Journal of Molecular Microbiology and Biotechnology

J Mol Microbiol Biotechnol 2009;16:81–90
DOI: 10.1159/000142896

Published online: October 29, 2008

# Rapid Access to Genes of Biotechnologically Useful Enzymes by Partial Genome Sequencing: The Thermoalkaliphile *Anaerobranca gottschalkii*

G. Antranikian[a] A. Ruepp[b, c] P.M.K. Gordon[g] M. Ballschmiter[d] A. Zibat[b] M. Stark[b] C.W. Sensen[g] D. Frishman[e] W. Liebl[d] H.-P. Klenk[b, f]

[a]Institute of Technical Microbiology, Hamburg University of Technology, Hamburg, [b]Formerly Epidauros Biotechnologie AG, Bernried, [c]Institute for Bioinformatics, GSF – Forschungszentrum für Umwelt und Gesundheit, Neuherberg, [d]Georg August University Göttingen, Institute for Microbiology and Genetics, Göttingen, [e]Technical University Munich, Genome Oriented Bioinformatics, Freising and [f]*e.*gene Biotechnologie GmbH, Feldafing, Germany; [g]University of Calgary, Faculty of Medicine, Department of Biochemistry and Molecular Biology, Calgary, Alta., Canada

**Key Words**

Rapid access to genes · Biotechnologically useful enzymes · Partial genome sequencing · *Anaerobranca gottschalkii* · Amylolytic enzymes

**Abstract**

*Anaerobranca gottschalkii* strain LBS3$^T$ is an extremophile living at high temperature (up to 65°C) and in alkaline environments (up to pH 10.5). An assembly of 696 DNA contigs representing about 96% of the 2.26-Mbp genome of *A. gottschalkii* has been generated with a low-sequence-coverage shotgun-sequencing strategy. The chosen sequencing strategy provided rapid and economical access to genes encoding key enzymes of the mono- and polysaccharide metabolism, without dilution of spare resources for extensive sequencing of genes lacking potential economical value. Five of these amylolytic enzymes of considerable commercial interest for biotechnological applications have been expressed and characterized in more detail after identification of their genes in the partial genome sequence: type I pullulanase, cyclodextrin glycosyltransferase (CGTase), two α-amylases (AmyA and AmyB), and an α-1,4-glucan-branching enzyme.



## Introduction

Some parts of Earth's biosphere are 'extreme', meaning uninhabitable for most 'regular' organisms. Yet, comparatively little is known about the physiology and genomics of the organisms naturally occurring in these environments, although these organisms (extremophiles) have a great potential as source for novel enzymes with unique biotechnological features. *Anaerobranca gottschalkii* strain LBS3$^T$ (DSM 13577), the thermoalkaliphilic bacterium of the *Clostridium/Bacillus* subphylum whose genome we describe here, was first isolated from a hot lake inlet at Lake Bogoriae, Kenya [Prowe and Antranikian, 2001]. Its name refers to the obligatory anaerobic *(anaero)* lifestyle and the branched *(branca)* cell shape of the organism, and to Dr. Gerhard Gottschalk in honor of his pioneer work in physiology and metabolism of anaerobes. *A. gottschalkii* grows optimally at 50–55°C and pH 9.5 and it is unique within the thermoalkaliphiles in its ability to grow heterotrophically on a variety of mono- and polysaccharides as energy and C-sources, as well as on proteinaceous substrates [Prowe and Antranikian, 2001]. The genes coding for *A. gottschalkii*'s amylolytic

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com


1464–1801/09/0162–0081$26.00/0

Accessible online at:
www.karger.com/mmb

G. Antranikian
Institute of Technical Microbiology, Hamburg University of Technology
Kasernenstrasse 12, DE–21073 Hamburg (Germany)
Tel. +49 531 2616 227, Fax +49 721 151 591 214
E-Mail antranikian@tuhh.de or hpk@dsmz.de

**Table 1.** Target genes or carbohydrate and peptide metabolism

| ORF number | Predicted function | Status |
|---|---|---|
| ago_00257/00490 | extracellular serine protease | sequence gap closed |
| ago_00315/01041 | prolyl endopeptidase | sequence gap closed |
| ago_00361 | similar to processing protease | complete |
| ago_00422 | ATP-dependent protease LA | sequence gap closed |
| ago_00440 | anositol-monophosphate dehydrogenase | complete |
| ago_00490 | extracellular serine protease | complete |
| ago_00634 | β-galactosidase | 3′-end missing |
| ago_00765 | lon protease | complete |
| ago_00808 | amidase | sequence gap closed |
| ago_00870 | spore protease | complete |
| ago_00942 | lysophospholipase | complete |
| ago_00958 | putative polysaccharide hydrolase | sequence gap closed |
| ago_01041 | prolyl endopeptidase | complete |
| ago_01042 | xaa-pro dipeptidase | sequence gap closed |
| ago_01036 | dimethyladenosine transferase | sequence gap closed |
| ago_01122 | periplasmic serine protease | sequence gap closed |
| ago_01295 | 1,4-α-glucan branching enzyme | complete |
| ago_01391 | amylopullulanase | complete |
| ago_01452 | similar to glycosyl hydrolase | complete |
| ago_01459 | amidase | complete |
| ago_01560 | sialoglycoprotease | complete |
| ago_01618 | ATP-dependent CLp protease proteolytic SU | complete |
| ago_01619 | ATP-dependent Clp protease | complete |
| ago_1620/00765 | ion-like ATP-dependent protease | sequence gap closed |
| ago_01698 | endoglucanase | complete |
| ago_01716 | putative protease | complete |
| ago_01740 | similar to endo-1,4-β-xylanase | complete |
| ago_01792 | α-amylase | sequence gap closed |
| ago_01851 | putative protease | sequence gap closed |
| ago_02265 | amidase enhancer | completed |
| ago_02288 | putative protease | sequence gap closed |
| ago_02425 | glucan 1,4-α-maltohydrolase | complete |
| ago_02489 | carboxyl-terminal protease | complete |
| ago_02679 | pullulanase | complete |
| ago_02739 | *N*-acetylmuromoyl-L-alanine amidase | sequence gap closed |
| Contig 818 ORF | cyclodextrin glycosyltransferase, CGTase | sequence gap closed |

