Final Report for DOE Grant FG0200ER62924

Annotation of the *Clostridium acetobutylicum* Genome

Principal Investigator:

Michael J. Daly, Ph.D.
Associate Professor
Department of Pathology
Uniformed Services University
of the Health Sciences
4301 Jones Bridge Road
Bethesda, MD 20814-4799
Tel: 301-295-3750
Fax: 301-295-1640

Co-Investigators:

Eugene Koonin, Ph. D.
NCBI
National Institutes of Health
Bethesda, MD 20814-4799

Kira S. Makarova, Ph.D.
NCBI
National Institutes of Health
Bethesda MD 20814-4799

Final Technical Report:

INTRODUCTION

The Clostridia are a diverse group of gram-positive, rod-shaped anaerobes that include several toxin-producing pathogens (notably *C. difficile, C. botulinum, C. tetani,* and *C. perfringens*) and a large number of terrestrial species that produce acetone, butanol, ethanol, isopropanol, and organic acids through fermentation of a variety of carbon sources[1-4]. Isolates of *C. acetobutylicum* were first identified between 1912 and 1914, and these were used to develop an industrial starch-based ABE fermentation process, to produce acetone for gunpowder production, by Chaim Weizmann during the First World War[5-8]. During the 1920s and 1930s, increased demand for butanol led to the establishment of large fermentation factories and a more efficient molasses-based process[6,9]. However, the establishment of more cost-effective petrochemical processes during the 1950s led to the abandonment of the ABE process in all but a few countries. The rise in oil prices during the 1970s stimulated renewed interest in the ABE process, and in the genetic manipulation of *C. acetobutylicum* and related species to improve the yield and purity of solvents from a broader range of fermentation substrates[7,10,11]. This has developed into an active research area over the past two decades.

The type strain, *C. acetobutylicum* ATCC824, was isolated in 1924 from garden soil in Connecticut[12], and is one of the best-studied solventogenic clostridia. Strain relationships among solventogenic Clostridia have been analyzed[13-15], and the ATCC 824 strain was shown to be closely related to the historical Weizmann strain. The ATCC824 strain has been characterized from a physiological point of view and used in a variety of molecular biology and metabolic engineering studies in the US and in Europe[16-24]. This strain is known to utilize a broad range of monosaccharides, disaccharides, starches, and other substrates such as inulin, pectin, whey and xylan but not crystalline cellulose[10,25-28]. Physical mapping of the genome demonstrated that this strain has a 4MB chromosome with 11 ribosomal operons[29] and harbors a large plasmid, about 200 kb in size, which carries the genes involved in solvent formation, hence the name pSOL1[30]. Much work has been done to elucidate the metabolic pathways by which solvents are produced, and to isolate solvent-tolerant, or solvent-overproducing strains[31-37]. Genetic systems have been developed that allow genes to be manipulated in *C. acetobutylicum* ATCC824 and related organisms[10,38-43], and these have been used to develop modified strains with altered solventogenic properties[43-46].

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

The complete genome sequence of *C. acetobutylicum* ATCC824 is expected to facilitate the further design and optimization of genetic engineering tools, and the subsequent development of novel, industrially useful organisms. The sequence also offers the opportunity to compare two moderately related, gram-positive bacterial genomes (*C. acetobutylicum* and *Bacillus subtilis*), and to examine the gene repertoire of a mesophile anaerobe with metabolic capacities that were not previously represented in the collection of complete genomes.

RESULTS

Genome organization. The genome consists of 3,940,880 bp. Genes are distributed fairly evenly with ~51.5% being transcribed from the forward strand and ~49.5% from the complementary strand. A total of 3739 polypeptide-encoding open reading frames (ORFs) and 107 stable RNA genes were identified, accounting for 88% of the genomic DNA, with intergenic regions averaging ~121 bp. A putative replication origin (base 1) and terminus were identified by GC- and AT-skew analysis[47]; the origin marks a strong inflection point in the coding strand and contains several dna-A boxes, as well as gyrA, gyrB, and dnaA genes that are adjacent to the replication origin in many other bacteria. Another strong inflection in the coding strand occurs at the diametrically opposed putative replication terminus (reminiscent of the *Mycoplasma genitalium* genome). The 11 ribosomal operons are clustered in general proximity to the origin of replication, and are all oriented in the same direction as the leading replication fork. The megaplasmid, pSOL1, consists of 192,000 base pairs and appears to encode 178 polypeptides. The single obvious skew inflection was placed at the origin (base 1), although there is no other support for a replication origin at this position (a repA homolog resides ~2.2kb away). In contrast to the genome, there is no obvious coding strand bias in the plasmid.

There appear to be two unrelated cryptic prophages in the genome. The first of these spans approximately 90 kb, and includes approximately 85 genes (CAC1113-CAC1197), with 11 phage-related genes, 3 XerC/XerD recombinase related genes, and a number of DNA processing enzymes. This region contains a strong coding-strand inflection point near its center, and has lower than average GC content. The second apparent prophage appears to span approximately 60 kb and displays similar coding characteristics in approximately 79 genes (CAC1878-CAC1957); slightly higher than average in GC content). Genes for three distinct insertion sequence (IS) related proteins (CAC0248, CAC3531, CAC0656-57) are present on the chromosome. Only one of these is intact, another one is a fragment and the third one has a frameshift. Another frameshifted gene coding for a TnpA-related transposase resides on pSOL1 (CAP0095-96). Thus, it appears that there are no active IS elements in *the C. acetobutylicum* genome.

