ELSEVIER

Journal of Biotechnology 95 (2002) 25–38

JOURNAL OF Biotechnology

www.elsevier.com/locate/jbiotec

Strategy to sequence the genome of *Corynebacterium glutamicum* ATCC 13032: use of a cosmid and a bacterial artificial chromosome library

Andreas Tauch [a,*], Iris Homann [b], Sascha Mormann [c], Silvia Rüberg [a], Alain Billault [d], Brigitte Bathe [c,f], Sven Brand [c], Olaf Brockmann-Gretza [c], Christian Rückert [c], Natalie Schischka [c,f], Carsten Wrenger [c], Jörg Hoheisel [e], Bettina Möckel [f,1], Klaus Huthmacher [f], Walter Pfefferle [f], Alfred Pühler [c], Jörn Kalinowski [a]

[a] *Zentrum für Genomforschung, Universität Bielefeld, Universitätsstraße 25, D-33615 Bielefeld, Germany*
[b] *Institut für Innovationstransfer an der Universität Bielefeld GmbH, Geschäftsbereich BioTech, Universitätsstraße 25, D-33615 Bielefeld, Germany*
[c] *Lehrstuhl für Genetik, Universität Bielefeld, Universitätsstraße 25, D-33615 Bielefeld, Germany*
[d] *Molecular Engines Laboratories, 20 Rue Bouvier, F-75011 Paris, France*
[e] *Deutsches Krebsforschungszentrum, Abteilung Funktionelle Genomanalyse, Im Neuenheimer Feld 506, D-69120 Heidelberg, Germany*
[f] *Degussa AG, Kantstraße 2, D-33788 Halle-Kunsebeck, Germany*

Received 23 October 2001; received in revised form 14 November 2001; accepted 25 November 2001

Abstract

The initial strategy of the *Corynebacterium glutamicum* genome project was to sequence overlapping inserts of an ordered cosmid library. High-density colony grids of approximately 28 genome equivalents were used for the identification of overlapping clones by Southern hybridization. Altogether 18 contiguous genomic segments comprising 95 overlapping cosmids were assembled. Systematic shotgun sequencing of the assembled cosmid set revealed that only 2.84 Mb (86.6%) of the *C. glutamicum* genome were represented by the cosmid library. To obtain a complete genome coverage, a bacterial artificial chromosome (BAC) library of the *C. glutamicum* chromosome was constructed in pBeloBAC11 and used for genome mapping. The BAC library consists of 3168 BACs and represents a theoretical 63-fold coverage of the *C. glutamicum* genome (3.28 Mb). Southern screening of 2304 BAC clones with PCR-amplified chromosomal markers and subsequent insert terminal sequencing allowed the identification of 119 BACs covering the entire chromosome of *C. glutamicum*. The minimal set representing a 100% genome coverage contains 44 unique BAC clones with an average overlap of 22 kb. A total of 21 BACs represented linking clones between

* Corresponding author. Fax: +49-521-1065626.
E-mail address: andreas.tauch@genetik.uni-bielefeld.de (A. Tauch).
[1] Present address: Qiagen GmbH, Max-Volmer-Straße 4, D-40724 Hilden, Germany.

PII: S0168-1656(01)00443-6

previously sequenced cosmid contigs and provided a valuable tool for completing the genome sequence of *C. glutamicum*.

Keywords: *Corynebacterium glutamicum*; Bacterial artificial chromosome library; Genome mapping; Genome sequence

1. Introduction

Corynebacterium glutamicum is a gram-positive soil micro-organism which is most widely used for the fermentative production of amino acids, e.g. L-glutamate and L-lysine (Leuchtenberger, 1996). Currently, about 1×10^6 tons of L-glutamate and 4.5×10^5 tons of L-lysine are produced annually with this organism. Fermentation processes using *C. glutamicum* strains have been successfully commercialized not only for L-glutamate and L-lysine but also for the production of other economically important amino acids (Leuchtenberger, 1996). Due to the importance of *C. glutamicum* as a biotechnological amino acid producer, extensive research has been carried out on strain improvement for amino acid fermentation. In early studies, the industrial amino acid production with *C. glutamicum* was primarily improved by using the iterative procedure of random mutagenesis and subsequent selection of enhanced production strains. With the advent of recombinant DNA techniques for *C. glutamicum*, strain improvement mainly focussed on rational metabolic engineering which was based on the molecular genetic characterization of single genes or gene clusters of the respective biosynthesis pathways (reviewed in Sahm et al. (1995), Eggeling et al. (1997)). Nowadays, the application of genomics and post-genomics is a promising approach for the construction of amino acid-producing *C. glutamicum* strains (Hodgson, 1998). Therefore, a sequencing project with the goal to determine the entire chromosomal sequence of the wildtype strain *C. glutamicum* ATCC 13032 was initiated (Hodgson, 1998).

Two general strategies for generating whole-genome sequences have frequently been used (Frangeul et al., 1999). The first strategy is a direct random-shotgun sequencing of the organism of choice, which does not require any preliminary data, such as a physical map, before starting the sequencing phase. The second strategy is the ordered-clone approach, which uses large-insert libraries to establish a contiguous set of overlapping clones covering the entire genome. Small-insert libraries of the assembled clones are then sequenced to obtain the whole-genome sequence. Large DNA fragments for genome mapping are usually cloned into cosmids or bacterial artificial chromosomes (BACs). BAC vectors contain the *Escherichia coli* F factor replication system which limits the number of BACs to one or a maximum of two copies per cell (Shizuya et al., 1992; Wang et al., 1997). This strict copy number control ensures the stable maintenance of large DNA inserts of up to 300 kb and avoids the lethal overexpression of cloned genes in the *E. coli* host cell.

As a preparatory work of the *C. glutamicum* ATCC 13032 sequencing project, total genomic digests with the meganuclease *Swa*I were separated by pulsed-field gel electrophoresis (PFGE) and the resulting macrorestriction fragments were physically linked by hybridization with cloned gene probes (Bathe et al., 1996). By summing up the lengths of the macrorestriction fragments, a genome size of 3082 ± 20 kb was determined. The initial strategy of the genome project was to order a cosmid library of *C. glutamicum* ATCC 13032 by hybridization mapping and to sequence small-insert libraries of overlapping cosmid clones. In the present study, we describe the assembly of a *C. glutamicum* cosmid library into contiguous genomic segments. In addition, a second large-insert library was constructed in the BAC vector pBeloBAC11 and used for genome mapping. The BAC library covered the entire *C. glutamicum* genome and provided a valuable tool for determining the complete genome sequence of this biotechnologically important bacterium.

2. Materials and methods

2.1. Bacterial strains and growth conditions

C. glutamicum ATCC 13032 was obtained from the American Type Culture Collection (Manassas, VA). *E. coli* DH10B (Grant et al., 1990) was used as host strain for BAC cloning. The BAC vector pBeloBAC11 (Wang et al., 1997) was used for BAC library construction. Cultures of *E. coli* and *C. glutamicum* were routinely grown in LB medium (Sambrook et al., 1989) at 37 and 30 °C, respectively. *E. coli* NM554 carrying recombinant cosmid clones was selected on LB agar supplemented with 50 μg ml^{-1} kanamycin. BAC recombinant clones were selected on LB agar containing 12.5 μg ml^{-1} chloramphenicol.

