HAL
open science

The complete genome sequence of Lactobacillus bulgaricus reveals extensive and ongoing reductive evolution

Maarten van de Guchte, Stéphanie Penaud, Christine Grimaldi, Valérie Barbe, K. Bryson, Pierre P. Nicolas, S. Otzas, S. Mangenot, A. Couloux, Valentin Loux, et al.

To cite this version:

Maarten van de Guchte, Stéphanie Penaud, Christine Grimaldi, Valérie Barbe, K. Bryson, et al.. The complete genome sequence of Lactobacillus bulgaricus reveals extensive and ongoing reductive evolution. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103 (24), pp.9274-9279. 10.1073/pnas.0603024103 . hal-02668654

HAL Id: hal-02668654

https://hal.inrae.fr/hal-02668654v1

Submitted on 31 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The complete genome sequence of *Lactobacillus bulgaricus* reveals extensive and ongoing reductive evolution

M. van de Guchte*†, S. Penaud*, C. Grimaldi*, V. Barbe‡, K. Bryson§¶, P. Nicolas§, C. Robert‡, S. Oztas‡, S. Mangenot‡, A. Couloux‡, V. Loux§, R. Dervyn*, R. Bossy§, A. Bolotin*, J.-M. Batto*, T. Walunas‖, J.-F. Gibrat§, P. Bessières§, J. Weissenbach‡**, S. D. Ehrlich*, and E. Maguin*

*Génétique Microbienne and §Mathématique, Informatique et Génome, Institut National de la Recherche Agronomique, 78352 Jouy en Josas Cedex, France; ‡Genoscope, 2 Rue Gaston Crémieux, 91057 Evry Cedex, France; ‖Integrated Genomics Inc., 2201 West Campbell Park Drive, Chicago, IL 60612; and **Centre National de la Recherche Scientifique, Unité Mixte de Recherche 8030, 2 Rue Gaston Crémieux, 91057 Evry Cedex, France

Communicated by Todd R. Klaenhammer, North Carolina State University, Raleigh, NC, April 14, 2006 (received for review October 14, 2005)

***Lactobacillus delbrueckii* ssp. *bulgaricus* (*L. bulgaricus*) is a representative of the group of lactic acid-producing bacteria, mainly known for its worldwide application in yogurt production. The genome sequence of this bacterium has been determined and shows the signs of ongoing specialization, with a substantial number of pseudogenes and incomplete metabolic pathways and relatively few regulatory functions. Several unique features of the *L. bulgaricus* genome support the hypothesis that the genome is in a phase of rapid evolution. (*i*) Exceptionally high numbers of rRNA and tRNA genes with regard to genome size may indicate that the *L. bulgaricus* genome has known a recent phase of important size reduction, in agreement with the observed high frequency of gene inactivation and elimination; (*ii*) a much higher GC content at codon position 3 than expected on the basis of the overall GC content suggests that the composition of the genome is evolving toward a higher GC content; and (*iii*) the presence of a 47.5-kbp inverted repeat in the replication termination region, an extremely rare feature in bacterial genomes, may be interpreted as a transient stage in genome evolution. The results indicate the adaptation of *L. bulgaricus* from a plant-associated habitat to the stable protein and lactose-rich milk environment through the loss of superfluous functions and protocooperation with *Streptococcus thermophilus*.**

adaptation | GC evolution | large inverted repeat | protocooperation | rRNA number

Lactobacillus delbrueckii ssp. *bulgaricus* (*L. bulgaricus*) is one of the economically most important representatives of the heterogeneous group of lactic acid bacteria, with a worldwide application in yogurt production. Yogurt has long been recognized as a nutritious, natural, and safe component of a healthy diet and is at the basis of the concept of probiotics (1, 2). A well documented health benefit of the consumption of yogurt containing live *L. bulgaricus* and *Streptococcus thermophilus* is an attenuation of lactose intolerance (3). In addition, immune modulation and diarrhea-alleviating effects have been reported (4), and both *L. bulgaricus* and *S. thermophilus* have been implicated in these effects (3, 5). During yogurt fermentations protocooperation between these two bacteria results in an accelerated acidification, but the mechanisms involved are not completely understood (6).

Among the lactic acid bacteria, *L. bulgaricus* belongs to the acidophilus complex, a group of lactobacilli related to *Lactobacillus acidophilus*, *Lactobacillus johnsonii*, and *Lactobacillus gasseri*, which have been used as probiotic cultures. Although within this group *L. bulgaricus* is considered unique because of its atypical GC content, until recently the lack of tools for genetic manipulation has severely hampered a more detailed analysis of this organism (7, 8, ††).

Here we present the genome sequence of *L. bulgaricus* strain ATCC11842, originally isolated from bulgarian yogurt by S. Orla-Jensen in 1919 (unpublished work). The analysis of this genome and comparison to other members of the acidophilus complex and *S. thermophilus* (9) have contributed to a more complete understanding of its phylogenetic position and revealed important details about its specialized adaptation to milk and protocooperation with *S. thermophilus*.

Results and Discussion

Primary Sequence Characteristics. The size of the *L. bulgaricus* genome was estimated at ≈1.8 Mbp on the basis of results obtained after pulsed-field gel electrophoresis of chromosomal DNA digests (data not shown). This estimation was confirmed in the present whole-genome sequencing project, where a circular chromosomal sequence of 1,864,998 bp has been assembled. The primary characteristics of this sequence are presented in Table 1.

The overall GC content (49.7%) differs significantly from that of the closely related species *L. acidophilus* (34.7%; ref. 10) and *L. johnsonii* (34.6%; ref. 11), a difference that is mainly due to very important differences at codon position 3 (GC3): 65.0% GC in *L. bulgaricus* as compared with 25.0% and 24.4% in *L. acidophilus* and *L. johnsonii*, respectively. Whereas the latter values fit the strong correlation that can be observed between GC3 and overall GC content in bacteria (Fig. 1), the *L. bulgaricus* GC3 value (65.0%) strongly deviates from the expected value (54.0%). Because the evolution at codon position 3 is generally much faster than at positions 1 and 2, the high GC3 value suggests that the *L. bulgaricus* genome is in an active phase of evolution toward a higher GC content or that in *L. bulgaricus* the correlation between GC3 and overall GC content has been lost or changed.

