

# Analysis of the clostridial hydrophobic with a conserved tryptophan family (ChW) of proteins in *Clostridium acetobutylicum* with emphasis on ChW14 and ChW16/17

Leighann Sullivan<sup>a</sup>, Carlos J. Paredes<sup>b</sup>, Eleftherios T. Papoutsakis<sup>b</sup>, George N. Bennett<sup>a,\*</sup>

<sup>a</sup> Department of Biochemistry and Cell Biology, Rice University, Houston, TX 77005, United States

<sup>b</sup> Department of Chemical Engineering, Northwestern University, Evanston, IL, United States

Received 10 May 2007; received in revised form 13 July 2007; accepted 27 July 2007

## Abstract

A novel protein family, clostridial hydrophobic with a conserved W (ChW), is specific to *Clostridium acetobutylicum*, thus, suggesting roles specific to its unique physiology. Functions for members of this protein family have not been characterized. The expression and promoter architecture of two genes, *chw14* (CAC1532) and *chw16/17* (CAC2584), encoding two members of the ChW protein family were characterized and their regulation explored. The genes *chw14* and *chw16/17* behave similarly under every condition tested in the DNA-microarray gene expression studies. Previous protein analysis suggested that the master transcriptional regulator, Spo0A, was required for their accumulation, as ChW14 and ChW16/17 proteins were absent in the *spo0A* null strain, SKO1. Primer extension assays showed a single transcript for each *chw14* and *chw16/17* detected from mid-exponential phase until early stationary phase. A predicted  $\sigma^A$  consensus motif is just upstream of the transcriptional start sites of both *chw14* and *chw16/17*, with a single putative Spo0A binding site, OA box, within the promoter region of both *chw14* and *chw16/17*. Using reporter analysis we showed that the promoters of *chw14* and *chw16/17* are highly active during mid-exponential phase in wild type *C. acetobutylicum*. Analysis of expression in the *spo0A* deficient SKO1 strain indicated the promoter activity of both *chw14* and *chw16/17* appears constitutive, thus the promoters do not appear to be inactivated in the absence of Spo0A. The relationship among ChW proteins and the ChW domains themselves was delineated. The ChW proteins appear to originate with *C. acetobutylicum* where the non-*C. acetobutylicum* ChW proteins are nested within the main *C. acetobutylicum* protein branch. Examination of the ChW domain alone shows that every third domain clusters together phylogenetically. Additionally, almost all of ChW proteins contain a multiple of three ChW domains. Taken together, these data suggest that ChW domains function in triplets of association.

© 2007 Elsevier Inc. All rights reserved.

**Keywords:** ChW domain; Spo0A transcriptional regulator;  $\beta$ -Galactosidase reporter assay; Primer extension; Phylogeny

## 1. Introduction

*Clostridium acetobutylicum* is grouped taxonomically with the Gram-positive, spore-forming, rod-shaped, obligate anaerobes containing a low G + C content. This microbe metabolizes sugars by fermentation and produces organic acid products. The accumulation of the acidic products acetate and butyrate increases the hydrogen ion concentration and thus threatens to disrupt the membrane chemiosmotic potential. As an immediate response to these adverse conditions *C. acetobutylicum* produces solvents, which act to diminish the pH stress, and later will initi-

ate the developmental program of sporulation. The ability of *C. acetobutylicum* to produce solvents has long been exploited for the microbial production of the industrially important chemicals, acetone and butanol.

In order to maximize our use of *C. acetobutylicum* as a solvent producer, it is of paramount importance to understand and control metabolic transitions because *C. acetobutylicum* only synthesizes solvents after it has transitioned from acid production and this transition stage is accompanied by many global cell physiological changes.

The transcriptional response to the sensed altered environment is carried out by  $\sigma$  factors and transcription factors to initiate large scale changes in patterns of gene expression based on the promoter structures present in the genome.  $\sigma$  factors regulate many post-exponential phase changes: sporulation,

\* Corresponding author. Tel.: +1 713 348 4920; fax: +1 713 348 5154.  
E-mail address: [gbennett@bioc.rice.edu](mailto:gbennett@bioc.rice.edu) (G.N. Bennett).

stationary phase, and repression of flagella synthesis [1,2]. The genes for the vegetative  $\sigma$  factor ( $\sigma^A$ ) and the sporulation specific  $\sigma$  factors ( $\sigma^E$ ,  $\sigma^G$ ,  $\sigma^K$ ) have been identified in *C. acetobutylicum* [1–3]. A  $\sigma$  factor may be required for transcription but the activity of the promoter may be regulated by an ancillary transcription factor. The transcription factor Spo0A is responsible for many of the large scale changes in gene expression. Spo0A may either positively or negatively affect transcription of genes that contain its recognition sequence, designated the 0A box, near promoters to regulate transcription. Spo0A has been better characterized in *Bacillus subtilis* than in *C. acetobutylicum*. In *C. acetobutylicum*, Spo0A has been shown to positively regulate the solvent synthesis operons *adc* and *adhE-ctfA-ctfB* [4,5] and suggested to negatively regulate the transition state regulator, *abrB* [6]. Spo0A has also been recently implicated as being required for expression of two genes, *chw14* and *chw16/17*, in a novel protein family, Clostridial hydrophobic with a conserved W (ChW), in *C. acetobutylicum* [7].

The ChW protein family is almost exclusively limited to the *C. acetobutylicum* species. No other organism has as many proteins (i.e., ChW proteins) containing multiple ChW domains. *C. acetobutylicum* has 20 such proteins. In species where a ChW protein was identified by BLASTP [8] it was limited to one protein, thus suggesting that the ChW domains originated with the *C. acetobutylicum* species. The ChW domain is approximately 47 amino acids long. Features of the domain are a conserved tryptophan, a high percentage of hydrophobic and small residues. ChW domains are found in proteins with a putative N-terminal signal sequence, domains for polysaccharide and protein degradation, or domains involved in cell adhesion [9]. ChW domains are additionally present in proteins with no other known domain. ChW proteins in *C. acetobutylicum* are surface proteins or secreted from the cell since they have putative N-terminal signal sequences. In general, little is known about the surface proteins of *C. acetobutylicum*, since only the non-cellulolytic cellulosome [10] and a few extracellular enzymes have been characterized: glycoside hydrolase [11], endoglucanase [12], muramidase [13], alpha-amylase [14], and metalloprotease [15]. Additionally, in other Gram-positive bacteria, such as the well-studied genera *Streptococcus* and *Staphylococcus*, surface proteins provide many diverse functions (for a review see Ref. [16]).

We characterized two genes, *chw14* (CAC1532) and *chw16/17* (CAC2584), as representative examples of the ChW family. These genes were chosen for in-depth analysis to provide a foundation regarding this undescribed family in *C. acetobutylicum*. We performed two types of analyses: gene level and protein level. At the gene level, we determined the expression patterns and promoter architectures for *chw14* and *chw16/17* using reporter, microarray, and primer extension analyses. The temporal expression pattern delineates the processes in which these genes may participate whereas promoter strength is an indication of the level of expressed product that may be available for those processes. Determining the exact start site of transcription allows the interpretation of binding sites of possible transcription factors which may modulate the gene expression of *chw14* and *chw16/17*. Reporter analysis performed in the

*spo0A* mutant strain adds to our understanding of regulation of *chw14* and *chw16/17*. At the protein level, we examined the entire amino acid sequence encoding the protein and the residues encoding only the ChW domains to discern whether there were any patterns in the predicted protein secondary structures and the phylogenetic relationships. Indeed, an apparent consensus protein secondary structure emerged among the ChW domains as well as the relationship between the ChW domains such that they are not homogenous, but rather form triplets of association.

## 2. Materials and methods

### 2.1. Bacterial strains and plasmids

All bacterial strains and plasmids used in this study are shown in Table 1.

### 2.2. Growth conditions and maintenance

All *Escherichia coli* strains were grown aerobically at 37 °C in liquid or agar-solidified Luria-Bertani (LB) medium. For recombinant *E. coli* strains, media were appropriately supplemented with ampicillin (100 µg/mL), erythromycin (300 µg/mL), chloramphenicol (35 µg/mL), or kanamycin (35 µg/mL). Both recombinant and wild type strains were stored at –80 °C in 50% (v/v) glycerol freezing solution (65% glycerol, 0.1 M MgSO<sub>4</sub>, 0.025 M Tris–HCl, pH 8). *C. acetobutylicum* ATCC 824 and recombinant strains were cultured anaerobically in Clostridial Growth Medium (CGM) at 37 °C [17] and stored frozen in 10% (v/v) glycerol at –80 °C or as horse serum-supplemented lyophilized stocks at room temperature. For recombinant *C. acetobutylicum* strains, liquid or agar-solidified medium was appropriately supplemented with erythromycin (40 µg/mL) or thiamphenicol (25 µg/mL).

Table 1  
Bacterial strains and plasmids

Strain or plasmid	Relevant characteristics <sup>a</sup>	Source or reference
Bacterial strains		
<i>C. acetobutylicum</i>		
ATCC 824	Wild type <i>C. acetobutylicum</i>	ATCC <sup>b</sup>
SKO1	ATCC 824 <i>spo0A::MLS<sup>r</sup></i>	Harris (2002)
<i>E. coli</i>		
DH5 $\alpha$	EndA1, recA1	Clontech <sup>c</sup>
DH10B	EndA1, recA1, Str <sup>r</sup>	Gibco <sup>d</sup>
TOP10	EndA1, recA1, Str <sup>r</sup>	Invitrogen <sup>e</sup>
Plasmids		
pAN1	Cm <sup>r</sup> $\Phi$ 3TI gene	Mermelstein (1993)
pDHKM	Km <sup>r</sup> $\Phi$ 3TI gene	Zhao (2003)
pHT3	Ap <sup>r</sup> MLS <sup>r</sup> <i>LacZ</i> ORF	Tummala (1999)
pHT3/MfeI	pHT3 with MfeI and XhoI sites	This study
pHT4	pHT3 with <i>ptb</i> promoter	Tummala (1999)
pHT14	pHT3/MfeI with <i>chw14</i> promoter	This study
pHT16	pHT3/MfeI with <i>chw16/17</i> promoter	This study
pThiLac	Cm <sup>r</sup> Thi <sup>r</sup> <i>lacZ</i> ORF	Scotcher (2005)
pTL14	pThiLac with <i>chw14</i> promoter	This study
pTL16	pThiLac with <i>chw16/17</i> promoter	This study

<sup>a</sup> Abbreviations: Cm<sup>r</sup>, chloramphenicol resistant; Thi<sup>r</sup>, thiamphenicol resistant; Ap<sup>r</sup>, ampicillin resistant; Str<sup>r</sup>, streptomycin resistant; Km<sup>r</sup>, kanamycin resistant; MLS<sup>r</sup>, macrolide lincosamide and streptogramin B resistant; endA1, mutant non-specific endonuclease1; recA1, mutant in homologous recombination;  $\Phi$ 3TI,  $\Phi$ 3TI methyltransferase; *ptb*, phosphotransbutyrylase.

<sup>b</sup> ATCC, American Type Culture Collection, Manassas, VA.

<sup>c</sup> Clontech, Palo Alto, CA.

<sup>d</sup> Gibco, Rockville, MD.