There are 73 tRNA genes. The isoleucine tRNA could not be identified using standard search methods; this correlates with the displacement of the typical bacterial form of isoleucyl-tRNA with the eukaryotic version, although for other similarly displaced aminoacyl-tRNA synthetases (see below), the cognate tRNA were readily identified.

Comparative analysis. The genome of *C. acetobutylicum* provides us with at least two unique opportunities: 1) compare, for the first time, two large and moderately related gram-positive bacterial genomes, those of *C. acetobutylicum* and *B. subtilis*; 2) investigate the genes that underlie the diverse set of metabolic capabilities so far not represented in the collection of complete genomes.

The median level of sequence similarity[49] between probable orthologs in *C. acetobutylicum* and *B. subtilis* was greater than between *C. acetobutylicum* and any other bacterium, but only by a rather small margin, indicating significant divergence[44]. Compared to the other pairs of evolutionarily relatively close genomes, the Clostridium-Bacillus pair is more distant than the species within the gamma-proteobacterial lineage (*E. coli, H. influenzae, V. cholerae and P. aeruginosa)*, or *H. pylori* and *C. jejunii*; in contrast, the level of divergence *between C. acetobutylicum* and *B. subtilis* is comparable to that between the two spirochetes, *T. pallidum* and *B. burgdorferi*[44]. The comparative analysis of the

spirochete genomes has proved to be highly informative for elucidating the functions of many of their genes and predicting previously undetected aspects of the physiology of these pathogens[50].

A taxonomic breakdown of the closest homologs for the *C. acetobutylicum* proteins immediately reveals the specific relationship with the low-GC Gram-positive bacteria, with the reliable best hits for 35% of the Clostridium protein sequences being to this bacterial lineage[44]. However, nearly as many proteins produced clear best hits to homologs from other taxa[44], which emphasizes the likely major role for lateral gene transfer, a hallmark of microbial evolution.

The same trends appear even more notable when the genome organizations of *C. acetobutylicum* and other bacteria are compared. Gene order is, in general, poorly conserved in the Bacteria, with no extended synteny detected even among relatively close genomes such as *E. coli* and *P. aeuginosa* or *H. influenzae*. In contrast, a genomic dot plot comparison *of C. acetobutylicum* against *B. subtilis* revealed several regions of co-linearity[44]. Thus, bacterial genomes separated by a moderate evolutionary distance, such as *C. acetobutylicum and B. subtilis*, appear to retain the memory of parts of the ancestral gene order, which gives us hope that reliable reconstruction of ancestral genome organizations may become possible once more genomes are sequenced that span the range of conservation from close to intermediate to distant. A systematic mapping of conserved gene strings (many of which form known or predicted operons) on the *Clostridium* genome shows the clear preponderance of shared gene clusters with *B. subtilis*, but also considerable complementary coverage by conserved operons from other bacterial and even archaeal genomes[44]. Altogether, 1243 Clostridium genes (32% of the total predicted number of genes and 40% of the genes with detectable homologs) belong to conserved gene strings; 779 of these are in 271 predicted operons shared with *B. subtilis*[44].

The genome region that shows the greatest level of gene order conservation *between C. acetobutylicum* and *B. subtilis* includes ~200 genes and includes primarily (predicted) operons encoding central cellular functions such as translation and transcription[44]. The multiple genome alignment for this region clearly shows numerous rearrangements of gene clusters, with large-scale co-linearity seen only between *C. acetobutylicum* and *B. subtilis*. The intermediate conservation of gene order seen between *C. acetobutylicum* and *B. subtilis* is likely to be particularly informative in terms of complementing functional predictions based on direct sequence conservation. For example, the predicted large 'superoperon', which contains genes for several components of the translation machinery (def, N-formylmethionyl-tRNA deformylase, fmt, methionyl-tRNA formyl transferase, and fmu, a predicted rRNA methylase) and also protein involved in transcription and replication, additionally include genes yloO (CAC1727), yloP (CAC1728) and yloQ (CAC1729). These genes encode predicted protein phosphatase, serine-threonine protein kinase and a GTPase, respectively. Based on the operon context, the readily testable predictions can be made that yloQ is a previously uncharacterized translation factor, whereas yloO and yloP are likely to play a role in the regulation of translation and/or transcription.

The mosaic picture of operon conservation can be explained by a combination of the processes of horizontal operon transfer, gene (operon) loss and operon disruption (rearrangement). Distinguishing between these phenomena is, in many cases, difficult, but in certain extreme situations, one of the evolutionary routes is clearly preferable. A striking example is the conservation of the nitrogen fixation operon (six genes in a row) between *C. acetobutylicum* and another nitrogen fixator, the *archaeon Methanobacterium thermoautotrophicum*. This particular gene organization so far has not been seen in any other genome except for another Clostridial species, *C. pasteurianum*, in which, interestingly, two genes of the operon are deleted[44]. Similarly, the aromatic amino acid biosynthesis operon is conserved, albeit with local rearrangements, *in C. acetobutylicum, Thermotoga maritima*, and partially in *Chlamydia*[44]. In these and similar cases, it is hard to imagine an evolutionary scenario that does not involve horizontal mobility of these operons, along with operon disruption in some of the bacterial and archaeal lineages.

In general, *C. acetobutylicum* carries the typical complement of genes that are conserved in most bacteria. The only gene that is present in all other bacteria (and, in fact, in all genomes sequenced to date), but is missing in *Clostridium* is that for thymidylate kinase.