In the *L. bulgaricus* genome, a large number of coding sequences (CDS) have been annotated as "fragments" on the basis of BLASTP and BLASTX (12) results. These fragments correspond to 270 different pseudogenes (of which 43 are remnants of transposase coding genes) that are regularly distributed over the genome, with a slight underrepresentation in the region from 283 kbp to 672 kbp (Fig. 2). This high number of pseudogenes suggests that the genome is in an active state of gene elimination and concomitant size reduction. Not counting pseudogenes, only 73% of the genome consists of (putative) coding sequences. A distinct 2.5-kbp noncoding region (starting at position 764 kbp) was identified that has all of the features of a CRISPR region (clustered regularly interspaced

Conflict of interest statement: No conflicts declared.

Abbreviations: CDS, coding sequences; IS, insertion sequence.

Data deposition: The *L. bulgaricus* genome sequence has been submitted to the European Molecular Biology Laboratory database (accession no. CR954253).

†To whom correspondence should be addressed. E-mail: maarten.vandeguchte@jouy.inra.fr.

¶Present address: Department of Computer Science, University College London, London WC1E 6BT, United Kingdom.

††Sasaki, T., Ito, Y. & Sasaki, Y. (1993) *FEMS Microbiol. Rev.* **12,** P8.

© 2006 by The National Academy of Sciences of the USA

Table 1. Characteristics of the *L. bulgaricus* ATCC11842 genome

Genome size, bp	1,864,998
Overall GC content, %	49.7
GC content of CDS,* %	51.6
GC content of CDS* at codon position 3, %	65.0†
Number of CDS*	1562
CDS* as % of genome sequence	73
Number of CDS* with unknown function	598
Number of pseudogenes‡	270
Number of rrn operons	9
Number of tRNA genes	95

*CDS not annotated as ''fragment.''
‡Corresponding to 534 CDS annotated as ''fragment.''
†64.4% if pseudogenes are included in the analysis.

short palindromic repeats) (13). CRISPRs have been interpreted as traces of past invasions by extrachromosomal elements and have been hypothesized to provide immunity against foreign DNA expression by coding antisense RNA (14).

L. bulgaricus contains a relatively high number of rRNA and tRNA genes. In the firmicutes, a clear correlation can be observed between the numbers of rRNA and tRNA genes and between each of these and genome size (Fig. 3 and Fig. 5, which is published as supporting information on the PNAS web site). Although in *L. bulgaricus* the first correlation is respected, the numbers of tRNA and rRNA genes are ≈50% higher than the average and 20–30% higher than the highest values observed so far for this genome size, corresponding to values found for genomes of 3–4 Mbp in size. Whereas variation in rRNA and tRNA gene copy numbers may be related to the capacity to respond to changing environmental conditions (15, 16), the exceptionally high copy numbers found in *L. bulgaricus* likely indicate that the genome has undergone a recent phase of size reduction.

Genome Structure and Organization: Replication Terminus with a 47.5-kbp Inverted Repeat and Duplicated *dif* Sites. A sharp change in the sign of the GC skew coincides with the location of the *dnaA* and *dnaN* genes (data not shown), indicating the likely presence of the origin of replication in this region. The presence of several DnaA boxes in this region confirms this hypothesis (17). On the opposite side of the genome, a second change in the sign of the GC skew locates the replication terminus in the region between positions 940000 and 946000 (data not shown). An intriguing feature of the replication terminus region of the ATCC11842 genome is that it

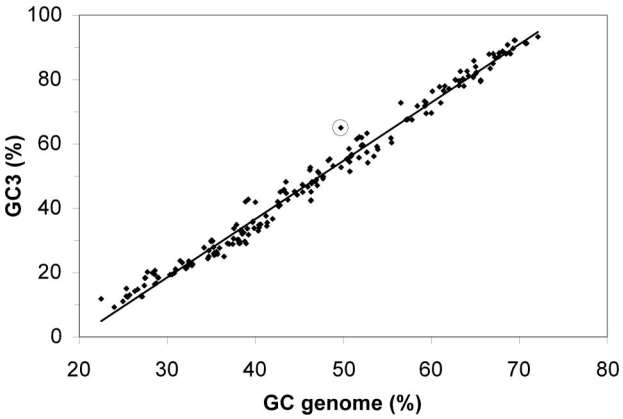

Fig. 1. Relationship between GC content at position 3 of coding sequences (GC3) and genomic GC content in 232 eubacterial genomes. The *L. bulgaricus* value was calculated excluding pseudogenes. The *L. bulgaricus* data point is circled.

Fig. 2. Genome atlas of the *L. bulgaricus* genome. The seven circles (outer to inner) show the following. Circle 1, pseudogenes on positive (red) or negative (blue) strand. Circle 2, IS elements (transposases or ISL4-related hypothetical genes). Elements with fewer than four copies are represented in gray, and elements with more than five copies are represented by separate colors of red (ISL7), purple (ISL4), blue (ISL5), and green (ISL4–5). (See also Table 2.) Circle 3, CDS (excluding pseudogenes and transposases) on positive (red) or negative (blue) strand. Circle 4, rRNA (red) and tRNA (green) genes. Circle 5, [(G-C)/window size (2000)], from less than −0.1 (cyan) to more than +0.1 (red). Circle 6, [(A+T)/window size (500)], from <0.3 (cyan) to >0.7 (red). Circle 7, position on the genome. The genome atlas was constructed by using GENEWIZ software (53).

contains an inverted repeat of 47.5 kbp (positions 918952–966484) enclosing a unique central region of 1.4 kbp between the repeated sequences (23 kbp each; see Fig. 6, which is published as supporting information on the PNAS web site). In bacterial genomes, inverted repeats of this size are extremely rare. To our knowledge, they have been observed only in the replication terminus region of certain *Streptomyces* chromosomes that have circularized after telomere deletion (18). Preliminary results of one primer PCR amplification in 30 different *L. bulgaricus* strains show that an inverted repeat is conserved in most or all strains (data not shown). The size of the unique central sequence and/or the size of the repeat may largely vary between strains, however, and this feature may represent a transient stage in the evolution of the *L. bulgaricus* genome.

