DNA.

bdm (biofilm-dependent modulation) whose expression is reduced in a biofilms,[49] and induced by the Rcs signaling system in response to low-temperature growth in glucose and to high external concentrations of zinc.[50] The *rpsV* gene was also identified in our regular DNA-microarray of *ymgB versus* wild-type (repressed 2.3-fold), which further supports the likelihood that YmgB regulates this gene directly. 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-D structural resemblance of YmgB to Hha, it was important to determine whether 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 kDa) and full-length protein (9.9 kDa) (Figure 5) bind both DNA targets either directly or indirectly *via* an as yet, despite several efforts, unidentified DNA-binding protein (potentially analogous to H-NS of the Hha-H-NS complex), 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, similar to Hha-H-NS, geometry-dependent.[36,37] 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 supported strongly 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 four- to fivefold by the *hha* deletion and repressed two- to threefold by the deletion of *ymgB* (our unpublished results).

Discussion

To summarize, we have shown 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, previously identified for the *gadABC* and *hdeABD* genes. In this study, we demonstrate that the *ymgABC* locus also follows these relationship rules and thus has 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 here, 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 proposed in the literature.[25]

To understand the how the hypothetical YmgB protein functions to accomplish these biological

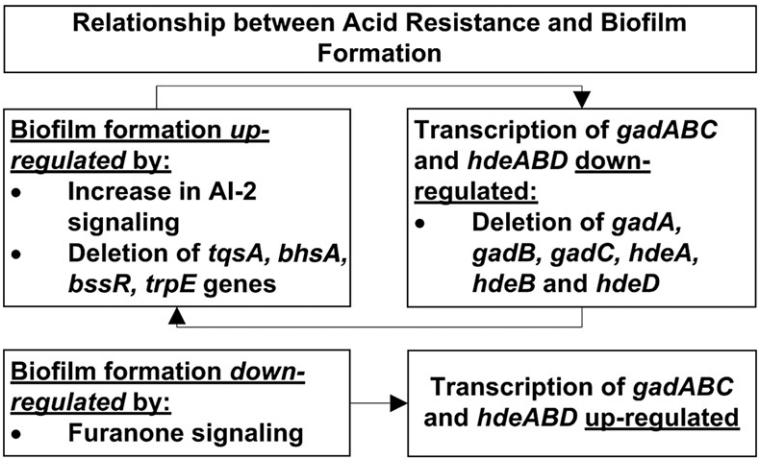

Figure 6. Relationship between acid-resistance and biofilm formation, which is up-regulated by increased AI-2 signaling and deletion of important biofilm regulatory genes. In parallel, acid-resistance is down-regulated. When genes important for acid-resistance, such as genes in the *gad* and *hde* loci,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

processes, we carried out a sequence analysis of this protein using NCBI blast‡ and ExPASy blast§. This analysis did not reveal any previously identified conserved domains and identified only a limited number of homologs, all of which were bacterial. *E. coli* YmgB shares 27–51% identity with several putative proteins of unknown function in Gram-negative plant/animal pathogenic bacteria, such as *Enterobacter* sp. 638, *S. typhimurium*, *Salmonella choleraesuis*, *Salmonella paratyphi*, *Salmonella typhi*, *E. carotovora* subsp. atroseptica, *Shigella sonnei*, and *Shigella flexneri*. Therefore, to gain additional insights into the biological function of YmgB, we elucidated its 3-D structure[51] and used this structure to identify structural homologs of the YmgB protein.

Using the 3-D structure of YmgB as our search scaffold in 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[38] that has been shown to interact with H-NS,[36,37] a two-domain protein that 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 show also that 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.[38] 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 mobility-shift assay data suggest strongly 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 has a critical role in its ability to regulate gene expression, by means of interaction with an H-NS-like protein.

Finally, despite the striking similarity of the 3-D structure between YmgB and Hha that strongly indicates a similar function, our data show convincingly that they fulfil 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 the same pH (~8.3) after 24 h as the 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 supported strongly by follow-up DNA microarray studies of an *hha* deletion mutant that show a very different gene expression pattern compared to that 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 Supplementary Data Table S1. 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 that described by 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/ml after acid treatment divided by the initial number of colonies/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 treatment with H_2O_2 was calculated as the number of colony-forming units/ml remaining after treatment with H_2O_2 divided by

‡ http://www.ncbi.nlm.nih.gov
§ http://expasy.org/sprot
‖ http://biunit.naist.jp/matras/

the initial number of colony-forming units/ml at time zero. Two independent experiments were conducted.