<sup>e</sup> Invitrogen, Carlsbad, CA.

### 2.3. Batch culture fermentation experiments

$\beta$ -galactosidase experiments were performed using batch culture fermentations. Single colonies of transformed *C. acetobutylicum* were grown in closed-cap batch fermentations of 200 or 300 mL (for pThiLac-based or pHT3-based transformed strains, respectively) CGM supplemented with the appropriate antibiotic at 37 °C in a Forma Scientific anaerobic chamber (Thermo Forma, Marietta, Ohio). To allow for differences in lag time following inoculation, 0 h is defined as the time when the culture had reached an OD<sub>600</sub> of 0.1. Cell growth was quantified by measuring the OD<sub>600</sub> using a Prim Light and Advanced spectrophotometer (Secomam, Domont, Cedex, France).

### 2.4. Controlled-pH fermentor experiments

Primer extension analysis was performed using controlled-pH fermentations. Large-scale batch fermentations of ATCC 824, 824(pHT14), and 824(pHT16) *C. acetobutylicum* strains were performed in a BioFlo 110 fermentor (New Brunswick Scientific, Edison, N.J.) with a culture volume of 2 L CGM. Recombinant strains were supplemented with appropriate antibiotics. Time 0 (0 h) is defined as in batch culture fermentation experiments. Cell growth was quantified by measuring the OD<sub>600</sub> using a Beckman DU64 spectrophotometer (Beckman, Coulter, Fullerton, Calif.).

### 2.5. Fermentation product analysis

For analysis of product formation, samples were taken at the time points specified for total RNA isolation or for  $\beta$ -galactosidase analysis. Batch fermentation samples were centrifuged at 3000×g for 5 min at room temperature in an EBA20 centrifuge (Hettich-Zentrifugen, Föhrenstr.12, Tuttingen, Germany). Controlled-pH fermentation samples were centrifuged at 2600×g for 5 min at 4 °C in a Sorvall RT6000B centrifuge (Dupont, Wilmington, Del.). The supernatant was collected, acidified and analyzed by a Hewlett-Packard 5890 Series II gas chromatograph (Hewlett-Packard, Palo Alto, Calif) for butanol, acetone, ethanol, butyrate, and acetate. Concentrations were determined from peak areas determined by Peak 2000 integration software (SRI Instruments, Torrance, Calif.).

### 2.6. DNA isolation, manipulation, and transformation into *C. acetobutylicum*

Plasmid isolation from *E. coli* was done by the QIAprep Miniprep method (QIAGEN, Valencia, Calif.). DNA was purified from agarose gels using the UltraClean 15 method (MO BIO Laboratories, Solana Beach, Calif.). PCR products or enzymatically manipulated DNA was purified by using the QIAquick PCR Purification method (QIAGEN, Valencia, Calif.). All commercial enzymes (Pfx

polymerase, restriction endonucleases, calf intestinal phosphatases, and T4 DNA ligase) were used according to the manufacturer's specifications. Pfx polymerase was obtained from Invitrogen while all other enzymes were obtained from New England Biolabs, Promega, or Fisher Scientific. Automated DNA sequencing was performed by LoneStar automated DNA sequencing Laboratory (LoneStar Laboratories, Houston, Tex). All oligonucleotides used in this study are listed in Table 2.

Previously published methods were used for electrotransformation of *C. acetobutylicum* [18]. Prior to transformation of *C. acetobutylicum*, plasmids pHT3, pHT4, pHT14, and pHT16 were methylated in *E. coli* (pAN1) [18] and plasmids pThiLac, pTL14, and pTL16 were methylated in *E. coli* (pDHKM) [19] by the *B. subtilis* phage $\Phi$ 3TI methyltransferase, which protects the plasmid DNA from restriction by the clostridial endonuclease *Cac824I* [20].

### 2.7. Isolation of total RNA for primer extension

Total RNA for time course primer extension studies was isolated from wild type and recombinant *C. acetobutylicum* [strains ATCC 824, 824(pHT14), and 824(pHT16)] cells collected during the following time points (OD<sub>600</sub> = 0.4, 0.8, 1.2, 1.6, 2.0 and time = 8, 12, 18 and 24 h) in controlled-pH fermentations. To ensure adequate RNA isolated from all time points, especially early ones, the sample volume harvested was dependent on growth stage as monitored by OD<sub>600</sub>. For the time points taken the volume harvested is indicated in parentheses: OD<sub>600</sub> = 0.4 (30 mL), 0.8 (15 mL), 1.2 (10 mL), 1.6 (10 mL), 2.0 and later time points (5 mL). Samples were centrifuged, cell pellets were resuspended in 200  $\mu$ L of lysis buffer (20 mg lysozyme per mL of 25% sucrose, 0.05 M Tris-HCl, 0.05 M EDTA), incubated at 37 °C for 5 min. 1 mL of TRIzol solution (Invitrogen, Carlsbad, Calif.) was added and the manufacturer's recommendations were followed for RNA isolation with one exception; an equal volume of 25 phenol:24 chloroform:1 isoamyl alcohol mixture pH 8 was used instead of 0.2 mL of chloroform in order to extract protein from the nucleic acids. RNA concentration was determined from OD<sub>260</sub> measurements (1 OD<sub>260</sub> unit = 40  $\mu$ g of RNA per mL).

### 2.8. Primer extension analysis

Time course primer extension reactions for *chw14* and *chw16/17* were performed with 14primext and lacZup primers by using an Avian Myeloblastosis Virus reverse transcriptase primer extension system (Promega, Madison, Wis.) using 100 and 20  $\mu$ g of total RNA, respectively, isolated from wild type *C. acetobutylicum* 824 and recombinant 824(pHT14), and 824(pHT16) strains, respectively.

The exact transcriptional start sites for *chw14* and *chw16/17* were mapped using sequencing reactions based on the Sanger dideoxy method. The sequencing

Table 2  
Oligonucleotides used in this study<sup>a</sup>

Oligonucleotide name	Sequence
MfeIXhoI	TCGAG <u>CAATTG</u> CTCGA
spo0A PE	CAAAATTCCTTATTATCATCTGCAATTAACAC
lacZup	TTAAATACCAATTATTATTAATAGGAATAATCTTTCTC
ptb primext	TTTCATTTCTTTGCTCTTTACCTTCATG
14primext	GCCCCAATCCAAAAGCAACTGCTGC
16primext	CCCAATATAAAAAACAAAAGCAGCTGC
Upst14F	CCCAATTG ATAAAATAAAGCCTAGTATGTTTTAATGC
Upst14R	CCCAATTG TTATTTCCCTTAAACTTG
Upst16F	CCCAATTG CTTGTTCTCCTTGAATGAATTTTATTG
Upst16R	CCCAATTG GTTTGAATTGTAGAAAACAGCATTGATACTG
pTL14F	CCG <u>CTCGAG</u> ATAAAATAAAGCCTAGTATGTT
pTL14R	GC <u>GGATCC</u> AACTTGAATATTATTGATTAATA
pTL16F	CG <u>CTCGAG</u> GTTTGAATTGTAGAAAACAGCATTG
pTL16R	GC <u>GGATCC</u> TGAATTTATTGTCAACATTTATATTTATA

<sup>a</sup> Underlined regions correspond to a restriction site for the restriction endonucleases MfeI (CAATTG), XhoI (CTCGAG), or BamHI (GGATCC).

reactions were performed on the corresponding DNA by using the same primer that was used for the primer extension reactions. Sequencing was performed using Sequenase Quick-denature plasmid sequencing protocol (USB Corporation, Cleveland, Ohio) with the radiolabeled lacZup primer complementary to the 5' end of the lacZ gene encoded on pHT14 and pHT16. Sequencing and primer extension products were visualized on a 6% polyacrylamide sequencing gel using the Otter sequencing system (Owl Separation Systems, Portsmouth, N.H.). Sequencing gels were run at 45 °C at 1100 V for 3 h. All reactions and sequencing were performed according to manufacturer's specifications with one exception; SUPERase-In RNase inhibitor was added to the reactions to prevent RNA degradation (Ambion, Austin, Tex.).

Total RNA sampling and isolation for microarray analysis have been previously described [21]. Reverse transcription and cDNA probe labeling of RNA have been previously described [21].

## 2.9. RNA sampling, isolation, and cDNA labeling for microarray

The methods used for RNA isolation and reverse transcription and cDNA probe labeling have been previously described [21].

## 2.10. Microarray construction and hybridizations

The methods used for construction and validation of the full-genome microarrays and hybridizations were described previously [21]. Briefly, microarrays were spotted with PCR-generated targets (designed to minimize non-specific hybridization) 150–500 bp in size. A total of 3802 genes are present on the microarrays (97% genome coverage). Samples from the stressed cultures were hybridized against oppositely labeled samples from the control culture at the same time point. At least two hybridizations were performed at each time point. To minimize dye biases, dyes (Cy3 and Cy5) were swapped for each replicate.

## 2.11. Microarray analysis

Microarray data were normalized and differential-gene expression was identified using a segmental nearest-neighbor approach [22] coded in MATLAB (MathWorks, Natick, Mass.). Genes showing intensity of 300 units or less were considered not expressed [21].

## 2.12. $\beta$ -Galactosidase analysis

Time course  $\beta$ -galactosidase assays were performed as previously reported [23]. Recombinant *C. acetobutylicum* [strains ATCC 824 (pHT3), 824 (pHT4), 824 (pHT14), and 824 (pHT16)] cells were collected every 2 h from 2 to 24 h. Recombinant *C. acetobutylicum* [strains ATCC 824 (pThiLac), 824 (pTL14), and 824 (pTL16)] cells were collected every 4 h from 4 to 24 h and then 6 h later at 30 h. Time 0 (0 h) is defined as in batch culture fermentation experiments.

## 2.13. Construction of plasmids

### 2.13.1. pHT3/MfeI

The pHT3 plasmid [23] was digested with SmaI creating blunt ends with a C on the 5' end and a G on the 3' end of the cut site. An oligonucleotide containing two XhoI sites, minus the 5' C at the 5' end of the primer and the 3' G at the 3' end of the primer, on both sides of a MfeI site (i.e., MfeIXhoI oligonucleotide in Table 2). The oligonucleotide was made doublestranded by annealing to itself. The oligonucleotide was then ligated to the SmaI digested pHT3. The resulting plasmid, pHT3/MfeI, contains two XhoI sites on either side of an MfeI site cloned into pHT3.

### 2.13.2. pHT14 and pHT16

The DNA fragment containing the intergenic region upstream of *chw14* was amplified by PCR with genomic wild type *C. acetobutylicum* as the template DNA. The upstream forward and reverse primers Upst14F and Upst14R, respectively, were generated by adding an MfeI restriction site and two pyrimidine bases 5' of the MfeI site. The 313-bp amplified DNA fragment was digested

with MfeI and ligated into pHT3/MfeI plasmid to yield pHT14. The plasmid pHT16 was generated in the same manner using the upstream forward and reverse primers Upst16F and Upst16R, respectively, to PCR amplify the 167-bp DNA fragment. The correct clone and forward orientation of the upstream region was confirmed by automated DNA sequencing.