In *L. bulgaricus* the repeated sequence includes the putative *dif* site (Figs. 6 and 7, which are published as supporting information on the PNAS web site), which is the recombination site involved in the resolution of chromosome dimers (19). Consequently, *L. bulgaricus* contains two putative *dif* sites in opposite orientation at a distance of 2.5 kbp (0.55 kbp of repeated sequence at either side of the unique central sequence of 1.4 kbp) instead of the unique site found in other microbial genomes, such as *Bacillus subtilis* (Fig. 7). One may hypothesize that this configuration could interfere with the efficient segregation of dimeric chromosomes that may be formed during replication, because recombination between two *dif* sites of opposite orientation in a chromosome dimer would result in the inversion of a part of the dimer rather than resolution of the dimer. Such FtsK XerCD-dependent inversions have been observed in plasmids containing *dif* sites of opposite orientation *in vitro* (20). The results of PCR analysis (data not shown) suggest that

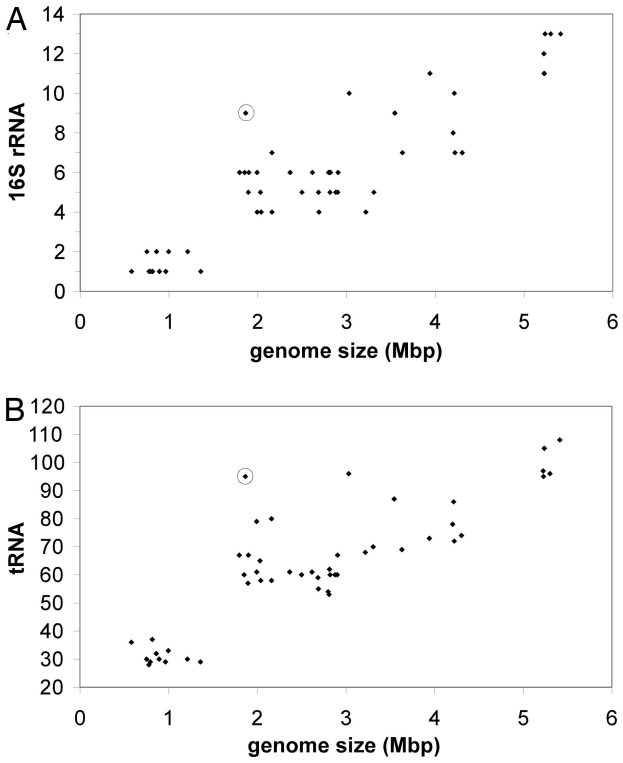

Fig. 3. rRNA and tRNA genes as a function of genome size. (*A*) Relationship between the number of 16S rRNA genes and genome size in 54 firmicutes genomes. (*B*) Relationship between the number of tRNA genes and genome size in 54 firmicutes genomes. *L. bulgaricus* (1.9 Mbp, 9 16S rRNA, and 95 tRNA) data points are circled. For comparison: *L. acidophilus* (2.0 Mbp, 4 16S rRNA, and 61 tRNA), *L. johnsonii* (2.0 Mbp, 6 16S rRNA, and 79 tRNA), and *L. plantarum* (3.3 Mbp, 5 16S rRNA, and 70 tRNA).

the 1.4-kbp unique sequence enclosed by the repeated sequences may be present in either of two orientations in a population of strain ATCC11842 (Fig. 6). This result would be expected as a consequence of such a site-specific recombination between two *dif* sites of opposite orientation or, alternatively, of homologous recombination between the inverted repeats.

Nonrandom Distribution of Mobile Elements. The *L. bulgaricus* ATCC11842 genome contains a large number of transposases and remnants thereof (Table 2, which is published as supporting information on the PNAS web site), most of which have been described earlier in *Lactobacillus delbrueckii* insertion sequence (IS) elements (ISL3, ISL4, ISL5, ISL4–5, ISL6, ISL7, and ISldl1) (21–23). In the present work, seemingly intact transposases of two new types could be detected [named ISL8 (Ldb0326 and Ldb2033) and ISL9 (Ldb1987)], as well as remnants of several other types.

Intriguingly, a region of 415 kbp (22% of the genome) between positions 279 kbp and 694 kbp is completely exempt of IS elements, whereas (remnants of) the more abundant IS elements seem randomly distributed over the remaining 78% of the genome (Fig. 2). This discontinuity in the distribution of IS elements coincides with the region, mentioned above, where pseudogenes are less abundant. The absence of IS elements does not, however, explain the difference in the frequency of pseudogenes, because in the whole genome inactivation of only 9 of 227 nontransposase pseudogenes occurred from insertion of IS elements. The positions of transposases often coincide with regions of locally elevated A+T content (Fig. 2 and Fig. 8, which is published as supporting information on the PNAS web site), often considered as possible signs of horizontal gene transfer.