Crystal violet biofilm assay

This assay was adapted from that described by Pratt and Kolter;[24] cells were grown in polystyrene 96-well plates at 37 °C for one day without shaking in LB and in 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 Student's *t*-test, and $p \leq 0.05$ was chosen as the level of statistical significance.[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, 500 μM indole in dimethylformamide was added to the motility agar. Dimethylformamide (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) dimethylformamide has no effect on motility.

Biofilm total RNA isolation

For the microarray experiments, 10 g of glass wool (Corning Glass Works, Corning, NY) was used to form biofilms[8] in 250 ml in 1 l Erlenmeyer shake flasks inoculated with overnight cultures 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.[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.[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 (Affymetrix company manual), which includes using premixed polyadenylated transcripts of the *Bacillus 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 (http://biocyc.org/ECOLI/).[59]

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 of glass wool to produce biofilms,[8] which were inoculated with overnight cultures 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, 1% (v/v) formaldehyde (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 for another 20 min with 100 rpm shaking at room temperature. The crosslinking was stopped by adding 0.125 M glycine (Sigma, St. Louis, MO) and shaking (100 rpm) the flasks for 5 min.

The glass wool was washed twice with 200 ml of cold 50 mM NaH_2PO_4 (pH 8), 0.5 M NaCl, 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) added to preserve the integrity of the proteins. The cells were lysed by sonication at a power level of 10 W using three 1 min pulses (Fisher Scientific, model 60 Dismembrator), and the lysate was centrifuged at 12,100***g***, at 4 °C for 20 min to pellet the cellular debris. Then 10 ml of Ni-NTA agarose gel resin (Invitrogen, San Diego, CA) was 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's protocol for the regular purification of His_6-tagged proteins (Invitrogen, Carlsbad, CA), and the purified complexes were eluted with 10 ml of 50 mM NaH_2PO_4 (pH 8), 0.5 M NaCl, 250 mM imidazole (Fisher Scientific). The eluate was incubated for 5 h at 70 °C with 75 μg/ml of 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 27,200***g*** for 25 min at 4 °C, and the pellet was dissolved in 1.8 ml of TE buffer (50 mM Tris–HCl (pH 7.5), 25 mM EDTA) with 100 μg of Proteinase K (Fisher Scientific) and incubated for 2 h at 45 °C to degrade the 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 100–1000 bp DNA fragments was confirmed by gel electrophoresis, the DNA fragments were labeled as described, and DNA microarrays were performed as indicated above.[19]

For analysis of the data, we chose as positive candidates genes and intergenic regions with at least a fourfold 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 sevenfold 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 assays were performed as described,[20] using purified 9.9 kDa (full-length) and 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 of 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 50 μg/ml of kanamycin and 34 μg/ml of chloramphenicol. 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 A_{600} of 0.6–0.9. The cultures were then placed at 4 °C and the shaker temperature adjusted to 18 °C. Cultures were induced with 1 mM IPTG and allowed to express overnight (~18 h) at 18 °C under vigorous shaking (250 rpm). Cultures were harvested by centrifugation and the pellets stored at −80 °C.

For purification, the pellets were resuspended in lysis buffer (50 mM Tris–HCl (pH 8.0), 5 mM imidazole, 500 mM NaCl, 0.1% (v/v) 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 was removed by centrifugation (40 000***g*** for 40 min at 4 °C). The clarified lysate was filtered through a 0.22 μm pore size membrane (Millipore, Billerica, MA) and loaded onto a HisTrap HP column (GE Healthcare, Piscataway, NJ) equilibrated with 50 mM Tris–HCl (pH 8.0), 5 mM imidazole, and 500 mM NaCl. The protein was eluted with a 5 mM–500 mM imidazole gradient. The fractions containing the target protein were identified by SDS-PAGE and were pooled and dialyzed against 50 mM Tris–HCl (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, was achieved overnight at room temperature with steady rocking. The cleaved sample was then dialyzed against 50 mM Tris–HCl (pH 7.5), 250 mM NaCl overnight and purified further 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–HCl (pH 7.8), 100 mM NaCl) and purified by size-exclusion chromatography (Sephadex 75 26/60, GE Healthcare) equilibrated in crystallization buffer. The protein elutes as a 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 of protein in 0.5 μl of crystallization liquid under 15 μl of 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–HCl (pH 8.0), 2.0 M NaCl) 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 β-OG at a final concentration of 0.5% (w/v) resulted in diffraction-quality crystals. The final crystallization conditions were 0.09 M Tris–HCl (pH 8.0), 1.8 M NaCl, 0.5% (w/v) β-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 obtained only with proteolytically cleaved protein, in which the N terminus had been cleaved, resulting in a shorter C-terminal YmgB protein fragment (~7 kDa instead of the full-length 9.9 kDa).