A differential genome display analysis for *C. acetobutylicum* and *B. subtilis*, which was performed using the COG system[51], revealed 186 conserved protein families (COGs) that are represented in *Clostridium*, but not in *Bacillus*. Many of these proteins are involved in redox chains that are characteristic of the anaerobic metabolism of Clostridia as opposed to the aerobic metabolism of *B. subtilis*, as well as oxidation/reduction that is required for assimilation of nitrogen and hydrogen. Another group of enzymes belongs to biosynthetic pathways that are present in *Clostridium,* but not in *B. subtilis*, primarily those for certain coenzymes, for example, cyancobalamin (see supplementary material). Conversely, 335 COGs were detected in which *B. subtilis* was represented, whereas *C. acetobutylicum* was not. An obvious part of this set are genes coding for components of aerobic redox chains such as cytochromes and proteins involved in the assembly of cytochrome complexes. Also missing are a variety of membrane transporters, the glycine cleavage system that is present in the majority of bacteria. Several metabolic pathways are incomplete, for example, a considerable part of the TCA cycle and molybdopterin biosynthesis is missing. The TCA cycle is incomplete in many prokaryotes, but in most of these cases, the chain of reactions producing three key precursors, 2-oxoglutarate, succinyl-CoA and fumarate, can proceed in either the oxidative or the reductive direction[52]. In *C. acetobutylicum*, citrate synthase, aconitase and isocitrate dehydrogenase are missing. It appears, however, that what remains of the TCA cycle could function in the reductive direction[44]. The counterparts of enzymes involved in succinyl-CoA and oxaloacetate formation in other organisms are missing in *C. acetobutylicum.* However, it encodes acetoacetyl:acyl CoA-transferase that catalyzes butyryl-CoA formation in solventogenesis (CAP0163-0164) and might also utilize succinate for the synthesis of succinyl-CoA and 2-oxoacid:ferredoxin oxidoreductase (CAC2458-2459) that could catalyze 2-oxoglutarate formation from succinyl-CoA[44]. Succinate dehydrogenase/fumarate reductase, the enzyme that normally catalyzes the reduction of fumarate to succinate, seems to be missing in *C. acetobutylicum.* However, this reaction is linked to the electron-transfer chain and might be supported by another dehydrogenase whose identity could not be easily determined.

The set of sporulation genes in *Clostridium acetobutylicum* surprisingly differs from the set that has been well studied in *Bacillus subtilis*[53]. The number and diversity of detectable sporulation genes in Clostridium is much smaller. The most dramatic difference was observed among the SpoV genes. Clostridium does not have *spoVF, spoVK, spoVM and spoVN* genes, the disruption of which in *B.subtilis* leads to formation of immature spores sensitive to heat, organic solvents, and lysozyme[53]. *B. subtilis* has 22 COT genes that are responsible for coat biosynthesis; only 14 of these genes are conserved in *C. acetobutylicum.* Similarly, *Bacillus* has 21 GER genes, 7 of which are represented by orthologs in *Clostridium.* Many of the missing GER genes encode various receptors of germination, which appear to be different in these bacteria. Furthermore, Clostridium does not have an ortholog of the cell-division-initiation gene *divIC*[53], which is essential in *B. subtilis*, suggesting differences in the mechanism of septum formation.

Many of the Clostridial genes that are missing in *B. subtilis* seem to show distinct evolutionary affinities and probably have been acquired via horizontal transfer. In particular, a significant number of Clostridial genes are conserved in all archaea whose genomes have been sequenced to date, but are present in bacteria only sporadically[44]. Many of these genes encode various redox proteins which reflects the similarity between the anaerobic redox chains in archaea and Clostridia. For most of these 'archaeal' genes found in bacteria, the probable evolutionary model is a single entry into the bacterial world by horizontal transfer from the Archaea followed by dissemination among the Bacteria. In several cases, however, direct gene transfer from archaea into the Clostridial genome seems likely; examples include the genes for a metal-dependent hydrolase of the metallo-beta-lactamase superfamily (CAC0535), a calcineurin-like phosphatase which has undergone duplication in *Clostridium*, probably

subsequent to the acquisition of an archaeal gene (CAC1010 and CAC1078), and a predicted DNA-binding protein (CAC3166). Another group of Clostridial genes includes probable eukaryotic acquisitions[44]. As with 'archaeal' genes, the scenario of a single entry into the bacterial world followed by horizontal dissemination is likely for many of these genes, for example, that for the FHA domain discussed below. However, about 50 genes in *C. acetobutylicum* could have been directly hijacked from eukaryotes[44]. An interesting example is the nucleotide pyrophosphatase, which is encoded within one of the gene clusters including genes for FHA-containing proteins[44] and therefore may be also implicated in signaling. A noticed previously, lateral acquisition of some of aminoacyl-tRNA synthetases from eukaryotes, accompanied by displacement of the original copies, seems to have occurred repeatedly in bacterial evolution. *Clostridium* is no exception, with its arginyl-tRNA synthetase showing a clear eukaryotic affinity.