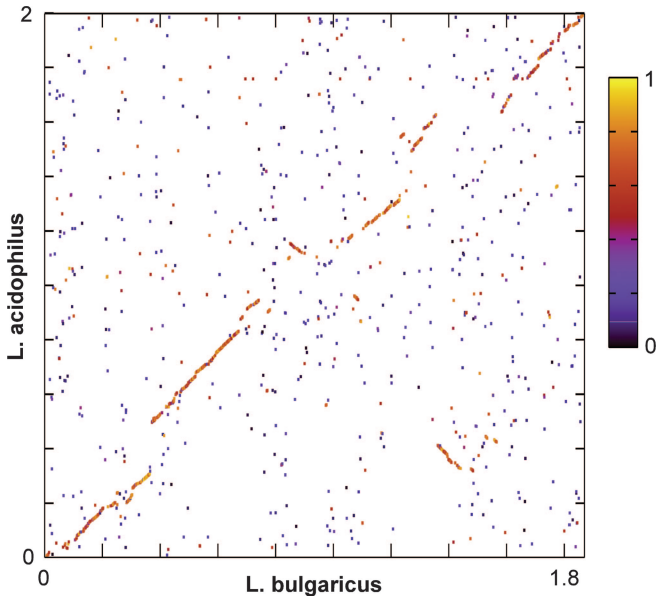

Fig. 4. Synteny between *L. bulgaricus* and *L. acidophilus* genomes. *x* axis, position on *L. bulgaricus* genome (Mbp); *y* axis, position on *L. acidophilus* genome (Mbp). 0, replication origin. Colors indicate protein similarity by BLAST score ratio (25) according to the scale on the right.

In contrast to many other lactic acid bacteria, *L. bulgaricus* ATCC11842 does not contain any prophage. Only four pseudogenes of putative phage origin were detected in a large region around the replication terminus, where the *attB* sites of the elsewhere-described bacteriophages mv4 (24) and JCL1032 (K.-A. Riipinen, personal communication) are found.

Synteny Among *L. bulgaricus*, *L. acidophilus*, and *L. johnsonii* Genomes. A clear global synteny is observed among the similarly sized genomes of *L. bulgaricus* and *L. acidophilus* or *L. johnsonii* (Fig. 4 and Fig. 9, which is published as supporting information on the PNAS web site). The most important perturbations of synteny are found in a region of ≈300 kbp around the replication terminus of the *L. bulgaricus* genome and in two smaller regions at ≈400 kbp at either side of the replication origin. By using the criterion of BLAST score ratio with a cutoff value of 0.4 (25) or the results of a BLASTP reciprocal best-hit analysis, ≈55–60% of the *L. bulgaricus* proteins can be considered as having homologues in *L. acidophilus* and *L. johnsonii* and constitute the common backbone of these three closely related genomes. Between 25% and 35% of the proteins appear unique *to L. bulgaricus* (Fig. 10 and Tables 3–5, which are published as supporting information on the PNAS web site). Although the differences between these genomes are likely to be of importance in their respective ecological niches, they mainly concern unknown functions. The most marked differences among known functions comprise the presence of complete pathways for the biosynthesis of folate and saturated fatty acids (see below) in *L. bulgaricus*, both of which are partially lacking in *L. acidophilus* and *L. johnsonii*. Likewise, *L. bulgaricus* contains all or most of the enzymes necessary for the biosynthesis of purines and pyrimidines, respectively, of which several are absent from *L. acidophilus* and *L. johnsonii* and can grow in a purine-free medium (data not shown). *L. bulgaricus* contains *ppx* and *ppk* genes that are involved in the turnover of polyphosphate and thereby may play a role in adaptation to stress conditions (26). Inversely, proteins of known function that are present in *L. acidophilus* or *L. johnsonii* but not in *L. bulgaricus* are mainly involved in sugar transport and metabolism (data not shown).

Restriction Modification Systems. The *L. bulgaricus* genome contains a complete type I restriction modification (R/M) system (Ldb1051, Ldb1052, and Ldb1053), an additional type I specificity subunit (Ldb1055), and a putative Mrr type restriction endonuclease (Ldb0474). The presence of these R/M systems may explain part of the difficulties encountered previously to transform this strain. A type III restriction modification system appears inactive as a consequence of mutations in the gene encoding the endonuclease (Ldb1227–Ldb1229), whereas the corresponding methyltransferase (Ldb1230) seems intact.

Regulatory Functions: Two SigA Homologues and Relatively Few Regulators. A unique feature of the *L. bulgaricus* genome is that it encodes two 58% identical SigA homologues in two genes that are organized in tandem and apparently cotranscribed. While one of the two encoded proteins (Ldb1246) shows −35 and −10 recognition domains that are 100% identical to those in *B. subtilis* SigA, the other (Ldb1245) shows some important differences in these domains (Fig. 11, which is published as supporting information on the PNAS web site). It is unknown whether this second copy is active and recognizes different promoters. If so, it may be considered a new σ factor developed after gene duplication. Apart from these SigA homologues, two putative extracytoplasmic function (ECF)-type σ factors (Ldb0066 and Ldb1881), one putative σ factor of unidentified type (Ldb1677), and one anti-σ factor (Ldb1880) were found.

Relatively few genes encoding transcriptional regulators were identified in *L. bulgaricus*, and the difference with *Lactobacillus plantarum* (27) was especially striking (53 and 234 predicted genes, respectively; Table 6, which is published as supporting information on the PNAS web site), even when considering the smaller genome size of *L. bulgaricus* (1.9 Mb as opposed to 3.3 Mb for *L. plantarum*). Both these features likely reflect adaptations to the stable and nutritionally rich milk environment, where fewer biosynthetic functions and less adaptive regulation are required. The largely different sugar metabolic capacities of these two species are reflected by the difference in the number of LacI type regulators. Notable differences in MarR-, MerR-, TetR-, and ArsR-type regulators may indicate significantly different levels of resistance to a variety of compounds and stress conditions or differing strategies to cope with them. Regulatory circuits may have a less complicated structure in *L. bulgaricus*, as exemplified by the predicted HrcA regulon (Table 7, which is published as supporting information on the PNAS web site) (28). In *B. subtilis*, *dnaK* and *groESL* on one hand and *clp* on the other hand are under the control of two different regulators, HrcA and CtsR, respectively. *L. bulgaricus* encodes only a HrcA homologue. CIRCE boxes, the operators for HrcA (29), are found upstream of *dnaK* and *groESL* and upstream of *clpP* and *clpE*, suggesting a broader role for HrcA in *L. bulgaricus*. A similar organization was found in *L. acidophilus* and *L. johnsonii* (Table 7) (10, 11).