ORF names from annotation based on the 14th assembly.

enzymes are promising candidates for biotechnological applications which require double-extreme enzymatic activity at high temperatures and high pH values. Putative industrial applications for such enzymes are: food processing, enzymatic liquefication of starch, and environmentally friendly production of cyclodextrins. Furthermore, the partial genome sequence of *A. gottschalkii* will also contribute to our general understanding of the mechanisms that allow microorganisms to grow at high pH and temperature.

Within the last decade, since the publication of the first microbial genome sequence [Fleischmann et al., 1995], the technology of whole-genome sequencing has significantly improved. Microbial genomics evolved within a relatively short time from an art that could be performed by only a few well-financed elite research institutions to a standard procedure that can be applied by many laboratories all over the world. However, compared to most other molecular techniques, the complete analysis of a microbial genome is still a relatively expensive and complex project that requires many resources for se-

**Table 2.** General genome features

| | |
|---|---|
| Estimated genome size, bp [100%] | 2,262,800 |
| Assembled unique sequence, bp [96%] | 2,173,793 |
| Annotated contigs, bp [85%] | 1,921,827 |
| G+C content, % | 30 |
| Predicted protein coding sequences | 1,907 |
| With homologues in databases | 1,598 |
| Belonging to at least one COG | 1,335 |
| Belonging to superfamilies | 1,168 |
| With homologous 3D structures known | 647 |
| Containing EC numbers | 383 |
| Containing ≥2 trans-membrane regions | 344 |
| Containing coiled-coil regions | 21 |

quencing, gap closure, sequence polishing, ORF prediction and annotation. The aim of our project was to achieve fast, fairly complete, and economical access to a suite of genes that code for amylolytic enzymes of biotechnological interest. We were not interested to determine the exact sequence the 400th version of any bacterial gene that might satisfy only academic interests, but may not serve commercial interests. With meanwhile (December 20, 2006) 399 published bacterial genomes in the public databases [GOLD, www.genomesonline.org, Liolios et al., 2006], and a further 997 ongoing bacterial genome-sequencing projects, we focused our project on the economical exploitation of those parts of the *A. gottschalkii* genome that serve our biotechnological targets, while providing only a rough overview of the remaining parts of the genome as basis for follow-up academic studies.

## Results and Discussion

*Partial Genome Sequence*

The 16S rRNA gene [Prowe and Antranikian, 2001] was the only tiny fraction of genetic information published from *A. gottschalkii* prior to the beginning of the here described low-path shotgun-sequencing approach. In order to gain fast insight into the genomic content of the *A. gottschalkii* genome, we chose a 'classical' shotgun approach [Fleischmann et al., 1995] by end-sequencing of randomly generated cloned chromosomal DNA fragments. The initial assembly of slightly less than 12,000 DNA sequences generated from our plasmid libraries resulted in 739 DNA contigs with a 3.6-fold average sequence coverage of more than two megabases representing about 95% of the genome. These DNA contigs were directly subjected into the MAGPIE annotation system [Gaasterland and Sensen, 1996] without any prior sequence editing. The MAGPIE system automatically performed all procedures required for ORF prediction and comparison of gene sequences as well as predicted amino acid sequences of the encoded proteins, with public domain databases. The functional predictions generated by MAGPIE were screened for enzymes involved in carbohydrate and peptide metabolism and compared to a list of about 60 enzymes that, if present in the genome, would be of interest for biotechnological applications at high temperatures and high pH values. As expected, not all genes of interest were completely encoded within these contigs, some were incomplete and located at the ends of one or two contigs. Almost 40 of these genes of interest were completed by primer walking on plasmid templates bridging the aforementioned sequencing gaps. Table 1 provides an overview of the identified genes coding for proteins of the carbohydrate and peptide metabolism and their sequencing status after two rounds of gap closure by primer walking on templates from our plasmid libraries. Several of the successfully completed genes from the central carbohydrate metabolism were subject of further functional characterization, as described below. Despite intensive search we could not detect genes encoding nitrilases, β-(1,4)-glucan glucanohydrolase, endo-1,4-β-D-glucanase, α-galactosidase, β-mannanase, levansucrase, glucoamylase, glucosyltransferase-I, chitinase, cycloisomaltooligosaccharide glucotransferase, endo-1,4-β-xylanase and trehalase.