Most of the essential functions *in C. acetobutylicum* and *B. subtilis* are associated with readily detectable orthologs, but there are also notable cases of non-orthologous gene displacement[44]. Examples include glycyl-tRNA synthetase, which is represented by the typical bacterial, two-subunit form in *B. subtilis* and by the one-subunit archaeal-eukaryotic version in *Clostridium*, and uracil-DNA glycosylase, similarly represented by the classical bacterial enzyme (ortholog of E. coli Ung) and by the archaeal version in *Clostridium*[44]. In many cases, while an apparent orthologous relationship was detected between a Clostridial protein and its counterpart from *B. subtilis*, there was nevertheless a clear difference in the domain architectures[44]. Notable examples of unusual domain organizations from *Clostridium* include the FtsK ATPase, which is fused to the FHA domain (see below), a Pkn2 family protein kinase fused to TPR repeats (CAC0404), and another ATPase fused to a LexA-like DNA-binding domain (CAC1793). The evolution of another set of genes seems to have involved xenologous gene displacement whereby a gene in one of the compared genomes (*C. acetobutylicum* or *B. subtilis*) is displaced by the ortholog from a distant branch of the phylogenetic tree, e.g. eukaryotes[44]. Characteristically, this evolutionary pattern was detected for three aminoacyl-tRNA synthetases, those for isoleucine, arginine and histidine; in each of these cases, *C. acetobutylicum* possesses the archaeal-eukaryotic version as opposed to the typical bacterial versions found in *B. subtilis.* Another interesting example of xenologous displacement involves the two forms of Clostridial ribonucleotide reductase, neither of which groups with the counterparts from *B. subtilis* in phylogenetic trees. One of the ribonucleotide reductase genes in *B. subtilis* contains the single intein in that organism; *C. acetobutylicum* has no inteins, however. These observations show that there had been significant horizontal exchange of genes between the Clostridium lineage and certain archaea and/or eukarya subsequent to its divergence from the Bacillus lineage.

The results of systematic analysis of the protein families that are specifically expanded in *C. acetobutylicum* are largely compatible with the current knowledge of the physiology of the bacterium[44]. For example, distinct families of proteins involved in sporulation, anaerobic energy conversion and carbohydrate degradation were identified[44]. A so far unique feature is the presence of four diverged copies of the single-stranded DNA-binding proteins, an essential component of the replication machinery that is present in one or two copies in all other sequenced bacterial genomes. In addition, this analysis revealed remarkable aspects of the signal transduction system in this bacterium. Of particular interest is the proliferation of the phosphopeptide-specific, protein-protein interaction module, the FHA domain, which is generally rare in the Bacteria[54]. *C. acetobutylicum* encodes five FHA-domain-containing proteins, which is comparable to the number of these domains in other bacteria with versatile Ser/Thr-phosphorylation-based signaling, namely *Mycobacterium tuberculosis* (10) and *Synechocystis sp.* (7); most of the other bacteria do not encode FHA domains or possess just one copy[55]. Four of the genes coding for FHA-domain-containing proteins in *C. acetobutylicum* belong to two partially similar gene clusters that are unique for *Clostridium* and additionally include genes for other phosphorylation-dependent signaling proteins, namely predicted protein kinases and a phosphatases[44]. The fusion of the FHA domain with the FtsK ATPases, which is involved in

chromosome segregation and the presence, in one of the clusters, of an ATPase of the MinD family, also involved in chromosome partitioning, suggests previously unsuspected regulation of cell division in *Clostridium* via reversible protein phosphorylation. The fifth FHA-domain-containing seems to belong to yet another operon involved in cell division[44], which is compatible with the hypothesis on the role of phosphorylation in the regulation of this process in *Clostridium*. Another signaling system that is predicted to play a prominent role in *Clostridium* on the basis of protein family expansion analysis includes the so-called HD-GYP domains that are suspected to possess cyclic diguanylate phosphoesterase activity[44]; the only comparable expansion of the HD-GYP domain is seen in *Thermotoga*. The HD-GYP proteins could play a major role in sensing the redox state of the environment in *Clostridium*[56].

The solventogenesis pathways of *Clostridium* involve the formation of acetone, acetate, butanol, butyrate and ethanol from acetyl-CoA[57]. Two mechanisms of butanol formation have been identified in *C. acetobutylicum,* one of which is associated with solventogenesis (production of butanol, ethanol and acetone) and the other one with alcohologenesis (production of butanol and ethanol only). The genes involved in solventogenesis have been previously identified on the megaplasmid and sequenced[56], but the genes responsible for alcohologenesis were unknown. The genome sequencing allows the identification of a second alcohol-aldehyde dehydrogenase (CAP0035), a pyruvate decarboxylase (CAP0025) and an ethanol dehydrogenase (CAP0059) probably involved in this alcohologenic metabolism[44]. The enzymes involved in the final steps of solvent formation show variable phylogenetic profiles, and in particular, several of them appear to be specifically related to the homologs from the archaeon *Archaeoglobus fulgidus* [44]. In contrast, the genes for the two subunits of another key enzyme of acetone and butyrate pathways, acetoacetyl-CoA: acyl-CoA transferase, show a clear proteobacterial affinity. Together with the fact that a significant subset of the solventogenesis enzymes is encoded on the Clostridial megaplasmid, these observations suggest that these pathways could have evolved via a complex sequence of gene/operon acquisition events. The megaplasmid also carries second copies of genes involved in PTS-type sugar transport (CAP0066-68), glycolysis (aldolase, CAP0064) and central metabolism (thiolase, CAP0078) illustrating how this microorganism has evolved an efficient system for solvent production and the functional complementarity between the chromosome and the plasmid in *C. acetobutylicum.*