Pseudogenes and Specialization Through Loss of Function. *L. bulgaricus* strain ATCC11842 was originally isolated from bulgarian yogurt. Traditional yogurt-making involves the sequential transfer of samples of yogurt cultures to fresh milk. With the first records of yogurt (kisim) dating to 3200 before Christ (30), one would logically predict that *L. bulgaricus* adapted to this environment over time. Traces of such an adaptation process were found throughout the *L. bulgaricus* genome in the form of pseudogenes and incomplete metabolic pathways.

A total of 270 pseudogenes have been detected in the *L. bulgaricus* genome, including 43 transposases. The remaining 227 genes represent 12% of the total number of protein-coding genes. This percentage of pseudogenes is remarkably high (31) and may indicate a recent and ongoing process of specialization. A function, albeit inactivated, could be assigned to 81 pseudogenes. For 30 of these genes, intact paralogues are present in the genome whereas for the remaining 51 genes no intact paralogues could be detected, indicating that the corresponding functions are completely lost (Table 8, which is published as supporting information on the PNAS web site). These include genes implicated in carbohydrate metabolism, amino acid and cofactor biosynthesis, and competence development. For 146 pseudogenes it was difficult to determine whether their function could be assumed by a paralogue as only a general function (e.g., "ABC transporter") or no function at all could be attributed to these genes.

Complete transport systems for the milk sugar lactose (lactose permease), mannose/glucose [phosphotransferase system (PTS) IIABCD], fructose (PTS IIABC), and glycerol (glycerol uptake facilitator) are present, but several other transport systems appeared incomplete. Remnants were found of cellobiose (PTS subunit IIC), sucrose (PTS subunit IIA), maltose (partial subunits of a dedicated ABC transporter), and some other sugar transport systems of undetermined specificity (several isolated PTS IIA subunits). Together with remnants (pseudogenes) of key enzymes in carbohydrate-specific metabolic pathways (glycerol kinase, 6-phospho-β-glucosidase/galactosidase, α-amylase, and mannitol-1-phosphate 5-dehydrogenase), these reveal the prior existence of metabolic capacities that may notably have served in a plant-associated environment.

Apart from the intrinsic evidence for specialization through loss of function provided by the presence of pseudogenes, the complete absence of a large number of enzymes involved in the biosynthesis of amino acids (incomplete or lacking pathways) also suggests an adaptation to the protein-rich milk environment. Here the presence of an extracellular protease, several amino acid transport systems, two complete peptide transport systems, and numerous peptidases (Tables 9 and 10, which are published as supporting information on the PNAS web site) would render most amino acid biosynthesis pathways superfluous. As a consequence, the extracellular protease has become essential for growth in milk (32).

Interestingly, only part of this adaptation to the milk environment is shared by *S. thermophilus*. Whereas *L. bulgaricus* has lost most of its amino acid biosynthesis capacity, *S. thermophilus* retains its ability to synthesize all amino acids, except histidine (33). This difference may be explained by the fact that *S. thermophilus* does not generally possess an extracellular protease to exploit the rich source of milk proteins. Perhaps *S. thermophilus* either adapted to the milk environment independently or coevolved with the protease producing *L. bulgaricus* but retained an advantage in conserving its capacity for biosynthesis of amino acids.

Central Carbohydrate Metabolism. As a consequence of the (partial) inactivation of several sugar transport and degradation pathways and inactivation of the LacR repressor (22), carbohydrate metabolism in *L. bulgaricus* appears to prefer the milk sugar lactose. After transport and hydrolysis, this organism selectively metabolizes the glucose moiety, because no enzymes are present to use galactose, which is known to accumulate in the culture medium (6, 34). Whereas glucose, fructose, and mannose can also be metabolized, utilization of lactose results in higher growth rates (35). A similar accumulation of galactose in the culture medium is observed when *S. thermophilus* is grown on lactose, albeit for a different reason. In contrast to *L. bulgaricus*, *S. thermophilus* does possess the necessary genes to direct catabolism of galactose, but these genes are generally repressed in the presence of lactose (36).

Although a putative transporter for ribose and a ribokinase are present, *L. bulgaricus* cannot grow on this sugar as the only carbon source (data not shown), presumably as a consequence of an incomplete pentose–phosphate pathway (see below). A complete glycolytic pathway is present leading to the production of pyruvate, which can subsequently be converted to L- or D-lactate to regenerate NAD^+.

Three particular features are noted relating to carbohydrate

metabolism (Fig. 12, which is published as supporting information on the PNAS web site).

The first substrate level phosphorylation step of glycolysis, in which glyceraldehyde-3-P is converted to 3-P-glycerate via 1,3-bi-P-glycerate, yielding NADH and ATP, can be bypassed by a nonphosphorylating $NADP^+$-dependent glyceraldehyde-3P dehydrogenase (GapN). This would result in glycolysis yielding pyruvate and NADPH with no net ATP production. The GapN protein is not present in other sequenced *Lactobacillus* species but is very common among streptococci. In *Streptococcus mutans*, the role of GapN has been suggested to be the production of NADPH because this bacterium lacks the NADPH-generating enzymes in the oxidative branch of the pentose phosphate pathway (37). GapN may also be implicated in an additional level of regulation of glycolysis, allowing the adjustment of metabolic and energy fluxes (38).