2.2. Cosmid library screening and cosmid sequencing

Cosmid colony filters were robotically arrayed from the cosmid library master plates (Bathe et al., 1996) essentially as described by Hoheisel et al. (1993). Colony hybridization followed standard procedures (Sambrook et al., 1989). Cosmid DNA was isolated from *E. coli* NM554 with the QIAprep Spin Miniprep Kit (Qiagen, Hilden, Germany). Terminal probes for Southern hybridization were generated on *Eco*RI-digested cosmid DNA by means of a single-strand labeling PCR with oligonucleotides corresponding to the T3 and T7 priming sites of the cosmid vector SuperCos I, *Taq* DNA Polymerase (Qiagen) and DIG-11-dUTP (Roche Diagnostics, Mannheim, Germany). Additional genetic markers for Southern hybridization were designed from genomic sequences with the Primer3 program (Whitehead Institute, Cambridge, MA). Labeled probes were amplified from chromosomal DNA of *C. glutamicum* ATCC 13032 (Tauch et al., 1995) with the *Taq* DNA polymerase and a PCT-100 thermocycler (MJ Research, Watertown, MA) using DIG-11-dUTP. The PCR conditions were: Initial denaturation at 94 °C for 30 s followed by 30 s denaturation, 30 s annealing at 55 °C, and extension at 72 °C for 2 min. This cycle was repeated 30 times followed by a final extension step for 2 min at 72 °C. Southern hybridization was performed with the DIG DNA labeling and detection kit, CSPD as chemiluminescent alkaline phosphatase substrate (Roche Diagnostics) and Cronex 4 films (Sterling Diagnostic, Newark, DE).

Selected cosmids from the *C. glutamicum* library were sequenced by LION bioscience AG (Heidelberg, Germany), MWG-Biotech AG (Ebersberg, Germany) and Qiagen GmbH (Hilden, Germany).

2.3. Preparation of high-molecular-weight DNA and BAC library construction

To prepare high-molecular-weight genomic DNA from *C. glutamicum* ATCC 13032, a saturated overnight culture grown in Brain Heart Infusion medium (Merck Eurolab, Darmstadt, Germany) was treated with 10 mg ml^{-1} lysozyme for 1 h at 37 °C. The cells were harvested and the pellet was used to prepare 100 μl DNA plugs embedded in low-melting agarose at a concentration of ≈10 μg per plug. The plugs were incubated twice in buffer containing 1.25 mg ml^{-1} proteinase K, 0.5 M EDTA (pH 8.5) and 1% (w/v) *N*-lauroyl sarcosine at 55 °C for 24 h. After inactivation of the proteinase K in 40 μg ml^{-1} PMSF for 30 min, plugs were rinsed twice in TE buffer (10 mM Tris, 1 mM EDTA) for 30 min and stored in 0.5 M EDTA (pH 8.0) at 4 °C. Preparation of pBeloBAC11 for library construction was carried out as described (Woo et al., 1994).

Partial digestion was carried out in agarose plugs, each containing ≈10 μg of high-molecular-weight DNA, after three 1-h equilibration steps in 10 ml of 1 × *Hin*dIII digestion buffer (Roche Diagnostics). The buffer was then removed and replaced by ice-cold enzyme buffer (1 ml per plug) containing 5 U of *Hin*dIII. After 2 h of incubation on ice, the plugs were transferred to a 37 °C water bath for 15 min. Digestions were stopped by adding 10 ml of 0.25 mM EDTA (pH 8.0). The partially digested DNA was subjected to contour-clamped homogeneous electric field (CHEF) electrophoresis on a 1% low-melting point agarose gel, using a DR III apparatus (Biorad, Hercules,

CA) in 1 × Tris–acetate–EDTA buffer at 12 °C, with a ramp from 5 to 15 s at 6 V cm^{-1} during 16 h. Agarose slices corresponding to size ranges from 50 to 75 kb and 75 to 100 kb were excised from the gel and subjected to a second size selection with a constant pulse of 5 s at 4 V cm^{-1} during 10 h at 15 °C. The size ranges were excised from the gel and stored in TE at 4 °C before use.

Agarose slices containing the genomic size fractions were melted at 65 °C for 10 min and digested with Gelase (Epicentre Technologies, Madison, WI) using 1 U per 100 mg of gel slice. Then 50–100 μg of each size-selected DNA were ligated in a molar ratio of 1:5 to 1:10 with *Hin*dIII-digested, dephosphorylated pBeloBAC11, using 10 U of T4 DNA ligase (New England Biolabs, Beverly, MA) at 12 °C for 36 h. Ligation mixtures were heated at 65 °C for 15 min and then drop-dialysed against 0.5 × TE (pH 8.0), using VS 0.025 μM membranes (Millipore, Bedford, MA).

Fresh electrocompetent *E. coli* DH10B cells (Sheng et al., 1995) were harvested from 200 ml of a mid-log culture grown in SOB medium. Cells were washed three times in ice-cold water and finally resuspended in ice-cold water to a cell density of 10^{11} cells ml^{-1}. One to two microliter of the ligation mix was used to transform 30 μl of electrocompetent *E. coli* DH10B cells in an Easyject Puls electroporator (Equibio, Ashford, UK) with electrical settings of 2.5 kV, 25 μF, and 99 Ω in 2-mm-wide electroporation cuvettes. After electrotransformation, cells were resuspended in 1 ml of SOC medium and allowed to recover for 45 min at 37 °C with gentle shaking. Selection of *E. coli* clones was performed on LB agar containing 12.5 μg ml^{-1} chloramphenicol, 50 μg ml^{-1} 5-bromo-4-chloro-3-indolyl β-D-galactoside (X-Gal), and 25 μg ml^{-1} isopropyl-β-D thiogalactopyranoside (IPTG). The plates were incubated for 20 h at 37 °C and recombinant clones (white colonies) were picked automatically into 96-well flatbottom microtiter plates (Greiner Bio-One, Solingen, Germany) by means of the Flexys robot system (Genomic Solutions, Ann Arbor, MI). Each colony was inoculated in a mixture of 170 μl selective LB medium and 17 μl 10 × freeze medium (13.2 mM KH_2PO_4, 36 mM K_2HPO_4, 0.4 mM $MgSO_4$, 6.8 mM $(NH_4)_2SO_4$, 1.7 mM tri-sodium citrate, 4.4% (v/v) glycerol) and incubated overnight at 37 °C. The library master plates were stored at −80 °C.

2.4. BAC library screening and BAC colony hybridization

Colony filters were prepared from the library master plates by the following procedure: Frozen master plates of the BAC library were incubated at room temperature for 20 min and BAC clones were then inoculated automatically (Flexys; Genomic Solutions) into fresh 96-well flatbottom microtiter plates containing 170 μl LB and 12.5 μg ml^{-1} chloramphenicol. The clones were grown overnight with moderate shaking at 37 °C. Then cell suspension from each well was transferred on Porablot Ny Amp membranes (Macherey-Nagel, Düren, Germany) covering Omnitray plates (Nalge Nunc, Rochester, NY) which were filled with selective LB medium. The cell transfer was performed with the Flexys gridding system (Genomic Solutions) in such a way that each BAC filter contains a doublet of 768 colonies. The colony filters were incubated for 16 h at 37 °C.