The cellulosome, the macromolecular complex for cellulose degradation, has been genetically and biochemically characterized in four *Clostridium species (C. thermocellum, C. cellulovorans, C. cellulolyticum*, and *C. josui*), but not in *C. acetobutylicum* (which is able to hydrolyze carboxy-methyl cellulose but not amorphous or crystalline cellulose[58]. The proteins of the cellulosome contain a C-terminal Ca^{2+}-binding dockerin domain, which is required for the binding to the cohesin domains of a scaffolding protein[59,60]. Genome sequence analysis revealed at least 11 proteins that are confidently identified as cellulosome components[44]. Most of these genes are organized in an operon-like cluster (CAC910-CAC919) with similar gene order to those in mesophilic *C. cellulolyticum* and *C. cellulovorans*, as distinct from the more dispersed organization in the thermophile, *C. thermocellum*[61,62]. The large glycohydrolase CAC3469 is the homolog of EngE of *C. cellulovorans,* which is also encoded away from the main cellulosome gene cluster. Unlike EngE CAC3469 possesses additional cell adhesion domain[44]. This protein contains S-layer homology domains and cell adhesion domains similar to SlpA, one of the anchoring proteins of *C. themocellum.* The presence of the short cohesion domain protein CAC914 suggests a role in cellulosome function related to that of the HbpA protein of *C. cellulovorans*[62]. The other dockerin-domain containing proteins, those of the GH48, GH5, and GH9 families, might interact with CAC910, the ortholog of the scaffolding protein CbpA. Generally, although the cellulosome has not been detected in *C. acetobutylicum*, the number of relevant proteins and domains would seem sufficient to encode the various combinations of cellulose-binding and hydrolytic proteins found in this complex. An interaction between CAC3469 and CAC910 could

be speculatively proposed as a means of anchoring a potential cellulosome-like structure to the peptidoglycan.

In work analyzing the cellulolytic activities of *C. acetobutylicum* strains it was found that NRRL B 527 could hydrolyze Avicel and acid swollen cellulose but *C. acetobutylicum* ATCC 824 could not[63]. The subsequent taxonomic and historical analysis of these strains[13,14] indicate a close relationship and suggest further investigation of the cluster from strain B 527 would be informative in elucidating the reason for the different cellulolytic activities of the two strains. Further work is required to resolve these issues and to determine the exact functions of the cellulosome subunits in *C. acetobutylicum.*

In addition to the known cellulosome components, *C. acetobutylicum* encodes numerous other proteins that are predicted to be involved in the degradation of xylan, levan, pectin, starch and other polysaccharides. Altogether, there seem to be over 90 genes encoding proteins implicated in these processes, including representatives of at least 14 distinct families of glycosyl hydrolases. In particular, a predicted operon located on the *C. acetobutylicum* megaplasmid (CAP0114-CAP0120) mostly consists of genes encoding xylan degradation enzymes. Similarly to the cellulosome components, these enzymes possess complex domain architectures, with the oligosaccharide-binding ricin domain[64] typically present at the C-terminus; the addition of ricin domain is (so far) a unique feature of this postulated novel system for xylan degradation in Clostridium[44]. Two of the putative xylanases presumably correspond to previously reported enzymes of xylan degradation isolated from *C. acetobutylicum* ATCC 824[65].

A number of sugar PTS transport system genes are found as well as the corresponding regulatory system analogs (eg Hpr, ptsK, CcpA) which couple transport signals to genetic regulation of degradative operons[66,67]. Non PTS mediated uptake of certain sugars especially pentoses has been found in several Clostridial species[10]. Many primary active transporters, including ABC-type transporters and P-type ATPases, electrochemical potential-driven transporters, channels and pores, and uncharacterized transporters were detected among the gene products of *C. acetobutylicum*[44]. There is, however, no ortholog of the glucose facilitator of *B. subtilis*[68].

Along with previously characterized molecular complexes involved in extracellular hydrolysis of organic polymers, a novel system possibly related to these processes was detected. The signature of this system is a previously undetected domain with a distinct repetitive structure, which we designated as "ChW repeats" (Clostridial hydrophobic, with a conserved W, tryptophan)[44]. So far, the only non-Clostridial protein containing similar repeats was detected in *Streptomyces*[44]. All proteins containing ChW repeats contain confidently predicted signal peptides at their N-termini and do not contain predicted transmembrane helices, which suggests that all of them are secreted[44]. Some of the ChW-repeat proteins contain additional enzymatic domains such as glycosyl hydrolases or proteases, which implicates them in the degradation of polysaccharides and proteins. Several ChW-repeat proteins also contain domains that are involved in cell interactions such as the cell-adhesion domain[69] and the leucine rich repeat (internalin) domain[70]. The internalin domain has been shown to play a critical role in the host cell invasion by the bacterial pathogen *Listeria monocytogenes*[70]. In *C. acetobutylicum,* these domains could be responsible for interactions with plant cells. ChW repeats also could function in either substrate-binding or protein-protein interactions. The specific expansion of this domain in *C. acetobutylicum* suggests the existence of a novel molecular system, which partially resembles the cellulosome, and could also form structurally distinct multisubunit complexes involved in polymer degradation and interaction with the environment. Elucidation of the function of this system is expected to shed light on the unique physiology of *C. acetobutylicum.*

The extreme diversity of the domain architectures of the proteins that comprise the cellulosome and other predicted polymer degradation systems suggests that such complexes are highly dynamic not only in terms of the subunit stoichiometry[58], but also with respect to the genetic organization, with horizontal gene transfer, domain shuffling and non-orthologous gene displacement playing pivotal roles in their evolution. *C. acetobutylicum* is the first sequenced bacterial genome with such a remarkable

abundance of polymer degradation systems, which makes it a model for future studies on other bacteria with similar lifestyles.