### 2.13.3. pTL14 and pTL16

The DNA fragment containing the intergenic region upstream of *chw14* was amplified by PCR with genomic wild type *C. acetobutylicum* as the template DNA. The upstream forward primer pTL14F was generated by adding a XhoI restriction site and three pyrimidine bases 5' of the XhoI site. The upstream reverse primer pTL14R was generated by adding a BamHI restriction site and three pyrimidine bases 5' of the BamHI site, again for optimized cleavage of the PCR product. The 298-bp amplified DNA fragment was digested with XhoI and BamHI and ligated into XhoI/BamHI-digested pThiLac plasmid to yield pTL14. The correct clone was confirmed by automated DNA sequencing. The plasmid pTL16 was generated in the same manner using the upstream forward and reverse primers pTL16F and pTL16R, respectively, to PCR amplify the 151-bp DNA fragment. The correct clone was confirmed by automated DNA sequencing.

## 2.14. Construction of alignments and trees

ChW proteins were identified from the SMART [24,25] and Pfam [26] architecture databases. Full sequences of ChW proteins were subjected to phylogenetic analysis. Alignments were constructed using the multiple sequence alignment mode of ClustalX [27]. Protein trees were constructed using the neighbor-joining method [28] implemented in the ClustalX program. Bootstrap analysis was performed using 200 and 750 iterations of tree building trials and 1000 random seed generations.

## 2.15. Protein secondary structure prediction

Protein secondary structure prediction was performed by Multivariate Linear Regression Combination as previously reported [29].

## 3. Results

When the *C. acetobutylicum* genome was sequenced [9] a protein family, Clostridial hydrophobic with a conserved tryptophan, specific to this microorganism was identified. Proteomic analysis [7] revealed that two members of the ChW protein family, ChW14 and ChW16/17, required Spo0A. This current study was undertaken in order to expound the requirement of Spo0A, characterize the expression and promoter architecture of *chw14* and *chw16/17*, and to delineate the relationship among all known members of the ChW protein family based on phylogenies and structure predictions. These characterizations used a reporter system, primer extension, phylogenies, and structure predictions.

### 3.1. Identification of ChW proteins

In order to identify all currently known members of the protein family database searches were employed. The identification of ChW proteins by protein architecture analysis using the SMART [24,25] and Pfam [26] databases revealed that 20 of the 29 proteins are *C. acetobutylicum* proteins. The other organisms that contain a single ChW protein are: two strains of *Listeria monocytogenes*, *Arthrobacter sp. FB24*, *Enterococcus faecalis*, *Streptococcus agalactiae*, *Streptomyces*

*coelicolor*, *Shewanella denitrificans*, and *Trichodesmium erythraeum*. Thus, ChW proteins reside almost exclusively in *C. acetobutylicum*.

To investigate how similar the amino acid sequence is between ChW domains in the same protein, since almost all proteins contain more than one domain, and between proteins in the same organism, and between proteins in different microorganisms BLASTP [8] was performed. These data indicate that within the *C. acetobutylicum* species the amino acid identity remains quite high (ranging from 46% to 100%), but diverges in non-*C. acetobutylicum* genera (with a range between 28% and 56%). The few microorganisms that encode only a single ChW protein may have acquired its encoding gene by genetic exchange with *C. acetobutylicum*.

### 3.2. ChW domains

To determine how similar the two members, ChW14 and ChW16/17, are to each other the protein architecture was examined. ChW14 and ChW16/17 have been previously analyzed [7] and the ChW16/17 protein was found to be covalently modified, whereas ChW14 was not. Nonetheless, the genes *chw14* and *chw16/17* appear to both encode the same general protein structure: an N-terminal signal sequence, six contiguous ChW domains, and a C-terminal tail. No other known domains have been identified in either protein. ChW14 and ChW16/17 are highly similar to each other with 70% identity and 82% similarity between the two proteins as analyzed by BLASTP [8]. The ChW domain structures for all 20 *C. acetobutylicum* and the 9 non-*C. acetobutylicum* proteins are illustrated (Fig. 1). Since phylogenetic analysis (Fig. 3B and C) indicates that the ChW domains cluster into groups of threes, i.e., that the ChW domains are not interchangeable but rather are specifically ordered, the individual ChW domains are coded to indicate their position as first (ChW1), second (ChW2), third (ChW3), et cetera along the primary sequence of the protein. Regarding specific amino acids, the absolutely conserved tryptophan may prove to be a key residue while the large number of hydrophobic and small residues may act to keep the interior environment of the ChW protein intact (Fig. 2).

In addition to the phylogenetic data of every third ChW domain clustering together, the number of ChW domains present in the proteins also suggests that the ChW domains function in triplets. Twenty-seven of the 29 ChW proteins contain a multiple of three ChW domains (Table 3). It is, thus, tempting to speculate that the ChW domains function as triplets.

### 3.3. Other protein domains found in ChW proteins

ChW proteins may harbor other domains in addition to the multiple ChW domains. SMART [24,25] recognizes domains that are protein degrading (serine protease, peptidases, extracellular neutral metalloprotease), sugar degrading (beta-mannase), cell wall hydrolyzing (*N*-acetylglucosaminidase), protein binding (leucine rich repeats), and cell adhering (BID<sub>2</sub>).

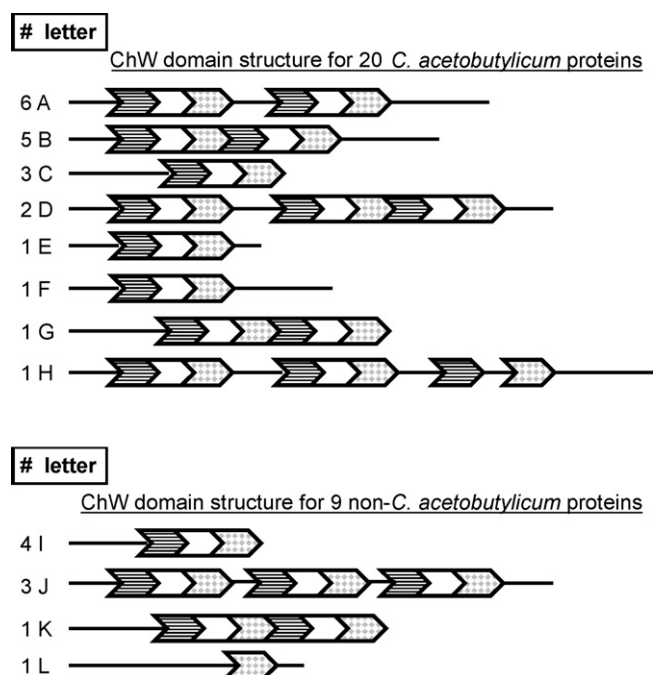


Fig. 1. ChW domain protein structures. The ChW domains are illustrated in the 20 *C. acetobutylicum* and nine non-*C. acetobutylicum* proteins. The symbols used indicate: (#) number of proteins with the indicated ChW domain structure; (▶) a single ChW domain with the pattern within the arrowhead indicating which domains are phylogenetically most related to each other; (—) protein sequence where no ChW domains are located, length of line indicates relative positions of ChW domains. The ChW domain indicated by a striped arrowhead, ▶, represents the first (ChW1), fourth (ChW4), or seventh (ChW7) ChW domain found in the linear protein sequence. The ChW domain indicated by a plain arrowhead, ▶, represents the second (ChW2), fifth (ChW5), or eighth (ChW8) ChW domain found in the linear protein sequence. The ChW domain indicated by a checkered arrowhead, ▶, represents the third (ChW3), sixth (ChW6), or ninth (ChW9) ChW domain found in the linear protein sequence. For simplicity, when a protein structure contains other domain types in addition to ChW domains, only the ChW domains are illustrated. The letter (A–L) labels the architecture type of protein based on how the ChW domains are clustered together. The number refers to the number of ChW proteins with the illustrated domain architecture: (A) proteins have 874, 701, 861, 849, 808 and 857 residues; (B) proteins have 765, 491, 481, 386, 511 residues; (C) proteins have 532, 505 and 531 residues; (D) proteins have 773, 752 residues; (E) protein has 259 residues; (F) protein has 500 residues; (G) protein has 836 residues; (H) protein has 1043 residues; (I) proteins have 438, 540, 450, 450 residues; (J) proteins have 783, 854, 854 residues; (K) protein has 537 residues; and (L) protein has 479 residues. ChW14 and ChW16/17 are both B architecture types of proteins.

Noteworthy among these domains is the cell adhesion domain BID<sub>2</sub>. This domain is found surface proteins such as intimins, which mediate bacterial adhesion to host-cells. Its lectin-like domain suggests that carbohydrate recognition may be important for its adhesive function. A total of 19 *C. acetobutylicum* proteins contain BID<sub>2</sub> domains. Among the other clostridial species, *C. thermocellum* and *C. perfringens* have a single protein with this domain, *C. tetani* has five, and *C. beijerincki* has seven. Because most bacterial species contain few proteins with this type of domain (according to Pfam database), the fact that *C. acetobutylicum* has 19 proteins with the BID<sub>2</sub> domain is curious. Furthermore, of the 20 ChW proteins in *C. acetobutylicum*, 11 also contain a BID<sub>2</sub> domain. Statistically, the association

Cac2367_ChW1	QIQNIGWQD-IRKTAGNTSGTVGQSLRVEAFKINLIN--APAGAKIKYSA
Cac1389_ChW1	QVENIGWQS-TAAQEQISGTGQGLRDEAFKINLVD--APDGAQIKYQA
Cac2290_ChW1	HVQNVGWQN-YVKGALAGTEGQGLRVEAFKINLNN--APAGLNKIKYT
Cap0003_ChW1	HVQNVGWQN-WFSDGVEAGTNGQGLRVEAFKIKLVN--APVGAQISYSA
Cap0002_ChW1	HVQNVGWQN-WVNDGAEAGTDGQALRVEALKVKLIN--APVGAQITYRT
Cac3275_ChW1	HVENIGWQDPWSKDGAEIGTDGKGLRVEALKIKLLN--APAGAKISYQA
Cac3274_ChW1	HVENIGWQDPWSKDGAEIGTDGKGLRVEALKIKLLN--APAGAKISYQA
Cac3280_ChW1	HVQNVGWQTPWAKDGETAGTDGKGLRVEALKIKLVN--APADAKILYQA
Cac3279_ChW1	HVENIGWQAPWAKDGEAEAGTDGKGLRVEALKIKLVN--APADAKITYQA
Cac3278_ChW1	HIQNVGWQNSWCINGEEAGTDGKSLRMEALKINLNT--APADARILYQA
Cac3273_ChW1	HVQNVGWQN-WVSDGEEAGTDGKGLRVEAFKINLNT--APSDAKILYQS
Cac3272_ChW1	HVQNVGWQNPWSNGEEIGTDGKGLRVEAFKIKLVN--APSDAKILYQA
Cac2584_ChW1	HVENIGWQGA-VENGQAEAGTDGKGLRVEALKIKLVN--APEGAHIQYQG
Cac1532_ChW1	HVENIGWQAP-KKDGEEAGTDGKGLRVEALKIKLVN--APAGAHIEYQG
Cac2533_ChW1	HVQNVGWQTA-VTDGAEAGTDGKGLRVEGLKLSLNT--APVGASILYQT
Cac2532_ChW1	HVENIGWQNP-VGDGEEIGTDGKGLRVEALKIKLVN--VPVGSATIQYET
Cac3235_ChW1	HVQNVGWQAFQ-DGDTSGTGGQDLRIEALKMNLNLSNVVPGATINYQV
Cac0539_ChW1	HVQNVGWQS-NVSDGDTAGTGGQGLRMEAIKINYN--NLGLHLHYQT
Cac0538_ChW1	HVSNIGWQD-YVKDAETAGTTGQTLSEAIQMNYPN---NIEHLHLYQA
Cac0540_ChW1	HVQNVGWQD-SVQDGAAGTIEKALRMEAIKINYPN---NAGLHVEYSQA

Fig. 2. ChW domain alignment. The first ChW domain (ChW1), of each of the 20 ChW proteins from *C. acetobutylicum* was aligned using CLUSTALW. The Cac and Cap numbers refer to the gene number of the chromosome and megaplasmid in the sequenced genome, respectively. The conserved tryptophan is highlighted in red.

of both the ChW and the BID<sub>2</sub> domains on *C. acetobutylicum* proteins is highly significant (probability of  $1.5 \times 10^{-22}$ ). Therefore, we cautiously speculate that the ChW domains and the BID<sub>2</sub> domains may operate cooperatively for carbohydrate binding on the cell surface of *C. acetobutylicum*.