L. bulgaricus contains an apparently intact oxidative, NADPH-generating branch of the pentose phosphate pathway. However, distinct from the related species *L. acidophilus* and *L. johnsonii*, the nonoxidative branch lacks a transketolase. Because pentose–phosphate cycling is generally regarded as a mechanism enabling the efficient synthesis of NADPH (39), the interruption of this cycle would support the hypothesis of an NADPH-producing role for GapN under conditions where the oxidative pentose–phosphate pathway is not able to meet the demand.

Similar to *L. acidophilus* and *L. johnsonii*, no pyruvate dehydrogenase complex is present, nor are other enzymes that could generate acetyl-CoA directly from pyruvate under anaerobic conditions. The most likely alternative route would convert the pentose phosphate pathway intermediate D-xylulose-5-P to acetyl-P, which can subsequently be used to generate acetyl-CoA. Acetyl-P may also be produced from acetate, which has been found to be an essential component of a chemically defined medium for *L. bulgaricus* (35). One may speculate that acetate fulfills this role under conditions where the pentose–phosphate pathway would be unable to supply sufficient substrates for acetyl-CoA generation.

Acetyl-CoA may subsequently be used for the production of saturated fatty acids. It is not clear whether *L. bulgaricus* would be able to produce unsaturated fatty acids, however, because a FabA or FabM homologue is missing (40). Failure to produce acetyl-CoA or the inability to produce unsaturated fatty acids may explain the need for Tween 80 in laboratory media for *L. bulgaricus*. Oleic acid, the unsaturated fatty acid in Tween 80, is also available in milk.

Protocooperation Between *L. bulgaricus* and *S. thermophilus*. *L. bulgaricus* and *S. thermophilus* have long been known to stimulate each other's growth and product acidification during milk fermentation in a process called protocooperation. Several factors have been shown or postulated to be responsible for this effect, of which the most obvious is that *L. bulgaricus* possesses an extracellular cell wall-bound proteinase (32) which, through degradation of milk proteins, can supply peptides and amino acids to *S. thermophilus*, which generally does not possess such a protease. Among the other protocooperative factors are formate and CO_2 produced by *S. thermophilus*, which stimulate the growth of *L. bulgaricus* (42).

Analysis of the *L. bulgaricus* genome revealed additional factors that may play a role in protocooperation. (*i*) The genome sequence encodes a full set of genes for the biosynthesis of folate (Fig. 13, which is published as supporting information on the PNAS web site), a cofactor in many metabolic reactions and an essential component of the human diet (42). *L. bulgaricus* does not, however, possess a means to produce *p*-aminobenzoic acid (PABA) which feeds into this pathway. *S. thermophilus* does possess the necessary enzymes to produce PABA and folate, and as a consequence yogurt is generally regarded as a source of PABA. *L. bulgaricus* may thus benefit from elevated levels of PABA (and folate) when cocultured with *S. thermophilus*. (*ii*) Two features in the genome suggest that *L. bulgaricus* has a particular need for polyamines, which participate in many cellular processes and may play a role in tolerance to oxidative stress (43, 44). First, two ABC transporters are found that could serve the uptake of putrescine and/or spermidine from the culture medium. One of these is encoded by a *potABCD* operon (52% GC) in which the genes are arranged in the same order as found in most other bacteria (ERGO). The other is encoded by atypical, AT-rich (38% GC) genes that are present in the order *potBCAD* found in very few bacteria and probably were acquired by horizontal transfer. Second, two genes are found that encode ornithine decarboxylases that catalyze the conversion of ornithine to putrescine. These enzymes are rare in Gram-positive bacteria, except for lactobacilli of the acidophilus group. Because *L. bulgaricus* does not possess the genes necessary to produce ornithine, one may hypothesize that during yogurt fermentation ornithine is provided by *S. thermophilus*, which does possess the enzymes necessary to produce ornithine, but no ornithine decarboxylase (Fig. 14, which is published as supporting information on the PNAS web site). Interestingly, *S. thermophilus* does also possess an ABC transporter dedicated to the uptake of spermidine/putrescine, and, thus, it could be speculated that *L. bulgaricus* and *S. thermophilus* mutually benefit from an exchange of ornithine and putrescine.

A prediction of protein localization (Tables 11 and 12, which are published as supporting information on the PNAS web site) revealed a limited number of putative cell wall-bound and extracellular proteins. Perhaps some of these could contribute to direct contacts between *L. bulgaricus* and *S. thermophilus* as visualized by Bolotin *et al.* (9). From a fermentation viewpoint, the exopolysaccharide (eps) gene clusters present in *L. bulgaricus* (Table 13, which is published as supporting information on the PNAS web site) and *S. thermophilus* appear highly diverse. Variation of these eps clusters among strains is likely important to the texture characteristics of yogurt and merits further investigation.

Stress Resistance. The *L. bulgaricus* genome sequence reveals few of the genes that are known to be involved in the resistance to oxidative stress or low pH, even though this lactic acid bacterium acidifies the culture medium and can exhibit optimal growth under microaerobic conditions (45). The only enzyme that could eliminate oxygen appears to be pyruvate oxidase, but no catalase is present to detoxify the H_2O_2 produced in this reaction. Strain ATCC11842 does not appear to possess the NADH oxidase described by Marty-Teysset *et al.* (46) in another *L. bulgaricus* strain. Other genes involved in oxidative stress resistance are two thioredoxins, two thioredoxin reductases, and possibly a homologue of the regulatory RNA polymerase-binding protein Spx (47). The uptake of polyamines or their precursors may also provide an alternative means of improving oxidative stress tolerance (see above). Acid stress caused by the production of lactic acid may be countered primarily by the action of an H+ transporting ATPase. The presence of two ornithine decarboxylases and several cation:proton antiporters may also assist in the stabilization of intracellular pH, as in *L. acidophilus* (48).