Re-assembly of the genome after gap closure for selected genes in carbohydrate and peptide metabolism resulted in 696 contigs with a total of 2,173,793 bp of unique chromosomal sequence with an average G+C content of 30.0% (table 2) and an average contig length of 3,123 bp. Most of these contigs (80%) were linked by templates with end sequences found in two adjacent contigs (555 sequencing gaps), with 141 remaining physical gaps (unlinked contigs). The largest contig had a length of 31,430 bp generated from 421 assembled sequences, with a total of 24 further contigs being longer than 10 kbp. From this assembly we estimate the total size of the *A. gottschalkii* strain LBS3$^T$ genome to be about 2.26 Mbp, without any known extrachromosomal elements (plasmids).

*Genome Annotation*

A second round of automated annotation, followed by subsequent expert annotation, was performed on a set of 510 long, high-sequence-quality contigs after visual sequence editing. The average length of these contigs was

3,768 bp, representing altogether more than 88% of the unique sequence in all our contigs. An overview of the genome features derived from the annotation of these contigs in PEDANT [Frishman et al., 2001] is provided in table 2. Unrestricted view-only-access to the latest annotation results is provided through the world wide web at http://pedant.gsf.de/cgi-bin/wwwfly.pl?Set=Agottschalkii&Page=index. The remaining set of 171 smaller contigs or contigs with lower sequence quality, respectively, were excluded from further annotation because predictions derived from these contigs would have lowered the overall quality of the data. The excluded contigs had an average length of only 1,473 bp, comprising a total of 252 kbp or less than 12% of the unique sequence.

All observations based on the annotated genome fragments can provide only a first insight into the genome content of *A. gottschalkii*. However, the facts and results reported in the following paragraphs might be of use to those who are interested on the non-biotechnological features of *A. gottschalkii*. The sequences of the contigs which we describe here might be helpful for rapid access to the genes located on them.

*Structural Components and Comparative Genomics*

The partial fragmented genome contains five copies of the 16S rRNA gene, with two of the five genes being completely located within large contigs (0652 and 0681) and five long fragments (three 3′- and two 5′-ends) being located at the ends of other contigs. With the exception of C-1421 instead of A-1421, all five copies of the gene are identical to the sequence published by Prowe and Antranikian [2001], a difference that is probably due to the PCR-mediated amplification used in the initial study. With 35 ORFs encoding predicted transposases, the genome is relatively rich in IS elements. Most of the predicted transposases are related to sequences clustered in COG0675 and frequently located in the genomes of *Thermoanaerobacter tengcongensis* and *Clostridium perfringens,* e.g. IS1136. The numerous copies of IS elements indicate that *A. gottschalkii* was repeatedly an acceptor for external genetic material. Identification of three genes associated with pilus formation further indicate the organisms' ability to exchange genetic material with other bacteria. Contig 603 contains a 5,412-bp-long stretch of clustered regularly interspaced short palindromic repeats, CRISPRs, as frequently found in bacterial genomes [Haft et al., 2006] and supposed to be involved in the organism's defense against foreign DNA (e.g. phages). The 82 G+C-rich (69%) repeats are 26–28 bp long and are separated by spacers of 36–43 bp length.

*A. gottschalkii*'s 16S rRNA shows the highest degree of sequence similarity with the respective genes in *Clostridium novi* and *Desulfitobacterium hafniense* (82% each, considering only bacteria with completely sequenced genomes). Comparison of a selection of conserved proteins usually used for phylogenetic analyses with the proteomes predicted from completely sequenced bacteria confirmed the placement of *A. gottschalkii* within the *Clostridia*, close to *C. thermocellum*, *Moorella thermoacetica* and *Thermoanaerobacter tengcongensis,* another thermophile growing at high pH conditions [Bao et al., 2002]. The coding density of 0.99 genes/kbp within the 510 annotated contigs corresponds very well to the 1 gene/kbp average known from other, completely sequenced, bacterial genomes.

The identification of *selA* (ctg. 0612), *selB* (also ctg. 0612) and *selD* (ctg. 0276) indicate the organisms' ability to incorporate the 21st amino acid, selenocysteine, into its proteins. Predicted ORFs that include an UGA codon have not yet been identified or extensively searched for in the partial genome. The use of selenocysteine by *A. gottschalkii* is further supported by the identification of a putative selenocysteine lyase (on contig 0464), an enzyme that exclusively decomposes L-selenocysteine into L-alanine and $H_2Se$. Some of the secreted exoenzymes of *A. gottschalkii* contain the TAT-secretion system recognition signal. However, none of the genes coding for the components of the TAT system could be identified, despite an intensive search. Subunits for the SEC preprotein translocase were identified on contigs 0650 (secY) and 0663 (secA).