REFERENCES:

1. Woods, D.R. The Clostridia and Biotechnology. (Butterworth-Heinemann, Boston; 1993)

2. Keis, S., Bennett, C.F., Ward, V.K. & Jones, D.T. Taxonomy and phylogeny of industrial solvent-producing clostridia. *Int J Syst Bacteriol* **45**, 693-705 (1995).

3. Stackebrandt, E., Rainey F.A. in The Clostridia: Molecular Biology and Pathogenesis. (ed. J. Rood, McClane, B.A., Songer, J.G., Titball, R.W.) (Academic Press, San Diego; 1997).

4. Stackebrandt, E., Kramer, I., Swiderski, J. & Hippe, H. Phylogenetic basis for a taxonomic dissection of the genus Clostridium. *FEMS Immunol. Med. Microbiol.* **24**, 253-258 (1999).

5. Weizmann, C. Improvements in the bacterial fermentation of carbohydrates and in bacterial culture for same. British Patent 4845, Filed Mar 29 1915, Issued Mar 6 1919.

6. Jones, D.T. & Woods, D.R. Acetone-butanol fermentation revisited. *Microbiol Rev* **50**, 484-524 (1986).

7. Woods, D.R. The genetic engineering of microbial solvent production. *Trends Biotechnol* **13**, 259-264 (1995).

8. Durre, P. New insights and novel developments in clostridial acetone/butanol/isopropanol fermentation. *Appl. Microbiol. Biotechnol.* **49**, 639-648 (1998).

9. Gabriel, C. L. Butanol fermentation process. *Ind. Eng. Chem.* **20**, 1063-1067 1928.

10. Mitchell, W.J. Physiology of carbohydrate to solvent conversion by clostridia. *Adv. Mirob. Physiol.* **39**, 31-130 (1998).

11. Qureshi, N. & Blaschek, H.P. Production of acetone butanol ethanol (ABE) by a hyper-producing mutant strain of Clostridium beijerinckii BA101 and recovery by pervaporation. *Biotechnol Prog* **15**, 594-602 (1999).

12. Weyer, E.R., Rettger, L.F. A comparative study of six different strains of the organism commonly concerned in large-scale production of butyl alcohol and acetone by the biological process. *J. Bacteriol.* **14**, 399-424 (1927).

13. Johnson, J.L. & Chen, J.-S. Taxonomic relationships among strains of *Clostridium acetobutylicum* and other phenotypically similar organisms. *FEMS Microbiol. Rev.* **17**, 233-240 (1995).

14. Jones, D. T. & Keis, S. Origins and relationships of industrial solvent-producing clostridial strains. *FEMS Microbiol. Rev.* **17**, 223-232 (1995).

15. Cornillot, E., & Soucaille, P. Solvent-forming genes in Clostridia. Nature **380**, 489 (1996)

16. Gottwald, M. & Gottschalk, G. The internal pH of *Clostridium acetobutylicum* and its effect on the shift from acid to solvent formation. *Archives of Microbiology* **143**, 42-46 (1985).

17. Terraciano, J.S., Rapaport, E. & Kashket, E.R. Stress and growth-phase associated proteins of *Clostridium acetobutylicum*. *Appl. Environ. Microbiol* **54**, 1989-1995 (1988).

18. Mattson, D.M. & Rogers, P. Analysis of Tn*916*-induced mutants of *Clostridium acetobutylicum* altered in solventogenesis and sporulation. *J.Ind. Microbiol.* **13**, 258-268 (1994).

19. Girbal L. & Soucaille, P. Regulation of Clostridium acetobutylicum metabolism as revealed by mixed substrate steady state continuous culture; role of NADH/NAD ratio and ATP pool. *J. Bacteriol.* **176**, 6433-6438 (1994).

20. Bahl, H., Muller, H., Behrens, S., Joseph, H. & Narberhaus, F. Expression of heat shock genes in *Clostridium acetobutylicum. FEMS Microbiology Reviews* **17**, 341-348 (1995).

21. Duerre, P. et al. Solventogenic enzymes of Clostridium acetobutylicum: catalytic properties, genetic organization, and transcriptional regulation. *FEMS Microbiol. Rev.* **17**, 251-262 (1995).

22. Petitdemange, H., Cherrier, C., Bengone, J.M. & Gay, R. Study of the NADH and NADPH-ferredoxin oxidoreductase activities in *Clostridium acetobutylicum. Can. J. Microbiol.* **23**, 152-160 (1997).

23. Girbal, L. & Soucaille, P. Regulation of solvent production in *Clostridium acetobutylicum. Trends in Biotechnology* **16**, 11-16 (1998).

24. Papoutsakis, E.T. & Bennett, G.N. Molecular regulation and metabolic engineering of solvent production by Clostridium acetobutylicum. *Bioprocess Technol.* **24**, 253-279 (1999).

25. Blanchet, D., Marchal, R. & Vandecasteele, J.P. in Fr. Demande 12 pp. ((Institut Francais du Petrole, Fr.)., Fr; 1985).

26. Lee, S.F., Forsberg, C.W. & Gibbins, L.N. Xylanolytic activity of Clostridium acetobutylicum. *Appl. Environ. Microbiol.* **50**, 1068-1076 (1985).