Data collection and structure determination

Data was collected at NSLS Beamline X6A at 100 K using a 210 CCD detector (Area Detector Systems Corporation, Poway, CA). Data was 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 *P*4222, *C*2221, *P*2221 and *P*42 (0.847, 0.842, 0.842 and 0.713, respectively). However, phases were only obtained only with data processed as *C*2221 (or *P*2221) and not in *P*42 or *P*4222. In addition, the phased model was later used as a template for molecular replacement against data processed in *P*42 and *P*4222; this also failed repeatedly. Thus, it was concluded that the true space group is *C*2221.

The structure of the YmgB was phased to 2.0 Å using a three-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 SeMET52. RESOLVE[68] was used immediately 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 RefMac[71] against the 1.8 Å resolution λ_3 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[a], which integrates seven validation tools: Procheck 3.5.4, SFcheck 4.0, Prove 2.5.1, ERRAT, WASP, DDQ 2.0, and Whatcheck.

Mass spectrometry

Crystals from the plate used for structure determination were harvested and washed (3×) 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 that did not. At various time-points (0, 2, 6, 8, 10, 13, and 38 days), 10 μl was removed, ethanol-precipitated and stored at −20 °C. Protein was resuspended in water immediately before mass spectrometry screening. Samples were diluted 1:1 with matrix solution (10 mg of sinapinic acid in 30% (v/v) acetonitrile with 0.1% (v/v) trifluoroacetic acid in water) and 1 μl plated onto a MALDI laser-etched stainless steel sample plate, allowed to dry and examined using a Voyager DE-Pro matrix-assisted laser desorption/ionization time-of-flight mass spectrometer MALDI-TOF-MS instrument (Applied Biosystems, Foster City, CA). Bovine serum albumin was used as a calibration standard.

Data Bank accession numbers

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] Atomic coordinates of the final model and experimental structure factors of have been deposited with the PDB and are accessible under the code 2OXL.

Acknowledgements

This research was supported by the NIH (EB003872-0IAI) to T.K.W. and Brown University start-up funds to W.P. and to R.P. The authors 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. We are grateful for the gift of *E. coli* mutants and plasmids by the National Institute of Genetics, Japan.