The elevated temperatures in *A. gottschalkii*'s regular habitat might already exert constraints on the stability of its mRNAs. Lao and Forsdyke [2000] reported greater Chargaff differences (as a measure of purine loading in mRNAs) in thermophiles than in mesophiles. The purine-loading index (PLI) of *A. gottschalkii* (0.562) is even in the range of the PLIs determined for true hyperthermophiles (0.552 ± 0.015) and well above the PLIs determined for a selection of meso- and psychrophiles (0.504 ± 0.008), as well as *Thermus thermophilus,* PLI 0.506 ($T_{opt}$ 75°C). The pressure on the purine load affects the codon choice in thermophiles, indicating that some features of their amino acid composition, e.g. a high level of glutamic acid (Glu), might reflect a purine-loading pressure resulting from constraints on mRNA [Lao and Forsdyke, 2000]. This hypothesis finds support by 7.4% Glu codons in the genome of *A. gottschalkii* (hyperthermophiles 7.7% ± 1.0), well above the respective values inferred from genomes of meso- and psychrophiles (5.6%

± 0.1). The surfaces of proteins from hyperthermophiles contain a strongly increased fraction of charged residues (Glu, arginine [Arg] and lysine [Lys]) at the expense of non-charged polar residues, mainly glutamine (Gln) [Cambillau and Claverie, 2000]. This imbalance can already be observed at the genomic level where Glu, Arg and Lys make up for 20.9% of the codons in *A. gottschalkii* (hyperthermophiles 20.2 ± 1.4%) as compared to an average of 15.9 ± 0.2% in mesophiles. With slightly more than 3% Gln codons in its genome (hyperthermophiles 1.9 ± 0.2%), the proteome of *A. gottschalkii* contains, however, less of the non-charged Gln than average meso- and psychrophiles (3.7 ± 0.8%). It might be noteworthy that the *A. gottschalkii* genome is unusually poor in tryptophan codons (0.78%) and alanine codons (5.65%), but rich in asparagine codons (5.1%) when compared to a selection of other whole-genome sequences from organisms with an optimal growth temperature range between 10 and 100°C.

*Metabolism and Motility*

A total of 39 genes coding for proteases and peptidases, including two collagenases, as well as 20 uptake systems for amino acids and peptides (19 ABC transporters, one symporter) were identified, confirming the organisms' appetite for proteinaceous substrates such as yeast extract, peptone and tryptone [Prowe and Antranikian, 2001]. All three genes coding for enzymes of the arginine deiminase pathway (arginine deiminase, ornithine transcarbamoylase, and carbamate kinase) were identified next to each other, as it is known from a wide range of organisms. *A. gottschalkii* might also be able to use nucleic acids imported from its environment. We have identified the genes for a DNA uptake system, as well as those coding for eight exonucleases, seven endonuclease and two restriction/modification systems. With only one phospholipase gene identified in the genome, lipids apparently do not play a major role in *A. gottschalkii*'s food supply. From the large variety for genes that enable *A. gottschalkii* to make use of carbohydrates (for details see below), proteins and nucleic acids out of its hot and alkaline environment, it might be suggested that the organism plays an ecological role as a scavenger in its natural habitat.

It is well known that alkaliphiles require sodium ions for their growth, and as Prowe and Antranikian [2001] demonstrated, potassium ions cannot replace the sodium ions in the growth medium of *A. gottschalkii*. Sodium/potassium-ATPases play an important role in the energy metabolism and intracellular pH regulation in *A. gottschalkii* [Prowe et al., 1996]. The identification of seven genes for sodium/potassium-ATPases in the partial genome sequence supports that observation. There are indications for up to ten iron ABC transporters and at least two complete (but possibly up to five) phosphate ABC transporters in the genome.

The identification of 43 putative genes related to spore formation as well as 14 putative genes possibly involved in germination, indicates that at least many of the genes used in other organisms for spore formation are also part of *A. gottschalkii*'s genetic makeup. However, there is no experimental evidence known from the literature that *A. gottschalkii* is able to form spores. Motility of *A. gottschalkii* cell has originally been described [Prowe and Antranikian, 2001]. The identification of 29 genes coding for response regulators indicate the organisms' ability to detect attractants and repellents in its environment and to move accordingly. Genes for some signaling transducers were identified (CheC, CheD, CheR, and CheW), whereas others are missing (CheA, CheB, CheY, and CheZ), possibly due to still missing parts of the genome sequence. 22 genes encoding components of the flagellar assembly have been clearly identified on the analyzed contigs.

No genes involved in dissimilatory sulfate reduction could be identified in the genome of *A. gottschalkii*, confirming the already described inability of the organism to perform this process [Prowe and Antranikian, 2001], just like its slightly less alkaliphilic close relative *Anaerobranca horikoshii*. The absence of any genes for the pyruvate dehydrogenase complex indicates that the organism is using pyruvate ferredoxin oxidoreductase instead, whose genes are located on contigs 0123 and 0319 (α-subunit), and 0533 (δ- and γ-subunits). No genes have been detected thus far for catalase and superoxide dismutase or for any cold shock proteins.