Table 3

The number of ChW domains present in *C. acetobutylicum* and non-*C. acetobutylicum* proteins

ChW domain	# Domains
<i>C. acetobutylicum</i> proteins	
Cac0538	3
Cac0539	3
Cac0540	3
Cac1389	9
Cac1532	6
Cac2290	6
Cac2325	6
Cac2367	9
Cac2532	6
Cac2533	6
Cac2584	6
Cac3272	6
Cac3273	3
Cac3274	6
Cac3275	8
Cac3278	3
Cac3279	6
Cac3280	6
Cap0002	6
Cap0003	6
Non- <i>C. acetobutylicum</i> proteins	
EF0123	9
gbs1279	9
LMOF2365_1900	3
LMOh7858_1996	3
SAG1206	9
SCO4256	1
TeryDRAFT_3421	3
ArthDRAFT_0637	6
SdenDRAFT_2946	3

### 3.4. Phylogeny of ChW proteins and of individual ChW domains

In order to investigate the relationship between ChW proteins a phylogeny was constructed using the entire amino acid sequence of the proteins (Fig. 3A). The phylogeny provides further evidence that ChW14 (CAC1532) and ChW16/17 (CAC2584) are most closely related to one another. Interestingly, in four instances, the most closely related proteins are ones whose genes are adjacent to each other (i.e., CAC0538–CAC0540; CAC3274 and CAC3275; CAP0002 and CAP0003; CAC2532 and CAC2533). Moreover, in three of the four cases, the same numbers of ChW domains are present in their respective proteins (Table 3). This arrangement suggests that there may have originally been fewer genes encoding ChW proteins and that they were subsequently copied during a less stable period in the history of the genome of *C. acetobutylicum*.

The non-*C. acetobutylicum* proteins cluster together and appear to be derived from the *C. acetobutylicum* proteins. Moreover, the genera listeria, enterococci, streptococci have been previously shown to cluster with clostridia in the low G + C phylogenetic branch [30,31]. There are a limited number of ChW domains in these genera. It is not an entirely unusual speculation that ChW domains may have transferred from *C. acetobutylicum* to other low G + C bacterial species.

In order to evaluate the relatedness of individual ChW domains to each other we constructed a phylogeny using the approximately 47 residues of the ChW domains alone. In proteins where more than one ChW domain is present, the domains were numbered from N- to C-terminal positions consecutively. The phylogeny that results from using all 159 ChW domains illustrates four major clusterings (Fig. 3B). In each cluster, the 1st, 4th, 7th, and 10th ChW domains group together, as do the 2nd, 5th, 8th, and 11th, as well as the 3rd, 6th, 9th, and 12th. Note that not all ChW proteins contain as many as 12 ChW domains, but the grouping of three is maintained in ChW proteins with fewer ChW domains as well. An example of the 1st, 4th, 7th, and 10th clustering is shown (Fig. 3C). This data indicates that the ChW domains appear to function as triplets of repeats and that the ChW domains may not be interchangeable with one another.

### 3.5. Predicted secondary structure of full length ChW proteins and of ChW domains

The overall secondary structure predictions for ChW14 and ChW16/17 are quite comparable in the percentage of specific secondary structures and the ordering of the  $\alpha$ -helices,  $\beta$ -sheets, and random coils. When the predicted secondary structures are examined over the entire protein length the percentage of  $\alpha$ -helices and  $\beta$ -sheets were both approximately 20% and the remaining 60% of the protein was predicted random coil. This result suggests that these two proteins likely have similar tertiary structures. Tertiary structure prediction is only available when a protein structure has been solved and reported, i.e., able to serve as a template for the protein of interest. No protein structures have been solved with sequence similarity to ChW

proteins. Thus, no protein is available in the non-redundant protein GenBank database to serve as a useful template for ChW14 and ChW16/17.

When the predicted secondary structures of the ChW domains were examined they were not all the same for this conserved segment. The 159 ChW domains in the 29 ChW proteins could be placed into six categories based on predicted secondary structure elements (Table 4). The number of ChW domains and the representative percentage of each secondary structure is detailed for all 159 ChW domains. Individual ChW domains appear to have three subsections. The most common structure within in a single ChW domain is  $\beta$ -sheet (69.2%),  $\alpha$ -helix (55.9%),  $\beta$ -

sheet (69.8%) (Table 4). Therefore, the most common predicted secondary structure of a single ChW domain is  $\beta$ - $\alpha$ - $\beta$  with each secondary structure up to approximately 15 amino acids long, for a total of about 45 residues. A total of 45 residues of secondary structure correlate well with the total average length of the ChW domain of 47 amino acids long.

### 3.6. $\beta$ -Galactosidase assays of *chw14* and *chw16/17* promoters in wild type *C. acetobutylicum* and SKO1

In order to assess when and to what extent *chw14* and *chw16/17* promoters were active we performed  $\beta$ -galactosidase

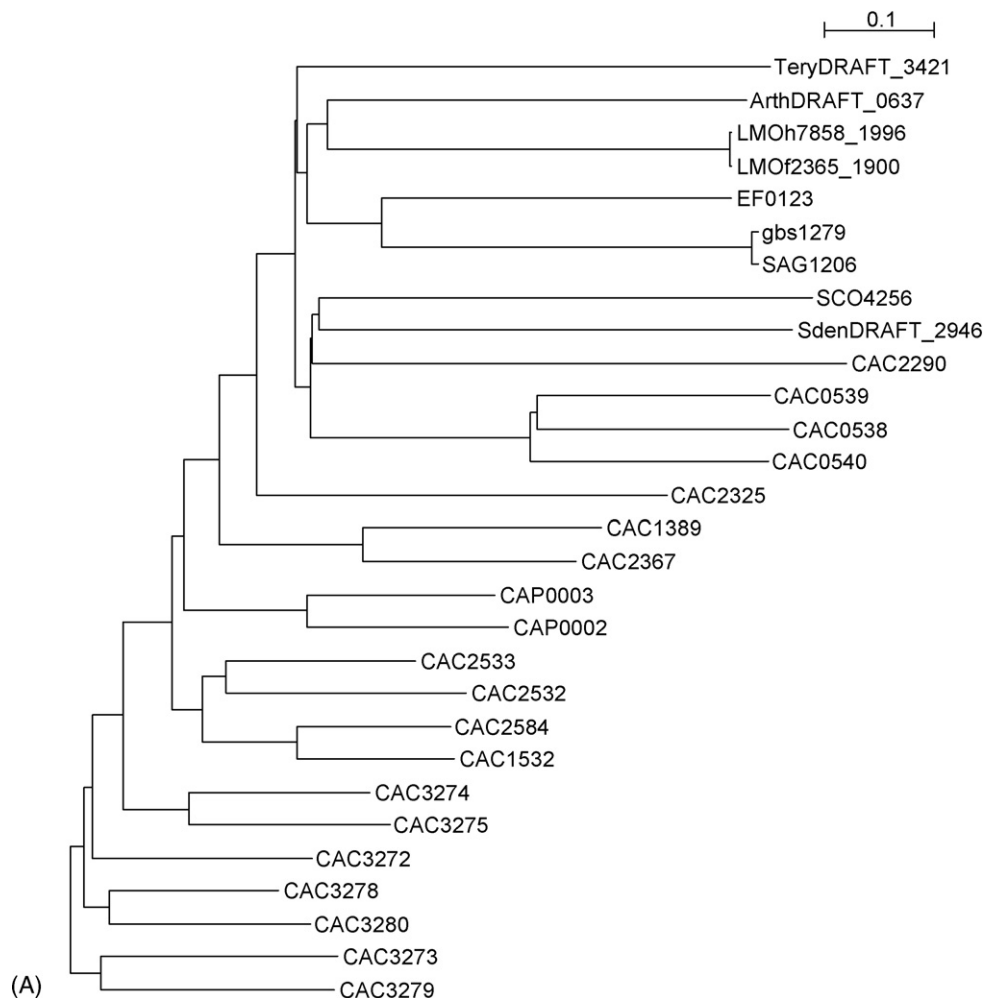


Fig. 3. Phylogenetic analyses. (A) Phylogenetic analysis of entire amino acid sequence of ChW proteins. Multiple alignments were performed on microbial protein sequences containing ChW domains using CLUSTALX and the phylogenetic relationships were calculated using the neighbor-joining approach. Bootstrap values correspond to the frequency of occurrence of node in 1000 bootstrap replicates. Bootstrap values (not shown) of a total of 26 nodes were calculated; 10 nodes had values of 1000 and 7 additional nodes had values greater than 900. Line length indicates percent sequence divergence (scale at top). The abbreviations indicate the microbial organism to which the protein belongs and the gene identification number in the genome encoding the specified proteins: CAC/P, *Clostridium acetobutylicum*; SCO, *Streptomyces coelicolor*; TeryDRAFT, *Trichodesmium erythraeum*; ArthDRAFT, *Arthrobacter* sp. FB24; SdenDRAFT, *Shewanella denitrificans*; LMOh/f, *Listeria monocytogenes*; EF, *Enterococcus faecalis*; and gbs/SAG, *Streptococcus agalactiae*. (B) Phylogenetic analysis of the amino acid sequence representing each individual ChW domain in each protein. Alignments and phylogeny was performed as specified in A with the exception that only the amino acid sequence for the ChW domains alone rather than the sequence for the entire protein was used. The abbreviations are the same as in (A) with the additional specification of the ChW domain position (e.g., ChW1–ChW12) at the end. The complete phylogeny included 159 sequences aligned and indicated three clusters of every third ChW domain (blue, green, and red) from *C. acetobutylicum* and a fourth cluster of non-*C. acetobutylicum* ChW domains (black). Shown are: the 1st (ChW1), 4th (ChW4), 7th (ChW7), and 10th (ChW10) ChW domains in *C. acetobutylicum* in blue; the 2nd (ChW2), 5th (ChW5), 8th (ChW8), and 11th (ChW11) ChW domains in *C. acetobutylicum* in green; the 3rd (ChW3), 6th (ChW6), 9th (ChW9), and 12th (ChW12) ChW domains in *C. acetobutylicum* in red; and the non-*C. acetobutylicum* ChW domains in black. (C) The clustering of the 1st, 4th, 7th, and 10th ChW domains in *C. acetobutylicum* in blue as shown in (B). The abbreviations are the same as in (B).