[a] www.jcsg.org

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.07.037

References

1. Kolter, R. & Losick, R. (1998). One for all and all for one. *Science*, **280**, 226–227.
2. Sauer, K. & Camper, A. K. (2001). Characterization of phenotypic changes in *Pseudomonas putida* in response to surface-associated growth. *J. Bacteriol.* **183**, 6579–6589.
3. Nickel, J. C., Ruseska, I., Wright, J. B. & Costerton, J. W. (1985). Tobramycin resistance of *Pseudomonas aeruginosa* cells growing as a biofilm on urinary catheter material. *Antimicrob. Agents Chemother.* **27**, 619–624.
4. Stanley, N. R. & Lazazzera, B. A. (2004). Environmental signals and regulatory pathways that influence biofilm formation. *Mol. Microbiol.* **52**, 917–924.
5. Costerton, B. (2004). Microbial ecology comes of age and joins the general ecology community. *Proc. Natl Acad. Sci. USA*, **101**, 16983–16984.
6. Whiteley, M., Bangera, M. G., Bumgarner, R. E., Parsek, M. R., Teitzel, G. M., Lory, S. & Greenberg, E. P. (2001). Gene expression in *Pseudomonas aeruginosa* biofilms. *Nature*, **413**, 860–864.
7. Beloin, C., Valle, J., Latour-Lambert, P., Faure, P., Kzreminski, M., Balestrino, D. *et al.* (2004). Global impact of mature biofilm lifestyle on *Escherichia coli* K-12 gene expression. *Mol. Microbiol.* **51**, 659–674.
8. Ren, D., Bedzyk, L. A., Thomas, S. M., Ye, R. W. & Wood, T. K. (2004). Gene expression in *Escherichia coli* biofilms. *Appl. Microbiol. Biotechnol.* **64**, 515–524.
9. Schembri, M. A., Kjaergaard, K. & Klemm, P. (2003). Global gene expression in *Escherichia coli* biofilms. *Mol. Microbiol.* **48**, 253–267.
10. Domka, J., Lee, J., Bansal, T. & Wood, T. K. (2007). Temporal gene-expression in *Escherichia coli* K-12 biofilms. *Environ. Microbiol.* **9**, 332–346.
11. Sauer, K., Camper, A. K., Ehrlich, G. D., Costerton, J. W. & Davies, D. G. (2002). *Pseudomonas aeruginosa* displays multiple phenotypes during development as a biofilm. *J. Bacteriol.* **184**, 1140–1154.
12. Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M. *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
13. Herzberg, M., Kaye, I. K., Peti, W. & Wood, T. K. (2006). YdgG (TqsA) controls biofilm formation in *Escherichia coli* K-12 through autoinducer 2 transport. *J. Bacteriol.* **188**, 587–598.
14. Ren, D., Bedzyk, L. A., Ye, R. W., Thomas, S. M. & Wood, T. K. (2004). Differential gene expression shows natural brominated furanones interfere with the autoinducer-2 bacterial signaling system of *Escherichia coli*. *Biotechnol. Bioeng.* **88**, 630–642.
15. Lee, J., Jayaraman, A. & Wood, T. K. (2007). Indole is an inter-species biofilm signal mediated by SdiA. *BMC Microbiol.* **7**, 42.
16. Lombardia, E., Rovetto, A. J., Arabolaza, A. L. & Grau,