Conclusions

The *L. bulgaricus* genome sequence reveals a number of features that support the hypothesis of a genome in a rapid phase of evolution. Among the lactobacilli of the acidophilus complex, the species *L. delbrueckii* is often regarded as atypical because of its strongly differing GC content. This difference has been the reason to rename what was originally called the *L. delbrueckii* group of lactobacilli and give it the name of a more representative member, *L. acidophilus* (49). The present genome sequence reveals that the observed difference in GC content is mainly due to a difference in GC content at codon position 3, which may be interpreted as the result of recent evolution and justify the conclusion that *L. bulgaricus* is less atypical than once thought. This conclusion is corroborated by the clear global synteny among the genomes of *L. bulgaricus*, *L. acidophilus*, and *L. johnsonii*, reflecting the presence of a common backbone.

The large number of pseudogenes and other features suggest that the *L. bulgaricus* genome is in a relatively recent state of evolution and adaptation to the dairy environment. For example, it has been reported that *L. bulgaricus* will deplete available folate sources rather than produce folate (50), an essential component of the human diet. The present genome sequence shows that all of the genes of the folate pathway are present and seemingly intact in *L. bulgaricus* ATCC11842.

In the light of a new turn in modern microbiological research with an increasing interest for the metagenome of the human gastrointestinal tract, the functional comparison of the industrial bacterium *L. bulgaricus* and the closely related bacteria from the gastrointestinal tract, *L. acidophilus* and *L. johnsonii*, offers an interesting perspective, particularly because *L. bulgaricus* has evolved away from the other members of the acidophilus complex by its rapid and specialized adaptation to an environment created by man, fermented milk.

Materials and Methods

Sequencing. The genome sequence of the *L. bulgaricus* type strain ATCC11842 was determined by using a shotgun sequencing and assembly strategy followed by multiplex long accurate PCR (51) and gap filling. Two types of libraries with inserts of 5 kbp (obtained by mechanical shearing) and 10 kbp (obtained by enzymatic digestion) were constructed in pcDNA2.1 (Invitrogen) and pBeloBAC11 (California Institute of Technology), respectively. Plasmid insert ends and PCR products were sequenced by using dye-primer and dye-terminator chemistries on LICOR4200L and ABI3700/ABI3730 sequencers. The genome sequence has been submitted to the European Molecular Biology Laboratory database under accession no. CR954253.

Annotation. A draft annotation of the genome sequence was generated by using self-training gene detection software (SHOW, http://ssb2.jouy.inra.fr/ssb/SHOW) and an interface for annotation developed at Institut National de la Recherche Agronomique (Agmial: K.B., V.L., R.B., P.N., S. Chaillou, M.v.d.G., S.P., E.M., M. Hoebeke, P.B., and J.-F.G., unpublished data). This draft annotation was manually checked and was updated where appropriate by using the tools provided in the Institut National de la Recherche Agronomique interface and the ERGO suite developed by Integrated Genomics (52).

Comparative Genome Analysis. Comparison of predicted *L. bulgaricus* proteins with predicted proteins from *L. acidophilus* and *L. johnsonii* and analysis of global synteny between the genomes of these bacteria were performed by using BLAST score ratio analysis software provided by Rasko *et al.* (25).

More details on the methods used are presented as *Supporting Materials and Methods*, which is published as supporting information on the PNAS web site.

We thank A. Sorokin, P. Serror, A. Fernandez, C. Chervaux, and T. Smokvina for useful discussions and N. Galleron, B. Quinquis, V. Brachet, M.-C. Beaussart, J. Musset, and L. Prieux for technical assistance. The work of S.P. and C.G. was in part financed by Danone Vitapole.