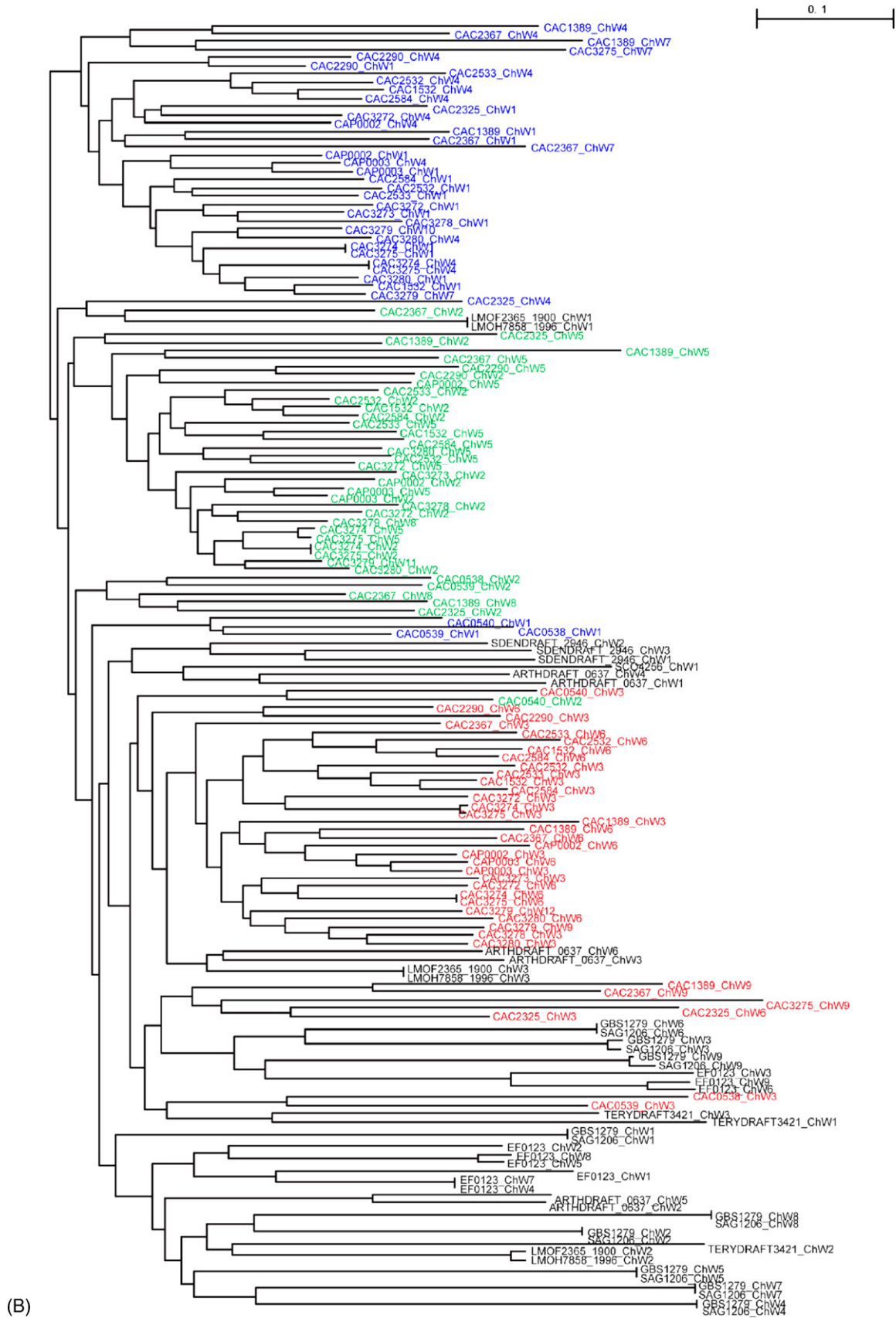


Fig. 3. (Continued)

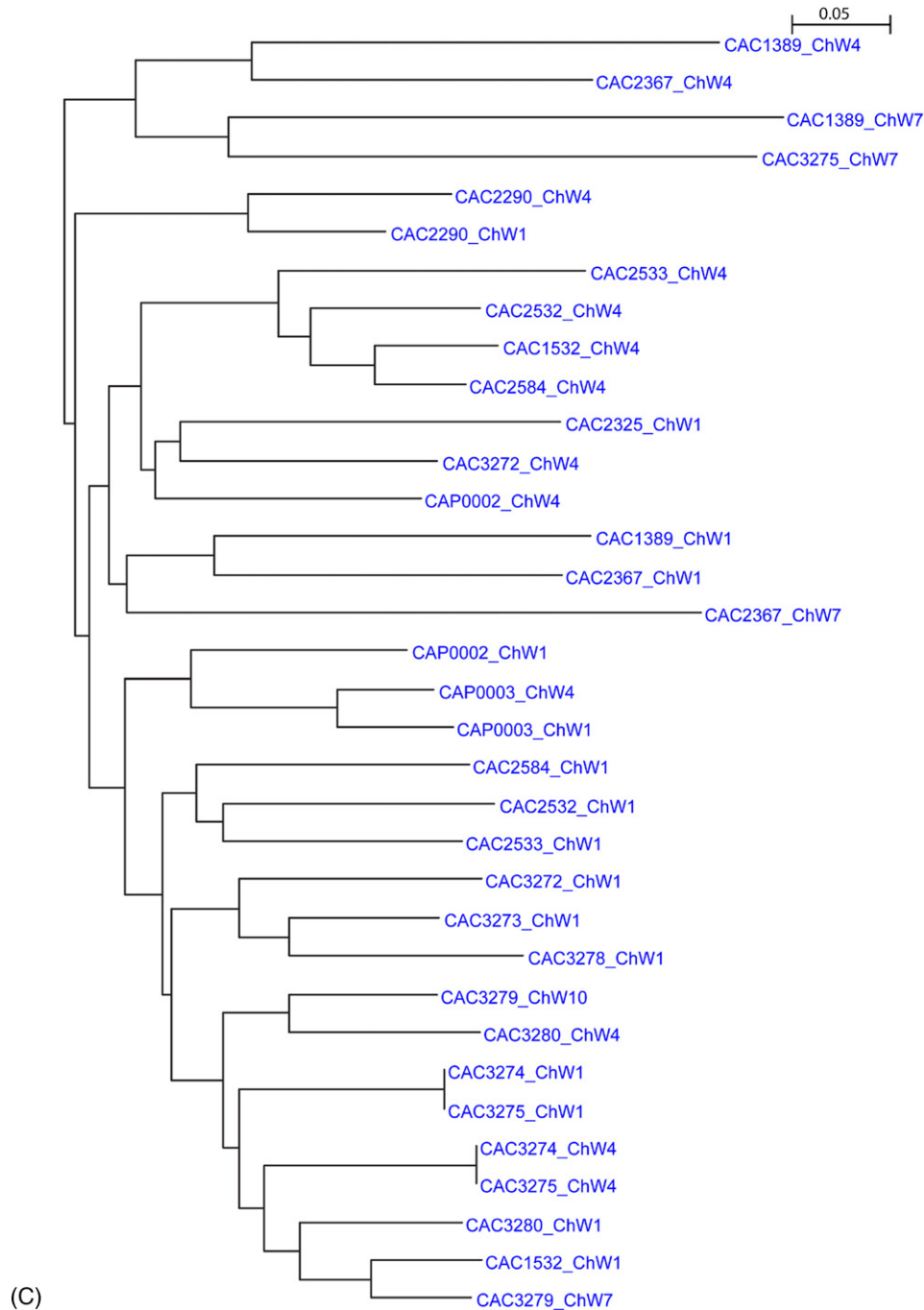


Fig. 3. (Continued).

reporter analyses.  $\beta$ -galactosidase stability tests [23] indicate that the turnover rate of the  $\beta$ -galactosidase enzyme allows it to function as a sensitive reporter.  $\beta$ -galactosidase reporter plasmids, pHT14 and pHT16, were constructed to assess promoter activity for *chw14* and *chw16/17*, respectively (Fig. 4A).  $\beta$ -galactosidase analysis in wild type 824 showed that *chw14* and *chw16/17* promoters are both active to approximately the same extent. The strength of the *chw14* promoter is somewhat less than that of *chw16/17*, with a 0.8-fold ratio of *chw14/chw16/17* at peak activity. The promoters for *chw14* and *chw16/17*, in wild type 824, are both active during mid-exponential phase.

Therefore, the timing and strength similarities between *chw14* and *chw16/17* indicate that both may be involved in the same process providing similar or perhaps redundant functions in *C. acetobutylicum*.

Our hypothesis was that *chw14* and *chw16/17* required Spo0A for expression because these proteins were not detected in SKO1 [7]. Thus, promoter activity was not expected in SKO1. Because SKO1 is an erythromycin resistant strain, a thiamphenicol  $\beta$ -galactosidase reporter plasmid, pThiLac, was used. pThiLac, had been previously shown to function properly in SKO1 [6]. Promoter activity of both *chw14* and *chw16/17*,

Table 4  
Predicted secondary structures of each single ChW domain

Structure	First section <sup>a</sup> # <sup>b</sup> (% <sup>c</sup> )	Second section <sup>a</sup> # <sup>b</sup> (% <sup>c</sup> )	Third section <sup>a</sup> # <sup>b</sup> (% <sup>c</sup> )
α-Helix	14 (8.8)	89 (55.9)	7 (4.4)
β-Sheet	110 (69.2)	37 (23.3)	111 (69.8)
Mix	0 (0)	33 (20.7)	3 (1.9)
Random coil	35 (22.0)	0 (0)	38 (23.9)

<sup>a</sup> Section, each ChW domain appears to have three sections of predicted secondary structure with the first section defined as the most N-terminal and the third section as the most C-terminal section of predicted secondary structure within one ChW domain.

<sup>b</sup> Number of members with a specific predicted secondary structure in the specified section of the ChW domain.

<sup>c</sup> Percentage of total number of members to which the specified number corresponds. The total number is 159 ChW domains examined.

using plasmids pTL14 and pTL16, respectively, appear constitutive in SKO1 (Fig. 4B).

Therefore, Spo0A either directly or via an intermediate player is not required for *chw14* and *chw16/17* expression, but may be required for stability or to delay protein turnover of ChW14 and ChW16/17. In contrast to β-galactosidase activity observed in wild type 824, in SKO1, the promoter of *chw14*

was 1.6-fold greater than that of *chw16/17* at its peak activity. The different β-galactosidase expression patterns in wild type 824 and SKO1 indicate that Spo0A or an intermediate indeed exerts control over *chw14* and *chw16/17* expression, albeit, not required for expression. Because *chw14* is 1.6-fold greater than *chw16/17* in SKO1, this result suggests that in wild type 824 either Spo0A directly or indirectly may act to repress *chw14* or may act to promote *chw16/17*. Copy number differences are not likely to dramatically affect expression level between the two strains, since the replication origin, *repL*, was present on both of the parent plasmids pThiLac and pHT3. Alternatively, Spo0A may affect general background mRNA levels.