Colony hybridization followed the standard procedure described by Sambrook et al. (1989) with the exception that the BAC DNA was crosslinked with an energy exposure of 70 000 μJ cm^{-2} by means of the Hoefer UV crosslinker (Amersham Pharmacia Biotech, Freiburg, Germany) before removal of the cell debris. Genetic markers for Southern hybridization were designed from cosmid sequences and from BAC terminal sequences with the Primer3 program (Whitehead Institute).

2.5. BAC vector preparation and BAC end sequencing

Each BAC clone was inoculated in 50 ml LB medium supplemented with 12.5 μg ml^{-1} chloramphenicol for 16 h at 37 °C. The bacterial culture was harvested by centrifugation for 15 min at 5000 *g* and 4 °C. DNA preparation was performed by means of an alkaline-SDS column purification procedure based on the Nucleobond

AX100 cartridge and folded filter system (Macherey-Nagel). The BAC DNA was digested with *Not*I and insert sizes were determined by PFGE on a 1% agarose gel (Bathe et al., 1996). MidRange and LowRange PFG markers (New England Biolabs) were used a size standards.

Purified BAC templates were sequenced by the dye-terminator sequencing method on a Prism ABI 377 DNA sequencer using the Prism Ready Reaction Dye Deoxy Termination Kit (Applied Biosystems, Foster City, CA). BAC terminal sequences were checked against a *C. glutamicum* genome database with the BLAST-2N algorithm (Altschul et al., 1997) to find a match. Linkage between sequence contigs was achieved when two BAC ends had a single match within each contig.

2.6. Size determination of sequence gaps by PCR

To determine the size of gaps between sequenced contigs, PCR experiments were performed with eLONGase (Life Technologies, Karlsruhe, Germany) and a PCT-100 thermocycler (MJ Research) using BAC DNA and *C. glutamicum* chromosomal DNA as template. Primer pairs were designed with the Primer3 program (Whitehead Institute) using cosmid contig ends and BAC terminal sequences, respectively. The PCR conditions were: initial denaturation at 94 °C for 30 s followed by 20 s denaturation at 94 °C, 30 s annealing at 49 °C, and extension at 68 °C for 12 min. This cycle was repeated nine times followed by 25 cycles with an increasing extension time of 10 s per cycle and final extension for 20 min at 68 °C. Size determination of the amplified products was performed on 0.8% agarose gels. Primer pairs used for the amplification of previously sequenced genes from the *C. glutamicum* chromosome were: proP1 5′-CGCTACGCAATGACTTTCAA-3′, proP2 5′-TGATTGGATAGGCTACCTCG-3′, cglIM1 5′-CCAGGACTTCTCCATGATCT-3′, and cglIM2 5′-ACAGGAGGAACAGCATTACC-3′. Primer pairs for *rel* and PCR conditions were as described (Wehmeier et al., 1998), with the exception that DIG-11-dUTP (Roche Diagnostics) was added to the PCR assay.

3. Results

3.1. Cosmid contig assembly of the C. glutamicum chromosome

The initial strategy in the *C. glutamicum* genome project was to utilize a cosmid library for genome mapping (Bathe et al., 1996) and to sequence overlapping cosmid clones deduced from an ordered library (Hodgson, 1998). The screening technique of Southern hybridization permits a simple identification of homologous stretches of DNA within a genomic library and thus allows the rapid identification of overlaps between individual library clones. In addition, a large number of library clones can be analyzed and ordered in parallel by high-density hybridization. Therefore, a total of 2304 genomic cosmid clones from the *C. glutamicum* library were robotically arrayed in high-density grids on nylon membranes (Hoheisel et al., 1993) for Southern screening. A size determination of a sample of randomly chosen cosmids by *Eco*RI digestion and subsequent agarose gel electrophoresis indicated an average insert size of 39.5 kb. Accordingly, the arrayed cosmid library represents a theoretical 28-fold coverage of the *C. glutamicum* genome.

For ordering of the *C. glutamicum* cosmid library, various hybridization procedures were applied. In the first round of hybridizations, eight cosmid clones previously shown to link adjacent *Swa*I macrorestriction fragments of the *C. glutamicum* genome (Bathe et al., 1996) were used for library screening (Fig. 1A). These cosmids were mapped at different positions around the *C. glutamicum* chromosome and thus represented ideal starting points for a contiguous cosmid assembly. Single-strand terminal labeling of the linking cosmids was carried out by a PCR technique with oligonucleotide primers corresponding to the T3 and T7 priming sites of the cosmid vector SuperCos I. In order to limit the length of the labeled cosmid termini and thus to reduce interference from any possible repetitive genomic sequence, DNA of linking cosmids was digested with *Eco*RI before performing the labeling PCR assay. The labeled cosmid probes were then hybridized against the library filters to identify overlapping

clones. *Eco*RI digests of DNAs from positive cosmids were separated on agarose gels and back-hybridized with a completely labeled linking cosmid for confirmation. In addition, the size of the overlapping cosmid portions was calculated by back-hybridization to eliminate redundant cos-