27. Bronnenmeier, K., Staudenbauer, W.L. in The Clostridia and Biotechnology. (ed. D. Woods) 443 (Butterworth-Heinemann, Reading, MA; 1993).

28. Morris, J.G. in The Clostridia and Biotechnology. (ed. D. Woods) 443 (Butterworth-Heinemann, Reading, MA; 1993).

29. Cornillot, E., Croux, C. & Soucaille, P. Physical and genetic map of the Clostridium acetobutylicum ATCC 824 chromosome. *J. Bacteriol.* **179**, 7426-7434 (1997).

30. Cornillot, E., Nair, R.V., Papoutsakis, E.T. & Soucaille, P. The genes for butanol and acetone formation in Clostridium acetobutylicum ATCC 824 reside on a large plasmid whose loss leads to degeneration of the strain. *J Bacteriol* **179**, 5442-5447 (1997).

31. Soucaille, P., Joliff, G., Izard, A. & Goma, G. Butanol tolerance and autobacteriocin production by Clostridium acetobutylicum. *Curr. Microbiol.* **14**, 295-299 (1987).

32. Rogers, P. & Gottschalk, G. Biochemistry and regulation of acid and solvent production in clostridia. *Biotechnol. Ser.* **25**, 25-50 (1993).

33. Girbal, L., Croux, C., Vasconcelos, I. & Soucaille, P. Regulation of metabolic shifts in Clostridium acetobutylicum ATCC 824. *FEMS Microbiol. Rev.* **17**, 287-297 (1995).

34. Vasconcelos, I., Girbal, L. & Soucaille, P. Regulation of carbon and electron flow in Clostridium acetobutylicum grown in chemostat culture at neutral pH on mixtures of glucose and glycerol. *J. Bacteriol.* **176**, 1443-1450 (1994).

35. Sierra, J., Acousta, R., Montoya, D., Buitrago, G. & Silva, E. Isolation of spontaneous butanol-resistant mutants of Clostridium acetobutylicum. *Rev. Colomb. Cienc. Quim.-Farm.* **25**, 26-35 (1996).

36. Junelles, A.M., Janati-Idrissi, R., El Kanouni, H., Petitdemange, H. & Gay, R. Acetone-butanol fermentation by mutants selected for resistance to acetate and butyrate halogen analogues. *Biotechnol. Lett.* **9**, 175-178 (1987).

37. Clark, S.W., Bennett, G.N. & Rudolph, F.B. Isolation and characterization of mutants of *Clostridium acetobutylicum* ATCC 824 deficient in acetoacetyl-Coenzyme A:acetate/butyrate:Coenzyme A transferase (EC 2.8.3.9) and in other solvent pathway enzymes. *App. Environ. Microb.* **55**, 970-976 (1989).

38. Minton, N.P. et al. Vector systems for the genetic analysis of Clostridium acetobutylicum. *Clin. Mol. Aspects Anaerobes, Proc. Bienn. Anaerobe Discuss. Group Int. Symp., 6th*, 187-201 (1990).

39. Mermelstein, L. D. Development and use of tools for the genetic analysis and metabolic engineering of Clostridium acetobutylicum ATCC 824. Thesis,Northwestern Univ.,Evanston,IL,USA. 233 pp. (1992).

40. Minton, N.P., Swinfield, T.J., Brehm, J.K., Whelan, S.M. & Oultram, J.D. Vectors for use in Clostridium acetobutylicum. *Genet. Mol. Biol. Anaerobic Bact.*, 120-140 (1993).

41. Mermelstein, L.D., Welker, N.E., Petersen, D.J., Bennett, G.N. & Papoutsakis, E.T. Genetic and metabolic engineering of Clostridium acetobutylicum ATCC 824. *Ann. N. Y. Acad. Sci.* **721**, 54-68 (1994).

42. Wilkinson, S. R. and M. Young (1994). Targeted integration of genes into the Clostridium acetobutylicum chromosome. Microbiology 140. 89-95.

43. Green, E.M. et al. Genetic manipulation of acid formation pathways by gene inactivation in Clostridium acetobutylicum ATCC 824. *Microbiology* **142**, 2079-2086 (1996).

44. Nolling J, et al. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *J Bacteriol.* 2001 Aug;**183**(16):4823-38.

45. Ravagnani, A. et al. Spo0A directly controls the switch from acid to solvent production in solvent-forming clostridia [In Process Citation]. *Mol Microbiol* **37**, 1172-1185 (2000).

46. Harris, L.M., Desai, R.P., Welker, N.E. & Papoutsakis, E.T. Characterization of recombinant strains of the Clostridium acetobutylicum butyrate kinase inactivation mutant: need for new phenomenological models for solventogenesis and butanol inhibition? *Biotechnol. Bioeng.* **67**, 1-11 (2000).

47. Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13**, 660-665 (1996).

48. Fraser, C.M. et al. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397-403 (1995).

49. Grishin NV, Wolf YI, Koonin EV. From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res.* **10**, 991-1000 (2000).

50. Subramanian G, Koonin EV, Aravind L. Comparative genome analysis of the pathogenic spirochetes Borrelia burgdorferi and Treponema pallidum. *Infect Immun.* **68**, 1633-1648 (2000).

51. Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin. The COG database. a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33-36 (2000).

52. Huynen MA, Dandekar T, Bork P. Variation and evolution of the citric-acid cycle. a genomic perspective. *Trends Microbiol.* **7**, 281-291 (1999).

53. Stragier P, Losick R. Molecular genetics of sporulation in Bacillus subtilis. *Ann Rev Genet.* **30**, 297-341 (1996).