*Overview on the Central Carbon Carbohydrate Metabolism*

A total of about 93 genes involved in sugar catabolism are predicted from the partial genome sequence. For the initial breakdown of hexoses and the central carbohydrate metabolism of *A. gottschalkii* it is of interest to know that the Embden-Meyerhof-Parnas pathway seems to be used for the utilization of starch or glucose, leading mainly to the production of acetate as fermentation end product [Prowe and Antranikian, 2001]. The obvious lack of any genes coding for alcohol dehydrogenases explains that only traces of ethanol were detected by Prowe and Antranikian [2001] as fermentation end product. No

**Table 3.** Overview over the recombinant enzymes in *A. gottschalkii*

| | Signal sequence | MW kDa | pH optimum | Temperature optimum, °C | Specific activity, U · mg$^{-1}$ | Reference |
|---|---|---|---|---|---|---|
| α-Amylase (AmyB) | no | 52 | 6.0 | 55 | 220[a] | Ballschmiter et al., 2005 |
| Branching enzyme (BE) | no | 72 | 7.0 | 50 | 291[b] | Thiemann et al., 2006 |
| α-Amylase (AmyA) | yes (lipoprotein) | 59 | 8.0 | 70 | 52[a] | Ballschmiter et al., 2005 |
| CGTase | yes | 78 | 6.0–9.0 | 65 | 308[c] | Thiemann et al., 2004 |
| Type I pullulanase | yes (lipoprotein) | 96 | 8.0 | 70 | 56[d] | Bertoldo et al., 2004 |

[a] One unit of α-amylase liberates 1 μmol · min$^{-1}$ reducing ends from starch.
[b] One unit of BE activity is defined as the decrease in absorbance of a glucan-iodine complex of 1 at 660 nm · min$^{-1}$ with amylose as substrate.
[c] One unit of CGTase catalyzes the formation of 1 μmol · min$^{-1}$ of β-cyclodextrin from starch.
[d] One unit of pullulanse liberates 1 μmol · min$^{-1}$ reducing ends from pullulan.

gene coding for pyruvate kinase could yet be identified in the known fraction of the genome; however, there is still a small probability for a pyruvate kinase gene in the yet unknown fraction of the genome. Most of the genes for the tricarboxylic acid cycle were identified, except for homologues for malate dehydrogenase and fumarase. Thus, as in most anaerobic, fermentative bacteria, *A. gottschalkii* probably operates an incomplete, reductive tricarboxylic acid cycle mainly used for anabolic purposes.

Previous physiological experiments supplemented with biochemical data obtained with authentic or recombinant enzymes have revealed that *A. gottschalkii* can utilize a broad variety of carbohydrate substrates and is capable of producing numerous polysaccharide depolymerases. The partial genome sequence, as presented here, nicely mirrors the physiological and biochemical situation by revealing genes (numbers in parentheses) for the breakdown of α-glucans (7), β-glucans (8), hemicelluloses (4), and agar (1), most of them predicted to be extracellular in accordance with their physiological role. In addition, genes coding for five ABC transporters, two permeases and 18 phosphotransferase systems were identified, which may be involved in the uptake of various oligo- and monosaccharides. In the following we will focus in more detail only on one group of depolymerases, whose genes were identified in the partial genome sequence and subsequently analyzed in more detail, the α-glucan degrading enzymes, for which a rather comprehensive picture is now available on the genetic and on the enzymatic level.

### Anaerobranca *as a Source of Amylolytic Enzymes*

Our sequencing effort has led to the identification and characterization of five amylolytic enzymes that are of considerable commercial interest for biotechnological applications: a type I pullulanase (gb: AY541591.1), a cyclodextrin glycosyltransferase (CGTase, emb: AJ850087.1), two α-amylases (gb: AY842299.1, gb: AY842298.1) and an α-1,4-glucan-branching enzyme (BE, emb: AM114416.1). For an overview, see table 3. All of them belong to the glycoside hydrolase family 13 (GH13), an enzyme family based on amino acid sequence similarity, which is commonly known as the α-amylase family, but in fact covers various amylolytic enzymes. Although the family harbors diverse enzyme activities and substrate specificities, the primary structures of all GH13 enzymes share four typical conserved regions, a $(\beta/\alpha)_8$-barrel fold as the characteristic super-secondary structure motif, and they act on glycosidic bonds with a retaining reaction mechanism. As this family is by far the largest family of glycoside hydrolases, recently a division into monofunctional subfamilies was suggested.

α-Amylases (EC 3.2.1.1) and pullulanases (EC 3.2.1.41) are hydrolyzing enzymes that differ in the choice of their target within a polysaccharide. While α-amylases such as the two *A. gottschalkii* amylases attack α-1,4-glycosidic bonds in glucose polymers with an endocleaving mechanism, i.e. in a random fashion within the α-glucan molecule, pullulanases attack α-1,6-glycosidic bonds within amylopectin or pullulan molecules. In the case of type I pullulanases like the *A. gottschalkii* pullulanase, the products liberated from pullulan and amylopectin are maltotriose and linear oligosaccharides, respectively.

CGTases (EC 2.4.1.19) and BEs (*glgB*, EC 2.4.1.18) have a transferase activity. Here, upon cleavage of a glycosidic bond, a new glycosidic linkage is introduced. In case of the CGTase after the cleavage of an α-1,4-glycosidic bond, the reducing end of the oligosaccharide product is fused to its non-reducing end by re-formation of an α-1,4-glycosidic bond, resulting in a circular dextrin of normally 6, 7 or 8 glucose residues (α-, β- and γ-cyclodextrin). The BE upon cleavage attaches an oligosaccharide moiety to a glucan via formation of an α-1,6 bond, thus introducing branches into the otherwise linear α-glucan.