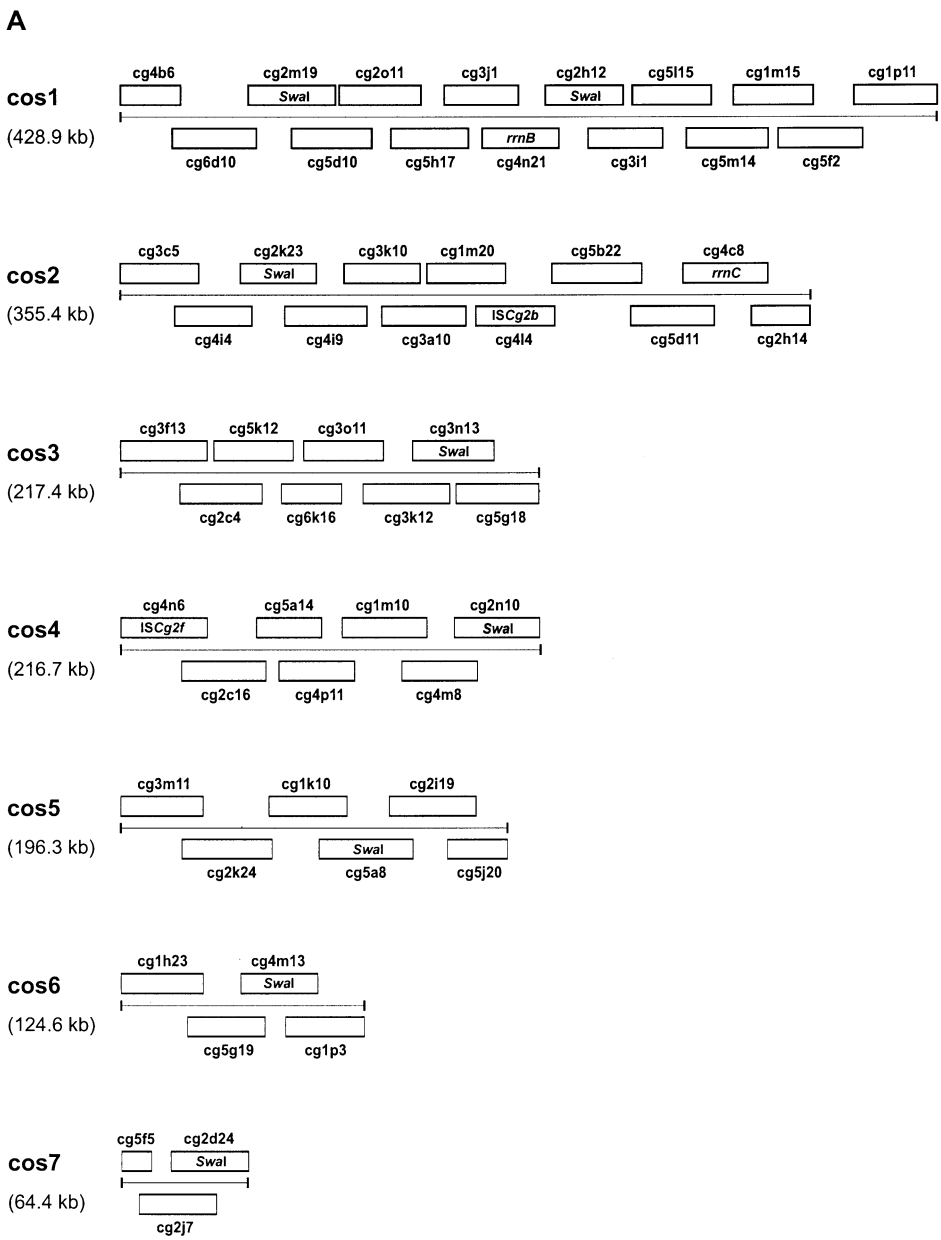

Fig. 1. Assembled cosmid contigs of the *C. glutamicum* chromosome. (A) *Swa*I linking clones (*Swa*I) were used for initial assembly of contigs cos1 to cos7. The order of the overlapping cosmid clone set is shown, clone names are indicated. The position of repetitive elements is marked. The nucleotide sequence length of the contigs is shown at the right. (B) Contigs assembled around repetitive elements of the *C. glutamicum* chromosome (IS*Cg1*, IS*Cg2*, *rrn*). (C) Contigs identified by hybridization with the specific gene probes *rplK*, *leuB* and *oriC*.

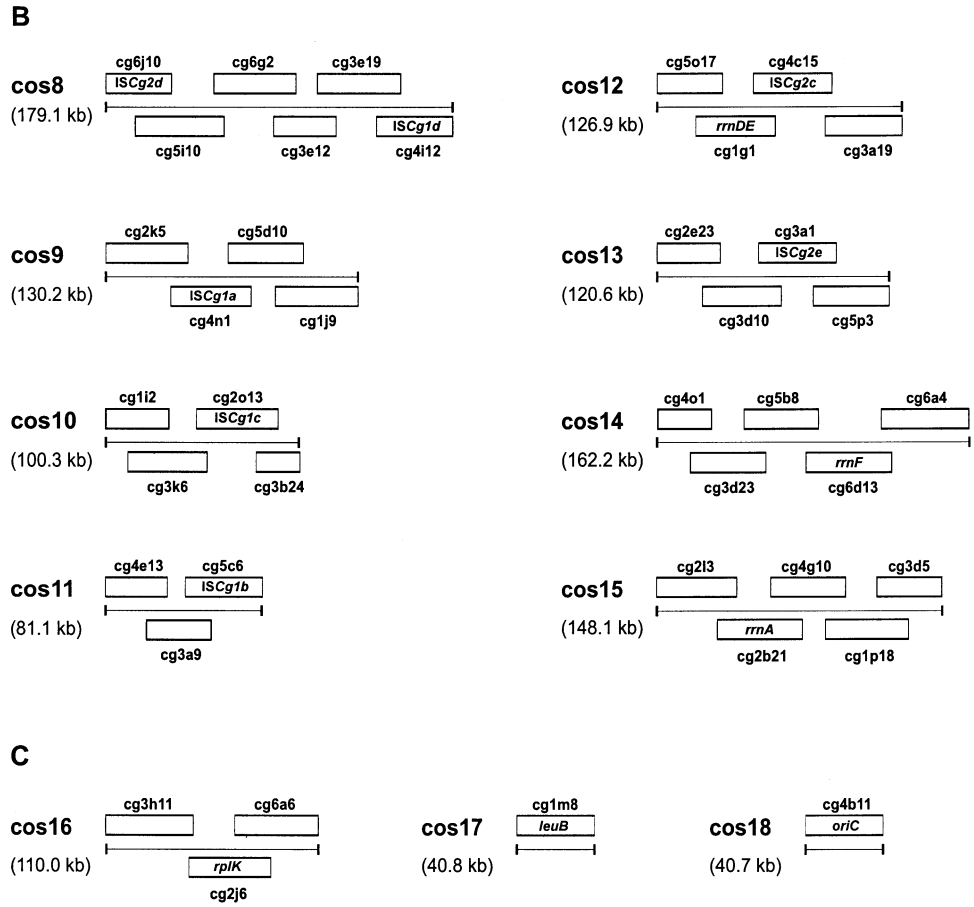

Fig. 1. (*Continued*)

mids from the contig assembly. This screening procedure allowed the identification of neighboring cosmids on each side of the labeled genomic segment and, when applied in an iterative manner, the contiguous assembly of minimally overlapping cosmid clones from the arrayed library. The Southern screening resulted in the assembly of 55 overlapping cosmids into seven contiguous genomic segments of *C. glutamicum*, designated cos1 to cos7 (Fig. 1A).

To avoid any potential cross-hybridization problem caused by repetitive elements of the *C. glutamicum* chromosome, such as *rrn* genes and insertion sequences, the assembled cosmid set of contigs cos1 to cos7 and the complete cosmid library were hybridized with labeled *rrn* probes and with labeled markers deduced from the insertion sequences IS*Cg1* and IS*Cg2* (Bathe et al., 1996; Quast et al., 1999). These Southern hybridizations identified cosmid clones with repetitive sequences, which were further analyzed by a fingerprint technique on digested cosmid DNA. Since each copy of the repetitive sequences is located on a chromosomal restriction fragment of defined size, each individual copy can be identified unequivocally by its unique hybridization pattern (Bathe et al., 1996; Quast et al., 1999; Tauch et al., 2001). Within the assembled contigs cos1 to cos7 only two of six *rrn* clusters, two of five IS*Cg2* elements and none of the four IS*Cg1* elements were detected (Fig. 1A). Therefore, cosmids carrying individual copies of the repetitive elements not present in the initially assembled contigs were used as new starting points for library screening

(Fig. 1B). This second iterative library screening resulted in the assembly of 35 overlapping cosmids into eight contiguous genomic segments of *C. glutamicum*, designated cos8 to cos15 (Fig. 1B). In addition, cosmids which were previously shown to hybridize to known gene probes such as *rplK*, *leuB* and *oriC* (Bathe et al., 1996) but which were not identified during assembly of cos1 to cos15 were used in a final library screening. This approach allowed the assembly of five overlapping cosmids into three additional contiguous genomic segments of *C. glutamicum*, designated cos16 to cos18 (Fig. 1C).

Since no further contig extensions were identified by cosmid library screening, the assembled set of 95 cosmids was subjected to systematic shotgun sequencing. DNA sequence assembly generated contigs ranging in size from 40.7 to 428.9 kb (Fig. 1) with an average cosmid overlap of 11.6 kb. A summation of the contig lengths revealed that only 2.84 Mb (86.6%) of the *C. glutamicum* genome, were represented by the cosmid library. Therefore, an alternative mapping strategy was necessary to complete the *C. glutamicum* genome sequence.