1. Heller, K. J. (2001) *Am. J. Clin. Nutr.* **73,** 374S–379S.
2. Metchnikoff, E. (1907) *The Prolongation of Life* (Heinemann, London).
3. Mercenier, A., Pavan, S. & Pot, B. (2003) *Curr. Pharm. Des.* **9,** 175–191.
4. Adolfsson, O., Meydani, S. N. & Russell, R. M. (2004) *Am. J. Clin. Nutr.* **80,** 245–256.
5. Perdigon, G., Maldonado Galdeano, C., Valdez, J. C. & Medici, M. (2002) *Eur. J. Clin. Nutr.* **56,** Suppl. 4, S21–S26.
6. Zourari, A., Accolas, J. P. & Desmazeaud, M. J. (1992) *Lait* **72,** 1–34.
7. Serror, P., Sasaki, T., Ehrlich, S. D. & Maguin, E. (2002) *Appl. Environ. Microbiol.* **68,** 46–52.
8. Ravin, V., Sasaki, T., Räisänen, L., Riipinen, K.-A. & Alatossava, T. (2006) *Plasmid* **55,** 184–193.
9. Bolotin, A., Quinquis, B., Renault, P., Sorokin, A., Ehrlich, S. D., Kulakauskas, S., Lapidus, A., Goltsman, E., Mazur, M., Pusch, G. D., *et al.* (2004) *Nat. Biotechnol.* **22,** 1554–1558.
10. Altermann, E., Russell, W. M., Azcarate-Peril, M. A., Barrangou, R., Buck, B. L., McAuliffe, O., Souther, N., Dobson, A., Duong, T., Callanan, M., *et al.* (2005) *Proc. Natl. Acad. Sci. USA* **102,** 3906–3912.
11. Pridmore, R. D., Berger, B., Desiere, F., Vilanova, D., Barretto, C., Pittet, A. C., Zwahlen, M. C., Rouvet, M., Altermann, E., Barrangou, R., *et al.* (2004) *Proc. Natl. Acad. Sci. USA* **101,** 2512–2517.
12. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
13. Jansen, R., Embden, J. D., Gaastra, W. & Schouls, L. M. (2002) *Mol. Microbiol.* **43,** 1565–1575.
14. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. (2005) *Microbiology* **151,** 2551–2561.
15. Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. (2000) *Appl. Environ. Microbiol.* **66,** 1328–1333.
16. Condon, C., Liveris, D., Squires, C., Schwartz, I. & Squires, C. L. (1995) *J. Bacteriol.* **177,** 4152–4156.
17. Yoshikawa, H. & Ogasawara, N. (1991) *Mol. Microbiol.* **5,** 2589–2597.
18. Uchida, T., Ishihara, N., Zenitani, H., Hiratsu, K. & Kinashi, H. (2004) *J. Bacteriol.* **186,** 3313–3320.
19. Kuempel, P. L., Henson, J. M., Dircks, L., Tecklenburg, M. & Lim, D. F. (1991) *N. Biol.* **3,** 799–811.
20. Aussel, L., Barre, F. X., Aroyo, M., Stasiak, A., Stasiak, A. Z. & Sherratt, D. (2002) *Cell* **108,** 195–205.
21. Germond, J. E., Lapierre, L., Delley, M., Mollet, B., Felis, G. E. & Dellaglio, F. (2003) *Mol. Biol. Evol.* **20,** 93–104.
22. Lapierre, L., Mollet, B. & Germond, J. E. (2002) *J. Bacteriol.* **184,** 928–935.
23. Ravin, V. & Alatossava, T. (2002) *Microbiol. Res.* **157,** 109–114.
24. Dupont, L., Boizet-Bonhoure, B., Coddeville, M., Auvray, F. & Ritzenthaler, P. (1995) *J. Bacteriol.* **177,** 586–595.
25. Rasko, D. A., Myers, G. S. & Ravel, J. (2005) *BMC Bioinformatics* **6,** 2.
26. Brown, M. R. & Kornberg, A. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 16085–16087.
27. Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O. P., Leer, R., Tarchini, R., Peters, S. A., Sandbrink, H. M., Fiers, M. W., *et al.* (2003) *Proc. Natl. Acad. Sci. USA* **100,** 1990–1995.
28. Chastanet, A. & Msadek, T. (2003) *J. Bacteriol.* **185,** 683–687.
29. Zuber, U. & Schumann, W. (1994) *J. Bacteriol.* **176,** 1359–1363.
30. Teuber, M. (1993) in *Biotechnology*, eds. Rehm, H.-J., Reed, G., Pühler, A. & Stadler, P. (Verlag Chemie, Weinheim, Germany), Vol. 1, pp. 325–366.
31. Lerat, E. & Ochman, H. (2004) *Genome Res.* **14,** 2273–2278.
32. Gilbert, C., Atlan, D., Blanc, B., Portailer, R., Germond, J. E., Lapierre, L. & Mollet, B. (1996) *J. Bacteriol.* **178,** 3059–3065.
33. Hols, P., Hancy, F., Fontaine, L., Grossiord, B., Prozzi, D., Leblond-Bourget, N., Decaris, B., Bolotin, A., Delorme, C., Dusko Ehrlich, S., *et al.* (2005) *FEMS Microbiol. Rev.* **29,** 435–463.
34. Welman, A. D. & Maddox, I. S. (2003) *J. Ind. Microbiol. Biotechnol.* **30,** 661–668.
35. Chervaux, C., Ehrlich, S. D. & Maguin, E. (2000) *Appl. Environ. Microbiol.* **66,** 5306–5311.
36. De Vin, F., Radstrom, P., Herman, L. & De Vuyst, L. (2005) *Appl. Environ. Microbiol.* **71,** 3659–3667.
37. Boyd, D. A., Cvitkovitch, D. G. & Hamilton, I. R. (1995) *J. Bacteriol.* **177,** 2622–2627.
38. Arutyunov, D. Y. & Muronetz, V. I. (2003) *Biochem. Biophys. Res. Commun.* **300,** 149–154.
39. Portais, J. C. & Delort, A. M. (2002) *FEMS Microbiol. Rev.* **26,** 375–402.
40. Marrakchi, H., Choi, K. H. & Rock, C. O. (2002) *J. Biol. Chem.* **277,** 44809–44816.
41. Driessen, F. M., Kingma, F. & Stadhouders, J. (1982) *Neth. Milk Dairy J.* **36,** 135–144.
42. Sybesma, W., Starrenburg, M., Tijsseling, L., Hoefnagel, M. H. & Hugenholtz, J. (2003) *Appl. Environ. Microbiol.* **69,** 4542–4548.
43. Igarashi, K. & Kashiwagi, K. (2000) *Biochem. Biophys. Res. Commun.* **271,** 559–564.
44. Chattopadhyay, M. K., Tabor, C. W. & Tabor, H. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 2261–2265.
45. Schiraldi, C., Adduci, V., Valli, V., Maresca, C., Giuliano, M., Lamberti, M., Carteni, M. & De Rosa, M. (2003) *Biotechnol. Bioeng.* **82,** 213–222.
46. Marty-Teysset, C., de la Torre, F. & Garel, J. (2000) *Appl. Environ. Microbiol.* **66,** 262–267.
47. Zuber, P. (2004) *J. Bacteriol.* **186,** 1911–1918.
48. Azcarate-Peril, M. A., Altermann, E., Hoover-Fitzula, R. L., Cano, R. J. & Klaenhammer, T. R. (2004) *Appl. Environ. Microbiol.* **70,** 5315–5322.
49. Schleifer, K. H. & Ludwig, W. (1995) *Syst. Appl. Microbiol.* **18,** 461–467.
50. Crittenden, R. G., Martinez, N. R. & Playne, M. J. (2003) *Int. J. Food Microbiol.* **80,** 217–222.
51. Sorokin, A., Lapidus, A., Capuano, V., Galleron, N., Pujic, P. & Ehrlich, S. D. (1996) *Genome Res.* **6,** 448–453.
52. Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov, E., Jr., Liolios, K., Joukov, V., Kaznadzey, D., Anderson, I., *et al.* (2003) *Nucleic Acids Res.* **31,** 164–171.
53. Pedersen, A. G., Jensen, L. J., Brunak, S., Staerfeldt, H. H. & Ussery, D. W. (2000) *J. Mol. Biol.* **299,** 907–930.