Sequence analysis of the corresponding genes revealed that one of the α-amylases (AmyA), the CGTase, and the pullulanase have N-terminal signal peptides and therefore are presumably extracellular enzymes [Ballschmiter et al., 2005; Bertoldo et al., 2004]. It can be assumed that all three of them have their physiological role in the breakdown of α-glucan polysaccharides which are utilized by *A. gottschalkii* as carbon source. AmyA and pullulanase show typical lipoprotein signal peptides. Therefore, it can be assumed that these enzymes are attached to the cell surface in the authentic host. Dextrins that are formed outside of the cell – mainly glucose and maltose by AmyA, malto-oligosaccharides by pullulanase, and cyclodextrins by CGTase – are subsequently transported into the cytoplasm, where they are further broken down and fed into the central carbohydrate metabolism.

CGTase was purified both as a recombinant protein expressed in *E. coli* and as a native protein from *A. gottschalkii* culture supernatant [Thiemann et al., 2004]. Both enzyme forms showed comparable temperature and pH profiles, thus most properties were investigated with the recombinant enzyme. $Ca^{2+}$ is bound to the enzyme with high affinity and has a stabilizing effect on the protein. The enzyme preferentially synthesized α-cyclodextrins and to a lesser amount β-cyclodextrins. The ratio of α- to β-cyclodextrins varied depending on the substrate composition. An increase in the substrate concentration and in the degree of α-1,6-glycosidic bonds shifted the ratio in favor of β-cyclodextrins. The *A. gottschalkii* CGTase was also found to be able to form cyclic glucans with more than eight glucose units. Analysis of the partial genome of *A. gottschalkii* revealed that the organism has the genetic outfit to produce the enzymes necessary to transport cyclodextrins formed by the CGTase into the cell and to subsequently hydrolyze them in the cytoplasm.

The extracellular α-amylase AmyA was heterologously overexpressed in *E. coli* without the signal peptide. Besides the typical hydrolytic activity of an α-amylase, the enzyme displayed high transglycosylation activity on malto-oligosaccharides at high substrate concentrations (25 mM). The combination of acting as a hydrolytic enzyme as well as – although often merely with low activity – as an enzyme with 4-α-glucanotransferase activity is not unusual for α-amylases, and vice versa typical 4-α-glucanotransferases (amylomaltases) mostly also display significant hydrolytic activity. Strikingly, *A. gottschalkii*'s AmyA also revealed a low but significant β-cyclodextrin glycosyltransferase activity, which is an unprecedented property for typical α-amylases. However, the physiological role of this cyclodextrin-forming activity is questionable, since the *A. gottschalkii* CGTase has a much higher activity.

Pullulanase was purified as the native enzyme from the supernatant of an *A. gottschalkii* culture as well as heterologously expressed in *E. coli*. Interestingly, two forms of the recombinant protein were observed. While the expression of the pullulanase without the lipoprotein signal peptide led to a 96-kDa protein, the expression of the full-length protein including the signal sequence led to a 70-kDa truncated protein, which was probably translated from an alternative start codon within the pullulanase ORF. Both enzyme forms were catalytically active. However, from the *A. gottschalkii* culture supernatant, only the 96-kDa full-length enzyme could be purified, whose pH and temperature profile were comparable to the recombinant full-length enzyme.

The three extracellular amylolytic enzymes, α-amylase AmyA, pullulanase and CGTase, are subjected to the harsh regime of the living conditions of *A. gottschalkii*. Consequently, all three enzymes can tolerate the alkaline pH and are thermostable. The pullulanase was found to have a half-life of 22 h at 70°C and for AmyA even a half-life of 48 h at 70°C was determined. All three enzymes are able to retain most of their activity up to pH 9.5. Yet, the optimal performance of the recombinant pullulanase, CGTase, and the extracellular α-amylase fall into the rather small window of pH 7–8 and 65–70°C. The recombinant pullulanase had a specific activity of 56 U/mg with pullulan as the substrate, while the recombinant CGTase formed β-cyclodextrins with a specific activity of 308 U/mg. The extracellular amylase (recombinant enzyme) cleaves starch with a specific activity of 52 U/mg.

The BE and the α-amylase AmyB are putatively located in the cytoplasm. BE is the key enzyme for the

formation of glycogen. Therefore, its presence in *A. gottschalkii* strongly suggests that the organism is able to synthesize branched storage glucans, and both the BE and AmyB may be involved in the metabolism of such a storage compound. This hypothesis is also supported by the identification of a gene coding for carbon storage regulator on contig 0641. Maltodextrins were found to be substrates for AmyB. Therefore, another likely function for AmyB is the breakdown of maltodextrins transported into the cell after the initial extracellular breakdown of starch. As with the extracellular enzymes discussed above, the range of maximal activity of the BE and the intracellular α-amylase were almost identical, here at pH 6–7 and 50–55°C. The recombinant α-amylase was overexpressed in *E. coli*. It cleaves starch with a specific activity of 220 U/mg. In contrast to the extracellular enzyme, no side activities like transglycoslyation were observed. The recombinant BE was expressed as secreted enzyme in *Staphylococcus carnosus*. A TAT signal sequence (trimethyl-amine *N*-oxide reductase gene from *E. coli*) was fused to the BE gene and the secreted protein was purified from the supernatant. The enzyme was able to transfer glucan chains with a DP4 to DP24 with a specific activity of 291 U/mg, while the minimal chain length of the donor required for cleavage was DP16. Cyclodextrins were neither hydrolyzed as oligosaccharides were, nor were they modified by transfer activity.