3.2. Construction of a BAC library of the C. glutamicum chromosome

To obtain a complete coverage of the *C. glutamicum* chromosome and to determine the nucleotide sequences of the remaining gaps, we have constructed and subsequently mapped a genomic BAC library by using the vector pBeloBAC11 and partial *Hin*dIII DNA fragments. Since no experimental data were available on the maximum insert size of *C. glutamicum* DNA that can be cloned efficiently in *E. coli* DH10B, two different BAC libraries were constructed containing genomic size fractions from 50 to 75 kb (library I) and from 75 to 100 kb (library II). Cloning of both genomic fractions into pBeloBAC11 and electrotransfer into *E. coli* DH10B resulted in $\approx 4 \times 10^3$ transformants (white colonies) in each case. BAC library II was used throughout this study and was further characterized for the presence of inserts. DNA preparations from 82 randomly selected white colonies and subsequent agarose gel electrophoresis revealed that 67% of the clones carried an insert of the appropriate size. A more precise size determination of 24 insert-carrying clones was performed by PFGE after cleavage with *Not*I. This restriction analysis indicated an average insert size of 97 kb with a minimum insert size of 77 kb and a maximum insert size of 112 kb. Finally, a total of 3168 white clones from the *C. glutamicum* BAC library was automatically picked into 96-well micotiter plates, representing a theoretical 63-fold coverage of the *C. glutamicum* genome.

3.3. Identification of BAC clones for sequence gap closure

The *C. glutamicum* BAC library II was used for the identification of BAC clones covering each of the 18 sequence gaps between the cosmid contigs. In a first screening phase, 35 BACs were randomly subjected to insert terminal sequencing. The resulting sequences were checked against the *C. glutamicum* genome sequence database with the BLAST-2N algorithm (Altschul et al., 1997). A linkage between two contigs was confirmed when both BAC end sequences revealed a single match within each contig. By means of this random sequencing approach eight BACs representing linking clones between sequenced cosmid contigs were identified (Fig. 4).

In a second screening phase, 2304 BACs corresponding to a theoretical 46-fold coverage of the *C. glutamicum* genome were applied for contig gap closure by a Southern screening technique (Fig. 2). For this purpose, three subsets of BAC colony filters, each one consisting of a doublet of 768 BAC clones, were automatically prepared by the Flexys robot system on nylon membranes. Labeled PCR products corresponding to the non-linked contig ends were amplified from chromosomal *C. glutamicum* DNA and subsequently used for Southern hybridization. On an average 34 hybridization signals were obtained with each of the labeled DNA probes, which is slightly lower than the theoretical genome coverage of the BAC library. A total of 43 BACs showing positive hybridization signals to two contig ends (linking BACs) was selected for terminal insert sequencing

(Fig. 2). The end sequences of these BACs were checked against the *C. glutamicum* genome sequence, which confirmed the predicted contig linkage deduced from the Southern blot. Altogether, the Southern screening technique identified nine groups of BACs within the *C. glutamicum* library which physically link nine additional cosmid contigs (Fig. 4).

The final gap between the contigs cos5 and cos3 was not closed by the systematic Southern strategy, suggesting that this gap may be too large to be covered by a single BAC clone. Therefore, the gap was closed by further rounds of Southern hybridization and PCR screening (Fig. 3). Initial hybridization with labeled DNA probes from contig ends of cos5 and cos3 identified a set of BAC clones, which produced contig matches at only one end, e.g. cgb07g09 and cgb08a04. Labeled PCR products derived from the non-matching ends of these BACs were used for an additional library screening. Colony hybridization with both DNA probes detected only 11 BACs indicating that this genomic region of *C. glutamicum* is underrepresented in the BAC library. Subsequently, PCR experiments were performed on the identified BACs with primer pairs designed from their terminal sequences. Overlaps were confirmed when the same primer pair amplified identical PCR products on different BAC templates. This PCR approach revealed that the BACs cgb15f07 and cgb08f02 have overlapping ends and that the gap between cos5 and cos3 could be closed by four overlapping BAC clones (Fig. 3). Thus, a total number of 21 BACs covered DNA regions of the *C. glutamicum* genome which were not represented by the cosmid library. This set of linking BACs also allowed the arrangement of the 18 assembled cosmid contigs in a circular physical map of the *C. glutamicum* chromosome (Fig. 4).

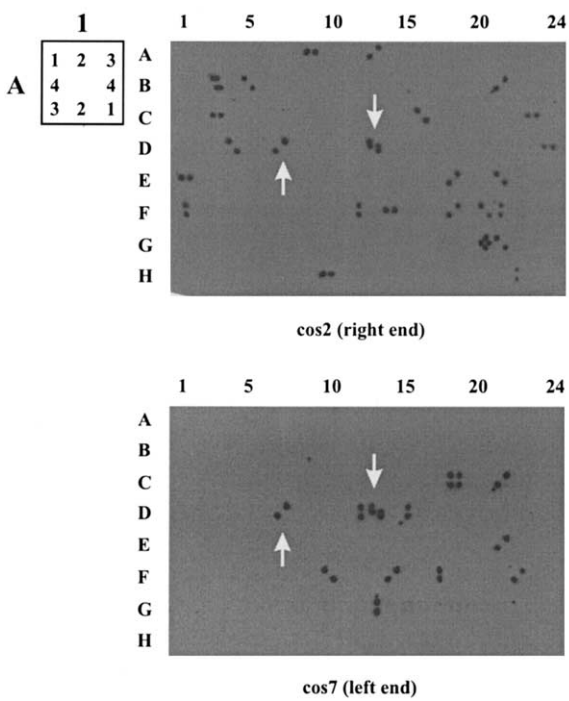

Fig. 2. Southern screening of the *C. glutamicum* BAC library. Two Cronex4 films showing chemiluminescence signals of BAC colony filters are shown. The filters were hybridized with DIG-labeled DNA probes designed from cosmid contig ends of cos2 (right end) and cos7 (left end). BACs showing positive signals with both DNA probes are marked by arrows. Each colony filter carries a doublet of 768 BAC clones. The scheme of clone arrangement of four different BACs (1–4) in each square of the filter is shown in upper left panel. The clone duplication produces two positive signals per BAC in diagonal (1 and 3), vertical (2) or horizontal (4) position.

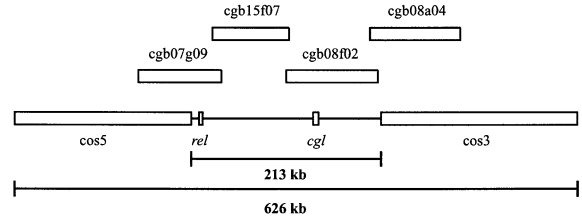

Fig. 3. Physical map of the gap between cosmid contigs cos5 and cos3. Initial library screening by Southern hybridization identified BACs whose terminal sequences matched only one contig (cgb07g09 and cgb08a04). A further round of PCR-amplified products from the non-matching BAC ends hybridized with two new BACs (cgb15f07 and cgb08f02). A third round of PCR deduced from their terminal sequences indicated that both BACs have overlapping ends. In addition, the position of the *rel* and *cgl* gene region of the *C. glutamicum* chromosome within the gap is indicated. The gap size given below the map was deduced from long-range PCR experiments.