54. Leonard CJ, Aravind L, Koonin EV. Novel families of putative protein kinases in bacteria and archaea. evolution of the eukaryotic protein kinase superfamily. *Genome Res.* **8**, 1038-1047 (1998).

55. Ponting CP, Aravind L, Schultz J, Bork P, Koonin EV. Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J Mol Biol.* **289**, 729-745 (1999).

56. Galperin MY, Natale DA, Aravind L, Koonin EV. A specialized version of the HD hydrolase domain implicated in signal transduction. *J Mol Microbiol Biotechnol.* **1**, 303-305 (1999).

57. Mitchell WJ. Physiology of carbohydrate to solvent conversion by clostridia. *Adv Microb Physiol.* **39**, 31-130 (1998).

58. Shoham Y, Lamed R, Bayer EA. The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides. *Trends Microbiol.* **7**, 275-281 (1999).

59. Kruus, K., Lua, A.C., Demain, A.L. & Wu, J.H. The anchorage function of CipA (CelL), a scaffolding protein of the Clostridium thermocellum cellulosome. *Proc Natl Acad Sci U S A* **92**, 9254-9258 (1995).

60. Kakiuchi, M. et al. Cloning and DNA sequencing of the genes encoding Clostridium josui scaffolding protein CipA and cellulase CelD and identification of their gene products as major components of the cellulosome. *J Bacteriol* **180**, 4303-4308 (1998).

61. Bayer, E.A., Shimon, L.J., Shoham, Y. & Lamed, R. Cellulosomes-structure and ultrastructure. *J Struct Biol* **124**, 221-234 (1998).

62. Tamaru, Y., Karita, S., Ibrahim, A., Chan, H. & Doi, R.H. A large gene cluster for the clostridium cellulovorans cellulosome [In Process Citation]. *J Bacteriol* **182**, 5906-5910 (2000).

63. Lee, S.F., Forsberg, C.W. & Gibbins, L.N. Xylanolytic activity of Clostridium acetobutylicum. *Appl. Environ. Microbiol.* **50**, 1068-1076 (1985).

64. Steeves RM, Denton ME, Barnard FC, Henry A, Lambert JM. Identification of three oligosaccharide binding sites in ricin. *Biochemistry* **38**, 11677-11685 (1999).

65. Lee, S.F., Forsberg, C.W. & Rattray, J.B. Purification of characterization of two endoxylanases from Clostridium acetobutylicum ATCC 824. *Appl. Environ. Microbiol.* **53**, 644-650 (1987).

66. Saier, M.H., Jr. et al. Catabolite repression and inducer control in Gram-positive bacteria. *Microbiology* **142**, 217-230 (1996).

67. Reizer, J. et al. Novel phosphotransferase system genes revealed by genome analysis - the complete complement of PTS proteins encoded within the genome of Bacillus subtilis. *Microbiology* **145**, 3419-3429 (1999).

68. Fiegler, H., Bassias, J., Jankovic, I. & Bruckner, R. Identification of a gene in Staphylococcus xylosus encoding a novel glucose uptake protein. *J Bacteriol* **181**, 4929-4936 (1999).

69. Kelly G, Prasannan S, Daniell S, Fleming K, Frankel G, Dougan G, Connerton I, Matthews S. Structure of the cell-adhesion fragment of intimin from enteropathogenic Escherichia coli. *Nat Struct Biol.* **6**, 313-318 (1999).

70. Marino M, Braun L, Cossart P, Ghosh P. Structure of the lnlB leucine-rich repeats, a domain that triggers host cell invasion by the bacterial pathogen L. onocytogenes. *Mol Cell.* **4**, 1063-1072 (1999).

71. Fleischmann, R.D. et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512 (1995).

72. Smith, D.R. et al. Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. *Journal of Bacteriology* **179**, 7135-7155 (1997).

73. Ewing, B. & Green, P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res* **8**, 186-194 (1998).

74. Green, P., University of Washington. http.//bozeman.mbt.washington.edu/phrap.docs/phrap.html

75. Shine, J. & Dalgarno, L. Terminal-sequence analysis of bacterial ribosomal RNA. Correlation between the 3'-terminal-polypyrimidine sequence of 16-S RNA and translational specificity of the ribosome. *Eur J Biochem* **57**, 221-230 (1975).

76. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

77. Altschul, S.F. & Koonin, E.V. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem Sci* **23**, 444-447 (1998).

78. Schaffer, A.A. et al. IMPALA. matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* **15**, 1000-1011 (1999).

79. Schultz, J., Milpetz, F., Bork, P. & Ponting, C.P. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* **95**, 5857-5864 (1998).

80. Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P. & Bork, P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res* **28**, 231-234 (2000).

81. Wootton, J.C. Non-globular domains in protein sequences. automated segmentation using complexity measures. *Comput Chem* **18**, 269-285 (1994).

82. Walker, D.R. & Koonin, in Int. Conf. Intell. Syst. Mol. Biol., 5th. (ed. T. Gaasterland) 333-339 (AAAI Press, Menlo Park, Calif, 1997).

83. Kanehisa, M., & Goto, S. KEGG; Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28,** 27-30 (2000).

84. Overbeek, R. et al. WIT; integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res* **28**, 123-125 (2000).

85. Higgins, D.G. Thompson, J.D. Gibson, T.J. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* **266**, 383-402 (1996).

86. Felsenstein, J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol* **266**, 418-427 (1996).