### 3.7. Primer extension analysis of *chw14* and *chw16/17* expression and transcription start site

Primer extension analysis has a two-fold objective: (i) to further characterize the expression pattern of *chw14* and *chw16/17* by determining when transcripts are detectable and (ii) to identify putative regulators of *chw14* and *chw16/17* by identifying the precise 5' end of the *chw14* and *chw16/17* transcripts. Results with total RNA from wild type 824 cells obtained from late

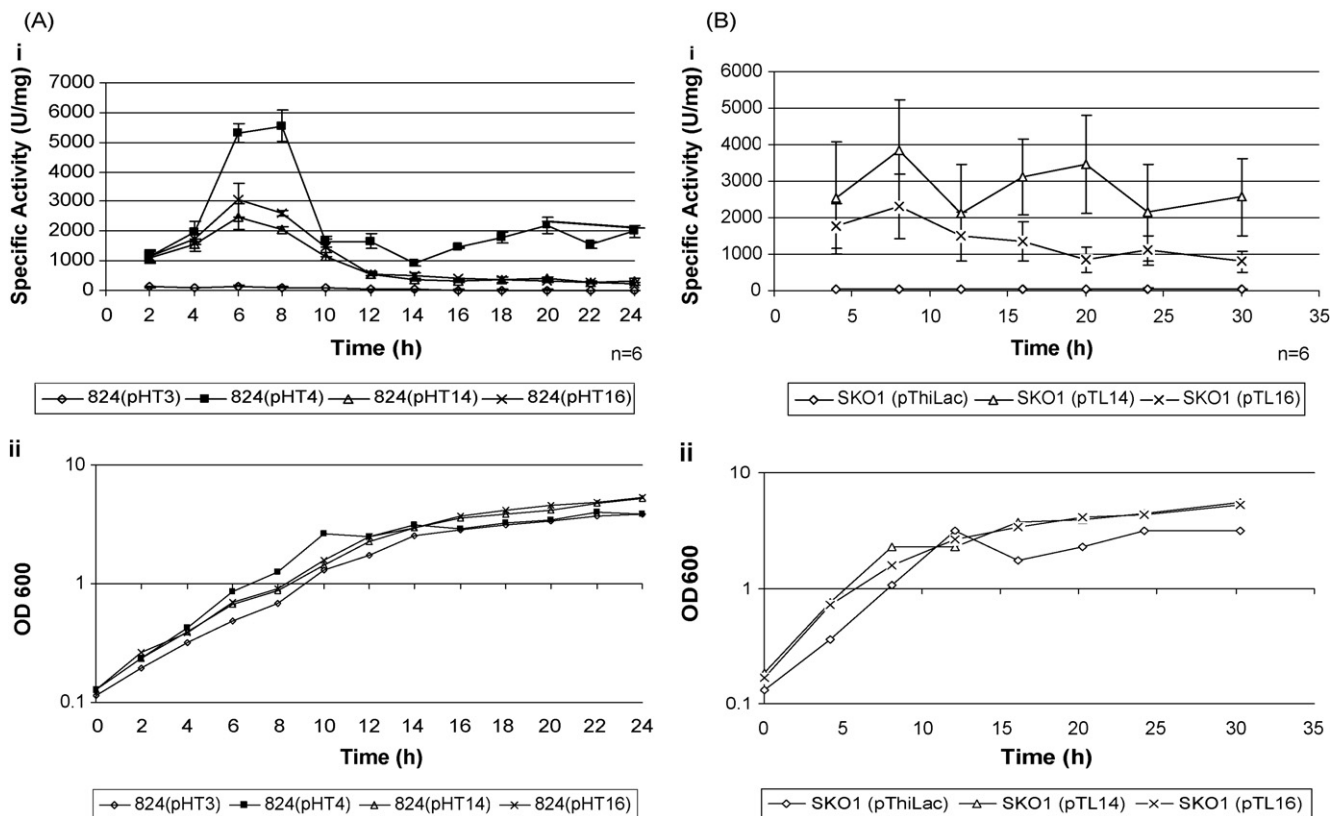


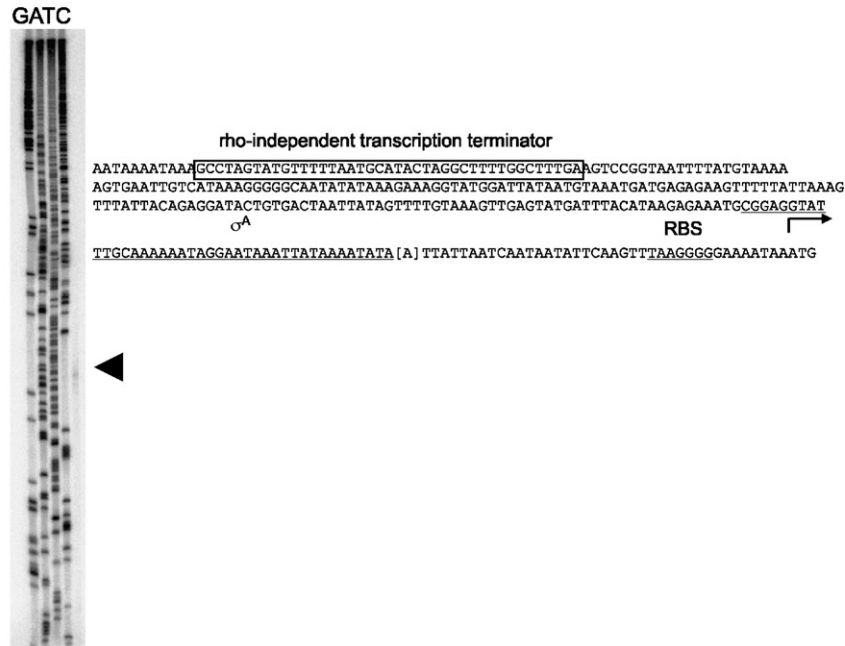
Fig. 4. (A) *In vivo* determinations of promoter activities of *chw14* and *chw16/17* by means of *lacZ* reporter for six replicate analyses of static flask cultures in *C. acetobutylicum* ATCC 824. (i) Specific β-galactosidase activities for the negative control with promoterless β-galactosidase, open diamond, 824(pHT3); the positive control, *ptb* promoter, closed square, 824(pHT4); *chw14* promoter, open triangle, 824(pHT14); and *chw16/17* promoter, X, 824(pHT16). (ii) Relationship of OD<sub>600</sub> averages during growth. (B) *In vivo* determinations of promoter activities of *chw14* and *chw16/17* by means of *lacZ* reporter for six replicate analyses of static flask cultures in SKO1. (i) Specific β-galactosidase activities for the negative control with no promoter β-galactosidase, open diamond, SKO1(pThiLac); *chw14* promoter, open triangle, SKO1(pTL14); and *chw16/17* promoter, X, SKO1(pTL16). (ii) Relationship of OD<sub>600</sub> averages during growth.

exponential through early stationary phase (OD<sub>600</sub> 2.0, 8 h, and 12 h) indicate that *chw14* mRNA is present at very low levels at these time points despite starting with five times the standard amount of total RNA (100 μg) (data not shown). The mRNA

transcripts for *chw16/17* were never detectable in wild type 824 using 100 μg of starting total RNA (data not shown).

In order to increase the sensitivity of detecting *chw14* and *chw16/17* mRNA transcripts the promoter regions of both genes

(A) *chw14*



(B) *chw16*

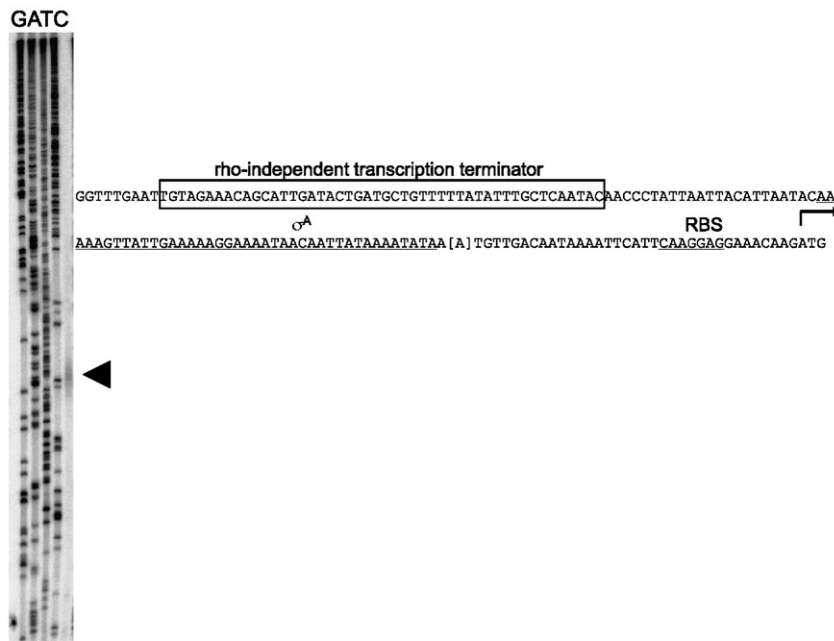


Fig. 5. Primer extension analysis for controlled pH (pH > 5) fermentation with strains 824(pHT14) and 824(pHT16). (A) Primer extension products made with primer designed to the 5' end of *lacZ* are shown. 20 μg of RNA sample from OD<sub>600</sub> 1.2 was used for these experiments and was obtained from strain 824(pHT14) and 824(pHT16). Transcription start sites and promoter region sequence of *chw14* and *chw16/17*. The transcripts (arrowhead) identified by primer extension for *chw14* (A) and *chw16/17* (B) correspond to the transcription start sites shown on the sequences in brackets. The sequencing reactions products (lanes A, T, C, and G) were generated with the same primers as those used for the extension reactions. Also identified within the promoter region sequences are a predicted  $\sigma^A$  binding site (underlined) [32], a putative binding site for Spo0A ( $\sigma^A$  box in lower case), a transcription terminator structure (boxed), and a ribosome-binding site (RBS underlined). The arrows indicate the first codon of the open reading frame.

were placed adjacent to the *lacZ* gene in the reporter plasmids. The higher sequence dosage of *chw14* and *chw16/17* in strains 824(pHT14) and 824(pHT16), respectively, allows primer extension products to be detected in early and late exponential phase, respectively, until early stationary phase in strains 824(pHT14) and 824(pHT16), respectively. In wild type 824, it appears that *chw14* transcripts correlate with the results for its corresponding recombinant reporter strain; transcripts are detectable in late exponential until early stationary phase (data not shown). For *chw14* and *chw16/17*, a single rho-independent terminator structure was predicted upstream of the predicted promoter elements [32], suggesting that for each gene the transcriptional unit begins with that gene. Thus, the transcript end likely represents the true transcriptional start site.