R. R. (2006). A LuxS-dependent cell-to-cell language regulates social behavior and development in *Bacillus subtilis*. *J. Bacteriol.* **188**, 4442–4452.
17. Ren, D., Sims, J. J. & Wood, T. K. (2001). Inhibition of Biofilm Formation and Swarming of *Escherichia coli* by (*5Z*)-4-Bromo-5-(Bromomethylene)-3-Butyl-2(*5H*)-Furanone. *Environ. Microbiol.* **3**, 731–736.
18. Ma, Z., Gong, S., Richard, H., Tucker, D. L., Conway, T. & Foster, J. W. (2003). GadE (YhiE) activates glutamate decarboxylase-dependent acid resistance in *Escherichia coli* K-12. *Mol. Microbiol.* **49**, 1309–1320.
19. González Barrios, A. F., Zuo, R., Hashimoto, Y., Yang, L., Bentley, W. E. & Wood, T. K. (2006). Autoinducer 2 controls biofilm formation in *Escherichia coli* through a novel motility quorum-sensing regulator (MqsR, B3022). *J. Bacteriol.* **188**, 305–316.
20. Zhang, X. S., García-Contreras, R. & Wood, T. K. (2007). YcfR (BhsA) influences *Escherichia coli* biofilm formation through stress response and surface hydrophobicity. *J. Bacteriol.* **189**, 3051–3062.
21. Domka, J., Lee, J. & Wood, T. K. (2006). YliH (BssR) and YceP (BssS) regulate *Escherichia coli* K-12 biofilm formation by influencing cell signaling. *Appl. Environ. Microbiol.* **72**, 2449–2459.
22. Masuda, N. & Church, G. M. (2003). Regulatory network of acid resistance genes in *Escherichia coli*. *Mol. Microbiol.* **48**, 699–712.
23. Wang, D., Ding, X. & Rather, P. N. (2001). Indole can act as an extracellular signal in *Escherichia coli*. *J. Bacteriol.* **183**, 4210–4216.
24. Pratt, L. A. & Kolter, R. (1998). Genetic analysis of *Escherichia coli* biofilm formation: roles of flagella, motility, chemotaxis and type I pili. *Mol. Microbiol.* **30**, 285–293.
25. Foster, J. W. (2004). *Escherichia coli* acid resistance: tales of an amateur acidophile. *Nature Rev. Microbiol.* **2**, 898–907.
26. Zheng, M., Wang, X., Templeton, L. J., Smulski, D. R., LaRossa, R. A. & Storz, G. (2001). DNA microarray-mediated transcriptional profiling of the *Escherichia coli* response to hydrogen peroxide. *J. Bacteriol.* **183**, 4562–4570.
27. Matthews, B. W. (1968). Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497.
28. Kantardjieff, K. A. & Rupp, B. (2003). Matthews coefficient probabilities: improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci.* **12**, 1865–1871.
29. Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G. *et al.* (2003). Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins: Struct. Funct. Genet.* **50**, 437–450.
30. von Freyberg, B., Richmond, T. J. & Braun, W. (1993). Surface area included in energy refinement of proteins. A comparative study on atomic solvation parameters. *J. Mol. Biol.* **233**, 275–292.
31. Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
32. Credle, J. J., Finer-Moore, J. S., Papa, F. R., Stroud, R. M. & Walter, P. (2005). On the mechanism of sensing unfolded protein in the endoplasmic reticulum. *Proc. Natl Acad. Sci. USA*, **102**, 18773–18784.
33. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
34. Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucl. Acids Res.* **33**, W89–W93.
35. Yee, A., Chang, X., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B. *et al.* (2002). An NMR approach to structural proteomics. *Proc. Natl Acad. Sci. USA*, **99**, 1825–1830.
36. Dorman, C. J. (2004). H-NS: a universal regulator for a dynamic genome. *Nature Rev. Microbiol.* **2**, 391–400.
37. Dorman, C. J. (2007). H-NS, the genome sentinel. *Nature Rev. Microbiol.* **5**, 157–161.
38. Madrid, C., Balsalobre, C., Garcia, J. & Juarez, A. (2007). 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.* **63**, 7–14.
39. Nieto, J. M., Madrid, C., Miquelay, E., Parra, J. L., Rodriguez, S. & Juarez, A. (2002). Evidence for direct protein-protein interaction between members of the enterobacterial Hha/YmoA and H-NS families of proteins. *J. Bacteriol.* **184**, 629–635.
40. Rodriguez, S., Nieto, J. M., Madrid, C. & Juarez, A. (2005). Functional replacement of the oligomerization domain of H-NS by the Hha protein of *Escherichia coli*. *J. Bacteriol.* **187**, 5452–5459.
41. Rico, M., Jimenez, M. A., Gonzalez, C., De Filippis, V. & Fontana, A. (1994). NMR solution structure of the C-terminal fragment 255-316 of thermolysin: a dimer formed by subunits having the native structure. *Biochemistry*, **33**, 14834–14847.
42. Tamimi, Y., Lines, M., Coca-Prados, M. & Walter, M. A. (2004). Identification of target genes regulated by FOXC1 using nickel agarose-based chromatin enrichment. *Invest. Ophthalmol. Vis. Sci.* **45**, 3904–3913.
43. Perrenoud, A. & Sauer, U. (2005). Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in *Escherichia coli*. *J. Bacteriol.* **187**, 3171–3179.
44. Hanamura, A. & Aiba, H. (1992). A new aspect of transcriptional control of the *Escherichia coli* crp gene: positive autoregulation. *Mol. Microbiol.* **6**, 2489–2497.
45. Taga, M. E., Miller, S. T. & Bassler, B. L. (2003). Lsr-mediated transport and processing of AI-2 in *Salmonella typhimurium*. *Mol. Microbiol.* **50**, 1411–1427.
46. Kim, M. K., Park, S. R., Cho, S. J., Lim, W. J., Ryu, S. K., An, C. L. *et al.* (2002). The effect of a disrupted *yhjQ* gene on cellular morphology and cell growth in *Escherichia coli*. *Appl. Microbiol. Biotechnol.* **60**, 134–138.
47. Zogaj, X., Nimtz, M., Rohde, M., Bokranz, W. & Romling, U. (2001). The multicellular morphotypes of *Salmonella typhimurium* and *Escherichia coli* produce cellulose as the second component of the extracellular matrix. *Mol. Microbiol.* **39**, 1452–1463.
48. Izutsu, K., Wada, C., Komine, Y., Sako, T., Ueguchi, C., Nakura, S. & Wada, A. (2001). *Escherichia coli* ribosome-associated protein SRA, whose copy number increases during stationary phase. *J. Bacteriol.* **183**, 2765–2773.
49. Francez-Charlot, A., Castanie-Cornet, M. P., Gutierrez, C. & Cam, K. (2005). Osmotic regulation of the *Escherichia coli bdm* (biofilm-dependent modulation) gene by the RcsCDB His-Asp phosphorelay. *J. Bacteriol.* **187**, 3873–3877.
50. Hagiwara, D., Sugiura, M., Oshima, T., Mori, H., Aiba, H., Yamashino, T. & Mizuno, T. (2003). Genome-wide analyses revealing a signaling network of the RcsC-YojN-RcsB phosphorelay system in *Escherichia coli*. *J. Bacteriol.* **185**, 5735–5746.
51. Lesk, A. M. (2001). *Introduction into Protein Architecture.* Oxford University Press, New York, NY, USA.