3.4. Investigating the sequence gaps

Size determination of the linking BACs by PFGE and additional PCR experiments showed

that the sequence gaps between cosmid contigs varied from 220 bp to ≈48.5 kb in size (Fig. 4), with an average gap size of 13.1 kb. The large gap between cos5 and cos3 was further investigated by a set of long-range PCR experiments with primer pairs deduced from neighbouring contig ends and a gap size of 213 kb was calculated (Fig. 3). A summation of the 18 gap sizes revealed that 436

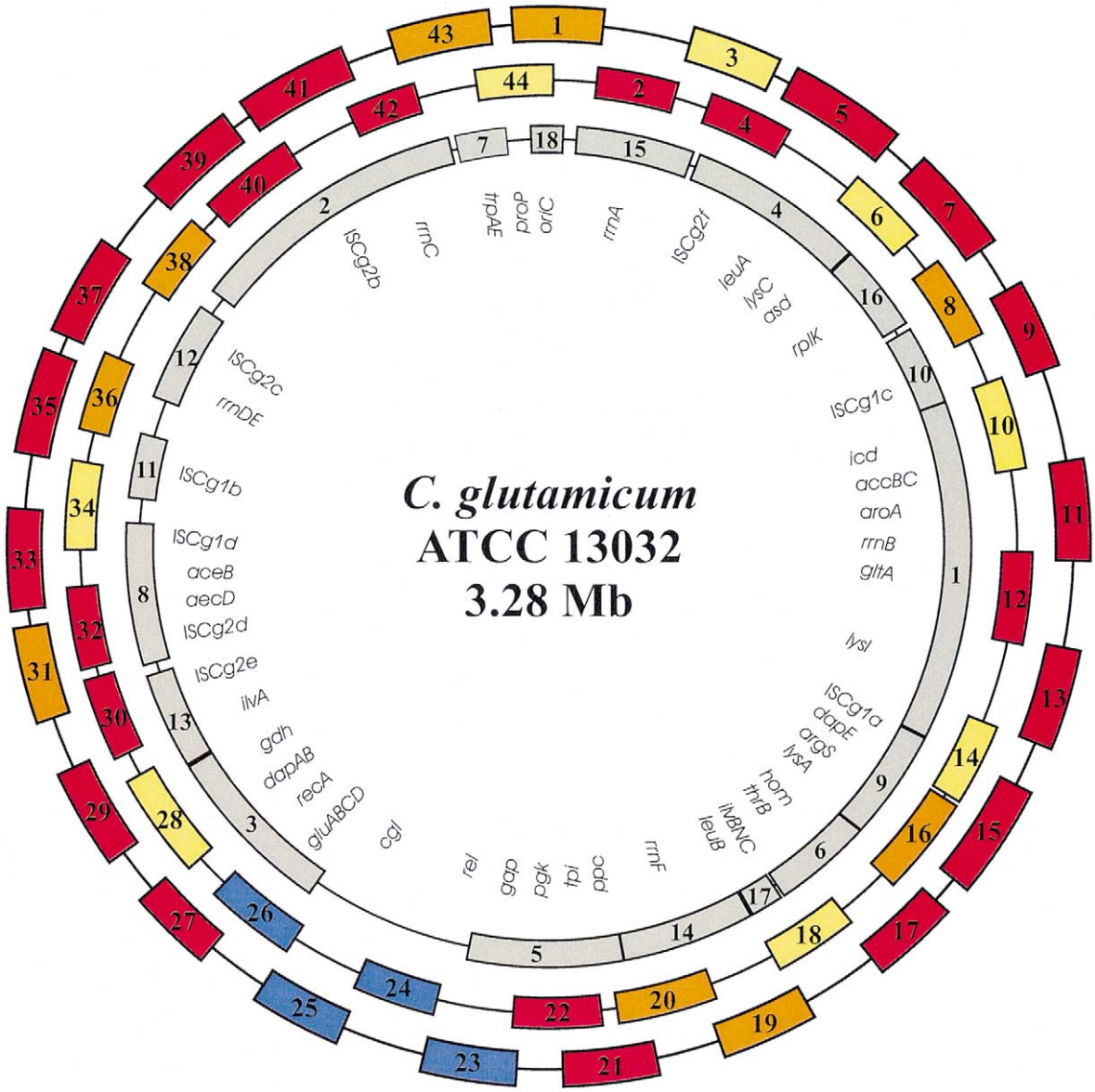

Fig. 4. Physical and genetic map of the *C. glutamicum* ATCC 13032 chromosome deduced from an ordered BAC library. The position of the initially sequenced cosmid contigs (cos1 to cos17) is shown in the inner circle. The outer circles represent the minimal set of 44 BAC clones (cgb1 to cgb44) covering the entire *C. glutamicum* chromosome. Linking BACs identified by random-sequencing (yellow) and Southern screening (orange) of the BAC library are marked. BACs covering the gap between cos5 and cos3 are shown in blue. BAC clones identified by Southern hybridization to obtain a 100% coverage of the *C. glutamicum* chromosome are shown in red color. The approximate position of a selected set of genes and repetitive elements is indicated.

Table 1
Mapping of previously sequenced genes not present in the cosmid contigs of the *C. glutamicum* ATCC 13032 chromosome

Gene	Product/proposed function	Genome position between contigs	GenBank No.
cglIM	5-cytosine methyltransferase	cos5–cos3	U13922
cglIR	Restriction endonuclease CglI	cos5–cos3	U13922
cglIIR	Restriction endonuclease CglII	cos5–cos3	U13922
rel	(p)ppGpp synthetase, (p)ppGpp degrading enzyme	cos5–cos3	AF038651
apt	Adenine phosphoribosyltransferase	cos5–cos3	AF038651
proP	Proline/ectoine uptake	cos7–cos18	Y12537

kb remained to be sequenced on the 21 linking BACs. Therefore, this set of BACs was selected to complete the genome sequence of *C. glutamicum* ATCC 13032 by BAC shotgun sequencing and primer walking strategies.

To obtain genetic information regarding the gaps present between the cosmid contigs, the GenBank database was searched for *C. glutamicum* ATCC 13032 nucleotide sequences, which were not identified in the course of the cosmid sequencing project. The deduced list of database entries is shown in Table 1. Labeled PCR probes of the respective gene regions were generated and hybridized against the *C. glutamicum* BAC library as well as against a new colony filter set carrying only linking BACs. The previously sequenced *proP* gene encoding a proline/ectoine uptake system (Peter et al., 1998) was identified in the gap between contigs cos7 and cos18 (Fig. 4). In addition, the *rel* gene, which is involved in (p)ppGpp metabolism and in the stringent response of *C. glutamicum* (Wehmeier et al., 1998), and the restriction–modification gene cluster *cglIM–cglIR–cglIIR* (Schäfer et al., 1997) were mapped within the gap between cos5 and cos3 (Fig. 4). Long-range PCR experiments allowed a more detailed localization of both gene regions within this large gap (Fig. 3).