This study examines the promoter structures of *chw14* and *chw16/17* according to the published transcriptional organization of *C. acetobutylicum* [32] and the mapped transcription start sites (this study). The gene *chw14* (CAC1532) and CAC1533 belong to the same transcriptional unit with a predicted  $\sigma^A$  promoter region located 80 to 40 nucleotides upstream of the ATG. A predicted rho-independent transcriptional terminator is

located at positions 292–252 upstream of the ATG. The mapped *chw14* transcriptional start site (Fig. 5A) is 39 nucleotides upstream from the putative translation initiation codon with A as the first transcribed nucleotide. A putative  $\sigma^A$  binding site, TTGCAA (17bp) TAAAAT began 32 nucleotides upstream of the transcriptional start site indicating that this is the promoter structure directing transcription of *chw14*. Additionally, a putative Spo0A binding site, 0A box, was identified 91 nucleotides upstream of the start site.

The gene *chw16/17* (CAC2584) is a predicted singleton with a  $\sigma^A$  promoter region located 77 to 37 nucleotides upstream of the ATG. Positions 148–101 upstream of the ATG harbors a predicted rho-independent transcriptional terminator. The mapped *chw16/17* transcriptional start site (Fig. 5B) is 36 nucleotides upstream from the putative translation initiation codon with A as the first transcribed nucleotide. A putative  $\sigma^A$  binding site, TTGAAA (17bp) TAAAAT began 33 nucleotides upstream of the transcriptional start site indicating that this is the promoter structure directing transcription of *chw16/17*. Additionally, a putative Spo0A binding site, 0A box, was identified 113 nucleotides upstream of the start site.

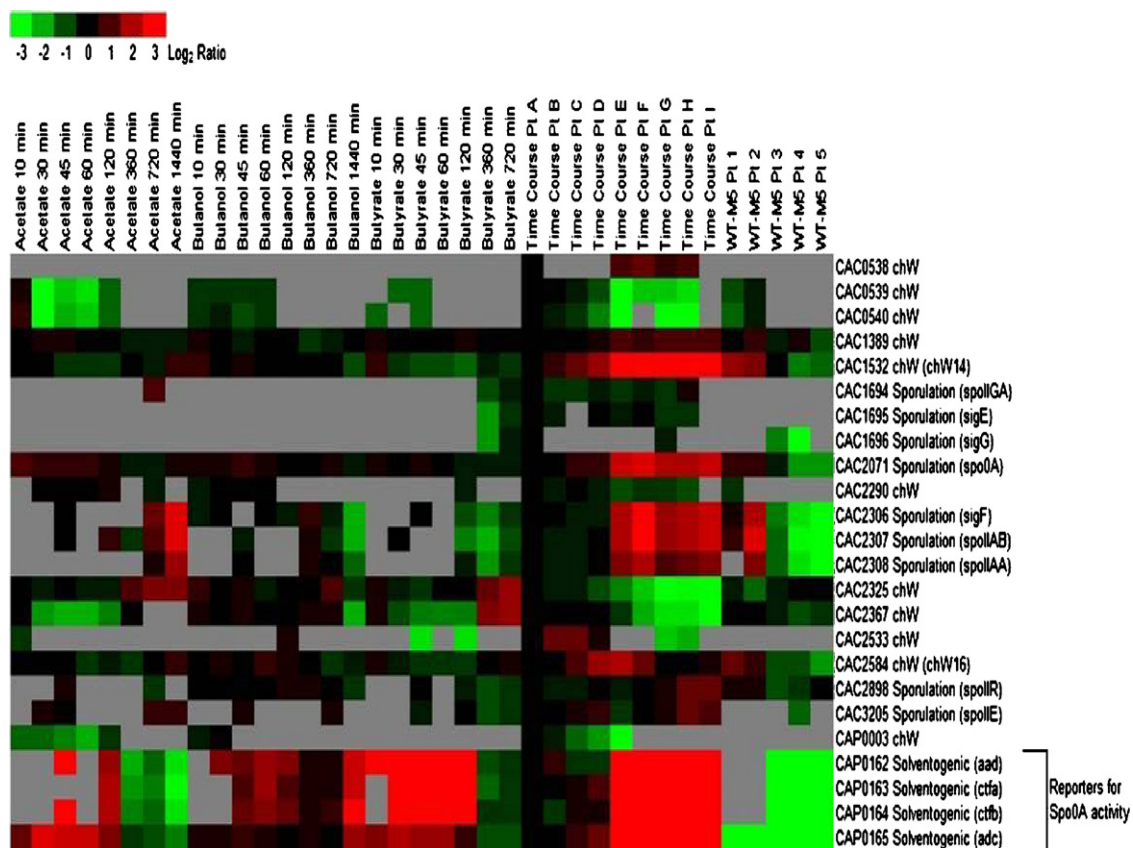


Fig. 6. Comparison of expression ratios of genes encoding ChW proteins. The expression ratios for acetate, butyrate and butanol stress challenges (GEO deposit number GSE5020 (<http://www.ncbi.nlm.nih.gov/geo/>)), time course [21] and M5/WT [22] are shown. For the acetate, butyrate and butanol experiments the sampling time is indicated with time=0 being the moment when the stress started whereas for the time course and M5/WT experiments, the sampling points are labeled as in their original paper. All values presented are log<sub>2</sub> ratios with red indicating a higher mRNA level than the control and green indicating a lower mRNA level than the control. Saturated red and saturated green indicate log<sub>2</sub> ratios equal or greater than 3 or equal or smaller than -3, respectively. Grey rectangles indicate no measurable ratio for the time point or a gene not expressed according to the criterion described [22]. Black rectangles indicate equal levels of expression. For comparison purposes, sporulation and solventogenic markers have been included.

### 3.8. DNA-microarray gene-expression studies

Because the function of genes encoding proteins with ChW repetitive motifs is as of yet undiscovered the significance of their expression patterns cannot yet be fully appreciated. The expression of these genes is summarized below and seen in Fig. 6. These microarray studies examined gene expression during a batch growth; challenges with acetate, butyrate, or butanol; and in a megaplasmid mutant strain (M5). First, the expression of many genes is altered at the transition from exponential to stationary phase. Five genes are downregulated (CAC0539, CAC0540, CAC2290, CAC2325, CAC2367, and CAC2533) and two upregulated (CAC0538 and CAC1389).

Secondly, five genes are downregulated in response to acid challenges, including both *chw14* (CAC1532) and *chw16/17* (CAC2584). The three remaining genes are CAC2367, CAC0539 and CAC0540. These last two genes (CAC0539 and CAC0540), which comprise a predicted two gene operon [32], are also downregulated as a response to butanol challenges. Of the six genes (CAC0539, CAC0540, CAC1389, *chw14* [CAC1532], CAC2367, *chw16/17* [CAC2584]) that show altered expression with the loss of the megaplasmid, two of them are expressed in the exponential phase (*chw14* and *chw16/17*) with the other four genes demonstrating downregulation.

Noteworthy among the genes in the ChW family, CAC3273 has an expression pattern very similar to that of *sigE*. The gene *sigE* is part of a bicistronic operon (*spoIIIGA-sigE*) and encodes a late stage sporulation sigma factor,  $\sigma^E$ . The expression of CAC3273 is either at a very low level or not expressed at all under all conditions tested.

It is also interesting to note that *chw14* and *chw16/17* behave similarly under every condition tested: with the loss of megaplasmid, with a challenge of acid or butanol, or during phases of growth. Thus, their strikingly similar expression pattern enforces the hypothesis that these two genes perform the same or an overlapping function.

Finally, not all of the genes in the ChW family display interesting expression patterns. In fact, 8 of the 20 genes (CAC2532, CAC3272, CAC3274, CAC3275, CAC3275, CAC3278, CAC3279, CAC3280, CAP0002) appear to not be expressed at all. Either these genes are not expressed at a detectable level under the conditions examined or they are possibly pseudogenes. While CAP0002 appears not to be expressed under the conditions of this study its expression has been reported to be upregulated in the absence of *spo0A* in exponential phase, as described below.

The expression of three genes in the ChW family was compared in the SKO1 strain versus wild type *C. acetobutylicum* microarray analysis: CAC2290, CAP0002, and CAP0003 [33]. That report encompassed approximately 25% of the *C. acetobutylicum* genome and thus many of the genes in the ChW family were not part of that analysis. CAC2290 was never differentially expressed in response to the presence or absence of *spo0A*. All three genes appeared to be truly expressed through the entire time course as their log mean intensity was above 300 units, with the exception of points A and H for CAC2290. CAP0003, and to a lesser extent CAP0002, were upregulated

in the SKO1 strain as compared to wild type until the transition to stationary phase [33].

## 4. Discussion

### 4.1. ChW domains

It is interesting to speculate what physiological relevance the novel ChW protein family might have in *C. acetobutylicum*. The functional role the ChW domains play is unknown but they may interact with other proteins or substances in the environment [9]. The high degree of amino acid similarity between ChW14 and ChW16/17 and the same encoded gene structure indicate that these proteins may provide analogous or redundant functions. The roles they play are likely to be specific to the unique physiology of *C. acetobutylicum* because proteins with these domains are largely limited to this microbe. Whereas little is known about the role of ChW proteins, one is known to be modified, i.e., ChW16/17, but the exact post-translational modification remains unknown [7]. While ChW14 protein is the next most closely related to ChW16/17, this protein is not apparently covalently modified [7]. It remains unknown how many other ChW proteins require post-translational modification for their activity, stability, or processing. ChW proteins contain N-terminal signal sequences suggesting a role on the cell surface.

### 4.2. Characteristics of surface proteins Gram-positive organisms

While the study of *C. acetobutylicum* surface proteins has been limited, the surface proteins of many other Gram-positive bacteria have been extensively studied, especially the pathogenic *Staphylococcus* and *Streptococcus* genera. The functions of surface proteins in these bacteria studied are many-fold (for a review see Ref. [16]) and have common requirements. The surface proteins of these non-sporulating Gram-positives have been well characterized and require an N-terminal signal peptide, a conserved LPXTG motif, and a C-terminal sorting signal for cell wall anchoring [34]. However, the same requirements for a signal sequence, LPXTG motif, and sorting signal do not necessarily hold true for the sporulating clostridia and bacilli. While N-terminal signal sequences may or may not be present, LPXTG motifs and C-terminal sorting signals appear to be absent. Regardless of the potential differences in surface proteins between non-sporulating and sporulating Gram-positive bacteria, the domain architectures of proteins in both are similar.

The surface proteins of sporulating and non-sporulating bacteria are modular, harbor repeat elements, are cleaved proteolytically, and may contain a large percentage of aromatic amino acids.

The function of the repeating modules is either to bind to a target or substrate, to cleave the protein, or is unknown. For example, the toxin proteins of the sporulating bacteria *Clostridium difficile*, TcdA and TcdB are not synthesized with an N-terminal signal sequence [35] and how they are secreted from the cell is unknown. TcdA and TcdB contain C-terminal 30-residue

tandem repeats which target eukaryotic cells [36] by binding to specific carbohydrate compounds [37,38] and may also act to cleave the toxin proteins to their mature and active form. Among the uncharacterized repeat structures with targeting elements, aromatic amino acids (W, Y, and F) are found regularly [39]. The aromatic residues may serve as stacking devices for the interaction with carbohydrate ring structures, as has been observed for sugar binding proteins in the periplasm of Gram-negative bacteria [39], but this function has yet to be demonstrated.