52. Kawabata, T. (2003). MATRAS: a program for protein 3D structure comparison. *Nucl. Acids Res.* **31**, 3367–3369.
53. Juarez, A., Nieto, J. M., Prenafeta, A., Miquelay, E., Balsalobre, C., Carrascal, M. & Madrid, C. (2000). Interaction of the nucleoid-associated proteins Hha and H-NS to modulate expression of the hemolysin operon in *Escherichia coli*. *Advan. Expt Med. Biol.* **485**, 127–131.
54. Nieto, J. M., Madrid, C., Prenafeta, A., Miquelay, E., Balsalobre, C., Carrascal, M. & Juarez, A. (2000). 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.* **263**, 349–358.
55. Garcia, J., Cordeiro, T. N., Nieto, J. M., Pons, I., Juarez, A. & Pons, M. (2005). Interaction between the bacterial nucleoid associated proteins Hha and H-NS involves a conformational change of Hha. *Biochem. J.* **388**, 755–762.
56. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning, A Laboratory Manual*, 2nd edit., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
57. Ross, S. M. (2004). *Introduction to Probability and Statistics for Engineers and Scientists.* Elsevier Academic Press, USA, Burlington, MA, USA.
58. Sperandio, V., Torres, A. G. & Kaper, J. B. (2002). 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.* **43**, 809–821.
59. Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I. T. *et al.* (2005). EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucl. Acids Res.* **33**, D334–D337.
60. Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P. *et al.* (2005). NCBI GEO: mining millions of expression profiles-database and tools. *Nucl. Acids Res.* **33**, D562–D566.
61. Edgar, R., Domrachev, M. & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucl. Acids Res.* **30**, 207–210.
62. Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H. & Mori, H. (2005). 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.* **12**, 291–299.
63. Peti, W. & Page, R. (2007). Strategies to maximize heterologous protein expression in *Escherichia coli* with minimal cost. *Protein Expr. Purif.* **51**, 1–10.
64. Luft, J. R., Collins, R. J., Fehrman, N. A., Lauricella, A. M., Veatch, C. K. & DeTitta, G. T. (2003). A deliberate approach to screening for initial crystallization conditions of biological macromolecules. *J. Struct. Biol.* **142**, 170–179.
65. Otwinowski, Z. & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326.
66. Evans, P. (2006). Scaling and assessment of data quality. *Acta Crystallog. sect. D*, **62**, 72–82.
67. Terwilliger, T. C. & Berendzen, J. (1999). Automated MAD and MIR structure solution. *Acta Crystallog. sect. D*, **55**, 849–861.
68. Terwilliger, T. C. (2000). Maximum-likelihood density modification. *Acta Crystallog. sect. D*, **56**, 965–972.
69. Lamzin, V. S., Perrakis, A., & Wilson, K. S., Eds. (2001). The ARP/wARP suite for automated construction and refinement of protein models. In *International Tables for Crystallography,* vol. F, Crystallography of biological macromolecules. Edited by (Rossman, M.G., & Arnold, E., eds), Kluwer Academic Publishers, Dordrecht, pp. 720-722.
70. Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallog. sect. D*, **60**, 2126–2132.
71. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallog. sect. D*, **53**, 240–255.
72. Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.