It was particularly rewarding to compare the biochemical properties of the α-amylases AmyA and AmyB of *A. gottschalkii*. These enzymes are closely related with each other at the primary structure level, but according to their primary sequences were proposed to be localized in different cellular microenvironments, i.e. in the extracellular milieu and the cytoplasm, respectively. Similar cases of enzymes from the same family putatively localized differently can also be found in other organisms, but generally such enzymes have not previously been directly compared with each other to find out if there are specific adaptations related to their subcellular localization. For AmyA and AmyB from *A. gottschalkii*, a remarkable evolutionary adaptation of physicochemical properties (pH dependence, stability) to their different cellular localizations was observed, while their biocatalytic properties (substrate cleavage spectrum, products, endomechanism) turned out to be very similar.

The pH profile of the extracellular amylase AmyA mirrors the living conditions of *A. gottschalkii*: while the enzyme is active from pH 6.0 to pH 9.5, *A. gottschalkii* grows from pH 6.0 to pH 10.5. The intracellular amylase AmyB has a much narrower pH range with an optimum at pH 6.0. In the cytoplasm, changes of pH are small due to cell homeostasis. There is no need for the intracellular enzyme to be as tolerant to drastic pH changes as its extracellular counterpart, which is directly and immediately exposed to any change in the environment of *A. gottschalkii*.

A similar adaptation can be observed in the temperature stabilities of AmyA and AmyB. While in vitro AmyA had a half-life of 48 h at 70°C, AmyB only had a half-life of about 10 min at this temperature, which is only 5°C above the growth range of *A. gottschalkii*. In vivo AmyB might be stabilized by cytoplasmic compounds, while again AmyA is subjected to the environmental conditions without a potentially protective cytoplasm.

## Conclusion

The wealth of information that has been deduced from the partial genome sequence of *A. gottschalkii* justified the low-sequence-coverage approach that we applied for our project. With only 3.6-fold sequence coverage over approximately 96% of the genome, we saved about two-thirds of the costs normally spent for the shotgun-sequencing phase in projects aimed for complete, high-quality genome sequences (8- to 12-fold sequence coverage). By closing only 43 sequencing gaps for genes of interest within the scope of our project, we saved a comparable fraction (two-thirds) of the costs usually required in gap closure of genomes with 8- to 12-fold sequence coverage. By extensively polishing of only those regions of the genome sequence which contain genes of interest for the scope of our project, we saved an even higher fraction (about 90%) of the respective costs in full (academic) genome projects. Nevertheless, with two rounds of automated annotation (with MAGPIE and PEDANT) and by intensive search for missing genes we can be confident that we achieved the identification of about 95% of the genes that are (i) encoded in the genome *of A. gottschalkii* and (ii) that are part of our initial list of biotechnologically relevant enzymes to look for. Decoding of the remaining 5% of the genome sequence and improvement of the sequence quality in regions of no interest for a commercially focused project would easily triple the overall cost for the project for sequencing, gap closure, polishing and annotation. This implies that the costs per newly gained information are about 35–40 times as high for the last 5% of the genome (two-thirds

of the costs), which we excluded for economical reasons from our analysis, as compared to the first 95% of the genome.

## Experimental Procedures

*Sequencing and Annotation*

The DNA for genome sequencing of *A. gottschalkii* strain LBS3$^T$ (DSM 13577; NCBI Taxonomy ID 108328) was prepared as previously described by Prowe and Antranikian [2001]. Genomic DNA was sequenced using a conventional whole-genome shotgun strategy. Fragments of 1–2.5 and 2.5–6 kbp, respectively, were isolated after mechanical shearing, end-repaired, and cloned via A-T cloning in pGEM-T Easy Vector (Promega). Templates for sequencing were isolated on BioRobots 9600 (Qiagen). Sequence reactions were generated from the plasmid templates using primers pGEM1-abi (forward) and pGEM2-abi (reverse) and either ABI BigDye™ Terminator chemistry (Applied Biosystems) or ET-Terminator chemistry (Amersham Biosciences). Sequence reactions were analyzed on automated sequencers ABI PRISM® models 377-96, 3100 and 3700, providing an average full read length of 859 nt. Sequence traces were first processed with Phred for base calling [Ewing and Green, 1998] and data quality assessment [Ewing at al., 1998], then assembled with Phrap [P. Green, University of Washington], and visualized and edited with Consed [Gordon et al., 1998]. The partial genome was assembled from 12,246 sequence reads with an average trimmed read length of 634 nt, resulting in 3.7-fold average sequence coverage. For gap closure and sequence editing, 43 primer walking reactions were performed on plasmid clones. Several sequence ambiguities were resolved by generating and sequencing appropriate PCR fragments.