3.5. Construction of a circular BAC contig of the *C. glutamicum* chromosome

To obtain a 100% coverage of the *C. glutamicum* genome with BAC clones, further labeled DNA probes localized within the previously sequenced regions of the chromosome were amplified for Southern screening of BAC colony filters. The terminal sequences of 41 BACs were determined and the position of these clones on the chromosome was identified by BLAST-2N searches against the *C. glutamicum* genome sequence. By using the Southern screening technique, a circular BAC contig of 119 clones covering the entire *C. glutamicum* chromosome was constructed. The 119 BAC clones have a total insert size of ≈ 11.5 Mb, which is equivalent to a 3.5-fold genome coverage. The *C. glutamicum* chromosome could be represented by a minimal set of 44 BACs (cgb1 to cgb44) with an average overlap of 22 kb. The resulting physical map of the *C. glutamicum* ATCC 13032 chromosome with an alignment of the minimal set of BAC clones is shown in Fig. 4. By summing-up the already sequenced portion of the *C. glutamicum* genome and the sizes of the identified gaps between the cosmid contigs, a genome size of 3.28 Mb was determined.

4. Discussion

The knowledge of the entire genome sequence of an industrially important micro-organism provides a wealth of information that can be used as starting point for the systematic functional analysis and for metabolic engineering (Hodgson, 1998). Two experimental strategies have been used thus far to determine whole-genome sequences: a random-sequencing approach, which is based on whole-genome shotgun sequencing, and a more complex approach comprising the construction of a contiguous set of recombinant clones as a prerequisite for ordered-clone sequencing (Frangeul et al., 1999). The major technical difficulty of the

ordered-clone approach is to obtain gene libraries with overlapping clones covering the complete genome. Especially cosmid libraries have played a crucial role in the genome sequencing projects of *Bacillus subtilis*, *Mycobacterium tuberculosis*, and *Sulfolobus solfataricus* (Kunst et al., 1997; Brosch et al., 1998; She et al., 2000). The finding that some regions of the respective chromosomes were apparently not represented in cosmid libraries has complicated the production of a contiguous set of clones covering the entire genome of these organisms. A cosmid library carrying chromosomal DNA of *Streptomyces coelicolor* represented nearly the entire genome with the exception of three short gaps (Redenbach et al., 1996) whereas a cosmid library of the *S. solfataricus* genome covered only 70% of the chromosome despite a high theoretical level of coverage (She et al., 2000). Likewise, the cosmid library of the *C. glutamicum* chromosome was very biased such that 18 regions of the genome with a total size of 436 kb (13.3%) were not covered, despite a 28-fold genome coverage of the library. Nowadays, this limitation can be overcome by BACs which are the most suitable cloning system for genome mapping and genome sequencing (Monaco and Larin, 1994; Frangeul et al., 1999). The main advantages of BACs over cosmid vectors are the high level of clone stability and the low copy number which reduces lethal effects of cloned genes in the *E. coli* host cell (Shizuya et al., 1992; Wang et al., 1997).

In this study, a representative BAC library of the *C. glutamicum* chromosome was constructed with a genomic size fraction from 75 to 100 kb. Library screening was successfully performed by a Southern technique with labeled chromosomal markers amplified from cosmid contig ends. Southern mapping is based on the idea that when a unique labeled PCR product generates signals from two BAC clones they must overlap. BAC terminal sequences were determined and used for both BAC end mapping and further library screening. By means of this Southern strategy 17 out of 18 gaps between cosmid contigs were covered by BACs, while the last gap was not closed by one round of screening. Efficient genome mapping with BAC libraries has also been reported for *Sinorhizobium meliloti* (Capela et al., 1999) and *S. solfataricus* (She et al., 2000). In contrast to the Southern screening strategy described for *C. glutamicum*, a PCR approach with pools of BAC DNA was used for genome mapping in these studies. A more complex experimental approach using *Hin*dIII fingerprints of BAC clones has been described for physical mapping of the archaeon *Methanosarcina thermophila* (Diaz-Perez et al., 1997).

The largest gap identified between contigs cos5 and cos3 of *C. glutamicum* has a size of 213 kb. This region of the genome could be covered by four BACs and was found to be underrepresented in the *C. glutamicum* BAC library. Mapping of previously sequenced genes showed that *rel* and the *cgl* gene cluster are located in this genomic region. Although the gap is too large to directly link underrepresentation with these genes, the cloning of *rel* and *cgl* has been shown to cause severe problems in an *E. coli* host (Wehmeier et al., 1998; Schäfer et al., 1997). The *rel* gene product is involved in the stringent response of *C. glutamicum* and the metabolism of (p)ppGpp. Imbalances in the cellular (p)ppGpp levels due to the expression of *rel* were shown to result in growth inhibition of *E. coli*. In addition, cloning of the *cgl* gene cluster of *C. glutamicum*, encoding a restriction–modification system with two different types of restriction endonucleases, was problematic when using standard cloning systems in *E. coli* (Schäfer et al., 1997).

Nevertheless, we have established a circular BAC map of the entire chromosome of *C. glutamicum* ATCC 13032. The complete map revealed a genome size of 3.28 Mb, which is 200 kb larger than the value estimated by Bathe et al. (1996). In addition, differences between the genetic maps of *C. glutamicum*, which were deduced from PFGE mapping (Bathe et al., 1996) and from whole-genome BAC mapping during this study, became obvious when comparing the approximate positions of the mapped genes. The differences in genome size and in the deduced genetic organization of the *C. glutamicum* chromosome can be explained by an underestimation of the number of *Swa*I macrorestriction fragments (especially of the number of doublets)

which were used for genome size determination. Therefore, it is most likely that the *Swa*I macrorestriction map of the *C. glutamicum* chromosome is different to that published by Bathe et al. (1996). Approximately 436 kb of the genome remained to be sequenced and 21 BACs were selected for walking over small gaps and for preparing shotgun template libraries for larger ones. The ordered BAC library of *C. glutamicum* was thus a valuable tool for finishing the genome sequence and it may now represent a fundamental resource for future research projects on functional genomics. The identified minimal set of 44 BACs representing a 100% genome coverage may play an important role in gene expression studies. Since the BAC library represents 97-kb size fractions of the *C. glutamicum* genome, gene clusters may be located on one or two BACs rather than on several individual clones. The minimal BAC set may be used for functional genomics and complementation studies in the heterologous host *E. coli* to derive knowledge about the identified gene products with unknown function. BAC DNA arrays could be used in hybridization experiments with labeled cDNA probes prepared from total RNA thus identifying gene regulation (induction or repression) under different physiological conditions.

The results of the *C. glutamicum* genome project can also be used to deduce an optimal sequencing strategy for micro-organisms. Since random-shotgun sequencing is nowadays the dominant method for generating whole-genome sequences, initial sequencing with high efficiency should be performed with small-insert shotgun libraries from either genomic or BAC DNA without previous physical mapping of the genome. In addition, BAC libraries with a sufficient genome coverage should be constructed and a set of BAC clones should be sequenced from both ends to allow the physical mapping on the assembled shotgun sequences. When the sequencing efficiency of the shotgun phase is decreasing, BAC clones covering sequence gaps can be immediately identified to perform an efficient and rapid sequence finishing.

Acknowledgements

The authors thank Sabrina Scheideler and Eva Schulte-Berndt (University of Bielefeld) for technical assistance during BAC DNA preparation and BAC colony hybridization.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Bathe, B., Kalinowski, J., Pühler, A., 1996. A physical and genetic map of the *Corynebacterium glutamicum* ATCC 13032 chromosome. Mol. Gen. Genet. 252, 255–265.