The two proteins examined in detail in this study, ChW14 and ChW16/17, have all the characteristics to be genuine surface proteins in *C. acetobutylicum* in that they both contain functionally uncharacterized repetitive ChW modules with a large number of aromatic residues. Moreover, ChW16/17 is cleaved *in vivo* at its N-terminus (data not shown), thus demonstrating a functional signal sequence. Evidence suggests that ChW14 and ChW16/17 are surface proteins, but because their location and function is currently indeterminate we instead investigated their regulation by transcriptional proteins as a means of characterization.

#### 4.3. Correlation between gene expression and protein accumulation

The protein accumulation pattern in wild type 824 of both ChW14 and ChW16/17 [7] agrees well with the gene expression data, as determined by reporter expression, primer extension time course, and microarray analysis (this study). The ChW14 protein increases in abundance during growth, especially at the end of exponential phase. At the transition from exponential to stationary phase the peak levels of ChW14 protein is present. By late stationary phase little ChW14 protein is detected, in contrast to ChW16/17 protein accumulation [7]. The *chw14* promoter is most active during mid-exponential phase and shut-off by the transition to stationary phase (Fig. 4A). The primer extension time course (data not shown) indicates that transcripts are detected from mid-exponential phase until the transition to stationary phase. The microarray data for *chw14* indicates that this gene is exponentially active and its transcripts may be somewhat more stable than *chw16/17* as *chw14* transcripts continue to be detected during late stationary phase at the same level as early stationary phase.

ChW16/17 expression patterns are similar to that of ChW14. Both isoforms of ChW16/17 protein increase somewhat in abundance during growth and especially accumulate during stationary phase with one isoform more abundant than the other [7]. The  $\beta$ -galactosidase data for *chw16/17* is similar in timing and strength to the promoter for *chw14*. The primer extension time course (data not shown) indicates a similar timing of transcript detection as that of *chw14*, with just a slightly earlier detection of transcripts from early-mid-exponential phase until the transition to stationary phase. The pattern of ChW14 and ChW16/17 protein accumulation can be explained by these proteins being more stable than their respective RNA transcripts, such that the ChW14 and ChW16/17 proteins synthesized remain detectable much later than when the promoters are active or the transcripts are detected. The microarray data for *chw16/17* show that this gene is expressed most highly during exponential phase and has

low levels of transcripts detected at late stationary phase when its promoter is shut-off. Thus, *chw16* may have shorter lived transcripts than *chw14* at this late stage.

#### 4.4. Concluding remarks

In summary, while it is unknown what the functional significance ChW14 and ChW16/17 and the entire ChW family provides for the physiology of *C. acetobutylicum*, this study provides a sound starting point for further research into this novel group. This study is a close examination of the expression pattern and promoters of two genes, *chw14* and *chw16/17*, belonging to the ChW family specific to *C. acetobutylicum*. Furthermore, this study portrays a detailed analysis of the relationships among all members of the protein family using phylogenies and structure predictions. Moreover, this study illustrates the different methods that can be utilized to analyze a family of proteins and the genes that encode them when many of the protein domains may be known and unknown.

#### Acknowledgments

We thank Kevin Williams for the addition of the restriction sites MfeI and XhoI to the plasmid pHT3. This work was supported by National Science Foundation grant # BES-0418289 (G.N.B.) and grant # BES-0418157 (E.T.P.).

#### References

- [1] Sauer U, Treuner A, Buchholz M, Santangelo JD, Durre P. Sporulation and primary sigma factor homologous genes in *Clostridium acetobutylicum*. J Bacteriol 1994;176:6572–82.
- [2] Sauer U, Santangelo JD, Treuner A, Buchholz M, Durre P. Sigma factor and sporulation genes in *Clostridium*. FEMS Microbiol Rev 1995;17:331–40.
- [3] Wong J, Sass C, Bennett GN. Sequence and arrangement of genes encoding sigma factors in *Clostridium acetobutylicum* ATCC 824. Gene 1995;153:89–92.
- [4] Harris LM, Welker NE, Papoutsakis ET. Northern, morphological, and fermentation analysis of *spo0A* inactivation and overexpression in *Clostridium acetobutylicum* ATCC 824. J Bacteriol 2002;184:3586–97.
- [5] Durre P, Bohringer M, Nakotte S, Schaffer S, Thormann K, Zickner B. Transcriptional regulation of solventogenesis in *Clostridium acetobutylicum*. J Mol Microbiol Biotechnol 2002;4:295–300.
- [6] Scotcher MC, Bennett GN. SpoIIE regulates sporulation but does not directly affect solventogenesis in *Clostridium acetobutylicum* ATCC 824. J Bacteriol 2005;187:1930–6.
- [7] Sullivan L, Bennett GN. Proteome analysis and comparison of *Clostridium acetobutylicum* ATCC 824 and Spo0A strain variants. J Ind Microbiol Biotechnol 2006;33:298–308.
- [8] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
- [9] Nolling J, Breton G, Omelchenko MV, Makarova KS, Zeng Q, Gibson R, et al. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. J Bacteriol 2001;183:4823–38.
- [10] Sabathe F, Croux C, Cornillot E, Soucaille P. amyP, a reporter gene to study strain degeneration in *Clostridium acetobutylicum* ATCC 824. FEMS Microbiol Lett 2002;210:93–8.
- [11] Lopez-Contreras AM, Gabor K, Martens AA, Renckens BA, Claassen PA, Van Der Oost J, et al. Substrate-induced production and secretion of cellulases by *Clostridium acetobutylicum*. Appl Environ Microbiol 2004;70:5238–43.

- [12] Lopez-Contreras AM, Martens AA, Szijarto N, Mooibroek H, Claassen PA, van der Oost J, et al. Production by *Clostridium acetobutylicum* ATCC 824 of CelG, a cellulosomal glycoside hydrolase belonging to family 9. *Appl Environ Microbiol* 2003;69:869–77.
- [13] Croux C, Canard B, Goma G, Soucaille P. Purification and characterization of an extracellular muramidase of *Clostridium acetobutylicum* ATCC 824 that acts on non-*N*-acetylated peptidoglycan. *Appl Environ Microbiol* 1992;58:1075–81.
- [14] Paquet V, Croux C, Goma G, Soucaille P. Purification and characterization of the extracellular alpha-amylase from *Clostridium acetobutylicum* ATCC 824. *Appl Environ Microbiol* 1991;57:212–8.
- [15] Croux C, Paquet V, Goma G, Soucaille P. Purification and characterization of acidolysin, an acidic metalloprotease produced by *Clostridium acetobutylicum* ATCC 824. *Appl Environ Microbiol* 1990;56:3634–42.
- [16] Navarre WW, Schneewind O. Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiol Mol Biol Rev* 1999;63:174–229.
- [17] Hartmanis MGN, Gatenbeck S. Intermediary Metabolism in *Clostridium acetobutylicum*: Levels of Enzymes Involved in the Formation of Acetate and Butyrate. *Appl Environ Microbiol* 1984;1277–83.
- [18] Mermelstein LD, Papoutsakis ET. In vivo methylation in *Escherichia coli* by the *Bacillus subtilis* phage phi 3T I methyltransferase to protect plasmids from restriction upon transformation of *Clostridium acetobutylicum* ATCC 824. *Appl Environ Microbiol* 1993;59:1077–81.
- [19] Zhao Y, Hindorff LA, Chuang A, Monroe-Augustus M, Lyristis M, Harrison ML, et al. Expression of a cloned cyclopropane fatty acid synthase gene reduces solvent formation in *Clostridium acetobutylicum* ATCC 824. *Appl Environ Microbiol* 2003;69:2831–41.
- [20] Mermelstein LD, Welker NE, Bennett GN, Papoutsakis ET. Expression of cloned homologous fermentative genes in *Clostridium acetobutylicum* ATCC 824. *Biotechnology (N Y)* 1992;10:190–5.
- [21] Alsaker KV, Paredes CJ, Papoutsakis ET. Design, optimization and validation of genomic DNA microarrays for examining the *Clostridium acetobutylicum* transcriptome. *Biotechnol Bioprocess Eng* 2005;10:432–43.
- [22] Yang H, Haddad H, Tomas C, Alsaker K, Papoutsakis ET. A segmental nearest neighbor normalization and gene identification method gives superior results for DNA-array analysis. In: *Proceedings of the National Academy of Sciences of the United States of America* vol. 100. 2003. p. 1122–7.
- [23] Tummala SB, Welker NE, Papoutsakis ET. Development and characterization of a gene expression reporter system for *Clostridium acetobutylicum* ATCC 824. *Appl Environ Microbiol* 1999;65:3793–9.
- [24] Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci USA* 1998;95:5857–64.
- [25] Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 2006;34:D257–60.
- [26] Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, et al. Pfam: clans, web tools and services. *Nucleic Acids Res* 2006;34:D247–51.
- [27] Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 1997;25:4876–82.
- [28] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4:406–25.
- [29] Guermeur Y, Geourjon C, Gallinari P, Deleage G. Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics* 1999;15:413–21.
- [30] Morse R, O'Hanlon K, Collins MD. Phylogenetic, amino acid content and indel analyses of the beta subunit of DNA-dependent RNA polymerase of gram-positive and gram-negative bacteria. *Int J Syst Evol Microbiol* 2002;52:1477–84.
- [31] Onyenwoke RU, Brill JA, Farahi K, Wiegell J. Sporulation genes in members of the low G+C Gram-type-positive phylogenetic branch (Firmicutes). *Arch Microbiol* 2004;182:182–92.
- [32] Paredes CJ, Rigoutsos I, Papoutsakis ET. Transcriptional organization of the *Clostridium acetobutylicum* genome. *Nucleic Acids Res* 2004;32:1973–81.
- [33] Tomas CA, Alsaker KV, Bonarius HP, Hendriksen WT, Yang H, Beamish JA, et al. DNA array-based transcriptional analysis of asporogenous, nonsolventogenic *Clostridium acetobutylicum* strains SKO1 and M5. *J Bacteriol* 2003;185:4539–47.
- [34] Schneewind O, Model P, Fischetti VA. Sorting of protein A to the staphylococcal cell wall. *Cell* 1992;70:267–81.
- [35] von Eichel-Streiber C, Boquet P, Sauerborn M, Thelestam M. Large clostridial cytotoxins—a family of glycosyltransferases modifying small GTP-binding proteins. *Trends Microbiol* 1996;4:375–82.
- [36] Sauerborn M, Leukel P, von Eichel-Streiber C. The C-terminal ligand-binding domain of *Clostridium difficile* toxin A (TcdA) abrogates TcdA-specific binding to cells and prevents mouse lethality. *FEMS Microbiol Lett* 1997;155:45–54.
- [37] Krivan HC, Wilkins TD. Purification of *Clostridium difficile* toxin A by affinity chromatography on immobilized thyroglobulin. *Infect Immun* 1987;55:1873–7.
- [38] Tucker KD, Wilkins TD. Toxin A of *Clostridium difficile* binds to the human carbohydrate antigens I, X, and Y. *Infect Immun* 1991;59:73–8.
- [39] Wren BW. A family of clostridial and streptococcal ligand-binding proteins with conserved C-terminal repeat sequences. *Mol Microbiol* 1991;5:797–803.