*Gene Identification and Functional Annotation*

Initial annotation was performed using MAGPIE [Gaasterland and Sensen, 1996]. Final annotation was performed with PEDANT-PRO [Frishman et al., 2001] from open reading frames predicted by ORPHEUS [Frishman et al., 1998]. PEDANT uses in addition to BLAST searches against a non-redundant database automatically collected results from analyses performed with BLOCKS, PROSITE, PFAM, 3D, SCOP and COGs. All automatically generated assignments were reviewed and curated in two manual annotation rounds. Frameshifts were detected and corrected after the first round of manual annotation. Small ORFs overlapping with validated ORFs were deleted from the final ORF list when no functional assignment could be justified. The N-termini of the predicted proteins were adjusted to the best database match only in cases where sufficient reason for shortening of the N-terminus was available. Gene designations were taken from SWISS-PROT [Boeckmann et al., 2003] or from respective references. tRNA genes were located using tRNAscan-SE [Lowe and Eddy, 1997].

## Acknowledgements

This work was supported by Deutsche Bundesstiftung Umwelt (DBU) grant No. AZ 13040/09. We thank Bettina Haberl and Romi Müller for technical assistance in DNA sequencing, Irmgard Becker, Klaus Gellner and Wolf-Dieter Leuschner (all formerly Epidauros Biotechnologie GmbH) for help with DNA sequence assembly and annotation. The development of the MAGPIE system is supported through Genome Canada and Genome Alberta. C.W.S. is the iCORE/Sun Microsystems industrial chair in Applied Bioinformatics.

## References

Ballschmiter M, Armbrecht M, Ivanova K, Antranikian G, Liebl W: AmyA, an α-amylase with β-cyclodextrin-forming activity, and AmyB from the thermoacidophilic organism *Anaerobranca gottschalkii:* two α-amylases adapted to their different cellular localizations. Appl Environ Microbiol 2005;71: 3709–3715.

Bao Q, Tian Y, Li W, Xu Z, Xuan Z, Hu S, Dong W, Yang J, Chen Y, Xue Y, Xu Y, Lai X, Huang L, Dong X, Ma Y, Ling L, Tan H, Chen R, Wang J, Yu J, Yang H: A complete sequence of the *T. tengcongensis* genome. Genome Res 12:689–700.

Bertoldo C, Armbrecht M, Becker F, Schäfer T, Antranikian G, Liebl W: Cloning, sequencing, and characterisation of a heat- and alkali-stable type I pullulanase from *Anaerobranca gottschalkii*. Appl Environ Microbiol 2004;70:3407–3416.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al: The SWISS-PROT protein knowledgebase and its supplement TrEMBL. Nucleic Acids Res 2003;31:365–370.

Cambillau C, Claverie JM: Structural and genomic correlates of hyperthermostability. J Biol Chem 2000;275:32383–32386.

Ewing B, Green P: Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998;8:186–194.

Ewing B, Hillier L, Wendl M, Green P: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998;8:175–185.

Fleischmann R, Adams MD, White O, et al: Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995;269:496–512.

Frishman D, Albermann K, Hani J, Heumann K, Metanomski A, Zollner A, Mewes HW: Functional and structural genomics using PEDANT. Bioinformatics 2001;17:44–57.

Frishman D, Mironov A, Mewes HW, Gelfand M: Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucleic Acids Res 1998;26:2941–2947.

Gaasterland T, Sensen CW: Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. Biochimie 1996;78:302–310.

Gordon D, Abajian C, Green P: Consed: a graphical tool for genome finishing. Genome Res 1998;8:195–202.

Haft DH, Selengut J, Mongodin EF, Nelson KE: A guild of 45 CRISPR-associated (cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLOS Comp Biol 2006;6:474–483.

Lao PJ, Forsdyke DR: Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with adenine and guanine. Genome Res 2000;10:1–20.

Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC: The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res 2006;34: D332–D334.

Lowe T, Eddy SR: tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucl Acids Res 1997;25: 955–964.

Prowe SG, Antranikian G: *Anaerobranca gottschalkii* sp. nov., a novel thermoalkaliphilic bacterium that grows anaerobically at high pH and temperature. Intl J Syst Evol Microbiol 2001;51:457–465.

Prowe SG, van de Vossenberg JLCM, Driessen AJM, Antranikian G, Konings WN: Sodium-coupled energy transduction in the newly isolated thermoalkaliphilic strain LBS3. J Bacteriol 1996;178:4099–4104.

Thiemann V, Donges C, Prowe SG, Sterner R, Antranikian G: Characterisation of thermoalkalistable cyclodextrin glycosyltransferases from the anaerobic thermoalkaliphilic bacterium *Anaerobranca gottschalkii*. Arch Microbiol 2004;182:226–235.

Thiemann V, Saake B, Vollstedt A, Schäfer T, Puls J, Bertoldo C, Freudl R, Antranikian G: Heterologous expression and characterisation of a novel branching enzyme from the thermoalkaliphilic anaerobic bacterium *Anaerobranca gottschalkii*. Appl Microbiol Biotechnol 2006;72:60–71.