Brosch, R., Gordon, S.V., Billault, A., Garnier, T., Eiglmeier, K., Soravito, C., Barrell, B.G., Cole, S.T., 1998. Use of *Mycobacterium tuberculosis* H37Rv bacterial artificial chromosome library for genome mapping, sequencing, and comparative genomics. Infect. Immun. 66, 2221–2229.

Capela, D., Barloy-Hubler, F., Gatius, M.-T., Gouzy, J., Galibert, F., 1999. A high-density physical map of *Sinorhizobium meliloti* 1021 chromosome derived from bacterial artificial chromosome library. Proc. Natl. Acad. Sci. USA 96, 9357–9362.

Diaz-Perez, S.V., Alatriste-Mondragon, F., Hernandez, R., Birren, B., Gunsalus, R.P., 1997. Bacterial artificial chromosome (BAC) library as a tool for physical mapping of the archaeon *Methanosarcina thermophila* TM-1. Microb. Comp. Genomics 2, 275–286.

Eggeling, L., Morbach, S., Sahm, H., 1997. The fruits of molecular physiology: engineering the L-isoleucine biosynthesis pathway in *Corynebacterium glutamicum*. J. Biotechnol. 56, 167–182.

Frangeul, L., Nelson, K.E., Buchrieser, C., Danchin, A., Glaser, P., Kunst, F., 1999. Cloning and assembly strategies in microbial genome projects. Microbiology 145, 2625–2634.

Grant, S.G.N., Jessee, J., Bloom, F.R., Hanahan, D., 1990. Differential plasmid rescue from transgenic mouse DNAs into *Escherichia coli* methylation-restriction mutants. Proc. Natl. Acad. Sci. USA 87, 4645–4649.

Hodgson, J., 1998. LION and Degussa apply genomics to fermentation. Nature Biotechnol. 16, 715.

Hoheisel, J.D., Maier, E., Mott, R., McCarthy, L., Grigoriev, A.V., Schalkwyk, L.C., Nizetic, D., Francis, F., Lehrach, H., 1993. High resolution cosmid and P1 maps spanning the 14 Mb genome of the fission yeast *S. pombe*. Cell 73, 109–120.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun,

M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.K., Codani, J.J., Connerton, I.F., Cummings, N.J., Daniel, R.A., Denizot, F., Devine, K.M., Dusterhoft, A., Ehrlich, S.D., Emmerson, P.T., Entian, K.D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.Y., Glaser, P., Goffeau, A., Golightly, E.J., Grandi, G., Guiseppi, G., Guy, B.J., Haga, K., Haiech, J., Harwood, C.R., Henaut, A., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M.F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.M., Levine, A., Liu, H., Masuda, S., Mauel, C., Medigue, C., Medina, N., Mellado, R.P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.H., Parro, V., Pohl, T.M., Portetelle, D., Porwollik, S., Prescott, A.M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Seror, S.J., Serror, P., Shin, B.S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, T., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanaka, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.F., Zumstein, E., Yoshikawa, H., Danchin, A., 1997. The complete genome sequence of the Gram-posiure bacterium *Bacillus subtiles*. Nature 390, 249–256.

Leuchtenberger, W., 1996. Amino acids-technichal production and use. In: Rehm, H.-J., Reed, G., Pühler, A., Stadler, P. (Eds.), Biotechnology, vol. 6. VCH, Weinheim, Germany, pp. 465–502.

Monaco, A.P., Larin, Z., 1994. YACs, BACs, PACs and MACs: artificial chromosomes as research tools. Trends Biotechnol. 12, 280–286.

Peter, H., Weil, B., Burkovski, A., Krämer, R., Morbach, S., 1998. *Corynebacterium glutamicum* is equipped with four secondary carriers for compatible solutes: identification, sequencing, and characterization of the proline/ectoine uptake system, ProP, and the ectoine/proline/glycine betaine carrier, EctP. J. Bacteriol. 180, 6005–6012.

Quast, K., Bathe, B., Pühler, A., Kalinowski, J., 1999. The *Corynebacterium glutamicum* insertion sequence IS*Cg2* prefers conserved target sequences located adjacent to genes involved in aspartate and glutamate metabolism. Mol. Gen. Genet. 262, 568–578.

Redenbach, M., Kieser, H.M., Denapaite, D., Eichner, A., Cullum, J., Kinashi, H., Hopwood, D.A., 1996. A set of ordered cosmids and detailed genetic and physical map for the 8 Mb *Streptomyces coelicolor* A3(2) chromosome. Mol. Microbiol. 21, 77–96.

Sahm, H., Eggeling, L., Eikmanns, B., Krämer, R., 1995. Metabolic design in amino acid producing bacterium *Corynebacterium glutamicum*. FEMS Microbiol. Rev. 16, 243–252.

Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. Molecular Cloning: A Laboratory Manual, second ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schäfer, A., Tauch, A., Droste, N., Pühler, A., Kalinowski, J., 1997. The *Corynebacterium glutamicum cglIM* gene encoding a 5-cytosine methyltransferase enzyme confers a specific DNA methylation pattern in an McrBC-deficient *Escherichia coli* strain. Gene 203, 95–101.

She, Q., Confalonieri, F., Zivanovic, Y., Medina, N., Billaut, A., Awayez, M.J., Thi-Ngoc, H.P., Thi Pham, B.-T., van der Oost, J., Duguet, M., Garrett, R.A., 2000. A Bac library and paired-PCR approach to mapping and completing the genome sequence of *Sulfolobus solfataricus* P2. DNA. Seq. 11, 183–192.

Sheng, Y., Mancino, V., Birren, B., 1995. Transformation of *Escherichia coli* with large DNA molecules by electroporation. Nucleic Acids Res. 23, 1990–1996.

Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, T., Simon, M., 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using F-factor-based vector. Proc. Natl. Acad. Sci. USA 89, 8794–8797.

Tauch, A., Kassing, F., Kalinowski, J., Pühler, A., 1995. The *Corynebacterium xerosis* composite transposon Tn5432 consists of two identical insertion sequences, designated IS1249, flanking the erythromycin resistance gene *ermCX*. Plasmid 34, 119–131.

Tauch, A., Wehmeier, L., Götker, S., Pühler, A., Kalinowski, J., 2001. Relaxed *rrn* expression and amino acid requirement of a *Corynebacterium glutamicum rel* mutant defective in (p)ppGpp metabolism. FEMS Microbiol. Lett. 201, 53–58.

Wang, K., Boysen, C., Shizuya, H., Simon, M.I., Hood, L., 1997. Complete nucleotide sequence of two generations of a bacterial artificial chromosome cloning vector. BioTechniques 23, 992–994.

Wehmeier, L., Schäfer, A., Burkovski, A., Krämer, R., Mechold, U., Malke, H., Pühler, A., Kalinowski, J., 1998. The role of the *Corynebacterium glutamicum rel* gene in (p)ppGpp metabolism. Microbiology 144, 1853–1862.

Woo, S.S., Jiang, J., Gill, B.S., Paterson, A.H., Wing, R.A., 1994. Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. Nucleic Acids Res. 22, 4922–4931.